Immune Transcriptome and B cell receptor repertoire in COVID-19

Prasanti Kotagiri

Hughes Hall, University of Cambridge

Department of Medicine

June 2022

Supervisors: Professor Kenneth GC Smith & Dr Paul A Lyons

This dissertation is submitted for the degree of Doctor of Philosophy.

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

The dissertation does not exceed the prescribed word limit specified by the Degree Committee for the Faculties of Clinical Medicine and Veterinary Medicine.

Prasanti Kotagiri
December 2021

ABSTRACT

Immune Transcriptome and B cell receptor repertoire in COVID-19

Prasanti Kotagiri

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) resulting in Coronavirus disease 2019 (COVID-19) was declared by the World Health Organization a global pandemic on March 11, 2020. SARS-CoV-2 primarily infects respiratory epithelial cells, and results in a range of clinical manifestations from asymptomatic disease to multi-organ failure.

We studied the immune response of SARS-CoV-2-infected individuals with a range of severities followed over 2- 6 months from symptom onset. We undertook deep immune-phenotyping and transcriptomic analysis. We demonstrated that an early robust immune response, without systemic inflammation, was characteristic of asymptomatic or mild disease. Immune recovery was complex, with profound persistent cellular abnormalities correlating with a change in the nature of the inflammatory response, where signatures characteristic of increased oxidative phosphorylation and reactive-oxygen species-associated inflammation replace those driven by TNF and IL-6.

In addition, we performed B cell receptor repertoire analysis of SARS-CoV-2 infected individuals and recipients of SARS-CoV-2 vaccine. B cells play a central role in the immune response to both SARS-CoV-2 infection and vaccination. We found marked differences in the global BCR repertoire after natural infection compared to vaccination. Following infection, the proportion of BCRs bearing IgG1/3 and IgA1 isotypes increased, somatic hypermutation (SHM) was markedly decreased and, in patients with severe disease, expansion of IgM and IgA clones were observed. In contrast, after vaccination the proportion of BCRs bearing IgD/M isotypes increased, SHM was unchanged and expansion of IgG clones was prominent.

Infection generated a broad distribution of SARS-CoV-2-specific clones predicted to target the spike protein whilst vaccination produced a more focused response mainly targeting the spike's receptor-binding domain. These findings offer insights into how different immune exposure to SARS-CoV-2 impacts upon BCR repertoire development, potentially informing vaccine strategies.

ACKNOWLEDGEMENTS

I would like to express my deep and sincere gratitude to my supervisors Professor Kenneth Smith and Dr Paul Lyons for giving me this research opportunity and providing me with invaluable guidance throughout my PhD study. Their vision, patience and enthusiasm have deeply inspired me. It was a great privilege and honour to be able to work with them.

My sincere thanks go to Prof Eoin McKinney, A/Prof Ravi Gupta, Prof John Bradley and Dr Rachael Bashford-Rogers for giving me the opportunity to collaborate on exciting projects.

I would like to thank my fellow colleagues in the Smith lab: Dr Will Rae, Dr John Sowerby, Dr Laura Bergamaschi, Dr Limy Wong, Sofie Tolmeijer, Diana Pombal, Dr Federica Mescia, Maddie Epping and Johanna Jung for stimulating discussion and insightful feedback.

Lastly, I would like to thank my family for their love and support.

This work was funded by the Jacquot Research Entry Fellowship.

PUBLICATIONS

Publications relevant to thesis

- Bergamaschi L*, Mescia F*, Turner L*, Hanson AL*, **Kotagiri P***, Dunmore BJ, Ruffieux H, De Sa A, Huhn O, Morgan MD, Gerber PP, Wills MR, Baker S, Calero-Nieto FJ, Doffinger R, Dougan G, Elmer A, Goodfellow IG, Gupta RK, Hosmillo M, Hunter K, Kingston N, Lehner PJ, Matheson NJ, Nicholson JK, Petrunkina AM, Richardson S, Saunders C, Thaventhiran JED, Toonen EJM, Weekes MP; Cambridge Institute of Therapeutic Immunology and Infectious Disease-National Institute of Health Research (CITIID-NIHR) COVID BioResource Collaboration, Göttgens B, Toshner M, Hess C, Bradley JR, Lyons PA, Smith KGC. Longitudinal analysis reveals that delayed bystander CD8+ T cell activation and early immune pathology distinguish severe COVID-19 from mild disease. **Immunity. 2021** Jun 8;54(6):1257-1275.e8. **Chapter 3)**

- Wang J, **Kotagiri P**, Lyons PA, Al-Lamki RS, Mescia F, Bergamaschi L, Turner L, Morgan MD, Calero-Nieto FJ, Bach K, Mende N, Wilson NK, Watts ER; Cambridge Institute of Therapeutic Immunology and Infectious Disease-National Institute of Health Research (CITIID-NIHR) Covid BioResource Collaboration, Maxwell PH, Chinnery PF, Kingston N, Papadia S, Stirrups KE, Walker N, Gupta RK, Menon DK, Allinson K, Aitken SJ, Toshner M, Weekes MP, Nathan JA, Walmsley SR, Ouwehand WH, Kasanicki M, Göttgens B, Marioni JC, Smith KGC, Pober JS, Bradley JR. Coagulation factor V is a T-cell inhibitor expressed by leukocytes in COVID-19. **iScience. 2022** Mar 18;25(3):103971. **(Chapter3)**

- Collier DA*, Ferreira IATM*, **Kotagiri P***, Datir RP*, Lim EY*, Touizer E, Meng B, Abdullahi A; CITIID-NIHR BioResource COVID-19 Collaboration, Elmer A, Kingston N, Graves B, Le Gresley E, Caputo D, Bergamaschi L, Smith KGC, Bradley JR, Ceron-Gutierrez L, Cortes-Acevedo P, Barcenas-Morales G, Linterman MA, McCoy LE, Davis C, Thomson E, Lyons PA, McKinney E, Doffinger R, Wills M, Gupta RK. Age-related immune response heterogeneity to SARS-CoV-2 vaccine BNT162b2. **Nature. 2021** Aug;596(7872):417-422. **Chapter 4)**

- **Kotagiri P**, Mescia F, Rae WM, Bergamaschi L, Tuong ZK, Turner L, Hunter K, Gerber PP, Hosmillo M; Cambridge Institute of Therapeutic Immunology and Infectious Disease-National Institute of Health Research (CITIID-NIHR) COVID BioResource Collaboration, Hess C, Clatworthy MR, Goodfellow IG, Matheson NJ, McKinney EF, Wills MR, Gupta RK, Bradley JR, Bashford-Rogers RJM, Lyons PA, Smith KGC. B cell receptor repertoire kinetics after SARS-CoV-2 infection and vaccination. **Cell Rep. 2022** Feb 15;38(7):110393. **Chapter 4)**

- Sowerby JM*, **Kotagiri P***…Smith KGC, McKinney EF, "Multi-Omic analysis identifies metabolic enhancement of immune memory by lysine deacetylase inhibition during immunisation and infection" **(in submission) (Chapter 4)**

- **Kotagiri P***, Rae W*...Smith K, McKinney E, Lyons P, "Shared hypermutated B cell clones in Crohn's disease" (**in submission**) **(Chapter 5)**

- **Kotagiri P***, Mescia F*, Hanson AL*, Turner L*, Bergamaschi L*, Peñalver A, Richoz N, Moore SD, Ortmann BM, Dunmore BJ, Morgan MD, Tuong ZK; Cambridge Institute of Therapeutic Immunology and Infectious Disease-National Institute of Health Research (CITIID-NIHR) COVID BioResource Collaboration, Göttgens B, Toshner M, Hess C, Maxwell PH, Clatworthy MR, Nathan JA, Bradley JR, Lyons PA, Burrows N, Smith KGC. The impact of hypoxia on B cells in COVID-19. **EBioMedicine. 2022** Mar;77:103878. **(Appendix A)**

Other publications

- Rae W, Sowerby J, Verhoeven D, **Kotagiri P**…. Kuijpers T, Smith K, "Immunodeficiency, autoimmunity, and increased risk of B cell malignancy in humans with TRAF3 mutations", (**in submission**)

- Mulcahy V*, Liaskou E*, Martin J, **Kotagiri P**...Hirschfield G, Mells G, "Transcriptional profiling shows stronger regulation of peripheral immune cells in well-controlled primary biliary cholangitis", (**in submission**)

- Kessler N, Viehmann S, Krollmann C, Mai C, Kirschner K, Luksch H, **Kotagiri P**… Heeringa P, Kurts C, Garbi N, "Monocyte-derived macrophages aggravate autoimmune vasculitis via cGAS/STING/IFN-mediated nucleic acid sensing", (**in submission**)

Contributors to Collaborative Data

Chapter 3	
Generation of Flow data	Laura Bergamaschi Lorinder Turner
Compilation of Clinical data	Federica Mescia
Generation of Bulk RNAseq libraries	Laura Bergamaschi Prasanti Kotagiri
Generation of Neutralisation data	Pehuén Pereyra Gerber
Chapter 4	
Generation of BCR libraries	Prasanti Kotagiri
Generation of Neutralisation data	Ravi Gupta lab

TABLE OF CONTENTS

LIST OF FIGS

LIST OF TABLES

ABBREVIATIONS

6-HB	Six-helix bundle
AID	Activation-induced cytidine deaminase
AIM	Activation induced marker
AP-1	Activator protein 1
BCR	B cell receptor
BH	Benjamini-hochberg
BLNK	B cell linker
Btk	Bruton tyrosine kinase
CD	Crohn's Disease
CDR	Complementarity-determining regions
CH	Central helix
CR	Connecting region
CRP	C-reactive protein
cNHEJ	Classical nonhomologous end joining
DAG	Diacylglycerol
EBI2	Epstein-Barr virus-induced molecule 2
ECMO	Extracorporeal membrane oxygenation
FDR	False discovery rate
FOXO1	Forkhead Box O1
FP	Fusion Peptide
gd	Gamma Delta
GSEA	Gene set enrichment analysis
HC	Healthy controls
HCW	Healthcare workers
HMG	High-mobility-group
HR	Heptad repeat
HV	Hypervariable regions
ICAM	Intercellular Adhesion Molecule
IBD	Inflammatory Bowel Disease
IFN	Interferon

Ig	Immunoglobulin
ISG	Interferon stimulated genes
IP3	Inositol-1,4,5-triphosphate
ISGF3	IFN-stimulated gene factor 3
ITAM	Immunoreceptor tyrosine-based activation motif
ITIM	Immunoreceptor tyrosine-based inhibitory motif
LN	Lymph node
LPS	Lipopolysaccharide
MAPK	Mitogen-activated protein kinase
MAVS	Mitochondrial antiviral signalling
MDA5	Melanoma differentiation gene 5
MDS	Multi-dimensional scaling
MHC II	Major histocompatibility complex class II
MOFA	Multi-omics Factor analysis
MyD88	Myeloid differentiation primary response 88
MZ	Marginal Zone
NAAT	nucleic acid amplification test
NFAT	Nuclear factor of activated T cells
NFKB	Nuclear factor kappa B
NET	neutrophil extracellular trap
NK	Natural killer
NTD	N-terminal domain
PBMC	Peripheral blood mononuclear cells
PCA	Principal component analysis
pDC	Peripheral dendritic cell
PI3	Phosphatidalyinositol-3
PLC-γ2	Phospholipase C-γ2
PLIER	Pathway-level information extractor
RAG	Recombination Activating Gene
RBD	Receptor binding domain
RIG	retinoic acid-inducible gene
RSS	Recombination signal sequences

S1P	Sphingosine-1-phosphate
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SHM	Somatic hypermutation
SVD	Singular Value Decomposition
Syk	Spleen tyrosine kinase
T_{FH}	T follicular helper
TI	Thymus-independent
TLR	Toll-like receptors
TMPRSS2	Transmembrane protease serine 2
TOM	Topological overlap matrix measure
Tregs	T regulatory cells
UC	Ulcerative Colitis
UH	Upstream helix
UMI	Unique Molecular Identifiers
UNG	uracil-DNA glycosylase
WGCNA	Weighted gene co-expression network analysis

1. Introduction

1.1 Foreword

The main aims of this thesis are threefold:

Aim One: Understand the immune transcriptome in SARS-CoV-2
Aim Two: Understand the BCR repertoire in SARS-CoV-2
Aim Three: Understand the BCR repertoire in Crohn's disease

I have organised the material such that each results chapter contains its own introduction to the relevant concepts, studies, and literature that it is concerned with. In contrast, this introductory material is of a more general nature covering key concepts and technologies that form a foundation for subsequent chapters.

In Section 1.3, to provide background to aims one and three, I provide an overview of B cells.
In Section 1.4, to provide background to aims one and two, I provide an overview of SARS-CoV-2.

1.2 Overview of the immune system

The role of the immune system is to protect the body against infectious organisms and their toxins. It can be divided into two broad categories, the innate and adaptive immune system. The innate immune system mounts a rapid, non-specific response to pathogens. The adaptative immune system mobilises antigen-specific lymphocytes and can generate immunological memory. The two major lymphocytes are T and B lymphocytes, and antigen-specificity is conferred through the presence of surface receptors. T cells contain T cell receptors (TCR) on their surface whilst B cells contains B cell receptors (BCR) on their surface.

T cells have three functional types: cytotoxic T cells, helper T cells and regulatory T cells. Cytotoxic T cells kill infected cells. Helper T cells facilitate the function of immune cells including B cells and macrophages. Regulatory T cells dampen the immune response. Immunoglobulins are proteins that are produced by B cells.

When immunoglobulins are membrane bound, they are known as BCR (Fig 1.1). When antigen successfully binds to the BCR, the cell is activated and undergoes clonal expansion and eventual differentiation into effectors cells which secrete antibodies with a specificity identical to the surface receptors. This is known as clonal section theory, coined by F. Macfarlane Burnet[1]. Terminally differentiated effector B cells are plasmablasts and plasma cells[2].

1.3 B cells

1.3.1 B cell receptor

The human BCR repertoire contains over 10^{13} sequences. This is achieved through DNA recombination[2–4]. An immunoglobulin is comprised of two heavy and light chains, giving it a 'Y' shape. The two heavy chains are linked by a disulfide bond at the stem and each to light chain at the ends. The immunoglobulin is divided into two Fab segments and a single Fc segment. The variable regions (V regions), present in Fab, binds to antigen. The strength of the interaction between a single antigen binding site and its antigen is called affinity[5]. The heavy chain variable region is encoded by V, D and J genes off Chromosome 14, whilst the light chain is only encoded by a V and J gene. There are two classes of light chains- lambda encoded off Chromosome 22 and kappa encoded off Chromosome 2[6,7]. The Fc region of the antibody is the constant region (C region), it interacts with effector molecules and cells. The constant region of the heavy chain determines the class of the antibody and falls under the broad grouping of immunoglobulin M (IgM), immunoglobulin D (IgD), immunoglobulin G (IgG), immunoglobulin A (IgA), and immunoglobulin E (IgE)[8]. When membrane bound, the carboxy terminus is hydrophobic, when secreted it is hydrophilic[2].

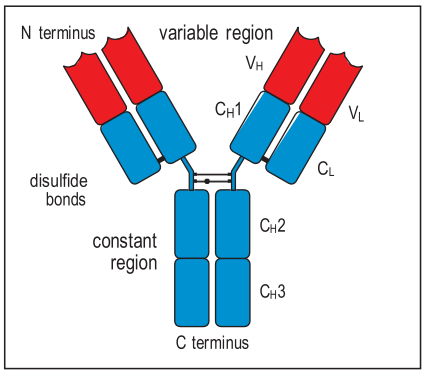


Fig 1.1 B cell receptor structure. Heavy and light chains depicted. The variable region is coloured in red and the constant region is coloured in blue[2].

The heavy and light chains are constructed from two beta pleated sheets which fold over, forming the immunoglobulin fold. Both the heavy and light chain variable regions contain three hypervariable regions (HV), with the heavy chain HV3 the most diverse and thought to dictate antigen binding properties[9–11]. The HV regions are flanked by 4 framework regions (Fig 1.2). The framework regions are heavily conserved between different antibodies. The framework regions form the beta pleated sheets whilst the HV regions spill out and take on loop structures, allowing the diversity to be localised to a region, forming a single hypervariable site. These sites are termed complementarity-determining regions (CDR), CDR1, CR2 and CR3 representing the three HV region of the heavy and light chains[2].

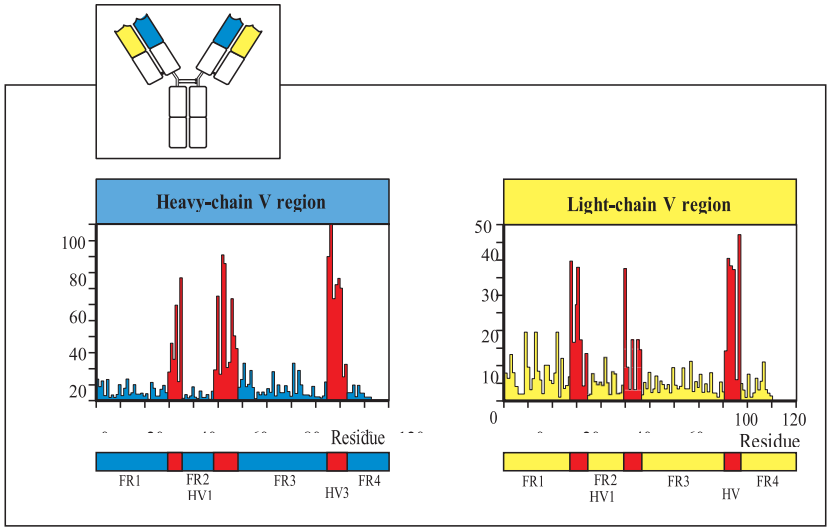


Fig 1.2 Variability in Heavy and light framework regions. The level of variability is presented on the y axis[2]

The combination of the CDRs of the heavy and light chain determines antigen binding and reflects the sequence and three-dimensional structure of the BCR[5]. Antibodies bind antigen whose surfaces are complementary with non-covalent bonds keeping antigen and antibody together[12]. Antibodies recognise only a small portion of the antigen; this is termed the antigenic determinant or epitope. Any chemical structure can be recognised by an antibody, but the usual antigens are proteins, glycoproteins, and polysaccharides. Conformational/discontinuous epitopes are where the segments of the epitope are discontinuous in the amino acid sequence but are brought together by the three-dimensional structure. Continuous/linear epitopes refers to the epitope representing a continuous single segment.

The binding between the BCR and antigen is a reversible non-covalent bond. The forces involved include[13]:

- Electrostatic forces, such as the bond between positively charged NH3 and negatively charged C02
- Hydrogen bonds- the sharing of hydrogen between two electronegative atoms
- Van der Waals forces
- Hydrophobic forces
- Cation-pi interaction- this is the bond formed between a cation and an electron cloud.

1.3.1.1 V, D and J rearrangement

V(D)J recombination allows for diversity of the BCR and T cell receptor [14]. This large combination along with insertion and deletions of nucleotides at junctional joining creates diversity. During somatic recombination of the heavy chain, the D – J regions are first re-arranged and then joined with the V region to form the primary transcript[15]. RNA splicing then occurs, removing introns separating the leader region and the constant region from the VDJ sequence. The leader peptide directs the protein to the endoplasmic reticulum (Fig 1.3).

The heavy chain encoded on chromosome 14, has 38-46 variable genes, 23 diversity genes and 6 joining genes. The kappa light chain encoded on chromosome 2 has 34-48 variable genes and 5 joining genes whilst the lambda light chain encoded on chromosome 22 has 29-33 variable genes and 4-5 joining genes[16]. V genes are commonly grouped into 7 big families based on sharing a minimum of 80% DNA sequence identity[2].

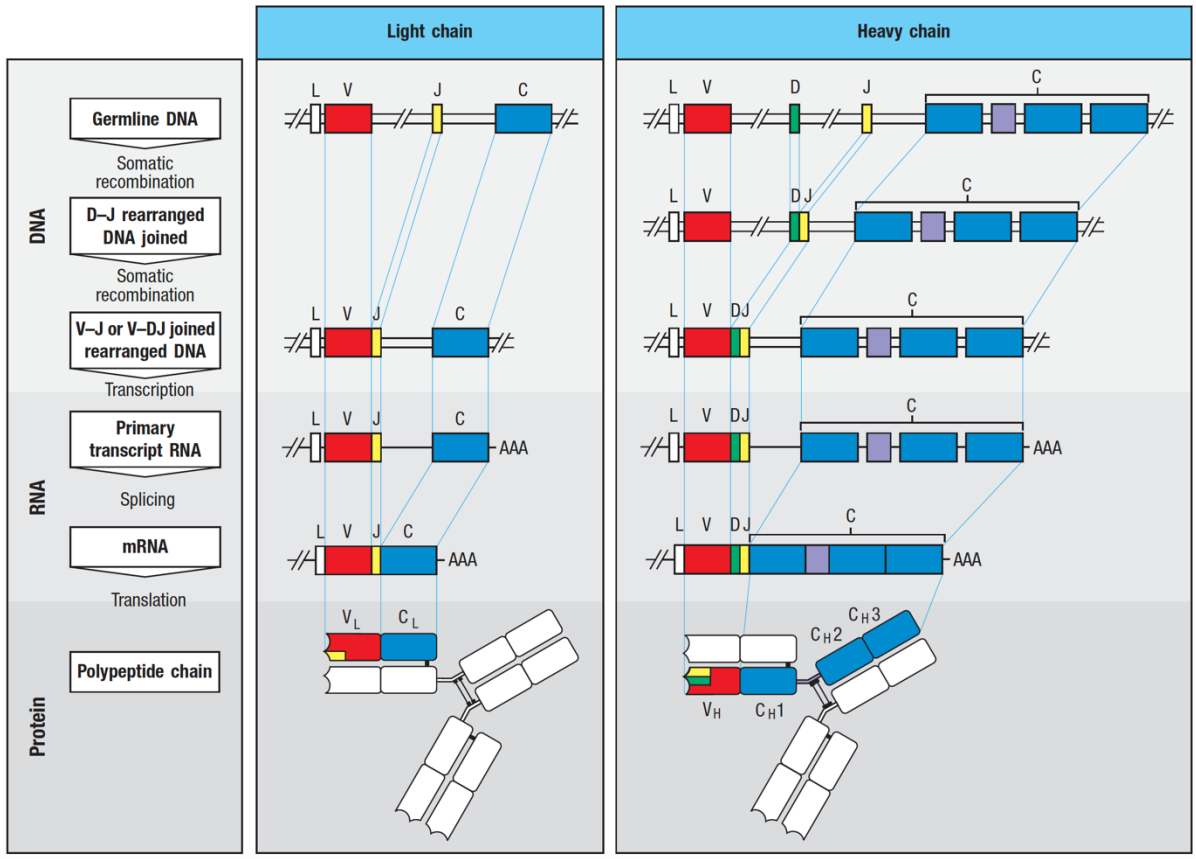


Fig 1.3 V(D)J recombination of heavy and light chains[2]

DNA rearrangement is guided by recombination signal sequences (RSS)[3]. This process needs to be finely tuned to prevent events such as V genes combining with each other. These noncoding DNA sequences are located at points of recombination and are comprised of a heptamer-spacer-nonamer[15]. The heptamer 5'CACAGTG3' is contiguous with the coding sequence, followed by the spacer, which is non-conserved and can be 12 or 23 base pairs long. This is subsequently followed by a conserved block of nine nucleotides knows as a nonamer 5'ACAAAAACC3'[16]. A gene segment flanked with an RSS with a 12 base pair spacer can only be joined to a fellow gene flanked by a 23 base pair spacer RSS. This rule prevents

the mis-binding of VDJ genes. (Fig 1.4). CDR1 and CDR2 are encoded by the V segment. CDR3 is encoded by the combination of the V, D and J segments[3].

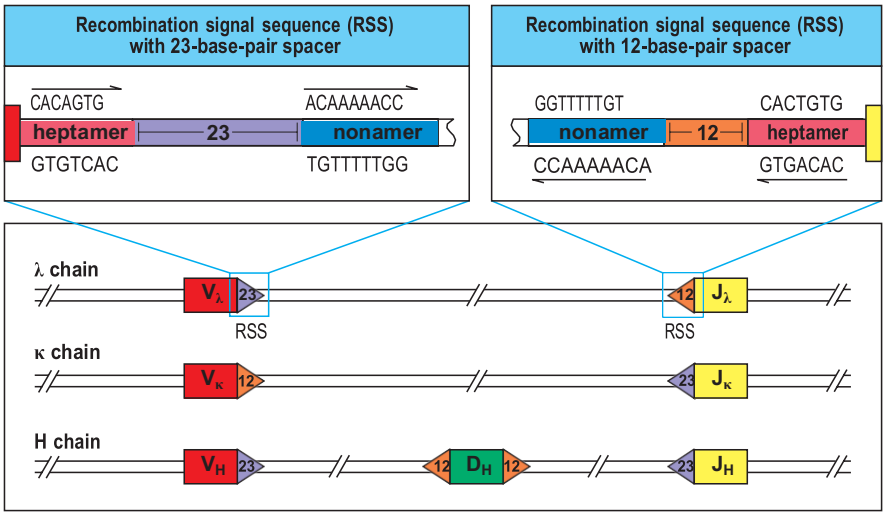


Fig 1.4 Recombination signal sequence overview for heavy and light chains [2].

V(D)J recombinase, a complex of enzymes carries out this recombination. In lymphocytes, Recombination Activating Gene (RAG) 1 and 2 are components of the recombinase complex that initiate recombination[17]. Double strand break repair proteins are also members of the recombinase complex. Their role is to imprecisely rejoin the ends of DNA post double break repair[18]. The imprecise joining results in junctional diversity[19].

RAG1 and 2 work together with high-mobility-group (HMG) proteins, they bind to the RSS and make a precise single-stranded cut to the DNA backbone (Fig 1.5) [17]. Cleavage occurs between the RSS and coding segment. The free 3'-OH group at the end of the cut strand immediately forms a phosphodiester bond with the opposite strand causing a double break and forming a hairpin coding end and blunt signal end. Post cleavage, the four DNA ends form a complex with the RAG proteins. The hairpin end is subsequently opened up by the DNA-PK:Artemis complex at random sites. This results in short single stranded extensions generating palindromic P-nucleotides. The cut end is then modified by terminal deoxynucleotidyl transferase which randomly delete and insert nucleotides. DNA ligase IV and XRCC4, components of the classical nonhomologous end joining(cNHEJ) repair pathway, ligate the two ends[3].

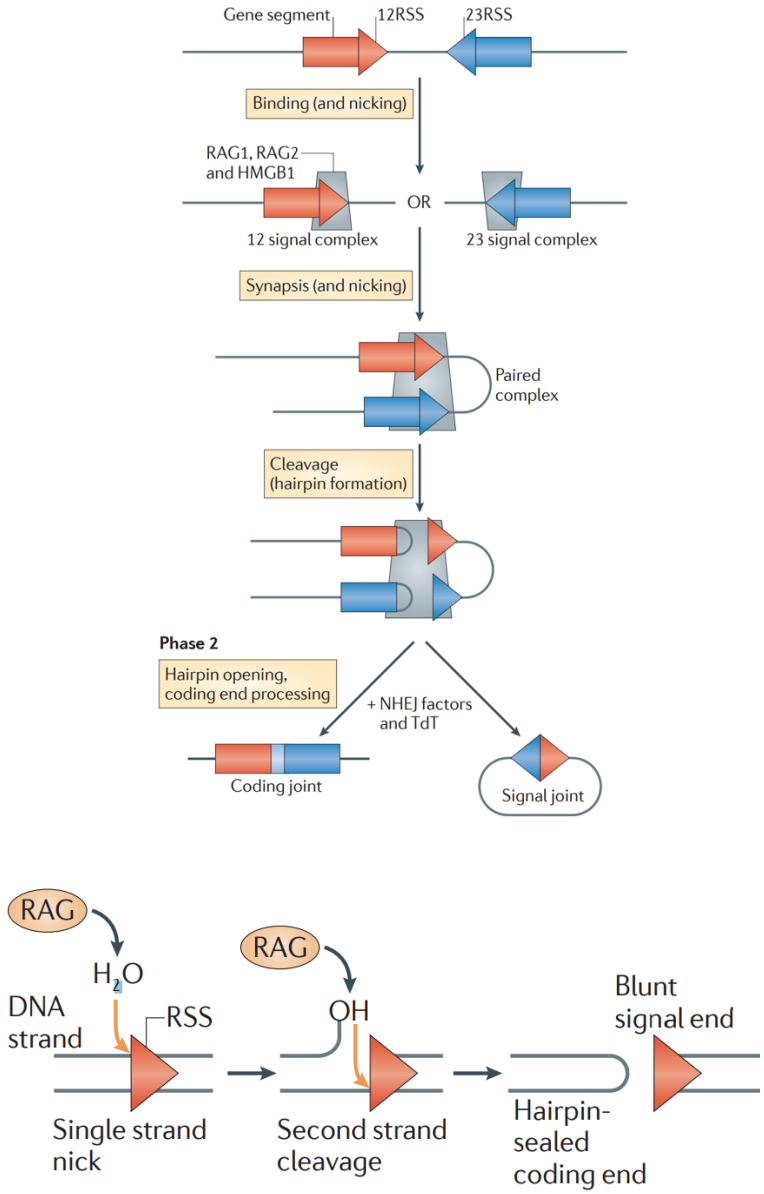


Fig 1.5 Recombination signal sequence illustrating hairpin formation[3]

1.3.1.2 Junctional diversity

CDR3 is formed by the junction of the V, D and J genes in the heavy chain and by the junction of the V and J genes of the light chain. The imprecise joining and addition and insertion of nucleotides adds to increased diversity[16]. After hairpin formation, Artemis catalyses a single-stranded break at a random point[20]. When this point is at a different position to the initial break, a single-stranded tail results consisting of a palindromic sequence. It is palindromic as it consists of nucleotides of the template sequence plus nucleotides from the complementary strand. Non-template encoded nucleotides, "N-

nucleotides" are added by enzyme TdT[21,22]. After the addition of 20 nucleotides, complementary basepairs from the single strands overlap. Repair enzyme remove non-complementary base pairs and further synthesize complementary basepairs for the unfilled gaps. From the addition of nucleotides, frameshift can occur leading to non-productive rearrangements (Fig 1.6).

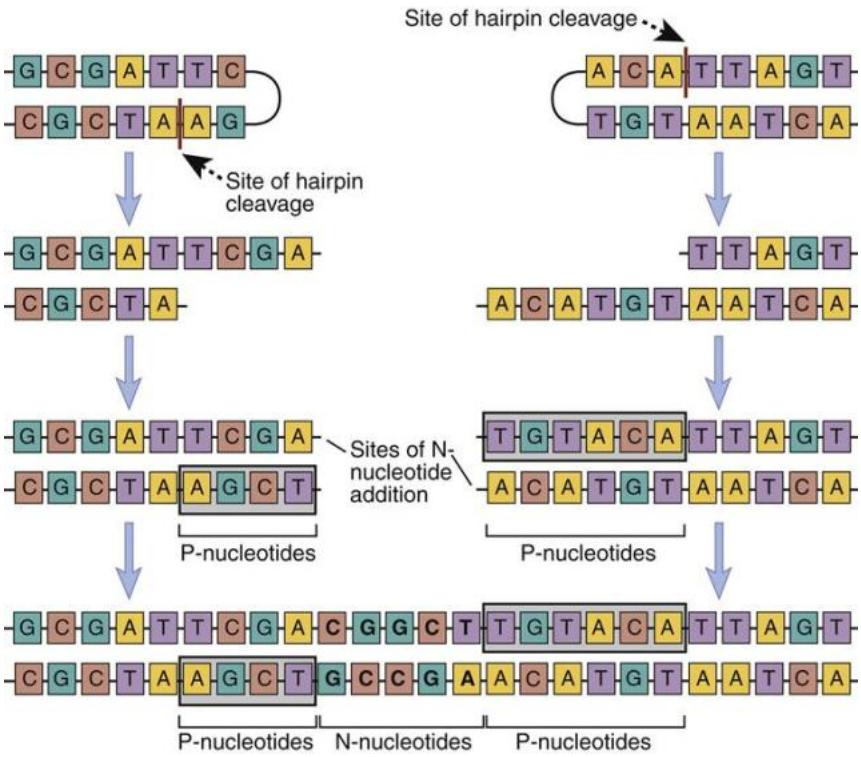


Fig 1.6 Recombination signal sequence illustrating P and N nucleotides [18].

Diversity occurs from the following:

- Combinatorial diversity:
    - Combination V/D/J genes
    - Combination of heavy chain with light chain
- Junctional diversity:
    - Imprecise joining
- Somatic hypermutation

1.3.1.3 Constant Region

All 5 classes of immunoglobulins can be membrane bound or secreted. IgG has four further subclasses, IgG1, 2, 3 and 4, named in order of serum abundance and IgA has two further subclasses IgA1 and 2. IgM and IgE have an extra C domain[2].

Post activation, IgM is the first immunoglobulin produced. IgM and IgD are both transcribed along the primary transcript[23,24]. Cleavage and polyadenylation at pA1 or pA2 lead to expression of IgM and IgD respectively. Similarly, transmembrane, and secreted forms of immunoglobulins undergo alternative RNA processing of the same heavy chain sequence[25]. IgA and IgM when secreted undergo polymerization which increases the avidity. IgA can form a dimer and IgM a pentamer[2]. Monomeric IgA1 is prominent in the serum whilst dimeric IgA2 is prominent in the gut[26]. Antibodies play three key roles, neutralization, opsonization and complement activation.

The Fc portion of the constant region has three main effector functions: Fc receptor binding, complement activation and secretion[27]. The Fc region of IgG1, 2 and 3 and IgM bind with C1q to activate the classical complement pathway[28,29]. Lastly, the Fc portion can bind to receptors enabling transport through cells such as in mucosal cells or across the placenta[30].

1.3.2 Development of B cells

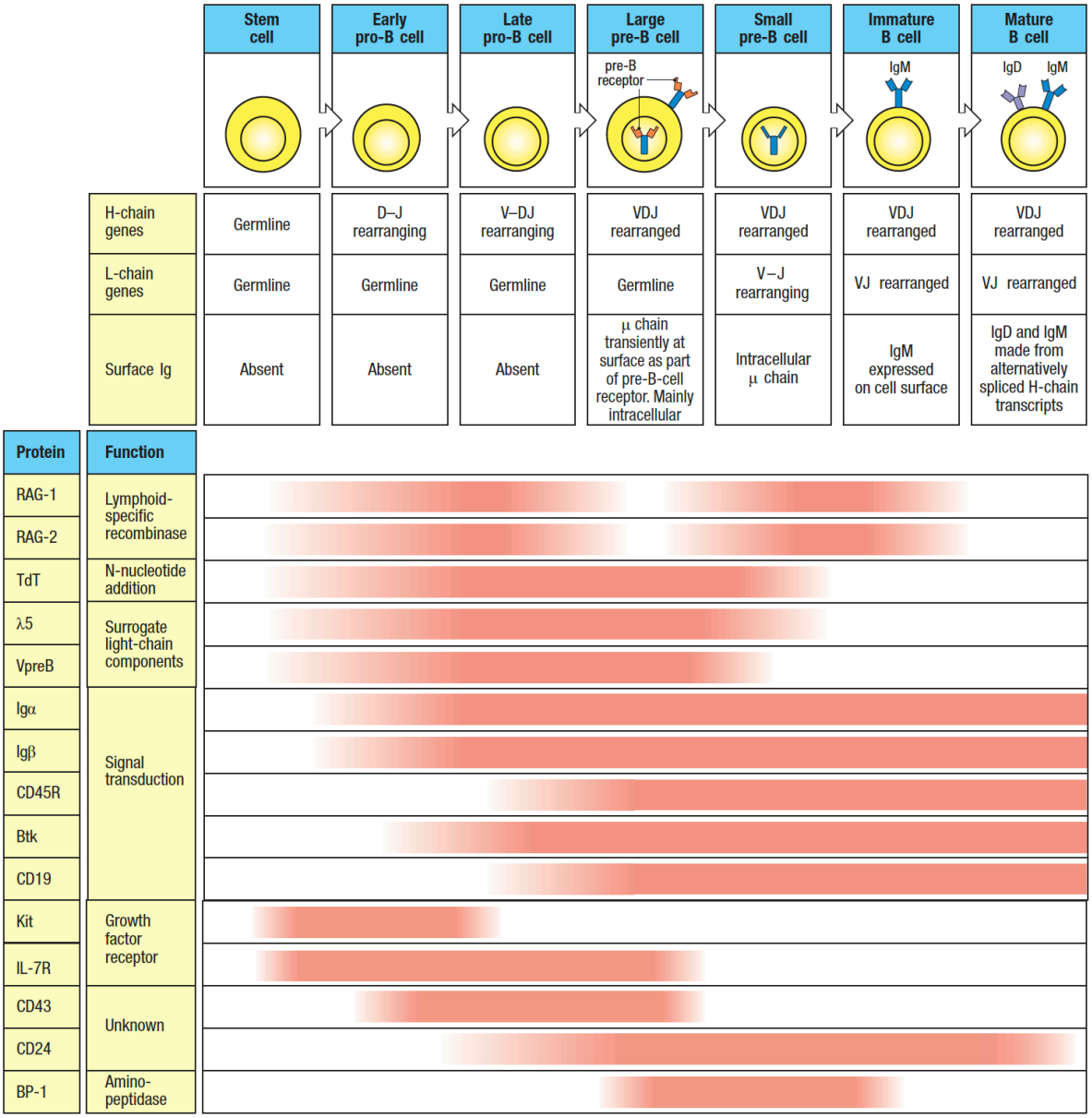


Fig 1.7 Stages of B cell development [2]

B and T lymphocytes are produced in the bone marrow[31]. B cells complete their maturation in the bone marrow whilst T cells migrate to the thymus to do so. New B cells are continually produced whilst T cell numbers are maintained by mature T cells in the periphery as the thymus atrophies with age. B cells are derived from common lymphoid progenitor cells. The stages of B cell development are early pro-B cell, late pro-B cell, large pre-B cell, small pre-B cell, immature B cell, and mature B cell (Fig 1.7)[2].

Heavy chain D-J rearrangement occurs during the early pro-B cell stage and occurs in both alleles[14,32]. As most D gene segments can be translated in all three reading frames and not generate a stop codon, this joining is mostly successful, and no special mechanism is in place to check. The V-DJ rearrangement occurs at the late pro-B cell stage and occurs initially in one chromosome. If successful, a μ heavy chain is formed, and the cell progresses to a pre-B cell. If this re-arrangement fails, rearrangement of the other chromosome occurs. Overall, there is a 55% chance of progressing to a pre-B cell. To test whether a functional μ chain has been formed, a temporary "surrogate" light chain is formed and assembles only with a successful μ chain[33]. The surrogate chain is encoded by lambda 5 and VpreB genes. Lambda 5 protein is a surrogate for the constant region of the light chain whilst protein VpreB is the variable gene surrogate[34–37]. Adjacent amino-terminal tails of VpreB and lambda 5 bind with each other from neighbouring BCRs leading to phosphorylation of Igβ and Igα (discussed further below)[38,39].

Heavy chain locus rearrangement halts by the pre-B-cell stage through the reduction in levels of RAG-1 and 2[40]. Pre-B cells are sensitive to IL-7 and in this environment undergo proliferation, expanding the population 30-60 fold[41]. Light chain rearrangement begins with re-expression of RAG-1 and 2. Thus a given heavy chain may have multiple different light chain pairs. Light chain re-arrangement also exhibits allelic exclusion[42]. If the initial rearrangement is non-functional, additional re-arrangement occurs on the same allele before re-arrangement occurs on the other allele. Isotypic exclusion also occurs in light chains where per cell either a κ or λ chain is expressed but not both[43]. The κ light chain in humans is the first to undergo rearrangement with initiation in the λ light chain 5 times less likely[44].

The now paired immunoglobulin chain- IgM is expressed on the cell surface and the cell becomes an immature B cell. B cells are tested for self-reactivity within the bone-marrow[45]. If there is no reaction, they will migrate to the periphery becoming mature B cells. If self-reactive, they may undergo clonal deletion, receptor editing, anergy or immunological ignorance[45,46]. RAG1-2 are still expressed in the immature B cell and if the B cell is self-reactive, further rearrangement of the light chain can occur[46]. Where the B cell remains self-reactive, clonal deletion occurs. When a B cell is self-reactive but only has weak cross-linking, anergy occurs. Anergic B cells are unreactive even in the presence of antigen and T cell help. Peripheral tolerance is a further checkpoint in place- where when self-antigen is encountered deletion, anergy or survival can occur[47]. The final maturation of B cells occurs in the spleen. Immature B cells express high levels of surface IgM and low levels of surface IgD whilst the converse is true for the mature naïve B cells[48].

The bone-marrow produces 5-10% of the total B lymphocyte population on a daily level. However, the size of the pool remains constant because of death of the immature B cells which have very short half-lives. The secondary lymphoid follicle is essential for immature B cell survival with the abundant production of BAFF by follicular dendritic cells promoting survival[49]. Lack of access to the follicle results in death after 2-3 days[50].

On entry to the spleen, an immature B cell transitions to a T1 and then T2 B cell[51]. A T2 B cell is defined by the presence of co-receptor CD21[52]. Weak activation of the BCR and BAFF-R simulation promotes B cell maturation. T2 B cells differentiate into either follicular (B-2) or marginal B cells, which are a much smaller population. Marginal B cells reside at the junction of the red and white pulp in the spleen and are poised to respond to pathogens present in the blood, thus acting as an early defence[53].

1.3.3 B cell receptor activation

Knowledge of successful antigen binding to receptor needs to be transduced into the cell. In the B cell, this is achieved through invariant protein chains, Igα and Igβ [54,55]. These proteins are single chained and consist of three portions, an extracellular immunoglobulin like

domain, a transmembrane domain and a cytoplasmic tail containing immunoreceptor tyrosine-based activation motifs (ITAMs). Igα and Igβ form a dimer and associate with the BCR. When the BCR is activated, tyrosine residues on the ITAM portion of Igα and Igβ are phosphorylated by Src-family kinases Lyn and spleen tyrosine kinase (Syk). This leads to recruitment of kinases including Syk, Bruton tyrosine kinase (Btk), and Lyn and proteins including Vav, Grb2 and B-cell linker (BLNK). The recruited kinases amplify the signal of activation. Src-family kinase Lyn also phosphorylates tyrosine residues on CD19 thus decreasing the activation threshold. All this combined leads to activation of 3 pathways including phosphatidalyinositol-3-kinase (PI3-Kinase), Btk and phospholipase C-γ2 (PLC-γ2). Syk and Lyn are recruited to the phosphorylated ITAM of Igα and Igβ whilst BLNK binds to non-ITAM portion of Igα via Src homology 2 (SH2) domain. BLNK is phosphorylated by Syk thus acting as a scaffold protein. PLC-γ2 now in close contact is phosphorylated by both Btk and Syk producing diacylglycerol (DAG) and inositol-1,4,5-triphosphate (IP3). DAG activates protein kinase C. IP3 generation leads to calcium influx from the endoplasmic reticulum and the extracellular compartment resulting in activation of NF-Kb, Jun and nuclear factor of activated T cells (NFAT). PI3K pathway is also activated. PI3K catalytic subunit p110 is held in check at rest by PI3K subunit p85. On BCR activation, p85 is recruited elsewhere including binding to CD19. PI3K facilitates continued BCR activation by recruiting more kinases to the site. The mitogen-activated protein kinase (MAPK) pathway is important in cell survival and proliferation and is activated post BCR receptor ligation [56].

For effective B cell activation, an additional activation signal is required beyond antigen. This is achieved with CD4 T cells. CD4 CD40 ligand binds with CD40 receptor on B cells. CD40 is a TNF receptor superfamily member. Activation leads to the recruitment of adaptor proteins termed TRAFs. These lead to the activation of the non-canonical nuclear factor kappa B (NFKB)-pathway. Receptors on B cells containing Immunoreceptor tyrosine-based inhibitory motif (ITIM) can inhibit activation such as FcγRIIB by inhibiting the functions of PI 3-kinase as discussed below.

1.3.4 Fcγ receptors

Fc receptors bind to the Fc portion of antibodies. They have two main purposes when not inhibitory. One is to remove the antibody/antigen complex via phagocytosis. This is led by macrophages, dendritic cells and neutrophils. The other is the release of cytokines and stored cellular components as seen with natural killer (NK) cells, eosinophils, basophils and mast cells[57].

Type 1 Fcγ receptors can be activating or inhibitory. Activating receptors include FcγRI, FcγRIIa, FcγRIIc, and FcγRIIIa. FcγRIIb is the sole inhibitory receptor. When activating FcγR are activated, this results in phosphorylation of their ITAM domains and activation of cytoplasmic kinases such as Src and Syk family[58]. When inhibitory receptors are activated, their ITIM domains undergo phosphorylation and SHIP phosphatases are recruited which inhibit the activation of Src kinases and Phospholipase C. Type II Fcγ receptors include DC-SIGN and CD23 and these interact with the CH2-CH3 interface[59].

Receptor	Constitutive Expression	Induced Expression
FcγRI	Monocytes	Neutrophils, eosinophils, Dendritic cells
FcγRIIa	Neutrophils, Monocytes, Eosinophils, Macrophages, Dendritic cells, Platelets	
FcγRIIb	Neutrophils, Monocytes, Eosinophils, Macrophages, Dendritic cells, B cells, Plasma cells	
FcγRIIc	NK cells	
FcγRIIIa	Macrophages, NK cells, Monocytes	Dendritic cells
FcγRIIIb	Neutrophils	Eosinophils
DC-SIGN	Dendritic cells, macrophages	
CD23	B cells	Neutrophils, Monocytes, Eosinophils, Macrophages, T cells

Table 1.1 Summary of Fcγ receptors

Most leucocytes co-express activating and inhibitory type I FcγRs and thus the outcome is determined by the dominant signal. IgG1 and IgG3 unlike IgG2 and IgG4 have high affinity for type 1 FcγRs[58].

FcRn modulates IgG half-life via endosomal recycling [60]. A large proportion of IgE is bound to mast cells via FcεR1[61]. When cross-linking occurs, degranulation results. IgA binds to the polymeric receptor of epithelial cells enabling translocation to the lumen[62].

1.3.5 Thymus dependent antigens

Naïve B cells are activated after binding of antigen to BCR. Accessory signalling is required which can be provided by T cells or the pathogen itself. Both pathways lead to activation of PI 3-kinases and activation of the Activator protein 1 (AP-1) pathway leading to cell proliferation and differentiation.

Protein antigens require T-cell help to induce an antibody response. T cell help is provided in the form of T follicular helper (T_{FH}) cells. T_{FH} cells are derived from naïve CD4+ T cells, reside in lymph nodes and express the chemokine receptor CXCR5[63,64]. IL-6 and ICOSL facilitate differentiation to T_{FH} whilst IL-2 is a potent inhibitor via its downstream induction of Blimp-1 and STAT5[65].

The BCR initiates a signalling cascade post activation by antigen. Protein antigen binds to a receptive BCR and is subsequently internalised. The protein is degraded, and peptide components are displayed on the MHCII complex. B cells act as antigen presenting cells by upregulating MHCII and co-stimulatory molecules such as CD80 and CD86 further stimulating T_{FH} differentiation. The peptide:MHC II complex is recognised by already primed (by the same antigen) T_{FH}. The T_{FH} cells are activated by the same antigen but likely different epitope in a process known as “linked recognition”[66].

For linked recognition to occur, shared antigen specific B and T cells need to meet, this occurs at the T-B junction[67]. When no antigen is encountered, B cells exit the lymphoid tissue after sensing sphingosine-1-phosphate (S1P). Upon encounter and binding of protein antigen, B cells upregulate CD69 which promotes retention in the lymph node. B cells then further upregulate Epstein-Barr virus-induced molecule 2 (EBI2), CXCR5 and CCR7 and downregulate SP1R1 which helps with the migration to the T-B border(Gatto et al., 2009; Pereira et al., 2009). Post priming by dendritic cells, CD4 T cells upregulate BCL-6 which leads to increased expression of CXCR5 and repression of CCR7 which enables migration to

the T-B border[70]. B cells express ICOSL which binds to ICOS on T_{FH} completing its differentiation[71]. T_{FH} cells subsequently increase their expression of BCL-6 and internalise SAP which facilitates sustained contact with B cells[72].

T_{FH} cell ligand, CD40L binds with CD40 on B cells activating the non-canonical NFKB pathway and release of anti-apoptotic molecules BCL-2[73]. Activated B cells can proliferate and either differentiate along the extrafollicular or follicular routes[74]. Activated B cells may migrate to the outer follicle and form the "primary focus" where they undergo differentiation and proliferation with some forming plasmablasts. These extra-follicular B cells give rise to short lived plasma cells which can undergo class-switching and carry fewer somatic hypermutations[74]. Other activated B cells along with activated germinal centre cells migrate to a primary lymphoid follicle forming a germinal centre. $CXCR5^{+}BCL\text{-}6^{+}PD\text{-}1^{+}T_{FH}$ produce IL-21, IL-4, CD40L, and CXCL13 which support germinal centre B cells. IL-21 and CD40L are required for B cell proliferation both in the germinal centre and extrafollicularly with IL-21 activating STAT3 resulting in proliferation and differentiation into plasma and memory B cells. The germinal centre ultimately produces high affinity long lived plasma cells. Germinal centre B cells express transcription factors BCL-6 and G-protein-coupled receptor S1P2.

BCL-6 has an important role in germinal centre formation, performing the following four roles[49,75]:

- Silencing the anti-apoptotic molecule BCL-2 thus promoting a pro-apoptotic state
- Reducing the expression p53 and ATR which increases the tolerance to DNA damage which occurs secondary to rapid proliferation.
- Represses Blimp-1 thus preventing exit from the germinal centre and differentiation into plasma cells.
- Downregulating mediators of BCR and CD40 signalling

1.3.5 Germinal centre

The germinal centre is a histological structure formed in the setting of a T dependent response. It is a site where affinity maturation and class-switching occur with the eventual

formation of memory B cells and long-lived plasma cells. Germinal centres can persist for months [75].

1.3.5.1 Affinity maturation

In the germinal centre, V region somatic hypermutation occurs (Fig 1.8)[49,76]. This is where random mutations at a rate of one base pair change per 10^3 base pairs occurs per cell division with the aim of improving antibody affinity. Most mutations are ill fated resulting in apoptosis either from the inability to make a functional BCR or the inability to compete for antigen with sibling B cells leading to negative selection. During every B cell division, there is a 50% chance that an amino acid will be altered. Positive selection results in replacement mutations in the CDR regions as this determines antigen affinity whilst silent mutations resulting in no amino acid change are scattered throughout the V region.

The germinal centre is split into a dark and light zone[77]. Dark zone B cells are $CXCR4^{hi}CD83^{low}CD86^{low}$ cells whilst light zone B cells are $CXCR4^{low}CD83^{hi}CD86^{h}$i cells[78,79]. In the dark zone, B cells proliferate. In the light zone, B cells exit the cell cycle, and their affinity is tested by follicular dendritic cells. B cells compete for T_{FH} help where B cells with higher affinity for antigen can internalise greater amounts of antigen presented by follicular dendritic cells and in turn present it to T_{FH} cells. Germinal centre B cells that do not receive help die[80,81]. In addition, germinal centre B cells express low levels of VLA4. VLA4 binds with VCAM1 on follicular dendritic cells which stabilises the interaction. Thus, low levels of expression require higher affinity for downstream signalling to occur. Expression of VCAM-1 on follicular dendritic cells occurs via activation of the NF-KB signalling pathway. Follicular dendritic cells have complement receptors CD21/CD35 which are critical in the trapping of CD3-tagged antigen and immune complexes whilst FcRIIB is dispensable[82].

Furthermore, a process of “antigen-masking” occurs where antibodies secreted by plasma cells generated early in the immune response bind to antigen of follicular dendritic cells and thus, only germinal centre B cells with high affinity BCRs are able to out compete for antigen acquisition, further enabling the selection of high-affinity B cells[83]. Masking of epitopes also encourages the generation of new clones that bind other epitopes. Integrins

also play an important role in cell-cell interaction. With Lymphocyte function-associated antigen 1 engaging Intercellular Adhesion Molecule (ICAM) 1 and 2 on B cells.

In the light zone, T_{FH} express CD40L and bind with high affinity B cells[84]. CD40 deficiency results in hyper IgM syndrome in humans with failure to form germinal centres[85,86]. In the germinal centre, B cell CD40 signals are transduced through NFKB whilst BCR signals are transduced through Forkhead Box O1 (FOXO1). When combined these activate c-Myc which promotes cell survival and cell cycle re-entry[87]. This is dissimilar to naïve B cells where these pathways are not siloed and Myc transmission can be induced by either BCR signalling or CD40. Inhibition of MYC leads to loss of established germinal centres.

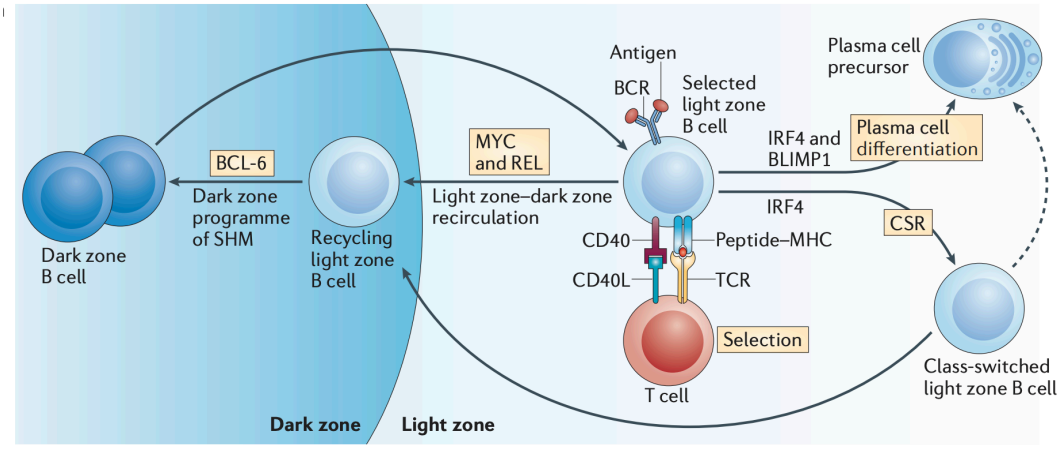


Fig 1.8 Germinal centre illustrating dark and light zones and affinity maturation[88].

CD40 stimulation of germinal centre B cells result in upregulation of IRF4 via NFKB pathway activation[89]. This leads to repression of BCL-6 transcription and terminates dark zone B cell transcriptional profiles. Germinal centre B cells have altered morphology compared with naïve B cells with a dendritic appearance potentially increasing ability to interact with antigen and have a stronger "tugging" ability when bound to antigen[90]. Germinal centre B cells may differentiate into long lived memory B cells which reside in the outer zones of secondary follicles or migrate to the marginal zones or differentiate into long-lived plasma cells that migrate to the bone marrow or gut. Long lived memory B cells reside in secondary lymphoid organs where they can be exposed to antigen on re-infection[91].

Long-lived plasma cells in the bone marrow secrete large amounts of antibody in the absence of antigen. It is important to note that although the primary aim of the germinal centre is to increase antibody affinity, it is also to increase diversity. This increases the breadth of the antibody response thus also allowing an effective response when mutations occur to the antigen. This breadth is in the B cells that differentiate into memory B cells, which are often low affinity and broadly reactive[91,92]. B memory cells are generated early in the germinal centre response contributing to diversity. In contrast, long lived plasma cells are less diverse and of a higher affinity. Germinal centre B cells favoured to differentiate into long lived plasma cells are BCL-6^{low}CD69^{hi}[93]. In mouse models, it has been shown that IgM+ memory B cells re-entered the germinal centre whilst IgG+ memory B cells were more likely to differentiate into plasma cells[94,95]. Germinal centre B cells reduce BCL-6, increase IRF4 and Blimp1 in order to differentiate into plasma cells[96]. Memory B cell generation requires BACH2 and CCR6 and occurs in the setting of "moderate" help from T_{FH} cells (thus not as high affinity). Upon exiting the germinal centre, plasmablasts continue to proliferate and eventually stop as they reach the medulla of the lymph node or the splenic red pulp.

1.3.5.2 Activation-induced cytidine deaminase

Activation-induced cytidine deaminase (AID) plays a key role in somatic hypermutation (SHM) and class switching[97,98]. AID deaminates cytidine, converting it to uridine. When this occurs in the V-region, SHM occurs and when this occurs in the switch regions, class switching occurs. Given AID can only work on single DNA strands, it acts during transcription when the DNA helix is unwound. When uridine is present in DNA it triggers mismatch or base excision repair resulting in somatic hypermutation[99].

In mismatch repair, MSH2 and 6 detect the error and recruit nucleases to remove the basepair and neighbouring nucleotides. This is subsequently repaired by error prone DNA polymerase[100]. In base-excision repair, uracil-DNA glycosylase (UNG) removes the uracil from uridine leaving an "abasic" nucleotide and subsequent insertion of a random nucleotide during DNA replication into the new DNA strand opposite the abasic site[101].

For SHM to occur higher amounts of AID are required compared with class-switching and thus is largely restricted to the germinal centre[49.]

1.3.5.3 Class switching

Class-switching only occurs post stimulation by antigen. Cytokines produced by T_{FH} cells determine isotype switching. Different antibody isotypes have different effector functions, with IgG1 and 3 being effective against viruses, IgG3 against encapsulated bacteria, IgG4 and IgE against parasites and IgA1/2 against mucosal infections. IgG is the most prominent isotype in serum and IgA in the mucosa[2].

IgM antibodies are produced first post infection and are often of low affinity. The ability to form pentamers results in great avidity. IgM is prominent in the blood but less so in tissue due to its size. Most IgM antibodies are produced by marginal zone B cells. IgA is not activated by complement, acts on epithelial surfaces, and acts as a neutralising antibody. The ratio of IgA1: IgA2 is 10:1 in the blood and 2:3 in the gut. Each constant region has a promoter which is sensitive to multiple inputs including the BCR, CD40, TLRs and cytokines which determines its transcription[2].

In class switch recombination, AID targets designated “switch regions” which flank each constant gene with exception of the IgD constant region[102]. Switch regions contain WGCW (A/T-G-C-A/T) motifs which are favoured targets of AID[103]. In this process, DNA is permanently deleted. Similar to somatic hypermutation, AID deaminates cytidine, converting it to uridine stimulating UNG to remove the uracil from uridine leaving an “abasic” nucleotide. Instead of a random nucleotide replacing the abasic nucleotide as seen with SHM, apurinic/apyrimidinic endonuclease 1 (APE1), excises the abasic nucleotide, causing a single stranded nick which is then converted into double stranded breaks. This occurs at both switch regions. Given switch regions are intronic, random deletions do not result in a frameshift. Similar to VDJ recombination, DNA recombination occurs via NHEJ[104].

After RNA is transcribed, the intron RNA segment containing the switch regions are spliced out. The sequence is G-rich and forms a G-quadruplex[105]. This complex associates with the DNA sequence from which it was transcribed from as it is complementary. It also binds with AID and thus guides AID to the switch region.

1.3.6 Thymus independent antigens

Thymus-independent (TI) antigens can be grouped into TI-1 and 2 antigens. TI-1 antigens can induce activation and proliferation of B cells regardless of their antigen specificity. Such antigens include Lipopolysaccharide (LPS) and bacterial DNA and high concentration of antigen is required to induce this polyclonal response. TI-2 antigens have highly repetitive structures and can activate mature B cells, especially marginal zone B cells[106]. Activation occurs through cross-linking of BCRs and antigen specificity enhances activation. Such antigens are often from capsular polysaccharides which have repetitive structures. Co-stimulation of TLRs have a synergistic effect on B cell activation[107].

1.3.7 Extrafollicular B cell responses

The extrafollicular response can be T dependent and T independent. It is where naïve B cells are activated and generate short lived plasmablasts outside of the follicle(MacLennan et al., 2003). Antibodies are rapidly produced within 3 days whilst a germinal response takes 7 days[109].

The T dependent response, similar to the germinal centre response, relies on BCL-6 + PD-1- T_{FH} cells [110,111]. These pre-germinal centre T_{FH} cells play an important role in class-switch recombination outside the germinal centre. It's unclear what determines whether a B cell enters the extrafollicular or germinal centre pathway with identical B cells clones (given they can proliferate prior to cell-fate decisions) found downstream of both pathways. Marginal zone B cells are key cells recruited to the extrafollicular pathway[53].

Extrafollicular B cells continue to express EBI2 and increase their expression of CXCR4 which results in localisation to the bridging channels of the spleen or the medullary cords of lymph nodes[69]. Extra-follicular B cells upregulate their expression of BLIMP-1 which leads to the direct suppression of PAX5, required for germinal centre formation[112]. Extrafollicular T_{FH}

cells mediate their interaction with B cells via CD40-CD40L and ICOS-ICOSL in the setting of IL-21. These interactions lead to plasma cell differentiation and class-switching[109].

Extra-follicular plasma cells are often but not always low affinity from a lack of SHM and are commonly short lived[113]. Memory cells can also be generated from this response and are commonly IgM and lowly mutated. AID expression in extrafollicular sites has been observed.

1.3.8 Humoral Memory

On antigen re-exposure, pre-existing protective antibodies which are secreted by long lived plasma cells are the first line of defence[91]. If antibody levels are not sufficiently high, pathogen-experienced memory B cells are mobilised as a second line of defence. Memory B cells on antigen re-exposure can differentiate into plasma cells and also re-enter the germinal centre undergoing further rounds of affinity maturation. Memory B cells can be germinal centre derived and germinal centre- independent and can be switched and unswitched[75].

1.3.9 Bone marrow homing

The process of antibody secreting cells leaving the lymph nodes, entering the blood and homing to bone marrow is incompletely understood. Activation of SP1R1 facilitates antibody secreting cells to the leave secondary lymphoid organs and enter the blood. CXCL12 and its receptor CXCR4 in turn facilitate recruitment to the bone marrow[114]. In contrast, inflammatory cytokines including CXCL9, CXCL10 and CXCL11 signal via receptor CXCR3 driving homing of antibody secreting cells to areas of inflammation[115]. Long lived plasma cells in the bone marrow are $CD19^{-}CD38^{hi}CD138^{+}$[116]. Long term survival of plasma cells in the bone marrow is dependent on the expression of antiapoptotic protein Mcl-1. It is unclear if plasmablasts are just short-lived secreting antibodies during acute infection or post vaccination or also contribute to the long-lived plasma cell pool. Post infection, the majority of plasmablasts undergo apoptosis. Plasmablasts are detected in the blood during

"steady state" and are likely secondary to mucosal immune reactions. IgA also represents a large percentage of long lived plasma cells in the bone marrow at 40%[113].

1.3.10 Marginal Zone B cells

Marginal zone (MZ) B cells reside in the spleen, GALT, lymph nodes and tonsillar crypts[117]. Both the spleen and the gut associated lymphoid tissue have an anatomically defined region called the MZ[118]. In the spleen, this region is at the interface of the circulation and the white pulp[117]. This location is strategically placed to provide the first line of defence against microbial antigens through the rapid production of IgM, IgG and IgA antibodies. MZ B cells have a lower threshold for activation and class-switching and are polyreactive[119].
MZ B cells express high levels of toll-like receptors (TLR) facilitating T cell independent activation[120]. MZ B cells recirculate with $IgM^{hi}IgD^{low}CD1c^{+}CD21^{hi}CD23{-}CD27^{+}$ $CD5^{-}$ cells found in the periphery[121]. Upon antigen exposure, MZ B cells can rapidly differentiate into plasmablasts and unlike in mice, human MZ B cells are mutated[118].

MZ B cells have distinguishing features from class-switched and non-class-switched memory B cells including expression of IgD, distinct IgV gene repertoire and fewer mutations[122]. MZ B cells are pre-diversified (not secondary to antigen exposure) utilising less VH1 and more VH3. The origin of MZ B cells is not clear. They may represent memory B cells derived at an early stage of germinal centre differentiation, memory B cells generated by T-independent responses resulting in fewer somatic hypermutations or a separate B cell lineage altogether[118]. MZ B cells express high levels of Blimp1 and low levels of Pax5 and Bcl6 aiding easy differentiation into antibody secreting cells[118,119]. In patients with hyperIgM syndrome secondary to a CD40 or CD40L deficiency resulting in lack of germinal centre, somatic mutation is still apparent in IgM+IgD+CD27+ B cell subsets suggestive of a germinal centre independent reaction. MZ B cells accrue mutations with time reaching the levels of adults by age 2-4 years old. MZ population can be replenished post depletion (e.g. bone marrow grafting)[118]

1.3.11 B-1 Cells

B-1 cells arise during fetal development and are present in peritoneal and pleural cavities in mice[123]. The BCR is germline-encoded, polyreactive, restricted in V-gene usage and of class IgM and IgA. These cells are activated by both T-1 and T-2 antigens. B-1 cells are further divided into B-1a (CD5+) and B-1b cells (CD5-). B-1a cells produce natural auto-antibodies and play a role in removing apoptotic debris as well as protect against streptococcus pneumonia and influenza[124]. B-1b cells generate IgM memory B cells, important in preventing recurrence of Borrelia hermsii, S. pneumoniae, and Salmonella[125].

1.3.12 Clonal redemption

Self-reactive B cells are present in the periphery. This occurs due to failure in identification during central tolerance and can occur because of reactivity to only monovalent antigen or weak binding to self-antigen[126]. Peripheral tolerance is the second line of defence with self-reactive B cells undergoing anergy. Anergic characteristics include down-regulated surface IgM but not surface IgD and decreased response to BCR stimulation[127,128]. The constant occupancy of the BCR by self-antigen is required for maintenance of anergy[126]. However, some auto-reactive B cells do enter the germinal centre and are redeemed, or self-reactivity is enhanced in the setting of an inflammatory milieu. T cell help is only provided to cross-reactive autoreactive B cells which leads to cross-reactive autoreactive memory and plasma cells.

Polyreactive naïve antibodies allow a broader coverage against potential antigens despite the increased risk of self-reactivity. Nemazee calculated that a minimum of 1/3 of nascent B cells that are self-reactive need to undergo anergy and not be deleted/edited to achieve optimal coverage against a range of pathogens. Cross-reactive B cells (against self and antigen) may play an important role in an initial pathogen specific antibody response or where a pathogen specific-non-self-reactive response is impossible[129].

Autoantibodies with IGHV4-34*01 heavy chains bind to poly-N-acetyllactosamine carbohydrates (I/i antigen) which is present on erythrocytes and B lymphocytes and can result in cold agglutinins disease[130]. This heavy chain is present in 5% of anergic B cells and is

autoreactive regardless of light chain pairing. Auto-reactivity is dependent on FWR1 where a hydrophobic patch (AVY) is recognised by anti-idiotype antibody 9G4[126]. These BCRs are under-represented in the germinal centre and memory compartment in health whilst over-represented in SLE[131]. Defective clonal redemption appears to occur in SLE. 9G4+ IGHV4-34 B cells are mutated away from binding to self-poly-N-acetyl-lactosamine but rather react against autoantigens such as dsDNA and other self-nuclear antigens. This failure to select progeny with removed self-reactivity may be due to plentiful T_{FH} cells or an intrinsic defect in the B cells.

Whereas Reed et al., illustrate that IGHV4-34*01 IgG antibodies generated post immunisation, no longer bound to I/i antigen, whilst the unmutated germline sequence did so. Self-reactivity was thus removed via somatic hypermutation. In addition, they illustrate that somatic hypermutation not only facilitated increased affinity to antigen but also reduced self-reactivity was prioritised at the expense of antigen-specificity. The benefit of decreasing self-reactivity is a decreased occupancy by self-antigen allowing increased binding to foreign antigen [132].

Similarly, IGHV1-69 is a poly-reactive antibody. Its long hydrophobic loop in the CDR2 segment allows it to bind to many structurally unrelated antigens including the Fc domain of self-IgG and hemagglutinin stalk of Influenza A virus[133]. This antibody is mobilised on initial infection but in the memory pool, somatic hypermutation alters this idiotype preventing long term auto-immune complications[134].

1.4 SARS-CoV-2

1.4.1 Background

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causing COVID-19 was declared a pandemic by the World Health Organisation on March 11, 2020 (https://covid19.who.int/). The origin of the virus was traced back to the Huanan Seafood Wholesale Market in Wuhan city, Hubei Province, China. Key clinical features include pneumonia with fever, cough, dyspnea and bilateral lung infiltrates[135].

The causative virus, SARS-CoV-2 is a novel betacoronavirus. Phylogenetic analysis shows that SARS-CoV-2 clusters with SARS-CoV, bat and pangolin coronaviruses placing it in the subgenus Sarbecovirus [136]. Although phylogenetically related, SARS-CoV-2 has likely undergone 20 years of sequence evolution and bats are unlikely the direct progenitor. Genomic analysis reveals ~96% nucleotide sequence similarity with the *Rhinolophus affinis* bat virus[137]. Despite the apparent similarity, *Rhinolophus affinis* bat virus lacks a polybasic cleavage site, a key component in increasing the infectivity of SARS-CoV-2, as discussed below[138]. In addition, this bat resides in the Yunnan province, over 1,500 km from Wuhan[139]. In contrast, HCoV-HKU1 which clusters in a different clade, contains a similar polybasic cleavage site[137]. Similarly, the *Rhinolophus* bat is divergent, sharing only ~72% sequence similarity but has high similarity of the long replicase gene at ~97% nucleotide sequence similarity. The Malayan pangolin imported to the Guangdong and Guangxi provinces have 97% RBD amino acid homology in the receptor binding domain region [140].

Human to human transmission of SARS-CoV-2 became evident with its rapid spread in people with no history of exposure to the Huanan Seafood Wholesale Market [141]. Coronaviruses have caused two previous large scale out breaks in the last 20 years with the SARS outbreak of 2002-2003 and the MERs outbreak since 2015[142]. In comparison, SARS-CoV-2 is less lethal but more infectious with transmission occurring during asymptomatic/pre-symptomatic phases [143]. With a higher case-fatality rate compared with seasonal influenza, it has resulted in a global pandemic [140].

Coronaviruses are enveloped, positive-sense single stranded RNA viruses[144,145]. These viruses have a lower mutation rate compared with other RNA viruses due to proof-reading ability via 3'-to-5' exoribonuclease[146].

The SARS-CoV-2 genomic sequence bears 79% homology to the SARS-COV sequence and 50% homology to the MERS sequence[147].

The SARS-CoV-2 genome has 14 open reading frames and encodes three classes of protein (Fig 1.9)[148].

- Polyproteins pp1a and pp1b which are cleaved into 16 non-structural proteins (nsp 1-16). These are vital for viral RNA synthesis.
- 9 accessory proteins which mitigate host defences
- 4 structural proteins: spike, envelope, membrane, and nucleocapsid. These are involved in viral entry and viral assembly.

Of the 4 structural proteins, the spike protein is the most dissimilar to its SARS-CoV counterpart with ~73% nucleotide sharing[147]. The spike protein mediates access to the cell via the ACE2 receptor. The spike protein is divided into S1 and S2 subunits. The S1 unit contains the c-terminal domain, also termed the receptor binding domain (RBD) and the n-terminal domain (NTD). The spike protein is trimeric, thus containing three RBDS. The S2 subunit is divided into an upstream helix (UH) region, fusion peptide (FP), connecting region (CR), heptad repeat 1 (HR1) and HR2 and a central helix (CH). The FP region is shielded by the UH domain. At the junction of the S1 and S2 protein are four amino acid residues (PRRA). This region plays an important role in ACE2 binding (discussed below) and is absent in SARS-CoV[149]. In addition, the ORF8 protein is functionally different between SARS-CoV and SARS-CoV-2 with only 40% amino acid homology. ORF8 of SARS-CoV-2 has the potential survival advantage of not triggering intracellular stress pathways in the host due to lacking motif VLVVL[147].

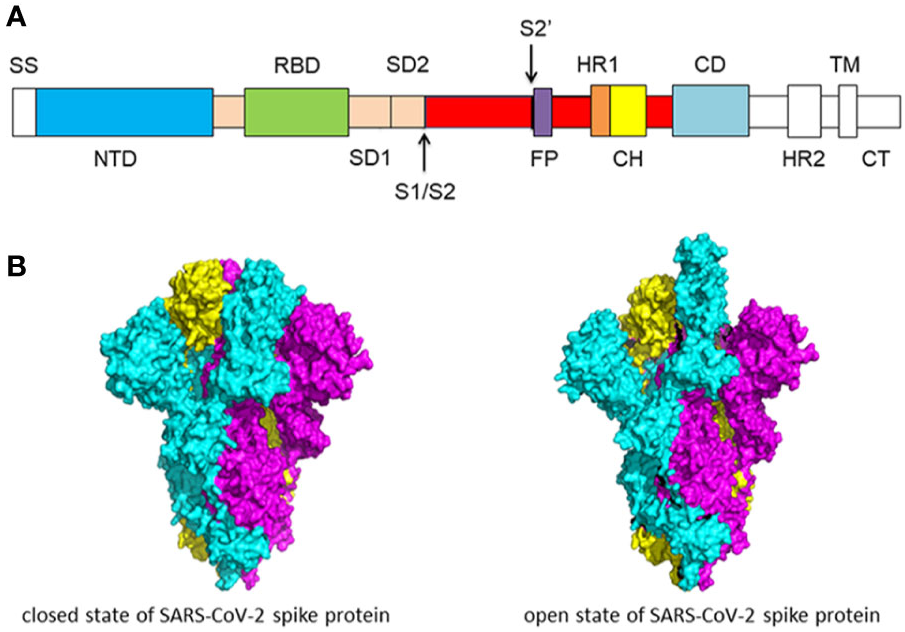


Fig 1.9 SARS-CoV-2 schematic A) Structure- S1/S2 protease cleavage site b) Cryo-EM structure showing the open and closed formation[150]

1.4.2 Receptor binding

Coronaviruses can enter a host cell via a membrane receptor, receptor-mediated plasma membrane fusion or endocytosis or by antibody dependent viral entry[144]. Angiotensin-converting enzyme II (ACE2) is the cell surface receptor for both SARS-CoV[151] and SARS-CoV-2[152,153].

RBD on the spike protein binds with the angiotensin-converting enzyme 2 (ACE2) on the cell surface, gaining entry into the cell[152,153]. The RBD takes on a closed and open conformation and not all three RBDs are synchronised in conformation (Fig 1.9)[154,155] . When in a closed conformation, RBD cannot interact with ACE2 receptor due to steric hindrance. This has the benefit of the RBD site not being constantly exposed to the adaptive immune system. A spike protein can bind 1-3 ACE2 receptors. Post binding to ACE2, S1 and S2 are cleaved at the S1-S2 and S2' cleavage site. This releases FP from its original conformation allowing it to protrude out and facilitate either cytoplasmic or endosomal membrane fusion. SARS-CoV-2 has a significantly higher affinity to ACE2 compared with SARS-CoV[156].

Furin proteases cleave S1 from S2 at the multibasic (presence of arginine) site at the S1/S2 boundary (Fig 1.10). The boundary has a furin cleavage site- proline-arginine-arginine-alanine. Furin is ubiquitously expressed. The multibasic cleavage site is present in MERs-COV but not SARS-CoV and its presence enables more efficient proteolysis[138].

Transmembrane protease serine 2 (TMPRSS2), a serine protease cleaves S1/S2 facilitating entry via endocytosis. TMPRSS2 similarly plays an important role in SARS-CoV. TMPRSS2 is co-expressed with ACE2 in nasal epithelial cells, lungs and bronchi helping to explain the virus' tissue tropism[157]. The TMPRSS2 cleavage site is within the S2 domain, termed the S2' site. This removes the UH domain which shields FP[152,153].

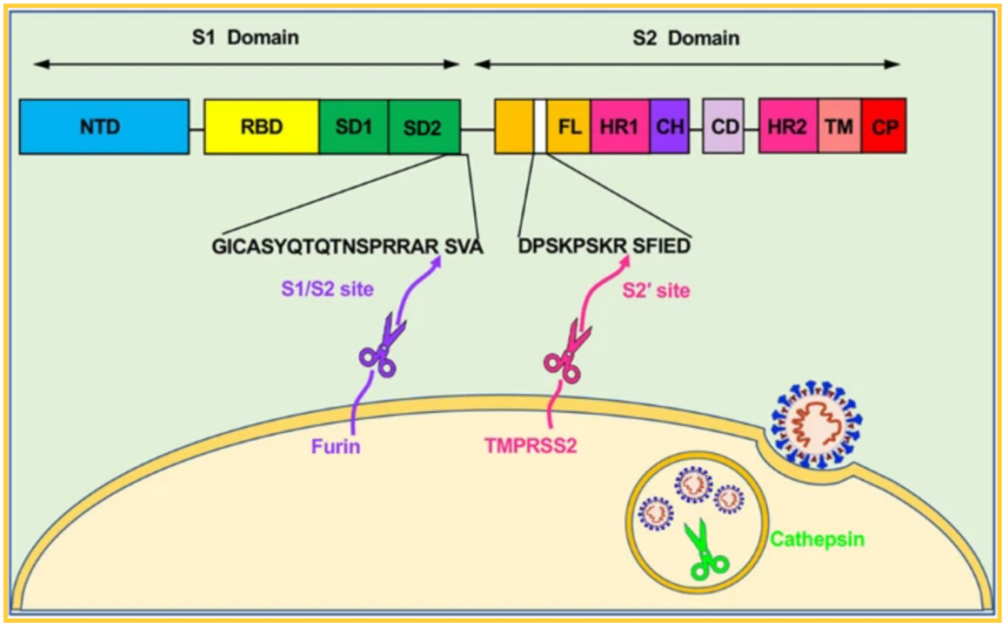


Fig 1.10 SARS-CoV-2 cleavage sites [140]

After viral attachment to ACE2, Furin cleaves the multibasic site and subsequently TMPRSS2 cleaves at the S2' site resulting in the freeing of the internal fusion protein. This is then followed by HR1 and HR2 interacting with one another forming a six-helix bundle (6-HB) fusion core. This combined draws the viral envelope and host membrane close to one another and finally membrane fusion[138](Fig 1.11).

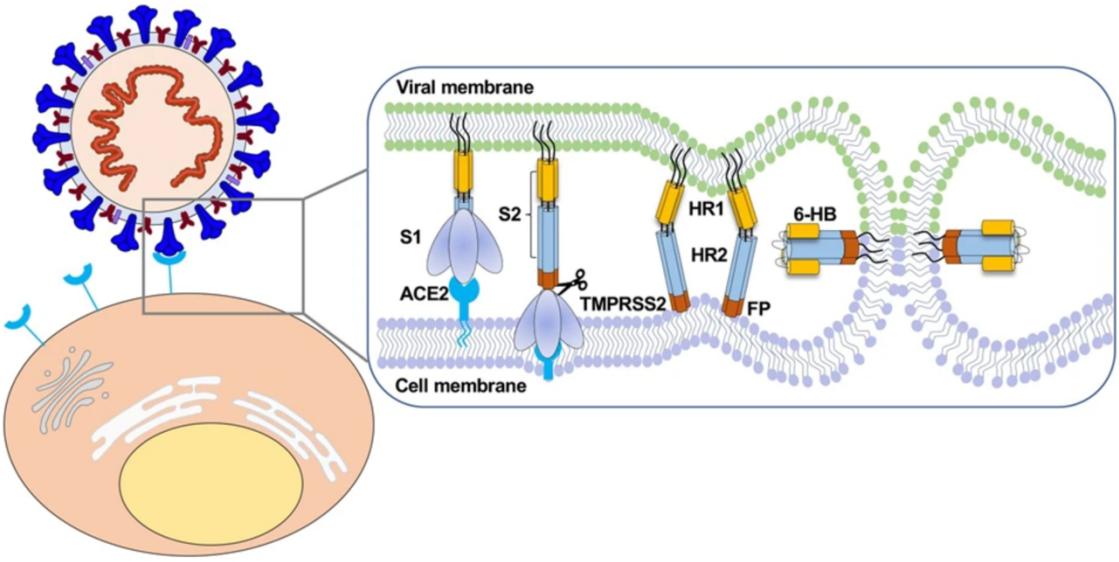


Fig 1.11 SARS-CoV-2 role of TMPRSS2 and Furin[140]

Cathepsin is localised in lysosomes and induces proteolysis post endocytosis[153]. It is an alternative to TMPRSS/Furin for viral entry into the host cell. Both entry mechanisms are utilised by SARS-CoV-2 with a preference for the endosomal pathway. Antibody- mediated SARS-CoV-2 cell entry is another form of potential viral entry although unproven unlike in SARS-CoV[144]. This process involves the binding of antigen to the Fab region and the Fc region interacting with the FcR leading to endocytosis.

Once the viral genome gains entry into the cell, ORF1a and b are translated into viral replicase proteins and are cleaved into individual NSPs thus forming RNA polymerase (nsp12). The endoplasmic reticulum is reorganised into double-membrane vesicles and is where replication occurs. The double membrane vesicle shield RNA from pattern recognition receptors. The positive strand serves as the template to generate the full length

negative-strand RNA and the subgenomic RNA are translated to structural and accessory proteins. In the ER-Golgi intermediate compartment, virion assembly occurs and is subsequently secreted from the plasma membrane. The nucleocapsid protein is important for viral packing into new virions[145].

1.4.3 Immune response

1.4.3.1 Innate response

A robust and regulated immune response is essential for control of SARS-CoV-2 (Fig 1.12).

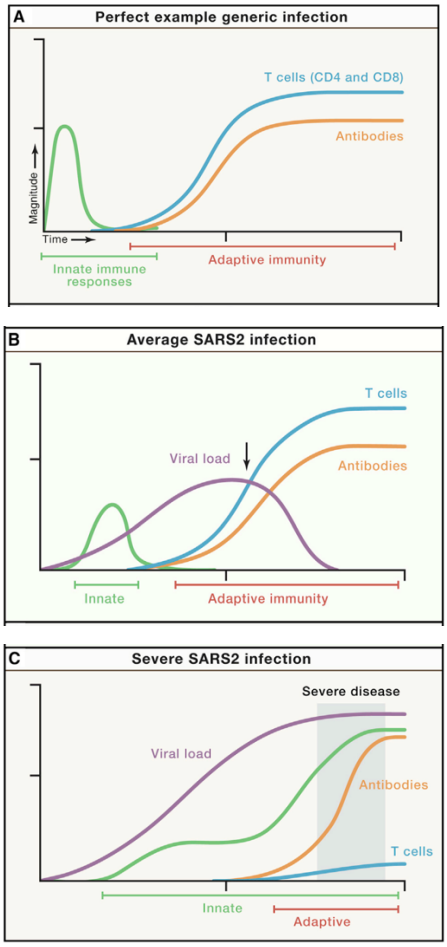


Fig 1.12 Potential immune responses to SARS-COV-2[159]

The innate immune system can control a virus by three key mechanisms:

1. Restriction of viral replication
2. Creating a hostile environment in local tissue
3. Priming of the adaptive immune system

SARS-CoV and MERs are both associated with a delayed Type 1 IFN response resulting in rapid viral replication, a heightened cytokine response and delayed adaptive immune response[160]. In some patients with SARS-CoV-2, there is evidence of a delayed/absent IFN-I and III response[161,162].

1.4.3.1.1 Interferon response in viral infection

Interferons are divided into three broad families, type I, II and III. Type I IFN consists of IFN-α, IFN-β, IFN-ε, IFN-ω and IFN-κ. Type II IFN includes IFN-γ. Type III IFN includes IFN-λ. Type I and III IFN play an important role in viral infection. Type I IFN receptors are widely expressed whilst type III IFN receptor expression is restricted to macrophages, dendritic cells, neutrophils and respiratory epithelial cells[163].

Post respiratory viral infection, the primary produces of IFNα/β are epithelial cells, endothelial cells, alveolar macrophages, NK cells, dendritic cells and inflammatory monocyte-macrophages. Type 1 IFN is produced in response to stimulation of pattern recognition receptors by microbials. The dsRNA of SARS-CoV-2 which is generated during replication is recognised by retinoic acid-inducible gene I (RIG-I) and/or melanoma differentiation gene 5 (MDA5) in the cytoplasm as well as TLRs in the endosome[164]. MDA5 and RIG-1 engage downstream receptors such as mitochondrial antiviral signalling (MAVS) protein leading to the activation of TRAF3, TBK1 and IKK. These phosphorylate IRF3/7. TLRs stimulate myeloid differentiation primary response 88 (MyD88) and TRIF. This similarly leads to the phosphorylation of IRF3 and 7. IRF3 and 7 bind IFN genes leading to the transcription of Type I IFN[165]. Type I IFN bind to IFNAR1/2 and activates TYK2 and JAK1. These phosphorylate STAT1/2 leading to the assembly of the IFN-stimulated gene factor 3 (ISGF3) complex, a trimeric complex containing IRF9 and STAT1/2. ISGF3 enters the nucleus binding to IFN-stimulated response elements and broad ISG production[166].

1.4.3.1.2 SARS-CoV-2 evasion of IFN

Accessory proteins generated by both SARS-CoV-1/2 virus aid in IFN production delay through evasion and antagonism. N and M protein, nsp10, nsp14, nsp15 and nsp16 prevent RIG-1/MDA-5 stimulation. ORF3b and ORF9b interfere with MAVS. ORF3a protein decreases IFNAR expression. Nsp1 inhibits STAT1 whilst ORF6 interferes with ISG production[167].

Study of the transcriptional response to SARS-CoV-2 after infecting normal human bronchial epithelial cells using gene enrichment analyses revealed a poor IFN-I response despite having a strong chemotactic and inflammatory response. Treatment with IFN did not

increase ISG expression suggesting that it is both IFN production and downstream function that is being antagonised(Blanco-Melo et al., 2020). In keeping with this, a study of 50 patients with COVID-19 at a median of 10 days from symptom onset of varying disease severity, identified a robust IFN response in patients with mild to moderate disease but was reduced in severe disease[162]. Arunachalam et al, similarly described a reduction in pDCs which was coupled with reduced expression of phosphorylated ribosomal protein S6, a canonical target of mTOR activation which is required in the production of IFN-α in pDCS in response to TLR stimulation. This was functionally confirmed with reduced generation of IFN in response to TLR stimuli[161]. In support, a weaker IFN-I signature was reported in the more severe COVID-19. This was associated with downregulation of translation and ribosome genes potentially due to suppressed protein translation seen in the setting of IFN. The low IFN response was attributed to reduced pDCs secondary to apoptosis[169]. Similarly, a diminished IFN-λ and type I IFN response in COVID-19 patients who become critically ill was reported[170]. Neutralising type I anti-IFN antibodies were detected in patients with severe COVID-19 at a rate of 10% and were absent in patients with mild and moderate disease. These auto-antibodies were present prior to disease suggesting that these patients represented an at-risk group[171]. Focusing on monogenic inborn errors affecting TLR3, IRF7, and IRF9 known to cause severe pneumonia and a further 10 loci associated with viral illnesses and connected to core loci, Zhang et al., showed an increase in variant predicted to cause a loss of function in a greater proportion of patients with severe COVID-19 compared with mild disease[171]. However, a defective interferon response has not been universally found. Schulte-Shrepping et al., reported an increase in ISG expression early in disease[172]. Lucas et al., did a longitudinal analysis of 113 COVID-19 patients illustrating a robust IFN response with elevated levels of IFN-α and IFN-λ in patients compared with healthy controls with duration correlating with length of hospitalisation and mortality[173].

1.4.3.1.3 Cytokine and chemokine response

Patients with severe COVID-19 have features of hyper-inflammation[174]. Heightened expression of IL-6, IL-8, IL-10, TNF, CCL2, CCL3 and CCL8 are frequently reported in the literature in association with severe disease[173,175]. SARS-CoV-2 enters and infect type II pneumocytes inducing a pro-inflammatory state with elevations in IL-1B, IL-6, CXCL8,

CXCL10, MIP1a/1b, VEGF, IFN-γ and TNF[176]. These cytokines promote migration of inflammatory neutrophils, CD4 and CD8 T cells and macrophages in the lung leading to severe lung pathology with the development of acute respiratory distress syndrome, pulmonary oedema and vascular damage[177]. Activated neutrophils produce ROS and neutrophil extracellular traps (NETs) promoting death of both virus and epithelial cells. NET formation may have a contributory role to immunothrombosis.

Immunosuppression with glucocorticoids has a survival benefit when administered in patients requiring oxygen[178]. In addition, anakinra an IL-1a/b inhibitor when administered in a select group of COVID-19 patients at risk of developing hypercytokinemia as determined by an elevated soluble urokinase plasminogen activator receptor (suPAR) serum level improved survival and decreased hospital stay. suPAR is a biomarker in early disease that predicts progression and is superior to CRP, IL-6, ferritin and D-dimers[179]. However, Canakinumab, an inhibitor in IL-1B did not show a survival benefit when an elevated CRP or ferritin was used as criteria for enrolment[180]. IL-6 blockade, on metanalysis is associated with a lower 28-day all-cause mortality[181].

1.4.3.2 Adaptive Immune system

The adaptive immune system consists broadly of CD4, CD8 and B cells. Due to its targeted approach in containing infection, its response takes time.

1.4.3.2.1 T cells

T cell lymphopenia, observed in COVID-19 is not secondary to redistribution to tissue with no evidence of T cell lymphocytosis on bronchioalveolar lavage or on quantitative analysis of lung imaging[182,183].

SARS-CoV-2 specific T cells target Spike, M, and nucleopcapsid proteins[159]. Grifoni et al., generated megapools based on predicted SARS-CoV-2 epitopes that were independent of ethnicity and HLA polymorphism. Using TCR dependent activation induced marker (AIM) assays, they found that spike specific CD4+ T cell responses were present in 100% of convalescent cases. Of the detected response, 50% was directed against spike and the

remainder against the other proteins. Given the important role CD4+ T cells play in providing "help" to B cells including promoting clonal expansion, as predicted, a correlation was present between spike-specific CD4+ T cell responses and anti-spike RBD IgG titres. SARS-CoV-2 specific CD4+ T cells associated most strongly with milder disease. Evidence of cross-reactivity was present with non-spike CD4+ T cell reactivity in healthy controls. In CD8+ T cells, given there is substantially less overlap between HLA class I allelic variants, only the 12 most prominent HLA class A and B alleles were targeted. These represent >85% of the general population. CD8 T cells responses using AIM were detected in 70% of cases. In the acute setting only 53% of cases had SARS-CoV-2 specific CD8 T+ cells. These T cells were IFN-γ producing. Similar to SARS-CoV-2 specific CD4 T+ cells, a robust SARS-CoV-2 specific CD8 T+ cell response was associated with milder disease[184].

1.4.3.2.2 B cells

Post infection with SARS-CoV-2, a rapid extrafollicular B cell response occurs, and short-lived antibody secreting cells are produced[185]. In parallel, a germinal centre response occurs, producing somatically hypermutated long lived plasma cells and class-switched B cells[186]. An increase in plasmablasts occurs within 7 days of hospitalisation, returning to baseline at 3-6 months[187]. B cell receptor repertoire sequencing reveals a polyclonal B cell population with minimal SHM [185]. This suggests that the early B cell response is generated predominantly from naïve B cells that differentiate into class-switched plasmablasts [188]. Woodruff et al., conducted an in-depth phenotypic analysis of B cell responses in COVID-19. They reported higher frequencies of activated naïve b cells and double negative B cells ($CD27^{-}IgD^{-}CD11c^{+}CD21^{-}$) compared with health which they surmised differentiated into lowly mutated plasmablasts extrafollicularly[189]. There was evidence of a small fraction of cross-reactive human coronavirus antibodies at 3 months post SARS-CoV-2 infection, identified by a high-level of SHM compared with other clones[186]. In severely ill patients, on post-mortem analysis, there was evidence of a decrease in germinal centres and in the number of BCL-6+ germinal centre B cells suggesting impairment of germinal centre formation and a sequela of impaired generation of long-lived memory B cells and plasma cells[190]. However, longitudinal studies on patients who have recovered from severe COVID-19 demonstrate

the generation of persisting high levels of SARS-CoV-2 specific antibodies and the generation of class-switched memory B cells with accumulating SHM[191,192].

Neutralising antibody titres positively correlate with disease activity[193]. This is the converse to what is seen in CD4 and CD8 SARS-CoV-2 specific responses[184]. Patients with more severe disease also have greater epitope spreading with evidence of a stronger and broader SARS-CoV-2 antibody response, using phage-display immunoprecipitation and sequencing technology[194]. The higher titres and complexity of antibodies in severe disease may be a function of time and persisting viral disease, giving rise to extended antibody evolution compared that seen in mild disease.

Seroconversion occurs in most people within 5-15 days of symptom onset[195]. The spike protein is target for neutralising antibodies with >90% targeting the receptor binding domain component[185]. Antibodies targeting the N-terminal domain can also be neutralising[196]. Antibodies are also formed against the nucleocapsid, an important component in virion replication and packing which is expressed heavily during active infection.

Instead of the usual temporal relationship between the formation of IgM and IgG antibodies, spike IgG, IgA and IgM develop simultaneously. IgM and IgA titres wane at 7-10 weeks whilst IgG titres remain elevated for 3-8 months before declining[195,197]. Antibodies patterns in serum appear similar with saliva with IgG titres remaining stable over serval months whilst IgM and IgA decline. IgM and IgG levels correlated well in serum and saliva whilst IgA does not[198].

1.4.3.2.3 B cell memory

Antibody titres in SARS-CoV and MERS appears to wain within 2-3 years[199]. Studies to date on SARS-CoV-2 are promising with evidence of RBD memory specific B cells in patients with all degrees of disease severity at 150 from symptom onset[186,192]. Dan et al., demonstrated the presence of S, RBD, and N memory B cells out at 8 months and further showed that the titres increased for the first 4 months before plateauing[200].

1.4.3.2.4 Pre-existing cross-reactivity

A small proportion of pre-existing antibodies, generated post human coronavirus infections can bind to SARS-CoV-2. <1% IgG antibodies can bind to SARS-CoV-2 RBD, 4-5% can bind to the full-length SARS-CoV-2 spike protein and 10-16% are able to bind to the N protein. Of the antibodies that can bind to the full-length SARS-CoV-2 spike protein, the main target is the S2 which has the greatest sequence homology to the human coronavirus infections [201]. Antibodies that bind to the N protein are non-neutralising. During SARS-CoV-2 infection, titres of these cross-reactive antibodies are boosted. It's unclear if cross-reactive antibodies play a key role in immune defence with Anderson et al., finding no correlation between pre-pandemic cross reactive antibodies and outcome[202]. Sagar et al., on the other hand found a correlation between previous reports of human coronavirus infections.

1.5 Aim and objectives

SARS-CoV-2 causing COVID-19 has resulted in a world-wide pandemic. The aim of this study was to better understand the virus and its impact on the immune system and disease severity. Specifically, the objectives were as follows:

I. To perform a deep immune-phenotyping and transcriptomic analysis to understand cellular changes;
II. To perform a BCR repertoire analysis to understand changes in clonality, class-switching and v gene usage.

2. Materials and Methods

This chapter contains a description of patient recruitment, immunophenotyping and of the concepts and techniques used in bulk RNAseq and B cell receptor sequencing.

2.1 Participant recruitment

2.1.1 COVID-19 participants

Study participants were recruited over a period between 31/3/2020 and 20/7/2020. Patients were recruited from Addenbrooke's and Royal Papworth Hospital with a confirmed positive nucleic acid amplification test (NAAT) COVID-19 test. In addition, Health Care Workers identified through staff screening as PCR positive for SARS-CoV-2[203] were included.

Recruitment of inpatients at Addenbrooke's Hospital and Health Care Workers was undertaken by the NIHR Cambridge Clinical Research Facility outreach team and the NIHR BioResource research nurse team. Ethical approval was obtained from the East of England – Cambridge Central Research Ethics Committee ("NIHR BioResource" REC ref 17/EE/0025, and "Genetic variation AND Altered Leucocyte Function in health and disease - GANDALF" REC ref 08/H0308/176). All participants provided informed consent.

Inpatients were sampled at study entry, and then at regular intervals whilst in hospital (approximately weekly up to 4 weeks, and then every 2 weeks up to 12 weeks). Discharged patients were invited to provide a follow-up sample 4-8 weeks after study enrolment. Health care workers were sampled at study entry, and subsequently after approximately 2 and 4 weeks. At each time-point, blood samples were drawn in EDTA, sodium citrate, serum and PAXgene Blood RNA tubes (BD Biosciences) and processed by the CITIID-NIHR COVID BioResource Collaboration group.

2.1.2 Vaccine Participants

2.1.2.1 SARS-CoV-2

Community participants or health care workers receiving the first dose of the BNT162b2 vaccine between the 14th of December 2020 to the 29th of January 2021 were consecutively recruited at Addenbrookes Hospital into the COVID-19 cohort of the NIHR Bioresource. The study was approved by the East of England – Cambridge Central Research Ethics Committee (17/EE/0025).

2.1.2.2 Influenza

Community participants receiving a dose of Adjuvanted Trivalent Influenza Vaccine (Surface Antigen, Inactivated) Adjuvanted with MF59C.1 (2020/2021 SEASON) were recruited. The study was approved by the East of England – Cambridge Central Research Ethics Committee (REC ref: 20/SW/0134, IRAS id: 287814, CBR#: 213). Samples were taken at baseline and subsequently day 7 and day 30 from vaccination. Paired analysis was performed with day 0 samples used as a healthy comparison.

2.1.3 Healthy controls

SARS-CoV-2 negative controls were recruited during Health Care Worker health screening after confirmation of negative NAAT and serology. Additional healthy control samples were obtained under the Gandalf ethics (08/H0308/176) and had been recruited prior to 2020. Thus, before the existence of COVID-19.

2.1.4 Inflammatory Bowel disease

2.1.4.1 Peripheral Blood BCR repertoire

Patients with inflammatory bowel disease were recruited from a specialist clinic at Addenbrooke's Hospital prior to starting treatment. Diagnosis was made based on endoscopic findings, histology, radiology and clinical history. Ethical approval was obtained from the Cambridge Local Research Ethics Committee (reference numbers 04/023,

08/H0306/21, 08/H0308/176) and Eastern NHS Multi Research Ethics Committee (07/MRE05/44).

2.1.4.2 Lymph Node BCR repertoire

Lymph nodes were collected from patients with Crohn's disease who required surgery secondary to disease complications including stenosis and fistula formation. A mesenteric lymph node was taken adjacent to an area of inflamed bowel (Medical University of Vienna's Institutional Review Board (EK number: 1480/2016).

2.2 Clinical data collection

Clinical data were retrospectively collected by review of medical files, laboratory test results and in-patient medications using Epic electronic health records (Addenbrooke's Hospital) and from MetaVision ICU (RPH ITU). Health care workers were classified into 2 groups (A and B) according to whether they were asymptomatic (group A) or had possible COVID-19 symptoms (group B) at the time of PCR testing. Symptomatic disease was defined as new-onset fever (> 37.8 C), cough, loss of sense of smell, hoarseness, nasal discharge/congestion, shortness of breath, wheeze, headache, muscle aches, nausea, vomiting and/or diarrhoea.

Participants in group A were further sub-grouped according to whether they were completely asymptomatic (n = 8) or had had any of the above COVID-19 symptoms before PCR testing (n = 10, median time from symptoms to COVID-19 PCR test 26 days, range 9-42 days).

Group B participants included both staff who were self-isolating because of COVID-19 symptoms (n = 30), and staff members who reported fit for duty but had symptoms that did not reach the threshold for self-isolation at the time (n = 10).

Hospital patients were assigned to 3 severity groups, reflective of the maximal intensity of respiratory support received during their hospital stay:

- group C: did not receive any supplemental oxygen.

- group D: received supplemental oxygen using low flow nasal prongs, simple face mask, Venturi mask or non-rebreather face mask.

- group E: received non-invasive ventilation (NIV), mechanical ventilation or ECMO. Deceased patients requiring supplemental oxygen (not ventilation) were also assigned to group E (see Table 2.1).

Oxygen requirements that were not related to COVID-19 were not considered during classification. In particular, 2 patients who received low flow supplemental oxygen for non-COVID-19 indications (ascitic splinting in decompensated cirrhosis in one case, and recovery from anaesthesia after orthopaedic surgery in the other) were assigned to group C. Cases in group C were further sub-classified according to chest radiology results (X-ray and, if available, CT scan), as:

- abnormal radiology: chest X-ray/ CT scan with changes compatible with COVID-19
- normal radiology: chest X-ray/ CT scan without abnormalities compatible with COVID-19 (reported as normal or showing lung changes diagnostic of conditions other than COVID-19).

Immunological parameters were analysed according to time since onset of symptoms, or otherwise time since positive SARS-CoV-2 NAAT (group A). Seven cases admitted to hospital for COVID-19 had no date of onset of symptoms documented in the medical records. In these cases, the date of onset of symptoms was estimated as follows: hospital admission date - median time from symptoms to hospital admission in patients admitted for COVID-19.

Clinical features of study participants, stratified by group A-E

	A	B	C	D	E
n	18	40	46	37	60
Gender (% male)	22.2%	22.5%	54.3%	64.9%	75.0%
Age (years, mean (SD))	32.9 (12.7)	36.0 (11.8)	58.0 (16.9)	64.4 (15.1)	57.0 (14.9)
Days from COVID-19 symptoms to enrollment (days, mean (SD))	NA	6.5 (2.9)	11.4 (6.7)	10.6 (8.1)	24.6 (14.3)
COVID-19 chest radiology	NA	NA	50.0%	89.2%	100%
Non-COVID19 admissions	NA	NA	30.2%	8.1%	6.7%
Haemoglobin (g/L, mean (SD))	NA	NA	124.8 (16.0)	121.6 (18.0)	95.2 (16.8)
Serum creatinine (µmol/L, mean (SD))	NA	NA	82.9 (40.1)	117.5 (154.7)	103.5 (129.3)
Serum albumin (g/L, mean (SD))	NA	NA	32.4 (7.1)	28.0 (6.3)	24.4 (7.2)
LOS (days, median (IQR))	NA	NA	4 (1.25-10)	10 (6-16)	44 (33.7-63.2)
Admitted to ITU	NA	NA	0%	13.5%	90.0%
Deceased in hospital	NA	NA	2.2%	0.0%	30.0%
Hypertension	NA	NA	47.8%	43.2%	48.3%
CAD	NA	NA	8.7%	24.3%	16.7%
Other heart condition	NA	NA	10.9%	18.9%	13.3%
Diabetes mellitus	NA	NA	26.1%	29.7%	43.3%
CKD	NA	NA	8.7%	16.2%	8.3%
PVD	NA	NA	6.5%	8.1%	8.3%
CVA/TIA	NA	NA	10.9%	2.7%	6.7%
COPD	NA	NA	6.5%	18.9%	5.0%
Asthma	NA	NA	21.7%	10.8%	10.0%
Other lung disease	NA	NA	10.9%	16.2%	10.0%
Cancer	NA	NA	4.4%	5.4%	1.7%
Haematological cancer	NA	NA	2.2%	5.4%	0.0%
Corticosteroids	NA	NA	19.6%	10.8%	10.0%
Immunosuppressive treatment	NA	NA	17.4%	16.2%	5.0%

SD is standard deviation, and IQR is interquartile range.
COVID-19 chest radiology: chest X-ray/ CT scan showed changes compatible with COVID-19, as opposed to normal findings or lung changes diagnostic of other conditions.
Non-COVID19 admissions: cases where COVID-19 was diagnosed during the hospital stay in patients initially admitted to hospital for reasons unrelated to COVID-19
Haemoglobin, serum albumin and serum creatinine: results from routine lab tests on the day of study enrollment, or closest result up to 2 days before. The included test results are available for at least for 75% of each severity group.
LOS: length of hospital stay (days from hospital admission to discharge, transfer or death in hospital)
Hypertension: history of hypertension, defined as blood pressure ≥140/80 on multiple occasions, or on treatment with any medication explicitly employed to reduce blood pressure
CAD: history of coronary artery disease, defined as myocardial infarction, angina, coronary artery stenting or coronary artery bypass grafting
Other heart condition: history of any other chronic cardiac disease (not CAD/hypertension), e.g. heart failure, congenital heart disease, cardiomyopathy, rheumatic heart disease
CKD: history of chronic kidney disease, defined as any of estimated glomerular filtration rate < 60 mL/min/1.73m^2, dialysis or kidney transplant
PVD: history of peripheral vascular disease, defined as intermittent claudication or past bypass for chronic arterial insufficiency, history of gangrene or acute arterial insufficiency, or thoracic/abdominal aneurysm (≥6 cm)
CVA/TIA: history of a cerebrovascular accident or transient ischemic attacks
COPD: history of chronic obstructive pulmonary disease
Other lung disease: history of other chronic pulmonary disease (non asthma/COPD), e.g. cystic fibrosis, bronchiectasis, interstitial lung disease
Cancer: current solid organ malignancy (active or in the last 5 years), except non-melanoma skin cancers
Corticosteroids: history of treatment with systemic corticosteroids in the 14 days prior to hospital admission/presentation
Immunosuppressive treatment: history of treatment with immunosuppressants (excluding corticosteroids) in the 14 days prior to hospital admission/presentation, or chemotherapy/biologic drugs in the previous 6 months

Table 2.1 Clinical features of study participants. Participants are grouped according to disease severity. Table made by F.Mescia.

2.3 Peripheral blood mononuclear cell preparation

Each participant provided 27 mL of peripheral venous blood collected into 9 mL sodium citrate tube. Peripheral blood mononuclear cells (PBMCs) were isolated using Leucosep tubes (Greiner Bio-One) with Histopaque 1077 (Sigma) by centrifugation at 800xg for 15min at room temperature. PBMCs at the interface were collected, rinsed twice with autoMACS running buffer (Miltenyi Biotech) and cryopreserved in FBS with 10% DMSO. All samples were processed within 4 hours of collection.

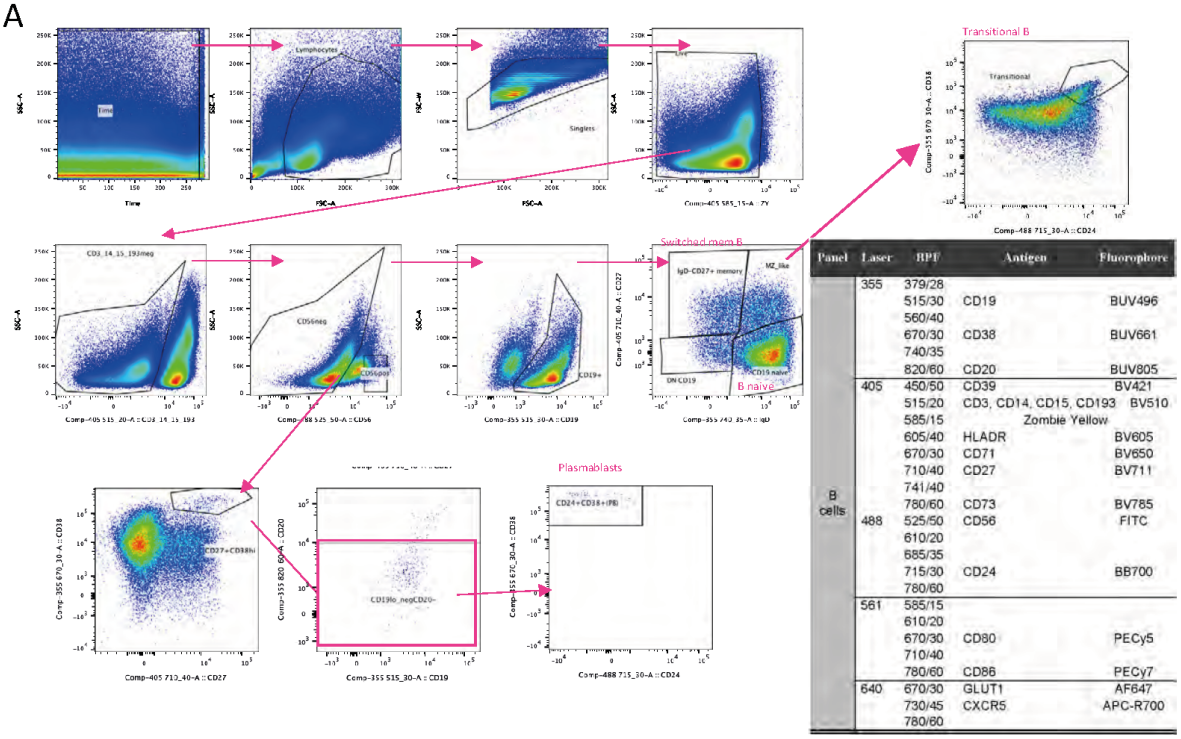
2.4 Flow immunophenotyping

Five distinct antibody cocktails were used to label approximately 10^6 PBMCs using standard methods. T regulatory cells were fixed and permeabilized following surface staining prior to the addition of intracellular antibodies. Samples were stored at 4°C and acquired within 4 h using a 5-laser BD Symphony X-50 flow cytometer. Single colour compensation tubes (BD CompBeads) or cells were prepared for each of the fluorophores used at the start of each flow cytometer run.

For direct enumeration of T, B and NK cells, an aliquot of whole blood (50 µl) was added to BD TruCount tubes with 20µl- BD Multitest 6-color TBNK reagent (BD Biosciences) and processed as per the manufacturer's instructions.

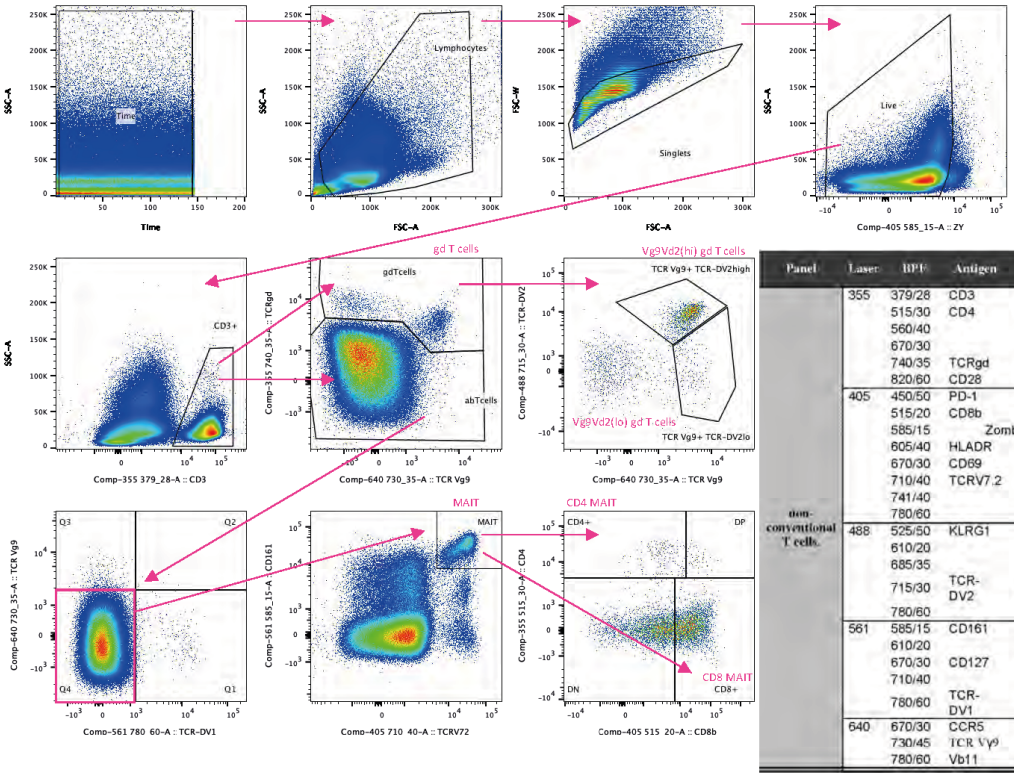
Samples were gated in FlowJo v10.2 according to the schema set out below (Fig 2.1). The number of cells falling within each gate were recorded. For analysis, these were expressed as an absolute concentration of cells per ml, calculated using the proportions of daughter populations present within the parent population determined using the BD TruCountsystem.

A



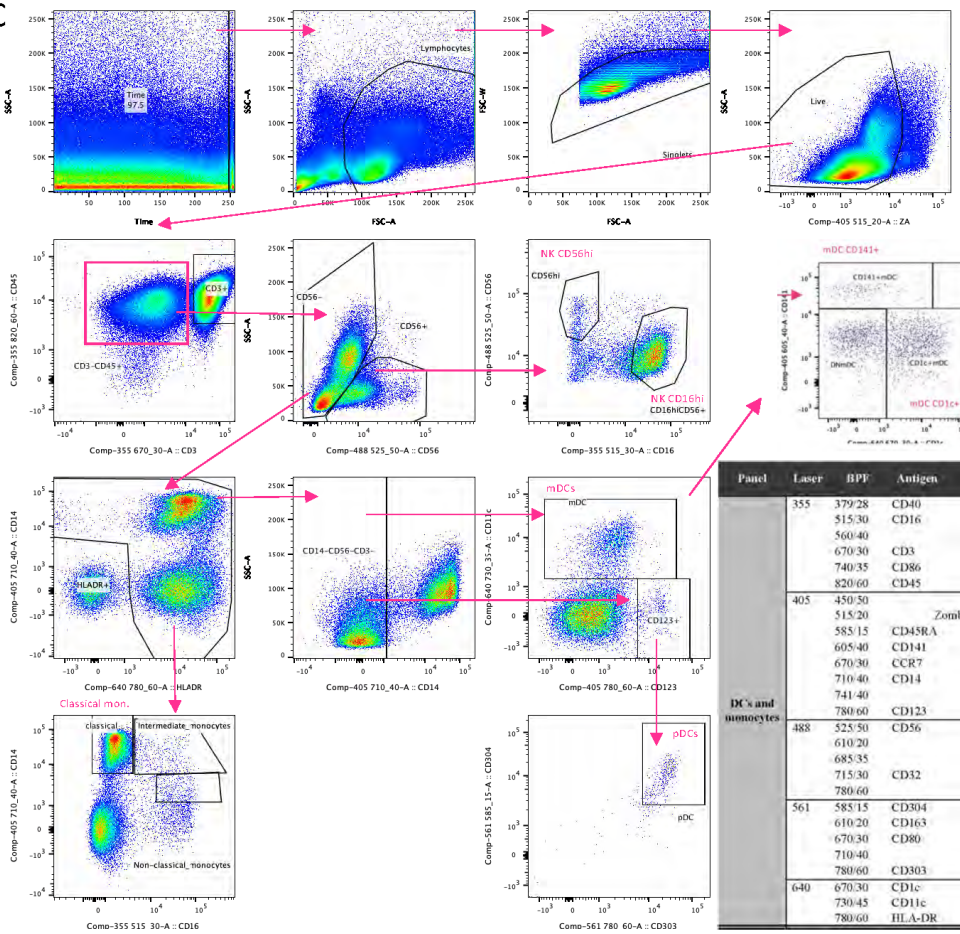
Panel	Laser	BPF	Antigen	Fluorophore
B cells	355	379/28		
		515/30	CD19	BUV496
		560/40		
		670/30	CD38	BUV661
		740/35		
		820/60	CD20	BUV805
	405	450/50	CD39	BV421
		515/20	CD3, CD14, CD15, CD193	BV510
		585/15	Zombie Yellow	
		605/40	HLADR	BV605
		670/30	CD71	BV650
		710/40	CD27	BV711
		741/40		
		780/60	CD73	BV785
	488	525/50	CD56	FITC
		610/20		
		685/35		
		715/30	CD24	BB700
		780/60		
	561	585/15		
		610/20		
		670/30	CD80	PECy5
		710/40		
		780/60	CD86	PECy7
	640	670/30	GLUT1	AF647
		730/45	CXCR5	APC-R700
		780/60		

B



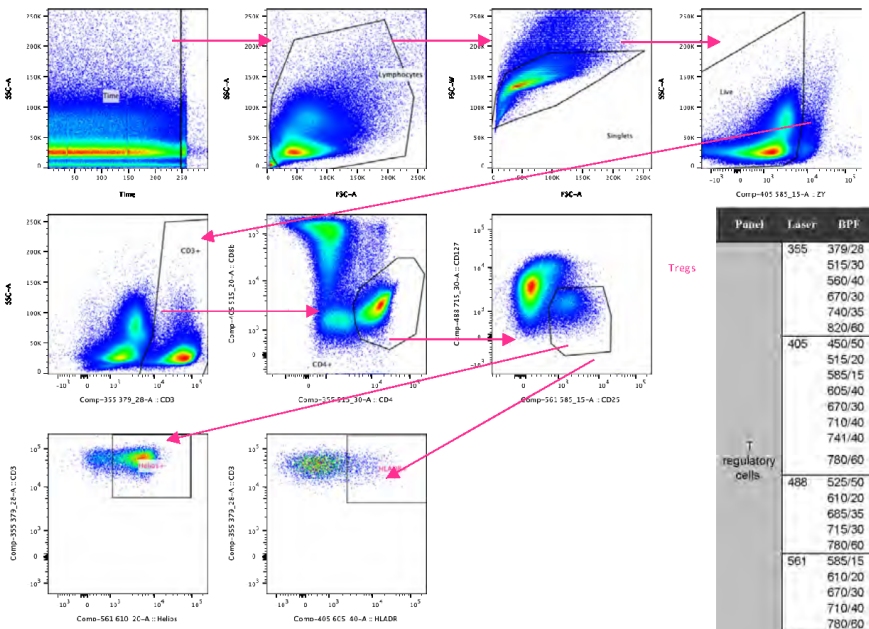
Panel	Laser	BPF	Antigen	Fluorophore
non-conventional T cells	355	379/28	CD3	BUV395
		515/30	CD4	BUV496
		560/40		
		670/30		
		740/35	TCRgd	BUV737
		820/60	CD28	BUV805
	405	450/50	PD-1	BV421
		515/20	CD8b	BV480
		585/15	Zombie Yellow	
		605/40	HLADR	BV605
		670/30	CD69	BV650
		710/40	TCRV7.2	BV711
		741/40		
		780/60		
	488	525/50	KLRG1	FITC
		610/20		
		685/35		
		715/30	TCR-DV2	PerCPCy5.5
		780/60		
	561	585/15	CD161	PE
		610/20		
		670/30	CD127	PECy5
		710/40		
		780/60	TCR-DV1	PECy7
	640	670/30	CCR5	AF647
		730/45	TCR Vγ9	AF700
		780/60	Vb11	APCVio770

C



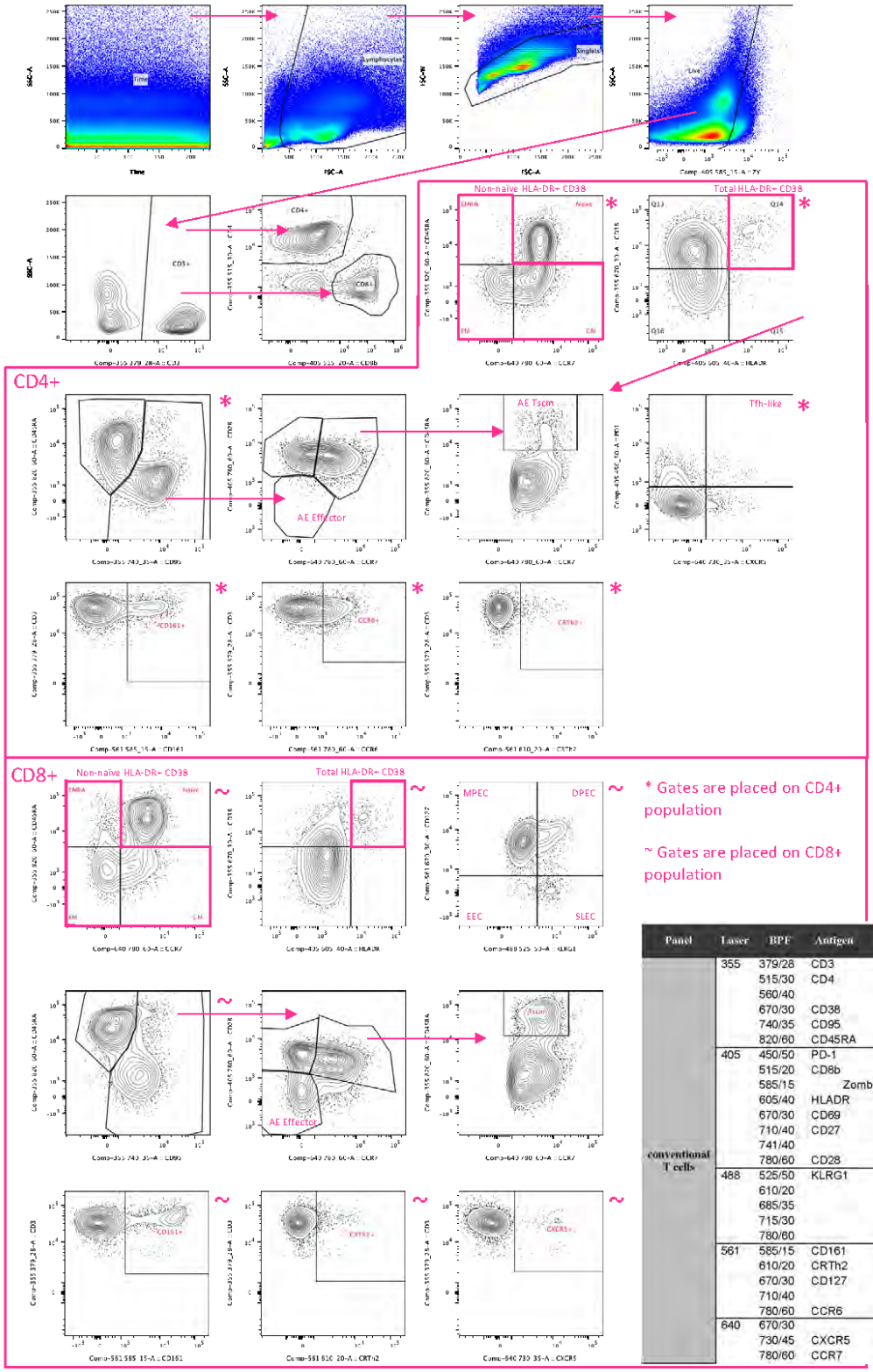
Panel	Laser	BPF	Antigen	Fluorophore
DCs and monocytes	355	379/28	CD40	BUV395
		515/30	CD16	BUV496
		560/40		
		670/30	CD3	BUV661
		740/35	CD86	BUV737
		820/60	CD45	BUV805
	405	450/50		
		515/20	Zombie Aqua	
		585/15	CD45RA	BV570
		605/40	CD141	BV605
		670/30	CCR7	BV650
		710/40	CD14	BV711
		741/40		
		780/60	CD123	BV786
	488	525/50	CD56	FITC
		610/20		
		685/35		
		715/30	CD32	BB700
		780/60		
	561	585/15	CD304	PE
		610/20	CD163	PE-CF594
		670/30	CD80	PECy5
		710/40		
		780/60	CD303	PE-Vio770
	640	670/30	CD1c	AF647
		730/45	CD11c	AF700
		780/60	HLA-DR	APC-H7

D



Panel	Laser	BPF	Antigen	Fluorophore
T regulatory cells	355	379/28	CD3	BUV395
		515/30	CD4	BUV496
		560/40		
		670/30		
		740/35		
		820/60	CD45RA	BUV805
	405	450/50	PD-1	BV421
		515/20	CD8b	BV480
		585/15	Zombie Yellow	
		605/40	HLADR	BV605
		670/30	CCR7	BV650
		710/40		
		741/40		
		780/60	CD73	Brilliant Violet 785™
	488	525/50		
		610/20		
		685/35		
		715/30	CD127	PerCP efluor710
		780/60		
	561	585/15	CD25	PE
		610/20	Helios	Pedazzle
		670/30		
		710/40		
		780/60	CCR4	PEVio770
	640	670/30	FoxP3	APC
		730/45	CXCR5	APC-R700
		780/60	CD39	APC-fire

E



Panel	Laser	BPF	Antigen	Fluorophore
conventional T cells	355	379/28	CD3	BUV395
		515/30	CD4	BUV496
		560/40		
		670/30	CD38	BUV661
		740/35	CD95	BUV737
		820/60	CD45RA	BUV805
	405	450/50	PD-1	BV421
		515/20	CD8b	BV480
		585/15	Zombie Yellow	
		605/40	HLADR	BV605
		670/30	CD69	BV650
		710/40	CD27	BV711
		741/40		
		780/60	CD28	BV785
	488	525/50	KLRG1	FITC
		610/20		
		685/35		
		715/30		
		780/60		
	561	585/15	CD161	PE
		610/20	CRTh2	PE-dazzle
		670/30	CD127	PECy5
		710/40		
		780/60	CCR6	PECy7
	640	670/30		
		730/45	CXCR5	APC-R700
		780/60	CCR7	APC-fire

Fig 2.1 Gating strategy used to define cell populations. A) B cell, B) non-conventional T cell, C) DCs and monocyte, D) T regulatory cells, and E) conventional T cell panels. Fig made by L.Turner.

2.5 CyTOF

Mass cytometric analysis was performed on a subgroup of patients and healthy controls (249 samples). Whole blood samples (270µl) were stained using the Fluidigm Maxpar® Direct Immune Profiling Assay according to the manufacturer's instructions. Samples were cryopreserved at −80°C following staining and thawed for analysis within 4 weeks. Samples were acquired using a Fluidigm Helios mass cytometer and normalized using the CyTOF Software v6.7.1016. FCS files generated were analyzed using the Maxpar® Pathsetter software v2.0.45 (Verity Software House, Topsham, ME). Standard settings were used to generate immune cell frequencies for 37 immune cell populations. Absolute cell numbers were calculated using the proportions of these immune cell populations within the parent populations determined by BD TruCount.

2.6 Reticulocyte counts

Reticulocyte numbers were measured using a Sysmex XN-1000 hematology analyzer according to manufacturer instruction and as previously described[204]. Briefly, Sysmex technology uses three signals to define the physiological and structural properties of cells and to distinguish reticulocytes from the other blood cells: forward scatter, side scatter and side fluorescent light. These measurements rely on the similar electromagnetic radiation and fluid dynamics concepts of a flow cytometer; reticulocyte specific fluorescent probes are covered by a patent deposited by Sysmex Corporation (i.e., Fluorocell RET, cat# BN-337-547).

2.7 Complement

Complement activation was assessed by measuring C3 activation products (C3a and C3c) together with the terminal complement complex (TCC) as an end product of the complement cascade. Concentrations of these complement components were measured in EDTA plasma from patients using commercially available enzyme-linked immunosorbent

assays (ELISA) kits (HK354 (C3a), HK368 (C3c), HK328 (TCC), Hycult Biotech, Uden, the Netherlands) according to the manufacturer's protocols.

2.8 CRP

High sensitivity CRP was measured using the standard assay by the Core Biochemical Assay Laboratory (CBAL) at Cambridge University Hospitals NHS Foundation Trust.

2.9 Cytokines

IL-6, IL-10, IL-1β, TNF-α and IFN-γ were measured in serum from patients and healthy controls by high sensitivity Base Kit HS Cytokine A Mag (cat# LHSCM000, R&D Systems / Biotechne) on a Luminex analyzer (Bio-Plex, Bio-Rad, UK) as standard clinical assay performed by the Clinical Immunology Laboratory at the Department of Biochemistry and Immunology, Addenbrooke's Hospital Cambridge.

2.10 SARS-CoV-2 serology

Quantification of Spike SARS-CoV-2 specific antibodies was performed by ELISA as described by Xiong X et al. [205]. Serum samples collected at time of enrolment and at 4-8 weeks follow-up with an AUC calculated.

2.11 SARS-CoV-2 neutralisation assays

2.11.1 SARS-CoV-2 neutralisation assay

The clinical isolate SARS-CoV-2/human/Liverpool/REMRQ0001/2020 was used (received from Ian Goodfellow, University of Cambridge), isolated by Lance Turtle (University of Liverpool) and David Matthews and Andrew Davidson (University of Bristol) [206,207]. Sera were heat-inactivated at 56°C for 30 min, then frozen in aliquots at −80°C. Neutralising antibody titers at 50% inhibition (NT50s) were measured [208]. HEK293T reporter cells expressing Renilla luciferase (Rluc) and SARS-CoV-2 Papain-like protease-activatable circularly permuted firefly luciferase (FFluc) were seeded in flat-bottomed 96-well plates. The following day, SARS-CoV-2 viral stock (MOI = 1) was pre-incubated with a 3-fold dilution series of each serum for 2 hours at 37°C, then added to the cells. After 24 hours, cells were lysed in Dual-Glo Luciferase Buffer (Promega) diluted 1:1 with PBS and 1% NP-40. Lysates were transferred to white half-area 96-well plates, and infectious virus quantitated as the ratio of FFluc/Rluc activity measured using the Dual-Glo kit (Promega) according to the manufacturer's instructions. Experiments were conducted in duplicate. To obtain NT50s, FFluc/Rluc ratios were analyzed using the Sigmoidal, 4PL, X is log(concentration) function in GraphPad Prism.

2.11.2 Pseudotyped virus neutralization assays used post SARS-COV-2 vaccination

Virus neutralization assays were performed on 293T cells transiently transfected with ACE2 and TMPRSS2 using SARS-CoV-2 spike pseudotyped virus expressing luciferase[209]. Pseudotyped virus was incubated with serial dilutions of heat-inactivated human serum samples or sera from vaccinated individuals in duplicate for 1 h at 37 °C. Virus and cell-only controls were also included. Then, freshly trypsinized 293T ACE2/TMPRSS2-expressing cells were added to each well. Following 48 h incubation with 5% CO2 at 37 °C, luminescence was measured using the Steady-Glo Luciferase assay system (Promega). Neutralization was calculated relative to virus-only controls. Dilution curves were presented as a mean neutralization with s.e.m. ID50 values were calculated in GraphPad Prism. The limit of detection for 50% neutralization was set at an ID50 of 20. The ID50 within groups were

summarized as a geometric mean titre (GMT) and statistical comparison between groups were made with Mann–Whitney or Wilxocon ranked sign tests.

2.12 Bulk RNA-Sequencing

2.12.1 Library preparation

RNA was isolated from blood samples stored in Paxgene tubes and quantified using the RNA HS assay on the Qubit. Libraries were prepared using SMARTer® Stranded Total RNA-Seq it v2 - Pico Input Mammalian (Takara) using 10ng of RNA as starting input. Library quality and quantity were validated by capillary electrophoresis on an Agilent 4200 TapeStation. Libraries were subsequently pooled at equimolar concentrations and paired-end sequenced (75bp) on 4 lanes of the Hiseq4000 (Illumina) to achieve 10 million reads per samples.

2.12.2 Reads mapping and quantification

The quality of raw reads was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). SMARTer adaptors were trimmed, along with sequencing calls with a Phred score below 24 using Trim_galore v.0.6.4 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.) Residual rRNA reads were depleted in silico using BBSplit (https://github.com/BioInfoTools/BBMap/blob/master/sh/bbsplit.sh).
Alignment was performed using HISAT2 v.2.1.0[210] against GRCh38 genome achieving more than 95% alignment rate. A count matrix was generated in R using featureCounts (Rsubreads - packages) and converted into a DGEList (EdgeR package) for downstream analysis. The analysis pipeline is depicted below in Fig 2.2.

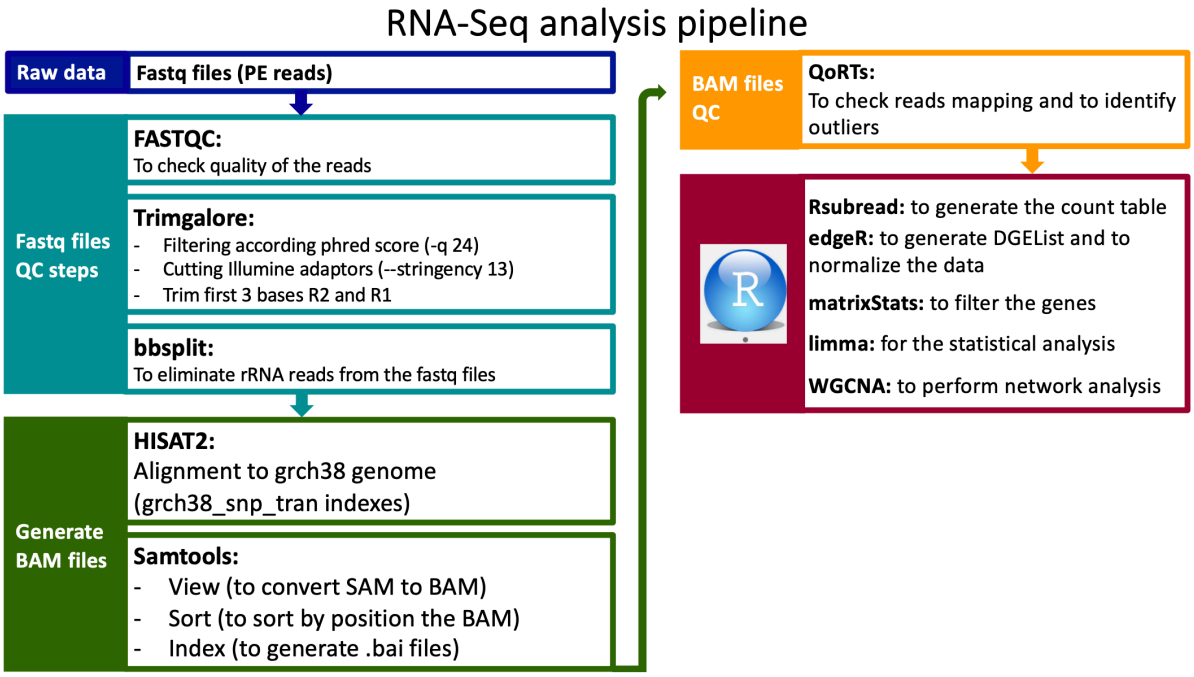


Fig 2.2 Bulk RNAseq processing and analysis pipeline.

2.13 Downstream analytical approaches in transcriptomics

2.13.1 Overview

In the following section we discuss the data analysis tools used in the analysis. These include machine learning and data mining tools.

2.13.1.1 Machine learning

Machine learning refers to computer-generated algorithms that model data without being explicitly programmed. In transcriptomics, machine learning can be used as a predictive tool or to aid in understanding biological processes.

2.13.1.2 Generative and Discriminative models

A generative model is used where the primary goal is interpretability whilst a discriminative model is used for higher prediction accuracy. The generative approach uses the full extent of information available to build a model and requires a large amount of data for accurate

modelling whilst a discriminative approach focuses on information at the boundary of the outcomes, as seen in support vector machines (see below).

2.13.1.3 Validation

Once a model is generated, validation is imperative. A model is generated on a "training set" and subsequently validated on a "test set". The test set may be an independent data set or a partition of the primary data set that is not used in the training. Partitioning can be achieved using cross-validation (no replacement of samples) or bootstrapping (replacement occurs).

High bias refers to underfitting of the data whilst high variance refers to overfitting of the data. If there is high bias, the algorithm will inadequately model the complexity of the data and will perform poorly in both the training and test set. If there is high variance, random noise in the training set will have influenced the training of the model resulting in a far more complex model and lead to poor fitting in the test data. Limiting the complexity of the model by reducing features or using regularisation can help prevent overfitting.

2.13.1.4 Dimension Reduction

Transcriptomics are high-dimensional data. It has the problem of p >> n where n refers to the number of samples and p the number of features.

Dimension reduction refers to projecting high dimensional data to a low dimensional space by reducing datasets to the features of greatest influence. Principal component analysis (PCA) and multi-dimensional scaling (MDS) are techniques used to visualise data in lower dimensions. For example, Principal component 1 (PC1), aligns the data along the axis of greatest variance and PC2 creates an orthogonal axis which accounts for the direction of the second greatest variance. Visualising the data along the primary axis of variance allows for more intuitive interpretation of the data.

2.13.1.5 Supervised Learning

Supervised learning involves the analysis of labelled data. That is, a model is built to understand the relationship between an input (e.g., samples' feature values) and a known

output (e.g., disease). Supervised learning can be grouped into two main problems, “regression” and “classification”. Regression is where the outcome that is being predicted is continuous, such as age or height. Linear modelling is a common approach. A classification problem is where the predicted outcome is discrete, such as sex. Logistic regression is a technique employed when the outcome is binary. A sigmoid curve is applied to a linear function to fit the data. K nearest neighbour can be used in both regression and classification. In regression, a continuous outcome is estimated from a data point’s ‘k’ nearest neighbours. The nearest neighbours are determined using a distance metric. The most used is Euclidean distance. In classification, group membership of an unlabelled sample is determined by the majority membership of its ‘k’ nearest neighbours. Other examples of supervised learning include, support vector machines, elastic net, decision trees and random forests. Supervised learning requires optimising the bias and variance trade off typically by using regularization.

2.13.1.6 Unsupervised Learning

Unsupervised learning is where an algorithm is used to find structure in the data with no specific outcome variable. Clustering is a technique used and in broad terms can be divided into a hierarchical approach where trees are built based on a distance metric and a partitioning approach which separates samples into non-overlapping groups as seen with k-means clustering and density-based spatial clustering of applications with noise.

2.13.1.7 Data Integration

Multi-omics, as the name suggests, describes a data set where for a given sample there are multiple ‘omic’ layers of information. This data can be analysed layer by layer, for example, analysing the neutrophil transcriptome across healthy and diseased. An alternative is using an integrative approach where multiple layers of information are simultaneously analysed. For example, in a data set incorporating transcriptomes, DNA methylation and chromatin accessibility, a shared axis of variation across all datasets can be used to give a more comprehensive view of biology.

2.13.2 Linear regression

Linear regression determines the linear relationship between an independent variable and one or more dependent variables. A simple example is shown below (Fig 2.3):

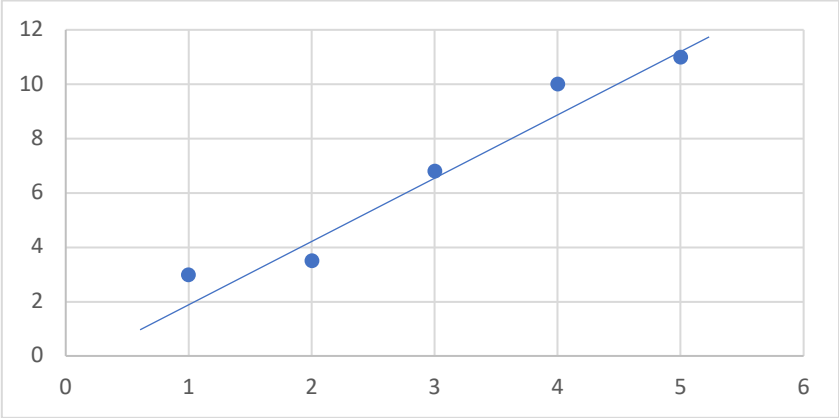


Fig 2.3 Linear Regression. Modelling of linear relationship of two variables.

The aim is to use a linear model to explain the relationship between x and y.

Hypothesis.

$$h_\theta(x) = \theta_0 + \theta_1 x$$

Parameters:

$$\theta_0, \theta_1$$

The value of the parameters is determined by what gives the lowest residual values (errors), where the residual value is $y(x)-h_\theta(x)$. This is the difference between the actual y value and the predicted y value. The cost function models the average residual for different values of θ_0 and θ_1.

Cost Function:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

The goal is to minimise $J(\theta_0, \theta_1)$ which can we achieved using a mathematical method called gradient descent, an iterative optimisation algorithm. Gradient descent takes the partial derivative of the cost function to find the local minima as shown in Fig 2.4.

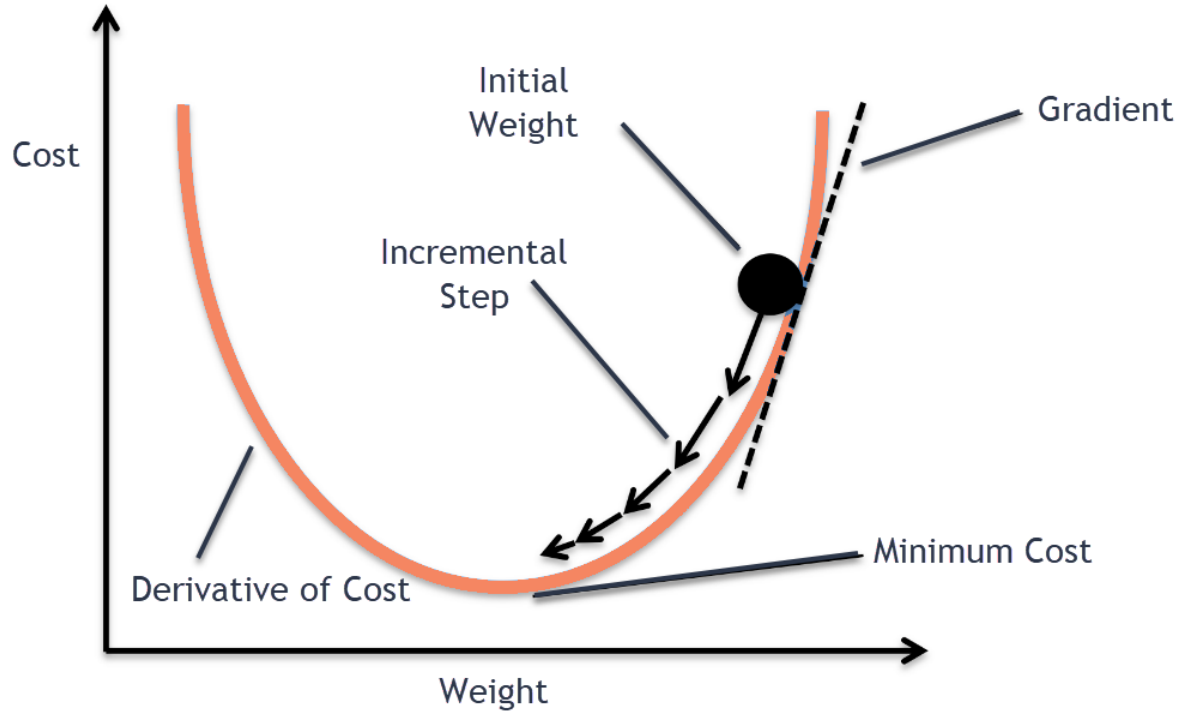


Fig 2.4 Gradient descent. Illustration of local minima (https://www.andrewng.org/).

Gradient descent algorithm

$$\theta j := \theta j - \frac{\alpha\delta}{\delta\theta j} J(\theta 0, \theta 1)$$

learning rate

derivative

(simultaneously update j = 0 and j =1)

α is the learning rate. If α is too small, the computational time is long whilst if too big, can miss the minima. The rate of learning values for θ_0 and θ_1 is determined by α. The change in values θ_0 and θ_1 in any given iteration step is computed by the product of the corresponding partial derivative (gradient) and α. If the gradient is large, a large change in the value of θ occurs whilst if small, a small change occurs allowing fine tuning of the local minima.

2.13.3 Clustering

Clustering is a method to group samples that are similar together. Similarity is measured using a distance metric. The distance metric is symmetrical where the distance from a to b is the same as b to a, it is positive and obeys triangle inequality. Triangle inequality is where the distance between a and c is equal or shorter than the sum of the distances between points a and b and b and c, i.e., the distance from a to c is the shortest route.

2.13.3.1 Distance

Euclidean distance is commonly used and is derived from Pythagorean theorem (Fig 2.5).

$$d_{(x,y)} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Manhattan distance takes the absolute value between values on a shared plain.

$$d_{(x,y)} = \sum_{i=1}^{n}|(x_i - y_i)|$$

Outlier distances will be larger when calculated using Manhattan distance versus Euclidean distance (Fig 2.5).

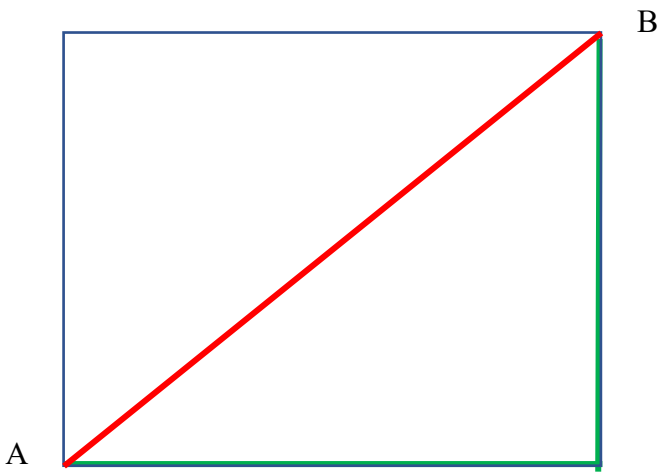


Fig 2.5 Distance metrics used in clustering. Distance between points A and B. Red line represents Euclidean distance, green line represents Manhattan distance.

Squared Euclidean distance, as the name suggests squares the distance. Thus, outliers are given more weight. Standardised Euclidean distance equalises the variance on each axis. Whilst Mahalanobis is an extension of the standardised Euclidean distance, but standardisation of variance is not limited by perpendicular axes. The Pearson Correlation distance is a measure of how well two features change under different conditions.

The distance metric used can greatly influence clustering.
For example,
P1= (1,2,3,4,5)
P2= (5,4,3,2,1)
P3= (100,200,300,400,500)

P1 and P2 will be clustered together if Euclidean distance is used whereas if Pearson correlation distance is used, P1 and P3 will we clustered together.

2.13.3.2 K means clustering

In K means clustering, a pre-determined number of clusters "K" are chosen. K random points are then selected to represent the centre of clusters. Samples are assigned to each cluster based on their distance to the centre of the cluster. New centres are then assigned based on the new cluster and samples are reassigned. This process is repeated multiple times until the clustering is stable. The quality of a cluster can be determined by comparing its intra-cluster distance versus inter-cluster distance. Larger inter-cluster distances are indicative of a good quality clustering.

2.13.3.3 Hierarchical clustering

In hierarchical clustering, a tree is formed representing a hierarchy of similarity between samples. The top-down approach is also known as a "divisive" approach as it works by splitting larger clusters into smaller clusters. Whereas the bottom-up approach known as the "agglomerative" method works by combining smaller clusters into larger clusters. Inter-cluster distances can be calculated using different methods including single linkage where

the distance is calculated between the two closest neighbours, complete linkage where the distance is calculated between the two farthest neighbours, centroid where the distance is calculated between the centre of two clusters and average linkage which is the average distance between all samples between clusters. Clusters are generated by "cutting the tree" which can be cut at different heights.

2.13.4 Singular Value Decomposition

Singular value decomposition (SVD) is used to construct key projections (Principal Components) that summarise the dataset.

2.13.4.1 Vectors

A vector is denoted as <a,b,c,,,,n>, it has a direction and a length. The unit vector of vector V is denoted as $\hat{V}$ and has a magnitude of 1.

$|\hat{V}|$ = 1 and is in direction of V

$\hat{V} = \frac{V}{|V|}$

The dot product of two-unit vectors gives the cos of the angle between vectors.

$\cos(\theta) = \hat{V}1.\hat{V}2$

Where two vectors are perpendicular to each other, the dot product = 0.

2.13.4.2 Projections

Vector V1 can be projected in the direction of V2. The resulting projection value 'p' is a scalar value in the direction of V2.

$V1.\widehat{V2} = V1.\frac{V2}{|V2|} = |V1| * \frac{V1.V2}{|V1|*|V2|} = \frac{V1.V2}{|V2|} = |V1|\cos(\theta) = p$

where θ is the angle between V1 and V2.

When V1 and V2 are perpendicular to one another, p = 0.

2.13.4.3 SVD

A is an n x m matrix. It contains gene-expression data with samples in rows and genes in columns.

	Gene1	Gene2	Gene3
Sample1	3	4	2
Sample2	4	5	3
Sample3	8	9	2

Matrix A can be decomposed to a product of U, S and V matrices as shown below.

$A = U*S*V^T$

U is a n x n orthonormal matrix with real numbers
S is a n x m diagonal matrix with elements of real numbers ordered from largest to smallest.
V is a m x m orthonormal matrix with real numbers
When a matrix is orthonormal the multiplication of it by its transverse gives an identity matrix.

$U^T*U = I$

$U*U^T = I$

$V^T*V = I$

$V*V^T = I$

If matrix A is projected into the eigen space V, where each column of vector V is a distinct eigenvector and the eigenvectors are ordered according to greatest importance. The projection matrix P is,

$P = A*V = U*S*V^T*V = U*S$

Given S is a diagonal matrix, it acts as a scalar and thus P and U are very similar to one another. Thus, the U matrix is a surrogate of the projection matrix P into the eigen space of V.

2.14 Downstream Analysis of Bulk RNAseq

All downstream analysis was performed in R. Counts were filtered using filterByExpr (EdgeR package) with a gene count threshold of 10CPM counts and the minimum number of samples set at the size of smallest disease group. Library counts were normalised using calcNormFactors (EdgeR package) using the method 'weighted trimmed mean of M-values'. The function 'voom' (limma package) was applied to the data to estimate the mean-variance relationship, allowing adjustment for heteroscedasticity.

2.14.1 Differential Expression

Assessment of differential gene expression was performed using the Limma package. A corrected p value cut-off of 0.05 (Benjamini-Hochberg) was used to assess significant genes that were upregulated or downregulated compared with healthy controls, with data grouped in time bins. The UpSetR package was used to visualise differentially expressed genes.

2.14.2 Clustering

To assess whether clinical severity was reflected on a transcriptional level in an unsupervised fashion, K-means clustering was utilised. Heat maps were created using the ComplexHeatmap package, with data scaled and centred prior to visualisation.

2.14.3 Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was used to assess if specific biological pathways were enriched in disease and how this changed with time [211]. A list of ranked genes, determined by Signal-To-Noise ratio was generated. This is the difference of means of the two phenotypes after scaling for standard deviation. The greater the difference, the more distinct the given gene expression is between the two groups.

$$\frac{\mu(a)-\mu(b)}{\delta(a)+\delta(b)}$$

where a and b are distinct phenotypes and μ = mean and δ= standard deviation.

An enrichment score was calculated, determined by how often genes from the geneset of interest appeared at the top or the bottom of the pre-ranked set of genes with the enrichment score representing the maximum deviation from zero. To assess statistical significance, an empirical phenotype- based permutation test was run where a collection of enrichment scores was generated from the random assignment of phenotype to samples and used to generate a null distribution (Fig 2.6). To account for multiple testing, an FDR rate q < 0.20 was deemed significant. HALLMARK gene sets from the Molecular Signatures Database (http://www.broadinstitute.org/gsea/msigdb) were used in analysis.

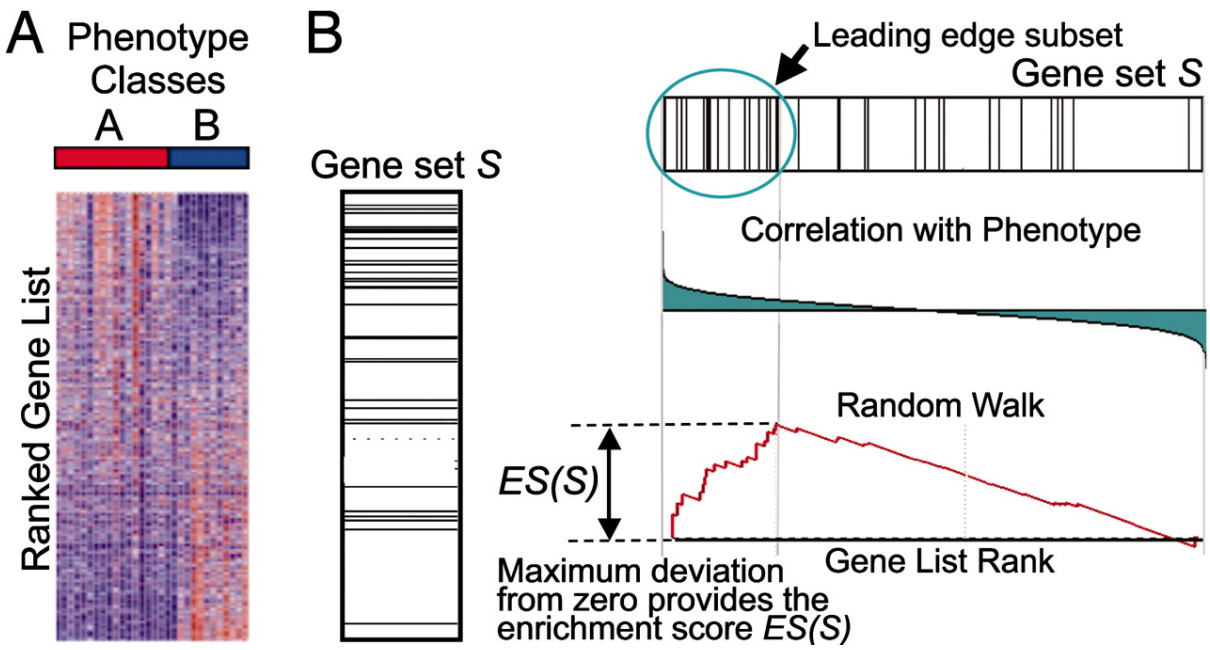


Fig 2.6 Scheme of GSEA. A, Ranked gene list comparing phenotypes A and B. B Gene set S used for assessment. There is enrichment of genes from Gene set S in the top ranked genes[211].

2.14.4 Weighted gene co-expression network analysis

To better understand the relationship between gene expression and clinical traits we used the weighted gene co-expression network analysis (WGCNA) package in R [212–214]. WGCNA overcomes the problem of multiple testing by grouping co-correlated genes into modules and then relating them to clinic traits. Modules are not comprised of a priori defined gene sets but rather are generated from unsupervised clustering.

Modules are summarized using singular value decomposition with the left singular vector (U matrix) used to represent the eigenvalues for a given eigenvector determined by matrix V. The eigengene of the module is then correlated with the sample traits. Significance of correlation between a given clinical trait and a modular eigengene is assessed using linear regression with Bonferroni adjustment to correct for multiple testing (Fig 2.7). Modules are then annotated using Enrichr (https://maayanlab.cloud/Enrichr/).

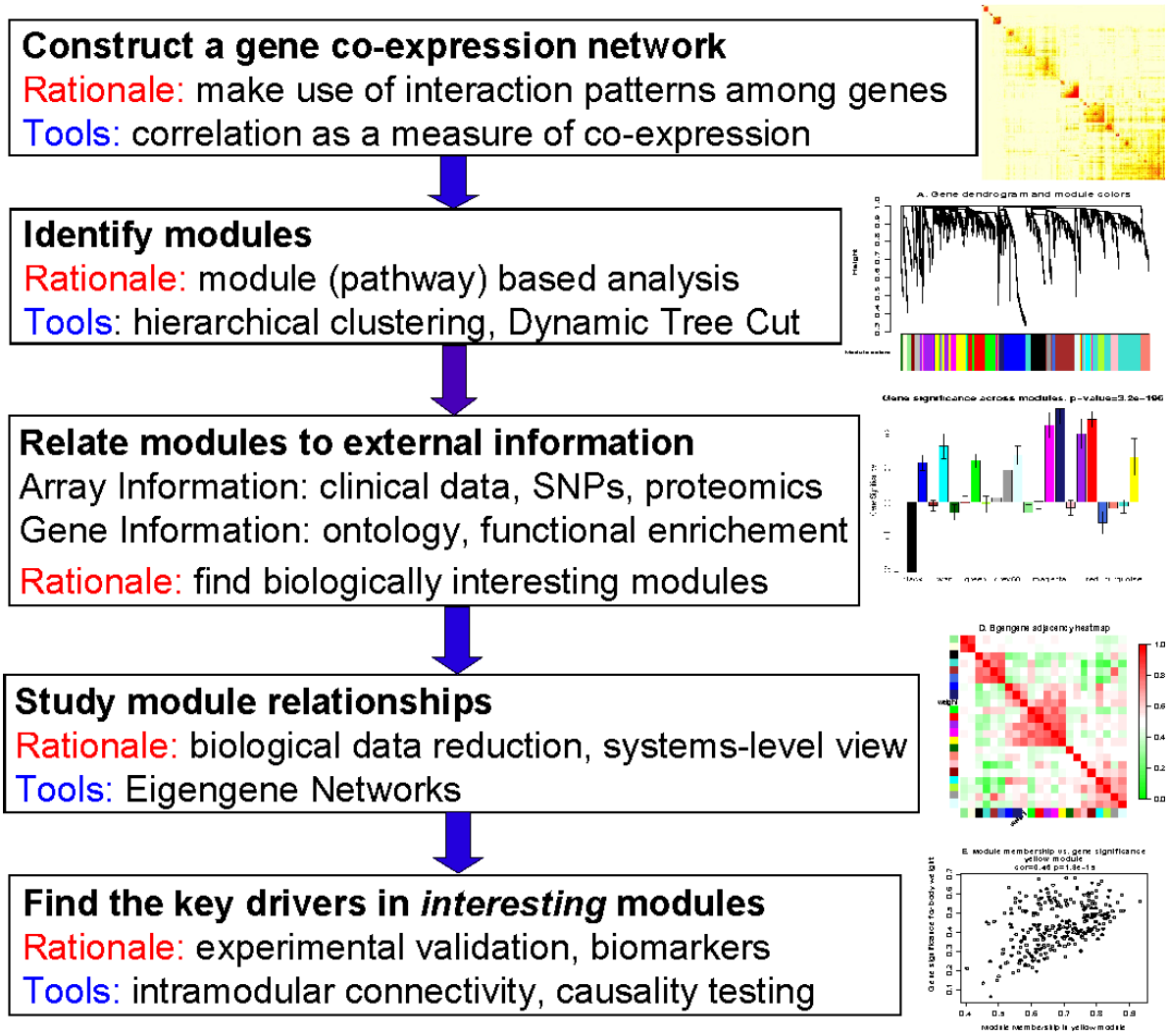


Fig 2.7 Scheme of WGCNA[212,213].

2.1.14.1 Module construction

First the correlation between gene pairs was quantified using a bi-weight mid correlation (See Fig 2.8). Bi-weight mid correlation uses the median instead of the mean when calculating co-variance. Weights are assigned to observations, with a higher value given to observations closer to the median. This method is less influenced by outliers. The gene pair correlations were not dichotomised (hard thresholding) but instead remained continuous (soft thresholding). Thus, avoiding loss in information.

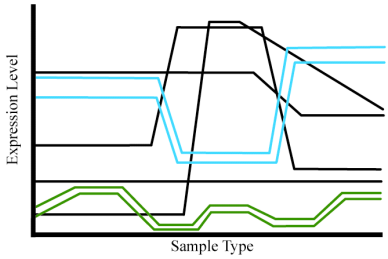


Fig 2.8 Gene expression correlation. Each line represents a single gene. The genes outlined in green are co-correlated and the genes outlined in blue are co-correlated[212,213].

In an unsigned network, only the absolute values are retained and thus a positive and negative correlation are treated the same, with genes ending up in the same module.

$$a_{ij} = | cor(x_i, x_j) |^{\beta}$$

A signed correlation network preserves the nature of the correlation with strongly negatively correlated genes resulting in a correlation matrix value close to 0 (Fig 2.9).

$$a_{ij} = | 0.5 + 0.5 \times cor(x_i, x_j) |^{\beta}$$

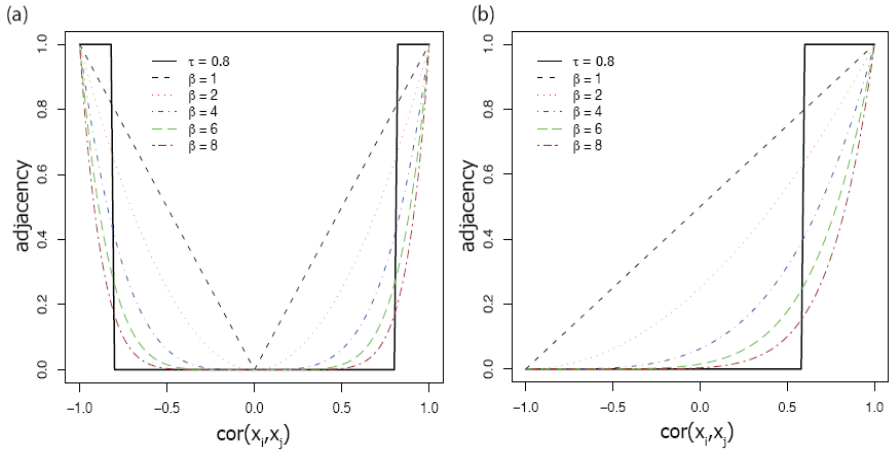


Fig 2.9 Correlation Networks. a) Unsigned correlation network raised to the power β. b) Signed correlation network raised to the power β[212,213].

A signed network was used in the analysis to provide more biologically interpretable modules. To create the adjacency matrix(a_{ij}), the correlation matrix was raised to a power (β). Without raising the correlation matrix to the power β, genes with no correlation are given an adjacency value of 0.5. By raising it to power β, only strong correlations remain and all else is down weighted.

The value of β is chosen to impose approximate scale-free topology. A "scale free topology" represents a network where a large number of nodes have a connectivity close to zero and a small number have a high connectivity. Gene connectivity, is defined as the sum of the elements in a row of the adjacency matrix, excluding the diagonal elements. It represents the gene's connection strength to all other genes. Genes with high connectivity are termed hub genes (Fig 2.10).

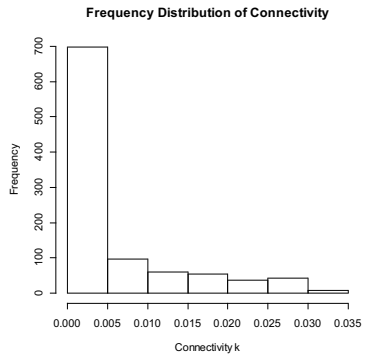


Fig 2.10 Frequency distribution of connectivity. Bar graph of frequency distribution of connectivity. A large number of nodes have low connectivity whilst a small number have high connectivity[212,213].

WGCNA imposes an 'approximate' scale free topology on the data, by raising the correlation matrix to a power, β. β is chosen by trialling various powers and choosing the smallest value that approximates scale free topology, i.e. results in a linear relationship when comparing the log(gene connectivity) versus log(frequency of gene connectivity) (Fig 2.11). This is determined using Pearson's correlation, aiming for an R^2 >0.9.

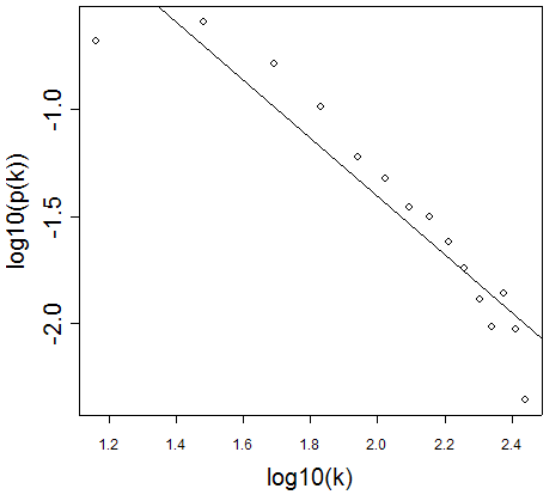


Fig 2.11 Scale free topology. Scatter plot illustrating scale free topology[212,213].

The topological overlap matrix measure (TOM) assesses the interconnectedness of the network. It is a similarity measure. It combines the adjacency measure of two genes (i,j) with the connection strengths that they shared with third party genes. This is subsequently converted to a dissimilarity measure (1- TOM).

$$TOM_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

$$DistTOM_{ij} = 1 - TOM_{ij}$$

k= connectivity

The dissimilarity measure is combined with clustering techniques including partitioning around medoids (PAM) measure and average linkage hierarchical clustering to create a cluster tree, with each branch representing a potential module. The hybrid dynamic tree cut method is then used to determine modules.

2.1.14.2 Module correlations

The module eigengene, broadly represents the first PCA of a given module. Each gene is standardised such that its expression has a mean of 0 and a variance of 1. The singular value decomposition (left singular vector) is used to calculate eigenvalues. Sample module eigenvalues are correlated with clinical traits and FDR corrected using the BH method, with an FDR p value <0.1 deemed significant.

The strength of a gene's module membership (kME) is determined by the following,

kME(i)= cor(module eigengene, gene)

This method is superior to using intermodule connectivity (kIN), as kIN is influenced by module size.

kIN(i)= Σa_{ij} *j* ε *module set*

2.14.5 Linear mixed Effects model

Longitudinal mixed modelling of log transformed absolute cell count changes over time (y_{ij}) was conducted using the nlme package in R, including time (t_{ij}) with a quadratic trend and disease severity category or unsupervised cluster ids (X_j) as fixed effects, and sampled individuals as random effects (u_j):

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \beta_{2j} t_{ij}^2 + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} X_j + u_j, \quad \beta_{1j} = \gamma_{10} + \gamma_{11} X_j, \quad \beta_{2j} = \gamma_{20} + \gamma_{21} X_j, \quad u_j \sim \mathcal{N}(0, \tau^2),$$

I.e., using the *lme* formula:

module_eigenvalue ~ (time + I(time^2)) * category, random = ~ 1|subject.

2.14.6 Multi-omics Factor analysis

Multi-omics Factor analysis (MOFA) is a dimension reduction method, used to assess data across multiple modalities, for the same or overlapping set of patients[215]. It learns 'factors' that represent major sources of variation across modalities and can be used to identify shared axes of variation. Unlike PCA components which are orthogonal, latent factors are oblique and perhaps better able to model biological data (Fig 2.12).

$$\mathbf{Y}^{\mathbf{m}} = \mathbf{Z}\mathbf{W}^{\mathbf{mT}} + \boldsymbol{\varepsilon}^{\mathbf{m}} \quad m = 1, \ldots, M.$$

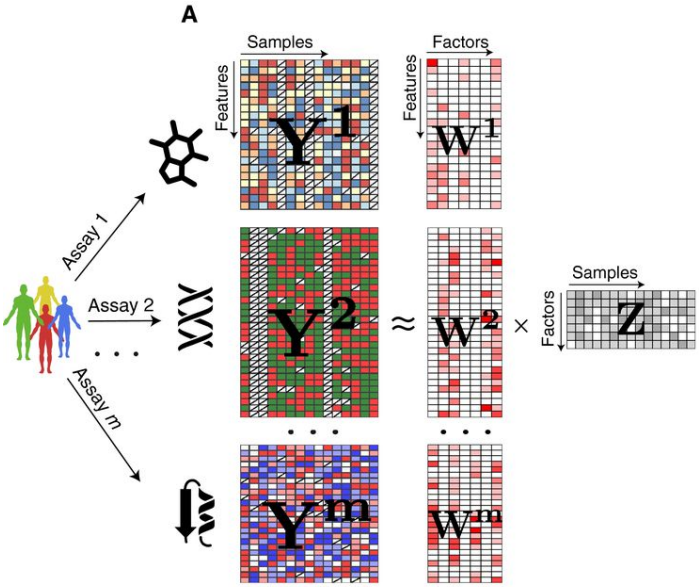


Fig 2.12 Multi-Omics Factor Analysis. Matrix decomposition of Y matrices into shared Z matrix of latent factors and M weight (W) matrices[215]

M data matrices, representative of different omics are decomposed into a shared Z matrix of latent factors and M weight (W) matrices. Data may be continuous, binary or count data. The Z matrix represents latent factors. The W matrices represent the weight of a feature for a given factor in an omic. The higher the weight, the more the sample expresses the latent factor. Features with large positive or negative weights are key components of the factor whilst features with no association have a weight of zero.

2.14.7 Pathway-level information extractor

Pathway-level information extractor (PLIER) (http://gobie.

csb.pitt.edu/PLIER), was used to perform cell subset deconvolution of the whole blood rnaseq dataset. Unlike MOFA, PLIER leverages off the pre-existing knowledge of cell specific pathways to generate pre-annotated pathways.

2.15 B Cell Receptor Repertoire

2.15.1 Background

There are 10^{10}-10^{11} B cells in a person[4]. The aim of B cell receptor (BCR) repertoire sequencing is to gain as much information of the BCR repertoire through exhaustive amplification whilst minimising sequencing error and bias. Two starting materials can be used in BCR repertoire generation, genomic DNA and mRNA. Genomic DNA is stable, and the gene copy is constant between cells. This is unlike mRNA where B cell subpopulations produce vastly different amounts of mRNA per cell, e.g., Plasmablast versus a naïve B cell. In addition, mRNA needs to be converted to cDNA creating a further opportunity for transcription error. mRNA however enables concurrent information of both the variable and constant regions to be gained, as the transcript is intronless and thus not prohibitively long[216].

Sequencing methods include forward/reverse primers where Vh family primers covering framework region 1 are employed as forward primers and J or constant segment primers are used as reverse primers. The use of multiple primers may lead to biases in priming, amplification and/or mask areas of somatic hypermutation. 5' RACE sequencing is an alternative where 5' rapid amplification of cDNA ends enables downstream PCR amplification of the known sequence and only requires one set of gene specific primers at the constant end. This method is limited by poor efficiency compared with direct priming. Lastly, a bait capture method can be used to specifically isolate Ig mRNA. Streptavidin magnetic beads are attached to a sequence of interest and used to bind to Ig sequences. Beads are then washed, and hybridized fragments are sequenced[216–218].

Unique Molecular Identifiers (UMI) are randomly generated sequences which can be used to tag individual sequences. It allows correction of PCR repeats and error. All PCR copies of a

sequence will have the same UMI. A UMI is usually between 8 to 22 nucleotides. If too short, insufficient variability will be present to enable the unique labelling of individual sequences. If too long, there is an increase chance of transcriptional error and primer-dimer formation. All sequences with the same UMI are collapsed to one in downstream processing and a consensus sequence is generated based on the majority. UMI therefore also allow correction of late PCR errors[216,218]. Early PCR errors cannot be corrected as they will affect the majority of sequences. Duplex sequencing is where UMIs are tagged to adaptors on both ends and both strands are sequenced. This allows sequencing errors to be distinguished from true mutations as a true mutation will be present on both strands. A less common technique involves utilising a Tn5 transposase attached to a primer which results in random insertions and subsequent Tn transposase-foreshortened sequences which can be overlapped to get the overall sequence. Errors due to incorrect base calling can be reduced with paired end sequencing which overcomes poor quality sequencing at the tail end[216].
An ideal sequencing platform allows accuracy, adequate read length and depth and is cost effective.

The BCR repertoire was generated from mRNA with UMI tagging and libraries sequenced using Illumina miseq which allows long reads at 250-300bp along with paired end sequencing (Fig 2.13).

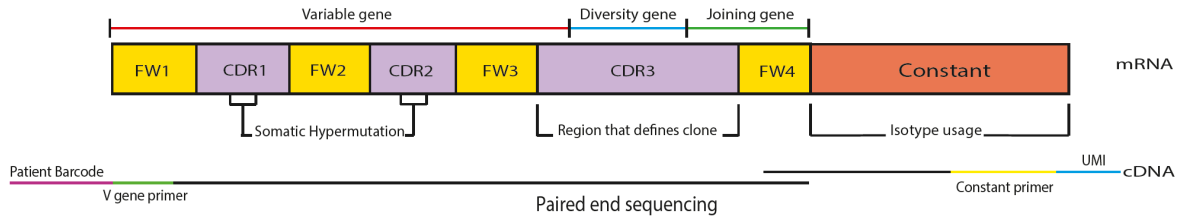


Fig 2.13 Cartoon of BCR sequencing method. UMI denoted in blue, patient barcode in red, forward variable gene primer in green and reverse constant primer in yellow. Region used to assess somatic hypermutation, clone identity and isotype marked.

2.15.2 BCR Library Preparation

PBMC were lysed and RNA extracted using Qiagen AllPrep® DNA/RNA mini and micro kits according to the manufactures protocol. RNA was extracted from PAXgenes using Qiagen PAXgene Blood RNA kit. The RNA was quantified using a Qubit.

B cell receptor repertoire libraries were generated using the protocol described by Bashford-Rogers et al[217].200ng of total RNA from PAXgenes/PBMCs (14ul volume) was combined with 1uL 10mM dNTP and 10uM reverse primer mix[219] (2uL) and incubated for 5 min at 70°C. The mixture was immediately placed on ice for 1 minute and then subsequently combined with 1uL DTT (0.1 M), 1uL SuperScriptIV (Thermo Fisher Scientific), 4ul SSIV Buffer (Thermo Fisher Scientific) and 1uL RNAse inhibitor. The solution was incubated at 50 °C for 60 min followed by 15 min inactivation at 70 °C. cDNA was cleaned with AMPure XP beads and PCR-amplified with a 5' V-gene multiplex primer mix[219] and 3' universal reverse primer using the KAPA protocol and the following thermal cycling conditions: 1cycle (95°C, 5min); 5cycles (98°C, 20s; 72°C, 30s); 5cycles (98°C, 15s; 65°C, 30s; 72°C, 30s); 19cycles (98 °C, 15s; 60°C, 30s; 72°C, 30s); 1 step (72°C, 5 min). Sequencing libraries were prepared using Illumina protocols and sequenced using 250 or 300-bp paired-end sequencing on a MiSeq.

2.15.3 Sequence Processing Theory

Fastq files are generated to assess the quality of base calling in the sequence and is represented in the form of a Phred score. The Phred score is an estimate of the probability of miscalling a nucleotide at each position and are encoded as ASCII characters. A Phred score of 20 means 1 error per 100 base pairs (p=10-Q/10). A high confidence in base pair calling in required in BCR analysis in order to accurately assess somatic hypermutation. As the length increases from the 5' toward the 3' prime end, the confidence in base calling declines. Thus pre-processing includes filtering out low quality reads and trimming sequences of low quality bases. Paired end reads are overlapped and where reads do not overlap, they are removed. This biases sequences with a shorter CDR3 length.

High quality merged reads are then grouped based on UMI. This corrects for PCR repeats and allows a consensus sequence to be created utilising all reads with the same UMI. An additional requirement may be a minimum number of reads required to construct a consensus sequence. In addition, a minimum number of unique UMIs may be required for a sequence to be retained. The constant region is initially annotated based on reverse primers used and then further isotype subgroups identified using kmer matching downstream of the primer.

The V primer is masked (base pairs changed to N) post identification and the VDJ region annotated using public available reference germline data. The IMGT database is the most comprehensive[220]. The difficulty with alignment arises from differentiating between allelic variation and somatic hypermutation in highly homologous V gene segments. The D gene due to its size and location makes identification very difficult. In addition, although comprehensive, the IMGT database is not complete. Efforts have been made to derive a germline sequence for a patient and use this as a reference to determine somatic hypermutation. Ideally, this would be from naïve b cells. Genomic DNA from a non-B cell lineage cell would contain the germline sequence but the length required for sequencing is too long and due to the presence of repetitive sequencing, short sequence alignment is not possible. Bio-informatics tools such TIgGER address this[221]. Determining haplotypes allows the restriction of V-J pairings, further increasing the accuracy of annotation.

2.15.4 Sequence Processing Pipeline

Raw reads were filtered for base quality using a median Phred score of >32 (http://sourceforge.net/projects/quasr/). Forward and reverse reads were merged where a minimum 8bp identical overlapping region was present. Sequences were retained where over 80% base sequence similarity was present between all sequences with the same UMI. The constant-region allele with highest sequence similarity was identified by 10-mer matching to the reference constant-region genes from the IMGT database. Sequences without complete reading frames and non-immunoglobulin sequences were removed and only reads with significant similarity to reference IGHV and J genes from the IMGT database

using BLAST were retained. Immunoglobulin gene use and sequence annotation were performed in IMGT V-QUEST[220].

2.15.5 BCR metrics

A summary of the key BCR metrics used in analysis is shown in Fig 2.14.

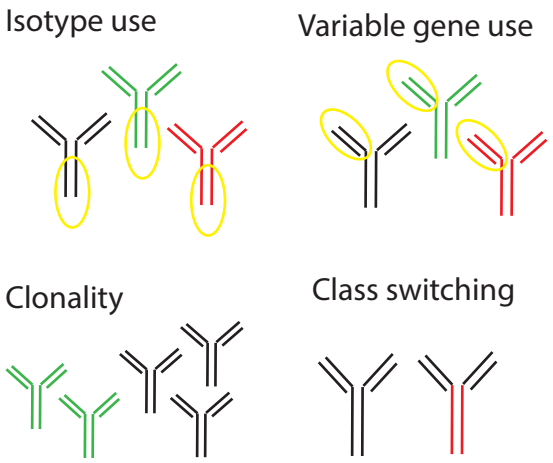


Fig 2.14 BCR metrics used in analysis. Illustration of isotype, variable region, clonality and class-switching.

2.15.5.1 Isotype and variable gene usage

The variable and junction genes and isotype can be recorded for each read. To determine for a given patient if there is skewing to a certain gene or isotype, the proportion of the repertoire taken up by each gene and isotype is determined per sample and can be compared between samples. Each unique VDJ region is counted only once to ensure results are not skewed by the differential mRNA content of B cell subsets (in particular plasmablasts which have increased immunoglobulin mRNA content).

2.15.5.2 Clonal grouping

A clone is defined as a group of cells descended from a common ancestor. Determining clones in BCR repertoire analysis is more complicated as somatic hypermutation introduces variability. Clones can be grouped based on identical V and J segments and junctional length and similarity of nucleotide sequence in the junction region. Similarity is determined by “hamming distance”. Hamming distance is a count of the number of positions that are

different between two strings of identical length. Clonal grouping is different to convergent evolution (discussed below) which is where independent clonal variants produce similar amino acid sequences. Clonality is based on nucleotide similarity. Using the Hamming distance measure, clustering of sequences is performed.

2.15.5.3 Somatic hypermutation

Mutational differences from germline sequence arise from VDJ recombination and somatic hypermutation. During SHM, RGYW/WRCY motifs (R = A/G, Y = C/T, W = A/T) are mutational hot-spots targeted by AID whilst SYC (S = C/G, Y = C/T) is a cold-spot motif. CDR regions have a larger number of mutational hotspots than FWR regions[4]. Given FWR regions have a structural role they are less likely to benefit from SHM. SHM uncommonly causes insertion and deletions as these are most likely to cause structural instability and thus such B cells are negatively selected. Somatic hypermutation levels are calculated based on divergence from the baseline and are corrected for length of the sequence assessed. Somatic hypermutations can be silent or replacement mutations. A silent mutation results in no amino-acid change whilst a replacement mutation results in the encoding of a different amino-acid. An increase in replacement/silent ratio may represent antigen driven selection pressure driving an increase in mutations that result in new amino-acids.

2.15.5.4 Stereotypic sequences, public clones and convergent evolution

The CDR3 region is the junction of the V, D and J genes and is thus the most variable. CDR3 length, hydrophobicity and aliphatic indexes are used in comparisons.
Public clones may also be identified, these are defined as shared clones present between individuals. A common definition of sharing is a clone with the same V and J gene, CDR-H3 length with a minimum 85% CDR-H3 amino acid sequence identity. These B cells are expected to respond to the same antigen and are termed “Stereotypic BCRs”. Clustering algorithms including CD-HIT can be used to cluster such sequences together. A “greedy algorithm” is where instead of the most optimal condition being chosen; the order the sequences are presented in determines how grouping occurs. To overcome this, and find the optimal clustering, multiple iterations are performed.

2.15.5.5 Subsampling

Subsampling, also known as rarefaction, is important when comparing certain metrics in BCR repertoire analysis which are influenced by library depth, in particular diversity. Subsampling is where a portion of reads are sampled with the lowest sample library depth is used. Subsampling has the limitations of not using all the available data and potentially limiting findings. Running multiple iterations is used to ensure generalisability. An alternative is to calculate metrics for a given library size and then extrapolate the asymptotic values[222,223].

2.15.5.6 Diversity

Richness

Richness refers to the number of unique clones in a repertoire. It is the most direct measure of diversity but does not consider the abundance of a given clone and hence does not reflect evenness/homogeneity. Library depth greatly influences this number, but subsampling reduces the ability to identify rare clones. Chao1 attempts to account for the unsampled clones[224].

$$\text{Chao1} = S_{obs} + n_1^2/2n^2$$

where S_{obs} is the observed number of species, n1 is the number of singletons (species with count = 1), and n2 is the number of doubletons (species with count = 2).

Shannon's index

Shannon's index is a measure of both "evenness" (the distribution of reads amongst clones) and "richness" whereby a dual increase in richness and evenness increases the index. It assumes that all species are present in the sample, and it assesses the proportion of total reads represented by each clone.

$$H = -\sum_{i=1}^{s} p_i \ln p_i,$$

Where H is entropy, pi = ni/N; ni is the number of individuals of the ith species; N is the total number of individuals and s is the total number of species[225].

Equitability = H/ Hmax where Hmax (maximal entropy) = log(S)

Equitability is not influenced by the number of unique of clones but rather the distribution of size of clones in the repertoire.

Simpson's index

The Simpson's index assesses the probability of two randomly sampled reads belonging to the same clone, the more expanded clones within the population, the greater the chance of clonal sharing. It increases as the dominance of clones increase and is not impacted by rare populations[226].

$$1 - D = 1 - \frac{\sum_{i=1}^{s} n_i(n_i-1)}{N(N-1)}$$

where ni is the number of individuals of the ith species and N is the total number of individuals, and s is the total number of species.

D50 index

The D50 index refers to the number of unique CDR3 sequences that are present in the top 50% of sequences. A small D50 index is suggestive of large dominant clones. Similar to the Simpson's index, D50 is not affected by rare populations.

The Hill indices

Hill indices models diversity as a function of a continuous parameter, q. q=0 corresponds to richness, q=1 is the exponential of Shannon index, q=2 is the reciprocal of Simpson's index

and as q approaches infinity, the y value approaches the reciprocal of the largest clone frequency[227].

Class switching

Class switching between isotype classes was quantified by assessing the frequency of unique VDJ regions that were shared amongst two different isotypes, having corrected for read depth by subsampling.

2.15.6 B cell Repertoire Analysis

BCR clones were assigned using the Change-O package using the single-nucleotide Hamming distance model[228]. The Alakazam package was used to analyse the BCR sequencing data for diversity estimation of CDR3 sequences; the diversity estimates were adjusted for sequencing depth by subsampling with multiple iterations [228]. Somatic hypermutation levels (including silent and non-silent mutations) per unique IGHV-D-J region per isotype were calculated over the CDR1/2 and FWR regions for each individual sample using the observedMutation function within the SHazaM package[228]. Lineage trees were generated using the buildPhylipLineage function within the Alakazam package after merging sequences from paired time points (Gupta et al., 2015). VDJtools was used to analyse the BCR sequencing data for diversity estimation of CDR3 sequences (Chao1); the diversity estimates were adjusted for sequencing depth via subsampling with 2,000 random iterations[229].

Convergent IGH clones were identified based on matching V and J gene and CDR-H3 length with a minimum 85% CDR-H3 amino acid sequence identity. CDR-H3 amino acid sequence clustering was performed using CD-HIT[230] with options -c 0.85 -l 4 -S 0 -g 1 -b 1. Clusters were identified as COVID-19 specific if they co-clustered with sequences from the CoV-AbDab database [231].

This summarises the techniques used in the analysis of bulk RNAseq and BCR repertoire in the subsequent chapters.

3. Whole blood transcriptomics and deep immunophenotyping in COVID-19

3.1 Introduction

COVID-19 caused by SARS-CoV-2 is a complex condition with a broad spectrum of disease severity (Wang et al., 2020; Zhou et al., 2020). Most patients are able to mount an adequate response and achieve viral control but in a minority of patients end-organ damage, and often death results[173].

Severe COVID-19 is associated with major perturbations in circulating immune cells with profound leukopenia affecting both the innate and adaptive immune cells[161,162,172,175,232–234]. The cause of these profound changes is unknown, with little evidence of virus directly infecting immune cells in the periphery. The combination of IFN-γ and TNF-α robustly induces cell death via STAT1/IRF1 axis, key cytokines involved in SARS-CoV-2 [235] and the presence of splenic atrophy may contribute to lymphopenia[190]. However, it is unclear whether leukopenia plays an active role in disease pathogenesis or whether it is simply a biomarker of severity.

SARS-CoV-2 is typified by the presence of elevated inflammatory cytokines juxtaposed with a comparatively lower IFN response, in contrast with other respiratory viruses(Blanco-Melo et al., 2020). Elevated cytokines include IL-1, IL-6, IL-8, TNFa and CXCL10[173,175] and may contribute to the predominant extra-follicular B cell response [189,190,237]. Low interferon is achieved through inhibition of the host innate interferon response via NSP1, NSP3, NSP12, NSP13, NSP14, ORF3, ORF6 and M proteins [238,239]and additionally, SARS-CoV-2 infected pDCs appear functionally impaired with reduced expression of phosphorylated ribosomal protein S6, a canonical target of mTOR activation, required in IFN production[161].

Severe COVID-19 is associated with higher viral titres [240], despite this, patients appear to have an initial lower IFN response and subsequent shorter duration of expression, compared in those with moderate disease[162]. Further illustrating the importance of host IFN response, a candidate gene approach revealed mutations in genes involved in the regulation

of type I and III IFN response in patients with severe COVID-19, with recapitulation of the variants resulting in decreased type I IFN gene and protein levels[171]. Additionally, in 10% of patients with severe COVID-19 pneumonia, neutralizing autoantibodies against interferon were detected whilst none were detected in mild disease[171].

Successful control of SARS-CoV-2 is achieved through the co-ordinated efforts of the immune response, with combined SARS-CoV-2-specific CD4+ and CD8+T cell responses resulting in milder disease [242]. A sustained anti-SARS-CoV-2-specific IgG response results in shorter duration of illness[243], however recovery is still achieved in the absence of a humoral response [244,245]. Pre-existing adaptive immunity may protect against severe disease with T cell reactivity present against SARS-CoV-2 in unexposed individuals [246] and evidence of antibody cross-reactivity between seasonal coronavirus and SARS-CoV-2[201].

The relationship between the initial immune response to SARS-CoV-2, viral clearance, and development of the ongoing inflammatory disease that drives severe COVID-19 is not clearly established, nor have the kinetics of the immune changes seen in COVID-19 been fully assessed as disease progresses. By analysing longitudinal samples from COVID-19 patients with a range of disease severities, for up to 3 months from symptom onset, we were able to address two important questions regarding the immune response to SARS-CoV-2: (i) How does the very early immune response in patients who cleared virus and recovered from disease with few or no symptoms, compare with those who progressed to severe inflammatory disease. This provided insight into what constitutes an effective versus an ineffective immune response, and whether systemic inflammation is an early or later development in those who progress to severe disease. (ii) How rapidly do the profound immune defects that accompany severe COVID-19 recover, and do the kinetics of recovery relate to ongoing inflammation and clinical status.

3.2 Results

3.2.1 Patient Cohort

SARS-CoV-2 positive participants were recruited between 31st March and 20th July 2020 from the routine screening of healthcare workers (HCW) at Addenbrooke's Hospital (Rivett et al., 2020) and from patients who presented to Addenbrooke's or Papworth hospitals. SARS-CoV-2 was confirmed by quantitative reverse transcription PCR (RT-qPCR). After recruitment patients were bled approximately weekly, and then at outpatient follow-up visits 4-12 weeks after study enrolment. HCWs were sampled at study entry, and then approximately 2 and 4 weeks later. Disease severity was graded into five categories, according to symptoms and oxygen requirements.

These were:

- A) asymptomatic HCWs.
- B) HCWs who either were still working with mild symptoms insufficient to meet the criteria for self-isolation (Rivett et al., 2020), or who were symptomatic and self-isolating.
- C) patients who presented to hospital but never required oxygen supplementation.
- D) patients who were admitted to hospital and whose maximal respiratory support was supplemental oxygen.
- E) patients who at some point required assisted ventilation. Three patients who died without admission to intensive care were also included in this severe group.

45 healthy controls were also recruited across a range of age and sex. Time is measured since the first positive swab for cohort A, and since the onset of symptoms for other cohorts. In total 605 blood samples were collected from 246 participants out to 90 days from the onset of symptoms (Fig 3.1A). Sex and age analysis revealed an increase in age in the more severe groups along with a bias towards the male sex (Fig 3.1B and C), as previously shown[247].

A high-sensitivity C reactive protein (CRP) assay, a marker of inflammation was performed. This demonstrated an increase in levels with increased disease severity as defined by maximal respiratory support (Fig 3.1D). A lower PCR cycle threshold indicative of higher viral titres was present in group E compared with other severity groups. Most patients cleared virus within 24 days from symptom onset regardless of severity group (Fig 3.1D). Of the 6 patients who remained positive out at 30 days, four were overtly immunosuppressed (3 solid organ transplants with recent induction/rejection treatment, 1 myeloma on B-cell depletion therapy) and one was a peritoneal dialysis patient admitted with peritonitis.

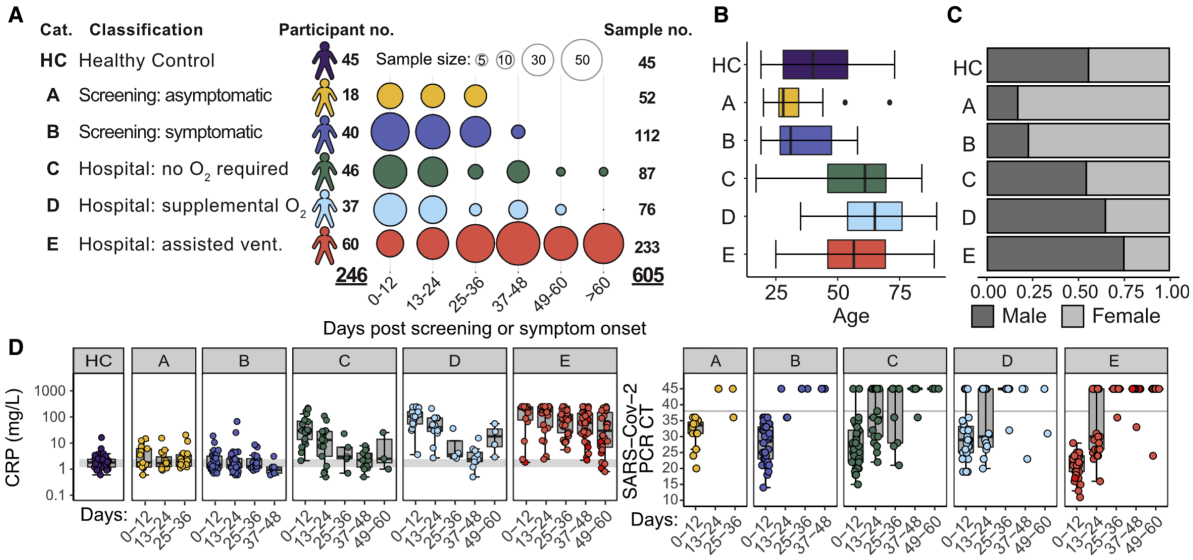


Fig 3.1 Cohort characteristics. A) Study participant and sample numbers split by severity categories and 12-day time bins post screening (group A) or symptom onset (group B-E). Distribution of participant age B) and gender C) across severity categories. D) Boxplots showing measured CRP (mg/L), complement proteins, cytokines and SARS-CoV-2 PCR cycle threshold (CT) for samples collected within 12-day time bins. Grey band indicates the interquartile range of the corresponding measure in HCs, or the SARS-CoV2 negative swab cycle threshold (CT > 38). Points are coloured based on asymptomatic or symptomatic classification for categories A and B respectively, normal or abnormal chest radiology (group C), and mode of respiratory support at sampling (group D and E); time points missing respiratory status are coloured grey. Fig generated by A.H.

3.2.2 Cytokines and complement components

Cytokine and complement components were measured from plasma at each time point. The heatmap in Fig 3.2 compares levels in severity groups at various time points with health. Asymptomatic HCWs in group A had no evidence of cytokine or complement dysregulation. Patients in group B similarly showed no increase in CRP or cytokine levels but rather an initial but only transient increase in C3c and the terminal complement complex (TCC).

Patients with more severe symptoms resulting in presentation to hospital showed elevations in CRP, cytokines and complement components with groups C, D and E having significantly raised CRP, IL-6, IL-1B, IL-10, TNFa, C3a and TCC. These abnormalities were maximal at the first bleed, and largely persisted in groups D and E. Interferon-gamma (IFN-γ) was raised in only groups D and E and this resolved within 12 days from symptom onset (Fig 3.2).

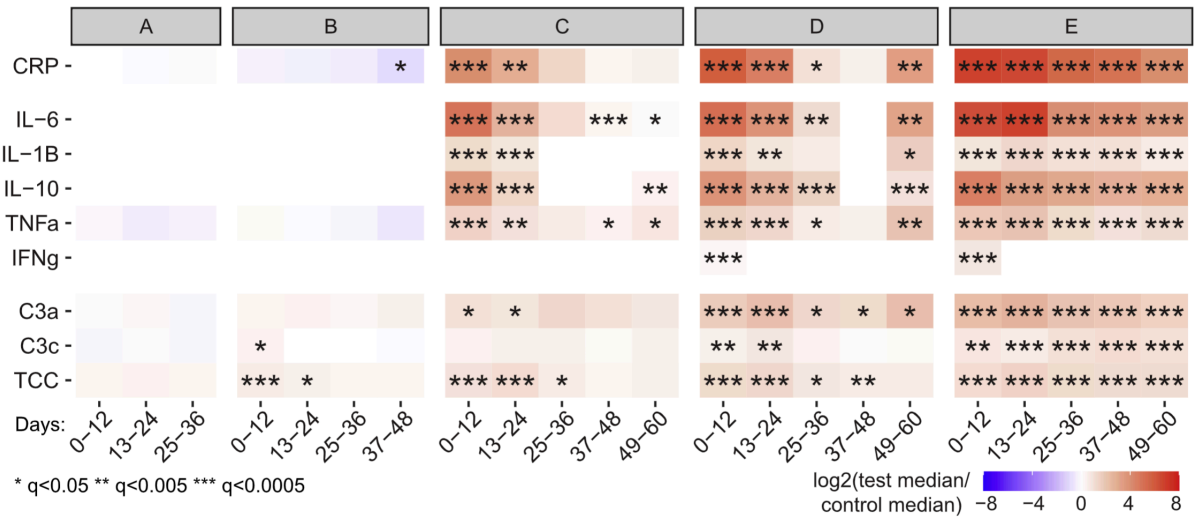


*Fig 3.2 Markers of disease activity. Heatmap showing log2 fold change in median measure between COVID-19 cases and HC, within severity categories and across 12-day time bins. Wilcoxon test FDR adjusted p-value: *<0.05, **<0.005, ***<0.0005 Fig generated by A.H.*

3.2.3 Immune cellular abnormalities

To understand immune cellular changes according to disease severity and over time, we used standardised flow cytometry panels (Fig 3.3). Trucount analysis enabled calculation of absolute cell numbers. Cellular changes were assessed across time "bins" of 12 days (using the earliest measure per patient per bin in instances of repeat sampling). Fig3.3A illustrates absolute cell numbers of plasmablasts and pDCs according to disease severity and time. Plasmablasts were markedly increased whilst pDCs were markedly suppressed proportionate to disease severity at the early time points. The outcomes for 30 cell types are summarised in a heat map, showing changes in cell population size relative to the median for healthy controls (Fig 3.3B) in terms of absolute counts as well as proportions.

CyTOF, which uses whole blood rather than peripheral blood mononuclear cells (PBMCs), was also used in a subset of patients to enable quantification of granulocytes (largely absent in PBMCs) and non-classical and intermediate monocytes. In keeping with CRP, cytokine and complement component findings, groups A and B patients had minimal cellular abnormalities when compared with health. The only aberration was an initial and transient increase in plasmablast numbers in group A. In group B, plasmablasts were also increased along with CD8+ CD38+ HLADR+ cells whilst memory B cells, pDCs, basophils and non-classical monocytes were decreased. Widespread abnormalities were seen in groups C, D and E and were more marked when absolute counts were examined compared with cellular proportions (Fig 3.3B). Almost all CD4 T cells subsets were reduced, as were many CD8 T cell subsets and both naive and memory B cells. A number of innate lymphoid subsets were also reduced, including MAIT cells, various γδ T cell subsets, and NK cells. The myeloid compartment was also affected, with a reduction in myeloid dendritic cells, and both non-classical and intermediate monocytes.

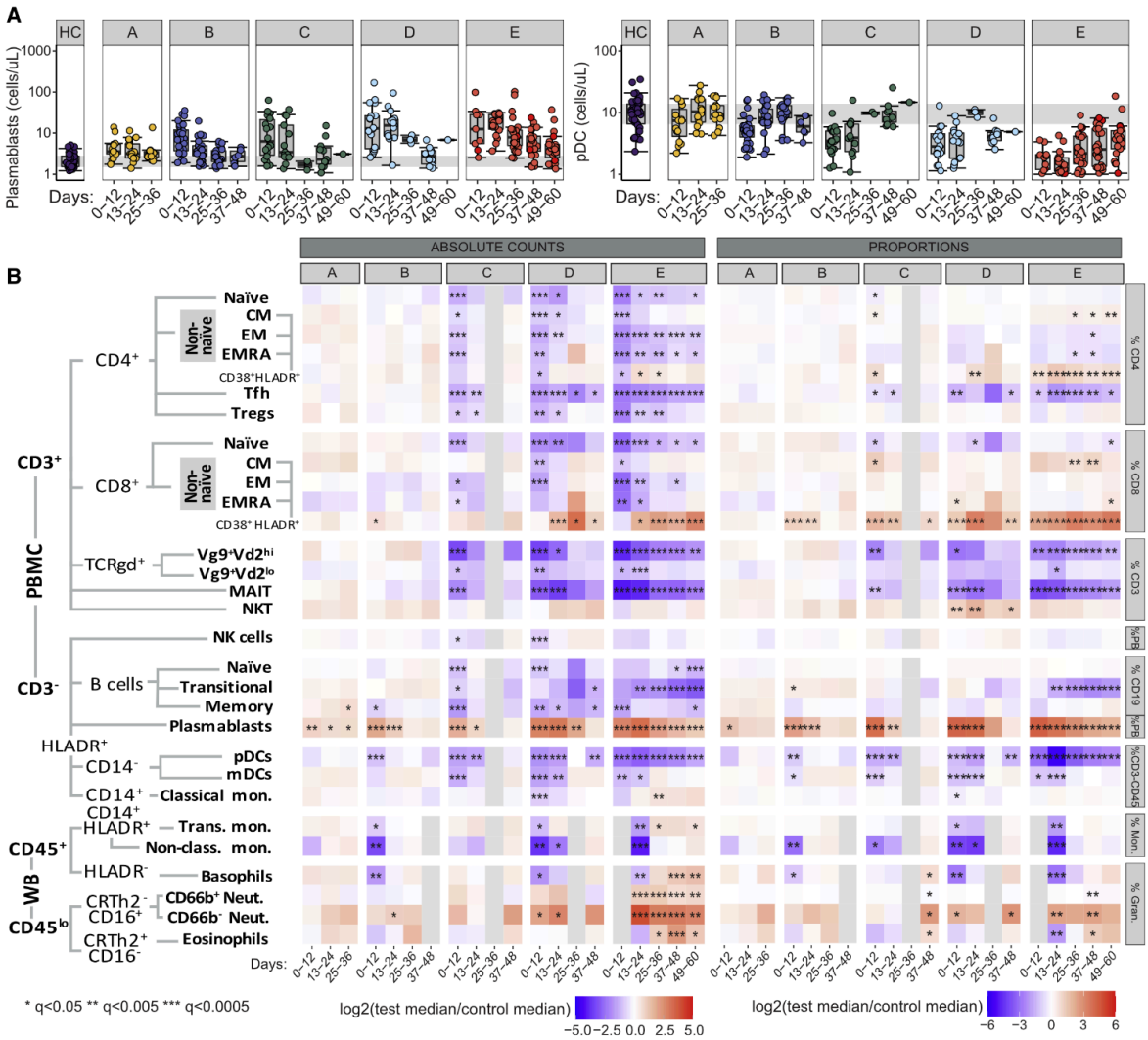


*Fig 3.3 Cellular changes over time. A) Boxplots showing absolute counts (cells/uL) for four representative cell populations, split by severity categories and 12-day time bins post screening (group A) or symptom onset (group B-E). Grey band indicates the interquartile range of the corresponding population in HC. Points are coloured based on asymptomatic or symptomatic classification for categories A and B respectively, normal or abnormal chest radiology (group C), and type of respiratory support at time of sampling (group D and E). B) Heatmap showing the log2 fold change in median absolute cell count between COVID-19 cases and HCs, within severity categories and across 12-day time bins. Wilcoxon test FDR adjusted p-value: *<0.05, **<0.005, ***<0.0005. Population hierarchy and associated cell surface markers are shown to the left. PBMC, peripheral blood mononuclear cells, analysed by flow cytometry; WB, whole blood, analysed by CyTOF. Fig generated by A.H.*

3.2.4 Blood transcriptomic inflammation-related signatures.

To examine changes in transcriptional signatures with disease severity and with resolution of inflammation, RNA was isolated, and whole blood transcriptomes generated by RNA-sequencing at select bleeds. The distribution of samples according to time intervals is shown in Fig 3.4. Due to minimal representation of samples after 48 days, for severity groups A-D, the analysis focused largely on the first 48 days from symptom onset.

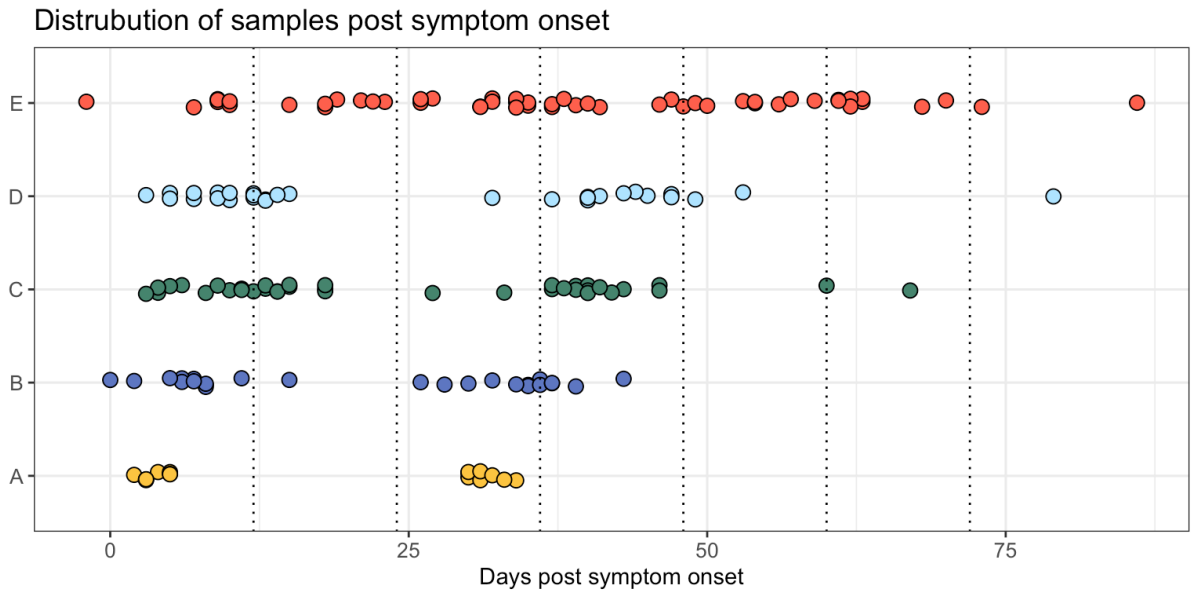


Fig 3.4 Distribution of samples according to disease severity and symptom onset.

3.2.41 Cell subset deconvolution

We first analysed the transcriptome data using Pathway-Level Information ExtractoR (PLIER) which performs matrix factorization to identify interpretable latent factors. Factors were learned using cell type specific pathways (Fig 3.5).

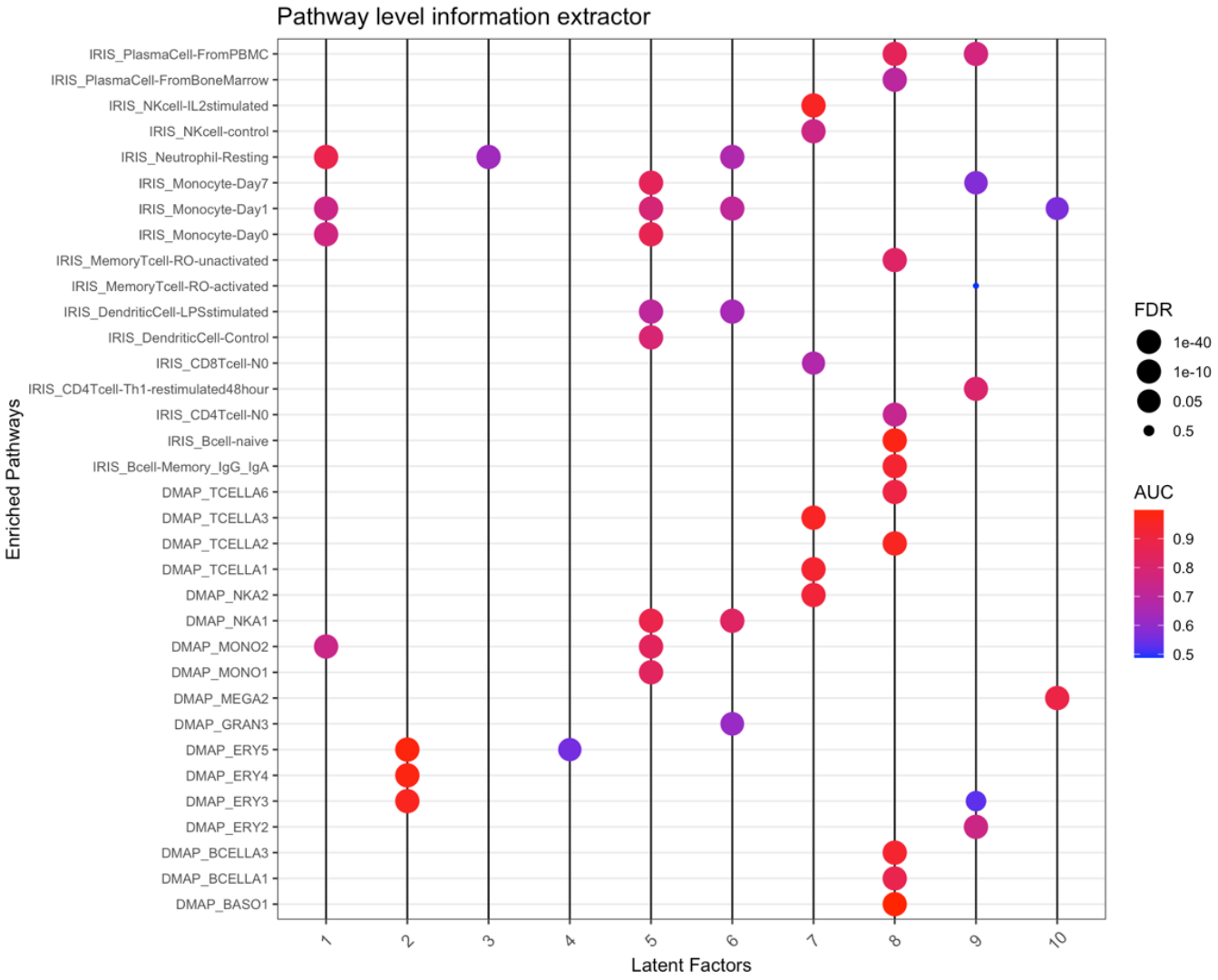


Fig 3.5 PLIER annotation of latent factors according to cell type specific pathways.

The contribution to each latent factor by immune cell subsets was then calculated across the severity groups and time points (Fig 3.6). These RNAseq-derived latent factors were broadly aligned with the pattern observed in the cell count data (Fig 3.3) with an elevation in the plasma cell signature (mirroring plasmablasts cellular findings) and a suppression of B cell memory and NK/T cells signatures. An exception to this was the pronounced neutrophil signature seen at day 0 to 24 across groups C to E, and persisting at day 25-48 in group E.

This transcriptomic analysis shows more pronounced neutrophil dysregulation across severity categories than is suggested by increasing neutrophil number alone. An erythrocyte gene expression-driven latent factor was also seen, and was prominent in group E at late times. This may be associated with heme metabolism, and is discussed below.

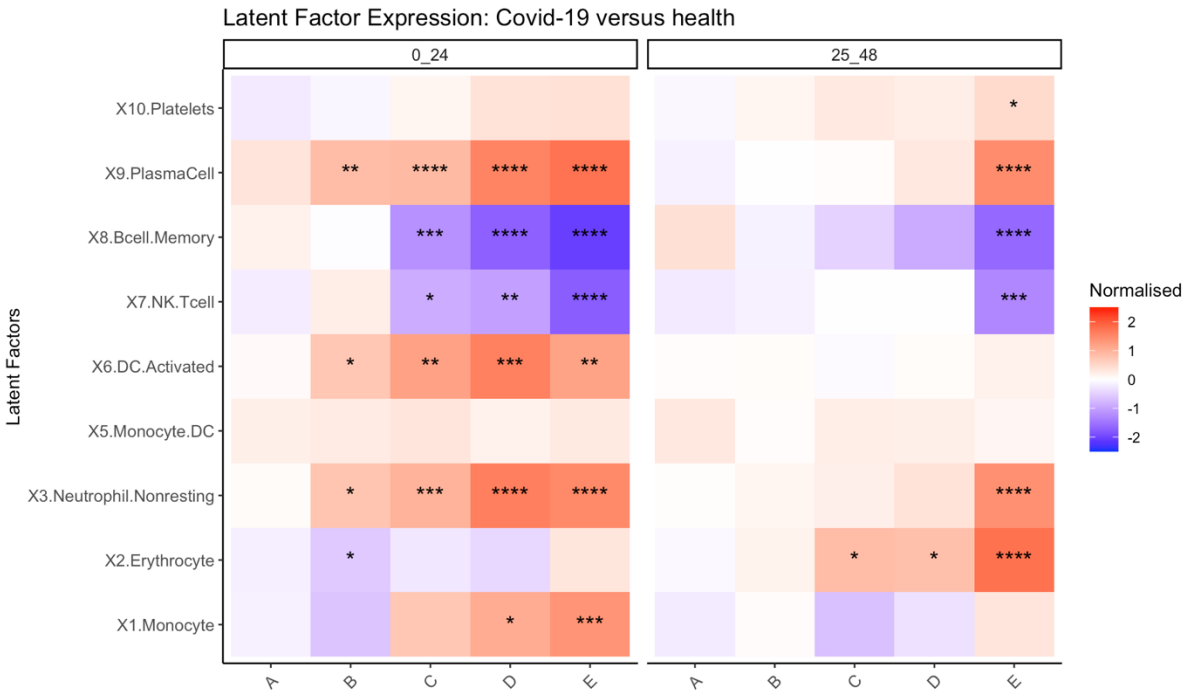
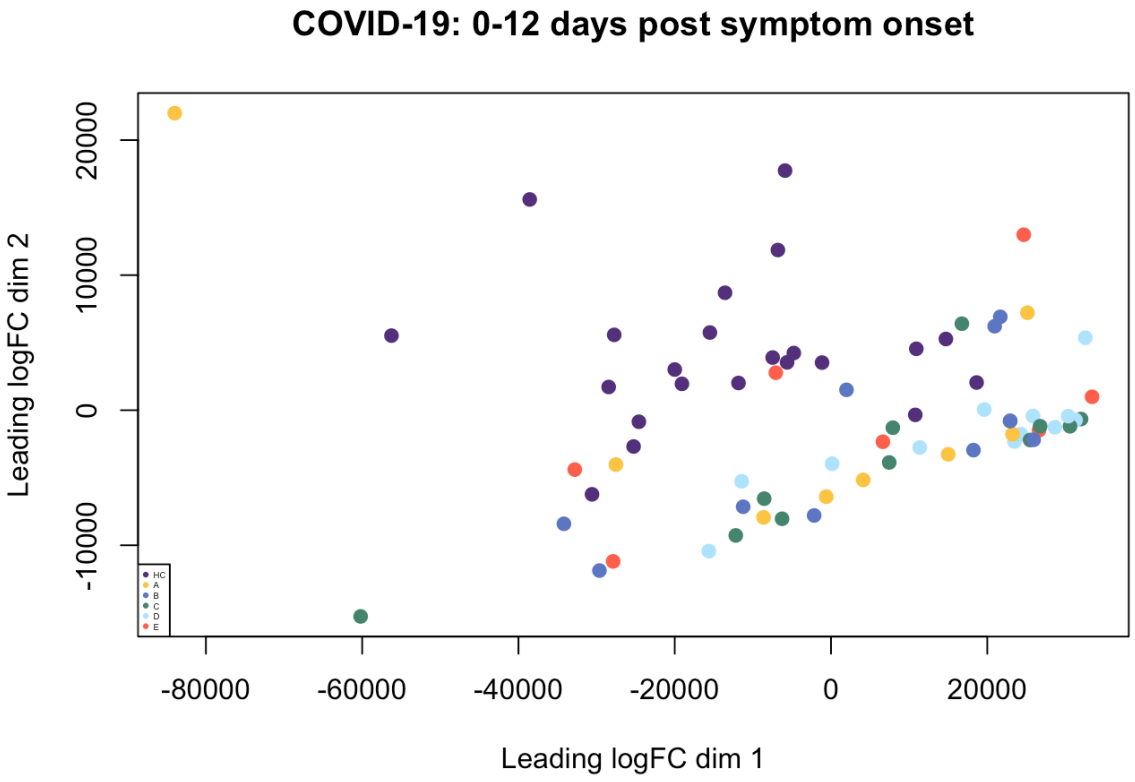


*Fig 3.6 PLIER latent factor enrichment. Cell subset deconvolution performed using PLIER, leveraging off prior knowledge of cell specific pathways. COVID-19 cases split by severity categories and 24-day time bins. Latent factor expression compared with HC, FDR adjusted p-value: *<0.0005.*

3.2.42 Principal component analysis and differential gene expression

Principal component Analysis (PCA) was performed to assess if disease severity explained a large proportion of variance. PCA revealed separation of healthy controls from SARS-CoV-2 patients at both 0-12 and 13-24 days from symptom onset (Fig 3.7) with PCA2 being the axis of greatest separation.



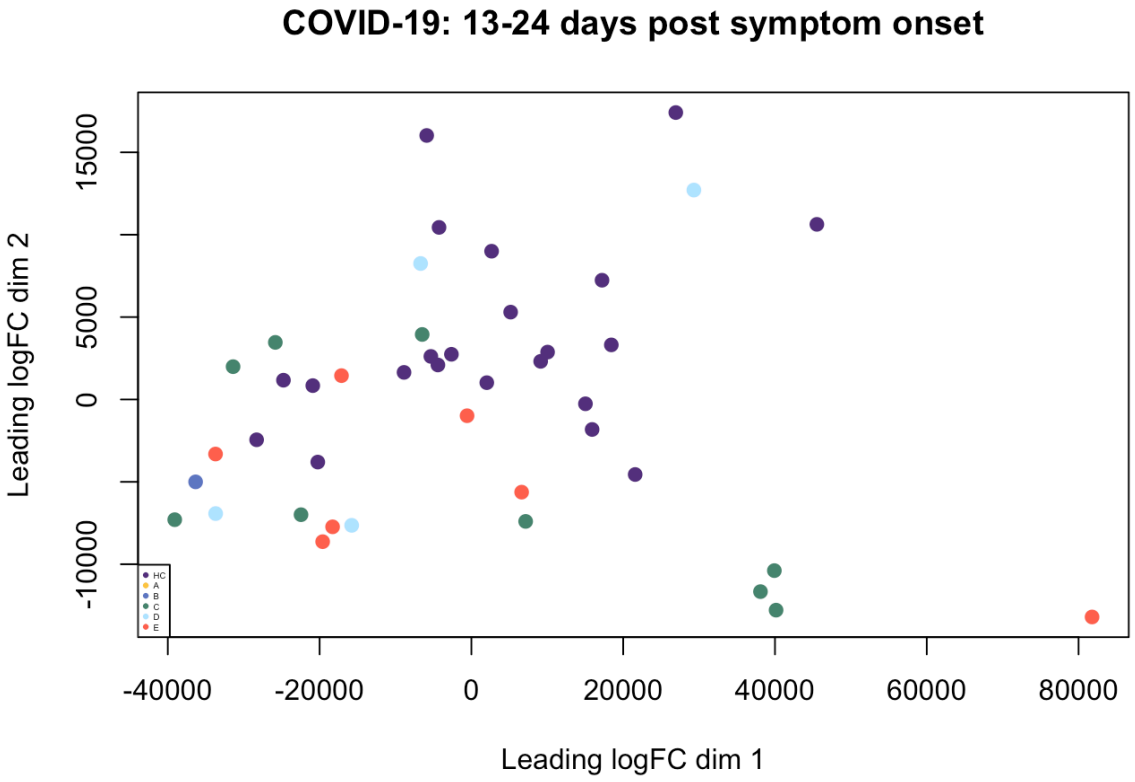


Fig 3.7 PCA at 0-12 and 13-24 days from symptom onset. HC in purple, A in yellow, B in dark blue, C in green, D in light blue and E in red.

Differential gene expression at 0-24 days from symptoms onset showed an increase in the number of differentially expressed genes according to disease severity with group E having the greatest number of differentially expressed genes and group A showing no differential gene expression.

Examining sharing of differentially expressed genes, we found 1070 downregulated genes shared between groups D and E. This was the greatest number shared downregulated genes in all pairwise comparisons with only 204 genes shared between C and D, 164 genes shared between C and E, 81 genes shared between B and D and 13 genes shared between B and C. 1753 genes were shared and downregulated between hospitalised groups C, D and E. 384 genes were shared amongst all 4 groups and 108 genes were downregulated in groups B, C and D. These findings highlight an overlapping transcriptomic signature between hospitalised groups C, D and E and a further similarity between the groups requiring oxygen support, D and E (Fig 3.8).

Once again, group E had the greatest number of differentially upregulated genes compared with healthy controls at just over 4000 genes. Similar to the pattern seen in downregulated genes, groups D and E had the greatest pairwise sharing of upregulated genes at 789 genes. The hospitalised groups C, D and E had the greatest overlap of genes at 1886 genes. A shared COVID-19 transcriptional signature was present with sharing of 231 genes between all 4 severity groups (Fig 3.8).

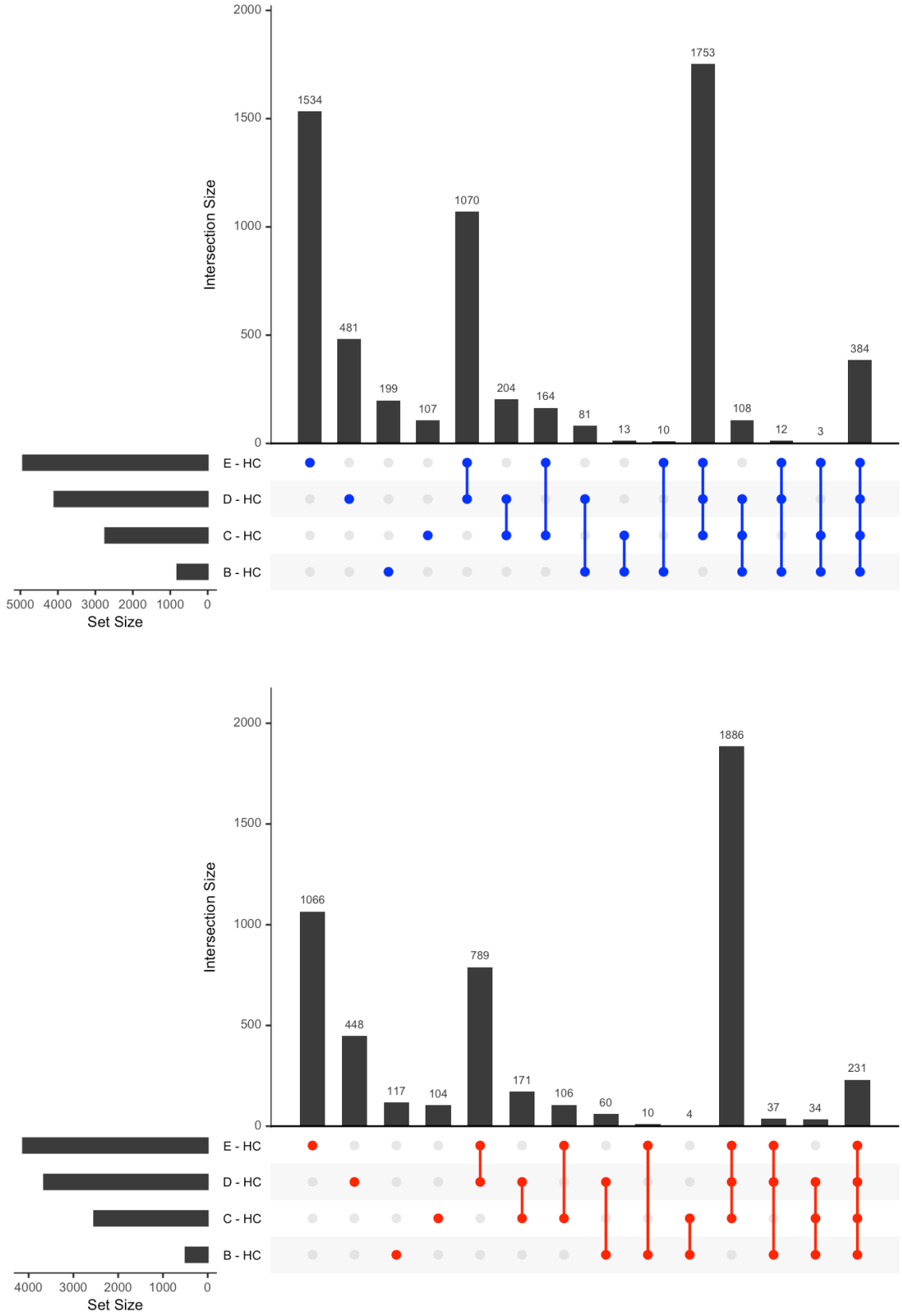


Fig 3.8 Differential gene expression 0-24 days from symptom onset. Blue represents genes downregulated and red represents genes upregulated.

We then examined the PCA plots at later time points to assess for signs of recovery. PCA revealed separation of healthy controls from group E patients at only 25-36 days from days from symptom onset (Fig 3.9) with PCA2 separating the two groups. PCA1 appeared to separate a portion of the SARS-CoV-2 infected patients. By 37-48 days from symptom onset, there was no clear pattern present (Fig 3.9).

Differential gene expression showed as expected, differentially expressed genes in group E with 5124 genes differentially downregulated and 3631 genes upregulated when compared with health (Fig 3.10).

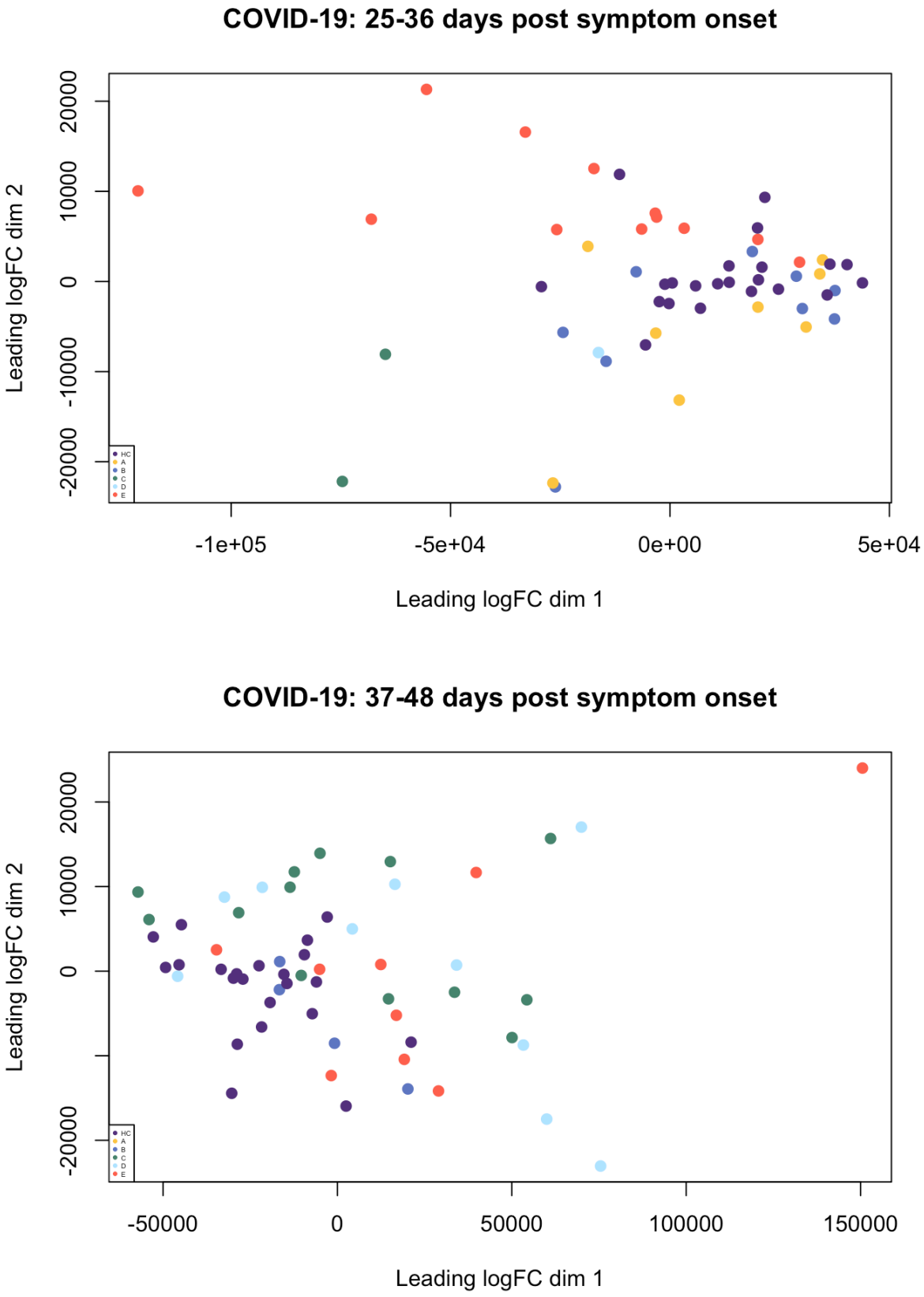


Fig 3.9 PCA at 25-36 and 37-48 days from symptom onset. HC in purple, A in yellow, B in dark blue, C in green, D in light blue and E in red.

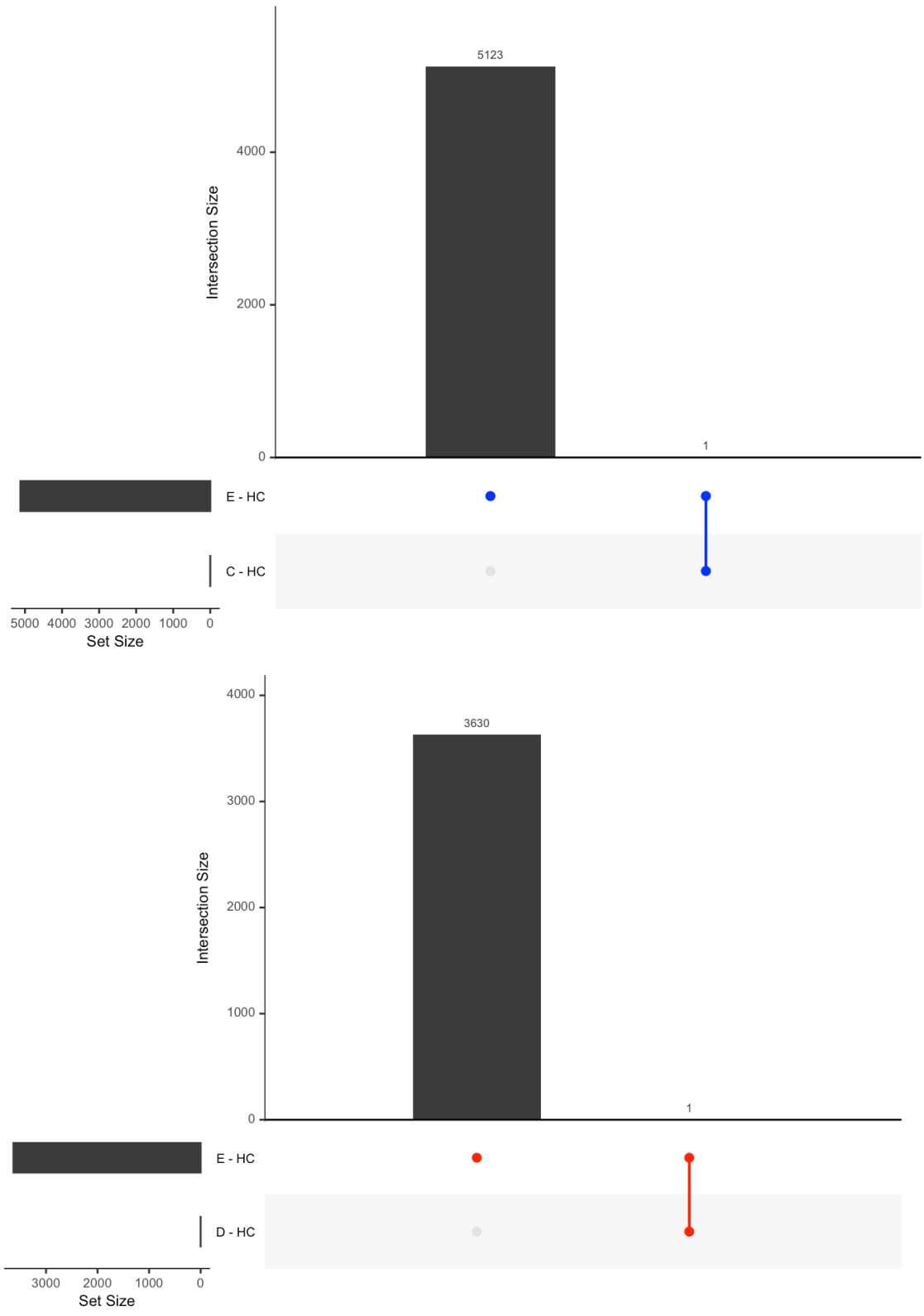


Fig 3.10 Differential gene expression 25-48 days from symptom onset. Blue represents genes downregulated and red represents genes upregulated.

3.2.43 Clustering

We then performed a clustering analysis at both 0-24 (Fig 3.11) and 25-48 (Fig 3.12) days from symptom onset to see how well health separated from disease and if grades of severity caused a further separation. At 0-24 days from symptom onset, using K means clustering, samples were divided into 4 big groups reflective of disease severity as seen with the annotation bars. Healthy controls contributed prominently to a single cluster with a smattering of group A patients (right most). Following this from right to left, the next cluster contained groups A, B and C. The third cluster predominantly contained group E whilst the final last cluster contained predominantly group D patients (Fig 3.11). Three transcriptional groups were formed with cluster 1 (row) enriching for TNFa, IL-6, and ISG pathways.

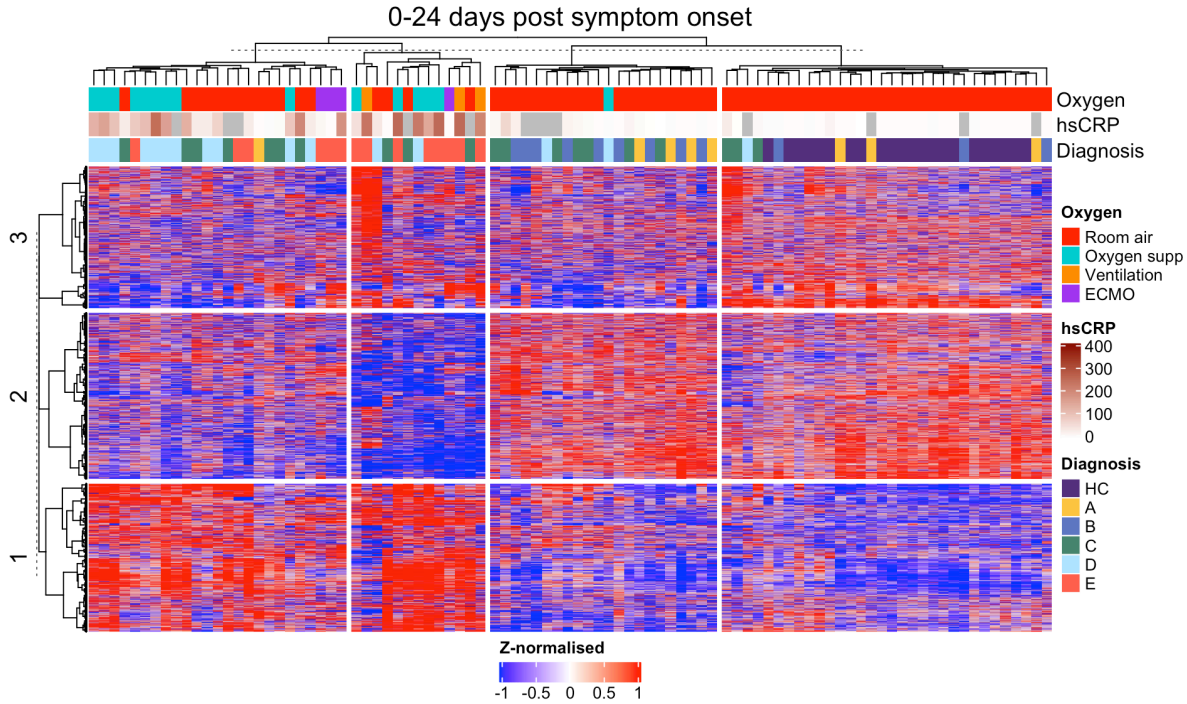


Fig 3.11 Kmeans clustering at 0-24 days from symptoms onset. K-means clustering of 18357 whole blood transcripts from COVID-19 samples.

At 25-48 days from symptom onset, using K means clustering, samples were divided into 6 big groups once again reflective of disease severity as seen with the annotation bars. Healthy controls contributed prominently to a single cluster (right most) with a smattering of group A, B, C and D patients likely reflecting recovery. A further three clusters contained groups A, B, C and D. The final two clusters contained group E (left most). The annotation

bar demonstrates ongoing disease activity in the two leftmost clusters with ongoing elevated CRP and the requirement of oxygen support (Fig 3.12).

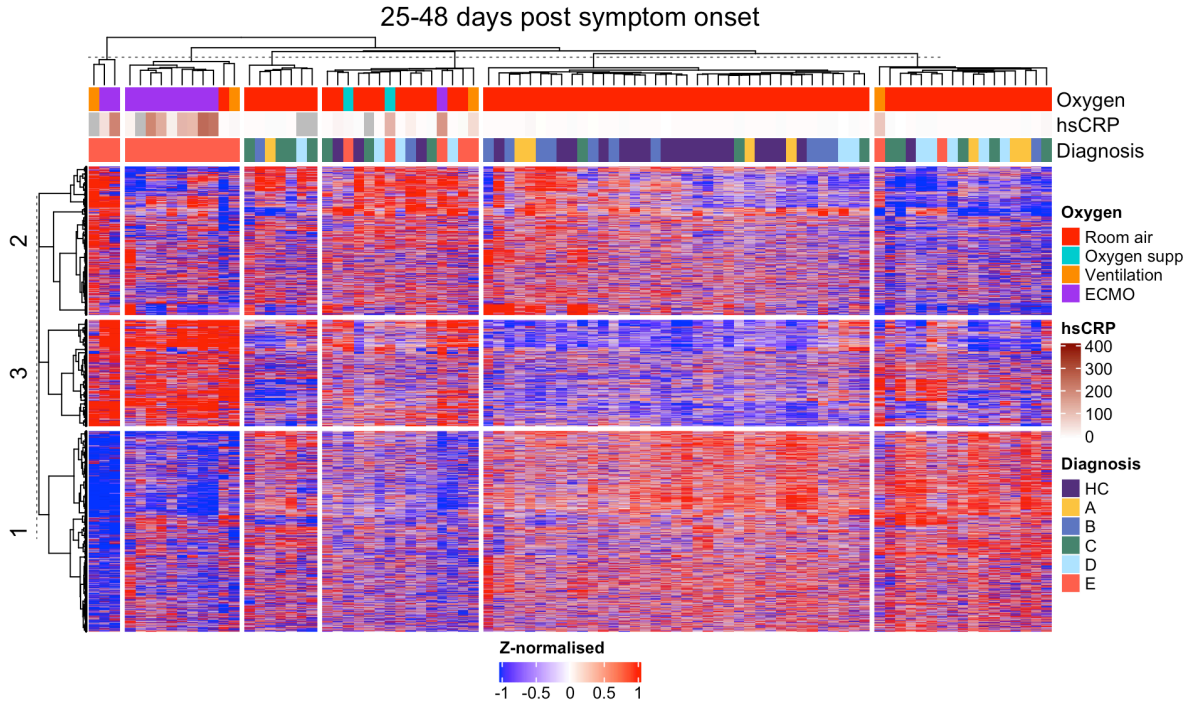


Fig 3.12 Kmeans clustering at 25-48 days from symptoms onset. K-means clustering of 18357 whole blood transcripts from COVID-19 samples.

Lastly, we wanted to assess if the pattern of inflammation changed in group E with time or remained the same. We performed Kmeans clustering with healthy controls and groups D and E at 0-24 and 25-48 days from symptom onset. From the analysis thus far, we expected to find markers of disease activity in groups D and E at 0-24 days from symptom onset and in group E at 25-48 days from symptom onset with signs of resolution in group D at 25-48 days from symptom onset.

Four large patient cluster groupings were generated. The central cluster (second from the right) contained healthy controls and recovered patients from group D at 25-48 days from symptom onset. Interestingly, two large clusters were present which associated with high levels of CRP but clustered away from each other. One cluster contained groups D and E at 0-24 and was driven by Interferon and TNFa whilst the other cluster contained groups D and E at 25-48 days from symptom onset and was driven by Haem metabolism and oxidative

phosphorylation transcriptional signatures (Fig 3.13A). A difference in respiratory support was apparent between these two clusters with the “early severe” cluster mostly on low flow oxygen and the “late severe” cluster mostly on ECMO. This difference in inflammatory patterns was further examined using geneset enrichment analysis and weighted gene correlation network analysis.

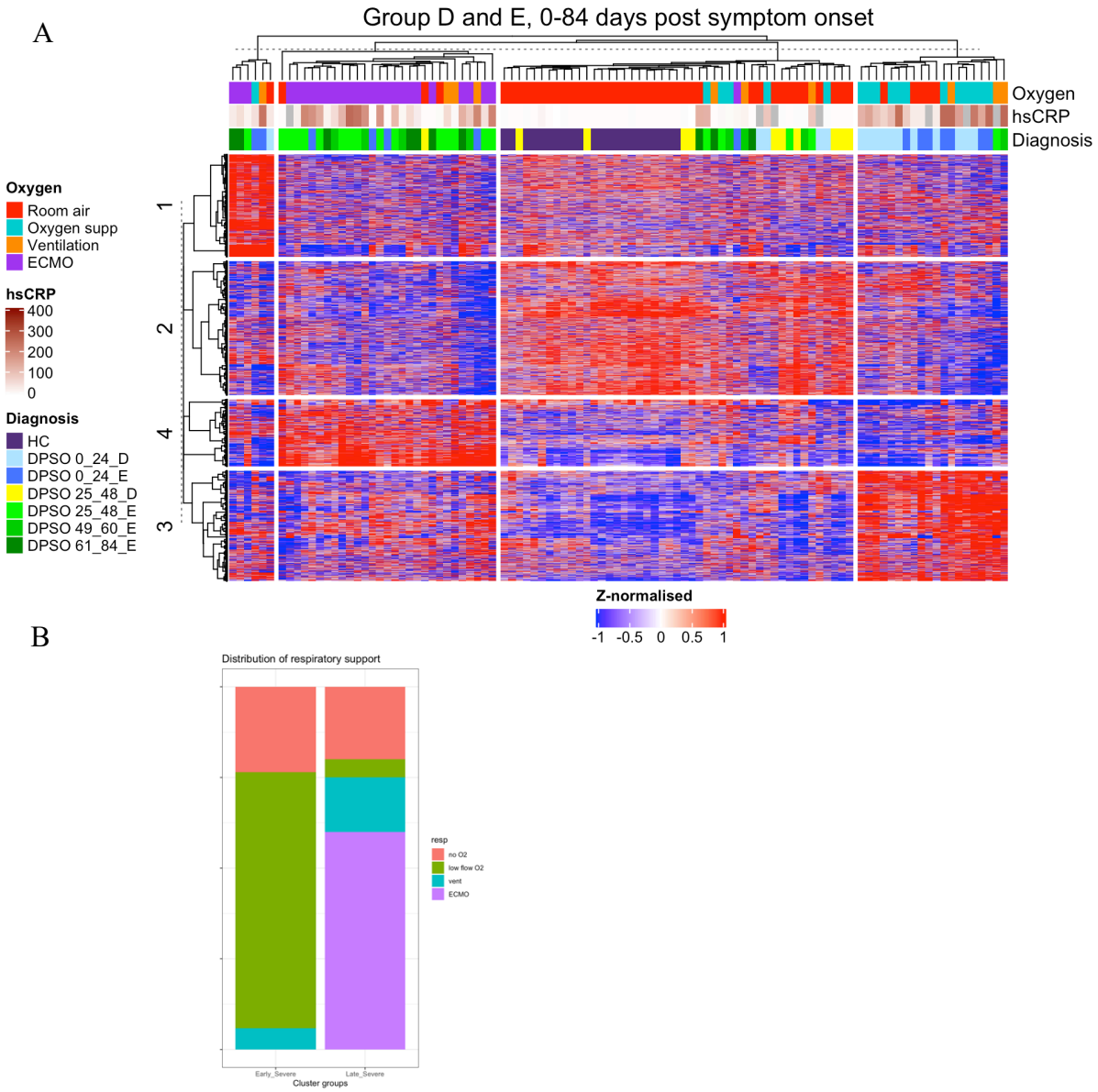


Fig 3.13 Kmeans clustering at 0-84 days from symptoms onset. A. K-means clustering of 18357 whole blood transcripts from COVID-19 samples. B. Distribution of respiratory support between “early severe” and “late severe” groups.

3.2.44 Weighted gene correlation network analysis

We then used weighted gene correlation network analysis (WGCNA) to identify, in an unbiased fashion, modules of co-regulated genes in the whole blood transcriptome data, where each module can be summarised as an "eigengene". Clustering of WGCNA modules was visualised (Fig 3.14) showing that some modules were closely correlated such as the lightgreen and magenta modules. Thus, even though these are treated as separate entities they are likely to enrich for similar or related pathways and have similar patterns of correlations with clinical traits.

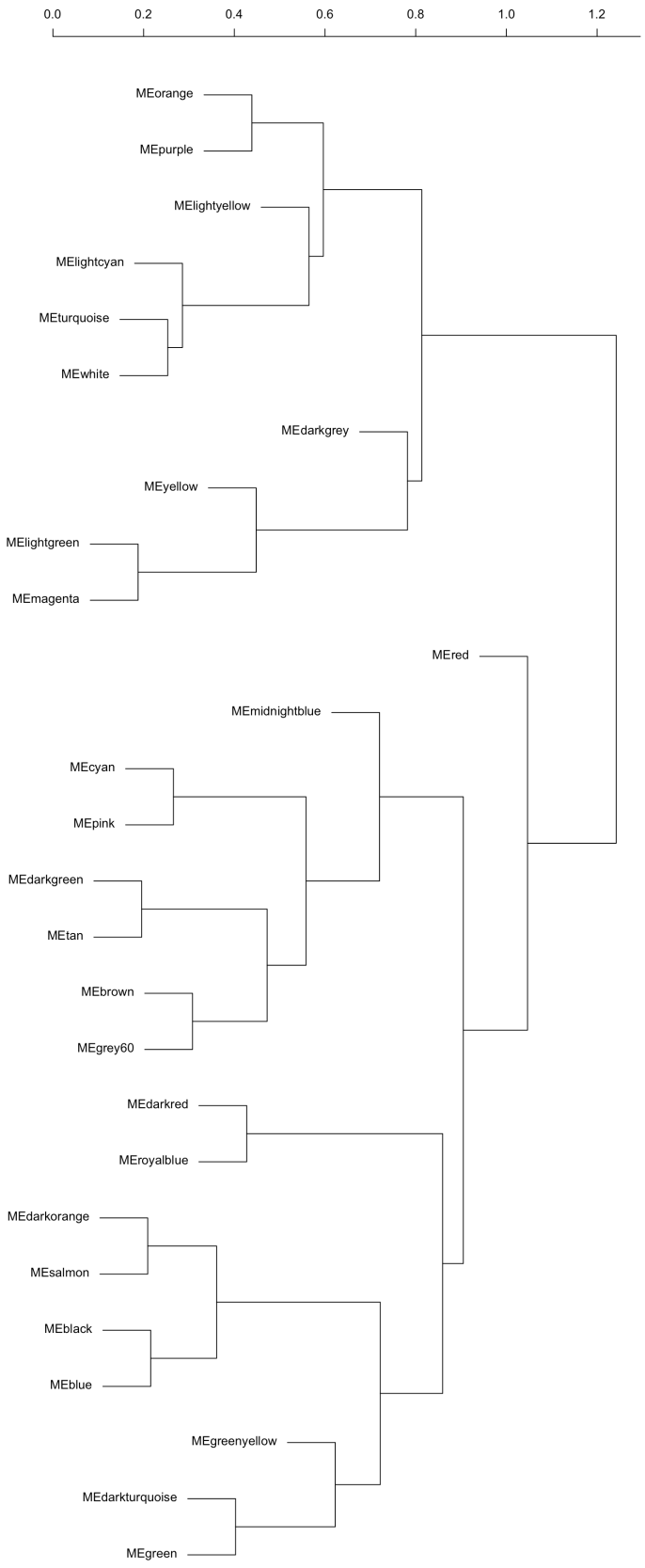


Fig 3.14 WGCNA module formation and clustering

WGCNA modules eigenvalues were then correlated with clinical traits (Fig 3.15). Age and hsCRP were recorded as a continuous variable whilst sex and steroids were treated as binary (sex: F=0, M=1, steroids: no=0, yes=1). COVID-19 categories were split according to disease severity groups and then further into time bins. Comparisons were made with health and the categories given binary values (Healthy=0, Disease=1). Group E at >48 days from symptom onset was included in this analysis to try and ascertain if recovery had occurred within this time window.

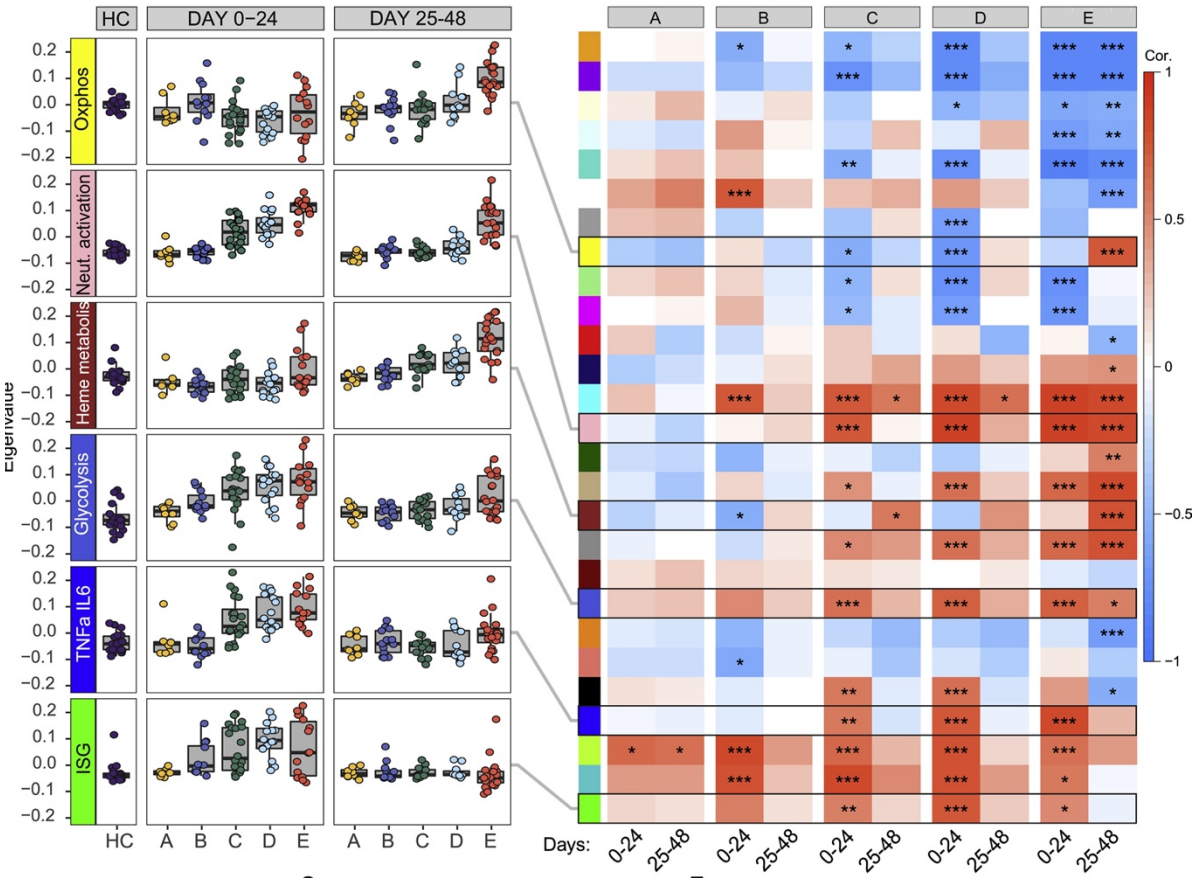


Fig 3.15 WGCNA module and trait correlations. Heatmap derived from WGCNA, illustrating the correlation of whole blood co-expression gene modules (coloured blocks, y axis) with COVID-19 severity groups (x axis) split by 24-day time bins. Boxplots displaying eigengene of key transcriptomic modules according to disease severity and time. Boxes are coloured by strength of correlation.

WGCNA modules were annotated using EnrichR and grouped according to correlation patterns. Prominent gene expression modules were observed, that correlated with both disease severity and time (Table 3.1). Gene modules that were upregulated in disease

compared with health within the first 24 days from symptom onset included histones, TNFa/IL-6, complement, coagulation, neutrophil degranulation, platelet activation, ferroptosis, glycolysis, interferon stimulated genes and immunoglobulins. Gene modules that remained elevated after 24 days in group E included histones, TNFa/IL-6, complement, coagulation, neutrophil degranulation, platelet activation, ferroptosis, glycolysis. Modules that became elevated at this later time point were heme metabolism and oxidative phosphorylation.

1.Positively correlated in all severity groups during early disease and remains positively correlated in group E in all later time bins.

MEcyan: Histones

MEblue: TNFa/IL-6

MEpink: Complement/Coagulation/Neut degranulation

MEmidnightblue: Platelet activation

MEgrey60: Ferroptosis

MEroyalblue: Glycolysis

2.Positively correlated at all time points except late severe

MEgreenyellow: Immunoglobulins

MEgreen: Interferon Stimulated Genes

3.Positively correlated in late mod/severe groups

MEbrown: Heme metabolism

4.Positively correlated in late severe group

MEyellow: Oxidative Phosphorylation

5.Negatively correlated with disease especially in early mod/severe disease

MEdarkgrey: GPCR

MElightgreen: Ribosomal proteins

MEmagenta: MYC targets

6.Negatively correlated with disease especially in late severe group

MEturquoise: Gene transcription

MEpurple: Splicesome

MElightyellow: BCR signalling

MElightcyan: IL-2/NK

Table 3.1 Annotation of WGCNA modules with further grouping according to correlation patterns.

Using bi-weighted correlation to model the relationship between a continuous and binary variable can be inaccurate. Logistic regression is a more appropriate method or a group comparison analysis. Thus, we compared eigenvalues between disease groups according to time windows as well as using a mixed effects model (Fig 3.16). This highlighted the increased expression of TNFa/Il-6, neutrophil activation and glycolysis early in disease in the hospitalised group which persisted in group E at the later time window. Interferon stimulated genes was upregulated in groups B, C, D and E only in early time windows. Lastly

oxidative phosphorylation and Heme metabolism were increased most prominently in group E at the later time window.

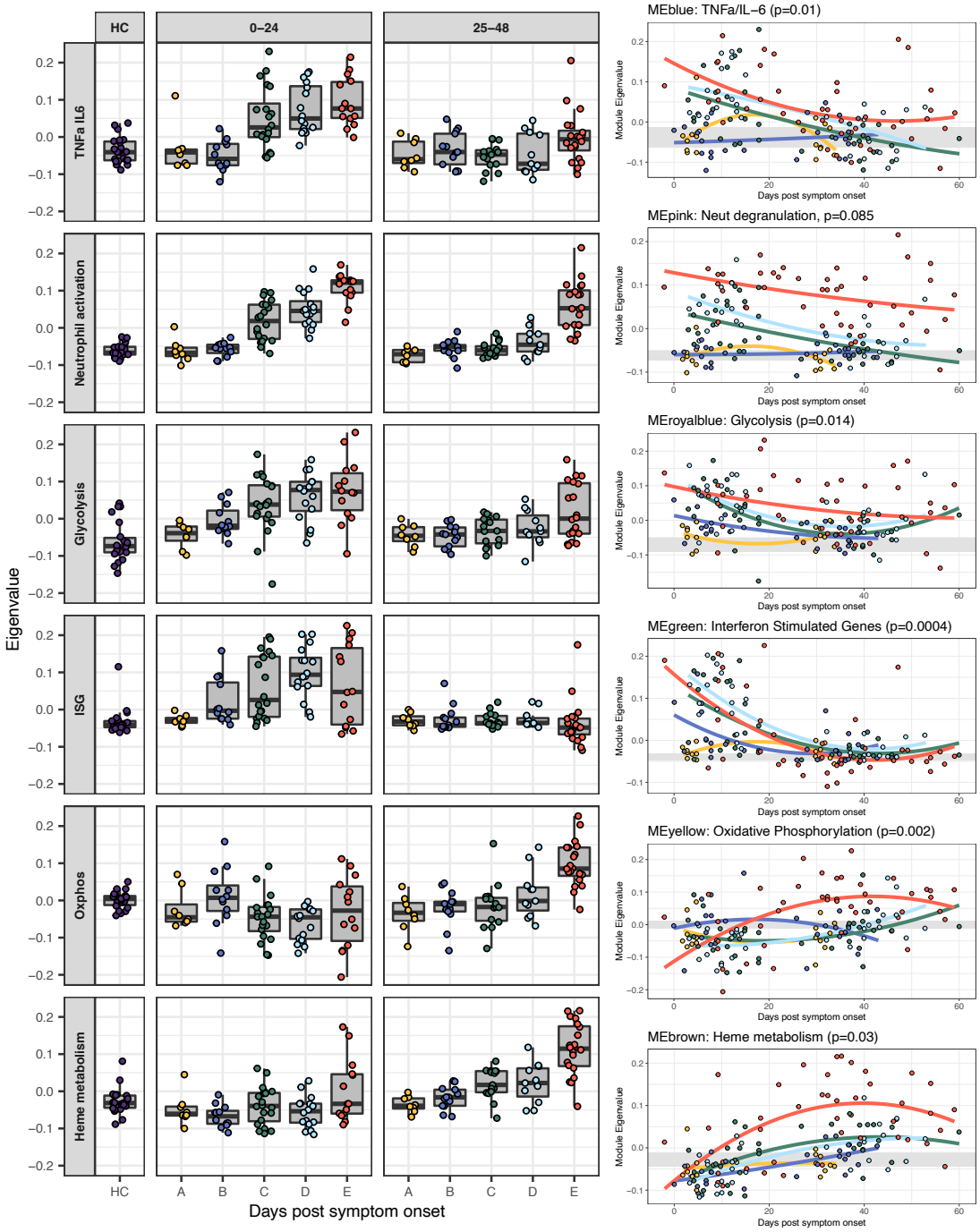


Fig 3.16 WGCNA eigen values across time and severity groups. Mixed-effects model with quadratic time trend showing the longitudinal expression of key eigengene over time, grouped by severity. Grey band indicates the interquartile range of the corresponding eigengene in HCs. Nominal and adjusted p-values for the time x severity group interaction term are reported.

The module enriching for TNF-α /IL-6 genes correlated well with the cytokine levels determined in Fig 3.2 – rising early in groups C-E and then largely resolving by 25-48 days. A neutrophil activation module was also prominent early across groups C to E and remained prominent in group E at 25-48 days, similar to what was found using PLIER. Thus, there is clear transcriptional evidence of activation of broad inflammatory pathways at early time points, and these largely recover in most patient groups (with the exception of group E, in which many patients have persistent disease).

In contrast, an interferon-related module is upregulated prominently in groups B-E at day 0 to 24 from symptom onset, but declines at later time points. As previously described[248], the relative contributions to this module by Type I, II and III interferons cannot be easily distinguished at the transcriptome level. Analysis of the kinetics of this interferon-stimulated gene module shows that, while expression peaks at different levels in each severity group it then declines in all of them by around 30 days coincident with viral clearance and occurring irrespective of clinical and inflammatory state (Fig 3.17).

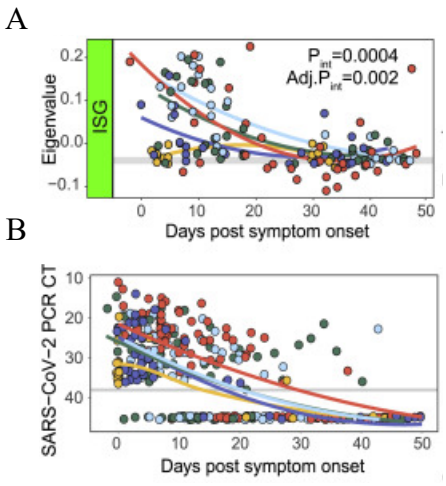


Fig 3.17 Mixed-effects model showing longitudinal expression of eigengene capturing interferon-stimulated genes (ISG) (A) and equivalent mixed-model showing changes in SARS-CoV-2 PCR cycle threshold (viral load) by time and severity (B). y axis inverted in (B).

There was a weak, non-significant correlation between viral clearance and the transcriptional signature of neutrophil activation which remained particularly prominent in the later time window for group E (Fig 3.18).

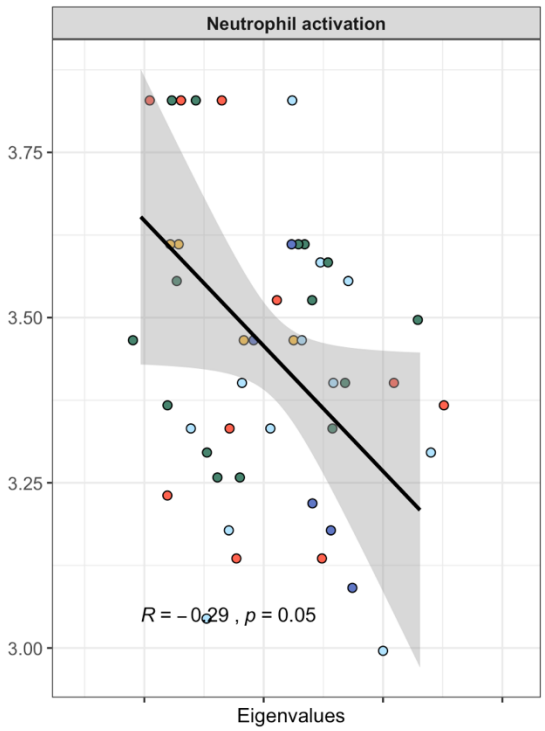


Fig 3.18 Correlation between module eigenvalues and PCR Cycle threshold

Mounting a sufficient Interferon response is necessary in viral control. We hypothesized that the strength of this early antiviral response may help govern outcome and thus an initial higher interferon response may be associated with a better prognosis. We stratified patients in group E at 0-24 days from symptom onset into two subgroups based on interferon expression. We found that those in group E with low interferon signatures in early disease were more likely to have persistently high and ongoing respiratory support. (Fig 3.19).

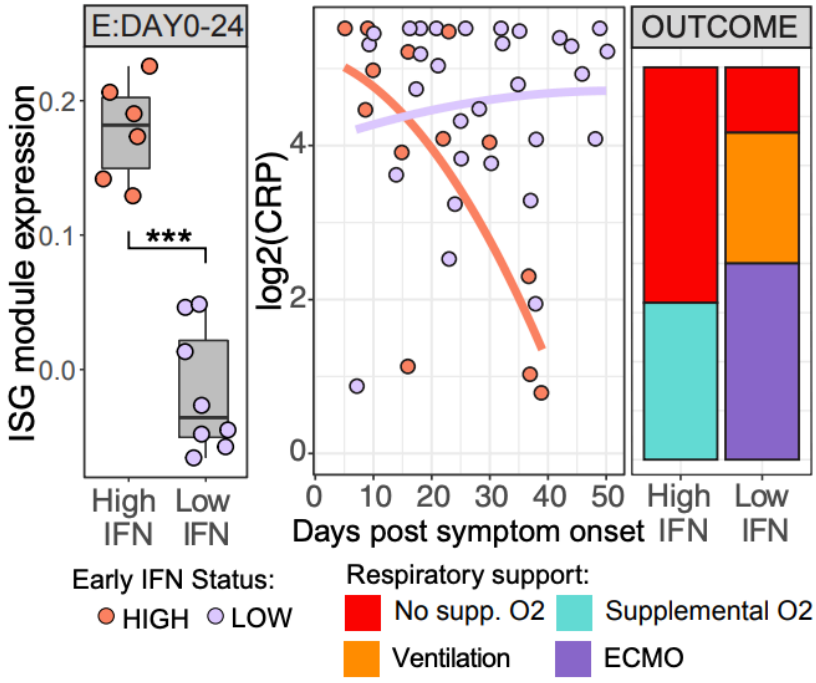


Fig 3.19 Interferon expression and recovery. Stratification of group E samples taken <24 days post symptom onset into high and low expression of interferon stimulated genes (ISG), with persisting and resolving CRP status and final respiratory status reported within 12 weeks shown by bar charts.

3.2.45 Geneset Enrichment Analysis

To further understand these transcriptional signatures, we performed a supervised gene set enrichment analysis (GSEA) using publicly available Hallmark gene signatures (Fig 3.20)[249]. These findings were largely consistent with those generated from the unbiased approaches above. At 0-24 days from symptom onset, there was enrichment of pathways; TNFa, Interferon alpha and gamma, IL6-Jak-stat, complement and coagulation in groups C, D and E groups. At the later time points, the nature of inflammation changed in groups C, D and E, moving from enrichment in interferon to a late upregulation of genes associated with reactive oxygen species, oxidative phosphorylation and heme metabolism.

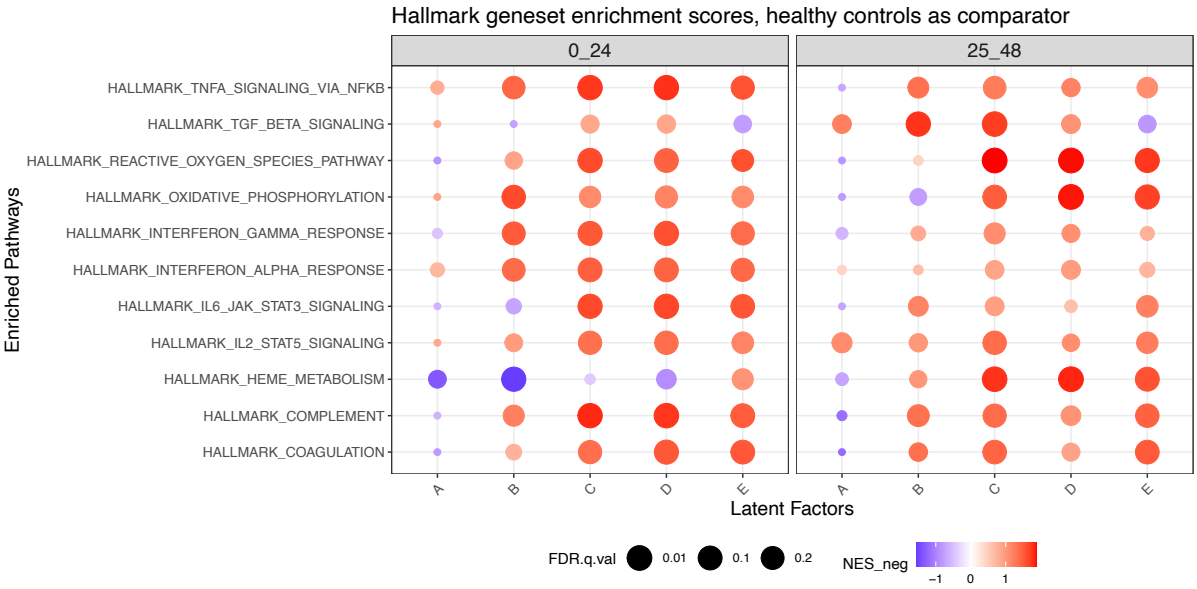


*Fig 3.20 GSEA using select hallmark genesets. GSEA assessing enrichment for HALLMARK genesets against HC in COVID-19 cases split by severity categories and 24-day time bins, FDR adjusted p-value: *0.2, **0.1, ***0.01 and ****0.001.*

3.2.5 Correlation between transcriptomics and immunophenotyping

In order to understand the relationship between transcriptional signatures, immunophenotyping, cytokines and viral load, we examined the correlation between these metrics at two time windows, 0-24 and 25-48 days from symptom onset. At 0-24 days from symptom onset, the following relationships were apparent (Fig 3.21).

- Gene modules TNFa, Neutrophil degranulation, Platelet activation, Ferroptosis, Glyclyosis, Immunoglobulins, Interferon and Heme metabolism positively correlated with one another. These modules in turn positively correlated with hsCRP, IgA, IgM and IgG spike titres, C3a, C3c and inflammatory cytokines IL6, IL1B, IL10, TNFa and IFNg. These modules had a negative correlation with CD4 and CD8 T cell subsets. This relationship is likely a readout of the level of disease severity and inflammation.
- Transcription signatures oxidative phosphorylation, GPCR, Ribosomal proteins and MYC targets positively correlate with CD4 and CD8 subsets, gd T cells and NK cells with a weaker relationship with B cells and a negative correlation with inflammatory cytokines. This relationship is once again likely a readout of the level of disease severity and inflammation.

At 25-48 days from symptom onset. The overall patterns of correlation appeared macroscopically similar although less marked. Of note, oxidative phosphorylation at this time point correlated with inflammatory cytokines and complement activation and is discussed further below (Fig 3.22).

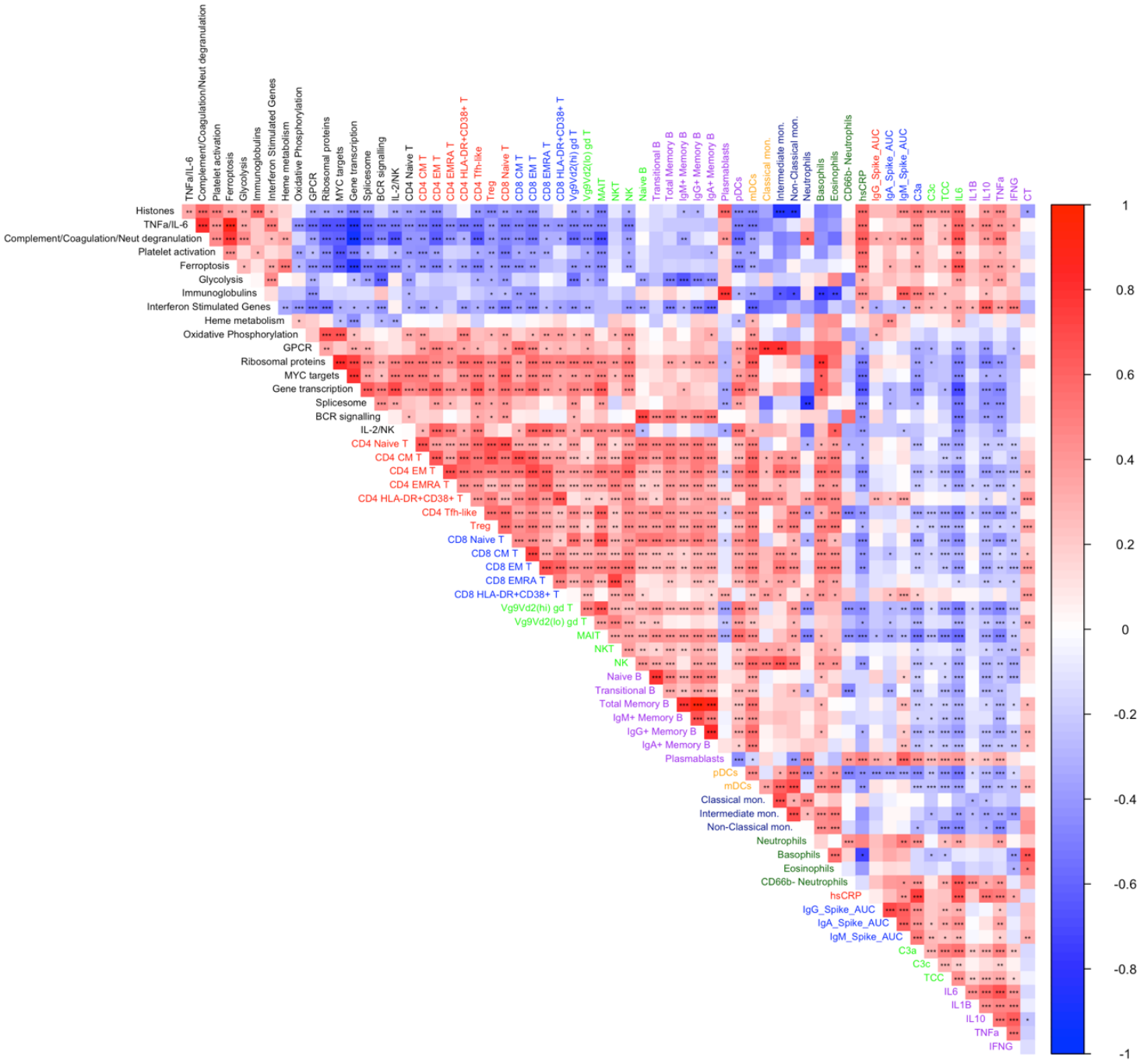


*Fig 3.21 Correlation heatmaps at 0-24 days from symptom onset. Heatmap showing the correlation between gene expression eigenvalues derived from whole blood RNA-Seq, absolute cell counts and inflammatory characteristics in COVID-19 patients collected within 0-24 days post screening (group A) or symptom onset (groups B-E). Pearson correlation p-values: *p<0.05, **p<0.01and ***p<0.001,*

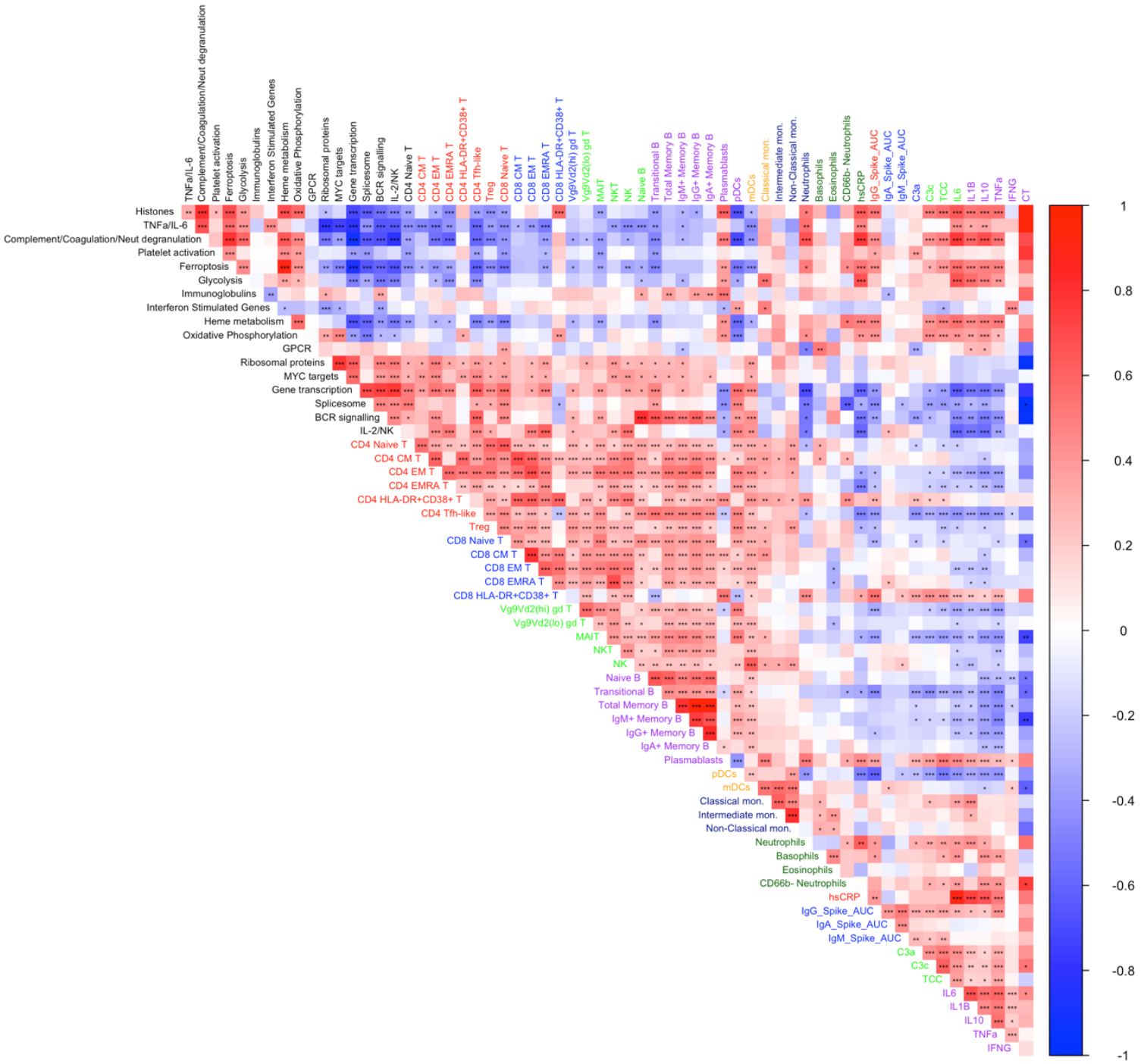


*Fig 3.22 Correlation heatmaps at 25-48 days from symptom onset. Heatmap showing the correlation between gene expression eigenvalues derived from whole blood RNA-Seq, absolute cell counts and inflammatory characteristics in COVID-19 patients collected within 25-48 days post screening (group A) or symptom onset (groups B-E). Pearson correlation p-values: *p<0.05, **p<0.01and ***p<0.001,*

3.2.6 Multi-omic analysis

Beyond just using linear correlations to understand how the data may relate to one another, we also employed a multi-omic factor analysis (MOFA). MOFA uses a Bayesian Group Factor Analysis framework to decompose data into a small number of latent factors. These latent factors capture the global source of variability and are oblique to one another.

Five data sets were layered, RNAseq, cell immunophenotyping and evolving datasets of small metabolites, lipoproteins and amino acids. These metabolomic/proteomic datasets were not individually analysed as they were still in their infancy at the time of analysis.

The latent factors produced by MOFA are influenced by the data set size. Given RNAseq has a much larger number of features, we applied a higher variance filter, reducing the gene universe to 4000 genes. This still resulted in RNAseq dominating the factors built so further dimensional reduction was required. We therefore applied MOFA to the RNAseq dataset on its own. Thus from 4000 genes, we condensed the features to 86 (latent factors) (Fig 3.23).

Fig 3.23 MOFA applied to RNAseq. Applied LF analysis to reduce RNAseq bulk features from 4000 genes to 86 latent factors. The purple grading indicates the level of variance explained by each factor with Factor 1 and 2 explaining the largest proportion represented by the darkness in purple.

Below is a graphical representation of datasets used in the analysis. The labels on the Y axis represent the omic and the number of features per omic. The x axis represents samples

used in the analysis. Where a continuous line of colour is present across all the omics, this indicates data present from all the omics for a given sample. Missing data for a given sample are coloured grey in a given omic (Fig 3.24).

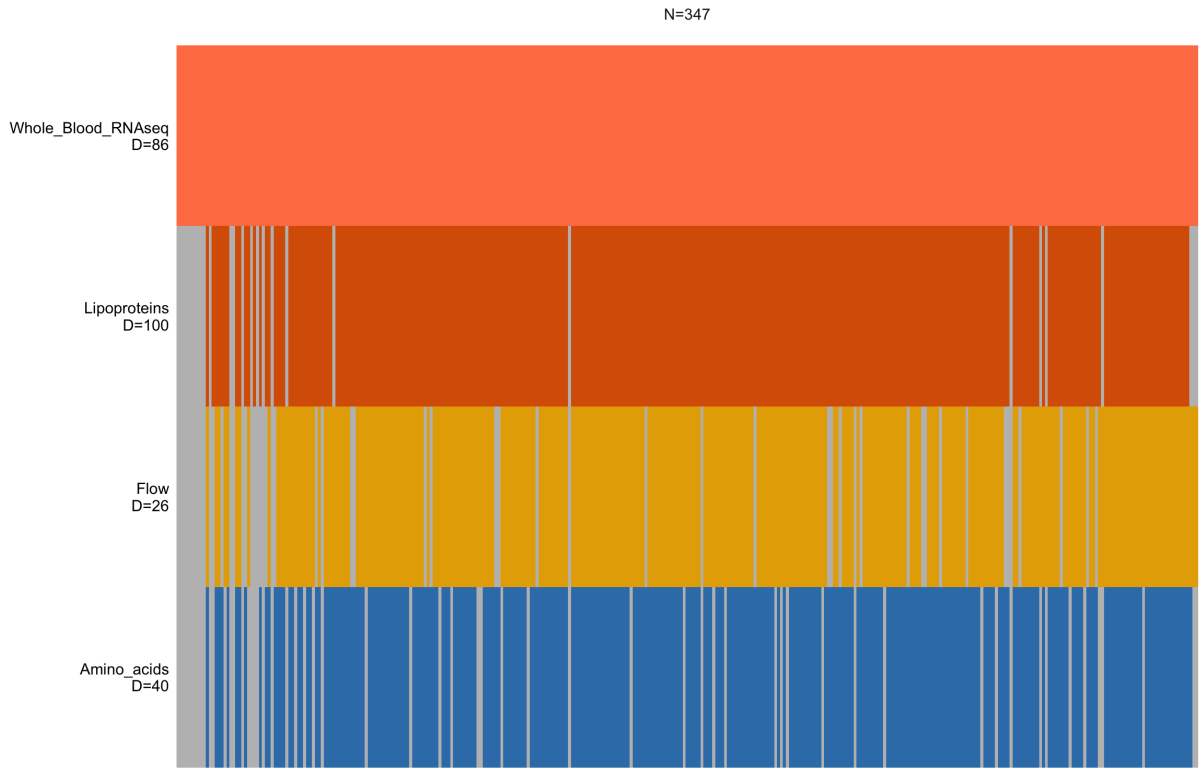


Fig 3.24 Representation of omics used in MOFA. Grey represents missing data for a given patient. N denotes the total number of samples across all omics. D denotes the number of features used per omic.

Using MOFA, we generated eight latent factors. For whole blood RNAseq, the eight latent factors explained less than 10% of the variance. However, given the 86 features themselves are latent factors and each do not carry equal variance weighting, this percentage is not an accurate representation of the total variance explained. The eight latent factors explained 80% of the variance for lipoprotiens, 60% of the variance of amino acids and 50% of the variance of flow data (immunophenotyping)(Fig 3.25 and Fig 3.26). This illustrates that the latent factors capture a large amount of variance in the data and thus adequately models the data.

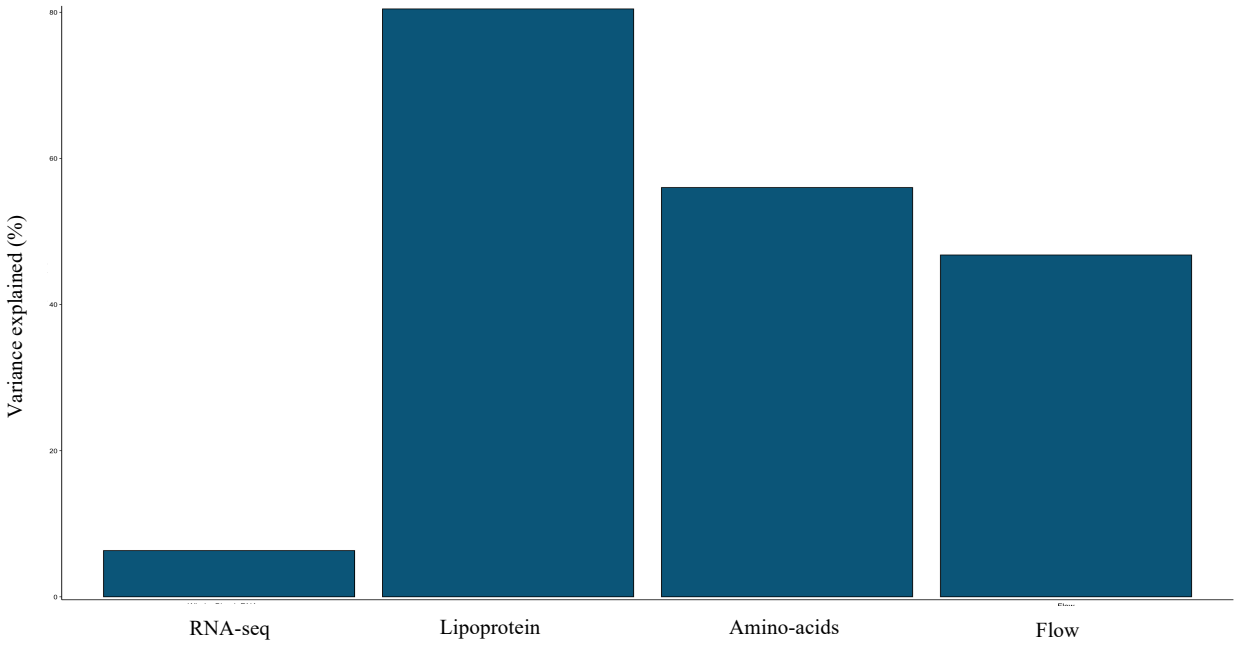


Fig 3.25 Total variance explained post factor decomposition. Variance on y axis and omics on x axis.

Fig 3.26 Shared variance across omics post factor decomposition. Factors on y axis and omics on x axis.

Fig 3.26 Illustrates the latent factors across the omics. Latent factor 1 explains a large proportion of the variance of lipoproteins. Latent factor 3 explains a large proportion of the variance for amino acids whilst latent factor 2 explains a large proportion of the variance for the immunophenotyping data. Latent factor 4 explains the large proportion of the variance for all omics at 2.19% for whole blood RNAseq (although inaccurate), 11.2% for lipoproteins, 5.7% for amino acids and 9.9% for immunophenotyping and is a shared axis of variance (Table 3.2).

	Whole_Blood_RNAseq	Lipoproteins	Amino_acids	Flow
Factor1	2.066037705	38.384461	4.970205e+00	4.1882268
Factor2	0.209660130	2.923329	9.622854e-05	31.4288813
Factor3	0.325693400	0.631666	3.281150e+01	0.1250459
Factor4	2.198818736	11.199691	5.713073e+00	9.8781905
Factor5	0.003172595	1.230850	1.204249e+01	0.2062573
Factor6	0.558028445	11.748009	1.417500e-01	0.6273550
Factor7	0.462334681	8.837407	1.199320e-03	0.1305940
Factor8	0.505595018	5.510050	3.397537e-01	0.2009570

Table 3.2 Variance explained per omic and latent factor

Latent factor 4 expression appeared reflective of disease severity with it being upregulated at 0-12 days and 13-24 days from symptom onset in groups C, D and E and remaining upregulated in group E at 25-36 and 37-48 days from symptom onset. A linear mixed effects model similarly illustrated the heighted expression of latent factor 4 within the first 24 days from symptom onset with it most marked in groups D and E. There was a marked decline in group D and ongoing heightened expression in group E (Fig 3.27).

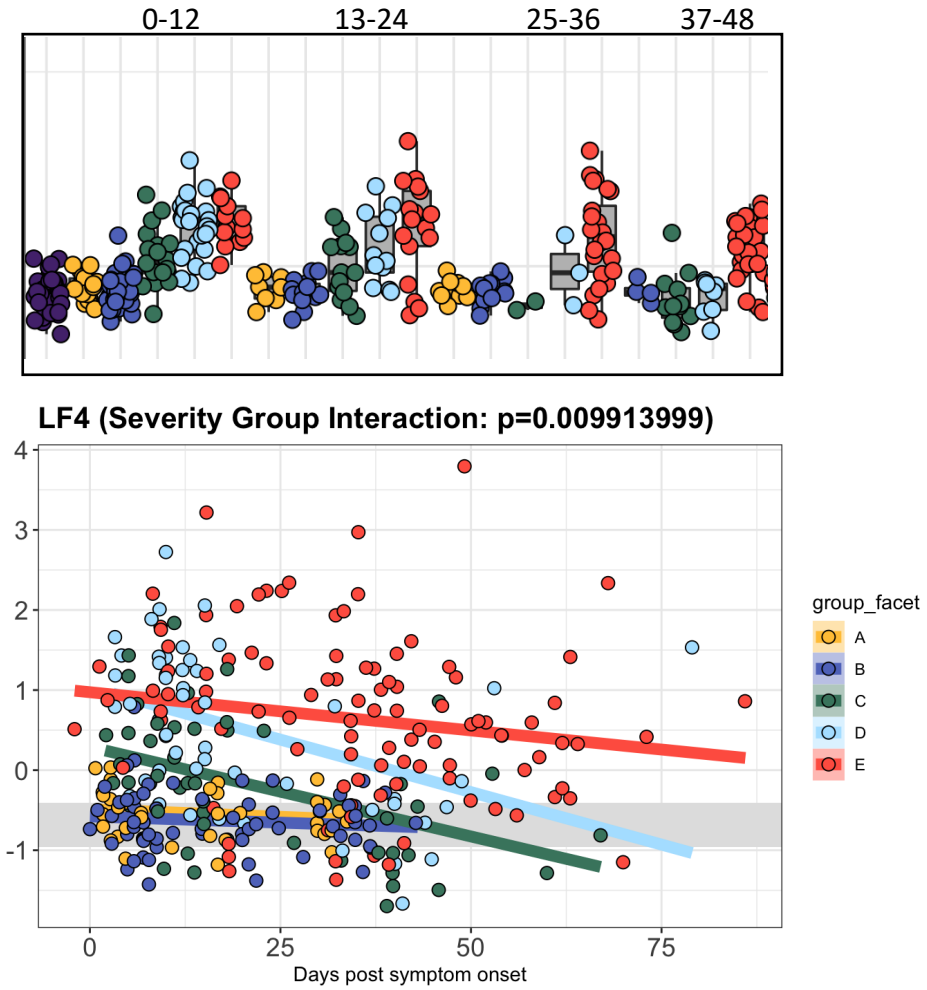


Fig 3.27 Latent Factor 4 represented in time bins and as a linear mixed effects model. The grey band in the linear mixed model is the interquartile range of healthy controls.

We then examined the top weighted features of latent factor 4 for each omic. For whole-blood RNAseq the top weighted feature was latent factor 2 which on initial matrix decomposition explained 9.5% of the variance and enriched for genes in keeping with neutrophil activation. This latent factor had a strong positive correlation with latent factor 4 eigenvalues (Fig 3.28 and Fig 3.29). This pattern of inflammation was mirrored in the WGCNA and PLIER analysis where the module representative of neutrophil activation was elevated post infection and remained so in group E at later timepoints whilst resolving in the other severity groups.

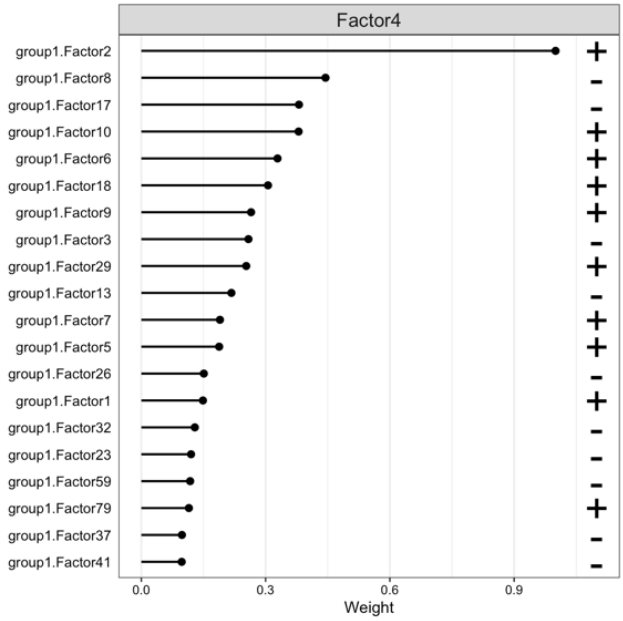


Fig 3.28 RNAseq Latent Factor 4 top weights. On initial dimensional reduction, latent factor 2 explained 9.52% of variance and was representative of neutrophil activation.

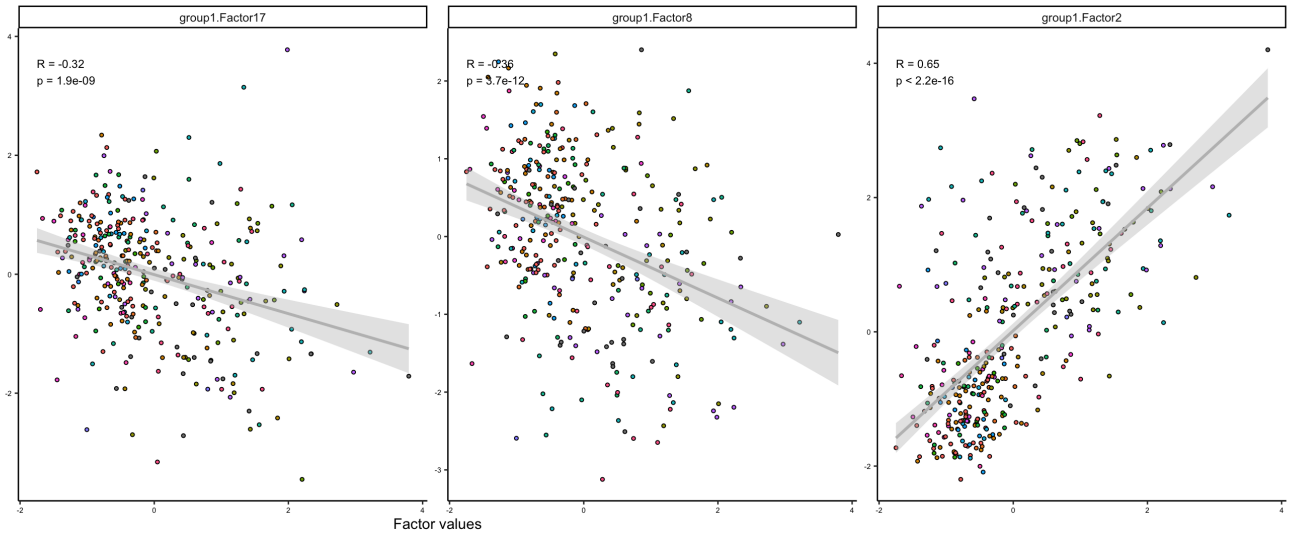


Fig 3.29 Correlations between top RNAseq Latent Factor 4 weights and Latent Factor 4 eigenvalues

For lipoproteins, the top weighted features were H4CH, H4A2, L5CH, L4CH, H4AL, H4PL, V1F2 and H4FC and were tightly negatively correlated with latent factor 4 eigenvalues with the level of correlation (R) ranging from -0.44 to -0.72 (Fig 3.30 and Fig 3.31).

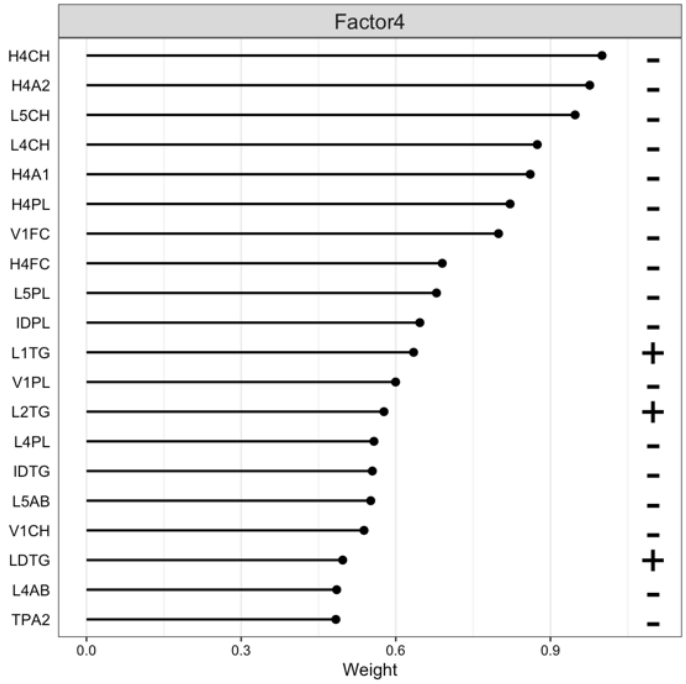


Fig 3.30 Lipoprotein LF4 top weights

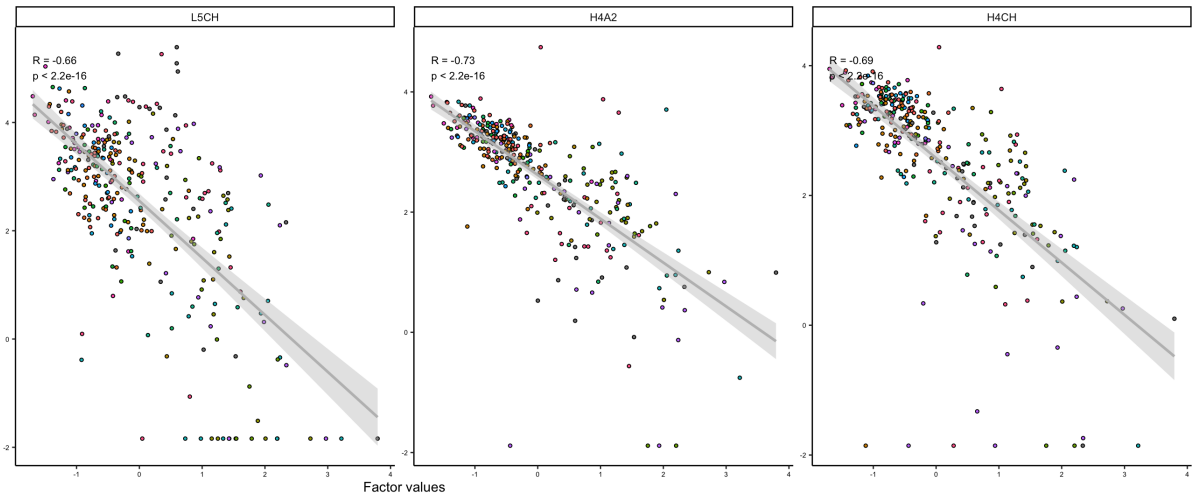


Fig 3.31 Correlations between top lipoproteins LF4 weights and LF4 eigenvalues

For amino acids the top weighted features were quinolinic acid, tryptophan and kynurenine for latent factor 4 (Fig 3.32). Quinolinic acid had a strong positive correlation with an R = 0.56 and tryptophan had a strong negative correlation at -0.56 (Fig 3.33).

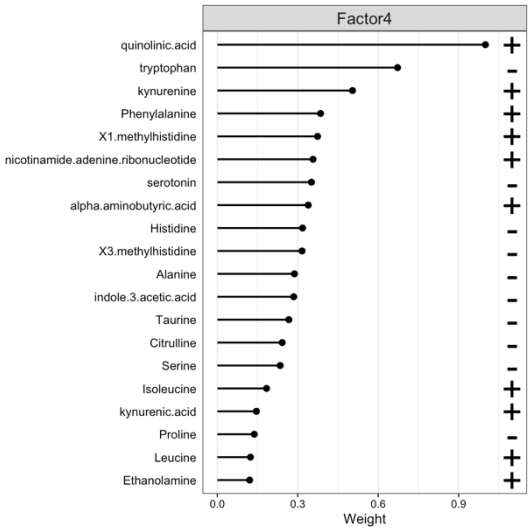


Fig 3.32 Amino-acids LF4 top weights

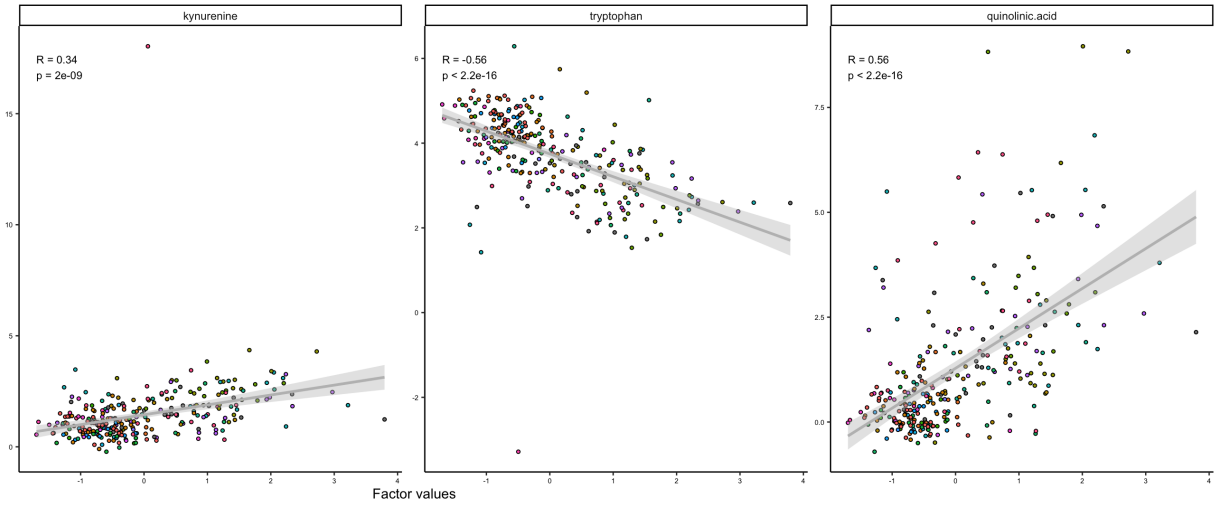


Fig 3.33 Correlations between top amino acids Latent Factor 4 weights and Latent Factor 4 eigenvalues

For cellular immune markers the top weighted features were Vg9Vd2(hi) gd T cells, pDC, CD8 HLA DR+CD38+ T cells and MAIT cells (Fig 3.34). Vg9Vd2(hi) gd T cells, pDC and MAIT cells had a strong negative correlation whilst CD8 HLA DR+CD38+ T cells had a strong positive correlation (Fig 3.35).

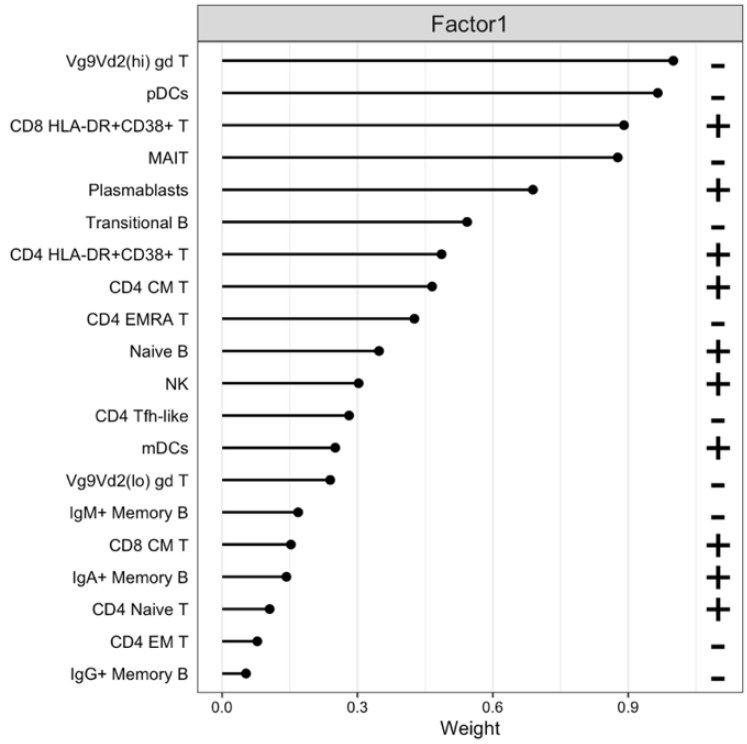


Fig 3.34 Immunophenotyping LF4 top weights

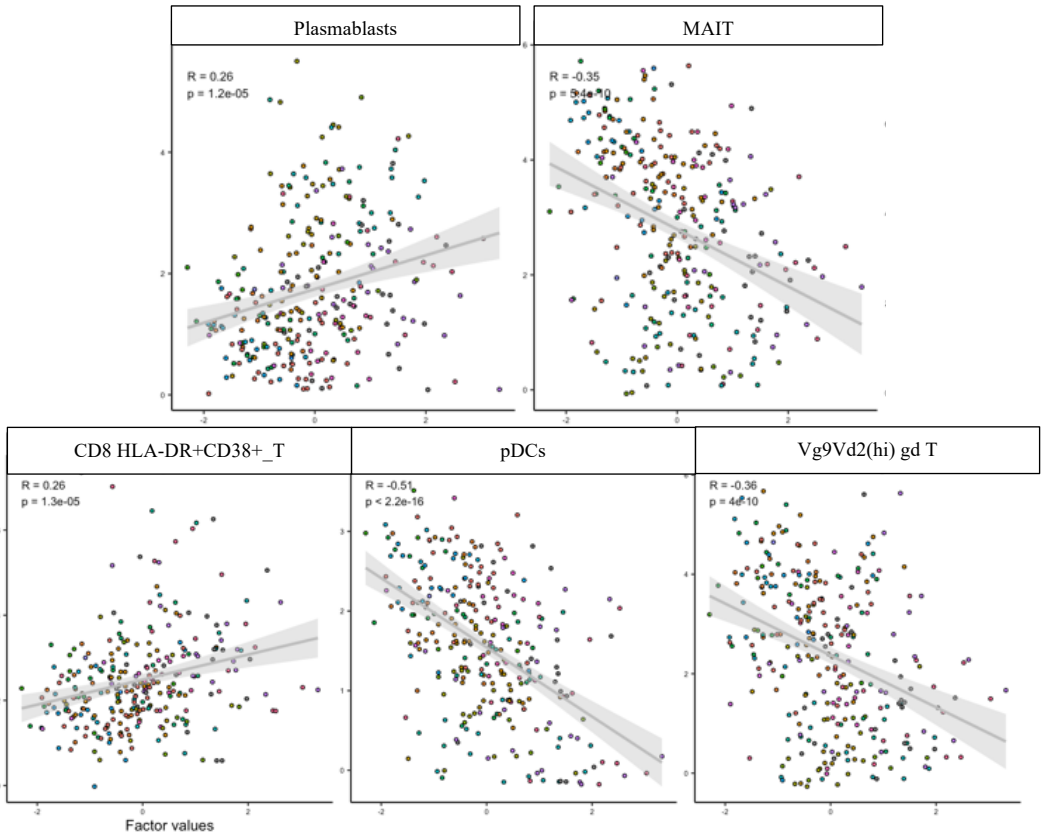


Fig 3.35 Correlations between top immunophenotyping LF4 weights and LF4 eigenvalues

We lastly created a UMAP using information derived from all 5 omic datasets. Samples were coloured according to disease severity and time from symptom onset (Fig 3.36). The UMAP illustrates the close clustering of healthy control samples and patients from group A at 25-36 weeks from symptom onset. This may suggest disease recovery and thus the co-clustering with health. The next closest neighbour to healthy control were samples from group B at 37-48 days from symptom onset and group A at 0-12 days from symptom onset. Peak severity occurs within the first 12 days of symptom onset for group A whilst 37-48 days from symptom onset was a point of recovery for group B. Group E remained far from healthy controls at all time windows. Group D at 0-12 and 13-24 days from symptom onset and group C at 0-12 days from symptom onset co-clustered with group E, likely representative of ongoing disease.

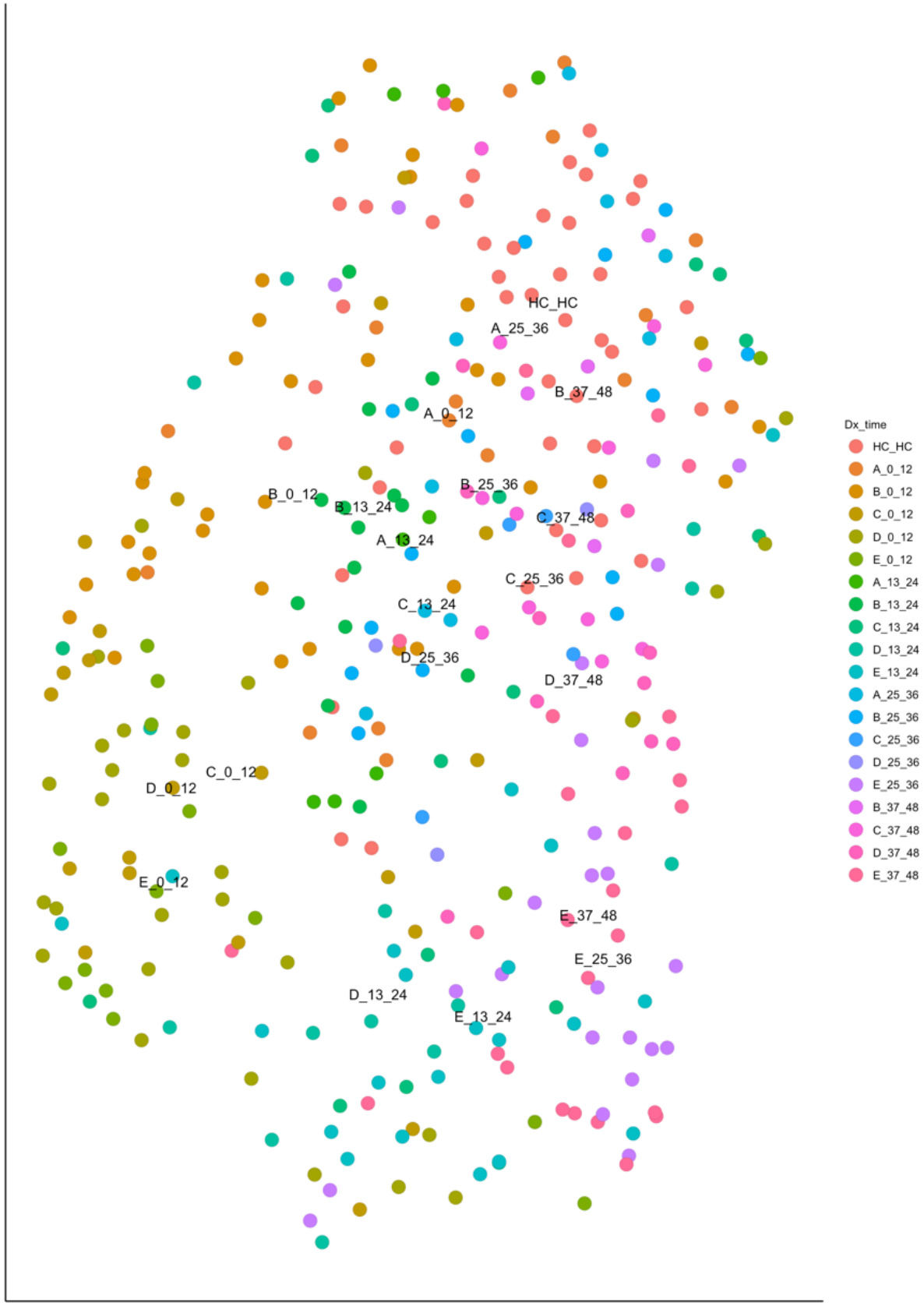


Fig 3.36 UMAP of all samples. UMAP illustrating spatial distribution of samples utilizing data from all 5 omics. Samples are coloured according to disease severity and time from symptom onset.

3.2.7 Transcriptional changes in persisting disease

We wished to further explore the changes in transcriptional signatures with time. At late time points, whole blood transcriptome analysis showed a change in inflammation-related signatures distinct from those that were prominent early in the disease course, particularly in severity groups C-E. These signatures were characteristic of oxidative phosphorylation, reactive oxygen species generation and heme. This contrasted with the neutrophil signature which persists in group E at both time points in patients with ongoing inflammation. These pathways were demonstrated in an un-biased fashion using WGCNA, where modules characterised by oxidative phosphorylation and heme metabolism signatures were prominent in samples analysed at days 25 to 48 post symptom onset, with oxidative phosphorylation most prominent in group E, and heme metabolism in C, D and E (Fig 3.16). Enrichment of Hallmark signatures in RNA-seq datasets confirmed the association of oxidative phosphorylation and heme metabolism in groups C, D and E, and also found association of a reactive oxygen species generation signature (Fig 3.37 and 3.38). Examination of leading edge genes showed the highest expression of the key genes in group E (Fig 3.38).

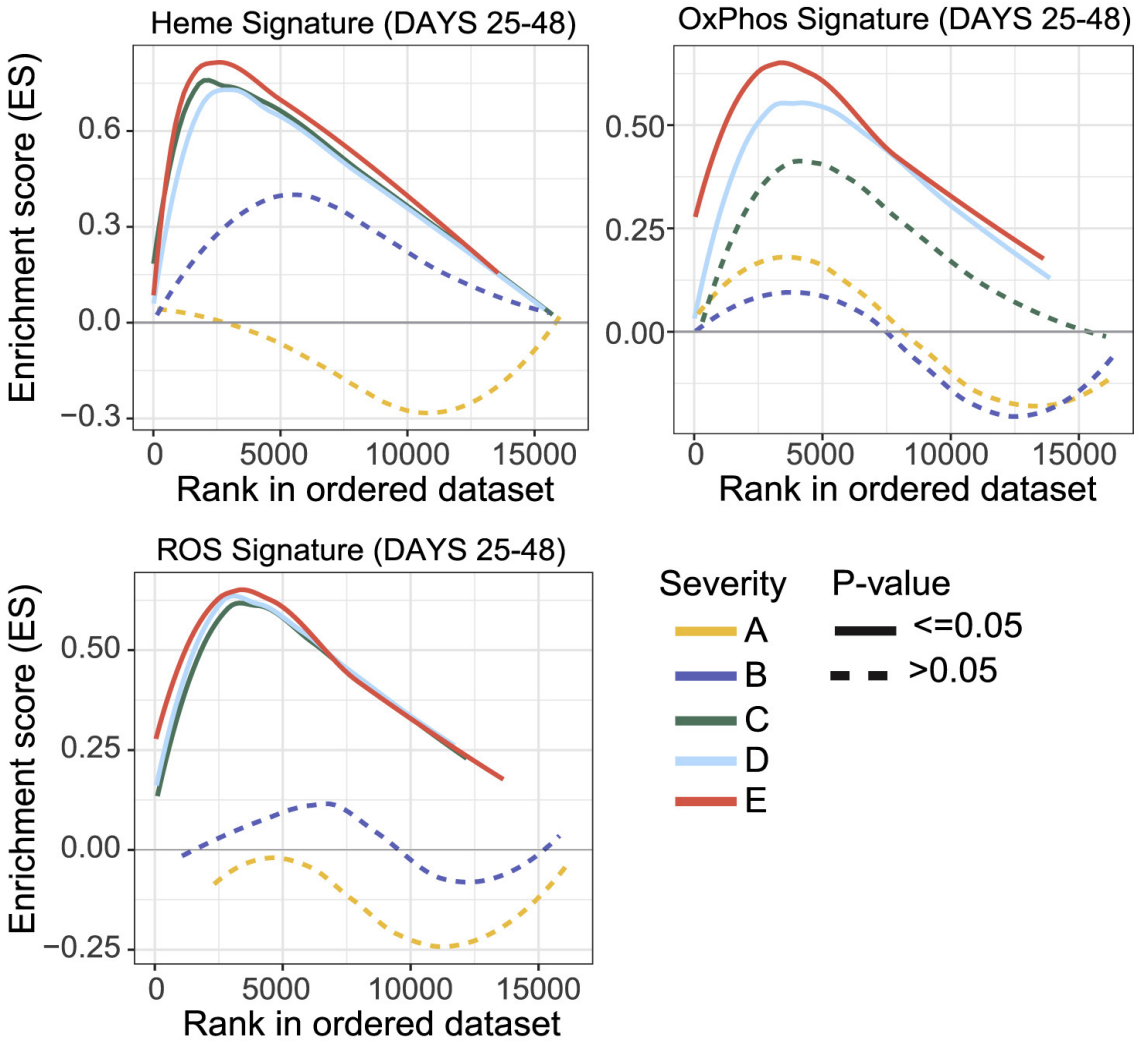


Fig 3.37 GSEA curves at 24-48 DPSO. Enrichment score for HALLMARK genesets capturing heme metabolism, oxidative phosphorylation and ROS related genes (as determined by GSEA) in group A-E samples taken 25-48 days post screening or symptom onset.

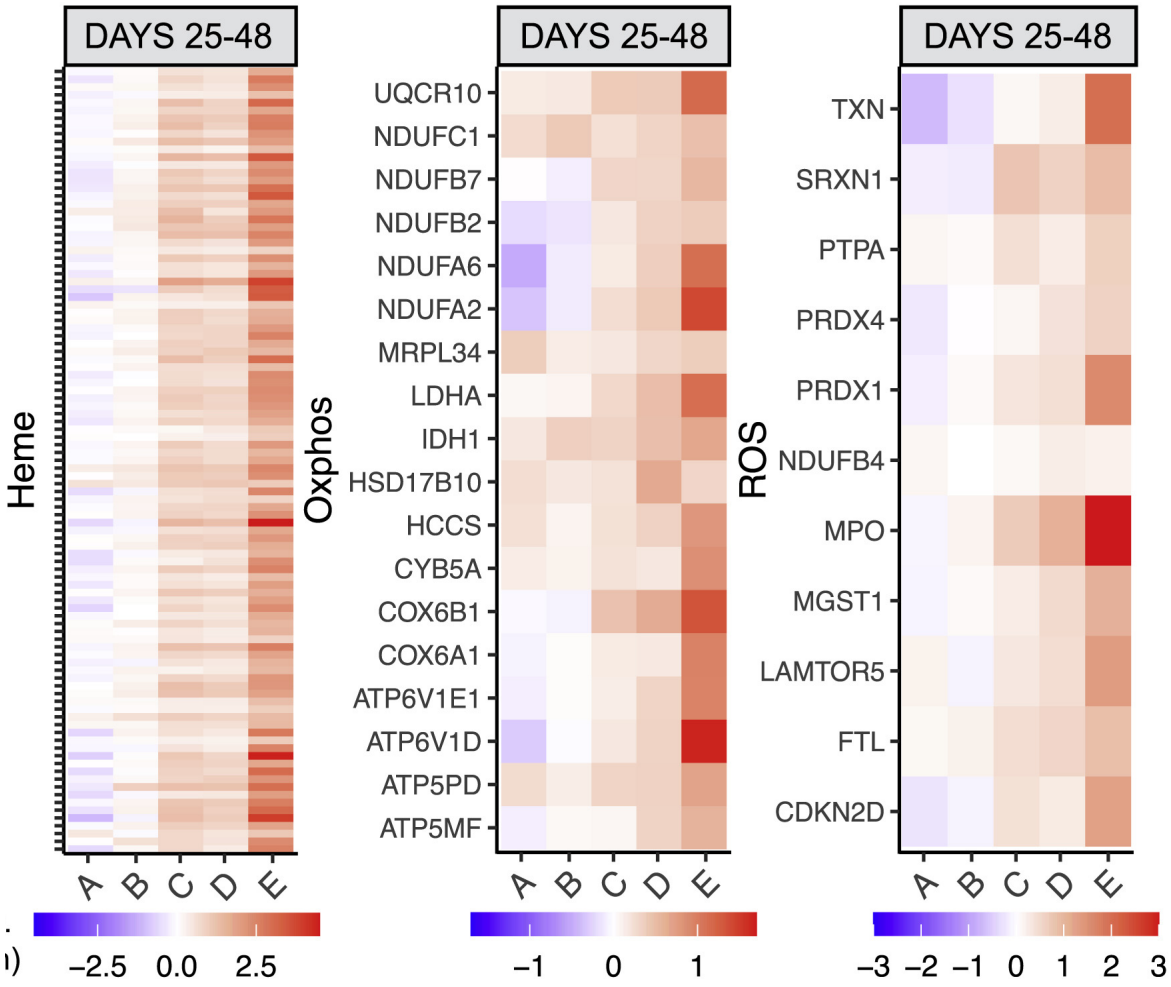


Fig 3.38 Leading edge genes. Heatmap showing enrichment for the intersection of GSEA leading edge genes from groups C, D and E, across all severity groups in samples taken 25-48 days post screening or symptom onset.

As previously mentioned, in the first 24 days after symptom onset, there was a strong association between TNF-α /IL-6, neutrophil degranulation and interferon signatures with most of the lymphoid cell types whose numbers fell in severe disease. However, at 24 and 48 days after symptom onset, these associations changed. While TNF-α /IL-6 and neutrophil degranulation signatures were still associated with many cell subsets that continue to be reduced, the interferon signature was no longer a significant player. Strikingly, the persistent increase seen in effector lymphocytes, both CD4 and CD8 activated T cells (HLA-DR+, CD38+) and plasmablasts, were now associated with the oxidative phosphorylation signature which, having become more prominent later in disease, has a much more restricted and specific association with immune dysfunction than other inflammatory signatures. It is thus clear that, for some cell types, the association with the inflammatory milieu changes over time, but for others it is more consistent. It is interestingly the

inflammatory signatures which appear late in disease, in particular oxidative phosphorylation, are specifically associated with persistent derangement of cell types of potential pathological importance, such as increased HLA-DR+CD38+ T cells and plasmablasts, and reduced pDCs (Fig 3.39). In addition, the heme module negatively correlated with hemoglobin and positively correlated with D.Dimer (Fig 3.40). At 25-48 days from symptom onset, group E showed a negative correlation between reticulocytes (Fig 3.41).

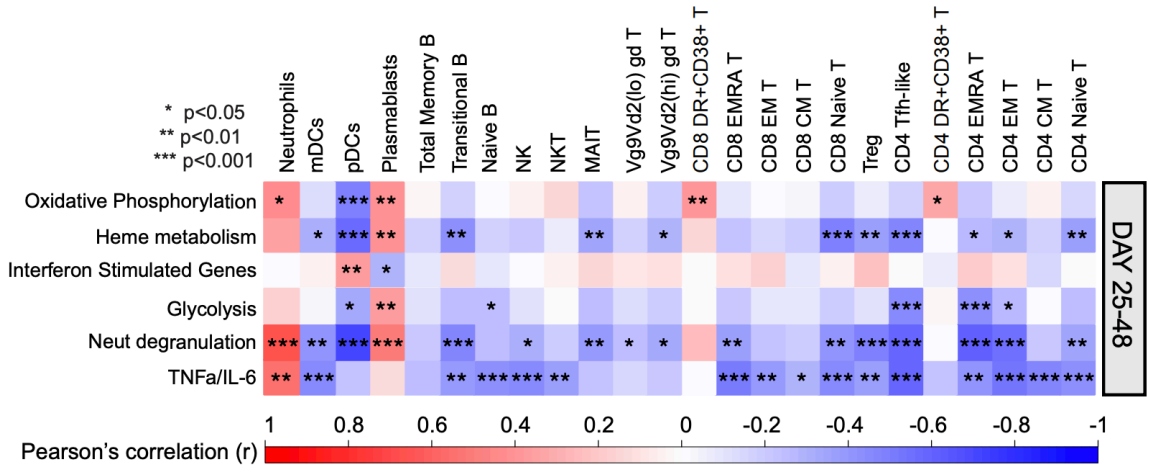


*Fig 3.39 Heatmap showing correlation between transcriptional eigengenes and absolute cell counts, at 25-48 days post symptom onset. Boxes are coloured by strength of correlation, Pearson correlation pvalues: *<0.05, **<0.01,***<0.001,*

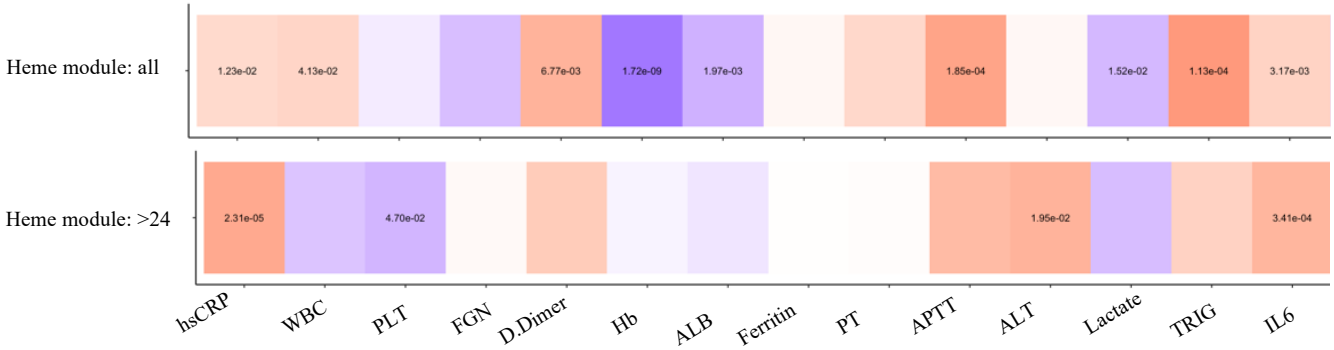


Fig 3.40 Heatmap of heme module correlations. Heatmap showing correlation between heme transcriptional eigenvalues and blood counts and cytokines. Boxes are coloured by strength of correlation, Pearson correlation.

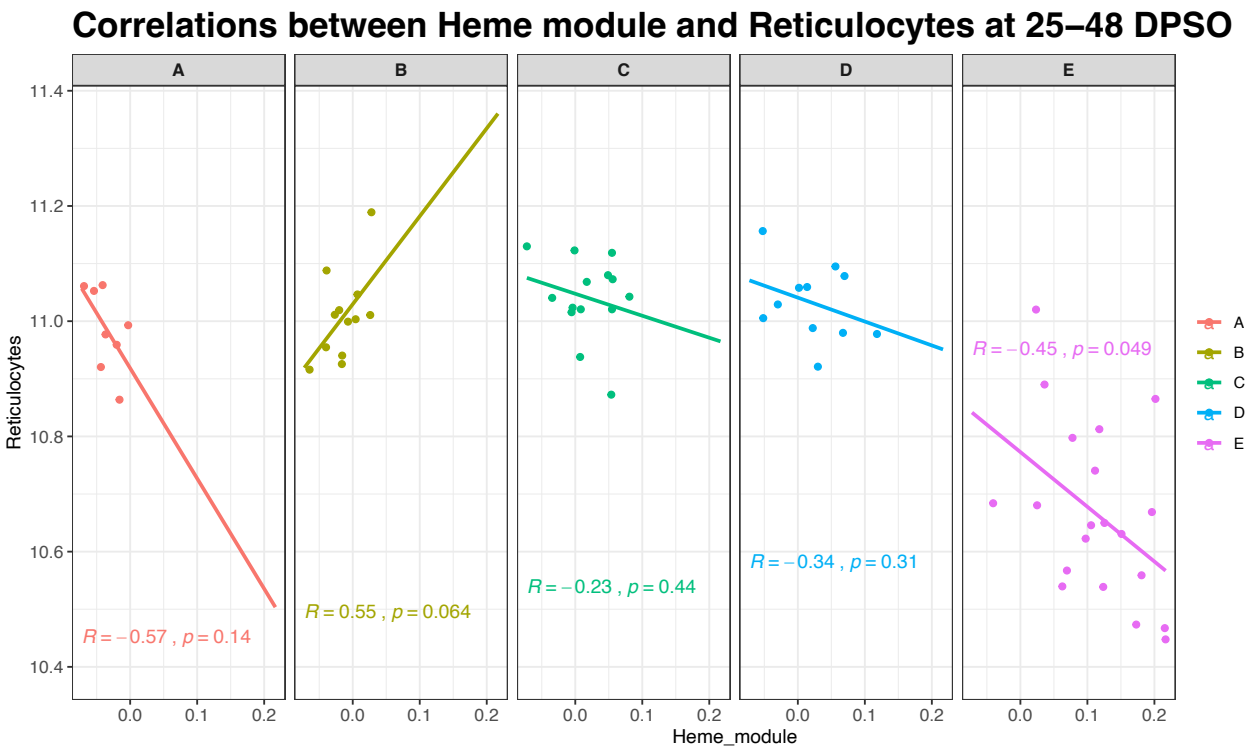


Fig 3.41 Correlation between reticulocyte counts and heme module at 25-48 DPSO.

3.3 Discussion

In this study, we compared the immune responses in 5 cohorts of patients with varying disease severity and tracked immunological changes over time by studying immunophenotyping, transcriptomics, cytokines and antibody responses.

In groups A and B, there was no evidence of systemic inflammation with normal levels of CRP, circulating TNF-a, IL-6 and no enrichment in transcriptional signatures associated with neutrophil activation. This is in stark contrast to groups C-E where marked derangement was noted with elevations in inflammatory markers, cytokines and profound leucopenia.

Transcriptionally, groups C-E clustered apart from healthy controls and enriched in gene signatures associated with TNFa, neutrophil activation and IL-6. Multi-omic factor analysis using data from all timepoints shows that neutrophil activation explained a large proportion of the variance in the transcriptional data and was associated with ongoing severe disease activity. Markers of ongoing disease activity on a cellular level were elevations in CD8 HLA

DR+CD38+ T cells and depression in Vg9Vd2(hi) gd T cells, pDC, and MAIT cells at all time points. An increase in quinolinic acid and kynurenine and decrease in tryptophan were top weighted features associated with disease activity on a metabolomic level. Derangement of tryptophan metabolism and the kynurenine pathway have previously been reported in COVID-19 with levels correlating with IL-6[250].

Interferon was only an early transcriptional marker of disease activity with levels resolving despite ongoing inflammation in patients in group E suggestive of viral clearance and of a different immunopathology driving inflammation. Ribosomal transcriptional pathways were downregulated in severe disease at the early time point and had a negative correlation with interferon stimulated genes expression. This has also been reported in the literature and has been surmised to relate to interferon stimulated genes suppressing protein translation and thus viral replication[169].

Three transcriptional signatures arose late in those with severe COVID-19 and were not present in early severe, nor mild disease. These included activation of oxidative phosphorylation, reactive oxygen species and heme related metabolic pathways. Although these pathways were enriched on GSEA for groups C, D and E at 25-48 days from symptom onset, the leading edge genes illustrated marked increased expression in group E at this late point supported also by WGCNA findings.

Activation of immune cells results in metabolic reprogramming that supports cell growth, proliferation and differentiation. Disruption of metabolic pathways can result in bioenergetic, anabolic, epigenetic or redox cellular crises – culminating in immune dysfunction[251]. It is unlikely that the metabolic signatures observed here simply reflect heightened bioenergetic requirements of activated immune cells, as one would expect that similar requirements are present also at early stages in the disease. The reactive oxygen species transcriptional signature may relate to more abundant production of reactive oxygen species inevitably accompanying increased oxidative phosphorylation. The oxidative phosphorylation and reactive oxygen species gene signatures may be associated with ECMO given at the early time window, 0-24 days from symptom onset, very few patients were on

ECMO whilst at 24-48 days from symptom onset, almost all patients in group E were on ECMO where this finding is most apparent.

Oxidative Phosphorylation is an aerobic form of cellular respiration taking place in the inner membrane of the mitochondria. It is the final step post glyocolsyis, pyruvate oxidation and the citric acid cycle. Electron carriers (NADH and FADH2) are oxidised, passing their electrons down the transport chain. Energy released during this process is used to pump H+ ions into the intermembrane space forming an electrochemical gradient. H+ ions move through the inner mitochondrial membrane via the ATP synthase channel which causes the ATP channel to rotate, catalysing the addition of a phosphate to ADP and thus forming ATP. At the end of the transport chain, electrons are transferred to oxygen which combines with H+ to form water. Reactive oxygen species are formed during oxidative phosphorylation. This occurs secondary to the leakage of electrons from electron transport chains resulting in the partial reduction of oxygen to form superoxide[252]. ECMO results in oxidative stress by multiple mechanisms. Exposure to the extracorporal circuit leading to an increase in IL-1B, TNFa and IL-B. Activation of the coagulation and complement cascade. Platelet and neutrophil activation (occurs in the oxygenator). Hemolysis occurs secondary to the circuit exacerbated by ROS driving damage to the cell membrane and thus altering permeability and deformability leading to lysis. As expected, patients have a large transfusion burden which may contribute to the heme metabolism finding further discussed below. Hyperoxia commonly occurs. The antioxidant ability is overwhelmed during hyperoxia resulting in the production of superoxide. This is especially so when it is preceded by a period of prolonged hypoxia as this causes damage to the mitochondrial electron chain transport. Continuous renal replacement therapy also exacerbates this through loss of anti-oxidants during filtering. Sequestering of antioxidants into the ECMO circuit causing increase in reactive oxygen species[253].

Mitochondria are also critically involved in heme biosynthesis. Heme serves as a prosthetic group for haemoglobin as well as many other proteins – including several that constitute the respiratory chain of mitochondria. While free heme can act as damage-associated molecular pattern and promote reactive oxygen species formation, the role of heme biosynthesis vs. catabolism in balancing cellular sensitivity to oxidants is complex and context dependent[254].

Here, given correlated regulation of heme and oxidative phosphorylation pathways in the clinical categories C, D and E, activity of these modules may be interrelated and possibly jointly reflective of dysfunctional mitochondria. How heme and oxidative phosphorylation transcriptional programmes are linked on a molecular level cannot be inferred from our data. Erythroid cell activation has recently been detected in severe COVID-19[255] and could also contributes to a heme transcriptional signature. However, the increase in heme metabolism in our cohort correlates strongly with a falling haemoglobin, and reticulocytes – suggesting suppression rather than activation of erythropoiesis in these individuals.

Hemophagocytic lymphohistiocytosis (HLH) secondary to SARS-CoV-2 has been reported in the literature. HLH is a life threatening, inflammatory syndrome associated with hypercytokinemia. The cardinal features are high fever, hepatomegaly, splenomegaly, anaemia, thrombocytopenia and neutropenia. Respiratory symptoms as severe as ARDS can occur[256,257]. These features were present in patients in group E whom enriched for the heme module with clinical features of requiring ECMO and laboratory findings of anaemia and thrombocytopenia and ongoing raised CRP. In addition, a positive correlation with D-Dimer, evidence of hypertriglyceridemia and a negatively albeit non-significant correlation was present with fibrinogen suggestive of a coagulopathy. A strong correlation was present at >24 days from symptom onset with inflammatory cytokines. A key feature of secondary HLH is hyperferritenemia with 90–100% of patients having this feature[258]. Hyperferritenemia is due to increased secretion of ferritin by macrophages and/or hepatocytes. A positive but non-significant relationship was present in our data and hyperferritenemia has been widely reported in the literature in severe COVID-19[257].

Many of the abnormalities we have observed in COVID-19 might also be features of other severe viral infections. To identify which are COVID specific will require a comparison with an appropriate disease control group. Continued follow-up of patients will be needed to determine the persistence of abnormalities still observed at late time points. Finally, because our patients were recruited during the first pandemic wave, a follow-up study examining the immune response to new SARS-CoV-2 strains with different virulence could be informative.

An analysis of the effects of hypoxia on B cells in SARS-CoV-2 is discussed in Appendix A.

4. B cell receptor repertoire kinetics after SARS-CoV-2 infection and vaccination

4.1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) resulting in coronavirus disease 2019 (COVID-19) has caused over 4.5 million deaths as of September 2021 (https://covid19.who.int/). It primarily infects respiratory epithelial cells, and results in a range of clinical manifestations from asymptomatic disease to multi-organ failure. B cells play a vital role in anti-viral defence[259]. B cell depletion can result in persisting viremia(Buckland et al., 2020; Kemp et al., 2021a), SARS-CoV-2 neutralizing monoclonal antibodies and convalescent plasma may have a therapeutic role [262,263] and neutralising antibodies may prevent re-infection and transmission[264].

These observations make a strong case for a central role for B cells in the defence against SARS-CoV-2. There is strong evidence that neutralising SARS-CoV-2-specific antibodies can protect against disease onset and progression[193,265,266] and potentially also through non-SARS-CoV-2 specific "natural" antibodies, or antibodies generated in response to other coronaviruses which may also cross-react with SARS-CoV-2[186,267–269]. It is also likely that B cells play a role through other functions, including antigen presentation to T cells, cytokine production and other regulatory mechanisms.

Severe COVID-19 is typified by major perturbations in circulating immune cells[161,172,175,233,234,242,270–272]. Together with other groups, we have shown that COVID-19 has a profound impact on B cell subsets. Increased numbers of recently generated circulating plasmablasts are seen early in disease irrespective of severity, and indeed is one of the few cellular abnormalities observed in asymptomatic SARS-CoV-2 infection[270].

Absolute numbers of almost all other B cell sub populations are reduced, including naive B cells, both switched and unswitched memory B cells, transitional B cells, and more recently,

marginal zone-like B cells. All of these B cell subsets are maximally reduced soon after symptom onset, with most gradually resolving thereafter (with the exception of transitional B cells, which continue to decline over the first two months after infection)[270]. Early histology reports also demonstrated reduced germinal centres in secondary lymphoid organs in COVID-19, and consistent with this, circulating T_{FH}-like cells are markedly reduced[190]. Most initial reports have underestimated the impact of COVID-19 on the B cell immune response, having examined proportions rather than absolute numbers of B cell subsets[270]. Changes between these subsets, as well as within them, will be reflected in the B cell receptor (BCR) repertoire.

The BCR repertoire refers to the range of individual BCRs that collectively provide the diversity of antigen receptors required by B cells to recognise new antigens, to minimise interaction with autoantigens, and, when certain specificities are expanded, to provide increased protection in the context of B cell memory. BCR diversity is driven by the rearrangement of the immunoglobulin receptor genes during B cell development in the bone marrow. During B cell development single variable (V), diversity (D) and joining (J) genes are selected from multiple distinct copies and imprecisely joined to create a BCR [273]. To prevent self-reactivity, B cells go through both central and peripheral tolerance checkpoints[274,275]. Further diversification of the BCR repertoire occurs post antigen exposure through somatic hypermutation (SHM) and subsequent selection of high affinity clones[49,76,92]. B cells may undergo a process termed immunoglobulin class-switching where, through stepwise DNA deletion and recombination of the constant region, downstream isotypes are generated[276]. During this process, the antigen binding region remains the same, and so therefore does antigen affinity but isotype switching confers a range of different effector functions[277]. High-throughput bulk RNA sequencing of BCR heavy chain genes allows us to assess isotype use, SHM, V gene usage and clonality.

The study of the repertoire has been illuminating in immune-mediated disease, infection and vaccination. In previous work, we described increased clonality, IgA proportion and shared IGHV gene usage in Systemic lupus erythematosus and Crohn's disease[278]. Early reports have similarly revealed substantial changes in the BCR repertoire in severe COVID-19. An increased representation of IgG1 and reduced IgM isotypes is seen, as is the over

representation of some specific heavy chain genes (such as the VH3 family). A global reduction in SHM has also been observed when compared with health[172,188,232,237,279,280]. Analysis of SARS-CoV-2-specific B cells has demonstrated some changes consistent with those seen in the global BCR repertoire, in particular low SHM early in disease with a subsequent increase in the memory population[186,187,191,192].

Reduced SHM levels in BCR repertoires have been seen early after Ebola [281], and Dengue infection[282], with the pattern of early low SHM followed by a late increase in SHM-high clones being seen in other infections (for example vesicular stomatitis virus –[283]. This has been attributed to an early extrafollicular response characterised by the initial rapid secretion of lowly mutated antibodies from naive unmutated B cells interacting with cognate T cells and differentiating into short-lived antibody secreting cells; followed by the generation of germinal centres and the production of long-lived plasma and memory B cells.

Increasingly more work is being conducted on the BCR repertoire post SARS-CoV-2 vaccination. Studies show that SARS-CoV-2 mRNA vaccination induces antibodies against NTD, RBD and S2[284] with anti-RBD clones showing high use of IGHV3-30 and IGHV3-53, similar to that seen in natural infection[285]. However, neutralising ability post vaccination appears targeted to the RBD domain with removal of RBD-specific antibodies abolishing neutralization of Wuhan-Hu-1 virus[286]. Vaccine-elicited antibodies appear more broadly distributed across the RBD compared with natural infection potentially preventing loss of efficacy when point mutations occur in the virus [287].

Increasing our understanding of the B cell immune response in the context of COVID-19 is important given its role in defence against SARS-CoV2 infection, and potentially in the prevention of secondary infection, re-infection and autoimmunity. We have some understanding of this early after SARS-CoV2 infection: little is known about how the BCR repertoire changes over time, varies with disease severity, or compares to that generated by vaccination. Studying the global BCR repertoire allows us to not only study antigen specific B cells but also "bystander" viral-associated clones that are often mobilised post infection and vaccination[288]. We have analysed the BCR repertoire in a large cohort of patients with varying disease severity, sampling at several timepoints to six months post symptom onset

and comparing these to BCR repertoire changes following vaccination with the BNT162B2 SARS-CoV-2 vaccine[289], and the Trivalent Influenza Vaccine (specific for influenza A (H3N2), A (H1N1) and B). Exploring how the BCR repertoire post mRNA SARS-CoV-2 vaccine compares with that seen in natural infection and post influenza vaccination allows us to gain insight into the nature of the vaccine response. A decrease in SHM may suggest a prominent extrafollicular response, as commonly seen with polysaccharide vaccines[290], leading to the generation of low-affinity and short lived plasmablasts. Alternatively, an increase in SHM is suggestive of a germinal centre response as seen post influenza vaccination[291] and is indicative of the generation of affinity matured, long lived plasma cells. In addition, the route that an antigen enters the body determines the class-switching patterns and thus systemic versus mucosal routes as seen in vaccination and natural infection may differentially influence the repertoire.

4.2 Results

4.2.1 Patient cohort

SARS-CoV-2 PCR-positive subjects (n = 171) were recruited between 31st March and 20th July 2020 and divided into five categories according to peak clinical severity.

A) asymptomatic healthcare workers (HCWs) recruited from routine screening.
B) HCWs either still working with mild symptoms, or symptomatic and self-isolating.
C) patients who presented to hospital but never required oxygen supplementation.
D) admitted patients whose maximal respiratory support was supplemental oxygen.
E) patients who required assisted ventilation (56) or died without ventilation (3).

Patients were bled weekly while inpatients, and less frequently thereafter. Patient time courses are measured since symptom onset for groups B to E, and from the first positive swab for group A (not having symptoms to trigger presentation, patients in group A were likely sampled, on average, later after infection than B-E). Recipients (n = 63) of the BNT162B2 (Pfizer/BioNTech) SARS-CoV-2 vaccine were bled post initial dose and before boosting. Recipients (n = 14) of the Trivalent Influenza Vaccine (TIV) were bled before

vaccination, and then at 7- and 30-days post. Heathy controls were recruited across a range of ages and included the TIV recipients prior to vaccination (Fig 4.1).

	Days	0-25	26-50	51-100	101-200
Healthy Donors	HC	74			
Screening asymptomatic	A	26 (17)	8		7
Screening symptomatic	B	48 (34)	14 (13)		20
Hospital: room air	C	36 (33)	14	3	17
Hospital: supplemental 02	D	37 (30)	10	2	9
Hospital: assisted vent	E	29 (24)	46 (38)	19 (16)	13
Vaccinated: SARS-C0V-2	VC	11	21	31	
Vaccinated: INFLUENZA	VI	14	9		

Fig 4.1 Study participants. Study participant and sample numbers split by severity categories and time bins post screening (cat. A), symptom onset (cat. B-E) or vaccination (cat. VC and VI). The number of samples are listed along with individuals in brackets where different.

For groups A-D, patients were well represented in the first 50 days from symptom onset/swab positive and then from 100 days onwards. Group E were well represented in the first 75 days from symptom onset and then from 125 days onwards (Fig 4.2).

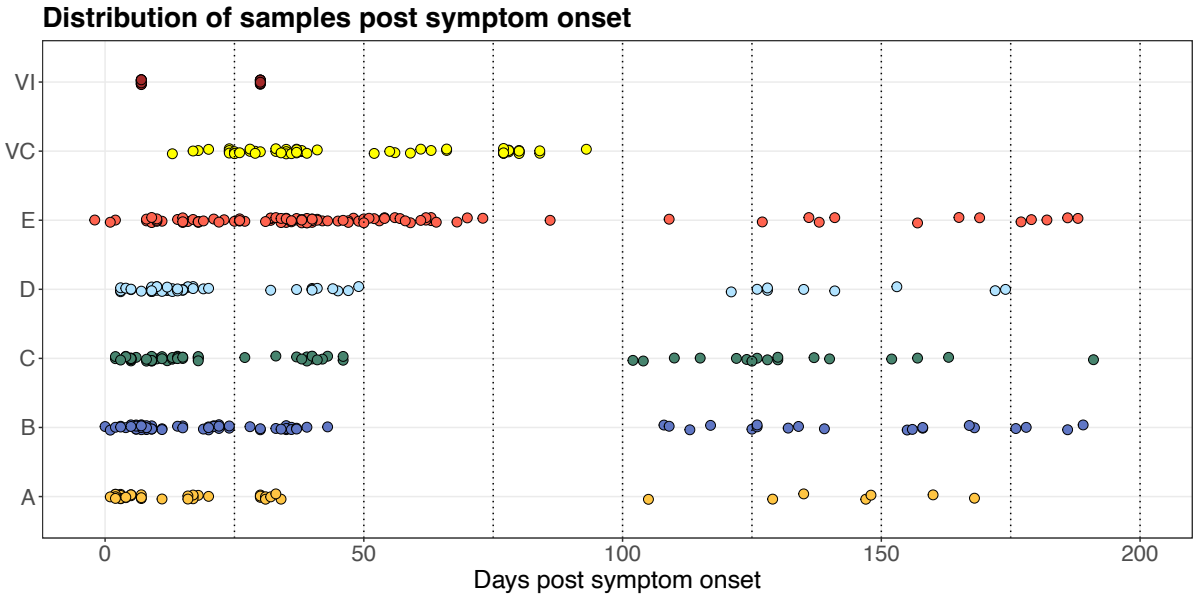


Fig 4.2 Sample distribution. Sample distribution according to days from symptom onset/swab or vaccination split according to disease severity group. Circles represent individual donors.

A skewing in age distribution was apparent in patients infected with SARS-CoV-2 with an older distribution in the hospitalised group. Similarly, an older age distribution was also notable in the SARS-CoV-2 vaccinated group, representative of the prioritisation of the elderly in vaccinating at the time of the study. A marked skewing in sex was also apparent with a great proportion of patients in group D and E being male (Fig 4.3).

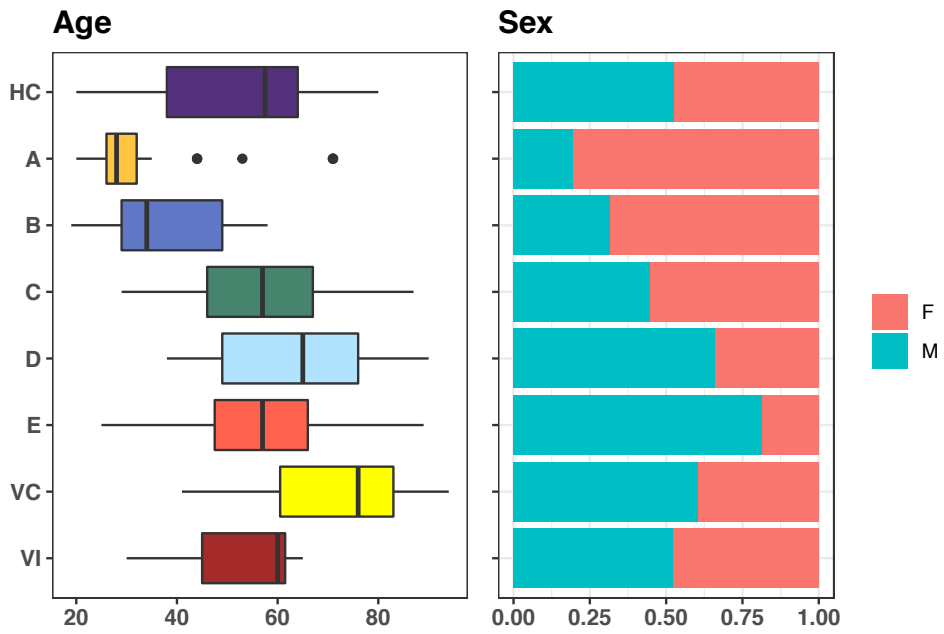


Fig 4.3 Distribution of participant across age, gender and severity categories.

To account for the potential effect of ageing on the BCR repertoire, all comparisons to health used age-matched controls. This was achieved by examining the age distribution of participants after grouping according to both disease severity/vaccination and time and then randomly selecting a subset of healthy controls to mirror this distribution for each time window and severity group (Fig 4.4).

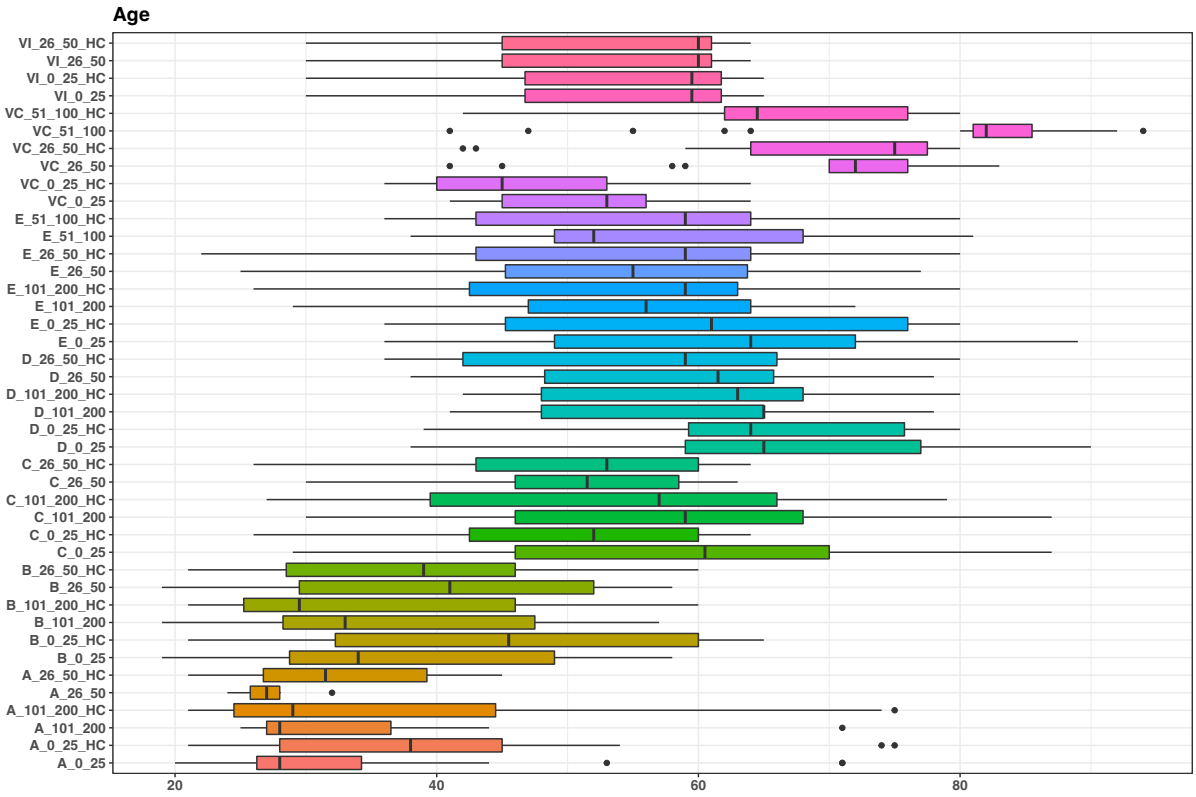


Fig 4.4 Age matched healthy controls. Age matched healthy controls used in analysis per time-window per disease group.

4.2.2 BCR repertoire reproducibility

Only a small portion of a person's BCR repertoire is sampled. Using 200ng of start RNA from blood, we generated libraries ranging from 5000-20000 reads per person.

In a recent study that deeply sequenced three individuals, starting from 13-30 billion PBMCs per person on average 5.5 X 10^8 BCR sequences (12.5 X 10^6 unique clones) were generated per person. Even at this depth, the rarefaction curves did not plateau indicating that not all clonotypes had been identified but was instead estimated at 50-60% coverage (Soto et al., 2019). Given BCR repertoire analysis is sensitive to sampling depth and does not fully capture the entirety of a patients' repertoire, we analysed the BCR generated from biological replicates to confirm the reproducibility of our data at the library depth we were generating. We compared BCR repertoire metrics from bleeds taken from patients on the same day with the libraries generated from both PBMC and whole blood (from Paxgene tubes). Fig 4.5 shows a strong correlation between isotype proportions generated from PBMCs versus whole blood. Only IGHG4 showed a weak correlation which is likely due to

the very few reads captured from this isotype compared with all other isotypes (proportion of total repertoire: 0.005-0.03). We similarly demonstrated a strong correlation between diversity metrics and SHM between PBMCs and Paxgenes (Fig 4.6 and Fig 4.7).

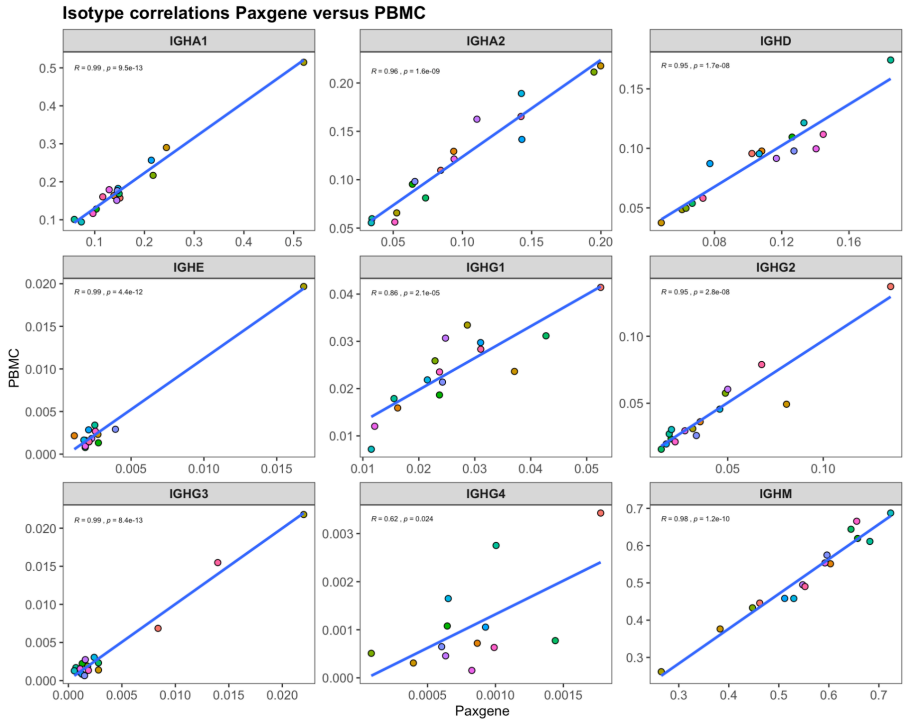


Fig 4.5 Isotype BCR repertoire reproducibility. Correlation between isotype proportions between paired samples with BCR repertoire generated from PBMC versus whole blood (Paxgene).

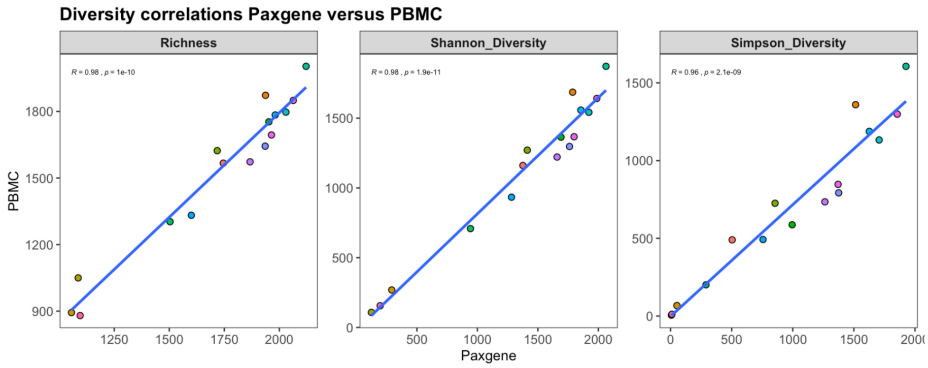


Fig 4.6 Diversity BCR repertoire reproducibility. Correlation between diversity metrics between paired samples with BCR repertoire processed from PBMC versus whole blood.

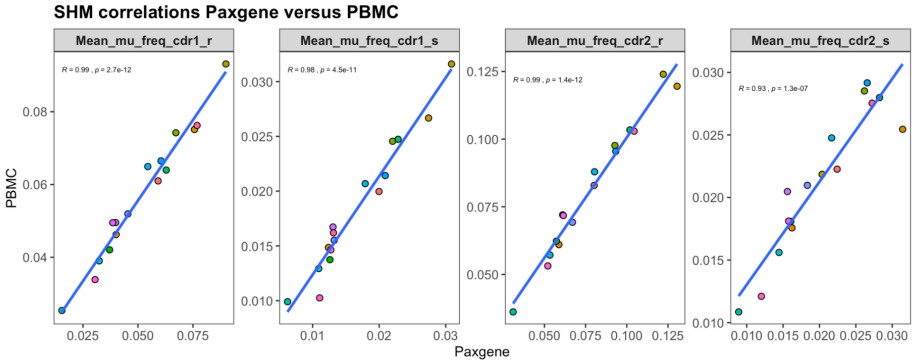


Fig 4.7 SHM BCR repertoire reproducibility. Correlation between SHM metrics between paired samples with BCR repertoire processed from PBMC versus whole blood.

Lastly, we use hierarchical clustering to groups samples according to CDR3 amino-acid sequence. This illustrated close clustering of samples from the same individual (Fig 4.8).

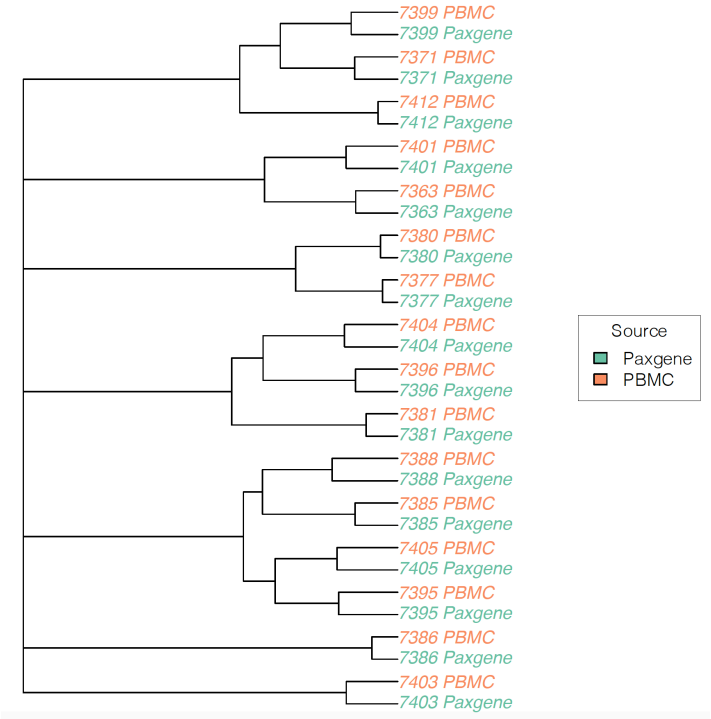


Fig 4.8 Hierarchical clustering of samples according to CDR3 amino-acid region.

4.2.3 B cell composition

To understand how compositional changes in B cell subsets might influence the global BCR repertoire, we compared B cell proportions in patients within 25 days from symptom onset to healthy controls (Fig 1C). The proportion of plasmablasts were increased in all severity

groups and in addition marginal zone, transitional and memory B cells were decreased in group E (Fig 4.9).

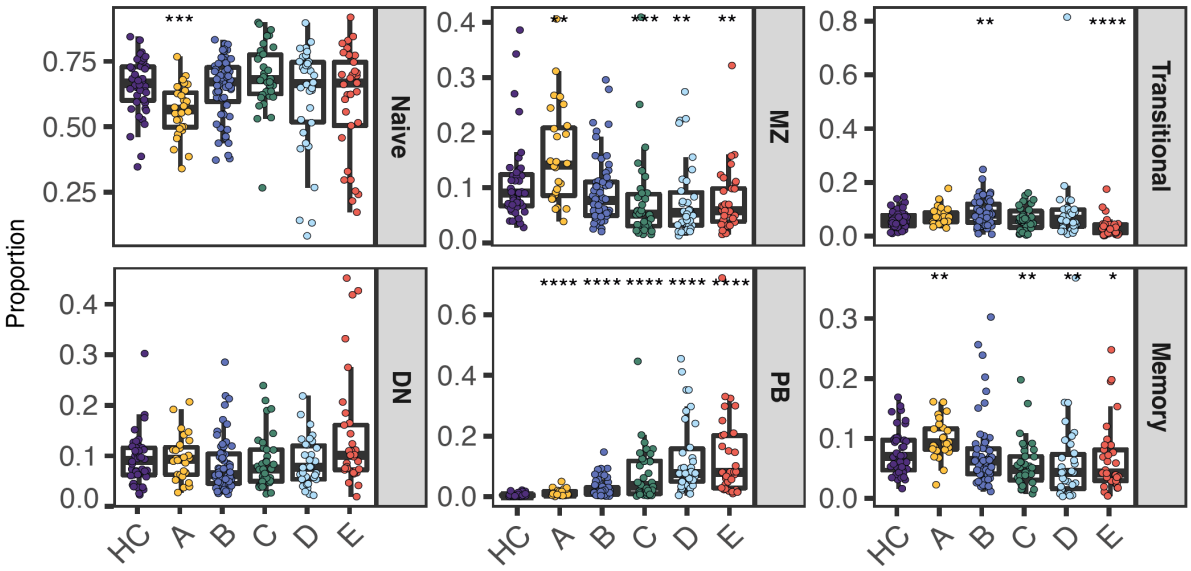


Fig 4.9 B cell subsets. *Boxplots showing B cell subset proportions according to disease severity at 0-25 days from symptom onset. Naïve (CD19+IgD+CD27-), Marginal Zone B cells (MZ) (CD19+IgD+CD27+), Double negative B cells (DN) (CD19+IgD-CD27-), Transitional (CD19+IgD+ CD27+CD24+CD38+), Memory (CD19+IgD-CD27+CD24+CD38+) and Plasmablasts (PB)(CD19-CD20-CD27+CD24+CD27+CD38+). Comparison with HC, unadjusted wilcox test p-value: *<0.05, **<0.005, ***<0.0005.*

To further illustrate the change in B cell composition with infection, we calculated the proportion of plasmablasts, memory, marginal zone, double negative, transitional and naïve B cells per person in the healthy control group and group E. We then summed up the proportions before recalculating the proportion. This is illustrated in the pie charts. This highlights the major proportional increase in plasmablasts and double negative B cells (Fig 4.10).

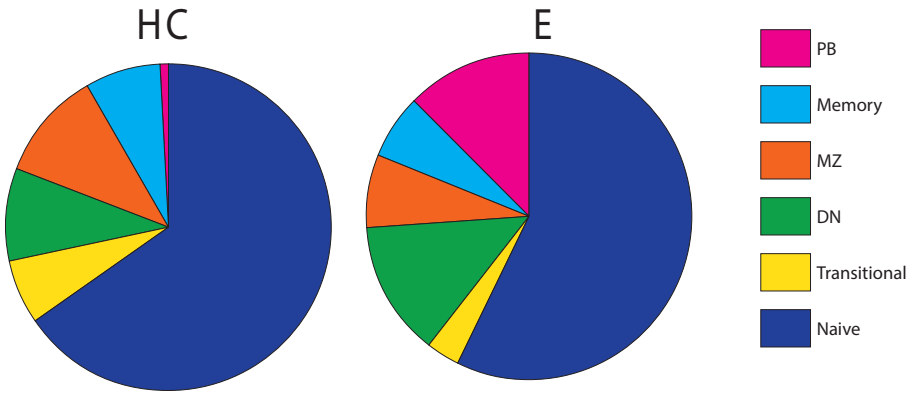


Fig 4.10 B cell proportions. *Pie chart comparing B cell proportions between HC and group E at 0-25 days from symptom onset.*

4.2.4 Isotype use

We assessed the proportion of unique B cell clones of different isotypes, counting each unique VDJ region only once to ensure results were not skewed by the differential mRNA content of B cell subsets (in particular plasmablasts which have increased immunoglobulin mRNA content). IGHG1 and IGHG3 proportions were increased across all severity groups, and were the only isotypes increased in the asymptomatic group A. IGHA1 was increased in a similar pattern, although changes were less pronounced. IGHA2 was only elevated in group E. Serum IGHA2 is more pro-inflammatory than IGHA1 with an increased ability to induce NET formation and the release of cytokines by neutrophils and macrophages[292]. This is keeping with our previous finding of increased neutrophil activation in group E [270]. The increase in IGHE seen in the hospitalised groups C, D and E is most reflective of an increase in IGHE plasmablasts given IGHE memory B cells are a transient cell type in the germinal centre response and most IGHE plasmablasts are derived from class switching IgG1 memory cells[293]. IGHE antibodies are known to be generated in other respiratory illness such as influenza A[294]. IGHD and IGHM were reduced, particularly in those with severe disease. All isotype proportions returned to normal over time, with recovery being delayed in more severe groups (Fig 4.11).

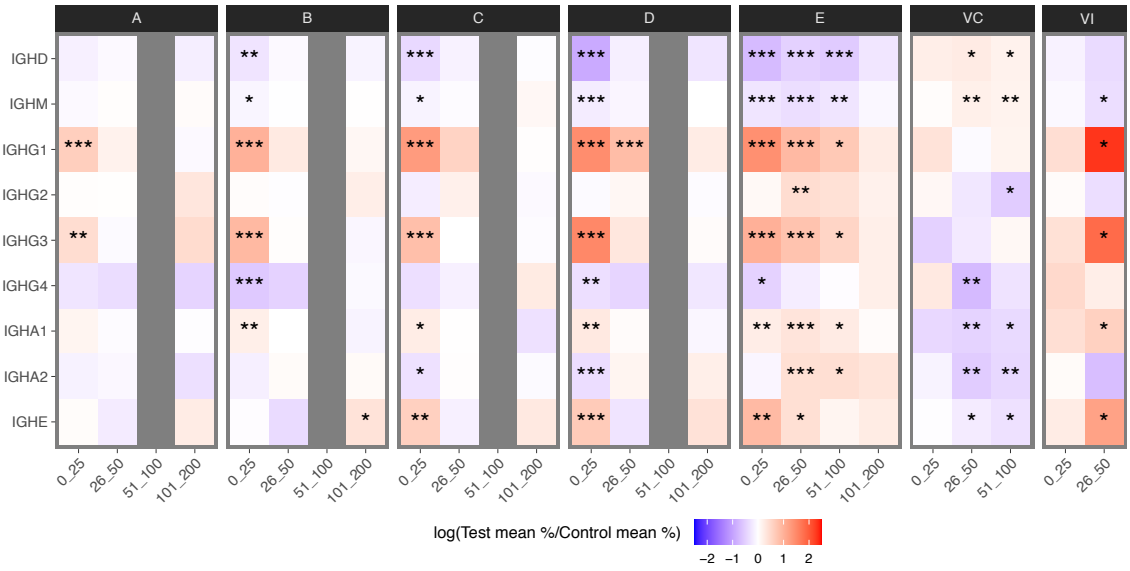


*Fig 4.11 Isotype Usage. Heatmap showing log2 fold change in mean proportion between SARS-CoV-2 and vaccine cases and HC, within severity categories and across 25-day time bins. Wilcoxon test FDR adjusted p-value: *<0.05, **<0.005, ***<0.0005.*

These changes in isotype proportion were illustrated assessing time as a continuous variable using a linear mixed effects model analysis (Fig 4.12). This once again highlights the decrease in proportion of IGHD and IGHM likely mirroring the decrease in naïve B cells with a concurrent increase in IGHG1 (Fig 4.12).

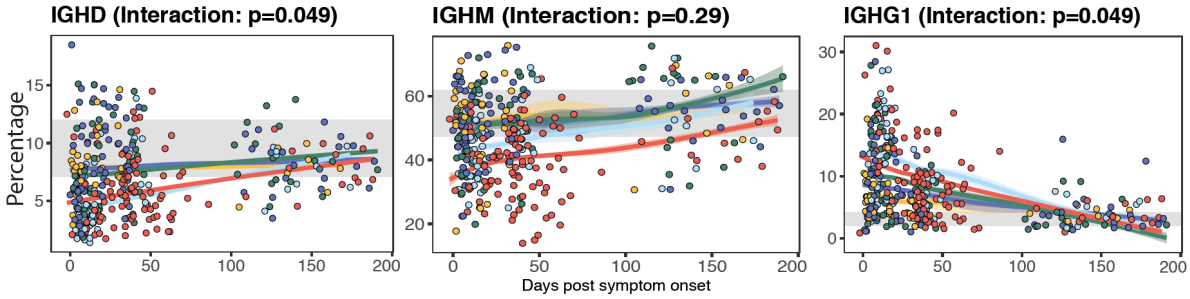


Fig 4.12 Linear mixed-effects model of isotype usage. Linear mixed-effects model showing the longitudinal expression of IGHD, IGHM and IGHG1 proportions over time, grouped by severity. Grey band indicates the interquartile range of the corresponding isotype in HCs. Nominal p-values for the time x severity group interaction term are reported.

Isotype changes in response to BNT162B2 SARS-CoV-2 vaccine were very different to those seen in SARS-CoV-2 infection. Increase in IGHD and M isotype proportions were apparent only after 25 days from vaccination with concurrent decreases in IGHG2/4, IGHA1/2 and IGHE (Fig 4.11). Similarly, isotype changes were only seen in response to the TIV beyond 25 days after vaccination. The prominent increase in IGHG1, IGHG3 and IGHA1 proportions mirrored that of SARS-CoV-2 infection (Fig 4.11). IGHG1 and IGHG3 are the key antibodies formed post viral infection[295].

Correlation of BCR isotype use derived from BCR repertoire sequencing with B cell subset numbers and with serum immunoglobulin titres was performed (Fig 4.13).
The strongest positive correlations were seen between IGHG1, IGHG3 and IGHA1 and plasmablast numbers and the strongest negative correlation between IGHM levels and plasmablast numbers, suggesting the increased proportion of these switched isotypes was, in large part, driven by an increase in clonally distinct plasmablasts (Fig 4.9 and 4.10). Consistent with this was the increased IGHG1 and IGHG3 proportions seen in group A, in which an early rise in plasmablasts is the only prominent change in B cell subpopulations seen (Bergamaschi et al., 2021). IGHD and IGHM correlate strongly with naive B cell number,

suggesting that their decline is in part a reflection of reduced naive and transitional B cell numbers in moderate to severe COVID-19. Correlation between IGHA1 isotype use and serum IgA was seen, but no such correlation was seen between the IgG isotypes and serum IgG (Fig 4.13). The lack of correlation is likely reflective of the extended half-life of IgG compared with other immunoglobulins. IgG takes time to build up in serum in the immune responses, lagging behind the cellular response, and then has a serum half-life of 21 days, so will persist after cellular resolution begins. In addition, antibody titres and isotype BCR proportions may not always correlate. “Steady state” serum immunoglobulin is made predominantly by long-lived plasma cells in the bone marrow. In the acute setting its rapid increase is driven largely by extrafollicular plasmablasts. In contrast, immunoglobulin transcripts in the blood, measured in our repertoire analysis, will be derived from not only plasmablasts in transit but also non-antibody secreting cells such as memory B cells. Thus, levels may not always correlate.

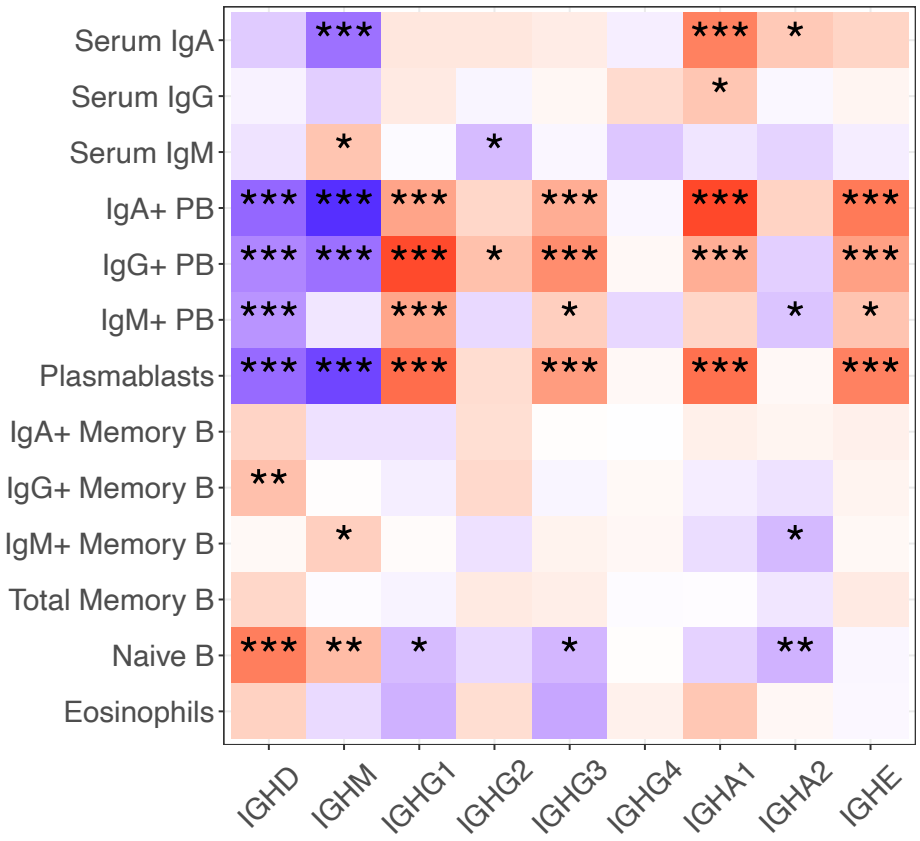


Fig 4.13 Correlation between BCR isotype proportions and B cell metrics. ***Heatmap depicting correlation between cell number, serum immunoglobulins and BCR isotypes at 0-25 days from symptom onset/swab. p-value: *<0.05, **<0.005, ***<0.0005.***

4.2.5 Class switching

We quantified the level of switching between classes in different disease subsets, by assessing the frequency of unique VDJ regions that shared two different isotypes, having corrected for read depth by subsampling. This demonstrated increased switching to IGHG1 and IGHA1 in all severity groups and post influenza vaccination in the first 25 days after symptom onset/swab positivity/vaccination (Fig 4.14).

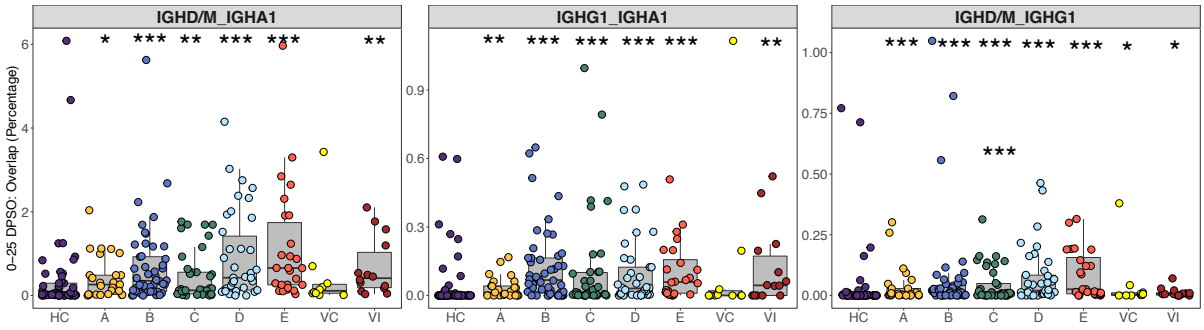


Fig 4.14 Class-switching at 0-25 days from symptom onset. Boxplots showing class-switching per patient in the first 0-25 days from symptom onset, swab or vaccination split according to severity. Circles represent individual donors.

Beyond 25 days, increased switching was prominent only to IGHA1, and predominantly in those with more severe disease (Fig 4.15). This is unlikely to be due to persisting virus, as clearance occurs within the first 25 days as measured by nasal/throat swab but rather could be associated with ongoing inflammation in group E evidenced by continued elevation in CRP(Bergamaschi et al., 2021a). In keeping with this, group E also demonstrated an ongoing class switching to IGHG1. Both vaccine groups demonstrated an increase in class switching to IGHA1 and IGHG1.

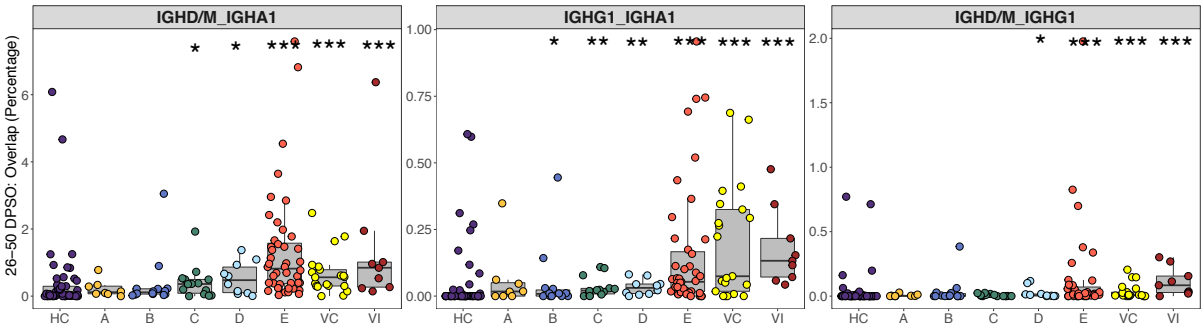


Fig 4.15 Class-switching at 26-50 days from symptom onset. Boxplots showing class-switching per patient at 26-50 days from symptom onset, swab or vaccination split according to severity. Circles represent individual donors.

4.2.6 Somatic hypermutation

SHM is the mechanism by which the BCR repertoire is diversified during the GC reaction, with the subsequent selection of high affinity mutants resulting in “affinity maturation”, and potentially also in an increased breadth of the memory B cell repertoire(Smith et al., 1997; Tonegawa, 1983; Victora and Nussenzweig, 2012). Reduced SHM has been seen in SARS-CoV-2 infection by others (Galson et al., 2020; Kreer et al., 2020; Kuri-Cervantes et al., 2020; Nielsen et al., 2020; Schultheiß et al., 2020; Seydoux et al., 2020), and this is confirmed in our cohort (Fig 4.16). Reduced SHM is most pronounced in IGHG1, IGHG3, IGHA1, and to a lesser extent IGHA2 and IGHE. This is most prominent early after symptom onset, occurs across all severity groups, and recovers over time. SHM is reduced in the isotypes most increased in the BCR repertoire, suggesting that most expansion occurs outside the GC.

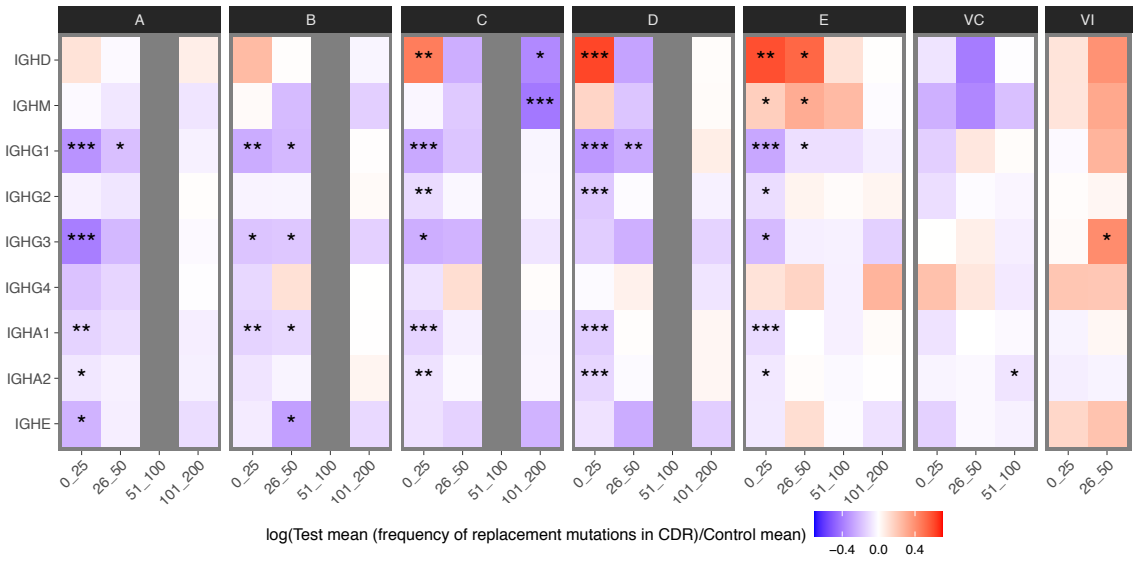


*Fig 4.16 Somatic Hypermutation. Heatmap showing the log2 fold change in mean frequency of replacement mutations covering regions CDR1 and CDR2 between SARS-CoV-2 and vaccine cases and HC, within severity categories and across time bins post screening (cat. A), symptom onset (cat. B-E) or vaccination (cat. VC and VI). Wilcoxon test FDR adjusted p-value: *<0.05, **<0.005, ***<0.0005.*

A mixed effects model of SHM illustrates the overall lower levels of SHM in IGHD and IGHM compared with IGHA1 and IGHG1 as expected. Modelling with time as continuum, IGHD and IGHM show an initial increase in SHM which normalises whilst IGHA1 and IGHG1 show the reverse with an initial decrease which normalises (Fig 4.17).

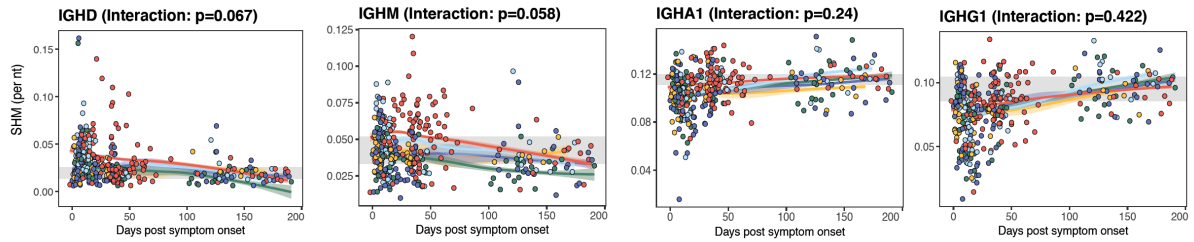


Fig 4.17 Linear mixed-effects model of SHM. Linear mixed-effects model showing the longitudinal levels of SHM over time, grouped by severity and isotype. Grey band indicates the interquartile range of the corresponding isotype in HCs. Nominal p-values for the time x severity group interaction term are reported.

When the kinetics of SHM reduction are considered in more detail, SHM reaches its nadir between 11 and 20 days after symptom onset in most groups (Fig 4.18).

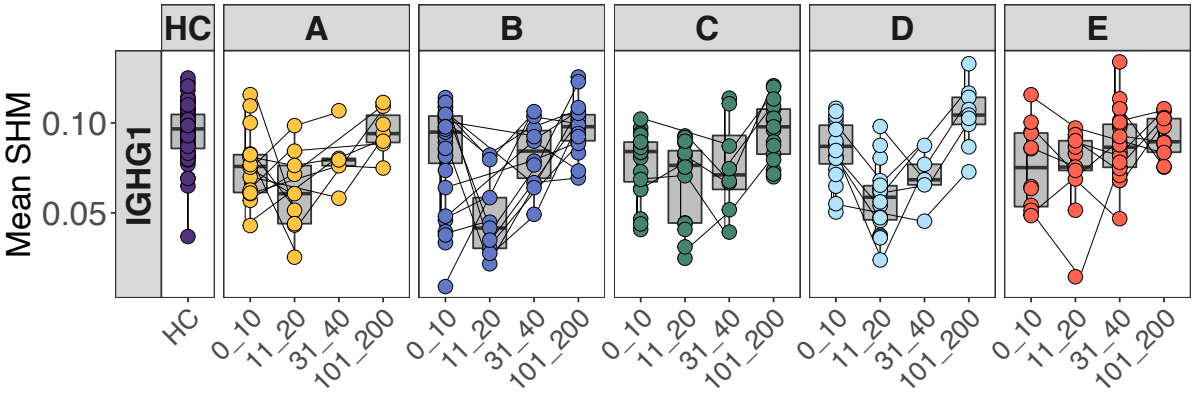


Fig 4.18 IGHG1 SHM. Boxplots showing the mean frequency of replacement mutations covering regions CDR1 and CDR2 in IGHG1 split by severity categories time bins. Circles represent individual donors. Lines connect matching patients.

A density plot of SHM split according to HC, disease severity and days from symptom onset, highlights the bi-modal distribution of SHM marked at 11-20 days from symptom onset which resolves out at 100-200 days from symptom onset (Fig 4.19 and Fig 4.20).

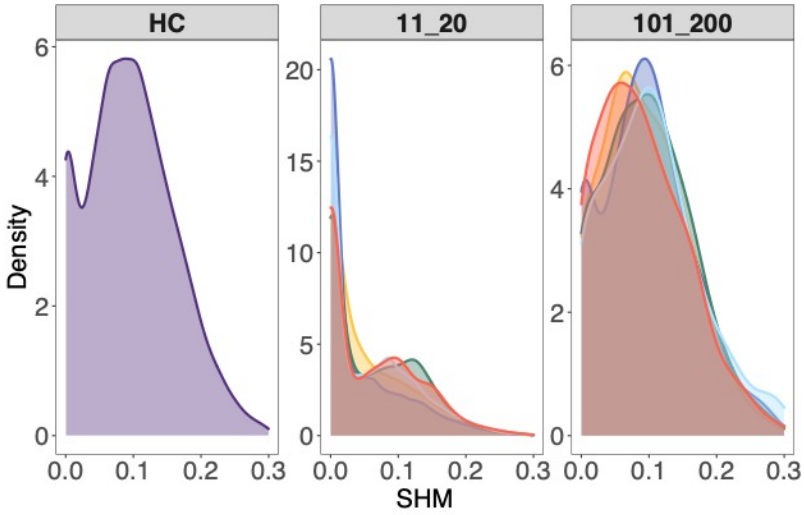


Fig 4.19 Density plot modelling IGHG1 SHM. Density plot modelling IGHG1 SHM across HC, infection and vaccination at 11-20 and 101-200 days from symptom onset.

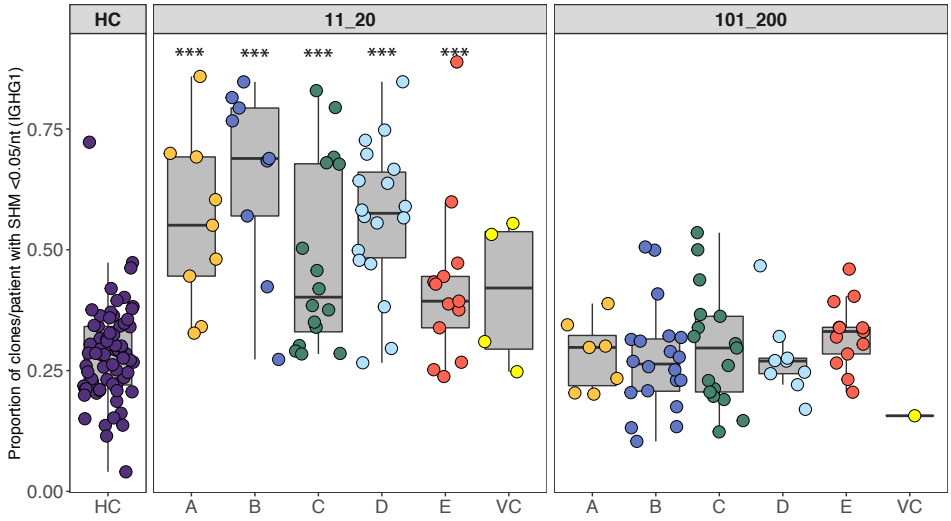


Fig 4.20 Distribution of SHM. Boxplots showing the proportion of clones per patient with a mean level of SHM <0.05nt across HC and infection and vaccination at 11-20 and 101-200 days from symptom onset.

Reduced SHM could reflect the relative increase in the proportion of unmutated B cell clones, highest in early time points compared to late ones (Fig 4.19), although less pronounced in group E (Fig 4.20).

In contrast, a marked increase in SHM in IGHD and to a lesser extent IGHM, is present in those with moderate to severe COVID-19 (Fig 4.16 and Fig 4.17). This may be reflective of a cellular compositional change in IgM and IgD positive cells. When compared with HC, severe

COVID-19 patients have a proportional increase in IgM+ plasmablasts and memory B cells which are expected to have a higher mutational load compared with naïve B cells (Fig 4.21).

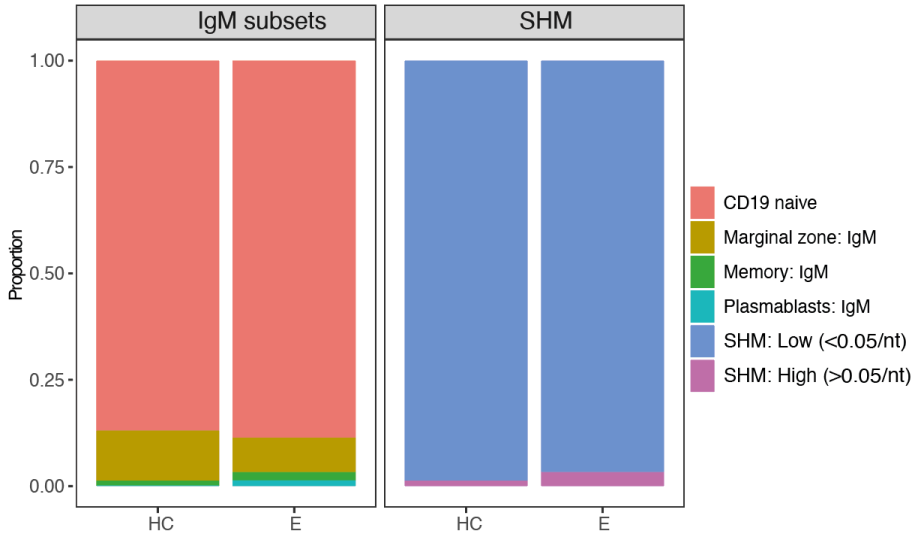


Fig 4.21 IgM+ cells according disease status. Density plot comparing cellular IgM+ cells proportions in HC and group E within 25 days from symptom onset.

This was confirmed with levels of SHM in IGHM/D clones having a negative correlation with CD19 naïve B cell numbers and a positive correlation with IgM memory and plasmablast proportions (Fig 4.22).

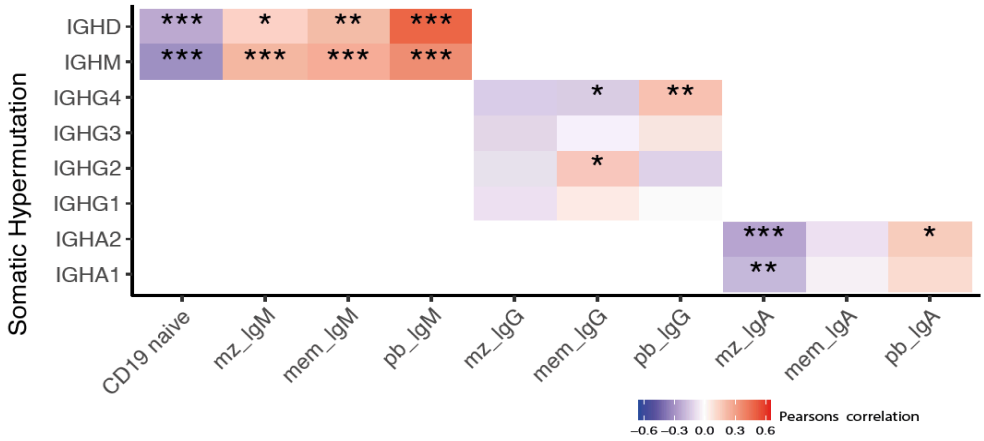


*Fig 4.22 Correlation between SHM and B cell subset proportions. Heatmap of correlations between isotype somatic hypermutations and B cell subset proportions within 25 days from symptom onset. p-value: *<0.05, **<0.005, ***<0.0005.*

To determine if SHM differed between expanded and unexpanded clones, we defined clones as expanded if they were ≥0.5% of the total repertoire and focused on groups C, D

and E within 25 days from symptom onset. There was also a comparative increase in SHM in expanded compared to unexpanded clones (Fig 4.23) consistent with generation in the germinal centre.

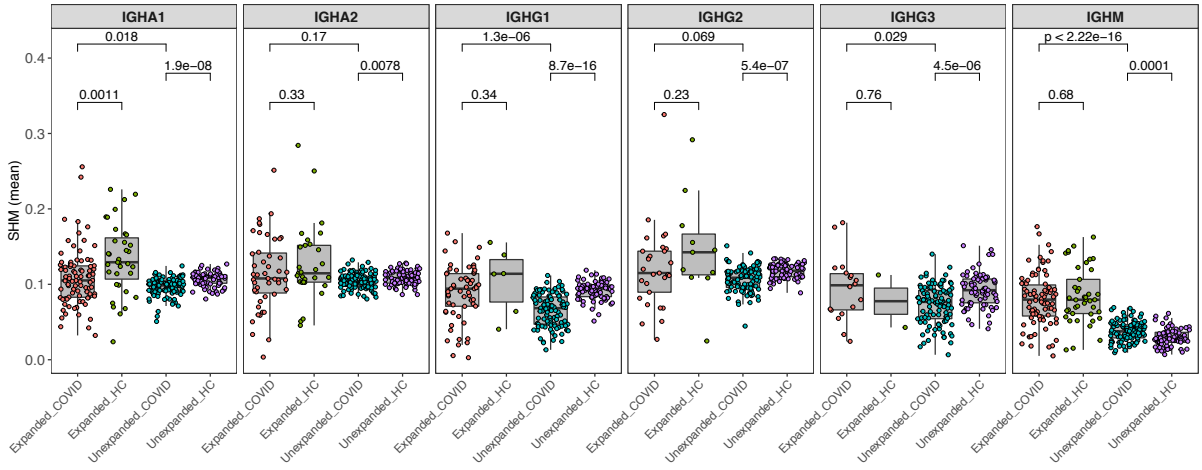


Fig 4.23 SHM in the first 25 days from symptom onset. SHM in the first 25 days from symptom onset in groups C, D and E compared with HC split according to isotype and expansion, defined as ≥0.5%.

SARS-CoV-2 vaccination did not appreciably alter SHM whilst Influenza vaccination showed an increase at 25-50 days in IGHG3 (Fig 4.16). Although not significant, IGHD and IGHM showed an increase in SHM post Influenza vaccination compared with healthy controls, similar to that seen in groups C, D and E.

Finally, the acquisition of anti-SARS-CoV-2 spike IgG antibodies was temporally associated with reduced global IGHG1 SHM and appeared independent of the time post symptom onset (Fig 4.24). Splitting samples into 10 day time windows according to IgG spike status consistently showed a decrease in SHM in the IgG spike seropositive group compared with the seronegative group within a given time window. Fig 4.25 is a visual representation of SHM, illustrating a decrease in SHM when going from seronegative status to seropositive status.

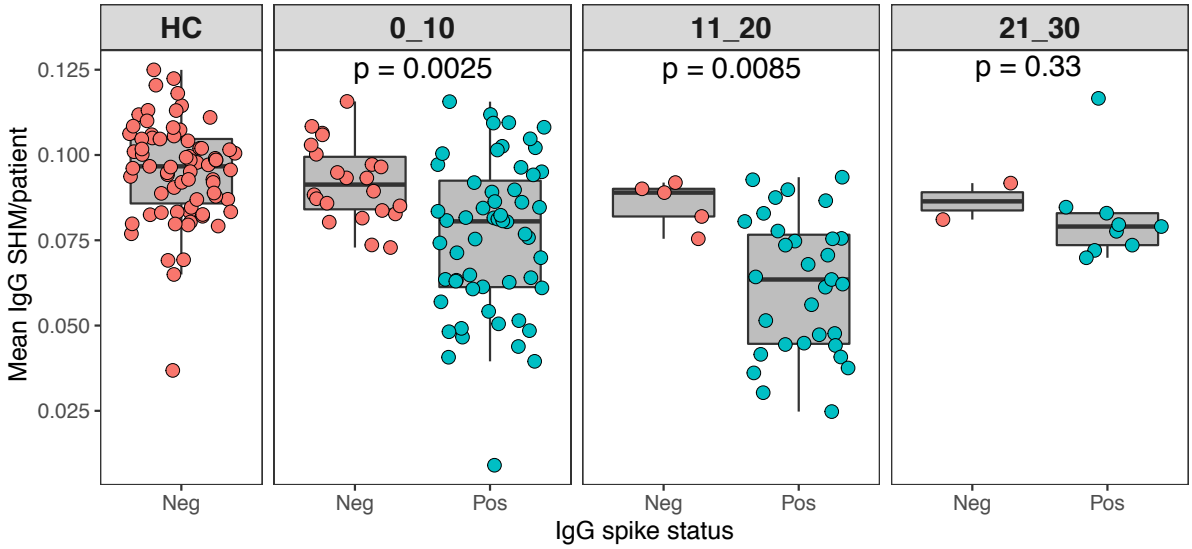


Fig 4.24 IgG SHM according to disease status. Boxplots showing mean IGHG1 SHM per patient split according to days from symptom onset/swab and IgG spike serostatus. Circles represent individual donors.

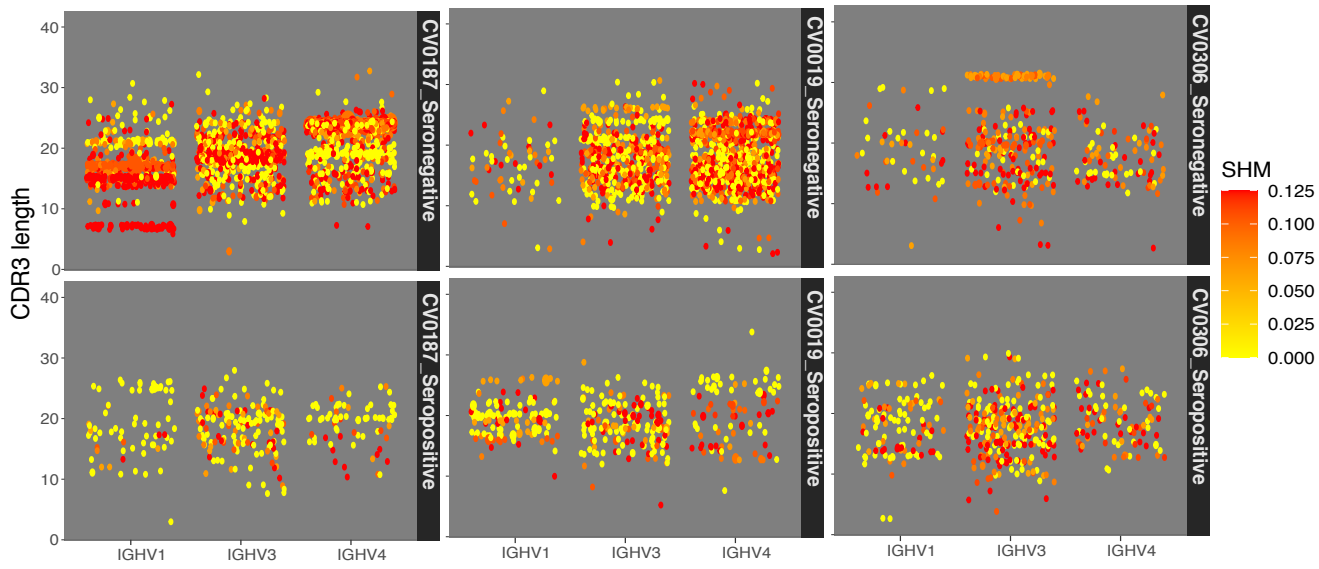


Fig 4.25 IgG SHM according to serostatus. Boxplots showing mean IGHG1 SHM per paired patient pre and post seroconversion. Visual representation of change in somatic hypermutation post seroconversion. Points represent B cell clones. Top row represents seronegative patients. Bottom row represents paired patient post seroconversion. IGHV gene is represented on the x axis, CDR3 length on the y axis and point colour represents level of SHM.

This observation also held true for serum neutralisation activity which showed a decrease in SHM upon positive neutralising activity (Fig 4.26).

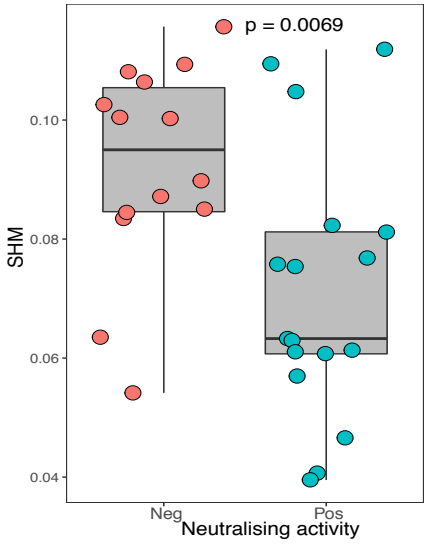


Fig 4.26 IgG SHM according to neutralising status. Boxplots showing mean SHM per patient split according to neutralising activity. Circles represent individual donors.

Increased switching to IgG1 with seroconversion was also seen within 25 days from symptom onset. When further split according to disease severity, this increase in switching was driven by groups C, D and E (Fig 4.27). These observations are consistent with recent evidence suggesting that the early neutralising anti-spike SARS-CoV-2 antibody response is not mutated [188,191,280] – with this antigen-specific observation reflected in the BCR repertoire as a whole.

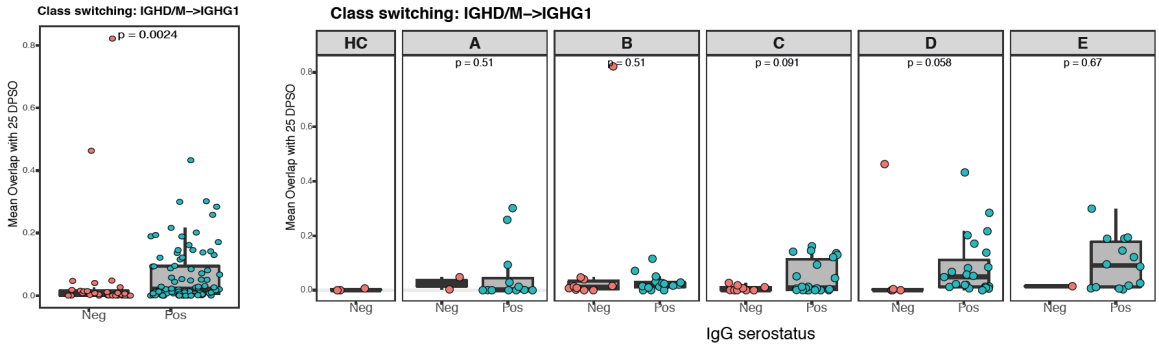


Fig 4.27 Class-switching between IGHD/M and IGHG1. Level of class switching from IGHD/M to IGHG1 at 25 days from symptom onset initially grouped and then split according to disease category

4.2.7 Clonal expansion

We next assessed clonal diversity and expansion, using a number of standard measures, after sub-sampling to correct for varying library depth. These measurements were the

repertoire richness, Simpson's, Shannon's and D50 indices. "Richness" refers to the abundance of unique clones in a repertoire (Chao1). The inverse Simpson's index assesses the probability of two randomly sampled reads belonging to the same clone, the more expanded clones within the population, the greater the chance of clonal sharing. The D50 index refers to the number of unique CDR3 sequences that are present in the top 50% of sequences. A small D50 index is suggestive of large dominant clones. Shannon's index is a measure of "evenness", whereby the proportion of total reads represented by each clone is assessed. This metric is not influenced by the number of unique of clones but rather the distribution of size of clones in the repertoire. Thus, a decrease in BCR repertoire diversity, corresponding to an increase in expanded clones, will usually be reflected in a decrease in all four indices.

There were no changes in BCR repertoire diversity in groups A, B and C. In contrast, there was a profound reduction in diversity in groups D and E. In both groups this was most pronounced in the first 25 days, but in group E the reduction persisted out to 100 days. By 200 days, diversity had been restored in all severity groups (Fig 4.28 and Fig 4.29). The persistence of reduced clonality in group E is most likely a product of severe disease and is associated with the persistence of SARS-CoV-2 specific clones (discussed below). It is unlikely to be driven by ongoing overt infection, as in severe disease viral clearance with broadly similar kinetics to milder disease is the rule in most patients[270]. More likely increased initial viral load[297] makes more antigen available on follicular dendritic cells which, in the context of ongoing systemic inflammation, results in a prolonged GC reaction and thus increased clonal expansion. The secondary infections commonly seen in the ICU setting would also contribute to increased B cell clonality. After a single dose of SARS-CoV-2 vaccine a decrease in diversity was similarly observed but was delayed until after 26 days after vaccination and persisted out to 100 days. Post Influenza vaccination, a similar trend was noted although not statistically significant (Fig 4.28 and Fig 4.29).

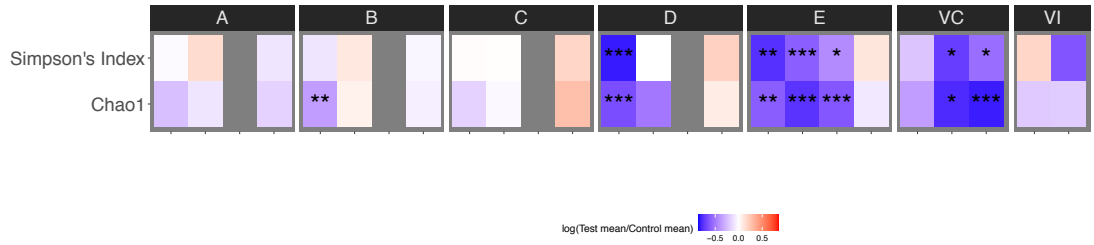


*Fig 4.28 Simpson's index and Chao1 diversity metrics. Heatmap showing log2 fold change in mean diversity indices between SARS-CoV-2 and vaccine cases and HC, within severity categories and across time bins post screening (cat. A), symptom onset (cat. B-E) or vaccination (cat. VC and VI). Wilcoxon test FDR adjusted p-value: *<0.05, **<0.005, ***<0.0005.*

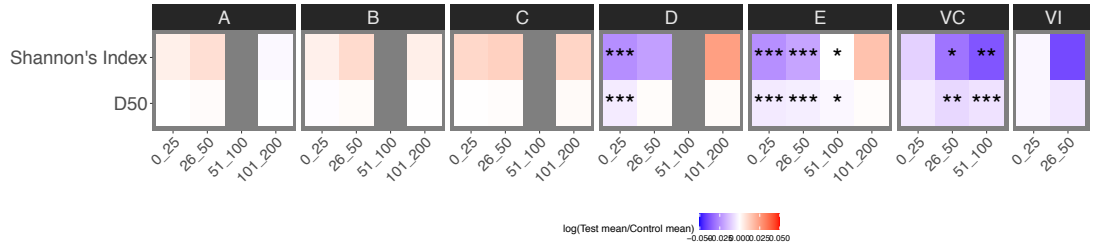


*Fig 4.29 Shannon's index and D50 diversity metrics. Heatmap showing log2 fold change in mean diversity indices between SARS-CoV-2 and vaccine cases and HC, within severity categories and across time bins post screening (cat. A), symptom onset (cat. B-E) or vaccination (cat. VC and VI). Wilcoxon test FDR adjusted p-value: *<0.05, **<0.005, ***<0.0005.*

Post SARS-CoV-2 infection, reduced diversity is most prominent in the severe groups in the IgM and IgA subgroups, and appears less pronounced for IgG, while in contrast, vaccination induces this reduction in IgM and IgG, and not IgA, perhaps reflecting the fact that vaccination does not engage mucosal immunity (Fig 4.30).

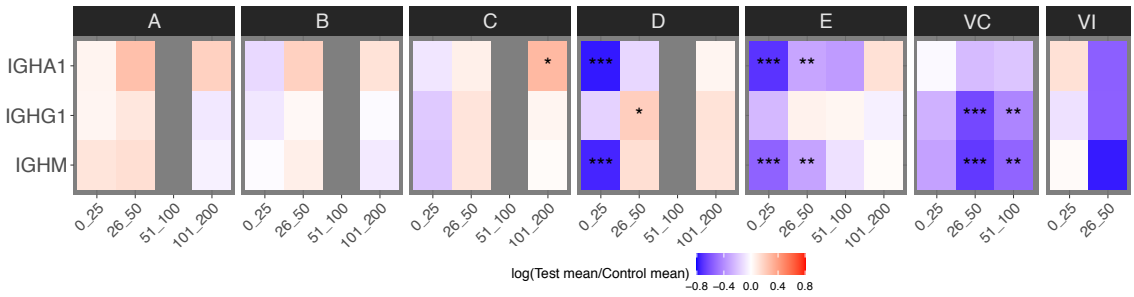


*Fig 4.30 Diversity metrics according to isotype. Heatmap showing log2 fold change in mean Simpson's diversity between SARS-CoV-2 and vaccine cases and HC, within severity categories and across isotypes and time bins. Wilcoxon test FDR adjusted p-value: *<0.05, **<0.005, ***<0.0005.*

The kinetic recovery of diversity is shown for different isotypes using the Simpson's index (Fig 4.31) in patients infected with SARS-CoV-2. This highlights the heterogeneity of group E with a portion showing no change in diversity.

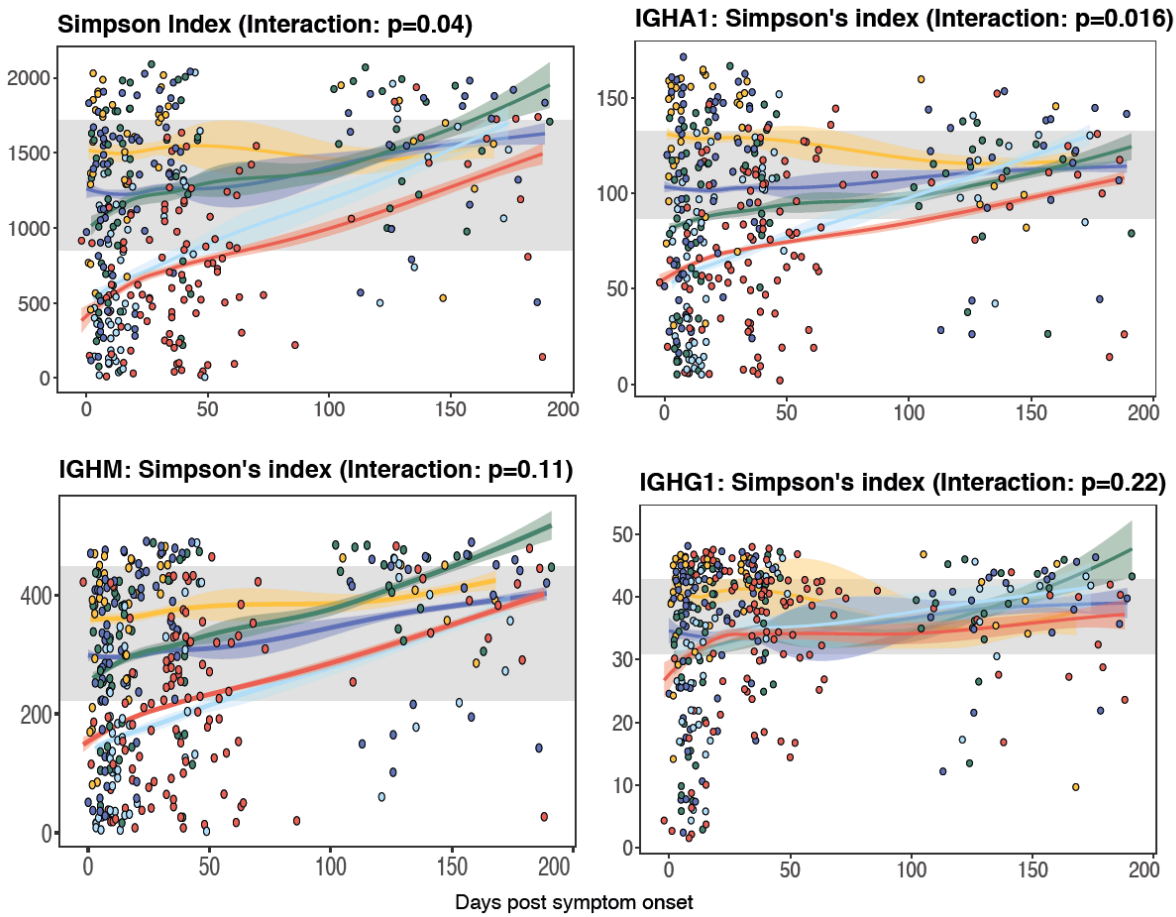


Fig 4.31 Linear mixed-effects model of Simpson's diversity index. Linear mixed-effects model showing Simpson's diversity index over time, grouped by severity and isotype. Grey band indicates the interquartile range of the corresponding isotype in HCs. Nominal p-values for the time x severity group interaction term are reported.

In keeping with IGHA displaying a decrease in diversity post SARS-CoV-2 infection and IGHG displaying a decrease in diversity post SARS-CoV-2 vaccination, there were significantly higher levels of IgG spike specific antibodies compared with IgA in patients post vaccination whilst in natural infection, levels were comparable (Fig 4.32).

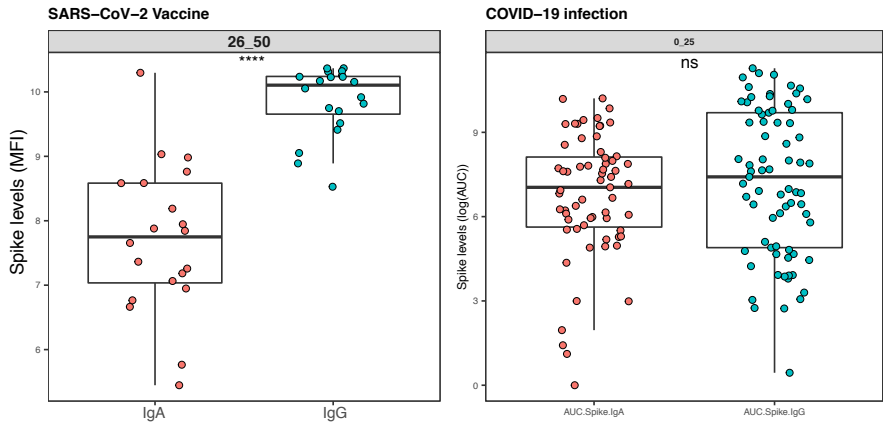


Fig 4.32 Anti-SARS-CoV-2 spike antibody level. Boxplots showing anti-SARS-CoV-2 spike antibody levels split according to isotype in COVID-19 patients (C,D and E) and post vaccination.

4.2.8 Variable gene usage

An examination of the contribution to the repertoire of various VH genes was then performed, with healthy controls matched by age with each of the severity groups (Fig 4.33).

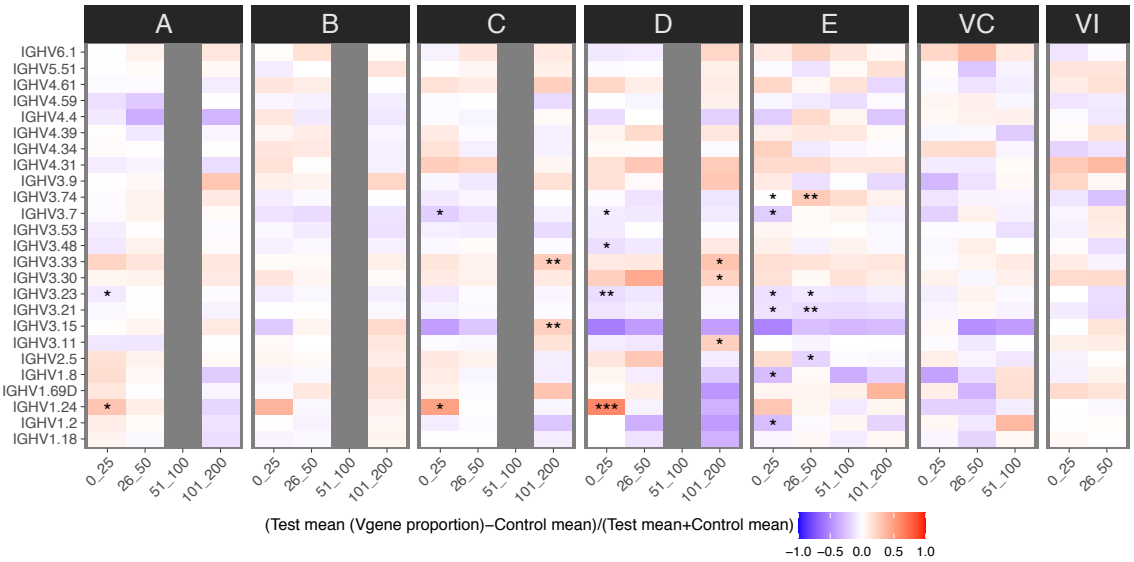
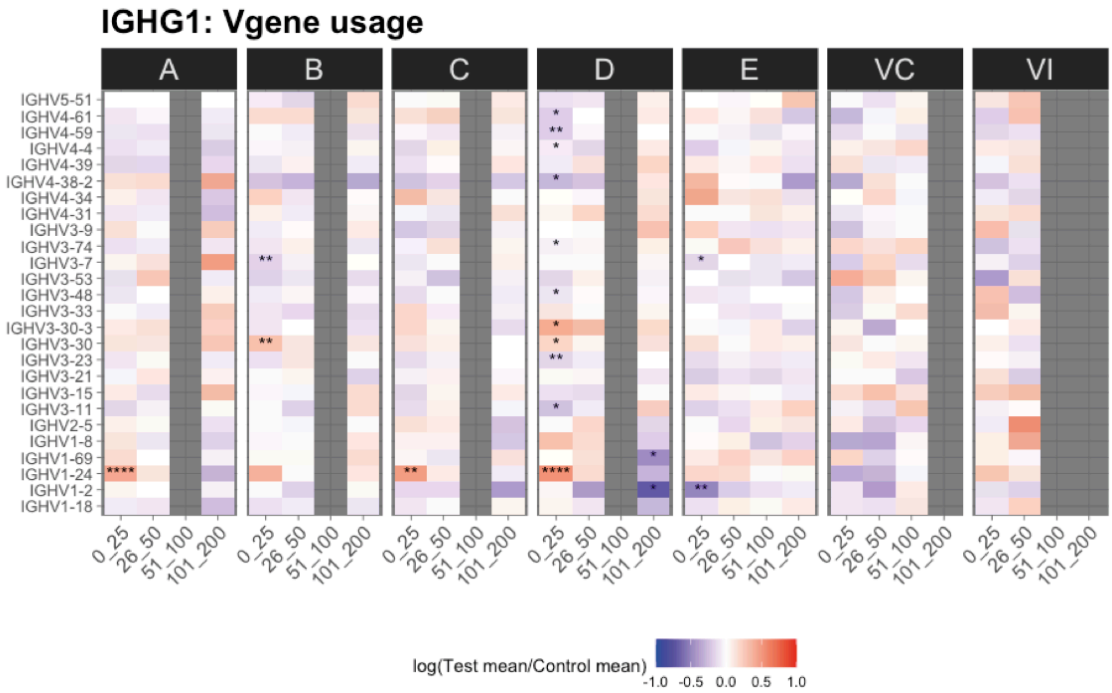
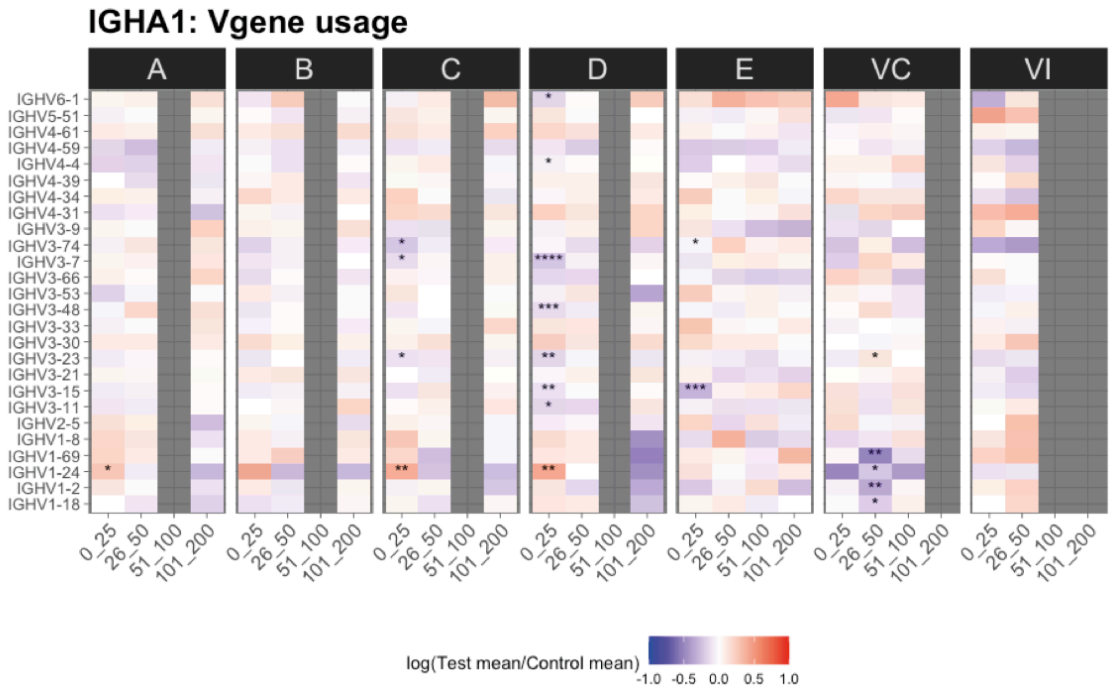


*Fig 4.33 Variable gene usage. Heatmap showing the difference between V gene proportion between SARS-CoV-2 and vaccine cases and HC, within severity categories and time bins. Difference calculated using the following, mean Vgene proportion of disease - mean Vgene proportion of HC/ mean Vgene proportion of disease + mean Vgene proportion of HC. Wilcoxon test FDR adjusted p-value: *<0.05, **<0.005, ***<0.0005.*

Two broad features were apparent. The first was that the majority of statistically significant changes in the VH gene usage were seen only in groups C, D and E, which were most prominent early. These were thought to be most likely a consequence of the major changes in B cell subsets which occur in those with severe disease. In contrast, one IGVH gene, VH1-24, was increased in all severity groups in the first-time window. VH1-24 was notably increased in IGHG1, IGHA1 and IGHM reads (Fig 4.34).





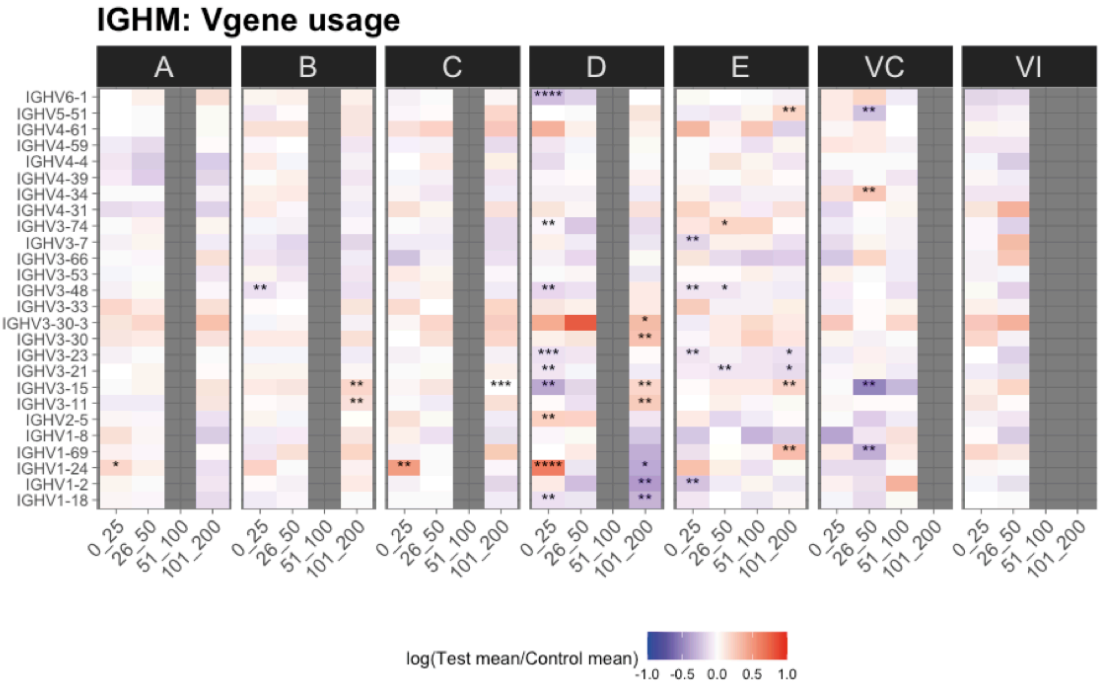


*Fig 4.34 Variable gene usage according to isotype. Heatmap showing the difference between V gene proportion between SARs-CoV-2 and vaccine cases and HC, within severity categories, isotypes and time bins. Difference calculated using the following, mean Vgene proportion of disease - mean Vgene proportion of HC/ mean Vgene proportion of disease + mean Vgene proportion of HC. Wilcoxon test FDR adjusted p-value: *<0.05, **<0.005, ***<0.0005.*

VH1-24 has been shown to be strongly associated with antibodies which recognise the N-terminal Domain (NTD) of the SARS-CoV-2 spike protein, conferring neutralisation even in the germline state[196]. Consistent with this VH1-24 proportion was strongly associated seroconversion and the development of neutralising antibodies (Fig 4.35 and 4.36), an observation not confounded by disease duration (Fig 4.37).

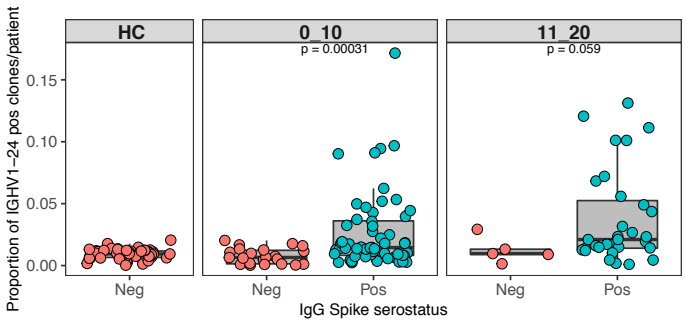


*Fig 4.35 IGHV1-24 positive clones. Boxplots showing proportion of IGHV1-24 positive clones/per patient split according to days from symptom onset/swab and IgG spike serostatus. p-value: *<0.05, **<0.005, ***<0.0005.*

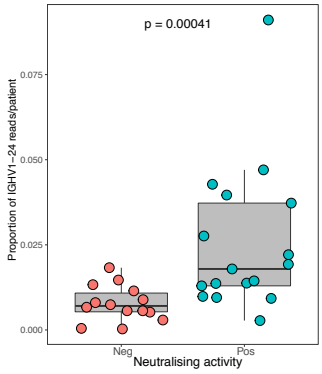


*Fig 4.36 IGHV1-24 positive clones according to neutralising activity. Boxplots showing proportion of IGHV1-24 positive clones/per patient split according to neutralisation ability. p-value: *<0.05, **<0.005, ***<0.0005.*

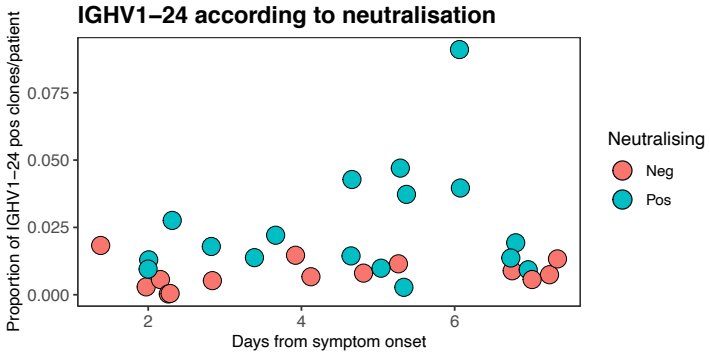


Fig 4.37 IGHV1-24 positive clones according to neutralising activity and time.

VH1-24 was increased in proportion at 0-25 days from symptom onset in groups A, C and E, but at this time there was no evidence of a concurrent increase in SHM or clonal expansion (Fig 4.38). There was a significant difference in VH1-24 proportion between groups D and E (Wilcox test, p value: 4.8e-03) at 0-25 days from symptom onset. No increase in VH1-24 was seen after SARS-CoV-2 vaccination.

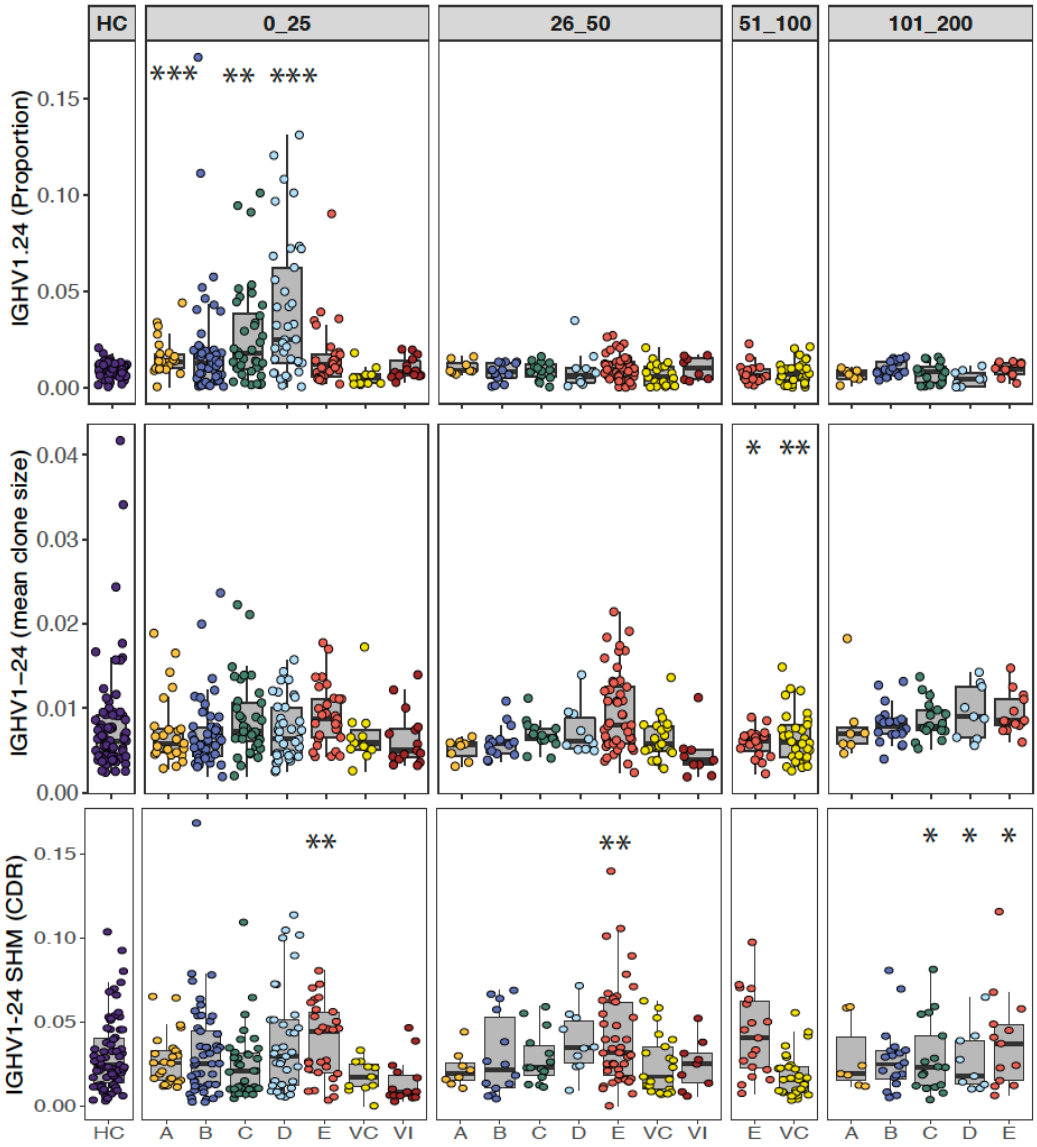


*Fig 4.38 IGHV1-24 metrics. Boxplots showing IGHV1-24 proportions, mean SHM and expanded IGHV1-24 clones per person by severity categories and time bins. p-value: *<0.05, **<0.005, ***<0.0005. Circles represent individual donors.*

4.2.9 Clonal convergence

We looked for overlap between BCR clones present in our study with the CoV-AbDab database, a resource detailing all published and patented antibodies shown to bind SARS-CoV-2 and other coronaviruses[231]. Convergent clones were defined by sharing of IGHV and IGHJ genes, having identical CDR-H3 region length and having CDR-H3 sequences that show 85% amino acid homology, and thus likely to have a similar antigen specificity to the reference antibodies.

We found clonotype convergence in both IGHD/IGHM clones in COVID-19 as well as in patients vaccinated against SARS-CoV-2, within the first 25 days from symptom onset/vaccination (median proportion of convergent clones in HC: 0, A: 0.00016, B: 0.000088, C: 0.00023, D: 0.00057, E: 0.00043, VC: 0, VI:0) and class-switched clones (median proportion HC: 0, A: 0.00039, B: 0.00047, C: 0.00047, D: 0.00048, E: 0.00046, VC: 0, VI:0). At 26-50 days clonal convergence decreased in all isotypes but most markedly in IGHD/IGHM (median proportion HC: 0, A: 0, B: 0.000037, C: 0, D: 0, E: 0.000059, VC: 0, VI:0) consistent with class-switching of antigen specific clones (Fig 4.39). Overall, there appeared to be an increase in convergence in groups C, D and E, compared with groups A and B, consistent with greater convergence being related to disease severity.

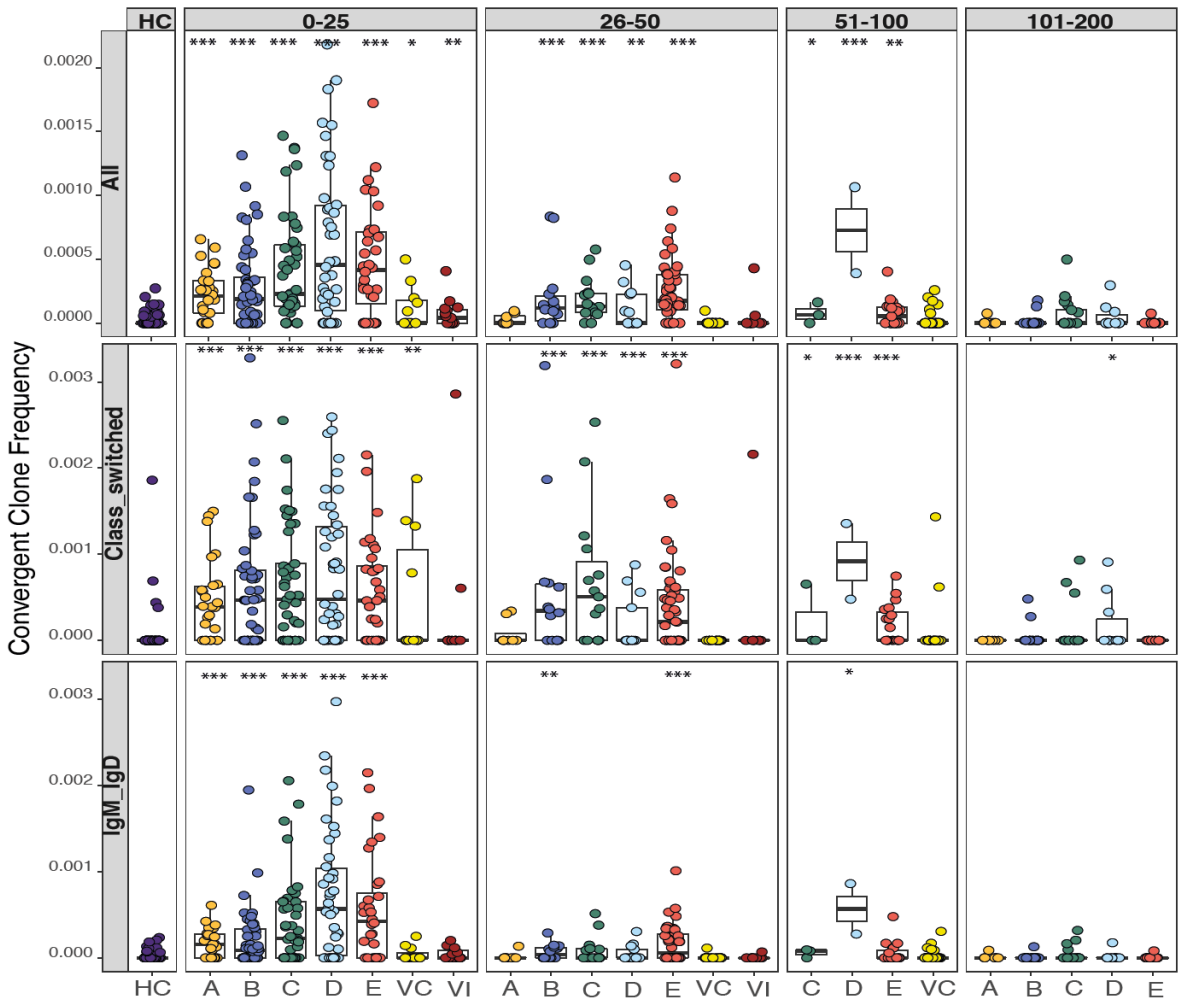


*Fig 4.39 Clonal convergence according to isotype. Convergent clone frequency between COVID-19 and vaccinated patients with the CoV-AbDab database. This represents the % of unique clones in each patient that are also found in the COV-AbDab database Samples split by severity categories and time bins post screening (cat. A), symptom onset (cat. B-E) or vaccination. One-sided Wilcoxon test FDR adjusted p-value: *<0.05, **<0.005, ***<0.0005. Circles represent individual donors.*

With infection, there was a significant convergence of class-switched and non-class-switched clones described to be neutralising and targeting the RBD and NTD of spike. In contrast, after SARS-CoV-2 vaccination, COVID-19 specific clones were mainly class-switched and targeted the RBD (Fig 4.40).

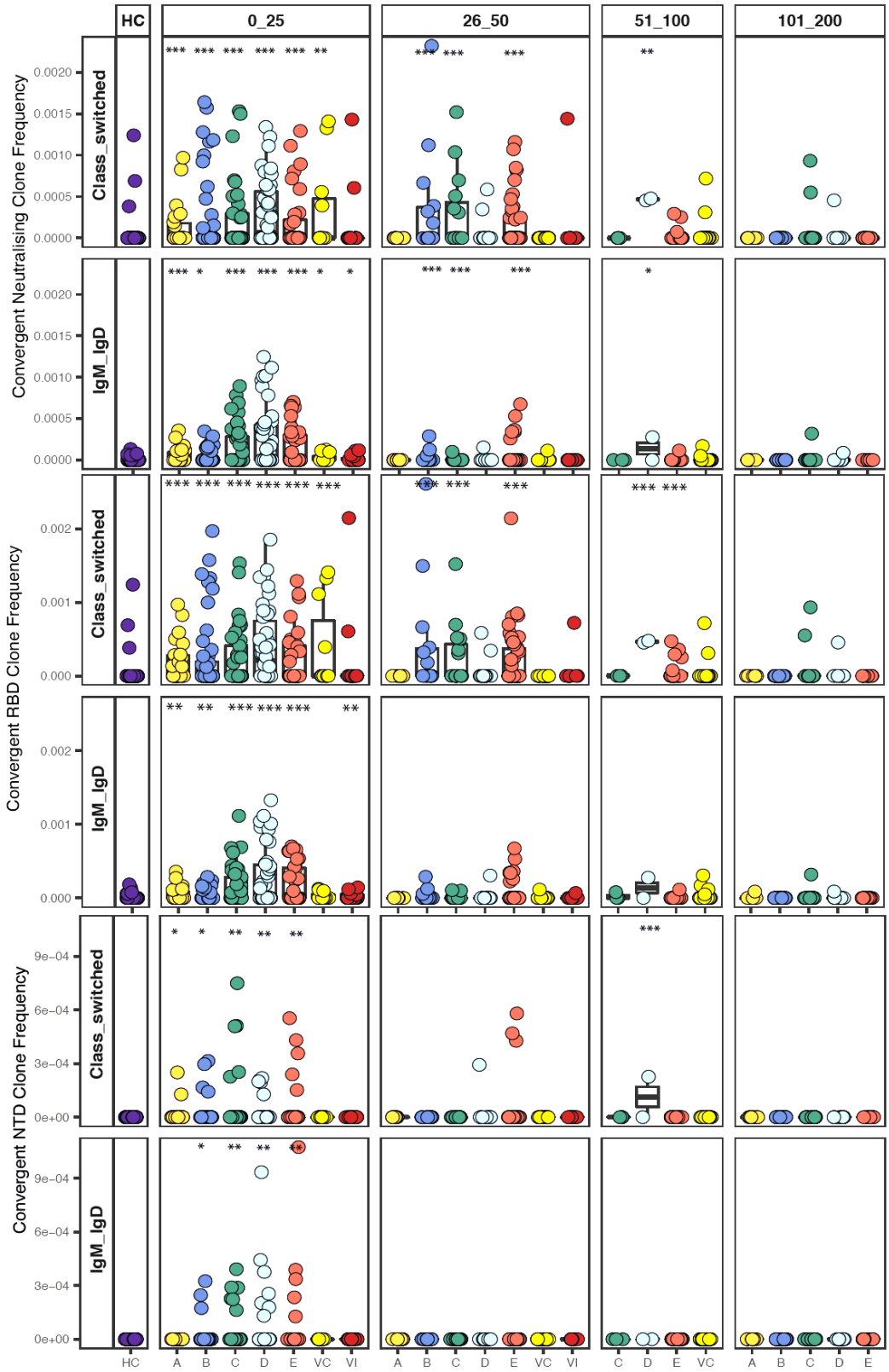


*Fig 4.40 Clonal convergence according to isotype and spike region. Convergent clone frequency of neutralising, RBD-specific and NTD-specific clones between COVID-19 and vaccinated patients with the CoV-AbDab database. Samples split by severity categories and time bins post screening (cat. A), symptom onset (cat. B-E) or vaccination. One-sided Wilcoxon test FDR adjusted p-value: *<0.05, **<0.005, ***<0.0005. Circles represent individual donors.*

Increased convergence with the CoV-AbDab database was present in patients who had seroconverted (Fig 4.41) as well as in patients with neutralizing antibodies (Fig 4.42).

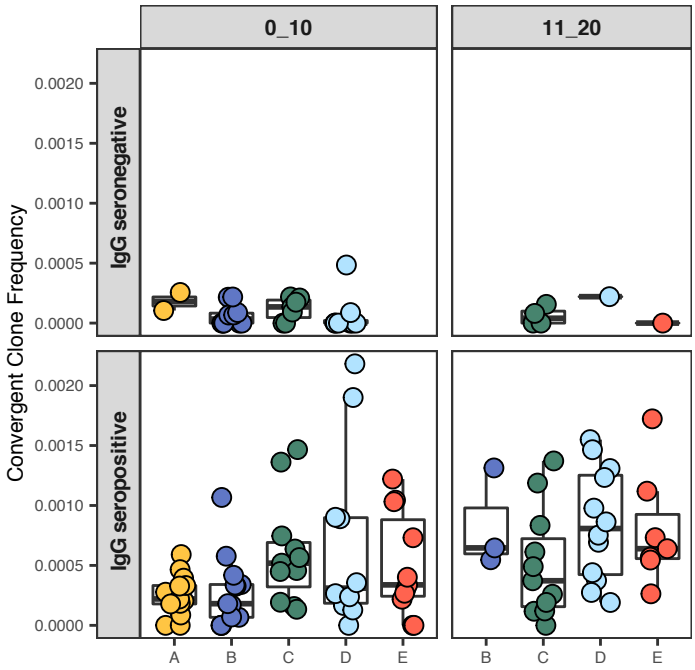


Fig 4.41 Clonal convergence according to serostatus. Boxplots showing convergence per patient split according to days from symptom onset and IgG spike serostatus. Circles represent individual donors.

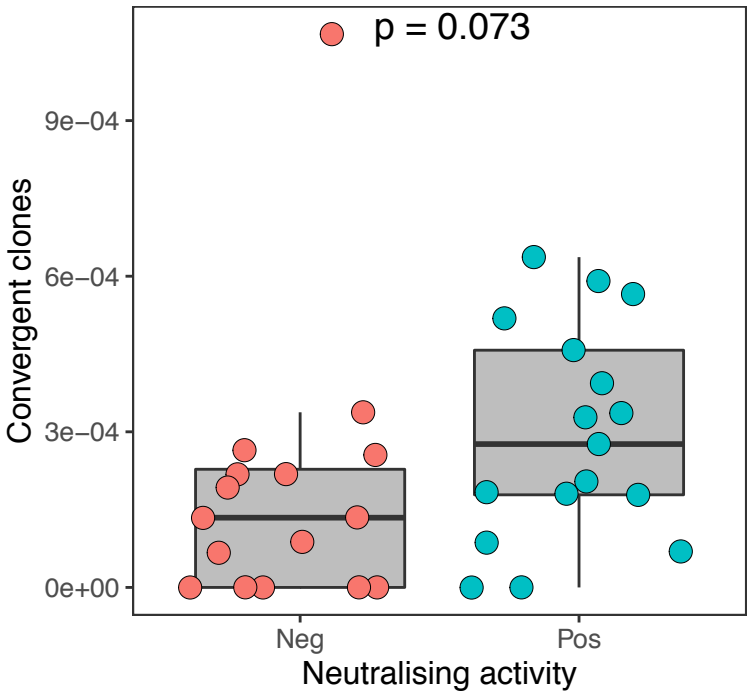


Fig 4.42 Clonal convergence according to neutralising activity. Boxplots showing convergence per patient split according to neutralisation ability. Circles represent individual donors.

In order to assess if there was an increase in somatic hypermutation with time, in clones that were convergent with the CoV-AbDab database, we created phylogenetic trees. We tracked the level of somatic hypermutation at serial time points of a given clone, sampled at multiple time points, for a given patient (Fig 4.43). Clone tracking showed progressive somatic hypermutation with time. Trees on the left that show that clones from the bleed taken early post symptom onset (blue) are close to the germline, denoted in black, whilst are more distal from the germline at later time points. The Figs on the right annotate the clones according to isotype. Patient CV0069 shows increased SHM but ongoing IGHM clones whilst patient CV0071 shows increased SHM with class switching to IGHG1 and IGHA1.

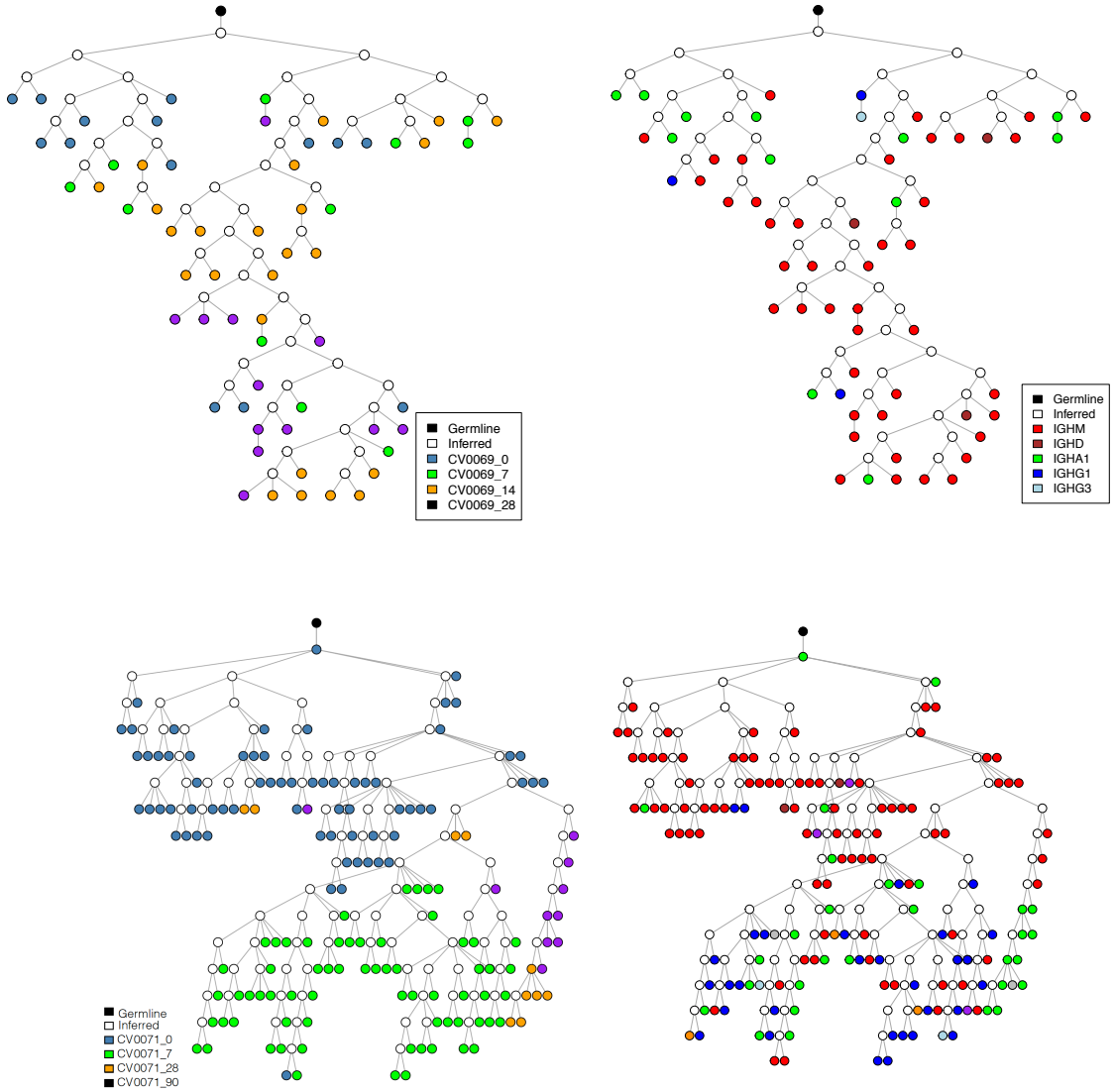


Fig 4.43 a CoV-AbDab convergent clone. Phylogenetic trees tracking a CoV-AbDab convergent clone across serial bleeds in a given patient.

To assess if there was increased clonal sharing within COVID-19 patients (0-25 from swab/symptom onset), vaccinated patients (26-50 days from vaccination) or health, we calculated the number of shared clusters pairwise up until 5 patients. We performed this in 9 patients to accommodate for the smallest group size. We performed 200 permutations where we randomly selected 9 patients within a disease group, and then 4000 unique clusters per person. This showed greater sharing in COVID-19 patients, suggestive of increased BCR overlap driven by shared antigen through exposure (Fig 4.44). Similarly, post SARS-CoV-2 vaccination, there was increased convergence, with a greater number of shared clusters amongst 4 or more people compared with natural infection or health (Fig 4.44).

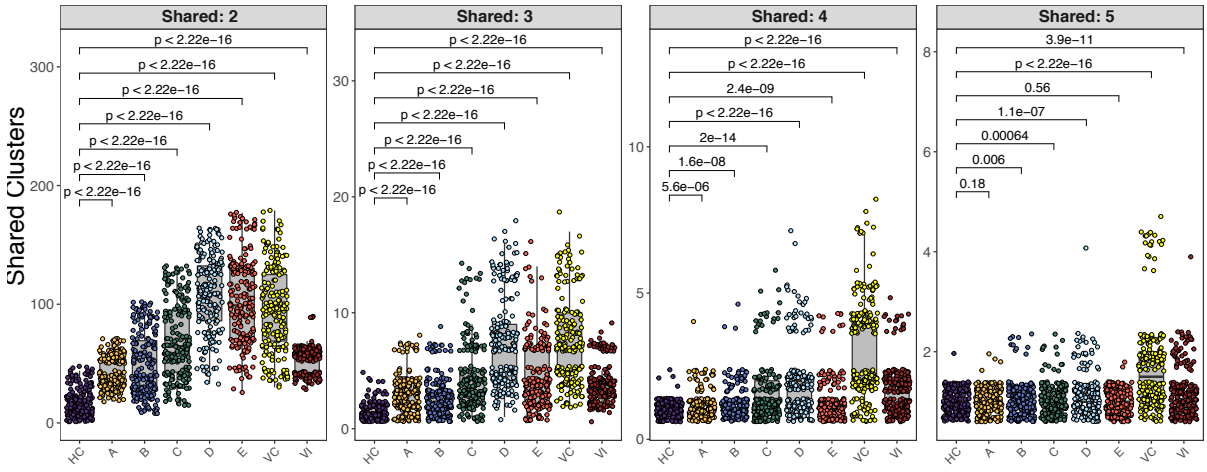


Fig 4.44 Shared clonotypes. Boxplot representing the number of clonotypes shared by patients in HC, COVID patients within 25 days from symptom onset/swab and within 25-50 post vaccination.

In order to identify new clones that might be COVID-19 specific, we looked for convergent clones that were shared amongst a minimum of 3 patients at a given time interval and that were not present in healthy controls (Fig 4.45). This revealed almost 6000 clones that were present within at least 3 patients at 0-25 days from symptom onset. Of these shared clones, 45 were also present in the CoV-AbDab database and 480 were also present at 26-50 days from symptom onset. To further understand these clones, cloning would need to be performed to understand whether they bound to SARS-CoV-2 virus, were neutralising or a bystander antibody response. The number of clones decreased at 26-50 days from symptom onset at just under 2000. Of these clones identified, 4 were present in the CoV-AbDab database and also happened to be present at 0-25 days from symptom onset. The number

of shared clones were especially low at 51-100 days from symptom onset. This is due to a decrease in clone size as well as the lower number of patients recruited at this time window. There were 50 clones that were present in at least three time windows. There was minimal sharing of clones in the vaccine groups and no overlap between SARS-CoV-2 infection and vaccination.

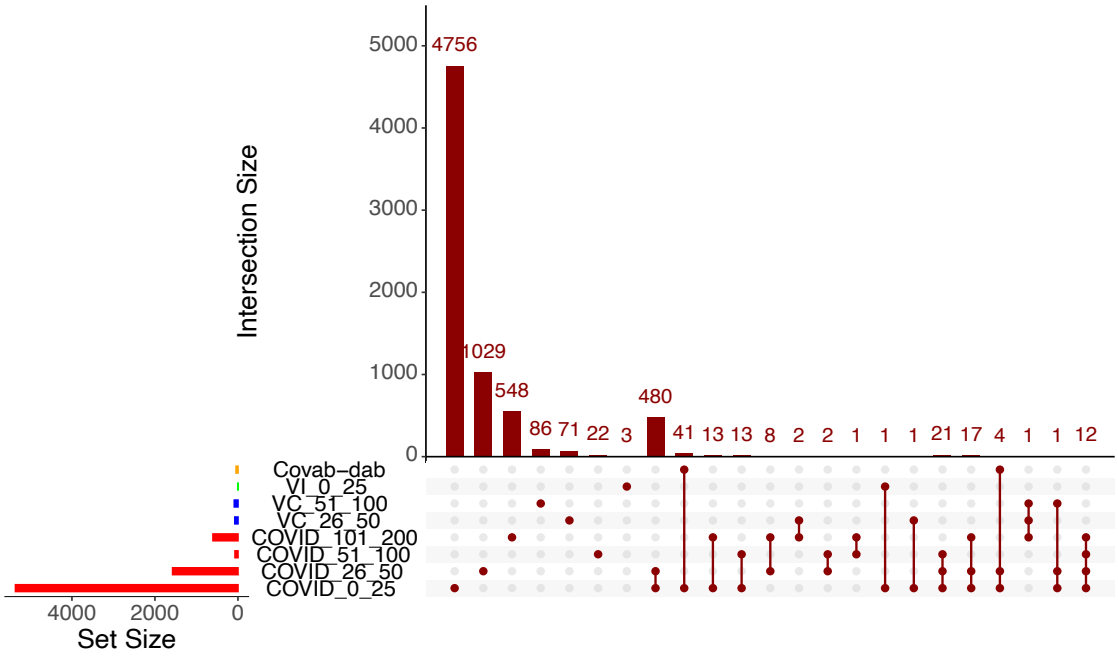


Fig 4.45 Convergent IGH clusters. Convergent IGH clusters among disease groups represented by the horizontal bars and shared across disease groups represented by the lines. A dot indicates no sharing and the total summated by the vertical histogram bars.

The convergence of clonotypes was consistent with shared antigen driving selection of clones. Although overlap with the CoV-AbDab database was minimal likely due to limitations of the database, clones were present that were shared in over 10 patients (Fig 4.46).

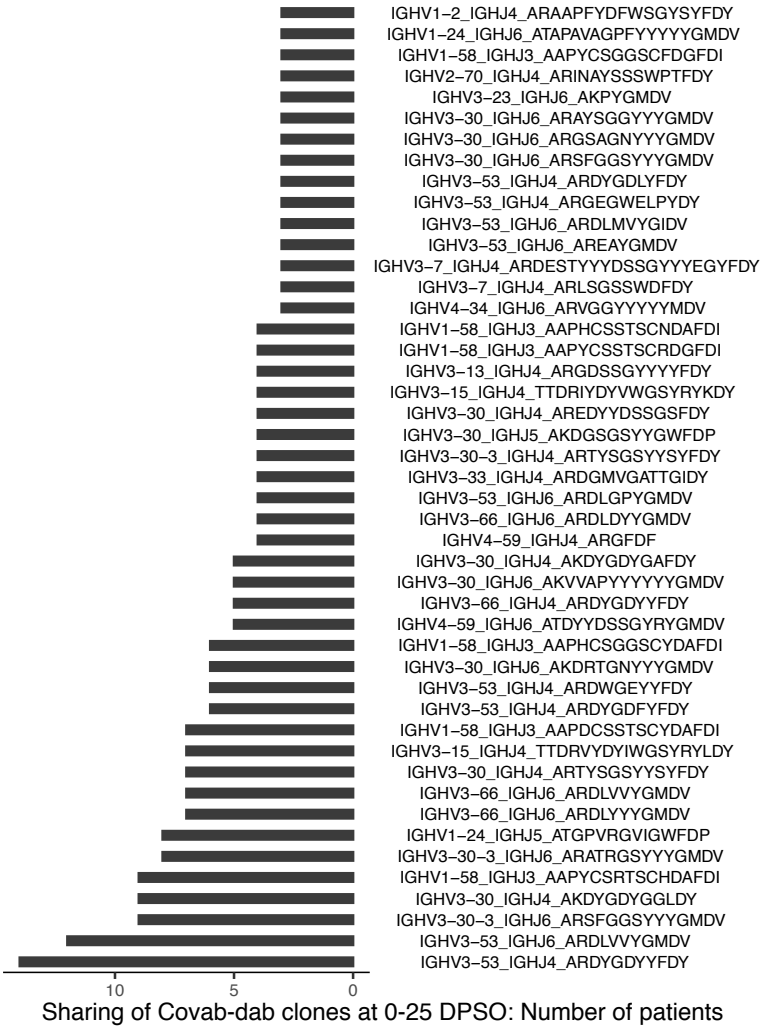


Fig 4.46 Clones present in 10 or more patients. Representation of the clones shared in ten or more COVID-19 patients and present in the CoV-AbDab database.

Examining the top 20 convergent IGH clusters from the first 25 days from symptom revealed a clones present in over 30 patients and not present in health. This level of sharing was highly suggestive of a shared antigen driven response (Fig 4.47).

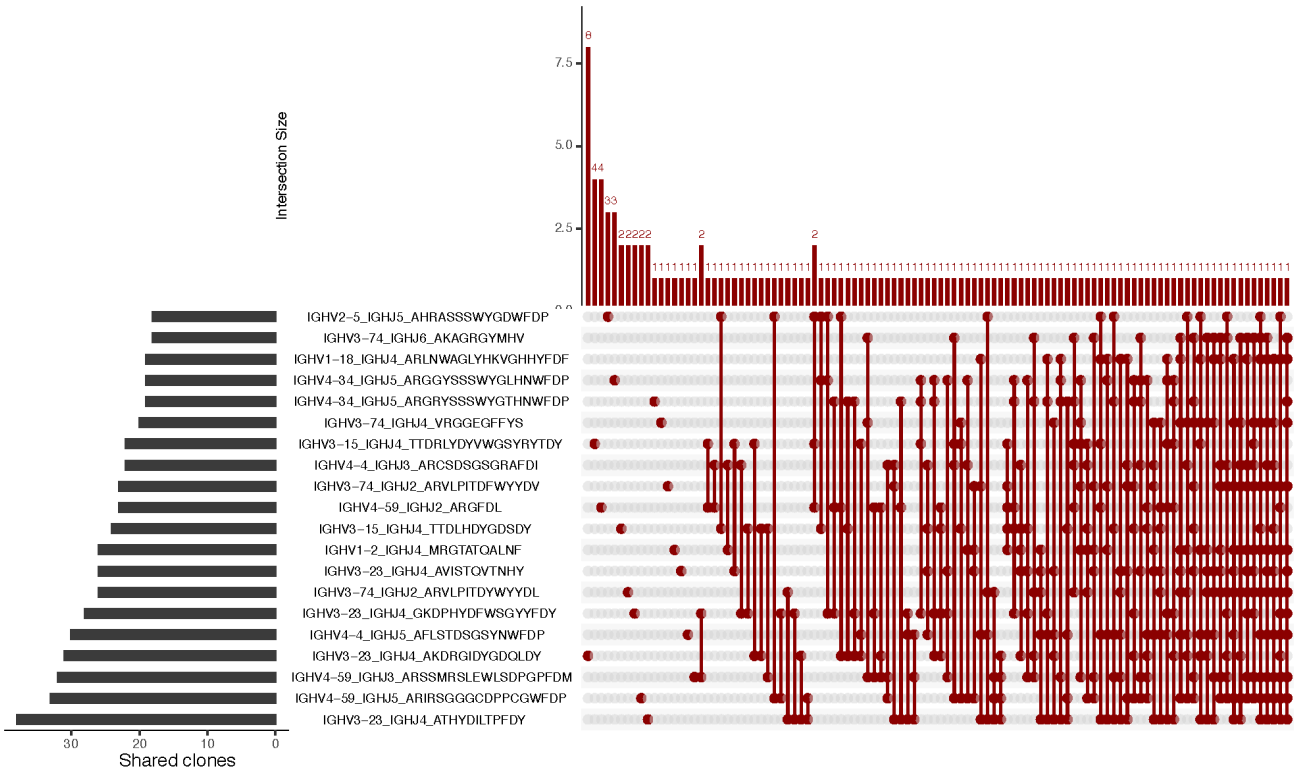


Fig 4.47 Top 20 Convergent IGH clusters. Top 20 Convergent IGH clusters in the first 25 days from symptom/swab. Number of patients that clusters are shared in is represented by the horizontal bars.

Assessing V gene usage of these disease-associated clones at 0-25 days from symptom onset, revealed an increase in representation of IGHV4-34 and IGHV1-24 compared with health (Fig 4.48 and Fig 4.49).

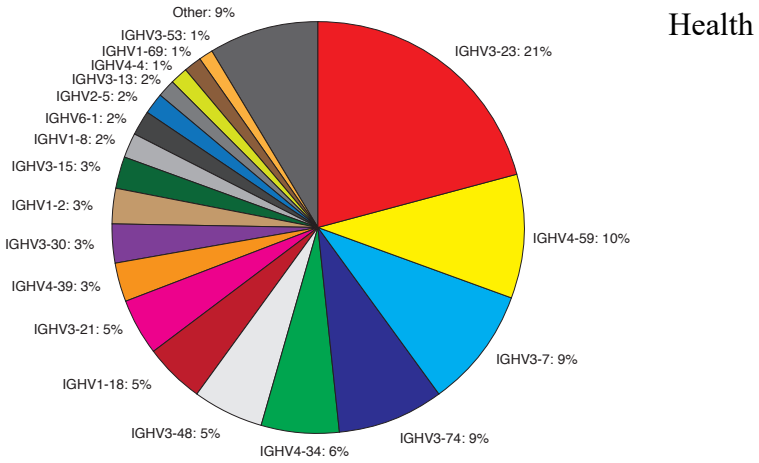


Fig 4.48 V gene usage in convergent clusters in health. Pie chart comparing V gene usage of convergent clusters in HC

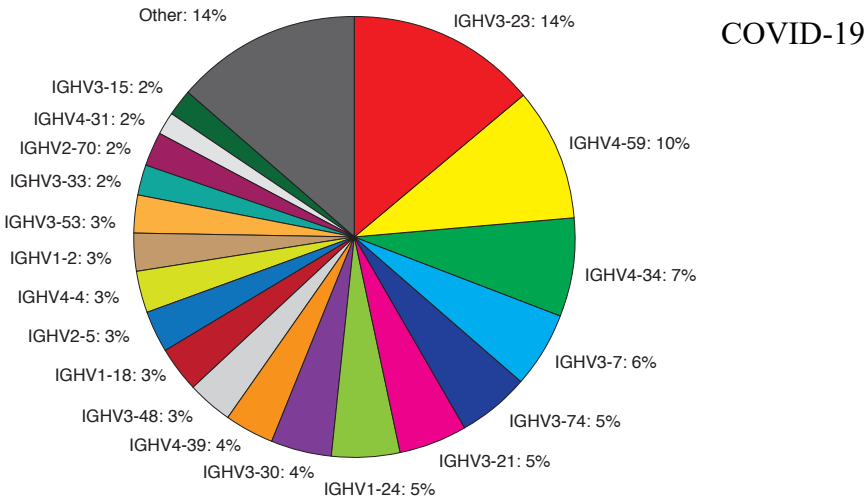


Fig 4.49 V gene usage in convergent clusters in COVID-19.. Pie chart comparing V gene usage of convergent clusters in COVID-19 within 25 days from symptom onset.

IGHV1-24 as previously mentioned plays an important role in neutralisation[196]. IGHV4-34 displays auto-reactivity against self-antigen and is associated with Systemic lupus erythematosus[109,278,298,299]. Parallels have been drawn between both conditions, with similarities in extrafollicular pathway activation and expansion of double negative B cells and the presence of lowly mutated clones [189,219,279,300].

4.2.10 Age related immune response to vaccination

We performed a subset analysis on SARS-CoV-2 vaccinated individuals to further understand the effects of age in mounting an immune response. Patients were subdivided into <80 year (n= 22) vs > 80 years old (n= 28). The age of 80 was used as a cut off as analysis of neutralisation titres showed a non-linear relationship with age with a sharp drop off at the age of 80. There were no differences in isotype proportions between the two age groups (Fig 4.50), or by neutralization (Fig 4.51).

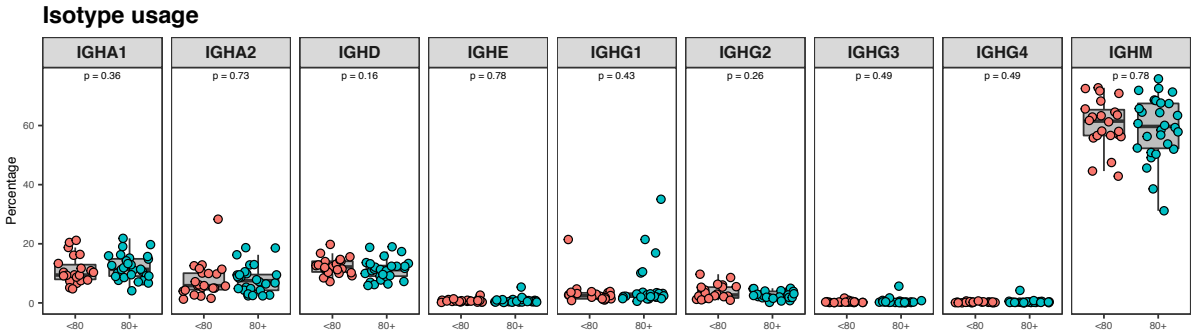


Fig 4.50 Isotype usage according to age. Isotype usage according to unique VDJ sequence comparing <80 year (n= 22) vs > 80 years old (n= 28). Differences between groups were calculated using Mann-Whitney U test.

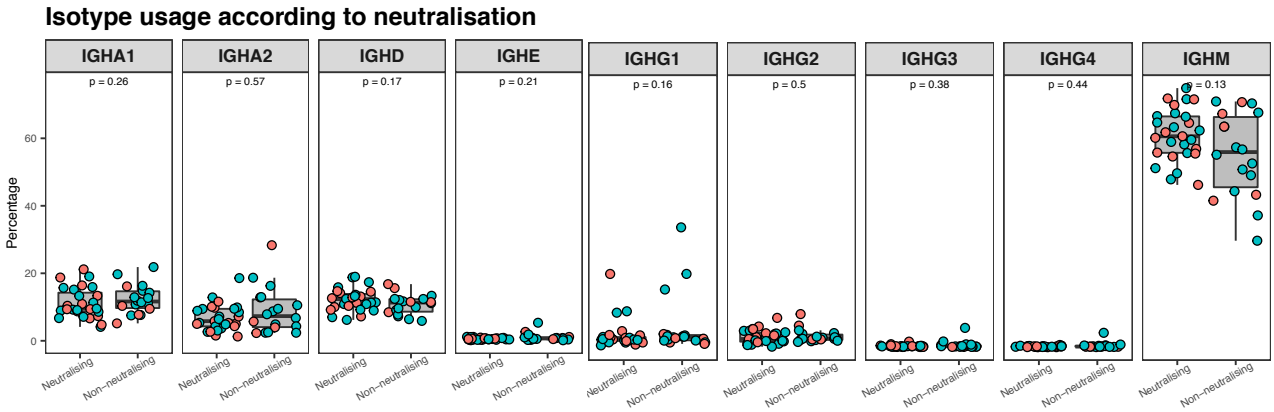


Fig 4.51 Isotype usage according to neutralising status. Boxplots showing Isotype usage according to unique VDJ sequence comparing participants <80 (n= 22) vs > 80 years old (n= 28) and association with neutralisation of spike pseudotyped virus. Neutralisation cut-off for 50% neutralisation was set at 20. Differences between groups were calculated using Mann-Whitney U test.

We found an increase in usage of the immunoglobulin heavy variable 4 (IGHV4) family in the younger age group, with an increased proportion of IGHV4-34, IGHV4-39, IGHV4-59 and IGHV4-61, whereas in the older age group there was an increase in usage of the IGHV1 family, with increases in IGHV1-18 and IGHV1-69D (Fig 4.52). IGHV4-34 is associated with autoimmunity as previously stated whilst IGHV1-69D is a paralog of IGHV1-69. IGHV1-69 is a common IGHV gene present broadly neutralizing antibodies against influenza virus, HCV, and HIV [301]

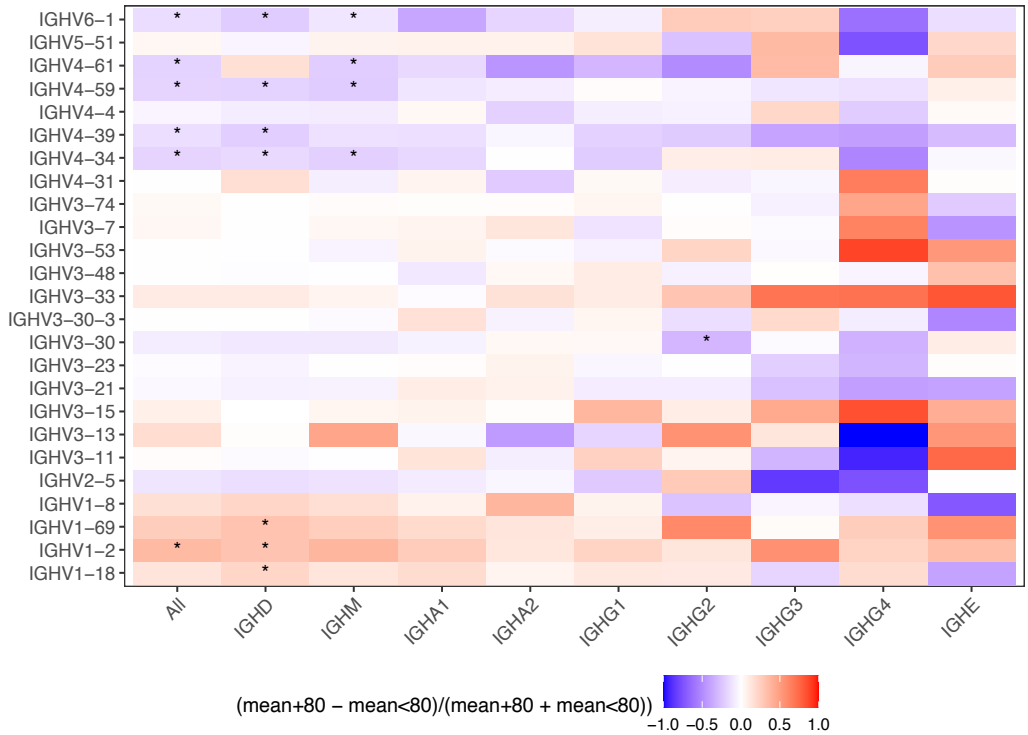


*Fig 4.52 V gene usage according to age. Heat map showing V gene usage, comparing participants <80 with >80 years old. Differences between groups were calculated using Mann-Whitney U test. A Benjamini Hochberg FDR correction was used, *P < 0.1.*

There were no significant differences in V gene usage associated with neutralization (Fig 4.53).

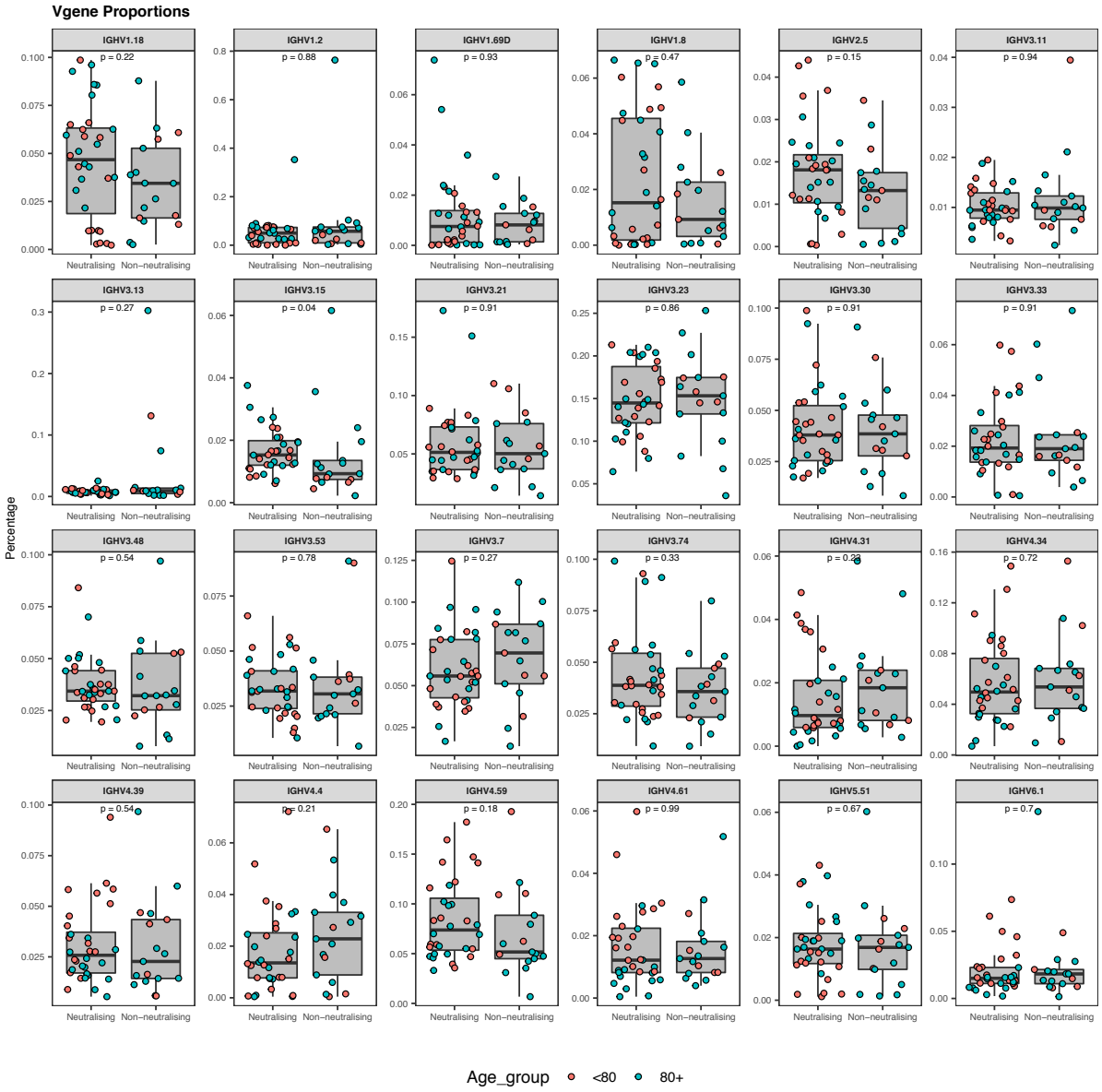


Fig 4.53 V gene usage according to neutralising status. Boxplots showing V gene usage as a proportion, comparing neutralisation of spike pseudotyped virus. Neutralisation cut-off for 50% neutralisation was set at 20. Differences between groups were calculated using Mann-Whitney U test.

Differences in somatic hypermutation could affect neutralization through antibody affinity maturation. We found that participants aged 80 years or more had a lower level of somatic hypermutation in class-switched B cell receptors (BCRs) than the younger group, and that the difference was driven by the IGHA1/2 isotype (Fig 4.54).

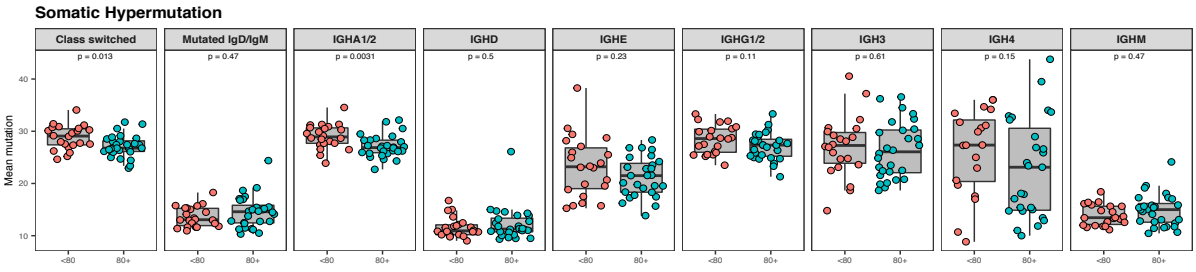


Fig 4.54 SHM according to age. Boxplots showing mean somatic hypermutation comparing participants <80 vs > 80 years old, grouped according to isotype class. Differences between groups were calculated using Mann-Whitney U test.

We also did not find any relationship between measures of diversity and age (Fig 4.55).

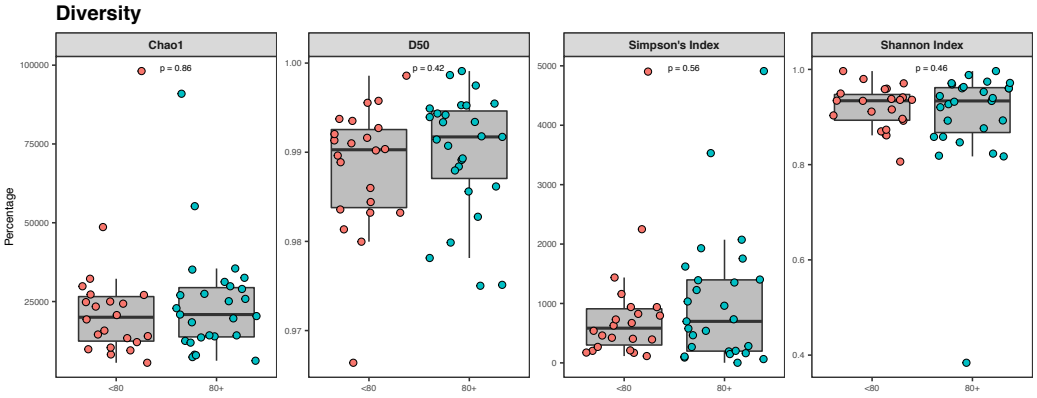


Fig 4.55 Diversity according to age. Diversity Indices comparing < 80 year olds vs > 80 year olds. The inverse is depicted for the Simpson's index and the Shannon-Weiner index is normalised. Differences between groups were calculated using a t test.

We next examined the B cell repertoire for public clones known to be associated with SARS-CoV-2 neutralization. We explored the convergence between BCR clones in our study and the CoV-AbDab database and found that participants under 80 years of age had a higher frequency of convergent clones, in keeping with increased neutralization, when compared with the older group (Fig 4.56).

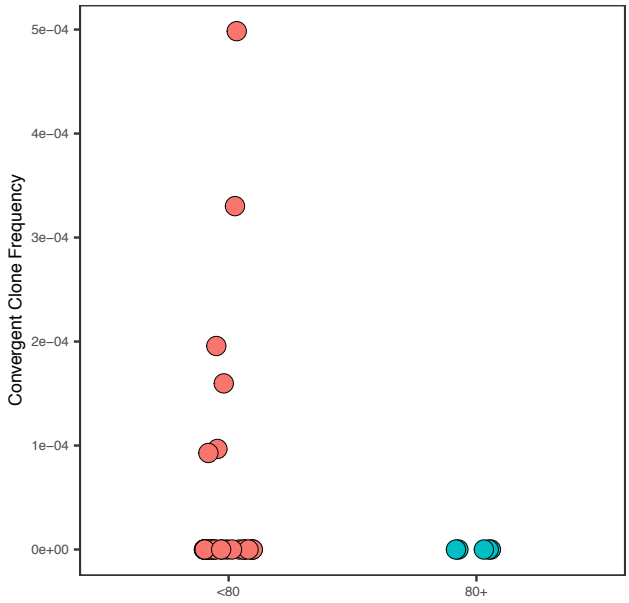


Fig 4.56 Convergent clones according to age. BCR comparison of patients in the first 50 days from vaccination <80 (n= 27) vs > 80 years old (n= 5) with public clones known to be associated with SARS-CoV-2 using the CoV-AbDab database. Clones from participants and the database were co-clustered based on matching IGHV and IGHJ segments, same CDR-H3 region length and 85% CDR-H3 sequence amino acid homology. Differences between groups were calculated using a one-sided t test.

4.3 Discussion

To assess the nature of the B cell response to SARS-CoV-2 infection we measured how BCR repertoires change in patients with COVID-19 stratified by both disease severity and time post infection. An increase in the proportion of BCRs bearing IgG1, IgG3 and IgA1 isotypes was seen in all groups soon after infection, including asymptomatic individuals. In parallel, decreased SHM was also seen in all groups, with the reduction particularly marked in the three expanded isotypes (IgG1, IgG3 and IgA1). These observations are consistent with the initial changes in the BCR repertoire being driven by an early plasmablast expansion, as they are seen in severity groups in which this is the only B cell subset which changes compared to healthy controls[270]. The implication, then, is that these early plasmablasts are unmutated, in keeping with the finding of early potent neutralising antibodies having near-germline sequences [186,188,191,280,302]. In keeping with this and previous studies, we show that a decrease in SHM correlates with seroconversion [237] and neutralisation. These plasmablasts could be derived from the rapid differentiation of already switched precursor cells from, for example, the marginal zone as the relative absence of SHM suggests that differentiating

memory B cells are not a major source. They may also be generated rapidly by isotype switching and differentiation of naive B cells to plasmablasts outside the germinal centre[303]. Given both the reduction in germinal centres and circulating as well as tissue CD4 T_{FH} cells observed in patients with severe covid-19, this seems a more likely source of plasmablasts than switching in early germinal centres[190,270,271]. Evidence of clonal evolution in the form of an increase in SHM, potency and broadening of the repertoire is however observed at 6 months post infection[186,191,192], which together with the detection of SARS-CoV-2 specific long lived plasma cells in bone marrow aspirates at 11 months post infection[304] suggests that a germinal centre response does occur post infection albeit delayed."

In those with more severe disease (groups D and E), increased clonal expansion, as evidenced by a reduction in repertoire diversity, is seen. The clonal expansion is most prominent in IgM and IgA, and not in IgG. This highlights the important role that IgA B cell memory may play post mucosal infection with SARS-CoV-2 first mainly infecting the upper respiratory tract. In addition, dimerised IgA antibodies, the form predominantly found in mucosal tissues, are highly potent against SARS-CoV-2 and are more so than IgG and monomeric IgA [305] and IgA-virus-immune complexes are potent induces of netosis via engagement of Fc-αRI on neutrophils[306].

Across all severity groups, VH1-24 was the dominant VH gene expanded after SARS-CoV-2 infection. Antibodies bearing this VH gene have been noted to make up the majority of neutralising IgG antibodies arising after SARS-CoV-2 infection, with a specificity for the NTD component of the spike antigen, rather than the receptor binding domain (RBD) which had been first assumed to be the main target of neutralising antibodies[196]. Given these observations, the fact that the proportional increase in VH1-24 in group E is substantially lower than all less severe groups raises the possibility that a robust early VH1-24 response might help prevent severe disease.

Changes in the repertoire following SARS-CoV-2 vaccination were less pronounced than in COVID-19 infection. Isotype usage showed an increase in IGHM and D and a decrease in IGHA1/2 which is reciprocal to that seen in natural infection. There were no changes in SHM or specific heavy chain usage. The lack of global changes to SHM rates, unlike the increase

seen post Influenza vaccine, likely highlights the lack of mobilisation of pre-existing memory B cells, which rapidly differentiate into plasmblasts on antigen re-exposure. Similarly, the pronounced decrease in SHM seen in infection is not apparent. This highlights how natural infection has a more overwhelming effect on the immune response that is both detectable at a global level and which potentially results in a more sustained memory response. This illustrates the important role of adjuvants in vaccination which are employed to increase the potency and longevity of the antigen specific immune response by prolonging exposure through a depot like affect and activating the innate immune system. Despite this, convergence analysis with the CoV-AbDab database revealed the generation of SARS-CoV-2 spike specific clones post vaccination and significantly increased clonal sharing compared to healthy controls. There was a larger number of clones shared post SARS-CoV-2 vaccination compared with any other group when looking at clonal overlap in a minimum of 5 people. This suggests a focusing of the immune response on a narrow range of antigens post SARS-CoV-2 vaccination. In contrast, Influenza vaccination resulted in clear global changes in the repertoire with isotype changes mirroring SARS-CoV2 infection and an increase in SHM, likely reflecting an expansion of cross-reactive antibodies from previous exposures and vaccinations [307]. Clonal expansion is, however, seen post SARS-CoV-2 vaccination and occurs with similar kinetics to that seen in response to natural infection. This appears to be driven primarily by expansion of clones bearing IgM and IgG isotypes rather than IgA that predominates in natural infection. A finding further supported by higher spike specific IgG antibody titres compared with IgA. This is reflective of the different anatomical compartments being mobilised in the early immune response with systemic vaccination being poor at generating a mucosal response[185]. The inability of systemic vaccine to induce mucosal IgA or tissue resident memory T cell responses limits efficacy against respiratory pathogens [308–310]. Given mucosal immune responses are compartmentalised with intra-nasal vaccines inducing a response in the upper and lower respiratory tracts, a vaccine utilising this approach would neutralise pathogens at the site of entry[308,309]. In support of this, a chimpanzee adenovirus-vectored vaccine administered in
Syrian hamsters intranasally resulted in less viral load and lung pathology upon challenge compared with intramuscular administration[311].

Vaccination elicited COVID-19 specific clones that were class-switched, likely neutralizing and mainly targeted the RBD of the spike protein. However, the level of convergence with CoV-AbDab database was far lower than that seen in severe disease, with minimal formation of antibodies targeting NTD compared to natural infection. The reason for the under-representation of NTD-specific clones is unclear given that the BNT162B2 SARS-CoV-2 vaccine utilises the sequence of the full-length SARS-CoV-2 spike protein, including the NTD. The lack of NTD convergent clones may represent a limitation of the database or that NTD targeting clones are rarer and thus less likely to be sampled. If a true difference in vaccination and infection is present, it might be that future vaccine design strategies might be developed to increase the immunogenicity of the NTD, but it is also worth bearing in mind that there is a divergence of views on the importance of this antigen in the generation of neutralising antibodies[286,287].

Global repertoire analysis is a useful adjunct to antigen specific B cell responses and can inform vaccine strategies. We show temporal changes in the BCR repertoire in response to natural SARS-CoV-2 infection and generation of antigen-specific B cell response, which is dynamically and compositionally distinct from vaccination. SARS-CoV-2 natural infection results in activation of mucosal immunity with clonal expansion in IgM and IgA isotypes and an increase in VH1-24. SARS-CoV-2 vaccination induces very different isotype changes to natural infection, results in clonal expansion of IgM and IgG and appears to focus the immune response to the RBD.

We recognize that a limitation of the study is that neither the antigenic specificity nor neutralising capacity of antibodies encoded by identified BCR sequences was determined experimentally. Instead, function was inferred by similarity to sequences deposited in the COV-AbDab database. While this is an excellent and growing resource, this approach is limited in breadth, with a particular bias towards the identification of RBD-binding clones, and likely results in the under reporting of SARS-CoV2 specific clones in our dataset. Future work would include generating monoclonal antibodies from convergent IGH sequences to allow further characterisation. In addition, the analysis of the BCR repertoire of flow sorted B cell subsets would have enabled a more granular delineation of how SARS-CoV-2 infection impacts the BCR repertoire.

5. B cell receptor Repertoire in Crohn's Disease

5.1 The gastrointestinal immune system

5.1.1 Introduction

Immune cells are present in defined anatomical compartments as well as being scattered throughout in the gastrointestinal tract. The gut associated lymphoid tissues include Peyer's patches which are present in the small intestine, lymphoid tissue of the appendix and lymphoid follicles which reside in the intestinal wall[312]. There are 100-200 Peyer's patches in the small intestine, and these contain a larger number of B cell follicles compared with peripheral lymphoid organs. Outside lymphoid organs, lymphocytes can be found in two compartments, the epithelium and the lamina propria which is the connective tissue that underlies the epithelium[2]. Mesenteric lymph nodes are the largest lymph nodes in the body and receive drainage from Peyer's patches and the lamina propria[312] (Fig 5.1)

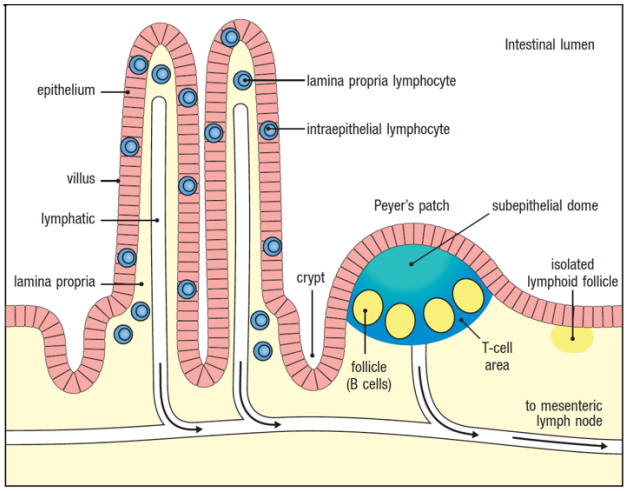


Fig 5.1 Gastrointestinal Immune system illustrating Peyer's patches and scattered lymphocytes.[2]

The gastrointestinal immune system can be divided functionally into inductive and effector sites[312]. The inductive sites are where antigen is sampled and results in activation of naïve and memory cells. This includes Peyer's patches, lymphoid follicles and mesenteric lymph nodes. The effector site is where post differentiation, effector cells perform their function and includes the epithelium and lamina propria.

5.1.2 Intestinal epithelial cells

Mucosal surfaces contain a single layer of intestinal epithelial cells. These cells along with a mucosal layer form the first barrier to pathogen entry[313]. Cells present in this layer include goblet cells which secrete mucus, paneth cells and microfold cells (M cells)[314]. Goblet cells, another specialized epithelial cell produce mucus which forms a barrier to invasion and in addition acts as a scaffold, retaining IgA antibodies and antimicrobial peptides[313]. Paneth cells express toll like receptors (TLR) and Nucleotide-binding and oligomerization domain-like receptors (NOD), are highly autophagic and produce anti-microbial peptides including RegIIIγ and defensins. M cells do not produce digestive enzymes or mucus and therefore allow direct passage of microbes from the lumen. The basal cell membrane is extensively folded forming a pocket enclosing lymphocytes. Thus, when microbes translocate, macrophages and dendritic cells are poised for uptake and subsequent presentation to the adaptive immune system[2].

Lymphocytes in the epithelium (intraepithelial lymphocytes) of the small intestine are predominantly CD8αβ T cells and bind to E-cadherin on epithelial cells via integrin CD103[315]. These activated CD8 T cells contain perforin and granzyme granules and are oligoclonal with restricted VDJ gene segments[2].

5.1.3 Lamina Propria

The lamina propria contains IgA-plasma cells, CD4 and CD8 memory and effector cells, innate lymphoid cells, dendritic cells, macrophages and mast cells[316]. These lymphocyte express integrin α4:β7[2]. CD4 T cells predominate in the lamina propria[312]. CD4 T cells are heterogeneous and differentiate depending on the cytokine milieu. Interleukin-12 causes the up-regulation of the transcription factor T-bet which leads to the differentiation to type 1 helper T (Th1) cells[317]. These cells are pro-inflammatory, secreting interferon-γ and TNF-α and recruit macrophages, natural killer cells, and CD8+ T cells[312]. Whereas, interleukin-6, TGF-β, and interleukin-1 cause the up-regulation of interleukin-23R and transcription factors including retinoic acid–related orphan receptor gamma t (RORγt). This makes the T cell more responsive to interleukin-23 and thus differentiate to Th17 cells[318]. Th17 cells recruit neutrophils, secrete IL-17 and IL-22. This pro-inflammatory state is balanced by presence of

Treg cells which produce IL-10 and modulate the expression of RORγt[319]. CD103+ dendritic cells are tolerogenic and play an important role in promoting tolerance by promoting the expansion and differentiation of Tregs[320].

5.1.4 Peyer's Patches

Peyer's patches are specialised immune niches formed in anatomical regions known as the subepithelial dome. B and T cells activated in Peyer's patches do not directly migrate to the adjacent lamina propria but rather drain into the mesenteric lymph nodes followed by the thoracic duct and into the circulation. The expression of CCR9 on their cell surface allows homing to the lamina propria via chemokine CCL25 produced by gut epithelial cells. α4:β7 binds to MAdCAM-1 expressed on endothelial cells of the blood vessels in mucosal tissues enabling emigration into the lamina propria[321] (Fig 5.2). Isotype switching to IgA occurs in Peyer's patches but do not differentiate fully into plasma cells until return back into the lamina propria[322].

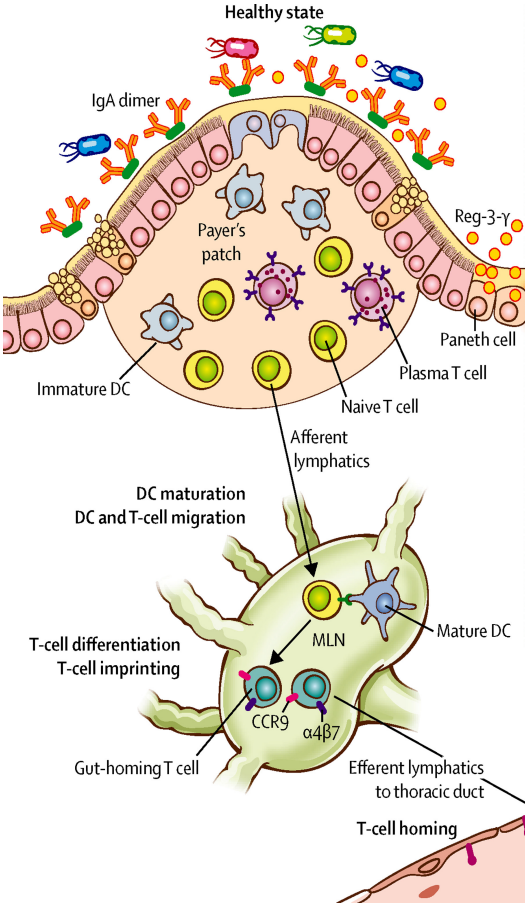


Fig 5.2 Passage of immune cells in the gut[323].

5.1.5 Immunoglobulin

IgA is the dominant mucosal antibody. In blood, IgA takes on a monomeric form whilst in mucosal tissues, IgA forms a dimer. Naïve B cells are activated in Peyer's patches and mesenteric lymph nodes. TGF-β stimulates class-switching to IgA and is T_{FH} cell dependent and IL-5, IL-6, IL-10 and IL-21 promotes expansion of the population[322]. B cells do not differentiate into plasma cells in the Peyer's patches but rather IgA lymphoblasts enter the circulation and then return to the intestinal lamina propria where they terminally differentiate.

For IgA and IgM antibodies to reach the luminal surface where antigen is present, they bind polymeric immunoglobulin receptor (pIgR) which transports them across to the luminal surface. pIgR is expressed constitutively in epithelial cells at the basolateral surface. pIgR binds to the J-chain of dimeric IgA and polymeric IgM and transports the antibody via endocytosis to the luminal surface. To release the antibody, proteolytic cleavage occurs and part of the cleaved pIgR remains bound to IgA, this is known as the "secretory component" (Fig 5.3)[2,324].

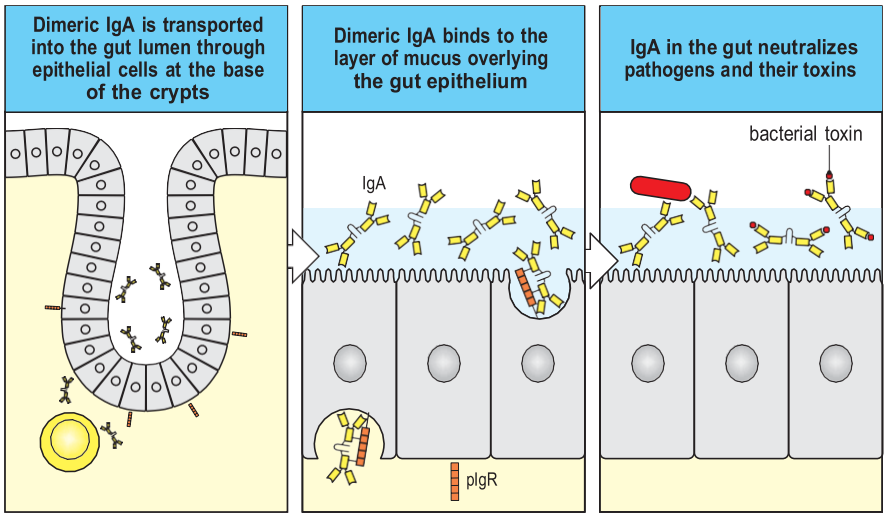


Fig 5.3 Passage of IgA and IgM to mucosal surfaces[2].

5.1.6 Tolerance

The human gastrointestinal tract houses more than 1000 species of microbes and the majority can be grouped into two broad categories, Bacteroidetes (gram negative) and Firmicutes (gram positive)[325]. The microbiota is imperative in maintaining gut health[326]. The microbiota assists in metabolizing dietary components such as cellulose and produce key vitamins such as vitamin K. Short-chain fatty acids are produced and act as an important source of energy for colonic enterocytes. The microbiota out competes pathogenic organisms and is non-invasive thus not causing tissue injury. 75% of commensal organisms are coated by IgA. The microbiota also strongly influences the host immune system[327]. A reduction in $\alpha\beta$ and $\gamma\delta$ intraepithelial lymphocytes and IgA antibodies occurs in germ-free mice which is reversed on colonization[328,329].

Germinal centres arise in a homeostatic manner in the gut and are highly influenced by the microbiota. Chen et al, show that in mice, there is a large overlap in repertoire across multiple mice in keeping with the presence of a "public repertoire"[330]. Public repertories are generated through common microbial antigens. Chronic germinal centres in Peyer's patches require the presence of antigen as illustrated by the lack of them in germ-free mice and the restoration of them post vaccination and colonisation of commensals.

The mucosal system needs to mount a response against infectious pathogens whilst remaining tolerant towards harmless antigens such as food and commensals.
Pathogens are sampled in multiple ways, including via M cells, dendritic cells which extend their dendrites through tight junctions and translocation of gut antigens through enterocytes[313].

Epithelial cells play an important role in preventing infection. They contain TLRs on their basolateral and apical surfaces as well as in intracellular vesicles which recognize PAMPs and DAMPs on invading bacteria. NOD1 and NOD2 recognise bacteria cell wall peptides, diaminopimelic-acid containing peptide and muramyl dipeptide respectively. Activation of TLR and NODs results in downstream activation of NFKB. This leads to epithelial cells releasing proinflammatory cytokines IL-1 and IL-6 and chemo-attractants CXCL8, CXCL1,

CCL1 and CCL2 which attract neutrophils and macrophages and CCL20 which attracts dendritic cells. The inflammasome is activated resulting in the release of caspase1 which cleaves pro-IL-1 and pro-IL-18 to produce IL-1 and IL-18. Bacteria that enter the epithelial cell cytoplasm may be ubiquinated which attracts the phagophore, forming an autophagosome. Fusion with the lysosome leads to its destruction. NOD-1 and NOD-2 promote autophagy[331].

IgA plays an important role in balancing host response to the microbiota. IgA binds to microbials preventing adherence to the cell epithelium and neutralizes toxins[324]. Once microbials have entered the epithelium, IgA can bind them in the endosome and the complex is subsequently re-exported into the gut lumen. IgA bound antigen can bind to Dectin-1 on M cells. The complex is subsequently taken up by DC- SIGN receptor on dendritic cells resulting in the production of anti-inflammatory cytokine IL-10. IgA's inability to fix complement also prevents inflammation[2].

5.2 Crohn's disease

Inflammatory bowel disease (IBD) affects 1 in 200 individuals with its global prevalence increasing since 2000[332]. It can be further broken up into two distinct disorders, Crohn's disease (CD) and Ulcerative colitis (UC). CD is characterised by intestinal skip lesions with are transmural in nature and can affect any region of the gastro-intestinal tract. Intestinal complications of CD include strictures, fistulas and abscess. UC on the other hand, affects only the colon and causes superficial, continuous regions of inflammation that extend proximally in a contiguous manner[333]. The combination of environmental factors coupled with host susceptibility and an aberrant immune response to the microbiome results in IBD (Fig 5.4).

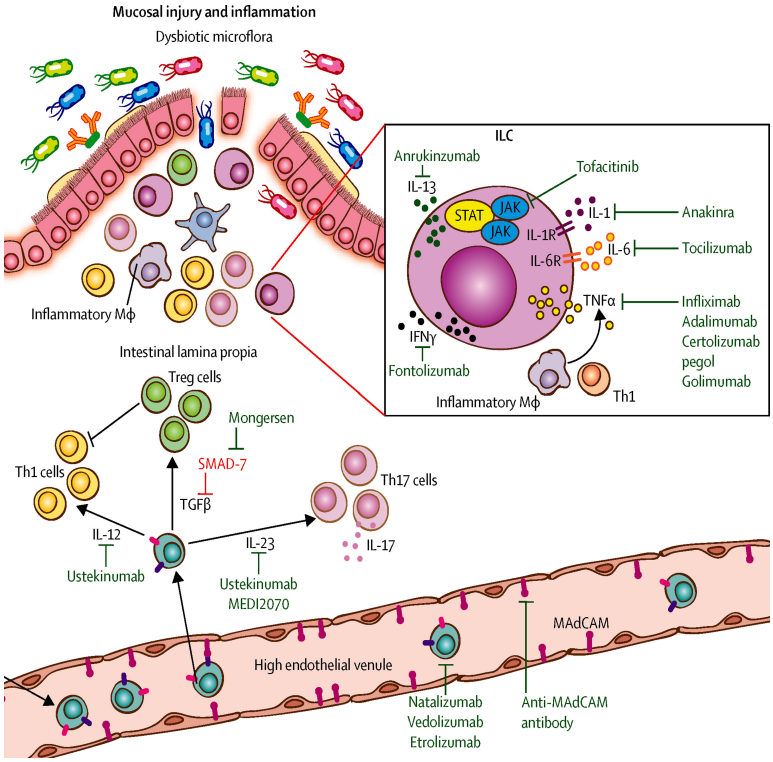


Fig 5.4 Immunopathology of Crohn's disease[323]

5.2.1 Genetic Factors

There is a 50% concordance of CD in monozygotic twins. Over 200 loci have been identified as CD risk loci but individually only modestly increase risk with odds ratios of 1.1-1.2[334]. Key at risk loci in Caucasian populations are *NOD2*[335,336] which decreases ability to kill *bacteria,*

ATG16L1 an autophagy gene and *IL23R.* However, in Asian populations, *TNFSF15* is the predominant risk locus[337].

5.2.2 Intestinal barrier

The intestinal barrier is the first line of defence as stated above. The small intestine contains a single layer of epithelial cells covered by a mucus biofilm which is secreted by goblet cells. It further is protected by antibacterial mediators and IgA. This is in contrast to the colon where the mucus layer is thicker and can be further divided into an inner layer which is impenetrable to bacteria and an outer layer which serves as a reservoir for distinct bacteria[338]. Comparisons of mucus gene expression derived from healthy individuals and patients with CD ileal disease reveals a decrease in MUC1 mRNA and comparisons between inflamed and uninflamed regions revealed a decrease in MUC3, MUC4 and MUC5B[339].

A defect in barrier components including junctional proteins, production of antimicrobial peptides and mucin increases permeability and the translocation of luminal antigens to the lamina propria. In CD, a decrease in claudin 5 and 8 which are sealing tight junctional proteins and an increase in pore-forming claudin 2 is observed[340]. Polymorphisms in NOD2(nucleotide-binding oligomerization domain 2), ATG16L1, IRGM and LRRK2 cause abnormalities in the secretory ability of Paneth cells[341]. 300T→A variant of the ATG16L1 results in fewer and functionally impaired granules secreted by Paneth cells. Autophagy gene ATG16L1 plays an important role in determining the tolerance of endoplasmic reticulum stress in intra-epithelial cells. Endoplasmic stress elicits a pathological unfolded protein response. ATG16L1 determines the activation level of inositol-requiring enzyme 1α (IRE1α), an unfolded protein response sensor. Hyperactivation of IRE1α leads to spontaneous ileitis in mice[342]. NOD2 senses conserved motifs on bacterial peptidoglycans and regulates the secretion of AMPs. NOD2 polymorphisms are strongly associated with Crohn's disease secondary to impaired ability to respond appropriately to microbial products [341].

5.2.3 Microbial dysbiosis

In CD there is a reduction in the number, diversity and richness of microbial species[343]. In particular, a decrease in Bacteriodetes such as *Bacteroides thetaiotaomicron* and the Clostridia class of Firmicutes and a concurrent increase in Gamma proteobacteria and Actinobacteria is reported [344]. A direct causal relationship between dysbiosis and IBD has not been proven and fecal transplants have had varying results[332]. However, in support of dysbiosis causing pathology, it has been shown that fecal material from patients with IBD resulted in an increased susceptibility to colitis in germ-free mice when compared with mice receiving fecal material from healthy donors with an increase in Th17 and a decrease in Treg cells being observed[345]. Adherent–invasive Escherichia coli (AIEC) have been implicated as a pathogenic bacterium in Crohn's disease. AIEC enter intra-epithelial cells by binding to carcinoembryonic antigen-related cell adhesion molecule 6 (CEACAM6) and subsequently reside in host macrophages driving production of pro-inflammatory cytokines and contribute to granuloma formation[346].

5.2.4 Adaptive immune response

An imbalance is present between effector T cells (controlled by transcription factor RORyT) and regulatory T cells (FOXP3)(Baumgart and Sandborn, 2012). Effector CD4 T cells are implicated in the pathology of IBD with excessive T helper 1 (Th1) and Th17 cell responses post stimulation by antigen presenting cells (presenting microbiota)[347]. IL-12 secreted by antigen presenting cells promotes differentiation into Th1 cells via STAT4[348] whilst IL-23 enhances Th17 differentiation[349]. Th1 and 17 cells in response secrete pro-inflammatory cytokines IFN-y, IL-17, TNFa and IL-22.

IL-12 is comprised of interleukin-12p35 and interleukin-12p40 subunits whilst IL-23 is comprised of interleukin-23p19 and interleukin-12p40 subunits[349]. Antibodies targeting anti–interleukin-12p40 allows dual blockage of both cytokines and are used in IBD. JAK inhibitors block downstream signalling of IL-12 and IL-23[350]. Blockage of IL-17 have not been effective in CD[347]. Tregs secrete IL-10, TGFB and IL-35. Studies are underway to upregulate Treg cells via IL-2[333].

The healthy gut mucosa is dominated by IgA which acts to limit microbial infiltration[351]. A dysregulated humoral response in inflammatory bowel disease (IBD) is supported by findings of increased IgG B cells[352,353] and anti-commensal antibodies[354]. An IgG+ plasma cell module was identified in patients refractory to anti-TNF therapy on scRNAseq of ileal biopsies in Crohn's disease[352]. IgG is potentially pro-inflammatory with the ability to fix complement and activate immune cells via Fcy receptors. *FCGR2A-R131* variant is protective against ulcerative colitis (UC), where the amino acid substitution lowers IgG binding affinity[355]. Furthermore, there is an increase in the activating to inhibitory ratio of FcgR mucosal immune cells in UC, lowering the threshold for activation by IgG which induces production of IL-1B in macrophages, a Th17 polarising cytokine[354]. In IBD, there is increased agalactosylated IgG in the serum which favour binding to activating FcγR and correlates with disease severity whilst serum IgG sialylation which is anti-inflammatory is reduced in Crohn's disease[356]. B cell clonal expansion with heightened use of auto-reactive heavy chain VH4-34 is evident in Crohn's disease[278]. Thus multiple lines of evidence suggest that B cells play a prominent role in the pathology of IBD. However, how B cells and their receptors may differ and contribute to IBD pathology remains unknown.

The BCR denotes the unique clonal identity of a B cell. Generation of the BCR first occurs during B cell development in the bone marrow with rearrangement of the immunoglobulin receptor genes, and then undergoes further diversification via somatic hypermutation (SHM) and class-switching recombination (CSR) in secondary lymphoid organs. The BCR repertoire refers to the range of individual BCRs that collectively provide the diversity of antigen receptors required by B cells to recognise antigens. The hypervariable complementarity-determining region 3 (CDR3) of the BCR is formed by the combination of V, D and J genes and is a key antigen binding determinant and thus informative when assessing for a shared antigen driven response. Using high resolution analysis of the BCR repertoire from blood and lymph nodes (LN) we identified clonal B cell populations unique to Crohn's disease LNs, local to sites of inflammation.

5.3 Results

5.3.1 Sample overview

We analysed the BCR repertoire from five independent cohorts.

Lymph Nodes

BCR repertoires were generated from 24 individuals with active Crohn's disease requiring small bowel resection. Samples were taken from mesenteric lymph nodes adjacent to areas of inflamed regions of bowel.

Post-mortem samples

A publicly available BCR dataset of 8 post-mortem individuals from 4 sites including the mediastinal LN (MDLN), spleen (SPL), mesenteric LN (MSLN) and peripheral blood (PBMC) was accessed [269].

PBMC

BCR repertoires were generated from the PBMCs of 24 individuals with active Crohn's disease who were off medications and 29 healthy controls.

Plasmablasts

BCR repertoires were generated from circulating plasmablasts of 24 individuals with active Crohn's disease who were off medications, 26 individuals with active UC who were off medications and 29 healthy controls.

5.3.2 Convergent Clones

We wished to assess for the presence of public clones indicative of shared antigen in Crohn's disease. Clones were considered "convergent" across patients if there was sharing of V and J genes with an identical CDR-H3 region length and an 85% amino acid CDR-H3 sequence homology (Fig 5.5A). We calculated the number of shared clones pairwise within a group where a high level of overlap is suggestive of a shared epitope (Fig 5.5B). In keeping with the expected high diversity of naive B cells there was minimal clonal sharing in blood.

However, there was considerable clonal sharing amongst repertoires generated from LN localised to areas of inflammation in Crohn's disease. This suggested Crohn's specific antigens generated from inflamed bowel are common amongst Crohn's disease patients and may contribute to B cell mediated disease pathology. To further assess the nature of these clones, we selected clones present in two or more Crohn's disease LN that were absent in post-mortem mesenteric LN (MSLN) (Fig 5.5A). The top ten most frequently identified clones were common to at least seven patients (Fig 5.5C). These "Crohn's specific clones" were enriched in all immunoglobulin isotypes, but with a greater proportion of class-switched isotypes suggesting that antigen-specific germinal centre derived B cell maturation has occurred (Fig 5.5D).

Antibody secreting cells generated in Peyer's patches and the MSLN do not immediately localise to gut tissue but systemically recirculate before homing back to the gut via expression of gut-specific addressins and lymphocyte receptors including MAdCAM and CCR9 and the a4b7 integrin respectively[321]. We were able to identify "Crohn's specific clones" derived from the LN in the peripheral blood mononuclear cells (PBMCs) BCR repertoire in an independent cohort with significant enrichment in class-switched Crohn's disease samples (Fig 5.5E).

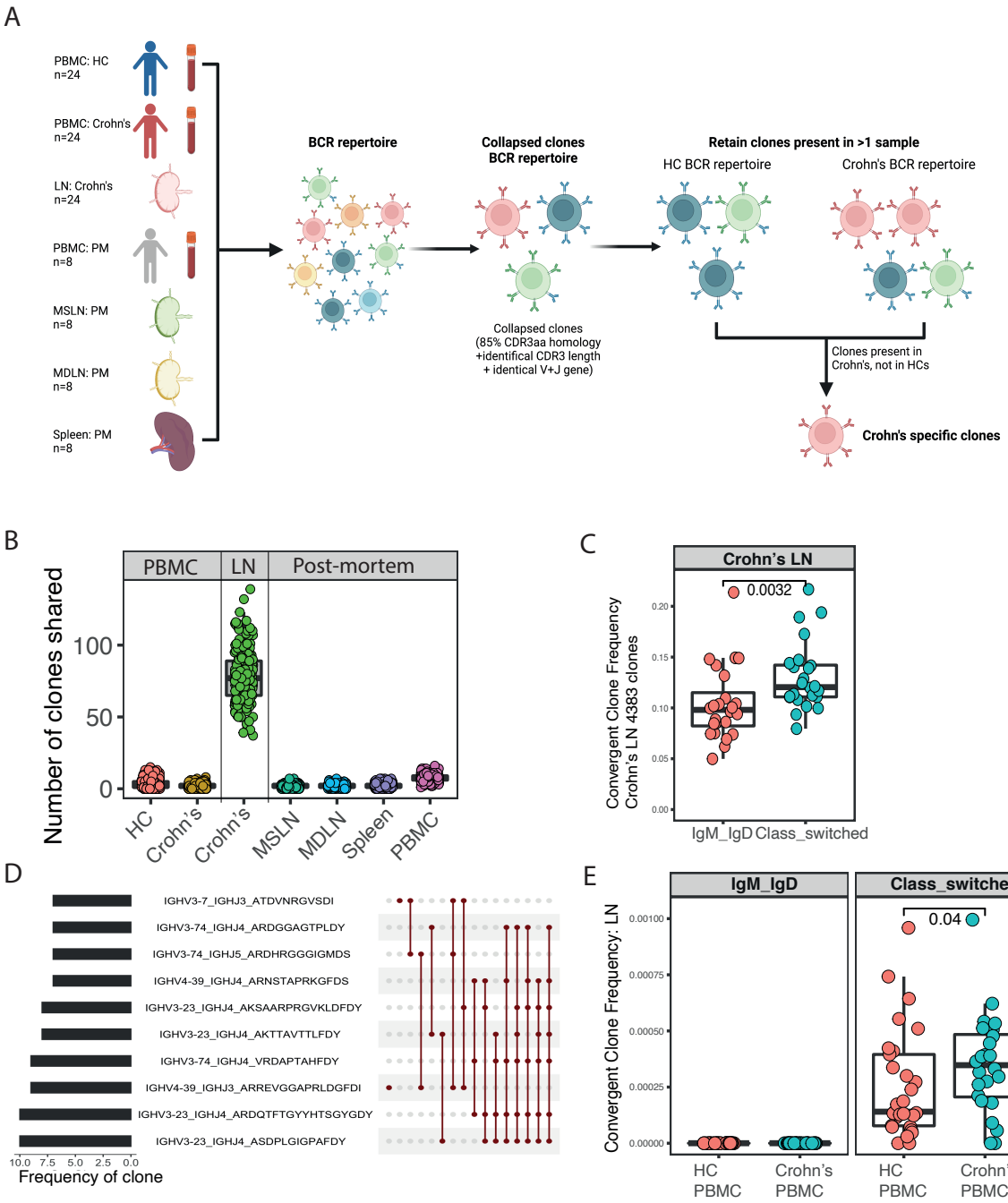


Fig 5.5 Public Clones in Crohn's disease in LN A. Schematic of convergent clones. Peripheral blood mononuclear cells (PBMC), Lymph nodes (LN), Healthy Controls (HC), mediastinal lymph nodes (MDLN), mesenteric lymph nodes (MSLN), post-mortem (PM) B. Assessment of clonal convergence. Randomly selected 8 patients per group and 1500 unique clones per patient (200 iterations). Boxplots of number of clones shared in at least two patients split according to group. Each dot represents an iteration C. Identified potentially Crohn's specific LN clones. Clones were identified based on presence in two or more Crohn's inflamed LN samples and absence in MSLN from the post-mortem cohort. Boxplot of convergence in IGHM-IGHD and class-switched clones in Crohn's LN BCR repertoire. Each dot represents a sample. Unpaired wilcox test. D. Dominant clones: Ten most frequent Crohn's clones identified in LN E. Validation of Crohn's specific LN clones: Boxplots of enrichment. Each dot represents a sample. One-sided wilcox test.

"Crohn's specific clones" were independently derived from PBMCS. They were defined as being present in 2 or more Crohn's samples and absent in HC. Similar to the LN, these clones were enriched in the class-switched clones (Fig 5.6A). Overall, the top ten clones were less frequent to that seen in the lymph nodes as predicted (Fig 5.6B). We validated the "Crohn's specific clones" identified from the PBMC BCR repertoire in the LN BCR repertoire with significant enrichment in the class-switched Crohn's LN BCR compared with the post-mortem MSLN BCR (Fig 5.6C).

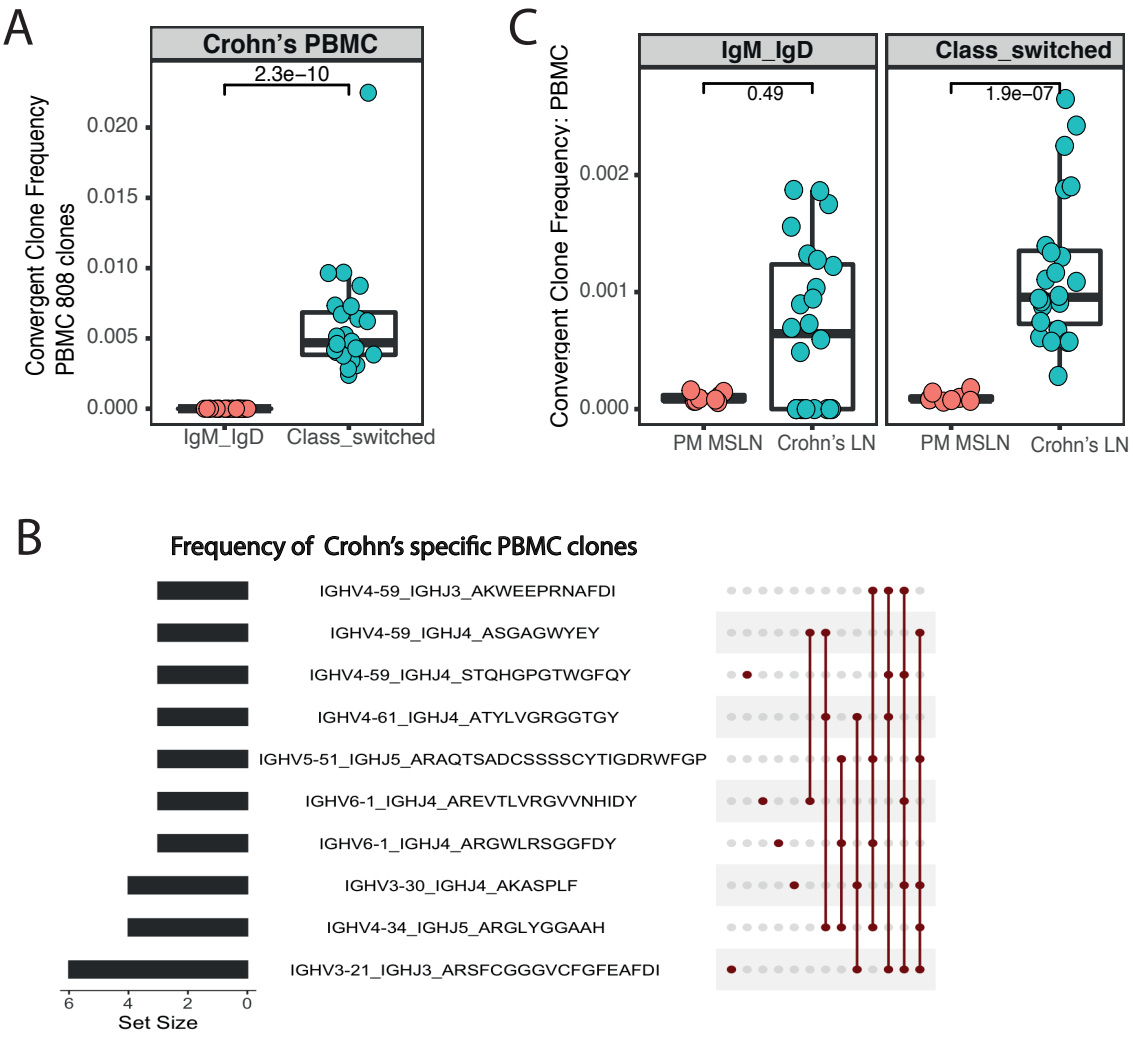


Fig 5.6 Public Clones in Crohn's disease in PBMCs A. Identified potentially Crohn's specific PBMC clones. Clones were identified based on presence in two or more Crohn's PBMC samples and absence in health. Boxplot of convergence in IGHM-IGHD and class-switched clones in Crohn's LN BCR repertoire. Unpaired wilcox test.B. Dominant clones: Ten most frequent Crohn's clones identified in PBMC. C. Validated of Crohn's specific PBMC clones: Boxplots of enrichment of "Crohn's specific clones" in LN dataset split according to isotype. Unpaired wilcox test.

In CD patients, an increase in both serum IgA and IgG1 immunoglobulin titres relative to health were present (Fig. 5.7A). To determine whether this might be driven by shared antigens, we generated BCR repertoires from plasmablasts isolated during active disease in another independent cohort of patients. Plasmablasts are short-lived and therefore are expected to be enriched for clones/sequences that are relevant to the current inflammatory response. We found that the CD-specific clones were uniquely found in CD plasmablasts, were more enriched compared to CD PBMCs, and were not present in either health or in an additional cohort of UC patients, confirming the CD specificity of the clones (Fig. 5.7B). While no single clone was shared between all CD patients, CD-specific clones were shared between multiple cases with convergence in IGHM, IGHA1/A2 and IGHG2 isotypes (Fig. 5.7C). Using the same clustering technique, we demonstrated shared clones in the UC plasmablast repertoire in keeping with recent findings[357]. These clones were not enriched in CD patients any more than in idiopathic pulmonary fibrosis, a disease control. This further highlights the difference in B cell antigenic responses in CD and UC (Fig. 5.7D).

In IBD, a global reduction in SHM was present across all class-switched isotypes when compared to healthy controls (Fig. 5.7E), as has been reported following infection[237,358]. This decrease in SHM may be attributable to an extrafollicular response secondary to pathogen-associated molecular pattern recognition and/or recently recruited clones from the naïve repertoire expanding in the germinal centre. When comparing SHM of CD-specific clones with that of healthy-clones defined by their presence in >7 HC, within the CD plasmablast repertoire, CD-specific clones had increased SHM across all isotypes (Fig. 5.7F) providing further evidence of antigen-targetted immunity given the occurrence of affinity maturation.

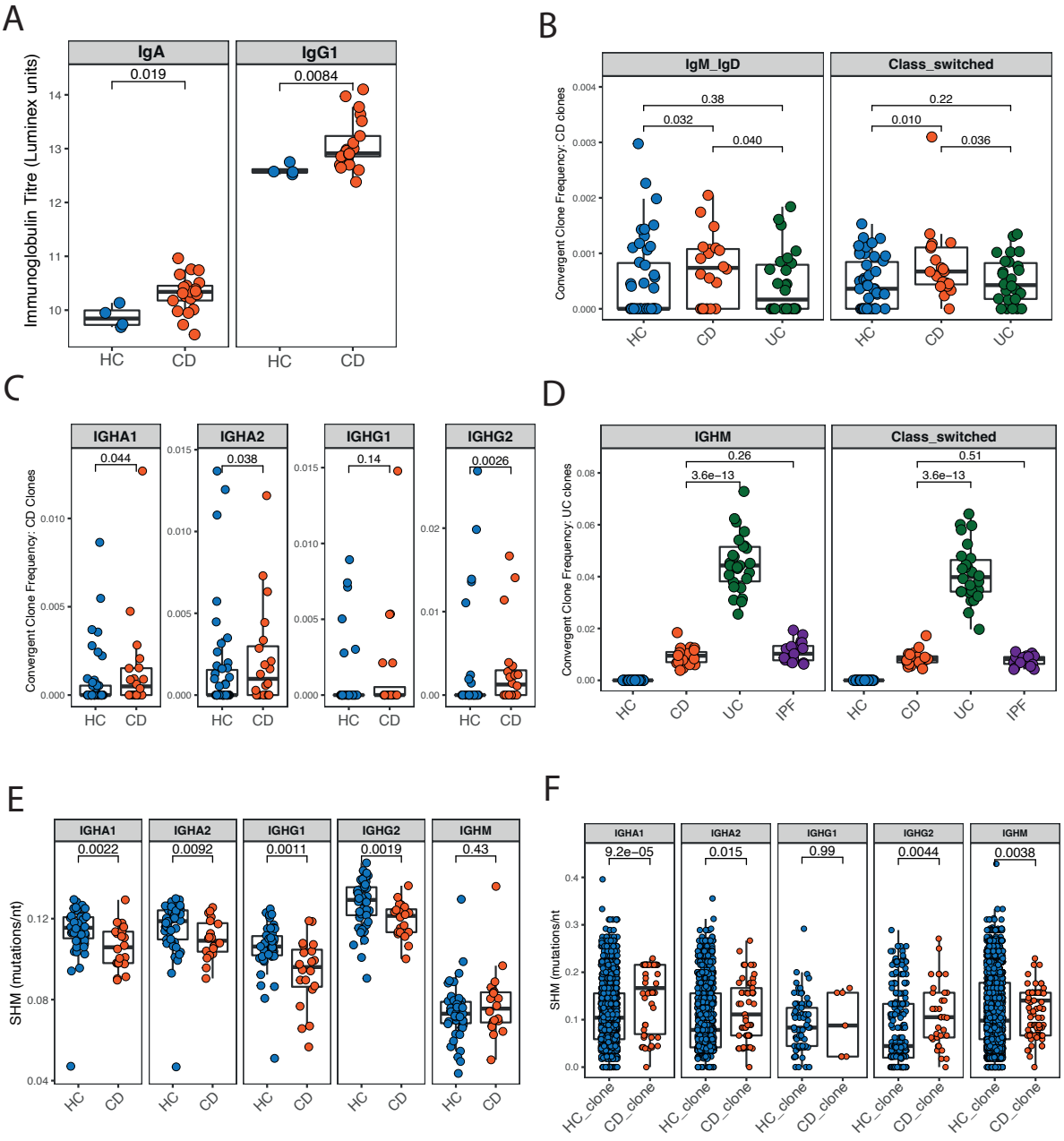


Fig 5.7 Plasmablast BCR repertoire in IBD A. Immunoglobulin titres: Boxplot of immunoglobulin titres split according to isotype and disease. Immunoglobulin titres in healthy individuals (n = 4), patients with CD (n = 20). Unpaired t test. B. Validation of CD specific clones: Assessed clonal convergence of CD specific clones derived from LN in plasmablasts. Boxplot representing convergence split according to isotype and disease. Each dot represents a sample. Unpaired one-tailed wilcox test. C. Validation of CD specific clones per isotype: Assessed clonal convergence of CD specific clones derived from LN in plasmablasts. Boxplot representing convergence split according to sub-isotype and disease. Each dot represents a sample. Unpaired one-tailed wilcox test. D. Shared clones in UC: Clones were identified based on presence in two or more UC plasmablast samples and absent in health. Boxplot of convergence in IGHM and class-switched clones in HC, UC, CD and Idiopathic pulmonary fibrosis (IPF) BCR repertoire. Each dot represents a sample. Unpaired wilcox test. E. SHM in plasmablasts: BCR repertoire in plasmablasts (CD19+IgD-CD27+CD24-CD38+). Boxplot of SHM split according to isotype and disease. Each dot represents a sample. Unpaired wilcox test. F. SHM of shared clones: Boxplot of SHM comparing HC clones and CD specific clones in CD patients split according to isotype. HC clones defined as being present in >7 HC. Each dot represents a unique clone per patient. Unpaired wilcox test. For b-f, n = 46 for healthy individuals and n = 20, n = 26 and n = 19, for patients with CD, UC and IPF respectively.

This study of the BCR repertoire in intestinal LNs and blood demonstrates the presence of CD-specific clones that are shared between multiple patients and present in both LNs and circulating plasmablasts during active disease. These clones are not seen in either healthy controls or during active UC. Their sharing between multiple patients suggests the presence of common, disease-specific antigens as opposed to a non-specific polyclonal activation of B cells. This not only provides further support for the role of pathogenic B cells in CD but provides the opportunity – providing epitope-specific antibodies can be fully elucidated – for a diagnostic antibody test for CD.

6. Future Directions

6.1 B cell receptor repertoire

BCR analysis of whole blood provides an understanding of the global changes in the repertoire including changes in isotype and V gene use, SHM and diversity. To gain an understanding on a disease specific level, we used clustering methods to identify shared sequences and used publicly available databases to functionally annotate sequences.

Our key findings on analysis of the BCR repertoire in SARS-CoV-2 infected individuals were,

- An increase in the proportion of BCRs bearing IgG1/3 and IgA1 isotypes;
- A decrease in SHM in class-switched isotypes whilst an increase occurred in IGHD and IGHM;
- A decrease in diversity;
- The generation of a broad distribution of SARS-CoV-2-specific clones predicted to target the spike protein.

Our key findings on analysis of the BCR repertoire in Crohn's disease were,

- Greater sharing of clones amongst Crohn's patients in inflamed LN compared with post-mortem LNs suggesting the presence of public clones;
- "Crohn's specific clones" were mostly class-switched in the LN and in PBMCs;
- Raised IgA and IgG immunoglobulin titres in Crohn's, further suggesting involvement of the humoral arm;
- Disease specificity with minimal enrichment in UC on assessment of plasmablasts;
- A global decrease in SHM occurred in class-switched plasmablasts clones in Crohn's;
- IgG2 and IgM "Crohn's specific clones" were highly mutated compared with clones that were shared with health.

Future work would involve the following:

- Performing BCR on antigen specific B cell subpopulations
- Performing single cell analysis to obtain paired light and heavy chain information of clones
- Functionally working up antibodies using techniques such as,

- Cloning
- Phage display

6.2 Functional assessment

6.2.1 Cloning

Cloning involves synthesizing double-stranded DNA fragments from heavy and light chain sequences and cloning these sequences into an expression vector. Antibody supernatants are harvested, and antibody purified. The antibody can be tested for binding to a specific antigen of interest such as SARS-CoV-2 RBD.

6.2.2 Phage display

Phage display is a technique that can be used to screen multiple antibodies against an antigen of interest. Lysogenic filamentous bacteriophages are modified to display polypeptides on their surface. This is achieved by the insertion of the DNA fragment of interest into the filamentous phage coat protein gene. The M13 Bacteriophage is commonly used. It infects Escherichia coli expressing the F pilus. This is because the phage coat protein needs to bind with the tip of the F pilus to enter the cell. Once it has entered the host cell, it continuously releases new phages.

The M13 phage genome is 6407 base pairs long and is coded by a single strand of DNA. M13 infects E.Coli through the binding of its G3P-N2 domain coat protein to the tip of a F pilus. Upon binding, a conformational change occurs facilitating binding of co-receptor G3P-N1 domain to TolA. This leads to injection of the phage genome into the bacterium. The single stranded DNA of the M13 phage is converted into double stranded DNA and takes on a super coiled structure known as the replicative form. Further replication occurs. With sufficient G5P, a DNA-binding protein is formed and the replicative formation is prevented. G5P then binds with the ssDNA changing the confirmation from a circular form to a rod-shaped form. G5P coats the entire sequence except for hairpin ends. A pore is formed in the membrane and the phage genome exits whilst its coat is assembled[359].

The M13 Phage is used to display antibody on its coat and enables the simultaneous linkage of protein and genomic sequence. This is achieved by incorporating the gene encoding the protein of interest in the M13 Phage genome abutting gene pIII(encodes G3P). This leads to the expression of the protein along with the phage coat protein. Antibody sequences are cloned in "recombinant antibody formats" as smaller fragments are more amenable to cell wall expression and can include segments from both the heavy and light chains. A modified form of the M13 phage is used to increase efficiency. The modified M13 phage does not contain all the required genomic sequences and co-infection of E.Coli with a "helper phage" is necessary.

Once the antibody phage display library is generated, a process of "phage biopanning" occurs to identify antigen binding. Purified antigens need to be immobilised on a solid surface such as beads, column matrices etc. Blocking agents are then used to block the remaining sites on the solid surface, thus preventing non-specific binding. The phage library is exposed to the immobilized antigens and unbound antibody is washed away and the high affinity-phage antibodies finally eluted. Antigen specific binding is confirmed with ELISA. Positive clone sequences can be derived from sequencing[359,360].

6.3 Future work

6.3.1 SARS-CoV-2

Extensive work has been performed on the humoral response to SARS-CoV-2. An interesting finding that goes beyond COVID-19 is the findings of increased SHM post infection with SARS-CoV-2 and vaccination with Influenza in IGHD. Enriching for IGHD positive cells on flow sorting and performing single cell analysis with concurrent cell surface labelling (CITEseq) would allow further characterisation of these cells of interest.

6.3.2 Public clones in Crohn's disease

Using cloning and phage display, sequences identified as public clones can be expressed as antibodies and antigen binding determined. Extensive work has already been performed showing the pathogenic role of anti-commensal IgG[361].

6.4 Future of BCR repertoire

Bulk BCR repertoire analysis is the ideal tool to study SHM, class-switching and diversity. We can generate sequences of such quality to confidently attribute deviations from the germline to mutations rather than sequencing error. The use of UMIs allows us to correct not only for PCR bias but also generate a consensus sequence adding confidence to the sequence generated. Class-switching requires identifying sequencing with shared VDJ sequences but different constant regions. This requires accuracy in sequencing calling and sufficient depth. To increase the accuracy, generating isotype specific libraries rather than pooling all isotype primers together would reduce barcode switching and over calling of class-switching. Diversity metrics are similar to class-switching requiring accurate VDJ calling and depth.

A limitation of BCR repertoire analysis is the quantification of isotype and V gene usage. As we need to correct for library depth, all metrics are proportions. In BCR repertoire analysis, if there is an uneven expansion in absolute numbers this will appear as a proportional reduction of the less expanded group. For example, in a healthy lymph node, is there were a total of 32 IgA cells and 16 IgG cells and if we sampled 21 cells, we would expect that 14 would be IgA, and 7 would be IgG. Assuming that there is an equal amount of mRNA produced by each cell, the BCR repertoire isotype usage would be 66% IgA and 33% IgG. If in an inflamed lymph node, there were a total of 48 IgA cells and 36 IgG cells and if we once again sampled 21 cells, we would expect 12 cells would be IgA and 9 cells would be IgG. Thus, the BCR isotype usage would be 57% IgA and 43% IgG. On comparing the two samples, an increase in IgG and a decrease in IgA appears to occur despite both being increased. This highlights the importance of not over interpreting repertoire proportions and instead using methods such as qPCR. Both bulk RNAseq and single cell sequencing are plagued with the same issue of correcting for library depth and thus each gene expression value not being independent of the other. To overcome this, normalisation techniques are used beyond correcting for library depth such as DESeq2. A similar technique needs to be adopted by the BCR community in analysis of V gene and isotype.

A further limitation in BCR repertoire analysis is library depth when identifying specific clones of interest. In our analysis, we achieved sequencing depths ranging from 5000-20000 unique clones. In the setting of acute infection where clones of interest are expanded, this depth appears adequate. However, when trying to track clones over time once disease recovery has occurred this proved to be difficult due to contraction in size. Sequencing methods need to be developed that allows larger depths at low cost whilst allowing long reads at 250-300 base pairs.

Bulk BCR repertoire analysis needs to be tailored to the correct scientific question. Using bulk BCR repertoire analysis on an unsorted population where the antigen is unknown is where it is best suited. However, where the antigen is known, such as is SARS-CoV-2 an antigen sorted population is more informative especially where longitudinal analysis is desirable.

References

1. Burnet, F. M. A modification of Jerne's theory of antibody production using the concept of clonal selection. *CA Cancer J Clin* **26**, 119–121 (1976).
2. Murphy, K., Weaver, C. & Janeway, C. *Janeway's immunobiology*. (2017).
3. Schatz, D. G. & Ji, Y. Recombination centres and the orchestration of V(D)J recombination. *Nature Reviews Immunology 2011 11:4* **11**, 251–263 (2011).
4. Yaari, G. & Kleinstein, S. H. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Medicine 2015 7:1* **7**, 1–14 (2015).
5. Padlan, E. A. Anatomy of the antibody molecule. *Mol Immunol* **31**, 169–217 (1994).
6. Early, P., Huang, H., Davis, M., Calame, K. & Hood, L. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. *Cell* **19**, 981–992 (1980).
7. Edelman, G. M. Antibody structure and molecular immunology. *Science* **180**, 830–840 (1973).
8. Matsuda, F. & Honjo, T. Organization of the Human Immunoglobulin Heavy-Chain Locus. **62**, 1–29 (1996).
9. Ban, N., Day, J., Wang, X., Ferrone, S. & McPherson, A. Crystal structure of an anti-anti-idiotype shows it to be self-complementary. *J Mol Biol* **255**, 617–627 (1996).
10. Davies, D. R. & Cohen, G. H. Interactions of protein antigens with antibodies. *Proc Natl Acad Sci U S A* **93**, 7–12 (1996).
11. Davies, D. R. & Metzger, H. Structural Basis of Antibody Function. *https://doi.org/10.1146/annurev.iy.01.040183.000511* **1**, 87–117 (2003).
12. Clackson, T. & Wells, J. A. A hot spot of binding energy in a hormone-receptor interface. *Science* **267**, 383–386 (1995).
13. Bhat, T. N. *et al.* Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proc Natl Acad Sci U S A* **91**, 1089 (1994).
14. Melchers, F. *et al.* Repertoire selection by pre-B-cell receptors and B-cell receptors, and genetic control of B-cell development from immature to mature B cells. *Immunological Reviews* **175**, 33–46 (2000).
15. Lewis, S. M. The Mechanism of V(D)J Joining: Lessons from Molecular, Immunological, and Comparative Analyses. *Advances in Immunology* **56**, 27–150 (1994).
16. Jung, D. & Alt, F. W. Unraveling V(D)J Recombination: Insights into Gene Regulation. *Cell* **116**, 299–311 (2004).
17. Swanson, P. C. The bounty of RAGs: recombination signal complexes and reaction outcomes. *Immunological Reviews* **200**, 90–114 (2004).
18. Fugmann, S. D., Lee, A. I., Shockett, P. E., Villey, I. J. & Schatz, D. G. The RAG Proteins and V(D)J Recombination: Complexes, Ends, and Transposition. *http://dx.doi.org/10.1146/annurev.immunol.18.1.495* **18**, 495–527 (2003).
19. Weigert, M. *et al.* The joining of V and J gene segments creates antibody diversity. *Nature 1980 283:5746* **283**, 497–499 (1980).
20. Moshous, D. *et al.* Artemis, a novel DNA double-strand break repair/V(D)J recombination protein, is mutated in human severe combined immune deficiency. *Cell* **105**, 177–186 (2001).
21. Weigert, M., Gatmaitan, L., Loh, E., Schilling, J. & Hood, L. Rearrangement of genetic information may produce immunoglobulin diversity. *Nature* **276**, 785–790 (1978).
22. Komori, T., Okada, A., Stewart, V. & Alt, F. W. Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes. *Science* **261**, 1171–1175 (1993).

23. Blattner, F. R. & Tucker, P. W. The molecular biology of immunoglobulin D. *Nature* **307**, 417–422 (1984).
24. Early, P. *et al.* Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**, 313–319 (1980).
25. Phillips, C., Jung, S. & Gunderson, S. I. Regulation of nuclear poly(A) addition controls the expression of immunoglobulin M secretory mRNA. *The EMBO Journal* **20**, 6443 (2001).
26. van Egmond, M. *et al.* IgA and the IgA Fc receptor. *Trends in Immunology* **22**, 205–211 (2001).
27. Nimmerjahn, F. & Ravetch, J. v. Fc-receptors as regulators of immunity. *Adv Immunol* **96**, 179–204 (2007).
28. Sharp, T. H. *et al.* Insights into IgM-mediated complement activation based on in situ structures of IgM-C1-C4b. *Proc Natl Acad Sci U S A* **116**, 11900–11905 (2019).
29. Hughes-Jones, N. C. & Gardner, B. Reaction between the isolated globular sub-units of the complement component C1q and IgG-complexes. *Mol Immunol* **16**, 697–701 (1979).
30. Palmeira, P., Quinello, C., Silveira-Lessa, A. L., Zago, C. A. & Carneiro-Sampaio, M. IgG placental transfer in healthy and pathological pregnancies. *Clin Dev Immunol* **2012**, (2012).
31. Funk, P. E. & Witte, P. L. Enrichment of primary lymphocyte-supporting stromal cells and characterization of associated B lymphocyte progenitors. *European Journal of Immunology* **22**, 1305–1313 (1992).
32. Alt, F. W. *et al.* Ordered rearrangement of immunoglobulin heavy chain variable region segments. *The EMBO Journal* **3**, 1209 (1984).
33. Lebien, T. W. & Tedder, T. F. B lymphocytes: how they develop and function. *Blood* **112**, 1570–1580 (2008).
34. Burrows, P., Lejeune, M. & Kearney, J. F. Evidence that murine pre-B cells synthesise mu heavy chains but no light chains. *Nature* **280**, 838–841 (1979).
35. Sakaguchi, N. & Melchers, F. λ5, a new light-chain-related locus selectively expressed in pre-B lymphocytes. *Nature 1986 324:6097* **324**, 579–582 (1986).
36. Kudo, A., Sakaguchi, N. & Melchers, F. Organization of the murine Ig-related lambda 5 gene transcribed selectively in pre-B lymphocytes. *The EMBO Journal* **6**, 103–107 (1987).
37. Kudo, A. & Melchers, F. A second gene, VpreB in the lambda 5 locus of the mouse, which appears to be selectively expressed in pre-B lymphocytes. *The EMBO Journal* **6**, 2267–2272 (1987).
38. Ohnishi, K. & Melchers, F. The nonimmunoglobulin portion of lambda5 mediates cell-autonomous pre-B cell receptor signaling. *Nat Immunol* **4**, 849–856 (2003).
39. Meixlsperger, S. *et al.* Conventional light chains inhibit the autonomous signaling capacity of the B cell receptor. *Immunity* **26**, 323–333 (2007).
40. Grawunder, U. *et al.* Down-regulation of RAG1 and RAG2 gene expression in preB cells after functional immunoglobulin heavy chain rearrangement. *Immunity* **3**, 601–608 (1995).
41. Osmond, D. G., Rolink, A. & Melchers, F. Murine B lymphopoiesis: towards a unified model. *Immunol Today* **19**, 65–68 (1998).
42. Mostoslavsky, R. *et al.* κ chain monoallelic demethylation and the establishment of allelic exclusion. *Genes & Development* **12**, 1801 (1998).
43. Löffert, D., Ehlich, A., Müller, W. & Rajewsky, K. Surrogate light chain expression is required to establish immunoglobulin heavy chain allelic exclusion during early B cell development. *Immunity* **4**, 133–144 (1996).

44. Meffre, E., Casellas, R. & Nussenzweig, M. C. Antibody regulation of B cell development. *Nature Immunology 2000 1:5* **1**, 379–385 (2000).
45. Nemazee, D. & Buerki, K. Clonal deletion of autoreactive B lymphocytes in bone marrow chimeras. *Proc Natl Acad Sci U S A* **86**, 8039 (1989).
46. Luning Prak, E. & Weigert, M. Light chain replacement: a new model for antibody gene rearrangement. *J Exp Med* **182**, 541–548 (1995).
47. Platt, J. L., Garcia de Mattos Barbosa, M. & Cascalho, M. The five dimensions of B cell tolerance. *Immunological Reviews* **292**, 180–193 (2019).
48. Nemazee, D. Mechanisms of central tolerance for B cells. *Nature Reviews Immunology 2017 17:5* **17**, 281–294 (2017).
49. Victora, G. D. & Nussenzweig, M. C. Germinal Centers. *https://doi.org/10.1146/annurev-immunol-020711-075032* **30**, 429–457 (2012).
50. Cyster, J. G., Hartley, S. B. & Goodnow, C. C. Competition for follicular niches excludes self-reactive cells from the recirculating B-cell repertoire. *Nature 1994 371:6496* **371**, 389–395 (1994).
51. Loder, F. *et al.* B cell development in the spleen takes place in discrete steps and is determined by the quality of B cell receptor-derived signals. *J Exp Med* **190**, 75–89 (1999).
52. Palanichamy, A. *et al.* Novel human transitional B cell populations revealed by B cell depletion therapy. *J Immunol* **182**, 5982–5993 (2009).
53. Sagaert, X., Sprangers, B. & de Wolf-Peeters, C. The dynamics of the B follicle: understanding the normal counterpart of B-cell-derived malignancies. *Leukemia 2007 21:7* **21**, 1378–1386 (2007).
54. Williams, G. T., Peaker, C. J. G., Patel, K. J. & Neuberger, M. S. The alpha/beta sheath and its cytoplasmic tyrosines are required for signaling by the B-cell antigen receptor but not for capping or for serine/threonine-kinase recruitment. *Proc Natl Acad Sci U S A* **91**, 474–478 (1994).
55. Sanchez, M. *et al.* Signal transduction by immunoglobulin is mediated through Ig alpha and Ig beta. *The Journal of Experimental Medicine* **178**, 1049 (1993).
56. Woyach, J. A., Johnson, A. J. & Byrd, J. C. The B-cell receptor signaling pathway as a therapeutic target in CLL. *Blood* **120**, 1175 (2012).
57. Nimmerjahn, F. & Ravetch, J. v. Fcγ receptors as regulators of immune responses. *Nature Reviews Immunology 2007 8:1* **8**, 34–47 (2008).
58. Bournazos, S. & Ravetch, J. v. Fcγ Receptor Function and the Design of Vaccination Strategies. *Immunity* **47**, 224 (2017).
59. Pincetic, A. *et al.* Type I and type II Fc receptors regulate innate and adaptive immunity. *Nature Immunology 2014 15:8* **15**, 707–716 (2014).
60. Ghetie, V. & Ward, E. S. Multiple roles for the major histocompatibility complex class I- related receptor FcRn. *Annu Rev Immunol* **18**, 739–766 (2000).
61. Turner, H. & Kinet, J. P. Signalling through the high-affinity IgE receptor FcεRI. *Nature 1999 402:6760* **402**, 24–30 (1999).
62. Turula, H. & Wobus, C. E. The Role of the Polymeric Immunoglobulin Receptor and Secretory Immunoglobulins during Mucosal Infection and Immunity. *Viruses* **10**, (2018).
63. Kim, C. H. *et al.* Subspecialization of Cxcr5+ T Cells: B Helper Activity Is Focused in a Germinal Center–Localized Subset of Cxcr5+ T Cells. *The Journal of Experimental Medicine* **193**, 1373 (2001).
64. Schaerli, P. *et al.* CXC chemokine receptor 5 expression defines follicular homing T cells with B cell helper function. *J Exp Med* **192**, 1553–1562 (2000).

65. Schmitt, N., Bentebibel, S. E. & Ueno, H. Phenotype and Functions of Memory Tfh cells in Human Blood. *Trends Immunol* **35**, 436 (2014).
66. McHeyzer-Williams, M. *et al.* Helper T-cell-regulated B-cell immunity. *Microbes Infect* **5**, 205–212 (2003).
67. Kerfoot, S. M. *et al.* Germinal center B cell and T follicular helper cell development initiates in the interfollicular zone. *Immunity* **34**, 947–960 (2011).
68. Pereira, J. P., Kelly, L. M., Xu, Y. & Cyster, J. G. EBI2 mediates B cell segregation between the outer and centre follicle. *Nature 2009 460:7259* **460**, 1122–1126 (2009).
69. Gatto, D., Paus, D., Basten, A., Mackay, C. R. & Brink, R. Guidance of B Cells by the Orphan G Protein-Coupled Receptor EBI2 Shapes Humoral Immune Responses. *Immunity* **31**, 259–269 (2009).
70. Reif, K. *et al.* Balanced responsiveness to chemoattractants from adjacent zones determines B-cell position. *Nature 2002 416:6876* **416**, 94–99 (2002).
71. Choi, Y. S. *et al.* ICOS receptor instructs T follicular helper cell versus effector cell differentiation via induction of the transcriptional repressor Bcl6. *Immunity* **34**, 932–946 (2011).
72. Basso, K. & Dalla-Favera, R. BCL6: Master Regulator of the Germinal Center Reaction and Key Oncogene in B Cell Lymphomagenesis. *Advances in Immunology* **105**, 193–210 (2010).
73. Banchereau. J. *et al.* The CD40 antigen and its ligand. *Annu Rev Immunol* **12**, 881–926 (1994).
74. MacLennan, I. C. M. *et al.* Extrafollicular antibody responses. *Immunol Rev* **194**, 8–18 (2003).
75. Tarlinton, D. & Good-Jacobson, K. Diversity Among Memory B Cells: Origin, Consequences, and Utility. *Science (1979)* **341**, 1205–1211 (2013).
76. Tonegawa, S. *Somatic generation of antibody diversity*. (1983).
77. Nieuwenhuis, P. & Opstelten, D. Functional anatomy of germinal centers. *American Journal of Anatomy* **170**, 421–435 (1984).
78. Allen, C. D. C. *et al.* Germinal center dark and light zone organization is mediated by CXCR4 and CXCR5. *Nature Immunology 2004 5:9* **5**, 943–952 (2004).
79. Victora, G. D. *et al.* Germinal Center Dynamics Revealed by Multiphoton Microscopy with a Photoactivatable Fluorescent Reporter. *Cell* **143**, 592–605 (2010).
80. Gitlin, A. D., Shulman, Z. & Nussenzweig, M. C. Clonal selection in the germinal centre by regulated proliferation and hypermutation. *Nature 2014 509:7502* **509**, 637–640 (2014).
81. Bannard, O. *et al.* Germinal center centroblasts transition to a centrocyte phenotype according to a timed program and depend on the dark zone for effective selection. *Immunity* **39**, 912–924 (2013).
82. Barrington, R. A., Pozdnyakova, O., Zafari, M. R., Benjamin, C. D. & Carroll, M. C. B Lymphocyte Memory Role of Stromal Cell Complement and FcγRIIB Receptors. *Journal of Experimental Medicine* **196**, 1189–1200 (2002).
83. Zhang, Y. *et al.* Germinal center B cells govern their own fate via antibody feedback. *The Journal of Experimental Medicine* **210**, 457 (2013).
84. Shulman, Z. *et al.* Dynamic signaling by T follicular helper cells during germinal center B cell selection. *Science* **345**, 1058–1062 (2014).
85. Ferrari, S. *et al.* Mutations of CD40 gene cause an autosomal recessive form of immunodeficiency with hyper IgM. *Proceedings of the National Academy of Sciences* **98**, 12614–12619 (2001).
86. Allen, R. C. *et al.* CD40 Ligand Gene Defects Responsible for X-Linked Hyper-IgM Syndrome. *Science (1979)* **259**, 990–993 (1993).

87. Luo, W., Weisel, F. & Shlomchik, M. J. B Cell Receptor and CD40 Signaling Are Rewired for Synergistic Induction of the c-Myc Transcription Factor in Germinal Center B Cells. *Immunity* **48**, 313-326.e5 (2018).
88. de Silva, N. S. & Klein, U. Dynamics of B cells in germinal centres. *Nature Publishing Group* (2015) doi:10.1038/nri3804.
89. Saito, M. *et al.* A signaling pathway mediating downregulation of BCL6 in germinal center B cells is blocked by BCL6 gene alterations in B cell lymphoma. *Cancer Cell* **12**, 280–292 (2007).
90. Laidlaw, B. J. & Cyster, J. G. Transcriptional regulation of memory B cell differentiation. *Nature Reviews Immunology 2020 21:4* **21**, 209–220 (2020).
91. Akkaya, M., Kwak, K. & Pierce, S. K. B cell memory: building two walls of protection against pathogens. *Nature Reviews Immunology 2019 20:4* **20**, 229–238 (2019).
92. Smith, K. G. C., Light, A., Nossal, G. J. V. & Tarlinton, D. M. The extent of affinity maturation differs between the memory and antibody-forming cell compartments in the primary immune response. *EMBO Journal* **16**, 2996–3006 (1997).
93. Ise, W. *et al.* T Follicular Helper Cell-Germinal Center B Cell Interaction Strength Regulates Entry into Plasma Cell or Recycling Germinal Center Cell Fate. *Immunity* **48**, 702-715.e4 (2018).
94. Kometani, K. *et al.* Repression of the transcription factor Bach2 contributes to predisposition of IgG1 memory B cells toward plasma cell differentiation. *Immunity* **39**, 136–147 (2013).
95. Kurosaki, T., Kometani, K. & Ise, W. Memory B cells. *Nature Reviews Immunology 2015 15:3* **15**, 149–159 (2015).
96. Weisel, F. J., Zuccarino-Catania, G. v., Chikina, M. & Shlomchik, M. J. A Temporal Switch in the Germinal Center Determines Differential Output of Memory B and Plasma Cells. *Immunity* **44**, 116–130 (2016).
97. Muramatsu, M. *et al.* Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme. *Cell* **102**, 553–563 (2000).
98. Petersen-Mahrt, S. K., Harris, R. S. & Neuberger, M. S. AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. *Nature* **418**, 99–103 (2002).
99. Durandy, A. Mini-review Activation-induced cytidine deaminase: a dual role in class-switch recombination and somatic hypermutation. (2003) doi:10.1002/eji.200324133.
100. Pećina-Šlaus, N., Kafka, A., Salamon, I. & Bukovac, A. Mismatch Repair Pathway, Genome Stability and Cancer. *Frontiers in Molecular Biosciences* **7**, 122 (2020).
101. Schormann, N., Ricciardi, R. & Chattopadhyay, D. Uracil-DNA glycosylases—Structural and functional perspectives on an essential family of DNA repair enzymes. *Protein Science* **23**, 1667–1685 (2014).
102. Saha, T., Sundaravinayagam, D. & di Virgilio, M. Charting a DNA Repair Roadmap for Immunoglobulin Class Switch Recombination. *Trends in Biochemical Sciences* **46**, 184–199 (2021).
103. Han, L., Masani, S. & Yu, K. Overlapping activation-induced cytidine deaminase hotspot motifs in Ig class-switch recombination. *Proc Natl Acad Sci U S A* **108**, 11584–11589 (2011).
104. Stavnezer, J. & Schrader, C. E. IgH chain class switch recombination: mechanism and regulation. *J Immunol* **193**, 5370–5378 (2014).

105. Chi, X., Li, Y. & Qiu, X. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. *Immunology* **160**, 233–247 (2020).
106. Balázs, M., Martin, F., Zhou, T. & Kearney, J. F. Blood dendritic cells interact with splenic marginal zone B cells to initiate T-independent immune responses. *Immunity* **17**, 341–352 (2002).
107. Stein, K. E. Thymus-independent and thymus-dependent responses to polysaccharide antigens. *J Infect Dis* **165 Suppl 1**, S49–S52 (1992).
108. MacLennan, I. C. M. *et al.* Extrafollicular antibody responses. *Immunological Reviews* **194**, 8–18 (2003).
109. Jenks, S. A., Cashman, K. S., Woodruff, M. C., Lee, F. E. H. & Sanz, I. Extrafollicular responses in humans and SLE. *Immunological Reviews* **288**, 136–148 (2019).
110. Kaji, T. *et al.* Distinct cellular pathways select germline-encoded and somatically mutated antibodies into immunological memory. *J Exp Med* **209**, 2079–2097 (2012).
111. Lee, J. S. *et al.* Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci Immunol* **5**, (2020).
112. Shaffer, A. L. *et al.* Blimp-1 orchestrates plasma cell differentiation by extinguishing the mature B cell gene expression program. *Immunity* **17**, 51–62 (2002).
113. Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody-secreting plasma cells. *Nature Reviews Immunology 2015 15:3* **15**, 160–171 (2015).
114. Tokoyoda, K., Egawa, T., Sugiyama, T., Choi, B. il & Nagasawa, T. Cellular niches controlling B lymphocyte behavior within bone marrow during development. *Immunity* **20**, 707–718 (2004).
115. Lacotte, S., Brun, S., Muller, S. & Dumortier, H. CXCR3, inflammation, and autoimmune diseases. *Ann N Y Acad Sci* **1173**, 310–317 (2009).
116. Halliley, J. L. *et al.* Long-Lived Plasma Cells Are Contained within the CD19-CD38hiCD138+ Subset in Human Bone Marrow. *Immunity* **43**, 132–145 (2015).
117. Cerutti, A., Cols, M. & Puga, I. Marginal zone B cells: virtues of innate-like antibody-producing lymphocytes. *Nature Reviews Immunology 2013 13:2* **13**, 118–132 (2013).
118. Weill, J. C., Weller, S. & Reynaud, C. A. Human Marginal Zone B Cells. *http://dx.doi.org/10.1146/annurev.immunol.021908.132607* **27**, 267–285 (2009).
119. Martin, F. & Kearney, J. F. Marginal-zone B cells. *Nature Reviews Immunology 2002 2:5* **2**, 323–335 (2002).
120. Pone, E. J. *et al.* BCR-signalling synergizes with TLR-signalling for induction of AID and immunoglobulin class-switching through the non-canonical NF-κB pathway. *Nature Communications 2012 3:1* **3**, 1–12 (2012).
121. Weller, S. *et al.* Human blood IgM "memory" B cells are circulating splenic marginal zone B cells harboring a prediversified immunoglobulin repertoire. *Blood* **104**, 3647–3654 (2004).
122. Berkowska, M. A. *et al.* Human memory B cells originate from three distinct germinal center-dependent and -independent maturation pathways. *Blood* **118**, 2150–2158 (2011).
123. Hardy, R. R. B-1 B Cell Development. *The Journal of Immunology* **177**, 2749–2754 (2006).
124. Cunningham, A. F. *et al.* B1b Cells Recognize Protective Antigens after Natural Infection and Vaccination. *Frontiers in Immunology* **5**, (2014).
125. Haas, K. M., Poe, J. C., Steeber, D. A. & Tedder, T. F. B-1a and B-1b Cells Exhibit Distinct Developmental Requirements and Have Unique Functional Roles in Innate and Adaptive Immunity to S. pneumoniae. *Immunity* **23**, 7–18 (2005).

126. Burnett, D. L., Reed, J. H., Christ, D. & Goodnow, C. C. Clonal redemption and clonal anergy as mechanisms to balance B cell tolerance and immunity. *Immunological Reviews* **292**, 61–75 (2019).
127. Goodnow, C. C., Crosbie, J., Jorgensen, H., Brink, R. A. & Basten, A. Induction of self-tolerance in mature peripheral B lymphocytes. *Nature 1989 342:6248* **342**, 385–391 (1989).
128. Duty, J. A. *et al.* Functional anergy in a subpopulation of naive B cells from healthy humans that express autoreactive immunoglobulin receptors. *The Journal of Experimental Medicine* **206**, 139 (2009).
129. Nemazee, D. Antigen receptor 'capacity' and the sensitivity of self-tolerance. *Immunol Today* **17**, 25 (1996).
130. Pugh-Bernard, A. E. *et al.* Regulation of inherently autoreactive VH4-34 B cells in the maintenance of human B cell tolerance. *J Clin Invest* **108**, 1061–1070 (2001).
131. Isenberg, D., Spellerberg, M., Williams, W., Griffiths, M. & Stevenson, F. Identification of the 9G4 idiotope in systemic lupus erythematosus. *Br J Rheumatol* **32**, 876–882 (1993).
132. Reed, J. H., Jackson, J., Christ, D. & Goodnow, C. C. Clonal redemption of autoantibodies by somatic hypermutation away from self-reactivity during human immunization. *Journal of Experimental Medicine* **213**, 1255–1265 (2016).
133. Wrammert, J. *et al.* Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. *J Exp Med* **208**, 181–193 (2011).
134. Pappas, L. *et al.* Rapid development of broadly influenza neutralizing antibodies through redundant mutations. *Nature 2014 516:7531* **516**, 418–422 (2014).
135. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**, 727–733 (2020).
136. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature 2020 579:7798* **579**, 265–269 (2020).
137. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature 2020 579:7798* **579**, 270–273 (2020).
138. Coutard, B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res* **176**, (2020).
139. Lin, X. D. *et al.* Extensive diversity of coronaviruses in bats from China. *Virology* **507**, 1–10 (2017).
140. Zhang, Y. Z. & Holmes, E. C. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell* **181**, 223–227 (2020).
141. Chan, J. F.-W. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet* **395**, 514–523 (2020).
142. Hu, B., Guo, H., Zhou, P. & Shi, Z. L. Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology 2020 19:3* **19**, 141–154 (2020).
143. Hui, D. S. *et al.* The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health — The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases* **91**, 264 (2020).
144. Zhang, Q. *et al.* Molecular mechanism of interaction between SARS-CoV-2 and host cells and interventional therapy. *Signal Transduction and Targeted Therapy* **6**, (2021).
145. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology 2020 19:3* **19**, 155–170 (2020).

146. Minskaia, E. *et al.* Discovery of an RNA virus 3′→5′ exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proceedings of the National Academy of Sciences* **103**, 5108–5113 (2006).
147. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **395**, 565–574 (2020).
148. Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology 2020 5:4* **5**, 536–544 (2020).
149. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine 2020 26:4* **26**, 450–452 (2020).
150. Wang, M. Y. *et al.* SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development. *Frontiers in Cellular and Infection Microbiology* **10**, 724 (2020).
151. Li, W. *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450–454 (2003).
152. Hoffmann, M., Kleine-Weber, H. & Pöhlmann, S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Molecular Cell* **78**, 779-784.e5 (2020).
153. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
154. Wrapp, D. *et al.* Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *bioRxiv* (2020) doi:10.1101/2020.02.11.944462.
155. Cai, Y. *et al.* Distinct conformational states of SARS-CoV-2 spike protein. *Science* **369**, (2020).
156. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature 2020 581:7807* **581**, 221–224 (2020).
157. Sungnak, W. *et al.* SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nature Medicine 2020 26:5* **26**, 681–687 (2020).
158. Sette, A. & Crotty, S. Adaptive immunity to SARS-CoV-2 and COVID-19. *Cell* **184**, 861–880 (2021).
159. Lim, Y., Ng, Y., Tam, J. & Liu, D. Human Coronaviruses: A Review of Virus-Host Interactions. *Diseases* **4**, 26 (2016).
160. Arunachalam, P. S. *et al.* Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science (1979)* **369**, 1210–1220 (2020).
161. Hadjadj, J. *et al.* Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science (1979)* **369**, 718–724 (2020).
162. McNab, F., Mayer-Barber, K., Sher, A., Wack, A. & O'Garra, A. Type I interferons in infectious disease. *Nature Reviews Immunology 2015 15:2* **15**, 87–103 (2015).
163. Li, J., Liu, Y. & Zhang, X. Murine Coronavirus Induces Type I Interferon in Oligodendrocytes through Recognition by RIG-I and MDA5. *Journal of Virology* **84**, 6472–6482 (2010).
164. Loo, Y. M. & Gale, M. Immune signaling by RIG-I-like receptors. *Immunity* **34**, 680–692 (2011).
165. Schindler, C., Levy, D. E. & Decker, T. JAK-STAT signaling: from interferons to cytokines. *J Biol Chem* **282**, 20059–20063 (2007).
166. Riberoid, M. S., Jouvenet, N., Dreuxid, M. & Bastien Nisoleid, S. Interplay between SARS-CoV-2 and the type I interferon response. (2020) doi:10.1371/journal.ppat.1008737.

167. Blanco-Melo, D. *et al.* Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* **181**, 1036 (2020).
168. Liu, C. *et al.* Time-resolved systems immunology reveals a late juncture linked to fatal COVID-19. *Cell* **184**, (2021).
169. Galani, I. E. *et al.* Untuned antiviral immunity in COVID-19 revealed by temporal type I/III interferon patterns and flu comparison. *Nature Immunology* **22**, 32–40 (2021).
170. Zhang, Q. *et al.* Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**, (2020).
171. Schulte-Schrepping, J. *et al.* Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. *Cell* **182**, (2020).
172. Lucas, C. *et al.* Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature 2020 584:7821* **584**, 463–469 (2020).
173. Cron, R. Q., Caricchio, R. & Chatham, W. W. Calming the cytokine storm in COVID-19. *Nature Medicine 2021 27:10* **27**, 1674–1675 (2021).
174. Laing, A. G. *et al.* A dynamic COVID-19 immune signature includes associations with poor prognosis. *Nature Medicine* **26**, 1623–1635 (2020).
175. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* **395**, 497–506 (2020).
176. Xu, Z. *et al.* Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *The Lancet Respiratory Medicine* **8**, 420–422 (2020).
177. Hornby, P. *et al.* Dexamethasone in Hospitalized Patients with Covid-19. *N Engl J Med* **384**, 693–704 (2021).
178. Kyriazopoulou, E. *et al.* Early treatment of COVID-19 with anakinra guided by soluble urokinase plasminogen receptor plasma levels: a double-blind, randomized controlled phase 3 trial. *Nature Medicine 2021 27:10* **27**, 1752–1760 (2021).
179. Caricchio, R. *et al.* Effect of Canakinumab vs Placebo on Survival Without Invasive Mechanical Ventilation in Patients Hospitalized With Severe COVID-19: A Randomized Clinical Trial. *JAMA* **326**, 230–239 (2021).
180. WHO Rapid Evidence Appraisal for COVID-19 Therapies (REACT) Working Group. Association Between Administration of IL-6 Antagonists and Mortality Among Patients Hospitalized for COVID-19: A Meta-analysis. *JAMA* **326**, 499–518 (2021).
181. Szabo, P. A. *et al.* Longitudinal profiling of respiratory and systemic immune responses reveals myeloid cell-driven lung inflammation in severe COVID-19. *Immunity* **54**, 797-814.e6 (2021).
182. Oja, A. E. *et al.* Divergent SARS-CoV-2-specific T- and B-cell responses in severe but not mild COVID-19 patients. *European Journal of Immunology* **50**, 1998–2012 (2020).
183. Grifoni, A. *et al.* Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. *Cell* **181**, 1489-1501.e15 (2020).
184. Röltgen, K. & Boyd, S. D. Antibody and B cell responses to SARS-CoV-2 infection and vaccination. *Cell Host & Microbe* (2021) doi:10.1016/j.chom.2021.06.009.
185. Sokal, A. *et al.* Journal Pre-proof Maturation and persistence of the anti-SARS-CoV-2 memory B cell response. *Cell* (2021) doi:10.1016/j.cell.2021.01.050.
186. Robbiani, D. F. *et al.* Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature* **584**, 437–442 (2020).
187. Kreer, C. *et al.* Longitudinal Isolation of Potent Near-Germline SARS-CoV-2-Neutralizing Antibodies from COVID-19 Patients. *Cell* **182**, 843-854.e12 (2020).
188. Woodruff, M. C. *et al.* Extrafollicular B cell responses correlate with neutralizing antibodies and morbidity in COVID-19. *Nature Immunology* **21**, 1506–1516 (2020).

189. Kaneko, N. *et al.* Loss of Bcl-6-Expressing T Follicular Helper Cells and Germinal Centers in COVID-19. *Cell* **183**, 143-157.e13 (2020).
190. Wang, Z. *et al.* Naturally enhanced neutralizing breadth against SARS-CoV-2 one year after infection. *Nature* 1–10 (2021) doi:10.1038/s41586-021-03696-9.
191. Gaebler, C. *et al.* Evolution of antibody immunity to SARS-CoV-2. *Nature* **591**, 639–644 (2021).
192. Garcia-Beltran, W. F. *et al.* COVID-19-neutralizing antibodies predict disease severity and survival. *Cell* **184**, 476-488.e11 (2021).
193. Shrock, E. *et al.* Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. *Science (1979)* **370**, (2020).
194. Long, Q.-X. *et al.* Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nature Medicine* **26**, 845–848 (2020).
195. Voss, W. N. *et al.* Prevalent, protective, and convergent IgG recognition of SARS-CoV-2 non-RBD spike epitopes. *Science (1979)* eabg5268 (2021) doi:10.1126/science.abg5268.
196. Röltgen, K. *et al.* Defining the features and duration of antibody responses to SARS-CoV-2 infection associated with disease severity and outcome. *Science Immunology* **5**, (2020).
197. Isho, B. *et al.* Persistence of serum and saliva antibody responses to SARS-CoV-2 spike antigens in COVID-19 patients. *Sci Immunol* **5**, (2020).
198. Sariol, A. & Perlman, S. Lessons for COVID-19 Immunity from Other Coronavirus Infections. *Immunity* **53**, 248–263 (2020).
199. Dan, J. M. *et al.* Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science (1979)* **371**, eabf4063 (2021).
200. Ng, K. W. *et al.* Preexisting and de novo humoral immunity to SARS-CoV-2 in humans. *Science* **370**, 1339–1343 (2020).
201. Anderson, E. M. *et al.* Seasonal human coronavirus antibodies are boosted upon SARS-CoV-2 infection but not associated with protection. *Cell* (2021) doi:10.1016/j.cell.2021.02.010.
202. Rivett, L. *et al.* Screening of healthcare workers for SARS-CoV-2 highlights the role of asymptomatic carriage in COVID-19 transmission. *Elife* **9**, (2020).
203. Akbari, P. *et al.* Genetic Analyses of Blood Cell Structure for Biological and Pharmacological Inference. *bioRxiv* **5**, 2020.01.30.927483 (2020).
204. Xiong, X. *et al.* A thermostable, closed SARS-CoV-2 spike protein trimer. *Nature Structural & Molecular Biology 2020 27:10* **27**, 934–941 (2020).
205. Daly, J. L. *et al.* Neuropilin-1 is a host factor for SARS-CoV-2 infection. *Science (1979)* **370**, 861–865 (2020).
206. Patterson, E. I. *et al.* Methods of Inactivation of SARS-CoV-2 for Downstream Biological Assays. *Journal of Infectious Diseases* **222**, 1462–1467 (2020).
207. Pereyra Gerber, P. *et al.* Protease-activatable biosensors of SARS-CoV-2 infection for cell. *bioRxiv* 2021.03.22.435957 (2021) doi:10.1101/2021.03.22.435957.
208. Mlcochova, P. *et al.* Combined Point-of-Care Nucleic Acid and Antibody Testing for SARS-CoV-2 following Emergence of D614G Spike Variant. *Cell Reports Medicine* **1**, 100099 (2020).
209. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology 2019 37:8* **37**, 907–915 (2019).
210. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–15550 (2005).

211. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
212. Horvath, S. & Langfelder, P. *Tutorials for the WGCNA package for R: WGCNA Background and glossary*. https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/Simulated-00-Background.pdf (2011).
213. Langfelder, P., Luo, R., Oldham, M. C. & Horvath, S. Is My Network Module Preserved and Reproducible? *PLoS Computational Biology* **7**, e1001057 (2011).
214. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* **14**, (2018).
215. Chaudhary, N. & Wesemann, D. R. Analyzing Immunoglobulin Repertoires. *Frontiers in Immunology* **0**, 462 (2018).
216. Bashford-Rogers, R. J. M., Smith, K. G. C. & Thomas, D. C. Antibody repertoire analysis in polygenic autoimmune diseases. *Immunology* vol. 155 3–17 (2018).
217. López-Santibáñez-Jácome, L., Avendaño-Vázquez, S. E. & Flores-Jasso, C. F. The Pipeline Repertoire for Ig-Seq Analysis. *Frontiers in Immunology* **10**, 899 (2019).
218. Consortium, Co.-19 M. B. At. (COMBAT) *et al.* Title: A blood atlas of COVID-19 defines hallmarks of disease severity and specificity COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium 1* 1. *medRxiv* 2021.05.11.21256877 (2021) doi:10.1101/2021.05.11.21256877.
219. Brochet, X., Lefranc, M. P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* **36**, (2008).
220. Gadala-Maria, D., Yaari, G., Uduman, M. & Kleinstein, S. H. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc Natl Acad Sci U S A* **112**, E862–E870 (2015).
221. Sanders, H. L. Marine Benthic Diversity: A Comparative Study. *https://doi.org/10.1086/282541* **102**, 243–282 (2015).
222. Willis, A. D. Rarefaction, alpha diversity, and statistics. *Frontiers in Microbiology* **10**, 2407 (2019).
223. Fisher, R. A., Corbet, A. S. & Williams, C. B. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *The Journal of Animal Ecology* **12**, 42 (1943).
224. Shannon, C. E. A Mathematical Theory of Communication. *The Bell System Technical Journal* **27**, 623–656.
225. Simpson, E. H. Measurement of Diversity. *Nature 1949 163:4148* **163**, 688–688 (1949).
226. Hill, M. O. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **54**, 427–432 (1973).
227. Gupta, N. T. *et al.* Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31**, 3356–3358 (2015).
228. Shugay, M. *et al.* VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLoS Computational Biology* **11**, 1004503 (2015).
229. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
230. Raybould, M. I. J., Kovaltsuk, A., Marks, C. & Deane, C. M. CoV-AbDab: the coronavirus antibody database. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa739.

231. Kuri-Cervantes, L. *et al.* Comprehensive mapping of immune perturbations associated with severe COVID-19. *Science Immunology* **5**, (2020).
232. Mann, E. R. *et al.* Longitudinal immune profiling reveals key myeloid signatures associated with COVID-19. *Science Immunology* **5**, (2020).
233. Mathew, D. *et al.* Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science (1979)* **369**, (2020).
234. Karki, R. *et al.* Synergism of TNF-α and IFN-γ Triggers Inflammatory Cell Death, Tissue Damage, and Mortality in SARS-CoV-2 Infection and Cytokine Shock Syndromes. *Cell* **184**, 149-168.e17 (2021).
235. Blanco-Melo, D. *et al.* Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* **181**, 1036-1045.e9 (2020).
236. Nielsen, S. C. A. *et al.* Human B Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2. *Cell Host and Microbe* **28**, 516-525.e5 (2020).
237. Yuen, C.-K. *et al.* SARS-CoV-2 nsp13, nsp14, nsp15 and orf6 function as potent interferon antagonists. *https://doi.org/10.1080/22221751.2020.1780953* **9**, 1418–1428 (2020).
238. Xia, H. *et al.* Evasion of Type I Interferon by SARS-CoV-2. *Cell Reports* **33**, 108234 (2020).
239. Fajnzylber, J. *et al.* SARS-CoV-2 viral load is associated with increased disease severity and mortality. *Nature Communications 2020 11:1* **11**, 1–9 (2020).
240. Rydyznski Moderbacher, C. *et al.* Antigen-Specific Adaptive Immunity to SARS-CoV-2 in Acute COVID-19 and Associations with Age and Disease Severity. *Cell* **183**, 996-1012.e19 (2020).
241. Chen, Y. *et al.* Quick COVID-19 Healers Sustain Anti-SARS-CoV-2 Antibody Production. *Cell* **183**, 1496-1507.e16 (2020).
242. Sekine, T. *et al.* Robust T Cell Immunity in Convalescent Individuals with Asymptomatic or Mild COVID-19. *Cell* **183**, 158-168.e14 (2020).
243. Shomuradova, A. S. *et al.* SARS-CoV-2 Epitopes Are Recognized by a Public and Diverse Repertoire of Human T Cell Receptors. *Immunity* **53**, 1245-1257.e5 (2020).
244. Sette, A. & Crotty, S. Pre-existing immunity to SARS-CoV-2: the knowns and unknowns. *Nature Reviews Immunology 2020 20:8* **20**, 457–458 (2020).
245. Ju, B. *et al.* Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* **584**, 115–119 (2020).
246. Banchereau, R., Cepika, A. M., Banchereau, J. & Pascual, V. Understanding Human Autoimmunity and Autoinflammation Through Transcriptomics. *Annu Rev Immunol* **35**, 337–370 (2017).
247. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417 (2015).
248. Thomas, T., Spitalnik, S. L. & Alessandro, A. D. '. COVID-19 infection alters kynurenine and fatty acid metabolism, correlating with IL-6 levels and renal status Graphical abstract Clinical Medicine COVID-19 Metabolism Find the latest version. (2020) doi:10.1172/jci.insight.140327.
249. Bantug, G. R., Galluzzi, L., Kroemer, G. & Hess, C. The spectrum of T cell metabolism in health and disease. *Nat Rev Immunol* **18**, 19–34 (2018).
250. Smeitink, J., van den Heuvel, L. & DiMauro, S. The genetics and pathology of oxidative phosphorylation. *Nature Reviews Genetics 2001 2:5* **2**, 342–352 (2001).
251. Millar, J. E., Fanning, J. P., McDonald, C. I., McAuley, D. F. & Fraser, J. F. The inflammatory response to extracorporeal membrane oxygenation (ECMO): A review of the pathophysiology. *Critical Care* **20**, 1–10 (2016).

252. Prestes, E. B. *et al.* Mitochondrial Reactive Oxygen Species Participate in Signaling Triggered by Heme in Macrophages and upon Hemolysis. *The Journal of Immunology* **205**, 2795–2805 (2020).
253. Bernardes, J. P., Mishra, N., Tran, F. & Rosenstiel, P. Longitudinal Multi-omics Analyses Identify Responses of Megakaryocytes, Erythroid Cells, and Plasmablasts as Hallmarks of Severe COVID-19. *Immunity* **53**, 1296-1314.e9 (2020).
254. Opoka-Winiarska, V., Grywalska, E. & Roliński, J. Could hemophagocytic lymphohistiocytosis be the core issue of severe COVID-19 cases? *BMC Medicine* **18**, (2020).
255. Dewaele, K. & Claeys, R. Hemophagocytic lymphohistiocytosis in SARS-CoV-2 infection. *Blood* **135**, 2323–2323 (2020).
256. George, M. R. Hemophagocytic lymphohistiocytosis: review of etiologies and management. *Journal of Blood Medicine* **5**, 69 (2014).
257. Quast, I. & Tarlinton, D. B cell memory: understanding COVID-19. *Immunity* vol. 54 205–210 (2021).
258. Buckland, M. S. *et al.* Treatment of COVID-19 with remdesivir in the absence of humoral immunity: a case report. *Nature Communications* **11**, 1–11 (2020).
259. Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592**, 277–282 (2021).
260. Libster, R. *et al.* Early High-Titer Plasma Therapy to Prevent Severe Covid-19 in Older Adults. *New England Journal of Medicine* **384**, 610–618 (2021).
261. Joyner, M. J. *et al.* Convalescent Plasma Antibody Levels and the Risk of Death from Covid-19. *New England Journal of Medicine* **384**, 1015–1027 (2021).
262. Kim, Y. Il *et al.* Critical role of neutralizing antibody for SARS-CoV-2 reinfection and transmission. *Emerging Microbes and Infections* **10**, 152–160 (2021).
263. Stephens, D. S. & McElrath, M. J. COVID-19 and the Path to Immunity. *JAMA - Journal of the American Medical Association* vol. 324 1279–1281 (2020).
264. Cox, R. J. & Brokstad, K. A. Not just antibodies: B cells and T cells mediate immunity to COVID-19. *Nature Reviews Immunology* vol. 20 581–582 (2020).
265. Hernandez, A. M. & Holodick, N. E. Editorial: Natural Antibodies in Health and Disease. *Frontiers in Immunology* **8**, 1795 (2017).
266. Song, G. *et al.* Cross-reactive serum and memory B-cell responses to spike protein in SARS-CoV-2 and endemic coronavirus infection. *Nature Communications* **12**, 2938 (2021).
267. Yang, F. *et al.* Shared B cell memory to coronaviruses and other pathogens varies in human age groups and tissues. *Science (1979)* **372**, 738–741 (2021).
268. Bergamaschi, L. *et al.* Longitudinal analysis reveals that delayed bystander CD8+ T cell activation and early immune pathology distinguish severe COVID-19 from mild disease. *Immunity* **54**, 1257-1275.e8 (2021).
269. Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine 2021 27:5* **27**, 904–916 (2021).
270. Wen, W. *et al.* Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. *Cell Discovery* **6**, (2020).
271. Schatz, D. G. & Swanson, P. C. V(D)J Recombination: Mechanisms of Initiation. (2011) doi:10.1146/annurev-genet-110410-132552.
272. Theofilopoulos, A. N., Kono, D. H. & Baccala, R. The multiple pathways to autoimmunity. *Nature Immunology* vol. 18 716–724 (2017).
273. Wardemann, H. *et al.* Predominant autoantibody production by early human B cell precursors. *Science (1979)* **301**, 1374–1377 (2003).

274. Stavnezer, J., Guikema, J. E. J. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annual Review of Immunology* vol. 26 261–292 (2008).
275. Xu, Z., Zan, H., Pone, E. J., Mai, T. & Casali, P. Immunoglobulin class-switch DNA recombination: Induction, targeting and beyond. *Nature Reviews Immunology* vol. 12 517–531 (2012).
276. Bashford-Rogers, R. J. M. *et al.* Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* **574**, 122–126 (2019).
277. Galson, J. D. *et al.* Deep Sequencing of B Cell Receptor Repertoires From COVID-19 Patients Reveals Strong Convergent Immune Signatures. *Frontiers in Immunology* **11**, 3283 (2020).
278. Seydoux, E. *et al.* Analysis of a SARS-CoV-2-Infected Individual Reveals Development of Potent Neutralizing Antibodies with Limited Somatic Mutation. *Immunity* **53**, 98-105.e5 (2020).
279. Davis, C. W. *et al.* Longitudinal Analysis of the Human B Cell Response to Ebola Virus Infection. *Cell* **177**, 1566-1582.e17 (2019).
280. Godoy-Lozano, E. E. *et al.* Lower IgG somatic hypermutation rates during acute dengue virus infection is compatible with a germinal center-independent B cell response. *Genome Medicine* **8**, (2016).
281. Kalinke, U. *et al.* The role of somatic mutation in the generation of the protective humoral immune response against vesicular stomatitis virus. *Immunity* **5**, 639–652 (1996).
282. Amanat, F. *et al.* SARS-CoV-2 mRNA vaccination induces functionally diverse antibodies to NTD, RBD and S2. *Cell* (2021) doi:10.1016/j.cell.2021.06.005.
283. Wang, Z. *et al.* mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature* **592**, 616–622 (2021).
284. Greaney, A. J. *et al. Antibodies elicited by mRNA-1273 vaccination bind more broadly to the receptor binding domain than do those from SARS-CoV-2 infection*. *Sci. Transl. Med* vol. 13 https://jbloomlab.github.io/SARS-CoV-2-RBD_MAP_Moderna/. (2021).
285. Stamatatos, L. *et al.* mRNA vaccination boosts cross-variant neutralizing antibodies elicited by SARS-CoV-2 infection. *Science (1979)* eabg9175 (2021) doi:10.1126/science.abg9175.
286. Horns, F., Dekker, C. L. & Quake, S. R. Memory B Cell Activation, Broad Anti-influenza Antibodies, and Bystander Activation Revealed by Single-Cell Transcriptomics. *Cell Reports* **30**, 905-913.e6 (2020).
287. Collier, D. A. *et al.* Age-related immune response heterogeneity to SARS-CoV-2 vaccine BNT162b2. *Nature 2021* 1–9 (2021) doi:10.1038/s41586-021-03739-1.
288. Defrance, T., Taillardet, M. & Genestier, L. T cell-independent B cell memory. *Current Opinion in Immunology* **23**, 330–336 (2011).
289. Turner, J. S. *et al.* Human germinal centres engage memory and naive B cells after influenza vaccination. *Nature* **586**, (2020).
290. Steffen, U. *et al.* IgA subclasses have different effector functions associated with distinct glycosylation profiles. *Nature Communications* **11**, 1–12 (2020).
291. He, J. S. *et al.* IgG1 memory B cells keep the memory of IgE responses. *Nature Communications* **8**, (2017).
292. Smith-Norowitz, T. A. *et al.* Long term persistence of IgE anti-influenza virus antibodies in pediatric and adult serum post vaccination with influenza virus vaccine. *International Journal of Medical Sciences* **8**, 239–244 (2011).
293. Vidarsson, G., Dekkers, G. & Rispens, T. IgG Subclasses and Allotypes: From Structure to Effector Functions. *Frontiers in Immunology* **0**, 520 (2014).

294. Victora, G. D. & Nussenzweig, M. C. Germinal centers. *Annual Review of Immunology* vol. 30 429–457 (2012).
295. Heesters, B. A. *et al.* Follicular Dendritic Cells Retain Infectious HIV in Cycling Endosomes. *PLoS Pathogens* **11**, (2015).
296. Scharer, C. D. *et al.* Epigenetic programming underpins B cell dysfunction in human SLE. *Nature Immunology* **20**, 1071–1082 (2019).
297. Tipton, C. M. *et al.* Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nature Immunology* **16**, 755–765 (2015).
298. Jenks, S. A. *et al.* Distinct Effector B Cells Induced by Unregulated Toll-like Receptor 7 Contribute to Pathogenic Responses in Systemic Lupus Erythematosus. *Immunity* **49**, 725-739.e6 (2018).
299. Chen, F., Tzarum, N., Wilson, I. A. & Law, M. VH1-69 antiviral broadly neutralizing antibodies: genetics, structures, and relevance to rational vaccine design. *Curr Opin Virol* **34**, 149 (2019).
300. Kim, S. Il *et al.* Stereotypic neutralizing VHantibodies against SARS-CoV-2 spike protein receptor binding domain in patients with COVID-19 and healthy individuals. *Science Translational Medicine* **13**, (2021).
301. Smith, K. G. C., Hewitson, T. D., Nossal, G. J. V. & Tarlinton, D. M. The phenotype and fate of the antibody-forming cells of the splenic foci. *European Journal of Immunology* **26**, 444–448 (1996).
302. Turner, J. S. *et al.* SARS-CoV-2 infection induces long-lived bone marrow plasma cells in humans. *Nature 2021 595:7867* **595**, 421–425 (2021).
303. Wang, Z. *et al.* Enhanced SARS-CoV-2 neutralization by dimeric IgA. *Science Translational Medicine* **13**, 1555 (2021).
304. Stacey, H. D. *et al.* IgA potentiates NETosis in response to viral infection. *Proceedings of the National Academy of Sciences* **118**, e2101497118 (2021).
305. Turner, J. S. *et al.* Human germinal centres engage memory and naive B cells after influenza vaccination. *Nature* **586**, 127–132 (2020).
306. Russell, M. W., Moldoveanu, Z., Ogra, P. L. & Mestecky, J. Mucosal Immunity in COVID-19: A Neglected but Critical Aspect of SARS-CoV-2 Infection. *Frontiers in Immunology* **11**, 3221 (2020).
307. Moradi-kalbolandi, S., Majidzadeh-A, K., Abdolvahab, M. H., Jalili, N. & Farahmand, L. The Role of Mucosal Immunity and Recombinant Probiotics in SARS-CoV2 Vaccine Development. *Probiotics and Antimicrobial Proteins* vol. 1 3 (2021).
308. Jeyanathan, M. *et al.* Immunological considerations for COVID-19 vaccine strategies. *Nature Reviews Immunology* vol. 20 615–632 (2020).
309. Bricker, T. L. *et al.* Journal Pre-proof A single intranasal or intramuscular immunization with chimpanzee adenovirus vectored SARS-CoV-2 vaccine protects against pneumonia in hamsters. *Cell Reports* (2021) doi:10.1016/j.celrep.2021.109400.
310. Ahluwalia, B., Magnusson, M. K. & Öhman, L. Mucosal immune system of the gastrointestinal tract: maintaining balance between the good and the bad. *http://dx.doi.org/10.1080/00365521.2017.1349173* **52**, 1185–1193 (2017).
311. Peterson, L. W. & Artis, D. Intestinal epithelial cells: regulators of barrier function and immune homeostasis. *Nature Reviews Immunology 2014 14:3* **14**, 141–153 (2014).
312. Scaldaferri, F., Pizzoferrato, M., Gerardi, V., Lopetuso, L. & Gasbarrini, A. The gut barrier: new acquisitions and therapeutic approaches. *J Clin Gastroenterol* **46 Suppl**, (2012).
313. Cepek, K. L. *et al.* Adhesion between epithelial cells and T lymphocytes mediated by E-cadherin and the alpha E beta 7 integrin. *Nature* **372**, 190–193 (1994).

314. Mowat, A. M. & Agace, W. W. Regional specialization within the intestinal immune system. *Nature Reviews Immunology 2014 14:10* **14**, 667–685 (2014).
315. Verbist, K. C. & Klonowski, K. D. Functions of IL-15 in Anti-Viral Immunity: Multiplicity and Variety. *Cytokine* **59**, 467 (2012).
316. Ouyang, W., Kolls, J. K. & Zheng, Y. The biological functions of T helper 17 cell effector cytokines in inflammation. *Immunity* **28**, 454–467 (2008).
317. Coombes, J. L. & Maloy, K. J. Control of intestinal homeostasis by regulatory T cells and dendritic cells. *Semin Immunol* **19**, 116–126 (2007).
318. Coombes, J. L. *et al.* A functionally specialized population of mucosal CD103+ DCs induces Foxp3+ regulatory T cells via a TGF-beta and retinoic acid-dependent mechanism. *J Exp Med* **204**, 1757–1764 (2007).
319. Berlin, C. *et al.* Alpha 4 beta 7 integrin mediates lymphocyte binding to the mucosal vascular addressin MAdCAM-1. *Cell* **74**, 185–195 (1993).
320. Bemark, M., Boysen, P. & Lycke, N. Y. Induction of gut IgA production through T cell-dependent and T cell-independent pathways. *Ann N Y Acad Sci* **1247**, 97–116 (2012).
321. Torres, J., Mehandru, S., Colombel, J. F. & Peyrin-Biroulet, L. Crohn's disease. *Lancet* **389**, 1741–1755 (2017).
322. Johansen, F. E. & Kaetzel, C. S. Regulation of the polymeric immunoglobulin receptor and IgA transport: new advances in environmental factors that stimulate pIgR expression and its role in mucosal immunity. *Mucosal Immunol* **4**, 598–602 (2011).
323. Kamada, N., Seo, S. U., Chen, G. Y. & Núñez, G. Role of the gut microbiota in immunity and inflammatory disease. *Nature Reviews Immunology 2013 13:5* **13**, 321–335 (2013).
324. Chistiakov, D. A., Bobryshev, Y. v., Kozarov, E., Sobenin, I. A. & Orekhov, A. N. Intestinal mucosal tolerance and impact of gut microbiota to mucosal tolerance. *Frontiers in Microbiology* **5**, (2014).
325. Ma, S., Wang, C., Mao, X. & Hao, Y. B Cell Dysfunction Associated With Aging and Autoimmune Diseases. *Frontiers in Immunology* **10**, 318 (2019).
326. Bauer, H., Horowoitz, R. E., Levenson, S. M. & Popper, H. The Response of the Lymphatic Tissue to the Microbial Flora. Studies on Germfree Mice. *The American Journal of Pathology* **42**, 471 (1963).
327. Hapfelmeier, S. *et al.* Reversible microbial colonization of germ-free mice reveals the dynamics of IgA immune responses. *Science* **328**, 1705–1709 (2010).
328. Chen, H. *et al.* BCR selection and affinity maturation in Peyer's patch germinal centres. *Nature 2020 582:7812* **582**, 421–425 (2020).
329. Travassos, L. H. *et al.* Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. *Nat Immunol* **11**, 55–62 (2010).
330. Chang, J. T. Pathophysiology of Inflammatory Bowel Diseases. *https://doi.org/10.1056/NEJMra2002697* **383**, 2652–2664 (2020).
331. Roda, G. *et al.* Crohn's disease. *Nature Reviews Disease Primers 2020 6:1* **6**, 1–19 (2020).
332. Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Research 2020 30:6* **30**, 492–506 (2020).
333. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
334. Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature 2001 411:6837* **411**, 603–606 (2001).
335. Baumgart, D. C. & Sandborn, W. J. Crohn's disease. *The Lancet* **380**, 1590–1605 (2012).

336. Johansson, M. E. V. & Hansson, G. C. Immunological aspects of intestinal mucus and mucins. *Nature Reviews Immunology 2016 16:10* **16**, 639–649 (2016).
337. Buisine, M. P. *et al.* Abnormalities in mucin gene expression in Crohn's disease. *Inflamm Bowel Dis* **5**, 24–32 (1999).
338. Zeissig, S. *et al.* Changes in expression and distribution of claudin 2, 5 and 8 lead to discontinuous tight junctions and barrier dysfunction in active Crohn's disease. *Gut* **56**, 61–72 (2007).
339. Petnicki-Ocwieja, T. *et al.* Nod2 is required for the regulation of commensal microbiota in the intestine. *Proceedings of the National Academy of Sciences* **106**, 15813–15818 (2009).
340. Tschurtschenthaler, M. *et al.* Defective ATG16L1-mediated removal of IRE1α drives Crohn's disease-like ileitis. *J Exp Med* **214**, 401–422 (2017).
341. Pascal, V. *et al.* A microbial signature for Crohn's disease. *Gut* **66**, 813–822 (2017).
342. Kaser, A. Genetic Risk of Severe Covid-19. *https://doi.org/10.1056/NEJMe2025501* **383**, 1590–1591 (2020).
343. Britton, G. J. *et al.* Microbiotas from Humans with Inflammatory Bowel Disease Alter the Balance of Gut Th17 and RORγt+ Regulatory T Cells and Exacerbate Colitis in Mice. *Immunity* **50**, 212-224.e4 (2019).
344. Barnich, N. *et al.* CEACAM6 acts as a receptor for adherent-invasive E. coli, supporting ileal mucosa colonization in Crohn disease. *J Clin Invest* **117**, 1566–1574 (2007).
345. Imam, T., Park, S., Kaplan, M. H. & Olson, M. R. Effector T helper cell subsets in inflammatory bowel diseases. *Frontiers in Immunology* **9**, 1212 (2018).
346. Hsieh, C. S. *et al.* Development of TH1 CD4+ T cells through IL-12 produced by Listeria-induced macrophages. *Science* **260**, 547–549 (1993).
347. McGeachy, M. J. & Cua, D. J. Th17 cell differentiation: the long and winding road. *Immunity* **28**, 445–453 (2008).
348. Baumgart, D. C. & le Berre, C. Newer Biologic and Small-Molecule Therapies for Inflammatory Bowel Disease. *New England Journal of Medicine* **385**, 1302–1315 (2021).
349. Moor, K. *et al.* High-avidity IgA protects the intestine by enchaining growing bacteria. *Nature* **544**, 498–502 (2017).
350. Martin, J. C. *et al.* Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* **178**, 1493-1508.e20 (2019).
351. Boland, B. S. *et al.* Heterogeneity and clonal relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. *Sci Immunol* **5**, (2020).
352. Castro-Dopico, T. *et al.* Anti-commensal IgG Drives Intestinal Inflammation and Type 17 Immunity in Ulcerative Colitis. *Immunity* **50**, 1099-1114.e10 (2019).
353. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature 2012 491:7422* **491**, 119–124 (2012).
354. Verhelst, X. *et al.* Protein Glycosylation as a Diagnostic and Prognostic Marker of Chronic Inflammatory Gastrointestinal and Liver Diseases. *Gastroenterology* **158**, 95–110 (2020).
355. Uzzan, M. *et al.* Ulcerative colitis is characterized by a plasmablast-skewed humoral response associated with disease activity. *Nature Medicine 2022* 1–14 (2022) doi:10.1038/s41591-022-01680-y.
356. Kotagiri, P. *et al.* B cell receptor repertoire kinetics after SARS-CoV-2 infection and vaccination. *Cell Rep* **38**, 110393 (2022).

357. Alfaleh, M. A. *et al.* Phage Display Derived Monoclonal Antibodies: From Bench to Bedside. *Frontiers in Immunology* **11**, 1986 (2020).
358. Ledsgaard, L., Kilstrup, M., Karatt-Vellatt, A., McCafferty, J. & Laustsen, A. H. Basics of Antibody Phage Display Technology. *Toxins (Basel)* **10**, (2018).
359. Castro-Dopico, T. & Clatworthy, M. R. Mucosal IgG in inflammatory bowel disease – a question of (sub)class? *Gut Microbes* **12**, (2020).
360. Cao, Y. *et al.* Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients' B Cells. *Cell* **182**, 73-84.e16 (2020).
361. Katz, L. M. (A Little) Clarity on Convalescent Plasma for Covid-19. *New England Journal of Medicine* **384**, 666–668 (2021).

Appendix A The impact of hypoxia on B cells in COVID-19

Introduction

The B cell response is a vital component of immune defence against SARS-CoV-2; neutralising antibodies contribute to protection from infection(Cao et al., 2020), monoclonal antibodies may be of benefit(Katz, 2021) and antibody deficiency predisposes to viral persistence(Buckland et al., 2020). Early and often persistent changes in B cell numbers are prominent in symptomatic COVID-19: increased plasmablasts and reduced memory B cells correlate with disease severity, and germinal centre (GC) responses, somatic hypermutation, and T follicular helper (T_{FH}) cells may be reduced(Bergamaschi et al., 2021; Kaneko et al., 2020; Nielsen et al., 2020; Stephenson et al., 2021b).

The reduction in B cell subsets in COVID-19 was reminiscent of the phenotype of mice with VHL-deficient B cells, which exhibit constitutive activation of Hypoxia-Inducible Factors (HIFs)(Burrows et al., 2020; Cho et al., 2016). These models had a Cre-mediated deletion of *Vhl*, targeted specifically to the B cell lineage, with Cre expression being driven by the B cell-specific promoter *Mb-1*, which deletes at the earliest Pro-B cell stage. A model of permanent Cre expression was used to assess developmental effects(Burrows et al., 2020) and a tamoxifen-inducible model used (*Mb1*-CreER^{T2})(Cho et al., 2016) to assess the effect on B cell immune responses.

This raised the possibility that hypoxia might contribute to B cell dysregulation in COVID-19. Hypoxemia is prominent in COVID-19, often occurring “silently”, with many patients presenting to hospital with profoundly low blood oxygen saturations(Couzin-Frankel, 2020).

The B cell abnormalities characteristic of COVID-19 might limit the efficiency of the anti-SARS-CoV-2 response(Kemp et al., 2021), predispose to secondary infection(Ripa et al., 2021), or contribute to sequelae such as autoimmunity(Wang et al., 2020a). Understanding if and how hypoxia impacts upon them could therefore inform management strategies.

Methods

Participant recruitment and clinical data collection

This cohort has been previously described by Bergamaschi et al(Bergamaschi et al., 2021). Briefly, study participants were recruited between 31/3/2020 and 20/7/2020 from patients attending Addenbrooke's Hospital, Royal Papworth Hospital NHS Foundation Trust or Cambridge and Peterborough Foundation Trust with a confirmed diagnosis of COVID-19, together with Health Care Workers identified through staff screening as PCR positive for SARS-CoV-2(Rivett et al., 2020). Controls were recruited among hospital staff attending Addenbrooke's for SARS-CoV-2 serology screening programme and having a negative serology result. All participants provided informed consent.

Inpatients were sampled at study entry, and then at regular intervals as long as they remained admitted to hospital (approximately weekly up to 4 weeks, and then every 2 weeks up to 12 weeks). Discharged patients were invited to provide a follow-up sample 4-8 weeks after study enrolment. Health care workers were sampled at study entry, and subsequently after approximately 2 and 4 weeks.

Clinical data were retrospectively collected by review of medical charts and extraction of data (laboratory test results, vital signs, medications) from Epic electronic health records (Addenbrooke's Hospital) and from MetaVision ICU (Royal Papworth Hospital).

Study volunteers were classified in 5 groups:

- Group A: health care workers who were asymptomatic at the time of positive SARS-CoV-2 testing. This group included 10 volunteers who had possible COVID-19 symptoms before PCR testing (median time from symptoms to COVID-19 PCR test 26 days, range 9-42 days).
- Group B: health care workers who had possible COVID-19 symptoms at the time of PCR testing.
- Group C: patients in hospital who did not receive any supplemental oxygen for COVID-19. Five patients were discharged soon after initial diagnosis and assessment but followed up as part of the study.
- Group D: patients in hospital who received supplemental oxygen using low flow nasal prongs, simple face mask, Venturi mask or non re-breather face mask
- Group E: patients in hospital who received any of non-invasive ventilation (NIV), mechanical ventilation or ECMO. Patients who received supplemental oxygen (but no ventilation) and deceased in hospital were also assigned to group E.

Study results were analysed according to time since onset of COVID-19 symptoms, or otherwise time since positive SARS-CoV-2 testing (in group A and in 4 asymptomatic patients in group C).

Peripheral blood mononuclear cell preparation and flow cytometry immunophenotyping

For direct enumeration of T, B and NK cells, an aliquot of whole blood EDTA (50μl) was added to BD TruCount™ tubes with 20μl BD Multitest™ 6-colour TBNK reagent (BD Biosciences) and processed as per the manufacturer's instructions.

Peripheral venous blood (up to 27 ml per sample) for isolation of Peripheral Blood Mononuclear Cells (PBMCs) was collected into 10% sodium citrate tubes. PBMCs were

isolated using Leucosep tubes (Greiner Bio-One) with Histopaque 1077 (Sigma) by centrifugation at 800x g for 15 minutes at room temperature. PBMCs at the interface were collected, rinsed twice with autoMACS running buffer (Miltenyi Biotech) and cryopreserved in FBS with 10% DMSO. All samples were processed within 4 hours of collection.

Approximately 10^6 cells have been stained with: anti-human IgM (clone: G20-127, BD), CD19 (clone: SJ25C1, BD), CD38 (clone: HIT2, BD), IgD (clone: IA6-2, BD), CD20 (clone: 2H7, BD), CD3 (clone: UCHT1, BioLegend), CD14 (clone: 63D3, BioLegend), CD15 (clone: W6D3, BioLegend), CD193 (clone: 5E8, BioLegend), CD27 (clone: O323, BioLegend), CD56 (clone: MEM188, Thermo), CD24 (clone: ML5, BD), IgA (polyclonal goat IgG, Jackson), IgG (clone: G18-145, BD), and Zombie Yellow (BioLegend) as described in detail by Bergamaschi et al(Bergamaschi et al., 2021), Samples were stored at 4°C and acquired within 4 hours using a 5-laser BD Symphony X-50 flow cytometer. Single colour compensation tubes (BD CompBeads) or cells were prepared for each of the fluorophores used and acquired at the start of each flow cytometer run.

Samples were gated in FlowJo v10.2 and number of cells falling within each gate was recorded. For analysis, these were expressed as an absolute concentration of cells per μl using the BD TruCount™ system. Some previously reported data, detailed by Bergamaschi et al is represented for comparison.

CRP, complement components and cytokines

As detailed in Bergamaschi et al(Bergamaschi et al., 2021), concentrations of complement components were measured in EDTA plasma using commercially available enzyme-linked immunosorbent assays (ELISA) kits. High sensitivity CRP and cytokines (IL-6, IL-10, IL-1β, TNFα and IFNγ) were assayed in serum using standard laboratory assays.

Total Immunoglobulin levels

Serum immunoglobulin levels were measured for 186 COVID-19 patients and 45 healthy controls at the time of enrolment using the standard assay by the Immunology Department at Peterborough City hospital.

Whole blood bulk RNA-Seq

Whole blood RNA was extracted from PAXgene Blood RNA tubes (BD Biosciences) of 188 COVID-19 patients at up to 2 time points and 42 healthy volunteers. RNA-Sequencing libraries were generated using the SMARTer® Stranded Total RNA-Seq v2 - Pico Input Mammalian kit (Takara) using 10ng RNA as input following the manufacturer's protocol. Libraries were pooled together (n = 96) and sequenced using 75bp paired-end chemistry across 4 lanes of a Hiseq4000 instrument (Illumina) to achieve 10 million reads per sample. Read quality was assessed using FastQC v.0.11.8 (Babraham Bioinformatics, UK), and SMARTer adaptors trimmed and residual rRNA reads depleted in silico using Trim galore v.0.6.4 (Babraham Bioinformatics, UK) and BBSplit (BBMap v.38.67(BBMap - Bushnell B. - sourceforge.net/projects/bbmap/)), respectively. Alignment was performed using HISAT2 v.2.1.0 (Kim et al., 2019) against the GRCh38 genome achieving a greater than 95% alignment rate. Count matrices were generated using featureCounts (Rsubreads package)(Liao et al., 2019) and stored as a DGEList object (EdgeR package)(Robinson et al., 2009) for further analysis.

All downstream data handling was performed in R (R Core Team, 2015). Counts were filtered using filterByExpr (EdgeR) with a gene count threshold of 10 CPM and the minimum number of samples set at the size of the smallest disease group. Library counts were

normalised using calcNormFactors (EdgeR) using the method 'weighted trimmed mean of M-values'. The function 'voom'(Law et al., 2014) was applied to the data to estimate the mean-variance relationship, allowing adjustment for heteroscedasticity.
The analyses were carried out splitting the samples in 12 days bins post screening (group A) or symptom onset (groups B-E).

Single cell RNA-seq

CITE-seq data were generated from frozen PBMCs of 36 COVID-19 patients and 11 healthy controls as described by Stephenson et al.(Stephenson et al., 2021b) Briefly, after thawing, pools of 4 samples were generated by combined 500,000 viable cells per individual (total of 2 million cells per pool). TotalSeq-C™ antibody cocktail (BioLegend 99813) was used to perform cell surface marker staining on 500,000 cells per pool. 50,000 live cells (up to a maximum of 60,000 total cells) for each pool were processed using Single Cell V(D)J 5' version 1.1 (1000020) together with Single Cell 5' Feature Barcode library kit (1000080), Single Cell V(D)J Enrichment Kit, Human B Cells (1000016) and Single Cell V(D)J Enrichment Kit, Human T Cells (1000005) (10xGenomics) according to the manufacturer's protocols. Samples were sequenced on NovaSeq 6000 (Illumina) using S1 flowcells. Droplet libraries were processed using Cellranger v4.0. Reads were aligned to the GRCh38 human genome concatenated to the SARS-Cov-2 genome (NCBI SARS-CoV-2 isolate Wuhan-Hu-1) using STAR(Dobin et al., 2013) and unique molecular identifiers (UMIs) deduplicated. CITE-seq UMIs were counted for GEX and ADT libraries simultaneously to generate feature X droplet UMI count matrices.

Statistics

All statistical analyses were conducted using custom scripts in R (R Core Team, 2015). Appropriately age matched healthy controls were included in all analyses. Absolute cell counts (cells/uL) were offset by +1 to allow subsequent log2 transformation of zero counts. Unless otherwise specified, longitudinally collected data was grouped by bins of 12 days from symptom onset or first positive SARS-CoV2 swab. Pairwise statistical comparisons of absolute cell counts and proportions and immunoglobulin levels between individuals in a given severity group at a given time bin and HCs, or between severity groups, was conducted by Wilcoxon test unless otherwise specified. For analyses involving repeated measures, false discovery rate corrected (Benjamini & Hochberg) p values were reported. For individuals sampled more than once within a given time bin, data from the earliest blood collection was used.

Gene set enrichment analysis (GSEA)(Subramanian et al., 2005) was used to identify biological pathways enriched in COVID-19 severity groups relative to healthy controls. Briefly, a list of ranked genes, determined by Signal-To-Noise ratio was generated. An enrichment score was calculated, determined by how often genes from the geneset of interest appeared at the top or the bottom of the pre-ranked set of genes with the enrichment score representing the maximum deviation from zero. To assess statistical significance, an empirical phenotype- based permutation test was run, where a collection of enrichment scores was generated from the random assignment of phenotype to samples and used to generate a null distribution. To account for multiple testing, an FDR rate q <

0.20 was deemed significant. HALLMARK gene sets from the Molecular Signatures Database (http://www.broadinstitute.org/gsea/msigdb) were used in analysis.

The relationships between immunological parameters and transcriptional data in the form of gene expression modules were assessed using Pearson's correlation (Hmisc package) and visualized with corrplot.

B Cell Receptor Repertoire

Library Preparation

B cell receptor repertoire libraries have been generated for 119 COVID-19 patients and 71 healthy controls using the protocol describe by Bashford-Rogers et al.[278]Briefly, 200ng of total RNA from PAXgenes (14ul volume) was combined with 1uL 10mM dNTP and 10uM reverse primer mix (2uL) and incubated for 5 min at 70°C. The mixture was immediately placed on ice for 1 minute and then subsequently combined with 1uL DTT (0.1 M), 1uL SuperScriptIV (Thermo Fisher Scientific), 4ul SSIV Buffer (Thermo Fisher Scientific) and 1uL RNAse inhibitor. The solution was incubated at 50 °C for 60 min followed by 15 min inactivation at 70 °C. cDNA was cleaned with AMPure XP beads and PCR-amplified with a 5' V-gene multiplex primer mix and 3' universal reverse primer using the KAPA protocol and the following thermal cycling conditions: 1cycle (95°C, 5min); 5cycles (98°C, 20s; 72°C, 30s); 5cycles (98°C, 15s; 65°C, 30s; 72°C, 30s); 19cycles (98 °C, 15s; 60°C, 30s; 72°C, 30s); 1 step (72°C, 5 min). Sequencing libraries were prepared using Illumina protocols and sequenced using 300-bp paired-end sequencing on a MiSeq.

Sequence analysis

Raw reads were filtered for base quality using a median Phred score of ≥32 (http://sourceforge.net/projects/quasr/). Forward and reverse reads were merged where a minimum 20bp identical overlapping region was present. Sequences were retained where over 80% base sequence similarity was present between all sequences with the same barcode. The constant-region allele with highest sequence similarity was identified by 10-mer matching to the reference constant-region genes from the IMGT database. Sequences without complete reading frames and non-immunoglobulin sequences were removed and only reads with significant similarity to reference IGHV and J genes from the IMGT database using BLAST were retained. Immunoglobulin gene use and sequence annotation were performed in IMGT V-QUEST, and repertoire differences were performed by custom scripts in Python.

Murine Models

Vhl$^{-/-}$ mice(Haase et al., 2001) were crossed with *Cd79a-cre* (*Mb1-cre*)(Hobeika et al., 2006) or *Cd19-cre* (JAX, stock no. 004126) to delete *Vhl* in the B cell lineage. All mice with *lox*P-flanked alleles were hemizygous for *Cre*. Deletion efficiency was determined via real-time PCR of genomic DNA. The degree of excision was calculated by comparison of *Vhl* intact DNA relative to an unexcised gene *Actb*. The primers and probes used were *Vhl* forward 5'-GCTTGCGAATCCGAGGG, *Vhl* reverse 5'-TCCTCTGGACTGGCTGCC, *Vhl* Probe 5'-E6-FAM−CCCGTTCCAATAATGCCCCGG (Life Technologies) and *Actb* (mouse assay ID: Mm00607939_s1; Life Technologies). The deletion efficiency for mature B cells was 52% (95% CI 30-75%) in *Vhl*$^{-/-}$*Cd19-cre* mice and 98% (95% CI, 97–99%) in *Vhl*$^{-/-}$*Mb1-cre* mice(Burrows et al., 2020). The mice were backcrossed for at least eight generations and maintained on a C57BL/6J background. These mice, along with C57BL/6J mice (JAX, stock no.

000664) were housed in specific pathogen-free animal facilities (at 20–23 °C, with 40–60% humidity, 12-h light:12-h dark cycle). All experiments included age- and litter-matched mice that were not selected for gender. Where possible, the resource equation was used to determine sample size for experiments. Randomization was genetic and, where possible, investigators were blinded to the genetic status. For hypoxic exposure studies, a randomization algorithm was used (Excel) to allocate mice into experimental groups. Mice were immunised with 100μg 4-hydroxy-3-nitrophenylaceyl-keyhole limpet hemocyanine (NP-KLH, loading 31-33) (Biosearch Technologies) adjuvanted with Alum (Thermo Scientific) via intraperitoneal injection. C57BL/6J mice were exposed to 10% O_2 in a hypoxic chamber for 1 day, then immunised. Mice remained in the hypoxic chamber for 10-, 14- or 20-days post immunisation. Normoxic (21% O_2) mice were treated the same way and were kept in standard conditions. The reoxygenation groups were removed from the hypoxic chamber on day 10 post immunisation to standard conditions for 4 or 10 days.

Tissue processing and immunophenotyping of murine cells by flow cytometry was performed as described(Burrows et al., 2020). B cells were gated as total B cells ($B220^+$), FO ($B220^+CD93^-CD23^+CD21^+$), MZ ($B220^+CD93^-CD23^-CD21^+$), GC ($B220^+CD95^{high}GL\text{-}7^{high}$), PCs ($B220^-CD138^+$), early memory ($B220^+IgD^{neg/lo}CD95^+GL7^-CD38^+CD73^+$), T cells ($CD3^+$) and Tfh cells ($CD3^+CD4^+PD\text{-}1^{high}CXCR5^{high}FoxP3^-$). Antibodies are listed in Supplementary table 1.

Murine total Immunoglobulin (Ig) and NP-specific ELISAs

Detection of total IgM and NP-specific IgG1 was performed as described(Brownlie et al., 2008).

Confocal microscopy

10µm sections were mounted on Superfrost Plus slides and air dried at RT for 1h. Samples were then fixed in -20°C acetone for 10 minutes and air dried again at RT for 1h before blocking in 0.1M Tris containing 1% BSA, 1% normal mouse serum and 1% normal rat serum. Samples were stained in a wet chamber at RT for 1h30 with the appropriate antibodies, washed 3 times in PBS and mounted in Fluoromount-G. Images were acquired using a TCS SP8 inverted confocal microscope on a 40x oil immersion objective. Raw imaging data were processed using Imaris.

Murine BCR amplification and sequencing

BCR amplification and sequencing was performed as described in Burrows et al.(Burrows et al., 2020) Data are available at the Sequence Research Archive (SRA) database (BioProject accession nos. PRJNA574931, PRJNA574906 and PRJNA574628). Briefly, total RNA was extracted from isolated plasma cells ($B220^{-}CD138^{+}$). Reverse transcription (RT) was performed using constant region-specific primers (including unique molecular identifiers (UMIs)), followed by cDNA cleanup and PCR amplification using V gene specific primers. Sequencing libraries were prepared using Illumina protocols and sequenced using 300bp paired-ended MiSeq (Illumina). Raw reads were filtered as Burrows et al.(Burrows et al., 2020) Ig gene sequence annotations were performed in IMGT V-QUEST, where somatic hypermutation repertoire and isotype usage differences were performed by custom scripts in python, and statistics were performed in *R* using Wilcoxon tests for significance (non-parametric test of differences between distributions).

Role of funding source

Financial support from CVC Capital Partners, the Evelyn Trust (20/75), Addenbrooke's Charitable Trust (12/20A), the UKRI/NIHR through the UK Coronavirus Immunology Consortium (UK-CIC) and NIHR Cambridge BioResource centre (Grant codes: RG85445 and RG94028) funded sample collection and processing. The Wellcome Trust (no. 19710) for supporting murine studies. Funders had no role in study design, data collection, data analyses, interpretation, or writing of report.

Ethics

Ethical approval was obtained from the East of England – Cambridge Central Research Ethics Committee ("NIHR BioResource" REC ref 17/EE/0025, and "Genetic variation AND Altered Leucocyte Function in health and disease - GANDALF" REC ref 08/H0308/176).
All procedures were ethically approved by the University of Cambridge Animal Welfare and Ethical Review Body and complied with the Animals (Scientific Procedures) Act 1986 Amendment Regulations 2012, under the authority of a UK Home Office Licence. The ARRIVE (Animal Research: Reporting In Vivo Experiments) guidelines (https://arriveguidelines.org/arrive-guidelines) were used for planning, conducting and reporting experiments.

Results

SARS-CoV-2 PCR-positive subjects were recruited between March and July 2020 and categorized by peak clinical severity(Bergamaschi et al., 2021) (**Fig 1a** and Supp Fig 1a):

A) asymptomatic healthcare workers (HCWs) recruited from routine screening (n=18).

B) HCWs either still working with mild symptoms, or symptomatic and self-isolating (n=40).

C) patients who presented to hospital but never required oxygen supplementation (n=46).

D) admitted patients whose maximal respiratory support was supplemental oxygen (n=37).

E) patients who required assisted ventilation (57 of 60) or died without ventilation (3 of 60).

We compared absolute B cell subset numbers in COVID-19 patients to 45 healthy controls (**Fig 1b**)(Bergamaschi et al., 2021). In more severe groups C-E, profound reductions in T_{FH}-like cells and many B cell subsets were seen at the first bleed, including memory and marginal zone like (MZL) B cells. Most then showed some recovery. Changes were far less pronounced in groups A and B (**Fig 1b**). Single-cell RNA-sequencing coupled with analysis of surface proteins on a subset of patients confirmed proportional differences (**Fig S1b**). We also explored cell kinetics in groups C–E, assigning patients to two categories based on whether their CRP concentrations remained elevated above 10 mg/L ("persisting CRP") or fell below 10 mg/L ("resolving CRP") by their final bleed within 3 months post symptom onset. The latter group included both individuals with early high CRP that then fell, together with those for which CRP remained low (10 mg/L) throughout (see Fig 6, Bergamaschi et al.6). B cell derangements including low transitional B cells and elevated plasmablasts persisted regardless of CRP, while MZ, memory and naïve B cell reductions recovered as the CRP did (**Fig S1**).

Total serum IgM fell as disease severity increased, with many patients in groups C-E having IgM levels below the normal range, while IgG and IgA were less impacted (**Figs 1c and S1d**). Anti- SARS-CoV-2 spike antibodies rose over time in all severity groups, reaching highest titres in the more severe groups(Bergamaschi et al., 2021). BCR sequencing showed reduced somatic mutation in COVID-19 patients, most prominent in IgA and IgG1/2 (**Figs 1d and S1e**), consistent with previous reports(Nielsen et al., 2020).

Having noted a similarity in B cell phenotype in COVID-19 and mice with constitutively active HIF(**Fig 1e)** (Burrows et al., 2020; Cho et al., 2016), we hypothesised that B cell loss might relate to hypoxia *in vivo*. The effects of acute hypoxia on immune responses in mice and humans have not been assessed. Immunised *Vhl*$^{-/-}$*Cd19-cre* mice (in which VHL is deleted at the pre-B cell stage) were studied to allow more granular comparison with the B cell pathology in COVID-19: they showed reductions in follicular (FO), MZ and GC B cells and increased plasma cell (PC) to B cell ratio (**Fig S2a**). Findings were confirmed in *Vhl*$^{-/-}$*Mb1-cre* mice (VHL deleted in pro-B cells) immunised with NP-KLH, in which reduced NP-specific GC and memory B cells, and T_{FH} cells were also observed (**Fig S2b-c**). Serum IgM, but not IgG and IgA, was reduced, as was affinity maturation and somatic hypermutation (SHM) in some isotypes (**Fig S2d-f**). Reduced GC and memory B cells, along with defects in affinity maturation, were similar to those observed in an inducible model of B cell specific *Vhl* deletion(Cho et al., 2016). Thus, HIF activation in multiple mouse models produces similar changes to those in patients with moderate to severe COVID-19 (**Fig 1e**), supporting the possibility that hypoxia could be implicated in COVID-19 B cell pathology.

Hypoxia in COVID-19 patients, as determined by monitoring peripheral oxygen saturation (SpO_2), was common early in disease and tended to improve with recovery, or with

ventilation or ECMO in intensive care (group E) (**Fig S1f**), but was hard to correlate directly with immune changes, as recorded SpO_2 is usually taken on oxygen replacement, which is commonly administered in the ambulance or immediately on arrival in hospital. Thus these data will underestimate the real hypoxia on admission, which is likely to have been sustained for hours or days before the patient presented. Furthermore, SpO_2 is not reliably reflective of tissue hypoxia, which may persist in severe pneumonia and acute lung injury. We therefore instead measured the impact of hypoxia on the transcriptome in COVID-19 blood samples. An eigengene representative of the curated Hallmark hypoxia signature in whole blood was associated with disease severity (**Fig 2a**). Hallmark hypoxia gene set enrichment in single-cell RNA-sequencing data(Stephenson et al., 2021b) showed enrichment in almost all B cell subsets in groups C, D and E but in only plasma cells in A and B (**Fig 2b**).

The “Hallmark hypoxia” signature was enriched for genes regulated by HIF, and could therefore be activated by reduced oxygen-tension and/or inflammatory stimuli. We therefore compared Hallmark signatures of hypoxia with inflammation in whole blood - the hypoxia signature was prominent in early severe disease (groups C-E) before declining, perhaps due to recovering disease and effective oxygen supplementation, but was not enriched in mild disease (A and B). In contrast inflammation-related signatures were often seen in these mild groups (**Fig 2c**). The differential enrichment of hypoxic and inflammatory signatures in asymptomatic and mild disease suggests a specific role for hypoxia, but an additional role for inflammation cannot be excluded. Finally the hypoxia eigengene correlated inversely with B cell number across most subsets, with the exception of plasma cells (see discussion) (**Fig 2d**).

To differentiate inflammatory-driven from hypoxia-driven HIF-mediated effects, we studied mice in hypoxic conditions (10% O_2) after immunisation with NP-KLH. After 11 days of hypoxia we observed reduced transitional, FO, MZ and GC B cells, whilst PCs were normal (**Fig S3a)**. These defects persisted when hypoxia was prolonged (20 days; **Fig 3a and S3b)**. Little effect was observed on B cell development in the bone marrow (**Fig S3c**) and as FO and MZ B cells turn over every 7-8 weeks(B et al., 2005; Sprent and Basten, 1973), it is unlikely these early reductions are due to a developmental defect. In hypoxic mice there was a tendency to reduced early memory B cells **(Fig 3a and S3d)** and serum Ig was normal but antigen-specific IgG1 was reduced **(Fig 3b and S3e)**. Histological analysis revealed that, in hypoxic conditions, B cells were almost absent from the MZ, which appeared otherwise structurally intact **(Fig 4a and S3f)**. Some mice were removed from the hypoxic chamber after 11 days: B cell subsets generally recovered following this reoxygenation. MZ and GC B cells were most prominently affected by hypoxia, continuing to decline under hypoxic conditions, and recovering more slowly following reoxygenation, than other subsets **(Fig 4b).** Hypoxia induced only minor reductions in T cells and NK cells, and no changes in macrophage numbers, indicating that B cells seem particularly sensitivity to perturbations in oxygenation and HIF **(Fig S4 and data not shown)**. Thus while it is likely that hypoxia will have other effects that will warrant more detailed examination, B cells seem particularly sensitivity to perturbations in oxygenation and HIF activity.

Discussion

We demonstrate profound B cell abnormalities in severe COVID-19 and provide evidence they may be the result of hypoxia. B cell lymphopenia extends across all subsets, is present soon after symptom onset, and is often persistent. There is an associated reduction in total

serum IgM and in somatic hypermutation in switched B cells. Despite this, patients in all groups develop neutralising anti-SARS-CoV-2 antibodies(Bergamaschi et al., 2021), known not to require affinity maturation(Clark et al., 2021). It nonetheless seems likely that these profound B cell deficits could have an impact on disease. As we have suggested for bystander CD8 T cell responses(Bergamaschi et al., 2021), early B cell defects may increase COVID-19 severity through non-antigen-specific mechanisms, perhaps reducing early antigen localisation, transport or presentation, or cytokine or "natural" antibody production(Hernandez and Holodick, 2017). Later, B cell defects may predispose to problematic secondary bacterial infection(Ripa et al., 2021). MZ B cells are key players in early defence against blood-borne bacterial infection (Nemazee, 2021), and MZL cells are profoundly reduced in COVID-19, and hypoxia almost ablates MZ B cells in mice. This may also predispose to re-infection by new variants, as affinity maturation might generate a broader spectrum of neutralizing antibodies and B cell memory(Clark et al., 2021). B cell dysregulation might also play a part in driving COVID-19-associated autoimmunity(Wang et al., 2020a).

Three HIF-α isoforms are known, with HIF-1α and HIF-2α providing the main transcriptional response to oxygen gradients. HIF-1α is ubiquitously expressed but HIF-2α is restricted to specific cell types, including B cells(Burrows et al., 2020). Both HIF-1α and HIF-2α are regulated by prolyl hydroxylation and VHL-mediated protein degradation, although the kinetics of their stabilisation in oxygen gradients differ, with a more prolonged HIF-2 response compared to HIF-1. HIF-1 and HIF-2 also share a number of target genes and are sometimes expressed in the same cells(Ratcliffe, 2007). How these distinct HIF isoforms regulate gene expression is an important area of ongoing study, but the hypoxic signature

we observe is entirely consistent with activation of HIF target genes. These findings are also supported by studies in mice with B cell-specific VHL deletion, and thus constitutively active HIF, which show abnormal B cell development and reduced GC B cells, antibody class-switching and affinity-maturation. Deleting HIF-1 or both HIF-1 and HIF-2 in these models rescued the defects, confirming a HIF-1/-2-dependent effect(Burrows et al., 2020; Cho et al., 2016) and suggesting that hypoxia-induced HIF stabilization might be physiologically important in B cell biology.

Hypoxia might drive transcriptional and cellular changes independently of HIFs via epigenetic modifications involving DNA/histone demethylation. These occur through the oxygen-dependence of several lysine demethylases (KDMs)(Barbarash et al., 1986; Chakraborty et al., 2019) or through impaired DNA methylcytosine hydroxylation via the oxygen-sensitive TET enzymes(Thienpont et al., 2016). The oxygen affinity of KDMs and TETs are higher than the HIF prolyl hydroxylases. Therefore HIF transcriptional responses will occur first, and epigenetic changes may only be observed in severe tissue hypoxia. The role of KDMs and TETs in hypoxic immune regulation have yet to be studied, but they could contribute to the longer-term defects observed on B cells once hypoxia has resolved.

While inflammation can contribute to activate HIF-1α, our demonstration that hypoxia alone can induce profound, reversible B cell abnormalities supports a major role for hypoxia in driving the B cell abnormalities in COVID-19. Hypoxia exerts its major effect on HIF-1α by preventing its degradation, while the main inflammatory impact is through enhancing HIF transcription, making a synergistic impact of hypoxia and inflammation on HIF function plausible(Burrows and Maxwell, 2017; Watts and Walmsley, 2019). While it is impossible to conclusively separate the two impacts in humans with COVID-19, the fact that a tighter correlation is seen between disease severity and hypoxia than with inflammatory

signatures, together with the demonstration that hypoxia alone can induce profound, reversible B cell abnormalities in mice, and supports a role for hypoxia in driving B cell abnormalities in COVID-19. Only plasmablasts showed HIF activation in the absence of hypoxia and these were the only B cell subset in which cell numbers do not correlate with the hypoxia signature, nor fall in mice subject to hypoxia, raising the possibility that HIF is constitutively active in these cells. This is consistent with a growing literature demonstrating that HIF-1α is active in multiple myeloma(Martin et al., 2011). Given this aspect of COVID-19 B cell pathology does not appear impacted by hypoxia, therapeutic approaches to plasma cell control may need to involve pharmacological antagonism of HIF rather than increased oxygenation.

Supplemental oxygen in established COVID-19 may not correct localised areas of hypoxia following acute lung injury/ARDS and could account for a persistent hypoxic transcriptional signal. Inflammation will exacerbate these transcriptional changes, and it is not possible to distinguish the relative contribution of hypoxia versus inflammation in severe disease. However, our corroboratory observations in mice clearly demonstrate that hypoxia is sufficient to drive these B cell changes, irrespective of inflammation, and the observation that hypoxia perturbs B cell immunity has implications in a wide range of clinical settings. In COVID-19, appropriate early oxygen therapy may lead to improved immune responsiveness, impacting on both the short-term and long-term outcomes of the disease, and this could be tested in clinical studies.

References

1 Cao Y, Su B, Guo X, *et al.* Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients' B Cells. *Cell* 2020; **182**: 73-84.e16.

2 Katz LM. (A Little) Clarity on Convalescent Plasma for Covid-19. *New England Journal of Medicine* 2021; **384**: 666–8.

3 Buckland MS, Galloway JB, Fhogartaigh CN, *et al.* Treatment of COVID-19 with remdesivir in the absence of humoral immunity: a case report. *Nature Communications* 2020; **11**: 1–11.

4 Kaneko N, Kuo HH, Boucau J, *et al.* Loss of Bcl-6-Expressing T Follicular Helper Cells and Germinal Centers in COVID-19. *Cell* 2020; **183**: 143-157.e13.

5 Nielsen SCA, Yang F, Jackson KJL, *et al.* Human B Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2. *Cell Host and Microbe* 2020; **28**: 516-525.e5.

6 Bergamaschi L, Mescia F, Turner L, Bradley JR, Lyons PA, Smith KGC. Longitudinal analysis reveals that delayed bystander CD8+ T cell activation and early immune pathology distinguish severe COVID-19 from mild disease. *Immunity* 2021; **54**: 1257-1275.e8.

7 Stephenson E, Reynolds G, Botting RA, *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine 2021 27:5* 2021; **27**: 904–16.

8 Cho SH, Raybuck AL, Stengel K, *et al.* Germinal centre hypoxia and regulation of antibody qualities by a hypoxia response system. *Nature* 2016; **537**: 234–8.

9 Burrows N, Bashford-Rogers RJM, Bhute VJ, *et al.* Dynamic regulation of hypoxia-inducible factor-1α activity is essential for normal B cell development. *Nature Immunology* 2020; **21**: 1408–20.

10 Couzin-Frankel J. The mystery of the pandemic's 'happy hypoxia.' Science. 2020; **368**: 455–6.

11 Kemp SA, Collier DA, Datir RP, *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* 2021; : 1–10.

12 Ripa M, Galli L, Poli A, *et al.* Secondary infections in patients hospitalized with COVID-19: incidence and predictive factors. *Clinical Microbiology and Infection* 2021; **27**: 451–7.

13 Wang EY, Mao T, Klein J, *et al.* Diverse functional autoantibodies in patients with COVID-19. medRxiv. 2020; : 2020.12.10.20247205.

14 Rivett L, Sridhar S, Sparkes D, *et al.* Screening of healthcare workers for SARS-CoV-2 highlights the role of asymptomatic carriage in COVID-19 transmission. *eLife* 2020; **9**. DOI:10.7554/eLife.58728.

15 Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research* 2019; **47**. DOI:10.1093/nar/gkz114.

16 Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2009; **26**: 139–40.

17 Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* 2014; **15**: R29.

18 Dobin A, Davis CA, Schlesinger F, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; **29**: 15–21.

19 Subramanian A, Tamayo P, Mootha VK, *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 2005; **102**: 15545–50.

20 Bashford-Rogers RJM, Bergamaschi L, McKinney EF, *et al.* Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* 2019; **574**: 122–6.

21 Haase VH, Glickman JN, Socolovsky M, Jaenisch R. Vascular tumors in livers with targeted inactivation of the von Hippel-Lindau tumor suppressor. *Proceedings of the National Academy of Sciences of the United States of America* 2001; **98**: 1583–8.
22 Hobeika E, Thiemann S, Storch B, *et al.* Testing gene function early in the B cell lineage in mb1-cre mice. *Proceedings of the National Academy of Sciences of the United States of America* 2006; **103**: 13789–94.
23 Brownlie RJ, Lawlor KE, Niederer HA, *et al.* Distinct cell-specific control of autoimmunity and infection by FcγRIIb. *Journal of Experimental Medicine* 2008; **205**: 883–95.
24 Kotagiri P, Mescia F, Hanson A, *et al.* The impact of hypoxia on B cells in COVID-19. *medRxiv* 2021; : 2021.07.12.21260360.
25 Sprent J, Basten A. Circulating T and B lymphocytes of the mouse: II. Lifespan. *Cellular Immunology* 1973; **7**: 40–59.
26 B S, WJ Q, K H, J E, D A. Characterization of marginal zone B cell precursors. *The Journal of experimental medicine* 2005; **202**: 1225–34.
27 Clark SA, Clark LE, Pan J, *et al.* SARS-CoV-2 evolution in an immunocompromised host reveals shared neutralization escape mechanisms. *Cell* 2021; published online March 16. DOI:10.1016/j.cell.2021.03.027.
28 Hernandez AM, Holodick NE. Editorial: Natural Antibodies in Health and Disease. *Frontiers in Immunology* 2017; **8**: 1795.
29 Nemazee D. Natural history of MZ B cells. Journal of Experimental Medicine. 2021; **218**. DOI:10.1084/JEM.20202700.
30 Ratcliffe PJ. HIF-1 and HIF-2: working alone or together in hypoxia? *The Journal of clinical investigation* 2007; **117**: 862–5.
31 Chakraborty AA, Laukka T, Myllykoski M, *et al.* Histone demethylase KDM6A directly senses oxygen to control chromatin and cell fate. *Science (New York, NY)* 2019; **363**: 1217–22.
32 Barbarash RA, Toll L, Sahn SA. Alpha-difluoromethylornithine infusion and cardiac arrest. *Annals of internal medicine* 1986; **105**: 141–2.
33 Thienpont B, Steinbacher J, Zhao H, *et al.* Tumour hypoxia causes DNA hypermethylation by reducing TET activity. *Nature* 2016; **537**: 63–8.
34 Burrows N, Maxwell PH. Hypoxia and B cells. *Experimental Cell Research* 2017; **356**: 197–203.
35 Watts ER, Walmsley SR. Inflammation and Hypoxia: HIF and PHD Isoform Selectivity. Trends in Molecular Medicine. 2019; **25**: 33–46.
36 Martin SK, Diamond P, Gronthos S, Peet DJ, Zannettino A. The emerging role of hypoxia, HIF-1 and HIF-2 in multiple myeloma. *Leukemia* 2011; **25**: 1533–42.

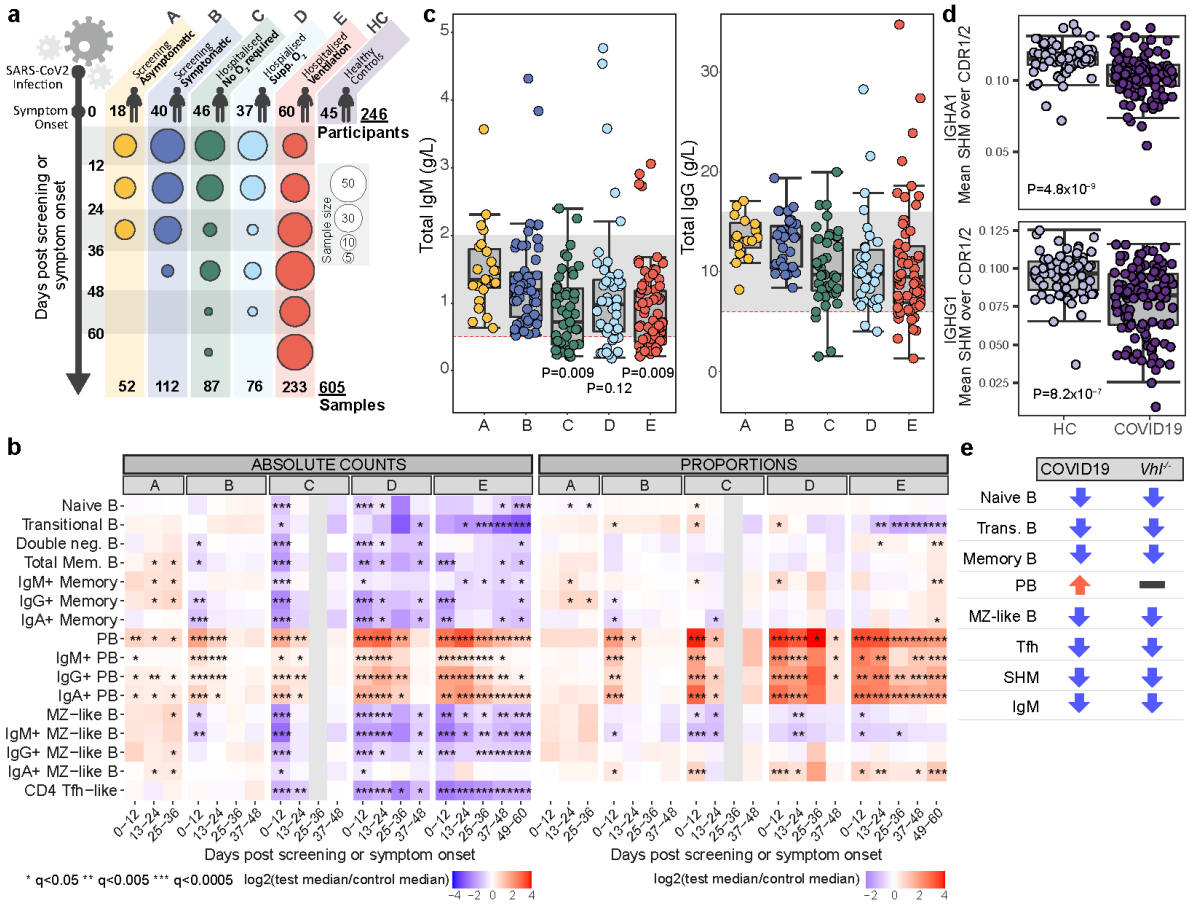


Fig 1. B cells in COVID-19 and VHL-deficient mice.

a, Cohort details. Time post positive swab (group A) or symptom onset (groups B-E). **b,** Median absolute cell counts (left) or proportions relative to total B cells (right) (log2 fold change relative to healthy controls). (Wilcoxon test FDR adjusted p-value (q)): *<0.05, **<0.005, ***<0.0005. **c,** Serum IgG and IgM (g/L) at enrolment. Grey band: 5-95th centiles of healthy reference range (see methods). Significant P values listed. **d,** Somatic hypermutation frequency in IgA and IgG within 0-12 days post symptom onset, calculated over the CDR1/CDR2 regions using BCR sequencing of whole blood. (Wilcoxon test). **c,d,** Circles represent individual donors. **e,** Phenotype comparison: COVID-19 patients versus mice with *Vhl*-deficient B cells.

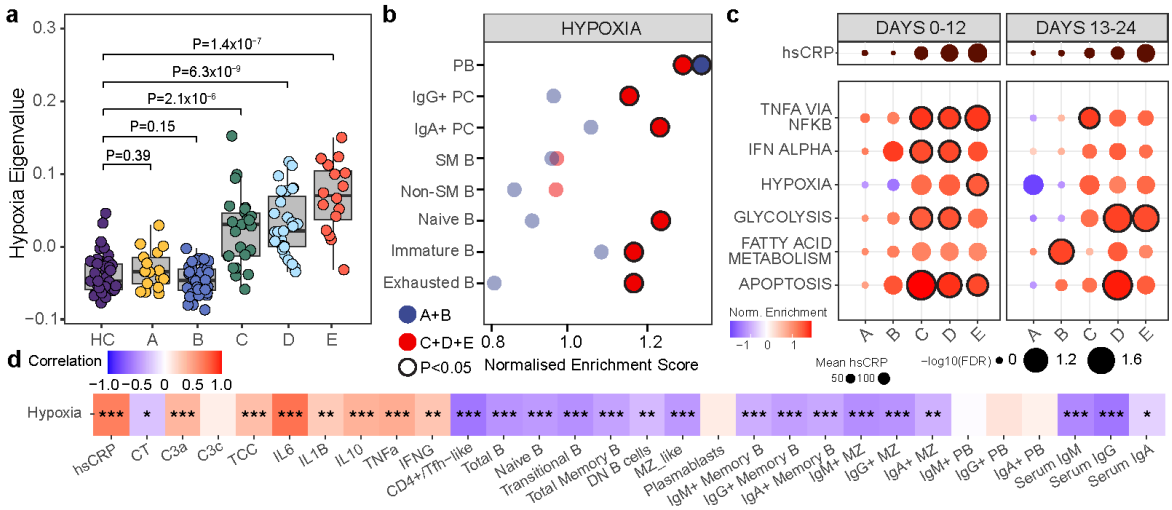


Fig 2. Hypoxia-related transcription signatures in COVID-19

a, Eigenvalues of Hallmark Hypoxia geneset grouped by severity at 0-12 days post symptom onset, (unpaired, two-sided Student's t-test). Circles represent individual donors. **b,** B cell subpopulations identified using CITEseq with Gene set enrichment analysis (GSEA) of Hallmark hypoxia geneset assessed on a single cell level comparing HC to COVID-19, grouped by severity (A/B n=8, C/D/E n =20), within 24 days of symptom onset (B-E)/positive swab(A). **c,** GSEA of Hallmark genesets in COVID-19 versus HC grouped by severity and time. Outlined circles: nominal P value <0.05 and FDR adjusted P <0.2. Mean hsCRP represented. **d,** Correlation between Hallmark hypoxia geneset eigengenes and parameters shown at 0-12 days post symptom onset in COVID-19 patients. Boxes coloured by strength of correlation, Pearson correlation.

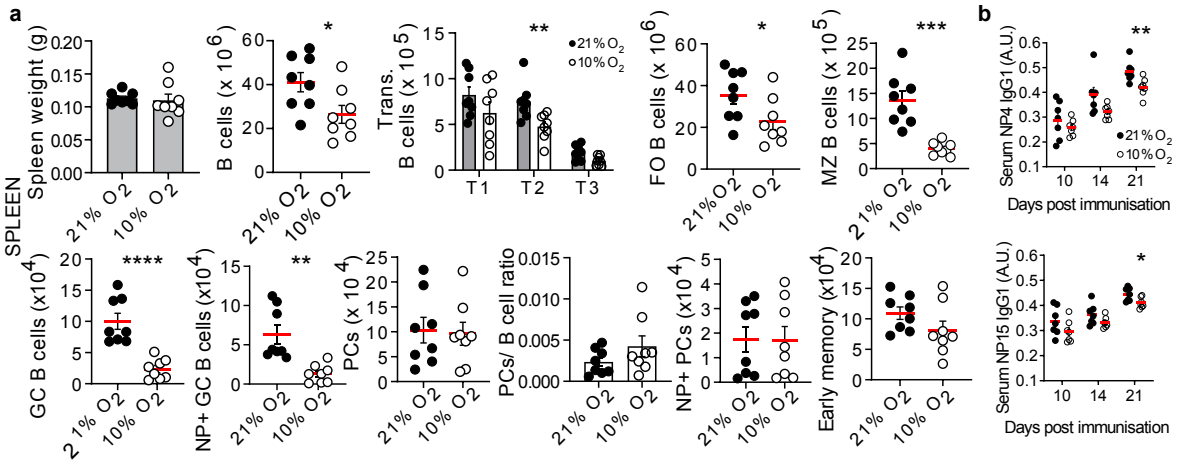


Fig 3. The response of mouse B cells to hypoxia *in vivo*.

a, WT mice exposed to 21% or 10% O_2, were immunized with NP-KLH at day 1, then absolute spleen B cells enumerated at day 21. (Unpaired, two-sided Student's t-test). FO, (Unpaired, two-sided Mann-Whitney U-test). Gating in methods, mean ± s.e.m, circles represent individual mice (n=8 per group), data pooled from two experiments, results confirmed in a third. **b,** Serum NP-specific antibodies after NP-KLH immunization, (unpaired, two-sided Student's t-test). mean ± s.e.m, circles represent individual mice, (n=8 per group), data pooled from two experiments, confirmed in a third. *$P < 0.05$,**$P < 0.01$,***$P < 0.001$,**** $P < 0.0001$

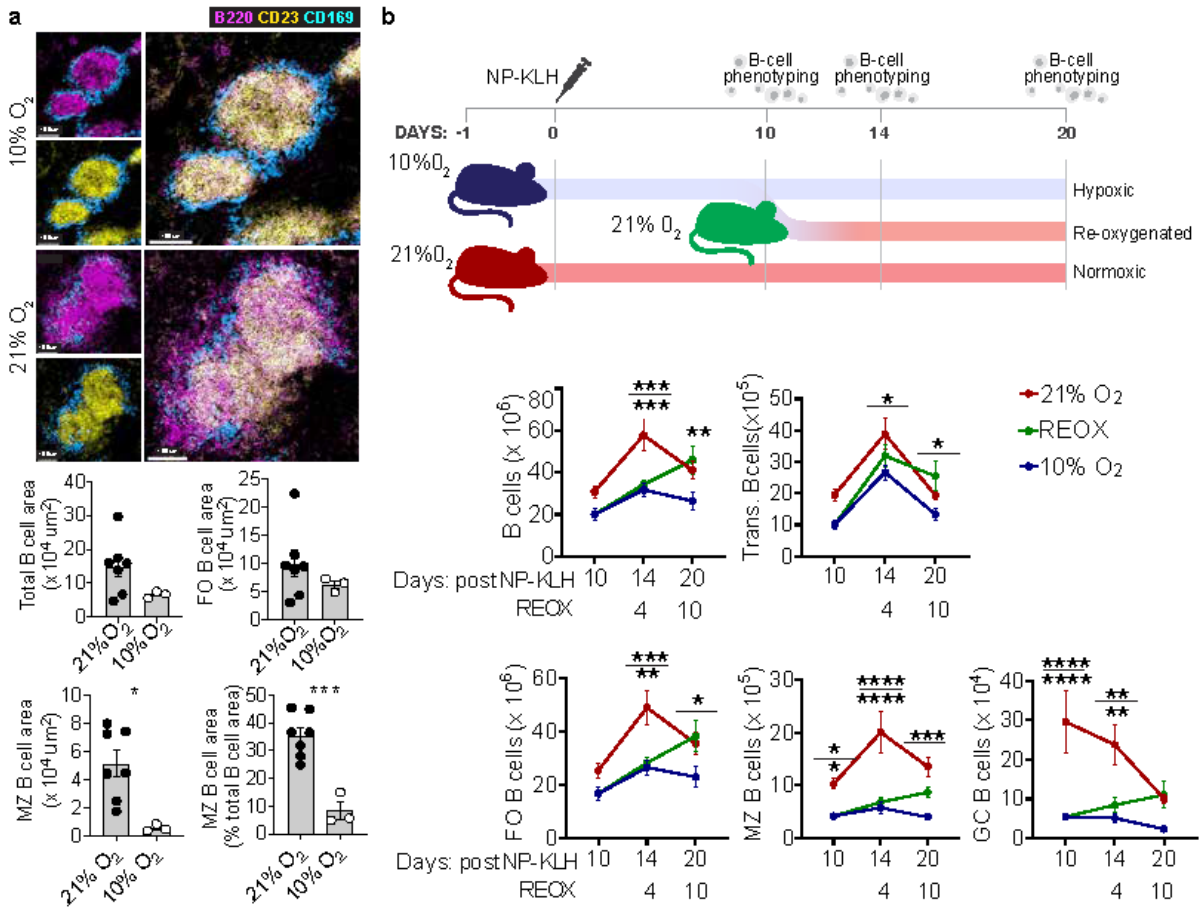


Fig 4. The response of mouse B cells to re-oxygenation *in vivo*.

a, Spleen confocal images from immunised mice (Fig 3a), MZ B cells (magenta, $B220^{+}CD23^{-}$), FO B cells (yellow, $B220^{+}CD23^{+}$) and MZ metallophilic macrophages (blue, $CD169^{+}$). B cell area, circles represent individual follicles from one spleen per condition, mean ± s.e.m. **b,** Experiment outline and absolute spleen B cells, (two-way ANOVA with Tukey's multiple comparisons **test).** Gating in methods. Mean ± s.e.m, data pooled from two experiments. $*P < 0.05, **P < 0.01, ***P < 0.001, **** P < 0.0001$.

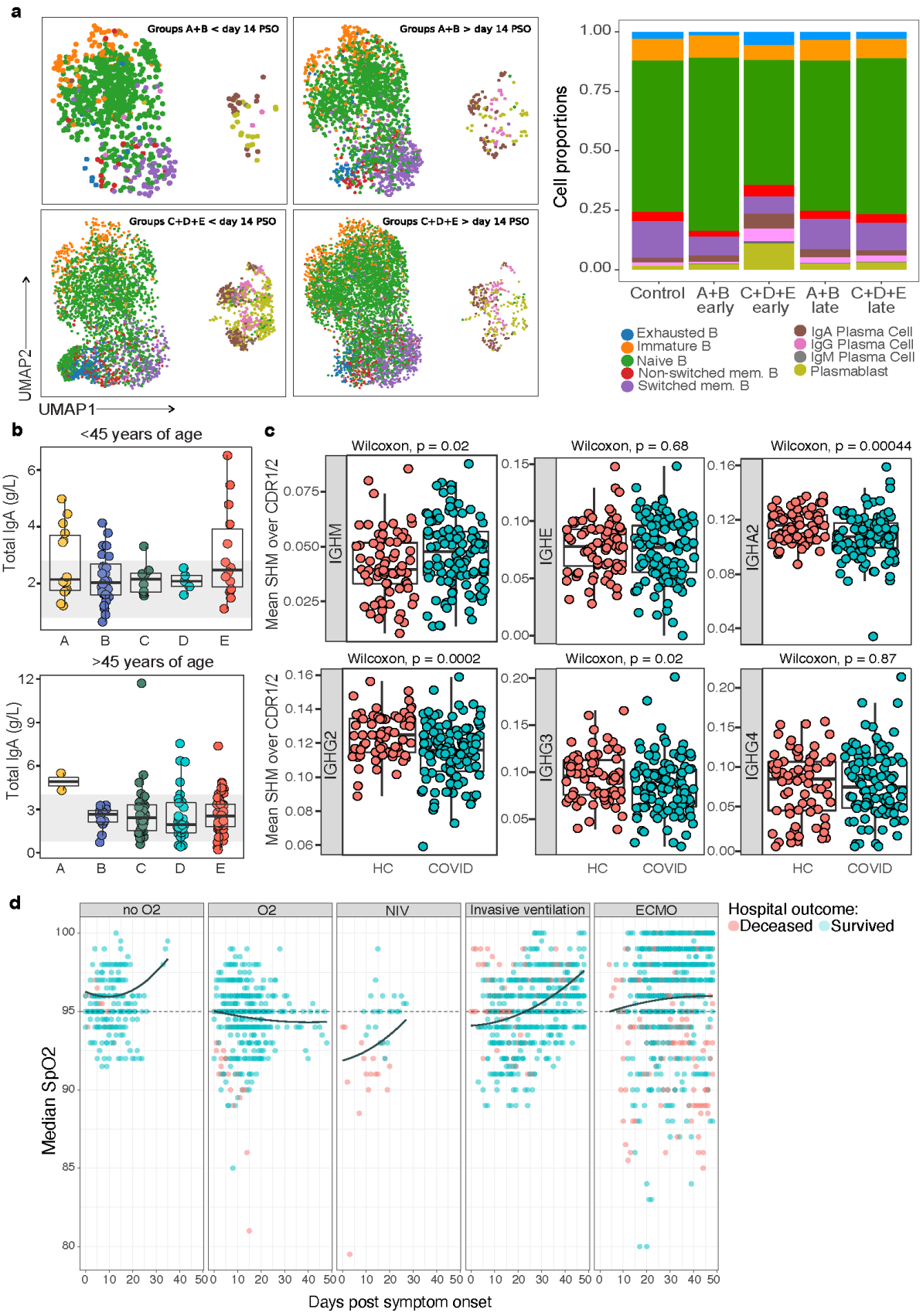

a
Groups A+B < day 14 PSO
Groups A+B > day 14 PSO
Groups C+D+E < day 14 PSO
Groups C+D+E > day 14 PSO
UMAP2
UMAP1
Cell proportions
1.00
0.75
0.50
0.25
0.00
Control
A+B early
C+D+E early
A+B late
C+D+E late
Exhausted B
Immature B
Naive B
Non-switched mem. B
Switched mem. B
IgA Plasma Cell
IgG Plasma Cell
IgM Plasma Cell
Plasmablast
b
<45 years of age
Total IgA (g/L)
A
B
C
D
E
>45 years of age
c
Wilcoxon, p = 0.02
Wilcoxon, p = 0.68
Wilcoxon, p = 0.00044
Mean SHM over CDR1/2
IGHM
IGHE
IGHA2
Wilcoxon, p = 0.0002
Wilcoxon, p = 0.02
Wilcoxon, p = 0.87
IGHG2
IGHG3
IGHG4
HC
COVID
d
no O2
O2
NIV
Invasive ventilation
ECMO
Hospital outcome:
Deceased
Survived
Median SpO2
Days post symptom onset

Fig S1. B cell changes, clinical severity and hypoxia in COVID-19 patients

a, Distribution of participant age and gender across severity categories. **b,** UMAP of B cell populations according to disease severity and days from symptom onset. Bar plot of the mean proportion of B cell populations. **c,** Heatmap showing the $\log_2$ fold change in median absolute cell count between COVID-19 cases in groups C, D, and E, split according to persisting or resolving CRP, and HCs. 12-day time bins. (Wilcoxon test FDR adjusted p value), $^{*}p < 0.05$, $^{**}p < 0.005$, $^{***}p < 0.0005$. **d,** Level of total IgA (g/L) detected in serum of COVID-19 cases at the time of enrolment divided into $\leq$ 45 and >45 years old age groups. Grey band correspond to 5-95th centile ranges based on UK Caucasian population published in the Protein Reference Unit handbook (9th Edn). **e,** Somatic hypermutation frequency calculated over the CDR1/CDR2 regions using BCR sequencing of whole blood, comparing COVID-19 cases and HCs, according to isotype, at 0-12 days post symptom onset (Wilcoxon test FDR adjusted p-value). **f,** Median Oxygen saturations of patients with COVID according to days from symptom onset and level of oxygen supplementation. **d,e** Circles represent individual donors.

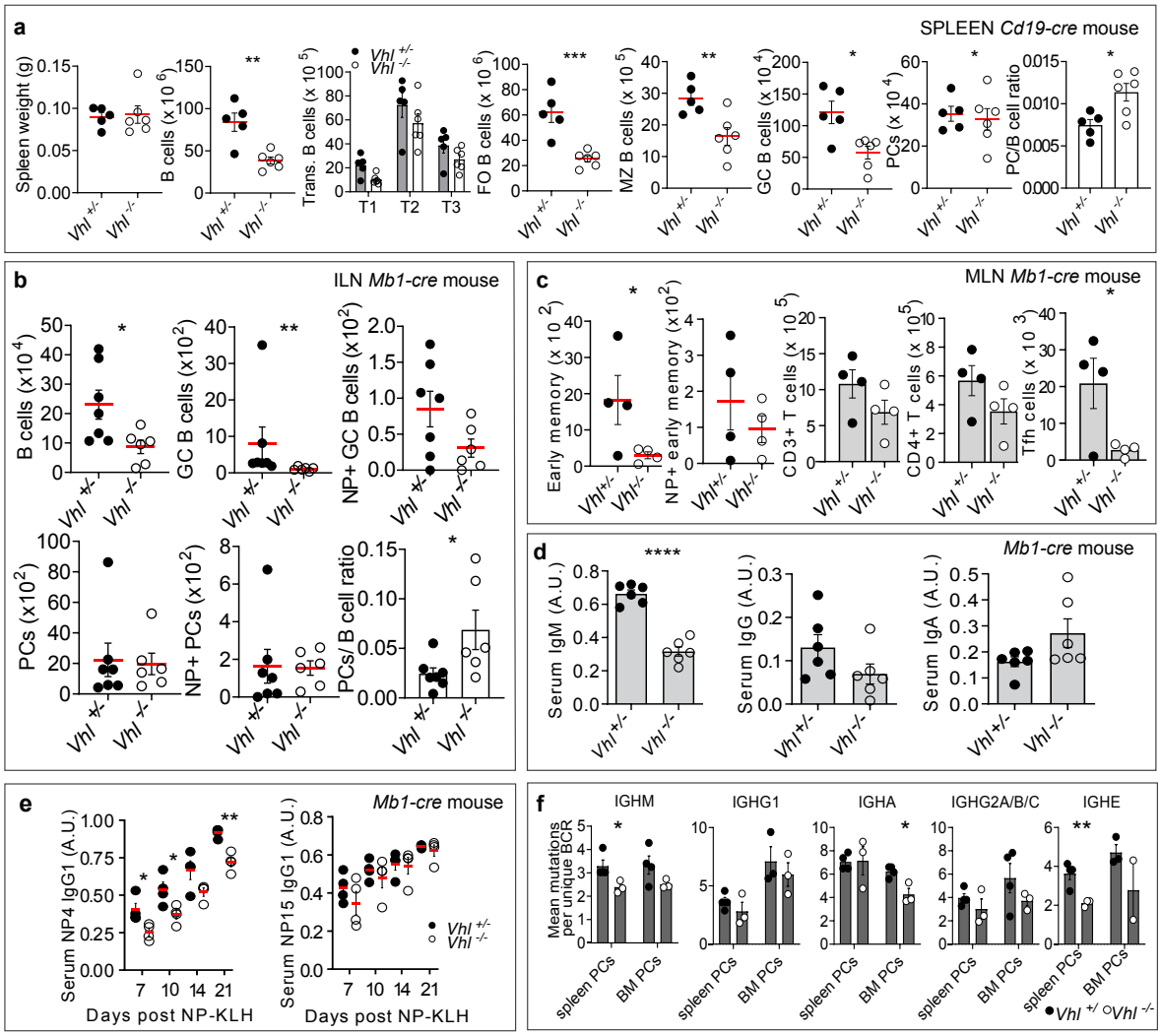


Fig S2. Constitutive HIF activation in mice leads to reduced antigen-specific B cells, T_{FH} cells and SHM.

a, B cell flow cytometric data from *Vhl*$^{+/-}$*Cd19-cre* and *Vhl*$^{-/-}$*Cd19-cre* mice 10 days post sheep red blood cell (SRBC) immunization; Spleen B cells, Transitional, FO, MZ, GC B cells and PCs. **P* < 0.05,***P* < 0.01,****P* < 0.001 (unpaired, two-sided Student's t-test)**.** B cell flow cytometric data from *Vhl*$^{+/-}$*Mb1-cre* and *Vhl*$^{-/-}$*Mb1-cre* mice 21 days post NP-KLH immunization in **b,** inguinal lymph node (ILN); total, GC, NP+ GC, PC, NP+ PC and PC:B cell ratio displayed. PCs were not increased in absolute number, but the PC:B cell ratio was consistently increased. **P* < 0.05,***P* < 0.01 (unpaired, two-sided Student's t-test); GC B cells

$P < 0.01$ (two-sided Mann-Whitney test), **c, mesenteric lymph node (MLN) total and NP+ early memory B cells, T cells and T follicular helper (T_{FH}) cells, displayed. *$P < 0.05$ (unpaired, two-sided Student's t-test). **d,** Serum IgM, IgG and IgA in naïve *Vhl*$^{+/-}$*Mb1-cre* and *Vhl*$^{-/-}$*Mb1-cre* mice, by ELISA (A.U., arbitrary units). ****$P < 0.0001$ unpaired, two-sided Student's t-test. **e,** Serum NP-specific antibody titres after NP-KLH immunization in *Vhl*$^{+/-}$*Mb1-cre* and *Vhl*$^{-/-}$*Mb1-cre* mice, by ELISA. *$P < 0.05$,**$P < 0.01$ (two-way ANOVA with Sidak's multiple comparisons test). **f,** Mean base-pair mutations per unique BCR per isotype (relative to reference germline IGHV gene) in naïve *Vhl*$^{+/-}$*Mb1-cre* and *Vhl*$^{-/-}$*Mb1-cre* mice. *$P < 0.05$,**$P < 0.01$ (unpaired, two-sided Student's t-test). **(a-e)** Gating in methods, mean ± s.e.m, circles represent individual mice. **a,** n= 5 *Vhl*$^{+/-}$*Cd19-cre* and 6 *Vhl*$^{-/-}$*Cd19-cre* mice, **b,** n= 7 *Vhl*$^{+/-}$*Mb1-cre* and n= 6 *Vhl*$^{-/-}$*Mb1-cre* mice, **c,e,** n= 4 per genotype **d,** n= 6 per genotype, **f,** n= 4 *Vhl*$^{+/-}$*Mb1-cre* and n= 3 *Vhl*$^{-/-}$*Mb1-cre* mice. **a,c,f,** from one experiment, **b,** data are pooled from two independent experiments, **d**, data represent three independent experiments, **e**, data represent two independent experiments.

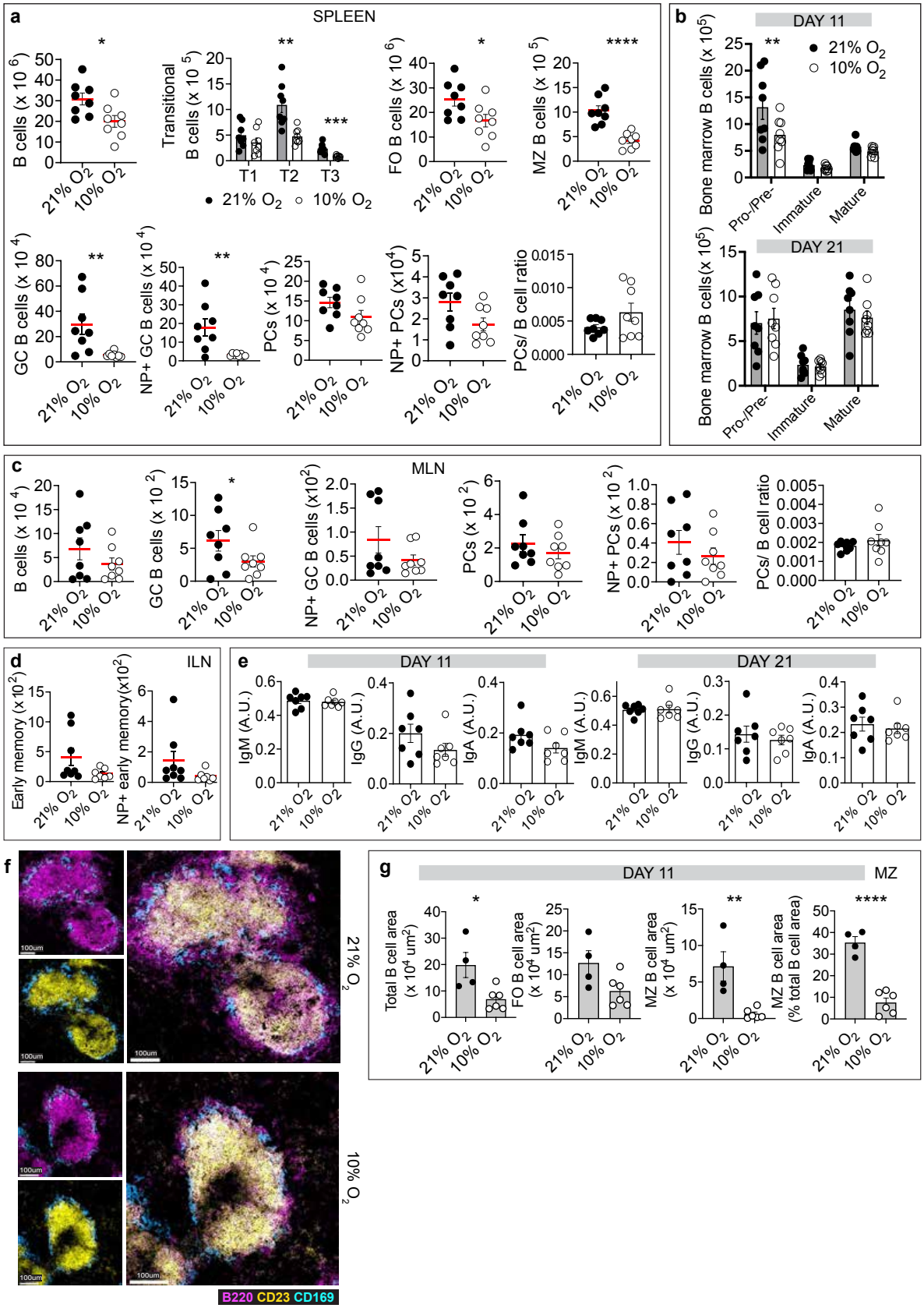


Fig S3. *In vivo* hypoxia leads to marked and persistent B cell defects in mice.

a, WT mice were exposed to 21% O_2 (normoxia) or 10% O_2 (hypoxia), were immunized with NP-KLH at d 1, then at d 11 splenic B cell subsets were enumerated by flow cytometry. Total, transitional, FO, MZ, GC, NP+ GC B cells, PCs, NP+ PCs and PC: B cell ratio is displayed. **b,** MLN; total, GC, NP+ GC, PC, NP+ PC and PC:B cell ratio displayed. **c,** bone marrow B cell subsets from normoxic and hypoxic mice immunised with NP-KLH on d 1 , then enumerated by flow cytometry on d 11 and d 21. Pro-/Pre-B, immature and mature B cells displayed. **d,** ILN; total and NP+ early memory B cells displayed. **e,** Serum IgM, IgG and IgA in normoxic and hypoxic mice immunised with NP-KLH on d 1 then harvested on d 11 or 21, by ELISA (A.U., arbitrary units). **f,** Representative spleen confocal images from normoxic and hypoxic mice immunised with NP-KLH on d 1 and harvested on d 11. MZ B cells (pink, $B220^+CD23^-$), FO B cells (yellow, $B220^+CD23^+$) and MZ metallophillic macrophages (blue, $CD169^+$). B cell area, circles represent individual follicles from one spleen per condition, mean ± s.e.m. **a,b,d-f**, $*P < 0.05$,$**P < 0.01$,$***P < 0.001$, $****P < 0.0001$ (unpaired, two-sided Student's t-test). **c,** $**P < 0.01$ (two-way ANOVA with Sidak's multiple comparisons test). **(a-d,f)** $n = 8$ 21% O_2 and 8 10% O_2, **(e)** $n = 7$ 21% O_2, 7 10% O_2, individual mice. **(a-d,f)** Data pooled from two independent experiments, results confirmed in a third. **(e)** Data, represents three independent experiments. **(a-e)** Circles represent individual mice, mean ± s.e.m.

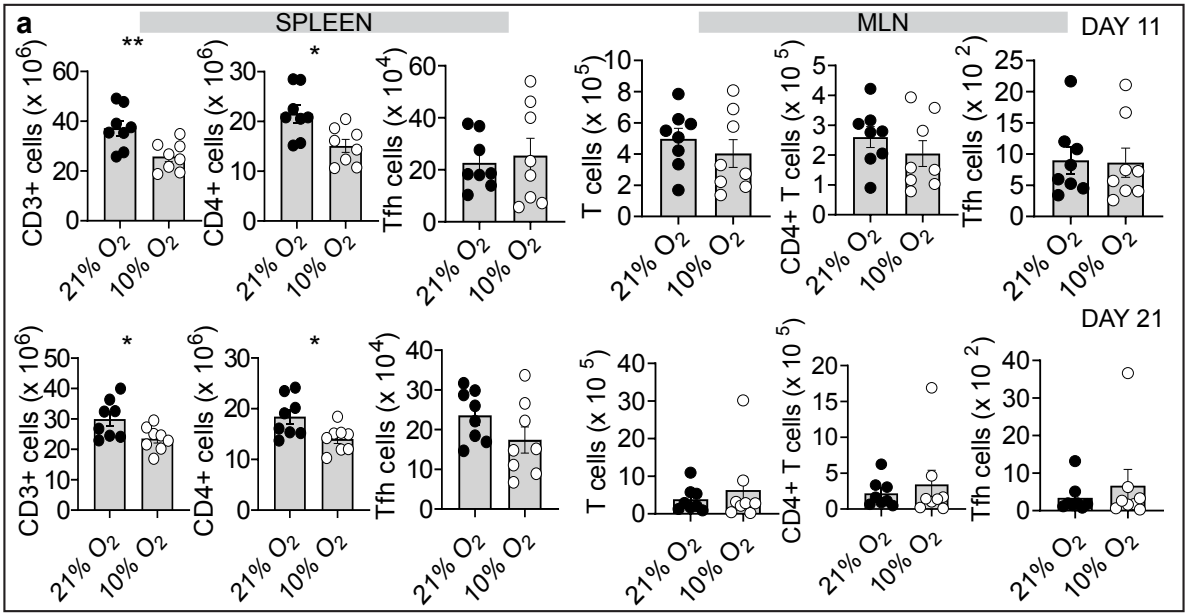


Fig S4. *In vivo* hypoxia has minor effects on T cells after 11 or 21 day exposures in mice

T cells and Tfh cells from normoxic and hypoxic mice immunised with NP-KLH on d 1 then harvested on d 11 or 21, gating in methods. **P* < 0.05,***P* < 0.01 (unpaired, two-sided Student's t-test). *n* = 8 21% O_2 and 8 10% O_2. Data pooled from two independent experiments. Circles represent individual mice, mean ± s.e.m.