

Doctor of Philosophy

Single-cell Analysis of Antigen-Specific T-cells in SARS-CoV2 Infection and Memory

by

Sam Farrar

October 2025

To Mama and Papa.

Declaration

Key parts of this work relied on collaborations. It is thanks to work carried out by others prior to my DPhil that the data exists for me to analyse in this thesis.

All data relating to Oxford SARS-CoV2 convalescent cohorts, with the exception of a single 4-year timepoint sample which I collected, were collected, processed, isolated and sequenced by a collaboration including Xuan Yao, Yanchun Peng, Elie Antoun, Zixi Yin, and Alexander J Mentzer.

Samples relating to the Sheffield Healthcare Worker Cohort were collected by Thushan de Silva and Hailey Hornsby. They were processed, isolated and sequenced by members of the Dong Lab, in particular including Xuan Yao, Yanchun Peng and Elie Antoun.

Parts of the Introduction chapter have been adapted for "The challenge and promise of studying antigen-specific T-cells in human infection, cancer and autoimmunity", under review at Nature Reviews Immunology.

This thesis is the product of work carried out between October 2022 and October 2025 under the supervision of Professor Tao Dong, Professor Julian C. Knight and Dr Yanchun Peng. I declare that, unless otherwise stated, all work presented in this thesis is my own.



Sam Farrar

Word Count: 49519

Abstract

T-cells play crucial roles in the response to viral infection, but studies of the formation, phenotypes and maintenance of T-cell memory have been limited by rare, heterogeneous infections and immature technologies for ex-vivo characterisation of antigen-specific cells. The emergence of SARS-CoV2, which rapidly achieved high levels of infection with dominant variants as well as the development and refinement of technologies for isolating and studying antigen-specific single cells presents a massive opportunity. In this thesis, I make use of human single-cell sequencing data of T-cells specific to various SARS-CoV2 epitopes to discern the phenotypes and repertoires of antigen-specific T-cells in memory. First, I study an immunodominant SARS-CoV2 membrane epitope - $M_{172\text{-}188}$, which is presented by multiple HLA-DR alleles. Supporting functional and structural data generated by my lab and collaborators, I find that DRB1*01:01 restricted cells show enhanced effector gene expression (*GZMB*, *XCL1*, *CCR5*) to DRB1*07:01 restricted cells that associates with higher interpolated antigen concentration of the epitope on DRB1*01:01 in live viral co-culture. Transcriptional differences were not maintained at four years post initial infection. Second, I investigate $CD8^+$ T-cells specific to five immunodominant viral epitopes including Spike and non-Spike longitudinally through breakthrough Omicron infection. Within the cohort some individuals were primed by infection or vaccination, and I found that exposure sequence rather than just number can shape memory phenotypes. Infection-primed individuals (infection followed by three mRNA vaccines) maintain cells expressing *EOMES*, *PRDM1*, and *XCL2*, whereas vaccine-primed individuals (three mRNA vaccines followed by Omicron breakthrough infection) express *IL10*. Moreover, I find that the persistence of clonotypes through Omicron breakthrough infection relates to their transcriptional cluster at the pre-Omicron timepoint, indicating that cells in effector-like clusters (GNLY (0), ABI3 (3)) were more likely to persist. Third, comparing Spike and non-Spike specific $CD4^+$ T-cells at four years post infection (n=16), I show that non-Spike repertoires experience significant Shannon diversity reductions over time whereas Spike-specific repertoires maintain their diversity. At four years, Spike-specific cells maintain interferon-stimulated gene signatures and cytotoxic markers while non-Spike cells express genes associated with anergy. The conditions of TCR stimulation such as HLA restriction and antigen load, as well as their sequence, create durable effects on T-cell memory. Across all studies, small sample sizes necessitated pseudobulking and led to very few differentially enriched pathways. This work provides insights into the development of long-term T-cell memory, as well as how vaccination and breakthrough infection shape T-cell phenotypes.

Acknowledgements

First, I would like to thank my supervisors Professor Tao Dong, Professor Julian Knight and Dr Yanchun Peng for their guidance and support. The data I have had the privilege of working with in this DPhil has been unique, and the product of immense work to recruit for, sample and process. I am especially grateful for the freedom and autonomy provided to me by you, allowing me to go on various tangents (like the structural bioinformatics summer school) and experiment with lots of things that didn't end up working, as well as some that did!

To the members of the Dong Lab and the COI, thank you for making it worth coming in every day despite my DPhil being entirely doable remotely. In particular, thank you to Maya Pidoux for welcoming me and showing me how to process blood on my first day and ensuring we in the lab had a social life. Thank you to Elie Antoun for all the work you do cataloguing and organising the data - just finding samples I needed wouldn't have been possible without your help. Thank you also for all the time you've offered to my various projects. Thank you also to Elie, Adam and Del for all the discussions around single-cell and for being such great sounding boards for my analyses. Thank you to Xuan Yao, whose experimental work underpins Chapter 3 and who has been a wonderful collaborator, and was beyond patient in my short re-introduction to the lab for some validation experiments.

To my parents: thank you for being inspirations to me and for everything you do for me. It would be impossible to list all the things to be grateful for here, but know that none of this would have been possible without you. Mama, danke für alles.

To Hattie, thank you for your patience in the periods where I had bitten off more than I could chew. Thank you for keeping me grounded and for your unwavering support even through that final submission push.

Table of contents

Declaration v

Abstract vii

Acknowledgements ix

List of Figures xvii

List of Tables xxxiii

Glossary xxxvii

Abbreviations xxxvii

Terminology xxxviii

1 Introduction 1

Summary 1

1.1 Adaptive Immune Memory 2

1.2 T-cell Biology 5

1.2.1 Structure, diversity and function of T-cell receptors 5

1.2.2 $CD4^+$ T-cells 11

1.2.3 $CD8^+$ T-cells 12

1.2.4 Control of T-cells by Innate Immunity 13

1.2.5 T-cell Activation and Priming 16

1.2.6 T-Cell Memory Subsets 18

1.3 T-Cell Responses to SARS-CoV2 19

1.3.1 A brief introduction to SARS-CoV2 . . . 19
1.3.2 Dynamics of the T-cell response to SARS-CoV2 . . . 22
1.3.3 T-cell Targets and Immunodominance in SARS-CoV2 . . . 32
1.4 Investigating Antigen-Specific T-cells using Single-Cell Techniques . . . 35
1.4.1 Challenges of Studying Antigen-Specific T-cells . . . 36
1.4.2 Methods for Defining Antigen Specific T-cells . . . 38
1.4.3 Applications of Studying Antigen-Specific T-cells . . . 53
1.5 Rationale, Knowledge Gaps and Thesis Aims . . . 61
1.5.1 How does HLA restriction shape T-cell memory in an immunodominant SARS-CoV2 Viral Epitope? . . . 61
1.5.2 How do spike-specific and non-spike-specific T-cells after vaccination evolve with breakthrough Omicron infection? . . . 61
1.5.3 How do the long-term phenotypes and repertoires of Spike and non-Spike specific CD4⁺ T-cells differ over time? . . . 62

2 Materials and Methods 63

Summary . . . 63
2.1 Data Collection . . . 64
2.1.1 Study Cohorts and Definitions . . . 64
2.1.2 Establishment of Antigen-Specific T-cell Lines . . . 65
2.1.3 Tetramer Sorting for Antigen-specific CD4⁺ T-cells . . . 65
2.1.4 Tetramer-associated magnetic enrichment and cell sorting for four-year CD4⁺ T-cells . . . 66
2.1.5 Tetramer Sorting for SARS-CoV2 specific CD8⁺ T-cells . . . 67
2.1.6 Intracellular Cytokine Staining for Antigen-Specific Responses . . . 67
2.1.7 HLA-Mismatched B-cells and Blocking Antibodies To Identify HLA-Restriction . . . 68
2.1.8 SARS-CoV2 Live Virus Propagation and Titration . . . 68
2.1.9 EBV-Transformed B-Cell Lines . . . 69
2.1.10 Intracellular Cytokine Staining (ICS) for SARS-CoV2 Infection . . . 69

2.1.11 Identification of Optimal Epitopes using IFN-γ ELISPOT 69
2.1.12 Single-Cell Sequencing of CD4⁺ T-cells of Different Restrictions . . 70
2.2 Data Analysis 70
2.2.1 Statistical Analysis 70
2.2.2 Prediction of HLA class II presentation of $M_{172\text{-}188}$ 70
2.2.3 Single-Cell RNA Sequencing 70

3 HLA Restriction shapes the phenotype and functional profile of CD4⁺ T-cells responsive to the same SARS-CoV2 antigenic peptide 75

Summary 75
3.1 Introduction 77
3.2 Results 78
3.2.1 SARS-CoV2 $M_{172\text{-}188}$ elicits a dominant CD4⁺ T-cell response across a cohort with diverse HLA class II types 78
3.2.2 Multiple HLA-DR alleles and HLA-DP present SARS-CoV2 $M_{172\text{-}188}$ 81
3.2.3 HLA DR-presented epitopes share a similar core region 85
3.2.4 $M_{172\text{-}188}$ responsive CD4⁺ clones' responses to live SARS-CoV2 virus co-culture associate with restricting HLA-allele. 86
3.2.5 Different HLA-DR alleles present different interpolated peptide concentrations of $M_{172\text{-}188}$, with DRB1*01:01 presenting the highest. 87
3.2.6 Single-cell sorting of HLA-DR restricted $M_{172\text{-}188}$ response in the Recovery Phase. 90
3.2.7 Single-cell transcriptomic characterisation reveals differential gene expression between DRB1*01:01 and DRB1*07:01 restricted $M_{172\text{-}188}$-specific T-cells 92
3.2.8 Repertoires are highly diverse, with frequent V-gene pairings shared between tetramers 95
3.2.9 Similar or identical TCRs sort to different tetramers, indicating the potential for cross-restriction 98
3.2.10 $M_{172\text{-}188}$ specific T-cells are observed 4 years after primary SARS-CoV2 exposure, but almost no T-cell clones are maintained. 100

3.3 Discussion . . . 102
3.4 Methods . . . 108
3.4.1 Interpolation of $M_{172\text{-}188}$ concentration in live viral co-culture . . . 108
3.4.2 Predicted Structural Analysis . . . 108

4 Comparing the evolution of Spike and non-Spike specific CD8⁺ T-cells in Breakthrough SARS-CoV2 Omicron infection 111

Summary . . . 111
4.1 Introduction . . . 112
4.2 Results . . . 114
4.2.1 Study cohort and epitope-specific responses . . . 114
4.2.2 Transcriptional clusters define CD8⁺ T-cell states . . . 115
4.2.3 Differential gene expression between order of exposure and number of exposures . . . 121
4.2.4 T-cell receptor repertoire diversity is similar across 1 to 5 exposures 127
4.2.5 Clonal persistence is associated with more clone members in Effector-like than Memory-like clusters . . . 127
4.3 Discussion . . . 131
4.4 Methods . . . 136
4.4.1 Single-Cell RNA Processing . . . 136
4.4.2 Dimensionality Reduction and Clustering . . . 136
4.4.3 Differential Gene Expression . . . 136
4.4.4 Data Transformation for Clonal Persistence Analysis . . . 137

5 The long term evolution of Spike specific CD4⁺ T-cells compared to those directed at an immunodominant Membrane epitope after SARS-CoV2 infection. 139

Summary . . . 139
5.1 Introduction . . . 140
5.2 Results . . . 144
5.2.1 Study cohort and epitope-specific responses . . . 144

5.2.2 Transcriptional profiles of Spike and Non-Spike specific CD4⁺ T-cells 146

5.2.3 Comparison of Epitope-specific Repertoires of Spike and Non-Spike specific CD4⁺ T-cells . . . 155

5.2.4 Differential gene expression between Spike and Non-Spike specific CD4⁺ T-cells . . . 161

5.2.5 Exploratory analysis of priming differences between Spike and non-Spike specific CD4⁺ T-cells in Acute SARS-CoV2 . . . 168

5.3 Discussion . . . 171

5.4 Methods . . . 176

5.4.1 Dimensionality Reduction and Clustering . . . 176

5.4.2 TCR Repertoire Analysis . . . 177

5.4.3 Differential Gene Expression . . . 178

6 Discussion **181**

Summary . . . 181

6.1 Context and Rationale . . . 182

6.2 Unifying Themes . . . 182

6.2.1 How Exposure shapes T-cell Phenotypes . . . 182

6.2.2 How Sequences of Exposures Combine . . . 184

6.2.3 How T-cell Phenotypes affect their Clonal Persistence . . . 185

6.3 Statistical Challenges in Small-n Single-Cell Studies . . . 186

6.4 Methodological Challenges . . . 188

6.5 Conclusions and Future Directions . . . 188

References **191**

Appendices **217**

A Appendix - Introduction **217**

A.1 Innate trained immunity . . . 217

B Appendix - Materials and Methods **219**

B.1 Representative Intracellular Cytokine Plots 219

C Appendix - Chapter 3 **221**

C.1 Gating strategy for isolating $M_{172\text{-}188}$ specific $CD4^+$ T-cells with tetramers . 221

C.2 T-cell based interpolation of effective pulsed peptide concentration in live viral $M_{172\text{-}188}$ concentration 222

C.3 $M_{172\text{-}188}$ specific $CD4^+$ T-cells isolated at each timepoint, by tetramer and by patient 224

C.4 QC metrics of single cell RNA-seq data of $M_{172\text{-}188}$ specific $CD4^+$ T-cells . . 224

D Appendix - Chapter 4 **227**

D.1 Dynamics of per-Sample Cluster proportion between pre- and post-Omicron timepoints 228

D.2 Gene Ontology 2025 Biological Process Enrichment of Differentially Expressed Genes 232

D.3 Paired T-cell Receptor Yield from Samples 242

D.4 Logistic Regression of Clonotype Maintenance 242

E Appendix - Chapter 5 **245**

E.1 Quality Control Metrics for Spike and Non-Spike specific $CD4^+$ T-cells . . 245

E.2 Principal Components Weightings by Principal Component 246

E.3 Interpretation of Transcriptional Clusters Genes 249

E.4 Cell Numbers by Epitope Group and Timepoint 249

E.5 Gene Usage Frequencies between 1-3 Month and 4 Year Timepoints after Initial SARS-CoV2 Infection 249

E.6 Gene Set Enrichment Analysis Tables 255

List of Figures

1.1 An illustrative schematic of how the primary and secondary responses compare in adaptive immunity 3

1.2 SARS-CoV2 infection and MHC class I and II presentation. Adapted from BioRender templates "MHC Class I and II Pathways" by Akiko Iwasaki and "Darunavir: Potential Repurposed Drug Candidate for COVID-19". 8

1.3 Key stages in dendritic cell activation and their priming of naive T-cells. Adapted from Biorender Template "How Dendritic Cells Become Competent to Stimulate T Cells" by Akiko Iwasaki 15

1.4 SARS-CoV2 genome organisation, structural proteins and immune evasion mechanisms. Image created using BioRender.com combining templates 'SARS-CoV-2 Genomic Organization' by Goldman-Israelow, B., Ona, S., and 'Coronavirus Proteins' Gu, J and information from Steiner et al. 2024 and Sette et al. 2021. 21

1.5 IEDB Immunome Browser (accessed 21/04/2025) of the SARS-CoV2 Membrane Protein, depicting the number of assays that have been carried out on a given region (top) and the frequency of responses observed in those assays (bottom). 34

1.6 Measures of epitope binding degeneracy from T-cell receptor repertoires by Mayer and Callan, PNAS, 2023, Figure 3[47]. Original Caption: "Excess coincidences follow a common functional form across experiments. Sequence similarity of specific T-cells for paired $\alpha\beta$-chain repertoires (Top), α-chain repertoires (Middle) and β-chain repertoires (Lower) compared with background expectations. In each panel, the assay type used to enrich for epitope-specific T-cells and the antigen source are noted in the upper right. Panel C is special as analysed TCRs are from unsorted blood and have not been explicitly selected for binding to a specific epitope. A common reference curve is plotted for visual guidance. Its parameter K is set equal to observed value at $\Delta = 0$. Z is determined by normalisation." 49

3.1 Bar Chart showing the distribution of $M_{172\text{-}188}$ responders and non-responders by clinical severity. Asymptomatic and mild cases were aggregated into mild, whereas severe and critical were aggregated into severe. Original ELISPOT data sourced from[95], analysis and visualisation carried out by me. 79

3.2 Scatter plot of SARS-CoV2 Membrane Protein Amino Acid residue number against COG-UK mutation frequency. Each dot is a particular amino acid mutation, e.g. (S187G). Original data sourced from COG-UK[354], data visualised by me. 80

3.3 SARS-CoV2 Membrane protein mutation frequencies across all residues in COG-UK database. Bar plot shows cumulative mutation frequencies at every position of the M protein. Four major hotspots are annotated, which represent mutations that are fixed in major lineages. The $M_{172\text{-}188}$ epitope is highlighted in yellow, illustrating its low mutation frequency compared to the rest of the protein. 81

3.4 Each Pie Chart represents the set of individuals that carry that specific HLA allele (individuals can be present in more than one due to heterozygosity), showing the distribution of clinical severity and whether individuals responded to $M_{172\text{-}188}$ by IFN-γ ELISPOT. Original ELISPOT data sourced from[95], analysis and visualisation carried out by me. 82

3.5 Proportions of cohort (n=44) responders and non-responders that carry a particular HLA-DRB1 allele. Orange bars indicate the proportion of HLA-DRB1 allele carriers that respond to $M_{172\text{-}188}$, whereas green indicates the proportion of carriers of that allele that do not respond to $M_{172\text{-}188}$. Grey bars indicate the proportion of the cohort as a whole that carry the indicated HLA-DRB1 allele. Original ELISPOT and HLA-typing data sourced from[95], analysis and visualisation carried out by me. 83

3.6 Using mismatched B-cell lines with known HLA-types as antigen-presenting cells for various $M_{172\text{-}188}$ responsive $CD4^+$ T-cell clones, we could identify clones restricted to $M_{172\text{-}188}$ presented by HLA-DRB1*01:01, *07:01 and *15:01. Work carried out by Dr Xuan Yao, figure generated by Dr Xuan Yao. 84

3.7 Positive, tetramer staining populations exist within $M_{172\text{-}188}$ responsive T-cells for tetramers of HLA-DRB1*01:01, DRB1*07:01 and DRB1*04:01. The first two panels are from the same donor, the final panel is from a different donor. Work carried out by Dr Xuan Yao, figure generated by Dr Xuan Yao. 84

3.8 Relative IFN-γ response (compared to full $M_{172\text{-}188}$ epitope) by T-cell clones of known restriction to various truncations of $M_{172\text{-}188}$ epitope. Work carried out by Dr Xuan Yao, figure generated by Dr Xuan Yao. 85

3.9 Structures of $M_{172\text{-}188}$ presented by HLA-DRB1*01:01 (left) and HLA-DRB1*07:01 (right), solved by Prof. Jamie Rossjohn's group. Middle panel represents the two structures overlaid, showing extensive overlap of the $M_{172\text{-}188}$ conformation across presenting HLA alleles. 86

3.10 Flow cytometry plots of bulk $M_{172\text{-}188}$ responsive T-cells under Intracellular Cytokine Staining for IFN-γ and TNF-α with different SARS-CoV2 Variants of Concern. For these experiments, B-cell lines were infected with live SARS-CoV2. Work carried out by Dr Xuan Yao. Figure generated by Dr Xuan Yao. 87

3.11 (a) Percentage viral RNA suppression by $CD4^+$ T-cell clones with different HLA-DRB1 restrictions in the live-viral co-culture assay. Suppression calculated as the percentage difference between control (no T-cells) and T-cell clone condition. (b) Percentage IFN-γ by ICS of $CD4^+$ T-cell clones with different restricting HLA-DRB1 alleles in the live-viral co-culture assay. Work carried out by Dr Xuan Yao, figures generated by Dr Xuan Yao. 88

3.12 Cartoon of experimental design for interpolation of effective pulsed concentration of live SARS-CoV2 infection. ACE2 expressing B-cell lines are pulsed with $M_{172\text{-}188}$ peptide at varying concentrations or infected with SARS-CoV2. IFN-γ positivity for the peptide titration is used to generate a three-parameter log-logistic curve, against which the IFN-γ positivity in the viral co-culture condition can be used to interpolate the effective pulsed concentration of the antigen. Experimental data generated by Dr Xuan Yao, figure and curve fitted by me. 90

3.13 Dot plot of Interpolated $M_{172\text{-}188}$ concentration of Live Viral Co-Culture. ACE2 transduced B-cells from donors with known HLA alleles that present $M_{172\text{-}188}$ (1180: DRB1*01:01 & DRB1*07:01, 1201: DRB1*04:01 & DRB1*15:01, 1493: DRB1*07:01 & DRB1*15:01) were pulsed with $M_{172\text{-}188}$ peptide at varying concentrations and a three-parameter log-logistic curve was fitted to each cell line with known restriction. Each dot reflects a particular $M_{172\text{-}188}$ responsive cell line with known restriction from a donor, indicated by their colour. Experimental data generated by Dr Xuan Yao, interpolation and generation of this figure was carried out by me. . . . 91

3.14 Proportion of Tetramer Sorted T-cells in samples that were stained with CCR7 and CD45RA via flow-cytometry during index-sorting that classify as Central Memory, Effector Memory, Naive or Terminal Effector Memory (TEMRA). Stratified by tetramer they sort to. Dots represent measurements, diamonds represent the mean. Ranges show the standard deviation of the proportions within that category. Flow Cytometry carried out by Dr Xuan Yao, analysis and plotting carried out by me. 93

3.15 A volcano plot summarising differentially expressed genes identified via DESeq2 over pseudobulked samples between DRB1*01:01 specific $CD4^+$ T-cells (greater expression on the left), and DRB1*07:01 specific $CD4^+$ T-cells (greater expression on the right). DESeq2 was carried out using the Wald test and a design formula integrating the patient of origin, and a combination term of tetramer and stimulation. 93

3.16 A volcano plot showing the DESeq2 differentially expressed genes over pseudobulked samples between DRB1*01:01 specific $CD4^+$ T-cells (greater expression on the left), and DRB1*07:01 specific $CD4^+$ T-cells (greater expression on the right). DESeq2 was carried out using the Wald test and a design formula integrating the patient of origin and a combination term of tetramer and stimulation. 94

3.17 Box and whisker plots showing the proportion of the overall paired repertoire with specific Alpha V genes and Beta V genes, with shaded boxes corresponding to the IQR range of V gene frequency within a sample (the cells specific for a given tetramer, isolated at the Recovery Phase time point, for a given patient). Genes that made up less than 5% of the repertoire were aggregated into "Other". 96

3.18 Box and whisker plots showing the proportion of the overall paired repertoire with specific Alpha J genes and Beta J genes, with shaded boxes corresponding to the IQR range of J gene frequency within a sample (the cells specific for a given tetramer, isolated at the Recovery Phase time point, for a given patient). Genes that made up less than 5% of the repertoire were aggregated into "Other". 97

3.19 Annotated graph of the adjacency matrix of $M_{172\text{-}188}$ clonotypes. Each node corresponds to a clonotype in a particular patient, with the radius corresponding to the log of the clonal size, and their colour corresponding to their tetramer-sorted restriction for $M_{172\text{-}188}$ presented by DRB1*01:01 (red) or DRB1*07:01 (blue). Clonotypes from cell lines responsive to $M_{172\text{-}188}$ with unknown restriction (due to maintenance via stimulation) are in grey. Edges connect TCRs with a combined edit-distance of less than or equal to 3, and a per-chain edit distance of less than or equal to 2. Edges are coloured by type: Public Clonotype (identical CDR3s, different patients; blue), Public Meta-Clonotype (similar TCRs, different patients; lightblue), or Meta-Clonotype (grey). Orange shaded areas highlight similar clonotypes where members of the meta-clonotype sort to both DRB1*01:01 and DRB1*07:01. (Right Panel) Representative paired sequence logos for three cross-restricted meta-clonotypes, displaying V/J gene usage and CDR3 amino acid sequences for alpha (top) and beta (bottom) chains within each cluster. Metadata indicates the HLA alleles (DRB1*01:01, DRB1*07:01) and patient identifiers present in each depicted cluster. 99

3.20 CDR3β contributes a larger proportion of the Buried Surface Area of the TCR:pMHC interaction in DRB1*07:01 (blue, labelled as DR7) than DRB1*01:01 (red, labelled as DR1) presented $M_{172\text{-}188}$. Violin Plots show the distribution of buried surface area (BSA) contributions from each CDR loop, as calculated by the Shrake Rupley algorithm for high-quality predicted structures. Significance was determined by Mann-Whitney U test with Bonferroni correction; ** = $p < 0.01$, ns = not significant. 101

3.21 The Clonotype dynamics between sampling after first infection and subsequent sampling at 4 years after the initial infection reveal minimal maintenance of paired clonotypes. Only one paired clone, in P1525, was present at both the initial and the four year timepoint. Where clones were maintained, they made up a larger proportion of the repertoire at the four year timepoint than at the recovery phase, but we could not exclude low sampling as a driver of this. 103

4.1 Schematic of the cohort longitudinally tracking epitope specific CD8⁺ T-cells through Omicron breakthrough infection. Healthcare workers primed by infection (n=10, top) or primed by vaccination (n=4, bottom) were sampled after the completion of a full course of 3 vaccinations and then again after Omicron infection. Infection-primed individuals were initially infected with SARS-CoV2, received 3 vaccine doses then had their first samples taken. In vaccine primed individuals, their only exposure to SARS-CoV2 antigen was via the vaccination course. Then, following PCR confirmed Omicron infection both groups were sampled again. 116

4.2 Number of tetramer-sorted CD8⁺ T-cells recovered before and after Omicron breakthrough infection labelled by epitope specificity. Tetramer-specific CD8⁺ T-cells specific for Spike (teal) and non-Spike (red) epitopes from 14 healthcare workers with either infection-primed (left) or vaccine primed (right) infection histories. Lines connect samples of the same epitope in the same patient before and after Omicron infection. Most paired samples (11/12) came from infection primed individuals. This is not a measure of the quantity or frequency of circulating cells, as the number isolated is usually multiples of a 96 well plate. 117

4.3 UMAP of 3157 CD8⁺ T-cells from Sheffield healthcare workers, coloured by cluster. The clusters are defined by the most significantly differentially expressed genes in each cluster, filtered to those with a log2 fold change greater than 0.1, and difference in percentage of cells expressing that transcript compared with other cells of 0.1. 118

4.4 Canonical CD8⁺ T-cell markers amongst differentially expressed genes between clusters. Heatmap shows scaled average expression across transcriptional clusters of SARS-CoV2 specific CD8⁺ T-cells. Expression levels are calculated as the mean log-normalised counts per cluster and scaled across clusters. Columns are hierarchically clustered. 120

4.5 Stacked bar chart coloured by clone size categories within each transcriptional cluster. Clone sizes are calculated within each sample on the basis of paired $\alpha\beta$ amino acid TCR sequences. For visualisation, clone sizes were categorised as: singleton (1 cell), small (2-5 cells), medium (6-10 cells), or large (>10 cells). 121

4.6 Cluster composition by priming exposure and timepoint. Stacked bar charts show the distribution of cell clusters separated by priming exposure (Vaccine-primed top row, Infection-primed bottom row) and timepoint (pre-Omicron left column, post-Omicron right column). The GNLY (0) cluster is more frequent in infection-primed individuals, while the ABI3 (3) cluster makes up a larger proportion of vaccine-primed cells. 122

4.7 UMAP visualisation of 3157 SARS-CoV2 antigen-specific CD8⁺ T-cells coloured by metadata categories. (1) Priming Exposure: Infection-primed (purple) vs Vaccine primed (orange). (2) Epitope Group: Spike (green) vs Non-Spike (red). (3) Pre- or Post- Omicron Timepoint: Samples collected before breakthrough infection at the completion of the vaccine course (pre, cyan) vs Samples collected after breakthrough infection (post, pink). (4) Paired $\alpha\beta$ TCR clone size. 123

4.8 Volcano plot showing differential gene expression between pre-Omicron and post-Omicron Spike-specific CD8⁺ T-cells in vaccine-primed individuals. Points represent individual genes, with significantly differentially expressed genes highlighted. Manually identified genes relevant to CD8⁺ T-cell state are labelled, showing enrichment of effector/exhaustion markers (*EOMES*, *PTPN12*, *DUSP6*) pre-Omicron and activation/metabolic markers (*CD38*, *NFATC1*, *RUNX1*) post-Omicron. 124

4.9 Volcano plot showing differential gene expression between pre-Omicron and post-Omicron Spike-specific CD8⁺ T-cells in infection-primed individuals. Points represent individual genes, with significantly differentially expressed genes highlighted. Manually identified genes relevant to CD8⁺ T-cell state are labelled, showing proliferation related genes and PD-L1 (*MKI67*, *CD274*). 125

4.10 Volcano plot showing differential gene expression between cells with four Spike exposures in different orders. Points represent individual genes differentially expressed between vaccine-first (three mRNA followed by omicron infection) and infection-first (infection, then three mRNA), with significantly differentially expressed genes highlighted. Manually identified genes relevant to CD8⁺ T-cell state are labelled, showing TCR negative regulator, CD8⁺ memory formation and chemokine genes (*DUSP6*, *EOMES*, *XCL2*) and regulatory cytokine (*IL10*). 126

4.11 Shannon diversity index of paired $\alpha\beta$ TCR sequences within each sample, coloured by epitope target and arranged by the inferred number of exposures. Samples where fewer than five paired TCRs were detected were excluded, leading to analysis of 28 samples including 8 pairs. 127

4.12 Shannon diversity index of paired $\alpha\beta$ TCR sequences within each sample, faceted by epitope type (Spike or Non-Spike), and infection or vaccine priming. Samples where fewer than five paired TCRs were detected were excluded, leading to analysis of 28 samples including 8 pairs. 128

4.13 UMAP Dimensionality Reduction showing only cells with a detected paired $\alpha\beta$ TCR. Each cell is overlaid with its detection status (newly detected, maintained, not detected again) based on its presence at pre- and post-Omicron timepoints. Not detected again means no cell with the same paired TCR as that cell at the pre-Omicron timepoint is detected at the post-Omicron timepoint in the same individual specific for the same epitope. Maintained means, with the same constraints, that the same paired TCR is detected at both timepoints. Newly detected indicates that that clonotype was not detected in that individual and epitope at the pre-Omicron timepoint 128

4.14 Mosaic of Cluster Annotation and Detection Status before Omicron. Mosaics represent all paired-chain TCR clonotypes detected in that cluster before Omicron infection. Mosaics are coloured by whether that clonotype sampled at the pre-Omicron timepoint was subsequently detected at the post-Omicron timepoint (detected again, green) or not (not detected again, orange). 129

4.15 Forest plot showing associations with paired clonotype maintenance from pre- to post-Omicron timepoints. Points represent odds ratios (OR) with 95% confidence intervals from logistic regression. Cluster coefficients represent the effect of having cells in that cluster relative to the CCR7 (1) reference cluster. Model controlled for priming exposure, post-Omicron cell count, and epitope specificity. 130

4.16 Mosaic of Cluster Annotation and Detection Status after Omicron. Mosaics represent all paired-chain TCR clonotypes detected in that cluster after Omicron infection. Mosaics are coloured by whether that clonotype sampled at the post-Omicron timepoint was previously detected at the pre-Omicron timepoint (detected again, green) or not (newly detected, purple). 131

5.1 Bar chart of severity distribution by epitope sampling group. “Both” refers to individuals where Spike and non-Spike epitopes were sampled. “Non-Spike only” refers to individuals where only non-Spike epitopes were sampled. “Spike only” refers to individuals where only Spike epitopes were sampled. 145

5.2 A linked dot plot showing cell counts for Spike and non-Spike specific $CD4^+$ T-cells. Non-Spike specific cells are shown in red ($M_{172\text{-}188}$ presented by HLA-DR1) and yellow ($M_{172\text{-}188}$ presented by HLA-DR7). Spike-specific cells are shown in shades of green (S_{166}, S_{751} and S_{866}) 147

5.3 Heatmap of manually curated genes characterising cell clusters. Colours reflect the average expression of gene over all cells in the cluster scaled and centred across clusters. Dendrogram above the heatmap represents hierarchical clustering over the scaled and normalised expression matrix. . . 148

5.4 UMAP Projection of 3429 Sampled CD4⁺ Single Cells, coloured by cluster. 149

5.5 Cell counts (left) and overall proportions (right) of CD4⁺ T-cells within each transcriptional cluster at each timepoint, split by epitope type (Spike and Non-Spike). 150

5.6 Whisker plot of the proportion of each sample within each transcriptional cluster. Box plots show the distribution of cluster proportions in samples, split by the epitope the sample is specific for. Each point represents an outlier sample (outside the IQR for the frequencies). 151

5.7 Distribution of transcriptional clusters across patients. (A) Cluster composition within each patient, stratified by timepoint (1-3 Months on the left, 4 Years on the right) and epitope type (Non-Spike above, Spike below). (B) Patient contribution to each cluster, showing that the ZNF165 (4) cluster is predominantly composed of cells from patients 1180 and 1525. 153

5.8 Stacked bar chart coloured by clone size categories within each transcriptional cluster. Clone sizes are calculated within each sample on the basis of paired $\alpha\beta$ amino acid TCR sequences. For visualisation, clone sizes were categorised as: singleton (1 cell), small (2-5 cells), medium (6-10 cells), or large (>10 cells). 154

5.9 Whisker plot of Shannon diversity indexes of paired $\alpha\beta$ TCR repertoires per sample within each transcriptional cluster. Box plots show the distribution of Shannon diversity indexes in samples, split by the transcriptional cluster the sample is specific for. Each point represents the diversity within a sample-cluster combination. 155

5.10 UMAP Projection of SARS-CoV2 specific CD4⁺ T-cells coloured by metadata categories including timepoint, epitope specificity, donor and the severity of initial infection by WHO classification. Epitopes are labelled S34, S151 and S174 in this figure, corresponding to $S_{166\text{-}180}$, $S_{751\text{-}765}$ and $S_{866\text{-}880}$ respectively, and M24DR01 and M24DR07 correspond to $M_{172\text{-}188}$ presented by DRB1*01:01 and DRB1*07:01. 156

5.11 Temporal changes in $\alpha\beta$ TCR Repertoire Diversity in Spike specific (n=26) and non-Spike specific (n=11) CD4⁺ TCR repertoires. Shannon diversity of paired $\alpha\beta$ repertoires was calculated for sample, and lines connect paired samples where the same epitope was sampled in the same individual at multiple timepoints. Each point represents one epitope-specific repertoire, coloured by Epitope identity. Box plots show the median and interquartile ranges of diversity. Epitopes are labelled S34, S151 and S174 in this figure, corresponding to $S_{166\text{-}180}$, $S_{751\text{-}765}$ and $S_{866\text{-}880}$ respectively, and M24DR01 and M24DR07 correspond to $M_{172\text{-}188}$ presented by DRB1*01:01 and DRB1*07:01. 157

5.12 Forest plot of the linear regression model fit to Shannon diversity of TCR repertoires directed at Spike and non-Spike epitopes at 1-3 months and four years. Points represent effect estimates, and horizontal bars indicate 95% confidence intervals. Spike-specific repertoires show moderate diversity loss (green), while the interaction term (red) suggests that Non-Spike specific repertoires show greater diversity loss over time. The estimated total change of Non-Spike repertoire diversity from 1-3 months to four years is the sum of these (purple). 158

5.13 Raw clonotype persistence rates and clone size distributions. Top panel shows the proportion of Alpha, Beta and Paired clonotypes that persist from 1-3 months to four years for Spike and non-Spike specific repertoires. Bottom panel shows the distribution of sample Simpson's clonalities at the early (1-3 Month) timepoint. Spike specific samples have more clones where more than one cell shows that paired clonotype, leading to greater clonalities of each sample. In the top panel, epitopes are labelled S34, S151 and S174, corresponding to $S_{166\text{-}180}$, $S_{751\text{-}765}$ and $S_{866\text{-}880}$ respectively. 159

5.14 Forest plot of predictors of Alpha, Beta and Paired clonotype persistence from the early (1-3 Month) to late (four years) timepoint. Pink, purple and blue bars represent the confidence intervals for Paired, Beta and Alpha clonotypes respectively. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ 160

5.15 The distribution of DESeq2 Normalised Pseudobulk counts of the 9 most significantly differentially expressed genes in the `T Cell Receptor Signaling Pathway (GO:0050852)` ontology. 162

5.16 Volcano plot showing differential gene expression between Spike and Non-Spike-specific CD4⁺ T-cells at 1-3 Months post-infection. Points represent individual genes, with significantly differentially expressed genes (adjusted p < 0.05, |log2FC| > 1) highlighted in red. Genes discussed in the text are highlighted in the figure. 163

5.17 Heatmap of manually curated genes differentially expressed between Spike (Green) and Non-Spike (Red) specific CD4⁺ T-cells at 1-3 Months post-infection. Heatmap elements reflect the pseudobulked, normalised expression of each gene in each sample (combination of donor, epitope and timepoint). . . . 164

5.18 Heatmap of manually curated genes differentially expressed between Spike (Green) and Non-Spike (Red) specific CD4⁺ T-cells at four years post-infection. Heatmap elements reflect the pseudobulked, normalised expression of each gene in each sample (combination of donor, epitope and timepoint). . . . 165

5.19 Volcano plot showing differential gene expression between Spike and Non-Spike specific CD4⁺ T-cells at four years post-infection. Points represent individual genes, with significantly differentially expressed genes (adjusted p < 0.05, |log2FC| > 1) highlighted in red. Genes discussed in the text are highlighted in the figure. . . . 166

5.20 Differential expression dynamics of genes differentially expressed at both early and late timepoints. Scatter plot comparing log2 fold changes under DESeq2 differential gene expression of genes significantly differently expressed (adjusted $p < 0.05$) at both early and late timepoints. Each point represents a gene, and genes relevant to CD4⁺ state or function are highlighted. Colours indicate expression patterns - genes consistently highly expressed by Spike are green, Non-Spike are red and those with switched greater expression are in purple or blue. . . . 167

5.21 Major T-cell subset distribution for COMBAT T-cells with TCRs similar to sampled Spike or non-Spike specific CD4⁺ T-cells. The vast majority (94.8%, 73/77) of matched clonotypes were annotated in the COMBAT CITE-Seq dataset as CD4⁺ T-cells. A small number of matches occur in cells annotated as CD8⁺ T-cells or MAIT cells in the COMBAT dataset. . . . 168

5.22 COMBAT disease category distribution for clones matched to sampled Spike or non-Spike specific CD4⁺ T-cells. . . . 169

5.23 Presence, absence or non-availability of the restricting HLA allele for unique similar clonotypes for epitopes sampled in this study. . . . 170

5.24 The distribution of `Annotation_minor_subset` of T-cells with similar clonotypes to sampled Spike and non-Spike specific CD4⁺ T-cells. The left panel reflects the exact counts of cells, whereas the right panel represents the proportions. . . . 171

B.1 Representative flow cytometry plot for intracellular cytokine staining of antigen-specific T-cells co-cultured with EBV transformed B-cell lines pulsed with $M_{172\text{-}188}$ peptide. 219

B.2 Representative flow cytometry plot for intracellular cytokine staining of antigen-specific T-cells co-cultured with EBV transformed, ACE2 transduced B-cell lines infected with SARS-CoV2. 220

C.1 Representative gating strategy for isolating $M_{172\text{-}188}$ specific $CD4^+$ T-cells restricted to DRB1*01:01, DRB1*07:01 and DRB1*04:01. 221

C.2 Dose response curves for DRB1*01:01 and DRB1*07:01 restricted $M_{172\text{-}188}$ specific $CD4^+$ T-cell lines of known restriction were co-cultured with EBV transformed B-cell lines pulsed with varying concentrations of $M_{172\text{-}188}$ peptide. IFN-γ production was measured by intracellular cytokine staining. Each panel shows a distinct B-cell line donor, T-cell line donor and HLA restriction. Curves represent the fit of the three-parameter log-logistic model with 95% confidence intervals (shaded regions). EC50 values (the concentration required for 50% of maximal response) are shown in each panel. Red dashed lines show interpolation of the IFN-γ response observed in co-culture with ACE2-transduced B-cells infected with SARS-CoV2. Experimental data generated by Dr Xuan Yao, model fitting and figure generation by me. 222

C.3 The number of cells isolated, at both the Recovery Phase (1-3 months after acute infection) and at Four Year Followup. Cells specific for DRB1*01:01 are highlighted in the red bars, whereas those specific for DRB1*07:01 are in blue. 223

C.4 A violin plot of QC metrics for $M_{172\text{-}188}$ specific $CD4^+$ T-cells isolated in the 1-3 month recovery phase split by donor id: the number of unique genes detected per cell (nFeature_RNA), the number of unique transcripts detected per cell (nCount_RNA), and the percentage of mitochondrial gene counts per cell (percent.mt). 224

C.5 A violin plot of QC metrics for $M_{172\text{-}188}$ specific $CD4^+$ T-cells isolated at the four year followup split by donor id: the number of unique genes detected per cell (nFeature_RNA), the number of unique transcripts detected per cell (nCount_RNA), and the percentage of mitochondrial gene counts per cell (percent.mt). 225

D.1 ABI3 (3) cluster dynamics between pre- and post-Omicron timepoints. Scatter plot showing changes in ABI3 (3) cluster proportions from pre- to post-Omicron timepoints for each sample. Where samples were taken for the same epitope in the same individual (longitudinal), they are connected. Lines are coloured by priming exposure (infection vs. vaccine), and point size represents absolute cell counts. Plots are faceted by spike vs. non-spike epitopes. 228

D.2 CCR7 (1) cluster dynamics between pre- and post-Omicron timepoints. Scatter plot showing changes in CCR7 (1) cluster proportions from pre- to post-Omicron timepoints for each sample. Where samples were taken for the same epitope in the same individual (longitudinal), they are connected. Lines are coloured by priming exposure (infection vs. vaccine), and point size represents absolute cell counts. Plots are faceted by spike vs. non-spike epitopes. 229

D.3 GNLY (0) cluster dynamics between pre- and post-Omicron timepoints. Scatter plot showing changes in GNLY (0) cluster proportions from pre- to post-Omicron timepoints for each sample. Where samples were taken for the same epitope in the same individual (longitudinal), they are connected. Lines are coloured by priming exposure (infection vs. vaccine), and point size represents absolute cell counts. Plots are faceted by spike vs. non-spike epitopes. 230

D.4 KIF23 (4) cluster dynamics between pre- and post-Omicron timepoints. Scatter plot showing changes in KIF23 (4) cluster proportions from pre- to post-Omicron timepoints for each sample. Where samples were taken for the same epitope in the same individual (longitudinal), they are connected. Lines are coloured by priming exposure (infection vs. vaccine), and point size represents absolute cell counts. Plots are faceted by spike vs. non-spike epitopes. 231

D.5 MALAT1 (2) cluster dynamics between pre- and post-Omicron timepoints. Scatter plot showing changes in MALAT1 (2) cluster proportions from pre- to post-Omicron timepoints for each sample. Where samples were taken for the same epitope in the same individual (longitudinal), they are connected. Lines are coloured by priming exposure (infection vs. vaccine), and point size represents absolute cell counts. Plots are faceted by spike vs. non-spike epitopes. 232

D.6 The yield of paired TCRs from samples. Each line represents a unique patient-epitope combination tracked from pre- to post-Omicron timepoints. Lines are coloured by priming exposure (red = infection, blue = vaccine). 29 patient-epitope combinations that included any TCRs are shown. 242

E.1 Violin Plot of control measurements, including number of features (genes detected per cell), counts (the total counts of all genes in a cell), and the percent of overall counts attributed to mitochondrial genes of all samples of Spike and non-Spike CD4⁺ T-cells. Violin Plots are grouped by the sequencing run in which the cells were sequenced. 245

E.2 Elbow plot of the standard deviation of each principal component in the PCA of Log Normalised RNA counts of Spike and Non-Spike specific CD4⁺ T-cells. 246

E.3 Heatmap representing the 5 most positively and 5 most negatively weighted genes for each principal component of components 1-10. 247

E.4 Heatmap of differentially expressed genes in each cluster. The top 5 most differentially expressed genes per cluster are shown. 247

E.5 Violin Plot of Percentage Mitochondrial RNA counts (percent.mt), Number of Genes Detected (nFeature_RNA) and Number of RNA counts (nCount_RNA) per cell by cluster. 248

E.6 Cell numbers by Epitope Group and Timepoint. 248

E.7 Volcano plot of mean change in TCR gene usage proportion against $\log_{10}$ p-value across all epitopes from early to four year timepoint. Each point represents an individual V or J gene, with red indicating genes that significantly changed in frequency. 249

E.8 TCR V gene usage temporal comparison for S166-specific CD4⁺ T-cells. Boxplots of the proportion of V genes at the early (1-3 months, blue) versus long-term (4 years, red) timepoints. 250

E.9 TCR V gene usage temporal comparison for S751-specific CD4⁺ T-cells showing alpha and beta chain V gene proportions at early versus late timepoints. 250

E.10 TCR V gene usage temporal comparison for S866-specific CD4⁺ T-cells. . . 251

E.11 TCR V gene usage temporal comparison for M24DR01-specific CD4⁺ T-cells showing relatively stable V gene repertoire composition over 4 years of follow-up. 251

E.12 Boxplots of TCR V-gene usage at early (1-3 months) and four year time-point for M24DR07-specific CD4⁺ T-cells. 252

E.13 TCR J gene usage temporal comparison for S166-specific CD4⁺ T-cells showing alpha and beta chain J gene proportions. 252

E.14 TCR J gene usage temporal comparison for S751-specific CD4⁺ T-cells. . . . 253

E.15 Boxplots of TCR J-gene usage at early (1-3 months) and four year timepoint for S866-specific CD4⁺ T-cells. 253

E.16 TCR J gene usage temporal comparison for M24DR01-specific CD4⁺ T-cells showing maintained J gene repertoire structure over long-term follow-up. . 254

E.17 TCR J gene usage temporal comparison for M24DR07-specific CD4⁺ T-cells with stable alpha and beta chain J gene usage frequencies between recovery and memory phases. 254

List of Tables

1.1 Table of key studies supporting Sette model of COVID-19 pathogenesis . . 25

3.1 Clinical metadata of the $M_{172\text{-}188}$ single-cell cohort. SARS-CoV2 Severity reflects the severity of the initial SARS-CoV2 infection. Age is the age (years) at the first SARS-CoV2 infection. Days between Symptoms and Sampling reflect the difference in self-reported initial symptom date and the date of sampling. Recovery phase indicates whether a sample was used for single-cell sequencing at the early (37-64 days post-infection) timepoint, and the Four Year sample indicates whether single-cell sequencing was carried out on $M_{172\text{-}188}$ responsive cells four years after the initial infection. . . 91

4.1 Clinical metadata of Sheffield healthcare worker cohort by priming exposure. Tetramer^{+} CD8^{+} T-cells recovered represents the numbers of wells, within the Smart-seq2 protocol, that contained quantifiable RNA (pre-QC). 114

4.2 Immunodominant SARS-CoV2 CD8 Epitopes used in Sheffield Healthcare worker study of Antigen-Specific CD8 T-cell evolution. 114

5.1 The HLA restrictions of SARS-CoV2 epitopes relevant to tetramer sorting and cell isolation in this study. Determined by, among others, Dr Yanchun Peng, Dr Guihai Liu and Dr Xuan Yao. 144

5.2 Summary of the sampling groups in this study, with the number of donors, median age (with interquartile range) and distribution of sex and severity of infection. Mod, Sev, Crit stand for Moderate, Severe and Critical respectively. 145

5.3 Logistic regression of clonotype persistence. Models composed of Epitope Type (Spike vs Non-Spike), log of clone size at 1-3 months, and log of overall sampling depth (i.e. the number of cells with the chain(s) observed) at four years. P-values are adjusted for multiple testing using the Benjamini-Hochberg method. OR = odds ratio; CI = confidence interval. 160

D.1 Table of top Gene Ontology Biological Process terms (2025 database) from gene set enrichment analysis of genes differentially expressed in Spike-specific CD8⁺ T-cells pre-Omicron vs post-Omicron breakthrough infection in vaccine primed individuals. Genes with adjusted p-value < 0.05 and positive `log2FoldChange` (more abundant pre-Omicron) were analysed using Enrichr. Shows the top 10 most enriched terms. 232

D.2 Table of top Gene Ontology Biological Process terms (2025 database) from gene set enrichment analysis of genes differentially expressed in Spike-specific CD8⁺ T-cells pre-Omicron vs post-Omicron breakthrough infection in vaccine primed individuals. Genes with adjusted p-value < 0.05 and negative `log2FoldChange` (more abundant post-Omicron) were analysed using Enrichr. Shows the top 10 most enriched terms. 234

D.3 Table of top Gene Ontology Biological Process terms (2025 database) from gene set enrichment analysis of genes differentially expressed in Spike-specific CD8⁺ T-cells pre-Omicron vs post-Omicron breakthrough infection in infection primed individuals. Genes with adjusted p-value < 0.05 and positive `log2FoldChange` (more abundant pre-Omicron) were analysed using Enrichr. Shows the top 10 most enriched terms. 236

D.4 Table of top Gene Ontology Biological Process terms (2025 database) from gene set enrichment analysis of genes differentially expressed in Spike-specific CD8⁺ T-cells pre-Omicron vs post-Omicron breakthrough infection in infection primed individuals. Genes with adjusted p-value < 0.05 and negative `log2FoldChange` (more abundant post-Omicron) were analysed using Enrichr. Shows the top 10 most enriched terms. 237

D.5 Table of top Gene Ontology Biological Process terms (2025 database) from gene set enrichment analysis of genes differentially expressed between Spike-specific CD8⁺ T-cells in individuals primed by vaccination post-Omicron (3x mRNA, infection) vs after completion of vaccine course in infection primed individuals (infection, 3x mRNA). Genes with adjusted p-value < 0.05 and positive `log2FoldChange` (more abundant in vaccine first) were analysed using Enrichr. Shows the top 10 most enriched terms. . 239

D.6 Table of top Gene Ontology Biological Process terms (2025 database) from gene set enrichment analysis of genes differentially expressed between Spike-specific CD8⁺ T-cells in individuals primed by vaccination post-Omicron (3x mRNA, infection) vs after completion of vaccine course in infection primed individuals (infection, 3x mRNA). Genes with adjusted p-value < 0.05 and negative `log2FoldChange` (more abundant in infection-first) were analysed using Enrichr. Shows the top 10 most enriched terms. 240

D.7 Example values for logistic regression model predicting clonotype maintenance. Each row represents a unique clonotype (paired $\alpha\beta$ TCR) detected at the pre-Omicron timepoint in a specific patient-epitope combination. . 242

D.8 Model coefficients for logistic regression of clonotype maintenance. For clonotypes detected at pre-Omicron timepoint, the model predicts the odds of detecting it again at the post-Omicron timepoint based on transcriptional cluster composition, priming exposure, epitope group, and post-Omicron cell count. CCR7 (1) serves as the reference cluster; coefficients for other clusters represent effects relative to CCR7 (1). Reference categories: vaccine priming, non-Spike epitopes. McFadden's $R^2 = 0.24$; n = 167. 243

E.1 Table of all significantly differentially expressed Gene Ontologies (Gene Ontology Biological Process 2025 database) as well as 10 not meeting the threshold, under gene set enrichment analysis of genes more abundant in Non-Spike specific CD4$^+$ T-cells at 1-3 Months. 52 gene sets were significantly enriched after multiple testing corrections. Genes differentially expressed under DESeq2 with an adjusted p-value below 0.05 and a negative `log2FoldChange` in the comparison between Spike-specific CD4$^+$ pseudobulks at 1-3 months and Non-Spike specific CD4$^+$ pseudobulks at 1-3 months (i.e. more abundant in Non-Spike specific cells) were entered into Enrichr. . 255

E.2 Table of the top 10 most significantly differentially expressed Gene Ontologies (Gene Ontology Biological Process 2025 database) under gene set enrichment analysis of genes more abundant in Spike specific CD4$^+$ T-cells at 1-3 Months. No gene sets reach significance after adjustment for multiple testing. Genes differentially expressed under DESeq2 with an adjusted p-value below 0.05 and a positive `log2FoldChange` in the comparison between Spike-specific CD4$^+$ pseudobulks at 1-3 months and Non-Spike specific CD4$^+$ pseudobulks at 1-3 months (i.e. more abundant in Spike specific cells) were entered into Enrichr. 279

E.3 Table of the top 10 most significantly differentially expressed Gene Ontologies (Gene Ontology Biological Process 2025 database) under gene set enrichment analysis of genes more abundant in non-Spike specific CD4$^+$ T-cells at 4 Years. No gene sets reach significance after adjustment for multiple testing. Genes differentially expressed under DESeq2 with an adjusted p-value below 0.05 and a negative `log2FoldChange` in the comparison between Spike-specific CD4$^+$ pseudobulks at 4 years and Non-Spike specific CD4$^+$ pseudobulks at 4 years (i.e. more abundant in non-Spike specific cells) were entered into Enrichr. 280

E.4 Table of the top 10 most significantly differentially expressed Gene Ontologies (Gene Ontology Biological Process 2025 database) under gene set enrichment analysis of genes more abundant in Spike specific CD4⁺ T-cells at 4 Years. No gene sets reach significance after adjustment for multiple testing. Genes differentially expressed under DESeq2 with an adjusted p-value below 0.05 and a positive `log2FoldChange` in the comparison between Spike-specific CD4⁺ pseudobulks at 4 years and Non-Spike specific CD4⁺ pseudobulks at 4 years (i.e. more abundant in Spike specific cells) were entered into Enrichr. 282

Glossary

Abbreviations

ACE2 Angiotensin-Converting Enzyme 2

AIM Activation-Induced Marker assay

BSA Buried Surface Area

CDR3 Complementarity-Determining Region 3

CI Confidence Interval

COVID-19 Coronavirus Disease 2019

DEG Differentially Expressed Gene

EBV Epstein-Barr Virus

ELISPOT Enzyme-Linked Immunosorbent Spot assay

HIV Human Immunodeficiency Virus

HLA Human Leukocyte Antigen

ICS Intracellular Cytokine Staining

IEDB Immune Epitope Database

IFN Interferon

IQR Interquartile Range

ISG Interferon-Stimulated Gene

$\log_2$FC Log_2 fold change

M Membrane protein of SARS-CoV2

MHC Major Histocompatibility Complex

MIS-C Multisystem Inflammatory Syndrome in Children

N Nucleocapsid protein of SARS-CoV2

OR Odds Ratio

PBMC Peripheral Blood Mononuclear Cell

PCR Polymerase Chain Reaction

pMHC Peptide-Major Histocompatibility Complex

pTM / ipTM Predicted / Interface Predicted Template Modelling score

S Spike protein of SARS-CoV2

SD Standard Deviation

TCR T-cell receptor

TEMRA Terminally differentiated effector memory T-cells re-expressing CD45RA

TNF Tumour Necrosis Factor

UMAP Uniform Manifold Approximation and Projection

Terminology

T-cell A conventional T-lymphocyte expressing α-β T-cell receptors.

Polyfunctionality Secretion, by a T-cell, of more than one cytokine in response to a stimulus.

Hybrid immunity The combination of natural and vaccine-induced immunity.

10x Genomics dataset The dataset of T-cells and counts for 44 DNA-barcoded pMHC multimers released by 10x Genomics as part of the whitepaper[1]

1 Introduction

Summary

Antigen-specific T-cells underlie adaptive immunity and immune memory to infections. However, the difficulty of collecting samples with comparable infection timing or history, on top of the inherent rarity of antigen-specific cells, has made detailed studies into antigen-specific T-cell memory challenging. The emergence of SARS-CoV2 as a global pandemic presents an opportunity whereby these challenges are reduced. Here, I summarise relevant T-cell immunology with respect to development, subtypes and the development of T-cell memory. Then, I introduce SARS-CoV2, and the role of T-cell immunity in COVID-19. By discussing methods of identifying antigen-specific T-cells, and highlighting case-studies of important insights that could be made by studying these rare subsets, I argue that these provide an important window into the immune response that unselected subsets cannot. Finally, I introduce the key aims and rationale of my thesis: to interrogate the role of HLA-restriction in the T-cell memory to an immunodominant $CD4^+$ SARS-CoV2 epitope, to understand how breakthrough infection alters $CD8^+$ T-cell memory to SARS-CoV2, and how $CD4^+$ T-cell responses to Spike and non-Spike epitopes evolve over time after primary infection.

1.1 Adaptive Immune Memory

The adaptive immune system has its name because it can alter its defensive strategies in response to a lifetime of encounters with pathogens or altered self-antigens[2]. The constant evolutionary battle with non- or altered self has selected a system that can adapt with greater speed than the slow process of genetic inheritance[3]. An adaptive immune system provides organisms with one characteristic: they can respond, specifically, to pathogens they have never encountered before in a way that potentiates a future response to the same insult. What did this make possible? Longer lifespans, more complex social structures and immunological strategies across populations. Today, this evolutionary legacy shapes human health. Genetic diversity in the immune response explains some of why individuals vary in their susceptibility to disease[4].

Immune memory allowed us to eradicate smallpox[5]. Vaccines introduce parts of pathogens, triggering immune memory that prevents future infections without causing illness. Now, by designing fragments of cancer cells rather than pathogens into vaccines, researchers can boost immunity to tumours[6]. Our growing ability to engineer the targets of immune memory has generated entire drug classes out of modified cells that attack targets we choose[7].

The immune system is an interconnected set of tissues and their constituent cells that protect the host from threats[2] and facilitate normal tissue development and homeostasis, such as in the brain[8]. Through millions of generations of experimentation, the immune system has developed specialised mechanisms to counter infectious pathogens, altered or mutated self, and controlling components of the immune system when they go wrong. Immunologists traditionally divide the immune system into two subsystems - innate and adaptive, delineated by how these responses are encoded genetically, their speed and whether they have memory. This clean division is convenient and broadly (Section A.1) accurate, and I will use this vocabulary in this thesis.

Effective host defence relies on a coordinated interplay between these two subsystems. The innate response provides immediate protection, whereas the naive adaptive immune system requires days to mobilise fully against a novel target but provides enhanced specificity and immunological memory[9,10]. This memory allows it to respond with greater magnitude and

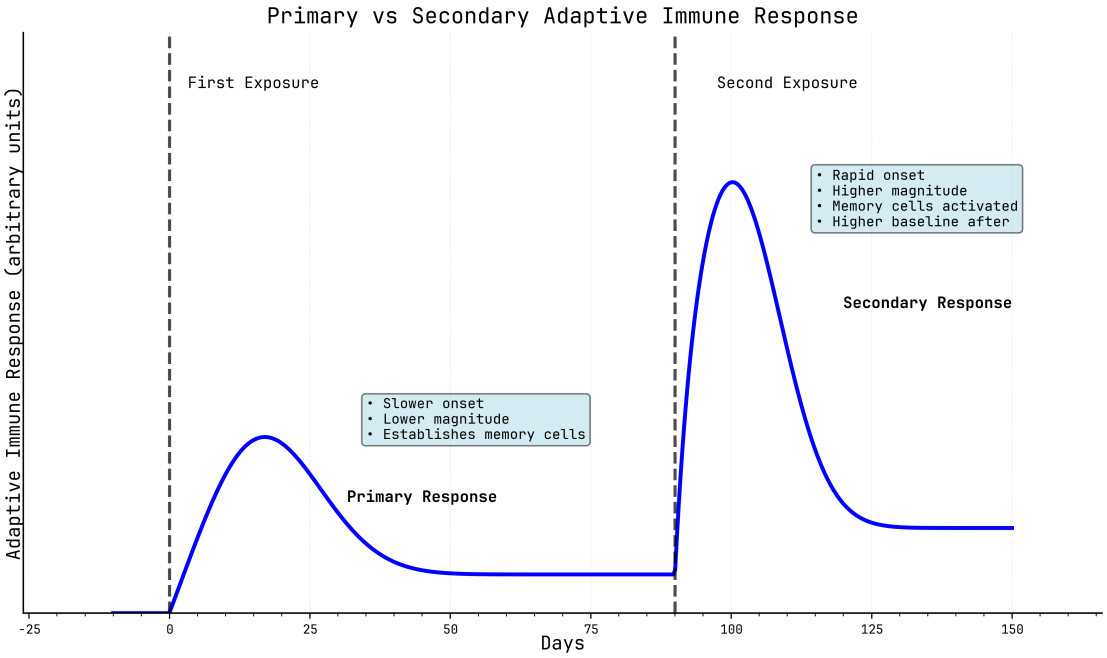


Figure 1.1: An illustrative schematic of how the primary and secondary responses compare in adaptive immunity

speed when it is recalled by subsequent exposure, in particular through resident memory responses which occupy the site of exposure. Following pathogen detection by the innate immune system, inflammatory mediators activate and direct the adaptive response through antigen presentation, co-stimulation and cytokine signalling[11]. Reciprocally, adaptive immune cells calibrate innate immunity through their cytokines and direct innate immune cells through antibody-dependent mechanisms[12]. Disruption of this coordinated response can cause both severe infectious disease[13] as well as problems regulating and resolving infection, leading to autoimmunity[14].

There are two major lineages to Adaptive Immunity, defined by expression of canonical receptors on their surface[2]. B-cells recognise native, unprocessed antigens, whereas T-cells recognise peptides processed and presented on Major Histocompatibility Complexes (MHCs) (although specific T-cell subtypes can recognise non-peptide epitopes). The adaptive immune system is an "anticipatory repertoire"[15] composed of diverse T- and B-cell receptors capable of recognition and memory of previously unseen epitopes. These cells can exert cellular effects, such as killing infected cells[16], facilitating the function of other

immune cells[17] or releasing mediators of humoral immunity like antibodies that directly bind the target antigen[18].

The T-cell receptor allows conventional, alpha-beta T-cells (from here on referred to as T-cells) to detect the intracellular state of other cells on MHC class I molecules, whereas MHC class II molecules on professional antigen-presenting cells can sample the extracellular environment. This provides T-cells a signal as to whether its binding target is present within that cell. To achieve this, MHC molecules on the surface of the surveyed cell act as a molecular display platform. All nucleated cells load peptides onto MHC class I (MHC-I) molecules in the endoplasmic reticulum and traffic them to the cell surface[19]. Passing T-cells scan the MHC molecules of their neighbouring cells, primed to respond to peptides derived from the proteins of the target cell. Since most of these peptides at any given time are derived from the host's self proteome, if a T-cell responds to a person's peptides this can cause autoimmune disease[20,21] or fatal tissue damage[22]. To achieve this fidelity the thymus filters and selects T-cells[23], with additional tolerance-inducing mechanisms in the liver and the periphery[24,25].

T-cells fulfil a variety of functions through specialised subsets. Within an individual, T-cells are specialised to occupy a variety of sites and carry out different but overlapping functions[26,27]. For example, $CD8^+$ cells that reside in tumours and respond to cancer cells have different surface markers and transcriptional profiles than the $CD4^+$ cells in germinal centres that facilitate the generation of antibodies[28]. This diversity in the proteins present on their surface, the effector and communication molecules they release and where they are located is on top of and combinatorial with the diversity of their T-cell receptors.

T-cell heterogeneity was initially explored with a small number of surface markers or functional characteristics[29–31]. For example, CD4 or CD8-expressing T-cells are classified separately based on divergent differentiation trajectories and coreceptor functions. However, the simple separation of T-cells into Cytotoxic $CD8^+$ and Helper $CD4^+$ cells overlooks the expanding nuance of these cells' diverse subsets and atypical subtypes such as Cytotoxic $CD4^+$ and $CD8^+$ T regulatory cells[32,33]. Modern multi-OMIC techniques allow simultaneous measurement of many thousands of features from each individual cell within hundreds of

individuals[34–36], with the promise of allowing us to identify pathology relevant cell types at a resolution previously impossible.

In this thesis, I make use of single-cell sequencing datasets collected longitudinally from individuals infected with SARS-CoV2. By looking specifically at peptide-MHC multimer sorted antigen-specific cells in memory, I can follow the development of antigen-specific T-cell memory and how it differs with presenting HLA-allele or SARS-CoV2 gene of origin.

1.2 T-cell Biology

T-cells are a fundamental component of adaptive immunity, orchestrating and effecting responses to non-self while maintaining tolerance to self[28]. Their importance to our control of pathogens, cancer and autoimmune disease is revealed in their absence. Individuals with Severe Combined Immunodeficiency, a genetic disorder of lymphocyte development caused by several possible mutations, are extremely vulnerable to infectious diseases. So vulnerable, in fact, that despite placement in sterile "bubble" environments they die within a year of birth due to recurrent infections[37]. Crucial T-cell subsets can also be infected, for example in Acquired Immunodeficiency Syndrome the causative virus directly infects $CD4^+$ T-cells as well as generally triggering inflammation. Over years, $CD4^+$ T-cell counts fall until AIDS defining illnesses take hold - opportunistic fungal and viral infections normally controlled by healthy immune systems[38]. While neither of these examples are rigorous evidence of a specific T-cell role, and are interwoven with non-T-cell pathophysiology, they illustrate the important role that T-cells play in normal homeostasis.

1.2.1 Structure, diversity and function of T-cell receptors

Conventional, $\alpha\beta$ TCRs that I will discuss in this thesis are a heterodimer of α and β chains, each of which contain a variable (V) and constant (C) domain on the outside of a T-cell. They are connected to the T-cell's cell membrane via connecting peptides and transmembrane domains, and have short cytoplasmic tails[39,40]. The two TCR chains are linked by disulfide bridges, and associate with CD3 subunits in 1:1:1:1 stoichiometry of TCR$\alpha\beta$: CD3$\gamma\epsilon$: CD3$\delta\epsilon$: CD3$\zeta\zeta$[41].

The diversity of TCRs stems hierarchically from multiple sources. The first layer of combinatorial diversity occurs through V(D)J recombination, where variable (V), diversity (D, in beta chains) and joining (J) gene segments undergo somatic recombination. This means that functional alpha chains can be formed when any of the 45 functional Vα rearrange with any of the 50 functional Jα, leading to a feasible diversity of 2250 alpha chain gene segment combinations. For the beta chain, it is formed of 50 Vβ, joining with one of two Dβ and one of 12 Jβ resulting in 1200 beta chain gene segment combinations[42].

On top of this, nucleotide addition or deletion at the junctions between V–D and D–J in beta chains and V–J in alpha chains creates diversity in these junctional regions within the CDR3 that vastly exceeds that of the gene segments[43]. This leads to entropy measures compatible with a theoretical repertoire size of 10^7 for the alpha chain, 10^9 for the beta chain and a paired repertoire size of 10^{16} possible unique receptors[44].

As a result of the diversity and the selection processes that transform the naive T-cell repertoire into one that can be sampled and sequenced, shared paired TCRs between individuals (public TCRs) are rare. Therefore, individuals largely possess private repertoires[45–47].

T-cell receptor peptide-MHC interaction

The canonical binding arrangement of a TCR with pMHC is that the β chain primarily binds the MHC α_1 helix, whereas the α chain binds either the α_2 on MHC-I or β_1 on MHC-II[48]. The three complementarity determining regions (CDRs) of each TCR chain are primarily responsible for ligand binding, with the first two loops (CDR1, CDR2) being determined by the V-gene segment. As a result, these loops are called the "germline" loops, whereas the CDR3 is hypervariable, owing to the junction(s) where nucleotide addition/-subtraction takes place. The CDR1 and CDR2 loops are "mainly focused on the recognition of the MHC molecule itself, while the hypervariable CDR3 loops are the main drivers of peptide recognition"[40].

In order to achieve coverage of the possible set of peptides the immune system might encounter, T-cells must be cross-reactive. A meta-analysis over various cell density studies in tissues of the human body estimates that a 73 kg male human contains approximately

5×10^{11} T-cells[49]. In the thought experiment proposed by Sewell in 2012, if we assume that class I MHC molecules present 1% of amino acid 10-mers and class II MHC presents 3% of 14-mers, the theoretical diversity of peptides a human could encounter is in the range of 10^{17}[21]. In order to have coverage of the wide range of feasible epitopes, each T-cell receptor would be required to bind 200,000 unique antigens. More than that in fact, if we take into account that observed richness of human naive T-cells is of the order of 10^8 sequences[50], leading to the estimate, supported by empirical observation, that an individual T-cell receptor can bind 1 million unique peptide sequences[51].

T-cell receptor cross-reactivity has a variety of implications in health and disease. One example is molecular mimicry, whereby the sensitisation of weakly self-reactive T-cells by danger signals and activation in infection generates potent self-reactive T-cells. Since memory T-cells are substantially more antigen-sensitive than naive T-cells[52], a "gap" exists in central tolerance whereby self-reactive T-cells can bypass negative selection while naive, despite exerting meaningful autoimmune effects once they form memory phenotypes[53]. Examples of T-cells specific to a pathogen epitope that could mediate disease when directed at a self-epitope have been identified in a variety of diseases, and in great detail in Type 1 Diabetes[54] and Multisystem Inflammatory Syndrome in Children[55].

In addition to broadening the targets of T-cell recognition, cross-reactivity delivers other benefits. T-cell immunity is considered robust to the evolution of pathogens due to this "local" cross-reactivity, whereby mutations in epitopes do not necessarily diminish TCR recognition[56]. This cross-reactivity has been implicated in a case of successful tumour-infiltrating lymphocyte therapy, whereby the same TCR can target various tumours expressing different antigens, all with a shared motif[57].

On top of the cross-reactivity of a single alpha-beta TCR, many cells express multiple unique alpha-beta TCR pairs due to expressing multiple individual chains[58]. Since specificity is jointly determined by the alpha and beta chains[47,59,60], each of these pairs expands the binding repertoire of that cell. As a result, the sequencing of T-cell receptors from antigen-specific cells alone is not sufficient to prove binding, as other or undetected T-cell receptor chains might mediate that binding.

A brief overview of antigen-processing and presentation via MHC-I and MHC-II

In order for TCRs to recognise antigens presented by other cells, these cells must process and present the antigen on their cell surfaces. These processes determine which parts of viruses are "visible" to the adaptive immune system, and contribute to an epitope's immunogenicity and therefore immunodominance[19,61].

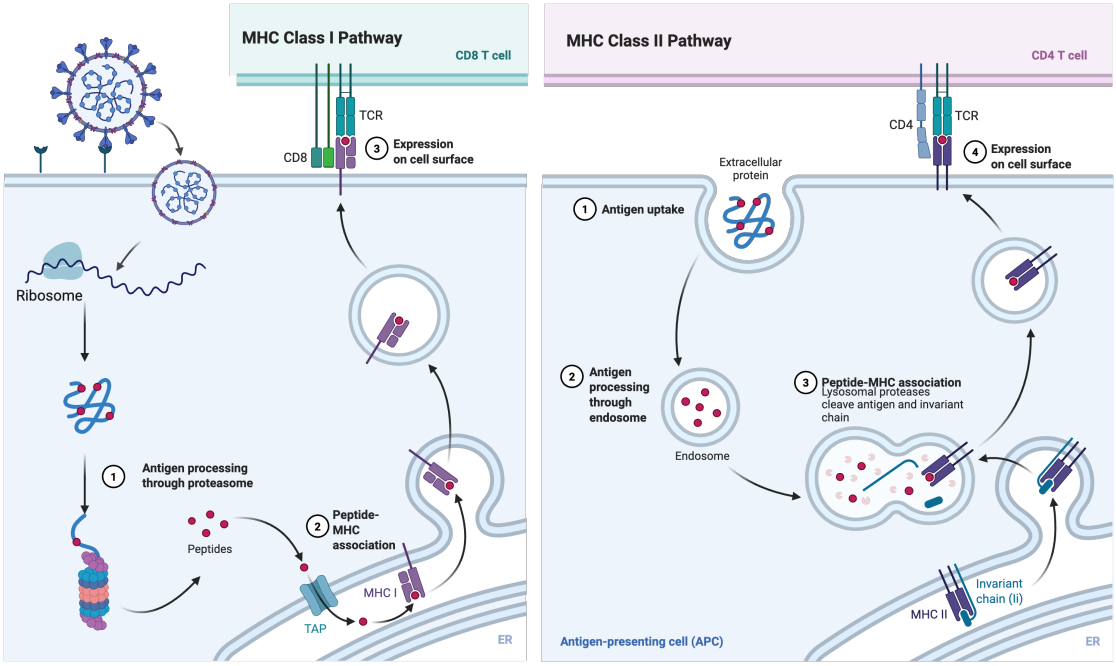


Figure 1.2: SARS-CoV2 infection and MHC class I and II presentation. Adapted from BioRender templates "MHC Class I and II Pathways" by Akiko Iwasaki and "Darunavir: Potential Repurposed Drug Candidate for COVID-19".

The Human Leukocyte Antigen (HLA) complex, the human ortholog of MHC is the "most genetically diverse locus in the human genome"[4]. The frequency of polymorphism within these genes reflects immunological adaptation to pathogen pressures, with HLA associated with a wealth of "immunological disorders [-] infectious and autoimmune"[4]. These associations could be due to their role in shaping the antigens presented to T-cells, or how these alleles influence "HLA expression ... and stability"[4].

MHC molecules differ in their structure and the mechanisms that "feed" them antigens to present. MHC class I (MHC-I) molecules are particularly important in the immune response to viruses, as they survey the intracellular environment that viruses hijack for

their replication. During protein turnover, which plays important roles in cellular maintenance as well as adaptive immunity[62], cytosolic proteins are tagged and broken down into constituent peptides. Some of these are then delivered through membranes of the endoplasmic reticulum into its lumen, where they are modified if needed and loaded onto MHC molecules. The vesicles with loaded MHC-I molecules are then trafficked to and fuse with the cell's membrane for external presentation[19].

MHC class II (MHC-II) predominantly survey the extracellular space and endosomal compartments of professional antigen-presenting cells such as macrophages, B-cells and importantly dendritic cells[2]. Through endocytosis or phagocytosis, these cells can sample their extracellular space for debris or viral particles which are degraded into 10-15 amino acid long peptides in lysosomes. The resulting peptides displace placeholder molecules on class II MHC in endosomes and are trafficked to the surface for presentation[63].

The structures of the two MHC classes are topologically similar - the groove in which peptides sit is composed of two domains which come from the heavy chain in MHC I and from both the α and β chains in the case of MHC II. MHC-II presents longer 10-15 residue peptides owing to its more open binding pocket, and being open at the ends of the peptide presenting groove. MHC-I on the other hand allows presentation of 8-10 amino acid peptides[64].

Central Tolerance and its influence on the T-cell repertoire

Thymic selection leaves a distinctive imprint on the TCR sequences. First, it drastically constrains the diversity of T-cell repertoires, with estimates of 10^8 unique nucleotide sequences for naive T-cells, 10^6 for $CD4^+$ memory cells and 10^5 for $CD8^+$ memory cells[50], orders of magnitude smaller than theoretical diversities of $10^{15} - 10^{16}$[43,65].

Second, it profoundly biases what T-cell receptors are present, with features of the repertoire strongly influenced by what MHC alleles are positively selected for. Expression quantitative trait locus mapping reveals associations between MHC polymorphisms and TCR V-gene usage[66], showing that the MHC alleles present in an individual profoundly shape what T-cell receptor genes carry through central tolerance and subsequent clonal selection.

Additionally, in paired chain analyses TCR sharing between individuals strongly correlates with the number of HLA alleles that they overlap, peaking in identical twins[46,47].

There is a question as to whether the MHC-reactivity of T-cell receptors is due to thymic selection (the selection theory) or whether TCRs in their germline V-genes encode MHC-reactive "interaction codons" (germline theory)[67]. Current evidence supports a contribution from both, because T-cell receptors tend to dock on peptide-MHC molecules in a conserved manner[68] and within an influenza epitope it appears that conserved manner is required for signalling[69]. Indeed, 5-20% of the unselected thymic repertoire of mice are MHC reactive[70], which is more than would be expected if rearrangement led to totally random binding to any peptides and MHC restriction was entirely imposed by selection.

However, the repertoire is further "optimised" by selection. A systematic comparison of pre-selection and post-selection mouse TCR repertoires has identified some constraints imposed on TCR sequences. One is a bias towards shorter CDR3 loops (8-13 amino acids) and restrictions on positively charged or hydrophobic amino acids in the CDR3β[71].

The initial naive T-cell repertoire provides the anticipatory cells for T-cell responses, but beyond just being a blank canvas, likely "imprints" the immunodominance of particular responses. The proportion of antigen-specific cells within the naive repertoire is very small[72]. Experiments using direct enumeration using peptide-MHC multimers estimate that in humans, the frequency of antigen-specific T-cells within the naive repertoire[73] is $1 \times 10^{-6} - 6 \times 10^{-5}$[74,75] for CD8$^+$ and $2 \times 10^{-7} - 1 \times 10^{-5}$ for CD4$^+$ T-cells specific to *B. anthracis* epitopes[76]. CD8$^+$ T-cells specific to immunodominant SARS-CoV2 epitope B7/N105 exist at even higher naive frequencies in pre-pandemic individuals of $3\text{-}7 \times 10^{-5}$. This elevated precursor frequency, as has been shown for other epitopes[73,74,76], likely relates to its immunodominance following infection. B7/N105-specific CD8$^+$ T-cells have highly diverse TCRs in gene usage and CDR3 features, suggesting that many different recognition strategies are feasible and the TCR features that generate this recognition survive thymic selection. They compare it with a subdominant epitope from the same protein, A2/S269, which has more restricted diversity and shared CDR3α motifs, as well as lower precursor frequencies[77].

1.2.2 CD4⁺ T-cells

CD4⁺ T-cells orchestrate immune responses through cytokine production and direct interactions with immune cells. Their classification comes from the expression of the CD4 co-receptor, which binds synergistically with the T-cell receptor to class II MHC[78]. Upon activation by TCR in the context of MHC class II molecules, naive CD4⁺ T-cells undergo clonal expansion and can differentiate into a variety of effector subsets. There are several well-characterised subsets including T-helper 1 (T_{h1}), T-helper 2 (T_{h2}), T-helper 17 (T_{h17}), T-follicular helper (T_{fh}) and regulatory T-cells (T_{reg})[26].

T-helper cells facilitate immunity by producing cytokines that specialise immune responses to particular pathogens. For example, T-helper 1 cells produce the cytokine IFN-γ, which stimulates paracrine responses against intracellular pathogens like viruses and some subsets of bacteria. T-helper 2 cells modulate tissue architecture, promoting tissue repair and the expulsion of helminth infections. These responses, when chronically triggered, can cause inflammatory diseases such as asthma and some allergies. Responses at mucosal barriers to extracellular bacteria and fungi are canonically mediated by T-helper 17 responses, but these cells can also contribute to autoimmunity[26].

One canonical role of CD4⁺ T-cells is to help B-cells to generate high affinity antibodies and memory B-cells. The subset responsible for this, T-follicular helper cells (T_{fh}), express high levels of CXCR5 which assist them in homing towards the B-cell follicles of secondary lymphoid organs. The interaction of T-follicular helper cells and follicular B-cells leads to the formation of germinal centres. Molecularly, this is mediated via signals through the CD40-CD40L axis and T-follicular helper secretion of IL-21 and IL-4. Through these signals, B-cells are prompted to divide and edit their B-cell receptors to achieve higher antigen affinity, which allows them to form more effective antibodies[79].

Another key subset are regulatory T-cells (T_{reg}), which maintain immune tolerance and prevent autoimmunity[80]. Tregs can promote tolerance to particular antigens by directly attenuating other cells' immune responses. Klawon et al. demonstrated in mice that T_{reg}s directly help discriminate self from non-self during infection, preventing the emergence of self-reactive T-cells. Using a model of bacterial infection where the pathogen expresses a

prostate antigen, they found that when they depleted prostate-antigen specific T_{reg}s, mice developed autoimmunity to the prostate. However, in the presence of antigen-specific Treg cells, self-reactive conventional T-cells were constrained[81]. T_{reg}s mediate the suppression of immune responses through cytokine release[82] and manipulation of the paracrine signalling environment[83].

$CD8^+$ T-cells were thought to be the sole mediator of cytotoxic responses that could kill cells presenting particular antigens. However, a subset of $CD4^+$ T-cells expressing cytotoxic genes and with cell-killing functions are increasingly identified in cell atlases[84,85] as well as explored in functional experiments[86]. Initially thought to be an artefact of long-term culture, these cells are consistently identified across groups and experimental methods[87]. Cytotoxic $CD4^+$ T-cells frequently express Granzyme B[87] and Perforin[84], and can directly eliminate cells[88].

1.2.3 $CD8^+$ T-cells

Referred to as Cytotoxic T-cells, $CD8^+$ T-cells are classified by their expression of CD8, a co-receptor that facilitates the TCR pMHC interaction by binding class I pMHC at a separate site[89]. Typically, $CD8^+$ T-cells exert their function by directly killing target cells, as well as releasing cytokines IFN-γ and TNF-α[2]. However, with improvements in the scale and resolution by which we can survey immune responses a diverse taxonomy of $CD8^+$ expressing cells have been discovered, including ones that release T_{h2} cytokines[90] or act to regulate immune responses in a manner similar to T_{reg}s[91].

Cytotoxic $CD8^+$ T-cells exert their effects through paracrine and cell surface mediators. They combine the release of perforins and granzymes, which open holes in target cell membranes and trigger programmed cell death[92]. Additionally, they express Fas ligand on their surface which triggers apoptosis in target cells on contact with Fas on those cells[93]. Moreover, by releasing cytokines they enhance antigen presentation, recruit immune cells and establish antiviral states in nearby cells[2,94,95].

$CD4^+$ and $CD8^+$ T-cells interact throughout immune responses. For example, $CD4^+$ T-cells "license" dendritic cells through CD40-CD40L interactions, which increases dendritic cell

antigen presentation and the expression of co-stimulatory molecules that direct and activate $CD8^+$ T-cells[96]. The importance of this in viral infection is unclear, however, because in the context of some infections innate immune stimulation of dendritic cells through their pattern-recognition receptors is able to achieve a similar phenotype[97,98]. Gressier et al. show that cytokines from $CD4^+$ T-cells change the antigen-presenting cell response to CD40 stimulation, creating a sequential activation of inflammatory and stimulatory molecules. In contrast to the lack of dependence on $CD4^+$ observed in mouse studies of other pathogens, signatures they observed of $CD4^+$ "help" of $CD8^+$ T-cells were correlated with mild SARS-CoV2 and activated or effector $CD8^+$ T-cells in their own single-cell cohort as well as published cohorts and cytometry by time of flight data[17].

1.2.4 Control of T-cells by Innate Immunity

Innate Immune Sensing and Recognition

The innate immune system is induced and activated by pattern recognition receptors that detect pathogens and damage[99]. These receptors target highly-conserved components of microbial species or molecules indicative of host-tissue damage or cell death. One pathogen-associated molecular pattern is Endotoxin, a component of Gram-negative bacteria's external cell membrane which activates the receptor TLR4 (Toll-like receptor 4)[100]. An example of a damage-associated molecular pattern is F-actin, which is normally a component of the cellular cytoskeleton. Depending on the mechanism, dying cells can lose membrane integrity and spill their cytoplasm and its content into the extracellular milieu. In this case, F-actin is present in the extracellular space and can be detected by DNGR-1 on dendritic cells[101].

Pattern recognition receptors, when activated, initiate a signalling cascade that meaningfully overlaps between pattern-recognition receptors but also encodes key features of its stimulation[11]. As illustration, viral RNA in the cytoplasm activates the MDA5 receptor, which leads to release of type I and type III interferons (IFNs)[102] triggering diverse antiviral functions[9]. Several of the components of this process are subject to immune evasion strategies, especially by SARS-CoV2, which are covered in more detail in Section 1.3.1.

Pathogen and danger are not the only patterns that the innate immune system can sense. Natural killer cells (NK cells) can provide cytotoxic responses to "missing-self"[103]. These cells are subject to both activatory and inhibitory signals from constitutively expressed cell surface molecules on other cells. In their resting state, the inhibitory receptors such as NKG2A are ligated by antigen-presentation molecules like HLA-E[104]. However, if these cells reduce their expression of these antigen-presenting molecules, as occurs in immune evasion by viral infections and cancer, the balance is shifted towards activation and NK cells can lyse the cells "missing-self"[105].

Bridging Innate to Adaptive Responses

One of the key roles of the innate immune system is to facilitate adaptive immunity, "bridging" between immediate innate responses and adaptive responses with specificity and memory[106]. Professional antigen-presenting cells, particularly conventional dendritic cells (from here referred to as dendritic cells), are a key interface between innate and adaptive immune responses[107]. These cells, and their various functional subtypes, reside in tissues throughout the body at "sites of pathogen entry" in one of two states - resting or activated[108]. When resting, subsets of dendritic cells sample their local environment by projecting cellular extensions[109] capable of reaching from the local tissues into the lumen of the airway[110], the small intestine[111] and peritubular capillaries in the kidney[112]. Stimulation via innate immune receptors makes dendritic cells capable of interacting with other cells, providing information about their activation to cells downstream in the immune response[11].

> **Note**
>
> **Dendritic Cell Response to Activation**
> Increased surface expression of MHC molecules[113,114].
> Increased expression of "maturation-marker" co-stimulatory molecules such as CD80, CD86[115] and CD40[114].
> Upregulation of CCR7, a chemokine receptor that promotes migration to lymphoid tissues[116].
> Remodelling of the cytoskeleton to enhance antigen capture[117].
> Release of "Level-1" cytokines in Iwasaki and Medzhitov's model of innate control of adaptive immunity[11]

Upon activation (historically "maturation"), Dendritic cells undergo a profound transformation into cells capable of priming naive T-cells. The key functional changes associated with bridging to adaptive immunity are summarised in Dendritic cell response to Activation.

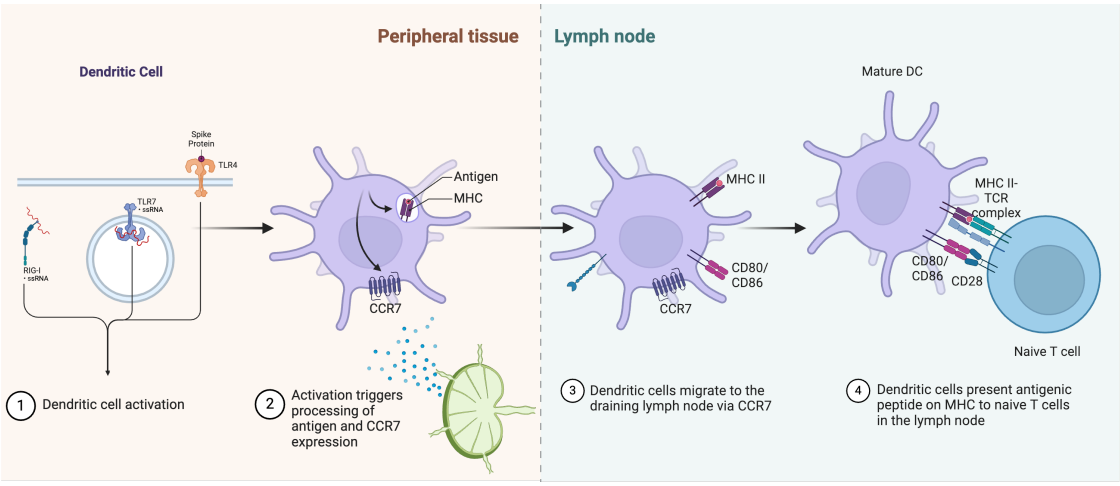


Figure 1.3: Key stages in dendritic cell activation and their priming of naive T-cells. Adapted from Biorender Template "How Dendritic Cells Become Competent to Stimulate T Cells" by Akiko Iwasaki

Dendritic cells propagate signals related to their cell state and the receptors that activated them to T-cells as they activate them. In Iwasaki and Medzhitov's[11] model of innate control of adaptive immunity, these are "Level 1" cytokines which act on signal propagating cells capable of releasing "Level 2" cytokines or directly effecting an immune response. For illustration, TLR3 ligation by viral double-stranded RNA in a subset of dendritic cells triggers the production of IL-12 which, in turn, biases naive T-cells towards Th1 differentiation and causes the cell to efficiently cross-present antigen to cytotoxic T-cells[118]. In contrast, specialised dendritic cells in the skin express Dectin-1, a receptor for fungal β-glucans. Stimulation triggers the release of IL-6, which primes the differentiation of the primary anti-fungal T-cell subset, Th17[119]. The location and migration of dendritic cell subsets to specialised niches in the lymph node is also important to the downstream immune response. In mice, a particular subset that expresses chemokine receptors directing them to the T-B cell border in lymph nodes is required for T-follicular helper cell differentiation and subsequent antibody response to viral challenge[120].

1.2.5 T-cell Activation and Priming

TCR dependent T-cell activation depends on three signals: TCR engagement, co-stimulatory receptors and the cytokine environment. The combination of factors reduces the probability of inappropriate activation, whereby multiple signals of danger are integrated to minimise the risk of autoimmunity[2]. The canonical second signal - CD28 on T-cells interacting with B7 on target cells distinguishes productive activation from anergy. Other co-stimulatory molecules fine-tune the T-cell response to priming, with ICOS, OX40 and 4-1BB generating different (but overlapping) downstream effects. Indeed, ligation of the T-cell receptor by its cognate peptide-MHC without the other signals leads to anergy, a state of minimal functionality[121].

Secondary lymphoid organs are the site of activation of naive and subsets of memory T-cells. These specialised structures enable contact between rare antigen-specific T-cells and the dendritic cells carrying their cognate antigens. High-endothelial venules, the blood vessels by which T-cells enter secondary lymphoid organs, are specialised for this function - expressing chemokines that attract naive and memory T-cells expressing CCR7[122].

While T-cell scanning of antigen-presenting cells occurs in homeostatic states, when lymph nodes become “reactive” they undergo several molecular and tissue level changes to encourage this. One example is that they induce the expression of CCR5 on naive $CD8^+$ T-cells, promoting their migration towards active $CD4^+$ T-cell and dendritic cell interactions[123]. Concomitantly, inflammation triggers widening of the blood vessels supplying lymph nodes, increasing the rate at which T-cells enter the secondary lymphoid organ[124].

In the peripheral tissues, tissue resident memory T-cells (T_{rm}) can be activated directly by antigen-presenting cells[125]. Tissue-resident T-cells are embedded within tissues and provide surveillance and an adaptive immune response that can be recalled more quickly[126]. Once re-activated, they can immediately exert effector functions without needing to migrate to or from lymphoid organs, allowing them to immediately prime tissue states and make organs more resilient to infection[127].

Priming of naive T-cells is facilitated by secondary lymphoid organs. Upon arrival in secondary lymphoid organs, dendritic cells migrate towards T-cell zones where they display processed antigen to naive T-cells[128]. While the initial signal is sufficient for T-cell proliferation and differentiation in vitro[129], T-cells can remain in contact with dendritic cells for days, integrating the signal strength, local cytokine milieu and modulatory surface molecules which influence the subsequent response[130].

Activation can induce a variety of pathways depending on the nature of the activation and the environment. Lineage commitment of activated T-cells is influenced by the expression of cytokines at the time of activation, with IL-12 promoting T_{h1} differentiation, IL-4 promoting T_{h2} and TGF-β promoting T_{h17} or T_{reg} differentiation[131] depending on the presence of IL-6. Metabolic signals are also involved, and nutrient availability can profoundly shape the subsequent state of T-cells[132]. Activation triggers transcriptional and epigenetic changes that promote proliferation, differentiation and migration[133].

When activated, many offspring form short-lived effector cells which exert important cytotoxic functions but contract rapidly. Short-lived effector cells can increase antigen-specific T-cell numbers by orders of magnitude within days, but have lifetimes constrained by intrinsic pathways and a strong dependence on cytokines[134].

Migration and Trafficking

The majority of activated T-cells leave secondary lymphoid organs with the aim of reaching their cognate antigen at the site of danger. Once they encounter their antigen and receive sufficient environmental signals T-cells are released from retention signals and exit the secondary lymphoid organ[135]. In the circulation, T-cells encounter the endothelium of various tissues, allowing them to sample the surface for adhesion molecules and inflammatory molecules that facilitate migration into the tissue[124].

One exception to this migration are T-follicular helper cells, which facilitate B-cell function and help form germinal centres. A diverse set of signals can promote T-follicular helper differentiation, with individual experiments identifying a role for various cytokines and TCR signalling strengths[136]. These signals converge at the lineage defining transcription factor

BCL6, which promotes the expression of CXCR5 and drives the developing T_{fh} towards the T-B cell border. Here, they interact with B-cells presenting their cognate antigen via their T-cell receptor, as well as other co-stimulatory and paracrine molecules.

The homing of T-cells towards sites of inflammation is thought to begin at priming, where dendritic cells carry markers of their tissue origin which influence the subsequent gene expression of the T-cells they trigger. For example, in mice, when T-cells are primed ex-vivo with dendritic cells from the lung, they migrate into bronchoalveolar lavage fluid and lung tissue more efficiently than T-cells activated by dendritic cells from other sites. Flow cytometry revealed that the lung-primed T-cells expressed CCR4 at higher levels than gut activated T-cells, with the lung-homing conferring an advantage upon influenza challenge[137].

The combination of chemokine release and chemokine-receptor expression creates a complex network that influences T-cell trafficking. On their surface, circulating activated T-cells express integrins and modulate their expression of selectins that facilitate their entry into tissues[138]. This allows the sequestration of antigen-specific T-cells at the site of danger and inflammation, which in concert with the T-cell's expansion and functional differentiation generates a potent infiltrating immune response 7-10 days after infection[84].

1.2.6 T-Cell Memory Subsets

Memory is a core attribute of adaptive immunity, and this is mediated by the small proportion of antigen-specific T-cells that participated in primary responses which differentiate and form memory populations. These memory T-cells maintain their antigen-specificity, but are specialised for continuous surveillance and rapid responses to re-exposure[139]. Subsets of memory T-cells are specialised for this role - lower activation thresholds, faster responses to antigenic stimulation, and banks of effector molecules for rapid responses[26,140].

Over an individual's lifetime, memory T-cells come to dominate their T-cell compartment as naive T-cells decline[125]. This reflects both the decrease in thymic output with age, whereby the rate at which naive T-cells are "added" to the repertoire slows down[141], as well as more exposures over a longer lifespan.

Since foundational flow-cytometry studies in the late 90s[142,143], both $CD8^+$ and $CD4^+$ T-cells are classified into 4 subsets on the basis of their expression of CCR7 and CD45RA. Both are initially expressed on naive T-cells ($CCR7^+CD45RA^+$) that circulate between secondary lymphoid organs, and upon activation they differentiate into central memory T-cells (T_{cm}, $CCR7^+CD45RA^-$) or effector memory T-cells (T_{em} $CCR7^-CD45RA^-$). Some effector memory T-cells re-express CD45RA once terminally differentiated (T_{emra}, $CCR7^-CD45RA^+$), and it is unclear whether this differentiation is linear (i.e. T-cells form effector cells and then central memory) or asymmetric (naive T-cells form either effector or central memory T-cells)[140].

1.3 T-Cell Responses to SARS-CoV2

1.3.1 A brief introduction to SARS-CoV2

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2) is the causative virus of Coronavirus Disease 2019 (COVID-19)[144]. First detected in late 2019 in Wuhan, China as part of a cluster of viral pneumonia cases, it has since caused 777 million infections and over 7 million deaths[145]. The world responded to a novel, viral threat with an international research effort on SARS-CoV2 biology, pathogenesis and epidemiology, culminating in diagnostics and vaccines produced in record time[146,147].

The emergence of a novel virus with minimal pre-existing immunity[148–151] provides, for all of its tragedy, a remarkable opportunity to study key questions in immunology that would be infeasible to study with other viruses. Unlike respiratory viruses such as Influenza or latent viruses such as Epstein-Barr Virus, SARS-CoV2 infection history and memory is not as confounded by prior vaccination or exposure. This means that even without rigorous longitudinal sampling, researchers can trace the development of immune memory from first exposure, offering a previously impossible window into how memory and immunity develop against a new pathogen.

Emergence and global impact

Following SARS-CoV (2003) and MERS-CoV (2012), SARS-CoV2 is the third highly pathogenic coronavirus to emerge in humans[152,153]. Unlike its predecessors, SARS-CoV2

achieved global spread within months[154] owing to its high transmissibility and capacity to transmit without reported symptoms[155,156].

Within months of its emergence, health systems worldwide struggled to scale care to the sudden increase in demand[157], leading to cancelled elective and outpatient procedures[158,159] still reflected in waiting lists worldwide[160].

There is massive variation in SARS-CoV2 disease outcomes, influenced by age, sex, comorbidities and genetics[161–165]. Indeed, the key results of a meta-analysis of genome-wide association studies on SARS-CoV2 severity are hits in inflammatory signalling genes, monocyte-macrophage activation, endothelial permeability genes, immune metabolism and host viral entry and replication factors[166], pointing to the diverse physiological functions that contribute to control of the virus.

The virology of SARS-CoV2

The life cycle of SARS-CoV2 begins with the Spike protein binding to the host cell's angiotensin-converting enzyme 2 receptor (ACE2). Following receptor binding, host proteases such as TMPRSS2 or cathepsins cut cleavage sites on the S protein, which triggers the cellular and viral membranes to fuse. As a result, the viral genome enters the cytoplasm and is translated to produce polyproteins pp1a and pp1b, which are cleaved by viral proteases into individual non-structural proteins. These proteins form SARS-CoV2 replication and transcription complexes on vesicles derived from the host-cell endoplasmic reticulum[153]. Newly synthesised viral genomes associate with nucleocapsid proteins before incorporating structural proteins and then being released via exocytosis[167].

SARS-CoV2 is a member of the Coronaviridae family of single-stranded, enveloped, positive sense RNA viruses. It shares 79% sequence similarity with SARS-CoV and 50% with MERS-CoV[153]. The genome comprises open reading frames that encode non-structural proteins, responsible for viral replication, transcription and virus-host interactions, and structural proteins (visualised in Figure 1.4).

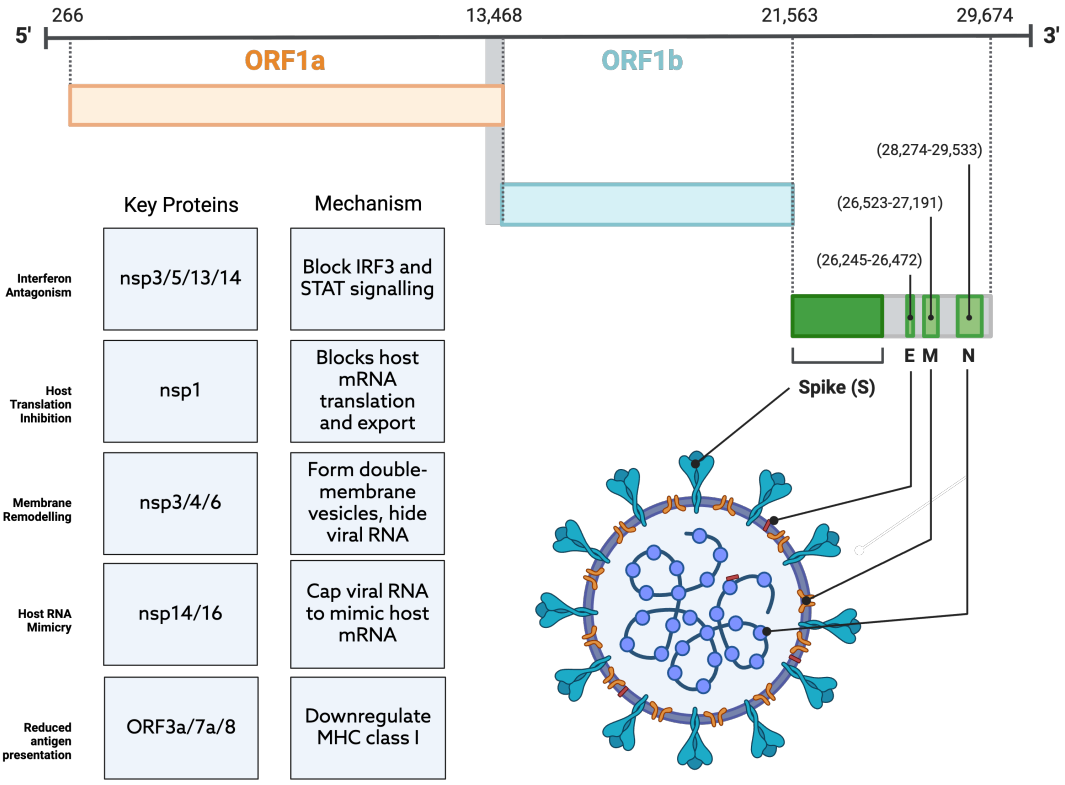


Figure 1.4: SARS-CoV2 genome organisation, structural proteins and immune evasion mechanisms. Image created using BioRender.com combining templates 'SARS-CoV-2 Genomic Organization' by Goldman-Israelow, B., Ona, S., and 'Coronavirus Proteins' Gu, J and information from Steiner et al. 2024 and Sette et al. 2021.

SARS-CoV2 mechanisms of immune evasion

The type I and III interferon responses are crucial to establishing an antiviral state in infected and neighbouring cells[168], and SARS-CoV2 has evolved multiple mechanisms to subvert this. Several viral proteins disrupt type I and III interferon responses, for example, nsp3 and nsp5 can function as proteases that cleave key factors that induce the interferon response. Concurrently, nsp6 and nsp13 can block the phosphorylation of key signal propagators in the JAK-STAT pathway, inhibiting the response to innate immune activating cytokines[153].

SARS-CoV2 forms specialised replication compartments by reorganising host cell membranes to evade various pattern recognition receptors in the cytoplasm[169]. Viral RNA is the target of conserved antiviral sensors in the cytoplasm such as RIG-I and Toll-like receptors. By forming viral replication complexes, which consist of a double-membrane vesicle within

the cytoplasm, the viral RNA is shielded from these host defences with the added benefit of concentrating key viral replication substrates and enzymes.

Even while in the cytoplasm, SARS-CoV2 viral RNA is adapted to avoid recognition by pattern-recognition receptors. Eukaryotic mRNAs have cap molecules at their 5' ends that facilitate splicing and transcription, while when uncapped RNA molecules are recognised they prompt innate immune responses and degradation[170]. Nonstructural proteins attach (nsp12) and then methylate (nsp14, nsp16) the RNA cap to mimic Eukaryotic mRNA[153].

While many of SARS-CoV2's immune evasion mechanisms disrupt the innate immune system, and thereby prevent its efficient priming of the adaptive immune system, some SARS-CoV2 proteins directly influence adaptive immune mechanisms. For example, ORF3a, ORF7a and ORF8 reduce cell surface expression of class I MHC, thereby attenuating presentation of viral peptides to $CD8^+$ T-cells capable of clearing virally infected cells. All three of these proteins target the adaptive immune system differently. ORF3a fragments the Golgi apparatus architecture, disrupting general cell surface protein expression and therefore reducing class I MHC presentation[171]. In contrast, ORF7a and ORF8 work via a more targeted mechanism directed at class I MHC. ORF7a is a molecular mimic of $\beta_2 M$, competing with it to interact with MHC-I heavy chains[171] while ORF8 binds MHC-I and marks it for degradation[172].

1.3.2 Dynamics of the T-cell response to SARS-CoV2

SARS-CoV2 likely initially infects ciliated cells in the nasopharynx through the inhalation of small airborne particles, droplets or aerosols. Ciliated cells have been identified as primary targets of SARS-CoV2 infections in autopsies[173,174], as well as human challenge trials[84], potentially due to their high expression of the major cell entry protein ACE2[175].

Mechanistic and OMIC studies have identified SARS-CoV2's capacity to delay innate type I and type III IFN responses even compared to other respiratory viruses. Hatton et al. generated a primary human cell organoid of the human nasal epithelium, and compared infection of the epithelium with SARS-CoV2 virus, Influenza A and Parainfluenza 3. By 24 hours after infection both the Influenza and Parainfluenza conditions induced expression of *IFNB* and

IFNL1 as well as several interferon stimulated genes by RT-PCR, whereas the SARS-CoV2 condition did not. Even when repeated at 20 times the original multiplicity of infection, interferon genes were either not or less induced by SARS-CoV2 than Influenza within 24 hours[176]. Blanco-Melo and colleagues also found low induction of interferon I/III and its stimulated genes by SARS-CoV2 compared to Influenza A and Respiratory Syncytial Virus in respiratory cell lines, but when looking at inflammatory cytokines such as *IL6*, *CCL20* and *CXCL5* SARS-CoV2 induced strong expression. This thread was shared within ferret and human serum experiments, whereby RNA reads of inflammatory cytokines in infected ferret nasal washes and blood showed high inflammatory cytokines but low type I and III interferons[177].

Human challenge trials provide us with the most precise insights into early infection dynamics as participants can be dosed equally and followed over known time points. Using an infectious dose that caused 6 of 16 participants to develop sustained SARS-CoV2, which is a higher but comparable infection rate to household transmission (25%)[156], Lindeboom et al carried out comprehensive single-cell analysis including per-cell viral RNA counts of the nasopharynx and blood[84]. Supporting a previous study that found a ciliated cell subset efficiently infected by SARS-CoV2[175], the trial found that a cluster of hyperinfected ciliated cells comprising 4% of cells had 67% of detectable viral RNA. The hyperinfected cluster was associated with lower expression of acute phase response, IFN response and MHC presentation related genes than other cilia[84].

The human challenge trial allows us to follow the natural history of infection in the nasopharynx and the blood. In individuals who never or only transiently became PCR positive, CITE-Seq cell clusters of immune cells increased in abundance in the nasopharynx within a day of challenge. Those with sustained infections saw increase in immune cell abundance by day 5 until day 10. The nature of the infiltrating cells varied between the infection outcomes - in abortive infections, the only cell clusters that grew were $CD4^+$ and $CD8^+$ T-cells, whereas in sustained and transient infections initial infiltration was mainly innate cell types such as monocytes, dendritic cells and $\gamma\delta$ T-cells[84].

Human challenge studies allow study of the immune response in individuals who are infected with SARS-CoV2 but never test positive on PCR. Using DESeq2 to compare pseudobulked RNA counts at the pre-challenge timepoint between abortive and sustained infections, the authors identify *HLA-DQA2* as the second most differentially expressed gene behind ribosomal protein *RPS26*[84]. The *HLA-DQA2* transcript has previously been identified as associated with milder COVID-19[178], but its function has not been characterised and it does not appear to present antigen like other class II molecules.

Interferon type I and III responses trigger antiviral states in neighbouring cells and enhance antigen-presentation to T-cells, and a dominant model of SARS-CoV2 pathogenesis suggests that by delaying this response, SARS-CoV2 is able to get a "head start" on replication without priming adaptive T-cell responses[10]. The key argument of Sette's model is that by subverting the early antiviral interferon responses, the virus is able to replicate unchecked and delay priming and activation of T-cells. The growth in viral counts leads to excessive innate immune activation and pathology, with the T-cell response unable to "catch up" to the exponential. The core lines of evidence are observational associations of T-cell responses with disease outcomes, severe COVID in individuals with diminished Interferon responses and evidence of innate lung pathology in COVID. Some of the studies he uses to support this model, their methods and key findings relevant to this are summarised in Table 1.1.

In line with this, interferon response signatures were apparent in nasopharyngeal ciliated cells in transient infections but not in sustained infections by day 1. In sustained infections interferon-induced gene expression modules were observed in blood samples by day 3, but not in the nasopharynx until day 5. This finding was validated in a separate, bulk-RNA sequenced human challenge cohort. Nasopharyngeal swab sampling does not collect nasal associated lymphoid tissue, so they could not study whether this reflects dynamics of migratory innate immune cells. Some level of innate immune recruitment likely occurs, as all infection outcomes see a decrease in circulating inflammatory monocytes by day 3[84].

By tracking clinical variables like symptoms, C-reactive protein concentrations and temperature, as well molecular characteristics longitudinally, Lindeboom et al. could identify the changes that correspond with symptoms of SARS-CoV2 infection. In sustained infections

the first reported symptoms mostly appeared by day 4, after the systemic interferon response signature[84]. Loss of smell coincided with the peak in infected cells at day 7, while resolution of symptoms coincided with "strong reduction in the number of infected cells at day 10 and a peak in the amount of nasopharyngeal immune infiltration".

Table 1.1: Table of key studies supporting Sette model of COVID-19 pathogenesis

Design	Category	Methods	Key Finding
Observational study of Antigen-Specific SARS-CoV2 T-cell responses in 50 subjects, 24 of which were acute SARS-CoV2[179]	T-cell association with milder disease	Activation induced marker assays and intracellular cytokine staining with SARS-CoV2 peptide pools. Analysed correlations between immune parameters, age, comorbidities and disease severity.	Responsive $CD4^+$ T-cells were significantly associated with less severe disease (AIM: $p = 0.0016$, IFN-γ ICS: $p = 0.0006$). Responsive $CD8^+$ T-cells also associated with less severe disease (AIM $p = 0.024$, IFN-γ ICS: $p = 0.005$)
Longitudinal observational study of 12 COVID-19 patients from disease onset to recovery/death.[180]	T-cell association with milder disease	Quantified T-cell responses with IFN-γ ELISPOT with overlapping peptide libraries of SARS-CoV2 proteins	Higher frequencies of IFN-γ secreting T-cells at <15 days and >15 days in mild patients. Earlier SARS-CoV2 T-cell ELISPOT responses were associated ($p=0.0011$) with shorter duration of infection.
Observational study of Bronchoalveolar Lavage samples from 4 healthy controls, 3 moderate and 6 severe COVID.[181]	T-cell association with milder disease, Innate Lung pathology in COVID	Single-cell RNA sequencing on Broncho-Alveolar Lavage Fluid cells.	Moderate COVID19 cases had a higher proportion of $CD8^+$ T-cells with Tissue resident phenotypes and more clonally expanded. Severe cases showed higher levels of macrophages expressing IL-1β, IL6 and TNF-α.

Design	Category	Methods	Key Finding
Longitudinal study of 32 individuals hospitalised with COVID, compared to 16 similar influenza patients. Also involved individuals with milder flu and healthy controls.[182]	Diminished interferon responses	Bulk RNA-seq of White Blood Cells in 9 COVID patients and healthy controls. Luminex and ELISA assays to quantify cytokines and chemokines.	COVID-19 patients had impaired induction of IFN-λ and IFN-α, with median levels at the lower limit of detection. Flu patients had higher expression of both IFNs ($p < 0.01$). Higher IFN-λ blood concentrations at ICU entry were associated with lower respiratory tract viral load ($p=0.0461$) and faster viral clearance ($p=0.0393$).
Study of 50 COVID-19 patients (17 severe, 18 critical) and 18 healthy controls, analysed around 10 days after symptom onset[183]	Diminished interferon responses	Mass cytometry of PBMCs. Digital ELISA to quantify cytokines. Droplet PCR to measure plasma viral load.	Critical COVID-19 patients had lower expression of Interferon stimulated genes compared to mild/moderate ($p<0.05$). Low IFN-α plasma levels were associated with a 12-fold (OR 12; 95% CI 1.21-118) increased risk of progression to critical status.
Longitudinal model study using rhesus macaques infected with high-dose SARS-CoV2[184]	Innate Lung Pathology in COVID	Histopathology and immunohisto-chemistry, Proteomics and Viral load quantification	SARS-CoV2 infection caused significant macrophage infiltration in lungs with endothelial damage and thrombosis. Strong positive correlation between viral loads and expression of inflammatory cytokines and complement genes ($p < 0.007$)

Design	Category	Methods	Key Finding
Observational study of 42 SARS-CoV2 infected individuals (28 severe) and 12 healthy controls.[13]	Innate Lung Pathology in COVID	Flow cytometry on whole blood.	Elevated neutrophil-to-T-cell ratio in severe COVID-19 (median 15) compared to other groups (<2.6). Neutrophil-to-T-cell ratio strongly correlated with APACHE III clinical score (higher means worse outcomes, p =0.022).
Postmortem analysis of lung specimens from 4 patients who died from COVID-19 and 4 control patients who died from non-COVID-19 causes[185].	Innate Lung Pathology in COVID	FFPE lung sections were analysed by immunofluorescence for neutrophil extracellular traps	NETs were present in lungs of all four COVID-19 patients but not in control patients ($p < 0.05$). In the vascular compartment, NETs were specifically found in arteriolar microthrombi containing "NET-prone primed neutrophils"

Studies published since continue to support this model, but require nuance with regards to T-cell subsets and the timing of Interferon measurements. For example, in the COVID-19 Multi-Omic Blood ATlas (COMBAT) they carried out single-cell sequencing of PBMCs from 78 COVID patients, 25 sepsis patients, 12 Influenza patients and 10 healthy controls. In a variety of celltypes, they found enrichment of interferon type I and II pathways in hospitalised but not severe COVID-19. However, interferon stimulated genes were enriched in many cell types when comparing COVID vs Flu and Sepsis samples, conflicting with studies suggesting ineffective priming of Interferons by SARS-CoV2[34].

In keeping with the findings of Table 1.1, Lucas et al.[186] found in 113 longitudinally sampled individuals hospitalised with moderate or severe COVID-19 that severe patients had higher levels of inflammatory cytokines IL-1α, IL-1β, IL-6 and reductions in both number and frequency of $CD4^+$ and $CD8^+$ T-cells. While they describe "similar" levels of type I (IFNα)

or type III (IFNλ) earlier than 10 days post symptom onset in moderate and severe COVID-19, they find a strong correlation in their study between nasopharyngeal viral load and blood IFNα, IFNγ and TNF.

One unifying explanation is that the type I and III interferon response is not ablated but delayed, and eventually high expression of interferon stimulated genes is achieved by the interaction of pattern-recognition receptors and high viral loads. The delayed, but eventually overwhelming IFN response could explain the lack of clonal expansion of $CD8^+$ effector and memory subsets in severe COVID in COMBAT[34] despite high interferon stimulated gene expression, as proliferation[129] reflects presentation and interferon responses at an earlier timepoint.

Supporting the delay argument, Lindeboom et al., used their temporally resolved human challenge data to train Gaussian process regression and latent variable models to predict from single-cell data the time since exposure. They then applied this to single-cell atlas[34,187–190] data across 361 COVID-19 patients and found that in more severe COVID, the difference between self-reported days since symptoms and model inferred time since exposure was less than in mild cases. In mild cases, this gap has a median of approximately 3 days (read from Extended Figure 6.o.) - close to the observed onset of symptoms. In more severe cases, the predicted time since exposure is much closer to the onset of symptoms (approximately 0 days)[84]. This suggests that the cellular immune response in more severe cases is up to 3 days "behind", providing the virus exponential growing time.

This theory also relates to age as a risk factor for severe SARS-CoV2. Yoshida et al. observed in a cross-sectional study of 93 individuals that healthy children at baseline had high expression of IFN-α, and it only increased slightly with infection. In contrast, healthy adults had the lowest IFN-α, and this increased significantly from baseline with infection, but was still lower than in infected children. When examining this by cell-types they found that IFN response signatures were robustly expressed by epithelial and immune types (including a $CD8^+$ memory T-cell subset) in healthy children, as well as being more strongly upregulated in COVID-19 than in adults. They also observed higher IFN-α expression in milder and asymptomatic disease than moderate and severe disease. Baseline activation of Interferon

type I and more potent upregulation of IFN response signatures in response to COVID-19 provide a mechanism for decreased COVID-19 risk amongst children[191], and elevating risk with age[192].

While the model indicates a protective role for T-cells, some studies support a pathological role of T-cells. A key piece of contradictory evidence is the observation that in mice, adaptive immune cells are required for SARS-CoV2 pathology[193]. In Imbiakha et al.'s experiments, $rag^{-/-}$ mice that lack B and T-cells do not develop weight loss when infected with a mouse-adapted SARS-CoV2. Indeed, the knockout mice had much lower concentrations of COVID-19 associated cytokines IL-6 and TNFα. The authors identified that at 3 days post-infection, $rag^{-/-}$ mice had significantly fewer neutrophils in the lungs and did not see the increase in monocyte derived macrophages in the lung that wild-type mice did. Knockout of the TCR alpha chain generated an intermediate phenotype, which had less severe disease than wild-type but greater weight loss than the $rag^{-/-}$ condition, suggesting $\gamma\delta$ T-cells and B-cells could contribute to inflammation. These results suggest that T-cells themselves play an important role in SARS-CoV2 pathogenesis, rather than simply being permissive to it when they are missing or inefficiently primed.

Some mechanistic hypotheses for this exist, including T-cells acting as part of a positive feedback loop with macrophages in severely infected lungs or through inappropriate cytotoxicity. Grant et al carried out flow cytometry on bronchoalveolar lavage fluid within 48h of intubation on SARS-CoV2 patients, and compared them to retrospective samples from severe pneumonia and other ICU patients. They found that severe COVID-19 was associated with bronchoalveolar neutrophil cell proportions over 50%, as well as a significant enrichment for $CD4^+$ and $CD8^+$ T-cells compared to other conditions. In bulk RNA sequencing of alveolar macrophages, they found expression of monocyte and T-cell chemoattractants and in single-cell analysis carried out on 10 samples, they showed expression of IFN-γ by T-cells. Alveolar macrophages contained SARS-CoV2 RNA and expressed T-cell and monocyte chemokines such as *CCL4*, *CCL20* and *CXCL10*. Explaining these results, the authors propose that infected alveolar macrophages express chemokines that recruit T-cells, leading to their activation and IFN-γ expression. Together, this drives a positive feedback

loop whereby monocyte-derived alveolar macrophages are recruited and infected, driving more T-cell activation and propagating virus throughout the lung[194].

Georg et al. identify a degranulating cytotoxic T-cell subset that could drive COVID-19 pathology. Using mass cytometry and single-cell sequencing of PBMCs in acute and convalescent COVID-19, they identify $CD16^+$ T-cell populations within $CD4^+$, $CD8^+$ and $\gamma\delta$ T-cells. CD16 is a Fcγ receptor that binds conserved sections of antibodies to mediate antibody-dependent cellular cytotoxicity. In their single-cell data, T-cell transcriptomic clusters were enriched for genes associated with degranulation and cytotoxicity and they showed that on stimulation with anti-CD16 antibodies severe patient's T-cells had greater upregulation of CD107a (a marker of degranulation) than mild patient's T-cells. Severe patient's T-cells, when stimulated with CD16, were capable of disrupting lung microvascular endothelial cells in co-culture. In a validation cohort, "proportions of activated $CD16^+$ T-cells ... were associated with worse outcomes [in] severe COVID-19"[195].

Contradictory evidence also exists of an association between SARS-CoV2 responsive T-cells and reduced severity. Olea et al. examined Spike-1 and Membrane peptide-pool responsive T-cells in peripheral blood of 71 individuals critically ill with COVID-19 via IFN-γ intracellular cytokine staining. All but one patient had detectable SARS-CoV2 responsive T-cells, and $CD4^+$ responses were more common. The frequency of reactive T-cells was associated with SARS-CoV2 RNA load in tracheal aspirates, and no correlation was found between their frequency and the dynamics of SARS-CoV2 RNA counts. They conclude that "peripheral blood levels of SARS-CoV-2-S1/M-reactive IFN-γ $CD4^+$ and $CD8^+$ T-cells does not predict viral clearance from the lower respiratory tract or poor clinical outcomes in critically ill COVID-19 patients"[196].

With some exceptions, studies of acute COVID-19 have focused on measurements of peripheral blood[197]. This only indirectly measures immune responses at the site of respiratory infection, and can lead to uncertain conclusions. For illustration, fewer SARS-CoV2 responsive T-cells in the blood between groups could reflect differences in priming or recruitment of T-cells, or could reflect that these T-cells more readily migrate to the lungs. Only recently have

studies emerged that combine acute disease, high resolution study of the immune response and the scale to identify nuanced differences.

One of these is work by Markov et al., who took "bronchoalveolar lavage fluid samples from 273 patients with severe pneumonia, including unvaccinated patients infected with [SARS-CoV2] or with respiratory failure not linked to pneumonia". Supporting the Sette model, flow cytometry and bulk RNA-sequencing finds that $CD4^+$, $CD8^+$ and T_{reg} proportions and an IFN-γ signature in those T-cells were associated with discharge. In contrast, lung T-cells expressing inflammatory and TNF-NF-κB signatures were associated with risk of death. The rate at which SARS-CoV2 RNA declined was greater in discharged patients[198].

Zhu et al. were able to link blood and mucosal immune responses in COVID-19, with 122 bronchoalveolar lavage and 280 peripheral blood samples, including 24 participants with both. Profiling SARS-CoV2 responsive cells with peptide pool stimulation and intracellular cytokine staining, they found that IFN-γ secreting $CD4^+$ and $CD8^+$ were more numerous in the lung than the blood. Casting some doubt on the conclusions drawn from PBMC experiments, the authors found no correlation between proportions of responsive T-cells in PBMCs and the lavage fluid. They find different relationships with severity between $CD4^+$ and $CD8^+$ T-cells. Overall, higher frequencies of responsive T-cells were associated with lower viral gene copies and for responsive $CD8^+$ their quantity was inversely associated with inflammatory cytokines IL-6, IL-8 and inflammation marker CRP. However, responsive $CD4^+$ frequency was "associated with markers of severe disease, such as ... neutrophilia, lymphopenia, and higher neutrophil:lymphocyte ratios (NLRs)"[199].

Many of the samples in Zhu et al. were taken after viral clearance, and thus they could study T-cell dynamics immediately after resolution. In bronchoalveolar lavage samples, single-cell clusters of $CXCR3^+CD4^+$ cells and $ITGA1^+CD8^+$ cells increased, whereas clusters defined by proliferation genes decreased in abundance after viral RNA became undetectable. Validating this subset via flow-cytometry, they find that after resolution T_{em} and T_{rm} subsets come to dominate the lavage fluid. These subsets were less frequent in PBMCs, which the authors suggest is the emergence of a tissue resident phenotype post infection[199].

Once viral counts begin to meaningfully decay, the frequency of T-cells with activated phenotypes contracts and begins the resolution stage of the immune response. In the human challenge trial cell clusters of activated T-cells, defined by surface and transcriptome expression of CD38, ICOS, CD27 and CD28 contracted rapidly by day 10, and were almost undetectable by day 28. Indeed, at day 14 in sustained infections where viral RNA was no longer detectable, there was a peak in the frequency of activated regulatory T-cells in the nasopharynx. These cells expressed *IL10*, an anti-inflammatory cytokine that may have a role in the resolution of immune responses and formation of $CD8^+$ T-cell memory[200]. Where clonotypes from activated clusters were detected at the 28 day timepoint, they had effector or memory T-cell phenotypes[84]. This represents the transition from the activated, acute immune response to the formation of T-cell memory.

1.3.3 T-cell Targets and Immunodominance in SARS-CoV2

T-cell responses to targets throughout the SARS-CoV2 proteome develop in most individuals following infection, with differences in their distribution and functionality depending on the nature of the infection. Peng et al. used IFN-γ ELISPOT assays with a matrix of overlapping peptide pools to compare T-cell responses between individuals around Oxford who had severe COVID, mild COVID or pre-pandemic samples at least 28 days after symptoms. In their study, 38/42 individuals recovered from COVID responded to at least one SARS-CoV2 peptide pool. By deconvoluting the matrix, they could identify the specific 15-18mer regions that each individual's T-cells were responding to. "[H]igher magnitude ... and broader ... T-cell responses were observed in severe cases compared with mild cases, in particular for responses to spike ... membrane, ... ORF 3 ... and ORF8."[95] However, mild cases showed a higher proportion of $CD8^+$ of the total T-cell response for spike and combined nucleocapsid and membrane proteins. Additionally, responses were qualitatively different in their phenotype and target hierarchy: in severe COVID-19 cases the proportion of responsive $CD8^+$ that were polyfunctional (express more than 1 cytokine from IFN-γ, TNF and IL-2) was no different between spike and combined membrane/nucleocapsid. In

contrast, in the overall cohort, but especially in mild cases, there was greater polyfunctionality in their membrane/nucleocapsid responses than spike responses, suggesting that these epitopes might prime more functional immune responses.

Some regions of SARS-CoV2 are immunodominant, meaning that they elicit immune responses that are more frequent than other regions whether across populations or more dominantly within an individual. Population level responses can be visualised through the IEDB (Immune Epitope Database) Immunome Browser, which aggregates T-cell response experiments across sources aggregated in the database and allows visualisation of cohort response rates at regions of a protein. For example, Figure 1.5 shows how assays within the IEDB and their observed response frequencies vary across the SARS-CoV2 membrane protein[201]. To some degree this is self-reinforcing, regions identified as immunodominant are studied more since they are more likely to detect an epitope-specific response. These regions are also biased by their source data - Sette, Sidney and Crotty aggregated the 85 studies of SARS-CoV2 T-cell epitopes in the IEDB in June 2022 and showed that the number of epitopes reported for a given HLA-allele correlates ($CD4^+$: Spearman's 0.79, $p<0.001$; $CD8^+$: Spearman's 0.81, $p<0.001$) with its population frequency[202].

Immunodominant regions of the SARS-CoV2 proteome are identified across different T-cell methods and studies. Some examples are the $CD8^+$ epitopes Spike 269-277, Nucleocapsid 105-113, and the $CD4^+$ regions within Membrane 182-190 and Spike 166-180[95,148,203]. These studies differ in whether they survey T-cell immune memory immediately ex-vivo, or whether they stimulate and expand to achieve greater cell numbers. This has the side effect of changing the ex-vivo phenotype, as these cells become activated, as well as potentially biasing the repertoire of the sampled cells in both TCR features and antigen target. Generally, studies that use "direct ex-vivo assays tended to find larger repertoires"[202]. They also differ in how they define antigen-responsive cells (described in more detail Section 1.4.2), the specific parameters within these methods, and the cohorts that they use.

Grifoni et al. used aggregated "megapools" of epitopes predicted by the IEDB analysis resource to measure whether $CD4^+$ and $CD8^+$ T-cells were responsive to SARS-CoV2 peptides via AIM (activation induced marker assay). Using the surface markers OX40 and

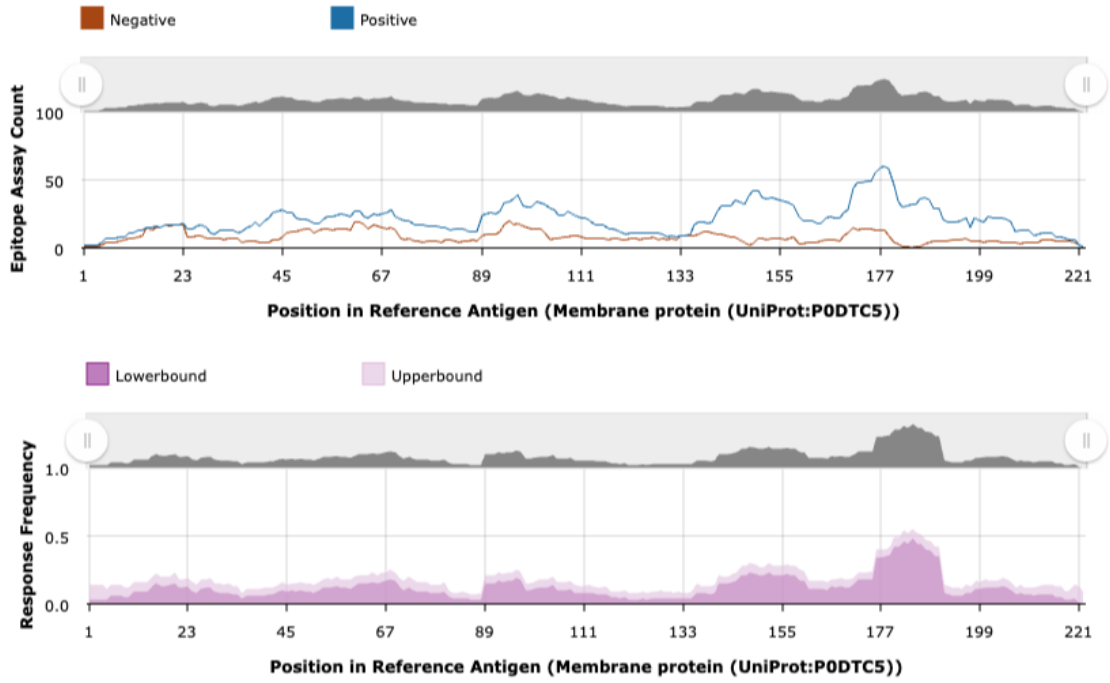


Figure 1.5: IEDB Immunome Browser (accessed 21/04/2025) of the SARS-CoV2 Membrane Protein, depicting the number of assays that have been carried out on a given region (top) and the frequency of responses observed in those assays (bottom).

CD137 in $CD4^+$ T-cells, and CD69 and CD137 in $CD8^+$ T-cells, they measured the presence and frequency of megapool-responsive T-cells between COVID-19 convalescent patients collected in a UC San Diego health clinic and pre-2019 blood samples from a commercial vendor. They found $CD4^+$ responses in all individuals, and $CD8^+$ responses in approximately 70% of individuals, but interestingly found megapool responsive $CD4^+$ T-cells in 40-60% of historic samples[148]. In contrast, the IFN-γ assays (at 200,000 PBMCs stimulated at 2 µg/ml of peptide) by Peng et al. did not show any responses to SARS-CoV2 epitopes in unexposed individuals[95]. A study that took place in Singapore and included individuals recovered from the original SARS as well as unexposed and COVID-recovered samples used 400,000 PBMCs per well and stimulated them at 5 µg/ml and found that 51% (19/37) of their unexposed group generated IFN-γ responses to SARS-CoV2 Nucleocapsid and/or NSP7/13 proteins. Their unexposed group did include 11 individuals who were SARS-CoV2 antibody negative, as well as historic samples.

Cross-reactive memory, where memory against a different, potentially related antigen,

provides memory to a different antigen may contribute to milder or abortive SARS-CoV2 infections. There are suggestions that milder and asymptomatic cases have pre-existing memory responses to epitopes shared with endemic human coronaviruses[149,151,204]. These studies, across "diverse settings and geographical locations" suggest that pre-existing SARS-CoV2 responsive memory T-cells exist and may influence the severity of subsequent infection[202]. Swadling and colleagues showed through a regularly tested healthcare worker cohort that some individuals displayed potent T-cell responses to NSP12, part of SARS-CoV2's replication and transcription complex, despite never becoming seropositive. These T-cells were detectable prior to exposure, expanded in-vivo following exposure and accumulated in greater number in individuals who did not become PCR-positive[149], suggesting they were expanding in response to antigen. Augusto et al. suggest that pre-existing cross-reactive immunity might mediate the controversial[205–207] association between the allele HLA-B*15:01 and asymptomatic SARS-CoV2 infection[151]. They used tetramer sorting to identify $CD8^+$ T-cells in pre-pandemic samples that respond to immunodominant B*15:01 epitopes, and found that five from nine samples had cells specific for one of these epitopes (NQKLIANQF). None of the cells that sorted to this epitope had a naive surface phenotype ($CCR7^+CD45RA^+$), which the authors took to indicate that these were pre-existing memory cells. Although this association was validated in two separate cohorts, no prospective work has been done to show that this allele is protective against symptomatic infection.

1.4 Investigating Antigen-Specific T-cells using Single-Cell Techniques

T-cell immunity hinges on antigen-recognition. The T-cell repertoire is able to respond to a variety of epitopes, and this specificity forms the basis of both protective immunity and some forms of autoimmunity. Therefore, understanding how antigen-recognition leads to pathogen clearance and immune memory are foundational to the study of immunology, and underlie key innovations in immunotherapies and vaccines. However, this requires the study of a very small proportion of cells, and in many cases prior knowledge of their MHC restriction and antigenic target. Multiple techniques have emerged to study antigen-specific

cells in diverse ways, including novel combinations of technologies to combine knowledge of specificity and phenotype. The study of antigen-specific T-cells provides insights that are not feasible from bulk T-cells and contributes to data that underlies future computational methods and the inference of specificity.

1.4.1 Challenges of Studying Antigen-Specific T-cells

Rarity of antigen-specific T-cells after contraction

Antigen-specific T-cell populations undergo 3 phases in response to exposure to their cognate antigen. Initially, approximately 1-2 days after infection, they undergo rapid expansion[208,209], with individual clonotypes expanding at least 7 times[129,210]. Following this initial expansion, the expanded population contracts, with approximately 95% of the cells undergoing apoptosis[211,212]. This contraction occurs predictably, whether the pathogen is cleared or persists, suggesting it is programmed early in the immune response[129,213]. This contraction then continues but at a lower rate, which can be approximated as an exponential decay process. Indeed, fitting measured ELISPOT responses against smallpox vaccine epitopes with time reveals a half life of 8-12 years[214]. Comparable methods applied to SARS-CoV2 in infected survivors estimate a half life of cytokine producing cells of 139 days[215].

Following pathogen clearance, most antigen-specific T-cells undergo apoptosis and only a small fraction form memory T-cells. Studies in mice infected with Lymphocytic Choriomeningitis Virus estimate that only 5% of the $CD8^+$ T-cells present at the peak response persist into the memory phase[72]. This contraction represents a technical challenge in studying antigen-specific T-cells because studies of immune memory have to sample from this small, contracted population. Indeed, estimated proportions of the T-cells within a PBMC $CD8^+$ T-cell repertoire specific to a given antigen (0.0005% to 18.7%)[216] are frequently below the limit of detection for techniques such as fluorescently-labelled peptide-MHC flow-cytometry (0.02%)[217].

The frequencies of T-cells specific to different epitopes varies, both within and across pathogen species. T-cell populations specific to latent viruses such as cytomegalovirus (CMV) and Epstein-Barr virus (EBV) are particularly abundant, connected to the

phenomenon of memory inflation. In this case, antigen-specific T-cell populations can expand over time rather than contracting. Kristensen et al. quantified these frequencies across 25 HLA-alleles, 929 epitopes and 29 viral species (as well as some human self-epitopes). Using DNA-barcoded peptide-MHC multimers, they estimated the frequency of epitope specific $CD8^+$ T-cell populations in patients that responded to the epitopes. Expectedly, the largest mean estimated frequencies were directed toward immunodominant CMV epitopes from proteins UL29/28 (7.16%) and pp65 (2.98%). Following this were epitopes from EBV (EBNA4, 1.25%), HPV (0.86%) and HIV (0.82%) in one donor. The next most frequent category are immunodominant antigens of viruses with repetitive exposures - examples of which include $M1_{58\text{-}66}$ of the Influenza A virus[218], some epitopes of seasonal coronaviruses and many SARS-CoV2 epitopes. Rarer still are subdominant epitopes with frequencies frequently below 0.1% and self-specific T-cells directed at tumour antigens such as MART1[216].

T-cell receptor cross-reactivity

The cross-reactivity of TCRs generates complexity in the interpretation of specificity assays. Some read-outs can (for reasons expanded on in Section 1.4.2) show binding evidence for multiple antigens by the same T-cell[219,220]. There is no clear answer as to how this should be processed, and what signals are indicative of genuine cross-reactivity rather than noise.

T-cell receptor independent reactivity

While conventionally the T-cell immune response is initiated by ligation of its receptor, activation markers and cytokines can be induced through mechanisms independent of the TCR[221]. Both $CD4^+$ and $CD8^+$ T-cells[222] can demonstrate IFN-γ[223], cytotoxic capacity[224], and genetic signatures indicative of migration to inflamed tissues[225] without evidence of TCR engagement. This is thought to be induced by inflammatory cytokines as well as pattern-recognition receptors expressed on T-cells. IL-15 in particular powerfully induces the proliferation, IFN-γ production and cytotoxicity on $CD8^+$ T-cells[226,227], whereas T-cell expression of TLR2[222] and IL-12[228] have been implicated in bystander activation of $CD4^+$ T-cells.

These "bystander" cells complicate assessments of antigen-specificity on the basis of response. For example, activation markers such as CD69, CD38 and HLA-DR are used in flow cytometry to identify activated cells, but can be expressed in TCR independent activation[229–231]. ELISPOT assays, which identify cytokine releasing T-cells, can also have their specificity affected by bystander T-cells since these potently produce IFN-γ. Likewise, attempting to identify antigen-specific clonotypes on the basis of their expansion in infection is also affected by the proliferation of bystander T-cells.

It is not only response based identification that can be confounded by TCR independent mechanisms. Some leukocyte Ig-like receptors such as ILT2/ILT4 can bind to most classical and non-classical MHC class I molecules[232–234]. Both HLA-A3 and HLA-A11 tetramers can bind natural killer lymphocyte receptors such as KIR3DL2[235], KIR2DL1 or KIR2DS1[236], all of which terminally differentiated $CD8^+$ T-cells can express[237]. This represents an unknown-unknown in TCR-specificity experiments, whereby the lack of understanding of the T-cell surface proteome that can bind peptide-MHC could lead to noise described in multiplexed peptide-MHC multimer assays[238]. To the best of my knowledge, no surface immunoprecipitation or analogous experiments have been carried out to systematically identify non-TCR binders of peptide-MHC.

1.4.2 Methods for Defining Antigen Specific T-cells

All methods have drawbacks relevant to the characterisation of antigen-specific cells. Experimental methods fall into two major categories - T-cell receptor binding and T-cell response based methods, which differ on the basis of whether they measure whether T-cells bind a given peptide-MHC or whether they respond to a particular peptide. The techniques underlying these methods give rise to their strengths and weaknesses, whereby binding based methods require prior knowledge of the specific peptide and MHC of a given T-cell response, but can detect antigen-specific T-cells of any phenotype. In contrast, response based methods use specific indicators of a T-cell response - cytokine secretion, proliferation or surface expression of activation markers to identify antigen-specific cells. None of these methods are infallible, and represent enrichment of antigen-specific cells

rather than pure isolation. Their advantages and potential blind spots or false positives stem from the preparation and assumptions made by the method.

Experimental Methods

T-cell receptor binding based methods

T-cell receptor binding based methods make use of modified, pMHC based constructs that preferentially bind to T-cells specific to that pMHC. They are theoretically capable of capturing the whole diversity of antigen-specific phenotypes - naive, T_{reg} or exhausted, by directly measuring TCR-p-MHC binding rather than a particular feature of the response. While it inherits its strengths of directly profiling TCRs that bind a given peptide-MHC, it also inherits the weaknesses from this approach. Most importantly, it requires researchers to know, in advance of the experiment, the specific antigen and presenting MHC.

The TCR-p-MHC interaction is low affinity, with fast dissociation rates that can trigger T-cell activation through serial engagement[239] but are insufficient to label cells with pMHC monomers[240]. To overcome the low affinity of the TCR-p-MHC interaction that rendered pMHC monomers ineffective for isolating antigen-specific T-cells, Altman et al. produced soluble pMHC monomers with a biotinylation tag, which allowed linking to a fluorochrome coated streptavidin in a 4:1 ratio of pMHC to streptavidin[240]. This multimeric arrangement of peptide-MHC constructs, named tetramers, could simultaneously bind multiple TCRs on the same cell leading to an accumulated binding strength, or avidity, that was sufficient to preferentially mark these cells. Combined with the ability to sort cells on the basis of their fluorescence, researchers could now isolate these cells, which enabled Altman and colleagues to study the phenotypes of $CD8^+$ T-cells specific to a HLA-A*02:01 HIV Gag epitope in seropositive donors. The multimeric innovation of Altman et al. has been extended to dodecamers[241] and spheromers[242] which similarly leverage the avidity principle to increase sensitivity.

The low affinity of the TCR-p-MHC interaction generates several challenges for receptor binding based methods. Firstly, binding based methods are biased towards higher-affinity TCRs, since lower-affinity TCRs are frequently below the affinity required to be identified

with pMHC multimers[243]. This means binding based methods are biased towards the highest affinity TCRs, which likely influences the subsets that are isolated[219]. Leading on from this, binding based methods are more frequently applied to $CD8^+$ T-cells, owing to their higher TCR affinities and greater ease of generating tetramers for the shorter epitopes[244]. Protein engineering approaches have been applied to improve the binding of class II MHC to the TCR to improve the sensitivity of class II tetramers. Second, despite aiming to isolate cells that bind pMHC, tetramer staining is not totally specific. Some cells that bind to a given pMHC will not be activated by it, owing to their phenotypic state or attributes of their T-cell receptor[245,246]. Additionally, some studies involving tetramers find that when they transduce cell lines with the identified "antigen-specific" TCRs, it does not bind the tetramer it was isolated with[55].

In contrast to response based methods, T-cells isolated using tetramers are not intentionally stimulated and thus more closely reflect the states they occupy in-vivo[247]. The manipulations done to these cells are those steps required for extraction, isolation and sequencing, all of which would occur in response based methods as well. As a result, T-cell receptor binding methods are best suited to measuring the phenotypes of cells.

T-cell receptor binding methods allow for the ex-vivo isolation of antigen-specific cells for further study by flow cytometry. In this case, cells are sorted using FACS for the fluorochrome coating the streptavidin. This can also be achieved using multimers attached to rare metals via mass spectrometry, which has greater sensitivity and capacity for multiplexing. However, cells are incinerated, so they cannot be collected for TCR sequencing or further functional profiling[248]. The use of flow-cytometry places limits on the sensitivity and multiplicity of fluorescent pMHC multimer methods. Spectral overlap between fluorophores limits the number of specificities and markers that can be measured simultaneously to around 25[249], and the limit of detection is approximately 10^{-4}, which is greater than the frequency of several antigen-specific populations (Section 1.4.1).

The multiplexing limits imposed by flow-cytometry can be overcome by attaching DNA barcodes to peptide-MHC multimers. Commonly used with gel-bead in emulsion sequencing (as provided by 10x Genomics), this method theoretically allows over 1,000

T-cell specificities to be measured simultaneously[250]. Each pMHC multimer (commonly dextramers[84,230,251]) reagent has a unique molecular identifier sequence (UMI) and they frequently share a fluorescent label (i.e. PE) to isolate or enrich for cells binding any antigen.

Since each droplet is formed from a cell and its immediate surface contacts, bound multimers are sequenced in a mixture with the cell's RNA. This allows bioinformatic mapping of pMHC (via the UMI sequence) to its bound cell, as well as quantification of that cell's transcriptome through single-cell sequencing. As a result, DNA barcoded dextramers facilitate scalable association of specificity and phenotype[251]. In dilution experiments the original authors of the method demonstrate they can detect populations specific to a HLA-B*07:02 CMV epitope in as low as 0.008% cells.

The process of assigning specificity from barcode read counts is inconsistent. The first paper to introduce DNA barcoded pMHCs, by Bentzen et al., models barcode counts as a negative binomial distribution and tests for an elevated signal compared to background that has not been sorted for the fluorescent label, and therefore not enriched for any of the pMHC multimers[250]. As part of a 2020 whitepaper[1], 10x Genomics released the largest publicly available dataset of this type (*the 10x Genomics dataset*), assessing the binding specificities of 150,000 $CD8^+$ T-cells from 4 donors to 44 pMHC multimers. This set of 15,214 unique clonotypes with annotated specificity remains the largest study within the VDJdb dataset of paired $\alpha\beta$ TCRs[252]. In their original paper, pMHC specific cells were defined by having a barcode count greater than 10 and a barcode count of at least 5 times the largest negative control pMHC.

Subsequent analyses showed that this led to an "unexpectedly high number of cross-reactive TCRs"[251], and suggested Integrative COntext-specific Normalization (ICON), whereby the data is stringently filtered on high-quality cells and paired TCRs, all pMHCs have the cell's highest negative control dextramer subtracted and multiple pMHCs binding the same cell are penalised. A "corrected dextramer signal" is generated from the negative-control subtracted count, the proportion of pMHC reads formed by that specific pMHC and the proportion of cells with the same clonotype that bind that pMHC. This signal is then

normalised across dextramers within a cell, then these normalised scores were normalised again over pMHCs, where the cutoff for binding was 0.

A different approach, Improved T Cell Receptor Antigen Pairing (ITRAP), also aims to identify genuine binders from noise. It does this by calibrating a pMHC UMI count threshold for genuine binders on large clonotypes where the signal to noise ratio is better. This approach ensembles "cells" (Gel-beads in emulsion, which can be doublets or empty droplets) across clonotypes, reducing the counts via the mean across cells with the same $\alpha\beta$ TCR. If a pMHC's dextramer counts, under the Wilcoxon signed-rank test, is significantly higher across an aggregated clonotype (>10 cells) than the next highest pMHC, those clonotypes are assigned that pMHC as their "true specificity". Then, for those large clonotypes an accuracy score can be computed across the dataset as the proportion of cells where the top pMHC is the clonotype's "true specificity". With the accuracy score as well as the fraction of retained cells as an objective, the authors exhaustively search pMHC UMI count thresholds for all the pMHCs in the dataset to identify the best performing cut-offs[238]. This approach is adaptable, which the authors demonstrate by filtering pairs to only those where the donor has the pMHC HLA allele.

While ground truth data for all these TCRs is implausible to collect, a subsequent benchmarking paper by the authors of ITRAP showed that both ICON and ITRAP substantially improve data quality compared with the original filtering. Both approaches label approximately 75% of the raw data as noise, but ITRAP retains more pMHCs, more cells and achieves greater within-pMHC TCR similarity than ICON. Retraining the NetTCR-2.1 model with antigen-specific cells defined by the original 10x cutoffs, ICON filtering or ITRAP filtering showed that ITRAP filtered data trains a model that best generalises to unseen VDJdb TCR-p-MHC data[220]. Owing to the stringent filtering and concerns about wasted data, subsequent papers using barcoded dextramers don't use either algorithm to identify genuine binders. Frequent approaches in the literature include taking the maximum pMHC count for a given cell[84], setting a per-cell proportion of reads threshold[253] or creatively fitting a negative binomial regression to technical variables and calculating residuals[219].

Despite the strengths of TCR binding based methods, they share a fundamental practical limitation - they cannot easily be used for antigen discovery since they require prior knowledge and design of the peptide and MHC. While both yeast display[254] and DNA-barcoded MHC complexes[255] can be used with libraries of peptide-MHC in binding based methods, they are less commonly used for this due to the specialised expertise and reagents required. An experiment involving peptide-MHC multimers to identify antigen-specific cells requires the researcher to know the peptide and restricting MHC of the response in advance[256]. In contrast, because response-based methods make use of the donor's own antigen-presenting cells and any of their HLA alleles, this can identify responses to peptide regardless of the specific k-mer or presenting MHC allele.

T-cell response-based methods

T-cell response-based methods aim to identify cells that change some measurable phenotype in response to T-cell receptor binding and activation. Typically, these methods involve stimulating cells in-vitro with antigen(s) of interest and then detecting responsive cells either through surface markers[148], the capture of release molecules[257] or tracking of clonotype frequencies[258]. The key advantages and drawbacks of these methods relate to them detecting response rather than binding: they are highly sensitive for cells that respond in the measured way and do not require prior knowledge of epitopes or HLA restriction due to the use of antigen-presenting cells. Since antigen-presenting cells are used, the cells can process and present "pulsed" antigens, allowing researchers to measure the response to a whole viral protein rather than a single epitope. However, as they modify cells through stimulation, they are less suited to studying ex-vivo phenotypes, as well as missing antigen-specific cells without the measured response. Moreover, since signatures of T-cell receptor-independent and dependent activation overlap (e.g. both can lead to IFN-γ and CD69 expression[231] or proliferation[259]), these methods can include non-specific "bystander" cells as discussed in Section 1.4.1.

Enzyme-Linked Immunosorbent Spot (ELISPOT)[260] uses the transformation of a substrate into measurable spots by trapped enzymes to quantify T-cells that produce a particular analyte (such as a cytokine, commonly IFN-γ). It, therefore, directly measures the secretory

function of cells rather than indirectly measuring the presence of molecules within the cell or their activation level. Researchers incubate cell mixtures containing T-cells and antigen-presenting cells with the stimulus of interest in wells coated with capture antibodies. The stimulus can be infected antigen-presenting cells, any combination of peptides or positive controls such as PMAi (phorbol myristate acetate and ionomycin). After incubation, researchers isolate the trapped molecules by washing the wells and add a biotinylated detection antibody and enzyme conjugate, which attaches to the trapped molecules. To form the spots, one washes unbound enzyme away and adds the enzyme's substrate, which is transformed into a visible and insoluble substance. Since the method does not require sophisticated equipment (beyond a plate reader and an incubator) and can operate on ten-fold less material than flow cytometry[261], it is cost-effective and can rapidly generate large volumes of important evidence about the immune response, as occurred in the SARS-CoV2 pandemic[256].

The ELISPOT reaction is quantitative since the number of "spots" can be measured at a given peptide stimulation concentration and cell count. A single set of capture antibodies, conjugated enzymes and substrates limits the assay to tracking one molecule secreted by T-cells, so while it is highly sensitive for that particular cytokine, it is less sensitive for antigen-specificity in general. Due to lower functionality or differentiation towards subsets that do not secrete that particular cytokine, the ELISPOT can miss meaningful numbers of antigen-responsive cells[262]. However, dual-colour and FluoroSpot assays allow measuring responses across multiple cytokines. In viral infection, since polyfunctionality is a correlate of T-cell efficacy against infection in both HIV and Leishmania[263], the ability to identify polyfunctional T-cell responses is of particular interest. Since the ELISPOT provides only a count of spots after stimulation, it cannot identify what particular cell type released the cytokines or provide cell-level information. Researchers overcome this by using flow cytometry to sort cells prior to the ELISPOT assay, such that they identify IFN-γ producing cells within sorted $CD4^+$ cells[95].

Since the ELISPOT is sensitive, scalable, cell-efficient and comparatively simple, researchers can apply it to complex factors of the T-cell response[261]. Three illustrative

examples include identifying T-cell targets of a pathogen, characterising the fine-peptide requirements of T-cell responses as well as quantifying the functional avidity of responsive T-cells. The SARS-CoV2 pandemic is full of examples of all 3, including identifying the T-cell targets of SARS-CoV2 within the proteome[95], how mutations in SARS-CoV2 evade T-cell immunity[264] and identifying particularly high functional avidity T-cell responses that dominate assessed clones in mild infection[16]. Beyond stimulating cells for cytokine secretion, the ELISPOT assay does not irreversibly modify the T-cells, meaning that they can be used for further assays. However, owing to its usual application to measure a single cytokine, it can systematically underestimate antigen-responsive cells, as was found in a meta-analysis with validation of SARS-CoV2 responsive T-cell studies across different methods[262].

Intracellular cytokine staining (ICS) is one of these flow-cytometry-based techniques that researchers use to more deeply phenotype-responsive cells. By stimulating cell mixtures containing antigen-presenting cells in the presence of protein transport inhibitors, newly synthesised cytokines are trapped within the Golgi apparatus. Cells can then be stained for surface markers for cell types or lineages (such as CD4, CD8 or CCR7) and are then fixed and permeabilised to allow labelled antibodies to reach intracellular cytokines[265].

Since ICS is multiparametric - meaning readouts reflect the presence of multiple label antibodies simultaneously, intracellular cytokine staining allows gating and identification of cell subsets that express combinations of cytokines and lineage markers. ICS can, therefore, inherently identify polyfunctionality but also relate T-cell cytokine expression to lineage markers. For example, Zuo et al. provided important evidence about the longevity of SARS-CoV2 cellular immunity in 2021 by characterising T-cell responses at 6 months post-infection. They found that cytokine-positive $CD4^+$ cells had twice the frequency of cytokine-positive $CD8^+$ within their overall subsets. They found IL-2 to be the dominant cytokine within $CD4^+$ and that polyfunctionality (defined as IL-2^+IFN-γ^+) was more frequent in non-spike than spike-responsive memory cells[266].

Owing to the use of flow cytometry, ICS is slower, less sensitive and requires more expertise and equipment for any given cytokine than ELISPOT. Additionally, due to the fixing and

permeabilisation necessary to label intracellular cytokines, the cells are made non-viable and cannot be used for secondary experiments[256].

During the SARS-CoV2 pandemic, many groups adopted stimulation of T-cells and flow-cytometric staining for surface proteins associated with activation to identify responsive T-cells. This method, the Activation Induced Markers (AIM) assay, relies on the principle that in response to TCR engagement, T-cells initiate various signalling cascades that influence the presentation of certain proteins on their surface[267]. Like ELISPOT and ICS, T-cells are incubated in mixtures containing antigen-specific cells and pulsed with the peptide(s) of interest. When looking for SARS-CoV2 spike-specific T-cells using AIM, the most commonly used markers were OX40 and CD137 for $CD4^+$ and CD69 and CD137 for $CD8^+$ T-cells, with CD25 also used for both[262]. Given that these activation markers have earlier response profiles to cytokines, their stimulation conditions and time need to be adapted to the marker used. The AIM assay can identify broad categories of activated T-cells without requiring them to secrete (a specific) cytokine, which means that they can more consistently identify subsets that are otherwise functionally specialised, such as circulating T_{fh} cells[256,268]. Since the assay makes use of multi-parameter flow cytometry, like ICS, it allows researchers to stratify analyses by combinations of surface markers, allowing detailed phenotypic characterisation of the stained cells[269,270].

This sensitivity to a variety of T-cell subtypes comes with drawbacks in terms of specificity, however. Zheng et al.[271] recently identified significant TCR-independent activation during AIM assays, showing that the conventional $CD4^+$ markers OX40 and CD25 frequently identify non-specific T-cell populations, particularly within $CCR6^+$ memory populations. They investigated the observation that $CCR6^+$ is reported amongst antigen-responsive cells in AIM assays[179,269,272], but largely absent on cells identified using peptide-MHC multimers[273,274]. Confirming this using $OX40^+CD25^+$ or $OX40^+CD137^+$ AIM and tetramer sorting for immunodominant $CD4^+$ spike epitopes in the same cohort, they show that $CCR6^+CXCR3^-$ cells make up more than 45% of AIM positive cells but less than 4% of ex-vivo tetramer-sorted cells. When they stimulated tetramer-sorted T-cells with peptide for 20 hours, fewer than 5% of tetramer-specific cells became $CCR6^+CXCR3^-$. Using an

in-vitro trans-well culture system to identify the role of contact versus soluble mediators in the activation of $CCR6^{+}CD4^{+}$ T-cells, they stimulate PBMCs in one chamber and show that PBMCs in the other chamber upregulate AIM markers (OX40, CD25, CD137) despite not being able to make physical contact with the stimulated cells. Using blocking experiments, they identify a key role for IL-2 secreted by an intermediate cell type in TCR-independent activation. Following this, they identify a gene signature of stimulation in their trans-well experiment and use this to identify a modified AIM approach using $CXCR4^{-}CD137^{+}$, which is much more specific for $tetramer^{+}$ cells (43% vs 6% for $CD25^{+}OX40^{+}$).

One final method that sits at the intersection of experimental and computational methods is following T-cell repertoires longitudinally and identifying clones that increase in frequency between time points where there has been an exposure. By quantifying the number of clones with a particular TCR sequence before or after an exposure, researchers can identify antigen-responsive clonotypes on the basis of what TCRs increased in frequency after stimulation. For example, Minervina et al. used this to investigate T-cell responses to yellow fever vaccination in one individual, tracking bulk α and β clones longitudinally before and after vaccination. They use edgeR[275] to identify clonotypes that were significantly expanded with exposure. Expanded clonotypes reach a peak of 0.5% of the TCR repertoire at 10-15 days after vaccination, whereas, in secondary responses, the expansion occurs by day 5 but with a lower magnitude.

Since this method does not require knowledge of the target antigen, it could be rapidly applied during the SARS-CoV2 pandemic. Milighetti et al. sequenced alpha and beta bulk repertoires to track clonotype frequencies in 41 healthcare workers who became PCR positive, as well as six individuals who remained seronegative. Defining expanded TCRs as those significantly greater at timepoint two than timepoint one under a Poisson distribution with $p<0.0001$, they identified a set of 4,075 alpha and 5,458 beta sequences of expanded TCRs. The TCRs peaked within the same week of the first positive PCR and, in some cases, before. Confirming the strong enrichment of this population for antigen-specific cells, the authors cross-referenced with known SARS-CoV2 specific TCR databases and found a much higher frequency of matches in expanded TCRs (3.1% vs 0.06% in random

non-expanded controls). Identifying these rapidly expanding clones led to the question - how do these virus-specific cells proliferate so quickly after exposure? Using a pre-pandemic dataset of beta TCR repertoires, the authors developed a statistical framework to estimate the frequencies of the expanded clonotypes pre-pandemic. Of the expanded TCRs, 52% were found in two or more pre-pandemic individuals, and their average estimated frequency was 1.8×10^{-6}, "over a thousand times higher than [10^{-9}] ... the estimated frequency of individual TCR sequences in the naive repertoire."[258]

Computational Methods

The most straightforward method for defining antigen-specificity computationally is a perfect amino acid match between a known antigen-specific TCR and a query TCR. As TCR specificity is defined by its amino acid sequences, if two TCRs share identical amino acid sequences over both chains, they can bind the same epitope. Extending this potential for binding to infer actual binding to the same epitope makes several assumptions: firstly, the individuals' HLA can present that epitope. Secondly, the same epitope is processed and presented across individuals, even in individuals with the same infection and presenting HLA allele, which is not always true. For example, in a cohort of patients recovered from COVID, 21% of individuals with the presenting HLA type of a SARS-CoV2 antigen did not generate IFN-γ responses to that antigen[95]. While this could represent inter-individual differences in the cytokine phenotype of surveyed cells, polymorphisms in non-HLA proteins can influence the immunopeptidome[276], meaning that even matching TCRs on a background of overlapping HLA alleles is insufficient for the presentation of a specific epitope to the same TCR.

Unsupervised Clustering Methods

Multiple techniques, broadly classified as unsupervised clustering techniques[277], aim to overcome the rarity of matching TCR sequences between individuals. Since any individual clonotype is so rare, the probability of matching can be boosted exponentially by allowing for a certain degree of sequence degeneracy. Analysis of "selected" repertoires taken from a variety of studies shows that the degree to which they show "selection" can be quantified by

the ratio of matching probabilities between a selected repertoire and a background repertoire. By analysing sequence coincidences (sequence matches allowing for Δ amino acid differences), Mayer and Callan show that TCR binding exhibits consistent degeneracy patterns across diverse experimental datasets. Sequence similarity is higher in epitope-selected populations, with the enhancement between a selected repertoire and a background repertoire decreasing exponentially as the number of allowed amino acid differences increases[47].

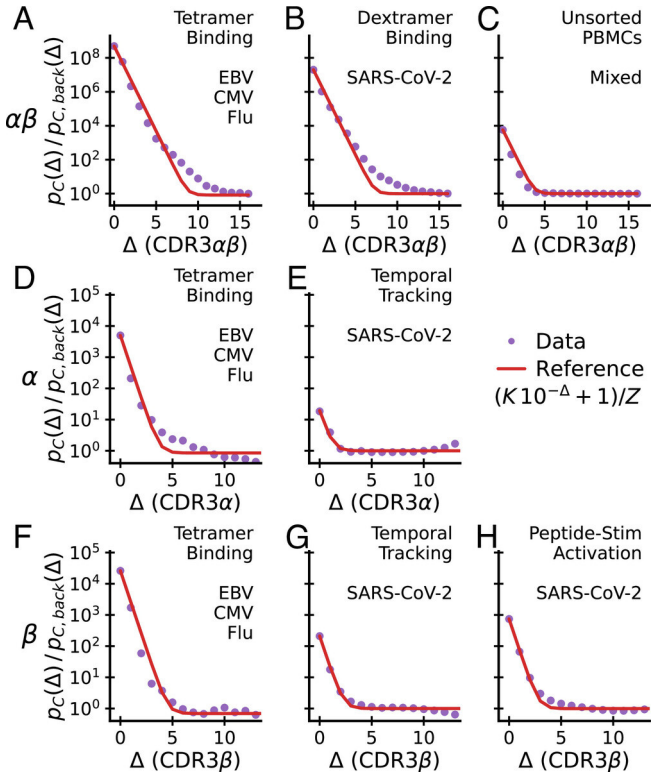


Figure 1.6: Measures of epitope binding degeneracy from T-cell receptor repertoires by Mayer and Callan, PNAS, 2023, Figure 3[47]. Original Caption: "Excess coincidences follow a common functional form across experiments. Sequence similarity of specific T-cells for paired $\alpha\beta$-chain repertoires (Top), α-chain repertoires (Middle) and β-chain repertoires (Lower) compared with background expectations. In each panel, the assay type used to enrich for epitope-specific T-cells and the antigen source are noted in the upper right. Panel C is special as analysed TCRs are from unsorted blood and have not been explicitly selected for binding to a specific epitope. A common reference curve is plotted for visual guidance. Its parameter K is set equal to observed value at $\Delta = 0$. Z is determined by normalisation."

The exponential falloff observed hints at a fundamental biophysical constraint - it estimates that approximately $\frac{1}{10}$ of single amino acid neighbours of an epitope-binding TCR maintain epitope specificity. Combined with the intuition that similar expanded sequences present

in activated cells across individuals with the same exposure likely share a target[84,253,278–280], this generates a powerful idea to identify antigen-responsive clonotypes.

One application of this is the concept of meta-clonotypes[279], a broadening of perfect sequence matches that address the rarity of public, paired TCR clonotypes. Using a distance metric calibrated for aligned TCR sequences, the "TCRdist" between TCRs is calculated. Using either a cutoff value or clustering algorithms over the distance matrix, such as DBSCAN or k-Nearest Neighbours, groups of similar T-cell receptors can be identified. The algorithm in its latest form[279] also constrains meta clonotypes to a CDR3 motif, requiring the presence of certain amino acids in sequence.

Since the initial papers demonstrated the utility and insights of clustering T-cell receptors, numerous methods evolved to group similar sequences. Jouannet et al. evaluated nine widely used TCR clustering algorithms for their performance over a collection of "as-is" datasets of publicly available annotated TCRs. Their analysis revealed meaningful differences in retention (how many known antigen-specific TCRs form clusters) and purity (whether a cluster of TCRs all share the same antigen-specificity). For example, DeepTCR, which forms large clusters, has high retention, whereas GLIPH2 excels at cluster purity but generates smaller clusters. GLIPH2 and DeepTCR performed the best in terms of sensitivity for antigen-specific TCRs[281].

A key bottleneck for some clustering algorithms is the need for pairwise comparisons when calculating distance metrics. If the algorithm has to compare every amino acid between two sequences for gaps or differences, it loops over sequences n^2 times. Innovations taken from spellchecking have sped up this process 100-fold[282]. However, an alternative is to directly embed sequences into a "representation", which is easier for computers to compare. Examples of this approach include GIANA[283], which claims "the same level of clustering specificity as TCRdist at 600 times its speed".

These representations have included domain knowledge - individual amino acids are turned into representations using their BLOSUM62 substitution matrices and aligned[283], or representations are encouraged to mirror Levenshtein distance[284] closely. More recently, the deep learning paradigm of learning useful representations directly from the data has been applied

to TCRs. DeepTCR, mentioned earlier, learns a representation of the T-cell receptor by learning to "reconstruct" the CDR3 sequence and V/(D)/J genes from a set of numbers. Eventually, this set of numbers can be used to cluster TCRs.

SCEPTR, on the other hand, applies contrastive learning to "censored" TCRs, where entire chains or elements of the CDR3 loop are hidden, and the model is trained to give censored and uncensored versions of the same TCR similar embeddings. These representations already achieve higher performance on the six epitopes tested than TCRdist, but they go further to fine-tune the model on epitope binding information. As a result, the model is encouraged to make the representations of TCRs that bind the same epitope more similar and those that bind different epitopes more different. This supervised contrastive learning sharpens the embedding, further improving the discrimination of different specificities[285].

Supervised Prediction Methods

Supervised prediction methods aim to map TCR sequences to their cognate peptide-MHC using training data of known TCR-peptide-MHC pairs[277]. Performance is measured by the area under the receiver operating characteristic curve (ROC–AUC), where random guessing with a balanced dataset of binders and non-binders would achieve 50% ROC–AUC. Supervised prediction methods are attractive because rather than mapping a TCR to the epitope for which there are similar TCRs, pan-peptide approaches can score TCRs for their likelihood of interaction with any peptide-MHC, even those not present in the training data. Beyond that, both Karthikeyan et al. and Fast et al. demonstrate generative approaches that allow the generation of TCRs binding a desired epitope[286,287].

Currently, these methods achieve strong performance on class I immunodominant epitopes with abundant examples in the training data[288], with the best-performing epitopes across 18 models being those with the most examples in public datasets[289]. Pre-trained models that first "learn" from general TCR-pMHC binding pairs and then are fine-tuned on a specific epitope they had not seen in "pre-training" indicate that high performance (>80% ROC-AUC) can be achieved with as few as 15 positive examples[290]. This is encouraging, as it suggests that a single study reporting a few hundred antigen-specific TCRs can make a

meaningful difference in the ability to identify antigen-specific cells in the future. This is a bonus of studies of antigen-specific T-cells - not only do they directly provide more detailed insights into antigen-specific subsets, but they also facilitate the identification of those subsets in future studies.

More broadly, the challenge with inferring specificity is a data problem[291]. Datasets of experimentally validated TCR-pMHC pairs are sparse and biased towards immunodominant viral epitopes[277], and various architectures converge on common (near random) performance for unseen epitopes.

One of the specific challenges for the development of supervised prediction methods is the absence of negative data. Since TCR-pMHC data involves labelled pairs (positive examples), very few negative examples exist with comparable experimental techniques[277]. Therefore, to generate negative examples for training, researchers shuffle epitopes or use unselected repertoires. As a result, negative data generation directly influences performance on downstream benchmarks - models trained on shuffled epitopes perform better where the test set was generated that way and vice versa[292]. Techniques that can better leverage data, i.e. by only training on positive examples and estimating how well a given sequence "fits" with its knowledge of positive examples[293], or by pre-training on protein interactions more generally, appear more resilient to different ways of generating test sets[294].

Phenotypic Inference

One can also attempt to infer antigen specificity from a T-cell's phenotype. This is common in the field of tumour immunology, where heterogeneous tumour mutational landscapes necessitate methods that are generalisable across patients to find tumour-reactive T-cells. However, it has been applied in SARS-CoV2 infection, for example, by Zhang et al., who use multimodal single-cell sequencing in a longitudinal mRNA vaccine cohort to identify clusters that expand after vaccination. Using these clusters to define a gene module, they find that the gene module is associated with clonal size and dextramer binding to immunodominant $CD8^+$ T-cell epitopes. Further refining their gene module through multiparametric flow-cytometry and pMHC dextramers, they identify $CD38^+HLA\text{-}DR^+KLRG1^-CD8^+$ as

"most highly enriched for antigen-specific CD8⁺ T-cells". The power of this approach was that they could subsequently apply it to CITE-Seq profiled datasets, and indeed, they found that within the COMBAT[34] dataset, both the frequency of this antigen-specific subset and the extent of clonal sharing of the subset with TEMRA cells was associated with improved outcomes in COVID-19[230].

1.4.3 Applications of Studying Antigen-Specific T-cells

Across infectious diseases, antigen-specific T-cells play important roles in the response to and memory from an infection. Isolating and studying these antigen-specific subsets, although challenging, enables insights into fundamental and translational immunology that are infeasible from unselected or mixed-specificity populations.

Quantifying Antigen-Specific T-cells

Despite their rarity, antigen-specific T-cells directed at different epitopes vary widely in their frequency[216]. Since these cells enact cellular adaptive immunity, their frequency is a key metric by which immune responses can be compared and evaluated. This is particularly true in the context of vaccine development, where the frequency of antigen-specific T-cells is used as a correlate of protection[146,295,296]. Where only the frequency is of interest, researchers approximate this using response-based methods such as ELISPOT, AIM or ICS assays. For instance, Haynes et al. analysed the immune correlates of protection of HIV vaccine and found that the frequency of Env (the HIV gene targeted by the vaccine) specific CD4⁺ T-cells measured by ICS did not correlate with protection[297], while markers of humoral immunity did. Similarly, Ishizuka et al. analysed immune correlates of a sporozoite vaccine against *Plasmodium falciparum* human challenge, and found also via ICS that TNF-α, IL-2 and/or IFN-γ producing T-cells at 3 or 21-25 weeks post vaccination did not correlate with outcome. The PBMC frequency of vaccine responsive T-cells decreased to pre-vaccine levels shortly after vaccination and participants who never became parasitaemic (despite challenge) had very low responsive T-cell frequency amongst PBMCs, leading the authors to suggest the antigen-specific cells might mediate protection in the liver rather than the blood. The authors measured liver levels of vaccine-responsive T-cells in non-human primates after

vaccination, and found that the frequency of responsive $CD8^+$ T-cells was over 100 times higher than in the blood[298].

Highlighting the popularity of ICS for its ability to simultaneously profile multiple cytokines in defined subsets, Arunachalam et al. used ICS to track T-cell responses to Spike following BNT162b2 vaccination. They found that $CD4^+$ and $CD8^+$ T-cells only minimally expanded following priming immunisations, but expanded significantly after boosting[299]. This highlights that quantifying antigen-specific cells can not only help to measure protective immunity, but also provides insights into dynamics and where protective immunity occurs. However, since these methods manipulate cells in order to measure their response, they are generally less suited to studying the ex-vivo phenotypes generated by vaccination or infection. To do this, researchers use binding-based methods such as pMHC multimers to identify antigen-specific cells. Schulien et al. used pMHC tetramers to estimate the ex-vivo frequency of SARS-CoV2 specific $CD8^+$ T-cells before, during and after infection. They followed $M_{322\text{-}330}$-B*44:03 and $ORF1ab_{3946\text{-}3954}$-B*44:03 specific cells as they expanded from undetectable pre-infection to expanded a day after symptom onset and subsequent contraction over the following 70 days. Following minimally manipulated cells allowed them to identify phenotypic transitions with time - for example the gradual acquisition of memory markers ($CD127^+$, $BCL\text{-}2^+$) over the contraction phase[300].

Understanding the Paired TCR Repertoire of Antigen-Specific Cells

How many exposures make up a mature T-cell repertoire? This is hard to investigate owing to heterogeneity in individual T-cell responses and even the degree of expansion and maintenance across individuals[258] or epitopes[301]. However, using coincidence analysis under the assumption that there will be negligible coincidences between different antigen-specific repertoires, the number of antigen exposures that form a mature repertoire mixture can be estimated as the ratio between the coincidence probability of an antigen-specific repertoire and the memory repertoire. Empirically, this appears to be in the order of magnitude of 10^4[47]. Supporting this is that the coincidence probability within twins is approximately the same within an individual as between twins, suggesting that their exposures overlap meaningfully[46,47].

Functionally Characterising Antigen-Specific Cells

One powerful way in which the study of antigen-specific T-cells contributes to novel insights is the potential to connect clonotypes to functional programs. While this is possible through transcriptomic atlases[219,302,303] that have made major contributions to the evidence linking TCRs to T-cell phenotypes, these atlases are mixtures of T-cell populations with unknown specificities. As a result, the different antigen-specific groups that make up the repertoire will be on different timelines since their last activation and may have quantitatively different activations. In contrast, when studying antigen-specific cells within an individual, these cells will likely have consistent exposure to antigens and a shared cytokine environment at activation. Moreover, since the antigen-specificity is known, these experiments can be more easily validated experimentally with the functional profiles of their T-cell receptors.

For example, van de Sandt et al.[304] identified age-dependent shifts in antigen-specific T-cell responses to an immunodominant Influenza epitope. They isolated $CD8^+$ T-cells specific for an immunodominant epitope of Influenza A viruses (HLA-A*02:01-$M1_{58}$) from 61 individuals across the age spectrum (0-88 years) using tetramer-associated magnetic enrichment (TAME). Generating a single-cell transcriptomic dataset from 3 individuals per group, they identified a different pseudotime trajectory in older adults than the one shared by newborns, children and adults. Child and adult $M1_{58}$ specific $CD8^+$ T-cells expressed cytotoxic gene signatures that strongly overlapped with cells expressing a known highly functional and public TRAV27/TRBV19 clonotype. On the other hand, older adult T-cells more closely resembled newborn gene signatures, with the dominance of a cluster defined by naive T-cell transcripts (*CCR7*, *SELL*, *TNFAIP3*).

Hypothesising that different clonotype profiles between older adults vs. children and adults might drive the observed phenotypes, they generated TCR-transduced SKW-3 cell lines and measured CD69 expression at titrations of peptide stimulation. TCRs derived from older adults had approximately three orders of magnitude higher EC_{50} values than public TCRs present in adults. Indeed, in an alanine scan of the $M1_{58}$ epitope, the TCR derived from an older adult was the least robust to mutations. This was confirmed in proliferation

and intracellular cytokine staining assays on patient-derived antigen-specific T-cells, where older adults had less functional T-cells than the other groups. Estimating generation probabilities for the repertoires of the various groups, they find that, as expected, public TCRs in the adult group have high generation probabilities. However, in older adults, their entirely private repertoires have low generation probabilities, which the authors take to represent a "clonal reset" in late adulthood that depletes highly functional antigen-specific clones.

The authors measured the frequency of $M1_{58}$ tetramer$^+$ T-cells early in their paper and found these to vary from 5×10^{-5} to 2×10^{-4} across their groups. By isolating these cells, they were able to uncover this "clonal reset" in older adults, the association of specific clonotypes with functional profiles and evidence that the clonal composition rather than cellular exhaustion is associated with reduced functionality in older adults. Given how rare these cells are and the resolution necessary to identify these nuanced differences, this finding would be unlikely in bulk or mixed-specificity samples. Indeed, without knowing what antigen these T-cells bound, the alanine scanning or TCR transduction experiments would not have been possible.

Another remarkable example of connecting antigen-specific TCR clonotypes with functional profiles is in HIV, where Galperin et al. explored the functional profiles of a public clonotype previously identified in long-term HIV controllers. In these individuals, the same HIV capsid epitope (Gag_{293}) was presented by multiple HLA-DR molecules. In structural and functional studies, they showed that public TCRs present across these individuals had broad HLA cross-restriction (the ability to recognise the same epitope presented by different HLA alleles) due to peptide-focused recognition. In the interaction between one of these TCRs (F24) and pMHC, the TCR-peptide interaction made up 45% of the buried surface area, the highest observed in TCR–MHC II complexes. In being so peptide-focused, the TCR was less dependent on specific HLA features, enabling it to bind across HLA-DR1, DR11, DR15 and DR5. Functionally, the high affinity of this TCR enabled cytotoxic function in CD4$^+$ and CD8$^+$ T-cells despite the CD8$^+$ cells not having the appropriate co-receptor. Indeed, this cytotoxicity was sufficient to effectively suppress HIV replication in infected dendritic cells across multiple HLA classes. By studying one particular antigen-specific

population, the authors could dissect the structure/function relationship of a particular TCR and suggest a mechanism contributing to the long-term control of HIV[305].

Characterising the impact of Interventions or Mutations

Since antigen-specific cells are key effectors of immune responses in both infection and autoimmune diseases, to quantify the impact of interventions on these cells, they have to be isolated. This is commonly used, for example, in vaccination, but also for novel therapeutic approaches.

One example is when a particular therapeutic approach targets an antigen-specific population, such as the work by Tremain et al. on synthetically glycosylated antigens for antigen-specific rather than global suppression of immune responses. Harnessing the role of the liver as a key site for peripheral tolerance, they build on prior work from their group to manipulate targets of autoimmunity for presentation in the liver[306]. Their theory is that presentation by hepatic antigen-presenting cells, tuned to drive immune tolerance to apoptotic debris and digestion intermediates[25], will allow them to desensitise immune responses to particular antigens. Using various mouse models, they show that targeting antigens for presentation in the liver drives depletion and/or exhaustion of antigen-specific cells in a manner that can attenuate experimental autoimmune encephalitis (a mouse model of antigen-specific demyelinating disease). While the mouse models showed the impact on antigen-specific cells, the importance of studying these cells is highlighted in their work on non-human primates. There, they show that they can reduce antigen-specific immune responses to a vaccine epitope as measured by ELISPOT, without any detectable effect on bulk "$CD3^+$, $CD4^+$, $CD8^+$, NK cells, monocytes or B cells."[307] This illustrates the power of studying the particular antigen-specific population to quantify the impact of an intervention. This approach has shown phase II promise in humans, where a synthetically glycosylated gliadin (gluten component)[308] significantly reduced "gluten induced symptoms and celiac-specific composite measures"[309], with an ongoing trial for Multiple Sclerosis[310].

While T-cell immunity is generally thought to be more robust to mutations in pathogens due to ensembles of antigen-specific T-cells against targets throughout the proteome and the inherent cross-reactivity of T-cell receptors, studies of antigen-specific T-cell populations

allow us to identify precise mechanisms of immune escape. Recently, Tian et al. investigated immune evasion by SARS-CoV2 variants BA.2.86 and JN.1. First, they used IFN-γ ELISPOT to identify previously characterised epitopes where mutations in these variants attenuated the T-cell response. They then deeply investigate $S_{448\text{-}456}$, showing that the combination of L455S, N450D and L452W mutations in JN.1 alters the conformation and hydrophobicity of the epitope when presented by HLA-A*24:01, associated with the loss of recognition by non-mutant specific TCRs. While they demonstrate that this evasion is detectable at a bulk level - 80% of donors carrying HLA-A*24 had declines in Spike subunit 1 ELISPOT counts, compared to 50% in A*2 and 12% in A*3, identifying the precise mechanism required isolating antigen-specific cells and testing them against these variants.

Vaccines aim, in part, to prime antigen-specific T-cell populations for more potent responses to subsequent infection. Since these antigen-specific T-cell populations are rare, studying these directly allows us to examine how vaccination or past exposure shapes the T-cell response. Minervina et al. examined 18 SARS-CoV2 epitopes presented by five HLA alleles using DNA-barcoded pMHC dextramers in individuals who had been previously infected, vaccinated or both, in different orders[253]. The mean frequency of dextramer-positive cells was 0.41% of $CD8^+$ T-cells, but because they could isolate this population, they could directly compare antigen-specific cells between exposure histories via CITE-Seq. Their experimental design allowed them to investigate how the order of hybrid immune exposures (infection then vaccination, or vaccination then infection) influenced the antigen-specific repertoires and their relative sizes. Vaccination after infection was associated with a terminally differentiated effector memory phenotype, with more cells in a cluster with surface expression of CD45RA and high transcriptome levels of *GZMB*, *GZMH* and *GNLY*. In previously infected individuals who were assessed with multiple HLA-A*01 dextramers, the $S_{865\text{-}873}$-HLA-A*01 epitope went from making up 0.8% of the A*01 restricted response before vaccination to 48% after vaccination. Breakthrough infections in their cohort generated strong non-spike responses comparable to infection-only individuals, showing that pre-existing spike-specific immunity in vaccinated individuals did not attenuate the generation of primary responses to other viral proteins.

Mechanistically Understanding Pathogenesis or Protective Immunity

Some children, who generally have milder COVID-19 infection, develop the autoimmune disorder Multisystem Inflammatory Syndrome in Children (MIS-C). The mechanisms by which MIS-C occurs are unknown. However, evidence from Lee et al. reports autosomal recessive deficiencies in the Mitochondrial Antiviral Signalling Pathway in ~1% of MIS-C patients, where defective OAS-RNase L pathway activation leads to excessive inflammatory cytokine production by mononuclear phagocytes upon SARS-CoV2 stimulation[14]. Additionally, other papers have identified an expansion of T-cells with particular V-J gene combinations and HLA risk alleles. Screening individuals with MIS-C and controls for autoantibodies and antibodies against SARS-CoV2, Bodansky et al. identify a set of similar epitopes in SNX8 (an intracellular trafficking protein, involved in the Mitochondrial Antiviral Signalling Pathway) and the SARS-CoV2 Nucleocapsid protein[55]. These epitopes are both targeted by antibodies in MIS-C patients but not controls. Since SNX8 is an intracellular protein, the authors hypothesise that cross-reactive T-cells primed against that Nucleocapsid epitope might cross-react with SNX8, causing cellular responses to cells engaging in antiviral signalling. Using both AIM and tetramers, they show that cross-reactive T-cells for SNX8 and the Nucleocapsid epitope exist in MIS-C. Some of the cross-reactive TCRs identified stain with both SNX8 and Nucleocapsid tetramers when expressed in Jurkats. In one of the patients where cross-reactive TCRs were isolated through tetramer staining ex-vivo, the cross-reactive clonotype was present in just six cells sequenced and sorted of more than 140000 $CD8^+$ T-cells analysed. Despite potentially mediating an important clinical mechanism, these cells are rare and would likely not have been identified in bulk $CD8^+$ T-cells since clustering and marker identification falter at tiny group sizes.

As with the Flu and Aging findings mentioned above[304], SARS-CoV2 provided an opportunity to study the contribution of antigen-specific T-cell clonotypes to COVID protection and disease. For example, Peng et al. isolated $NP_{105-113}$-B07:02-specific $CD8^+$ T-cells from patients with the allele using pentamers and compared their phenotypes in severe and mild COVID. Cells specific for this epitope were present at much higher frequencies in mild disease than in severe disease at 1-3 months convalescence. Additionally, and supporting our

other work finding broader targets in severe disease, individuals with mild COVID had T-cell responses much more "focused" on the epitope - their responses made up 60% overall NP ELISPOT counts, while in severe disease this was substantially lower at 19.5%. Severe cases had higher expression of cytotoxicity genes (*GZMK*, *GNLY*) and inhibitory receptors (*TIGIT*, *CTLA4*, and *TIM3*) in single-cell sequencing of these antigen-specific cells. A challenge of single-cell data is that a lot of data is generated about the transcriptome of individual cells, but it can be challenging to relate that to their functionality. However, in this study, using cultured T-cell clones with the same β TCRs as those found in the single-cell sequencing they were able to profile the functional avidity of these clones via EC_{50} measurements. High avidity clones made up a larger proportion of clones generated from mild patients than severe patients, suggesting that the expansion of these clonotypes may contribute to protection from disease.

While studying antigen-specific cells in the blood offers important insights into circulating T-cell memory, the most important responses are likely at sites of infection[149,199]. Markov et al used bulk TCR sequencing in conjunction with multiparameter flow cytometry and computational inference of T-cell specificity. Using GLIPH2 to map $CD4^+$ and $CD8^+$ TCR sequences to the MIRA dataset allowed them to infer antigen-specificity within 432 bronchoalveolar lavage samples from 273 unvaccinated, critical COVID-19 patients. They found that the quantity of $CD4^+$ and $CD8^+$ T-cells was associated with favourable clinical parameters such as respiratory system compliance and shorter durations of mechanical ventilation. High T-cell numbers early after intubation (<48h) were associated with discharge, whereas declining T-cell frequencies were associated with mortality. Where they could infer antigen-specificity from similarity between the MIRA dataset and GLIPH 2 enriched clusters, they found that of the annotated $CD8^+$ TCRs discharged patients had high specificity to Spike and Nucleocapsid epitopes (41.4%), whereas in deceased patients ORF1ab was the dominant target (43.2%). Potentially related was the relationship between age and BAL $CD8^+$ TCR targets they found - individuals over 65 more frequently targeted non-structural epitopes[198]. This difference in targeting suggests that the balance of antigen-specific T-cells in the lungs of COVID-19 patients can influence their clinical outcomes - further highlighting the importance of studying antigen-specific cells.

1.5 Rationale, Knowledge Gaps and Thesis Aims

In this thesis, I aim to apply single-cell analyses to better characterise antigen-specific T-cell immunity to SARS-CoV2. I do so by using single-cell sequenced, longitudinal data collected of antigen-specific cells after COVID-19.

1.5.1 How does HLA restriction shape T-cell memory in an immunodominant SARS-CoV2 Viral Epitope?

HLA restriction and its downstream factors, such as antigen-presentation quantity, stability and the binding TCR repertoire, are thought to influence phenotypes and functional profiles of responding T-cells[26,301,311]. However, these studies are frequently confounded by the use of different epitopes[301] and comparisons across individuals. While elegant studies to pick apart contributions of these factors are possible with animal models[311–316], so far this has not been possible in humans.

I make use of an immunodominant viral epitope, $M_{172\text{-}188}$, that my group and others discovered as a frequent target of $CD4^+$ T-cells in unselected populations of people recovered from COVID-19. My colleagues identify, through functional, flow-cytometry and structural studies, that multiple HLA-DRB1 alleles, including *01:01, *04:01, *07:01 and *15:01, present the epitope. This provides the unique opportunity to isolate T-cells specific for the same epitope but presented by different HLA alleles. In fact, due to heterozygosity, it is even possible to make paired comparisons of T-cells specific for the same epitope presented by different HLA alleles within the same individual. Leveraging this, my first results chapter compares $M_{172\text{-}188}$-DRB1*01:01 and $M_{172\text{-}188}$-DRB1*07:01 specific $CD4^+$ T-cells via single-cell analysis in individuals recovered from COVID-19.

1.5.2 How do spike-specific and non-spike-specific T-cells after vaccination evolve with breakthrough Omicron infection?

Hybrid immunity (a combination of immunity generated by vaccines and natural infection) generates different T-cell responses to those generated by either alone[253,317,318]. However,

these have generally been studied through snapshot observational studies that cannot compare the evolution of antigen-specific T-cells with multiple exposures and are "hindered by the lack of preinfection samples in most cases"[202].

My second results chapter leverages a single-cell dataset generated by longitudinally sampling healthcare workers for T-cells specific to 5 immunodominant viral epitopes, including Spike and non-spike antigens. These healthcare workers were part of a regular workplace PCR testing scheme, allowing us to identify COVID-19 recovered and COVID-19 naive individuals. These individuals were sampled after completion of their SARS-CoV2 vaccine course (3 doses) and again shortly after PCR confirmed Omicron breakthrough infection. By focusing on antigen-specific T-cell subsets, we can compare the transcriptomes and repertoires of T-cells specific to the same epitope but generated by different exposures.

1.5.3 How do the long-term phenotypes and repertoires of Spike and non-Spike specific CD4⁺ T-cells differ over time?

Since the SARS-CoV2 pandemic, my laboratory has generated large amounts of antigen-specific T-cell data, from the same individuals across multiple epitopes. To augment the longitudinal study of CD8⁺ T-cells in the healthcare worker cohort, I compare Spike and non-Spike specific CD4⁺ T-cells in their transcriptome, repertoire and phenotype and how these change in long-term follow-up samples.

Studies this far have characterised SARS-CoV2 specific CD4⁺ T-cell responses for up to two years after infection[270,319], but the longer term evolution in particular regard to repertoire narrowing[320] is unclear. Spike-specific and non-Spike specific CD4⁺ T-cells have different recall patterns owing to the approved vaccines in the UK. Spike-specific CD4⁺ T-cells are boosted by vaccination or infection, whereas non-Spike specific CD4⁺ T-cells are only recalled by infection. This divergence may lead to differences in their long term trajectories.

My third results chapter uses single-cell sequencing of MHC class II tetramer-sorted CD4⁺ T-cells from individuals recovered from COVID-19 at 1-3 months and four years after infection. By comparing transcriptomes and repertoires specific to different antigen targets over four years, I can explore how recall patterns shape long-term CD4⁺ T-cell memory.

2

Materials and Methods

Summary

This chapter describes the materials and methods used in this thesis. As this thesis was a bioinformatic thesis involving the analysis of previously generated experimental data, the methods described here reflect both experimental work not carried out by myself as well as the computational methods I used to analyse the data. Hence, all methods covered in the "Data Collection" section were performed by colleagues, but the methods are described in detail as they are relevant to the understanding of data I have analysed. The "Data Analysis" section describes the computational work I carried out alone, including analysis of this data and re-analysis of existing datasets.

2.1 Data Collection

2.1.1 Study Cohorts and Definitions

Longitudinal ISARIC/Sepsis Immunomics Cohort

Hospitalised, RT-PCR confirmed SARS-CoV2 infected individuals were recruited from the Oxford University Hospitals NHS Trust, UK, between March 2020 and September 2021 and recruited into the Sepsis Immunomics study and International Severe Acute Respiratory and Emerging Infection Consortium World Health Organization Clinical Characterisation Protocol UK (IRAS260007 and IRAS126600). Patients were sampled at least 28 days after symptom onset. Written informed consent was obtained from all patients. Ethical approvals were given by the South Central-Oxford C Research Ethics Committee in England (reference: 19/SC/0296 and 13/SC/0149). Clinical definitions, recruitment, sampling, and phenotyping followed previously described methods[95]. WHO criteria defined clinical classification of COVID-19[321], where the threshold for severe COVID was hospitalisation where oxygen was delivered by a Non-Invasive Ventilation Mask or high flow. The participants' four-digit HLA typing was carried out by sequencing genomic DNA at the Weatherall Institute for Molecular Medicine's genomic sequencing facility.

Longitudinal Sheffield Healthcare Worker Cohort

The Sheffield Healthcare Worker cohort was a subset of the PITCH (Protective Immunity from T-cells to COVID-19 in Health Workers)[322] cohort that were more deeply phenotyped with antigen-specific single-cell sequencing. Sampling details and criteria inherit from the above study, but in brief - "Eligible participants were adults aged 18 years and over, currently working as healthcare workers, including allied support and laboratory staff" and "participants were recruited under the Observational Biobanking study STHObs (18/YH/0441), which was amended for this study on 10 September 2020"[322].

Naive or previously infected status was defined "at the time of enrollment in the PITCH study based on documented PCR and/or serology from local NHS trusts, or from MSD analysis of S and N plasma IgG levels of PITCH samples[323]". "Patients were aimed to be

sampled 28 days following their third vaccine dose, and 28 days following documented SARS-CoV2 infection"[322].

2.1.2 Establishment of Antigen-Specific T-cell Lines

$M_{172\text{-}188}$-specific $CD4^+$ T-cells were identified and isolated based on cytokine secretion using FACS from donors within the cohort (Section 2.1.1). Cells were thawed from frozen PBMCs and stimulated with 10 µg/ml $M_{172\text{-}188}$ peptide (ProImmune) for 4hrs. The $M_{172\text{-}188}$ specific $CD4^+$ T-cells were labelled with TNF-α, IL-2, IFN-γ Secretion Assay – Detection Kit (PE), human (Miltenyi) and sorted using BD FACSAria™ III or BD FACSAria™ Fusion Cell Sorter. The isolated $M_{172\text{-}188}$-specific $CD4^+$ T-cells were then expanded and maintained as either T-cell clones or bulk T-cell lines following previously described methods[324]. $M_{172\text{-}188}$ responsive T-cells were cultured with irradiated cone blood cells in RPMI-1640 supplemented with heat-inactivated human AB serum, L-glutamine, penicillin-streptomycin and recombinant human IL-2 at 37°C for 14 days. To generate $M_{172\text{-}188}$-responsive $CD4^+$ T-cell lines with known restriction, the $M_{172\text{-}188}$ responsive cell lines were sorted with $M_{172\text{-}188}$-DRA1*01:01: DRB1*01:01, DRB1*04:01, or DRB1*07:01 tetramers. Known restriction T-cell clones and bulk cell lines were maintained and their purity was checked by $M_{172\text{-}188}$ tetramer staining after each round of expansion. Only T-cells with over 95% tetramer-positive staining were used for the experiments.

All cell lines were tested for mycoplasma every four weeks using the MycoAlert detection kit (Lonza).

2.1.3 Tetramer Sorting for Antigen-specific $CD4^+$ T-cells

$M_{172\text{-}188}$-specific $CD4^+$ T-cells

Frozen PBMCs were thawed and rested for 2 hours. The live cells were counted and re-suspended in R10 (RPMI 1640 medium with 10% fetal calf serum, 2 mM glutamine and 100 U/ml penicillin and 100 µg/ml streptomycin). Half of the cells from each sample was stimulated with 2 µg/ml $M_{172\text{-}188}$ peptide in R10 for 4 hours in the incubator. The other half was maintained in R10 for 4 hours in the incubator. The unstimulated and stimulated cells were washed and stained with Live/dead fixable Aqua dye (Invitrogen) in DPBS on ice

for 15 minutes. Then, the cells were washed and stained with fluorophore-conjugated HLA-DRA1:DRB1*01:01 $M_{172\text{-}188}$ or HLA-DRA1:DRB1*07:01 $M_{175\text{-}188}$ tetramer(s) (ProImmune) in R10 at 37 °C for 30min. Subsequent peptide pulsing experiments verify that cells sorted with $M_{175\text{-}188}$ DRB1*07:01 tetramers generate responses to pulsed $M_{172\text{-}188}$ (Figure B.1). No negative control or irrelevant tetramer was used in this experiment. Representative gating is shown in Figure C.1.

We then washed and stained with the following surface antibodies: CD3-BV785, CD4-PE (or PercP-Cy5.5 when using duo-tetramer), HLA-DR-BB515 (BD Biosciences), CD14-BV510, CD19-BV510, CD16-BV510, CCR7-BV421, CD45RA-APC-H7, PD-1-BV650, and CD27-PE-Cy7 (BioLegend) on ice for 20 minutes. After washing, the cells were resuspended in R10 and kept on ice before flow cytometric sorting. A BD Fusion sorter or BD FACS Aria III (BD Biosciences) were used to sort $CD3^+CD4^+$tetramer$^+$ cells with the exclusion of nonviable/$CD19^+$/$CD14^+$/$CD16^+$ cells. The cells were sorted into 96-well PCR plates (Starlab (UK) LTD) and stored at -80 °C for subsequent analysis.

Gating for canonical $CD4^+$ T-cell compartment markers (CCR7, CD45RA) was performed per patient. Fluorescence intensity thresholds for each individual were used to assign positive or negative surface protein expression based on each cell's fluorescence intensity in single-cell analysis.

Spike-specific $CD4^+$ T-cells

"3–5 $\times$ 10^6 cells were stained with APC-conjugated HLA-DRB1*15:01 $S_{751\text{–}765}$ and $S_{866\text{–}880}$ tetramers (ProImmune)" and "$S_{166\text{–}180}$ tetramer (NIH Tetramer Core Facility)"[325]. The protocol for tetramer staining and sorting into plates for sequencing was as described above for $M_{172\text{-}188}$-specific $CD4^+$ T-cells.

2.1.4 Tetramer-associated magnetic enrichment and cell sorting for four-year $CD4^+$ T-cells

As described by my colleagues[325], antigen-specific $CD4^+$ T-cells from four-year samples were enriched using tetramer-associated magnetic enrichment prior to single-cell sequencing. Between $1.5 - 3 \times 10^7$ PBMCs were labelled with APC- or PE-conjugated peptide-MHC

class II tetramers for 30 minutes at 37°C. Magnetic enrichment was then performed using anti-APC or anti-PE microbeads with magnetic-activated cell sorting technology (Miltenyi Biotec) following the manufacturer's protocol. Enriched tetramer-positive $CD4^+$ T-cells were subsequently sorted for single-cell RNA sequencing.

2.1.5 Tetramer Sorting for SARS-CoV2 specific $CD8^+$ T-cells

Frozen PBMCs from the Sheffield Healthcare Worker cohort (Section 2.1.1) were thawed, rested and stained with "PE-conjugated SARS-CoV2 spike or non-spike specific MHC pentamer/dextramers"[322], sourced from ProImmune. The specific multimers used to stain $CD8^+$ T-cells for Spike specific responses were `KCYGVSPTK`:HLA-A*03:01 (S_{378}), `YLQPRTFLL`:HLA-A*02:01 (S_{269}), `LTDEMIAQY`:HLA-A*01:01 (S_{856}). In addition, $CD8^+$ T-cells were stained for non-spike responses against a Replicase Protein epitope - `TTDPSFLGRY`:HLA-A*01:01 (RP_{1637}), and a Nucleocapsid epitope `SPRWYFYYL`:HLA-B*07:02 (NCP_{105}).

2.1.6 Intracellular Cytokine Staining for Antigen-Specific Responses

Intracellular cytokine staining was conducted as described previously[326]. Briefly, antigen-presenting EBV-B cells were pulsed with $M_{172\text{-}188}$ peptide at concentration per condition for one hour and washed to remove the non-presented peptide. T-cells were co-cultured with the peptide-pulsed B cells in the presence of GolgiPlug and GolgiStop (BD Biosciences) for another 4-5h. Dead cells were labelled using LIVE/DEAD Fixable Aqua dye from Invitrogen to stain the cells for FACS. Cells were then washed, fixed with Cytofix/Cytoperm and stained with PE-anti-CD4 (BD Biosciences), PE-Cy7-anti-IFNγ (BD Biosciences), APC-anti-TNFα (eBioscience) and BV421-anti-IL-2 (BioLegend). Negative controls of T-cells co-cultured with BCLs without peptide stimulation were run for each sample. A representative intracellular cytokine staining flow-cytometry plot for $M_{172\text{-}188}$ peptide pulsed B cells is shown in Figure B.1. All samples were acquired on Attune NxT Cytometer (Life Technologies).

2.1.7 HLA-Mismatched B-cells and Blocking Antibodies To Identify HLA-Restriction

To determine the HLA restriction of the $M_{172\text{-}188}$-specific $CD4^+$ T-cells, peptide presentation was first blocked with anti-DR or anti-DQ antibodies. The T-cells were counted, aliquoted, and treated with anti-DR or anti-DQ antibodies for 0.5 hours at room temperature. 2 µg/ml $M_{172\text{-}188}$ (ProImmune, UK) was added to stimulate the T-cells for 16 hours. Then the response was measured using IFN-γ ELISPOT to determine DR/DQ/DP.

BCLs with known HLA types were used as antigen-presenting cells, and MHC restriction was determined by comparing T-cell responses across BCLs sharing different HLA alleles. Recognition patterns were analysed using a matrix of shared HLA alleles between the BCLs that did or did not stimulate T-cell responses. The specific restricting HLA allele was identified as shared between all stimulatory B-LCLs and absent in non-stimulatory B-LCLs.

2.1.8 SARS-CoV2 Live Virus Propagation and Titration

SARS-CoV2 Victoria 01/20 strain (BVIC01) and variants of concern—Alpha (Lineage B.1.1.7, 20I/501Y.V1.HMPP1) and Beta (Lineage B.1.351, 201/501.V2.HV001)—were originally from Public Health England and provided by J. McKeating. SARS-CoV2 Gamma (Lineage P.1) was supplied by G. Screaton. The Delta (B.1.617.2) variant was originally from W. Barclay and T. De Silva (G2P-UK) and provided by G. Screaton. The BA.2 variant was supplied by G. Screaton. Victoria 01/20, Alpha and Beta variants were propagated with Vero E6 cells, whereas SARS-CoV2 Gamma, Delta and BA.2 variants were propagated with Vero E6/TMPRSS2 (provided by A. Townsend). Naive Vero E6 or Vero E6/TMPRSS2 cells were plated overnight and infected with SARS-CoV2 at an MOI of 0.003. Cultures were harvested when visible cytopathic effects were observed 48–72 h later. The virus-containing supernatant was centrifuged to remove debris and then aliquoted and stored at −80 °C. As previously described, the viral titre was determined by plaque assay with Vero E6 or Vero E6/TMPRSS2, and plaque-forming units per ml were used to calculate the MOI.

2.1.9 EBV-Transformed B-Cell Lines

B cell lines (BCLs) and the human ACE2-transduced BCLs were generated as described previously[16]. The complementary DNA for the human ACE2 gene was cloned into a lentiviral vector, allowing the expression of enhanced green fluorescent protein and a puromycin resistance marker. The plasmids were transfected with packaging plasmids into HEK293-TLA cells. Lentiviral supernatant was collected and concentrated by ultracentrifugation. EBV-transformed BCLs were infected by ACE2 lentivirus at a multiplicity of infection of 0.1 with polybrene overnight, then washed and cultured for 3–5 days. ACE2-expressing B cells were stained using primary goat anti-human ACE2 antibody and donkey anti-goat AF647 secondary antibody, followed by cell sorting via flow cytometry. B cells with stable expression of ACE2 were maintained with 0.5 μg ml–1 of puromycin (Thermo Fisher Scientific). Mycoplasma testing was also carried out as described above every four weeks.

2.1.10 Intracellular Cytokine Staining (ICS) for SARS-CoV2 Infection

T-cells were cocultured with SARS-CoV2 Victoria 01/20, Delta or Omicron-infected BCLs-ACE2 at an E:T ratio of 1:2 for 6 h with GolgiPlug, GolgiStop, and PE-anti-CD107a (BD Biosciences, 1:20). Dead cells were labelled using Live/Dead Fixable Aqua dye (Invitrogen, 1:1000). After fixation with Cytofix/Cytoperm (BD Biosciences), the cells were stained with APC-anti-CD4 (Thermofisher, 1:33), PE-Cy7-anti-IFNγ (BD Biosciences, 1:33), APC-H7-anti-TNFα (Biolegend, 1:33) and BV421-anti-IL-2 (Biolegend, 1:33). Negative controls without virus infection were run for each sample, and a representative plot is shown in Figure B.2. All samples were acquired on Attune NxT Flow Cytometer (software v.3.2.1).

2.1.11 Identification of Optimal Epitopes using IFN-γ ELISPOT

Autologous B cell lines were loaded with $M_{172\text{-}188}$ truncated peptide (>70% purity; Thermo) at 10, 2, 0.4, 0.08 and 0.016 µg/ml concentrations for 1 hour. B cell lines without peptide treatment were used to determine the background reactivity. The peptide-loaded B cells were washed three times. Then 400,000 $M_{172\text{-}188}$-tetramer-specific $CD4^+$ T-cells for each restriction were co-cultured with 20,000 B cells for 8hrs. The ELISPOT was developed following the manufacturer's instructions.

2.1.12 Single-Cell Sequencing of CD4⁺ T-cells of Different Restrictions

Stimulated and unstimulated $M_{172\text{-}188}$ cells sorted with $M_{172\text{-}188}$-DRB1*01:01 and $M_{172\text{-}188}$-DRB1*07:01 tetramer, as well as cells identified by cytokine capture were single-cell sequenced using Smart-seq2[327] with the following modifications: reverse-transcription (RT) and PCR amplification were performed as described except using ISPCR primer with biotin tagged at the 5′ ends and increasing the number of cycles to 25. Sequencing libraries were prepared using the Nextera XT Library Preparation Kit (Illumina) and sequencing was performed on the Illumina NextSeq sequencing platform with NextSeq Control Software v.4.

2.2 Data Analysis

2.2.1 Statistical Analysis

ELISPOT counts were analysed in R v4.4.1[328] and plotted in ggplot2[329] with the statistical tests described in the figure or text. T-cell receptor data was visualised using circlize[330], pyrepseq[47] and igraph[331].

2.2.2 Prediction of HLA class II presentation of $M_{172\text{-}188}$

The online MixMHC2pred (v.2.0.2) server was used to make predictions for the sequence `TSRTLSYYKLGASQRVA` without context encoding for HLA-DR alleles `DRB1_01_01`, `DRB1_04_01`, `DRB1_07_01` and `DRB1_15_01`.

2.2.3 Single-Cell RNA Sequencing

Extracting Transcript Count Matrices

Raw scRNA-seq data were trimmed for adapter sequences using `cutadapt v2.13.0`[332], specifying the adapter sequence used for PCR as well as a minimum length of `20`. RNA transcripts were aligned to a human genome reference[333] using `STAR/2.7.8a`[334]. featureCounts, as part of `subread v2.0.6` collected counts for each gene[335].

Aligning T-cell Receptor Sequences

For the Smart-seq2 sequenced T-cells, TCR sequences were aligned to reference V-, D-, J- and C-gene segment databases and assembled into clones based on their Smart-seq2 well IDs using `MIXCR v4.7.0`[336]. For the short-term cell lines sequenced using 10x Genomics, TCR sequences were aligned and extracted using `cellranger-multi`. These were demultiplexed using antibody feature barcodes except for one sample, in which case vireo and cellSNP were used. Clone tables were cleaned and imported using Python 3.11.3 and `pyrepseq v1.5.1`[47].

When more than one sequence for a TCR chain was obtained from an individual cell, sequences that did not make up more than 95% of reads for that cell were excluded. This aimed to remove dual alpha or dual beta T-cells, for which we cannot determine which chain combination generates the TCR's specificity.

Comparison of TCR Repertoires

Paired $\alpha\beta$ clonotypes are defined by the unique combination of Vα gene, Vβ gene, CDR3α and CDR3β sequences, whereas single-chain indices were defined on the basis of CDR3 and V gene.

Diversity indices were calculated using either Shannon diversity (effective number of clonotypes) or Simpson diversity (coincidence probability), using the repertoire of a given sample. This means all cells sampled specific to a given epitope, from a given patient, at a given timepoint where a TCR sequence could be aligned and extracted. In order to minimise variance based on sampling number, all samples required at least 5 cells with an aligned TCR sequence to be included in diversity analyses. The formulas used for these indices are:

$$\text{Shannon} = \exp\left(-\sum_{i=1}^{n} p_i \log(p_i)\right)$$

$$\text{Simpson} = \frac{1}{\sum_{i=1}^{n} p_i^2}$$

Where p_i represents the frequency (proportion) of clonotype i within the repertoire, n is the total number of unique clonotypes in the repertoire, and the Shannon index is calculated as the exponential of Shannon entropy to give the effective number of species. The effective number of clonotypes (species) can be estimated as the exponent of the Shannon entropy.

Filtering of Low-Quality Cells

scRNA data was analysed using Seurat v4.3[337]. Quality control metrics such as RNA counts per cell and percentage mitochondrial gene expression were evaluated. Thresholds for quality control metrics are defined per experiment, and laid out in the relevant chapter. Features expressed in fewer than five cells were excluded.

Normalisation

Data were normalised using the `LogNormalize` method in Seurat with default parameters (normalisation scale factor 10,000). The top 3,000 most variable features were identified using the `vst` method in Seurat. These data were used for dimensionality reduction and clustering. Raw count matrices were used for differential gene expression.

Dimensionality Reduction and Clustering

Dimensionality reduction and clustering are described in detail in the relevant chapter, but principal component analysis was performed on scaled data and the first 50 principal components were visualised on an elbow plot. Visual inspection of the elbow plot determined the number of principal components to use for downstream analysis. The `FindNeighbors` function was used to construct a 20-nearest neighbour graph, and the UMAP and Louvain algorithms were applied to the graph for visualisation and clustering respectively.

The resolution parameter for clustering was determined by the stability of clusters across a range of resolutions, and then putative resolutions were selected based on whether marker genes reflected canonical $CD4^+$ T-cell subtypes, states or functions. Clusters were annotated using literature search for differentially expressed genes as well as inspecting the comparative expression of canonical markers, as listed in the 2023 PBMC atlas covering 901,152 $CD4^+$ T-cells[36].

When annotating T-cell clusters, genes are presented alongside their $\log_2$ fold change ($\log_2$FC) and adjusted p-value (p) when tested via a Wilcoxon rank-sum test against all other cells. This is done using `FindAllMarkers` in Seurat including negative markers, with a minimum difference percentage of `0.1` and a minimum $\log_2$FC of `0.25`.

Differential Gene Expression

Differential gene expression was performed by pseudobulking RNA count matrices using the Seurat[337] `AggregateExpression` function, which aggregates the counts of cells in a given group. TCR genes (genes beginning with `TRA`, `TRB`, `TRD`, `TRG`) were excluded from the expression matrix, as TCR features were analysed separately. The groupings are specified in the main text of each chapter. Expression was compared using DESeq2[338] using the Wald test and a design specified in the text. The DESeq α (target false-discovery rate) was set to 0.05, and this was also used as a cutoff for significance in the Bonferroni adjusted p-values.

To identify differentially enriched gene sets to interpret identified genes, we used EnrichR with the Gene Ontology Biological Process database[339].

3

HLA Restriction shapes the phenotype and functional profile of CD4$^+$ T-cells responsive to the same SARS-CoV2 antigenic peptide

Summary

The presentation of viral antigens by HLA class II molecules to CD4$^+$ T-cells generates diverse functional phenotypes that contribute to pathogen responses and immunological memory. Here, we investigate the immunodominance of SARS-CoV2 epitope $M_{172\text{-}188}$, to which over 60% of our cohort responded, regardless of HLA typing. We find that multiple MHC class II alleles present this epitope with shared fine peptide requirements. Building on this, we solve structures of the epitope presented by DRB1*01:01 and DRB1*07:01, and find remarkably similar core regions. Using patient-derived $M_{172\text{-}188}$-specific T-cell clones, we demonstrate that the functional peptide presentation level of the epitope by DRB1*01:01 exceeds that of DRB1*07:01 in a model of viral infection. DRB1*01:01-restricted CD4$^+$ T-cells exhibit enhanced responses to and superior ability to suppress live SARS-CoV2 compared with other DRB1 restrictions. Using the differential presentation of this epitope as a natural experiment, we carry out single-cell transcriptomic analysis of CD4$^+$ T-cells in

the recovery phase (~1 month) post-infection, revealing elevated expression of $CD4^+$ effector genes in DRB1*01:01-restricted cells and T helper 2 differentiation genes in DRB1*07:01-restricted cells. Upon re-stimulation with $M_{172\text{-}188}$, MHC restriction influences the response, with DRB1*01:01 maintaining greater expression of effector and Th1 genes. Four years after their initial infection, we resampled the same patients and did not find differences in gene modules that separated DRB1*01:01 and DRB1*07:01 specific cells in the recovery phase. These findings demonstrate that multiple HLA alleles can present $M_{172\text{-}188}$ and that the antigen load presented by different MHC class II molecules is associated with their functional profile and transcriptomes in the recovery phase. Promiscuous presentation of the epitope by multiple HLA-DR alleles and HLA-DP makes $M_{172\text{-}188}$ a strong candidate for non-Spike vaccines, as multiple presentation gives the epitope greater population coverage.

3.1 Introduction

CD4⁺ T-cells are critical to coordinating the cellular and humoral immune response to viral infections[26,268,340–342]. In the case of SARS-CoV2, the adaptive immune system mounts responses to combat the virus, including CD4⁺ and CD8⁺ T-cells that recognise diverse viral peptides presented by human leukocyte antigen (HLA) molecules[95,148,343].

Upon recognising viral epitopes presented by HLA class II molecules on antigen-presenting cells, naïve CD4⁺ T-cells receive multiple signals: their T-cell receptor (TCR), costimulatory or coinhibitory molecules, and the cytokine environment[313]. However, specific signals such as antigen load and TCR signalling strength and their consequences on CD4⁺ T-cell differentiation and memory formation are not fully understood.

The emergence of SARS-CoV2 as a novel pathogen with minimal pre-existing immunity provides a unique opportunity to investigate various fundamental immunology questions in humans. Cohorts with clear primary infection history, comparable sampling, and temporally similar viral strains, such as ours[16,95] and others[34,230,274], have significantly contributed to our understanding of the establishment and memory of T-cell responses, including in CD4⁺ T-cells.

Antigen load, or the quantity of peptide-MHC complexes presented to T-cells, has been shown to influence CD4⁺ T-cell differentiation and memory formation in vitro[312,344,345] and mouse models of recombinant LCMV[313] or *L. monocytogenes* infection[311,314,346]. The effects of antigen load from these studies were unclear. Different studies found that the highest antigen loads or TCR signalling strengths were associated with Th1[313–316,347,348], Tfh[346,349,350] or Th2[312,345] differentiation. Diverse models, definitions of antigen load versus TCR signalling strength and other variables might drive different results, and fundamental differences between mice and humans may confound the interpretation of this in human immunity.

We and others previously identified residues 172-188 of the Membrane protein ($M_{172\text{-}188}$, `TSRTLSYYKLGASQRVA`) as an immunodominant CD4⁺ T-cell epitope. This epitope is conserved in the Delta and Omicron strains of SARS-CoV2[351], and is shared with SARS-CoV,

RaTG13 and Pangolin coronavirus (UniProt: A0A6B9WDP7)[352]. In convalescent SARS-CoV2-infected individuals at 30 days, 63.6% (28/44) generated responses to $M_{172\text{-}188}$ in IFN-γ ELISPOT assays[95]. This was the most immunodominant epitope in our study. An overlapping epitope, $M_{176\text{-}190}$, elicited responses in 95% (21/22) of individuals tested with the same assay in a different population[353]. Intrigued by this epitope's immunodominance, we undertook detailed characterisation of the epitope's presentation.

In this study, we use our deeply characterised, longitudinal cohort to profile the immune memory to SARS-CoV2 $M_{172\text{-}188}$, identifying multiple HLA alleles capable of presenting the epitope. Further, we assess the structural and functional differences between presentation of the epitope by different alleles and use this as a natural experiment to compare the functional responses and transcriptomes of memory $CD4^+$ T-cells specific to these epitopes but presented by different class II MHC molecules, finding implications for T-cell function and short term differentiation.

3.2 Results

3.2.1 SARS-CoV2 $M_{172\text{-}188}$ elicits a dominant $CD4^+$ T-cell response across a cohort with diverse HLA class II types

Of the 44 individuals we deeply profiled after their SARS-CoV2 infection, one was asymptomatic, 27 were mild, 13 were severe, and 3 were critical. As shown by the ex-vivo interferon (IFN)-γ ELISPOT assays using convalescent samples 1–3 months post-infection, 52% (14/27) of mild, 77% (10/13) of severe cases and all critical cases, as well as the asymptomatic individual, responded to $M_{172\text{-}188}$ (Figure 3.1). Overall, 63.64% (28/44) cases responded to $M_{172\text{-}188}$.

This epitope is also highly conserved. In the COG-UK database of 3,079,046 UK SARS-CoV2 genomes[354] sequenced between April 2020[355] and December 2024 the cumulative frequency of missense mutations to the $M_{172\text{-}188}$ epitope is 0.11%, and when including 5 flanking amino acids on either side the cumulative frequency is 0.14% (Figure 3.2).

The epitope's mutation rate represents 73% fewer mutations than would be expected based on the average mutation frequency across the Membrane protein (excluding major variant

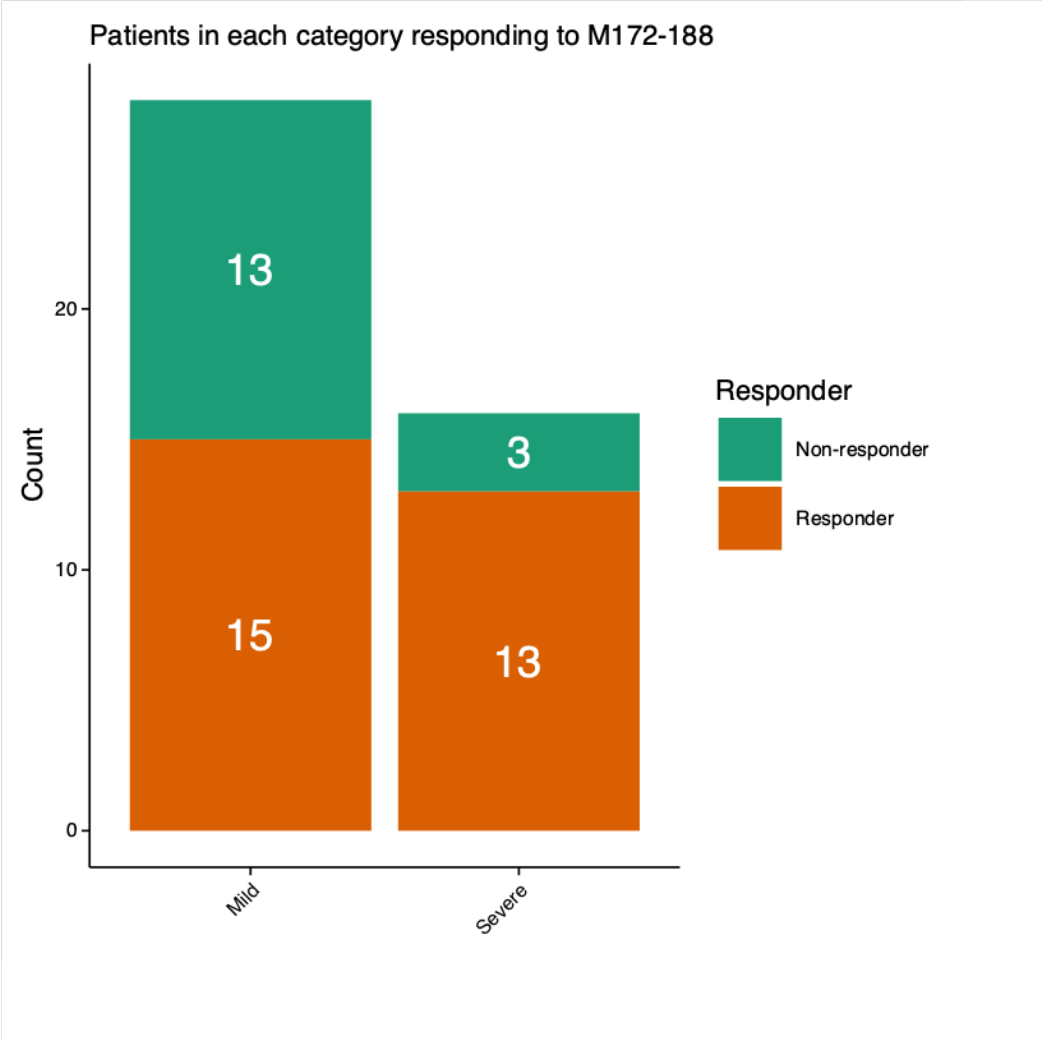


Figure 3.1: Bar Chart showing the distribution of $M_{172\text{-}188}$ responders and non-responders by clinical severity. Asymptomatic and mild cases were aggregated into mild, whereas severe and critical were aggregated into severe. Original ELISPOT data sourced from[95], analysis and visualisation carried out by me.

defining mutations), putting $M_{172\text{-}188}$ in the 32nd percentile of least mutated 17-residue windows (Figure 3.3). This, as well as its conservation with the closest known relatives of SARS-CoV2, despite its immunodominance, suggests that this epitope is under functional constraint.

By comparing the distribution of these HLA-DR alleles with our clinical data, we could dissect how individuals that carry specific HLA-DRB1 alleles were distributed between IFN-γ responders to $M_{172\text{-}188}$ and their clinical severity. Amongst DRB1*01:01 responders,

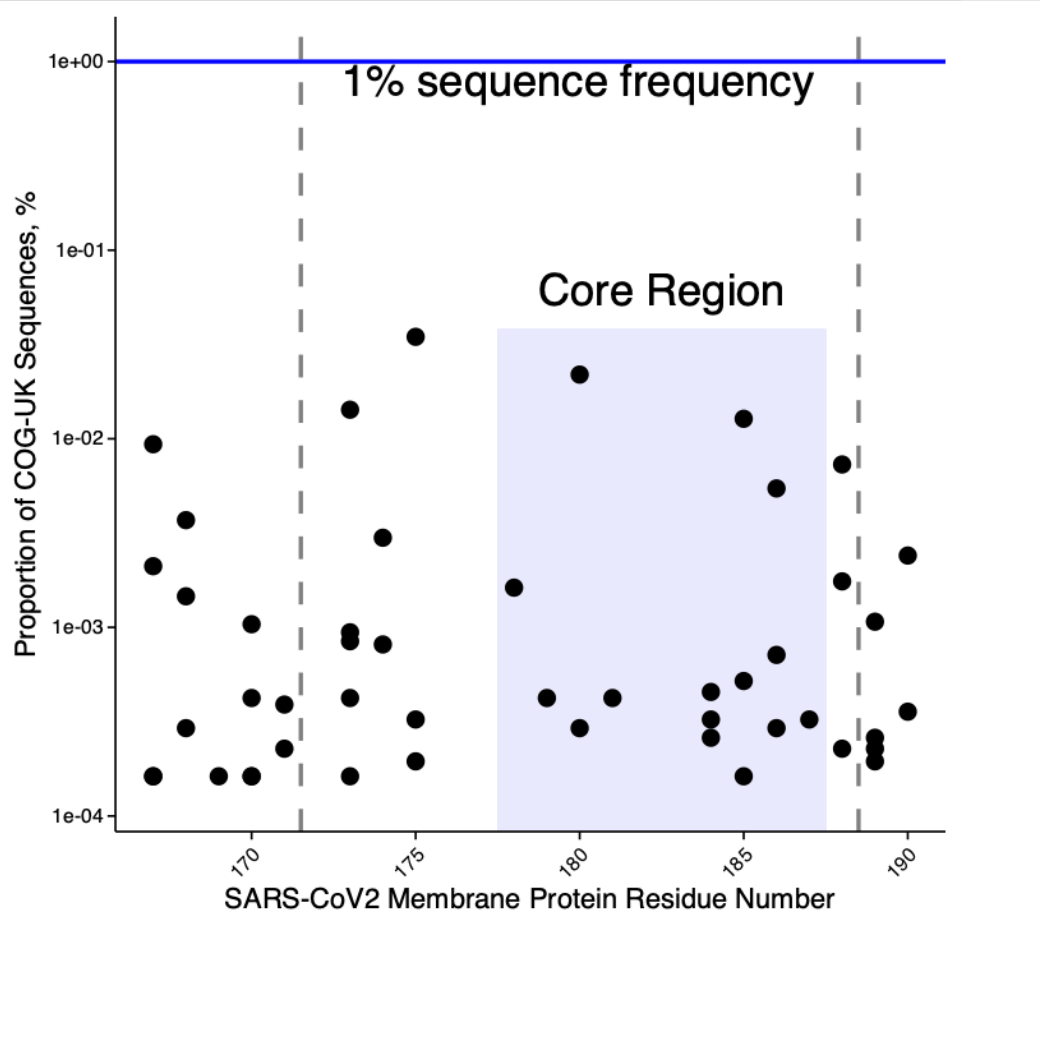


Figure 3.2: Scatter plot of SARS-CoV2 Membrane Protein Amino Acid residue number against COG-UK mutation frequency. Each dot is a particular amino acid mutation, e.g. (S187G). Original data sourced from COG-UK[354], data visualised by me.

six out of nine (66.7%) were severe or critical cases, whereas for DRB1*04:01 (40%), *07:01 (33.3%), and *15:01 (45.5%) severe responders made up the minority (Figure 3.4). Consistent with the finding that severe SARS-CoV2 is associated with broader antigen-specific responses[95], severe-non-responders to $M_{172\text{-}188}$ were rare among the surveyed HLA-types. Only in DRB1*15:01 did we observe severe patients who in convalescence did not respond to $M_{172\text{-}188}$.

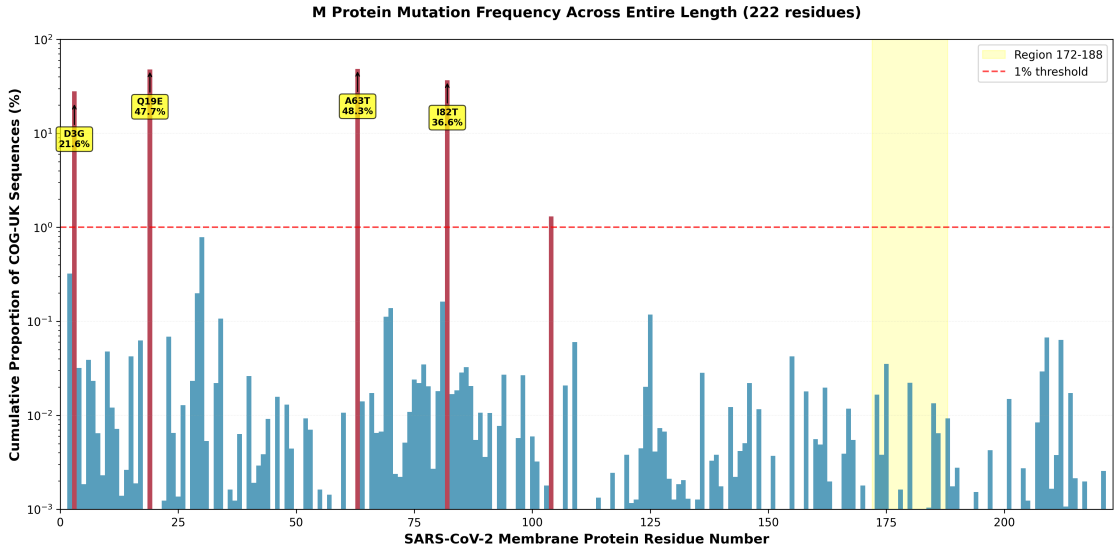


Figure 3.3: SARS-CoV2 Membrane protein mutation frequencies across all residues in COG-UK database. Bar plot shows cumulative mutation frequencies at every position of the M protein. Four major hotspots are annotated, which represent mutations that are fixed in major lineages. The $M_{172\text{-}188}$ epitope is highlighted in yellow, illustrating its low mutation frequency compared to the rest of the protein.

3.2.2 Multiple HLA-DR alleles and HLA-DP present SARS-CoV2 $M_{172\text{-}188}$

Responders to $M_{172\text{-}188}$ had heterogeneous HLA genetic makeup (Figure 3.5). Compared to the cohort as a whole, HLA-DRB1*15:01 and 01:01 HLA-types were more frequent amongst responders than non-responders. We hypothesised this would be compatible with multiple HLA class II molecules presenting the $M_{172\text{-}188}$ peptide.

To identify the specific HLA-DR alleles presenting $M_{172\text{-}188}$, my colleagues first generated responsive cell lines through intracellular cytokine staining for IFN-γ, TNF-α or IL-2. They then used IFN-γ ELISPOT assays with partially mismatched B cell lines (BCLs) to stimulate CD4⁺ clones specific to $M_{172\text{-}188}$ (Figure 3.6). Since each set of B-cell lines stems from a heterozygous donor with known HLA-DR alleles, through a process of elimination we can assign the restricting HLA type for each clone. This demonstrates that clones exist that are responsive to $M_{172\text{-}188}$ restricted to HLA-DRB1*01:01, DRB1*07:01 and DRB1*15:01.

They also stained bulk $M_{172\text{-}188}$ responsive CD4⁺ T-cells with different MHC-II tetramers, identifying T-cells that stain for DRB1*01:01, DRB1*07:01 and DRB1*04:01 (Figure 3.7).

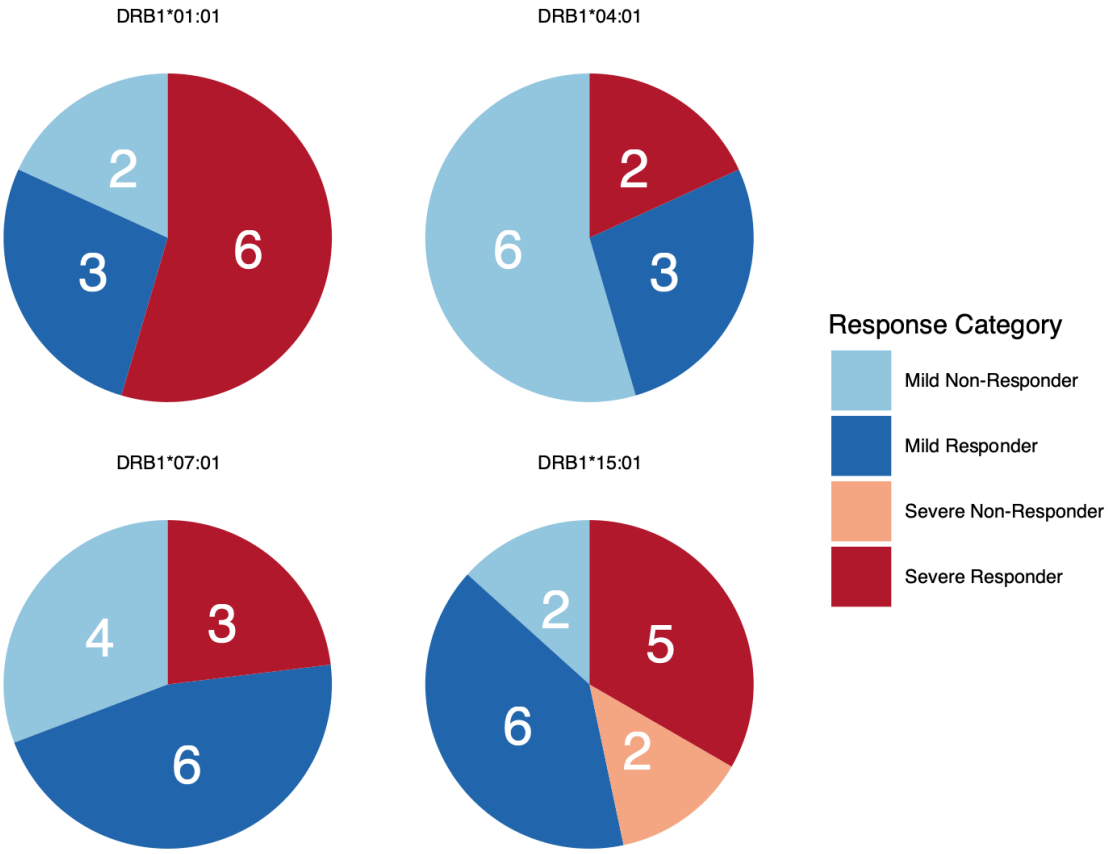


Figure 3.4: Each Pie Chart represents the set of individuals that carry that specific HLA allele (individuals can be present in more than one due to heterozygosity), showing the distribution of clinical severity and whether individuals responded to $M_{172\text{-}188}$ by IFN-γ ELISPOT. Original ELISPOT data sourced from[95], analysis and visualisation carried out by me.

Their results support that DRB1*01:01, DRB1*04:01, DRB1*07:01 can all present this epitope to $CD4^+$ T-cells.

$M_{172\text{-}188}$ fits the criteria for an HLA-DR super motif[356], and when I used MixMHC2pred[357] to predict the likelihood of presentation by various HLA-DRB1 alleles, $M_{172\text{-}188}$ gets <20% rankings for presentation by DRB1*01:01 (13.1%), DRB1*07:01 (19.1%) and DRB1*15:01 (15.2%) but not for DRB1*04:01 (60.0%). The best ranking was predicted for DRB1*01:01, but DRB1*15:01 was predicted to have a different core-region (`LSYYKLGASQ`) to the other DR alleles (`YKLGASQRVA`).

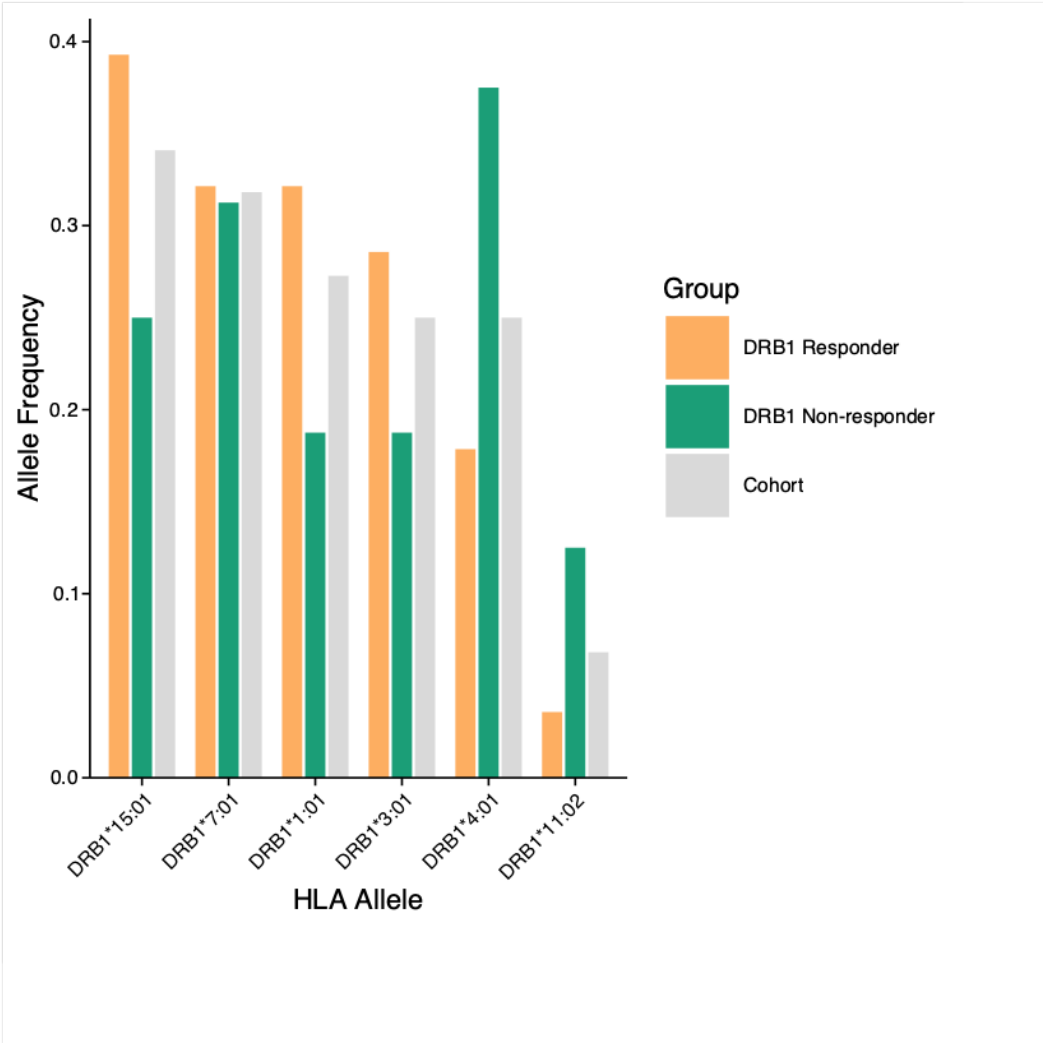


Figure 3.5: Proportions of cohort (n=44) responders and non-responders that carry a particular HLA-DRB1 allele. Orange bars indicate the proportion of HLA-DRB1 allele carriers that respond to $M_{172\text{-}188}$, whereas green indicates the proportion of carriers of that allele that do not respond to $M_{172\text{-}188}$. Grey bars indicate the proportion of the cohort as a whole that carry the indicated HLA-DRB1 allele. Original ELISPOT and HLA-typing data sourced from[95], analysis and visualisation carried out by me.

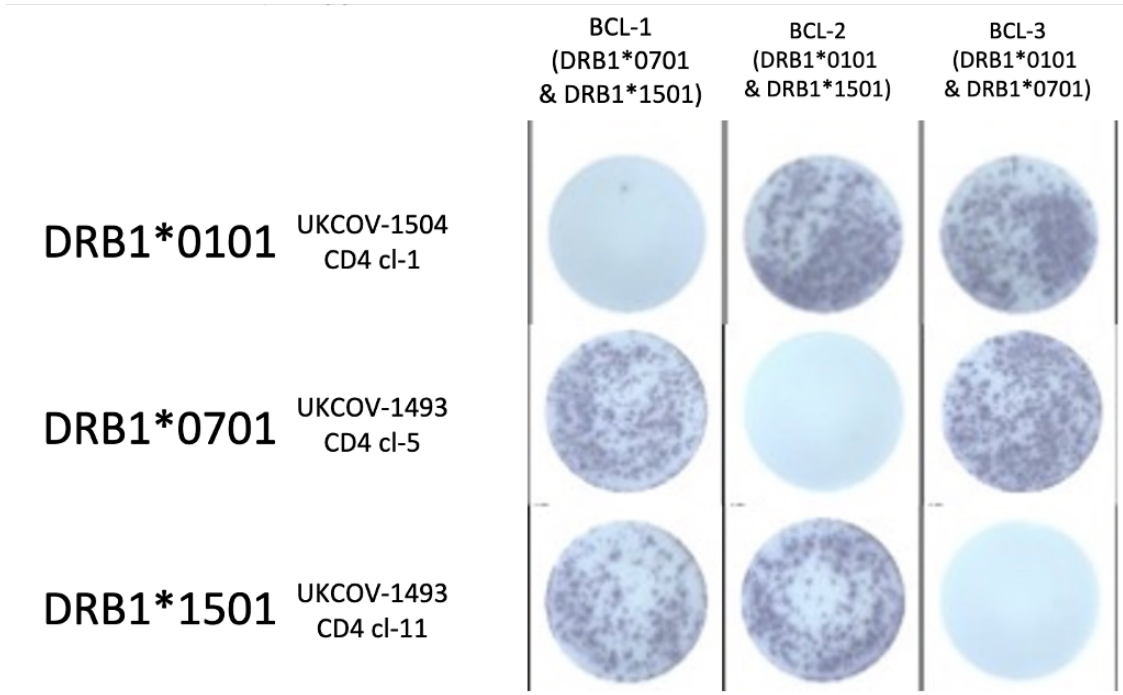


Figure 3.6: Using mismatched B-cell lines with known HLA-types as antigen-presenting cells for various $M_{172\text{-}188}$ responsive $CD4^+$ T-cell clones, we could identify clones restricted to $M_{172\text{-}188}$ presented by HLA-DRB1*01:01, *07:01 and *15:01. Work carried out by Dr Xuan Yao, figure generated by Dr Xuan Yao.

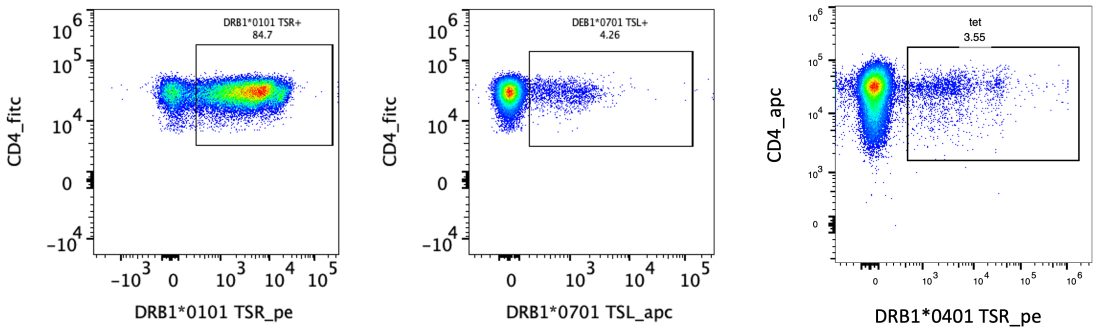


Figure 3.7: Positive, tetramer staining populations exist within $M_{172\text{-}188}$ responsive T-cells for tetramers of HLA-DRB1*01:01, DRB1*07:01 and DRB1*04:01. The first two panels are from the same donor, the final panel is from a different donor. Work carried out by Dr Xuan Yao, figure generated by Dr Xuan Yao.

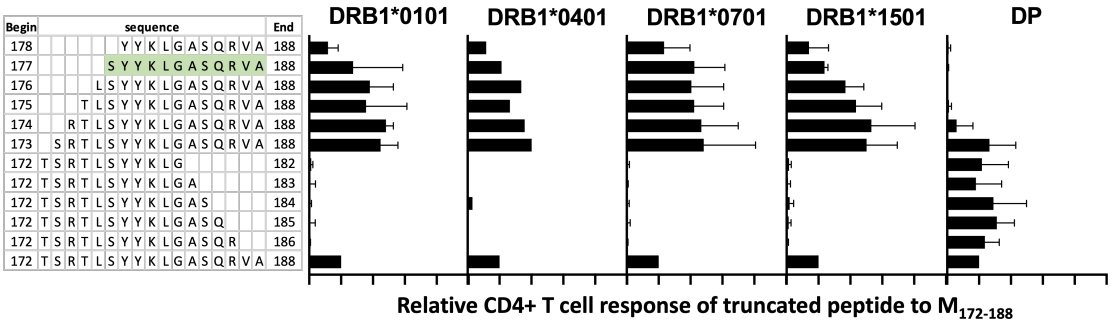


Figure 3.8: Relative IFN-γ response (compared to full $M_{172\text{-}188}$ epitope) by T-cell clones of known restriction to various truncations of $M_{172\text{-}188}$ epitope. Work carried out by Dr Xuan Yao, figure generated by Dr Xuan Yao.

3.2.3 HLA DR-presented epitopes share a similar core region

To understand if the different HLA restrictions of $M_{172\text{-}188}$-specific CD4 T-cells alter the fine peptide requirements of the T-cell response, we wanted to identify the required core region in $M_{172\text{-}188}$ that is recognised by T-cell receptors. My colleagues generated 11 different truncated peptides of the $M_{172\text{-}188}$ epitope and used autologous BCLs to present these peptides to $M_{172\text{-}188}$ responsive T-cells with known restriction. Similar patterns of peptide recognition are observed by all HLA-DR-restricted T-cells. Peptides with C-terminal truncation (TSR_C_11, 12, 13, 14 and 15) did not elicit responses by HLA-DRB1*01:01, 04:01, 07:01 or 15:01 restricted T-cells. The Alanine on the C terminus of the peptide is indispensable for the DRB1*01:01 T-cell response. In contrast, Threonine on the N terminus is necessary to elicit a DP-restricted response (Figure 3.8).

Collaborators Dr. Dene Littler, Dr. Jan Petersen and Prof. Jamie Rossjohn solved the structures of HLA-DRA*01:01:DRB1*01:01 and HLA-DRA*01:01:DRB1*07:01 to visualise the peptide-MHC conformation (Figure 3.9). This supports the shared fine peptide requirements, suggesting the three dimensional arrangement of the epitope is similar when presented by DRB1*01:01 or DRB1*07:01. This confirms the shared core region we identified through T-cell assays with truncated peptides. The conformations of DRB1*01:01 and DRB1*07:01 are similar, but polymorphisms at the base of the binding cleft might explain the wider binding cleft in DRB1*07:01.

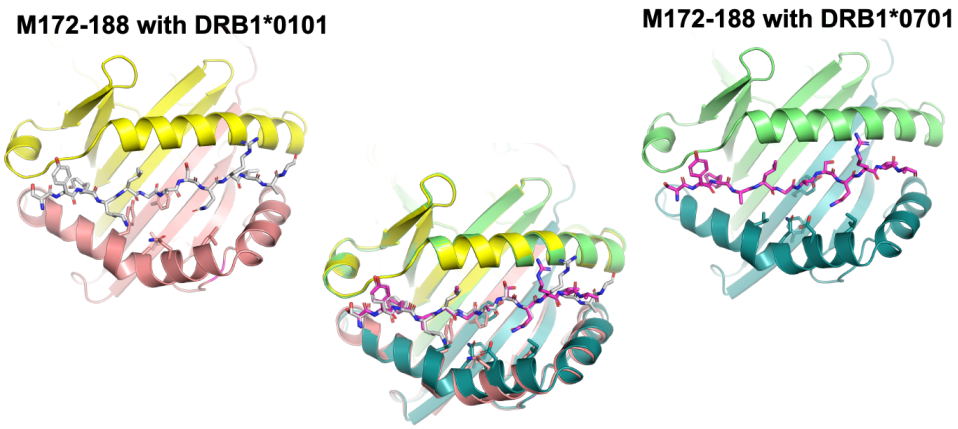


Figure 3.9: Structures of $M_{172\text{-}188}$ presented by HLA-DRB1*01:01 (left) and HLA-DRB1*07:01 (right), solved by Prof. Jamie Rossjohn's group. Middle panel represents the two structures overlaid, showing extensive overlap of the $M_{172\text{-}188}$ conformation across presenting HLA alleles.

3.2.4 $M_{172\text{-}188}$ responsive $CD4^+$ clones' responses to live SARS-CoV2 virus co-culture associate with restricting HLA-allele.

To examine whether the memory $M_{172\text{-}188}$-specific CD4 T-cells established after natural infection could respond to SARS-CoV2 infection, my colleagues assessed the antiviral efficacy and cytokine production of the T-cell clones restricted to DRB1*01:01, DRB1*07:01 or DRB1*15:01 in our in vitro SARS-CoV2 co-culture assays[16]. To test whether $M_{172\text{-}188}$-specific T-cells are robust to SARS-CoV2 variants of concern, they carried out live viral co-culture assays with $M_{172\text{-}188}$ responsive $CD4^+$ T-cell lines tetramer-sorted for each DRB1 allele we identified (Figure 3.10). Cytokine production by T-cells was assessed by intracellular cytokine staining. DRB1*01:01, DRB1*07:01, and DRB1*15:01 were robust against both the Victoria and Delta strains. The DRB1*04:01 responses were much less potent.

Then, we wanted to examine the quantitative features of T-cells with different restrictions, so my colleagues measured both viral copy number suppression and IFN-γ positivity by $M_{172\text{-}188}$ responsive $CD4^+$ T-cells with varying restrictions in live viral co-culture assays. DRB1*01:01 or DRB1*07:01-restricted T-cells reduced the viral copy number in the supernatant more than DRB1*15:01 after 48 hours of co-culture with the SARS-CoV2-infected ACE2 BCLs (Figure 3.11a). Consistent with the suppression result, DRB1*01:01-restricted clones produced the most IFN-γ when co-cultured with SARS-CoV2-infected ACE2 BCLs

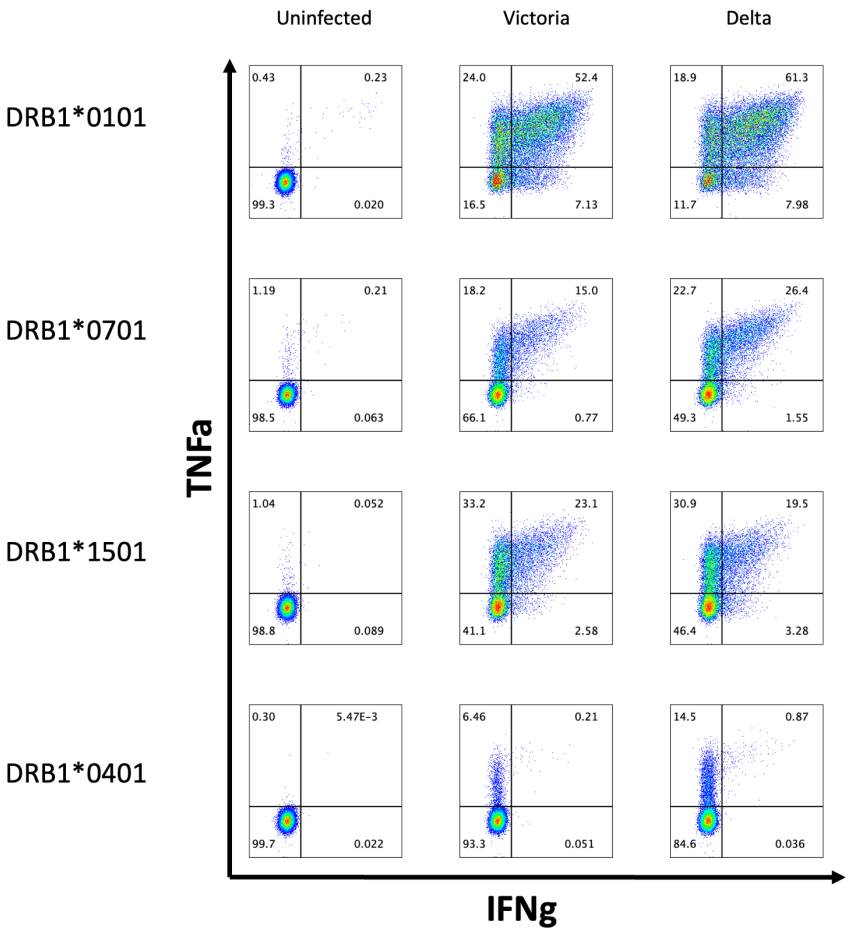


Figure 3.10: Flow cytometry plots of bulk $M_{172\text{-}188}$ responsive T-cells under Intracellular Cytokine Staining for IFN-γ and TNF-α with different SARS-CoV2 Variants of Concern. For these experiments, B-cell lines were infected with live SARS-CoV2. Work carried out by Dr Xuan Yao. Figure generated by Dr Xuan Yao.

(Figure 3.11b).

3.2.5 Different HLA-DR alleles present different interpolated peptide concentrations of $M_{172\text{-}188}$, with DRB1*01:01 presenting the highest.

Given the similarity of the epitope's presented structure and required core regions between the DRB1 alleles, we hypothesised that different antigen-presentation abundance might underlie the observed differences in functional responses to live virus.

Estimating the amount of antigen-presentation in live-viral infection is challenging, as polyclonal T-cell responses may have different functional avidities and comparisons between individuals are challenging owing to heterogeneity of antigen-presenting cells. In order to

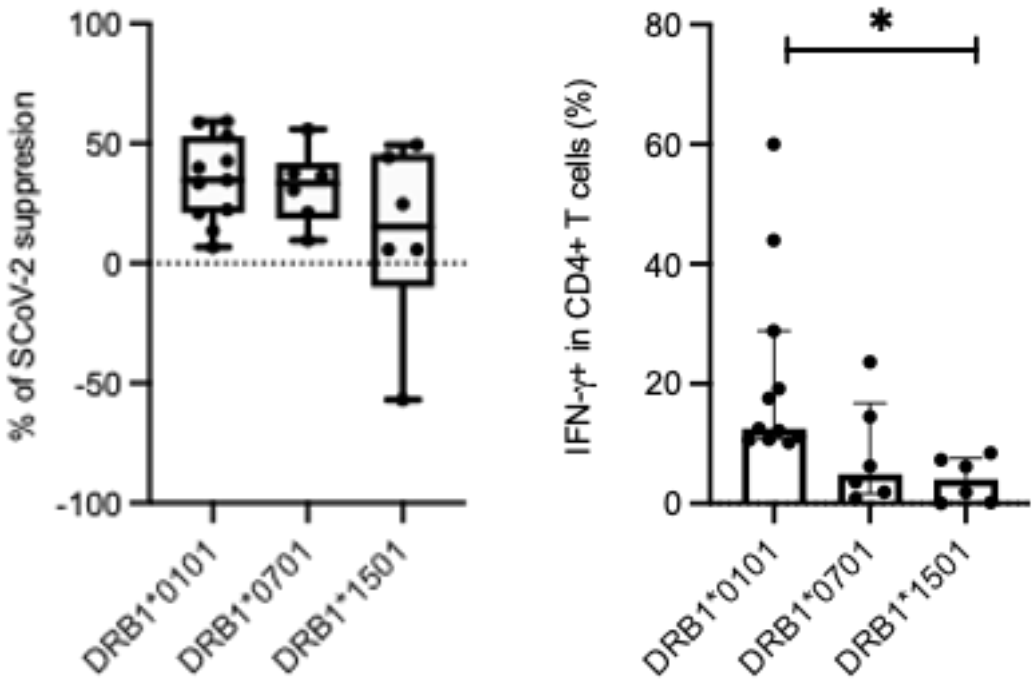


Figure 3.11: (a) Percentage viral RNA suppression by $CD4^+$ T-cell clones with different HLA-DRB1 restrictions in the live-viral co-culture assay. Suppression calculated as the percentage difference between control (no T-cells) and T-cell clone condition. (b) Percentage IFN-γ by ICS of $CD4^+$ T-cell clones with different restricting HLA-DRB1 alleles in the live-viral co-culture assay. Work carried out by Dr Xuan Yao, figures generated by Dr Xuan Yao.

overcome these challenges, we use a T-cell response based method to interpolate the equivalent pulsed antigen-concentration of live-viral infection. In brief, my colleagues generated individual $M_{172\text{-}188}$ pulsing concentration vs. IFN-γ^+ curves for T-cell lines (Section 2.1.2) of known restriction (Figure 3.12). The curves are shown in Figure C.2. As B-cell lines for antigen presentation were chosen to be heterozygous at HLA-DRB1, the degree of activation in live viral co-culture could be compared between two different HLA-DRB1 restrictions within the same antigen-presenting cell.

I then fit log-logistic curves to the pulsed concentration and degree of IFN-γ of the T-cell lines, and then queried the curves with the observed IFN-γ^+ frequency in live viral coculture. Since the B-cells naturally process and present the epitope, this represents an estimate of the degree of antigen processing and presentation during viral infection of that specific HLA allele to antigen-specific T-cell lines. The interpolated antigen-concentration can be directly compared between pairs of HLA-DRB1 alleles, allowing a relative ranking to be

established. One strength of this method is that since each curve is generated for each donor's antigen-specific T-cell lines individually, confounders such as the functional avidity, donor or severity of each clone are inherently controlled.

This enables these comparisons across T-cell lines and epitopes or restrictions because each T-cell line's sensitivity to antigen is represented in its own dose-response curve. For example, if in response to SARS-CoV2 co-culture 65% of T-cells are IFN-γ^+, and the T-cell line's dose-response curve indicates that 65% of T-cells are IFN-γ^+ at a pulsed concentration of 0.3 µg/ml, then the interpolated effective concentration of $M_{172\text{-}188}$ in the live viral co-culture is 0.3 µg/ml. For another epitope and T-cell line, but the same antigen-presenting cells, the degree of T-cell activation might be 43%. This could have an equivalent pulsed concentration of 0.4 µg/ml, and therefore the interpolated effective concentration of $M_{172\text{-}188}$ in the live viral co-culture is 0.4 µg/ml. As a result, this approach can quantitatively compare the effective concentration of different epitopes or restrictions in live viral co-culture assays, and therefore the antigen load presented by different HLA-DR alleles.

The left panel of Figure 3.13 compares, in DRB1*01:01 & *07:01 expressing B-cells the interpolated $M_{172\text{-}188}$ concentration of DRB1*01:01 and DRB1*07:01 restricted T-cell lines. In this case, both DRB1*01:01 lines have higher interpolated $M_{172\text{-}188}$ concentrations than the DRB1*07:01 lines. This means the DRB1*01:01 T-cell lines are activated to a degree consistent with higher $M_{172\text{-}188}$ pulsing concentrations than the DRB1*07:01 lines in live viral co-culture on a single set of transformed B-cells. Since the same B-cell line is used in both cases, this suggests that the DRB1*01:01 B-cells present $M_{172\text{-}188}$ more effectively than the DRB1*07:01 B-cells in natural infection on this cell line.

When comparing the different antigen loading within individual experiments, DRB1*01:01 had greater antigen load than DRB1*07:01. In the live viral co-culture assays DRB1*01:01 and DRB1*07:01 had higher suppression and IFN-γ positivity, followed by DRB1*15:01 and then DRB1*04:01.

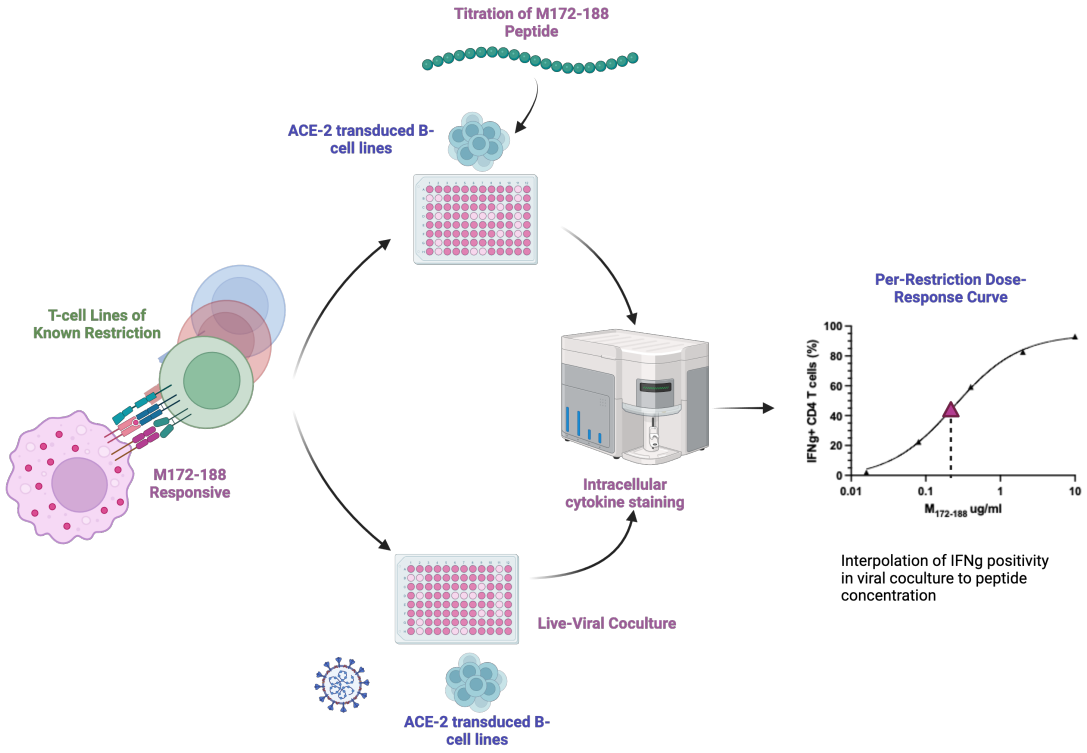


Figure 3.12: Cartoon of experimental design for interpolation of effective pulsed concentration of live SARS-CoV2 infection. ACE2 expressing B-cell lines are pulsed with $M_{172\text{-}188}$ peptide at varying concentrations or infected with SARS-CoV2. IFN-γ positivity for the peptide titration is used to generate a three-parameter log-logistic curve, against which the IFN-γ positivity in the viral co-culture condition can be used to interpolate the effective pulsed concentration of the antigen. Experimental data generated by Dr Xuan Yao, figure and curve fitted by me.

3.2.6 Single-cell sorting of HLA-DR restricted $M_{172\text{-}188}$ response in the Recovery Phase.

I characterised the sorted cells' single-cell transcriptome and TCR repertoire to explore how viral neutralisation and antigen load differences interact with the cell's phenotypes and differentiation. To do this, my colleagues sorted $M_{172\text{-}188}$-DRA1*01:01/DRB1*01:01 or $M_{175\text{-}188}$-DRA1*01:01/DRB1*07:01-specific T-cells at a single-cell level with tetramers from four PBMC samples with or without 4-hr stimulation of $M_{172\text{-}188}$. Our samples came from individuals with $M_{172\text{-}188}$ response restricted to DRB1*01:01 (1343, severe, age 48), one to DRB1*07:01 (1493, mild, age 74), and two to both DRB1*01:01 and 0701 (1180, severe, age 43 and 1525, severe, age 73). Donors for which single-cell sequencing was carried out are summarised in Table 3.1. Cell counts isolated from these individuals per tetramer, as well

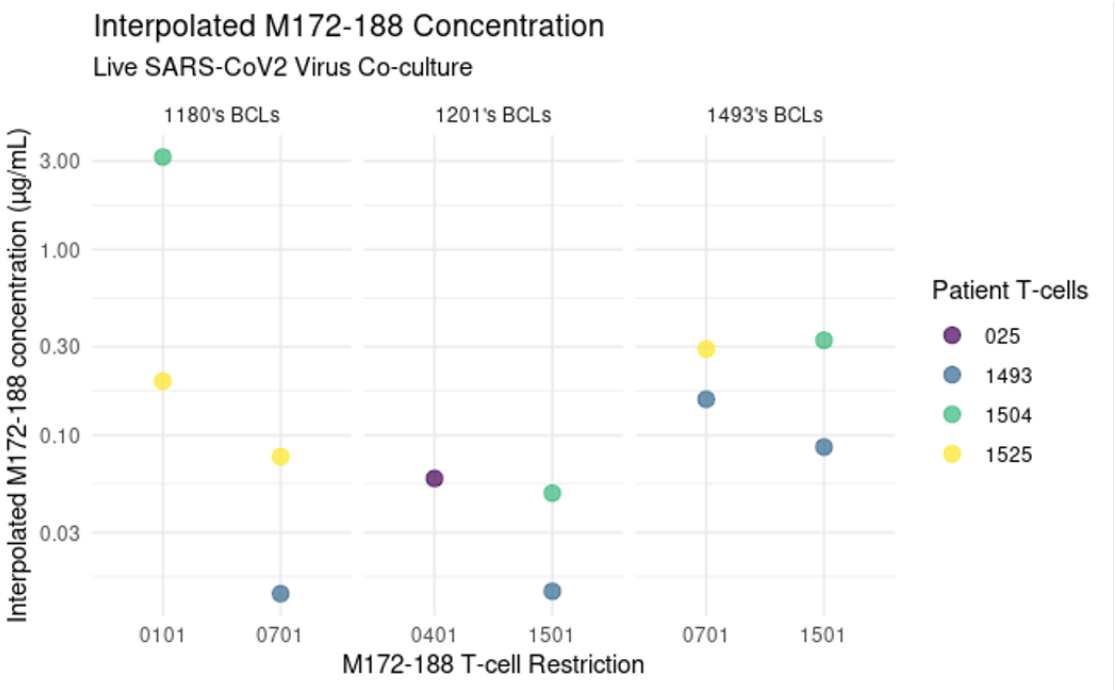


Figure 3.13: Dot plot of Interpolated $M_{172\text{-}188}$ concentration of Live Viral Co-Culture. ACE2 transduced B-cells from donors with known HLA alleles that present $M_{172\text{-}188}$ (1180: DRB1*01:01 & DRB1*07:01, 1201: DRB1*04:01 & DRB1*15:01, 1493: DRB1*07:01 & DRB1*15:01) were pulsed with $M_{172\text{-}188}$ peptide at varying concentrations and a three-parameter log-logistic curve was fitted to each cell line with known restriction. Each dot reflects a particular $M_{172\text{-}188}$ responsive cell line with known restriction from a donor, indicated by their colour. Experimental data generated by Dr Xuan Yao, interpolation and generation of this figure was carried out by me.

as those sorted at the four-year timepoint are shown in Figure C.3.

Table 3.1: Clinical metadata of the $M_{172\text{-}188}$ single-cell cohort. SARS-CoV2 Severity reflects the severity of the initial SARS-CoV2 infection. Age is the age (years) at the first SARS-CoV2 infection. Days between Symptoms and Sampling reflect the difference in self-reported initial symptom date and the date of sampling. Recovery phase indicates whether a sample was used for single-cell sequencing at the early (37-64 days post-infection) timepoint, and the Four Year sample indicates whether single-cell sequencing was carried out on $M_{172\text{-}188}$ responsive cells four years after the initial infection.

Donor	SARS-CoV2 Severity	Age	Sex	DRB1-1	DRB1-2	Recovery Phase Sample	Days between Symptoms and Sampling	Four Year Sample
0198	Severe	42	Male	01:01	03:01	No		Yes
1180	Severe	43	Male	01:01	07:01	Yes	43	Yes
1343	Severe	48	Male	01:01	04:01	Yes	61	No

Donor	SARS-CoV2 Severity	Age	Sex	DRB1-1	DRB1-2	Recovery Phase Sample	Days between Symptoms and Sampling	Four Year Sample
1493	Mild	74	Female	07:01	15:01	Yes	64	Yes
1504	Severe	60	Male	01:01	15:01	No		Yes
1525	Severe	73	Male	01:01	07:01	Yes	37	Yes

After index sorting these cells based on their tetramer staining, we performed single-cell sequencing using the Smart-seq2 protocol. As part of the index sorting, 36 of 45 plates (733/983) had cells stained for surface markers that allowed categorisation of cells into canonical $CD4^+$ memory subsets.

As shown in Figure 3.14, cell proportions by sample were highly variable. Central Memory cells dominated Recovery Phase samples, with a mean proportion of 58.4% (n=3, std. $\pm$50.8%) and 55.5% (n=3, std. $\pm$48.1%) in DRB1*01:01 and DRB1*07:01 specific cells respectively. The two donors that were sampled for $M_{172\text{-}188}$ presented by one allele at a time had their samples dominated by effector memory (1343, 1493), and were sampled later after their test (61 and 64 days, respectively) than the other donors (1525: 37 days, 1180: 43 days). The later collection of these samples may have allowed for a greater degree of central memory cell migration to the lymph nodes, leading to fewer present in the circulation.

3.2.7 Single-cell transcriptomic characterisation reveals differential gene expression between DRB1*01:01 and DRB1*07:01 restricted $M_{172\text{-}188}$-specific T-cells

To investigate potential differences in gene expression between T-cells specific for $M_{172\text{-}188}$ presented on DRB1*01:01 and DRB1*07:01, we aggregated post-QC (Section C.4) RNA count matrices by the patient of origin, tetramer sorting and stimulation condition and used DESeq2 to identify differentially expressed genes in the pseudobulks. Our design formula accounted for the patient of origin and a combined treatment factor of stimulation (stimulated or ex-vivo) and tetramer sorting (DRB1*01:01 and DRB1*07:01).

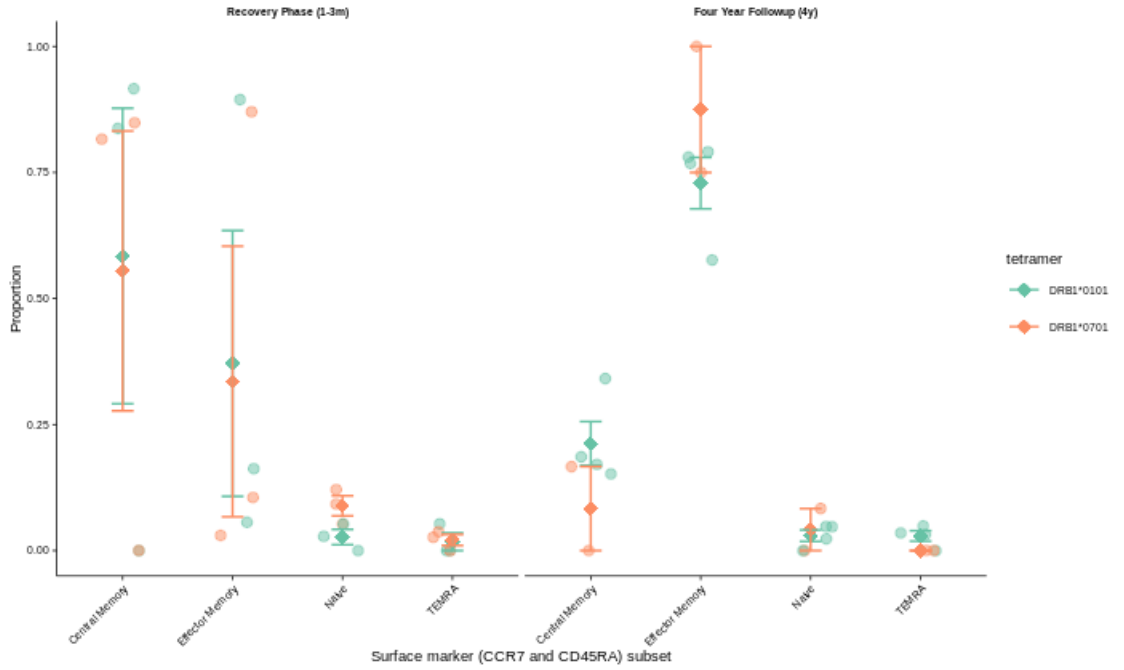


Figure 3.14: Proportion of Tetramer Sorted T-cells in samples that were stained with CCR7 and CD45RA via flow-cytometry during index-sorting that classify as Central Memory, Effector Memory, Naive or Terminal Effector Memory (TEMRA). Stratified by tetramer they sort to. Dots represent measurements, diamonds represent the mean. Ranges show the standard deviation of the proportions within that category. Flow Cytometry carried out by Dr Xuan Yao, analysis and plotting carried out by me.

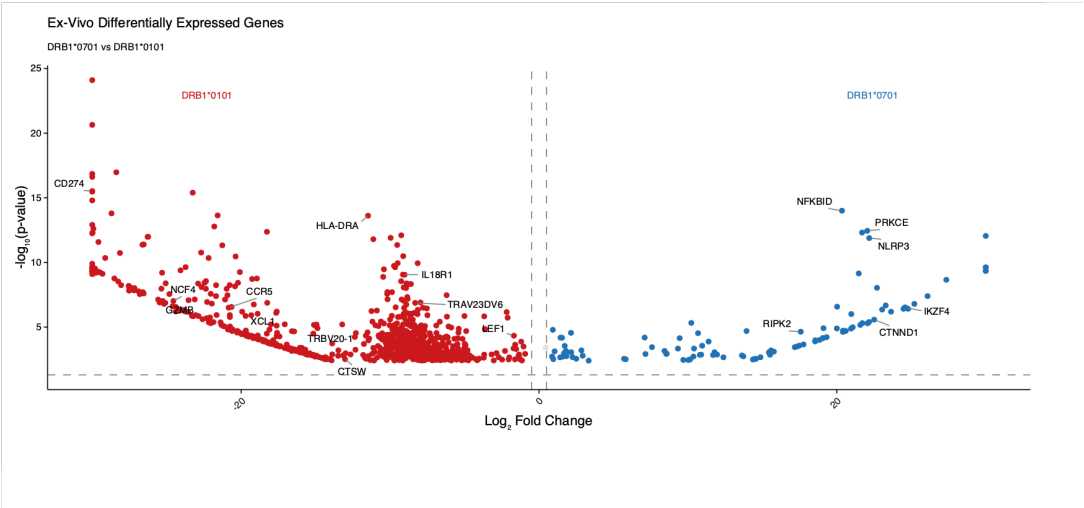


Figure 3.15: A volcano plot summarising differentially expressed genes identified via DESeq2 over pseudobulked samples between DRB1*01:01 specific CD4⁺ T-cells (greater expression on the left), and DRB1*07:01 specific CD4⁺ T-cells (greater expression on the right). DESeq2 was carried out using the Wald test and a design formula integrating the patient of origin, and a combination term of tetramer and stimulation.

Differential gene expression analysis (Figure 3.15) between DRB1*01:01 and DRB1*07:01-restricted $M_{172\text{-}188}$-specific CD4$^+$ T-cells revealed distinct transcriptomic profiles. DRB1*01:01 T-cells had higher expression of effector molecules *GZMB*, *XCL1*, *CCR5* and *CCR6*, *IL18R1*, as well as inhibitory molecules or ligands *CD274*, *PDCD1* and *HAVCR2*. *XCL1*, *IL18R1*, *CCR5* and *CCR6* play roles in promoting Th1 responses, Th1 differentiation and IFN-γ production. DRB1*01:01-sorted T-cells expressed more *NFATC1* and *CD83*, both coordinators of the T-cell response expressed after T-cell receptor activation, potentially supporting more potent activation in cells sorted to this tetramer. DRB1*01:01 restricted cells also expressed transcripts associated with active proliferation (*CDK1*) and metabolic reprogramming (*PDK2*) at greater levels than DRB1*07:01 restricted cells.

DRB1*07:01-restricted cells however differentially expressed *NLRP3*, *PRKCE*, *IKZF4* and *NFKBID*, genes involved in IL-1β processing, T-cell signal transduction and Treg differentiation. Although more difficult to functionally interpret, many metabolic genes were differentially expressed.

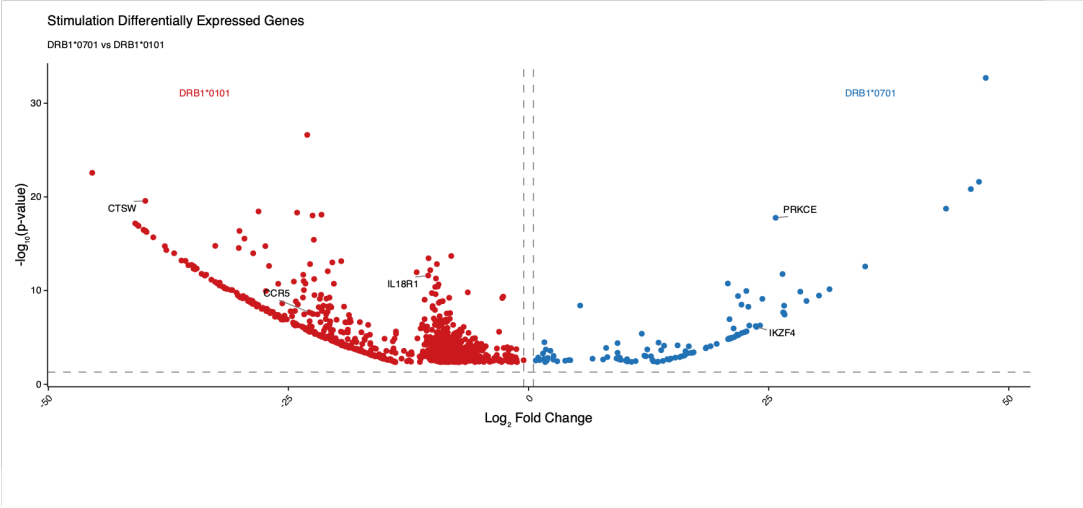


Figure 3.16: A volcano plot showing the DESeq2 differentially expressed genes over pseudobulked samples between DRB1*01:01 specific CD4$^+$ T-cells (greater expression on the left), and DRB1*07:01 specific CD4$^+$ T-cells (greater expression on the right). DESeq2 was carried out using the Wald test and a design formula integrating the patient of origin and a combination term of tetramer and stimulation.

Upon stimulation (Figure 3.16), DRB1*01:01 cells highly expressed *GNLY* (granulysin, an antimicrobial cytolytic protein), *CTSW* (cathepsin W, a lymphocyte-specific protease), and

additional Th1-associated genes: *CCL5* and *TXK* (a tyrosine kinase essential for Th1 development) in comparison to stimulated DRB1*07:01. Similarly to non-stimulated DRB1*01:01, stimulated DRB1*01:01 differentially expressed *CCR5* and *IL18R1* compared to stimulated DRB1*07:01. The fact that *CCR5* is differentially expressed by DRB1*01:01 in both stimulation and ex-vivo suggests that a different, potentially more migratory $CD4^+$ phenotype is maintained upon stimulation. The cells also upregulated *CD103* (αE integrin, mediating tissue retention) and *LAG3* (an inhibitory receptor that regulates T-cell expansion). While this could suggest recent activation in-vivo, it could also represent a greater *CD103* and *LAG3* response to the in-vitro stimulation. This may represent greater capacity for tissue-resident memory differentiation in stimulated DRB1*01:01 cells. DRB1*07:01 cells under stimulation showed fewer canonical Th1 effector or cytolytic genes but notably expressed *IKZF4*, suggesting potential regulatory T-cell characteristics.

Pathway analysis using these differentially expressed genes, both as a complete set and restriction-specific groups, yielded no significant Gene Ontologies or Pathways after multiple testing corrections.

3.2.8 Repertoires are highly diverse, with frequent V-gene pairings shared between tetramers

I used MIXCR[336] to extract T-cell receptor (TCR) sequences and genes in each sequenced cell, and pre-processed them to remove dual alpha-chain T-cells. The T-cells sorted to both tetramers in the recovery phase had diverse T-cell receptor sequences with broad V and J gene usage. Many V (Figure 3.17) and J (Figure 3.18) genes had similar proportions between DRB1*01:01 and DRB1*07:01 specific cells. Exceptions to this were only seen with V-genes, which constitute 2/3 of the CDR loops, where *TRAV24* is more frequent in DRB1*07:01 specific T-cells, and *TRBV19* is more frequent in DRB1*01:01 specific T-cells. Otherwise the interquartile range of gene usage overlapped for most genes. This suggests that beyond *TRAV24* and *TRBV19*, the TCRs that sort to either of the tetramers are similar in their gene segment usage - suggesting that the requirements for binding are similar between the two tetramers.

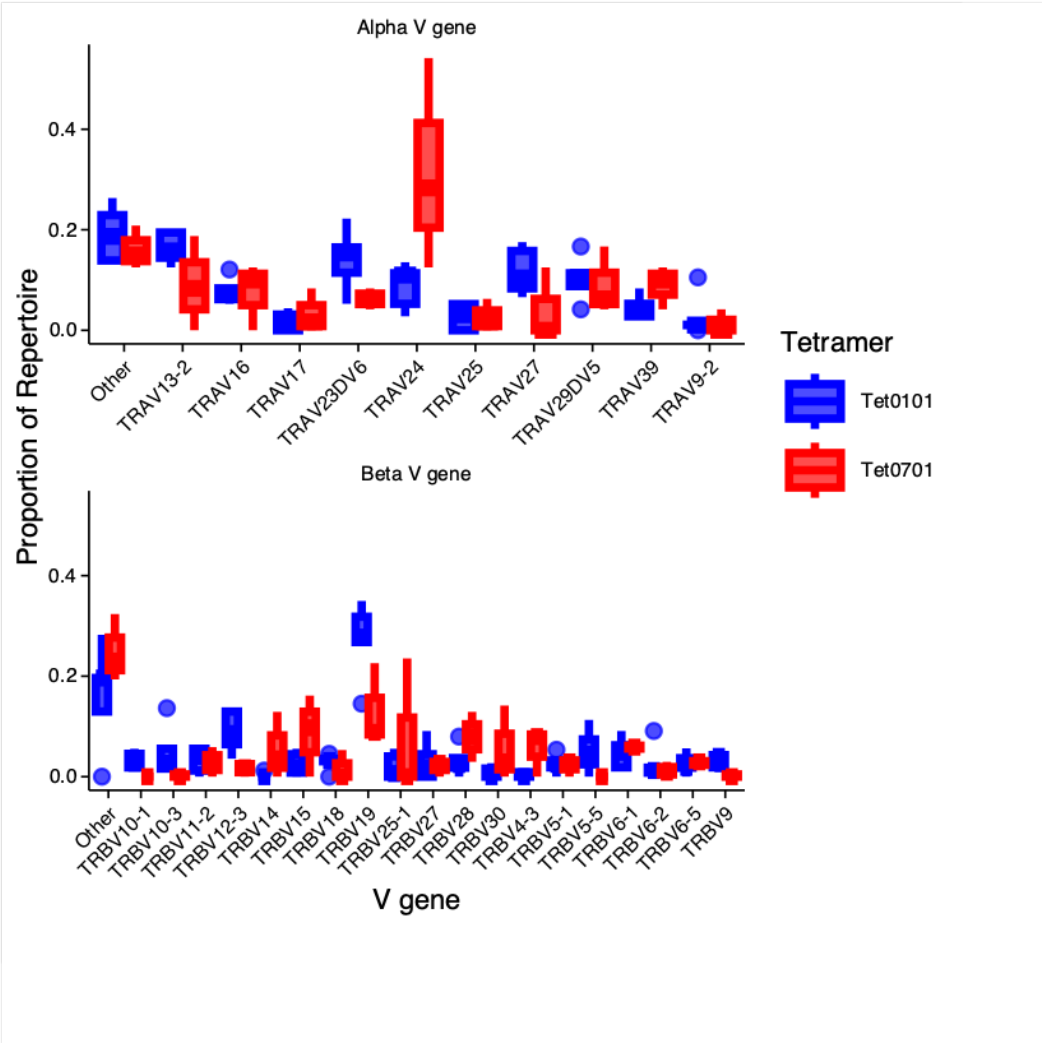


Figure 3.17: Box and whisker plots showing the proportion of the overall paired repertoire with specific Alpha V genes and Beta V genes, with shaded boxes corresponding to the IQR range of V gene frequency within a sample (the cells specific for a given tetramer, isolated at the Recovery Phase time point, for a given patient). Genes that made up less than 5% of the repertoire were aggregated into "Other".

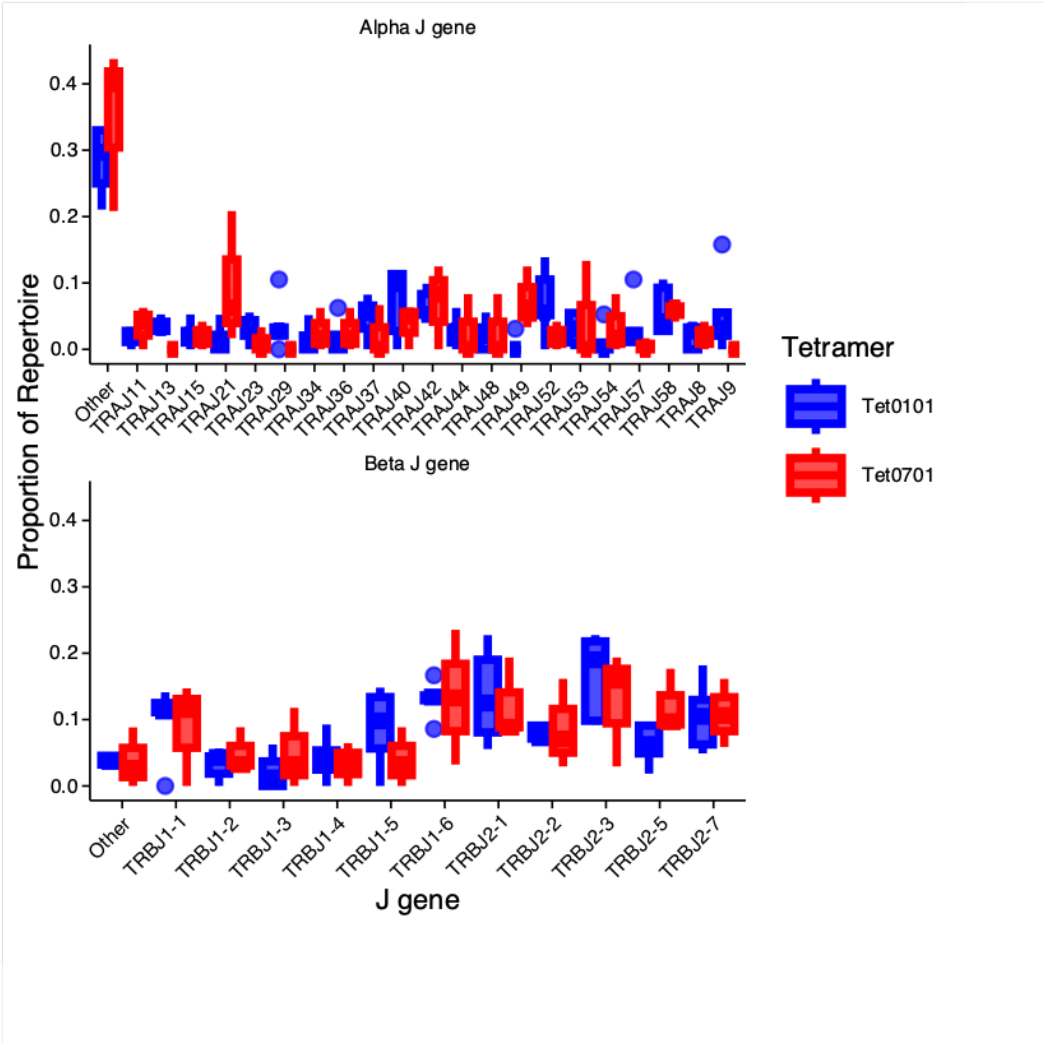


Figure 3.18: Box and whisker plots showing the proportion of the overall paired repertoire with specific Alpha J genes and Beta J genes, with shaded boxes corresponding to the IQR range of J gene frequency within a sample (the cells specific for a given tetramer, isolated at the Recovery Phase time point, for a given patient). Genes that made up less than 5% of the repertoire were aggregated into "Other".

I observed a single public TCR, where both cells with that TCR sorted to DRB1*01:01. Clones were small in size, with the largest paired alpha-beta clone containing 5 cells. Unlike a highly immunodominant $CD4^+$ Spike epitope, $S_{166\text{-}180}$, where public TCRs are common especially amongst the alpha chain, this epitope contains very few public TCRs.

3.2.9 Similar or identical TCRs sort to different tetramers, indicating the potential for cross-restriction

Given the structural similarity of the core regions of $M_{172\text{-}188}$ presented by DRB1*01:01 and DRB1*07:01 and the shared core-region requirements of $M_{172\text{-}188}$ specific T-cells restricted to those alleles, we wanted to explore whether similar T-cell receptors bound both MHC alleles.

The full-length TCR sequences enabled us to group and compare TCR sequences (Figure 3.19) to identify similar TCRs that sorted to both tetramers. I clustered TCRs together using pyrepseq[47], where T-cell receptors formed a cluster if their Levenshtein distance between individual chains was no more than 2 and across both chains was no more than 3. Edit distance has competitive, explainable performance in TCR clustering[358], and edit distance or its derivatives over paired chains are associated with shared antigen specificity and selection[47,279].

I combined all our single-cell TCR sequencing data of $M_{172\text{-}188}$-responsive T-cells. This comprises the described Smart-seq2 sequenced tetramer-sorted cells and 136 Smart-seq2 sequenced cells from 5 individuals identified through cytokine capture on stimulation with $M_{172\text{-}188}$. In addition, we include 1119 cells sequenced from short-term $M_{172\text{-}188}$ responsive $CD4^+$ T-cell lines derived from 3 patients. I deduplicated clonotypes in the cell lines based on their paired CDR3 amino acid sequences, V-, and J-genes, as these short-term cell lines have been maintained and expanded.

I identify 22 meta-clonotypes or clusters of similar TCRs from multiple individuals. I also identify 11 clusters (not necessarily across multiple individuals) where similar or identical TCRs are sorted into different tetramers (Figure 3.19). While the motifs and gene usage of the TCRs sorting to multiple tetramers were diverse, this suggests

that the structural and core-region similarity of $M_{172\text{-}188}$ presentation on DRB1*01:01 and DRB1*07:01 has functional consequences – namely that some TCRs could be "cross-restricted" to both. Within CD4⁺ T-cell clones generated and maintained by my colleagues, one was a perfect match to a TCR that sorted to both tetramers in the single-cell data (`CAASSLGHQFYF/CASSQGGTSGRLTGELFF`, *TRAV29/DV5/TRBV3-1*), and another was similar with a single amino acid difference in the CDR3 beta (`CAANRGSTLGRLYF/CASS[P|A]RLAGDSTDTQYF`, *TRAV23/TRBV11*). Both of these clones were originally identified as restricted to $M_{172\text{-}188}$-DRB1*01:01.

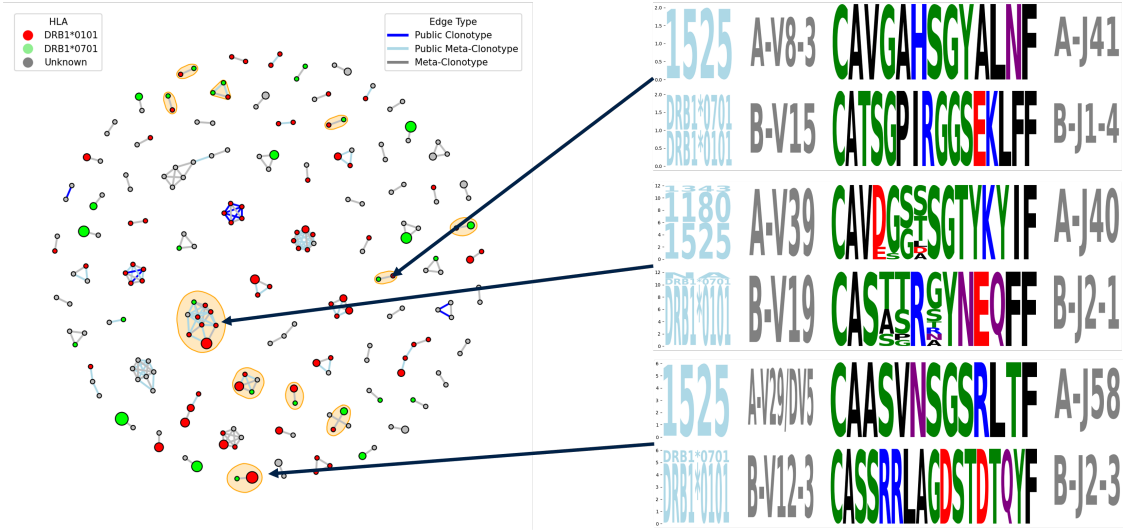


Figure 3.19: Annotated graph of the adjacency matrix of $M_{172\text{-}188}$ clonotypes. Each node corresponds to a clonotype in a particular patient, with the radius corresponding to the log of the clonal size, and their colour corresponding to their tetramer-sorted restriction for $M_{172\text{-}188}$ presented by DRB1*01:01 (red) or DRB1*07:01 (blue). Clonotypes from cell lines responsive to $M_{172\text{-}188}$ with unknown restriction (due to maintenance via stimulation) are in grey. Edges connect TCRs with a combined edit-distance of less than or equal to 3, and a per-chain edit distance of less than or equal to 2. Edges are coloured by type: Public Clonotype (identical CDR3s, different patients; blue), Public Meta-Clonotype (similar TCRs, different patients; lightblue), or Meta-Clonotype (grey). Orange shaded areas highlight similar clonotypes where members of the meta-clonotype sort to both DRB1*01:01 and DRB1*07:01. (Right Panel) Representative paired sequence logos for three cross-restricted meta-clonotypes, displaying V/J gene usage and CDR3 amino acid sequences for alpha (top) and beta (bottom) chains within each cluster. Metadata indicates the HLA alleles (DRB1*01:01, DRB1*07:01) and patient identifiers present in each depicted cluster.

To investigate the potential structural nature of recognition in the context of different HLA alleles, we used the AlphaFold3[359] server to predict the structure of the T-cell receptor, peptide and MHC complex for all the unique full-length paired TCRs that sorted to the different tetramers - 183 for DRB1*01:01, 48 for DRB1*07:01 and 5 where cells sorted

to both tetramers. The mean predicted Template Modelling score (pTM) is a measure of AlphaFold's confidence in its prediction where "overall predicted fold for the complex might be similar to the true structure"[360]. The predicted Template Modelling score (pTM) was $0.812(\pm 0.085)$ for DRB1*01:01 and $0.800(\pm 0.090)$ for DRB1*07:01 sequences. The mean interface predicted Template Modelling score (ipTM), which predicts the accuracy of the predicted relative positions of the subunits was 0.762 (± 0.121) for DRB1*01:01, and 0.741 (± 0.130) for DRB1*07:01. EMBL-EBI guidelines suggest "values higher than 0.8 represent high-quality predictions ... ipTM values between 0.6 and 0.8 are a grey zone"[360].

For high quality predictions (those with ipTM values > 0.8), the sum of contact probabilities of CDR loop residues was not significantly different between DRB1*01:01 sorting TCRs and DRB1*07:01 sorting TCRs. However, in DRB1*07:01 sorting TCRs (n=16), the CDR3β contributed a larger proportion of the buried surface area (median 33.1%) of the interaction than in DRB1*01:01 sorting TCRs (median 28.1%, n=56) with an adjusted p-value of 0.008 under Mann-Whitney U test with 6 comparisons. The distribution of these values is shown in Figure 3.20. Some T-cell receptors formed structural clusters where the RMSD between every CDR loop was less than 2Å for two different paired T-cells. Among the 9 clusters we identified of this type, two contained DRB1*01:01 and DRB1*07:01 sorting T-cell receptors. However, none of the TCRs that sorted to both DRB1*01:01 and DRB1*07:01 appeared in the same structural cluster, which could represent a different binding mechanism but could also reflect drawbacks of structure prediction.

3.2.10 $M_{172\text{-}188}$ specific T-cells are observed 4 years after primary SARS-CoV2 exposure, but almost no T-cell clones are maintained.

A key question surrounding antigen-specific immunity in SARS-CoV2 is how the response evolves. To our knowledge, no study has followed the long-term evolution of the antigen-specific immune response to any SARS-CoV2 epitope. Here, we report the single-cell phenotypes and repertoire of $M_{172\text{-}188}$-specific $CD4^+$ T-cells at 4 years after their initial infection.

First, I wanted to investigate whether the differences we observed in the recovery phase between T-cells specific to $M_{172\text{-}188}$-DRB1*01:01 and $M_{172\text{-}188}$-DRB1*07:01 were maintained

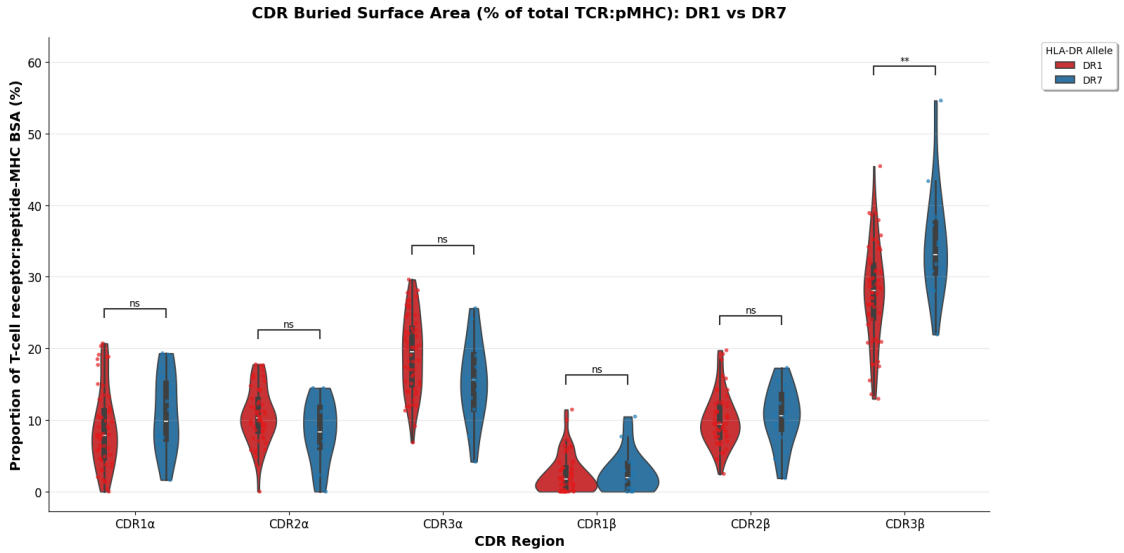


Figure 3.20: CDR3β contributes a larger proportion of the Buried Surface Area of the TCR:pMHC interaction in DRB1*07:01 (blue, labelled as DR7) than DRB1*01:01 (red, labelled as DR1) presented $M_{172\text{-}188}$. Violin Plots show the distribution of buried surface area (BSA) contributions from each CDR loop, as calculated by the Shrake Rupley algorithm for high-quality predicted structures. Significance was determined by Mann-Whitney U test with Bonferroni correction; ** = $p < 0.01$, ns = not significant.

at the 4-year timepoint. I aimed to identify whether differences in gene expression observed at the Recovery Phase timepoint were maintained at the 4-year timepoint. Using the same DESeq2 design formula as in the Recovery Phase, except without grouping by stimulation because we did not stimulate cells at the Four Year timepoint, I tested 131 cells sorting to the DRB1*07:01 tetramer against 588 sorting to the DRB1*01:01 tetramer. None of the differentially expressed genes at the Recovery Phase timepoint were significantly differentially expressed at the Four Year timepoint. This suggests that the transcriptional differences in the HLA-presentation of the epitope are only maintained in the short term, or that complex migratory or differentiation processes have occurred in the intervening period. Additionally, it is possible that with repeated exposures due to subsequent infections, the differences in DRB1*01:01 and DRB1*07:01 are attenuated.

Paired T-cell receptors maintained from the recovery phase timepoint to 4 years were very rare, with only a single paired clone identified at both timepoints in the two donors sampled at both timepoints. However, in the four year timepoint, yield of paired TCRs was much lower (36/350, 10.3%) than at the recovery phase (320/663), which impairs our ability to

detect maintained clones.

3.3 Discussion

$M_{172\text{-}188}$ is a highly immunodominant SARS-CoV2 epitope to which people with diverse HLA-II types generate $CD4^+$ T-cell responses. Using ELISPOT with mismatched B-cell lines and tetramer sorting, we found that individuals could respond to $M_{172\text{-}188}$ in the context of multiple HLA-DR alleles and HLA-DP. Promiscuous binding of epitopes to multiple HLA-DR1 alleles has been described[356]. The $M_{172\text{-}188}$ epitope meets the criteria for a DR-supermotif with a large aromatic residue at position 1 of its core region (Y) and a small, non-charged residue at position 6 (S). The prediction hierarchy (DRB1*01:01 > DRB1*07:01 ≈ DRB1*15:01 » DRB1*04:01) aligns with our functional observations (Section 3.2.4), where DRB1*04:01-restricted responses were weakest.

Extensive studies of SARS-CoV2 epitopes and their immunodominance[95,203,343,361] have been conducted; however, these have generally been studied in the context of one restricting MHC allele. A detailed survey of binding characteristics and structures of immunodominant SARS-CoV2 epitopes presented by DR1 showed an overlapping epitope to $M_{172\text{-}188}$, namely $M_{176\text{-}190}$ has extremely high affinity to HLA-DR1. The core region in the study of the $M_{176\text{-}190}$ epitope presented by HLA-DRA1/DRB1*01:01 was found to be YKLGASQRV, and it was the highest affinity binder of HLA-DR1 of the immunodominant SARS-CoV2 epitopes tested[362].

Examples from malaria[363] and HIV[364,365] show that immunodominance can result from favourable binding to multiple HLA alleles and thus the capacity to be presented by larger proportions of the population. Here, we show for the first time in SARS-CoV2 that one of the most immunodominant epitopes, $M_{172\text{-}188}$, generates frequent responses through promiscuous presentation.

In this study, we leverage our deeply characterised, longitudinal cohort of individuals infected with SARS-CoV2 in the first wave (ancestral variant, early 2020). I used this cohort to explore the T-cell response to $M_{172\text{-}188}$, identifying that individuals who responded to this

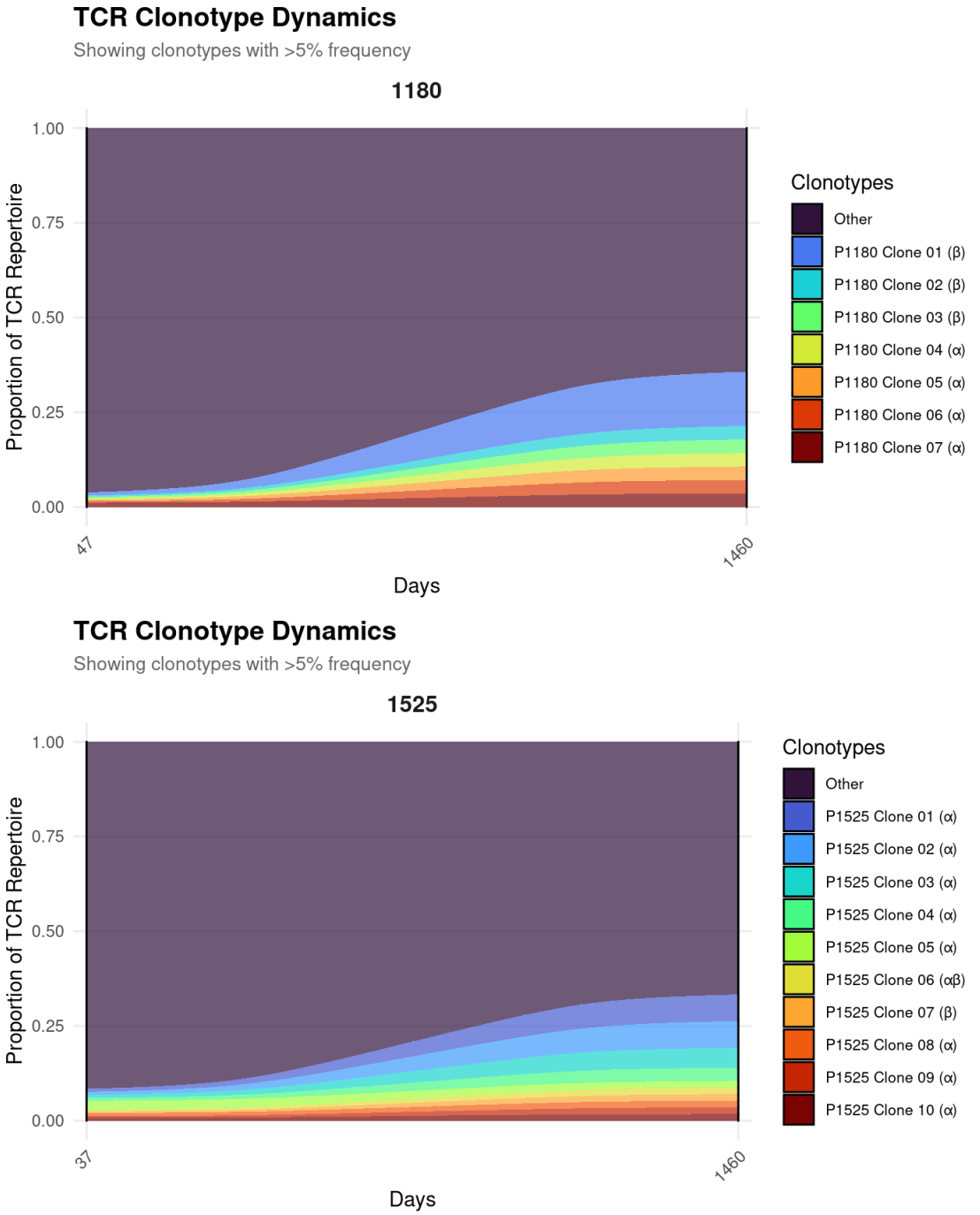


Figure 3.21: The Clonotype dynamics between sampling after first infection and subsequent sampling at 4 years after the initial infection reveal minimal maintenance of paired clonotypes. Only one paired clone, in P1525, was present at both the initial and the four year timepoint. Where clones were maintained, they made up a larger proportion of the repertoire at the four year timepoint than at the recovery phase, but we could not exclude low sampling as a driver of this.

epitope had diverse HLA class II genotypes. We also observed trends in the clinical severity of SARS-CoV2 between responders with different DRB1 alleles.

First, my colleagues confirmed this epitope's physiological presentation and recognition in the context of multiple HLA class II alleles and molecules using blocking antibodies and HLA-mismatched EBV-transformed B-cell lines. To characterise the residues of this epitope involved in antigen presentation and recognition, they generated truncated variants and screened them, identifying shared fine peptide requirements between HLA-DR alleles. The solved structures of the epitope presented by DRB1*01:01 and DRB1*07:01 showed similarity.

Second, my colleagues used functional assays to characterise the mechanisms and implications of differences in $M_{172\text{-}188}$ presentation by different HLA class II alleles. Fitting peptide concentration against T-cell activation curves for multiple $M_{172\text{-}188}$ specific clones allowed us to interpolate the equivalent pulsed antigen concentration presented in live viral infection by the different HLA-DR alleles. We found that interpolated antigen concentration was consistently the highest when presented by DRB1*01:01. This suggests that the same epitope might be presented in different quantities on different HLA class II molecules in a similar conformation. This offers a novel approach to understanding the impact of antigen load on $CD4^+$ T-cell memory in a physiologically relevant context.

This hierarchy of presentation levels fit with functional outcomes in bulk patient-derived cell lines - DRB1*01:01-restricted T-cells showed superior viral suppression and stronger IFN-γ responses compared to those restricted to other alleles. In contrast, DRB1*04:01 restricted cells showed the weakest viral suppression and were most affected by variants of concern.

Comparing the epitope presentation between the different HLA-DR alleles and HLA-DP, we found that core regions of $M_{172\text{-}188}$ presentation by HLA-DR alleles had shared crystal structures and fine-peptide requirements for T-cell responses. In contrast, the core region for HLA-DP response showed stringent requirements at the N-terminus. This suggests that HLA-DP would not present the overlapping epitope ($M_{176\text{-}190}$).

The structural basis for these differences could lie in the polymorphisms at the base of the peptide-binding cleft. While the core region of $M_{172\text{-}188}$ adopts similar conformations when presented by DRB1*01:01 and DRB1*07:01, the wider binding cleft in DRB1*07:01 could affect peptide loading or stability. The peptide-MHC structure and T-cell function relationship shares parallels with an immunodominant HIV epitope. A Gag epitope presented by multiple DR alleles generates the most potent responses when presented by HLA-DR11, with potency descending with the opening of the hinge region in DR1 and DR15[305]. In the DRB1*07:01 structure, we see a similar opening of the hinge region corresponding to weaker functional phenotypes, indicating that perhaps, like the Gag epitope, the TCR interaction is firmly "peptide focused".

These quantitative differences in antigen presentation appear to shape the phenotype and function of responding T-cells. Leveraging the natural experiment of the same epitope presented by multiple MHC alleles including in heterozygous individuals, we conducted ex-vivo and post-stimulation single-cell sequencing to compare phenotypes between T-cells restricted to different alleles. DRB1*01:01-restricted cells exhibit higher expression of effector genes like *GZMB*, *CCR5* and *CCR6*, cytokines such as *XCL1*, and regulatory molecules such as PD-1 (*PDCD1*), PD-L1 (*CD274*) and *IL10*. The expression of PD-1 in SARS-CoV2 specific CD8⁺ T-cells has been associated with functionality[366], but it is unclear whether this relationship holds for CD4⁺ cells studied here. In contrast, DRB1*07:01-restricted cells showed enrichment for fewer genes with clear roles in CD4⁺ T-cells, but they included *NFKBID* and *NLRP3*. This finding connects with previous work showing that antigen load can influence CD4⁺ T-cell differentiation in in-vitro and mouse models[311–314,345] though the directionality of this effect has been debated.

We sequenced paired T-cell receptors from these cells to investigate whether these HLA-dependent differences in antigen presentation influenced clonal selection and maintenance. Examining the T-cell receptor repertoire and its evolution over this longitudinal time point, we observed highly diverse repertoires, with only one publicly paired T-cell receptor. Across the sequenced T-cells, we observed very little maintenance of $M_{172\text{-}188}$ specific T-cell clones from the recovery phase to the four-year timepoint.

We found some evidence for TCR cross-restriction between DRB1*01:01 and DRB1*07:01. Eleven clusters of similar or identical TCRs sorted to both tetramers, suggesting that the structural similarity in peptide presentation can enable recognition by the same TCR. This raises interesting questions about whether such cross-restricted T-cells might exhibit intermediate phenotypes or respond differently depending on which allele presents the antigen, and more studies of T-cell function in this circumstance are needed. It is worth noting that our assignment of antigen-specificity is highly dependent on tetramer-sorting, which is one of the most specific methods that exist[367] but can identify non-reactive clonotypes[55].

While $M_{172\text{-}188}$-specific T-cells persisted four years after their donor's primary SARS-CoV2 exposure, individual T-cell clones were minimally maintained from the initial recovery phase. Our donors likely had subsequent SARS-CoV2 infections, and new clones may have been recruited, or previous clones were no longer identifiable in the blood compartment. A limitation of our study is that we could not track potential exposures or infections of our long-term follow-up donors.

The divergence between early and late T-cell populations may reflect repertoire evolution, as a study of antigen-specific mouse T-cell populations has shown that initial recruitment of broad TCR repertoires provides a polyclonal niche from which flexible secondary responses can be recalled[368]. While in their model high avidity clones dominated the initial response, TCRs better suited to recognise variant epitopes were recruited by subsequent heterologous challenge. This may be what we are observing at the 4-year time point, and further studies of the functional properties of T-cell receptors sampled longitudinally may help confirm this.

The lack of maintained transcriptional differences between DRB1*01:01 and DRB1*07:01-restricted cells at 4 years after primary infection could imply that initial differences in antigen load may primarily influence the acute response rather than determining long-term memory characteristics.

These results expand our understanding of the $CD4^+$ T-cell response to SARS-CoV2 beyond spike protein epitopes. The conservation of the $M_{172\text{-}188}$ epitope and its T-cell response

across emerging viral variants highlights its potential relevance for vaccine design and immune monitoring. Incorporating highly immunogenic epitopes presented by multiple HLA types may help induce broad T-cell immunity or allow monitoring of T-cell responses across individuals in a single assay. The ability of $M_{172\text{-}188}$ to be presented by multiple HLA-DR alleles, the combination of which has greater population coverage than an epitope presented by a single HLA allele, makes it an exciting candidate for both vaccines and targeting antigen-specific immune responses to non-spike epitopes.

While spike protein-focused immunity has been central to vaccine development, our study highlights the importance of non-spike epitopes in natural immunity to SARS-CoV2. The $M_{172\text{-}188}$ epitope shows the potential advantages of targeting non-spike regions: conservation across variants, promiscuous HLA presentation, and generation of durable T-cell responses that could still be measured four years after primary infection.

As new variants emerge, monitoring responses to conserved epitopes like $M_{172\text{-}188}$ becomes increasingly important. Our data show that $M_{172\text{-}188}$-specific T-cells maintain recognition of multiple variants of concern, likely due to the epitope's high conservation. This contrasts with spike-specific responses, where escape variants are currently more strongly selected for as Spike is a target of neutralising antibodies[369]. The preservation of $M_{172\text{-}188}$ across SARS-CoV2 variants in the UK and related coronaviruses suggests it could contribute to stable, resilient immunity.

Our study's limitations include its deep focus on a few convalescent patients from a single cohort. While two individuals can serve as their own control due to the expression of DRB1*01:01 and DRB1*07:01 alleles, the generalisability of altered peptide presentation by different alleles should be expanded to larger, more clinically varied cohorts. Expanding to other geographic populations and HLA alleles not covered by our cohort could identify further HLA class II alleles that present $M_{172\text{-}188}$. Also, while we characterised functional profiles of $M_{172\text{-}188}$-specific cells in convalescence, more work is needed to assess how these functional profiles relate to disease outcomes or protection. More extensive cohort studies that separate individuals by their responses to SARS-CoV2 epitopes could clarify how HLA genotype and CD4⁺ T-cell functionality relate to clinical outcomes.

In conclusion, our findings demonstrate that SARS-CoV2 membrane protein epitope $M_{172\text{-}188}$ is immunodominant across multiple HLA class II alleles and gives rise to $CD4^+$ T-cells with distinct functional profiles. The ability of a single viral epitope to generate T-cell responses restricted by multiple HLA types may be an important determinant of its immunodominance and population-level impact. Further defining the factors that underlie heterogeneous T-cell responses to immunodominant epitopes may inform strategies for vaccination and monitoring of immune responses.

3.4 Methods

3.4.1 Interpolation of $M_{172\text{-}188}$ concentration in live viral co-culture

To investigate how the HLA-allele impacts the degree of activation of $M_{172\text{-}188}$-specific T-cells, my group developed an interpolated live-infection pulsing concentration assay. First, my colleagues generated a panel of B-cell lines with known HLA-DRB1 alleles that present $M_{172\text{-}188}$. Then, they co-cultured short term cell lines specific for $M_{172\text{-}188}$ presented by particular HLA-DRB1 alleles with these B-cell lines pulsed at varying concentrations of $M_{172\text{-}188}$ peptide. Pulsing was carried out according to Section 2.1.6.

Antigen concentration-response curves were fitted using a 3-parameter log-logistic model (`LL.3`) using the `drc` R package, whereby the curve was fit on pulsed concentrations (0.016, 0.08, 0.4, 2, 10 μg/ml) and queried with the observed ICS IFN-γ positive frequency.

3.4.2 Predicted Structural Analysis

For all observed combinations of paired T-cell receptors and peptide-MHC ($M_{172\text{-}188}$-DRA*01:01-DRB1*01:01/DRB1*07:01), we made predictions using the AlphaFold Server[359]. I gathered all the unique paired T-cell receptors from our tetramer sorting experiments and used stitchr[370] to obtain full-length amino-acid sequences where MIXCR did not cover specific regions. Inputs to the server were the IMGT numbered subsequences of the T-cell receptor (via ANARCI[371]), the sub-sequences of DRA*01:01 and DRB1*01:01 present in the structure 8PJE[372], and the homologous subsequence in DRB1*07:01 (for which no structures exist). For comparison of binding patterns, we excluded all structures with a predicted interface TM-score of below 0.8 or a clash (as determined by has_clash).

To investigate which elements of the T-cell receptor-peptide-MHC interaction are forming contacts, we used the `contact_probs` matrix returned by AlphaFold3 for each prediction. I extracted the contact probabilities from each matrix between the ANARCI annotated CDR regions (CDR1, CDR2, CDR3) and summed the contact probabilities with the peptide over those regions within each T-cell receptor to identify the contribution of each complementarity-determining region to the peptide-TCR interface.

I calculated buried surface area (BSA) using the Shrake-Rupley algorithm implemented in BioPython. The BSA was computed by calculating solvent accessible surface area (SASA) for isolated components and the multimer, and then BSA was calculated as half the difference between the sum of individual component SASA and the complex SASA. For CDR region-specific analyses, we calculated region-specific SASA in isolation and in complex, then determined the BSA contribution of each CDR region to both peptide-specific and total peptide-MHC interactions.

To calculate RMSD differences of CDR loops between structures for clustering, we used spline up-sampling[373] to adjust for CDR loops with different lengths. CDR loops were first aligned on the 5 residues before and after the CDR region using the BioPython `Superimposer` class, and then RMSDs over the loop were calculated. Sequences were greedily clustered together if the RMSD between all of their CDR loops was less than 2Å.

4

Comparing the evolution of Spike and non-Spike specific CD8⁺ T-cells in Breakthrough SARS-CoV2 Omicron infection

Summary

Breakthrough SARS-CoV2 infections allow us to study how T-cell memory adapts to repeated antigen exposure. I profile CD8⁺ T-cells specific for five immunodominant epitopes in 14 healthcare workers with different exposure histories before and after breakthrough infection. Single-cell RNA and TCR sequencing generated 3,157 high-quality cells from infection-primed (n=10) and vaccine-primed (n=4) individuals.

Five transcriptional clusters make up the sampled CD8⁺ T-cells, representing different states: terminal effector (GNLY (0)), central memory (CCR7 (1), KIF23 (4)), transitional (MALAT1 (2)) and effector memory (ABI3 (3)). Differential gene expression shows that exposure sequence, not just number, influences CD8⁺ T-cell phenotypes. Vaccination following infection produces cells that highly express *EOMES*, *PRDM1* and *XCL2* while individuals primed by vaccination and subsequently infected generated cells expressing *IL10*.

TCR repertoire diversity is similar across 1 to 5 exposures. Logistic regression modelling of 12 paired samples showed clonotype persistence through breakthrough infection is strongly associated with pre-Omicron transcriptional cluster. Clonotypes with cells in effector clusters GNLY (0) and ABI3 (3) were more likely to persist. Exposure combinations shape memory CD8⁺ T-cell phenotypes, although demographic confounding between exposure groups limits causal inference.

4.1 Introduction

Breakthrough infections provide a model to study how T-cell memory adapts to repeated antigen exposure. SARS-CoV2 has infected hundreds of millions worldwide[145], and the Omicron variant has been associated with significant infection in vaccinated and previously infected individuals[374]. Through 37 mutations in the Spike protein, Omicron evades neutralising antibodies generated by previous infection or vaccination[375]. The resulting breakthrough infections have been associated with milder disease[376,377] but represent a public health challenge.

Cross-reactive T-cells primed by ancestral strains recognise the Omicron variant. Of 280 CD4⁺ epitopes and 523 CD8⁺ epitopes[343], 93-97% are unchanged across B.1.1.7, B.1.351, P.1, and CAL.20C[378]. Intracellular cytokine staining of spike peptide megapools in both vaccinated and convalescent individuals has shown that 70-80% of the CD4⁺ and CD8⁺ T-cell response to Spike peptides is maintained between the ancestral strain and Omicron[379]. Work by Gao et al. showed that "spike-specific CD4⁺ and CD8⁺ T-cells induced by prior infection or ... vaccination provide extensive immune coverage against [Omicron]". In their study comparing individuals vaccinated with BNT162b2 (6 months), recovering from infection (9 months) or seronegative controls they measured responses via AIM of Spike megapools. Comparing the activated proportion between Omicron and ancestral Spike peptides they found no significant difference in median relative frequency of responses to Omicron compared to Wildtype peptides in any subset (CD4⁺, CD8⁺, vaccinated, convalescent)[380]. Therefore, breakthrough infections represent recall responses to many of the antigen-specific T-cells responsive to ancestral strains.

Several studies have dissected how breakthrough infection influences T-cell memory. Tarke et al. used AIM with megapools and ICS in 100 double-vaccinated individuals with symptomatic Delta or Omicron infection to identify and characterise Spike specific and non-Spike specific T-cell responses. They detected unknown prior breakthrough infection through non-Spike specific T-cell responses, which approved vaccines do not induce. S-responsive CD4⁺ T-cells increased only in individuals without prior non-Spike immunity after breakthrough infection, indicating that natural infection might generate a "maximum" response. Similarly, only non-Spike directed CD8⁺ responses increased in frequency with breakthrough infection[381]. Studying the present cohort prior to this work, Hornsby et al. found that "Spike-specific T-cells increase only in infection-naive vaccinees" but that "post-omicron T-cell responses are significantly higher in the previously infected". Sorting cells for immunodominant epitopes, they find that the phenotypes of Spike specific cells did not change significantly from before to after Omicron infection, but that Spike specific CD8⁺ T-cells in individuals infected then triple vaccinated had fewer Central Memory, Transitional Memory cells, but more Terminal Effector cells. The frequency of Granzyme B positive cells by flow-cytometry of Spike-specific infection primed cells was similar to those they sorted for CMV epitopes, much higher than the GZMB⁺ frequency in only vaccinated individuals[322].

Bulk measurements cannot track whether clonotypes persist over exposures. Single-cell analysis is required to identify persistent clonotypes and assess how breakthrough infection shapes the paired repertoire. Minervina et al. used DNA-barcoded dextramers to follow over 4,000 epitope specific T-cells across individuals with diverse exposure histories. They found that vaccination after infection expanded Spike specific clonotypes and shifted these cells towards high expression of GZMB, GZMH and GNLY compared to single-exposure groups (both vaccine and infection)[253]. For breakthrough infection, their study was cross-sectional, and therefore lacked the ability to track individual clonotypes.

In this study, I characterise antigen-specific CD8⁺ T-cells longitudinally with breakthrough infection, profiling antigen-specific cells to five immunodominant epitopes to understand how different combinations of exposure influence individual cell phenotypes and their

repertoire. By performing single-cell RNA and TCR sequencing before and after Omicron breakthrough infection, I track clonotypes across timepoints and identify transcriptional clusters associated with persistence. Sequential sampling in the context of different exposures allows us to investigate how the order influences the phenotype of antigen-specific CD8⁺ T-cells.

4.2 Results

4.2.1 Study cohort and epitope-specific responses

Table 4.1: Clinical metadata of Sheffield healthcare worker cohort by priming exposure. Tetramer⁺ CD8⁺ T-cells recovered represents the numbers of wells, within the Smart-seq2 protocol, that contained quantifiable RNA (pre-QC).

Characteristic	Overall	infection	vaccine
N	14	10	4
Age, mean (SD)	45.3 (9.3)	48.9 (8.5)	36.2 (2.4)
Female, n (%)	11 (78.6%)	9 (90.0%)	2 (50.0%)
Tetramer⁺ CD8⁺ T-cells recovered, mean (SD)	231.4 (223.1)	280.6 (246.1)	108.2 (76.0)
Days post-vaccine, mean (SD)	40.1 (27.5)	38.7 (32.6)	43.8 (7.1)
Days post-symptoms, mean (SD)	33.6 (6.3)	33.1 (6.8)	34.8 (5.6)

We studied 14 healthcare workers from the Sheffield cohort with a history of infection and/or vaccination (Section 2.1.1). All individuals were triple vaccinated with mRNA vaccines, and subsequently became infected with BA.1 or BA.2 infections. Ten individuals were infection-primed (infected then triple-vaccinated) and four were vaccine-primed (triple-vaccinated only). We sorted CD8⁺ T-cells specific for five HLA-restricted epitopes (Table 4.2) before and after Omicron breakthrough infection.

Table 4.2: Immunodominant SARS-CoV2 CD8 Epitopes used in Sheffield Healthcare worker study of Antigen-Specific CD8 T-cell evolution.

Epitope	HLA	gene
KCYGVSPTK	HLA-A*03:01	Spike
LTDEMIAQY	HLA-A*01:01	Spike

Epitope	HLA	gene
YLQPRTFLL	HLA-A*02:01	Spike
TTDPSFLGRY	HLA-A*01:01	ORF1ab
SPRWYFYYL	HLA-B*07:02	Nucleoprotein

We obtained 32 unique patient-epitope-timepoint combinations, including 12 paired samples (same patient and epitope pre- and post-Omicron). This enables longitudinal analysis for a subset of individuals and epitopes, although only one paired sample came from a vaccine primed individual. We obtained no paired samples for non-Spike epitopes in individuals primed by vaccination, reflecting the very low frequency of naive circulating antigen-specific cells in these individuals at the pre-infection timepoint. Subclinical SARS-CoV2 exposures are challenging to measure, and we cannot definitively exclude them in this study. However, previous infection status was defined on the basis of workplace PCR or serology, or from MSD analysis of S and N plasma IgG levels. The metadata of the cohort is imbalanced, with infection primed individuals being older (mean age 48.9 vs 36.2) and more likely to be female (90% vs 50%).

4.2.2 Transcriptional clusters define CD8⁺ T-cell states

After quality control I generated principal components over the 1500 most variable genes, selecting the first 6 based on inspection of the elbow plot. I identified 5 clusters at resolution 0.5. These clusters are defined by different marker genes, and labelled on the basis of the gene with the lowest p-value on a Wilcoxon rank-sum test compared to the other clusters with a `log2FoldChange` greater than 0.1.

The GNLY (0) cluster most differentially expresses tissue-resident marker *ZNF683* (or HOBIT, $\log_2$FC = 2.0, $p < 0.001$), terminal effector markers *GNLY* ($\log_2$FC = 1.5, $p < 0.001$), *GZMH* ($\log_2$FC = 1.4, $p < 0.001$), *FGFBP2* ($\log_2$FC = 1.2, $p < 0.001$), and *KLRG1* ($\log_2$FC = 1.4, $p < 0.001$). The cluster highly expresses cytotoxic genes *GZMA* ($\log_2$FC = 0.90, $p < 0.001$), *GZMB* ($\log_2$FC = 0.81, $p < 0.001$), *NKG7* ($\log_2$FC = 1.1, $p < 0.001$) and chemokine *CCL4* ($\log_2$FC = 1.0, $p < 0.001$). This cluster could represent terminal effector or TEMRA CD8⁺ T-cells with cytotoxic capacity and the expression of genes implying

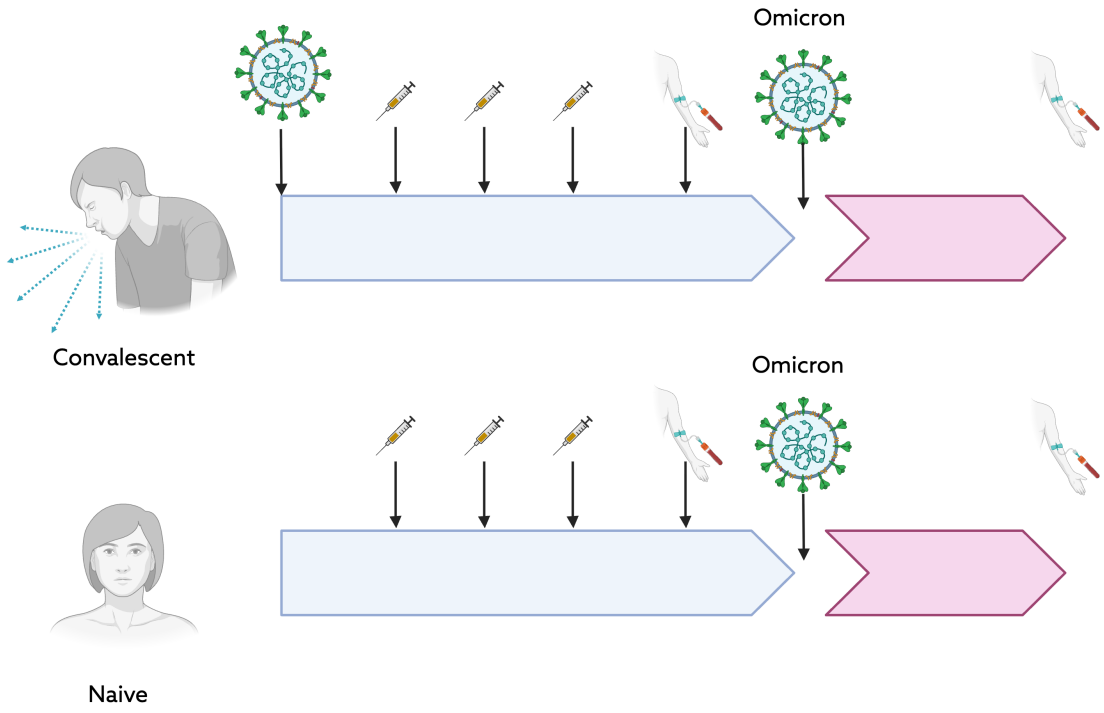


Figure 4.1: Schematic of the cohort longitudinally tracking epitope specific CD8⁺ T-cells through Omicron breakthrough infection. Healthcare workers primed by infection (n=10, top) or primed by vaccination (n=4, bottom) were sampled after the completion of a full course of 3 vaccinations and then again after Omicron infection. Infection-primed individuals were initially infected with SARS-CoV2, received 3 vaccine doses then had their first samples taken. In vaccine primed individuals, their only exposure to SARS-CoV2 antigen was via the vaccination course. Then, following PCR confirmed Omicron infection both groups were sampled again.

tissue residency. All epitopes and timepoints contained cells in this cluster. The cluster shows negative $\log_2$FC for *CCR7* ($\log_2\text{FC} = -5.7$, $p < 0.001$) and *SELL* ($\log_2\text{FC} = -0.68$, $p < 0.001$) indicating that these cells are unlikely to re-enter lymphoid tissues.

The CCR7 (1) cluster most differentially expresses naive/central memory markers *CCR7* ($\log_2\text{FC} = 1.9$, $p < 0.001$), *TCF7* ($\log_2\text{FC} = 1.9$, $p < 0.001$), *SELL* ($\log_2\text{FC} = 1.0$, $p < 0.001$), and *IL7R* ($\log_2\text{FC} = 1.3$, $p < 0.001$). The cluster shows modest upregulation of *CD69* ($\log_2\text{FC} = 0.81$, $p < 0.001$) but not expression of cytotoxic effector molecules. Also distributed between timepoints and epitopes, this cluster likely represents central memory cells. Almost opposite to the GNLY (0) cluster, this cluster shows negative $\log_2$FC for *GZMH* ($\log_2\text{FC} = -4.1$, $p < 0.001$), *GNLY* ($\log_2\text{FC} = -2.0$, $p < 0.001$), and *NKG7* ($\log_2\text{FC} = -2.8$, $p < 0.001$) as well as other genes defining Cluster 0.

The MALAT1 (2) cluster shows modest differential expression across markers, with the long

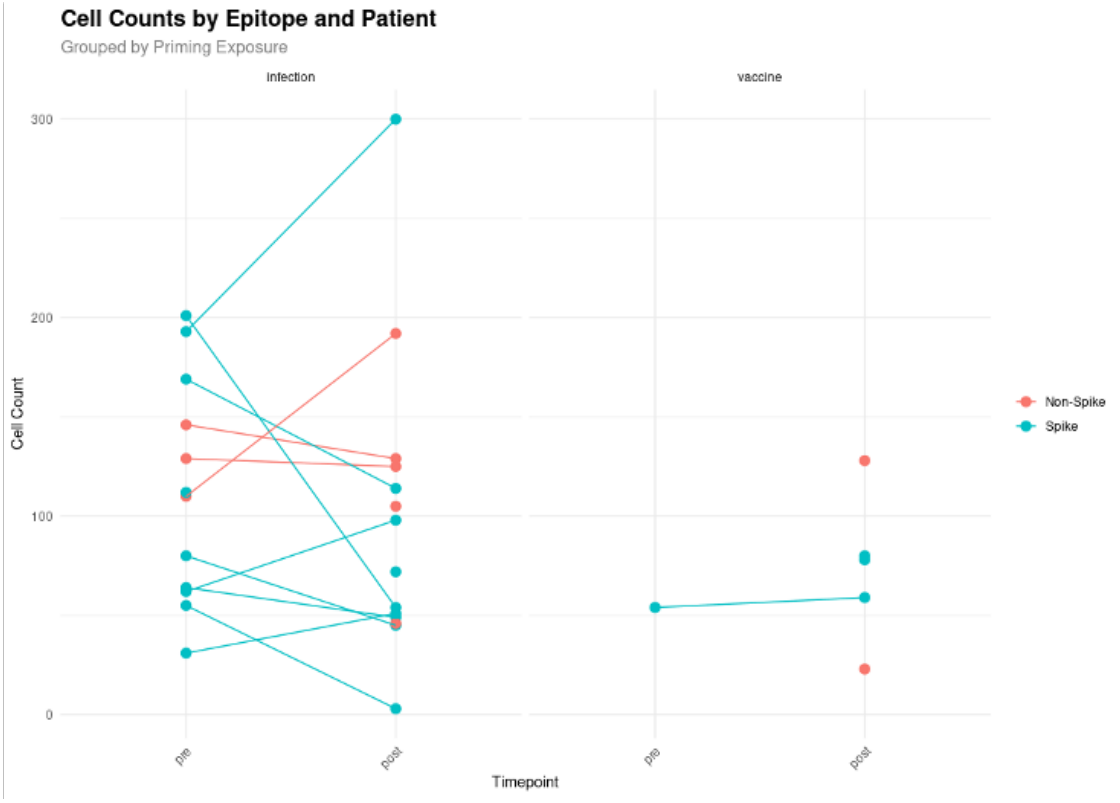


Figure 4.2: Number of tetramer-sorted CD8⁺ T-cells recovered before and after Omicron breakthrough infection labelled by epitope specificity. Tetramer-specific CD8⁺ T-cells specific for Spike (teal) and non-Spike (red) epitopes from 14 healthcare workers with either infection-primed (left) or vaccine primed (right) infection histories. Lines connect samples of the same epitope in the same patient before and after Omicron infection. Most paired samples (11/12) came from infection primed individuals. This is not a measure of the quantity or frequency of circulating cells, as the number isolated is usually multiples of a 96 well plate.

non-coding RNA *MALAT1* ($\log_2$FC = 0.64, $p < 0.001$) as its top marker and several ribosomal or pseudogenes amongst its top 20. The absence of positive expression of canonical memory, effector, or exhaustion markers suggests this cluster represents a homeostatic or transitional population, present at both timepoints. Reduced expression of *CD8A*, *CD8B* and *CD3E* ($\log_2$FC = −0.41, $p < 0.001$; $\log_2$FC = −0.27, $p < 0.001$; $\log_2$FC = −0.40, $p < 0.001$ respectively) suggests that this cluster may contain low quality or contaminant cells, although all cells were flow-sorted and passed quality control.

The ABI3 (3) cluster differentially expresses the terminal effector marker *S1PR5* ($\log_2$FC = 2.6, $p < 0.001$), transcriptional repressor *BCL6* ($\log_2$FC = 2.6, $p < 0.001$), and *ABI3* ($\log_2$FC = 2.5, $p < 0.001$). The cluster shows high expression of cytotoxic genes including

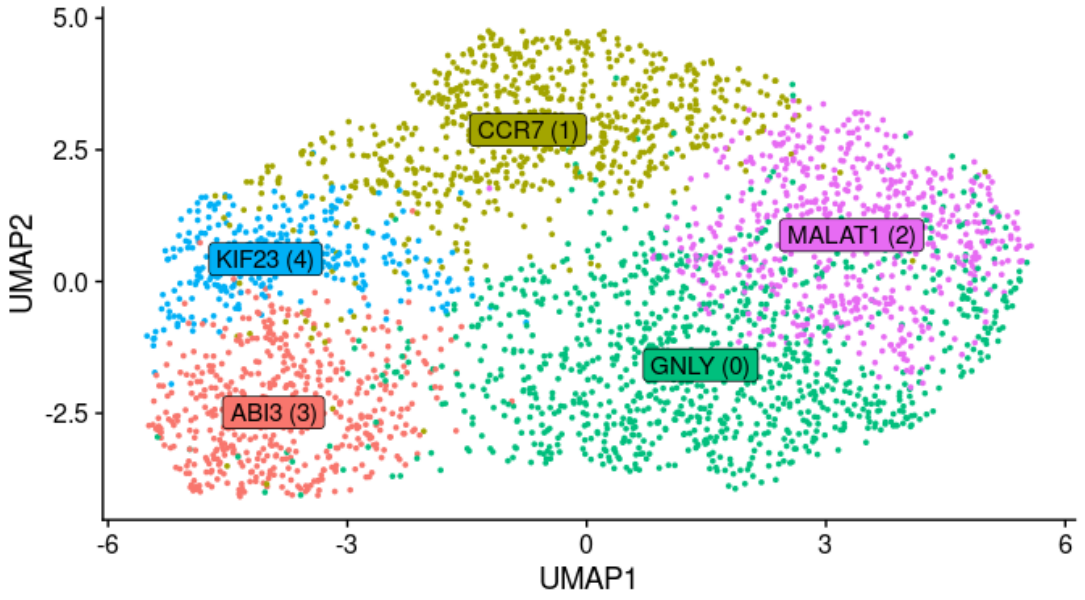


Figure 4.3: UMAP of 3157 CD8⁺ T-cells from Sheffield healthcare workers, coloured by cluster. The clusters are defined by the most significantly differentially expressed genes in each cluster, filtered to those with a log2 fold change greater than 0.1, and difference in percentage of cells expressing that transcript compared with other cells of 0.1.

PRF1 ($\log_2$FC = 1.2, $p < 0.001$), *GZMB* ($\log_2$FC = 0.56, $p < 0.001$), *NKG7* ($\log_2$FC = 0.10, $p < 0.001$), as well as exhaustion markers *HAVCR2* ($\log_2$FC = 1.3, $p < 0.001$) and *CREM* ($\log_2$FC = 0.78, $p < 0.001$). This set of cells differentially express glycolytic gene *LDHA* ($\log_2$FC = 1.3, $p < 0.001$) and related pseudogenes *LDHAP3*, *LDHAP4*, and *LDHAP5* ($\log_2$FC = 1.5, 1.4, 1.4, $p < 0.001$ respectively), which could suggest metabolic reprogramming of effector differentiation. Slightly more frequent in pre-Omicron samples with strong enrichment for the Spike YLQPRTFLL (YLQ) epitope (301/501 cells), this cluster likely represents effector memory cells primed for glycolytic metabolism upon recall and retained cytotoxic capacity.

The KIF23 (4) cluster has proliferation-associated *KIF23* as its most enriched gene ($\log_2$FC = 2.9, $p < 0.001$) and *CCR7* as its second ($\log_2$FC = 2.7, $p < 0.001$), with naive/central memory markers *LEF1* ($\log_2$FC = 1.8, $p < 0.001$), *TCF7* ($\log_2$FC = 1.2, $p < 0.001$), *IL7R* ($\log_2$FC = 1.2, $p < 0.001$), and *SELL* ($\log_2$FC = 1.1, $p < 0.001$). It also highly expresses costimulatory molecule *CD28* ($\log_2$FC = 2.1, $p < 0.001$) and chemokine receptor *CXCR3*

($\log_2$FC = 1.7, $p < 0.001$). The cluster expresses exhaustion markers *TIGIT* ($\log_2$FC = 1.1, $p < 0.001$) and *CREM* ($\log_2$FC = 0.83, $p < 0.001$), and cytotoxic or effector memory marker *GZMK* ($\log_2$FC = 0.92, $p < 0.001$). The cluster shows strongly negative $\log_2$FC for terminal effector markers *FGFBP2* ($\log_2$FC = -5.3, $p < 0.001$), *GZMH* ($\log_2$FC = -5.0, $p < 0.001$), and *GNLY* ($\log_2$FC = -2.9, $p < 0.001$), and cytotoxic markers *GZMB* ($\log_2$FC = -3.6, $p < 0.001$), *NKG7* ($\log_2$FC = -2.6, $p < 0.001$), and *GZMA* ($\log_2$FC = -1.4, $p < 0.001$), suggesting memory rather than effector differentiation.

These clusters represent distinct states of antigen-specific CD8⁺ T-cells in memory, as they are present across individuals, vary in their clonal structure (Figure 4.5) and are enriched for different gene sets (Figure 4.4). Clusters ABI3 (3) and GNLY (0) have the highest proportion of cells belonging to large clones (>10 cells with identical paired $\alpha\beta$ TCR sequences) with 33% and 19.4% respectively, consistent with clonal expansion and their effector gene expression. In contrast, the KIF23 (4) cluster is made up of some large clones, but has singletons (21.9%) and small clones (2-5 cells, 19.6%) making up most of its cells. Low TCR detection limits interpretation of CCR7 (1) and MALAT1 (2) clusters, with 83.1% and 85.5% of cells respectively lacking paired $\alpha\beta$ TCR sequences. Although the clusters are made up of a mixture of patients, batches and timepoints, both clusters are dominated by cells from few patients (CCR7 (1): 41.7% from patient 3071; MALAT1 (2): 34.9% from patient 3163, 34.5% from patient 3072), all of which have low paired TCR yields ($< 20\%$).

Cluster abundance differs between individuals primed by infection and those primed by vaccination (Figure 4.6), in particular in the frequency of the cytotoxic GNLY (0) subset, reflecting prior flow-cytometry[322] and single-cell work[253]. No clear patterns emerged of differential abundance or consistent shifts in cluster proportions between pre- and post-Omicron timepoints (Section D.1). However, clustering on the shared nearest neighbour graph of all cells may not capture shifts in individual or small sets of genes.

Owing to the unique nature of our cohort containing individuals with different priming status sampled for multiple epitopes, I used a `DESeq2` design of a grouped variable incorporating epitope specificity, pre- or post-Omicron status and priming exposure. This allows for two key comparisons. First, the evolution of Spike-specific cells in vaccine or infection

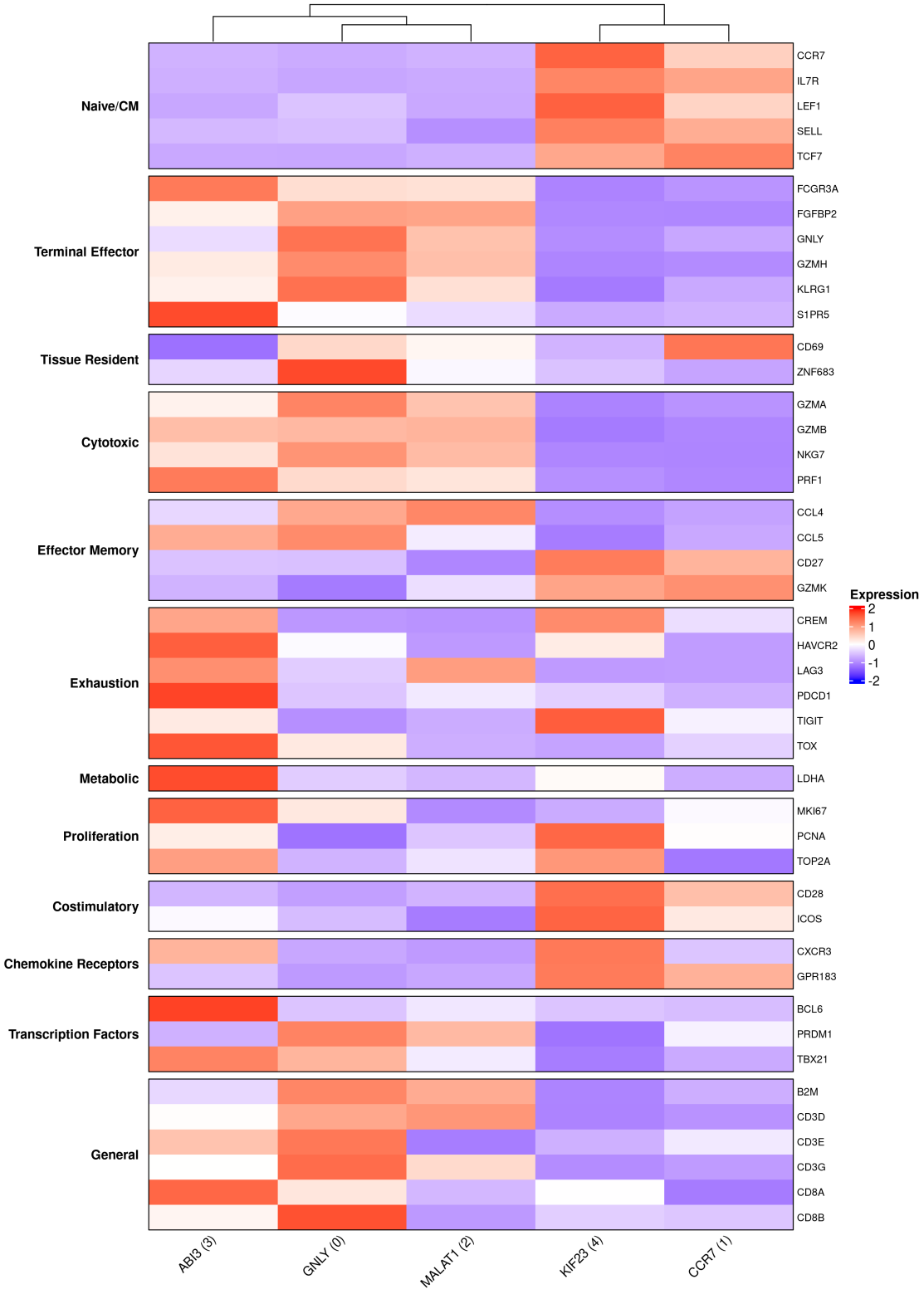


Figure 4.4: Canonical CD8⁺ T-cell markers amongst differentially expressed genes between clusters. Heatmap shows scaled average expression across transcriptional clusters of SARS-CoV2 specific CD8⁺ T-cells. Expression levels are calculated as the mean log-normalised counts per cluster and scaled across clusters. Columns are hierarchically clustered.

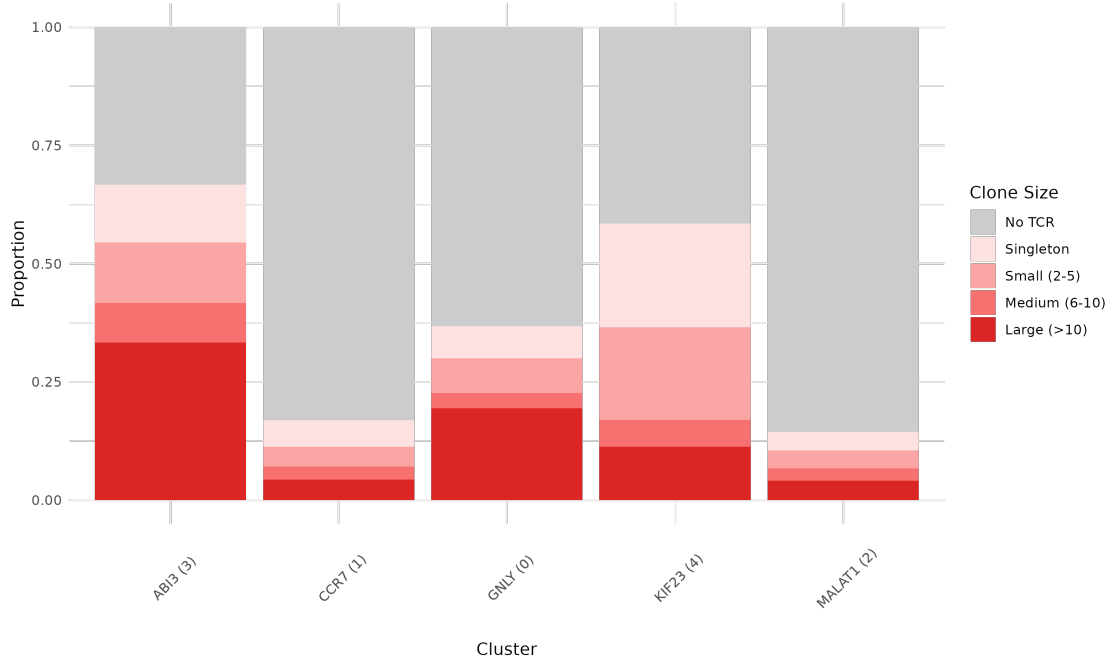


Figure 4.5: Stacked bar chart coloured by clone size categories within each transcriptional cluster. Clone sizes are calculated within each sample on the basis of paired $\alpha\beta$ amino acid TCR sequences. For visualisation, clone sizes were categorised as: singleton (1 cell), small (2-5 cells), medium (6-10 cells), or large (>10 cells).

primed individuals with Omicron infection. Second, Spike-specific cells in individuals primed by infection, sampled after their 3rd vaccination vs Spike-specific cells in individuals primed by vaccination after Omicron infection, comparing the same number of exposures (3x mRNA, 1x infection) but in different orders.

4.2.3 Differential gene expression between order of exposure and number of exposures

I performed differential gene expression analysis using DESeq2[338] on pseudobulked counts of samples. First, I compared pre- and post-Omicron samples of Spike-specific CD8⁺ T-cells in vaccine primed individuals. `DESeq2` identifies 36 genes highly expressed pre-Omicron and 581 genes highly expressed post-Omicron in vaccine-primed Spike-specific cells (Figure 4.8).

Gene set enrichment analysis of the genes against the Gene Ontology Biological process database identified no significantly enriched biological processes between these groups (Table D.2, Table D.1). Individual genes are differentially expressed but do not cluster into

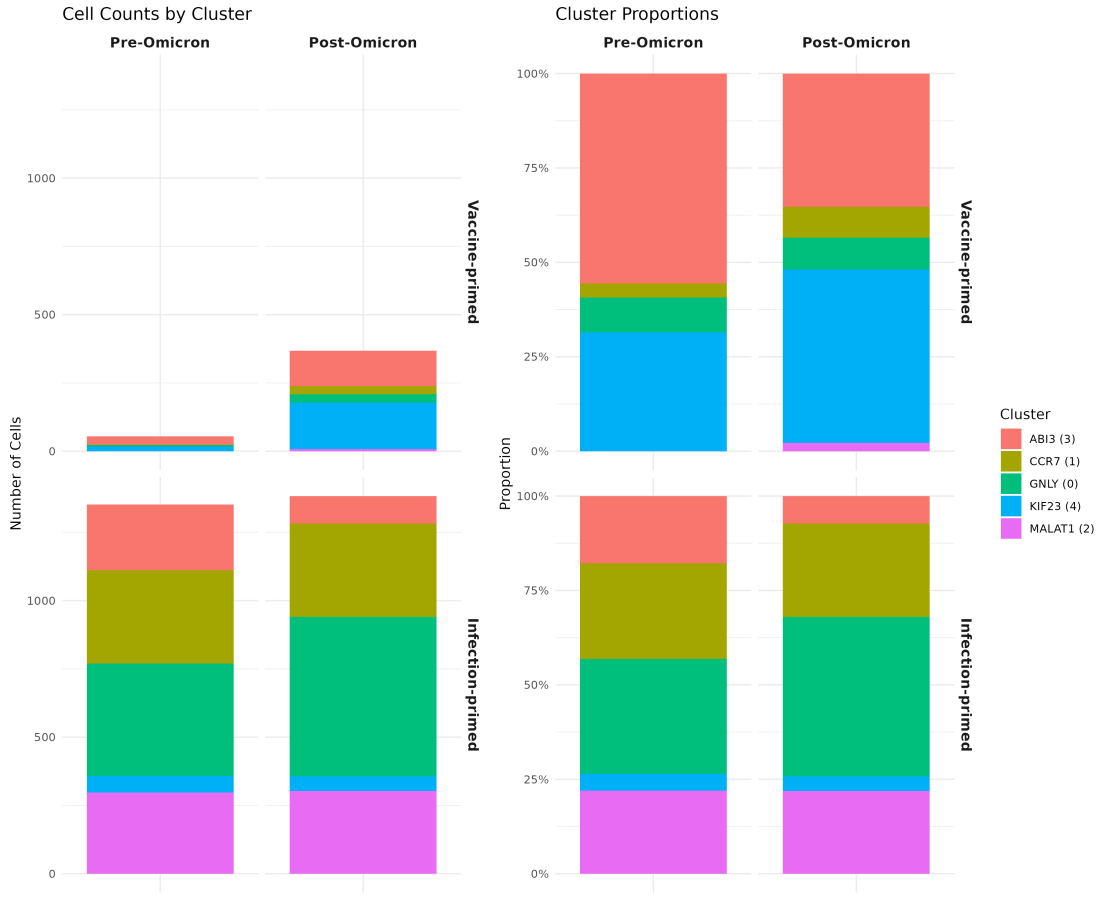


Figure 4.6: Cluster composition by priming exposure and timepoint. Stacked bar charts show the distribution of cell clusters separated by priming exposure (Vaccine-primed top row, Infection-primed bottom row) and timepoint (pre-Omicron left column, post-Omicron right column). The GNLY (0) cluster is more frequent in infection-primed individuals, while the ABI3 (3) cluster makes up a larger proportion of vaccine-primed cells.

biological processes consistently altered across patients, suggesting that the cellular response is heterogeneous with a sample size insufficient to detect consistent changes at the level of processes or pathways.

Compared to pre-Omicron samples, post-Omicron Spike-specific $CD8^+$ T-cells from vaccine-primed individuals show highly enriched expression of activation markers *CD38* (log2FC = 21.5, $p < 0.05$) and *DPP4* (log2FC = 23.6, $p < 0.05$), along with critical T-cell activation transcription factor *NFATC1* (log2FC = 21.6, $p < 0.01$) and T-cell development regulator *RUNX1* (log2FC = 22.4, $p < 0.001$). Post-Omicron cells additionally show enrichment of interferon-stimulated gene *IFI44L* (log2FC = 22.0, $p < 0.01$), T-cell survival

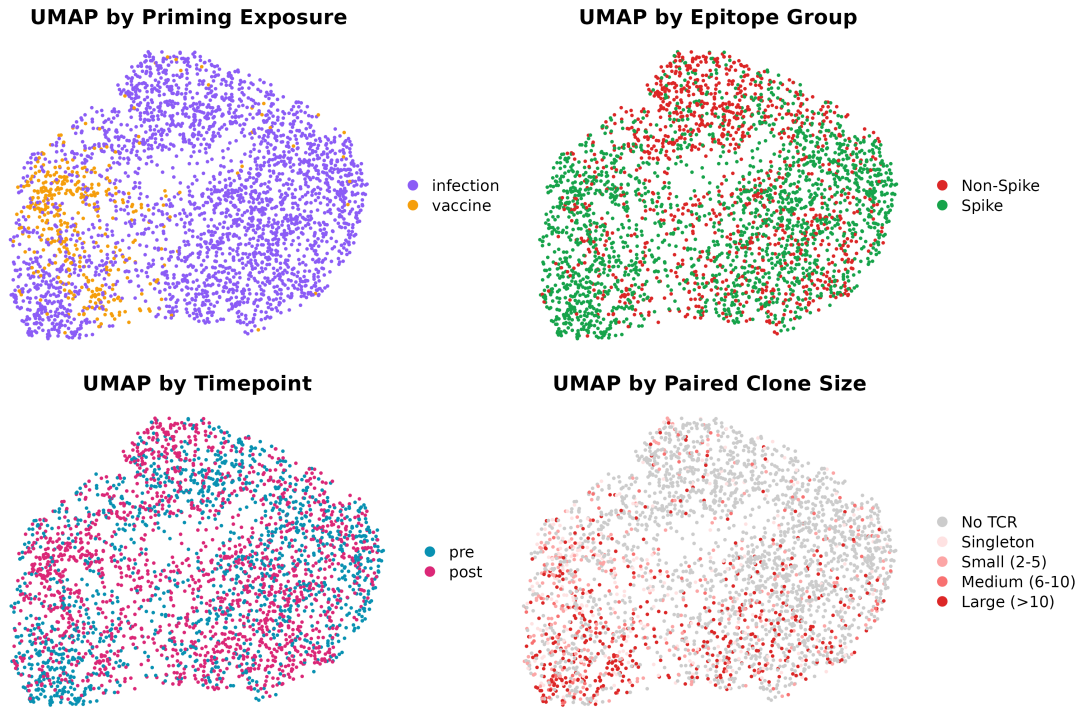


Figure 4.7: UMAP visualisation of 3157 SARS-CoV2 antigen-specific CD8⁺ T-cells coloured by metadata categories. (1) Priming Exposure: Infection-primed (purple) vs Vaccine primed (orange). (2) Epitope Group: Spike (green) vs Non-Spike (red). (3) Pre- or Post- Omicron Timepoint: Samples collected before breakthrough infection at the completion of the vaccine course (pre, cyan) vs Samples collected after breakthrough infection (post, pink). (4) Paired $\alpha\beta$ TCR clone size.

factor *GIMAP6* (log2FC = 21.1, $p < 0.05$), and cell cycle regulator *CDKN1B* (log2FC = 20.8, $p < 0.001$). In contrast, pre-Omicron samples show strong enrichment of terminal effector or cytotoxicity transcription factor *EOMES* (log2FC = -27.7, $p < 0.001$), and TCR signalling regulator *DUSP6* (log2FC = -24.3, $p < 0.001$).

Then, I carried out the same comparison in individuals primed by infection. 128 genes were enriched pre-Omicron and 86 genes were enriched post-Omicron in infection-primed Spike-specific cells (Figure 4.9). Again, gene set enrichment analysis identified no significantly enriched biological processes between these groups (Table D.4, Table D.3). Exploring the differentially expressed genes post-hoc, pre-Omicron samples show strong enrichment of proliferation marker *MKI67* (log2FC = 27.3, $p < 0.001$) and immune checkpoint molecule *CD274* (PD-L1, log2FC = 24.3, $p < 0.001$).

Three genes overlapped between those differentially expressed pre- or post-Omicron be-

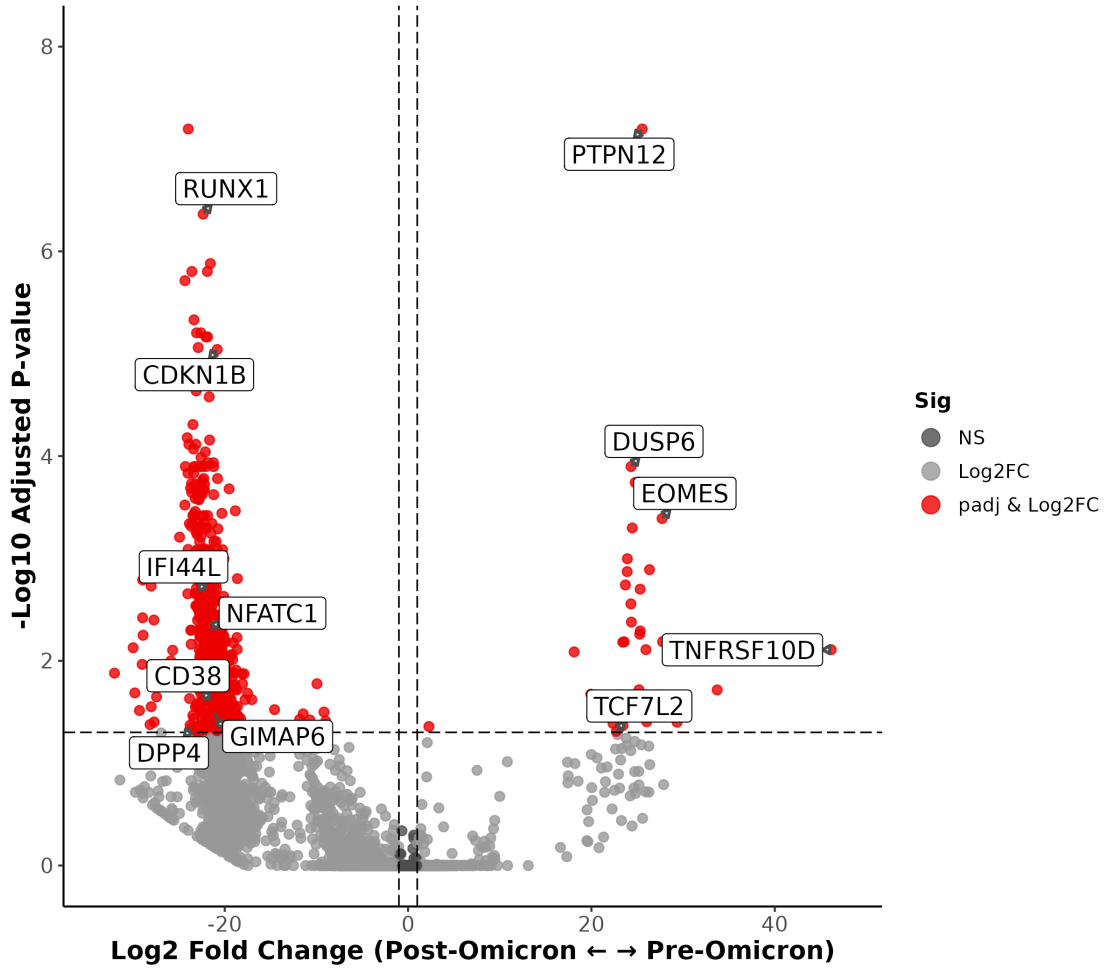


Figure 4.8: Volcano plot showing differential gene expression between pre-Omicron and post-Omicron Spike-specific CD8⁺ T-cells in vaccine-primed individuals. Points represent individual genes, with significantly differentially expressed genes highlighted. Manually identified genes relevant to CD8⁺ T-cell state are labelled, showing enrichment of effector/exhaustion markers (*EOMES*, *PTPN12*, *DUSP6*) pre-Omicron and activation/metabolic markers (*CD38*, *NFATC1*, *RUNX1*) post-Omicron.

tween vaccine primed and infection primed individuals. These were *SERPINI1* and *SBF2*, which were enriched post-Omicron in both groups, and *ZNF682*, which was enriched pre-Omicron in infection primed individuals and post-Omicron in vaccine primed individuals.

Finally, I investigated the genes differentially expressed in cells with four exposures but in different orders. Spike-specific cells primed by infection sampled pre-Omicron have had infection followed by three mRNA vaccinations (infection-first), in contrast, Spike-specific cells primed by vaccination sampled post-Omicron have had three vaccinations followed by infection (vaccine-first).

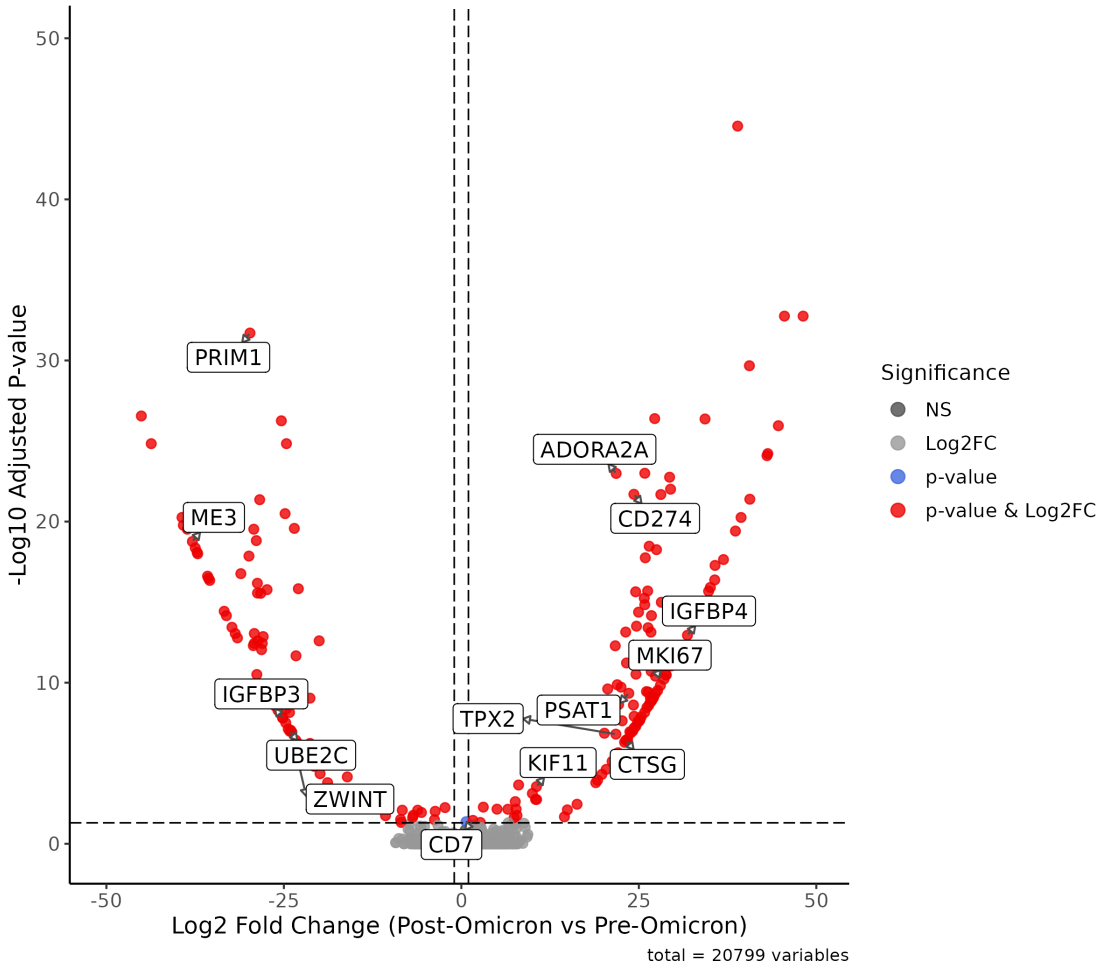


Figure 4.9: Volcano plot showing differential gene expression between pre-Omicron and post-Omicron Spike-specific CD8⁺ T-cells in infection-primed individuals. Points represent individual genes, with significantly differentially expressed genes highlighted. Manually identified genes relevant to CD8⁺ T-cell state are labelled, showing proliferation related genes and PD-L1 (*MKI67*, *CD274*).

Vaccine-first cells had significantly higher expression of 71 genes, whereas infection-first cells were enriched for expression of 280 genes (Figure 4.10). Again, no gene-sets were differentially enriched between these groups (Table D.6, Table D.5).

Spike-specific CD8⁺ T-cells primed by infection (sampled pre-Omicron) show high expression of terminal effector memory transcription factor *EOMES* (log2FC = 25.2, $p < 0.001$), chemokine *XCL2* (log2FC = 29.8, $p < 0.001$), alongside adhesion molecule *ALCAM/CD166* (log2FC = 22.4, $p < 0.001$) and negative regulator of TCR signalling *DUSP6* (log2FC = 24.7, $p < 0.001$). These cells additionally express terminal differentiation regulator *PRDM1/BLIMP-1* (log2FC = 1.93, $p < 0.05$), inflammatory cytokine *OSM* (log2FC = 31.2,

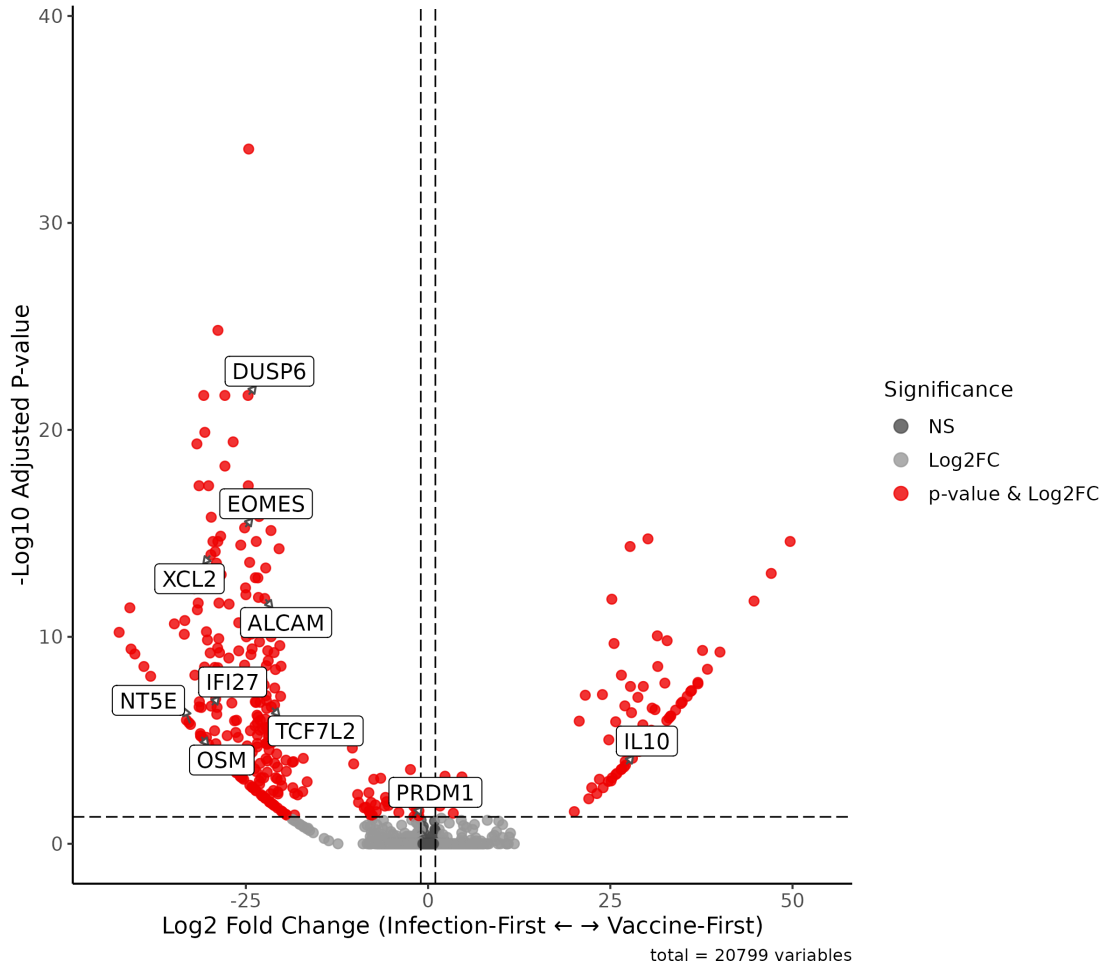


Figure 4.10: Volcano plot showing differential gene expression between cells with four Spike exposures in different orders. Points represent individual genes differentially expressed between vaccine-first (three mRNA followed by omicron infection) and infection-first (infection, then three mRNA), with significantly differentially expressed genes highlighted. Manually identified genes relevant to $CD8^+$ T-cell state are labelled, showing TCR negative regulator, $CD8^+$ memory formation and chemokine genes (*DUSP6*, *EOMES*, *XCL2*) and regulatory cytokine (*IL10*).

$p < 0.001$), and interferon-stimulated gene *IFI27* (log2FC = 29.8, $p < 0.001$). In contrast, vaccine-primed cells (sampled post-Omicron) show enrichment of immunoregulatory cytokine *IL10* (log2FC = 27.1, $p < 0.001$) and matrix metalloproteinase *MMP9* (log2FC = 28.9, $p < 0.001$).

Although no unifying biological processes were identified, the infection primed cells could be described as a more highly differentiated effector memory phenotype characterised by strong *EOMES* expression and effector function (*XCL2*). The co-expression of *DUSP6*, a negative feedback regulator, could suggest repeated or sustained TCR stimulation. Vaccine

priming appears to maintain cells in a less terminally differentiated state with regulatory capacity (*IL10*).

4.2.4 T-cell receptor repertoire diversity is similar across 1 to 5 exposures

I aggregated paired $\alpha\beta$ TCRs within each sample to explore how the repertoire evolves with different exposures (Figure D.6). Shannon diversity index was stable across timepoints (Figure 4.11) in infection and vaccine primed samples. Shannon diversity did not differ by number of exposures (Kruskal-Wallis, $\chi^2 = 3.6527$, $p = 0.455$).

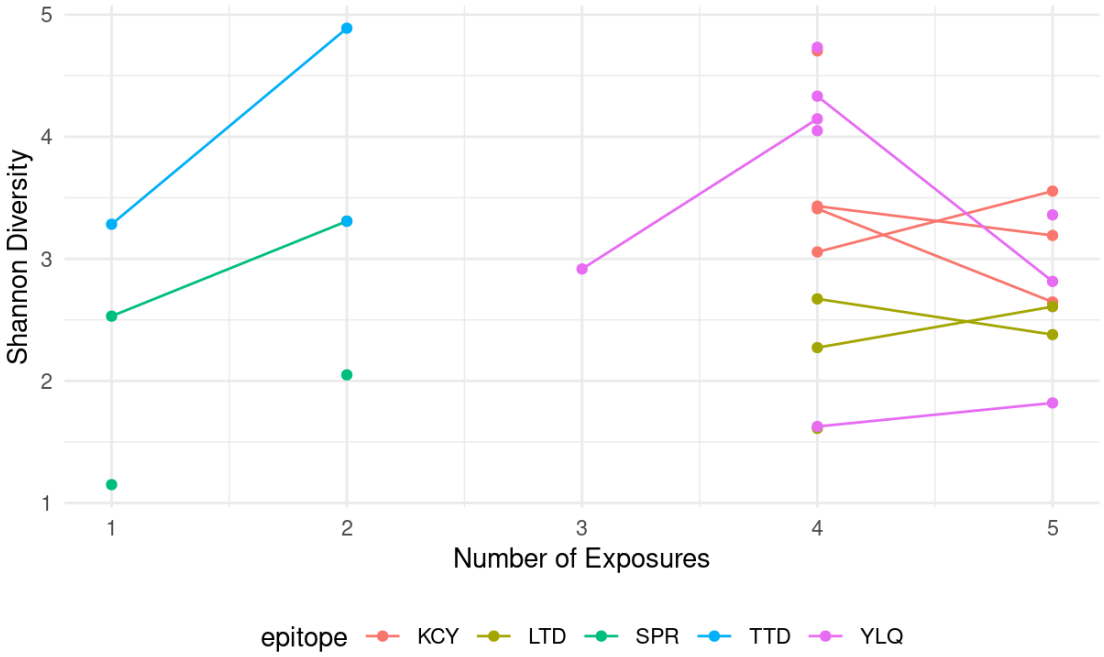


Figure 4.11: Shannon diversity index of paired $\alpha\beta$ TCR sequences within each sample, coloured by epitope target and arranged by the inferred number of exposures. Samples where fewer than five paired TCRs were detected were excluded, leading to analysis of 28 samples including 8 pairs.

Both paired samples involving infection primed, Non-Spike specific CD8⁺ T-cells increased in TCR repertoire diversity after Omicron infection (Figure 4.12), but these changes were not statistically significant.

4.2.5 Clonal persistence is associated with more clone members in Effector-like than Memory-like clusters

In this longitudinal dataset, we are able to track clones across timepoints to identify clonotypes that are newly detected, maintained and not detected again. Exploring the dimen-

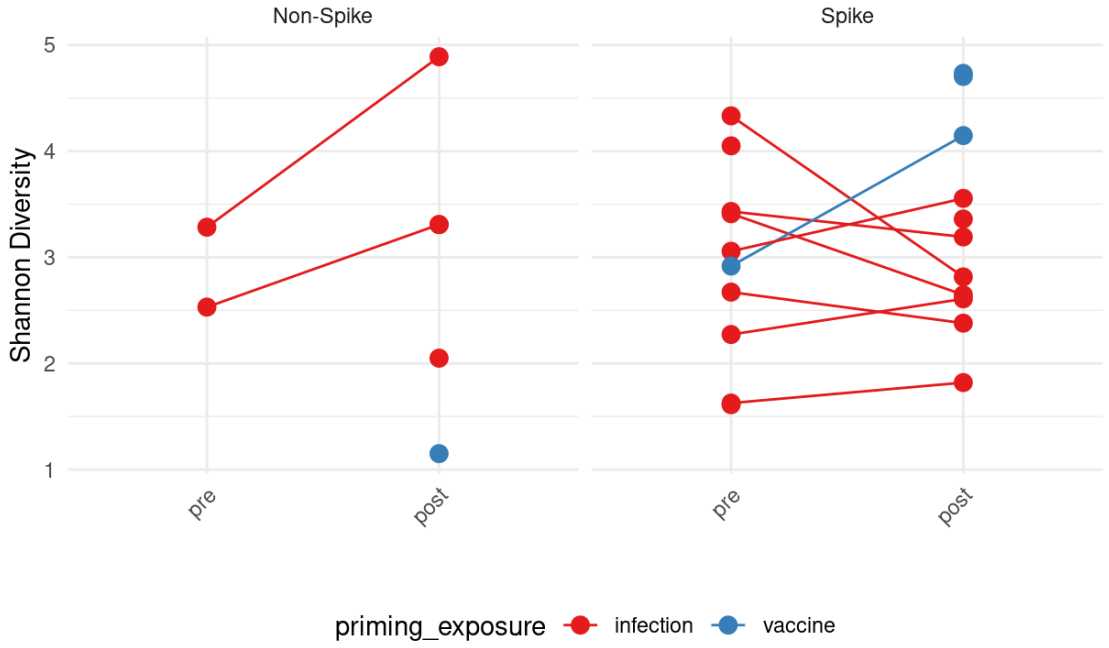


Figure 4.12: Shannon diversity index of paired $\alpha\beta$ TCR sequences within each sample, faceted by epitope type (Spike or Non-Spike), and infection or vaccine priming. Samples where fewer than five paired TCRs were detected were excluded, leading to analysis of 28 samples including 8 pairs.

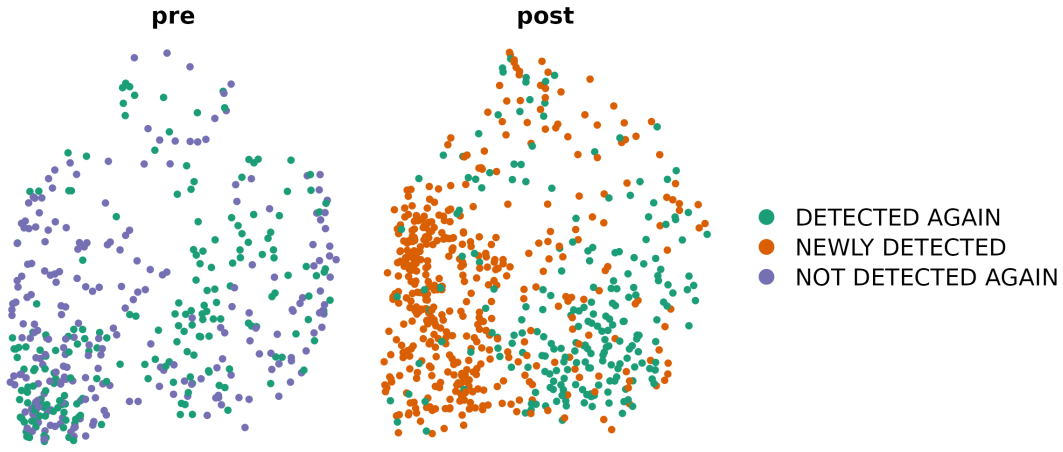


Figure 4.13: UMAP Dimensionality Reduction showing only cells with a detected paired $\alpha\beta$ TCR. Each cell is overlaid with its detection status (newly detected, maintained, not detected again) based on its presence at pre- and post-Omicron timepoints. Not detected again means no cell with the same paired TCR as that cell at the pre-Omicron timepoint is detected at the post-Omicron timepoint in the same individual specific for the same epitope. Maintained means, with the same constraints, that the same paired TCR is detected at both timepoints. Newly detected indicates that that clonotype was not detected in that individual and epitope at the pre-Omicron timepoint

sionally reduced set of cells, the distribution of clonotypes that are maintained, newly detected or not detected again after Omicron infection show spatial patterns on the UMAP (Figure 4.13). To investigate this further, I categorised cells by their transcriptional cluster and, for pre-Omicron samples, whether they were detected again after Omicron infection (Figure 4.14).

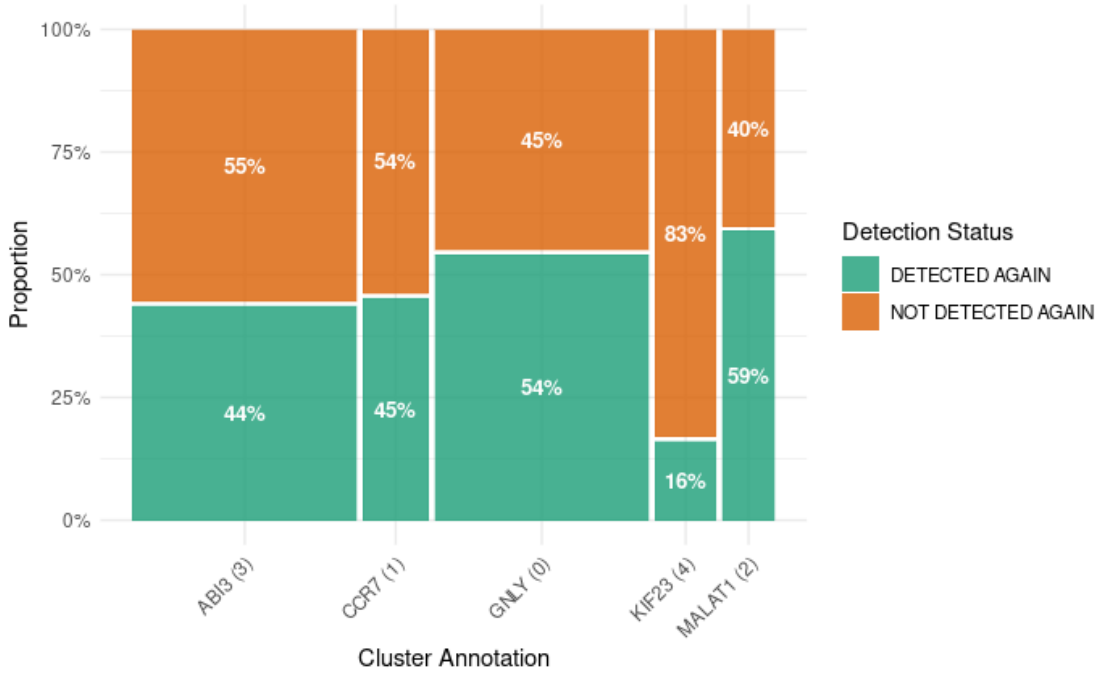


Figure 4.14: Mosaic of Cluster Annotation and Detection Status before Omicron. Mosaics represent all paired-chain TCR clonotypes detected in that cluster before Omicron infection. Mosaics are coloured by whether that clonotype sampled at the pre-Omicron timepoint was subsequently detected at the post-Omicron timepoint (detected again, green) or not (not detected again, orange).

MALAT1 (2) and GNLY (0) clusters have the highest proportions of clonotypes detected again at the post-Omicron timepoint. However, this does not account for the number of cells of that clonotype in each cluster, the number of cells sampled at each timepoint or the priming exposure. To account for this, I fitted a logistic regression model to paired samples where TCRs were detected to predict whether pre-Omicron clonotypes were detected again post-Omicron. Each clonotype specific for a given epitope in an individual was an element, and features included the number of cells with that clonotype in each cluster, priming exposure (vaccine vs infection), epitope type, and the total number of cells where a TCR

was detected post-Omicron (Figure 4.15). As illustration, some example values are shown in Table D.7.

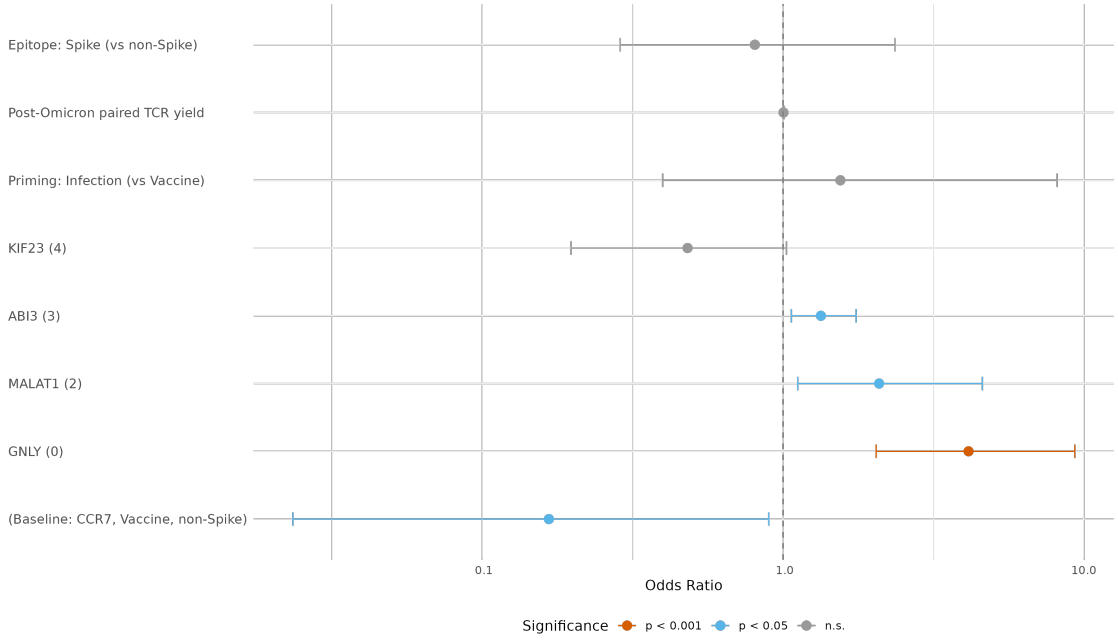


Figure 4.15: Forest plot showing associations with paired clonotype maintenance from pre- to post-Omicron timepoints. Points represent odds ratios (OR) with 95% confidence intervals from logistic regression. Cluster coefficients represent the effect of having cells in that cluster relative to the CCR7 (1) reference cluster. Model controlled for priming exposure, post-Omicron cell count, and epitope specificity.

The logistic regression model (Table D.8) fit the data moderately well (McFadden's R^2 = 0.24) and was significantly better than the null model (likelihood ratio test: χ^2 = 52.35, df = 7, p < 0.001).

Transcriptional cluster composition significantly influences clonotype maintenance. Clonotypes with more cells in the GNLY (0) cluster had substantially higher odds of being detected again compared to those with cells in the reference cluster (OR = 4.14, 95% CI: 2.04–9.33, $p < 0.001$). Similarly, cells in the MALAT1 (2) cluster were associated with increased maintenance (OR = 2.09, 95% CI: 1.12–4.60, $p = 0.038$), as were cells in the ABI3 (3) cluster (OR = 1.34, 95% CI: 1.07–1.75, $p = 0.021$).

In contrast, the KIF23 (4) cluster, with modestly raised expression of the proliferation marker *PCNA* ($\log_2$FC = 0.49, $p < 0.001$) alongside central memory markers *CCR7*, *SELL* and *IL7R*, showed a trend toward reduced odds of clonotype maintenance (OR = 0.48, 95%

CI: 0.20–1.03, p = 0.076). Neither priming exposure (infection vs vaccine, OR = 1.55, p = 0.557), post-Omicron cell count (OR = 1.00, p = 0.169), nor epitope specificity (Spike vs non-Spike, OR = 0.81, p = 0.685) showed significant associations with clonotype maintenance in this model. Post-Omicron TCR yields were adequate (Figure D.6), suggesting that once a baseline number of paired TCRs are detected, the clonotypes detected in the first sample are sufficiently likely to be detected again that differences in yield have little effect.

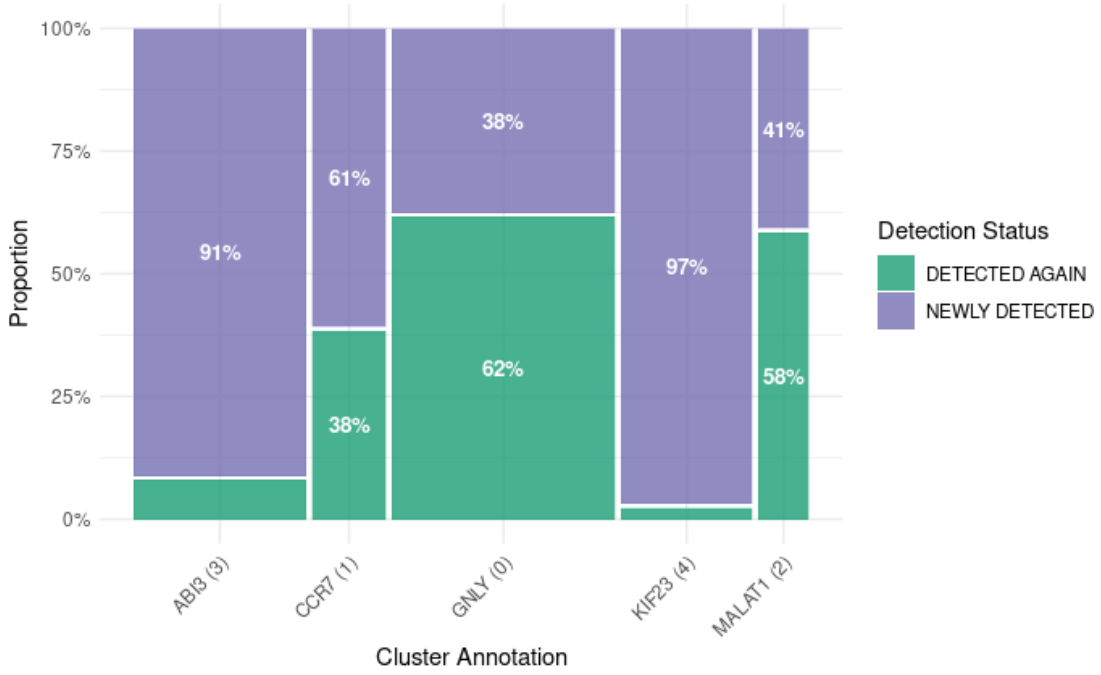


Figure 4.16: Mosaic of Cluster Annotation and Detection Status after Omicron. Mosaics represent all paired-chain TCR clonotypes detected in that cluster after Omicron infection. Mosaics are coloured by whether that clonotype sampled at the post-Omicron timepoint was previously detected at the pre-Omicron timepoint (detected again, green) or not (newly detected, purple).

When considering post-Omicron clonotypes, ABI3 (3) and KIF23 (4) have the highest raw rates of newly detected clonotypes (Figure 4.16).

4.3 Discussion

This work uses a longitudinal cohort of healthcare workers with single-cell sequencing of CD8⁺ T-cells directed at five epitopes to identify potential correlates and consequences of different exposures on the T-cell memory against SARS-CoV2. First, I identify 5 clusters

that reflect different T-cell subsets after vaccination and infection. Then, I identified differentially expressed genes between groups of cells likely to have experienced different exposures. Finally, I explored the longitudinal repertoires of $CD8^+$ T-cells specific to different epitopes, identifying phenotypic characteristics that associate with detecting that clonotype again at subsequent time points.

The transcriptional clusters describe cells with distinct $CD8^+$ T-cell states. The GNLY (0) cluster exhibited markers of TEMRA cells, with high expression of cytotoxic molecules (*GNLY*, *GZMH*, *GZMB*, *GZMA*) alongside the tissue residency marker *HOBIT* and the absence of lymphoid homing receptors (*CCR7* or *SELL*). This phenotype aligns with TEMRA clusters in studies of hybrid SARS-CoV2 T-cell memory, such as by Minervina et al.[253] and Mayer-Blackwell et al[382]. Both identify clusters with high expression of cytotoxic molecules as expanded by vaccination following infection. Large clones are most frequent in this cluster, supporting its identity as a differentiated effector population. These populations accumulate with age[383], and thus interpretation of their different abundance in priming groups is complicated by their different metadata.

Clusters CCR7 (1) and KIF23 (4) express central memory markers (upregulation of *CCR7*, *TCF7*, *SELL*). Where they differ is the expression of exhaustion and cytotoxic/effector markers, with KIF23 (4) expressing *TIGIT*, *CREM* and *GZMK*.

ABI3 (3) expresses *BCL6* - a transcription factor required for the generation of memory $CD8^+$ T-cells[384], alongside a short-lived effector signature (*LDHA*, *PRF1*, *GZMB*, *NKG7*, *HAVCR2*). This could reflect an intermediate state on the way to terminal effector differentiation, with the *LDHA* and *BCL6* signalling in opposing directions of memory or terminal effector formation[385,386].

Differential gene expression analysis suggests the sequence of infection and vaccination exposures, not simply the number of exposures, influences the transcriptional phenotype of $CD8^+$ T-cells. Breakthrough infection in vaccine primed individuals represents an activation and expansion event, showing strong upregulation of activation markers (*CD38*, *DPP4*) and activation related transcription factors (*NFATC1*, *RUNX1*) alongside interferon stimulated genes. $CD8^+$ T-cells primed by vaccination prior to breakthrough infection highly express

EOMES and *DUSP6*. *EOMES* may promote the formation of self-renewing memory cells rather than terminally differentiated cells[387], but breakthrough infection may have interrupted this process.

Infection primed individuals before breakthrough infection showed strong expression of *MKI67* and *CD274* (PD-L1), suggesting a proliferative and regulatory state. Breakthrough infection in individuals primed by infection does not further induce effector or proliferation genes. This aligns with Hornsby et al.'s findings in this cohort that flow-cytometry determined cytotoxic phenotypes cap out at three vaccine doses and are not further induced by breakthrough infection[322].

Comparing infection-first versus vaccine-first cells with four total exposures showed that infection-first cells had higher *EOMES*, *XCL2*, *ALCAM/CD166*, *DUSP6* and *PRDM1*. Taken together, the expression of *PRDM1* (BLIMP-1), a transcription factor driving terminal differentiation including expression of cytotoxic molecules, suggests that infection-primed cells are biased towards forming TEMRA-like cells. Vaccine-first cells are enriched for *IL10* expression, suggesting that they are biased towards a regulatory capacity.

These findings support that infection followed by vaccination produces activated, highly cytotoxic CD8⁺ T-cells with limited potential for further boosting. Minervina et al. similarly found that infection-vaccination order drives TEMRA differentiation[253]. This could underlie the finding that TEMRA phenotypes progressively grow to dominate the SARS-CoV2 specific circulating CD8⁺ compartment over 6 months to a year after infection[388].

Repertoire diversity metrics showed no significant differences between groups experiencing one to five antigen exposures. This is consistent with findings in SARS-CoV2 that epitope-specific repertoire diversity is maintained by the complex process of clonotype loss, recruitment and proliferation[389].

Pre-Omicron transcriptional state predicts clonotype persistence. Clonotypes where more cells were detected in GNLY (0) and ABI3 (3) clusters pre-Omicron were more likely to be detected post-Omicron, reflecting the impact of larger clone sizes but also differences in the persistence by transcriptional profile. Adamo et al. found persistent clones one year after

priming infection highly expressed IFN response genes and members of the TNF pathway at the acute timepoint. In contrast, non-persistent clones highly expressed *CTLA4*, *HAVCR2* and *MKI67*[388]. Other than *MKI67*, these genes were not significantly differentially expressed in any of our analyses, although their sampling began at 5 days post symptom onset after initial infection rather than mean 43.8 days after third vaccination, so this could reflect an earlier or specifically infection primed phenotype.

These findings have implications for vaccination strategy and our understanding of immune memory. Minimal transcriptional responses to breakthrough infection after three vaccine doses suggest that these populations are saturated in terms of effector differentiation and additional boosting might only have limited benefit. This is in line with meta-analysis of protection provided by hybrid immunity, where hybrid immunity provides "the highest magnitude and durability of protection", and that the influence of infection on immune phenotypes "might be able to extend the period before booster vaccinations are needed compared to individuals who have never been infected"[390].

The concern that sequential boosting narrows the response is not supported by these data. Shannon diversity did not differ across 1-5 exposures. However, repertoire narrowing is associated both with age and repetitive exposure over the long term[320], and so this study's short duration longitudinal cohort in younger individuals (mean 45.3 SD ± 9.3 years) may not capture this effect.

While this study provides a variety of findings, including correlates of clonotype maintenance and the consequences of different exposures on $CD8^+$ T-cell phenotypes, key limitations constrain its external validity. Most importantly the infection and vaccine primed individuals differ meaningfully in their clinical metadata (Table 4.1), and therefore while the identified differences could represent different states of antigen-specific $CD8^+$ T-cells, individuals with vaccine-primed samples have distinctive age and sex profiles that confound the analysis. For example, populations of TEMRA cells expressing cytotoxic markers accumulate with age[383], and females mount more robust T-cell activation during SARS-CoV2 infection[391].

Since patient of origin, priming exposure and metadata variables such as age and sex are collinear, it is not possible to disentangle their effects in this dataset. Additionally, the findings within this cohort reflect the evolution of SARS-CoV2 in a relatively small cohort of individuals outside the highest burden of COVID-19 disease with high occupational exposure. Therefore while these findings may support existing literature or provide hypotheses for directed testing, larger, pre-registered studies with matched cohorts are required to validate these findings.

Another important limitation is the influence the yield of paired T-cell receptor sampling has on the underlying conclusions. Repertoire diversity and the likelihood of clonotype maintenance are both significantly affected by the number of clonotypes sampled. The low TCR recovery in two clusters - CCR7 (1) and MALAT1 (2), limits interpretation of how clonal architecture and transcriptome interrelate. This appears to be patient or batch specific, as these clusters contain large contributions from individuals with low TCR yields. While quality control was stringent and required flow cytometric sorting for live/dead, $CD3^+$, $CD8^+$, $tetramer^+$ cells along with single-cell quality control metrics, technical factors may have influenced the yield of paired TCRs.

Longer followup of this cohort with further antigen exposures would demonstrate whether identified phenotypes are stable or evolve further. Do terminal effector enriched responses in infection-primed individuals eventually adopt more memory-like phenotypes? Do vaccine-primed individuals eventually develop more cytotoxic phenotypes as exposure to endemic SARS-CoV2 continues?

Expanding this work to larger, matched cohorts with balanced age, sex and clinical metadata would allow more definitive conclusions and potentially yield more interpretable biological processes than this small, heterogeneous cohort. Additionally, investigating tissue resident memory in parallel with the circulation would provide a more complete picture, including where “disappearing” clonotypes actually go.

While limitations including cohort size, demographic confounding and technical factors mean that these findings require cautious interpretation, this work contributes insight into

how sequential exposure to SARS-CoV2 shapes CD8⁺ T-cell memory, and identifies phenotypic correlates of clonal persistence that could be validated in larger cohorts.

4.4 Methods

4.4.1 Single-Cell RNA Processing

Single-cell count matrices were extracted as specified in Section 2.2.3. I used the `NormalizeData` function in Seurat with default parameters, and used the 1500 most variable genes for dimensionality reduction.

4.4.2 Dimensionality Reduction and Clustering

I performed dimensionality reduction using PCA on the 1500 most variable genes, selecting the first 6 principal components on the basis of the explained variance on an elbow plot. I clustered cells on the Nearest Neighbour graph built on the first 6 principal components using the `FindClusters` function in Seurat with a range of resolutions between 0.3 and 1.5. The final resolution for clustering was chosen on the basis of visual inspection of cluster stability across a range of resolutions.

Clusters were annotated by identifying the most differentially expressed genes in each cluster using the `FindAllMarkers` function in Seurat, filtering to those with a log2 fold change greater than 0.1, and difference in percentage of cells expressing that transcript compared with other cells of 0.1. This was done using default parameters, therefore using the Wilcoxon rank-sum test over scaled data.

4.4.3 Differential Gene Expression

Differential gene expression was performed using DESeq2[338] on pseudobulked counts of samples (epitope, patient, timepoint). The design formula integrated a group variable comprised of spike specificity and timepoint, and a patient variable. We corrected for multiple testing using the Benjamini-Hochberg method, and considered genes with adjusted p-values < 0.05 and $|log_2(\text{Fold Change})| > 0.1$ as differentially expressed.

4.4.4 Data Transformation for Clonal Persistence Analysis

Here, a clonotype is defined at the level of an individual and an epitope, this means it is defined as persisted if it is detected at both timepoints in the same individual for the same epitope.

To model the impact of transcriptional cluster composition on clonal persistence, we transformed the data into a wide format where each row represented a paired clonotype at the pre-Omicron timepoint. First, we excluded patient epitope combinations where the post-Omicron paired TCR yield was less than 5. For each clonotype, we identified the number of cells belonging to each transcriptional cluster. Cluster counts are compositional (each cell can only have one cluster) and sum to the total number of cells for that clonotype. We excluded CCR7 (1) from the model and used it as the reference cluster to avoid multicollinearity. Cluster coefficients therefore represent relative odds in comparison to CCR7 (1). This allowed us, for every clonotype, to assign a binary `maintained` variable, which was 1 if the clonotype was present at both timepoints, and 0 otherwise.

The logistic regression model included the number of cells in clusters ABI3 (3), GNLY (0), KIF23 (4), and MALAT1 (2), as well as covariates for priming exposure (vaccine vs infection), total post-Omicron cell count for that patient-epitope combination, and epitope specificity (Spike vs non-Spike). With this design, we could estimate how clonotype persistence is related to the cluster of those clones.

Logistic Regression of Clonal Persistence

We used the `glm` function in R to fit a logistic regression model to predict clonal persistence at the post-Omicron timepoint. The model included the following predictors: number of cells of that clonotype in each cluster, priming exposure, epitope specificity and post-Omicron paired TCR yield for that patient-epitope combination. The response variable was whether the clonotype was maintained (1) or not (0). We fit the model with binomial family.

5

The long term evolution of Spike specific CD4$^+$ T-cells compared to those directed at an immunodominant Membrane epitope after SARS-CoV2 infection.

Summary

With the endemicity of SARS-CoV2, understanding the long term evolution of T-cell memory in the context of heterogeneous immunity could help inform vaccination strategies and predicting or measuring immunity. Spike-specific CD4$^+$ T-cells can be recalled by both vaccination and infection, whereas non-Spike specific cells are only recalled by subsequent infections. Here, I analyse single-cell sequencing data from CD4$^+$ T-cells isolated using Class II MHC tetramers to characterise Spike-specific and non-Spike specific T-cells from 16 individuals at 1-3 months and four years post-infection. We sequenced 3,429 single cells, and we clustered cells into seven clusters on the basis of their gene expression. Early timepoints were dominated by clusters representing recently activated, central memory or interferon-stimulated states (*LMNA*, *TIGIT*, *IL2*, *IFIT3*), while late timepoints were dominated by potentially quiescent cells or cytotoxic phenotypes. In the 37 epitope-specific repertoires studied, non-Spike specific repertoires had significant reduction in Shannon diversity over four

years (a decrease of 1.45 in Shannon diversity, p = 0.002) whereas Spike repertoires did not. The strongest predictor of clonotype persistence was clone size at the early timepoint (OR = 8.3 per log-fold increase, p = 0.007). Spike specific repertoires showed greater clonality at the early timepoint. At four years post-infection, Spike-specific cells maintained expression of interferon-stimulated genes and PD-1, while expressing cytotoxic molecules (*GZMB*, *GZMH*), alongside co-stimulatory receptors and migratory chemokine receptors. Non-Spike specific cells exhibited more quiescent phenotypes with higher expression of anergy-associated genes (*EGR2*, *CREM*). Overall, this supports continued vaccination to maintain the diversity of Spike-specific $CD4^+$ T-cells, while highlighting the potential vulnerability of non-Spike specific cells to clonal contraction.

5.1 Introduction

The durability and evolution of $CD4^+$ T-cell memory to SARS-CoV2 infection informs our knowledge of long-term immunity and the consequences of booster vaccinations and breakthrough infections. SARS-CoV2 infection elicits $CD4^+$ T-cells that facilitate antibody production, enhance other adaptive immune cells and calibrate innate immunity[17,270]. The frequency and state of $CD4^+$ cells associate with reduced COVID-19 disease severity in bronchoalveolar lavage fluid[198], as well as in peripheral blood[179,180]. SARS-CoV2-specific $CD4^+$ T-cells persist as memory populations that can be identified at 2-years after initial infection[319].

In memory, these cells have decaying frequencies but diverse memory phenotypes. Nelson et al. followed $CD4^+$ T-cells specific to two Spike and two Nucleocapsid epitopes in a cohort of 40 individuals for up to 10 months, finding that the half-life of tetramer-specific cells ranged between 122 and 165 days[392]. Where response-based methods are used, such as AIM assays for $CD137^+OX40^+$ $CD4^+$ T-cells, Dan et al. estimate an overall half life of 94 days, with 92% of the cohort maintaining detectable $CD4^+$ responses to SARS-CoV2 peptide pools at 6 months[270]. Nelson et al's work used multiparameter flow cytometry to categorise cell subtypes and compare their dynamics, finding that PD-1^{hi}CXCR5^+ cT_{fh} cells decayed the fastest due to their transition to PD-1^{lo} cells. Dan et al find stable cT_{fh} populations after a contraction of the "recently activated" PD-1^{hi}CXCR5^+ cT_{fh} population

at 6 months. In terms of more conventional subtypes, Jung et al. used AIM and found that T_{cm} ($CCR7^+CD45RA^-$) cells made up approximately half of SARS-CoV2 peptide responsive cells throughout their sampling range (1-317 days), whereas T_{em} cells ($CCR7^-CD45RA^-$) increased from less than 25% to 35% from the onset of symptoms to 60 days after[393]. Zuo et al. performed detailed characterisation of the functional profiles of these phenotypes via ICS and ELISPOT supernatants at six-months post infection, finding that the cytokine profile was indicative of Th1 subsets with IL-2, TNF and IFN-γ the dominant cytokines[266].

These cells maintain their functionality upon re-stimulation, with Jung et al. and Guo et al. finding via ICS at 10-months and 2-years respectively that polyfunctional cytokine responses can be elicited by maintained $CD4^+$ T-cells. An important and consistent finding from the first few months after SARS-CoV2 infection is that the frequency of SARS-CoV2 responsive $CD4^+$ T-cells is increased in individuals who had more severe prior infections[95,179,394]. At the six month timepoint, Zuo et al. similarly suggest that the priming severity provides a "set point" of cellular immunity on the basis of their results that symptomatic infections had 50% higher ELISPOT responses than asymptomatic ones at 6 months. However Jung et al. found no differences in IFN-γ ELISPOT resolved T-cell frequencies between moderate, severe and critical cases beyond 200 days from symptom onset, suggesting that either time or repeated exposures after priming could normalize responsive T-cell frequencies.

Divergent exposures of Spike-specific and non-Spike specific T-cells to vaccines and infections shape their phenotypes and functionality. For example, at 6 months post infection, significantly more non-spike responsive $CD4^+$ T-cells were $IL\text{-}2^+$ than spike responsive $CD4^+$ T-cells when measured by ICS[266]. However, at 1 and 2-years Guo et al. found that when measured via ELISPOT counts or ICS, cytokine profiles and frequencies did not differ between Spike and non-Spike specific $CD4^+$ T-cells[319].

Single-cell studies of SARS-CoV2-specific T-cells link phenotype to specificity, and allow long-term tracking of individual clones. Wragg et al. developed a class II HLA-DRB1*15:01-S_{751} tetramer to track an immunodominant Spike epitope in 21 individuals from 23-450 days after symptom onset[274]. Using this, they could resolve distinct stages of contraction at

the single-epitope level, finding an early ($T_{1/2}$ of 20) and late ($T_{1/2}$ of 377) rate of decay, with the frequency stabilising at 0.0038% of $CD4^+$ cells. This approach also allows more detailed, ex-vivo characterisation of the phenotypes where, in contrast to AIM work, $CCR6^+$ cells were rare and $CXCR3^+$ cells grew with time. This relates to other work by this group identifying a TCR-independent activation within classical AIM methods[271]. Looking into a different epitope, S_{236}, they find that despite both epitope specific populations having similar frequencies in early convalescence, their phenotypes are strikingly different, with naive phenotypes much more frequent in S_{236} specific cells. They find dominant TRBV24-1 usage, with five public clonotypes shared between individuals, and in one individual identify that nine post-infection clonotypes characterise 20% of the repertoire after subsequent vaccination. By simultaneously studying multiple epitopes longitudinally, they were able to detect subdominance patterns and poor recruitment of one epitope of the same protein compared to another, as well as explore the dynamics of clonotypes after multiple exposures.

The longer-term evolution of SARS-CoV2-specific $CD4^+$ T-cells remains unclear. Studies of phenotype and function have been limited to at maximum two years after first infection, and studies have generally relied on individual surface markers or ICS to annotate cells. The transcriptional profiles of ex-vivo Spike-specific and non-Spike specific $CD4^+$ T-cells long after infection are unknown. This is of particular interest because Spike specific $CD4^+$ T-cells are recalled by approved SARS-CoV2 vaccines, that only include Spike antigens, whereas non-Spike specific $CD4^+$ T-cells are only recalled by subsequent infections. Initial priming exposure can imprint phenotypes - Nelson et al. found that milder cases have higher flow-cytometry-defined cT_{fh} and Th1 proportions at less than 60 days after symptom onset. Gray Gaillard et al. investigated vaccine versus infection imprinting by studying 127,000 $CD4^+$ T-cells after AIM against Spike peptides[395]. Individuals primed by infection approximately 20 months earlier and subsequently vaccinated had Spike-responsive $CD4^+$ T-cells which highly expressed Interferon Stimulated Genes (ISGs) and Cytotoxicity and Chemokine genes (*GZMB*, *CCL3-5*), whereas those primed by vaccination 9 months earlier expressed higher *IL2*, *CCR7* and NF-κB pathway genes. Subsequent booster vaccinations in infection primed individuals were insufficient to alter the profiles of infection primed cells, whereas individual clonotypes with vaccine primed phenotypes were subsequently

detected with infection primed phenotypes after breakthrough infections. This suggests that infection is more "potent" at imprinting the phenotype of Spike-specific CD4⁺ T-cells, and that this is maintained over time. The Gray Gaillard study, in making this comparison also highlights the utility of TCR sequencing to follow clonotypes and their evolution with time.

Vaccination diversifies the TCR repertoire by expanding pre-existing memory clones and recruiting cells from naive pools. Dykema et al. used a response based assay to show that in convalescent COVID-19 patients (median days since symptoms: 101) vaccination induced mainly new, vaccine-induced clones "distinct from the repertoire induced by natural infection"[396]. Ford et al identified Spike-specific CD8⁺ clonotypes using a response-based assay, then tracked the TCRβ of those clonotypes in longitudinal bulk repertoire sampling over five timepoints including two vaccinations. "Vaccinations led to large expansions in memory T_S cell clonotypes... while also eliciting diverse T_S clonotypes not observed before vaccination"[397], supporting maintained repertoire diversity with these exposures. Studies of repertoire narrowing in endemic respiratory viral exposure over many years[320], raise whether SARS-CoV2 repertoires will follow a different trajectory, or whether two years of SARS-CoV2 studies is sufficient to detect this phenomenon.

These studies taken together raise questions about the maintenance of CD4⁺ memory phenotypes and clones in the longer term, past 2 years, but also about how non-Spike specific CD4⁺ T-cells, which are not recalled by UK approved vaccines, evolve over time. Direct comparison of Spike and non-Spike specific CD4⁺ T-cells is challenging, owing to different cohorts and non-overlapping HLA alleles of immunodominant epitopes. In this study, we use Class II MHC tetramers to isolate Spike-specific and non-Spike specific CD4⁺ T-cells from a cohort of 16 donors initially infected with SARS-CoV2 between May 2020 and February 2021. Longitudinally sampling these individuals, at both 1-3 months and four years after initial infection, we use single-cell RNA sequencing to characterise the transcriptomes and TCR repertoires of Spike-specific and non-Spike specific CD4⁺ T-cells. Following cell populations with different targets longitudinally allows us to compare the evolution of Spike-specific and non-Spike specific CD4⁺ T-cells within the same individuals,

and provides insight into how Spike-specific and non-Spike-specific repertoires differ when sampled four years post-infection.

5.2 Results

5.2.1 Study cohort and epitope-specific responses

Following SARS-CoV2 infection, we and others have identified dominant responses to conserved SARS-CoV2 epitopes $S_{166\text{-}180}$, $S_{751\text{-}765}$, $S_{866\text{-}880}$, as well as the membrane epitope $M_{172\text{-}188}$[95,148,203,273,362,398]. Subsequent work by my colleagues using IFN-γ ELISPOT assays and MHC-II tetramer staining has identified the restricting HLA alleles for these epitopes, summarised in Table 5.1 (see also Section 3.2.1).

Table 5.1: The HLA restrictions of SARS-CoV2 epitopes relevant to tetramer sorting and cell isolation in this study. Determined by, among others, Dr Yanchun Peng, Dr Guihai Liu and Dr Xuan Yao.

Epitope	HLA Restriction
$S_{166\text{-}180}$	HLA-DRB1*15:01
$S_{751\text{-}765}$	HLA-DPB1*04:01
$S_{866\text{-}880}$	HLA-DRB1*15:01
$M_{172\text{-}188}$	HLA-DRB1*01:01, HLA-DRB1*07:01

As part of our studies of convalescent SARS-CoV2 infection, my colleagues assembled a cohort of individuals infected with SARS-CoV2 in the early stages of the pandemic (May 2020-February 2021)[95] to profile immune memory to SARS-CoV2. This included study of immunodominant SARS-CoV2 membrane epitope $M_{172\text{-}188}$ (Section 3.2.1), as well as Spike epitopes $S_{166\text{-}180}$ (`CTFEYVSQPFLMDLE`), $S_{751\text{-}765}$ (`NLLLQYGSFCTQLNR`) and $S_{866\text{-}880}$ (`TDEMIAQYTSALLAG`). My colleagues performed single-cell sequencing of Tetramer sorted $CD4^+$ T-cells from 16 individuals. The sampling groups are summarised in Table 5.2. The cohort includes 2 individuals with responses to both Spike and non-Spike epitopes, 4 individuals with responses to non-Spike epitopes only, and 10 individuals with responses to Spike epitopes only. Importantly, the distribution of WHO COVID-19 severity classification differs between these groups, with all non-Spike only individuals having had severe infections, whereas the

Spike only group includes 2 mild, 4 severe and 4 critical cases (Figure 5.1). The median age of the non-Spike only group is lower than the Spike only group (49.5 vs 61.5 years), and the both epitopes group is older still (71 years). All individuals in this cohort were infected prior to any vaccinations being available in the UK. We were not able to document and assemble vaccination and infection histories between the early and four year timepoints, meaning that we could not evaluate the relationship between the observed changes and documented exposures.

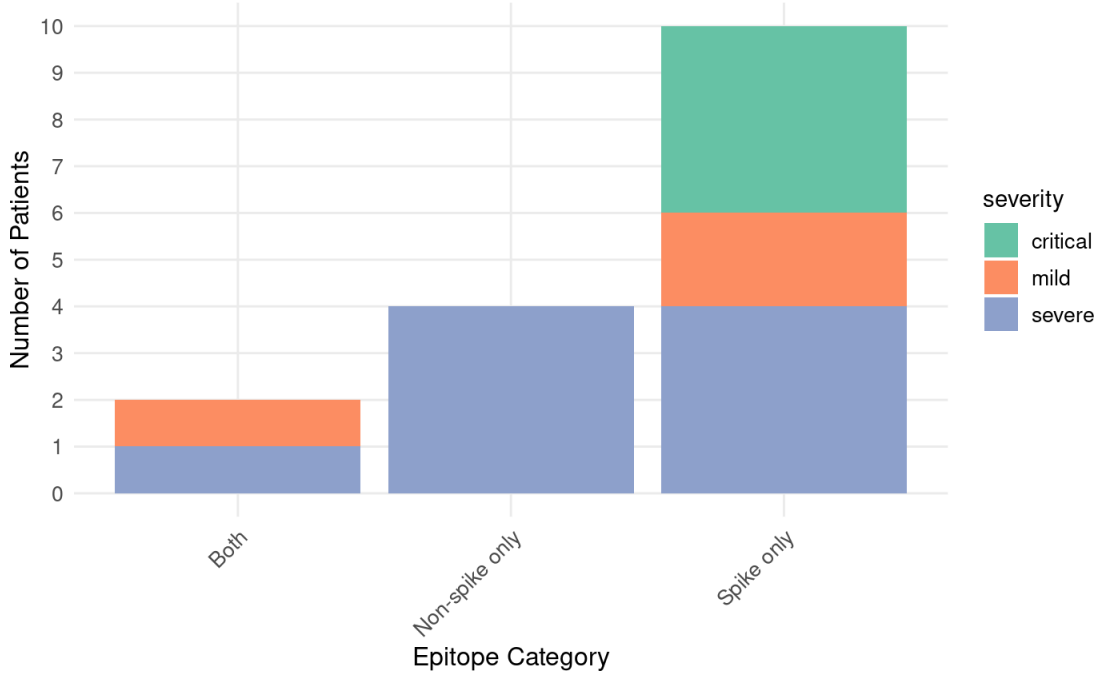


Figure 5.1: Bar chart of severity distribution by epitope sampling group. "Both" refers to individuals where Spike and non-Spike epitopes were sampled. "Non-Spike only" refers to individuals where only non-Spike epitopes were sampled. "Spike only" refers to individuals where only Spike epitopes were sampled.

Table 5.2: Summary of the sampling groups in this study, with the number of donors, median age (with interquartile range) and distribution of sex and severity of infection. Mod, Sev, Crit stand for Moderate, Severe and Critical respectively.

Epitope Category	Number of Donors	Median Age (IQR)	Sex Ratio	Severity Distribution	Median Cells (IQR)
Both	2	71 (67.5-74.5)	1M/1F	1 mild, 0 mod, 1 sev, 0 crit	482.5 (424.75-540.25)

Epitope Category	Number of Donors	Median Age (IQR)	Sex Ratio	Severity Distribution	Median Cells (IQR)
Non-Spike only	4	49.5 (45.75-58.25)	4M/0F	0 mild, 0 mod, 4 sev, 0 crit	232.5 (175.75-292)
Spike only	10	61.5 (54.75-72.25)	7M/3F	2 mild, 0 mod, 4 sev, 4 crit	115 (88.5-210.5)

5.2.2 Transcriptional profiles of Spike and Non-Spike specific CD4⁺ T-cells

3,429 cells across 40 samples from 16 individuals passed quality control (quality control metrics, including the number of features, genes and percentage mitochondrial reads are summarised in Section E.1). The median number of included cells per sample was 86.5, with the largest sample yielding 211 cells and the smallest sample yielding 2 cells (Figure 5.2).

Dimensionality reduction and clustering revealed seven clusters of cells (Figure 5.4). Clusters were annotated on the basis of their unique most significantly differentially expressed genes (lowest `p_adj`), given a `log2FoldChange > 1`, excluding ribosomal or mitochondrial genes. Canonical CD4⁺ T-cell marker gene expression (Figure 5.3) facilitated cluster interpretation. The clusters were annotated as follows: LTB (0), LMNA (1), TIGIT (2), IL2 (3), ZNF165 (4), FGFBP2 (5), IFIT3 (6).

The frequency of these clusters within a sample associates with the timepoint and epitope-specificity of that sample. LMNA (1), TIGIT (2) and IL2 (3) clusters are almost entirely found in the early timepoint, with 99.6%, 90.0% and 100% of cells of respective clusters from the 1-3 month timepoint. The IFIT3 (6) cluster is also enriched for the early timepoint at 83.5%. In contrast, LTB (0) is almost entirely made up of 4 year timepoint cells (94.3%). FGFBP2 (5) and ZNF165 (4) clusters are also enriched for the late timepoint at 84.3% and 88.7% respectively (Figure 5.5). In terms of their association with epitope specificity, Spike-specific samples have higher frequencies of LMNA (1) and FGFBP2 (5) cells, whereas Non-Spike specific samples have higher frequencies of TIGIT (2) and ZNF165 (4) cells (Figure 5.6). Clusters differed in clonal expansion. In particular, the FGFBP2 (5) cluster is the only cluster

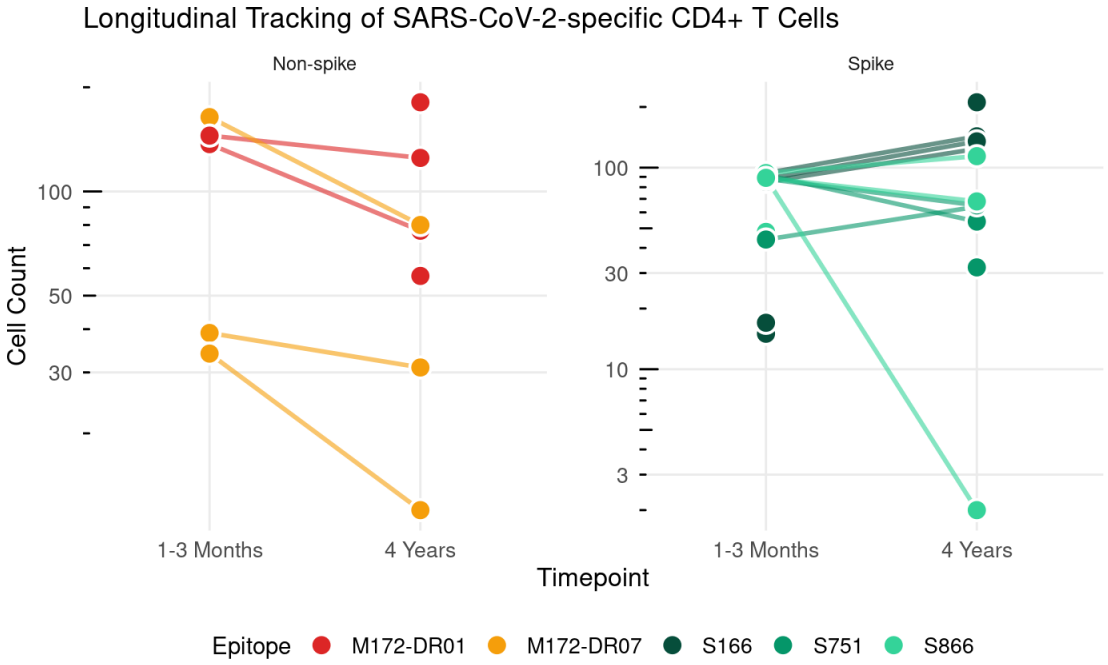


Figure 5.2: A linked dot plot showing cell counts for Spike and non-Spike specific CD4⁺ T-cells. Non-Spike specific cells are shown in red ($M_{172\text{-}188}$ presented by HLA-DR1) and yellow ($M_{172\text{-}188}$ presented by HLA-DR7). Spike-specific cells are shown in shades of green (S_{166}, S_{751} and S_{866})

where large paired clonotypes (>10 cells) are observed (Figure 5.8), and as a result shows lower Shannon diversity indexes than the other clusters (Figure 5.9).

The combination of cluster markers and dynamics between timepoints allow putative phenotypes to be assigned to each cluster. The LTB (0) cluster most differentially expresses *LTB* ($\log_2$FC = 3.0, $p < 0.001$), *TXNIP* ($\log_2$FC = 2.4, $p < 0.001$), *KLF2* ($\log_2$FC = 1.9, $p < 0.001$), *JUNB* ($\log_2$FC = 1.9, $p < 0.001$) and *IL32* ($\log_2$FC = 1.8, $p < 0.001$). There is significant upregulation of *CD69* ($\log_2$FC = 0.6, $p < 0.001$) but with a smaller $\log_2$FC. It also expresses *CCR7* at lower levels than other cells ($\log_2$FC = -1.4, $p < 0.001$). The cluster shows negative differential expression of activation marker *CD44* ($\log_2$FC = -1.8, $p < 0.001$), exhaustion-associated *CREM* ($\log_2$FC = -2.2, $p < 0.001$), and markers associated with recent activation including *LMNA* ($\log_2$FC = -4.7, $p < 0.001$), *SKIL* ($\log_2$FC = -3.3, $p < 0.001$) and *NR3C1* ($\log_2$FC = -2.5, $p < 0.001$). Taken together, the LTB (0) cluster likely represents a quiescent, long-lived memory population, potentially representing TEMRA cells.

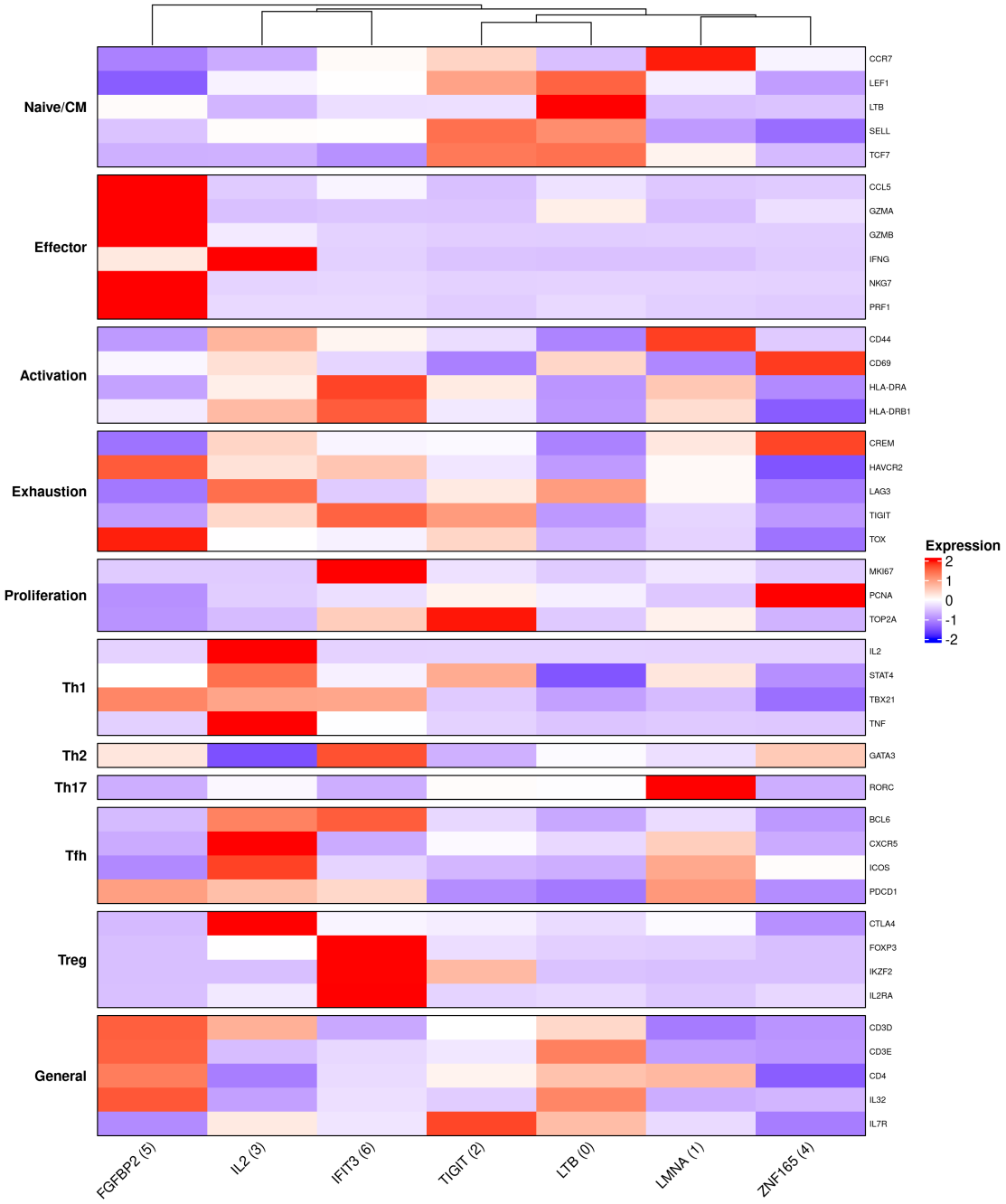


Figure 5.3: Heatmap of manually curated genes characterising cell clusters. Colours reflect the average expression of gene over all cells in the cluster scaled and centred across clusters. Dendrogram above the heatmap represents hierarchical clustering over the scaled and normalised expression matrix.

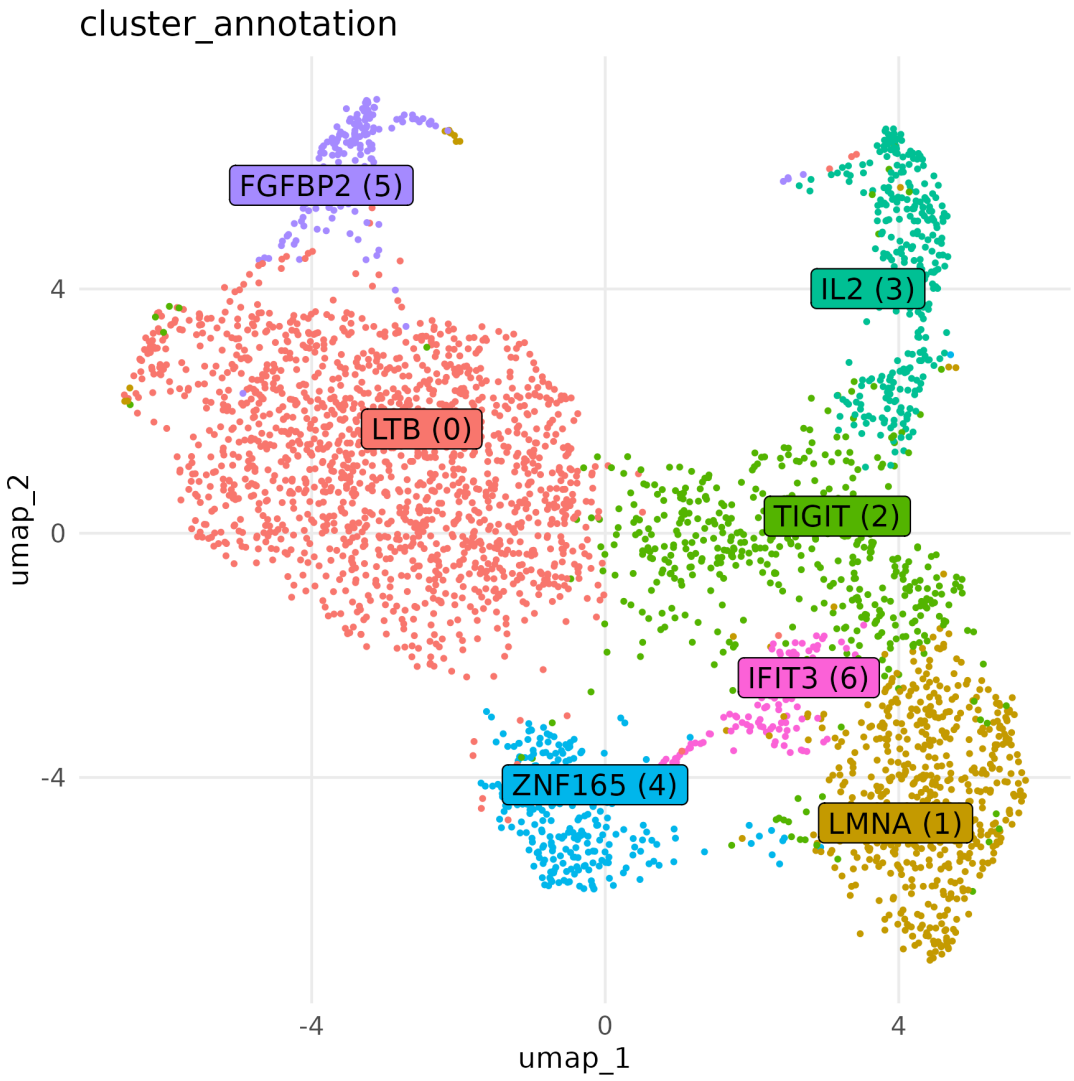


Figure 5.4: UMAP Projection of 3429 Sampled CD4⁺ Single Cells, coloured by cluster.

The LMNA (1) cluster most differentially expresses *LMNA* ($\log_2$FC = 4.5, $p < 0.001$), *PPP1R15A* ($\log_2$FC = 3.0, $p < 0.001$), *MYADM* ($\log_2$FC = 2.9, $p < 0.001$), *ARID5A* ($\log_2$FC = 2.9, $p < 0.001$). This cluster shows the highest expression of *CCR7* ($\log_2$FC = 1.9, $p < 0.001$) and activation marker *CD44* ($\log_2$FC = 1.7, $p < 0.001$), alongside modest upregulation of Th1 marker *STAT4* ($\log_2$FC = 0.35, $p < 0.001$) and Tfh marker *ICOS* ($\log_2$FC = 0.92, $p < 0.001$). Conversely, it shows reduced expression of *LTB* ($\log_2$FC = −2.9, $p < 0.001$), *TXNIP* ($\log_2$FC = −2.6, $p < 0.001$), and *CD69* ($\log_2$FC = −1.6, $p < 0.001$). This cluster is predominantly found at 1-3 months post-infection (565 of 567 cells) and likely represents recently activated central memory cells with proliferative capacity.

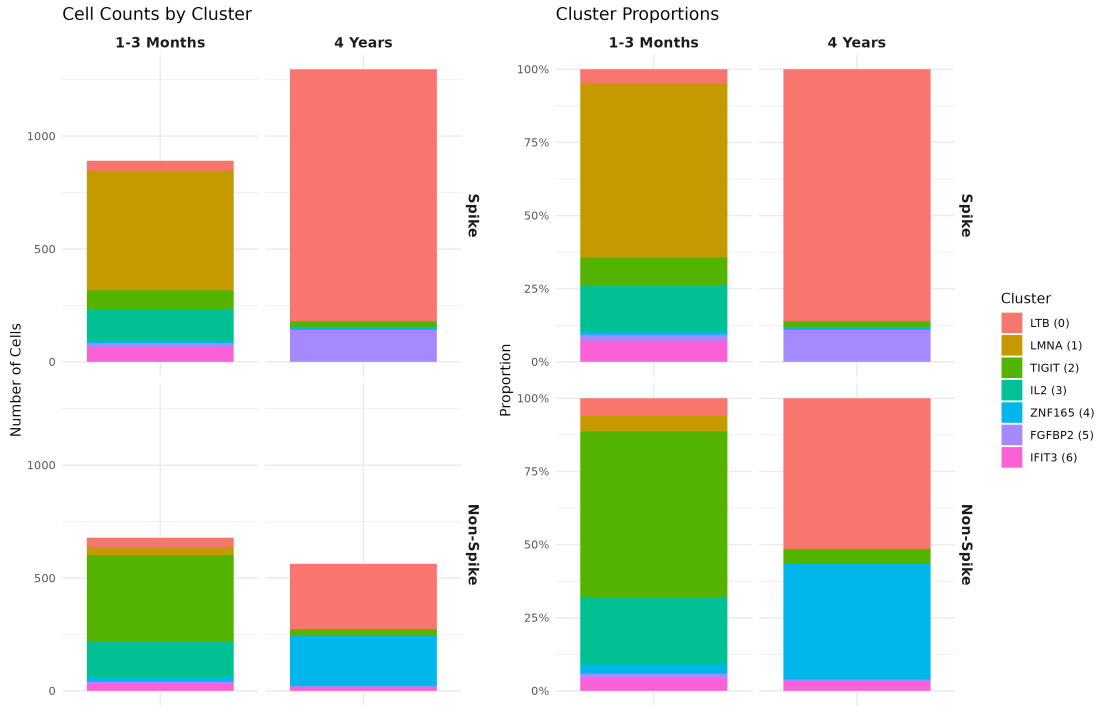


Figure 5.5: Cell counts (left) and overall proportions (right) of CD4⁺ T-cells within each transcriptional cluster at each timepoint, split by epitope type (Spike and Non-Spike).

The TIGIT (2) cluster most differentially expresses mitochondrial genes *MT1X* ($\log_2$FC = 3.1, $p < 0.001$), *MT1XP1* ($\log_2$FC = 2.6, $p < 0.001$), and *MT2A* ($\log_2$FC = 2.6, $p < 0.001$), alongside exhaustion marker *TIGIT* ($\log_2$FC = 2.0, $p < 0.001$). The cluster shows naive/-central memory markers including *SELL* ($\log_2$FC = 0.58, $p < 0.001$), *CCR7* ($\log_2$FC = 0.52, $p < 0.001$), and *TCF7* ($\log_2$FC = 0.38, $p < 0.001$). Immediate early genes *FOS* ($\log_2$FC = -2.6, $p < 0.001$) and *JUNB* ($\log_2$FC = -2.2, $p < 0.001$), activation marker *CD69* ($\log_2$FC = -1.7, $p < 0.001$), and *LMNA* ($\log_2$FC = -2.6, $p < 0.001$) are negatively expressed in this population. Predominantly found at 1-3 months (468 of 520 cells), this cluster is potentially a transitional central memory population with exhaustion features.

The IL2 (3) cluster shows an activation signature, with upregulation of *IL2* ($\log_2$FC = 10.0, $p < 0.001$), *IL21* ($\log_2$FC = 6.8, $p < 0.001$), and early growth response genes *EGR3* ($\log_2$FC = 6.4, $p < 0.001$), *EGR2* ($\log_2$FC = 6.1, $p < 0.001$), and *EGR1* ($\log_2$FC = 5.4, $p < 0.001$). The cluster is the only cluster with expression of Th1 effector cytokines *IFNG* ($\log_2$FC = 5.1, $p < 0.001$) and *TNF* ($\log_2$FC = 4.8, $p < 0.001$), alongside cytotoxic molecules *FASLG*

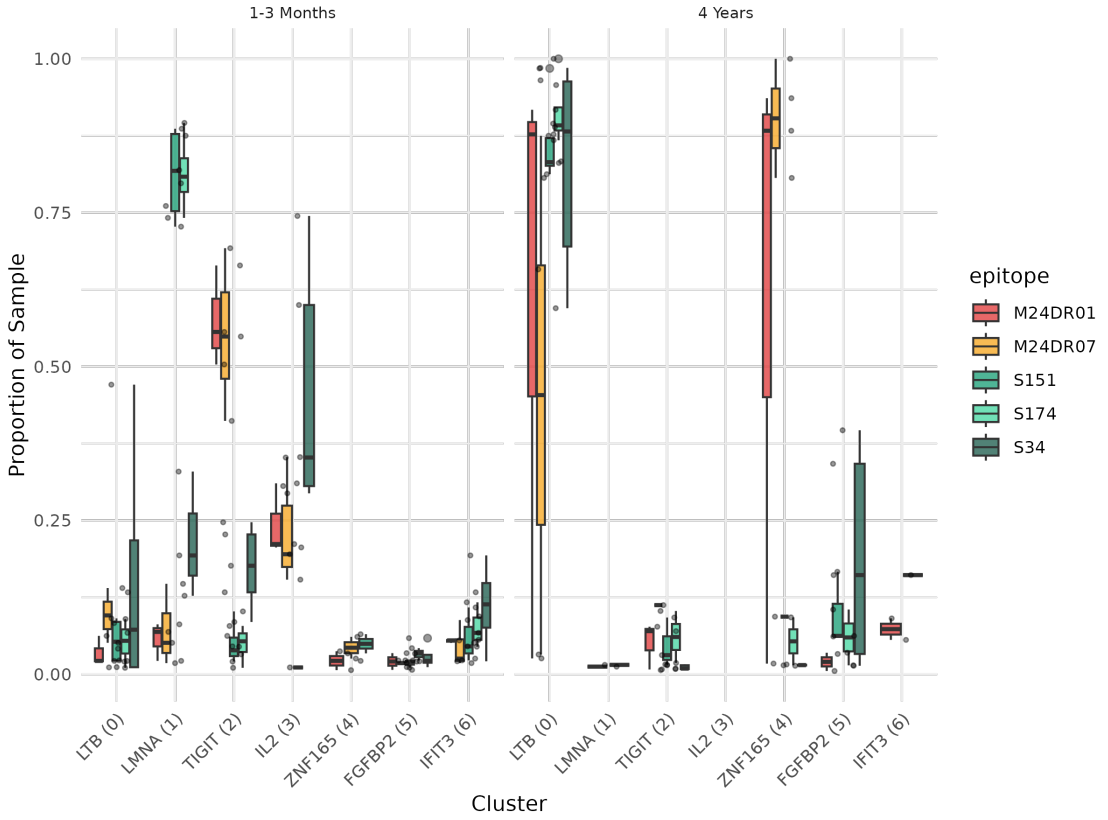


Figure 5.6: Whisker plot of the proportion of each sample within each transcriptional cluster. Box plots show the distribution of cluster proportions in samples, split by the epitope the sample is specific for. Each point represents an outlier sample (outside the IQR for the frequencies).

($\log_2$FC = 5.4, $p < 0.001$) and *TNFSF14* ($\log_2$FC = 5.0, $p < 0.001$). It shows modest upregulation of *STAT4* ($\log_2$FC = 0.76, $p < 0.001$), *ICOS* ($\log_2$FC = 1.3, $p < 0.001$), and *CTLA4* ($\log_2$FC = 1.5, $p < 0.001$), but reduced expression of *LTB* ($\log_2$FC = -3.4, $p < 0.001$) and *CCR7* ($\log_2$FC = -1.5, $p < 0.001$). Exclusively present at 1-3 months, this cluster likely represents Th1 effector cells in the recovery phase after infection.

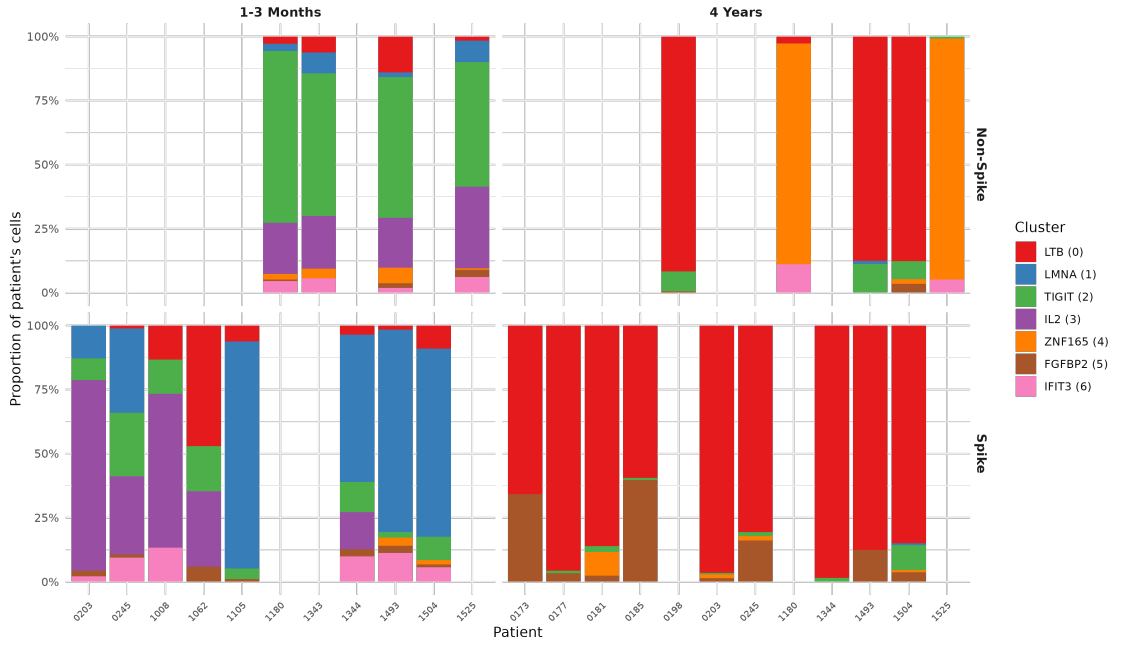
The ZNF165 (4) cluster most differentially expresses *ZNF165* ($\log_2$FC = 4.0, $p < 0.001$), *ELL2* ($\log_2$FC = 3.5, $p < 0.001$), *RASGEF1B* ($\log_2$FC = 3.4, $p < 0.001$), and *SKIL* ($\log_2$FC = 3.2, $p < 0.001$), alongside pseudogenes, ribosomal genes and mitochondrial genes. It shows high expression of stress-associated markers *BIRC3* ($\log_2$FC = 2.3, $p < 0.001$) and *SERPINB9* ($\log_2$FC = 2.1, $p < 0.001$), and upregulation of *CREM* ($\log_2$FC = 1.7, $p < 0.001$).

The ZNF165 (4) cluster contains a lower rate of paired clonotype detection than the other clusters (Figure 5.8), and while it is composed of a mixture of 7 patients, both timepoints

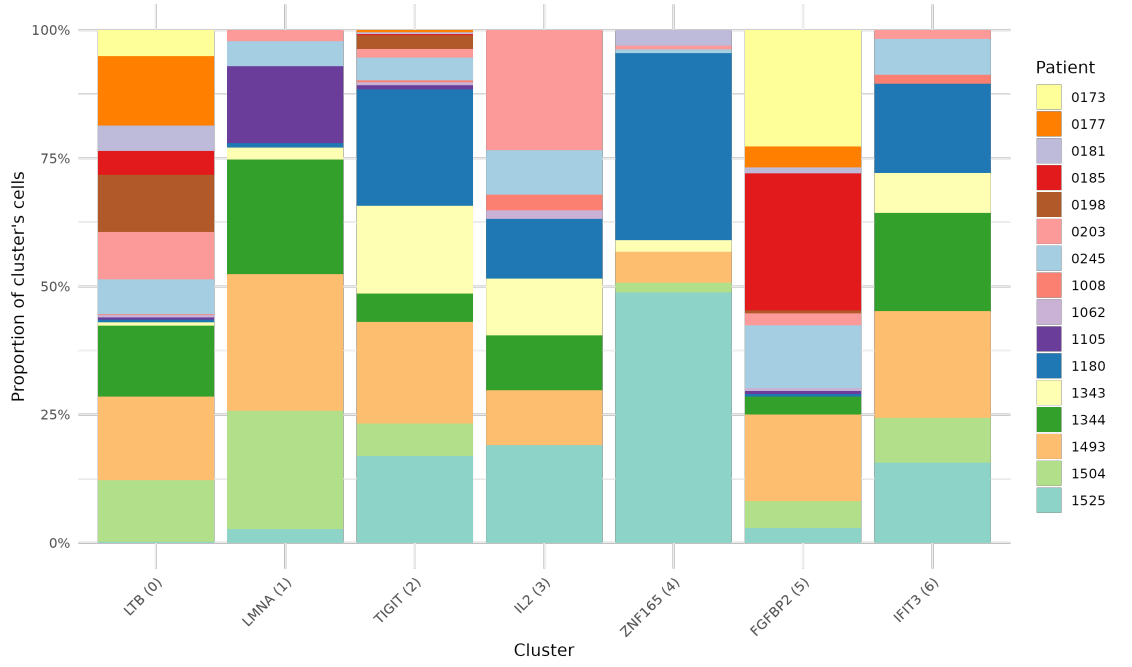
and 9 sequencing runs the majority come from two patients (1180 and 1525) sequenced in the same sequencing run (Figure 5.7). This could suggest a technical artefact where the ZNF165 (4) cluster reflects a batch effect with lower yield of unique mRNA features while having greater RNA counts (Figure E.5). While the cluster does differentially express T-cell related genes such as *CREM* and *CD69*, none of its top 20 most differentially expressed genes are canonical T-cell genes, it is enriched for ribosomal genes and is mostly defined by negative differential expression of *LCK* ($\log_2$FC = -3.0, $p < 0.001$), *SELL* ($\log_2$FC = -1.8, $p < 0.001$), *CD3D* ($\log_2$FC = -0.7, $p < 0.001$), *CD3E* ($\log_2$FC = -1.5, $p < 0.001$) and *IL32* ($\log_2$FC = -1.7, $p < 0.001$). I retained this cluster in descriptive analyses of cluster frequencies and UMAP visualisations to maintain transparency about the full dataset composition but exclude this cluster from later differential expression analysis, as the systematic non-detection of many genes will bias the resulting pseudobulk.

The FGFBP2 (5) cluster exhibits a cytotoxic phenotype, with high expression of *FGFBP2* ($\log_2$FC = 10.5, $p < 0.001$), *GZMH* ($\log_2$FC = 10.0, $p < 0.001$), *NKG7* ($\log_2$FC = 8.7, $p < 0.001$), *GZMB* ($\log_2$FC = 6.7, $p < 0.001$), and *GNLY* ($\log_2$FC = 6.1, $p < 0.001$). The cluster also highly expresses chemokines *CCL4* ($\log_2$FC = 5.5, $p < 0.001$) and *CCL5* ($\log_2$FC = 3.7, $p < 0.001$), cytotoxic effector molecules *PRF1* ($\log_2$FC = 4.0, $p < 0.001$) and *GZMA* ($\log_2$FC = 2.7, $p < 0.001$), and *KLRG1* ($\log_2$FC = 2.6, $p < 0.001$). It shows very low *CCR7* expression ($\log_2$FC = -3.8, $p < 0.001$) and reduced *CREM* ($\log_2$FC = -2.3, $p < 0.001$). Predominantly made up of cells sampled at four years (84%) and dominated by cells specific for S_{166}, this cluster represents long-lived cytotoxic $CD4^+$ T-cells with effector memory characteristics.

The IFIT3 (6) cluster shows a type I interferon response signature, with upregulation of interferon-stimulated genes *IFIT3* ($\log_2$FC = 7.9, $p < 0.001$), *IFIT2* ($\log_2$FC = 7.6, $p < 0.001$), *IFIT1* ($\log_2$FC = 7.3, $p < 0.001$), and *HERC5* ($\log_2$FC = 4.8, $p < 0.001$). The cluster also expresses antiviral effectors *MX1* ($\log_2$FC = 3.3, $p < 0.001$), *OAS1* ($\log_2$FC = 3.3, $p < 0.001$), *OAS2* ($\log_2$FC = 2.6, $p < 0.001$), and *ISG15* ($\log_2$FC = 3.3, $p < 0.001$). It shows modest upregulation of *TIGIT* ($\log_2$FC = 1.9, $p < 0.001$). Predominantly at 1-3 months (96 of 115 cells), this cluster represents recently activated cells experiencing active



(a)



(b)

Figure 5.7: Distribution of transcriptional clusters across patients. (A) Cluster composition within each patient, stratified by timepoint (1-3 Months on the left, 4 Years on the right) and epitope type (Non-Spike above, Spike below). (B) Patient contribution to each cluster, showing that the ZNF165 (4) cluster is predominantly composed of cells from patients 1180 and 1525.

interferon stimulation, likely reflecting recent viral antigen encounter or stimulation with interferons.

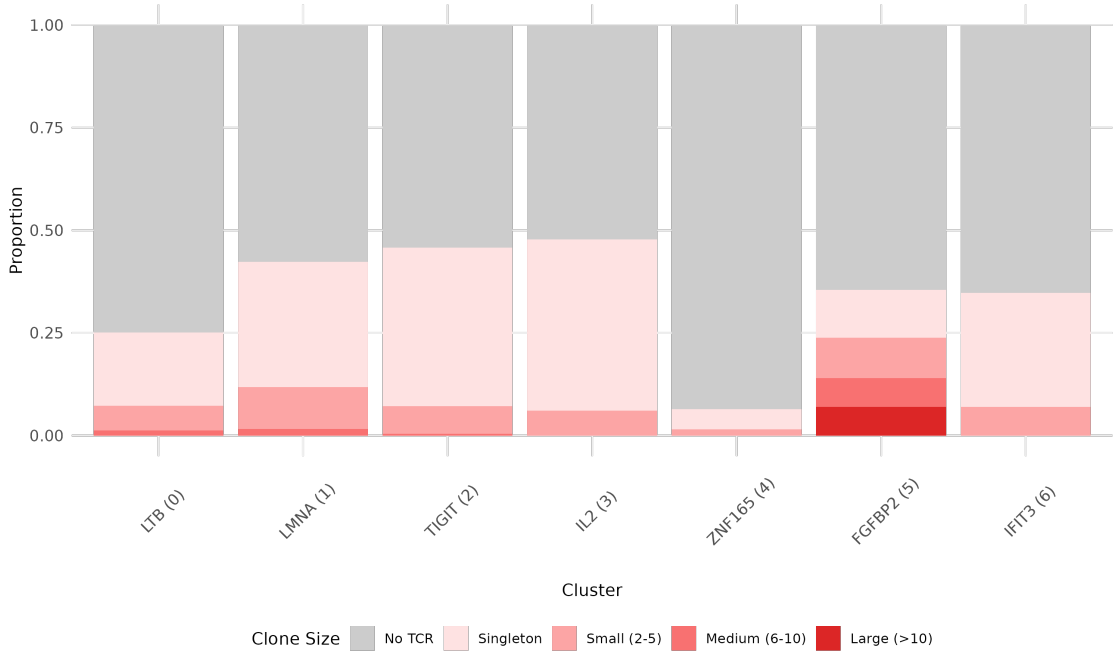


Figure 5.8: Stacked bar chart coloured by clone size categories within each transcriptional cluster. Clone sizes are calculated within each sample on the basis of paired $\alpha\beta$ amino acid TCR sequences. For visualisation, clone sizes were categorised as: singleton (1 cell), small (2-5 cells), medium (6-10 cells), or large (>10 cells).

The distribution of clusters by timepoint and epitope specificity (Figure 5.10) provides hypotheses about the evolution of these cells over time. While UMAP coordinates are not directly interpretable, the overlap of metadata categories with transcriptional clusters suggest that dimensionality reduction and clustering has effectively grouped similar cells from across samples, while maintaining biological variation (Figure 5.10). Cells sampled from mild, severe and critical patients are distributed between clusters, whereas the Spike and non-Spike specificity are more clearly separated. Most clearly separated are the timepoints, with a strong separation between early and late timepoints from left to right. This reflects both the likely biological evolution of these cells over time, but importantly could also reflect differences in technical factors between samples taken and sequenced over 3 years apart. I do not attempt to formally compare gene expression between timepoints, but rather focus on the differences between Spike and non-Spike specific cells, which are sampled and sequenced in parallel.

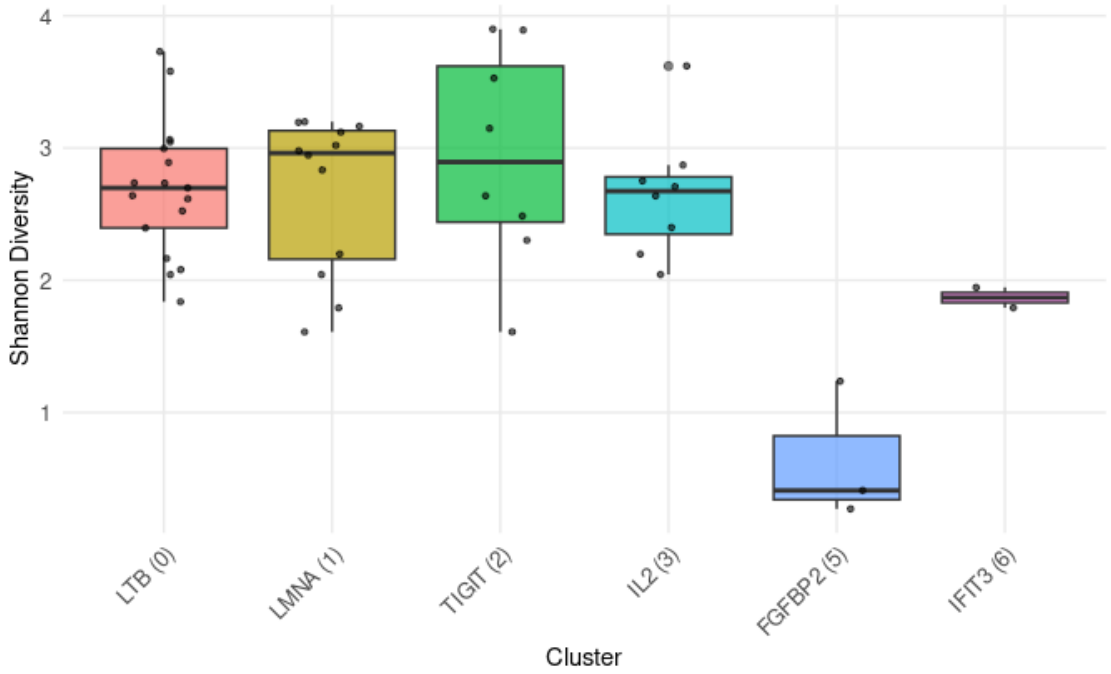


Figure 5.9: Whisker plot of Shannon diversity indexes of paired $\alpha\beta$ TCR repertoires per sample within each transcriptional cluster. Box plots show the distribution of Shannon diversity indexes in samples, split by the transcriptional cluster the sample is specific for. Each point represents the diversity within a sample-cluster combination.

5.2.3 Comparison of Epitope-specific Repertoires of Spike and Non-Spike specific CD4⁺ T-cells

Longitudinal sampling of repertoires directed at multiple epitopes allows us to track how antigen-specific CD4⁺ TCRs evolve over time. In the 7 paired samples with more than 3 paired TCRs at both timepoints there were no statistically significant changes in V- or J- gene frequency (Adjusted p-values between $0.635 - 1.0$) between repertoires from 1-3 months to four years timepoints (FDR-adjusted $p > 0.05$ for all comparisons, Wilcoxon signed-rank test). Our cohort provides no evidence that specific genes are preferentially maintained at higher rates for any epitopes in particular. Whisker plots of gene usage frequencies by epitope are provided in Section E.5.

To test whether repertoire dynamics differed between Spike and non-Spike epitopes at the 1-3 month timepoint and the four-year timepoint I compared 37 epitope-specific CD4⁺ TCR repertoires from 16 individuals between Spike specific and non-Spike specific T-cells (Figure 5.11). I fit a linear model to the Shannon diversity of the paired $\alpha\beta$ repertoire of

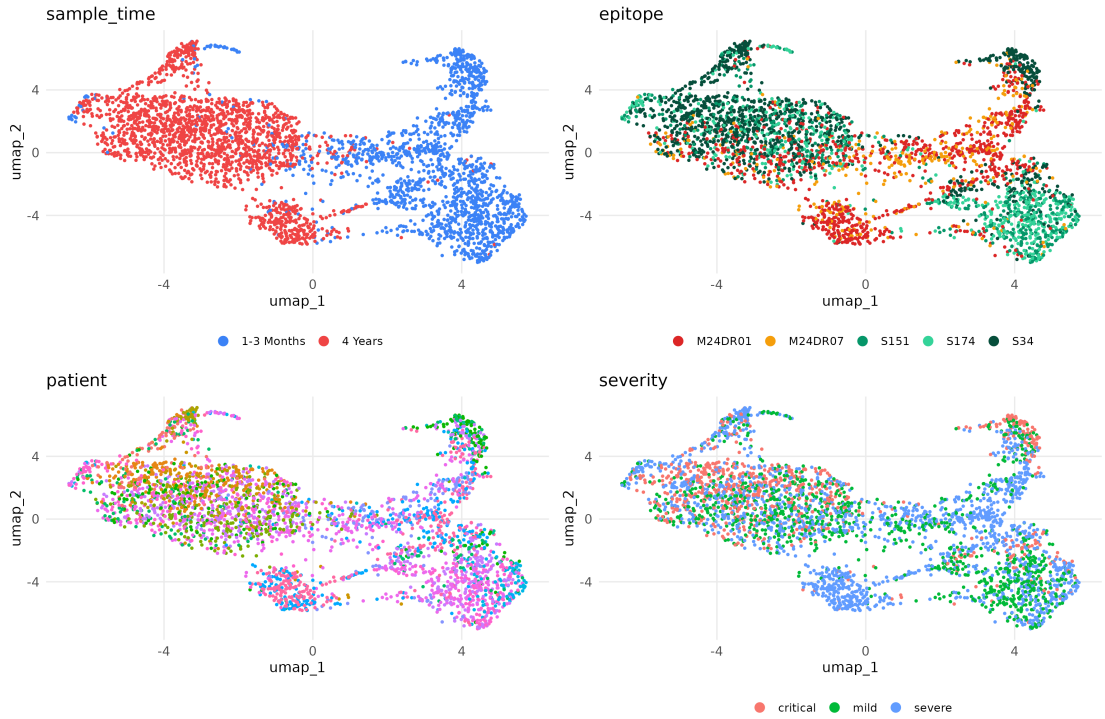


Figure 5.10: UMAP Projection of SARS-CoV2 specific CD4⁺ T-cells coloured by metadata categories including timepoint, epitope specificity, donor and the severity of initial infection by WHO classification. Epitopes are labelled S34, S151 and S174 in this figure, corresponding to $S_{166\text{-}180}$, $S_{751\text{-}765}$ and $S_{866\text{-}880}$ respectively, and M24DR01 and M24DR07 correspond to $M_{172\text{-}188}$ presented by DRB1*01:01 and DRB1*07:01.

each sample (combination of donor, timepoint and epitope) as a function of the interaction between protein target (Spike and non-Spike) and timepoint. The model explained 27.4% of the variance in diversity ($R^2 = 0.274$). Spike epitope specific repertoires showed no significant change in diversity between the early and late timepoint ($\beta = -0.250$, 95% CI: -0.816 to 0.315). However, the interaction between timepoint and non-Spike specificity was significant ($\beta = -1.202$, 95% CI: -2.243 to -0.162, $p = 0.025$), indicating that non-Spike specific repertoires show greater diversity loss over time than Spike-specific repertoires (Figure 5.12). The estimated total change of Non-Spike repertoire diversity from 1-3 months to four years was therefore -1.453 (95% CI: -2.326 to -0.580), $p = 0.002$.

This pattern suggests that Spike-specific CD4⁺ T-cell repertoires maintain their diversity, whereas non-Spike specific repertoires reduce in repertoire diversity. In line with work by Ford et al.[397] and Dykema et al[396], I speculate that if Spike-specific cells experienced

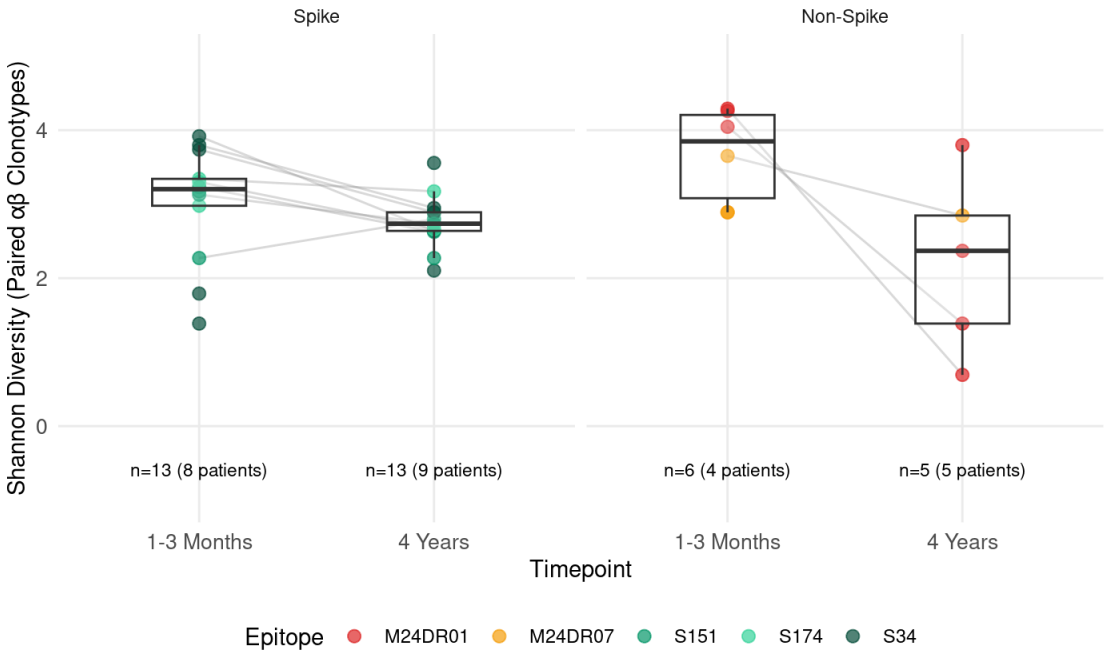


Figure 5.11: Temporal changes in $\alpha\beta$ TCR Repertoire Diversity in Spike specific (n=26) and non-Spike specific (n=11) CD4⁺ TCR repertoires. Shannon diversity of paired $\alpha\beta$ repertoires was calculated for sample, and lines connect paired samples where the same epitope was sampled in the same individual at multiple timepoints. Each point represents one epitope-specific repertoire, coloured by Epitope identity. Box plots show the median and interquartile ranges of diversity. Epitopes are labelled S34, S151 and S174 in this figure, corresponding to $S_{166\text{-}180}$, $S_{751\text{-}765}$ and $S_{866\text{-}880}$ respectively, and M24DR01 and M24DR07 correspond to $M_{172\text{-}188}$ presented by DRB1*01:01 and DRB1*07:01.

more frequent antigen exposure (e.g., through vaccination), this could provide proliferative, recruitment or maintenance signals that preserve diversity.

The trend of greater diversity loss over time by Non-Spike epitopes led me to investigate factors that might influence clonal persistence or loss. To assess the long-term persistence of individual clonotypes, I tracked individual clonotypes (defined either by their TCR α, TCR β or their paired TCR) across timepoints. Raw persistence rates suggested differences between Spike and non-Spike repertoires, but this could be driven by greater clonality in Spike-specific repertoires (Figure 5.13). Additionally, sampling depth at both early and late timepoints varied between epitopes and donors, so I fit a logistic regression model to the persistence of an individual clonotype on the basis of (1) whether it was specific for a Spike or non-Spike epitope, (2) the number of cells with that clonotype sampled at the early timepoint, and (3) the total number of cells with an observed TCR in the late timepoint.

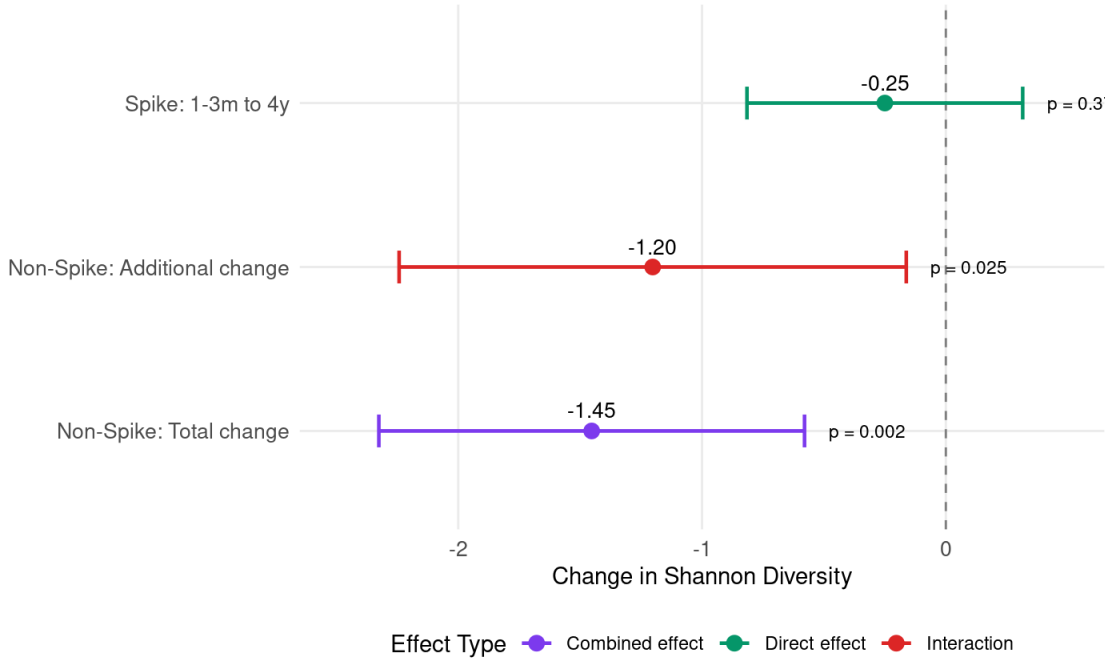


Figure 5.12: Forest plot of the linear regression model fit to Shannon diversity of TCR repertoires directed at Spike and non-Spike epitopes at 1-3 months and four years. Points represent effect estimates, and horizontal bars indicate 95% confidence intervals. Spike-specific repertoires show moderate diversity loss (green), while the interaction term (red) suggests that Non-Spike specific repertoires show greater diversity loss over time. The estimated total change of Non-Spike repertoire diversity from 1-3 months to four years is the sum of these (purple).

When accounting for these variables and applying Benjamini-Hochberg correction for multiple testing across 9 comparisons, the persistence advantage of Spike-specific clonotypes was not statistically significant for alpha chains (OR=2.98, 95% CI: 0.94 to 9.42, $p = 0.082$), beta chains (OR=1.97, 95% CI: 0.97 to 4.01, $p = 0.082$) or paired chains (OR=2.99, 95% CI: 0.61 to 14.52, $p = 0.197$), suggesting that the stability of Spike-specific repertoires is driven by their more clonal structure and therefore a higher probability of encountering the same clone, rather than differences in clonal longevity compared to non-Spike specific $CD4^+$ T-cells (Figure 5.14). Clone size was the strongest predictor of persistence, with each log-fold increase in the number of cells with a given clonotype increasing persistence odds significantly in all clonotype definitions (alpha: OR=10.0, $p < 0.001$; beta: OR=8.3, $p < 0.001$; paired: OR=8.3, $p = 0.007$). Late timepoint sampling depth also influenced persistence detection, with significant effects for beta chains (OR=2.9, $p = 0.007$) and paired chains (OR=8.7, $p = 0.024$), but not for alpha

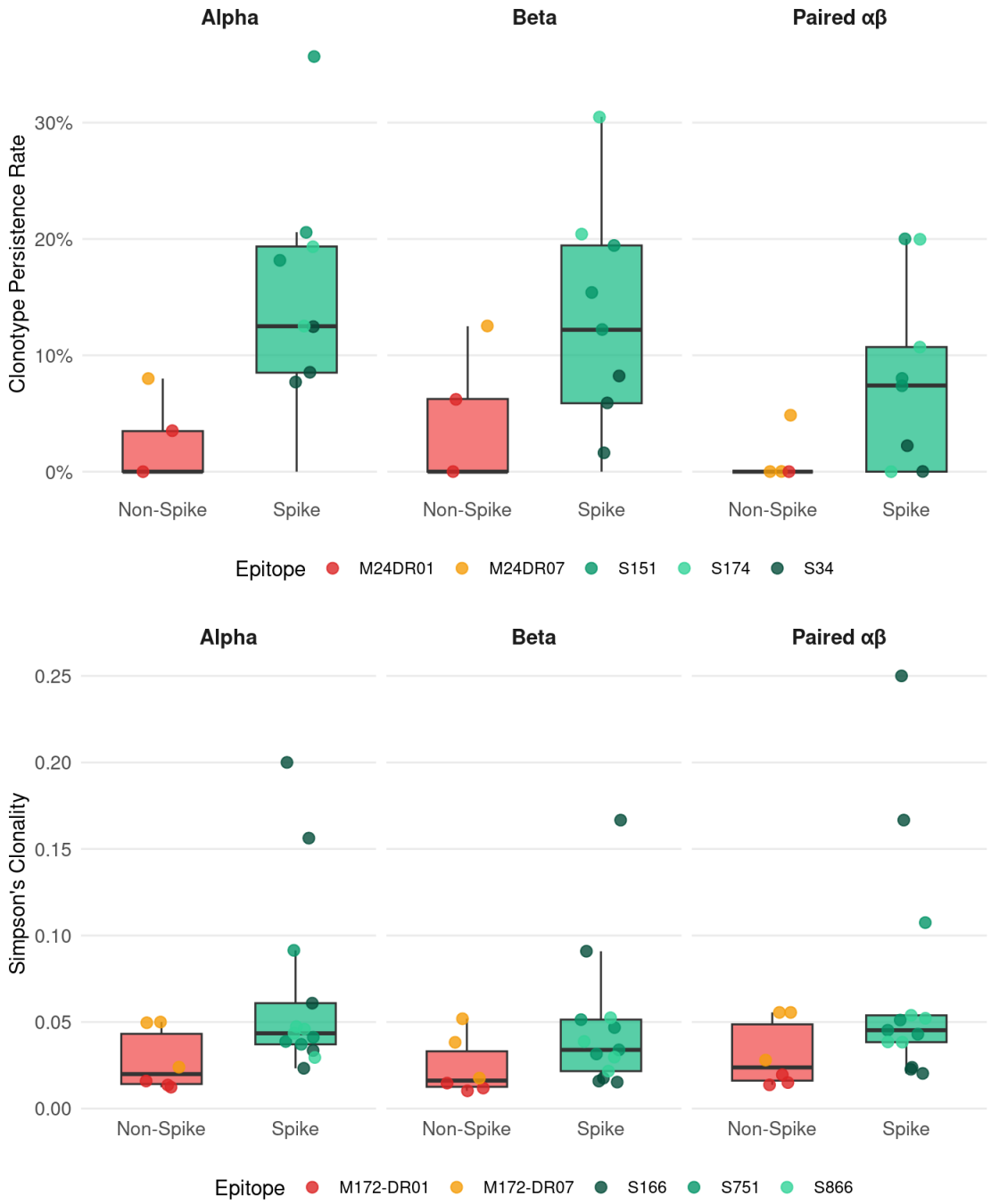


Figure 5.13: Raw clonotype persistence rates and clone size distributions. Top panel shows the proportion of Alpha, Beta and Paired clonotypes that persist from 1-3 months to four years for Spike and non-Spike specific repertoires. Bottom panel shows the distribution of sample Simpson's clonalities at the early (1-3 Month) timepoint. Spike specific samples have more clones where more than one cell shows that paired clonotype, leading to greater clonalities of each sample. In the top panel, epitopes are labelled S34, S151 and S174, corresponding to $S_{166\text{-}180}$, $S_{751\text{-}765}$ and $S_{866\text{-}880}$ respectively.

chains (OR=1.5, $p = 0.400$). This suggests that the persistence advantage of Spike-specific clonotypes is attributable to their greater clonal expansion at the early timepoint, rather than differential clonal longevity.

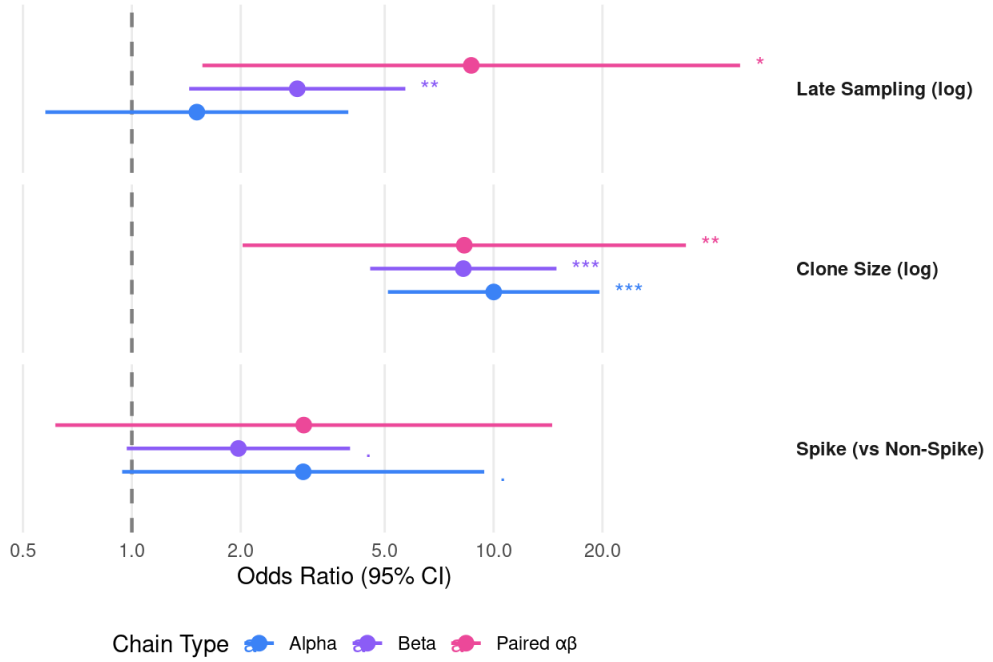


Figure 5.14: Forest plot of predictors of Alpha, Beta and Paired clonotype persistence from the early (1-3 Month) to late (four years) timepoint. Pink, purple and blue bars represent the confidence intervals for Paired, Beta and Alpha clonotypes respectively. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5.3: Logistic regression of clonotype persistence. Models composed of Epitope Type (Spike vs Non-Spike), log of clone size at 1-3 months, and log of overall sampling depth (i.e. the number of cells with the chain(s) observed) at four years. P-values are adjusted for multiple testing using the Benjamini-Hochberg method. OR = odds ratio; CI = confidence interval.

Parameter	Chain Type	Odds Ratio	95% CI	BH adjusted p-value
Late Sampling Depth (per log unit)				
	Alpha	1.51	0.58–3.97	0.400
	Beta	2.87	1.44–5.70	0.007
	Paired $\alpha\beta$	8.68	1.57–48.03	0.024
Clone Size (per log unit)				
	Alpha	10.01	5.10–19.62	<0.001
	Beta	8.25	4.56–14.92	<0.001
	Paired $\alpha\beta$	8.30	2.02–34.02	0.007
Spike vs Non-Spike				

Parameter	Chain Type	Odds Ratio	95% CI	BH adjusted p-value
	Alpha	2.98	0.94–9.42	0.082
	Beta	1.97	0.97–4.01	0.082
	Paired $\alpha\beta$	2.99	0.61–14.52	0.197

5.2.4 Differential gene expression between Spike and Non-Spike specific CD4⁺ T-cells

One of the core questions that this cohort allows us to investigate is whether Spike specific CD4⁺ T-cells have different transcriptional profiles to non-Spike specific CD4⁺ T-cells in the long term after their initial infection. While Spike specific CD4⁺ T-cells might be recalled by both vaccinations and infections, non-Spike specific CD4⁺ T-cells are only recalled by infections, which raises questions as to their long term phenotypes and evolution. To address this question, I performed pseudobulked differential gene expression using `DESeq2`, aggregating transcript counts by donor, timepoint and epitope as covariates. As we have separately analysed clonotype dynamics, I excluded all T-cell receptor genes to prevent epitope-specific signatures from dominating the differential expression signal. Comparing Spike-specific and non-Spike specific cells at the early (1-3 Months) timepoint, there are 3706 differentially expressed genes (DEGs) at a significance threshold of $FDR < 0.05$, with 539 of them more abundant in Spike-specific CD4⁺ T-cells, and 3167 more abundant in non-Spike specific CD4⁺ T-cells (Figure 5.16). At the 4-year timepoint, there are 2002 DEGs, with 1792 more abundant in Spike-specific CD4⁺ T-cells, and 210 more abundant in non-Spike specific CD4⁺ T-cells (Figure 5.19).

The differentially expressed genes between Spike and non-Spike specific CD4⁺ T-cells at the 1-3 month timepoint could reflect consequences of different epitope targets or presentation patterns. In gene set enrichment analysis, no gene ontologies were significantly enriched amongst the genes highly expressed in Spike samples, but among the highly expressed genes among Non-Spike samples 52 gene ontologies were significantly enriched after multiple testing (Section E.6). The only ontology mentioning T-cells or lymphocytes amongst these was `T Cell Receptor Signaling Pathway (GO:0050852)`, which was the 22nd most

significantly enriched gene set (Figure 5.15). While these genes reached statistical significance in DESeq2 analysis adjusted for donor, 29 of the 88 genes in this gene set were more highly expressed in Non-Spike specific cells. The other significant gene sets relate to broad molecular biology processes.

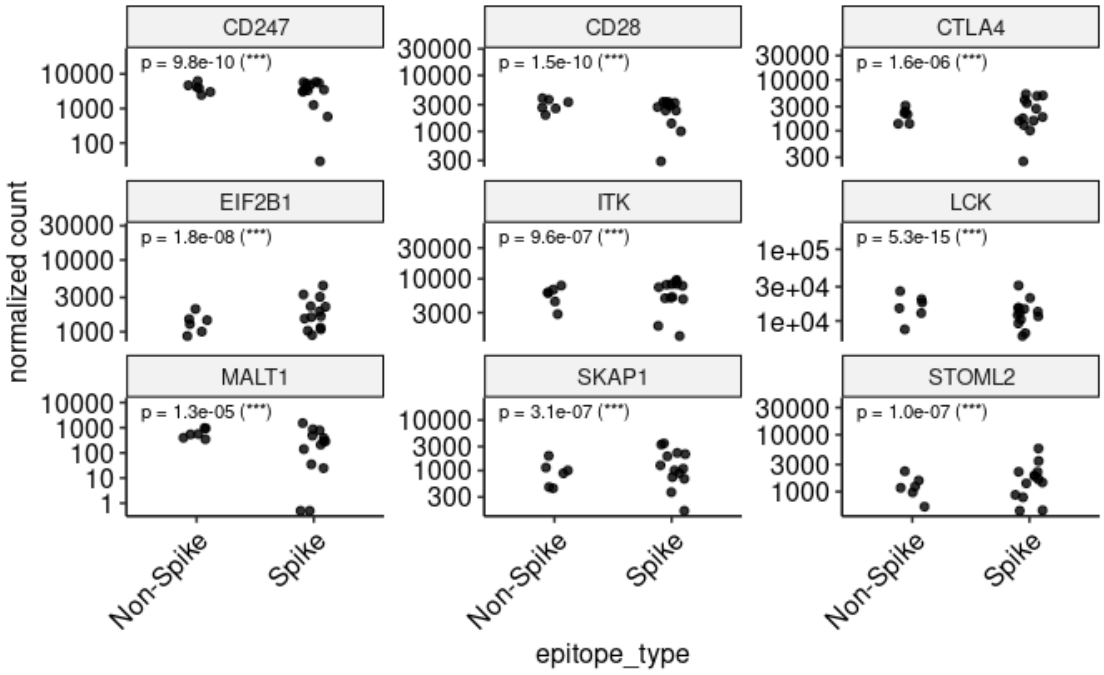


Figure 5.15: The distribution of DESeq2 Normalised Pseudobulk counts of the 9 most significantly differentially expressed genes in the `T Cell Receptor Signaling Pathway (GO:0050852)` ontology.

Post-hoc, I also visualise differentially expressed genes relevant to CD4⁺ T-cell function or subset for annotation of cell types and states between Spike specific and non-Spike specific cells. This is shown in Figure 5.17.

At the four year timepoint, differentially expressed genes between Spike and non-Spike specific CD4⁺ T-cells potentially reflect consequences of the epitope targets and the different recall patterns. Gene set enrichment analysis of Spike and non-Spike specific cells at the four year timepoint also does not reveal any ontologies that are significantly enriched after multiple testing corrections (Section E.6). So while individual genes are differentially expressed between Spike and non-Spike specific samples, no evidence for differentially expressed ontologies is observed in this dataset. That no ontologies reach significance could

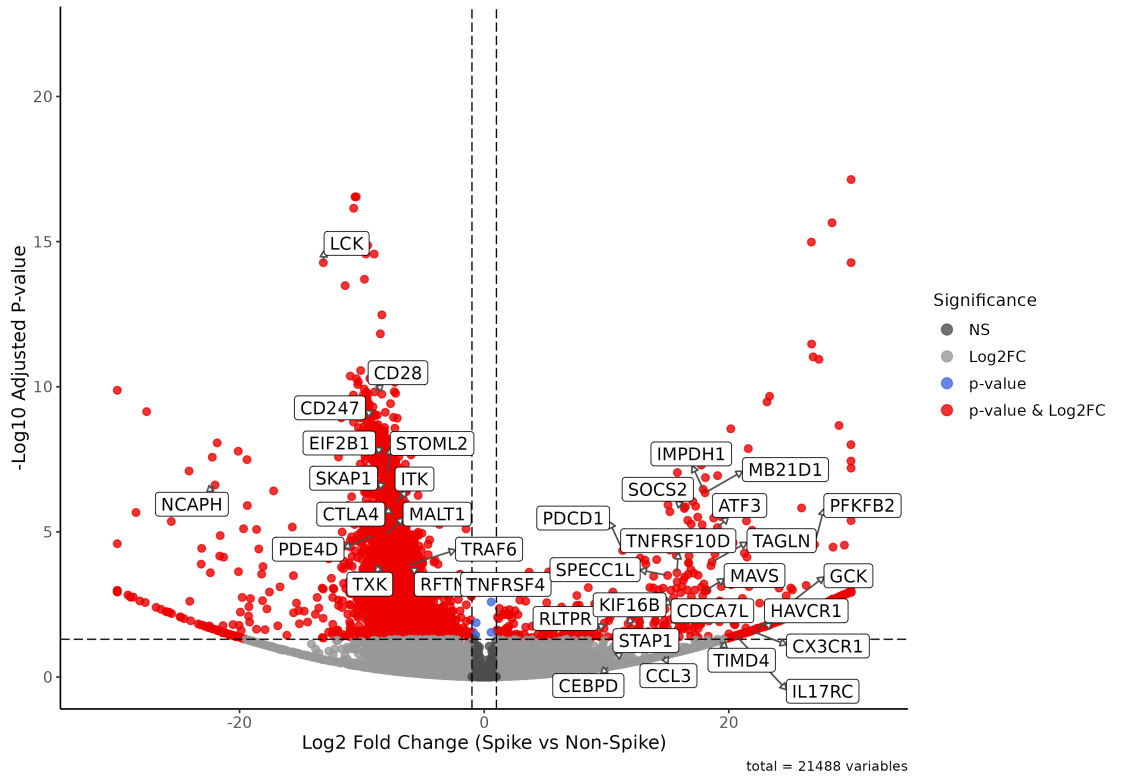


Figure 5.16: Volcano plot showing differential gene expression between Spike and Non-Spike-specific CD4⁺ T-cells at 1-3 Months post-infection. Points represent individual genes, with significantly differentially expressed genes (adjusted $p < 0.05$, $|\log2FC| > 1$) highlighted in red. Genes discussed in the text are highlighted in the figure.

reflect the heterogeneity of the cohort, whereby individual variation in disease severity, time since infection and other factors could dilute signals from epitope specificity.

Without significant gene set enrichment, I performed exploratory analysis of the differentially expressed genes, interpreting genes I identified as relevant to CD4⁺ T-cell state, subset or function (Figure 5.18). Spike-specific cells highly express Interferon stimulated genes (ISGs) and antiviral effectors at four years post-infection including *OAS2*, *OAS3*, *RNASEL*, *TRIM25*, *RSAD2* (Viperin) and the signalling adaptor *MAVS*. Interferon Regulatory Factors *IRF5* and *IRF7* were also highly expressed in Spike-specific samples. *IRF7* regulates Type I interferon-dependent immune responses and *IRF5* promotes pro-inflammatory cytokine production, suggesting that these cells maintain a heightened antiviral state with capacity for interferon production and responsiveness to interferon signalling. A sustained interferon response signature four years after priming infection could reflect repeated antigen

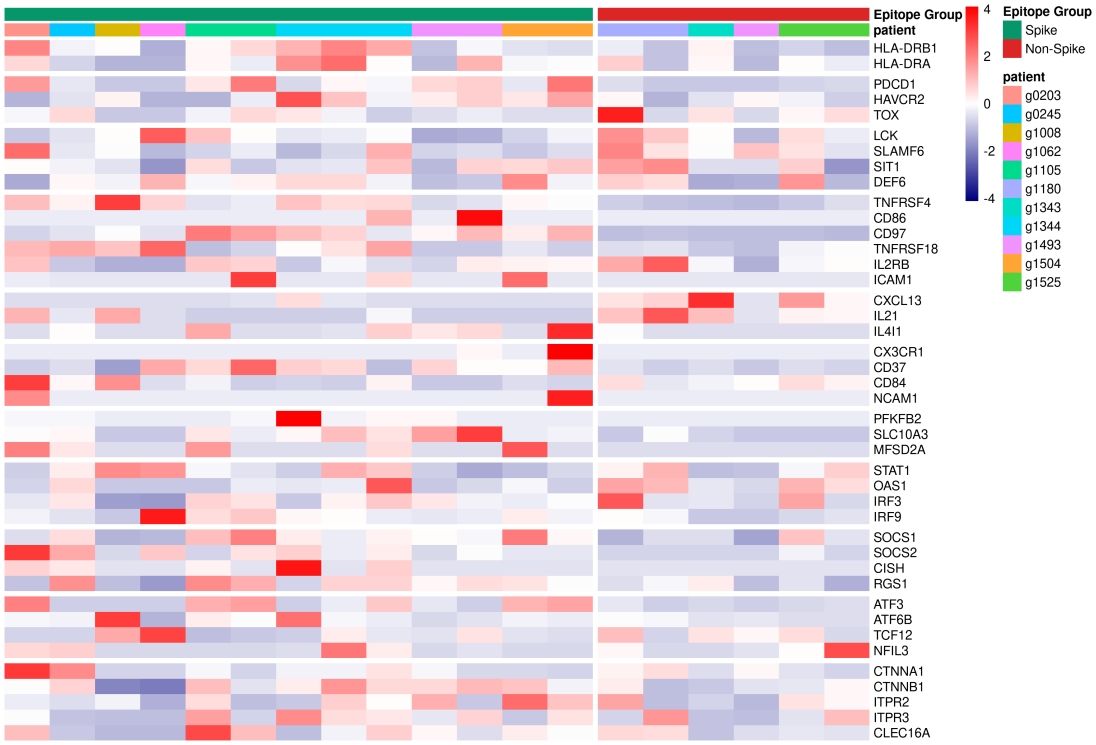


Figure 5.17: Heatmap of manually curated genes differentially expressed between Spike (Green) and Non-Spike (Red) specific CD4⁺ T-cells at 1-3 Months post-infection. Heatmap elements reflect the pseudobulked, normalised expression of each gene in each sample (combination of donor, epitope and timepoint).

exposure. Supporting their maintained antiviral state, Spike-specific cells highly express stress-response factor *ATF3*, which could represent management of cellular stress from repeated antigen exposure.

Non-Spike specific cells at the late timepoint highly express genes that suggest a regulatory or anergic signature. They have high expression of *EGR2*, *CREM* and *DUSP6*. Non-Spike specific samples were enriched for *IFIT1* and *OASL* counts, suggesting a subset of interferon-response genes might be triggered or regulated differently.

Exhaustion, co-stimulation and functional markers were also differentially expressed between Spike and non-Spike specific cells at the four year timepoint. Spike specific cells highly express a variety of effector molecules including cytotoxic molecules *GZMB* and *GZMH*, chemokines and chemokine receptors that mediate tissue trafficking (*CCL3*, *CXCR5*, *CXCR6*) and receptors involved in T-helper differentiation (*IL17RA*, *IL17RC*, *IL12RB1*). The presence of granzymes as well as *PRF1* suggests that some Spike-specific

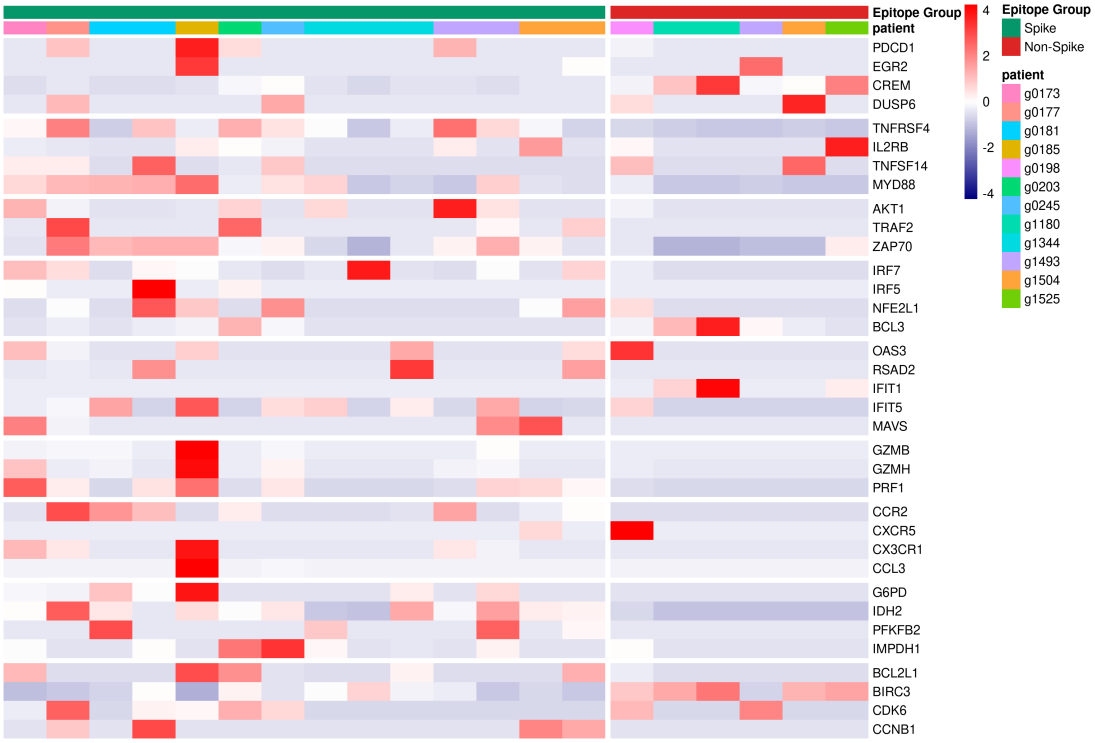


Figure 5.18: Heatmap of manually curated genes differentially expressed between Spike (Green) and Non-Spike (Red) specific CD4⁺ T-cells at four years post-infection. Heatmap elements reflect the pseudobulked, normalised expression of each gene in each sample (combination of donor, epitope and timepoint).

cells, in particular those from patients 0185 and 0173 (Figure 5.18) may have differentiated into CD4-CTLs, which have been observed in SARS-CoV2 infection and viral memory to a variety of viruses. This is also reflected by the FGFBP2 (5) cluster, present most prominently in Spike samples at the late timepoint. High expression of *CX3CR1* and *CXCR5* indicates capacity for migration to peripheral or lymphoid tissues as well as inflamed sites, whereas *CCL3* expression suggests a capacity to recruit other immune cells. Compared to non-Spike specific cells, high expression of IL17 and IL12 receptor subunits suggests that the pseudobulks of Spike-specific cells maintain a mixture of cells with capacity for differentiation into T_{h17} or T_{h1} subsets on cytokine stimulation. The absence of functional or trafficking markers in non-Spike specific cells suggests a more quiescent phenotype.

Spike-specific pseudobulks show a complex activation profile, expressing *PDCD1* (PD-1), *TNFRSF4* (OX40) and *TNFRSF10D* indicating co-stimulatory capacity and responsiveness to

TNF. Elevated expression of *BCL2L1* (Bcl-xL) alongside activation of pro-survival pathways (*AKT1*) and MAPK signalling (*MAP2K1*, *MAP2K3*, *MAP2K6*, *MAP2K7*) suggests that Spike-specific cells resist apoptosis despite exhaustion marker expression. The NF-κB pathway also shows evidence of engagement in Spike-specific cells, with high expression of *NFKBIL1*, *NFKBID*, *TRAF2* and *TRAF3IP2* (which is also a mediator of T_{h17} inflammatory responses).

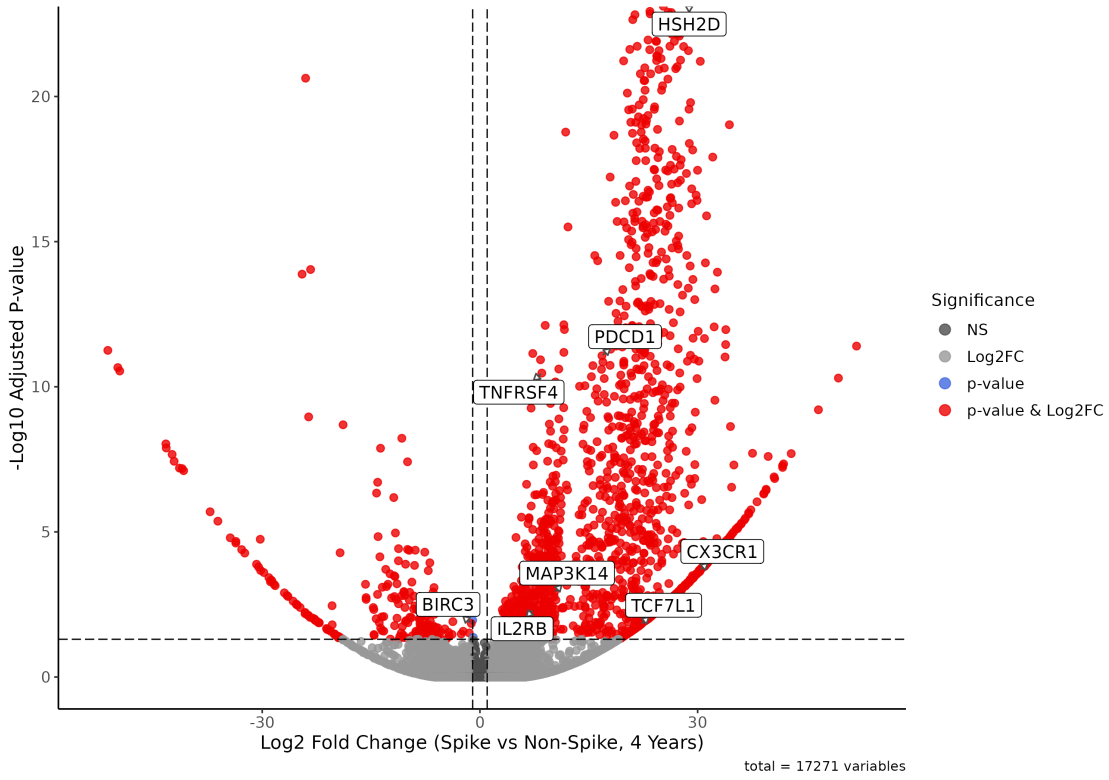


Figure 5.19: Volcano plot showing differential gene expression between Spike and Non-Spike specific CD4⁺ T-cells at four years post-infection. Points represent individual genes, with significantly differentially expressed genes (adjusted $p < 0.05$, $|\log2FC| > 1$) highlighted in red. Genes discussed in the text are highlighted in the figure.

I identified genes consistently differentially expressed between Spike and non-Spike specific CD4⁺ T-cells at both timepoints. These genes might reflect "imprinted" differences in their early exposure. The intersection of DEGs at both timepoints reveals 364 genes differentially expressed in the same direction at both timepoints, but 70 genes differentially expressed in opposite directions (Figure 5.20, 434 in total).

Genes highly expressed in Spike-specific cells at both timepoints make up the majority of these genes (300/434), including immune checkpoint *PDCD1*, co-stimulatory molecule *TN-*

FRSF4 (OX40) and NF-κB activator *TNFRSF12A*. The cytokine receptor *IL2RB* and *CX3CR1* are also more highly expressed in Spike-specific cells at both timepoints. Transcriptional regulators associated with T-cell state are also maintained - *ATF3*, a regulator of T-cell quiescence and exhaustion; *CEBPD*, which regulates T-cell differentiation; and *TRIB1/3* which regulate T-cell proliferation and survival. *MAVS*, involved in antiviral signalling, maintains higher Spike-specific expression at both timepoints.

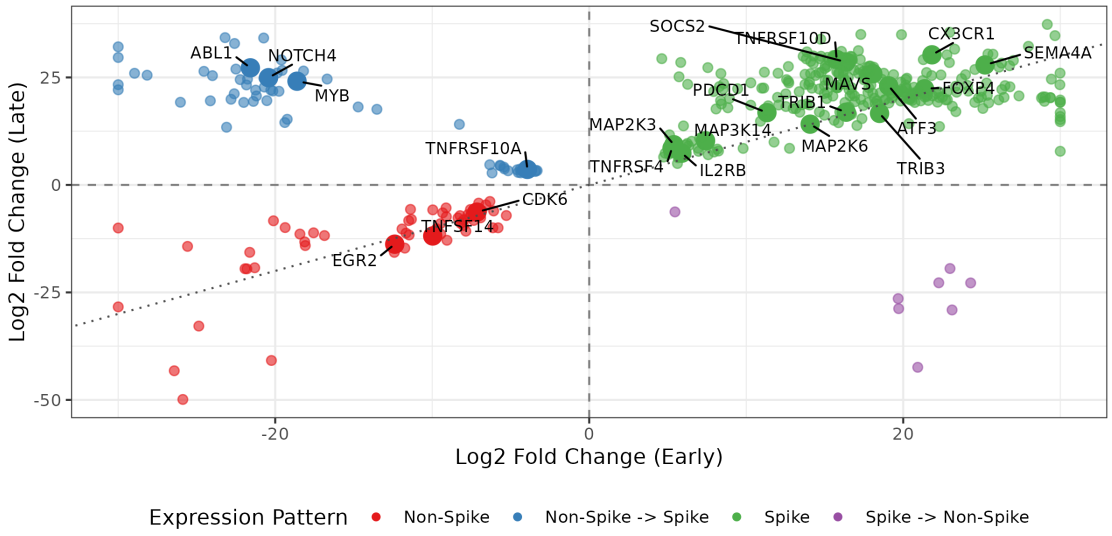


Figure 5.20: Differential expression dynamics of genes differentially expressed at both early and late timepoints. Scatter plot comparing log2 fold changes under DESeq2 differential gene expression of genes significantly differently expressed (adjusted $p < 0.05$) at both early and late timepoints. Each point represents a gene, and genes relevant to CD4⁺ state or function are highlighted. Colours indicate expression patterns - genes consistently highly expressed by Spike are green, Non-Spike are red and those with switched greater expression are in purple or blue.

Only two labelled T-cell genes switch expression patterns between timepoints. *ABL1* (a kinase involved in T-cell receptor signalling) transitions from Non-Spike high to Spike high expression. The swap of *ABL1* could suggest more recent or frequent TCR stimulation in Spike-specific cells at the four year timepoint, possibly due to more frequent recall. *MYB* is highly expressed at the early timepoint in Non-Spike samples, but becomes highly expressed in Spike samples at the late timepoint. The gene has recently been implicated in the regulation of T-cell exhaustion through downregulating effector function and preserving self-renewal capacity[399].

5.2.5 Exploratory analysis of priming differences between Spike and non-Spike specific CD4⁺ T-cells in Acute SARS-CoV2

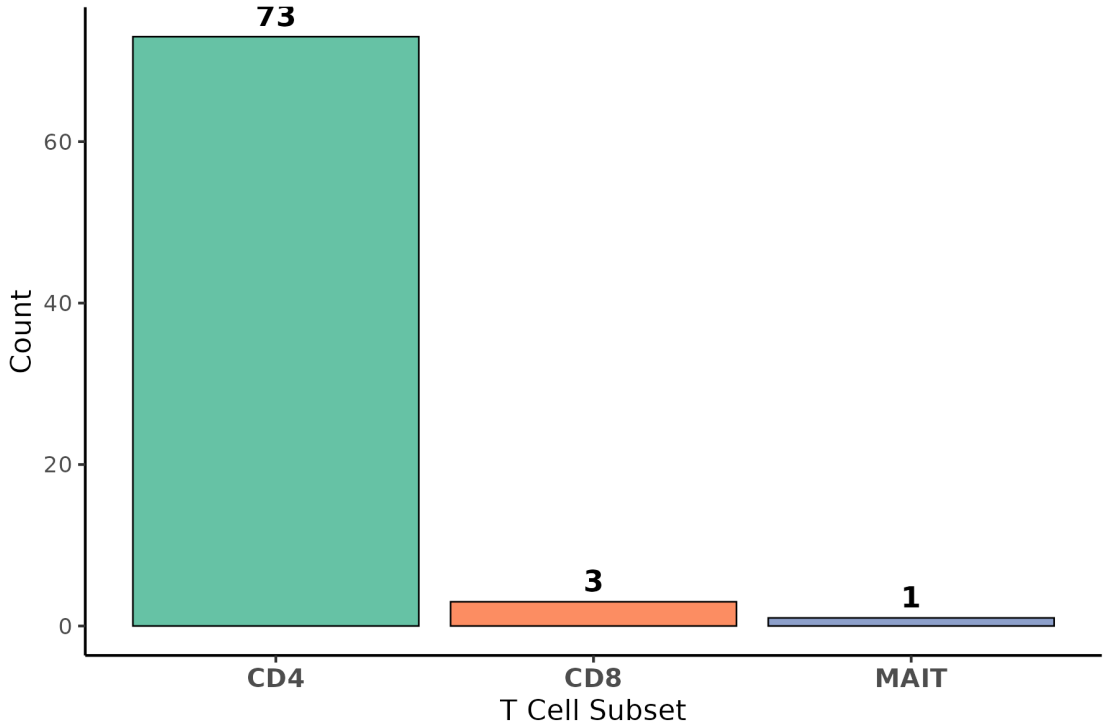


Figure 5.21: Major T-cell subset distribution for COMBAT T-cells with TCRs similar to sampled Spike or non-Spike specific CD4⁺ T-cells. The vast majority (94.8%, 73/77) of matched clonotypes were annotated in the COMBAT CITE-Seq dataset as CD4⁺ T-cells. A small number of matches occur in cells annotated as CD8⁺ T-cells or MAIT cells in the COMBAT dataset.

To generate hypotheses about these priming related differences, I re-analysed the COMBAT[34] dataset of single-cell RNA and TCR sequencing of PBMCs. Since this dataset was sampled in the acute phase of infection, I hypothesised that it would contain recently primed SARS-CoV2 specific CD4⁺ T-cells, particularly those epitopes we have sequenced TCRs for. By comparing cells with matched TCRs to Spike epitopes to those with matched TCRs to the Non-Spike epitopes there is an opportunity to explore how their priming differs. I identified 82 putative Spike-specific and 19 putative non-Spike specific CD4⁺ T-cells across 42 patients within a dataset of 316927 cells and 122 patients on the basis of paired TCR similarity to our experimental TCRs. Supporting the validity of this approach, of the 77 unique clonotypes (within patients) with TCRs similar to the SARS-CoV2 TCRs, 89.6% (69/77) were from patients with COVID-19 (Figure 5.22), as opposed to Flu, Healthy Controls or Sepsis. Additionally, 94.8% (73/77) of these matched clonotypes were

annotated as CD4⁺ T-cells under `Annotation_major_subset`. As shown in Figure 5.23, a subset of COMBAT patients have HLA typing available, but within these just under half (49.2%, 31/63) of unique clonotype matches with available HLA typing have the relevant HLA alleles for the epitopes of interest (i.e. the match with a M172-DR01 specific TCR occurs within a COMBAT patient expressing *HLA-DRB1*01:01*).

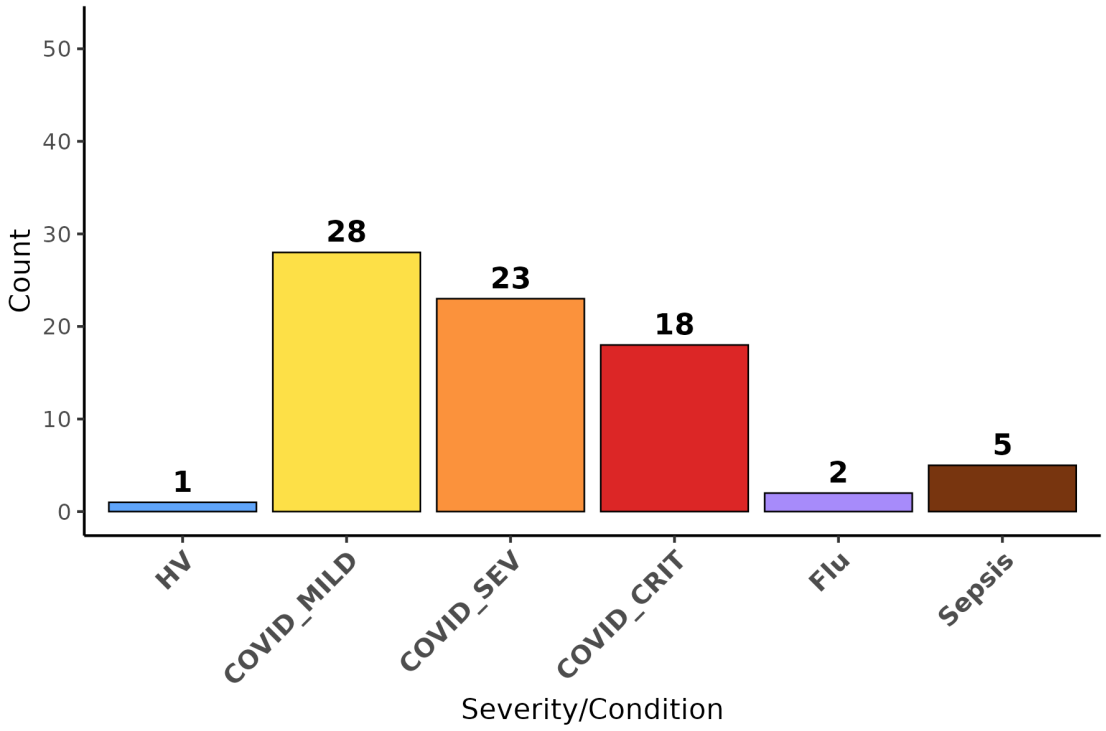


Figure 5.22: COMBAT disease category distribution for clones matched to sampled Spike or non-Spike specific CD4⁺ T-cells.

Acknowledging the uncertainty around HLA presentation and the small number of matches, I compared the distribution of these putative Spike and non-Spike specific CD4⁺ T-cells across the transcriptional clusters defined in the COMBAT dataset (Figure 5.24). While cells annotated as `CD4.TEFF` make up 11% of non-Spike matches, they make up 34% of Spike matches. Similarly, `CD4.TEM/TEMRA` cells make up 24% of Spike matches but none of the non-Spike matches. Conversely while `CD4.NAIVE` make up 26% of non-Spike matches, they make up 11% of Spike matches. This supports further investigation of the priming of these cells, as it could provide a link between priming, clonal expansion within effector subsets and long-term phenotypes. However, the small number of matches and uncertainty around HLA

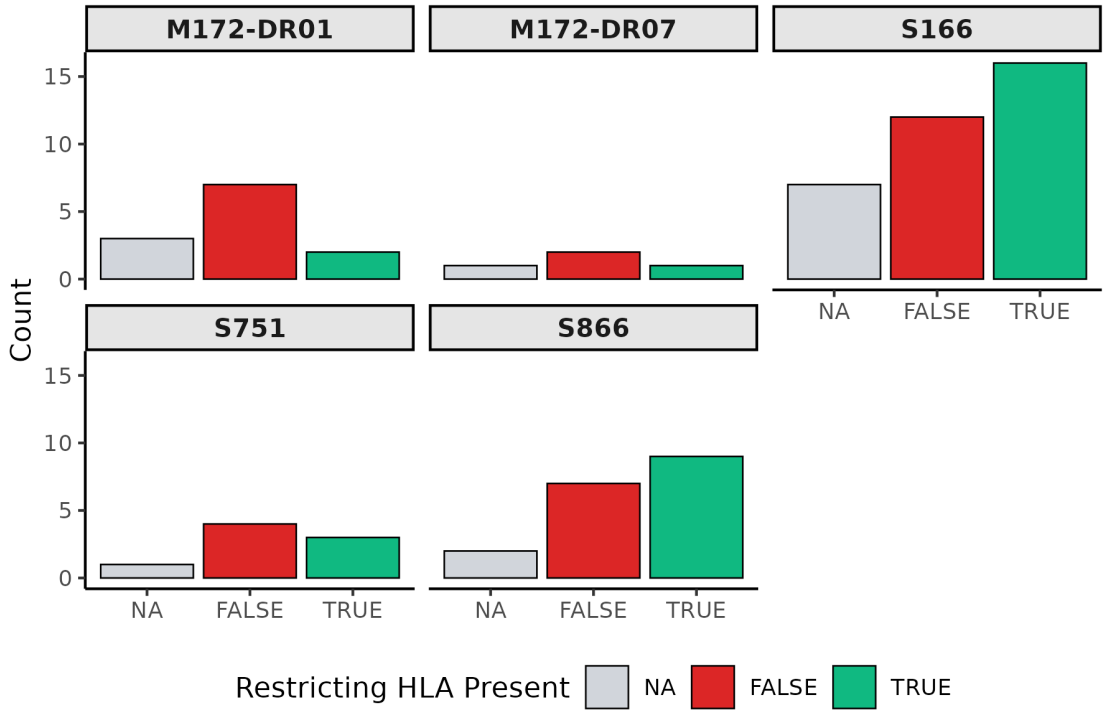


Figure 5.23: Presence, absence or non-availability of the restricting HLA allele for unique similar clonotypes for epitopes sampled in this study.

presentation means these results should be interpreted with caution and more as hypothesis generating.

In this chapter I have described the long-term evolutionary trajectories for Spike and non-Spike specific CD4⁺ T-cells following SARS-CoV2 infection. At the four year timepoint, Spike-specific CD4⁺ T-cells maintain their diversity, whereas non-Spike specific CD4⁺ T-cells show a significant reduction in diversity. While one key difference between these cell groups is an expectation that the Spike-specific cells are recalled by vaccination, leading to more frequent stimulation and therefore maintenance, in this data we find a role for clonal size. Spike-specific repertoires are more clonal with more expanded clones present from the first sample. Since, in our data, clonal size at the early timepoint is the strongest predictor of clonotype persistence to the four year timepoint, Spike-specific clones persist more frequently than non-Spike specific clones. At the four year timepoint, Spike-specific CD4⁺ T-cells express a range of interferon-stimulated genes, co-stimulatory molecules and cytotoxic genes - reflecting a cytotoxic CD4⁺ cluster only observed in Spike-specific samples.

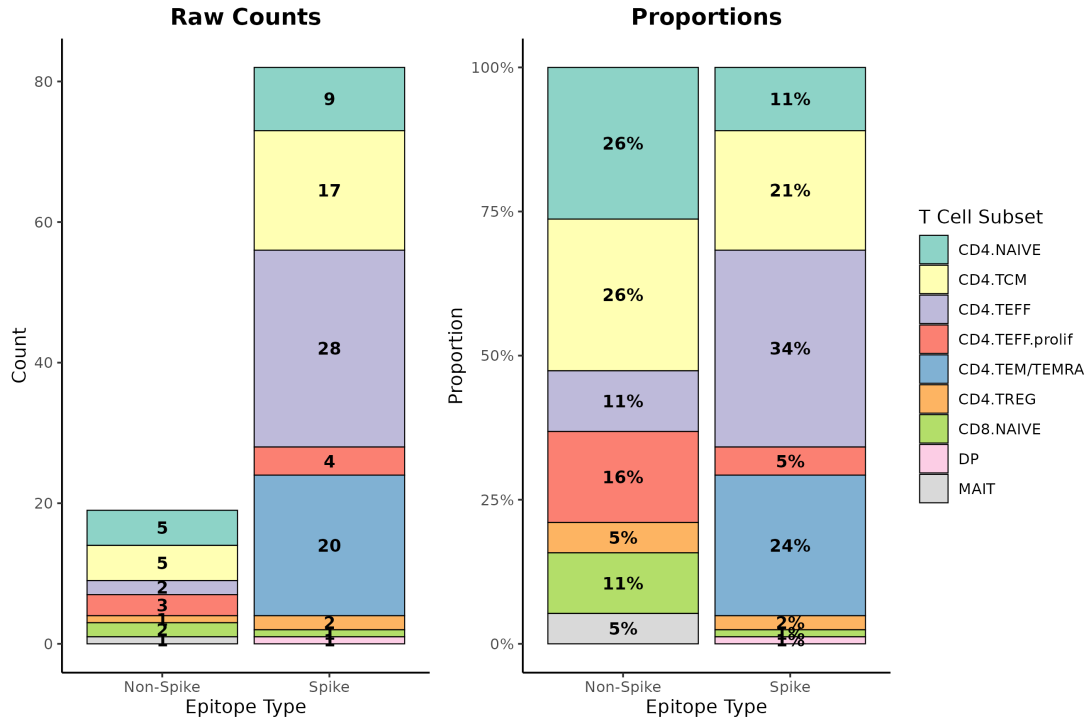


Figure 5.24: The distribution of `Annotation_minor_subset` of T-cells with similar clonotypes to sampled Spike and non-Spike specific CD4⁺ T-cells. The left panel reflects the exact counts of cells, whereas the right panel represents the proportions.

Other differentially expressed genes included migratory chemokine receptors and exhaustion markers. Non-Spike specific CD4⁺ T-cells expressed fewer activation, co-stimulatory or migratory genes, potentially reflecting a more quiescent phenotype. However, no gene ontologies were significantly enriched between Spike and non-Spike specific cells at either timepoint, which could reflect the heterogeneity of the cohort. Some genes were consistently differentially expressed between Spike and non-Spike specific cells at both timepoints, including *PDCD1* (PD-1) and *TNFRSF4* (OX40) which were more highly expressed in Spike-specific cells at both timepoints. Re-analysis of the COMBAT dataset provides some hypotheses about the acute state of these cells, potentially providing a link between their priming, initial phenotypes, clonal expansion and persistence.

5.3 Discussion

The durability and evolution of T-cell memory following SARS-CoV2 infection has implications for predicting long-term immunity and informing vaccine strategies[202]. Antibodies,

a core correlate of protection, wane over time[400]. Cellular immunity, in particular CD4⁺ T-cell memory, provides sustained protection through facilitating antibody production, enhancing CD8⁺ T-cell responses[95] and direct antiviral effects[401]. Previous studies characterised responses up to two years post-infection[270,319], and few had directly compared how epitope-specific repertoires evolve in the context of different recall patterns.

This study uses a unique cohort involving longitudinal sampling of individual's CD4⁺ T-cell responses to Spike and a non-Spike epitope over four years - spanning the period when mRNA vaccines (containing only Spike antigen) were broadly deployed. Our cohort provides observational data on how Spike-specific and non-Spike-specific repertoires evolve differently, though we cannot definitively attribute differences to specific recall patterns due to the lack of documented exposure histories.

At four years post initial infection, non-Spike repertoires showed substantially greater Shannon diversity decline (a decrease of 1.45) compared to Spike repertoires (a decrease of 0.25, non-significant). This might have functional implications, as clonal diversity of responses against an epitope at the initial timepoint can predict the persistence of the overall epitope-specific response[402]. In the short term, following Spike and non-Spike specific repertoires at most 80 days after their most recent vaccination, differences in the diversity of Spike-specific and non-Spike specific TCRβ are not apparent[253], highlighting the importance of long term data.

While mRNA vaccination both expands pre-existing memory clones and recruits from naive pools[396,397], I applied our cohort to investigate the maintenance of individual clonotypes long term. The clonal size at the early timepoint was the strongest predictor of persistence to the four year timepoint, with each log-fold increase in clone size increasing the odds of persistence by 8.3 fold. While Spike-specific clonotypes had higher raw persistence rates, this was not significant when accounting for clonal size and sampling depth, suggesting that part of the greater stability of Spike-specific repertoires is driven by their more clonal structure. Clone size hierarchies remain stable over time in healthy individuals[403], likely explained by random sampling[404], supporting that larger clones are more likely to be observed.

Relating clonal size, persistence and late timepoint phenotypes - Spike-specific cells highly express *CXCR5* at the late timepoint as well as the gene for PD-1. Phenotyping cells on the basis of their gene expression is challenging, particularly as cells highly expressing these genes were insufficiently different or common to form a separate cluster. However, the protein products of these genes define T_{fh} cells, so it is possible these represent circulating T_{fh} cells which are effectively induced by SARS-CoV2 infection[274]. Their proportion is associated with SARS-CoV2 directed antibodies[405], and given the differential expression design attempts to adjust for patient specific factors, this would provide circumstantial evidence that circulating Spike specific cells in memory have a more T_{fh} phenotype. Work by Wragg et al, albeit defining this group based on surface proteins, identifies that these respond rapidly to vaccination after SARS-CoV2 convalescence but wane by day 12[274]. Identifying these differentially expressed genes could reflect more recent exposure by Spike epitopes (e.g. by more recent vaccination), and therefore Spike-specific circulating T_{fh} cells being earlier on the decay trajectory, or that more frequent exposure might aggregate these cells in the circulation. Further work relating this subset as defined by flow cytometry to its expression profile and dynamics in hybrid immunity could resolve these questions.

The cluster most clonally expanded, FGFBP2 (5), is predominantly found at the late timepoint suggesting that the highly expanded clones maintained or developed effector phenotypes as they persist in the circulation. We do not investigate the affinity or avidity of the T-cell receptors in this study, but identifying where on the spectrum these clonally expanded long-persisting cells lie could provide important insights into how to drive long-lasting protective immunity[313,314].

Maintenance of repertoire diversity in Spike-specific repertoires challenges notions of repertoire narrowing with repeated antigen exposure. While long term evidence from endemic viruses such as Influenza suggest that repeated exposures[406] narrow repertoires[320], this work implies findings about vaccine-driven diversification of the repertoire continue to be relevant four years after initial infection.

At more than 3 years after initial infection Spike-specific cells express interferon-stimulated genes, PD-1 expression and cytotoxicity markers, suggesting functional and migratory ca-

pacity. The Spike-specific ISG signatures and cytotoxic phenotypes could represent long term infection imprinting, as shown using chromatin-accessibility assays 20 months after recovery from COVID-19[395], vaccine mediated recall[253] or breakthrough infection responses. Co-expression of exhaustion and activation markers (*PDCD1*, *TNFRSF4*) might reflect poised readiness for effector activation, rather than exhaustion, something that could be validated by isolating and functionally characterising this population in vitro.

Work by Liu et al. on the same group of Spike-specific samples finds that a subset expressing a dominant TCR alpha have expression of *GZMA* with functional viral suppression[325], however this population is insufficiently common to separate Spike and non-Spike specific samples on the basis of *GZMA* expression. The cytotoxic phenotype identified in this study (*GZMB*, *GZMH*), present most strikingly in two samples, could have capacity for protective respiratory viral suppression in Influenza[401] but equally has been implicated in "immunopathogenesis of severe SARS-CoV2"[85,325].

That the FGFBP2 (5) cluster is more frequent at four years than 1-3 months after initial infection suggests that these cytotoxic $CD4^+$ T-cells are a durable component of long-term immunity rather than a transient effector population in acute responses. Speculatively, accumulation of these over time with repeated exposures to diverse viral or other epitopes and the observed clonal expansion may explain their frequency and clonal dominance in supercentenarians[407].

Non-Spike samples showed higher expression of anergy-associated genes such as *EGR2*[408], *CREM*[409] and *DUSP6*[410], all of which associate with reduced responses to TCR stimulation. In the absence of repeated antigenic stimulation with SARS-CoV2 spike vaccines and boosters, non-Spike specific $CD4^+$ T-cells may drift towards quiescent or tolerogenic states.

A fundamental limitation of this work constrains causal interpretation of this data: we do not have vaccination or reinfection histories between the 1-3 month and 4 year timepoints. Exposure history shapes T-cell phenotypes[253,322,411], and without this information the plausible heterogeneity of the cohort in terms of number of infections and vaccinations, as well as their timing means I cannot definitively attribute the differences in repertoire dynamics and phenotypes to vaccination. Given the cohort's location in the UK, the "participants

likely underwent multiple vaccinations and possible infections by SARS-CoV2 variants"[325], and inherently since infection recalls both Spike and non-Spike T-cells whereas vaccination only recalls Spike T-cells, Spike T-cells will have at least as many recall events as non-Spike T-cells.

An additional concern is that the Spike-only and non-Spike-only groups differ substantially in age (median 61.5 vs 49.5), sex distribution (70% vs 100% male) and severity of initial infection (mixed vs all severe). While the pseudobulked differential expression analysis included patient of origin as a covariate as well as patients sampled for both epitope groups, the heterogeneity of the cohort could dilute epitope specific signals, potentially explaining why no gene ontologies reached significance. This cohort was assembled early in the SARS-CoV2 pandemic to generate evidence about immune memory to SARS-CoV2, and hence the attributes of the cohort such as age, sex and severity were not a priority to balance.

Single-cell profiling of CD4⁺s from PBMCs has inherent limitations. As the number of measurements greatly exceeds the number of independent trials (i.e. number of donors), and groups of samples were taken and sequenced in batches there is a meaningful risk of Type I error that can confound biological interpretation. Peripheral blood sampling also misses tissue-resident compartments, which could lead to paradoxical findings and misses key effectors in protective antiviral immunity. For example, groups of cells more primed to form niches in tissues such as the lung may show depletion in the circulation despite forming protective immunity[199]. Unfortunately, owing to the limitations of antigen-specific cell count and Smart-seq2 sequencing insufficient paired clonotypes were sampled for more complex analyses. TCR repertoire diversity measurements are highly sensitive to the number of cells sequenced, and while I have compromised by excluding samples with fewer than 5 paired clonotypes this can require more than 50 sampled receptors for these measurements to stabilize[412,413].

Further work on this should prospectively track exposure histories in age, sex and severity matched participants with vaccine diaries and serological or PCR confirmation of infection, allowing us to attribute phenotypic changes to immunological events. This cohort's

value is that it has followed individuals after their initial infection from the very beginning of the pandemic, a window that can't be recreated for endemic infections. Tissue sampling, particularly airway through bronchoalveolar lavage or lymphoid organs through fine-needle aspiration, would enable us to follow dynamics of circulating and tissue memory and explore how exposures influence the development and phenotypes of protective tissue immunity. While maintained clones directed at Spike epitopes[325] and non-Spike epitopes (Section 3.2.4) have been investigated with viral suppression assays, confounding on the basis of fewer samples and different individuals prevents direct comparisons being made between their functional characteristics. Future work on this should involve isolating clones directed at these epitopes from the same individuals systematically, such that intra-individual comparisons can be made.

Overall, these findings suggest that Spike-specific $CD4^+$ T-cell repertoires maintain diversity and activation phenotypes four years post infection, whereas non-Spike specific repertoires contract. While clone size is significant in the maintenance of clonotypes longitudinally, diversity is maintained through partial maintenance and the recruitment of new clonotypes to make up for lost ones. If repeated antigen exposure maintains repertoire diversity long term inclusion of non-Spike antigens in vaccines may facilitate maintained diverse immune responses more resilient to viral evolution. These dynamics may help inform strategies for broad, functional and long-term T-cell immunity.

5.4 Methods

5.4.1 Dimensionality Reduction and Clustering

Inspection of the first 50 principal components (PCs) on an elbow plot suggests that the first 7 PCs have the most meaningful standard deviation (Figure E.2) and clearly separate the top 500 cells of each PCA with interpretable weightings that include $CD4^+$ T-cell marker genes (Figure E.3). I therefore used the first 7 PCs for UMAP dimensionality reduction and clustering. Clustering was performed using the Louvain algorithm with a resolution of 0.4, yielding 7 clusters. Marker genes per cluster were identified via `FindAllMarkers` in `Seurat`, using default parameters. Clusters were manually annotated on the basis of these genes,

as well as canonical marker genes for CD4⁺ T-cell subsets. Visualisation of cluster marker genes was performed using the `pheatmap` package, and the resulting dendrogram reflects hierarchical clustering of the scaled, normalised and aggregated expression matrix.

5.4.2 TCR Repertoire Analysis

Gene Usage Analysis

V- and J- gene usage frequencies were calculated for each sample (combination of donor, timepoint and epitope) as the proportion of cells with that gene out of all cells where the appropriate chain was observed. V- and J- genes were only included in this analysis if they were present in more than 3 paired samples. For example, TRAV usage was calculated as the number of cells with a given TRAV gene divided by the total number of cells with an observed TRAV gene. Only samples with at least 3 cells with the relevant chain were included in the analysis, and only genes detected in 2 or more patients were considered. Genes with mean proportions below 3% (V genes) or 5% (J genes) across samples for that epitope were grouped into "Other" for visualisation. For patients with samples at both timepoints, temporal changes in gene usage were calculated as the difference in proportion of each gene between the two timepoints. Statistical significance of these changes was assessed using two-tailed Wilcoxon signed-rank tests, and multiple testing correction was performed using the Benjamini-Hochberg method.

TCR Repertoire Diversity Analysis

To quantify the diversity of TCR repertoires, we calculated the Shannon diversity for each sample (Section 2.2.3), excluding samples that contained fewer than 5 paired clonotypes. I fit the linear model by ordinary least squares using the `lm` function in R. The model included timepoint (1-3 Months vs four years), epitope type (Spike vs Non-Spike) and their interaction.

The design was `lm(shannon_diversity ~ timepoint * epitope)`.

TCR Clonotype Persistence Analysis

In the 7 patients for which we had samples at both timepoints, I tracked the persistence of individual clonotypes defined by their TCR α, TCR β or paired TCR chains. Only cells

with a CDR3 sequence and defined V- gene for the chain(s) considered were included in each analysis. Clonotypes observed in a given sample were aggregated into their clonal size, and were defined as persistent if the same clonotype (CDR3 sequence and V-gene) was observed in both timepoints.

Logistic regression models were fit using the `glm` function in R (binomial family, no random effects), and included epitope type (Spike vs Non-Spike), log of clone size at 1-3 months, and log of overall sampling depth (i.e. the number of cells with observed $\alpha\beta$ paired clonotypes) at four years as fixed effects.

Searching of Similar TCRs in the COMBAT Dataset

The COMBAT[34] dataset, including all single-cell RNA and TCR sequencing data was downloaded from Zenodo. HLA typing data was kindly made available on request from the authors. The COMBAT dataset was subset to those where the major cell type was `T`, and all cells that showed more than one TCR α or TCR β chain were excluded. Clonotypes were defined by their CDR3 amino acid sequence and V-gene for each chain. For each clonotype observed in our dataset, I searched for identical or similar clonotypes in the COMBAT dataset using the `pyrepseq` package's `nn.nearest_neighbour` function for both chains separately. Where both the alpha and beta chains had an edit distance of 2 or less, the COMBAT clonotype was considered a match. To interpret the subsets of matched cells, I used the `Annotation_minor_subset`.

5.4.3 Differential Gene Expression

Gene expression matrices were aggregated into Pseudobulks using `AggregateExpression` by combination of patient, timepoint and epitope. Patient of origin was included in the `DESeq2` design formula to account for inter-individual variation. The design was `~ patient + epitope_timepoint`, where `epitope_timepoint` is a combined factor of epitope (Spike vs Non-Spike) and timepoint (1-3 Months vs four years). Differential gene expression between Spike and Non-Spike specific cells was performed separately at each timepoint using the `results()` function in `DESeq2`. Genes with an adjusted $p < 0.05$ were considered significantly differentially expressed.

Gene set enrichment analysis was performed using `EnrichR`[339] with the `GO_Biological_Process_2025` database. Genes significantly differentially expressed (adjusted $p < 0.05$) in each direction were used as input gene sets separately. Terms with an adjusted $p < 0.05$ were considered significantly enriched.

6

Discussion

Summary

This thesis characterises antigen-specific T-cell responses to SARS-CoV2 across three experiments. Chapter 3 demonstrates that the same epitope presented by different HLA-DR alleles generates $CD4^+$ T-cells with distinct transcriptional profiles that associate with interpolated antigen load and functional studies. This includes the expression of cytotoxic molecules such as *GZMB*. Chapter 4 shows that exposure sequence influences $CD8^+$ T-cell phenotypes, with infection followed by vaccination associating with expression of *EOMES* and *XCL2*, whereas vaccination followed by breakthrough infection associates with *IL10* expression. The clones that were maintained through Omicron breakthrough infection were more likely to be found in effector-like clusters such as GNLY (0) and ABI3 (3). Chapter 5 shows that four years after infection, Spike-specific $CD4^+$ T-cells maintain interferon stimulated gene signatures and begin expressing cytotoxic markers while non-Spike specific cells express anergy associated genes. Spike-specific repertoires maintain their diversity over the four years, whereas non-Spike specific repertoires lose this. In all chapters, small sample sizes and the curse of dimensionality present statistical challenges that I have addressed through rigorous pseudobulking and conservative interpretation of results. In the future,

pre-registered analyses with power calculations and multi-modal single-cell data could help definitively profile T-cell memory using single-cell sequencing.

6.1 Context and Rationale

This thesis studies antigen-specific T-cells following SARS-CoV2 infection using single-cell analysis. By isolating these rare populations via peptide-MHC tetramers, I can explore how exposures influence the phenotypes and repertoires of these cells. My thesis has contributed findings across three themes: how dimensions of the exposure influence the short term phenotypes of SARS-CoV2 T-cells, how sequences of exposures combine, and how T-cell phenotypes influence their persistence. Then, I discuss the statistical challenges of small studies involving deeply characterised samples present across all chapters. Finally I conclude with the future directions for this work.

SARS-CoV2 emerging in late 2019 presented a unique opportunity: primary human T-cell responses to a novel pathogen could be studied in real time. T-cell dynamics and responses, whether from previous related exposures[257] or de-novo responses[95,414] could be tracked from the earliest timepoints of infection through to memory formation.

6.2 Unifying Themes

6.2.1 How Exposure shapes T-cell Phenotypes

TCR signalling involves a cacophony of interactions. The TCR itself[219,311], co-stimulatory or inhibitory molecules, and the cytokine milieu[129] all contribute to the signal delivered to the T-cell and its subsequent response. Even nuclear envelope invaginations at the time of antigen-stimulation bias the differentiation of naive T-cells[415]. This immense complexity is poorly addressed by the measurement of individual surface markers. This thesis uses single-cell analysis in T-cells with diverse exposures varying by interpolated antigen load, target protein and inferred cytokine milieu at the time of stimulation to explore some of this complexity.

In Chapter 3, $M_{172\text{-}188}$ specific $CD4^+$ T-cells restricted by different HLA-DRB1 alleles (that present the same epitope) showed distinct transcriptional profiles. DRB1*01:01 restricted

cells highly expressed effector genes (*GZMB*, *XCL1*, *CCR5*) and checkpoint molecules (*PDCD1*, *HAVCR2*) whereas DRB1*07:01 restricted cells showed more regulatory and inflammatory signatures (*IKZF4*, *NFKBID*, *NLRP3*). Interpolated antigen load in live viral co-culture implied that DRB1*01:01 presents higher effective $M_{172\text{-}188}$ concentrations than DRB1*07:01, providing evidence that more potent TCR-p-MHC signalling associates with more effector functions and states. We also observed differences in the binding cleft between the alleles, that could influence the quality of TCR engagement.

Chapter 5 examined how different epitope targets - Spike and non-Spike, influence the long-term phenotypes of SARS-CoV2 specific $CD4^+$ T-cells in memory. While late differences inherently associate to sequence and number, the most striking differences were found between Spike and non-Spike specific $CD4^+$ T-cells at the early timepoint, with 3706 differentially expressed genes apparent at 1-3 months. This could be potentially inflated, and a reflection of confounders in the underlying cohort such as the age, sex and severity distribution. However, this was the only analysis in this thesis that identified significantly enriched pathways, including `T Cell Receptor Signaling Pathway (GO:0050852)` enriched among Non-Spike specific cells. The enrichment of this pathway, despite stringent differential expression and pathway enrichment criteria, suggests that Non-Spike specific $CD4^+$ T-cells in the recovery phase from infection more highly express T-cell receptor signalling associated genes than Spike-specific ones. Whether this reflects better antigen-presentation, higher avidity TCRs or any of the cacophony of mediators involved is not clear from this data, but in-vitro work should explore this using patients that generate responses to both.

However, even single-cell sequencing, which can profile 20,000 genes, can be insufficient to resolve important subsets. Work by Zhang et al. shows the subset of circulating $CD8^+$ T-cells most potently responding to antigen after vaccination is not readily apparent under single-cell analysis and clustering, and requires the integration of ATAC-Seq data to identify[230]. This presents another natural avenue for the expansion of work on exposures and phenotypes - the integration of multi-modal data.

6.2.2 How Sequences of Exposures Combine

In Chapter 4, I find that different exposure order is associated with different CD8⁺ T-cell phenotypes. Comparing four exposures to Spike epitope in different orders - infection followed by three vaccinations (infection-first) versus three vaccinations followed by breakthrough infection (vaccine-first) showed that 280 genes are enriched in infection-first cells and 71 in vaccine-first cells. Infection-first cells highly expressed *EOMES*, *XCL2* and *PRDM1*, suggesting differentiation towards central memory cells, whereas vaccine-first cells were enriched for immunoregulatory cytokine *IL10*.

The distribution of transcriptional clusters reinforced sequence-dependent effects. Terminal effector-like cluster GNLY (0) made up a large proportion of infection-primed samples at both timepoints, whereas more effector-memory-like ABI3 (3) made up the majority of vaccine-primed samples. Minervina et al. also found that infection followed by vaccination expanded a TEMRA-like cluster expressing *GZMB* and *GNLY*[253]. Despite different epitopes, sampling times and technical approaches infection priming associates with durable cytotoxic effector programs that are not reprogrammed by subsequent vaccination.

Chapter 5 implicitly compares sequences of exposures that include infections and vaccinations against infections only by comparing Spike and non-Spike specific CD4⁺ T-cells. Since UK-approved vaccines contain only Spike antigens, Spike-specific cells probably encountered more frequent recall than non-Spike specific cells. At four years after initial infection, Spike CD4⁺ T-cells maintained ISG expression (*OAS2*, *OAS3*, *RNASEL*, *MAVS*) alongside cytotoxic molecules (*GZMB*, *GZMH*) and exhaustion marker *PDCD1* (PD-1). The maintained ISG signature corresponds to that Gray-Gaillard et al. identify in single-cell sequencing and chromatin accessibility studies as imprinted on Spike-specific cells by infection priming[395]. My findings show that this persists years after initial priming, and that these effector phenotypes could require recall (via vaccination) for their maintenance. Non-Spike specific cells expressed anergy associated genes (*EGR2*, *CREM*, *DUSP6*). The maintained activation and ISG signature in Spike-specific cells four years post-infection suggest that repeated antigen exposure in the context of vaccination sustains effector-like phenotypes.

Mechanisms linking exposure type to phenotype are not fully resolved - while studies have characterised the immediate responses to human challenge[84] and mRNA vaccination[299,416], direct comparison in the same cohorts is lacking. It is likely that inflammation during natural infection provides signals absent during vaccination, which could include the interferon signalling activated in the blood by day 2-3 and then in the nasopharynx by day 5 in the human challenge study. In fact, the profound interferon signalling induced by infection could imprint the ISG signature seen in infection-primed $CD4^+$s long term. If individuals in the human challenge trial were followed up long-term, it would be fascinating to compare how the magnitude of the initial ISG signature relates to their Spike-specific $CD4^+$ ISG profile and chromatin accessibility.

6.2.3 How T-cell Phenotypes affect their Clonal Persistence

Clone size at initial sampling was a dominant predictor of clonotype persistence in Chapters 4 and 5. In Chapter 5, each log-fold increase in clone size increased persistence odds by 8.3 fold for paired $\alpha\beta$ clonotypes over four years. Chapter 4 explored how the distribution of clone members across different transcriptional clusters relates to persistence, finding that clonotypes with more cells in effector clusters such as the Cytotoxic TEMRA-like GNLY (0) and Effector-Memory-like ABI3 (3) showed the highest maintenance odds. Since these clusters contained the highest proportion of large clones, transcriptional state and clonal expansion strongly relate.

In Chapter 5, I identify that large clones at 1-3 months predict persistence at four years, supporting work that large clones persist because of their size[403,404] and the relative stability of the TCR repertoire. However, could the phenotypic features that enable expansion also promote longevity? The finding that more cells in the GNLY (0) cluster mean that clonotype is more likely to be observed after Omicron infection suggests that terminal effector-like phenotypes (with GNLY, *GZMH*, *GZMB*) have survival advantages beyond expansion alone. Detecting a clonotype at a subsequent sample requires it to be at a measurable frequency amongst PBMCs, meaning that clusters with markers relating to localisation or trafficking to other systems should persist less as they vacate the blood.

While this holds true for KIF23 (4) with high expression of lymphoid homing markers *CCR7* and *SELL* and a trend towards reduced clonotype maintenance, GNLY (0) highly expresses the tissue-residency marker *ZNF683*. Novel experimental techniques such as mitochondrial DNA lineage tracing[417] could help identify the origin cell for these highly expanded populations, identifying whether they originate from a highly proliferative set of cells that release TEMRA-like clones to populate tissues, or whether the TEMRA-like cells are just highly persistent.

6.3 Statistical Challenges in Small-n Single-Cell Studies

As technologies improve for the deep profiling of rare cell populations, we generate extremely rich data to explore key immunological phenomena. However, as these technologies generate more measurements per independent sample, they suffer from the curse of dimensionality. The most frequently used and default methods to compare cell types "can discover hundreds of differentially expressed genes in the absence of biological differences"[418]. In essence, statistical power depends on biological replicates, not cell numbers[419]. The nature of sampling humans is that immeasurable diversity exists between every donor, and hence making thousands of comparisons between small groups introduces variance that obscures "genuine" biological programs.

Single cell analysis, particularly of antigen-specific cells, reflects a deep but narrow view of the immune response. The cells captured can be analysed in great detail revealing their T-cell receptors and their gene expression in the context of their antigen-specificity. However, these cells represent a small fraction of the total immune response, and only those that happen to be circulating near the sampling site and amenable to sorting with tetramers. As a result, the cells analysed do not necessarily reflect the overall immune response generated in concert with a variety of cell types and tissues, and may be biased towards particular phenotypes more likely to circulate. Indeed, this finding may be paradoxical[199]. For example, by sampling only circulating cells we may observe larger quantities of cells circulating in one condition than another - here, it is unclear whether there are more cells, or whether these cells are more likely to circulate rather than migrating to tissues.

Since statistical power is limited by the number of independent samples, which in this thesis is low, the risk of false-positives is very high. The end goal of purely computational work should be to provide testable hypotheses to drive future experimental work, and thus a false positive represents wasted time, resources and effort. As such, I have aimed to maximise the specificity of my work, even where that risks its sensitivity. To do so I have rigorously applied pseudobulked comparisons to avoid the fallacy that multiple cells sampled from the same individual are independent trials[418]. I have explicitly described the interpretation of manually curated genes or gene sets as post-hoc, and to be interpreted not as a finding in itself but a descriptor of the underlying data. Additionally, I have ensured that where feasible patient of origin is a covariate in all analyses to account for inter-individual variation.

Chapter 3 included four donors sampled with single-cell sequencing, with two donors sampled for both DRB1*01:01 and DRB1*07:01 restricted responses. Chapter 5's comparison of Spike and non-Spike specific $CD4^+$ T-cells involved groups differing in age and sex distribution. Other than the two donors where Spike-specific and non-Spike specific samples were jointly present in Chapter 5, the severity distribution of Spike-specific and non-Spike specific cells is very different. While in all cases the pseudobulked differential expression analyses were designed to minimise the impact of this, such as by using patient of origin as a covariate, the small number of patients means individual variation can meaningfully influence results.

These cohorts were put together early in the pandemic to investigate primary responses, not to test specific hypotheses that require matched cohorts. The value of these cohorts lies in the longitudinal, long-term tracking of their T-cell immune response including the beginning of the pandemic, something infeasible for endemic infections. Prospective studies involving symptom or vaccine diaries and balanced demographics could support these conclusions.

6.4 Methodological Challenges

Isolation of antigen-specific T-cells via peptide-MHC tetramers, while well suited to the study of rare, ex-vivo populations[367], presents several challenges. First, tetramer sorting can identify non-reactive clonotypes[55], and since tetramers are often used as the "Gold Standard", it has not been quantified exactly how sensitive or specific they are.

Second, tetramers are not completely sensitive, and are biased towards high-affinity TCRs[420]. While one would expect the bias to be consistent across timepoints, and therefore not influence persistence analyses, it may artificially constrain clonotype diversity or alter the profile of cells sampled.

The lack of documented vaccination and reinfection histories between timepoints in Chapter 5 limits the interpretation of results. Exposure history shapes T-cell phenotypes[253,322,395,411] and participants likely underwent multiple vaccinations as well as possible breakthrough infections in the period 2021-2024.

6.5 Conclusions and Future Directions

This thesis makes three core findings relevant to the immunology of SARS-CoV2 specific T-cells in memory. First, that different HLA alleles presenting the same immunodominant Class II epitope can shape early T-cell differentiation in humans, with higher interpolated antigen-presentation concentration associating with effector gene expression. Second, that depending on the pre-existing exposures the $CD8^+$ T-cell response to breakthrough infection differs, and that maintenance of clonotypes through breakthrough infection is related to pre-infection phenotypes. Third, that four years after infection, Spike-specific $CD4^+$ T-cells maintain interferon stimulated gene signatures and begin expressing cytotoxic markers while non-Spike specific cells express anergy associated genes.

For vaccination, these results suggest priming context has durable consequences that are not erased by boosting. It is not clear from these results whether the Spike phenotypes result from this imprinting or repeated recall via vaccination. While I can speculate, and make the assumption that greater mRNA levels reflect function, single-cell RNA work

cannot demonstrate functional effects. Do cytotoxic Spike CD4⁺ T-cells provide superior protection, or do they contribute to immunopathology? Could vaccine-first *IL10* expression represent suboptimal priming or appropriate regulation?

Future work should involve (1) prospective cohorts with documented exposures, balanced demographics and adequate power that is pre-calculated[419]. This is possible[421], but extremely ambitious. Part of the reason I have returned to clinical school is to bring what I have learnt about the design and implementation of high-dimensional experiments into the recruitment of clinical cohorts throughout my career. (2) Tissue sampling via bronchoalveolar lavage[199] or fine-needle aspiration[273] to analyse circulating and tissue-resident populations in parallel. (3) Functional validation through viral suppression assays of identified T-cell populations, including comparison of Chapter 4's GNLY (0) cluster and KIF23 (4) cluster.

These findings suggest initial priming creates lasting effects on subsequent immune trajectories.

References

1. 10x Genomics. A new way of exploring immunity–linking highly multiplexed antigen recognition to immune repertoire and phenotype. *Tech. rep* (2019).
2. Murphy, K., Weaver, C., Berg, L. & Janeway, C. *Janeway's immunobiology*. (W.W. Norton; Company, New York, NY, 2022).
3. Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics* **11**, 47–59 (2010).
4. Liston, A., Humblet-Baron, S., Duffy, D. & Goris, A. Human immune diversity: from evolution to modernity. *Nature Immunology* **22**, 1479–1489 (2021).
5. Strassburg, M. A. The global eradication of smallpox. *American Journal of Infection Control* **10**, 53–59 (1982).
6. Rojas, L. A. *et al.* Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. *Nature* **618**, 144–150 (2023).
7. Klebanoff, C. A., Chandran, S. S., Baker, B. M., Quezada, S. A. & Ribas, A. T cell receptor therapeutics: immunological targeting of the intracellular cancer proteome. *Nature Reviews Drug Discovery* 1–22 (2023) doi:10.1038/s41573-023-00809-z.
8. Kovacs, M., Dominguez-Belloso, A., Ali-Moussa, S. & Deczkowska, A. Immune control of brain physiology. *Nature Reviews Immunology* 1–13 (2025) doi:10.1038/s41577-025-01129-6.
9. Diamond, M. S. & Kanneganti, T.-D. Innate immunity: the first line of defense against SARS-CoV-2. *Nature Immunology* **23**, 165–176 (2022).
10. Sette, A. & Crotty, S. Adaptive immunity to SARS-CoV-2 and COVID-19. *Cell* **184**, 861–880 (2021).
11. Iwasaki, A. & Medzhitov, R. Control of adaptive immunity by the innate immune system. *Nature Immunology* **16**, 343–353 (2015).
12. Nimmerjahn, F. & Ravetch, J. V. Fcγ receptors as regulators of immune responses. *Nature Reviews Immunology* **8**, 34–47 (2008).
13. Kuri-Cervantes, L. *et al.* Comprehensive mapping of immune perturbations associated with severe COVID-19. *Science Immunology* **5**, eabd7114 (2020).
14. Lee, D. *et al.* Inborn errors of OAS–RNase l in SARS-CoV-2–related multisystem inflammatory syndrome in children. *Science* **379**, eabo3627 (2022).
15. Cooper, M. D. & Alder, M. N. The evolution of adaptive immune systems. *Cell* **124**, 815–822 (2006).
16. Peng, Y. *et al.* An immunodominant NP105–113-B*07:02 cytotoxic T cell response controls viral replication and is associated with less severe COVID-19 disease. *Nature Immunology* **23**, 50–61 (2022).

17. Gressier, E. *et al.* CD4+ T cell calibration of antigen-presenting cells optimizes antiviral CD8+ T cell immunity. *Nature Immunology* 1–12 (2023) doi:10.1038/s41590-023-01517-x.
18. Cele, S. *et al.* Omicron extensively but incompletely escapes pfizer BNT162b2 neutralization. *Nature* **602**, 654–656 (2022).
19. Pishesha, N., Harmand, T. J. & Ploegh, H. L. A guide to antigen processing and presentation. *Nature Reviews Immunology* **22**, 751–764 (2022).
20. Thomas, O. G. *et al.* Cross-reactive EBNA1 immunity targets alpha-crystallin B and is associated with multiple sclerosis. *Science Advances* **9**, eadg3032 (2023).
21. Sewell, A. K. Why must T cells be cross-reactive? *Nature Reviews Immunology* **12**, 669–677 (2012).
22. Linette, G. P. *et al.* Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* **122**, 863–871 (2013).
23. Ashby, K. M. & Hogquist, K. A. A guide to thymic selection of T cells. *Nature Reviews Immunology* 1–15 (2023) doi:10.1038/s41577-023-00911-8.
24. ElTanbouly, M. A. & Noelle, R. J. Rethinking peripheral T cell tolerance: checkpoints across a T cell's journey. *Nature Reviews Immunology* **21**, 257–267 (2021).
25. Li, F. & Tian, Z. The liver works as a school to educate regulatory immune cells. *Cellular & Molecular Immunology* **10**, 292–302 (2013).
26. Künzli, M. & Masopust, D. CD4+ T cell memory. *Nature Immunology* **24**, 903–914 (2023).
27. Koh, C.-H., Lee, S., Kwak, M., Kim, B.-S. & Chung, Y. CD8 T-cell subsets: heterogeneity, functions, and therapeutic potential. *Experimental & Molecular Medicine* **55**, 2287–2299 (2023).
28. Farber, D. L. Form and function for T cells in health and disease. *Nature Reviews Immunology* **20**, 83–84 (2020).
29. Reinherz, E. L., Kung, P. C., Goldstein, G. & Schlossman, S. F. Separation of functional subsets of human T cells by a monoclonal antibody. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 4061–4065 (1979).
30. Reinherz, E. L. *et al.* Antibody directed at a surface structure inhibits cytolytic but not suppressor function of human T lymphocytes. *Nature* **294**, 168–170 (1981).
31. Shiku, H. *et al.* Expression of T-cell differentiation antigens on effector cells in cell-mediated cytotoxicity in vitro. Evidence for functional heterogeneity related to the surface phenotype of T cells. *Journal of Experimental Medicine* **141**, 227–241 (1975).
32. Takeuchi, A. & Saito, T. CD4 CTL, a cytotoxic subset of CD4+ T cells, their differentiation and function. *Frontiers in Immunology* **8**, (2017).
33. Zhang, S., Wu, M. & Wang, F. Immune regulation by CD8+ Treg cells: novel possibilities for anticancer immunotherapy. *Cellular & Molecular Immunology* **15**, 805–807 (2018).
34. Ahern, D. J. *et al.* A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* **185**, 916–938.e58 (2022).
35. Chu, Y. *et al.* Pan-cancer T cell atlas links a cellular stress response state to immunotherapy resistance. *Nature Medicine* **29**, 1550–1562 (2023).

36. Terekhova, M. *et al.* Single-cell atlas of healthy human blood unveils age-related loss of NKG2C+GZMB–CD8+ memory T cells and accumulation of type 2 memory T cells. *Immunity* **56**, 2836–2854.e9 (2023).
37. Roifman, C. M., Zhang, J., Chitayat, D. & Sharfe, N. A partial deficiency of interleukin-7Rα is sufficient to abrogate T-cell development and cause severe combined immunodeficiency. *Blood* **96**, 2803–2807 (2000).
38. Bekker, L.-G. *et al.* HIV infection. *Nature Reviews Disease Primers* **9**, 1–21 (2023).

39. Garboczi, D. N. *et al.* Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* **384**, 134–141 (1996).
40. Szeto, C., Lobos, C. A., Nguyen, A. T. & Gras, S. TCR recognition of peptide–MHC-I: Rule makers and breakers. *International Journal of Molecular Sciences* **22**, 68 (2020).
41. Dong, D. *et al.* Structural basis of assembly of the human T cell receptor–CD3 complex. *Nature* **573**, 546–552 (2019).
42. Giudicelli, V. *et al.* IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Research* **34**, D781–D784 (2006).
43. Sethna, Z., Elhanati, Y., Callan, C. G. J., Walczak, A. M. & Mora, T. OLGA: Fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics* **35**, 2974–2981 (2019).
44. Isacchini, G., Walczak, A. M., Mora, T. & Nourmohammad, A. Deep generative selection models of T and B cell receptor repertoires with soNNia. *Proceedings of the National Academy of Sciences* **118**, e2023141118 (2021).
45. Raybould, M. I. J. *et al.* The Observed T cell receptor Space database enables paired-chain repertoire mining, coherence analysis and language modelling. doi:10.1101/2024.05.20.594960.
46. Tanno, H. *et al.* Determinants governing T cell receptor α/β-chain pairing in repertoire formation of identical twins. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 532–540 (2020).
47. Mayer, A. & Callan, C. G. Measures of epitope binding degeneracy from T cell receptor repertoires. *Proceedings of the National Academy of Sciences* **120**, e2213264120 (2023).
48. Rossjohn, J. *et al.* T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. *Annual Review of Immunology* **33**, 169–200 (2015).
49. Sender, R. *et al.* The total mass, number, and distribution of immune cells in the human body. *Proceedings of the National Academy of Sciences* **120**, e2308511120 (2023).
50. Qi, Q. *et al.* Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 13139–13144 (2014).
51. Wooldridge, L. *et al.* A single autoimmune T cell receptor recognizes more than a million different peptides. *The Journal of Biological Chemistry* **287**, 1168–1177 (2012).
52. Veiga-Fernandes, H., Walter, U., Bourgeois, C., McLean, A. & Rocha, B. Response of naïve and memory CD8+ T cells to antigen stimulation in vivo. *Nature Immunology* **1**, 47–53 (2000).

53. Garrido-Mesa, J. & Brown, M. A. Antigen-driven T cell responses in rheumatic diseases: insights from T cell receptor repertoire studies. *Nature Reviews Rheumatology* **21**, 157–173 (2025).
54. Dolton, G. *et al.* HLA A*24:02–restricted T cell receptors cross-recognize bacterial and preproinsulin peptides in type 1 diabetes. *The Journal of Clinical Investigation* **134**, (2024).
55. Bodansky, A. *et al.* Molecular mimicry in multisystem inflammatory syndrome in children. *Nature* **632**, 622–629 (2024).
56. Nesamari, R. *et al.* Post-pandemic memory T cell response to SARS-CoV-2 is durable, broadly targeted, and cross-reactive to the hypermutated BA.2.86 variant. *Cell Host & Microbe* **32**, 162–169.e3 (2024).
57. Dolton, G. *et al.* Targeting of multiple tumor-associated antigens by individual T cell receptors during successful cancer immunotherapy. *Cell* **0**, (2023).
58. Zhu, L. *et al.* scRNA-seq revealed the special TCR β & α V(D)J allelic inclusion rearrangement and the high proportion dual (or more) TCR-expressing cells. *Cell Death & Disease* **14**, 1–11 (2023).
59. McMaster, B., Thorpe, C. J., Rossjohn, J., Deane, C. M. & Koohy, H. Quantifying conformational changes in the TCR:pMHC-I binding interface. *Frontiers in Immunology* **15**, (2024).
60. Henderson, J., Nagano, Y., Milighetti, M. & Tiffeau-Mayer, A. Limits on inferring T cell specificity from partial information. *Proceedings of the National Academy of Sciences* **121**, e2408696121 (2024).
61. Wellington, D., Yin, Z., Kessler, B. M. & Dong, T. Immunodominance complexity: lessons yet to be learned from dominant T cell responses to SARS-COV-2. *Current Opinion in Virology* **50**, 183–191 (2021).
62. Toyama, B. H. & Hetzer, M. W. Protein homeostasis: Live long, won't prosper. *Nature reviews. Molecular cell biology* **14**, 55–61 (2013).
63. Roche, P. A. & Furuta, K. The ins and outs of MHC class II-mediated antigen processing and presentation. *Nature reviews. Immunology* **15**, 203–216 (2015).
64. Wieczorek, M. *et al.* Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation. *Frontiers in Immunology* **8**, (2017).
65. Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
66. Sharon, E. *et al.* Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nature Genetics* **48**, 995–1002 (2016).
67. Garcia, K. C. Reconciling views on T cell receptor germline bias for MHC. *Trends in Immunology* **33**, 429–436 (2012).
68. Gray, G. I. *et al.* The Evolving T Cell Receptor Recognition Code: The Rules Are More Like Guidelines. *Immunological Reviews* **329**, e13439 (2025).
69. Zareie, P. *et al.* Canonical T cell receptor docking on peptide–MHC is essential for T cell signaling. *Science* **372**, eabe9124 (2021).

70. Krovi, S. H., Kappler, J. W., Marrack, P. & Gapin, L. Inherent reactivity of unselected TCR repertoires to peptide-MHC molecules. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 22252–22261 (2019).
71. Lu, J. *et al.* Molecular constraints on CDR3 for thymic selection of MHC-restricted TCRs from a random pre-selection repertoire. *Nature Communications* **10**, 1019 (2019).
72. Blattman, J. N. *et al.* Estimating the precursor frequency of naive antigen-specific CD8 T cells. *The Journal of Experimental Medicine* **195**, 657–664 (2002).
73. Jenkins, M. K. & Moon, J. J. The role of naive T cell precursor frequency and recruitment in dictating immune response magnitude. *The Journal of Immunology* **188**, 4135–4140 (2012).
74. Legoux, F. *et al.* Impact of TCR reactivity and HLA phenotype on naive CD8 T cell frequency in humans. *The Journal of Immunology* **184**, 6731–6738 (2010).
75. Alanio, C., Lemaitre, F., Law, H. K. W., Hasan, M. & Albert, M. L. Enumeration of human antigen–specific naive CD8+ T cells reveals conserved precursor frequencies. *Blood* **115**, 3718–3725 (2010).
76. Kwok, W. W. *et al.* Frequency of epitope-specific naive CD4+ T cells correlates with immunodominance in the human memory repertoire. *The Journal of Immunology* **188**, 2537–2544 (2012).
77. Nguyen, T. H. O. *et al.* CD8+ T cells specific for an immunodominant SARS-CoV-2 nucleocapsid epitope display high naive precursor frequency and TCR promiscuity. *Immunity* **54**, 1066–1082.e5 (2021).
78. Yin, Y., Wang, X. X. & Mariuzza, R. A. Crystal structure of a complete ternary complex of T-cell receptor, peptide–MHC, and CD4. *Proceedings of the National Academy of Sciences* **109**, 5405–5410 (2012).
79. Crotty, S. Follicular helper CD4 T cells (TFH). *Annual Review of Immunology* **29**, 621–663 (2011).
80. Dikiy, S. & Rudensky, A. Y. Principles of regulatory T cell function. *Immunity* **56**, 240–255 (2023).
81. Klawon, D. E. J. *et al.* Regulatory T cells constrain T cells of shared specificity to enforce tolerance during infection. *Science* **0**, eadk3248 (2025).
82. Fiorentino, D. F. *et al.* IL-10 acts on the antigen-presenting cell to inhibit cytokine production by Th1 cells. *Journal of Immunology (Baltimore, Md.: 1950)* **146**, 3444–3451 (1991).
83. Deaglio, S. *et al.* Adenosine generation catalyzed by CD39 and CD73 expressed on regulatory T cells mediates immune suppression. *Journal of Experimental Medicine* **204**, 1257–1265 (2007).
84. Lindeboom, R. G. H. *et al.* Human SARS-CoV-2 challenge uncovers local and systemic response dynamics. *Nature* 1–10 (2024) doi:10.1038/s41586-024-07575-x.
85. Meckiff, B. J. *et al.* Imbalance of regulatory and cytotoxic SARS-CoV-2-reactive CD4+ T cells in COVID-19. *Cell* **183**, 1340–1353.e16 (2020).
86. Weiskopf, D. *et al.* Dengue virus infection elicits highly polarized CX3CR1+ cytotoxic CD4+ T cells associated with protective immunity. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E4256–E4263 (2015).

87. Cenerenti, M., Saillard, M., Romero, P. & Jandus, C. The era of cytotoxic CD4 T cells. *Frontiers in Immunology* **13**, (2022).
88. Adhikary, D., Damaschke, J., Mautner, J. & Behrends, U. The epstein-barr virus major tegument protein BNRF1 is a common target of cytotoxic CD4+ T cells. *Journal of Virology* **94**, e00284–20 (2020).
89. Gao, G. F. *et al.* Crystal structure of the complex between human CD8$\alpha\alpha$ and HLA-A2. *Nature* **387**, 630–634 (1997).
90. Bangert, C. *et al.* Persistence of mature dendritic cells, TH2A, and Tc2 cells characterize clinically resolved atopic dermatitis under IL-4Rα blockade. *Science Immunology* **6**, eabe2749 (2021).
91. Agle, K. *et al.* Bim regulates the survival and suppressive capability of CD8+ FOXP3+ regulatory T cells during murine GVHD. *Blood* **132**, 435–447 (2018).
92. Thiery, J. *et al.* Perforin pores in the endosomal membrane trigger the release of endocytosed granzyme B into the cytosol of target cells. *Nature Immunology* **12**, 770–777 (2011).
93. Fu, Q. *et al.* Structural Basis and Functional Role of Intramembrane Trimerization of the Fas/CD95 Death Receptor. *Molecular Cell* **61**, 602–613 (2016).
94. Tewari, K., Nakayama, Y. & Suresh, M. Role of direct effects of IFN-γ on T cells in the regulation of CD8 T cell Homeostasis1. *The Journal of Immunology* **179**, 2115–2125 (2007).
95. Peng, Y. *et al.* Broad and strong memory CD4+ and CD8+ T cells induced by SARS-CoV-2 in UK convalescent individuals following COVID-19. (2020) doi:10.1038/S41590-020-0782-6.
96. Laidlaw, B. J., Craft, J. E. & Kaech, S. M. The multifaceted role of CD4+ T cells in CD8+ T cell memory. *Nature Reviews Immunology* **16**, 102–111 (2016).
97. Hamilton, S. E., Tvinnereim, A. R. & Harty, J. T. Listeria monocytogenes infection overcomes the requirement for CD40 ligand in exogenous antigen presentation to CD8+ T Cells1. *The Journal of Immunology* **167**, 5603–5609 (2001).
98. Marzo, A. L. *et al.* Fully functional memory CD8 T cells in the absence of CD4 T Cells1. *The Journal of Immunology* **173**, 969–975 (2004).
99. Takeuchi, O. & Akira, S. Pattern recognition receptors and inflammation. *Cell* **140**, 805–820 (2010).
100. Poltorak, A. *et al.* Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. *Science (New York, N.Y.)* **282**, 2085–2088 (1998).
101. Ahrens, S. *et al.* F-actin is an evolutionarily conserved damage-associated molecular pattern recognized by DNGR-1, a receptor for dead cells. *Immunity* **36**, 635–645 (2012).
102. Sampaio, N. G. *et al.* The RNA sensor MDA5 detects SARS-CoV-2 infection. *Scientific Reports* **11**, 13638 (2021).
103. Raulet, D. H. Missing self recognition and self tolerance of natural killer (NK) cells. *Seminars in Immunology* **18**, 145–150 (2006).
104. Braud, V. M. *et al.* HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature* **391**, 795–799 (1998).

105. Karlhofer, F. M., Ribaudo, R. K. & Yokoyama, W. M. MHC class I alloantigen specificity of Ly-49+ IL-2-activated natural killer cells. *Nature* **358**, 66–70 (1992).
106. Paul, W. E. Bridging innate and adaptive immunity. *Cell* **147**, 1212–1215 (2011).
107. Collin, M. & Bigley, V. Human dendritic cell subsets: An update. *Immunology* **154**, 3–20 (2018).
108. Cabeza-Cabrerizo, M., Cardoso, A., Minutti, C. M., Costa, M. P. da & Sousa, C. R. e. Dendritic Cells Revisited. *Annual Review of Immunology* **39**, 131–166 (2021).
109. Bošnjak, B., Do, K. T. H., Förster, R. & Hammerschmidt, S. I. Imaging dendritic cell functions. *Immunological Reviews* **306**, 137–163 (2022).
110. Jahnsen, F. L. *et al.* Accelerated antigen sampling and transport by airway mucosal dendritic cells following inhalation of a bacterial Stimulus1. *The Journal of Immunology* **177**, 5861–5867 (2006).
111. Farache, J. *et al.* Luminal bacteria recruit CD103+ dendritic cells into the intestinal epithelium to sample bacterial antigens for presentation. *Immunity* **38**, 581–595 (2013).
112. Yatim, K. M., Gosto, M., Humar, R., Williams, A. L. & Oberbarnscheidt, M. H. Renal dendritic cells sample blood-borne antigen and guide T-cell migration to the kidney by means of intravascular processes. *Kidney International* **90**, 818–827 (2016).
113. Inaba, K. *et al.* The formation of immunogenic major histocompatibility complex class II–peptide ligands in lysosomal compartments of dendritic cells is regulated by inflammatory stimuli. *The Journal of Experimental Medicine* **191**, 927–936 (2000).
114. Udeshi, N. D. *et al.* Cell-surface milieu remodeling in human dendritic cell activation. *Journal of immunology (Baltimore, Md. : 1950)* **213**, 1023–1032 (2024).
115. Liang, F. *et al.* Efficient targeting and activation of antigen-presenting cells in vivo after modified mRNA vaccine administration in rhesus macaques. *Molecular Therapy* **25**, 2635–2647 (2017).
116. Ohl, L. *et al.* CCR7 governs skin dendritic cell migration under inflammatory and steady-state conditions. *Immunity* **21**, 279–288 (2004).
117. West, M. A. *et al.* Enhanced dendritic cell antigen capture via toll-like receptor-induced actin remodeling. *Science (New York, N.Y.)* **305**, 1153–1157 (2004).
118. Jongbloed, S. L. *et al.* Human CD141+ (BDCA-3)+ dendritic cells (DCs) represent a unique myeloid DC subset that cross-presents necrotic cell antigens. *The Journal of Experimental Medicine* **207**, 1247–1260 (2010).
119. Kashem, S. W. *et al.* Candida albicans morphology and dendritic cell subsets determine T helper cell differentiation. *Immunity* **42**, 356–366 (2015).
120. Krishnaswamy, J. K. *et al.* Migratory CD11b+ conventional dendritic cells induce T follicular helper cell–dependent antibody responses. *Science immunology* **2**, eaam9169 (2017).
121. Shah, K., Al-Haidari, A., Sun, J. & Kazi, J. U. T cell receptor (TCR) signaling in health and disease. *Signal Transduction and Targeted Therapy* **6**, 1–26 (2021).
122. Blanchard, L. & Girard, J.-P. High endothelial venules (HEVs) in immunity, inflammation and cancer. *Angiogenesis* **24**, 719–753 (2021).
123. Castellino, F. *et al.* Chemokines enhance immunity by guiding naive CD8+ T cells to sites of CD4+ T cell-dendritic cell interaction. *Nature* **440**, 890–895 (2006).

124. Masopust, D. & Schenkel, J. M. The integration of T cell migration, differentiation and function. *Nature Reviews Immunology* **13**, 309–320 (2013).
125. Farber, D. L., Yudanin, N. A. & Restifo, N. P. Human memory T cells: generation, compartmentalization and homeostasis. *Nature Reviews Immunology* **14**, 24–35 (2014).
126. Schenkel, J. M., Fraser, K. A., Vezys, V. & Masopust, D. Sensing and alarm function of resident memory CD8+ T cells. *Nature Immunology* **14**, 509–513 (2013).
127. Ariotti, S. *et al.* Skin-resident memory CD8+ T cells trigger a state of tissue-wide pathogen alert. *Science* **346**, 101–105 (2014).
128. Andrian, U. H. von & Mempel, T. R. Homing and cellular traffic in lymph nodes. *Nature Reviews Immunology* **3**, 867–878 (2003).
129. Kaech, S. M. & Ahmed, R. Memory CD8+ T cell differentiation: initial antigen encounter triggers a developmental program in naïve cells. *Nature Immunology* **2**, 415–422 (2001).
130. Mempel, T. R., Henrickson, S. E. & Von Andrian, U. H. T-cell priming by dendritic cells in lymph nodes occurs in three distinct phases. *Nature* **427**, 154–159 (2004).
131. Rowell, E. & Wilson, C. B. Programming perpetual T helper cell plasticity. *Immunity* **30**, 7–9 (2009).
132. Pearce, E. L. *et al.* Enhancing CD8 T-cell memory by modulating fatty acid metabolism. *Nature* **460**, 103–107 (2009).
133. Szabo, P. A. *et al.* Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nature Communications* **10**, 4706 (2019).
134. Joshi, N. S. *et al.* Inflammation directs memory precursor and short-lived effector CD8+ T cell fates via the graded expression of T-bet transcription factor. *Immunity* **27**, 281–295 (2007).
135. Pham, T. H. M., Okada, T., Matloubian, M., Lo, C. G. & Cyster, J. G. S1P1 receptor signaling overrides retention mediated by G alpha i-coupled receptors to promote T cell egress. *Immunity* **28**, 122–133 (2008).
136. Crotty, S. T follicular helper cell biology: A decade of discovery and diseases. *Immunity* **50**, 1132–1148 (2019).
137. Mikhak, Z., Strassner, J. P. & Luster, A. D. Lung dendritic cells imprint T cell lung homing and promote lung immunity through the chemokine receptor CCR4. *Journal of Experimental Medicine* **210**, 1855–1869 (2013).
138. Nourshargh, S. & Alon, R. Leukocyte migration into inflamed tissues. *Immunity* **41**, 694–707 (2014).
139. Wherry, E. J. *et al.* Lineage relationship and protective immunity of memory CD8 T cell subsets. *Nature Immunology* **4**, 225–234 (2003).
140. Buggert, M., Price, D. A., Mackay, L. K. & Betts, M. R. Human circulating and tissue-resident memory CD8+ T cells. *Nature Immunology* **24**, 1076–1086 (2023).
141. Goronzy, J. J. & Weyand, C. M. Understanding immunosenescence to improve responses to vaccines. *Nature Immunology* **14**, 428–436 (2013).
142. Hamann, D. *et al.* Phenotypic and functional separation of memory and effector human CD8+ T cells. *The Journal of Experimental Medicine* **186**, 1407–1418 (1997).

143. Sallusto, F., Lenig, D., Förster, R., Lipp, M. & Lanzavecchia, A. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature* **401**, 708–712 (1999).
144. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine* **382**, 727–733 (2020).
145. COVID-19 cases | WHO COVID-19 dashboard.

146. Sahin, U. *et al.* COVID-19 vaccine BNT162b1 elicits human antibody and TH1 T cell responses. *Nature* **586**, 594–599 (2020).
147. Peeling, R. W., Heymann, D. L., Teo, Y.-Y. & Garcia, P. J. Diagnostics for COVID-19: Moving from pandemic response to control. *The Lancet* **399**, 757–768 (2022).
148. 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).
149. Swadling, L. *et al.* Pre-existing polymerase-specific T cells expand in abortive seronegative SARS-CoV-2. *Nature* **601**, 110–117 (2022).
150. Diniz, M. O. *et al.* Airway-resident T cells from unexposed individuals cross-recognize SARS-CoV-2. *Nature Immunology* **23**, 1324–1329 (2022).
151. Augusto, D. *et al.* A common allele of HLA is associated with asymptomatic SARS-CoV-2 infection. *Nature* (2023) doi:10.1038/S41586-023-06331-X.
152. Coleman, C. M. & Frieman, M. B. Coronaviruses: Important emerging human pathogens. *Journal of Virology* **88**, 5209–5212 (2014).
153. Steiner, S. *et al.* SARS-CoV-2 biology and host interactions. *Nature Reviews Microbiology* **22**, 206–225 (2024).
154. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020.
155. Wölfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature* **581**, 465–469 (2020).
156. Singanayagam, A. *et al.* Community transmission and viral load kinetics of the SARS-CoV-2 delta (B.1.617.2) variant in vaccinated and unvaccinated individuals in the UK: A prospective, longitudinal, cohort study. *The Lancet Infectious Diseases* **22**, 183–195 (2022).
157. Karagiannidis, C. *et al.* Case characteristics, resource use, and outcomes of 10 021 patients with COVID-19 admitted to 920 german hospitals: An observational study. *The Lancet Respiratory Medicine* **8**, 853–862 (2020).
158. Emanuel, E. J. *et al.* Fair allocation of scarce medical resources in the time of covid-19. *New England Journal of Medicine* **382**, 2049–2055 (2020).
159. *COVID-19 rapid guideline: critical care in adults*. (National Institute for Health; Care Excellence (NICE), London, 2021).
160. Health at a Glance 2021. (2021).

161. Wu, Z. & McGoogan, J. M. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in china: Summary of a report of 72 314 cases from the chinese center for disease control and prevention. *JAMA* **323**, 1239–1242 (2020).

162. Aquino, Y. *et al.* Dissecting human population variation in single-cell responses to SARS-CoV-2. *Nature* **621**, 120–128 (2023).
163. O'Driscoll, M. *et al.* Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590**, 140–145 (2021).
164. Xie, J. *et al.* Relationship between HLA genetic variations, COVID-19 vaccine antibody response, and risk of breakthrough outcomes. *Nature Communications* **15**, 4031 (2024).
165. Russell, C. D., Lone, N. I. & Baillie, J. K. Comorbidities, multimorbidity and COVID-19. *Nature Medicine* **29**, 334–343 (2023).
166. Pairo-Castineira, E. *et al.* GWAS and meta-analysis identifies 49 genetic variants underlying critical COVID-19. *Nature* **617**, 764–768 (2023).
167. Ghosh, S. *et al.* β-coronaviruses use lysosomes for egress instead of the biosynthetic secretory pathway. *Cell* **183**, 1520–1535.e14 (2020).
168. Park, A. & Iwasaki, A. Type I and type III interferons – induction, signaling, evasion, and application to combat COVID-19. *Cell Host & Microbe* **27**, 870–878 (2020).
169. Cortese, M. *et al.* Integrative imaging reveals SARS-CoV-2-induced reshaping of subcellular morphologies. *Cell Host & Microbe* **28**, 853–866.e5 (2020).
170. Decroly, E., Ferron, F., Lescar, J. & Canard, B. Conventional and unconventional mechanisms for capping viral mRNA. *Nature Reviews Microbiology* **10**, 51–65 (2012).
171. Arshad, N. *et al.* SARS-CoV-2 accessory proteins ORF7a and ORF3a use distinct mechanisms to down-regulate MHC-I surface expression. *Proceedings of the National Academy of Sciences* **120**, e2208525120 (2023).
172. Zhang, Y. *et al.* The ORF8 protein of SARS-CoV-2 mediates immune evasion through down-regulating MHC-I. *Proceedings of the National Academy of Sciences* **118**, e2024202118 (2021).
173. Hou, Y. J. *et al.* SARS-CoV-2 reverse genetics reveals a variable infection gradient in the respiratory tract. *Cell* **182**, 429–446.e14 (2020).
174. Khan, M. *et al.* Visualizing in deceased COVID-19 patients how SARS-CoV-2 attacks the respiratory and olfactory mucosae but spares the olfactory bulb. *Cell* **184**, 5932–5949.e15 (2021).
175. Ahn, J. H. *et al.* Nasal ciliated cells are primary targets for SARS-CoV-2 replication in the early stage of COVID-19. *The Journal of Clinical Investigation* **131**, (2021).
176. Hatton, C. F. *et al.* Delayed induction of type I and III interferons mediates nasal epithelial cell permissiveness to SARS-CoV-2. *Nature Communications* **12**, 7092 (2021).
177. Blanco-Melo, D. *et al.* Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* **181**, 1036–1045.e9 (2020).
178. Israelow, B. & Iwasaki, A. First encounter with SARS-CoV-2: immune portraits of COVID susceptibility. *Nature* **631**, 33–35 (2024).
179. 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).

180. Tan, A. T. *et al.* Early induction of functional SARS-CoV-2-specific T cells associates with rapid viral clearance and mild disease in COVID-19 patients. *Cell Reports* **34**, 108728 (2021).
181. Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nature Medicine* **26**, 842–844 (2020).
182. 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).
183. Hadjadj, J. *et al.* Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science* **369**, 718–724 (2020).
184. Aid, M. *et al.* Vascular disease and thrombosis in SARS-CoV-2-infected rhesus macaques. *Cell* **183**, 1354–1366.e13 (2020).
185. Radermecker, C. *et al.* Neutrophil extracellular traps infiltrate the lung airway, interstitial, and vascular compartments in severe COVID-19. *Journal of Experimental Medicine* **217**, (2020).
186. Lucas, C. *et al.* Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* **584**, 463–469 (2020).
187. Liu, C. *et al.* Time-resolved systems immunology reveals a late juncture linked to fatal COVID-19. *Cell* **184**, 1836–1857.e22 (2021).
188. Ren, X. *et al.* COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184**, 1895–1913.e19 (2021).
189. Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine* **27**, 904–916 (2021).
190. Yoshida, M. *et al.* Local and systemic responses to SARS-CoV-2 infection in children and adults. *Nature* **602**, 321–327 (2022).
191. Swann, O. V. *et al.* Clinical characteristics of children and young people admitted to hospital with covid-19 in United Kingdom: prospective multicentre observational cohort study. *BMJ (Clinical research ed.)* **370**, m3249 (2020).
192. Starke, K. R. *et al.* The isolated effect of age on the risk of COVID-19 severe outcomes: a systematic review with meta-analysis. *BMJ Global Health* **6**, (2021).
193. Imbiakha, B. *et al.* Adaptive immune cells are necessary for SARS-CoV-2–induced pathology. *Science Advances* **10**, eadg5461 (2024).
194. Grant, R. A. *et al.* Circuits between infected macrophages and T cells in SARS-CoV-2 pneumonia. *Nature* **590**, 635–641 (2021).
195. Georg, P. *et al.* Complement activation induces excessive T cell cytotoxicity in severe COVID-19. *Cell* **185**, 493–512.e25 (2022).
196. Olea, B. *et al.* SARS-CoV-2-reactive IFN-γ-producing CD4+ and CD8+ T cells in blood do not correlate with clinical severity in unvaccinated critically ill COVID-19 patients. (2022) doi:10.1038/S41598-022-18659-X.
197. Moss, P. Protective SARS-CoV-2-specific T cells take up residence. *Nature Immunology* **26**, 331–332 (2025).
198. Markov, N. S. *et al.* Distinctive evolution of alveolar T cell responses is associated with clinical outcomes in unvaccinated patients with SARS-CoV-2 pneumonia. *Nature Immunology* **25**, 1607–1622 (2024).

199. Zhu, A. *et al.* Robust mucosal SARS-CoV-2-specific T cells effectively combat COVID-19 and establish polyfunctional resident memory in patient lungs. *Nature Immunology* **26**, 459–472 (2025).
200. Laidlaw, B. J. *et al.* Production of IL-10 by CD4+ regulatory T cells during the resolution of infection promotes the maturation of memory CD8+ T cells. *Nature Immunology* **16**, 871–879 (2015).
201. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2024 update. *Nucleic Acids Research* **53**, D436–D443 (2025).
202. Sette, A., Sidney, J. & Crotty, S. T Cell Responses to SARS-CoV-2. *Annual Review of Immunology* **41**, 343–373 (2023).
203. Grifoni, A. *et al.* SARS-CoV-2 human T cell epitopes: Adaptive immune response against COVID-19. *Cell Host & Microbe* (2021) doi:10.1016/J.CHOM.2021.05.010.
204. Loyal, L. *et al.* Cross-reactive CD4+ T cells enhance SARS-CoV-2 immune responses upon infection and vaccination. *Science* **374**, eabh1823 (2021).
205. Karp-Tatham, E., O'Neill, C. R., Knight, J. C., Mentzer, A. J. & Chong, A. Y. Lack of association between classical HLA genes and asymptomatic SARS-CoV-2 infection. *Human Genetics and Genomics Advances* **6**, (2025).
206. Marchal, A. *et al.* Lack of association between classical HLA genes and asymptomatic SARS-CoV-2 infection. *Human Genetics and Genomics Advances* **5**, (2024).
207. Marchal, A. *et al.* Response to Karp-Tatham et al. *Human Genetics and Genomics Advances* **6**, (2025).
208. Stipdonk, M. J. B. van, Lemmens, E. E. & Schoenberger, S. P. Naïve CTLs require a single brief period of antigenic stimulation for clonal expansion and differentiation. *Nature Immunology* **2**, 423–429 (2001).
209. Kinjyo, I. *et al.* Real-time tracking of cell cycle progression during CD8+ effector and memory T-cell differentiation. *Nature Communications* **6**, 6301 (2015).
210. Lawrence, C. W. & Braciale, T. J. Activation, differentiation, and migration of naive virus-specific CD8+ T cells during pulmonary influenza virus Infection1. *The Journal of Immunology* **173**, 1209–1218 (2004).
211. Badovinac, V. P., Porter, B. B. & Harty, J. T. Programmed contraction of CD8+ T cells after infection. *Nature Immunology* **3**, 619–626 (2002).
212. Harbertson, J., Biederman, E., Bennett, K. E., Kondrack, R. M. & Bradley, L. M. Withdrawal of stimulation may initiate the transition of effector to memory CD4 Cells1. *The Journal of Immunology* **168**, 1095–1102 (2002).
213. Badovinac, V. P., Porter, B. B. & Harty, J. T. CD8+ T cell contraction is controlled by early inflammation. *Nature Immunology* **5**, 809–817 (2004).
214. Hammarlund, E. *et al.* Duration of antiviral immunity after smallpox vaccination. *Nature Medicine* **9**, 1131–1137 (2003).
215. Lin, J. *et al.* Longitudinal assessment of SARS-CoV-2-specific T cell cytokine-producing responses for 1 year reveals persistence of multicytokine proliferative responses, with greater immunity associated with disease severity. *Journal of Virology* **96**, e00509–22.

216. Kristensen, N. P. *et al.* Simultaneous analysis of pMHC binding and reactivity unveils virus-specific CD8 T cell immunity to a concise epitope set. *Science Advances* **10**, eadm8951 (2024).
217. Sims, S. & Klenerman, P. and. MHC–peptide tetramers for the analysis of antigen-specific T cells. *Expert Review of Vaccines* **9**, 765–774 (2010).
218. Choo, J. A. L., Liu, J., Toh, X., Grotenbreg, G. M. & Ren, E. C. The immunodominant influenza a virus M158–66 cytotoxic T lymphocyte epitope exhibits degenerate class I major histocompatibility complex restriction in humans. *Journal of Virology* **88**, 10613–10623 (2014).
219. Lagattuta, K. A. *et al.* The T cell receptor sequence influences the likelihood of T cell memory formation. *Cell Reports* **44**, 115098 (2025).
220. Montemurro, A., Povlsen, H. R., Jessen, L. E. & Nielsen, M. Benchmarking data-driven filtering for denoising of TCRpMHC single-cell data. *Scientific Reports* **13**, 16147 (2023).
221. Whiteside, S. K., Snook, J. P., Williams, M. A. & Weis, J. J. Bystander T cells: A balancing act of friends and foes. *Trends in Immunology* **39**, 1021–1035 (2018).
222. Whiteside, S. K. *et al.* IL-10 Deficiency Reveals a Role for TLR2-Dependent Bystander Activation of T Cells in Lyme Arthritis. *The Journal of Immunology* **200**, 1457–1470 (2018).
223. Beadling, C. & Slifka, M. K. Differential regulation of virus-specific T-cell effector functions following activation by peptide or innate cytokines. *Blood* **105**, 1179–1186 (2005).
224. Chu, T. *et al.* Bystander-activated memory CD8 T cells control early pathogen load in an innate-like, NKG2D-dependent manner. *Cell Reports* **3**, 701–708 (2013).
225. 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).
226. Liu, K., Catalfamo, M., Li, Y., Henkart, P. A. & Weng, N. IL-15 mimics T cell receptor crosslinking in the induction of cellular proliferation, gene expression, and cytotoxicity in CD8+ memory T cells. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6192–6197 (2002).
227. Correia, M. P., Costa, A. V., Uhrberg, M., Cardoso, E. M. & Arosa, F. A. IL-15 induces CD8+ T cells to acquire functional NK receptors capable of modulating cytotoxicity and cytokine secretion. *Immunobiology* **216**, 604–612 (2011).
228. Kawabe, T. *et al.* Memory-phenotype CD4+ T cells spontaneously generated under steady-state conditions exert innate TH1-like effector function. *Science Immunology* **2**, eaam9304 (2017).
229. Doisne, J.-M. *et al.* CD8+ T cells specific for EBV, cytomegalovirus, and influenza virus are activated during primary HIV Infection1. *The Journal of Immunology* **173**, 2410–2418 (2004).
230. Zhang, B. *et al.* Multimodal single-cell datasets characterize antigen-specific CD8+ T cells across SARS-CoV-2 vaccination and infection. *Nature Immunology* **24**, 1725–1734 (2023).

231. Freeman, B. E., Hammarlund, E., Raué, H.-P. & Slifka, M. K. Regulation of innate CD8+ T-cell activation mediated by cytokines. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 9971–9976 (2012).
232. Shiroishi, M. *et al.* Human inhibitory receptors ig-like transcript 2 (ILT2) and ILT4 compete with CD8 for MHC class I binding and bind preferentially to HLA-g. *Proceedings of the National Academy of Sciences* **100**, 8856–8861 (2003).
233. Colonna, M. *et al.* A common inhibitory receptor for major histocompatibility complex class I molecules on human lymphoid and myelomonocytic cells. *Journal of Experimental Medicine* **186**, 1809–1818 (1997).
234. Borges, L., Hsu, M. L., Fanger, N., Kubin, M. & Cosman, D. A family of human lymphoid and myeloid ig-like receptors, some of which bind to MHC class I molecules. *The Journal of Immunology* **159**, 5192–5196 (1997).
235. Hansasuta, P. *et al.* Recognition of HLA-A3 and HLA-A11 by KIR3DL2 is peptide-specific. *European Journal of Immunology* **34**, 1673–1679 (2004).
236. Stewart, C. A. *et al.* Recognition of peptide–MHC class I complexes by activating killer immunoglobulin-like receptors. *Proceedings of the National Academy of Sciences* **102**, 13224–13229 (2005).
237. Björkström, N. K. *et al.* CD8 T cells express randomly selected KIRs with distinct specificities compared with NK cells. *Blood* **120**, 3455–3465 (2012).
238. Povlsen, H. R. *et al.* Improved T cell receptor antigen pairing through data-driven filtering of sequencing information from single cells. *eLife* **12**, e81810 (2023).
239. Corr, M. *et al.* T cell receptor-MHC class I peptide interactions: Affinity, kinetics, and specificity. *Science* **265**, 946–949 (1994).
240. Altman, J. D. *et al.* Phenotypic analysis of antigen-specific T lymphocytes. *Science* **274**, 94–96 (1996).
241. Huang, J. *et al.* Detection, phenotyping, and quantification of antigen-specific T cells using a peptide-MHC dodecamer. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E1890–1897 (2016).
242. Mallajosyula, V. *et al.* CD8+ T cells specific for conserved coronavirus epitopes correlate with milder disease in patients with COVID-19. *Science Immunology* **6**, eabg5669.
243. Laugel, B. *et al.* Different T cell receptor affinity thresholds and CD8 coreceptor dependence govern cytotoxic T lymphocyte activation and tetramer binding properties*. *Journal of Biological Chemistry* **282**, 23799–23810 (2007).
244. Dileepan, T. *et al.* MHC class II tetramers engineered for enhanced binding to CD4 improve detection of antigen-specific T cells. *Nature Biotechnology* **39**, 943–948 (2021).
245. Chandran, S. S. *et al.* Immunogenicity and therapeutic targeting of a public neoantigen derived from mutated PIK3CA. *Nature Medicine* **28**, 946–957 (2022).
246. Parkhurst, M. *et al.* Isolation of T-cell receptors specifically reactive with mutated tumor-associated antigens from tumor-infiltrating lymphocytes based on CD137 expression. *Clinical Cancer Research* **23**, 2491–2505 (2017).
247. Dolton, G. *et al.* More tricks with tetramers: a practical guide to staining T cells with peptide–MHC multimers. *Immunology* **146**, 11–22 (2015).

248. Newell, E. W. *et al.* Combinatorial tetramer staining and mass cytometry analysis facilitate T-cell epitope mapping and characterization. *Nature Biotechnology* **31**, 623–629 (2013).
249. Hadrup, S. R. *et al.* Parallel detection of antigen-specific T-cell responses by multi-dimensional encoding of MHC multimers. *Nature Methods* **6**, 520–526 (2009).
250. Bentzen, A. K. *et al.* Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nature Biotechnology* **34**, 1037–1045 (2016).
251. Zhang, W. *et al.* A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity. *Science Advances* **7**, eabf5835 (2021).
252. Goncharov, M. *et al.* VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. *Nature Methods* **19**, 1017–1019 (2022).
253. Minervina, A. A. *et al.* SARS-CoV-2 antigen exposure history shapes phenotypes and specificity of memory CD8+ T cells. *Nature Immunology* **23**, 781–790 (2022).
254. Birnbaum, M. E. *et al.* Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* **157**, 1073–1087 (2014).
255. Saini, S. K. *et al.* SARS-CoV-2 genome-wide T cell epitope mapping reveals immunodominance and substantial CD8+ T cell activation in COVID-19 patients. *Science Immunology* **6**, eabf7550 (2021).
256. Law, J. C. & Watts, T. H. Considerations for choosing T cell assays during a pandemic. *The Journal of Immunology* **211**, 169–174 (2023).
257. Le Bert, N. *et al.* SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature* **584**, 457–462 (2020).
258. Milighetti, M. *et al.* Large clones of pre-existing T cells drive early immunity against SARS-COV-2 and LCMV infection. *iScience* **26**, 106937 (2023).
259. Lodolce, J. P., Burkett, P. R., Boone, D. L., Chien, M. & Ma, A. T cell–independent interleukin 15rα signals are required for bystander proliferation. *The Journal of Experimental Medicine* **194**, 1187–1194 (2001).
260. Czerkinsky, C. C., Nilsson, L.-Å., Nygren, H., Ouchterlony, Ö. & Tarkowski, A. A solid-phase enzyme-linked immunospot (ELISPOT) assay for enumeration of specific antibody-secreting cells. *Journal of Immunological Methods* **65**, 109–121 (1983).
261. Lehmann, P. V. & Zhang, W. Unique strengths of ELISPOT for T cell diagnostics. *Methods in Molecular Biology (Clifton, N.J.)* **792**, 3–23 (2012).
262. Binayke, A. *et al.* A quest for universal anti-SARS-CoV-2 T cell assay: systematic review, meta-analysis, and experimental validation. *npj Vaccines* **9**, 1–17 (2024).
263. Boyd, A. *et al.* Pathogen-Specific T Cell Polyfunctionality Is a Correlate of T Cell Efficacy and Immune Protection. *PLOS ONE* **10**, e0128714 (2015).
264. Tian, J. *et al.* T cell immune evasion by SARS-CoV-2 JN.1 escapees targeting two cytotoxic T cell epitope hotspots. *Nature Immunology* **26**, 265–278 (2025).
265. Silva, T. I. de *et al.* Correlates of T-cell–mediated viral control and phenotype of CD8+ T cells in HIV-2, a naturally contained human retroviral infection. *Blood* **121**, 4330–4339 (2013).

266. Zuo, J. *et al.* Robust SARS-CoV-2-specific T-cell immunity is maintained at 6 months following primary infection. *Nature immunology* **22**, 620–626 (2021).
267. Zaunders, J. J. *et al.* High levels of human antigen-specific CD4+ T cells in peripheral blood revealed by stimulated coexpression of CD25 and CD134 (OX40) 1. *The Journal of Immunology* **183**, 2827–2836 (2009).
268. Sekine, T. *et al.* Robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19. *Cell* (2020) doi:10.1016/J.CELL.2020.08.017.
269. Juno, J. A. *et al.* Humoral and circulating follicular helper T cell responses in recovered patients with COVID-19. *Nature Medicine* **26**, 1428–1434 (2020).
270. Dan, J. M. *et al.* Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science* **371**, eabf4063 (2021).
271. Zheng, M. Z. M. *et al.* Deconvoluting TCR-dependent & -independent activation is vital for reliable ag-specific $CD4^+$ T cell characterization by AIM assay. (2024).
272. Cai, C. *et al.* SARS-CoV-2 vaccination enhances the effector qualities of spike-specific T cells induced by COVID-19. *Science Immunology* **8**, eadh0687 (2023).
273. Mudd, P. A. *et al.* SARS-CoV-2 mRNA vaccination elicits a robust and persistent T follicular helper cell response in humans. *Cell* **185**, 603–613.e15 (2022).
274. Wragg, K. M. *et al.* Establishment and recall of SARS-CoV-2 spike epitope-specific CD4+ T cell memory. (2022) doi:10.1038/S41590-022-01175-5.
275. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
276. Yao, Y., Liu, N., Zhou, Z. & Shi, L. Influence of ERAP1 and ERAP2 gene polymorphisms on disease susceptibility in different populations. *Human Immunology* **80**, 325–334 (2019).
277. Hudson, D., Fernandes, R. A., Basham, M., Ogg, G. & Koohy, H. Can we predict T cell specificity with digital biology and machine learning? *Nature Reviews Immunology* **23**, 511–521 (2023).
278. Ford, E. S. *et al.* CD8+ T cell clonotypes from prior SARS-CoV-2 infection predominate during the cellular immune response to mRNA vaccination. *Research Square* rs.3.rs–2146712 (2022) doi:10.21203/rs.3.rs-2146712/v1.
279. Mayer-Blackwell, K. *et al.* TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *eLife* **10**, e68605 (2021).
280. Dash, P. *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).
281. Jouannet, C., Vantomme, H., Gouge, K. L., Klatzmann, D. & Mariotti-Ferrandiz, E. Benchmarking unsupervised methods for inferring TCR specificity. *NAR Genomics and Bioinformatics* **7**, lqaf150 (2025).
282. Chotisorayuth, T. & Tiffeau-Mayer, A. Lightning-fast adaptive immune receptor similarity search by symmetric deletion lookup. doi:10.48550/arXiv.2403.09010.
283. Zhang, H., Zhan, X. & Li, B. GIANA allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation. *Nature Communications* **12**, (2021).

284. Jiang, Y., Huo, M., Zhang, P., Zou, Y. & Li, S. C. TCR2vec: a deep representation learning framework of T-cell receptor sequence and function. doi:10.1101/2023.03.31.535142.
285. Nagano, Y. *et al.* Contrastive learning of T cell receptor representations. *Cell Systems* **16**, (2025).
286. Fast, E., Dhar, M. & Chen, B. TAPIR: a T-cell receptor language model for predicting rare and novel targets. doi:10.1101/2023.09.12.557285.
287. Karthikeyan, D., Raffel, C., Vincent, B. & Rubinsteyn, A. *TCR-TRANSLATE: Conditional Generation of Real Antigen Specific T-Cell Receptor Sequences*. (2024).
288. Nielsen, M. *et al.* Lessons learned from the IMMREP23 TCR-epitope prediction challenge. *ImmunoInformatics* **16**, 100045 (2024).
289. Drost, F. *et al.* Benchmarking of T-Cell Receptor - Epitope Predictors with ePytope-TCR. doi:10.1101/2024.11.06.622261.
290. Jensen, M. F. & Nielsen, M. Enhancing TCR specificity predictions by combined pan- and peptide-specific training, loss-scaling, and sequence similarity integration. *eLife* **12**, RP93934 (2024).
291. Mason, D. M. & Reddy, S. T. Predicting adaptive immune receptor specificities by machine learning is a data generation problem. *Cell Systems* **15**, 1190–1197 (2024).
292. Dens, C., Laukens, K., Bittremieux, W. & Meysman, P. The pitfalls of negative data bias for the T-cell epitope specificity challenge. *Nature Machine Intelligence* **5**, 1060–1062 (2023).
293. Meynard-Piganeau, B., Feinauer, C., Weigt, M., Walczak, A. M. & Mora, T. TULIP — a Transformer based Unsupervised Language model for Interacting Peptides and T-cell receptors that generalizes to unseen epitopes. doi:10.1101/2023.07.19.549669.
294. Ullanat, V., Jing, B., Sledzieski, S. & Berger, B. Learning the language of protein-protein interactions. doi:10.1101/2025.03.09.642188.
295. Ewer, K. J. *et al.* Protective CD8+ T-cell immunity to human malaria induced by chimpanzee adenovirus-MVA immunisation. *Nature Communications* **4**, 2836 (2013).
296. Brasu, N. *et al.* Memory CD8+ T cell diversity and B cell responses correlate with protection against SARS-CoV-2 following mRNA vaccination. *Nature Immunology* **23**, 1445–1456 (2022).
297. Haynes, B. F. *et al.* Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *New England Journal of Medicine* **366**, 1275–1286 (2012).
298. Ishizuka, A. S. *et al.* Protection against malaria at 1 year and immune correlates following PfSPZ vaccination. *Nature medicine* **22**, 614–623 (2016).
299. Arunachalam, P. S. *et al.* Systems vaccinology of the BNT162b2 mRNA vaccine in humans. *Nature* **596**, 410–416 (2021).
300. Schulien, I. *et al.* Characterization of pre-existing and induced SARS-CoV-2-specific CD8+ T cells. *Nature Medicine* **27**, 78–85 (2021).
301. Francis, J. M. *et al.* Allelic variation in class I HLA determines CD8+ T cell repertoire shape and cross-reactive memory responses to SARS-CoV-2. *Science Immunology* **7**, eabk3070 (2022).

302. Schattgen, S. A. *et al.* Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). *Nature Biotechnology* **40**, 54–63 (2022).
303. Drost, F. *et al.* Multi-modal generative modeling for joint analysis of single-cell T cell receptor and gene expression data. *Nature Communications* **15**, (2024).
304. Sandt, C. E. van de *et al.* Newborn and child-like molecular signatures in older adults stem from TCR shifts across human lifespan. *Nature Immunology* **24**, 1890–1907 (2023).
305. Galperin, M. *et al.* CD4+ T cell–mediated HLA class II cross-restriction in HIV controllers. *Science Immunology* **3**, eaat0687 (2018).
306. Wilson, D. S. *et al.* Synthetically glycosylated antigens induce antigen-specific tolerance and prevent the onset of diabetes. *Nature Biomedical Engineering* **3**, 817–829 (2019).
307. Tremain, A. C. *et al.* Synthetically glycosylated antigens for the antigen-specific suppression of established immune responses. *Nature Biomedical Engineering* **7**, 1142–1155 (2023).
308. Murray, J. A. *et al.* Safety and tolerability of KAN-101, a liver-targeted immune tolerance therapy, in patients with coeliac disease (ACeD): A phase 1 trial. *The Lancet Gastroenterology & Hepatology* **8**, 735–747 (2023).
309. Anokion Announces Positive Symptom Data from its Phase 2 Trial Evaluating KAN-101 for the Treatment of Celiac Disease – Anokion. https://anokion.com/press_releases/anokion-announces-positive-symptom-data-from-its-phase-2-trial-evaluating-kan-101-for-the-treatment-of-celiac-disease/.
310. Anokion SA. *A Phase 1 Study of the Safety and Tolerability of Single and Multiple Doses of ANK-700 in Patients With Relapsing Remitting Multiple Sclerosis*. https://clinicaltrials.gov/study/NCT04602390 (2024).
311. Keck, S. *et al.* Antigen affinity and antigen dose exert distinct influences on CD4 T-cell differentiation. *Proceedings of the National Academy of Sciences* **111**, 14852–14857 (2014).
312. Hosken, N. A., Shibuya, K., Heath, A. W., Murphy, K. M. & O'Garra, A. The effect of antigen dose on CD4+ T helper cell phenotype development in a T cell receptor-alpha beta-transgenic model. *Journal of Experimental Medicine* **182**, 1579–1584 (1995).
313. Künzli, M., Reuther, P., Pinschewer, D. D. & King, C. G. Opposing effects of T cell receptor signal strength on CD4 T cells responding to acute versus chronic viral infection. *eLife* **10**, e61869 (2021).
314. Kotov, D. I. *et al.* TCR affinity biases th cell differentiation by regulating CD25, Eef1e1, and Gbp2. *The Journal of Immunology* **202**, 2535–2545 (2019).
315. Vanguri, V., Govern, C. C., Smith, R. & Huseby, E. S. Viral antigen density and confinement time regulate the reactivity pattern of CD4 T-cell responses to vaccinia virus infection. *Proceedings of the National Academy of Sciences* **110**, 288–293 (2013).
316. Snook, J. P., Kim, C. & Williams, M. A. TCR signal strength controls the differentiation of CD4+ effector and memory T cells. *Science Immunology* **3**, eaas9103 (2018).
317. Crotty, S. Hybrid immunity. *Science* **372**, 1392–1393 (2021).

318. Reynolds, C. J. *et al.* Heterologous infection and vaccination shapes immunity against SARS-CoV-2 variants. *Science* **375**, 183–192 (2022).
319. Guo, L. *et al.* Durability and cross-reactive immune memory to SARS-CoV-2 in individuals 2 years after recovery from COVID-19: A longitudinal cohort study. *The Lancet Microbe* **5**, e24–e33 (2024).
320. Gil, A., Yassai, M. B., Naumov, Y. N. & Selin, L. K. Narrowing of human influenza a virus-specific T cell receptor α and β repertoires with increasing age. *Journal of Virology* **89**, 4102–4116 (2015).
321. Marshall, J. C. *et al.* A minimal common outcome measure set for COVID-19 clinical research. *The Lancet Infectious Diseases* **20**, e192–e197 (2020).
322. Hornsby, H. *et al.* Omicron infection following vaccination enhances a broad spectrum of immune responses dependent on infection history. *Nature Communications* **14**, 5065 (2023).
323. Angyal, A. *et al.* T-cell and antibody responses to first BNT162b2 vaccine dose in previously infected and SARS-CoV-2-naive UK health-care workers: a multicentre prospective cohort study. *The Lancet. Microbe* **3**, e21–e31 (2022).
324. Abd Hamid, M. *et al.* Self-maintaining CD103+ cancer-specific T cells are highly energetic with rapid cytotoxic and effector responses. *Cancer Immunology Research* **8**, 203–216 (2020).
325. Liu, G. *et al.* Long-persisting SARS-CoV-2 spike-specific CD4+ T cells associated with mild disease and increased cytotoxicity post COVID-19. *Nature Communications* **16**, 8743 (2025).
326. Peng, Y. *et al.* Boosted influenza-specific T cell responses after H5N1 pandemic live attenuated influenza virus vaccination. *Frontiers in Immunology* **6**, (2015).
327. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10**, 1096–1098 (2013).
328. *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/ (2022).
329. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).
330. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics (Oxford, England)* **30**, 2811–2812 (2014).
331. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695 (2005).
332. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
333. Homo sapiens (ID 31257) - BioProject - NCBI.
334. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21 (2013).
335. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)* **30**, 923–930 (2014).
336. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nature Methods* **12**, 380–381 (2015).

337. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
338. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
339. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Current Protocols* **1**, e90 (2021).
340. Moss, P. The T cell immune response against SARS-CoV-2. *Nature Immunology* **23**, 186–193 (2022).
341. Kedzierska, K. & Thomas, P. G. Count on us: T cells in SARS-CoV-2 infection and vaccination. (2022) doi:10.1016/J.XCRM.2022.100562.
342. Chou, J., Thomas, P. G. & Randolph, A. G. Immunology of SARS-CoV-2 infection in children. *Nature Immunology* **23**, 177–185 (2022).
343. Tarke, A. *et al.* Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. *Cell reports medicine* (2021) doi:10.1016/J.XCRM.2021.100204.
344. Govern, C. C., Paczosa, M. K., Chakraborty, A. K. & Huseby, E. S. Fast on-rates allow short dwell time ligands to activate T cells. *Proceedings of the National Academy of Sciences* **107**, 8724–8729 (2010).
345. Constant, S., Pfeiffer, C., Woodard, A., Pasqualini, T. & Bottomly, K. Extent of T cell receptor ligation can determine the functional differentiation of naive CD4+ T cells. *Journal of Experimental Medicine* **182**, 1591–1596 (1995).
346. Tubo, N. J. *et al.* Single naive CD4+ T cells from a diverse repertoire produce different effector cell types during infection. *Cell* **153**, 785–796 (2013).
347. Ploquin, M. J.-Y., Eksmond, U. & Kassiotis, G. B cells and TCR avidity determine distinct functions of CD4+ T cells in retroviral infection. *The Journal of Immunology* **187**, 3321–3330 (2011).
348. Krishnamoorthy, V. *et al.* The IRF4 gene regulatory module functions as a read-write integrator to dynamically coordinate T helper cell fate. *Immunity* **47**, 481–497.e7 (2017).
349. DiToro, D. *et al.* Differential IL-2 expression defines developmental fates of follicular versus nonfollicular helper T cells. *Science* **361**, eaao2933 (2018).
350. Fazilleau, N., McHeyzer-Williams, L. J., Rosen, H. & McHeyzer-Williams, M. G. The function of follicular helper T cells is regulated by the strength of T cell antigen receptor binding. *Nature Immunology* **10**, 375–384 (2009).
351. Syed, A. M. *et al.* Omicron mutations enhance infectivity and reduce antibody neutralization of SARS-CoV-2 virus-like particles. *Proceedings of the National Academy of Sciences* **119**, e2200592119 (2022).
352. Zhang, Z. *et al.* Structure of SARS-CoV-2 membrane protein essential for virus assembly. *Nature Communications* **13**, 4399 (2022).
353. Nelde, A. *et al.* SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. *Nature Immunology* (2021) doi:10.1038/S41590-020-00808-X.
354. Wright, D. W. *et al.* Tracking SARS-CoV-2 mutations and variants through the COG-UK-mutation explorer. *Virus Evolution* **8**, veac023 (2022).

355. Marjanovic, S. *et al.* COVID-19 genomics UK (COG-UK) consortium: Final report. *Rand Health Quarterly* **9**, 24 (2022).
356. Southwood, S. *et al.* Several common HLA-DR types share largely overlapping peptide binding Repertoires1 2. *The Journal of Immunology* **160**, 3363–3373 (1998).
357. Racle, J. *et al.* Machine learning predictions of MHC-II specificities reveal alternative binding mode of class II epitopes. *Immunity* **56**, 1359–1375.e13 (2023).
358. Hudson, D., Lubbock, A., Basham, M. & Koohy, H. A comparison of clustering models for inference of T cell receptor antigen specificity. *ImmunoInformatics* **13**, (2024).
359. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
360. EMBL-EBI. Confidence scores in AlphaFold-Multimer | AlphaFold.

361. Sette, A. & Crotty, S. Immunological memory to SARS-CoV-2 infection and COVID-19 vaccines. *Immunological Reviews* **310**, 27–46 (2022).
362. Chen, Y. *et al.* Structural definition of HLA class II-presented SARS-CoV-2 epitopes reveals a mechanism to escape pre-existing CD4+ T cell immunity. *Cell Reports* (2023) doi:10.1016/J.CELREP.2023.112827.
363. Sinigaglia, F. *et al.* A malaria T-cell epitope recognized in association with most mouse and human MHC class II molecules. *Nature* **336**, 778–780 (1988).
364. Kaufmann, D. E. *et al.* Comprehensive analysis of human immunodeficiency virus type 1-specific CD4 responses reveals marked immunodominance of gag and nef and the presence of broadly recognized peptides. *Journal of Virology* **78**, 4463–4477 (2004).
365. Wilson, C. C. *et al.* Identification and antigenicity of broadly cross-reactive and conserved human immunodeficiency virus type 1-derived helper T-lymphocyte epitopes. *Journal of Virology* **75**, 4195–4207 (2001).
366. Rha, M.-S. *et al.* PD-1-Expressing SARS-CoV-2-Specific CD8+ T Cells Are Not Exhausted, but Functional in Patients with COVID-19. *Immunity* **54**, 44–52.e3 (2021).
367. Joglekar, A. V. & Li, G. T cell antigen discovery. *Nature Methods* **18**, 873–880 (2021).

368. Straub, A. *et al.* Recruitment of epitope-specific T cell clones with a low-avidity threshold supports efficacy against mutational escape upon re-infection. *Immunity* **56**, 1269–1284.e6 (2023).
369. Tye, E. X. C. *et al.* Mutations in SARS-CoV-2 spike protein impair epitope-specific CD4+ T cell recognition. (2022) doi:10.1038/S41590-022-01351-7.
370. Heather, J. M. *et al.* Stitchr: stitching coding TCR nucleotide sequences from V/J/CDR3 information. *Nucleic Acids Research* **50**, e68 (2022).
371. Dunbar, J. & Deane, C. M. ANARCI: Antigen receptor numbering and receptor classification. *Bioinformatics* **32**, 298–300 (2016).
372. Hulin-Curtis, S. *et al.* A targeted single mutation in influenza a virus universal epitope transforms immunogenicity and protective immunity via CD4+ T cell activation. *Cell Reports* **43**, 114259 (2024).

373. Quast, N. P. *et al.* T-cell receptor structures and predictive models reveal comparable alpha and beta chain structural diversity despite differing genetic complexity. doi:10.1101/2024.05.20.594940.
374. Chemaitelly, H. *et al.* Differential protection against SARS-CoV-2 reinfection pre- and post-Omicron. *Nature* **639**, 1024–1031 (2025).
375. Cao, Y. *et al.* Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657–663 (2022).
376. Andeweg, S. P. *et al.* Protection of COVID-19 vaccination and previous infection against Omicron BA.1, BA.2 and Delta SARS-CoV-2 infections. *Nature Communications* **13**, 4738 (2022).
377. Mayr, F. B. *et al.* COVID-19 disease severity in US Veterans infected during Omicron and Delta variant predominant periods. *Nature Communications* **13**, 3647 (2022).
378. Tarke, A. *et al.* Impact of SARS-CoV-2 variants on the total CD4+ and CD8+ T cell reactivity in infected or vaccinated individuals. *Cell Reports Medicine* **2**, 100355 (2021).
379. Keeton, R. *et al.* T cell responses to SARS-CoV-2 spike cross-recognize Omicron. *Nature* **603**, 488–492 (2022).
380. Gao, Y. *et al.* Ancestral SARS-CoV-2-specific T cells cross-recognize the Omicron variant. *Nature Medicine* **28**, 472–476 (2022).
381. Tarke, A. *et al.* SARS-CoV-2 breakthrough infections enhance T cell response magnitude, breadth, and epitope repertoire. *Cell Reports Medicine* **5**, 101583 (2024).
382. Mayer-Blackwell, K. *et al.* mRNA vaccination boosts S-specific T cell memory and promotes expansion of CD45RAint TEMRA-like CD8+ T cells in COVID-19 recovered individuals. *Cell Reports Medicine* **4**, 101149 (2023).
383. Sturmlechner, I. *et al.* Antigen specificity shapes distinct aging trajectories of memory $CD8^+$ T cells. *Nature Communications* **16**, 6394 (2025).
384. Liu, Z. *et al.* Cutting Edge: Transcription Factor BCL6 Is Required for the Generation, but Not Maintenance, of Memory CD8+ T Cells in Acute Viral Infection. *Journal of Immunology (Baltimore, Md.: 1950)* **203**, 323–327 (2019).
385. Xu, K. *et al.* Glycolysis fuels phosphoinositide 3-kinase signaling to bolster T cell immunity. *Science* **371**, 405–410 (2021).
386. Sukumar, M. *et al.* Inhibiting glycolytic metabolism enhances CD8+ T cell memory and antitumor function. *The Journal of Clinical Investigation* **123**, 4479–4488 (2013).
387. Banerjee, A. *et al.* The transcription factor eomesodermin enables CD8+ T cells to compete for the memory cell niche. *Journal of immunology (Baltimore, Md. : 1950)* **185**, 4988–4992 (2010).
388. Adamo, S. *et al.* Signature of long-lived memory CD8+ T cells in acute SARS-CoV-2 infection. *Nature* **602**, 148–155 (2022).
389. Zornikova, K., Sheetikov, S. A., Rusinov, A. Y., Iskhakov, R. N. & Bogolyubova, A. Architecture of the SARS-CoV-2-specific T cell repertoire. *Frontiers in Immunology* (2023) doi:10.3389/FIMMU.2023.1070077.
390. Bobrovitz, N. *et al.* Protective effectiveness of previous SARS-CoV-2 infection and hybrid immunity against the omicron variant and severe disease: A systematic review and meta-regression. *The Lancet Infectious Diseases* **23**, 556–567 (2023).

391. Takahashi, T. *et al.* Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature* **588**, 315–320 (2020).
392. Nelson, R. W. *et al.* SARS-CoV-2 epitope–specific CD4+ memory T cell responses across COVID-19 disease severity and antibody durability. *Science Immunology* **7**, eabl9464 (2022).
393. Jung, J. H. *et al.* SARS-CoV-2-specific T cell memory is sustained in COVID-19 convalescent patients for 10 months with successful development of stem cell-like memory T cells. *Nature Communications* **12**, 4043 (2021).
394. Yang, G. *et al.* Initial COVID-19 severity influenced by SARS-CoV-2-specific T cells imprints T-cell memory and inversely affects reinfection. *Signal Transduction and Targeted Therapy* **9**, 141 (2024).
395. Gray-Gaillard, S. L. *et al.* SARS-CoV-2 inflammation durably imprints memory CD4 T cells. *Science Immunology* **9**, eadj8526 (2024).
396. Dykema, A. G. *et al.* SARS-CoV-2 vaccination diversifies the CD4+ spike-reactive T cell repertoire in patients with prior SARS-CoV-2 infection. *eBioMedicine* **80**, (2022).
397. Ford, E. S. *et al.* Repeated mRNA vaccination sequentially boosts SARS-CoV-2-specific CD8+ T cells in persons with previous COVID-19. *Nature Immunology* **25**, 166–177 (2024).
398. Rowntree, L. C. *et al.* SARS-CoV-2-specific T cell memory with common TCR$\alpha\beta$ motifs is established in unvaccinated children who seroconvert after infection. *Immunity* **55**, 1299–1315.e4 (2022).
399. Tsui, C. *et al.* MYB orchestrates T cell exhaustion and response to checkpoint inhibition. *Nature* **609**, 354–360 (2022).
400. Movsisyan, M. *et al.* Tracking the evolution of anti-SARS-CoV-2 antibodies and long-term humoral immunity within 2 years after COVID-19 infection. *Scientific Reports* **14**, 13417 (2024).
401. Wilkinson, T. M. *et al.* Preexisting influenza-specific CD4+ T cells correlate with disease protection against influenza challenge in humans. *Nature Medicine* **18**, 274–280 (2012).
402. Zornikova, K. V. *et al.* Clonal diversity predicts persistence of SARS-CoV-2 epitope-specific T-cell response. *Communications Biology* **5**, 1351 (2022).
403. Gaimann, M. U., Nguyen, M., Desponds, J. & Mayer, A. Early life imprints the hierarchy of T cell clone sizes. *eLife* **9**, e61639 (2020).
404. Pavlova, A. V., Zvyagin, I. V. & Shugay, M. Detecting T-cell clonal expansions and quantifying clone survival using deep profiling of immune repertoires. *Frontiers in Immunology* **15**, 1321603 (2024).
405. Gong, F. *et al.* Peripheral CD4$^+$ T cell subsets and antibody response in COVID-19 convalescent individuals. *The Journal of Clinical Investigation* **130**, 6588–6599 (2020).
406. Cornberg, M. *et al.* Narrowed TCR repertoire and viral escape as a consequence of heterologous immunity. *Journal of Clinical Investigation* **116**, 1443–1456 (2006).
407. Hashimoto, K. *et al.* Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 24242–24251 (2019).

408. Safford, M. *et al.* Egr-2 and Egr-3 are negative regulators of T cell activation. *Nature Immunology* **6**, 472–480 (2005).
409. Maine, C. J., Teijaro, J. R., Marquardt, K. & Sherman, L. A. PTPN22 contributes to exhaustion of T lymphocytes during chronic viral infection. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E7231–E7239 (2016).
410. Bertin, S. *et al.* Dual specificity phosphatase 6 (DUSP6) regulates CD4+ T cell functions and restrains the spontaneous colitis in IL-10 deficient mice. *Mucosal immunology* **8**, 505–515 (2015).
411. Painter, M. M. *et al.* Prior vaccination promotes early activation of memory T cells and enhances immune responses during SARS-CoV-2 breakthrough infection. *Nature Immunology* **24**, 1711–1724 (2023).
412. Tiffeau-Mayer, A. Unbiased estimation of sampling variance for simpson's diversity index. doi:10.48550/arXiv.2310.03439.
413. Laydon, D. J., Bangham, C. R. M. & Asquith, B. Estimating T-cell repertoire diversity: Limitations of classical estimators and a new approach. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20140291 (2015).
414. Minervina, A. A. *et al.* Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. *eLife* **10**, e63502 (2021).
415. Hale, B. D. *et al.* Cellular architecture shapes the naïve T cell response. *Science* **384**, eadh8697 (2024).
416. Kramer, K. J. *et al.* Single-cell profiling of the antigen-specific response to BNT162b2 SARS-CoV-2 RNA vaccine. *Nature Communications* **13**, 3466 (2022).
417. Wang, X. *et al.* Clonal expansion dictates the efficacy of mitochondrial lineage tracing in single cells. *Genome Biology* **26**, 70 (2025).
418. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nature Communications* **12**, 5692 (2021).
419. Zimmerman, K. D. & Langefeld, C. D. Hierarchicell: An r-package for estimating power for tests of differential expression with single-cell data. *BMC Genomics* **22**, 319 (2021).
420. Rius, C. *et al.* Peptide–MHC class I tetramers can fail to detect relevant functional T cell clonotypes and underestimate antigen-reactive T cell populations. *The Journal of Immunology* **200**, 2263–2279 (2018).
421. Yin, K. *et al.* Long COVID manifests with T cell dysregulation, inflammation and an uncoordinated adaptive immune response to SARS-CoV-2. *Nature Immunology* **25**, 218–225 (2024).
422. Netea, M. G. *et al.* Trained immunity: A program of innate immune memory in health and disease. *Science* **352**, aaf1098 (2016).
423. van 't Wout, J. W., Poell, R. & Furth, R. van. The role of BCG/PPD-activated macrophages in resistance against systemic candidiasis in mice. *Scandinavian Journal of Immunology* **36**, 713–719 (1992).
424. Benn, C. S., Netea, M. G., Selin, L. K. & Aaby, P. A small jab – a big effect: Nonspecific immunomodulation by vaccines. *Trends in Immunology* **34**, 431–439 (2013).

425. Aaby, P. *et al.* Non-specific effects of standard measles vaccine at 4.5 and 9 months of age on childhood mortality: randomised controlled trial. *BMJ* **341**, c6495 (2010).

A

Appendix - Introduction

A.1 Innate trained immunity

Trained innate immunity[422] challenges the neat separation of innate and adaptive immunity on the basis of memory. It represents the phenomenon where an initial insult with one pathogen or vaccine can potentiate responses against other, unrelated infections. For example, when vaccinated against Tuberculosis, mice develop protection against Candida albicans in a manner that does not involve T-cell memory[423]. In a clinical trial of an additional, earlier, measles vaccine, the additional vaccine had a 26% reduction in all-cause mortality beyond that which measles infection could explain[424,425].

B

Appendix - Materials and Methods

B.1 Representative Intracellular Cytokine Plots

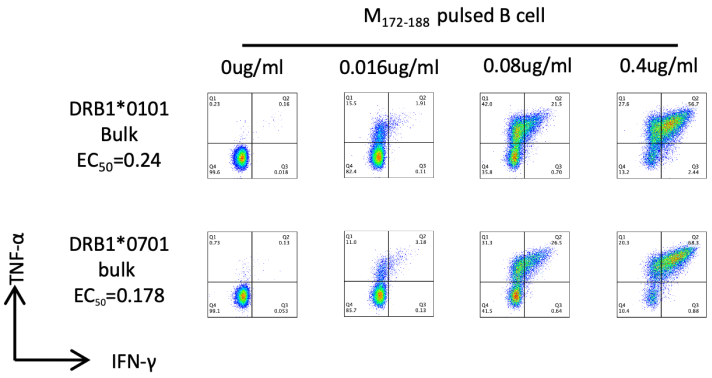


Figure B.1: Representative flow cytometry plot for intracellular cytokine staining of antigen-specific T-cells co-cultured with EBV transformed B-cell lines pulsed with $M_{172\text{-}188}$ peptide.

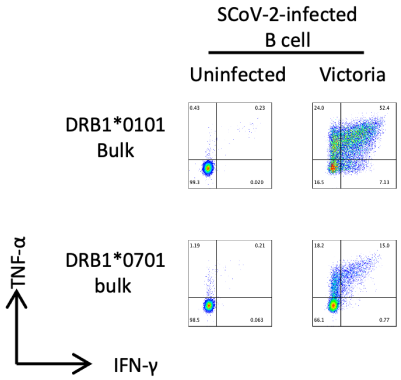


Figure B.2: Representative flow cytometry plot for intracellular cytokine staining of antigen-specific T-cells co-cultured with EBV transformed, ACE2 transduced B-cell lines infected with SARS-CoV2.

C

Appendix - Chapter 3

C.1 Gating strategy for isolating $M_{172\text{-}188}$ specific $CD4^+$ T-cells with tetramers

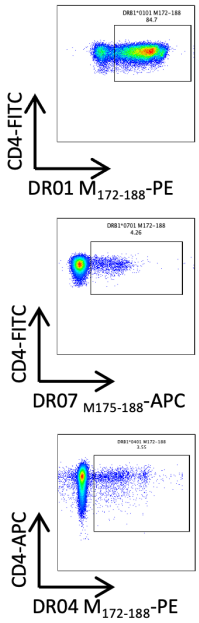


Figure C.1: Representative gating strategy for isolating $M_{172\text{-}188}$ specific $CD4^+$ T-cells restricted to DRB1*01:01, DRB1*07:01 and DRB1*04:01.

C.2 T-cell based interpolation of effective pulsed peptide concentration in live viral $M_{172\text{-}188}$ concentration

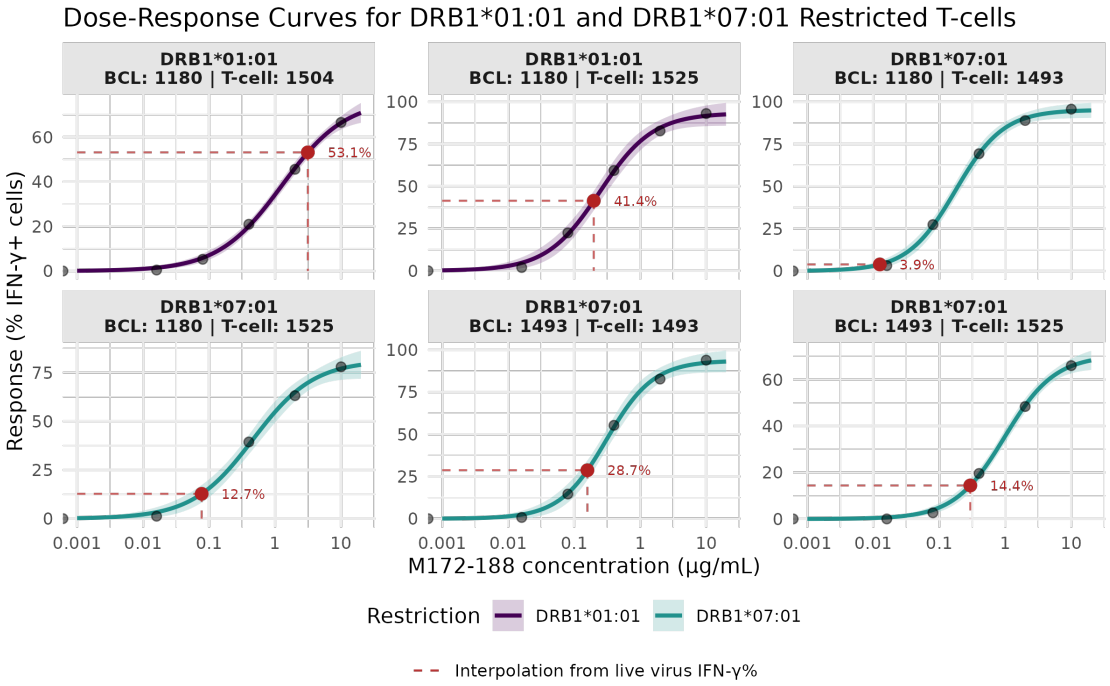


Figure C.2: Dose response curves for DRB1*01:01 and DRB1*07:01 restricted $M_{172\text{-}188}$ specific $CD4^+$ T-cell lines of known restriction were co-cultured with EBV transformed B-cell lines pulsed with varying concentrations of $M_{172\text{-}188}$ peptide. IFN-γ production was measured by intracellular cytokine staining. Each panel shows a distinct B-cell line donor, T-cell line donor and HLA restriction. Curves represent the fit of the three-parameter log-logistic model with 95% confidence intervals (shaded regions). EC50 values (the concentration required for 50% of maximal response) are shown in each panel. Red dashed lines show interpolation of the IFN-γ response observed in co-culture with ACE2-transduced B-cells infected with SARS-CoV2. Experimental data generated by Dr Xuan Yao, model fitting and figure generation by me.

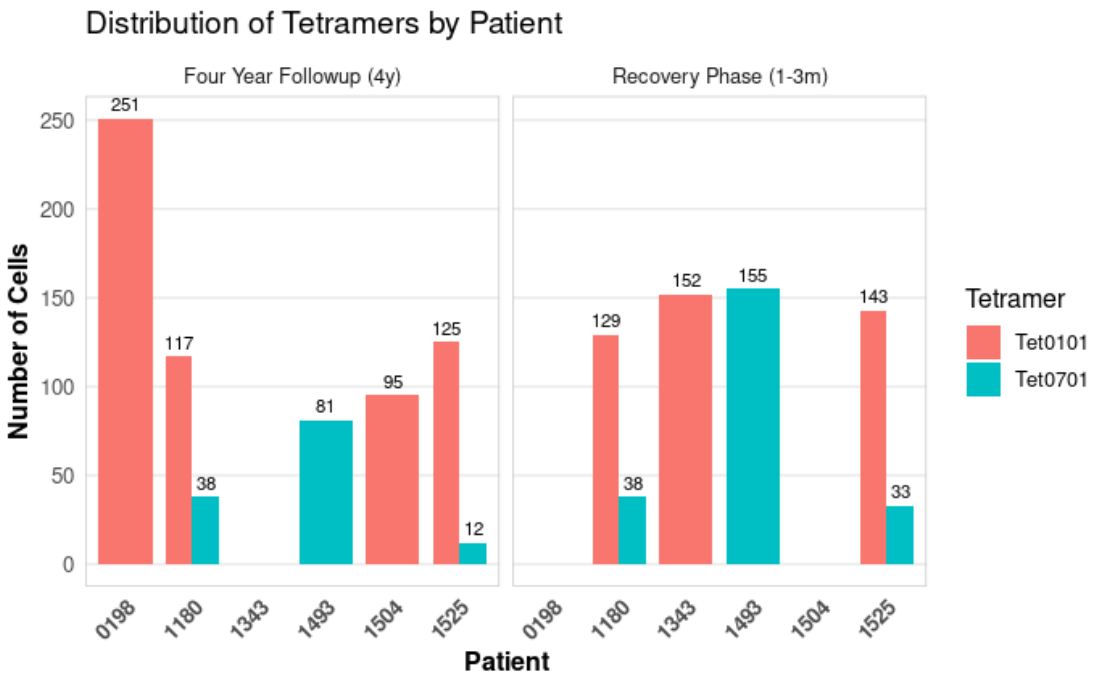


Figure C.3: The number of cells isolated, at both the Recovery Phase (1-3 months after acute infection) and at Four Year Followup. Cells specific for DRB1*01:01 are highlighted in the red bars, whereas those specific for DRB1*07:01 are in blue.

C.3 $M_{172\text{-}188}$ specific CD4⁺ T-cells isolated at each timepoint, by tetramer and by patient

C.4 QC metrics of single cell RNA-seq data of $M_{172\text{-}188}$ specific CD4⁺ T-cells

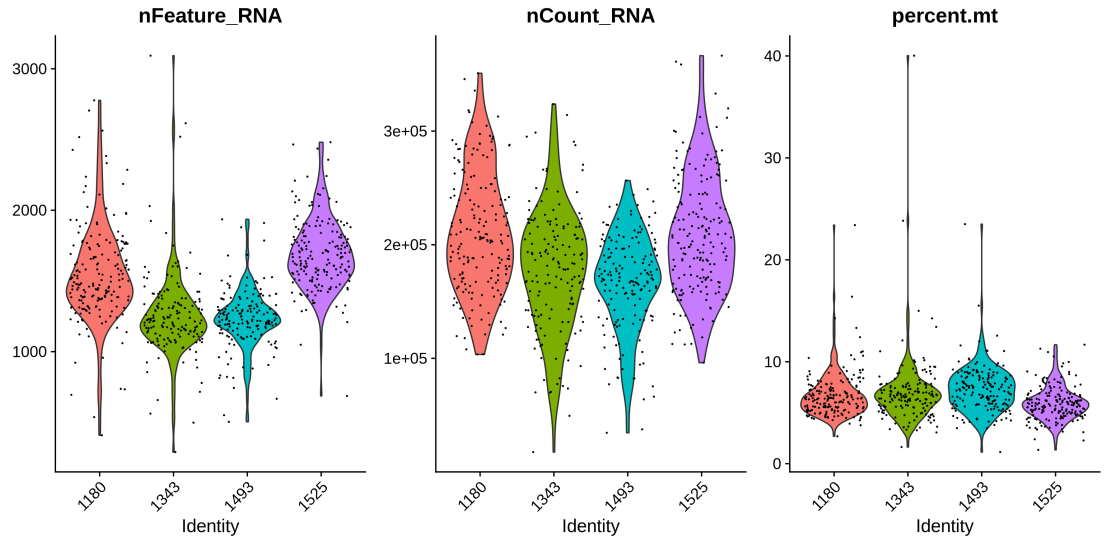


Figure C.4: A violin plot of QC metrics for $M_{172\text{-}188}$ specific CD4⁺ T-cells isolated in the 1-3 month recovery phase split by donor id: the number of unique genes detected per cell (nFeature_RNA), the number of unique transcripts detected per cell (nCount_RNA), and the percentage of mitochondrial gene counts per cell (percent.mt).

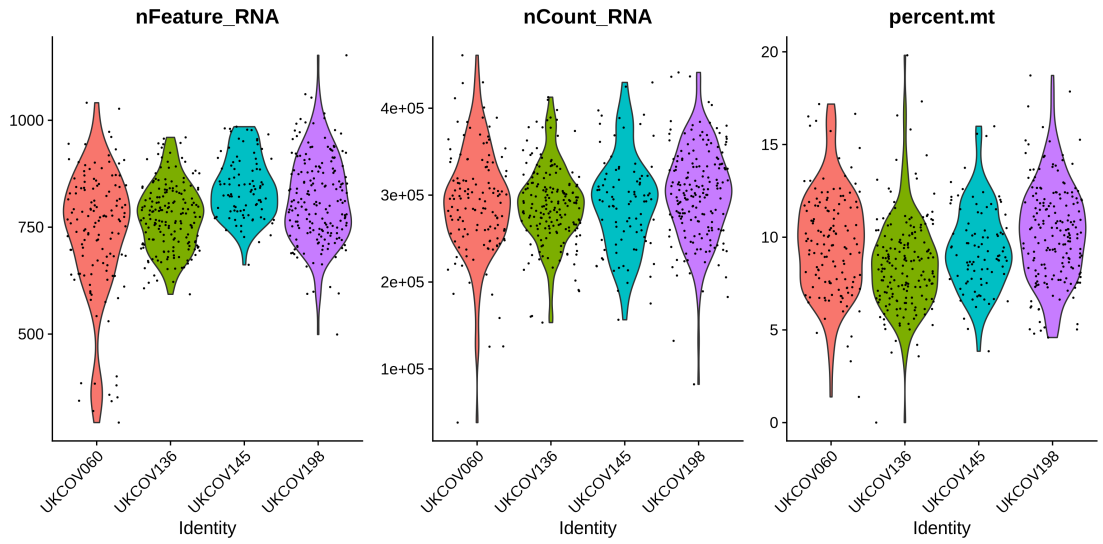


Figure C.5: A violin plot of QC metrics for $M_{172\text{-}188}$ specific $CD4^+$ T-cells isolated at the four year followup split by donor id: the number of unique genes detected per cell (nFeature_RNA), the number of unique transcripts detected per cell (nCount_RNA), and the percentage of mitochondrial gene counts per cell (percent.mt).

On inspecting these QC metrics, and in line with best practices for single-cell RNA-seq data analysis, cells with a percentage of mitochondrial gene counts greater than 15% were removed prior to analysis or visualisation. Cells were required to have between 300 and 3500 unique genes detected. Genes expressed in fewer than five cells were also removed from the data matrix.

D

Appendix - Chapter 4

D.1 Dynamics of per-Sample Cluster proportion between pre- and post-Omicron timepoints

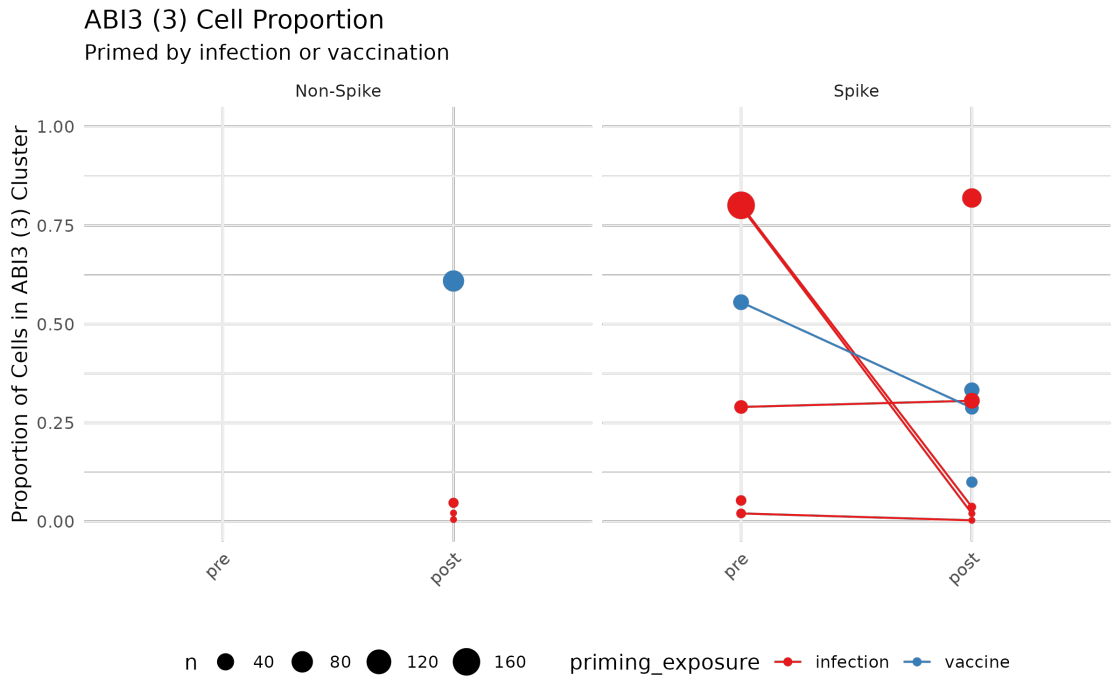


Figure D.1: ABI3 (3) cluster dynamics between pre- and post-Omicron timepoints. Scatter plot showing changes in ABI3 (3) cluster proportions from pre- to post-Omicron timepoints for each sample. Where samples were taken for the same epitope in the same individual (longitudinal), they are connected. Lines are coloured by priming exposure (infection vs. vaccine), and point size represents absolute cell counts. Plots are faceted by spike vs. non-spike epitopes.

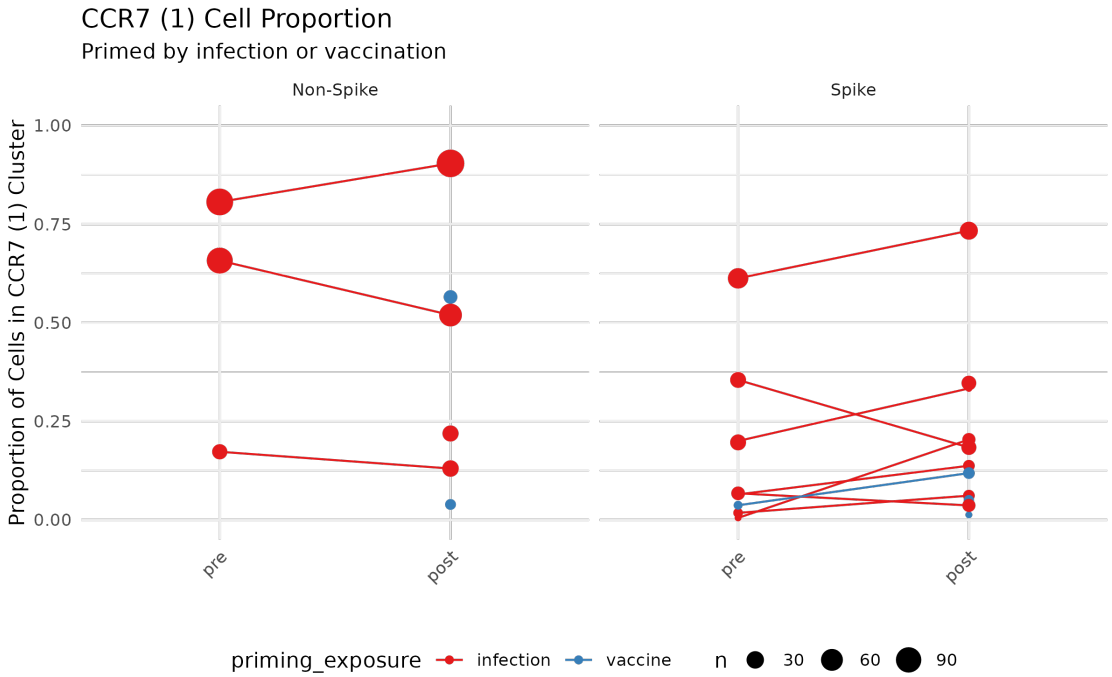


Figure D.2: CCR7 (1) cluster dynamics between pre- and post-Omicron timepoints. Scatter plot showing changes in CCR7 (1) cluster proportions from pre- to post-Omicron timepoints for each sample. Where samples were taken for the same epitope in the same individual (longitudinal), they are connected. Lines are coloured by priming exposure (infection vs. vaccine), and point size represents absolute cell counts. Plots are faceted by spike vs. non-spike epitopes.

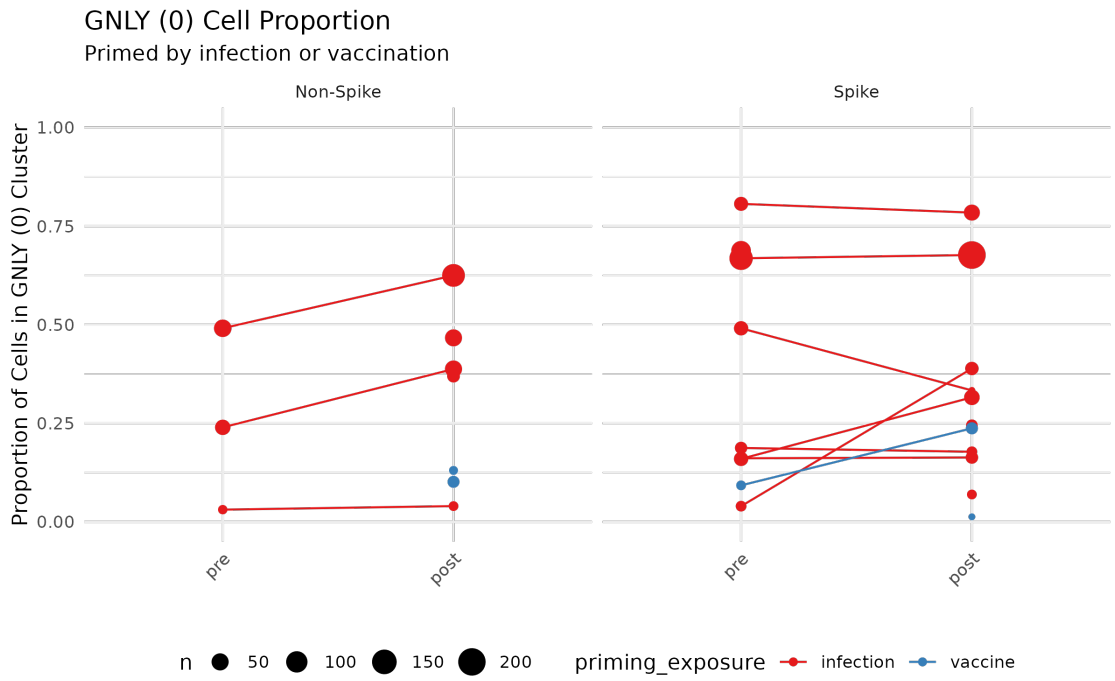


Figure D.3: GNLY (0) cluster dynamics between pre- and post-Omicron timepoints. Scatter plot showing changes in GNLY (0) cluster proportions from pre- to post-Omicron timepoints for each sample. Where samples were taken for the same epitope in the same individual (longitudinal), they are connected. Lines are coloured by priming exposure (infection vs. vaccine), and point size represents absolute cell counts. Plots are faceted by spike vs. non-spike epitopes.

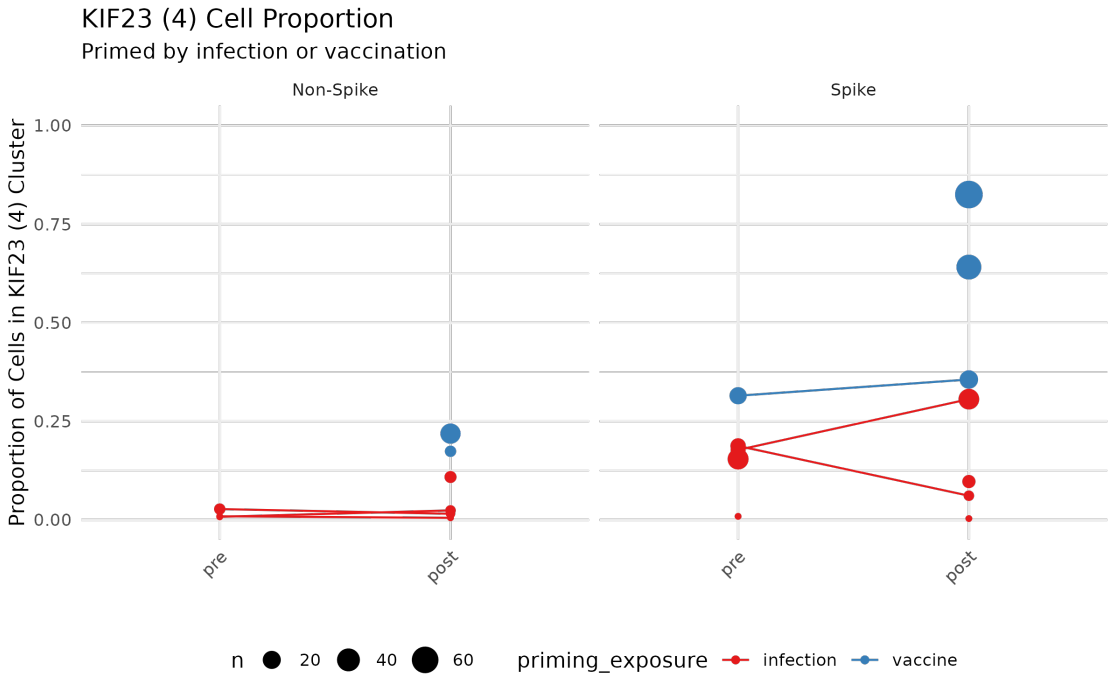


Figure D.4: KIF23 (4) cluster dynamics between pre- and post-Omicron timepoints. Scatter plot showing changes in KIF23 (4) cluster proportions from pre- to post-Omicron timepoints for each sample. Where samples were taken for the same epitope in the same individual (longitudinal), they are connected. Lines are coloured by priming exposure (infection vs. vaccine), and point size represents absolute cell counts. Plots are faceted by spike vs. non-spike epitopes.

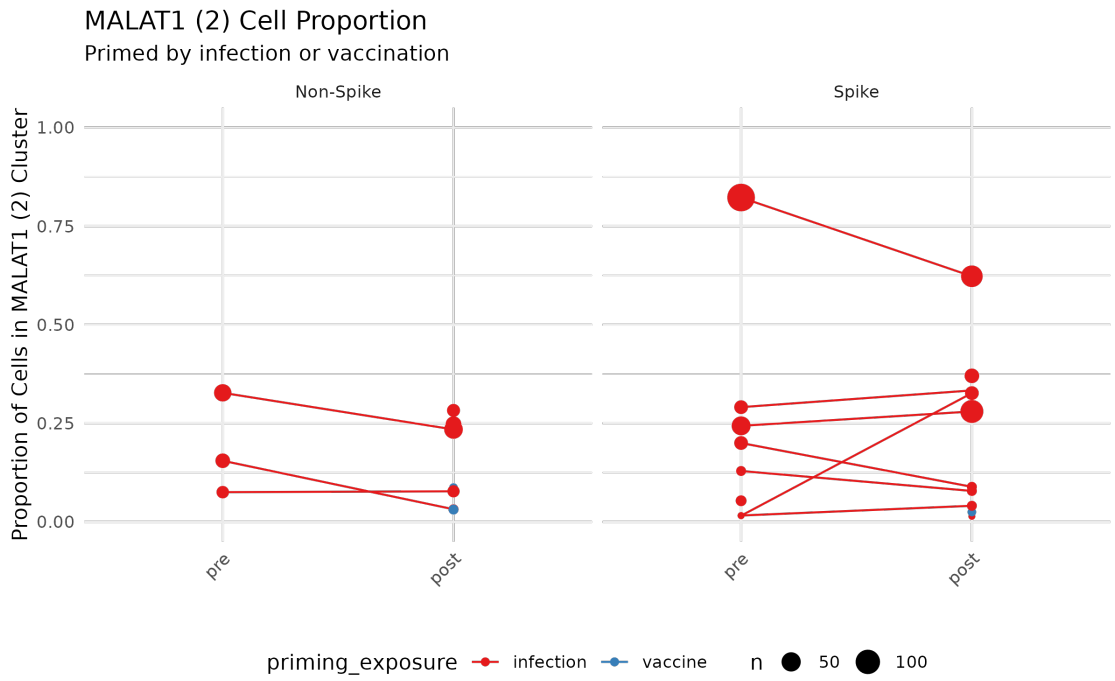


Figure D.5: MALAT1 (2) cluster dynamics between pre- and post-Omicron timepoints. Scatter plot showing changes in MALAT1 (2) cluster proportions from pre- to post-Omicron timepoints for each sample. Where samples were taken for the same epitope in the same individual (longitudinal), they are connected. Lines are coloured by priming exposure (infection vs. vaccine), and point size represents absolute cell counts. Plots are faceted by spike vs. non-spike epitopes.

D.2 Gene Ontology 2025 Biological Process Enrichment of Differentially Expressed Genes

Table D.1: Table of top Gene Ontology Biological Process terms (2025 database) from gene set enrichment analysis of genes differentially expressed in Spike-specific $CD8^+$ T-cells pre-Omicron vs post-Omicron breakthrough infection in vaccine primed individuals. Genes with adjusted p-value < 0.05 and positive `log2FoldChange` (more abundant pre-Omicron) were analysed using Enrichr. Shows the top 10 most enriched terms.

Term	Overlap	P-value	Adjusted P-value	Genes
Regulation of Mast Cell Degranulation (GO:0043304)	2/11	0.0001714	0.0338146	FGR; SPHK2

Term	Overlap	P-value	Adjusted P-value	Genes
Positive Regulation of Leukocyte Degranulation (GO:0043302)	2/12	0.0002055	0.0338146	FGR; SPHK2
Carbohydrate Metabolic Process (GO:0005975)	3/85	0.0004779	0.0524186	FGGY; GNPDA1; SLC2A8
Peptidyl-Tyrosine Dephosphorylation (GO:0035335)	2/22	0.0007113	0.0585098	PTPN12; DUSP6
Negative Regulation of DNA-binding Transcription Factor Activity (GO:0043433)	3/122	0.0013650	0.089821	EOMES; TCF7L2; PTCH1
Positive Regulation of Production of Molecular Mediator of Immune Response (GO:0002702)	2/43	0.0027155	0.129659	GNPDA1; SPHK2
+ Reg of Phosphatidylinositol 3-Kinase/Prot Kinase B Signal Transduction (GO:0051897)	3/162	0.0030613	0.129659	FGR; TCF7L2; AMBRA1
Cellular Response to Alcohol (GO:0097306)	2/56	0.0045637	0.129659	SPHK2; PTCH1

Term	Overlap	P-value	Adjusted P-value	Genes
Regulation of Phosphatidylinositol 3-Kinase/Protein Kinase B Signal Transduction (GO:0051896)	3/219	0.0070862	0.129659	FGR; TCF7L2; AMBRA1
Negative Regulation of DNA-templated Transcription (GO:0045892)	6/1006	0.0085131	0.129659	EOMES; DNAJB1; TCF7L2; ZHX3; PTCH1; CPEB3
Maintenance of DNA Repeat Elements (GO:0043570)	1/5	0.0089684	0.129659	TCF7L2
N-acetylneuraminat Catabolic Process (GO:0019262)	1/5	0.0089684	0.129659	GNPDA1

Table D.2: Table of top Gene Ontology Biological Process terms (2025 database) from gene set enrichment analysis of genes differentially expressed in Spike-specific CD8⁺ T-cells pre-Omicron vs post-Omicron breakthrough infection in vaccine primed individuals. Genes with adjusted p-value < 0.05 and negative `log2FoldChange` (more abundant post-Omicron) were analysed using Enrichr. Shows the top 10 most enriched terms.

Term	Overlap	P-value	Adjusted P-value	Genes
Protein Glycosylation (GO:0006486)	13/145	0.0003201	0.240916	B3GALNT2; GALNT6; COG7; ALG6; GALNT2; FUT10; MOGS; POMGNT1; C1GALT1C1; B3GNT7; MGAT5; RFT1; B4GALT7
Autophagy of Mitochondrion (GO:0000422)	8/61	0.0003669	0.240916	SPATA33; CTSK; CLEC16A; ATG4A; ULK1; USP30; FBXL4; NIPSNAP1

Term	Overlap	P-value	Adjusted P-value	Genes
Phospholipid Dephosphorylation (GO:0046839)	6/34	0.0003937	0.240916	FIG4; MTMR10; MTMR3; INPP5F; MTMR4; SGPP1
Regulation of Catabolic Process (GO:0009894)	9/81	0.0005623	0.240916	SCARB2; DRAM1; USP32; FAF1; KAT8; ABL2; SIRT6; ULK1; FURIN
Protein Export From Nucleus (GO:0006611)	5/25	0.0006669	0.240916	EGR2; SMURF1; STRADB; XPO7; PKD1
Recombinational Repair (GO:0000725)	10/101	0.0007113	0.240916	MBTD1; ZFYVE26; RAD51C; SFR1; ERCC4; MCM9; TOP3A; NABP2; AP5S1; SPIDR
Macroautophagy (GO:0016236)	13/162	0.0009193	0.266887	MTMR3; CLEC16A; ZFYVE1; NIPSNAP1; ZFYVE26; SPATA33; CTSK; ULK1; ATG4A; VIPAS39; USP30; TECPR1; UBXN6
Retrograde Transport, Vesicle Recycling Within Golgi (GO:0000301)	3/8	0.0012245	0.276481	COG7; COG6; COG4
Necroptotic Process (GO:0070266)	3/8	0.0012245	0.276481	MLKL; RIPK3; IPMK
Branched-Chain Amino Acid Catabolic Process (GO:0009083)	4/18	0.0015589	0.316777	HMGCL; ALDH6A1; BCKDHB; DBT

Table D.3: Table of top Gene Ontology Biological Process terms (2025 database) from gene set enrichment analysis of genes differentially expressed in Spike-specific $CD8^+$ T-cells pre-Omicron vs post-Omicron breakthrough infection in infection primed individuals. Genes with adjusted p-value < 0.05 and positive `log2FoldChange` (more abundant pre-Omicron) were analysed using Enrichr. Shows the top 10 most enriched terms.

Term	Overlap	P-value	Adjusted P-value	Genes
tRNA Modification (GO:0006400)	3/71	0.0106647	0.389668	ANKRD16; PUS1; TRMT61A
Activin Receptor Signaling Pathway (GO:0032924)	2/25	0.0110733	0.389668	GDF11; ACVR1C
Plasma Membrane Bounded Cell Projection Organization (GO:0120036)	4/141	0.0128831	0.389668	MAP1B; CDNF; PTPRM; PTPRG
Regulation of Microtubule Depolymerization (GO:0031114)	2/28	0.0137793	0.389668	TPX2; MAP1B
Regulation of Calcium Ion Transmembrane Transport (GO:1903169)	2/28	0.0137793	0.389668	CACNB1; KCNN4
Neuron Development (GO:0048666)	4/159	0.0191882	0.389668	MAP1B; CDNF; PTPRM; PTPRG
Regulation of Synapse Organization (GO:0050807)	2/38	0.0245861	0.389668	ABHD17C; MDGA1
Metal Ion Transport (GO:0030001)	4/180	0.028611	0.389668	CACNB1; KCNC4; KCNN4; RAMP1

Term	Overlap	P-value	Adjusted P-value	Genes
Positive Regulation of Mitochondrial Fusion (GO:0010636)	1/5	0.0315959	0.389668	PLD6
Negative Regulation of Morphogenesis of an Epithelium (GO:1905331)	1/5	0.0315959	0.389668	CTNND1

Table D.4: Table of top Gene Ontology Biological Process terms (2025 database) from gene set enrichment analysis of genes differentially expressed in Spike-specific $CD8^+$ T-cells pre-Omicron vs post-Omicron breakthrough infection in infection primed individuals. Genes with adjusted p-value < 0.05 and negative `log2FoldChange` (more abundant post-Omicron) were analysed using Enrichr. Shows the top 10 most enriched terms.

Term	Overlap	P-value	Adjusted P-value	Genes
Nucleotide-Sugar Metabolic Process (GO:0009225)	2/14	0.0016081	0.351573	PARG; ENTPD5
Neuromuscular Junction Development (GO:0007528)	2/20	0.0033018	0.351573	CHRNA1; PPFIBP2
Mitotic Sister Chromatid Segregation (GO:0000070)	3/120	0.015025	0.351573	PSRC1; SPDL1; ZWINT
Mitotic Nuclear Division (GO:0140014)	2/47	0.0174293	0.351573	UBE2C; ZWINT

Term	Overlap	P-value	Adjusted P-value	Genes
Purine Deoxyribonucleotide Catabolic Process (GO:0009155)	1/5	0.0213179	0.351573	NUDT18
UDP Metabolic Process (GO:0046048)	1/5	0.0213179	0.351573	ENTPD5
Positive Regulation of cGAS/STING Signaling Pathway (GO:0141111)	1/5	0.0213179	0.351573	ZDHHC9
Metaphase Chromosome Alignment (GO:0051310)	2/55	0.0234202	0.351573	PSRC1; SPDL1
Mitotic Metaphase Chromosome Alignment (GO:0007080)	2/56	0.0242207	0.351573	PSRC1; SPDL1
Peptidyl-Lysine Monomethylation (GO:0018026)	1/6	0.0255272	0.351573	SETD7

Table D.5: Table of top Gene Ontology Biological Process terms (2025 database) from gene set enrichment analysis of genes differentially expressed between Spike-specific $CD8^+$ T-cells in individuals primed by vaccination post-Omicron (3x mRNA, infection) vs after completion of vaccine course in infection primed individuals (infection, 3x mRNA). Genes with adjusted p-value < 0.05 and positive `log2FoldChange` (more abundant in vaccine first) were analysed using Enrichr. Shows the top 10 most enriched terms.

Term	Overlap	P-value	Adjusted P-value	Genes
Negative Regulation of B Cell Proliferation (GO:0030889)	2/10	0.0005489	0.20966	IL10; TYROBP
Regulation of Activin Receptor Signaling Pathway (GO:0032925)	2/18	0.0018324	0.20966	SHH; ACVR2A
Neuromuscular Junction Development (GO:0007528)	2/20	0.0022651	0.20966	CHRNA1; PPFIBP2
Negative Regulation of B Cell Activation (GO:0050869)	2/20	0.0022651	0.20966	IL10; TYROBP
Positive Regulation of Vascular Associated Smooth Muscle Cell Proliferation (GO:1904707)	2/21	0.0024978	0.20966	IL10; MMP9
Negative Regulation of Smooth Muscle Cell Proliferation (GO:0048662)	2/36	0.0072408	0.20966	IL10; IGFBP3

Term	Overlap	P-value	Adjusted P-value	Genes
Regulation of Vascular Associated Smooth Muscle Cell Proliferation (GO:1904705)	2/38	0.008043(	0.20966	IL10; MMP9
Negative Regulation of Type I Interferon Production (GO:0032480)	2/41	0.0093177	0.20966	IL10; TYROBP
Negative Regulation of Lymphocyte Proliferation (GO:0050672)	2/43	0.0102142	0.20966	IL10; TYROBP
Regulation of B Cell Proliferation (GO:0030888)	2/45	0.0111474	0.20966	IL10; TYROBP

Table D.6: Table of top Gene Ontology Biological Process terms (2025 database) from gene set enrichment analysis of genes differentially expressed between Spike-specific $CD8^+$ T-cells in individuals primed by vaccination post-Omicron (3x mRNA, infection) vs after completion of vaccine course in infection primed individuals (infection, 3x mRNA). Genes with adjusted p-value < 0.05 and negative `log2FoldChange` (more abundant in infection-first) were analysed using Enrichr. Shows the top 10 most enriched terms.

Term	Overlap	P-value	Adjusted P-value	Genes
Regulation of Protein Depolymerization (GO:1901879)	2/6	0.0028226	0.64727	MAP1S; MAP1B
Embryonic Skeletal Joint Morphogenesis (GO:0060272)	2/7	0.0039152	0.64727	OSR2; NOG

Term	Overlap	P-value	Adjusted P-value	Genes
Embryonic Skeletal Joint Development (GO:0072498)	2/8	0.0051722	0.64727	OSR2; NOG
Regulation of Ceramide Biosynthetic Process (GO:2000303)	2/10	0.0081604	0.64727	SPHK2; PLA2G6
Dorsal/Ventral Pattern Formation (GO:0009953)	3/30	0.0083353	0.64727	BMP1; PTCH1; NOG
Phosphorylation (GO:0016310)	10/298	0.0095585	0.64727	BCR; FGR; NUAK1; NUAK2; FGGY; CDK2; CTSG; NEK3; NLK; MAP3K12
Regulation of Mast Cell Degranulation (GO:0043304)	2/11	0.0098822	0.64727	FGR; SPHK2
acetyl-CoA Biosynthetic Process (GO:0006085)	2/12	0.0117499	0.64727	PDHX; ACSS2
Positive Regulation of Leukocyte Degranulation (GO:0043302)	2/12	0.0117499	0.64727	FGR; SPHK2
Protein Autoubiquitination (GO:0051865)	4/65	0.0130534	0.64727	SH3RF2; TRIM68; ERCC8; RNF122

D.3 Paired T-cell Receptor Yield from Samples

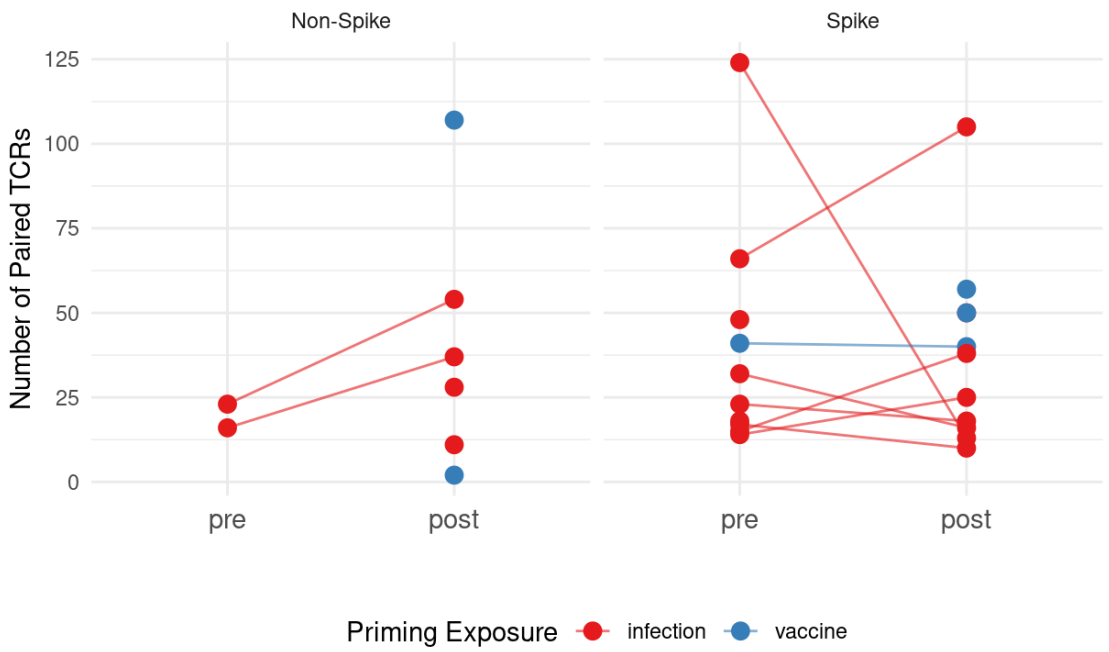


Figure D.6: The yield of paired TCRs from samples. Each line represents a unique patient-epitope combination tracked from pre- to post-Omicron timepoints. Lines are coloured by priming exposure (red = infection, blue = vaccine). 29 patient-epitope combinations that included any TCRs are shown.

D.4 Logistic Regression of Clonotype Maintenance

Table D.7: Example values for logistic regression model predicting clonotype maintenance. Each row represents a unique clonotype (paired $\alpha\beta$ TCR) detected at the pre-Omicron timepoint in a specific patient-epitope combination.

Priming exposure	Epitope group	Post-Omicron cells	ABI3 (3)	CCR7 (1)	KIF23 (4)	MALAT (2)	Maintained
infection	Spike	15	1	0	0	0	0
infection	Non-Spike	44	0	1	0	0	1
infection	Spike	15	4	0	0	0	0
infection	Spike	25	0	0	1	0	0
infection	Spike	14	0	0	0	1	0

Table D.8: Model coefficients for logistic regression of clonotype maintenance. For clonotypes detected at pre-Omicron timepoint, the model predicts the odds of detecting it again at the post-Omicron timepoint based on transcriptional cluster composition, priming exposure, epitope group, and post-Omicron cell count. CCR7 (1) serves as the reference cluster; coefficients for other clusters represent effects relative to CCR7 (1). Reference categories: vaccine priming, non-Spike epitopes. McFadden's $R^2 = 0.24$; n = 167.

Term	Coefficient	SE	Odds Ratio	95% CI	z value	p value
(Intercept)	-1.792	0.913	0.167	0.02–0.9	-1.964	0.05
ABI3 (3)	0.290	0.126	1.336	1.07–1.75	2.301	0.021
GNLY (0)	1.420	0.391	4.136	2.04–9.33	3.632	<0.001
KIF23 (4)	-0.731	0.413	0.481	0.2–1.03	-1.772	0.076
MALAT1 (2)	0.735	0.354	2.086	1.12–4.6	2.077	0.038
Priming exposure (infection)	0.439	0.747	1.551	0.4–8.15	0.587	0.557
Post-Omicron cell count	0.003	0.002	1.003	1–1.01	1.375	0.169
Epitope group (Spike)	-0.216	0.532	0.806	0.29–2.36	-0.406	0.685

E

Appendix - Chapter 5

E.1 Quality Control Metrics for Spike and Non-Spike specific CD4⁺ T-cells

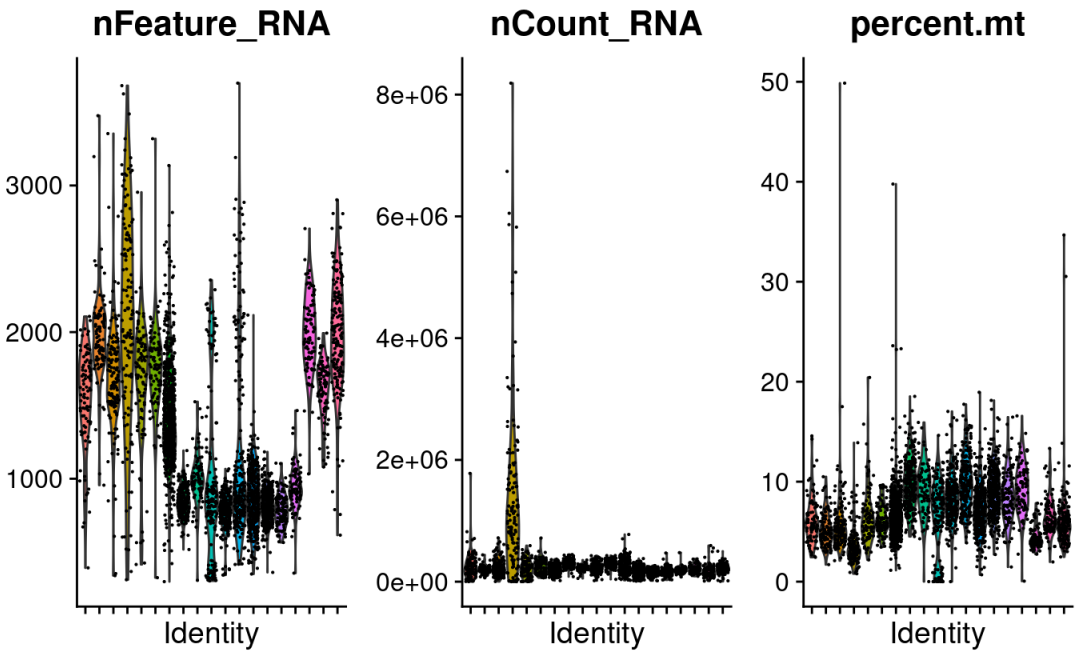


Figure E.1: Violin Plot of control measurements, including number of features (genes detected per cell), counts (the total counts of all genes in a cell), and the percent of overall counts attributed to mitochondrial genes of all samples of Spike and non-Spike CD4⁺ T-cells. Violin Plots are grouped by the sequencing run in which the cells were sequenced.

On the basis of inspecting these metrics, and adapting recommendations in single-cell best practices, we filtered cells to include: Features detected per cell > 500, Counts per cell $< 5.1 \times 10^5$, Percent mitochondrial reads $< 15\%$. After filtering, 3429 cells from 16 individuals were retained for analysis. The median number of features detected per cell was 1740 (IQR 1300-2200), the median number of counts per cell was 3.6 x 10^4 (IQR 2.7 x 10^4 - 4.8 x 10^4) and the median percentage of mitochondrial reads was 2.3% (IQR 1.5-3.5%).

E.2 Principal Components Weightings by Principal Component

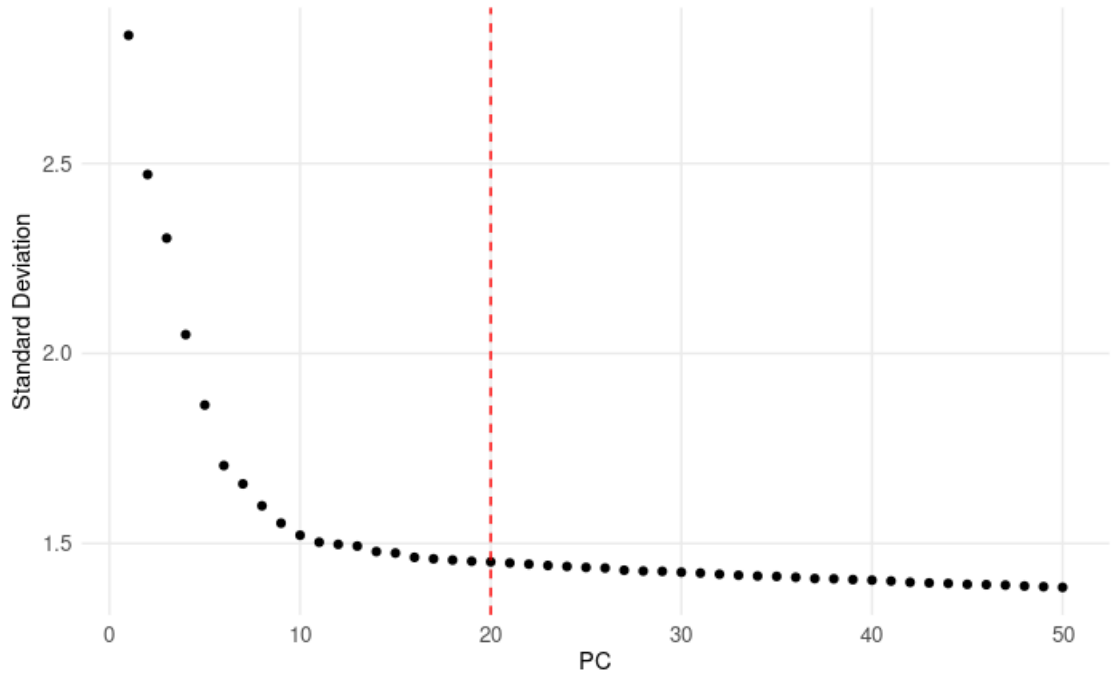


Figure E.2: Elbow plot of the standard deviation of each principal component in the PCA of Log Normalised RNA counts of Spike and Non-Spike specific CD4⁺ T-cells.

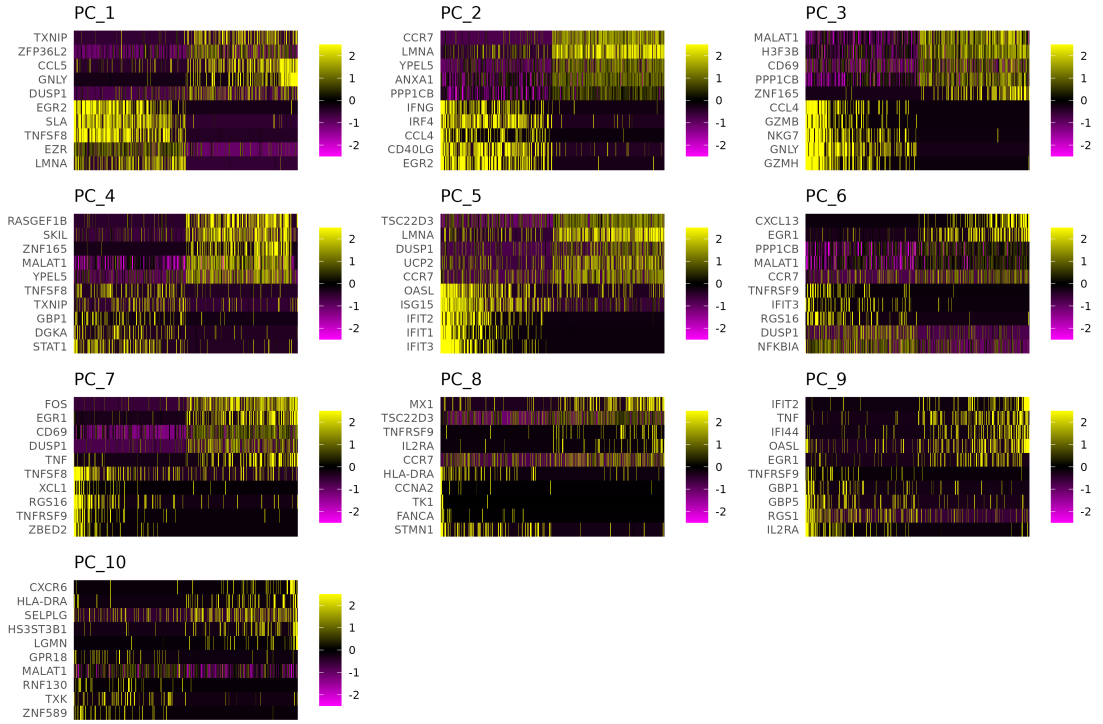


Figure E.3: Heatmap representing the 5 most positively and 5 most negatively weighted genes for each principal component of components 1-10.

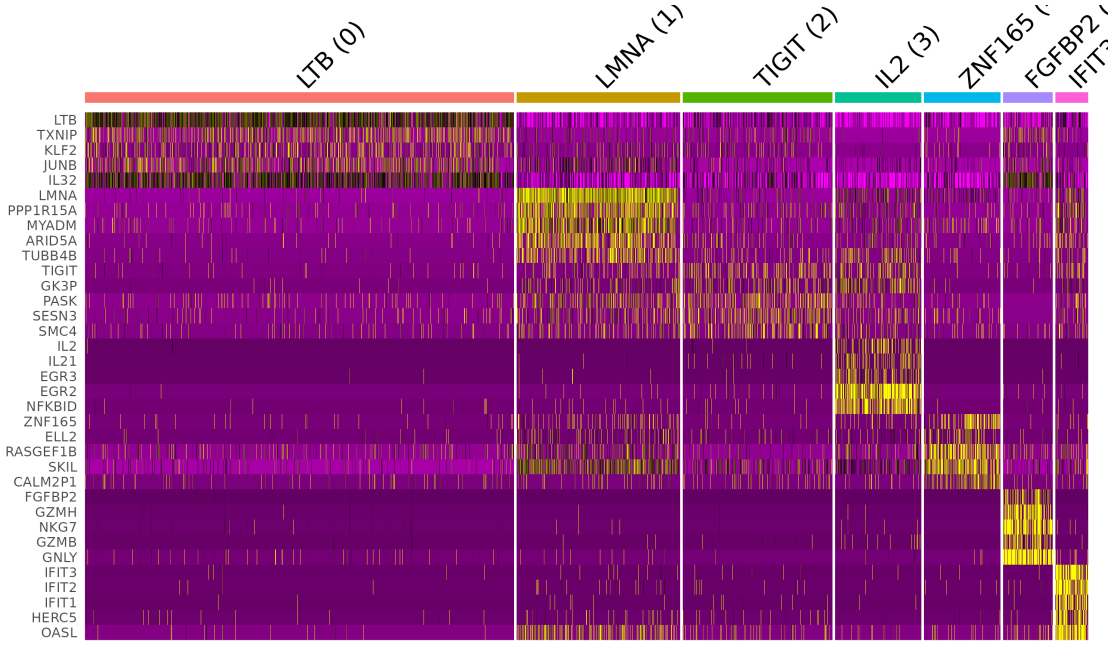


Figure E.4: Heatmap of differentially expressed genes in each cluster. The top 5 most differentially expressed genes per cluster are shown.

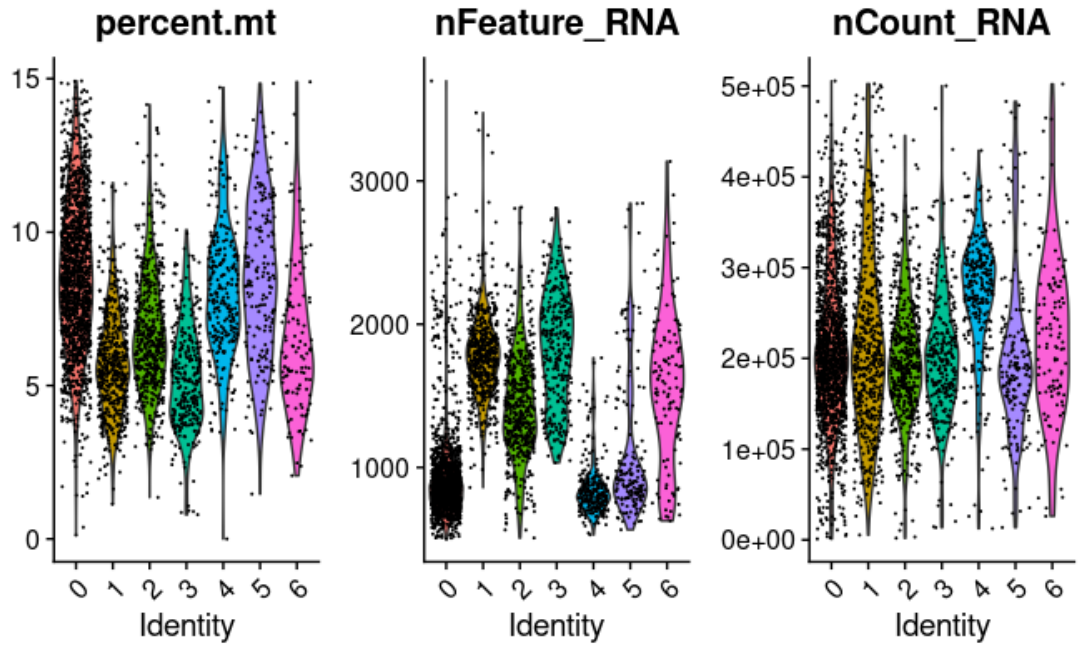


Figure E.5: Violin Plot of Percentage Mitochondrial RNA counts (percent.mt), Number of Genes Detected (nFeature_RNA) and Number of RNA counts (nCount_RNA) per cell by cluster.

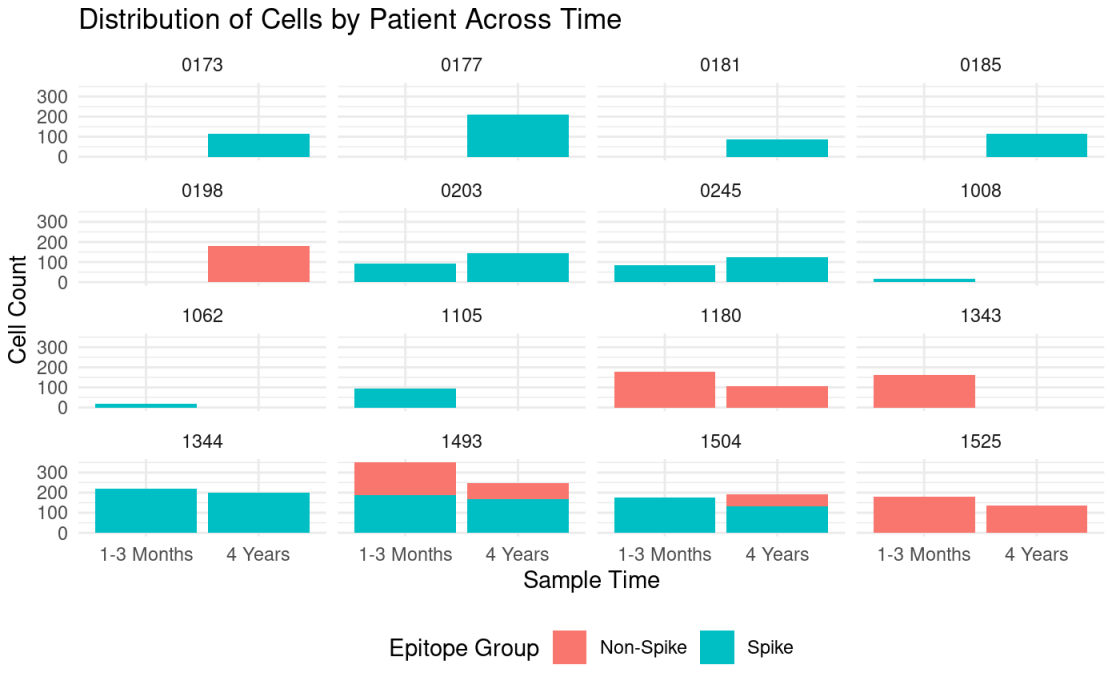


Figure E.6: Cell numbers by Epitope Group and Timepoint.

E.3 Interpretation of Transcriptional Clusters Genes

E.4 Cell Numbers by Epitope Group and Timepoint

E.5 Gene Usage Frequencies between 1-3 Month and 4 Year Timepoints after Initial SARS-CoV2 Infection

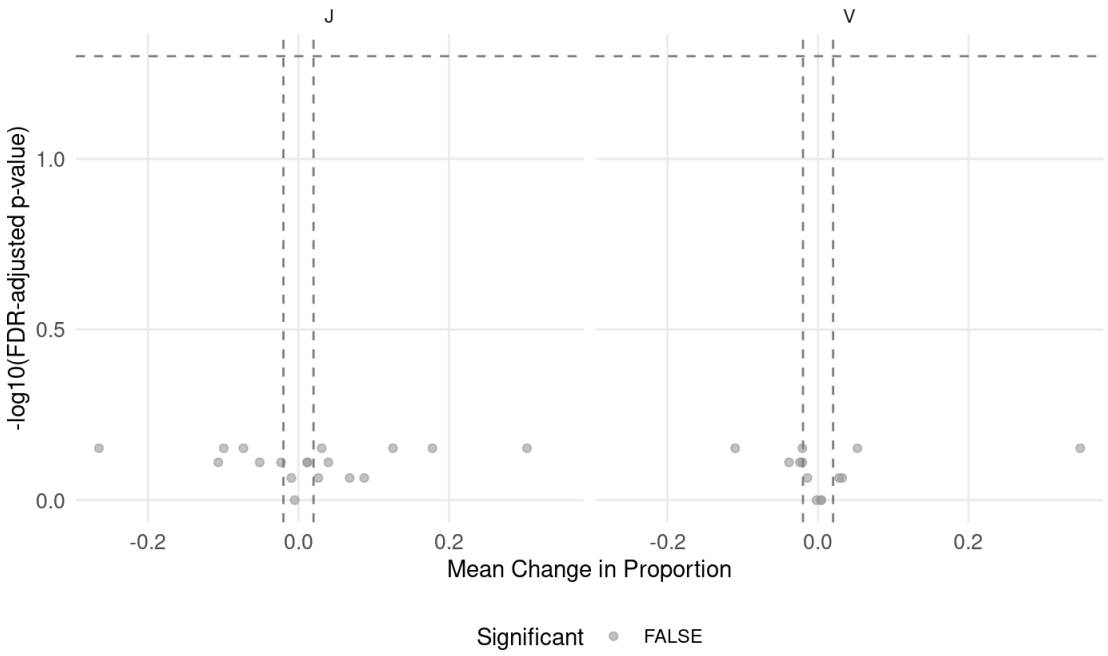


Figure E.7: Volcano plot of mean change in TCR gene usage proportion against $\log_{10}$ p-value across all epitopes from early to four year timepoint. Each point represents an individual V or J gene, with red indicating genes that significantly changed in frequency.

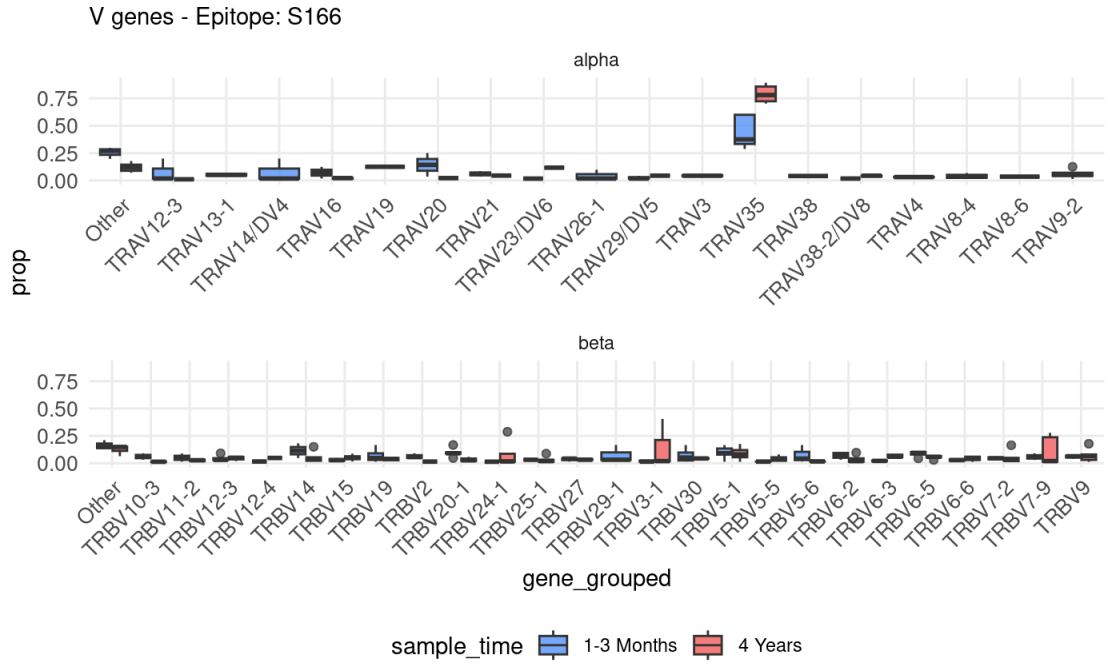


Figure E.8: TCR V gene usage temporal comparison for S166-specific CD4⁺ T-cells. Boxplots of the proportion of V genes at the early (1-3 months, blue) versus long-term (4 years, red) timepoints.

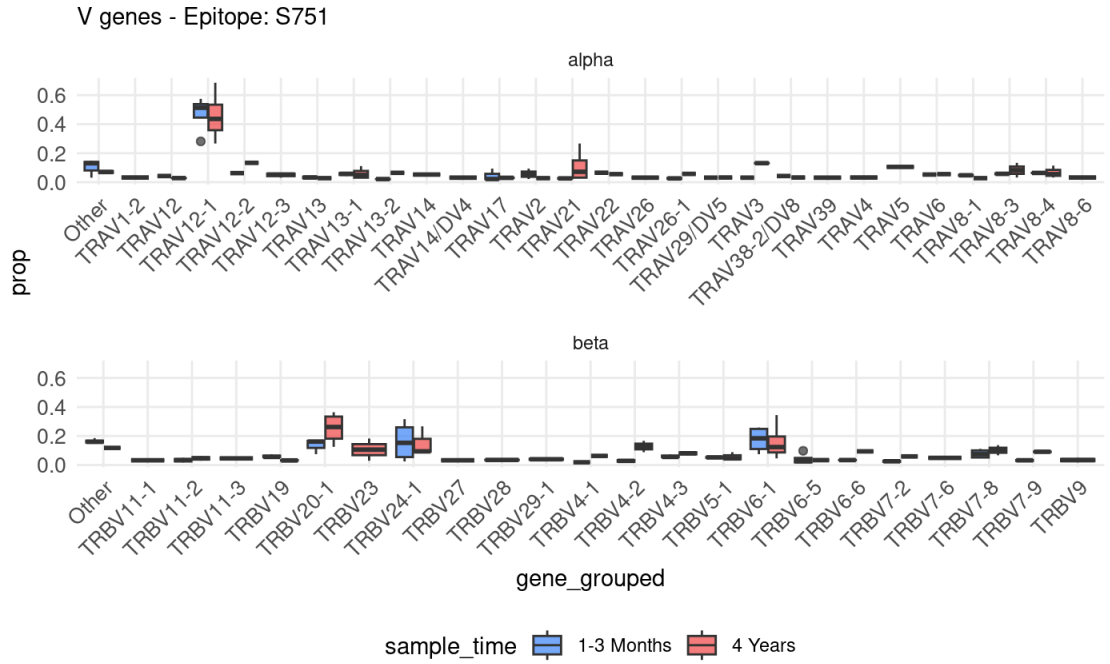


Figure E.9: TCR V gene usage temporal comparison for S751-specific CD4⁺ T-cells showing alpha and beta chain V gene proportions at early versus late timepoints.

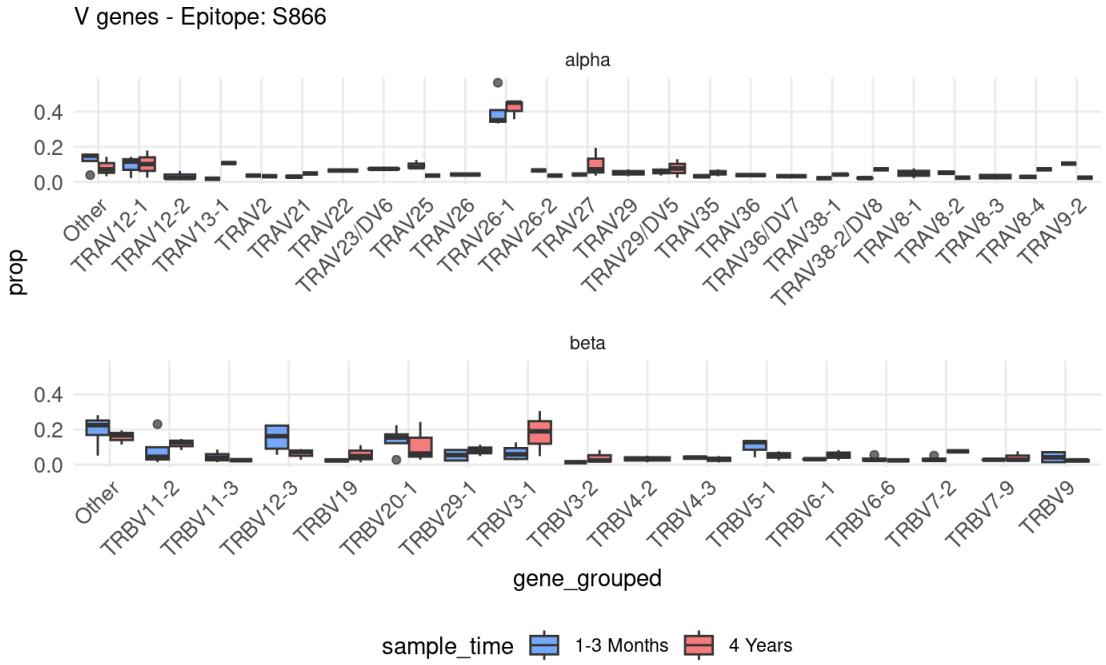


Figure E.10: TCR V gene usage temporal comparison for S866-specific CD4⁺ T-cells.

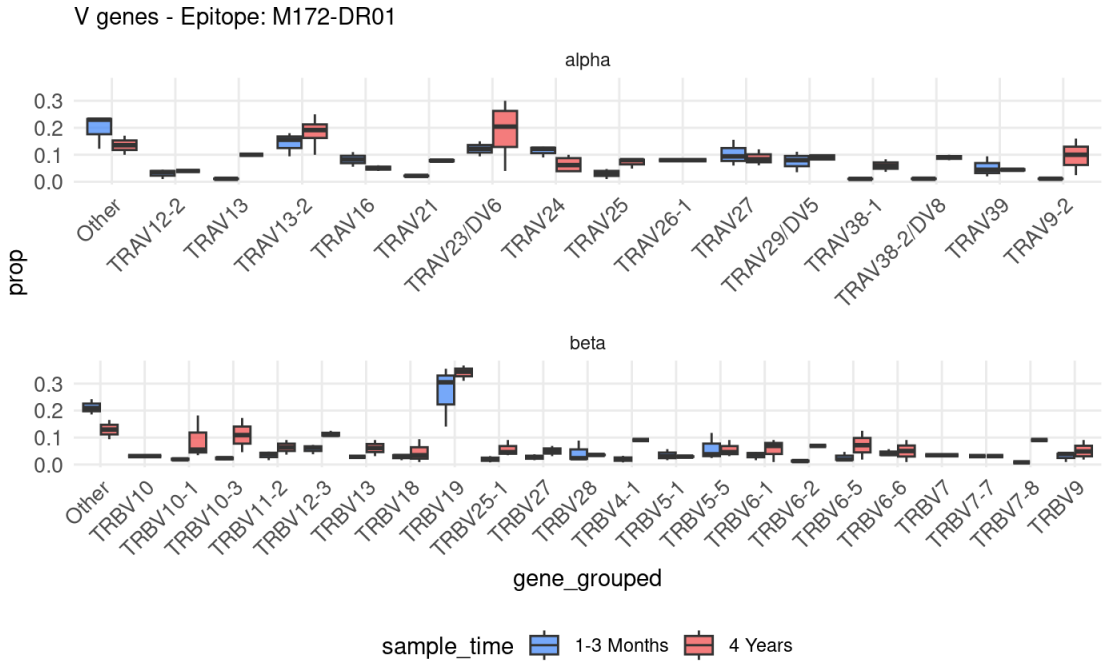


Figure E.11: TCR V gene usage temporal comparison for M24DR01-specific CD4⁺ T-cells showing relatively stable V gene repertoire composition over 4 years of follow-up.

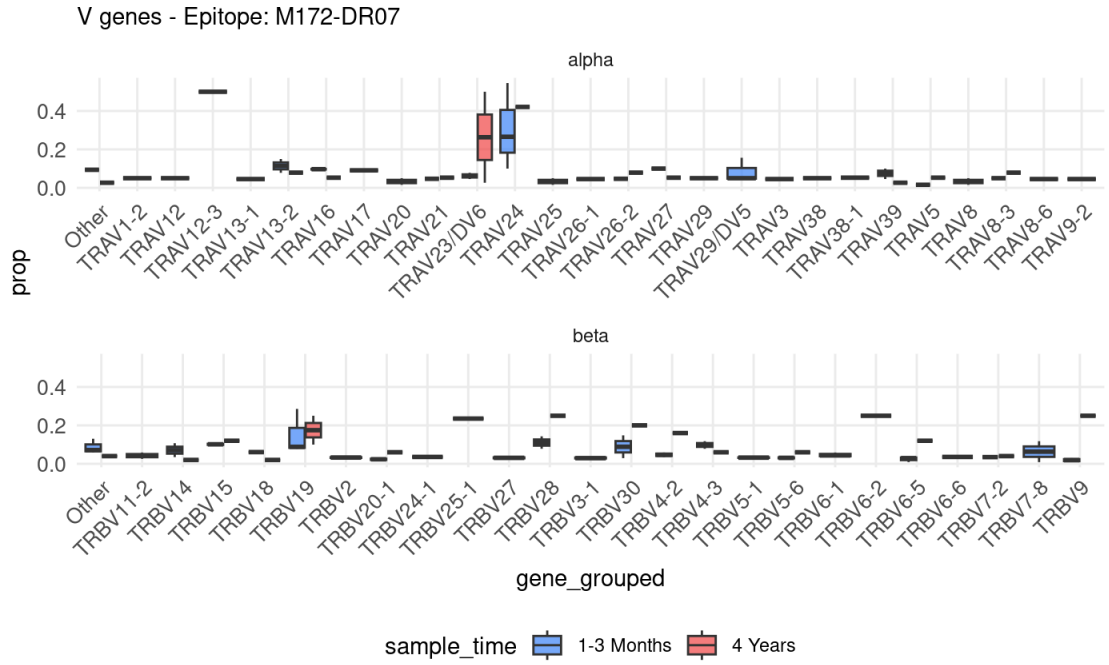


Figure E.12: Boxplots of TCR V-gene usage at early (1-3 months) and four year timepoint for M24DR07-specific CD4⁺ T-cells.

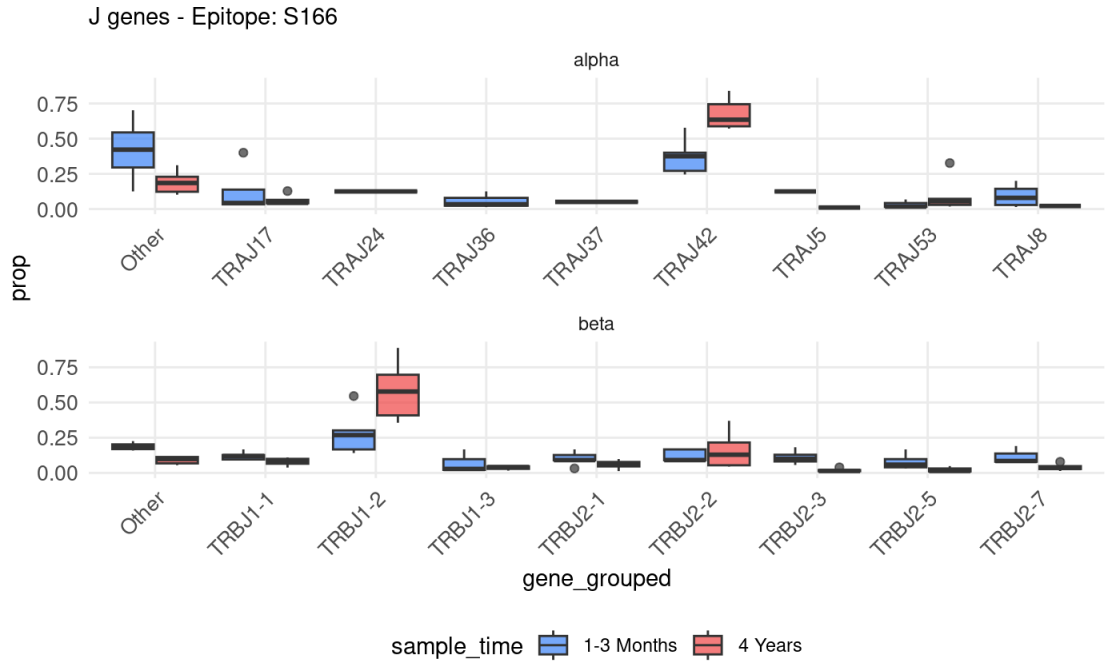


Figure E.13: TCR J gene usage temporal comparison for S166-specific CD4⁺ T-cells showing alpha and beta chain J gene proportions.

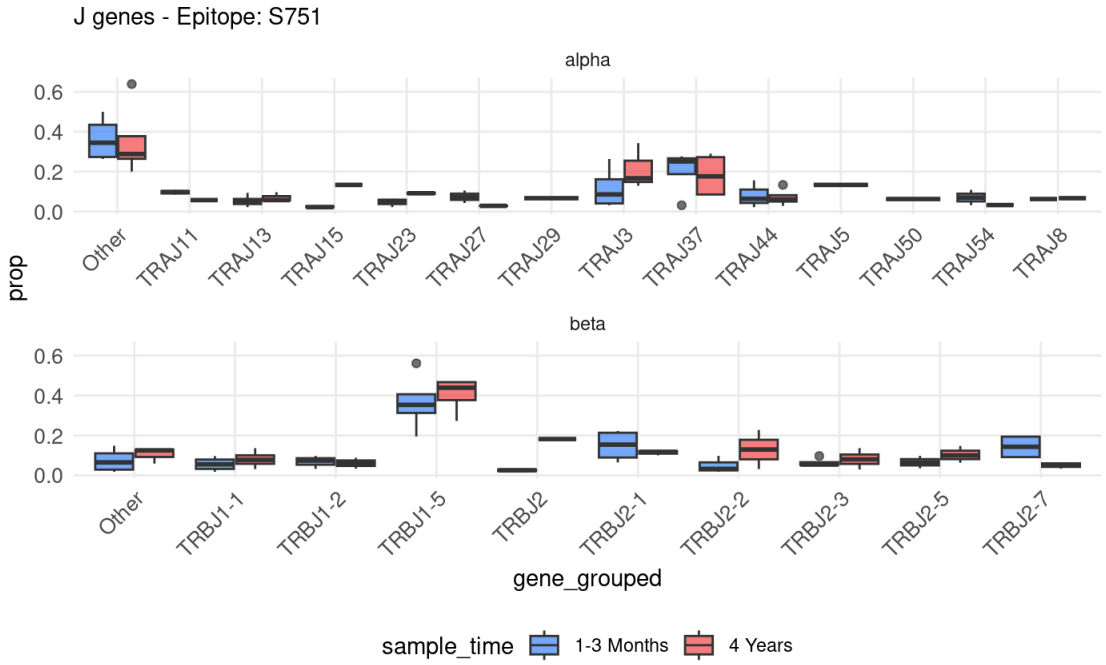


Figure E.14: TCR J gene usage temporal comparison for S751-specific CD4⁺ T-cells.

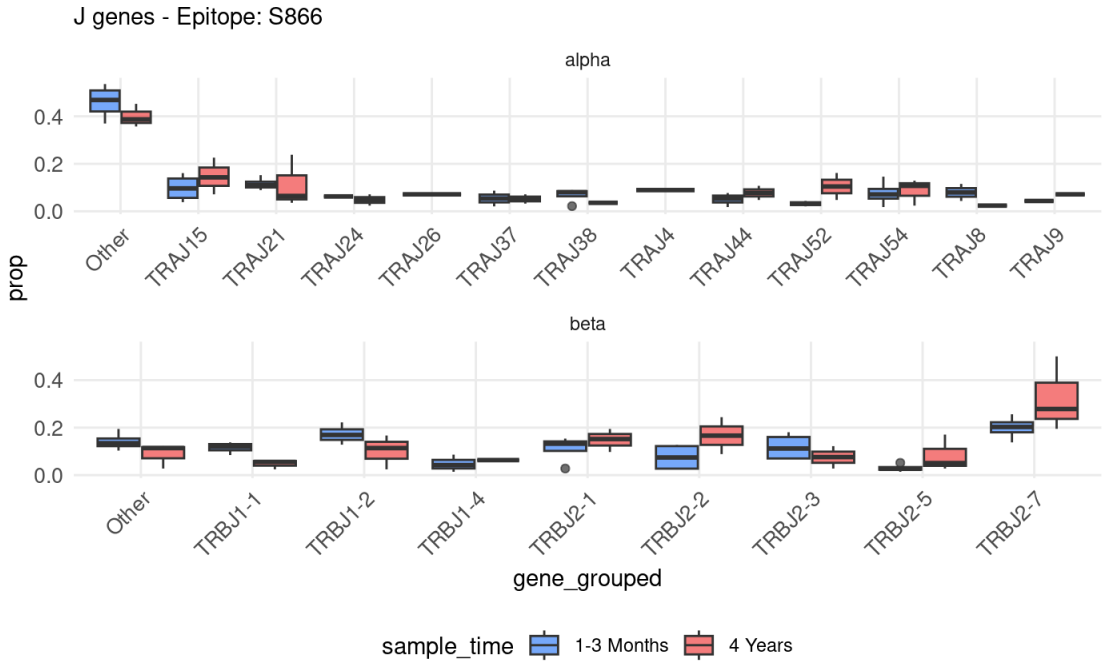


Figure E.15: Boxplots of TCR J-gene usage at early (1-3 months) and four year timepoint for S866-specific CD4⁺ T-cells.

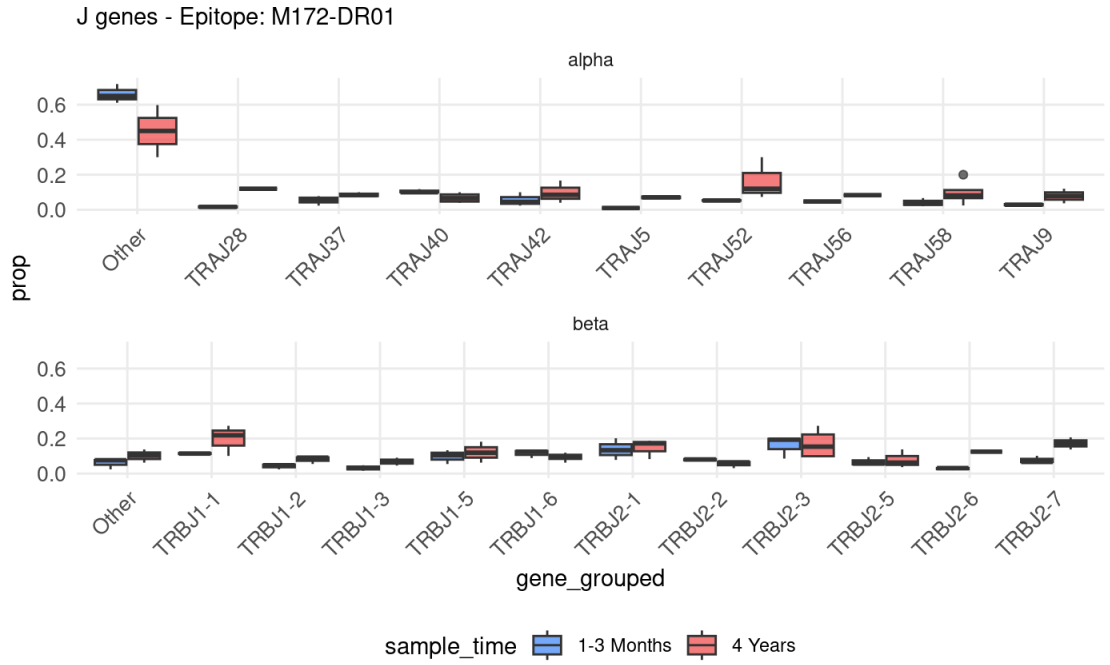


Figure E.16: TCR J gene usage temporal comparison for M24DR01-specific CD4⁺ T-cells showing maintained J gene repertoire structure over long-term follow-up.

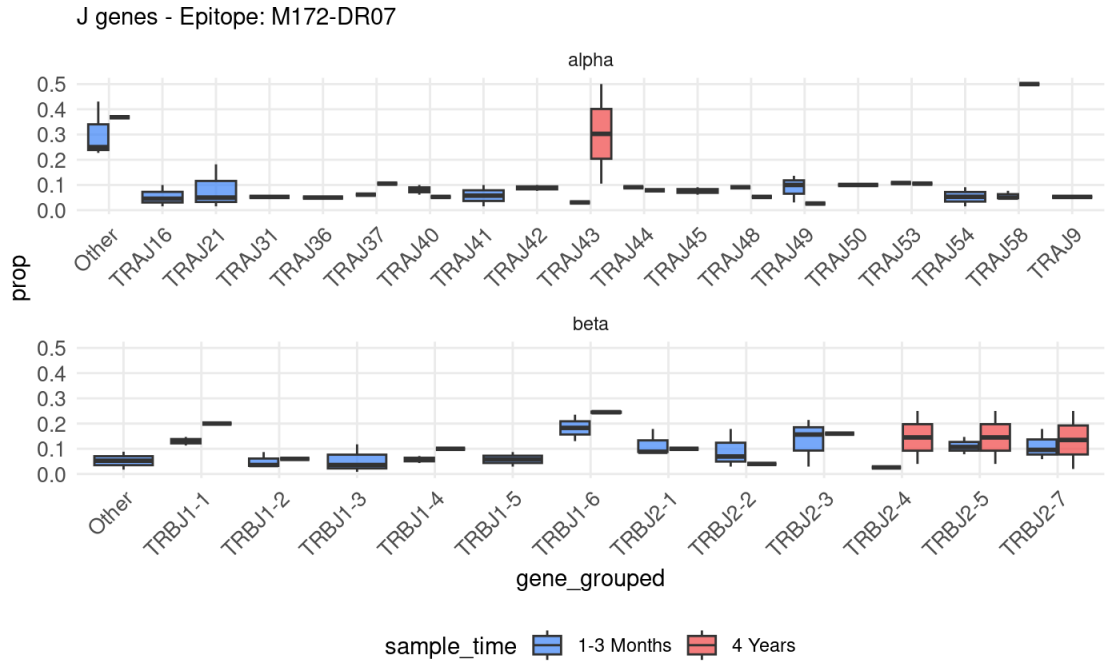


Figure E.17: TCR J gene usage temporal comparison for M24DR07-specific CD4⁺ T-cells with stable alpha and beta chain J gene usage frequencies between recovery and memory phases.

E.6 Gene Set Enrichment Analysis Tables

Table E.1: Table of all significantly differentially expressed Gene Ontologies (Gene Ontology Biological Process 2025 database) as well as 10 not meeting the threshold, under gene set enrichment analysis of genes more abundant in Non-Spike specific $CD4^+$ T-cells at 1-3 Months. 52 gene sets were significantly enriched after multiple testing corrections. Genes differentially expressed under DESeq2 with an adjusted p-value below 0.05 and a negative `log2FoldChange` in the comparison between Spike-specific $CD4^+$ pseudobulks at 1-3 months and Non-Spike specific $CD4^+$ pseudobulks at 1-3 months (i.e. more abundant in Non-Spike specific cells) were entered into Enrichr.

Term	Overlap	P-value	Adjusted P-value	Genes
Intracellular Protein Transport (GO:0006886)	98/326	6.30136e-11	2.69068e-07	MTCH1; IPO11; MTCH2; NUP107; CSE1L; PHB2; IPO5; IPO9; CHORDC1; FAM53C; KPNA6; ARFIP1; KPNA4; KPNA5; VPS36; NUP214; VPS33A; VPS33B; FAM91A1; NUP93; NUP98; VPS28; KPNB1; COPB2; COPA; RAB5C; PTTG1IP; VPS26B; TIMM50; POM121C; NUP85; GRPEL2; MYO6; NUP88; RAB6A; SNX6; CTSA; AKIRIN2; ACD; EGR2; NUP155; SURF4; ASPSCR1; RABL2B; TIMM44; TIMM10B; IFT27; RAB5A; ARF3; ARF4; C2CD5; ADAR; XPO1; XPO5; TMED3; TIMM17B; TMED7; APPL2; TMED5; TNPO3; APPL1; YWHAH; ATP6AP1; STRADB; TIMM23; AP3B1; TIMM22; VPS37A; ARFGAP3; COPZ1; ARL4A; PEX3; SLU7; ARF5; STX16; TIMM9; USO1; TIMM13; NAPG; HDAC6; STX10; GGA2; LMAN1; GGA3; SAMM50; SNX27; STX6; RAB28; PEX16; TRMT10B; MTX2; RANBP6; WDR11; SEC23IP; RAB14; SNX17; VPS41; RANGRF

Term	Overlap	P-value	Adjusted P-value	Genes
Mitochondrial Translation (GO:0032543)	41/94	1.44045e-10	3.07535e-07	MRPS17; MRPS16; GFM2; MRPS14; MRPL19; MRPS34; MRPL16; MRPS31; MRPL39; MRPL17; MRPL36; MRPL14; MTIF2; MRPL37; MRPL15; MRPS30; MTIF3; MRPL34; MRPL35; MRPL13; MRPL11; MRPL3; MRPL20; MTG1; RARS2; GADD45GIP1; PTCD3; TSFM; MRPS26; MRPS25; MRPS22; MRPS23; MRPS18A; MRPL45; QRSL1; MRPL23; MRPS6; MRPL24; MRPS18C; TUFM; OXA1L
Mitochondrial Gene Expression (GO:0140053)	41/100	1.41524e-09	2.01436e-06	MRPS17; MRPS16; FASTKD1; GFM2; MRPS14; MRPL19; MRPS34; MRPL16; MRPS31; MRPL17; MRPL39; MRPL14; MRPL36; MRPL15; MRPL37; MRPS30; FOXO3; MRPL34; MRPL13; MRPL35; MRPL11; TBRG4; MRPL20; MRPL3; MTG1; RARS2; GADD45GIP1; PTCD3; MRPS26; MRPS25; MRPS22; MRPS23; MRPS18A; MRPL23; MRPL45; QRSL1; MRPL24; MRPS6; MRPS18C; OXA1L; FASTKD5

Term	Overlap	P-value	Adjusted P-value	Genes
Protein Ubiquitination (GO:0016567)	102/369	4.38382e-09	4.49444e-06	UBE2D4; UBE3C; FBXO25; RNF19B; UBE2L6; UBE3B; FBXO22; HERC4; NPEPPS; HERC3; NUB1; CDC23; FBXO4; TRIM21; RNF111; HERC6; ATG3; RNF44; NHLRC3; WSB2; USP5; UBE2E3; LMO7; FBXO11; RC3H2; RNF41; RNF126; RNF125; KCTD13; TRIM13; UBE2V2; RBCK1; CUL9; CBFB; TNKS; UBA6; ZC4H2; UBR3; KLHL12; UBR1; UBE2J2; RNF213; UBR5; VPS11; HACE1; BFAR; UBAC1; UBE2W; UBOX5; RNF103; C6ORF89; CDC16; MKRN1; MDM2; FBXL3; BLMH; UBA1; RNF187; DET1; RNF10; RNF13; ENC1; ARIH1; SHPRH; PEX2; SIRT1; DDB2; PJA1; RNF144A; RNF167; TRAF6; BIRC6; RNF166; TRIM52; NEDD8; RNF6; CBL; RCHY1; RNF4; RNFT1; RNF113A; CAND1; OS9; SPOP; RNF20; TRIM41; RNF25; SYVN1; FBXL17; KLHL2; WWP1; KLHL3; WWP2; UBE2G2; KLHL7; KLHL8; RLIM; TRIP12; STUB1; RNF141; TRIM33; CUL4B

Term	Overlap	P-value	Adjusted P-value	Genes
RNA Processing (GO:0006396)	85/291	5.26281e-09	4.49444e-06	TCERG1; INTS12; POP7; FASTKD1; PDCD7; CELF1; NVL; SPPL2A; HNRNPR; ADAR; TSEN2; TSEN54; RRP9; SYNCRIP; FBL; RSRC1; MTPAP; METTL16; RBM6; UTP15; WDR36; PRMT1; DIS3; NCBP2; PIH1D1; PABPC4; METTL3; WDR74; DDX10; IMP4; PRKCA; TFIP11; ZCRB1; KIN; EBNA1BP2; PRPF6; PRPF3; XRN2; FASTKD5; RPP21; SRSF4; RP9; SNRNP25; ZNF830; SLU7; SRSF9; ZCCHC8; TSEN34; SF3B3; DDX27; DHX8; CSTF3; NOP2; SRSF1; WBP4; ISG20L2; DDX20; NOLC1; NOL8; PRPF8; DEDD2; TBRG4; RRP7A; EXOSC5; EMG1; EXOSC4; ZMAT5; RRP36; UTP20; LYAR; HNRNPA0; SF3A1; RNGTT; ADAT1; SRPK1; WDR55; SETX; NUDT21; SNRNP40; LAS1L; INTS4; ESF1; INTS7; RNPS1; TRMT61B

Term	Overlap	P-value	Adjusted P-value	Genes
mRNA Processing (GO:0006397)	67/226	1.14733e-07	8.16516e-05	TCERG1; DDX46; CELF1; HNRNPR; CASC3; DDX41; CWC27; SYNCRIP; RSRC1; RBMX2; SNRNP70; TXNL4A; METTL16; CMTR2; RBM6; AQR; NCBP2; METTL3; PRKCA; TFIP11; PLRG1; CRNKL1; CDC40; ERN1; KIN; DDX39A; PRPF6; PRPF3; PPIE; SNRNP27; PPIH; SRSF4; SNRNP200; SLU7; SRSF9; RBM22; ZCCHC8; SF3B3; DHX8; RNMT; CSTF2; SRSF1; WBP4; HTATSF1; PRPF8; TBRG4; U2AF1L4; RNF113A; ZMAT2; LEO1; RALY; RPRD1A; HNRNPA0; SF3A1; HNRNPA3; PPIL1; RNGTT; NUDT21; SNRNP40; PAF1; RBM42; RNPS1; PRPF31; PPIL3; HNRNPA1L2; FRA10AC1; TRMT61B
Nuclear Transport (GO:0051169)	20/38	1.69981e-07	0.000103688	NDC1; NUP214; AHCTF1; NUP107; SEH1L; ANP32A; NUP155; NUP210; NUP160; NUP93; RBM15B; XPO1; NUP85; RSRC1; HCLS1; NUP43; NUP88; NUP54; NUP98; NUP37

Term	Overlap	P-value	Adjusted P-value	Genes
Chromatin Remodeling (GO:0006338)	73/262	4.77049e-07	0.000254625	KDM5A; SETD2; KDM1A; EHMT1; RSF1; CTCF; CHD3; NUDT5; YY1; KAT5; ZNF83; VPS72; RUVBL2; UIMC1; RUVBL1; BCL7B; DPF2; PWWP2A; JARID2; NFRKB; BRD3; PRMT5; SS18; BRD1; PBRM1; SMARCC2; GPX1; PRMT1; PIH1D1; PRMT3; DDX56; VRK1; SIRT7; CHD1L; ARID1A; SIRT1; ARID1B; HCFC1; CFDP1; DPEP2; KDM7A; INO80C; ZNF274; HDAC3; SATB1; HDAC1; BAZ2B; HDAC6; IWS1; PRIMPOL; CHRAC1; ZNF827; UBR5; BRD9; DMAP1; SUDS3; BRD7; SMARCE1; KDM4C; MCRS1; DEK; SMARCA2; SMARCA4; MYO1E; MCM3AP; RAD50; TADA2A; YEATS2; C6ORF89; APLF; TRIP12; BCOR; EZH1
Organelle Disassembly (GO:1903008)	21/44	7.10211e-07	0.000336956	MCTS1; HBS1L; GFM2; FAF2; TRIP4; MTIF2; UFSP2; VRK1; MTIF3; ZFAND1; KLC1; HDAC6; RRP7A; KAT5; DDRGK1; ASCC2; DENR; PLIN2; ASCC3; TRIM21; MRRF

Term	Overlap	P-value	Adjusted P-value	Genes
Regulation of DNA Repair (GO:0006282)	43/130	8.55751e-07	0.000365406	INO80C; PARG; SF3B3; MEAF6; YY1; SMCHD1; PPP4C; KAT5; VPS72; RUVBL2; RUVBL1; BCL7B; UIMC1; DPF2; EP400; EPC1; EPC2; SUPT20H; DMAP1; BRD7; NFRKB; SMARCE1; PBRM1; SMARCC2; PARP1; TAF12; EYA3; MCRS1; TAF10; ACTL6A; CCDC117; SIRT7; DEK; ARID1A; SMARCA2; ARID1B; SIRT1; SUPT7L; SMARCA4; RAD52; UBE2V2; TADA1; TAF2
Protein Modification by Small Protein Conjugation (GO:0032446)	83/316	1.19402e-06	0.000463497	DET1; ZNF451; UBE2D4; FBXO25; UBE2L6; SMC6; HERC4; HERC3; NUB1; CDC23; ENC1; FBXO4; TRIM27; ARIH1; TRIM21; RNF111; HERC6; ATG3; SHPRH; RNF44; USP5; NSMCE2; LMO7; SIRT1; FBXO11; PIAS2; PJA1; SENP1; RNF144A; RNF126; EID3; NSMCE4A; KCTD13; UBE2V2; BIRC6; CUL9; TRIM52; UBA7; UBA6; URM1; DCUN1D4; UBR3; NEDD8; RNF6; UBR1; CBL; RCHY1; RNF113A; RNF213; CAND1; OS9; DDRGK1; VPS11; UBE2F; EGR2; HACE1; UBE2I; RNF25; SYVN1; FBXL17; KLHL2; WWP1; KLHL3; WWP2; UBAC1; COPS3; COPS6; RNF103; KLHL7; COPS2; C6ORF89; CDC16; KLHL8; MKRN1; RLIM; MDM2; FBXL3; UBA1; STUB1; COPS8; SAE1; TRIM33; CUL4B

Term	Overlap	P-value	Adjusted P-value	Genes
Nucleocytoplas Transport (GO:0006913)	22/50	2.17837e-06	0.000775138	NDC1; NUP214; AHCTF1; NUP107; SEH1L; UBE2I; NUP210; ANP32A; NUP155; MALT1; NUP160; NUP93; RBM15B; XPO1; RSRC1; NUP85; NUP54; NUP43; NUP88; NUP98; HTATIP2; NUP37
Purine-Containing Compound Biosynthetic Process (GO:0072522)	14/24	2.44041e-06	0.000801581	ADSL; PANK2; DCAKD; GMPS; PFAS; PPCDC; PRPSAP1; SLC4A7; PRPSAP2; MTHFD1; PPCS; HPRT1; GART; ADA
Modification-Dependent Protein Catabolic Process (GO:0019941)	59/208	3.10443e-06	0.00094685	UBE2D4; UBE3C; RNF13; USP33; RNF19B; UBE2L6; UBE3B; FBXO22; HERC4; HERC3; NUB1; CDC23; FBXO4; ARIH1; FBXO8; RNF111; HERC6; UBL7; FBXW4; FBXO11; RC3H2; RNF144A; RNF167; RNF126; RNF125; RNF166; UBA6; UBR3; NEDD8; RNF6; UBE2J2; CBL; RCHY1; RNF5; BAG6; ATXN3; RNF213; OS9; RFFL; UBQLN2; RNF20; HACE1; ANKIB1; UBE2I; SMURF2; RNF25; SIAH2; SYVN1; KLHL3; BFAR; UBE2G2; NTAN1; COPS3; KLHL8; RLIM; MDM2; TRIP12; UBA1; STUB1

Term	Overlap	P-value	Adjusted P-value	Genes
mRNA Splicing, via Spliceosome (GO:0000398)	59/211	5.08557e-06	0.00144769	DDX46; HNRNPR; CASC3; DDX41; CWC27; SYNCRIP; RSRC1; RBMX2; SNRNP70; TXNL4A; RBM6; PRMT5; RBM17; NCBP2; METTL3; TFIP11; PLRG1; CRNKL1; CDC40; DDX39A; PRPF6; HNRNPUL2; PRPF3; PPIE; SNRNP27; SRFBP1; PPIH; SRSF4; SNRNP200; SLU7; SRSF9; RBM22; ZCCHC8; SF3B3; DHX8; SRSF1; WBP4; DDX20; HTATSF1; PRPF8; U2AF1L4; RNF113A; ZMAT2; PUF60; RALY; DCPS; ZC3H14; SF3A1; HNRNPA3; PPIL1; DNAJC17; SRPK1; SNRNP40; RBM42; RNPS1; PRPF31; PPIL3; HNRNPA1L2; FRA10AC1
Protein Polyubiquitination (GO:0000209)	62/228	7.90289e-06	0.00210908	UBE2D4; UBE3C; UBE2L6; UBE3B; FBXO22; NPEPPS; CDC23; FBXO4; TRIM26; TRIM27; TRIM21; TRIM22; SHPRH; NHLRC3; WSB2; UBE2E3; RC3H2; DDB2; RNF41; RNF144A; RNF167; RNF126; RNF125; TRAF6; UBE2V2; ANAPC4; RBCK1; RNF166; ANAPC15; CBFB; TNKS; RNF6; UBE2J2; NT5C2; RNF4; RNF5; ANAPC10; RNF213; RNF216; UBR5; SPOP; RFFL; RNF20; HACE1; RNF25; FBXO38; SYVN1; FBXL17; KLHL3; WWP2; BFAR; UBE2G2; UBE2W; UBOX5; CDC16; KBTBD2; MKRN1; MDM2; TRIP12; BLMH; STUB1; RNF187

Term	Overlap	P-value	Adjusted P-value	Genes
Chromatin Organization (GO:0006325)	80/323	1.95226e-05	0.00490363	KDM5A; SETD7; EHMT1; RSF1; NUCKS1; SMC6; CTCF; CHD3; NUDT5; ING4; YY1; HMGN4; VPS72; RUVBL2; RUVBL1; BCL7B; DPF2; EPC1; PWWP2A; BANF1; HMG20A; JARID2; NFRKB; BRD3; PRMT5; SS18; BRD1; PBRM1; SMARCC2; PRMT1; PIH1D1; NSMCE2; PRMT3; DDX56; VRK1; CHD1L; ARID1A; SAFB; SIRT1; ARID1B; HCFC1; CFDP1; EID3; NSMCE4A; KAT6B; DPEP2; KDM7A; INO80C; ZNF274; HDAC3; SATB1; HDAC1; BAZ2B; SPTY2D1; HDAC8; IWS1; ATAD2B; CHRAC1; ZNF827; BRD9; DMAP1; SUDS3; BRD7; SMARCE1; KDM4C; MCRS1; SETDB2; ATAD2; DEK; SMARCA2; SMARCA4; SRPK1; MYO1E; TADA2A; YEATS2; C6ORF89; BCOR; EZH1; EZH2; FAM50A
Fatty Acid Catabolic Process (GO:0009062)	23/60	2.08961e-05	0.00495703	ACOT8; ACBD5; ECHS1; ABHD3; ECI2; PEX2; CROT; PEX13; HADHB; HADHA; BDH2; SLC25A17; CPT2; ACOX1; PCCB; NUDT19; ECHDC1; ACADM; ACAA1; LPIN1; LPIN2; PPARD; DECR1

Term	Overlap	P-value	Adjusted P-value	Genes
Ribosome Biogenesis (GO:0042254)	46/161	2.9601e-05	0.00665244	DDX27; RPS19BP1; NVL; NIP7; SPPL2A; NOP2; HEATR1; NOLC1; NOL8; MALSU1; RPF2; SDAD1; WDR46; TMA16; RRP9; RRP7A; FBL; EMG1; EXOSC5; XPO1; EXOSC4; RRP36; UTP20; ZNF622; LYAR; METTL17; ABCF1; UTP15; WDR36; UTP6; DIS3; KRR1; PIH1D1; WDR74; DDX10; IMP4; WDR12; UTP18; WDR55; EBNA1BP2; LAS1L; XRN2; CINP; ESF1; DHX29; CUL4B
RNA Splicing, via Transesterification Reactions With Bulged Adenosine as Nucleophile (GO:0000377)	48/171	3.32103e-05	0.0070904	ZCCHC8; SF3B3; DHX8; DDX46; SRSF1; WBP4; HNRNPR; CASC3; DDX41; CWC27; HTATSF1; PRPF8; U2AF1L4; SYNCRIP; RNF113A; ZMAT2; RBMX2; RSRC1; SNRNP70; RALY; TXNL4A; RBM6; HNRNPA3; PPIL1; SF3A1; NCBP2; METTL3; TFIP11; PLRG1; CRNKL1; CDC40; SNRNP40; DDX39A; PRPF6; PRPF3; PPIE; SNRNP27; RBM42; PPIH; SRSF4; RNPS1; PRPF31; SNRNP200; PPIL3; HNRNPA1L2; SLU7; FRA10AC1; RBM22
Ribosome Disassembly (GO:0032790)	17/40	5.32738e-05	0.0103609	MCTS1; HBS1L; GFM2; RNF25; ELAC1; TRIP4; MTIF2; UFSP2; MTIF3; RCHY1; DDRGK1; ASCC2; DENR; ASCC3; EIF4E2; MRRF; ANKZF1

Term	Overlap	P-value	Adjusted P-value	Genes
T Cell Receptor Signaling Pathway (GO:0050852)	29/88	5.33815e-05	0.0103609	DENND1B; ITK; STOML2; TXK; CLSTN1; MALT1; IKBKB; FCHO1; INPP5D; SLC39A6; CTLA4; CSK; EIF2B1; EIF2B5; EIF2B3; RIPK2; PDE4D; RFTN1; THEMIS; BTN3A3; RC3H2; LCK; TRAF6; CD28; PRKD2; RBCK1; CD247; SKAP1; HLA-DRB1
Fatty Acid Beta-Oxidation (GO:0006635)	19/48	6.42091e-05	0.0119206	ACADVL; ACOT8; ECHS1; ECI2; ETFA; PEX2; CROT; ETFB; HADHB; HADHA; BDH2; SLC25A17; CPT2; ACOX1; ECHDC1; ACADM; ACAA1; DECR1; PPARD
rRNA Metabolic Process (GO:0016072)	31/98	7.24856e-05	0.0128964	DDX27; NVL; SPPL2A; NOP2; HEATR1; NOLC1; NOL8; RRP9; DEDD2; RRP7A; FBL; EMG1; EXOSC5; EXOSC4; RRP36; UTP20; LYAR; UTP15; WDR36; DIS3; PIH1D1; WDR74; DDX10; IMP4; SIRT7; WDR55; EBNA1BP2; XRN1; LAS1L; XRN2; ESF1
Positive Regulation of Double-Strand Break Repair (GO:2000781)	30/94	7.83845e-05	0.0133577	MEAF6; SMCHD1; KAT5; VPS72; RUVBL2; BCL7B; RUVBL1; DPF2; EP400; EPC1; EPC2; DMAP1; BRD7; SMARCE1; PBRM1; CREBBP; SMARCC2; HELQ; PARP1; PRMT1; MGMT; ARID1A; ARID1B; SIRT1; SMARCA2; SMARCA4; MAD2L2; RNF126; RAD50; UBE2V2

Term	Overlap	P-value	Adjusted P-value	Genes
Endoplasmic Reticulum to Golgi Vesicle-Mediated Transport (GO:0006888)	34/112	8.41644e-05	0.0133577	COPB2; ARF4; COPA; ATL3; PEF1; GOSR2; GOSR1; USO1; KLHL12; IER3IP1; LMAN1; SPAST; LMAN2; VTI1A; TMED3; BCAP29; TRAPPC13; TMED7; TMED5; RAB2A; TRAPPC2; TRAPPC3; SURF4; NRBP1; YIF1A; YIF1B; TRAPPC6A; VAPB; ZW10; RANGRF; PREB; GOPC; ERGIC3; SEC24C
Mitochondrial Respiratory Chain Complex Assembly (GO:0033108)	29/90	8.44632e-05	0.0133577	COX19; NDUFB3; COX7A2; HIGD1A; FOXRED1; PET117; TMEM70; TMEM242; TMEM223; SAMM50; COA5; ATPAF2; SURF1; UQCC1; BCS1L; OXA1L; LYRM2; NDUFS7; SDHAF2; NDUFAF6; NDUFS4; NDUFAF5; NDUFS3; TMEM126A; NDUFS2; NDUFS1; NDUFAF1; FXN; CEP89
Protein Localization to Nucleus (GO:0034504)	37/127	0.000109637	0.0167196	IPO11; NUP107; CSE1L; HEATR1; XPA; ARL2; NOL8; ADAR; PTTG1IP; RPF2; PHB2; IPO5; RRP7A; WRN; IPO9; POM121C; NUP85; FAM53C; KPNA6; KPNA4; NUP88; KPNA5; APPL2; TNPO3; APPL1; NUP214; AKIRIN2; MORC3; NUP155; RANBP6; SUPT7L; NUP93; ARL2BP; MDM2; PAF1; NUP98; KPNB1
Regulation of Microtubule Nucleation (GO:0010968)	10/18	0.000124864	0.0179155	MECP2; HAUS8; NME7; HAUS4; HAUS6; HAUS5; CKAP5; HAUS2; GIT1; HAUS1

Term	Overlap	P-value	Adjusted P-value	Genes
NADH Dehydrogenase Complex Assembly (GO:0010257)	15/35	0.000130066	0.0179155	NDUFB3; BCS1L; FOXRED1; TMEM70; OXA1L; LYRM2; NDUFAF6; NDUFS7; NDUFAF5; NDUFS4; NDUFS3; TMEM126A; NDUFS2; NDUFS1; NDUFAF1
Mitochondrial Respiratory Chain Complex I Assembly (GO:0032981)	15/35	0.000130066	0.0179155	NDUFB3; BCS1L; FOXRED1; TMEM70; OXA1L; LYRM2; NDUFAF6; NDUFS7; NDUFAF5; NDUFS4; NDUFS3; TMEM126A; NDUFS2; NDUFS1; NDUFAF1
rRNA Processing (GO:0006364)	33/110	0.000136674	0.0179564	DDX27; NVL; SPPL2A; NOP2; MAK16; NOLC1; NOL8; NOL9; TFB1M; RRP9; RRP7A; FBL; EMG1; EXOSC5; EXOSC4; ERI1; RRP36; UTP20; LYAR; UTP15; WDR36; DIS3; PIH1D1; WDR74; DDX10; IMP4; WDR55; EBNA1BP2; LAS1L; XRN2; C1D; ESF1; SRFBP1
Import Into Nucleus (GO:0051170)	25/75	0.000138773	0.0179564	IPO11; NUP107; CSE1L; PTTG1IP; PHB2; IPO5; IPO9; POM121C; NUP85; FAM53C; KPNA6; KPNA4; NUP88; KPNA5; APPL2; TNPO3; APPL1; HTATIP2; NUP214; AKIRIN2; NUP155; RANBP6; NUP93; NUP98; KPNB1

Term	Overlap	P-value	Adjusted P-value	Genes
Macroautophagy (GO:0016236)	44/162	0.000158527	0.0199091	MTMR3; TEX264; CALCOCO2; WDR45B; EI24; VPS4A; AUP1; PHB2; MFSD8; TBK1; RNF213; DDRGK1; LAMP2; VTI1A; PIP4K2A; LGALS8; VPS36; SNAP29; SPG11; VPS39; UBXN6; RAB2A; ATG3; ACBD5; PDCD6IP; VPS33A; LRBA; VPS33B; LIX1L; PEX2; VPS37A; KLHL3; WIPI2; ATG13; ATG16L1; SNX14; NBR1; VPS41; PLAA; PGAM5; PIK3C3; VPS28; TECPR1; ATG2B
Protein Import Into Nucleus (GO:0006606)	26/80	0.000164559	0.0200762	IPO11; NUP107; CSE1L; PTTG1IP; PHB2; IPO5; IPO9; POM121C; NUP85; FAM53C; KPNA6; KPNA4; NUP88; KPNA5; APPL2; TNPO3; APPL1; NUP214; AKIRIN2; NUP155; MMAB; RANBP6; NUP93; NUP54; NUP98; KPNB1
Establishment of Protein Localization to Organelle (GO:0072594)	28/90	0.000217388	0.0257846	IPO11; RALA; NUP107; CSE1L; PTTG1IP; PHB2; IPO5; IPO9; POM121C; NUP85; FAM53C; KPNA6; KPNA4; NUP88; KPNA5; APPL2; TNPO3; APPL1; NUP214; AKIRIN2; NUP155; GOLPH3L; RANBP6; RAB11A; NUP93; HOOK2; NUP98; KPNB1

Term	Overlap	P-value	Adjusted P-value	Genes
Regulation of Double-Strand Break Repair (GO:2000779)	28/91	0.000267152	0.0308308	SETD2; KDM1A; MEAF6; PPP4C; KAT5; VPS72; RUVBL2; BCL7B; RUVBL1; DPF2; EP400; EPC1; EPC2; DMAP1; BRD7; SMARCE1; PBRM1; SMARCC2; MGMT; DEK; ARID1A; ARID1B; SIRT1; SMARCA2; SMARCA4; RAD50; UBE2V2; TERF2IP
Organelle Fission (GO:0048285)	11/23	0.000324809	0.0351127	MED12; FIS1; MTFP1; ACOX1; PEX11B; MTFR1L; MTFR2; TMCC1; MFF; DNM1L; SLC25A46
Translesion Synthesis (GO:0019985)	11/23	0.000324809	0.0351127	MAD2L2; POLDIP2; POLD1; POLD2; REV1; REV3L; SPRTN; POLI; POLK; RCHY1; PRIMPOL
Protein Transport (GO:0015031)	72/308	0.000328924	0.0351127	ARF3; ARF4; ARHGAP1; LAMP2; TMED3; TIMM17B; ARFIP1; TMED7; TMED5; YWHAH; NDC1; TANGO6; VPS33A; VPS33B; GOLPH3L; TIMM23; AP3B1; FAM91A1; ARFGAP3; SCAMP3; COPZ1; SCAMP2; PDIA4; RUFY1; PREPL; ARL4A; C1D; PREB; SLU7; VAMP5; ARF5; MACF1; COPB2; COPA; RAB5C; STX16; USO1; ASAP1; VPS26B; NAPG; HDAC6; STX10; TIMM50; GGA2; SNX4; GGA3; SNX27; KIF3B; KIF3A; GRPEL2; MYO6; STX6; RAB28; RAB6A; SNX6; BBS2; CTSA; GSDMD; ACD; SURF4; CMTM6; ASPSCR1; RABL2B; TIMM44; RAB11A; WDR11; SEC23IP; RAB14; SNX17; PPT1; IFT27; RAB5A

Term	Overlap	P-value	Adjusted P-value	Genes
Regulation of Macroautophagy (GO:0016241)	31/106	0.000353827	0.0362296	ATP6V1A; VPS26B; HDAC6; MAPK8; TBK1; CAPNS1; SESN3; SESN1; CASP3; ATP6V0A2; ATP6V1H; ATP6V1E1; DNM1L; ATP6V0E2; ATP6V1C1; SNX5; SNX6; UBQLN2; SH3GLB1; ATP6V0B; ATP6V0E1; CSNK2A2; PRKCA; TSC1; SIRT1; ERN1; TRIM13; PIK3C3; ATP6V0D1; TECPR1; EXOC1
Translation (GO:0006412)	57/232	0.000356356	0.0362296	MRPS17; HBS1L; MRPS16; GFM2; MRPS14; MRPL39; MRPL36; MRPL37; MRPL34; APEH; MRPL35; YARS2; MRPL3; EEF2K; MTG1; RARS2; TSFM; MRPS26; MRPS25; MRPS22; PABPC4; RPL22; MRPS23; DHPS; MRPS18A; MRPL45; QRSL1; MRPS6; MRPS18C; TUFM; OXA1L; EEF1D; AURKAIP1; MRRF; EIF5A2; MRPL19; MRPS34; SRBD1; MRPL16; MRPS31; MRPL17; MRPL14; MRPL15; MRPS30; MRPL13; GSPT1; MRPL11; MRPL20; RPS2; DTD1; ABCF1; GADD45GIP1; PTCD3; MRPL23; MRPL24; RPS26; FARSB
Inner Mitochondrial Membrane Organization (GO:0007007)	15/38	0.000386634	0.0383937	TRMT10B; MTX2; TIMM9; TIMM13; IMMT; TIMM22; BCS1L; MED12; TIMM10B; APOO; OXA1L; CHCHD3; SAMM50; TMEM126A; SLC25A46

Term	Overlap	P-value	Adjusted P-value	Genes
Regulation of DNA Metabolic Process (GO:0051052)	33/116	0.000405415	0.0388091	ANKRD17; INO80C; PARG; SF3B3; MEAF6; NUCKS1; JADE2; SMC6; HDAC8; YY1; ORC5; ING4; HELB; CHRAC1; RUVBL2; RUVBL1; UIMC1; MPV17; SUPT20H; NFRKB; TAF12; MCRS1; TAF10; NSMCE2; ACTL6A; SIRT7; SUPT7L; EID3; NSMCE4A; UBE2V2; TADA1; TERF2IP; TAF2
DNA Damage Response (GO:0006974)	93/421	0.000408995	0.0388091	DYRK2; MCM7; KDM1A; SETD7; SMC6; ALKBH7; YY1; RASSF1; HELB; KAT5; ZNF83; HMCES; UIMC1; CTLA4; DGCR8; POLI; POLK; IKBKE; RFC5; SHPRH; RFC4; PARP4; PARP1; PRMT1; CSNK2A2; METTL3; HUS1; VRK1; SIRT7; CHD1L; VRK2; CSNK1E; VRK3; REXO2; SMC1A; SIRT1; WDR76; NRBP1; DDB2; MSH6; KIN; EID3; NSMCE4A; RAD51C; MSH2; RRM2B; IRF3; PPP1R15A; HDAC3; UBA7; UBA6; BOD1L1; XPA; XPC; FOXO1; CDKN2AIP; ZBTB40; WRN; OARD1; POLD1; STK38; UBR5; EP400; LRIG1; BID; MICA; MUTYH; MOAP1; RRM1; CRADD; MGMT; MAPK14; HINFP; FANCF; PPP1CA; SETX; MAD2L2; FAM111A; RAD50; ERCC3; TTC5; TDP1; ERCC1; TAOK2; APLF; SPRTN; NABP2; NABP1; BAX; GRB2; TRIP12; UBA1; CUL4B

Term	Overlap	P-value	Adjusted P-value	Genes
RNA Export From Nucleus (GO:0006405)	22/67	0.000433406	0.0402313	NUP214; NUP107; NUP155; NCBP2; CASC3; NXT2; SMG7; THOC7; NUP160; THOC6; DDX19B; RBM15B; MCM3AP; XPO1; DDX39A; POM121C; XPOT; NUP85; XPO5; NUP88; NUP98; RBM33
Ubiquitin-Dependent Protein Catabolic Process (GO:0006511)	99/455	0.000473843	0.0424862	UBE2D4; UBE3C; RNF19B; UBE2L6; UBE3B; FBXO22; HERC4; NUB1; HERC3; CDC23; KAT5; FBXO4; PSMD1; TRIM26; FBXO8; RNF111; HERC6; FBXW4; NHLRC3; FBXO11; RC3H2; RNF126; RNF125; KCTD13; TRIM13; PLAA; RBCK1; VPS28; UBA6; UBR3; KLHL12; UBR1; UBE2J2; BAG6; ATXN3; RNF213; RNF216; UBR5; KLHDC2; HACE1; UBE2I; SMURF2; SIAH2; BFAR; NTAN1; COPS3; UBE2W; MDM2; UBA1; RNF187; UBXN1; RNF13; FAF1; USP33; ZFAND2B; ARIH1; UBL7; RAD23A; VPS37A; SIRT1; RNF144A; RNF167; RNF166; VPS4A; NEDD8; RNF6; RCHY1; CBL; RNF4; RNF5; PSMB10; PSMB2; OS9; HECTD3; SPOP; TBL1X; DCAF11; RFFL; UBQLN2; RNF20; ANKIB1; CSNK1A1; RNF25; SYVN1; FBXL17; WWP1; KLHL2; WWP2; KLHL3; UBE2G2; KBTBD3; KLHL6; KLHL7; GID8; KBTBD2; KLHL8; RLIM; STUB1; TRIP12

Term	Overlap	P-value	Adjusted P-value	Genes
Nuclear Pore Organization (GO:0006999)	8/14	0.000477597	0.0424862	NDC1; FXR1; NUP93; NUP107; SEH1L; NUP54; NUP98; RTN4
Gene Expression (GO:0010467)	85/381	0.000511742	0.0441734	MRPS17; ISCA2; HBS1L; NUP107; MRPS16; MRPS14; MRPL36; ADAR; MRPL37; CASC3; MRPL34; SMG7; NXT2; MRPL35; YARS2; CISD3; IKBKB; SYNCRIP; MRPL3; XPO1; CASP3; CASP1; EPC1; KPNA4; AGFG1; RBM6; MEF2A; NUP214; DIS3; NCBP2; PABPC4; RPL22; DHPS; MRPS18A; TFIP11; MRPS6; THOC7; MRPS18C; THOC6; DDX19B; DDX39A; XRN2; AURKAIP1; ZNF830; MRRF; ZCCHC8; DHX8; MRPS34; SRBD1; ISG20L2; DDX20; CREBL2; MRPL13; GSPT1; NUP160; CIAO1; DEDD2; MRPL11; ZMPSTE24; EXOSC5; RBM15B; EXOSC4; TAF1B; NUP85; TSPAN5; TASP1; NUP88; RPS2; DTD1; ABCF1; MBTPS1; RNGTT; MRPL23; MRPL24; UBAC1; RPS26; SETX; CCS; MCM3AP; SNRNP40; FUBP3; MMS19; RNPS1; METAP1; FARSB
Protein Localization to Ciliary Membrane (GO:1903441)	5/6	0.000517253	0.0441734	RABEP1; ARL13B; ARL3; WDR19; BBIP1

Term	Overlap	P-value	Adjusted P-value	Genes
Fatty Acid Oxidation (GO:0019395)	18/51	0.000531129	0.0444691	ECHS1; ECI2; PEX2; CROT; ADIPOR1; PEX13; ADH5; HADHB; HADHA; BDH2; SLC25A17; CPT2; ACOX1; ECHDC1; ACADM; ACAA1; PPARD; DECR1
Regulation of RNA Splicing (GO:0043484)	30/104	0.000563666	0.0462857	SF3B3; CELF1; SRSF1; PIK3R1; POLR2A; SNRNP70; HABP4; METTL4; CCNL2; SUPT20H; RBM12; METTL16; TIA1; TAF12; TAF10; DYRK1A; RRP1B; SUPT7L; CLK3; SRPK1; PTBP3; ERN1; RPS26; SETX; ZNF638; TADA1; GRSF1; RBM23; RBM22; FAM50A
Lysosome Organization (GO:0007040)	16/44	0.000733447	0.0563992	MBTPS1; ACD; SNAPIN; AKTIP; VPS33A; HPS1; CORO1A; CLN5; TMEM165; MFSD8; GNPTAB; HOOK2; RAB14; TMEM106B; LAMTOR1; SPG11
Response to Amino Acid Starvation (GO:1990928)	16/44	0.000733447	0.0563992	SH3GLB1; SEH1L; PRKCH; WDR59; MIOS; NPRL2; EIF2AK3; EIF2AK2; PEX2; ITFG2; ATXN3; MAPK8; IMPACT; SESN3; SESN1; MAP3K5
mRNA Metabolic Process (GO:0016071)	29/101	0.00074974	0.0563992	TCERG1; CELF1; SRSF1; HNRNPR; PRPF8; TBRG4; SYNCRIP; EXOSC5; RNASEH1; METTL16; DCPS; HNRNPA0; SF3A1; PNPT1; YTHDF2; NCBP2; METTL3; PRKCA; NUDT16L1; HNRNPAB; SND1; ERN1; KIN; NUDT21; IRF3; PRPF3; SRSF4; TRMT61B; SRSF9
Melanosome Localization (GO:0032400)	9/18	0.000766079	0.0563992	BBS2; BLOC1S5; MKKS; BLOC1S6; VPS33A; VPS33B; BLOC1S3; BBS7; RAB11A

Term	Overlap	P-value	Adjusted P-value	Genes
Negative Regulation of Endoplasmic Reticulum Unfolded Protein Response (GO:1900102)	9/18	0.000766079	0.0563992	PPP1R15A; PPP1R15B; PTPN1; ATF6B; WFS1; DDRGK1; CREBRF; NCK2; BFAR
Nucleotide Biosynthetic Process (GO:0009165)	9/18	0.000766079	0.0563992	PRPSAP1; SLC4A7; RRM1; PRPSAP2; MTHFD1; RRM2B; HPRT1; PFAS; GART
Positive Regulation of T Cell Activation (GO:0050870)	31/111	0.000839926	0.0587074	IL21; CD81; DOCK8; SIRPG; SLC7A1; PYCARD; HLA-DMA; FCHO1; NCK2; FLOT2; JAK2; BRD7; HLA-DPA1; SMARCE1; HLA-DRB5; PBRM1; SMARCC2; FBXO38; ACTL6A; ARID1A; SMARCA2; ARID1B; SMARCA4; CD40LG; CD6; IL23A; LCK; CD28; HLA-DRA; PRKCQ; HLA-DRB1

Term	Overlap	P-value	Adjusted P-value	Genes
DNA Metabolic Process (GO:0006259)	69/302	0.000841065	0.0587074	MCM7; HMGB3; YY1; SYNCRIP; HELB; ZNF83; RUVBL2; RUVBL1; FBXO4; POLI; POLK; NFRKB; RFC5; PARP4; RFC4; PARP1; ACTL6A; REV1; HUS1; CSNK1E; CHD1L; SMC1A; DDB2; MSH6; KIN; EID3; RAD51C; NSMCE4A; MSH2; RRM2B; UBE2V2; MCM3; TERF2IP; DNASE2; MCM5; DNASE1; MCM6; DCTPP1; INO80C; SET; PARG; XPA; XPC; ORC5; PURA; WRN; ZNF827; POLD1; POLD2; EP400; LRIG1; MUTYH; ACD; RRM1; MGMT; MCRS1; DONSON; NOC3L; HNRNPAB; HINFP; RAD52; FAM111A; RAD50; ERCC3; TDP1; ERCC1; APLF; NABP2; NABP1

Term	Overlap	P-value	Adjusted P-value	Genes
DNA Repair (GO:0006281)	69/302	0.000841065	0.0587074	SETD2; MPG; NUCKS1; KAT5; ZNF83; HMCES; UIMC1; POLI; POLK; RFC5; PARP4; RFC4; PARP1; AKTIP; CSNK2A2; HUS1; RAD23A; CSNK1E; COMMD1; CHD1L; PARP9; SMC1A; SIRT1; DDB2; WDR33; MSH6; EID3; RAD51C; NSMCE4A; MSH2; RRM2B; UBE2V2; RHNO1; RNASEH2C; RNASEH2B; TEX264; XPA; XPC; PRIMPOL; ATXN3; WRN; POLD1; EP400; LRIG1; ASCC2; ASCC3; MUTYH; RRM1; ERCC6L2; MGMT; EYA3; REC8; HINFP; FANCF; SETX; MAD2L2; RAD52; SMUG1; FAM111A; RAD50; ERCC3; TDP2; TDP1; ERCC1; APLF; SPRTN; NABP2; NABP1; CUL4B
Error-Prone Translesion Synthesis (GO:0042276)	6/9	0.000853464	0.0587074	POLD2; REV1; REV3L; POLI; POLK; PRIMPOL

Table E.2: Table of the top 10 most significantly differentially expressed Gene Ontologies (Gene Ontology Biological Process 2025 database) under gene set enrichment analysis of genes more abundant in Spike specific $CD4^+$ T-cells at 1-3 Months. No gene sets reach significance after adjustment for multiple testing. Genes differentially expressed under DESeq2 with an adjusted p-value below 0.05 and a positive `log2FoldChange` in the comparison between Spike-specific $CD4^+$ pseudobulks at 1-3 months and Non-Spike specific $CD4^+$ pseudobulks at 1-3 months (i.e. more abundant in Spike specific cells) were entered into Enrichr.

Term	Overlap	P-value	Adjusted P-value	Genes
Aromatic Amino Acid Transport (GO:0015801)	3/8	0.000985188	0.76357	SLC25A29; SLC7A8; SLC38A3
Purine Nucleobase Transport (GO:0006863)	3/9	0.00144828	0.76357	SLC23A3; SLC29A3; SLC29A2
Negative Regulation of Fat Cell Differentiation (GO:0045599)	5/32	0.00153891	0.76357	WNT10B; E2F1; RORA; TRIB3; WNT1
p38MAPK Cascade (GO:0038066)	3/12	0.003571	0.76357	MAP2K3; ZFP36; MAP2K6
Regulation of Fat Cell Differentiation (GO:0045598)	7/75	0.00405648	0.76357	WNT10B; ZFP36; ZBTB16; E2F1; RORA; TRIB3; WNT1
Negative Regulation of Response to Stimulus (GO:0048585)	6/57	0.00422254	0.76357	SOCS2; USP25; GGT7; UBASH3B; IGFBP3; PDCD1
Cellular Response to Estrogen Stimulus (GO:0071391)	3/14	0.00567624	0.76357	BCAS3; ARID5A; RARA

Term	Overlap	P-value	Adjusted P-value	Genes
Regulation of Glucose Metabolic Process (GO:0010906)	4/28	0.00639123	0.76357	PMAIP1; RORA; MLYCD; FBP1
Negative Regulation of Signaling (GO:0023057)	4/28	0.00639123	0.76357	SOCS2; UBASH3B; IGFBP3; LDLR
Organic Cation Transport (GO:0015695)	6/62	0.00639249	0.76357	SLC25A29; SEC14L1; SLC7A8; SLC25A42; SLC38A3; SLC29A3

Table E.3: Table of the top 10 most significantly differentially expressed Gene Ontologies (Gene Ontology Biological Process 2025 database) under gene set enrichment analysis of genes more abundant in non-Spike specific $CD4^+$ T-cells at 4 Years. No gene sets reach significance after adjustment for multiple testing. Genes differentially expressed under DESeq2 with an adjusted p-value below 0.05 and a negative `log2FoldChange` in the comparison between Spike-specific $CD4^+$ pseudobulks at 4 years and Non-Spike specific $CD4^+$ pseudobulks at 4 years (i.e. more abundant in non-Spike specific cells) were entered into Enrichr.

Term	Overlap	P-value	Adjusted P-value	Genes
N-acetylneuramir Catabolic Process (GO:0019262)	2/5	0.00107463	0.477957	GNPDA2; NPL
Amino Sugar Catabolic Process (GO:0046348)	2/6	0.0016008	0.477957	GNPDA2; NPL
Type B Pancreatic Cell Development (GO:0003323)	2/6	0.0016008	0.477957	CDK6; DLL1

Term	Overlap	P-value	Adjusted P-value	Genes
Regulation of Myeloid Leukocyte Mediated Immunity (GO:0002886)	2/8	0.00294703	0.477957	DNASE1; UNC13D
Autophagosome Lysosome Fusion (GO:0061909)	2/8	0.00294703	0.477957	PLEKHM1; PIP4K2B
Positive Regulation of Mesenchymal Cell Proliferation (GO:0002053)	2/9	0.0037629	0.477957	PDGFA; FGFR2
G1/S Transition of Mitotic Cell Cycle (GO:0000082)	4/64	0.00459005	0.477957	CDK6; CCNG2; HINFP; CDKN3
N-acetylneuramin Metabolic Process (GO:0006054)	2/10	0.00467121	0.477957	GNPDA2; NPL
Type B Pancreatic Cell Differentiation (GO:0003309)	2/10	0.00467121	0.477957	CDK6; DLL1
Cell Cycle G1/S Phase Transition (GO:0044843)	4/66	0.0051228	0.477957	CDK6; CCNG2; HINFP; CDKN3

Table E.4: Table of the top 10 most significantly differentially expressed Gene Ontologies (Gene Ontology Biological Process 2025 database) under gene set enrichment analysis of genes more abundant in Spike specific $CD4^+$ T-cells at 4 Years. No gene sets reach significance after adjustment for multiple testing. Genes differentially expressed under DESeq2 with an adjusted p-value below 0.05 and a positive `log2FoldChange` in the comparison between Spike-specific $CD4^+$ pseudobulks at 4 years and Non-Spike specific $CD4^+$ pseudobulks at 4 years (i.e. more abundant in Spike specific cells) were entered into Enrichr.

Term	Overlap	P-value	Adjusted P-value	Genes
Cellular Response to Hydroxyurea (GO:0072711)	5/8	0.000255214	0.493449	BLM; DDI2; DDX11; TIMELESS; SPIDR
Regulation of Melanin Biosynthetic Process (GO:0048021)	5/9	0.000531761	0.493449	ZEB2; GIPC1; CTNS; RAPGEF2; MFSD12
Response to Hydroxyurea (GO:0072710)	5/9	0.000531761	0.493449	BLM; DDI2; DDX11; TIMELESS; SPIDR
Regulation of Cytoskeleton Organization (GO:0051493)	25/139	0.000577979	0.493449	GSK3B; ARHGAP18; ARHGAP17; SLC4A2; BICD1; CLN3; SH3BP1; DVL2; MLST8; ABL2; DVL3; PLXNA3; BCAS3; ARHGEF19; HIP1R; ARAP1; RHOF; BAIAP2; SORBS3; P2RX7; CAPN10; MAP6D1; PKN1; ARHGEF7; PHLDB2
Negative Regulation of Epidermal Growth Factor Receptor Signaling Pathway (GO:0042059)	8/25	0.00110107	0.752028	ZFYVE28; DUSP3; ZGPAT; MVP; PTPRJ; MMACHC; SOCS4; SOCS5

Term	Overlap	P-value	Adjusted P-value	Genes
Glycerophosph Biosynthetic Process (GO:0046474)	20/110	0.0017123	0.974586	MTMR1; VAC14; DGKD; PCYT2; ABHD4; SH2D3C; AGPAT1; AGPAT2; PLD2; INPP4A; CLN3; PLSCR1; PIKFYVE; GPAM; LPCAT4; CHEK2; PITPNM2; PIP5K1A; CDIPT; PI4K2A
Regulation of Epidermal Growth Factor Receptor Signaling Pathway (GO:0042058)	11/51	0.00480235	0.999996	ZFYVE28; DGKD; DUSP3; RALB; PDE6G; MVP; NUP62; RHBDF2; HIP1R; PTPRJ; MMACHC
Regulation of Actin Filament-Based Process (GO:0032970)	17/96	0.00484771	0.999996	BCAS3; ARHGEF19; HIP1R; ARHGAP18; ARHGAP17; SLC4A2; ARAP1; RHOF; BAIAP2; SH3BP1; CAPN10; MLST8; DVL2; ABL2; DVL3; PKN1; ARHGEF7
Regulation of Actin Cytoskeleton Organization (GO:0032956)	18/104	0.00489862	0.999996	BCAS3; ARHGEF19; HIP1R; ARHGAP18; ARHGAP17; NEB; SLC4A2; ARAP1; RHOF; BAIAP2; SH3BP1; CAPN10; DVL2; MLST8; ABL2; DVL3; PKN1; ARHGEF7
Nucleic Acid Metabolic Process (GO:0090304)	7/25	0.00519344	0.999996	EXD2; ATXN2; PDE12; TREX1; SETMAR; DXO; DNASE1L1