



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Investigating the Genetic and Immunological Aetiology of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome

Joshua James Dibble



THE UNIVERSITY
of EDINBURGH

Thesis submitted for the award of:

Doctor of Philosophy (Genetics and Molecular Medicine)

The University of Edinburgh

2022

Dedicated to my family - I wouldn't have made it this far without you.

I don't intend to stop here.

Declaration

I confirm that this thesis has been composed by me, and that the work presented is my own, except where otherwise noted. This work has not been submitted for any other degree or professional qualification, and work taken from publications is clearly indicated.

Extracts from the publication “Genetic risk factors of ME/CFS: a critical review”, J.J. Dibble, S.J McGrath, C.P. Ponting, *Hum. Mol. Genet.* vol 29(R1) (2020) are included in **Chapters 1** and **2**. I wrote the first draft of this paper, and have obtained permission from my co-authors for its use in this thesis. The inclusion of work from this review is clearly indicated.

Joshua James Dibble

Date: 30/06/22

Abstract

This thesis describes two investigations into the disease Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS), specifically its genetic aetiology and immune system alterations. The first study investigated the genetic basis of ME/CFS using Genome-wide Association Studies (GWAS) by attempting to replicate and extend results previously found using UK Biobank cohort data. GWAS attempt to identify associations between DNA variants and phenotypes. This GWAS was novel, conducted on new phenotypes constructed by combining those in the most up-to-date UK Biobank data release. A new, previously unseen, genome-wide significant association was found on chromosome 6 for males with ME/CFS within the gene *PDE10A*. Further results were not genome-wide significant, but many were suggestive and hence independent replication may justify further research.

A previous analysis on the UK Biobank cohort had identified an indicative association in females between variants around the *SLC25A15* gene at genome-wide significance. I adopted a hypothesis that the dietary protein intake of people with the CFS risk variants would be lower than those with the alternative alleles, due to potentially reduced production of mitochondrial ornithine transporter 1 (*ORNT1*). However, this association with dietary protein intake was not supported by UK Biobank data. Additionally, I investigated associations between the human leukocyte antigen (HLA) alleles and the ME/CFS phenotype using UK Biobank data. Associations between alleles within the *HLA-C* and *-DQB1* genes had previously been found in a cohort of Norwegian people with ME/CFS, and my goal was to seek replication of these results in a larger dataset. None of the associations found in the UK Biobank proved to be genome-wide significant.

In my second study I investigated the use of T-cell clonal diversity as a potential biomarker for ME/CFS. This project used cells from CureME Biobank samples in collaboration with Systems Biology Laboratory (SBL). I developed a data analysis pipeline to analyse T-cell receptor (TCR) genomic DNA data based on the best practices currently used in the fields of immunology and mathematical biology. This approach used a mathematical notion of entropy as a measure for the diversity of TCR repertoires, in this way combining all of the most commonly used metrics in mathematical biology. When combined, these measures form a profile for each repertoire, a set of which can be sorted using a machine learning algorithm to partition the repertoires into subgroups.

My hypothesis was that the T-cell clonal expansion of people with ME/CFS would be greater than for healthy controls, and comparable to disease (multiple sclerosis) controls. Although this method was able to effectively classify TCR chains using simulated data, results from experimentally-derived data did not support the hypothesis, with the most effective classifications for both CD4⁺ and CD8⁺ cells failing to pass corrections for multiple hypothesis significance testing.

Lay Summary

Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) is a disease that affects millions of people around the world. Very little is understood about the cause or progression of the disease, and there is no known cure. At present, there is also no reliable clinical test to determine whether a person has ME/CFS. This thesis explores the potential for a genetic or immunological basis for ME/CFS, with the goal to eventually find a biomarker that could be used in diagnosis.

The first part of this thesis investigates whether genetic variants are more (or less) common among those with ME/CFS than in the general population. In particular, the region of the genome that encodes immune system proteins was of interest, as previous studies have shown associations between this region and the disease. Using strict statistical thresholds, none of the previously found associations were replicated. However, one new association was found, with the gene *PDE10A*, which is implicated in central nervous system diseases, such as Parkinsons and Huntingtons disease. This association has never been seen before, and would require replication in a new cohort before its role in ME/CFS could be confirmed. However, it represents a promising avenue for new research.

The second part of this thesis investigates T-cells. These are highly specialised immune cells in the blood, each of which targets an antigen (foreign substance) such as from a virus. When a T-cell recognises this antigen, it clones itself repeatedly. This clonal expansion is measurable, and can serve as evidence of immune system activation. My hypothesis was that this immune signature could be used to distinguish people with ME/CFS from healthy controls and others diagnosed with another disease. I used a mathematical measure of diversity and a machine learning method to sort their immune profiles into groups. However, the pattern of immune activation was not sufficiently clear to provide consistent classification. Hence, the role of the immune system in ME/CFS is still unclear, and the utility of this method as a diagnostic biomarker is not proved.

Contents

1	Introduction	23
1.1	Case Definitions	24
1.2	Prevalence, Diagnosis and Biomarkers	26
1.3	Treatment	27
1.4	Causes	29
1.5	Similar and Co-morbid Diseases	30
1.5.1	Fibromyalgia	30
1.5.2	(Chronic) Lyme Disease	31
1.5.3	Post-Covid Syndrome (Long Covid)	31
1.5.4	Postural Tachycardia Syndrome	32
1.5.5	Migraine Syndromes	33
1.5.6	Irritable Bowel Syndrome	33
1.5.7	Gulf War Syndrome	33
1.5.8	Mental Illnesses	34
1.5.9	Multiple Sclerosis	35
1.5.10	Guillain-Barré Syndrome	36
2	Areas of Current ME/CFS Research	37
2.1	Genetics	37
2.1.1	Relatedness	38
2.1.2	Heritability	38
2.1.3	ME/CFS Heritability	41
2.1.4	ME/CFS and UK Biobank	43
2.1.5	Studies Not Using the UK Biobank Data	46
2.1.6	Candidate Gene Studies	48
2.1.7	Autoimmunity (Genetics)	49
2.2	Immunology (Lymphocytes)	50
2.3	Summary, and the Basis for Further Research	53

3	Genetic Aetiology of ME/CFS and Fibromyalgia	54
3.1	Genome-wide Association Studies	54
3.1.1	Phenotypes in the UK Biobank	56
3.1.2	Combined Phenotypes	62
3.1.3	Power	65
3.1.4	GWAS Methods	66
3.1.5	Original Phenotypes Results	70
3.1.6	Combined Phenotypes Results	76
3.1.7	Summary of Results and Interpretation	79
3.2	Human Leukocyte Antigen Loci	87
3.2.1	Association of HLA Alleles with ME/CFS	89
3.2.2	Association of HLA Alleles with Fibromyalgia	90
3.2.3	Combined Cohorts	91
3.3	Discussion	92
4	T-cell Receptor Repertoire Diversity: Background and Methods	94
4.1	Biological Background	94
4.1.1	Adaptive Immune System	95
4.1.2	VDJ Recombination	95
4.1.3	Clonal Activation in Disease	97
4.2	Experimental Design	98
4.3	Systems Biology Laboratory Processing	99
4.4	Assigning Clonotypes	103
4.5	T-cell Receptor Diversity Estimation	104
4.5.1	Diversity Measures	104
4.5.2	Evenness Measures	109
4.6	Computational Implementation	110
4.6.1	Selection of the α -range	111
4.6.2	Generation of the Distance Matrix	118
4.6.3	Multidimensional Scaling	119
4.6.4	Classification (Potential Support Vector Machine)	121
4.6.5	Validation of the Analysis via Simulation	124
4.7	Hypotheses	126
4.7.1	Parameters	128
5	T-cell Receptor Repertoire Diversity: Results	130
5.1	Data Summary	130
5.2	Post-unblinding Analysis Adjustments	131

5.3	CD8 ⁺ Cells	132
5.3.1	Combined Analysis (α -, β -, and γ -chains)	133
5.3.2	Separated by Chain Type	139
5.4	CD4 ⁺ Cells	141
5.4.1	Separated by Chain Type	143
5.5	Axis Correlations by Chain Type	144
5.6	Summary	145
5.6.1	Discussion	146
5.7	Additional Analysis I: Power Law Models	147
5.8	Additional Analysis II: Gene Segment Usage	150
5.9	Additional Analysis III: CMV Seropositivity	152
6	Discussion	154

List of Figures

2.1	An ME/CFS-associated GWAS locus on chromosome 13 (x-axis). Genome-wide significant SNPs from Neale <i>et al.</i> [120] are those lying above the horizontal dashed line. The location of the <i>SLC25A15</i> gene is indicated below (exons are shown as vertical bars). The left-hand y-axis reflects the statistical confidence ($-\log_{10} p$ -value) of the association between a DNA variant and ME/CFS case status. The right-hand y-axis (blue data curves) indicates the low extent of recombination within this locus. rs7337312 is highlighted as the reference variant (purple diamond). The degree of linkage disequilibrium of rs7337312 with neighbouring SNPs is indicated as r^2 (right-hand legend).	45
2.2	Histograms of replication p -values for association of SNPs for CFS risk. These p -values were obtained from the UK Biobank CFS GWAS of Canela-Xandri <i>et al.</i> [100]. The variants tested had been previously identified as associated with CFS by published analyses in separate datasets. Only variants tested in the UK Biobank were considered. No evidence of a skew towards lower p -values was observed. A) Association p -values for SNPs identified by Wang <i>et al.</i> [138]. B) Association p -values for SNPs identified by Marshall-Gradisnik <i>et al.</i> [139] and Schlauch <i>et al.</i> [136].	47
2.3	Slides from Dr Mark Davis' presentation, OMF Symposium 2017 [156].	51
3.1	Intersections between the CFS and Fibromyalgia phenotypes from the UK Biobank (data from February 2022), and the ME/CFS and Fibromyalgia phenotypes from the "Experience of Pain" questionnaire (issued in 2019).	57

3.2	Set intersections for those excluded from analyses. The majority (132,607) were removed due to relatives in the cohort (kinship). Restricting to the White British subset results in the loss of some power ($\sim 15\%$ of the cohort), but reduces the likelihood of confounding due to genetic ancestry. Other genetic ancestries can be considered separately, or accounted for using a genetic relationship matrix approach (see Section 2.1.1).	59
3.3	Intersections between the four phenotypes after filtering.	60
3.4	Age distributions (in years) for each of the four phenotypes, by case status and sex [175]. ME/CFS, Fibromyalgia (Pain Questionnaire), Fibromyalgia, and CFS are shown clockwise from the top left. . . .	61
3.5	Intersection for the constructed ME/CFS and Fibromyalgia phenotypes. These shall be used for analyses in subsequent sections. Venn diagram drawn with the “venn” R package [177].	64
3.6	Manhattan and QQ plots showing the results of GWAS on the CFS phenotype in the UK Biobank, for the full (1,358 cases, 274,949 controls), male (368 cases, 128,678 controls), and female (990 cases, 146,271 controls) cohorts. Genomic inflation factors were $\lambda = 1.01$, $\lambda = 1.01$, and $\lambda = 1.01$ for these groups, respectively. The red lines on the Manhattan plots indicate genome-wide significance at $p < 5 \times 10^{-8}$ and blue indicates the suggestive threshold of $p < 1 \times 10^{-5}$. No genome-wide significant associations were detected, however, there is a suggestive signal on chromosome 1 for the CFS phenotype in males. There is also a suggestive association on chromosome 13 in females with the same phenotype.	72
3.7	Manhattan and QQ plots showing the results of GWAS on the ME/CFS phenotype in the UK Biobank, for the full (1,561 cases, 94,720 controls), male (443 cases, 42,131 controls), and female (1,118 cases, 52,589 controls) cohorts. Genomic inflation factors were $\lambda = 1.02$, $\lambda = 1.02$, and $\lambda = 1.00$ for these groups, respectively. The red lines on the Manhattan plots indicate genome-wide significance at $p < 5 \times 10^{-8}$ and blue indicates the suggestive threshold of $p < 1 \times 10^{-5}$. No genome-wide significant associations were detected.	73

- 3.8 Manhattan and QQ plots showing the results of GWAS on the Fibromyalgia phenotype in the UK Biobank, for the full (554 cases, 275,753 controls), male (66 cases, 128,980 controls), and female (488 cases, 146,773 controls) cohorts. Genomic inflation factors were $\lambda = 1.01$, $\lambda = 1.01$, and $\lambda = 1.01$ for these groups, respectively. The red lines on the Manhattan plots indicate genome-wide significance at $p < 5 \times 10^{-8}$ and blue indicates the suggestive threshold of $p < 1 \times 10^{-5}$. No genome-wide significant associations were detected. 74
- 3.9 Manhattan and QQ plots showing the results of GWAS on the Fibromyalgia (PQ) phenotype in the UK Biobank, for the full (1,697 cases, 93,308 controls), male (291 cases, 41,597 controls), and female (1,406 cases, 51,711 controls) cohorts. Genomic inflation factors were $\lambda = 1.00$, $\lambda = 1.00$, and $\lambda = 1.02$ for these groups, respectively. The red lines on the Manhattan plots indicate genome-wide significance at $p < 5 \times 10^{-8}$ and blue indicates the suggestive threshold of $p < 1 \times 10^{-5}$. No genome-wide significant associations were detected. 75
- 3.10 Manhattan and QQ plots showing the results of GWAS on the ME/CFS (Combined) phenotype, for the full (2,350 cases, 273,903 controls), male (666 cases, 128,365 controls), and female (1,684 cases, 145,538 controls) cohorts. Genomic inflation factors were $\lambda = 1.02$, $\lambda = 1.02$, and $\lambda = 1.00$ for these groups, respectively. The red lines on the Manhattan plots indicate genome-wide significance at $p < 5 \times 10^{-8}$ and blue indicates the suggestive threshold of $p < 1 \times 10^{-5}$. There was a genome-wide significant association on chromosome 6 for males (shown in green). 77
- 3.11 Manhattan and QQ plots showing the results of GWAS on the Fibromyalgia (Combined) phenotype, for the full (1,990 cases, 274,308 controls), male (322 cases, 128,724 controls), and female (1,668 cases, 145,584 controls) cohorts. Genomic inflation factors were $\lambda = 1.01$, $\lambda = 1.00$, and $\lambda = 1.02$ for these groups, respectively. The red lines on the Manhattan plots indicate genome-wide significance at $p < 5 \times 10^{-8}$ and blue indicates the suggestive threshold of $p < 1 \times 10^{-5}$. No genome-wide significant associations were detected. 78

3.12	LocusZoom plot [201] of a section of chromosome 6, showing the risk variant rs76346913 (purple diamond). Recombination hotspots are visible as spikes in the recombination rate (blue line).	81
3.13	Part of chromosome 13 from Figure 3.6C , with the three SNPs with $p < 1 \times 10^{-5}$ around <i>SLC25A15</i> highlighted in green.	86
3.14	Violin plots of average self-reported dietary protein intake (in grams) for CFS cases and controls, separated by sex.	87
3.15	Distribution of alleles per locus [175]. Low quality imputations had been removed ($Q < 0.7$), resulting in < 22 alleles in total for some participants.	89
3.16	Scatter plots showing the significance values of HLA alleles with frequency ≥ 0.05 for A) CFS and B) ME/CFS. No alleles were significant after genome-wide multiple testing correction ($p < 5 \times 10^{-8}$).	90
3.17	Scatter plots showing the significance values of HLA alleles with frequency ≥ 0.05 for A) Fibromyalgia and B) Fibromyalgia (PQ). No alleles were significant after genome-wide multiple testing correction ($p < 5 \times 10^{-8}$).	91
3.18	Scatter plots showing the significance values of HLA alleles with frequency ≥ 0.05 for A) ME/CFS (Combined) and B) Fibromyalgia (Combined). No alleles were significant after genome-wide multiple testing correction ($p < 5 \times 10^{-8}$).	91
4.1	Graphic showing the recombination process. A V and a J gene segment are brought together by folding the genome. The resulting “loop” of DNA is excised, and the ends can merge to form a circle.	97
4.2	An overview of the project workflow.	98
4.3	Plot of sorted PBMCs stained for CD8-FITC (x-axis) vs CD4-PE (y-axis), from SBL sample number 143. FITC and PE are fluorophores used in cell staining. The gates shown were used to verify the sorting process was successful. Approximately 2.53×10^6 cells in total are shown. The mean purity per sample is shown in the data exploration of Chapter 5 (Table 5.2) . Figure courtesy of SBL.	100
4.4	Hair-pinned adapter. Figure courtesy of SBL.	100
4.5	Adapter design. Figure courtesy of SBL.	101

4.6	A. Diversity profiles of XPBF647 (red) and XPBF648 (blue), both CD8 ⁺ libraries from ATCC healthy controls. XPBF647 had 1,644 recombinants, and XPBF648 1,454. The intersection of the profiles is magnified in the upper right corner of the plot. B. Corresponding clonotype count plots for the same libraries, with the clone index on the horizontal axis and the total number of that clone found on the vertical. Clonotype number was assigned after sorting by size, so a larger clonotype ID indicates fewer instances of that clonotype found.	107
4.7	Analysis of simulated clonotype data. A) Rényi entropies (H_α , Equation 4.1) calculated for three simulated datasets Red, Blue and Green. The Red and Green samples have the same number of simulated sequences, but Green has been artificially constructed to demonstrate that even when the sample size is normalised, profile intersection is not guaranteed. B) Clonotype distributions for the corresponding profiles in A) , ordered from most expanded clonotype to least along the x-axis. Note that the “tails” for Red and Blue are singletons, and the “tail” of Green is singletons and doubles.	108
4.8	Analysis pipeline for the computational implementation of the diversity analysis.	110
4.9	Recombinants recovered from the initial NovaSeq sequencing run. There was a two order of magnitude difference in the number recovered across the range of samples. Figure courtesy of SBL.	112
4.10	Diversity profiles for subsamples of a TCR dataset (taken at 10% size intervals). Red are from library XPBF647, blue are from XPBF648 (as in Figure 4.6). Significant variance of D_α is seen for low α values, reflecting that rare clonotypes will drop out as subsamples of decreasing size are taken.	113
4.11	Effects of the Horvitz-Thompson correction for: A) Libraries XPBF647 and XPBF648 (as in Figure 4.6). B) Subsamples of library XPBF647 (as in Figure 4.10).	115
4.12	Visual representation of the α -optimisation algorithm.	117
4.13	Pairwise distance between the ATCC test samples (y-axis) as a function of resample number, n , on a log scale. This was used to indicate the number of resamples needed for accurate calculation of the distance matrix M	120
4.14	Flowchart showing the PSVM algorithm.	122

4.15	Example clonotype distributions for simulated severe ME/CFS cases (red, yellow), mild/moderate ME/CFS cases (green), and healthy controls (blue).	125
4.16	Multidimensional scaling plot (projection) of the distance matrix generated from the simulated data shown in Figure 4.15 . Calculated using non-optimal parameters of $n = 100$, $\alpha \in [30, 50]$ and step size 1.	126
4.17	Multidimensional scaling plot (projection) of the distance matrix generated from the “red” and “yellow” simulated data shown in Figure 4.15 . Calculated using non-optimal parameters of $n = 100$, $\alpha \in [30, 50]$ and step size 1. A. Diversity profiles for samples from the red and yellow groups from Figure 4.16, with the region used in the distance matrix calculation highlighted. B. Multidimensional scaling plot (projection) of their distance matrix.	127
5.1	Repertoires of the 154 CD8 ⁺ samples. The first four are shown in greater detail, with subgroup labels. Colours are assigned to each clonotype in a sample, starting with red and gradually desaturating to white. More monochrome plots therefore indicate samples with a greater richness.	133
5.2	Multidimensional scaling plot of all CD8 ⁺ chains. ME(sa) are in red, ME(mm) are in yellow, MS are in green and HC are in blue. Axes indicate projected distances, such that the linear distance between two samples reflects the value in the distance matrix. There is no clear visual separation between the groups in two dimensions, but there are notable disease case (ME(sa), ME(mm) and MS) outliers for $x > 2,000$	134
5.3	The MDS plot of Figure 5.2 , with a circle of radius 2,800 in this coordinate space. 12 points were outside of the circle, all disease cases.	138
5.4	The dartboards for the CD8 ⁺ outliers, as initially shown in Figure 5.1	138
5.5	MDS plots showing the changes to the distance matrix caused by adjusting the step size and α -range used in the calculations. Plots are shown for the α -chain CD8 ⁺ data (compare with Figure 5.6) with: A) An α -range of $[0, 20]$ and a step size of 0.1. B) An α -range of $[0, 10]$ and a step size of 0.2, as in Greiff <i>et al.</i>	139

5.6	Multidimensional scaling plots for the three chain types of CD8 ⁺ cells. α , β and γ are shown clockwise from the top left. ME(sa) are in red, ME(mm) are in yellow, MS are in green and HC are in blue. No clear separation is visible in two dimensions, although it appears that the outliers of Figure 5.2 are primarily driven by diversity differences in the α chains, consistent with the correlation found in Table 5.4	140
5.7	Repertoires of the 148 CD4 ⁺ samples. The first four are shown in greater detail, with subgroup labels. Colours are assigned to each clonotype in a sample, starting with red and gradually desaturating to white. More monochrome plots therefore indicate samples with a greater richness.	142
5.8	Multidimensional scaling plots for the three chain types of CD4 ⁺ cells. α , β and γ are shown clockwise from the top left. ME(sa) are in red, ME(mm) are in yellow, MS are in green and HC are in blue. No clear separation was visible in two dimensions.	143
5.9	Log-log plot showing a linear relationship between cumulative clonotypes and the recombinants in a given clonotype.	148
5.10	Histograms showing the distributions of exponents and intercepts from linear models fit to both CD4 ⁺ and CD8 ⁺ cell data. All distributions were unipolar.	149
5.11	Histograms showing the p -value distributions for each of the six hypotheses tested (Case vs HC, ME vs MS, and MEsa vs MEmm, for both CD8 ⁺ and CD4 ⁺ cells).	151

List of Tables

1.1	Summary of studies reporting a proportion of their subjects with ME/CFS to have a suspected viral onset.	29
2.1	Summary of SNPs identified as significant in the UK Biobank CFS Cohort [6] (updated to include the finding of Hajdarevic <i>et al.</i> [118]).	44
3.1	Summary of the numbers of people in the filtered cohort with each phenotype, separated by sex. The female to male ratio is markedly increased compared with the controls (none of the four phenotypes).	60
3.2	The logical rationale for considering the exclusion of some participants from analyses. “PQ” indicates the “Experience of Pain” questionnaire. The numbers of participants from the filtered cohort in each group are included. Those without a definitive response to the pain questionnaire were counted as “No” for ME/CFS. This table was used for the construction of the “Combined” phenotypes of Section 3.1.2 . The same section also defines the criteria for pre- and post-PQ inclusion.	62
3.3	Power to detect associations, shown for various cases of risk allele frequency and relative risk, using an additive model. Estimates are for the ME/CFS (Combined) phenotype, and an assumption was made that the phenotyping process ensured none of the controls had the disease.	65
3.4	Investigation using <i>glm()</i> in <i>R</i> of the covariates to be used in the GWAS. The results shown are for ME/CFS (Combined), but were similar for the other phenotypes. Sex and age were the only co-variates which significantly affected all six phenotypes. Significance levels are shown for $p < 0.001$ (***), $p < 0.01$ (**), and $p < 0.05$ (*).	68

3.5	Suggestive and genome-wide significant SNPs from the GWAS of Sections 3.1.5 and 3.1.6 ($p < 1 \times 10^{-5}$). The genome-wide significant result is marked in bold. Note that the MAF for each SNP was calculated using the filtered cohort of Section 3.1.1 . The distance to the nearest gene is shown for each variant, in base pairs (bp), as identified by Open Targets [161].	80
3.6	Minor allele frequencies for rs76346913 ($C \rightarrow T$).	82
3.7	Results of fitting a GLM (using PLINK2 [171]) to the most promising SNPs from the female CFS GWAS of Section 3.1.5 , as highlighted in Figure 3.13 . No significant associations were detected between average dietary protein intake in females and any of the three variants tested.	86
3.8	Number of alleles imputed for each locus.	88
4.1	Pairwise distances between the diversity profiles of two sets of subsamples of XPBF647 and XPBF648 (shown in Figure 4.7). Sample sizes are shown both as absolute values, and as a percentage of the total, for each subsample. The distances where sample sizes most closely matched for each column are shown in green.	113
4.2	Pairwise distances between the diversity profiles ($10 \leq \alpha \leq 50$) of two distinct sets of subsamples of XPBF647 only (shown in Figure 4.7 , red). Sample sizes are shown both as absolute values, and as a percentage of the total, for each subsample. Matching sample sizes are shown in green.	114
4.3	Decision boundaries calculated. "ME" indicates combined ME(mm) and ME(sa), "Case" indicates disease case (MS and ME combined)	128
5.1	A two-way table of sample counts by disease subgroup and age bracket. Severely affected people with ME/CFS are significantly overrepresented in younger age groups.	130
5.2	Means of key experimental parameters, shown for all samples, and by disease subgroup. Comparison by ANOVA (<i>aov()</i> in <i>R</i>) showed only one significant association – for CD8 ⁺ cell number ($p = 0.0257$).131	
5.3	Number of CD8 ⁺ samples, by subgroup, available for analysis. . . .	132

5.4	Significance values for the variables tested as potential confounders with the MDS plot x-axis. Significance levels are shown for $p < 0.001$ (***), $p < 0.01$ (**), and $p < 0.05$ (*). “ME/CFS(sa)”, “ME/CFS(mm)” and “MS” were testing whether being a member of the given subgroup was correlated with the axis, relative to HC. “Age” was the linear term of the model fit for that covariate. “PBMC Number” was the total number of cells for that sample, and “CD8 ⁺ % Input” was the percentage of input PBMCs that were CD8 ⁺ cells. “CD8 ⁺ Cells Post-selection” was the number of CD8 ⁺ cells after the SureSelect process (see Section 4.3), and “Purity” the corresponding CD8 ⁺ purity at this stage. “DNA Extracted (μ g)” was the total amount of DNA initially recovered from each sample, and “DNA Yield (ng)” was the amount remaining after size selection (see Section 4.3). “Demultiplexed Reads [Library]” were the total reads recovered for the two Illumina flowcells of sufficient quality (library numbers 641 and 782), and “PEAR Assembled” were the percentages assembled into clonotypes for the same libraries using the Paired-End Read Merger tool [286]. “[Chain Type] Coding Junctions” were the total number of coding junction of each chain type found, and “[Chain Type] %” was the percentage of each chain type, relative to the other two types (such that the sum was 100%).	135
5.5	Significance values for the variables tested as potential confounders with the MDS plot y-axis. Significance levels are shown for $p < 0.001$ (***), $p < 0.01$ (**), and $p < 0.05$ (*). Parameters are the same as for Table 5.4	136
5.6	p -values for the primary hypotheses for the TCRs from CD8 ⁺ cells. “ME” indicates ME(sa) and ME(mm) combined.	141
5.7	p -values for the secondary hypotheses for the TCRs from CD8 ⁺ cells. “Cases” indicates ME(sa), ME(mm) and MS combined, “ME” indicates ME(sa) and ME(mm) combined. One hypothesis had $p < 0.05$ (*), and three failed to reach 1,000 permutations (†).	141
5.8	Number of CD4 ⁺ samples, by subgroup, available for analysis. . . .	142
5.9	p -values for the primary hypotheses for the TCRs from CD4 ⁺ cells. “ME” indicates ME(sa) and ME(mm) combined. One hypothesis had $p < 0.05$ (*).	144

- 5.10 p -values for the secondary hypotheses for the TCRs from CD4⁺ cells. “Cases” indicates ME(sa), ME(mm) and MS combined, “ME” indicates ME(sa) and ME(mm) combined. Two hypotheses had $p < 0.05$ (*), and three failed to reach 1,000 permutations (†). 144
- 5.11 Significance values from the GLM fit (*glm()* in *R*) for the variables tested as potential confounders with the MDS plot x- and y-axes. Parameters are the same as for **Table 5.4**, with the addition of “Demultiplexed Reads 948”, which were the total reads recovered for the Illumina flowcell (library number 948) used for the CD4⁺ cells of sufficient quality. PEAR assembly was not used for the CD4⁺ cell data. Associations with $p < 0.05$ are marked in bold. . . 145
- 5.12 p -values from permutation testing for CD4⁺ and CD8⁺ cells, by chain type. Asterisks (*) indicate results with $p < 0.05$, and † indicates $< 1,000$ permutations were used to generate the p -value. 145
- 5.14 Gene segments where caseness was a significant predictor ($p < 0.05$) under a negative binomial model (*glm.nb()* in *R*, from the *MASS* package [296]), used to indicate differences in usage among the subgroups. No segments remained significantly associated with differential usage after multiple hypothesis testing correction. Hence, these data do not support rejection of the null hypothesis. 152
- 5.15 CMV seropositivity results, shown by disease subgroup. Samples were sorted into “Positive” and “Negative” based upon the level of CMV immunoglobulin G (IgG) measured in their plasma. “Equivocal” were those with indeterminate levels of IgG. 152

List of Abbreviations

AIDP – Acute Inflammatory Demyelinating Polyradiculopathy
APC(s) – Antigen-Presenting Cell(s)
ATCC – American Type Culture Collection
(c)AMP – (Cyclic) Adenosine Monophosphate
CBT – Cognitive Behavioural Therapy
CCC – Canadian Consensus Criteria
CDC – Centers for Disease Control (and Prevention)
CDR(3) – Complementarity-Determining Region (3)

CFS – Chronic Fatigue Syndrome
(c)GMP – (Cyclic) Guanosine Monophosphate
CI – Confidence Interval
CNS – Central Nervous System
CNV – Copy Number Variation
EBV – Epstein-Barr Virus
(e)QTL – Expression Quantitative Trait Locus
FACS – Fluorescence Activated Cell Sorting
(g)DNA – (Genomic) Deoxyribonucleic Acid
GET – Graded Exercise Therapy
GLM – Generalised Linear Model
GRM – Genetic Relationship Matrix
GWAS – Genome-Wide Association Study (Studies)
HC – Healthy Control
(H)CMV – (Human) Cytomegalovirus
HHH – Hyperornithinemia-Hyperammonemia-Homocitrullinuria
HHV – Human Herpesvirus
HLA – Human Leukocyte Antigen
HWE – Hardy-Weinberg Equilibrium
IBD – Inflammatory Bowel Disease
IBS – Irritable Bowel Syndrome
ICC – International Consensus Criteria
(i)CPET – (Invasive) Cardiopulmonary Exercise Test(ing)
IgG4-RD – IgG4-Related Disease
IOM – (USA) Institute of Medicine
JSON – JavaScript Object Notation
 k -NN – k -Nearest Neighbours
LD – Linkage Disequilibrium
MACS – Magnetic Activated Cell Sorting
MAF – Minor Allele Frequency
MDS – Multidimensional Scaling
ME – Myalgic Encephalomyelitis
ME(mm) – ME (Mild/Moderate)
ME(sa) – ME (Severely Affected)
MHC – Major Histocompatibility Complex
MIGEC – Molecular Identifier Guided Error Correction
ML – Machine Learning

mRNA – Messenger Ribonucleic Acid
MSE – Mean Squared Error
MS – Multiple Sclerosis
NAM – National Academy of Medicine
NEB – New England Biolabs
NICE – (UK) National Institute of Health and Care Excellence
NK – Natural Killer (Cells)
OR – Odds Ratio
PACE – Pacing, graded Activity, and Cognitive behaviour therapy: a randomised Evaluation
PAMP(s) – Pathogen-Associated Molecular Pattern(s)
PBMC(s) – Peripheral Blood Mononuclear Cell(s) PCR – Polymerase Chain Reaction
PEAR – Paired-End reAd mergeR
PEM – Post-Exertional Malaise
PheWAS – Phenome-Wide Association Study
POTS – Postural Tachycardia Syndrome
PQ – (Experience of) Pain Questionnaire
PRR(s) – Pattern-Recognition Receptors
(P)SVM – (Potential) Support Vector Machine
QC – Quality Control
QQ – Quantile-Quantile
RAF – Risk Allele Frequency
RAG – Recombination Activating Gene
RALD – Ras-Associated Lymphoproliferative Disease
RR – Relative Risk
RSS – Recombination Signal Sequence
RTCR – Recover T-Cell Receptor
SBL – Systems Biology Laboratory
SF-36 – Short Form-36
SNP(s) – Single Nucleotide Polymorphism(s)
SPRI – Solid Phase Reversible Immobilisation
TCR(s) – T-Cell Receptor(s)
TREC(s) – T-cell Recombination Excision Circle(s)
TSH – Thyroid Stimulating Hormone
UCSC – University of California, Santa Cruz (Genome Browser)
UMI – Unique Molecular Identifier

VIF – Variance Inflation Factor

Estimated Word Count: 55,849

All analytical code and data made publicly accessible on publication of the project manuscripts will be available at <https://github.com/JoshuaDibble108>.

Chapter 1

Introduction

Myalgic Encephalomyelitis (ME), sometimes known as Chronic Fatigue Syndrome (CFS), is a multi-systemic disabling disease of unknown aetiology [1]. To avoid ambiguity, the usual convention is to refer to the illness as Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS), with CFS sometimes referring to earlier definitions (such as the 1994 CDC “Fukuda” criteria) [2].

The disease is highly heterogeneous, with many proposed case definitions differentiated by groupings of primary and secondary symptoms (see **Section 1.1**). It is not yet known whether the conditions described by these case criteria are different clinical presentations of one disease, or a group of related diseases.

People with ME/CFS, in common with other “invisible” conditions, frequently experience disbelief as to their lived experience, despite a clear and measurable reduction in quality of life. Research indicates that the effect on quality of life is worse than multiple sclerosis (MS) [3], mental illness, osteoarthritis, and many cancers [4]. People with ME/CFS are at significant risk of attempting or committing suicide as compared to the general population [5].

Despite its high prevalence, ME/CFS research receives relatively little research funding. As a result, almost nothing is known regarding the aetiology or progression of the illness. In particular, there is no reliable biomarker or diagnostic test to definitively determine whether someone has ME/CFS. The search for a biomarker forms a major component of current research efforts, and its discovery would represent an important advancement in our understanding of this disease.

1.1 Case Definitions

The most commonly used definitions in research this century are shown in **Box 1** (adapted from Dibble *et al.* [6]).

This lack of consistency in definition is an extremely significant problem present in ME/CFS research. Inability to define clear and consistent cohorts for research purposes has hampered progress, and is likely a leading cause for the lack of replication in results seen in the field (for example, in genetics research [6]). It is possible, or even likely, that ME/CFS is an umbrella term for a heterogeneous group of related diseases with different mechanisms and pathways, and similar clinical presentation [7].

However, the primary distinction between ME/CFS and other fatiguing illnesses is usually considered to be the “post-exertional malaise” (PEM) experienced by most (69 – 100%) people with the former [1]. This led to the illness’ proposed renaming to “Systemic Exertion Intolerance Disease” (SEID) [1]. PEM is defined as “an inability to replicate an exertion at the second attempt” experienced by people with ME/CFS, and is commonly understood to take the form of an “energy crash” after any kind of physical or mental exertion. This symptom forms the basis of the cardiopulmonary exercise test (CPET) (discussed in **Section 1.2**), and may eventually prove to be a tool for defining research cohorts with ME/CFS.

Other key symptoms include persistent fatigue, not alleviated by rest, and with no clear cause (such as excessive exercise). Cognitive impairment, (typically described as “brain fog”), orthostatic intolerance and sleep dysfunction are also common. These five symptoms sit at the intersection of almost all case criteria [8].

Box 1.1 ME/CFS Case Inclusion/Exclusion Criteria (adapted from Dibble *et al.* [6]).

1. The Fukuda Criteria: developed in 1994 by the International Chronic Fatigue Syndrome Study Group [2]. To be diagnosed with CFS, a person must display unexplained, persistent or relapsing chronic fatigue, as well as at least four of eight additional symptoms. Note that the 1994 CDC Fukuda criteria for CFS is occasionally still used in modern research as an equivalent to ME/CFS, despite this definition lacking the characteristic post-exertional malaise as a core symptom.
2. The Canadian Consensus Criteria [9]. To be diagnosed, people with ME/CFS must exhibit a broad array of symptoms in specific combinations, namely:
 - (a) cardinal symptoms of fatigue, post-exertional malaise or post-exertional fatigue, sleep dysfunction and pain (myalgia, often including headaches);
 - (b) two or more neurological and/or cognitive symptoms from a given list;
 - (c) at least one symptom from two of the following categories: autonomic, neuroendocrine and immune manifestations (with a list of relevant symptoms provided for each); and,
 - (d) symptoms must have persisted for at least six months.

The authors also provided a list of co-morbidities, and another of illnesses that exclude the diagnosis of ME/CFS. These criteria seek to define patients with diverse symptoms, and have been proposed to preferentially select individuals with more severe symptoms [10].

3. International Consensus Criteria [11]: This modifies the Canadian Consensus Criteria by revising criteria, adding others and removing both chronic fatigue as a criterion and the requirement of a six-month period prior to diagnosis. For diagnosis, a patient must display:
 - (a) post-exertional neuroimmune exhaustion (broadly equivalent to post-exertional malaise);
 - (b) three neurological impairment symptoms – at least one each from three of four categories;
 - (c) three immune, gastro-intestinal or genitourinary symptoms – at least one from three of five categories; and,
 - (d) at least one energy production/transportation symptom.
4. USA Institute of Medicine (IOM, now called the National Academy of Medicine [NAM]) Criteria, developed for clinical (not research) purposes [1]. Criteria were streamlined and focussed on core symptoms:
 - (a) substantial reduction from pre-illness activity levels, evidence of post-exertional malaise and unrefreshing sleep; and,
 - (b) cognitive impairment and/or orthostatic intolerance.

1.2 Prevalence, Diagnosis and Biomarkers

Meta-analyses show that ME/CFS is estimated to affect close to 1 in 100 people worldwide (for example, estimates of 0.76% [12] and 0.89% prevalence [8] for Fukuda criteria CFS), but this can vary considerably by study. Examples of estimation by country include 0.2% of people in the UK [13], $\sim 0.5 - 1\%$ people in the United States [14], 0.77% in Korea, and 0.76% in Japan [15]. Among children, paediatric ME/CFS has been shown to have a similar prevalence (0.76%) [16].

ME/CFS preferentially affects women (approximately 60 – 65% of cases) [14]. It has a bimodal age distribution, shown by Bakken *et al.* to have distinct peaks for adolescents (ages 10 – 19) and adults (ages 30 – 39) [17]. The same study found that the second, adult, peak was only present for women. Due to the highly heterogeneous nature of the disease, paediatric ME/CFS is often treated separately in clinical and research settings. In autoimmune diseases, sex biases are hypothesised to be due to hormonal differences (in estrogens and androgens) [18]. This would be consistent with the two age peaks seen for ME/CFS in adolescence and middle age (for women only), which are both correlated with hormonal changes.

The prognosis for those with ME/CFS is extremely poor, with only 7% of those diagnosed making a full recovery [19]. Quality of life for those affected is dramatically reduced, with impairment comparable to many cancers [4][20].

Measurement of prevalence is impaired by the difficulty inherent to diagnosing a disease with neither a clear and consistent definition, nor a reliable biomarker. Physicians will typically diagnose ME/CFS via a process of elimination, ruling out all other diseases with similar symptoms first. They are then able to compare symptoms with the case definitions of **Section 1.1** to provide a positive diagnosis. This is clearly an extremely subjective approach, which likely contributes to the heterogeneity of the current ME/CFS label. Hence, the search for a reliable biomarker (especially one with clinical relevance) forms a major component of current research efforts. It is also very likely that there are distinct subtypes of ME/CFS, which may be determined by identifying clusters of symptoms. This is another promising research area for determining the aetiology of the disease.

There are two proposed clinical tests at present, albeit neither form part of diagnostic criteria. The first is a cardiopulmonary exercise test (CPET) [21]. This protocol requires test subjects to complete an exercise test, typically on a stationary bike, twice, within a short time period (24 – 48 hours, usually). This test will trigger symptoms of post-exertional malaise in people with ME/CFS, and is now consid-

ered a viable method to detect this symptom [22]. An invasive cardiopulmonary exercise test (iCPET), which allows for the collection of further hemodynamics data using arterial catheters, is therefore seeing increased use in research settings (for example, Joseph *et al.* [23]).

There are two major disadvantages to CPET as a diagnostic tool. First, the inconvenience of requiring people with suspected ME/CFS to attend in-person physical testing, which also requires the expense and difficulty of maintaining a controlled laboratory setting for tests. More seriously however, is that the test deliberately attempts to induce PEM symptoms in subjects. This can sometimes lead to a general worsening of disease symptoms, which is a notable ethical concern. However, this approach is becoming increasingly accepted as a potential method to distinguish people with ME/CFS from similar diseases [22].

The other proposed clinical test is the “nanoneedle device” developed at Stanford University [24]. This device measures electrical impedance across a peripheral blood mononuclear cell (PBMC) sample under stress conditions induced by the addition of salt. The researchers found a clear and significant difference between serum and plasma samples from people with ME/CFS compared with healthy controls. Although a promising initial study, the result has neither been validated with a distinct cohort, nor has it been replicated by a separate research group. However, this may eventually prove a useful tool for the diagnosis of ME/CFS.

It is notable that ME/CFS tends to be more prevalent within families, and hence appears to have a heritable component. Heritability is discussed further in **Section 2.1**.

1.3 Treatment

There is no known effective treatment for ME/CFS [1].

For 14 years, guidelines from the UK National Institute of Health and Care Excellence (NICE) promoted Cognitive Behavioural Therapy (CBT) and Graded Exercise Therapy (GET) as treatments for ME/CFS [25]. CBT is a widely used tool in the treatment of depression and other mental health issues and illnesses [26], and conflation of ME/CFS with these led to the proposed use of CBT. GET is a programme that has seen use in rehabilitation from other illnesses [27] and injuries [28], which involves the gradual increase of activity levels with the aim of increasing these over time. This is similar to the “progressive overload” approach used by athletes in

training [29]. However, while evidence suggests that muscular ability is not compromised in ME/CFS, signalling through the central nervous system (CNS) may be affected [30]. Both therapies proved ineffective in the case of ME/CFS, with the failure of GET consistent with this hypothesis, as it is possible that the typical markers of fitness (muscular strength and cardiovascular health) are not affected any more than through deconditioning, but that recruitment of these systems by the CNS is compromised. In other words, there may be no measurable impairment of muscle cells *in vitro*, and the effects of ME/CFS may only be observed during exercise testing.

NICE supported these therapies for ME/CFS based on the work of White *et al.* [31], known as the “PACE Trial”. This study has since been widely discredited, principally for changing the criteria used to class individuals as “recovered”. Wilshire, *et al.* show that statistical standards were changed such that the “recovered” score on the SF-36 scale was set lower than that initially considered the minimum threshold for recruitment into the trial, with a similar modification to the Chalder Fatigue Questionnaire [32]. Because of this, participants could record lower scores after the trial on the health metrics used, and still be classed as recovered. This alteration was so extreme that some participants would have qualified as recovered after recruitment but before the trial commenced (13% using the SF-36 scale, seven people by the modified Chalder Fatigue Questionnaire) [32]. Unsurprisingly, this post-hoc adjustment resulted in a “statistically significant difference in recovery between the CBT and GET groups, as compared with the control group.

Other major concerns reported with the study included no objective measure of recovery (progress was measured through self-reported scores), and how heavily these would be influenced by the lack of blinding in the study design. Even more seriously, those in the GET and CBT arms of the trial were told these interventions were “highly effective”, which almost certainly introduced a degree of bias into the study [32].

Despite these significant flaws, and reports of illness regression among people with ME/CFS prescribed CBT and GET, NICE guidelines were only updated to remove this recommendation in 2021 [33]. A meta-analysis by Vink *et al.* confirmed that CBT and GET result in more patients unable to work after intervention than before [34].

No other trial has reported a treatment that improves outcomes for ME/CFS that has proved robust to replication. Initial results using rituximab to induce B-cell depletion in Phase II trials [35], based on the hypothesis that ME/CFS is an

Study	Year	Inclusion Criteria	% reporting viral infection
Jason <i>et al.</i> [37]	2021	Mixed	60.3
Ghali <i>et al.</i> [38]	2020	ICC	49.2
Chu <i>et al.</i> [39]	2019	1994 CDC (Fukuda)	64
Evans <i>et al.</i> [40]	2018	1994 CDC (Fukuda) - Sudden Onset	73.5
Evans <i>et al.</i> [40]	2018	1994 CDC (Fukuda) - Gradual Onset	66.5
Naess <i>et al.</i> [41]	2010	1994 CDC (Fukuda)	77
Ray <i>et al.</i> [42]	1998	Oxford, plus additional exercise intolerance	93
Salit [43]	1997	1988 CDC (Holmes)	72
		Mean	69.4
		Median	69.3

Table 1.1: Summary of studies reporting a proportion of their subjects with ME/CFS to have a suspected viral onset.

autoimmune condition, were not replicated [36].

1.4 Causes

Although the biological mechanism which triggers ME/CFS is unknown, studies commonly show that more than half of people with ME/CFS attribute the cause of their disease to be a viral or bacterial infection, with some estimates over 90% (average across studies $\approx 69\%$, **Table 1.1**).

Infectious mononucleosis (glandular fever, triggered by the Epstein-Barr virus [EBV]) is the most commonly reported viral trigger [37][40], but others implicated include human cytomegalovirus (HCMV) and other human herpesviruses (HHVs, for example, Human Herpesvirus-6 [HHV-6]) [44]. Herpesviruses may be particularly relevant due to their ability to enter a “latent” phase, with no viral replication for a period of time, which would be consistent with periods of ME/CFS remission [45]. Of the eight herpesviruses which frequently affect humans (EBV, HCMV, Herpes simplex virus 1 and 2, HHV-6, HHV-7, HHV-8, and varicella-zoster virus) [46], sporadic research has suggested that any of them may be implicated in ME/CFS, but the evidence is inconsistent [45]. A review of suggested causal viruses is provided by Rasa *et al.* [47]. EBV has also been strongly implicated as the trigger of multiple sclerosis [48]. Supporting this is a significant subset of the research literature, where studies have found alterations to levels of natural killer (NK), B-, and T-cells, as well as altered cytokine levels [49]. See **Section 2.2** for further details on research into the immunology of ME/CFS.

Other potential triggers of ME/CFS reported by the study of Chu *et al.* [39] include acute stress (39% of respondents) and environmental triggers (20%). The latter implies a potential link with the similar, and equally poorly-understood, “Gulf war

syndrome” reported by veterans of the 1991 war. As yet unproven hypotheses for the onset of this disease include environmental triggers, such as the substances used in chemical warfare (see **Section 1.5.7**).

1.5 Similar and Co-morbid Diseases

ME/CFS is frequently diagnosed alongside many other co-morbidities, for which there is often considerable symptom overlap [50]. A selection of the most common co-diagnoses are presented in this section.

1.5.1 Fibromyalgia

There is considerable overlap between the definitions of ME/CFS and other pain and fatigue disorders – most notably with fibromyalgia. Fibromyalgia is a pain disorder of unknown aetiology, but which likely involves a degree of nervous system dysfunction [51]. The syndrome is defined by Kodner [52] as:

“Fibromyalgia is caused by central amplification of peripheral sensory signals, so that normal sensations are perceived as painful.”

It was previously diagnosed by examination of tender points, but newer criteria bring the disease even closer to the definitions used for ME/CFS. The principal difference is that fibromyalgia may improve with exercise [52], whereas this will usually trigger PEM for people with ME/CFS.

In **Section 3.1.1** I consider and discuss the diagnosis and overlap of CFS and fibromyalgia in the UK Biobank, and find that while the literature considers them to be commonly co-morbid (15.6% of people with CFS met fibromyalgia criteria according to Jason *et al.* [53], albeit with a small CFS sample), the intersection of the two groups in these data is actually quite small. This could be due to physicians preferring to diagnose one or the other, as they are still regarded as different diseases with potentially distinct causal mechanisms. An alternative explanation is simply the lack of power in the UK Biobank to detect the genuine intersection between these diseases, where more cases would show greater co-incidence.

1.5.2 (Chronic) Lyme Disease

Lyme disease is caused by *Borrelia* bacteria (typically *B. burgdorferi* in North America, and additionally *B. afzelii* and *B. garinii* in Europe [54]), typically transmitted via tick bite [55]. A characteristic skin lesion (Erythema migrans) typically develops around the bite location, aiding diagnosis. Symptoms progress to include secondary skin lesions, fatigue, myalgia, fever and headaches. Symptoms typically clear within two to four weeks with antimicrobial treatment, but late Lyme Disease, where symptoms persist for weeks or months after infection, is also possible. The most common symptom of late-stage infection is arthritis, and it is established that the immune system plays some role in this [55]. In particular, a small study [56] has established that certain *HLA-DRB* alleles were more common among those with antibiotic resistant arthritis after Lyme infection, similar to other autoimmune diseases. However, this study fails to account for any multiple testing correction across the hundreds of alleles, and so it is likely that this association would not be replicated in a different cohort. More recently, strong associations have been detected between infection risk and alleles within the *HLA-DQB1* locus using GWAS [57] (see **Section 2.1.3** for details on GWAS).

Although the progression and incidence of Lyme disease are well understood, with a known vector and cause, there are some who do not respond to antibiotic treatment and continue to experience symptoms long after infection. This disease is usually referred to as “Chronic Lyme Disease”, and has many symptoms in common with post-viral diseases [58]. Researchers sometimes have difficulty in establishing cohorts that have a clear history of Lyme disease, and symptoms persisting beyond six months after infection, underlining that many of those with Chronic Lyme Disease may have been misdiagnosed. To emphasise this, the research literature does not support the hypothesis of persistent *Borrelia* infection [58].

Treatment for Chronic Lyme Disease suggested by some physicians involves repeated courses of antibiotics, however this is not supported by the research literature [59].

1.5.3 Post-Covid Syndrome (Long Covid)

A similar post-viral syndrome is common after infection with SARS-CoV-2 [60]. The UK Office of National Statistics (ONS) published Infection Survey data showing that 3.1% of people in the UK self-reported symptoms of Covid-19 more than

four weeks after suspected or confirmed infection [61]. 42% of respondents reported symptoms still present despite infection occurring at least 12 months previously. Symptoms appear to overlap with ME/CFS – particularly fatigue (55% of respondents) and difficulty concentrating ($\sim 21\%$). Halpin, *et al.* also noted a degree of exercise intolerance in their study [60]. However, loss of taste is not a core symptom of ME/CFS (18% of respondents), indicating that at least a subset of people with Long Covid may be distinguishable from ME/CFS. However, some of the risk factors overlap (in particular, Long Covid is more common among females [61]), so while it is currently unknown whether this disorder is distinct from ME/CFS, it is plausible that it may be another form of this heterogeneous disease.

Viral triggers of ME/CFS are common, and there is currently no reason to believe that SARS-CoV-2 would not lead to a similar outcome as the previously mentioned viruses (see **Section 1.4**). It is plausible that acute and Long Covid act through fundamentally different pathways. This is consistent with the fact that not all cases of Covid-19 lead to Long Covid, as is also the case for the proposed viral incidences of ME/CFS. One further consequence of the 2020 pandemic is that the consistent, typically traceable, incidence of Long Covid has created a cohort with a complex disease that may prove useful for future research into ME/CFS.

1.5.4 Postural Tachycardia Syndrome

Postural Tachycardia Syndrome (POTS) is a syndrome with significant symptom and onset overlap with ME/CFS, and in some cases diagnosed as a co-morbidity (reported at 11% of cases in one study [62]). It is a heterogeneous syndrome, but typical diagnosis is via testing of heart rate and blood pressure in different postures to establish the presence of orthostatic intolerance (such as a tilt-table test) [63]. However, the symptom overlap is notable – orthostatic intolerance is common to almost all definitions of ME/CFS (**Section 1.1**), and 73% of people with POTS report fatigue as an additional symptom [64]. It is, however, still typically considered a distinct syndrome.

Risk factors are also similar to ME/CFS - females account for the majority of cases (female to male ratio $\sim 4 - 5 : 1$). Age is also a factor, with typical onset between the ages of 15 – 50. It is also common for onset to occur after viral infection (50% of cases) [64].

1.5.5 Migraine Syndromes

Migraine syndromes [65] are frequently seen alongside and diagnosed co-morbidly with ME/CFS [66][67]. Like ME/CFS, they are more common among women (3 : 1 female to male ratio), and have a heritable component [68]. Migraine is typically considered to be a neurovascular disorder, with symptoms caused by dilation of blood vessels in the brain [65].

Although the headache symptoms of ME/CFS are similar, migraines fall into two distinct subtypes of with and without aura. The migraine aura is experienced by $\sim 30\%$ of migraineurs as (typically visual) neurological symptoms. Characteristic symptoms which separate migraines from tension headaches include photophobia and audiophobia (light and sound sensitivities), nausea and vomiting [65]. Migraines are also sometimes preceded by a prodrome phase, where common symptoms (such as fatigue) precede an acute headache phase [68].

As a common (prevalence $\sim 15 - 18\%$ [65][68]) complex disease with a genetic component, and symptom overlap with ME/CFS, further understanding of the mechanisms of migraines may aid research into the aetiology of ME/CFS.

1.5.6 Irritable Bowel Syndrome

Irritable Bowel Syndrome (IBS) [69] is a chronic condition of bowel dysregulation, distinct from Inflammatory Bowel Disease (IBD) [70]. There is a significant symptom overlap with the gastrointestinal symptoms of ME/CFS (see **Section 1.1**), and the diagnoses are often co-morbid ($\sim 51\%$ of people with CFS also have IBS, according to Whitehead, *et al.* [71]). The prevalence ($\sim 15\%$), 3 : 1 female to male ratio [69], and common reports of an infectious onset ($\sim 10\%$ of cases) [72] are in common with the other complex diseases discussed in this section.

As with those diseases, very little is understood regarding the aetiology of IBS [69].

1.5.7 Gulf War Syndrome

Gulf War Syndrome is a fatiguing illness of unknown aetiology seen among veterans of the 1991 Persian Gulf War [73]. Its symptoms overlap substantially with ME/CFS, and the two diseases are sometimes studied in tandem [74].

Exposure to chemical weapons or depleted uranium, or a reaction to the many vaccinations given to troops prior to deployment have all been proposed as the cause of Gulf War Syndrome [73]. These can be partially refuted by the widespread incidence among members of multiple nations' armed forces, each of which administered different vaccinations and were exposed to different environmental hazards [73]. Stress, particularly combat stress, or chronic stress related to chemical weapon attacks, could be considered an underlying common factor [73]. However, these stressors are not unique to this conflict, and there is no recognition of a similar "2003 Iraq War Syndrome", where troops had reason to believe the opposition possessed nuclear weapons or technology.

A recent, more comprehensive study by Haley *et al.* [75] demonstrated a potential gene-environment interaction between a mutation in the gene *PON1*, and low-level exposure to sarin gas, among US military veterans of the war. However, this has yet to be replicated in a cohort from another nation's military that also participated in the campaign.

1.5.8 Mental Illnesses

Symptoms of ME/CFS will often overlap with mental illnesses, such as depression. These illnesses are also often co-morbid with ME/CFS [76][77], most notable for mood disorders and anxiety [78]. Finding a clinically useful biomarker for ME/CFS would substantially aid differentiation of the disease from mental illnesses.

For a significant period of time since the disease was identified, ME/CFS was treated as a mental illness, where symptoms were characterised as originating from either a mental health disorder or from physical deconditioning. Under the psychosocial model, fatigue was considered to be perpetuated by psychological factors such as "focusing on symptoms" [79]. The related biopsychosocial model considers there to be a biological origin, but a similar pattern of psychological and additional biological factors which prolong fatigue [80]. Both models are now considered to be an incorrect interpretation of the disease [81], however, it was within this framework that most research was conducted, including the now discredited PACE trial (see **Section 1.3**). Researchers now widely acknowledge the role of an as yet unknown biological pathway that distinguishes ME/CFS from mental illnesses [81].

However, this does not mitigate the profound mental health impacts of ME/CFS. There are two proposed models for this interaction. The first is that mental illnesses

are a symptom of ME/CFS, where the biological pathway that triggers PEM also triggers the same effects as major depressive disorder, anxiety or similar. In this case, mental health issues could either be considered within the diagnostic criteria of ME/CFS, or as separate, but highly co-morbid diseases.

The second model is that mental illness such as depression follows the onset of ME/CFS. It is known that traumatic events can lead to mental health disorders (for example, illness leading to admission to an intensive care unit [82]), and as noted, ME/CFS symptoms cause a profound reduction in quality of life (see **Section 1.2**). Under this hypothesis, the distinct disease ME/CFS would act as the trigger for a subsequent decline in mental health, or the diagnosis of a mental illness.

It is currently unknown which of these two hypotheses best describes the relationship between ME/CFS and mental illnesses.

1.5.9 Multiple Sclerosis

Multiple sclerosis (MS) is an autoimmune condition mediated by T- and potentially B-cells, which presents as lesions in the brain and spinal cord [83].

The exact causal pathway that leads to MS is currently unknown, and many genes have been associated with the disease (using GWAS, see **Section 2.1.7**) [83]. In particular, in common with other autoimmune diseases, there is a strong association with an HLA allele (*HLA-DRB1*15*). Heterozygotes for *HLA-DRB1*15:01* have an average $OR = 3.08$ (see **Section 2.1**) for MS with an additional allele increasing the risk under an additive model [84] (typical for genetics studies, see **Section 3.1.4**). There are also another 150 single nucleotide polymorphisms with more modest odds ratios ($OR \approx 1.1$ or 1.2) linked to MS [83].

Alongside EBV infection, which increases the risk as much as 32-fold, and may in fact be causal [48], other risk factors are well understood. Smoking, sunlight exposure and vitamin D levels [85] are all known to be associated with the risk of developing MS [83].

Although distinct from ME/CFS, as a complex autoimmune disease with many associated genetic loci, people with MS will act as disease controls in the immunological study of **Chapters 4** and **5**.

1.5.10 Guillain-Barré Syndrome

Guillain-Barré syndrome is a rare (on the order of 0.001% prevalence [86]) autoimmune syndrome, caused by activation of the immune system in response to an external pathogen (typically *Campylobacter jejuni*, EBV or cytomegalovirus) [87]. It is hypothesised to be the result of self-targeting of the immune system against nerve cells. There are four known subtypes [86], of which the most common among those of European ancestry is Acute Inflammatory Demyelinating Polyradiculopathy (AIDP), where damage is restricted to myelin sheaths [87]. Other subtypes are known to cause damage to axons, which repair even more slowly than myelin [87].

It has a distinct pathology from other autoimmune diseases, where the attack on nerves is proposed to be time-limited, and further symptoms are the result of the slow rate of repair of nervous tissues, rather than ongoing immune system activation [87]. The syndrome is severe (25% of cases require hospitalisation for assistance with breathing [86]), and leads to death in 4 – 15% of cases. Hence, it has similarities, but also quite pronounced differences, with the other complex diseases discussed in this section.

Chapter 2

Areas of Current ME/CFS Research

This chapter surveys and reviews the research literature on ME/CFS in two key areas: genetics and immunology. **Section 2.1** draws heavily upon the published review “Genetic risk factors of ME/CFS: a critical review” [6], in which I reviewed and analysed the ME/CFS genetics literature (co-authored with S.J. McGrath and C.P. Ponting). In particular; **Sections 2.1.4** and **2.1.5** and **Section 2.1.6** are updated extracts from this review.

2.1 Genetics

As previously mentioned, ME/CFS appears to have a heritable component, which implies there is merit to investigating a potential genetic aetiology. There are many ways to measure heritability and relatedness in different contexts. Common study designs include twin studies (which control for environmental effects by investigating monozygotic [identical] and dizygotic [fraternal] twins), family studies (which investigate pedigree), and the use of unrelated populations (where genetic ancestry is most prominent) [88]. Measurement of heritability varies with each of these.

2.1.1 Relatedness

Individuals within populations will typically be related in some way, whether through immediate families or shared ancestry. To quantify relatedness, shared genetic information can be investigated. Two haplotypes are identical by state if their sequences are identical [89]. Two haplotypes are identical by descent if they were also obtained from a common ancestor [90]. Identical by descent haplotypes are a subset of those identical by state [89].

A definition of relatedness between two individuals x and y follows:

$$r_{xy} = 2\theta_{xy} \quad (2.1)$$

where θ_{xy} is the coefficient of ancestry [91]. This is the probability that an allele [92] at any locus is identical by descent between individuals x and y . Extending this concept, one can find the proportion of the genome that is identical by descent between two individuals, and use this as the measure of relatedness. Estimating the degree of relatedness between every individual in a dataset, and forming a matrix of the r_{xy} , generates a genetic relationship matrix (GRM). This matrix $\mathbf{K}$ will capture both family relations and wider genetic ancestry.

An alternative in some models is to provide some account of the relatedness in a population by incorporating a number of principal components as covariates (see **Section 3.1.4**) [93]. This will typically account only for stratification by genetic ancestry, and will not usually correct for family structure or unseen (“cryptic”) relations [94].

For diseases, relatedness leads to questions of the risk of inheriting a trait from parents.

2.1.2 Heritability

A typical model of heritability assumes that a phenotype is some combination of individual genotype and environmental factors. From Tenesa *et al.* [88]:

$$P = G + E_T \quad (2.2)$$

where P is the phenotype, G is the genetic component, and E_T is the total

environmental component. G can be considered as a combination of an additive component, dominance component and interaction component:

$$G = A + D + I \quad (2.3)$$

and E_T to be a combination of common environment and a more general environment term:

$$E_T = C + E \quad (2.4)$$

Since variance is an additive quantity, using **Equation 2.2**, one can write:

$$V_P = V_G + V_{E_T} \quad (2.5)$$

where V is the variance (standard deviation squared $V = \sigma^2$ [95]). Through the equivalents of **Equation 2.5** for **2.3** and **2.4**, the phenotypic variance can then be written in terms of these components:

$$V_P = V_A + V_D + V_I + V_C + V_E \quad (2.6)$$

With these definitions, one can define “broad-sense heritability”, H^2 , to be the variance explained by genetics irrespective of environment:

$$H^2 = \frac{V_G}{V_P} \quad (2.7)$$

This is most often considered in twin studies, and is the most useful for predicting the disease risk for an individual based solely on their genome [88]. The “narrow-sense heritability”, h^2 , is defined:

$$h^2 = \frac{V_A}{V_P} \quad (2.8)$$

and indicates the disease risk incurred from an individuals' parents [88]. The broad-sense heritability is an upper bound on the narrow-sense [95]:

$$h^2 \leq H^2 \quad (2.9)$$

and they will be equal only if the dominance and interaction components of the variance are zero ($V_D = V_I = 0$).

Other types of models such as ADE and ACE consider the relevant subset of variance components in **Equation 2.6**. For ACE models, Falconer's Equations are a well-known method of estimating heritability in twins [96]. Under the ACE model for monozygotic and dyzygotic twins [97]:

$$h^2 = \frac{V_{AMZ}}{V_{AMZ} + V_{CMZ} + V_{EMZ}} = \frac{V_{ADZ}}{V_{ADZ} + V_{CDZ} + V_{EDZ}} \quad (2.10)$$

where the heritability is assumed to be equal in both cases (but not necessarily the individual variance components). Falconer's estimate of the heritability is then:

$$\hat{h}^2 = 2(r_{MZ} - r_{DZ}) \quad (2.11)$$

where r_{MZ} and r_{DZ} are the correlation coefficients for the samples of monozygotic and dizygotic twins, respectively [97].

It is also possible to consider restricting the additive variance component to that contributed by a subset S of observed single nucleotide polymorphisms (SNPs), which are single base changes in genomic DNA.

$$h_g^2 = \frac{V_{SNPs \in S}}{V_P} \quad (2.12)$$

This is known as a "SNP-based estimate" [95]. SNPs are the typical unit of consideration in GWAS, as discussed in **Section 2.1.3**. This estimate is realistic for data that was not generated using whole-genome sequencing, and it is bounded above by the narrow-sense heritability, such that:

$$h_g^2 \leq h^2 \leq H^2 \quad (2.13)$$

as the additive term will typically be smaller for h_g^2 , given it incorporates only a subset of the genomic variants [95].

Finally, a more recent approach is to use the previously defined genetic relationship matrix in a linear mixed model to estimate heritability. First described by Yang *et al.* [98], this uses the model:

$$\mathbf{y} \sim MVN(\mathbf{M}, \mathbf{K}V_g + \mathbf{I}V_E) \quad (2.14)$$

where the phenotype vector $\mathbf{y}$ is modelled as a multivariate normal distribution, $\mathbf{M}$ is the mean vector including the covariates, $\mathbf{K}$ is the genetic relationship matrix, and $\mathbf{I}$ is the identity matrix. $\mathbf{K}$ can be estimated using genetic covariance [92]. This method is typically used for family and pedigree studies [88], as well as in unrelated populations [98].

2.1.3 ME/CFS Heritability

The narrow-sense heritability of ME/CFS has been shown to be non-zero, using three independent cohorts. Narrow sense heritability refers only to the additive component of heritability, but it has been shown to be the most useful for determining the inheritance of traits from parents [88]. Lakhani *et al.* estimated narrow-sense heritability to be $h^2 = 0.476$ using a dataset derived from US health insurance claims [99]. However, the GeneATLAS project estimated narrow-sense heritability in the UK Biobank dataset to be considerably lower ($h^2 = 0.0816$) [100]. Further, a study of Swedish twins concluded that narrow-sense heritability was 0.03, with a confidence interval including zero ($CI = [0, 0.65]$), however this study did not use formal CFS diagnoses [101]. Hence, more research is required to conclude that ME/CFS is heritable. Non-zero heritability would imply a genetic component to the disease, which is discussed in this section.

Genome-wide Association Studies (GWAS) attempt to identify associations between a phenotype (a measurable characteristic) and variation in the genome [102], where the typical unit of consideration is a single nucleotide polymorphism (SNP). Genotyping can be performed on the whole genome [103], or by sampling SNPs across the genome [102]. If the latter, SNPs are not sampled randomly, but selected to account for linkage disequilibrium (LD). LD refers to the fact that the probability of possessing a given variant is not independent – they are typically inherited in blocks of correlated SNPs (haplotypes). This correlation is useful for assessing whether associations are true positives, or technical artefacts, as for the former, one expects to see other nearby SNPs with decreased p -values. This can be seen by considering two SNPs in perfect LD, where if one is associated with a phenotype, then we would expect to see the same association for the second SNP (with an identical p -value). For less than perfect LD, it therefore follows that the second SNP will also be associated, but with an increased p -value.

LD is typically measured with D' [104], an adjusted form of D :

$$D = p_{AB} - p_A p_B \quad (2.15)$$

where p_A and p_B are the frequencies of alleles A and B (for biallelic loci A/a and B/b), and p_{AB} is the frequency of the two-locus haplotype A and B [105]. Because D is calculated from allele frequencies, rarer and more common alleles may lead to different values for the same degree of LD. D' normalizes the measure as a fraction of its maximal value to maintain consistency across allele frequencies:

$$D' = \frac{D}{D_{max}} \quad (2.16)$$

Other common choices for measurement of LD (such as r^2) follow the same format of:

$$\frac{f(D)}{g(p_A, p_B)} \quad (2.17)$$

where $f(D)$ is a function of D , and $g(p_A, p_B)$ is a function of the individual locus frequencies [105].

In data not from whole-genome sequencing, the missing SNPs can be imputed – inferred with probability of inclusion. In the studies presented in this section, the data discussed are from imputations or exclusively from the genotyping. For genotyping microarrays, the chosen SNPs are common in the general population. A measure of this is the minor allele frequency (MAF), which for a biallelic SNP reports the frequency of the rarer of the two variants. A MAF filter of > 0.05 [106] is typical in GWAS when investigating common diseases (rare diseases typically have prevalences $< 0.1\%$ [107]), as it is implausible that a rare variant would cause a disease with greater prevalence in the general population (the “Common disease, common variant” hypothesis) [108][109]. While it is plausible that a rare variant may be a risk factor for a complex disease [110], in the case of ME/CFS, this is a secondary research goal to finding the main set of genetic variants that explain most of the risk. Rare variants must be studied using genotyping arrays or sequencing technologies that specifically include them ($MAF \lesssim 0.0012\%$ for microarray studies) [110].

The effect of SNPs on a trait can be inferred from the odds ratio (OR). This

is the ratio comparing the odds that an event occurs, to the odds that it does not, where the odds are defined as the ratio between the probability that an event occurs, and the probability that it does not [111]. Hence, in the case of genetics, the OR represents the changed odds of having a phenotype, compared with not, given a specific allele [112]. A related concept is the relative risk (RR) which, intuitively, represents the increase in risk (having a phenotype) conferred by a SNP as a multiplier, compared with not having that allele. Note that these are similar, but not interchangeable [113].

To account for the large number of tests performed in a GWAS, an adjusted threshold for p should be used. In the population genetics literature, this is typically taken to be $p < 5 \times 10^{-8}$, based on permutation testing by the International HapMap Consortium [106][114]. A “suggestive” threshold is sometimes considered to be $p < 1 \times 10^{-5}$ [115]. The appropriateness of these thresholds varies based on the number of tests performed in a given analysis, but $p < 5 \times 10^{-8}$ has been shown to be robust for $MAF \geq 0.05$ in European ancestry cohorts [106].

The phenotypes tested can be continuous (such as height), or categorical (typically binary, such as positive or negative for a disease), with different statistical models used for each. The inverse of a GWAS is a phenome-wide association study (PheWAS), which compare multiple phenotypes against a single SNP [116].

2.1.4 ME/CFS and UK Biobank

The UK Biobank provides a cohort that has been genotyped and exceptionally well phenotyped and hence is well suited to GWAS [117]. One of the over 750 anthropometric and disease-related traits captured by the UK Biobank is self-reported clinical diagnosis of CFS. This generated the largest cohort of people collected to date who have self-reported a clinical diagnosis of CFS (2,281 people). There is also an ME/CFS phenotype in the UK Biobank, for which participants were asked whether they had ever received a clinical diagnosis of ME/CFS or CFS. There are 2,721 people in the UK Biobank with this phenotype. It is unknown how many of these participants meet the ME/CFS criteria of **Section 1.1**, as they are self-reporting their diagnosis; hence, all results discussed below are preliminary and, as ever, require statistically robust replication. The UK Biobank phenotypes, and their overlaps with the fibromyalgia equivalents, are discussed extensively in **Section 3.1.1**, with replication addressed in **Section 3.1.5**.

Six groups have performed a case-control GWAS on CFS cases in the UK Biobank.

DNA Variant	Chromosome	Nearby Gene	Allele Frequency (GnomAD)	p (Neale) Female	p (Neale) Male	p (Neale) Both	p (SAIGE) Both	p (GeneAtlas) Both	p (Global Biobank Engine)	p (Pan-UKBB)	p (Hajdarevic) Both
rs7337312	13	SLC25A15	0.54	2.6×10^{-8}	0.74	4.0×10^{-6}	0.29	0.00017	9.3×10^{-5}	2.1×10^{-4}	not reported
rs150954845	10	PAHA1	0.00029	not reported	not reported	not reported	0.37	2.6×10^{-12}	not reported	not reported	not reported
rs148723539	10	EBF3	0.0093	9.7×10^{-7}	6.1×10^{-4}	2.3×10^{-9}	0.29	not reported	not reported	3.7×10^{-6}	not reported
rs564809936	4	COX7B2	0.00016	not reported	not reported	not reported	3.8×10^{-8}	not reported	not reported	4.8×10^{-3}	not reported
rs77381650	6	NOX3	0.019	2.0×10^{-3}	8.4×10^{-2}	4.2×10^{-4}	not reported	4.1×10^{-4}	not reported	0.83	4.4×10^{-8}

Table 2.1: Summary of SNPs identified as significant in the UK Biobank CFS Cohort [6] (updated to include the finding of Hajdarevic et al. [118]).

Unfortunately, they reach no consensus and their results are far from being definitive. One study found no DNA variant to pass the $p < 5 \times 10^{-8}$ threshold typically used in GWAS (**Section 2.1**) [119]. Another study reported results both partitioned and unpartitioned by sex [120], finding one variant associated with female CFS status and another with male or female CFS status (**Table 2.1**). Three other studies [100][121][118] highlighted a single variant each (**Table 2.1**). The final study [122] found no variants associated with either the Biobank CFS phenotype (**Table 2.1**), or the “phecode” definition of CFS [123]. Consequently, despite these six studies analyzing the same data set, not a single associated DNA variant was replicated by multiple analyses. Two explanations of this lack of replication are likely. First, as alleles become rarer the likelihood increases that all people with the minor allele are placed among the cases just by chance. Three of the highlighted SNPs that are very rare in the population ($MAF < 0.5\%$) are possible examples of this. Second, DNA sites that appear to have three or more alleles are usually excluded from analyses because they likely reflect technical genotyping artefacts and hence may require more careful analysis [124]. One of the highlighted variants (rs148723539) is one such multi-allelic site (**Table 2.1**), in addition to being rare ($MAF = 0.0093$). These factors, combined with each study using slightly different statistical analyses, are the likely explanations for this inconsistency.

The variant rs7337312 is neither rare ($MAF \approx 0.5$) nor multi-allelic and was identified in the female-only CFS GWAS by Neale et al. [120]. This association was only found in females, and was not replicated in the full or male only cohorts, but as there are more females than males with the CFS phenotype, this may be due to the corresponding greater power to detect associations due to larger sample size. This variant, together with adjacent significant variants that are in linkage disequilibrium, occur within a 51kb region containing the *SLC25A15* gene (**Figure 2.1**). Recent analysis of whole exome sequencing from the UK Biobank has shown that proximity to the sentinel variant is a strong predictor of association for a gene, and hence the closest gene is the most likely to be affected by DNA variation [125][126]. However, in some cases, the gene affected by DNA variation is not the

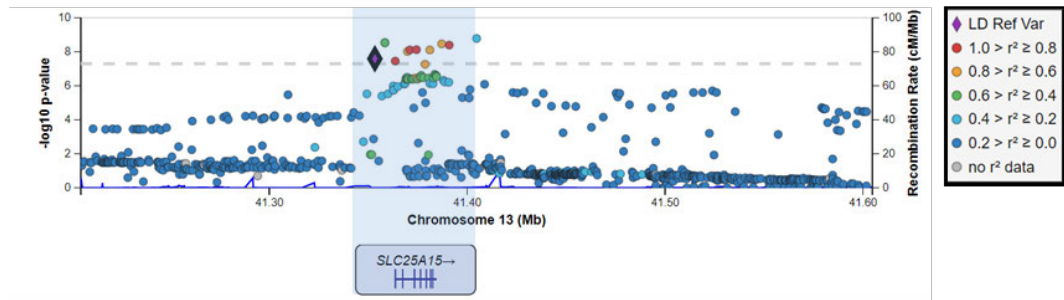


Figure 2.1: An ME/CFS-associated GWAS locus on chromosome 13 (x-axis). Genome-wide significant SNPs from Neale *et al.* [120] are those lying above the horizontal dashed line. The location of the *SLC25A15* gene is indicated below (exons are shown as vertical bars). The left-hand y-axis reflects the statistical confidence ($-\log_{10} p$ -value) of the association between a DNA variant and ME/CFS case status. The right-hand y-axis (blue data curves) indicates the low extent of recombination within this locus. *rs7337312* is highlighted as the reference variant (purple diamond). The degree of linkage disequilibrium of *rs7337312* with neighbouring SNPs is indicated as r^2 (right-hand legend).

nearest [127][128] because of long-range genetic regulation [129]. Nonetheless, because the *rs7337312* variant is an expression quantitative trait locus (eQTL) for this gene, it predicts the amount of *SLC25A15* mRNA transcribed in some tissue samples, *SLC25A15* could be a causal gene of altered CFS risk. *SLC25A15* encodes the Ornithine Transporter type 1 protein that transports ornithine (as well as lysine and arginine) across the inner membrane of mitochondria to the mitochondrial matrix. Ornithine is an amino acid that plays a role in the urea cycle. A person with the *rs7337312* CFS risk allele is expected to produce lower amounts of *SLC25A15* mRNA, as this SNP is an eQTL for this gene, resulting in reduced transport of ornithine into the mitochondrion and higher amounts of ornithine in blood. Studies have reported some evidence in support of this prediction, including Yamano *et al.* [130], who found the ornithine/citrulline ratio to be elevated in plasma from people with CFS, and Naviaux *et al.* [131], who found plasma arginine levels to be increased in both males and females with CFS. Both of these results are consistent with reduced ORNT1 protein levels. However, Armstrong *et al.* [132] found that ornithine levels were reduced, albeit with a very small sample size, and without correcting for multiple tests. Hence, this is still the most plausible association identified in the literature, and so warrants further investigation (see **Section 3.1.7**).

Using the same UK Biobank data, Aguirre *et al.* [133] tested for association with copy number variations (CNVs, the number of duplications of sections of the genome [134]), finding two genes (*TCOF1* and *THUMP2*) to have a greater number of CNVs (either gains or losses) in the UK Biobank CFS cases than in controls after applying a multiple-testing significance threshold of $p < 3.1 \times 10^{-6}$.

Nevertheless, these results need to be treated with caution owing to CNV calling artefacts being prevalent at the low allele frequencies ($< 0.1\%$) used in the study.

Another study performed a PheWAS on the exome sequencing data released by the UK Biobank in December 2019 to investigate the contribution of rare variants to common diseases [110]. As noted in **Section 2.1**, it is not anticipated that a rare variant, or set of rare variants, will prove the primary risk factor for ME/CFS, given its prevalence. In agreement with this, the study found no associations with the UK Biobank CFS phenotype.

2.1.5 Studies Not Using the UK Biobank Data

Smith *et al.* [135] undertook a GWAS on CFS (defined using Fukuda criteria) for very low numbers (40 cases and 40 non-fatigued control subjects) and reported 65 DNA variants as being associated at a non-standard highly permissive threshold of $p < 10^{-3}$. When the well-established threshold of $p < 5 \times 10^{-8}$ is applied, however, none remain significant. This is expected because, as discussed above, GWAS using this number of cases are only well powered to identify population-frequent alleles with strong effects.

A second GWAS also used similarly small cohort sizes (42 cases, defined using both Canadian and Fukuda criteria, and 38 controls) and reported 299 DNA variants as associated with ME/CFS status at $p < 1 \times 10^{-5}$ [136]. The authors justified this permissive threshold as being an inclusion criterion of the GWAS Catalogue [137] (see **Section 3.1.4**). Nevertheless, $p < 1 \times 10^{-5}$ is used by the GWAS Catalogue for reporting purposes only, and only for results from the overall (initial GWAS plus replication) population, when this study lacked a replication cohort. Fifteen variants were reported as genome-wide significant at $p < 5 \times 10^{-8}$, but these associations are not replicated in the UK Biobank cohort (**Figure 2.2**).

Perez *et al.* [140] conducted a genetics study using 383 people who self-reported a clinical diagnosis of ME/CFS (Fukuda criteria). The study only considered DNA variants that disrupted genes despite $> 90\%$ of variants associated with diseases or traits lying elsewhere in the genome [141]. Using non-standard thresholds on variant frequency, the authors reported 5,693 DNA variants that are two-fold more (or less) frequent in cases versus controls. However, such case-control allele frequency inequalities could have two technical, rather than biological, explanations. Firstly, errors stemming from poor DNA sample quality, incomplete DNA hybridization to the genotyping array or poorly performing array probes. Controlling for these

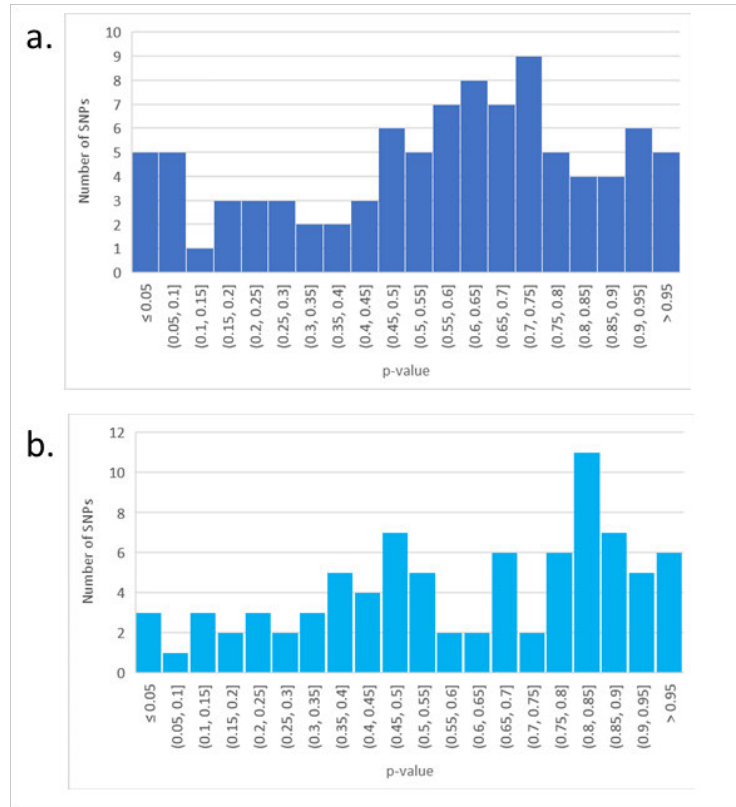


Figure 2.2: Histograms of replication p -values for association of SNPs for CFS risk. These p -values were obtained from the UK Biobank CFS GWAS of Canela-Xandri et al. [100]. The variants tested had been previously identified as associated with CFS by published analyses in separate datasets. Only variants tested in the UK Biobank were considered. No evidence of a skew towards lower p -values was observed.

A) Association p -values for SNPs identified by Wang et al. [138].

B) Association p -values for SNPs identified by Marshall-Gradisnik et al. [139] and Schlauch et al. [136].

errors imperfectly, or inconsistently between case and control genotypes, has led to retraction of publications reporting genetic associations (for example, Sebastiani *et al.* [142]). Secondly, errors arising because confounding effects from the controls such as differences in genetic ancestry were not accounted for.

Hajdarevic *et al.* [118] incorporated their UK Biobank analysis (**Section 2.1.4**) into a larger project. 427 Norwegian people with ME/CFS acted as a discovery cohort, with a 460 person Danish group, and their UK Biobank dataset acting as replication sets. This is a non-standard design for GWAS, and none of their findings using the non-UK Biobank cohorts reached genome-wide significance.

2.1.6 Candidate Gene Studies

A review by leading human geneticists stated that: “initial efforts targeting variants within ‘candidate’ genes were plagued by inadequate power, unduly liberal thresholds for declaring significance and scant attention to sources of bias and confounding, resulting in overblown claims and failed replication” [143]. Their confidence stems from knowledge that tens of thousands of associations have been identified by GWAS that subsequently are often replicated independently [137]. An example of a candidate gene approach in ME/CFS research is a study of nine DNA variants in the gene *NR3C1* for 43 people with CFS and 58 controls [144]. Four of these variants (rs1866388, rs2918419, rs860458 and rs6188) passed their statistical threshold for significance ($p < 0.05$). Nevertheless, none of these variants survive a Bonferroni multiple testing correction ($p < \frac{0.05}{9}$ or < 0.006) and none are replicated with the UK Biobank CFS cohort GWAS of Canela-Xandri *et al.* [100] ($p = 0.96, 0.71, 0.71$ and 0.97 , respectively).

To generalize this point, we again exploited the large cohort of self-reported CFS participants of the UK Biobank. We tested for replication a large set of genetic findings reported in 16 CFS studies published between 2003 and 2015 that met the systematic review criteria of Wang *et al.* [138]. Looking up associations between DNA variants (**Table 1** of reference [138]) and CFS status in the Canela-Xandri *et al.* GWAS of the UK Biobank [100] for the 11 studies that had readily available SNP references yielded a p -value distribution between 0 and 1 (**Figure 2.2**). Replication would require these p -values to be skewed towards small values. No such skew is observed (**Figure 2.2**), consistent with random samples; hence, these initial findings show no evidence of being replicated.

This conclusion is further substantiated from plotting the replication p -values for

a further set of 77 DNA variants from Marshall-Gradisnik *et al.* [139], and 23 from Schlauch *et al.* [136]: these p -values also show a lack of skew towards small values (**Figure 2.2**). Hence, it is clear that candidate gene studies are typically insufficiently robust to false positive results. Although GWAS are not completely hypothesis-free [145], their more agnostic approach should be preferred, even where there is prior evidence of association with a particular genomic region, to ensure an appropriate threshold for false positives is used.

2.1.7 Autoimmunity (Genetics)

Associations have been found between the human leukocyte antigen (HLA) region of the genome and the ME/CFS phenotype [146]. The major histocompatibility complex (MHC) region (HLA in humans) on chromosome 6 encodes the MHC molecules used by antigen-presenting cells (APCs), which present pathogenic peptides to immune system cells [147]. This molecule transports peptide fragments to the surface of the APC, formed after an antigen is processed into fragments inside the cell [148]. The HLA region is the most gene dense (252 within 7.6 megabases) and polymorphic section of the genome [147].

There are two major HLA types, Class I and Class II [147]. Class I includes the *HLA-A*, *-B* and *-C* genes, and present antigens to CD8⁺ T- cells and Natural Killer (NK) cells. Class II includes the *HLA-DR*, *-DQ* and *-DP* genes, and present antigens to CD4⁺ T-cells. Hence dysfunction in a particular gene may imply causal effects in a specific T-cell type. HLA genes have a specific nomenclature [149][150], described in further detail for the UK Biobank in **Section 3.2**.

This region of the genome has been heavily implicated in autoimmune disease incidence [147]. Associations have been detected between HLA loci and multiple sclerosis (see **Section 1.5.9**), rheumatoid arthritis, type I diabetes, and other diseases. The precise mechanism through which disruption of normal HLA function leads to autoimmune disease is hypothesized, but not yet fully understood [147]. An example of a suggested mechanism includes HLA variants inducing alterations in binding grooves. This could affect which peptides are able to bind, leading to incomplete presentation of self-molecules, and hence some autoreactive T-cells incorrectly passing thymic selection. Another hypothesis is that HLA variants could lead to “cross-presentation”, where endogenous antigens are presented by HLA class II molecules, the opposite to the convention where they are presented only by Class I MHC molecules. This may lead to recognition of a self-antigen by

CD4⁺ T-cells, which could drive an immune response against self-tissues [147].

Lande *et al.* found overrepresentation of the alleles *HLA-C*07:04* and *HLA-DQB1*03:03* among people with ME/CFS [146]. The study included people diagnosed in Norway with CCC (422 people) or ICC (four people) ME/CFS, and 4,511 ancestry-matched healthy controls. *HLA-B*08:01* was also found to be less frequent than in the control group. The *HLA-C* locus association had not previously been reported in the literature, and the *HLA-B* and *-D* results were consistent with previous results. However, Lande *et al.* used locus-wise multiple testing correction due to the high linkage disequilibrium between alleles of the same HLA locus [146]. Locus-wise correction fails to account for the rest of the genome however, and so is equivalent to the “candidate gene” approach discussed in **Section 2.1.6**, where a prior association is assumed between the HLA loci and ME/CFS. This is not an unreasonable assumption for an autoimmune disease, which tend to be strongly associated with at least one HLA locus [147]. However, it is not yet known whether ME/CFS is an autoimmune disease, and so this threshold is likely to be unduly permissive to false positives. The standard GWAS threshold of $p < 5 \times 10^{-8}$ would be more appropriate, given the unknown aetiology of ME/CFS.

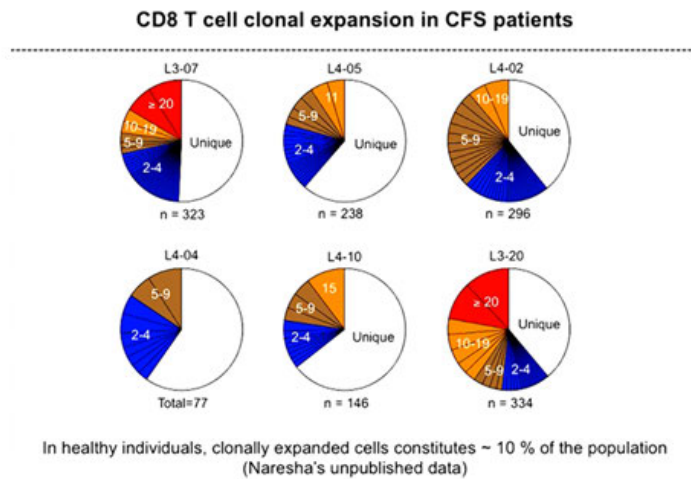
The overall picture is one in which there is a potential association between ME/CFS and immune dysfunction implied by the genetics, although this requires replication, which shall be addressed in **Section 3.2**.

2.2 Immunology (Lymphocytes)

As noted in **Section 2.1.7**, associations with the HLA loci might imply an expectation of immune system alterations in people with ME/CFS, as with autoimmune diseases. There is evidence to suggest that ME/CFS may be an autoimmune disease [49], and immune system alteration would be consistent with many studies in the ME/CFS field, albeit ones that are not necessarily in agreement with each other.

All types of lymphocytes have been implicated in ME/CFS. Studies have found altered levels of NK cells (increased [151] and decreased [152]), as well as decreased levels of T regulatory cells [151]. Sato *et al.* found evidence of skewed B-cell receptor (BCR) gene usage in ME/CFS [153], and Milivojevic *et al.* found evidence of altered use of the *IGHV3-23/30* gene segment (a component of BCRs) [154]. Differences in proportions of T-cell subtypes have also been reported [155].

A



B

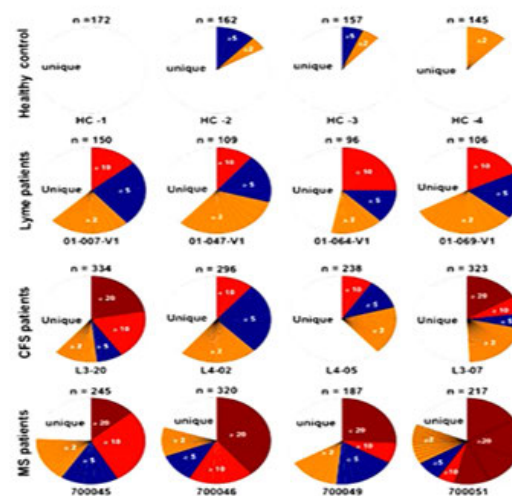


Figure 2.3: Slides from Dr Mark Davis' presentation, OMF Symposium 2017 [156].

In 2017, Dr Mark Davis of Stanford University presented results to the Open Medicine Foundation's (OMF) Community Symposium that suggested ME/CFS may also be an autoimmune disease. His presentation [156] cited evidence of inflammation in ME/CFS [157] which showed a linear association between cytokine activity and disease severity. Initial work by Jacob Glanville and Jose Montoya in bulk sequencing $TCR\beta$ chains had also shown more similarities between receptors in people with ME/CFS than controls.

From these, Dr Davis' lab investigated whether T- or B-cells might be involved in this systemic inflammation [156]. His previous work had showed evidence of T-cell clonal expansion in colon cancer ($CD4^+$ cells), multiple sclerosis ($CD4^+$ and $CD8^+$ in mouse models, $CD8^+$ in humans), Lyme disease ($CD4^+$ and $CD8^+$), and he extended this to ME/CFS ($CD8^+$, see **Figure 2.3**).

The combination of the previous studies [157], alongside these results, led Dr Davis to suggest that ME/CFS may be an autoimmune disease. This could make it tractable to methods frequently used when treating similar diseases, including immune suppressant therapies such as monoclonal antibodies which bind immune cells [158]. His work to find the antigen triggering clonal expansion is ongoing, however this evidence supports the implications of the genetics study discussed in **Section 2.1.7**, and a confirmed association with an HLA locus would provide strong evidence that ME/CFS is an autoimmune disease.

Dr Davis' work also indicated that there is a potential biomarker from the immune system by considering the diversity of the T-cell repertoire. Evidence from the literature supports this conclusion for a variety of diseases (see **Section 4.1.3**). **Chapters 4** and **5** are hence focused on the viability of TCR expansion as a biomarker in the case of ME/CFS.

2.3 Summary, and the Basis for Further Research

This chapter has provided the background of current ME/CFS research, and why further work is so important. Very little is understood about the aetiology or mechanisms of this debilitating disease. This thesis will focus on furthering the most promising evidence from the fields of genetics and immunology, and the links between these.

As discussed in **Section 2.1**, some of the most promising research on ME/CFS has been from population genetics studies, and given the typically causal nature of genetics, there is strong justification for starting research efforts in this field. I focus on replication of the results from GWAS, particularly the most plausible association found by the Neale group UK Biobank study (**Section 2.1.4**). Although this is not formally an independent replication, as it uses the same population, it may prove useful to compare the outputs from my analysis with those a previous study. Further, I investigate a new phenotype released by the UK Biobank (ME/CFS, as opposed to CFS), and combine the two into a larger, better powered cohort suitable for new discoveries. I also follow the implications of the Neale group results to investigate whether women with CFS report dietary protein reduction to mitigate the effects of reduced *ORNT1* protein production (**Chapter 3**).

Due to their highly polymorphic nature, HLA alleles are not typically included in GWAS. However, confirmation of the associations found by Lande *et al.* would suggest immune system dysfunction in ME/CFS, and potentially imply an autoimmune basis for the disease (**Section 2.1.7**). Hence, these were investigated in a separate study, again using the UK Biobank cohort, and made more rigorous by implementing a more appropriate multiple testing correction.

There is sufficient evidence in the literature to suggest that ME/CFS causes, or is the result of, immune dysfunction. Hence, this thread was followed to the most promising work in the field of immunology; the T-cell repertoire profiling of Dr Davis and his research group (**Section 2.2**). This thesis (**Chapters 4 and 5**) will demonstrate a method of measuring a mathematical “distance” between disease cases and controls using a refined version of a method established in the TCR literature. This is accomplished by quantifying the levels of clonal activation seen in the immune system using mathematical metrics, and separating them into subgroups by applying a machine learning algorithm.

The goal was to investigate the aetiology and find a potential biomarker (or set of genetic risk factors) for this terrible disease.

Chapter 3

Genetic Aetiology of ME/CFS and Fibromyalgia

3.1 Genome-wide Association Studies

The field of genetics is well-placed to make discoveries into the aetiology of ME/CFS, due to the causal nature of genes and genetic variation [6]. As described in **Section 2.1**, of the many genetics studies on ME/CFS, the most promising for further research are UK Biobank genome-wide association studies (GWAS), and the HLA study of Lande *et al.* [146]. The most plausible GWAS result (genome-wide significant, for a common allele) was that found by Neale *et al.* [120]. The SNPs they found associated with CFS in females lie near the gene *SLC25A15*, which is mutated in the genetic illness Hyperornithinemia-hyperammonemia-homocitrullinuria (HHH) syndrome [159]. Since HHH symptoms include some overlap with ME/CFS (in particular lethargy and cognitive dysfunction [159]) this potentially lends further support to the Neale group finding.

SLC25A15 encodes mitochondrial ornithine transporter 1 (ORNT1), which plays an important role in the urea cycle by transporting ornithine, lysine and arginine across cell membranes [159]. Disrupting the production of ORNT1 could imply a build-up of toxic ammonia in the blood and cytoplasm of affected persons, as in the extreme case of HHH syndrome. People with HHH syndrome will often avoid sources of protein in their diet, as this tends to cause a relapse or worsening of symptoms. A strictly maintained and clinically prescribed low-protein diet is, in fact, one of the treatments for HHH syndrome [159]. It is sufficiently effective that there are studies involving adults with *SLC25A15* mutations who did not realise

they had HHH syndrome, due to lifelong vegetarianism [160]. This represents another commonality with ME/CFS – people predisposed to the condition are unaffected before an inciting incident.

There is evidence to suggest that rs7337312 is an eQTL for *SLC25A15*, including in adipose and arterial tissues, among others [161]. It is therefore plausible that those with the CFS risk variants might avoid protein sources in their diet, as a single mutation would potentially lead to reduced *SLC25A15* mRNA, and hence reduced ORNT1 protein production. This may then lead to an accumulation of ammonia in blood which may then present clinically in a similar manner to a milder form of HHH syndrome.

In parallel, the HLA locus associations found by Lande *et al.* support the systematic immune dysfunction reported in the literature (see **Section 2.1.7**). Replication of these associations in an independent and larger cohort would lend credibility to the hypothesis that ME/CFS is principally an immune system disease.

This chapter addresses both of these research topics. I first address the fatiguing complex disease phenotypes in the UK Biobank, and discuss the creation of more rigorously defined ME/CFS and fibromyalgia cohorts. These individuals were then used for case-control GWAS. The goal was to attempt to replicate these previous studies with the UK Biobank phenotypes (**Section 3.1.1**), and with the aim to discover new associations using the constructed phenotypes (**Section 3.1.2**). The association with *SLC25A15* was then investigated by fitting linear models to detect whether females with the CFS risk variants show a significant tendency to alter their dietary protein intake. Finally, associations to variants in the HLA region were investigated, with the aim to further strengthen the evidence base for the immunological study of **Chapters 4 and 5**.

The hypotheses to be investigated are, in order:

1. There are significant associations between complex fatiguing disease phenotypes in the UK Biobank (ME/CFS, CFS or fibromyalgia) and common single nucleotide polymorphisms in the genome.
2. There is a significant alteration of dietary protein intake in those with the *SLC25A15* risk alleles, and those without (conditional on the first hypothesis being true, and the association being consistent with the findings of Neale *et al.*).
3. There is a significant association between complex fatiguing disease pheno-

types in the UK Biobank (ME/CFS, CFS or fibromyalgia), and at least one HLA allele.

3.1.1 Phenotypes in the UK Biobank

There are two ME/CFS phenotypes reported in the UK Biobank. The first, “Chronic fatigue syndrome”, was released among the initially acquired phenotypes (“Non-cancer illness code, self-reported”, classified under “systemic/immunological disorders”) and is regularly updated with in-person assessments. As of the February 2022 data release, there were 2,281 people with this CFS phenotype.

Those with this phenotype have self-reported that they have been told by a doctor that they have chronic fatigue syndrome [162]. If the participant was unsure of their illness, they described it to their interviewer (a nurse) who attempted to classify it within the available categories. If the interviewer was unable to classify the described illness, they entered a description of it for review by a doctor [163]. These free-text descriptions were then either classified within the available phenotypes, or were marked as unclassifiable (if a category could not be identified with high probability) [163]. This multi-step procedure may hence have introduced some false diagnoses into the CFS group, or misclassified them under an incorrect code.

The “ME/CFS” phenotype derives from a follow up “Experience of Pain” questionnaire, released in 2019. This questionnaire was sent to all UK Biobank participants who provided a contact e-mail address (335,587 people [164]), providing 167,200 responses in total (including those who accessed the questionnaire through other methods) [165]. Participants were asked “Have you ever been told by a doctor that you have had Chronic Fatigue Syndrome or Myalgic Encephalomyelitis (M.E.)?”, with 2,721 responding “Yes” [166]. Participants returned responses over the course of January 2019 – April 2020. This phenotype should capture both current and recovered ME/CFS and CFS cases. This is particularly useful as diagnoses of ME/CFS, CFS and ME are often given interchangeably.

There are also equivalents for the disease fibromyalgia, with the initial phenotype “Fibromyalgia” classified under “Non-cancer illness” and “musculoskeletal/-trauma” as a “muscle/soft tissue problem”, including 1,032 individuals as of their most recent site visit. The Pain Questionnaire phenotype (“Fibromyalgia (PQ)”) includes 3,013 cases. As with ME/CFS and CFS, the Pain Questionnaire version should capture both current and recovered cases.

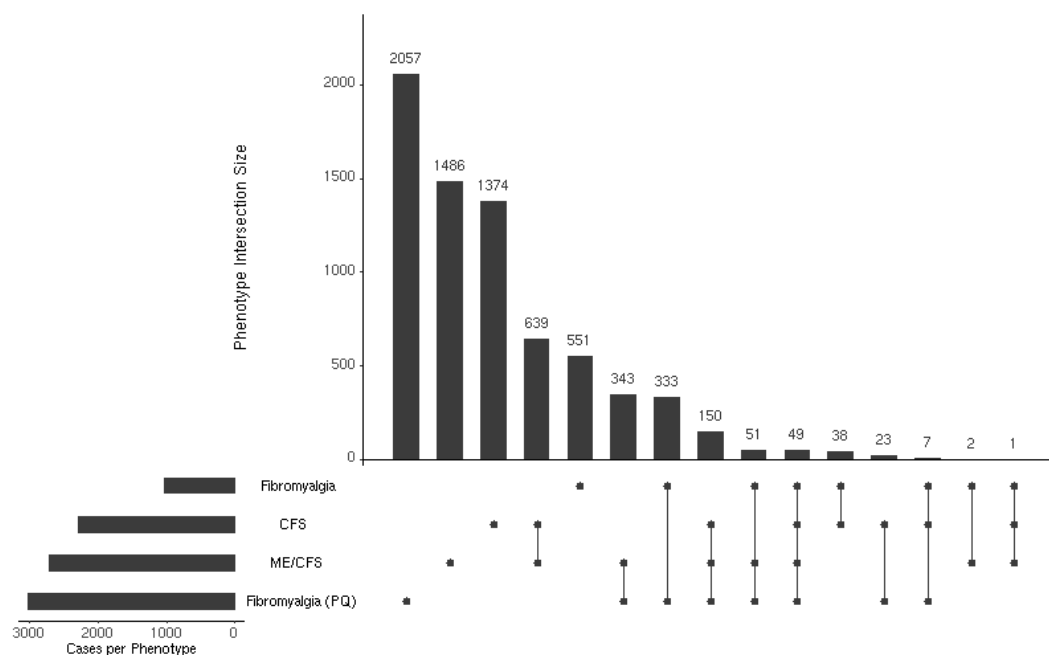


Figure 3.1: Intersections between the CFS and Fibromyalgia phenotypes from the UK Biobank (data from February 2022), and the ME/CFS and Fibromyalgia phenotypes from the “Experience of Pain” questionnaire (issued in 2019).

The overlaps between these four phenotypes are shown in **Figure 3.1** using an UpSet plot [167].

Intersections between these four phenotypes, and hence the two diseases, are small. There are 7,104 people who reported at least one of these four phenotypes, of the 502,413 in this UK Biobank data release. Of these, 4,163 reported either ME/CFS or CFS (or a previous diagnosis of ME/CFS or CFS), and 3,605 reported fibromyalgia (or a previous diagnosis), giving prevalences of $\sim 0.8\%$ for ME/CFS and $\sim 0.7\%$ for fibromyalgia. However, those missing a pain questionnaire response (*NA*), will only be included for their non-cancer phenotypes, so these intersections may be undercounting the true values.

Multiple quality control (QC) criteria were applied. Samples assessed to have low-quality genotyping (marked as heterozygosity or missingness rate outliers) by the Biobank were removed, as were those with sex chromosome aneuploidy (fewer, or additional, sex chromosomes [168]). Heterozygosity and missingness were calculated by Bycroft *et al.* [169] using the PLINK [170][171][172] `--het` and `--miss` flags. Their missingness threshold flagged any samples with a rate > 0.05 ($> 5\%$ markers missing), which I removed from my analyses. The samples to be considered heterozygosity outliers required additional calculations. Bycroft *et al.* defined the raw heterozygosity (the proportion of non-missing heterozygous genotypes) as:

$$h = \frac{N_{nm} - N_{hom}}{N_{nm}} \quad (3.1)$$

where N_{nm} is the number of non-missing genotypes, and N_{hom} is the number of homozygous genotypes. Heterozygosity rate depends upon population structure, as allele frequencies will vary across subpopulations. They controlled for this using a linear model:

$$h(x) = h_0 + \beta(x) \quad (3.2)$$

where x is a set of features related to ancestry, and β is the bias due to population stratification. The authors chose to use the first six principal components of the dataset, so $x = (x_1, x_2, x_3, x_4, x_5, x_6)$ here. They selected a quadratic for β , giving it the form:

$$\beta(x) = \sum_{i=1}^6 \beta_i x_i + \sum_{i=1}^6 \beta_{ii} x_i^2 + \sum_{i=1}^5 \sum_{j>i}^6 \beta_{ij} x_i x_j \quad (3.3)$$

where the β are scalars. This includes all x_i and x_i^2 terms, plus their products $x_i x_j$. h_0 was then the corrected linear model for heterozygosity. Those in the first four largest genetic ancestry groups with significantly greater heterozygosity (using this model) than the mean (0.1903) were flagged, and other ancestry groups were checked manually with plots of the missing rate and h_0 . This procedure, along with the missing rate filter, identified 968 individuals as outliers.

The UK Biobank also provided inferred kinship for close relations, and so any participant with at least one first, second, or third degree relative in the dataset was excluded from my analyses (in line with the Neale group, who also filtered based on relatedness [173]). Those who had not had their kinship investigated (with an *NA* entry) were also removed. Finally, the “White British” subset was used, where UK Biobank checks self-reported ethnicity against the genotyping results to create a large subpopulation (84% of the UK Biobank) with common genetic ancestry [169]. Although there are ethical implications to using ancestrally homogeneous cohorts [174], it is important that false associations are not found due to overrepresentation of certain alleles among particular populations.

Biobank participants in the groups excluded from further analyses are shown in **Figure 3.2**. After this filtering, 276,307 participants remained. This cohort shall

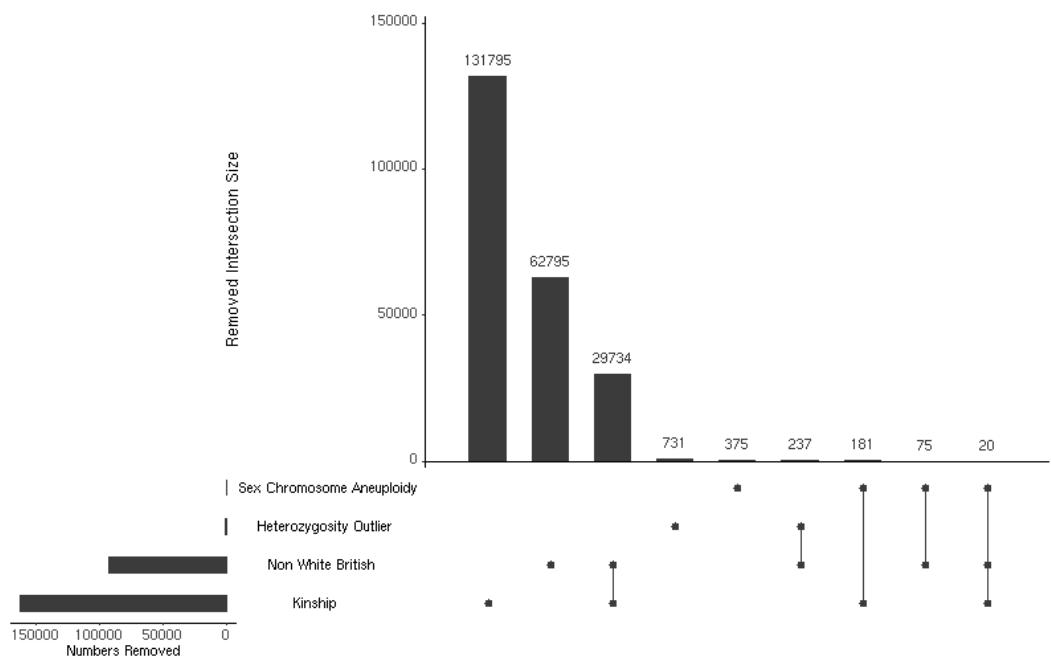


Figure 3.2: Set intersections for those excluded from analyses. The majority (132,607) were removed due to relatives in the cohort (kinship). Restricting to the White British subset results in the loss of some power ($\sim 15\%$ of the cohort), but reduces the likelihood of confounding due to genetic ancestry. Other genetic ancestries can be considered separately, or accounted for using a genetic relationship matrix approach (see **Section 2.1.1**).

be referred to as the “filtered cohort”, and was used for all subsequent main analyses.

Updated case numbers for each phenotype after filtering are shown in **Figure 3.3**. There were 1,358 people with CFS, 1,561 with ME/CFS, 554 with Fibromyalgia, and 1,697 with Fibromyalgia identified by the pain questionnaire. This represents a dramatic reduction, but is in line with the reduced size of the filtered cohort, and is a result of the QC measures. Again, those who did not respond definitively (“Yes” or “No”) to the Pain Questionnaire will only be categorised based on their non-cancer phenotype status. Hence, for example, the intersection of people with both CFS and ME/CFS may be greater than shown here, as someone with CFS and “Do not know” or “Prefer not to answer” for ME/CFS will only be counted under the CFS intersections even if they would have answered “Yes” to that phenotype had they completed the Pain Questionnaire.

A summary of the numbers of people with each phenotype in the cohort is shown in **Table 3.1**, separated by sex. The female to male ratios were noticeably increased when compared with the group of distinct controls (the group which had none of the above phenotypes). The relative risk (RR) increase of ~ 2 for females is consistent with the well documented sex imbalance noted in the literature for both

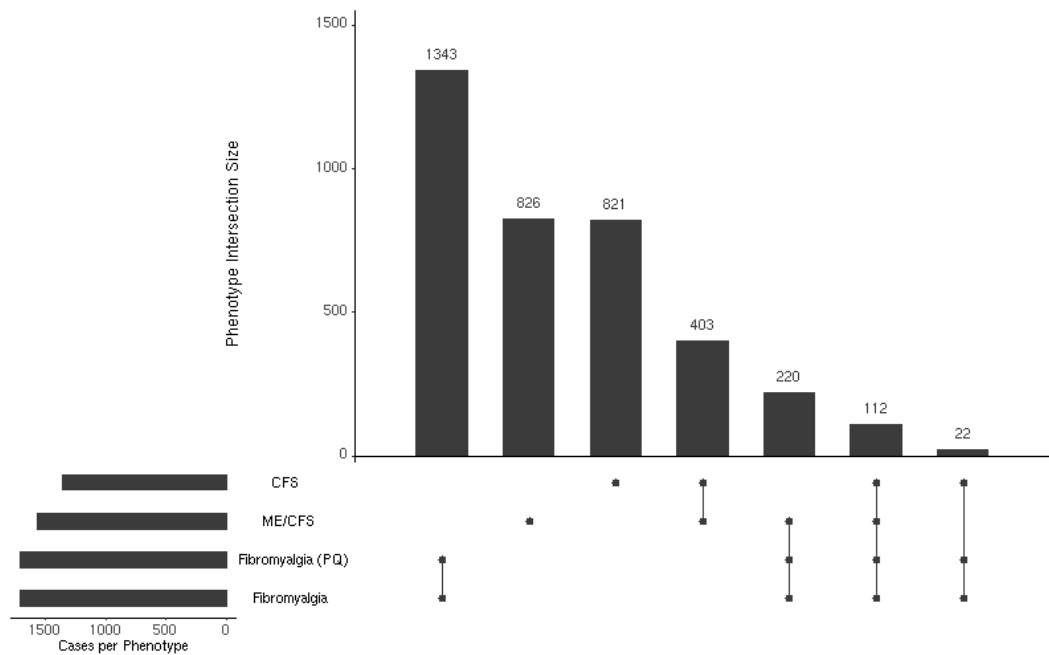


Figure 3.3: *Intersections between the four phenotypes after filtering.*

	ME/CFS	Fibromyalgia (PQ)	CFS	Fibromyalgia	Control
Female	1,118	1,406	990	448	144,169
Male	443	291	368	66	128,107
Ratio	2.52	4.83	2.69	6.79	1.26

Table 3.1: *Summary of the numbers of people in the filtered cohort with each phenotype, separated by sex. The female to male ratio is markedly increased compared with the controls (none of the four phenotypes).*

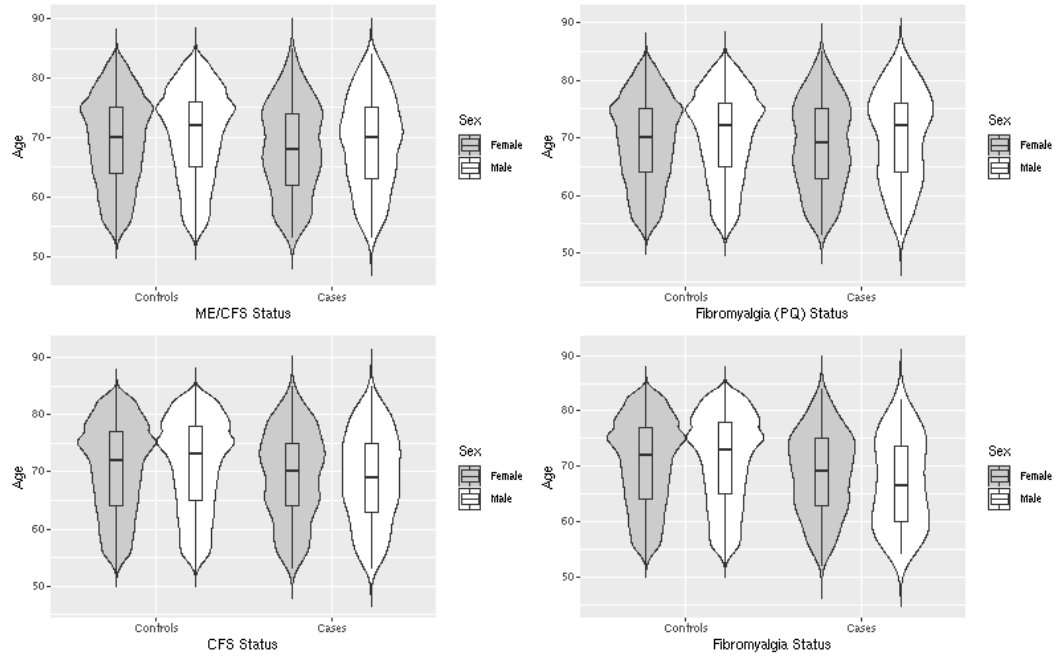


Figure 3.4: Age distributions (in years) for each of the four phenotypes, by case status and sex [175]. ME/CFS, Fibromyalgia (Pain Questionnaire), Fibromyalgia, and CFS are shown clockwise from the top left.

ME/CFS and fibromyalgia (approximately a 3 : 1 female to male ratio [17]), and is therefore to be expected for these Biobank phenotypes.

The age distribution for each phenotype (as of the February 2022 data release), separated by sex, is shown in **Figure 3.4**.

Age distributions do not differ visibly by sex, but the age peak at around 75 years in the controls is less prominent in the cases. This could indicate an effect of ME/CFS on lifespan, or quality of life in advanced age (and hence ability to participate in research). However this could also potentially reflect a historical lack of diagnosis of the disease, as although reports of a similar disease date to 1934, the first use of “Benign Myalgic Encephalomyelitis” was after an outbreak at the Royal Free Hospital in London in 1955 [176]. This is only ~ 7 years after the birth year of the age peak in the UK Biobank data, and hence the diagnosis was unlikely to be widespread for a considerable period of this cohorts’ lives.

However, the numbers for each phenotype are not fixed. Participants are permitted to withdraw at any time, and new cases were added to the CFS and fibromyalgia phenotypes over time at follow-up appointments, after the initial assessment. Note that this implies the prevalences reported in the literature, and here, for UK Biobank phenotypes may not remain constant as a result. Some of the intersections in **Figure 3.3** lead to questions of less reliable participants, such as

Reference	CFS (pre-PQ)	ME/CFS	CFS before PQ	Result	Number
1	No	No	N/A	Control	273903
2	No	Yes	N/A	ME/CFS Case (current ME, or a history of ME/CFS, pre-Biobank)	1046
3	Yes	No	No	CFS Case (incidence after PQ)	2
4	Yes	No	Yes	Potential anomaly	841
5	Yes	Yes	No	ME/CFS Case (incidence after final visit before PQ)	0
6	Yes	Yes	Yes	ME/CFS Case	515

Table 3.2: *The logical rationale for considering the exclusion of some participants from analyses. “PQ” indicates the “Experience of Pain” questionnaire. The numbers of participants from the filtered cohort in each group are included. Those without a definitive response to the pain questionnaire were counted as “No” for ME/CFS. This table was used for the construction of the “Combined” phenotypes of Section 3.1.2. The same section also defines the criteria for pre- and post-PQ inclusion.*

those reporting CFS before the Pain Questionnaire, then answering “No” to “Have you ever been told by a doctor that you have had Chronic Fatigue Syndrome or Myalgic Encephalomyelitis (M.E.)?”. The CFS and fibromyalgia phenotypes were introduced before the Pain Questionnaire equivalents were defined, but have been continuously updated. Hence, there are several logical outcomes, as shown in **Table 3.2**.

The UK Biobank also records the year of onset for each disease for the non-cancer illness phenotypes, and the time of completion of the Pain Questionnaire. By systematically testing whether they report their diagnosis of CFS before or after their completion of the Pain Questionnaire, it is in theory possible to determine to which of the above categories each participant belongs (see **Section 3.1.2**). The numbers of each are recorded in **Table 3.2**. This will be expanded to define new ME/CFS and fibromyalgia phenotypes in **Section 3.1.2**, which are used in subsequent analyses, in addition to the initial four phenotypes shown in **Figure 3.3**.

Potential anomalies will be discussed further in **Section 3.1.2**. There is an equivalent table formed by replacing CFS with Fibromyalgia, and ME/CFS with Fibromyalgia (PQ).

3.1.2 Combined Phenotypes

As mentioned, there are questions over the reliability of the non-cancer illness codes, and responses to the Pain Questionnaire, as some combinations of answers imply anomalies (**Table 3.2**). This section will investigate the possibility of defining new ME/CFS and fibromyalgia cohorts based upon additional data supplied by the UK Biobank.

The non-cancer illness codes are stored in a four-section matrix, numbered 0 – 3

for each assessment visit, with 34 (numbered 0 – 33) entries for each visit. These are stored in the format “visit.entry”, and each entry can hold a single illness code. For example, the code 1482 in column 1.0 indicates that an individual reported CFS at their second assessment (1 is the second assessment, 0 is the first illness code for that assessment). If they also have 1542 in column 1.1, this indicates that they reported fibromyalgia at the same assessment.

In addition, the UK Biobank stores the date of illness onset (encoded by “year.decimal fraction”) in a separate matrix constructed in the same manner. Therefore, an entry of 2012.5 in column 1.0 indicates that the individual above reported their CFS onset to be halfway through 2012. UK Biobank interpolates to the mid-point of the year if only a year or age was provided.

The UK Biobank also records the time of completion of the Pain Questionnaire. 90% of participants had completed the questionnaire by June 2019, but it is in theory possible to compare the date of completion with their reported onset of non-cancer illness (where both are available). This is particularly useful to exclude participants who report the onset of CFS to have been before their completion of the Pain Questionnaire, and yet who also reported no history of ME/CFS or CFS, as these are likely anomalies. The logic shown in **Table 3.2** was used to exclude them if their answers indicated an anomaly (row 4).

I wrote *R* code to find the date of onset for all participants in the filtered cohort who reported CFS or fibromyalgia, and planned to compare this with their date of completion of the Pain Questionnaire. The date that each individual participant completed their Pain Questionnaire was not available, however meta-data showed that all participants had completed the questionnaire by June 2020 [165]. Hence, any participant who reported their onset of CFS to be before 2020.5, without also being in the ME/CFS group, was excluded from the combined cohort. The same was separately done for the Fibromyalgia (PQ) and Fibromyalgia phenotypes.

There were 54 participants considered anomalies for ME/CFS, with almost all reporting CFS onset before 2019. Only one case was ambiguous, with their CFS onset in 2019, but without the data to confirm the date of their PQ completion, this case was also excluded. Note that those without a response to the Pain Questionnaire (a “Do not know” or “Prefer not to answer”) were retained, as there was no reason to consider their disease status anomalous, and so the actual number excluded does not match the number in **Table 3.2**. Hence there were $841 - 54 = 787$ people considered potential anomalies for ME/CFS who were eventually retained. For fibromyalgia, there were nine anomalies, with one ambiguous (2019

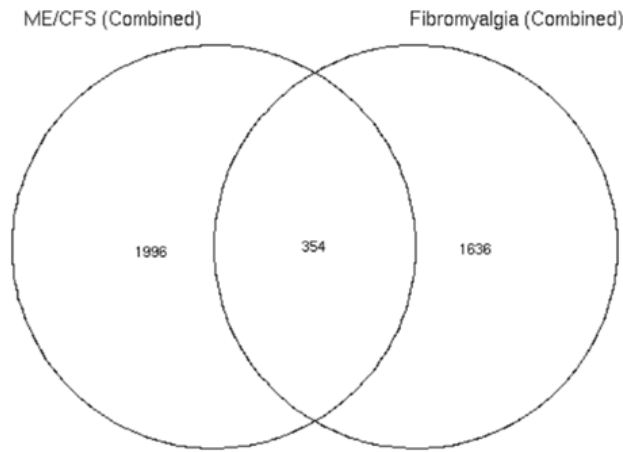


Figure 3.5: Intersection for the constructed ME/CFS and Fibromyalgia phenotypes. These shall be used for analyses in subsequent sections. Venn diagram drawn with the “venn” R package [177].

onset). Again, all of these were excluded.

By excluding these individuals and taking the union of the ME/CFS and CFS phenotypes, and the Fibromyalgia (PQ) and Fibromyalgia phenotypes, this yielded two new phenotypes. These will henceforth be referred to as “ME/CFS (Combined)” and “Fibromyalgia (Combined)”, and their intersection is shown in **Figure 3.5**. Combining these four in some manner to form a “complex disease” phenotype was considered, but ultimately decided against, as ME/CFS and fibromyalgia are typically considered distinct diseases (they have different ICD-11 codes, for example).

There were 2,350 people with the ME/CFS (Combined) phenotype, and 1,990 with the Fibromyalgia (Combined) phenotype. The Jaccard index for the overlap between sets A and B [178] is:

$$J = \frac{A \cap B}{A \cup B} \quad (3.4)$$

which is small for these two phenotypes ($J \approx 0.09$). This implies perhaps that clinicians will tend to diagnose only one of the two diseases, despite the frequent co-morbidity reported in the literature (see **Section 1.5.1**), and poorly defined case definitions for both diseases. The prevalence of ME/CFS in the UK Biobank filtered cohort can now be estimated as $\sim 0.9\%$, consistent with both the pre-filtering estimates and those in the literature (see **Section 1.2**). The prevalence was $\sim 0.7\%$ for fibromyalgia.

RAF	RR	Power	Cases for 80% Power
0.05	1.05	< 1%	165975
0.05	1.1	< 1%	41700
0.05	1.2	< 1%	10528
0.05	1.3	15.5%	4725
0.05	1.4	66.9%	2684
0.05	1.5	97.0%	1734
0.3	1.05	< 1%	38484
0.3	1.1	< 1%	9907
0.3	1.2	69.3%	2623
0.3	1.3	99.9%	1232
0.3	1.4	100%	732
0.3	1.5	100%	493

Table 3.3: Power to detect associations, shown for various cases of risk allele frequency and relative risk, using an additive model. Estimates are for the ME/CFS (Combined) phenotype, and an assumption was made that the phenotyping process ensured none of the controls had the disease.

3.1.3 Power

The power to detect associations for these cohorts should be calculated, to determine whether there is a possibility of missing genuine associations (false negatives, or Type II errors [179]). Spencer *et al.* report that common alleles affecting common diseases will typically have effect sizes (relative risk per allele) in the range of 1.1 – 1.2 [180].

Using the Genetic Power Calculator of Purcell *et al.* [181] for the ME/CFS (Combined) cohort for a range of inputs suggested more disease cases would be needed to detect a genome-wide significant result with 80% power (20% probability of a Type II error). The extreme cases of risk allele frequency (RAF), alongside a slightly broader range of RR than given by Spencer *et al.*, are shown in **Table 3.3**. Calculations were for a 1 degree-of-freedom allele test, with the Type I (false positive [179]) error rate set at the genome-wide significance level of 5×10^{-8} , and a control:case ratio of 117 : 1. The risks here are shown for a single allele, additive model [182] (where the relative risks for allele *A* of genotypes *Aa* and *AA* are not equivalent), which was chosen to match the GWAS analysis (see **Section 3.1.4**).

Case numbers in the UK Biobank are sufficient only to detect associations with SNPs that incur a relatively larger risk of ME/CFS (greater than ~ 1.2 for $RAF = 0.30$). However, for greater levels of risk (~ 1.5) associations should be found even for the rarer variants (closer to $RAF = 0.05$). **Table 3.3** is consistent with the findings of Spencer *et al.*, who report that 2,000 – 3,000 cases and controls are

sufficient for detecting associations in the 1.3 – 1.5 RR range, but studies with even 5,000 cases and controls are insufficient to detect associations for lower RR alleles [180]. Since ME/CFS is a complex trait, which likely involves many genetic associations, this is a clear limitation of investigating this disease using the UK Biobank cohort.

All other phenotypes should give similar results, as the case numbers are of approximately the same order (10^3).

3.1.4 GWAS Methods

A GWAS for each of the four Biobank phenotypes was completed using PLINK 2.00 [170][171][172][183][184]. The cohort was filtered as previously described (**Section 3.1.1**), and a logistic model was fit to each genotyped SNP (784,256 across the 22 autosomes) for the four phenotypes. The model takes the form:

$$\mathbf{y} = \mathbf{G}\beta_G + \mathbf{X}\beta_X + \mathbf{e} \quad (3.5)$$

where $\mathbf{y}$ is the phenotype vector, $\mathbf{G}$ is the genotype matrix, $\mathbf{X}$ is the covariate matrix and $\mathbf{e}$ is the error term (from least-squares minimization) [185]. The phenotype vector is logit transformed:

$$\text{logit}(p) = \log \frac{p}{1-p} \quad (3.6)$$

where p is the proportion of individuals with a phenotype [186]. Logistic models are more appropriate for disease phenotypes, as their typically binary nature [186] may require the use of an underlying continuous risk variable to fit a linear model. The model used was additive, where each additional copy of the risk variant was assumed to linearly increase the risk on a log scale. Therefore, for a reference allele a and risk allele A , the genotype Aa was assumed to confer a risk (odds ratio) of X , and AA was assumed to confer a risk of X^2 [182].

Age, age^2 , assessment centre, batch (and hence array type), genotyping plate, genotyping well, and the first 40 principal components (calculated by the UK Biobank) were investigated as covariates. Batch proved to have too many levels (categories), and age^2 was too highly correlated with age, for PLINK to include these as covariates. PLINK also failed to fit assessment centre for all phenotypes,

so this was removed as well. These failures to fit were seen as missing p -values in the output file when that covariate was included in the model. As three analyses were completed for these four phenotypes (for the entire cohort, and for males and females separately), genetic sex was also included as a covariate for the full cohort, in the manner of Neale *et al.* 40 principal components (PCs) would likely lead to issues of overfitting, so only $PC1 - 20$ were included in the final model (again, as for the Neale *et al.* study). Likewise, genotyping plate and well covariates were removed, due to similar overfitting concerns. Overfitting would lead to issues of generalisation, where the model had fit to the noise in the dataset, and hence the result would not be reproducible in another cohort [187]. A general rule is that it is ideal to have at least 10 – 15 cases per covariate in the model [188], and hence including all 40 PCs may have led to bias.

Testing of the remaining covariates was completed using a (generalised) linear model ($glm()$ in *R* [189]):

$$\mathbf{y} \sim Sex + Age + \sum_{X=1}^{20} PCX \quad (3.7)$$

where $\mathbf{y}$ is the phenotype vector. The results for the ME/CFS (Combined) phenotype are shown in **Table 3.4**. Sex and age were the only covariates to remain significant for all four phenotypes, but the inclusion of some principal components is considered standard in GWAS [93].

The variance inflation factor (VIF) was also investigated, which is measure of multicollinearity [190]. For a linear model:

$$y = \beta_0 + \sum_{j=1}^k \beta_j x_j + \epsilon \quad (3.8)$$

(where the x_j are the covariates), VIF is defined as [190]:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (3.9)$$

where:

$$R_j^2 = 1 - \frac{\sum (x_j - x_{jc})^2}{\sum (x_j - \bar{x}_j)^2} \quad (3.10)$$

Covariate	Estimate	Std. Error	t-value	$Pr(> t)$	Sig.	VIF
(Intercept)	2.53E-02	2.10E-03	12.0	$< 2e - 16$	***	-
Sex	-6.15E-03	3.50E-04	-17.6	$< 2e - 16$	***	1.00
Age	-2.37E-04	2.19E-05	-10.8	$< 2e - 16$	***	1.00
PC1	-2.39E-04	1.14E-04	-2.09	0.0364	*	1.12
PC2	1.08E-04	1.19E-04	0.907	0.364		1.04
PC3	4.36E-05	1.15E-04	0.380	0.704		1.09
PC4	-1.51E-04	8.76E-05	-1.73	0.0841		2.19
PC5	8.95E-05	3.83E-05	2.34	0.0194	*	2.09
PC6	4.69E-05	1.10E-04	0.428	0.669		1.07
PC7	-7.76E-05	9.84E-05	-0.79	0.430		1.12
PC8	2.98E-04	9.86E-05	3.02	0.00250	**	1.26
PC9	3.62E-06	3.93E-05	0.092	0.927		1.14
PC10	4.34E-06	9.46E-05	0.046	0.963		1.54
PC11	-7.50E-06	7.04E-05	-0.107	0.915		2.12
PC12	5.00E-05	9.42E-05	0.530	0.596		1.63
PC13	-1.55E-04	1.08E-04	-1.44	0.150		1.03
PC14	-3.02E-05	5.67E-05	-0.532	0.595		1.15
PC15	1.26E-04	9.96E-05	1.26	0.206		1.22
PC16	4.09E-05	5.86E-05	0.697	0.486		1.17
PC17	-1.01E-04	8.69E-05	-1.16	0.247		1.00
PC18	-9.62E-05	5.89E-05	-1.63	0.102		1.01
PC19	1.83E-05	5.98E-05	0.306	0.760		1.00
PC20	-8.19E-05	6.02E-05	-1.36	0.174		1.00

Table 3.4: Investigation using `glm()` in R of the covariates to be used in the GWAS. The results shown are for ME/CFS (Combined), but were similar for the other phenotypes. Sex and age were the only covariates which significantly affected all six phenotypes. Significance levels are shown for $p < 0.001$ (***), $p < 0.01$ (**), and $p < 0.05$ (*).

with x_{jc} the regression line data, and $\bar{x}_j$ the mean value [191]. R_j^2 is the coefficient of determination of the model given by the j^{th} covariate written as a function of all the other covariates (**Equation 3.7**, with y removed, and the relevant x_j moved to the left-hand side [192]). This should be calculated for all covariates $j \in [1, k]$ to give a set of VIF scores. The literature gives varying numbers for an acceptable VIF, however 5 [193] or 10 [193][194] are considered typical upper bounds. For these data none of the covariates used in **Equation 3.7** had $VIF > 5$, as shown for ME/CFS (Combined) in **Table 3.4**, and so these were all retained for the PLINK and R analyses.

There are potential issues of reproducibility introduced by using current age (where $Age = Current\ Year - Year\ of\ Birth$) as a covariate in this manner, rather than “Age at Diagnosis”. However, a covariate such as this would leave open the question of how to address controls, who lack this data entry. “Age at Initial Site Visit”, may have proved an acceptable alternative, however since current age proved highly significant here, and the inclusion of a measure of age is standard practice in GWAS, I retained it as a covariate, while being aware of its limitations.

A minor allele frequency filter was included (alleles with $MAF < 0.05$ were removed), as this is a typical threshold for excluding rare variants in GWAS [106]. As previously discussed in **Section 2.1**, it is unlikely for a relatively common disease such as ME/CFS or fibromyalgia to be the result of a single rare variant, and rarer alleles are more likely to lead to false associations (especially with so few cases). Therefore, filtering these alleles from the results is appropriate. This will also account for the extremely unbalanced control:case ratios of $\sim 117 : 1$. For this level of imbalance, statistical correction or subsampling would usually be required [121], and the latter may reduce the power (**Section 3.1.3**). However, unbalanced control:case ratios only have a notable effect on rarer variants with a $MAF < 0.05$ [121], and so this can be avoided here.

Additionally, a Hardy-Weinberg Equilibrium (HWE) test was applied, removing variants with $p < 1 \times 10^{-9}$ in the manner of Neale *et al.* [120][173]. The Hardy-Weinberg Theorem states that the genotype frequencies between heterozygous and homozygous individuals will not change in a population in HWE [195]. SNPs associated with a trait might deviate from HWE among the cases, but extreme departures (typically on the order of $p \sim 10^{-50}$) are more likely due to genotyping errors, and these variants should be removed from analyses [185].

Results will be displayed using Manhattan and Quantile-Quantile (QQ) plots in **Sections 3.1.5** and **3.1.6**. The Manhattan plots show the $-\log_{10}(p)$ of each

SNP against its position in GRCh37/hg19 coordinates, and the QQ plots show the actual $-\log_{10}(p)$ against the expected values (where points above the red line are those with smaller p -values than expected, and below those with greater p -values than expected). Genomic inflation factors (λ) will be reported for each cohort [196][197]. P -values will be assessed using two significance levels. The first are those considered genome-wide significant at $p < 5 \times 10^{-8}$, which is usually considered the standard in GWAS for common variants [106]. This significance threshold will be shown on the Manhattan plots as a horizontal red line. The second is sometimes referred to as the “suggestive” threshold of $p < 1 \times 10^{-5}$ [115], and shall be indicated on the Manhattan plots as a horizontal blue line. This significance threshold is the level reported by Open Targets Genetics [161][198], a portal aggregating the results of GWAS from many sources. Their disease associated loci data are obtained from the Neale group [120], GWAS catalog [137], and FinnGen [199], all of which are GWAS projects that have made their data available for research. Results from the paper describing the SAIGE method [121] shall also be used. SNPs identified by the analyses presented here will be compared with the Open Targets resource to determine whether they have been associated with any other traits. This may provide some indication of any potentially causal biological mechanisms relevant to the disease phenotypes.

Associations will also be compared with those of the GeneATLAS project [100], which conducted GWAS and PheWAS on the UK Biobank cohort before the release of the Pain Questionnaire. This will enable validation of results, where similar significance levels should be seen for SNPs associated with CFS and fibromyalgia. It will also allow investigation of any other phenotypes associated with the SNPs identified, as with Open Targets.

3.1.5 Original Phenotypes Results

GWAS were performed for each of the four phenotypes provided by the UK Biobank. The results are shown below for CFS, ME/CFS, Fibromyalgia and Fibromyalgia (PQ). **Figure 3.6** shows Manhattan and QQ plots for CFS [200]. No SNP reached genome-wide significance. There was a small suggestive signal for CFS on chromosome 1 in the full cohort, where the SNP with the smallest p -value was rs12145346, with $p = 9.49 \times 10^{-8}$. For males, there was a signal on chromosome 11 for the CFS phenotype, with rs10896588, rs35008227, and rs12225155 all reaching the suggestive threshold. Of these, rs10896588 was the closest to genome-wide significance ($p = 6.50 \times 10^{-7}$). The QQ plots for both of these cohorts indicate an

increased number of p -values near the suggestive threshold than would be expected by chance.

There were also three associations for females with the CFS phenotype on chromosome 13, at the suggestive $p < 1 \times 10^{-5}$ level, with a corresponding deviation in the QQ plot. Two of these are in the same location (near *SLC25A15*) as detected by Neale *et al.*, and so the hypothesis of an association between reduced dietary protein intake and females with CFS remains plausible. This remains the most likely hypothesis, due to the known function of the lead SNP as an eQTL for *SLC25A15*. The implications of this are investigated further in **Section 3.1.7**. Once again, the QQ plot indicates more than expected associations close to the suggestive threshold.

Figure 3.7 shows the plots for the ME/CFS phenotype. No genome-wide significant associations were detected, which is supported by the minimal deviation observed in the QQ plots for all three cohorts.

Associations were investigated for the two Fibromyalgia phenotypes (**Figures 3.8 and 3.9**). Again, no genome-wide significant associations were found, with very few associations found meeting the suggestive threshold. In addition, the QQ plots for both only showed deviations below the line, indicating fewer than expected p -values around the suggestive threshold.

The suggestive associations found for these four phenotypes are summarised in **Section 3.1.7**.

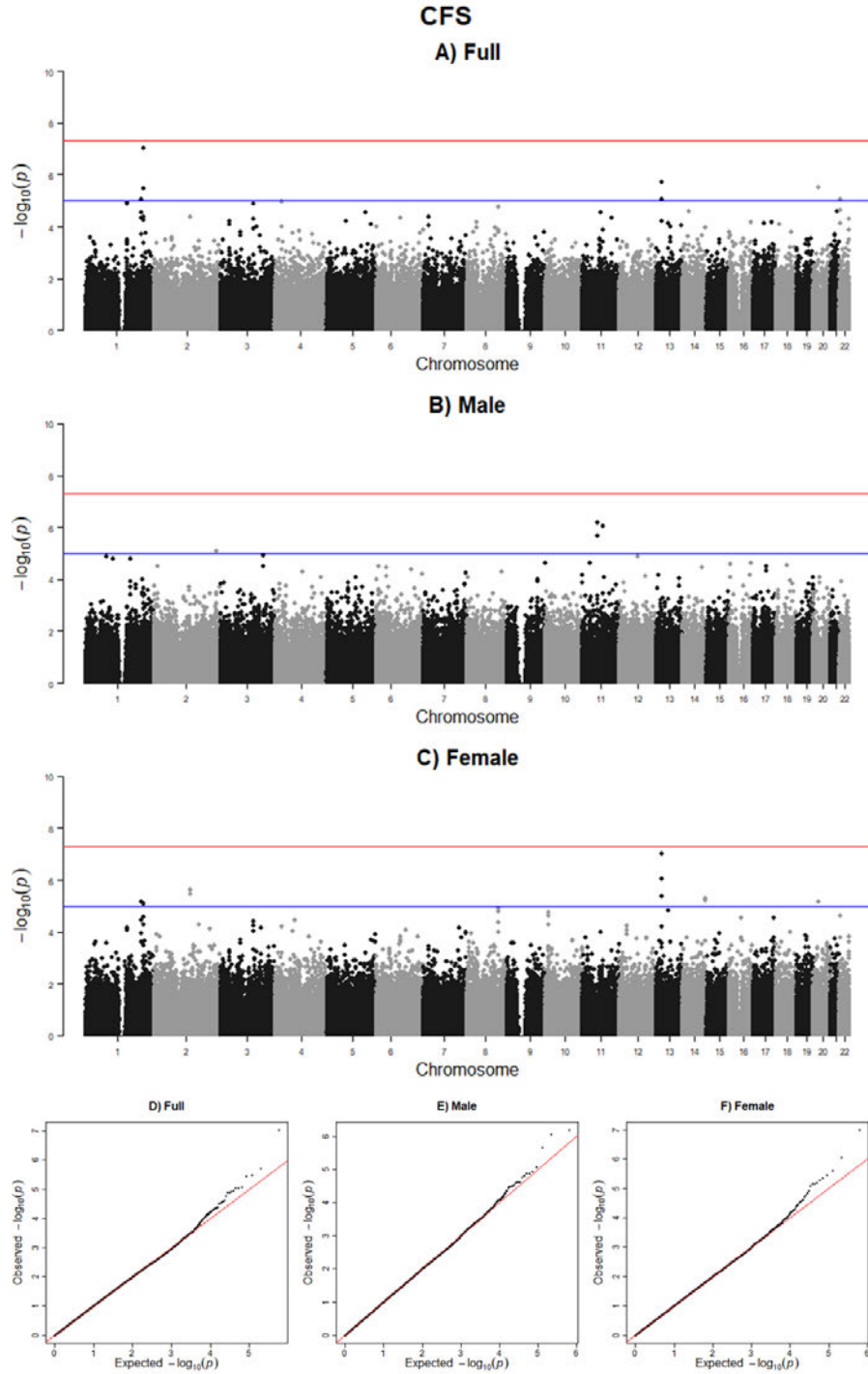


Figure 3.6: Manhattan and QQ plots showing the results of GWAS on the CFS phenotype in the UK Biobank, for the full (1,358 cases, 274,949 controls), male (368 cases, 128,678 controls), and female (990 cases, 146,271 controls) cohorts. Genomic inflation factors were $\lambda = 1.01$, $\lambda = 1.01$, and $\lambda = 1.01$ for these groups, respectively. The red lines on the Manhattan plots indicate genome-wide significance at $p < 5 \times 10^{-8}$ and blue indicates the suggestive threshold of $p < 1 \times 10^{-5}$. No genome-wide significant associations were detected, however, there is a suggestive signal on chromosome 1 for the CFS phenotype in males. There is also a suggestive association on chromosome 13 in females with the same phenotype.

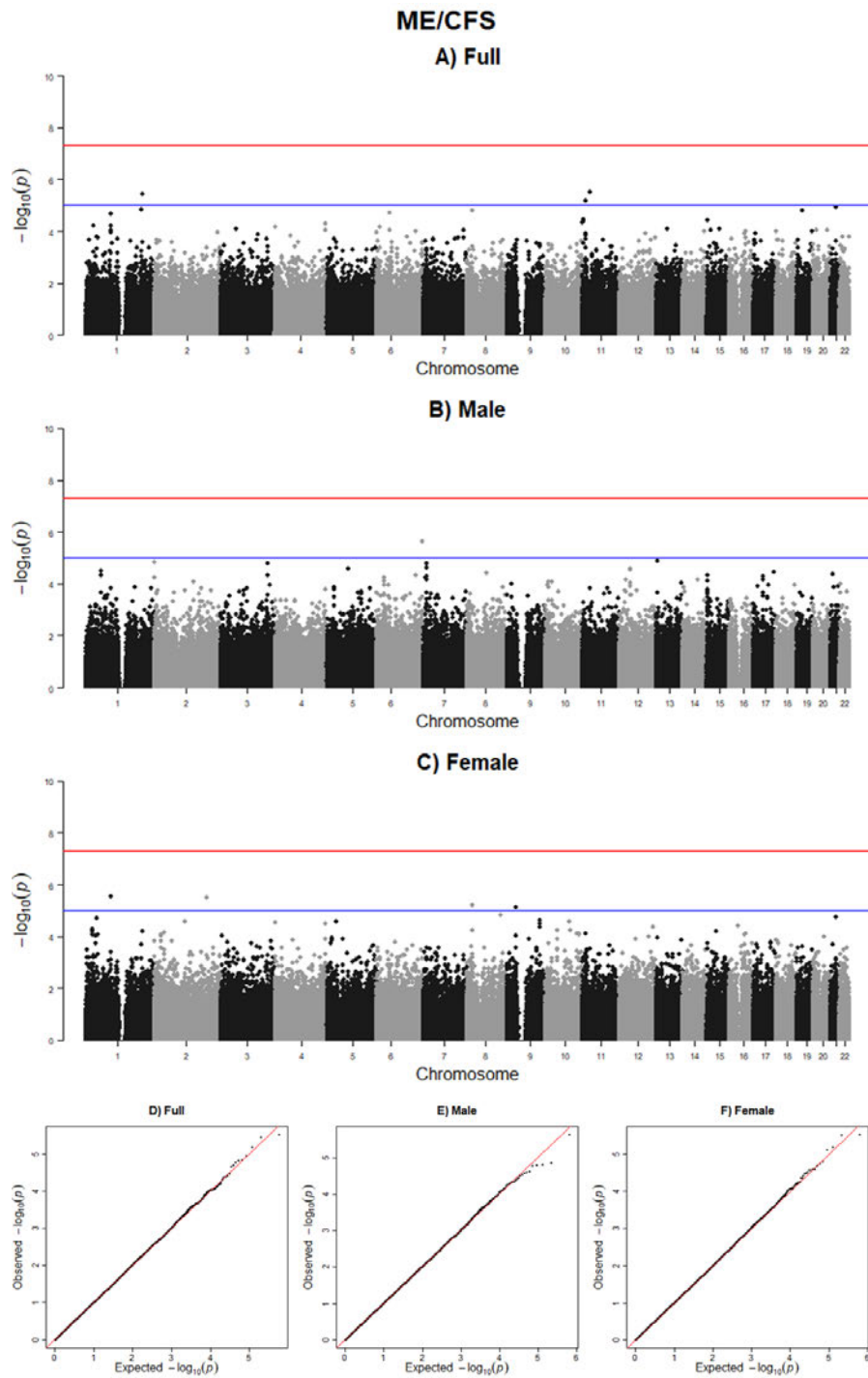


Figure 3.7: Manhattan and QQ plots showing the results of GWAS on the ME/CFS phenotype in the UK Biobank, for the full (1,561 cases, 94,720 controls), male (443 cases, 42,131 controls), and female (1,118 cases, 52,589 controls) cohorts. Genomic inflation factors were $\lambda = 1.02$, $\lambda = 1.02$, and $\lambda = 1.00$ for these groups, respectively. The red lines on the Manhattan plots indicate genome-wide significance at $p < 5 \times 10^{-8}$ and blue indicates the suggestive threshold of $p < 1 \times 10^{-5}$. No genome-wide significant associations were detected.

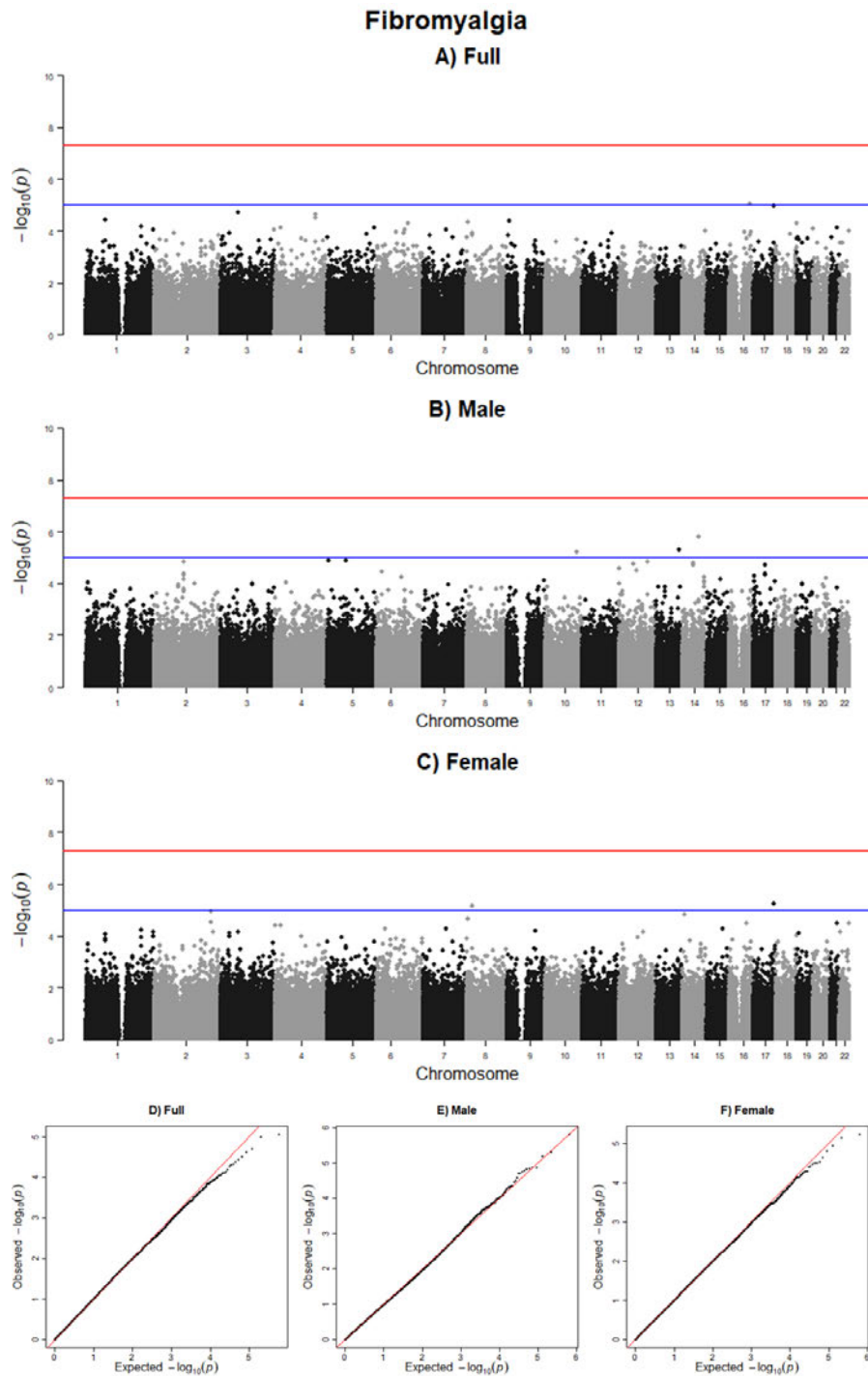


Figure 3.8: Manhattan and QQ plots showing the results of GWAS on the Fibromyalgia phenotype in the UK Biobank, for the full (554 cases, 275,753 controls), male (66 cases, 128,980 controls), and female (488 cases, 146,773 controls) cohorts. Genomic inflation factors were $\lambda = 1.01$, $\lambda = 1.01$, and $\lambda = 1.01$ for these groups, respectively. The red lines on the Manhattan plots indicate genome-wide significance at $p < 5 \times 10^{-8}$ and blue indicates the suggestive threshold of $p < 1 \times 10^{-5}$. No genome-wide significant associations were detected.

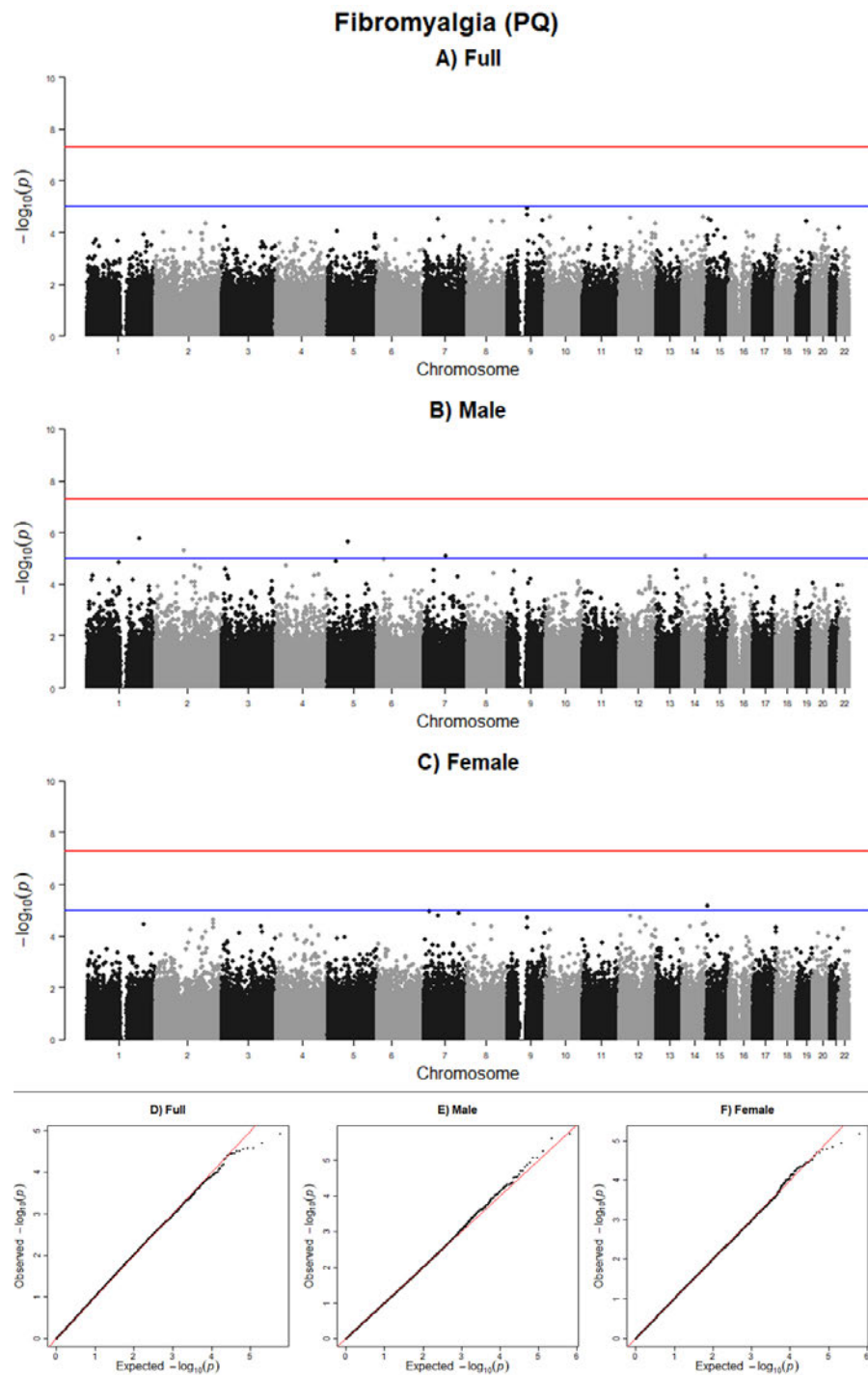


Figure 3.9: Manhattan and QQ plots showing the results of GWAS on the Fibromyalgia (PQ) phenotype in the UK Biobank, for the full (1,697 cases, 93,308 controls), male (291 cases, 41,597 controls), and female (1,406 cases, 51,711 controls) cohorts. Genomic inflation factors were $\lambda = 1.00$, $\lambda = 1.00$, and $\lambda = 1.02$ for these groups, respectively. The red lines on the Manhattan plots indicate genome-wide significance at $p < 5 \times 10^{-8}$ and blue indicates the suggestive threshold of $p < 1 \times 10^{-5}$. No genome-wide significant associations were detected.

3.1.6 Combined Phenotypes Results

The analyses were repeated for the constructed ME/CFS (Combined) and Fibromyalgia (Combined) phenotypes described in **Section 3.1.2**. Manhattan and QQ plots for the results for ME/CFS (Combined) are shown in **Figure 3.10**, with the same for Fibromyalgia in **Figure 3.11**.

No significant results were detected for fibromyalgia. However, the genome-wide significant association for ME/CFS in males (rs76346913) is as yet unreported in the literature. The implications of this result are discussed further in **Section 3.1.7**. The other suggestive results are summarised in **Section 3.1.7**. The QQ plots show slightly more than expected p -values around the suggestive threshold for males and females separately, but not when analysed together.

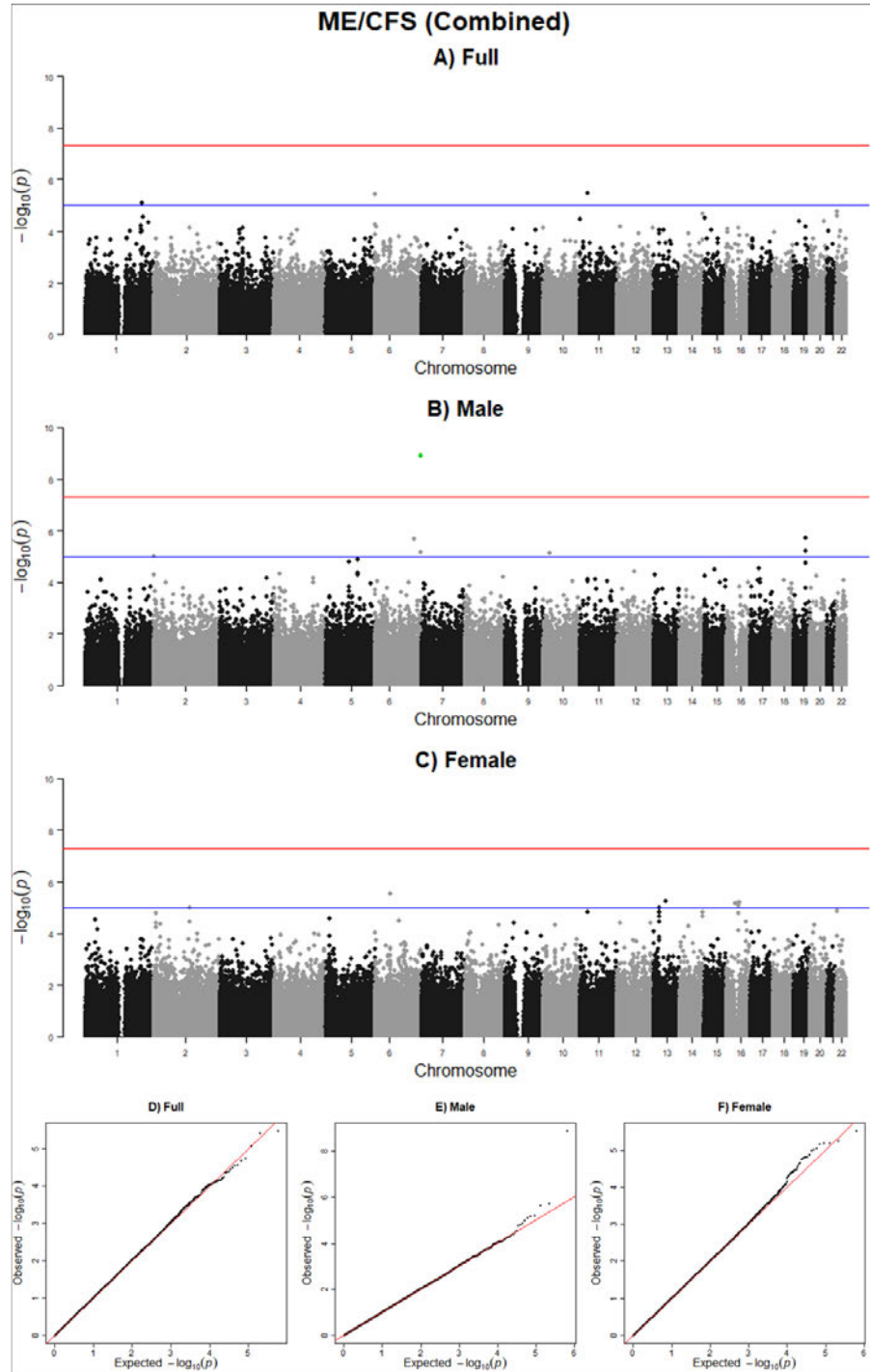


Figure 3.10: Manhattan and QQ plots showing the results of GWAS on the ME/CFS (Combined) phenotype, for the full (2,350 cases, 273,903 controls), male (666 cases, 128,365 controls), and female (1,684 cases, 145,538 controls) cohorts. Genomic inflation factors were $\lambda = 1.02$, $\lambda = 1.02$, and $\lambda = 1.00$ for these groups, respectively. The red lines on the Manhattan plots indicate genome-wide significance at $p < 5 \times 10^{-8}$ and blue indicates the suggestive threshold of $p < 1 \times 10^{-5}$. There was a genome-wide significant association on chromosome 6 for males (shown in green).

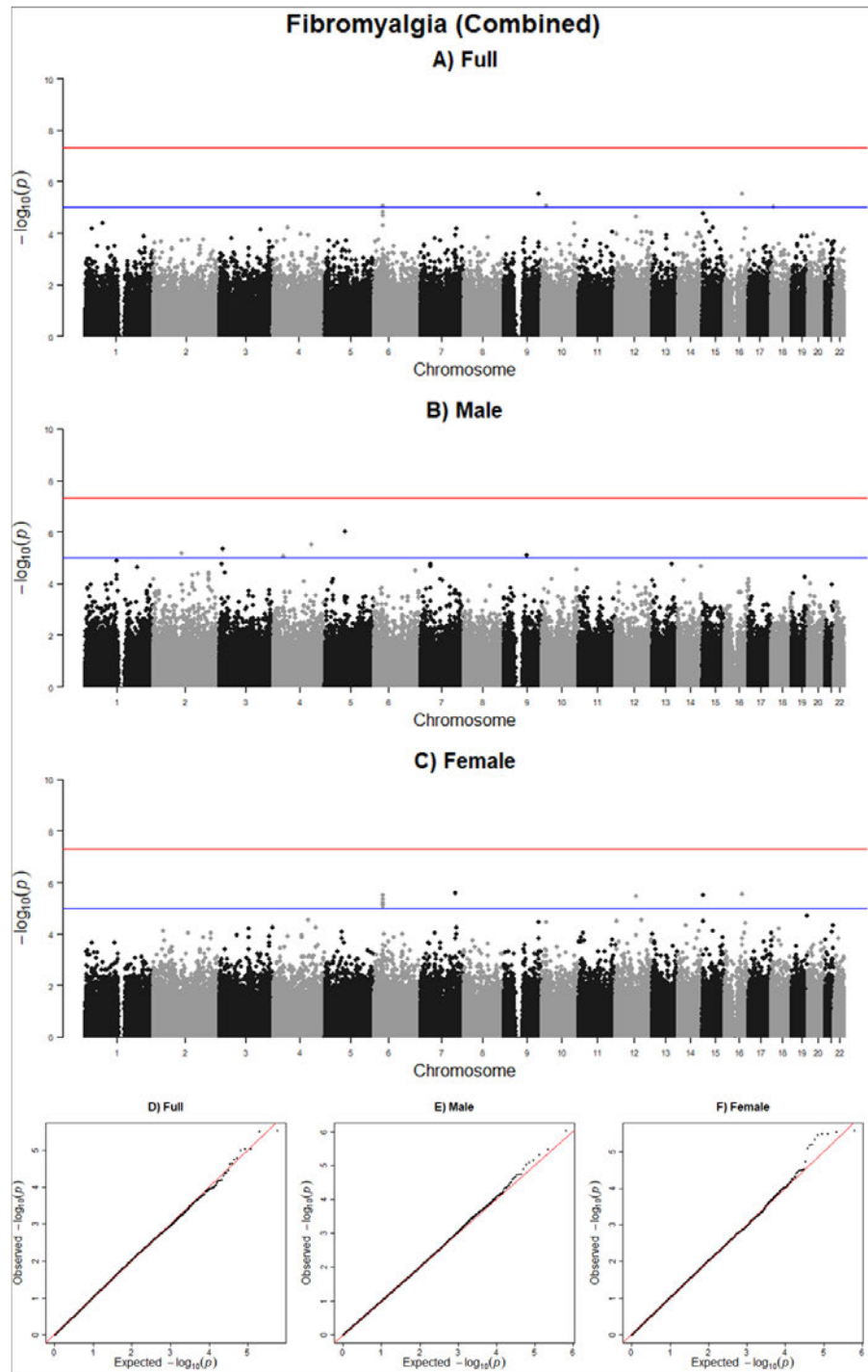


Figure 3.11: Manhattan and QQ plots showing the results of GWAS on the Fibromyalgia (Combined) phenotype, for the full (1,990 cases, 274,308 controls), male (322 cases, 128,724 controls), and female (1,668 cases, 145,584 controls) cohorts. Genomic inflation factors were $\lambda = 1.01$, $\lambda = 1.00$, and $\lambda = 1.02$ for these groups, respectively. The red lines on the Manhattan plots indicate genome-wide significance at $p < 5 \times 10^{-8}$ and blue indicates the suggestive threshold of $p < 1 \times 10^{-5}$. No genome-wide significant associations were detected.

3.1.7 Summary of Results and Interpretation

All suggestive SNPs ($p < 1 \times 10^{-5}$) from the analyses of **Sections 3.1.5** and **3.1.6** are summarised in **Table 3.5**. Most of these are not expected to be causal for ME/CFS or fibromyalgia, as they fail to reach genome-wide significance ($p < 5 \times 10^{-8}$). The notable exception of rs76346913 on chromosome 6 is also not guaranteed to be causal for ME/CFS in males, however it is the most interesting of these results and shall be discussed further, alongside two other variants near the genome-wide significance threshold for CFS.

Phenotype	Cohort	Chromosome	RefSNP	MAF	OR	p-value	Nearest Gene	Distance (bp)
CFS	Full	1	rs7518201	0.147	1.25	9.16E-06	ZC3H11A	8,345
		1	rs12145346	0.0987	1.36	9.49E-08	NEK2	76,486
		1	rs116109664	0.0558	1.40	3.55E-06	LPAT1	4,989
		13	rs7337312	0.493	1.20	1.97E-06	MRPS31	7,971
		13	rs1536805	0.404	0.837	9.34E-06	SLC25A15	6,110
		20	rs2092944	0.108	0.719	3.21E-06	RALGAP2	121,507
		22	rs73387688	0.119	1.28	8.67E-06	ADA2	44,199
	Male	2	rs1569090	0.160	1.49	8.31E-06	PID1	420,744
		11	rs12225155	0.431	0.679	2.13E-06	APLN	23,885
		11	rs10896588	0.374	0.665	6.50E-07	ANLN	18,240
		11	rs35008227	0.113	1.62	8.63E-07	WNT11	32,612
	Female	1	rs7518201	0.147	1.30	6.83E-06	ZC3H11A	8,345
		1	rs4844445	0.0924	1.36	8.06E-06	PLXNA2	581,404
		2	rs4954069	0.440	0.803	2.40E-06	NCKAP5	20,589
		2	rs10209819	0.494	0.801	3.45E-06	NCKAP5	32,279
		13	rs7337312	0.493	1.27	1.03E-07	MRPS31	7,971
		13	rs1536805	0.404	0.792	8.97E-07	SLC25A15	6,110
		13	rs2282025	0.400	0.804	4.38E-06	SLC25A15	17,527
		14	rs35165446	0.116	1.34	5.14E-06	SIVA1	4,369
		14	rs10142660	0.126	1.32	6.55E-06	SIVA1	6,136
		20	rs2092944	0.108	0.683	6.79E-06	RALGAP2	121,507
ME/CFS	Full	1	rs2761437	0.224	1.21	3.64E-06	CD46	2,301
		11	rs63357560	0.342	1.18	6.62E-06	-	-
		11	rs11030125	0.0794	0.693	3.08E-06	BDNF	11,873
	Male	6	rs76346913	0.0861	1.61	2.35E-06	PDE10A	251,943
	Female	1	rs55696855	0.0678	1.41	2.96E-06	GBPT	10,458
		2	rs12987796	0.150	1.29	3.05E-06	STAT1	54,819
		8	rs7836800	0.211	1.25	6.28E-06	PSD3	36,408
		9	rs10972298	0.0857	0.668	7.66E-06	C9orf131	9,748
	Full	-	-	-	-	-	-	-
Fibromyalgia (PQ)	Male	1	rs12142777	0.110	1.72	1.77E-06	RGS18	47,721
		2	rs10194273	0.219	1.51	5.32E-06	FHL2	12,068
		5	rs2314754	0.227	0.571	2.32E-06	ENC1	14,753
		7	rs572104	0.196	1.53	8.11E-06	MAGI2	617,174
		14	rs78706926	0.0978	1.69	8.38E-06	INF2	4,292
	Female	15	rs78861252	0.0660	1.36	6.73E-06	NPAP1	34,436
Fibromyalgia	Full	16	rs17707630	0.175	1.38	8.74E-06	MAF	691,367
	Male	10	rs290487	0.161	2.35	6.38E-06	TCF7L2	199,725
		13	rs9519664	0.274	2.22	5.14E-06	DAOA	47,001
		14	rs178348	0.353	2.33	1.56E-06	DIO2	784,388
	Female	8	rs75182652	0.0588	1.64	7.15E-06	GFRA2	201,045
		17	rs7216027	0.498	0.744	5.87E-06	DNAH17	82,861
ME/CFS (Combined)	Full	1	rs2761437	0.224	1.16	8.26E-06	CD46	2,301
		6	rs75294181	0.0944	1.24	3.69E-06	FOX2	22,228
		11	rs11030125	0.0794	0.752	3.34E-06	BDNF	11,873
	Male	2	rs56046241	0.0726	1.49	9.96E-06	TMEM18	119,782
		6	rs1013902	0.0818	1.50	2.20E-06	ADGRG6	96,737
		6	rs76346913	0.0861	1.64	1.31E-09	PDE10A	251,943
		6	rs78375762	0.178	1.35	6.70E-06	PDE10A	212,731
		10	rs2297325	0.382	1.28	7.58E-06	ARHGAP21	189,740
		19	rs349036	0.137	1.38	6.04E-06	ZNF283	8,487
		19	rs12459186	0.137	1.40	1.92E-06	ZNF404	30,190
	Female	6	rs603164	0.217	1.21	3.01E-06	COL21A1	14,496
		13	rs7337312	0.493	1.17	9.76E-06	MRPS31	7,971
		13	rs74239375	0.170	0.785	5.51E-06	PCDH9	1,602,938
		16	rs117331002	0.0550	1.35	6.63E-06	IGHV3OR16-17	1,294,922
		16	rs77305868	0.0563	1.35	8.69E-06	SHCBP1	145,236
		16	rs12925077	0.0570	1.35	6.40E-06	DNAJA2	9,202
Fibromyalgia (Combined)	Full	16	rs77578510	0.112	1.27	6.39E-06	ITFG1	335,916
		6	rs707936	0.0574	1.31	9.22E-06	SAPCD1	3,217
		9	rs10819272	0.425	1.16	3.07E-06	GARNL3	32,210
		10	rs12217500	0.272	1.17	9.14E-06	PTER	85,850
		16	rs112907595	0.0786	1.29	2.97E-06	CDH5	305,343
	Male	18	rs4798686	0.0562	1.31	9.73E-06	MTCL1	80,752
		2	rs10194273	0.219	1.48	6.82E-06	FHL2	12,068
		3	rs6442468	0.0925	1.67	4.68E-06	C3orf20	7,102
		4	rs645469	0.301	1.44	9.41E-06	PTTG2	160,440
		4	rs299133	0.0930	1.69	3.24E-06	SLC7A11	278,214
		5	rs2314754	0.227	0.576	9.40E-07	ENC1	14,753
	Female	9	rs76893567	0.0500	4.63	7.96E-06	IDNK	28,406
		6	rs376510	0.0574	1.35	6.34E-06	LY6G6C	1,323
		6	rs400547	0.0574	1.35	4.59E-06	DDAH2	4,316
		6	rs707915	0.0575	1.34	8.18E-06	MSHS-SAPCD1	3,171
		6	rs1150793	0.0576	1.34	6.71E-06	MSHS-SAPCD1	9,899
		6	rs707936	0.0574	1.36	3.25E-06	SAPCD1	3,217
		7	rs11768964	0.0810	1.31	2.73E-06	SMKR1	2,622
		12	rs1691590	0.265	0.825	3.50E-06	CSRP2	53,815
		15	rs78861252	0.0660	1.34	3.20E-06	NPAP1	34,436
		16	rs112907595	0.0786	1.31	2.93E-06	CDH5	305,343

Table 3.5: Suggestive and genome-wide significant SNPs from the GWAS of Sections 3.1.5 and 3.1.6 ($p < 1 \times 10^{-5}$). The genome-wide significant result is marked in bold. Note that the MAF for each SNP was calculated using the filtered cohort of Section 3.1.1. The distance to the nearest gene is shown for each variant, in base pairs (bp), as identified by Open Targets [161].

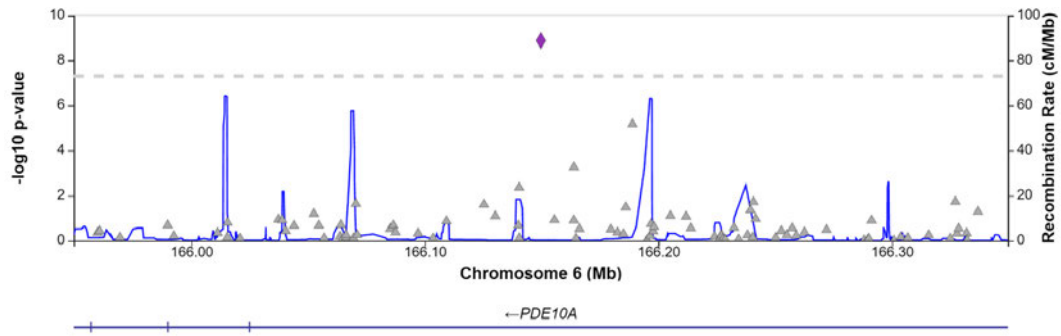


Figure 3.12: LocusZoom plot [201] of a section of chromosome 6, showing the risk variant *rs76346913* (purple diamond). Recombination hotspots are visible as spikes in the recombination rate (blue line).

rs76346913, Chromosome 6

There was only one genome-wide significant result from these analyses: *rs76346913* for the ME/CFS (Combined) phenotype, for males only, with $p = 1.31 \times 10^{-9}$ (**Figure 3.12**). It is notable that this result remains significant even after correcting for the multiple phenotypes tested (since these phenotypes are not independent, $\frac{5 \times 10^{-8}}{6} = 8.33 \times 10^{-9}$ would be overly strict). This association was not found in the equivalent female cohort ($p = 0.737$). When investigating this SNP in the uncombined male ME/CFS and CFS analyses, it reaches the suggestive threshold for the ME/CFS phenotype ($p = 2.35 \times 10^{-6}$) and is reduced for CFS ($p = 6.33 \times 10^{-5}$). These are not independent of the ME/CFS (Combined) phenotype, but their intersection (ME/CFS and CFS) is small relative to their union, and so they are somewhat independent of each other. Note that there should be no duplication of individuals in the ME/CFS (Combined) phenotype.

This is a multi-allelic site, which complicates the interpretation. There are two possibilities at this site: a $C \rightarrow T$ nucleotide substitution, or a G deletion ($CG \rightarrow C$). The deletion is rare (~ 0.0090 in the non-Finnish European population [202]), and therefore this variant would have been filtered during the PLINK analyses. Hence, only the more common variant is considered. In the UK Biobank genotyping, the two alleles at this position are coded as reference (T) and alternate (C), with T being the minor and risk allele for ME/CFS. Its frequency was consistent across the subgroups, as shown by the MAFs in **Table 3.6**, and with the value reported in gnomAD (~ 0.085 in non-Finnish Europeans) [202]. Note that gnomAD reports the values for non-Finnish Europeans as Finns are distinct from other European genetic ancestries [203]. This variant also passed the HWE test ($p = 0.917$, as calculated in PLINK2 [204]).

Cohort	Full	Male	Female
ME/CFS (Combined)	0.101	0.133	0.088
Controls	0.086	0.086	0.086
Both	0.086	0.086	0.086

Table 3.6: Minor allele frequencies for *rs76346913* ($C \rightarrow T$).

The odds ratio of 1.64 is among the largest in **Table 3.5** (joint 9th greatest), with a low standard error of 0.0814 ($OR = 1.02$, $SE = 0.0614$ in females). Given the notably high odds ratio for a variant associated with a complex disease, it might be expected that this association would have been found in previous studies. For a risk variant with $OR = 1.64$, $RAF = 0.133$, and phenotype prevalence 0.0052 (as for males in the ME/CFS (Combined) cohort), the power is 97.0% to detect a genome-wide significant association, and hence this study is appropriately powered for this finding [181]. However, to date, only one other GWAS (Neale, *et al.*) has separated a CFS cohort by sex, and they had notably fewer cases available (451 compared with 666). Very few studies are conducted on males alone due to the pronounced sex bias of ME/CFS. It is therefore plausible that this is the first such investigation with sufficient power to detect this association.

This SNP lies within the gene *PDE10A*, which encodes a protein responsible for regulating cyclic adenosine monophosphate (cAMP) and cyclic guanosine monophosphate (cGMP) levels in striatal medium spiny neurons [205], which are localized to the striatum in the basal ganglia [206]. This SNP is not an eQTL for any other genes, and hence *PDE10A*, as the nearest gene, is the most likely to be affected [161]. This is confirmed by the UCSC Genome Browser “GeneHancer” tool [207], which shows no evidence of longer-range interactions with any other genes [208]. The striatum in particular is believed to play a role in cognition [209], and the basal ganglia as a whole is associated with many neuropsychiatric disorders. In particular, it is linked with motor dysfunction in Huntington’s and Parkinson’s diseases [210], where levels of *PDE10A* protein and mRNA are reduced [205]. Mutations in *PDE10A* have also been found to cause Chorea (an involuntary movement disorder) [205][211], and it is clear therefore that this gene plays a significant role in motor control. Inhibition of *PDE10A* results in an increase in cAMP and a similar clinical effect to D2 (dopamine) antagonists, which are a standard antipsychotic medication [206]. This means there may already be a potential treatment available for clinical trials, should this genetic association be replicated.

It is surprising to find this association only in the male cohort. ME/CFS is notably more prevalent in females, and this result was not replicated in that group. In

addition, no other study has previously found an association between this SNP and ME/CFS. Hence, this could imply a technical error or artefact, or merely be a consequence of the heterogeneity of ME/CFS. However, to support this as a true association, it is encouraging that there are two other neighbouring SNPs on chromosome 6 with p -values near the suggestive threshold for this phenotype in males: rs1013902 and rs78375762. rs1013902 has the more notable p -value ($p = 2.20 \times 10^{-6}$), but is not in LD with rs76346913 ($D' = 0.0203$). However the second SNP (rs78375762, $p = 6.70 \times 10^{-6}$) is in strong linkage disequilibrium with rs76346913 ($D' = 0.912$). The MAF for rs78375762 is also in agreement with that reported by gnomAD [202] for non-Finnish Europeans (0.18 in both these data and gnomAD). Linkage disequilibrium with another SNP in the same region, also with a noticeable p -value, increases the likelihood that the finding for rs76346913 is a true association.

Lending further support, research has shown that variants within *PDE10A* are associated with thyroid stimulating hormone (TSH) levels to a much greater degree in males [212]. rs78375762 is not in LD with rs753760 ($D' = 0.037$) [213], the lead SNP identified in this study, implying that neither is rs76346913. However, this reported male-bias in *PDE10A* function is further evidence that a true male-specific genetic association for rs76346913 has been detected, given that this association was not also found in females.

In addition, *PDE10A* function has been associated with blood pressure, which is consistent with the symptoms of ME/CFS that overlap with POTS (see **Section 1.5.4**). Warren, *et al.* report that the SNP rs147212971 is protective of elevated diastolic blood pressure [214], and this SNP appears to be in perfect linkage disequilibrium with rs78375762 ($D' = 1$) [213], and hence is in LD with rs76346913.

No other significant associations with this SNP are reported in Open Targets or GeneAtlas. As always, any significant association requires replication in a distinct study and population, and a collaboration with 23andMe using their Research Cohort was established [215]. Their analysis in this population did not replicate the association. Comments on other potential follow-up studies are included in **Chapter 6**.

rs12145346, Chromosome 1

For the full CFS cohort, rs12145346 on chromosome 1 lies just below the threshold for genome-wide significance ($p = 9.49 \times 10^{-8}$). As with the other highlighted

SNPs, it is in strong LD ($D' = 0.973$) with a suggestive SNP (rs116109664, $p = 3.55 \times 10^{-6}$), implying that it is worthy of further consideration.

This SNP lies near, and is an eQTL (principally in thyroid tissue) for, the gene *LPGAT1* [161]. Knockout of this gene has been shown to lead to effects associated with metabolic impairment [216] and disease [217].

No genome-wide significant associations are reported for rs12145346 in Open Targets or GeneATLAS. However, both report that this SNP is associated with CFS at, or near the suggestive level. Open Targets references two studies, one with $p = 2.4 \times 10^{-6}$ [218], and another with $p = 2.0 \times 10^{-5}$ [120]. GeneATLAS reports a p -value of $p = 6.84 \times 10^{-7}$, comparable to the significance in this analysis. Since all of these results were generated using the UK Biobank data, replication in a distinct cohort is required.

rs10896588, Chromosome 11

For males with CFS, the SNP rs10896588 had the smallest p -value. Although not genome-wide significant, with $p = 6.50 \times 10^{-7}$, it is very clearly above the suggestive threshold. However, any conclusions for this SNP must be treated with a great degree of caution.

Once again, it is not clear why this variant has a distinct p -value for males only ($p = 0.801$ in females with CFS). But to support that this is a true association, it is also in very strong LD ($D' = 0.997$) with another nearby variant with a suggestive p -value (rs12225155, $p = 2.13 \times 10^{-6}$).

rs10896588 is an eQTL for the genes *TIMM10* and *SLC43A3* (in blood cells). *TIMM10* encodes a mitochondrial inner membrane transporter [219], and *SLC43A3* encodes a nucleobase transporter [220]. Knockdown of *SLC43A3* results in an increase in intracellular cAMP production, and overexpression results in an increase in extracellular cAMP production [221]. Notably, *PDE10A* (from **Section 3.1.7**) also affects cAMP regulation by hydrolysing it into non-signalling AMP [206].

No genome-wide significant associations for rs10896588 are reported in Open Targets or GeneATLAS. This association should be treated with additional caution until replication is achieved using a larger, distinct, cohort.

Dietary Protein Analysis (rs7337312, Chromosome 13)

As a result of the GWAS of **Section 3.1.5 (Figure 3.6C)**, I also observed a plausible, albeit not genome-wide significant, association between the *SLC25A15* gene and the CFS phenotype in females. This is somewhat consistent with the results of Neale *et al.* [120], who reported a genome-wide significant association for rs7337312. It is likely that the differences in p -values observed are due to study design. Although the model fit (logistic) was the same [222], the Neale analysis included more cases (1208 vs 990) as a result of the strict kinship filtering used in this project. This is the most likely cause of the differences in p -values, as having fewer cases has potentially reduced the power here. In addition, their analysis included further covariates (age^2 for all analyses, and additionally $age \times sex$ and $age^2 \times sex$ for the sex-specific analyses), although these will have likely reduced the probability of a type I error rather than increased it.

Hence, given the rationale presented above relating to dietary effects due to the HHH phenotype, I proceeded to investigate whether females with the potential risk alleles reported significantly altered dietary protein intake. **Figure 3.13** shows genetic coordinates 4.0×10^7 through 4.3×10^7 on chromosome 13, highlighting and annotating the SNPs with $p < 1 \times 10^{-5}$. These were the only variants within *SLC25A15* with notable p -values, and hence were the only SNPs investigated further.

To investigate this hypothesis required an estimate of individuals' typical dietary protein intake. Two methods are proposed in the literature for estimating protein intake. The first is to simply take the self-reported estimate provided in the questionnaire given to UK Biobank participants (the "Estimated nutrients yesterday - Protein" phenotype) [223]. This is limited by the time frame (previous 24 hours, by recall), and by individuals' ability to estimate their macronutrients. The second is to attempt to estimate individuals' protein intake from their answers to the food intake questionnaire [224]. This method is substantially more complicated and prone to the same errors in self-reporting. Hence, I chose the first method, averaging protein intake over the five time points provided by the UK Biobank (where relevant data was available). Plotting the distribution of these protein intakes by case/control status shows that they do not visibly differ by group (**Figure 3.14**). A t-test showed that the mean intake of protein was lower for the cases than the controls ($p = 0.024$). However, when separated by sex, differences were not significant. It is clear that males report significantly greater consumption of dietary protein than females (mean 87.2g versus 78.4g, $p < 2.2 \times 10^{-16}$), and

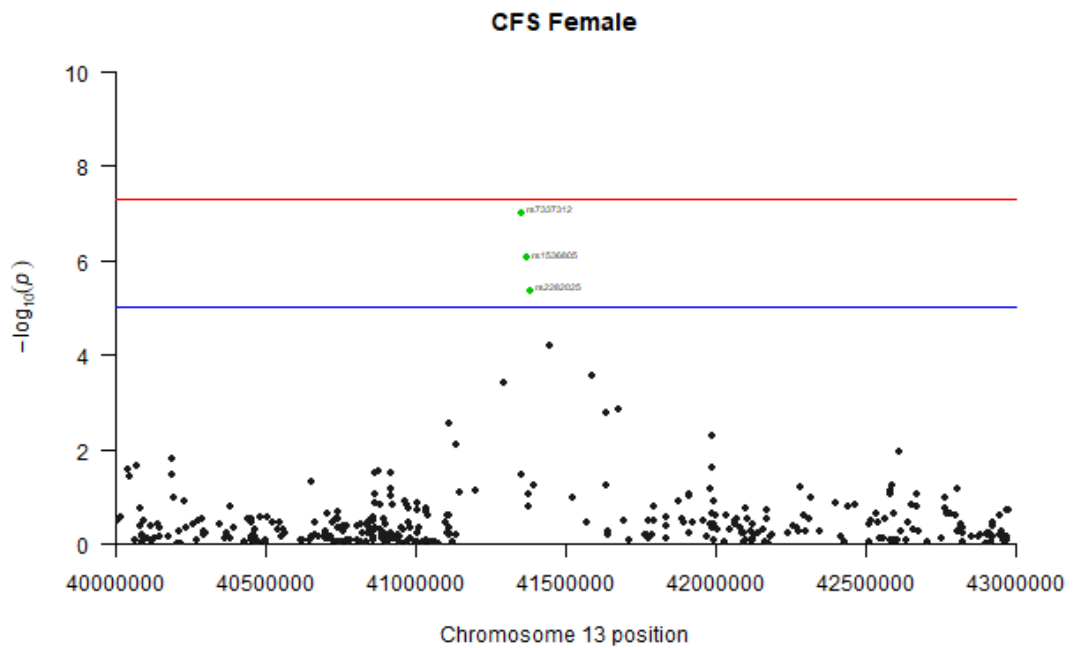


Figure 3.13: Part of chromosome 13 from **Figure 3.6C**, with the three SNPs with $p < 1 \times 10^{-5}$ around *SLC25A15* highlighted in green.

Chromosome	Position	rsID	Counts	β	Std. Error	t-statistic	p-value
13	41353297	rs7337312	65494	0.0580	0.126	0.461	0.645
13	41369658	rs1536805	65385	-0.191	0.128	-1.49	0.136
13	41381075	rs2282025	65071	-0.204	0.129	-1.58	0.114

Table 3.7: Results of fitting a GLM (using PLINK2 [171]) to the most promising SNPs from the female CFS GWAS of **Section 3.1.5**, as highlighted in **Figure 3.13**. No significant associations were detected between average dietary protein intake in females and any of the three variants tested.

that this is driving the case/control difference. This is to be expected given the sex imbalance in ME/CFS, however there may also be other confounders (such as socio-economic status) impacting dietary protein intake that are not considered here. Case/control imbalance may also have affected these estimates [225].

Further, a (generalised) linear model was fit to each of the three SNPs with the smallest p -values on chromosome 13 around *SLC25A15* ($p < 1 \times 10^{-5}$), against the estimated protein intake calculated as above. The covariates used were as for the GWAS of **Sections 3.1.5** and **3.1.6**. This found that there was no significant association between protein intake and the suggestive SNPs (**Figure 3.13**) on chromosome 13 in females (**Table 3.7**).

None of these SNPs were significant for any acceptable threshold, hence this investigation failed to reject the null hypothesis. There is no evidence to support

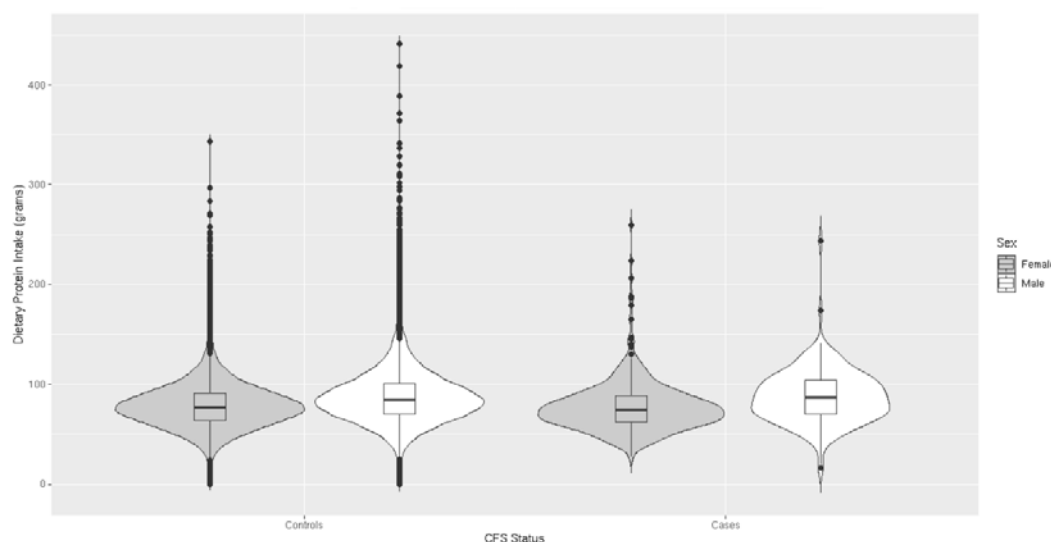


Figure 3.14: Violin plots of average self-reported dietary protein intake (in grams) for CFS cases and controls, separated by sex.

alteration of dietary protein intake in females carrying the potential CFS risk alleles.

Further investigation of these SNPs in Open Targets and GeneATLAS even suggests that they might be protective of CFS in the full cohort. rs7337312 was not significantly associated with CFS in GeneATLAS, but was suggestive in Open Targets ($p = 4.0 \times 10^{-6}$) with a slightly negative effect [120]. rs1536805 had three suggestive results; two in Open Targets ($p = 7.0 \times 10^{-6}$ [120] and $p = 9.2 \times 10^{-6}$ [218], both with a slightly negative effect), and another from GeneATLAS ($p = 4.30 \times 10^{-6}$). rs2282025 was slightly below the suggestive level for the same three studies, again with a negative effect in Open Targets.

This implies that there is still considerable uncertainty as to the role of *SLC25A15* in CFS, despite the initially encouraging result found by Neale *et al.* The analyses presented here do not support the hypothesis, and no other study on UK Biobank data has replicated the Neale group result at the same level of significance.

3.2 Human Leukocyte Antigen Loci

The human leukocyte antigen (HLA) loci are located on chromosome 6, in the most gene dense region of the genome. They encode HLA molecules that present antigens to immune cells (see **Section 2.1.7**).

The goal of this part of the project was to use the larger UK Biobank cohort to replicate the work of Lande *et al.*, which identified locus-wise significant associa-

HLA-	A	B	C	DRB1	DRB3	DRB4	DRB5	DQA1	DQB1	DPA1	DPB1	Total
Alleles	55	127	32	59	6	3	4	14	18	8	36	362

Table 3.8: *Number of alleles imputed for each locus.*

tions between HLA loci and ME/CFS [146]. The authors do not explicitly state the model of choice for their study, hence here I fit logistic models to each allele, using the same covariates as described in **Section 3.1.4**.

The reference datasets for the HLA imputations were collected by the UK Biobank from 8 sources, with each dataset genotyped on separate arrays [169]. Each had independent quality control measures applied, and then a further round of QC after combining into a single dataset. Due to the relatively homogeneous genetic background of the UK Biobank (94% self-reported white British), the reference panel was optimised for those with European ancestry by excluding some of the reference datasets at some HLA loci [169]. HLA imputations (from the genotype data) are available for almost all participants in the UK Biobank.

HLA nomenclature varies, but typically includes at least the locus, allele group and allele number. The format used in the UK Biobank is:

$$HLA-X^{*ab:cd}$$

where X is the locus, ab is the two-digit allele group, and cd is the two-digit allele number (further information is sometimes included in additional, colon separated, fields). Allele groups (or supertypes) define sets of alleles where a peptide which binds to an allele in the group should bind to all others in that group (in practice however, this is often not the case) [226].

The UK Biobank imputation includes 11 HLA loci: $HLA-A$, $-B$, $-C$, $-DRB1$, $-DRB3$, $-DRB4$, $-DRB5$, $-DQA1$, $-DQB1$, $-DPA1$ and $-DPB1$, with 362 alleles in total. The number of alleles imputed for each locus is shown in **Table 3.8**.

The HLA imputations are not stored with the other imputed genotypes in the UK Biobank. This is due to the multiple polymorphisms at each locus making them difficult to encode in the same way as regular SNPs (see **Section 2.1.7**). Some judgement must be exercised with respect to their interpretation. Each score in the imputation table represents the posterior probability of a given allele being present in an individual [227]. The alleles with greatest posterior probability are reported for each locus. 0 indicates that the allele is not present, heterozygotes have the maximum posterior probability reported for each allele, and homozygotes have the sum of the probabilities for both alleles recorded. Following the recommendation

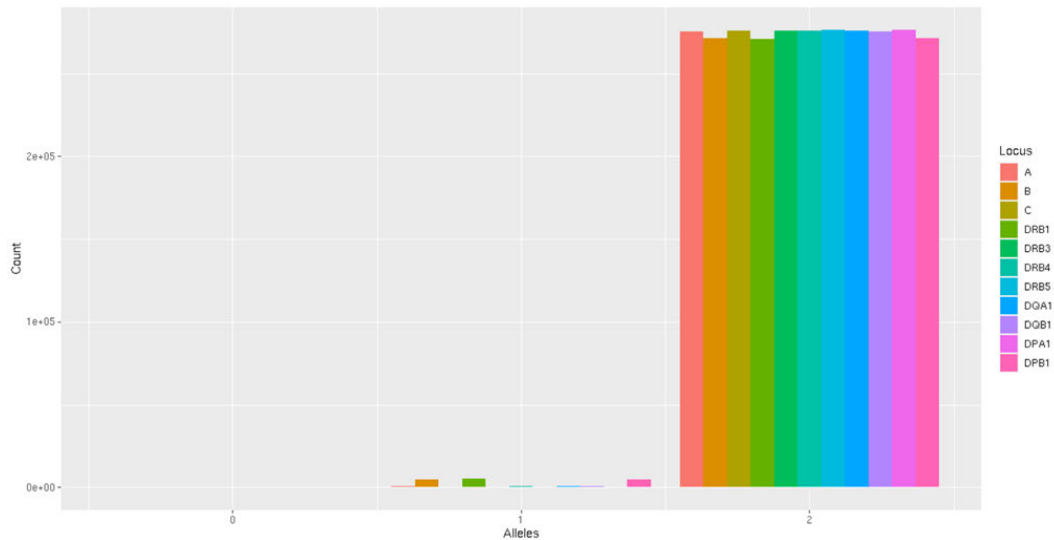


Figure 3.15: Distribution of alleles per locus [175]. Low quality imputations had been removed ($Q < 0.7$), resulting in < 22 alleles in total for some participants.

in the UK Biobank documentation, I imposed posterior probability thresholding at 0.7, by setting any entry with a lower probability to *NA* [227]. The remainder were rounded, giving possible outcomes of 0, 1 or 2. These were respectively interpreted as “allele not present”, heterozygous for the allele, and homozygous for the allele, in an additive model (see **Section 3.1.4**).

Figure 3.15 shows the allele distributions for each locus. No participant reported more than two alleles per locus. Those with a low-quality imputation, with < 2 alleles for a given locus, were removed for the relevant allele test. All other participants were retained.

3.2.1 Association of HLA Alleles with ME/CFS

The objective was to determine whether there were any associations between the HLA alleles and the fatiguing disease phenotypes. I wrote *R* code to fit a logistic model to each allele in turn, for all six phenotypes of **Section 3.1.1**. All 11 loci and 362 alleles were used in the analysis.

Since the alleles within each HLA locus tend to be in very strong linkage disequilibrium with each other, Lande *et al.* considered a locus-wise Bonferroni multiple testing correction to be more appropriate than allele-wise [146]. Since 11 HLA loci have been imputed in the UK Biobank data, this implies the significance threshold would be $p < \frac{0.05}{11} \approx 0.00455$. However, this analysis is more similar to a “Candidate gene” approach, where an a priori hypothesis is made that these particular

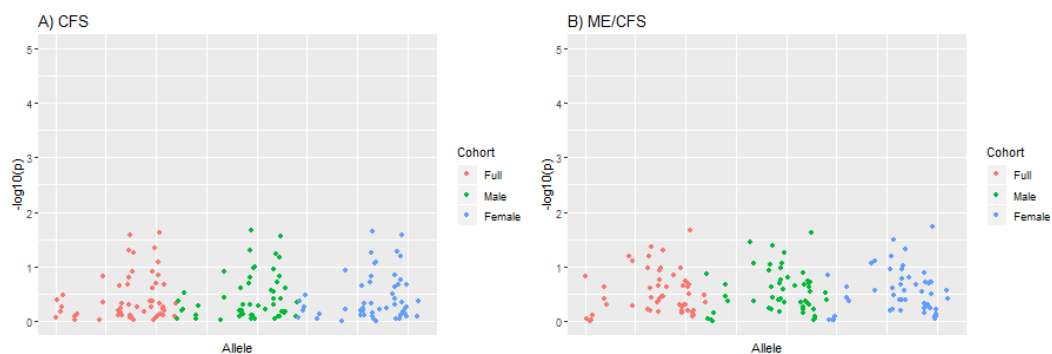


Figure 3.16: Scatter plots showing the significance values of HLA alleles with frequency ≥ 0.05 for **A) CFS** and **B) ME/CFS**. No alleles were significant after genome-wide multiple testing correction ($p < 5 \times 10^{-8}$).

loci will be associated with a phenotype. This is not an unreasonable assumption for autoimmune disorders, which tend to be strongly associated with at least one HLA locus [228]. However, the candidate gene approach is now considered largely defunct in favour of the GWAS [143]. In any case, given the unknown aetiology of ME/CFS, it is more appropriate to apply the genome-wide significance threshold of $p < 5 \times 10^{-8}$, or at the very least the suggestive threshold of $p < 1 \times 10^{-5}$.

The CFS and ME/CFS phenotypes were tested independently. **Figure 3.16** (panels **A** and **B** respectively) shows the p -values for alleles with frequencies ≥ 0.05 .

None of these alleles were genome-wide significant, and hence I conclude that the null hypothesis of no association between ME/CFS (and CFS) and the HLA alleles should not be rejected.

3.2.2 Association of HLA Alleles with Fibromyalgia

The HLA association tests were repeated for the fibromyalgia phenotypes (**Figure 3.17**). No recent investigation into the genetics of fibromyalgia has found an association with an HLA locus, despite earlier attempts using a candidate gene approach [229].

There were no significant associations between either fibromyalgia phenotype in the UK Biobank and any HLA loci after filtering (allele frequency ≥ 0.05) and the genome-wide multiple hypothesis testing correction. Hence, there is insufficient evidence to reject the null hypothesis of no association between the HLA loci and fibromyalgia.

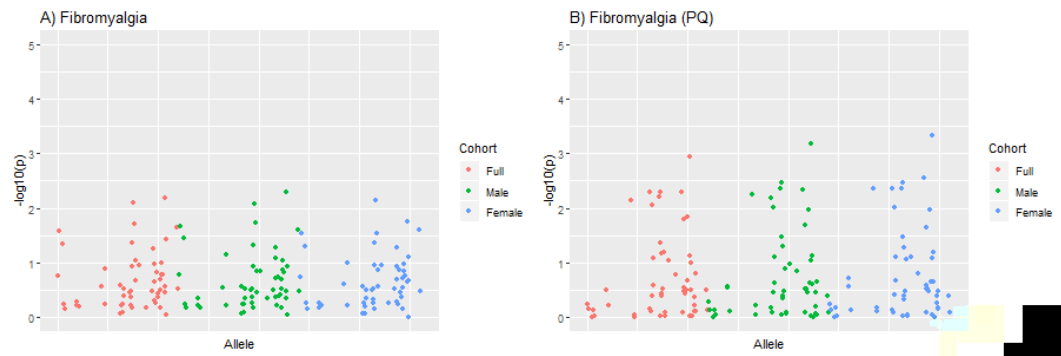


Figure 3.17: Scatter plots showing the significance values of HLA alleles with frequency ≥ 0.05 for **A) Fibromyalgia** and **B) Fibromyalgia (PQ)**. No alleles were significant after genome-wide multiple testing correction ($p < 5 \times 10^{-8}$).

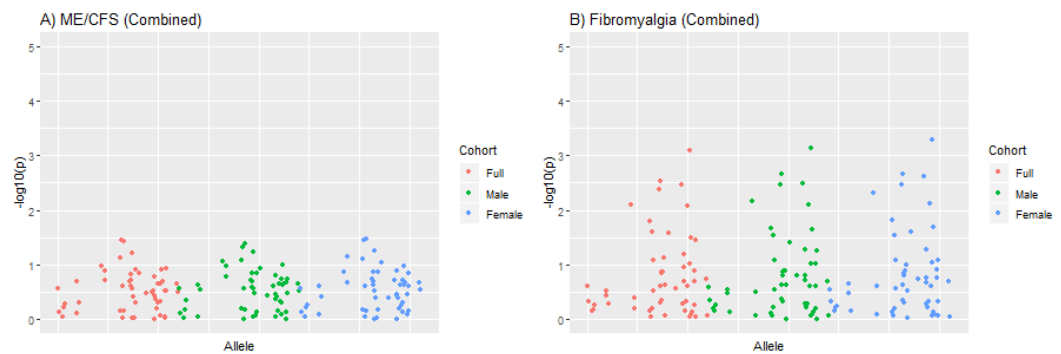


Figure 3.18: Scatter plots showing the significance values of HLA alleles with frequency ≥ 0.05 for **A) ME/CFS (Combined)** and **B) Fibromyalgia (Combined)**. No alleles were significant after genome-wide multiple testing correction ($p < 5 \times 10^{-8}$).

3.2.3 Combined Cohorts

Finally, these analyses were repeated with the ME/CFS (Combined) and Fibromyalgia (Combined) phenotypes of **Section 3.1.2**. The results are shown in **Figure 3.18**.

As with the standard UK Biobank phenotypes, no genome-wide significant associations were detected. Hence, there remains no evidence to support a link between these diseases and the HLA loci. The null hypothesis of no association should not be rejected.

3.3 Discussion

From these analyses, by far the most noteworthy was the genome-wide significant association found between rs76346913 and males with ME/CFS.

Interpreting the exact biological pathway through which this may influence ME/CFS is not straightforward. This variant is closest to the gene *PDE10A*, but is not an eQTL for it. It hence presumably has a regulatory effect, but exactly how this would cause the symptoms of ME/CFS is unclear. Disruption to the hydrolysis of cAMP and cGMP is a plausible cause for the neurological symptoms inherent to ME/CFS, but these are fundamentally different to those seen in the other diseases and disorders associated with this gene (ME/CFS does not typically include involuntary movements). Another potential pathway would be through endothelial tissue and blood pressure, but the evidence here is weak – relying solely upon the association seen here, and the work of one other study linking protection against high blood pressure with *PDE10A*. However, other variants of *PDE10A* show a marked sex-bias (related to TSH), which although not directly linked to rs76346913, is consistent with this sex-specific finding. Significant research into the functions of *PDE10A* would be needed in order to construct a more robust causal link to ME/CFS.

A suggestive association was seen at the same position as Neale, *et al.* on chromosome 13 for females. As discussed, the difference in *p*-value was likely due to slight differences in study design, and hence there remained sufficient justification for the follow-up analysis investigating dietary protein intake. The results here did not support the hypothesis, however, self-reported protein consumption data should be regarded with a great degree of caution. To fully refute this hypothesis would require a controlled diet study, or better, direct measurement of *SLC25A15* mRNA or ORNT1 protein levels in the relevant (particularly arterial) tissues of those with the risk allele.

There were limitations inherent to following a previous study design. Using an arbitrary number of principal components is common practice in GWAS, but the Biobank-provided set were calculated on the whole cohort rather than the White British subset. Although some remained significant predictors (see **Table 3.4**), an improvement in methodology would be to recalculate them for the “filtered” cohort, or even for each individual analysis. Alternatively, the genetic relationship matrix approach would avoid this limitation entirely, and a future study on the “Combined” phenotypes will use this approach. Incorporating a GRM will also

avoid the overly conservative filtering based on kinship, which will boost the power available ~ 2 -fold, and potentially enable the discovery of other variants associated with ME/CFS.

Further implications of this result, as well as discussions of potential follow-up studies, are included in **Chapter 6**.

Chapter 4

T-cell Receptor Repertoire Diversity: Background and Methods

4.1 Biological Background

The vertebrate immune system is divided between the innate and adaptive immune systems [230]. The innate immune system includes physical barriers (such as epithelial cells of the skin), and antimicrobial molecules (defensins), which form the first line of defence. There are also pathogen-associated molecular patterns (PAMPs) – features unique to pathogens which are not found in self-tissues. Examples include formylated methionine in the first amino acid position of a peptide instead of methionine, and certain cell surface characteristics, such as flagellum (both are characteristics unique to bacteria). Recognition of these by pattern-recognition receptors (PRRs) in blood or on cell-surfaces activates non-specific processes to target and destroy pathogens. These include cytokines (which can trigger inflammation), phagocytic cells, the complement system, interferons, and natural killer (NK) cells.

The link between these immune system levels is provided by antigen-presenting cells (APCs), which present pathogen protein peptides to the cells of the adaptive immune system. Activating T-cells is the only known function of dendritic cells, but macrophages and B-cells can also present antigens [148].

4.1.1 Adaptive Immune System

The adaptive immune system comprises T- and B-cells (lymphocytes), of which a large pool are produced in the thymus and bone marrow, respectively [230]. Each of these expresses a chain of amino acids which is able to bind to a particular peptide or antigen produced by a pathogen. There is considerable overlap in recognition of pathogens – a particular lymphocyte will recognise multiple antigens, and each antigen can be recognised by more than one receptor [231]. The antigen is presented to the T- or B-cell by an antigen-presenting cell (usually the aforementioned dendritic cells) [230]. B-cells can bind to whole antigens, whereas T-cells react only to peptide fragments [148]. These are formed by antigen-presenting cells, which fragment the antigen internally. These peptides are then transported to the surface of the APC by MHC proteins (encoded by the HLA region, see **Section 3.2**) for presentation to T-cells [148].

There are multiple types of T-cell, including cytotoxic, helper and regulatory [230]. The former directly kill cells infected by pathogens, and the latter two stimulate and suppress the activity of other immune cells, respectively. Alongside the receptor chain, T-cells express a co-receptor. This is a type of accessory receptor that aids binding, but also plays a role in activating the cell. Cytotoxic T-cells express CD8, and helper and regulatory cells express CD4 [230].

4.1.2 VDJ Recombination

To efficiently store the information needed to counteract the diversity of antigens encountered, the mammalian genome has developed methods of compression. By using combinations of separate gene segments, the human immune system is able to recognise the antigens it encounters without using a separate gene for each [148].

On the surface of each T-cell sits the receptor (TCR) – a chain of amino acids in two parts; the constant (C) region, and the hypervariable region. The latter is encoded by three gene segments – variable (V), diversity (D), and joining (J) [232]. There are four chains that are formed from these segments – α , β , γ and δ . They are paired $\alpha - \beta$ or $\gamma - \delta$, with the α and γ chains formed only from V and J segments, and β and δ formed from all three [233]. These are analogous to the light and heavy immunoglobulin chains of B-cells, which are formed from V-J and V-D-J recombination, respectively [148]. The D segment is not present

in α or γ chains as this is the first part to recombine - once the cell commits to this line, the recombination process excises the D segment [234]. In this project, a recombinant is defined as the V(D)J gene combination, plus the CDR3 region sequence (addition/deletion of P and N nucleotides; see below), and without considering the pairing of chains ($\alpha - \beta$ or $\gamma - \delta$). Note that there should be no sequences in the dataset that have not undergone recombination after assembly into clonotypes (see **Section 4.4**).

As examples, there are 54 known human V_β segments [231], 27 D segments [235], and 61 J_α segments [236]. Addition and deletion of nucleotides occurs at the joins of these gene segments through both random (N nucleotides) and templated (P nucleotides) processes [237], which increase the space of possible nucleotide base combinations to potentially 10^{18} in humans [238]. However, the number of combinations in practice is estimated to be much lower (for example, $\sim 10^6 - 10^8$ for β -chains [239]) This hypervariability allows the higher vertebrate immune system [240] to respond to a huge and diverse array of potential pathogens.

The recombination process folds the DNA to bring the correct gene segments together, then cuts to remove the unneeded loop (see **Figure 4.1**). Depending on the recombination process (whether the δ locus is excised), the join on the TREC can be a signal or coding junction [234]. In this project, analysis was performed on the coding junctions remaining on the genome (hence, there are no results for δ chains, as these are on the excised signal junctions here). The removed loops (sometime called T-cell recombination excision circles, or TRECs [234]) can form a useful measure of thymopoiesis (thymic activity) [241].

Excision is carried out by double-strand cleavage of the genome at specified recombination signal sequence (RSS) sites by RAG proteins [242]. The RSS are composed of fixed-length sequences of seven and nine base pairs lengths separated by 12- or 23-base pair sequences [243]. There is a strong preference after cleavage for these sites to combine in 12/23 length pairings (known as the 12/23 rule) [243], although this is not seen in some cases [244].

Activation occurs when a peptide from an antigen is presented to the TCR by a dendritic cell [230]. Upon activation, a T-cell will begin to clone itself repeatedly and produce a cytokine that further activates genes relevant to the differentiation process. This ensures a mass cloning event is triggered when multiple T-cells react to the same peptide, and the resulting swarm of effector cells perform their role against the pathogen [148]. This is the basis for immune repertoire analysis, as a picture of the whole repertoire can indicate an immune system challenge.

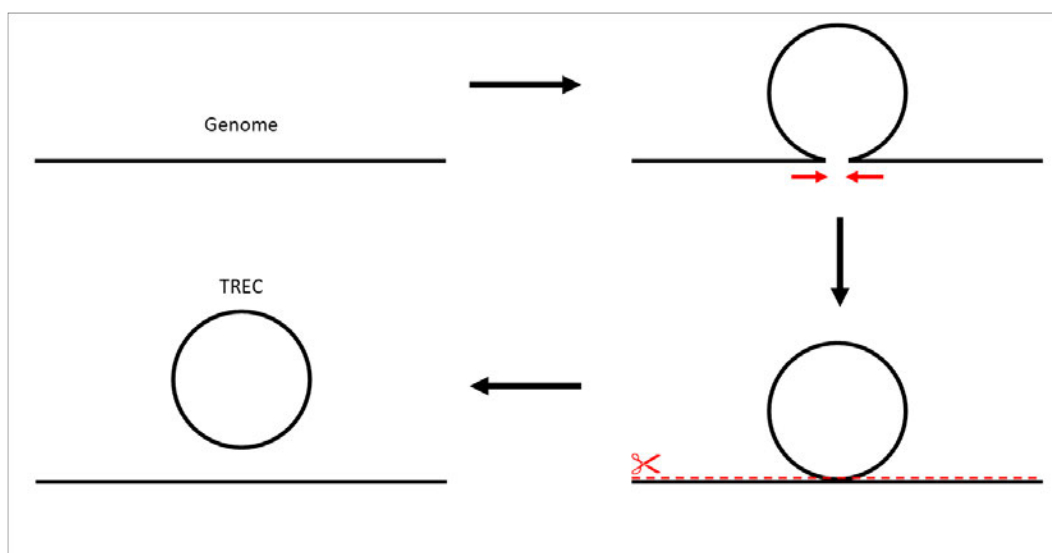


Figure 4.1: Graphic showing the recombination process. A V and a J gene segment are brought together by folding the genome. The resulting “loop” of DNA is excised, and the ends can merge to form a circle.

4.1.3 Clonal Activation in Disease

It has previously been suggested that during acute disease phases, there may be significant clonal expansion of T-cells. This has been indicated for Kawasaki Disease [245], IG4-related disease (IG4-RD) [246], and Ras-associated lymphoproliferative disease (RALD) [247]. Perhaps most relevant for ME/CFS though is the expansion seen in CMV infection [248][249], which as previously mentioned (see **Section 1.4**) is a proposed infectious trigger for the disease. CMV status shall be investigated further in this project in **Section 5.9**.

There are also indications that there may be preferential gene segment usage in multiple sclerosis (found in brain lesions [250] and cerebrospinal fluid [251]), although the evidence is not yet conclusive [252]. Non-random selection of gene segments would indicate that some T-cell receptors are more common in people with MS, which similar to TCR clonal expansion, could indicate the specific antigen (and hence pathogen) involved in disease incidence.

As discussed in **Section 2.2**, Dr Mark Davis had previously shown preliminary results that corroborated clonal expansion in Lyme disease, MS and CFS. The remainder of this thesis aimed to replicate this finding in an independent cohort of people with ME/CFS.

4.2 Experimental Design

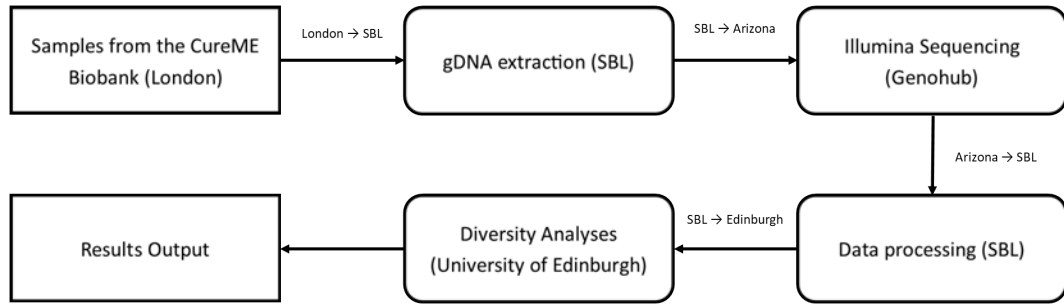


Figure 4.2: *An overview of the project workflow.*

The project used 160 peripheral blood mononuclear cell (PBMC) samples from the CureME Biobank [253], from the same number of donors. From these, $CD4^+$ (helper/regulatory) and $CD8^+$ (cytotoxic) T-cells [230] were separately isolated by Systems Biology Laboratory (SBL), project collaborators in Oxfordshire. The intent was to determine the clonotype repertoire of each sample. Samples were divided into four groups of 40:

- Severely affected (MEsa) people with ME/CFS (either house- or bed-bound);
- Mildly/moderately (MEmm) affected people with ME/CFS;
- Multiple Sclerosis (MS; disease controls); and,
- Healthy controls (HC).

The ME/CFS samples were from donors diagnosed as meeting either or both of the CCC and Fukuda criteria (see **Section 1.1**) [253]. As much as possible, these were age-matched, with the majority being between 40 and 60 years old at the time of sample collection. All were from female donors. These four categories shall henceforth be referred to as ME(sa), ME(mm), MS and HC, respectively. As described in **Section 1.5.9**, MS is relevant to the study of ME/CFS, and hence samples are already available in the CureME Biobank. The assignment of individuals to each of these four categories was concealed from all involved in the project for the initial analyses.

The PBMCs were processed by SBL using a multi-step pipeline to isolate the T-cells. An overview of these steps is provided in **Section 4.3**. Sequencing was conducted by Genohub [254] in Arizona, USA. SBL received the sequencing output from Genohub and conducted the initial data processing. The resulting JSON file was forwarded to me for conversion into a clonotype vector, and estimation of sample diversity.

4.3 Systems Biology Laboratory Processing

In this section, I describe the work carried out by collaborators at SBL.

The cells arrived at Systems Biology Laboratory in Milton, Oxfordshire by courier, stored in dry ice. From there, they were refrigerated in liquid nitrogen at -180°C . The samples were anonymised, randomly assigned a unique SBL identifier (*SBL1*–*160*), and organised into groups of three (the maximum number that could be processed in one day). Processing at this stage was done by Dr Matt Roddis.

The PBMCs were first sorted by the Magnetic Activated Cell Sorting (MACS) system [255], which uses antibodies attached to magnetic beads to select the desired cell types. Fluorescence Activated Cell Sorting (FACS) [256] can be more efficient, but would have required off-site equipment, and would have introduced transport issues into the workflow. The antibodies were chosen to either positively select the cell type of interest, or negatively select the unwanted cells. First, the system was set for positive selection of CD8^{+} cells, then positive selection of CD4^{+} , then $\gamma\delta$ negative selection. After each selection, the material was refrigerated in bead form. CD8^{+} cells spent longest in refrigeration, but not sufficiently long that loss of DNA due to cell apoptosis was observed. The remaining material was mostly B-cells, which were refrozen for use in another project.

In almost all cases, adequate numbers of CD4^{+} and CD8^{+} cells were obtained, with only five samples yielding $< 2 \times 10^6$ PBMCs (only one sample failed the sort completely and was discarded). Unfortunately, amounts of $\gamma\delta$ T-cells were insufficient for analysis, but are being stored in the hope that pipeline improvements enable analysis at a later stage.

For verification, sorting results were visualised with plots of co-receptor expression (**Figure 4.3**), and gates for cell types were set manually. It was clear that there was visible clustering, so gates dividing by cell type were set by eye as a first order check that the sorting was successful. A minor batch effect was noted here, but variability was more pronounced between individuals than between batches, so was considered to have minimal impact.

After sorting, cells were lysed to release the DNA. A sonication step followed, which fragmented the chromosomes into smaller lengths to ensure that they eluted from the column. This step helped to prevent a bias towards smaller fragments, as the longer fragments can stick to the column and would be excluded from the results. DNA was then extracted using a Qiagen gDNA kit. Another 30 rounds

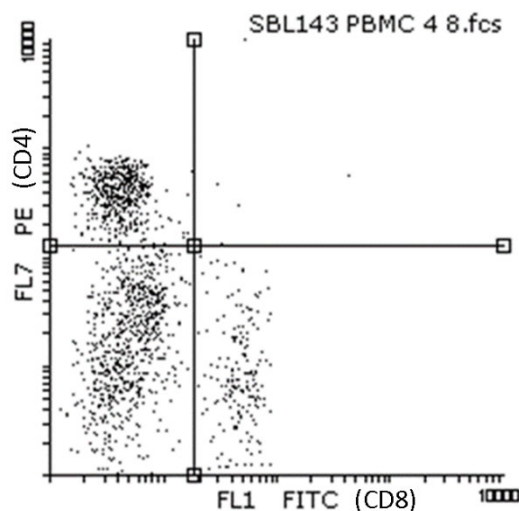


Figure 4.3: Plot of sorted PBMCs stained for CD8-FITC (x-axis) vs CD4-PE (y-axis), from SBL sample number 143. FITC and PE are fluorophores used in cell staining. The gates shown were used to verify the sorting process was successful. Approximately 2.53×10^6 cells in total are shown. The mean purity per sample is shown in the data exploration of **Chapter 5 (Table 5.2)**. Figure courtesy of SBL.

of sonication at 20 second intervals on the Bioruptor fragmented the DNA further into 200 – 400 nucleotide lengths. At this point, the DNA was stored at -80°C while it awaited processing.

Box 4.1 SBL Adapter Design

SBL designed custom “hair-pinned” adapters, which attach to both ends of each fragment. These are not typical, as they prevent PCR amplification. To work around this limitation, a cleavable Uracil base was attached to the end of each strand, just after the 5 thymine bases that loop to close the hairpin. Thus, the closed loop can be opened at will.

To prevent adapter/adapter or fragment/fragment ligation each was left with a single nucleotide overhang – A on the adapter, and T on the fragment. The ends were polished before adapters were added.

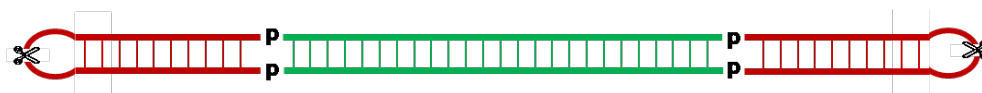
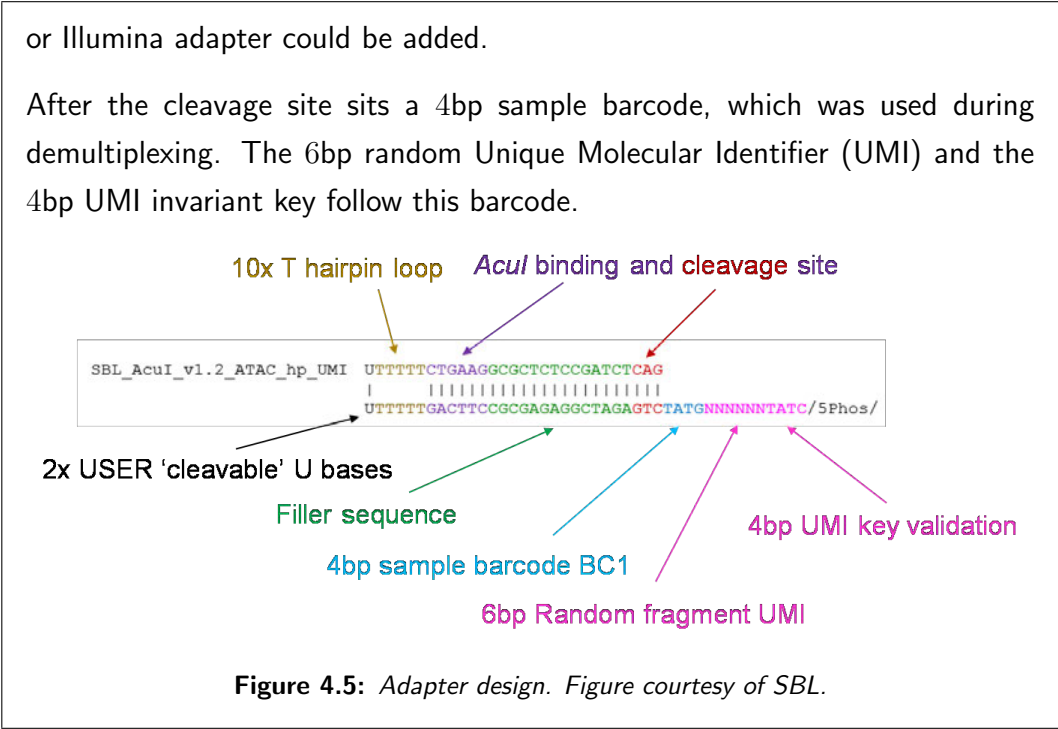


Figure 4.4: Hair-pinned adapter. Figure courtesy of SBL.

To remove the entire adapter, an Acul binding site was added just after the string of T-bases. This allows the adapter to be cut at a specific position – 16 bases away on the top strand, and 14 bases away on the bottom strand. This gave a predictable, “sticky” cleavage site, to which either an Ion Torrent



Unfortunately, this portion of the cell processing protocol was resistant to scaling up. The method required that cells were processed manually in small batches, because automation attempts provided minimal improvements. This could frustrate attempts to increase the sample size number in any future projects.

Once ready, the DNA was removed from the freezer and size-separated using gel electrophoresis by Dr Ben Ferneyhough. Fragments of approximate size 250 – 350 base pairs were selected. The DNA yield was measured using an Invitrogen Quant-IT Picogreen DNA assay kit. An “end-polishing” kit (NEB NextNEB End Repair Module) was used to ensure blunt end DNA fragments. The custom UMI adapter (**Box 4.1**) was then added to the gDNA, which ligated to both ends of each fragment. Solid Phase Reversible Immobilisation (SPRI) bead cleanup was used to size select and “clean” the DNA, in a procedure similar to MACS.

The hair-pinned adapter ends were then opened (**Figure 4.4**) with a NEB USER kit to enable five rounds of PCR with a Kappa Hifi Polymerase 2× Mastermix kit. The gDNA yield was again measured with PicoGreen and size selected on a gel for a second time.

The DNA was removed from the freezer and a SureSelect process used to enrich for the regions of interest. This was performed on batches of blended samples (from more than one donor) for efficiency, with four 300ng samples per batch. SureSelect hybridises a “baits” library to the gDNA samples, to which a solution

of magnetic beads will bind. This allows the selection (with a magnet) of specific regions of the gDNA fragments – here the V and J regions of the T-cell receptor.

The J region was used for selection before the V region for two reasons. First, it is smaller, and hence easier to bait in the first instance. Also, and more importantly, it has been found in mice that when the DNA forms a TREC (see **Figure 4.1**), the V region was sometimes deleted, resulting in J segments with no associated V [244]. To a lesser extent, this has also been observed with V segments [244], but selecting J first ensured that the more common version was not lost if this phenomenon also occurs in humans.

Once the J regions were selected, the fragments were bound to streptavidin beads. These were then washed and added to a multi-well plate with an air “foam” to ensure even coverage across the wells. Although increasing the salt content of the wash increases the exactness of the match between the library and the baits, the washing was done at low stringency at this stage to maximise retention of material. The coverage achieved with this method, by hand, was found to be comparable to the best automated systems available. 15 rounds of PCR amplification were then performed, and the result was again gel purified and size selected.

Once complete, the entire SureSelect process was repeated for the V regions, with two exceptions: the final gel purification and size selection step was skipped, and the washing was performed with high stringency. The resulting material was then ready to have adapters added for either the Ion Torrent S5 or Illumina NovaSeq sequencing platform.

The first step of this was to remove the SBL adapter using the AclI restriction enzyme. The platform-specific adapter was then ligated onto each end. Five more cycles of PCR were run, and the result was gel purified and size selected one final time, before it was ready for sequencing on the appropriate platform. The Ion Torrent S5 was used for test runs (see the ATCC samples of **Section 4.5.1**), but the Illumina NovaSeq platform was used for the final dataset.

SBL estimate the enrichment from these processes to be $\sim 308,000$ -fold by comparison of coding junctions in the final libraries, as compared with the input material.

4.4 Assigning Clonotypes

Many programmes have been developed for the analysis of TCR sequences, with several of the leading tools compared and summarised by Afzal et al. [257]. Only two of the tools they assessed offered options for demultiplexing using UMIs – MIGEC/MiXCR [258][259] and Recover T-cell Receptor (RTCR) [260], which were also their most highly recommended. RTCR showed slightly better accuracy, and MiXCR had more detailed output reports and more flexibility regarding programme parameters [257]. Both tools would likely be suitable, but by citations, MiXCR is the more commonly used, and was chosen for this project. Hence a four-step pipeline based on MIGEC [258] and MiXCR [259] was constructed to demultiplex the reads, align the sequences to genome, and then assign them into clonotypes.

The pipeline was tested with library XCBF649 – sample data provided by SBL from a control purchased from ATCC (American Type Culture Collection, Virginia, USA). Initially, using only MiXCR, it was unable to distinguish duplicates using the unique molecular identifiers (UMIs, see **Box 4.1**). PCR amplification can over-duplicate some sequences in the process of ensuring a sufficient quantity of usable DNA, which then complicates the distinction of true clones (due to clonal expansion) from PCR duplicates. UMIs enable the detection of PCR artefacts, as amplification also copies the unique adapter sequence.

I verified the presence of PCR duplicates in the dataset using multiple alignment and a manual check of the known adapter design (**Figure 4.5**) against the output file. An additional step with MIGEC proved necessary, which has tools to filter these duplicates. MIGEC has two main functions – it will detect and append the UMI sequence to the read header (“Checkout”), and use these to group duplicates into a single instance (“Assemble”).

Unfortunately, results from this approach were unreliable. A MiXCR approach was used with ATCC CD8⁺ test data (see **Figure 4.6**), however it became clear late in the project that some data from the transfer from SBL to Edinburgh was missing or corrupted. Those libraries that did return results showed evidence of inflated clonotypes, potentially indicative of PCR-amplified sequences which had not been filtered correctly by MIGEC.

In parallel, Dr Nick Parkinson, Dr Sam Millington and Sean West at SBL had developed their own custom suite of Java programmes for demultiplexing, alignment and assigning clonotypes. Due to the questionable reliability of the more conventional MIGEC/MiXCR approach in this case, it was decided to move forward

with their version. SBL ran this locally, then forwarded the sequences grouped by clonotype to Edinburgh for diversity estimation and further analyses.

4.5 T-cell Receptor Diversity Estimation

Many metrics are proposed in the literature as appropriate for estimating the diversity of the human T-cell repertoire [238]. Most were adapted from standard measures used in ecology – estimating the relative abundance of animal species in an environment, and were later discovered to be individual instances of a more general measure [261]. Much work has been done in deriving more appropriate ways to analyse T-cells, given the large number of unseen clonotypes [238].

Diversity metrics most commonly used in the T-cell literature are Shannon entropy [262] and Simpson’s diversity [263]. These can be mathematically derived from the same source; the Rényi entropy [261]. Reviews in the mathematical biology literature point out that the Shannon entropy and Simpson’s index are used uncritically, without considering what is actually being measured [264].

To avoid biasing results by choosing a specific diversity metric pre-analysis, this general Rényi entropy is a more appropriate choice. This section defines and describes the analysis, starting from its theoretical basis.

4.5.1 Diversity Measures

Diversity is a general notion of the distribution of members of a population across subcategories. This description of diversity starts from the Rényi entropy, derives the commonly-used Shannon entropy and Simpson’s diversity, and discusses the idea of a diversity profile. The Shannon entropy was defined first, but a more general form of entropy was introduced by Rényi [261]. The Rényi entropy of order α is:

$$H_\alpha = \frac{1}{1 - \alpha} \log \sum_{i=1}^s p_i^\alpha \quad (4.1)$$

where α is the weighting parameter, s is the number of subcategories, $\log$ denotes the natural logarithm, and p_i is the frequency of the i^{th} category:

$$p_i = \frac{x_i}{\sum_i x_i} \quad (4.2)$$

and the input is a vector of counts for a sample x :

$$\mathbf{x} = (x_1, x_2, x_3, \dots, x_s) \quad (4.3)$$

Each entry x_i is a count of the instances of the i^{th} category. This may be modified to account for undersampling in various ways [238]. This is especially relevant for T-cells due to the significant undersampling that occurs when attempting to profile the repertoire. In practice, for example in this project, only on the order of $10^3 - 10^4$ human cells are typically recovered, which is many orders of magnitude fewer than that contained in the whole repertoire ($> 10^{11}$) [265]. As a result, most of the clonotypes remain “unseen”, which is also a typical problem in ecology (many species will not be observed during fieldwork, and population numbers will fluctuate) [266]. As a result, estimators have been developed in the mathematical biology literature. For example, the Horvitz-Thompson correction to the Rényi entropy for a sample of size N [238] is:

$$H_\alpha^N = \frac{1}{1-\alpha} \log \sum_{i=1} \frac{p_i^\alpha}{1 - (1-p_i)^N} \quad (4.4)$$

with p_i and α as before. As the sample size $N \rightarrow \infty$, the correction term $(1 - p_i)^N \rightarrow 0$, and the typical form of the Rényi entropy is recovered. This is based on a similar modification to the Shannon entropy, demonstrated by Chao and Shen [267].

Of special note for both measures are the cases $\alpha = 0$, $\alpha \rightarrow 1$, $\alpha = 2$, and $\alpha \rightarrow \infty$.

Richness ($\alpha = 0$)

The simplest metric available, this is merely the number of distinct categories (or species, s), in the population. It can be recovered from the Rényi entropy at $\alpha = 0$:

$$H_0 = \frac{1}{1-(0)} \log \sum_{i=1}^s p_i^{(0)} = \log(s) \quad (4.5)$$

which is just the log-transformed richness. This corresponds to assigning equal weight to all clonotypes, irrespective of how many instances of each are found.

Shannon Entropy ($\alpha \rightarrow 1$)

The Shannon entropy represents the uncertainty in which category a chosen element belongs – a greater Shannon entropy implies greater uncertainty. It was introduced in 1948 in the context of information theory [262] and is written:

$$H' = - \sum_{i=1}^s p_i \log p_i \quad (4.6)$$

with the p_i as before. This expression is obtained from the Rényi entropy as the limiting case of $\alpha \rightarrow 1$ [268].

Simpson's Diversity ($\alpha = 2$)

Simpson's diversity index [263] (a transformed version of H_2 [238]):

$$\lambda = e^{-H_2} = \sum_{i=1}^s p_i^2 \quad (4.7)$$

is the probability that two randomly drawn (with replacement) sequences are of the same clonotype.

Berger-Parker Index ($\alpha \rightarrow \infty$)

The Berger-Parker index:

$$H_\infty = - \log \max_i p_i \quad (4.8)$$

is the limit of the Rényi entropy as α tends to infinity [238]. This is a measure of the relative size of the most prominent category.

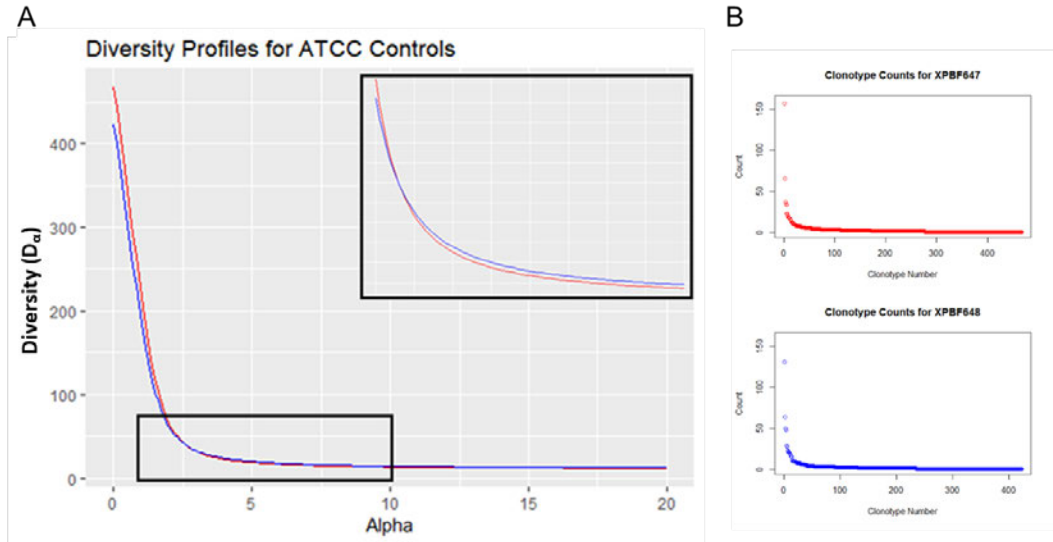


Figure 4.6:
A. Diversity profiles of XPBF647 (red) and XPBF648 (blue), both $CD8^+$ libraries from ATCC healthy controls. XPBF647 had 1,644 recombinants, and XPBF648 1,454. The intersection of the profiles is magnified in the upper right corner of the plot.
B. Corresponding clonotype count plots for the same libraries, with the clone index on the horizontal axis and the total number of that clone found on the vertical. Clonotype number was assigned after sorting by size, so a larger clonotype ID indicates fewer instances of that clonotype found.

Diversity Profiles

The extension of this collection of individual cases is a continuous range of diversity measures encoded by the α parameter. Rather than consider these indices individually, it is better to consider the diversity profile defined by the Rényi entropy over a range of values [269][270]. Increasing α will gradually increase the weight assigned to the most frequently occurring clonotypes. Selecting an appropriate finite range will provide a good approximation of all relevant measures, suitable for computational approaches.

Taking the exponential of the Rényi entropy for an input vector x leads to a diversity profile D_α :

$$D_\alpha = e^{H_\alpha} \quad (4.9)$$

Figure 4.6A shows two diversity profiles across $0 \leq \alpha \leq 20$ for two $CD8^+$ healthy control samples, purchased by SBL from ATCC [271]. Library XPBF647 is shown in red, and XPBF648 in blue. **Figure 4.6B** shows the clonotype counts (x_i) for the recombinants across the clonotypes identified.

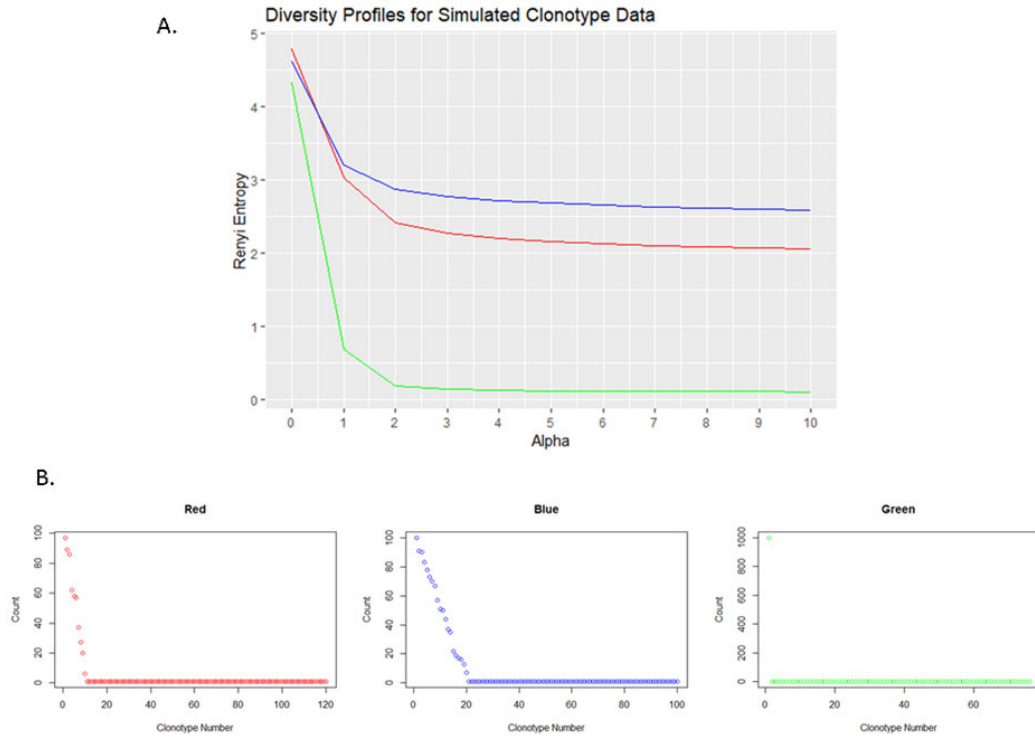


Figure 4.7:

Analysis of simulated clonotype data.

A) Rényi entropies (H_α , Equation 4.1) calculated for three simulated datasets Red, Blue and Green. The Red and Green samples have the same number of simulated sequences, but Green has been artificially constructed to demonstrate that even when the sample size is normalised, profile intersection is not guaranteed.

B) Clonotype distributions for the corresponding profiles in A), ordered from most expanded clonotype to least along the x-axis. Note that the “tails” for Red and Blue are singletons, and the “tail” of Green is singletons and doubles.

Note that it is possible for these diversity profiles to intersect (**Figure 4.6A**, box). The consequence of this is that choosing any one of the individual measures described above may give an entirely opposite result concerning which sample has greater diversity. Indeed, this is extremely common, although I demonstrated with simulations that intersection is not guaranteed even when sample sizes are normalised (**Figure 4.7**).

For comparison of these diversity profiles, it is necessary to define a method to measure a generalised “distance” between them. A typical approach is to measure the Euclidean distance [269]:

$$d_{ab} = \sqrt{\sum_{\alpha} (a_{\alpha} - b_{\alpha})^2} \quad (4.10)$$

between each diversity vector $\mathbf{a}$ and $\mathbf{b}$, each containing a finite number of entries

(note that the α step size can be made arbitrarily small). This measure works analogously to an n -dimensional Pythagoras' theorem by measuring the linear distance between each point a_α and b_α contained within the vectors $\mathbf{a}$ and $\mathbf{b}$, calculating the sum of the squares, and taking the square root. This can then be interpreted as a measure of the similarity of two samples ($d_{ab} = 0$ implies the samples are precisely identical). The Euclidean distance is one of a family of L_k -norms [272]:

$$L_k(x, y) = \sum_{i=1}^d (||x_i - y_i||^k)^{\frac{1}{k}} \quad (4.11)$$

where k defines the measure ($k = 2$ for the Euclidean distance), and d indicates the dimension of the space. Any of these could be used for the distance measurement here. For example, the L_1 -norm (or “Manhattan distance”) would be found by taking the sum of the distances at each α -point. However, due to the “Curse of Dimensionality” in high-dimensional spaces, lower k -measures are preferable [272]. Hence, given many α -values will be used to capture the shape of the profiles, L_2 will be used here.

By applying **Equation 4.10** to each pair of samples in turn, one can generate a distance matrix M :

$$M = \begin{pmatrix} 0 & d_{12} & d_{13} & \dots & d_{1N} \\ d_{21} & 0 & d_{23} & \dots & d_{2N} \\ d_{31} & d_{32} & 0 & \dots & d_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & d_{N3} & \dots & 0 \end{pmatrix} \quad (4.12)$$

Note that $d_{ab} = d_{ba}$ (the matrix is symmetric).

4.5.2 Evenness Measures

Evenness measures are also found in the literature [273]. Evenness is a measure of the distribution of sequences across the clonotypes. Evenness is often measured from 0 (all sequences are of one clonotype) to 1 (exactly equal numbers of each clonotype are found). Thus, a greater evenness metric usually implies counts are more equally distributed across the categories.

Formally, there is a definition relating the concept of evenness to diversity [269]:

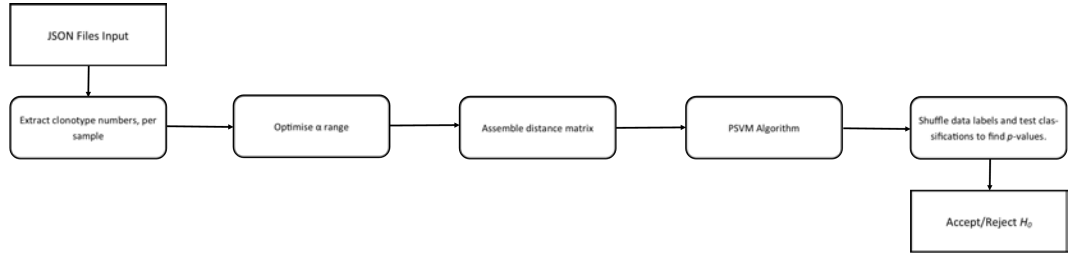


Figure 4.8: Analysis pipeline for the computational implementation of the diversity analysis.

$$E_{\alpha} = H_0 \times D_{\alpha} \quad (4.13)$$

where H_0 is the richness. Written in this form, these quantities are not independent (H_0 is a scalar, and $H_0 \in D_{\alpha}$). In practice, this means a linear transformation of the diversity curve by the number of categories leads to the evenness curve (as noted by Greiff *et al.* [269]), and hence there is little point in plotting both.

There are other interpretations and forms of the evenness. See Jost [274] for a rigorous overview of writing a quantity that expresses the concept of evenness which is genuinely independent of D_{α} .

4.6 Computational Implementation

My aim, based on the method described in Greiff *et al.* [269], was to find the distance between each pair of samples (**Equation 4.10**), and use this as the basis for clustering. For T-cells, the “categories” of **Section 4.5** are the distinct clonotypes. Here, a clonotype is defined to be the CDR3 region, plus the full V, D and J gene segments, without considering pairing between $\alpha - \beta$ or $\gamma - \delta$ chains.

There were four steps:

1. Select a range for α (because the domain of D_{α} is infinite, taking a subset is required), and ensure this provides the most accurate measure for the data;
2. Use this range to calculate a diversity profile for each sample, and calculate the pairwise distances between each of these profiles;
3. Cluster the profiles with a machine learning approach; and,
4. Shuffle the labels, and rerun the clustering to determine whether the original data clusters represent genuine (or else random) associations.

The first two steps require repeated calculation of D_α . This was accomplished using a custom function of the form:

Algorithm 4.1: *Pseudocode for the Diversity function.*

```

INPUT clonotype vector  $v$ .
SET  $\alpha$  vector for range to calculate.
SET  $p$  to the probability vector.
SET  $total$  to  $\sum v$ 
SET  $x$  to empty.
SET  $denom$  to  $1 - (1 - p)^{total}$ .
IF using Horvitz–Thompson correction, adjust the
    equations with  $denom$ .
    BLOCK 1
    FOR each entry in the  $\alpha$  vector.
        IF calculating  $\alpha$  equal to 1, use formula
            for the Shannon entropy, take
            exponential.
            APPEND to  $x$ .
        ELSE use general Renyi entropy, take
            exponential.
            APPEND to  $x$ .
    ELSE RUN BLOCK 1, unadjusted with  $denom$ .
RETURN  $x$ .

```

4.6.1 Selection of the α -range

I was concerned that the Rényi entropy may not be as robust to undersampling as claimed in the literature [269], and that this may have an effect on calculating accurate pairwise distances between samples. Sensitivity to sample size would likely have substantial consequences because the range of recombinants recovered in the CD8⁺ dataset initially varied by two orders of magnitude from smallest to largest (**Figure 4.9**).

Empirical Testing

Consequently, the two ATCC controls of **Figure 4.6** were investigated further. As these data were acquired from healthy controls, it is easily explicable why their

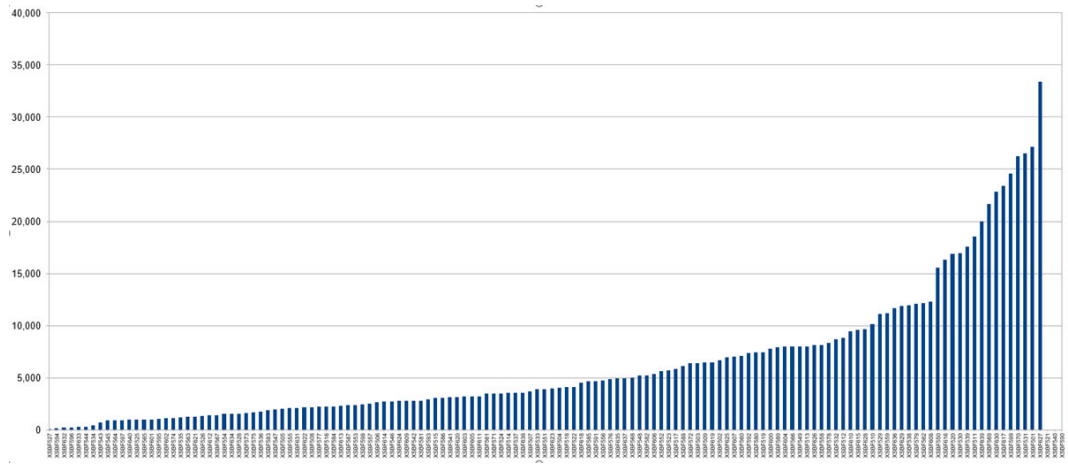


Figure 4.9: *Recombinants recovered from the initial NovaSeq sequencing run. There was a two order of magnitude difference in the number recovered across the range of samples. Figure courtesy of SBL.*

profiles differ significantly only for low α , where the intercept ($\alpha = 0$) is the number of clonotypes found. This is the region where the measure D_α (**Equation 4.9**) is most sensitive to different sample sizes, due to the distributions of T-cell receptor repertoires containing large proportions of singletons. Here, XPBF647 had 1,644 recombinants (with 40.9% of clonotypes being singletons), and XPBF648 had 1,454 recombinants (with 44.4% of clonotypes singletons).

To investigate the effect of sample size on the D_α measure, I bootstrapped these by subsampling at 10% intervals (from 90 – 10% of total recombinants) to provide $2 \times 10 = 20$ total “sub-samples” (including the original samples). We expect these profiles to be similar across all subsamples of each library. However, plotting their diversities (**Figure 4.10**) revealed this measure to be highly sensitive to sample size for low α values (less than $\alpha \approx 5$). So as to consider including these low α values in my analysis to differentiate between samples, it was clear that sample sizes must be made equivalent. This could be done when making pairwise comparisons by downsampling (randomly sampling, with replacement) the sample with the greater number of recombinants to form a new subsample with the same number of recombinants as the smaller of the pair. This approach was further supported by calculating the pairwise Euclidean distances (**Equation 4.10**) between the diversity profiles for each subsample, shown in **Table 4.1**.

When the larger sample is downsampled to approximately the size of the smaller, the distance between their diversity profiles is approximately 131.4; when both samples are maximal in size, the distance between their profiles is approximately 142.7. Note that it is not expected that this distance would be zero, as these

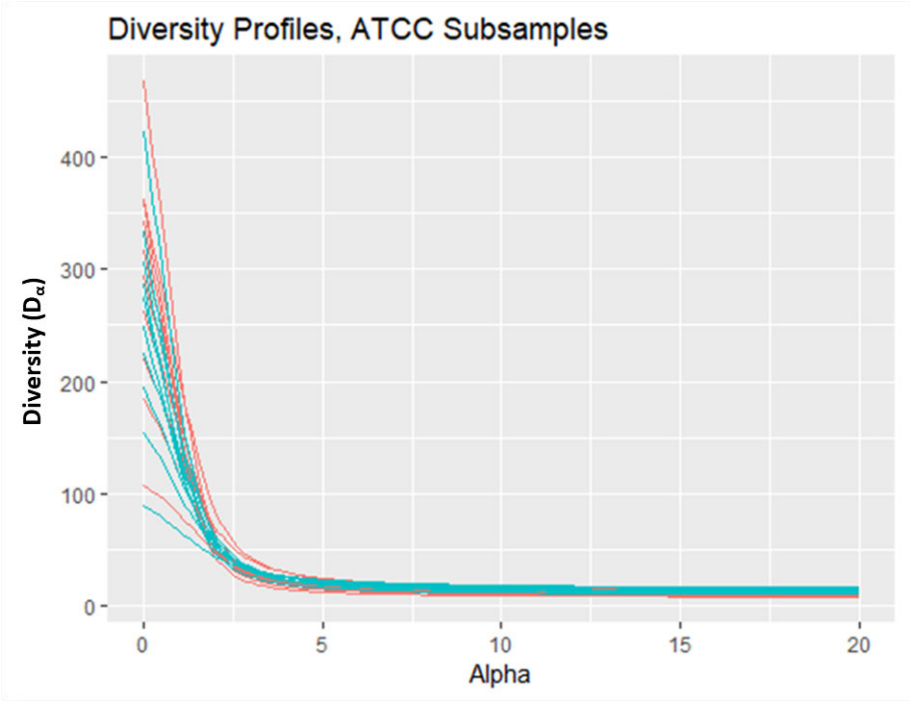


Figure 4.10: Diversity profiles for subsamples of a TCR dataset (taken at 10% size intervals). Red are from library XPBF647, blue are from XPBF648 (as in **Figure 4.6**). Significant variance of D_α is seen for low α values, reflecting that rare clonotypes will drop out as subsamples of decreasing size are taken.

		XPBF648 (1454 Recombinants)										Sample Size
XPBF647 (1644 Rec.)		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
	10%	94.9	130	218	300	327	395	403	482	519	739	164
	20%	276	107	31.7	100	130	198	206	283	319	538	329
	30%	379	201	120	77.3	101	113	169	183	236	442	493
	40%	479	299	195	112	104	66.3	90.7	88.1	131	338	658
	50%	533	360	255	180	154	126	79.3	113	100	294	822
	60%	597	417	315	232	212	147	157	61.5	64.8	217	986
	70%	629	453	349	270	240	186	162	118	66.2	187	1151
	80%	743	563	467	386	370	299	322	213	209	156	1315
	90%	674	495	394	312	285	219	218	132	97.8	131	1480
	100%	938	760	659	576	550	484	479	396	359	143	1644
Sample Size		145	291	436	582	727	872	1018	1163	1309	1454	

Table 4.1: Pairwise distances between the diversity profiles of two sets of subsamples of XPBF647 and XPBF648 (shown in **Figure 4.7**). Sample sizes are shown both as absolute values, and as a percentage of the total, for each subsample. The distances where sample sizes most closely matched for each column are shown in green.

		XPBF647 (1644 Recombinants)										Sample Size
XPBF647 (1644 Rec.)		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
	10%	34.0	126	41.2	12.1	42.7	10.8	5.79	29.8	34.6	5.13	164
	20%	41.2	133	48.4	19.3	49.9	18.0	1.41	37.0	41.8	12.3	329
	30%	37.6	54.0	30.4	59.6	28.9	60.8	77.4	41.8	37.1	66.5	493
	40%	37.3	129	44.5	15.4	46.0	14.1	2.49	33.1	37.9	8.42	658
	50%	91.5	183	98.7	69.5	100	68.3	51.7	87.3	92.0	62.6	822
	60%	25.9	117	33.1	3.91	34.6	2.67	13.9	21.7	26.4	3.03	986
	70%	68.0	160	75.2	46.0	76.7	44.8	28.2	63.8	68.5	39.1	1151
	80%	30.5	61.1	23.3	52.5	21.8	53.7	70.3	34.7	29.9	59.4	1315
	90%	21.4	113	28.6	0.582	30.1	1.81	18.4	17.2	21.9	7.52	1480
	100%	28.9	121	36.1	6.94	37.6	5.71	10.9	24.7	29.4	0	1644
Sample Size		164	329	493	658	822	986	1151	1315	1480	1644	

Table 4.2: Pairwise distances between the diversity profiles ($10 \leq \alpha \leq 50$) of two distinct sets of subsamples of XPBF647 only (shown in **Figure 4.7**, red). Sample sizes are shown both as absolute values, and as a percentage of the total, for each subsample. Matching sample sizes are shown in green.

two samples, although both from healthy controls, are not identical (**Figure 4.6**). When sample sizes are most closely matched (**Table 4.1**, marked in green), distances between profiles (mean: 128.0) were much lower than pairwise distances between subsamples whose sizes differed greatly (mean value of the whole table: 290.3). This provided the justification for downsampling the larger sample to the size of the smaller, for each entry in the distance matrix.

Repeating this approach with two sets of subsamples of XPBF647 gave similar results (**Table 4.2**). No entry was zero (except for the comparison of the original sample with itself), but distances for similarly-sized subsamples tended to be smaller than for very differently-sized subsamples. This approach of comparing sets of subsamples forms the basis of the α -optimisation algorithm.

Another attempted adjustment was to implement the Horvitz-Thompson correction. This accounts for the undersampling of clonotypes by modifying the diversity profile. However, this only modifies the measure for low α -values (**Figure 4.11**). This was verified by taking the difference between the output vector of the Horvitz-Thompson corrected, and unadjusted, D_α code (**Algorithm 4.1**). This difference was found to be zero for almost all of the α -values tested, except those near zero. While expected (the metrics near zero account most for the number of clonotypes found), this led to the conclusion that the Horvitz-Thompson correction was unsuitable for correcting the issues observed with the diversity profiles. Therefore, a new approach was needed, and so I designed an algorithm to restrict the set of measures used in the distance calculations.

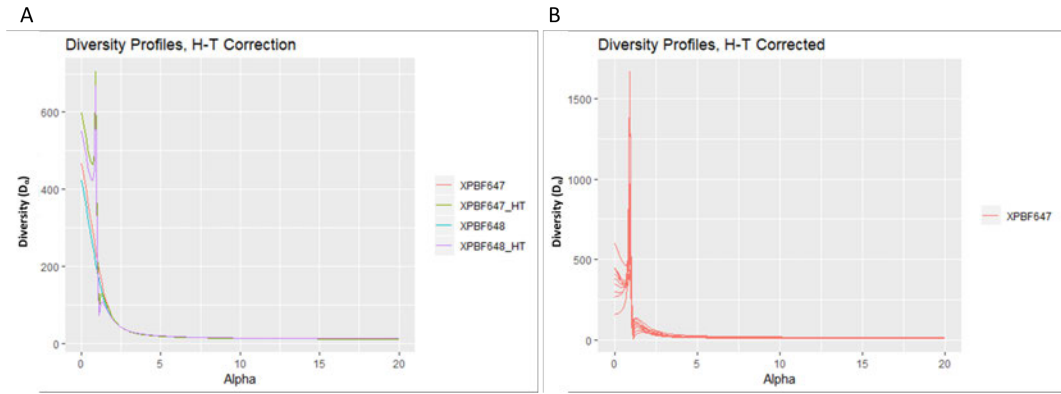


Figure 4.11: Effects of the Horvitz-Thompson correction for:
A) Libraries XPBF647 and XPBF648 (as in **Figure 4.6**).
B) Subsamples of library XPBF647 (as in **Figure 4.10**).

The α -range Algorithm

The code for the α -range algorithm took the form:

Algorithm 4.2: Pseudocode for the α -range algorithm.

```

INPUT clonotype vector  $v$ .
SET  $step\_size$ 
SET  $min\_diff$       #Minimum value of  $|b - a|$ , in loop below.
SET  $start\_max$      #Maximum  $\alpha$  value to test.
SET  $best\_start = 0$ 
SET  $best\_dist = 1000$       #Maximum permitted distance
                           between samples.
SET  $best\_diff = min\_diff$ 
SET  $i = number\ of\ clonotypes$ 
BLOCK 1
#Create first subsample.
SET  $sam =$  subsample with  $0.1 \times$  number of recombinants of
input sample.
SET  $cv$  to zeros, length equal to  $|i|$ .
  FOR  $j$  in 1 to  $|i|$ .
    SET  $j^{th}$  entry of  $cv$  to the number of
    elements  $== j$  in  $sam$ .
FOR  $k$  in 2 to 9.
  SET  $sam =$  subsample with  $k \times 0.1$  number of
  recombinants of input sample.
  FOR  $m$  in 1 to  $|i|$ .

```

```

    SET  $m^{th}$  entry of  $cv$  to the number of
        elements  $== m$  in  $sam$ .
REPEAT BLOCK 1, to create another set of subsamples.
FOR  $a$  in 0 to  $start\_max$ .
    FOR  $b$  in 0 to  $start\_max$ .
        IF  $a < b$  and  $b - a \geq min\_diff$ .
            SET  $D_{100}$  to  $D_\alpha$  of original
                sample.
            SET  $D_1$  to  $D_\alpha$  of the first 9
                subsamples, plus  $D_{100}$ .
            SET  $D_2$  to  $D_\alpha$  of the second 9
                subsamples, plus  $D_{100}$ .
            SET  $distance\_vector$  to a vector of
                pairwise distances between  $D_1$ 
                and  $D_2$ .
            IF  $median(distance\_vector) < best\_dist$ 
                AND  $median(distance\_vector) > 0$ .
                SET  $best\_dist$  to
                     $median(distance\_vector)$ .
                SET  $best\_diff$  to  $|b - a|$ .
                SET  $best\_start$  to  $a$ .
                SET  $best\_max$  to  $b$ .
RETURN a vector of  $best\_start$ ,  $best\_max$ ,  $best\_dist$ .

```

The alpha range to be used in calculation of pairwise distances was optimised using a “zero-point” calibration. This selected a sample i , and drew two sets of 9 subsamples, at 10% size intervals of the original (90%, 80%, etc). The diagonal of the distance matrix between these 20 samples (two sets of nine, plus the original sample in both sets) was calculated, for all α -intervals $[x, y]$, where $x < y$ and with step size 0.5 (to minimise the computational time required). The algorithm chose the range which gave the minimum non-zero median distance between the two sets, and this was repeated for all samples i . These distances should be minimal, as these subsamples are drawn from the same initial sample, and the metric would ideally see them as equivalent (and so have a minimal distance between their diversity profiles), except for a small variance introduced by the sampling (and so not have a pairwise distance of exactly zero).

Visually, this is represented in **Figure 4.12**. These represent the 10 resamples

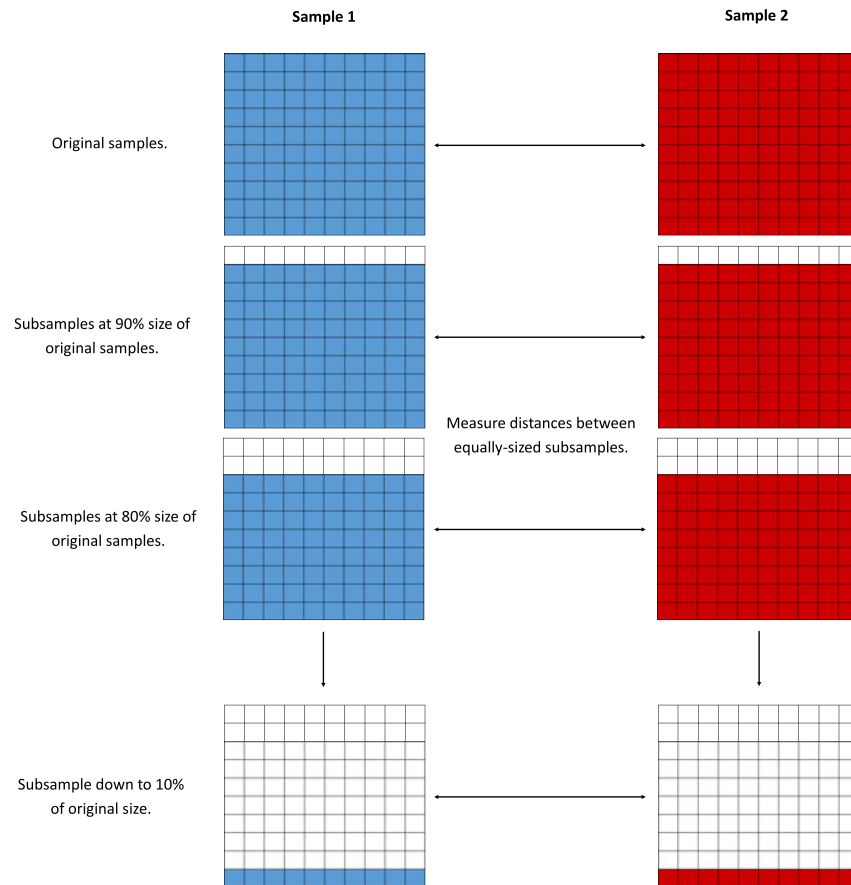


Figure 4.12: Visual representation of the α -optimisation algorithm.

for an individual shown in **Table 4.2**, where the diagonal was calculated (green), and the median of these 10 values taken to be the distance between the sample and itself. This distance will be non-zero (due to the variability introduced by the sampling procedure), but ideally this should be as close to zero as possible. This acts as a calibration for the measure, and the entire procedure was repeated for different intervals of α . The range $[x, y]$ which minimised the median of these distances was taken to be the best estimate of the optimal α range.

This step was performed before unblinding, to check for any anomalies in the results. In particular, since testing revealed that the distribution of intervals was sometimes bimodal, the concern was that the algorithm would choose a highly optimal α -range for half of the samples, which would prove highly suboptimal for the other half. Testing this with the CD8⁺ data confirmed that this was not an issue – the mean of the lower bound a was 27.1, and the mean of the upper bound b was 47.1, indicating that relatively few samples deviated from the chosen values ($a = 30$, $b = 50$, based on these results).

It also proved impossible to also optimise the step size with the chosen method, as the algorithm attempts to use the largest step possible (removing the extra information between the end points seems to give a closer fit). Since this approach ignores large amounts of information, I chose instead to use a step size of 0.1, based upon the value (0.2) used by Greiff *et al.* (selecting the same order of magnitude, but with twice the resolution).

4.6.2 Generation of the Distance Matrix

The aim was to construct a distance matrix that captured the actual distance between each pair of samples, irrespective of the degree of undersampling. The distance matrix is an $N \times N$ symmetric matrix, with a zero diagonal. Each entry represents the distance between the diversity profiles of the corresponding samples.

This was implemented using code with the following form:

Algorithm 4.3: *Pseudocode for generation of the distance matrix.*

```

INPUT set of clonotype vectors  $V$ .
SET  $dist\_mat$  to  $N \times N$  matrix of zeros.
SET  $n$  = user defined resample number.
FOR each row  $i$  in the matrix.
    FOR each column  $j$  in the matrix.
        BLOCK 1
        IF  $i > j$ , and  $|i| \geq |j|$ .
            BLOCK 2
            SET  $sam$  to a random subsample (
                with replacement) of sample  $i$ ,
                with  $|j|$  elements.
            SET  $cv$  to zeros, length equal to
                 $|i|$ .
            FOR  $k$  in 1 to  $|i|$ .
                SET  $k^{th}$  entry of  $cv$  to
                    the number of
                    elements ==  $k$  in  $sam$ .
            SET  $downsampled\_counts$  to  $cv$ .
            FOR  $l$  in 1 to  $n - 1$ .
                Repeat BLOCK 2 to
                    generate a total of  $n$ 

```

```

        samples .
    APPEND results to
        downsampled_counts .
    SET ds_da to vector of  $n$  zeros .
    FOR  $o$  in 1 to  $n$  .
        APPEND distance between
             $D_{\alpha}(\text{downsampled\_counts}(o))$ 
            and  $D_{\alpha}(\text{downsampled\_counts}(j))$  to dist_mat
            entries  $(i, j)$  and  $(j, i)$  .
    ELSE IF  $i > j$ 
        REPEAT BLOCK 1, with  $i$  and  $j$ 
            interchanged .
    RETURN dist_mat , an  $N \times N$  symmetric distance matrix with
        0 diagonal .

```

Similarly as for the α -optimisation algorithm (**Section 4.6.1**), to mitigate the effects of undersampling, a resampling procedure was included. Two samples a and b were taken, where sample a contains more recombinants than sample b (denoted by $|a| > |b|$). To calculate the entry d_{ab} , sample a was downsampled to $|b|$ recombinants by drawing n random subsamples from sample a repeatedly, calculating the distance (**Equation 4.10**) between sample b and each subsample of a , and taking the median of these to find d_{ab} and d_{ba} . This process was then repeated separately for all a and b in the matrix, downsampling the larger to the smaller in each case. Where sample sizes were equal, one (sample a , where $a > b$) was resampled to the same size, n times. The resample number n was determined empirically using the ATCC test data (of **Figure 4.6**), by calculating the distance matrix for the two samples across a range of values. This testing demonstrated that the distance between the samples converged to a fixed value for a number of resamples $n \approx 1000$ (**Figure 4.13**).

4.6.3 Multidimensional Scaling

For ease of interpretation of the results, a visual representation of the samples and distances between them is useful. One method is to produce a multidimensional scaling (MDS) plot. MDS is a class of analysis, where the axes explain the greatest variation in the distance matrix [275], but the resulting plots provide an intuitive

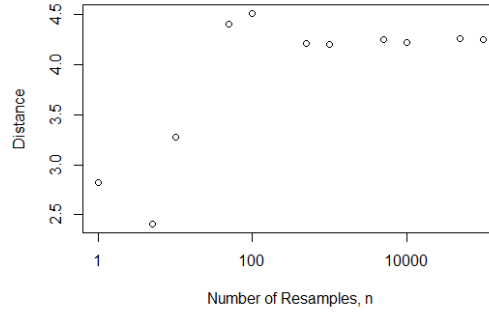


Figure 4.13: Pairwise distance between the ATCC test samples (y-axis) as a function of resample number, n , on a log scale. This was used to indicate the number of resamples needed for accurate calculation of the distance matrix $\mathbf{M}$.

understanding of the relative distances between samples.

Intuitively, MDS plots are lower-dimensional representations of the data [276]. For example, place sample 1 at the origin of a two- or three-dimensional space. Now draw a circle (or sphere, in 3D) of radius d_{12} from the matrix $\mathbf{M}$ (**Equation 4.12**). Sample 2 can be placed arbitrarily on the circumference (or surface). Repeating this process with d_{13} and d_{23} for samples 1 and 2, respectively will lead to intersections between the circles/spheres. Sample 3 may be placed at any of these intersections. Repeating for all N samples in the dataset will lead to a projection of the distances in the matrix $\mathbf{M}$. Note that this is unlikely to work perfectly, and more dimensions may be required to preserve pairwise distances (for example, one may need to “elevate” a sample into the third dimension, even if all preceding samples fit exactly in 2D). Hence, MDS is typically only a projection into the specified low-dimensional space, and is only defined up to rotations [276].

More formally, the MDS algorithms use eigendecomposition to form a coordinate matrix $\mathbf{X}$ from a distance matrix $\mathbf{M}$, choosing the best fit by minimising a stress function [276]:

$$\sigma_r(\hat{\mathbf{d}}, \mathbf{X}) = \sum_{i < j} w_{ij} (d_{ij}(\mathbf{X}) - \hat{d}_{ij})^2 \quad (4.14)$$

where d_{ij} is the distance between points i and j in the distance matrix $\mathbf{M}$, $\hat{d}_{ij}$ is the transformed distance in the lower-dimensional MDS space, and the w_{ij} are non-negative weights. In metric MDS, as used in **Chapter 5**, the $\hat{d}_{ij}$ are continuous transformations (for example, for ratio MDS, $\hat{d}_{ij} = b d_{ij}$ are the distances of the matrix $\mathbf{M}$, scaled by a factor b). A thorough discussion of the matrix algebra and

algorithms involved in MDS is given by Borg *et al.* [276].

4.6.4 Classification (Potential Support Vector Machine)

A classifier was required to separate the cases and controls into groups, and several machine learning (ML) algorithms are frequently used for this purpose. These algorithms can be supervised or unsupervised, which refers to whether the ML algorithm has access to the data labels, or not (respectively) [187]. Supervised algorithms are more appropriate here, as the subgroup data labels are available (MEsa, MEmm, MS, HC) for all samples. Algorithms can be further divided into parametric (which make assumptions on the distribution), and non-parametric (which do not) [187]. Here, given the unknown distribution, a non-parametric method is more appropriate.

Commonly used non-parametric supervised algorithms include convolutional neural networks (CNNs), decision trees (for example, random forests), k -nearest neighbours (k -NN), and support vector machines (SVMs) [187]. Given that the measures of dissimilarity between samples are the distances between diversity profiles, a distance-based algorithm would be best for these data. Hence, either k -nearest neighbours or a SVM would have been an acceptable choice. However, k -NN is prone to overfitting (fitting too closely to the training data, and hence performing worse on test data) [187], is more computationally intensive, and the boundary prediction may be unstable for small values of k (increasing k increases the linearity of the boundary) [277]. A SVM draws a classification boundary between two labelled sets of data, and will accept a kernel function to allow non-linear boundaries [187][277][278]. It achieves this by projecting the data into a higher dimension, drawing a linear boundary, and then transforming back to the original space [278]. Hence, although prone to the same overfitting concerns as k -NN [187], a SVM is a good choice for these data. An additional improvement is to use the Potential Support Vector Machine (PSVM) of Hochreiter *et al.* [279], a modified SVM which was formulated to work on pairwise relationship data (such as distance matrices) [279]. This has been shown to typically outperform similar SVM approaches [279], and has the added advantage of having been tested previously on T-cell clonotype counts data [269]. Hence, a PSVM was used to attempt to accurately classify the repertoires into the four groups (severe ME/CFS, mild/moderate ME/CFS, multiple sclerosis, and healthy controls).

The PSVM software used allows the use of leave-one-out cross-validation (rather

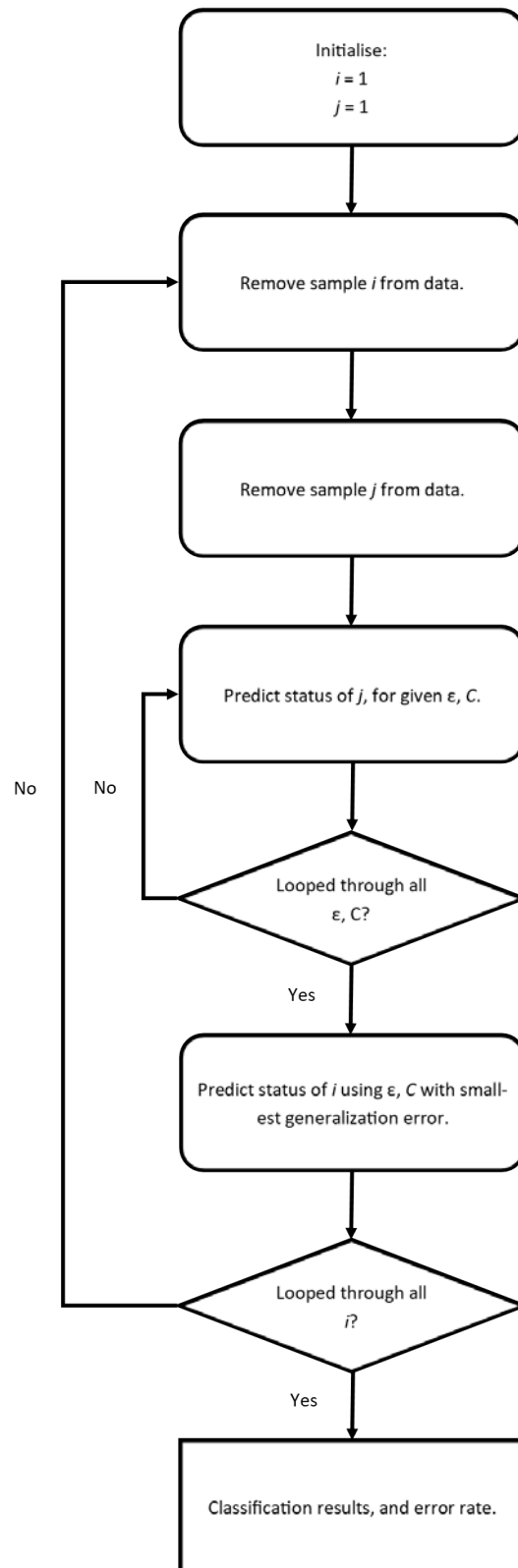


Figure 4.14: Flowchart showing the PSVM algorithm.

than training and testing datasets), with optimised hyperparameters [280]. For this, a single data point was removed, the model was trained on the remaining data, and then the model predicted the status of the removed point (**Figure 4.14**). This was repeated for each data point in turn (**Figure 4.14**, “outer” loop). The algorithm also used another leave-one-out loop (**Figure 4.14**, “inner” loop) to select the optimal model hyperparameters, ϵ and C . The ϵ parameter is the level of tolerance for errors in the training data – if it is too high the model will fail to capture nuance in the training set, and if it is too small it will overfit. The C parameter controls the maximum weighting of the support vectors. If this is not set ($C = \infty$), then there will be few support vectors, which leaves the model more prone to outliers. Setting C too low will average across too many vectors, and provide a less accurate classification boundary [280]. Optimisation of these hyperparameters was achieved by minimising the generalization error [281], here the mean squared error (MSE) [280]:

$$MSE = \frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2 \quad (4.15)$$

where $x_i \in \{0, 1\}$ is the case/control status of the i^{th} sample, $y_i \in \{0, 1\}$ is the predicted case/control status of the i^{th} sample, and n is the total number of samples [282].

This procedure should lead to robust classification boundaries between the four classes for each sample in the dataset. The PSVM handles multiple classes iteratively, finding one classification boundary at a time. To verify the results, significance testing was applied. In the context of machine learning, this is permutation testing, performed by randomly shuffling the data labels and training the machine on these datasets. This generates a distribution of accuracy scores, within which the original model with correct labels should be significant. The accuracy score A can be summarised as:

$$A = T_P + T_N \quad (4.16)$$

where T_P is the number of true positives, and T_N is the number of true negatives. Any permutation with a greater accuracy score than the original dataset is noted, and the total number of these “more accurate” permutations, A_{Total} , is used for the level of significance. For example, for 100 permutations, $A_{Total} < 5$ is significant at $p < 0.05$. Ideally, this permutation testing should be performed on the set of

all possible permutations, but as this number is large here ($\approx 4^{40} \times 3^{40} \times 2^{40} \approx 1.62 \times 10^{55}$), a subset was more practical [283].

An advantage of this implementation of the PSVM is that the hyperparameters are set per individual permutation. This dramatically increases the computation time, but is more robust to false positives, as using the hyperparameters from the initial data for the permutations will lead to a better fit on the actual data than on the permutations. A trial run by fixing the initial parameters for an optimal fit to the real data immediately generated a significant result after fitting the same model to the permutations, underlining the importance of reassessing the hyperparameters for each permuted dataset individually.

To provide further reassurance that this method was robust, the entire analysis pipeline was tested on simulated data.

4.6.5 Validation of the Analysis via Simulation

To demonstrate the validity of this approach, simulations were run to establish that the PSVM could determine the difference between clonotype repertoires that showed signs of expansion, and those that did not. 160 samples were generated, to match the number of samples available from the CureME biobank, each with 10,000 recombinants (**Figure 4.15**) (comparable with the average of 8,547 combined α , β , and γ chains from the CD8⁺ dataset). There were:

- 80 “severe” cases, 40 drawn from a power law distribution with exponent -1.5 (red), and 40 from a distribution with exponent -1.4 (yellow). These exponents were chosen by trial-and-error to give distributions that resembled example clonotype distributions, with a more extreme peak (compare with **Figure 4.6B**).
- 40 “mild/moderate” cases, drawn from a power law distribution with exponent -0.2 (green).
- 40 “control” samples drawn from a uniform distribution (blue).

Analysis was performed as described above, by optimisation of the α -range, generation of the distance matrix, and input into the PSVM. The differing distributions reflect the range of possibilities for an immune repertoire, which are typically represented using power law models (Zipf distributions in Grieff *et al.* [269]), but could also be more uniform in nature. This ensured that the distance matrix and PSVM

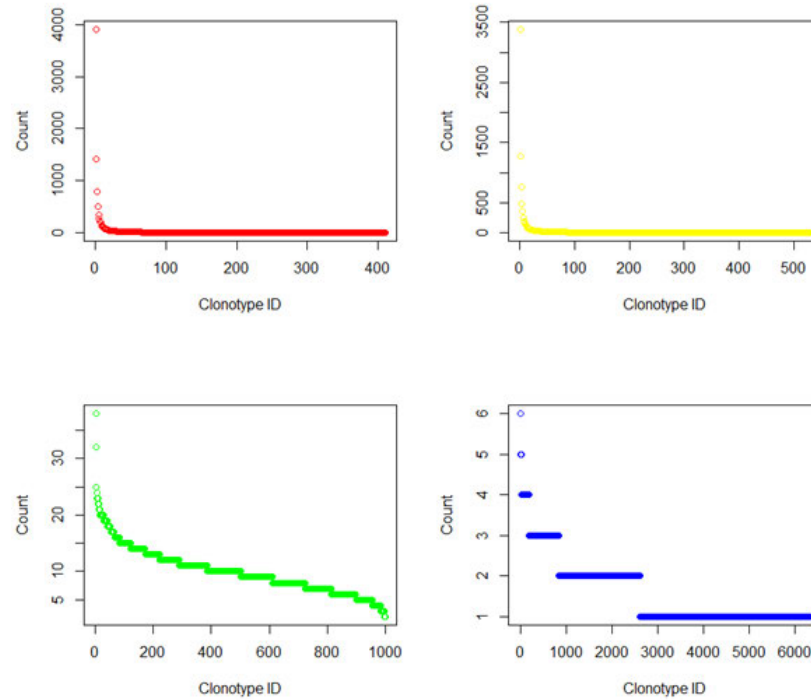


Figure 4.15: Example clonotype distributions for simulated severe ME/CFS cases (red, yellow), mild/moderate ME/CFS cases (green), and healthy controls (blue).

classification method was robust, regardless of the structure of the measured immune repertoires. In other words, it was possible in practice that the healthy controls would more closely resemble the “yellow” category, and the analysis protocol would need the precision to distinguish four groups clustered between the “red” and “yellow” groups (four groups of power law distributions, with exponents in the range $[-1.4, -1.5]$).

The α -range distribution was highly bimodal. The algorithm generally chose a 0 – 20 range for the less expanded samples (mild/moderate cases and healthy controls), and a 30 – 50 range for the (red) severe cases. This was surprising, but is explicable by the qualitative explanation of the α -range. For less uniform distributions, the algorithm minimised the distance between subsamples by emphasising the largest clonotypes. For the more uniform distributions, it did the same by favouring lower α -values, which correspond to the number of clonotypes. This was investigated as an alternative method of classification, but did not prove to be any more accurate than chance ($\sim 50\%$ probability of assignment into one of two categories).

The distance matrix was calculated with $n = 100$ resamples, and a step size of 1 (to minimise the computational time required). An α -range of 30 – 50 was chosen

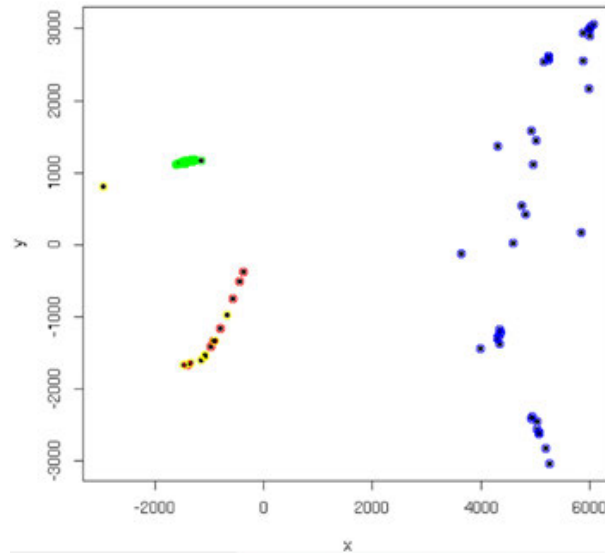


Figure 4.16: *Multidimensional scaling plot (projection) of the distance matrix generated from the simulated data shown in Figure 4.15. Calculated using non-optimal parameters of $n = 100$, $\alpha \in [30, 50]$ and step size 1.*

to work around the bimodality issues described with the optimisation step. Even testing with these approximations showed striking results, as seen in **Figure 4.16**. The simulated “severe cases” (red and yellow) visibly separate from the “mild cases” (green) and controls (blue), demonstrating that this method of calculation of pairwise distances was viable for at least first-order classification.

To further demonstrate the accuracy of this method, the “red” and “yellow” samples were investigated in more detail. With non-optimised parameters, perfect separation was still observed in the distance matrix (**Figure 4.17B**), for diversity profiles that are almost completely indistinguishable by eye (**Figure 4.17A**). This was evidence that the method would remain robust, even if **Algorithm 4.2** selected a sub-optimal α -range, or if it was not possible to complete the full number of resamples (n).

The PSVM algorithm achieved a perfect result for all classifications tested, and none of the ≥ 100 permutation shuffled-label datasets used in significance testing gave a better result.

4.7 Hypotheses

In order, my three primary hypotheses were that there is a significant T-cell activation (expansion of clonotypes inferred from TCR sequencing) difference between:

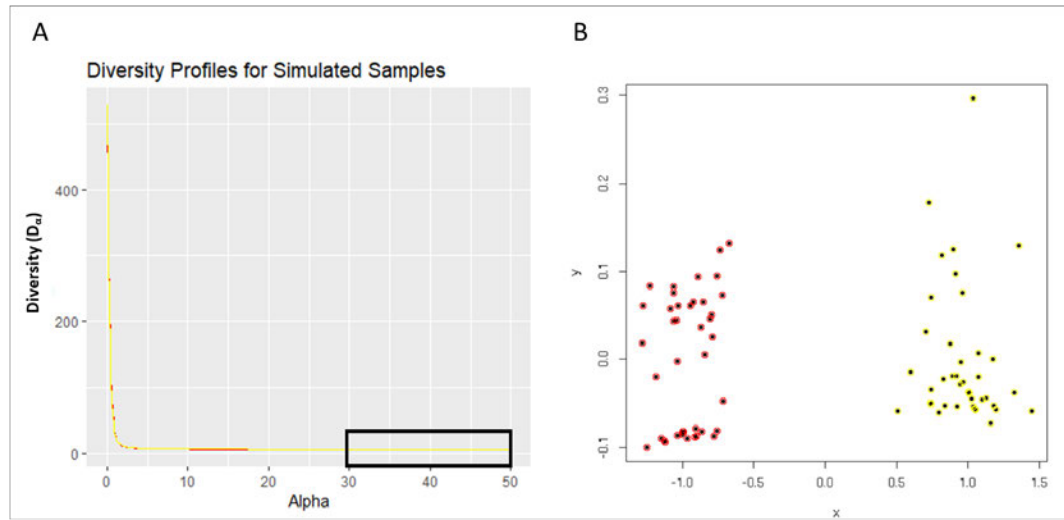


Figure 4.17: Multidimensional scaling plot (projection) of the distance matrix generated from the “red” and “yellow” simulated data shown in **Figure 4.15**. Calculated using non-optimal parameters of $n = 100$, $\alpha \in [30, 50]$ and step size 1.

A. Diversity profiles for samples from the red and yellow groups from **Figure 4.16**, with the region used in the distance matrix calculation highlighted.

B. Multidimensional scaling plot (projection) of their distance matrix.

- Healthy controls and ME/CFS cases;
- Multiple sclerosis controls and ME/CFS cases;
- ME/CFS (mild/moderate) cases and ME/CFS (severely affected) cases.

These were controlled with a Bonferroni-corrected significance threshold of $\frac{0.05}{3} \approx 0.017$. The secondary hypotheses were that there is a significant activation difference between:

- Healthy controls and cases (ME and MS);
- Healthy controls and MS cases;
- Healthy controls and ME (mild/moderate) cases;
- Healthy controls and ME (severe) cases;
- MS cases and ME (mild/moderate) cases;
- MS cases and ME (severe) cases.

These hypotheses are marked in **Table 4.3**. The secondary hypotheses were to be considered only “indicative” if they met the full Bonferroni-corrected p -threshold of $\frac{0.05}{9} \approx 0.006$. Note that the people with MS are considered to be positive controls – it was expected that they would display some degree of clonal expansion, based on the work of Dr Davis (see **Figure 2.3**). The measurements described in **Section**

	HC	Case	MS	ME	ME(mm)	ME(sa)
HC		Secondary	Secondary	Primary	Secondary	Secondary
Case				Primary	Secondary	Secondary
MS						Primary
ME						
ME(mm)						
ME(sa)						

Table 4.3: Decision boundaries calculated. “ME” indicates combined ME(mm) and ME(sa), “Case” indicates disease case (MS and ME combined)

4.5 are merely considering the degree of expansion.

4.7.1 Parameters

Before unblinding, certain parameters were set in advance.

- Primary hypotheses were set, as shown in **Table 4.3**.
- Use of the Horvitz-Thompson correction was decided against. The correction only adjusts the value of D_α for low α -values. Since the algorithm preferred larger values, and the correction introduces an awkward singularity at $\alpha = 1$, I considered it unnecessary.
- The resample number n was set to 1,000, based upon the testing shown in **Figure 4.13**.
- Based upon observations of the algorithm of **Algorithm 4.2**, the step size was set manually. This algorithm preferred the largest step size, which may have been less accurate when not applied to the subsamples. Hence, based on the literature [269], I chose a step size of 0.1 for calculation of the distance matrix.
- The number of label shuffles to determine significance was intended to be 10,000. This was based on the computation time required by the simulations of **Section 4.6.5**. However, this was revised down to 1,000 after computation times proved unrealistic (approximately six months) with the actual dataset.
- The first two steps (α -optimisation and calculation of the distance matrix) were completed before unblinding, to allow for any adjustments required without biasing the results.

Unblinding occurred after calculation of the combined chains CD8⁺ distance matrix.

The following chapter (**Chapter 5**) shall present the results from this project.

Chapter 5

T-cell Receptor Repertoire Diversity: Results

This chapter contains results obtained from the analyses of the T-cell receptor repertoire study described in **Chapter 4**.

5.1 Data Summary

After unblinding, the data were tested for any systematic biases unrelated to the diversity analyses. A two-way table of the counts of samples in each age bracket and subgroup was generated (**Table 5.1**), which showed a marked bias for severe ME/CFS in younger age groups. A Kruskal-Wallis test for the comparison of categorical and ordinal data [284] (*kruskal.test()* in *R*) showed that this significantly differed from the null of equivalent distributions for at least one group ($p = 0.00177$). The effects of age on these analyses will be discussed further in **Section 5.3**.

Further, **Table 5.2** shows the means of the key experimental parameters, per subgroup, for both the CD4⁺ and CD8⁺ data.

Subgroup	$\geq 18 \ \& \ \leq 29$	$\geq 30 \ \& \ \leq 39$	$\geq 40 \ \& \ \leq 49$	$\geq 50 \ \& \ \leq 60$
ME/CFS(sa)	8	9	6	17
ME/CFS(mm)	0	0	20	20
Multiple Sclerosis	0	0	12	28
Healthy Controls	0	1	16	23

Table 5.1: A two-way table of sample counts by disease subgroup and age bracket. Severely affected people with ME/CFS are significantly overrepresented in younger age groups.

Subgroup	PBMC Cells ($\times 10^6$)	CD8 ⁺ Cells ($\times 10^6$)	CD8 ⁺ Purity (%)	CD4 ⁺ Cells ($\times 10^6$)	CD4 ⁺ Purity (%)
ME/CFS(sa)	5.27	0.632	70.7	0.665	72.0
ME/CFS(mm)	4.50	0.541	70.3	0.703	76.0
Multiple Sclerosis	4.62	0.473	68.0	0.652	73.5
Healthy Controls	4.72	0.585	69.8	0.698	73.6
All Samples	4.78	0.558	69.7	0.679	73.7

Table 5.2: Means of key experimental parameters, shown for all samples, and by disease subgroup. Comparison by ANOVA (`aov()` in R) showed only one significant association – for CD8⁺ cell number ($p = 0.0257$).

Analysis of variance (ANOVA) [284] for the experimental parameters showed that the only association with subgroup was CD8⁺ cell number ($p = 0.0257$). Inspection of the box plots showed that the MS samples had slightly fewer CD8⁺ cells. However, even allowing for only two independent tests (correction to $p < 0.025$) shows that this would not remain significant. Hence, even without considering the dependence of the variables, it is clear that this association is not of concern for the following analyses.

5.2 Post-unblinding Analysis Adjustments

In **Section 4.7**, multiple testing adjustments to the significance threshold were set, using the Bonferroni multiple testing correction [285] $p < \frac{0.05}{3} \approx 0.017$ for the primary hypotheses and $p < \frac{0.05}{9} \approx 0.006$ for the secondary hypotheses. These will be referred to as the pre-analysis corrected p -thresholds.

However, these were based on the corrections needed for a combined analysis of the CD8⁺ data only. After unblinding, I decided to investigate each set of α -, β - and γ -chain sequences separately, which trebled the number of hypotheses to be tested. Furthermore, with the CD4⁺ data from the same samples becoming available, the number of hypotheses to be tested again doubled. This increased the total number of hypotheses from 9 to $9 \times 32 = 54$. Recalculating the corrected p -thresholds in the same way led to the new adjusted thresholds of $p = \frac{0.05}{18} \approx 0.0028$ (primary hypotheses) and $p = \frac{0.05}{54} \approx 0.0009$ (secondary hypotheses). These will henceforth be referred to as the post-analysis adjusted p -thresholds.

The pre-analysis p -thresholds were overly liberal given the new number of hypotheses, and hence would have been prone to generating false positive results. In addition, an adjustment to the analysis plan was necessary in order to compute p -values within a reasonable time frame. The planned 10,000 label permutations (**Section 4.7.1**) were reduced to 1,000, because initial runs timed out in the 300 – 400 shuffle range. Given the multiple testing corrections, for some hy-

Subgroup	Final Number
ME/CFS(sa)	38
ME/CFS(mm)	38
Multiple Sclerosis	39
Healthy Controls	39

Table 5.3: *Number of $CD8^+$ samples, by subgroup, available for analysis.*

potheses tested this number of permutations was sufficient. This was because it generated sufficiently many decision boundaries that were an improvement upon the one calculated for the real data that even had permutation testing continued to 10,000 shuffles without adding to their number, the hypothesis would not have been significant at a level of $p < 0.05$ after Bonferroni multiple testing correction. Even so, all hypotheses were rerun with an increased time limit until 1,000 permutations were reached (with one set of exceptions, noted below).

Post-analysis multiple testing correction is much stricter than the pre-analysis version, hence comparisons with both are included in **Section 5.6**, **Table 5.13**. In particular, although calculated in the usual manner, the post-analysis correction may be unrealistic (as $p < 0.0009$ implies rejecting a hypothesis if even one among 1,000 permutations is superior to the real data). Hence, for comparison purposes, I also report the threshold used by Greiff *et al.* ($p < 0.01$, on 1,000 shuffles) [269].

Hypotheses with $p < 0.05$, before multiple testing correction, are marked with an asterisk (*) in the following tables of results, and summarised in **Section 5.6**, **Table 5.12**. Those hypotheses that were not tested with 1,000 permutations are marked with a dagger (†) in the results tables below.

I started by analysing recombinants from $CD8^+$ cells (**Section 5.3**) and then repeated with the data from $CD4^+$ cells (**Section 5.4**).

5.3 $CD8^+$ Cells

For $CD8^+$ cells, from the initial 160 individuals' samples, 154 were successfully processed and sequenced (**Figure 5.1**). One sample had insufficient DNA, and another five libraries had been cross-contaminated (material from one library was found in another). These were distributed relatively evenly across the four groups (**Table 5.3**).

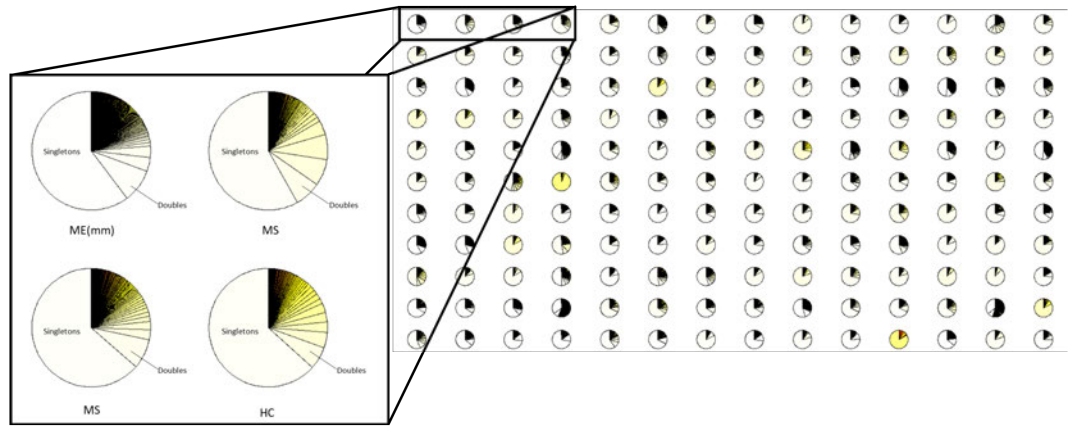


Figure 5.1: Repertoires of the 154 $CD8^+$ samples. The first four are shown in greater detail, with subgroup labels. Colours are assigned to each clonotype in a sample, starting with red and gradually desaturating to white. More monochrome plots therefore indicate samples with a greater richness.

5.3.1 Combined Analysis (α -, β -, and γ -chains)

The distance matrix for all three $CD8^+$ chain types combined was visualised using a multidimensional scaling (MDS) plot with `cmdscale()` in *R*, shown in **Figure 5.2**. This plot showed no clear visual separation between the four groups, although this MDS plot is only a two-dimensional representation of the matrix, and so clearer separation might have been apparent if this were extended to include a third dimension. However, two dimensional projections of the three-dimensional case of **Figure 5.2** were also not visibly separable.

There was concern that the outliers for $x > 2,000$ may have indicated that some dependence on the recombinant numbers recovered in the lab remained in the distance matrix (**Equation 4.12**). Investigation of the axes using a generalized linear model (with `glm()` in *R*) revealed correlations for each, shown in **Tables 5.4** and **5.5**.

The x-axis appeared to have correlations as a result of differences in age among the sample donors. The CureME biobank only provides ~ 10 year age brackets, so it is likely that the effect is much stronger than the model p -value ($p = 0.00199$) suggests. It is well established that age has a pronounced effect on immune repertoires, and so this is an unsurprising correlation [287]. It might be expected that there would be fewer recombinants from those in the older age categories, as reduced thymic activity (thymopoiesis) with age is well documented in the literature [288]. As stated in **Section 4.1.2**, TRECs can provide a measure of thymopoiesis, and hence this correlation is supported by modelling the total $CD8^+$ TREC counts per sample against age group with `glm()` in *R*, and noting that age group is a

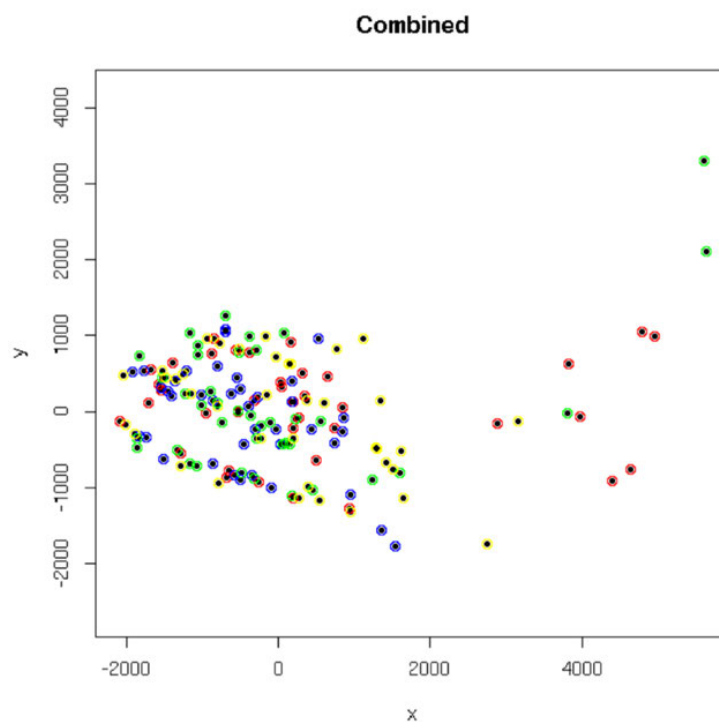


Figure 5.2: Multidimensional scaling plot of all $CD8^+$ chains. $ME(sa)$ are in red, $ME(mm)$ are in yellow, MS are in green and HC are in blue. Axes indicate projected distances, such that the linear distance between two samples reflects the value in the distance matrix. There is no clear visual separation between the groups in two dimensions, but there are notable disease case ($ME(sa)$, $ME(mm)$ and MS) outliers for $x > 2,000$.

Variable	Estimate	Std. Error	t-value	$Pr(> t)$	Sig.
(Intercept)	1.25E+11	2.22E+11	0.564	0.574	
ME/CFS (mm)	1.89E+02	3.20E+02	0.590	0.556	
ME/CFS (sa)	4.51E+01	3.68E+02	0.123	0.903	
MS	4.73E+01	3.23E+02	0.147	0.884	
Age	-1.36E+03	4.29E+02	-3.16	0.00199	**
PBMC Number	1.35E+02	1.02E+02	1.32	0.190	
CD8 ⁺ % Input	-4.54E+01	2.87E+01	-1.58	0.116	
CD8 ⁺ Cells Post-selection	-7.90E+02	7.85E+02	-1.01	0.316	
Purity	6.47E+00	8.34E+00	0.776	0.439	
DNA Extracted (μ g)	-9.18E+01	1.51E+02	-0.608	0.544	
DNA Yield (ng)	6.11E-03	3.90E-01	0.016	0.988	
Demultiplexed Reads (641)	1.23E-04	5.47E-05	2.25	0.0263	*
PEAR Assembled (641)	1.19E+01	1.69E+01	0.701	0.485	
Demultiplexed Reads (782)	-1.56E-05	4.27E-05	-0.366	0.715	
PEAR Assembled (782)	-1.39E+01	1.87E+01	-0.741	0.460	
α Coding Junctions	5.74E-01	2.38E-01	2.41	0.0174	*
β Coding Junctions	-4.18E-01	3.32E-01	-1.26	0.211	
γ Coding Junctions	-8.88E-01	3.07E-01	-2.90	0.00452	**
α %	-1.25E+11	2.22E+11	-0.564	0.574	
β %	-1.25E+11	2.22E+11	-0.564	0.574	
γ %	-1.25E+11	2.22E+11	-0.564	0.574	

Table 5.4: Significance values for the variables tested as potential confounders with the MDS plot x-axis. Significance levels are shown for $p < 0.001$ (***), $p < 0.01$ (**), and $p < 0.05$ (*). “ME/CFS(sa)”, “ME/CFS(mm)” and “MS” were testing whether being a member of the given subgroup was correlated with the axis, relative to HC. “Age” was the linear term of the model fit for that covariate. “PBMC Number” was the total number of cells for that sample, and “CD8⁺ % Input” was the percentage of input PBMCs that were CD8⁺ cells. “CD8⁺ Cells Post-selection” was the number of CD8⁺ cells after the SureSelect process (see **Section 4.3**), and “Purity” the corresponding CD8⁺ purity at this stage. “DNA Extracted (μ g)” was the total amount of DNA initially recovered from each sample, and “DNA Yield (ng)” was the amount remaining after size selection (see **Section 4.3**). “Demultiplexed Reads [Library]” were the total reads recovered for the two Illumina flowcells of sufficient quality (library numbers 641 and 782), and “PEAR Assembled” were the percentages assembled into clonotypes for the same libraries using the Paired-End Read Merger tool [286]. “[Chain Type] Coding Junctions” were the total number of coding junction of each chain type found, and “[Chain Type] %” was the percentage of each chain type, relative to the other two types (such that the sum was 100%).

Variable	Estimate	Std. Error	t-value	$Pr(> t)$	Sig.
(Intercept)	4.93E+10	8.91E+10	0.553	0.581	
ME/CFS (mm)	-1.48E+01	1.29E+02	-0.115	0.908	
ME/CFS (sa)	1.77E+02	1.48E+02	1.20	0.235	
MS	1.50E+02	1.30E+02	1.16	0.249	
Age	3.49E+01	1.72E+02	0.202	0.840	
PBMC Number	-2.28E+01	4.11E+01	-0.554	0.580	
CD8 ⁺ % Input	2.70E+00	1.15E+01	0.234	0.815	
CD8 ⁺ Cells Post-selection	3.02E+02	3.15E+02	0.959	0.340	
Purity	-4.29E+00	3.35E+00	-1.28	0.203	
DNA Extracted (μ g)	2.42E+00	6.07E+01	0.04	0.968	
DNA Yield (ng)	1.88E-01	1.57E-01	1.20	0.232	
Demultiplexed Reads (641)	5.19E-05	2.20E-05	2.36	0.0199	*
PEAR Assembled (641)	-5.77E+00	6.79E+00	-0.85	0.397	
Demultiplexed Reads (782)	-6.02E-06	1.72E-05	-0.351	0.726	
PEAR Assembled (782)	-1.04E+00	7.51E+00	-0.138	0.890	
α Coding Junctions	1.56E-01	9.56E-02	1.63	0.105	
β Coding Junctions	-4.50E-01	1.33E-01	-3.38	0.000998	***
γ Coding Junctions	-2.47E-01	1.23E-01	-2.00	0.0474	*
α %	-4.93E+10	8.91E+10	-0.553	0.581	
β %	-4.93E+10	8.91E+10	-0.553	0.581	
γ %	-4.93E+10	8.91E+10	-0.553	0.581	

Table 5.5: Significance values for the variables tested as potential confounders with the MDS plot y-axis. Significance levels are shown for $p < 0.001$ (***), $p < 0.01$ (**), and $p < 0.05$ (*). Parameters are the same as for **Table 5.4**.

significant predictor ($p = 0.00775$ for the linear term).

For the y-axis, significant associations were found with the number of β and γ chain coding junctions recovered. For the x-axis, the same was true for the α and γ chains. This provides further justification for performing analyses separated by chain type in subsequent sections.

Both axes were associated with the number of coding junctions recovered in library XPBF641. This means that despite the multiple steps taken to reduce the impact of the widely varying recovery rates by sample (by restricting the range of measures used [Section 4.6.1], and downsampling each entry in the distance matrix), the number of recombinants in each sample has still influenced the pairwise distances calculated between samples. The most likely reason for this residual association is that the smallest samples have been downsampled less frequently compared to the larger samples (in fact, the smallest sample is downsampled in at most one pairwise comparison), and so they will tend to yield a less accurate estimate of their diversity. This association may also have led to the presence of outliers.

Figure 5.2 shows that it is possible to draw a decision boundary for $x > 2,000$ to separate outlier data points for the CD8⁺ cell TCRs. To make this more rigorous, the MDS plot is centred (at (0, 0)) on the average of the columns in the distance matrix. Hence, drawing a circle in this space shows points that are greater than a fixed distance from the average of the data (in a similar manner to Mahalanobis distance used in principal component analysis [289]). Choosing a distance that captured the visible outliers while including as few points in the main cluster as possible leads to the boundary shown in **Figure 5.3**.

Although not statistically significantly distinct, the 12 samples that are noticeable outliers in **Figure 5.3** are all disease cases (ME/CFS and MS), and their diversity plots are shown in **Figure 5.4**.

All seem to have a large number of singletons (and hence a high richness). This is potentially consistent with the x-axis of **Figure 5.3** correlating with the downstream effects of age, as research has shown that there is a sudden contraction in T-cell diversity around 65 – 75 years of age [290], which is only slightly above the upper age limit in this study. As these samples tend to be in younger age groups than those in the main cluster, it is hence possible that this may be an observation of more diverse T-cell repertoires. This would be a logical consequence of their relatively increased thymic activity discussed previously (where age was found to be a significant predictor of CD8⁺ TREC count). In other words, younger peo-

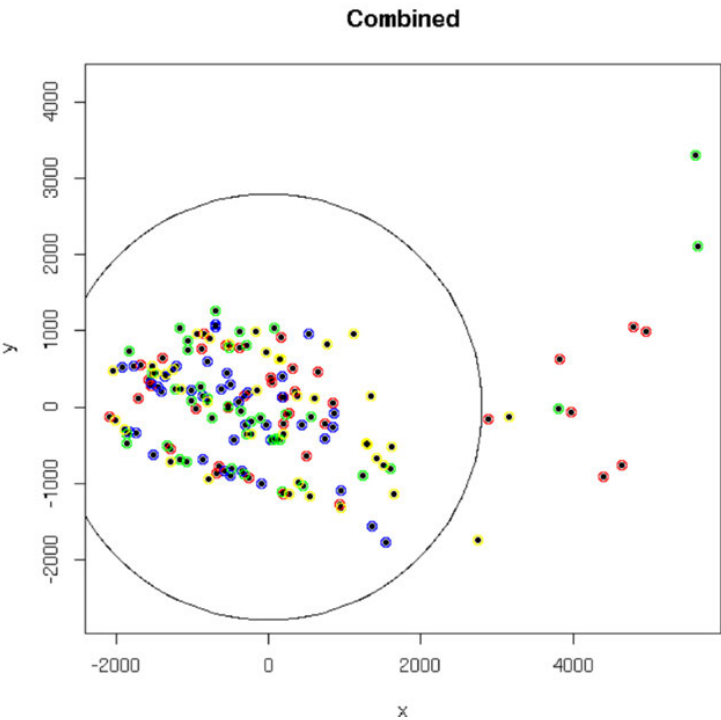


Figure 5.3: The MDS plot of **Figure 5.2**, with a circle of radius 2,800 in this coordinate space. 12 points were outside of the circle, all disease cases.

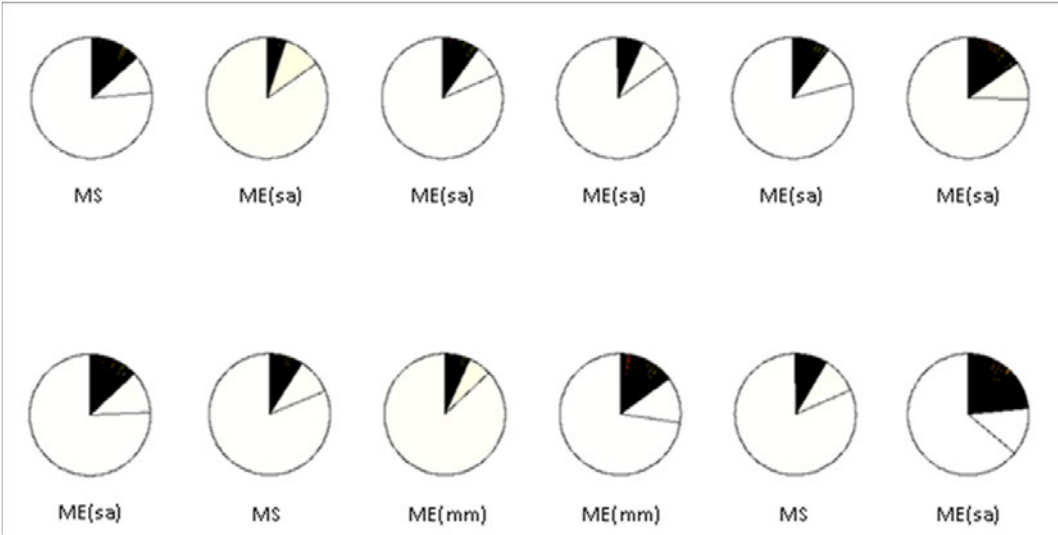


Figure 5.4: The dartboards for the CD8⁺ outliers, as initially shown in **Figure 5.1**.

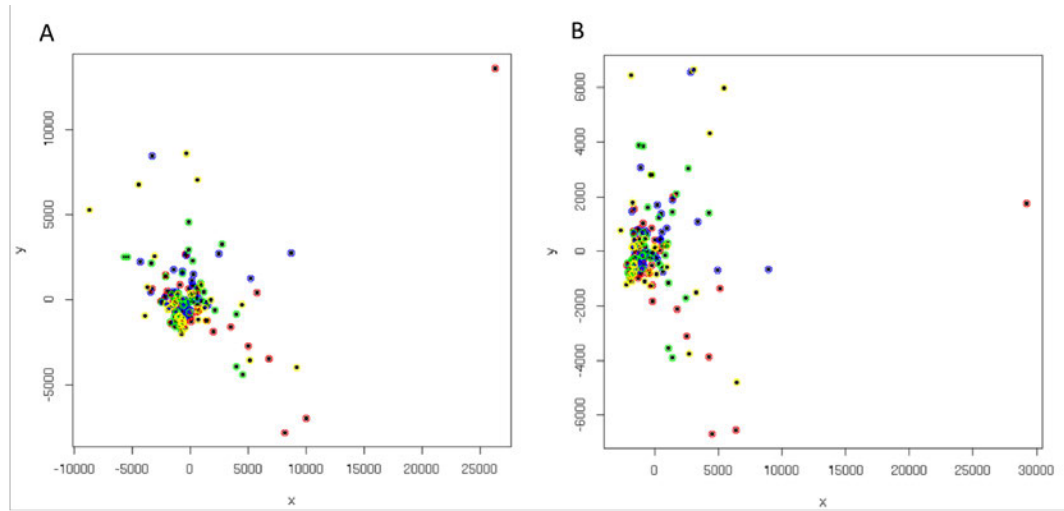


Figure 5.5: MDS plots showing the changes to the distance matrix caused by adjusting the step size and α -range used in the calculations. Plots are shown for the α -chain $CD8^+$ data (compare with **Figure 5.6**) with:

- A)** An α -range of $[0, 20]$ and a step size of 0.1.
B) An α -range of $[0, 10]$ and a step size of 0.2, as in Greiff *et al.*

ple have greater thymopoiesis, which leads to greater T-cell turnover, and hence greater clonotype richness. The distances in the matrix M reflect differences in diversities, and this would cause separation by age in the MDS plot along the x-axis (which shows the greatest component of the distance). Hence, there are outliers from the main group, separated by differences in TCR diversity, which are explicable by their age.

5.3.2 Separated by Chain Type

Analyses were then performed on α , β , and γ chains separately, to investigate whether true diversity differences in one chain type was hidden by the others. For example, it would be plausible that the α chains (which tended to have the most recombinants in total), when added to the fewer γ chains, could obscure any diversity differences in the latter. The effects of different α -ranges and step sizes (see **Section 4.6.1**) were also investigated. **Figure 5.5**, and its comparison with **Figure 5.2**, shows the dramatic effects of using different parameters in the generation of the distance matrix, and hence underlies the importance of using the more impartial algorithm of **Section 4.6.1** to set these in advance.

Individual MDS plots were generated (with the same parameters as before) to visualise the separated matrices (**Figure 5.6**). Note that MDS plots are invariant under rotation and reflection.

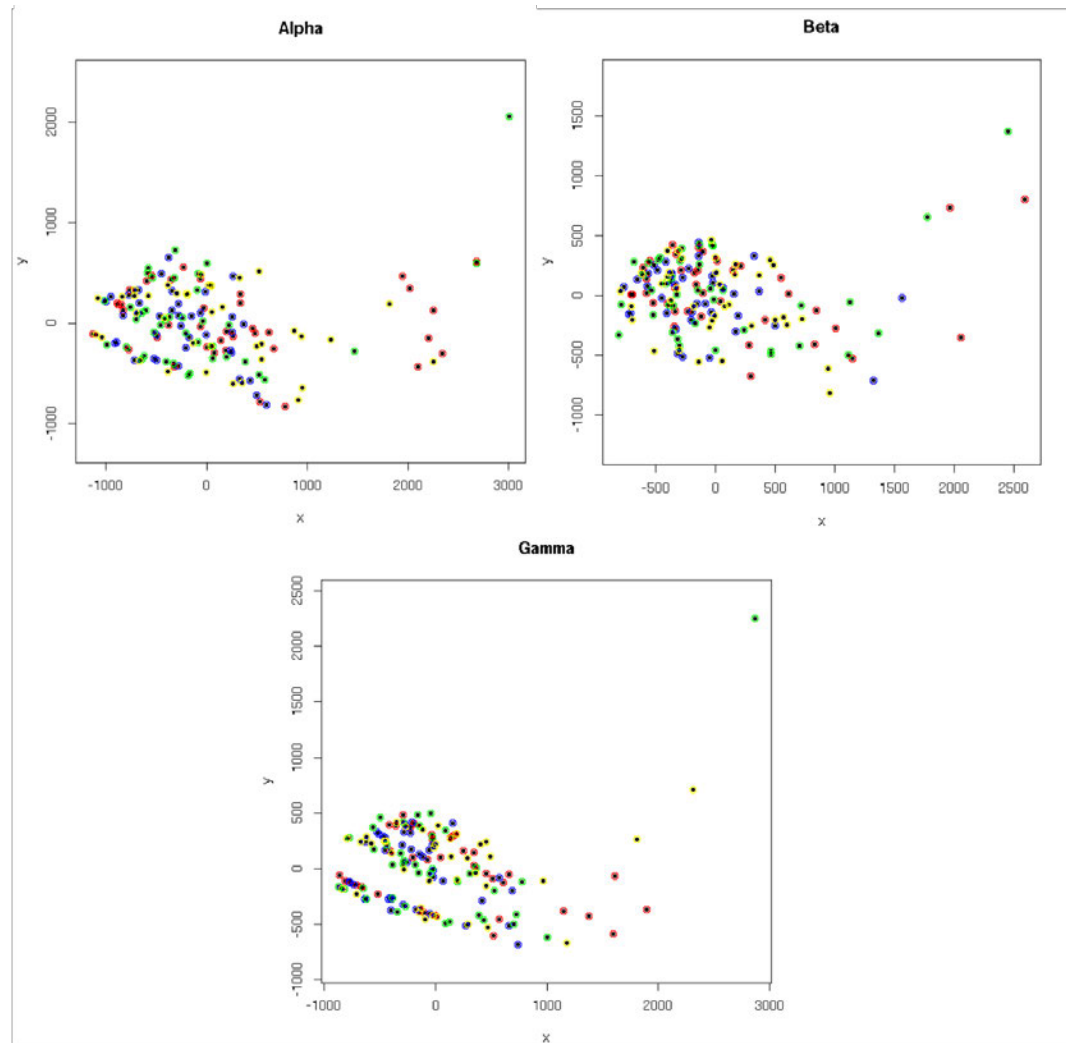


Figure 5.6: Multidimensional scaling plots for the three chain types of $CD8^+$ cells. α , β and γ are shown clockwise from the top left. $ME(sa)$ are in red, $ME(mm)$ are in yellow, MS are in green and HC are in blue. No clear separation is visible in two dimensions, although it appears that the outliers of **Figure 5.2** are primarily driven by diversity differences in the α chains, consistent with the correlation found in **Table 5.4**.

CD8⁺ Primary Hypotheses	α	β	γ
ME vs HC	0.224	1.00	1.00
ME vs MS	0.236	1.00	1.00
ME(sa) vs ME(mm)	0.395	0.224	0.433

Table 5.6: *p*-values for the primary hypotheses for the TCRs from CD8⁺ cells. “ME” indicates ME(sa) and ME(mm) combined.

CD8⁺ Secondary Hypotheses	α	β	γ
Cases vs HC	1.00 [†]	1.00 [†]	1.00 [†]
MS vs HC	0.998	0.715	0.709
ME(mm) vs HC	0.547	0.207	0.301
ME(sa) vs HC	0.020*	0.597	0.133
ME(mm) vs MS	0.193	0.451	0.553
ME(sa) vs MS	0.129	0.597	0.267

Table 5.7: *p*-values for the secondary hypotheses for the TCRs from CD8⁺ cells. “Cases” indicates ME(sa), ME(mm) and MS combined, “ME” indicates ME(sa) and ME(mm) combined. One hypothesis had $p < 0.05$ (*), and three failed to reach 1,000 permutations (†).

Testing with the PSVM revealed no significant differences ($p < 0.05$) between the groups for the three primary hypotheses (**Table 5.6**).

The *p*-values for the secondary analyses are tabulated in **Table 5.7**. The only comparison that was significant at $p < 0.05$ was for the α chains of severely affected ME/CFS cases and healthy controls. However, this result was not significant after the post-analysis Bonferroni correction across all 54 hypotheses ($p < 0.0009$).

5.4 CD4⁺ Cells

An equivalent analysis, separated by chain type, was performed for recombinants acquired from CD4⁺ cells.

148 samples of the initial 160 were sequenced. Of the remaining 12, six had insufficient DNA, three failed the UMI ligation step, two had a barcoding error, and one failed the pooling stage. Once again, the samples were distributed relatively evenly across the four groups, although with slightly fewer severely affected ME/CFS cases (**Table 5.8**). The clonotype plots demonstrated that the CD4⁺ samples had fewer clonotypes in general than the CD8⁺ samples (a lower richness). This can be seen by noting the greater number of yellow plots in **Figure 5.7** compared with **Figure 5.1** (the colour for each sample desaturates as more clonotypes are added).

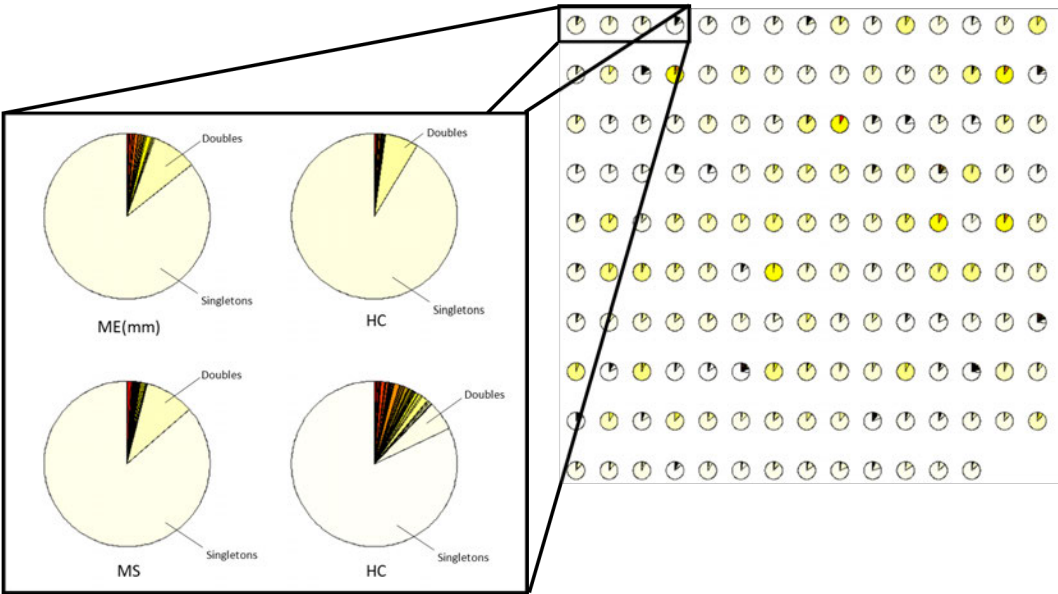


Figure 5.7: Repertoires of the 148 $CD4^+$ samples. The first four are shown in greater detail, with subgroup labels. Colours are assigned to each clonotype in a sample, starting with red and gradually desaturating to white. More monochrome plots therefore indicate samples with a greater richness.

Subgroup	Final Number
ME/CFS(sa)	35
ME/CFS(mm)	37
Multiple Sclerosis	39
Healthy Controls	37

Table 5.8: Number of $CD4^+$ samples, by subgroup, available for analysis.

5.4.1 Separated by Chain Type

MDS plots for the α , β and γ chain matrices were generated (**Figure 5.8**). Although some structure was visible, there were no obvious outliers which merited further analysis.

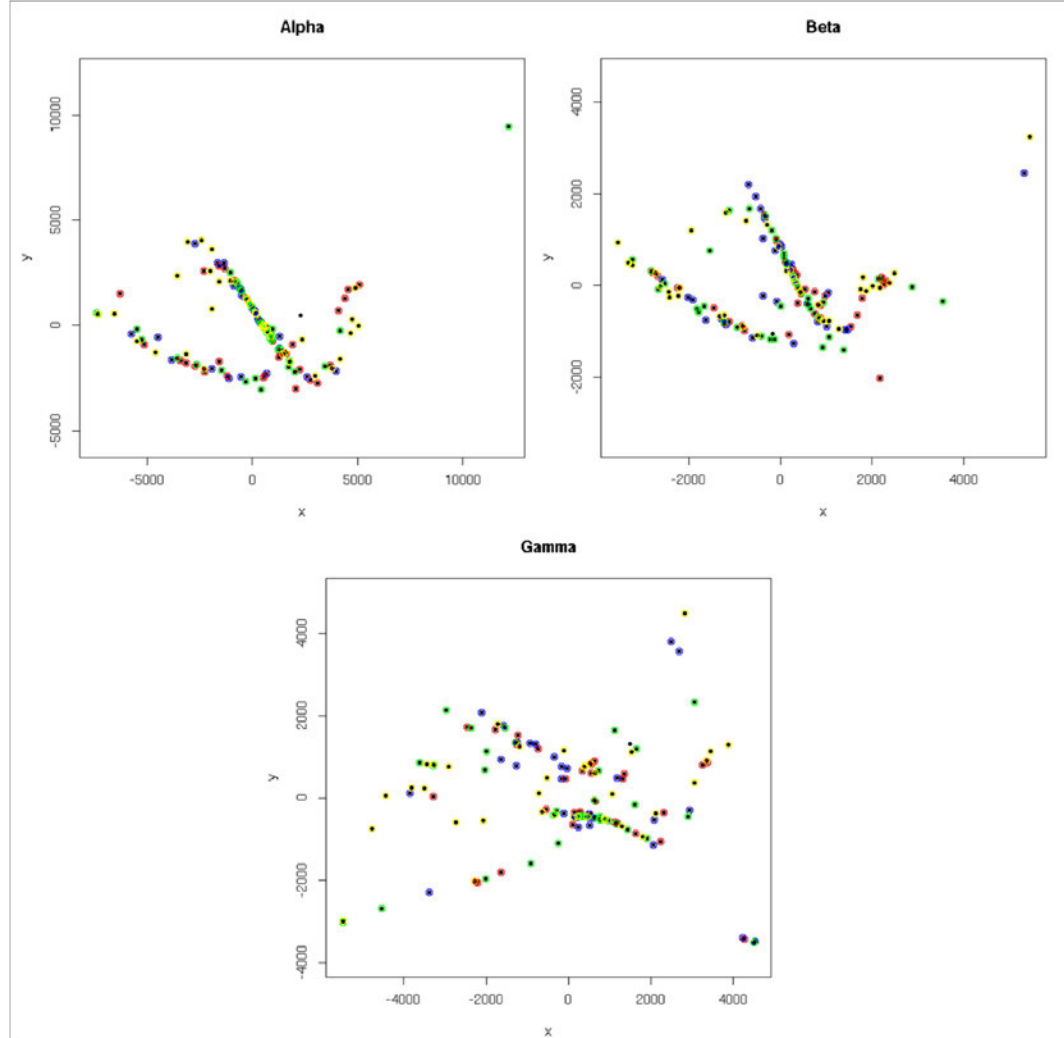


Figure 5.8: Multidimensional scaling plots for the three chain types of $CD4^+$ cells. α , β and γ are shown clockwise from the top left. ME(sa) are in red, ME(mm) are in yellow, MS are in green and HC are in blue. No clear separation was visible in two dimensions.

Testing with the PSVM revealed one significant difference at the $p < 0.05$ level, for the β chains of ME/CFS cases compared with healthy controls (**Table 5.9**). However, this did not pass the post-analysis Bonferroni multiple testing correction across the 18 primary hypotheses ($p < 0.0028$). The p -values for the secondary analyses are tabulated in **Table 5.10**.

Two results were significant at the $p < 0.05$ level. For the β chains, severely affected ME/CFS cases differed from multiple sclerosis, and for the γ chains all

CD4 ⁺ Primary Hypotheses	α	β	γ
ME vs HC	0.137	0.030*	0.053
ME vs MS	0.663	0.351	0.293
ME(sa) vs ME(mm)	0.396	0.875	0.735

Table 5.9: *p*-values for the primary hypotheses for the TCRs from CD4⁺ cells. “ME” indicates ME(sa) and ME(mm) combined. One hypothesis had $p < 0.05$ (*).

CD4 ⁺ Secondary Hypotheses	α	β	γ
Cases vs HC	0.301 [†]	0.104 [†]	0.024* [†]
MS vs HC	0.379	0.827	0.140
ME(mm) vs HC	0.454	0.385	0.231
ME(sa) vs HC	0.064	0.193	0.318
ME(mm) vs MS	0.135	0.491	0.514
ME(sa) vs MS	0.662	0.020*	0.588

Table 5.10: *p*-values for the secondary hypotheses for the TCRs from CD4⁺ cells. “Cases” indicates ME(sa), ME(mm) and MS combined, “ME” indicates ME(sa) and ME(mm) combined. Two hypotheses had $p < 0.05$ (*), and three failed to reach 1,000 permutations ([†]).

disease cases were separated from healthy controls. Neither result passed the post-analysis Bonferroni correction ($p < 0.0009$).

5.5 Axis Correlations by Chain Type

Further to the initial analyses of axis correlations for the CD8⁺ cells in **Section 5.3.1**, similar associations were tested for both cell types, separated by chain type. The results of these are summarised in **Table 5.11**.

The associations by chain type are consistent with those seen in **Tables 5.4** and **5.5** – the age correlation along the x-axis is particularly clear for the CD8⁺ α -chains, which are the majority in these data. The number of reads demultiplexed is also associated with the x-direction for all three chain types, as seen previously.

For the CD4⁺ cells, the y-axis is very strongly associated with input material. Significance is seen for the CD4⁺ cell percentage selected for the β -chains, and for input PBMC number, CD4⁺ cells recovered, and purity for the α -chains. This may help to explain the structure seen in **Figure 5.8**, where lines in the MDS plots are seen clearly for the α - and β -chains, but not the γ -chains. If this is the case, differing quantities of input material, or reduced efficiency of the SureSelect procedure (see **Section 4.3**), may have biased the estimates of diversity. Although the resampling procedure was intended to protect against this type of bias, as

Cell Type MDS Axis	CD8 ⁺						CD4 ⁺					
	α	β	γ	α	β	γ	α	β	γ	α	β	γ
ME/CFS(mm)	0.408	0.593	0.367	0.965	0.691	0.872	0.737	0.782	0.123	0.158	0.0666	0.985
ME/CFS(sa)	0.815	0.326	0.998	0.247	0.117	0.671	0.480	0.459	0.0814	0.505	0.549	0.395
MS	0.937	0.460	0.350	0.220	0.496	0.216	0.482	0.717	0.919	0.757	0.744	0.467
Age	0.000709	0.370	0.0119	0.490	0.0212	0.0810	0.398	0.423	0.132	0.318	0.164	0.233
PMBC	0.567	0.116	0.259	0.729	0.367	0.782	0.708	0.427	0.911	0.0475	0.184	0.637
(Cell Type) % Input	0.178	0.296	0.349	0.449	0.556	0.596	0.739	0.475	0.0771	0.000805	0.217	0.0977
(Cell Type) % Post Selection	0.429	0.359	0.536	0.402	0.387	0.328	0.905	0.719	0.915	0.0860	0.0393	0.409
(Cell Type) Purity	0.773	0.863	0.966	0.157	0.437	0.144	0.965	0.393	0.118	0.000900	0.760	0.676
(Cell Type) Extracted DNA	0.998	0.282	0.865	0.884	0.879	0.797	0.133	0.762	0.506	0.520	0.281	0.897
(Cell Type) Yield DNA	0.678	0.456	0.433	0.345	0.701	0.794	0.528	0.758	0.577	0.805	0.798	0.229
Demultiplexed Reads 641	0.029	0.00895	0.0536	0.0207	0.720	0.303	NA	NA	NA	NA	NA	NA
PEAR 641	0.701	0.363	0.114	0.861	0.0766	0.589	NA	NA	NA	NA	NA	NA
Demultiplexed Reads 782	0.785	0.869	0.506	0.632	0.197	0.731	NA	NA	NA	NA	NA	NA
PEAR 782	0.335	0.354	0.102	0.468	0.922	0.906	NA	NA	NA	NA	NA	NA
Demultiplexed Reads 948	NA	NA	NA	NA	NA	NA	0.313	0.341	0.326	0.922	0.914	0.787

Table 5.11: Significance values from the GLM fit (`glm()` in R) for the variables tested as potential confounders with the MDS plot x- and y-axes. Parameters are the same as for Table 5.4, with the addition of “Demultiplexed Reads 948”, which were the total reads recovered for the Illumina flowcell (library number 948) used for the CD4⁺ cells of sufficient quality. PEAR assembly was not used for the CD4⁺ cell data. Associations with $p < 0.05$ are marked in bold.

		CD8 ⁺			CD4 ⁺		
		α	β	γ	α	β	γ
Primary Hypotheses	ME vs HC	0.224	1.00	1.00	0.137	0.030*	0.053
	ME vs MS	0.236	1.00	1.00	0.663	0.351	0.293
	ME(sa) vs ME(mm)	0.395	0.224	0.433	0.396	0.875	0.735
Secondary Hypotheses	Cases vs HC	1.00 [†]	1.00 [†]	1.00 [†]	0.301 [†]	0.104 [†]	0.024* [†]
	MS vs HC	0.998	0.715	0.709	0.379	0.827	0.140
	ME(mm) vs HC	0.547	0.207	0.301	0.454	0.385	0.231
	ME(sa) vs HC	0.020*	0.597	0.133	0.064	0.193	0.318
	ME(mm) vs MS	0.193	0.451	0.553	0.135	0.491	0.514
	ME(sa) vs MS	0.129	0.597	0.267	0.662	0.020*	0.588

Table 5.12: p -values from permutation testing for CD4⁺ and CD8⁺ cells, by chain type. Asterisks (*) indicate results with $p < 0.05$, and [†] indicates $< 1,000$ permutations were used to generate the p -value.

previously discussed, there may be an issue with the precision of the distance calculations for the smaller samples (see Section 5.3.1).

It is also possible that the sample batches used by SBL has had an impact. Although SBL did not detect serious differences between batches, this effect was not modelled here due to the large number of batches ($\sim \frac{160}{3}$ per cell type). These are unlikely to be the cause of the structure seen in Figure 5.8, as the lines there contain many more than the 3 samples used per batch, but the influence of this effect cannot be ruled out completely.

5.6 Summary

P -values from the testing of all hypotheses are summarised in Table 5.12.

The four results with $p < 0.05$ are summarised in Table 5.13, with multiple testing

Hypothesis Type	Hypothesis	T-cell Type	Chain	p -value	p -value (pre)	p -value (post)	$p < 0.01$
Primary	ME vs HC	CD4 ⁺	β	0.030	FALSE	FALSE	FALSE
Secondary	ME(sa) vs MS	CD4 ⁺	β	0.020	FALSE	FALSE	FALSE
Secondary	Cases vs HC	CD4 ⁺	γ	0.024	FALSE	FALSE	FALSE
Secondary	ME(sa) vs MS	CD8 ⁺	α	0.020	FALSE	FALSE	FALSE

Table 5.13: Table summarising the $p < 0.05$ results of **Table 5.12**. This shows (TRUE/FALSE) which of the multiple testing thresholds they pass. These are (rounded for 1,000 permutations):

- The pre-analysis threshold (“pre”) established in **Section 4.7** ($p < 0.017$ for primary hypotheses, and $p < 0.006$ for secondary hypotheses);
- The stricter post-analysis threshold (“post”) recalculated in **Section 5.2** ($p < 0.003$ for primary hypotheses, and $p < 0.001$ for secondary hypotheses);
- The literature established level of $p < 0.01$ [269].

thresholds as established pre-analysis, post-analysis, and as reported by Greiff, *et al.* (who used $p < 0.01$, with 1,000 shuffles) [269].

5.6.1 Discussion

The testing of four hypotheses yielded $p < 0.05$ (*, **Table 5.12**). None of these remained significant after Bonferroni multiple hypothesis testing correction, for any of the three corrections proposed (**Table 5.13**). From these, I conclude that none of the null hypotheses should be rejected.

As discussed in **Section 5.2**, due to computational constraints, permutation tests were run for all hypotheses until they completed 1,000 label shuffles, or hit a time-out. Tests were rerun in all cases to establish a p -value for 1,000 permutations, except for the Cases vs Healthy Controls (Cases vs HC) hypotheses (marked with a † in **Table 5.12**). These reached a time-out even after 25 days of calculations, and hence are less precise than the other values. However, only one of these had $p < 0.05$; the classification boundary testing the hypothesis of Cases vs HC for the CD4⁺ gamma chains. Due to the order of locus rearrangements in T-cell receptors, most gamma loci from these data are expected to be non-productive. Given that there is no pattern of more accurate classification for the other CD4⁺ chain Cases vs HC hypotheses, this is likely of little biological relevance, and so reduced precision for this classification should not have an impact on the conclusions drawn.

Also of note are the $p = 1$ values for the CD8⁺ cells. These indicate the lower bound for the accuracy scores generated by the PSVM. The worst possible classification is equal to the size of the largest category divided by the total number of samples, which is equivalent to drawing a classification boundary outside of the

dataset. Consider an example with 18 samples, where 12 are cases and 6 controls. At worst, the accuracy score will be $\frac{12}{18} \approx 0.67$, if no boundary is drawn and all samples are classified as cases. This represents truly random data, and is seen in the case of the four $p = 1$ hypotheses. For the CD8⁺ β chains, ME vs MS hypothesis, the mean accuracy was 0.661, with 39 misclassified. There are 39 MS samples in the CD8⁺ dataset, consistent with the reasoning above.

The implication is that if any label permutation results in one or more extra points correctly classified, it has a greater accuracy score. Therefore, only permutations that are equally random will not have a better classification boundary assigned. Across the $< 1,000$ permutations used for these non-significant hypotheses, it is plausible that none of these random permutations were generated, and they would only be seen if the p -value were calculated with greater precision. Indeed, it appears that this was the case for the MS vs HC CD8⁺ α chain hypothesis ($p = 0.998$), where in two cases the original data was more easily classified than the background permutations.

To summarise, although this method was able to robustly classify simulated clonotypes (**Section 4.6.5**), it was unable to detect differences between any tested combination of cases and controls in the biologically-derived data. Pipeline refinements may improve the classifications on real data, such as through the testing of alternative distance metrics (the L_1 norm, or even the fractional norms of Aggarwal *et al.* [272], may show clearer separations between groups). However, from these analyses, I conclude that the null hypothesis should not be rejected in favour of the alternate, for any of the tested subgroup combinations.

5.7 Additional Analysis I: Power Law Models

In addition to the primary analysis using the diversity measures and PSVM, I also fit power-law models to the data, which are a commonly used model for T-cell receptor repertoires (see **Section 4.6.5**). This approach of plotting log-log plots of cumulative clonotype counts is used in the TCR repertoire literature [291].

A log-log plot of the cumulative number of recombinants in each clonotype was generated for each sample in the dataset (an example is shown in **Figure 5.9**). A linear model was then fit to this plot, and the gradient and intercept were noted.

A linear model in log-space indicates an exponential relationship, as expected when drawing from clonotype distributions [269]. Plotting histograms for the exponent

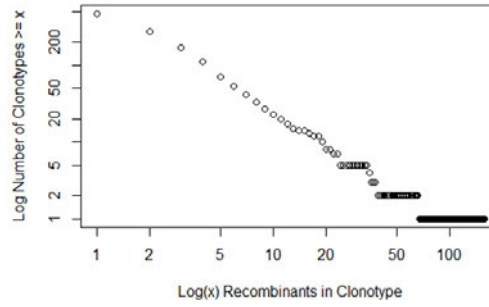


Figure 5.9: Log-log plot showing a linear relationship between cumulative clonotypes and the recombinants in a given clonotype.

(gradient) and intercept for the models fit to all samples in both the $CD4^+$ and $CD8^+$ datasets (**Figure 5.10**) showed unimodal distributions. Multimodal distributions may have been an indication of immune activation, and hence this is consistent with the diversity analyses of **Sections 5.3** and **5.4** showing no significant clonal expansion. However, the skew seen for the $CD4^+$ cells is unexplained. For the exponents, this shows that the $CD4^+$ samples have more even distributions, since a lower exponent implies a more balanced distribution (see the simulated data of **Section 4.6.5**). de Greef *et al.* [292] showed that the type of T-cell may be less notable in these analyses than the differences between chain types (however this was for mRNA rather than gDNA data).

It is likely that the sampling noise in clonotype counts was too great to find a clear signal, and a notable increase in recombinant numbers would be needed to avoid the issues inherent in extrapolating from $10^4 - 10^5$ sequences per donor to the full $\sim 10^{11}$ cells [265] in the human TCR repertoire. Further discussion of issues of power in both this, and the genetics studies of **Chapter 3**, are included in **Chapter 6**.

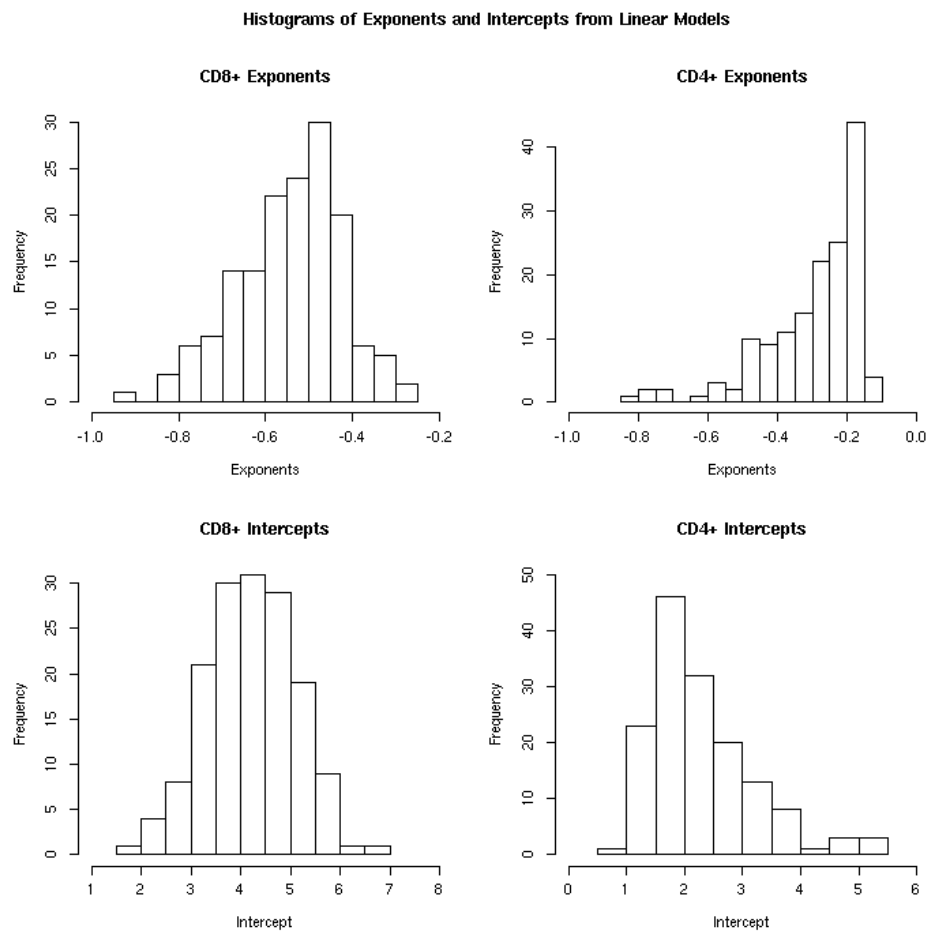


Figure 5.10: Histograms showing the distributions of exponents and intercepts from linear models fit to both $CD4^+$ and $CD8^+$ cell data. All distributions were unipolar.

5.8 Additional Analysis II: Gene Segment Usage

A second analysis was conducted to determine whether any individual V, D or J gene segment was over- or under-represented among the diseases cases. In ME/CFS, it has been shown that there is biased usage of certain B-cell gene segments by Sato *et al.* [153], and biased TCR gene segment usage can indicate an immune response [293]. To investigate whether that was the case in these data, counts of the number of occasions each segment was detected in a V-J or V-D-J recombination were assembled for each participant. This is an investigation of Type I bias (see Turner *et al.* [293]), but also investigates the D and J segments. Note that of interest here was how frequently each segment was used in a pairing event, not which pairings clonally expanded. Therefore, the counts for each segment were not multiplied by the coding junction count for the relevant recombination.

There are several distributions that can be used to model counts, including Poisson, Quasi-Poisson or Negative Binomial [294]. Poisson distributions have only a single parameter, and hence are prone to over-dispersion (where the data variance is greater than the mean) [295]. That was the case here, and hence either a Quasi-Poisson or Negative Binomial model was more appropriate. As the more formal approach [294], a negative binomial model was chosen for these data.

Fitting this distribution across subsets of the data corresponding to the primary hypotheses used in the diversity analysis (**Table 4.3**), and determining whether caseness was a significant predictor in the model, was used to determine whether there was a difference in gene segment usage for the subgroups. For each hypothesis, the subgroup labels were recoded to indicate whether a sample was a case or control. For the disease Case versus Healthy Control hypothesis (Case vs HC), MEsa, MEmm and MS samples were recoded as “Case” and HC were recoded as “Control”. For ME vs MS, MEsa and MEmm were recoded as “Case”, MS were recoded as “Control”, and HC were set to missing (*NA*). For MEsa vs MEmm, MEsa were set to “Case”, MEmm were set to “Control”, and MS and HC were set to *NA*. Before each model fit, the data subsets were checked to ensure that a particular segment was present in at least one sample, or else the result was set to *NA*.

Fitting the models across the gene segments (using *glm.nb()* from the *MASS* package in *R* [296]), and adding the Case/Control variables as covariates, allowed for the determination as to whether caseness significantly affected gene segment usage for each hypothesis. Histograms of the *p*-values for the models which provided an

acceptable fit to the data (where the effect estimate was not NA) are shown in **Figure 5.11**. A slight bias towards smaller p -values was only seen for the $CD8^+$ ME vs MS and the $CD4^+$ MEsa vs MEMm hypotheses. All others tested showed either uniform distributions, or a bias towards greater p -values. However, the peaks of the distributions for the two hypotheses showing a skew to smaller p -values are found for $p \geq 0.05$, indicating that few of these segments are significantly associated, even before multiple hypothesis testing correction. The slight bias seen could be explained by the smaller sample sizes for the ME vs MS and MEsa vs MEMm hypotheses – removal of samples due to both the recoding procedure and model fit failure could be obscuring uniform distributions.

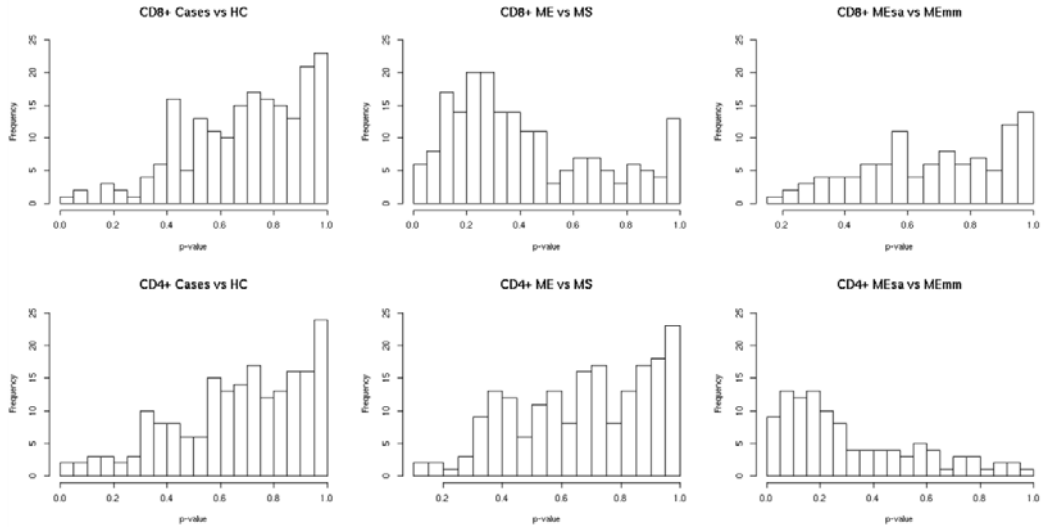


Figure 5.11: Histograms showing the p -value distributions for each of the six hypotheses tested (Case vs HC, ME vs MS, and MEsa vs MEMm, for both $CD8^+$ and $CD4^+$ cells).

Further, all segments with $p < 0.05$ are shown in **Table 5.14**. Assuming that each segment was used in pairing events independently of all others, that each segment can be used independently in both $CD4^+$ and $CD8^+$ cells, and that each hypothesis is independent of the other two, an appropriate multiple testing correction would be $p < \frac{0.05}{(194+193) \times 3} = 4.31 \times 10^{-5}$ for the number of segments (194 $CD8^+$, 193 $CD4^+$, with three hypotheses tested for each). None of the gene segments shown in **Table 5.14** met this threshold. Hence the null hypothesis is not rejected – these data do not support the hypothesis of altered TCR gene segment usage in people with ME/CFS.

Cell Type	Gene Segment	Hypothesis	Estimate	Std. Error	p-value
CD8 ⁺	TRBV7-5	Case vs HC	-0.579	0.257	0.0241
	TRAV35	ME vs MS	0.145	0.0728	0.0468
	TRBJ2-2P	ME vs MS	0.468	0.177	0.00828
	TRBV10-2	ME vs MS	0.259	0.109	0.0178
	TRBV14	ME vs MS	0.294	0.114	0.0100
	TRBV24-1	ME vs MS	0.141	0.0707	0.0457
	TRBV28	ME vs MS	0.124	0.062	0.0471
CD4 ⁺	TRBJ2-2P	Case vs HC	0.355	0.151	0.0188
	TRDJ4	Disease vs HC	0.981	0.486	0.0435
	TRAJ21	MEsa vs MEmm	-0.239	0.117	0.0404
	TRAJ29	MEsa vs MEmm	-0.173	0.0844	0.0404
	TRAJ34	MEsa vs MEmm	-0.202	0.0942	0.0322
	TRAJ38	MEsa vs MEmm	-0.182	0.0869	0.0360
	TRAJ43	MEsa vs MEmm	-0.135	0.0667	0.0437
	TRAJ54	MEsa vs MEmm	-0.145	0.0734	0.0486
	TRAV25	MEsa vs MEmm	-0.193	0.0955	0.0430
	TRBV21-1	MEsa vs MEmm	-0.217	0.0880	0.0135
	TRBV28	MEsa vs MEmm	-0.187	0.0853	0.0279

Table 5.14: Gene segments where caseness was a significant predictor ($p < 0.05$) under a negative binomial model (`glm.nb()` in R, from the MASS package [296]), used to indicate differences in usage among the subgroups. No segments remained significantly associated with differential usage after multiple hypothesis testing correction. Hence, these data do not support rejection of the null hypothesis.

5.9 Additional Analysis III: CMV Seropositivity

As discussed in **Section 1.4**, human herpesviruses are a proposed trigger for ME/CFS, and CMV has been shown to affect clonal diversity (see **Section 4.1.3**). CMV is very common in the UK [297], and hence may have impacted the results of the diversity analyses. CMV seropositivity results were available for 98 of the initially provided 160 samples, and the prevalence by disease subgroup is shown in **Table 5.15**.

Seropositivity	ME(sa)	ME(mm)	MS	HC
Positive	14	4	11	12
Negative	19	9	14	13
Equivocal	2	0	0	0

Table 5.15: CMV seropositivity results, shown by disease subgroup. Samples were sorted into “Positive” and “Negative” based upon the level of CMV immunoglobulin G (IgG) measured in their plasma. “Equivocal” were those with indeterminate levels of IgG.

A χ^2 test [284] showed that this distribution did not deviate significantly from the expected values ($p = 0.575$). Hence, unless there is significant unseen CMV in a particular subgroup in the samples for which data was not available, the presence

or absence of CMV in samples should not have confounded the other results in this chapter.

Chapter 6

Discussion

In this thesis, I have presented studies investigating two different aspects of the disease ME/CFS. All except one result failed to pass multiple hypothesis testing correction, and that association merits further study. However, the other previously found associations, and new hypotheses, remain unproven. There are several possible reasons for this.

First, ME/CFS is a highly heterogeneous disease. Given the wide variety of potential causes and symptoms (see **Chapter 1**), it is possible, and even likely, that the label is applied to a set of related diseases. Since the aetiology is still unknown, it is difficult to define cohorts that can be used for research purposes. Discovery of a biomarker or clinical test remains among the most urgent topics for researchers, to enable the more precise definition of ME/CFS cohorts for future studies. In the meantime, research should focus on cohorts diagnosed with as clearly defined ME/CFS as currently possible. In particular, this might involve strict application of one of the more recent case criteria (see **Section 1.1**), and would ideally also involve testing for post-exertional malaise.

This problem of heterogeneity was almost certainly present in the TCR study of **Chapters 4** and **5**, where donors were accepted if they met either the Canadian Consensus and/or Fukuda Criteria [253]. However, this heterogeneity is more visible in the GWAS of **Chapter 3**. There are two phenotypes included in the UK Biobank, with some overlap, and these involve participants' self-reporting of a physician's diagnosis. Evidence has shown that self-reported phenotypes are typically accurate representations of disease [298], but without clear definitions of that disease, the result may still be unreliable. There are also inconsistencies in the way ME/CFS is diagnosed, as seen in the almost completely disjoint ME/CFS and

fibromyalgia phenotypes, despite their overlapping case criteria and symptoms. It is therefore likely that these phenotypes are not accurately portraying genuinely distinct complex disease cohorts.

The second problem is that sample sizes remain small, despite the prevalence of ME/CFS. Although the TCR project had approximately an order of magnitude more samples than Dr Davis' study (~ 160 vs 16), with a similar increase in recombinants ($\gtrsim 10^3$ vs $\sim 10^2$) [156], we were unable to detect differences in TCR clonal diversity. It is likely that there are either no differences to be found, or that we were underpowered to detect them.

Future work could consider refining the pipeline using samples from a disease with a more clearly established incidence of clonal expansion. A good candidate may be T-cell leukemia/lymphoma, where research has shown that more than 50% of sample repertoires can consist of a single clonotype [299]. Refining the approach in this way could then allow for the method to detect the potentially subtler differences in disease cases here. However, it is likely that all current TCR repertoire studies are severely underpowered at present, due to the limitations of current technology. Extrapolating from $10^3 - 10^4$ cells to $\sim 10^{11}$ is likely too many orders of magnitude of undersampling for replicable results. This issue is most relevant for investigating specific clonotypes [300], but has relevance to diversity analysis. To see this, consider a repertoire that is composed of one clonotype expanded to 50%, with the remainder singletons. Samples taken from this repertoire will be representative of the population, with $p = \frac{1}{2}$ for either the expanded clonotype, or a random singleton. Now consider the same, but including a clonotype that has expanded two orders of magnitude (to hundreds of recombinants). This is biologically relevant, as it represents genuine clonal expansion, but assuming equal probabilities of inclusion, it is very unlikely to see this specific clonotype ($p = \frac{10^2}{10^{11}} = \frac{1}{10^9}$) in a sample of size $\sim 10^4$. Even if it is present, it is even less likely to see sufficient such recombinants to recognise that clonotype as expanded.

Power was also an issue for the GWAS, where calculations (**Section 3.1.3**), showed that detecting variants which convey a lower relative risk would be very difficult with the disease cohort sizes available in the UK Biobank. This was mitigated somewhat by creating new larger "Combined" phenotypes, which led to the only genome-wide significant association of the project. However, the UK Biobank will likely remain of limited use in detecting the full genetic aetiology of ME/CFS, as although more diagnoses may be added over time, there would need to be an

approximately order of magnitude increase in cases to detect variants with lower relative risks. Since there are almost certainly multiple risk loci of this nature involved in the aetiology of ME/CFS, GWAS of this greater scale are an urgent research priority. Studies such as DecodeME are currently attempting to create datasets of this size [301].

The genome-wide significant association detected was between the SNP rs76346913 and males with ME/CFS. There is cause for optimism that this is a true association, as the version of the data used in this analysis is the most up-to-date, and has greater power than those in the literature presented in **Chapter 2**. In addition, as a common variant ($MAF = 0.086$), it ranks among the most plausible associations yet found for ME/CFS. However, since this association has never been found in the UK Biobank (or otherwise) previously, it requires replication in an independent cohort before it can be definitively claimed that this is a risk variant for the disease ME/CFS, rather than just for this specific phenotype. As noted above, the mapping between the phenotypes and the actual disease cohorts is not yet fully clear. It is also very unlikely to be the sole variant associated with this disease. Although the odds ratio is notable ($OR = 1.64$, $SE = 0.0814$), the disease is still expected to be complex, and hence have other as-yet-undetected associated variants.

Should the association with the variant be replicated, and *PDE10A* remains implicated, then there are potential follow-up clinical studies that could be conducted. This variant is non-coding, and if causal, is hence likely to have a regulatory function. Given the known functions of *PDE10A*, this may result in differences in mRNA transcription or other measurable changes in nervous system tissue. Notably, lesions have been reported in the brains of those with *PDE10A* mutations [211], and so an MRI-based study may be of interest, to examine whether there are measurable differences in the brains of those with ME/CFS. Previous MRI studies on the disease have found consistent abnormalities in the brainstem [302], but few have focused on the basal ganglia. Those that have found reduced activity in this region to be correlated with fatigue [303]. This brain region is strongly associated with fatiguing diseases and disorders [304], and so is a strong candidate area for further research. *PDE10A* inhibitors were trialed, unsuccessfully, for the treatment of schizophrenia [206] (and are proposed as a treatment for obesity and diabetes [209]), but if the effect of the allele is to cause an increase in cAMP or cGMP hydrolysis, they may prove to have a therapeutic effect here by mitigating the neurological symptoms.

The other associations detected had greater p -values, but are still worthy of note. rs10896588 has not been associated with ME/CFS before, but its effects on cAMP (via *SLC43A3*) and notable expression levels in the thyroid [305] may link it to a similar pathway to *PDE10A*. *SLC43A3* is also expressed in EBV-transformed lymphocytes [305], providing an alternative potential link to an ME/CFS pathway (see **Section 1.4**). rs12145346 has seen associations in GWAS previously, but also only at a suggestive level, and only using the UK Biobank. Given its potential link to metabolic disease, it may be worthy of further study. As previously discussed (**Section 2.1.6**), candidate gene approaches are not, in general, robust, and hence both of these would require GWAS replication in a non-UK Biobank population before their implications for ME/CFS can be fully considered.

The SNPs near *SLC25A15* detected by Neale *et al.* were again found to be associated only in the female cohort, but only at the suggestive level. The female bias may be explained by the fact that there are more females than males with the CFS phenotype, which may provide greater power to detect associations due to a larger sample size (however the result was not seen in the even larger “Combined” cohorts). Although other GWAS using the UK Biobank data did not identify rs7337312 as a risk variant, it was still the most plausible association identified in the literature (see **Section 2.1.4**), and so warranted further investigation. Although the replication presented here successfully identified SNPs in this region as suggestive, the hypothesis regarding modified dietary protein intake due to reduced *ORNT1* protein was not supported. Further work would need to measure levels of *ORNT1* mRNA or protein directly to definitively refute this hypothesis.

For the HLA loci, the associations found by Lande *et al.* have greater p -values than would usually be expected in population genetics. Locus-wise multiple testing correction is insufficient to ensure robust results in this case, as it requires a prior expectation of associations between a phenotype and a specific genomic region. Hence, I used the same suggestive and genome-wide significance thresholds as in the GWAS. In addition, an allele frequency comparison method fails to account for confounding factors. The analysis here uses the approach of fitting individual logistic models to each allele, allowing for the addition of covariates. With these modifications, no allele reached the threshold for even suggestive significance, and hence the hypothesis that ME/CFS is a T-cell associated autoimmune disease is not supported by the UK Biobank data.

Future Hypotheses

There are three plausible hypotheses for the aetiology of ME/CFS as a result of this research.

The first is that ME/CFS is a disease of the central nervous system (CNS), mediated through either *PDE10A* or *SLC25A15* (or both). Specifically, that ME/CFS is caused by disruption to *PDE10A*, which affects nervous system signalling in a similar fashion to Parkinson's or Huntington's disease, or Chorea (see **Section 3.1.7**). This would be consistent with the neurological symptoms of ME/CFS, and with evidence suggesting that neurologically-impaired muscle recruitment is relevant to the results of exercise testing [30]. For this hypothesis to be viable, further research would need to demonstrate that the rs76346913 polymorphism affects neuronal signalling. This may explain at least some of the very severe cases who never recover, as repairing damage or impairment to the nervous system is a very slow process. As an example, see Guillan-Barré Syndrome (**Section 1.5.10**).

In the second CNS case, that ME/CFS is caused by ammonia build-up affecting CNS signalling [30]. This would be consistent with the associations detected for *SLC25A15*, and with observations that increases in ammonia affect fatigue levels. The rs7337312 polymorphism may be associated with reduced *ORNT1* protein levels, and hence an extended period of elevated ammonia levels in blood. It is possible that such a build-up of ammonia in blood would increase levels in the brain, leading to CNS signalling problems [30], and it is unclear how long this would take to dissipate. Hence this effect would be consistent with the symptoms of PEM under repeated exercise challenges. Further investigation of this hypothesis should involve direct measurement of ammonia levels before, during, and after exercise testing. This as with the first hypothesis, may help to explain severe cases who do not recover. Equally, it is plausible that a degree of tolerance to ammonia accumulation may be acquired over time, indicating a path to the unexplained symptom improvement, and in some cases full recovery, observed in ME/CFS (see **Section 1.2**).

Third, that ME/CFS is an endothelial disease, mediated through (among others) *PDE10A*. This would be consistent with the blood-pressure associated variant found by Warren *et al.* [214], orthostatic intolerance symptoms, and PEM symptoms induced by cardiovascular testing. *PDE10A* protein has been found in some endothelial tissues [306][307], and hence may be relevant to the proposed blood-pressure pathway. However, this hypothesis relies upon the connection (through

LD) between two not yet replicated results, one of which was seen in males only. Further research would need to confirm the connections between *PDE10A* and ME/CFS, and *PDE10A* with endothelial cell dysfunction. This is likely best accomplished with replication GWAS.

As ME/CFS is a complex disease, these hypotheses are not mutually exclusive, and some combination of these is more plausible than any one individually.

Concluding Remarks

As predicted, GWAS proved the most useful tool for investigating this disease, and found the only statistically significant association across all of the studies. GWAS remains the best starting point for ME/CFS research, given its proven track record for finding genetic associations in other diseases (for example, Alzheimer's disease [6][308]). The creation of new composite phenotypes means that this study is the best-powered investigation into ME/CFS using the UK Biobank to date. Alongside the *SCL25A15* association previously reported in the literature, this represents one of the strongest genetic signals yet detected for ME/CFS. Further work continuing from this thesis should focus on the gene *PDE10A* and the effects associated with its disruption. A replication GWAS should be the first step, with the hope of confirming rs76346913 as a risk allele. Future GWAS should focus on recruitment of larger numbers of people with ME/CFS and related diseases (such that power is sufficient), and accurate phenotyping. Research should then focus on investigating the clinical implications of this variant, with the hope of finding a treatment for this terrible disease.

Acknowledgements

I would like to thank my family and Audrey Sherr for their support throughout my PhD.

I would like to thank the Chief Scientist's Office (Scotland) and Action for ME (AfME), who jointly funded this research.

I also need to thank Dr Nick Parkinson, Dr Matt Roddis, Dr Ben Ferneyhough, Dr Sam Millington, Sean West and Mike Fischer at Systems Biology Laboratory (SBL) for their work processing the samples for the T-cell project, as well as the additional resources provided beyond that allowed by the project funding.

I would like to thank the current, and former, members of the Ponting group from my time in Edinburgh. In particular, Øyvind Almelid for his assistance with various coding problems, Talitha Bromwich for help with the data management required for the genetics project, and Gemma Samms for many discussions regarding GWAS and PLINK. Thibaud Boutin provided advice on the GWAS project, and Kenton D'Mellow and Albert Tenesa assisted with UK Biobank data access. I also need to thank Audrey Ryback, Natasha Smith and Catherine Sutherland of the Cowan group for their advice on the TCR project. Michael Nicholson also provided helpful advice on the TCR project methods, suggested the additional analysis, and provided literature references. I would also like to thank Alex Levine for informally discussing mathematical entropy with me throughout the project.

I need to thank John Ireland and the Research Support team for their help with the computing and data management.

I would like to thank my supervisors, Dr Veronique Vittart for her advice on the genetics project, and Dr Graeme Cowan and Dr Andy Sims, for their comments on the immunology project. I would also like to thank Prof. Chris Haley, and Prof. Cathy Abbott for joining my thesis committee.

This study was carried out using samples from the UK ME/CFS Biobank, which

were processed and stored at the UCL-RFH Biobank. This research has also been conducted using the UK Biobank resource. I would like to thank all of the sample donors from both Biobanks, but especially the severely ill donors of the CureME Biobank – your time and energy makes research such as this possible.

Finally, I need to thank my principal supervisor and mentor, Prof. Chris Ponting. This research would not have been possible without your help, advice and guidance over the past few years. I leave the University of Edinburgh a far better scientist than when I arrived.

References

- [1] H. Wickham. *Committee on the Diagnostic Criteria for Myalgic Encephalomyelitis/Chronic Fatigue Syndrome; Board on the Health of Select Populations; Institute of Medicine. Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Redefining an Illness*. Washington (DC): National Academies Press (US), 2015.
- [2] K. Fukuda, S. E. Straus, I. Hickie, et al. "The chronic fatigue syndrome: a comprehensive approach to its definition and study". *Annals of Internal Medicine* 121.12 (1994), pp. 953–959.
- [3] C. C. Kingdon, E. W. Bowman, H. Curran, et al. "Functional status and well-being in people with myalgic encephalomyelitis/chronic fatigue syndrome compared with people with multiple sclerosis and healthy controls". *Pharmacoeconomics Open* 2.4 (2018), pp. 381–392.
- [4] M. F. Hvidberg, L. S. Brinth, A. V. Olesen, et al. "The health-related quality of life for patients with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)". *PloS One* 10.7 (2015), e0132421.
- [5] E. Roberts, S. Wessely, T. Chalder, et al. "Mortality of people with chronic fatigue syndrome: a retrospective cohort study in England and Wales from the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Clinical Record Interactive Search (CRIS) register". *The Lancet* 387.10028 (2016), pp. 1638–1643.
- [6] J. J. Dibble, S. J. McGrath, and C. P. Ponting. "Genetic risk factors of ME/CFS: a critical review". *Human Molecular Genetics* 29.R1 (2020), R117–R124.
- [7] K. A. Huber, M. Sunnquist, and L. A. Jason. "Latent class analysis of a heterogeneous international sample of patients

- with myalgic encephalomyelitis/chronic fatigue syndrome".
Fatigue: Biomedicine, Health & Behavior 6.3 (2018), pp. 163–178.
- [8] E.-J. Lim, Y.-C. Ahn, E.-S. Jang, et al.
"Systematic review and meta-analysis of the prevalence of chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME)".
Journal of Translational Medicine 18.1 (2020), pp. 1–15.
- [9] B. M. Carruthers, A. K. Jain, K. L. De Meirleir, et al.
"Myalgic encephalomyelitis/chronic fatigue syndrome: clinical working case definition, diagnostic and treatment protocols".
Journal of Chronic Fatigue Syndrome 11.1 (2003), pp. 7–115.
- [10] L. A. Jason, A. Brown, E. Clyne, et al. "Contrasting case definitions for chronic fatigue syndrome, myalgic encephalomyelitis/chronic fatigue syndrome and myalgic encephalomyelitis".
Evaluation & the Health Professions 35.3 (2012), pp. 280–304.
- [11] B. M. Carruthers, M. I. van de Sande, K. L. De Meirleir, et al.
"Myalgic encephalomyelitis: International consensus criteria".
Journal of Internal Medicine 270.4 (2011), pp. 327–338.
- [12] S. Johnston, E. W. Brenu, D. Staines, et al. "The prevalence of chronic fatigue syndrome/myalgic encephalomyelitis: a meta-analysis".
Clinical Epidemiology 5 (2013), p. 105.
- [13] L. C. Nacul, E. M. Lacerda, D. Pheby, et al. "Prevalence of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) in three regions of England: a repeated cross-sectional study in primary care".
BMC Medicine 9.1 (2011), pp. 1–12.
- [14] A. R. Valdez, E. E. Hancock, S. Adebayo, et al.
"Estimating prevalence, demographics, and costs of ME/CFS using large scale medical claims data and machine learning".
Frontiers in Pediatrics 6 (2019), p. 412.
- [15] E.-J. Lim and C.-G. Son. "Prevalence of chronic fatigue syndrome (CFS) in Korea and Japan: a meta-analysis".
Journal of Clinical Medicine 10.15 (2021), p. 3204.
- [16] L. A. Jason, B. Z. Katz, M. Sunnquist, et al.
"The prevalence of pediatric myalgic encephalomyelitis/chronic fatigue syndrome in a community-based sample". *Child & Youth Care Forum*. Vol. 49. 4. Springer. 2020, pp. 563–579.
- [17] I. J. Bakken, K. Tveito, N. Gunnes, et al. "Two age peaks in the incidence of chronic fatigue syndrome/myalgic encephalomyelitis: a

- population-based registry study from Norway 2008–2012”.
BMC Medicine 12.1 (2014), pp. 1–7.
- [18] O. L. Quintero, M. J. Amador-Patarroyo, G. Montoya-Ortiz, et al.
“Autoimmune disease and gender: Plausible mechanisms for the female predominance of autoimmunity”.
Journal of Autoimmunity 38.2 (2012), J109–J119.
- [19] R Cairns and M Hotopf. “A systematic review describing the prognosis of chronic fatigue syndrome”.
Occupational Medicine 55.1 (2005), pp. 20–31.
- [20] L. C. Nacul, E. M. Lacerda, P. Campion, et al.
“The functional status and well being of people with myalgic encephalomyelitis/chronic fatigue syndrome and their carers”.
BMC Public Health 11.1 (2011), pp. 1–11.
- [21] K. Albouaini, M. Egred, A. Alahmar, et al.
“Cardiopulmonary exercise testing and its application”.
Postgraduate Medical Journal 83.985 (2007), pp. 675–682.
- [22] E.-J. Lim, E.-B. Kang, E.-S. Jang, et al.
“The prospects of the two-day cardiopulmonary exercise test (CPET) in ME/CFS patients: a meta-analysis”.
Journal of Clinical Medicine 9.12 (2020), p. 4040.
- [23] P. Joseph, C. Arevalo, R. K. Oliveira, et al.
“Insights from invasive cardiopulmonary exercise testing of patients with myalgic encephalomyelitis/chronic fatigue syndrome”.
Chest 160.2 (2021), pp. 642–651.
- [24] R. Esfandyarpour, A Kashi, M. Nemat-Gorgani, et al.
“A nanoelectronics-blood-based diagnostic biomarker for myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)”. *Proceedings of the National Academy of Sciences* 116.21 (2019), pp. 10250–10257.
- [25] R Baker and E. Shaw.
“Diagnosis and management of chronic fatigue syndrome or myalgic encephalomyelitis (or encephalopathy): summary of NICE guidance”.
Bmj 335.7617 (2007), pp. 446–448.
- [26] S. G. Hofmann, A. Asnaani, I. J. Vonk, et al.
“The efficacy of cognitive behavioral therapy: a review of meta-analyses”.
Cognitive Therapy and Research 36.5 (2012), pp. 427–440.
- [27] C. X. Sandler, D. Goldstein, S. Horsfield, et al.
“Randomized evaluation of cognitive-behavioral therapy and graded

- exercise therapy for post-cancer fatigue".
Journal of Pain and Symptom Management 54.1 (2017), pp. 74–84.
- [28] J. J. Geraets, M. E. Goossens, I. J. de Groot, et al.
"Effectiveness of a graded exercise therapy program for patients with chronic shoulder complaints".
Australian Journal of Physiotherapy 51.2 (2005), pp. 87–94.
- [29] N. A. Ratamess, B. A. Alvar, W. B. Kibler, et al.
"Progression models in resistance training for healthy adults" (2009).
- [30] J. M. Davis and S. P. Bailey.
"Possible mechanisms of central nervous system fatigue during exercise."
Medicine and Science in Sports and Exercise 29.1 (1997), pp. 45–57.
- [31] P. D. White, K. A. Goldsmith, A. L. Johnson, et al.
"Comparison of adaptive pacing therapy, cognitive behaviour therapy, graded exercise therapy, and specialist medical care for chronic fatigue syndrome (PACE): a randomised trial".
The Lancet 377.9768 (2011), pp. 823–836.
- [32] C. Wilshire, T. Kindlon, A. Matthees, et al.
"Can patients with chronic fatigue syndrome really recover after graded exercise or cognitive behavioural therapy? A critical commentary and preliminary re-analysis of the PACE trial".
Fatigue: Biomedicine, Health & Behavior 5.1 (2017), pp. 43–56.
- [33] N. I. for Health Care and Excellence. *Myalgic encephalomyelitis (or encephalopathy)/chronic fatigue syndrome: diagnosis and management*. 2021. URL: <https://www.nice.org.uk/guidance/ng206>.
- [34] M. Vink and F. Vink-Niese. "Is it useful to question the recovery behaviour of patients with ME/CFS or Long COVID?" *Healthcare*. Vol. 10. 2. MDPI. 2022, p. 392.
- [35] Ø. Fluge, K. Risa, S. Lunde, et al. "B-lymphocyte depletion in myalgic encephalopathy/chronic fatigue syndrome. An open-label phase II study with rituximab maintenance treatment".
PloS One 10.7 (2015), e0129898.
- [36] Ø. Fluge, I. G. Rekeland, K. Lien, et al. "B-lymphocyte depletion in patients with myalgic encephalomyelitis/chronic fatigue syndrome: a randomized, double-blind, placebo-controlled trial".
Annals of Internal Medicine 170.9 (2019), pp. 585–593.
- [37] L. A. Jason, S. Yoo, and S. Bhatia.
"Patient perceptions of infectious illnesses preceding Myalgic

- Encephalomyelitis/Chronic Fatigue Syndrome".
Chronic Illness (2021), p. 17423953211043106.
- [38] A. Ghali, P. Richa, C. Lacout, et al. "Epidemiological and clinical factors associated with post-exertional malaise severity in patients with myalgic encephalomyelitis/chronic fatigue syndrome".
Journal of Translational Medicine 18.1 (2020), pp. 1–8.
- [39] L. Chu, I. J. Valencia, D. W. Garvert, et al. "Onset patterns and course of myalgic encephalomyelitis/chronic fatigue syndrome".
Frontiers in Pediatrics 7 (2019), p. 12.
- [40] M. A. Evans and L. A. Jason. "Onset patterns of chronic fatigue syndrome and myalgic encephalomyelitis".
Research on Chronic Diseases 2 (2018), pp. 1–30.
- [41] H. Naess, E. Sundal, K.-M. Myhr, et al.
"Postinfectious and chronic fatigue syndromes: clinical experience from a tertiary-referral centre in Norway". *In Vivo* 24.2 (2010), pp. 185–188.
- [42] C. Ray, S. Jeffeues, W. Weir, et al.
"Making sense of chronic fatigue syndrome: patients' accounts of onset".
Psychology and Health 13.1 (1998), pp. 99–109.
- [43] I. E. Salit. "Precipitating factors for the chronic fatigue syndrome".
Journal of Psychiatric Research 31.1 (1997), pp. 59–65.
- [44] E. Shikova, V. Reshkova, n. Kumanova, et al.
"Cytomegalovirus, Epstein-Barr virus, and human herpesvirus-6 infections in patients with myalgic ncephalomyelitis/chronic fatigue syndrome".
Journal of Medical Virology 92.12 (2020), pp. 3682–3688.
- [45] M. E. Ariza. "Myalgic encephalomyelitis/chronic fatigue syndrome: the human herpesviruses are back!" *Biomolecules* 11.2 (2021), p. 185.
- [46] R. J. Whitley.
Medical Microbiology, 4th Edition, Chapter 68 Herpesviruses.
University of Texas Medical Branch at Galveston, 1996.
- [47] S. Rasa, Z. Nora-Kruble, N. Henning, et al. "Chronic viral infections in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)".
Journal of Translational Medicine 16.1 (2018), pp. 1–25.
- [48] K. Bjornevik, M. Cortese, B. C. Healy, et al.
"Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis".
Science 375.6578 (2022), pp. 296–301.

- [49] F. Sotzny, J. Blanco, E. Capelli, et al. "Myalgic encephalomyelitis/chronic fatigue syndrome – evidence for an autoimmune disease". *Autoimmunity Reviews* 17.6 (2018), pp. 601–609.
- [50] J. Castro-Marrero, M. Faro, L. Aliste, et al. "Comorbidity in chronic fatigue syndrome/myalgic encephalomyelitis: a nationwide population-based cohort study". *Psychosomatics* 58.5 (2017), pp. 533–543.
- [51] E. Bellato, E. Marini, F. Castoldi, et al. "Fibromyalgia syndrome: etiology, pathogenesis, diagnosis, and treatment". *Pain research and treatment* 2012 (2012).
- [52] C. Kodner. "Common questions about the diagnosis and management of fibromyalgia". *American Family Physician* 91.7 (2015), pp. 472–478.
- [53] L. A. Jason, R. R. Taylor, C. L. Kennedy, et al. "Chronic fatigue syndrome, fibromyalgia, and multiple chemical sensitivities in a community-based sample of persons with chronic fatigue syndrome-like symptoms". *Psychosomatic Medicine* 62.5 (2000), pp. 655–663.
- [54] S. Shor, C. Green, B. Szantyr, et al. "Chronic Lyme disease: an evidence-based definition by the ILADS Working Group". *Antibiotics* 8.4 (2019), p. 269.
- [55] T. S. Murray and E. D. Shapiro. "Lyme disease". *Clinics in Laboratory Medicine* 30.1 (2010), pp. 311–328.
- [56] A. C. Steere, W. Klitz, E. E. Drouin, et al. "Antibiotic-refractory Lyme arthritis is associated with HLA-DR molecules that bind a *Borrelia burgdorferi* peptide". *The Journal of Experimental Medicine* 203.4 (2006), pp. 961–971.
- [57] S. Strausz, G. Blacker, S. Galloway, et al. "Secretoglobin family 1D member 2 (SCGB1D2) protein inhibits growth of *Borrelia burgdorferi* and affects susceptibility to Lyme disease". *bioRxiv* (2022). DOI: 10.1101/2022.05.27.493784. URL: <https://www.biorxiv.org/content/early/2022/06/02/2022.05.27.493784>.
- [58] H. M. Feder Jr, B. J. Johnson, S. O'Connell, et al. "A critical appraisal of chronic Lyme disease". *New England Journal of Medicine* 357.14 (2007), pp. 1422–1430.
- [59] L. B. Krupp, L. G. Hyman, R. Grimson, et al. "Study and treatment of post Lyme disease (STOP-LD): a randomized double masked clinical trial". *Neurology* 60.12 (2003), pp. 1923–1930.

- [60] S. J. Halpin, C. McIvor, G. Whyatt, et al.
“Postdischarge symptoms and rehabilitation needs in survivors of COVID-19 infection: a cross-sectional evaluation”.
Journal of Medical Virology 93.2 (2021), pp. 1013–1022.
- [61] D. Ayoubkhani and P. Pawelek. *Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK: 1 June 2022*. 2022.
URL: <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19/infectionintheuk/1june2022>.
- [62] G. Reynolds, D. Lewis, A. Richardson, et al.
“Comorbidity of postural orthostatic tachycardia syndrome and chronic fatigue syndrome in an Australian cohort”.
Journal of Internal Medicine 275.4 (2014), pp. 409–417.
- [63] M. Lamarre-Cliche and J. Cusson. “The fainting patient: value of the head-upright tilt-table test in adult patients with orthostatic intolerance”.
Canadian Medical Association Journal 164.3 (2001), pp. 372–376.
- [64] P. A. Low and P. Sandroni.
“Chapter 106 - Postural Tachycardia Syndrome (POTS)”.
Primer on the Autonomic Nervous System (Third Edition).
Ed. by D. Robertson, I. Biaggioni, G. Burnstock, et al.
San Diego: Academic Press, 2012, pp. 517–519.
ISBN: 978-0-12-386525-0.
- [65] P. J. Goadsby, R. B. Lipton, and M. D. Ferrari.
“Migraine – current understanding and treatment”.
New England Journal of Medicine 346.4 (2002), pp. 257–270.
- [66] C.-I. Lau, C.-C. Lin, W.-H. Chen, et al. “Increased risk of chronic fatigue syndrome in patients with migraine: a retrospective cohort study”.
Journal of Psychosomatic Research 79.6 (2015), pp. 514–518.
- [67] M. Peres, E. Zukerman, W. Young, et al.
“Fatigue in chronic migraine patients”.
Cephalalgia 22.9 (2002), pp. 720–724.
- [68] R. Burstein, R. Nosedá, and D. Borsook.
“Migraine: multiple processes, complex pathophysiology”.
Journal of Neuroscience 35.17 (2015), pp. 6619–6629.

- [69] B. J. Horwitz and R. S. Fisher. "The irritable bowel syndrome". *New England Journal of Medicine* 344.24 (2001), pp. 1846–1850.
- [70] D. C. Baumgart and S. R. Carding. "Inflammatory bowel disease: cause and immunobiology". *The Lancet* 369.9573 (2007), pp. 1627–1640.
- [71] W. E. Whitehead, O. Palsson, and K. R. Jones. "Systematic review of the comorbidity of irritable bowel syndrome with other disorders: What are the causes and implications?" *Gastroenterology* 122.4 (2002), pp. 1140–1156.
- [72] R. Spiller and K. Garsed. "Postinfectious irritable bowel syndrome". *Gastroenterology* 136.6 (2009), pp. 1979–1988.
- [73] S. Wessely et al. "Ten years on: what do we know about the Gulf War syndrome?" *Clinical Medicine* 1.1 (2001), p. 28.
- [74] H. K. Kang, B. H. Natelson, C. M. Mahan, et al. "Post-traumatic stress disorder and chronic fatigue syndrome-like illness among Gulf War veterans: a population-based survey of 30,000 veterans". *American Journal of Epidemiology* 157.2 (2003), pp. 141–148.
- [75] R. W. Haley, G. Kramer, J. Xiao, et al. "Evaluation of a gene–environment interaction of PON1 and low-level nerve agent exposure with Gulf War illness: a prevalence case–control study drawn from the US military health surveys national population sample". *Environmental Health Perspectives* 130.5 (2022), p. 057001.
- [76] M. E. Loades, K. A. Rimes, S. Ali, et al. "The presence of co-morbid mental health problems in a cohort of adolescents with chronic fatigue syndrome". *Clinical Child Psychology and Psychiatry* 23.3 (2018), pp. 398–408.
- [77] M. E. Loades, R. Read, L. Smith, et al. "How common are depression and anxiety in adolescents with chronic fatigue syndrome (CFS) and how should we screen for these mental health co-morbidities? A clinical cohort study". *European Child & Adolescent Psychiatry* 30.11 (2021), pp. 1733–1743.
- [78] A. Mariman, L. Delesie, E. Tobback, et al. "Undiagnosed and comorbid disorders in patients with presumed chronic fatigue syndrome". *Journal of Psychosomatic Research* 75.5 (2013), pp. 491–496.
- [79] J. Vercoulen, C. Swanink, J. Galama, et al. "The persistence of fatigue in chronic fatigue syndrome and multiple

- sclerosis: Development of a model".
Journal of Psychosomatic Research 45.6 (1998), pp. 507–517.
- [80] S. B. Harvey and S. Wessely.
"Chronic fatigue syndrome: identifying zebras amongst the horses".
BMC Medicine 7.1 (2009), pp. 1–4.
- [81] M. Maes and F. N. Twisk. "Chronic fatigue syndrome: Harvey and Wessely's (bio) psychosocial model versus a bio (psychosocial) model based on inflammatory and oxidative and nitrosative stress pathways".
BMC Medicine 8.1 (2010), pp. 1–13.
- [82] R. Hatch, D. Young, V. Barber, et al.
"Anxiety, depression and post traumatic stress disorder after critical illness: a UK-wide prospective cohort study".
Critical Care 22.1 (2018), pp. 1–13.
- [83] R. Dobson and G. Giovannoni. "Multiple sclerosis – a review".
European Journal of Neurology 26.1 (2019), pp. 27–40.
- [84] J. A. Hollenbach and J. R. Oksenberg.
"The immunogenetics of multiple sclerosis: a comprehensive review".
Journal of Autoimmunity 64 (2015), pp. 13–25.
- [85] L. E. Mokry, S. Ross, O. S. Ahmad, et al. "Vitamin D and risk of multiple sclerosis: a Mendelian randomization study".
PLoS Medicine 12.8 (2015), e1001866.
- [86] R. A. Hughes and D. R. Cornblath. "Guillain-barré syndrome".
The Lancet 366.9497 (2005), pp. 1653–1666.
- [87] T. A. Hardy, S. Blum, P. A. McCombe, et al.
"Guillain-barré syndrome: modern theories of etiology".
Current allergy and asthma reports 11.3 (2011), pp. 197–204.
- [88] A. Tenesa and C. S. Haley.
"The heritability of human disease: estimation, uses and abuses".
Nature Reviews Genetics 14.2 (2013), pp. 139–149.
- [89] L. Henden, S. Lee, I. Mueller, et al. "Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens".
PLoS genetics 14.5 (2018), e1007279.
- [90] S. R. Browning and B. L. Browning. "Identity by descent between distant relatives: detection and applications".
Annual review of genetics 46 (2012), pp. 617–633.

- [91] M. Lynch and K. Ritland.
“Estimation of pairwise relatedness with molecular markers”.
Genetics 152.4 (1999), pp. 1753–1766.
- [92] N. Zaitlen and P. Kraft.
“Heritability in the genome-wide association era”.
Human genetics 131.10 (2012), pp. 1655–1664.
- [93] A. L. Price, N. J. Patterson, R. M. Plenge, et al. “Principal components analysis corrects for stratification in genome-wide association studies”.
Nature Genetics 38.8 (2006), pp. 904–909.
- [94] A. L. Price, N. A. Zaitlen, D. Reich, et al. “New approaches to population stratification in genome-wide association studies”.
Nature reviews genetics 11.7 (2010), pp. 459–463.
- [95] R. Walters, C. Churchouse, and R. Hosking.
Heritability 201: types of heritability and how we estimate it. 2017.
URL: <http://www.nealelab.is/blog/2017/9/13/heritability-201-types-of-heritability-and-how-we-estimate-it>.
- [96] D. Falconer. “Introduction to quantitative genetics” (1985).
- [97] J. Arbet, M. McGue, and S. Basu.
“A robust and unified framework for estimating heritability in twin studies using generalized estimating equations”.
Statistics in medicine 39.27 (2020), pp. 3897–3913.
- [98] J. Yang, B. Benyamin, B. P. McEvoy, et al. “Common SNPs explain a large proportion of the heritability for human height”.
Nature genetics 42.7 (2010), pp. 565–569.
- [99] C. M. Lakhani, B. T. Tierney, A. K. Manrai, et al.
“Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes”.
Nature Genetics 51.2 (2019), pp. 327–334.
- [100] O. Canela-Xandri, K. Rawlik, and A. Tenesa.
“An atlas of genetic associations in UK Biobank”.
Nature Genetics 50.11 (2018), pp. 1593–1599.
- [101] P. F. Sullivan, B. Evengård, A. Jacks, et al.
“Twin analyses of chronic fatigue in a Swedish national sample”.
Psychological Medicine 35.9 (2005), pp. 1327–1336.
- [102] W. Wang, B. J. Barratt, D. G. Clayton, et al.
“Genome-wide association studies: theoretical and practical concerns”.
Nature Reviews Genetics 6.2 (2005), pp. 109–118.

- [103] P. C. Ng and E. F. Kirkness. "Whole genome sequencing". *Genetic Variation* (2010), pp. 215–226.
- [104] R. C. Lewontin. "The interaction of selection and linkage. I. General considerations; heterotic models". *Genetics* 49.1 (1964), p. 49.
- [105] J. T. Kang and N. A. Rosenberg.
"Mathematical properties of linkage disequilibrium statistics defined by normalization of the coefficient $d = p_{ab} - p_a p_b$ ".
Human Heredity 84.3 (2019), pp. 127–143.
- [106] J. Fadista, A. K. Manning, J. C. Florez, et al. "The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants". *European Journal of Human Genetics* 24.8 (2016), pp. 1202–1205.
- [107] S. Nguengang Wakap, D. M. Lambert, A. Olry, et al.
"Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database".
European Journal of Human Genetics 28.2 (2020), pp. 165–173.
- [108] E. S. Lander. "The new genomics: global views of biology". *Science* 274.5287 (1996), pp. 536–539.
- [109] D. J. Smith and A. J. Lusk. "The allelic structure of common disease". *Human Molecular Genetics* 11.20 (Oct. 2002), pp. 2455–2461.
- [110] Q. Wang, R. S. Dhindsa, K. Carss, et al. "Rare variant contribution to human disease in 281,104 UK Biobank exomes". *Nature* 597.7877 (2021), pp. 527–532.
- [111] J. M. Bland and D. G. Altman. "The odds ratio". *British Medical Journal* 320.7247 (2000), p. 1468.
- [112] L. R. Lloyd-Jones, M. R. Robinson, J. Yang, et al.
"Transformation of summary statistics from linear mixed model association on all-or-none traits to odds ratio".
Genetics 208.4 (2018), pp. 1397–1408.
- [113] C. O. Schmidt and T. Kohlmann.
"When to use the odds ratio or the relative risk?"
International Journal of Public Health 53.3 (2008), p. 165.
- [114] D. Altshuler, P. Donnelly, International HapMap Consortium, et al.
"A haplotype map of the human genome".
Nature 437.7063 (2005), nature04226.
- [115] P. Duggal, E. M. Gillanders, T. N. Holmes, et al.
"Establishing an adjusted p-value threshold to control the family-wide

- type 1 error in genome wide association studies".
BMC Genomics 9.1 (2008), pp. 1–8.
- [116] J. C. Denny, M. D. Ritchie, M. A. Basford, et al.
"PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations".
Bioinformatics 26.9 (Mar. 2010), pp. 1205–1210.
- [117] C. Sudlow, J. Gallacher, N. Allen, et al.
"UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age".
PLoS Medicine 12.3 (2015), e1001779.
- [118] R. Hajdarevic, A. Lande, J. Mehlsen, et al.
"Genetic association study in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) identifies several potential risk loci".
Brain, Behavior, and Immunity 102 (2022), pp. 362–369.
- [119] Y. Tanigawa, J. Li, J. M. Justesen, et al.
"Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology".
Nature Communications 10.1 (2019), pp. 1–14.
- [120] B. Neale et al. *UK Biobank GWAS*. 2018.
URL: <http://www.nealelab.is/uk-biobank/>.
- [121] W. Zhou, J. B. Nielsen, L. G. Fritsche, et al.
"Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies".
Nature Genetics 50.9 (2018), pp. 1335–1341.
- [122] Pan-UKB team. *Pan-UK Biobank*. 2020.
URL: <https://pan.ukbb.broadinstitute.org>.
- [123] W.-Q. Wei, L. A. Bastarache, R. J. Carroll, et al. "Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record".
PloS One 12.7 (2017), e0175508.
- [124] Y. Jiang, S. Chen, X. Wang, et al. "Association analysis and meta-analysis of multi-allelic variants for large-scale sequence data".
Genes 11.5 (2020), p. 586.
- [125] J. D. Backman, A. H. Li, A. Marcketta, et al.
"Exome sequencing and analysis of 454,787 UK Biobank participants".
Nature 599.7886 (2021), pp. 628–634.

- [126] M. Schipper and D. Posthuma. “Demystifying non-coding GWAS variants: an overview of computational tools and methods”. *Human Molecular Genetics* 31.R1 (2022), R73–R83.
- [127] A. Gusev, A. Ko, H. Shi, et al. “Integrative approaches for large-scale transcriptome-wide association studies”. *Nature Genetics* 48.3 (2016), pp. 245–252.
- [128] E. R. Gamazon, A. V. Segrè, M. Van De Bunt, et al. “Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation”. *Nature genetics* 50.7 (2018), pp. 956–967.
- [129] D. A. Kleinjan and V. van Heyningen. “Long-range control of gene expression: emerging mechanisms and disruption in disease”. *The American Journal of Human Genetics* 76.1 (2005), pp. 8–32.
- [130] E. Yamano, M. Sugimoto, A. Hirayama, et al. “Index markers of chronic fatigue syndrome with dysfunction of TCA and urea cycles”. *Scientific Reports* 6.1 (2016), pp. 1–9.
- [131] R. K. Naviaux, J. C. Naviaux, K. Li, et al. “Metabolic features of chronic fatigue syndrome”. *Proceedings of the National Academy of Sciences* 113.37 (2016), E5472–E5480.
- [132] C. W. Armstrong, N. R. McGregor, J. R. Sheedy, et al. “NMR metabolic profiling of serum identifies amino acid disturbances in chronic fatigue syndrome”. *International Journal of Clinical Chemistry* 413.19-20 (2012), pp. 1525–1531.
- [133] M. Aguirre, M. A. Rivas, and J. Priest. “Phenome-wide burden of copy-number variation in the UK biobank”. *The American Journal of Human Genetics* 105.2 (2019), pp. 373–383.
- [134] J. L. Freeman, G. H. Perry, L. Feuk, et al. “Copy number variation: new insights in genome diversity”. *Genome Research* 16.8 (2006), pp. 949–961.
- [135] A. K. Smith, H. Fang, T. Whistler, et al. “Convergent genomic studies identify association of GRIK2 and NPAS2 with chronic fatigue syndrome”. *Neuropsychobiology* 64.4 (2011), pp. 183–194.
- [136] K. Schlauch, S. F. Khaiboullina, K. L. De Meirleir, et al. “Genome-wide association analysis identifies genetic variations in subjects with myalgic encephalomyelitis/chronic fatigue syndrome”. *Translational Psychiatry* 6.2 (2016), e730–e730.

- [137] A. Buniello, J. A. L. MacArthur, M. Cerezo, et al.
“The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019”.
Nucleic Acids Research 47.D1 (2019), pp. D1005–D1012.
- [138] T. Wang, J. Yin, A. H. Miller, et al. “A systematic review of the association between fatigue and genetic polymorphisms”.
Brain, Behavior, and Immunity 62 (2017), pp. 230–244.
- [139] S. Marshall-Gradisnik, S. Johnston, A. Chacko, et al. “Single nucleotide polymorphisms and genotypes of transient receptor potential ion channel and acetylcholine receptor genes from isolated B lymphocytes in myalgic encephalomyelitis/chronic fatigue syndrome patients”.
Journal of International Medical Research 44.6 (2016), pp. 1381–1394.
- [140] M. Perez, R. Jaundoo, K. Hilton, et al. “Genetic predisposition for immune system, hormone, and metabolic dysfunction in myalgic encephalomyelitis/chronic fatigue syndrome: a pilot study”.
Frontiers in Pediatrics (2019), p. 206.
- [141] M. T. Maurano, R. Humbert, E. Rynes, et al. “Systematic localization of common disease-associated variation in regulatory DNA”.
Science 337.6099 (2012), pp. 1190–1195.
- [142] P. Sebastiani, N. Solovieff, A. T. DeWan, et al.
“Genetic signatures of exceptional longevity in humans”.
PloS One 7.1 (2012), e29848.
- [143] M. Claussnitzer, J. H. Cho, R. Collins, et al.
“A brief history of human disease genetics”.
Nature 577.7789 (2020), pp. 179–189.
- [144] B. N. Goertzel, C. Pennachin, L. de Souza Coelho, et al.
“Combinations of single nucleotide polymorphisms in neuroendocrine effector and receptor genes predict chronic fatigue syndrome” (2006).
- [145] G. D. Kitsios and E. Zintzaras.
“Genome-wide association studies: hypothesis-free or engaged?”
Translational Research 154.4 (2009), pp. 161–164.
- [146] A. Lande, Ø. Fluge, E. B. Strand, et al.
“Human leukocyte antigen alleles associated with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)”.
Scientific Reports 10.1 (2020), pp. 1–8.

- [147] S. Gough and M. Simmonds. "The HLA region and autoimmune disease: associations and mechanisms of action".
Current Genomics 8.7 (2007), pp. 453–465.
- [148] B. Alberts, A. Johnson, J. Lewis, et al.
Molecular Biology of the Cell. 4th edition. Garland Science, 2002.
- [149] S. G. Marsh, E. Albert, W. Bodmer, et al.
"Nomenclature for factors of the HLA system, 2010".
Tissue Antigens 75.4 (2010), p. 291.
- [150] C. K. Hurley. "Naming HLA diversity: a review of HLA nomenclature".
Human Immunology 82.7 (2021). Defining and Characterizing HLA Diversity, pp. 457–465.
- [151] J. L. Rivas, T. Palencia, G. Fernández, et al.
"Association of T and NK cell phenotype with the diagnosis of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)".
Frontiers in Immunology 9 (2018).
- [152] E. W. Brenu, M. L. Van Driel, D. R. Staines, et al.
"Longitudinal investigation of natural killer cells and cytokines in chronic fatigue syndrome/myalgic encephalomyelitis".
Journal of Translational Medicine 10.1 (2012), pp. 1–11.
- [153] W. Sato, H. Ono, T. Matsutani, et al. "Skewing of the B cell receptor repertoire in myalgic encephalomyelitis/chronic fatigue syndrome".
Brain, Behavior, and Immunity 95 (2021), pp. 245–255.
- [154] M. Milivojevic, X. Che, L. Bateman, et al.
"Plasma proteomic profiling suggests an association between antigen driven clonal B cell expansion and ME/CFS".
PloS One 15.7 (2020), e0236148.
- [155] J. M. Cliff, E. C. King, J.-S. Lee, et al. "Cellular immune function in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)".
Frontiers in Immunology 10 (2019).
- [156] M. Davis. *Is CFS/ME an autoimmune disease? [Presentation]*. 2017.
URL: <https://www.youtube.com/watch?v=SbnCMOHOL28>.
- [157] J. G. Montoya, T. H. Holmes, J. N. Anderson, et al. "Cytokine signature associated with disease severity in chronic fatigue syndrome patients".
Proceedings of the National Academy of Sciences 114.34 (2017), E7150–E7158.

- [158] L. Fugger, L. T. Jensen, and J. Rossjohn. “Challenges, progress, and prospects of developing therapies to treat autoimmune diseases”. *Cell* 181.1 (2020), pp. 63–80.
- [159] D. Martinelli, D. Diodato, E. Ponzi, et al. “The hyperornithinemia–hyperammonemia–homocitrullinuria syndrome”. *Orphanet Journal of Rare Diseases* 10.1 (2015), pp. 1–17.
- [160] K. Tezcan, K. T. Louie, Y. Qu, et al. “Adult-onset presentation of a hyperornithinemia-hyperammonemia-homocitrullinuria patient without prior history of neurological complications”. *JIMD Reports – Case and Research Reports*. Springer, 2011, pp. 97–102.
- [161] M. Ghoussaini, E. Mountjoy, M. Carmona, et al. “Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics”. *Nucleic Acids Research* 49.D1 (2021), pp. D1311–D1320.
- [162] UK Biobank. *Verbal Interview stage, Version 1.1*. 2012. URL: <https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/Interview.pdf>.
- [163] UK Biobank. *Data-Field 20002*. 2022. URL: <https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20002>.
- [164] UK Biobank. *Pain Web Questionnaire*. 2022. URL: https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/pain_questionnaire.pdf.
- [165] UK Biobank. *Data-Field 120128*. 2022. URL: <https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=120128>.
- [166] UK Biobank. *Data-Field 120010*. 2022. URL: <https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=120010>.
- [167] J. R. Conway, A. Lex, and N. Gehlenborg. “UpSetR: an R package for the visualization of intersecting sets and their properties”. *Bioinformatics* 33.18 (June 2017), pp. 2938–2940.
- [168] D. Skuse, F. Printzlau, and J. Wolstencroft. “Chapter 24 - Sex chromosome aneuploidies”. *Neurogenetics, Part I*. Ed. by D. H. Geschwind, H. L. Paulson, and C. Klein. Vol. 147. Handbook of Clinical Neurology. Elsevier, 2018, pp. 355–376.
- [169] C. Bycroft, C. Freeman, D. Petkova, et al. “The UK Biobank resource with deep phenotyping and genomic data”. *Nature* 562.7726 (2018), pp. 203–209.
- [170] S. Purcell, B. Neale, K. Todd-Brown, et al. “PLINK: a tool set for whole-genome association and population-based linkage analyses”. *The American Journal of Human Genetics* 81.3 (2007), pp. 559–575.

- [171] C. C. Chang, C. C. Chow, L. C. Tellier, et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". *Gigascience* 4.1 (2015), s13742–015.
- [172] S. Purcell and C. Chang. *PLINK 2.00*. 2015.
URL: www.cog-genomics.org/plink/2.0/.
- [173] B. Neale et al. *Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK biobank*. 2017. URL: <http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank>.
- [174] C. D. Bustamante, F. M. De La Vega, and E. G. Burchard. "Genomics for the world". *Nature* 475.7355 (2011), pp. 163–165.
- [175] H. Wickham. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4.
URL: <https://ggplot2.tidyverse.org>.
- [176] N. Crowley, M. Nelson, and S. Stovin. "Epidemiological aspects of an outbreak of encephalomyelitis at the Royal Free Hospital, London, in the summer of 1955". *Epidemiology & Infection* 55.1 (1957), pp. 102–122.
- [177] A. Dusa. *venn 1.11*. 2022.
URL: <https://CRAN.R-project.org/package=venn>.
- [178] P. Jaccard. "The distribution of the flora in the alpine zone. 1". *New Phytologist* 11.2 (1912), pp. 37–50.
- [179] A. Banerjee, U. Chitnis, S. Jadhav, et al. "Hypothesis testing, type I and type II errors". *Industrial Psychiatry Journal* 18.2 (2009), p. 127.
- [180] C. C. Spencer, Z. Su, P. Donnelly, et al. "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip". *PLoS Genetics* 5.5 (2009), e1000477.
- [181] S. Purcell, S. S. Cherny, and P. C. Sham. "Genetic power calculator: design of linkage and association genetic mapping studies of complex traits". *Bioinformatics* 19.1 (2003), pp. 149–150.
- [182] W. S. Bush and J. H. Moore. "Chapter 11: Genome-wide association studies". *PLoS Computational Biology* 8.12 (2012), e1002822.

- [183] J. E. Wigginton, D. J. Cutler, and G. R. Abecasis.
“A note on exact tests of Hardy-Weinberg equilibrium”.
The American Journal of Human Genetics 76.5 (2005), pp. 887–893.
- [184] A. Hill, P.-R. Loh, R. B. Bharadwaj, et al.
“Stepwise distributed open innovation contests for software development:
acceleration of genome-wide association analysis”.
Giga Science 6.5 (2017), gix009.
- [185] C. Chang. *PLINK2: Association Analysis*. 2022.
URL: <https://www.cog-genomics.org/plink/2.0/>.
- [186] G. Tripepi, K. Jager, F. Dekker, et al.
“Linear and logistic regression analysis”.
Kidney International 73.7 (2008), pp. 806–810.
- [187] H. H. Rashidi, N. K. Tran, E. V. Betts, et al.
“Artificial intelligence and machine learning in pathology: the present
landscape of supervised methods”.
Academic pathology 6 (2019), p. 2374289519873088.
- [188] E. Vittinghoff and C. E. McCulloch. “Relaxing the rule of ten events per
variable in logistic and Cox regression”.
American journal of epidemiology 165.6 (2007), pp. 710–718.
- [189] R. C. Team et al.
“R: a language and environment for statistical computing” (2013).
- [190] R. A. Stine. “Graphical interpretation of variance inflation factors”.
The American Statistician 49.1 (1995), pp. 53–56.
- [191] A. Hughes and D. Grawoig. *Statistics; a foundation for analysis*. 1971.
- [192] A. Alin. “Multicollinearity”. *Wiley Interdisciplinary Reviews:
Computational Statistics* 2.3 (2010), pp. 370–374.
- [193] R. M. O'Brien.
“A caution regarding rules of thumb for variance inflation factors”.
Quality & Quantity 41.5 (2007), pp. 673–690.
- [194] J. D. Curto and J. C. Pinto. “The corrected VIF (CVIF)”.
Journal of Applied Statistics 38.7 (2011), pp. 1499–1507.
- [195] O. Mayo. “A century of Hardy-Weinberg equilibrium”.
Twin Research and Human Genetics 11.3 (2008), pp. 249–256.
- [196] B. Devlin and K. Roeder. “Genomic control for association studies”.
Biometrics 55.4 (1999), pp. 997–1004.

- [197] S.-A. Bacanu, B. Devlin, and K. Roeder.
“The power of genomic control”.
The American Journal of Human Genetics 66.6 (2000), pp. 1933–1944.
- [198] E. Mountjoy, E. M. Schmidt, M. Carmona, et al.
“An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci”.
Nature Genetics 53.11 (2021), pp. 1527–1533.
- [199] M. I. Kurki, J. Karjalainen, P. Palta, et al.
“FinnGen: Unique genetic insights from combining isolated population and national health register data”. *medRxiv* (2022).
- [200] S. Turner. “qqman: an R package for visualizing GWAS results using Q-Q and Manhattan plots”. *The Journal of Open Source Software* (2018).
- [201] A. P. Boughton, R. P. Welch, M. Flickinger, et al.
“LocusZoom.js: Interactive and embeddable visualization of genetic association study results”. *Bioinformatics* 37.18 (2021), pp. 3017–3018.
- [202] K. J. Karczewski, L. C. Francioli, G. Tiao, et al. “The mutational constraint spectrum quantified from variation in 141,456 humans”.
Nature 581.7809 (2020), pp. 434–443.
- [203] M. F. Seldin, R. Shigeta, P. Villoslada, et al. “European population substructure: clustering of northern and southern populations”.
PLoS genetics 2.9 (2006), e143.
- [204] J. Graffelman and V. Moreno.
“The mid p-value in exact tests for Hardy-Weinberg equilibrium”.
Statistical Applications in Genetics and Molecular Biology 12.4 (2013), pp. 433–448.
- [205] G. S. Tejeda, E. L. Whiteley, T. Z. Deeb, et al.
“Chorea-related mutations in PDE10A result in aberrant compartmentalization and functionality of the enzyme”. *Proceedings of the National Academy of Sciences* 117.1 (2020), pp. 677–688.
- [206] F. S. Menniti, T. A. Chappie, and C. J. Schmidt. “PDE10A inhibitors – clinical failure or window into antipsychotic drug action?”
Frontiers in Neuroscience (2021), p. 1442.
- [207] S. Fishilevich, R. Nudel, N. Rappaport, et al. “GeneHancer: genome-wide integration of enhancers and target genes in GeneCards”.
Database 2017 (2017).

- [208] W. J. Kent, C. W. Sugnet, T. S. Furey, et al.
“The human genome browser at UCSC”.
Genome research 12.6 (2002), pp. 996–1006.
- [209] M. K. Hankir, M. Kranz, T. Gnad, et al. “A novel thermoregulatory role for PDE10A in mouse and human adipocytes”.
EMBO Molecular Medicine 8.7 (2016), pp. 796–812.
- [210] A. M. Graybiel. “The basal ganglia”.
Current Biology 10.14 (2000), R509–R511.
- [211] N. E. Mencacci, E.-J. Kamsteeg, K. Nakashima, et al.
“De novo mutations in PDE10A cause childhood-onset chorea with bilateral striatal lesions”.
The American Journal of Human Genetics 98.4 (2016), pp. 763–771.
- [212] E. Porcu, M. Medici, G. Pistis, et al.
“A meta-analysis of thyroid-related traits reveals novel loci and gender-specific differences in the regulation of thyroid function”.
PLoS Genetics 9.2 (2013), e1003266.
- [213] T. A. Alexander and M. J. Machiela. “LDpop: an interactive online tool to calculate and visualize geographic LD patterns”.
BMC Bioinformatics 21.1 (2020), pp. 1–4.
- [214] H. R. Warren, E. Evangelou, C. P. Cabrera, et al.
“Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk”.
Nature Genetics 49.3 (2017), pp. 403–415.
- [215] 23andMe. *23andMe Research Innovation Collaborations Program*. 2022.
URL: <https://research.23andme.com/research-innovation-collaborations/>.
- [216] Y. Xu, P. C. Miller, C. K. Phoon, et al. “LPGAT1 controls the stearate/palmitate ratio of phosphatidylethanolamine and phosphatidylcholine in sn-1 specific remodeling”.
Journal of Biological Chemistry 298.3 (2022).
- [217] X. Zhang, J. Zhang, H. Sun, et al.
“Defective phosphatidylglycerol remodeling causes hepatopathy, linking mitochondrial dysfunction to hepatosteatosis”. *Cellular and Molecular Gastroenterology and Hepatology* 7.4 (2019), pp. 763–781.
- [218] H. M. Dönertaş, D. K. Fabian, M. Fuentealba, et al.
“Common genetic associations between age-related diseases”.
Nature Aging 1.4 (2021), pp. 400–412.

- [219] N. Mühlenbein, S. Hofmann, U. Rothbauer, et al.
“Organization and function of the small Tim complexes acting along the import pathway of metabolite carriers into mammalian mitochondria”.
Journal of Biological Chemistry 279.14 (2004), pp. 13540–13546.
- [220] J. Furukawa, K. Inoue, J. Maeda, et al.
“Functional identification of SLC43A3 as an equilibrative nucleobase transporter involved in purine salvage in mammals”.
Scientific Reports 5.1 (2015), pp. 1–11.
- [221] K. B. Hasbargen, W.-J. Shen, Y. Zhang, et al.
“SLC43A3 is a regulator of free fatty acid flux1”.
Journal of Lipid Research 61.5 (2020), pp. 734–745.
- [222] L. A. Millard, N. M. Davies, T. R. Gaunt, et al.
“Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank”.
International journal of epidemiology (2017).
- [223] K. E. Bradbury, T. Y. Tong, and T. J. Key. “Dietary intake of high-protein foods and other major foods in meat-eaters, poultry-eaters, fish-eaters, vegetarians, and vegans in UK Biobank”.
Nutrients 9.12 (2017), p. 1317.
- [224] L. Steell, A. Sillars, P. Welsh, et al.
“Associations of dietary protein intake with bone mineral density: an observational study in 70,215 UK Biobank participants”.
Bone 120 (2019), pp. 38–43.
- [225] T. K. Kim and J. H. Park.
“More about the basic assumptions of t-test: normality and sample size”.
Korean journal of anesthesiology 72.4 (2019), pp. 331–335.
- [226] M. Wang and M. H. Claesson.
“Classification of human leukocyte antigen (HLA) supertypes”.
Immunoinformatics. Springer, 2014, pp. 309–317.
- [227] UK Biobank.
Imputation of classical HLA types from UK Biobank genotype data. 2016.
URL: https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/HLA_imputation.pdf.
- [228] M. Simmonds and S. Gough. “The HLA region and autoimmune disease: associations and mechanisms of action”.
Current Genomics 8.7 (2007), pp. 453–465.

- [229] J. Ablin, L. Neumann, and D. Buskila.
“Pathogenesis of fibromyalgia – a review” .
Joint Bone Spine 75.3 (2008), pp. 273–279.
- [230] B. Alberts, A. Johnson, J. Lewis, et al.
Molecular Biology of the Cell, 6th edition. Garland Science, 2014.
ISBN: 9780815344643.
- [231] A. K. Sewell. “Why must T cells be cross-reactive?”
Nature Reviews Immunology 12.9 (2012), pp. 669–677.
- [232] S. Tonegawa. “Somatic generation of antibody diversity” .
Nature 302.5909 (1983), pp. 575–581.
- [233] M.-P. Lefranc and G. Lefranc. *The T cell receptor FactsBook*.
Elsevier, 2001.
- [234] C. M. Steffens, G. Marchetti, A. Landay, et al. “The human thymus: a
new perspective on thymic function, aging, and HIV infection” .
Clinical Immunology Newsletter 19.6 (1999), pp. 65–79.
- [235] S. J. Corbett, I. M. Tomlinson, E. L. Sonnenhammer, et al.
“Sequence of the human immunoglobulin diversity (D) segment locus: a
systematic analysis provides no evidence for the use of DIR segments,
inverted D segments, “minor” D segments or D-D recombination.”
Journal of Molecular Biology 270.4 (1997), pp. 587–597.
- [236] C. Janeway Jr, P Travers, M Walport, et al.
Immunobiology: the immune system in health and disease. 5th edition.
Garland Science, 2001.
- [237] D. G. Schatz, M. A. Oettinger, and M. S. Schlissel.
“V(D)J recombination: molecular biology and regulation” . *Annual Review
of Immunology* 10.1 (1992). PMID: 1590991, pp. 359–383.
DOI: 10.1146/annurev.iy.10.040192.002043.
- [238] G. A. Rempala and M. Seweryn. “Methods for diversity and overlap
analysis in T-cell receptor populations” .
Journal of Mathematical Biology 67.6 (2013), pp. 1339–1368.
- [239] Q. Qi, Y. Liu, Y. Cheng, et al.
“Diversity and clonal selection in the human T-cell repertoire” .
Proceedings of the National Academy of Sciences 111.36 (2014),
pp. 13139–13144.
- [240] K. Buchmann. “Evolution of innate immunity: clues from invertebrates
via fish to mammals” . *Frontiers in Immunology* 5 (2014).

- [241] C. M. Steffens, L. Al-Harthi, S. Shott, et al. "Evaluation of thymopoiesis using T cell receptor excision circles (TRECs): differential correlation between adult and pediatric TRECs and naïve phenotypes". *Clinical Immunology* 97.2 (2000), pp. 95–101.
- [242] C. H. Bassing, W. Swat, and F. W. Alt. "The mechanism and regulation of chromosomal V(D)J recombination". *Cell* 109.2, Supplement 1 (2002), S45–S55.
- [243] D. G. Schatz and Y. Ji. "Recombination centres and the orchestration of V(D)J recombination". *Nature Reviews Immunology* 11.4 (2011), pp. 251–263.
- [244] N. J. Parkinson, M. Roddis, B. Ferneyhough, et al. "Violation of the 12/23 rule of genomic V(D)J recombination is common in lymphocytes". *Genome Research* 25.2 (2015), pp. 226–234.
- [245] I. H. Choi, Y. J. Chwae, W. S. Shim, et al. "Clonal expansion of CD8+ T cells in Kawasaki disease." *The Journal of Immunology* 159.1 (1997), pp. 481–486.
- [246] H. Mattoo, V. S. Mahajan, T. Maehara, et al. "Clonal expansion of CD4+ cytotoxic T lymphocytes in patients with IgG4-related disease". *Journal of Allergy and Clinical Immunology* 138.3 (2016), pp. 825–838.
- [247] S. Levy-Mendelovich, A. Lev, E. Rechavi, et al. "T and B cell clonal expansion in Ras-associated lymphoproliferative disease (RALD) as revealed by next-generation sequencing". *Clinical & Experimental Immunology* 189.3 (2017), pp. 310–317.
- [248] G. C. Wang, P. Dash, J. A. McCullers, et al. "T cell receptor $\alpha\beta$ diversity inversely correlates with pathogen-specific antibody levels in human cytomegalovirus infection". *Science translational medicine* 4.128 (2012), 128ra42–128ra42.
- [249] T. H. Nguyen, L. C. Rowntree, D. G. Pellicci, et al. "Recognition of distinct cross-reactive virus-specific CD8+ T cells reveals a unique TCR signature in a clinical setting". *The Journal of Immunology* 192.11 (2014), pp. 5039–5049.
- [250] R. Planas, I. Metz, R. Martin, et al. "Detailed characterization of T cell receptor repertoires in multiple sclerosis brain lesions". *Frontiers in Immunology* 9 (2018), p. 509.
- [251] K Usuku, N Joshi, C. Hatem Jr, et al. "Biased expression of T cell receptor genes characterizes activated T cells

- in multiple sclerosis cerebrospinal fluid".
Journal of Neuroscience Research 45.6 (1996), pp. 829–837.
- [252] R. Amoriello, M. Chernigovskaya, V. Greiff, et al. "TCR repertoire diversity in Multiple Sclerosis: high-dimensional bioinformatics analysis of sequences from brain, cerebrospinal fluid and peripheral blood".
EBioMedicine 68 (2021), p. 103429.
- [253] E. M. Lacerda, K. Mudie, C. C. Kingdon, et al. "The UK ME/CFS biobank: a disease-specific biobank for advancing clinical research into myalgic encephalomyelitis/chronic fatigue syndrome".
Frontiers in Neurology 9 (2018), p. 1026.
- [254] *Genohub*. URL: <https://genohub.com/>.
- [255] B Schmitz, A Radbruch, T Kümmel, et al. "Magnetic activated cell sorting (MACS) – a new immunomagnetic method for megakaryocytic cell isolation: comparison of different separation techniques".
European Journal of Haematology 52.5 (1994), pp. 267–275.
- [256] W. Bonner, H. Hulett, R. Sweet, et al.
"Fluorescence activated cell sorting".
Review of Scientific Instruments 43.3 (1972), pp. 404–409.
- [257] S. Afzal, I. Gil-Farina, R. Gabriel, et al. "Systematic comparative study of computational methods for T-cell receptor sequencing data analysis".
Briefings in bioinformatics 20.1 (2019), pp. 222–234.
- [258] M. Shugay, O. V. Britanova, E. M. Merzlyak, et al.
"Towards error-free profiling of immune repertoires".
Nature Methods 11.6 (2014), pp. 653–655.
- [259] D. A. Bolotin, S. Poslavsky, I. Mitrophanov, et al.
"MiXCR: software for comprehensive adaptive immunity profiling".
Nature Methods 12.5 (2015), pp. 380–381.
- [260] B. Gerritsen, A. Pandit, A. C. Andeweg, et al.
"RTCR: a pipeline for complete and accurate recovery of T cell repertoires from high throughput sequencing data".
Bioinformatics 32.20 (2016), pp. 3098–3106.
- [261] A. Rényi. "On measures of entropy and information". *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Vol. 4. University of California Press. 1961, pp. 547–562.
- [262] C. E. Shannon. "A mathematical theory of communication".
The Bell System Technical Journal 27.3 (1948), pp. 379–423.

- [263] E. H. Simpson. "Measurement of diversity".
Nature 163.4148 (1949), pp. 688–688.
- [264] R. V. Alatalo. "Problems in the measurement of evenness in ecology".
Oikos 37.2 (1981), pp. 199–204.
- [265] D. R. Clark, R. J. de Boer, K. C. Wolthers, et al.
"T cell dynamics in HIV-1 infection".
Advances in Immunology 73 (1999), pp. 301–327.
- [266] R. K. Peet. "The Measurement of Species Diversity".
Annual Review of Ecology and Systematics 5 (1974), pp. 285–307.
- [267] A. Chao and T.-J. Shen. "Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample".
Environmental and Ecological Statistics 10.4 (2003), pp. 429–443.
- [268] P. Bromiley, N. Thacker, and E. Bouhova-Thacker.
"Shannon entropy, Rényi entropy, and information".
Statistics and Infinite Series 9 (2004), pp. 10–42.
- [269] V. Greiff, P. Bhat, S. C. Cook, et al.
"A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status".
Genome Medicine 7.1 (2015), pp. 1–15.
- [270] B. Tóthmérész. "Comparison of different methods for diversity ordering".
Journal of Vegetation Science 6.2 (1995), pp. 283–290.
- [271] *American Type Culture Collection (ATCC)*.
URL: <https://www.atcc.org/>.
- [272] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. "On the surprising behavior of distance metrics in high dimensional space".
International conference on database theory. Springer. 2001, pp. 420–434.
- [273] B. Smith and J. B. Wilson. "A consumer's guide to evenness indices".
Oikos 76.1 (1996), pp. 70–82.
- [274] L. Jost. "The relation between evenness and diversity".
Diversity 2.2 (2010), pp. 207–232.
- [275] "Principal coordinate analysis and non-metric multidimensional scaling".
Analysing Ecological Data. New York, NY: Springer New York, 2007, pp. 259–264. ISBN: 978-0-387-45972-1.
- [276] I. Borg and P. J. Groenen.
Modern multidimensional scaling: Theory and applications. Springer Science & Business Media, 2005.

- [277] D. Bzdok, M. Krzywinski, and N. Altman.
“Machine learning: supervised methods” .
Nature methods 15.1 (2018), p. 5.
- [278] W. S. Noble. “What is a support vector machine?”
Nature Biotechnology 24.12 (2006), pp. 1565–1567.
- [279] S. Hochreiter and K. Obermayer.
“Support vector machines for dyadic data” .
Neural Computation 18.6 (2006), pp. 1472–1510.
- [280] T. Knebel and S. Hochreiter.
Software documentation of the potential support vector machine. 2014.
URL: <http://ni.cs.tu-berlin.de/software/psvm>.
- [281] C. Nadeau and Y. Bengio. “Inference for the generalization error” .
Machine Learning 52 (239-281 2003).
- [282] G. James, D. Witten, T. Hastie, et al. *An introduction to statistical learning: with applications in R, second edition*. Springer, 2021.
- [283] M. Ojala and G. C. Garriga.
“Permutation tests for studying classifier performance.”
Journal of Machine Learning Research 11.6 (2010).
- [284] T. D. V. Swinscow et al. *Statistics at square one*.
- [285] W. S. Noble. “How does multiple testing correction work?”
Nature Biotechnology 27.12 (2009), pp. 1135–1137.
- [286] J. Zhang, K. Kobert, T. Flouri, et al.
“PEAR: a fast and accurate Illumina Paired-End reAd mergeR” .
Bioinformatics 30.5 (2014), pp. 614–620.
- [287] M. A. Blackman and D. L. Woodland.
“The narrowing of the CD8 T cell repertoire in old age” .
Current Opinion in Immunology 23.4 (2011), pp. 537–542.
- [288] D. B. Palmer. “The effect of age on thymic function” .
Frontiers in Immunology 4 (2013), p. 316.
- [289] P. C. Mahalanobis. “On the generalized distance in statistics” .
2.1 (1936).
- [290] J. J. Goronzy, W.-W. Lee, and C. M. Weyand.
“Aging and T-cell diversity” .
Experimental Gerontology 42.5 (2007), pp. 400–406.
- [291] K. Joshi, M. R. de Massy, M. Ismail, et al.
“Spatial heterogeneity of the T cell receptor repertoire reflects the

- mutational landscape in lung cancer".
Nature Medicine 25.10 (2019), pp. 1549–1559.
- [292] P. C. de Greef, T. Oakes, B. Gerritsen, et al. "The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes".
Elife 9 (2020), e49900.
- [293] S. J. Turner, P. C. Doherty, J. McCluskey, et al.
"Structural determinants of T-cell receptor bias in immunity".
Nature reviews immunology 6.12 (2006), pp. 883–894.
- [294] A. Zeileis, C. Kleiber, and S. Jackman.
"Regression models for count data in R".
Journal of statistical software 27.8 (2008), pp. 1–25.
- [295] J. M. Ver Hoef and P. L. Boveng. "Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?"
Ecology 88.11 (2007), pp. 2766–2772.
- [296] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002.
URL: <https://www.stats.ox.ac.uk/pub/MASS4/>.
- [297] A. Vyse, L. Hesketh, and R. Pebody.
"The burden of infection with cytomegalovirus in England and Wales: how many women are infected in pregnancy?"
Epidemiology & Infection 137.4 (2009), pp. 526–533.
- [298] M. Muñoz, R. Pong-Wong, O. Canela-Xandri, et al.
"Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank".
Nature Genetics 48.9 (2016), pp. 980–983.
- [299] A. Farmanbar, R. Kneller, and S. Firouzi.
"RNA sequencing identifies clonal structure of T-cell repertoires in patients with adult T-cell leukemia/lymphoma".
NPJ Genomic Medicine 4.1 (2019), pp. 1–9.
- [300] R. L. Warren, J. D. Freeman, T. Zeng, et al.
"Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes".
Genome research 21.5 (2011), pp. 790–797.
- [301] DecodeME Study. *DecodeME, the ME/CFS study*. 2022.
URL: <https://www.decode-me.org.uk/>.

- [302] Z. Y. Shan, L. R. Barnden, R. A. Kwiatek, et al.
“Neuroimaging characteristics of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS): a systematic review”.
Journal of Translational Medicine 18.1 (2020), pp. 1–11.
- [303] A. H. Miller, J. F. Jones, D. F. Drake, et al.
“Decreased basal ganglia activation in subjects with chronic fatigue syndrome: association with symptoms of fatigue”.
PLoS One 9.5 (2014), e98156.
- [304] A. Chaudhuri and P. O. Behan. “Fatigue and basal ganglia”.
Journal of the Neurological Sciences 179.1 (2000), pp. 34–42.
- [305] L. J. Carithers, K. Ardlie, M. Barcus, et al. “A novel approach to high-quality postmortem tissue procurement: the GTEx project”.
Biopreservation and Biobanking 13.5 (2015), pp. 311–319.
- [306] M. Uhlén, L. Fagerberg, B. M. Hallström, et al.
“Tissue-based map of the human proteome”.
Science 347.6220 (2015), p. 1260419.
- [307] *Human Protein Atlas*. 2021.
URL: [https://www.proteinatlas.org/ENSG00000112541-PDE10A/cell+line{}\(v21.1.proteinatlas.org\)](https://www.proteinatlas.org/ENSG00000112541-PDE10A/cell+line{}(v21.1.proteinatlas.org)).
- [308] A. Nott, I. R. Holtman, N. G. Coufal, et al. “Brain cell type-specific enhancer-promoter interactome maps and disease-risk association”.
Science 366.6469 (2019), pp. 1134–1139.