Computational methods for rapid structural modelling of antibody-antigen interactions to improve identification of antigen-specific antibodies from BCR-seq data

Eve Richardson
St. Hilda's College



University of Oxford
A thesis presented for the degree of
Doctor of Philosophy

Michaelmas 2022

Copyright © 2022 by Eve Richardson
All Rights Reserved

Declaration

I declare that none of the work presented in this thesis has been reproduced or accepted for any other degree at any other university. All the work presented is my own unless otherwise stated, where the contributors have been named.

Eve Richardson, Michaelmas 2022

This thesis is dedicated to my Mum. My Mum did a lot of her PhD in the evenings, when her three young children were asleep. I hope to one day have just a fraction of her strength.

Abstract

Antibodies are immune proteins that are the basis of humoral immunity in jawed vertebrates, permitting highly-specific, mutable and lasting recognition of diverse foreign molecules. In their membrane-bound form, they are referred to as B cell receptors (BCRs): the collection of antibodies or BCRs in an individual is a record of their antigenic history, and tells us how their B cells have recognised and interacted with pathogens over their lifespan. Since the advent of next-generation sequencing and its application to these repertoires, we can sample and sequence the receptors of 10^3 to 10^6 B cells from a given individual. However, there remains an unmet need for mapping the mounting number of these BCR sequences to their putative antigen specificity. In this thesis, we aim to demonstrate how computational structural methods can be applied to BCR repertoires to improve our ability to identify antigen-specific antibodies.

In the first chapter, we describe a novel method for identifying antigen-specific antibodies from repertoire data using paratope prediction which we call "paratyping", and then use paratyping to discover novel Pertussis toxoid-binding antibodies from the BCR repertoires of transgenic mice.

Transgenic mice are a common source of therapeutic antibodies. To improve our understanding of the structural landscape of the BCR repertoires of these workhorses of antibody discovery, in the second chapter we apply structural annotation and modelling approaches to naive repertoires from humans, mice and transgenic mice. We show that the starting structural repertoires of transgenic mice are intermediate between humans and mice, despite being encoded by human genes, as a result of deficient junctional diversification in the primary repertoire of transgenic mice.

In the third chapter, we describe the creation of a novel sequence database for Ebolavirus-binding antibodies called EBOV-AbDab. Using this data alongside other techniques, we analysed the longitudinal BCR repertoires of 40 in-

dividuals vaccinated with Ad26.ZEBOV/MVA-BN-Filo. We used paratyping to predict the epitopes of the majority of the most expanded clonotypes after the booster vaccine, providing evidence of a highly convergent response to Ebolavirus vaccination that correlated with anti-Ebolavirus glycoprotein IgG titre.

Finally, in the fourth chapter we examine the growing body of natively-paired data to try to understand VH:VL pairing preferences. We re-examine the VH:VL interface in the context of germline gene usage and subsequently analyse at scale the genetic pairing preference of antibodies in the largest paired sequencing repository, the Observed Antibody Space, in combination with 3x as much novel data produced by an alternative single-cell sequencing technology.

In the final chapter, I outline further work which can extend or complete the work presented in the previous chapters, as well as general outlooks for the field.

Table of Contents

1 Introduction **1**

1.1 Antibody genetics and structure 1

1.1.1 Antibody genetics and protein structure 1

1.1.2 B cell development 5

1.1.3 Antibody function 9

1.2 Antibody repertoire sequencing 11

1.2.1 Isolation of B cells 12

1.2.2 Library preparation 13

1.2.3 Bulk vs. paired sequencing 15

1.2.4 Sequencing 17

1.2.5 Repertoire analysis 18

1.2.5.1 Processing of raw reads 18

1.2.5.2 Germline databases 19

1.2.5.3 Germline alignment 20

1.2.5.4 Clonotypes 22

1.2.6 Structural annotation of BCRs 24

1.2.6.1 Antibody numbering 25

1.2.6.2 Complementarity-determining regions 26

1.2.6.3 Canonical forms 27

1.2.7 Antibody structure prediction 28

1.2.7.1 Homology-based and *ab initio* modelling 28

1.2.7.2 Next-generation deep learning methods 30

1.2.8 Structural annotation 32

1.3 The antibody-antigen interface 33

1.3.1 Wet-lab techniques for epitope mapping 34

1.3.1.1 Competition assays 34

1.3.1.2 Mutagenesis 34

1.3.1.3 Hydrogen-deuterium exchange mass spectrometry 35

1.3.1.4 Display technologies 35

1.3.1.5 Antibody/antigen complex imaging 36

1.3.2 Computational techniques . . . 37
1.3.2.1 Paratope prediction . . . 37
1.3.2.2 Epitope prediction . . . 40
1.3.2.3 Antibody-antigen docking . . . 42
1.3.2.4 Computational epitope binning . . . 43
1.4 Antibody databases . . . 45
1.4.1 Sequence databases . . . 45
1.4.2 Antigen-specific antibody databases . . . 45
1.4.3 Antibody structure databases . . . 46
1.5 Antibody discovery . . . 46
1.5.1 mAbs derived from in vivo systems . . . 46
1.5.1.1 Humanisation of murine antibodies . . . 46
1.5.1.2 Humanised mice . . . 47
1.5.1.3 Hybridoma and cell culture . . . 48
1.5.2 Display technologies . . . 49
1.5.3 BCR-seq and antibody discovery . . . 50
1.6 Publications and preprints . . . 52

2 A novel method that uses high-throughput paratope prediction to identify antigen-specific antibodies from repertoire data 53
2.1 Abstract . . . 53
2.2 Introduction . . . 54
2.3 Materials and Methods . . . 56
2.3.1 Datasets . . . 56
2.3.1.1 PTx datasets . . . 56
2.3.1.2 Single-cell dataset . . . 56
2.3.1.3 Bulk dataset . . . 57
2.3.1.4 SAbDab dataset . . . 57
2.3.1.5 CoV-AbDab benchmark dataset . . . 58
2.3.2 Methods . . . 58
2.3.2.1 Alignment and numbering . . . 58
2.3.2.2 Clonotyping . . . 59
2.3.2.3 Paratope prediction . . . 59
2.3.2.4 Structure prediction . . . 59

2.3.2.5 Paratope identity calculation 60
2.3.2.6 Repertoire mining experiment 60
2.3.2.7 Experimental validation of predicted binders . 61
2.3.2.8 Performance evaluation 63
2.3.2.9 In silico developability assessment 64
2.3.2.10 Ab-Ligity and SPACE 64
2.4 Results 65
2.4.1 Epitope convergence can be identified at the level of paratope residues 65
2.4.2 Paratyping and clonotyping successfully cluster PTx binders in a single-cell dataset 66
2.4.3 Paratyping enables the identification of novel anti-PTx antibodies from different clonotypes to known binders in prospective experiment 70
2.4.3.1 Paratyping and clonotyping identify different binders in a bulk repertoire experiment 70
2.4.3.2 Prospective experimental validation of novel PTx-binding sequences 73
2.4.3.3 Discovery of novel anti-PTx antibodies from different clonotypes 74
2.4.3.4 Repertoire mining can improve in silico developability metrics 76
2.4.4 Paratyping groups antibodies to the same epitope in five antigen test cases 79
2.4.5 Paratyping in the context of new methods 81
2.5 Discussion 84
2.6 Chapter summary 90

3 Applying structure prediction to human, mouse and transgenic mouse BCR repertoires 93
3.1 Abstract 93
3.2 Introduction 93
3.3 Materials and Methods 95
3.3.1 Datasets 95

3.3.1.1 Human BCR sequencing dataset 95
3.3.1.2 Kymouse BCR sequencing dataset 95
3.3.1.3 Mouse BCR sequencing dataset 96
3.3.2 Methods 96
3.3.2.1 Sequence alignment and analysis and clonal inference 96
3.3.2.2 Structural annotation 97
3.3.2.3 Structural clustering 99
3.3.2.4 Structural comparison 99
3.3.2.5 Humanness scoring of VH sequences 99
3.4 Results 100
3.4.1 Kymouse and human repertoires are distinguishable on the basis of sequence properties 100
3.4.1.1 Germline gene usage 100
3.4.1.2 CDRH3 sequence analysis 102
3.4.2 Kymouse, mouse and human repertoires can be compared via structural annotation and modelling 106
3.4.2.1 Homology modelling coverage inversely correlates with the number of junctional insertions . 106
3.4.2.2 The Kymouse repertoires are structurally more human-like than mouse-like 108
3.4.2.3 Comparison of structural distance via SAAB+ and ABlooper 114
3.4.3 A state-of-the art humanization tool scores Kymouse sequences as fully human 116
3.5 Discussion 117
3.6 Chapter summary 121

4 Paratyping and clonotyping applied to BCR repertoires following Ebolavirus vaccination 123
4.1 Abstract 123
4.2 Introduction 123
4.3 Methods 129
4.3.1 Trial and samples 129

4.3.2 BCR sequencing . . . 131
4.3.3 Immunogenicity data . . . 131
4.3.4 Sequence analysis . . . 131
4.3.5 EBOV-AbDab curation . . . 132
4.3.6 Other antigen databases . . . 133
4.3.7 Human OAS dataset . . . 134
4.3.8 Comparison to antigen databases including EBOV-AbDab 134
4.3.9 Structural modelling . . . 135
4.3.10 Ab-Ligity for structural paratope comparisons . . . 136
4.3.11 Statistics . . . 136
4.4 Results . . . 136
4.4.1 BCR repertoires of vaccinees show evidence of antigen exposure . . . 136
4.4.2 Analysis of an Ebolavirus-binding antibody database . . 141
4.4.3 Ebolavirus-binding antibodies appear more frequently in vaccinee repertoires than in healthy repertoires . . . 143
4.4.4 EBOV GP-reactive serum IgG titres correlate with repertoire overlap with EBOV-AbDab . . . 147
4.4.5 Analysis of the hit clonotypes from vaccinees . . . 149
4.4.6 Comparison of EBOV-AbDab to the Human OAS dataset 152
4.4.7 Comparison to 6666/6662/3T0245 to influenza antibodies 154
4.5 Discussion . . . 155
4.6 Chapter summary . . . 160

5 Sequence and structural analysis of VH:VL pairing 162
5.1 Abstract . . . 162
5.2 Introduction . . . 162
5.3 Methods . . . 164
5.3.1 Human structural dataset . . . 164
5.3.2 Human sequencing datasets . . . 165
5.3.2.1 Human Paired data from the Observed Antibody Space . . . 165
5.3.2.2 Frater/Dustin single-cell sequencing data . . . 165
5.3.3 Structural analysis . . . 168

5.3.3.1 Contact analysis . . . 168
5.3.3.2 Energetics . . . 169
5.3.3.3 VH:VL orientation . . . 170
5.3.3.4 Elbow angle . . . 171
5.3.4 Statistical analysis . . . 171
5.3.4.1 Identification of over- and underrepresented gene pairings . . . 171
5.3.4.2 Other statistical tests . . . 172
5.4 Results . . . 173
5.4.1 The VH:VL interface is dominated by CDR3-mediated hydrophobic interactions . . . 173
5.4.2 Analysis of interfaces on the basis of genetic origin suggests positions that may be differentially involved . . . 177
5.4.3 Kappa and lambda antibodies differ significantly in some interface properties . . . 180
5.4.4 A number of gene pairings are over- or underrepresented, but we are unable to correlate with this with predicted interface energetics . . . 188
5.4.5 Crystal structures of human lambda antibodies are biased towards longer CDRH3s in comparison to sequencing datasets . . . 190
5.4.6 VH:VL pairing in a novel single-cell sequencing dataset . 191
5.5 Discussion . . . 193
5.6 Chapter summary . . . 198

6 Conclusions and future work **199**
6.1 Conclusions . . . 199
6.2 Future work . . . 200

7 Appendices **204**

Bibliography **214**

List of Figures

1.1 The BCR complex and antibody structure 2
1.2 Variable domain structure in relation to genetics 6
1.3 Simplified diagrams of the human IgH, IgK and IgL loci 7
1.4 BCR development and V(D)J recombination 9
1.5 Schematic diagram of isotypes and isotype subtypes 11
1.6 Schematic of current antibody structural annotation techniques 33

2.1 Method of calculating predicted paratope identity 60
2.2 Graphical illustration of the process of repertoire mining via paratyping 62
2.3 Example 1 of two antibodies that bind to the same epitope but derive from different clonotypes 66
2.4 Example 2 of two antibodies that bind to the same epitope but derive from different clonotypes 67
2.5 IGHV, IGKLV and SHM frequencies and CDRH3 length distribution in known PTx binders 68
2.6 Precision-recall curves for clonotyping and paratyping in the task of predicting PTx binding 71
2.7 Representative dendrograms from the single-cell repertoire mining experiment 72
2.8 CDRH3, total CDR and total VH amino acid identity of novel PTx-binding antibodies to the known PTx-binding antibody by which they were identified 75
2.9 Dendrograms showing two examples of immune repertoire mining, using known PTx binders CL-95906 and CL-95940 76
2.10 An example of a probe antibody with a flagged developability issue significantly improved by immune repertoire mining . . . 79
2.11 Three further examples of probe antibodies with flagged developability issues 80

3.1 IGHV, IGHD and IGHJ usage in the Kymouse versus human repertoires 101

3.2 Comparison of IGHJ usage between naive human and Kymouse IgM BCR repertoires . . . 102
3.3 The CDRH3 length distribution in murine, Kymouse and human sequences . . . 103
3.4 VD and DJ insertion length distributions in the Kymice and humans 103
3.5 Examination of diversity of human, Kymouse and murine repertoires . . . 105
3.6 Coverage by SAAB+ and ABlooper as a function of CDRH3 length 108
3.7 Canonical class usage in Kymouse, mouse and human repertoires 109
3.8 CDRH3 structural clusters observed by repertoire type . . . 110
3.9 Hierarchical clustering on SAAB+ CDRH3 structural usage . . . 112
3.10 Hierarchical clustering of ABlooper CDRH3 structural cluster usage . . . 113
3.11 Hierarchical clustering on SAAB+ CDRH3 structural cluster usage in the subset modelled by ABlooper . . . 115
3.12 Comparison of ABlooper and SAAB+ predictions . . . 116
3.13 RMSD between ABlooper models and SAAB+ CDRH3 structural cluster centres . . . 117

4.1 Overview of the domains of EBOV GP and representation of its structure . . . 125
4.2 EBOV GP colored by amino acid conservation across Ebolavirus species . . . 126
4.3 EBOV-AbDab curation . . . 133
4.4 Sensitivity of clonal diversity measure to the number of clonotypes included in the calculation . . . 138
4.5 Sensitivity of three clonal expansion measures to subsampling . 139
4.6 Evidence of repertoire perturbation in vaccinees with respect to clonality, CDRH3 length and isotype subtype frequency . . . 140
4.7 Summary statistics for EBOV-AbDab and CoV-AbDab . . . 144
4.8 Increase in hits to EBOV-AbDab post-boost . . . 146
4.9 Correlation between EBOV-AbDab hit rate and serological data 148
4.10 Clonal expansion in relation to serological data . . . 149
4.11 Convergent clonotypes and hits to EBOV-AbDab . . . 151

5.1 ABangle diagram . . . 171
5.2 Interaction types in the VH:VL interface . . . 174
5.3 Commonness of VH:VL residue pairings in the interface . . . 175
5.4 VH interface positions by IGHV subgroup . . . 179
5.5 VL interface positions by kappa or lambda locus . . . 180
5.6 Positions whose inclusion in the interface depends on encoding gene . . . 181
5.7 Entropy of VH:VL positions in interface . . . 182
5.8 Visualisation of two antibodies with opposite degrees of CDRH3/CDRL3 contact . . . 182
5.9 Length distributions of lambda and kappa-paired CDRH3s and lambda and kappa CDRL3s . . . 183
5.10 Difference in VH:VL interface contacts according to kappa or lambda locus . . . 184
5.11 FoldX interaction energy differs between kappa and lambda antibodies . . . 186
5.12 VH:VL orientation differs between kappa and lambda antibodies 187
5.13 Difference in elbow angle distribution for kappa and lambda antibodies . . . 187
5.14 CDRH3 length distribution stratified by light chain locus in a human structural dataset and the Observed Antibody Space . . . 192

1 Cluster map of maximum amino acid identity between IGHV genes (IMGT, as of January 2023) . . . 205
2 Predicted paratope identity across CDRH1 and CDRH2 of germline alleles (IMGT, as of February 14th 2020) . . . 206
3 The IgG repertoires of subjects who received placebo did not show changes in expansion, CDRH3 length or IgG subtype frequencies . . . 207
4 Elbow angle calculation . . . 208
5 CDRH3 length distribution of kappa vs. lambda antibodies in diverse databases . . . 208
6 Interaction coefficient and difference in average interface energy 209

7 VH:VL interface energetics in models produced by ABlooper and IgFold 210
8 Kappa/lambda split in H1 canonical forms 210
9 Kappa/lambda split in H2 canonical forms 211
10 Kappa/lambda split in L1 canonical forms 211
11 Kappa/lambda split in L2 canonical forms 212
12 Kappa/lambda split in L3 canonical forms 212
13 RMSD distribution between pairs of CDRH3s of matched length between kappa and lambda antibodies 212

List of Tables

1.1 Propensities of different isotypes for Fc-mediated functions . . . 12
1.2 VH CDR definitions . . . 27
1.3 VL CDR definitions . . . 27
1.4 Performance of different paratope predictors in the same benchmark set . . . 40
1.5 Residue tokenization used by Ab-Ligity . . . 44

2.1 IMGT numbering ranges for CDRs 1 - 3 for heavy and light chains with IMGT or North definitions. . . . 59
2.2 Precision-recall values for prediction of PTx-binding . . . 70
2.3 Two pairs of antibodies from CoV-AbDab (May 2020) where the sequences derived from different clonotypes . . . 81
2.4 Evaluation of the hypothesis that paratyping groups antibodies which bind to the same epitope on five antigens from SAbDab (October 2020) . . . 82
2.5 Performance of SPACE, Ab-Ligity, paratyping and clonotyping at the task of grouping antibodies that bind to PTx . . . 83
2.6 Performance in terms of precision and recall of methods in epitope binning in CoV-AbDab . . . 85
2.7 Sequence identity of antibodies grouped by clonotyping, paratyping, SPACE and Ab-Ligity on the CoV-AbDab benchmark . . . 85

3.1 Percentages of sequences which could be annotated via SAAB+ 107
3.2 Numbers of sequences chosen to model using ABlooper . . . 108

4.1 Relevant Ebolavirus immunotherapeutics as of August 2022 . . . 127
4.2 Dosing intervals in vaccinees . . . 130
4.3 Ethnicity breakdown of vaccinees . . . 130
4.4 Formulae for clonal expansion metrics . . . 132
4.5 The number of antibodies to each of the GP protomers . . . 142
4.6 Non-6666-lineage hit antibodies in the repertoire data . . . 151

4.7 IGHV3-15 antibodies and their frequency in the Human OAS dataset . . . 153
4.8 Non 6666/6662/3T0245 antibodies in the Human OAS dataset . 153

5.1 Overview of Human Paired OAS datasets . . . 165
5.2 The number of non-singleton and therefore highest confidence clonotypes obtained for each of 23 subjects (pooled across longitudinal time points for the vaccinees given little evidence of an antigen-specific response in terms of clonal expansion or similarity to CoV-AbDab); samples used were from subjects vaccinated with ChAdOx-1/nCoV-19 (6), a subject vaccinated with MenACWY and 16 subjects who were PCR positive for COVID in 2020. . . . 168
5.3 FoldX interaction terms . . . 170
5.4 The proportion of the interface in each VH and VL region . . . 173
5.5 The ten most highly conserved VH:VL interactions . . . 176
5.6 Hydrogen bonds in the VH:VL interface . . . 176
5.7 Ionic interactions in the VH:VL interface . . . 177
5.8 The number of significant over- or underrepresentations of gene pairings . . . 190
5.9 The significant interaction terms in each of the four gene-wise interaction models . . . 191
5.10 Significant gene pairings and their associated expectation values 193

1 VH interface positions seen in more than 50% of complexes in the human structural dataset . . . 204
2 VL interface positions seen in more than 50% of complexes in the human structural dataset . . . 213

1 | Introduction

1.1 Antibody genetics and structure

Antibodies are immune proteins that are the basis of humoral immunity in jawed vertebrates, permitting highly-specific, mutable and lasting recognition of diverse foreign molecules. They are the secreted form of the antigen receptor component of the B cell receptor (BCR) complex (Figure 1.1). Antigen recognition by the BCR complex leads to, in combination with T-cell help for protein antigens, activation of B cells and differentiation into either memory B cells or antibody-secreting cells (ASCs) which can produce between 10 and 1,000 antibodies per second (Nguyen et al., 2019). In blood, lymph and the mucosae, antibodies enable recognition and destruction of pathogens with diverse effector functions. Antibodies are a critical component of adaptive immunity: people who suffer from B cell deficiencies suffer more frequently and more chronically from infectious diseases, in particular bacterial infections (Groth et al., 2020). The extreme specificity of antibodies in combination with their range of effector functions has made them the most successful class of biologics, used for treating diverse indications including cancers, autoimmune diseases and infectious diseases (Kaplon et al., 2022).

1.1.1 Antibody genetics and protein structure

The antibody monomer consists of a heavy chain and a light chain. The heavy and light chain each consist of a single variable domain, which mediates antigen recognition, and a number of constant domains (one for light chains, and three or four for heavy chains) which determine the effector function. These two parts of the antibody are referred to as the variable and constant regions.

The constant region is the product of a single constant gene; meanwhile, the variable regions are encoded by three (heavy chain) or two (light chain) genes. These genes are found in the IGH locus (on chromosome 14 in humans) for the heavy chain. The light chain can be encoded by one of two loci each with

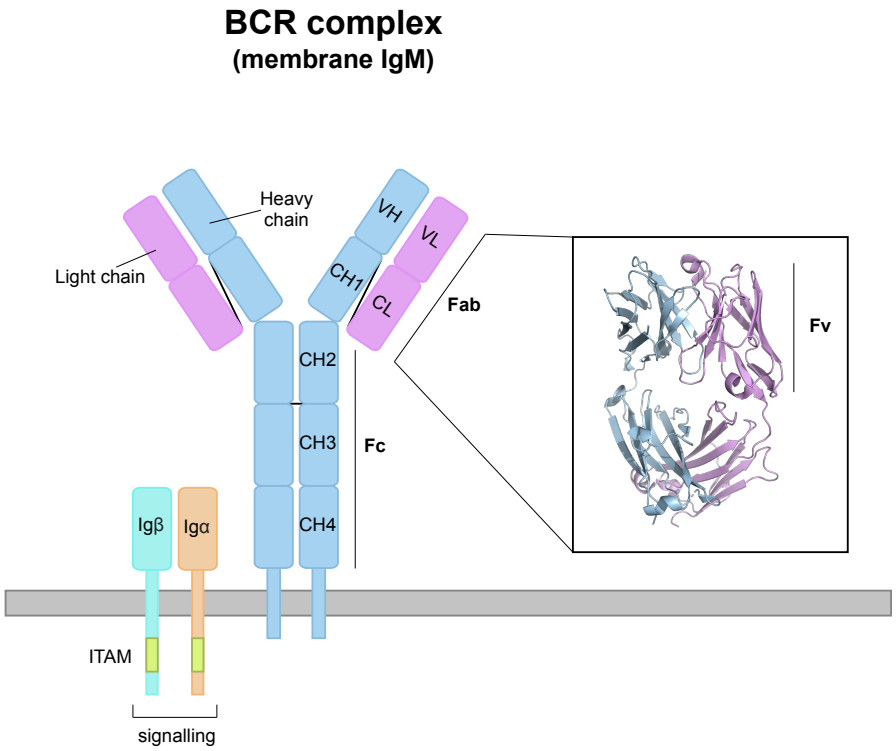

Figure 1.1: | **The BCR complex and antibody structure.** A schematic of the B cell receptor (BCR) complex which consists of the BCR itself, the antigen receptor, and the Igα/Igβ dimer which is vital for both successful membrane expression and signal transduction. Signal transduction is achieved via the immunoreceptor tyrosine-based activation motif (ITAM). The antibody is the secreted form of the BCR without the transmembrane and cytosolic domains. The antibody can be split into the Fc (fragment crystallisable) and Fab domains (fragment antigen-binding) domains. The Fab domain consists of the variable domains (the Fv) and one constant domain for each of the chains. The three-dimensional structure of the Fab, with the Fv highlighted, is shown in cartoon rendering (PDB ID: 4eow_HL).

their own complement of variable- and constant region genes, referred to as the kappa (on chromosome 2) and lambda (chromosome 22) loci. A single chain can only be encoded by genes on the same chromosome, i.e. a light chain will be encoded by either all kappa or lambda genes (Janeway et al., 2001).

The heavy chain variable region (VH) is encoded by a variable gene (IGHV), a diversity gene (IGHD) and a joining gene (IGHJ). The light chain variable region (VL) is encoded by a variable gene (IGKLV, K or L referring to the separate kappa and lambda loci) and a joining gene (IGLJ). These genes recombine (the heavy chain first, forming a pre-BCR with a surrogate light chain) in developing B cells in a process referred to as V(D)J recombination. This is the first step in the generation of the extreme diversity observed in the antibody repertoire: combinatorial diversity formed by the combining of one of each of the multiple genes. There is also significant diversity at the population level in terms of allelic and haplotype diversity, as well as copy number variation and structural variations of each of the loci (Gidoni et al., 2019; Jackson et al., 2012; Scheepers et al., 2015; Watson et al., 2017; Watson et al., 2013).

As the genes recombine, there is significant diversity added due to the addition of palindromic (P) nucleotides (after DNA repair of overhang of the ligating ends) and non-templated (N) nucleotides (by the activity of the enzyme terminal deoxynucleotidyl transferase (TdT)). These added nucleotides at the junctions of IGHV/IGHD, IGHD/IGHJ and IGKLV/IGKLJ are referred to as NP nucleotides. There are significantly fewer N nucleotides added in light chains due to the lower activity of TdT during light chain recombination (Janeway et al., 2001).

There is further or "secondary" diversification upon antigen exposure, for B cells that migrate to the lymphoid organs and form germinal centres (GC) (described further in Section 1.1.2). In activated B cells, the activity of activation-induced cytidine deaminase (AID) increases both the diversity of the variable domain and of the constant region-mediated effector functions by deaminating

cytosine into uracil in DNA that is being transcribed. The subsequent lesions in the DNA result in somatic hypermutation (SHM), gene conversion and class-switch recombination (CSR).

Somatic hypermutation introduces point mutations into the Ig loci at a rate of around 10^{-3} per base pair per division (Kleinstein et al., 2003; McKean et al., 1984) with a clear bias for transitions over transversions (at a ratio of 1.3) (Dörner et al., 1997; Messmer et al., 2004). In addition to clear substitution biases, the targeting is non-random and depends on microsequence context (resulting in mutation hotspots/coldspots) as well as position within the variable gene (Cowell & Kepler, 2000; Cui et al., 2016; Krantsevich et al., 2021; Schramm & Douek, 2018; Smith et al., 1996; Spisak et al., 2020; Tang et al., 2022; Yaari et al., 2013; Zhou & Kleinstein, 2020). AID activity can also result in a process referred to as gene conversion or replacement where parts of the V gene segment are replaced with sequences from another gene segment (Arakawa et al., 2002; Harris et al., 2002). AID mediates class switch recombination (CSR), whereby the heavy chain locus recombines further, deleting the IgM gene segment and replacing the constant region with a downstream constant region segment. This results in switching of the isotype of antibody produced by the cell (Janeway et al., 2001).

The complex recombination process of the variable region genes results in the VH and VL; these domains can then be mutated further during secondary diversification of the antibody repertoire. The VH and VL form a heterodimer referred to as the Fv, shown in Figure 1.2. Both heavy and light chain variable domains have the standard immunoglobulin fold, i.e. a two-layered structure consisting of seven anti-parallel β-strands (Figure 1.2) stabilised by a conserved disulphide bond. Unlike other structures with an immunoglobulin fold, each variable domain has three hypervariable loops which are called complementarity-determining regions (CDRs) (Chothia & Lesk, 1987). CDRH1 and CDRL1 link adjacent β-sheets, while the remainder are β-turns. The most diverse of these loops in terms of both sequence and structure is CDRH3 (Marks & Deane, 2020; Regep et al., 2017; Weitzner et al., 2015): the sequence diversity of the CDRH3

is explained by its position within the VH exon, spanning the IGHV, IGHD and IGHJ gene junctions, where there are two junctions for NP addition. The CDRH3 appears to have a special role in determining antigen-specificity as a result (Xu & Davis, 2000).

The VH and VL are each associated with constant regions, which determine the effector function of the antibody. There are nine human heavy chain constant genes (IGHC) downstream of the variable region genes, which can be divided into five subclasses which encode five isotypes: IgM, IgA, IgG, IgE and IgD (encoded by IGHM, IGHA, IGHG, IGHE and IGHD genes) (Figure 1.3). There are four subclasses of IgG (IgG1 through IgG4) and two subclasses of IgA (IgA1 and IgA2). There is a single kappa chain constant gene, IGKC, with a similar locus arrangement to the IGH locus (Collins & Watson, 2018). The lambda locus contains between seven and eleven IGLC genes: the lambda locus differs significantly from these IGH and IGK loci as each of these IGLC genes are preceded by a single IGLJ gene. Simplifications of these loci are represented in Figure 1.3.

1.1.2 B cell development

A schematic diagram of B cell development is shown in Figure 1.4. B cells begin their development in the bone marrow from common lymphoid progenitors (CLPs). The CLP gives rise to the earliest B-lineage cell, the pro-B cell, in which V(D)J recombination begins. After recombination of the D- and JH segments, the cell is called a late pro-B cell. The transition from pro-B cell to the next stage, the large pre-B cell, occurs after recombination of the V with the -DJ segment. The V(D)J chain is expressed with an IgM constant chain as a pre-B cell receptor with the surrogate light chain. Expression of the pre-BCR results in signalling that enforces allelic exclusion, whereby the recombined V(D)J region of only a single chromosome is transcribed to prevent dual receptor specificity. This is achieved, partly, by reduction in recombinase activity by direct downregulation of RAG-1 and RAG-2. Single-cell transcriptomics experiments have however found evidence of B cells expressing BCRs derived from multiple V(D)J recom-

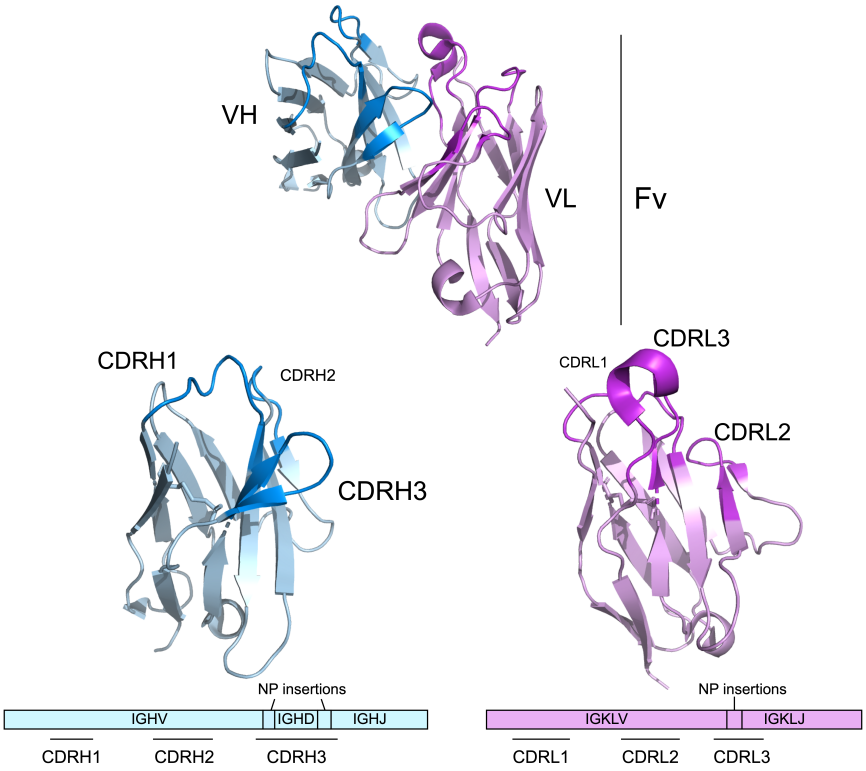

Figure 1.2: | **Variable domain structure in relation to genetics.** Shown are cartoon renderings of the Fv structure, with the heavy chain in blue and light chain in pink. The relation of the highlighted CDRs, which are shown in darker blue and pink respectively, to the genes encoding these domains are shown. CDRs 1 and 2 are encoded by the variable (V) genes. CDRL3 encompasses the end of the IGKLV gene, the IGLJ gene and the junction of the two genes. CDRH3 is encoded by the IGHV gene, the IGHD gene and IGHJ gene, and the two junctions here. Due to the contribution of the diversity gene (IGHD), greater number of junctions and higher rate of non-templated (N) nucleotide addition, the CDRH3 is considerably more diverse than the CDRL3.

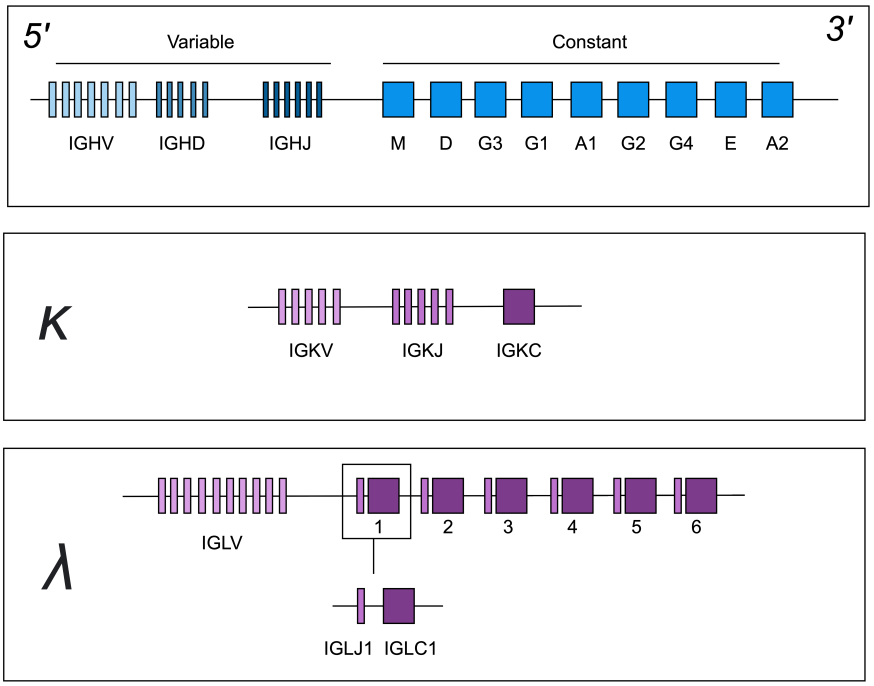

Figure 1.3: | **Simplified diagrams of the human IgH, IgK and IgL loci.** Simplified diagrams of the canonical locus representations of the IgH, IgK and IgL loci are shown as per the IMGT (Lefranc et al., 2003). Only representations of functional gene exons are shown and distances are non-representative. The number of genes pictured corresponds to the number of IMGT-recognised gene subgroups as of 14th September 2022. For example, there are seven IGHV gene segments drawn to reflect the seven human IGHV subgroups - it should be noted that genes from different subgroups are interspersed. The heavy chain (blue) and kappa loci are fundamentally similar in that they consist of separate clusters of variable domain genes upstream of the constant region gene(s). The human lambda locus is different and consists of the IGLV cluster and then multiple mini-cassettes of IGLJ/IGLC genes (Collins & Watson, 2018).

bination events, however multiple heavy chains per cell have not been reported by other studies (Shi et al., 2019). The large pre-B cell undergoes rounds of cell division, meaning that a given large pre-B cell subsequently results in multiple small pre-B cells. At the small pre-B cell stage, RAG-1/2 expression increases again, and the light chain recombines separately in each of these descendant cells. The first recombination attempts ordinarily occur on the kappa locus in mice and humans. Multiple attempts at the production of an intact light chain can occur; if such attempts fail, the lambda locus recombines (ordinarily following a non-productive kappa rearrangement, though not always) (Collins & Watson, 2018). After successful recombination of a light chain, and allelic and isotypic exclusion, such cells are classified as immature B cells. The light chain is then expressed, with the heavy chain, as a surface or membrane IgM (mIgM). When the heavy chain first recombines, the constant region is always IgM, and remains IgM until the B cell escapes the bone marrow and enters the periphery. In the periphery, the mature B cell also begins to express IgD (in mammals and fish) primarily via alternative splicing (as opposed to recombination); the reason for this co-expression is not clear (Geisberger et al., 2006). Subject to central tolerance, IgM+/IgD+ naive B cells leave the bone marrow and enter the periphery.

In the next developmental stage, B cells enter the spleen (with this stage referred to as "transitional") and either complete development as follicular (FO) or marginal zone (MZ) B cells. MZ B cells develop into short-lived plasma cells upon antigen exposure, while follicular cells form a germinal centre upon antigen and T-cell cross-linking and develop into long-lived memory or plasma cells. It is here that the BCR repertoire undergoes secondary diversification through the activity of AID and DNA repair mechanisms. Affinity may be improved by introduction of mutations in the variable domain and subsequent selection for higher-affinity variants, with survival controlled by access to limited helper T-cells and antigen in the GC (Janeway et al., 2001).

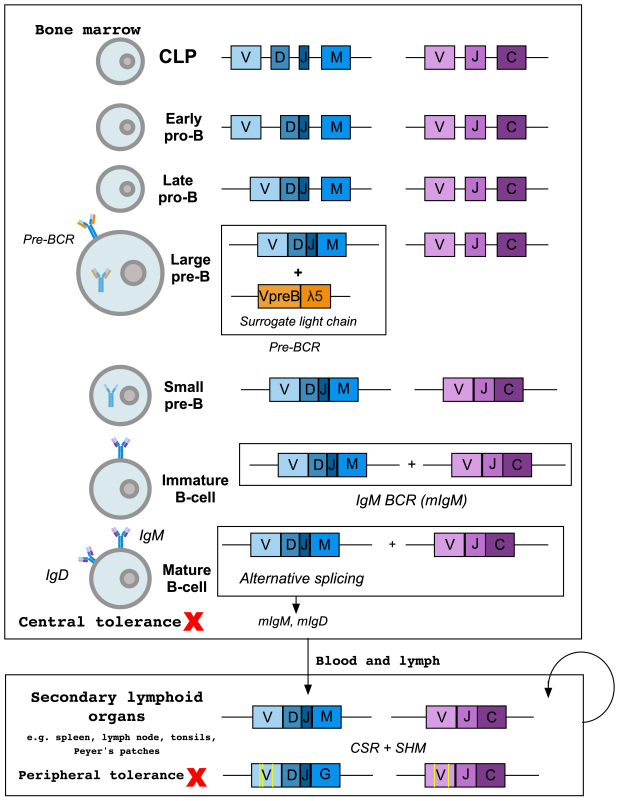

Figure 1.4: | **BCR development and V(D)J recombination.** V(D)J recombination and junctional diversification define the primary repertoire. As B cells develop from a common lymphoid progenitor (CLP), the IGH locus recombines first in the order D-J and then V-DJ. This V(D)J recombinant is expressed by the large pre-B cell as a surface IgM with a surrogate light chain encoded by the VpreB, also referred to as the iota chain, which as an Ig V-domain-like structure, and $\lambda 5$ which is an Ig C-domain-like structure. The large pre-B cell divides to form many small pre-B cells in which the IGK locus recombines (in humans). In the mature B cell, alternative splicing of the IGHM constant domain results in the co-expression of IGHM and IGHD. B cells which pass central tolerance checks in the bone marrow (i.e. are not autoreactive) move to the secondary lymphoid organs (SLOs), and circulate between the blood/lymph for 4 - 6 weeks until they die or encounter antigen. Upon activation, B cells proliferate in germinal centres and undergo SHM and CSR.

1.1.3 Antibody function

Antibodies are involved in the removal of pathogens by a number of means. Antibodies may either inhibit the infectivity of a pathogen or the toxicity of a

toxin by directly impeding function of a molecule, for example by blocking the receptor-binding domains required for cell entry (e.g. the ACE2-binding site on the SARS-CoV-2 spike receptor-binding domain). This is referred to as neutralisation and is independent of, though not mutually exclusive with, interaction with any other components of the immune system. Antibodies can also work by interacting with other cells in the immune system, targeting pathogens for clearance by effector cells such as macrophages, monocytes, natural killer (NK) cells, eosinophils, neutrophils and dendritic cells. Effector cells recognise the Fc regions of antibodies via Fc receptors (FcR). Opsonisation is an example of such behaviour, where pathogenic molecules are marked for phagocytosis. Certain classes of antibody can also initiate the complement cascade via the classical pathway, with the Fc interacting with Complement 1Q (C1Q). Antibodies can therefore trigger antibody-dependent cellular phagocytosis (ADCP), antibody-dependent cell-mediated cytoxicity (ADCC) and complement-dependent cytoxicity (CDC) (Reusch & Tejada, 2015), all mediated via the Fc.

As described in Section 1.1.1, the Fc is encoded by one of standardise constant gene segments, which can be divided into five isotypes. The five isotypes differ in number and position of disulfide bonds, the extent and nature of glycosylation, the length of the hinge region, the number of constant domains and multimerisation (Figure 1.5). These properties result in varying affinity of interaction with different FcRs and different binding stoichiometries and immune complex stability, which results in different effector functions (Lu et al., 2018). Furthermore, isotype has a significant effect on pharmacokinetic properties of the antibodies, such as half-life (Lobo et al., 2004). For example, IgG antibodies have particularly long half-lives due to their ability to be recycled by the neonatal Fc receptor (FcRn) (Stapleton et al., 2011), while antibodies with longer hinge regions are more susceptible to proteolysis (Brezski et al., 2009; Ryan et al., 2008). There is also evidence that the constant domain's effect on Fab angle and flexibility significantly affects antibody-antigen interactions as well as interactions with FcRs (McLean et al., 2002). An overview of the isotypes and their propensities for different functions is shown in Table 1.1.

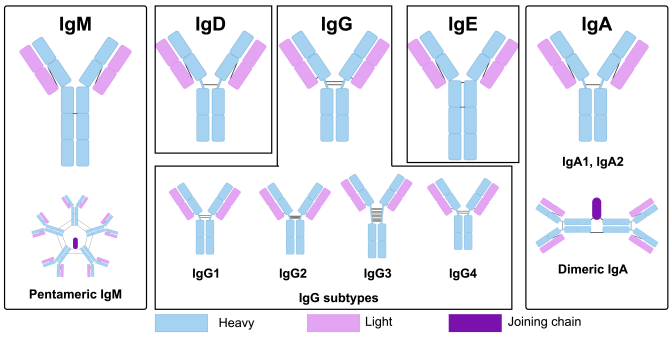

Figure 1.5: | **Schematic diagram of isotypes and isotype subtypes.** Schematic diagrams of each of the five mammalian isotypes, the four IgG isotype subtypes and two IgA isotypes.

1.2 Antibody repertoire sequencing

Next-generation immune repertoire sequencing (BCR-seq or Rep-seq (Georgiou et al., 2014; Yaari & Kleinstein, 2015) can provide comprehensive information about adaptive immune repertoires across individuals (Briney et al., 2019) and immune states (Bashford-Rogers et al., 2019). Progress has been made in the task of interrogating the vast diversity of B cell receptor (BCR) repertoires, primarily through the analysis of predicted clonal relationships inferred via clonotyping (Greiff et al., 2015).

Despite advances in *de novo* proteomics sequencing (Peng et al., 2021; Schulte et al., 2022), inferring the primary sequence of antibodies from serum via mass-spectrometry is still extremely limited in coverage and more time- and cost-intensive than cDNA sequencing. cDNA sequencing was first applied to the BCR repertoire in 2009 using 454 pyrosequencing (Boyd et al., 2009) and was coined as "Ig-seq": while some continue to refer to cDNA sequencing of BCR transcripts as Ig-seq (Kovaltsuk et al., 2018; Tian et al., 2020), others now use this term to refer to protein sequencing of serum antibodies (Avram et al., 2021; Lavinder et al., 2015). Next-generation sequencing of genetic material encod-

Function	Isotypes
Classical complement activation	IgM > IgG3 > IgG1 > IgG2
Binding to macrophage and phagocyte FcRs	IgG1, IgG3, IgG4, IgA1, IgA2, IgE
Binding to mast cell and basophil FcRs	IgE, IgG
Placental transfer	IgG1 > IgG3 > IgG2 > IgG4

Table 1.1: | **Propensities of different isotypes for Fc-mediated functions.** A number of Fc-mediated effector functions and the isotypes that are most strongly associated with them, in order ranked by Janeway's Immunobiology (Janeway et al., 2001). Complement cascade activation by interaction with C1q is most strongly associated with IgM but can also be observed for highly glycosylated, hexameric IgG complexes. IgG, IgA and IgE interact with macrophages and phagocytes such as dendritic cells, which as antigen-presenting cells (APCs) are critical for the induction of the T-cell response. IgE has the highest association with mast cell and basophils Fc receptors.

ing BCRs is referred to as BCR-seq here to avoid confusion; BCR-seq is a subset of AIRR-seq (Adaptive Immune Receptor Repertoire sequencing) and Rep-seq (Repertoire Sequencing) (Babrak et al., 2022; Kim & Park, 2019; Olsen, Boyles, et al., 2022). The experimental and computational steps required to sequence and analyse the BCR repertoire are outlined here.

1.2.1 Isolation of B cells

The first step in sequencing the BCR repertoire (apart from subject selection) is selection of the cells from which the genetic material is to be sequenced, which may require decisions about B cell compartment, cell type and antigen-specificity. The vast majority of B cell samples from human subjects are derived from peripheral blood: peripheral blood monocytes (PBMCs) of which B cells are a subset can be isolated by density centrifugation of peripheral blood (Ghraichy et al., 2020) and can be used as direct input for RNA extraction. Fine-needle aspiration of lymph nodes has been used to sample germinal centre B cells (Havenar-Daughton et al., 2020). Where applicable e.g. for neurodegenerative diseases, B cells have also been sampled from the cerebrospinal fluid via lumbar puncture (Lanz et al., 2022).

Specific B cell types can be isolated according to their expression of known

cell surface markers or binding to specific antigens via magnet- or fluorescence-activated cell-sorting (MACS or FACS). In MACS, magnetic beads are coated with antibodies to the cell surface marker of interest, and are incubated with the sample. This mixture is passed through a column with a strong magnetic field. Cells which are not bound to the magnetic beads pass through, allowing for selection of or removal of cells expressing a particular surface marker. In FACS, cells are incubated with fluorophore-coupled antibodies to cell markers of interest. The resultant cell-antibody conjugates are sorted into different populations according to their fluorescence intensity; this can be carried out with multiple cell surface markers, each recognised by a separate fluorophore-coupled antibody, to sort cells using complex gating strategies. Antigens of interest can also be fluorophore-conjugated in order to allow antigen-specific cell sorting, though it should be noted that this has a high false positive rate due to B cell binding to fluorophores and purification tags (Pedrioli & Oxenius, 2021).

1.2.2 Library preparation

Given a population of cells of interest, genetic material for sequencing is extracted. Either genomic DNA (gDNA) or messenger RNA (mRNA) can be the starting material for sequencing library preparation. While each has their own advantages, mRNA is more widely used (Babrak et al., 2022). gDNA provides a better estimate of clonality as, primer biases aside, the number of reads will be proportional to the number of gDNA molecules and therefore cells (Boyd et al., 2009; Maecker et al., 2012), while the relationship between cell count and mRNA transcript counts is confounded by differential expression across cell types and activation states, with plasma cells having up to 100-fold greater immunoglobulin transcription rates than naive cells (Klein et al., 1997). However, gDNA provides a lower concentration of genetic material (some of which will consist of non-productive rearrangements) and therefore more polymerase chain reaction (PCR) cycles are required to produce sufficient material for library preparation which introduces more error. mRNA also has an advantage in that the beginning of the constant region can be sequenced even using short-

read sequencing like Illumina technologies, because the switch region between the variable and constant genes is excised during transcription; for gDNA, the switch region extends the beginning of the constant region beyond reach. For error correction, unique molecular identifiers (UMIs) are now used to identify sequence derivations of the same starting mRNA or gDNA molecule which have been produced by PCR or sequencing error; while UMIs can be added in both mRNA and gDNA library preparation protocols, they are more common in mRNA preparations as this can be straightforwardly incorporated in the mRNA to cDNA synthesis step (Babrak et al., 2022).

As mentioned, gDNA and mRNA must both be amplified via PCR. mRNA must first be converted to cDNA via reverse transcription (RT) using constant-region specific primers which is an important bottleneck that can introduce bias (Schwaber et al., 2019; Zucha et al., 2020). Amplification of cDNA and gDNA is difficult because it generally requires the selection of primers for both termini; primer choice at the 3′ end is simple with a universal constant region primer, but sets of V-region specific primers are required for the 5′ end. Usage of complex primer sets in PCR (multiplex PCR) is not only technically difficult (often resulting in the formation of primer dimers) but also introduces bias in the amplification of particular cDNA species due to differing primer efficiencies. 5′ RACE (Rapid Amplification of cDNA Ends), also referred to as anchored PCR (Douek et al., 2002), can be used to amplify cDNA in a non-V-gene biased manner (Freeman et al., 2009; Li et al., 2013). RACE is used where only part of a sequence is known that can be the anchor for the PCR reaction - here, the constant region is the "known" sequence. An oligo is added to the 3′ end of the cDNA (e.g. a homopolymeric tail added by TdT) which acts as a "universal primer site". Thereafter, only a constant region primer in addition to the universal primer is required for amplification. At the PCR amplification stage, error can occur due to base misincorporation (proofreading DNA polymerases have error rates between 10^{-7} to 10^{-6} (Cline et al., 1996)) and template switching (which results in chimerisation of different template molecules (Georgiou et al., 2014)).

There are also sequencing platform-specific adaptors which are added during PCR. For Illumina paired-end sequencing, P7 and P5 adaptors are generally ligated to the amplicon during a secondary PCR reaction (Babrak et al., 2022). One or both of the adaptors may contain an index which can be used to tag cDNAs originating from the same sample to allow sample multiplexing.

1.2.3 Bulk vs. paired sequencing

In the described steps for library preparation for next-generation sequencing, the genetic material encoding the heavy and light chain expressed by a single cell (referred to as a "natively paired VH and VL") is amplified and sequenced together with the genetic material encoding the heavy and light chains of any other of the hundreds of thousands to millions of other cells in the sample. Thus, the native VH:VL pairing is lost during RNA or gDNA extraction; for this reason, such protocols are referred to as "bulk sequencing".

Knowing the cognate light chain for a particular heavy chain and vice-versa is highly desirable, as changing the light chain of an antigen-specific antibody can completely abrogate binding. Single-cell RT-PCR and Sanger sequencing have been possible since the 90s (Küppers et al., 1993) however this does not provide the throughput required to interrogate the paired repertoire at any depth due to the cost and time intensiveness of Sanger sequencing.

scRNA-seq is a large field: there have been multiple platforms developed with focus on BCRs, all predicated upon the ability to separate individual B cells prior to lysis and PCR, into either wells or single-cell emulsions. Well-based methods are more limited in their throughput (Curtis & Lee, 2020), and the reduced reaction volumes of single-cell emulsions results in improved yields of cDNA, reduced technical variability and savings in the cost of reagents (Zilionis et al., 2017).

One class of methods from the Georgiou lab physically links the heavy and

light chain mRNA from the same cell into the same cDNA via extension overlap RT-PCR (DeKosky et al., 2016; McDaniel et al., 2016). Alternatively, cell-unique barcodes can be added during PCR, which can then be used to link transcripts from the same cell after sequencing: an earlier form of this method involved sorting single cells into individual wells and using a "primer barcode matrix", whereby each well is defined by a combination of row- and column-unique barcodes added to the forward and reverse primers (Busse et al., 2014). The most successful cell barcoding methods are emulsion or "droplet"-based: cells are encapsulated in nanolitre-scale droplets with barcoded beads. inDrop, Drop-seq and 10xGenomics are three widely used platforms which differ in bead properties, barcode design and cDNA amplification (Macosko et al., 2015; Zhang et al., 2019; Zheng et al., 2017; Zilionis et al., 2017).

10xGenomics Chromium is the most widely-used platform for single-cell sequencing of natively paired VH and VL transcripts (Goldstein et al., 2019). In 10xGenomics V(D)J, single cells are sorted into emulsion droplets containing a bead that is barcoded with a unique DNA oligo referred to as a "cell barcode". Each cell barcode is also ligated to one of several UMIs per cell. Reverse transcription and PCR of the cell- and UMI-barcoded mRNA proceeds within this emulsion droplet, followed by enzymatic shearing of the cDNAs and ligation of the adaptors and indexes as described in Section 1.2.2 (Goldstein et al., 2019). The resultant library is sequenced via Illumina MiSeq and the cell- and UMI-barcodes are used for consensus building by the proprietary CellRanger software to produce error-corrected, natively-paired VH:VL reads. An important derivation of this is LIBRA-seq which enables direct sequencing of antigen-specificity. A pool of antigens of interest are fluorophore-conjugated and uniquely barcoded; the antigens are incubated with the B cells, which are then FAC-sorted to first derive a putatively antigen-specific population of cells. When the cells are co-encapsulated with the cognate antigen, both the antigen barcode and VH:VL transcripts are tagged with the cell barcode during reverse transcription. The resultant sequencing data links VH and VL sequence with their antigen-specificity (Setliff et al., 2019; Walker et al., 2022). These data can be integrated with other cell-barcoded data such as gene expression that can be

used to ascertain transcriptional state (Hurtado et al., 2022).

1.2.4 Sequencing

Following sequencing library preparation, the resultant cDNA fragments can be sequenced via a number of methods. The first BCR-seq experiments used Roche 454 (since withdrawn), which is a form of pyrosequencing (Boyd et al., 2009): this platform was advantageous due to its longer read lengths ($>$ 400 bp) capable of covering the full V(D)J amplicon. This sequencing platform of choice in this field is now by Illumina-MiSeq which produces large numbers of reads at low cost with a benchtop sequencer.

Sanger sequencing is still applied where accuracy exceeds throughput as an objective, given its minimal error rate and long read length ($>$ 500bp). It continues to be used for cosmid- and artificial chromosome- based sequencing of Ig loci (Watson et al., 2015; Watson et al., 2013). Capillary-based Sanger sequencing works by cycles of stochastic termination of primer extension with fluorescently-labelled ddNTPs, resulting in mixtures of extension products where the identity of the terminal position is revealed by the fluorescent label. Electrophoretic separation of these extension products in capillary-based polymer gel combined with laser excitation and detection of the emission spectra of exiting fragments is the basis of sequencing (Shendure & Ji, 2008). Throughput can be increased by parallelisation in 96- or 384 independent capillaries.

The next-generation sequencing technologies that have been used in BCR-sequencing operate via "sequencing-by-synthesis". They are conceptually similar though have key differences in error profiles, read length and cost, with Illumina Mi-Seq currently favourable in terms of both error rate and cost (Friedensohn et al., 2017). Roche 454 and Illumina both use imaging: Roche 454 used chemiluminescence resulting from the enzymatic processing of pyrophosphate released upon base incorporation (referred to as pyrosequencing), while Illumina measures fluorescence from fluorophore-labelled dNTPs. IonTorrent is

another short-read technology that uses change in pH upon nucleotide incorporation as a nucleotide readout (Merriman et al., 2012): this has a significantly higher error rate however it has the fastest turnaround (Friedensohn et al., 2017). Both IonTorrent and Roche 454 suffer more from insertions and deletions particularly following homopolymeric stretches, while base miscalling is Illumina's primary form of error.

Alternatively, long-read sequencing technologies such as PacBio's single-molecule real-time sequencing (SMRT) platform (promising read lengths of greater than 1kb) have been used to sequence the long cDNAs which result from ligation of heavy and light chain amplicons (as per Section 1.2.3) and rearranged Ig loci for the purpose of allele inference (Gibson et al., 2022; Jackson et al., 2022; Watson et al., 2019). Oxford Nanopore (Deamer et al., 2016) promises ultra-long reads of greater than 100 kb which will allow more widespread analyses of the Ig loci themselves, either recombined or in somatic cells.

An important part of sequencing for downstream analysis, in addition to base calling itself, are quality scores. All major sequencing platforms use the *Phred* score, defined as -10 times the base 10 logarithm of the probability that a base call is incorrect (e.g. a Phred score of 20 corresponds to an 0.01% error rate) (Ewing & Green, 1998). This score is determined using features of the readout such as the spacing between fluorescence peaks.

1.2.5 Repertoire analysis

1.2.5.1 Processing of raw reads

There are a number of BCR-specific tools for the downstream analysis of sequencer outputs, such as Immcantation, ImmuneDB and MiXCR (Bolotin et al., 2015; Rosenfeld et al., 2018; Vander Heiden et al., 2014). pRESTO, designed for use within Immcantation but also part of the ImmuneDB pipeline, processes raw reads from the FASTQ and FASTA file formats (Vander Heiden et al., 2014). Illumina MiSeq can be applied as a paired-end sequencing platform, i.e. se-

quencing from both 5' and 3' directions, and both MiXCR and pRESTO have support for merging of these reads in order to increase confidence in the overlapping region which is usually the CDR3 (Friedensohn et al., 2017).

In addition to merging of paired-end reads, pRESTO enables read quality control by filtering by average Phred score, masking of primers for isotype assignment, UMI annotation and consensus building of UMI read groups and various deduplication and filtering steps to retain high-confidence UMI read groups (Vander Heiden et al., 2014).

1.2.5.2 Germline databases

A critical step in all repertoire analyses is germline alignment. As described in Section 1.1.1, the heavy chain is encoded by an IGHV, IGHD and IGHJ gene, while the light chain is encoded by an IGKLV and an IGKLJ gene. There is extensive allelic diversity in these genes, and new allelic variants are inferred from repertoire sequencing datasets and, more rarely, identified in long-read genomic sequencing data (Gibson et al., 2022; Watson et al., 2019) One major direction of computation in this field is assignment of the genes given the nucleotide or amino acid sequence of a heavy or light chain, and a database of possible genes. Both the construction of the gene database and the gene alignment algorithm used are critical for accurate assignment of germlines and inference of mutation.

The first step in the problem is to define the possible set of genes encoding the sequences, i.e. selection of a germline database. Ideally, the Ig loci of the subjects from which the antibody sequences were derived would be sequenced to create personalised germline databases: this would allow the most accurate assignment of germline gene usage and calculation of SHM. In practice, the difficulty of assembling the repetitive and diverse Ig locus means that genomic studies have only been performed on tens of subjects (Ohlin et al., 2019). This means that either a standard germline gene database (IMGT or Genbank) are used or a personalised germline gene database is inferred. The Partis algorithm

is capable of performing this joint inference/alignment (Ralph & Iv, 2019).

The IMmunoGeneTics system (IMGT) was devised in 1989, curating Ig and TCR genes which had been sequenced via Sanger sequencing from human genomic cosmid and artificial chromosome libraries (Kawasaki et al., 1997; Kawasaki et al., 2001; Matsuda et al., 1998; Watson et al., 2013). Sanger sequencing of genomic constructs is the gold standard of information, but is not without error (Wang et al., 2008). There has been significant work in inferring alleles based on short-read BCR-seq data (Corcoran et al., 2016; Gadala-Maria et al., 2019; Gadala-Maria et al., 2015; Ralph & Iv, 2019) and long-read sequencing data (Gibson et al., 2022; Watson et al., 2017). The Inferred Allele Review Committee (IARC) of the AIRR community now has a review process to consider the balance of evidence for a particular allelic variants and alleles that pass review are added to the IMGT database (Ohlin et al., 2019). OGRDB compiles these inferred alleles affirmed by IARC (Lees et al., 2020), while VDJBase contains the complete set of inferred genotypes and haplotypes (Omer et al., 2020).

IMGT is the most widely-used database, working directly with IARC (Lefranc et al., 2003). Genbank is also used, for example, this is the standard germline database for MiXCR, and is preferable for some commercial applications where the IMGT database cannot be used.

1.2.5.3 Germline alignment

The three most widely used germline assignment tools are IMGT/High-V-QUEST, IgBLAST and MiXCR (Alamyar et al., 2012; Bolotin et al., 2015; Ye et al., 2013). A benchmark in 2020 revealed significant differences between each method on the same datasets, even at the level of the CDR3 sequences extracted (Smakaj et al., 2020). IgBLAST was recommended as the most accurate due to lowest frequency of germline mis-hits in simulated data; IMGT/High-V-QUEST is recommended for non-command line users as it is accessible as a webserver, however this imposes a throughput limit of 500,000 sequences per job. MiXCR was sig-

nificantly faster but does not assign alleles, and uses a less complete Genbank-derived database (Smakaj et al., 2020). Another tool in use is abstar, part of the ab[x] suite of packages for antibody analysis, which performed favorably in benchmarks against IgBLAST, IMGT/High-V-QUEST, MiXCR and Partis. In this benchmark, the BLAST-based IgBLAST, Abstar and IMGT/High-V-QUEST were more consistent with one another than with MiXCR and Partis (Briney & Burton, 2018).

The algorithm used by IgBLAST is just a chaining of multiple BLAST searches which are optimised for the different lengths of the gene segments (290 bp for the V genes vs. as few as 10 bp for the D genes) (Ye et al., 2013). The BLAST (Basic Local Alignment Search Tool) algorithm is a widely-used algorithm in bioinformatics for the identification of homology between an input sequence and any sequence in a search database (Altschul et al., 1990; McGinnis & Madden, 2004). The algorithm searches for shared k-mers between the input sequence and search database sequences, of user-specified length ("word size"). Each k-mer match is scored (using a simple match/mismatch scheme for DNA and a scoring matrix such as BLOSUM62 (Henikoff & Henikoff, 1992) for amino acids). K-mers with a score above a user-specified threshold are used to seed alignments in the query database: the alignment is extended in each k-mer-containing sequence in the query database, with a new score calculated after each extension, until the score falls below a threshold. Expectation values (the probability of observing such a match of the given length in a random database of the same size as the query database) are calculated for each sequence pair and reported.

Unlike other tools, IgBLAST can use both nucleotide (igblastn) and amino acid (igblastp) sequences, however only performs junctional alignments given the nucleotide sequences (Ye et al., 2013). A number of wrappers for IgBLAST exist, including the Immcantation pipeline's *AssignGenes* command-line tool (Gupta et al., 2015) and PyIR, a Python wrapper (Soto et al., 2020). Briney and colleagues modified the IgBLAST protocol by masking the J gene region and performing a Smith-Waterman algorithm over all D gene segments (as op-

posed to doing a concomittant D/J gene search in standard IgBLAST), giving better resolution over the junctional region (Briney & Burton, 2018).

1.2.5.4 Clonotypes

Clonotypes are intended to be groups of clonally-related antibody sequences derived from common progenitor B cells (Hershberg & Luning Prak, 2015). They are a key component of BCR repertoire analysis and as such, are the starting point for nearly all analyses (Greiff et al., 2015). Definitions vary, primarily through treatment of the CDRH3. Until recently, most clonotyping methods used heavy chain information only as bulk heavy chain sequencing does not provide the native light chain (Section 1.2.3). The heavy chain is usually considered sufficient to capture most clonal relationships (Zhou & Kleinstein, 2019).

A requirement for two sequences to be predicted to share the same clonotype is common IGHV- and IGHJ-germline gene assignment (Greiff et al., 2015). The IGHD gene is not usually included in standard clonotype definitions because its assignment is both difficult and redundant in the clonotype definition, as it is wholly contained within the CDRH3 (Chen et al., 2010; Smakaj et al., 2020). The nucleotide or amino acid sequence identity across the CDRH3 required for two sequences to be considered in the same clonotype varies across studies - in studies performing clonotyping with length-normalized amino acid sequence identity thresholds, sequence identity thresholds vary between 80% and 100% (Greiff et al., 2015).

A number of publicly available, well-supported pipelines have made clonotype analysis standard practice (Greiff et al., 2015; Hershberg & Luning Prak, 2015). This has permitted large advances in the practical utility of BCR-seq data (Robins, 2013). Clinically, it has found use in tracking minimal residual disease in blood cancers, (Boyd et al., 2009) monitoring vaccination responses (Galson et al., 2014; Jackson et al., 2014; Jiang et al., 2013) and providing mechanistic insights into immune-mediated diseases(Bashford-Rogers et al., 2019;

Bashford-Rogers et al., 2018; Roskin et al., 2015; Stern et al., 2014). Clonotyping has also proven useful in antibody discovery as a means of selecting candidate sequences for expression as monoclonal antibodies (Reddy et al., 2010; Tsioris et al., 2015; Zhu, Ofek, et al., 2013) and as a method of lead antibody optimization via repertoire mining (Hsiao et al., 2019).

Most clonotyping tools have been benchmarked on heavy chain sequences alone. Clonal inference for paired VH:VL sequences from single-cell sequencing has largely used heavy chains only (Horns et al., 2020; Roy et al., 2017); clonal inference within the BraCeR tool defines heavy and light chain clones separately (Lindeman et al., 2018). In some applications clonotypes are defined on the basis of the heavy chain alone (Jaffe et al., 2022) while others use the light chain too, e.g. Cao and colleagues defined clonotypes as sharing the same CDRH3 and CDRL3 (Cao et al., 2020) which is also the basis of clonotyping in the R package scRepertoire (Borcherding et al., 2020). Reddy and colleagues also define clonotypes in paired data in this way in recent work (Neumeier et al., 2022).

As well as permitting inferences about clonal relationships between B cells, clonotyping is often used to make functional inferences: when the same clonotype is observed across multiple subjects (referred to as convergence or "public clonotypes"), it is hypothesised that these antibodies are binding to the same epitope (Greiff et al., 2015; Hershberg & Luning Prak, 2015). In practice, "binding to the same epitope" can be defined at multiple levels of resolution, in much the same way that the epitope itself is defined at multiple levels of resolution (as per Section 1.3.1: at the highest level of resolution, it can be defined in terms of an identity calculation over the contacting residues (or indeed contacting residues with some threshold energetic contribution to the interaction). It is often defined as competition for binding between two mAbs, although the relationship between epitope competition and overlap in contacting residues is complicated by possible steric or allosteric effects (Westwood & Hay, 2001). There is also the question of whether a given epitope, and binding to the same epitope, remains across closely related antigenic variants: a competition assay

may indicate close epitope sharing between two mAbs on one viral strain which would not be found when comparing another strain of the virus, so there is a degree of antigenic relativity.

Convergent clonotypes have been identified for many viral pathogens such as SARS-CoV-2 (Galson et al., 2020; Nielsen et al., 2020), Ebolavirus (Chen et al., 2022; Ehrhardt et al., 2019), influenza (Forgacs et al., 2021), HIV (Setliff et al., 2018) and dengue (Zhao et al., 2019), suggesting that this convergence may be the rule rather than the exception. The majority of antibodies binding to the same epitope in antibody-antigen complex structures in the Structural Antibody Database (SAbDab, a database of experimentally solved antibody and antibody-antigen complex structures) (Dunbar et al., 2014; Schneider et al., 2022) have highly similar CDRH3s (Wong et al., 2021).

However, it has also been observed that multiple clonotypes may converge on the same epitope. For example, Scheid and colleagues identified clonotypes from distinct immunoglobulin heavy chain variable (IGHV) gene subgroups converging on the CD4 binding site in gp120 (Scheid et al., 2011). Separate clonotypes have also been observed to bind to overlapping epitopes on the hemagglutinin stem (Joyce et al., 2016) or globular head, and on multiple epitopes on the Ebola virus glycoprotein (Rijal et al., 2019). Wong and colleagues identified 190 pairs of antibodies with sub-80% CDRH3 amino acid identity binding to the same epitope within SAbDab (Dunbar et al., 2014; Schneider et al., 2022; Wong et al., 2021). This convergence between clonotypes offers the potential to improve our understanding of the functional landscape of BCR repertoires.

1.2.6 Structural annotation of BCRs

In addition to understanding and describing antibodies in relation to their genetic origin and clonal relatedness, structural immunoinformatics as a field describes antibodies as proteins, allowing more directly functional descriptions

and comparisons. Antibody numbering, though often germline-based in terms of algorithm, is intended to establish positions of structural equivalence across antibodies. This is most often the first step in structural annotation, a class of methods including canonical class assignment and homology modelling template selection. Finally, full structural modelling of antibodies including side chain orientation provides the greatest resolution for structural analyses.

1.2.6.1 Antibody numbering

Antibody amino acid sequences can be "numbered", which is effectively an alignment that establishes equivalence across all numbered sequences of the same chain type. A number of different schemes have been proposed and are used by different subdisciplines. The first numbering scheme, the Kabat scheme was based on aligning regions of high sequence variation. No antibody structures were available at this time, however, meaning that the Kabat scheme did not ensure structural equivalence of residues. The Chothia scheme came as a derivation of the Kabat scheme once structural information became available, correcting for insertions in the CDR1s, and this scheme has been enhanced further to make structural corrections for indels (Abhinandan & Martin, 2007; Al-Lazikani et al., 1997; Chothia & Lesk, 1987; Chothia et al., 1989). These schemes are referred to as "Kabat-like" schemes and are primarily in use in structural studies.

LeFranc devised a numbering scheme that could be applied to the variable domains of antibodies and TCRs alike, known as IMGT numbering. The IMGT scheme, which did not cover the CDRH3, was extended in Honegger and Pluckthun's completely structural derivation, AHo, which introduced the idea of symmetrical insertions (Honegger & Plückthun, 2001). The IMGT scheme was later extended to include the CDRH3 and symmetrical insertions (Lefranc et al., 2003).

There are a number of webservers to implement these numbering schemes.

Abnum implements Chothia and the extended Chothia/Martin numbering (Abhinandan & Martin, 2007). DomainGapAlign is the IMGT's webserver for IMGT numbering (Ehrenmann et al., 2010; Ehrenmann & Lefranc, 2011). PyIgClassify implements AHo numbering (Adolf-Bryfogle et al., 2015). DIGIT implements Chothia numbering (Chailyan et al., 2012). ANARCI is the only freely-available downloadable tool capable of numbering antibodies with all of these schemes, and works on the basis of Hidden Markov Models (HMMs) built on the germline sequences (Dunbar & Deane, 2016).

1.2.6.2 Complementarity-determining regions

The complementarity-determining regions or CDRs are hypervariable regions of the antibody that mediate the majority of interactions with the antigen (Kunik, Ashkenazi, et al., 2012). Hypervariable regions were delineated using the first antibody sequence alignments (Wu & Kabat, 1970).

Kabat's variability-based CDR definition was extended by Chothia and colleagues in their 1987 study of four Fabs and two Bence-Jones proteins (single light chains) (Chothia & Lesk, 1987), and was subsequently updated as structural databases grew (Al-Lazikani et al., 1997; Chothia et al., 1989). Martin and Thornton showed that CDRs could be delineated based on structural clustering (Martin & Thornton, 1996), and in 2008 Ofran and colleagues provided a method for CDR delineation via structural alignment (Ofran et al., 2008). The Chothia CDR definitions in IMGT numbering are shown in Tables 1.2 and 1.3.

North and colleagues provided a novel CDR definition based on selecting anchor residues which showed tightly-clustered conformations relative to the framework, such that the anchoring residues were opposite one another in structure (North et al., 2011). This is referred to as the North scheme. The main difference between the North and Chothia definitions is in the CDRH1, which extends an additional 3 residues either side of the Chothia definition.

The IMGT CDR definition is another scheme based on sequence- and struc-

Scheme	FWR1	CDR1	FWR2	CDR2	FWR3	CDR3	FWR4
Chothia	1 - 26	27 - 37	38 - 56	57 - 64	65 - 106	107 - 117	118 - 128
North	1 - 23	24 - 40	41 - 54	55 - 66	67 - 104	105 - 117	118 - 128
IMGT	1 - 26	27 - 38	39 - 55	56 - 65	66 - 104	105 - 117	118 - 128

Table 1.2: | **VH CDR definitions.** CDR definitions in IMGT numbering.

Scheme	FWR1	CDR1	FWR2	CDR2	FWR3	CDR3	FWR4
Chothia	1 - 23	24 - 40	41 - 54	55 - 69	70 - 104	105 - 117	118 - 128
North	1 - 23	24 - 40	41 - 54	55 - 69	70 - 104	105 - 117	118 - 128
IMGT	1 - 26	27 - 38	49 - 55	56 - 65	66 - 104	105 - 117	118 - 128

Table 1.3: | **VL CDR definitions.** CDR definitions in IMGT numbering.

tural alignments referred to as the IMGT scheme (Lefranc et al., 2003). Unlike North and Chothia definitions, the positions are the same for the VH and VL. These CDR definitions are shown for the heavy chain in Table 1.2 and the light chain in Table 1.3. It can be seen that the largest difference is in the definition of CDRH1, which is between IMGT positions 24 and 40 for the North scheme and between 27 and 37 or 38 for the Chothia and IMGT schemes respectively. CDRs can be assigned on the basis of these ranges in the numbering schemes, or can be assigned by tools such as IgBLAST by mapping to the top germline alignment (Ye et al., 2013).

1.2.6.3 Canonical forms

When antibody structures became available, it was noted that the light chain CDRs, and heavy chain CDRs 1 and 2 each adopt a limited set of backbone conformations (Chothia & Lesk, 1987). These are referred to as the canonical CDRs, and the conformations they adopt are referred to as canonical classes or clusters (Long et al., 2019; North et al., 2011; Nowak et al., 2016; Wong et al., 2019). The canonical CDRs are strongly predicted by the IGHV or IGKLV gene encoding the antibody sequence (Kovaltsuk et al., 2019; North et al., 2011). As structural databases have grown, so has the number of observed canonical forms. There were 10 canonical forms in 1997, 26 by 2016 and 27 in 2018 (Wong et al., 2021).

The first automated classification algorithm was PyIgClassify which clus-

ters CDRs in torsional space using an affinity propagation algorithm (North et al., 2011). Nowak and colleagues extended canonical forms to include CDRs differing in length by up to one amino acid but still with highly similar backbone conformations using a dynamic-time warping (DTW) algorithm to cluster backbone coordinates (Nowak et al., 2016). The most rapid and accurate algorithm is SCALOP which uses position-specific scoring matrices (PSSMs) to assign sequences to an auto-updating, preclustered database, ensuring maximum coverage (Wong et al., 2019).

The CDRH3 is the most sequence- and structurally-diverse CDR (Regep et al., 2017; Weitzner et al., 2015). It has been noted that CDRH3s can be partially classified according to the conformation of their C-terminal end, which can either be kinked or extended (Shirai et al., 1996). While there were successful early attempts to predict the presence or absence of the C-terminal kink (Kuroda et al., 2008; Morea et al., 1997; Morea et al., 1998; Oliva et al., 1998; Shirai et al., 1996), more structural data made these prediction rules obselete (North et al., 2011).

1.2.7 Antibody structure prediction

Antibody structure prediction was, until the advent of the newer deep learning models, a task that was split into multiple subtasks which had different error profiles. The most difficult step in antibody modelling continues to be CDRH3 structure prediction: this is the most variable part of the antibody structure, and even the most up-to-date deep learning methods still achieve on average 2.5Å RMSD over this loop versus sub-angstrom RMSD for the canonical loops (Abanades et al., 2022; Ruffolo & Gray, 2022; Ruffolo et al., 2020).

1.2.7.1 Homology-based and *ab initio* modelling

In this previous generation of antibody structure prediction pipelines, the task is split into the structurally conserved framework, canonical CDRs and then the

challenging CDRH3. Homology-based methods (using similarity to templates found in databases of crystallographically-solved structures, described in Section 1.4.3) are preferable due to their speed in comparison to physics-based methods which are highly computationally expensive. There are a number of such pipelines which vary in their accuracy, speed and availability as web-servers or standalone applications (Berrondo et al., 2014; Bujotzek et al., 2015; Fasnacht et al., 2014; Kemmish et al., 2017; Klausen et al., 2015; Kuroda et al., 2012; Leem et al., 2016; Maier & Labute, 2014; Marcatili et al., 2008; Weitzner et al., 2014; Yamashita et al., 2014; Zhu et al., 2014). Algorithmically, there is greatest variance in the approach to CDRH3 structure prediction which may be completely *ab initio*, such as RosettaAntibody (Weitzner et al., 2017), or completely homology-based, which is the default for ABodyBuilder (Leem et al., 2016). At the time of this thesis, ABodyBuilder was the fastest (30 seconds per model) and most accurate freely-available antibody structure prediction pipeline, and as such was used for antibody structure prediction in this thesis for Chapter 2 (Leem et al., 2016).

ABodyBuilder first selects an antibody framework template for the VH and VL separately based purely on sequence identity. The orientation of these domains is selected based on sequence identity over the full VH:VL, and the domains are aligned to this template. Then, homology templates are selected for all CDRs, consecutively to prevent creating clashes in the order L2-H2-L1-H1-L3-H3, and grafted onto the framework. Template selection is performed with the algorithm FREAD which uses a combination of anchor residue (residues either side of the C- and N-termini of the CDR) RMSD and an environment-specific substitution score (ESS) to rank potential homology templates. ESS considers the frequency of residue substitutions given a particular dihedral angle neighbourhood, CDR loop and CDR length and position within the CDR (Choi & Deane, 2010). If a given ESS score (which is calculated as a function of CDR type and length) cannot be obtained given loops of the same length, the algorithm is then used to query different lengths of the same CDR type and finally, all CDR loops in the structural database. Where the ESS score still does not meet the ESS threshold, the loop is passed to MODELLER (Eswar et al.,

2008) or, in the case to CDRH3s, to the hybrid homology/*ab initio* algorithm Sphinx (Marks et al., 2017). Recourse to even hybrid *ab initio* algorithms substantially increases runtime from 30 seconds per model to 14 hours: RosettaAntibody which uses the kinetic modelling approach (KIC) described by Stein and colleagues (Stein & Kortemme, 2013) can take up to 500 CPU hours (Marks et al., 2017) to model a single CDRH3 and 1,000 CPU hours for a full Fv model (Marks et al., 2017; Weitzner et al., 2017). The final step in ABodyBuilder, and in other pipelines, is to remodel side chains for residues that have been substituted from the homology templates (in some pipelines all side-chains are remodelled to avoid clashes). ABodyBuilder samples side chain conformations for the given residue and IMGT position from an antibody-specific rotamer library, an algorithm called PEARS (Position-dEpendent Antibody Rotamer Swapper) (Leem et al., 2018). Other methods learn from general protein rotamer libraries (Miao et al., 2011; Nagata et al., 2012; Shapovalov & Dunbrack, 2011; Towse et al., 2016).

1.2.7.2 Next-generation deep learning methods

There was a significant shift in algorithmic approaches to protein structure prediction during this thesis, which has led to the widespread adoption of deep learning algorithms for protein structure prediction. This started at CASP13 where AlphaFold (Senior et al., 2020) and RaptorX (Xu, 2019) showed that deep algorithms could learn inter-residue distances and in doing so, improve protein structure prediction. The value of deep learning for protein structure prediction was cemented by the performance of AlphaFold-2 in CASP14 (Jumper et al., 2021).

Yang and colleagues extended the original deep learning algorithms by predicting not only inter-residue distances but inter-residue geometries, resulting in a geometric potential that could be minimised as part of the Rosetta pipeline. This method is referred to as trRosetta (Yang et al., 2020). The first deep learning CDRH3 structure predictor applied a similar approach to CDRH3s resulting in the model DeepH3, which could accurately predict the backbone structure

of CDRH3s (Ruffolo et al., 2020). This was later extended to prediction of the entire Fv with DeepAb (Ruffolo et al., 2022).

Some subsequent deep learning methods have moved away from the prediction of inter-residue distances and geometries to direct prediction of coordinates. ABlooper is an ensemble of five equivariant graph neural networks (eGNNs) trained to predict backbone atom coordinates. The predictions are averaged at the final step, and the RMSD between each of the eGNNs in the ensemble acts as a diversity score which can be used to filter low-confidence predictions (Abanades et al., 2022). ABlooper performed comparably with DeepAb and AlphaFold-Multimer (Evans et al., 2022) and outperformed AlphaFold-2 (Jumper et al., 2021) on the CDRH3.

ABlooper reported huge improvements in speed (AF-2 takes 3 hours on average, while DeepAb takes 10 minutes and ABlooper just 10 seconds or 30 seconds with a backbone relaxation step and side-chain grafting) (Abanades et al., 2022). IgFold followed in the direct prediction of backbone coordinates using a pretrained language model and graph neural networks (Ruffolo & Gray, 2022), with comparable speed to ABlooper. Its prediction of not only the CDRs but the full Fv means that it does not require an up-to-date structural database for framework template selection, which may also contribute to error (Ruffolo & Gray, 2022). ImmuneBuilder is a new whole Fv predictor which outperforms IgFold, AlphaFold-2 and AlphaFold-Multimer (Abanades-Kenyon et al, in prep), and also directly performs side-chain prediction. IgFold and ABlooper are sufficiently rapid as to enable accurate structure prediction at the scale of entire repertoires: the main limitation to their application in these cases is the absence of the light chain, which is required by both IgFold and ImmuneBuilder.

1.2.8 Structural annotation

Structure prediction is the prediction of atomic coordinates of both backbone and side chains. We distinguish this from structural annotation, which labels sequences with structural information permitting comparison without the computational expense of full prediction (Kovaltsuk et al., 2017; Kovaltsuk et al., 2020; Krawczyk et al., 2018). The earliest form of structural annotation was canonical form assignment (Section 1.2.6.3) which was automated for the first time in 1996 (Martin & Thornton, 1996). Annotation of heavy chain sequences with homology templates assigned on the basis of ESS score was the first method explicitly labelled as antibody structural annotation, in the format of the SAAB pipeline (Structural Annotation of Antibodies), with the purpose of allowing structural characterisation of large repertoires by removing the side-chain prediction, light chain modelling and VH:VL orientation that would be required for full structural modelling (Krawczyk et al., 2018). In comparison to maximally 2,000 antibody sequences which had previously been structurally modelled (DeKosky et al., 2016), this approach allowed structural annotation of some 35 million unique heavy chain sequences (Krawczyk et al., 2018). Kovaltsuk and colleagues later streamlined this algorithm, replaced the canonical template assignment with canonical class assignment using the rapid SCALOP algorithm (Wong et al., 2019) and clustering the CDRH3 FREAD templates into "CDRH3 structural clusters" to allow a measure of similarity between closely-related FREAD templates (Kovaltsuk et al., 2020). This study linked these structural annotation profiles to biological signal for the first time: the predicted distribution of CDRH3 shapes (in terms of CDRH3 cluster usage) was shown to change according to B cell development stage, providing repertoire-scale validation of ideas about how CDRH3s change with increasing mutation. "Repertoire structural profiling" was also introduced as a concept on the basis of CDR template usage of randomly paired heavy and light chain repertoires (Raybould, Marks, Kovaltsuk, et al., 2021). An overview of the structural annotation for a nominal heavy chain sequence is shown in Figure 1.6.

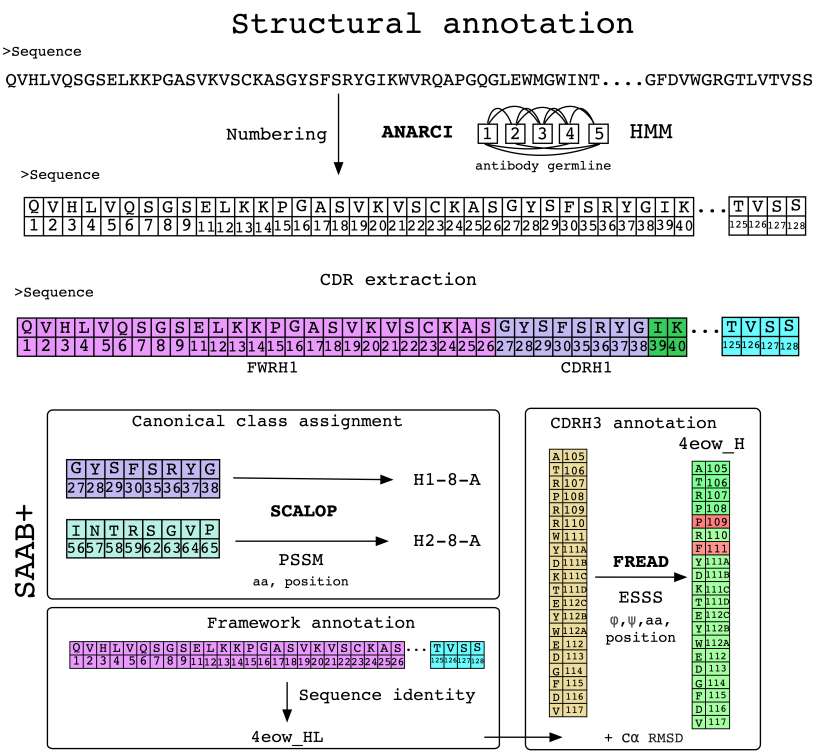

Figure 1.6: | **Schematic of current antibody structural annotation techniques.** The first step is usually antibody numbering, which establishes positions of structural equivalence across antibodies. This is required for the delineation of CDRs, which have numerous different schemes. Shown here are the IMGT CDRs. These delineations are then used to identify canonical classes for the non-H3 CDRs and structural templates for the framework region and subsequently the CDRH3. The names of the tools used are highlighted with a brief indication of the algorithm used. ϕ and ψ correspond to the dihedral angles.

1.3 The antibody-antigen interface

The part of the antibody that binds to the antigen is referred to as the paratope, while the cognate antigen surface is referred to as the epitope. There are multiple techniques for assessing what antigen an antibody binds to, the epitope it binds to and via what paratope residues.

1.3.1 Wet-lab techniques for epitope mapping

1.3.1.1 Competition assays

Competition or cross-blocking assays consist of pairwise comparisons of antibodies to identify groups of antibodies competing for the same epitope (Abdiche, 2017; Kwak & Yoon, 1996). In practice, this can be achieved by sandwich assays with different reporting mechanisms. One antibody in a pair is immobilised on a sensor chip. Antigen and the analyte antibody is added, and the binding response is measured and compared to the response observed in the absence of the analyte antibody (Kwak & Yoon, 1996). Decreases in binding response are characterised as blocking responses. Label-free (i.e. neither antibody or antigen is labelled) optical techniques such as surface plasmon resonance (SPR) and biolayer inferometry (BLI) are widely used: here, a change in refractive index for SPR and a wavelength shift in BLI are correlated to binding kinetics. Carterra-LSA is the best-in-class using a novel one-on-many fluidics approach that enables the construction of a 384x384 competition matrix within a week using minimal sample (Abdiche, 2017; Yeung et al., 2016). Alternatives to these biosensor-based sandwich assays are in use - e.g. testing for binding to cells of biotin-labelled antibodies in the presence and absence of the analyte antibody, reported by fluorescence upon adding fluorophore-conjugated streptavidin (Rijal et al., 2019) - but the speed and small quantity of material required by label-free techniques means that they are generally favourable (Syahir et al., 2015).

1.3.1.2 Mutagenesis

Mutagenesis is a standard technique for epitope mapping. One technique is alanine-scanning mutagenesis, in which a set of putative epitope residues are systematically mutated to alanine (Gershoni et al., 2007). Alanine has just a methyl side chain so mutated epitope residues may result in a reduction in binding measured by assays like the enzyme-linked immunosorbant assay (ELISA) or BLI and SPR (Zhou et al., 2019). Combinatorial mutagenesis aims to identify synergy between epitope residues by mutating multiple positions at once via

comprehensive randomisation of particular antigenic regions and then screening on phage display (Section 1.5.2) (Infante et al., 2014). Shotgun mutagenesis is a method for large-scale mutagenesis and testing, but that relies on single-site mutants (Davidson & Doranz, 2014).

1.3.1.3 Hydrogen-deuterium exchange mass spectrometry

Hydrogen-deuterium exchange mass spectrometry (HDX-MS) is a technique that allows conformational epitope mapping but not at the level of single residues. HDX is the process of replacement of hydrogen with the heavier isotope, deuterium: in proteins, this concerns the exchange of amide backbone hydrogens for deuterium in solution with a deuterated solvent. This exchange is a function of solvent accessibility, therefore would be expected to occur at lower rates in epitope residues due to steric hindrance and hydrogen bonding. The rates of hydrogen-deuterium exchange of particular positions within the antigen can be queried via mass spectrometry, and compared in the presence and absence of the antibody to identify putative epitope residues (Sun et al., 2021). However, this does not currently permit residue-level resolution so it is ordinarily combined with other techniques, including computational techniques (Pandit et al., 2012).

1.3.1.4 Display technologies

For linear or continuous epitopes (i.e. residues that are contiguous in the antigen sequence), peptide display or array technologies can be highly informative. Linear segments of the antigen are expressed and binding is assessed to identify the most strongly reactive linear peptides. Pepscan is one such method in which the peptides are immobilised on polypropylene pins and binding is assessed via ELISA (Geysen et al., 1984).

A significant limitation of linear peptide display systems such as Pepscan is that a given peptide segment is unlikely to be expressed in its native conformation, meaning that the signal measured during a Pepscan does not necessarily

correlate with binding to the equivalent residues in the full protein: false positives due to non-specific binding are a known caveat of the system (Rebollo et al., 2022; Westwood & Hay, 2001). This is an argument for using display systems which are more likely to recapitulate native conformations.

Yeast or phage surface display libraries can also be used to express protein fragments (or mutants for mutagenesis experiments as in the original combinatorial mutagenesis protocol (Infante et al., 2014)) which can contain within them conformational epitopes (Guo et al., 2015; Rijal et al., 2019; Zuo et al., 2011). For phage libraries, bound phage are eluted and used to infect *E.coli* cells which can then be sequenced (or used to generate material for subsequent rounds as in biopanning described in Section 1.5.2). For yeast display, yeast cells expressing tagged antigen fragments can be sorted via FACS and sequenced.

1.3.1.5 Antibody/antigen complex imaging

There are multiple methods for structure determination of antibody/antigen complexes. X-ray crystallography has historically provided the highest possible resolution insight into an antibody-antigen interaction as it allows the delineation of specific contacts, from which interactions can be inferred: prior to 2013, the majority of such structures deposited in the Protein Data Bank (PDB) were solved via X-ray crystallography. Multiple algorithms trained on protein structural data continue to use X-ray crystallographically-solved structures as the gold-standard. X-ray crystallography has become a routine method, however it creates a significant bottleneck as it requires crystallisation of the antibody/antigen complex.

Electron cryo-microscopy (cryo-EM) is an alternative technique that has found growing application as its resolution, previously at nanometer to sub-nanometer scale, is now competitive with X-ray crystallography (Cheng, 2015). As a result of this, cryo-EM is now the dominant method for solving antibody structures and it is predicted that by 2024 the majority of structures in the PDB will be solved by cryo-EM (Callaway, 2020). It has been applied at the level of

polyclonal sera in an assay called EM-PEM (EM polyclonal epitope mapping), which allows direct imaging of immune complexes purified from sera (Bangaru et al., 2022; Bianchi et al., 2018; Cirelli et al., 2019; Han et al., 2021; Moyer et al., 2020; Nogal et al., 2020). EM-PEM allows measurement of the dominant epitopes targeted by serum antibodies without the bottleneck of sequencing and expression of mAbs, however the exact sequences of the mAbs are not known.

Nuclear magnetic resonance (NMR) is an alternative technique but is limited to complexes below 30 kDa (Ahmad et al., 2016; Valente & Manzano-Rendeiro, 2021).

1.3.2 Computational techniques

Given the expense of expression of antibodies, a range of computational techniques have been developed to predict properties about the antibody-antigen interface.

1.3.2.1 Paratope prediction

Paratope prediction is the problem of predicting which of the residues in the antibody participate in binding to the antigen, which is usually measured on the basis of proximity of interacting residues (Chinery et al., 2022). The earliest paratope prediction method was simply the CDR concept, and some work used paratope and CDR as interchangeable terms (Ofran et al., 2008). The paratope is certainly related to the CDRs, as around 80% of the paratope residues are contained within the CDRs (Kunik, Ashkenazi, et al., 2012). Due to the fact that the binding site of antibodies is so clearly related to the CDRs, the paratope is the more trivial component of the antibody/antigen interface to predict.

However, on average only around 40% of CDR residues participate in the paratope, meaning that assigning the totality of the CDRs as the paratope is low precision. Algorithms since have largely aimed to predict the subset of the CDRs that are involved in binding, excluding the framework regions, as this reduces the class imbalance (with the exception of Paratome (Kunik, Ashkenazi,

et al., 2012) and PECAN (Pittala & Bailey-Kellogg, 2020)).

Recent methods for paratope prediction generally involve: 1. Selection of a set of high-resolution (commonly resolution less than 3Å), antibody-antigen complex structures (with protein antigens) solved by X-ray crystallography; 2. Annotation of the paratope residues, often only those paratope residues within the CDRs. This is recently defined as those residues with a minimum heavy atom to heavy atom distance less than or equal to between 4.5Å and 5Å (Ambrosetti et al., 2020; Chinery et al., 2022; Deac et al., 2019; Liberis et al., 2018; Pittala & Bailey-Kellogg, 2020), though in previous work these contacts were defined on an interaction basis (Olimpieri et al., 2013); 3. Featurising each residue in the variable region (or often CDRs) and training a model to classify each residue as paratope or non-paratope residue.

The form of the model and featurisation varies. Featurisation often includes one-hot encoding of the amino acid, some other physicochemical featurisation for example the Meiler features (Meiler et al., 2001), and spatial information encoded via graphs (Pittala & Bailey-Kellogg, 2020). Recently, equivariant graph neural networks (eGNNs) with simple feature vectors (one-hot encoded amino acids) outperformed the previous state-of-the-art (Chinery et al., 2022). Increases in algorithmic and feature complexity introduced by the most recent generation of structure-based paratope predictors have resulted in reported improvements in performance over sequence-based paratope predictors (e.g. PECAN (Pittala & Bailey-Kellogg, 2020) and Paragraph (Chinery et al., 2022)). However, as such models require structure prediction in addition to paratope prediction, their applicability is currently limited in throughput.

There are a number of paratope predictors which do not require full structural modelling. The first of these methods, beyond the CDR definition, was Paratome. Paratome identifies antigen-binding regions on the basis of structural alignments and uses a sequence identity look-up to map these predefined antigen-binding regions to a given query sequence (Kunik, Ashkenazi, et al.,

2012; Kunik, Peters, et al., 2012). This method was able to identify over 95% of paratope residues, but at significant cost to precision and specificity, and furthermore is only available as a webserver.

The next method developed was ProABC, which used random forests and a number of features including V germline subgroup, canonical forms and one-hot encoded amino acid identity at each position, to predict the paratope and the nature of the interaction in which the paratope residue was engaging (Olimpieri et al., 2013). ProABC was only available as a webserver making it unsuitable for high-throughput applications. Antibody i-Patch was the first freely-available structure-based paratope predictor based on the i-Patch algorithm, which takes into account the structural context of each residue when considering its propensity to be involved in the interface (Hamer et al., 2010; Krawczyk et al., 2013)

Parapred, published in 2018, reported significant improvements over ProABC and i-Patch. Parapred takes as input each CDR sequence (Chothia definition $\pm$ 2 CDR residues) separately. Each amino acid is featurised by its one-hot-encoded amino acid label and a set of physicochemical features consisting of a steric parameter, polarisability, volume, hydrophobicity, isoelectric point, helix probability and sheet probability (Meiler et al., 2001). The model consists of a convolutional neural network (CNN) followed by a bidirectional long-short term memory (LSTM) recurrent neural network (RNN) followed by a fully-connected layer, with dropout and L2 normalization to combat overfitting to the small dataset. The model was trained to predict a probability for each residue within the CDR regions being in contact with the antigen. This is converted to a binary label in downstream applications such as docking: the threshold chosen by the authors was 0.67 as this best recapitulated the total number of binding residues in the dataset (Liberis et al., 2018).

Since the development of Parapred in 2018, other models have been developed for deep learning-based paratope prediction on antibody sequences such as AG-Parapred-Fast, Para-EPMP and proABC-2 (Ambrosetti et al., 2020; Deac

Method	Precision	Recall	F-score	Model required
Parapred	64.5	77.2	0.70	No
Pro-ABC2	65.8	64.7	0.65	No
iPatch	42.0	89.6	0.57	Yes

Table 1.4: | **Performance of different paratope predictors in the same benchmark set.** Precision, recall and F-score at the optimal thresholds reported in the original papers for different paratope prediction methods on a set of 552 non-redundant (<95% CDRH or CDRL identity) antibody/antigen complex structures.

et al., 2019; Del Vecchio et al., 2021). Para-EPMP was not publicly available at the time of writing, while pro-ABC2, only available as a webserver, appeared to show a significant performance drop-off in our independent benchmark (Table 1.4). Parapred remains the best-performing publicly available sequence-based paratope predictor.

Some models include the antigen in addition to the antibody. ProABC used antigen volume as a feature (Olimpieri et al., 2013) while AG-Fast-Parapred and PECAN use attention mechanisms to include the antigen structure (Deac et al., 2019; Pittala & Bailey-Kellogg, 2019). Inclusion of the antigen doesn't appear to on average improve performance of paratope prediction, which is expected under the hypothesis that the majority of antibodies bind with high specificity, i.e. their binding can be abrogated by even small changes to their epitope.

1.3.2.2 Epitope prediction

Antibodies can bind to all classes of biomolecule: from haptens (small molecules conjugated to proteins), to nucleic acids, to lipids, to carbohydrates, to peptides, and proteins. While anti-glycan antibodies are of extreme importance to disease outcome (Temme et al., 2021), most crystallographically-solved antibody-antigen complexes feature protein antigens. Out of 4,435 PDB entries containing antibody-antigen complexes in SAbDab as of July 2022, 3,435 are protein antigen complexes. 604 are peptides, 221 are haptens, 128 are carbohydrates and 37 are nucleic acids. While there have been some productive efforts towards the

prediction of non-peptidic epitopes (Zierep et al., 2022), the majority of epitope prediction tools only apply to protein antigens.

Protein epitope prediction is a considerably more difficult task than paratope prediction, given that any surface exposed residue on an antigen can be immunogenic. Different studies have arrived at different conclusions as to how different epitope and non-epitope protein surfaces are. There is an observed enrichment of epitope resides in loops versus α-helix and β-strands (Rubinstein et al., 2008), there are not clear sequence differences in epitope- and non-epitope protein surfaces in most analyses (Peng et al., 2014). This is in contrast to the paratope which is notably enriched in aromatic residues, particularly tyrosine, (Fellouse et al., 2004; Kabat et al., 1977; Kringelum et al., 2013; Ramaraj et al., 2012), often in the context of neighbouring residues with short hydrophilic side chains (serine, threonine, aspartic acid and asparagine) (Benjamin et al., 1984; Mian et al., 1991; Ofran et al., 2008). Peng and colleagues found that this apparent mismatch is due to the fact that aromatic side chains in the paratope largely interact with backbone atoms and side chain carbons, which will be common across all 20 and 19 amino acids (excluding glycine) respectively. This means that the antibody's preference for aromatic side chains need not correlate with a particular side chain on the epitope side (Peng et al., 2014).

Epitope prediction as a field includes the prediction of epitopes for specific antibodies or the prediction of immunogenic regions. The majority of models have focussed on predicting either linear or conformational epitopes. Models can be further categorised according to whether they are sequence- or structure-based (Kringelum et al., 2012).

Non-antibody specific models include the sequence-based BepiPred family of models (Clifford et al., 2022; Jespersen et al., 2017; Larsen et al., 2006) as well as the most recent best-in-class for structure-based epitope prediction, ScanNet (Tubiana et al., 2022). Antibody-specific epitope prediction models include Epipred (Krawczyk et al., 2014) and the PEASE method by Sela-Culang

and colleagues (Sela-Culang et al., 2014; Sela-Culang et al., 2015).

1.3.2.3 Antibody-antigen docking

Antibody-antigen docking is the process of predicting how an antibody interacts with an antigen, and is a subset of protein-protein docking. The algorithms can be constrained such that the epitope and paratope prediction steps are separated from the docking itself (a form of constrained docking), or this can be part of the docking process (global docking). Docking algorithms include the generation of multiple guesses of the binding pose, which are referred to as decoys, and subsequent ranking of these decoys according to a scoring function (Raybould, Wong, et al., 2019). These algorithms tend to be extremely compute intensive with a clear trade-off between algorithm complexity and accuracy.

The most rapid docking algorithms are "rigid-body", meaning that neither of the binding partners can undergo conformational changes during pose generation. An example of such an algorithm is ZDock which rapidly explores hundreds of thousands of poses via Fast Fourier Transform (FFT) (Chen & Weng, 2002; Pierce et al., 2014). ZDock then ranks these poses by shape complementarity, electrostatics and statistical potentials. ZDock can be tailored to antibody-antigen docking using the Antibody i-Patch potential which reflects the observed differences in interaction propensities between antibody-antigen and general protein-protein interfaces (Krawczyk et al., 2013). While ZDock and novel scoring functions such as those used by DLAB (Schneider et al., 2021) are useful, particularly where the global search can be constrained to a known epitope region, rigid-body docking does remove an important part of the binding process which is induced fit: quite extreme conformational changes can occur in both the antibody and antigen upon binding (Fernández-Quintero et al., 2019; Stanfield et al., 1993; Wilson & Stanfield, 1994).

Other algorithms such as HADDOCK allow flexibility in both side chains and backbones (Dominguez et al., 2003). Rosetta SnugDock is an antibody-

specific algorithm which allows structural optimisation of both the antibody-antigen interface and the VH:VL interface, as well as remodelling of the CDRH2 and CDRH3 loops (Sircar & Gray, 2010), which has been successful even with homology models. However, these methods are significantly more compute-intensive requiring 250 CPU hours per complex (Weitzner et al., 2017).

Docking can be circumvented altogether by predicting the antibody and antigen as a multimeric complex using AlphaFold-Multimer (Evans et al., 2022). However, while AlphaFold-Multimer has worked exceptionally well for general protein-protein docking, it does not perform well on antibody/antigen docking with a success rate of 11% (Yin et al., 2022), likely a result of the described differences between antibody/antigen and general protein-protein interfaces (Section 1.3.2.2).

1.3.2.4 Computational epitope binning

Computational epitope binning is the process of prediction of epitope competition, as opposed to direct epitope prediction, and as a result is significantly more computationally tractable. Successful approaches for *in silico* epitope binning include Ab-Ligity (Wong et al., 2021), SPACE (Robinson et al., 2021) and EpiBin (Mahita et al., 2022).

Ab-Ligity uses the distance between pharmacophoric interaction points (PIPs) in known or predicted structures to create hash table representations of binding sites. It can be applied to group similar epitopes or to group similar paratopes, which can be predicted. In Ab-Ligity on the "paratope" side, Parapred (Liberis et al., 2018) is first used to predict the subset of the CDR residues that are to be treated as "paratope" residues. Each paratope residue is assigned a physicochemical bin and is reduced to its Cα coordinate. Distances are calculated between all pairs of PIPs and these pairs of PIPs are put into bins according to distance with a bin size of 1Å. All possible triplets (i.e. triangles of PIPs) that can be built using these hashes are enumerated, and a similarity is calculated

Residue bin	Amino acids (one-letter code)
Aliphatic	G, A, V, L, I, P
Hydroxyl	S, T
Sulfur	C, M
Aromatic	F, Y, W
Acidic	D, E
Amine	N, Q
Basic	H, K, R

Table 1.5: | **Residue tokenization used by Ab-Ligity.** Physicochemical residue bins used in Ab-Ligity.

using the Tversky index:

$$S(X,Y) = \frac{X \cap Y}{X \cup Y + \alpha|X - Y| + \beta|Y - X|} \tag{1.1}$$

This Tversky index, where sets X and Y are the sets of PIP triplets in each binding site in the pair and α and β are set to 1, is the binding site similarity score. For both epitope and paratope comparison, a similarity score of 0.1 was found by Wong and colleagues to maximise the precision and recall of grouping antibodies to the same epitope (Wong et al., 2021). Antibody-antigen complexes with an epitope similarity score of greater than or equal to 0.1 were considered to be the same epitope, while antibodies with a paratope similarity score of greater than or equal to 0.1 were predicted to bind to the same epitope (Wong et al., 2021).

In SPACE, Cα RMSD is calculated across the CDRs of pairs of antibodies which have the same length CDRs. This distance matrix is then clustered via a greedy clustering algorithm to produce clusters of antibodies predicted to bind to the same epitope (Robinson et al., 2021).

EpiBin is a method so-far only applied to repebodies (protein scaffolds based on the non-immunoglobulin antibodies found in jawless vertebrates), which used rigid-body docking competition scores to define competing repebodies (Mahita et al., 2022).

1.4 Antibody databases

1.4.1 Sequence databases

There are a number of repositories specifically for repertoire sequencing data. iReceptor contains both BCR- and TCR-seq data and is the sequencing repository most tightly linked with the AIRR community. As of September 2022, it contains over 5 billion TCR and BCR sequences (Corrie et al., 2018).

The Observed Antibody Space (OAS) is an alternative repository currently containing 2.4 billion sequences of antibodies alone. The database contains sequences processed according to a standardised protocol in a miAIRR (Minimal Information about Adaptive immune Receptor Repertoire) compliant format, in combination with ANARCI numbering (Kovaltsuk et al., 2017; Olsen, Boyles, et al., 2022).

Both databases now also include paired data, with OAS currently only containing paired data from the 10x Genomics platform (Olsen, Boyles, et al., 2022).

1.4.2 Antigen-specific antibody databases

The largest database of epitope-specific antibody identifiers is the Immune Epitope Database (IEDB). Where available, receptor sequences are collected (Rosenfeld et al., 2018). As of September 2022, there are 2,272 antibody sequence groups (defined by CDRH3/CDRL3 amino acid sequences) available.

There are a number of disease-specific databases containing varying degrees of antigen or epitope information. CATNAP is a database of neutralising HIV antibodies and contains 362 amino acid sequences (Yoon et al., 2015). CoV-AbDab is the largest antigen-specific database containing over 10,000 sequences of antibodies and nanobodies with known binding to any coronavirus protein

(Raybould, Kovaltsuk, et al., 2021).

1.4.3 Antibody structure databases

There are several databases dedicated to the curation of structures of antibodies which are pulled down or mirrored from the PDB. The two most widely used are AbDb (Allcorn & Martin, 2002; Ferdous & Martin, 2018) and SAbDab (Dunbar et al., 2014), both of which are available as webservers. The IMGT also has its own structural database IMGT/3D-Structure-DB (Ruiz & Lefranc, 2002).

1.5 Antibody discovery

1.5.1 mAbs derived from in vivo systems

Deriving monoclonal antibodies from *in vivo* systems, i.e. from B cells of animals immunised with the antigen of interest or human survivors of the disease of interest, has key advantages over display systems using synthetic libraries. Namely, these antibodies are affinity-matured in vivo and so are generally high-affinity and specific; furthermore, they on average have more favourable developability properties (Garripelli et al., 2021; Kaleli et al., 2019; Nixon et al., 2014; Sun & Benet, 2020) though it should be noted that therapeutic antibodies differ from natural repertoires in a number of regards (Raybould, Marks, Krawczyk, et al., 2019).

1.5.1.1 Humanisation of murine antibodies

Where antibodies are derived from non-human animals, extensive modification to reduce the likelihood of immunogenicity in human subjects is required. This is referred to as humanisation (Marks et al., 2021; Safdari et al., 2013) and generally is achieved by CDR grafting onto the closest human antibody framework, often with back-mutations in the Vernier zones to preserve CDR conformations (Foote & Winter, 1992; Jones et al., 1986). More systematic computational approaches based on iterative optimisations of humanness scores, such as over-

lap with reference human antibody sequence datasets (Abhinandan & Martin, 2007; Lazar et al., 2007), scores from classifiers such as Hu-mAb (Marks et al., 2021) and BioPhi/Sapiens (Prihoda et al., 2022), have been developed in recent years.

1.5.1.2 Humanised mice

The challenge of humanising non-human antibodies can be minimised by using organisms which either have transgenic immunoglobulin loci or which host transplanted human hematopoetic cells (referred to as humanised or "human immune system" organisms). This means that any antibodies isolated thereafter will be encoded by human germline genes and will be less likely to elicit anti-drug antibodies (ADAs).

The immunoglobulin loci posed a challenge for geneticists due to their size and complexity, being among the largest genomic regions that have been successfully humanised in mice (Zhu et al., 2019). The first Ig transgenic mouse platform was published in 1989 (Brüggemann et al., 1989). It was not until 25 years later that the full set of variable human Ig genes were successfully inserted into the genomes of mice, replacing the endogeneous loci, which was achieved in parallel by two separate groups (Lee et al., 2014; Macdonald et al., 2014; Murphy et al., 2014). Humanised immunoglobulin (Ig) loci-transgenic animal models have proven extremely useful in therapeutic antibody discovery; twenty of the 127 therapeutic antibodies licensed in the US or EU as of April 2022 were derived from transgenic mouse platforms (data from Thera-SAbDab (Raybould, Marks, Lewis, et al., 2020)). Transgenic platforms have also found a new application in vaccine response modelling (Pantophlet et al., 2017; Sok et al., 2016; Walls et al., 2020). The two most advanced platforms are Kymab's Kymouse (Lee et al., 2014) and Regeneron's VelocImmune mice (Macdonald et al., 2014; Murphy et al., 2014). Earlier platforms include the Xenomouse (Jakobovits et al., 2007) and the OmniAb platform (Osborn et al., 2013).

Contemporary Ig transgenic animal models vary according to the number of genes and localisation of the inserted human Ig loci (Brüggemann et al., 2015; Green, 2014). In Kymab's Intelliselect Transgenic mouse (Kymouse), a complete set of human variable (V), diversity (D) and junction (J) genes of the IGH locus as well as the V and J genes of the IGL and IGK loci were inserted at the sites of the endogenous mouse loci. The mouse constant regions were retained, preserving downstream interactions with endogenous intracellular signalling components and cell membrane Fc receptors, resulting in functional, fully active chimeric antibodies. Kymice exhibit normal B cell production and maturation and the resulting B cell receptors (BCRs) are diverse, with human-like CDRH3 lengths and evidence of somatic hypermutation (Lee et al., 2014).

HIS animals are animals with severe primary immunodeficiencies (in most strains due to the severe combined immunodeficiency (*scid*) mutation) that have been engrafted with human hematopoetic cells. More recent strains are transgenic for several cytokine genes, as the native mouse cytokines do not interact properly with human receptors, hampering B cell development (Zhu et al., 2019). An additional problem is the lack of lymph node development in these mouse strains which is an obvious limitation on affinity maturation. Nonetheless, HIS mice have been used for the discovery of antigen-specific antibodies from memory populations (Becker et al., 2010).

1.5.1.3 Hybridoma and cell culture

The next challenge in isolating therapeutic antibodies from an immunised animal or disease-exposed human subject is screening for antigen-specific antibodies. This requires a method to go from cells or sequences to sufficient quantities of antibody for functional assays.

The classic method for this is hybridoma technology. Hybridomas are fusions of ASCs and immortal myeloma cell lines, which provide a stable, long-living cell line with the antibody-production capability of the ASC. The hybridoma was devised in 1975 by Kohler and Milstein; in 1986, the first therapeu-

tic antibody (Muromonab-CD3) was licensed by the FDA, after being isolated from a hybridoma derived from ASCs from mice immunised with CD3 (Miller et al., 1981). The hybridoma is however a relatively slow process, taking on average four to six months (Zhang, 2012); furthermore, only 1 in 100,000 ASCs successfully fuse with the myeloma lines on average (Yu et al., 2008), meaning that on average only 0.001% of the B cell response to an antigen can be queried. Single-cell culture is an alternative to immortalisation that has been successfully used to isolate antigen-specific antibodies; its lower attrition rate means that rare B cell lineages such as HIV bnAbs can be sampled (Corti et al., 2013; Macagno et al., 2010; Walker et al., 2009; Yu et al., 2008). For antigen-specific memory cells (as opposed to ASCs), antigen-specific FACS or MACS followed by activation and analysis of secreted antibody, or VH:VL cloning and expression, enables isolation of antigen-specific antibodies (Pedrioli & Oxenius, 2021).

Most recently, paired sequencing of antigen-sorted B cells followed by cloning and expression of select antibody sequences, often chosen due to their expansion, degree of mutatedness and isotype subtype, has enabled antibody discovery from cells isolated from human survivors of disease. Cao and colleagues identified numerous potent SARS-CoV-2 neutralising antibodies from convalescent patients using single-cell sequencing (Cao et al., 2020). This remains a fairly high-cost approach to antibody discovery.

1.5.2 Display technologies

While antibodies derived from immunised animal or disease-exposed human subjects generally have excellent target affinity and specificity, the processes for identifying and screening them are time- and cost-intensive and there are limitations in terms of the antigens that can be used (antigens have to be tolerable and sufficiently immunogenic). Furthermore, there are ethical concerns for using animal immunisation. Many have called for a decreased reliance on animal immunisation by focussing on *in vitro* discovery methods such as display technologies. Gray and colleagues argue that the described techniques only allow us access to the "tip of the antibody iceberg", and that *in vitro* techniques should

be encouraged going forward due to their greater versatility and reproducibility (Gray et al., 2020).

Display technologies involve surface expression of proteins (in this context, a library of antibody variable domains) onto a phage virion, ribosome, or bacterial, yeast or mammalian cell. Phage display was the first such display technique developed in 1985 (Smith, 1985). In combination with methods to iteratively enrich for high-affinity antibodies, referred to as "panning" (Parmley & Smith, 1988), phage display was used in 1990 to display human antibodies derived from recombinant antibody libraries (McCafferty et al., 1990; Winter & Milstein, 1991). Phage display has been used to discover at least 29 therapeutic antibodies that have reached clinical development, including the most widely used antibody therapeutic adalimumab (Alfaleh et al., 2020). Eukaryotic display systems (yeast and mammalian cells) have the advantage of producing glycosylated antibodies and can display full-length antibodies, which will provide an earlier insight into potential developability problems (Parthiban et al., 2019).

The major disadvantage of antibodies produced via *in vitro* technologies is they tend to pose greater developability risks, in terms of immunogenicity, biophysical attributes and polyspecificity (Jain et al., 2017; Spencer et al., 2012). Polyspecificity of phage-derived antibodies appears to result from a higher proportion of aliphatic residues in the CDRs than non-phage-derived antibodies (Kaleli et al., 2019). Either experimental (Bailly et al., 2020) or computational screens for developability issues (such as the Therapeutic Antibody Profiler (TAP) (Raybould, Marks, Krawczyk, et al., 2019)) are more critical for these technologies (Laustsen et al., 2021).

1.5.3 BCR-seq and antibody discovery

The standardisation and streamlining of BCR-seq protocols means that BCR-seq and associated clonal analysis has found its place in antibody discovery both as a method of identification of antigen-specific antibodies, and as a tool to be used in antibody optimisation.

The first application of BCR-seq to antibody discovery came in 2010, when Reddy and colleagues demonstrated that the majority (78%) of the most abundant clones of bone marrow-resident plasma cells following immunisation were antigen-specific - they were able to identify high-affinity antibodies to three antigens without costly screening steps (Reddy et al., 2010). There have been multiple subsequent studies confirming that expansion in BCR-seq data can be a reliable indicator of antigen reactivity (Galson et al., 2015; Lanz et al., 2022; Saggy et al., 2012; Valdés-Alemán et al., 2014; Wang et al., 2015). Neumeier and colleagues reported a reliable link between antigen-specificity and expansion for around the top ten largest clones in bone marrow-resident plasma cells in mice (Neumeier et al., 2022), though this appears to only be the case for young mice (Agrafiotis et al., 2022). In combination with expansion, isotype subtype appears to be a good predictor of antigen-specificity from repertoire data (Agrafiotis et al., 2022) and is used, in combination with factors like mutation count, to select antibodies from repertoire sequencing data (Cao et al., 2020).

In addition to being used as sources of novel therapeutic antibodies, repertoire sequencing datasets (in particular bulk repertoires due to their superior depth) can be used in the process of lead antibody optimization by allowing the identification of clonal relatives which may have superior affinity or improved developability profiles. Hsiao and colleagues used this technique, which they refer to as immune repertoire mining, to identify multiple variants of their known binders with orders of magnitude improvements in affinity (Hsiao et al., 2019). Zhou and colleagues used a similar repertoire mining process to identify variants of a broadly neutralizing HIV antibody (Zhu, Ofek, et al., 2013). However, such techniques are limited to identifying close sequence relatives of the starting or "probe" antibodies: other techniques such as identifying relatives according to presence of important CDRH3 structural motifs, such as disulfide bonds in IGHV1-69 HCV neutralizing antibodies, have recently been published (Bozhanova et al., 2022; Finn et al., 2020), however are still restricted to identifying antibodies with the same IGHV genes.

These repertoire mining techniques are the same in principle as *in silico* epitope binning (Section 1.3.2.4), as they relate similarity in some metric (e.g. clonal relation) to the probability of binding to the same epitope. The main difference in these approaches is scale and treatment of the light chain. Computational epitope binning need only be applied to panels of hundreds of antibodies and the light chain is ordinarily known, whereas algorithms used in immune repertoire mining on bulk repertoires must be able to operate on 10^4 - 10^6 heavy chain sequences.

1.6 Publications and preprints

Some of this thesis is reproduced from following publication and preprint, respectively:

Chapter 2 — Richardson, E., Galson, J. D., Kellam, P., Kelly, D. F., Smith, S. E., Palser, A., ... Deane, C. M. (2021). A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. In MAbs (Vol. 13, No. 1, p. 1869406). Taylor Francis.

Chapter 3 — Richardson, E., Binter, S., Kosmac, M., Ghraichy, M., von Niederhäusern, V., Kovaltsuk, A., ... Watson, S. (2022). Characterisation of the immune repertoire of a humanised transgenic mouse through immunophenotyping and high-throughput sequencing. bioRxiv. (Under review)

2 | A novel method that uses high-throughput paratope prediction to identify antigen-specific antibodies from repertoire data

2.1 Abstract

Characterizing the functional relationship between sequence-distinct antibodies is an important step in our understanding of the adaptive immune landscape. Mapping antigen preference to antibody repertoires will allow us to identify epitope convergence of antibodies at a large scale. In antibody discovery, such computational methods are the basis of "immune repertoire mining" (Section 1.5.3). In BCR repertoire mining, the BCR sequences of a repertoire of interest are compared to an antibody of interest to find related sequences, which may have improved binding or a better developability profile.

The baseline method for this task is clonotyping (as per Section 1.2.5.4) and relies on the hypothesis that antibodies that derive from B cells from the same lineage often bind to the same epitope with varying affinity, as they arise during the somatic hypermutation process which occurs when B cells enter the germinal centre following engagement with the antigen. This method continues to be used with great success to identify sequences from the same genetic background.

We sought an alternative method that is not predicated upon the ability to group antibodies from the same lineage, which itself is a difficult computational problem (Section 1.2.5.2). We developed a method that works with similar precision to the baseline method but allows the grouping of more sequence-diverse antibodies that may have happened upon the same epitope. This method was used to isolate a number of novel PTx-binding antibodies from different genetic backgrounds and with predicted improvements in properties related to developability.

The following chapter contains reproduced material from the following paper:

Richardson, E., Galson, J. D., Kellam, P., Kelly, D. F., Smith, S. E., Palser, A., ... Deane, C. M. (2021). A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. In MAbs (Vol. 13, No. 1, p. 1869406). Taylor Francis.

All work presented in this chapter is mine unless otherwise stated.

2.2 Introduction

In antibody discovery, clonotyping is used to search for clonal relatives of lead antibodies in bulk BCR-seq datasets in order to identify antibodies that target the same epitope, but which have either an increased affinity or a superior developability profile. This process is referred to as "immune repertoire mining" (Section 1.5.3). Hsiao and colleagues performed clonotyping on a set of bulk heavy-chain repertoires and used the resultant clonotypes for hit expansion against two targets (Hsiao et al., 2019). They achieved greater than an order of magnitude improvement in affinity for both targets and between 48% and 100% of tested heavy chain variants retained target-binding (Hsiao et al., 2019). This suggests that sampling within a clonotype can be highly effective as a means of repertoire mining. However, the method does not allow the identification of binders to the same epitope that derive from different clonotypes, which currently limits the sequence distinctness of novel binders that can be recovered from immune repertoires.

In this chapter, we describe a new method to identify functional convergence of antibody sequences that is germline-independent and that considers only the binding site of the antibody sequences, the paratope. We call this approach "paratyping". We show how paratyping allows grouping of antigen-specific sequences from different clonotypes, and is a rapid, structurally intuitive way

of grouping functionally related antibodies. Paratyping simplifies the complex phenomenon of antibody–antigen interaction into sets of shared residues. Learning the complexities of antibody–antigen interactions as part of a predictive model of antigen binding has been achieved on an antigen-by-antigen basis, for example in the case of the antigen interaction of one therapeutic monoclonal antibody (Mason et al., 2019). However, such approaches rely on a large (on the order of 10^4) library of experimentally validated binding and non-binding variants. Paratyping removes the need for large amounts of training data and is generalizable across protein antigens.

We first show the rationale for our paratyping method using the structures of a pair of antibodies from different clonotypes that bind to the same epitope. Paratyping is then applied to a single-cell dataset of sequences raised against pertussis toxoid (PTx) in a transgenic mouse platform where it is as accurate as clonotyping but identifies different binders. We then perform a prospective experimental test of the method by expressing as monoclonal antibodies and experimentally testing predicted PTx-binding and non-binding antibodies mined from a set of non-enriched bulk heavy chain sequencing repertoires. Our experimental test demonstrates that paratyping identifies PTx-binding antibodies from different clonotypes to our original hits. This expands the sequence space available through repertoire mining and permits favorable shifts in in silico developability metrics. Of particular advantage is paratyping's ability to predict common antigen reactivity of antibodies from different IGHV and IGHJ gene backgrounds, which has implications for large-scale repertoire analysis. To put paratyping into the context of other methods which have been developed since paratyping, we perform an up-to-date benchmark of paratyping's ability to group antibodies to the same epitope using data collated in the coronavirus antibody database, CoV-AbDab, demonstrating that paratyping continues to have comparable precision to clonotyping at this task.

2.3 Materials and Methods

2.3.1 Datasets

Experimental datasets (sections entitled: "Single-cell dataset" and "Bulk dataset") were generated by the Infectious Diseases unit at Kymab Ltd, with special thanks to Dr. Sarah Smith and Dr. Simon Watson.

2.3.1.1 PTx datasets

2.3.1.2 Single-cell dataset

Five genetically engineered mice that have a full set of human immunoglobulin variable region genes (Intelliselect transgenic mice) were immunized with PTx. This study was carried out under the Project Licenses 70/8718 issued by the UK Government Home Office under Animal (Scientific Procedures) Act (A(SP)A), 1986, incorporating Directive 2010/63/EU of the European Parliament, and with the approval of the Sanger Institute Animal Welfare and Ethical Review Body. The institute complied with the Code of Practice issued by the UK Government which aids compliance with the A(SP)A. The institute has a PHS assurance F16-00128 (WTSI).

A total of 1290 paired (VH:VL) sequences were recovered from antigen-sorted, plasma and memory cells. These 1290 sequences were expressed in HEK293 cells. Antibody supernatant was collected on d8 after transfection and screened for binding to PTx by HTRF. Positive control anti-PTx antibody (ab37574, abcam) was diluted in Expi293TM Expression Medium (Gibco) over an 11-point titration using one in three dilutions to generate a standard curve. Titrations of 5 μl of antibody ab37574 were added to a 384-well white-walled assay plate (Greiner Bio-One). Negative control wells received 5 μl of Expi293TM Expression Medium only. Five μl of HEK293 antibody supernatants (undetermined concentration) were added to one well of the 384-well plate. Five μl of PTx conjugated to Alexa 647 (Lightning-Link, Innova Bioscience) (3.75 nM final concentration) was added to all wells of the assay plate except negative con-

trol wells, which instead received 5 μl Expi293TM Expression Medium. Finally, 10 μl of anti-mouse IgG donor antibody (Southern Biotech; to bind to murine constant region in control and chimeric Intelliselect transgenic mouse antibodies) labeled with europium cryptate (Cis Bio), (1:4000 final concentration) was added to each well and the assay was left in the dark at room temperature to incubate for 2 h. After incubation, the assay was read on an Envision plate reader (Perkin Elmer) using a standard HTRF protocol. Then, 620 and 665 nm channel values were exported to Microsoft Excel (Microsoft) and F calculations performed ((665/620 nm ratio – signal negative control)/signal negative control) × 100). Percent effect values were calculated for each antibody by comparing its F value against a positive control antibody (ab37574) at 6.66 nM.

Antibody sequences with greater than 10% effect value relative to the positive control were labeled as binders. This resulted in 364 PTx-binders and 926 non-binders. SPR was performed on each of the 364 binders.

2.3.1.3 Bulk dataset

Heavy chains from sorted splenic B cells from the same individual five mice as the single-cell dataset were sequenced using standard protocols (Stern et al., 2014) and processed using the pRESTO/Change-O (Gupta et al., 2015; Vander Heiden et al., 2014) pipeline. This resulted in 259,151 heavy chain sequences. For quality control, only sequences with read count (reads with a particular unique molecular identifier (UMI)) above or equal to two or a consensus count (reads with different UMIs but the same nucleotide sequence) above or equal to 10 were considered, reducing the size of the dataset to 59,107 sequences.

2.3.1.4 SAbDab dataset

We identified sets of antibodies from SAbDab (Dunbar et al., 2014; Schneider et al., 2022) that bound to different epitopes on the same antigen. We curated five separate antigen-specific databases with non-redundant (100% level) heavy chain CDRs (IMGT definition): influenza hemagglutinin (n = 79), HIV

gp120 (n = 77) and gp160 (n = 39), the SARS-CoV-2 spike protein (n = 44) and lysozyme (n = 43). For each antigen, the epitopes were converted to a hashtable representation via Ab-Ligity (Wong et al., 2021) and compared. Antibodies to the same antigen with epitope similarity scores equal to or greater than 0.1 were considered to be positive pairs as in the original Ab-Ligity application.

2.3.1.5 CoV-AbDab benchmark dataset

CoV-AbDab is the most complete dataset of coronavirus-binding antibodies with epitope information as described in Section 1.4.2.
For an up-to-date benchmark, we analysed CoV-AbDab (date stamped to 14th June 2022). To compare performance of the methods with different levels of epitope resolution (domain vs. residue-level), we only considered the 308 antibodies for which there is an antibody:antigen co-crystal structure. The majority (263) are to the receptor-binding domain (RBD). For each antibody/antigen complex, the epitopes were converted to hashtable representation via Ab-Ligity and compared as in Section 2.3.1.4 to define positive pairs (antibodies binding to the same epitope).

2.3.2 Methods

2.3.2.1 Alignment and numbering

Where the nucleotide sequence was available, the Change-O (Gupta et al., 2015) AssignGenes module (which is a wrapper for IgBLAST) was used to align sequences to an IMGT germline database (Lefranc et al., 2003) and extract CDRH3s. Where only amino acid sequences are known, such as in SAbDab, ANARCI was used to assign germline genes (Dunbar & Deane, 2016) and extract CDRs according to the IMGT or North definitions given in Table 2.1.

Definition	Chain	CDR1	CDR2	CDR3
IMGT	H	27 - 38	56 - 65	105 - 117
IMGT	L	27 - 38	56 - 65	105 - 117
North	H	24 - 40	55 - 66	105 - 117
North	L	24 - 40	55 - 69	105 - 117

Table 2.1: | **IMGT numbering ranges for CDRs 1 - 3 for heavy and light chains with IMGT or North definitions..**

2.3.2.2 Clonotyping

We aimed to group antibodies into clonotypes according to the most common definitions in the field as per Section 1.2.5.4 (Greiff et al., 2017; Greiff et al., 2015; Hershberg & Luning Prak, 2015; Yaari & Kleinstein, 2015). This requires matching of the IGHV and IGHJ genes and a sequence identity threshold over length-matched CDRH3s. The specific sequence identity threshold was tuned during the single-cell PTx experiment: this was chosen as the sequence identity which simultaneously maximised the precision and recall in the task of grouping PTx-binding antibodies in the experiment described in Section 2.4.2. It should be noted that this is distinct from the task of grouping clonally related antibodies.

2.3.2.3 Paratope prediction

At the time of writing, Parapred was both the state-of-the-art in antibody structure prediction and the quickest method. We used the Parapred model trained by E. Liberis available at https://github.com/eliberis/parapred. Probabilities were converted to binary labels using the 0.67 threshold deemed optimal by the original authors.

2.3.2.4 Structure prediction

At the time of the development of paratyping, ABodyBuilder was the fastest and best-performing antibody modelling tool (Leem et al., 2016), therefore any models of the PTx-binding antibodies were built using ABodyBuilder.

For the CoV-AbDab benchmark which was performed in July 2022, ABlooper

was used for CDR prediction (Abanades et al., 2022) and SAbDab was used for antibody framework template and orientation selection. We considered an antibody to be successfully modelled where the diversity of the five models was below 0.87 for CDRH1, 0.82 for CDRH2, 3.25 for CDRH3, 1.01 for CDRH1, 0.92 for CDRH2 and 1.10 for CDRL3 (recommendations by Brennan Abanades-Kenyon, personal communcations). 270 of these 308 antibodies could be modelled with high confidence using ABlooper (Abanades et al., 2022). The effect of these quality control measures on downstream performance was not assessed.

2.3.2.5 Paratope identity calculation

Paratope identity was calculated across sequences with the same length CDRs (North). The calculation is as outlined in Figure 2.1. It is a simple identity calculation where the numerator is positions that are i) predicted to be in the paratope in both sequences and ii) an identical amino acid, and the denominator is the minimum predicted paratope size in the pair of sequences.

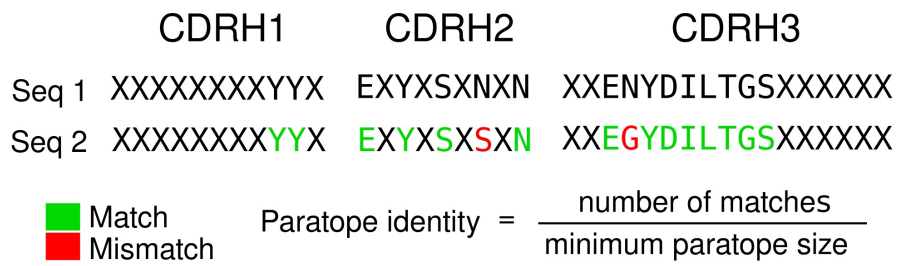

Figure 2.1: | **Method of calculating predicted paratope identity.** X indicates CDR residues not predicted to constitute the paratope. In the example shown, the paratope identity is 87.5%.

2.3.2.6 Repertoire mining experiment

In the repertoire mining experiment, we selected a number of hit antibodies from the single-cell dataset to use as probes. SPR was carried out on the HTRF-positive antibodies and the highest affinity representatives were selected from each clonotype containing only sequences labeled as binders equating to 97 antibodies, in order to align with previous repertoire mining experiments (Hsiao

et al., 2019). We also selected a number of non-PTx binding antibodies to be used as a negative control. The non-PTx binding antibodies were selected as representatives of clonotypes containing only non-PTx binding sequences (551 antibodies).

The heavy chains from these probe antibodies were used as probes to mine the bulk repertoires via both paratyping and clonotyping, using the optimal sequence identity thresholds from the single-cell dataset (75% predicted paratope identity for paratyping and 72% CDRH3 sequence identity for clonotyping). These thresholds were selected as the values that, for the relevant method, maximised precision and recall in the task of grouping PTx-binding antibodies in the single-cell dataset. The repertoire mining process using paratyping is illustrated in Figure 2.2.

Predictions were labeled as "paratype-only", "clonotype-only" or "both" as detailed in the "Results" section. If an antibody is within the same paratype as a particular probe but not within its clonotype, it is a "paratype-only" prediction and vice-versa. A prediction is labeled as "both" if it is within both the paratype and clonotype of the same probe. The more stringent "paratype-unique" and "clonotype-unique" definitions were used to select antibodies to test for the repertoire mining experiment.

2.3.2.7 Experimental validation of predicted binders

We selected 192 heavy chain sequences from the bulk dataset for expression. A total of 144 of the sequences were predicted to be PTx binders using the 97 PTx-binding probe sequences. We created a new way of categorizing predictions in order to fully test the method in the context of an antibody discovery experiment. We define "paratype-unique" predictions as predictions which are not within the paratype of not only the probe in question (paratype-only) but of any of the full complement of 97 probes, and similarly for "clonotype-unique" predictions. Paratype-unique predictions are a more stringent subset of paratype-

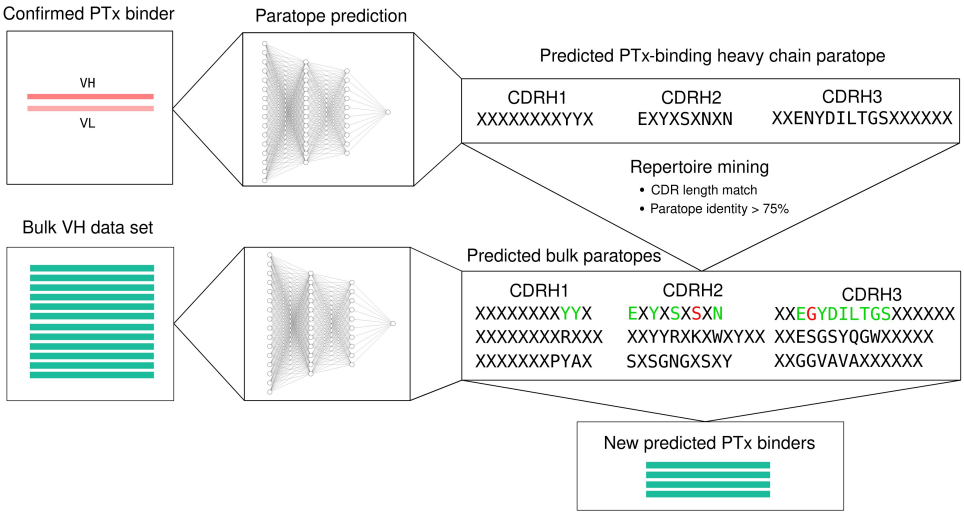

Figure 2.2: | **Graphical illustration of the process of repertoire mining via paratyping.** A probe antibody is selected. Paratope prediction is performed both on the probe antibody and the bulk dataset. Heavy chain predicted paratope identity is used to mine the bulk repertoire for new predicted binders.

only predictions and present a greater challenge for the method (as such predictions tend to be corroborated by fewer probes). Forty-eight predicted PTx-binding heavy chain sequences were selected from each of the three categories of prediction ("paratype-unique", "clonotype-unique" and "both"). Five of the "both" predictions were identical to their probes. The remaining 48 sequences assayed were predicted non-PTx binding heavy chain sequences with 16 sequences selected from "clonotype-unique", "paratype-unique" and "both" categories of prediction.

The heavy chain sequences selected for expression from the bulk repertoire were paired with the cognate light chain of the probe sequence by which they were identified. Within-clonotype VH:VL pairing has been validated as a method of reconstituting binding where the cognate light chain is not sequenced (Hsiao et al., 2019; Zhu, Wu, et al., 2013). Pairings were only made where there was greater than 82% sequence identity across the residues considered to constitute the VH:VL interface based on solvent accessibility (Raybould, Marks, Krawczyk, et al., 2019) in order to maximize the probability of

expression. These positions lie outside of the CDRs and therefore do not enforce any constraint on binding site sequence identity.

The predicted binding and non-binding antibodies were expressed in HEK293 cells. The assay is as described in the "Single-cell dataset" subsection under the "Materials and methods" section with the exception that the anti-PTx antibody 1B7 was used as positive control. Antibody sequences with an F value exceeding 100 were labeled as PTx-binding antibodies. The expression and testing of the antibodies was performed by the Infectious Disease team at Kymab Ltd.

2.3.2.8 Performance evaluation

To evaluate the performance of either method in grouping PTx-binding sequences, we calculated precision and recall of the method in the single-cell dataset according to the following definitions of true positives, false positives and false negatives.

The task is to group antibodies that bind PTx. However, the method is hypothesized to work by grouping by epitope. There will be multiple epitopes on PTx so not all PTx-binding antibodies will be grouped into a single paratype or clonotype. As a result, we do not expect perfect recall. Further, we do not classify antibodies as non-binders if they do not group with a particular binder – we only predict that they do not bind at the same epitope as the binder in question. As a result, we do not calculate a "true negative" rate. Precision and recall are calculated as per standard definitions (TP/(FP+TP), TP/(TP+FN), respectively).

True positive (TP): A PTx-binding sequence that was identified by another PTx-binding sequence. False positive (FP): A non-PTx-binding sequence that was identified by a PTx-binding sequence. False negative (FN): A PTx-binding sequence that was not identified by any PTx-binding sequence.

2.3.2.9 In silico developability assessment

We calculated the *in silico* developability metrics of total CDR length, PSH, patches of positive charge, patches of negative charge and structural Fv charge symmetry parameter using the Therapeutic Antibody Profiler tool (TAP) (Raybould, Marks, Krawczyk, et al., 2019). The tool calculates these metrics from homology models built using ABodyBuilder (Leem et al., 2016) and compares them to a database of all clinical-stage therapeutics (there were 377 in the CST database at the time of the experiment) (Raybould, Marks, Krawczyk, et al., 2019). According to the metric in question, values are flagged as "amber" (extreme) if they lie outside the upper and lower, or upper or lower, 5% of observed values among CSTs. An antibody is flagged as "red" if the value of a particular metric is outside of the observed range in CSTs.

2.3.2.10 Ab-Ligity and SPACE

For the CoV-AbDab benchmark, we benchmarked against two other methods: Ab-Ligity and SPACE, as well as a baseline which was grouping antibodies with the same length CDRs.

Ab-Ligity was used for two applications: to delineate which antibodies were binding to the same epitope (using Ab-Ligity's "epitope" mode on crystal structures) and to make predictions about which antibodies bound to the same epitope (using Ab-Ligity's "predicted paratope" mode on ABlooper models). As per the original publication, we used a score of 0.1 to group both epitopes and paratopes (Wong et al., 2021). We explored (paratope) clustering with CDR length matching and without.

SPACE is an algorithm for computational epitope binning via CDR RMSD clustering (Robinson et al., 2021). This was originally applied to FREAD templates but we implemented it for ABlooper models. For the CDR RMSD (root-mean-square deviation) calculation, the Cα atoms of framework residues were aligned using single-value decomposition (implemented with the Biopython

SVDSuperimposer module (Cock et al., 2009)). RMSD was calculated over the Cα atoms of length-matched CDRs. A threshold of 0.75Å CDR RMSD was selected as per the SPACE paper (Robinson et al., 2021).

2.4 Results

2.4.1 Epitope convergence can be identified at the level of paratope residues

Antibodies from different clonotypes have been observed to converge on the same binding site (Joyce et al., 2016; Rijal et al., 2019; Scheid et al., 2011; Wong et al., 2021). We hypothesize that these functionally convergent antibodies may use the same paratope for interaction, and examined SAbDab for evidence that antibodies with similar paratopes bind to the same epitope. We defined the epitope and paratope as those residues with any atom within 4.5Å of any residue in the cognate antibody or antigen, respectively, and consider only the heavy chains of the antibodies because in most BCR sequencing experiments only heavy chain sequences are available. The threshold of 4.5Å was selected as it is a common threshold used in previous work e.g. (Krawczyk et al., 2014; Liberis et al., 2018; Wong et al., 2021). Other thresholds are used, most commonly 4Å (Clifford et al., 2022; Jespersen et al., 2019) or 5Å (Akbar et al., 2021; Jubb et al., 2017).

Figure 2.3 shows an example of a pair of antibody-antigen complex structures where antibodies binding to the same epitope derive from different clonotypes based on their heavy chains, but have very similar heavy chain paratopes. The two monoclonal antibodies 4C2 (PDB ID: 5do2) and D12 (PDB ID: 4zpv) bind to the receptor-binding domain of the MERS-CoV spike protein at the same epitope (94.7% of epitope residues are shared across the pair of structures). In the standard clonotyping definition, an exact IGHV and IGHJ match is required for grouping of two sequences regardless of treatment of the CDRH3; in this instance, the antibodies differ in both IGHV (IGHV5-6-4 vs IGHV5-9-1) and IGHJ gene (IGHJ2/IGHJ4) and have lower CDRH3 identity (66.7%) than common

definitions. However, 4C2 and D12 use largely the same paratope residues to achieve this epitope convergence, with 81.3% of heavy chain paratope residues conserved across the structures. Another example of epitope convergence between antibodies with differing IGHJ genes (IGHJ1 versus IGHJ3 respectively) and sub-80% (57.1%) CDRH3 identity can be seen in the anti-lysozyme antibodies HyHel-10 and HyHel-63 (Figure 2.4).

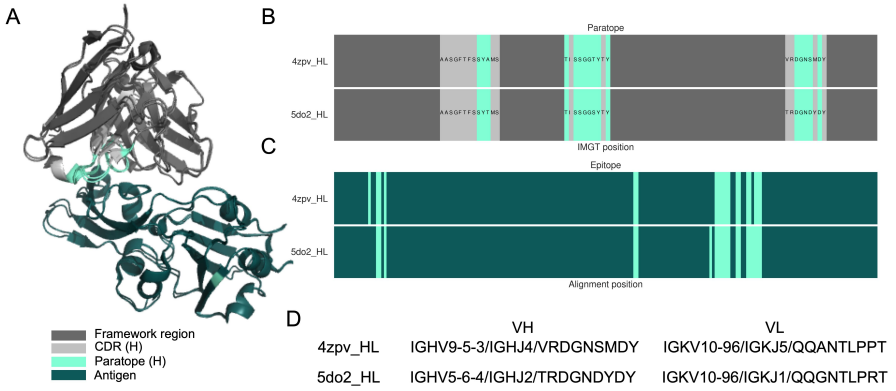

Figure 2.3: | **Example 1 of two antibodies that bind to the same epitope but derive from different clonotypes.** The murine 4C2 (PDB ID: 5do2) and D12 (PDB ID: 4zpv) anti-MERS-CoV antibodies target the same residues on the receptor-binding domain of the spike protein, which can be seen from alignment of their antibody/antigen complex structures (a). 4C2 and D12 use over 80% of the same heavy chain paratope residues (b) and target 95% of the same epitope residues (c). The antibodies are derived from different IGHV and IGHJ genes and display CDRH3 amino acid identity (66.7%) below standard clonotyping definitions (80 - 100%) (d), so would not be considered within the same clonotype.

2.4.2 Paratyping and clonotyping successfully cluster PTx binders in a single-cell dataset

In our novel method, paratyping, antibodies with the same length CDRs (North definition) and above a threshold level of sequence identity across the predicted paratope residues are grouped into the same paratype.

To test the ability of paratyping and clonotyping to group antibodies that target the same epitope, we performed a test in a single-cell (paired VH:VL) dataset of 1290 antibodies isolated from genetically engineered mice that have

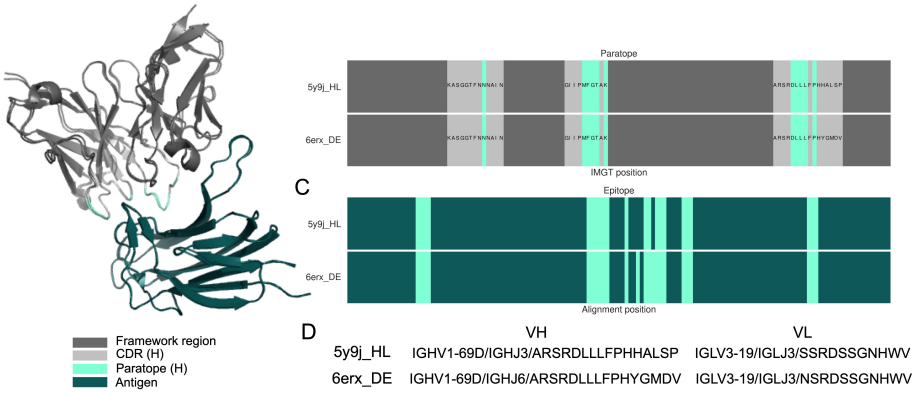

Figure 2.4: | **Example 2 of two antibodies that bind to the same epitope but derive from different clonotypes.** HyHel-10 (PDB ID: 1c08) and HyHel-63 (PDB ID: 1dqj) anti-hen egg white lysozyme antibodies target the same residues, despite being derived from different J genes and displaying CDRH3 amino acid identity (57.1%) below the standard clonotyping definition (80 - 100%) (C). The antibodies use the same paratope residues (100% paratope identity) (B) to achieve this functional convergence (95.7% epitope identity). The paratope and epitope are defined as those residues with any atom within 4.5Å of any residue in the antigen or antibody respectively. Epitope and paratope identity correspond to sequence identity at equivalent epitope and paratope residues in the antigen and antibody alignments respectively, where the denominator is the minimum number of epitope or paratope residues in the pair of structures being compared.

a full set of human immunoglobulin variable region genes(Lee et al., 2014) immunized with PTx. Although we had pairing information within the single-cell dataset, given that the majority of repertoire sequencing data to date is heavy-chain only (Kovaltsuk et al., 2018; Olsen, Boyles, et al., 2022), we demonstrated the method using only the heavy chain information. The sequences were annotated with a PTx-binding (364) or non-binding label (926) (using HTRF and SPR), and we used paratyping (our new method) and clonotyping (the conventional approach) to identify PTx-binding sequences. VH and VK/VL gene frequencies of these antibodies as well as somatic hypermutation counts can be seen in Figure 2.5.

For each of the 364 PTx-binders in turn, we mimicked a repertoire mining experiment by using paratyping or clonotyping to identify binders amongst the remaining 1289 sequences (one-vs-all cross-validation). Each of the PTx-binders is referred to as a "probe" antibody; sequences that are within the same

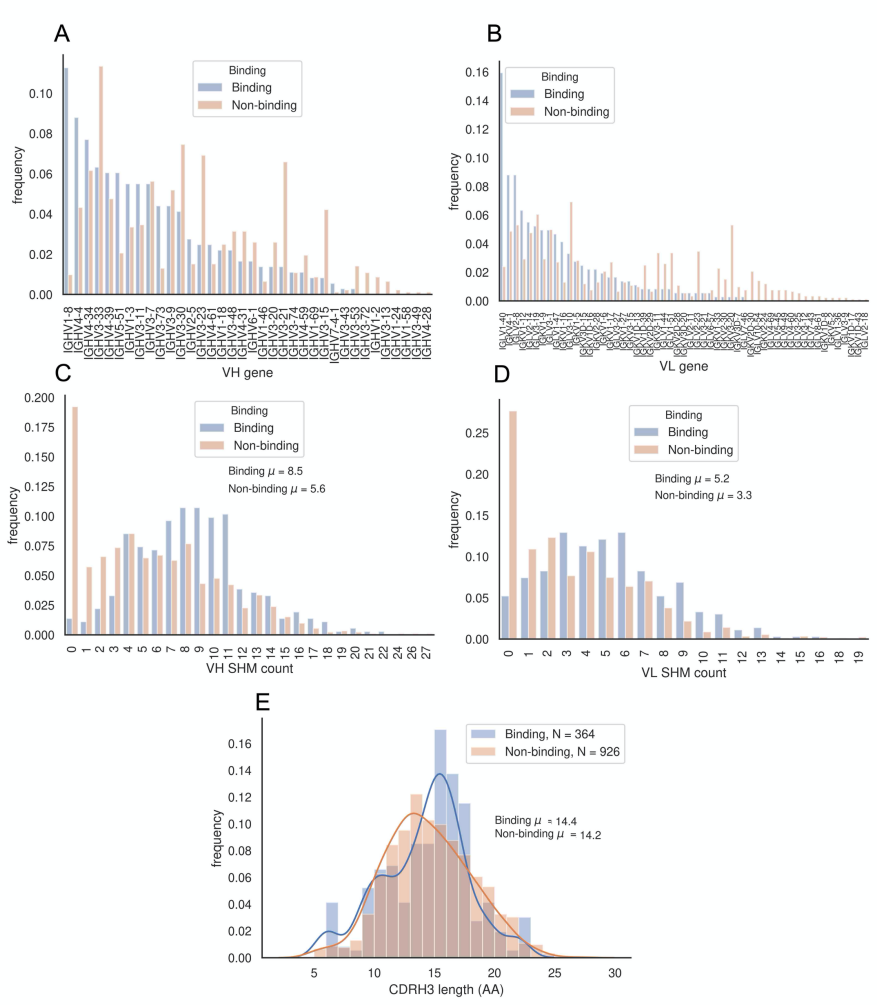

Figure 2.5: | **IGHV, IGKLV and SHM frequencies and CDRH3 length distribution in known PTx binders.** (A) IGHV frequencies in the binding and non-binding populations of antibodies from the single-cell experiment (N = 1290, of which 364 are binding and 926 non-binding). Over 10% of PTx-binding antibodies are derived from IGHV1-8. (B) IGKLV frequencies in the binding and non-binding antibodies. The most common light chain in binding antibodies is IGLV1-40. Somatic hypermutation (SHM) counts across the IGHV (C) and IGKLV (D) counts reveal that the non-binding antibodies have on average lower SHM counts across both VH and VL chains. (E) There is a small but significant difference in the average CDRH3 lengths in the two populations, with CDRH3 lengths of the binding antibodies being significantly greater than those of the non-binding antibodies (average of 14.4 vs. 14.2 amino acids respectively).

paratype or clonotype as the probe are predicted to bind PTx. Our hypothesis is that antibodies with similar paratopes will bind to similar epitopes, but our

method does not require that there are unique paratope/epitope pairings; antibodies with different paratopes may bind to the same epitope, and this would contribute to the false-negative rate of the method. The precision and recall of the two methods (calculated over the aggregate of predictions) for repertoire mining are comparable (Figure 2.6, Table 2.2). The precision and recall using clonotyping and paratyping with varying CDRH3 sequence identity or paratope sequence identity thresholds, respectively, are shown in Figure 2.6. The methods require different sequence identity thresholds for optimum performance but have similar precision-recall profiles.

For clonotyping, the optimal heavy-chain only threshold is 72% CDRH3 amino acid identity. For paratyping, the optimum occurs at 75% paratope identity. These thresholds are optimal as they maximize both precision and recall, corresponding to the "shoulder" of the precision-recall curve. At these optimal thresholds, clonotyping recovers binders with 83% precision and 79% recall (meaning that 21% of the binders in this dataset are not related by clonotype to any other). Paratyping recovers binders with a precision of 84% and a recall of 76% (meaning that 24% of binders in this dataset have distinct paratopes). We expect that this would be an overestimation of performance in the bulk dataset due to the enrichment of PTx-binders created through antigen-specific sorting.

For each probe, a prediction can be made by both paratyping and clonotyping (hereafter labeled a "Both" prediction), paratyping alone (labeled "Paratype-only") or clonotyping alone (labeled "Clonotype-only"). Out of all of these predictions, 76.6% are made by both methods with 89 "paratype-only" predictions (precision: 76%) and 345 "clonotype-only" predictions (precision: 84%). Paratyping and clonotyping make a number of method-exclusive predictions, as shown in Figure 2.7, where the probe antibodies with the largest number of "paratype-only" and "clonotype-only" predictions are shown in dendrograms with these predictions. Figure 2.7 also shows the variability of the method's performance across probes, with the probe antibodies yielding the lowest and highest precision shown. One probe identified 20 other PTx-binding antibodies with 100% precision, while another had just 33% precision. The heterogeneity

in predictions across probe emphasizes the utility of either method when only a small number of binders are known, despite the similar behavior of the methods over the aggregate of probe antibodies.

Paratyping-only predictions are PTx-binding heavy chain sequences from different clonotypes to the probe antibody. There are 89 predictions, which can be split into three groups: 1) those with different V genes (4); 2) those with the same V genes but different J genes (38); and 3) those with the same V and J genes but with CDRH3 identity below 72.0% (47). For example, a pair of binders with just 40% H3 sequence identity but 80% predicted paratope identity were clustered together. Homology modeling of the sequences using ABodyBuilder suggests that the pair of sequences would also be predicted to be highly structurally similar. Antibodies within the same paratype are predicted to bind to the same epitope, as it is highly improbable that sequence and structurally similar antibodies would bind to completely separate epitopes on the same antigen. This hypothesis could be experimentally validated by solving structures of antibodies in complex with antigen, or with epitope competition data.

Method	Sequence identity threshold	Precision	Recall
Paratyping	75%	84%	76%
Clonotyping	72%	83%	79%

Table 2.2: | **Precision-recall values for prediction of PTx-binding.** Precision-recall values for prediction of PTx-binding according to paratyping and clonotyping at the optimal thresholds of 75% and 72%, respectively.

2.4.3 Paratyping enables the identification of novel anti-PTx antibodies from different clonotypes to known binders in prospective experiment

2.4.3.1 Paratyping and clonotyping identify different binders in a bulk repertoire experiment

In order to test how the methods would scale in a larger and non-enriched dataset, we performed a prospective repertoire mining experiment in which the heavy chain sequences from PTx-binding antibodies were used to identify

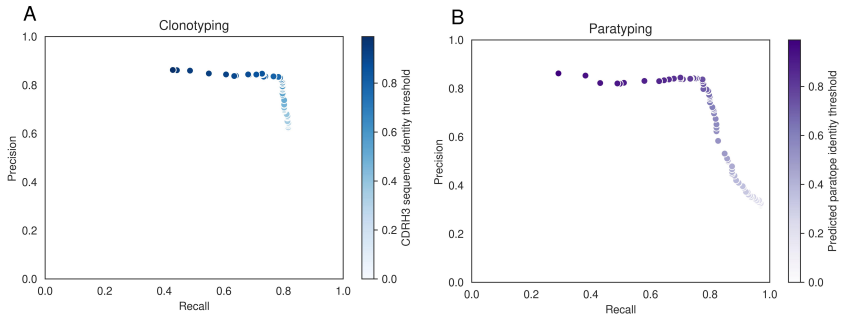

Figure 2.6: | **Precision-recall curves for clonotyping and paratyping in the task of predicting PTx binding.** PR curves for clonotyping (a) and paratyping (b) in the task of predicting PTx binding. Precision and recall values are calculated over a range of CDRH3 sequence identity and predicted paratope identity thresholds respectively, between 0.0% and 100.0% identity. The precision–recall profiles are similar.

novel PTx-binding heavy chains from a set of bulk heavy chain repertoires.

We first looked at how the methods scale in terms of number of predictions, and then experimentally validated the PTx-reactivity of a number of the predictions made by both clonotyping and paratyping.

We used the thresholds determined in the single-cell section. Sequences were considered to be in the same paratype if they have the same length CDRs (North definition) and above 75% amino acid sequence identity in the predicted paratope. Sequences were considered to be in the same clonotype if they have the same V and J genes, the same length CDRH3 (North definition) and above 72% amino acid sequence identity across the CDRH3. The 72% value is below the most common thresholds used for clonotyping (80 - 100%), but it was chosen for fair comparison with paratyping as the threshold providing the best performance in the single-cell section of "Results".

Using the 364 known PTx binders as probes, 59,107 heavy chain sequences from bulk sequencing repertoires were searched for paratype- or clonotype-related sequences. A total of 4269 sequences were identified by both clonotyping and paratyping, 1113 by paratype-only and 1077 by clonotype-only.

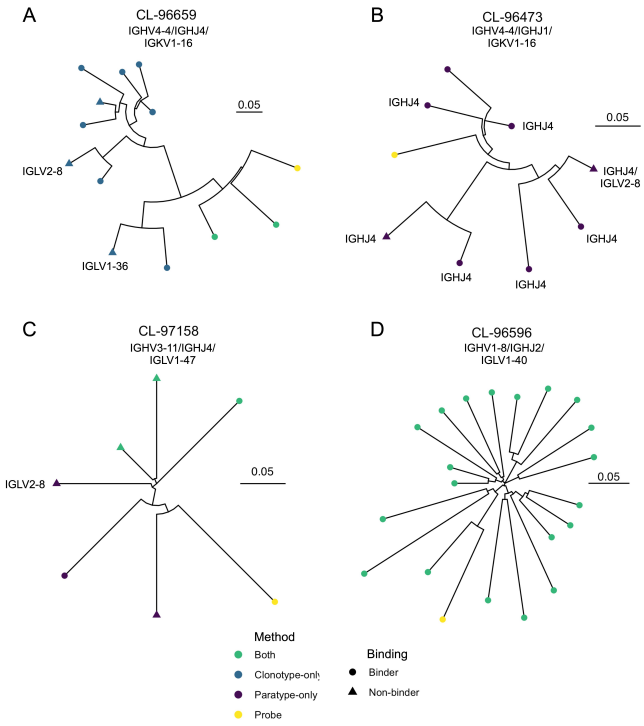

Figure 2.7: | **Representative dendrograms from the single-cell repertoire mining experiment.** Probe sequences (yellow) are known PTx binding sequences. Other sequences which are in the same paratype or clonotype are predicted to also bind PTx. These predicted PTx binding sequences are colored according to whether they are identified by both paratyping and clonotyping ("Both"), paratyping but not clonotyping ("Paratype-only"), or clonotyping but not paratyping ("Clonotype-only"). Circular leaves represent true PTx binding antibodies (i.e. true positives) while triangular leaves represent sequences that do not bind PTx (false positives). Dendrogram A shows the probe which had the most "clonotype-only" predictions, of which 70% are true positives; dendrogram B shows the probe with the most "paratype-only" predictions, of which 75% are true positives; dendrogram C is the probe antibody which is associated with the most false positives (33% true positives) while dendrogram D is the probe antibody associated with the most true positives (20 true positives). Performance is heterogeneous across probes ranging between 33% and 100% precision; precision and recall values reported elsewhere consider the performance in aggregate of all 364 probes. IGHV, IGHJ or IGK/LV genes are annotated where this changes within a dendrogram.

The 1113 paratype-only predictions can be categorized as those with a different IGHV gene with respect to the probe (179); of the 934 predictions with the same IGHV gene, 396 have a different IGHJ gene; the remaining 538 predic-

tions have sub-72% CDRH3 identity to the probe.

2.4.3.2 Prospective experimental validation of novel PTx-binding sequences

For the experimental validation of predicted PTx binders and non-binders, we created a category of prediction more stringent than the "paratype-only" or "clonotype-only" categories to show the utility of paratyping in a real antibody discovery experiment context where multiple probe sequences are available. As shown in Figure 3, a particular probe antibody may make predictions via clonotype or paratype alone. As reflected in the similarity in the precision-recall values calculated over the aggregate of probe antibodies in the single-cell one-versus-all cross-validation, these "method-unique" predictions become rarer when considering a larger number of probes. Such predictions, which could not be found by another method even when using the full complement of known binders, are referred to as "paratype-unique" or "clonotype-unique" predictions.

Using the 97 probes of highest SPR affinity from the single-cell dataset, 2193 heavy chains were predicted to bind PTx. Of the potential 2193 predicted PTx binders, 139 were selected for expression and PTx-binding assay. An additional 48 antibodies predicted to not bind PTx (due to paratope identity or shared clonotype with confirmed non-binders) were also tested. The predictions were split into thirds according to whether they were predicted by both methods (labeled "both") or were unique across all binders for paratyping ("paratype-unique") or clonotyping ("clonotype-unique").

Of the 43 novel heavy chains predicted to bind PTx by both paratyping and clonotyping, 39 (90%) were experimentally confirmed PTx binders. Paratope identity between known and predicted binders ranged between 83% and 100%, with CDRH3 identity ranging between 77% and 100%. Thirty-one of the 48 (65%) of the clonotype-unique predictions were true PTx binders, with minimally 57% predicted paratope identity to a known binder, 72% CDRH3 identity

and 74% CDRH identity (amino acid identity calculated over the heavy chain CDRs). Fourteen of the 48 (30%) paratype-unique PTx binders bound PTx. The minimal CDRH3 identity of a PTx binder to any known binder was 56% with 76% paratope identity. The distribution of CDRH3, total CDR and total VH amino acid identity of novel PTx-binding heavy chains to known PTx-binding antibodies is shown in Figure 2.8. None of the 48 predicted PTx non-binders bound PTx.

As in the single-cell dataset, the success rate in the predictions that were made by both clonotyping and paratyping is higher than either method alone. The success rate of paratype-unique predictions is significantly lower than that of clonotype-unique predictions. However, it may not be appropriate to compare performance across different probe antibodies, some of which may be liable to activity cliffs (a concept from small-molecule chemistry where a compound exhibits a large change in activity given only a small change in structure (Hu et al., 2013)). A direct comparison can be made where both clonotype-unique and paratype-unique predictions were made using the same known PTx binder. This occurred for six PTx probes, and across these, an average precision of 75% for paratype-unique and 92% for clonotype-unique was observed.

2.4.3.3 Discovery of novel anti-PTx antibodies from different clonotypes

Paratyping identified PTx-binding antibodies that could not be found using clonotyping ("paratype-unique"), for example, those using a different IGHV gene to any of the known PTx binders. An example is shown in Figure 2.9, where the original antibody used the inherently autoreactive IGHV gene IGHV4-34 (Schickel et al., 2017) which may be problematic in development. However, paratyping recovers seven PTx-reactive antibodies that use the IGHV4-59 gene segment instead.

Paratyping also recovered novel PTx-binding heavy chains that derive from different J genes and examples with CDRH3 identities well below most clono-

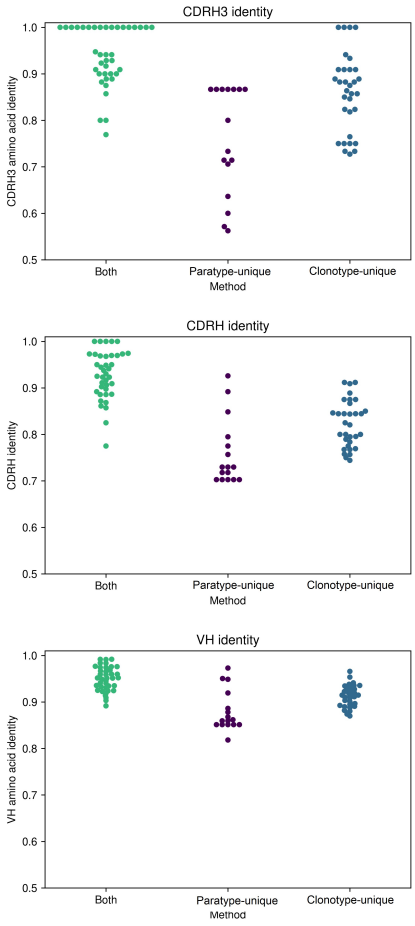

Figure 2.8: | **CDRH3, total CDR and total VH amino acid identity of novel PTx-binding antibodies to the known PTx-binding antibody by which they were identified.** Paratyping enables the discovery of PTx-binding sequences with lower sequence identity across each of these regions with minimally 56% CDRH3 identity, 70% CDRH identity and 80% total VH identity. The line of points in all plots for the paratype-unique predictions correspond to 7 predictions which were made with a single probe.

typing thresholds (commonly 80 - 100%). The minimal CDRH3 identity of a validated PTx-binding antibody to any known binder was 56%, suggesting that paratyping can identify antibodies that bind to the same epitope that could not be found by any clonotyping method.

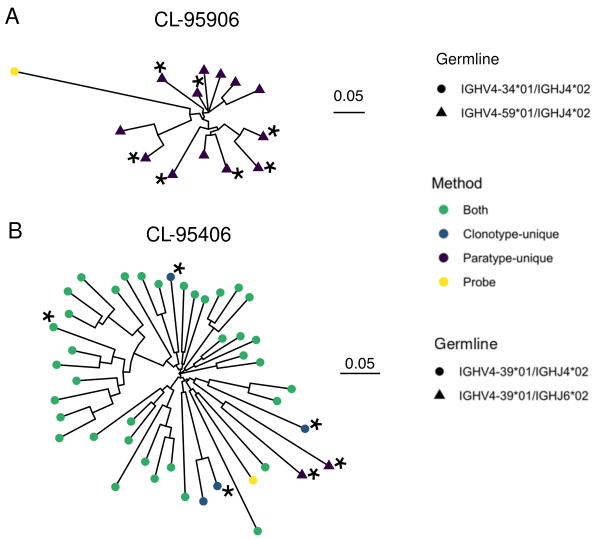

Figure 2.9: | **Dendrograms showing two examples of immune repertoire mining, using known PTx binders CL-95906 and CL-95940.** Heavy chains predicted to bind PTx are colored as green, blue or purple depending on whether they are predicted to bind via both methods or were clonotype- or paratype-unique, while the probe antibody is coloured yellow. Asterisks indicate heavy chains selected for testing, all of which were validated as PTx binding. A shows sequences predicted to bind PTx that use a different IGHV gene (IGHV4-59) than the known PTx binder used for prediction (IGHV4-34). Sequences using a different J gene to the known PTx binder are shown in B.

2.4.3.4 Repertoire mining can improve in silico developability metrics

One of the limitations of clonotyping as a method for immune repertoire mining is the relatively narrow sequence space within which it is capable of making predictions, meaning that the discovered antibodies may have conserved developability problems. We have already seen one example where paratyping's ability to jump between germlines allows us to avoid an autoreactive IGHV gene-derived antibody; other developability problems such as aggregation propensity may also be improved by using paratyping.

Of the original 97 antibodies used for repertoire mining, 38 antibodies were flagged by the Therapeutic Antibody Profiler (TAP) tool as having possible developability issues due to CDR length, high density of charge or hydrophobicity, or charge asymmetry between the heavy and light chains. As paratyping only groups antibodies with the same length CDRs, we considered only the latter four developability metrics. Twenty-six of the original probe antibodies were flagged with extreme (2) or unobserved (24) values. Values are considered extreme (flagged as amber by TAP) if they fall within either the top or bottom 5% of the distribution of values observed in a set of 377 clinical-stage therapeutics (CSTs) (Raybould, Marks, Krawczyk, et al., 2019). Values are unobserved (flagged as red by TAP) if they are outside of the range of values observed in existing CSTs. Seventeen of these probes were used in the identification of novel antibodies assayed for PTx binding, of which 13 successfully identified new PTx binders. For four of these probes, one of the new PTx-binding antibodies identified showed a sufficiently large change in the flagged developability metric that the flag was removed. There were a further five new PTx-binding antibodies showing a shift in the flagged developability metric toward the mean value among CSTs.

Figure 2.10 shows the improvement in patch surface hydrophobicity achieved by immune repertoire mining using CL-95375 as a probe antibody. CL-95375 had an amber flag for this metric. Repertoire mining was used to identify a number of predicted PTx-binding antibodies. It can be seen that the more sequence-distinct paratype-only predictions are able to achieve greater changes in patch surface hydrophobicity (PSH). These predictions were not assayed as they were within the clonotype of another known binder, and therefore not "paratype-unique".

We identified a further three examples in which either PSH (Figures 2.11A and C) or charge asymmetry between the heavy and light chains, SFvCSP (Figure 2.11B) was found to be more CST-like in novel PTx-binding antibodies mined

from the bulk repertoires.

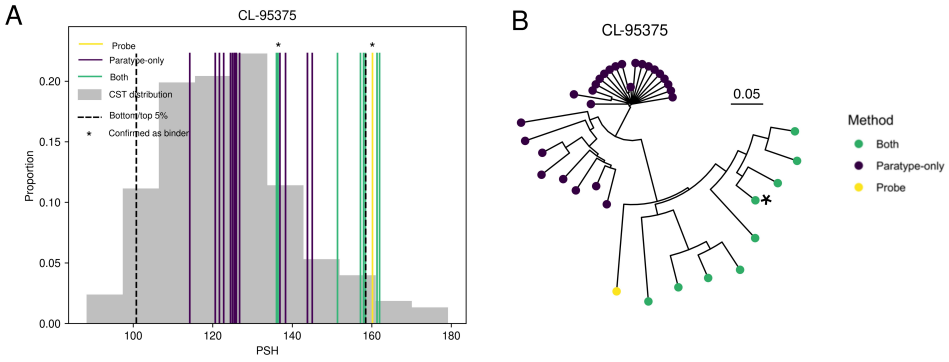

Figure 2.10: | **An example of a probe antibody with a flagged developability issue significantly improved by immune repertoire mining.** The original PTx antibody had high patch surface hydrophobicity (PSH). Clonotyping and paratyping were used to discover a novel PTx-binding antibody without a developability flag. This shift is shown in A, where asterisks indicate the probe antibody, CL-95375 (yellow), and the validated novel PTx-binding antibody which has a shift in PSH toward the mean value among 377 clinical-stage therapeutics. The extremes of the distribution (upper and lower 5% for PSH) are shown as the black dashed lines. The distribution of the metric among antibodies within the same paratype and/or clonotype (colored according to the legend) is shown; the only antibody assayed, which was validated as PTx-binding, is marked with an asterisk. The predicted PTx binders in A are shown in the dendrogram in B, with the confirmed novel PTx binder marked by an asterisk.

2.4.4 Paratyping groups antibodies to the same epitope in five antigen test cases

The PTx experiments demonstrated that paratyping was able to group antibodies binding to the same antigen. We hypothesise that we achieved success in this binary binding/non-binding prediction because of the hypothesis that antibodies with similar paratopes bind to the same epitope. We did not have epitope resolution for these PTx-binding antibodies to test the hypothesis that the antibodies with similar paratopes competed for the same epitope.

As we expect our method to generalise across protein antigens, we looked at two separate sources of epitope data: CoV-AbDab, a database of antibodies and nanobodies known to bind to betacoronavirus proteins labelled with course epitope details (Raybould, Kovaltsuk, et al., 2021), and crystallographically-solved

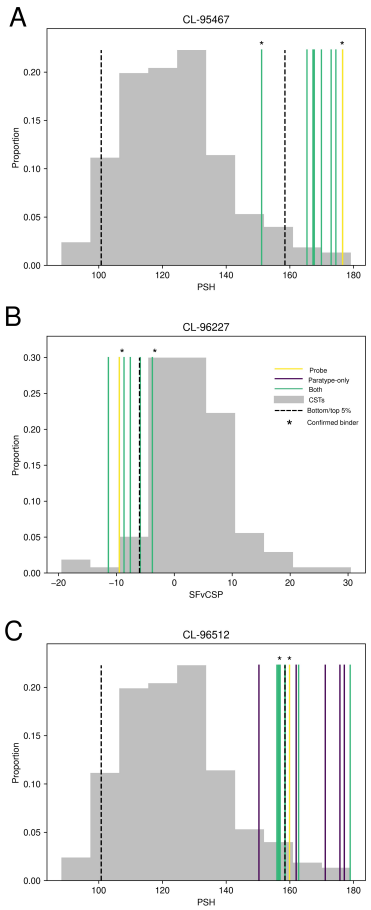

Figure 2.11: | **Three further examples of probe antibodies with flagged developability issues.** In figures A and C, patch surface hydrophobicity (PSH) is flagged as an issue; in figure B, charge asymmetry between the heavy and light chains (SFvCSP) is predicted to be problematic. These developability flags were removed in putative PTx binders identified by immune repertoire mining. Asterisks indicate confirmed PTx-binding antibodies, the probe antibody (yellow) and binders identified through prospective experimentation. The putative PTx-binding antibodies were identified by both clonotyping and paratyping (green), paratyping only (purple) or clonotyping only (blue). In these three instances, the change in the property is significant enough that the identified PTx-binding antibody discovered through repertoire mining is no longer flagged.

antibody:antigen complexes in SAbDab (Dunbar et al., 2014; Schneider et al., 2022).

In CoV-AbDab (download dated to May 2020), paratyping identified two

Name	IGHV	IGHJ	CDRH3	Epitope
MERS-4	IGHV3-30	IGHJ4	AGNDY	S; RBD
MERS-4V2	IGHV3-30-3	IGHJ4	TNTYY	S; RBD
27D	IGHV3-53	IGHJ4	ARNFGEDFDY	S; RBD
6A	IGHV3-66	IGHJ4	AKDYGEGFDY	S; RBD

Table 2.3: | **Two pairs of antibodies from CoV-AbDab (May 2020) where the sequences derived from different clonotypes.** Two examples where antibodies with different IGHV genes and sub 80% CDRH3 identity, belong to the same paratype and both show binding to SARS-CoV-2 spike RBD. It must be noted that IGHV3-30-3 and IGHV3-30, and IGHV3-53 and IGHV3-66 are among the most similar IGHV genes having up to 100% and 99.0% amino acid identity to one another respectively: see Appendix Figure 1 for a visualisation in context with other IGHV genes.

pairs of antibodies from different clonotypes binding to the same epitope, MERS-4 and MERS-4-V2, and 27D and 6A (Table 2.3).

We further analyzed five separate antigen examples for which there were multiple unique antibodies solved in complex with antigen (hemagglutinin, gp120, SARS CoV-2 spike protein, gp160 and lysozyme) to evaluate the hypothesis that antibodies within the same paratype bind to the same epitope. All pairs of antibodies within the same paratype shared some epitope residues, and would therefore be likely to compete for the same epitope (see Table 2.4).

2.4.5 Paratyping in the context of new methods

Concurrent with the development of paratyping, Ab-Ligity was developed by Dr. Catherine Wong (Wong et al., 2021). Ab-Ligity creates a structural and physicochemical representation of a given protein surface that allows an alignment-free comparison of those surfaces. While the method was developed and benchmarked on SAbDab (with access to crystal structures and known paratopes and epitopes), it was also shown to work on homology models of antibodies using paratope prediction with Parapred (Liberis et al., 2018). A key advantage of the method is that it does not require CDR length-matching for the comparison.

Later, Robinson and Raybould demonstrated that clustering antibody model

Antigen	# Antibodies	Positives	Paratype predictions	Paratype positives	Minimum epitope identity (%)
Hemagglutinin	79	84	11	8	23.1
gp120	77	132	16	7	16.7
SARS CoV-2 spike protein	44	21	2	2	58.3
gp160	39	98	33	30	33.3
Lysozyme	33	221	144	144	53.8

Table 2.4: | **Evaluation of the hypothesis that paratyping groups antibodies which bind to the same epitope on five antigens from SAbDab (October 2020).** For each antigen, the dataset of antibody/antigen structures is made non-redundant at the level of 100% sequence identity across the heavy chain CDRs. Epitope similarity for each pair of antibodies binding to the same antigen is evaluated using Ab-Ligity (Wong et al., 2021), a stringent definition considering physicochemical and structural similarity. True positives are pairs of antibodies with an Ab-Ligity epitope score greater than or equal to 0.1. The heavy chain predicted paratope of each pair of antibodies with the same length CDRHs is compared. As per the paratype definition, antibodies with greater than 75% predicted paratope identity are within the same paratype and predicted to bind to the same epitope. All pairs of antibodies predicted to bind to the same epitope share epitope residues (minimally 16.7%) and could be argued to be at the same epitope. However, using the Ab-Ligity definition there is a false positive rate between 0% (SARS-CoV-2 spike, lysozyme) and 56.3% (gp120).

Method	Chain	Precision	Recall	Threshold
SPACE	H	0.32	0.94	3
SPACE	H + L	0.50	0.83	3
Ab-Ligity (CDR length matched)	H	0.74	0.73	0.1
Ab-Ligity (CDR length matched)	H + L	0.61	0.61	0.1
Ab-Ligity (non-length matched)	H	0.49	0.78	0.1
Ab-Ligity (non-length matched)	H + L	0.35	0.83	0.1
Paratyping	H	0.84	0.76	0.75
Paratyping	H + L	0.85	0.75	0.75
Clonotyping	H	0.83	0.79	0.72
Clonotyping	H + L	0.87	0.73	0.72

Table 2.5: | **Performance of SPACE, Ab-Ligity, paratyping and clonotyping at the task of grouping antibodies that bind to PTx.** Paratyping and clonotyping achieve considerably higher precision than either RMSD-based clustering or Ab-Ligity.

CDRs by Cα RMSD with a threshold of 0.75Å (where all CDRs were of the same length) could group antibodies to the same epitope domains on the SARS-CoV-2 spike protein with high precision (92% of non-singleton clusters contained only antibodies to the same domain). This method is referred to as SPACE.

Ab-Ligity and SPACE clustering methods' performance on the PTx benchmark are shown in Table 2.5. Paratyping achieves the greatest precision, while recall is higher for CDR RMSD-based methods such as SPACE but precision considerably lower. Ab-Ligity shows the worst performance in terms of poor precision and recall that is comparable to or lower than paratyping or clonotyping.

We performed a further test to benchmark these methods using CoV-AbDab, which is the largest database of antigen-specific antibodies with labelled epitope information and sequences (Raybould, Kovaltsuk, et al., 2021).

As of July 2022, there were 308 VH:VL antibody sequences solved in complex with their antigen. 270 of these 308 antibodies could be modelled with high confidence using ABlooper (Abanades et al., 2022). We assessed the performance of paratyping, SPACE (CDR Cα RMSD), Ab-Ligity and clonotyping

at grouping these antibodies based on i) their Ab-Ligity epitope bins and ii) the CoV-AbDab antibody domain annotations. This benchmark allows comparison of the methods with standardised modelling and allows us to assess the effect of different levels of epitope resolution on the performance of the methods.

In this benchmark, Ab-Ligity outperforms clonotyping, paratyping and SPACE in the task of grouping epitope-specific antibodies, particularly when considering the higher-resolution Ab-Ligity epitope definition. Ab-Ligity achieves 75% precision with four and six times higher recall than SPACE or paratyping respectively. When imposing CDR length matching, the performance of Ab-Ligity falls and is comparable with SPACE. This is in contrast to the PTx benchmark in which CDR length matching increased precision considerably. This demonstrates the importance of the benchmark.

Part of the value of each of the structure-based methods is how sequence-diverse the antibodies are that can be identified as binding to the same epitope. The minimum CDR sequence identities of true positive pairs identified by each method are shown in Table 2.7 where it can be seen that Ab-Ligity identifies the most distinct pair of antibodies binding to the same epitope, and paratyping the least.

2.5 Discussion

Characterizing the functional relationship between sequence-distinct antibodies is an important step in our understanding of the adaptive immune landscape. Mapping antigen preference to antibody repertoires will allow us to identify epitope convergence of antibodies at a large scale. In a test system of transgenic mice immunized with PTx, we show for the first time that prediction and comparison of paratopes can be used to group antigen-specific antibodies in both an enriched, single-cell dataset and non-enriched bulk heavy chain repertoires. We demonstrate the utility of the method in the context of an antibody discovery experiment alongside the conventional approach of clono-

Method	Epitope definition	# Positives	# Predictions	Precision	Recall
Clonotyping	Domain	27,700	54	1.0	0.002
Paratyping	Domain	27,700	64	1.0	0.002
SPACE	Domain	27,700	140	0.98	0.005
Ab-Ligity	Domain	27,700	387	0.99	0.01
Ab-Ligity (CDR length match)	Domain	27,700	123	0.98	0.004
CDR length matching	Domain	27,700	207	0.88	0.007
Clonotyping	Ab-Ligity	2,429	54	0.72	0.02
Paratyping	Ab-Ligity	2,429	64	0.69	0.02
SPACE	Ab-Ligity	2,429	140	0.57	0.03
Ab-Ligity	Ab-Ligity	2,429	387	0.75	0.12
Ab-Ligity (CDR length match)	Ab-Ligity	2,429	123	0.57	0.02
CDR length matching	Ab-Ligity	2,429	207	0.49	0.04

Table 2.6: | **Performance in terms of precision and recall of methods in epitope binning in CoV-AbDab.** Clonotyping, paratyping and the structure-based methods of SPACE (Robinson et al., 2021) and Ab-Ligity (Wong et al., 2021) in identifying pairs of antibodies to the same epitope. Epitope is either defined by CoV-AbDab domain annotation ("Domain") or via physicochemical similarity of the crystallographically-solved epitope ("Ab-Ligity"). "Positives" refers to the number of pairs of antibodies to the same epitope. "Predictions" refers to how many pairs of antibodies are predicted to bind to the same epitope. The maximum recall is found for Ab-Ligity for both domain and Ab-Ligity epitope categories, but only when not imposing CDR length matching. CDR length matching, which is used for both paratyping and CDR RMSD, caps the maximum recall.

Method	CDR identity	Antibody 1	Antibody 2
Clonotyping	67.4%	D27	m396
Paratyping	76.5%	Omi-12	PDI-222
SPACE	60.0%	BD-508	COVOX-269
Ab-Ligity	50.0%	Beta-27	P4A1
Ab-Ligity (CDR length match)	65.2%	COVOX-150	PDI-42

Table 2.7: | **Sequence identity of antibodies grouped by clonotyping, paratyping, SPACE and Ab-Ligity on the CoV-AbDab benchmark.** The minimum CDR identity of any of the true positives (Ab-Ligity epitope definition) identified by each method. The minimum CDR identity of a true positive pair was found by non-CDR length matched Ab-Ligity. Beta-27 and P4A1 have different lengths CDRL3 and CDRH3 and therefore would not be recovered by any other method. Paratyping exhibited the highest minimum CDR identity within minimally 76.5% CDR identity verus 67.4% for clonotyping.

typing and discover new anti-PTx antibodies from different clonotypes to any known binders.

We first developed the method in a single-cell dataset, where paratyping and clonotyping were able to group PTx-specific antibody heavy chains with high precision (84% and 82%, respectively). These results may not map to the bulk sequencing dataset given that the sequences derived from both plasma-, memory- and antigen-sorted cells (leading to ca. 30% of sequences being PTx-reactive). To validate the method in a less enriched dataset, we performed a prospective experimental test of paratyping in a non-antigen sorted set of bulk heavy chain sequencing repertoires.

In the prospective experimental test, paratyping allowed us to discover new PTx-binding antibodies that we could not have found using clonotyping. These include antibodies that use different germline genes as well as antibodies with lower CDRH3 identity than in common definitions (80 - 100%). In terms of antibody discovery, paratyping allows us to identify sequence-distinct antibodies that bind to the same epitope and that can differ significantly in developability or affinity. In the terms of repertoire analysis, paratyping expands our ability to functionally group antibodies beyond clonotypes, and therefore allows us to detect specific cases of epitope convergence between clonotypes. For example, we found epitope convergence between IGHV4-34/IGHJ6 and IGHV4-59/IGHJ6 clonotypes, IGHV4-39/IGHJ6 and IGHV4-39/IGHJ4 clonotypes, IGHV4-34/IGHJ5 and IGHV4-34/IGHJ4 clonotypes and IGHV5-51/IGHJ4 and IGHV5-51/IGHJ6 clonotypes in PTx binders. We did not observe pairs of antibodies in the same paratype using different IGHV gene subgroups, but antigen-reactive antibodies with identical CDRH3s deriving from different IGHV gene subgroups have been observed, suggesting that this convergence does occur (D'Angelo et al., 2018). Paratope identity across the germline-encoded CDRHs 1 and 2 is equal to or in excess of 75% across members of IGHV1 and IGHV7, and IGHV3 and IGHV4 (Appendix Figure 2), so we predict that it should be possible to use paratyping to find binders from different IGHV gene subgroups, should a large enough sequencing dataset be mined.

The implications of this for immune repertoire clustering are as yet unexplored. Usually, clonal diversity is considered a proxy for so-called "functional diversity" of the repertoire, the number of unique epitopes that can be recognised by the BCR repertoire. Closely-related clonotypes are known to be important in recognising closely-related antigenic variants, e.g. in chronic infections like HIV. If convergence in epitope between unrelated clonotypes is widespread, it is possible that clonotyping overestimates functional diversity, and this may account for low proportions of clonotypes shared across individuals (Briney et al., 2019; Soto et al., 2019).

Success rates in the prospective experimental validation were variable across the categories of prediction. We found that 90% of sequences predicted by both clonotyping and paratyping to be binders bound PTx. The success rate was considerably lower in clonotype-unique predictions (65%) and even lower in paratype-unique predictions (30%). It should be noted that these method-unique predictions form the minority of predictions for either method (15 - 22% of predictions). The lower success rate of paratyping versus clonotyping may be attributable to the particularly low CDRH3 identity of these predictions to known binders - paratyping does not give special weight to the CDRH3 in the paratope identity calculation despite the particular role it plays in antigen complementarity. Predictions with as little as 33% CDRH3 identity to a known binder were assayed, but no antibody with a CDRH3 amino acid identity below 56% bound to PTx, suggesting that paratyping could be further improved by the use of CDRH3 weighting.

The method relies upon the accuracy of the paratope prediction step. We have used a previously published sequence-based paratope prediction model, Parapred (Liberis et al., 2018), which achieved the highest F-score in our benchmark and is fully open-source. We tested other paratope prediction methods as inputs to paratyping using the data from the single-cell experiment, and found that the results were comparable to those with Parapred. Parapred's

convolutional and recurrent neural networks were trained and validated on 277 antibody-antigen complexes with protein antigens. While it is reasonable to expect paratyping to generalize across protein antigens, it may not be reasonable to expect paratyping to generalize across classes of antigens that did not feature in Parapred's training set. Among a number of metrics shown to discern carbohydrate- and protein-binding antibodies, paratope size is a key discriminant. The application of paratope prediction models trained on protein antigens to carbohydrate or peptide antigens could lead to lower success rates in paratyping's predictions.

Paratope prediction only takes around 0.1 s per sequence (of which 0.02 s correspond to CDR extraction) as opposed to 0.05 s per sequence for VDJ annotation using IgBlast (Ye et al., 2013)), which means it is tractable for large datasets unlike homology modeling (on average 30 seconds per sequence, 600 times slower than germline gene annotation). It has the advantage that it does not rely on the upkeep of consistent and complete germline databases (a leading cause of disagreement between germline annotation tools (Smakaj et al., 2020), but rather on the distribution of a pretrained paratope prediction model. The paratope prediction step is purely sequence-based and structural modeling is not required, meaning that immune repertoires can be annotated without accurate germline alignment and without access to large computing power. The method is likely to improve as paratope prediction methods improve.

There is a limitation to the applicability of paratyping which is that it requires CDR length matching. A degree of positional equivalence (within the CDRs) can be obtained by comparing the same IMGT positions regardless of CDR length, but this has not been explored. This is the benefit of the alignment-free method used in Ab-Ligity (Wong et al., 2021).

Deciphering the functional landscape of immune repertoires will greatly improve our understanding of the adaptive immune system. Improving our ability to group antibodies binding to the same epitope is a step toward this. Our

results show here that the simple and computationally rapid abstraction of the antibody binding site used by paratyping is sufficient to group antigen-specific antibodies in a way that provides us with additional information beyond clonotyping. This additional information is particularly significant in the context of antibody discovery, where it allows us to recover different novel antigen-specific antibodies from immune repertoires.

The idea of clustering antibodies that are predicted to share the same epitope has latterly been referred to as *in silico* or computational "epitope binning" (Ghanbarpour et al., 2022; Mahita et al., 2022) or "epitope profiling" (Robinson et al., 2021). Ghanbarpour and colleagues used a similar approach to paratyping of comparing paratope representations. They trained a new paratope predictor via fine-tuning a protein language model and created a "paratope similarity" measure by calculating the Euclidean distance between the resultant embeddings (the embeddings consisted of the language model's embedding of the antibody sequence weighted by the output probability from their paratope predictor). This showed a good correlation with epitope binning on SARS-CoV-2 RBD and EGFR antibodies.

As discussed, the method of Robinson and colleagues (referred to as SPACE) showed that Cα RMSD across length-matched homology modelled CDRs was sufficient to group antibodies to the same domain of SARS-CoV-2 spike or nucleoprotein. Mahita and colleagues explored Cβ RMSD in hypervariable regions, sequence identity and a novel docking-based competition score, referred to as the "EpiBin score" and the ability of each of these metrics to predict competition between 11 anti-IL-6 repebodies. While the EpiBin score was shown to be predictive, the best competition clustering appeared to be achieved by the Cβ RMSD measure indicating that this simple measure can be highly predictive (though it must be noted that the repebody variable regions are considerably less variable than antibody CDRs) (Mahita et al., 2022).

Comparison of paratyping with Ab-Ligity (Wong et al., 2021) and SPACE-

inspired CDR RMSD clustering (Robinson et al., 2021) showed that the relative performance of these methods changed considerably upon considering different benchmarks. In the PTx benchmark, paratyping is the most precise of the structure-based methods. In the CoV-AbDab benchmark, relative performance depends on the epitope definition. At the highest resolution definition (Ab-Ligity epitope based on the crystal structure), Ab-Ligity has the highest precision and recall of all methods and groups the most sequence-distinct antibodies. There is increasing computational cost going from paratyping to RMSD clustering (as this requires modelling) to Ab-Ligity (as this requires both paratope prediction and modelling). The results from the CoV-AbDab benchmark indicate that it may be important to make comparisons across antibodies with different CDR lengths to improve recall.

As described above, the work in this field since the development of paratyping has provided further evidence that paratope comparison can group antibodies to the same epitope. Furthermore, there is evidence that an orthogonal approach which does not use sequence information at all can provide promising results in specific cases.

2.6 Chapter summary

In this chapter, we detail the development of paratyping, a novel structure-based method for identifying antigen-specific antibodies from repertoire data. This involves high-throughput paratope prediction and comparison of these predictions to antibodies where the antigen specificity is known. We developed and validated the model with PTx-binding antibodies and bulk BCR repertoires isolated from the Kymouse transgenic platform (Lee et al., 2014).

We first demonstrated that paratyping could group PTx-binding antibodies with similar precision and recall to the baseline method, clonotyping, but could group antibodies from different genetic backgrounds and with sub-80% CDRH3 amino acid identity. We used this set of antibodies to tune the pre-

dicted paratope identity threshold for paratyping, which was optimal at 75%. For fair comparison, we also tuned the CDRH3 identity threshold and found it to be optimal at 72%.

We then devised a prospective experimental test of the methods by comparing the known PTx-binding antibodies to bulk repertoires of heavy chain sequences from the same immunised Kymice that the antibodies were isolated from. We created three categories of positive prediction based on relationship of predicted binders to known binders: "paratype" or "clonotype-unique" sequences and sequences that were predicted to bind via both methods. These 144 antibodies as well as 48 antibodies predicted not to bind were expressed and tested against PTx. We found that 90% of our predictions based on both clonotyping and paratyping, 65% of clonotype-unique predictions and 30% of paratype-unique predictions bound to PTx. We therefore isolated and report a further 88 PTx-binding antibodies of which 14 were not clonally related to any of the previously known PTx-binding antibodies.

To provide evidence for our hypothesis that these antibodies are clustered on the basis of epitope, we identified five case studies where there are multiple non-clonally related antibodies to the same epitope on the same antigen. This demonstrated that paratyping grouped antibodies that shared at least 16.7% of the same epitope residues.

Finally, we performed an up-to-date benchmark on a database of coronavirus-binding antibodies, CoV-AbDab, to compare our method "paratyping" with two other structure-based methods, CDR RMSD clustering and Ab-Ligity. This demonstrated that paratyping groups antibodies to the same epitope (with both domain-level and residue-level epitope resolution) with high precision. Ab-Ligity with no CDR length constraints outperformed paratyping however, indicating that it is important to compare antibodies with differing CDR lengths in order to maximise recall at this task.

In the following chapter, we explore the application of higher-throughput structure prediction to naive BCR repertoires of humans, mice and the Kymouse, to provide further insight into using human and Ig transgenic repertoires for antibody discovery.

3 | Applying structure prediction to human, mouse and transgenic mouse BCR repertoires

3.1 Abstract

In this chapter we describe our combined sequence and structural analysis of the BCR repertoires of humans, Kymice and mice. We report that while there are differences in CDRH3 length (driven by both differences in germline gene usage and reduced junctional diversification) between naive Kymouse and human repertoires, the structural space occupied by the Kymouse CDRH3s is more human-like than mouse-like. Our combined sequence and structural analysis indicates that the naive Kymouse BCR repertoire is diverse with key similarities to human repertoires.

This chapter contains material reproduced from the following preprint under review:

Richardson, E., Binter, S., Kosmac, M., Ghraichy, M., von Niederhäusern, V., Kovaltsuk, A., ... Watson, S. (2022). Characterisation of the immune repertoire of a humanised transgenic mouse through immunophenotyping and high-throughput sequencing. bioRxiv. (Under review)

All the described work is mine unless otherwise stated.

3.2 Introduction

The Kymouse platform is one of the most advanced transgenic platforms used for antibody discovery and vaccine response modelling (Section 1.5.1.2) and in the 2014 publication of the platform, Kymice were shown to produce somatically-mutated, high-affinity antibodies to a diverse set of antigens (Lee et al., 2014).

The increased availability of BCR sequencing data since 2014 now allows a more comprehensive comparison of Kymouse, mouse and human BCR repertoires. Antibodies with different sequences can have similar structures, and vice-versa, meaning that comparison of these repertoires based on sequence alone does not give a complete picture of the structural and thus functional diversity of these repertoires. Regardless of how well we are able to predict the structure-function link (which was explored in the previous chapter), it is logical that more structurally diverse sets of antibodies are able to target more structurally diverse epitopes.

There have been recent significant advances in the speed and accuracy of antibody structure prediction (Hummer et al., 2022) that have significantly extended our capacity for "repertoire structural annotation" (Kovaltsuk et al., 2017) as described in Section 1.2.8. Here, we apply these structural annotation and prediction techniques to compare the primary BCR repertoires of humans, mice and an Ig-transgenic platform, the Kymouse.

In this chapter, we analyse naive IgM BCR repertoires from humans, mice and Kymice. We first report key differences in germline gene usage and junctional diversity that have implications for the utility of the platform in vaccine response modelling. We then apply high-throughput homology-based structural annotation and deep learning-based structure prediction to representative sequences from these naive IgM BCR repertoires. We find that despite the observed sequence differences, the CDR structures are more human-like than mouse-like.

3.3 Materials and Methods

3.3.1 Datasets

The experimental datasets in sections 3.3.1.1 and 3.3.1.2 were generated by the Infectious Disease team at Kymab (with special thanks to Dr. Simon Watson and Dr. Spela Binter) and the Trück lab (with thanks to Dr. Marie Ghraichy, Valentin von Niederhäusern and Dr. Johannes Trück) respectively.

3.3.1.1 Human BCR sequencing dataset

Buffy coat samples were obtained from ten healthy individuals as described previously (Ghraichy et al., 2021). In the previous study, B cells were FAC-sorted into naive, marginal zone, plasma and switched memory cell populations. In the present study, we analysed IgM sequences from the naive subset of B cells with no mutations. There was a total of 338,677 sequences (mean: 33,867 per human, range: 20,653 - 48,293).

3.3.1.2 Kymouse BCR sequencing dataset

Bone marrows from the femur and tibia of 22 Kymice were processed into single-cell suspension. Seven were FAC sorted to recover CD19+ B220+ B cells into a single tube. The cells were lysed and RT-PCR was performed to amplify the VH domain, followed by standard Illumina library generation, before sequencing on an Illumina MiSeq. The Change-O pipeline (Gupta et al., 2015) was used to process the sequences generated by the MiSeq sequencers. IgM sequences with zero mutations were selected for further analysis resulting in a total of 412,493 unique (nucleotide) sequences across the 7 Kymice (average 58,928 range 31,905 - 100,240).

3.3.1.3 Mouse BCR sequencing dataset

6,763,480 IgM nucleotide sequences from naive B cells of healthy unvaccinated C57BL/6 wild-type mice were downloaded from the Observed Antibody Space (OAS) (Kovaltsuk et al., 2018; Olsen, Boyles, et al., 2022). Those sequences with any nucleotide mutations were removed, and the remaining sequences were down-sampled via stratified sampling aiming to preserve the original clonal structure in the complete dataset. 150,000 sequences with redundancy were randomly selected from each of the five C57BL/6 mice. 150,000 was selected because, based on rarefaction at intervals at 50,000 sequences (from 50,000 to 200,000), it gave the most similar range of sample sizes after redundancy filtering to the Kymouse and human dataset. Collapsing to unique (nucleotide) sequences resulted in a total of 268,285 sequences (mean: 53,657 sequences per mouse; range: 20,026 - 87,041).

3.3.2 Methods

3.3.2.1 Sequence alignment and analysis and clonal inference

Human sequences were aligned via the DefineClones Change-O module (Gupta et al., 2015) to the human IMGT germline gene database, C57BL/6 mice to the mouse IMGT germline gene database and Kymouse sequences to the proprietary Kymouse germline gene database (Lee et al., 2014; Lefranc et al., 2009).

Alignment to these germline databases was used to determine IGHV, IGHD and IGHJ germline origin for each sequence, predict the NP1 and NP2 insertions (the combined N/P region between the IGHV and IGHD and IGHD and IGHJ gene alignments respectively), and delineate the CDRH3. While IGHD gene alignments are generally not used in e.g. clonal lineage assignment due to their inaccuracy, we used IGHD gene information as this was an important point in the previous analysis of Briney and colleagues (Joyce et al., 2020); furthermore, non-mutated sequences from naïve B-cells will be the least challenging subset of sequences for IGHD assignment. As our method relies not on absolute accuracy of IGHD gene assignment but a comparison of statistical

propensities towards different IGHD gene usages, we do not believe this to be problematic in our analysis. We did not look for or explicitly exclude possible IGHD-IGHD gene fusion events, with these representing a very small proportion of the naïve peripheral repertoire (Briney et al., 2012).

Clonotypes were defined as sequences sharing the same IGHV, IGHJ and with greater than or equal to 90% amino acid identity across the CDRH3. This was implemented using the DefineClones module of Change-O (Gupta et al., 2015). It should be noted that this CDRH3 identity threshold is significantly higher than that in Chapter 2: here, it is intended to group clonally-related sequences as opposed to a functional relationship, with the previous threshold of 72% based on a functional task (grouping binders to the same antigen) in a single dataset, vs. a standard clonotyping definition (80 - 100%) (Yaari & Kleinstein, 2015).

For diversity calculations, Shannon diversity (H) was calculated using the *stats.entropy* function from Python's SciPy library.

The formula is as follows:

$$H = -\sum_{k=1}^{s} p_i \ln p_i \tag{3.1}$$

Where p_i is the proportion of sequences in the clonotype i of s clonotypes.

3.3.2.2 Structural annotation

Sequences that IgBLAST identified as coding for a productive immunoglobulin were translated into amino acids and aligned to the IMGT antibody numbering scheme (Lefranc et al., 2003) using ANARCI (Dunbar & Deane, 2016). IMGT CDR sequences were extracted and structurally annotated using the SAAB+ version 1.01 pipeline (Kovaltsuk et al., 2020). As described in Section 1.2.8, SAAB+ uses SCALOP (Wong et al., 2019) to assign each sequence's CDRH1

and CDRH2 loops to a structural canonical class and uses FREAD (Choi & Deane, 2010) to identify if the CDRH3 loop has a similar structure to any crystallographically-solved set of 4,544 CDRH3 structures (referred to as templates, downloaded from SAbDab (Dunbar et al., 2014; Schneider et al., 2022) on 16th February 2022).

To reduce dimensionality, templates are clustered with a 0.6Å RMSD cut-off, producing a set of 1,944 templates. In this set of templates, 41% of the antibody structures are of murine-origin and 37% are of human-origin. The SAAB+ pipeline outputs for each sequence the canonical class of the CDRH1 and CDRH2 loops, and the Protein Data Bank (PDB) ID of the structure that contains the best matched CDRH3 structure for homology modelling from the 1,944 templates.

As a complementary approach, we also modelled representatives of all non-singleton clonotypes in each of the repertoires using a recent deep learning method, ABlooper (Abanades et al., 2022). ABlooper is competitive with AlphaFold2 on the canonical CDRs and shows better performance on the CDRH3 which was the target of our analysis. It is also more than 1000x faster than AlphaFold2 and was therefore suitable for our large-scale analysis.

For full structural modelling with ABlooper, a light chain must be supplied. As we did not know the cognate light chain for any of the heavy chains from bulk sequencing, all heavy chains were paired with a single light chain which was selected from the paired dataset. This light chain was the most commonly observed light chain in the Kymouse dataset. We selected a common light chain so that each heavy chain was modelled in the context of the same light chain: however, this could feasibly create a bias towards lower prediction accuracy for the heavy chains which are not observed with this light chain. An alternative would be to model everything with a rare light chain. ABlooper outputs a decoy diversity score for each CDR that it models. We used models which had CDRH3 decoy diversity scores below 3.2 (threshold from personal communi-

cations with Brennan Abanades-Kenyon).

3.3.2.3 Structural clustering

To cluster the ABlooper CDRH3 structures, each pair of modelled antibodies with CDRH3s of the same length were aligned via the Cα coordinates of their framework regions (IMGT residues 1-26,39-54,67-104,118-128). RMSD was calculated over the Cα coordinates of the CDRH3s given the framework alignment.

We clustered each set of antibodies of the same CDRH3 length via a greedy clustering algorithm as used in previous work (Raybould, Marks, Kovaltsuk, et al., 2020; Robinson et al., 2021). Clusters are formed by successively selecting the antibody sequence with the largest number of structural neighbours and removing the sequence and its neighbours from the pool of clusters until only singleton structures remain. We used a 0.6Å threshold in line with the threshold chosen by Kovaltsuk and colleagues in the SAAB+ application.

3.3.2.4 Structural comparison

To compare usage of either canonical forms or CDRH3 structural clusters (designated by either SAAB+ or RMSD clustering on ABlooper models), we followed the protocol from Kovaltsuk and colleagues whereby each repertoire is represented as a vector of frequencies of all observed clusters in the set of repertoires being compared. These vectors are Z-normalized on a per-cluster (versus per-repertoire) basis. A distance is calculated between each pair of repertoires (Euclidean distance) to produce a pairwise distance matrix in structural space over the set of clusters considered.

3.3.2.5 Humanness scoring of VH sequences

The human, Kymouse and C57BL/6 mouse VH sequences were assigned a "humanness score" using the random forest regressors from Hu-mAb (Marks et al., 2021). Sequences were first IMGT numbered using ANARCI as above. While

the C57BL/6 mouse and Kymouse sequences were mostly full length (IMGT positions 1-128), the human sequences were in most cases missing FWR1 (IMGT positions 1-26 in the IMGT CDR definition). As the human sequences were non-mutated, we considered it reasonable to simply fill in FWR1 according to the sequence found in the assigned germline. For the human and Kymouse sequences, the RF model trained on the IGHV gene assigned by IgBLAST were used for scoring, while for the murine sequences, all seven (IGHV1-7) RF models were used to score the sequence and the highest score was selected. We used the IGHV-specific classification thresholds reported in the Hu-mAb paper to annotate if a sequence was considered human or not (Marks et al., 2021).

3.4 Results

3.4.1 Kymouse and human repertoires are distinguishable on the basis of sequence properties

3.4.1.1 Germline gene usage

We used hierarchical clustering to compare the gene usage profiles of individual Kymice and humans, building phylogenetic trees to show the relationships between the individuals' gene usage profiles. Although Kymice and humans are the same in the numbers of IGHV genes used the frequencies as determined by sequence abundance can differ. The hierarchical clustering of the IGHV genes showed that the Kymice and humans form nearly separate monophyletic clusters except for a single outlier human subject. Most of the variation in IGHV gene usage is explained by the IGHV gene subgroup usage: clustering by this separates humans and Kymice without the outlier, with Kymice using a lower proportion of IGHV1 and IGHV2 genes compensated for by increased IGHV3, IGHV4 and IGHV6 usage (Figure 3.1A).

The IGHJ gene usage profile is similar: Kymice and nine out of the ten humans form monophyletic clades with one outlier, a different outlier human subject from the IGHV gene usage clustering. On average, the Kymouse uses IGHJ4 more frequently than humans, and IGHJ5 and IGHJ1 less frequently (Figure 3.2). Both the IGHV and IGHJ gene usage profiles of naive Kymouse reper-

toires are more similar to one another than to any human repertoire. However, Kymice are closer to some of the human repertoires than others. This therefore indicates Kymice fall within the range of human IGHV and IGHJ usage frequencies.

The hierarchical clustering of the IGHD genes revealed that the Kymice formed a monophyletic sister group distinct to that of the humans (Figure 3.2D). As can be seen from the heatmap, the IGHD germline genes used by the Kymice (e.g. IGHD3-22, IGHD2-15) are infrequently used by humans and vice versa.

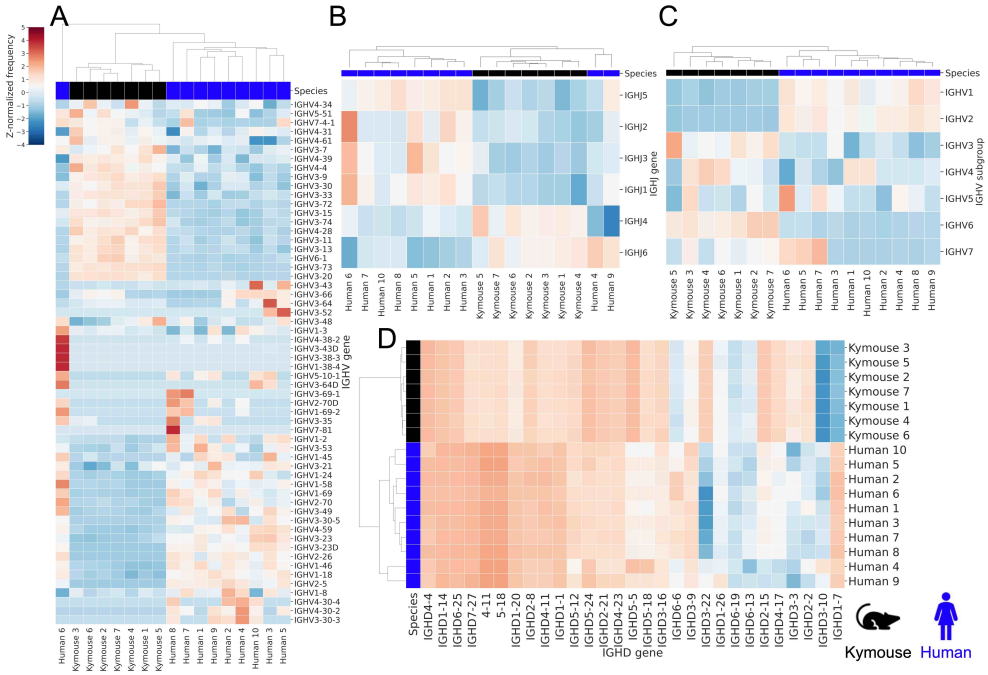

Figure 3.1: | **IGHV, IGHD and IGHJ usage in the Kymouse versus human repertoires.** Gene usage clustermaps for A) IGHV genes, B) IGHV subgroups, C) IGHJ genes and D) IGHD genes. The IGHV clustermaps show a separation between human (blue) and Kymouse (black) repertoires, with lower usage of IGHV1 and IGHV2 in the Kymouse. Downstream, there are also differences in usage of IGHJ genes with a preference in the Kymouse repertoires for IGHJ4. The IGHD gene usage shows the clearest distinction.

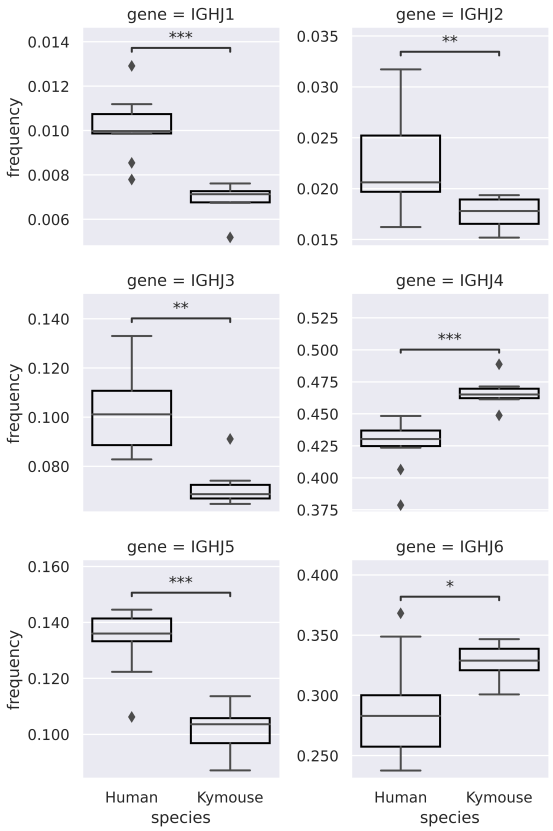

Figure 3.2: | **Comparison of IGHJ usage between naive human and Kymouse IgM BCR repertoires.** Kymouse repertoires disproportionately use IGHJ4 versus human repertoires, and use less IGHJ1 and IGHJ5. A Mann-Whitney U-test is performed for each of the six IGHJ genes: *, ** and *** indicate significance at the 5%, 1% and 0.1% levels respectively.

3.4.1.2 CDRH3 sequence analysis

We compared the distribution of the CDRH3 lengths in each species' naive repertoire (Figure 3.3). We calculated the length of the CDRH3 loop (under the IMGT antibody numbering scheme) from the Kymouse, human and C57BL/6 mouse heavy-chain sequences. This revealed that Kymice have an average CDRH3 length in between that of humans and mice, with a mean CDRH3 length of 14.25. In comparison, the C57BL/6 mouse dataset have a mean length of 12.39 amino acids, and the human dataset have a mean CDRH3 length of 16.56 amino acids. Kymouse CDRH3 loops are on average 2.36 aa shorter than humans (95% CI: 2.26, 2.48; $p<0.001$), while WT mice CDRH3 loops are on average 4.21 aa

shorter than humans (95% CI: 4.12, 4.30; p<0.001).

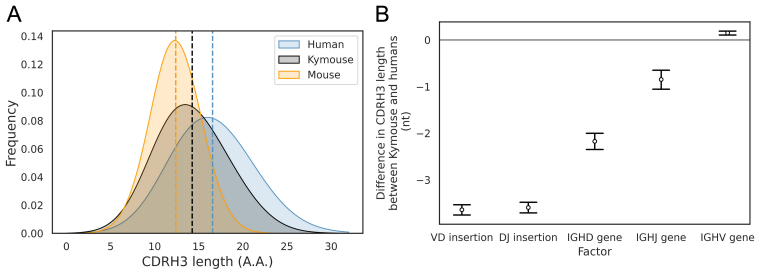

Figure 3.3: | **The CDRH3 length distribution in murine, Kymouse and human sequences.** The Kymouse CDRH3 length distribution is intermediate between the mouse and human distributions (A). Plus or minus indicates the standard error of the mean. We used estimation statistics to reveal that the major factor leading to this reduction in CDRH3 length despite access to the same germline repertoire in Kymouse is the relative lack of VD and DJ insertions (B).

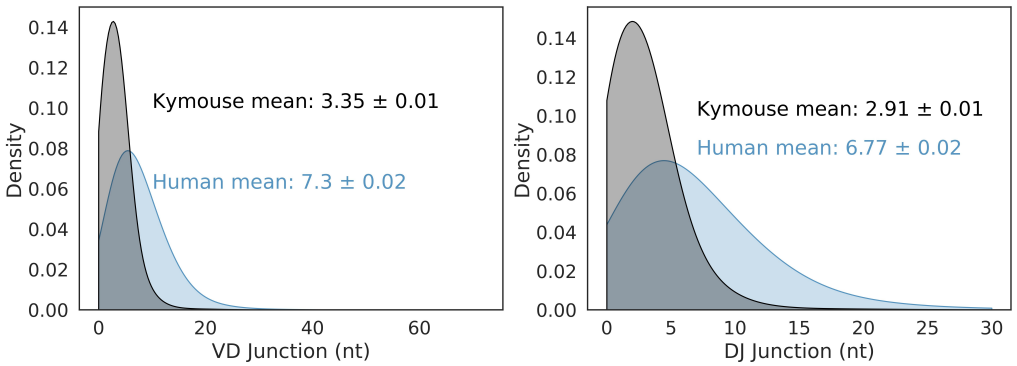

Figure 3.4: | **VD and DJ insertion length distributions in the Kymice and humans.** VD and DJ insertions in Kymice (grey) versus human (blue) repertoires. We considered only sequences where a satisfactory IGHD germline alignment was achieved.

The length of a CDRH3 loop is determined by four main factors: (i) the choice of IGHV gene; (ii) the choice of IGHD gene; (iii) the choice of IGHJ gene; (iv) the number of nucleotides inserted or deleted in the V-D and D-J junctional regions during B cell development. In order to investigate further why the Kymouse tends to have shorter CDRH3 loops on average than humans, we considered each of these factors in turn. We first looked at whether the differential IGHV gene usage by Kymouse had a significant impact on the length of the CDRH3 loop. For each heavy-chain sequence in the human and Kymouse datasets we determined the number of nucleotides that the IGHV germline gene

contributes to the CDRH3 loop. The results showed a statistically significant but small difference of 0.14 nt (95% CI: 0.12, 0.16; p<0.001) between humans and Kymice. Therefore, it does not appear that the differential choice of IGHV between humans and Kymice greatly affects the CDRH3 length. We next investigated the effect of the differential usage of IGHD genes between humans and Kymice on the length of the CDRH3 loop. This showed that the human IGHD germline genes used by the Kymouse are, on average, 2.34 nt shorter than humans (95% CI: 2.26, 2.41; p<0.001). We then compared the relative usage of the IGHJ genes of humans and Kymice (Figure 3.2). Kymice tended to use IGHJ4 (47.3% in Kymice versus 44.2% in humans) and IGHJ6 (32.8% in Kymice versus 28.6% in humans) while using the other genes (IGHJ1, IGHJ2, IGHJ3, IGHJ5) slightly less frequently, in particular IGHJ5 (11.1% in Kymice versus 15.1% in humans). Estimation statistics revealed that the differential IGHJ gene usage between Kymouse and human resulted in a decrease in the CDRH3 length for Kymouse of 0.85 nt (95% CI: 0.77, 0.94; p<0.001).

Finally, we looked at the number of nucleotide insertions in the V-D and D-J junctions in the antibody heavy chains. The results showed that Kymice V-D junctions are on average 3.73 nucleotides shorter than humans (95% CI: 3.68, 3.78 p<0.001), with a mean insertion size of 3.35 nucleotides compared to 7.30 nucleotides for humans (Figure 3.4A). Equally, the Kymice D-J junctions are on average 3.68 nucleotides shorter than humans (95% CI: 3.63, 3.73; p<0.001), with a mean insertion size of 2.91 nucleotides compared to 6.77 nucleotides for humans (Figure 3.4B). Overall, the number of junctional nucleotides inserted in the Kymouse heavy chain is 7.33 fewer than in humans (95% CI: 7.25, 7.40; p<0.001), with an average of 7.55 in Kymice compared to 14.88 in humans. These results show that the main factors that give rise to the shorter CDRH3 lengths in Kymouse compared to humans are the reduced number of nucleotide insertions in the V-D and D-J junctional regions, and the differential usage of IGHD germline genes between the species (Figure 3.3B).

A correlate of shorter CDRH3 length is a reduction in the theoretical CDRH3 diversity obtainable, with each amino acid less in the CDRH3 leading to a further 20-fold reduction in the number of possible CDRH3s. We do not expect the difference in diversity to approach this magnitude because of the CDRH3

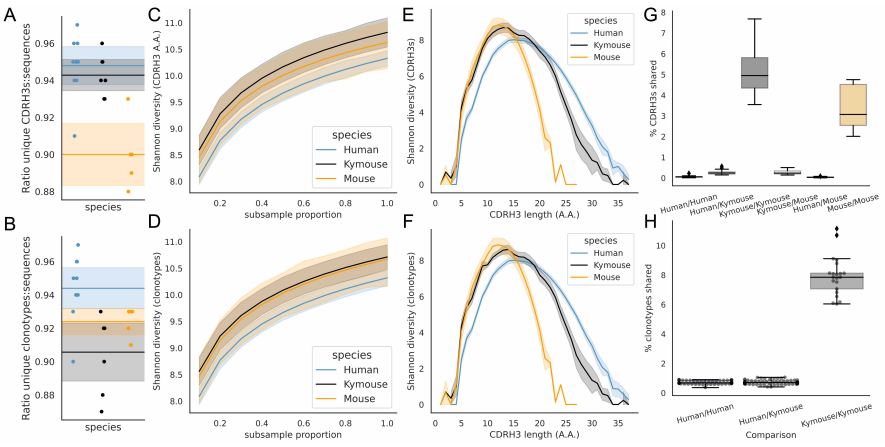

Figure 3.5: | **Examination of diversity of human, Kymouse and murine repertoires.** The top rows pertain to exact CDRH3 (amino acids) and the bottom rows to clonotypes (same IGHV, IGHJ and greater than 90% amino acid identity across length-matched CDRH3s). At the level of CDRH3s, the Kymouse repertoires have more unique CDRH3s per sequence sampled (A) are more diverse in their usages (C), despite their limited VD and DJ insertion rates. Diversity is reduced relative to human sequences at longer CDRH3 lengths which in unmutated repertoires depend on VD/DJ insertions (E). Kymouse repertoires show an opposite pattern in unique sampling rate and diversity when looking at clonotypes (B and D respectively) but still show reduced diversity versus humans at longer lengths. Overlap among CDRH3s (G) and clonotypes (H) between individuals is considerably higher in Kymice than humans, and more comparable to mice (G).

sequence patterns created by combinations of germline genes, however a reduction in the non-templated insertion of nucleotides will lead to less diverse repertoires in the Kymouse.

To examine the diversity of the repertoires, we considered how many unique CDRH3s or clonotypes can be found in a given repertoire, normalized by the total number of sequences observed. Figure 3.5A shows that the ratio of the number of unique CDRH3s to the number of sequences is comparable in Kymice and humans, and that both are considerably higher than in C57BL/6 mice. When clustering similar CDRH3s in combination with common IGHV and IGHJ genes (clonotypes), there are comparable unique clonotypes per sequence in mice, humans and the Kymouse (Figure 3.5B). We calculated the Shannon diversity of the CDRH3 and clone distributions, a measure which takes into account both richness and abundance and found comparable diversity across all repertoire

types (Figures 3.5C and 3.5D). The diversity observed depends on CDRH3 length considered, with the mouse and Kymouse repertoires having greater diversity at shorter lengths (peak diversity at lengths 12 and 13 respectively) and the human repertoires having greater diversity at lengths greater than 18 amino acids (peak diversity at length 15) for both CDRH3s (Figure 3.5E) and clonotypes (Figure 3.5F). Given that junctional insertions are required to reach lengths of 23 or greater in the human and Kymouse data, this supports the hypothesis that the reduced junctional diversification in naive Kymouse repertoires limits CDRH3 diversity at the longest lengths, however the greater diversity at shorter lengths may compensate. It is also clear that the Kymouse occupies a region in CDRH3 diversity and length between wild type mice and humans.

Finally, we considered the CDRH3 and clonotype overlap (Figures 3.5G and 3.5H). The overlap among CDRH3s is highest between individual mice and individual Kymice. On average 5.1% of CDRH3s were shared between any two Kymice and 3.4% of CDRH3s shared between any two mice compared to 0.12% of CDRH3s shared between any two human subjects. Despite the greater CDRH3 diversity reported for individual Kymice, over 25 times more CDRH3s are shared between pairs of Kymice on average than between humans. Clonotype sharing was also higher between individual Kymice (average 7.92% vs 0.84% in humans).

3.4.2 Kymouse, mouse and human repertoires can be compared via structural annotation and modelling

3.4.2.1 Homology modelling coverage inversely correlates with the number of junctional insertions

Human repertoires had significantly lower SAAB+ coverage than Kymouse or mouse repertoires (see Table 3.1). This has been reported previously (Kovaltsuk et al., 2020), though was ascribed to length differences.

In our analysis, we find that this difference appears when considering the same CDRH3 lengths range (between 5 and 16 amino acids) as in Table 3.1 and when stratifying by exact CDRH3 length (Figure 3.6A).

Another potential reason for this is that similar CDRH3s are not represented in SAbDab. The probability of a CDRH3 occurring multiple times is inversely correlated with the number of junctional insertions (Jaffe et al., 2022), so we hypothesised that the lower modellability is due to the significantly greater number of junctional insertions in human CDRH3s than murine or Kymouse CDRH3s.

We find that there is a significant difference in the total NP insertion length for each repertoire type between annotated and non-annotated CDRH3s (Mann-Whitney U-test: $p \ll 0.05$). This indicates that part of the lower modellability of human CDRH3s could be due to the higher number of junctional insertions than in mice or Kymice.

In the ABlooper analysis, just 61/43,378 CDRH3s could not be modelled (i.e. had a CDRH3 diversity score exceeding 3.25) (Table 3.2). 54 (88.5%) of these failures had NP insertion lengths greater than the species-specific median (Figure 3.6B). The coverage over this set of antibodies is significantly greater for ABlooper than for SAAB+, indicating that this new type of deep learning-based modelling is significantly less affected by non-templated junctional diversity than template-based methods.

Species	Total	Sequences with CDRH3 length between 5 and 16 amino acids	CDRH1 canonical class	CDRH2 canonical class	CDRH3 template
Human	338,677	175,907 (51.2%)	100.0%	99.0%	59.2%
Kymouse	412,493	292,946 (71.0%)	100.0%	95.1%	79.8%
Mouse	268,277	254,298 (94.7%)	99.5%	99.8%	92.0%

Table 3.1: | **Percentages of sequences which could be annotated via SAAB+.** Coverage over CDRH1 and CDRH2 exceeded 99% for all but the Kymouse CDRH2s. Humans have significantly worse SAAB+ coverage over the CDRH3 than Kymice or mice, even over the modellable range of 5 - 17 amino acids.

Species	Number of sequences	ABlooper coverage	SAAB+ coverage
Human	10,634	99.8%	59.8%
Kymouse	15,662	99.9%	87.3%
Mouse	17,057	99.9%	93.4%

Table 3.2: | **Numbers of sequences chosen to model using ABlooper.** Just 61 of 43,378 CDRH3s between the lengths of 5 and 17 amino acids could not be modelled successfully. SAAB+ coverage over the equivalent sequences is significantly lower.

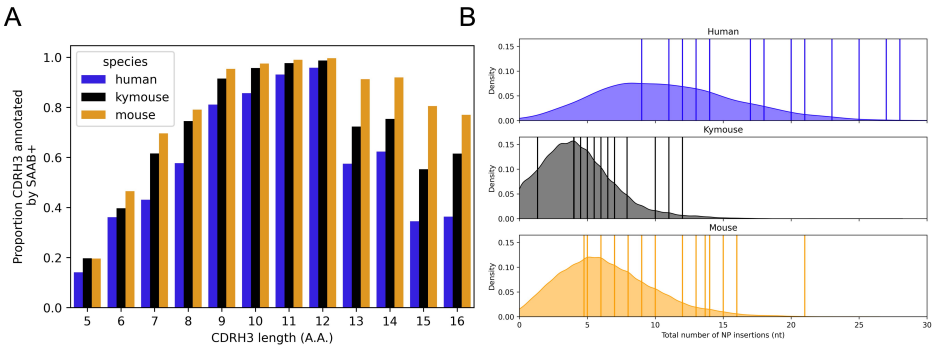

Figure 3.6: | **Coverage by SAAB+ and ABlooper as a function of CDRH3 length.** A shows the proportion of sequences for which a CDRH3 structural cluster could be assigned via SAAB+ as a function of CDRH3 length and species. Human repertoires are the least modellable at all lengths, which we hypothesise is due to the larger number of NP insertions. B shows where the failed ABlooper CDRH3 models fall in the distribution of NP insertion lengths for all models of a particular repertoire type.

3.4.2.2 The Kymouse repertoires are structurally more human-like than mouse-like

We compared the structural similarities of the BCRs via structural annotation of the CDRs. We first compared the canonical class distribution (CDRH1 and CDRH2). For both CDRH1 and CDRH2 the Kymouse repertoires group separately from human repertoires, but group with human repertoires before the mouse repertoires, in their usage of canonical forms (Figure 3.7). All canonical forms are strongly predicted by IGHV germline gene subgroup especially as these sequences are non-mutated. Despite the effects of the differential IGHV usage observed in humans and Kymice, this does not make the Kymouse canonical class usage more similar to mice as the murine repertoires use different canonical classes than either the human or humanised repertoires. Interest-

ingly, six of the nine murine IGHV germlines encode a single canonical form, H1-8-A, suggesting more limited structural diversity in murine CDRH1s.

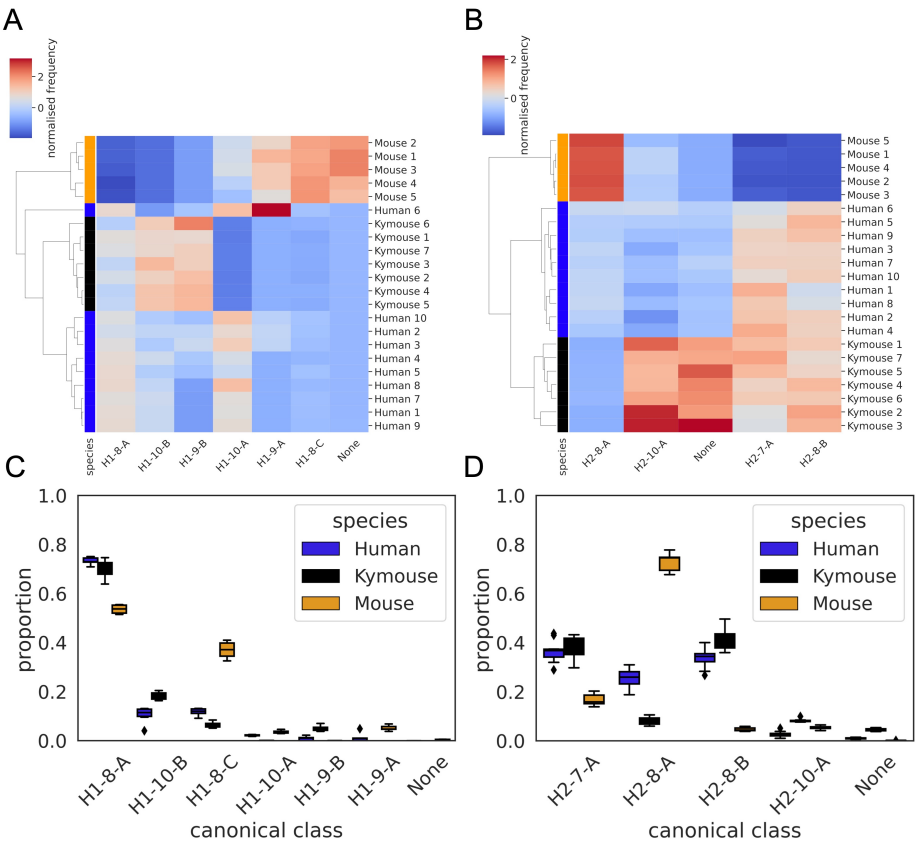

Figure 3.7: | **Canonical class usage in Kymouse, mouse and human repertoires.** CDRH1 (A and C) and CDRH2 (B and D) canonical class usage clusters Kymice and humans separately from mice. The key differences between humans and Kymice and mice in CDRH1 canonical forms are the greater usage of H1-9-A and H1-8-C in mice verus H1-10-B and H1-9-B in humans and Kymice. Between humans and Kymice, the Kymouse repertoires have lower usage of H1-10-A and greater usage of H1-10-B than humans. Focussing on the CDRH2 canonical forms, the largest difference between mice and humans/Kymice is in the greater usage of H2-8-A and lower usage of H2-8-B and H2-7-A. The Kymouse repertoires use significantly less H2-8-A than the human repertoires.

We then performed structural comparisons of the shapes of the human, Kymouse and mouse CDRH3s. CDRH3s do not adopt canonical conformations, so we used two different approaches: firstly, structural annotation which consists of comparison to a CDRH3 structural database and annotation with a structural

cluster ID, and secondly full CDRH3 modelling.

Of the 1,944 possible structural clusters in the CDRH3 structural database, 1,594 were observed in at least one repertoire. The majority (1,270) of these clusters were observed in all species. There was an observable difference among repertoires in the species origin of the structural clusters observed, i.e. each species was biased in the structural space it tended to use. The majority of structural clusters used in human and Kymouse repertoires were of human origin (57.5 and 55.0% respectively), while the majority of structural clusters used in mouse repertoires were of murine origin (64.0%) (Figure 3.8). The CDRH3 structural database was 41% of murine origin, 37% of human origin.

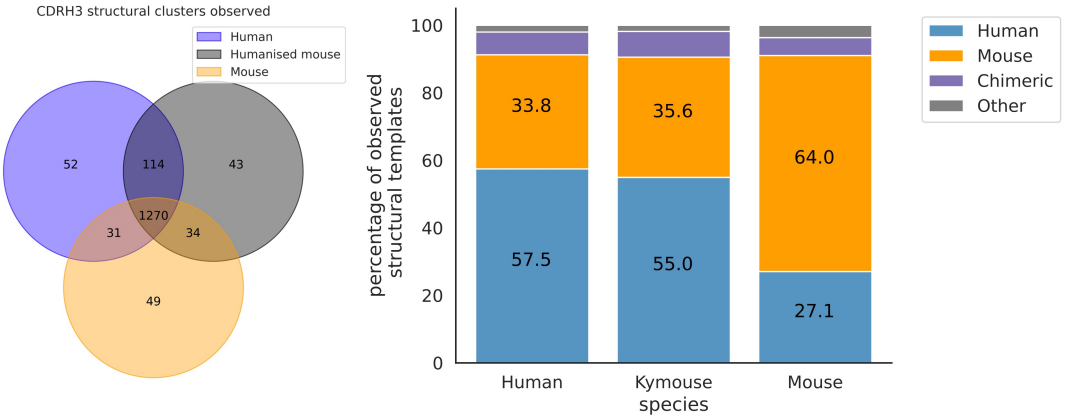

Figure 3.8: | **CDRH3 structural clusters observed by repertoire type.** The majority of CDRH3 structural clusters are observed across all three repertoire types. Focussing on the origin of the antibody representing the structural cluster, 57.5% of templates used in the human repertoires and 55% in the Kymouse repertoires are of human origin. The majority of templates observed in the mouse repertoires are of murine-origin.

As described, the CDRH3 lengths in the human dataset were on average 2.36 aa longer than those in the Kymouse, and 4.21 aa longer than those in the C57BL/6 mice. As we considered only CDRH3s between length 4 and 16 in this structural analysis, this difference was reduced to a difference of 0.45 amino acids between human and Kymouse repertoires (CI 0.43, 0.47) and 0.63 between human and mouse repertoires (CI 0.62, 0.65). The average CDRH3

lengths of the human templates in the FREAD database were longer than those in the murine templates (12.6 aa compared to 11.2 aa respectively). We checked that the observed differences in structural template usage by humans, Kymice, and C57BL/6 mice were not just reflecting differences in the availability of templates at the different CDRH3 lengths. Therefore, we stratified the FREAD database by CDRH3 length and ran the datasets against each CDRH3 length separately. If the preferential usage of human-specific templates by humans and humanized mice, and vice-versa for mice, is simply a result of random sampling from the FREAD database, we expect the ratio of the proportion of human templates in human repertoires to the proportion of human templates in the FREAD database to be approximately 1 across all CDRH3 lengths, and below 1 across all CDRH3 lengths for mice. Instead, we see a significant enrichment (at the 1% level) of species-specific templates at multiple CDRH3 lengths.

The human repertoires are more structurally variable than are the humanized murine or murine repertoires (Figures 3.9A, B and C). The murine repertoires are the least variable (there is the smallest range of distances between any two given subjects for mice, and on average the smallest average distance between pairs of subjects). All the Kymouse repertoires were structurally closer (Euclidean distance in Z-normalized CDRH3 structural cluster usage) to any given human repertoire than to any C57BL/6 mouse repertoire (Figures 3.9A, B and C). With no correction for sample size or for CDRH3 length distribution, the humanized murine and human repertoires form a monophyletic cluster that is sister group to the murine repertoires (Figure 3.9D).

Labelling cluster members as "murine" or "human-like", the adjusted Rand index of this clustering is 1.0 (perfect correspondence). We calculated Rand indices for 100 subsamples with equal numbers of each CDRH3 length to control for any length effect; in all subsamples, the human and humanized repertoires could be clustered separately from the murine repertoires resulting in an adjusted Rand index of 1.0. Further, in all 10 subsamples, at least one human repertoire was more structurally similar to a given humanized murine repertoire than to at least one other human repertoire (between 5 and 9 of 10 subjects

per subsample).

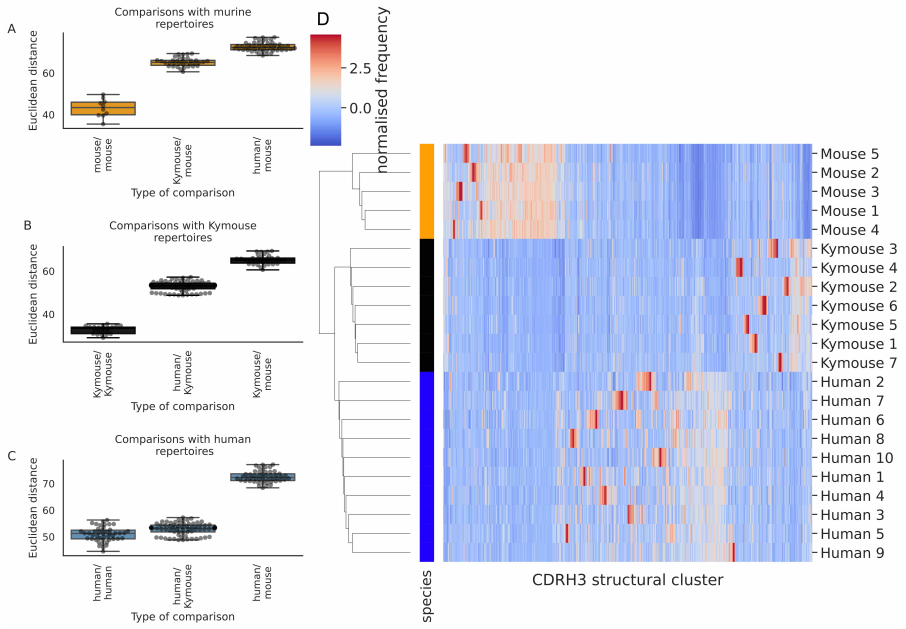

Figure 3.9: | **Hierarchical clustering on SAAB+ CDRH3 structural usage.** Distance between CDRH3 structural cluster usage profiles is measured with Euclidean distance in Z-normalized proportions. This is calculated pairwise between subjects and these distances are clustered hierarchically. Figures A through C show these pairwise distances stratified by the type of comparison. The leftmost box shows the range in distance between individuals of the same species. For the mouse and Kymouse repertoires, this range is smaller than the range in distances for any other species, meaning that they cluster monophyletically (D). Human repertoires have less self-similar CDRH3 structural cluster usages with ranges overlapping with Kymouse repertoires. In the hierarchical clustering solution with these pairwise distances that is shown in D, the human and Kymouse repertoires form a monophyletic clade separately from the murine repertoires.

In conclusion, even when adjusting for sample size and differing CDRH3 length distribution, the repertoires of Kymouse are structurally more similar to human repertoires than they are to murine repertoires according to the homology-based structural annotation technique.

The species bias we observed in structural cluster usage, which is assigned via sequence similarity to a structural template, meant that a template-free mod-

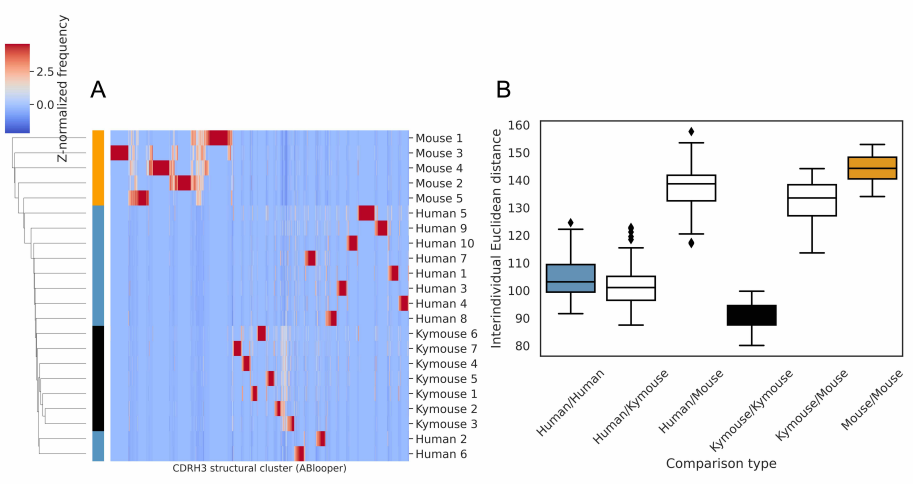

Figure 3.10: | **Hierarchical clustering of ABlooper CDRH3 structural cluster usage.** A shows the clustermap of Z-normalized usages of CDRH3 structural clusters which were derived from greedy clustering with a 0.6Å cut-off of CDRH3 Cα RMSDs calculated between models built by ABlooper. The humans and Kymice form a monophyletic cluster, which the C57BL/6 mouse repertoires do not. The distribution of distances shown in figure B reveals that the Kymouse repertoires are the least variable. Similarly to the SAAB+ CDRH3 structural cluster usage comparison, the ranges of human/human and human/Kymouse intersubject distances are overlapping however the extent of overlap is greater than observed with SAAB+. There is also overlap between the distances observed between humans and Kymice with mice, and between individual mice. The murine distances are greater than that observed with the SAAB+ clustering because of the greater diversity sampled at shorter lengths in this limited dataset.

elling procedure might lead to different results. We modelled non-singleton clonotype representatives from humans, Kymice and mice using ABlooper, then compared the resultant CDRH3s via Cα RMSD. Analogously to clustering of the CDRH3 template database, we performed greedy clustering with a 0.6Å cut-off. This clustered the 43,378 models with 41,397 unique CDRH3s into 6,546 clusters. The modelled human CDRH3s were on average longer than Kymouse CDRH3s by 0.99 amino acids (CI: 0.94, 1.05) and longer than mouse CDRH3s by 1.23 amino acids (CI: 1.17, 1.29). We observed a slightly different clustering by usage than via the homology approach (Figure 3.4.2.2A), with monophyly of Kymice, and Kymice and humans, but not of mice. The distribution of intersubject differences was such that the human/Kymouse and human/human comparison distributions are largely overlapping. Indeed, the differences be-

tween all distributions were less extreme with overlap between most of the distributions (Figure 3.4.2.2B). Mice exhbit the greatest variance here: this is likely partly due to our sampling, which resulted in a greater number of unique CDRH3s modelled in this subsection of the analysis. The monophyly of Kymice and humans versus mice is observed when subsampling in order to equalize the CDRH3 length distribution, i.e., it is not driven by differences in CDRH3 length. The deep learning-based modelling approach supports the earlier finding that Kymice are more structurally similar to humans than mice, and vice-versa; the extent of this similarity is greater than observed with the homology modelling approach. It must be noted that the sample size for this comparison is significantly lower (total 43,378 ABlooper models vs. 571,200 structurally annotated CDRH3s).

In conclusion, the sequence differences in the repertoires which were described in the germline and diversity sections impact upon the repertoire of CDR structures which are observed in the Kymouse. The distribution of canonical forms in CDRH1 and CDRH2 is distinguishable, but usage is human-like. In the CDRH3, both structural annotation and full structure prediction indicate that the naive Kymouse CDRH3 structural space is human-like, indicating that Kymice repertoires offer comparable structural starting points for the production of antigen-specific antibodies.

3.4.2.3 Comparison of structural distance via SAAB+ and ABlooper

As described in Section 3.4.2.2, the clustering produced by comparing SAAB+ and ABlooper cluster usages are slightly different; the ranges of inter-subject distances show considerably more overlap, and the murine repertoires appear the most variable versus the human repertoires in the SAAB+ analysis.

This could be ascribable to i) the lower sample size in the ABlooper analysis or ii) overestimation of the structural distance between CDRH3s. Rerunning the analysis using just the 43,378 sequences which were included in the

ABlooper analysis, we indeed see a different clustering than that observed when clustering the full dataset and a greater degree of overlap in interindividual differences (Figure 3.11). However, there is significantly less overlap between the different categories of subjects than observed when using ABlooper structural clusters. This supports the hypothesis that SAAB+ may be overestimating the structural distance between CDRH3s.

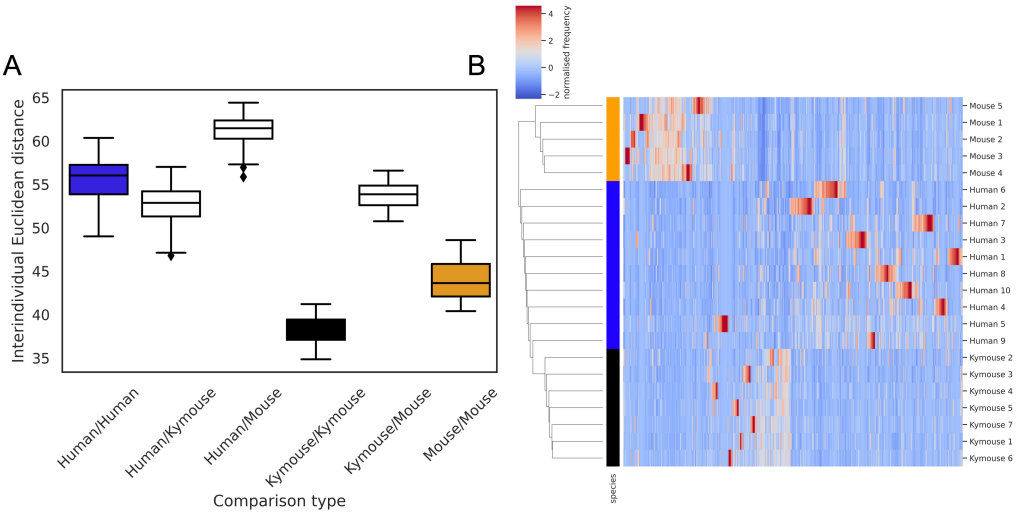

Figure 3.11: | **Hierarchical clustering on SAAB+ CDRH3 structural cluster usage in the subset modelled by ABlooper.** Interindividual distances were recalculated using the SAAB+ annotations of the reduced set of sequences used in the ABlooper analysis, in order to ascertain whether the difference between the ABlooper and SAAB+ results can be ascribed to a reduction in sample size. While there is more overlap between repertoire types than observed with the full dataset, there is less overlap than observed using ABlooper structural clusters.

To this end, we compared all pairs of CDRH3s which were modelled and which were assigned a SAAB+ template, resulting in 1.2 x 10^8 comparisons. Pairs were considered as positives if the ABlooper RMSD was less than or equal to 0.6Å or if the SAAB+ templates were the same (as this is the basis of the clustering). We found that a significantly greater proportion of shorter CDRH3s (lengths 5- 7) were structurally clustered using ABlooper than SAAB+, while SAAB+ clustered a greater proportion of CDRH3s at lengths above 13 amino acids (Figure 3.12A). SAAB+ has access to relatively few templates above length

13, increasing the probability that any two annotated CDRH3s will share the same annotation. ABlooper was more likely than SAAB+ to cluster two CDRH3s from the same species (Figure 3.12B). We also found that there was relatively little overlap in terms of which pairs of CDRH3s were predicted to be structurally similar by both methods (Figure 3.12C), indicating that the results may simply differ due to differing predictions. Despite these differing results, the RMSD between the template structural clusters and the ABlooper models was reasonable (Figure 3.13).

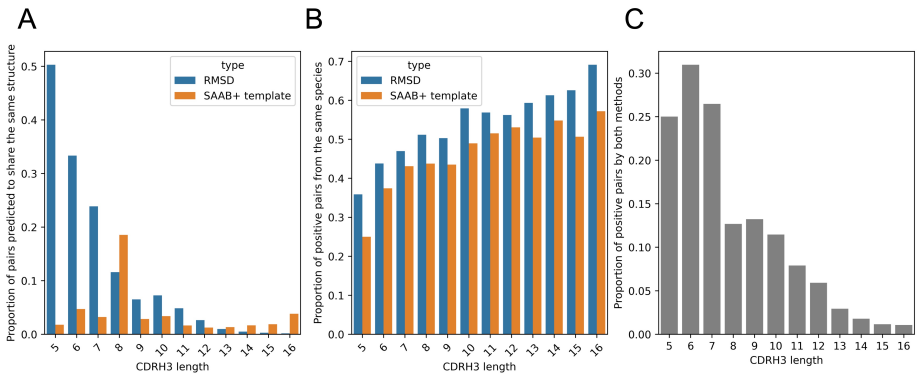

Figure 3.12: | **Comparison of ABlooper and SAAB+ predictions.** All pairs of CDRH3s of the same length modelled by ABlooper and successfully annotated with SAAB+ were compared. A positive pair was a pair of CDRH3s with less than 0.6Å between the ABlooper models or two CDRH3s with the same SAAB+ cluster. A greater proportion of pairs of shorter CDRH3s are structurally similar via ABlooper than via SAAB (A). ABlooper groups proportionally more CDRH3s from the same species at each length, suggesting that homology modelling does not create an artificial structural distance between repertoires (B). Only the minority of predictions are shared between the two methods at all lengths (C).

3.4.3 A state-of-the art humanization tool scores Kymouse sequences as fully human

We next tested whether the Kymouse sequences are considered human by a state-of-the-art humanization tool. We used the random forest classifiers within Hu-mAb to score heavy chain amino acid sequence humanness. 100% of human and Kymouse heavy chain sequences were classified as human, with the

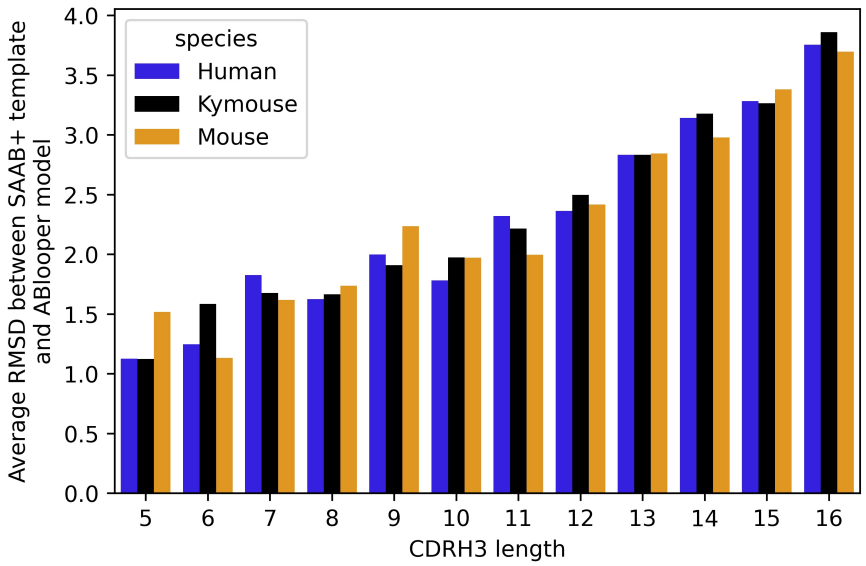

Figure 3.13: | **RMSD between ABlooper models and SAAB+ CDRH3 structural cluster centres.** For each sequence modelled via ABlooper, RMSD was calculated over the CDRH3 of the structural cluster assigned via SAAB+ and the modelled CDRH3. An average was calculated for each CDRH3 length bin.

maximum humanness score assigned to 99.1% of human sequences and 98.3% of Kymouse sequences. All sequences produced scores in the "Positive (High Score)" category which had minimal anti-drug antibody events reported among therapeutic antibodies (Marks et al., 2021). No murine sequences were classified as human.

3.5 Discussion

Sequence analysis of immunoglobulin heavy chains showed that naive Kymouse BCRs have a CDRH3 length distribution that is intermediate between human and mouse repertoires (on average 14 aa versus 16 aa for human repertoires and 12 aa for mice). We compared the IGHV, IGHJ and IGHD gene usage profiles of human and Kymouse repertoires and showed that the IGHV and IGHJ gene usage profiles in Kymice are distinct from human profiles but human-like, while the IGHD gene usage profiles are sufficiently different that Kymouse and human repertoires are distinct. This contrasts with the previous NGS character-

ization of the OmniRat in which both IGHV and IGHD gene usage was distinct from humans (Joyce et al., 2020).

It is an ongoing concern that different germline gene distributions may affect how representative transgenic models are of immune responses to germline-targeting immunogens. In this work, the frequency of certain germline genes which have been identified as targets for priming for the development of broadly neutralizing antibodies, such as IGHV1-2 and IGHV1-46 (for priming of VRC01 class antibodies against HIV (Zhou et al., 2010; Zhou et al., 2015)) and IGHV1-69 (which is enriched in broadly neutralizing antibodies to the HCV E1/E2 antigen, influenza hemagglutinin head and stem, and HIV-1 gp120 and gp41 (Chen et al., 2019)) are lower in Kymouse repertoires than in human repertoires. Despite this, Kymice have been shown to be able to produce VRC01-class antibodies in response to a VRC01-class bnAb germline-targeting immunogen (Sok et al., 2016). Perhaps of more significance to the use of Kymice for response modelling is that they all have the same Ig loci and therefore provide no insight into the robustness of vaccine response to the considerable genetic diversity in human populations. This is of particular relevance for IGHV1-69 which is the most polymorphic gene in the IGH locus with considerable repercussions for disease-specific antibody repertoires (Avnir et al., 2016; Watson et al., 2017).

After establishing key differences in the germline gene usage profiles, we examined the extent to which these differing distributions contribute to the shorter CDRH3s observed in Kymice. The Kymouse repertoires display a preference for shorter IGHD genes: this was also observed in the NGS characterization of OmniRat BCR repertoires suggesting that a preference for shorter IGHD genes may be common across transgenic rodent platforms (Joyce et al., 2020). While the differences in germline gene usage distributions do appear to contribute to the differing CDRH3 length distributions, a greater proportion of the effect is ascribable to fewer nucleotide insertions at both the V(D) and (D)J junctions during junctional diversification in the Kymouse with on average 7.33 nt fewer inserted in Kymouse over the two junctions. This reduced junctional diversification in the Kymouse leads to lower diversity in longer CDRH3s and

greater clonotype overlap between individual Kymice than between individual humans.

There are four biological caveats for the comparison of the BCR repertoires between mice, Kymice and humans. Firstly, the average age of the humans in the study was 43.1 years, while the mice and Kymice were sacrificed at an age of 8-10 weeks which is considerably younger in proportion to average lifespan. The effect of ageing on the B cell repertoire has been extensively studied (Dunn-Walters & Ademokun, 2010): it has been shown that human neonatal CDRH3s are shorter in length due to selective usage of IGHD7-27 and reduced TdT activity (Souto-Carneiro et al., 2005; Wang et al., 2014) and shorter CDRH3s in neonatal murine repertoires have also been reported (Feeney, 1992; Ivanov et al., 2005; Schroeder, 2006). Ghraichy and colleagues also reported changes in IGHV and IGHJ usage with age, however they reported downregulation of IGHV1 genes which are in any case expressed at higher levels in the human naive repertoires than the Kymouse repertoires (Ghraichy et al., 2020). As a second biological caveat, we sampled the naive repertoire from peripheral blood in humans vs. splenic B cells in mice and Kymice. This increases the probability of sampling IgM memory cells, which increase in frequency in age (Dunn-Walters & Ademokun, 2010). It has also been reported that CDRH3 lengths are shorter in splenic B cells than PBMCs in humans (Pickman et al., 2013) and that there is greater clonal overlap among individual mice in splenic B cells than PBMCs (Wu et al., 2022). Thirdly, the Kymice and C57BL/6 mice are inbred meaning that we would expect less variation among individuals in their germline gene usage due to the considerable structural diversity in the human Ig locus that continues to be uncovered (Gidoni et al., 2019). There is the final obvious caveat in terms of the degree of antigen exposure of adult humans versus mice kept in sterile conditions, however the restriction of the analysis to non-mutated, naive B cells minimises the importance of this consideration.

Changes in CDRH3 structural cluster usage have been previously observed along the B cell development axis (Kovaltsuk et al., 2020) and allow comparison of repertoires derived from different species. Despite these genetic differences,

modelled structural comparison of the human and humanized repertoires to the murine repertoires, by annotating the VH sequences with predicted CDRH3 structural template clusters showed murine repertoires use mostly CDRH3 structural clusters that have been identified from murine antibodies, while human and Kymouse repertoires use CDRH3 structural clusters identified from a more even distribution of species of which more than 50% were identified from human antibodies. Further, grouping of the exact distribution of CDRH3 structural clusters reveals that Kymouse structural cluster usage is closer to human usage than murine usage, accounting for CDRH3 length differences. When modelling the CDRH3s (as opposed to performing an approximation via structural annotation), part of the "structural distance" between the human, mouse and Kymouse repertoires disappeared suggesting part of the signal observed in homology modelling is due to different sequences with similar predicted shapes. This suggests that despite the observed differences at the sequence level, the CDRH3 structural shapes adopted by the BCRs are within the distribution of observed human shapes. Finally, using the Hu-mAb humanness classifiers, all Kymouse and human sequences are classified as human, meaning that naive sequences isolated from the Kymouse are predicted to have similar immunogenicity in humans to sequences isolated from humans themselves.

There are two significant limitations of the structural analysis. The first limitation is that predictions above length 16 are considerably less accurate and are therefore excluded from the analysis. For the human repertoires, this excludes around 50% of the data, while just 5% of murine CDRH3s were longer than length 16. The second limitation is the finding here that coverage with the homology-based structural annotation tool SAAB+ appears to decline with the number of non-palindromic/palindromic nucleotides inserted at the VD and DJ junctions. This is logical given that these CDRH3s are less likely to be observed publicly (Jaffe et al., 2022) and therefore less likely to have been previously studied. While most of ABlooper's failures (models where the decoy diversity exceeded 3.2) had higher than average NP insertion lengths, this did not lead to significantly lower coverage in human repertoires than mouse or Kymouse repertoires, and therefore this limitation is less severe for deep learning based

modelling than structural annotation. The net result of these limitations is that the majority of human CDRH3s are discarded for the structural analysis while the majority of murine CDRH3s are retained, which may create an artificial relative diversity of murine CDRH3 structures. Our analysis is robust to the number of sequences at each length due to the subsampling we performed, however we can only extend our conclusions to a subset of the primary BCR repertoire (though this may counteract some of the previously described CDRH3 length bias introduced by comparing peripheral to splenic B cells).

In conclusion, although naive BCR repertoires of the Kymouse have key distinctions from human repertoires at the sequence level they are more human-like than mouse-like in terms of CDRH3 structural usage. A number of studies have shown how the Kymouse is able to elicit equivalent antibodies to those found in humans exposed to the same antigen (McLeod et al., 2019; Oyen et al., 2020; Scally et al., 2017; Sok et al., 2016). This supports the hypothesis that the structural templates available for antigen binding are indeed human-like as we show here.

3.6 Chapter summary

In this chapter, we performed a combined sequence and structural analysis of naive BCR repertoires of humans, mice and Kymice.

Our sequence analysis revealed key differences in the primary repertoires of humans and Kymice that should be kept in mind when comparing their immune responses, as they have different IGHV, IGHD and IGHJ gene preferences that might bias their antigen-experienced BCR repertoires in different directions. Kymice also have shorter CDRH3s which is mainly attributable to a lack of VD and VJ insertions, which will limit the frequency of longer CDRH3s.

We then performed homology-based structural annotation with SAAB+ and

concluded that the predicted distribution of CDRH3 shapes found in Kymice is more human-like than mouse-like, and indeed falls within the range of observed CDRH3 structural cluster usage.

Modelling of a subset of the BCR repertoires with ABlooper and structural clustering of these models provided further evidence that Kymouse CDRH3s are structurally more human-like than mouse-like. Comparison of SAAB+ and ABlooper's predictions showed that that the degree of structural convergence found between subjects depends on the method used with greater overlap observed between ABlooper structural cluster usage. We also found that ABlooper modelling does not appear to be as affected by NP insertions at the V(D) and (D)J junctions as does homology modelling, suggesting that it will represent a major advance in our ability to structurally characterise human repertoires.

4 | Paratyping and clonotyping applied to BCR repertoires following Ebolavirus vaccination

4.1 Abstract

In Chapter 2, I presented a novel method for immune repertoire mining "paratyping" that was able to identify antibodies that bind to the same antigen from different clonotypes. We observed that the precision of the method was highest combined with clonotyping.

In this chapter, we use this combination of clonotyping and paratyping to annotate the predicted antigen-specific component of BCR repertoires responding to vaccination with an Ebolavirus vaccine candidate, Ad26.EBOV/MVA-BN-Filo. We first curate and analyse the largest known database of Ebolavirus-binding antibody sequences, called EBOV-AbDab, identifying the sequences of nearly 1,000 anti-EBOV antibodies (in comparison to just 34 found in the largest epitope database, the IEDB). These known binders are then used to annotate the predicted EBOV-specific component of the BCR repertoire response to Ebolavirus vaccination. We find that our predicted "hits" are disproportionately found in expanded and convergent clonotypes; we note that one group of antibodies is particularly convergent, and found in most vaccinees. This work further validates the utility of computational immune repertoire mining.

4.2 Introduction

Ebolaviruses are highly infectious filoviruses which can cause severe disease (Ebolavirus disease, referred to as EVD) with mortality rates of up to 90% (Van Kerkhove et al., 2015). There are six species currently within the Ebolavirus genus: Zaire (Ebola virus; ZEBOV), Sudan (SUDV), Bundibugyo (BDBV), Tai

forest (TFV), Reston (RESTV) and the most recently described Bombali ebolavirus (BOMV). All but Reston and Bombali virus have been associated with severe disease in humans (Goldstein et al., 2020; Martell et al., 2019). Only three species (ZEBOV, SUDV, BDBV) have caused outbreaks, with ZEBOV and SUDV in particular responsible for tens of thousands of deaths in over thirty separate outbreaks in West and equatorial Africa since 1976 (Malvy et al., 2019; Misasi & Sullivan, 2021).

The 2013-16 Ebola virus outbreak in the DRC was the largest of these outbreaks, causing in excess of 28,000 cases and 11,000 deaths (Garske et al., 2017). This epidemic hastened the process of human safety and efficacy testing of Ebola vaccine candidates (Osterholm, 2016; Whitty, 2017) and the first Ebola virus vaccine was approved for use in 2019. The rVSV-based vaccine (Ervebo) is currently used to immunise at-risk individuals during active outbreaks but these continue to occur with regularity: there have been three distinct Ebolavirus outbreaks in the DRC between May 2018 and November 2020, an outbreak in Guinea in 2021 and again in the DRC between April and June of 2022. As of the time of writing, there is an outbreak of closely related filovirus Marburgvirus in Ghana. The continued outbreaks of Ebolavirus in West and Central Africa demonstrate that vaccination is not currently controlling the spread of EBOV and related species (Woolsey & Geisbert, 2021).

The majority of the B cell response to Ebolaviruses is directed at the surface glycoprotein (GP) which is the only antigen on the virion surface. GP is a trimer of GP1-GP2 heterodimers (Lee et al., 2008); GP1 can be divided into the core, glycan cap and a mucin-like domain (MLD). GP2 can be divided into an internal fusion loop (IFL), two sets of heptad repeats (HR1 and HR2) and the membrane-proximal external region (MPER) (Figure 4.1). A key epitope among neutralizing antibodies is the receptor-binding region (RBR) which is where the glycoprotein interacts with Niemann-Pick C1 (NPC1); prior to cleavage by cathepsins in the late endosome, the RBR is masked by the glycan cap and MLD. Antibodies to the glycan cap and MLD are rarely neutralising (Saphire, 2019). In addition to this membrane-bound GP, there is a secreted

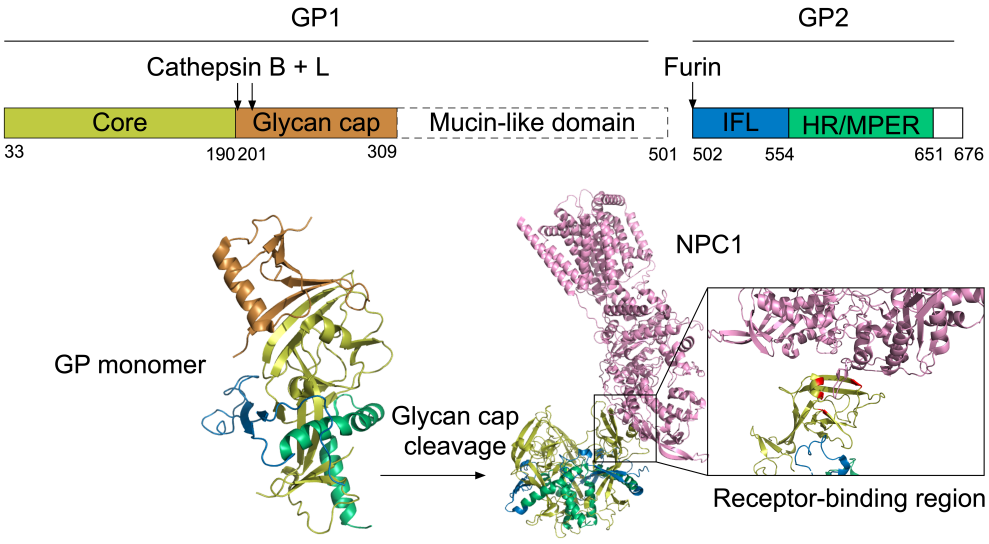

Figure 4.1: | **Overview of the domains of EBOV GP and representation of its structure.** The Ebolavirus glycoprotein (GP) is a homotrimer. The GP monomer consists of GP1 and GP2 which are produced from the precursor GP through furin cleavage. The larger subunit, GP1, consists of a core which is involved in receptor binding, and the heavily glycosylated glycan cap and mucin-like domain. The glycan cap and mucin-like domain are cleaved by cathepsins B and L in the late endosome. This exposes the receptor-binding region (RBR) which binds to the Niemann-Pick C1 (NPC1) protein, after which the viral and host cell membranes fuse. GP2 consists of an internal fusion loop, which is crucial for membrane fusion, heptad-repeat regions HR1 and HR2 and the membrane-proximal external region (MPER). Figure adapted from (Misasi & Sullivan, 2021).

GP called soluble GP (sGP) which is a heterodimer of GP1 that lacks the MLD. While it has been proposed that sGP acts as an "antibody sink" (Misasi & Sullivan, 2021), GP/sGP cross-reactive mAbs can still be protective (Saphire, 2019). Antibodies and nanobodies have been elicited to four other EBOV proteins (of at least ten proteins which the Ebolavirus genome is predicted to encode (Misasi & Sullivan, 2021)). These include antibodies to the polymerase co-factor VP35, matrix protein VP40, transcription factor VP30 and the nucleoprotein (NP). Currently, all immunotherapies directly target the virion GP (O'Donnell & Marzi, 2021).

The EBOV glycoprotein exhibits considerable diversity across Ebolavirus

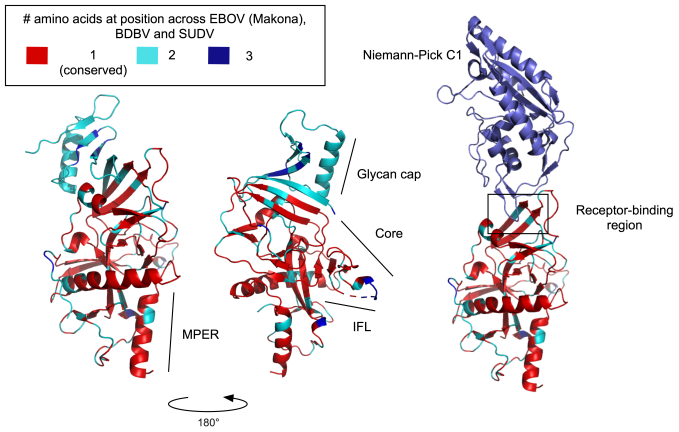

Figure 4.2: | **EBOV GP colored by amino acid conservation across Ebolavirus species.** GP1 (lacking the MLD) and GP2 colored by amino acid conservation between EBOV Makona 2014 (KJ660346.2), BDBV 2007 (accession NC_014373.1) and SUDV Gulu 2000 (accession NC_006432.1), as single heterodimers and in complex with NPC1 in order to highlight the receptor-binding region. The majority of the core is conserved. The GP structure corresponds to PDB ID 7kf9 (Murin et al., 2021); GP in complex with NPC1 is from PDB ID 5f1p (Wang et al., 2016).

species, with just 55% amino acid sequence identity between the EBOV Makona (2014) and SUDV glycoproteins (Rijal et al., 2019). The glycan cap and MLD are the least conserved across Ebolavirus species: amino acid conservation across a GP1/GP2 dimer with deleted MLD are shown in Figure 4.2. This poses a significant challenge to vaccination and immunotherapy, both of which are entirely GP-targeted.

Immunotherapeutics occupy much of the Ebolavirus therapeutic space (O'Donnell & Marzi, 2021). The first potently neutralizing antibodies such as KZ52 offered no protection against EBOV challenge in NHPs (Wilson et al., 2000); the first pre-clinical proof of protection via immunotherapy came from transfer of IgG isolated from hyperimmune macaques in 2009 (Dye et al., 2009). Subsequently, a number of monoclonal antibody cocktails have shown high protective efficacy in non-human primates, with diversity and potency of effector functions being key in protective efficacy. The details of these cocktails are shown in Table 4.1.

Name	Components	Source	Stage	PALM 2018
ZMab	1H3, 2G4, 4G7	Mouse	NA	NA
MB-003	13C6, 6D8, 13F6	Mouse (chimerized)	NA	NA
ZMapp	2G4, 4G7, 13C6	Mouse (chimerized)	FDA-approved (October 2014)	Standard of care
mAb114 (Ansuvimab)	mAb 114	Survivor of 1995 EVD pandemic	FDA-approved (December 2020)	» ZMapp
REGN-EB3 (Inmazeb)	REGN3470, REGN3471, REGN3479	Humanised mouse	FDA-approved (October 2020)	» ZMapp
MBP134	ADI-15878, ADI-23774	Survivor of 2014 EVD pandemic	Phase 1	NA

Table 4.1: | **Relevant Ebolavirus immunotherapeutics as of August 2022.** There are two therapeutic cocktails and one monoclonal therapy that have been approved by the FDA and were included in the 2018 PALM trials. ZMapp was licensed in 2014. Included in this table are the predecessors of ZMapp, ZMab and MB-003. MBP134 is in Phase 1 clinical trials; data on its efficacy was published post-PALM trials, so it was not included.

The first cocktails consisted of murine (ZMab) (Qiu et al., 2011) or human-mouse chimeric (MB-003) (Olinger et al., 2012) neutralizing antibodies. The murine ZMab antibodies were chimerized and combined with the most potent of the MB-003 mAbs, c13C6, to form ZMapp (Qiu et al., 2014). Finally, REGN-EB3 (Inmazeb) consists of three fully human mAbs (Pascal et al., 2018). In 2019, the first cocktail which exhibits pan-ebolavirus protection in NHPs was published, consisting of ADI-15878 and ADI-23774, an antibody matured specifically for SUDV GP binding (Wec et al., 2019). The greatest protective efficacy in NHPs that has been so-far observed for a single mAb is mAb114, an RBR-binding mAb (Corti et al., 2016). In the Pamoja Tulinde Maisha (Swahili for "Together Save Lives"; PALM) trials (the Phase 2/3 trials for mAbs 114, REGN-EB3 and ZMapp in combination with the anti-viral small molecule drug remdesivir), mAbs 114 and REGN-EB3 were found to be superior to ZMapp in reducing fatality (Mulangu et al., 2019). REGN-EB3 was licensed by the FDA in October 2020 and mAb114 was licensed in December 2020 (O'Donnell & Marzi, 2021). MBP134, which is the first cocktail to report protection against both EBOV and SUDV, is in Phase I clinical trials.

In addition to improving our ability to treat EVD, vaccination will be key in the control of this disease. Currently, Ervebo is the only vaccine with proven clinical efficacy and FDA approval (Woolsey & Geisbert, 2021): however, this efficacy is only shown for EBOV. Protection against the other species of Ebolavirus is desirable, and could be produced by a multivalent vaccine. This is the approach of the Ad26.ZEBOV and MVA-BN-Filo vaccine which contains the glycoproteins from EBOV, SUDV and MARV and the TFV NP in the boost. Safety and immunogenicity data have been recently presented for the Ad26.ZEBOV and MVA-BN-Filo vaccination regime as part of a consortium called EBOVAC2. Trials in Europe and Africa have provided evidence that this regime produces lasting anti-EBOV-GP titre a year post-vaccination (Anywaine et al., 2022; Ishola et al., 2022; Pollard et al., 2021).

With over three decades of work characterising the B cell response to Ebolavirus glycoproteins, a considerable number of EBOV-reactive antibody sequences are

publicly available. In this chapter, we use the methods described in Chapter 2 (clonotyping and paratyping) to computationally annotate the antigen-specific component of BCR repertoires from Ad26.ZEBOV/MVA-BN-Filo vaccinees using a novel database of experimentally-confirmed EBOV-binding antibody and nanobody sequences collected from the literature and patents (termed EBOV-AbDab).

We demonstrate that the predicted antigen-specific subset of the repertoire only increases significantly post-boost, despite significant clonal expansion post-prime. Predicted antigen-specific sequences are enriched for in expanded and convergent clonotypes. The proportion of the BCR repertoire predicted to be EBOV-reactive correlates significantly with serum titres of anti-EBOV-GP IgG. We identify a particularly convergent family of IGHV3-15/IGHJ6 antibodies via their similarity to previously described RBR-binding antibodies isolated from survivors of EVD and ChAd.ZEBOV/MVA-BN-Filo vaccinees, indicating that the response to Ad26.ZEBOV/MVA-BN-Filo vaccinees is targeted towards the receptor-binding region.

The work presented here is mine unless otherwise stated.

4.3 Methods

4.3.1 Trial and samples

The EBOVAC2 trials correspond to the Phase 2 trials of Ad26.ZEBOV and MVA-BN-Filo vaccination (Pollard et al., 2021).

In this two-dose heterologous vaccination regime, the prime consists of Zaïre EBOV (ZEBOV or EBOV) glycoprotein vectored with adenovirus serotype 26 (termed "Ad26.ZEBOV" or the "Zabdeno" component). The boost, termed "MVA-BN-Filo" or "Mvabea", consists of a modified Vaccinia Ankara virus (MVA) with glycoproteins from EBOV, SUDV and Marburg virus (MARV), and TAFV nucleoprotein.

Group	Interval (days)	N subjects
1	28	13
2	56	15
3	85	12
Placebo	NA	5

Table 4.2: | **Dosing intervals in vaccinees.** The three dosing intervals between Ad26.EBOV prime and MVA-BN-Filo boost range between 28 to 85 days.

Group	Ethnicity
Placebo	White (5)
Vaccinee	White (37); Asian (1); Multiple (1); Other (1)

Table 4.3: | **Ethnicity breakdown of vaccinees.** Subjects were all UK-based; the majority of subjects (42/45) were reported to be of white ethnicity; the remainder were of Asian (1), multiple (1) or described other ethnicity (1).

An eight-week interval (56 days) was found to be optimal in the licensure by the EMA for protective efficacy in non-human primates. The human subjects in this trial received the boost either 28 days (D28), 56 days (D56) or 85 days (D85) after prime.

Peripheral blood monocytes (PBMCs) were collected from 45 subjects at baseline (D0), eleven days post-prime (D11) and seven days post-boost (D35, D64 or D92). Forty of these subjects received the Ad26.ZEBOV/MVA-BN-Filo treatment and five received placebo for both doses (Table 4.2).

Peripheral blood was taken from volunteers at baseline, seven days post-prime and eleven days post-boost. This work was performed by the Oxford Vaccine Group.

PBMCs were isolated by routine Ficoll-Paque density gradient centrifugation and lysed. RNA extraction was performed on the lysate using the RNeasy Mini Kit (Qiagen, Hilden, Germany).

The majority of this work was performed by Sagida Bibi at the Oxford Vac-

cine Group.

4.3.2 BCR sequencing

Libraries were prepared using IgG-specific primers as described by Ghraichy and colleagues (Ghraichy et al., 2020) and sequenced on the Illumina MiSeq platform (2 x 300 bp protocol).

This work was carried out by Marie Ghraichy and Valentin Von Niederhäusern in the Trück lab at the University Children's Hospital in Zurich.

4.3.3 Immunogenicity data

For 42/45 subjects, titre of EBOV GP-reactive IgG was measured at baseline and 21 days post-boost via ELISA. A fold-change was calculated from baseline to post-boost. This work was carried out by the Oxford Vaccine Group.

4.3.4 Sequence analysis

Sequences were processed using the Immcantation pipeline (Gupta et al., 2015). Reads were assembled and UMIs processed. This work was carried out by the Trück lab with thanks to Marie Ghraichy and Valentin von Niederhäusern.

Sequences were aligned to an IMGT human germline database using the AssignGermlines module of Change-O (Gupta et al., 2015). Clonotypes were assigned using the DefineClones module of Change-O with IGHV and IGHJ matching and a 90% CDRH3 amino acid identity threshold.

Clonal expansion was primarily measured as the percentage of sequences in the repertoire among the ten largest clonotypes. This method choice is further justified in Section 4.4.1.

Metric	Formula
Proportion of sequences in ten largest clonotypes	$\sum_{i=1}^{10} v_i$
Gini index	$\frac{\sum_{i=1}^{n}(2i-n-1)v_i}{n\sum_{i=1}^{n} v_i}$
Shannon entropy	$-\sum_{i=1}^{n} v_i log_2 v_i$

Table 4.4: | **Formulae for clonal expansion metrics.** Formulae for the three clonal expansion metrics explored in this work: a simple metric defining the proportion of sequences in the ten largest clonotypes, the Gini index of inequality and Shannon diversity or entropy. p_i is the proportion of sequences in clonotype i. Vector v is the vector of proportions of sequences in n clonotypes, $v = p_1, p_2..., p_n$, where n is the number of clonotypes. The vector is sorted by clonotype size (i.e. $p_i > p_{i+1}$), such that v_i is the ith largest clonotype.

We verified our results with two separate clonal expansion metrics, Gini index and Shannon entropy. The formulae for these measures are shown in Table 4.4.

4.3.5 EBOV-AbDab curation

To be able to compare the sequenced BCR repertoires to known Ebolavirus-binding antibodies, we curated sequences from the literature and patents.

We collected antibody or nanobody sequences with experimentally-proven reactivity to any Ebolavirus protein from 24 publications and seven patent applications amounting to 984 sequences. We also collected binary neutralization labels and epitope or competition data. For the binding and neutralization fields, we recorded the Ebolavirus species for which binding or neutralization had been confirmed. Where a structure of an antibody or antibody:antigen complex has been solved via X-ray crystallography or cryo-EM, the PDB accession code was also collected.

In the curation process, the maximal amount of information was retained where possible (Figure 4.3). If nucleotide sequences were available, these were recorded and germline gene assignments were performed on these nucleotide sequences using nucleotide IgBLAST (Ye et al., 2013). If amino acid sequences were available, IGV gene assignment was performed using amino acid IgBLAST; IgBLAST-aa does not perform J gene assignment due to the difficulty of germline alignment in the junctional region without nucleotide information (Ye et al., 2013). In these instances, J gene assignment was performed using ANARCI (Dunbar & Deane, 2016) for completeness. Where only the IGV and IGJ genes and CDR3s were reported, these were collected as they permit clonal comparisons (albeit without verification or standardisation of the germline assignments).

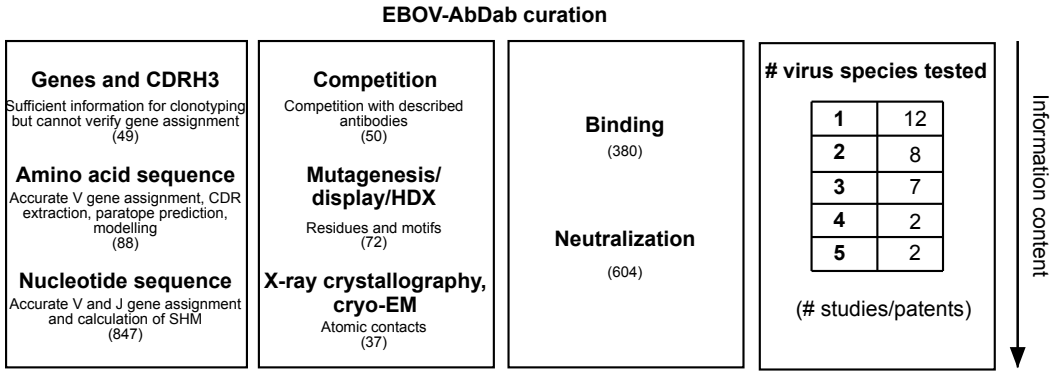

Figure 4.3: | **EBOV-AbDab curation.** An overview of information curated in EBOV-AbDab and the prioritisation of this with regards to information content.

4.3.6 Other antigen databases

We also compared repertoires to antibodies from CoV-AbDab, the largest database of sequences of antibodies binding to coronaviruses, (Raybould, Kovaltsuk, et al., 2021) and an influenza database which was collected from the literature and shared by Dr. Jacob Galson (private communication).

CoV-AbDab contained 9,054 human antibody sequences with IGHV gene and CDRH3 extracted as of 27th July 2022. The influenza database contained

513 antibodies with IGHV gene and CDRH3 amino acid sequences from 10 publications and 2 patents (Ahmed et al., 2013; Corti et al., 2010; Kaur et al., 2015; Krause et al., 2012; Krause et al., 2011; Li et al., 2012; Marasco et al., 2011; Meng et al., 2013; Ohshima et al., 2014; Okada et al., 2010; Pappas et al., 2014; Thomson et al., 2012).

4.3.7 Human OAS dataset

To establish a baseline for how common predicted Ebolavirus-binding antibodies are in non-exposed BCR repertoires, we used all of the human data in the Observed Antibody Space as of April 2022. This amounted to 5 x 10^9 heavy chain sequences from 69 studies. The majority of these sequences were from healthy repertoires (84.2%). The remainder of the datasets came from 24 disease indications, including EVD which was excluded, and 13 vaccines. Ethnicity is not reported in the Observed Antibody Space, so we were unable to attempt ethnicity matching to our vaccinee pool (described in Table 4.3).

4.3.8 Comparison to antigen databases including EBOV-AbDab

Repertoire sequences were compared to EBOV-AbDab, CoV-AbDab and an in-house influenza antibody database.

Antibodies were compared via i) clonotyping and ii) paratyping in the case of EBOV-AbDab.

For clonotyping, antibodies had to be human, share the same IGHV gene and have 72% or greater CDRH3 identity (amino acids). The 72% threshold was chosen as per the paratyping experiment described in Chapter 2. We did not impose an IGHJ gene matching requirement, as for a minority of EBOV-AbDab (137 antibodies out of 978) and the entirety of CoV-AbDab, IGHJ gene assignment was performed using amino acid sequences. As IgBLAST does not perform IGHJ gene assignment on amino acid sequences, ANARCI germline

gene assignment was used for IGHJ gene assignment for CoV-AbDab (private communication with Dr. Matthew Raybould). ANARCI IGHJ gene assignment is not reliable, so was excluded from the clonotype match definition. Clonotyping was used to compare repertoires to the human subset of EBOV-AbDab.

For paratyping, sequences were numbered via ANARCI (Dunbar & Deane, 2016). Chothia/Chothia CDRs were extracted with ± 2 residues either side, and paratope probabilities were assigned using Parapred (Liberis et al., 2018) with the same weights as in Chapter 2. Where two sequences had the same length CDRs, sequence identity was calculated over residues with paratope probabilities exceeding 0.67 as per the calculation in Chapter 2. Paratyping was used to compare repertoires to all EBOV-AbDab antibodies which could be numbered via ANARCI; antibodies with the same length CDRs and greater than 75% sequence identity across the predicted paratope to an antibody in EBOV-AbDab were predicted as EBOV-reactive.

As discussed in Chapter 2, maximal accuracy in the annotation of antigen-specific antibodies is attained when considering the intersection of predictions by paratyping and clonotyping. We label sequences in the repertoire as predicted binders if they are in the same clonotype and paratype of a given binder.

4.3.9 Structural modelling

Where applicable, antibody sequences were modelled using a combination of ABodyBuilder (Leem et al., 2016) for framework and orientation selection, and ABlooper for CDR modelling (Abanades et al., 2022). Models were considered high-quality if the diversity of five models was below 0.87 for CDRH1, 0.82 for CDRH2, 3.25 for CDRH3, 1.01 for CDRH1, 0.92 for CDRH2 and 1.10 for CDRL3 (recommendations by Brennan Abanades-Kenyon, personal communcations).

4.3.10 Ab-Ligity for structural paratope comparisons

In the work comparing influenza and Ebolavirus antibodies, we used Ab-Ligity to compare binding sites with a threshold exceeding 0.1 resulting in a prediction that two antibodies bind to the same epitope, as per the original paper (Wong et al., 2021). Ab-Ligity was used as its allows comparison of antibodies with differing CDR lengths.

4.3.11 Statistics

We performed a number of ANOVA tests using the *anova_test* function in R. We used mixed ANOVAs allowing us to compare both across groups and within subjects. Where applicable, we used post-hoc paired t-tests with Bonferroni correction.

4.4 Results

4.4.1 BCR repertoires of vaccinees show evidence of antigen exposure

We sequenced the heavy chain BCR repertoires of 45 subjects at three time points: at baseline, eleven days post-prime and seven days post-boost: 40 subjects who were vaccinated with Ad26.EBOV/MVA-BN-Filo at varying prime-boost intervals, and five subjects who received a placebo. This resulted in on average 19,512 IgG sequences and 4,519 clonotypes per sample, and in total 2,634,138 sequences.

A number of properties of the BCR repertoires changed significantly over time in the vaccinated groups, and did not change at equivalent time points in subjects which received the placebo, indicating that these effects result from vaccination. Repertoires became more clonally "top heavy", indicating clonal expansion. CDRH3s increased in length and isotype subtype frequencies changed

with increases in IgG1 and decreases in IgG2 following vaccination that exceed any fluctuations observed in the placebo group. Increases in IgG1 are consistent with exposure to viral protein antigens while IgG2 is largely associated with bacterial capsular polysaccharide antigens (Vidarsson et al., 2014), increasing our confidence that we are observing an anti-viral response.

To measure how clonally expanded a repertoire was, we calculated the percentage of sequences which were in the N largest clonotypes. N was selected to be ten in order to account for a degree of oligoclonality; to test how sensitive our ranking of repertoires in terms of clonality would be while varying N, we calculated the Spearman's rank correlation coefficient between each pair of rankings varying N between 1 and 100 and averaged over all N. N = 10 falls within the shoulder of the curve relating N to average correlation between different values of N, indicating that it is robust (Figure 4.4). Furthermore, Neumeier and colleagues found that only around the top ten expanded clonotypes were antigen-specific in bone marrow plasma cells in mice (Neumeier et al., 2022) indicating that this is an appropriate order of magnitude to select.

This simple measure was chosen as opposed to more complex measures such as Shannon diversity or Gini index due to the lower sensitivity of the measure to subsampling, with a near exact correlation across ten subsamples of 10,000 sequences resulting in a Pearson's correlation coefficient of 1.0 vs. 0.97 for Shannon diversity and 0.64 for Gini index. These relationships and corresponding R^2 values are shown in Figure 4.5.

The percentage of sequences in the largest ten clonotypes significantly increased from baseline to post-prime (Figure 4.6A). At baseline, on average 5.8% of sequences were found in the ten largest clonotypes vs. 14.2% post-prime and 10.5% post-boost. There was no signficant difference between baseline and post-boost expansion when including subject and group as variables within the regression. There was no significant group effect. The repertoires are most expanded at prime across all expansion metrics explored.

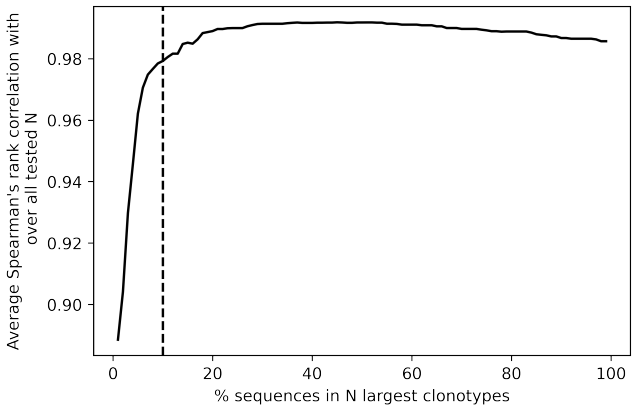

Figure 4.4: | **Sensitivity of clonal diversity measure to the number of clonotypes included in the calculation.** For every value between 1 and 100 (which we call N), the percentage of sequences in the N largest clonotypes was calculated per repertoire. The repertoires were then ranked. For each pair of values of N, the repertoire rankings were compared via Spearman's rank coefficient. For each value of N, the Spearman's rank coefficient was averaged over all other values of N. This is plotted here as a function of N. Naturally, this increases with increasing N, however large values of N do not reflect the proportion of the repertoire that is genuinely expanded in response to antigen. Ten was chosen as it sits in the shoulder of this relationship.

CDRH3 length increases significantly over time, from an average of 15.4 to 15.9 to 16.0 amino acids at baseline, post-prime and post-boost respectively (Figure 4.6B). There was no group effect. In the placebo group, average CDRH3 length was 15.0, 15.0 and 15.1 amino acids respectively.

There were also changes in the proportion of IgG1 and IgG2 relative to other IgG subtypes (Figure 4.6C). In vaccinees, the average proportion of IgG1 at baseline was 49.4% increasing to 65.1% post-prime and 71.3% post-boost. The majority of this change was compensated for by a decrease in IgG2, decreasing from 40.3% to 23.2% to 19.6%. In the placebo repertoires, IgG1 was more stable, appearing at a frequency of 41.0% at baseline, 43.9% post-prime and 41.5% post-boost on average. These repertoire measures for subjects who received placebo are shown in Appendix Figure 3.

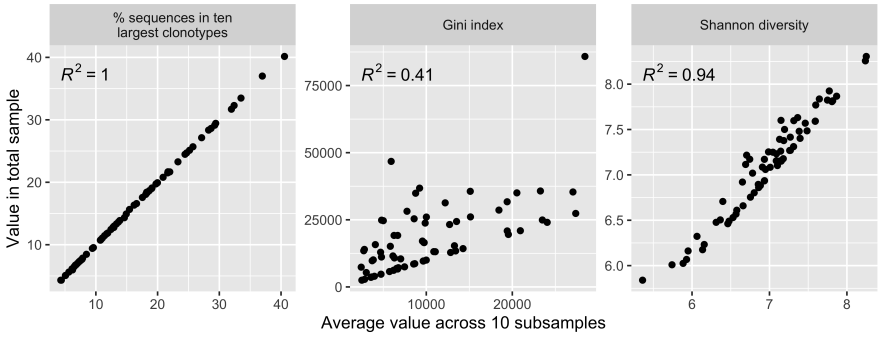

Figure 4.5: | **Sensitivity of three clonal expansion measures to subsampling.** A Pearson's correlation was calculated between each metric calculated over the full dataset and its value in subsamples of 10,000 sequences from the dataset, for each repertoire. This was repeated 10 times and the value was averaged per repertoire. Coefficient of determination (R^2) was calculated between the subsample average and the value in the total population. The percentage of sequences in the N largest clonotypes and the Shannon entropy were robust to subsampling. The Gini index was significantly less robust across subsamples.

The degree of SHM, measured by nucleotide identity to the reference IGHV allele, appears to decrease over time accounting for group effect. The median SHM decreases from 11.9% at prime to 11.0% at post-prime and 10.7% post-boost. However, this decrease in SHM is within the range observed between time points among the placebo cohort (minimally 11.8% at baseline, 12.2% post-prime and 9.1% post-boost).

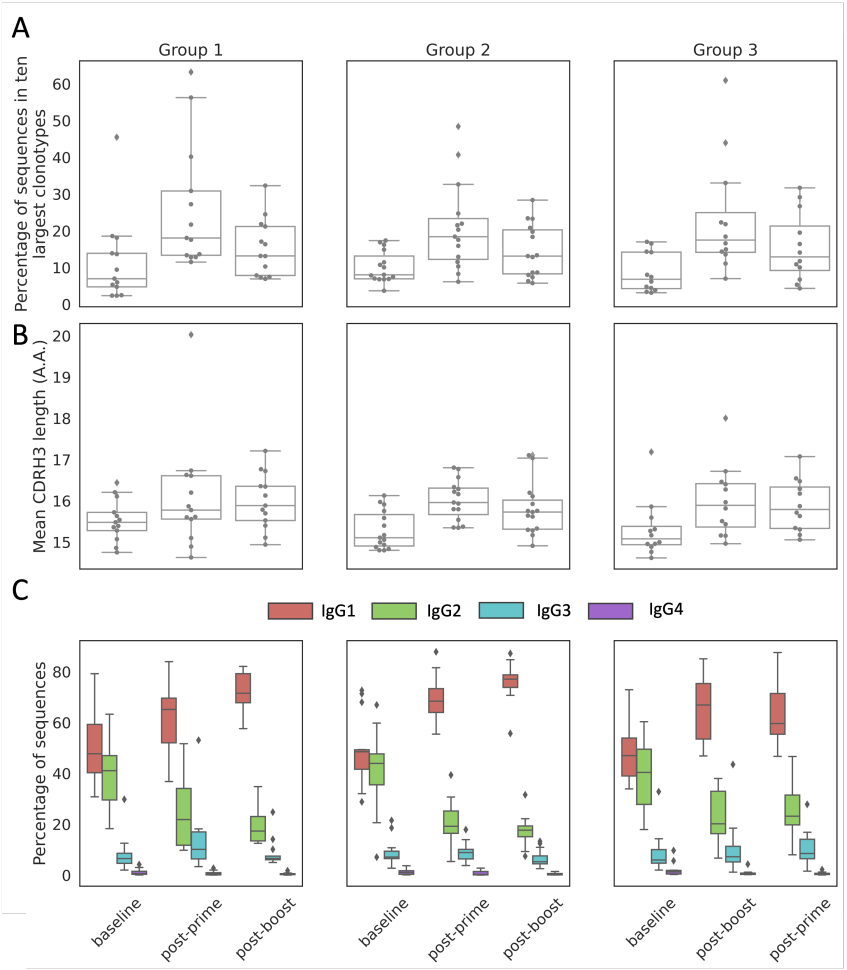

Figure 4.6: | **Evidence of repertoire perturbation in vaccinees with respect to clonality, CDRH3 length and isotype subtype frequency.** Percentage of sequences in the ten largest clonotypes increased significantly from prime to post-prime and significantly decreases from post-prime to post-boost (A). There is no group effect observed. CDRH3 length significantly increases as a function of time after vaccination (B). IgG subtype frequencies change with increases in IGHG1 compensated for by decreases in IGHG1.

4.4.2 Analysis of an Ebolavirus-binding antibody database

We wanted to see whether the clonotypes that we saw expanding in response to Ad26.EBOV/MVA-BN-Filo vaccination were predicted to bind to any of the vaccine components (multiple Ebolavirus glycoproteins and a TAFV nucleoprotein) via the clonotyping and paratyping methods described in Chapter 2. In order to do this, we created a database of antibodies that have already been validated as binding to Ebolavirus proteins. At the time of writing, there were just 37 receptor sequences for antibodies binding to Ebolavirus epitopes collected in the Immune Epitope Database (Rosenfeld et al., 2018). To overcome this, we curated our own database ("EBOV-AbDab" following the nomenclature established by "CoV-AbDab") of Ebolavirus-binding antibody and nanobody sequences.

We collected information about 984 characterised Ebolavirus-binding antibodies (978) and nanobodies (6) as per Methods Section 4.3.5 from 24 studies and 7 patent applications. The majority of antibodies (951/978) where characterised as binding to the glycoprotein with the remainder to VP40 (11), NP (5), VP35 (1), VP30 (1) and sGP alone (as opposed to primarily to GP with sGP cross-reactivity) (2). The majority of antibodies are of human origin (927), with the remainder of murine (44) or macaque (2) origin. All described nanobodies (6) were to NP and derived from llamas.

Where available, we recorded the epitope of each antibody. Epitope information was reported by authors for just 135 of the 978 antibodies, with varying resolution. As discussed in the literature, the majority of GP antibodies are to epitopes on GP1 (see Table 4.5). There are 33 where the epitope spans both the GP1 and GP2, and a small number targeting epitopes on the GP2 alone, primarily targetting the MPER.

The largest epitope class of antibodies with sub-GP protomer resolution epitope labels includes glycan cap epitopes (37/135 antibodies). Epitope information was stored separately from epitope competition data. Including antibodies

Protomer	Count	Most common epitope
GP1	85	Glycan cap
GP2	17	MPER
Both	33	GP1-GP2 base

Table 4.5: | **The number of antibodies to each of the GP protomers.** Among GP-binding antibodies with epitope labels in EBOV-AbDab (135), the majority bind to GP1 alone (85) of which the largest epitope group is the glycan cap. 17 antibodies in the GP2 largely target the MPER. 33 antibodies bind to epitopes in both the GP1 and GP2, predominantly at the base.

that are known to compete with any glycan cap antibody, there are 218 glycan cap antibodies in EBOV-AbDab. The MLD is also well represented with 18 labelled with an MLD epitope and 19 competing with any MLD-labelled antibody.

The receptor-binding region which is on the GP1 core masked by the glycan cap/MLD has been identified as a good candidate for breadth of neutralization as it is the most conserved region of GP. There are 13 RBR-binding antibodies in EBOV-AbDab including 13C6, a constituent of both MB-003 and ZMapp, and mAb 114. Note that 13C6 is labelled as either a glycan cap or RBR antibody (Rijal et al., 2019).

The majority of antibodies (815) have described reactivity just to EBOV. There are 134 with reactivity with BDBV, 109 with reactivity to SUDV, 16 with reactivity to RESTV, eight with reactivity to TAFV and five with reactivity to closely related MARV. Among antibodies with reactivity to more than one virus (130), diverse epitopes are represented.

143 of the antibodies have been shown to neutralize at least one Ebolavirus species. 19 antibodies have described neutralization against multiple species of Ebolavirus. Reactivity and neutralizing data are not definitive, in that not all antibodies are tested against all Ebolavirus species.

The average human CDRH3 length in EBOV-AbDab was 16.9 amino acids

which is longer than the average in non-mutated, naive human B cell repertoires that was found in Chapter 3 (16.6). This is in contrast to the coronavirus-binding antibody database (CoV-AbDab) in which the average CDRH3 length was 15.8 (Figure 4.7). The majority of human antibodies in the database are mutated (779/807). A direct comparison cannot be made with CoV-AbDab because nucleotide sequences are not curated by CoV-AbDab, but the literature suggests that many described antibodies are very close to germline (Clark et al., 2021; Galson et al., 2020; Kreer et al., 2020; Yan et al., 2021; Zhu et al., 2021). In terms of IGHV gene usage, IGHV3 antibodies are overrepresented in comparison to baseline human repertoires. In one of the deepest sequencing studies of the baseline human repertoire, IGHV3 usage was found to be between 39.1% and 47.4% (Briney et al., 2019). This contrasts with 55.2% of EBOV-AbDab and 51.0% of CoV-AbDab. IGHV1-69 also occupies a large proportion of these databases, as in anti-viral antibodies generally (Chen et al., 2019).

4.4.3 Ebolavirus-binding antibodies appear more frequently in vaccinee repertoires than in healthy repertoires

Given our diverse database of EBOV-binding antibodies, EBOV-AbDab, we attempted to predict which sequences in the repertoires were Ebolavirus-specific. We compared each of the vaccinee and control group repertoire sequences to EBOV-AbDab via clonotyping with a 72% CDRH3 sequence identity threshold and paratyping with a CDR length threshold and 75% predicted paratope identity threshold. To establish a baseline, we compared CoV-AbDab and an influenza antibody database (described in Section 4.3.6) to both the vaccinee and control group repertoires. Sequences which were predicted to bind an Ebolavirus protein due to clonotype- or paratype-similarity with an EBOV-AbDab antibody are referred to as "hit sequences".

We found that hit sequences increased in frequency post-boost in vaccinees from both prime and postprime (pairwise t-test, $p \ll 0.05$) while non-EBOV-

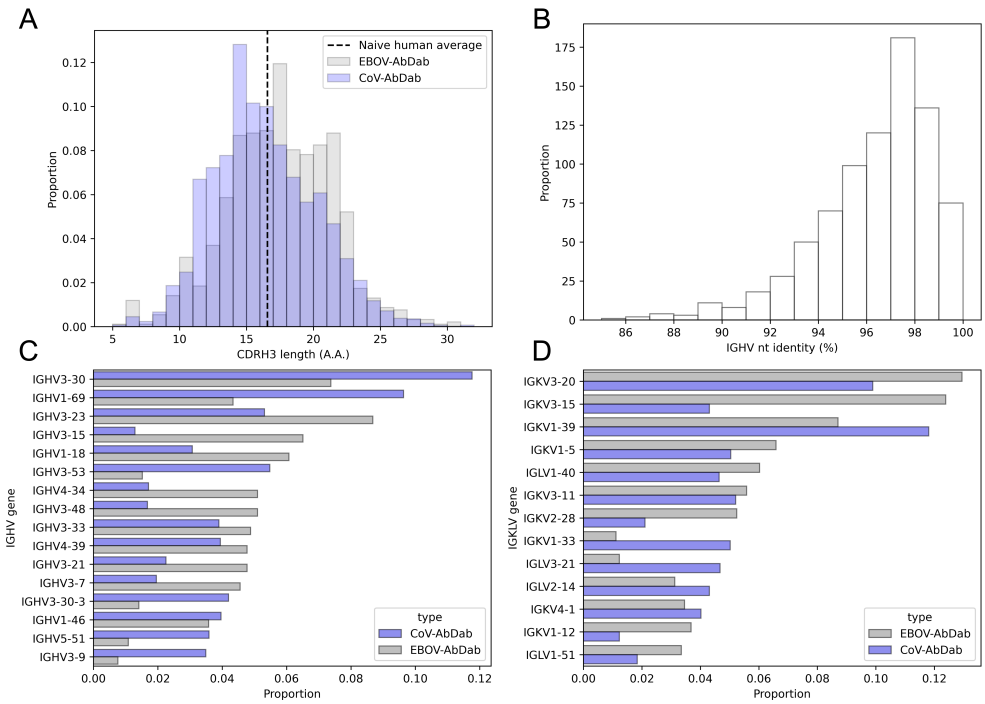

Figure 4.7: | **Summary statistics for EBOV-AbDab and CoV-AbDab .** A) CDRH3 length, B) somatic hypermutation in the heavy chain, C) IGHV gene usage and D) IGKLV gene usage are shown for EBOV-AbDab and another antiviral antibody database, CoV-AbDab. Antibodies with confirmed binding to Ebolavirus proteins have significantly longer CDRH3s than antibodies binding coronaviruses (A). Somatic hypermutation cannot be compared as nucleotide data was not collected for CoV-AbDab and the IGHV gene assignment of ANARCI is inaccurate at the allele level (B). Both COVID and Ebola antibodies have an enrichment of IGHV3 genes (C). IGKV3-20 appears at high frequency in both databases (D).

binding antibodies did not increase in frequency (Figure 4.8): the average change in hit rate against EBOV-AbDab from baseline to post-boost was on average 3.0% and was positive in 38/40 subjects while the average hit rate against CoV-AbDab and the influenza database was 0.003% on average, and was positive (with a maximal increase in 0.9%) in 19 of 40 subjects. There was not a significant group effect.

These hit sequences were disproportionately found in expanded clonotypes post-prime and post-boost. At prime, the average clonotype size was 4.50 and the average clonotype size of hit sequences was 4.95; post-prime, average clonotype size was 4.48 while average clonotype size of hit sequences was 20.0. Post-boost, average clonotype size was 4.36 while the average size of a hit clonotype was 42.7 sequences. At baseline, none of the hit clonotypes were in the ten largest clonotypes for the majority of subjects (43/45). The highest hit clonotype rank observed was the 8th largest. 4/40 vaccinees had hit clonotypes in the top ten sequences at prime, while post-boost, 27 out of the 40 vaccinees had hit clonotypes within the ten largest clonotypes, with 13 vaccinees where the largest clonotype was a hit.

A corollary of this is that despite significant expansion observed post-prime, the expanded clonotypes in most subjects do not resemble EBOV-AbDab antibodies. These expansions could be due to immunity to the adenoviral vector; alternatively, it could be that sufficient somatic hypermutation or CDRH3 length increase is required to resemble anything within EBOV-AbDab. We find instead that the average IGHV mutatedness of hits is lower than the wider repertoire; furthermore, the average CDRH3 length of the hit sequences is 12.6 amino acids compared to an average of 15.6 amino acids repertoire-wide. These data do not support the hypothesis that sequences are required to mutate from the germline or reach longer CDRH3 lengths before they attain similarity to EBOV-AbDab antibodies. It may be that post-prime, the responses are biased to previously unobserved responses to EBOV GP, or indeed to the adenoviral vector.

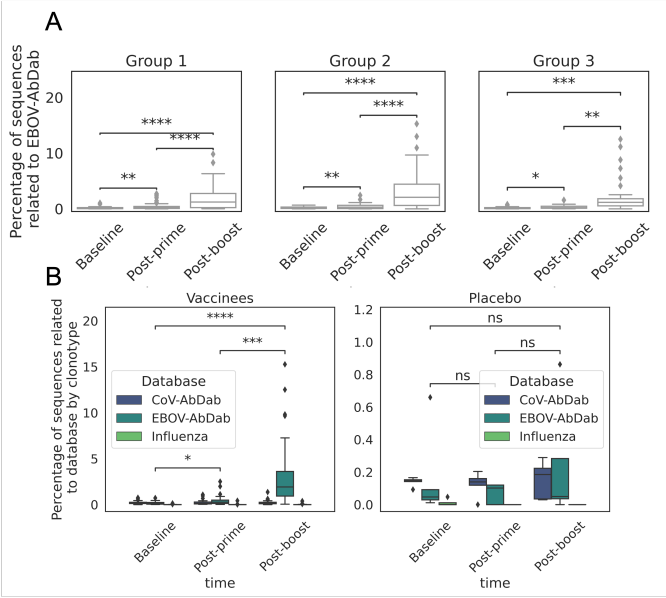

Figure 4.8: | **Increase in hits to EBOV-AbDab post-boost.** Figure A shows the percentage of sequences which are in the same paratype and clonotype to any antibody in EBOV-AbDab, stratified by group. There is no group effect with respect to the proportion of the repertoire annotated as EBOV-specific: this increases significantly from baseline to post-prime (at the 5% level, paired t-test) and most obviously from baseline and post-prime to post-boost (at the 0.01% and 0.1% level respectively, paired t-test). Figure B shows the percentage of sequences related to either EBOV-AbDab, CoV-AbDab or an influenza database by clonotype. While there is a non-zero hit rate for these antigen databases, this does not increase significantly over time in response to vaccination ($p = 0.83$, 0.60 and 0.36 for hits to the Influenza database; $p = 0.75$, 0.84 and 0.91 for the hits to CoV-AbDab from baseline to post-prime, post-prime to post-boost and baseline to post-boost respectively). In the placebo group, there is not an increase in the percentage of sequences mapping to any antigen database. Significance bars are shown for the change in the proportion of hits to EBOV-AbDab only: significant increases are observed for vaccinees (as in Figure A) but not for the placebo subjects ($p = 0.80$, repeated measures ANOVA; significance bars show post-hoc paired t-tests for graphical purposes).

4.4.4 EBOV GP-reactive serum IgG titres correlate with repertoire overlap with EBOV-AbDab

There was a significant correlation between EBOV GP-reactive serum IgG titre and the percentage of sequences in the repertoire which were annotated by paratype- and clonotype-similarity as related to EBOV-AbDab antibodies, i.e. the predicted antigen-specific proportion of the repertoire (Figure 4.9).

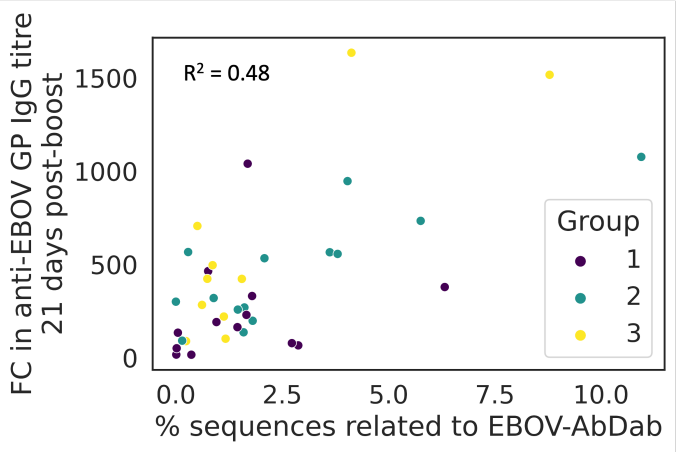

Figure 4.9: | **Correlation between EBOV-AbDab hit rate and serological data.** The anti-EBOV GP IgG titre as measured via ELISA is significantly correlated with the percentage of sequences which are predicted to be EBOV GP reactive ($p \ll 0.001$) with an adjusted R^2 value of 0.48.

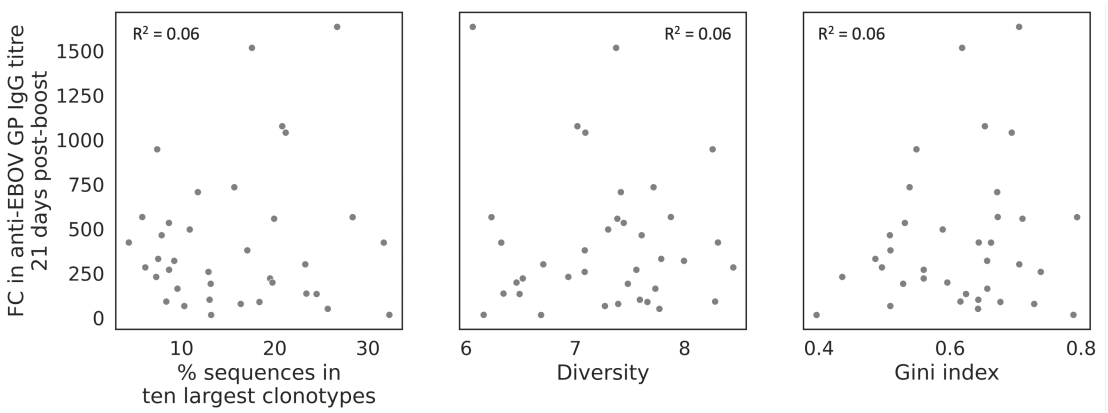

Figure 4.10: | **Clonal expansion in relation to serological data.** Clonal expansion measured by any of three metrics is not correlated with anti-EBOV GP IgG titre.

In our cohort of 40 subjects, the group effect according to boost interval was not signficant at the 5% level ($p = 0.06$). However, given that the full dataset of 345 subjects there was a significant group effect on fold-change in ELISA concentration with increasing fold-change as a function of interval ($p < 0.01$), we included group effect as a variable in our regression. While we note that the correlation is significant ($p \ll 0.001$), we also note that three of the subjects have disproportionate leverage in the regression due to their extreme values with regards to both variables.

There is no significant correlation between degree of clonal expansion post-boost, measured by the top heaviness of the repertoires, Shannon diversity or Gini index, and serum anti-EBOV GP IgG titre (Figure 4.10). This indicates the proportion of sequences with similarity to EBOV-AbDab is more predictive of serum anti-EBOV GP titre than repertoire perturbation measures like clonal expansion alone, given the described caveat about the outliers.

4.4.5 Analysis of the hit clonotypes from vaccinees

The majority of hit sequences observed at baseline correspond to three IGHV3-15/IGHJ6 antibodies, 6666 and 6662 (discovered in ChAd3-ZEBOV-vaccinated

individuals (Rijal et al., 2019)), and 3T0245 (discovered in rVSV vaccinees (Ehrhardt et al., 2019)) (507 of 697 hits). These antibodies will be referred to as 6666-like antibodies. Post-prime, 2,308 out of a total of 2,954 hit sequences show similarity to these antibodies; post-boost, 18,898 of these sequences from 22,853 total hit sequences correspond to 6666-like antibodies. The majority of the annotated antigen-specific component of the repertoire in vaccinees is this specific lineage of antibodies which have been described as binding to the receptor-binding region (Rijal et al., 2019).

Clonotype relatives (shared IGHV and greater than 72% CDRH3 amino acid identity) of 6666-like antibodies were found in 38/45 subjects at baseline, in 26 subjects with more than two reads, and 11 of these subjects with more than one read per UMI. Relatives were found in all forty vaccinees at either post-prime or post-boost. Contamination among time points at the RNA extraction or library preparation stage was ruled out by performing these steps separately for baseline, post-prime and post-boost samples for 23 of the 45 subjects. However, we cannot rule out Illumina index hopping at the sequencing stage as multiple time points were multiplexed. In this set of subjects, there were relatives of 6666-like antibodies in all 23 samples, with 3 subjects with at least one UMI represented by multiple reads.

Besides the 6666-like antibodies, there were a further 184 hit clonotypes that were identified convergently (i.e. across more than one subject). Many convergent or "public" clonotypes contained hits to EBOV-AbDab (Figure 4.11). The ten most convergent of these clonotypes are described in Table 4.6. Using information curated in EBOV-AbDab, it can be seen that among these highly convergent clonotypes are multiple neutralizing antibodies, from epitopes including the receptor-binding region (similarly to 6666-like antibodies) and glycan cap, mucin-like domain, and GP1/GP2 base or IFL.

There was also convergence in hit clonotypes between time points. In total, 4.3% of clonotypes were observed at more than one time point within an

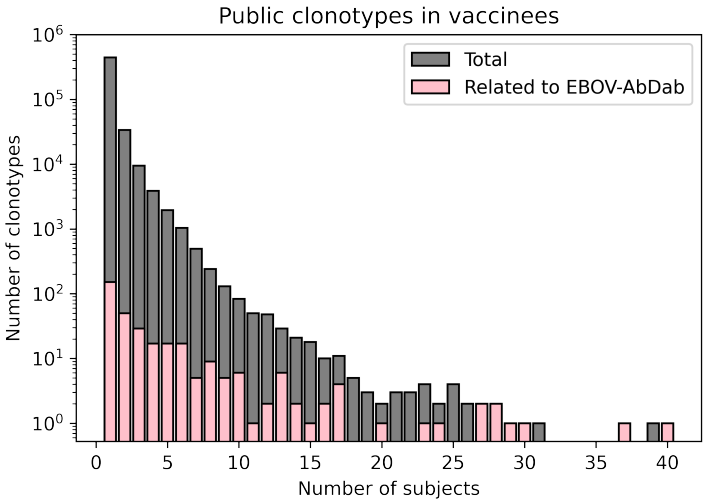

Figure 4.11: | **Convergent clonotypes and hits to EBOV-AbDab.** Public or convergent clonotypes observed according to the number of subjects they were observed in (post-prime or post-boost) and whether any clonotype members were hits to EBOV-AbDab. 51,382 clonotypes of 495,497 total clonotypes were identified across more than one subject. 185 of these 51,382 clonotypes contained hits to EBOV-AbDab (out of a total 336 hit clonotypes).

Antibody	# of subjects	Epitope	IGHV	Neutralizing
ADI-16015	25	RBR/Glycan cap	IGHV3-33	Unknown
121	23	Unknown	IGHV3-33	EBOV (partial)
3T0202	21	Unknown	IGHV3-13	EBOV
66-3-2B	18	GP1/GP2 base or IFL	IGHV3-21	EBOV
1T0451	17	MLD	IGHV1-3	EBOV
4T0444	16	Unknown	IGHV3-13	EBOV
5.1.7D03	15	GP1 core	IGHV3-13	EBOV
4T0578	12	Unknown	IGHV3-13	EBOV

Table 4.6: | **Non-6666-lineage hit antibodies in the repertoire data.** Besides antibodies related to 6666/6662/3T0245, the ten most convergent hit clonotypes and their associated EBOV-AbDab metadata are shown.

individual, whereas 12.5% of the clonotypes containing hits were observed at multiple time points.

In conclusion, the sequences which were highlighted via their similarity to EBOV-AbDab antibodies by combined clonotyping and paratyping are enriched for in expanded and convergent clonotypes. There is exceptional convergence of one particular RBR-binding lineage of antibodies, 6666-like antibodies.

Among other convergent clonotypes multiple epitopes are represented; where tested, these antibodies are largely neutralizing.

4.4.6 Comparison of EBOV-AbDab to the Human OAS dataset

In order to rationalise the ubiquity of 6666-like antibodies in both baseline and antigen-exposed BCR repertoires, we compared EBOV-AbDab to all human sequences in the Observed Antibody Space (described in Section 4.3.6), referred to as the Human OAS dataset. If relatives of 6666-like antibodies are prevalent in a large number of sequenced repertoires, their commonness in baseline vaccinee repertoires would not be surprising. Because of the size of this search dataset (5×10^9 total sequences from 69 studies and 786 subjects), we only compared sequences via clonotyping due to the computational expense of paratope prediction.

We found just 1,012 heavy chain sequences clonally related to 6666-like antibodies, described in Table 4.7. Table 4.7 also shows the frequency of antibody 2.1.1D05, which was one of the most commonly observed antibodies in EBOV-AbDab in the Human OAS dataset that is related to the 6666-like antibodies but has a CDRH3 identity of just 45.5% to 6666. Of the 6666-like antibodies, 6666 was the most common, observed in the repertoires of 35 subjects which were most commonly sampled in a "healthy" i.e. not recently disease or vaccine exposed repertoires. If we were to take the representativeness of 6666-like antibodies in healthy repertoires in the Human OAS dataset as representative of the population generally, we would estimate a frequency of 5.6% (25 subjects out of 267 healthy subjects); the probability of observing this antibody in 38 healthy subjects would therefore be 2.7×10^{-38}, indicating that this is highly improbable.

We then looked at generally how common clonal relatives of EBOV-binding antibodies in the Human OAS dataset were. We first report that the baseline frequency of EBOV-binding antibodies in non-EBOV-exposed repertoires on aver-

mAb	CDRH3	# observations in Human OAS dataset	Most common repertoire type	# subjects
6666	TTRWVYNGMDV	792	Healthy	35
6662	TTRIVLNGMDV	485	Healthy	23
3T0245	ATRVYHYNGMDV	51	TIV vaccinee	10
2.1.1D05	TTVYRYNYDSV	42,670	TIV vaccinee	5

Table 4.7: | **IGHV3-15 antibodies and their frequency in the Human OAS dataset.** The three most commonly observed "hit" EBOV-AbDab antibodies in the data from the vaccinees are 6666, 6662 and 3T0245. 2.1.1D05 is a related IGHV3-15/IGLV1-40 antibody that did not have clonal relatives in the vaccinees, but was the most commonly observed antibody from EBOV-AbDab in the Human OAS dataset. Shown are the CDRH3s, the number of observations in the Human OAS dataset at the level of sequences, the most common repertoire type in which the antibody was observed and the number of subjects it was observed in.

age was similar to our observation of baseline vaccinees. In non-Ebolavirus repertoires containing more than 10,000 sequences, on average 0.09% of sequences were clonally related to EBOV-AbDab while our average baseline hit rate among the vaccinees and control group was 0.11%, and maximally 0.66%. Maximally 15.2% of a repertoire was predicted to be EBOV-binding, corresponding to a baseline in a TIV vaccination study (Ellebedy et al., 2016); 12,099 sequences out of the total repertoire of 79,104 sequences were clonally related to an EBOV-AbDab antibody, the majority of which were hits to 2.1.1D05 (Table 4.7). The three most commonly observed antibodies across all non-EVD repertoires were 3T0478 (69,923 sequences), 5T0378 (61,429 sequences) and 5.1.5E07 (41,803 sequences) (Table 4.8).

Antibody	IGHV gene	CDRH3	Number of subjects
5T0378	IGHV4-30	ARFFDY	167
3T0478	IGHV4-59	ARDKRYYDSSGYSHFDY	163
5.1.5E07	IGHV1-69	ARDLGTPSDY	134

Table 4.8: | **Non 6666/6662/3T0245 antibodies in the Human OAS dataset.** The three EBOV-AbDab antibodies that most commonly had relatives at the 72% CDRH3 identity level in non-EVD related human BCR repertoires in the human OAS dataset. Their corresponding IGHV and CDRH3 are shown, as are the number of subjects in which a relative of the antibody appears.

It can be seen that 6666-like antibodies are disproportionately represented at baseline in comparison to non-EVD repertoires in the Human OAS dataset, and a number of other EBOV-AbDab antibodies are much more commonly seen in non-EVD repertoires. 154 antibodies, out of the database of 927 human antibodies, were seen in more subjects than 6666-like antibodies. In conclusion, this lineage of antibodies does not appear to be so widespread in the vaccinees purely due to its frequency in the general population.

We then looked at the repertoires in which the IGHV3-15 antibodies (6666, 6662, 3T0245 and 2.1.1D05) were found. The majority appear in TIV (Trivalent Influenza Vaccine) vaccinees (29,631 sequences out of a total 43,206 across 6 subjects which received the TIV vaccine, in two studies (Ellebedy et al., 2016; Sevy et al., 2019)). All four antibodies were found in TIV vaccinees. Of particular interest is that the majority of the IGHV3-15 antibodies were found in HA-sorted memory and antigen-secreting cells in the TIV vaccinees, which suggests that these antibodies may be cross-reactive with HA. Given the appearance of these IGHV3-15 antibodies, primarily of related antibodies to 2.1.1D05, in HA-sorted B cells from TIV vaccinees, we hypothesised that these antibodies are so common in the response to Ebolavirus vaccination due to activation of HA-reactive memory B cells.

4.4.7 Comparison to 6666/6662/3T0245 to influenza antibodies

There was no clonal overlap at the level of shared IGHV and greater than 72% sequence identity in CDRH3s between EBOV-AbDab and the influenza database. However, as discussed in Chapter 2, related antibodies with low CDR identity and differing CDR lengths are observed to bind to the same epitopes.

To examine the possibility that IGHV3-15 antibodies play a dual role in B cell immunity to influenza and Ebolaviruses, we modelled the IGHV3-15 sequences from the influenza database and EBOV-AbDab and compared their predicted binding sites using Ab-Ligity (Wong et al., 2021).

We were able to identify sequences (as opposed to gene combinations and CDRH3 sequence) for eight IGHV3-15 HA-reactive mAbs from a 2010 study (Okada et al., 2010). We used Ab-Ligity to compare the IGHV3-15 EBOV mAbs with these flu reactive mAbs, and with all full sequences from CoV-AbDab as a baseline with no hypothesis of cross-reactivity. We identified four mAbs where heavy chain Ab-Ligity similarity exceeded 0.1, while no other non-RBR binding mAbs in the complete dataset did. In combination with a separate influenza antibody sequence with an identical IGLV1-40 light chain (Dugan et al., 2020), we decided to test for cross-reactivity of the known RBR-binding antibodies to H3N2 hemagglutinin.

Pramila Rijal tested 6666, 6662, 66-2-BC, 105 and P7 (all IGHV3-15/IGLV1-40 mAbs) for binding to H3N2 infected cells with A/Victoria/3/1975, A/Sydney/5/1997 and A/Perth/16/2009 using Medi8852 as a positive control; no binding was observed.

4.5 Discussion

Continued outbreaks of Ebolavirus and closely-related filoviruses such as Marburgvirus demonstrate that there is a clear and unmet need for widespread prophylactic vaccination in endemic regions (Bausch, 2021). Ebolaviruses are challenging due to the diversity of their surface glycoprotein. Both rVSV-EBOV (Ervebo) and the heterologous adenoviral prime and multivalent MVA-vectored boost (Ad26.ZEBOV and ChAd3.ZEBOV (Woolsey & Geisbert, 2021)) have shown promising results in terms of safety, immunogenicity and protection, however, we currently lack experimental methods to be able to characterise epitopes of B cells at scale, so the epitopes contributing to this protection (and therefore the sensitivity of protection to mutation) are unknown. Furthermore, the bulk BCR repertoire of Ebolavirus vaccinees has never been studied at scale, so the genetic features of these antibodies and the reproducibility of the B cell response to vaccination has not been described. In this work, we apply computational methods

to annotate sequences with predicted epitope and antigen specificity that were developed in Chapter 2, namely immune repertoire mining, to predict the epitopes of over 2.5 x 10^8 sequences from vaccinees at baseline, post-prime and post-boost.

We showed that the IgG BCR repertoires of forty subjects vaccinated with Ad26. ZEBOV/MVA-BN-Filo showed evidence of antigen exposure in terms of clonal expansion as well as increases in CDRH3 length and changes in isotype subtype frequencies which were not observed in the repertoires of five subjects who received placebo. We were then interested in whether we were able to relate any of these sequences to previously described Ebolavirus-binding antibodies.

We collected antibody and nanobody sequences from 24 studies and 7 patents, resulting in the largest known database of EBOV-binding antibodies, which we call EBOV-AbDab. EBOV-AbDab, which currently contains 984 sequences from humans, mice, macaques and camelids to five of the ten known Ebolavirus proteins, was used as the input database for immune repertoire mining. Database curation is currently a task that relies on manual curation: the IEDB described their process which includes some automation but still ultimately relies on expert curation (Salimi et al., 2020). Use of natural language processing to supplement the database curation step will allow the continued upkeep of EBOV-AbDab, which will also be made available as a website similarly to CoV-AbDab in order to maximise its utility in the community of BCR researchers working on Ebolaviruses (Raybould, Kovaltsuk, et al., 2021). The Saphire lab continues to create standardised datasets for both COVID (COVIC-DB) (Hastie et al., 2021) and Ebolaviruses (the VHF consortium) (Saphire et al., 2018), however the mAb identifiers are blinded so this information cannot be combined with EBOV-AbDab.

We then compared EBOV-AbDab to bulk IgG BCR repertoires of the vaccinees. We were able to find 207 of the 927 human EBOV-AbDab antibodies in

these repertoires. The proportion of the repertoire with similarity to described EBOV-GP-binding antibodies increased significantly post-boost; at least one of the ten most expanded clonotypes post-boost from 27 out of 40 vaccinees was a hit to EBOV-AbDab. Despite signficant expansion of the repertoires post-prime, we were unable to annotate the majority of the expanded clonotypes (i.e. they showed no similarity to EBOV-AbDab). This demonstrates a significant limitation of the method: database size and diversity. As these antibodies do not appear similar to anything in our database, we cannot make any predictions about their functionality. Our inability to make predictions about these sequences suggests that either there is prevalent vector immunity post-prime or that our database is missing a significant number of antibodies important in the earliest response to EBOV GP. A database of antibodies with known reactivity to common viral vectors such as adenoviruses would be extremely useful here in order to exclude vector immunity: the IEDB did not contain any adenovirus serotype 26 antibodies, so we were unable to pursue this. Expression and functional characterisation of these post-prime antibodies will be necessary to elucidate how this response differs from the post-boost response.

The most common "hit" antibody in EBOV-AbDab was 6666 and related 3T0245 and 6662. 6666-like lineages appeared at baseline and showed remarkable convergence, appearing in all forty vaccinees post-prime or post-boost. The antibodies were also present at baseline in 38 subjects, and in more than half of subjects where there were multiple reads or reads represented by multiple UMIs (26/45). While we cannot completely rule out either contamination or Illumina index hopping, we processed the baseline, post-prime and post-boost samples on separate days for 23 subjects and again found these antibodies at baseline in multiple subjects. Given the anti-contamination measures we took, the most likely explanation would therefore be index hopping. A more sophisticated statistical analysis would be required to calculate the probability that the widespread occurrence of this antibody is not simply due to this phenomenon. However, we note that it seems improbable that in these two separate experiments (sequencing of the repertoires of the first 22 subjects, and then the subsequent 23 subjects with timepoint-separated RNA extraction and library prepa-

ration), the same lineage of antibodies is the only lineage we suspect of index hopping (the next most commonly observed lineage is seen in 25 subjects). In any case, the ubiquity of this antibody is consistent with the dominant role such antibodies appear to play in the response to EBOV vaccination and infection, as judged by their discovery in four separate studies (Bornholdt et al., 2019; Davis et al., 2019; Ehrhardt et al., 2019; Rijal et al., 2019).

IGHV3-15 antibodies have been isolated in multiple EVD survivors, rVSV-ZEBOV vaccinees and ChAd3-EBOV/MVA-BN-Filo vaccinees. Single-cell sequencing of EBOV-GP-sorted B cells from four rVSV-EBOV vaccinated individuals and subsequent cloning, expression and testing of 93 mAbs indicated that IGHV3-15/IGLV1-40 antibodies may play an important role in rVSV-EBOV vaccine-induced immunity; IGHV3-15/IGLV1-40 mAbs were found in each of the four vaccinees (Ehrhardt et al., 2019). A number of IGHV3-15/IGLV1-40 mAbs were also isolated independently from two survivors of EVD (Bornholdt et al., 2019; Davis et al., 2019). Cloning of EBOV-GP-reactive antibodies from ChAd3.ZEBOV vaccinees also revealed prevalence of IGHV3-15/IGLV1-40 antibodies with 22.0% of the 83 antibodies isolated deriving from IGHV3-15/IGLV1-40 (Rijal et al., 2019). We only sequenced the heavy chains which means that we do not know that all of our 6666/6662/3T0245 antibodies are natively paired with IGLV1-40 light chains: paired sequencing of these samples will be vital for confirming this and providing information about the evolution of the cognate light chains, which appear to play a critical role in breadth of reactivity (Cohen-Dvashi et al., 2020). Our data provides further evidence that this class of antibodies plays a major role in the response to Ebolavirus vaccination. Making predictions about the breadth of binding engendered by these antibodies, perhaps by docking of antibody model structures onto the known IGHV3-15/IGLV1-40 antibody epitopes using any of a number of docking and reranking approaches (Davila et al., 2022; Jin et al., 2022; Myung et al., 2022; Pierce et al., 2014; Schneider et al., 2021; Weitzner et al., 2017), would be extremely useful for assessing whether targeting IGHV3-15 antibodies is conducive to a desirable antibody response. If not, this has implications for immunogen design.

A major concern in the development of an Ebolavirus vaccine candidate is the breadth of protection provided. The multivalent boost in the Ad26.ZEBOV/MVA-BN-Filo regime is intended to increase the breadth of the protection engendered by the monovalent prime. As discussed, we cannot make predictions about the nature of the EBOV-specific response after the monovalent prime as the most expanded sequences do not show similarity to described antibodies; among the antibodies at post-boost, 6666 is not broadly neutralizing. Of related IGHV3-15/IGLV1-40 antibodies, the majority are not broadly reactive or neutralizing despite the relative sequence conservation of the RBR (Cohen-Dvashi et al., 2020). This suggests that the most convergent response is unlikely to be broadly reactive; however, the fact that largely different sequences are expanded post-prime than are identified as hits post-boost, could indicate that the B cell response changes signficantly as a result of exposure to the glycoproteins of other species of Ebolavirus.

We found related antibodies to another IGHV3-15 antibody, 2.1.D105, in the Human OAS dataset. These antibodies were found commonly across TIV vaccinees, including at baseline but also in HA-sorted B cell samples (Ellebedy et al., 2016). This led to our hypothesis that these antibodies are cross-reactive with influenza hemagglutinin (which is also a class I fusion protein); their convergence could be a result of activation of HA-reactive memory cells. This hypothesis was supported by predicted binding site similarity (measured with Ab-Ligity (Wong et al., 2021)) in the heavy chain between a number of IGHV3-15 influenza and EBOV-AbDab antibodies; no similarity was observed to IGHV3-15 antibodies in CoV-AbDab. While this hypothesis of cross-viral reactivity seems unlikely, such a phenomenon has been observed before for influenza and HIV, where anti-HIV antibody 2G12 was found to also bind and neutralize H3N2 due to a shared glycan epitope (Lee et al., 2021). As a result of the combined evidence of an antibody related to 6666, 2.1.D105, found in HA-sorted cells and Ab-Ligity similarity to a number of H3N2 antibodies, we tested a number of the IGHV3-15/IGHJ6 RBR-binding antibodies isolated from vaccinees against cells infected with three strains of H3N2. No cross-reactivity was observed. It would be interesting to test 2.1.D105 but we did not have access to

any material.

Besides IGHV3-15/IGHJ6 antibodies, antibodies related to EBOV-AbDab were generally enriched for in convergent and expanded clonotypes. Among these antibodies, the majority display neutralization against EBOV, and multiple epitopes are represented. IGHV3-33 and IGHV3-13 were common; IGHV3-13 has also been previously highlighted in the literature as RBR-targeting mAb114 is an IGHV3-13 antibody; it is prevalent in rVSV-EBOV vaccinees (Ehrhardt et al., 2019) and EVD survivors (Davis et al., 2019), and the macaque orthologue IGHV3-13 appears to play a critical role in the NHP immune response (Cagigi et al., 2018). This provides further evidence of a highly convergent response to Ebolavirus vaccination.

It is worth noting when considering the generalisability of these data that the subject pool was not ethnically representative of Ebola virus endemic regions, nor was it ethnically diverse. There is growing evidence of the influence of allelic polymorphism on the genetic composition of the BCR response (Avnir et al., 2016; Pennell et al., 2023; Pushparaj et al., 2023; Sangesland et al., 2022), so it is not necessarily the case that the notable IGHV3-15 convergence we observe here would be found in other cohorts (although IGHV3-15 antibodies have been isolated from EVD survivors (Davis et al., 2019)). A more diverse, haplotyped subject cohort would be required to establish whether our data are representative of non-Caucasian cohorts and to what extent the response is linked to particular allelic variants.

4.6 Chapter summary

In this chapter, we applied the methods discussed in Chapter 2 to BCR repertoires from Ebolavirus vaccinees.

We first showed that the BCR repertoires of Ebolavirus vaccinees show sig-

nificant clonal expansion, increases in CDRH3 length and changes in isotype subtype frequencies that are not observed in the repertoires of subjects who received placebo, suggesting that antigen exposure is observable in the BCR repertoire of these vaccinees.

We collated a database of antibodies with experimentally-confirmed binding to any Ebolavirus protein, and associated metadata such as binary neutralization labels, competition data and epitope labels; this database is referred to as EBOV-AbDab. EBOV-AbDab was then compared to the IgG repertoires of the vaccinees and control group subjects via clonotyping and paratyping, as this was established as the most precise method of group antigen-specific antibodies in Chapter 2.

Via immune repertoire mining, we were able to describe some of the most expanded clonotypes observed in vaccinees in terms of their similarity to described EBOV-binding antibodies. The proportion of sequences related to EBOV-AbDab increased significantly post-boost; the proportion of sequences related to EBOV-AbDab post-boost showed better correlation with anti-EBOV GP serum IgG titre than clonal expansion measures.

We showed extreme convergence of IGHV3-15/IGHJ6 antibodies with similarity to previously described RBR-binding antibodies, 6666 and 3T0245. Comparison to human sequences in the Observed Antibody Space revealed that related antibodies are found in multiple subjects and particularly in TIV vaccinees; these antibodies have predicted binding site similarity to a number of influenza antibodies, however did not show cross-reactivity when tested against a number of H3N2 strains. Further analysis of this class of antibodies will be required to understand their prevalence in the immune response to Ebola.

This work demonstrates that immune repertoire mining is a valuable technique for annotating the antigen-specific component of BCR repertoires.

5 | Sequence and structural analysis of VH:VL pairing

5.1 Abstract

The prediction of native VH:VL pairing preferences remains an intractable problem, however single-cell sequencing is producing an increasing body of natively paired VH:VL sequences. At the same time, the database of crystallographically-solved structures of antibodies has grown extensively since the last general analyses of VH:VL interfaces (Chothia & Lesk, 1987; Vargas-Madrazo & Paz-García, 2003). The sum of these data has yet to be analysed in the context of VH:VL pairing. We first analysed the VH:VL interfaces of human antibodies in the structural database SAbDab by examining their interactions and energetics. We were able to relate some of the variance in these parameters to the genetic origins of these antibodies. We subsequently analysed the paired human data in the Observed Antibody Space to identify potentially over- or underrepresented gene pairings. Finally, we present data from a novel sequencing method which produced over 3x as much data as was available prior, and validated one of our previously identified overrepresented gene pairings.

The work here is mine unless otherwise stated.

5.2 Introduction

The majority of the work presented in this thesis has dealt exclusively with the heavy chain; as described in Section 1.2.3, until the last decade, the heavy and light chain of an antibody could only be sequenced at the throughput of a handful of B cells of interest, creating bias towards a small number of antigen-specific antibodies. Due to the special role of the CDRH3 in the determination of antibody specificity (Xu & Davis, 2000), heavy chain sequences have received preferential sequencing attention resulting in over 2 billion heavy chain sequences

having accumulated over the last decade (2.1×10^9 unique sequences from 911 subjects and 86 studies) with significantly fewer studies choosing to sequence the light chain (3.6×10^8 unique sequences from 352 subjects and 43 studies, data collected from the Observed Antibody Space as of August 2022).

However, the light chain is involved in the paratope for the majority of solved antibody/antigen complexes, and was indeed dominant in terms of number of paratope residues in 14% of complexes in a non-redundant structural dataset (Richardson, 2020). Knowledge of the light chain is critical for the correct inference of an antibody's binding profile, and this is a fundamental limitation of bulk heavy chain sequencing. For example, in the previous chapter we identified relatives of a convergent EBOV GP-binding antibody, 6666, in multiple vaccinees. All of the other IGHV3-15 antibodies that bind the EBOV GP receptor-binding region curated in EBOV-AbDab are paired with an IGLV1-40 light chain. The light chain appears to play the most critical role in the breadth of these antibodies, according to a structural study on three such antibodies (Cohen-Dvashi et al., 2020). As we did not sequence the light chain repertoire, we were not able to make predictions about the breadth of binding of these important antibodies.

The development of single-cell sequencing now enables sequencing of VH:VL pairs from 10^2 - 10^3 B cells at a time via 10x Genomics, a microfluidics-based method described in Section 1.2.3. Despite the relatively low depth of this method in comparison to bulk sequencing, nearly 90,000 paired VH:VL sequences have been curated in the Observed Antibody Space across six studies (Olsen, Boyles, et al., 2022). In addition to 10x Genomics, alternative microfluidics-based single-cell sequencing platforms have been developed (DeKosky et al., 2016). This platform has recently been extended by the labs of John Frater and Lynn Dustin and applied to samples from 28 subjects, resulting in the sequencing of over 300,000 VH:VL clonotypes which is more than three times the existing corpus of paired data in the Observed Antibody Space. These two sets of data allow us to analyse VH:VL pairing at an unprecedented scale.

In addition to the increase in VH:VL sequencing data, significantly more structural data has been deposited since the most recent analyses of VH:VL complex interfaces (Chailyan et al., 2011; Chothia & Lesk, 1987; Vargas-Madrazo & Paz-García, 2003). In this chapter, we analyse human VH:VL paired structures within SAbDab (Dunbar et al., 2014) to provide an up-to-date overview of the VH:VL interface in terms of contact profiles, predicted interface energetics and orientation. We then relate these contact profiles to the genetic origins of these antibodies, focussing on the significant differences in contact profiles and VH:VL domain orientation between kappa and lambda antibodies. We then analyse the currently available paired human data in the Observed Antibody Space (Kovaltsuk et al., 2018; Olsen, Boyles, et al., 2022) to identify over- and underrepresented gene pairings. In the final section, we describe an analysis of a novel single-cell sequencing dataset and find some correlation in these results to the results from paired OAS.

5.3 Methods

5.3.1 Human structural dataset

IMGT-numbered antibody structures solved by X-ray crystallography, with complete heavy and light chains, resolution below or equal to 3Å and labelled as of human origin in their original PDB files were extracted from SAbDab in January of 2022. Excluding scFvs, this resulted in 3,506 Fvs from 1,874 complexes.

The 3,506 Fvs were clustered to 100% amino acid identity across the heavy and light chain using cd-hit (Li & Godzik, 2006), resulting in 1,323 Fvs.

IGHV and IGKLV were assigned using IgBLAST-aa (Ye et al., 2013). IGHJ and IGKLJ were assigned using ANARCI but were not used elsewhere (Dunbar & Deane, 2016); as discussed, IgBLAST-aa does not make IGJ assignments.

Study	N subjects	N sequences	Disease
Mor et al, 2021	15	47,455	SARS-CoV-2
King et al, 2020	6	31,071	Tonsillitis
Setliff et al, 2019	2	5,547	HIV
Woodruff et al, 2020	1	3,430	SARS-CoV-2
Eccles et al, 2020	1	771	None

Table 5.1: | **Overview of Human Paired OAS datasets.** Studies, number of subjects and sequences, and disease status of subjects from each study in the Human Paired OAS dataset.

5.3.2 Human sequencing datasets

5.3.2.1 Human Paired data from the Observed Antibody Space

All human, paired data was extracted from the Observed Antibody Space as of March 2022 resulting in 88,274 VH:VL (87,312 unique) sequences from 25 subjects, outlined in 5.1. In each of the studies, sequencing libraries were prepared using the 10x Genomics Chromium platform which is described in Section 1.2.3. Annotations such as IGHV, IGKLV, IGHJ, IGKLJ and CDRs were used as provided (these are inferred using IgBLAST (Ye et al., 2013)).

In order to prevent certain gene pairings appearing overrepresented due to clonal expansion in antigen-experienced repertoires, we clustered the sequences into clonotypes using the DefineClones module of Change-O (Gupta et al., 2015). Clonotypes were defined as sequences with the same IGHV, IGHJ, IGKLV and IGKLJ genes and with greater than 90% amino acid sequence identity across the CDRH3s. This clustered the 88,274 sequences into 85,111 unique clonotypes.

5.3.2.2 Frater/Dustin single-cell sequencing data

All experimental work in the production of this sequencing method was performed by the labs of Professor Lynn Dustin and Professor John Frater, and the read merging was performed by Dr. Nicholas Grayson.

A microfluidics platform built on the method developed by the lab of George

Georgiou (as described in their 2016 paper (DeKosky et al., 2016)) was developed and optimised by the labs of Professor John Frater and Professor Lynn Dustin. Unlike in 10x (where the VH and VL transcripts are sequenced separately but can be paired on the basis of cell barcoding), the VH and VL transcripts are linked into a single cDNA transcript. Paired end reads were assembled by Dr. Nicholas Grayson and I was provided with nucleotide sequences with unique identifiers.

We used the Change-O MaskPrimers.py script to assign isotypes using the constant region primer set published by DeKosky and colleagues (DeKosky et al., 2016). VH and VL sequences were paired on the basis of their identifiers using an in-house script. Nucleotide sequences were aligned to a reference IMGT germline database using the AssignGenes module of Change-O (wrapper for IgBLAST) (Gupta et al., 2015; Ye et al., 2013). Clonotypes were assigned using the DefineClones module of Change-O, grouping sequences with the same IGHV, IGHJ, IGKLV and IGKLJ gene and with greater than 90% amino acid identity across the CDRH3.

For the data used in this study, there were no unique molecular identifiers (UMIs) as this signficantly increases the cost of the method. UMIs are used for error correction in most recent BCR-sequencing applications (Friedensohn et al., 2017): to find an appropriate error correction method, we explored different redundancy filtering methods (VH:VL amino acid and nucleotides with either 100% or 90% identity, as well as clonotypes with either 100% or 90% CDRH3 amino acid identity) and their effect on:

i) the overlap observed between a pair of samples from the same subject eight weeks apart compared to the overlap seen using 10x. Sequences/clonotypes observed in two samples from the same subject are considered likely to be real.

ii) the number of different sequences/clonotypes belonging to a CBH5 hybridoma which was spiked-in to the above samples, and their overlap over time.

The same samples had libraries prepared and sequenced via this method and 10x Genomics and there were many more hybridoma variants proportionally in the present method. The variants were consistent with sequencing error (due to an increase in the number of mutations along the 3' to 5' direction of the heavy chain sequence, which corresponds to the sequencing direction) as opposed to AID mutants in the hybridoma (the transition:transversion ratio was half that reported in the literature).

iii) clonal diversity in the full dataset.

Retaining only clonotypes (same IGHV, IGKLV, IGHJ and IGLJ and >90% CDRH3 amino acid identity) that were observed more than once was found to maximise overlap in the longitudinal samples and remove 95% of likely erroneous singleton hybridoma variants. Removing singleton clonotypes as opposed to singleton sequences had a significantly lower impact on the clonal diversity retained after error correction than removing singletons at the level of nucleotide or amino acid sequences. Clonal deduplication is a standard error correction procedure in the absence of UMIs (Friedensohn et al., 2017).

This sequencing method was applied to peripheral blood samples from 16 subjects who had tested positive for SARS-CoV-2 by PCR (dated to 2020), as well as five subjects who were vaccinated with ChAdOx-1/nCoV-19 at baseline and four, five and eight weeks post-vaccination (D0, D28, D35 and D56). An additional ChAdOx-1 vaccinee was sampled at D0 and D56. Samples from a MenACWY vaccinee were sequenced at equivalent time points (D0, D28, D35 and D56). In the present analysis, all samples for a given subject were grouped together, as there was no significant clonal expansion relative to at baseline.

Between 431 and 52,316 clonotypes were obtained per sample after (on average 6,788.3) resulting in a total of 285,107 clonotypes across the 42 samples after error correction. The number of non-singleton clonotypes obtained in each sample is shown in Table 5.2.

Subject	Type	clonotypes
690	Vaccine	5802
891	Vaccine	7825
912	Vaccine	30135
1013	Vaccine	4478
1059	Vaccine	12219
1166	Vaccine	21626
1197	Control	32422
2693	COVID	19547
5072	COVID	2686
5080	COVID	3482
5082	COVID	1318
5085	COVID	4673
5086	COVID	5454
5087	COVID	2494
5088	COVID	25134
5090	COVID	3428
5097	COVID	4140
5104	COVID	4345
5106	COVID	3878
5171	COVID	5706
5419	COVID	52316
5427	COVID	31524
5478	COVID	475

Table 5.2: The number of non-singleton and therefore highest confidence clonotypes obtained for each of 23 subjects (pooled across longitudinal time points for the vaccinees given little evidence of an antigen-specific response in terms of clonal expansion or similarity to CoV-AbDab); samples used were from subjects vaccinated with ChAdOx-1/nCoV-19 (6), a subject vaccinated with MenACWY and 16 subjects who were PCR positive for COVID in 2020.

5.3.3 Structural analysis

5.3.3.1 Contact analysis

To delineate interactions in the VH:VL interface, we used Arpeggio (Jubb et al., 2017).

Arpeggio performs interaction type assignment based on atom type, distance and angular terms. The output is a feature vector for each interatomic contact in which the first five features correspond to mutually exclusive con-

tact types that describe potential steric or Van der Waals clashes and whether the atoms are covalently-linked, forming a Van der Waals interaction or simply within 4.5Å ("proximal"). The interaction is then classified as one or multiple of: hydrogen, weak hydrogen, ionic, aromatic-aromatic, hydrophobic, carbonyl, halogen or metal. Polar and weak polar bonds are also included which are hydrogen bonds without angle constraints. These are defined as in CREDO (Schreyer & Blundell, 2013).

We only considered variable domain contacts, excluding contacts derived from residues with IMGT numberings exceeding 128 (on both chains).

5.3.3.2 Energetics

There are a number of computational methods for the prediction and analysis of protein-protein interaction energetics: such methods exist on opposing sliding scales of accuracy and time intensiveness. The most accurate but time-intensive methods include free energy calculation via perturbation (FEP) or thermodynamic integration; at the other end of the scale, rapid but less accurate methods use statistical potentials or empirical force fields (e.g. Rosetta (Humphris & Kortemme, 2008; Kortemme et al., 2003; Wedemeyer & Baker, 2003) and FoldX (Guerois et al., 2002)).

To assess VH:VL energetics at scale, we used the FoldX AnalyseComplex functionality which can calculate an interaction energy for the VH:VL domain interaction based on an empirical force field (Schymkowitz et al., 2005). FoldX was among the best performing force fields for prediction of binding affinity in an antibody-antigen benchmark, though it must be noted that the correlation between predicted and observed free energy changes in this benchmark was still "very weak" with an Rr of 0.11 (Sirin et al., 2016).

As in the contact analysis, we trimmed the structures to only include VH and VL domains by excluding all residues with IMGT numbers over 128.

Term	Description
Van der Waals	Sum of experimentally-derived VdW energies of all atoms with respect to the solvent
Hydrophobic solvation	Sum of solvation energy for non-polar groups
Polar solvation	Sum of solvation energy for polar groups
Water bridge	Marginal change in free energy due to a water molecule with more than one hydrogen bond to the protein
Hydrogen bond	Free energy difference between an interface hydrogen bond and a hydrogen bond with the solvent
Electrostatics	Electrostatic contribution of charged groups
Electrostatic k_{on}	Relationship between electrostatic term and k_{on} rate
Main chain entropy	Entropic cost of fixing backbone residues in observed conformation
side chain entropy	Entropic cost of fixing side chain in observed conformation
Steric clash	Measure of steric overlap

Table 5.3: | **FoldX interaction terms.** Short summaries of each of the ten interaction terms in the calculation of the FoldX interaction energy of the VH:VL complexes, which is a $\Delta\Delta$G between the complex ΔG and the sum of ΔGs of the individual VH and VL domains.

FoldX outputs a total interaction energy, a $\Delta\Delta$G upon complex formation, which is a linear combination of ten empirical terms. These are described in Table 5.3 (Schymkowitz et al., 2005).

5.3.3.3 VH:VL orientation

VH:VL orientation is described using the six ABangle parameters which provide a complete description of VH:VL orientation (Dunbar et al., 2013). The ABangle parameters consist of five angles describing the relative orientations of five vectors that define a common reference plane among structures (two orthogonal vectors on each domain and a vector *C* between the VH and VL domains). These vectors are shown in Figure 5.1.

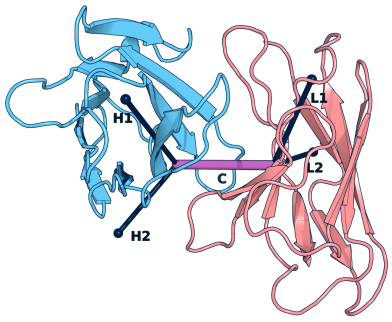

Figure 5.1: | **ABangle diagram.** Diagram of the vectors used to describe the five ABangle parameters and distance (*dc*). *HL* is the torsion angle between *H1* and *L1*. *HC1* is the bend angle between *H1* and *C*. *HC2* is the bend angle between *H2* and *C*. *LC1* and *LC2* are defined using the equivalent light chain vectors. *dc* is the length of the vector *C* in Å ngstrom. These vectors are fit through a set of structurally-conserved residues at the VH:VL interface. Image courtesy of the OPIG website.

5.3.3.4 Elbow angle

For 1,259 of the 1,323 complexes, the CH1 and CL domains were solved as well as the VH and VL. This enables calculation of the antibody elbow angle, which is the angle between the two axes relating the VL and VH domains, and the CH1 and CL domains (Stanfield et al., 2006) (Appendix Figure 4). This was calculated in pymol using the *elbow_angle.py* script written by Jared Sampson.

5.3.4 Statistical analysis

5.3.4.1 Identification of over- and underrepresented gene pairings

Previously, authors have identified over- and underrepresented gene pairings in their published single-cell datasets via a X^2 test. This test computes the probability of the observed pairing frequencies under the null hypothesis which would be random pairing (i.e. $P(H:L) = P(H)P(L)$, where H is a VH gene of choice and L a VL gene of choice and *H:L* is an observation of a heavy chain H with light chain L).

Typically, significantly over- or underrepresented pairings identified within a single subject disappear when pooling data across multiple subjects and are therefore unlikely to represent biophysical biases.

We note that subjects are reported to have their own germline gene usage preferences which would result in differing probabilities of observing germline genes across different subjects. Furthermore, given that study centres vary by primer usage, there are also likely to be study-specific biases in observed gene usage. Instead of applying multiple statistical tests on a subject-by-subject basis, we can borrow statistical power across multiple samples by modelling the probability of observing a heavy/light chain gene pairing via a negative binomial generalized linear model of the form:

$$H_i L_j = \beta_0 + \beta_1 H_i + \beta_2 L_j + \beta_3 H_i L_j + \beta_4 I_m + \beta_5 S_n + \epsilon \tag{5.1}$$

Where H_i is heavy chain gene *i*, L_j is light chain gene *j*, I_m is individual *m* and S_n is study *n*. This models the count of a particular gene pairing (e.g. IGHV4-34 and IGLV1-40) as a function of the count of the heavy gene, the count of the light gene, the subject in which the count is observed and the study, to identify pairings which are over- or underrepresented. Significance at the 5% level is reported after correcting for multiple testing with a Bonferroni correction. The model was fit in R using negative binomial linear regression function (*nbm*) from the MASS package (Ripley et al., 2022).

This statistical model was suggested by Dr. Oliver Crook.

5.3.4.2 Other statistical tests

Where the distributions appeared approximately normally distributed, a student's t-test was used to compare the distributions. Elsewhere, a permutation t-test was performed using the Python package *dabest* with 5,000 permutations (Ho et al., 2019).

Region	Heavy	Light
FWR1	4.1%	4.6%
CDR1	4.3%	6.5%
FWR2	34.0%	42.7%
CDR2	4.2%	4.4%
FWR3	9.8%	9.5%
CDR3	41.7%	27.4%
FWR4	14.3%	14.6%

Table 5.4: | **The proportion of the interface in each VH and VL region.** The proportion of interface residues in each of the regions of the VH and VL chains, averaged across the full dataset.

5.4 Results

5.4.1 The VH:VL interface is dominated by CDR3-mediated hydrophobic interactions

We used Arpeggio to annotate each of the interactions in the VH:VL interface of our human structural dataset. Each pair of interacting atoms for each residue pair is assigned one of multiple interaction types 5.3. The majority of interacting VH:VL residue pairs had hydrophobic interactions (66.9%) (Figure 5.2A). 39.2% included hydrogen or polar interactions. 12.4% of residue pairs exhibited aromatic interactions. 1.1% of interactions were ionic. The majority of interface residues are in FWR2 and CDR3 (Table 5.4). On average, the majority of residues in contact are in the CDR3s (Figure 5.2C and D): between CDRH3 and FWRL2 (19.3%), CDRH3 and CDRL3 (13.3%) and CDRL3 and FWRH2 (9.9%).

We were interested in which positions could allow us to identify structural causes for VH:VL pair over- or underrepresentation. We first looked at how often each residue pair (VH IMGT position: VL IMGT position) appeared in VH:VL interfaces. The majority of the total observed interacting positions (534 of the 607 VH:VL residue pairs) were observed in less than 5% of complexes. There were 56 residue pairings which appeared in more than 10% of complexes, and 21 residue pairings which appeared in more than 50% of complexes (Figure 5.3).

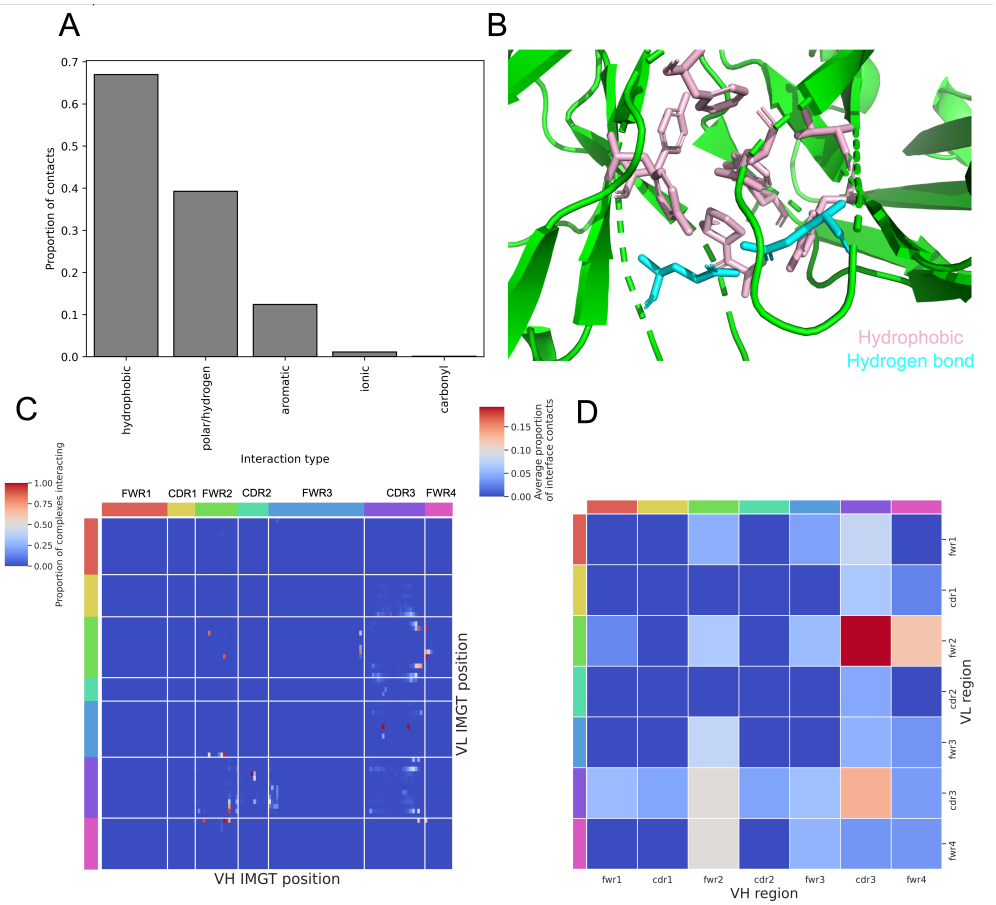

Figure 5.2: | **Interaction types in the VH:VL interface.** The majority of contacts in the VH:VL interfaces identified by Arpeggio were hydrophobic. The majority of the remainder of interactions are hydrogen or polar, and the rarest interaction types were ionic and carbonyl (A). B shows the most highly conserved (>95%) VH:VL interactions, hydrophobic in pink and the bidentate H50:L50 hydrogen bond in cyan. Heatmap (C) shows the conservation of interacting positions between the VH and VL. Heatmap (D) shows the average proportion of VH:VL interactions between each region of the VH and VL.

The majority of highly conserved residue contacts were hydrophobic interactions (Table 5.5). These positions form a conserved core of hydrophobic interactions between FWR2 (positions 42, 50 and 52) and the residues that either lie either side of the CDR3s (positions 103 and 118) or lie towards the end of the CDRH3 (115 in the VH and 116 in the VL). In addition to hydrophobic interactions, all 1,323 complexes have at least two hydrogen or polar bonds. 329/1,323 (nearly a quarter) of complexes have at least one and maximally four ionic in-

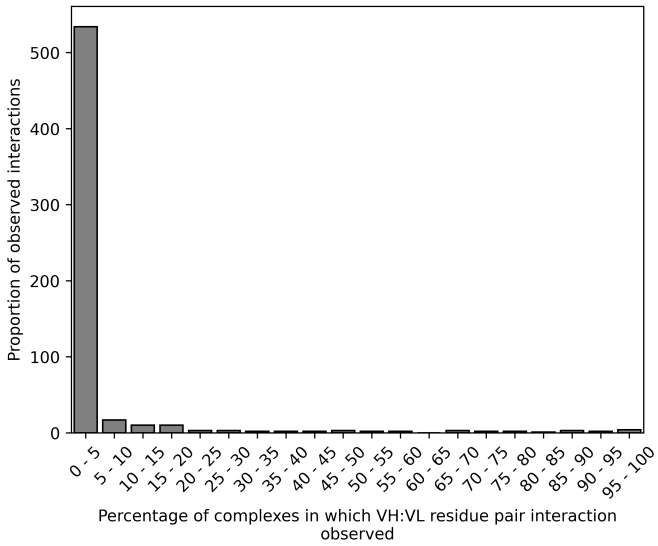

Figure 5.3: | **Commonness of VH:VL residue pairings in the interface.** The distribution of the number of complexes in which a particular VH:VL residue pairing (by IMGT position, regardless of amino acid identity at that position) are observed to interact. Most pairings (88.0%) are seen in less than 5% of complexes, meaning that there are just 73 VH:VL residue pairs that are observed to interact in more than 5% of complexes.

teractions (average 1.3). The positions of the most conserved hydrogen bonds and ionic interactions are shown in Tables 5.6 and 5.7 respectively.

The heavy chain interface residues used in more than 50% of complexes are shown in Appendix Table 1. Also shown is the proportion of complexes in the dataset in which the residue is charged. Appendix Table 2 shows equivalent measures for the VL. Nine of these positions are charged in at least 10% of complexes.

The presence of charged residues in the interface could lead to VH:VL complementarity rules, as it would be energetically unfavourable to have repulsive charges in the interface. Ionic interactions appear in around 1/4 of VH:VL interfaces, though these are not particularly conserved. The most commonly observed ionic interaction is between H107 and L116 (40/469 complexes in which H107 is charged). Furthermore, H66 (which is in the interface of 58.6% of com-

plexes) exhibits an ionic bond in 26/336 complexes in which it is a charged residue, interacting largely with L114. While ionic interactions appear to play a role in around a quarter of human VH:VL interfaces, not all charged residues in frequently interface positions are involved in ionic interactions; furthermore, the residues with repulsive charges are observed at these positions at similar rates. These ionic interactions therefore have limited predictive power in determining VH:VL complementarity. Table 5.7 shows how well the charge at one position predicts the charge at the corresponding interface position.

VH position	VL position	%	VH A.A	VL A.A	Interaction type
50	118	99.3%	L	F	Hydrophobic
118	50	99.0%	W	P	Hydrophobic
50	103	97.6%	L	Y	Hydrophobic
118	42	96.2%	W	Y	Aromatic/Hydrophobic
52	116	93.0%	W	Y	Aromatic/Hydrophobic
42	118	91.4%	V	F	Hydrophobic
115	42	88.7%	F	Y	Hydrogen/Hydrophobic
50	50	88.4%	L	P	Hydrophobic
44	44	85.7%	Q	Q	Hydrogen
103	49	83.7%	Y	A	Hydrophobic

Table 5.5: | **The ten most highly conserved VH:VL interactions.** The ten most highly conserved VH:VL interactions are observed in more than 80% of complexes and are largely hydrophobic in nature. The most common amino acid at each position is shown in the column "VH A.A" or "VL AA".

Pair	Complexes	Most frequent H	Most frequent L
44/44	85.7%	Q	Q
115/42	76.8%	F	Y
44/103	27.0%	Q	Y
113/105	24.1%	W	Q
103/44	14.6%	Y	Q

Table 5.6: | **Hydrogen bonds in the VH:VL interface.** The five most frequently observed VH:VL strong hydrogen bonds as identified by Arpeggio. The most frequently observed is between the backbone nitrogen atom of most frequently phenylalanine at position 115 with the hydroxyl group of tyrosine at position 42.

Pair	Complexes	Opposite charge	Same charge	VH accuracy	VL accuracy
107/116	40	64	21	12.0%	38.5%
108/68	34	61	22	17.7%	24.9%
66/114	21	39	7	12.8%	30.7%
113/107	17	21	11	7.8%	13.8%
69/115	16	34	31	2.0%	10.9%

Table 5.7: | **Ionic interactions in the VH:VL interface.** The five most frequently observed VH:VL ionic interactions, seen maximally across 40 complexes (H107 and L116). Not all complexes with opposite charges at these positions have an ionic interaction (minimally 47.1% of complexes), and a similar number of complexes have amino acids at these positions of the same charge. Accuracy is reported for the prediction that if the heavy chain residue is charged, the light chain residue is of opposite charge ("VH accuracy") and vice-versa ("VL accuracy").

5.4.2 Analysis of interfaces on the basis of genetic origin suggests positions that may be differentially involved

Following the identification of the positions and physicochemical nature of interacting pairs in human VH:VL interfaces, we looked at how the involvement of these positions in the interface relates to the germline genes encoding the VH and VL chains. As IGJ gene assignment is unreliable on amino acid sequences and further, no publicly available tool provides IGJ alignments to validate whether CDRH3 positions are IGJ-encoded, we only looked for correlations with V genes.

53 IGHV genes and 55 IGKLV genes are represented in the human structural dataset, in 541 combinations. The dataset is skewed towards common anti-viral IGHV genes (IGHV1-69 (10.3%), IGHV3-23 (10.1%) and IGHV1-2 (9.0%)). The light chain gene usage is also skewed, towards IGKV1-39 (12.1%) and IGKV3-20 (12.0%). The kappa to lambda ratio in the dataset was 2.2, versus on average 1.5 reported in the literature in serum and mature B cells, indicating a skew towards kappa antibodies that is also found in therapeutic antibodies (personal communications with Matthew Raybould). A χ^2 statistic of 4590.6 and associated p-value of 7.1 x 10^{-9} support non-random pairing within these skewed gene usages.

We first looked at amino acid profiles stratified by IGHV subgroup usage (Figure 5.4) and kappa/lambda locus split (Figure 5.5) for positions which were identified as interface for more than 50% of complexes. The sequence logos, colored by physicochemical residue bin, show that position 66 is positively charged in most IGHV2 and IGHV5 sequences, negatively charged in most IGHV6 sequences and more physicochemically variable across IGHV1 and IGHV3. Position 66 most commonly interacts with position 114, which is more frequently charged in lambda than kappa sequences. This could create a bias towards kappa or lambda antibodies for IGHV genes in which position 66 is charged.

In addition to amino acid identity, we analysed interface inclusion of each position with respect to IGHV and IGKLV genes and K/L loci. We did this by calculating a mutual information score for each position with respect to a binary interface classification and its IGHV, IGKLV and K/L locus. High mutual information score would indicate that a particular interface position was restricted in its appearance to particular IGV genes or K/L loci. This can occur by chance, so the absolute adjusted mutual information score is compared to the distribution of mutual information scores obtained by 1,000 random permutations of the gene label. Mutual information scores exceeding the 95th percentile of this random distribution are regarded as significant. We excluded positions seen in less than 50% of the sequences, which consisted of framework or CDR1 insertions. We also excluded sequences beyond IMGT position 107, which is the maximal IMGT position which is IGHV-encoded in the dataset.

For heavy chains, this resulted in 13 positions with significant MI scores, i.e. where the interface position appears to show some correlation with IGHV gene. Across the light chain, there were 13 positions which were significant with respect to IGKLV gene, and 8 which were significant with respect to the kappa or lambda locus. These positions and their associated frequency in IGHV and IGKLV gene subgroups are shown in Figure 5.6.



Figure 5.4: | **VH interface positions by IGHV subgroup.** Sequence logos of the 13 heavy chain positions which are in the interface for more than 50% of complexes, stratified by IGHV gene group, for the human structural dataset.

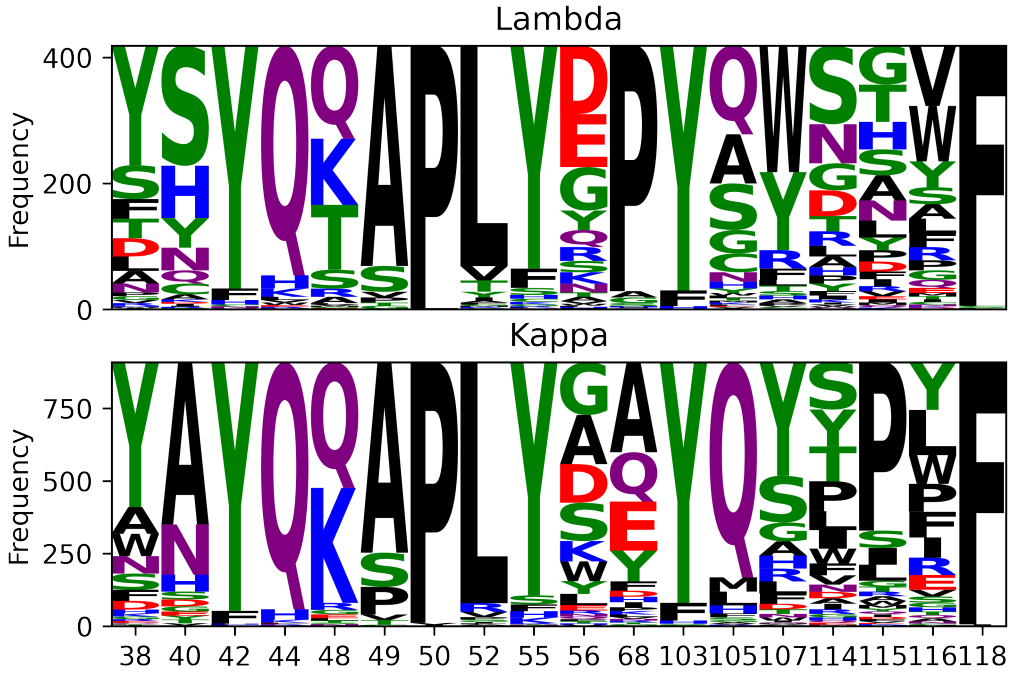

Figure 5.5: | **VL interface positions by kappa or lambda locus.** Sequence logos of the 15 light chain positions which are in the interface for more than 50% of complexes, stratified by kappa or lambda light chain class.

5.4.3 Kappa and lambda antibodies differ significantly in some interface properties

The contacting residues in the CDR3s, in particular involving CDRH3, are the most variable interface contacts among complexes (Figure 5.8). Given differences in length and physicochemical nature of kappa and lambda CDRL3s which have been described elsewhere (DeKosky et al., 2016; Townsend et al., 2016), we examined the interfaces of kappa and lambda antibodies for differences.

In the structural dataset, we observe the previously described tendancy for lambda antibodies to have longer CDRL3s (Figure 5.9A). We also observed a previously undescribed tendancy to pair with VH chains with longer CDRH3 sequences (Figure 5.9B). Kappa light chains pair with heavy chains with CDRH3s with an average length of 15.9 amino acids compared to lambda with an average of 17.4 amino acids. Sequence logos for both CDRH3s and CDRL3s are shown

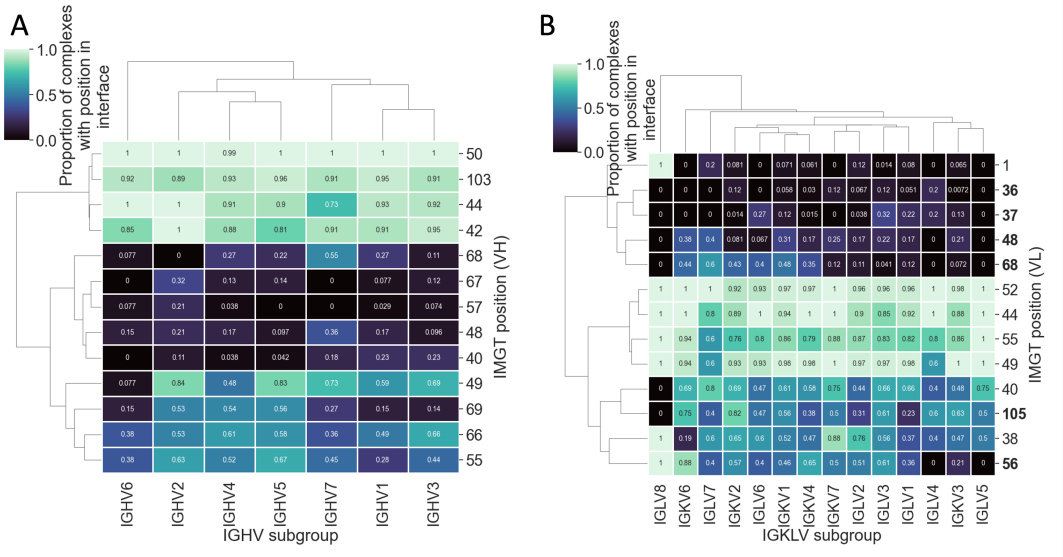

Figure 5.6: | **Positions whose inclusion in the interface depends on encoding gene.** There were thirteen heavy chain (A) and thirteen light chain (B) positions which had significant mutual information scores with respect to interface inclusion and IGHV (A) or IGKLV (B) gene, i.e. differed signficantly across IGHV and IGKLV genes. The proportion of complexes in which the position is in the interface for each IGHV (A) or IGKLV (B) subgroup are shown. In plot B, the y-labels corresponding to light chain positions that were differentially in the interface in kappa vs. lambda antibodies according to mutual information are bolded.

in Figure 5.9C and D respectively. Physicochemical differences in the CDRL3s of kappa and lambda antibodies have been previously described (Townsend et al., 2016). These length and physicochemical differences were hypothesised to lead to different interface properties. Indeed, we find significant differences in the contact profiles, predicted energetics and VH:VL orientation between kappa and lambda antibodies.

The average number of heavy chain interface residues is greater for lambda antibodies (estimated difference of 0.5 amino acids, p = 0.001, permutation t-test) while the average number of light chain interface residues is greater for kappa antibodies (estimated difference of 0.3 amino acids, p = 0.03, permutation t-test). The number of total interface residues is not significantly different however (p = 0.26, permutation t-test).

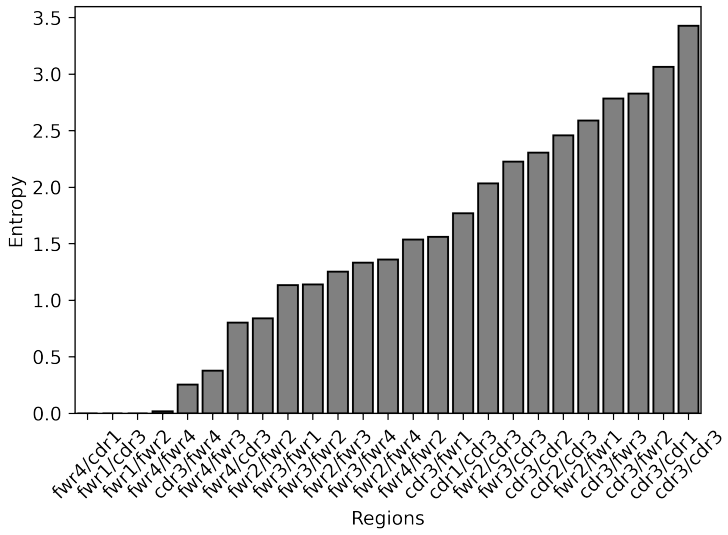

Figure 5.7: | **Entropy of VH:VL positions in interface.** Shannon entropy in the usage of particular VH:VL IMGT positions in the interface. The most variable types of contact (i.e. most variable between different complexes) are between the CDRH3 and CDRL3; there are 138 unique VH:VL residue pairings observed to interact, of which 15.2% are only observed in a single complex. This makes sense given the greater structural and sequence diversity of the CDRH3 versus the remainder of the antibody.

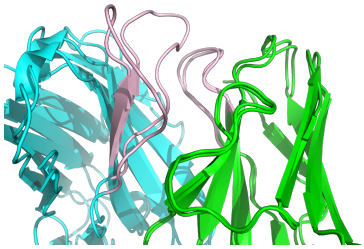

Figure 5.8: | **Visualisation of two antibodies with opposite degrees of CDRH3/CDRL3 contact.** Visualisation of two antibodies, one in which there were no heavy atom contacts between the CDRH3 and CDRL3 in the interface (1axs_HL) and a complex in which the CDRH3 makes the most extensive contact with the CDRL3 itself (6x4t_BE). The CDRH3 and CDRL3 are coloured in light pink, the VH in green and the VL in cyan.

Despite shorter CDRH3s and CDRL3s, proportionally more of the interface is occupied by the CDRH3 and CDRL3 in kappa than lambda antibodies; on average, 13.4% of the contacts in the VH:VL interface of kappa antibodies are between CDRH3 and CDRL3, versus 11.8% in lambda antibodies. The distri-

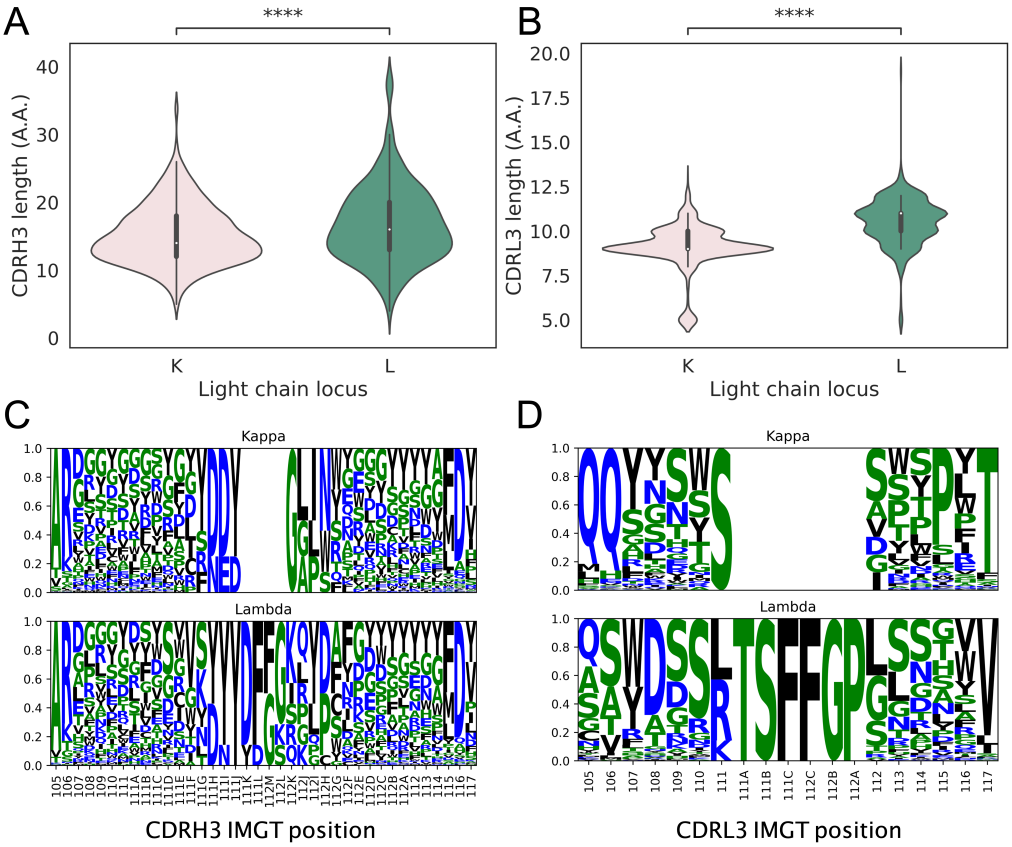

Figure 5.9: | **Length distributions of lambda and kappa-paired CDRH3s and lambda and kappa CDRL3s.** CDRH3s (A) and CDRL3s (B). Lambda antibodies have significantly longer CDRH3s (17.4 vs. 15.9 amino acids) and CDRL3s (average 10.5 vs. 8.9 amino acids). Sequence logos, colored by hydrophobicity, are shown for kappa and lambda CDRH3s (C) and CDRL3s (D). Hydrophobic residues are colored in black, while hydrophilic are coloured in blue and polar in green.

butions of contact profiles in terms of proportions of contacts between regions are shown in Figure 5.10A. Using these inter-region contact profiles as feature vectors, a random forest can be used to classify antibodies as kappa or lambda with on average 85% accuracy (five-fold cross-validation). Figures 5.10B and C show the mean difference between the proportion of VH:VL contacts in each of the CDR regions, subtracting kappa from lambda, with the confidence intervals for this difference derived from bootstrap resampling, across the VH and VL respectively. It can be seen that a significantly greater proportion of the interface is found in FWRH2 in lambda than kappa antibodies, and on average

less of the interface is contributed by the CDRH3, though the 95% confidence interval spans zero (Figure 5.10B). Across the VL, more lambda contacts are found in the CDRL1, FWRL2 and CDRL2, and proportionally fewer are found in FWRL3 and CDRL3 (Figure 5.10C).

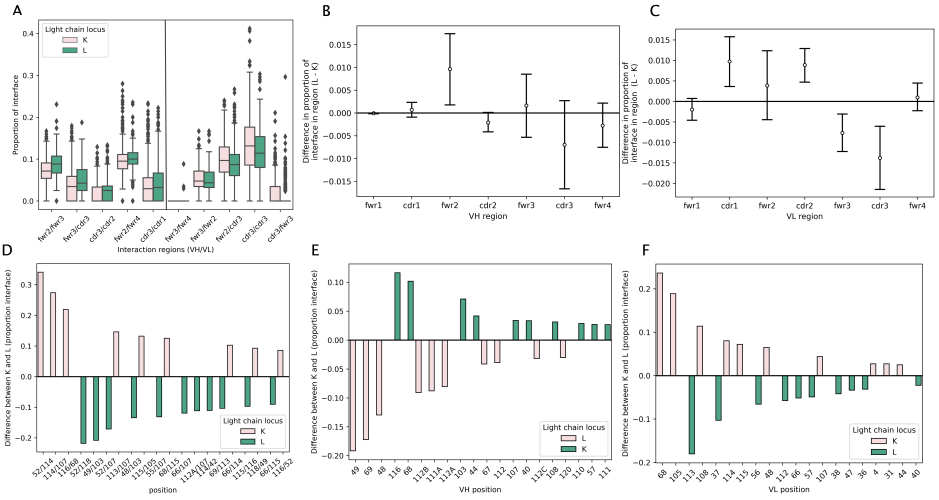

Figure 5.10: | **Difference in VH:VL interface contacts according to kappa or lambda locus.** Figure A, sorted along the x axis by the difference in the mean proportions observed in kappa vs. lambda antibodies, shows the distribution of proportions of contacts between each region (VH:VL). On the left-hand side of the solid black line, the kappa mean is higher than the lambda mean, and vice-versa on the RHS. Figures B and C show the difference in the proportion of contacts in each region on the VH and VL respectively with error bars indicating the confidence interval estimated by bootstrapping. Figure D shows the twenty VH:VL positions that are most differentially used in the interface by kappa and lambda antibodies, while figure E and F show this with regards to individual VH and VL positions.

The twenty VH:VL residue pairs, VH residues and VL residues with the greatest absolute difference in interface frequency between kappa and lambda are shown in Figures 5.10D, E and F respectively. For example, H52 and L114 interact in over 30% fewer lambda than kappa complexes proportionally. In addition to the position of the contacts differing among kappa and lambda chains on average, we also report proportionally more ionic VH:VL interface interactions in lambda than kappa antibodies ($p = 0.012$, permutation t-test).

We next looked at the possible effects of these different contact profiles on interface energetics using FoldX (Schymkowitz et al., 2005). Excluding complexes in which FoldX detected VH:VL steric clashes (261/1,323), we found that FoldX predicted total interaction energies were more negative (i.e. predicted greater $\Delta\Delta G$ upon complex formation) for kappa than lambda antibodies with a mean $\Delta\Delta G$ difference of 0.93 (95% confidence interval of 0.38 to 1.49, p = 0.001 under a permutation t-test). Delving into the specific energetic terms responsible for this difference, a number of the terms used in the calculation of the interface energy significantly differed between kappa and lambda antibodies (significant at the 5% level with permutation t-tests and a Bonferroni correction for multiple testing). The contribution of Van der Waals interactions and electrostatics (measured by an electrostatics term and an empirical equation shown to relate to k_{on} rate) were significantly greater in lambda antibodies, while in total the main chain entropy and polar solvation contributions were greater for kappa antibodies. There were no significant differences in hydrogen bonding or hydrophobic solvation (p = 0.87 and 0.68 for backbone and side-chain mediated hydrogen bonding respectively; p = 0.24 for hydrophobic solvation after multiple testing correction). These mean differences are shown in Figure 5.11. Total interaction energy was not significantly correlated with CDRH3 length (Pearson's correlation coefficient of -0.03, p = 0.21) indicating that this is not the primary driver of the difference in interface energy. Interaction energy was inversely correlated with CDRL3 length (Pearson's correlation coefficient of -0.13, $p < 0.01$).

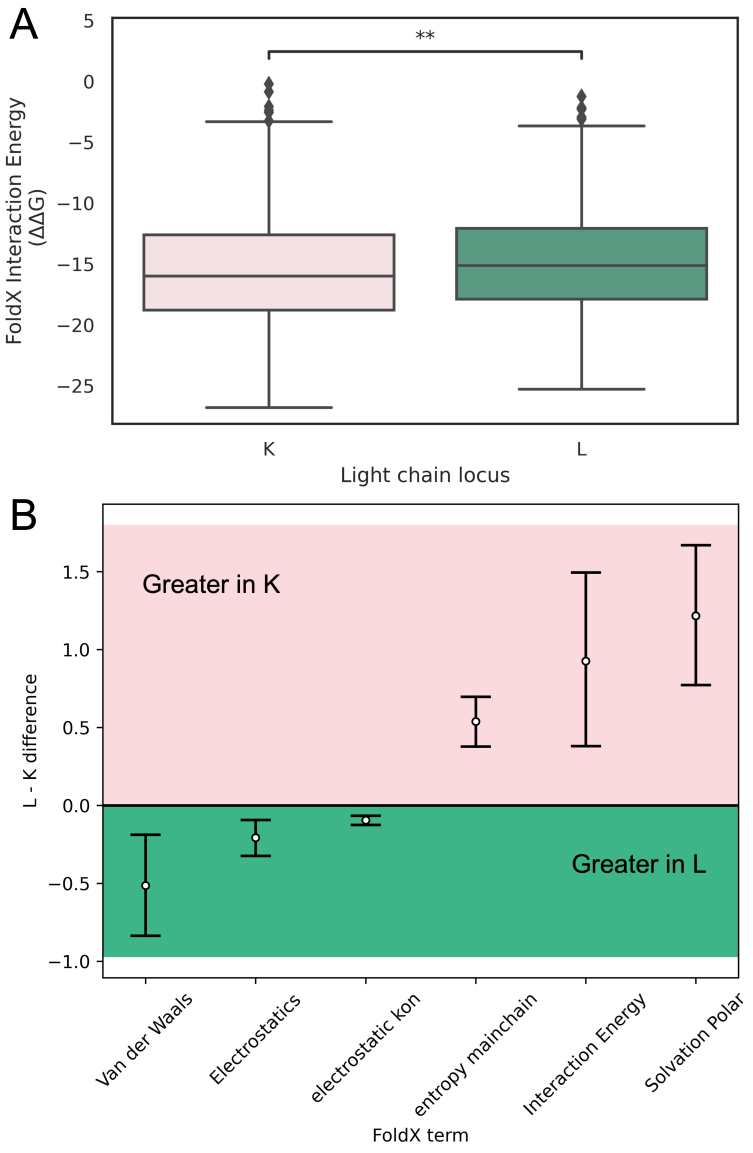

Figure 5.11: | **FoldX interaction energy differs between kappa and lambda antibodies.** Total FoldX interaction energy was calculated using the AnalyseComplex functionality of FoldX. The total FoldX interaction energy differs signficantly between kappa and lambda antibodies (A). The total interaction energy is a linear combination of ten energetic terms. Six differed signficantly, indicating that there is a greater contribution to VH:VL interface binding from electrostatic and VdW interactions in lambda antibodies, and a greater contribution of polar solvation energy and backbone entropy changes to kappa antibodies.

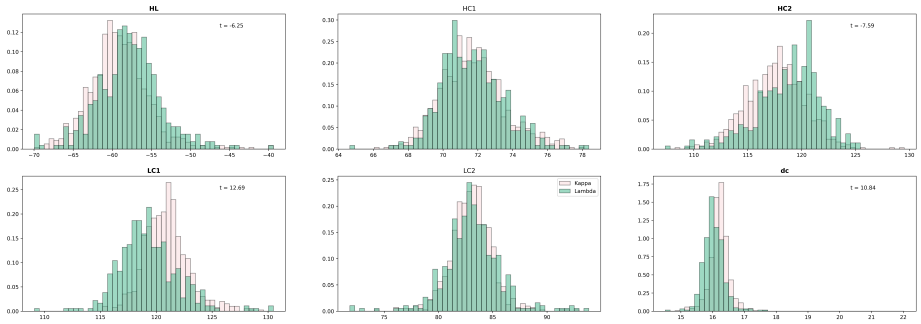

Figure 5.12: | **VH:VL orientation differs between kappa and lambda antibodies.** The six ABangle parameters as described by Dunbar and colleagues as an alignment-free method for the description of VH:VL orientation (Dunbar et al., 2013). Three of these six parameters differ significantly between kappa and lambda antibodies. These are HL, HC2, LC1 and dc.

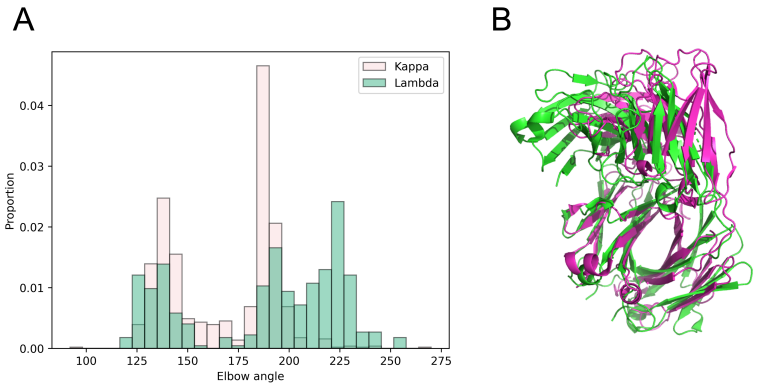

Figure 5.13: | **Difference in elbow angle distribution for kappa and lambda antibodies.** Elbow angles between the VH:VL and CH1/CL domains were calculated for 1,194 (/1,323) complexes with CH1 and CL domains solved. Kappa and lambda antibodies follow different distributions with more density at larger elbow angles (A). The pair of antibodies with the largest difference in elbow angle are shown (green, 6cr1_HL: pink, 6mty_HL).

Kappa and lambda antibodies differed further in terms of their VH:VL orientation as described by the ABangle parameters (Figure 5.12), and elbow angle (Figure 5.13). Kappa antibodies have significantly smaller *HL* and HC2 angles, and signicantly greater *LC1* angles, as well as significantly greater *dc* distances (Student's t-test). The original ABangle publication reported the critical role of L50 (L44 in the Chothia numbering scheme of the original publication) in determining orientation, distinguishing two clusters in VH:VL orientation space distinguishable on the basis of the presence of proline or phenylalanine at this

position. 99.0% of the human antibodies in our dataset have a proline at L50, so the observed kappa/lambda distinction is not attributable to this.

While this does not relate to VH:VL energetics but rather the orientation of the Fv relative to the Fc, we also analysed elbow angles, as previous work on a limited set of antibodies (365 complexes versus our dataset of 1,323) reported that lambda antibodies adopt a wider range of elbow angles than kappa antibodies and that a greater proportion adopt larger elbow angles (Stanfield et al., 2006). We confirmed this result in our up-to-date set of human antibodies (Figure 5.13).

This work demonstrates that there is a clear distinction in the VH:VL interfaces of kappa and lambda antibodies with regards to which positions are most frequently involved and the VH:VL orientation (larger distances between VH:VL domains in lambda antibodies and differences in a number of angles describing torsion and tilt). We also noted that lambda light chains were paired with heavy chains that had longer CDRH3s on average than the heavy chains paired with kappa light chains. These differences in the interfaces appear to lead to differing energetics with significantly lower total interaction energies for lambda antibodies (i.e. less energetically favourable interfaces).

5.4.4 A number of gene pairings are over- or underrepresented, but we are unable to correlate with this with predicted interface energetics

Our structural analyses indicated that there are some interface positions which share significant mutual information with the encoding V gene. The clearest statistical differences can be observed between kappa and lambda antibodies. However, our dataset of 1,323 human antibody structures is limited both in terms of statistical power and in terms of representativeness of natural repertoires. Single-cell sequencing of native VH:VL pairs provides information on many more VH:VL pairings.

We analysed a dataset of 88,274 VH:VL sequences from the Observed Antibody Space from five studies (Human Paired OAS data) to identify over- or underrepresented gene pairings. As the majority of these sequences were from antigen-experienced repertoires, we deduplicated by clonotype to prevent overrepresentation of certain pairings due to clonal expansion resulting in 85,111 clonotypes.

We first checked with a simple X^2 whether there were any well expressed IGHV or IGKLV gene pairings that are never observed together: the greatest expectation value of a non-observed pairing was 12.8 instances corresponding to IGHV3-49 and IGKV1-8. This expectation value is small given the dataset size of 85,111.

The X^2 test is often performed on a subject-by-subject basis, which does not allow us to borrow statistical power between different samples and studies. Furthermore, we are interested in potential biophysical biases as opposed to individual-specific pairing preferences. We used a negative binomial linear regression to identify interactions between a) IGHV and IGKLV genes, b) IGHV and IGKLJ genes, c) IGHJ and IGKLV genes and d) IGHV and K vs. L genes. Significant interactions, after correcting for multiple testing via Bonferroni correction, were identified in all four models. The number of significant interactions is outlined in Table 5.8. The identified genes are shown in Table 5.9. As expected, the sign of the effect size is consistent with the sign of the (*observed - expected*) count calculated under a X^2 distribution.

There were significant interactions shown for IGHV3-72 genes, with an enrichment of IGHV3-72 heavy chains paired with light chains encoded for by IGKJ4 and IGKJ5. IGHV3-72 showed the only significant kappa/lambda bias (i.e. preference for kappa chains as opposed to lambda chains) in the dataset. Given this, we examined our structural dataset to see if there is anything in particular about IGHV3-72 structures that could explain a preference for pairing

with kappa over lambda light chains.

There are 39 IGHV genes which have structures in complex with both kappa and lambda light chains (corresponding to 1,283 structures, with maximally 137 IGHV1-69 structures). There were eight IGHV3-72 complexes, seven of which were solved in complex with kappa antibodies. While the average kappa antibody FoldX interaction energy was marginally greater than the lambda antibody interaction energy, the lambda interaction energy was within the range observed for IGHV3-72/IGK interactions. A single lambda complex is insufficient to make claims about the interface energetics of kappa/lambda antibodies. IGHV3-72 in the context of other IGHV genes is shown in Appendix Figure. To explore further, either greater structural data or modelled data of IGHV3-72 in complex with kappa and lambda chains would be required.

Type	Number total	Number under	Number over
IGHV:IGKLV	3,318	11	0
IGHV:IGLJ	439	0	3
IGKLV:IGHJ	340	2	0
IGHV:KL	98	1	0

Table 5.8: | **The number of significant over- or underrepresentations of gene pairings.** The number of observed pairings in the Observed Antibody Space and the corresponding number of pairings which were identified as significantly over- or underrepresented.

5.4.5 Crystal structures of human lambda antibodies are biased towards longer CDRH3s in comparison to sequencing datasets

In Section 5.4.3, we described our finding that lambda chains were disproportionately paired with heavy chains with longer CDRH3s in the human structural dataset. The average CDRH3 length of all natively paired antibodies in the Observed Antibody Space is 15.7 amino acids for both kappa and lambda chains. This contrasts with a difference of 1.5 amino acids observed in the structural dataset. On a sample-by-sample basis, the maximum observed length difference was lambda CDRH3s being longer by 0.7 amino acids - in another sam-

Model	H	L	Expected	Observed	Effect
IGHV:IGKLV	IGHV3-66	IGLV1-47	17.6	6	-2.63
IGHV:IGKLV	IGHV2-26	IGKV3-15	36.4	17	-2.24
IGHV:IGKLV	IGHV4-39	IGLV7-46	33.4	15	-2.18
IGHV:IGKLV	IGHV3-66	IGLV1-44	24.0	15	-2.05
IGHV:IGKLV	IGHV2-26	IGKV3-11	33.0	19	-2.04
IGHV:IGKLV	IGHV4-59	IGKV2-30	61.6	20	-2.01
IGHV:IGKLV	IGHV4-59	IGLV7-46	26.8	21	-2.01
IGHV:IGKLV	IGHV2-5	IGKV2-39	30.0	17	-1.99
IGHV:IGKLV	IGHV3-23	IGLV7-46	59.6	47	-1.88
IGHV:IGKLV	IGHV4-39	IGKV2-30	76.8	30	-1.84
IGHV:IGKLV	IGHV2-5	IGKV3-11	92.7	52	0.05
IGHV:IGKLJ	IGHV3-72	IGKJ4	55.9	157	1.7
IGHV:IGKLJ	IGHV3-72	IGKJ5	22.5	42	1.38
IGHV:IGKLJ	IGHV7-4-1	IGLJ1	49.0	80	1.36
IGHJ:IGKLV	IGHJ5	IGKV1-5	532.7	332	-1.45
IGHJ:IGKLV	IGJH6	IGKV4-1	1081.1	595	-1.37
IGHV:KL	IGHV3-72	λ	173.3	84	-1.03

Table 5.9: | **The significant interaction terms in each of the four gene-wise interaction models.** The significant interactions are shown with their expectation and observed values calculated over the entire dataset as well as the coefficient of the interaction in the model. This includes accounting for multiple testing.

ple corresponding to the only non-disease-associated dataset, kappa antibodies have CDRH3 lengths 0.5 amino acids longer on average. The total CDRH3 length distribution of the human structural dataset is shown in Figure 5.14.

This lack of the CDRH3 length bias observed in the paired sequencing dataset suggests that interface energetics are non-representative. Any subsequent interface energetics analysis should control for this. We also found that in some databases of antibodies which are confirmed as being able to be expressible, such as CoV-AbDab and the IEDB, longer CDRH3s tend to pair with lambda chains, while in EBOV-AbDab the opposite was true (Appendix Figure 5).

5.4.6 VH:VL pairing in a novel single-cell sequencing dataset

In addition to the paired sequencing dataset in the Observed Antibody Space, we also processed and analysed a large single-cell sequencing dataset generated by the labs of Professor Lynn Dustin and Professor John Frater referred to

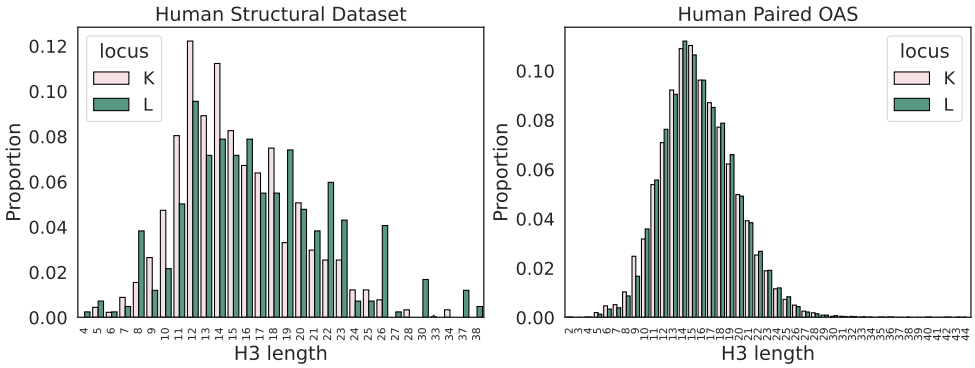

Figure 5.14: | **CDRH3 length distribution stratified by light chain locus in a human structural dataset and the Observed Antibody Space.** CDRH3 length distributions of kappa and lambda-paired heavy chains in the human structural dataset and Human Paired OAS: the length difference observed for lambda antibodies in the human structural dataset disappears in repertoire sequencing data.

as the Frater/Dustin single-cell sequencing dataset.

This alternative method provides another window into VH:VL pairing. The dataset differed most obviously in its lambda bias: between 46.0% and 90.0% of clonotypes recovered were lambda (with an average of 73.2%, a κ:λ ratio of 0.4), versus between 31.7% to 63.2% of clonotypes in the 10x data from the Observed Antibody Space (average of 41.2%, a κ:λ ratio of 1.5). Despite the relative paucity of kappa data, there were still 88,650 kappa clonotypes alone in the dataset which is greater than the total number of paired clonotypes in the Observed Antibody Space.

We found that for the majority of IGHV:IGKLV, IGHV:IGKLJ and IGHJ:IGKLV gene pairings that were identified as significantly under- or overrepresented in the Observed Antibody Space data, the corresponding frequencies in this separate dataset were also either lower or higher than expected under a null distribution of random pairing. Interestingly, IGHV3-72 is the IGHV gene with the highest score with regards to kappa/lambda pairing (with the same kappa preference observed in the OAS datasets). We modelled the IGHV:KL frequencies using the same negative binomial regression model as before but with-

out the "study" term (given that all the data was produced at the same study centre) and found that IGHV3-72 again displayed a significant preference for kappa vs. lambda light chains. There was also a significant interaction in the IGHV:IGKLJ model for IGHV3-72 and IGKJ4, as per the previous model (in addition, IGHV3-20 and IGKJ4 were overrepresented, and IGHV3-30-5 and IGLJ2). There were no significant interactions between IGHJ and IGKLV genes in the present dataset.

H	L	Expected	Observed	OAS	Dustin
IGHV3-66	IGLV1-47	39.4	33	Under	Under
IGHV2-26	IGKV3-15	2.2	0	Under	Under
IGHV4-39	IGLV7-46	280.0	247	Under	Under
IGHV3-66	IGLV1-44	53.7	44	Under	Under
IGHV2-26	IGKV3-11	4.2	3	Under	Under
IGHV4-59	IGKV2-30	9.7	7	Under	Under
IGHV4-59	IGLV7-46	535.7	447	Under	Under
IGHV2-5	IGKV2-39	NA	NA	Under	NA
IGHV3-23	IGLV7-46	142.6	116	Under	Under
IGHV4-39	IGKV2-30	5.1	1	Under	Under
IGHV2-5	IGKV3-11	14.0	15.0	Under	Over
IGHV3-72	IGKJ4	63.7	232	Over	Over
IGHV3-72	IGKJ5	19.2	46	Over	Over
IGHV7-4-1	IGLJ1	135.4	84	Over	Under
IGHJ5	IGKV1-5	1794.2	1510	Under	Under
IGHJ6	IGKV4-1	514.0	350	Under	Under
IGHV3-72	λ	609.8	334	Under	Under

Table 5.10: | **Significant gene pairings and their associated expectation values.** 17 gene pairings were identified as signficantly over- or underrepresented in the Observed Antibody Space, accounting for any subject- or study-specific biases. We compared the expected and observed counts of these genes in this novel single-cell sequencing dataset, and found that for the majority of predicted over- or underrepresentated pairings, the gene pairing is observed correspondingly in this data. Most interestingly, IGHV3-72 shows a significant preference away from lambda.

5.5 Discussion

The aim of this chapter was to characterise VH:VL interfaces in a human structural dataset, and identify possible correlates in the paired data available in the Observed Antibody Space and a novel dataset produced by the labs of Professor John Frater and Professor Lynn Dustin.

In the first part of this work, we characterised VH:VL interfaces of human antibody crystal structures. This revealed a conserved core of largely hydrophobic interactions, with a large number of interface contacts mediated by the CDRH3 or CDRL3. We reported a number of positions which appear to differ significantly in their interface frequency in antibodies encoded by different IGHV and IGKLV genes suggesting that there may be germline-encoded interaction patterns. The clearest difference (due to the large number of structures available for both classes) was between kappa and lambda antibodies. We found that they differed significantly in their contact profiles, predicted interface energetics and VH:VL orientation.

We then identified significantly over- and underrepresented IGHV:IGKLV, IGHV:IGKLJ, IGHJ:IGKLV and IGHV:KL gene pairings from the nearly 90,000 paired sequences available in the Observed Antibody Space. This constitutes the most extensive analysis of its kind. We focussed on the only IGHV:KL gene pairing that was significant, IGHV3-72 which displayed a preference for kappa antibodies, and referred to our structural dataset to look for evidence of a structural correlate. Unfortunately, there was only a single IGHV3-72 lambda antibody in the dataset, and this showed only a marginally lower VH:VL interaction energy than the kappa antibodies in the dataset. However, we also found that this IGHV3-72 displayed a kappa preference in a dataset produced by a different method, the Frater/Dustin single-cell sequencing dataset, which itself was lambda-biased, suggesting that this may be a real preference.

It should be noted that the Frater/Dustin single-cell sequencing dataset is an unpublished method, for which the error rate of pairing has been significantly less well characterised than the 10x platform which was used to produce all of the datasets in the Observed Antibody Space. Analysis of pairing accuracy on hybridoma spike-ins, measured as the percentage of VH hybridoma reads paired with the appropriate VL clonotype, suggests accuracy between 62.6% and 99.5%; on average, 23.5% of VH clonotypes are paired with more than one

IGKLV gene. For comparison, analysis of the largest dataset of 10x data yet released suggested that 9.3% of cells had more than one light chain associated with them (Jaffe et al., 2022). This data also has a non-physiological lambda bias. This should be kept in mind in any future analysis of the present datasets produced by the method.

The work was perhaps ambitious in attempting to identify what would likely be a subtle biophysical bias among a number of significantly stronger biases, both in the data and in the biology underlying VH:VL pairing itself. Individuals have their own IGHV and IGLV gene usage preferences (Briney et al., 2019) and V(D)J recombination models (Slabodkin et al., 2021). There is furthermore significant allelic diversity, particularly in IGHV1-69 and IGHV3-30 (Corcoran et al., 2016; Gadala-Maria et al., 2019; Gidoni et al., 2019) which the current work does not consider. Furthermore, sequences mutate from germline (surprisingly around 3/8 of the Paired OAS data was completely germline) meaning that any initial liabilities in pairing can be rescued. Finally, CH1/CL energetics will contribute to the pairing propensity of different heavy and light chains. While there is a single IGKC gene, there are multiple IGLC genes each of which are paired with IGLJ genes in tandem cassettes on the lambda locus, which means that IGLJ choice will covary with IGLC choice. There is also the hypothesis that the lambda locus, which recombines after the kappa, exists to rescue autoreactive heavy chains through receptor editing and this shapes the observed light chain repertoire (Collins & Watson, 2018). Our results about orientation could be interesting here, as the differing orientations between kappa and lambda chains may result in different parts of the heavy chain becoming surface-exposed. The significant number of stronger biological correlates means it will likely always be difficult to predict pairings of antibodies based solely on their genetic origin.

Another issue with using this gene-based system for looking at VH:VL pairing is that the ability to correlate the structural and sequencing data is limited by the lack of solved structures for any specific particular gene pairing: as the likelihood of observing a pairing in the sequencing data declines, as does the probability that such a pairing exists in the structural dataset. We originally

envisaged that this problem could be solved by structural modelling, allowing us to model unobserved pairings and examine the effect on the quality of the model and the predicted interfaces. We found however that models produced by both ABlooper and IgFold were deemed energetically unfavourable by FoldX in the VH:VL interface, with evidence that interface side chain minimisation can only ameliorate this in around 50% of complexes (Appendix Figure 7). Improvements in modelling tools will likely be necessary to perform these experiments at high throughput.

The natural next step would be to dispense with the information about genetic origin altogether and focus on covariation in amino acid usage. One of the most powerful ways to do this would be to use a language model: language models learn the syntax of a language via masking of regions in a sentence or sequence and subsequent prediction of masked regions (Vaswani et al., 2017; Wolf et al., 2020). Language models have been applied to amino acid sequences, and the embeddings that the models learn have been used as features for secondary structure prediction, contact prediction and identifying remote structural homologs (Alley et al., 2019; Lin et al., 2022; Rao et al., 2019; Rao et al., 2020). Antibody-specific models have been trained, mostly on heavy chain data, and have been used to reconstruct missing sequence regions (Olsen, Moal, et al., 2022), predict paratopes (Leem et al., 2022), and provide insight into antibody evolution (Ruffolo et al., 2021). These models are interpretable through their attention layers, which tell us how important one feature is for the prediction of another (Vig et al., 2021). Current antibody language models have been trained on just heavy chain sequences (Leem et al., 2022; Ruffolo et al., 2022) or heavy chain and light chain sequences separately (Olsen, Moal, et al., 2022). A language model could be trained in which the masked region for prediction is on the opposing chain, and attention matrices could be analysed to decipher the language that relates VH and VLs. We hope that the positions identified here as common in the interface but varying according to the genetic origin of the antibody in question, can be related to these attention matrices.

The current structural analysis contains a lot of uncertainty due to the large

number of possible covariates. Furthermore, we have only considered static snapshots of the VH:VL interface: VH:VL orientation appears to be flexible and affected by CDR conformation in molecular dynamics simulations (Fernández-Quintero, Hoerschinger, et al., 2020; Fernández-Quintero, Pomarici, et al., 2020), and furthermore has been observed to change significantly upon antigen binding (Stanfield et al., 1993). Changes in the conformations of CDRs when paired with different light chains have been observed, particularly in the CDRH3 (Teplyakov et al., 2016), and this will clearly affect the interface. The canonical forms observed for the kappa and lambda antibodies in the human structural dataset are reported in the Appendices. Furthermore, while we did not note significantly higher RMSD between length-matched CDRH3s compared across chain types (Appendix Figure 13), molecular dynamics experiments suggest that CDRH3 packing in the interface is significantly affected by pairing with a kappa or lambda light chain (personal communications with Bora Guloglu). It would be interesting to see how the VH:VL interface contacts themselves change as a function of CDR conformations, and whether this would be a more informative explanatory scheme for VH:VL pairing preferences than germline origin.

In terms of generating predictive models of VH:VL pairing in the future, a fundamental limitation moving forward is the absence of public negative data. While a model could be trained to correctly re-pair sequences from large paired sequencing datasets such as the novel dataset presented here, it is not known whether the "false positives" that the model would learn to discriminate are truly negatives, or if they simply have not been sequenced yet given the "needle in a haystack" nature of BCR sequencing. While failure of particular VH:VL antibody pairs to be recombinantly expressed is a common phenomenon, there are no public datasets containing this information and this is generally not considered of interest. Some correlation with biophysical properties of therapeutic antibodies, as reported by Jain and colleagues, could be an interesting route, however only offers 137 examples (Jain et al., 2017).

The data collected here about VH:VL pairing is not only useful in terms of our basic understanding of BCR repertoires, but also in terms of bispecific en-

gineering. For heterotetrameric bispecifics (two different heavy chains and two different light chains), light chain mispairing remains a problem. One means of solving this is redesign of the separate VH:VL/CH1:CL interfaces to promote association of the correct heavy and light chain pairs (Klein et al., 2012; Krah et al., 2017). Lewis and colleagues used computational design to introduce mutations in both the VH:VL and CH1:CL interfaces to enforce cognate pairing (Lewis et al., 2014), as did Liu and colleagues (Liu et al., 2015), and had success in improving yield of the appropriately paired heavy and light chains. Furthermore, Lewis and colleagues reported that pairing specificity appears to be dominated by differences in VH:VL interactions (as opposed to CH1:CL) (Lewis et al., 2014). The interface positions reported here and their variation with IGHV and IGLKV gene origin could be useful in identifying candidate mutations.

5.6 Chapter summary

In this chapter we performed both sequence and structural analyses in an effort to extend our understanding of the VH:VL pairing problem. While we were not able to link the results of our structural and sequence analyses, we identified a number of interface positions which appear to be differentially involved according to the genetic origins of the antibody in question as well as a number of gene pairings which appear to be over- or underrepresented in the current corpus of paired sequencing data. Finally, we presented data from a novel sequencing method which produced over 3x as much data as was available prior, and validated one of our previously identified overrepresented gene pairings.

6 | Conclusions and future work

6.1 Conclusions

The present thesis has gone someway towards showing how high-throughput, structurally-informed methods can be applied to BCR repertoire data to identify antigen-specific antibodies.

In Chapter 2, we showed how paratope prediction of bulk heavy chain repertoires could be applied to discover novel antigen-specific antibodies via immune repertoire mining, leading to the discovery of fourteen anti-PTx mAbs which were clonally unrelated to any previously described PTx antibody. We called this method "paratyping".

In Chapter 3, we showed how structural annotation and its successor, rapid structural modelling, can be applied to BCR repertoires to enable more directly functional comparisons of BCR repertoires, which we applied to a commonly used transgenic discovery platform, the Kymouse. Combined with sequence analysis, this resulted in some insights into both how homology-based structural annotation performs on large sequencing datasets and the transgenic platform itself. The Kymouse appears to have reduced CDRH3 diversity due to a relative paucity of NP insertions: however, the repertoire of CDRH3 structural shapes used by the Kymouse is distinctly more human-like than mouse-like.

In Chapter 4, we curated the largest known database of Ebolavirus-binding antibodies ("EBOV-AbDab"). We used this database to apply paratyping and clonotyping to BCR repertoires from individuals vaccinated with a candidate Ebolavirus vaccine, Ad26.ZEBOV/MVA-BN-Filo. In doing so, we identified putative EBOV-binding sequences enriched in convergent and expanded clonotypes, in particular IGHV3-15 lineages which have been identified as of interest in previous work in the field. We therefore showed how immune repertoire mining can be used to annotate BCR repertoires with predicted antigen-

specificity, leading to important insights.

Finally, in Chapter 5 we performed an up-to-date structural analysis of VH:VL interfaces of human antibodies, with reference to their genetic origin. We found that kappa and lambda antibodies differ with respect to a number of interface properties. We then looked at the mounting 10x data curated in the Observed Antibody Space, and a large, novel dataset to identify over- or underrepresented pairings of heavy and light chain genes, and found a single IGHV gene which showed a significant kappa preference in both of these different datasets.

We hope that the result within Chapters 2, 3 and 4 will be informative in the future as to how paratope prediction and structure prediction can be useful for antibody discovery and repertoire analysis generally. In the new era of paired BCR-sequencing data, Chapter 5 will provide a reference for subsequent VH:VL pairing studies.

6.2 Future work

There are two categories of "future work": specific outstanding questions directly related to the work in this thesis and more general future directions of work that will advance our ability to predict antigen-specificity. I will first outline the most specific advances that can be made to work in each chapter.

In the work described in Chapter 2, we found that paratope prediction could be a useful addition to repertoire mining, either to increase confidence in predictions made by more sequence-conservative methods like clonotyping, or to allow the discovery of more sequence-diverse antibodies predicted to bind to the same epitope. Ghanbarpour and colleagues have arguably explored one of the obvious next steps for computational epitope binning: they created a more sophisticated paratope similarity metric based on similarity in language model embeddings (Ghanbarpour et al., 2022). As the authors of this work did not

compare their work to results that they would have obtained using paratyping, it is unknown if this sophistication has added value. In the benchmark we performed in this chapter (comparing the performance of paratyping, AbLigity and SPACE on a dataset of coronavirus-binding antibodies), we showed that a key limitation to computational epitope binning methods is CDR length matching which is currently a requirement of both paratyping and SPACE. AbLigity without CDR length matching displayed the best performance, so perhaps future work should focus on elaborations of this class of method, combining physicochemical paratope and structural information. A fundamental concern is the stability of deep learning methods trained on small structural datasets, such as Parapred, when applied to repertoire data which have diversity that is unparalleled in the training data.

In Chapter 3, we showed that the naive BCR repertoires of humans, humanised mice and mice were distinguishable on the basis of their predicted distribution of CDRH3 structural shapes. There are certain limitations of this work that are unlikely to be solved, partly due to the inherent difficulty in comparing the BCR repertoires of species separated by nearly 100 million years of evolution, such as the problem of comparing subjects and B cells of equivalent developmental stages, in addition to controlling for prior antigen exposure. Limitations aside, it would be of interest to compare the structural and sequence diversity of BCR repertoires of humans and humanised mice post-exposure to a common antigen, given the baseline we have established here. Similarity in the kinds of antibodies produced in response to a common antigen is the key question in terms of the future use of transgenic models in vaccine response modelling, and while our method indicates the repertoires of humans and transgenic models will start from different places, we cannot answer this question conclusively until we compare repertoire responses to the same antigen. Khatri and colleagues recently released a single-cell, longitudinal BCR sequencing dataset of DTaP (diphtheria-tetanus-acellular Pertussis) vaccinees. PTx is one of the aP antigens, so this dataset could possibly be compared with the bulk PTx data from the previous chapter (Khatri et al., 2021).

In Chapter 4, we showed that the BCR repertoires of Ebolavirus vaccinees appear to change in response to vaccination, and via comparison to our novel Ebolavirus antibody database, EBOV-AbDab, we found an interesting insight into vaccination: the continued prevalence of IGHV3-15 antibodies in the response to both natural infection and vaccination. The lineage of antibodies we identified, 6666/6662/3T0245, appears to be very common both at baseline and in the most expanded clonotypes following vaccination. Such antibodies are not highly common in the Observed Antibody Space, however were found in HA-sorted cells of TIV vaccinees. We showed that such antibodies are not indeed HA-reactive to common influenza strains, which would explain their prevalence at baseline. We do not yet have a convincing explanation as to why they are so common. The field does not currently have a statistical framework for describing how common an antibody should be that does not rely on large amounts of non-productive data required to fit personalised recombination models. A simple measure here to describe the distribution of "antibody commonness" would be extremely useful here.

There is the most future work to be done for Chapter 5 in which we attempted to identify statistically over- and underrepresented gene pairings in sequencing data and find structural correlates from contact and energetics analysis of human VH:VL interfaces. Looking at the attention matrices of language models trained on paired data is the most actionable future work in this area: otherwise, more data will need to be generated (or rather collected in one place) to elucidate which particular VH:VL pairings are energetically unfavourable. Recent improvements in structure prediction such as ImmuneBuilder (Abanades-Kenyon et al, in prep) may allow us to generate some of this data synthetically, at scale.

These are specific suggestions as to how to either improve or complete the presented work. More generally, one avenue for identifying antigen- or epitope-specific antibodies from repertoire data that has not been explored in this thesis is docking-based virtual screening (Schneider et al., 2021). Both computational expense and low accuracy mean that these methods are generally not

feasible at the scale of BCR repertoire sequencing data: the fastest modelling tools still take 30 seconds per model when performing side chain minimisation (Abanades et al., 2022), and even rigid-body docking is on the order of minutes per antibody/antigen complex (Pierce et al., 2014). In the virtual screening work by Schneider and colleagues, reasonable pose accuracy could only be obtained where the approximate epitope is already known (Schneider et al., 2022) and binder/non-binder pose classification is an additional complex problem that would need to be solved. AlphaFold models have not yet shown accuracy in antibody/antigen structure prediction (Evans et al., 2022). There are other deep learning-based models which attempt antibody/antigen docking (Jin et al., 2022; Myung et al., 2022) but their application to virtual screening remains to be seen.

Given the continued challenges posed by antibody/antigen docking (especially at scale), the most actionable direction for future work continues to lie in immune repertoire mining (e.g. paratyping and clonotyping). This class of method aligns with the experimental methods that are becoming more commonplace: derivations of single-cell sequencing. One such derivation enables direct sequencing of antigen- (Setliff et al., 2019) and epitope-specificity (Walker et al., 2022) at the same time as native VH:VL pairing, a protocol called LIBRA-seq. These methods are still costly and scale-limited, however, so the challenge of rapidly performing epitope assignment of entire BCR repertoires at low cost still remains, reinforcing the utility of computational tools that can link known binders to larger bulk sequencing repertoires. Deep learning will continue to play a role moving forward, however any such methods should be assessed for their generalisability and performance on diverse sequencing datasets. Language models trained on large bodies of repertoire data will likely be useful here. Finally, deciphering the features of antibodies that result in binding to the same epitope is an important task which is generally considered secondary to prediction itself but which would be of considerable interest to the community.

7| Appendices

Position	Frequency	Entropy	Gene MI	Contact MI	% charged
50	99.6%	0.25	0.02	0.0007	0.08
52	99.3%	0.14	0.01	-0.004	0.08
118	99.2%	0.01	-0.0003	-0.001	0.00
115	95.2%	1.67	0.03	0.05	2.33
103	92.8%	0.44	0.01	0.07	0.45
44	91.6%	0.20	0.01	0.02	1.36
42	91.5%	0.64	0.15	-0.002	0.08
114	83.0%	2.41	0.07	0.13	10.69
116	78.9%	1.01	0.05	0.04	83.58
113	82.1%	2.48	0.07	0.03	19.12
112	72.5%	2.64	0.06	0.03	14.91
49	63.3%	0.40	0.07	0.02	5.57
66	58.9%	2.2	0.28	0.03	25.30

Table 1: | **VH interface positions seen in more than 50% of complexes in the human structural dataset.** The thirteen heavy chain positions which were observed in the interface of more than 50% of complexes are shown with their frequency, the entropy in their amino acid usage (minimum value of 0.0), the mutual information between IGHV gene and amino acid identity at that position, and the mutual information between amino acid identity at that position and occurrence in the interface or not. The percentage of sequences in the dataset in which a charged residue is observed is also shown.

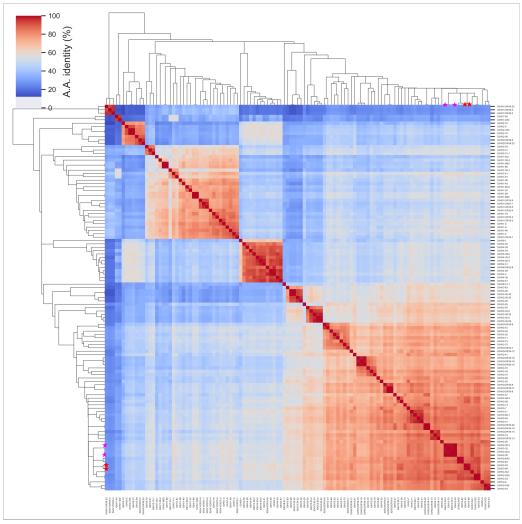

Figure 1: | **Cluster map of maximum amino acid identity between IGHV genes (IMGT, as of January 2023).** Amino acid identity was calculated between all pairs of alleles in the IMGT gapped human IGHV reference set; for each pair of genes, the maximum sequence identity between any pair of alleles was retained. The cluster map shows this distance matrix and an associated single-linkage clustering dendrogram. The pink stars indicate genes IGHV3-30 and IGHV3-30-3; the red stars indicate IGHV3-53 and IGHV3-66. These can be seen to cluster together, and are among the most closely related pairs of IGHV genes.

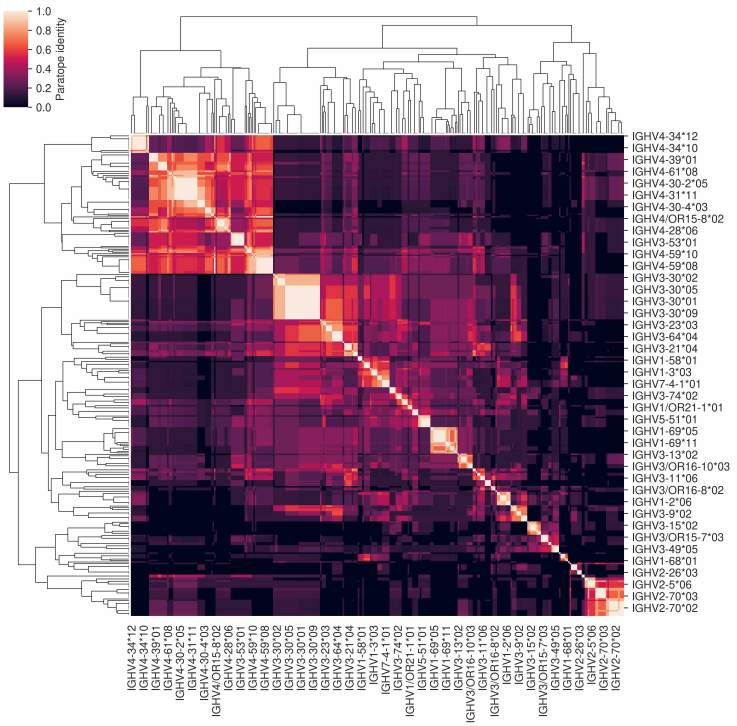

Figure 2: | **Predicted paratope identity across CDRH1 and CDRH2 of germline alleles (IMGT, as of February 14th 2020).** Paratope identity may be equal to or in excess of 75% across V genes within the same and different families - this was detected in alleles from genes IGHV3-53 and IGHV4-59, IGHV3-66 and IGHV4-59 and IGHV7-4-1 and IGHV1-8, suggesting that in large enough sequencing datasets paratyping could cluster antibodies from different V gene families.

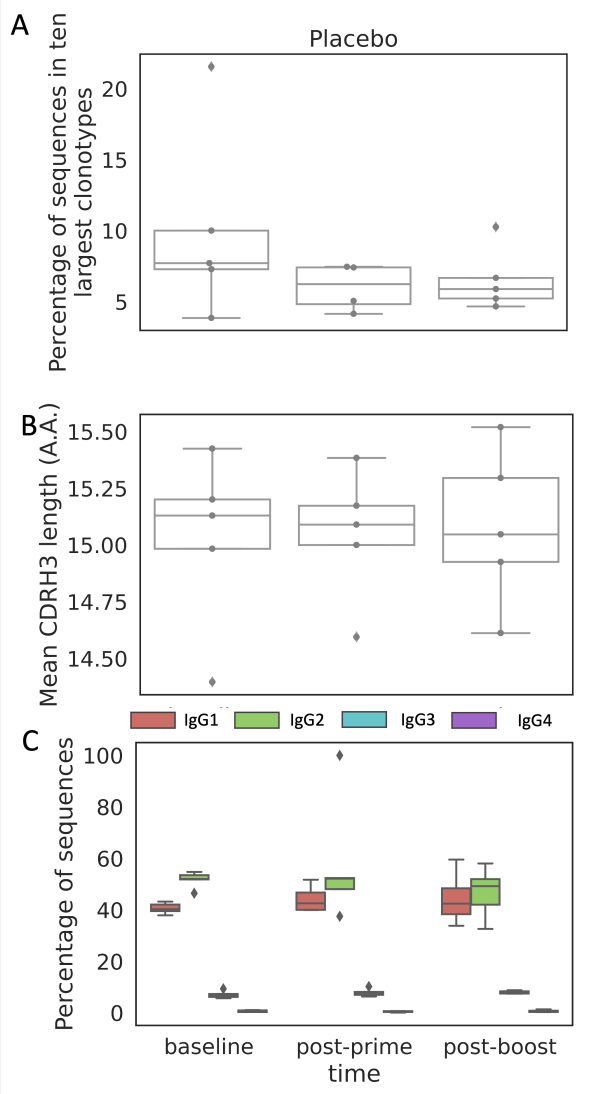

Figure 3: | **The IgG repertoires of subjects who received placebo did not show changes in expansion, CDRH3 length or IgG subtype frequencies.** Repertoire perturbation is not observed in the placebo repertoires. Expansion, CDRH3 length and IgG subtype usage do not change significantly over time.

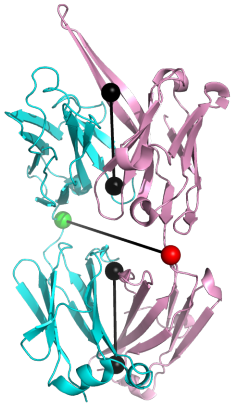

Figure 4: | **Elbow angle calculation.** Elbow angle is calculated as the arccos of the dot product of the two vectors, the first describing the pseudo-twofold rotation of the VH onto the VL domain and the CH1 onto the CL domain. Where the direction of the cross-product of these vectors is the same as the vector relating the Cαs of the residues delimiting the variable/constant region, an angle less than 180 is reported. Otherwise, the obtuse angle is reported. This figure was generated using the *elbow_angle.py* script.

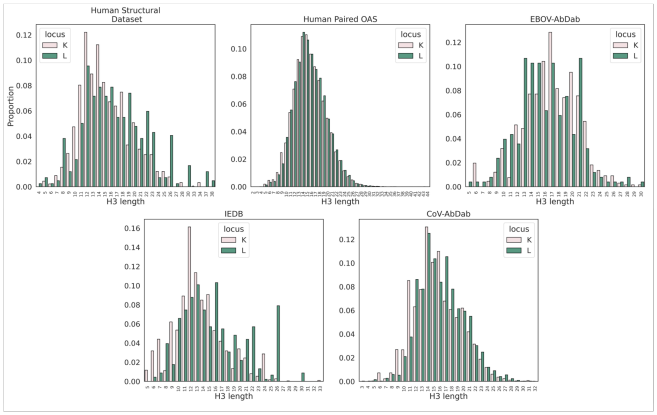

Figure 5: | **CDRH3 length distribution of kappa vs. lambda antibodies in diverse databases.** The CDRH3 distributions of antibodies with kappa or lambda antibodies are shown for the human structural dataset, Human Paired OAS, EBOV-AbDab, the IEDB and CoV-AbDab. Unlike Human Paired OAS which involves sequencing of mRNA, all antibodies in EBOV-AbDab, IEDB and CoV-AbDab are energetically favourable enough to be expressed. For the IEDB (2,595 sequences), the average CDRH3 length of kappa antibodies is 13.4 vs. 15.8 for lambda. For CoV-AbDab (9,022 sequences), the kappa average is 15.6 vs. 16.3 for lambda. EBOV-AbDab is the other way round with a kappa average of 17.0 vs. 16.2 amino acids. All of these databases have a kappa bias however, with between 1.6 to 4.1 times as many kappa antibodies as lambda.

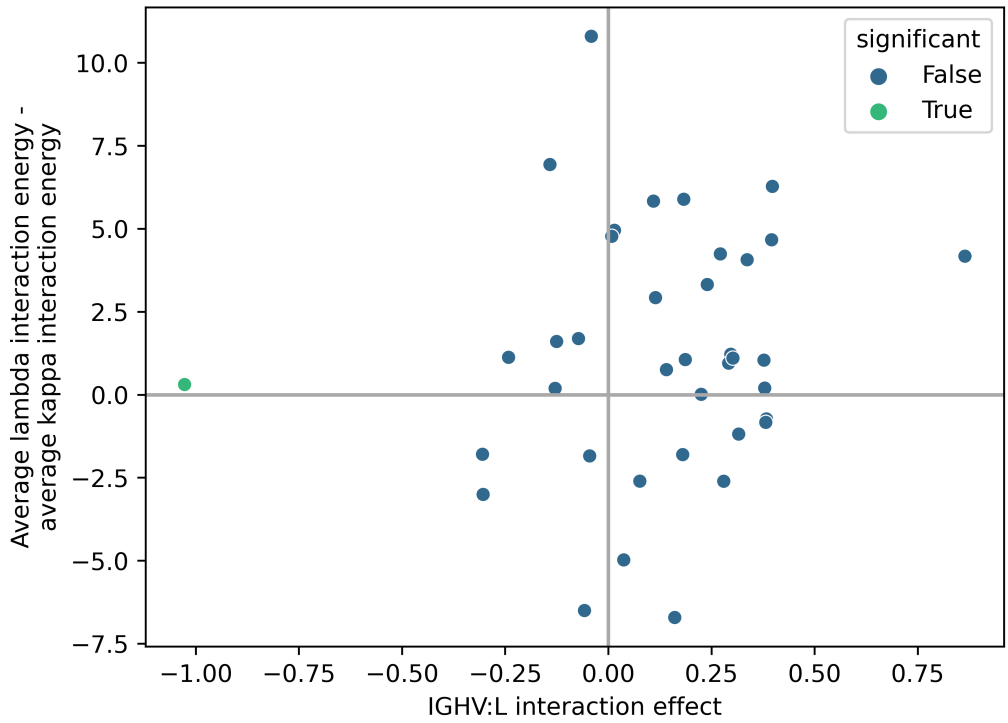

Figure 6: | **Interaction coefficient and difference in average interface energy.** For each of the 39 IGHV genes which have representatives solved in structure with both kappa and lambda light chains, the average difference in FoldX interaction energy between the kappa and lambda complexes is plotted against the corresponding IGHV:K coefficient in the regression model in Section 5.3.4.1. The two are in agreement in the upper right and lower left quadrants (i.e. in the upper right quadrant, the regression indicates that there is an enrichment of lambda antibodies of that IGHV gene, and the interfaces of lambda complexes are more energetically favourable than kappa interfaces in structures of that complex on average). IGHV3-72, the only significant term in the regression, is indicated.

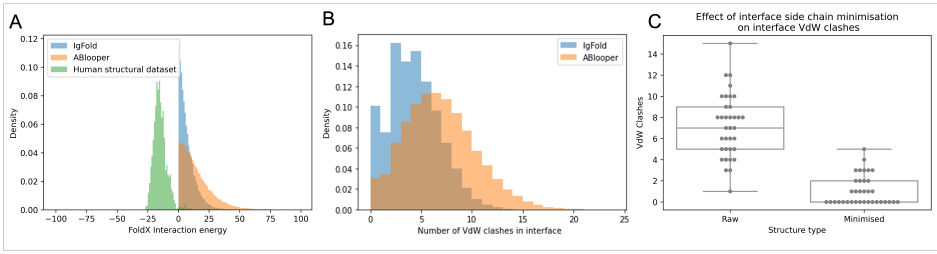

Figure 7: | **VH:VL interface energetics in models produced by ABlooper and IgFold.** Current high-throughput modelling techniques (ABlooper and IgFold), which would be required to produce sufficient data to make broad statements about the energetics of kappa/lambda interfaces for particular IGHV genes, do not produce energetically favourable VH:VL interfaces i.e. positive FoldX interaction energies (A). This is the result of VdW clashes between side chains in the majority of interfaces (B). The distributions are derived from 83,495 ABlooper models of Human Paired OAS provided by Brennan Abanades-Kenyon (personal communication) and 104,992 IgFold models of Paired OAS (murine sequences included) (Ruffolo & Gray, 2022). To assess whether energy minimisation of the side chains can resolve this, we used OpenMM to minimise side chains on interface residues on IGHV3-72 antibodies from the ABlooper set. Of 35 models, side chain minimisation removed VdW clashes in 18/35 models. This indicates that for around 50% of models side chain minimisation may be insufficient to produce energetically favourable interfaces (Eastman et al., 2017).

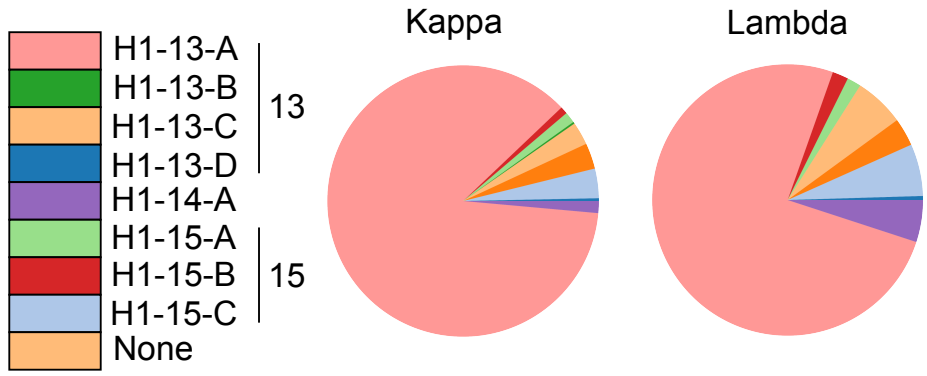

Figure 8: | **Kappa/lambda split in H1 canonical forms.** A greater proportion of kappa-paired heavy chain sequences have H1-13-A conformations than lambda-paired heavy chain sequences. This analysis does not account for different heavy chain sequences which are observed multiply.

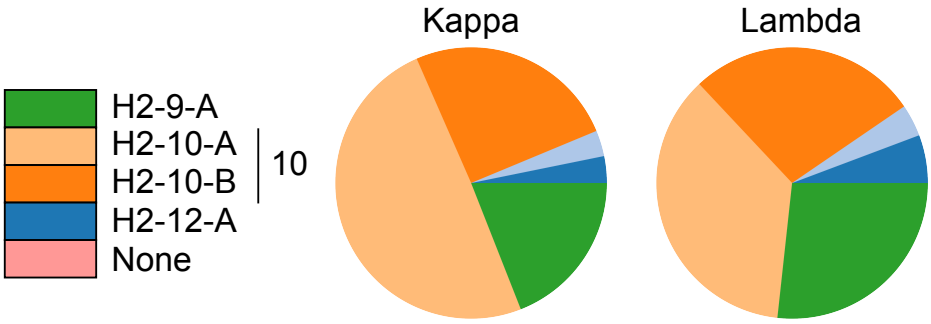

Figure 9: | **Kappa/lambda split in H2 canonical forms.** A greater proportion of lambda-paired H2s are length 9 and adopt the H2-9-A conformation than kappa-paired H2s. This analysis does not account for different heavy chain sequences which are observed multiply.

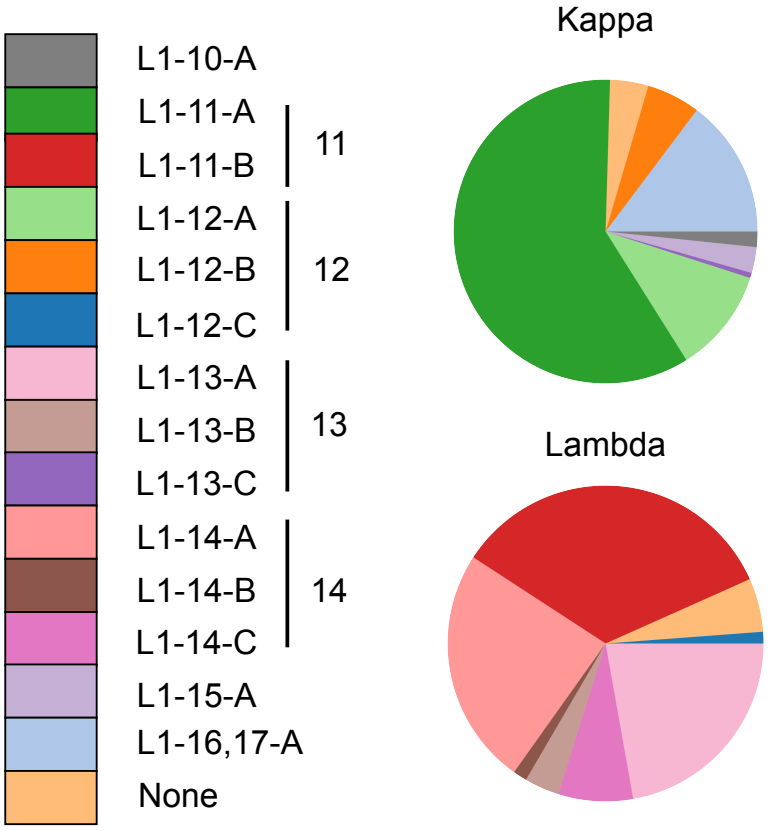

Figure 10: | **Kappa/lambda split in L1 canonical forms.** Kappa and lambda antibodies adopt nearly entirely different L1 canonical forms.

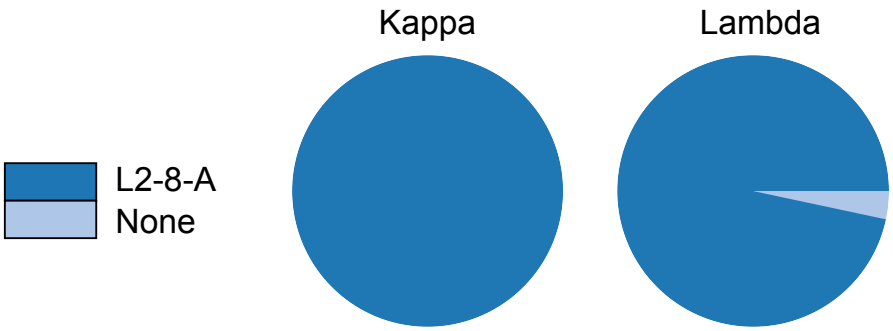

Figure 11: | **Kappa/lambda split in L2 canonical forms.** All CDRL2s in the dataset adopted L2-8-A or were unassigned.

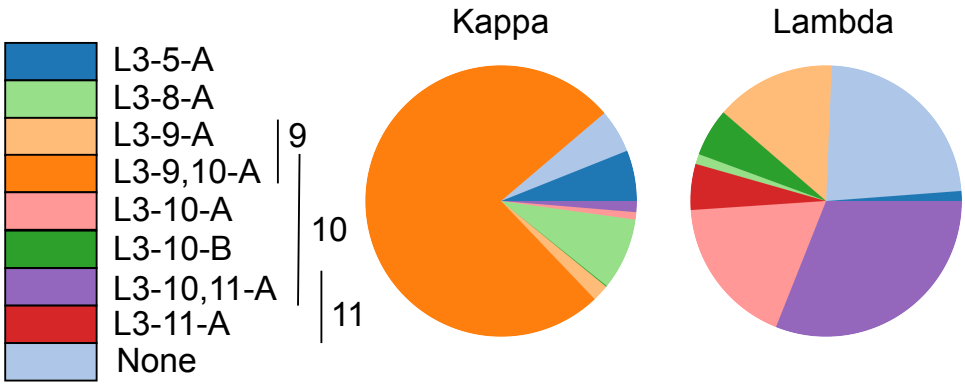

Figure 12: | **Kappa/lambda split in L3 canonical forms.** CDRL3s adopt largely different conformations in kappa and lambda antibodies.

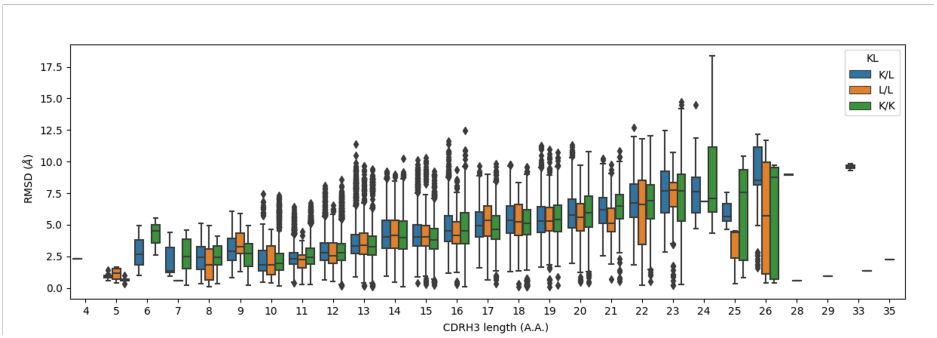

Figure 13: | **RMSD distribution between pairs of CDRH3s of matched length between kappa and lambda antibodies.** RMSD was calculated between all pairs of antibodies of the same length CDRH3. Each comparison is classified as "K/K" (between two kappa antibodies), "K/L" (a kappa and a lambda antibody) and "L/L" (between two lambda antibodies).

Position	Frequency	Entropy	Gene MI	Contact MI	Locus MI	% charged
50	99.9%	0.07	0.005	-0.001	0.005	0.00
118	99.2%	0.06	0.002	-0.004	0.004	0.00
42	98.8%	0.31	0.042	-0.002	0.015	0.83
103	98.1%	0.35	0.011	0.000	0.000	7.76
49	97.3%	0.85	0.195	0.003	0.020	0.00
116	96.4%	2.54	0.106	0.001	0.070	12.12
52	96.4%	0.59	0.105	0.002	0.051	2.48
44	91.0%	0.42	0.026	0.079	0.016	3.31
55	83.7%	0.62	0.054	0.004	0.017	4.29
107	82.1%	2.04	0.393	0.006	0.163	10.92
115	79.8%	2.07	0.203	0.010	0.201.	8.28
114	69.8%	2.52	0.268	0.002	0.077	10.54
40	58.2%	1.77	0.402	0.010	0.285	14.23
105	53.0%	1.41	0.334	0.010	0.220	3.61
38	52.0%	1.84	0.281	-0.000	0.065	7.61

Table 2: | **VL interface positions seen in more than 50% of complexes in the human structural dataset.** The fifteen light chain positions which were observed in the interface of more than 50% of complexes are shown with their frequency, the entropy in their amino acid usage, the mutual information of their amino acids with respect to IGKLV gene ("Gene MI") and whether or not they are in the contact ("Contact MI") and the kappa/lambda locus ("Locus MI"). The percentage of sequences in the dataset in which a charged residue is observed is also shown.

Bibliography

Abanades, B., Georges, G., Bujotzek, A., & Deane, C. M. (2022). ABlooper: Fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics*, *38*(7), 1877–1880. https://doi.org/10.1093/bioinformatics/btac016

Abdiche, Y. N. (2017). High-Throughput Antibody Characterization. *Genetic Engineering & Biotechnology News*, *37*(15), 28–29. https://doi.org/10.1089/gen.37.15.13

Abhinandan, K. R., & Martin, A. C. R. (2007). Analyzing the "Degree of Humanness" of Antibody Sequences. *Journal of Molecular Biology*, *369*(3), 852–862. https://doi.org/10.1016/j.jmb.2007.02.100

Adolf-Bryfogle, J., Xu, Q., North, B., Lehmann, A., & Dunbrack, R. L., Jr. (2015). PyIgClassify: A database of antibody CDR structural classifications. *Nucleic Acids Research*, *43*(D1), D432–D438. https://doi.org/10.1093/nar/gku1106

Agrafiotis, A., Neumeier, D., Hong, K.-L., Chowdhury, T., Ehling, R., Kuhn, R., Sandu, I., Kreiner, V., Cotet, T.-S., Laslo, D., Anzböck, S., Starkie, D., Lightwood, D. J., Oxenius, A., Reddy, S. T., & Yermanos, A. (2022). Generation of a single-cell B cell atlas of antibody repertoires and transcriptomes to identify signatures associated with antigen-specificity. https://doi.org/10.1101/2021.11.09.467876

Ahmad, T. A., Eweida, A. E., & Sheweita, S. A. (2016). B-cell epitope mapping for the design of vaccines and effective diagnostics. *Trials in Vaccinology*, *5*, 71–83. https://doi.org/10.1016/j.trivac.2016.04.003

Ahmed, R., Miller, J., Wilson, P. C., Capra, J. D., & Wrammert, J. (2013). *Rapid generation of antibodies* (US8563305B2). Retrieved January 13, 2023, from https://patents.google.com/patent/US8563305B2/en

Akbar, R., Robert, P. A., Pavlović, M., Jeliazkov, J. R., Snapkov, I., Slabodkin, A., Weber, C. R., Scheffer, L., Miho, E., Haff, I. H., Haug, D. T. T., Lund-Johansen, F., Safonova, Y., Sandve, G. K., & Greiff, V. (2021). A compact vocabulary of paratope-epitope interactions enables predictability of

antibody-antigen binding. *Cell Reports*, *34*(11), 108856. https://doi.org/10.1016/j.celrep.2021.108856

Alamyar, E., Duroux, P., Lefranc, M.-P., & Giudicelli, V. (2012). 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 F. T. Christiansen & B. D. Tait (Eds.), *Immunogenetics: Methods and Applications in Clinical Practice* (pp. 569–604). Humana Press. https://doi.org/10.1007/978-1-61779-842-9_32

Alfaleh, M. A., Alsaab, H. O., Mahmoud, A. B., Alkayyal, A. A., Jones, M. L., Mahler, S. M., & Hashem, A. M. (2020). Phage Display Derived Monoclonal Antibodies: From Bench to Bedside. *Frontiers in Immunology*, *11*. Retrieved September 25, 2022, from https://www.frontiersin.org/articles/10.3389/fimmu.2020.01986

Al-Lazikani, B., Lesk, A. M., & Chothia, C. (1997). Standard conformations for the canonical structures of immunoglobulins11Edited by I. A. Wilson. *Journal of Molecular Biology*, *273*(4), 927–948. https://doi.org/10.1006/jmbi.1997.1354

Allcorn, L. C., & Martin, A. C. R. (2002). SACS—Self-maintaining database of antibody crystal structure information. *Bioinformatics*, *18*(1), 175–181. https://doi.org/10.1093/bioinformatics/18.1.175

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, *16*(12), 1315–1322. https://doi.org/10.1038/s41592-019-0598-1

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ambrosetti, F., Olsen, T. H., Olimpieri, P. P., Jiménez-García, B., Milanetti, E., Marcatilli, P., & Bonvin, A. M. J. J. (2020). proABC-2: PRediction Of AntiBody Contacts v2 and its application to information-driven docking. *bioRxiv*, 2020.03.18.967828. https://doi.org/10.1101/2020.03.18.967828

Anywaine, Z., Barry, H., Anzala, O., Mutua, G., Sirima, S. B., Eholie, S., Kibuuka, H., Bétard, C., Richert, L., Lacabaratz, C., McElrath, M. J., Rosa, S. C. D., Cohen, K. W., Shukarev, G., Katwere, M., Robinson, C., Gaddah, A., Heerwegh, D., Bockstal, V., … Group, o. b. o. t. E. S. (2022). Safety and immunogenicity of 2-dose heterologous Ad26.ZEBOV, MVA-BN-Filo Ebola vaccination in children and adolescents in Africa: A randomised, placebo-controlled, multicentre Phase II clinical trial. *PLOS Medicine*, *19*(1), e1003865. https://doi.org/10.1371/journal.pmed.1003865

Arakawa, H., Hauschild, J., & Buerstedde, J.-M. (2002). Requirement of the Activation-Induced Deaminase (AID) Gene for Immunoglobulin Gene Conversion. *Science*, *295*(5558), 1301–1306. https://doi.org/10.1126/science.1067308

Avnir, Y., Watson, C. T., Glanville, J., Peterson, E. C., Tallarico, A. S., Bennett, A. S., Qin, K., Fu, Y., Huang, C.-Y., Beigel, J. H., Breden, F., Zhu, Q., & Marasco, W. A. (2016). IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Scientific Reports*, *6*(1), 20842. https://doi.org/10.1038/srep20842

Avram, O., Kigel, A., Vaisman-Mentesh, A., Kligsberg, S., Rosenstein, S., Dror, Y., Pupko, T., & Wine, Y. (2021). PASA: Proteomic analysis of serum antibodies web server. *PLOS Computational Biology*, *17*(1), e1008607. https://doi.org/10.1371/journal.pcbi.1008607

Babrak, L., Marquez, S., Busse, C. E., Lees, W. D., Miho, E., Ohlin, M., Rosenfeld, A. M., Stervbo, U., Watson, C. T., & Schramm, C. A. (2022). Adaptive Immune Receptor Repertoire (AIRR) Community Guide to TR and IGImmunoglobulin (IG)T cell receptor (TR)Gene annotationGene Annotation. In A. W. Langerak (Ed.), *Immunogenetics: Methods and Protocols* (pp. 279–296). Springer US. https://doi.org/10.1007/978-1-0716-2115-8_16

Bailly, M., Mieczkowski, C., Juan, V., Metwally, E., Tomazela, D., Baker, J., Uchida, M., Kofman, E., Raoufi, F., Motlagh, S., Yu, Y., Park, J., Raghava, S., Welsh, J., Rauscher, M., Raghunathan, G., Hsieh, M., Chen, Y.-L., Nguyen, H. T., … Fayadat-Dilman, L. (2020). Predicting Antibody Developa-

bility Profiles Through Early Stage Discovery Screening. *mAbs*, *12*(1), 1743053. https://doi.org/10.1080/19420862.2020.1743053

Bangaru, S., Antanasijevic, A., Kose, N., Sewall, L. M., Jackson, A. M., Suryadevara, N., Zhan, X., Torres, J. L., Copps, J., de la Peña, A. T., Crowe, J. E., & Ward, A. B. (2022). Structural mapping of antibody landscapes to human betacoronavirus spike proteins. *Science Advances*, *8*(18), eabn2911. https://doi.org/10.1126/sciadv.abn2911

Bashford-Rogers, R. J. M., Bergamaschi, L., McKinney, E. F., Pombal, D. C., Mescia, F., Lee, J. C., Thomas, D. C., Flint, S. M., Kellam, P., Jayne, D. R. W., Lyons, P. A., & Smith, K. G. C. (2019). Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature*, *574*(7776), 122–126. https://doi.org/10.1038/s41586-019-1595-3

Bashford-Rogers, R. J. M., Smith, K. G. C., & Thomas, D. C. (2018). Antibody repertoire analysis in polygenic autoimmune diseases. *Immunology*, *155*(1), 3–17. https://doi.org/10.1111/imm.12927

Bausch, D. G. (2021). The need for a new strategy for Ebola vaccination. *Nature Medicine*, *27*(4), 580–581. https://doi.org/10.1038/s41591-021-01313-w

Becker, P. D., Legrand, N., Geelen, C. M. M. v., Noerder, M., Huntington, N. D., Lim, A., Yasuda, E., Diehl, S. A., Scheeren, F. A., Ott, M., Weijer, K., Wedemeyer, H., Santo, J. P. D., Beaumont, T., Guzman, C. A., & Spits, H. (2010). Generation of Human Antigen-Specific Monoclonal IgM Antibodies Using Vaccinated "Human Immune System" Mice. *PLOS ONE*, *5*(10), e13137. https://doi.org/10.1371/journal.pone.0013137

Benjamin, D. C., Berzofsky, J. A., East, I. J., Gurd, F. R. N., Hannum, C., Leach, S. J., Margoliash, E., Michael, J. G., Miller, A., Prager, E. M., Reichlin, M., Sercarz, E. E., Smith-Gill, S. J., Todd, P. E., & Wilson, A. C. (1984). The Antigenic Structure of Proteins: A Reappraisal. *Annual Review of Immunology*, *2*(1), 67–101. https://doi.org/10.1146/annurev.iy.02.040184.000435

Berrondo, M., Kaufmann, S., & Berrondo, M. (2014). Automated Aufbau of antibody structures from given sequences using Macromoltek's SmrtMolAntibody. *Proteins: Structure, Function, and Bioinformatics*, *82*(8), 1636–1645. https://doi.org/10.1002/prot.24595

Bianchi, M., Turner, H. L., Nogal, B., Cottrell, C. A., Oyen, D., Pauthner, M., Bastidas, R., Nedellec, R., McCoy, L. E., Wilson, I. A., Burton, D. R., Ward, A. B., & Hangartner, L. (2018). Electron-Microscopy-Based Epitope Mapping Defines Specificities of Polyclonal Antibodies Elicited during HIV-1 BG505 Envelope Trimer Immunization. *Immunity*, *49*(2), 288–300.e8. https://doi.org/10.1016/j.immuni.2018.07.009

Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., & Chudakov, D. M. (2015). MiXCR: Software for comprehensive adaptive immunity profiling. *Nature Methods*, *12*(5), 380–381. https://doi.org/10.1038/nmeth.3364

Borcherding, N., Bormann, N. L., & Kraus, G. (2020). scRepertoire: An R-based toolkit for single-cell immune receptor analysis. *F1000Research*, *9*, 47. https://doi.org/10.12688/f1000research.22139.2

Bornholdt, Z. A., Herbert, A. S., Mire, C. E., He, S., Cross, R. W., Wec, A. Z., Abelson, D. M., Geisbert, J. B., James, R. M., Rahim, M. N., Zhu, W., Borisevich, V., Banadyga, L., Gunn, B. M., Agans, K. N., Wirchnianski, A. S., Goodwin, E., Tierney, K., Shestowsky, W. S., … Dye, J. M. (2019). A Two-Antibody Pan-Ebolavirus Cocktail Confers Broad Therapeutic Protection in Ferrets and Nonhuman Primates. *Cell Host & Microbe*, *25*(1), 49–58.e5. https://doi.org/10.1016/j.chom.2018.12.005

Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., Jones, C. D., Simen, B. B., Hanczaruk, B., Nguyen, K. D., Nadeau, K. C., Egholm, M., Miklos, D. B., Zehnder, J. L., & Fire, A. Z. (2009). Measurement and Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel V-D-J Pyrosequencing. *Science Translational Medicine*, *1*(12), 12ra23–12ra23. https://doi.org/10.1126/scitranslmed.3000540

Bozhanova, N. G., Flyak, A. I., Brown, B. P., Ruiz, S. E., Salas, J., Rho, S., Bombardi, R. G., Myers, L., Soto, C., Bailey, J. R., Crowe, J. E., Bjorkman, P. J., & Meiler, J. (2022). Computational identification of HCV neutralizing antibodies with a common HCDR3 disulfide bond motif in the antibody repertoires of infected individuals. *Nature Communications*, *13*(1), 3178. https://doi.org/10.1038/s41467-022-30865-9

Brezski, R. J., Vafa, O., Petrone, D., Tam, S. H., Powers, G., Ryan, M. H., Luongo, J. L., Oberholtzer, A., Knight, D. M., & Jordan, R. E. (2009). Tumor-associated and microbial proteases compromise host IgG effector functions by a single cleavage proximal to the hinge. *Proceedings of the National Academy of Sciences*, *106*(42), 17864–17869. https://doi.org/10.1073/pnas.0904174106

Briney, B., & Burton, D. R. (2018). Massively scalable genetic analysis of antibody repertoires. *bioRxiv*, 447813. https://doi.org/10.1101/447813

Briney, B., Inderbitzin, A., Joyce, C., & Burton, D. R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, *566*(7744), 393–397. https://doi.org/10.1038/s41586-019-0879-y

Briney, B. S., Willis, J. R., Hicar, M. D., Thomas II, J. W., & Crowe Jr, J. E. (2012). Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology*, *137*(1), 56–64. https://doi.org/10.1111/j.1365-2567.2012.03605.x

Brüggemann, M., Caskey, H. M., Teale, C., Waldmann, H., Williams, G. T., Surani, M. A., & Neuberger, M. S. (1989). A repertoire of monoclonal antibodies with human heavy chains from transgenic mice. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(17), 6709–6713. https://doi.org/10.1073/pnas.86.17.6709

Brüggemann, M., Osborn, M. J., Ma, B., Hayre, J., Avis, S., Lundstrom, B., & Buelow, R. (2015). Human antibody production in transgenic animals. *Archivum Immunologiae Et Therapiae Experimentalis*, *63*(2), 101–108. https://doi.org/10.1007/s00005-014-0322-x

Bujotzek, A., Fuchs, A., Qu, C., Benz, J., Klostermann, S., Antes, I., & Georges, G. (2015). MoFvAb: Modeling the Fv region of antibodies. *mAbs*, *7*(5), 838–852. https://doi.org/10.1080/19420862.2015.1068492

Busse, C. E., Czogiel, I., Braun, P., Arndt, P. F., & Wardemann, H. (2014). Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *European Journal of Immunology*, *44*(2), 597–603. https://doi.org/10.1002/eji.201343917

Cagigi, A., Misasi, J., Ploquin, A., Stanley, D. A., Ambrozak, D., Tsybovsky, Y., Mason, R. D., Roederer, M., & Sullivan, N. J. (2018). Vaccine Generation of Protective Ebola Antibodies and Identification of Conserved B-Cell Signatures. *The Journal of Infectious Diseases*, *218*(suppl_5), S528–S536. https://doi.org/10.1093/infdis/jiy333

Callaway, E. (2020). Revolutionary cryo-EM is taking over structural biology. *Nature*, *578*(7794), 201–201. https://doi.org/10.1038/d41586-020-00341-9

Cao, Y., Su, B., Guo, X., Sun, W., Deng, Y., Bao, L., Zhu, Q., Zhang, X., Zheng, Y., Geng, C., Chai, X., He, R., Li, X., Lv, Q., Zhu, H., Deng, W., Xu, Y., Wang, Y., Qiao, L., … Xie, X. S. (2020). Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients' B Cells. *Cell*, *182*(1), 73–84.e16. https://doi.org/10.1016/j.cell.2020.05.025

Chailyan, A., Marcatili, P., & Tramontano, A. (2011). The association of heavy and light chain variable domains in antibodies: Implications for antigen specificity. *The FEBS Journal*, *278*(16), 2858–2866. https://doi.org/10.1111/j.1742-4658.2011.08207.x

Chailyan, A., Tramontano, A., & Marcatili, P. (2012). A database of immunoglobulins with integrated tools: DIGIT. *Nucleic Acids Research*, *40*(D1), D1230–D1234. https://doi.org/10.1093/nar/gkr806

Chen, E. C., Gilchuk, P., Zost, S. J., Ilinykh, P. A., Binshtein, E., Huang, K., Myers, L., Bonissone, S., Day, S., Kona, C. R., Trivette, A., Reidy, J. X., Sutton, R. E., Gainza, C., Monroig, S., Davidson, E., Saphire, E. O., Doranz, B. J., Castellana, N., … Crowe, J. E. (2022). Systematic analysis of human antibody response to ebolavirus glycoprotein reveals high prevalence of neutralizing public clonotypes. https://doi.org/10.1101/2022.01.12.476089

Chen, F., Tzarum, N., Wilson, I. A., & Law, M. (2019). VH1-69 antiviral broadly neutralizing antibodies: Genetics, structures, and relevance to rational vaccine design. *Current Opinion in Virology*, *34*, 149–159. https://doi.org/10.1016/j.coviro.2019.02.004

Chen, R., & Weng, Z. (2002). Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins: Structure, Function, and Bioinformatics*, *47*(3), 281–294. https://doi.org/10.1002/prot.10092

Chen, Z., Collins, A. M., Wang, Y., & Gaëta, B. A. (2010). Clustering-based identification of clonally-related immunoglobulin gene sequence sets. *Immunome Research*, *6*(1), S4. https://doi.org/10.1186/1745-7580-6-S1-S4

Cheng, Y. (2015). Single-Particle Cryo-EM at Crystallographic Resolution. *Cell*, *161*(3), 450–457. https://doi.org/10.1016/j.cell.2015.03.049

Chinery, L., Wahome, N., Moal, I., & Deane, C. M. (2022). Paragraph - Antibody Paratope prediction using Graph Neural Networks with minimal feature vectors. https://doi.org/10.1101/2022.06.10.495640

Choi, Y., & Deane, C. M. (2010). FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins: Structure, Function, and Bioinformatics*, *78*(6), 1431–1440. https://doi.org/10.1002/prot.22658

Chothia, C., & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *Journal of Molecular Biology*, *196*(4), 901–917. https://doi.org/10.1016/0022-2836(87)90412-8

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M., & Poljak, R. J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature*, *342*(6252), 877–883. https://doi.org/10.1038/342877a0

Cirelli, K. M., Carnathan, D. G., Nogal, B., Martin, J. T., Rodriguez, O. L., Upadhyay, A. A., Enemuo, C. A., Gebru, E. H., Choe, Y., Viviano, F., Nakao, C., Pauthner, M. G., Reiss, S., Cottrell, C. A., Smith, M. L., Bastidas, R., Gibson, W., Wolabaugh, A. N., Melo, M. B., … Crotty, S. (2019). Slow Delivery Immunization Enhances HIV Neutralizing Antibody and Germinal Center Responses via Modulation of Immunodominance. *Cell*, *177*(5), 1153–1171.e28. https://doi.org/10.1016/j.cell.2019.04.012

Clark, S. A., Clark, L. E., Pan, J., Coscia, A., McKay, L. G. A., Shankar, S., Johnson, R. I., Brusic, V., Choudhary, M. C., Regan, J., Li, J. Z., Griffiths, A., & Abraham, J. (2021). SARS-CoV-2 evolution in an immunocom-

promised host reveals shared neutralization escape mechanisms. *Cell*, *184*(10), 2605–2617.e18. https://doi.org/10.1016/j.cell.2021.03.027

Clifford, J., Høie, M. H., Nielsen, M., Deleuran, S., Peters, B., & Marcatili, P. (2022). BepiPred-3.0: Improved B-cell epitope prediction using protein language models. https://doi.org/10.1101/2022.07.11.499418

Cline, J., Braman, J. C., & Hogrefe, H. H. (1996). PCR Fidelity of Pfu DNA Polymerase and Other Thermostable DNA Polymerases. *Nucleic Acids Research*, *24*(18), 3546–3551. https://doi.org/10.1093/nar/24.18.3546

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Cohen-Dvashi, H., Zehner, M., Ehrhardt, S., Katz, M., Elad, N., Klein, F., & Diskin, R. (2020). Structural Basis for a Convergent Immune Response against Ebola Virus. *Cell Host & Microbe*, *27*(3), 418–427.e4. https://doi.org/10.1016/j.chom.2020.01.007

Collins, A. M., & Watson, C. T. (2018). Immunoglobulin Light Chain Gene Rearrangements, Receptor Editing and the Development of a Self-Tolerant Antibody Repertoire. *Frontiers in Immunology*, *9*, 2249. https://doi.org/10.3389/fimmu.2018.02249

Corcoran, M. M., Phad, G. E., Bernat, N. V., Stahl-Hennig, C., Sumida, N., Persson, M. A. A., Martin, M., & Hedestam, G. B. K. (2016). Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nature Communications*, *7*(1), 13642. https://doi.org/10.1038/ncomms13642

Corrie, B. D., Marthandan, N., Zimonja, B., Jaglale, J., Zhou, Y., Barr, E., Knoetze, N., Breden, F. M. W., Christley, S., Scott, J. K., Cowell, L. G., & Breden, F. (2018). iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunological Reviews*, *284*(1), 24–41. https://doi.org/10.1111/imr.12666

Corti, D., Bianchi, S., Vanzetta, F., Minola, A., Perez, L., Agatic, G., Guarino, B., Silacci, C., Marcandalli, J., Marsland, B. J., Piralla, A., Percivalle, E., Sallusto, F., Baldanti, F., & Lanzavecchia, A. (2013). Cross-neutralization of four paramyxoviruses by a human monoclonal antibody. *Nature*, *501*(7467), 439–443. https://doi.org/10.1038/nature12442

Corti, D., Misasi, J., Mulangu, S., Stanley, D. A., Kanekiyo, M., Wollen, S., Ploquin, A., Doria-Rose, N. A., Staupe, R. P., Bailey, M., Shi, W., Choe, M., Marcus, H., Thompson, E. A., Cagigi, A., Silacci, C., Fernandez-Rodriguez, B., Perez, L., Sallusto, F., … Sullivan, N. J. (2016). Protective monotherapy against lethal Ebola virus infection by a potently neutralizing antibody. *Science*, *351*(6279), 1339–1342. https://doi.org/10.1126/science.aad5224

Corti, D., Suguitan, A. L., Pinna, D., Silacci, C., Fernandez-Rodriguez, B. M., Vanzetta, F., Santos, C., Luke, C. J., Torres-Velez, F. J., Temperton, N. J., Weiss, R. A., Sallusto, F., Subbarao, K., & Lanzavecchia, A. (2010). Heterosubtypic neutralizing antibodies are produced by individuals immunized with a seasonal influenza vaccine. *The Journal of Clinical Investigation*, *120*(5), 1663–1673. https://doi.org/10.1172/JCI41902

Cowell, L. G., & Kepler, T. B. (2000). The Nucleotide-Replacement Spectrum Under Somatic Hypermutation Exhibits Microsequence Dependence That Is Strand-Symmetric and Distinct from That Under Germline Mutation. *The Journal of Immunology*, *164*(4), 1971–1976. https://doi.org/10.4049/jimmunol.164.4.1971

Cui, A., Niro, R. D., Heiden, J. A. V., Briggs, A. W., Adams, K., Gilbert, T., O'Connor, K. C., Vigneault, F., Shlomchik, M. J., & Kleinstein, S. H. (2016). A Model of Somatic Hypermutation Targeting in Mice Based on High-Throughput Ig Sequencing Data. *The Journal of Immunology*, *197*(9), 3566–3574. https://doi.org/10.4049/jimmunol.1502263

Curtis, N. C., & Lee, J. (2020). Beyond bulk single-chain sequencing: Getting at the whole receptor. *Current Opinion in Systems Biology*, *24*, 93–99. https://doi.org/10.1016/j.coisb.2020.10.008

D'Angelo, S., Ferrara, F., Naranjo, L., Erasmus, M. F., Hraber, P., & Bradbury, A. R. M. (2018). Many Routes to an Antibody Heavy-Chain CDR3: Nec-

essary, Yet Insufficient, for Specific Binding. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.00395

Davidson, E., & Doranz, B. J. (2014). A high-throughput shotgun mutagenesis approach to mapping B-cell antibody epitopes. *Immunology*, *143*(1), 13–20. https://doi.org/10.1111/imm.12323

Davila, A., Xu, Z., Li, S., Rozewicki, J., Wilamowski, J., Kotelnikov, S., Kozakov, D., Teraguchi, S., & Standley, D. M. (2022). AbAdapt: An adaptive approach to predicting antibody–antigen complex structures from sequence. *Bioinformatics Advances*, *2*(1), vbac015. https://doi.org/10.1093/bioadv/vbac015

Davis, C. W., Jackson, K. J. L., McElroy, A. K., Halfmann, P., Huang, J., Chennareddy, C., Piper, A. E., Leung, Y., Albariño, C. G., Crozier, I., Ellebedy, A. H., Sidney, J., Sette, A., Yu, T., Nielsen, S. C. A., Goff, A. J., Spiropoulou, C. F., Saphire, E. O., Cavet, G., … Ahmed, R. (2019). Longitudinal Analysis of the Human B Cell Response to Ebola Virus Infection. *Cell*, *177*(6), 1566–1582.e17. https://doi.org/10.1016/j.cell.2019.04.036

Deac, A., VeliČković, P., & Sormanni, P. (2019). Attentive Cross-Modal Paratope Prediction. *Journal of Computational Biology*, *26*(6), 536–545. https://doi.org/10.1089/cmb.2018.0175

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. *Nature Biotechnology*, *34*(5), 518–524. https://doi.org/10.1038/nbt.3423

DeKosky, B. J., Lungu, O. I., Park, D., Johnson, E. L., Charab, W., Chrysostomou, C., Kuroda, D., Ellington, A. D., Ippolito, G. C., Gray, J. J., & Georgiou, G. (2016). Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proceedings of the National Academy of Sciences*, *113*(19), E2636–E2645. https://doi.org/10.1073/pnas.1525510113

Del Vecchio, A., Deac, A., Liò, P., & Veličković, P. (2021). Neural message passing for joint paratope-epitope prediction. *arXiv:2106.00757 [cs, q-bio]*. Retrieved September 8, 2021, from http://arxiv.org/abs/2106.00757

Dominguez, C., Boelens, R., & Bonvin, A. M. J. J. (2003). HADDOCK: A ProteinProtein Docking Approach Based on Biochemical or Biophysical In-

formation. *Journal of the American Chemical Society*, *125*(7), 1731–1737. https://doi.org/10.1021/ja026939x

Dörner, T., Brezinschek, H. P., Brezinschek, R. I., Foster, S. J., Domiati-Saad, R., & Lipsky, P. E. (1997). Analysis of the frequency and pattern of somatic mutations within nonproductively rearranged human variable heavy chain genes. *The Journal of Immunology*, *158*(6), 2779–2789. Retrieved September 21, 2022, from https://www.jimmunol.org/content/158/6/2779

Douek, D. C., Betts, M. R., Brenchley, J. M., Hill, B. J., Ambrozak, D. R., Ngai, K.-L., Karandikar, N. J., Casazza, J. P., & Koup, R. A. (2002). A Novel Approach to the Analysis of Specificity, Clonality, and Frequency of HIV-Specific T Cell Responses Reveals a Potential Mechanism for Control of Viral Escape. *The Journal of Immunology*, *168*(6), 3099–3104. https://doi.org/10.4049/jimmunol.168.6.3099

Dugan, H. L., Guthmiller, J. J., Arevalo, P., Huang, M., Chen, Y.-Q., Neu, K. E., Henry, C., Zheng, N.-Y., Lan, L. Y.-L., Tepora, M. E., Stovicek, O., Bitar, D., Palm, A.-K. E., Stamper, C. T., Changrob, S., Utset, H. A., Coughlan, L., Krammer, F., Cobey, S., & Wilson, P. C. (2020). Preexisting immunity shapes distinct antibody landscapes after influenza virus infection and vaccination in humans. *Science Translational Medicine*, *12*(573), eabd3601. https://doi.org/10.1126/scitranslmed.abd3601

Dunbar, J., Fuchs, A., Shi, J., & Deane, C. (2013). ABangle: Characterising the VH–VL orientation in antibodies. *Protein Engineering, Design and Selection*, *26*(10), 611–620. https://doi.org/10.1093/protein/gzt020

Dunbar, J., & Deane, C. M. (2016). ANARCI: Antigen receptor numbering and receptor classification. *Bioinformatics*, *32*(2), 298–300. https://doi.org/10.1093/bioinformatics/btv552

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., & Deane, C. M. (2014). SAbDab: The structural antibody database. *Nucleic Acids Research*, *42*(D1), D1140–D1146. https://doi.org/10.1093/nar/gkt1043

Dunn-Walters, D. K., & Ademokun, A. A. (2010). B cell repertoire and ageing. *Current Opinion in Immunology*, *22*(4), 514–520. https://doi.org/10.1016/j.coi.2010.04.009

Dye, J. M., Muhammad, M., Bakken, R., Barth, J., Kuehne, A., & Olinger, G. G. (2009). Protection of cynomolgus macaques from Ebola Sudan infection (45.10). *The Journal of Immunology*, *182*(1 Supplement), 45.10–45.10. Retrieved August 10, 2022, from https://www.jimmunol.org/content/182/1_Supplement/45.10

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, *13*(7), e1005659. https://doi.org/10.1371/journal.pcbi.1005659

Ehrenmann, F., Kaas, Q., & Lefranc, M.-P. (2010). IMGT/3Dstructure-DB and IMGT/DomainGapAlign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Research*, *38*(suppl_1), D301–D307. https://doi.org/10.1093/nar/gkp946

Ehrenmann, F., & Lefranc, M.-P. (2011). IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harbor Protocols*, *2011*(6), pdb.prot5636. https://doi.org/10.1101/pdb.prot5636

Ehrhardt, S. A., Zehner, M., Krähling, V., Cohen-Dvashi, H., Kreer, C., Elad, N., Gruell, H., Ercanoglu, M. S., Schommers, P., Gieselmann, L., Eggeling, R., Dahlke, C., Wolf, T., Pfeifer, N., Addo, M. M., Diskin, R., Becker, S., & Klein, F. (2019). Polyclonal and convergent antibody response to Ebola virus vaccine rVSV-ZEBOV. *Nature Medicine*, *25*(10), 1589–1600. https://doi.org/10.1038/s41591-019-0602-4

Ellebedy, A. H., Jackson, K. J. L., Kissick, H. T., Nakaya, H. I., Davis, C. W., Roskin, K. M., McElroy, A. K., Oshansky, C. M., Elbein, R., Thomas, S., Lyon, G. M., Spiropoulou, C. F., Mehta, A. K., Thomas, P. G., Boyd, S. D., & Ahmed, R. (2016). Defining antigen-specific plasmablast and mem-

ory B cell subsets in human blood after viral infection or vaccination. *Nature Immunology*, *17*(10), 1226–1234. https://doi.org/10.1038/ni.3533

Eswar, N., Eramian, D., Webb, B., Shen, M.-Y., & Sali, A. (2008). Protein Structure Modeling with MODELLER. In B. Kobe, M. Guss, & T. Huber (Eds.), *Structural Proteomics: High-Throughput Methods* (pp. 145–159). Humana Press. https://doi.org/10.1007/978-1-60327-058-8_8

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., … Hassabis, D. (2022). Protein complex prediction with AlphaFold-Multimer. https://doi.org/10.1101/2021.10.04.463034

Ewing, B., & Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research*, *8*(3), 186–194. https://doi.org/10.1101/gr.8.3.186

Fasnacht, M., Butenhof, K., Goupil-Lamy, A., Hernandez-Guzman, F., Huang, H., & Yan, L. (2014). Automated antibody structure prediction using Accelrys tools: Results and best practices. *Proteins: Structure, Function, and Bioinformatics*, *82*(8), 1583–1598. https://doi.org/10.1002/prot.24604

Feeney, A. J. (1992). Comparison of Junctional Diversity in the Neonatal and Adult Immunoglobulin Repertoires. *International Reviews of Immunology*, *8*(2-3), 113–122. https://doi.org/10.3109/08830189209055567

Fellouse, F. A., Wiesmann, C., & Sidhu, S. S. (2004). Synthetic antibodies from a four-amino-acid code: A dominant role for tyrosine in antigen recognition. *Proceedings of the National Academy of Sciences*, *101*(34), 12467–12472. https://doi.org/10.1073/pnas.0401786101

Ferdous, S., & Martin, A. C. R. (2018). AbDb: Antibody structure database—a database of PDB-derived antibody structures. *Database*, *2018*, bay040. https://doi.org/10.1093/database/bay040

Fernández-Quintero, M. L., Hoerschinger, V. J., Lamp, L. M., Bujotzek, A., Georges, G., & Liedl, K. R. (2020). VH-VL interdomain dynamics observed by computer simulations and NMR. *Proteins: Structure, Function, and Bioinformatics*, *88*(7), 830–839. https://doi.org/10.1002/prot.25872

Fernández-Quintero, M. L., Loeffler, J. R., Kraml, J., Kahler, U., Kamenik, A. S., & Liedl, K. R. (2019). Characterizing the Diversity of the CDR-H3 Loop Conformational Ensembles in Relationship to Antibody Binding Properties. *Frontiers in Immunology*, *9*. Retrieved September 26, 2022, from https://www.frontiersin.org/articles/10.3389/fimmu.2018.03065

Fernández-Quintero, M. L., Pomarici, N. D., Math, B. A., Kroell, K. B., Waibl, F., Bujotzek, A., Georges, G., & Liedl, K. R. (2020). Antibodies exhibit multiple paratope states influencing VH–VL domain orientations. *Communications Biology*, *3*(1), 1–14. https://doi.org/10.1038/s42003-020-01319-z

Finn, J. A., Dong, J., Sevy, A. M., Parrish, E., Gilchuk, I., Nargi, R., Scarlett-Jones, M., Reichard, W., Bombardi, R., Voss, T. G., Meiler, J., & Crowe, J. E. (2020). Identification of Structurally Related Antibodies in Antibody Sequence Databases Using Rosetta-Derived Position-Specific Scoring. *Structure*, *28*(10), 1124–1130.e5. https://doi.org/10.1016/j.str.2020.07.012

Foote, J., & Winter, G. (1992). Antibody framework residues affecting the conformation of the hypervariable loops. *Journal of Molecular Biology*, *224*(2), 487–499. https://doi.org/10.1016/0022-2836(92)91010-M

Forgacs, D., Abreu, R. B., Sautto, G. A., Kirchenbaum, G. A., Drabek, E., Williamson, K. S., Kim, D., Emerling, D. E., & Ross, T. M. (2021). Convergent antibody evolution and clonotype expansion following influenza virus vaccination. *PLOS ONE*, *16*(2), e0247253. https://doi.org/10.1371/journal.pone.0247253

Freeman, J. D., Warren, R. L., Webb, J. R., Nelson, B. H., & Holt, R. A. (2009). Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Research*, *19*(10), 1817–1824. https://doi.org/10.1101/gr.092924.109

Friedensohn, S., Khan, T. A., & Reddy, S. T. (2017). Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires. *Trends in Biotechnology*, *35*(3), 203–214. https://doi.org/10.1016/j.tibtech.2016.09.010

Gadala-Maria, D., Gidoni, M., Marquez, S., Vander Heiden, J. A., Kos, J. T., Watson, C. T., O'Connor, K. C., Yaari, G., & Kleinstein, S. H. (2019). Iden-

tification of Subject-Specific Immunoglobulin Alleles From Expressed Repertoire Sequencing Data. *Frontiers in Immunology*, *10*. Retrieved August 29, 2022, from https://www.frontiersin.org/articles/10.3389/fimmu.2019.00129

Gadala-Maria, D., Yaari, G., Uduman, M., & Kleinstein, S. H. (2015). Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proceedings of the National Academy of Sciences*, *112*(8), E862–E870. https://doi.org/10.1073/pnas.1417683112

Galson, J. D., Clutterbuck, E. A., Trück, J., Ramasamy, M. N., Münz, M., Fowler, A., Cerundolo, V., Pollard, A. J., Lunter, G., & Kelly, D. F. (2015). BCR repertoire sequencing: Different patterns of B-cell activation after two Meningococcal vaccines. *Immunology & Cell Biology*, *93*(10), 885–895. https://doi.org/10.1038/icb.2015.57

Galson, J. D., Pollard, A. J., Trück, J., & Kelly, D. F. (2014). Studying the antibody repertoire after vaccination: Practical applications. *Trends in Immunology*, *35*(7), 319–331. https://doi.org/10.1016/j.it.2014.04.005

Galson, J. D., Schaetzle, S., Bashford-Rogers, R. J. M., Raybould, M. I. J., Kovaltsuk, A., Kilpatrick, G. J., Minter, R., Finch, D. K., Dias, J., James, L. K., Thomas, G., Lee, W.-Y. J., Betley, J., Cavlan, O., Leech, A., Deane, C. M., Seoane, J., Caldas, C., Pennington, D. J., … Osbourn, J. (2020). Deep Sequencing of B Cell Receptor Repertoires From COVID-19 Patients Reveals Strong Convergent Immune Signatures. *Frontiers in Immunology*, *11*. Retrieved September 25, 2022, from https://www.frontiersin.org/articles/10.3389/fimmu.2020.605170

Garripelli, V. K., Wu, Z., & Gupta, S. (2021). Developability assessment for monoclonal antibody drug candidates: A case study. *Pharmaceutical Development and Technology*, *26*(1), 11–20. https://doi.org/10.1080/10837450.2020.1829641

Garske, T., Cori, A., Ariyarajah, A., Blake, I. M., Dorigatti, I., Eckmanns, T., Fraser, C., Hinsley, W., Jombart, T., Mills, H. L., Nedjati-Gilani, G., Newton, E., Nouvellet, P., Perkins, D., Riley, S., Schumacher, D., Shah, A., Van Kerkhove, M. D., Dye, C., … Donnelly, C. A. (2017). Heterogeneities

in the case fatality ratio in the West African Ebola outbreak 2013–2016. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1721), 20160308. https://doi.org/10.1098/rstb.2016.0308

Geisberger, R., Lamers, M., & Achatz, G. (2006). The riddle of the dual expression of IgM and IgD. *Immunology*, *118*(4), 429–437. https://doi.org/10.1111/j.1365-2567.2006.02386.x

Georgiou, G., Ippolito, G. C., Beausang, J., Busse, C. E., Wardemann, H., & Quake, S. R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology*, *32*(2), 158–168. https://doi.org/10.1038/nbt.2782

Gershoni, J. M., Roitburd-Berman, A., Siman-Tov, D. D., Freund, N. T., & Weiss, Y. (2007). Epitope Mapping. *BioDrugs*, *21*(3), 145–156. https://doi.org/10.2165/00063030-200721030-00002

Geysen, H. M., Meloen, R. H., & Barteling, S. J. (1984). Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. *Proceedings of the National Academy of Sciences*, *81*(13), 3998–4002. https://doi.org/10.1073/pnas.81.13.3998

Ghanbarpour, A., Jiang, M., Foster, D., & Chai, Q. (2022). Structure-Free Antibody Paratope Similarity Prediction for *in silico* Epitope Binning via Protein Language Models. https://doi.org/10.2139/ssrn.4147931

Ghraichy, M., Galson, J. D., Kovaltsuk, A., von Niederhäusern, V., Pachlopnik Schmid, J., Recher, M., Jauch, A. J., Miho, E., Kelly, D. F., Deane, C. M., & Trück, J. (2020). Maturation of the Human Immunoglobulin Heavy Chain Repertoire With Age. *Frontiers in Immunology*, *11*, 1734. https://doi.org/10.3389/fimmu.2020.01734

Ghraichy, M., Niederhäusern, V. v., Kovaltsuk, A., Galson, J. D., Deane, C. M., & Trück, J. (2021). *Different B cell subpopulations show distinct patterns in their IgH repertoire metrics* (tech. rep.). https://doi.org/10.1101/2021.09.01.458522

Gibson, W. S., Rodriguez, O. L., Shields, K., Silver, C. A., Dorgham, A., Emery, M., Deikus, G., Sebra, R., Eichler, E. E., Bashir, A., Smith, M. L., & Watson, C. T. (2022). Characterization of the immunoglobulin lambda chain

locus from diverse populations reveals extensive genetic variation. https://doi.org/10.1101/2022.07.20.500849

Gidoni, M., Snir, O., Peres, A., Polak, P., Lindeman, I., Mikocziova, I., Sarna, V. K., Lundin, K. E. A., Clouser, C., Vigneault, F., Collins, A. M., Sollid, L. M., & Yaari, G. (2019). Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nature Communications*, *10*(1), 628. https://doi.org/10.1038/s41467-019-08489-3

Goldstein, L. D., Chen, Y.-J. J., Wu, J., Chaudhuri, S., Hsiao, Y.-C., Schneider, K., Hoi, K. H., Lin, Z., Guerrero, S., Jaiswal, B. S., Stinson, J., Antony, A., Pahuja, K. B., Seshasayee, D., Modrusan, Z., Hötzel, I., & Seshagiri, S. (2019). Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Communications Biology*, *2*(1), 1–10. https://doi.org/10.1038/s42003-019-0551-y

Goldstein, T., Belaganahalli, M. N., Syaluha, E. K., Lukusa, J.-P. K., Greig, D. J., Anthony, S. J., Tremeau-Bravard, A., Thakkar, R., Caciula, A., Mishra, N., Lipkin, W. I., Dhanota, J. K., Smith, B. R., Ontiveros, V. M., Randhawa, N., Cranfield, M., Johnson, C. K., Gilardi, K. V., & Mazet, J. A. K. (2020). Spillover of ebolaviruses into people in eastern Democratic Republic of Congo prior to the 2018 Ebola virus disease outbreak. *One Health Outlook*, *2*(1), 21. https://doi.org/10.1186/s42522-020-00028-1

Gray, A., Bradbury, A. R. M., Knappik, A., Plückthun, A., Borrebaeck, C. A. K., & Dübel, S. (2020). Animal-free alternatives and the antibody iceberg. *Nature Biotechnology*, *38*(11), 1234–1239. https://doi.org/10.1038/s41587-020-0687-9

Green, L. L. (2014). Transgenic mouse strains as platforms for the successful discovery and development of human therapeutic monoclonal antibodies. *Current Drug Discovery Technologies*, *11*(1), 74–84. https://doi.org/10.2174/15701638113109990038

Greiff, V., Menzel, U., Miho, E., Weber, C., Riedel, R., Cook, S., Valai, A., Lopes, T., Radbruch, A., Winkler, T. H., & Reddy, S. T. (2017). Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of An-

tibody Repertoires throughout B Cell Development. *Cell Reports*, *19*(7), 1467–1478. https://doi.org/10.1016/j.celrep.2017.04.054

Greiff, V., Miho, E., Menzel, U., & Reddy, S. T. (2015). Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. *Trends in Immunology*, *36*(11), 738–749. https://doi.org/10.1016/j.it.2015.09.006

Groth, D., Wright, H., Marsh, R., Fuleihan, R., Cunningham-Rundles, C., Sullivan, K., & Feuille, E. (2020). X-Linked Agammaglobulinemia: Infection Frequencies in 226 Patients from the USIDNET Registry. *Journal of Allergy and Clinical Immunology*, *145*(2, Supplement), AB80. https://doi.org/10.1016/j.jaci.2019.12.676

Guerois, R., Nielsen, J. E., & Serrano, L. (2002). Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *Journal of Molecular Biology*, *320*(2), 369–387. https://doi.org/10.1016/S0022-2836(02)00442-4

Guo, J., Zuo, T., Cheng, L., Wu, X., Tang, J., Sun, C., Feng, L., Chen, L., Zhang, L., & Chen, Z. (2015). Simian Immunodeficiency Virus Infection Evades Vaccine-Elicited Antibody Responses to V2 Region. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, *68*(5), 502–510. https://doi.org/10.1097/QAI.0000000000000530

Gupta, N. T., Vander Heiden, J. A., Uduman, M., Gadala-Maria, D., Yaari, G., & Kleinstein, S. H. (2015). Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics*, *31*(20), 3356–3358. https://doi.org/10.1093/bioinformatics/btv359

Hamer, R., Luo, Q., Armitage, J. P., Reinert, G., & Deane, C. M. (2010). I-Patch: Interprotein contact prediction using local network information. *Proteins: Structure, Function, and Bioinformatics*, *78*(13), 2781–2797. https://doi.org/10.1002/prot.22792

Han, J., Schmitz, A. J., Richey, S. T., Dai, Y.-N., Turner, H. L., Mohammed, B. M., Fremont, D. H., Ellebedy, A. H., & Ward, A. B. (2021). Polyclonal epitope mapping reveals temporal dynamics and diversity of human antibody responses to H5N1 vaccination. *Cell Reports*, *34*(4), 108682. https://doi.org/10.1016/j.celrep.2020.108682

Harris, R. S., Sale, J. E., Petersen-Mahrt, S. K., & Neuberger, M. S. (2002). AID Is Essential for Immunoglobulin V Gene Conversion in a Cultured B Cell Line. *Current Biology*, *12*(5), 435–438. https://doi.org/10.1016/S0960-9822(02)00717-0

Hastie, K. M., Li, H., Bedinger, D., Schendel, S. L., Dennison, S. M., Li, K., Rayaprolu, V., Yu, X., Mann, C., Zandonatti, M., Diaz Avalos, R., Zyla, D., Buck, T., Hui, S., Shaffer, K., Hariharan, C., Yin, J., Olmedillas, E., Enriquez, A., … Saphire, E. O. (2021). Defining variant-resistant epitopes targeted by SARS-CoV-2 antibodies: A global consortium study. *Science*, *374*(6566), 472–478. https://doi.org/10.1126/science.abh2315

Havenar-Daughton, C., Newton, I. G., Zare, S. Y., Reiss, S. M., Schwan, B., Suh, M. J., Hasteh, F., Levi, G., & Crotty, S. (2020). Normal human lymph node T follicular helper cells and germinal center B cells accessed via fine needle aspirations. *Journal of Immunological Methods*, *479*, 112746. https://doi.org/10.1016/j.jim.2020.112746

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915

Hershberg, U., & Luning Prak, E. T. (2015). The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1676), 20140239. https://doi.org/10.1098/rstb.2014.0239

Ho, J., Tumkaya, T., Aryal, S., Choi, H., & Claridge-Chang, A. (2019). Moving beyond P values: Data analysis with estimation graphics. *Nature Methods*, *16*(7), 565–566. https://doi.org/10.1038/s41592-019-0470-3

Honegger, A., & Plückthun, A. (2001). Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool. *Journal of Molecular Biology*, *309*(3), 657–670. https://doi.org/10.1006/jmbi.2001.4662

Horns, F., Dekker, C. L., & Quake, S. R. (2020). Memory B Cell Activation, Broad Anti-influenza Antibodies, and Bystander Activation Revealed by Single-Cell Transcriptomics. *Cell Reports*, *30*(3), 905–913.e6. https://doi.org/10.1016/j.celrep.2019.12.063

Hsiao, Y.-C., Shang, Y., DiCara, D. M., Yee, A., Lai, J., Kim, S. H., Ellerman, D., Corpuz, R., Chen, Y., Rajan, S., Cai, H., Wu, Y., Seshasayee, D., & Hötzel, I. (2019). Immune repertoire mining for rapid affinity optimization of mouse monoclonal antibodies. *mAbs*, *11*(4), 735–746. https://doi.org/10.1080/19420862.2019.1584517

Hu, Y., Stumpfe, D., & Bajorath, J. (2013). Advancing the activity cliff concept. *F1000Research*, *2*. https://doi.org/10.12688/f1000research.2-199.v1

Hummer, A. M., Abanades, B., & Deane, C. M. (2022). Advances in computational structure-based antibody design. *Current Opinion in Structural Biology*, *74*, 102379. https://doi.org/10.1016/j.sbi.2022.102379

Humphris, E. L., & Kortemme, T. (2008). Prediction of Protein-Protein Interface Sequence Diversity Using Flexible Backbone Computational Protein Design. *Structure*, *16*(12), 1777–1788. https://doi.org/10.1016/j.str.2008.09.012

Hurtado, J., Flynn, C., Lee, J. H., Salcedo, E., Cottrell, C. A., Skog, P. D., Nemazee, D., Schief, W. R., Landais, E., Sok, D., & Briney, B. (2022). Efficient isolation of rare B cells using next-generation antigen barcoding. https://doi.org/10.1101/2022.06.06.495029

Infante, Y. C., Pupo, A., & Rojas, G. (2014). A combinatorial mutagenesis approach for functional epitope mapping on phage-displayed target antigen. *mAbs*, *6*(3), 637–648. https://doi.org/10.4161/mabs.28395

Ishola, D., Manno, D., Afolabi, M. O., Keshinro, B., Bockstal, V., Rogers, B., Owusu-Kyei, K., Serry-Bangura, A., Swaray, I., Lowe, B., Kowuor, D., Baiden, F., Mooney, T., Smout, E., Köhn, B., Otieno, G. T., Jusu, M., Foster, J., Samai, M., … Lee, S. (2022). Safety and long-term immunogenicity of the two-dose heterologous Ad26.ZEBOV and MVA-BN-Filo Ebola vaccine regimen in adults in Sierra Leone: A combined open-label, non-randomised stage 1, and a randomised, double-blind, controlled stage 2 trial. *The Lancet Infectious Diseases*, *22*(1), 97–109. https://doi.org/10.1016/S1473-3099(21)00125-0

Ivanov, I. I., Schelonka, R. L., Zhuang, Y., Gartland, G. L., Zemlin, M., & Schroeder, H. W. (2005). Development of the Expressed Ig CDR-H3 Repertoire Is Marked by Focusing of Constraints in Length, Amino Acid Use, and

Charge That Are First Established in Early B Cell Progenitors. *The Journal of Immunology*, *174*(12), 7773–7780. https://doi.org/10.4049/jimmunol.174.12.7773

Jackson, K. J. L., Kos, J. T., Lees, W., Gibson, W. S., Smith, M. L., Peres, A., Yaari, G., Corcoran, M., Busse, C. E., Ohlin, M., Watson, C. T., & Collins, A. M. (2022). A BALB/c IGHV Reference Set, Defined by Haplotype Analysis of Long-Read VDJ-C Sequences From F1 (BALB/c x C57BL/6) Mice. *Frontiers in Immunology*, *13*. Retrieved September 22, 2022, from https://www.frontiersin.org/articles/10.3389/fimmu.2022.888555

Jackson, K. J. L., Liu, Y., Roskin, K. M., Glanville, J., Hoh, R. A., Seo, K., Marshall, E. L., Gurley, T. C., Moody, M. A., Haynes, B. F., Walter, E. B., Liao, H.-X., Albrecht, R. A., García-Sastre, A., Chaparro-Riggers, J., Rajpal, A., Pons, J., Simen, B. B., Hanczaruk, B., … Boyd, S. D. (2014). Human Responses to Influenza Vaccination Show Seroconversion Signatures and Convergent Antibody Rearrangements. *Cell Host & Microbe*, *16*(1), 105–114. https://doi.org/10.1016/j.chom.2014.05.013

Jackson, K. J. L., Wang, Y., Gaeta, B. A., Pomat, W., Siba, P., Rimmer, J., Sewell, W. A., & Collins, A. M. (2012). Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. *Immunogenetics*, *64*(1), 3–14. https://doi.org/10.1007/s00251-011-0559-z

Jaffe, D. B., Shahi, P., Adams, B. A., Chrisman, A. M., Finnegan, P. M., Raman, N., Royall, A. E., Tsai, F., Vollbrecht, T., Reyes, D. S., & McDonnell, W. J. (2022). Functional antibodies exhibit light chain coherence. https://doi.org/10.1101/2022.04.23.489267

Jain, T., Sun, T., Durand, S., Hall, A., Houston, N. R., Nett, J. H., Sharkey, B., Bobrowicz, B., Caffry, I., Yu, Y., Cao, Y., Lynaugh, H., Brown, M., Baruah, H., Gray, L. T., Krauland, E. M., Xu, Y., Vásquez, M., & Wittrup, K. D. (2017). Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences*, *114*(5), 944–949. https://doi.org/10.1073/pnas.1616408114

Jakobovits, A., Amado, R. G., Yang, X., Roskos, L., & Schwab, G. (2007). From XenoMouse technology to panitumumab, the first fully human anti-

body product from transgenic mice. *Nature Biotechnology*, *25*(10), 1134–1143. https://doi.org/10.1038/nbt1337

Janeway, C., Travers, P., Walport, M., & Schlomchik, M. (2001). *Immunobiology: The Immune System in Health and Disease*. Garland Pub.

Jespersen, M. C., Mahajan, S., Peters, B., Nielsen, M., & Marcatili, P. (2019). Antibody Specific B-Cell Epitope Predictions: Leveraging Information From Antibody-Antigen Protein Complexes. *Frontiers in Immunology*, *10*. Retrieved January 8, 2023, from https://www.frontiersin.org/articles/10.3389/fimmu.2019.00298

Jespersen, M. C., Peters, B., Nielsen, M., & Marcatili, P. (2017). BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Research*, *45*(W1), W24–W29. https://doi.org/10.1093/nar/gkx346

Jiang, N., He, J., Weinstein, J. A., Penland, L., Sasaki, S., He, X.-S., Dekker, C. L., Zheng, N.-Y., Huang, M., Sullivan, M., Wilson, P. C., Greenberg, H. B., Davis, M. M., Fisher, D. S., & Quake, S. R. (2013). Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination. *Science Translational Medicine*, *5*(171), 171ra19–171ra19. https://doi.org/10.1126/scitranslmed.3004794

Jin, W., Barzilay, R., & Jaakkola, T. (2022). Antibody-Antigen Docking and Design via Hierarchical Equivariant Refinement. https://doi.org/10.48550/arXiv.2207.06616

Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S., & Winter, G. (1986). Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature*, *321*(6069), 522–525. https://doi.org/10.1038/321522a0

Joyce, C., Burton, D. R., & Briney, B. (2020). Comparisons of the antibody repertoires of a humanized rodent and humans by high throughput sequencing. *Scientific Reports*, *10*(1), 1–9. https://doi.org/10.1038/s41598-020-57764-7

Joyce, M. G., Wheatley, A. K., Thomas, P. V., Chuang, G.-Y., Soto, C., Bailer, R. T., Druz, A., Georgiev, I. S., Gillespie, R. A., Kanekiyo, M., Kong, W.-P., Leung, K., Narpala, S. N., Prabhakaran, M. S., Yang, E. S., Zhang, B.,

Zhang, Y., Asokan, M., Boyington, J. C., … McDermott, A. B. (2016). Vaccine-Induced Antibodies that Neutralize Group 1 and Group 2 Influenza A Viruses. *Cell*, *166*(3), 609–623. https://doi.org/10.1016/j.cell.2016.06.043

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., & Blundell, T. L. (2017). Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of Molecular Biology*, *429*(3), 365–371. https://doi.org/10.1016/j.jmb.2016.12.004

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kabat, E. A., Wu, T. T., & Bilofsky, H. (1977). Unusual distributions of amino acids in complementarity determining (hypervariable) segments of heavy and light chains of immunoglobulins and their possible roles in specificity of antibody-combining sites. *Journal of Biological Chemistry*, *252*(19), 6609–6616. Retrieved September 20, 2022, from http://www.scopus.com/inward/record.url?scp=0017747498&partnerID=8YFLogxK

Kaleli, N. E., Karadag, M., & Kalyoncu, S. (2019). Phage display derived therapeutic antibodies have enriched aliphatic content: Insights for developability issues. *Proteins: Structure, Function, and Bioinformatics*, *87*(7), 607–618. https://doi.org/10.1002/prot.25685

Kaplon, H., Chenoweth, A., Crescioli, S., & Reichert, J. M. (2022). Antibodies to watch in 2022. *mAbs*, *14*(1), 2014296. https://doi.org/10.1080/19420862.2021.2014296

Kaur, K., Zheng, N.-Y., Smith, K., Huang, M., Li, L., Pauli, N. T., Dunand, C. J. H., Lee, J.-H., Morrissey, M., Wu, Y., Joachims, M. L., Munroe, M. E., Lau, D., Qu, X., Krammer, F., Wrammert, J., Palese, P., Ahmed, R., James, J. A., & Wilson, P. C. (2015). High Affinity Antibodies against Influenza Characterize the Plasmablast Response in SLE Patients After Vaccina-

tion. *PLOS ONE*, *10*(5), e0125618. https://doi.org/10.1371/journal.pone.0125618

Kawasaki, K., Minoshima, S., Nakato, E., Shibuya, K., Shintani, A., Schmeits, J. L., Wang, J., & Shimizu, N. (1997). One-megabase sequence analysis of the human immunoglobulin lambda gene locus. *Genome Research*, *7*(3), 250–261. https://doi.org/10.1101/gr.7.3.250

Kawasaki, K., Minoshima, S., Nakato, E., Shibuya, K., Shintani, A., Asakawa, S., Sasaki, T., Klobeck, H.-G., Combriato, G., Zachau, H. G., & Shimizu, N. (2001). Evolutionary dynamics of the human immunoglobulin locus and the germline repertoire of the V genes. *European Journal of Immunology*, *31*(4), 1017–1028. https://doi.org/10.1002/1521-4141(200104)31:4<1017::AID-IMMU1017>3.0.CO;2-3

Kemmish, H., Fasnacht, M., & Yan, L. (2017). Fully automated antibody structure prediction using BIOVIA tools: Validation study. *PLOS ONE*, *12*(5), e0177923. https://doi.org/10.1371/journal.pone.0177923

Khatri, I., Diks, A. M., van den Akker, E. B., Oosten, L. E. M., Zwaginga, J. J., Reinders, M. J. T., van Dongen, J. J. M., & Berkowska, M. A. (2021). Longitudinal Dynamics of Human B-Cell Response at the Single-Cell Level in Response to Tdap Vaccination. *Vaccines*, *9*(11), 1352. https://doi.org/10.3390/vaccines9111352

Kim, D., & Park, D. (2019). Deep sequencing of B cell receptor repertoire. *BMB Reports*, *52*(9), 540–547. https://doi.org/10.5483/BMBRep.2019.52.9.192

Klausen, M. S., Anderson, M. V., Jespersen, M. C., Nielsen, M., & Marcatili, P. (2015). LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic Acids Research*, *43*(W1), W349–W355. https://doi.org/10.1093/nar/gkv535

Klein, C., Sustmann, C., Thomas, M., Stubenrauch, K., Croasdale, R., Schanzer, J., Brinkmann, U., Kettenberger, H., Regula, J. T., & Schaefer, W. (2012). Progress in overcoming the chain association issue in bispecific heterodimeric IgG antibodies. *mAbs*, *4*(6), 653–663. https://doi.org/10.4161/mabs.21379

Klein, U., Küppers, R., & Rajewsky, K. (1997). Evidence for a Large Compartment of IgM-Expressing Memory B Cells in Humans. *Blood*, *89*(4), 1288–1298. https://doi.org/10.1182/blood.V89.4.1288

Kleinstein, S. H., Louzoun, Y., & Shlomchik, M. J. (2003). Estimating Hypermutation Rates from Clonal Tree Data. *The Journal of Immunology*, *171*(9), 4639–4649. https://doi.org/10.4049/jimmunol.171.9.4639

Kortemme, T., Morozov, A. V., & Baker, D. (2003). An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein–Protein Complexes. *Journal of Molecular Biology*, *326*(4), 1239–1259. https://doi.org/10.1016/S0022-2836(03)00021-4

Kovaltsuk, A., Krawczyk, K., Galson, J. D., Kelly, D. F., Deane, C. M., & Trück, J. (2017). How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data. *Frontiers in Immunology*, *8*. https://doi.org/10.3389/fimmu.2017.01753

Kovaltsuk, A., Leem, J., Kelm, S., Snowden, J., Deane, C. M., & Krawczyk, K. (2018). Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *The Journal of Immunology*, *201*(8), 2502–2509. https://doi.org/10.4049/jimmunol.1800708

Kovaltsuk, A., Raybould, M. I. J., Wong, W. K., Marks, C., Kelm, S., Snowden, J., Trück, J., & Deane, C. M. (2019). Structural Diversity of B-Cell Receptor Repertoires along the B-cell Differentiation Axis in Humans and Mice. *bioRxiv*, 762880. https://doi.org/10.1101/762880

Kovaltsuk, A., Raybould, M. I. J., Wong, W. K., Marks, C., Kelm, S., Snowden, J., Trück, J., & Deane, C. M. (2020). Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice. *PLOS Computational Biology*, *16*(2), e1007636. https://doi.org/10.1371/journal.pcbi.1007636

Krah, S., Sellmann, C., Rhiel, L., Schröter, C., Dickgiesser, S., Beck, J., Zielonka, S., Toleikis, L., Hock, B., Kolmar, H., & Becker, S. (2017). Engineering bispecific antibodies with defined chain pairing. *New Biotechnology*, *39*, 167–173. https://doi.org/10.1016/j.nbt.2016.12.010

Krantsevich, A., Tang, C., & MacCarthy, T. (2021). Correlations in Somatic Hypermutation Between Sites in IGHV Genes Can Be Explained by Interactions Between AID and/or Pol Hotspots. *Frontiers in Immunology*, *11*. Retrieved September 14, 2022, from https://www.frontiersin.org/articles/10.3389/fimmu.2020.618409

Krause, J. C., Tsibane, T., Tumpey, T. M., Huffman, C. J., Albrecht, R., Blum, D. L., Ramos, I., Fernandez-Sesma, A., Edwards, K. M., García-Sastre, A., Basler, C. F., & Crowe, J. E. (2012). Human Monoclonal Antibodies to Pandemic 1957 H2N2 and Pandemic 1968 H3N2 Influenza Viruses. *Journal of Virology*, *86*(11), 6334–6340. https://doi.org/10.1128/JVI.07158-11

Krause, J. C., Tsibane, T., Tumpey, T. M., Huffman, C. J., Briney, B. S., Smith, S. A., Basler, C. F., & Crowe, J. E. (2011). Epitope-Specific Human Influenza Antibody Repertoires Diversify by B Cell Intraclonal Sequence Divergence and Interclonal Convergence. *The Journal of Immunology*, *187*(7), 3704–3711. https://doi.org/10.4049/jimmunol.1101823

Krawczyk, K., Baker, T., Shi, J., & Deane, C. M. (2013). Antibody i-Patch prediction of the antibody binding site improves rigid local antibody–antigen docking. *Protein Engineering, Design and Selection*, *26*(10), 621–629. https://doi.org/10.1093/protein/gzt043

Krawczyk, K., Kelm, S., Kovaltsuk, A., Galson, J. D., Kelly, D., Trück, J., Regep, C., Leem, J., Wong, W. K., Nowak, J., Snowden, J., Wright, M., Starkie, L., Scott-Tucker, A., Shi, J., & Deane, C. M. (2018). Structurally Mapping Antibody Repertoires. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.01698

Krawczyk, K., Liu, X., Baker, T., Shi, J., & Deane, C. M. (2014). Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics*, *30*(16), 2288–2294. https://doi.org/10.1093/bioinformatics/btu190

Kreer, C., Gruell, H., Mora, T., Walczak, A. M., & Klein, F. (2020). Exploiting B Cell Receptor Analyses to Inform on HIV-1 Vaccination Strategies. *Vaccines*, *8*(1), 13. https://doi.org/10.3390/vaccines8010013

Kringelum, J. V., Lundegaard, C., Lund, O., & Nielsen, M. (2012). Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking. *PLOS Computational Biology*, *8*(12), e1002829. https://doi.org/10.1371/journal.pcbi.1002829

Kringelum, J. V., Nielsen, M., Padkjær, S. B., & Lund, O. (2013). Structural analysis of B-cell epitopes in antibody:protein complexes. *Molecular Immunology*, *53*(1), 24–34. https://doi.org/10.1016/j.molimm.2012.06.001

Kunik, V., Ashkenazi, S., & Ofran, Y. (2012). Paratome: An online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. *Nucleic Acids Research*, *40*(W1), W521–W524. https://doi.org/10.1093/nar/gks480

Kunik, V., Peters, B., & Ofran, Y. (2012). Structural Consensus among Antibodies Defines the Antigen Binding Site. *PLOS Computational Biology*, *8*(2), e1002388. https://doi.org/10.1371/journal.pcbi.1002388

Küppers, R., Zhao, M., Hansmann, M., & Rajewsky, K. (1993). Tracing B cell development in human germinal centres by molecular analysis of single cells picked from histological sections. *The EMBO Journal*, *12*(13), 4955–4967. https://doi.org/10.1002/j.1460-2075.1993.tb06189.x

Kuroda, D., Shirai, H., Jacobson, M. P., & Nakamura, H. (2012). Computer-aided antibody design. *Protein Engineering, Design and Selection*, *25*(10), 507–522. https://doi.org/10.1093/protein/gzs024

Kuroda, D., Shirai, H., Kobori, M., & Nakamura, H. (2008). Structural classification of CDR-H3 revisited: A lesson in antibody modeling. *Proteins: Structure, Function, and Bioinformatics*, *73*(3), 608–620. https://doi.org/10.1002/prot.22087

Kwak, J.-W., & Yoon, C.-S. (1996). A convenient method for epitope competition analysis of two monoclonal antibodies for their antigen binding. *Journal of Immunological Methods*, *191*(1), 49–54. https://doi.org/10.1016/0022-1759(95)00287-1

Lanz, T. V., Brewer, R. C., Ho, P. P., Moon, J.-S., Jude, K. M., Fernandez, D., Fernandes, R. A., Gomez, A. M., Nadj, G.-S., Bartley, C. M., Schubert, R. D., Hawes, I. A., Vazquez, S. E., Iyer, M., Zuchero, J. B., Teegen, B., Dunn, J. E., Lock, C. B., Kipp, L. B., … Robinson, W. H. (2022). Clonally ex-

panded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM. *Nature*, *603*(7900), 321–327. https://doi.org/10.1038/s41586-022-04432-7

Larsen, J. E. P., Lund, O., & Nielsen, M. (2006). Improved method for predicting linear B-cell epitopes. *Immunome Research*, *2*(1), 2. https://doi.org/10.1186/1745-7580-2-2

Laustsen, A. H., Greiff, V., Karatt-Vellatt, A., Muyldermans, S., & Jenkins, T. P. (2021). Animal Immunization, in Vitro Display Technologies, and Machine Learning for Antibody Discovery. *Trends in Biotechnology*, *39*(12), 1263–1273. https://doi.org/10.1016/j.tibtech.2021.03.003

Lavinder, J. J., Horton, A. P., Georgiou, G., & Ippolito, G. C. (2015). Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. *Current Opinion in Chemical Biology*, *24*, 112–120. https://doi.org/10.1016/j.cbpa.2014.11.007

Lazar, G. A., Desjarlais, J. R., Jacinto, J., Karki, S., & Hammond, P. W. (2007). A molecular immunology approach to antibody humanization and functional optimization. *Molecular Immunology*, *44*(8), 1986–1998. https://doi.org/10.1016/j.molimm.2006.09.029

Lee, C.-C. D., Watanabe, Y., Wu, N. C., Han, J., Kumar, S., Pholcharee, T., Seabright, G. E., Allen, J. D., Lin, C.-W., Yang, J.-R., Liu, M.-T., Wu, C.-Y., Ward, A. B., Crispin, M., & Wilson, I. A. (2021). A cross-neutralizing antibody between HIV-1 and influenza virus. *PLOS Pathogens*, *17*(3), e1009407. https://doi.org/10.1371/journal.ppat.1009407

Lee, E.-C., Liang, Q., Ali, H., Bayliss, L., Beasley, A., Bloomfield-Gerdes, T., Bonoli, L., Brown, R., Campbell, J., Carpenter, A., Chalk, S., Davis, A., England, N., Fane-Dremucheva, A., Franz, B., Germaschewski, V., Holmes, H., Holmes, S., Kirby, I., … Bradley, A. (2014). Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. *Nature Biotechnology*, *32*(4), 356–363. https://doi.org/10.1038/nbt.2825

Lee, J. E., Fusco, M. L., Hessell, A. J., Oswald, W. B., Burton, D. R., & Saphire, E. O. (2008). Structure of the Ebola virus glycoprotein bound to an an-

tibody from a human survivor. *Nature*, *454*(7201), 177–182. https://doi.org/10.1038/nature07082

Leem, J., Dunbar, J., Georges, G., Shi, J., & Deane, C. M. (2016). ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation. *mAbs*, *8*(7), 1259–1268. https://doi.org/10.1080/19420862.2016.1205773

Leem, J., Georges, G., Shi, J., & Deane, C. M. (2018). Antibody side chain conformations are position-dependent. *Proteins: Structure, Function, and Bioinformatics*, *86*(4), 383–392. https://doi.org/10.1002/prot.25453

Leem, J., Mitchell, L. S., Farmery, J. H. R., Barton, J., & Galson, J. D. (2022). Deciphering the language of antibodies using self-supervised learning. *Patterns*, *3*(7), 100513. https://doi.org/10.1016/j.patter.2022.100513

Lees, W., Busse, C. E., Corcoran, M., Ohlin, M., Scheepers, C., Matsen, F. A., IV, Yaari, G., Watson, C. T., The AIRR Community, Collins, A., & Shepherd, A. J. (2020). OGRDB: A reference database of inferred immune receptor genes. *Nucleic Acids Research*, *48*(D1), D964–D970. https://doi.org/10.1093/nar/gkz822

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G., & Duroux, P. (2009). IMGT®, the international ImMunoGeneTics information system®. *Nucleic Acids Research*, *37*(suppl_1), D1006–D1012. https://doi.org/10.1093/nar/gkn838

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., & Lefranc, G. (2003). IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental & Comparative Immunology*, *27*(1), 55–77. https://doi.org/10.1016/S0145-305X(02)00039-3

Lewis, S. M., Wu, X., Pustilnik, A., Sereno, A., Huang, F., Rick, H. L., Guntas, G., Leaver-Fay, A., Smith, E. M., Ho, C., Hansen-Estruch, C., Chamberlain, A. K., Truhlar, S. M., Conner, E. M., Atwell, S., Kuhlman, B., & Demarest, S. J. (2014). Generation of bispecific IgG antibodies by structure-based design of an orthogonal Fab interface. *Nature Biotechnology*, *32*(2), 191–198. https://doi.org/10.1038/nbt.2797

Li, G.-M., Chiu, C., Wrammert, J., McCausland, M., Andrews, S. F., Zheng, N.-Y., Lee, J.-H., Huang, M., Qu, X., Edupuganti, S., Mulligan, M., Das, S. R., Yewdell, J. W., Mehta, A. K., Wilson, P. C., & Ahmed, R. (2012). Pandemic H1N1 influenza vaccine induces a recall response in humans that favors broadly cross-reactive memory B cells. *Proceedings of the National Academy of Sciences*, *109*(23), 9047–9052. https://doi.org/10.1073/pnas.1118979109

Li, S., Lefranc, M.-P., Miles, J. J., Alamyar, E., Giudicelli, V., Duroux, P., Freeman, J. D., Corbin, V. D. A., Scheerlinck, J.-P., Frohman, M. A., Cameron, P. U., Plebanski, M., Loveland, B., Burrows, S. R., Papenfuss, A. T., & Gowans, E. J. (2013). IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nature Communications*, *4*(1), 2333. https://doi.org/10.1038/ncomms3333

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Liberis, E., Veličković, P., Sormanni, P., Vendruscolo, M., & Liò, P. (2018). Parapred: Antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics*, *34*(17), 2944–2950. https://doi.org/10.1093/bioinformatics/bty305

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Costa, A. d. S., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction. https://doi.org/10.1101/2022.07.20.500902

Lindeman, I., Emerton, G., Mamanova, L., Snir, O., Polanski, K., Qiao, S.-W., Sollid, L. M., Teichmann, S. A., & Stubbington, M. J. T. (2018). BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nature Methods*, *15*(8), 563–565. https://doi.org/10.1038/s41592-018-0082-3

Liu, Z., Leng, E. C., Gunasekaran, K., Pentony, M., Shen, M., Howard, M., Stoops, J., Manchulenko, K., Razinkov, V., Liu, H., Fanslow, W., Hu, Z., Sun, N., Hasegawa, H., Clark, R., Foltz, I. N., & Yan, W. (2015). A Novel

Antibody Engineering Strategy for Making Monovalent Bispecific Heterodimeric IgG Antibodies by Electrostatic Steering Mechanism. *Journal of Biological Chemistry*, *290*(12), 7535–7562. https://doi.org/10.1074/jbc.M114.620260

Lobo, E. D., Hansen, R. J., & Balthasar, J. P. (2004). Antibody Pharmacokinetics and Pharmacodynamics. *Journal of Pharmaceutical Sciences*, *93*(11), 2645–2668. https://doi.org/10.1002/jps.20178

Long, X., Jeliazkov, J. R., & Gray, J. J. (2019). Non-H3 CDR template selection in antibody modeling through machine learning. *PeerJ*, *7*, e6179. https://doi.org/10.7717/peerj.6179

Lu, L. L., Suscovich, T. J., Fortune, S. M., & Alter, G. (2018). Beyond binding: Antibody effector functions in infectious diseases. *Nature Reviews Immunology*, *18*(1), 46–61. https://doi.org/10.1038/nri.2017.106

Macagno, A., Bernasconi, N. L., Vanzetta, F., Dander, E., Sarasini, A., Revello, M. G., Gerna, G., Sallusto, F., & Lanzavecchia, A. (2010). Isolation of Human Monoclonal Antibodies That Potently Neutralize Human Cytomegalovirus Infection by Targeting Different Epitopes on the gH/gL/UL128-131A Complex. *Journal of Virology*, *84*(2), 1005–1013. https://doi.org/10.1128/JVI.01809-09

Macdonald, L. E., Karow, M., Stevens, S., Auerbach, W., Poueymirou, W. T., Yasenchak, J., Frendewey, D., Valenzuela, D. M., Giallourakis, C. C., Alt, F. W., Yancopoulos, G. D., & Murphy, A. J. (2014). Precise and in situ genetic humanization of 6 Mb of mouse immunoglobulin genes. *Proceedings of the National Academy of Sciences*, *111*(14), 5147–5152. https://doi.org/10.1073/pnas.1323896111

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Maecker, H. T., Lindstrom, T. M., Robinson, W. H., Utz, P. J., Hale, M., Boyd, S. D., Shen-Orr, S. S., & Fathman, C. G. (2012). New tools for classifica-

tion and monitoring of autoimmune diseases. *Nature Reviews Rheumatology*, *8*(6), 317–328. https://doi.org/10.1038/nrrheum.2012.66

Mahita, J., Kim, D.-G., Son, S., Choi, Y., Kim, H.-S., & Bailey-Kellogg, C. (2022). Computational epitope binning reveals functional equivalence of sequence-divergent paratopes. *Computational and Structural Biotechnology Journal*, *20*, 2169–2180. https://doi.org/10.1016/j.csbj.2022.04.036

Maier, J. K. X., & Labute, P. (2014). Assessment of fully automated antibody homology modeling protocols in molecular operating environment. *Proteins: Structure, Function, and Bioinformatics*, *82*(8), 1599–1610. https://doi.org/10.1002/prot.24576

Malvy, D., McElroy, A. K., de Clerck, H., Günther, S., & van Griensven, J. (2019). Ebola virus disease. *The Lancet*, *393*(10174), 936–948. https://doi.org/10.1016/S0140-6736(18)33132-5

Marasco, W. A., Sui, J., & Liddington, R. C. (2011). *Antibodies against influenza virus and methods of use thereof* (US20110038935A1). Retrieved January 13, 2023, from https://patents.google.com/patent/US20110038935A1/en

Marcatili, P., Rosi, A., & Tramontano, A. (2008). PIGS: Automatic prediction of antibody structures. *Bioinformatics*, *24*(17), 1953–1954. https://doi.org/10.1093/bioinformatics/btn341

Marks, C., & Deane, C. M. (2020). How repertoire data are changing antibody science. *The Journal of Biological Chemistry*, *295*(29), 9823–9837. https://doi.org/10.1074/jbc.REV120.010181

Marks, C., Hummer, A. M., Chin, M., & Deane, C. M. (2021). Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics*, (btab434). https://doi.org/10.1093/bioinformatics/btab434

Marks, C., Nowak, J., Klostermann, S., Georges, G., Dunbar, J., Shi, J., Kelm, S., & Deane, C. M. (2017). Sphinx: Merging knowledge-based and *ab initio* approaches to improve protein loop prediction. *Bioinformatics*, btw823. https://doi.org/10.1093/bioinformatics/btw823

Martell, H. J., Masterson, S. G., McGreig, J. E., Michaelis, M., & Wass, M. N. (2019). Is the Bombali virus pathogenic in humans? *Bioinformatics*, *35*(19), 3553–3558. https://doi.org/10.1093/bioinformatics/btz267

Martin, A. C. R., & Thornton, J. M. (1996). Structural Families in Loops of Homologous Proteins: Automatic Classification, Modelling and Application to Antibodies. *Journal of Molecular Biology*, *263*(5), 800–815. https://doi.org/10.1006/jmbi.1996.0617

Mason, D. M., Friedensohn, S., Weber, C. R., Jordi, C., Wagner, B., Meng, S., Gainza, P., Correia, B. E., & Reddy, S. T. (2019). Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space. *bioRxiv*, 617860. https://doi.org/10.1101/617860

Matsuda, F., Ishii, K., Bourvagnet, P., Kuma, K.-i., Hayashida, H., Miyata, T., & Honjo, T. (1998). The Complete Nucleotide Sequence of the Human Immunoglobulin Heavy Chain Variable Region Locus. *Journal of Experimental Medicine*, *188*(11), 2151–2162. https://doi.org/10.1084/jem.188.11.2151

McCafferty, J., Griffiths, A. D., Winter, G., & Chiswell, D. J. (1990). Phage antibodies: Filamentous phage displaying antibody variable domains. *Nature*, *348*(6301), 552–554. https://doi.org/10.1038/348552a0

McDaniel, J. R., DeKosky, B. J., Tanno, H., Ellington, A. D., & Georgiou, G. (2016). Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nature Protocols*, *11*(3), 429–442. https://doi.org/10.1038/nprot.2016.024

McGinnis, S., & Madden, T. L. (2004). BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, *32*(suppl_2), W20–W25. https://doi.org/10.1093/nar/gkh435

McKean, D., Huppi, K., Bell, M., Staudt, L., Gerhard, W., & Weigert, M. (1984). Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proceedings of the National Academy of Sciences*, *81*(10), 3180–3184. https://doi.org/10.1073/pnas.81.10.3180

McLean, G. R., Torres, M., Elguezabal, N., Nakouzi, A., & Casadevall, A. (2002). Isotype Can Affect the Fine Specificity of an Antibody for a Polysac-

charide Antigen. *The Journal of Immunology*, *169*(3), 1379–1386. https://doi.org/10.4049/jimmunol.169.3.1379

McLeod, B., Miura, K., Scally, S. W., Bosch, A., Nguyen, N., Shin, H., Kim, D., Volkmuth, W., Rämisch, S., Chichester, J. A., Streatfield, S., Woods, C., Schief, W. R., Emerling, D., King, C. R., & Julien, J.-P. (2019). Potent antibody lineage against malaria transmission elicited by human vaccination with Pfs25. *Nature Communications*, *10*(1), 4328. https://doi.org/10.1038/s41467-019-11980-6

Meiler, J., Müller, M., Zeidler, A., & Schmäschke, F. (2001). Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Molecular modeling annual*, *7*(9), 360–369. https://doi.org/10.1007/s008940100038

Meng, W., Pan, W., Zhang, A. J. X., Li, Z., Wei, G., Feng, L., Dong, Z., Li, C., Hu, X., Sun, C., Luo, Q., Yuen, K.-Y., Zhong, N., & Chen, L. (2013). Rapid Generation of Human-Like Neutralizing Monoclonal Antibodies in Urgent Preparedness for Influenza Pandemics and Virulent Infectious Diseases. *PLOS ONE*, *8*(6), e66276. https://doi.org/10.1371/journal.pone.0066276

Merriman, B., R&D Team, I. T., & Rothberg, J. M. (2012). Progress in Ion Torrent semiconductor chip based sequencing. *ELECTROPHORESIS*, *33*(23), 3397–3417. https://doi.org/10.1002/elps.201200424

Messmer, B. T., Albesiano, E., Messmer, D., & Chiorazzi, N. (2004). The pattern and distribution of immunoglobulin VH gene mutations in chronic lymphocytic leukemia B cells are consistent with the canonical somatic hypermutation process. *Blood*, *103*(9), 3490–3495. https://doi.org/10.1182/blood-2003-10-3407

Mian, I. S., Bradwell, A. R., & Olson, A. J. (1991). Structure, function and properties of antibody binding sites. *Journal of Molecular Biology*, *217*(1), 133–151. https://doi.org/10.1016/0022-2836(91)90617-f

Miao, Z., Cao, Y., & Jiang, T. (2011). RASP: Rapid modeling of protein side chain conformations. *Bioinformatics*, *27*(22), 3117–3122. https://doi.org/10.1093/bioinformatics/btr538

Miller, R., Maloney, D., McKillop, J., & Levy, R. (1981). In vivo effects of murine hybridoma monoclonal antibody in a patient with T-cell leukemia. *Blood*, *58*(1), 78–86. https://doi.org/10.1182/blood.V58.1.78.78

Misasi, J., & Sullivan, N. J. (2021). Immunotherapeutic strategies to target vulnerabilities in the Ebolavirus glycoprotein. *Immunity*, *54*(3), 412–436. https://doi.org/10.1016/j.immuni.2021.01.015

Morea, V., Tramontano, A., Rustici, M., Chothia, C., & Lesk, A. M. (1997). Antibody structure, prediction and redesign. *Biophysical Chemistry*, *68*(1), 9–16. https://doi.org/10.1016/S0301-4622(96)02266-1

Morea, V., Tramontano, A., Rustici, M., Chothia, C., & Lesk, A. M. (1998). Conformations of the third hypervariable region in the VH domain of immunoglobulins11Edited by I. A. Wilson. *Journal of Molecular Biology*, *275*(2), 269–294. https://doi.org/10.1006/jmbi.1997.1442

Moyer, T. J., Kato, Y., Abraham, W., Chang, J. Y. H., Kulp, D. W., Watson, N., Turner, H. L., Menis, S., Abbott, R. K., Bhiman, J. N., Melo, M. B., Simon, H. A., Herrera-De la Mata, S., Liang, S., Seumois, G., Agarwal, Y., Li, N., Burton, D. R., Ward, A. B., … Irvine, D. J. (2020). Engineered immunogen binding to alum adjuvant enhances humoral immunity. *Nature Medicine*, *26*(3), 430–440. https://doi.org/10.1038/s41591-020-0753-3

Mulangu, S., Dodd, L. E., Davey, R. T., Tshiani Mbaya, O., Proschan, M., Mukadi, D., Lusakibanza Manzo, M., Nzolo, D., Tshomba Oloma, A., Ibanda, A., Ali, R., Coulibaly, S., Levine, A. C., Grais, R., Diaz, J., Lane, H. C., Muyembe-Tamfum, J.-J., & the PALM Writing Group. (2019). A Randomized, Controlled Trial of Ebola Virus Disease Therapeutics. *New England Journal of Medicine*, *381*(24), 2293–2303. https://doi.org/10.1056/NEJMoa1910993

Murin, C. D., Gilchuk, P., Ilinykh, P. A., Huang, K., Kuzmina, N., Shen, X., Bruhn, J. F., Bryan, A. L., Davidson, E., Doranz, B. J., Williamson, L. E., Copps, J., Alkutkar, T., Flyak, A. I., Bukreyev, A., Crowe, J. E., & Ward, A. B. (2021). Convergence of a common solution for broad ebolavirus neutralization by glycan cap-directed human antibodies. *Cell Reports*, *35*(2), 108984. https://doi.org/10.1016/j.celrep.2021.108984

Murphy, A. J., Macdonald, L. E., Stevens, S., Karow, M., Dore, A. T., Poboursky, K., Huang, T. T., Poueymirou, W. T., Esau, L., Meola, M., Mikulka, W., Krueger, P., Fairhurst, J., Valenzuela, D. M., Papadopoulos, N., & Yancopoulos, G. D. (2014). Mice with megabase humanization of their immunoglobulin genes generate antibodies as efficiently as normal mice. *Proceedings of the National Academy of Sciences*, *111*(14), 5153–5158. https://doi.org/10.1073/pnas.1324022111

Myung, Y., Pires, D. E. V., & Ascher, D. B. (2022). CSM-AB: Graph-based antibody–antigen binding affinity prediction and docking scoring function. *Bioinformatics*, *38*(4), 1141–1143. https://doi.org/10.1093/bioinformatics/btab762

Nagata, K., Randall, A., & Baldi, P. (2012). SIDEpro: A novel machine learning approach for the fast and accurate prediction of side-chain conformations. *Proteins: Structure, Function, and Bioinformatics*, *80*(1), 142–153. https://doi.org/10.1002/prot.23170

Neumeier, D., Yermanos, A., Agrafiotis, A., Csepregi, L., Chowdhury, T., Ehling, R. A., Kuhn, R., Cotet, T.-S., Brisset-Di Roberto, R., Di Tacchio, M., Antonialli, R., Starkie, D., Lightwood, D. J., Oxenius, A., & Reddy, S. T. (2022). Phenotypic determinism and stochasticity in antibody repertoires of clonally expanded plasma cells. *Proceedings of the National Academy of Sciences*, *119*(18), e2113766119. https://doi.org/10.1073/pnas.2113766119

Nguyen, D. C., Joyner, C. J., Sanz, I., & Lee, F. E.-H. (2019). Factors Affecting Early Antibody Secreting Cell Maturation Into Long-Lived Plasma Cells. *Frontiers in Immunology*, *10*. Retrieved September 7, 2022, from https://www.frontiersin.org/articles/10.3389/fimmu.2019.02138

Nielsen, S. C. A., Yang, F., Jackson, K. J. L., Hoh, R. A., Röltgen, K., Jean, G. H., Stevens, B. A., Lee, J.-Y., Rustagi, A., Rogers, A. J., Powell, A. E., Hunter, M., Najeeb, J., Otrelo-Cardoso, A. R., Yost, K. E., Daniel, B., Nadeau, K. C., Chang, H. Y., Satpathy, A. T., … Boyd, S. D. (2020). Human B Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2. *Cell Host & Microbe*, *28*(4), 516–525.e5. https://doi.org/10.1016/j.chom.2020.09.002

Nixon, A. E., Sexton, D. J., & Ladner, R. C. (2014). Drugs derived from phage display. *mAbs*, *6*(1), 73–85. https://doi.org/10.4161/mabs.27240

Nogal, B., Bianchi, M., Cottrell, C. A., Kirchdoerfer, R. N., Sewall, L. M., Turner, H. L., Zhao, F., Sok, D., Burton, D. R., Hangartner, L., & Ward, A. B. (2020). Mapping Polyclonal Antibody Responses in Non-human Primates Vaccinated with HIV Env Trimer Subunit Vaccines. *Cell Reports*, *30*(11), 3755–3765.e7. https://doi.org/10.1016/j.celrep.2020.02.061

North, B., Lehmann, A., & Dunbrack, R. L. (2011). A new clustering of antibody CDR loop conformations. *Journal of molecular biology*, *406*(2), 228–256. https://doi.org/10.1016/j.jmb.2010.10.030

Nowak, J., Baker, T., Georges, G., Kelm, S., Klostermann, S., Shi, J., Sridharan, S., & Deane, C. M. (2016). Length-independent structural similarities enrich the antibody CDR canonical class model. *mAbs*, *8*(4), 751–760. https://doi.org/10.1080/19420862.2016.1158370

O'Donnell, K. L., & Marzi, A. (2021). Immunotherapeutics for Ebola Virus Disease: Hope on the Horizon. *Biologics : Targets & Therapy*, *15*, 79–86. https://doi.org/10.2147/BTT.S259069

Ofran, Y., Schlessinger, A., & Rost, B. (2008). Automated Identification of Complementarity Determining Regions (CDRs) Reveals Peculiar Characteristics of CDRs and B Cell Epitopes. *The Journal of Immunology*, *181*(9), 6230–6235. https://doi.org/10.4049/jimmunol.181.9.6230

Ohlin, M., Scheepers, C., Corcoran, M., Lees, W. D., Busse, C. E., Bagnara, D., Thörnqvist, L., Bürckert, J.-P., Jackson, K. J. L., Ralph, D., Schramm, C. A., Marthandan, N., Breden, F., Scott, J., Matsen IV, F. A., Greiff, V., Yaari, G., Kleinstein, S. H., Christley, S., … Collins, A. M. (2019). Inferred Allelic Variants of Immunoglobulin Receptor Genes: A System for Their Evaluation, Documentation, and Naming. *Frontiers in Immunology*, *10*. Retrieved September 14, 2022, from https://www.frontiersin.org/articles/10.3389/fimmu.2019.00435

Ohshima, N., Kubota-Koketsu, R., Iba, Y., Okuno, Y., & Kurosawa, Y. (2014). Two Types of Antibodies Are Induced by Vaccination with A/California/2009pdm Virus: Binding near the Sialic Acid-Binding Pocket and Neutralizing

Both H1N1 and H5N1 Viruses. *PLOS ONE*, *9*(2), e87305. https://doi.org/10.1371/journal.pone.0087305

Okada, J., Ohshima, N., Kubota-Koketsu, R., Ota, S., Takase, W., Azuma, M., Iba, Y., Nakagawa, N., Yoshikawa, T., Nakajima, Y., Ishikawa, T., Asano, Y., Okuno, Y., & Kurosawa, Y. (2010). Monoclonal antibodies in man that neutralized H3N2 influenza viruses were classified into three groups with distinct strain specificity: 1968–1973, 1977–1993 and 1997–2003. *Virology*, *397*(2), 322–330. https://doi.org/10.1016/j.virol.2009.11.025

Olimpieri, P. P., Chailyan, A., Tramontano, A., & Marcatili, P. (2013). Prediction of site-specific interactions in antibody-antigen complexes: The proABC method and server. *Bioinformatics*, *29*(18), 2285–2291. https://doi.org/10.1093/bioinformatics/btt369

Olinger, G. G., Pettitt, J., Kim, D., Working, C., Bohorov, O., Bratcher, B., Hiatt, E., Hume, S. D., Johnson, A. K., Morton, J., Pauly, M., Whaley, K. J., Lear, C. M., Biggins, J. E., Scully, C., Hensley, L., & Zeitlin, L. (2012). Delayed treatment of Ebola virus infection with plant-derived monoclonal antibodies provides protection in rhesus macaques. *Proceedings of the National Academy of Sciences*, *109*(44), 18030–18035. https://doi.org/10.1073/pnas.1213709109

Oliva, B., Bates, P. A., Querol, E., Avilés, F. X., & Sternberg, M. J. E. (1998). Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction11Edited by J. Thornton. *Journal of Molecular Biology*, *279*(5), 1193–1210. https://doi.org/10.1006/jmbi.1998.1847

Olsen, T. H., Boyles, F., & Deane, C. M. (2022). Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, *31*(1), 141–146. https://doi.org/10.1002/pro.4205

Olsen, T. H., Moal, I. H., & Deane, C. M. (2022). AbLang: An antibody language model for completing antibody sequences. *Bioinformatics Advances*, *2*(1), vbac046. https://doi.org/10.1093/bioadv/vbac046

Omer, A., Shemesh, O., Peres, A., Polak, P., Shepherd, A. J., Watson, C. T., Boyd, S. D., Collins, A. M., Lees, W., & Yaari, G. (2020). VDJbase: An adaptive immune receptor genotype and haplotype database. *Nucleic Acids Research*, *48*(D1), D1051–D1056. https://doi.org/10.1093/nar/gkz872

Osborn, M. J., Ma, B., Avis, S., Binnie, A., Dilley, J., Yang, X., Lindquist, K., Ménoret, S., Iscache, A.-L., Ouisse, L.-H., Rajpal, A., Anegon, I., Neuberger, M. S., Buelow, R., & Brüggemann, M. (2013). High-Affinity IgG Antibodies Develop Naturally in Ig-Knockout Rats Carrying Germline Human IgH/Ig/Ig Loci Bearing the Rat CH Region. *The Journal of Immunology*, *190*(4), 1481–1490. https://doi.org/10.4049/jimmunol.1203041

Osterholm, M. T. (2016). Ebola and Zika: Cautionary tales. *Science*, *353*(6304), 1073–1073. https://doi.org/10.1126/science.aai9078

Oyen, D., Torres, J. L., Aoto, P. C., Flores-Garcia, Y., Binter, Š., Pholcharee, T., Carroll, S., Reponen, S., Wash, R., Liang, Q., Lemiale, F., Locke, E., Bradley, A., King, C. R., Emerling, D., Kellam, P., Zavala, F., Ward, A. B., & Wilson, I. A. (2020). Structure and mechanism of monoclonal antibody binding to the junctional epitope of Plasmodium falciparum circumsporozoite protein. *PLOS Pathogens*, *16*(3), e1008373. https://doi.org/10.1371/journal.ppat.1008373

Pandit, D., Tuske, S. J., Coales, S. J., E, S. Y., Liu, A., Lee, J. E., Morrow, J. A., Nemeth, J. F., & Hamuro, Y. (2012). Mapping of discontinuous conformational epitopes by amide hydrogen/deuterium exchange mass spectrometry and computational docking. *Journal of Molecular Recognition*, *25*(3), 114–124. https://doi.org/10.1002/jmr.1169

Pantophlet, R., Trattnig, N., Murrell, S., Lu, N., Chau, D., Rempel, C., Wilson, I. A., & Kosma, P. (2017). Bacterially derived synthetic mimetics of mammalian oligomannose prime antibody responses that neutralize HIV infectivity. *Nature Communications*, *8*(1), 1601. https://doi.org/10.1038/s41467-017-01640-y

Pappas, L., Foglierini, M., Piccoli, L., Kallewaard, N. L., Turrini, F., Silacci, C., Fernandez-Rodriguez, B., Agatic, G., Giacchetto-Sasselli, I., Pellicciotta, G., Sallusto, F., Zhu, Q., Vicenzi, E., Corti, D., & Lanzavecchia, A. (2014).

Rapid development of broadly influenza neutralizing antibodies through redundant mutations. *Nature*, *516*(7531), 418–422. https://doi.org/10.1038/nature13764

Parmley, S. F., & Smith, G. P. (1988). Antibody-selectable filamentous fd phage vectors: Affinity purification of target genes. *Gene*, *73*(2), 305–318. https://doi.org/10.1016/0378-1119(88)90495-7

Parthiban, K., Perera, R. L., Sattar, M., Huang, Y., Mayle, S., Masters, E., Griffiths, D., Surade, S., Leah, R., Dyson, M. R., & McCafferty, J. (2019). A comprehensive search of functional sequence space using large mammalian display libraries created by gene editing. *mAbs*, *11*(5), 884–898. https://doi.org/10.1080/19420862.2019.1618673

Pascal, K. E., Dudgeon, D., Trefry, J. C., Anantpadma, M., Sakurai, Y., Murin, C. D., Turner, H. L., Fairhurst, J., Torres, M., Rafique, A., Yan, Y., Badithe, A., Yu, K., Potocky, T., Bixler, S. L., Chance, T. B., Pratt, W. D., Rossi, F. D., Shamblin, J. D., … Kyratsous, C. A. (2018). Development of Clinical-Stage Human Monoclonal Antibodies That Treat Advanced Ebola Virus Disease in Nonhuman Primates. *The Journal of Infectious Diseases*, *218*(suppl_5), S612–S626. https://doi.org/10.1093/infdis/jiy285

Pedrioli, A., & Oxenius, A. (2021). Single B cell technologies for monoclonal antibody discovery. *Trends in Immunology*, *42*(12), 1143–1158. https://doi.org/10.1016/j.it.2021.10.008

Peng, H.-P., Lee, K. H., Jian, J.-W., & Yang, A.-S. (2014). Origins of specificity and affinity in antibody–protein interactions. *Proceedings of the National Academy of Sciences*, *111*(26), E2656–E2665. https://doi.org/10.1073/pnas.1401131111

Peng, W., Pronker, M. F., & Snijder, J. (2021). Mass Spectrometry-Based De Novo Sequencing of Monoclonal Antibodies Using Multiple Proteases and a Dual Fragmentation Scheme. *Journal of Proteome Research*, *20*(7), 3559–3566. https://doi.org/10.1021/acs.jproteome.1c00169

Pennell, M., Rodriguez, O. L., Watson, C. T., & Greiff, V. (2023). The evolutionary and functional significance of germline immunoglobulin gene vari-

ation. *Trends in Immunology*, *44*(1), 7–21. https://doi.org/10.1016/j.it.2022.11.001

Pickman, Y., Dunn-Walters, D., & Mehr, R. (2013). BCR CDR3 length distributions differ between blood and spleen and between old and young patients, and TCR distributions can be used to detect myelodysplastic syndrome. *Physical Biology*, *10*(5), 056001. https://doi.org/10.1088/1478-3975/10/5/056001

Pierce, B. G., Wiehe, K., Hwang, H., Kim, B.-H., Vreven, T., & Weng, Z. (2014). ZDOCK server: Interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics*, *30*(12), 1771–1773. https://doi.org/10.1093/bioinformatics/btu097

Pittala, S., & Bailey-Kellogg, C. (2019). Learning Context-aware Structural Representations to Predict Antigen and Antibody Binding Interfaces. *bioRxiv*, 658054. https://doi.org/10.1101/658054

Pittala, S., & Bailey-Kellogg, C. (2020). Learning context-aware structural representations to predict antigen and antibody binding interfaces. *Bioinformatics*, *36*(13), 3996–4003. https://doi.org/10.1093/bioinformatics/btaa263

Pollard, A. J., Launay, O., Lelievre, J.-D., Lacabaratz, C., Grande, S., Goldstein, N., Robinson, C., Gaddah, A., Bockstal, V., Wiedemann, A., Leyssen, M., Luhn, K., Richert, L., Bétard, C., Gibani, M. M., Clutterbuck, E. A., Snape, M. D., Levy, Y., Douoguih, M., … Molenberghs, G. (2021). Safety and immunogenicity of a two-dose heterologous Ad26.ZEBOV and MVA-BN-Filo Ebola vaccine regimen in adults in Europe (EBOVAC2): A randomised, observer-blind, participant-blind, placebo-controlled, phase 2 trial. *The Lancet Infectious Diseases*, *21*(4), 493–506. https://doi.org/10.1016/S1473-3099(20)30476-X

Prihoda, D., Maamary, J., Waight, A., Juan, V., Fayadat-Dilman, L., Svozil, D., & Bitton, D. A. (2022). BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs*, *14*(1), 2020203. https://doi.org/10.1080/19420862.2021.2020203

Pushparaj, P., Nicoletto, A., Sheward, D. J., Das, H., Castro Dopico, X., Perez Vidakovics, L., Hanke, L., Chernyshev, M., Narang, S., Kim, S., Fischbach, J., Ekström, S., McInerney, G., Hällberg, B. M., Murrell, B., Corcoran, M., & Karlsson Hedestam, G. B. (2023). Immunoglobulin germline gene polymorphisms influence the function of SARS-CoV-2 neutralizing antibodies. *Immunity*, *56*(1), 193–206.e7. https://doi.org/10.1016/j.immuni.2022.12.005

Qiu, X., Alimonti, J. B., Melito, P. L., Fernando, L., Ströher, U., & Jones, S. M. (2011). Characterization of Zaire ebolavirus glycoprotein-specific monoclonal antibodies. *Clinical Immunology*, *141*(2), 218–227. https://doi.org/10.1016/j.clim.2011.08.008

Qiu, X., Wong, G., Audet, J., Bello, A., Fernando, L., Alimonti, J. B., Fausther-Bovendo, H., Wei, H., Aviles, J., Hiatt, E., Johnson, A., Morton, J., Swope, K., Bohorov, O., Bohorova, N., Goodman, C., Kim, D., Pauly, M. H., Velasco, J., … Kobinger, G. P. (2014). Reversion of advanced Ebola virus disease in nonhuman primates with ZMapp. *Nature*, *514*(7520), 47–53. https://doi.org/10.1038/nature13777

Ralph, D. K., & Iv, F. A. M. (2019). Per-sample immunoglobulin germline inference from B cell receptor deep sequencing data. *PLOS Computational Biology*, *15*(7), e1007133. https://doi.org/10.1371/journal.pcbi.1007133

Ramaraj, T., Angel, T., Dratz, E. A., Jesaitis, A. J., & Mumey, B. (2012). Antigen–antibody interface properties: Composition, residue interactions, and features of 53 non-redundant structures. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1824*(3), 520–532. https://doi.org/10.1016/j.bbapap.2011.12.007

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., & Song, Y. (2019). Evaluating Protein Transfer Learning with TAPE. *Advances in Neural Information Processing Systems*, *32*. Retrieved August 29, 2022, from https://proceedings.neurips.cc/paper/2019/hash/37f65c068b7723cd7809ee2d31d7861c-Abstract.html

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., & Rives, A. (2020). Transformer protein language models are unsupervised structure learners. https://doi.org/10.1101/2020.12.15.422761

Raybould, M. I. J., Kovaltsuk, A., Marks, C., & Deane, C. M. (2021). CoV-AbDab: The coronavirus antibody database. *Bioinformatics*, *37*(5), 734–735. https://doi.org/10.1093/bioinformatics/btaa739

Raybould, M. I. J., Marks, C., Kovaltsuk, A., Lewis, A. P., Shi, J., & Deane, C. M. (2020). Evidence of Antibody Repertoire Functional Convergence through Public Baseline and Shared Response Structures. *bioRxiv*, 2020.03.17.993444. https://doi.org/10.1101/2020.03.17.993444

Raybould, M. I. J., Marks, C., Kovaltsuk, A., Lewis, A. P., Shi, J., & Deane, C. M. (2021). Public Baseline and shared response structures support the theory of antibody repertoire functional commonality. *PLOS Computational Biology*, *17*(3), e1008781. https://doi.org/10.1371/journal.pcbi.1008781

Raybould, M. I. J., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A. P., Bujotzek, A., Shi, J., & Deane, C. M. (2019). Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences*, *116*(10), 4025–4030. https://doi.org/10.1073/pnas.1810576116

Raybould, M. I. J., Marks, C., Lewis, A. P., Shi, J., Bujotzek, A., Taddese, B., & Deane, C. M. (2020). Thera-SAbDab: The Therapeutic Structural Antibody Database. *Nucleic Acids Research*, *48*(D1), D383–D388. https://doi.org/10.1093/nar/gkz827

Raybould, M. I. J., Wong, W. K., & Deane, C. M. (2019). Antibody–antigen complex modelling in the era of immunoglobulin repertoire sequencing. *Molecular Systems Design & Engineering*, *4*(4), 679–688. https://doi.org/10.1039/C9ME00034H

Rebollo, A., Fliedel, L., & Tuffery, P. (2022). PEPscan: A Broad Spectrum Approach for the Characterization of Protein-Binder Interactions? *Biomolecules*, *12*(2), 178. https://doi.org/10.3390/biom12020178

Reddy, S. T., Ge, X., Miklos, A. E., Hughes, R. A., Kang, S. H., Hoi, K. H., Chrysostomou, C., Hunicke-Smith, S. P., Iverson, B. L., Tucker, P. W., Ellington, A. D., & Georgiou, G. (2010). Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nature Biotechnology*, *28*(9), 965–969. https://doi.org/10.1038/nbt.1673

Regep, C., Georges, G., Shi, J., Popovic, B., & Deane, C. M. (2017). The H3 loop of antibodies shows unique structural characteristics. *Proteins: Structure, Function, and Bioinformatics*, *85*(7), 1311–1318. https://doi.org/10.1002/prot.25291

Reusch, D., & Tejada, M. L. (2015). Fc glycans of therapeutic antibodies as critical quality attributes. *Glycobiology*, *25*(12), 1325–1334. https://doi.org/10.1093/glycob/cwv065

Richardson, E. (2020). Re-educating myself about the light chain | Oxford Protein Informatics Group. Retrieved September 22, 2022, from https://www.blopig.com/blog/2020/09/re-educating-myself-about-the-light-chain/

Rijal, P., Elias, S. C., Machado, S. R., Xiao, J., Schimanski, L., O'Dowd, V., Baker, T., Barry, E., Mendelsohn, S. C., Cherry, C. J., Jin, J., Labbé, G. M., Donnellan, F. R., Rampling, T., Dowall, S., Rayner, E., Findlay-Wilson, S., Carroll, M., Guo, J., … Townsend, A. R. (2019). Therapeutic Monoclonal Antibodies for Ebola Virus Infection Derived from Vaccinated Humans. *Cell Reports*, *27*(1), 172–186.e7. https://doi.org/10.1016/j.celrep.2019.03.020

Ripley, B., Venables, B., Bates, D. M., ca 1998), K. H. (p., ca 1998), A. G. (p., & Firth, D. (2022). MASS: Support Functions and Datasets for Venables and Ripley's MASS. Retrieved August 28, 2022, from https://CRAN.R-project.org/package=MASS

Robins, H. (2013). Immunosequencing: Applications of immune repertoire deep sequencing. *Current Opinion in Immunology*, *25*(5), 646–652. https://doi.org/10.1016/j.coi.2013.09.017

Robinson, S. A., Raybould, M. I. J., Schneider, C., Wong, W. K., Marks, C., & Deane, C. M. (2021). Epitope profiling using computational structural modelling demonstrated on coronavirus-binding antibodies. *PLOS Computational Biology*, *17*(12), e1009675. https://doi.org/10.1371/journal.pcbi.1009675

Rosenfeld, A. M., Meng, W., Luning Prak, E. T., & Hershberg, U. (2018). ImmuneDB, a Novel Tool for the Analysis, Storage, and Dissemination of

Immune Repertoire Sequencing Data. *Frontiers in Immunology*, *9*, 2107. https://doi.org/10.3389/fimmu.2018.02107

Roskin, K. M., Simchoni, N., Liu, Y., Lee, J.-Y., Seo, K., Hoh, R. A., Pham, T., Park, J. H., Furman, D., Dekker, C. L., Davis, M. M., James, J. A., Nadeau, K. C., Cunningham-Rundles, C., & Boyd, S. D. (2015). IgH sequences in common variable immune deficiency reveal altered B cell development and selection. *Science translational medicine*, *7*(302), 302ra135. https://doi.org/10.1126/scitranslmed.aab1216

Roy, B., Neumann, R. S., Snir, O., Iversen, R., Sandve, G. K., Lundin, K. E. A., & Sollid, L. M. (2017). High-Throughput Single-Cell Analysis of B Cell Receptor Usage among Autoantigen-Specific Plasma Cells in Celiac Disease. *The Journal of Immunology*, *199*(2), 782–791. https://doi.org/10.4049/jimmunol.1700169

Rubinstein, N. D., Mayrose, I., Halperin, D., Yekutieli, D., Gershoni, J. M., & Pupko, T. (2008). Computational characterization of B-cell epitopes. *Molecular Immunology*, *45*(12), 3477–3489. https://doi.org/10.1016/j.molimm.2007.10.016

Ruffolo, J. A., & Gray, J. J. (2022). Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Biophysical Journal*, *121*(3), 155a–156a. https://doi.org/10.1016/j.bpj.2021.11.1942

Ruffolo, J. A., Gray, J. J., & Sulam, J. (2021). Deciphering antibody affinity maturation with language models and weakly supervised learning. https://doi.org/10.48550/arXiv.2112.07782

Ruffolo, J. A., Guerra, C., Mahajan, S. P., Sulam, J., & Gray, J. J. (2020). Geometric potentials from deep learning improve prediction of CDR H3 loop structures. *Bioinformatics*, *36*(Supplement_1), i268–i275. https://doi.org/10.1093/bioinformatics/btaa457

Ruffolo, J. A., Sulam, J., & Gray, J. J. (2022). Antibody structure prediction using interpretable deep learning. *Patterns*, *3*(2), 100406. https://doi.org/10.1016/j.patter.2021.100406

Ruiz, M., & Lefranc, M.-P. (2002). IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics*, *53*(10), 857–883. https://doi.org/10.1007/s00251-001-0408-6

Ryan, M. H., Petrone, D., Nemeth, J. F., Barnathan, E., Björck, L., & Jordan, R. E. (2008). Proteolysis of purified IgGs by human and bacterial enzymes in vitro and the detection of specific proteolytic fragments of endogenous IgG in rheumatoid synovial fluid. *Molecular Immunology*, *45*(7), 1837–1846. https://doi.org/10.1016/j.molimm.2007.10.043

Safdari, Y., Farajnia, S., Asgharzadeh, M., & Khalili, M. (2013). Antibody humanization methods – a review and update. *Biotechnology and Genetic Engineering Reviews*, *29*(2), 175–186. https://doi.org/10.1080/02648725.2013.801235

Saggy, I., Wine, Y., Shefet-Carasso, L., Nahary, L., Georgiou, G., & Benhar, I. (2012). Antibody isolation from immunized animals: Comparison of phage display and antibody discovery via V gene repertoire mining. *Protein Engineering, Design and Selection*, *25*(10), 539–549. https://doi.org/10.1093/protein/gzs060

Salimi, N., Edwards, L., Foos, G., Greenbaum, J. A., Martini, S., Reardon, B., Shackelford, D., Vita, R., Zalman, L., Peters, B., & Sette, A. (2020). A behind-the-scenes tour of the IEDB curation process: An optimized process empirically integrating automation and human curation efforts. *Immunology*, *161*(2), 139–147. https://doi.org/10.1111/imm.13234

Sangesland, M., Torrents de la Peña, A., Boyoglu-Barnum, S., Ronsard, L., Mohamed, F. A. N., Moreno, T. B., Barnes, R. M., Rohrer, D., Lonberg, N., Ghebremichael, M., Kanekiyo, M., Ward, A., & Lingwood, D. (2022). Allelic polymorphism controls autoreactivity and vaccine elicitation of human broadly neutralizing antibodies against influenza virus. *Immunity*, *55*(9), 1693–1709.e8. https://doi.org/10.1016/j.immuni.2022.07.006

Saphire, E. O. (2019). A glimpse into immune responses evolving against Ebola virus. *Nature Medicine*, *25*(10), 1470–1471. https://doi.org/10.1038/s41591-019-0607-z

Saphire, E. O., Schendel, S. L., Fusco, M. L., Gangavarapu, K., Gunn, B. M., Wec, A. Z., Halfmann, P. J., Brannan, J. M., Herbert, A. S., Qiu, X., Wagh, K., He, S., Giorgi, E. E., Theiler, J., Pommert, K. B. J., Krause, T. B., Turner, H. L., Murin, C. D., Pallesen, J., … Dye, J. M. (2018). Systematic Analy-

sis of Monoclonal Antibodies against Ebola Virus GP Defines Features that Contribute to Protection. *Cell*, *174*(4), 938–952.e13. https://doi.org/10.1016/j.cell.2018.07.033

Scally, S. W., McLeod, B., Bosch, A., Miura, K., Liang, Q., Carroll, S., Reponen, S., Nguyen, N., Giladi, E., Rämisch, S., Yusibov, V., Bradley, A., Lemiale, F., Schief, W. R., Emerling, D., Kellam, P., King, C. R., & Julien, J.-P. (2017). Molecular definition of multiple sites of antibody inhibition of malaria transmission-blocking vaccine antigen Pfs25. *Nature Communications*, *8*(1), 1568. https://doi.org/10.1038/s41467-017-01924-3

Scheepers, C., Shrestha, R. K., Lambson, B. E., Jackson, K. J. L., Wright, I. A., Naicker, D., Goosen, M., Berrie, L., Ismail, A., Garrett, N., Karim, Q. A., Karim, S. S. A., Moore, P. L., Travers, S. A., & Morris, L. (2015). Ability To Develop Broadly Neutralizing HIV-1 Antibodies Is Not Restricted by the Germline Ig Gene Repertoire. *The Journal of Immunology*, *194*(9), 4371–4378. https://doi.org/10.4049/jimmunol.1500118

Scheid, J. F., Mouquet, H., Ueberheide, B., Diskin, R., Klein, F., Oliveira, T. Y. K., Pietzsch, J., Fenyo, D., Abadir, A., Velinzon, K., Hurley, A., Myung, S., Boulad, F., Poignard, P., Burton, D. R., Pereyra, F., Ho, D. D., Walker, B. D., Seaman, M. S., … Nussenzweig, M. C. (2011). Sequence and Structural Convergence of Broad and Potent HIV Antibodies That Mimic CD4 Binding. *Science*, *333*(6049), 1633–1637. https://doi.org/10.1126/science.1207227

Schickel, J.-N., Glauzy, S., Ng, Y.-S., Chamberlain, N., Massad, C., Isnardi, I., Katz, N., Uzel, G., Holland, S. M., Picard, C., Puel, A., Casanova, J.-L., & Meffre, E. (2017). Self-reactive VH4-34–expressing IgG B cells recognize commensal bacteria. *The Journal of Experimental Medicine*, *214*(7), 1991–2003. https://doi.org/10.1084/jem.20160201

Schneider, C., Buchanan, A., Taddese, B., & Deane, C. M. (2021). *DLAB - Deep learning methods for structure-based virtual screening of antibodies* (tech. rep.). https://doi.org/10.1101/2021.02.12.430941

Schneider, C., Raybould, M. I. J., & Deane, C. M. (2022). SAbDab in the age of biotherapeutics: Updates including SAbDab-nano, the nanobody struc-

ture tracker. *Nucleic Acids Research*, *50*(D1), D1368–D1372. https://doi.org/10.1093/nar/gkab1050

Schramm, C. A., & Douek, D. C. (2018). Beyond Hot Spots: Biases in Antibody Somatic Hypermutation and Implications for Vaccine Design. *Frontiers in Immunology*, *9*. Retrieved September 14, 2022, from https://www.frontiersin.org/articles/10.3389/fimmu.2018.01876

Schreyer, A. M., & Blundell, T. L. (2013). CREDO: A structural interactomics database for drug discovery. *Database: The Journal of Biological Databases and Curation*, *2013*, bat049. https://doi.org/10.1093/database/bat049

Schroeder, H. W. (2006). Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Developmental & Comparative Immunology*, *30*(1), 119–135. https://doi.org/10.1016/j.dci.2005.06.006

Schulte, D., Peng, W., & Snijder, J. (2022). Template-Based Assembly of Proteomic Short Reads For De Novo Antibody Sequencing and Repertoire Profiling. *Analytical Chemistry*, *94*(29), 10391–10399. https://doi.org/10.1021/acs.analchem.2c01300

Schwaber, J., Andersen, S., & Nielsen, L. (2019). Shedding light: The importance of reverse transcription efficiency standards in data interpretation. *Biomolecular Detection and Quantification*, *17*, 100077. https://doi.org/10.1016/j.bdq.2018.12.002

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Research*, *33*(suppl_2), W382–W388. https://doi.org/10.1093/nar/gki387

Sela-Culang, I., Benhnia, M. R.-E.-I., Matho, M. H., Kaever, T., Maybeno, M., Schlossman, A., Nimrod, G., Li, S., Xiang, Y., Zajonc, D., Crotty, S., Ofran, Y., & Peters, B. (2014). Using a Combined Computational-Experimental Approach to Predict Antibody-Specific B Cell Epitopes. *Structure*, *22*(4), 646–657. https://doi.org/10.1016/j.str.2014.02.003

Sela-Culang, I., Ofran, Y., & Peters, B. (2015). Antibody specific epitope prediction—emergence of a new paradigm. *Current Opinion in Virology*, *11*, 98–102. https://doi.org/10.1016/j.coviro.2015.03.012

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Setliff, I., McDonnell, W. J., Raju, N., Bombardi, R. G., Murji, A. A., Scheepers, C., Ziki, R., Mynhardt, C., Shepherd, B. E., Mamchak, A. A., Garrett, N., Karim, S. A., Mallal, S. A., Crowe, J. E., Morris, L., & Georgiev, I. S. (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. https://doi.org/10.1016/j.chom.2018.05.001

Setliff, I., Shiakolas, A. R., Pilewski, K. A., Murji, A. A., Mapengo, R. E., Janowska, K., Richardson, S., Oosthuysen, C., Raju, N., Ronsard, L., Kanekiyo, M., Qin, J. S., Kramer, K. J., Greenplate, A. R., McDonnell, W. J., Graham, B. S., Connors, M., Lingwood, D., Acharya, P., … Georgiev, I. S. (2019). High-Throughput Mapping of B Cell Receptor Sequences to Antigen Specificity. *Cell*, *179*(7), 1636–1646.e15. https://doi.org/10.1016/j.cell.2019.11.003

Sevy, A. M., Wu, N. C., Gilchuk, I. M., Parrish, E. H., Burger, S., Yousif, D., Nagel, M. B. M., Schey, K. L., Wilson, I. A., Crowe, J. E., & Meiler, J. (2019). Multistate design of influenza antibodies improves affinity and breadth against seasonal viruses. *Proceedings of the National Academy of Sciences*, *116*(5), 1597–1602. https://doi.org/10.1073/pnas.1806004116

Shapovalov, M. V., & Dunbrack, R. L. (2011). A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure*, *19*(6), 844–858. https://doi.org/10.1016/j.str.2011.03.019

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/nbt1486

Shi, Z., Zhang, Q., Yan, H., Yang, Y., Wang, P., Zhang, Y., Deng, Z., Yu, M., Zhou, W., Wang, Q., Yang, X., Mo, X., Zhang, C., Huang, J., Dai, H., Sun, B.,

Zhao, Y., Zhang, L., Yang, Y.-G., & Qiu, X. (2019). More than one antibody of individual B cells revealed by single-cell immune profiling. *Cell Discovery*, *5*(1), 1–13. https://doi.org/10.1038/s41421-019-0137-3

Shirai, H., Kidera, A., & Nakamura, H. (1996). Structural classification of CDR-H3 in antibodies. *FEBS Letters*, *399*(1-2), 1–8. https://doi.org/10.1016/S0014-5793(96)01252-5

Sircar, A., & Gray, J. J. (2010). SnugDock: Paratope Structural Optimization during Antibody-Antigen Docking Compensates for Errors in Antibody Homology Models. *PLOS Computational Biology*, *6*(1), e1000644. https://doi.org/10.1371/journal.pcbi.1000644

Sirin, S., Apgar, J. R., Bennett, E. M., & Keating, A. E. (2016). AB-Bind: Antibody binding mutational database for computational affinity predictions. *Protein Science*, *25*(2), 393–409. https://doi.org/10.1002/pro.2829

Slabodkin, A., Chernigovskaya, M., Mikocziova, I., Akbar, R., Scheffer, L., Pavlović, M., Bashour, H., Snapkov, I., Mehta, B. B., Weber, C. R., Gutierrez-Marcos, J., Sollid, L. M., Haff, I. H., Sandve, G. K., Robert, P. A., & Greiff, V. (2021). Individualized VDJ recombination predisposes the available Ig sequence space. *Genome Research*, *31*(12), 2209–2224. https://doi.org/10.1101/gr.275373.121

Smakaj, E., Babrak, L., Ohlin, M., Shugay, M., Briney, B., Tosoni, D., Galli, C., Grobelsek, V., D'Angelo, I., Olson, B., Reddy, S., Greiff, V., Trück, J., Marquez, S., Lees, W., & Miho, E. (2020). Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences. *Bioinformatics*, *36*(6), 1731–1739. https://doi.org/10.1093/bioinformatics/btz845

Smith, D. S., Creadon, G., Jena, P. K., Portanova, J. P., Kotzin, B. L., & Wysocki, L. J. (1996). Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells. *The Journal of Immunology*, *156*(7), 2642–2652. Retrieved September 21, 2022, from https://www.jimmunol.org/content/156/7/2642

Smith, G. P. (1985). Filamentous Fusion Phage: Novel Expression Vectors That Display Cloned Antigens on the Virion Surface. *Science*, *228*(4705), 1315–1317. https://doi.org/10.1126/science.4001944

Sok, D., Briney, B., Jardine, J. G., Kulp, D. W., Menis, S., Pauthner, M., Wood, A., Lee, E.-C., Le, K. M., Jones, M., Ramos, A., Kalyuzhniy, O., Adachi, Y., Kubitz, M., MacPherson, S., Bradley, A., Friedrich, G. A., Schief, W. R., & Burton, D. R. (2016). Priming HIV-1 broadly neutralizing antibody precursors in human Ig loci transgenic mice. *Science (New York, N.Y.)*, *353*(6307), 1557–1560. https://doi.org/10.1126/science.aah3945

Soto, C., Bombardi, R. G., Branchizio, A., Kose, N., Matta, P., Sevy, A. M., Sinkovits, R. S., Gilchuk, P., Finn, J. A., & Crowe, J. E. (2019). High frequency of shared clonotypes in human B cell receptor repertoires. *Nature*, *566*(7744), 398–402. https://doi.org/10.1038/s41586-019-0934-8

Soto, C., Finn, J. A., Willis, J. R., Day, S. B., Sinkovits, R. S., Jones, T., Schmitz, S., Meiler, J., Branchizio, A., & Crowe, J. E. (2020). PyIR: A scalable wrapper for processing billions of immunoglobulin and T cell receptor sequences using IgBLAST. *BMC Bioinformatics*, *21*(1), 314. https://doi.org/10.1186/s12859-020-03649-5

Souto-Carneiro, M. M., Sims, G. P., Girschik, H., Lee, J., & Lipsky, P. E. (2005). Developmental Changes in the Human Heavy Chain CDR3. *The Journal of Immunology*, *175*(11), 7425–7436. https://doi.org/10.4049/jimmunol.175.11.7425

Spencer, S., Bethea, D., Raju, T. S., Giles-Komar, J., & Feng, Y. (2012). Solubility evaluation of murine hybridoma antibodies. *mAbs*, *4*(3), 319–325. https://doi.org/10.4161/mabs.19869

Spisak, N., Walczak, A. M., & Mora, T. (2020). Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. *Nucleic Acids Research*, *48*(19), 10702–10712. https://doi.org/10.1093/nar/gkaa825

Stanfield, R. L., Takimoto-Kamimura, M., Rini, J. M., Profy, A. T., & Wilson, I. A. (1993). Major antigen-induced domain rearrangements in an antibody. *Structure*, *1*(2), 83–93. https://doi.org/10.1016/0969-2126(93)90024-B

Stanfield, R. L., Zemla, A., Wilson, I. A., & Rupp, B. (2006). Antibody Elbow Angles are Influenced by their Light Chain Class. *Journal of Molecular Biology*, *357*(5), 1566–1574. https://doi.org/10.1016/j.jmb.2006.01.023

Stapleton, N. M., Andersen, J. T., Stemerding, A. M., Bjarnarson, S. P., Verheul, R. C., Gerritsen, J., Zhao, Y., Kleijer, M., Sandlie, I., de Haas, M., Jonsdottir, I., van der Schoot, C. E., & Vidarsson, G. (2011). Competition for FcRn-mediated transport gives rise to short half-life of human IgG3 and offers therapeutic potential. *Nature Communications*, *2*(1), 599. https://doi.org/10.1038/ncomms1608

Stein, A., & Kortemme, T. (2013). Improvements to Robotics-Inspired Conformational Sampling in Rosetta. *PLOS ONE*, *8*(5), e63090. https://doi.org/10.1371/journal.pone.0063090

Stern, J. N. H., Yaari, G., Heiden, J. A. V., Church, G., Donahue, W. F., Hintzen, R. Q., Huttner, A. J., Laman, J. D., Nagra, R. M., Nylander, A., Pitt, D., Ramanan, S., Siddiqui, B. A., Vigneault, F., Kleinstein, S. H., Hafler, D. A., & O'Connor, K. C. (2014). B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Science Translational Medicine*, *6*(248), 248ra107–248ra107. https://doi.org/10.1126/scitranslmed.3008879

Sun, A., & Benet, L. Z. (2020). Late-Stage Failures of Monoclonal Antibody Drugs: A Retrospective Case Study Analysis. *Pharmacology*, *105*(3-4), 145–163. https://doi.org/10.1159/000505379

Sun, H., Ma, L., Wang, L., Xiao, P., Li, H., Zhou, M., & Song, D. (2021). Research advances in hydrogen–deuterium exchange mass spectrometry for protein epitope mapping. *Analytical and Bioanalytical Chemistry*, *413*(9), 2345–2359. https://doi.org/10.1007/s00216-020-03091-9

Syahir, A., Usui, K., Tomizaki, K.-y., Kajikawa, K., & Mihara, H. (2015). Label and Label-Free Detection Techniques for Protein Microarrays. *Microarrays*, *4*(2), 228–244. https://doi.org/10.3390/microarrays4020228

Tang, C., Krantsevich, A., & MacCarthy, T. (2022). Deep learning model of somatic hypermutation reveals importance of sequence context beyond hotspot targeting. *iScience*, *25*(1), 103668. https://doi.org/10.1016/j.isci.2021.103668

Temme, J. S., Butler, D. L., & Gildersleeve, J. C. (2021). Anti-glycan antibodies: Roles in human disease. *Biochemical Journal*, *478*(8), 1485–1509. https://doi.org/10.1042/BCJ20200610

Teplyakov, A., Obmolova, G., Malia, T. J., Luo, J., Muzammil, S., Sweet, R., Almagro, J. C., & Gilliland, G. L. (2016). Structural diversity in a human antibody germline library. *mAbs*, *8*(6), 1045–1063. https://doi.org/10.1080/19420862.2016.1190060

Thomson, C., Wang, Y., Jackson, L., Olson, M., Wang, W., Liavonchanka, A., Keleta, L., Silva, V., Diederich, S., Jones, R., Gubbay, J., Pasick, J., Petric, M., Jean, F., Allen, V., Brown, E., Rini, J., & Schrader, J. (2012). Pandemic H1N1 Influenza Infection and Vaccination in Humans Induces Cross-Protective Antibodies that Target the Hemagglutinin Stem. *Frontiers in Immunology*, *3*. Retrieved January 13, 2023, from https://www.frontiersin.org/articles/10.3389/fimmu.2012.00087

Tian, X., Li, C., Wu, Y., & Ying, T. (2020). Deep Mining of Human Antibody Repertoires: Concepts, Methodologies, and Applications. *Small Methods*, *4*(10), 2000451. https://doi.org/10.1002/smtd.202000451

Townsend, C. L., Laffy, J. M. J., Wu, Y.-C. B., Silva O'Hare, J., Martin, V., Kipling, D., Fraternali, F., & Dunn-Walters, D. K. (2016). Significant Differences in Physicochemical Properties of Human Immunoglobulin Kappa and Lambda CDR3 Regions. *Frontiers in Immunology*, *7*. Retrieved May 26, 2022, from https://www.frontiersin.org/article/10.3389/fimmu.2016.00388

Towse, C.-L., Rysavy, S. J., Vulovic, I. M., & Daggett, V. (2016). New Dynamic Rotamer Libraries: Data-Driven Analysis of Side-Chain Conformational Propensities. *Structure*, *24*(1), 187–199. https://doi.org/10.1016/j.str.2015.10.017

Tsioris, K., Gupta, N. T., Ogunniyi, A. O., Zimnisky, R. M., Qian, F., Yao, Y., Wang, X., Stern, J. N. H., Chari, R., Briggs, A. W., Clouser, C. R., Vigneault, F., Church, G. M., Garcia, M. N., Murray, K. O., Montgomery, R. R., Kleinstein, S. H., & Love, J. C. (2015). Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. *Integrative Biology*, *7*(12), 1587–1597. https://doi.org/10.1039/c5ib00169b

Tubiana, J., Schneidman-Duhovny, D., & Wolfson, H. J. (2022). ScanNet: An interpretable geometric deep learning model for structure-based protein

binding site prediction. *Nature Methods*, *19*(6), 730–739. https://doi.org/10.1038/s41592-022-01490-7

Valdés-Alemán, J., Téllez-Sosa, J., Ovilla-Muñoz, M., Godoy-Lozano, E., Velázquez-Ramírez, D., Valdovinos-Torres, H., Gómez-Barreto, R. E., & Martinez-Barnetche, J. (2014). Hybridization-based antibody cDNA recovery for the production of recombinant antibodies identified by repertoire sequencing. *mAbs*, *6*(2), 493–501. https://doi.org/10.4161/mabs.27435

Valente, A. P., & Manzano-Rendeiro, M. (2021). Mapping conformational epitopes by NMR spectroscopy. *Current Opinion in Virology*, *49*, 1–6. https://doi.org/10.1016/j.coviro.2021.04.001

Van Kerkhove, M. D., Bento, A. I., Mills, H. L., Ferguson, N. M., & Donnelly, C. A. (2015). A review of epidemiological parameters from Ebola outbreaks to inform early public health decision-making. *Scientific Data*, *2*(1), 150019. https://doi.org/10.1038/sdata.2015.19

Vander Heiden, J. A., Yaari, G., Uduman, M., Stern, J. N. H., O'Connor, K. C., Hafler, D. A., Vigneault, F., & Kleinstein, S. H. (2014). pRESTO: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, *30*(13), 1930–1932. https://doi.org/10.1093/bioinformatics/btu138

Vargas-Madrazo, E., & Paz-García, E. (2003). An improved model of association for VH–VL immunoglobulin domains: Asymmetries between VH and VL in the packing of some interface residues. *Journal of Molecular Recognition*, *16*(3), 113–120. https://doi.org/10.1002/jmr.613

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. Retrieved August 29, 2022, from https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vidarsson, G., Dekkers, G., & Rispens, T. (2014). IgG Subclasses and Allotypes: From Structure to Effector Functions. *Frontiers in Immunology*, *5*. Retrieved September 25, 2022, from https://www.frontiersin.org/articles/10.3389/fimmu.2014.00520

Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., & Rajani, N. F. (2021). BERTology Meets Biology: Interpreting Attention in Protein Language Models. https://doi.org/10.48550/arXiv.2006.15222

Walker, L. M., Phogat, S. K., Chan-Hui, P.-Y., Wagner, D., Phung, P., Goss, J. L., Wrin, T., Simek, M. D., Fling, S., Mitcham, J. L., Lehrman, J. K., Priddy, F. H., Olsen, O. A., Frey, S. M., Hammond, P. W., Protocol G Principal Investigators, Kaminsky, S., Zamb, T., Moyle, M., … Burton, D. R. (2009). Broad and Potent Neutralizing Antibodies from an African Donor Reveal a New HIV-1 Vaccine Target. *Science*, *326*(5950), 285–289. https://doi.org/10.1126/science.1178746

Walker, L. M., Shiakolas, A. R., Venkat, R., Liu, Z. A., Wall, S., Raju, N., Pilewski, K. A., Setliff, I., Murji, A. A., Gillespie, R., Makoah, N. A., Kanekiyo, M., Connors, M., Morris, L., & Georgiev, I. S. (2022). High-Throughput B Cell Epitope Determination by Next-Generation Sequencing. *Frontiers in Immunology*, *13*. Retrieved September 5, 2022, from https://www.frontiersin.org/articles/10.3389/fimmu.2022.855772

Walls, A. C., Fiala, B., Schäfer, A., Wrenn, S., Pham, M. N., Murphy, M., Tse, L. V., Shehata, L., O'Connor, M. A., Chen, C., Navarro, M. J., Miranda, M. C., Pettie, D., Ravichandran, R., Kraft, J. C., Ogohara, C., Palser, A., Chalk, S., Lee, E.-C., … King, N. P. (2020). Elicitation of Potent Neutralizing Antibody Responses by Designed Protein Nanoparticle Vaccines for SARS-CoV-2. *Cell*, *183*(5), 1367–1382.e17. https://doi.org/10.1016/j.cell.2020.10.043

Wang, B., Kluwe, C. A., Lungu, O. I., DeKosky, B. J., Kerr, S. A., Johnson, E. L., Tanno, H., Lee, C.-H., Jung, J., Rezigh, A. B., Carroll, S. M., Reyes, A. N., Bentz, J. R., Villanueva, I., Altman, A. L., Davey, R. A., Ellington, A. D., & Georgiou, G. (2015). Facile Discovery of a Diverse Panel of Anti-Ebola Virus Antibodies by Immune Repertoire Mining. *Scientific Reports*, *5*(1), 13926. https://doi.org/10.1038/srep13926

Wang, C., Liu, Y., Xu, L. T., Jackson, K. J. L., Roskin, K. M., Pham, T. D., Laserson, J., Marshall, E. L., Seo, K., Lee, J.-Y., Furman, D., Koller, D., Dekker, C. L., Davis, M. M., Fire, A. Z., & Boyd, S. D. (2014). Effects of Aging, Cytomegalovirus Infection, and EBV Infection on Human B Cell Reper-

toires. *The Journal of Immunology*, *192*(2), 603–611. https://doi.org/10.4049/jimmunol.1301384

Wang, H., Shi, Y., Song, J., Qi, J., Lu, G., Yan, J., & Gao, G. F. (2016). Ebola Viral Glycoprotein Bound to Its Endosomal Receptor Niemann-Pick C1. *Cell*, *164*(1), 258–268. https://doi.org/10.1016/j.cell.2015.12.044

Wang, Y., Jackson, K. J. L., Sewell, W. A., & Collins, A. M. (2008). Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. *Immunology & Cell Biology*, *86*(2), 111–115. https://doi.org/10.1038/sj.icb.7100144

Watson, C. T., Steinberg, K. M., Graves, T. A., Warren, R. L., Malig, M., Schein, J., Wilson, R. K., Holt, R. A., Eichler, E. E., & Breden, F. (2015). Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. *Genes & Immunity*, *16*(1), 24–34. https://doi.org/10.1038/gene.2014.56

Watson, C. T., Glanville, J., & Marasco, W. A. (2017). The Individual and Population Genetics of Antibody Immunity. *Trends in Immunology*, *38*(7), 459–470. https://doi.org/10.1016/j.it.2017.04.003

Watson, C. T., Kos, J. T., Gibson, W. S., Newman, L., Deikus, G., Busse, C. E., Smith, M. L., Jackson, K. J., & Collins, A. M. (2019). A comparison of immunoglobulin IGHV, IGHD and IGHJ genes in wild-derived and classical inbred mouse strains. *Immunology & Cell Biology*, *97*(10), 888–901. https://doi.org/10.1111/imcb.12288

Watson, C. T., Steinberg, K. M., Huddleston, J., Warren, R. L., Malig, M., Schein, J., Willsey, A. J., Joy, J. B., Scott, J. K., Graves, T. A., Wilson, R. K., Holt, R. A., Eichler, E. E., & Breden, F. (2013). 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*, *92*(4), 530–546. https://doi.org/10.1016/j.ajhg.2013.03.004

Wec, A. Z., Bornholdt, Z. A., He, S., Herbert, A. S., Goodwin, E., Wirchnianski, A. S., Gunn, B. M., Zhang, Z., Zhu, W., Liu, G., Abelson, D. M., Moyer, C. L., Jangra, R. K., James, R. M., Bakken, R. R., Bohorova, N., Bohorov, O., Kim, D. H., Pauly, M. H., … Chandran, K. (2019). Development of

a human antibody cocktail that deploys multiple functions to confer pan-ebolavirus protection. *Cell host & microbe*, *25*(1), 39–48.e5. https://doi.org/10.1016/j.chom.2018.12.004

Wedemeyer, W. J., & Baker, D. (2003). Efficient minimization of angle-dependent potentials for polypeptides in internal coordinates. *Proteins: Structure, Function, and Bioinformatics*, *53*(2), 262–272. https://doi.org/10.1002/prot.10525

Weitzner, B. D., Dunbrack, R. L., & Gray, J. J. (2015). The Origin of CDR H3 Structural Diversity. *Structure*, *23*(2), 302–311. https://doi.org/10.1016/j.str.2014.11.010

Weitzner, B. D., Jeliazkov, J. R., Lyskov, S., Marze, N., Kuroda, D., Frick, R., Adolf-Bryfogle, J., Biswas, N., Dunbrack, R. L., & Gray, J. J. (2017). Modeling and docking of antibody structures with Rosetta. *Nature Protocols*, *12*(2), 401–416. https://doi.org/10.1038/nprot.2016.180

Weitzner, B. D., Kuroda, D., Marze, N., Xu, J., & Gray, J. J. (2014). Blind prediction performance of RosettaAntibody 3.0: Grafting, relaxation, kinematic loop modeling, and full CDR optimization. *Proteins: Structure, Function, and Bioinformatics*, *82*(8), 1611–1623. https://doi.org/10.1002/prot.24534

Westwood, O. M. R., & Hay, F. C. (2001). *Epitope Mapping: A Practical Approach*. Oxford University Press.

Whitty, C. J. M. (2017). The contribution of biological, mathematical, clinical, engineering and social sciences to combatting the West African Ebola epidemic. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1721), 20160293. https://doi.org/10.1098/rstb.2016.0293

Wilson, I. A., & Stanfield, R. L. (1994). Antibody-antigen interactions: New structures and new conformational changes. *Current Opinion in Structural Biology*, *4*(6), 857–867. https://doi.org/10.1016/0959-440X(94)90267-4

Wilson, J. A., Hevey, M., Bakken, R., Guest, S., Bray, M., Schmaljohn, A. L., & Hart, M. K. (2000). Epitopes Involved in Antibody-Mediated Protection from Ebola Virus. *Science*, *287*(5458), 1664–1666. https://doi.org/10.1126/science.287.5458.1664

Winter, G., & Milstein, C. (1991). Man-made antibodies. *Nature*, *349*(6307), 293–299. https://doi.org/10.1038/349293a0

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wong, W. K., Georges, G., Ros, F., Kelm, S., Lewis, A. P., Taddese, B., Leem, J., & Deane, C. M. (2019). SCALOP: Sequence-based antibody canonical loop structure annotation. *Bioinformatics*, *35*(10), 1774–1776. https://doi.org/10.1093/bioinformatics/bty877

Wong, W. K., Robinson, S. A., Bujotzek, A., Georges, G., Lewis, A. P., Shi, J., Snowden, J., Taddese, B., & Deane, C. M. (2021). Ab-Ligity: Identifying sequence-dissimilar antibodies that bind to the same epitope. *mAbs*, *13*(1), 1873478. https://doi.org/10.1080/19420862.2021.1873478

Woolsey, C., & Geisbert, T. W. (2021). Current state of Ebola virus vaccines: A snapshot. *PLOS Pathogens*, *17*(12), e1010078. https://doi.org/10.1371/journal.ppat.1010078

Wu, H., Zhou, Z., Xie, S., Yan, R., Gong, M., Tian, X., & Wang, Z. (2022). Similarity measurements of B cell receptor repertoire in baseline mice showed spectrum convergence of IgM. *BMC Immunology*, *23*(1), 11. https://doi.org/10.1186/s12865-022-00482-8

Wu, T. T., & Kabat, E. A. (1970). AN ANALYSIS OF THE SEQUENCES OF THE VARIABLE REGIONS OF BENCE JONES PROTEINS AND MYELOMA LIGHT CHAINS AND THEIR IMPLICATIONS FOR ANTIBODY COMPLEMENTARITY. *Journal of Experimental Medicine*, *132*(2), 211–250. https://doi.org/10.1084/jem.132.2.211

Xu, J. (2019). Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences*, *116*(34), 16856–16865. https://doi.org/10.1073/pnas.1821309116

Xu, J. L., & Davis, M. M. (2000). Diversity in the CDR3 Region of VH Is Sufficient for Most Antibody Specificities. *Immunity*, *13*(1), 37–45. https://doi.org/10.1016/S1074-7613(00)00006-6

Yaari, G., & Kleinstein, S. H. (2015). Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Medicine*, *7*(1), 121. https://doi.org/10.1186/s13073-015-0243-2

Yaari, G., Vander Heiden, J., Uduman, M., Gadala-Maria, D., Gupta, N., Stern, J., O'Connor, K., Hafler, D., Laserson, U., Vigneault, F., & Kleinstein, S. (2013). Models of Somatic Hypermutation Targeting and Substitution Based on Synonymous Mutations from High-Throughput Immunoglobulin Sequencing Data. *Frontiers in Immunology*, *4*. Retrieved September 14, 2022, from https://www.frontiersin.org/articles/10.3389/fimmu.2013.00358

Yamashita, K., Ikeda, K., Amada, K., Liang, S., Tsuchiya, Y., Nakamura, H., Shirai, H., & Standley, D. M. (2014). Kotai Antibody Builder: Automated high-resolution structural modeling of antibodies. *Bioinformatics*, *30*(22), 3279–3280. https://doi.org/10.1093/bioinformatics/btu510

Yan, Q., He, P., Huang, X., Luo, K., Zhang, Y., Yi, H., Wang, Q., Li, F., Hou, R., Fan, X., Li, P., Liu, X., Liang, H., Deng, Y., Chen, Z., Chen, Y., Mo, X., Feng, L., Xiong, X., … Chen, L. (2021). Germline IGHV3-53-encoded RBD-targeting neutralizing antibodies are commonly present in the antibody repertoires of COVID-19 patients. *Emerging Microbes & Infections*, *10*(1), 1097–1111. https://doi.org/10.1080/22221751.2021.1925594

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, *117*(3), 1496–1503. https://doi.org/10.1073/pnas.1914677117

Ye, J., Ma, N., Madden, T. L., & Ostell, J. M. (2013). IgBLAST: An immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, *41*(W1), W34–W40. https://doi.org/10.1093/nar/gkt382

Yeung, Y. A., Foletti, D., Deng, X., Abdiche, Y., Strop, P., Glanville, J., Pitts, S., Lindquist, K., Sundar, P. D., Sirota, M., Hasa-Moreno, A., Pham, A., Melton Witt, J., Ni, I., Pons, J., Shelton, D., Rajpal, A., & Chaparro-

Riggers, J. (2016). Germline-encoded neutralization of a Staphylococcus aureus virulence factor by the human antibody repertoire. *Nature Communications*, *7*(1), 13376. https://doi.org/10.1038/ncomms13376

Yin, R., Feng, B. Y., Varshney, A., & Pierce, B. G. (2022). Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Science*, *31*(8), e4379. https://doi.org/10.1002/pro.4379

Yoon, H., Macke, J., West, A. P., Foley, B., Bjorkman, P. J., Korber, B., & Yusim, K. (2015). CATNAP: A tool to compile, analyze and tally neutralizing antibody panels. *Nucleic Acids Research*, *43*(W1), W213–W219. https://doi.org/10.1093/nar/gkv404

Yu, X., McGraw, P. A., House, F. S., & Crowe, J. E. (2008). An optimized electrofusion-based protocol for generating virus-specific human monoclonal antibodies. *Journal of Immunological Methods*, *336*(2), 142–151. https://doi.org/10.1016/j.jim.2008.04.008

Zhang, C. (2012). Hybridoma Technology for the Generation of Monoclonal Antibodies. In G. Proetzel & H. Ebersbach (Eds.), *Antibody Methods and Protocols* (pp. 117–135). Humana Press. https://doi.org/10.1007/978-1-61779-931-0_7

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., & Wang, J. (2019). Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell*, *73*(1), 130–142.e5. https://doi.org/10.1016/j.molcel.2018.10.020

Zhao, H., Xu, L., Bombardi, R., Nargi, R., Deng, Z., Errico, J. M., Nelson, C. A., Dowd, K. A., Pierson, T. C., Crowe, J. E., Jr., Diamond, M. S., & Fremont, D. H. (2019). Mechanism of differential Zika and dengue virus neutralization by a public antibody lineage targeting the DIII lateral ridge. *Journal of Experimental Medicine*, *217*(2), e20191792. https://doi.org/10.1084/jem.20191792

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H.

(2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 14049. https://doi.org/10.1038/ncomms14049

Zhou, H., Chen, Y., Zhang, S., Niu, P., Qin, K., Jia, W., Huang, B., Zhang, S., Lan, J., Zhang, L., Tan, W., & Wang, X. (2019). Structural definition of a neutralization epitope on the N-terminal domain of MERS-CoV spike glycoprotein. *Nature Communications*, *10*(1), 3068. https://doi.org/10.1038/s41467-019-10897-4

Zhou, J. Q., & Kleinstein, S. H. (2019). Cutting Edge: Ig H Chains Are Sufficient to Determine Most B Cell Clonal Relationships. *The Journal of Immunology*, *203*(7), 1687–1692. https://doi.org/10.4049/jimmunol.1900666

Zhou, J. Q., & Kleinstein, S. H. (2020). Position-Dependent Differential Targeting of Somatic Hypermutation. *The Journal of Immunology*, *205*(12), 3468–3479. https://doi.org/10.4049/jimmunol.2000496

Zhou, T., Georgiev, I., Wu, X., Yang, Z.-Y., Dai, K., Finzi, A., Do Kwon, Y., Scheid, J. F., Shi, W., Xu, L., Yang, Y., Zhu, J., Nussenzweig, M. C., Sodroski, J., Shapiro, L., Nabel, G. J., Mascola, J. R., & Kwong, P. D. (2010). Structural Basis for Broad and Potent Neutralization of HIV-1 by Antibody VRC01. *Science*, *329*(5993), 811–817. https://doi.org/10.1126/science.1192819

Zhou, T., Lynch, R. M., Chen, L., Acharya, P., Wu, X., Doria-Rose, N. A., Joyce, M. G., Lingwood, D., Soto, C., Bailer, R. T., Ernandes, M. J., Kong, R., Longo, N. S., Louder, M. K., McKee, K., O'Dell, S., Schmidt, S. D., Tran, L., Yang, Z., … Kwong, P. D. (2015). Structural Repertoire of HIV-1-Neutralizing Antibodies Targeting the CD4 Supersite in 14 Donors. *Cell*, *161*(6), 1280–1292. https://doi.org/10.1016/j.cell.2015.05.007

Zhu, F., Nair, R. R., Fisher, E. M. C., & Cunningham, T. J. (2019). Humanising the mouse genome piece by piece. *Nature Communications*, *10*(1), 1845. https://doi.org/10.1038/s41467-019-09716-7

Zhu, J., Wu, X., Zhang, B., McKee, K., O'Dell, S., Soto, C., Zhou, T., Casazza, J. P., NISC Comparative Sequencing Program, Mullikin, J. C., Kwong, P. D., Mascola, J. R., Shapiro, L., Becker, J., Benjamin, B., Blakesley, R., Bouffard, G., Brooks, S., Coleman, H., … Young, A. (2013). De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-

generation sequencing of B-cell transcripts. *Proceedings of the National Academy of Sciences*, *110*(43), E4088–E4097. https://doi.org/10.1073/pnas.1306262110

Zhu, J., Ofek, G., Yang, Y., Zhang, B., Louder, M. K., Lu, G., McKee, K., Pancera, M., Skinner, J., Zhang, Z., Parks, R., Eudailey, J., Lloyd, K. E., Blinn, J., Alam, S. M., Haynes, B. F., Simek, M., Burton, D. R., Koff, W. C., … Kwong, P. D. (2013). Mining the antibodyome for HIV-1–neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proceedings of the National Academy of Sciences*, *110*(16), 6470–6475. https://doi.org/10.1073/pnas.1219320110

Zhu, K., Day, T., Warshaviak, D., Murrett, C., Friesner, R., & Pearlman, D. (2014). Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins: Structure, Function, and Bioinformatics*, *82*(8), 1646–1655. https://doi.org/10.1002/prot.24551

Zhu, X., Yu, F., Wu, Y., & Ying, T. (2021). Potent germline-like monoclonal antibodies: Rapid identification of promising candidates for antibody-based antiviral therapy. *Antibody Therapeutics*, *4*(2), 89–98. https://doi.org/10.1093/abt/tbab008

Zierep, P. F., Vita, R., Blazeska, N., Moumbock, A. F. A., Greenbaum, J. A., Peters, B., & Günther, S. (2022). Towards the prediction of non-peptidic epitopes. *PLOS Computational Biology*, *18*(2), e1009151. https://doi.org/10.1371/journal.pcbi.1009151

Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A. M., & Mazutis, L. (2017). Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, *12*(1), 44–73. https://doi.org/10.1038/nprot.2016.154

Zucha, D., Androvic, P., Kubista, M., & Valihrach, L. (2020). Performance Comparison of Reverse Transcriptases for Single-Cell Studies. *Clinical Chemistry*, *66*(1), 217–228. https://doi.org/10.1373/clinchem.2019.307835

Zuo, T., Shi, X., Liu, Z., Guo, L., Zhao, Q., Guan, T., Pan, X., Jia, N., Cao, W., Zhou, B., Goldin, M., & Zhang, L. (2011). Comprehensive Analysis of Pathogen-specific Antibody Response in Vivo Based on an Antigen

Library Displayed on Surface of Yeast. *Journal of Biological Chemistry*, *286*(38), 33511–33519. https://doi.org/10.1074/jbc.M111.270553

Acknowledgements

I have so many people to thank for my PhD. Firstly, my supervisors: to Professor Charlotte M. Deane and to Dr. Dominic Kelly for their invaluable guidance. To my Kymab supervisors, Dr. Jacob Galson and Professor Paul Kellam, who were never short of insights, clever ideas I wouldn't have thought of, and enthusiasm, and who were critical for multiple steps in the work presented, including working at Kymab for 2 months which was a formative experience.

Besides my supervisors, I want to thank Dr. Sagida Bibi and Dr. Simon Watson in particular: Sagida for performing the experiments for the Ebola work (including teaching me RNA extraction), and Simon for his role in the Kymouse work. Furthermore, I would like to thank Professors John Frater and Lynn Dustin for including me in their single-cell BCR sequencing collaboration. I couldn't have asked for a better set of supervisors and collaborators.

Many thanks to my colleagues and friends. Above all, thanks to my PhD soulmate Sarah Robinson, I can honestly say I couldn't have done this without you. Thank you to Alissa Hummer for your continual support and being such an inspiration to me. I can't wait to see the amazing stuff you will both continue to do!

I would also like to thank my undergraduate tutor and research supervisor Professor Graham K. Taylor. Without his mentorship, I wouldn't have had the confidence to put myself forward for this PhD.

I would not be here without my family and my partner, who have been a huge support to me through this process. Thanks from the bottom of my heart to my parents and my sisters, to my Mum in particular who this thesis is dedicated to - I wouldn't have got to either the starting or finishing line without you. Thank you to my partner Christian for his patience, support and belief in me. Finally, thank you to Bella and Sooty Richardson and Arnie Mustaklem for

being responsible for my happiness.