

Genetic variation contributing to autoimmune disease

Daniel Rainbow

Brasenose College

Trinity 2019

Supervisor: Professor Linda S. Wicker

Second supervisor: Professor John A. Todd

This dissertation is submitted for the degree of Doctor of Philosophy in Clinical Medicine

Abstract

Genetic variants influencing the IL-2 signaling pathway, including *IL2RA*, which encodes CD25, the alpha chain of the IL-2 receptor, are enriched in autoimmune diseases. Regulatory T cells (Tregs) are vital to the maintenance of a balanced immune system. They require IL-2 for their survival and by expressing high levels of CD25 are exquisitely sensitive to this cytokine. My thesis has focused on (1) accurate quantification and understanding of the heterogeneous nature of Tregs and (2) analysing phenotypes determined by *IL2RA* variants associated with autoimmune disease.

FOXP3 is the lineage-defining transcription factor of Tregs and stable expression is dependent on the demethylation of the Treg Specific Demethylated Region (TSDR) within intron 1 of the *FOXP3* gene. I developed and validated a high-throughput and sensitive method to measure the TSDR methylation and used the assay to define the heterogeneous nature of Tregs.

The genetic association of *IL2RA* with autoimmune disease is complex and is explained by five independent signals, which have been termed the A, C, D, E and F haplotypes. Genotype-to-phenotype correlations have previously been identified for the A and D haplotypes but here I dissected IL2-RA gene polymorphisms and its mRNA expression in detail. I showed that the C haplotype results in decreased *IL2RA* mRNA expression, seen only in activated $CD4^+$ T cells and the F haplotype correlates with decreased expression in *ex vivo* and activated B cells, and activated monocytes. The sequence relationship between disease-protective haplotypes was assessed across divergent populations. Finally, I have identified *IL2RA* mRNA expression phenotypes not associated with disease serving as a reminder that not all changes in gene expression correlate with a known disease haplotype.

Acknowledgments

I would like to thank Professor Linda Wicker and Professor John Todd for the opportunity to conduct this research, and for their continual support and supervision.

I gratefully acknowledge the participation of all Cambridge NIHR BioResource volunteers, without which this work would not be possible. I thank the Cambridge BioResource staff for their help with volunteer recruitment and the National Institute for Health Research Cambridge Biomedical Research Centre and NHS Blood and Transplant for funding.

I would also like to thank all members of the Diabetes and Inflammation Laboratory (DIL) who have helped me along the way. In particular, I acknowledge the Sample Preparation Team who processed the samples from the Cambridge BioResource, and performed the immunophenotyping, which was analysed Dr Marcin Pekalski (DIL). I thank Dr Ricardo Ferreria for his help and advice with purifying Treg populations, Jamie Inshaw for his statistical analysis using the stochastic search method and Dr Tony Cutler for general help and advice.

I would like to thank Dr Jennifer Asimit and Dr Chris Wallace (Department of Medicine, University of Cambridge and MRC Biostatistics Unit) for performing the *IL2RA* haplotype analysis and for many helpful discussions.

This research was supported by the Cambridge NIHR BRC Cell Phenotyping Hub and I am grateful for all their advice and support in cell sorting.

This work has been supported by funding from the Wellcome and the JDRF.

I would like to thank Dr Joanne Jones and Professor Alasdair Coles and all members of their lab for being so welcoming to me whilst being a visitor in their lab in Clinical Neurology, University of Cambridge.

Finally, I am forever grateful for the love and support I have received from my family: my wife Kara, and children Jessica and Benjamin who have been there for me and helped me through the many ups and downs encountered on this journey.

Abbreviations

AITD Autoimmune thyroid disease
APC Antigen presenting cell
ASE Allele-specific expression
B2M Beta-2-microglobulin
bisDNA bisulfite converted DNA
BSA Bovine serum albumin
bp base pair
CD Crohn's disease
cDNA complementary DNA
CVD Cardiovascular disease
Ct Cycle threshold
DN Double negative
DNA Deoxyribonucleic acid
dNTP Deoxynucleotide
DP Double positive
eQTL Expression quantitative trait loci
ERV Endogenous retrovirus
GTEx Genotype-Tissue Expression Project
HLA Human leukocyte antigen
HPRT hypoxanthine phosphoribosyltransferase 1
IL2RA interleukin 2 receptor subunit alpha
iPSC Induced pluripotent stem cell
iTreg Induced Treg
Kb Kilobase
LD Linkage disequilibrium
MAF Minor allele frequency
MFI Mean fluorescence intensity
MS Multiple sclerosis
NGS Next Generation Sequencing
NOD Non-obese diabetic
OR Odds ratio
PBMC Peripheral blood mononuclear cell
PCR Polymerase chain reaction
qPCR Quantitative PCR
PMA Phorbol myristate acetate
PS Psoriasis
PSC Primary sclerosing cholangitis
RA Rheumatoid arthritis
RNA Ribonucleic acid
RT Reverse transcriptase
sCD25 Soluble CD25
SP single positive
T1D Type 1 diabetes
Teff T effector cell
Th T helper cell
Treg Regulatory T cell
TSDR Treg Specific Demethylated Region
tTreg Thymically-derived Treg

Table of Contents

Abstract i

Acknowledgments ii

Abbreviations iv

Table of contents 1

Chapter 1 Introduction

1.1 Background 6

1.2 Autoimmunity 7

1.3. Genetic control of autoimmune disease 7

1.3.1. Limitations of GWAS and problems with fine mapping 10

1.4. Strategies to fine-map GWAS loci 12

1.4.1. Prioritisation of cell types to test SNPs selected by fine-mapping 14

1.5. Association of genes in the IL-2 pathway with autoimmune disease 16

1.6. Regulated expression of the IL-2 receptor 17

1.7. Regulatory T cells 20

1.8. *FOXP3* Treg-specific demethylated region 22

1.9. Objectives and research value 23

Chapter 2 Material and Methods

2.1 Ethics statement 24

2.2 Methods relating to Chapter 3 and 4: *FOXP3* Treg Specific Demethylated Region (TSDR) analysis 24

2.2.1 Cell sorting for *FOXP3* TSDR experiments 24

2.2.2 Compensation controls for cell sorting 27

2.2.3 DNA extraction and bisulfite conversion 27

2.2.4 Targeted measurement of DNA methylation 28

2.2.5 Generation of a standard curve to normalise TSDR methylation levels 34

2.2.6 Sorts based on Cytokine production 35

2.3 Methods relating to Chapter 5, 6 and 7: *IL2RA* **35**
2.3.1 Recall of genotyped individuals 35
2.3.2 Cell sorting for allele-specific expression 36
2.3.3 RNA extraction and cDNA synthesis 40
2.3.4 *IL2RA* mRNA expression 41
2.3.5 *IL2RA* Allele-Specific Expression 41
2.3.6 Analysis of *IL2RA* ASE data 44
2.3.7 Immunophenotyping of whole blood samples 44
2.3.8 Normalisation of immunophenotyping data 46
2.3.9 Targeted methylation analysis of *IL2RA* 48
2.3.10 Phylogenetic analysis of the *IL2RA* region 49
2.3.11 1000 Genomes Project genotype comparison across the *IL2RA* region 51
2.3.12 Genetic association of the A, A^{-660} and susceptible haplotypes with T1D 52

Chapter 3 Evaluation of *FOXP3* TSDR measurement

3.1 Introduction and aims **53**

3.2 Survey of *FOXP3* TSDR demethylation within CD4+ T cells subsets **54**

3.3 Understanding the overestimation of *FOXP3* TSDR measurement **57**
3.3.1 Does the cell cycle affect *FOXP3* TSDR measurement? 57
3.4.2 Does 'leaky' X chromosome affect *FOXP3* TSDR measurement? 59
3.4.3 Bias in DNA polymerase 62
3.4.4 Using a *FOXP3* TSDR standard curve to correct overestimation 64

3.5 Discussion **65**

Chapter 4 Using the *FOXP3* TSDR to assess Treg heterogeneity

4.1 Introduction and aims **68**

4.2 Survey of FOXP3 and HELIOS expressing cells **69**
4.2.1 Multiplex analysis of CTLA4 and CD4 methylation 74
4.2.2 Reproducibility of data 77
4.2.3 Characterising $FOXP3^{-}$ $HELIOS^{-}$ cells within the Treg gate 79
4.2.4 Can TSDR demethylated cells make cytokines? 81
4.2.5 Demethylation of lineage defining Th-1 and Th-17 regions 83

4.3 Application of *FOXP3* TSDR measurement **85**
4.3.1 Are the $CD25^{low}$ $FOXP3^{+}$ cells that are increased in autoimmunity demethylated at the *FOXP3* TSDR? 86

4.3.2 Are the Tregs expanded in T1D patients receiving ultra-low dose IL-2 demethylated at the *FOXP3* TSDR? 90

4.4 Discussion 92

Chapter 5 *IL2RA* autoimmune disease associated haplotypes

5.1 Introduction and aims 96

5.2 Haplotype analysis and comparison 99

5.2.3 How similar are donors with the same haplotype 103

5.2.4 How similar are the haplotypes between each other? 106

5.2.5 Frequency of *IL2RA* haplotypes in different parts of the world 108

5.3 Analysis of the *IL2RA* group A haplotype 110

5.3.1 Introduction 110

5.3.2 Epigenetic evidence identifies rs61839660 as the most likely causal variant of the group A SNPs 111

5.3.3 Identification of rare *IL2RA* haplotypes 114

5.3.4 High throughput allele-specific expression with the A haplotype 117

5.3.5 Expression of *IL2RA* in purified cells 121

5.3.6 The A haplotype makes more *IL2RA* mRNA 123

5.3.7 Immunophenotyping the DIL PBMC bank collection 126

5.3.7.1 Data collection and normalisation 128

5.3.7.2 Group A immunophenotyping: genotype to phenotype 131

5.3.8 Interaction of the *IL2RA* promoter with a region containing rs61839660 136

5.3.9 The A^{-660} haplotype is not protective for T1D 138

5.3.10 Summary of the A haplotype 139

5.4 The *IL2RA* group C haplotype 140

5.4.1 Introduction 140

5.4.2 Epigenetic analysis of the *IL2RA* C haplotype 141

5.4.3 High throughput allele-specific expression with the C haplotype 145

5.4.4 Immunophenotyping the C haplotype for CD25 expression 149

5.4.5 Summary of the C haplotype 150

5.5 Evidence that the AITD association is two separate signals 151

5.6 The *IL2RA* group D haplotype 156

5.6.1 Explaining the genetic association of the *IL2RA* D haplotype 156

5.6.2 The D haplotype makes less *IL2RA* in naive $CD4^+$ T cells 159

5.6.3 The D haplotype correlates with decreased frequency of $CD4^+$ naive T cells that express CD25 161

5.6.4 Expression phenotypes confirm that the A and D haplotypes best explain the association with MS 162
5.6.5 Which of the *IL2RA* group D SNPs is most likely to be causal? 167
5.6.6 Summary of the D haplotype 171

5.7 The *IL2RA* group E haplotype 172
5.7.1 Introduction 172
5.7.2 The E SNP breaks a methylation site 173
5.7.3 Does the E haplotype affect the expression of *IL2RA* or *RBM17*? 174
5.7.4 Summary of the *IL2RA* E haplotype 179

5.8 The *IL2RA* group F haplotype 180
5.8.1 Introduction 180
5.8.2 The F haplotype reduces *IL2RA* expression in B cells and monocytes 181
5.8.2 The F haplotype causes reduced CD25 expression on B cells 185
5.8.3 Summary of the F haplotype 186

5.9 Summary of *IL2RA* disease related phenotypes 187

5.10 Discussion 190

Chapter 6 Non-disease associated variants that alter *IL2RA* expression

6.1 Introduction 194

6.2 Soluble CD25 195
6.2.1 Haplotype analysis of variants associated with a change in sCD25 concentration 197
6.2.2 Comparison of alleles found on the C haplotype split by rs12722497, that alter the levels of sCD25 199
6.2.3 Which cell type does rs12722497 affect? 200
6.2.4 CD25 expression on B cells is affected by three SNPs 203
6.2.5 Repeat ASE for sCD25 controlling for rs1887027 207
6.2.6 The region surrounding rs12722497 is demethylated in $CD14^+$ monocytes 209

6.3 B cell phenotypes 211

6.4 The F haplotype is associated with protection from T1D and decreased levels of sCD25 214

6.5 Summary of non-disease related phenotypes 214

6.6 Discussion 215

Chapter 7 *IL2RA* Haplotype structure and history

7.1 Introduction and aims **218**

7.2 Assessment of haplotype similarity **220**
7.2.1 Using pre-developed methods 220
7.2.2 Development of method to visualise haplotype similarity 222
7.2.3 *PTPN22* as a proof of principal 223
7.2.3 *TYK2* as a proof of principal 227

7.3 The *IL2RA* region **229**
7.3.1 The *IL2RA* A haplotype 229
7.3.2 The *IL2RA* C haplotype 231
7.3.3 The *IL2RA* D haplotype 234

7.3.4 The *IL2RA* E haplotype **239**
7.3.5 The *IL2RA* F haplotype 242

7.4 Discussion **245**

Chapter 8 General conclusions **248**

Appendix 1 Publications with data from my thesis **252**

References **255**

Chapter 1

Introduction

1.1 Background

My thesis contains two distinct projects, the first utilises a high-throughput assay I developed measuring the methylation status of the *FOXP3* Treg Specific Demethylated Region (TSDR) to assess the heterogeneity of Treg populations and the second focuses on mRNA and protein expression phenotypes determined by multiple, independent genetic associations of the *IL2RA* region with autoimmune disease. These projects are linked through the IL-2 pathway, genes of which are a major focus of research within my laboratory because many are associated with type 1 diabetes (T1D) and other autoimmune diseases. Tregs are dependent on IL-2, and function to maintain homeostasis within the immune system, and *IL2RA* encodes one of the chains of the high affinity IL-2 receptor, therefore connecting these two projects that aim to understand Treg diversity and to define the cellular basis of the genetic association of *IL2RA* with autoimmune diseases.

I first started working on developing and validating the assay to measure the *FOXP3* TSDR, and soon realised that it could be modified to allow the measurement of allele-specific expression (ASE), which had been performed previously in our lab using an extremely powerful but laborious PCR cloning and sequencing technique[1]. This coincided with a publication by our close collaborator Dr Chris Wallace who had implemented a stochastic search method to fine map the *IL2RA* region, resolving some of the complex genetic associations[2]. Therefore, I took on the challenge to identify the genotype to

phenotype associations of *IL2RA* with autoimmune disease using ASE, followed up with protein expression experiments when feasible.

1.2 Autoimmunity

There are approximately four million people in the UK diagnosed with at least one of the 80 autoimmune conditions known, with up to a third living with more than one[3]. As well as the cost to the individual's health, there is also a huge financial cost to the UK with three common autoimmune diseases (T1D, RA and MS) costing £13 billion per year[3]. With the incidence of autoimmune disease increasing by 5% per year[3], there is a real need to understand the mechanisms that lead to their development and use this knowledge to develop better treatments.

The feature common to all autoimmune diseases is the break in self-tolerance that leads to the body's immune system attacking healthy tissue. Each autoimmune disease has its own characteristics including the age of onset, target organ and incidence, but are generally more common in first degree relatives suggesting a genetic element[4], with the concordance rates in twin studies varying between 12-67%[5]. In this thesis I will use genetic data from three autoimmune conditions: type 1 diabetes (T1D), multiple sclerosis (MS) and autoimmune thyroid disease (AITD).

1.3. Genetic control of autoimmune disease

Almost 50 years ago, the first region to be associated with T1D was identified as the Human leukocyte antigen (HLA), and as the effect size was large only 50 T1D cases and

233 controls were needed to detect the disease association[6,7]. The original HLA association predates the development of Sanger sequencing[8,9] or other direct genotyping methods, and used a lymphocyte cytotoxicity test to determine HLA genotype. This test was originally developed for matching organ donors and recipients in transplantation[10]. The Human leukocyte antigen (HLA) region on chromosome 6 is the most gene dense region of the genome with 252 expressed loci across a 7.6 Mb region including those that encode the MHC proteins[11], which bind and present antigens. These are some of the most associated genes with either susceptibility to or protection from autoimmune disease; with the HLA-DR3/DR4 alleles conferring susceptibility to T1D, whereas DR14/DR15 are protective. For MS, it is the DR15 allele that is susceptible and DR14 protective, and for AITD the DR3 allele is susceptible and DR4 protective. The HLA region remains by far the most strongly associated region, with an odds ratio greater than 16[REF12].

Just over 10 years ago there was a boom in the field of genetic association studies that resulted from the development of genome-wide genotyping arrays that allowed hundreds of thousands of variants to be genotyped simultaneously, and marked the end of candidate gene approaches. The principle of genome-wide associated studies (GWAS) is to assess if the frequency of a particular allele is different between people with disease compared to those who do not have the disease being investigated, and is repeated on many hundred of thousands of variants across the genome. Each variant is reported as the odds ratio which is a measure of the effect size that is calculated as the ratio of the allele in the cases versus the ratio of allele in the controls (e.g. $(A_{case}/B_{case})/(A_{control}/B_{control})$). Immunochip was a custom genotyping array that covered 186 regions[13] of the genome that had evidence of

association with autoimmune disease and every SNP with a MAF greater than 1% included (if it met the sequence requirements of the chip).

Studies have been performed for many autoimmune diseases using the Immunochip genotyping array that have led to the identification of many new associations, some of which revealed previously unknown sharing of regions by two or more diseases. Outside of the MHC, many genes have been associated with the development of autoimmune disease, with some shared across many diseases like *PTPN22*, *TYK2*, *STAT4* and *IL2RA*, whereas others associations are specific to one autoimmune disease, like the gene *Ins*, which encodes insulin, for T1D. However, even though the same gene might be associated with multiple diseases the direction of effect might be the opposite, for example the *IL2RA* gene is protective for T1D but susceptible for IBD[14], and interestingly many other associations are also discordant between T1D and IBD[15].

There is some genetic overlap between T1D, AITD and MS, and early work showed that AITD and T1D were genetically more similar to each other than to MS[16], however this work was conducted prior to the large GWAS studies being published and used only 15 SNPs that were significantly associated with the five autoimmune diseases examined. A more recent study of 15 immune-mediated diseases using GWAS data showed those which produced autoantibodies (e.g. type 1 diabetes (T1D), juvenile idiopathic arthritis and rheumatoid arthritis) were genetically more similar when compared to auto-inflammatory disorders, such as ulcerative colitis and Crohn's disease (CD)[17]. Although, MS and AITD had less genetic enrichment for shared loci with T1D than JIA, the *IL2RA* gene association was shown to be shared by the three diseases.

GWAS is an unbiased approach and consequently unexpected associations are identified that reveal novel biological processes not previously linked with disease. For example, the genes *NOD2*, *ATG16L1* and *IRGM* [18,19] that are all disease risk factors for CD are all linked to the autophagy pathway and modulate the response to microbial products[20]. Additionally, GWAS can implicate disease pathways and in the case of T1D has focused our attention on the IL-2 pathway, which will be discussed later.

1.3.1. Limitations of GWAS and problems with fine mapping

The use of genotyping arrays has been a phenomenal success at identifying regions associated with autoimmune disease, however there are still many challenges to resolve. Firstly, the number of loci identified by GWAS is driven by a combination of allele frequency, effect size and sample size. Larger sample sizes allow for loci with smaller effect size and lower allele frequency to be identified. As an extreme example, the largest GWAS performed to date was on insomnia in which 1.3 million people were analysed and over 200 loci identified[21], whereas the previous smaller but still substantial study of 113,000 people only identified three loci[22,23]. Although there is a substantial cost for performing additional GWAS to increase sample size, the likelihood of understanding disease mechanisms increases when more associated loci are identified and accurately fine-mapped. The GWAS performed for common autoimmune diseases like T1D[17], MS[24] and IBD[14], have compared tens of thousands of cases and controls. As the analysis is genome-wide, correction for multiple testing needs to be applied and typically uses Bonferroni correction that sets the threshold for evidence of significance at typically p =

$<5x10^{-08}$ [REF25], (which assumes an $\alpha = 0.05$ and 1,000,000 million variants: 0.05/1,000,000 = $5x10^{-08}$).

Secondly, the LD structure of the population, which aids in the initial identification of associated loci, hinders the fine-mapping and identification of the causal variant[26], and the most associated variant is not necessarily the causal variant[27]. The majority of GWAS studies have been performed on the European population and the use of a more divergent populations would aid fine-mapping as there would be increased recombination rates that reduces the size of LD blocks[28,29]. However, the incidence of autoimmune disease is not uniform across all regions of the world. For example, the country with the highest incidence of T1D diagnosed in children under 14 is Finland with 57 cases per 100,000 whereas Venezuela had the lowest incidence with only 0.1 cases per 100,000 people[30]. This large variation in the incidence of T1D results from a combination of genetic and environmental factors, which leads to another consideration for GWAS which is population stratification.

Population stratification is the systematic ancestry differences between cases and controls[31] and needs to be taken into account to stop spurious associations resulting from sub-population ancestry differences. This is especially true in large GWAS when samples could be collected from many different countries[31,32].

As well as performing ever larger GWAS studies, there should be increased emphasis on the fine-mapping and understanding of the genetic association of the regions identified that will increase our understanding of disease pathology[33].

1.4. Strategies to fine-map GWAS loci

A GWAS will result in a number of loci associated with either protection or susceptibility to the trait being studied, and will generally report the lead variant for each locus. The two main strategies to analyse GWAS are stepwise regression, and Bayesian stochastic search methods such as GUESSFM[2] or FINEMAP[34]. Stepwise regression identifies the single most associated SNP within a region, then by reiteratively conditioning on the first and any subsequent signals identifies any additional associations meeting the significance threshold. However, due to LD the SNP with the lowest p-value is not necessarily the causal variant and either p-value or LD with the lead SNP are used to set a threshold to determine which SNPs are included in the group of potential causal variants[35]. Bayesian methods differ as they simultaneously explore multiple SNP models for each region and identify if more than one signal exists and generates credible sets of SNPs that are likely to contain the causal variant. The process of fine-mapping attempts to refine the association to a limited number of SNPs that can be followed up with functional experiments[27,36].

In order to fine-map a region, all variants within a region need to be genotyped or imputed to understand the LD structure and relatedness of the SNPs. Stepwise regression does not apply a probability of causality to each SNP, which is an advantage that Bayesian methods bring to fine-mapping[27]. Bayesian methods calculate the posterior probability of each variant being causal for the association in that region, information that can be used to help prioritise a particular SNP within a credible SNP group. This has been applied to the IBD GWAS with great success, and 18 association signals have been refined to a single causal variant with greater than 95% certainty[14].

Once a limited number of SNPs have been identified through statistical approaches, then each SNP can be interrogated to see if it has any attribute that provides functional evidence for causality such as being a non-synonymous SNP, a SNP that disrupts a transcription factor binding site or a SNP residing in a region of epigenetic interest in a disease relevant cell type, thus prioritising a particular variant.

Large scale GWAS have identified tens to hundreds of loci associated with disease and the next challenge is understanding how each locus contributes to disease, enabling the identification of pathways or cell types that are enriched for GWAS hits, providing potential mechanisms that could be targeted therapeutically.

Identifying the function or phenotype of each disease associated region is challenging and takes many years of work to fully understand. The ability to recruit people carrying either the protective or susceptible allele at specific variants is essential in order to perform genotype to phenotype studies. The development of regional BioResources, like the Cambridge BioResource and the Oxford BioResource that have recruited thousands of volunteers with dense genotyping information, has allowed the recall of participants with relevant genotypes to assess genotype to phenotype correlations, and this strategy has been applied successfully to the *IL2RA*[37] and *TYK2*[REF38] regions.

1.4.1. Prioritisation of cell types to test SNPs selected by fine-mapping

Most GWAS hits are located in non-coding regions of the DNA[39], suggesting they are residing in enhancer or insulator regulatory elements. In order to prioritise which cell type to analyse for the disease being investigated, the credible SNPs from each fine-mapped region are overlaid with RNA and chromatin information to observe variants located within cell type specific enhancer elements, which for autoimmune diseases clustered in T and B cells[39]. When the credible SNPs identified by GWAS for T1D were aligned to 15 chromatin states in 127 cell types from the RoadMap Epigenomics Program[40] and ENCODE[41], there was strong enrichment for SNPs to be located in enhancer elements in thymus, $CD4^+$ T cells, $CD8^+$ T cells and B cells[17]. Assessing a wide variety of cell types and activation conditions is key to identifying where each GWAS variant is exerting its effect, for example a study using 13 different activation conditions for $CD4^+$ T cells found that IBD variants were enriched in Th1 cells[42]. However, to scale this work to all immune cells and their activation conditions would be extremely time consuming and expensive.

As an alternative, large scale eQTL studies, which correlates gene expression with genotype, are useful in identifying if any of the credible SNPs are more or less associated with a change in gene expression of nearby genes. Additionally, an eQTL might be identified that has no link to the credible SNP group, but needs to be considered when designing genotype to phenotype studies. The largest eQTL studies are performed in whole blood that lacks cell-type specificity, but the increased sample size does help

identify eQTLs and highlights that there is variation in gene expression that is not necessarily related to your GWAS SNP[43].

The Genotype-Tissue Expression (GTEx) project[44] is an excellent resource that has performed whole-genome sequencing and RNA-Seq on 54 tissues from almost 1,000 individuals and contains gene and splice isoform expression information and QTL analysis. The best strategy would be to combine both expression analyses and QTL studies within single cells. Single cell RNA-Sequencing is now commonplace, and determines the unique transcriptomic profile of an individual cell, rather than a bulk population and has the potential to reveal rare or heterogeneous cell populations[45]

With the Human Cell Atlas project aiming to 'create comprehensive reference maps of all human cells' by analysing the gene expression of every cell in the body[46], it will soon be possible to understand cell-type-specific expression. For example, the single-cell atlas of the lung epithelium identified a previously unknown cell population, called the pulmonary ionocyte that is the major expresser of CFTR, the protein encoded by the gene that is mutated in cystic fibrosis[47,48]. Therefore, there is great potential single cell eQTL studies, which are in their infancy, but from the approach shown in Figure 1.1, it is clear that single cell eQTL will be a powerful technique to identify the effect of each GWAS variant and how it changes the regulatory network of which it is part[49].

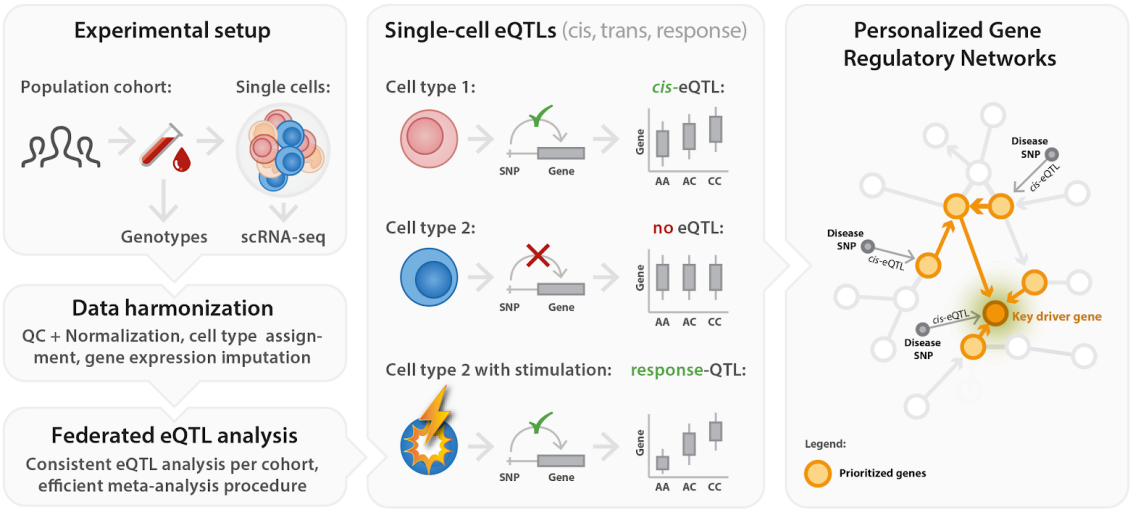


Figure 1.1. Single cell eQTL identify the upstream interactors and downstream consequences of trait-related genetic variants in specific immune cell types. Figure reproduced from van der Wijst *et al*. "Single-cell eQTLGen Consortium: a personalized understanding of disease", 2019[REF49].

1.5. Association of genes in the IL-2 pathway with autoimmune disease

The T1D GWAS results identified that many of the implicated regions contain genes involved with the immune system, and are enriched in T cell function[17,50], several of which are fundamental proteins within the IL-2 pathway. These include IL-2 itself, the alpha chain of the IL2 receptor (IL2RA), the transcription factor BACH2 and the Protein Tyrosine Phosphatase PTPN2[17], that function to promote T-cell differentiation and the development of a regulatory phenotype[17]. Replacement of the susceptible *Il2* allele in the non-obese diabetic (NOD) mouse model of T1D, in which 80% of mice spontaneously develop T1D with a protective *Il2* allele that produces two fold more IL-2 is extremely protective and results in only 10% of mice developing T1D[51].

This evidence highlighting the importance of IL-2 in the pathogenesis of T1D led our laboratory to pursue *in vivo* human studies to administer ultra-low doses of exogenous IL-2 with the aim to increase the frequency and function of Tregs, which are exquisitely sensitive to IL-2, without activating effector T cells. The first study called DILT1D, aimed to find the optimal low dose of IL-2 needed to increase Treg frequency by 20%[REF52,53] and the second study called DILfrequency was designed to determine the frequency at which IL-2 needed to be administered to maintain the increased frequency of Tregs, without activating other immune cells that are responsive to IL-2[54,55]. This work has culminated in ITAD, Interleukin-2 Therapy of Autoimmunity in Diabetes ITAD trial (www.clinicaltrials.gov: NCT03782636) that will assess if twice weekly injection of low-dose IL-2 for six months can preserve residual insulin production in children and young adults recently diagnosed with T1D (recruitment started September 2019). The primary outcome is the preservation of C-peptide over the 6 month-treatment period between those receiving IL-2 and a placebo group. We know from both DILT1D and DILfrequency that administration of low-dose IL-2 increased Treg frequency, and hope that the increased Treg frequency expected in ITAD will tip the balance of the immune system to a more regulatory phenotype and prevent further autoimmune destruction of the insulin producing cells of the pancreas.

1.6. Regulated expression of the IL-2 receptor

The high affinity IL-2 receptor is a trimeric receptor that is encoded by the genes *IL2RA*, *IL2RB* and *IL2RG* that translate to the proteins CD25, CD122 and CD132. The gene *IL2RG* is also called the common cytokine receptor gamma chain as it is shared with five other receptors, IL-4R, IL-7R, IL-9R, IL-15R, and IL-21R. The IL-2 receptor can bind

IL-2 with varying degrees of affinity depending on which chains of the receptor are present: CD25 alone can bind IL-2 with low-affinity, the dimer of CD122 and CD132 binds IL-2 with intermediate affinity, but it is the trimeric receptor of CD25, CD122 and CD132 that bind IL-2 with the highest affinity. Only the intermediate and high affinity receptors have the ability to signal, which requires the heterodimerization of the cytoplasmic domain of CD122 and CD132[REF56,57].

The IL-2 receptor is expressed on $CD4^+$ T cells, $CD8^+$ T cells, B cells, NK cells and neutrophils and promotes cell proliferation, differentiation and function by signaling through the JAK STAT, PI3 kinase or MAP kinase pathways[58]. Dendritic cells also express *IL2RA* and might sequester IL-2, thus inhibiting T cell responses[59], but have also been shown to trans-present IL-2, enabling high-affinity IL-2 signalling prior to *IL2RA* upregulation on T cells[60].

Even though IL-2 signals through the same pathway, each cell type does not respond in the same way and can promote either proinflammatory or anti-inflammatory differentiation in $CD4^+$ T cells. IL-2 signalling does not directly promote or inhibit the direction of $CD4^+$ T helper differentiation that happens as a consequence of IL-2, but rather modulates the expression of specific receptors and transcription factors, which in turn specify a particular Th lineage. Overall IL-2 provides positive signals for the differentiation of naive $CD4^+$ T cells towards a Th1 or Th2 phenotype, but an inhibitory signal for Th17 or Tfh differentiation[61,62] (Fig. 1.2).

Tregs express the highest levels of *IL2RA*, but are unable to produce IL-2 themselves, and therefore are reliant on IL-2 produced predominantly by T cells[63] for functional maturation in the thymus[64], and in the periphery for Treg homeostasis[65].

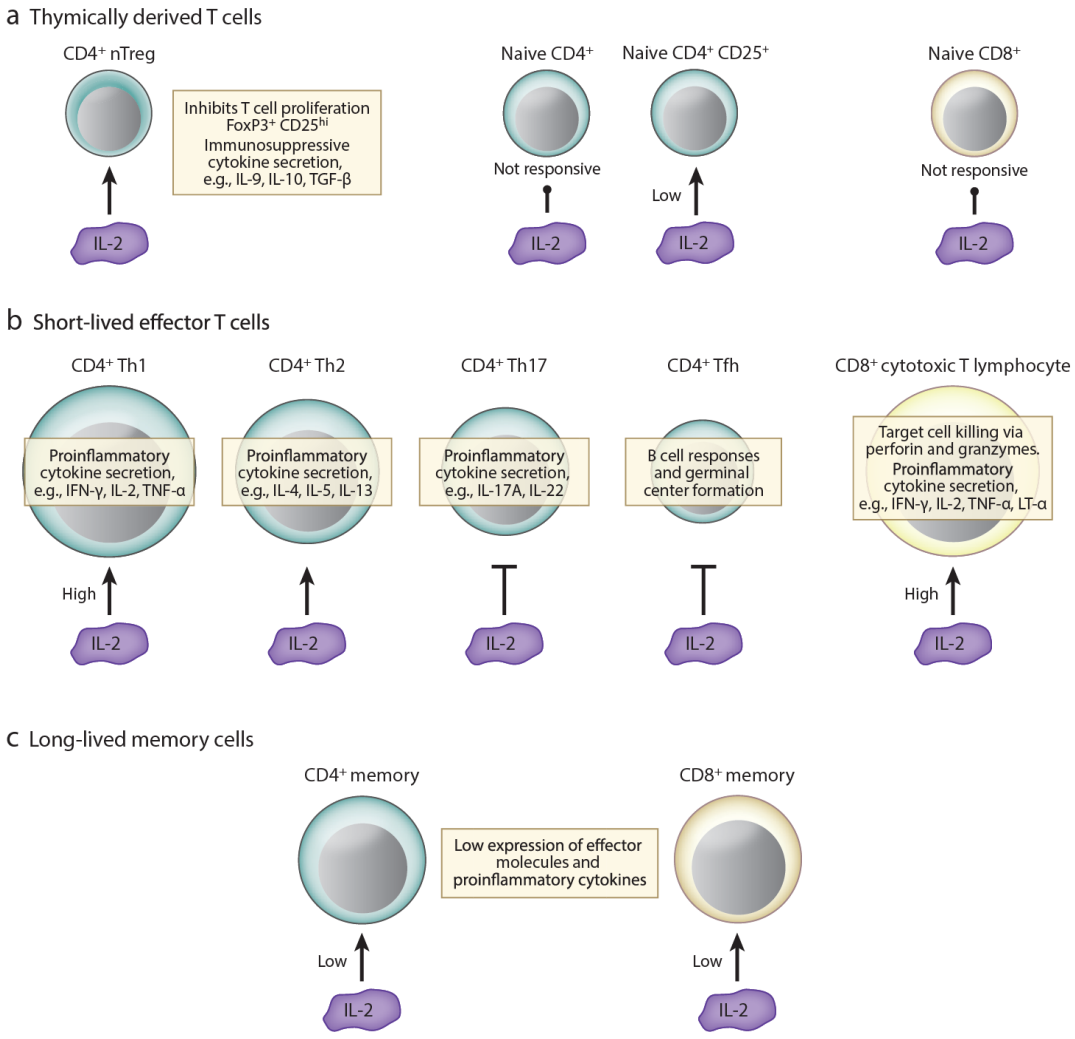


Figure 1.2. IL-2 biology: IL-2 positively influences the homeostasis and development of a number of T cell lineages. Expression of IL-2R in nTregs (*a*) and antigen-activated T cells (*b*,*c*) influences their functions. IL-2-responsive T cell subsets have extremely diverse characteristics and proinflammatory and anti-inflammatory biological roles. Moreover, IL-2 inhibits the differentiation of Th17 and Tfh cells (*b*), thus making IL-2 an important regulator of T cell lineage commitment. In addition, the levels of IL-2 signalling can define the effector/memory fates of $CD4^+$ and $CD8^+$ T cells: High levels of IL-2 favour the development of short-lived effector cells (*b*), while low levels of IL-2 signalling promote the differentiation of memory T cells (*c*). IL-2 stimulation of T cell function is depicted by

arrows and IL-2 inhibition of lineage development is represented by inhibitory lines, and no response to IL-2 is indicated by a line with a circle at the end. Abbreviations: IFN-γ, interferon gamma; IL-2R, interleukin-2 receptor; LT-α, lymphotoxin alpha; nTreg, thymically derived regulatory T cell; Tfh, follicular helper T; TNF-α, tumour necrosis factor alpha. Figure and legend reproduced with slight modification from Ross and Cantrell "Signalling and Function of Interleukin-2 in T Lymphocytes", 2018[REF62.]

1.7. Regulatory T cells

Regulatory T cells (Tregs) are a heterogeneous population of cells that are essential for the maintenance of self-tolerance that are characterised by the constitutive expression of the transcription factor FOXP3[REF66–68]. The critical regulatory role Tregs play in balancing the immune system is highlighted in the rare disease Immunodysregulation polyendocrinopathy enteropathy X-linked (IPEX) in which patients do not express the FOXP3 protein and the suppressive function of their Tregs is severely impaired resulting in multiple autoimmune disorders[69]. In common autoimmune diseases there is a growing literature describing either a decrease in Treg frequency or function[70,71], and in T1D, MS and AITD there is evidence that the ability of Tregs to suppress has been reduced[72–74] (Fig. 1.3).

Tregs function to suppress autoreactive T cells through four known mechanisms (Fig. 1.3). The first is through a process called trans-endocytosis in which the costimulatory molecules CD80/CD86 are pulled off the antigen presenting cells (APC) and internalised by Tregs expressing the inhibitory receptor CTLA4 thereby reducing the capacity for costimulation[75]. Secondly, Tregs expressing Galectin-9 can induce apoptosis of T effector cells (Teff) expressing the T cell immunoglobulin and mucin domain-3 (TIM-3)[76]. Thirdly, Tregs can secrete the anti-inflammatory cytokines TGFβ, IL10 and IL35[REF77,78]. Lastly,

Tregs express the ectonucleotidases CD39 and CD73 that convert ATP to AMP and then to adenosine, which can bind to the A2a receptor on Teff cells that suppresses their function[79].

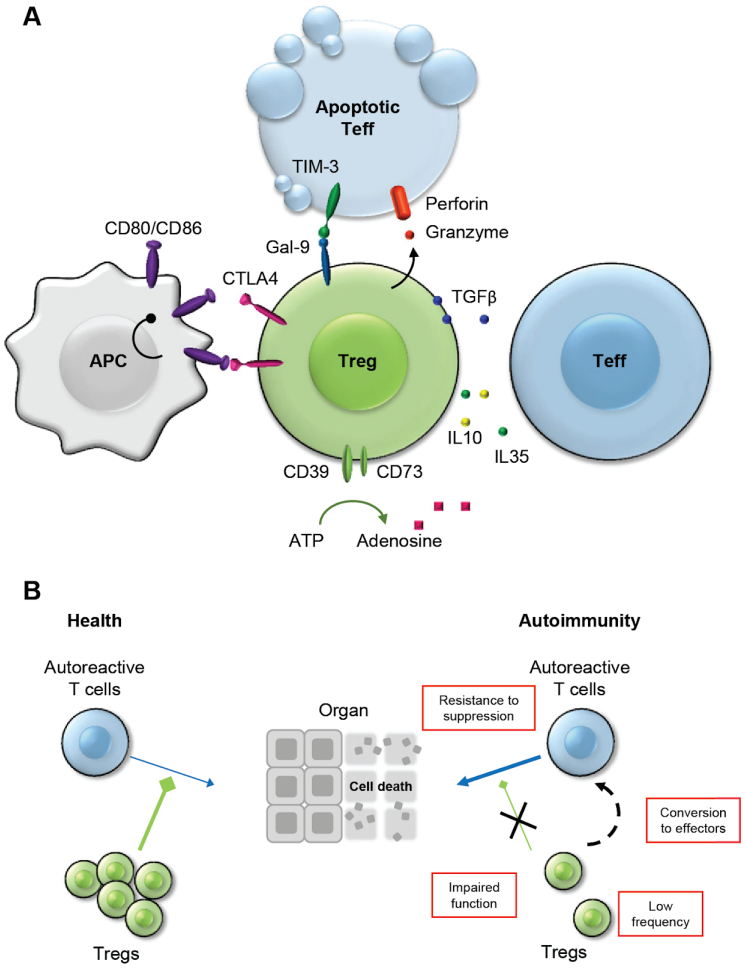


Figure 1.3. A) Regulatory T cell mechanisms of suppression. Regulatory T cell (Treg) can suppress by four basic mechanisms. The interaction between cytotoxic T lymphocyte antigen-4 (CTLA4) and CD80/CD86, expressed by antigen presenting cells (APCs), leads to CD80/CD86 down-regulation. Removal of these co-stimulatory molecules modulates APC function, limiting the initiation of an adaptive immune response. Tregs induce effector T cell (Teff) apoptosis by the interaction between Galectin-9 (Gal-9) and the T cell immunoglobulin and mucin domain-3 (TIM-3), and by the release of granzymes which enter Teffs via perforin pores. Tregs release the anti-inflammatory cytokines TGFβ, IL10 and IL35. Treg expression of the ecto-enzymes CD39 and CD73 enables the hydrolysis of pro-inflammatory adenosine triphosphate (ATP) into anti-inflammatory adenosine (ADO). B) Regulatory T cell defects in autoimmunity. In health, Tregs maintain tolerance by exerting suppression of effector T cells. In organ specific autoimmune disease, Tregs fail to suppress autoreactive effector T cells, therefore leading to target cell death. Reported reasons for this include inadequate numbers of Tregs, impaired suppressive ability, Treg conversion into effector cells and resistance of effector T cells to Treg-mediated suppression. Figure and legend reproduced from Grant, C. *et al.* "Regulatory T-cells in autoimmune diseases: Challenges, controversies and yet unanswered questions", 2015[REF71].

Clinical research has now focused on how to redress the reduced function and frequency of Tregs seen in autoimmune disease, through either boosting Tregs *in vivo* through administration of exogenous IL-2, or by adoptively transferring expanded autologous Tregs, with the aim to rebalance the immune system in favour of a more regulatory phenotype[80,81]. In order to fully understand the outcome, good methods are needed to measure the frequency and heterogeneity of the population.

1.8. *FOXP3* Treg-specific demethylated region

Regulatory T cells (Tregs) are characterised by the constitutive and stable expression of the transcription Factor FOXP3. A conserved region within intron 1 of *FOXP3* that contains 15 CpG sites has been identified as specifically demethylated in $FOXP3^+$ Tregs, whereas in all other cell types including those that transiently express FOXP3, such as induced Tregs or activated effector T cells, the region is methylated and has been named the Treg specific demethylated region (TSDR)[82–85]. Methods have been developed to exploit this differential methylation between Tregs and non-Tregs as a way to enumerate the proportion of *bona fide* Tregs within a sample and include qPCR[86,87], Pyrosequencing[88], analysis of Sanger sequencing[89,90] and PCR cloning and sequencing[84]. All of these methods use bisulfite converted DNA (bisDNA), which results from a chemical reaction in which DNA is treated with sodium bisulfite and all Cytosine nucleotides are converted to Uracil (which is read as Thymine during PCR) unless they are methylated, which causes resistance to the sodium bisulfite reaction and retention of Cytosine. All of the methods noted above average the methylation of each of the CpG sites except the most laborious method (cloning and sequencing) that does yield single-base resolution for each

sequencing read. Additionally, the results are generally presented as a heat map, which does not permit accurate assessment of the degree of TSDR methylation. Therefore, I have developed a high-throughput method to accurately quantity the methylation of the TSDR at single-base resolution which addresses these issues[91] and the validation and application of the method will be discussed in Chapter 3 and 4 of this thesis.

1.9. Objectives and research value

As mentioned, my thesis contains two different projects. First I will present the work to develop and validate a high-throughput method to measure the methylation of the *FOXP3* TSDR. During assay development I realised there was an overestimation of demethylation, and investigated potential sources and developed ways to correct for the overestimation. Next I applied this method to dissect the Treg population and show that not all Tregs have a demethylated *FOXP3* TSDR. This work is important as it shows that Tregs can be classified based on the methylation of the *FOXP3* TSDR, and have applied this assay to answer various biological questions.

Secondly, I have comprehensively analysed the complex genetic association of the *IL2RA* region with autoimmune disease. I have attempted to identify genotype to phenotype correlations that explain the genetic association, as well as identify the most likely causal variant for each association signal. This work is relevant as the fine mapping of GWAS regions is the next milestone, and this work highlights many of the tools and techniques needed.

Chapter 2

Material and Methods

2.1 Ethics statement

The NHS Cambridgeshire Research Ethics Committee approved all work involving human participants. Samples were obtained from the Cambridge BioResource (www.cambridgebioresource.org.uk) as part of the 'Genes and Mechanisms of Type 1 Diabetes' study, LREC number 05/Q0106/20; or through ‘CAMSAFE - A study of the long term safety and mechanism of action of alemtuzumab (Campath-1H), LREC number 11/EE/0007. Participants were of self-reported white ethnicity and informed consent was obtained from all volunteers for the collection and use of peripheral blood samples. Data and samples are always treated anonymously and confidentially.

2.2 Methods relating to Chapter 3 and 4: *FOXP3* Treg Specific Demethylated Region (TSDR) analysis

2.2.1 Cell sorting for *FOXP3* TSDR experiments

$CD4^+$ T cells were purified from blood using RosetteSep human $CD4^+$ T-Cell Enrichment Cocktail (Stemcell Technologies). The $CD4^+$ T cells were stained in PBS supplemented with 0.2% BSA and brilliant stain buffer (BD Biosciences) at room temperature for 30

minutes with the panel of antibodies to surface markers detailed in Table 2.1. Two different antibody clones to CD25 on APC were used to increase the sensitivity of the CD25 staining. An intracellular stain was performed to enable sorting of various FOXP3 and HELIOS positive and negative cell populations, which used the eBioscience Foxp3 Transcription Factor Staining Buffer Set[92] and the intracellular antibodies listed in Table 2.1. The cells were sorted on a BD FACSAria Fusion with the laser and filter configuration shown in Fig. 2.1. Some additional sorts were performed using Ki-67 as a marker of cell proliferation, and for these sorts Ki-67 antibody (clone B56 on BUV395, BD Horizon) was added to the antibody panel listed in Table 2.1. Examples of the gating strategy used for cell sorting are shown in Figure 4.1. Cells were sorted in 1.5 ml lo-bind tubes (Eppendorf) containing 500 µl of PBS supplemented with 0.2% BSA.

Table 2.1. Antibodies used to sort $CD4^+$ T cell populations based on FOXP3 and HELIOS expression.

Marker	Clone	Fluorochrome	Supplier	µl per 30 million CD4
CD3	UCHT1	BV510	BioLegend	15
CD4	RPA-T4	AF700	BioLegend	15
CD25	M-A251 2A3	APC APC	BD Biosciences BD Biosciences	60 15
CD127	eBioRDR5	PE-Cy7	eBioscience	15
CD27	O323	BV605	BioLegend	15
CD45RA	HI100	BV785	BioLegend	15
FOXP3	259D	PE	BioLegend	15
HELIOS	22F6	FITC	BioLegend	15
Viability dye	65-0865-14	eFlou780	eBioscience	1

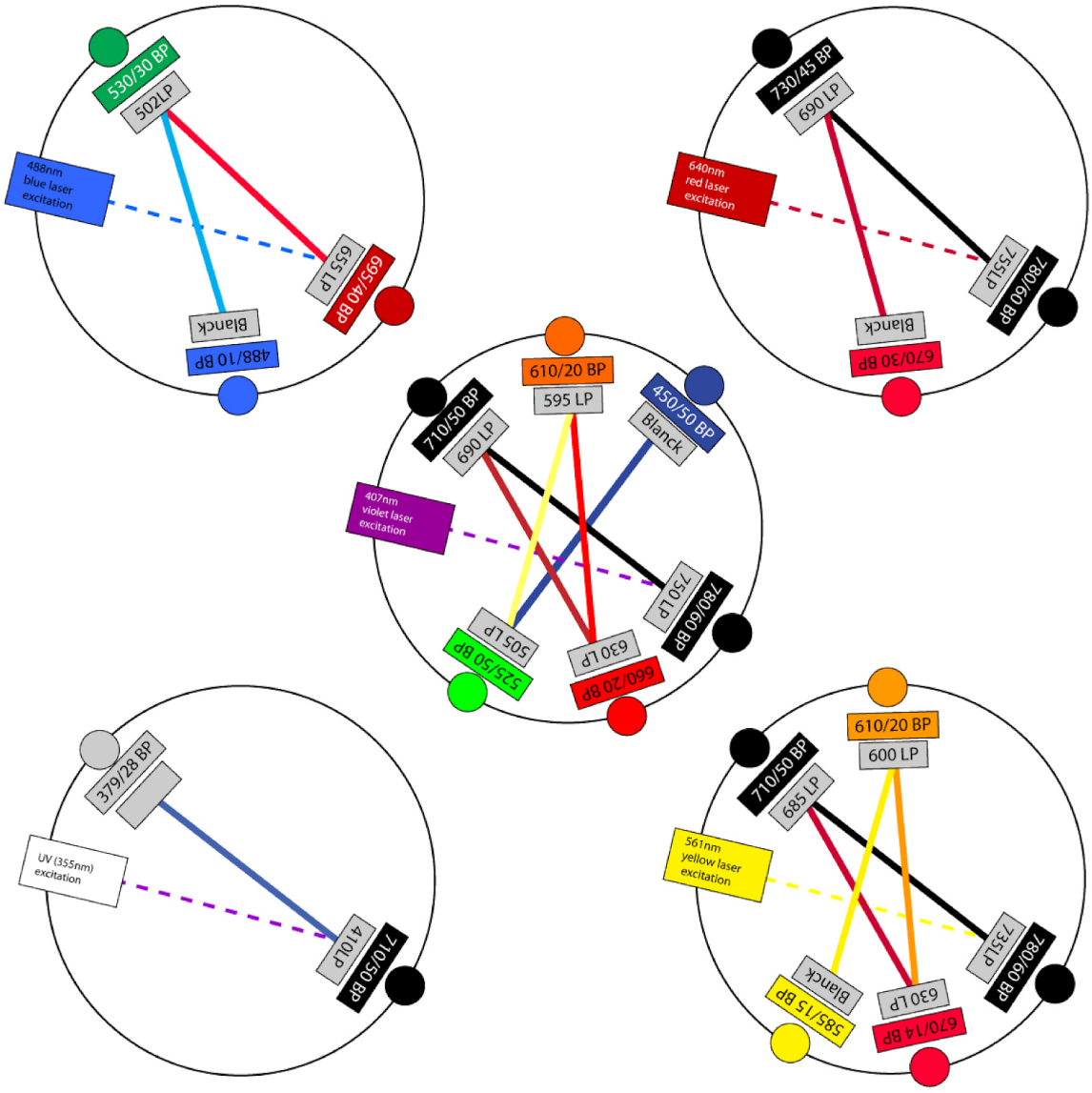


Figure 2.1. Laser and filter configuration of the cell sorters used. The BD FACSAria Fusion and BD Influx cell sorters housed within the NIHR Cambridge BRC Cell Phenotyping Hub were used and have the laser configuration shown. The BD Influx does not have a UV laser (355 nm). Figure taken from http://www.med.cam.ac.uk/nihr-cambridge-brc-cell-phenotyping-hub/aria-fusion/.

2.2.2 Compensation controls for cell sorting

Compensation beads (BD Bioscience Comp Beads) were used to set up the compensation settings of the cell sorter. The beads bind mouse-derived antibodies and contain both a positive and negative fraction. To 100 µl of PBS/0.2% BSA one drop of positive and one drop of negative compensation beads were added, followed by vortexing. Each antibody used in the sorting panel (generally 2 µl) was added as a single stain to the compensation beads and incubated at room temperature for 10 minutes. The tubes were topped up with 2 ml of PBS/0.2% BSA, vortexed and spun at 600 x g for 5 minutes. The supernatant was poured off, the tube flicked and the beads resuspended in 500 µl of PBS/0.2% BSA.

2.2.3 DNA extraction and bisulfite conversion

DNA methylation is measured by bisulfite sequencing; this is where DNA is treated with sodium bisulfite resulting in all cytosine nucleotides that are not methylated being converted to uracil (read as thymine during PCR), whereas methylated cytosine nucleotides are protected from bisulfite conversion. Bisulfite conversion was performed using the Qiagen Epitech Fast lyse all kit; this method processes up to 100,000 cells in a one step cell lysis and bisulfite conversion. Bisulfite-converted DNA was quantified using a nanodrop spectrophotometer with the single strand DNA setting. For all steps lo-bind DNA tubes (Eppendorf) were used to reduce the loss of DNA from adsorption onto the tube wall, especially from low cell input samples.

2.2.4 Targeted measurement of DNA methylation

I developed a method to measure the level of methylation at the *FOXP3* Treg Specific Demethylated Region (TSDR, Chr X:49260689-49260861) using high-throughput sequencing, which has been published as a technical comment in the European Journal of Immunology[91] and a schematic of the workflow is shown in Figure 2.2. In summary the method is based on a two stage PCR of bisulfite converted DNA and uses deep sequencing to count the ratio of methylated (read as cytosine) and demethylated (read as thymine) nucleotides at nine different CpG sites within the *FOXP3* TSDR. A Treg is classified as containing eight or nine demethylated CpG sites within the *FOXP3* TSDR. I have developed similar assays to measure methylation of regions within *CTLA4*, *CD4*, *IFNG*, *IL17A*, *RORC* and *HPRT*. The region in *CTLA4* shares strong homology with the orthologous region in mouse that was identified as having preferential demethylation within Tregs[93]. The regions targeted in *CD4* and *HPRT* I designed based on epigenetic evidence from the Roadmap Epigenomics project[40]. The genes *IFNG*, *IL17A* and *RORC* have been previously identified as being specifically demethylated in Th1 and Th17 cells for which they are lineage defining markers using pyrosequencing[94].

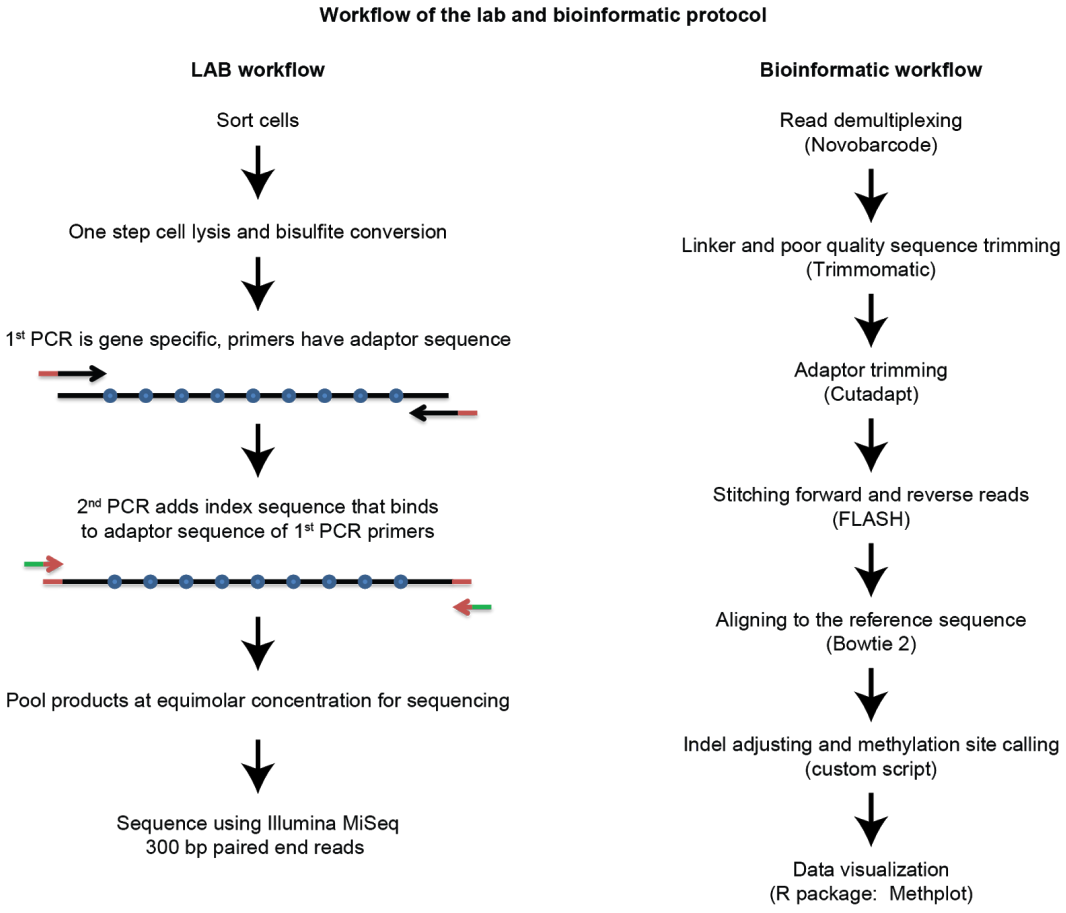


Figure 2.2. Schematic of targeted methylation workflow. I produced this figure which has been published as Supplemental Figure 8[REF91].

Primers were designed to flank these regions (based on bisulfite converted DNA) using Primer3[REF5–7] and the primers were synthesised (Sigma Aldrich) and are listed in Table 2.2, where lowercase indicates sequence complementary to the second round primer and uppercase is specific to the gene of interest. The first round PCR consisted of 10 µl of Qiagen multiplex master mix, 0.4 µl 10 µM forward primer, 0.4 µl 10 µM reverse primer and 4 µl of bisulfite DNA brought up to 20 µl with nuclease-free water. I have multiplexed up to six PCR primer sets in the same reaction, to do this I added 0.4 µl 10 µM forward primer and 0.4 µl 10 µM reverse primer of each primer set and reduced the volume of nuclease-free water needed to give a total volume of 20 µl. The PCR cycling

conditions start with a 15 minute 95ºC activation step followed by 30 cycles of 95ºC for 30 seconds, 60ºC for 90 seconds and 72ºC for 60 seconds and concludes with a 72ºC for 10 minutes final step. The long annealing time is needed to allow efficient binding of all PCR primers in the multiplex PCR. All samples were run in quadruplicate, and the first round PCR products were cleaned up with Ampure XP beads (Beckman Coulter) which allows removal of unincorporated primer and dNTPs. The second round PCR uses a set of 480 primers which each contain a unique 9 bp index, a sequence complementary to the adaptor sequence added to the first round primers as well as Illumina compatible ends. The second round PCR products were pooled at equimolar concentration and the library quantified using the Kapa library quantification qPCR kit, before being sequenced on an Illumina MiSeq using v3 reagents with 2 x 300bp paired end reads.

The pipeline used for data analysis was developed by Xin Yang (PhD Thesis, Cambridge) and further functionality was added by Oliver Burren (DIL) and has been described in detail[91] and summarised in Figure 2.2. The bioinformatic analysis starts by demultiplexing the FASTQ data files using Novobarcode (www.novocraft.com), generating a unique file corresponding to each of the index sequence used. Trimmomatic[95] was used to remove adaptor sequence and trim the read based on a quality score of 20, Cutadapt[96] removed the Illumima adaptor sequence, FLASH[97] stitched pair end reads together and Bowtie2[REF98] was used to align each read to a library containing the expected amplicon sequences. The pipeline is available at https://github.com/ollyburren/Methpup. The pipeline outputs the number of reads with the same methylation status at all of the CpG sites within the amplicon tested, which are also plotted with an R script developed by Xin Yang (PhD Thesis, Cambridge) and examples of all the amplicons tested are shown in Figure 2.3.

Table 2.2. Primer sequence for methylation.

Gene	Primer sequences	Size (bp)	Number of CpG sites
FOXP3	For - tgtaaaacgacggccagtTTGAATGGGGGATGTTTTTG Rev - caggaaacagctatgaccAAAATATCTACCCTCTTCTCTTCCTC	173	9
CTLA4	For - tgtaaaacgacggccagtGTGGTTTAGTTTGTTGTGGTATTG Rev - caggaaacagctatgaccCTCCACCTTACAAATATAAAATCCC	279	7
CD4	For - tgtaaaacgacggccagt TGATGGGGGAAGTTATTGTTATTT Rev - caggaaacagctatgaccTACCCCCAAACTCTATCTAACACC	242	4
IFNG	For - tgtaaaacgacggccagtGAGGAGGTGTAGTATATTGTTTTGA Rev - caggaaacagctatgaccTTCAATCATTTTCAACCACAAAC	301	3
IL17A	For - tgtaaaacgacggccagtAATGGTGTTATTTTTGAATTTATTG Rev - caggaaacagctatgaccACCAAAACTTTCATTTCCTATCCT	318	10
RORC	For - tgtaaaacgacggccagtAAGTATTGGGGGAGAGAGTTAGGT Rev - caggaaacagctatgaccACTATTCCCAAAACATCCTCCTTT	243	7
HPRT	For - tgtaaaacgacggccagtGGGGAAGTATTGGGTTAAGAGTTAG Rev - caggaaacagctatgaccTAAAAATTTTCTCCCAAAAAAAACC	251	14

Figure 2.3. Location and examples of targeted methylation analysis. Legend is on page 31.

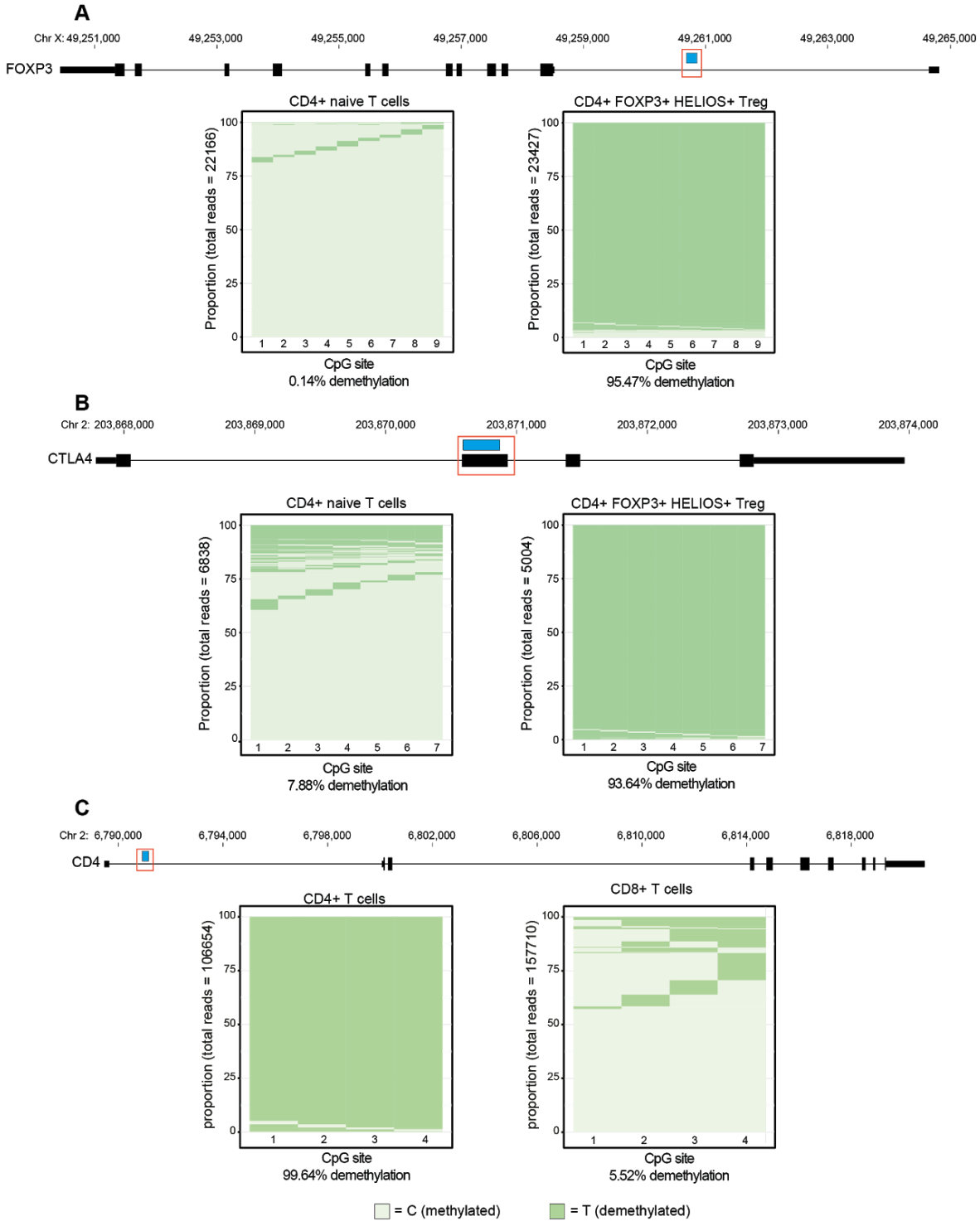
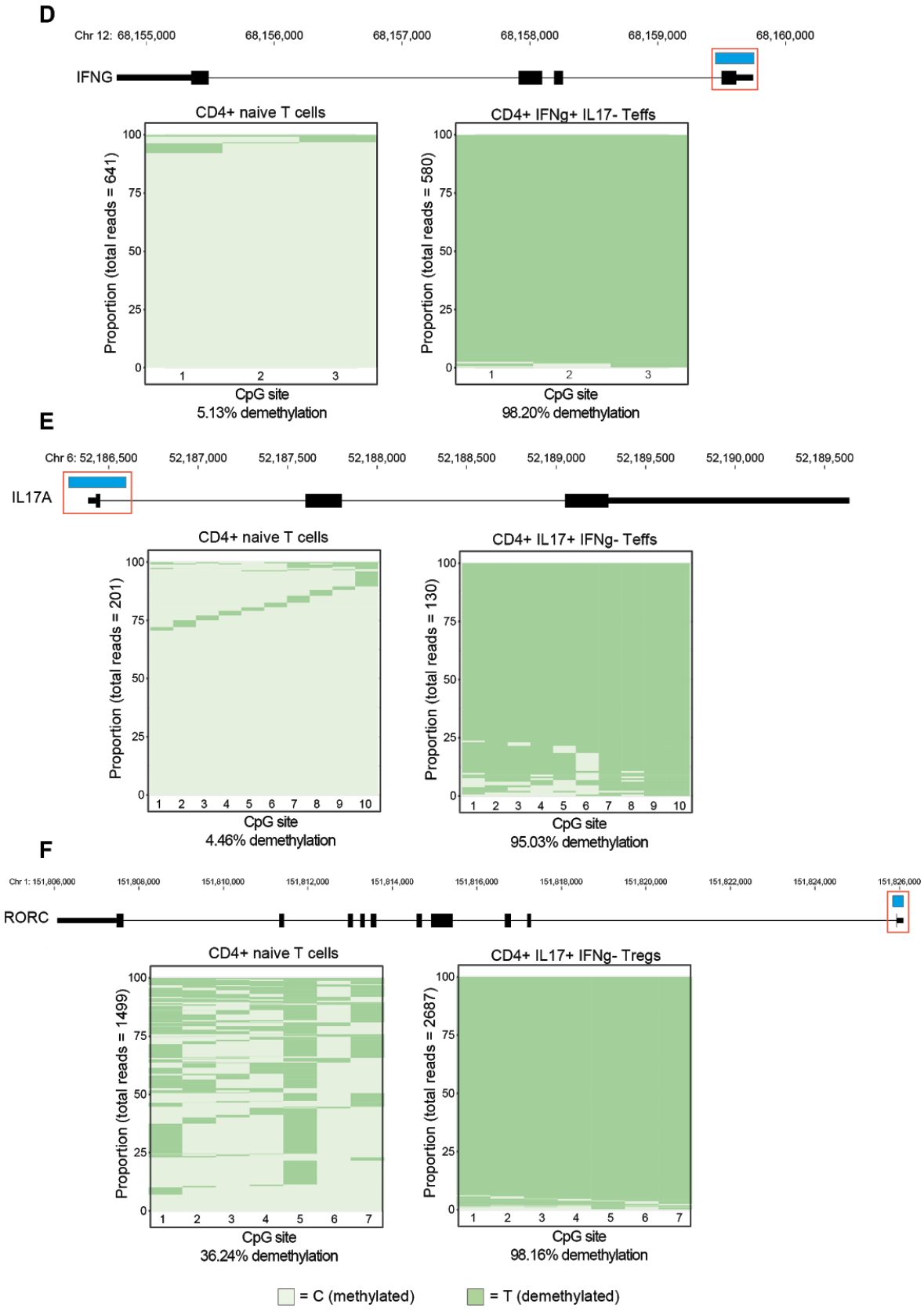


Figure 2.3. Location and examples of targeted methylation analysis. Legend is on page 31.



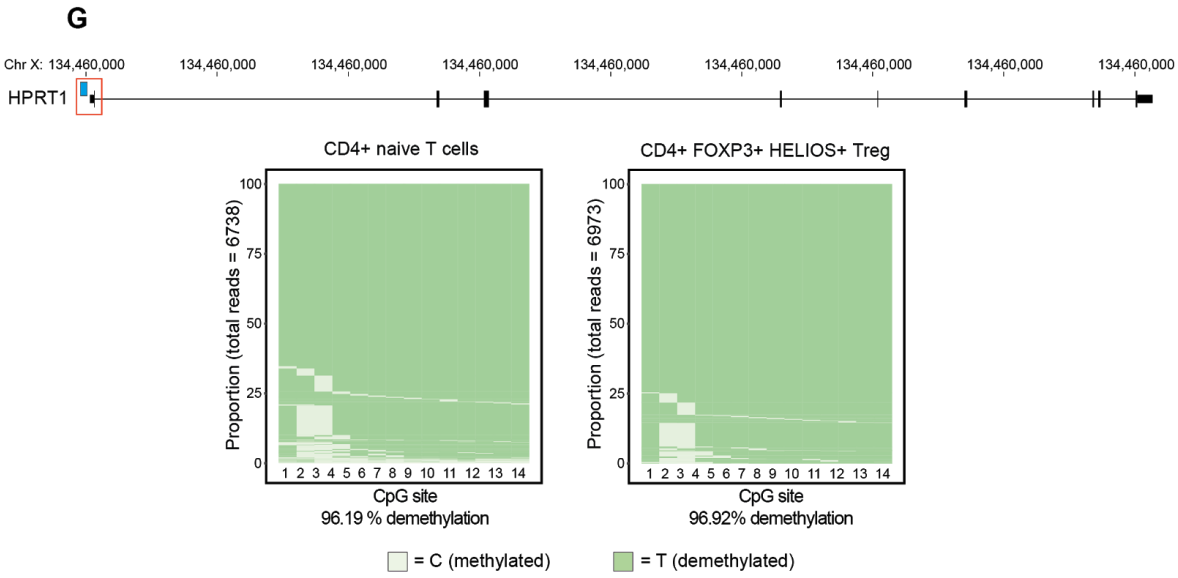


Figure 2.3. Location and examples of targeted methylation analysis. The location of the region targeted for methylation analysis is shown by the blue bar and red box for the genes (A) *FOXP3* (B) *CTLA4* (C) *CD4* (D) *IFNG* (E) *IL17A* (F) *RORC* and (G) *HPRT*, coordinates are based on GRCh38. Below the gene structure, are two plots which display the most methylated and demethylated cell type analysed, with the left hand plot from naive or total CD4 T cells and the right hand plot differing in each assay. The x axis displays the CpG sites interrogated for each gene and each row on the y axis displays the status of each CpG site within a single read (the reads are stacked) with light green representing a C (methylated) and dark green representing a T (demethylated) CpG site. (G) HPRT is essential for all cells and is ubiquitously expressed and is therefore demethylated in all cell types, and is included as an additional gene located on the X chromosome.

2.2.5 Generation of a standard curve to normalise TSDR methylation levels

A standard curve of methylated and demethylated DNA was generated by sorting FOXP3$^-$ HELIOS$^-$ non-Tregs (which are fully methylated at the TSDR) and FOXP3$^+$ HELIOS$^+$ Tregs (which are fully demethylated at the TSDR) in various ratios from 0% FOXP3$^-$

HELIOS$^-$ non-Tregs : 100% FOXP3$^+$ HELIOS$^+$ Tregs at 10% increments to 100% FOXP3$^-$ HELIOS$^-$ non-Tregs : 0% FOXP3$^+$ HELIOS$^+$ Tregs. Cells at various ratios of FOXP3$^-$ HELIOS$^-$ non-Tregs and FOXP3$^+$ HELIOS$^+$ Tregs were sorted into the same tube and bisulfite conversion of the DNA was performed, which generated a mixture of *FOXP3* TSDR methylated and demethylated DNA at the same ratio as the mixed cells.

2.2.6 Sorts based on Cytokine production

The sorting of cells based on cytokine production was performed by Ricardo Ferreria (DIL), using previously published protocols[99–101]. In brief, CD4$^+$ T cells were cultured for five hours with cell stimulation cocktail (eBiosiences), which contains phorbol myristate acetate (PMA), ionomycin, and protein transport inhibitors (brefeldin A and monensin). Cells were then stained with intracellular antibodies to IL-10, IL-21, IFNɣ and IL-2, all conjugated to PE, and IL-17 on AF700. I performed the TSDR analysis on the cell populations sorted based on their production of cytokines.

2.3 Methods relating to Chapter 5, 6 and 7: *IL2RA*

2.3.1 Recall of genotyped individuals

Samples were obtained from the Cambridge BioResource (www.cambridgebioresource.org.uk) and recalled based on a specific genotype. All samples have been genotyped using the Immunochip genotyping array (Illumina), allowing genotype to phenotype studies to be performed. Immunochip is a custom genotyping array with 195,806 SNPs and 718 small insertion–deletions from 186 autoimmune

disease-associated loci[13], and was designed by collaborators studying 11 distinct autoimmune diseases to replicate and fine map all association signals known at the time the platform was developed. Plink[102,103] was used to extract and quality control the genotype calls from Immunochip. Donors were selected to carry a particular *IL2RA* haplotype based on a previous analysis[2,104] and the SNP genotypes used to select donors are shown in Table 2.3.

Table 2.3. The *IL2RA* haplotypes analysed and the tag SNPs that identify each haplotype. *The allele frequency was calculated from the 1000 Genomes Project[105,106] that contains 503 individuals of European ancestries.

***IL2RA* Haplotype**	**Tag SNP (selected allele)**	**Frequency in European population***	**Associated disease and direction**
A	rs61839660 (T) rs12722495 (C)	5.2%	T1D, MS, AITD, PSC - Protective Allergic disease - Susceptible CD, PS - Susceptible
A^{-660}	rs61839660 (C) rs12722495 (C)	0.8%	None
C	rs11597367 (G)	29.0%	T1D, AITD - Protective
D	rs62626317 (G)	16.2%	MS - Protective
E	rs6602437 (T)	53.6%	T1D - Protective
F	rs41295121 (T)	0.7%	T1D - Protective

2.3.2 Cell sorting for allele-specific expression

Cell sorting via fluorescence-activated flow cytometry was used to purify immune cell subsets from PBMCs that had been cryopreserved. Frozen PBMCs (10-15 million) were thawed at 37°C in a waterbath, gently swirling the vial until only a small amount of frozen

material was seen. All steps were carried out at 4ºC or on ice. The contents of the vial was transferred to a 15 ml Falcon tube (Corning) and topped up drop by drop with RPMI 1640 (Thermo Fisher) supplemented with 10% AB serum (Sigma). The empty vial was rinsed out with media and added to the 15 ml Falcon tube; this was then spun at 600 x g for 5 minutes to pellet the cells. The supernatant was poured off and 4 ml of RPMI 1640 with 10% AB serum added and a cell count was performed with trypan blue using a haemocytometer to check cell viability. The cells were transferred to a polypropylene FACS tube (Corning) and spun at 600 x g for 5 minutes. The supernatant was poured off and the cells resuspended in 100 µl PBS (Thermo Fisher) with 0.2% BSA (Sigma) and 50 µl of brilliant stain buffer (BD Biosciences) .

The cells were stained at 4ºC for 30 minutes with antibodies to the surface proteins listed in Table 2.4. Two different CD25 antibodies with the same fluorochrome were used to increase sensitivity. After staining the tubes were topped up with 3 ml of PBS (no BSA) and spun at 600 x g for 5 minutes to wash out the BSA as it binds the viability dye. Viability dye (Zombie Aqua, Thermo Fisher) was diluted 1:1000 with PBS and added to the cells and incubated for 10 minutes at 4ºC. The cells were spun at 600 x g for 5 minutes, the supernatant poured off and the pellet resuspended at approximately 10 million cells per ml in PBS 0.2% BSA. The sample was passed through a sterile 30 µM filter (CellTrics, Sysmex) to remove any clumps of cells to prevent the cell sorted becoming blocked.

Table 2.4. Antibodies used for sorting cell populations for *IL2RA* ASE experiments.

Marker	Clone	Fluorochrome	Supplier	µl per 15 million PBMC
CD14	M5E2	BV421	BioLegend	7.5
Viability dye Aqua	L34957	BV510	Thermo Fisher	1
CD27	O323	BV605	BioLegend	7.5
CD4	RPA-T4	BV650	BioLegend	7.5
CD45RA	HI100	BV785	BioLegend	7.5
CD19	HIB19	AF488	BioLegend	7.5
CD3	UCHT1	PerCP/Cy5.5	BioLegend	7.5
CD56	HCD56	PE	BioLegend	7.5
HLA-DR	L243	PE dazzle	BioLegend	7.5
CD127	eBioRDR5	PE-Cy7	eBioscience	7.5
CD25	M-A251 2A3	APC APC	BD Biosciences BD Biosciences	30 7.5
CD16	3G8	AF700	BioLegend	7.5
CD8	RPA-T8	APC/Cy7	BioLegend	7.5

Up to eight different immune cell populations were sorted (Table 2.5, Fig. 5.9) using either a BD Influx or BD FACSAria Fusion cell sorter (Fig 2.1). The compensation controls for cell sorting were the same as described in section 2.2.2. The cells were sorted into 1.5 ml lo-bind DNA/RNA tubes (Eppendorf) and for T cells contained 500 µl X-VIVO 15 (Lonza) supplemented with 1% AB serum, and for B cells, NK cells and monocytes 500 µl RPMI 1640 supplemented with 1% AB serum. The lo-bind DNA/RNA tubes were used to

reduce the adsorption of RNA onto the tube once the cells were lysed. A proportion of the sorted cells were pelleted at 600 x g for 5 minutes, lysed in 300 µl RNA lysis buffer (Zymo Research) and stored at -80ºC until used for RNA extraction. The remaining cells were activated using the conditions listed in Table 2.5 for 4, 16 or 40 hours; the number of time points analysed depended on the yield of the sort, with approximately 100,000 cells used per time point. After culturing, the cells were harvested, spun at 600 x g for five minutes and lysed in 300ul of RNA lysis buffer (Zymo Research).

Table 2.5: The marker characteristics of the immune cell subsets sorted for *IL2RA* ASE. An example flow cytometry gating strategy is shown in Figure 5.9.

Cell type	Sorted as:	Stimulation
$CD4^+$ naïve T cells	$CD3^+$ $CD19^-$ $CD8^-$ $CD4^+$ $CD127^+$ $CD45RA^+$ $CD27^+$	0.3 - 1 CD3/CD28 T cell activation bead per cell
$CD4^+$ memory T cells	$CD3^+$ $CD19^-$ $CD8^-$ $CD4^+$ $CD127^+$ $CD45RA^-$ $CD27^+$	0.3 - 1 CD3/CD28 T cell activation bead per cell
CD4+ Tregs	$CD3^+$ $CD19^-$ $CD8^-$ $CD4^+$ $CD127^{low}$ $CD25^+$	Not tested
$CD8^+$ T cells	$CD3^+$ $CD19^-$ $CD4^-$ $CD8^+$	0.3 - 1 CD3/CD28 T cell activation bead per cell
NK cells	$CD3^-$ $CD19^-$ HLA-DR^- $CD56^+$	IL-12/IL-15/IL-18
NK $CD56^{bright}$	$CD3^-$ $CD19^-$ HLA-DR^- $CD56^{++}$ $CD16^{low}$	Not tested
NK $CD56^{dim}$	$CD3^-$ $CD19^-$ HLA-DR^- $CD56^+$ $CD16^+$	Not tested
B cells	$CD3^-$ $CD19^+$	2.5 µM CpG
$CD14^+$ monocytes	$CD3^-$ $CD19^-$ $CD56^-$ HLA-DR^+ $CD14^+$ $CD16^-$	IFNɣ and GM-CSF
$CD16^+$ monocytes	$CD3^-$ $CD19^-$ $CD56^-$ HLA-DR^+ $CD14^-$ $CD16^+$	IFNɣ and GM-CSF

2.3.3 RNA extraction and cDNA synthesis

The RNA was extracted from the stored cell lysates in 300 µl RNA lysis buffer using a Zymo Research Quick RNA micro kit following the manufacturer's instructions. To summarise, the RNA lysis buffer-containing lysates were thawed and allowed to reach room temperature and vortexed to ensure that they were fully resuspended. All centrifuge spins were performed in a microfuge at 10,000 x g. An equal volume (300 µl) of 100% ethanol (Sigma) was added to the RNA lysis buffer, the tubes vortexed and then briefly spun to remove drops from the lids. The contents was transferred to a Zymo-Spin IC column and spun for 30 seconds. The column was washed with 400 µl of RNA wash buffer and spun for 30 seconds. An on-column DNA digestion was performed to remove all genomic DNA from the sample using 5 µl of DNase I and 35 µl of DNA digestion buffer and incubated at room temperature for 15 minutes. The column was washed with 400 µl of RNA Prep buffer, followed by two washes with 700 and 400 µl, respectively, of RNA wash buffer. The final spin was for two minutes to remove all traces of ethanol, before the RNA was eluted in 15 µl of nuclease-free water.

Synthesis of cDNA was performed using Superscript III reverse transcriptase (Thermo Fisher). 13.2 µl of RNA was added to 1.2 µl of 10 mM dNTPs (Bioline) and 1.2 µl of 0.5 µM oligo dT (18T, Sigma) and incubated at 65ºC for 5 minutes before quenching on ice. A master mix was made containing 1.2 µl Superscript III, 1.2 µl RNase out, 1.2 µl DTT and 4.8 µl of 5x Superscript buffer (all from Thermo Fisher); 8.4 µl of the master mix was added to each cDNA synthesis reaction and transferred to a Thermocycler for one hour at 50ºC followed by 15 minutes at 70ºC to heat inactivate the reverse transcriptase.

2.3.4 *IL2RA* mRNA expression

To measure the expression of *IL2RA* mRNA, quantitative-PCR (qPCR) was performed. As the main experiment to be performed on the cDNA was allele-specific expression, 4 µl of cDNA was diluted into 8 µl of water. This 1:2 dilution of the cDNA would generate enough cDNA for qPCR, and would only result in a decreased cycle threshold (Ct) of 1.6 cycles compared to neat cDNA. qPCR assays to beta-2-microglobulin (*B2M*, housekeeping gene) and *IL2RA* were purchased from Thermo Fisher and run using a 20 µl reaction in a 384 well plate (Thermo Fisher) with the following recipe: 10 µl qPCR master mix (Thermo Fisher), 7.3 µl water, 0.7 µl qPCR primer/probe mix and 2 µl of cDNA, with samples being run in duplicate. The Quant Studio 7 qPCR instrument (Thermo Fisher) was used with the following cycling conditions: 50ºC for 2 minutes, 95ºC for 10 minutes, then 40 cycles of 95ºC for 15 seconds and 60ºC for 1 minute. Data were analysed using QuantStudio Real-Time PCR software, and the mean Ct for *IL2RA* being subtracted from the B2M Ct to generate the ΔCt, which is the normalised expression value.

2.3.5 *IL2RA* Allele-Specific Expression

To measure the ASE of *IL2RA* a modified version of the protocol developed to examine targeted bisulfite sequencing was used[91]. The DNA sequence for each SNP analysed was downloaded from Human Ensembl GRCh38 (http://www.ensembl.org/Homo_sapiens) and Primer3[REF107–109] was used to design primers with standard primer picking conditions and amplicon length set to 100-250 bp. To ensure that no known SNPs were present within the primer sequences, since this could affect the primer binding to DNA and alter the ASE results, the program SNPCHECK (https://secure.ngrl.org.uk/SNPCheck/snpcheck.htm)

was used. Primers used in this study are shown in Table 2.6. The 5′ end of the first round primer contains an adaptor sequence which allows binding of the second round primers, shown as lowercase in Table 2.6.

Table 2.6. Details of PCR primers used for allele-specific experiments.

SNP	Primer sequence	Size (bp)	Location
rs12244380	For - tgtaaaacgacggccagtTTCGTGGGAGTTGAGAGTGG Rev - caggaaacagctatgaccTTAAAAGAGTTCGCTGGGCC	180	*IL2RA* 3′ UTR
rs12722495	For - tgtaaaacgacggccagtGTGAGTTTCAATCCTAAGTGCGA Rev - caggaaacagctatgaccATTAAGCGGACTCTCTGGGG	97	*IL2RA* Intron 1
rs12722606	For - tgtaaaacgacggccagtTAACACTCTCCTGCCCCTTC Rev - caggaaacagctatgaccATTTCGGTGGAAGTCCTGGT	165	*IL2RA* Intron 1
rs2104286	For - tgtaaaacgacggccagtCTCCCTTTCCATGTCACCCT Rev - caggaaacagctatgaccTTGGTGAGGAGGAGAAAGGC	116	*IL2RA* Intron 1
rs12722497	For - tgtaaaacgacggccagtGAGGAAAAAGCCTGGGAAAC Rev - caggaaacagctatgaccGGCAAGATACGGATCAGGAG	121	*IL2RA* Intron 1
rs2476491	For - tgtaaaacgacggccagtAACAGGCCCTGATTTCCTTT Rev - caggaaacagctatgaccGGCAATGGATATTCACTGTGG	191	*IL2RA* Intron 1
rs8463	For - tgtaaaacgacggccagtGGTCATGTTTCTACTGGCAAAA Rev - caggaaacagctatgaccGGCAATGAACCTCAGTACCAA	217	*RBM17* 3′ UTR

The first round PCR contains 10 µl of Qiagen multiplex master mix, 0.5 µl containing 10 nmol forward and reverse primers and 5 µl of cDNA brought up to 20 µl with nuclease-free water. Most analyses were performed as a duplex reaction with two *IL2RA* SNPs being examined in the same PCR reaction. First-round PCR cycling conditions were 15 minute 95ºC hotstart, followed by 30 cycles of: 95ºC for 30 seconds, 60ºC for 90 seconds, 72ºC for 60 seconds, finishing with a 72ºC for 10 minute cycle, with each sample being amplified in quadruplicate.

The first-round PCR product was cleaned up to remove unincorporated primers with Ampure XP beads (Beckman Coulter). To the 20 µl PCR reaction, 22 µl of Ampure XP beads were added and incubated for 5 minutes. A 96 well plate magnet was used to trap the beads that had bound the PCR products, and two washes with 70% ethanol were performed before the PCR product was eluted off the beads in 20 µl of nuclease-free water. Only 18 µl of sample was collected to prevent the carryover of Ampure XP beads.

A second-round PCR was performed to add Illumina sequence-compatible ends to the first round amplicons. Additionally a nine base pair unique index sequence was incorporated into the second round primer that is specific to each sample. The second-round PCR contains 8 µl Qiagen multiplex master mix, 2 µl nuclease-free water, 5.3 µl of first-round PCR product and 0.35 µl of each forward and reverse second-round primers. A list of the index sequences has been described[91]. The second-round PCR cycle consists of a 15 minute 95ºC hotstart, followed by 7 cycles of: 95ºC for 30 seconds, 56ºC for 90 seconds, 72ºC for 60 seconds, finishing with a 72ºC for 10 minute cycle.

All PCR products were quantified on a Shimadzu Multina gel electrophoresis system to allow pooling of equimolar PCR products. The pooled PCR products were purified using Ampure XP beads and eluted in nuclease-free water. The library was diluted to 10 nmol based on Qubit DNA quantification (Thermo Fisher) and accurately quantified using the Kapa Bioscience library quantification kit, which only detects Illumina compatible ends. A Thermo Fisher Quant Studio 7 real time PCR instrument was used and a dilution of 1:1000, 1:2000, 1:4000 and 1:8000 of the library was compared to a standard curve at dilutions from 20 pmol to 0.0002 pmol. The libraries were sequenced on an Illumina

MiSeq at the Stratified Medicine Core Laboratory NGS hub that is hosted by the Department of Medical Genetics, University of Cambridge using v3 reagents producing 300 bp paired-end reads. The libraries were spiked with 5% PhiX to add diversity to the sequencing run and the average output was 20 million reads.

2.3.6 Analysis of *IL2RA* ASE data

A modified pipeline that was previously used to analyse targeted bisulfite sequencing, Methpup[91], was used for the analysis of ASE data. Additional functionality was added to Methpup by Oliver Burren (DIL) (https://github.com/ollyburren/Methpup) that allowed the 'relaxed' calling of any base within the amplicon. The Methpup pipeline was originally written for analysis of methylation and only reported on sites that were either C or T, the methylated or demethylated sequence base product of bisulfite sequencing. The Methpup pipeline had already been developed and tested (Xin Yang, Cambridge PhD Thesis) and Oliver Burren (DIL) added the ability to call the base at any position within the amplicon.

2.3.7 Immunophenotyping of whole blood samples

The collection and staining of the samples was performed by the Sample Processing Team within the DIL and flow cytometric analysis was performed in FlowJo by Marcin Pekalski (DIL). Immunophenotyping was performed on fresh blood samples from 522 volunteers recruited through the Cambridge BioResource. DNA isolated from each of the participants was genotyped using the Immunochip genotyping array (Illumina), allowing genotype to phenotype studies to be performed.

For immunophenotyping using flow cytometry, 200 µl of whole blood was incubated with two panels of fluorochrome-conjugated antibodies (Table 2.7 and Table 2.8) at room temperature in the dark for 45 minutes. Red blood cells were lysed with BD FACS Lysing Solution (which also fixes the remaining blood cells) and the samples were washed in PBS/0.2% BSA, and then analysed on a BD Fortessa flow cytometer (details of lasers and filters in Fig. 2.4) using FACSDiva software (BD Biosciences). FlowJo software (FlowJo LLC) was used to analyse the immunophenotyping data.

Table 2.7. Antibodies used for immunophenotyping of $CD4^+$ naive, central memory and Tregs, $CD8^+$ T cells and NK cells.

Marker	Clone	Fluorochrome	Supplier	µl per 200 µl of blood
CD4	RPA-T4	AF700	BioLegend	2
CD8	RPA-T8	APC-Cy7	BioLegend	2
CD25	M-A251 2A3	APC APC	BD Biosciences BD Biosciences	25 5
CD127	eBioRDR5	PE-Cy7	eBioscience	5
CD45RA	HI100	BV785	BioLegend	2
CD62L	DREG-56	BV605	eBioscience	5
CCR10	314305	PE	R&D Systems	10
CXCR3	G025H7	PerCP-Cy5.5	BioLegend	5
CCR6	G034E3	AF488	BioLegend	5
CCR4	IG1	BV421	BD Biosciences	5
CD56	HCD56	BV711	BioLegend	5

Table 2.8. Antibodies used for immunophenotyping of B cells.

Naive tube Marker	Clone	Fluorochrome	Supplier	µl per 200 µl of blood
CD25	M-A251 2A3	APC APC	BD Biosciences BD Biosciences	25 5
CD8	RPA-T8	APC-Cy7	BioLegend	2
CD45RA	HI100	BV785	BioLegend	2
CD21	Bu32 B98	PE PE	BioLegend BD Biosciences	5 20
CD127	eBioRDR5	PE-Cy7	eBioscience	5
CD95	DX2	PerCP-Cy5.5	BioLegend	5
CD31	WM-59	FITC	eBioscience	5
CD27	M-T271	AF700	BioLegend	5
CD4	RPA-T4	BUV395	BD Biosciences	2
CD3	OKT3	BV510	BioLegend	5
CD56	HCD56	BV711	BioLegend	5
CCR7	G043H7	BV421	BioLegend	2
CD62L	DREG-56	BV605	eBioscience	5

2.3.8 Normalisation of immunophenotyping data

It took 708 days to collect immunophenotyping data from the 522 participants. At day 430 an essential upgrade was required on the BD Fortessa flow cytometry instrument that was used to collect immunophenotyping data resulting in different instrument baseline settings that meant the data needed to be processed as two different batches (pre and post upgrade). Additionally a decline in CD25 fluorescence (on the APC fluorochrome) was observed

between day 1 and day 430, and a slight increase in CD25 fluorescence from days 430 to 708 regardless of genotype. Therefore, I needed to normalise the fluorescence seen pre and post Fortessa upgrade allowing it to be combined into a single dataset. The method used to normalise this dataset is described in Chapter 5.3.5.

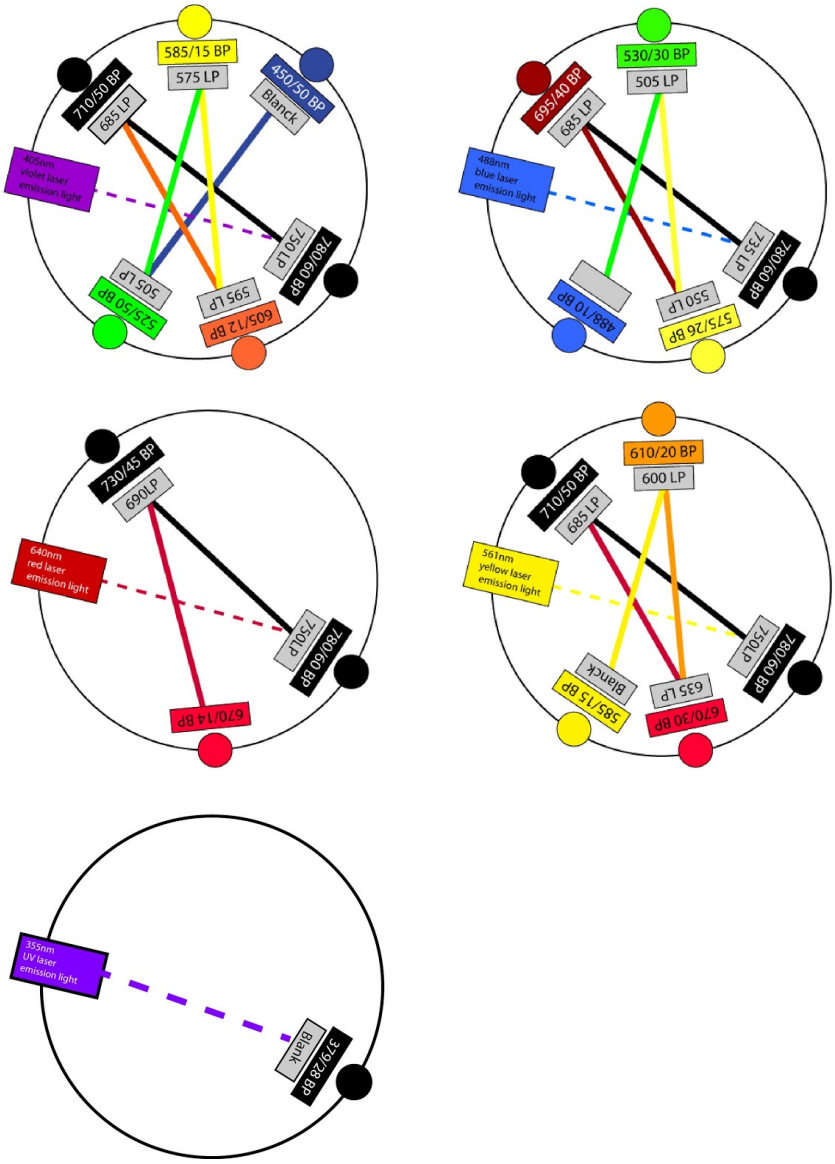


Figure 2.4. Laser and filter configuration of the BD Fortessa flow cytometer. Figure modified from http://www.med.cam.ac.uk/nihr-cambridge-brc-cell-phenotyping-hub/cytometry-fortessa/.

2.3.9 Targeted methylation analysis of *IL2RA*

To measure the methylation status of specific CpG sites around the IL-2RA gene, a targeted bisulfite sequencing approach was undertaken, which followed the same protocol as described for *FOXP3* TSDR analysis in section 2.2.4. The primers listed in Table 2.9 were used to amplify the region containing the CpG sites of interest, and the percent demethylation was calculated for each amplicon. Where a SNP was contained within the amplicon, allele-specific methylation could be performed with the SNP allele being called along with the DNA sequence of each CpG site, and the percent methylation was calculated for all sequencing reads carrying the same SNP allele.

Table 2.9. Primers used for methylation analysis around *IL2RA*. The number of CpG sites sequenced in each primer set is listed, and in parentheses indicates if the SNP creates a CpG site.

Name	Primer sequences	Size (bp)	No. of CpG sites
IL2RA alt-1 660	For - tgtaaaacgacggccagtTTTAGAGTAGTTTGGGGGTATT Rev - caggaaacagctatgaccTCTCTCTAAAAACCTTCCCAAAAA	300	8
rs11597367	For - tgtaaaacgacggccagtGGTTGGTTTTAGATTTTTGGATTT Rev - caggaaacagctatgaccCTTCATACATAATAATTAAACACCCATA	247	3 (4)
rs62626317	For - tgtaaaacgacggccagtTGTATTGGATGTTTTTGTTAGGTTT Rev - caggaaacagctatgaccTAAACAAAACTCCCCATTCCTAAA	223	1
rs62626317 _Proximal	For - tgtaaaacgacggccagtAAATGAATTATTGTAGGATAGTTTGGA Rev - caggaaacagctatgaccAATAACACAACCCACAAATCAACA	397	7
rs62626317 _Distal	For - tgtaaaacgacggccagtTGTTAGAATTTGTGAGGAGAGTTAAGA Rev - caggaaacagctatgaccCAAAAACATCCAATACATAAACAACA	351	8
rs41295055	For - tgtaaaacgacggccagtGGTTAGGGAGTTTTGTTGTTATTTTT Rev - caggaaacagctatgaccTCTAATTCCCAAACCACTATTAACTT	245	2
rs6602437	For - tgtaaaacgacggccagtTTGTGAGATATTTGGTTTTGTTTTG Rev - caggaaacagctatgaccCACCCACACATACATACATACACA	146	(1)
IL2RA-10	For - tgtaaaacgacggccagtGGTAATAAAGTGTGGGTGGTATG Rev - caggaaacagctatgaccCTAATCCAATTACACCAACTTATTCT	245	7
IL2RA Prom-2	For - tgtaaaacgacggccagtTGGATGGATTTATAAGGGTGATAGT Rev - caggaaacagctatgaccCCTTACCTACCTAACAACCAAACAC	222	9

2.3.10 Phylogenetic analysis of the *IL2RA* region

The 1000 Genomes Project has whole-genome sequenced 2,504 individuals from five continental populations, African, American, East Asian, European and South Asian[105,106], Fig. 2.5, and identified genetic variants within the sequence data. The phase 3 release of the 1000 Genomes Project contains fully phased haplotypes from 2,504 individuals[105].



Figure 2.5. The location of the 26 different populations from five world regions that form the 1000 Genomes Project. Approximately 100 individuals (2504 individuals in total) from each population were whole genome sequenced and variants identified. Figure reproduced from http://www.1000genomes.org/.

To identify if there was any common ancestry between the *IL2RA* A, C, D, E and F haplotypes from European individuals identified from the 1000 Genomes Project, I downloaded the phased haplotype information from the *IL2RA* region (GRCh37 Chr10: 6,020,000-6,250,000). Most phylogenetic analysis programs use DNA sequence as input and I converted the phased genotypes into DNA sequence using the Ensembl reference sequence as the backbone sequence and replaced each SNP position with the phased haplotype allele of the variant. I tested BEAST[110], PHYLIP (http://evolution.genetics.washington.edu/phylip.html) and Mega-X[111] programs for phylogenetic analysis and they all generated similar outputs, and Mega-X was used to generate Figure 7.1.

2.3.11 1000 Genomes Project genotype comparison across the *IL2RA* region

To assess how similar European individuals from the 1000 genomes projects carrying the same A, C, D, E, F and susceptible haplotype were to each other, the A, C, D, E, F and susceptible haplotype tags SNPs (Table 2.3) were used to identify which haplotype was present on each phased chromosome. The susceptible haplotype was defined as carrying none of the A, C, D, E or F protective alleles. I calculated the frequency of the major allele at each SNP on the A, C, D, E, and susceptible haplotypes; the major allele for each haplotype is the most common allele, and each haplotype has a different combination of major alleles. A score of zero means all individuals are homozygous for the major allele. The maximum score of 50 means there is an equal proportion of the major and minor alleles in the haplotype, which could result from all donors being heterozygous for the SNP, or a Mendelian distribution of alleles (1 AA : 2 Aa : 1 aa). Scores between 0 and 50 reveals a gradient of increasing segregation. This method of analysis can be applied to any group of individuals tagged by a variant of interest to assess how homogenous the population carrying the variant is compared to the population carrying the alternate allele.

Comparison of how similar the A, C, D, E, F and susceptible haplotypes are to each other was performed by comparing the frequency of the major allele in the first haplotype to the frequency of the same allele found in the comparator haplotype, with a score of zero showing the major allele is homozygous in the first haplotype and comparator haplotype, whereas the maximum score of 100 shows the comparator haplotype is homozygous for the minor allele found in the first haplotype. A score between 0 and 100 indicates there is segregation occurring and the degree of segregation is reflected by the size of the score.

For example a score of 60 could be achieved by the first haplotype being homozygous and the comparator having the minor allele present at 60%; a score of 60 is also achievable when there is segregation in both haplotypes, for example if the first haplotype had the minor allele present at 20% and the comparator had the minor allele present at 80%.

2.3.12 Genetic association of the A, A^{-660} and susceptible haplotypes with T1D

This analysis was performed by Dr Chris Wallace (University of Cambridge). Genotype data from the Immunochip SNP microarray platform[13] on 15,557 controls and 6,681 T1D cases were extracted for the region surrounding *IL2RA*[17] (chr10:6030243-6169685, build 36). Data were prephased using SHAPEIT[112] and imputed to 1000 Genomes (phase 3) using IMPUTE2[ref113]. Haplotypes of group A SNPs and other T1D associated tag SNPs found in Wallace *et al*.[2] were phased using an EM (expectation-maximization) algorithm. Ten multiply-imputed datasets were generated from the converged model and T1D haplotype risk analysed under a logistic regression model, with inference aggregated across the multiply-imputed datasets to allow for uncertainty in phasing. These steps are implemented in the R package snpStatsHaps (https://github.com/chr1swallace/snpStatsHaps). Effects of the A and A^{-660} haplotypes on T1D risk were compared by setting A as reference in the logistic regression.

Chapter 3

Evaluation of *FOXP3* TSDR measurement

3.1 Introduction and aims

Regulatory T cells (Tregs) are an essential component of the immune system and maintain peripheral tolerance. Therefore accurate methods to measure the frequency of Tregs is critical, and classically, a $CD4^+$ Treg as been defined by the high expression of CD25, the low expression of CD127 and the constitutive expression of the transcription factor FOXP3[REF66–68]. However, these same markers can also identify an activated T effector that has transiently upregulated the expression of FOXP3[REF114,115]. Additionally, any functional analysis requires viable Tregs which excludes the use of intracellular staining protocol required for FOXP3. The constitutive expression of FOXP3 seen in Tregs is dependent on the demethylation of a CpG dense conserved region within intron 1 of *FOXP3* named the Treg Specific Demethylated Region (TSDR)[82,83,88]. This region binds the transcription factors CREB/ATF, ETS-1, NF-κB[116,117], and STAT5[REF 118–120] as well as FOXP3 itself to maintain transcription[121]. Therefore, the specific demethylation of the *FOXP3* TSDR has been used as a method to identify and validate a Treg population as being thymically derived, as opposed to a Treg population induced through TGF-b that has a methylated TSDR[122]. Furthermore the ratio of demethylated to methylated DNA measured at the TSDR has been used as a way to enumerate the frequency of $FOXP3^+$ Tregs[87].

Recently I have developed a method to analyse at single base resolution the methylation status of each CpG dinucleotide in the *FOXP3* TSDR using high-throughput sequencing[91]. My method uses a two stage PCR, the first stage targets the *FOXP3* TSDR and the second round primers add a unique barcode to each sample, allowing up to 960 samples to be processed per sequencing run (Fig. 2.2, Methods). I have optimised the protocol and reproducible results can be obtained from as little as 10,000 cell input[91], which is essential given the often limited sample from clinical trials, as well as the low frequency of Tregs (~5% of $CD4^+$ T cells). This method has only been tested on Tregs sorted using markers to surface proteins and therefore the aims of this chapter are to:

1. Test my *FOXP3* TSDR protocol on cells sorted with the use of an intracellular stain to FOXP3 in order to identify $FOXP3^+$ cells.
2. Understand the limitations and considerations of my method to measure the methylation of the *FOXP3* TSDR.

3.2 Survey of *FOXP3* TSDR demethylation within $CD4^+$ T cells subsets

To enable accurate enumeration of $FOXP3^+$ Tregs using TSDR demethylation relies on the specificity of the TSDR demethylation to Tregs. There has not been a detailed characterisation of the TSDR status of all FOXP3 expressing cells in the peripheral blood; most Treg studies are based on the surface expressed molecules CD127 and CD25. Miyara *et al*[84] showed that cells expressing the highest levels of CD25 were the most demethylated at the TSDR, which correlated well with FOXP3 expression, and were the most likely to be Tregs. However, in the CD25 intermediate expressing population there

was a mixture of cells, those that were fully demethylated at the TSDR and those that were fully methylated at the TSDR indicating that the cells sorted as CD25 intermediate contained at least two distinct subpopulations. I have repeated this gating strategy and sorted purified $CD4^+$ T cells defined by CD45RA and FOXP3 (Fig. 3.1A) and have reproduced the original finding that cells expressing the highest levels of FOXP3 show the greatest demethylation at the TSDR. As the level of FOXP3 expression decreases then so does the proportion of cells that are fully demethylated at the TSDR, with essentially all cells not expressing FOXP3 being fully methylated (Fig. 3.1B-C). A higher proportion of cells expressing intermediate levels of FOXP3 were demethylated in memory $CD4^+$ T cells as compared to those expressing CD45RA.

As *FOXP3* is located on the X chromosome I had to take into consideration that in females one X chromosome is inactivated and the majority of genes (85%) are silenced[123] and methylated and the active X chromosome is demethylated, whereas a male has only one X chromosome, which is active. Consequently, the measured TSDR demethylation values of the two female individuals analysed in Figure 3.1 were approximately half that of the male donor and were doubled to allow the values to be compared to the male, . However, for the brightest FOXP3 expressing memory Tregs, this results in values of 106.2% and 110.8% for the female participants, whereas the male participant value was 98.8% (Fig. 3.1 C). Doubling of the female *FOXP3* TSDR values results in values of more than 100% demethylated, which implies there is an overestimation of demethylation in my assay methodology which needs to be understood and if possible corrected.

The three mechanisms that I investigated to account for the overestimation were (1) the cell cycle, (2) leaky X chromosome and (3) a bias in the polymerase used in the PCR.

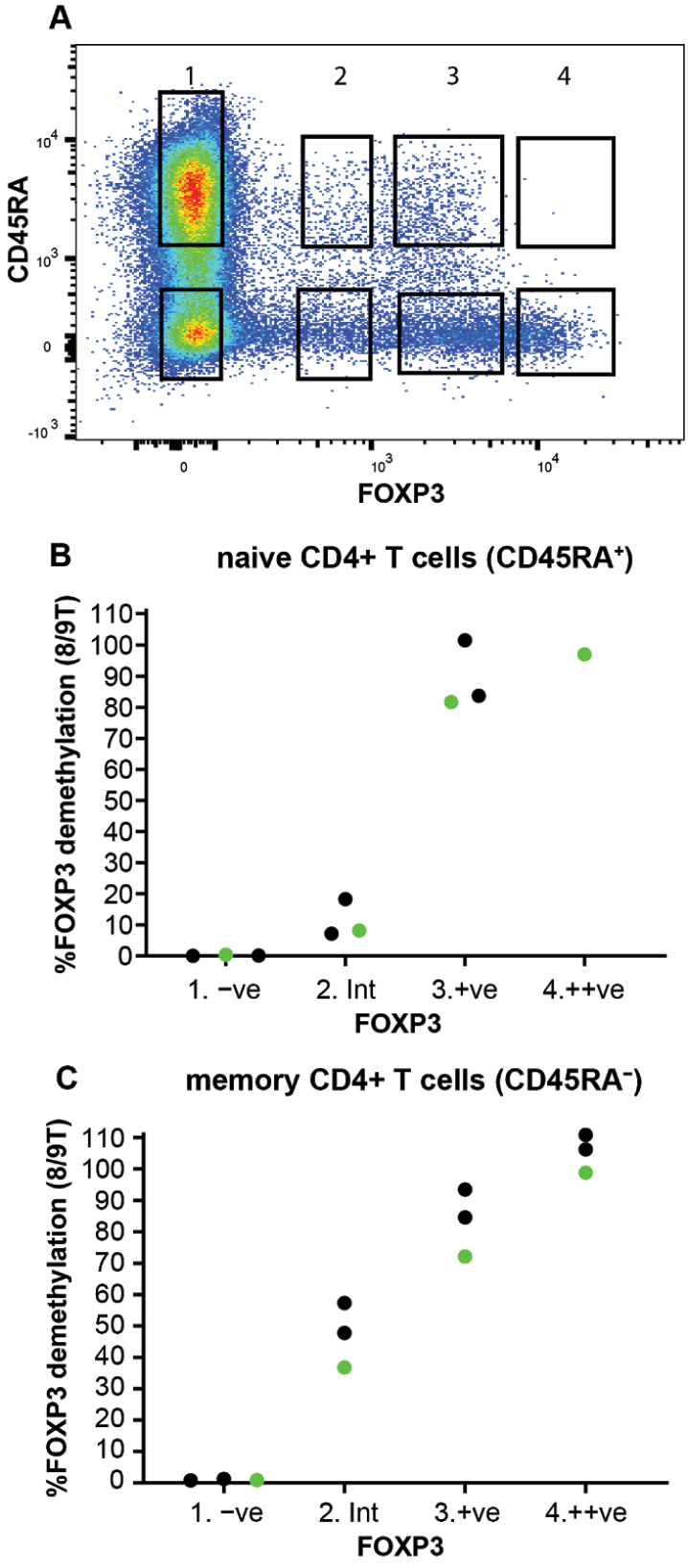


Figure 3.1. FOXP3 TSDR measurement in cells sorted by FOXP3 and CD45RA. (A) Gating strategy used to purify cells based on the expression of FOXP3 and CD45RA from $CD4^+$ T cells. The boxes indicate the four populations sorted based on the level of FOXP3 expression (1 = negative, 2 = intermediate, 3 = bright and 4 = very bright), which is the same strategy used in Miyara *et al*[84]. FOXP3 TSDR methylation levels were measured in purified naive (B, $CD45RA^+$) and memory (C, $CD45RA^-$) $CD4^+$ T cell populations separated by the expression of FOXP3 from three healthy participants. The green circle represents a male participant and the black circles are two female participants, for which the *FOXP3* TSDR demethylation values were doubled to take into account the inactive X chromosome.

3.3 Understanding the overestimation of *FOXP3* TSDR measurement

3.3.1 Does the cell cycle affect *FOXP3* TSDR measurement?

As a cell goes through division, the DNA is first replicated and the pattern of methylation on CpG dinucleotides is not maintained and the newly synthesised DNA is unmodified[124,125]. The DNA methyltransferase DNMT1 is responsible for the maintenance of DNA methylation and ensures the methylation pattern of the parental DNA is copied to the progeny strand[126]. Therefore there is a time during cell replication where there is increased global DNA demethylation, and if a reasonable number of cells are in cell cycle, that could account for the increased TSDR demethylation seen in female donors. This would not affect male donors as there is only one copy of the X chromosome and in Tregs would be fully demethylated all of the time.

To address the question of whether cell proliferation altered the level of DNA demethylation seen at the *FOXP3* TSDR, I sorted Tregs using the proliferation marker Ki-67. Ki-67 is a nuclear protein that is expressed in proliferating cells, for which a monoclonal antibody is available[127], enabling intracellular staining and sorting of Ki-67 expressing cells. From three female participants, I sorted proliferating mTregs ($CD4^+$ $CD25^+$ $CD127^{low}$ $CD45RA^-$ $FOXP3^+$ $HELIOS^+$ $Ki\text{-}67^+$) and non-proliferating mTregs ($Ki\text{-}67^-$); approximately 20% of the mTregs were in cell cycle (Fig. 3.2A). There was no difference in the level of demethylation measured at the *FOXP3* TSDR between proliferating ($Ki\text{-}67^+$) and non-proliferating ($Ki\text{-}67^-$) mTregs (54.0% vs 54.2%, Fig. 3.2B). Additionally, naive Tregs were analysed and none were found to be in cell cycle ($Ki\text{-}67^-$),

but showed the same level of TSDR demethylation proliferating mTregs (53.2% vs 54.0% Fig. 3.2B). These data together provide strong evidence that the cell cycle status of the cell does not alter the estimate of TSDR demethylation.

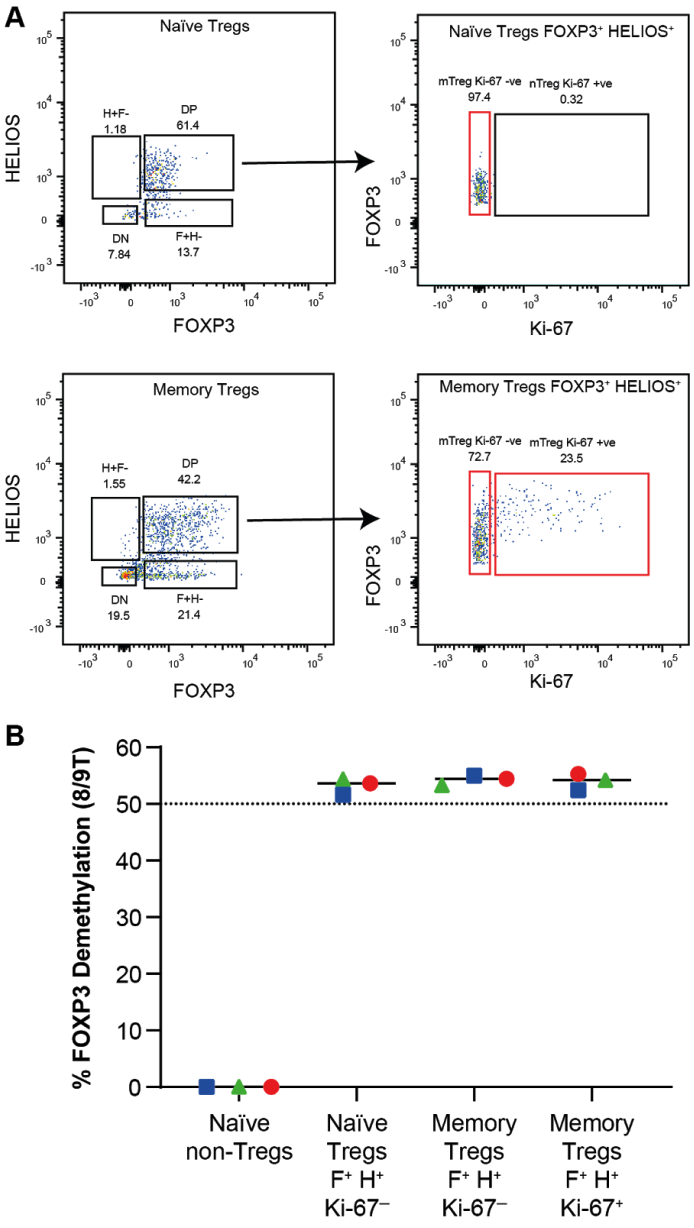


Figure 3.2. Cell cycle does not alter *FOXP3* TSDR measurement. (A) Representative flow cytometric plot starting with naive Tregs (upper panel, gated on $CD4^+$, $CD25^+$, $CD127^{low}$, $CD45RA^+$, $CD62L^+$) and memory Tregs (lower panel, gated on $CD4^+$, $CD25^+$, $CD127^{low}$, $CD45RA^-$, $CD62L^+$). $FOXP3^+$ $HELIOS^+$ cells were gated and divided based on the expression of the proliferation marker, Ki-67. The red gates are the Treg subsets analysed. (B) *FOXP3* TSDR measurement from sorted Tregs subpopulations from three female donors (blue square, green triangle and red circle). The dashed line indicates 50%, which is the maximum value that should be obtained. Naive non-tregs are included as a negative control and have a methylated *FOXP3* TSDR.

3.4.2 Does 'leaky' X chromosome affect *FOXP3* TSDR measurement?

A second possible explanation for the overestimation of TSDR demethylation in females is that the *FOXP3* gene on the X chromosome is 'leaky'. Approximately 10% of genes escape X inactivation due to incomplete methylation of the inactive X chromosome[128,129]. Studies that have assessed if a gene escapes X inactivation have used cell lines that display non-random X inactivation and therefore the silenced allele is known, and measure allelic expression to determine if the gene is expressed by just the active X chromosome, or the inactive X chromosome as well [123,130]. With the development of single cell RNA sequencing, it is now possible to measure X inactivation within individual cells, and these results support the earlier findings that approximately 10% of genes display some degree of escape from X inactivation [131,132].

To measure if *FOXP3* does escape X inactivation is a difficult problem to address as I am working with a gene that is specifically demethylated in Tregs and methylated in all other cell types. I would like to measure the methylation of different parts of the X chromosome, but the active X chromosome will display a demethylation profile restricted to the cell type that a particular gene is expressed in. I failed to find any other X-linked genes that might have a Treg restricted expression profile, instead I looked for genes that are expressed in all cell types and therefore should be demethylated in all cell types including Tregs. The gene hypoxanthine phosphoribosyltransferase 1 (*HPRT1*) is ubiquitously expressed[44,133] and as such, is used as a housekeeping gene in some qPCR experiments[134].

Although *HPRT1* is located 85 Mb distal of *FOXP3* and will not show if any effects are due to the local chromatin landscape. I looked for a region within the *HPRT1* gene that could be subject to alterations in DNA methylation using epigenetic data from the Roadmap Epigenomics program[40] and found a region in the promoter that contained 14 CpG dinucleotides (Fig 2.3G, Methods). As there is also no evidence that *HPRT1* escapes X chromosome inactivation in humans[132], one copy of *HPRT1* should be methylated (inactive) and one copy demethylated (active) in female individuals. I designed primers to this region and assessed the methylation in both male and female donors, and found it to be fully demethylated in all immune cell types examined from a male donor (99.7% demethylation averaged over all cell types assessed), whereas the female donors showed a bimodal methylation profile with approximately half the sequencing reads coming from the methylated inactive X chromosome and half from the active X chromosome (Fig 3.3A).

When sorted Tregs were analysed, the level of demethylation of *HPRT1* was very similar to that of the *FOXP3* TSDR, with approximately 55% demethylation seen in both assays (Fig 3.3B). Given how similar the measured demethylation of *HPRT1* is to that of the *FOXP3* TSDR in highly enriched Tregs from females, I concluded that a 'leaky' *FOXP3* gene on the X chromosome was not likely leading to overestimation of TSDR demethylation, instead a technical issue with the assay could be accounting for the overestimation. Therefore, I next assessed if the DNA polymerase used in the TSDR assay was responsible for the overestimation of TSDR demethylation in female samples.

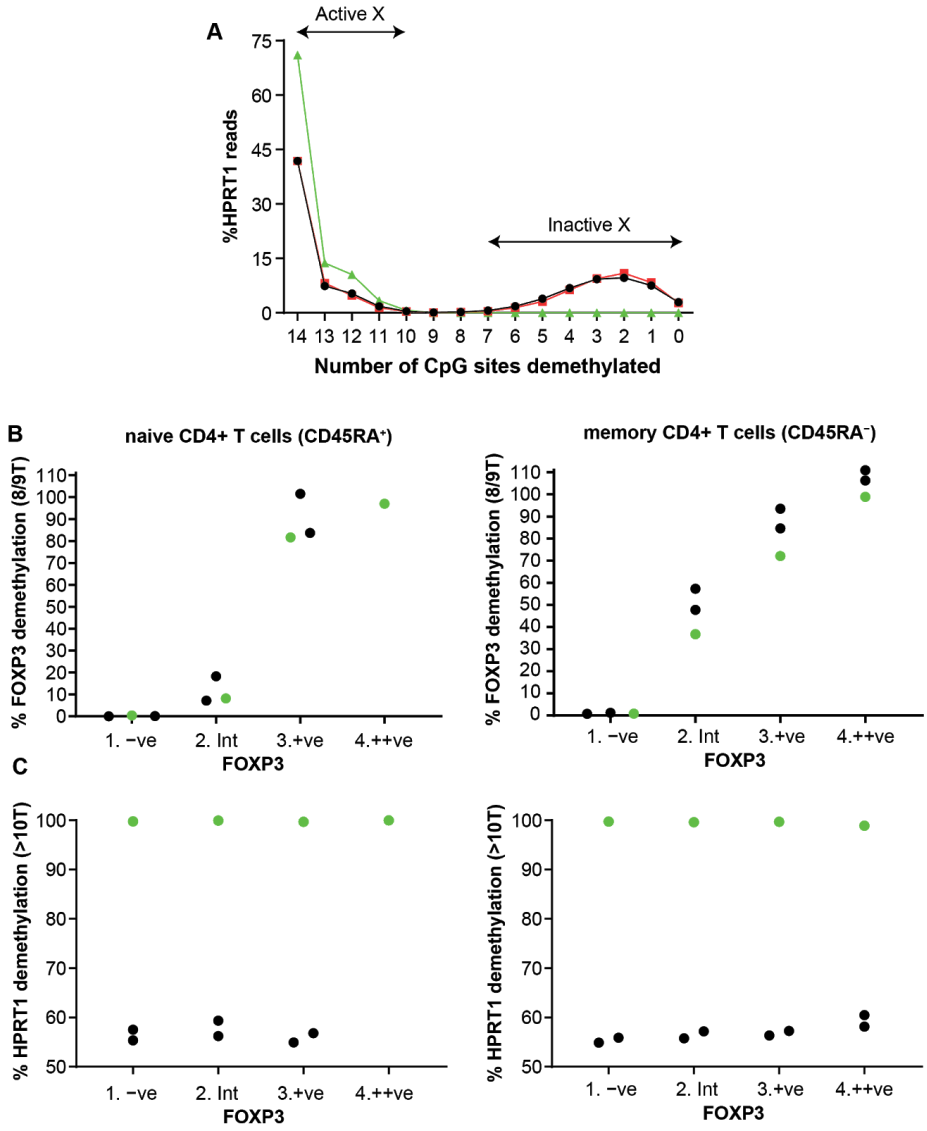


Figure 3.3. *HPRT1* shows a similar overestimation of DNA demethylation as the *FOXP3* TSDR. (A) The proportion of sequencing reads showing the indicated number of demethylation sites within each *HPRT1* read was measured from one male (green) and two female donors (red and black). The male donor had reads with 10 or more CpG sites demethylated indicating the pattern of the active X chromosome, whereas the two female donors display a bimodal pattern that includes the inactive X chromosome which is methylated (0-6 sites demethylated) and the active X chromosome (10-14 sites demethylated). (B) The data presented in Fig. 3.1B-C is repeated showing the degree of *FOXP3* TSDR demethylation. These samples have also been analysed for *HPRT1* demethylation (C) and *HPRT1* is fully demethylated in all cell subsets tested. Female show approximately 55% demethylation, very similar to that of the TSDR in Tregs (values before doubling, due to X inactivation).

3.4.3 Bias in DNA polymerase

Whilst performing the experiments to assess the first two possible explanations for the overestimation of TSDR demethylation, I discovered a literature on the potential bias introduced by the DNA polymerase used in PCR when amplifying bisulfite converted DNA. In a study on the methylation of ten gene promoters[135] it was shown that there was preferential amplification of unmethylated DNA (examples shown in Fig. 3.4), and they used samples with defined levels of methylation to correct for the overestimation of DNA demethylation.

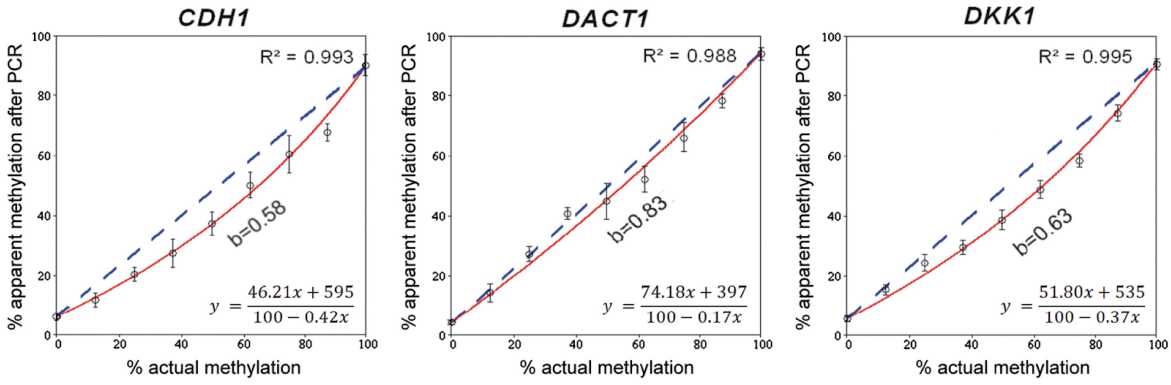


Figure 3.4. This is part of Figure 1 taken from 'Correction of PCR-bias in quantitative DNA methylation studies by means of cubic polynomial regression' that was published in Nucleic Acid Research in 2011[REF135]. The methylation of the gene promoters were measured in samples with defined levels of DNA methylation (dashed line) and compared to the actual measured methylation level (red line).

Therefore, I decided to address the *FOXP3* TSDR in a similar manner, to assess if there is a bias with the DNA polymerase that preferentially amplifies demethylated DNA. The DNA I use for the *FOXP3* TSDR analysis has been bisulfite treated, which results in all cytosine nucleotides being converted to uracil, unless they are methylated in which case they remain as cytosine. Consequently, the DNA has an unusual base composition and

contains approximately 25% adenine, 25% guanine, 25% thymine, 24% uracil and 1% cytosine. During PCR uracil is read as thymine by the DNA polymerase, but only some polymerases can accommodate uracil. I use HotStarTaq DNA polymerase (Qiagen) which can amplify DNA containing uracil. As the primers I use to assay the *FOXP3* TSDR flank 10 CpG sites, there will be 10 nucleotides different between a fully methylated and demethylated TSDR sequence and the altered composition of nucleotides could allow for bias in the PCR amplification efficiency. I sequence 10 CpG sites but only use nine in the analysis, as the most proximal CpG site does not show the same TSDR methylation profile as the other nine CpG sites, which I have previously described (published as Supplemental Figure 1 in [REF91]).

To assess if the DNA polymerase used preferentially amplifies demethylated DNA, I sorted naive T effectors and Tregs that should be fully methylated and demethylated at the TSDR, respectively, at various ratios from 100:0, 10:90, 20:80 through to 80:20, 90:10 and 0:100 as a way of generating a standard curve of methylated and demethylated TSDR DNA. If there was no bias in the DNA polymerase, then I would expect to fit a straight line through the expected *FOXP3* TSDR methylation values compared to the measured value for these mixed ratios of cells. There was a difference in the measured TSDR methylation values compared to those expected, with demethylation being overestimated, especially when the ratio was close to 50:50 (Fig. 3.5). Females have one active and one inactive X chromosome and in sorted Tregs from female donors I would expect a ratio around 50% demethylated and 50% methylated at the TSDR, and this is where I see the overestimation of TSDR demethylation from the mixed T effectors and Tregs. Therefore,

bias in the way the DNA polymerase amplifies bisDNA seems the most likely explanation for the overestimation of TSDR demethylation in purified Tregs from female donors.

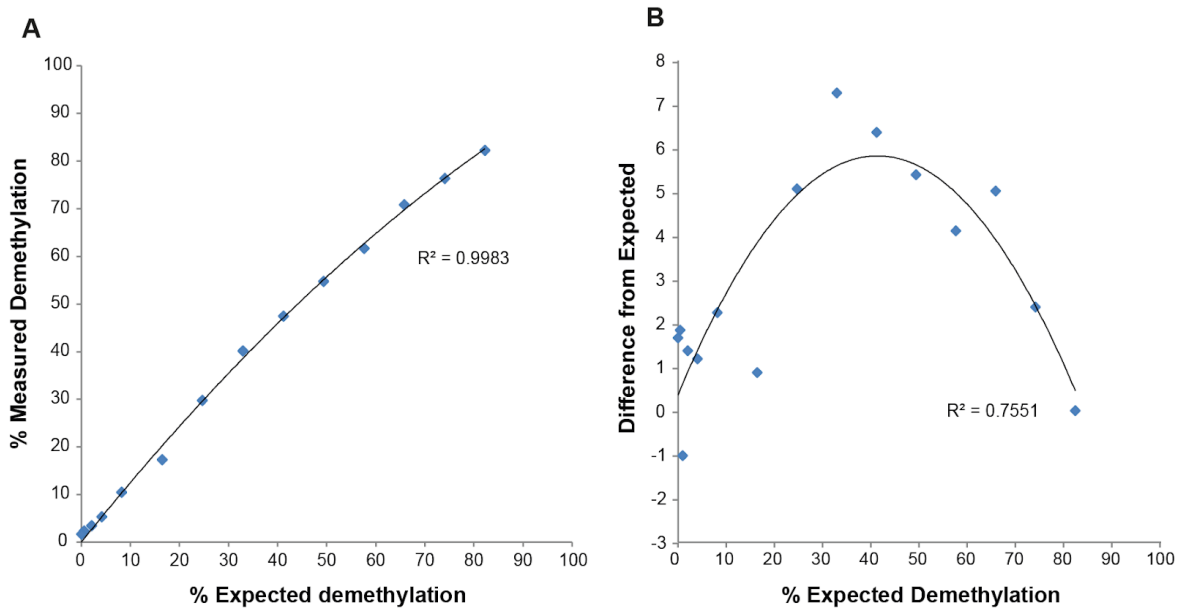


Figure 3.5. *FOXP3* TSDR demethylation is overestimated. (A) *FOXP3* TSDR methylation values from mixed Teffector : Treg populations based on surface staining, with the expected *FOXP3* TSDR methylation value based on the ratio of mixed cells shown on the x axis and the measured *FOXP3* TSDR demethylation values on the y axis. (B) To show the overestimation of demethylation values, the difference between the expected and measured demethylation value is plotted.

3.4.4 Using a *FOXP3* TSDR standard curve to correct overestimation

I have generated a standard curve of *FOXP3* TSDR measurements from a mixture of T effectors:Tregs and have shown that there is a bias in the amplification of demethylated TSDR leading to an overestimation in the level of TSDR demethylation, especially around the 50% value. I have fitted a line through the data (Fig. 3.5) and have used this equation to correct for the overestimation of *FOXP3* TSDR demethylation, using the *FOXP3* TSDR measurements from three female donors assessed for Ki-67 expression (Fig. 3.2). This

correction reduces the overestimation of the female *FOXP3* TSDR values to below 100% demethylated (Fig. 3.6). However, the standard curve and actual samples were prepared on different occasions and there could be differences in the quality of the FOXP3 intracellular stain, and in future experiments it would be optimal to prepare both at the same time to reduce any batch effects.

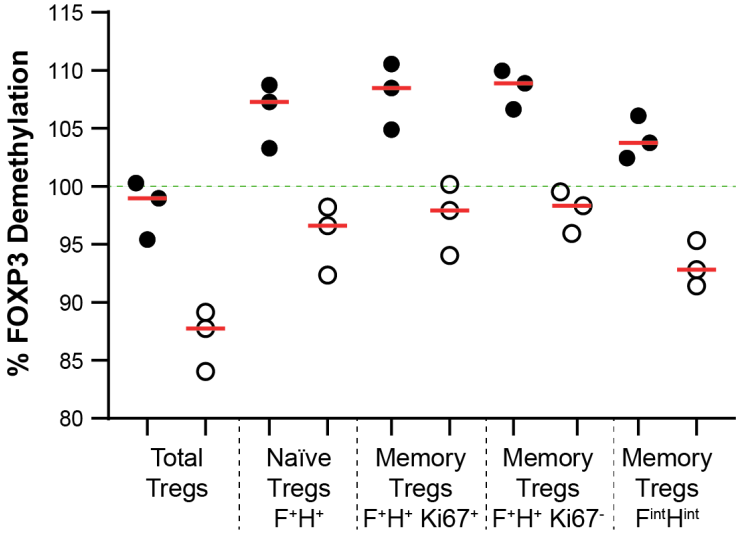


Figure 3.6. Applying a correction factor to *FOXP3* TSDR measurement reduces overestimation. Filled circles display the same data as presented in Figure 3.2 and the open circles are the same data with a correction factor applied based on the mixed Teffector:Treg standard curve presented in Figure 3.5.

3.5 Discussion

In this chapter I have shown that the methylation status of the *FOXP3* TSDR correlates with FOXP3 expression; the cell population sorted as having the highest level of FOXP3 protein also had the greatest proportion of cells with a demethylated TSDR. Cells which do not express FOXP3 have a fully methylated *FOXP3* TSDR, and cells with intermediate levels of FOXP3 contain a mixture of methylated and demethylated TSDR sequences. There is no loss in performance of the TSDR assay when using cells that have been fixed for intracellular staining, and I routinely start with 10,000 cells input.

One limitation of my *FOXP3* TSDR assay is the overestimation of TSDR demethylation in female individuals, especially when sorted Tregs are analysed as a value of 55% is often achieved. Due to X chromosome inactivation this value is doubled resulting in values greater than 100% of the cells analyzed have a demethylated TSDR, an impossible result. I have shown that cell cycle is not the cause as the TSDR methylation status is the same in both proliferating and non-proliferating cells. Secondly, I have shown that a different gene on the X chromosome, *HPRT1*, shows the same overestimation of DNA methylation meaning that escape from X inactivation is not the likely cause. Studies looking at escape from X chromosome inactivation in humans have found no evidence that *FOXP3* is affected[123,129,136], although other immune related genes *CD40L*, *CXCR3* and *TLR7* do show a degree of escape[123,137,138]. The last and most probable explanation examined was a bias in the way in which the DNA polymerase preferentially amplifies the demethylated DNA leading to an overestimation in TSDR demethylation. By using a standard curve I have shown there is a bias in the DNA polymerase and this effect is greatest around the 50% TSDR demethylation value, which is where sorted Tregs from female donors map. I have applied a correction to the data using the standard curve which reduces the bias, and this approach has been used before to correct for PCR-bias seen in bisulfite sequencing[135].

There are limited DNA polymerases available that amplify bisDNA, which need the ability to copy uracil that results from the bisulfite conversion of a demethylated cytosine nucleotide. Also, none of the current polymerases that can copy uracil are classified as high-fidelity, but in the coming months New England Biolabs will be releasing a high-fidelity polymerase capable of copying uracil (personal communication), and this new polymerase should be tested to see if it overcomes the bias seen in this TSDR assay.

Other options to remove the bias seen in this assay would be to move to an amplification free system that has been shown to be the least biased approach, but currently it is only available as a genome-wide protocol that still relies on the bisulfite conversion of DNA[139]. Direct detection of DNA methylation without the need for bisulfite conversion is possible using Nanopore[140,141] or PacBio[142] sequencing technologies and should be free from many of the biases associated with bisulfite sequencing. Neither of these methods allow for the specific targeting of the *FOXP3* TSDR, as they have been developed to generate long reads, which means they will be much more expensive than my protocol.

As I have only performed the standard curve of mixed Tregs and non-Tregs once, I plan to run the same experiment from more donors to assess how consistent the overestimation is, and if the same standard curve can be used to correct all data. Therefore all the TSDR data presented in Chapter 4 has been doubled to account for X inactivation, but has not been corrected for the overestimation of TSDR methylation resulting from PCR bias.
The technical challenges associated with measuring the *FOXP3* TSDR have been described in this chapter. In the next chapter I will utilise this assay to define the *FOXP3* TSDR methylation status of many different $CD4^+$ T cell subsets to aide in their functional characterisation.

Chapter 4

Using the *FOXP3* TSDR to assess Treg heterogeneity

4.1 Introduction and aims

Regulatory T cells (Tregs) can be divided into two broad categories depending if they originated in the thymus (tTregs) or were induced in the periphery (iTregs). HELIOS, which is encoded by the gene *IKZF2*, is a transcription factor that is expressed by approimately 60% of Tregs and has been proposed as a marker of thymically-derived Tregs[143]. However, this observation has been questioned[144–147] and the HELIOS expression was suggested to be a consequence of the experimental conditions. Induced Tregs are derived outside of the thymus from $CD4^+$ naive T cells[148,149] and are phenotypically similar to tTregs, with a major exception of being methylated at the *FOXP3* TSDR[122,150]. *In vitro*, induced Tregs are generated in response to TCR stimulation, TGF-β and IL-2.

The demethylation of the *FOXP3* TSDR is now widely assessed using qualitative methods, but little work has been performed to determine in a precise, quantitative manner the methylation status of the *FOXP3* TSDR in Treg and non-Treg cell subsets in order to determine how specific *FOXP3* TSDR demethylation is to classically-defined Tregs defined by FOXP3 protein expression and particular intracellular and cell surface molecules including HELIOS and CD25.

Therefore the aims of this chapter are to:

1. To survey all $CD4^+$ $FOXP3^+$ cell populations for the level of TSDR demethylation, and in particular determine if demethylation at the TSDR is restricted to the classical $CD4^+$ $CD25^+$ $CD127^{low}$ Treg gate.
2. To survey Tregs that express both FOXP3 and HELIOS to assess how homogeneous the population is.
3. To measure the methylation of the FOXP3 TSDR from experimental samples, in order to gain further insight into the cell population studied.

4.2 Survey of FOXP3 and HELIOS expressing cells

In order to complete a comprehensive survey of *FOXP3* TSDR demethylation, I first gated cells based on surface markers CD45RA, CD127 and CD25, that would divide cells into the typically-defined subsets of naive T effectors, memory T effectors, $CD127^{low}$ $CD25^{low}$ T cells and $CD127^{low}$ $CD25^{high}$ Tregs. From each of these populations I also sorted cells based on the expression of FOXP3 and HELIOS, and an example sort strategy is shown in Figure 4.1.

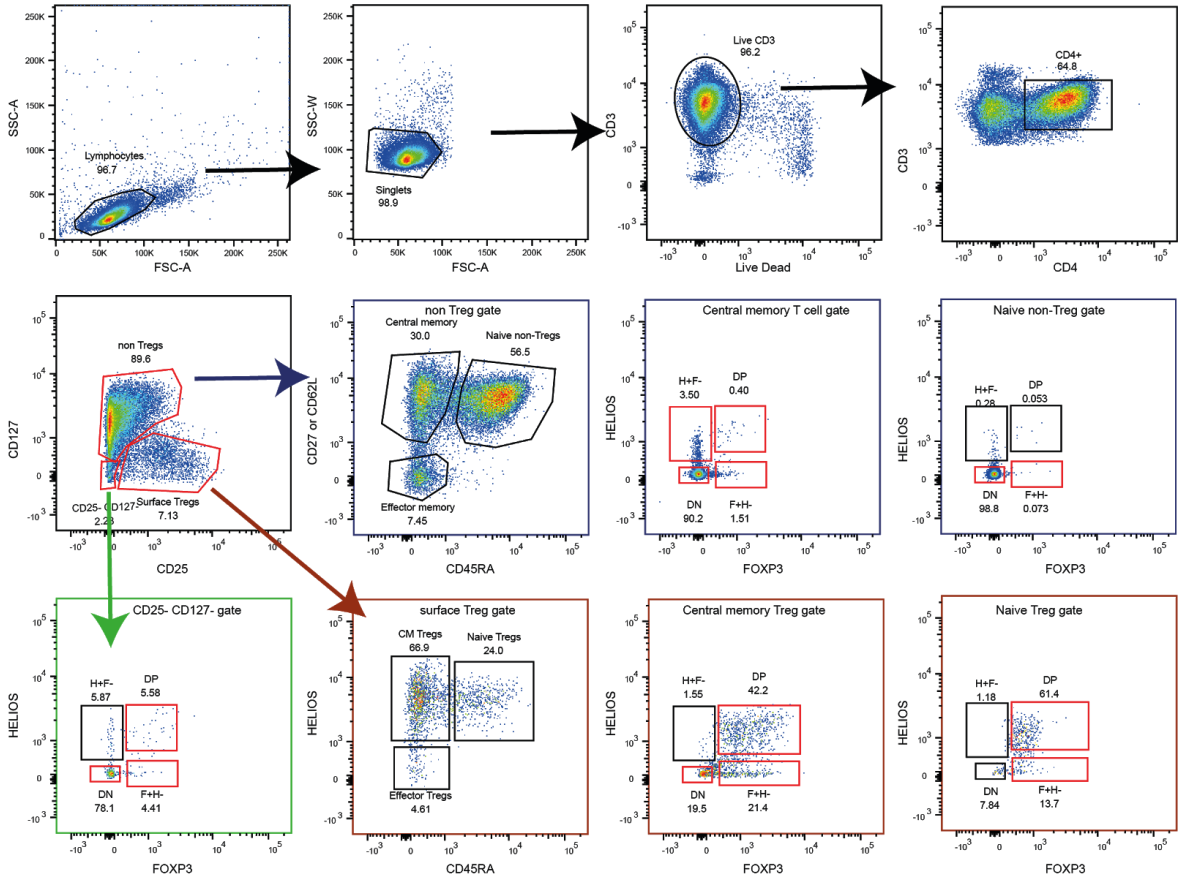


Figure 4.1. Gating strategy to purify CD4$^+$ T cell subsets based on the expression of FOXP3 and HELIOS. CD4$^+$ T cells were enriched from whole blood using a RosetteSep CD4$^+$ T cell enrichment kit (Stem Cell) and stained with the antibodies listed in Table 2.1. A BD Aria Fusion flow sorter was used to purify cells based on the expression of CD25 and CD127 followed by FOXP3 and HELIOS. Red gates indicate cell populations where both *FOXP3* TSDR methylation and cell frequency was measured, black gates indicate cell populations where cell frequency was measured.

Analysis of the *FOXP3* TSDR in these sorted cell populations showed that in any cell population expressing both FOXP3 and HELIOS protein essentially 100% of the cells have a demethylated TSDR, regardless of where it lies within the classical CD25 / CD127 surface gate (Fig. 4.2), with the largest populations located within the Treg and CD25^{low} CD127^{low} gates as well as a rare population within the central memory T effectors (less than 0.2% of CD4$^+$ T cells, Fig. 4.2). There is a similar population within the naive T

effector gate, although this population was too rare to sort and measure. I have backgated this rare population of FOXP3 and HELIOS double positive naive T effector cells and they are very close to the boundary between Tregs and non-Tregs using CD25 and CD127 gating, and it is therefore likely that this rare population represents naive Tregs with relatively low levels of CD25. It has been previously reported that there are $FOXP3^{+}$ cells within the $CD25^{low}$ $CD127^{low}$ T cell gate, the frequency can be variable and has been shown to be increased in autoimmune disease, in particular SLE[101,151–154], MS[155], RA[156] and T1D[101], but the TSDR had not been clearly defined previously as I have now done for healthy controls[101] and SLE patients (unpublished).

Cell populations single positive for FOXP3 show a lower proportion of demethylated cells than $FOXP3^{+}$ $HELIOS^{+}$ cells, especially outside of the surface Treg gate, where levels below 25% demethylation were observed; within the Treg gate the average level of demethylation at the TSDR was 70%. The cell populations which are negative for both FOXP3 and HELIOS are fully methylated at the TSDR, with the exception of cells within the surface Treg gate that show an intermediate methylation profile, with a wide range of values seen between the donors analysed (Fig. 4.2).

From all samples analysed any cell contains either a fully methylated TSDR or a fully demethylated TSDR, and in all the samples I have analysed I have not observed a cell population which contains TSDR sequences that contains a mixture of methylated and demethylated CpG sites. Therefore, when a cell population displays an intermediate methylation profile at the TSDR, it means that a proportion of cells contain a fully methylated TSDR and a proportion show a fully demethylated TSDR, suggesting it is a

heterogeneous population. Only the FOXP3^{+} HELIOS^{+} Treg populations are homogeneous for TSDR methylation status, with essentially all cells expressing both proteins having a demethylated TSDR. Cells expressing only FOXP3 protein are much more heterogeneous with different proportions of demethylation observed based on CD25 levels.

Also, I thought that cells within the surface Treg gate that are FOXP3^{-} HELIOS^{-} would be T effectors effectively "contaminating" the Treg population and therefore possess a fully methylated *FOXP3* TSDR, however I saw a wide range of methylation values, which indicates that a proportion of these cells contain a demethylated TSDR. In section 4.2.3 I will further investigate these cell populations to try and understand the relationship between the different TSDR methylated and demethylated cell populations.

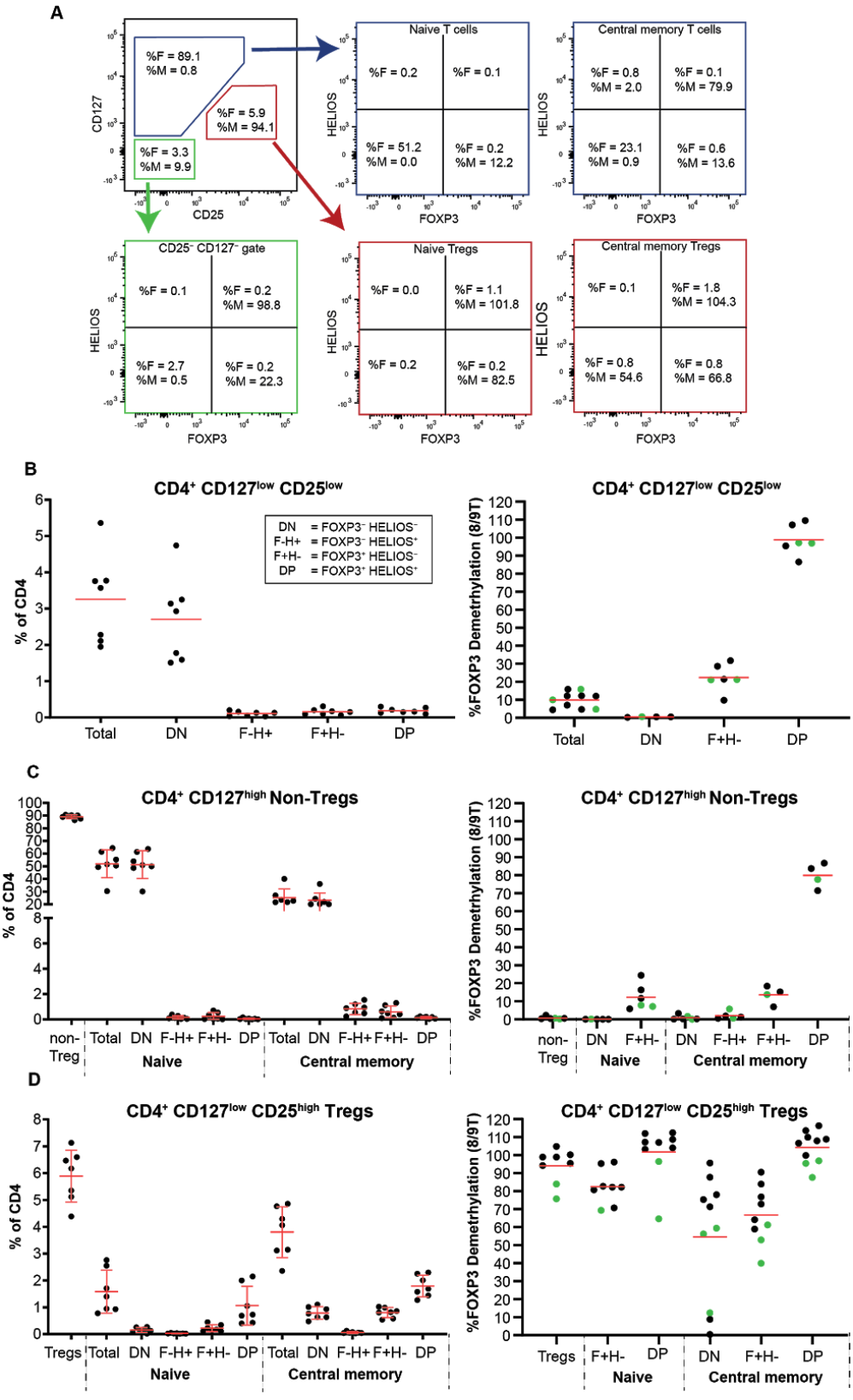


Figure 4.2. *FOXP3* TSDR demethylation in purified CD4$^+$ T cells populations. (A) The average cell frequency of CD4$^+$ T cells (%F) and percent *FOXP3* TSDR demethylation (%M) measured in different FOXP3$^{+/-}$ and HELIOS$^{+/-}$ populations. Data

presented using the same sort strategy as the representative flow plot in Fig. 4.1. (B, C, D) The same *FOXP3* TSDR data presented in (A) shown as a dot plot with the left graph displaying cell frequency out of the $CD4^+$ parent population and the right graph displaying the percent demethylation at the *FOXP3* TSDR from (B) $CD25^{low}$ $CD4^+$ T cells (C) Non-Tregs and (D) Treg subpopulations. *FOXP3* TSDR demethylation values from the female participants (black circles) were doubled to take into account the inactive X chromosome. Green circles represent the male donors. Up to ten individuals were analysed, but due to limited cell number, not all populations were analysed from all individuals.

4.2.1 Multiplex analysis of *CTLA4* and *CD4* methylation

The assay I have developed to measure TSDR demethylation is PCR based and other methylated regions of interest can be analysed at the same time with the same bisulfite DNA sample through a multiplexed PCR reaction. I have previously published a region within exon 2 of *CTLA4* that shows enrichment of DNA demethylation in Treg populations compared to non-Tregs[91], which is based on an orthologous mouse region[93]. This *CTLA4* region is not as Treg specific as the *FOXP3* TSDR, as the non-Treg populations are not fully methylated at this locus, with naive T cells and memory T cells, 10 and 20% demethylated, respectively (Fig. 4.3). I have also designed an assay to assess methylation within the *CD4* gene, to be used as a control to assess sort purity as it is specifically demethylated in $CD4^+$ T cells and methylated in all other immune cell populations (Fig. 2.3C). I have tested the samples analysed for *FOXP3* TSDR methylation in Fig. 4.2 for both *CTLA4* and *CD4* methylation and the results are displayed in Figure 4.3.

The maximum demethylation seen in the region of *CTLA4* analysed was in Tregs, and unlike *FOXP3*, *CTLA4* demethylation was much more uniform across the different Treg

subsets, with a minimum of 85% demethylation seen. In non-Tregs, only the $CD25^{-}$ $CD127^{-}$ $FOXP3^{+}$ $HELIOS^{+}$ cells displayed a demethylation profile similar to that of Tregs, again reflecting the close relationship between these populations (Fig. 4.3).

Analysis of the region assessed for *CD4* methylation reveals that all of the samples show greater than 98% demethylation and the vast majority being greater than 99.5% demethylated at CD4. This means that there is virtually no contamination from non-$CD4^{+}$ T cells within these sorted cell populations (Fig. 4.3).

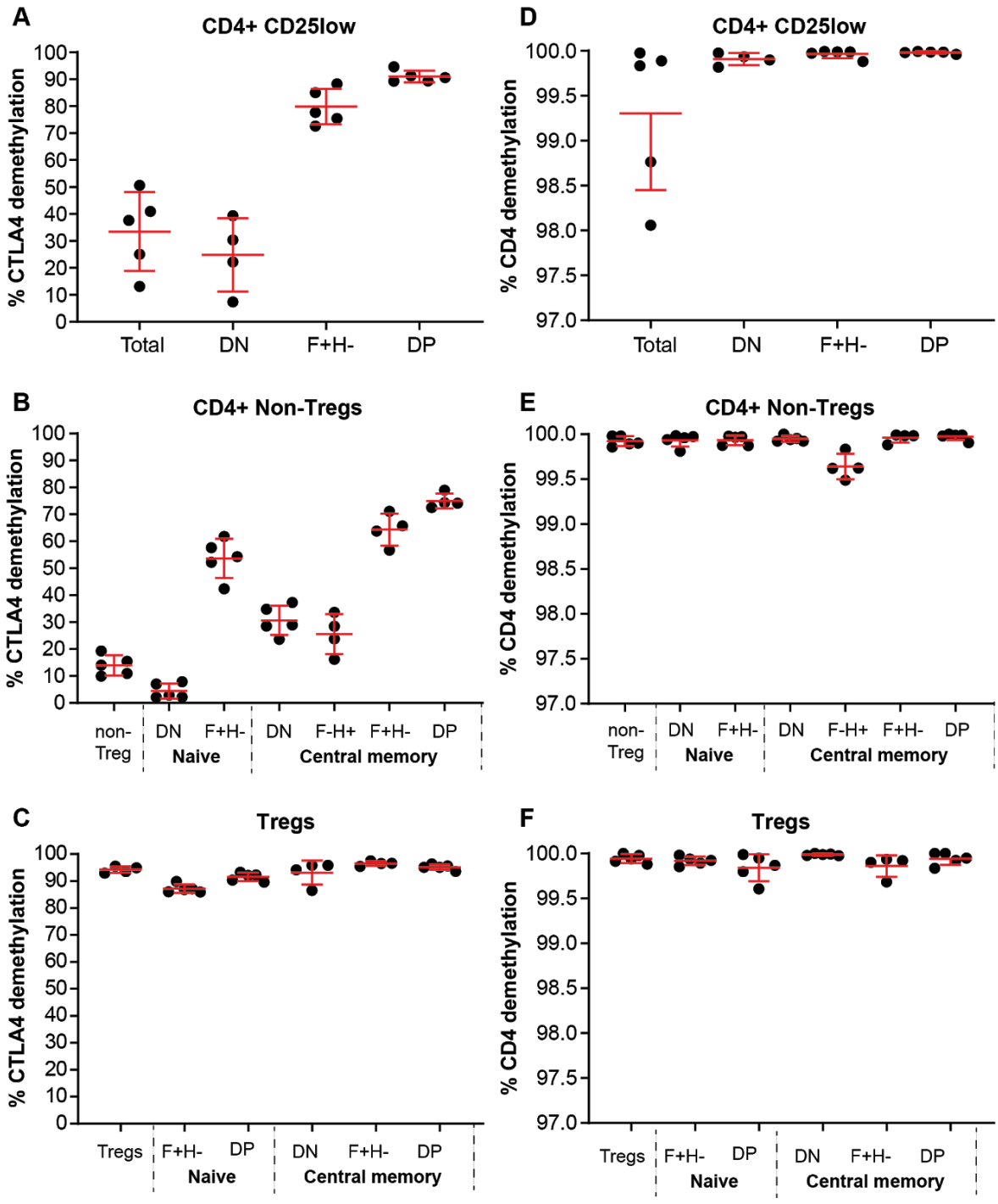


Figure 4.3. Analysis of *CTLA4* and *CD4* methylation in purified cells. Measurement of methylation within exon 2 of *CTLA4* (A-C) and intron 1 of *CD4* (D-F) in the same samples measured for *FOXP3* TSDR methylation (Fig. 4.2).

4.2.2 Reproducibility of data

Before further investigating the cells within the Treg gate that do not express FOXP3 or HELIOS but still contain a proportion of TSDR demethylated cells, I wanted to assess how reproducible the TSDR measurement was. Could any of these cell populations result from inconsistent immunostaining? The same female individual was bled 15 months apart and cells were stained and sorted using the same protocol, and the *FOXP3* TSDR measured. The data in Table 4.1 shows that most of the TSDR measurements were quite similar between bleed one and two, with the exception of the naive $FOXP3^{+}$ $HELIOS^{-}$ Tregs and memory $FOXP3^{-}$ $HELIOS^{-}$ Tregs.

Table 4.1. Reproducibility of *FOXP3* TSDR measurement from the same female participant taken 15 months apart. The values in the 1st bleed and 2nd bleed columns are the percent *FOXP3* TSDR demethylation.

Cell population	1st Bleed	2nd Bleed	Difference
Naïve non tregs $F^{-}H^{-}$	0.065	0.0074	0.057
CM non tregs $F^{-}H^{-}$	0.76	0.14	0.62
$CD25^{-}$ $CD127^{-}$ $F^{-}H^{-}$	Not tested	0.17	
$CD25^{-}$ $CD127^{-}$ $F^{+}H^{+}$	Not tested	104.78	
Tregs Naïve $F^{+}H^{-}$	70.77	49.72	21.05
Tregs Naïve $F^{+}H^{+}$	107.28	108.09	0.81
Tregs memory $F^{-}H^{-}$	71.37	24.41	46.96
Tregs memory $F^{+}H^{-}$	72.86	79.03	6.17
Tregs memory F^{+} H^{+}	108.88	112.66	3.78

The FOXP3$^-$ HELIOS$^-$ cells within the Treg gate, again showed an intermediate methylation profile, however the second measure was considerably lower than the first. When looking at the quality of the flow cytometry staining, there was a clear difference with the second sample having a much larger dynamic range of both the FOXP3 and HELIOS markers (Fig. 4.4), and consequently much easier to gate the positive and negative cells. The level of *FOXP3* TSDR demethylation was 71.4% in the first sample compared to 24.4% in the second, and the difference could be due to the presence of cells expressing low levels of FOXP3 being classified as negative due to the poor stain performance, when in the second sample there is clear separation between the positive and negative population. Even though there is a difference in the performance of the intracellular stain for FOXP3 and HELIOS, both samples identified a population within the FOXP3$^-$ HELIOS$^-$ population that had a demethylated *FOXP3* TSDR. One variable that has been shown to affect the performance of the FOXP3 intracellular stain is temperature, and a higher incubation temperature (30ºC) results in a better FOXP3 stain compared to 20ºC (personal communication from Dr Lorna Jarvis, Cambridge University), and the lab was particularly hot the day the second sample was processed.

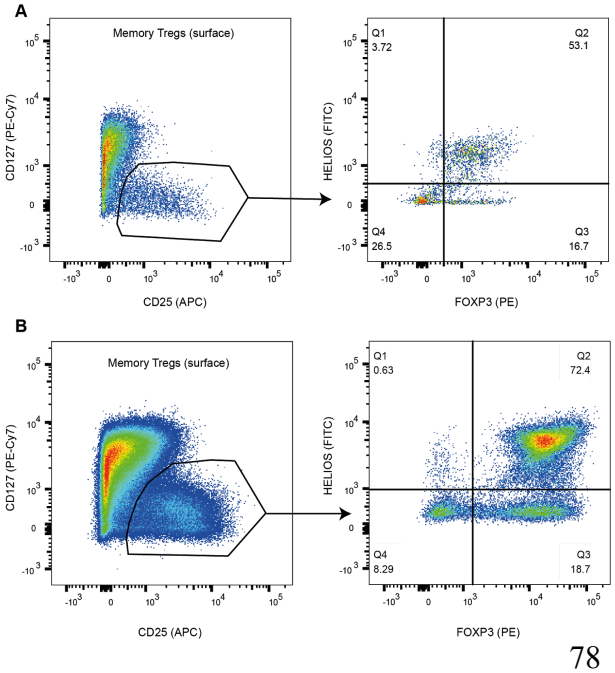


Figure 4.4. Variability in the FOXP3 and HELIOS intracellular flow cytometry stain. The same donor was bled 15 months apart to test the reproducibility of the FOXP3$^+$ cell population frequency and *FOXP3* TSDR measurement. The left panel is gated on memory T cells, from which the Tregs are gated and the right panel displays the FOXP3 and HELIOS staining. The FOXP3 and HELIOS stain is less bright in the first bleed (A) compared to the second bleed

(B). The CD25 and CD127 stain are less bright in the first bleed compared to the second, but not to the same extent as FOXP3 and HELIOS.

4.2.3 Characterising FOXP3$^-$ HELIOS$^-$ cells within the Treg gate

In Figure 4.2 I identified that there were cells within both the naive and memory surface Treg gate that did not express FOXP3 and HELIOS. Due to limited cell number, I have only analysed the TSDR from FOXP3$^-$ HELIOS$^-$ memory Tregs, and as a bulk population displayed a partial TSDR demethylation profile. To better characterise these cells, I analysed the flow cytometric data generated as the samples were sorted to look at the expression of CD25 and CD127 to see if these cells were on the edge of the Treg gate and part of the population may be contaminating T effectors. I have taken the surface Treg population (CD25$^+$, CD127^{low}) and divided it by FOXP3 and HELIOS expression and back gated these cells onto the surface Treg gate (Fig. 4.5). In the two donors I have analysed, one showed that the FOXP3$^-$ HELIOS$^-$ cells had a lower CD25 MFI than the FOXP3$^+$ Treg populations, where the second donors showed no difference in CD25 MFI between subsets (Fig. 4.5D). Although I need to assess more donors using this backgating strategy, as there was no consistent difference in the expression of CD25 (or CD127, data not shown) from the four Treg populations spilt by FOXP3 and HELIOS, and cells were distributed throughout the surface Treg gate, rather than being concentrated on the edge of the Treg gate, these data indicate that at least a portion of the FOXP3$^-$ HELIOS$^-$ cells are not contaminating T effectors. Although all cells express CD25, there is still a range of CD25 expression and in future experiments I would like to analyse more donors to see how the CD25 MFI varies in these Treg subsets and sort the highest and lowest CD25 expressing Tregs to see if TSDR demethylation is enriched in the CD25 brightest cells.

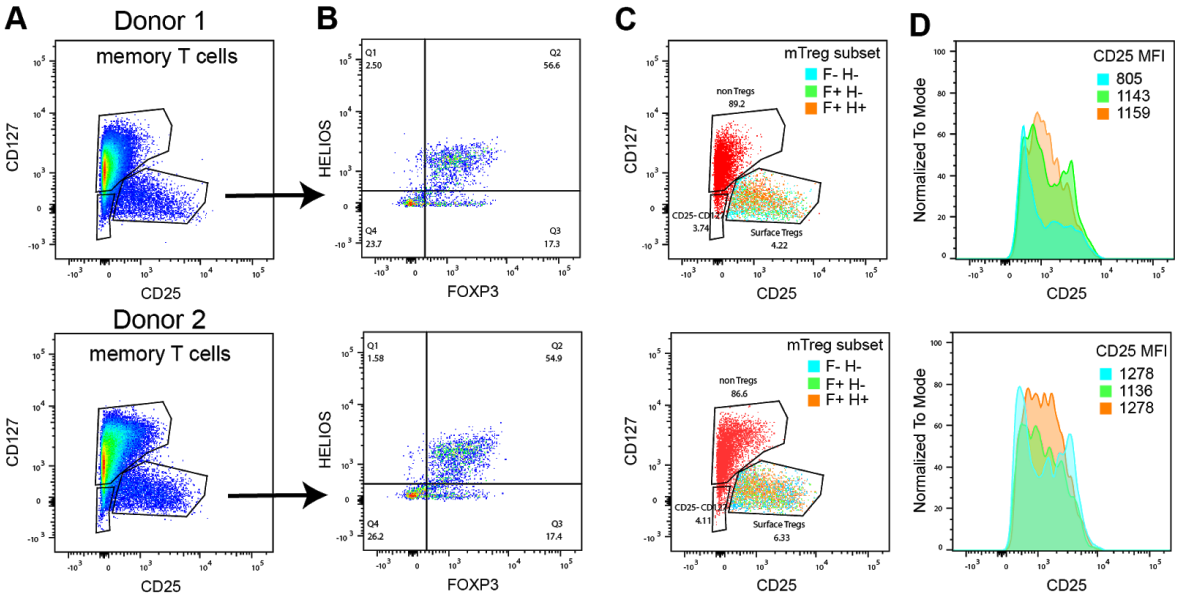


Figure 4.5. FOXP3$^-$ cells are evenly distributed throughout the surface Treg gate. Two donors with 70% (top panel) and 85% (bottom panel) *FOXP3* TSDR demethylation in the CD4$^+$ CD25$^+$ CD127^{low} FOXP3$^-$ HELIOS$^-$ cell populations were analysed for distribution within the surface gate. (A) Initial gating of memory Tregs based on CD25 / CD127 expression (pregated on CD45RA$^-$). (B) Memory Tregs analysed for the expression of FOXP3 and HELIOS, and these populations were backgated (C) onto the CD25 / CD127 gate to see if the Treg subsets differed in their distribution. (D) CD25 median fluorescence intensity (MFI) measured in the three Treg subsets split by FOXP3 and HELIOS. The FOXP3$^-$ HELIOS$^+$ population was not analysed as the frequency was too low.

As the FOXP3$^-$ HELIOS$^-$ cells within the Treg gate do not look like contaminating T effectors based on the expression of CD25 and CD127, I considered if they could be a population of Tregs that are transitioning from FOXP3$^+$ HELIOS$^+$ to FOXP3$^-$ HELIOS$^-$ as the flow cytometry data in some cases suggest a trajectory from the double positive to the double negative and perhaps represents Tregs in the process of apoptosis. To assess this, I have sorted cells that are intermediate for FOXP3 and HELIOS expression for TSDR methylation to see if there is a mixture of methylated and demethylated cells, indicating the cells were transitioning to a methylated TSDR. Figure 4.6 shows that in the three

donors tested, the FOXP3 and HELIOS intermediate expressing cells are nearly as demethylated at the TSDR as the FOXP3 and HELIOS bright cells and more demethylated than the FOXP3^{+} HELIOS^{-} cells. This suggests that the remethylation of the *FOXP3* TSDR does not occur when there is moderate expression of FOXP3. The TSDR methylation levels of FOXP3 and HELIOS double negative cells in the three donors were very similar to the levels seen in the FOXP3^{+} HELIOS^{-} population suggesting the possibility that the FOXP3^{+} HELIOS^{-} population can lose FOXP3 protein expression but retain a demethylated TSDR.

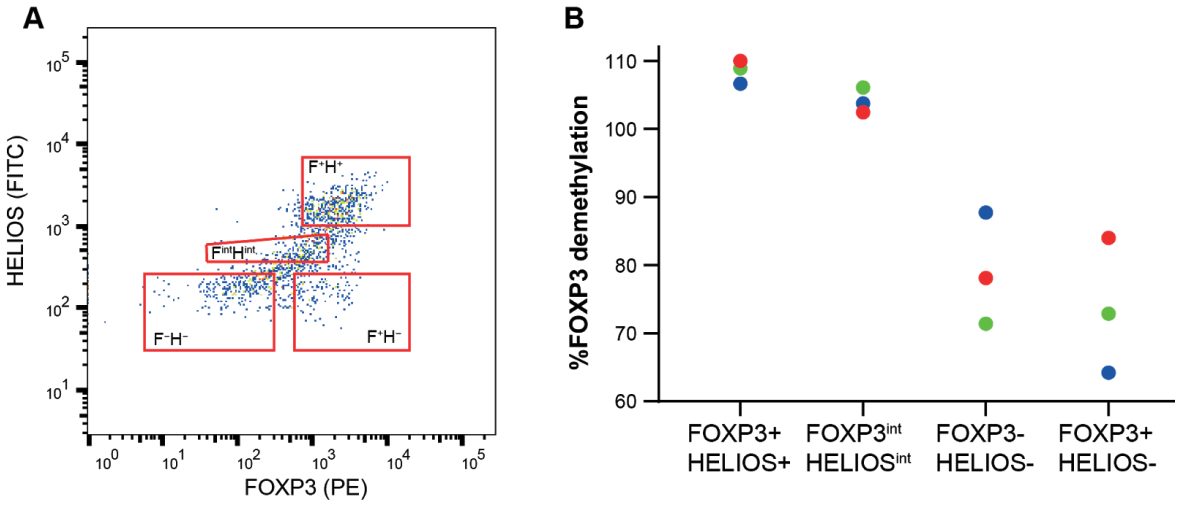


Figure 4.6. The TSDR of FOXP3^{int} HELIOS^{int} Tregs is nearly as demethylated as FOXP3^{+} HELIOS^{+} Tregs. (A) The gating strategy used to sort cells from the memory Treg gate based on the expression of FOXP3 and HELIOS from three female donors. (B) The level of *FOXP3* demethylation at the TSDR was measured, and the individual values from the three donors are shown in red, blue and green. The values were doubled due to X chromosome inactivation.

4.2.4 Can TSDR demethylated cells make cytokines?

The identification of these Treg subsets relies on the intracellular staining of the transcription factors, FOXP3 and HELIOS, which renders the cells unsuitable for any downstream functional assays, making the acquisition of biological information about

these subsets challenging. Therefore as an alternative approach, we stimulated and sorted based on cytokine production as well as FOXP3 and HELIOS, in order to assess the *FOXP3* TSDR from cells that can or cannot produce cytokines. Ricardo Ferreira (DIL) stimulated purified $CD4^+$ T cells with PMA and Ionomycin for 5 hours from three donors and then sorted Treg subsets based on the expression of FOXP3, HELIOS and the ability to produce cytokines. Since cell numbers were limiting, precluding sorting each type of cytokine secreting cell individually, PE-conjugated antibodies to IL-10, IL-21, IFNɣ and IL-2 were added together during intracellular staining to increase the sensitivity to detect production of any of these four cytokines. The antibody to IL-17 was conjugated with a different fluorochrome to enable the specific isolation of IL-17 producing cells, which was achieved from one donor who had a relative abundance of these cells.

I performed analysis of the *FOXP3* TSDR from these cytokine and non-cytokine producing Treg populations and showed that cells which do not make cytokines have a fully demethylated TSDR, regardless of the expression of FOXP3 and HELIOS (Fig. 4.7). Conversely, the cells which produce cytokine have a more methylated TSDR, with the $FOXP3^-$ $HELIOS^-$ double negative population being fully methylated. Cells that are FOXP3 single positive and produce cytokine show an intermediate methylation profile, and need further work to understand this heterogeneous population.

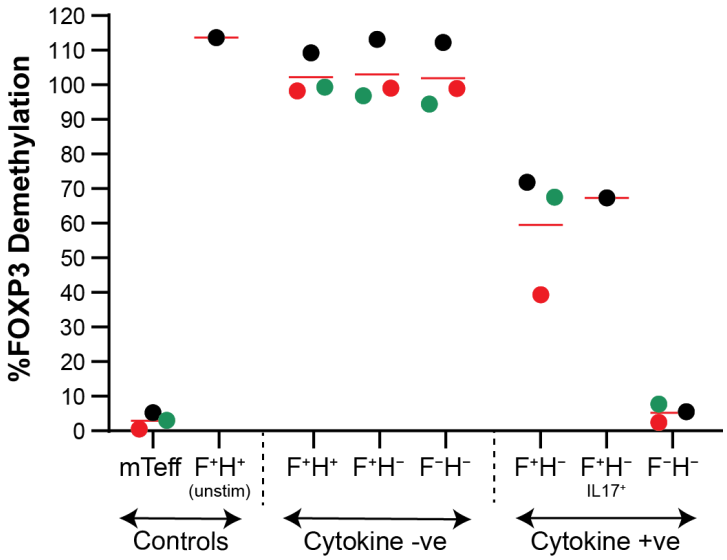


Figure 4.7. Tregs that do not produce cytokines have a fully demethylated *FOXP3* TSDR. Total CD4$^+$ T cells were stimulated with PMA and ionomycin for five hours from two male (red and green circles) and one female donor (black circles), and then stained and sorted based on being either cytokine positive (+ve for a combination of IL-2, IL-21, IFNɣ, IL-10 and IL-17, or in one case IL-17 alone) or negative (-ve), as well as on FOXP3 and HELIOS to define the Treg subsets. The level of the demethylation at the *FOXP3* TSDR was measured and memory T effectors and unstimulated Tregs were included as negative controls. The values for the female donors were doubled to account for X inactivation.

4.2.5 Demethylation of lineage defining Th-1 and Th-17 regions

As well as being able to assess the methylation status of the TSDR from cytokine-producing and non-producing Treg subsets to gain insight into these populations, it is also possible to assess the methylation status of the key Th-1 and Th-17 lineage defining genes; for Th-1 cells a region within *IFNG* and for Th-17 cells regions within *IL17A* and *RORC* have been identified[94] as being specifically demethylated in Th-1 and Th-17 cells, respectively. The aim is to assess if these Treg subsets already have the epigenetic landscape necessary to support the production of IFNɣ and IL17A cytokines. Ricardo Ferreira (DIL) purified CD4$^+$ T cells from three individuals and stimulated them for five hours with PMA and Ionomycin prior to immunostaining for IFNɣ and IL17 production as well as markers to define Treg and Teff populations, thereby allowing naive,

Th-1 and Th-17 Tregs and Teff cells to be sorted. I designed primers to assess the methylation status of the *FOXP3* TSDR and the promoters of *IFNG*, *IL17A* and *RORC* with bisulfite sequencing (Fig. 2.3, Methods), based on previously identified regions[94].

As expected the Teff populations displayed a methylated TSDR, whereas, only the naive Tregs displayed a demethylated TSDR, with the Th-1 and Th-17 Tregs displaying intermediate levels of methylation (Fig. 4.8), which confirms the results from the previous experiment (Fig. 4.7). These Th-1 and Th-17 populations are heterogeneous, and contain a mixture of methylated and demethylated TSDR, which may represent a population of induced Tregs or Tregs with a less committed fate[157], which are able to react to their local immune environment[158].

In naive Tregs and T effectors the promoters of *IFNG*, *IL17A* and *RORC* are primarily methylated, indicating that these cells have not committed to a particular cell lineage yet. Substantial demethylation of the *IFNG* promoter was seen in both Th-1 Tregs and Teffector cells, which is the defining cytokine of Th-1 cells, whereas the *IFNG* promoter in Th-17 cells is methylated to the same degree as naive T cells. Conversely, the promoters of IL17A and RORC are specifically demethylated in Th-17 cells, which are the lineage defining cytokine and transcription factor of Th-17 cells, and methylated in Th-1 cells (Fig. 4.8). These methylation assays provide additional capability to assess the lineage commitment of Treg subsets.

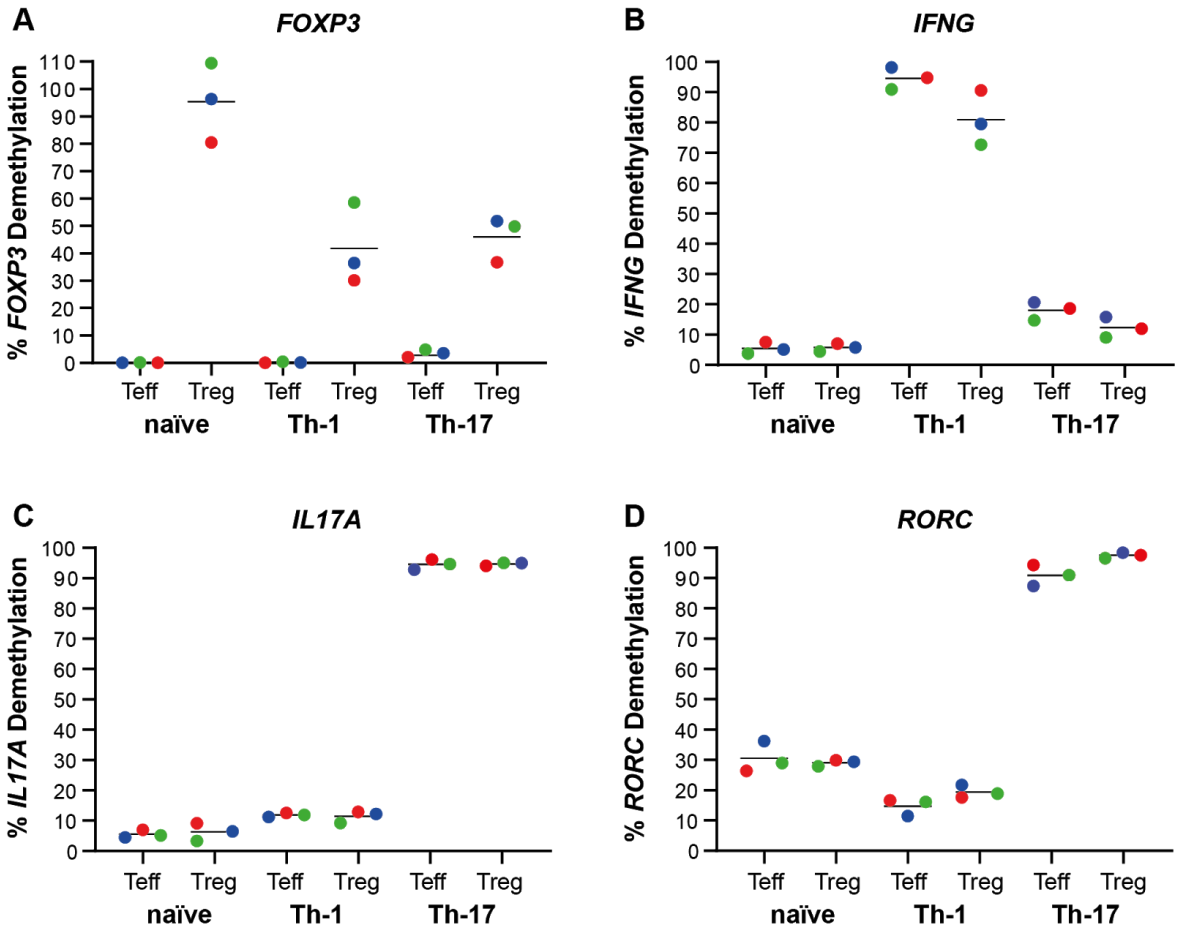


Figure 4.8. Measurement of (A) *FOXP3*, (B) *IL17A*, (C) *IFNG* and (D) *RORC* demethylation in cells based on IL17 and IFNɣ production following PMA and ionomycin stimulation. The regions analysed for methylation of *IFNG*, *IL17A* and *RORC* were previously identified[94] and contain 3, 10 and 7 CpG sites, respectively. Th-1 cells were sorted as CCR6$^-$ CD161$^-$ IL17$^-$ IFNɣ$^+$, and Th-17 cells were sorted as CCR6$^+$ IL17$^+$ IFNɣ$^-$ from either the T effector or Treg population. Red circle is a male donor and the green and blue circles are female donors for which the *FOXP3* demethylation has been doubled as this is the only gene analysed which is located on the X chromosome.

4.3 Application of *FOXP3* TSDR measurement

Now that I have a good understanding of the technical performance of the TSDR assay as well as comprehensive data on which cells have a methylated or demethylated TSDR, we have applied this assay to answer specific questions. I have performed all of the *FOXP3*

TSDR methylation analyses in this section, which have formed a key part of the associated publications.

4.3.1 Are the CD25^{low} FOXP3^{+} cells that are increased in autoimmunity demethylated at the *FOXP3* TSDR?

A population of cells that express FOXP3 but with low expression of CD25 have been identified as being expanded in patients with autoimmune diseases including systemic lupus erythematosus[101,151–154], multiple sclerosis[155], rheumatoid arthritis[156] and type 1 diabetes[101]. These cells do not fit the standard definition of a Treg as they have low expression of CD25. Therefore, the first question to address was what is the methylation status of the *FOXP3* TSDR in these cells? Cell populations were sorted based on their expression of CD25$^{(low/high)}$ and FOXP3$^{(+/-)}$ from four healthy individuals and I analysed their methylation status at the *FOXP3* TSDR (Fig. 4.9) . The majority of cells within the CD25^{low} FOXP3^{+} and CD25^{high} FOXP3^{+} cells were demethylated at the *FOXP3* TSDR (57.1% vs 80.5%), suggesting these cells are thymically-derived Tregs and not activated T effectors that have upregulated FOXP3 as a consequence of immune activation.

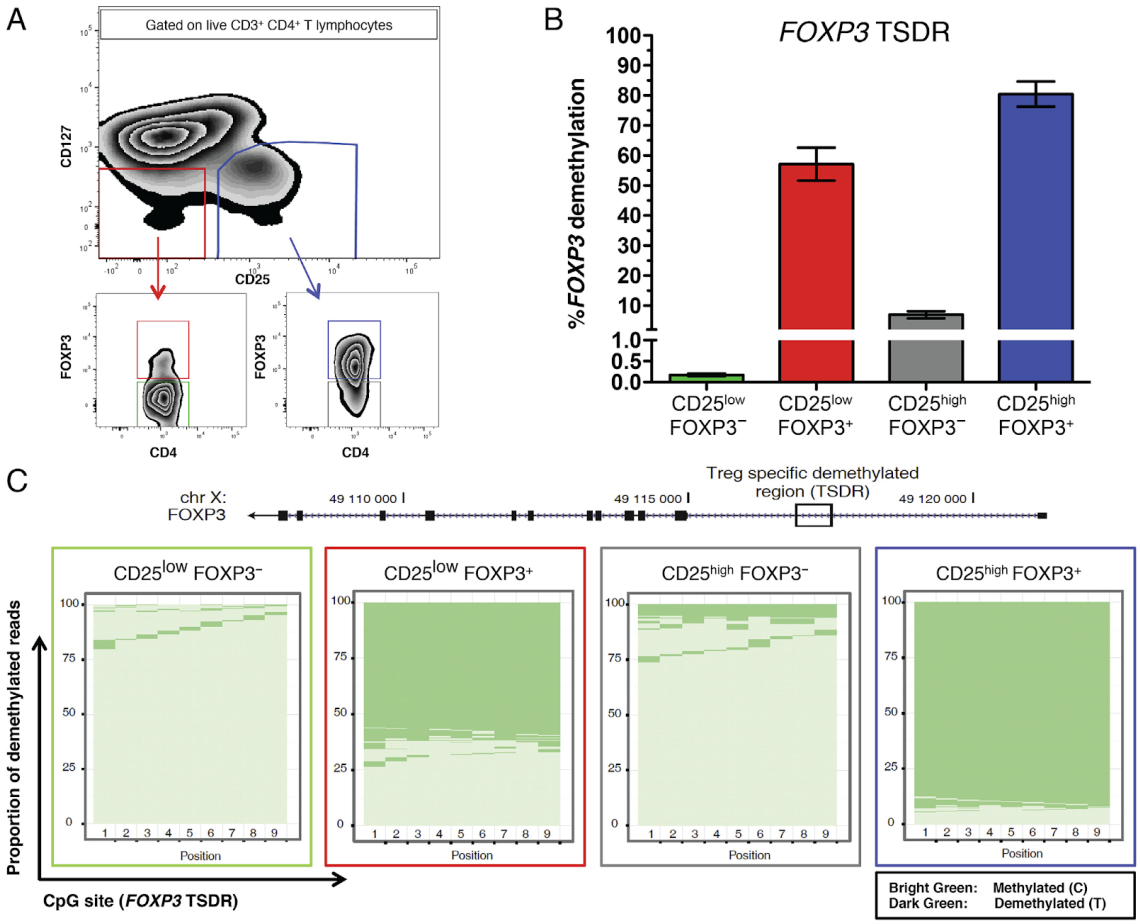


Figure 4.9. This is Figure 2 from "Cells with Treg-specific *FOXP3* demethylation but low CD25 are prevalent in autoimmunity" published in the Journal of Autoimmunity[101] for which I performed the experiments in this figure. The legend reads: Fig. 2. $CD25^{low}FOXP3^{+}$ cells are demethylated at the *FOXP3* Treg-specific demethylated region (TSDR). (A) Gating strategy for FACS sorting of four $CD4^{+}$ T-cell subsets: (i) $CD127^{low}CD25^{low}FOXP3^{-}$ (depicted in green), (ii) $CD127^{low}CD25^{low}FOXP3^{+}$ (depicted in red), (iii) $CD127^{low}CD25^{high}FOXP3^{-}$ (depicted in grey), and (iv) $CD127^{low}CD25^{high}FOXP3^{+}$ (depicted in blue). (B) Frequency (mean ± SEM) of reads demethylated at eight or nine of the nine interrogated CpG sites in the *FOXP3* TSDR. The data were obtained from sorted cells from four independent healthy donors. (C) Graphic depicts the proportion of demethylated reads at the nine interrogated CpG sites from the *FOXP3* TSDR in one illustrative donor. Each horizontal line represents one sequencing read, with light green representing a methylated read (C) and dark green representing a demethylated read (T). Note that the plot is representative of a male donor. For female donors, X-chromosome inactivation causes half of the reads to be methylated and a correction factor of two was applied to obtain the frequency of demethylated reads.

The above sorts purified cells based on the expression of FOXP3 and not in conjunction with HELIOS, which might explain why complete demethylation of the TSDR was not observed; as I have shown in Fig. 4.2, the level of methylation of the *FOXP3* TSDR is variable when sorting on FOXP3 alone. Therefore a second sort was performed and cells were purified based on the expression of FOXP3 and HELIOS. When the *FOXP3* TSDR was assessed from cells purified as double positive for FOXP3 and HELIOS, regardless of CD25 expression ($CD25^{low}$ or $CD25^{high}$), a fully demethylated *FOXP3* TSDR was observed, whereas the $FOXP3^{+}$ $HELIOS^{-}$ cells displayed an intermediate methylation profile (Fig. 4.10).

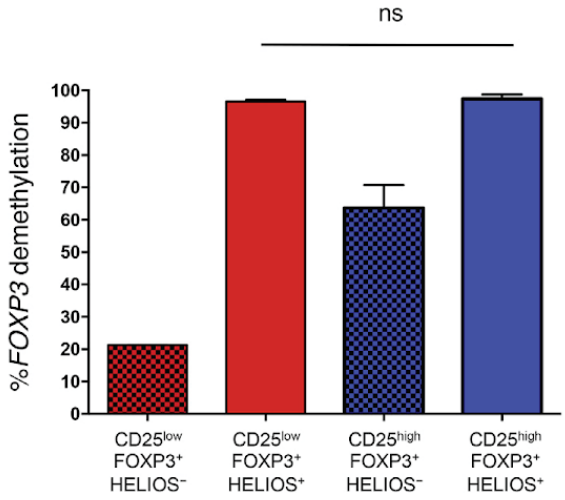


Figure 4.10. This is Figure 5A from "Cells with Treg-specific *FOXP3* demethylation but low CD25 are prevalent in autoimmunity" published in the Journal of Autoimmunity[101] for which I performed the experiments in this figure. The legend reads: Fig. 5. $HELIOS^{+}$ $CD45RA^{-}$ $CD25^{low}$ $FOXP3^{+}$ cells are demethylated at TSDR as much as conventional $HELIOS^{+}$ $CD45RA^{-}$ $CD25^{high}$ $FOXP3^{+}$ Tregs. Frequency (mean ± SEM) of reads demethylated at eight or nine of the nine interrogated CpG sites in the *FOXP3* TSDR in $CD45RA^{-}$ $CD25^{low}$ $FOXP3^{+}$ cells and $CD45RA^{-}CD25^{high}FOXP3^{+}$ Tregs stratified by the expression of HELIOS. The data were obtained from sorted cells from three independent healthy donors.

The TSDR data presented in Figures 4.9 and 4.10 was measured in healthy controls, however the largest expansion of $CD25^{low}$ $FOXP3^{+}$ cells was seen in SLE patients. Therefore, $CD25^{low}$ $FOXP3^{+}$ $HELIOS^{+}$ cells were purified from three SLE patients and the

methylation of the TSDR measured. Consistent, with all the previous data I have generated, the TSDR of CD25^{low} FOXP3^{+} HELIOS^{+} cells was fully demethylated in the SLE patients to the same degree to that of healthy individuals (Fig. 4.11).

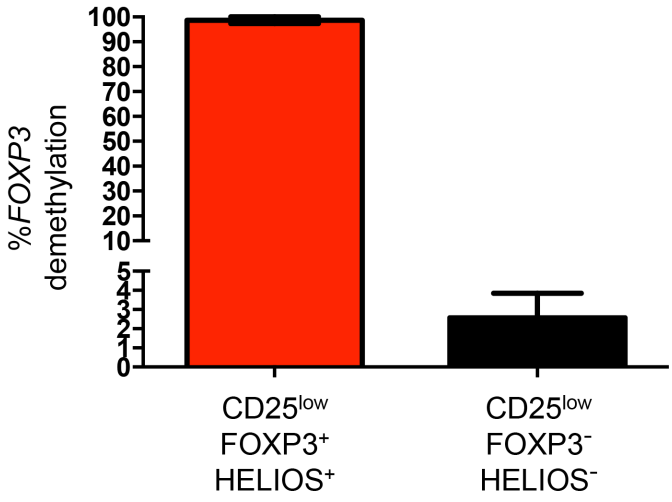


Figure 4.11. This is Supplemental Figure 4 from "Cells with Treg-specific *FOXP3* demethylation but low CD25 are prevalent in autoimmunity" published in the Journal of Autoimmunity[101] for which I performed the experiments in this figure. The legend is: Supplementary Fig. 4. TSDR methylation profile of HELIOS^{+} CD45RA^{-} CD25^{low} FOXP3^{+} cells is maintained in SLE patients. Frequency (mean ± SEM) of reads demethylated at eight or nine of the nine interrogated CpG sites in the *FOXP3* TSDR in CD45RA^{-} HELIOS^{+} CD25^{low} FOXP3^{+} cells and CD45RA^{-} HELIOS^{-} CD25^{low} FOXP3^{-} Teffs. The data were obtained from sorted cells from three independent SLE donors.

To further characterise the CD25^{low} FOXP3^{+} population, Ricardo Ferreria (DIL, Oxford) has performed detailed immunophenotyping on a cohort of SLE patients, CID patients, T1D patients and a healthy control group. The CD25^{low} FOXP3^{+} population expressed the same markers as CD25^{high} FOXP3^{+}, but contained a higher frequency of CD45RA^{-} (memory) and Ki-67 expression, and lower expression of FOXP3, PD-1, HELIOS, TIGIT, CD15s and CTLA4 and may represent a subset of Tregs that has expanded in response to an autoimmune reaction[101].

4.3.2 Are the Tregs expanded in T1D patients receiving ultra-low dose IL-2 demethylated at the *FOXP3* TSDR?

As mentioned in the introduction, the GWAS results for T1D have identified several key proteins within the IL-2 pathway which include *IL2*, *IL2RA*, *BACH2* and *PTPN2*[59]. These IL-2 pathway associations, together with our work showing that replacement of the susceptible *Il2* allele in the non-obese diabetic (NOD) mouse model of T1D, in which 80% of mice spontaneously develop T1D, with a protective *Il2* allele that produces 2-fold more IL-2 is hugely protective and results in only 10% of mice developing T1D[51]. This provided the rationale for our laboratory to conduct a study that administered a single ultra-low dose of IL-2 (Aldesleukin, Novartis) to adults who had developed T1D[55]. The objective of the study (DILT1D) was to discover the optimal dose of Aldesleukin to achieve a 20% increase in Treg frequency, and therefore included an initial dose escalation learning phase followed by an adaptive phase that allowed refinement of the administered dose.

Each participant was followed at multiple time points post administration of Aldesleukin and flow cytometric analysis showed that immediately after injection there was a reduction in the frequency of Tregs in the periphery, which recovered, and increased to 20% above baseline frequency, peaking at day 5[REF53] (Fig. 4.12), which was maintained with the higher doses of Aldesleukin for at least seven days. Flow cytometry analysis of Treg frequency by both surface staining for $CD4^+$ T cells expressing $CD25^{high}$ and $CD127^{low}$, or in conjunction with the intracellular stain to FOXP3 showed the same initial dip followed by

an increase in Treg frequency. However, as both CD25 and FOXP3 can be upregulated transiently in activated T effector cells, I therefore measured the methylation status of the *FOXP3* TSDR to confirm the flow cytometric analysis that the expanded cells were genuine Tregs and not expanded T effectors.

At day 0 and 3 and 60 days post Aldesleukin administration, $CD4^+$ Tregs, central memory and effector memory T cells were sort purified and I measured the methylation status of the *FOXP3* TSDR. The level of demethylation measured at the TSDR remained consistent throughout the time points tested with Tregs (sorted by surface markers $CD25^{high}$ $CD127^{low}$) being 80% demethylated and T effectors showing less than 2% demethylation, confirming that there was a *bona fide* increase in Tregs frequency in the periphery (Fig. 4.12). Due to sample limitation, as the blood collected daily from DILT1D trial participants was used for numerous assays, cells were only sorted as $CD4^+$ $CD25^{high}$ $CD127^{low}$, as the intracellular protocol required to stain for FOXP3 leads to increased cell loss, and requires a higher starting number of cells. Therefore the TSDR was only 80% demethylated in the assessed Treg populations as I have shown there are a proportion of cells within the surface Treg gate that are methylated at the TSDR (Fig. 4.2).

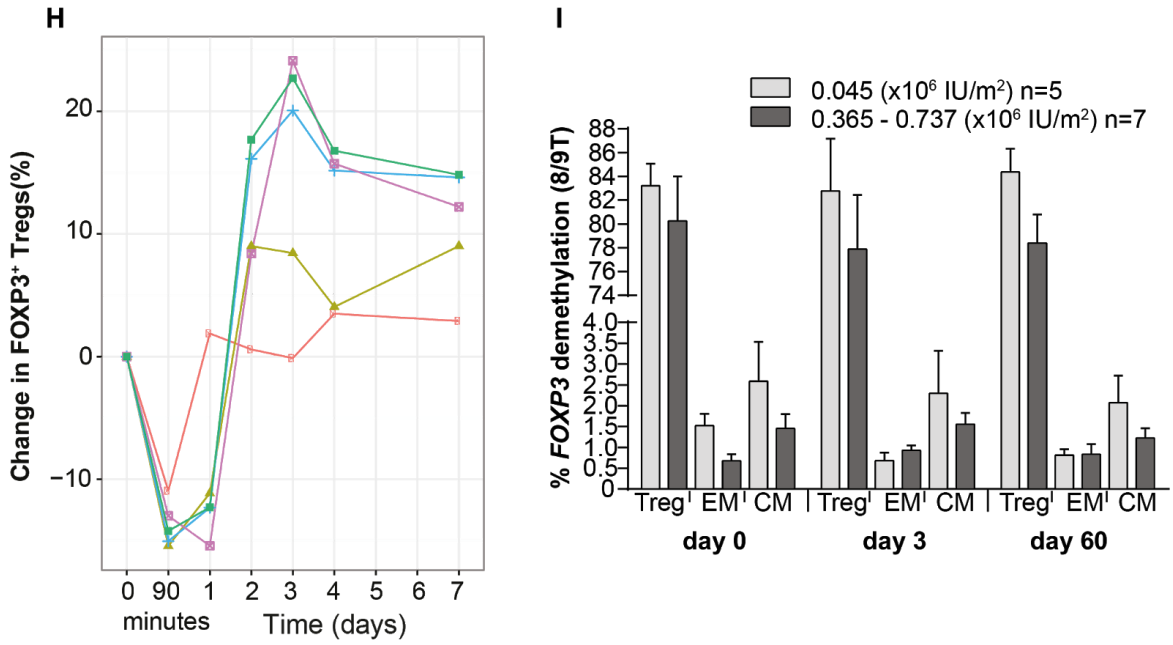


Figure 4.12. This is Figure 7H and 5I taken from "Regulatory T Cell Responses in Participants with Type 1 Diabetes after a Single Dose of Interleukin-2: A Non-Randomised, Open Label, Adaptive Dose-Finding Trial" published in PLOS Medicine[53], for which I performed the TSDR measurement in part I. The legend reads: (H) Intracellular staining of Tregs from whole blood for FOXP3 showed an increase in FOXP3$^+$ Tregs on day 3 (FOXP3$^+$ Tregs/CD4$^+$ T cells = 6.44%, SE = 0.25%, range 4.03%±10.30%, n = 37). (I) Analysis of *FOXP3* gene demethylation on total Tregs and CD62L^{low} (effector memory) and CD62L^{high} (central memory) CD4$^+$ memory T cells sorted from whole blood at pretreatment, post-treatment (day 3), and the last visit (day 60) showing stability of this Treg phenotype.

4.4 Discussion

In this chapter I have shown that the *FOXP3* TSDR is the most demethylated in Tregs which express both FOXP3 and HELIOS. However, there is heterogeneity in the FOXP3$^+$ HELIOS$^-$ and FOXP3$^-$ HELIOS$^-$ Treg cell populations. All cells which coexpress the transcription factors FOXP3 and HELIOS are extremely demethylated regardless where they are located within the CD4 gate, although their frequency outside of the Treg gate is

low. Unexpectedly, there are TSDR demethylated cells within the surface Treg gate in which cells do not express FOXP3 and HELIOS; these demethylated cells are further confirmed not to be contaminating T effector cells as they do not produce cytokines in response to PMA and ionomycin stimulation.

In the mouse, there is a region in *Ikzf2* (which encodes Helios) that is specifically demethylated in Tregs[93,159], and marks if a cell has the ability to express HELIOS. Unfortunately this region is not conserved with humans, but if a similar region could be identified, it could possibly determine if FOXP3$^+$ HELIOS$^-$ Tregs have the epigenetic potential to express HELIOS like FOXP3$^+$ HELIOS$^+$ Tregs or are epigenetically distinct suggesting they do not upregulate HELIOS and cannot transition to the FOXP3$^+$ HELIOS$^+$ population and are a separate lineage of Tregs. A recent murine study showed little overlap of the TCR repertoire between Helios$^+$ and Helios$^-$ Tregs, suggesting they are functionally distinct populations[160].

A consequence of sorting cells using intracellular markers is that you render the cells unsuitable for functional assays such as suppression assays, which is unfortunate since it would be fascinating to see if the cells which are demethylated at the *FOXP3* TSDR but do not express FOXP3 or HELIOS can suppress cell proliferation to the same extent as FOXP3$^+$ Tregs. To circumvent not being able to perform functional assays on fixed cells, an alternative approach would be to perform RNA-Seq on fixed cells and see if there are any genes (which translates to a surface protein) whose expression is enriched in either of these FOXP3 and HELIOS expressing cells allowing for an 'enrichment' sort. The downside of sorting fixed cells is that the RNA extraction protocol is more challenging and

yields low quality and quantity RNA[161], and an amplification free protocol, such as Nanostring would be better for assessing mRNA expression differences. Unfortunately Nanostring is a targeted approach and prevents a non-hypothesis driven search for differentially expressed genes.

As an alternative strategy, our laboratory is undertaking ATAC-Seq, which is possible in fixed cells, to assess how the chromatin landscape differs between FOXP3^{+} HELIOS^{+}, FOXP3^{+} HELIOS^{-} and FOXP3^{-} HELIOS^{-} cells sorted from the CD4^{+} CD25^{+} CD127^{low} surface Treg gate. The ATAC-Seq protocol works on 50,000 cells and provides information about the region of open chromatin within each of these Treg subsets and can be compared to assess the relatedness of each population. A recent publication has demonstrated the potential of using the mitochondrial DNA sequencing reads generated as a byproduct of ATAC-Seq data for lineage tracing of cell populations, whereby cells with the same mutation in the mitochondrial DNA sequences share a common progenitor, therefore it would be possible to see if there was any overlap between the different Treg populations[162].

There are now numerous clinical trials in the setting of autoimmunity to either boost the frequency of Tregs using injection of IL-2[REF53,54,163,164] or expanding Tregs *ex vivo* and transplanting them back into patients[165,166]. Therefore, assessment of the *FOXP3* TSDR is critical to ensure accurate quantification of *bona fide* Tregs to be transferred and to continually monitor the frequency of Treg changes during and post treatment with either expanded Tregs or with Treg-expanding therapeutics. Such treatments could prove to be interesting settings to monitor the frequencies of different FOXP3- and HELIOS-

expressing Tregs to see if there is evidence of transitioning between the $FOXP3^+$ $HELIOS^+$, $FOXP3^+$ $HELIOS^-$ or $FOXP3^-$ $HELIOS^-$ cell populations.

The two stage PCR amplification system I have used to assess the *FOXP3* TSDR has been adapted and used to genotype SNPs which were not possible by Taqman or genotyping chips, like rs1800521 in AIRE[91]. This same strategy could be used for allele-specific expression studies, and would revolutionise the way in which these experiments are currently being performed, allowing much higher throughput, and this method will be used routinely in Chapter 5 to dissect the genetic association of *IL2RA* with autoimmune disease.

Chapter 5

IL2RA autoimmune disease associated haplotypes

5.1 Introduction and aims

IL2RA (the gene encoding the α chain of the heterotrimeric IL-2 receptor, CD25) has been associated with 39 traits in GWAS Catalogue[167], the majority of which are autoimmune diseases, including type 1 diabetes[17,168,169] (T1D), multiple sclerosis[170–172] (MS) and autoimmune thyroid disease[173] (AITD). Interestingly, there is no common genetic association signal, instead, many different variants in the *IL2RA* region, which are not in strong LD, have been identified as being associated with autoimmune disease. Where variants have been identified that are shared between diseases, the direction of association is not always concordant, for example, the T allele rs61839660 that is protective for T1D[17], is a susceptibility allele for CD[14]. These initial observations highlight the complexity of the genetic association of the *IL2RA* region, which I will attempt to explain in this chapter.

Traditionally, stepwise regression is used to identify the single most associated variant within a region, and then iteratively test for secondary signals. However, Dr Chris Wallace (Department of Medicine and MRC Biostatistics Unit, University of Cambridge), adapted a Bayesian stochastic search algorithm to enable fine mapping, GUESSFM[10], with the advantage that it simultaneously assesses multiple groups of SNPs to find which best

explain the association signal. This stochastic search method (GUESSFM) was applied to the *IL2RA* region[2] for both T1D and MS, and forms the cornerstone of this chapter.

The stochastic search analysis of the *IL2RA* region in T1D has revealed there are multiple association signals, which have been called A, C, E and F and they lie within intron 1 of *IL2RA*, intergenic or within the distal genes *RBM17*, *PFKFB3* and the microRNA *MIR3155*, spanning a 190 kb region (Fig. 5.1). For MS, the SNP rs2140286 (Figure 5.1, called B) has been associated with both T1D and MS[5,6], but using the stochastic search method it has shown that this SNP is actually tagging two independent effects in moderate LD ($r^2 \simeq 0.3$), which are the T1D A association, as well as an MS specific signal termed D[10,11], that will be discussed further in section 5.6.1. Recently, this method has been applied to AITD, and again the lead SNP identified by stepwise regression, rs706779[REF173] was shown to be joint tagging two independent signals, which are the A and C signals identified as being associated with T1D[104].

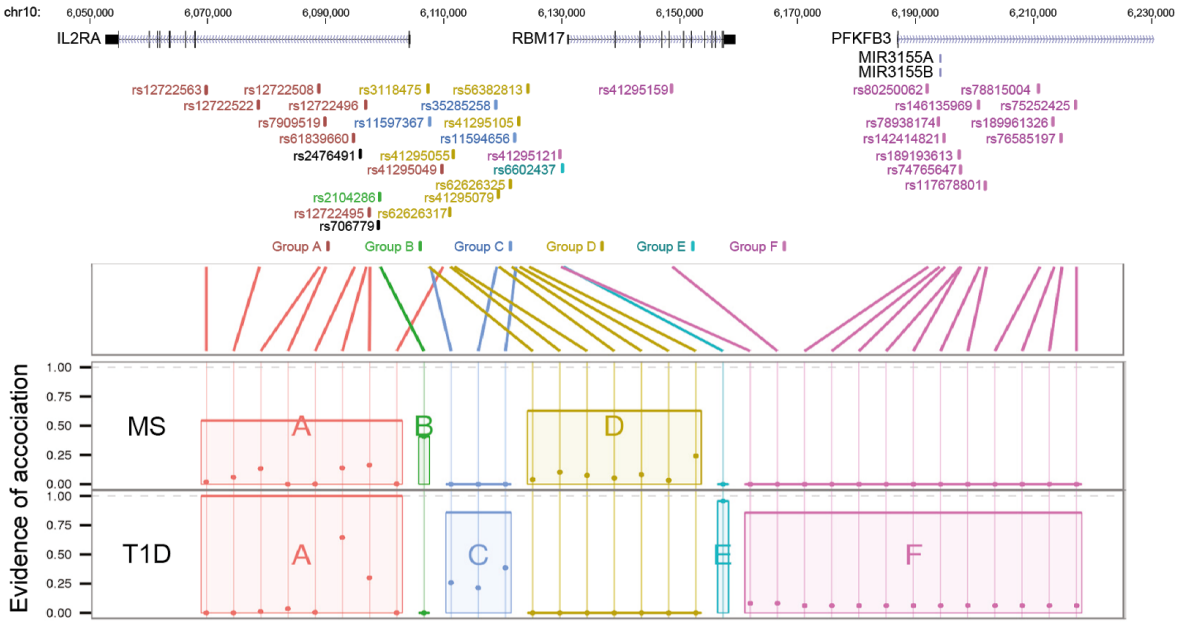


Figure 5.1. The genomic location of the *IL2RA* group A, B, C, D, E and F SNPs[2] as well as the most associated AITD[173] (rs706779) SNP. Data are displayed in custom tracks uploaded to the UCSC browser. Coordinates are based on genome build GRCh37. The lower part of the figure is modified from Wallace, C. *et al.*[2] "Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping". Plos Genetics, (2015) and shows the marginal posterior probability of inclusion for MS and T1D for each SNP within each group and the height of the rectangle is the group marginal posterior probability of inclusion. SNP groups are labelled by the letters A-F and SNPs in this track are ordered by SNP group for ease of visualisation.

For T1D, each association signal could be explained by any of between one and 13 SNPs, which throughout this report will be termed the A, C, E and F haplotypes[2]. The association that is specific to MS, is a group of seven SNPs that will be referred to as the D haplotype. The SNPs that form each group are shown in Figure 5.1, along with their location in the *IL2RA* region.

The aim of this chapter of my thesis was to investigate each of the disease associated haplotypes, with the following goals:

1. To understand the haplotype structure of each *IL2RA* SNP group and assess heterogeneity amongst donors carrying the same haplotype and compare this variation to that present in individuals carrying a different haplotype.
2. To understand how each group of SNPs exerts its effect by analysing the expression of *IL2RA* in purified cell populations.
3. To understand the genetic relationship between T1D, MS and AITD to see which disease association signals are shared between diseases and which are unique to a particular disease.
4. To identify which SNP within each group is the most likely disease variant.

5.2 Haplotype analysis and comparison

The most recent publications on genetic association of the *IL2RA* region with T1D used a novel Bayesian approach to identify the most likely credible set of SNPs[2,104]. Dr Chris Wallace (Department of Medicine and MRC Biostatistics Unit, University of Cambridge) has performed a haplotype analysis (Fig. 5.2) using the most credible sets of SNPs[2,104] to define haplotypes. This analysis has also been applied to the GWAS data from AITD[173] and MS[171], which found that the A haplotype is protective for T1D, but also for AITD and MS. Although the size of the protective effect is different with the group A haplotype being most protective for T1D, AITD and MS are similarly protected (Fig. 5.2). The C haplotype is shared between AITD and T1D, the D haplotype association is unique to MS, and the E and F associations appear to be specific to T1D.

The haplotype analysis identified some discrepancies between the lead SNPs identified by stepwise and stochastic search analysis. In the case of AITD, stepwise analysis identified rs706779 as the most associated SNP[173], however, it is clear from the haplotype analysis that this association could be explained by the fact that the A and C haplotypes, which are both protective for AITD, share the minor allele at rs706779 (Fig. 5.2). It is likely that single-SNP regression analysis fails to distinguish two separate signals from the A and C haplotypes (as it did in T1D[169]) since the two haplotypes offer similar levels of protection for AITD (as compared to the distinct T1D ORs for the A and C haplotypes). In section 5.5 I will explore the relationship between rs706779 and the A and C haplotypes and provide functional data that supports the results of the haplotype analysis showing that the A and C haplotypes better explain the association with AITD.

A similar result was seen in MS, with the lead SNP identified by stepwise analysis, rs2104286[170,171] (termed B, Fig. 5.1 and 5.2) being located primarily on the A and D haplotypes, which were identified by stochastic search (Fig. 5.1). Haplotype analysis confirms that the A and D haplotypes are both protective to a similar extent and direction for MS (Fig 5.2) and shows why rs2104286 (B) was identified as the single top variant by stepwise analysis. The relationship between rs2104286 (B) and the A and D haplotypes will be further explored in section 5.6.

Haplotype analysis has been extremely useful in dissecting the association signals at *IL2RA* as I have shown for AITD and MS. For T1D, one of the signals identified, termed

E, is a single SNP, rs6602437, and through haplotype analysis we realised that the protective T allele of rs6602437 is present on the A haplotype and the vast majority of the time on the D haplotype, but rarely on the C or F haplotypes, as well as being found in isolation with none of the A, C, D or F protective alleles present (Fig. 5.2). Haplotype analysis has also been very helpful for assessing the F haplotype, which is the rarest signal identified (~1% frequency) and is found solely on a small proportion of donors who also carry the C haplotype. The relationships of the protective SNP groups in the population are essential to understand when testing for phenotypes associated with each of the haplotypes.

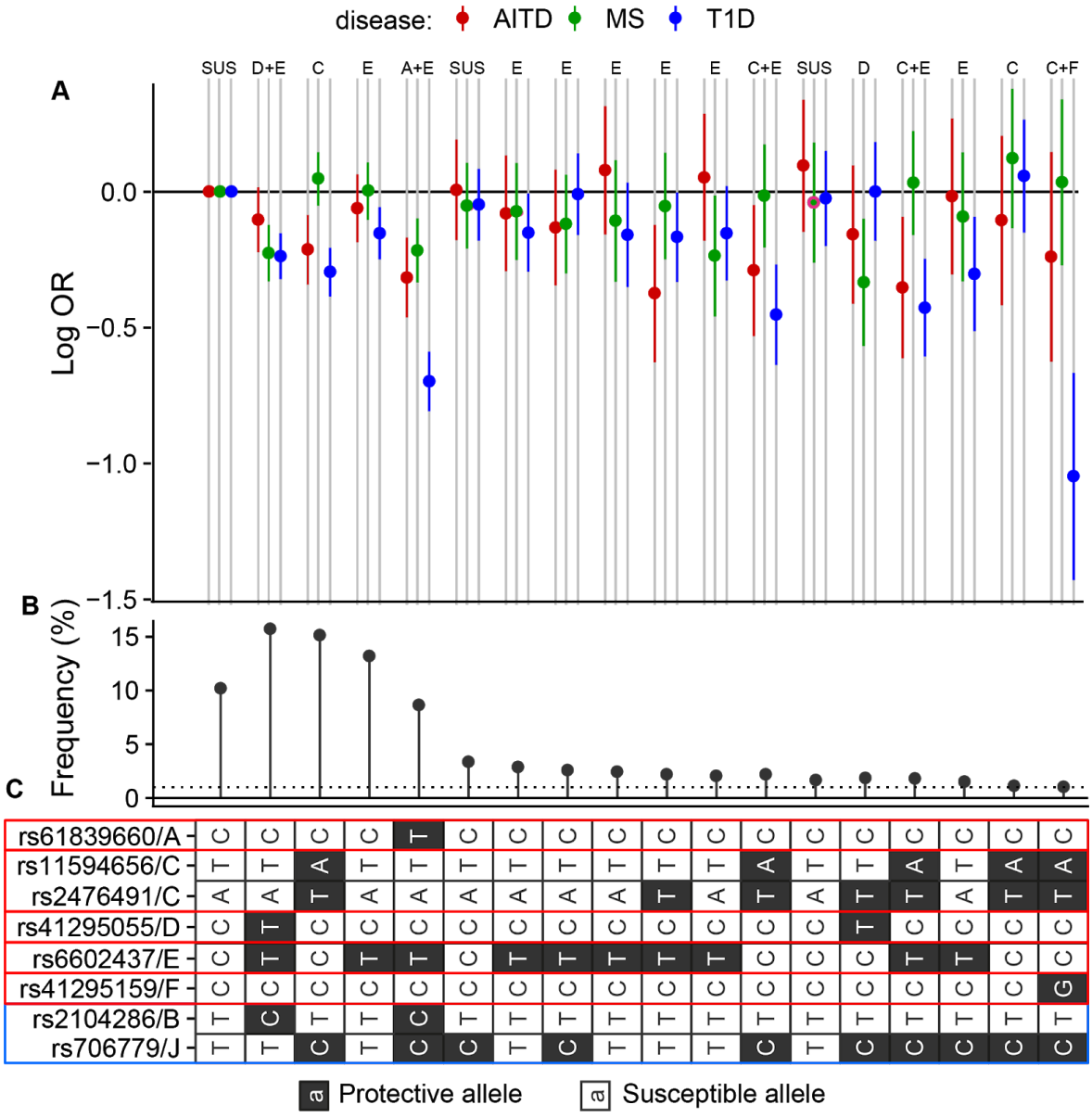


Figure 5.2. Haplotype analysis of the *IL2RA* association with T1D, AITD and MS. This figure was modified from Asimit *et al.*[104] "Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases" published in Nature Communications, for which I was an author. (A) The log OR of three autoimmune diseases with confidence intervals are shown for each haplotype (vertical grey lines separates each haplotype). A negative log OR means a protective haplotype, whereas a log OR greater than or equal to 0 is susceptible to disease. (B) The frequency of each haplotype within the population is shown with decreasing frequency from left to right with the exception of the susceptible haplotype (Sus) to which the others are compared. (C) The haplotype tag SNPs of groups A, C, D, E and F credible SNP groups[2,104], as well as the top SNP for MS[171] (rs2104286, also termed 'B') and AITD[173]

(rs706779). The SNP rs2476491 is included as it is in moderate LD with the C haplotype and has been associated with decreased *IL2RA* in activated $CD4^+$ T cells[174]. The common susceptible haplotype is set as the reference on the left, and the protective allele of each SNP is shown with black background.

This haplotype analysis looks at the relationship of a SNP from each of the disease associated groups (A-F) as well as the most associated SNPs from AITD (rs706779) and MS (rs2104286). However there is no comparison of any of the other SNPs found on each of the haplotypes and I wanted to understand how similar the alleles at all of the other SNPs are amongst donors carrying the same haplotype and to donors carrying a different haplotype at the *IL2RA*. My first aim was to test if all disease protective haplotypes have arisen from a single haplotype with just the disease associated variants differing, or are the haplotypes different at many SNPs suggesting they are ancient. For this analysis, I will use the 503 individuals of European ancestries from the 1000 Genomes Project[105] for which there is phased genetic information. I use the term European ancestries to reflect that the European population is not a homogenous well defined group, and was suggested by @ewanbirney (9th July 2019, https://twitter.com/ewanbirney/status/1148574119438573568) as a better description of the analysed population.

5.2.3 How similar are donors with the same haplotype

Haplotype analysis shows that the T1D/MS protective alleles of the group A, C and D SNPs are on different haplotypes and the protective allele at the E SNP can either be found in isolation or linked together with the A, C or D haplotypes. Apart from the tag SNPs that

define the A, C, D, E and F T1D/MS protective haplotypes and the susceptible haplotype (defined as having susceptible alleles at A, C, D, E and F defining SNPs), how similar is the rest of the genetic landscape of these haplotypes?

From the 1000 Genomes Project[105] I identified the A, C, D, E and F phased haplotypes and then calculate within each haplotype at each SNP how similar the individuals are with the same haplotype and how far the similarity extends physically in the region (Fig. 5.3). Each of the haplotypes had a very different pattern of similarity across the *IL2RA* region, although all haplotypes analysed were similar over a core region which covered the *IL2RA* promoter and exon 1 and the extent of their haplotype tag SNPs. The length that each haplotype was similar varied considerably, with the D haplotype similar over a 140 kb region, whereas the A and C haplotypes were similar over 70 kb and 40 kb regions, respectively (Fig. 5.3). Outside of the regions of similarity, recombination and segregation has occurred multiple times in evolutionary history, which has introduced diversity into the extended haplotypes suggesting that the core regions of similarity, which includes the causal SNP for that haplotype, have been subject to selection.

Within the region that each haplotype is similar must be located elements of DNA that are crucial to the transcriptional regulation of the *IL2RA* gene, a gene essential for immune function, and functional attributes that have facilitated particular haplotypes to become more frequent in the population due to positive selection.

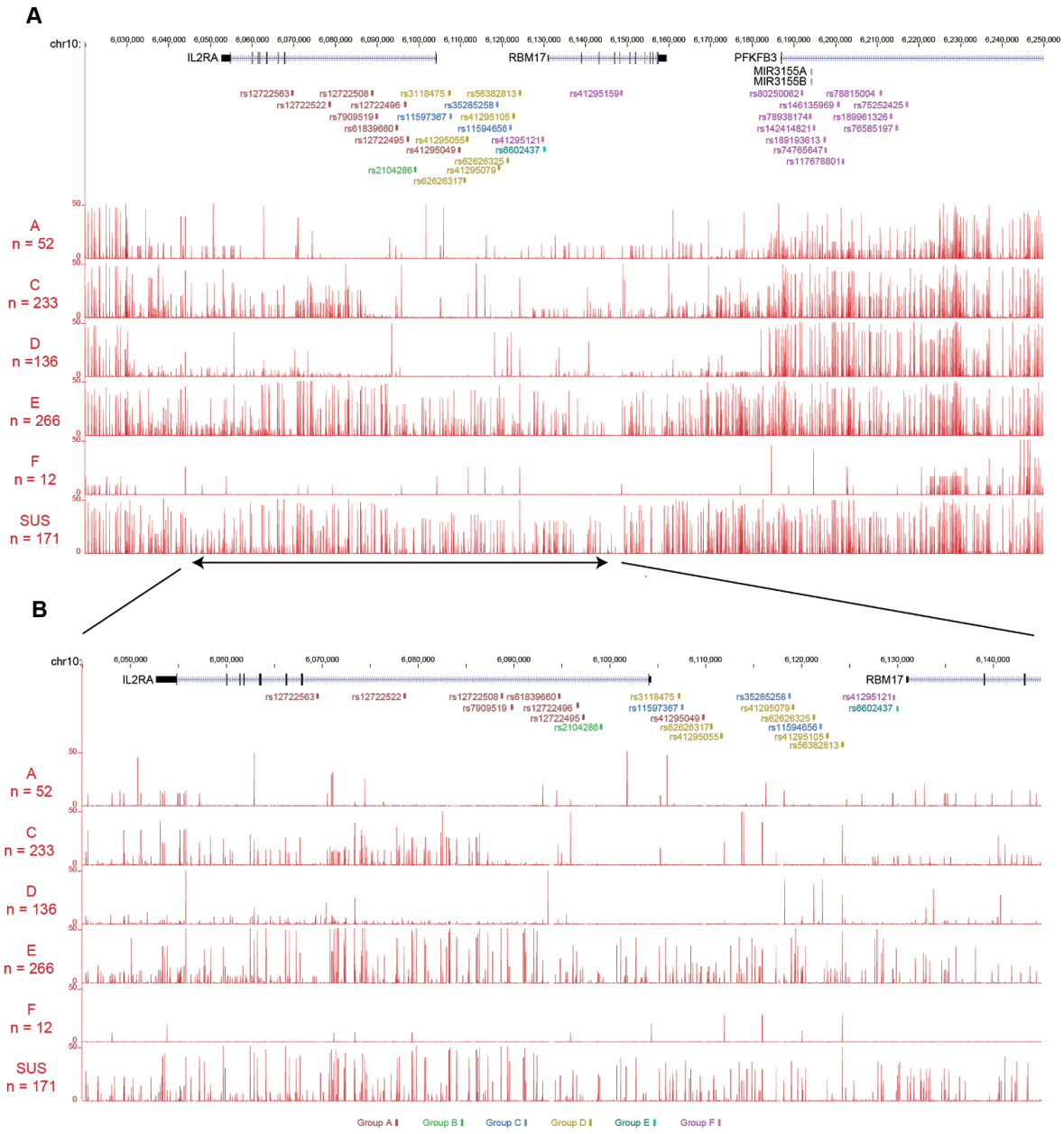


Figure 5.3. Assessment of similarity of each *IL2RA* haplotype from the European population within the 1000 Genomes Project. The location of the group A, C, D, E and F SNPs are displayed along with the *IL2RA* and *RBM17* genes. Each haplotype is displayed as a separate row and each point is one of the (A) 7751SNPs across 230 Kb or (B) 3154 SNPs across 100 Kb analysed. The height of each point is related to the proportion of individuals having the same alleles for the SNP (i.e. homozygous for the SNP; score of zero) through to having equal counts of the two SNP alleles (i.e. heterozygous for the SNP; score of 50).

5.2.4 How similar are the haplotypes between each other?

Within each of the protective A, C, D, E and F as well as susceptible haplotypes are regions of similarity, but are these regions the same between haplotypes? By comparing the difference in frequency between the major alleles found on the two haplotypes being compared, I have calculated a score of how different each haplotype is, with a score of zero when the same allele is found on both haplotypes being compared and a score of 100 being achieved when the opposite allele is found in all comparisons.

As well as each haplotype's defining SNPs, there are many other SNPs that are different between the haplotypes (Fig. 5.4). However, there is a 2.5 Kb region around the *IL2RA* promoter that contains no variants between any of the haplotypes reinforcing the notion that this DNA sequence is essential and subject to purifying selection. I have performed a comparison of all haplotypes and there are only a few regions where the different haplotypes show any similarity, which are highlighted in Figure 5.4. These are the C and F haplotypes that share the same alleles over 80 Kb, but in Europe the F haplotype is found only at the distal end of a rare group of donors carrying the C haplotype, therefore they naturally share a similar distribution of alleles, proximal to where the F haplotype defining SNPs are located. Also, the susceptible and E haplotypes share a similar pattern of alleles across a region of 85 Kb, and this is located proximal of the E SNP, suggesting that recombination or gene conversion has moved a small piece of DNA onto the susceptible haplotype, or vice versa. The last region which shares similarity between haplotypes, is a small 3 kb region containing the E SNP; all haplotypes that carry the E SNP (A, D and E)

have a similar pattern of alleles, which is different to all haplotypes not carrying the E SNP (C, F and Susceptible, Fig. 5.4); this observation will be discussed further in Chapter 7.3.4, which delves deeper into the haplotype comparisons.

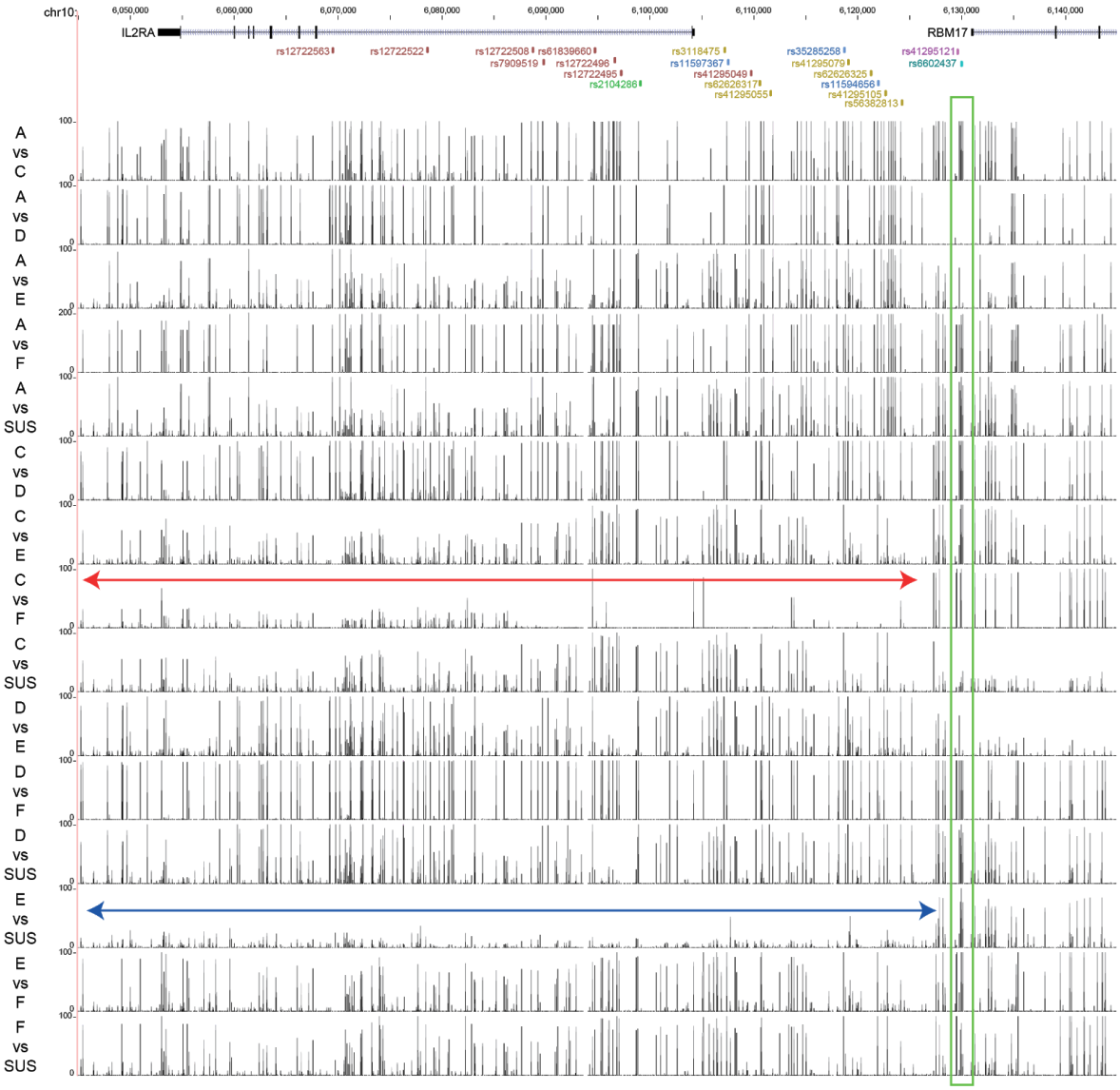


Figure 5.4. Variation between the *IL2RA* haplotypes. The haplotype tag SNPs as defined in Wallace *et al*[2,104] are shown as red for group A, blue for group C, light brown for group D, turquoise for group E and pink for group F. Using phased haplotypes from the 1000 Genomes project, the distribution of alleles between *IL2RA* haplotypes was calculated, and each vertical line shows the proportion of each SNP allele that is common between the two haplotypes being compared, with a score of 0 meaning they have the same allele and a score of 100 meaning they have opposite alleles. Red arrow indicates a large region where the C and F haplotypes share the same alleles, the blue arrow a

region where the E and Susceptible haplotypes share the same alleles. The green box highlights the region containing the E SNP, and those haplotypes that carry the E SNP (A, D and E) share a similar pattern of alleles, compared to the haplotypes that do not carry the E SNP, but share a different set of alleles. Custom tracks were uploaded to the UCSC browser to display the data, and genome coordinates are based upon the GRCh37/hg19 build.

5.2.5 Frequency of *IL2RA* haplotypes in different parts of the world

The data presented in Figure 5.4 used donors of European ancestries from the 1000 Genomes Project, that also contains genetic information from individuals with African, American, East Asian and South Asian ancestries. I have assessed the frequency of the *IL2RA* A, C, D, E and F haplotypes within these populations to compare with my observations made with donors in Europe (Table 5.1). The frequency of the A haplotypes is highest in Europe and rare or not seen in Africa and East Asia, with an intermediate frequency in South Asia and America. The C haplotype shows a similar distribution to the A haplotype, whereas the D haplotype is more similar between Europe, East Asia and South Asia, and has a substantially lower frequency in America and Africa. The E haplotype, which is only variant where the protective allele is also the ancestral allele is most common in the African population. The F haplotype, as it exists as a credible SNP set of 13 SNPs, is extremely rare, and only seen in Europe, suggesting a very strong founder effect. Whilst producing this table, it became clear that the *IL2RA* haplotypes are not always found complete and in certain situations incomplete haplotypes are more common, which I have shown as the frequency of any haplotype defining SNP being present (Table 5.1). Within Europe, the difference between the complete and incomplete *IL2RA* haplotypes is small, and this is reflected by the inability to refine the number of

SNPs that define each of the *IL2RA* haplotypes. The opposite occurs in Africa, where the frequency of incomplete haplotypes is more common than the complete haplotypes and may reflect either increased recombination breaking up the haplotypes or the fact the haplotypes have not formed yet, and Chapter 7 will discuss these hypotheses.

Table 5.1. The frequency of the A, C, D, E and F *IL2RA* haplotypes calculated from the 1000 Genomes Project[105] from five divergent populations across the world. Frequency is calculated as either carrying all of the haplotype defining SNPs or the presence of any haplotype defining SNP, with the difference between these frequencies indicating the presence of partial haplotypes (resulting from recombination, or the haplotype as it is found in Europe had not been formed at the time humans migrated to a different region of the world). The A and D haplotypes primarily occur in combination with E, whereas the C haplotype is most common without E, however in this analysis, the frequency of the A, C and D haplotypes is calculated irrespective of the E SNP. The frequency of E includes A+E, C+E and D+E, as well as E in isolation.

***IL2RA* Group**	**No. of SNPs**	**Frequency of:**	**Africa %**	**America %**	**East Asia %**	**Europe %**	**South Asia %**
A	8	All 8 A	0.4	2.6	0.0	5.2	2.2
		Any A	5.8	4.3	0.0	9.9	7.5
C	3	All 3 C	2.3	26.2	3.0	29.0	19.9
		Any C	8.3	26.8	3.3	29.5	24.7
D	7	All 7 D	0.9	6.3	13.0	16.2	13.6
		Any D	23.7	24.6	20.4	20.1	16.4
E	1	E	61.3	48.4	45.0	53.6	37.4
F	13	All 13 F	0.0	0.0	0.0	0.7	0.0
		Any F	0.2	1.0	2.3	1.3	21.3

5.3 Analysis of the *IL2RA* group A haplotype

5.3.1 Introduction

Common variants of the gene *IL2RA* encoding CD25, the alpha chain of the IL-2 receptor, are associated with multiple autoimmune diseases including T1D[17] and MS[2,171]. The *IL2RA* genetic association is influenced by several distinct haplotypes, with the most protective T1D haplotype shared between T1D (OR = 0.63) and MS (OR = 0.85) and termed "group A"[2]. The group A haplotype is defined by eight SNPs in high LD (Table 5.2)[2], including SNPs rs12722495 and rs61839660 that correlate with higher CD25 expression on $CD4^+$ memory T cells when the minor alleles are present[37,175,176]. One study demonstrated increased *IL2RA* mRNA from the A haplotype in the same cell type[37], with the effect large enough to be seen as an eQTL in whole blood[177], and the same SNP correlating with a *cis*-meQTL with the *IL2RA* promoter[178]. Two eQTLs were seen in activated $CD4^+$ T cells; one with SNPs in high LD with the C haplotype, and a second signal with the A haplotype, in both cases there was a correlation of the T1D-protective alleles with less *IL2RA* mRNA produced with activation[174,179]. Although SNPs in the group A haplotype have been identified as the most associated in genetic association studies of T1D[17,180] and CD (rs61839660)[14] as well as juvenile idiopathic arthritis (rs7909519)[181], owing to the high degree of LD amongst the eight SNPs (minimum pairwise r^2=0.65, Table 5.2), the most associated SNP is not necessarily the causal variant[27], and resolution of high LD between the most disease-associated variants in a region remains a major barrier to causal variant identification. Therefore each of the eight

IL2RA groups A SNPs were assessed to see if they were located in regions of epigenetic interest.

Table 5.2. LD between the group A SNPs compared to rs61839660 was calculated by Haploreg (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) using 503 individuals of European descent from The 1000 Genomes Project.

Chr	Position (hg37)	variant	LD (r^2)	LD (D′)	Ref	Alt	European frequency
10	6069561	rs12722563	0.66	0.96	G	A	9.24 %
10	6078553	rs12722522	0.82	0.96	G	A	7.95 %
10	6088743	rs12722508	0.82	0.96	A	T	8.05 %
10	6089841	rs7909519	0.82	0.96	T	G	8.05 %
10	6094697	rs61839660	1	1	C	T	6.96 %
10	6096667	rs12722496	0.88	1	A	G	7.85 %
10	6097283	rs12722495	0.86	0.98	T	C	7.75 %
10	6109676	rs41295049	0.65	0.84	G	A	6.46 %
10	Frequency of all eight protective alleles on the same chromosome						5.17 %
10	Frequency of A^{-660} haplotype						0.6%

5.3.2 Epigenetic evidence identifies rs61839660 as the most likely causal variant of the group A SNPs

Previously, it has been reported that the SNPs within the *IL2RA* group A haplotype are associated with increased expression of CD25 on central memory $CD4^+$ T cells. Therefore, each of the *IL2RA* group A SNPs was assessed to see if they colocalised with epigenetic signals within memory $CD4^+$ T cells or total $CD4^+$ T cells. Analysis of the ENCODE[41,182] and Roadmap Epigenetic[40] databases suggests rs61839660 is the most likely

causal variant as it was the only A haplotype SNP that lies within epigenetic marks in a relevant cell type (Fig. 5.5). The SNP rs61939660 sits in a region of conserved DNA sequence with open chromatin, potential enhancer activity and able to bind transcription factors (Fig. 5.5). The binding of 91 different transcription factors was assessed in 161 cell lines as part of the ENCODE project, however this does not necessarily reflect what is seen in primary cells, however it has been shown that Tbet binds to the region containing rs61839660 in Th-1 cells[183]. Several regions have been identified through a CRISPR screen for enhancer activity in the *IL2RA* region called CRISPRa-responsive elements (CaRE), and of the A haplotype SNPs only rs61839660 resides in one[174]. Additionally, the Regulome[184] and Haploreg[185] databases assigned rs61839660 as the most likely SNP from the eight SNP group A haplotype to alter transcriptional regulation of *IL2RA*. However, intron 1 of *IL2RA* along with the intergenic region between *IL2RA* and *RBM17* has been designated a super-enhancer[186,187] and therefore additional regulatory elements could exist that I have not identified in this analysis.

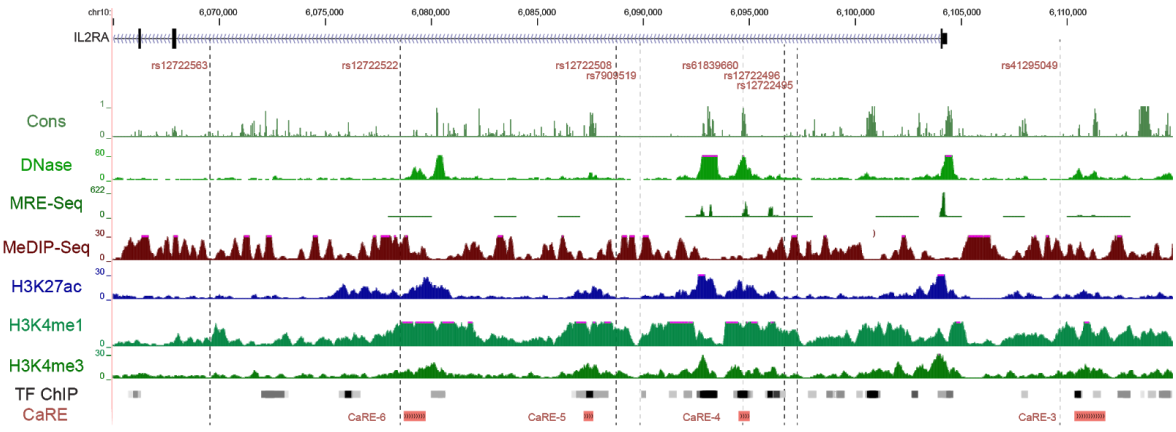


Figure 5.5. Epigenetic analysis of the *IL2RA* region spanning the eight group A SNPs. From the ENCODE and Roadmap Epigenetic data hosted by the UCSC browser, tracks displaying epigenetic data from either total $CD4^+$ T cells or $CD4^+$ central memory T cells were selected. The eight group A SNPs are displayed as vertical dashed lines. From the ENCODE Project, conservation (Cons) between 100 vertebrate species identifies regions

that are conserved calculated by PhastCons and represent probabilities of negative selection. The transcription factor ChIP-seq (TF ChIP) is a merge of data from 161 transcription factors in 91 cell types and a grey box encloses each peak cluster of transcription factor occupancy, with the darkness of the box being proportional to the maximum signal strength observed. From the Roadmap Epigenomics Project, DNase I hypersensitivity (DNase) shows regions of open chromatin, MRE-seq shows regions of demethylated DNA, whereas MeDIP-seq shows regions of methylated DNA. Histone H3K4me1 peaks are suggestive of enhancer activity, H3K27ac distinguishes active enhancers from poised enhancers and H3K4me3 is enriched in transcriptionally active promoters. CRISPRa-responsive elements (CaRE) were identified in Jurkat cells transduced with guide RNAs that tiled across the *IL2RA* region and measured guide RNA enrichment in *IL2RA* expressing cells[174]. A custom track was uploaded to the UCSC browser to display the position of the group A SNPs. Genome coordinates are based upon the GRCh37/hg19 build.

From this analysis, it was clear that rs61839660 was located in an area of epigenetic interest, and based on the MRE-Seq and MeDIP-Seq data was located in a region of demethylation in memory $CD4^+$ T cells. Inspection of the DNA sequence revealed there were six 6 CpG sites within close proximity of rs61839660, but the SNP did not make or break a CpG site. Therefore I chose to see if I could validate and extend this data by analysing the methylation of the six CpG sites flanking rs61839660 in multiple purified immune populations *ex vivo* as well as post 16 hour stimulation (Fig. 5.6). The six CpG sites were extremely demethylated in all cell types examined regardless of cell activation state. $CD14^+$ monocytes displayed a slightly lower level of demethylation, with the 6 CpG sites being 91% demethylated, whereas all of the other populations examined were more than 98% demethylated (Fig. 5.6). As the cell populations were completely demethylated, I did not pursue allele-specific methylation experiments to test if there was differential methylation by the alleles of rs61839660. However, as this region is demethylated in all

cell types examined, it does raise the question if the protective allele of rs61839660 has an effect in other cell populations, apart from the published increase in CD25 expression on $CD4^+$ central memory T cells[37,175,176]. Therefore, the next steps were to test the candidacy of rs61839660 as the causal variant of the *IL2RA* A haplotype, and see if rs61839660 alters the expression of *IL2RA* / CD25 in other cell populations.

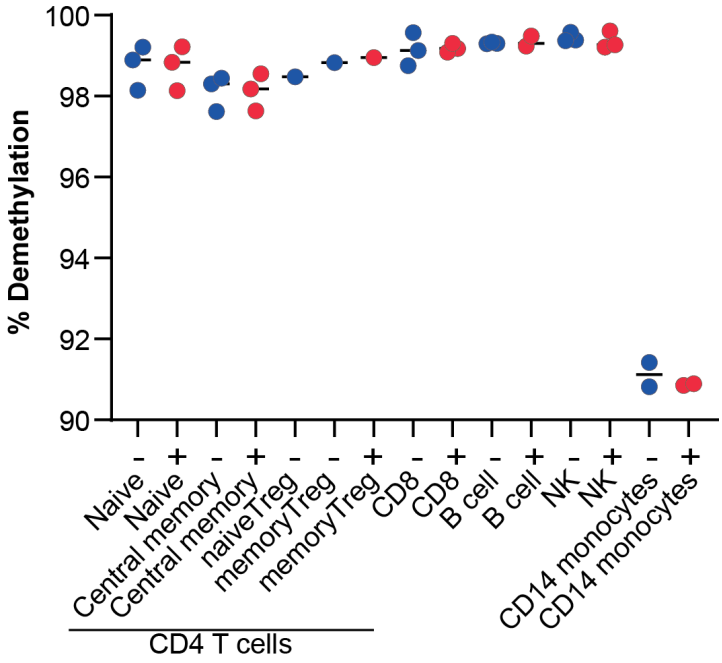


Figure 5.6. Assessment of the methylation status of the six CpG sites flanking rs61839660. Within 180 bp of DNA that contains the SNP rs61839660 lie six CpG sites, for which I have measured the average level of demethylation in purified immune cells *ex vivo* (-, blue circles) and after 16 hours of stimulation (+, red circles) from up to three donors.

5.3.3 Identification of rare *IL2RA* haplotypes

We chose a novel approach to test whether SNP rs61839660 causally affects the expression of *IL2RA* mRNA, using cells from individuals with a rare *IL2RA* haplotype identified from the Cambridge BioResource, a large panel of volunteers recallable by specific genotypes. By phasing dense genotype information across the *IL2RA* region we identified a rare variation (0.6% frequency, Table 5.2) of the T1D-protective group A haplotype in which

all of the protective alleles are present except at rs61839660, where the protective T allele has been replaced with the susceptible C allele (called A^{-660}, Fig. 5.7, Table 5.2).

Comparison of the A and A^{-660} haplotypes using 1000 Genomes Project[105] whole-genome sequence data revealed that the A and A^{-660} haplotypes share a remarkably similar genetic background (across 70 kb, chr10:6,055,500-6,126,500; Fig. 5.7) suggesting the A and A^{-660} haplotypes have arisen together through an ancient bottleneck event. Through this many alleles (including the A haplotype SNPs) have been fixed, when compared to the susceptible haplotype that shows a greater proportion of segregating alleles (A compared to susceptible, Fig. 5.4; A compared to A^{-660}, Fig. 5.7). One additional SNP (rs143054291) differed between the A and A^{-660} haplotypes across all individuals (Fig. 5.7B). As this A^{-660} haplotype-unique SNP is found on all donors having the A^{-660} haplotype from different world populations, it suggests a single gene conversion event[188] created the A^{-660} haplotype and SNP rs143054291 arose at a very similar point in time. We infer that the gene conversion event introduced a short piece of DNA from a non-A haplotype replacing the DNA segment including rs61839660 as well as an in/del polymorphism 92 bp centromeric to the SNP (Fig. 5.7). Based on this sequence comparison, we hypothesised that if rs61839660 does alter CD25 expression on $CD4^+$ memory T cells[37,175,176], we would expect A^{-660} donors to have CD25 expression levels comparable to the non-A haplotypes, which carry none of the group A T1D protective alleles[2]).

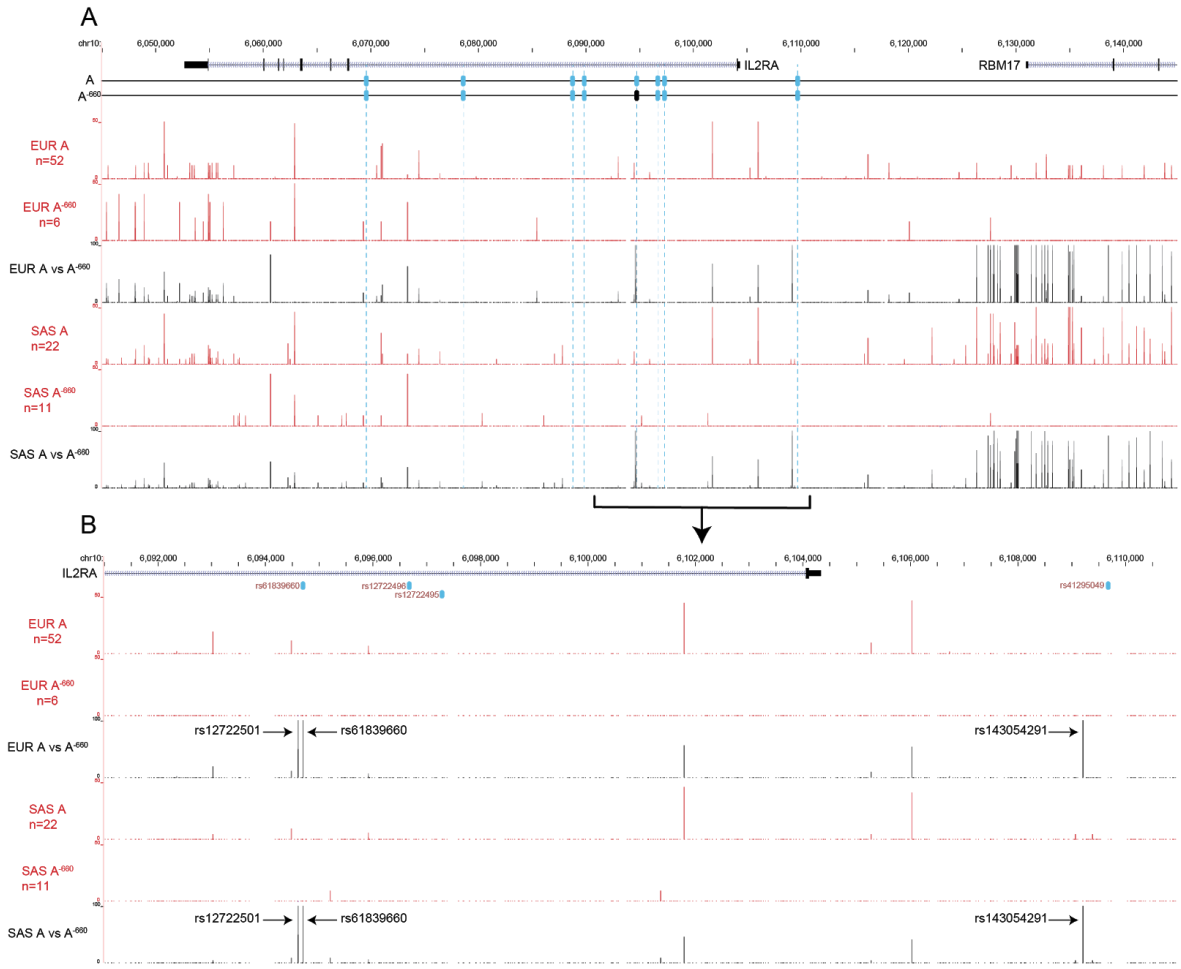


Figure 5.7. Genotype similarity across 3,154 SNPs from donors carrying either the A or the A^{-660} haplotypes identified from the 1000 Genomes Project using the European (EUR) and South Asian (SAS) populations. (A) 100 Kb region centered on the *IL2RA* group A SNPs and (B) zoomed in view highlighting that as well as rs61839660, two other variants differ between the A and A^{-660} haplotypes. The indel rs12722501 is 92 bp centromeric of rs61839660 was most likely introduced along with the gene conversion event that replaced the protective T allele of rs61839660 with the susceptible C allele. The A haplotype always has the insertion, whereas the A^{-660} haplotype always has the deletion at rs12722501. The susceptible haplotype which probably donated the inserted DNA to the A^{-660} haplotype, is segregating equal proportions of the insertion and deletion of rs12722501. The SNP rs143054291 must have arisen at a similar time as the A^{-660} haplotype formed. The height of each line represents the minor allele frequency on the A and A^{-660} haplotypes for each SNP.

5.3.4 High throughput allele-specific expression with the A haplotype

The rarity of the A^{-660} haplotype (Table 5.2), makes the study of A^{-660} homozygous individuals impractical even in the large panel of donors available through the Cambridge BioResource (n=2748). We therefore developed a high throughput method to measure allele-specific expression (ASE), that was modified from the method presented in Chapter 3 to assess the methylation of the *FOXP3* TSDR to assess the relative amount of *IL2RA* mRNA produced from each chromosome within individuals heterozygous for the A or A^{-660} haplotypes and a non-A haplotype (A x non-A, and A^{-660} x non-A, respectively). The ASE method developed uses high-throughput sequencing and is based on a protocol I developed to perform targeted analysis of methylation[91] and allows hundreds of samples to be analysed in a cost effective manner and is a great improvement over previous methods I have developed to measure ASE that used either PCR cloning and sequencing[1] or measuring fluorescence intensity of an RFLP digest[189]. Using ASE to assess the relative amount of *IL2RA* mRNA produced by the each chromosome is a powerful technique and controls for the between individual biological variation resulting from other genetic and environmental differences.

To validate our high throughput method to measure ASE, we first reproduced the finding that the group A protective haplotype produces more *IL2RA* mRNA in $CD4^+$ memory T cells compared to non-A haplotypes[37] when measured by NGS in A x non-A donors using a reporter SNP in intron 1 (rs12722495, allelic ratio 1.44, p=2.6x10^{-07}, Fig. 5.8A) or in the

3′ UTR (rs12244380, allelic ratio 1.44, p=2.1x10^{-07}, Fig. 5.8B). Next we assessed the ASE in A^{-660} x non-A donors in CD4^{+} memory T cells and found significantly reduced allelic ratios (0.84 at rs12722495, p=5.7x10^{-06}, Fig. 5.8A; 0.92 at rs12244380, p=1.7x10^{-04}, Fig. 5.8B), a level close to that from the genomic DNA control (allelic ratio ~1 at rs12722495 and rs12244380 Fig. 5.8A-B).

From within the Cambridge BioResource, we identified one recallable donor out of 2,748 individuals heterozygous for the A^{-660} and A haplotypes. ASE was performed using mRNA from purified CD4+ central memory T cells using only the reporter SNP rs12244380 since the donor was homozygous at rs12722495. Compared to cells from two A-homozygous donors, which displayed nearly equal mRNA expression from both alleles, cells from the A^{-660} x A donor produced approximately twice the amount of *IL2RA* mRNA from the A allele as compared to the A^{-660} allele (Fig. 5.8C).

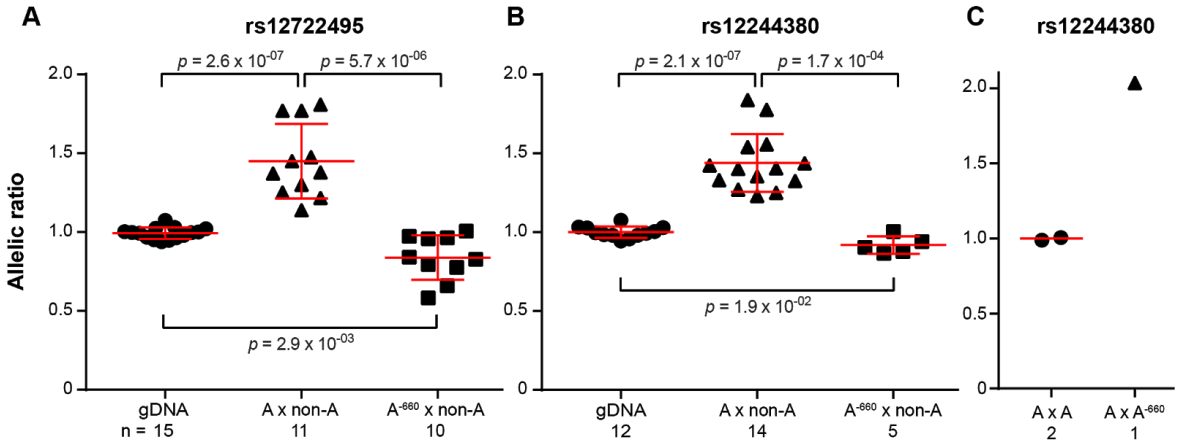


Figure 5.8. This is Figure 1 from "A rare *IL2RA* haplotype identifies SNP rs61839660 as causal for autoimmunity" that I wrote and is published on BioRxiv[190]. The legend reads: Fig.1 SNP rs61839660 determines allele-specific expression of *IL2RA*. (A-C) ASE of *IL2RA* mRNA in CD4^{+} central memory T cells. The allelic ratio of *IL2RA* mRNA made by the chromosome carrying the A or A^{-660} haplotype compared to that carrying the non-A

haplotype was measured using reporter SNP rs12722495 (A) located in intron 1 of *IL2RA*, and rs12244380 (B-C) located within the 3′ UTR of *IL2RA*. Red lines show mean and SD. P values were calculated using a Wilcoxon rank sum test. The gating strategy used to sort $CD4^+$ central memory T cells is displayed in Figure 5.9.

These ASE results, which show donors carrying the A haplotype make more *IL2RA* in $CD4^+$ central memory T cells when compared to non-A donors, and donors that lack the protective (T) allele of rs61839660 on an otherwise complete A haplotype (A^{-660}) do not show an allelic imbalance in *IL2RA* expression provides evidence that rs61839660 has a key role in the regulation of *IL2RA* expression in $CD4^+$ central memory T cells.

In addition to memory $CD4^+$ T cells, *IL2RA* mRNA / CD25 protein is expressed on many different immune populations; therefore to extend the analysis of the A haplotype and to test if the effect of the A haplotype is restricted to $CD4^+$ central memory T cells, up to ten different immune populations were purified by cell sorting and tested for *IL2RA* ASE. The cell populations sorted were $CD4^+$ naive and central memory T cells, Tregs, $CD8^+$ T cells, B cells, NK cells, $CD14^+$ monocytes and $CD16^+$ monocytes from frozen PBMC, recalled by *IL2RA* genotype from the Cambridge BioResource. An example gating strategy is shown in Figure 5.9, and where there were enough cells, I also activated the cell populations for 4 and 16 hours.

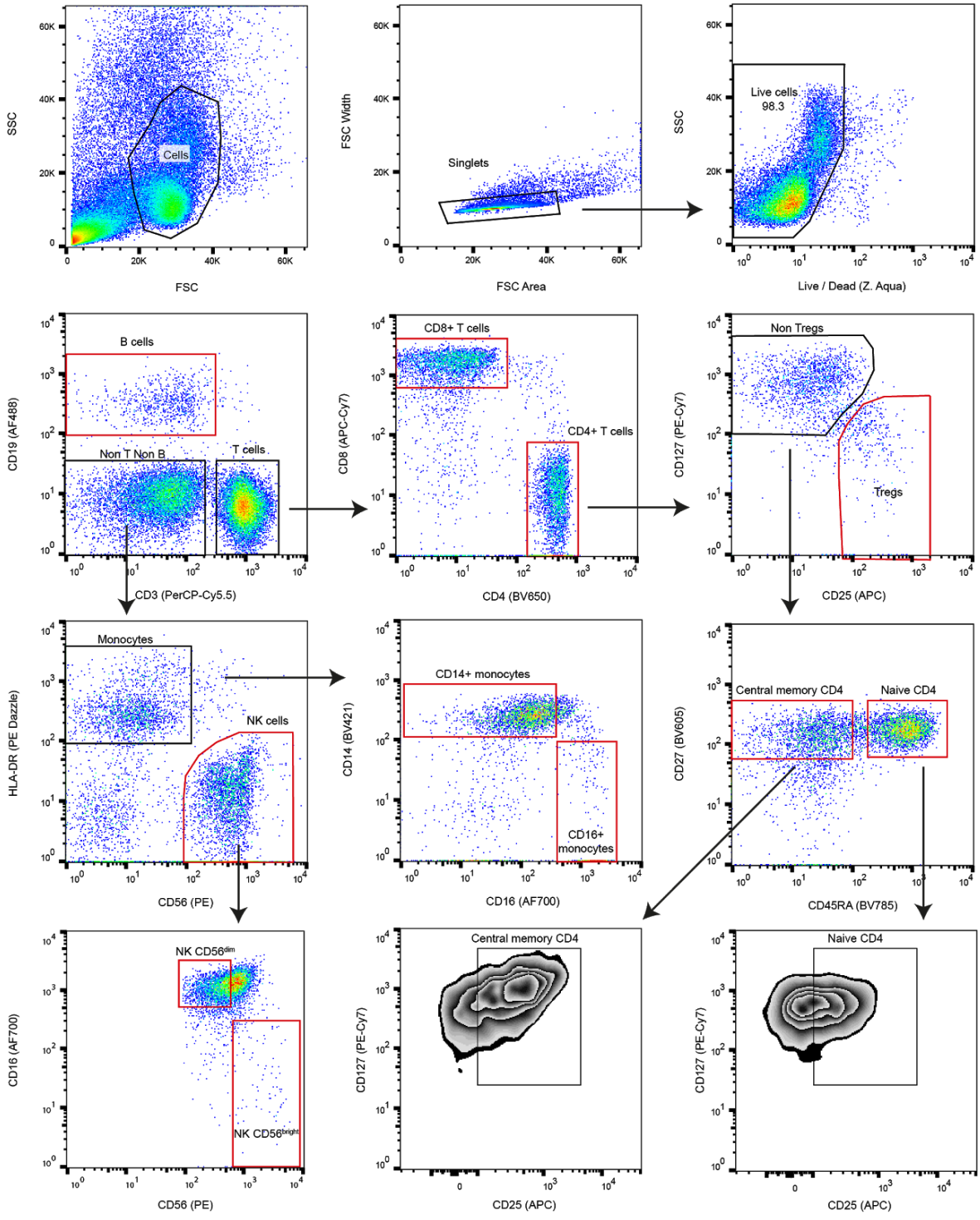


Figure 5.9. Example gating strategy used to sort up to ten immune populations starting from frozen PBMCs using a BD Influx cell sorter. The red gates depict the final populations collected, and the x and y axes display the antibody and fluorochrome used. The zebra plots display the expression of CD25 on naive and central memory $CD4^+$ T

cells with a gate highlighting cells expressing CD25. The methods section 2.3.2 contains detailed information regarding the staining protocol.

5.3.5 Expression of *IL2RA* in purified cells

From the purified immune cell populations I have extracted RNA and made cDNA (Method section 2.3.3) from both the stimulated and unstimulated cells. Before proceeding to ASE experiments, I analysed the expression by qPCR of *IL2RA*, *RBM17* and *PFKFB3*, which are the genes spanned by the *IL2RA* A, C, D, E and F haplotypes, to ensure these genes are expressed to detectable levels and that the cell activation had worked. The qPCR data shows that there is a wide range of *IL2RA* expression that depends on cell type and activation state. Monocytes have extremely low to undetectable levels of *IL2RA ex vivo*, whereas Tregs have approximately 4,000 fold higher expression and are known to constitutively express high levels of *IL2RA* mRNA / CD25. When the cells were activated, the expression of *IL2RA* goes up dramatically in all populations, and the large differences seen *ex vivo* are reduced (Fig. 5.10A). The genes *RBM17* and *PFKFB3* are distal to *IL2RA* and in close proximity to some of the *IL2RA* group E and F SNPs and could potentially exert their effect on the expression of *RBM17* or *PFKFB3*, therefore their expression was measured (Fig. 5.10B and C). The expression of *RBM17* and *PFKFB3* is more stable across all populations examined and only a small (approximately double) increase with activation was seen; although, the basal expression of *RBM17* and *PFKFB3* are substantially higher than the level of *IL2RA* measured in most of the *ex vivo* populations.

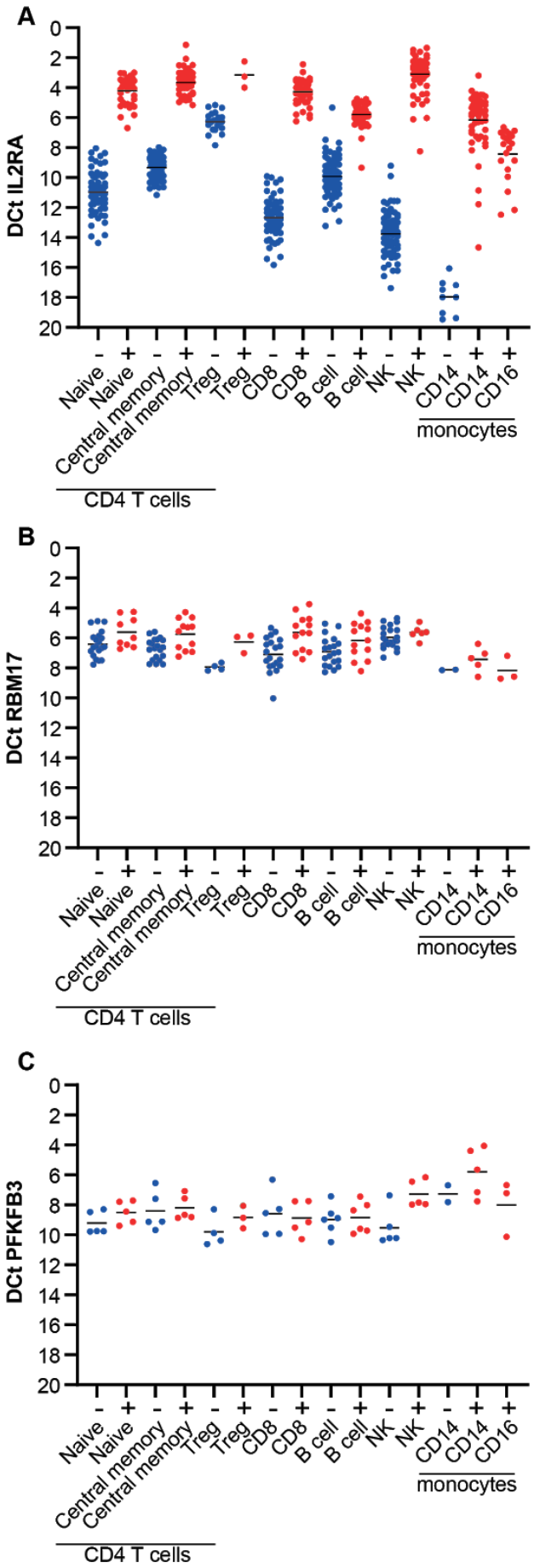


Figure 5.10. Expression of *IL2RA*, *RBM17* and *PFKFB3* measured by qPCR from flow sorted cells. Blue circles are unstimulated cells (-) and red circles are stimulated (+) for 16 hours. T cells were stimulated with CD3/CD28 beads, B cells with CpG, NK cells with IL-12, IL-15 and IL-18, and monocytes with GMCSF and IFNɣ. The Ct of *IL2RA*, *RBM17* and *PFKFB3* was subtracted from the Ct of the housekeeping gene, *B2M*, to generate the DCt, where a lower value means higher expression, and consequently the y axis is reversed.

5.3.6 The A haplotype makes more *IL2RA* mRNA

The high throughput method I have developed to measure ASE has repeated the previous findings that the *IL2RA* A haplotype makes more *IL2RA* mRNA / CD25 (Fig. 5.8). As *IL2RA* is expressed in many different immune populations, I have sorted purified immune populations of cells from donors heterozygous for the *IL2RA* A haplotype, with either the susceptible or E haplotype, and showed by qPCR that *IL2RA* is detectable in all cell populations, but the degree of expression varies widely (Fig. 5.10). I have only used donors that are heterozygous with either the susceptible or E haplotype to make interpretation of the results simpler, by omitting the influence of either the C or D haplotype which are protective from T1D and MS, respectively. Results from donors heterozygous for the A and C haplotypes as well as the A and D haplotypes will be discussed later in this chapter (section 5.5 and 5.6, respectively).

I have performed ASE on these purified populations, as well as on genomic DNA, which serves as a control to show there is no bias in the assay as the average of all the gDNA samples is 50%, however there is a spread in the data with any measure within 2% of the 50:50 being in the range of 'noise' for the assay (Fig. 5.11). As well as the A haplotype making more *IL2RA* in central memory $CD4^+$ T cells, the same effect is also seen in naive $CD4^+$, effector memory $CD4^+$, $CD8^+$ T cells and most strikingly in $NKCD56^{bright}$ cells *ex vivo* (Fig. 5.11), where the A haplotype makes 1.75 fold more *IL2RA* mRNA. The A haplotype does not appear to affect $NKCD56^{dim}$ cells, although the level of *IL2RA* expression is about 8 fold lower than $NKCD56^{bright}$ cells, and much harder to measure

robustly. Tregs and B cells do not show either a consistent or significant difference in expression of *IL2RA* driven by the A haplotype (Fig. 5.11).

When cells were stimulated, the level of *IL2RA* expression increases more than 30 fold and the effect of the A haplotype is lost and it even makes less *IL2RA* mRNA than the haplotype it is heterozygous with (Fig. 5.11). This is seen in all cell populations where there was enough cells to test: naive and memory $CD4^+$ T cells, $CD8^+$ T cells and $CD14^+$ monocytes, where there is virtually no *IL2RA* mRNA expression *ex vivo* (Fig. 5.10).

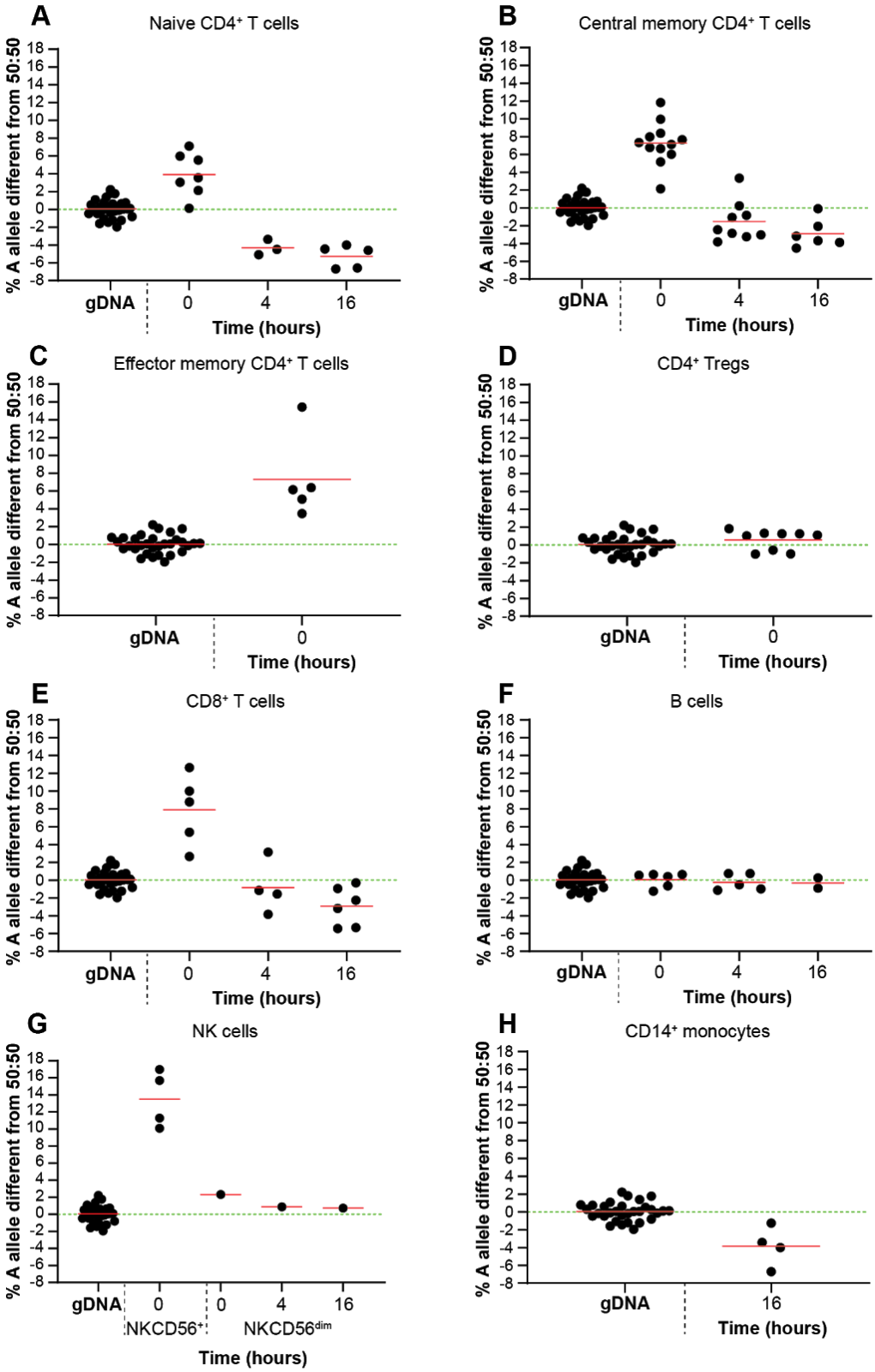


Figure 5.11. Allele-specific expression (ASE) analysis of the *IL2RA* A haplotype in purified immune cells. Various immune cell populations were purified by cell sorting using the gating strategy in Figure 5.9 and were tested for ASE *ex vivo* or after cell stimulation

(4 and 16 hours). T cells were stimulated with CD3/CD28 activation beads, B cells with CpG, NK cells with IL-12, IL-15 and IL-18 and $CD14^+$ monocytes with GMCSF and IFNɣ. Donors were heterozygous for the A haplotype with either the susceptible or E haplotype and the SNP rs12244380 in the 3′ UTR of *IL2RA* was used to measure the degree of allelic imbalance that is reported as the percent made by the A haplotype. The gDNA control shows equal amplification of the *IL2RA* alleles.

5.3.7 Immunophenotyping the DIL PBMC bank collection

It has been previously published that the *IL2RA* haplotype tagged by the SNP rs12722495 (now known as the group A haplotype) is associated with increased expression of CD25 on the surface of $CD4^+$ memory T cells[37] and this observation has been replicated by two different groups[175,176]. Our group has recently completed a new collection of immunophenotyping data from 522 individuals recruited through the Cambridge BioResourse called the DIL PBMC bank collection. All participants have been genotyped using the Illumina Immunochip genotyping array that allows the analysis of genotype to phenotype correlations. The collection and staining of the samples was performed by the Sample Processing Team within the DIL and flow cytometric analysis was performed by Marcin Pekalski (DIL).

The flow cytometric gating strategy is shown in Figure 5.12, and I received the mean fluorescence intensity (MFI) values of CD25 expression on $CD4^+$ central memory T cells, naive Tregs, memory Tregs, $CD8^+$ central memory T cells, $CD27^+$ memory B cells and the frequency of CD25 expressing $CD4^+$ naive T cells from the 522 donors collected over a 708 day period.

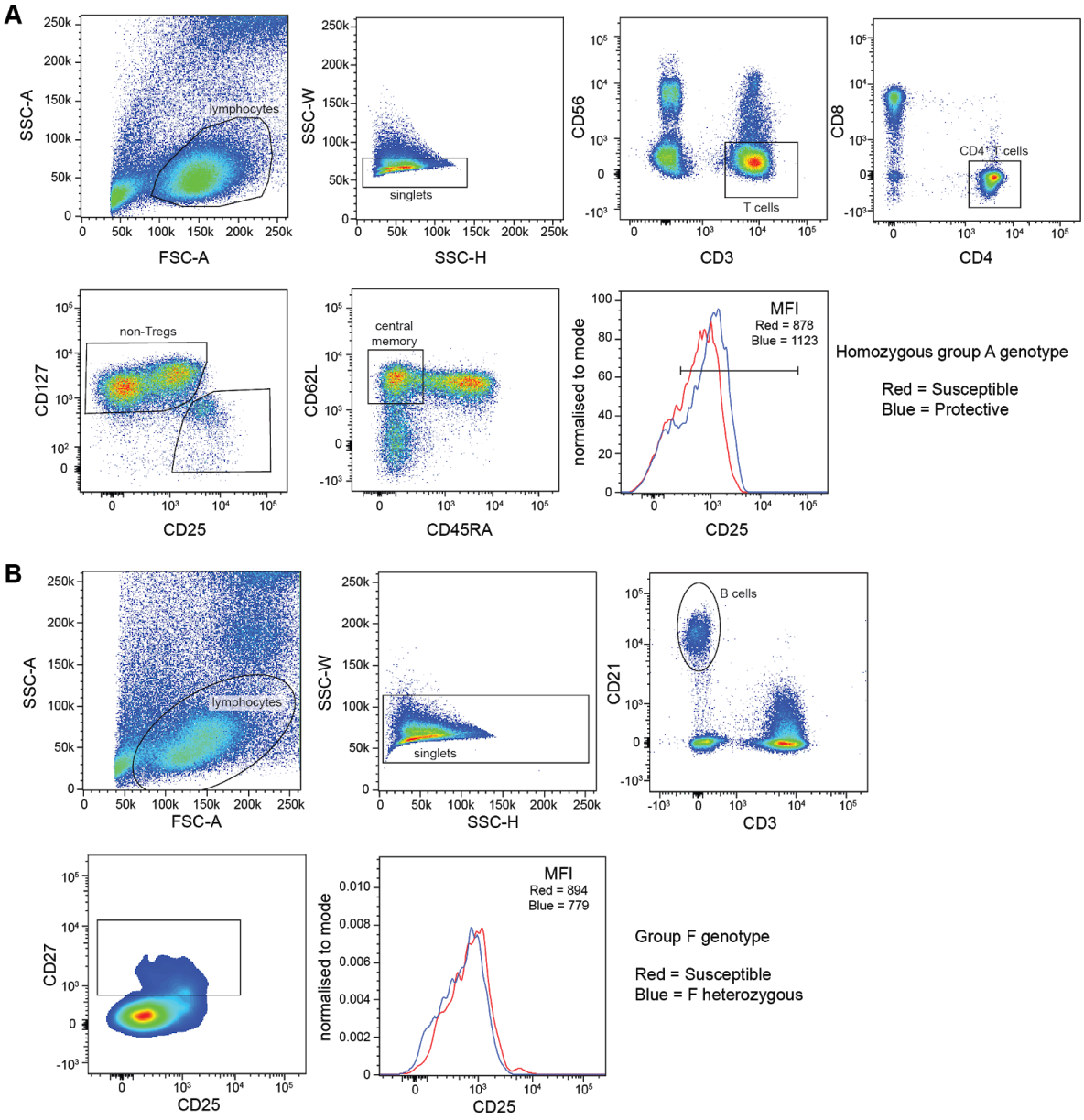


Figure 5.12: CD25 expression on CD4$^+$ central memory T cells and CD27$^+$ memory B cells. (A) Example of the gating strategy used to identify CD4$^+$ central memory T cells from blood, and representative examples of CD25 median fluorescence intensity from a homozygous group A protective and susceptible individual. The horizontal bar indicates the level of CD25 expression used to calculate median fluorescence intensity, with naïve CD4$^+$ T cells used to set the background expression level (lower limit of bar). (B) Example of the gating strategy used to identify CD27$^+$ memory B cells from blood, and representative examples of CD25 median fluorescence intensity from a heterozygous group F protective donor versus a susceptible individual.

5.3.7.1 Data collection and normalisation

The DIL PBMC collection took 708 days to collect blood samples from 522 participants from the Cambridge BioResource. From each blood sample, flow cytometry was performed on the same day as the blood sample was received. Unfortunately, at day 458 of the collection, the red laser of the flow cytometer failed and was replaced as well as additional essential maintenance being performed. The antibody to CD25 is conjugated to APC, which is excited by the red laser; therefore the data were analysed to determine if any batch effects were present related to the performance of the red laser within the flow cytometer. The data I am presenting is for $CD4^+$ central memory T cells, however the same analysis was performed on all cell populations analysed. Clearly there had been a decline in the performance of the red laser in the months prior to it failing, as a linear decrease in the MFI of CD25 was observed (Fig. 5.13).

After the red laser had been replaced, there was a slight increase through time of the CD25 MFI. Therefore the data collected prior to day 458, when the red laser was replaced and the data collected after needed to be treating as two batches (ANOVA P value $2.2x10^{-8}$, Fig. 5.13). The initial analysis and normalisation I performed on $CD4^+$ central memory T cells and is described below, and I then subsequently applied the same normalisation strategy to naive $CD4^+$ T cells, naive and memory Tregs and B cells.

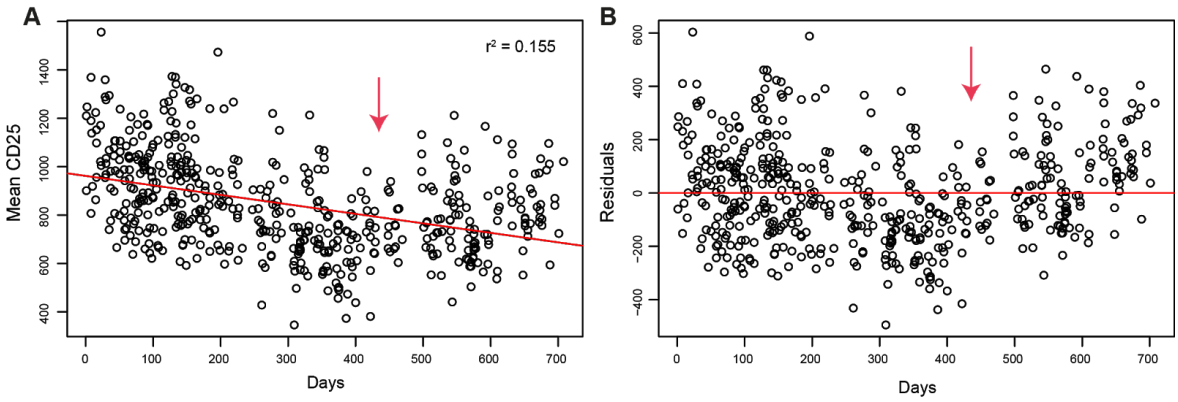


Figure 5.13. Decline in CD25 MFI on $CD4^+$ central memory T cells due to failure of the red laser within the flow cytometer. (A) The CD25 MFI on $CD4^+$ central memory T cells was measured in 522 samples collected over a 708 day period. The red laser failed and was replaced on day 458 (red arrow). Prior to the red laser failing there was a steady decrease in the MFI of CD25. (B) The residuals from the linear regression in (A) show a non-random distribution around zero, therefore the data should be treated as two batches (ANOVA P value 2.2×10^{-8}).

I will analyse the data as two batches; batch 1 will be prior to the red laser failure and batch 2 will be samples collected after the red laser was replaced. Batch 1 (pre red laser failure) contains 384 participants and during collection the average CD25 MFI declines from 1100 to 700, and I assessed if the decline was independent of genotype before assessing the genotype affects CD25 expression. It is known that the A haplotype alters the expression of CD25 expression on central memory $CD4^+$ T cells, but the frequency of donors who are homozygous for the A haplotype in the collection are rare, therefore I tested donors heterozygous for the A haplotype compared to donors who do not carry the A haplotype. Figure 5.14 shows that genotype does not alter the rate of decline of CD25 expression in Batch 1 (ANOVA P=0.65), therefore all samples in Batch 1 can be treated together to normalise for the decrease in CD25 MFI. Batch 2 contained 138 samples and

showed a smaller change in CD25 expression during the collection from day 458 to day 700 than Batch 1. I performed the same analysis to test if the change in CD25 expression during Batch 2 collection was related to genotype and showed that genotype had no affect (ANOVA P=0.57, Fig. 5.14) and so all samples in Batch 2 could be normalised together.

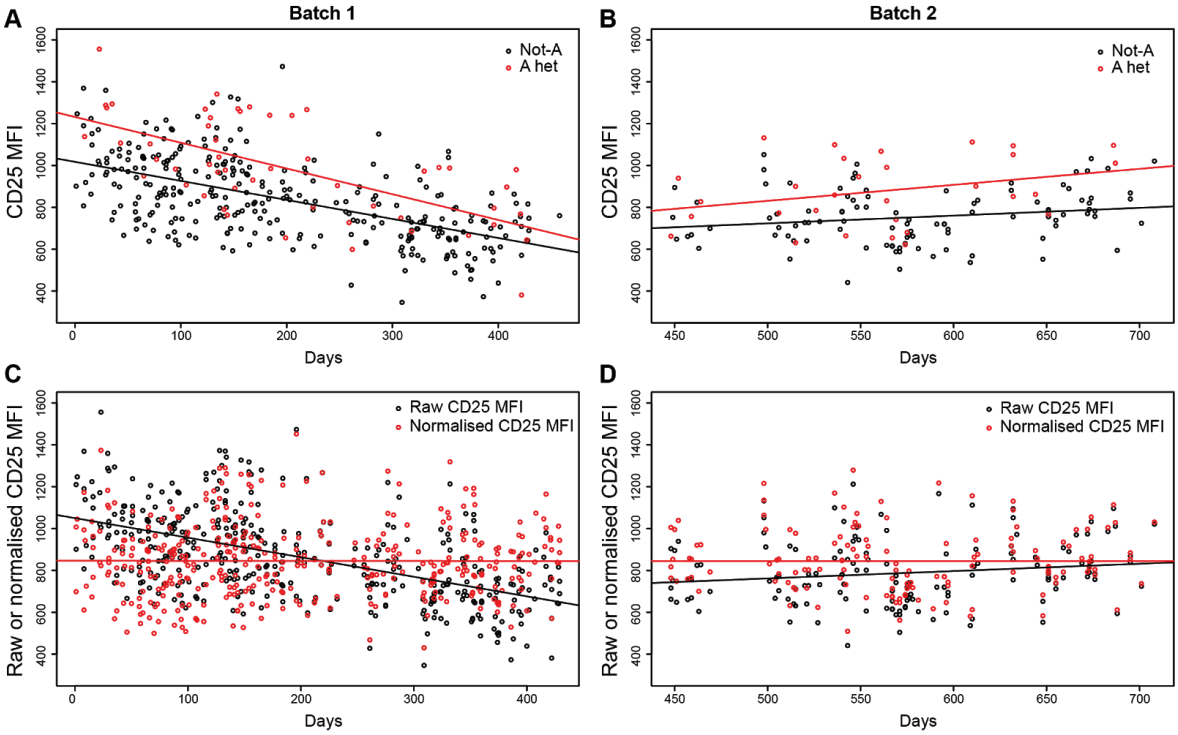


Figure 5.14. Testing if *IL2RA* genotype affects the decline of CD25 expression in batch 1 and batch 2 of the collection. (A and C) are samples from batch 1 with (B and D) representing samples from batch 2. Donors heterozygous for the A haplotype (A and B, red circles) were compared to donors not carrying the A haplotype (C and D, black circles) to see if there was any difference in the rate of change of CD25 MFI based on genotype, and no difference was observed, Batch1 ANOVA p=0.65, Batch 2 ANOVA p=0.57. As there was no difference in the change of CD25 expression based on genotype, all samples within each batch were combined (C and D, black circles). The slope from linear regression was used to normalise for the change in CD25 expression through time (C and D, red circles).

The CD25 MFI for batch 1 and batch 2 were normalised separately using the slope of the linear regression for each batch, and then were normalised to the mean of the entire dataset so both batches can be combined (Fig. 5.15). This normalisation was performed to allow genotype to phenotype analysis to be performed on the maximum number of samples.

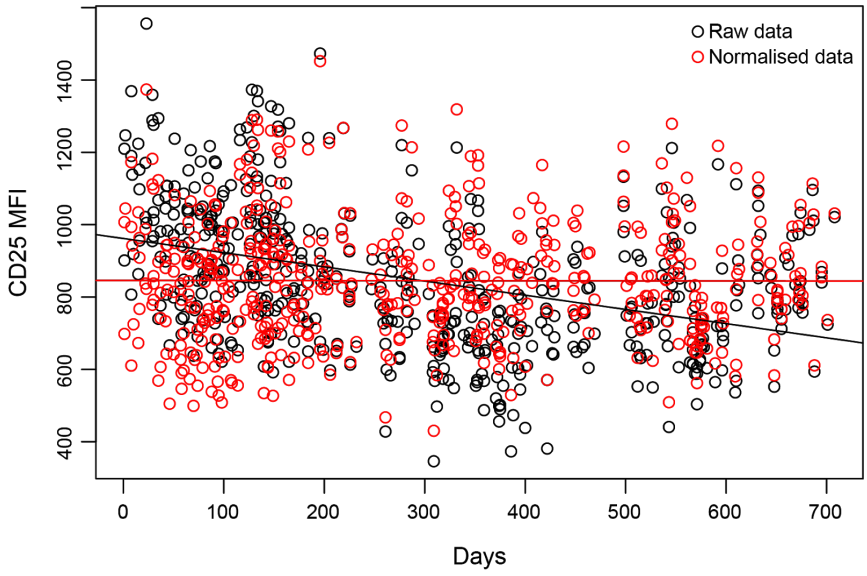


Figure 5.15. Comparison of the raw and normalised CD25 MFI from 522 samples collected over 700 days. The raw values (black circle) show a decline in CD25 MFI, whereas the normalised values (red circles) have been corrected to account for decline in performance of the red laser in the flow cytometer and show no change over time.

5.3.7.2 Group A immunophenotyping: genotype to phenotype

I have completed the normalisation of CD25 expression on $CD4^+$ naive and central memory T cells, $CD4^+$ naive and memory Tregs and B cells from 522 donors who all have dense genotyping information from the Immunochip genotyping platform. Jamie Inshaw (DIL) then parsed this data through the stochastic search program, that was used successfully to dissect the genetic association of *IL2RA* in an unbiased way[2,104], which was

a novel use of this method, and I hope that any complex genetic association with CD25 expression would be successfully deconvoluted.

The immunophenotyping results have replicated two previously identified genotype-phenotype associations which are the *IL2RA* A haplotype with higher CD25 expression on $CD4^+$ central memory T cells[37,175,176], and the D haplotype with a lower frequency of $CD4^+$ naive T cells that express CD25[REF2,37] (Fig. 5.16A-B). The D haplotype will be presented in section 5.6, but I am including the results to show there is no association of the A haplotype with the frequency of $CD4^+$ naive T cells that express CD25. The stochastic search analysis did not detect any further genetic association with CD25 expression in $CD4^+$ naive and central memory T cells.

Interestingly, there was no association of the *IL2RA* A haplotype, or any of the haplotypes with the expression of CD25 on either $CD4^+$ naive or memory Tregs. The CD25 MFI on these Treg subsets is much higher than that on non-Treg cells by at least three fold (Fig. 5.16C-D), a similar difference to that seen in the qPCR results for *IL2RA* expression (Fig. 5.10), suggesting that there must be more or stronger enhancer elements that regulate the expression of CD25 on Tregs which may override any of the effects of the SNPs associated with the *IL2RA* A haplotype.

The immunophenotyping of B cells included CD27, which allowed the analysis of $CD27^-$ naive B cells and $CD27^+$ memory B cells, however I am only presenting the results from the $CD27^+$ B cells, although similar trends were seen in the $CD27^-$ naive B cells, but as the frequency of CD25 expressing cells, and the amount of CD25 they expressed is lower than $CD27^+$ memory B cells, there was more variability. Initial analysis of CD25 expression on $CD27^+$ memory B cells, showed that the *IL2RA* A haplotype had no effect, however it appears that there was lower expression of CD25 on the C and F haplotypes (Fig. 5.16E). However, when this data was parsed through the stochastic search algorithm, three signals were identified, which were not in high LD with any of the *IL2RA* A, C, D, E or F haplotypes and these were tagged by rs1887027, rs12722502 and rs71479745, and this observation will be discussed fully in Chapter 6, which focuses on *IL2RA* mRNA / CD25 phenotypes not associated with disease.

As there was an increase in *IL2RA* mRNA made from the A haplotype in $CD8^+$ T cells by ASE (Fig. 5.16F), the immunophenotyping dataset was investigated to see if this could be confirmed by the expression of CD25 on $CD8^+$ T cells. However, the flow cytometric data of the $CD8^+$ T cells showed larger variation over time that could not be normalised, therefore these data were analysed differently. On the same day that a donor for the A haplotype was recruited, all other donors recruited on the same day that did not carry any copies of the A haplotype were analysed as a group (called Non-A), and in this scenario there is no need for normalisation as the A and Non-A donors were subject to the same staining and instrument variability. In Figure 5.16F I show confirmation of the ASE results showing that the A haplotype is associated with increased expression of CD25 on $CD8^+$ T cells.

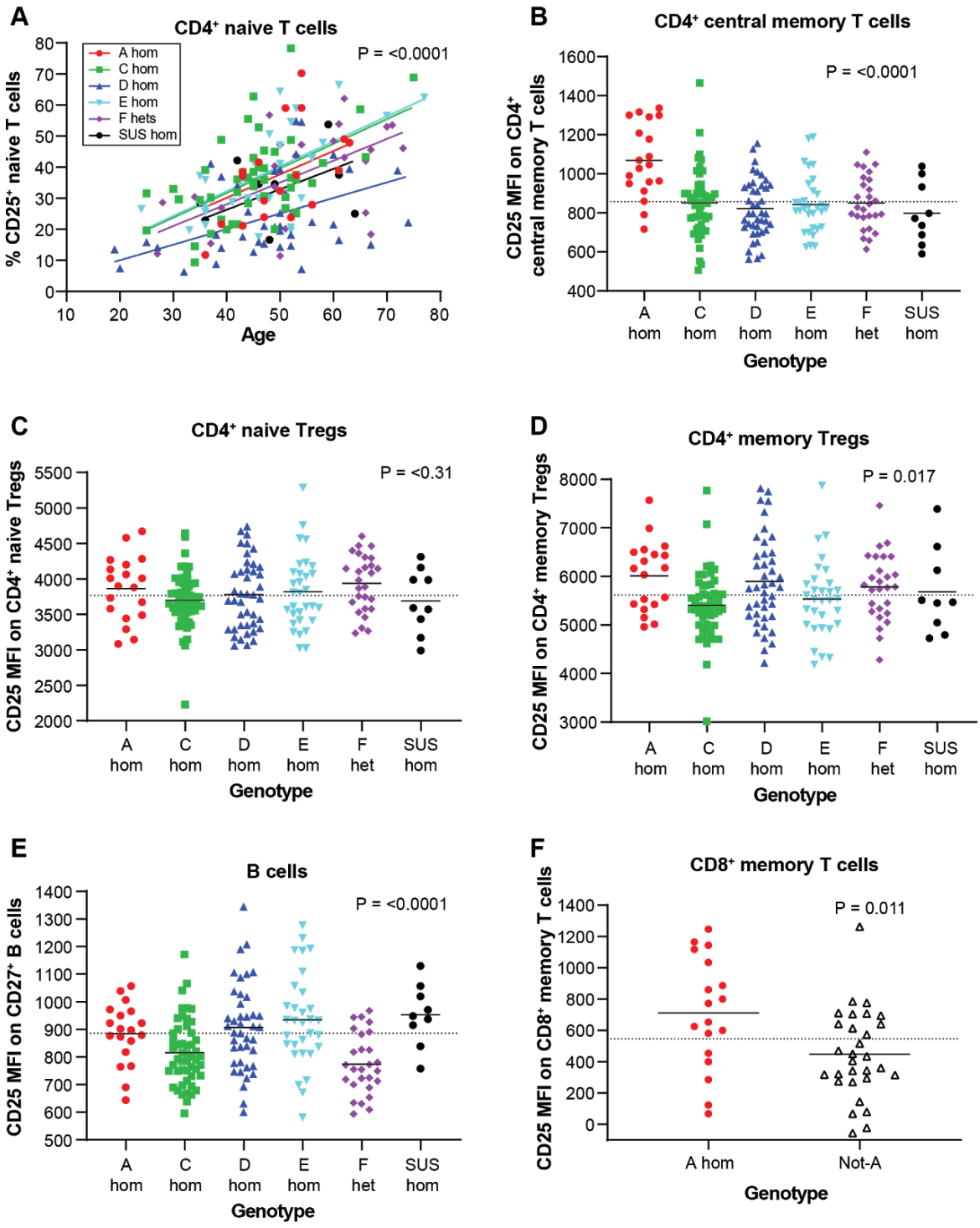


Figure 5.16. Genotype-phenotype correlations of CD25 expression with *IL2RA* genotypes in CD4⁺ (A) naive and (B) memory T cells, (C) naive and (D) memory Tregs, (E) CD27⁺ memory B cells and (F) CD8⁺ T cells. Donors were recruited from the Cambridge BioResource with specific *IL2RA* genotypes, and fresh blood from 522 participants was immunostained with antibodies listed in Tables 2.7 and 2.8 and example staining profiles are shown in Figure 5.12. The immunophenotyping data was split

according to *IL2RA* genotype, and donors homozygous for the A haplotype are displayed as red circles, the C haplotype as green squares, the D haplotype as blue triangles, the E haplotype as inverted cyan triangles, the F haplotype is rare and only found as heterozygous in this dataset and is displayed as purple diamonds. The susceptible haplotype carries no A, C, D, E or F protective alleles and is shown as black circles. The black dashed line represents the mean of the data set. P values for (A-E) were calculated as a one-way ANOVA, and in (F) using a Mann-Whitney test.

There is a gene dosage effect of the protective A haplotype on CD25 expression on $CD4^+$ central memory T cells, with heterozygous and homozygous protective haplotypes increasing surface expression of CD25 by approximtely 15% and 22% respectively compared to the homozygous susceptible haplotype. There is no difference in the CD25 surface expression amongst the C, D, E and susceptible haplotypes, although there is a large variation amongst individuals having a particular genotype, suggesting other factors apart from genotype are affecting the surface levels of CD25 (Fig. 5.16B). Analysis of donors with the A^{-660} haplotype heterozygous with a non-A haplotype (either C, D, E or susceptible) showed decreased CD25 expression compared to the A heterozygous haplotype and equivalent levels as compared to the C, D, E or susceptible haplotype confirming the ASE result that the A^{-660} haplotype makes less *IL2RA* mRNA than the full A haplotype (Fig. 5.17).

Two donors who carried the A^{-660} haplotype heterozygous for the A haplotype displayed CD25 expression equivalent to the A heterozygous group, reinforcing the notion that rs61839660 is key regulator of CD25 expression.

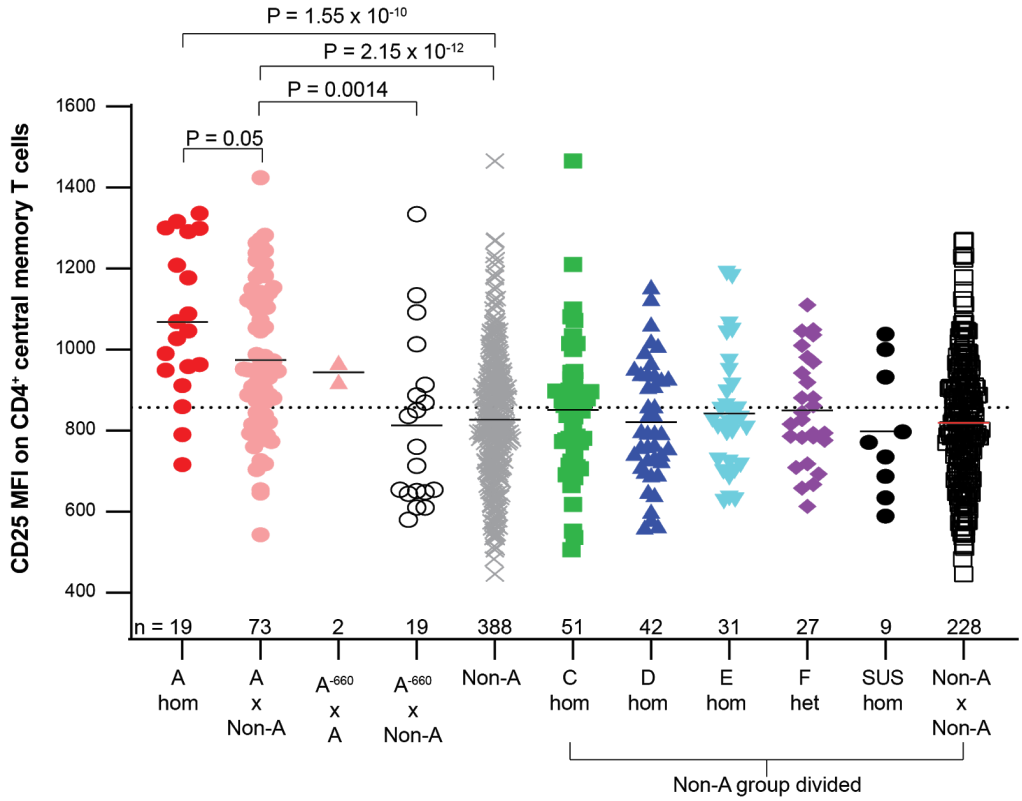


Figure 5.17. The mean fluorescence intensity of the surface CD25 expression on $CD4^+$ central memory T cells is displayed per *IL2RA* haplotype. The non-A haplotype group contains all individuals which have no A haplotype alleles, and this group is also shown broken down into the C, D, E, F and susceptible homozygous haplotypes and all heterozygous haplotypes not involving the A haplotype. P values were calculated in Prism using an unpaired T test.

5.3.8 Interaction of the *IL2RA* promoter with a region containing rs61839660

To understand the importance of how gene transcription can be controlled by enhancer elements that are located at a distance from the transcription start site of a gene, Tony Cutler (DIL) in collaboration with Peter Fraser (Babraham Institute) has performed promoter capture Hi-C (PCHi-C) on activated and non-activated $CD4^+$ T cells[191,192]. Analysis of the *IL2RA* promoter enhancer interaction in resting $CD4^+$ T cells shows that

there is an interaction with the promoter of *IL2RA* to a region within intron 1 of *IL2RA* that contains three group A SNPs including our proposed causal SNP rs61839660 (Fig. 5.18). However, when the $CD4^+$ T cells are activated many more promoter enhancer interactions are seen, suggesting additional control elements are required to drive the increased expression of *IL2RA* seen when cells are activated and may explain why the effect of rs61839660 is lost when cells are activated. In addition to an interaction with the intron 1 region containing the SNP rs61839660, there are also several interactions with the regions upstream of the promoter, some of which contain SNPs from the other *IL2RA* protective haplotypes (Fig. 5.18) and will be discussed further in the section focussing on the C, E and F haplotypes.

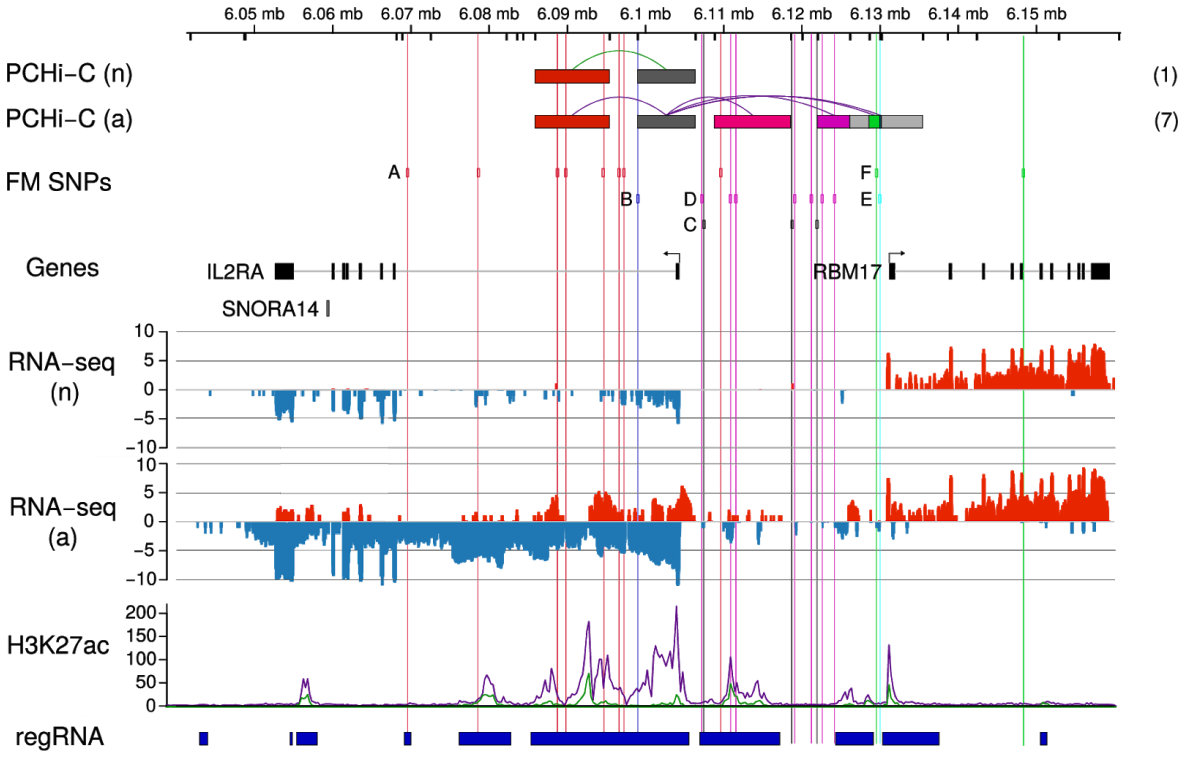
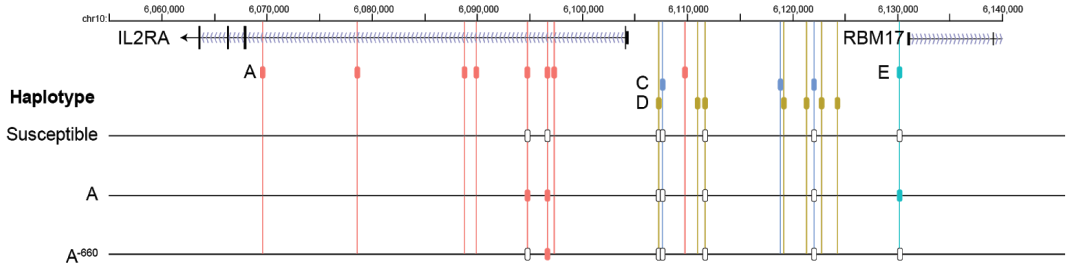


Figure 5.18. Promoter enhancer interactions within the *IL2RA* region. This is figure 6 taken from 'Chromosome contacts in activated T cells identify autoimmune disease candidate genes[191]' for which I was joint first author. The PCHi-C track displays the interaction of the *IL2RA* promoter (grey box) with regions containing A SNPs (red), D

SNPs (magenta), F SNPs (green) or regions with no T1D associated SNPs (light grey). The locations of the T1D associated SNPs are shown along with the *IL2RA* and *RMB17* genes. The RNA-Seq track displays directional reads, with the reads mapping to the forward strand shown in red and reverse strand in blue. The *IL2RA* gene is located on the reverse strand and *RBM17* on the forward strand. Data are shown for unactivated $CD4^+$ T cells (n) and $CD4^+$ T cells following a 4-hour activation with CD3/CD28 T cell activator beads (a).

5.3.9 The A^{-660} haplotype is not protective for T1D

Having shown that increased CD25 expression on central memory $CD4^+$ T cells as well as increased *IL2RA* mRNA levels determined by the A haplotype were not attributes of the A^{-660} haplotype, Dr. Chris Wallace (University of Cambridge) tested the A^{-660} haplotype for T1D disease association. Despite the rarity of A^{-660} haplotypes, we observed lower protection from T1D for the A^{-660} compared to the A haplotype (OR = 1.37, CI=1.015-1.840, p=0.04, Fig. 5.19).



Haplotype	T1D count	T1D %	Controls count	Controls %	Odds ratio	LCI	UCI	P value
Susceptible	2806	21.7	4176	17.7	1.92	1.731	2.12	$<2.00\times10^{-16}$
A+E (reference)	804	6.2	2323	9.8	1.000	-	-	-
A^{-660}	92	0.7	194	0.8	1.37	1.015	1.84	0.0396

Figure 5.19. This is modified Supplemental Figure 5 from "A rare *IL2RA* haplotype identifies SNP rs61839660 as causal for autoimmunity" that I wrote and is published on BioRxiv[190]. The legend reads: Genetic association of the different *IL2RA* haplotypes with

T1D. The schematic shows the location of the group A, C, D and E SNPs that alter the risk of T1D identified in Wallace et al.[2,104]. For the genetic association, only two tag SNPs from group A, rs61839660 and rs12722495 were used and are shown as ovals, with open ovals denoting the T1D susceptible allele and filled ovals denoting the T1D protective allele being each tag SNP. The A^{-660} haplotype was defined as carrying the protective allele at rs12722495 and the susceptible allele at rs61839660, and the susceptible haplotype carries none of the group A, C, D or E protective alleles. The A haplotype (which also carries the E protective allele) was set as the reference (OR = 1) and all other haplotypes were compared to it (see Methods section 2.3.12 for more details). The data used to generate this figure came from Onengut-Gumuscu *et al.*[17].

5.3.10 Summary of the A haplotype

I have shown that the *IL2RA* A haplotype is protective for T1D, MS, AITD, and used a novel strategy to assess which of the eight SNPs identified as tagging the A haplotype is the most likely causal variant. Epigenetic evidence suggested rs61839660 was most likely, and individuals with a rare haplotype in which the protective allele of rs61839660 was lost from an otherwise complete A haplotype were recruited and through ASE showed they lost the increased *IL2RA* mRNA expression associated with the A haplotype. Lower CD25 expression on $CD4^+$ central memory T cells was also observed in the rare A haplotype lacking the protective allele at rs61839660. I have tested other immune cell populations that express *IL2RA*, to see if the effect of the A haplotype is restricted to $CD4^+$ central memory T cells, and found that $CD8^+$ T cells and $NKCD56^{bright}$ cells display the same 1.5 fold increase in *IL2RA* expression with the protective A haplotype. Additionally, I have tested the effect of the A haplotype in activated immune cell populations, and the reverse effect is seen, with the A haplotype making less *IL2RA* in activated $CD4^+$ naive and central

memory T cells, $CD8^+$ T cells and $CD14^+$ monocytes. The *IL2RA* A haplotype did not affect expression on Tregs or B cells.

5.4 The *IL2RA* group C haplotype

5.4.1 Introduction

The *IL2RA* C haplotype has been associated with T1D and AITD, and is defined by a group of three credible SNPs[2,104] which are located intergenic between *IL2RA* and *RBM17* (Fig. 5.2). The *IL2RA* C haplotype could be described as modular, in terms that as well as being found in isolation (44.6%), it is also commonly carried with a variant (rs12722497) that alters the expression of soluble CD25[REF169,193] (sCD25); less commonly the C haplotype is carried with the E and F haplotypes (7.3% and 3.8%, respectively). Figure 5.20 is a cartoon representation of these different flavours of the C haplotype, and shows that rs12722497 and the E and F haplotypes are located at the proximal and distal ends of the C haplotype, suggesting that recombination has allowed these additional features to move between extended haplotypes.

Currently no phenotype has been described for the C haplotype, although variants in good LD ($r^2 \sim 0.7$) have been associated with an activation phenotype, in which less *IL2RA* mRNA is made from stimulated $CD4^+$ T cells[34,26]. Therefore, I plan to analyse the available epigenetic databases for two reasons. Firstly to see if there are any clues as to which cell type might be affected by the C haplotype and secondly if there is any evidence which of the three C SNPs is the more likely causal variant.

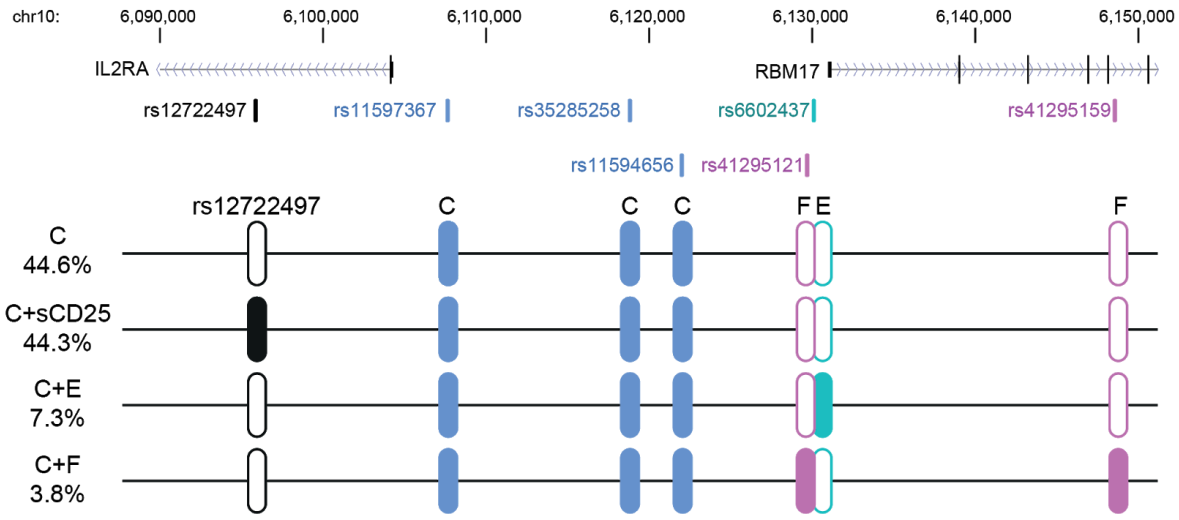


Figure 5.20. The C haplotype is 'modular' and can be found with the E and F haplotypes as well as a SNP associated with higher levels of soluble CD25, rs12722497[REF169,193]. The location of these SNPs are shown relative the *IL2RA* and *RBM17* genes. The frequency of combination of alleles that are present on each of the C haplotypes is shown as a proportion of the total number of C haplotypes. For simplicity, only the two most proximal F haplotype SNPs are shown; the location of all 13 F haplotype SNPs are shown in Fig. 5.1 and span 84 Kb. Filled ovals represent the minor protective alleles and open ovals the major susceptibility allele.

5.4.2 Epigenetic analysis of the *IL2RA* C haplotype

Analysis of epigenetic information from primary cells, reveals that only one of the *IL2RA* C haplotype SNPs, rs11597367, intersects features of interest in immune cells (Fig. 5.21). The SNP rs11597367 lies in both peaks of DNase hypersensitivity (Fig. 5.21A) and ATAC-Seq from NK cells (Fig. 5.21B), as well as near a small, but reproducible peak for activated $CD4^+$ T cells (Fig. 5.21B-C). This information suggests that rs11597367 is the more likely of the three *IL2RA* C haplotype SNPs to be the causal variant, and secondly that NK cells and activated $CD4^+$ T cells need to be assessed as the potential cell type that is affected.

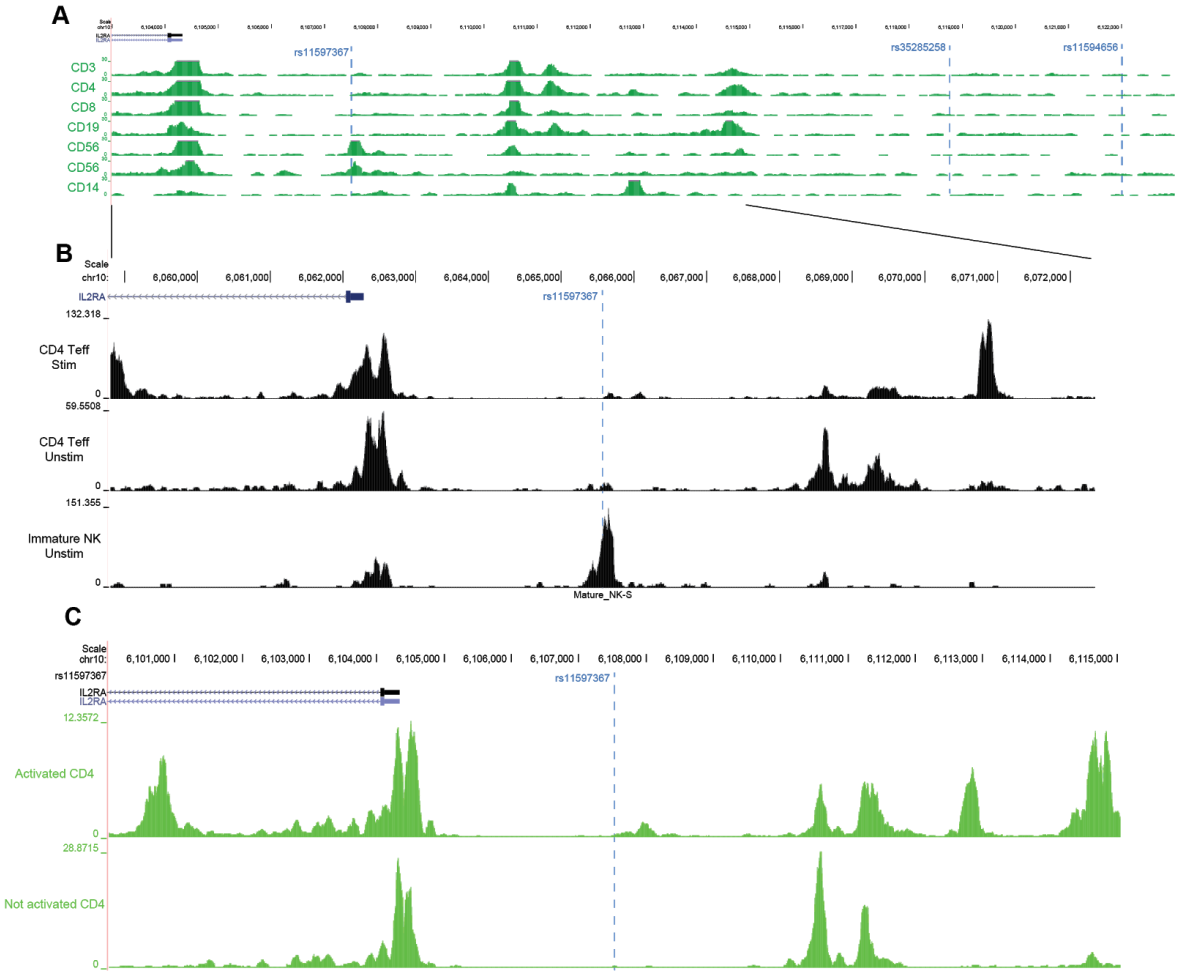


Figure 5.21. Analysis of epigenetic marks around the *IL2RA* C haplotype SNPs. (A) From the Roadmap Epigenomics Project, DNase I hypersensitivity (DNase) shows regions of open chromatin in six broad immune cell populations. (B) ATAC-Seq data from Calderon *et al.*[194] from activated and not-activated $CD4^+$ T effector cells and Immature NK cells (defined as $CD56^{bright}$ and $CD16^{low}$), which displays chromatin accessibility and is focused around rs11597367 due to the epigenetic evidence in (A). (C) Displays ATAC-Seq generated by Tony Cutler (DIL) and displays chromatin accessibility inactivated and not-activated $CD4^+$ T cells. A custom track was uploaded to the UCSC browser to display the position of the group C SNPs (blue dashed lines). Genome coordinates are based upon the GRCh37/hg19 build.

In addition to DNA hypersensitivity and ATAC-Seq data, there is evidence of increased demethylation around rs11597367 in NK cells (Blueprint Project, data not shown).

Inspection of the DNA sequence around rs11597367 showed there were three CpG sites within close proximity to rs11597367, as well as the minor allele of rs11597367 (G) creating a CpG site (Fig. 5.22A). Interestingly, there are examples of SNPs that alter the formation of a CpG site being enriched with disease[195]. Therefore I chose to assess the methylation of the three CpG sites that flank rs11597367, as well as the CpG site that is created by the minor allele of rs11597367.

Analysis of the methylation status of the three CpG sites that flank rs11597367 show that NK cells are the most demethylated, with $CD4^+$ central memory T cells, $CD8^+$ T cells and Tregs with an intermediate level of demethylation and naive T cells, B cells and monocytes predominantly methylated (Fig. 5.22B). From one donor, there was enough cells to perform an activation condition and then measure the methylation around rs11597367 and there was no difference to the unstimulated samples (Fig. 5.22B). As the C haplotype SNP, rs11597367, is located between these CpG sites (Fig. 5.22A), I assessed the allele-specific methylation of the CpG site located on the DNA strand defined by either the minor or major allele of rs11597367 (Fig. 5.22C). In none of the purified cells assessed was there a difference in the level of methylation seen based on the alleles of rs11597367. As the SNP itself, rs11597367 makes (minor allele C**G**) or breaks a CpG site (major allele C**A**), I measured the methylation of the cytosine base prior to the SNP. In the case of the major allele, only a demethylated site was detected, which was expected as the major A allele broke the CpG site (Fig. 5.22D). The minor allele created a CpG site, and the methylation of this site was the same as that measured for the three flanking CpG sites. Examples of the methylation profile measured at these CpG sites are displayed (Fig.

5.22E). The methylation data supports the DNA hypersensitivity and ATAC-Seq data presented in Fig.5.21, which suggested NK cells were an important population to study for the effect of the C haplotype as these were the most demethylated and had the most accessible chromatin meaning this region was active in NK cells.

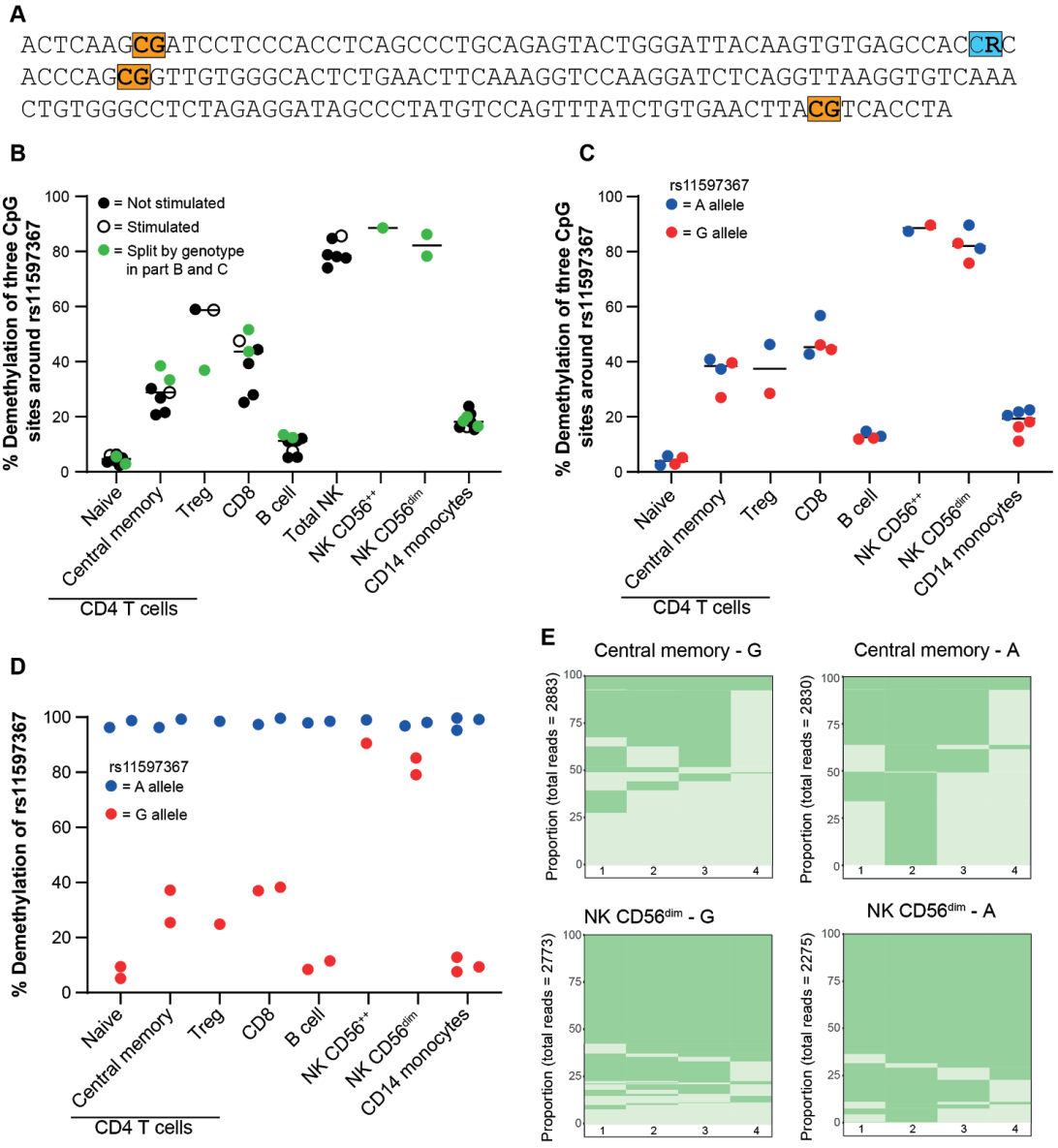


Figure 5.22. Analysis of the methylation sites around rs11597367. (A) The genomic DNA sequence surrounding rs11597367 contains three CpG sites (highlighted orange) and the CpG site created by the minor allele of rs11597367 (highlighted blue). rs11597367 is an A/G SNP (IUB code R) and the major allele breaks a CpG site (C**A**), but the minor

allele creates a CpG site (C**G**). (B) The methylation status of the three CpG sites flanking rs11597367 was assessed in purified cell populations. (C) Assessment of allele-specific methylation of the three CpG sites flanking rs11597367, the data presented in (B) has been split by the minor allele of rs11597367. (D) Assessment of the methylation profile of the CpG site created by the minor allele of rs11597367 (red circles), whereas the major A allele does not make a methylation site (blue circles), therefore the cytosine prior to rs11597367 is not methylated and the bisulfite conversion reaction converts the cytosine to uracil which is read as a thymine, and therefore 100% demethylated. (E) Example output of the methylation analysis of the three CpG sites (site numbers 1, 3 and 4) as well as the CpG site created by the minor allele of rs11597367 (site number 2), measured in central memory T cells and NKCD56^{dim} cells.

5.4.3 High throughput allele-specific expression with the C haplotype

Epigenetic evidence suggested that the DNA region around rs11597367 was active in NK cells and should be followed up in expression analysis. Additionally, I will survey a panel of purified immune populations, as the effect of the C haplotype might be in several different populations, like I have shown for the A haplotype. As a CD4$^+$ T cell activation eQTL has been described for *IL2RA* with SNPs in moderate LD with the C haplotype[174,179], I will also activate each cell population. I will use allele-specific expression to measure the ratio of *IL2RA* mRNA made from donors heterozygous for the C haplotype with a susceptible haplotype that carries none of the A, C, D, E or F haplotype SNPs.

From all of the cell populations assessed for ASE *ex vivo*, none showed a consistent difference in the level of *IL2RA* mRNA made from the C haplotype as compared to the susceptible haplotype (Fig. 5.23). This includes NK cells, which I thought would show a

genotype dependent difference in *IL2RA* based on the C haplotype, given the weight of epigenetic evidence. However, with activation, both naive and central memory $CD4^+$ T cells displayed reduced levels of *IL2RA* mRNA made by the C haplotype, when compared to the susceptible allele (Fig. 5.23). This small but consistent decrease agrees with the previous published eQTL studies which identified a SNP in LD with the C haplotype making 1.16 fold less *IL2RA* mRNA[179].

The presence of recombination within the C haplotype is extremely rare within the European population, however within the Cambridge BioResouce one donor was identified that carried only the minor allele at rs11597367, but the major alleles at the other two C SNPs. Analysis of this rare recombinant haplotype by ASE showed the same phenotype as the full C haplotype, with reduced levels of *IL2RA* mRNA in activated naive and central memory $CD4^+$ T cells (Fig. 5.23, green circles). Therefore, this provides additional evidence that the SNP rs11597367 is the most likely causal variant for the C haplotype.

The ASE data generated from B cells shows that *ex vivo* there are large differences in the amount of *IL2RA* made, but not in a consistent direction. This is due to the presence of SNPs that alter the expression of *IL2RA* in B cells (Fig. 5.23E, red circles) that are not associated with disease, and will be discussed further in Chapter 6, Non-disease associated variants.

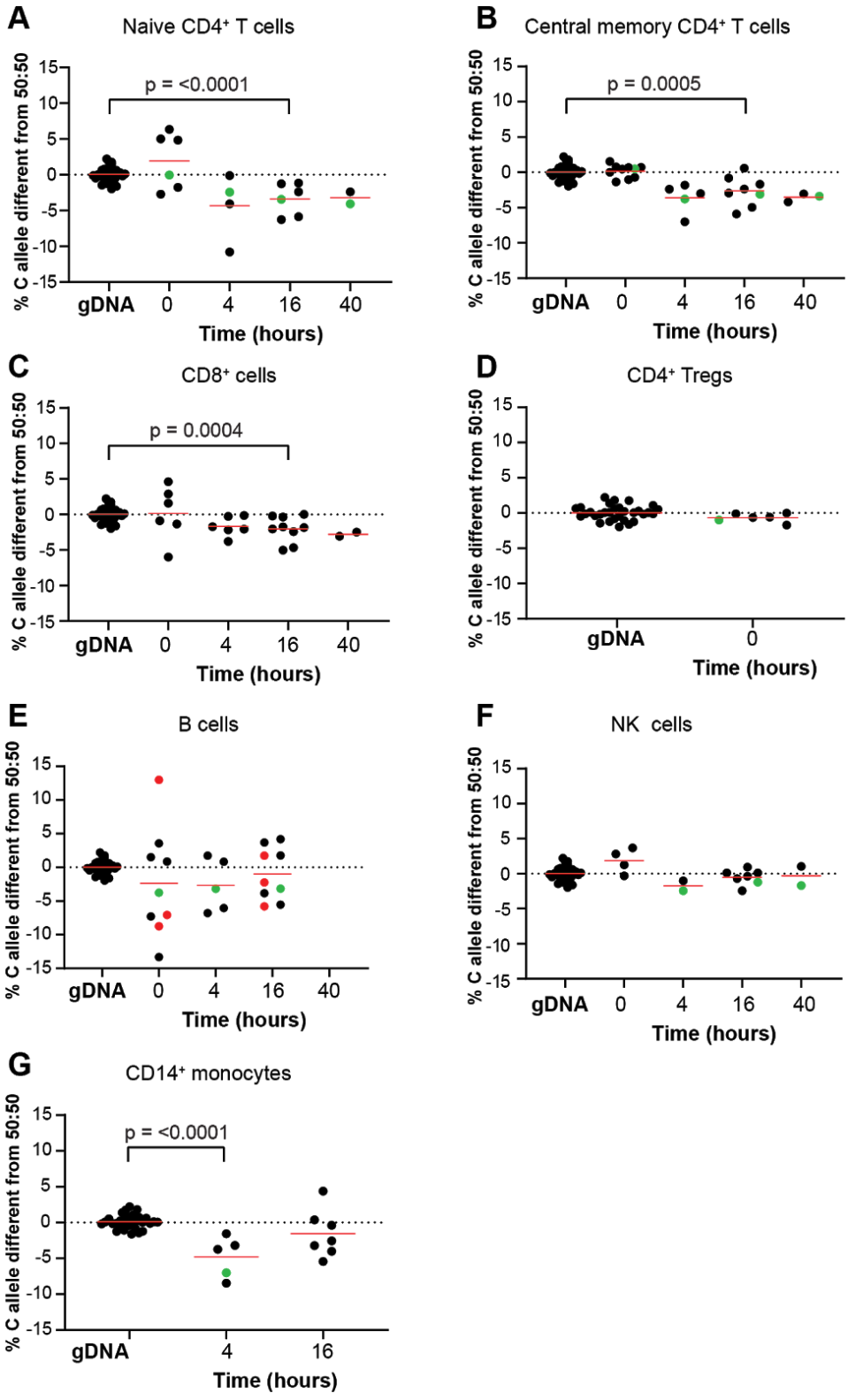


Figure 5.23. The C haplotype makes reduced levels of *IL2RA* mRNA in CD4$^+$ naive and memory T cells with cell activation. Various immune cell populations were purified by cell sorting using the gating strategy in Figure 5.9 and were tested for ASE *ex vivo* or after

stimulation (4 and 16 hours). T cells were stimulated with CD3/CD28 activation beads, B cells with CpG, NK cells with IL-12, IL-15 and IL-18 and $CD14^+$ monocytes with GMCSF and IFNγ. Donors were heterozygous for the C haplotype and the susceptible haplotype (black circles) and a SNP in the 3′ UTR of *IL2RA* was used to measure the degree of allelic imbalance that is reported as the percent made by the C haplotype. One donor with a rare haplotype where only the minor allele of rs11597367 was present (and not the minor alleles at the other two C haplotype SNPs) and heterozygous for the susceptible haplotype was tested (green circles) for ASE. A genotype-phenotype correlation was identified in B cells that is not associated with disease (red circles) and will be discussed in Chapter 6. Red line displays the mean of the dataset, and P values calculated using a Mann-Whitney test.

As mentioned in the introduction, about half of the C haplotypes carry an allele that influences the level of soluble CD25, which will be discussed in Chapter 6. However, this variant does not alter the expression of *IL2RA* on T cells, therefore to increase the number of samples analysed for the C haplotype, I am presenting the ASE data from donors heterozygous for the C haplotype which carry the minor allele of rs12422497 (which increases the level of soluble CD25 in the blood) with the susceptible haplotype. This data shows the same decrease in the production of *IL2RA* mRNA from the C haplotype with activation from both $CD4^+$ naive and central memory T cells (Fig. 5.24), whereas there is no difference in allelic expression without activation.

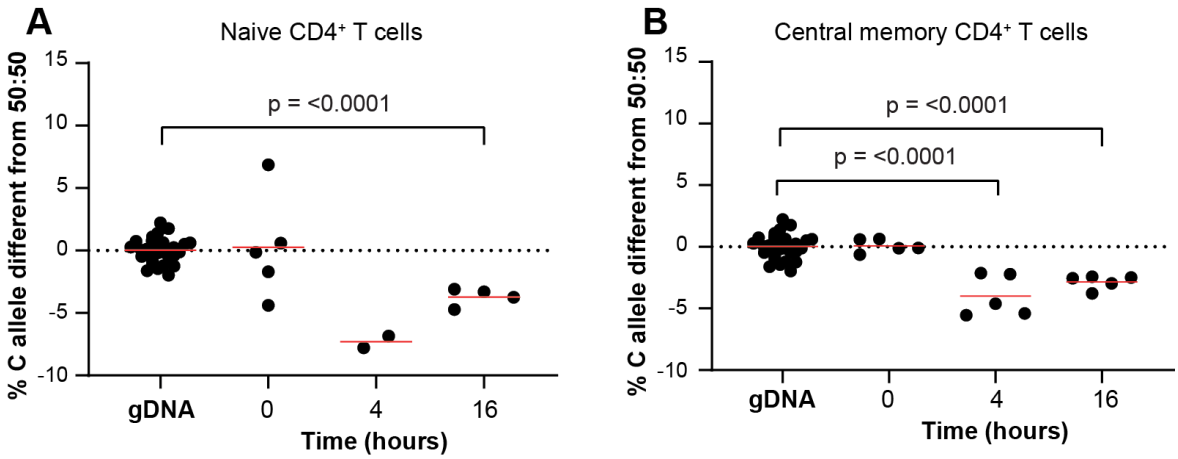


Figure 5.24. The C haplotype makes reduced levels of *IL2RA* mRNA in CD4$^+$ naive and memory T cells with cell activation. CD4$^+$ naive and central memory T cells were purified by cell sorting using the gating strategy in Fig. 5.9 and were tested for ASE *ex vivo* or after stimulation (4 and 16 hours) with CD3/CD28 activation beads. Donors were heterozygous for the C haplotype containing the minor allele of rs12422497 (which increases the amount of soluble CD25 present in the blood) and the susceptible haplotype and a SNP in the 3′ UTR of *IL2RA* was used to measure the degree of allelic imbalance that is reported as the percent made by the C haplotype. Red line displays the mean of the dataset, and P values calculated using a Mann-Whitney test.

5.4.4 Immunophenotyping the C haplotype for CD25 expression

In the DIL PBMC collection, 522 individuals were immunophenotyped and CD25 expression assessed in various immune cells types from fresh blood, and subsequently analysed for any genetic correlations. No association of CD25 expression was seen for the C haplotype (Fig. 5.12), which was reassuring, as the only ASE I observed was with an activation phenotype in which the C haplotype made less *IL2RA* mRNA in activated CD4$^+$ T cells, and the immunophenotyping did not include an activation condition.

5.4.5 Summary of the C haplotype

A phenotype has not previously been described for the C haplotype, but an activation eQTL in $CD4^+$ T cells was identified with SNPs that were in high LD[174,179], as well as epigenetic evidence pointing to NK cells. Therefore, a comprehensive survey of purified immune cells using ASE was performed in donors heterozygous for the C haplotype and I confirmed the activation phenotype showing that the C haplotype makes less *IL2RA* in $CD4^+$ naive and central memory T cells. No consistent phenotype was seen in NK cells with the C haplotype, although there was strong epigenetic evidence, which could be interpreted as the C SNP was located in an enhancer for NK cells but did not disrupt it, or the conditions tested did not reveal the effect of the C haplotype. Due to limited sample availability I tested bulk NK cells rather than the $CD56^{bright}$ population which represents 10% of the NK population and expresses the highest levels of CD25, and if the effect was on NK $CD56^{bright}$ cells it might well be obscured.

The T cell activation phenotype originally identified was not assigned to the C haplotype, as the C haplotype was not identified until the stochastic search method was applied to the GWAS data. This problem was particularly acute for AITD, where the lead SNP has now been identified as joint-tagging the A and C haplotypes[104], a finding that will be discussed in the next section.

5.5 Evidence that the AITD association is two separate signals

The GWAS for AITD indicates that the SNP rs706779 is the best single SNP to explain the genetic association[173]. However, analysis of the same data using a Bayesian stochastic search approach (GUESSFM) indicates that the association is best explained with two independent associations, which are the A and C *IL2RA* SNP groups[104]. In Figure 5.2, I presented a haplotype analysis using all the identified *IL2RA* SNP groups associated with T1D, MS and AITD. To highlight how important the haplotype analysis was to explain the genetic association of *IL2RA* to AITD, I have included a haplotype analysis that includes one SNP from the *IL2RA* A group (rs61839660), C group (rs11597367) and the best SNP identified by stepwise regression (rs706779). This analysis shows that only the A and C haplotypes are protective for AITD, to a similar degree and direction. Whereas when rs706779 is assessed without the presence of the A or C haplotypes, it offers no protection from AITD (Fig. 5.25).

The fact that rs706779 sits primarily on the A and C haplotypes and these are equally protective from disease, explains why rs706779 would be picked as the single most protective SNP.

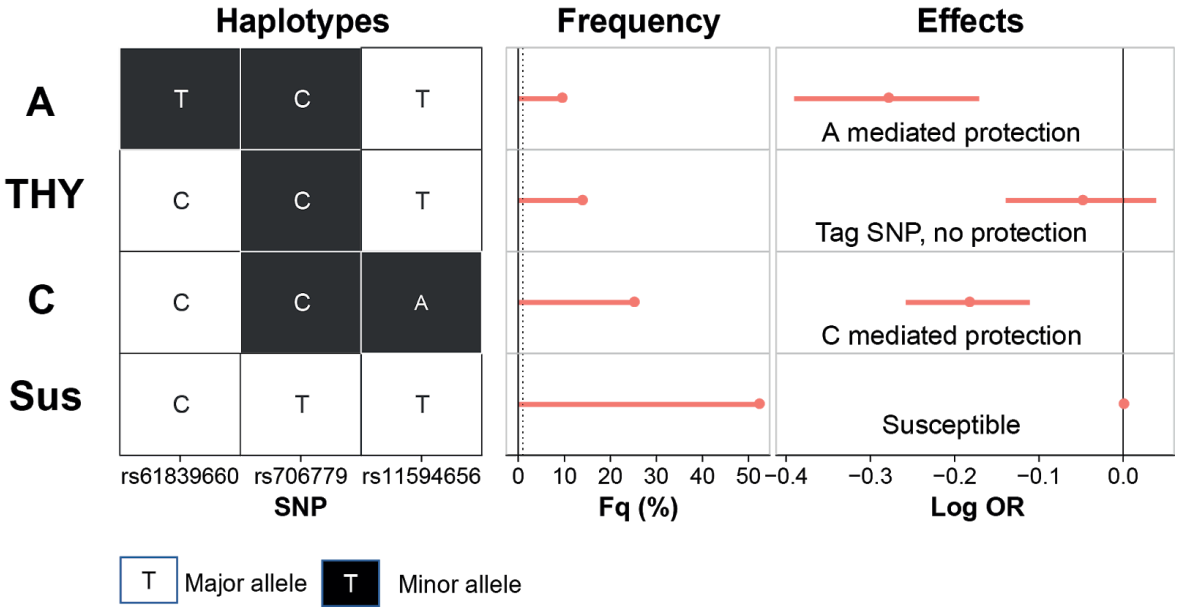


Figure 5.25. This is Figure 1A taken from "Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases", for which I was an author and is published in Nature Communications[104]. The legend reads: Fig. 1 Evidence for joint tagging. a Haplotype analysis of SNPs selected by stepwise search and GUESSFM for ATD in region 10p-6030000-6220000. A representative SNP from each SNP group is shown. Each column represents one SNP, with possible alleles colour coded according to major or minor. Each row is a haplotype—a specific combination of alleles across all SNPs (A, C or THY which is defined as rs706779) — with frequency in UK controls and effect on disease risk (log OR +95% CI). There are four common haplotypes. Three carry the minor allele at rs706799, but only those that also carry minor allele at A or C show a significant effect on disease risk.

Another way to show how the A and C haplotypes better explain the AITD association is to look at the distribution of the alleles at all variants across the *IL2RA* region. In Figure 5.3, I showed that each of the *IL2RA* A, C, D, E and F haplotypes contained a region which contained the haplotype defining SNPs, and all of the other alleles carried on that haplotype were similar, and there was very little variation within donors carrying the same

haplotype, suggesting they had arisen either through a bottleneck event or selection. This is true for the A and C haplotype, however, donors identified as carrying the minor allele at rs706779 (regardless of status at the A and C SNPs), display a very different distribution of alleles, in which most are segregating (Fig. 5.26).

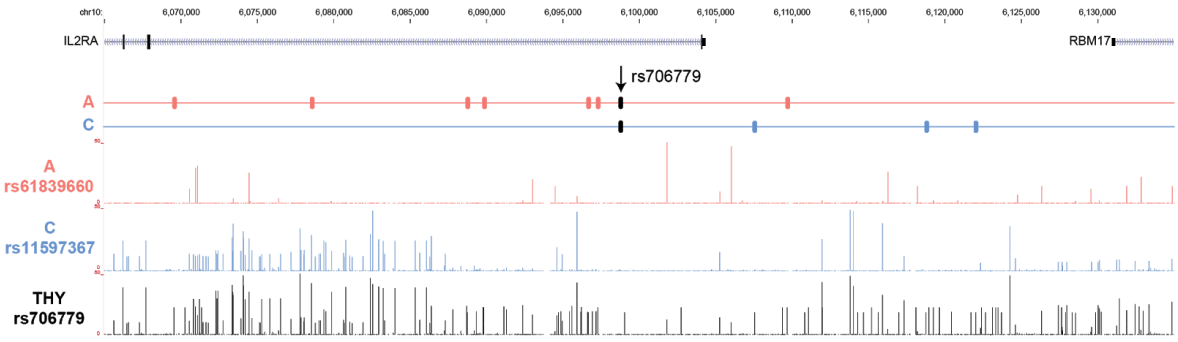


Figure 5.26. Assessment of similarity within the *IL2RA* A, C and THY haplotypes from the European population within the 1000 Genomes Project. The schematic shows the location of the group A and group C credible SNP groups. The black arrow shows the location of rs706779, which was identified as being the lead SNP for AITD and is located on both the A and C haplotypes. Each haplotype is displayed as a separate row and the height of each point is related to the proportion of individuals having the same alleles for the SNP (score of zero) through to having equal counts of the two SNP alleles (score of 50).

Allele-specific expression measures the ratio of mRNA made from the maternal and paternal chromosome, and I have shown robust phenotypes for the A and C haplotypes, when the A or C haplotype was heterozygous with either the E or susceptible haplotype. To provide biological proof that the AITD association is best explained by a two signal model, which is the A and C haplotype and not rs706779 that is the single best AITD association, I have performed ASE on donors heterozygous for the A and C haplotype. The rationale is that as rs706779 is found on both the A and C haplotype, and donors

heterozygous for both the A and C haplotype will be homozygous for rs706779 and therefore will not display the phenotypes described for the A and C haplotypes (Fig. 5.27A).

As previously shown in $CD4^+$ central memory T cells, the A haplotype produces more *IL2RA* mRNA than the susceptible haplotype, whereas there is no difference in the ratio of *IL2RA* made from either the C or susceptible haplotype. When donors heterozygous for the A and C haplotype were analysed for ASE, the same increased production of *IL2RA* from the A haplotype was observed. When the cells were activated, both the A and C haplotypes when heterozygous with the susceptible haplotype showed decreased expression of *IL2RA*, and when donors heterozygous for both the A and C haplotypes were analysed, no difference in the ratio of *IL2RA* made was observed (Fig. 5.27B). This data clearly shows that in $CD4^+$ central memory T cells, only the A haplotype is responsible for the increased expression of *IL2RA* and not either the C haplotype or the presence of rs706779. As both the A and C haplotypes display the same decreased *IL2RA* expression with activation, no comment can be made about the effect of rs706779.

From our DIL PBMC bank collection of 522 individuals, the expression of CD25 has been measured on $CD4^+$ central memory T cells. Here I present the CD25 expression from the same genotype groups that I performed ASE on, and show that only donors carrying the A haplotype have increased expression of CD25, and is not dependent on the C haplotype or rs706779, confirming the ASE result (Fig. 5.27C).

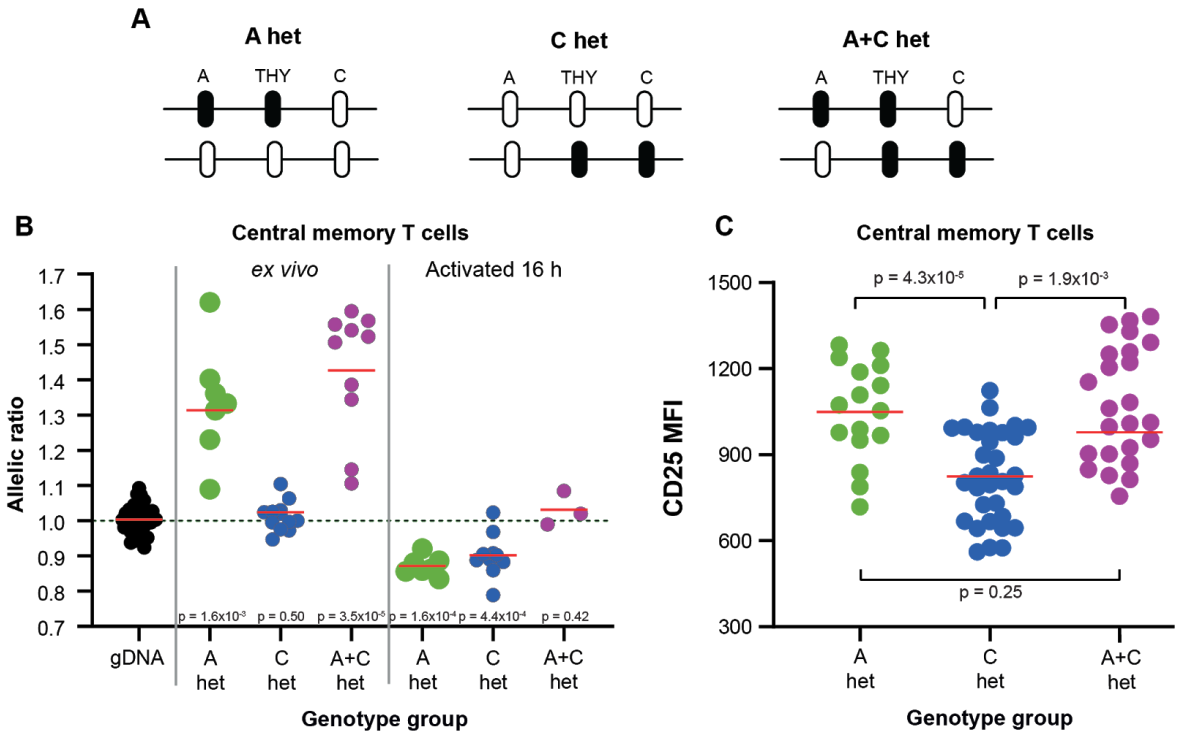


Figure 5.27. The *IL2RA* A and C haplotypes control the expression of *IL2RA* and not rs706779. (A) Cartoon representation of the A, C and THY (rs706779) genotypes of the heterozygous individuals tested. Both the A and C haplotypes carry THY (rs706779), with filled ovals representing the minor protective allele and open ovals the major alleles. (B) The ratio of *IL2RA* made by donors heterozygous for the A and Susceptible, C and Susceptible or A and C haplotypes was measured in $CD4^+$ central memory T cells *ex vivo* or activated for 16 hours with CD3/CD28 activation beads. (C) The CD25 MFI was measured by flow cytometry in donors with the same genotype as that in (B). Red lines depict the mean, and P values were calculated by an unpaired T Test.

5.6 The *IL2RA* group D haplotype

5.6.1 Explaining the genetic association of the *IL2RA* D haplotype

The *IL2RA* D haplotype association appears to be specific to MS[2,104], and the history of how the association was discovered shares a similar path to that of AITD. The results of the MS GWAS, which used stepwise regression for the analysis, reported that the lead *IL2RA* SNP was rs2104286[REF170–172]. When the data was analysed using a stochastic search approach the association was resolved into two separate signals[2,104], which were the *IL2RA* A and D haplotypes.

A haplotype analysis was key to resolving the AITD association, and again proved critical to understanding the reasons why rs2104286 was picked as the lead SNP by stepwise regression, whereas the Bayesian analysis picked two different variants. In Figure 5.2, I show the relationship between rs2104286 to a SNP from both the A and D sets of credible SNPs (A is rs61839660, D is rs62626317). The majority of individuals carrying the D haplotype also carry the minor allele of rs2104286 and all individuals carrying the A haplotype also carry the minor allele of rs2104286. As the protective effect of the A and D haplotype is both in the same direction and size, it is not surprising that as rs2104286 tags both of these haplotypes it is chosen as the single best SNP by stepwise. There is a rare D haplotype in which the minor allele of rs2104286 is not present, and it has equivalent protection to that of the more common D haplotype that carries rs2104286, again showing it is the D haplotype that is protective and not rs2104286 (Fig. 5.28).

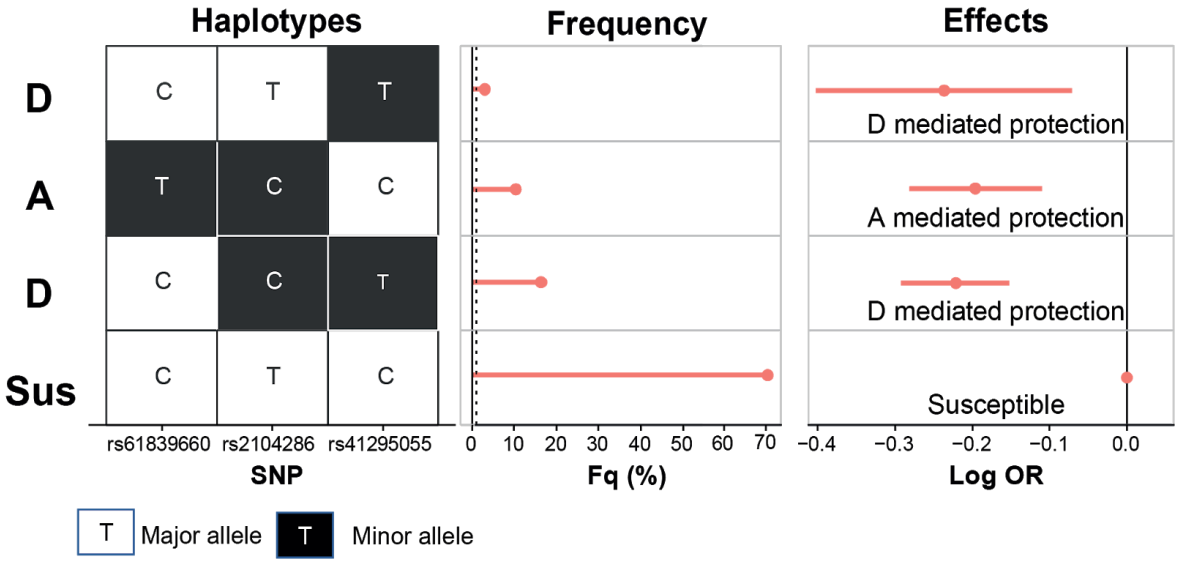


Figure 5.28. This is Figure 6B taken from "Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases", for which I was an author and is published in Nature Communications[104]. The legend reads: Fig. 6B Analysis of chromosome 10p region containing *IL2RA*. (B) Haplotype analysis of SNPs selected by stepwise search and GUESSFM for MS. Each row represents one SNP, with possible alleles colour coded according to major or minor. Each column is a haplotype—a specific combination of alleles across all SNPs—with frequency in UK controls and effect on disease risk (log OR + 95% CI). There are four common haplotypes. Three appear protective, carrying the minor allele at either A or D, but only two carry the minor allele at B (rs2014286).

The haplotype analysis uses only a single SNP from the A and D haplotypes as well as rs2104286. When you look at the distribution of alleles across the whole *IL2RA* region, a very different distribution of alleles is seen, depending on which SNPs you use to define the haplotypes. I have shown in Figure 5.3 that each of the *IL2RA* A, C, D, E and F haplotypes contained a region which contained the haplotype defining SNPs, and all of the other alleles carried on that haplotype were mostly shared, and there was very little variation within donors carrying the same haplotype, suggesting this haplotype had been

fixed in the population either by a migratory bottleneck or positive selection. This can be seen for the A and D haplotypes, however, donors identified as carrying the minor allele at rs2104286 (regardless of status at the A and D SNPs), display a very different distribution of alleles, in which most are segregating (Fig. 5.29). If the haplotype defined by rs2104286 was indeed causal for disease, I would conclude that this variant was ancient and via recombination had become associated with distinct haplotypes, in this case the A and D haplotypes, by chance. Next I will describe the phenotype of the D haplotype, and then I will show the biological evidence to confirm the statistical analysis of the GWAS data that the A and D haplotypes best explain the association with MS.



Figure 5.29. Assessment of similarity of the *IL2RA* A, D and B haplotypes from the European population within the 1000 Genomes Project. The schematic shows the location of the group A and group D credible SNP groups. The green arrow shows the location of rs2104286, which was identified as being the lead SNP for MS and is located on both the A and D haplotypes. Each haplotype is displayed as a separate row and the height of each point is related to the proportion of individuals having the same alleles for the SNP (score of zero) through to having equal counts of the two SNP alleles (score of 50).

5.6.2 The D haplotype makes less *IL2RA* in naive CD4$^+$ T cells

Using the results from the GWAS studies of MS, which identified rs2104286 as the lead SNP for MS, a genotype-phenotype interaction was identified that resulted in donors carrying the minor allele of rs2104286 having a reduced frequency of CD4$^+$ naive T cells that express CD25[REF37]. When the MS association was reanalysed through a Bayesian method, and the A and D haplotypes identified, the genotype-phenotype data from Dendrou *et al.*[37] showed a stronger correlation with the D haplotype than rs2104286[REF2], again providing more evidence that the A and D haplotypes better explain the association than rs2104286.

I have performed the same high-throughput ASE experiments as done for the A and C haplotypes to firstly reproduce the published decreased in frequency of CD4$^+$ naive T cells that express CD25[REF37], and secondly to assess if the D haplotype influences the expression of *IL2RA* mRNA / CD25 in any other immune cell population *ex vivo* and with activation. The results from the ASE experiments clearly show that the D haplotype produces at least 1.5 fold less *IL2RA* mRNA when compared to either the susceptible or E haplotype (Fig. 5.30). The expression of *IL2RA* is approximately four-fold lower in CD4$^+$ naive T cells compared to CD4$^+$ central memory T cells, and this has made the ASE assay more challenging to achieve consistent results, and more variability is seen in the ASE results, although the direction is always consistent with the D haplotype making less *IL2RA*. Unlike the A haplotype, no other ASE was seen in any of the other immune cell populations assessed, and unlike the A and C haplotypes, no difference in expression was seen with cell activation.

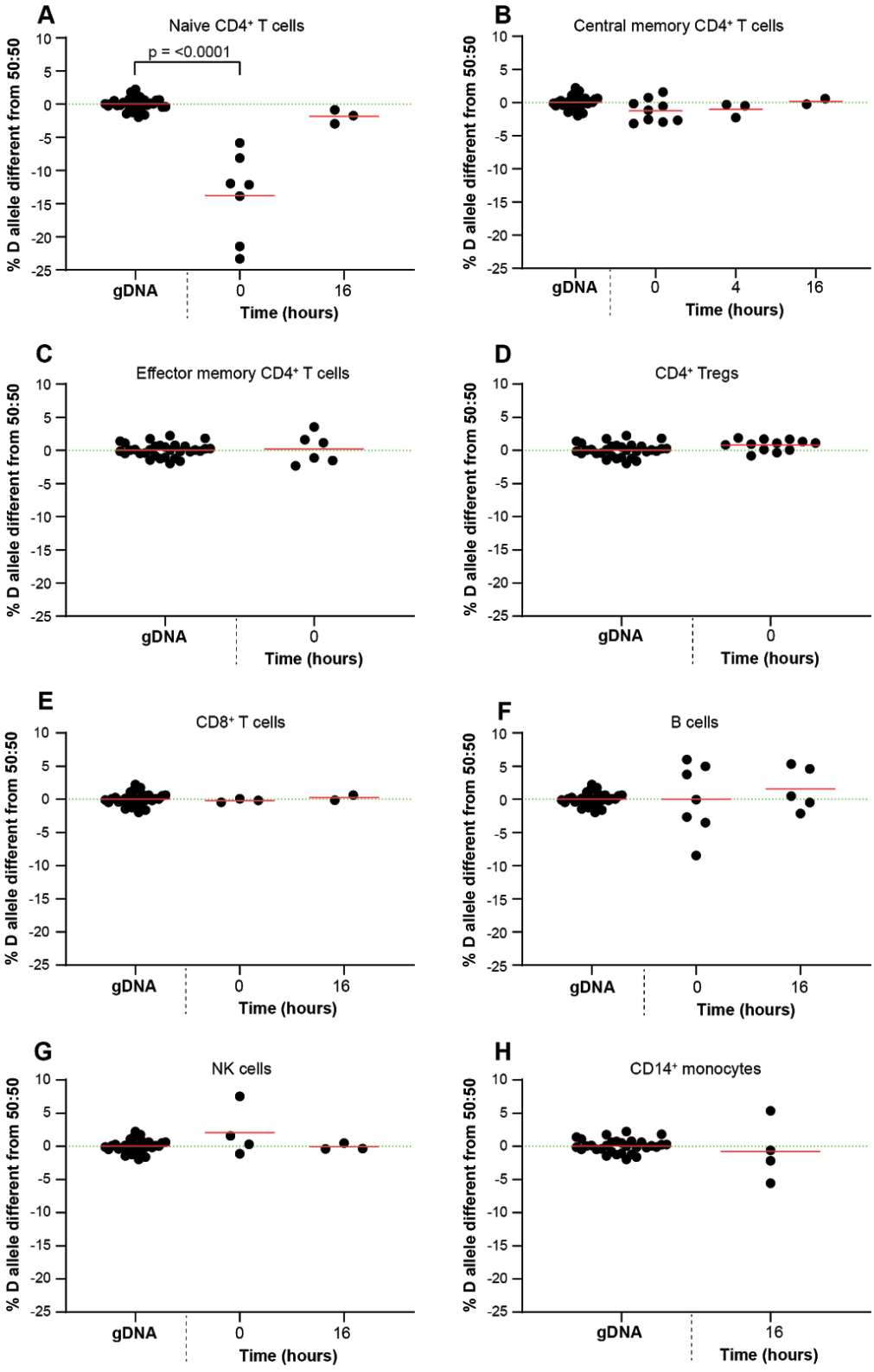


Figure 5.30. The D haplotype causes reduced levels of *IL2RA* mRNA in CD4$^+$ naive T cells. Various immune cell populations were purified by cell sorting using the gating strategy in Figure 5.9 and were tested for ASE *ex vivo* or after cell stimulation (4 and 16

hours). T cells were stimulated with CD3/CD28 activation beads, B cells with CpG, NK cells with IL-12, IL-15 and IL-18 and $CD14^+$ monocytes with GMCSF and IFNɣ. Donors were heterozygous for the D haplotype with either the susceptible or E haplotypes and the SNP rs12244380 in the 3′ UTR of *IL2RA* was used to measure the degree of allelic imbalance that is reported as the percent made by the D haplotype. The gDNA control was included to show each allele was amplified equally efficiently. The red line depicts the mean of the dataset and P values were calculated with a Mann-Whitney test.

5.6.3 The D haplotype correlates with decreased frequency of $CD4^+$ naive T cells that express CD25

The association of the D haplotype with a decreased percentage of $CD4^+$ naive T cells that express CD25 has been published[2,37] and we sought to replicate the finding using samples from the DIL PBMC bank. The CD25 frequency phenotype is challenging to assess for two reasons, the first is that low expression of CD25 on naive T cells requires careful placement of the gate delineating positive from negative cells, and secondly the frequency of naive T cells that express CD25 increases with age. An example of CD25 expression on naive cells is shown in Figure 5.9. We assessed if CD25 MFI, either on the total naive CD4 T cell population or on cells gated as $CD25^+$, provides a stronger association with the D haplotype as compared to the percentage of $CD4^+$ naive T cells that express CD25. The CD25 MFI of $CD25^+$ naive T cells was measured on the first 308 samples collected in the DIL PBMC bank and although a trend was observed in the same direction as the frequency phenotype (lower CD25 MFI on naive CD4 T cells having the D haplotype), the association of the MFI phenotypes with the D haplotype was lower than that of the frequency phenotype (data not shown).

In Figure 5.16A I showed that only the D haplotype showed a decrease frequency of naive T cells that express CD25, and the A, C, E, F and susceptible haplotypes had the same frequency. Therefore all samples were combined, and divided solely on the basis of the genotype of the D haplotype, and classed as D homozygous (two copies), D heterozygous (1 copy) and Not-D (no copies of the D haplotype). This pooling increased the group size of the Not-D and D heterozygous samples and clearly replicates the previous finding that the D haplotype associates with decreased frequency of $CD4^+$ naive T cells that express CD25.

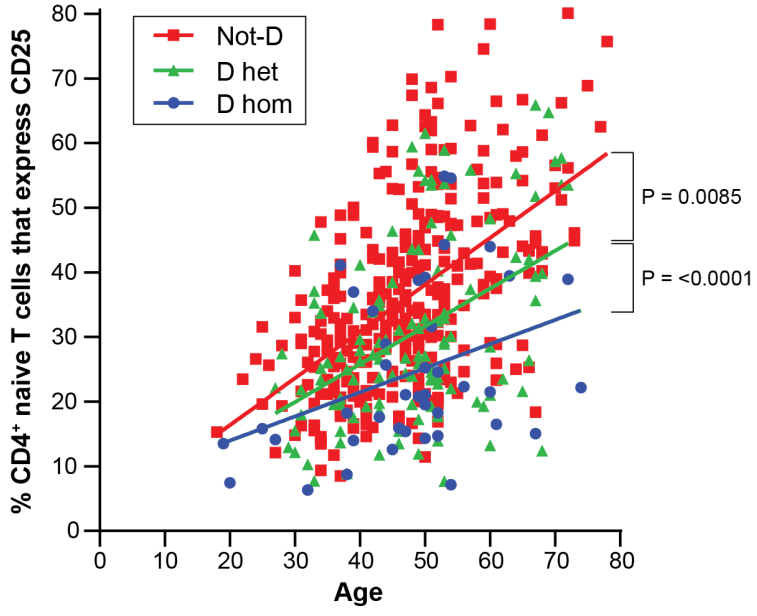


Figure 5.31. The *IL2RA* D haplotype correlates with the decreased frequency of $CD4^+$ naive T cells that express CD25. From the DIL PBMC bank collection of 496 donors, the frequency of $CD4^+$ naive T cells that express CD25 was measured in a blinded fashion and then the data split by genotype into donors homozygous for the D haplotype (D hom, blue), heterozygous for the D haplotype (D het, green) and carrying no copies of the D haplotype (Not-D, red).

5.6.4 Expression phenotypes confirm that the A and D haplotypes best explain the association with MS

I have shown that the A and D haplotypes alter the expression of *IL2RA* in $CD4^+$ central memory and naive T cell, respectively, however due to LD structure in the region both

carry the minor allele of rs2104286. Here I will test whether rs2104286 affects the expression of *IL2RA* in the context of the A and D haplotypes using ASE. Donors heterozygous for the A (A-het) or D (D-Het) haplotypes are also heterozygous for rs2104286. Donors that are heterozygous for the A and D haplotypes (A+D het) are homozygous for rs2104286 and tests the effect of rs2104286 in the context of the A and D haplotypes, thereby providing evidence to either support or contradict our statistical model that the A and D haplotypes and not rs2104286 best explain the association with MS.

In $CD4^+$ central memory T cells, only the A-Het and A+D individuals show an allelic imbalance with the A haplotype making more *IL2RA* mRNA; since A+D individuals are homozygous for rs2104286 and an imbalance is seen, it negates any effect of rs2104286 (Fig. 5.32). Secondly, D-het individuals show no *IL2RA* imbalance but are also heterozygous for rs2104286, thereby providing more evidence that rs2104286 does not alter *IL2RA* expression.

In naive $CD4^+$ T cells, a similar pattern is seen, with the D-het and A+D het individuals showing that the D haplotype makes less *IL2RA* mRNA (Fig. 5.32). This result also contradicts the causality of rs2104286, since only the D-het individuals are heterozygous for rs2104286, whereas the A+D het individuals do show allelic imbalance but are homozygous at rs2104286.

The identification of individuals that carry rare recombination events in the *IL2RA* region have proved extremely useful in identification of the causal variant for the A haplotype,

therefore we looked to see if there were any donors that separated rs2104286 from either the A or D haplotype. We identified four donors, three of which carry rare *IL2RA* haplotypes that are defined as (Fig 5.32A): Donor 1 is a control for donors 2 and 3 and does not contain any A, D or rs2104286 minor alleles, but is heterozygous for E; donor 2 carries the minor allele at rs2104286, but no A or D minor alleles; donor 3 carries the minor alleles at D but not at rs2104286; donor 4 carries the same D haplotype that lacks the minor allele at rs2104286 as donor 3 but is heterozygous with the A haplotype that does carry rs2104286. This combination of rare recombinants isolates the B and D haplotypes and allows the testing of their individual effects. As expected as donor 1 carries no A, B or D alleles, no allelic imbalance in $CD4^+$ naive and central memory T cells was observed. Donor 2 which is is heterozygous for rs2104286, but not does not carry the A or D haplotype, also shows no allelic imbalance in the cell types tested. Donors 3 and 4 which carry the D haplotype that lacks rs2104286 display an allelic imbalance (Fig. 5.32C), like that seen for the common D haplotype that carries rs2104286 (Fig. 5.32). The use of donors carrying relatively rare recombination events has again proved useful in showing that the allelic imbalance seen in the A and D haplotypes is not a consequence of rs2104286, but rather the effects of the A and D haplotypes.

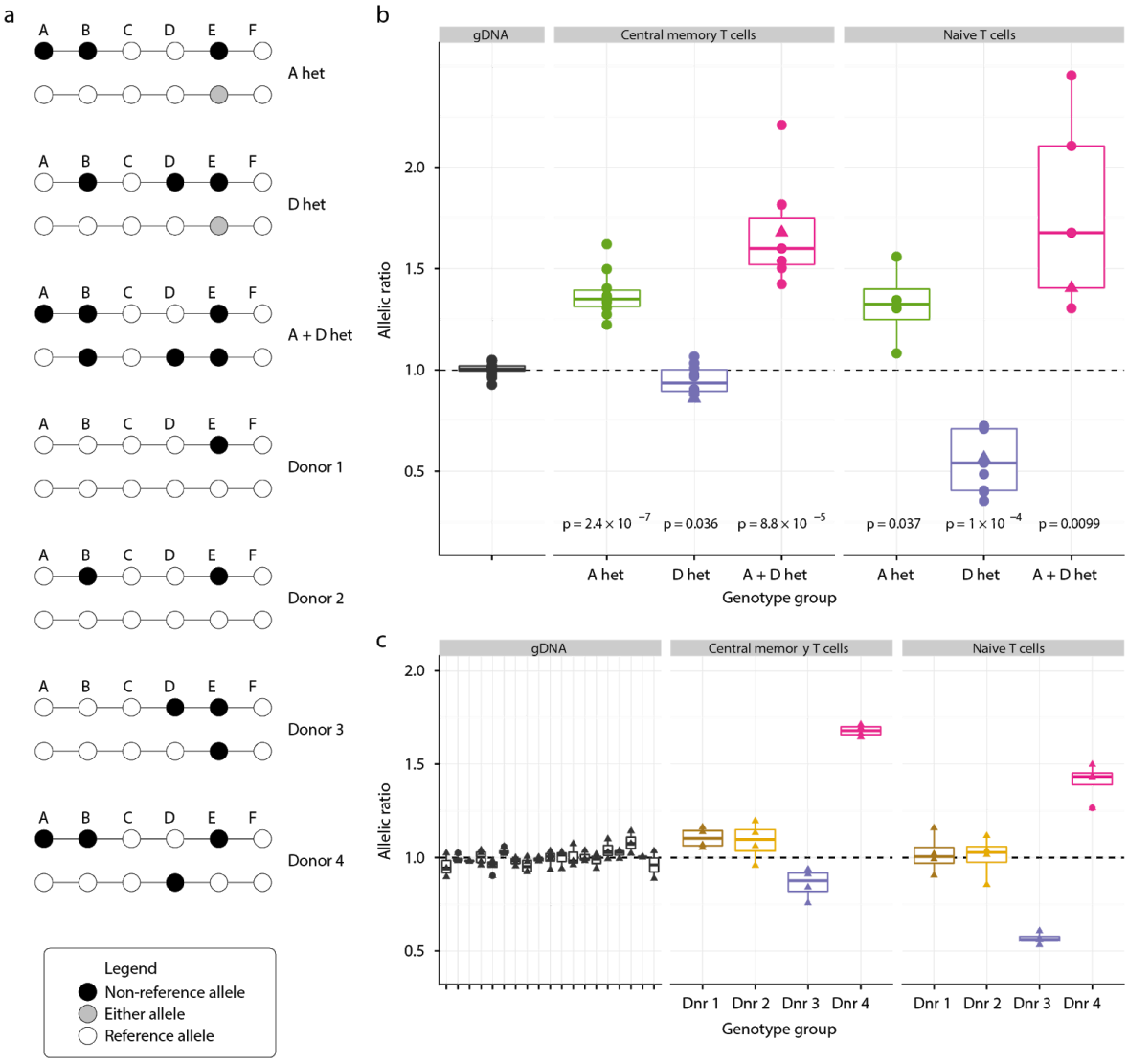


Figure 5.32. This is Figure 7 taken from "Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases", for which I was an author and is published in Nature Communications[104]. The legend reads: Fig. 7 Allele-specific expression analysis of *IL2RA*. This shows that there are two phenotypes that map to the A and D SNP groups and not the B group, providing functional evidence that the stochastic search better explains the genetic association than stepwise. a Schematic of donor *IL2RA* genotypes used in allele-specific expression studies. As the minor alleles for both A and D each usually co-occur with the minor B allele, in A-het and D-het individuals, the B SNP is heterozygous but in A + D-het individuals, the B SNP is homozygous. There are rare exceptions as seen in donors 3 and 4. b Allele-specific expression of *IL2RA* in CD4+ central memory T cells and CD4+ naive T cells in A-het, D-het and A + D het donors. The allelic ratios (top:bottom haplotypes shown in panel a) are calculated from 3–4 replicates per individual. P-values were

calculated by t-test of allelic ratio in each group compared to gDNA. Donors that contributed to panel c are indicated by a triangle. c Donors with rare *IL2RA* haplotypes confirm that the B SNP does not explain genotype-phenotype expression of *IL2RA*. As there is only one donor per genotype we cannot perform statistical testing. We show each of the four replicate allelic ratios (ratio order as in panel b) to indicate the variability of the assay. The bounds of each box correspond to the first and third quartiles, and the whiskers extend to the smallest/largest value at most 1.5*IQR from the bounds; IQR is the distance between the first and third quartiles. Data beyond the whisker ends are plotted individually.

Recently a very large eQTL dataset has been published which used RNA-Seq data from over 30,000 individuals[43], and I wondered if the effect of the *IL2RA* A and D haplotypes would be seen even though these data are not from purified cell types, but from blood which contains many cell types that express *IL2RA* mRNA. Table 5.3 shows that indeed the effects of the *IL2RA* A and D haplotype are seen, and reassuringly the eQTLs are in opposite directions with the minor alleles of the A and D haplotypes correlating with increased and decreased *IL2RA* expression, respectively. In addition, no eQTL was seen for rs2104286, which confirms all the results I have presented that it is the A and D haplotypes that influence *IL2RA* expression and not rs2104286.

Table 5.3. *IL2RA* eQTL in whole blood correlates with the A and D haplotypes and not rs2104286. A large eQTL analysis has been performed using RNA-Seq data from whole blood in over 30,000 individuals[43] and the information is available at http://eqtlgen.org/ and SNPs from the *IL2RA* A and D haplotypes as well as rs2104286 were investigated. Only SNPs from the A and D haplotypes showed an *IL2RA* eQTL. This database has also been mined for the C and E haplotypes and no eQTLs were found. The F haplotype is too rare to be included in the eQTL analysis.

Haplotype	SNP	Gene	Allele assessed	Major allele	Z score	P value	Direction
A	rs61839660	*IL2RA*	T	C	15.73	8.81×10^{-56}	A makes more *IL2RA*
D	rs62626317	*IL2RA*	G	A	-10.12	4.46×10^{-24}	D makes less *IL2RA*
B	rs2104286	*IL2RA*	C	T	-	-	No effect - FDR > 0.05

5.6.5 Which of the *IL2RA* group D SNPs is most likely to be causal?

The current statistical analysis has identified that the group D set of credible SNPs contains 7 candidates[2]. The method used to identify the most likely causal variant of the group A set of SNPs using rare recombinants is not possible for the group D as interestingly there are none seen in the European population. Analysis of epigenetic marks surrounding each of the seven group D SNPs showed that rs62626317 and rs41295055, which are only 697 bp apart, were in close proximity to regions of demethylation in $CD4^+$

naive T cells as well as increased presence of histone H3K4me1 that indicates an active enhancer (Fig. 5.33). Additionally, a recent publication that used CRISPR to tile across the *IL2RA* region identified 6 control elements that when modified altered the expression of CD25[REF 196], and one of the elements called CaRE4 (1441 bp) contains the same two D SNPs, rs62626317 and rs41295055 suggesting that either one or both of these SNPs are important for the regulation of CD25 expression (Fig. 5.33).

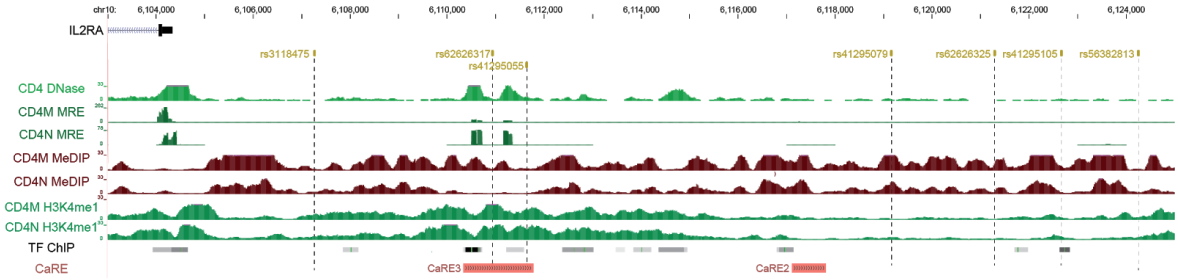


Figure 5.33. Epigenetic analysis of the *IL2RA* region spanning the seven group D SNPs. From the ENCODE and Roadmap Epigenetic data hosted by the UCSC browser, tracks displaying epigenetic data from either total $CD4^+$ T cells, $CD4^+$ memory or naive T cells were selected. The seven group D SNPs are displayed as vertical dashed lines. From the ENCODE Project, the transcription factor ChIP-seq (TF ChIP) is a merge of data from 161 transcription factors in 91 cell types and a grey box encloses each peak cluster of transcription factor occupancy, with the darkness of the box being proportional to the maximum signal strength observed. From the Roadmap Epigenomics Project, DNase I hypersensitivity (DNase) shows regions of open chromatin, MRE-seq shows regions of demethylated DNA, whereas MeDIP-seq shows regions of methylated DNA. Histone H3K4me1 peaks are suggestive of enhancer activity. CRISPRa-responsive elements (CaRE) were identified in Jurkat cells transduced with guide RNAs that tiled across the *IL2RA* region and measured guide RNA enrichment in *IL2RA* expressing cells[174]. A custom track was uploaded to the UCSC browser to display the position of the group D SNPs. Genome coordinates are based upon the GRCh37/hg19 build.

Close examination of the DNA sequence surrounding rs62626317 and rs41295055 showed they were close to sites of DNA methylation, with rs62626317 being only five bases away from a CpG site and rs41295055 is a C/T SNP with the next base being a G, therefore the major allele is a CpG site whereas the minor allele breaks the CpG site. I investigated the methylation status of these two CpG sites as well as other CpG site in close proximity (Fig. 5.34A), from purified cell populations to see if either SNP had increased demethylation in $CD4^+$ naive T cells, thereby providing evidence that one of these SNPs is more likely to be causal.

A similar pattern of demethylation was seen in the regions left and right of rs62626317 that contained eight and seven CpG sites respectively, in which $CD4^+$ naive T cells and Tregs as well as B cells were extremely demethylated at all CpG sites (Fig. 5.34B). The CpG site closest to rs62626317 showed the same high level of demethylation in $CD4^+$ naive T cells and Tregs, but B cells showed a lower level of demethylation and similar to all other cell types tested. The two CpG sites assessed in close proximity to rs41295055 displayed a lower level of demethylation in $CD4^+$ naive T cells when compared to those close to rs62626317, although I have not yet tested the methylation status of the CpG site created by the major allele of rs41295055. From these results I would conclude that rs62626317 is the more likely of the D haplotype SNPs to be causal for the D haplotype.

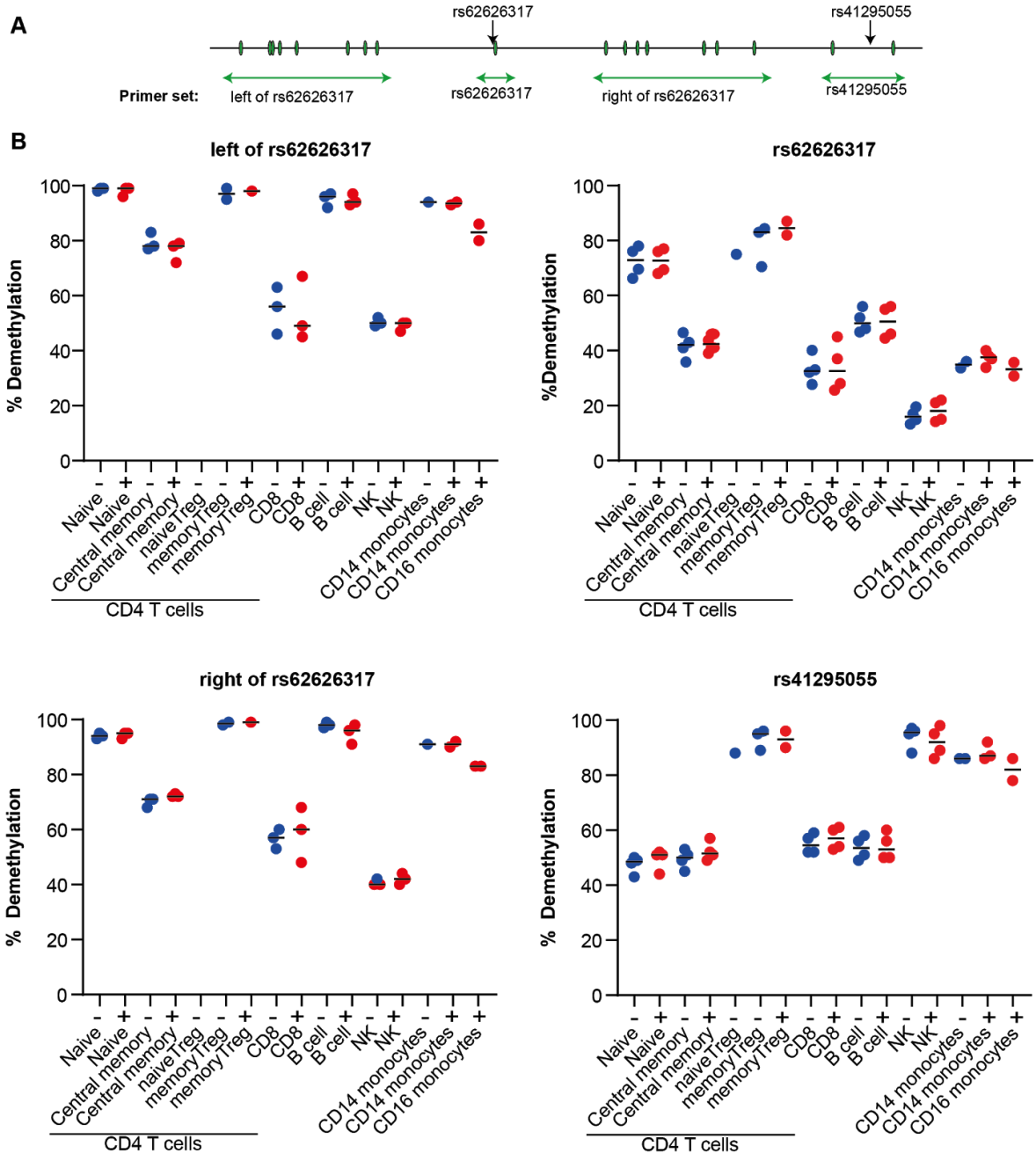


Figure 5.34. Assessment of methylation at sites flanking the group D SNPs rs6262317 and rs41295055 in purified cell populations. (A) The location of the 18 CpG sites (green ovals) that methylation was assessed together with the *IL2RA* group D SNPs rs62626317 and rs41295055. The green arrows display the location of the four primer sets needed for bisulfite sequencing. (B) For each primer set the percent demethylation was calculated as an average over the number of CpG sites within the primer set from purified immune cell populations, with (+) and without (-) 16 hour stimulation.

5.6.6 Summary of the D haplotype

The protection of the D haplotype appears to be specific to MS and has been shown to correlate with decreased frequency of $CD4^+$ naive T cells expressing $CD25^{REF37}$. Using ASE I have shown that the D haplotype makes less *IL2RA* mRNA in $CD4^+$ naive T cells providing a mechanism for the reduced protein expression. There are seven SNPs defining the D haplotype, and epigenetic analysis suggests that either rs62626317 or rs41295055 are the most likely causal variants.

Having a defined phenotype for the A and D haplotypes has allowed for biological proof that the statistical method used to identify the D haplotype was correct. It has been proposed that rs2104286 (termed B) is most associated SNP in the region for MS, but we propose that it is jointly tagging the A and D haplotypes. Using donors heterozygous for the A and D haplotypes we show that their identified phenotypes do not segregate with the minor allele at rs2104286 that is present on both the A and D haplotypes.

5.7 The *IL2RA* group E haplotype

5.7.1 Introduction

The association which has been termed ‘E’ differs from all of the other *IL2RA* haplotypes identified as being protective from T1D, MS or AITD in that only a single variant, rs6602437 has been associated, whereas the A, C, D and F haplotypes range from groups of three to 13 SNPs[2,104]. As only a single variant has been identified for the E signal, it implies this is the causal variant. This protective allele is found on the vast majority of individuals carrying the A and D haplotypes as well as being found on its own haplotype, although as shown in Fig. 5.3 the distribution of all the other alleles on this haplotype again differs from the A, C, D and F haplotypes in that there is no region where donors show a homogenous pattern of alleles. However, closer analysis reveals there are at least seven independent haplotypes that carry only the protective allele at rs6602437, and these each have their own distribution of alleles data not shown). As the E SNP is found on so many different haplotypes it suggests this variant is both ancient and located in a recombination hotspot, that has allowed its movement between haplotypes. The age of human genetic variants have been estimated[197] (https://human.genome.dating) and predicts the E SNP arose 49,690 generations ago confirming that it is the oldest of the *IL2RA* associated variants. In addition, the E SNP is located within a LINE repeat (L1 family), which are associated with increased recombination rate[198]. Finally, the protective T allele of rs6602437 is the more common allele, with a frequency of 54% within the European population, with a small effect size. Chapter 7 deveals into greater detail about the

haplotype similarity of the haplotypes that carry E (A, D and E) to those that do not (C, F and the susceptible haplotypes).

The E SNP is located 25 Kb from the start of *IL2RA*, but is only 1 Kb from *RBM17*, which is a splicing factor that is ubiquitously expressed[44] and to higher levels than *IL2RA* in immune cells (Fig. 5.10), with only a small increase with activation. Given the proximity of E to *RBM17* and complex regulation seen for *IL2RA*, I will assess whether either of these genes is influenced by the E SNP.

5.7.2 The E SNP breaks a methylation site

Analysis of the DNA sequence surrounding the E SNP, rs6602437, reveals that susceptibility allele C makes a CpG site (**C**G), whereas the major protective allele T breaks a CpG site (**T**G). There are no other potential sites of DNA methylation for at least 140 bp either side of the E SNP. Therefore I decided to assess the methylation status of the E SNP in heterozygous donors to see if there is any difference in methylation by cell type. As the donors are heterozygous, the major T allele would be sequenced as a T through bisulfite sequencing, whereas the alternate C allele would be sequenced as a C if methylated or as a T if demethylated. In Figure 5.35, I show that the level of DNA demethylation at the E SNP is approximately 50% in all cell types assessed, which means the C allele is fully methylated as the major T allele can only be sequenced as T.

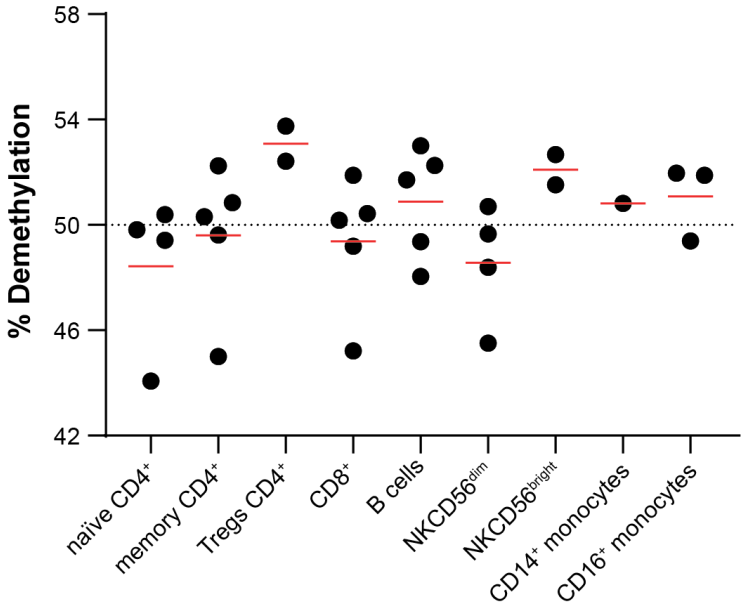


Figure 5.35. Assessment of the DNA methylation status of the E SNP, rs6602437. The major protective T allele breaks a CpG site (**T**G), whereas the minor C allele makes a CpG site (**C**G). Bisulfite sequencing was used to measure DNA methylation in donors heterozygous for the E and Susceptible haplotypes.

5.7.3 Does the E haplotype affect the expression of *IL2RA* or *RBM17*?

Analysis of the region surrounding the E SNP, reveals that it does not reside in a region of epigenetic interest in any immune cell population, although given it is located within a LINE repeat, I wondered if the lack of epigenetic signal was due to the inability to map sequence reads to the region. I have checked the 'mapability' of the region[199], which assesses the level of sequence uniqueness, and reads longer than 35 bp would map solely to this region, therefore the lack of epigenetic signal is valid and not due to an artifact of the alignment process.

As the epigenetic analysis did not provide any clues as to which cell type the E SNP might affect, I have assessed the allelic expression of *IL2RA* and *RBM17* in a panel of immune

cells sort purified using the gating the strategy shown in Fig. 5.9. I am using rs12244380, which is a SNP located in the 3′ UTR of *IL2RA* and rs8463 located in the 3′ UTR of *RBM17* to assess the allelic expression, however both the E and susceptible haplotypes have a very similar allele frequency (65%) and given the LD structure of the region it makes phasing the direction of any allelic imbalance very difficult.

The ASE results show there is no consistent association between the E haplotype and allelic imbalance of *IL2RA* in immune cell populations, with or without cell activation (Fig. 5.36). I have not been able to phase all of the data (green circles Fig. 5.36), and these are reported as the ratio of G to A allele, which occurs most commonly on the E haplotype (65%), although for some samples this is likely to be the wrong direction. If I do not include the samples I cannot phase, then there is still no consistent effect of the *IL2RA* E haplotype with the expression of *IL2RA*.

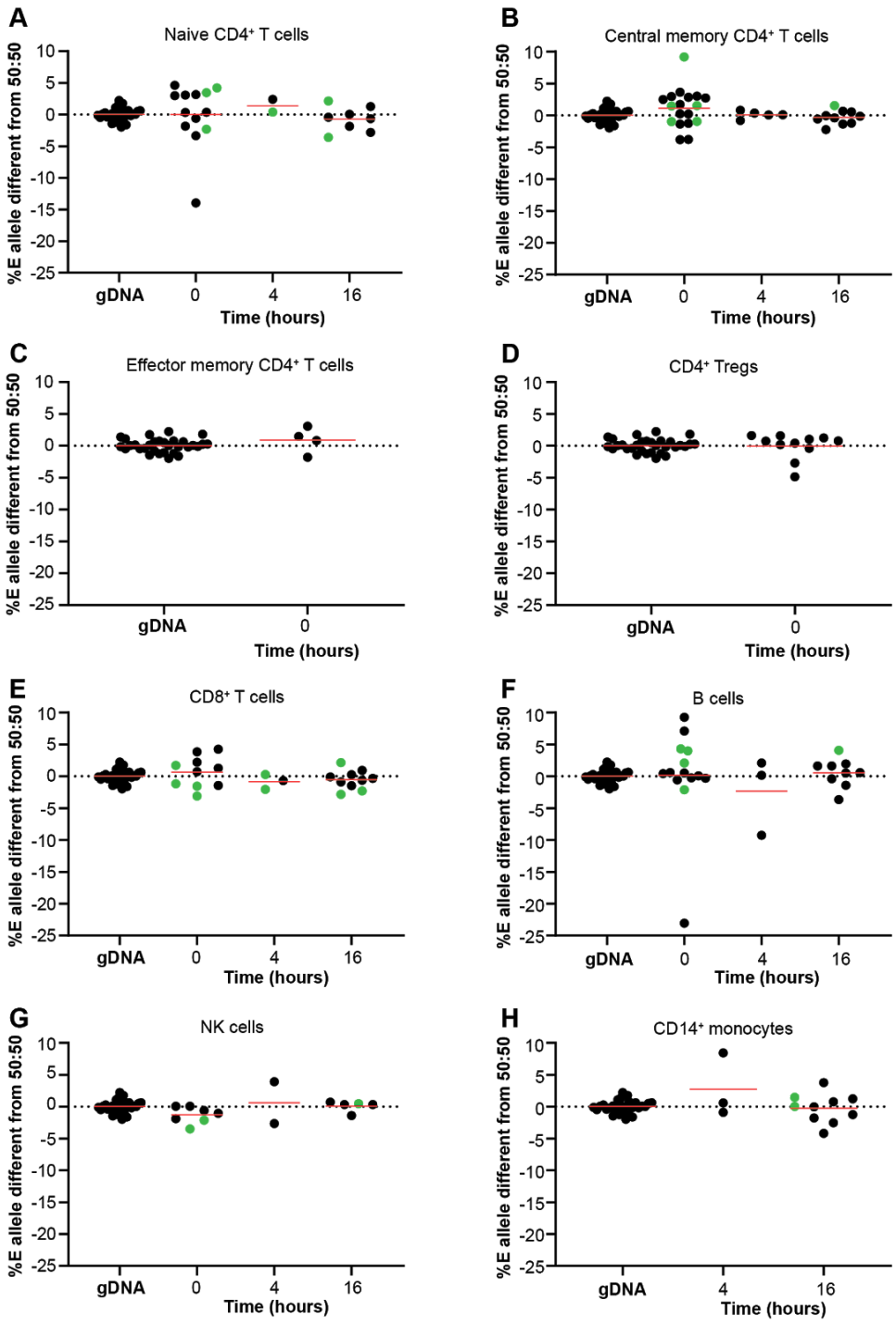


Figure 5.36. The E haplotype do not alter the expression of *IL2RA*. Various immune cell populations were purified by cell sorting using the gating strategy in Fig. 5.9 and were tested for ASE ex vivo or after cell stimulation (4 and 16 hours). T cells were stimulated with CD3/CD28 activation beads, B cells with CpG, NK cells with IL-12, IL-15 and IL-18 and CD14^{+} monocytes with GMCSF and IFNγ. Donors were heterozygous for the E haplotype and the susceptible haplotype (black circles) and SNP rs12244380 in the 3′

UTR of *IL2RA* was used to measure the degree of allelic imbalance that is reported as the percent made by the E haplotype. For some donors it was not possible to phase the E SNP and ASE readout SNP (green circles) and these are phased as the most common direction.

As the E SNP was located 1 Kb from the transcription start site of *RBM17*, this gene was also assessed for allelic imbalance. *RBM17* is well expressed in all immune cells to a similar degree with only a modest increase in expression with activation (Fig. 5.10). The only SNP in either the exons or 3′ UTR that could be used to assess allelic imbalance is rs8463 (MAF 14.7%), however distal of the E SNP, the LD structure breaks down (Fig. 5.3A) and there is more segregation of alleles on each haplotype that makes the phasing of the direction of rs8463 difficult. Additionally, the E SNP is located on and A and D haplotypes as well as being found in isolation, therefore when I performed ASE I have plotted the results for *RMB17* as the difference from 50:50, which removes the difficulty of phasing the data, but does not permit assessment of the direction of effect. The results show that for most cell populations assessed there is a spread in the size of the ASE seen for *RBM17* (Fig. 5.37A), with the mean difference being upto 2% from 50:50, but some individual samples being upto 9% different, with no consistent difference from 50:50 seen that could be attributed to the E haplotype, suggesting there could be other alleles at other SNPs, not in LD with E, that could be affecting the expression of *RBM17*. I have also assessed the allelic imbalance of *RBM17* in donors who are homozygous for either the minor or major allele of E, as a measure of how much variability there is in the assessment of *RBM17* ASE. A similar result is seen with the mean difference from 50:50 being upto 2% (Fig. 5.37B), although there are no donors with an ASE larger than 3%, suggesting the

variants that are causing the larger allelic imbalance seen in donors heterozygous for E (Fig. 5.37A) are rare.

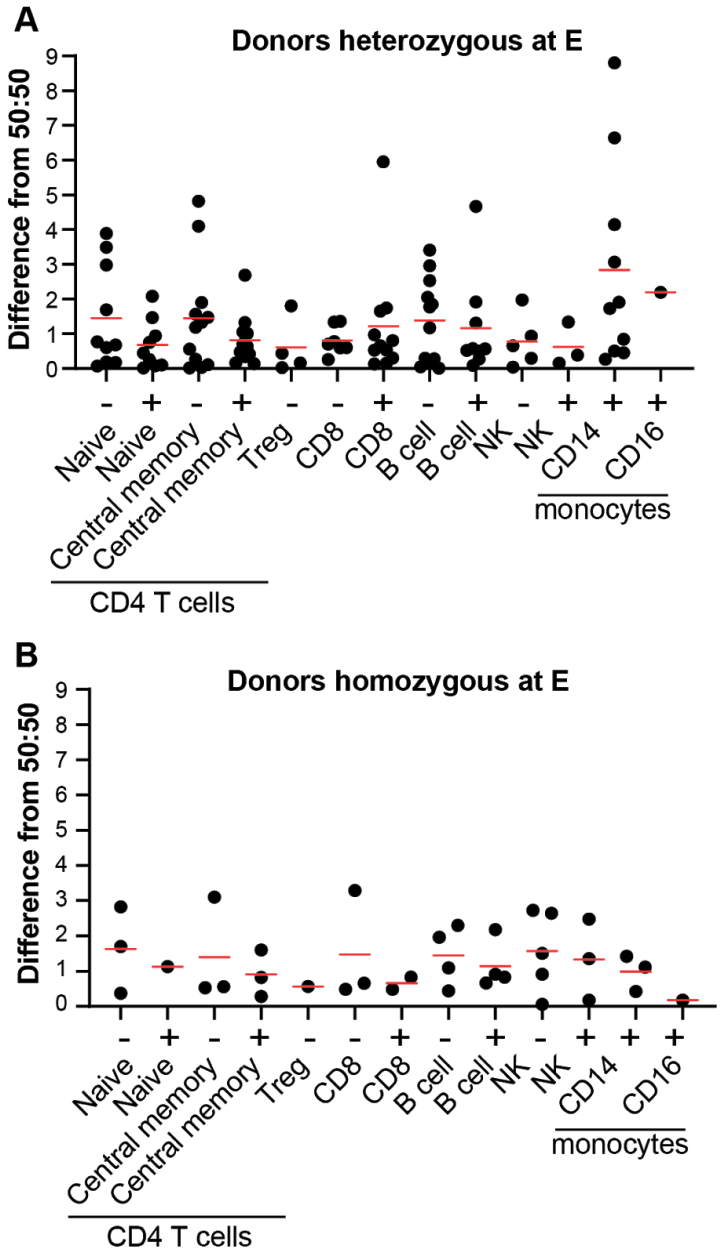


Figure 5.37. The E haplotype does not alter the expression of *RBM17*. Various immune cell populations were purified by cell sorting using the gating strategy in Figure 5.9 from donors heterozygous for E (A) or homozygous for either the minor or major alleles of E (B), and were tested for ASE *ex vivo* or after cell stimulation (4 and 16 hours). T cells were stimulated with CD3/CD28 activation beads, B cells with CpG, NK cells with IL-12, IL-15 and IL-18 and $CD14^+$ monocytes with GMCSF and IFNɣ. (A) Donors heterozygous for E are a mixture of A+E, D+E and E heterozygous with the susceptible haplotype. (B) Donors are homozygous for E and are a mixture of A+E x D+E, A+E x E and D+E x E, whereas donors homozygous for the minor allele of E are C x SUS or SUS x SUS. The SNP rs8463 in the 3′ UTR of *RBM17* is used to calculate allelic imbalance which is reported as difference from 50:50 due to difficulties in phasing.

5.7.4 Summary of the *IL2RA* E haplotype

In summary, there is no clear effect of the E haplotype with either *IL2RA* or *RBM17* mRNA expression in the cell types and activation conditions assessed. The E haplotype has the weakest protection from T1D out of the A, C, E and F haplotypes, implying that its effect likely will be small, and possibly beyond the sensitivity of the ASE method. Additionally, it could be that the effect is in a cell type or activation condition not assessed; as this work was performed on frozen PBMCs from genotype recalled donors, there is a limit to the number of cell populations and conditions that can be assessed. Single-cell transcriptomic analysis of the human pancreas reveals that *RBM17* is expressed in the majority of cells analysed[200,201], therefore the effect of the E SNP could possibly be on the insulin-secreting beta cell. A third explanation is that the effect of the E SNP is through a long range interaction to a gene beyond *IL2RA* and *RBM17*, although analysis of Capture HiC from immune cells shows that the E SNP only interacts with *IL2RA*[192]. Due to the proximity of the E SNP to the promoter of *RBM17*, the Capture HiC technique could not detect if there was an interaction. Lastly, it is possible that the E signal is actually a tag of two or more signals, and the variability seen in the ASE experiments (Fig. 5.36, Fig. 5.37) is due to E joint-tagging different effects; however a larger T1D GWAS that contains almost double the number of cases and controls still identifies E as the same single SNP (unpublished, personal communication, Jamie Inshaw, DIL).

5.8 The *IL2RA* group F haplotype

5.8.1 Introduction

The *IL2RA* F haplotype is the rarest of the haplotypes identified as being protective for T1D, with a frequency of 0.7% within the European population. For a haplotype this rare to be detected the effect size needs to be large (Odds ratio, 0.49[REF17]) , which is the most protective haplotype for T1D at *IL2RA*. There is no evidence of the F haplotype in either the MS or AITD GWAS, however these are smaller collections than for T1D. The UK biobank is a tremendous study of over 500,000 participants with many health details collected, including genetic information (https://www.ukbiobank.ac.uk/), and there are over 14,000 people self-reporting hypothyroidism as well as being prescribed sodium levothyroxine, a medication used to treat hypothyroidism. Analysis of this data identified an association of the F haplotype with sodium levothyroxine medication (odds ratio 0.689, p value 5.4×10^{-11}) as well as hypothyroidism (odds ratio 0.751, p value 3.1×10^{-6}), suggesting the F haplotype does have a protective role in the thyroid disease.

The F haplotype is located solely at the distal end of the *IL2RA* C haplotype in the European population, although this is different across diverse world populations and discussed in Chapter 7.3.5. Due to its rarity and lack of recombination, the F haplotype is defined by a group of 13 SNPs that spans 87 Kb distal of *IL2RA* and includes *RBM17* and *PFKFB3*.

5.8.2 The F haplotype reduces *IL2RA* expression in B cells and monocytes

To assess the effect of the F haplotype I have performed ASE on sort purified immune cells and tested the allelic imbalance of *IL2RA* in donors heterozygous for the C+F haplotype with the C haplotype. As the F haplotype is located at the distal end of the C haplotype, donors who also carry a second C haplotype were used to fix the C haplotype as homozygous, so any affect seen is the result of the F haplotype and not the C haplotype it is linked to, for which I have shown an activation phenotype in $CD4^+$ T cells (Fig. 5.23, Fig. 5.24).

The ASE results show two striking phenotypes with the *IL2RA* F haplotype correlated with: (1) reduced expression of *IL2RA* in B cells *ex vivo* and following activation and (2) reduced expression of *IL2RA* in activated $CD14^+$ and $CD16^+$ monocytes (Fig. 5.38). The results seen in B cells is intriguing as the F haplotype causes a two-fold reduction of *IL2RA* mRNA both *ex vivo* and once the B cells are activated with CpG producing a 30-fold increase in expression, and no other *IL2RA* phenotypes I have identified display a differential expression that is maintained through cell activation. I have only tested monocytes for ASE after stimulation as I have shown by qPCR the expression of *IL2RA ex vivo* is undetectable (Fig. 5.10). I believe it is the decreased expression of *IL2RA* in B cells that is responsible for the protective effect of the *IL2RA* F haplotype; as I will explain in Chapter 6, the F haplotype is also associated with a reduction in the level of soluble CD25 in the blood, which I hypothesise is derived from activated monocytes.

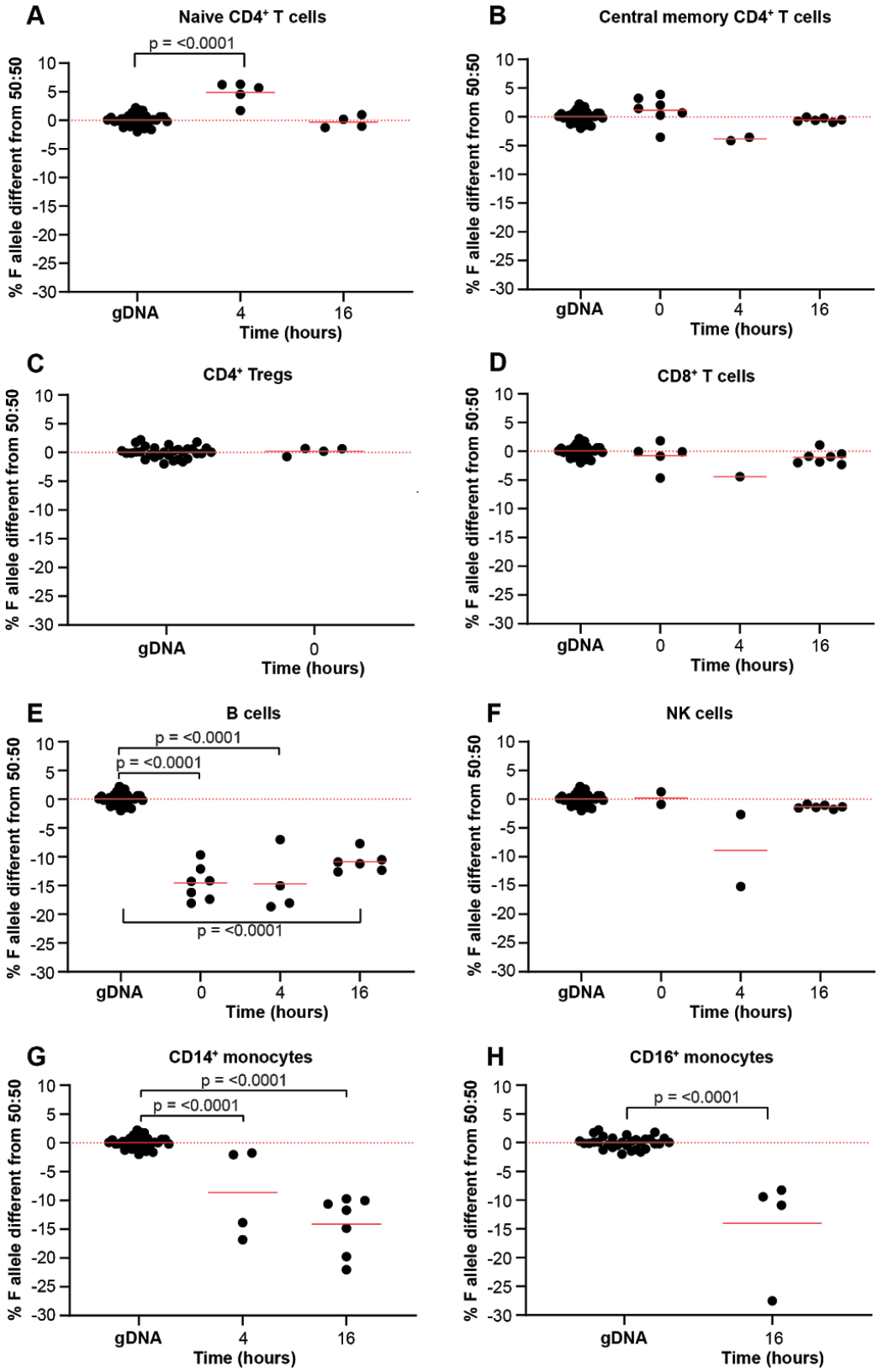

A
Naive CD4+ T cells
p = <0.0001
% F allele different from 50:50
gDNA
4
16
Time (hours)
B
Central memory CD4+ T cells
gDNA
0
4
16
Time (hours)
C
CD4+ Tregs
gDNA
0
Time (hours)
D
CD8+ T cells
gDNA
0
4
16
Time (hours)
E
B cells
p = <0.0001
p = <0.0001
p = <0.0001
gDNA
0
4
16
Time (hours)
F
NK cells
gDNA
0
4
16
Time (hours)
G
CD14+ monocytes
p = <0.0001
p = <0.0001
gDNA
4
16
Time (hours)
H
CD16+ monocytes
p = <0.0001
gDNA
16
Time (hours)

Figure 5.38. The F haplotype makes reduced levels of *IL2RA* mRNA in B cells and monocytes. Various immune cell populations were purified by cell sorting using the gating strategy in

The rarity and LD structure of the F haplotype means that it is currently defined as 13 SNPs, however I have identified two donors carrying minor alleles at only the first 2 or 3 SNPs of the F haplotype. As I have shown in Fig. 5.38, there is a strong phenotype associated with the F haplotype in B cells and both $CD14^+$ and $CD16^+$ monocytes. Therefore I performed ASE on the two donors with recombinant F haplotypes to see if they displayed the F phenotype (Donor A and B, Fig. 5.39A), which would suggest the 2 or 3 F SNPs they carry drive the phenotypes and would refine the number or potential causal variants to explain the F haplotype. The results show that the same phenotype of reduced *IL2RA* made from the F haplotype was seen in $CD14^+$ and $CD16^+$ monocytes from both donors carrying the recombinant haplotypes as that of the standard F haplotype (Fig. 5.39C-D). The results of the *IL2RA* ASE for B cells was harder to interpret as donor A showed the same phenotype as that of the F haplotype with reduced production *IL2RA* mRNA, but donor B did not show an allelic imbalance. Given, that both donor A and B carry the same recombinant *IL2RA* F haplotype this observation was hard to reconcile, and hence a second independent sample was tested and the same result returned (Fig. 5.39B). It was not until the CD25 expression on B cells from the DIL PBMC bank was analysed did the explanation become apparent. In Chapter 6, I will discuss genotype-phenotype correlations with CD25 expression that are not associated with disease, and one of the strongest signals is rs12722502 that causes a large decrease in CD25 expression on B cells. This variant, rs12722502, which is rare (0.7% in the European population) just happens to

be present on the non-F allele of donor B and as this allele is on the opposite DNA strand as the F haplotype and both cause a similar decrease in the production of *IL2RA*, thereby creating a situation where the effect of these haplotypes neutralise each other and a 50:50 ratio is observed (Fig. 5.39B). The F haplotype and the effect of rs12722502 differs in that only the F haplotype effect is maintained with B cell activation, and this is seen in Figure 5.39B, where an ASE is observed in B cells from donor B following activation.

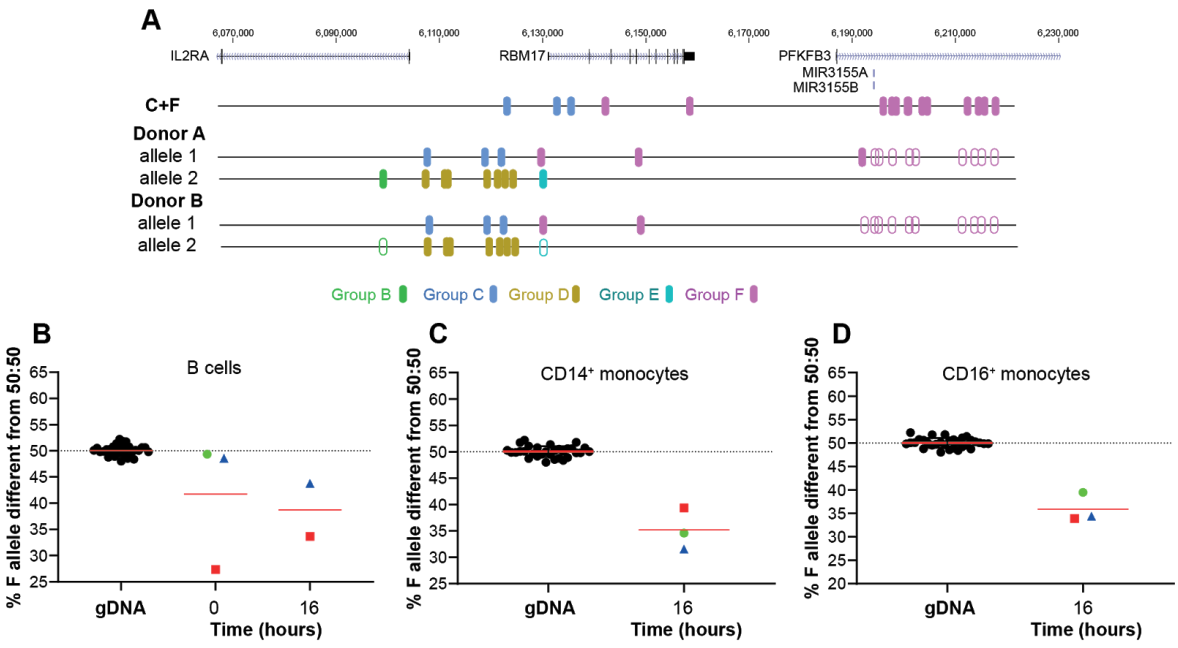


Figure 5.39. Analysis of ASE from two donors carrying rare *IL2RA* F haplotypes refine the number of potential causal variants. (A) Is a schematic of the alleles present in the donors assessed for ASE. The C+F track displays the C and F allele present on the common F haplotype. Donor A contains the C and F haplotypes in which the first three protective F alleles are present heterozygous with the D haplotype. Donor B carries the C and F haplotypes in which the first two protective F alleles are present heterozygous with a rare D haplotype in which the B and E alleles not present. Open ovals represent the major allele and filled ovals the minor allele. The ASE results from (B) B cells, (C) $CD14^+$ monocytes and (D) $CD16^+$ monocytes were measured in donor A (red square) and in two

independent samples from donor B (green circle and blue triangle). A SNP in the 3′ UTR of *IL2RA* was used to assess the allelic imbalance.

5.8.2 The F haplotype causes reduced CD25 expression on B cells

To confirm the results of the ASE seen with the *IL2RA* F haplotype, I analysed the DIL PBMC bank immunophenotyping data for the expression of CD25 on $CD27^+$ B cells from 494 individuals. This analysis showed that there was a substantial decrease in the expression of CD25 associated with the F haplotype. Due to the rarity of the F haplotype, only donors heterozygous for the F haplotype could be compared to donors with no copies of the F haplotype, and the CD25 MFI dropped from an average of 891 in donors with no copies of the F haplotype to 774 in donors with one copy of the F haplotype (Fig. 5.40, $p=4.14x10^{-5}$). However, when this data was parsed through the stochastic search method, three non-disease associated SNP were identified that altered the expression of CD25 on B cells, but surprisingly the F haplotype did not appear in the results. In Chapter 6, I will analyse this phenotype identified as being not associated with disease, and describe how one of the variants altering CD25 on B cells happens to be present on the F haplotype as well as other haplotypes.

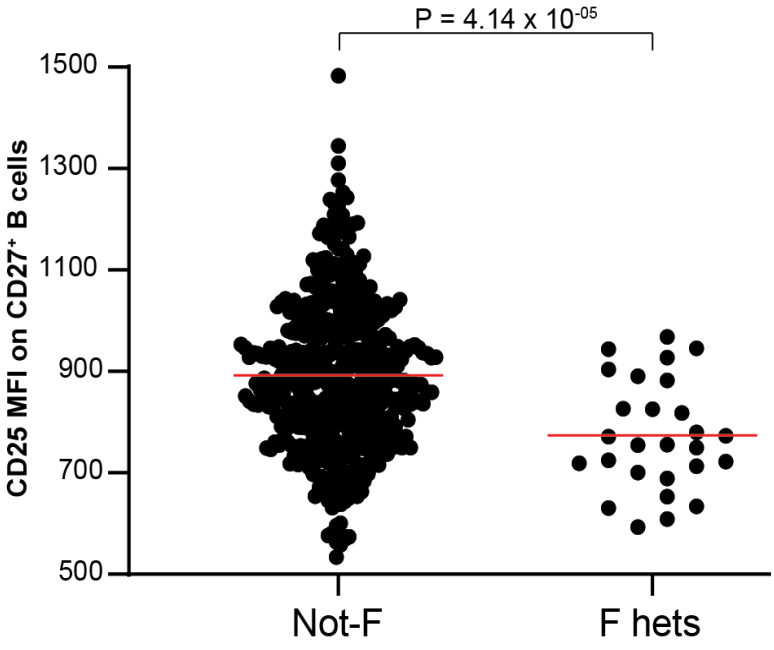


Figure 5.40. The F haplotype correlates with decreased CD25 expression on CD27$^+$ B cells. Immunophenotyping data from 494 individuals was split based on the genotype of the *IL2RA* F haplotype into those carrying no copies of the F haplotype (Not-F) and those with one copy of the F haplotype (F hets). Due to the rarity of the F haplotype no donors homozygous for the F haplotype were seen. Red line depicts the mean of the dataset and P value was calculated using a T Test.

5.8.3 Summary of the F haplotype

The F haplotype is rare and has a large protective effect from T1D, and evidence from UK Biobank indicates that it may be protective for AITD. Through ASE, I have shown that the F haplotype results in decreased expression of *IL2RA* from both B cells (*ex vivo* and with activation) and monocytes. Two donors with rare recombinant F haplotypes show that the most proximal three out of the 13 credible SNPs control the identified phenotype. In Chapter 6, I will present work on *IL2RA* phenotypes not-associated with disease and will show the F haplotype is also associated with decreased levels of soluble CD25 in the blood, as well as the complexity of non-disease associated SNPs affecting CD25 expression on B cells.

5.9 Summary of *IL2RA* disease related phenotypes

Here I have summarised all of the phenotypes identified for the A, C, D and F haplotypes (Table 5.4). Unfortunately, I was unable to find a phenotype that was associated with the E haplotype. I have also proposed which SNP I think is the most likely causal variant for each of the *IL2RA* haplotypes based on the evidence I have presented throughout this chapter.

Table 5.4. Summary of the phenotypes detected for each of the *IL2RA* disease associated haplotypes, as well as the number of SNPs that define each haplotype and which of them is the most likely causal variant.

***ILRA* haplotype**	**No. of credible SNPs**	**Most likely causal variant and evidence**	**Phenotype: Change in *IL2RA* mRNA and CD25 protein expression**
A	8	rs61839660 Rare haplotype and epigenetics	Increased expression *ex vivo* in: $CD4^+$ central memory T cells, $CD8^+$ T cells, NK cells. Decreased expression with activation in: $CD4^+$ central memory T cells, $CD8^+$ T cells, $CD14^+$ monocytes.
C	3	rs11597367 Rare haplotype and epigenetics	Decreased expression with activation in $CD4^+$ naive and memory T cells
D	7	rs62626317 Epigenetics	Decreased expression in $CD4^+$ naive T cells / decreases frequency of CD4+ naive T cells expressing CD25
E	1	rs6602437	Not identified
F	13	rs41295121 rs41295159 Rare haplotype	Decreases expression in B cells *ex vivo* and with activation and monocytes

The consequence of the group A protective haplotype is increased steady state expression of *IL2RA* mRNA in $CD4^+$ central memory T cells, $CD8^+$ T cells and $NKCD56^{bright}$ cells, which was confirmed at the CD25 protein level for $CD4^+$ central memory T cells and $CD8^+$ T cells. The consequence of cell activation was decreased percentages of *IL2RA* mRNA from the A haplotype in $CD4^+$ central memory T cells, $CD8^+$ T cells and $CD14^+$ monocytes. Identifying which of these phenotypes leads to the protective effect of the A haplotype will be a challenge to detangle. The A haplotype is substantially more protective for T1D, than for MS and AITD, leading me to ask if this is because several of these phenotypes are important for the development of T1D but not for MS and AITD? The original finding of the A haplotype was increased expression of CD25 on $CD4^+$ central memory T cells, however, this cell population can be subdivided into Th1, Th2, Th17 and Th22 cells[202]. Further analysis of CD25 expression by these cell types might give a better understanding of how the increased expression of CD25 might be protective for T1D, as IL-2 influences Th cell differentiation and that increased IL-2 signaling could lead to preferential expansion of an anti-inflammatory response.

The A haplotype also correlated with increased *IL2RA* expression on $NKCD56^{bright}$ cells, which is interesting as treatment of MS patients with the CD25 blocking therapy, daclizumab, resulted in positive outcomes that correlated with an expansion of $NKCD56^{bright}$ cells[203]. The increased *IL2RA* expression could result in increased responsiveness to IL-2, which in turn would boost their function and immunoregulatory properties[204].

The C haplotype shared the same activation phenotype as the A haplotype, with decreased *IL2RA* mRNA expression when $CD4^+$ T cells were stimulated. A consequence of decreased CD25 on activated $CD4^+$ T cells would be decreased IL-2 signaling causing less effector T cell expansion as well as less IL-2 consumption by the activated $CD4^+$ T cells, making more IL-2 available for Tregs thereby facilitating regulation and suppression of autoreactive effector T cells.

Decreased expression of *IL2RA* mRNA in $CD4^+$ naive T cells is the phenotype associated with the D haplotype, which was confirmed by immunophenotyping, where a decreased frequency of CD25 expressing naive $CD4^+$ T cells was observed, and reproduces the earlier publication of this phenotype[37]. Naive T cells do not initially express CD25, but through successive rounds of homeostatic proliferation[205], CD25 becomes upregulated and consequently the naive cells become more responsive to IL-2[92], which could lead to accelerated differentiation of autoreactive naive T cells towards effector or memory T cell subsets.

The F haplotype results in decreased expression of *IL2RA* mRNA in both B cells and monocytes, although I believe it is the effect in B cells that is associated with disease, as I show Chapter 6 that the F haplotype also correlates with decreased soluble CD25 levels in the blood that can be attributed to $CD14^+$ monocytes. CD25 is predominantly expressed on memory B cells, which have the ability after activation to present antigen to T cells and

secrete proinflammatory cytokines[206–208], and the protection associated with the F haplotype could result from less IL-2 signaling and less B cell activation. Treatment to deplete B cells with either Rituximab or Ocrelizumab have been used to treat several autoimmune diseases including MS and RA, supporting the hypothesis that B cells have a pathogenic role in disease progression.

5.10 Discussion

Currently in the GWAS catalogue there are 5,687 GWAS studies that have identified 70,000 variant-trait associations[209]. Given that over 90% of these variants will lie in non-coding regions[39], thereby altering the efficiency of enhancer elements that drive gene transcription, either locally or through long-range interactions, eQTL analysis has become a useful tool to assess if a particular variant is associated with gene expression. However, in the GTEx project, 92% of common variants were associated with the expression level of at least one nearby gene[210]. Therefore the accurate identification of the causal variant is key, before embarking on functional experiments to determine the consequence of the polymorphism.

The finding that many autoimmune diseases have been associated with *IL2RA* with only modest overlap between the lead variants suggests that there are either distinct associations for each disease, which seems unlikely, or that the GWAS analysis did not find the true association(s). To address this question, Dr Chris Wallace (University of Cambridge)

modified a stochastic search algorithm to facilitate fine mapping and the results were key to understanding the genetic association of *IL2RA*[2]. From this analysis, in conjunction with haplotype analyses using the results of the stochastic search, it was clear that the *IL2RA* A haplotype was associated not only with protection from T1D, but also MS and AITD. The minor allele at rs61839660, which I identified as the most likely causal variant for the A haplotype, has also been identified as protective for primary sclerosing cholangitis[45] but in contrast is associated with susceptibility to Crohn's disease[14], hayfever and eczema[211,212].

Apart from the A haplotype, additional protective signals were identified for T1D (A, C, E and F), MS (A and D) and AITD (A and C), some of these were shared like the C haplotype in T1D and AITD, and some were disease specific, like D for MS and E and F to T1D. This analysis identified signals different to what the original GWAS found, and the use of stochastic search was better at dissecting the complex genetic association at *IL2RA*, where GWAS identified a SNP that joint-tagged two independent haplotypes for both MS and AITD. We could not access the raw genotyping data for the GWAS from ulcerative colitis[213], rheumatoid arthritis[214] or alopecia areata[215], in order to apply the stochastic search method to fine map the *IL2RA* region, however when we added the lead variants rs4147359 for UC, rs10795791 for RA and rs3118470 for AA to a haplotype analysis, it appears they may tagging three signals, which are the *IL2RA* A, C and D haplotypes. Since instigating this work, a much larger GWAS for MS has been published[24] and *IL2RA* was one of the most associated genes with the lead SNP identified as rs11256593, where the minor allele, present at a 45% frequency, is protective. This new

lead SNP is in moderate LD with the previous best signal, rs2104286 (D′ 0.88, r^2 0.28, minor allele frequency 22%). This new result in not compatible with the two haplotype model we proposed (A and D) and suggests that with a larger number of MS and control samples that the tagging SNP has changed because it now encompasses another association. It would be exciting if this data could be analysed through the stochastic search method and perhaps identify a third association signal in addition to the A and D haplotypes.

The field of epigenetics is rapidly advancing with robust high throughput and genome wide methods available to measure open chromatin with DNAse-Seq or ATAC-Seq, quantify histone modifications defining regions of DNA with enhancer, repressor or promoter marks and assess the 3D formation of DNA with looping measured by Hi-C or promoter capture Hi-C. Analysis of the *IL2RA* region for epigenetic marks provided additional evidence as to which of the A and D haplotype SNPs are our best candidates. For the A haplotype, there is evidence that the promoter of *IL2RA* loops into a region of intron 1 which contains our best SNP rs61839660 providing a mechanism of how the SNP could influence the expression of *IL2RA*, with the protective allele providing a stronger enhancer sequence.

However, for the C haplotype NK cells had the strongest epigenetic evidence, but in the experimental conditions analysed I could not see any difference in *IL2RA* expression by genotype. There was less epigenetic evidence available for activated $CD4^+$ T cells, but this

is where we saw the effect of the C haplotype. A recent publication performed detailed analysis of chromatin activity in $CD4^+$ T cells cultured in 13 and monocytes in six different conditions at different time points and is the kind of dataset needed to thoroughly examine genetic overlap with chromatin marks[42].

How relevant is this work to the fine mapping of other regions or is the *IL2RA* region the most complex? For T1D, a third of the regions fine mapped contain at least two signals and in situations where two signals are of similar effect size and in the same direction the stochastic search algorithm performs better than stepwise regression. Additionally, the stochastic search method is not just limited to GWAS data, but can be applied to identify genetic correlations with quantitative measures and in Chapter 6 will be used to discover genotype-phenotype correlations that are not associated with disease.

Chapter 6

Non-disease associated variants that alter *IL2RA* expression

6.1 Introduction

This Chapter will describe two different *IL2RA* mRNA / CD25 phenotypes which are under genetic control, but appear to have no influence on susceptibility to autoimmune disease. The first relates to the amount of soluble CD25 (sCD25) detected in the peripheral circulation and the second was discovered serendipitously while trying to understand the ASE results presented in Chapter 5, during which several SNPs were discovered that alter the expression of *IL2RA* in B cells.

Across a 150 Kb region of the genome that includes *IL2RA*, there are 6,102 SNPs with an allele frequency greater than 1%, and hundreds of these have the potential to influence gene expression as they bind transcription factors, reside in enhancer regions or are associated with an eQTL. The Regulome database has scored all variants (>1% frequency) based on these parameters and aids interpretation of regulatory variants in the human genome[184]. I have plotted the regulome scores for all SNPs across the *IL2RA* region and there are hundreds of SNPs with a score the same or greater than the lead SNP for the *IL2RA* A haplotype (rs61839660, regulome score 3), therefore it is not surprising that some

additional SNPs other than those currently known to be associated with disease will alter *IL2RA* expression (Fig. 6.1).

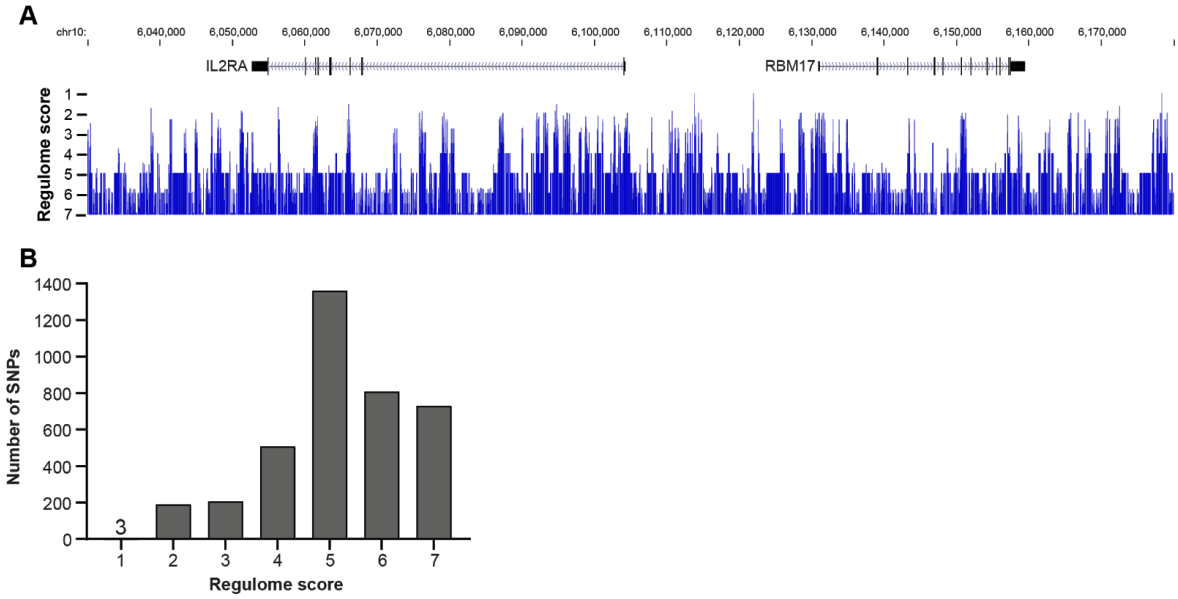


Figure 6.1. There are hundreds of SNPs across the *IL2RA* region that have the potential to alter its expression. (A) From the Regulome database[184], the scores of all SNPs with allele frequency >1% across a 150 Kb region were downloaded and converted to a custom track that could be loaded into the UCSC browser. The Regulome scores range from 1 (most evidence) to 7 (least evidence) that the SNP has the potential to be functional. (B) The number of SNPs assigned each of the Regulome scores from 1 through to 7 are graphed. Only three SNPs have a Regulome score of 1, but there are hundreds with scores equal to or less than 3 which is the same as the *IL2RA* group A haplotype.

6.2 Soluble CD25

It has been known for over thirty years that soluble CD25 (sCD25) is released from the surface of activated human T cells[216] by the cleavage of membrane bound CD25 by matrix metalloproteinases (MMP-9[REF217]), rather than through splicing of the transmembrane

domain of *IL2RA*[218]. The levels of sCD25 are increased in patients with autoimmune disease, including T1D[219] and MS[170] and provide a useful inflammatory marker which is a diagnostic criteria for hemophagocytic lymphohistiocytosis[220] where extremely high levels are seen.

Previous reports have shown that the production of sCD25 is under genetic control, with a SNP that is part of the *IL2RA* C haplotype (rs11594656) correlating with higher levels[169], a more recent study refined this signal to rs12722497, which is found 50% of the time on the *IL2RA* C haplotype[193]. Ricardo Ferreira (DIL) has performed an ELISA for sCD25 on 4082 T1D cases and 5197 healthy controls individuals and Jamie Inshaw (DIL) performed the statistical analysis using the same stochastic search method applied to the GWAS data allowing for multiple signals to be identified. In this dataset, the SNP rs12722497 was confirmed as the only signal that caused an increase in the serum levels of sCD25, but in addition, several independent signals, one of which was the *IL2RA* F haplotype, resulted in a ~20% decrease in the serum levels of sCD25.

My part in this project was to identify which cell type was responsible for the increased sCD25 seen in individuals carrying the minor allele at rs12722497 and the decreased sCD25 associated with the F haplotype. My hypothesis is that it is either increased (for rs12722497) or decreased (for the F haplotype) levels of sCD25 resulted from cleavage of increased or decreased surface expression of CD25. Since all of the associated SNPs were either intronic or intergenic, a mutation in the cleavage site of the CD25 protein was not a

potential explanation for the altered levels. Additionally there is no evidence that a splice form of *IL2RA* that skips the transmembrane exon generates sCD25.

Therefore, the aims for this Chapter are to:

1. Understand the relationship between the SNPs associated with protection from autoimmune disease and those associated with changes in the levels of sCD25.
2. Investigate in which cell type rs12722497 affects *IL2RA* mRNA expression.
3. Investigate the F haplotype correlation with decreased sCD25 levels and compare to phenotypes that are associated with protection from T1D.

6.2.1 Haplotype analysis of variants associated with a change in sCD25 concentration

First, a haplotype analysis was performed by Jamie Inshaw (DIL), which compared all of the alleles identified by stochastic search that correlated with a change in the levels of sCD25. This showed that the minor allele of rs12722497, which caused a ~25% increase in sCD25 was found on approximately 50% of the individuals carrying the *IL2RA* C haplotype. No difference in protection from T1D was seen when the C haplotype with or without rs12722497 were compared (Fig. 6.2, column 4 vs 6), indicating that rs12722497 is not associated with T1D. The *IL2RA* F haplotype was unique in that it showed both protection from T1D and a change in sCD25 concentration (Fig. 6.2, column 14). The A and D haplotypes were not associated with a change in sCD25 levels, but carried alleles that did alter sCD25 levels (for A haplotype, rs11256611, column 5 vs 8; for D haplotype, rs12722588, column 2 vs 11 or 13; Fig. 6.2).

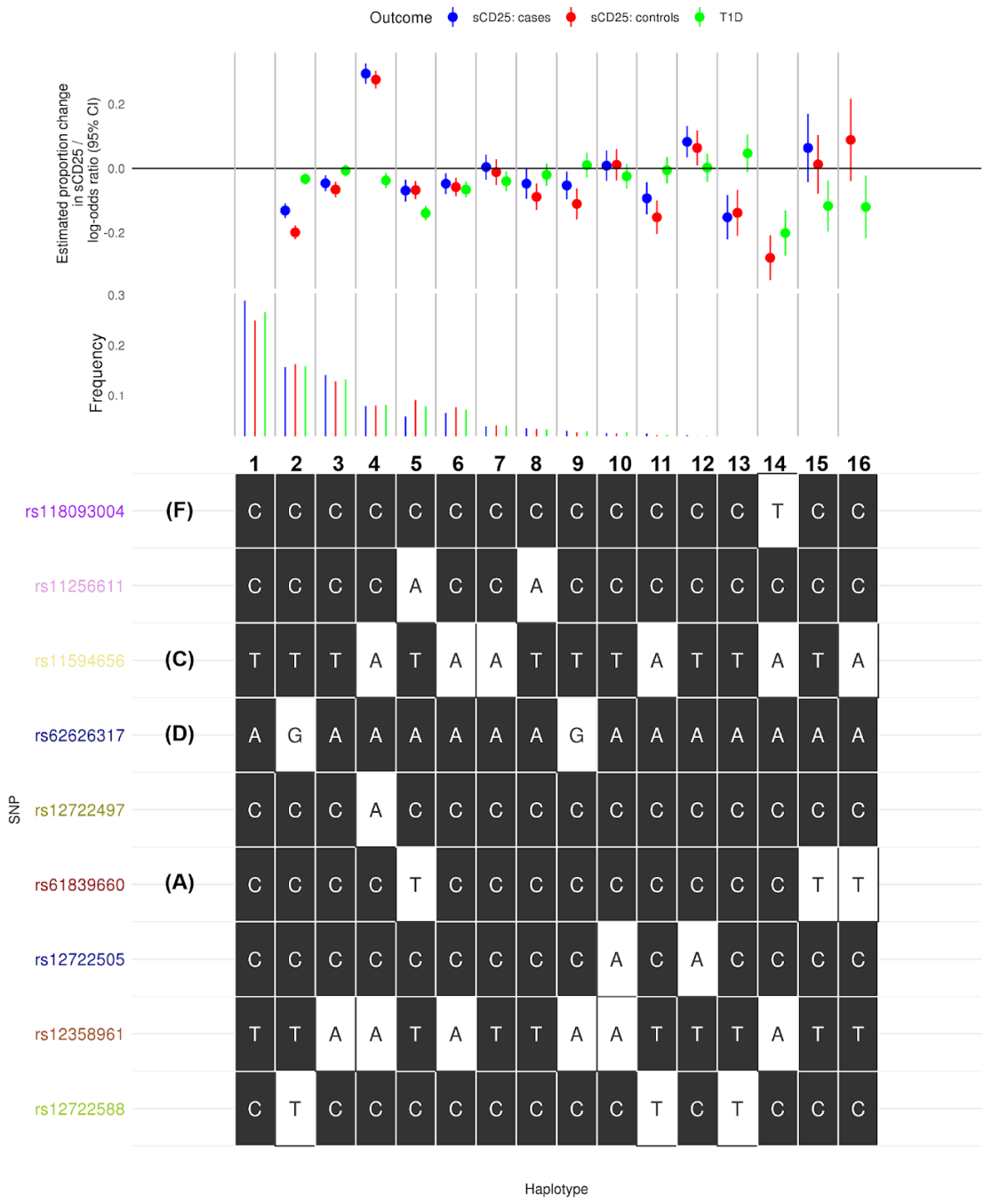


Figure 6.2. Haplotype analysis of the variants identified as correlating with sCD25 levels. The concentration of sCD25 was measured by ELISA in 4082 T1D cases and 5197 healthy controls and stochastic search was used to identify correlated variants. Haplotype analysis was performed using these variants as well as a tag for the A, C, D and F haplotypes (shown in parentheses). The change in sCD25 concentration is plotted for the T1D cases and healthy controls as well as the odds ratio for T1D susceptibility in haplotypes of decreasing frequency from left to right.

6.2.2 Comparison of alleles found on the C haplotype split by rs12722497, that alter the levels of sCD25

To understand the context of rs12722497 on the C haplotype, I have taken the analysis I presented in Figure 5.3, where I showed how similar donors carrying the C haplotype were to each other and split the C haplotype by rs12722497. The results are fascinating and show two different patterns of alleles; the first is the C haplotype which carries the major allele at rs12722497 that displays the same pattern of alleles seen in Figure 5.3 with the region spanning the C haplotype SNPs being fixed, but outside the C SNPs there is increased segregation of alleles. The C haplotype that carries the minor allele at rs12722497 shows that the vast majority of alleles across a 100 Kb region are not segregating, which implies that a single founder event occured and all haplotypes carrying the minor allele at rs12722497 are descendants of the mutation.

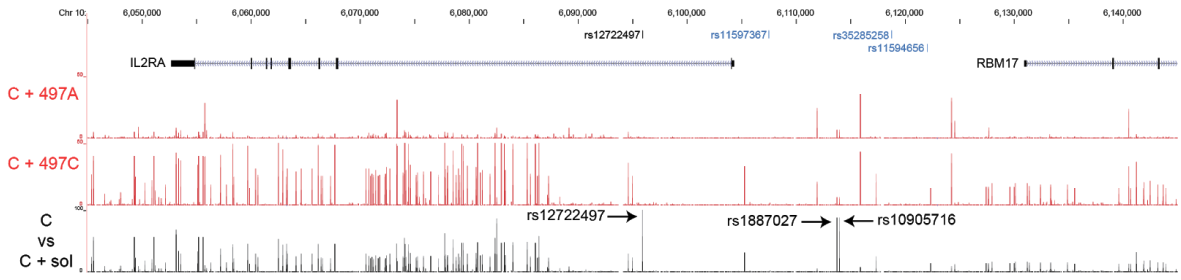


Figure 6.3. Comparison of alleles found on the C haplotype split by rs12722497. From the 1000 Genomes Project, the 233 C haplotypes identified in Figure 5.3 were split by rs12722497: 112 carried the major C allele (+497C) and 121 the minor A allele (+497A). The location of the group C SNPs are displayed in blue and rs12722497 in black along with the *IL2RA* and *RBM17* genes. Each haplotype is displayed (in red) as a separate row and the height of each point is related to the proportion of individuals having the same alleles for the SNP (score of zero) through to having equal counts of the two SNP alleles (score of 50). The distribution of alleles between the *IL2RA* C haplotype split by rs12722497 was compared (black) and each vertical line shows the proportion of each

SNP allele that is common between the two haplotypes, with a score of 0 meaning they have the same allele and a score of 100 meaning that all individuals in the haplotype comparison have the opposite allele. Two SNPs, rs1887027 and rs10905716 are highlighted (black arrows) as these variants segregate almost perfectly on the opposite haplotypes with rs12722497.

6.2.3 Which cell type does rs12722497 affect?

The method chosen to assess which cell type was responsible for the genetic control of sCD25 by rs12722497 was ASE. In Chapter 5, I have described disease related phenotypes for the A, C, D and F haplotypes, therefore I will analyse donors that carry the C haplotype and carry the major allele of rs12722497 heterozygous with the C haplotype and the minor allele of rs12722497, thereby fixing the effect of the C haplotype as homozygous, therefore any allelic imbalance seen will result from rs12722497.

I have used the same strategy used to assess ASE in Chapter 5 from purified immune cell populations using donors recalled as having the genotype of interest. The results of the ASE show that B cells and both $CD14^+$ and $CD16^+$ monocytes display increased *IL2RA* production from the haplotype carrying the minor (A) allele of rs12722497 (Fig. 6.4). This is the same direction as the increased production of sCD25, and implies the increase in *IL2RA* would translate to an increase in CD25 expression on B cells and monocytes, thereby more protein available to be cleaved resulting in a higher sCD25. The increase in *IL2RA* expression on B cells is not maintained with activation, which causes a 30-fold increase in expression. This differs from the B cell phenotype described for the F

haplotype, which was maintained with activation. In $CD14^+$ monocytes, there is a large spread in the size of the allelic imbalance seen, although the direction is consistent. I have compared the level of *IL2RA* mRNA expression to the size of the ASE, and those samples with the lowest *IL2RA* expression have the biggest allelic imbalance (data not shown). The activation phenotype described in Chapter 5.4.3 for $CD4^+$ naive and memory T cells was not seen since the C haplotype was homozygous.

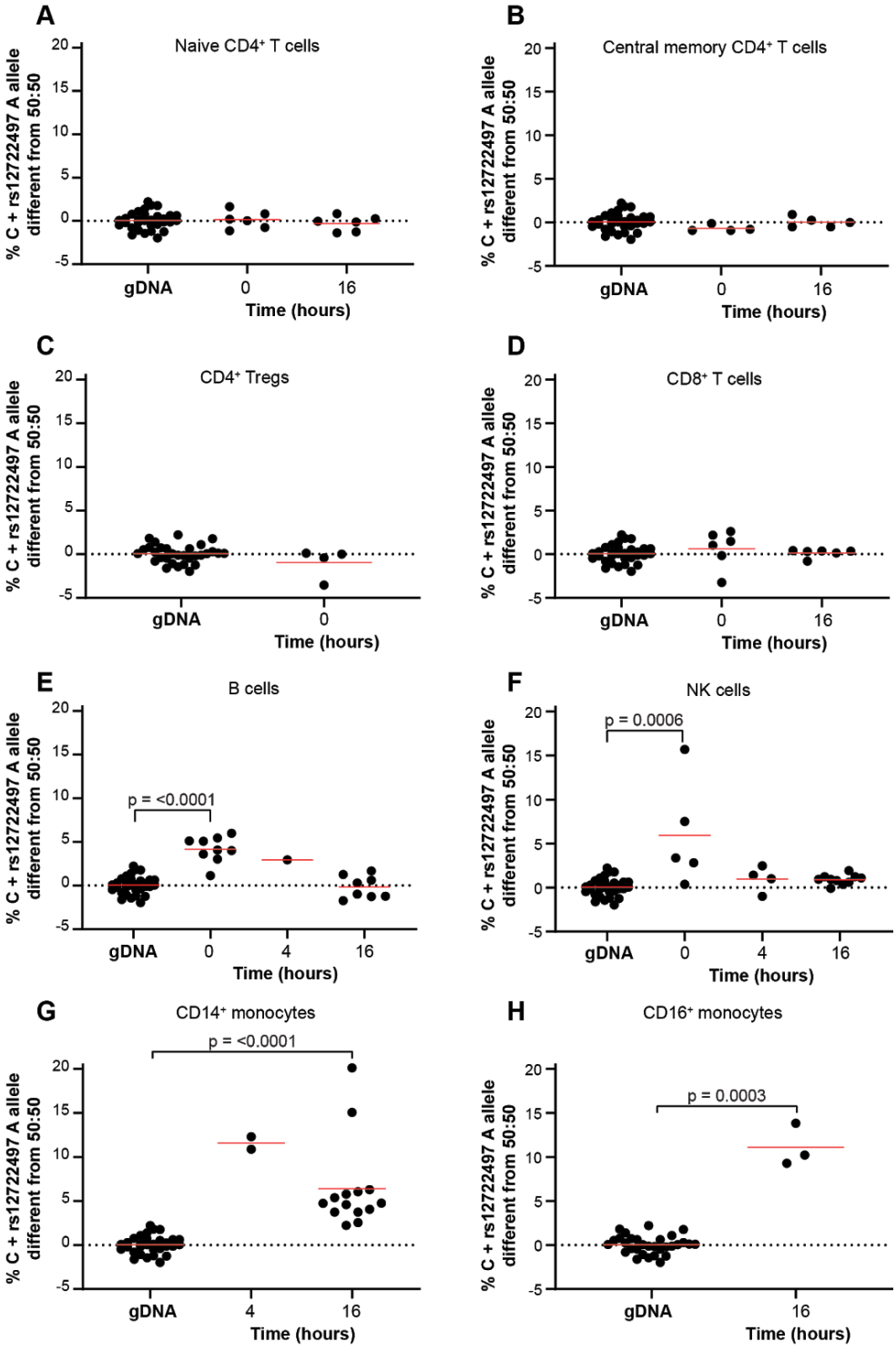


Figure 6.4. rs12722497 increases the expression of *IL2RA* mRNA in B cells, CD14$^+$ and CD16$^+$ monocytes. Various immune cell populations were purified by cell sorting using the gating strategy in Figure 5.9 and were tested for ASE *ex vivo* or after activation (4 and 16 hours). T cells were stimulated with CD3/CD28 activation beads, B cells with CpG, NK

cells with IL-12, IL-15 and IL-18 and $CD14^+$ monocytes with GMCSF and IFNɣ. Donors were homozygous for the C haplotype and the SNP rs12244380 in the 3′ UTR of *IL2RA* was used to measure the degree of allelic imbalance that is reported as the percent made by the C haplotype. Red line displays the mean of the dataset and P-values calculated with a Mann-Whitney test.

6.2.4 CD25 expression on B cells is affected by three SNPs

We have collected a large immunophenotyping dataset from over 500 individuals who have dense genotyping information that allows genotype to phenotype studies to be performed. In Chapter 5.3.5 I described the method used to normalise the dataset as it took 708 days to complete the collection. Immunophenotyping was performed on fresh blood and staining for $CD4^+$ T cell, $CD8^+$ T cell and B cells was performed. The ASE results showed an allelic imbalance with rs12722497 in B cell as well as $CD14^+$ and $CD16^+$ monocytes. Only B cells and not monocytes were analysed for the expression of CD25, and the normalised CD25 expression values I produced were analysed by Jamie Inshaw (DIL) using the same Bayesian method used to analyse the GWAS data in order to identify any variants that altered the expression of CD25 on B cells in an unbiased way. I expected that rs12722497 and the F haplotype to be the top signals as I have shown these affected the expression of *IL2RA* mRNA in B cells. Surprisingly, neither the F haplotype or rs12722497 signals were detected, however three different signals were seen, which are rs1887027, rs12722502 and rs71479745.

To show the direction of effect of these three SNPs (rs1887027, rs12722502 and rs7149745) a haplotype analysis was performed (Fig. 6.5A). A second haplotype analysis including a variant from each of the *IL2RA* A, C, E and F haplotypes was performed to assess the distribution of SNPs affecting CD25 expression on B cells versus those affecting T1D (Fig. 6.5A). Both haplotype analyses were performed by Jamie Inshaw (DIL). The first analysis showed that the minor alleles of rs1887027 and rs12722502 resulted in 10% and 20% reduction in CD25 expression on B cells, respectively (column 2 and 4, Fig. 6.5A). Whereas the minor allele of rs71479745 resulted in a 15% increase in CD25 expression on B cells (column 3, Fig. 6.5A). The second haplotype analysis which included the *IL2RA* A, C, E and F haplotypes was key to understanding these surprising results. First, the minor allele at rs1887027 is located on approximately 50% of the *IL2RA* C haplotypes, and segregates in the opposite direction to the minor allele at rs12722497 (column 2 vs 5, Fig. 6.5B, and cartoon representation of the C haplotypes Fig. 6.6), and as rs1887027 is associated with decreased CD25 expression on B cells, the ASE results would look like the minor allele of rs12722497 makes more (as in Fig. 6.4E). Second, rs1887027 is located on the F haplotype, so the effect of F could be masked by the effect of rs1887027, as they both cause a decrease in CD25 expression (column 2 and 6, Fig. 6.5B). In the ASE experiments with the F haplotype, rs1887027 was fixed as homozygous, and therefore the effect of the F haplotype on B cells is independent of rs1887027.

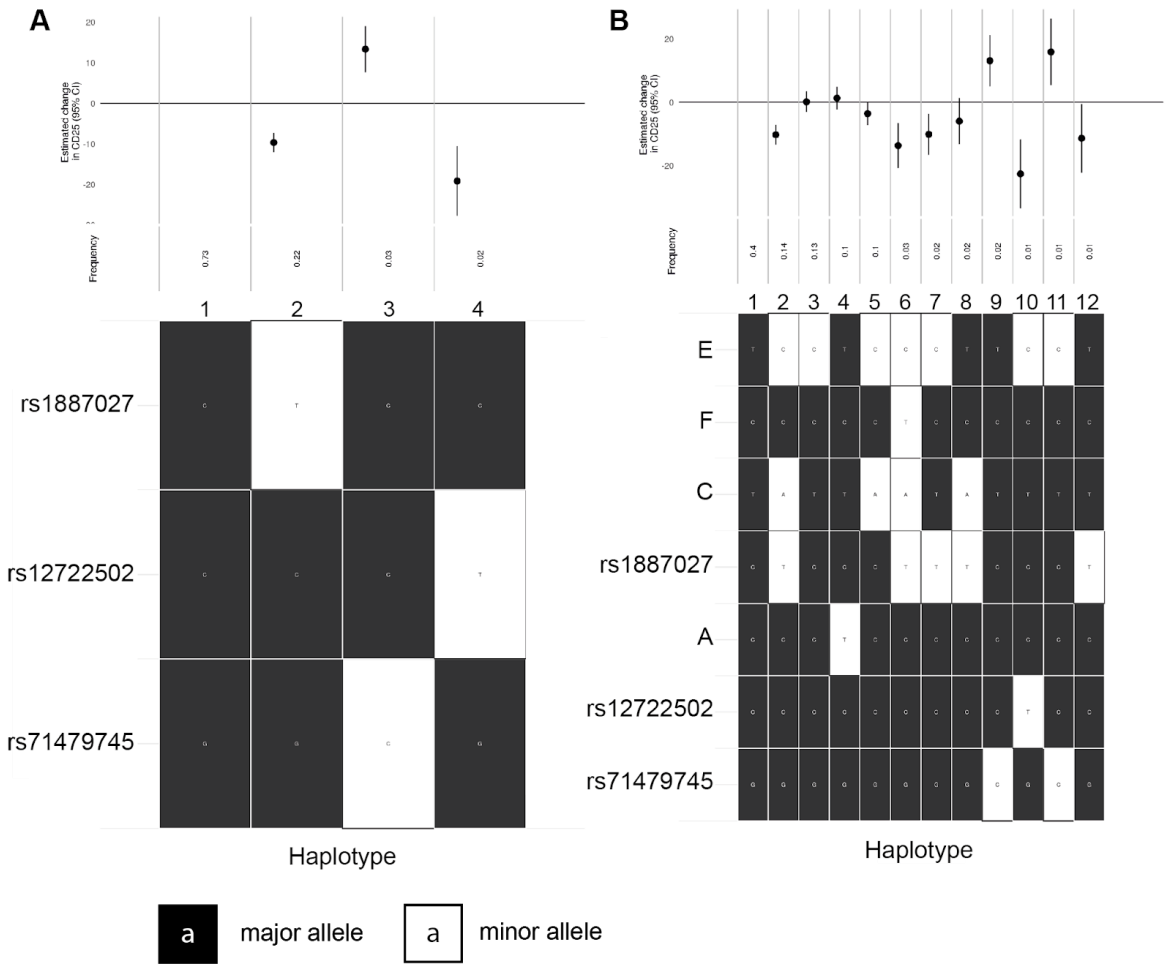


Figure 6.5. Haplotype analysis of the SNPs associated with altered expression of CD25 on B cells. (A) The three SNPs, rs1887027, rs12722502 and rs71479745 alter the expression of CD25 on B cells, with the minor alleles of rs1887027 and rs12722502 (column 2 and 4) being associated with decreased CD25 expression on B cells, whereas the minor allele of rs71479745 is associated with increased CD25 expression on B cells (column 3). The haplotypes defined by these three SNPs are compared to a haplotype in which the major alleles of these SNPs are present and set as the reference (column 1). (B) Haplotype analysis of the three SNPs identified in (A) as well as the lead SNPs for the T1D associated A, C, E and F *IL2RA* haplotypes. The minor allele of rs12722497 that results in increased sCD25 protein is not shown, but would be found on the C haplotype depicted in column 2. Figure 6.6 shows the relationship of rs12722497 and rs1887027.

The SNP rs1887027 sounded familiar, and it was one of two variants identified in Figure 6.3, that differed between the C haplotypes split by rs12722497. This variant is also one of the three variants (out of 6102 SNPs) in the *IL2RA* region that has a Regulome score of 1 (Fig. 6.1), the highest score, which signifies that the SNP has the highest level of evidence supporting the hypothesis that it resides in a regulatory region. At this point I realised that the B cell ASE I had observed (Fig. 6.4) may not be a consequence of rs12722497, but could actually be caused by rs1887027. The minor allele of rs1877027 is almost always carried with the major allele of rs12722497, and the reverse is seen with the major allele of rs1887027 present on the same haplotype as the minor allele of rs1227497 (Fig. 6.6). The higher level of *IL2RA* mRNA seen from the haplotype carrying the minor allele of rs12722497 (Fig. 6.4) could also be achieved if the strand carrying the minor allele of rs1887027 made less *IL2RA* mRNA, which could be the case given the CD25 expression result (Fig. 6.5).

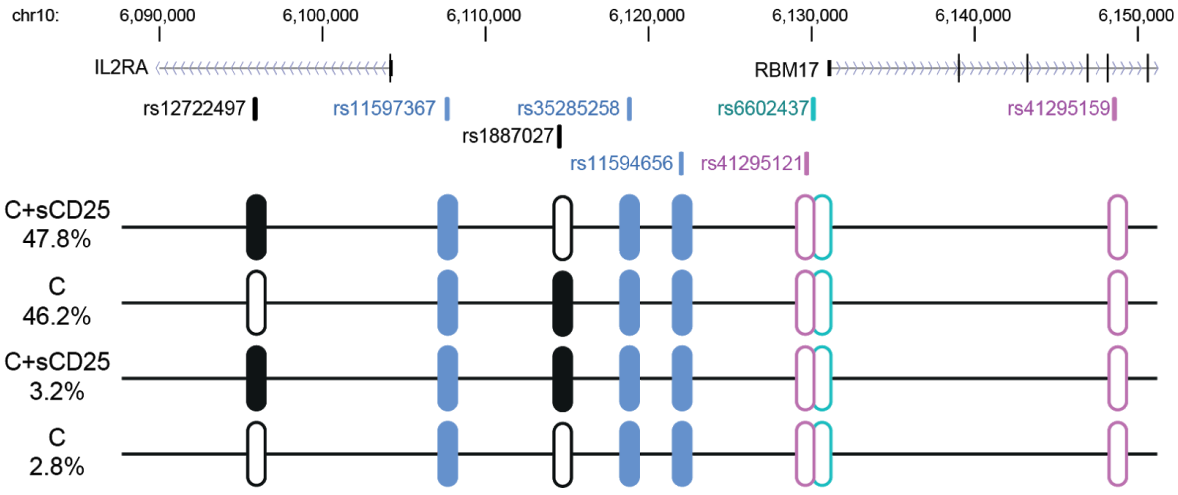


Figure 6.6. A cartoon representation of the *IL2RA* C haplotype showing the relationship between rs12722497 and rs1887027. Predominantly, the minor alleles of rs12722497 and rs1887027 are not on the same haplotype, even though they occur at a similar

frequency in the European population. The location of these SNPs are shown relative to the *IL2RA* and *RBM17* genes. The frequency of each allelic combination present on the C haplotypes is shown as a proportion of the total number of C haplotypes. For simplicity, only the two most proximal F haplotype SNPs are shown; the location of all 13 F haplotype SNPs are shown in Figure 5.1 and span 84 Kb. Filled ovals represent the minor protective alleles and open ovals the major susceptibility allele.

6.2.5 Repeat ASE for sCD25 controlling for rs1887027

A new strategy was needed to test if the allelic effect seen in B cells was caused by rs1272497 or rs1887027. The results presented in Figure 6.4 used donors homozygous for the C haplotype and heterozygous for rs12722497, thereby controlling for the effect of the C haplotype. However, this design allowed rs1887027 (not known at that time) to influence the results. The new strategy selected donors carrying the C haplotype and minor allele at rs12722497 heterozygous with the susceptible haplotype (therefore homozygous for the major allele at rs1887027). Such donors allowed the activation-dependent genotype to phenotype effect of the C haplotype to be observed in $CD4^+$ T cells; heterozygosity at rs12722497 did not alter this phenotype (Fig. 6.7 vs Fig. 5.24).

Using sorted cells from donors based on this new genotype-selection strategy, the allelic imbalance seen in $CD14^+$ monocytes (Fig. 6.4) was confirmed with the minor allele at rs12722497 making more *IL2RA* mRNA (Fig. 6.7). Due to limited sample availability, it was not possible to test for ASE in $CD16^+$ monocytes. Importantly, no ASE was seen in B cells from these donors, which means the B cell ASE results shown in Figure 6.4 were caused by the presence of the minor allele at rs1887027 (Fig. 6.5).

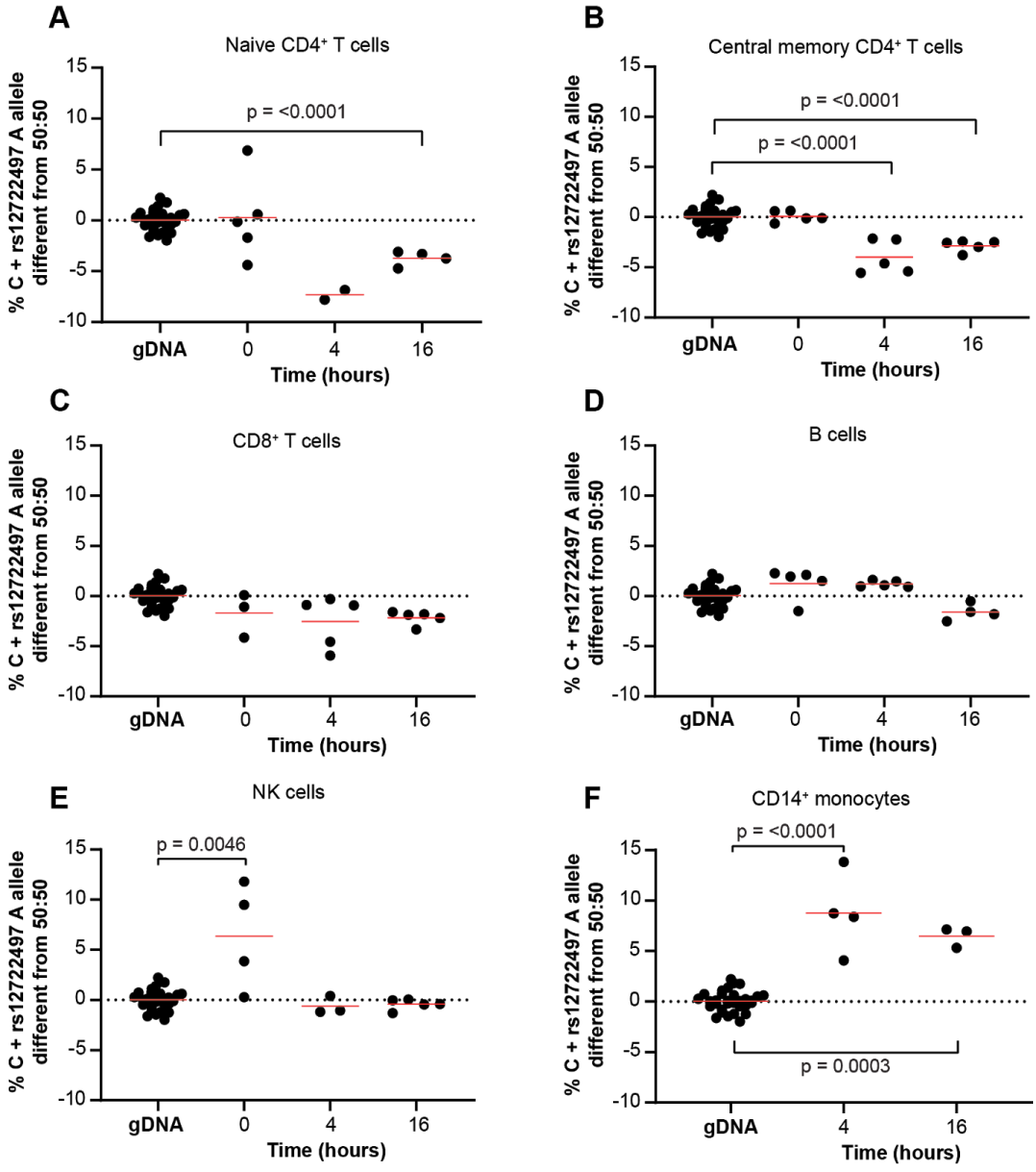


Figure 6.7. rs12722497 increases the expression of *IL2RA* mRNA in CD14$^+$ monocytes. Various immune cell populations were purified by cell sorting using the gating strategy in Figure 5.9 and were tested for ASE *ex vivo* or after cell stimulation (4 and 16 hours). T cells were stimulated with CD3/CD28 activation beads, B cells with CpG, NK cells with IL-12, IL-15 and IL-18 and CD14$^+$ monocytes with GMCSF and IFNɣ. Donors carried the C haplotype and the minor allele at rs12722497 heterozygous with the susceptible haplotype. A SNP in the 3′ UTR of *IL2RA* was used to measure the degree of allelic imbalance that is reported as the percentage produced by the minor allele of rs12722497. Red line displays the mean of the dataset and P-values were calculated with a Mann-Whitney test.

From the ASE data, it was the activated $CD14^+$ monocytes that showed increased *IL2RA* mRNA correlated with the rs12722497 A allele. Since activated monocytes are normally a tissue resident population, I hypothesised that monocytes expressing CD25, located in tissues, would contribute to the circulating levels of sCD25 via cleavage of surface CD25. The rs12722497-dependent ASE we observed after *in vitro* activation of monocytes is supported by eQTL evidence from the GTEx Project[44], which shows there is an eQTL for rs12722497 in spleen ($p=9.1x10^{-26}$) and adipose tissue ($p=2.0x10^{-15}$) but not blood. Additionally the cell culture supernatant of activated monocytes has been measured for sCD25 and were found to produce more than activated T cells[221,222]. Furthermore, a whole blood eQTL dataset containing information from 30,000 individuals[43] showed there was a small eQTL with rs12722497, but this was the same size and direction as the eQTL for rs11597367 (*IL2RA* C haplotype) suggesting that a stochastic search of this expression data might reveal further insights.

6.2.6 The region surrounding rs12722497 is demethylated in $CD14^+$ monocytes

The DNA sequence surrounding rs12722497 is CpG dense with seven methylation sites within 170 bp (Fig. 6.8A) with rs12722497 located in the middle. Therefore I decided to assess the methylation status of these seven CpG sites in purified immune cells, and as rs12722497 is located between the CpG sites, I can also assess the methylation status of each allele. I am using the method described in Chapter 3 that I developed to measure the methylation of the *FOXP3* TSDR, but modified to divide the results based on rs12722497 alleles. The results show that the seven CpG sites surrounding rs12722497 are almost

completely demethylated in CD14$^+$ monocytes, with decreasing levels of demethylation seen in the other cell types assessed, with CD4$^+$ naive T cells, CD8$^+$ T cells and NK cells the most methylated, with approximately 20% of the reads being demethylated (Fig. 6.8B). If these results are split based on the genotype at rs12722497, then it is the minor allele of rs12722497 which is more demethylated in all cell types assessed (Fig. 6.8C) with approximately 20% more demethylation seen (Fig. 6.8D). I hypothesise that monocytes produce the rs12722497-dependent increase of sCD25, and these methylation results support this notion since CD14$^+$ monocytes are most demethylated at rs12722497, with the allele associated with increased sCD25 being more demethylated.

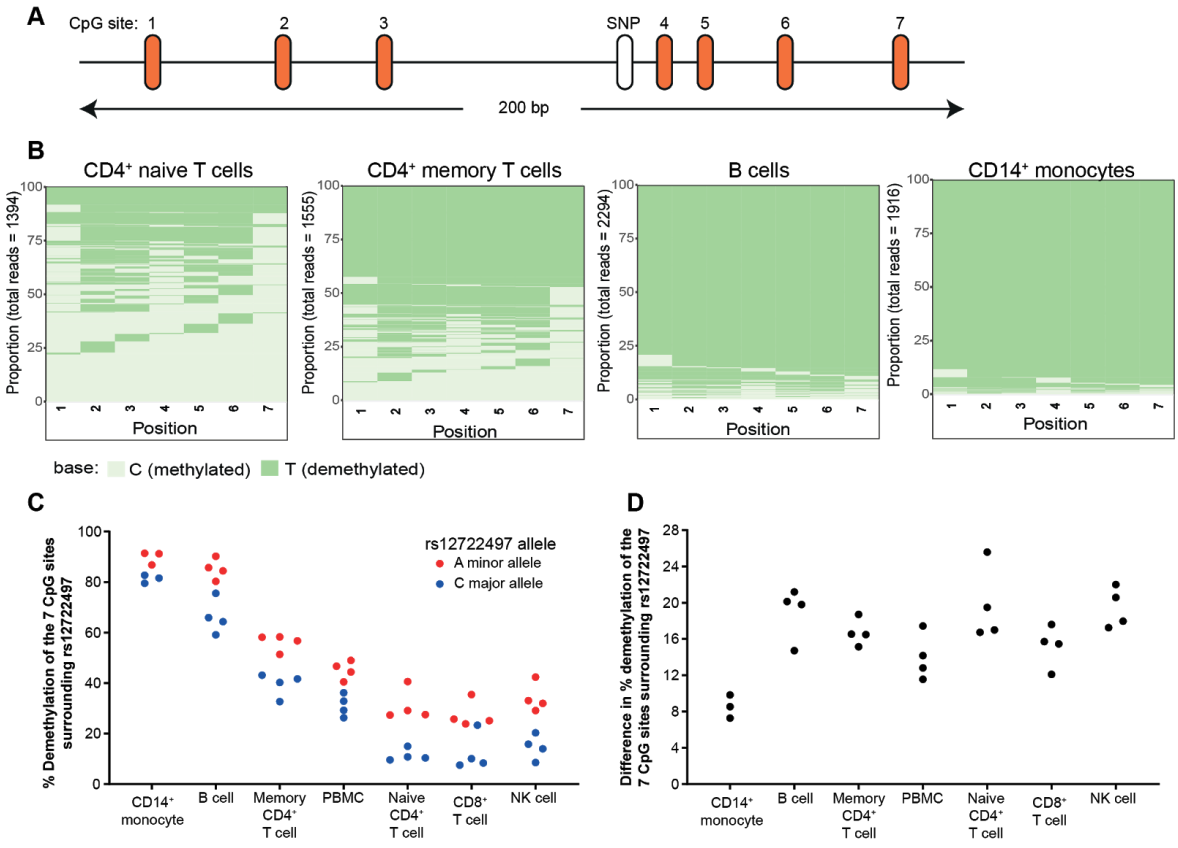


Figure 6.8. Allele-specific methylation of the CpG sites surrounding rs12722497. (A) Schematic of the seven CpG sites (orange ovals) in relation to rs12722497 (white oval). (B) Examples of the methylation profile of CD4$^+$ naive and memory T cells, B cells and

$CD14^+$ monocytes, with the x axis displaying the seven CpG sites and the y axis are stacked sequencing reads displaying the methylation status of each CpG site, with dark green representing a demethylated and light green a methylated CpG site. (C) The percent demethylation of the seven CpG sites surrounding rs12722497 was calculated per allele of rs12722497, from which the minor A allele (red circles) is more demethylated than the major C allele (blue circles). (D) From each sample, the difference in demethylation measured from the CpG sites shared on the major and minor alleles of rs12722497 was calculated.

6.3 B cell phenotypes

When trying to understand which cell type was producing the sCD25 under genetic control of rs12722497 I discovered that there were at least three SNPs associated with changes in CD25 expression on B cells (Section 6.2.4): rs1887027, rs12722502 and rs71479745. To explain the results of the B cell ASE I presented a haplotype analysis with these three SNPs in conjunction with the A, C, E and F *IL2RA* haplotypes (Fig. 6.5), which showed that the B cell ASE result was due to the presence of rs1887027 and not rs12722497.

In this section, I will present data showing how each of the three SNPs identified affect the expression of CD25 on B cells (Fig. 6.9). This is immunophenotyping from the DILPBMC bank, where 498 individuals were analysed for B cell phenotypes and the largest difference was seen with $CD27^+$ B cells, which have a memory phenotype. As mentioned in section 6.4, rs1887027 is found on about half of the individuals carrying the C haplotype, but is not found on the C haplotype that also carries the minor allele of rs12722497 that increases the concentration of sCD25 in blood. For each minor T allele of rs1887027, there is a 66 unit drop in CD25 MFI on B cells (Fig. 6.9A).

The SNP rs12722502 has a Regulome[184] score of 2, again making this SNP a good candidate to be responsible for altering CD25 expression on B cells. Indeed, it has been shown that the minor T allele disrupts an ELF1 transcription factor binding site, and has five-times weaker binding than the major C allele[223]. The minor allele of rs12722502 is rare, with an allele frequency of 0.7%, consequently only seen as heterozygous in our collection and there is over a 100 unit decrease in CD25 MFI on B cells with one copy of this minor allele (Fig. 6.9B).

The third SNP identified, rs71479745, was unlike the previous two in that there was no epigenetic evidence for its effect (nor for the six other SNPs in high LD, $r^2 > 0.88$), and secondly as it was associated with an increase in CD25 expression on B cells. This low frequency variant (rs71479745, MAF 5%) was tested as heterozygous, and only one donor homozygous for the minor C allele was identified. One copy of the minor C allele of rs71479745, resulted in a 117 unit increase in CD25 MFI on $CD27^+$ B cells (Fig. 6.9C).

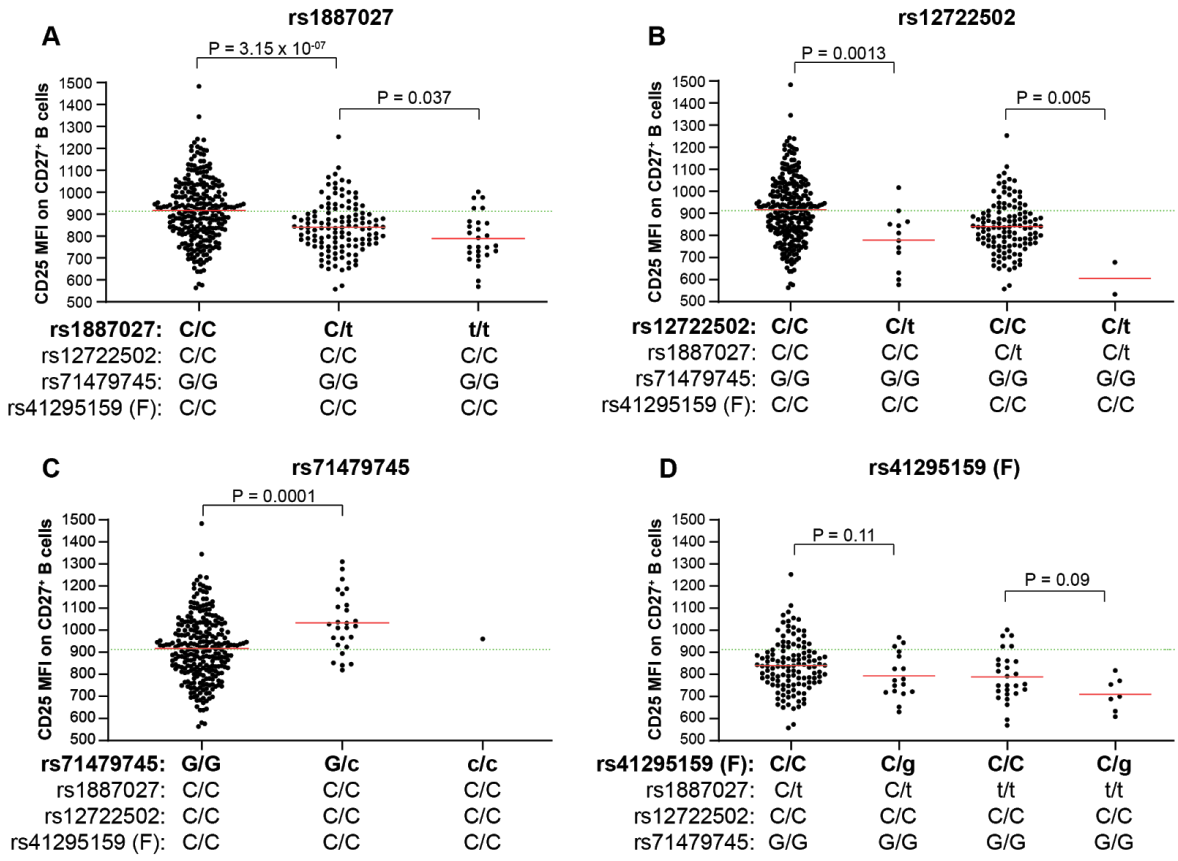


Figure 6.9. Variants that affect the CD25 MFI on CD27⁺ B cells. (A-D) Shows the expression of CD25 on CD27⁺ B cells, split by the SNPs that were identified as affecting the expression of CD25 on B cells, which are (A) rs1887027, (B) rs12722502, (C) rs71479745 and (D) the *IL2RA* F haplotype tagged by rs41295159. As each of these SNPs influence the expression of CD25 on B cells, the data was split so that each plot fixes the alleles not being assessed, and only the assessed allele splits into homozygous for the major alleles (capitals, eg. C/C), homozygous for the minor alleles (lower case, eg. t/t) or heterozygous (eg. C/t). The green dashed line represents the mean CD25 expression of the data set, which is an MFI of 912.9. The red line displays the mean of each dataset and the P values were calculated using a T-test.

6.4 The F haplotype is associated with protection from T1D and decreased levels of sCD25

I am including this section to tie up the F haplotype as data is spread over section 5.8, which explains the association with T1D, and section 6.2 that showed the F haplotype was also associated with decreased levels of sCD25 (Fig. 6.2).

The ASE results showed that the F haplotype produced less *IL2RA* mRNA in B cells *ex vivo* and also with activation, as well as less *IL2RA* mRNA in activated $CD14^+$ monocytes (Fig. 5.38). This monocyte phenotype is likely responsible for the reduced sCD25 protein levels in blood associated with the F haplotype. Through ASE I have shown that the effect of rs12722497, a SNP not associated with disease that causes a 20% increase in sCD25 protein in blood, was in activated $CD14^+$ monocytes (Fig. 6.4 and Fig. 6.7). I have also identified three SNPs that are associated with CD25 expression on B cells that are not associated with disease (Fig. 6.9). Therefore, combining these observations, I would conclude that the protection from T1D associated with the F haplotype is related to decreased expression of *IL2RA* by activated B cells, and decreased sCD25 levels *in vivo* coming from the $CD14^+$ monocytes.

6.5 Summary of non-disease related phenotypes

In this chapter I have shown that it is likely to be $CD14^+$ activated monocytes that are responsible for the increase in sCD25 associated with rs12722497, as well as the decreased sCD25 related to the F haplotype. Whilst undertaking this research, I discovered that

CD25 expression on B cells was also under genetic control from at least three variants. In Table 6.1 I have summarised the phenotypes identified.

Table 6.1. Summary of the non-disease associated phenotypes detected. The SNP that tags each phenotype, the allele assessed, a description of the phenotypes and the affected cell type are listed. The minor allele frequency (MAF) for these variants was calculated from the genomAD browser[224] for the European non-Finnish population.

Associated SNP	Allele assessed	MAF %	Phenotype / cell type
rs12722497	A	9.9	Increases sCD25 from activated monocytes
rs1887027	T	21.3	Decreases CD25 expression on B cells *ex vivo*
rs12722502	T	1.2	Decreases CD25 expression on B cells *ex vivo*
rs71479745	C	4.9	Increases CD25 expression on B cells *ex vivo*
F rs41295159	G	0.73	Decreases expression of CD25 on B cells *ex vivo* and with activation Decreases sCD25 from activated monocytes

6.6 Discussion

The results in this Chapter have been particularly informative and have clearly underlined the fact that not all phenotypes detected are related to the disease you are studying. Here I have shown that the expression of CD25 on B cells as well as the production of sCD25 are under genetic control, neither of which are associated with T1D, MS and AITD. This does not mean that these phenotypes are not associated with any disease. For example, in asthma, in which there is a strong aetiological link with B cells, the minor T allele of

rs12722502 (C>T), which is associated with a decrease in CD25 expression on B cells, was found to be associated with decreased risk of asthma[225]. Additionally, rs12722497 that causes an increase in the production of sCD25, is weakly associated with the incidence of streptococcus throat infections and tonsillectomy, with the minor A allele that correlates with increased sCD25 expression having a higher incidence of infection[226]. A GWAS on sCD25 levels in cardiovascular disease (CVD) identified rs7911500 in European Americans, which is in good LD with rs12722497 ($r^2 \sim 0.79$), reporting that the T allele of rs7911500 is associated with increased levels of sCD25 and increased risk for CVD traits[227]. Therefore in the right context, variants we have found to be associated with particular phenotypes but not the autoimmune diseases currently studied, could well be associated with other diseases.

Both the measurements of the sCD25 concentration and CD25 MFI on B cells were analysed using the stochastic search method[2] that has proved so powerful in analysing GWAS data. The same issues identified in Chapter 5 relating to the analysis of GWAS data also applies when trying to link quantitative traits to genetic data, and this novel application of the stochastic search method will perform better than stepwise, especially in a setting where there are multiple signals in the region, and stepwise risks identifying a SNP that jointly tags two or more associations.

The *IL2RA* region contains thousands of SNPs, and using information from the Regulome database, which provides a score of how likely it is located in a region of epigenetic interest, resulted in hundreds scoring as highly as the *IL2RA* disease associated variants.

Although the method I have developed to measure ASE is high-throughput, it is not scalable in order to test the effect of all these variants. There are methods that have tiled across a 178 Kb region that encompasses *IL2RA* to look for enhancers using CRISPR activation, which identified six stimulus-responsive enhancers, two of which overlapped with SNPs of the *IL2RA* A and D haplotypes[26] in Jurkat cells. An alternate approach is to use a high throughput reporter assay[228], in which millions of fragments of DNA are cloned into 3′ UTR of a reporter gene on the self-transcribing enhancer reporter vector, and fragments with enhancer activity can promote self-transcription and through RNA-Seq these transcribed fragments can be identified, mapped and quantified allowing the generation of enhancer maps[228]. Both of these techniques are powerful and unbiased approaches that would enable the identification of many enhancer elements, however these techniques have only been applied to cell lines, and the next step would be to move to primary cells, which could prove technically difficult as it requires transfection. The use of CRISPR-mediated gene editing on induced pluripotent stem cells (iPSC) to study the effect of specific variants has been tested using risk variants from Schizophrenia, and following successful editing the iPSCs have been differentiated to a neuronal cell type[57]. In this system it is also possible to test the combination of risk alleles, and for T1D that could be the IL-2 pathway to see if there is an additive effect of each risk allele.

The next Chapter will describe the method I have developed to visualise the *IL2RA* haplotype similarity that have been used extensively in my studies, and have helped to explain the relationship of rs12722497 and rs1887027 on the C haplotype.

Chapter 7

IL2RA Haplotype structure and history

7.1 Introduction and aims

This chapter describes the development and validation of a method to visualise haplotype similarity, some of which has been presented earlier in Chapter 5.2, when comparing the *IL2RA* haplotypes. Classically a GWAS study reports back what is the most associated variant within a region, but gives no context of how many other SNPs are in strong LD, or how similar the haplotypes carrying the disease protective and susceptible alleles are. This information is useful as it indicates how the haplotypes have evolved, and also for more complicated associations, like the *IL2RA* region, if the haplotype under study is relatively homogeneous.

Additionally, the majority of GWAS studies have focused on the European population, which can limit the ability to fine map disease variants due to linkage equilibrium. There is now a growing realisation that the genetic analyses of diverse populations improves discovery for complex traits[28,29], and as a consequence there are large studies being undertaken to redress the balance, and focusing on non-European ancestries. For example, the Trans-Omics for Precision Medicine (TopMed) study[229] contains 144,000 participants, of which 32% are of African ancestry, 16% Hispanic / Latino, 10% Asian and the All of Us Research Program containing 175,000 participants with more than 80% from groups

that have been historically underrepresented in biomedical research[230]. The GWAS data analysed in Chapter 5 for T1D, MS and AITD are from individuals who are self reported white British, and not from a diverse population. However, I am interested if the *IL2RA* A, C, D, E and F haplotypes are present in different parts of the world and at what frequency, meaning that if GWAS data were available, would the same variants even be present to influence disease or would different variants be influential in other populations? In reality, since T1D is relatively rare in people with non-European ancestries, such a GWAS likely would be small and underpowered, but understanding when variants within credible SNP sets have arisen in evolution and subjected to selective pressures can also inform on their candidacy as a causal SNP. At the start of Chapter 5, I presented a method to compare the different *IL2RA* haplotypes, and this could be used to compare the same haplotype across different populations to see if it is possible to chart the evolution of the haplotypes. However the method I developed has not been validated outside of the *IL2RA* region.

Therefore the aims of this chapter is to:

1. Develop a method to show how similar donors with the same *IL2RA* haplotype are to each other, and then compare how similar the different *IL2RA* haplotypes are to each other.
2. Validate this method using the genes *PTPN22* and *TYK2*.
3. Apply this method to the *IL2RA* region using the 1000 Genomes Project data with different ancestries and speculate which is the most likely causal variant for each haplotype.

7.2 Assessment of haplotype similarity

7.2.1 Using pre-developed methods

One method to compare the similarity of donors carrying the same *IL2RA* protective haplotypes to donors carrying a different protective or susceptible haplotype would be through a phylogenetic analysis. By entering the haplotype sequence from a few donors of each haplotype, a phylogenetic analysis can be computed to show how related individuals with the same protective haplotype are to each other and if they are related to any of the other protective or susceptible haplotypes, suggesting they could have arisen from a common ancestor. There are many software programs that perform phylogenetic tree analysis, and I have looked at PHYLIP[231], BEAST[110] and MEGA[111]. I have chosen to use MEGA because it is a well supported package, that has just been updated and has excellent documentation and performs the analysis I need.

To draw the phylogenetic tree I am using phased data from the European population of the 1000 Genomes Project, where I have selected the haplotypes I wish to analyse based on the A, C, D, E and F tag SNPs and am using 100 Kb of sequence centred around *IL2RA* (Chr 10: 6,040,000 - 6,140,000, build hg19).

Phylogenetic analysis shows that individuals that carry the A or D haplotype, respectively, cluster on the same branch of the phylogenetic tree as donors with the same haplotype, but are on a different branch to each other. This implies that there is little heterogeneity between all donors carrying either the A or D *IL2RA* haplotype. The majority of donors carrying the C haplotype or the C + F haplotype cluster together, however there are several

donors located on an unrelated branch suggesting that a recombination event might have occurred that has introduced a different haplotype outside of the core C protective SNPs. The E and susceptible haplotype cluster on several divergent branches of the tree suggesting that the E and susceptible groups do not contain a homogenous population, and that there are many 'flavours' of these haplotypes (Fig. 7.1).

This analysis method shows the relatedness of the haplotypes, but does not show information at the individual SNP level, which would be useful to see how many SNPs are shared between haplotypes. Additionally, the results of the phylogenetic tree analysis are dependent on the length of input sequence used. Here, 100 Kb of each haplotype was used to calculate the phylogenetic tree, which was chosen as it covers all the SNPs from groups A, C, D, E and only the most proximal group F SNPs. I used only the proximal group F SNPs in my analysis because the group F SNPs cover a region of 87 Kb and given that the F haplotype is rare and there is high LD between SNPs in the group, suggesting there has been little recombination between them, including all of the F SNPs would be unlikely to generate a better understanding of the relatedness of the haplotypes. In addition there are practical implications of using less DNA sequence. Analysis with less DNA input sequence was performed that contained a region spanning only the haplotype defining SNPs, and this showed that donors with the same haplotype clustered together and there was no divergence (data not shown).

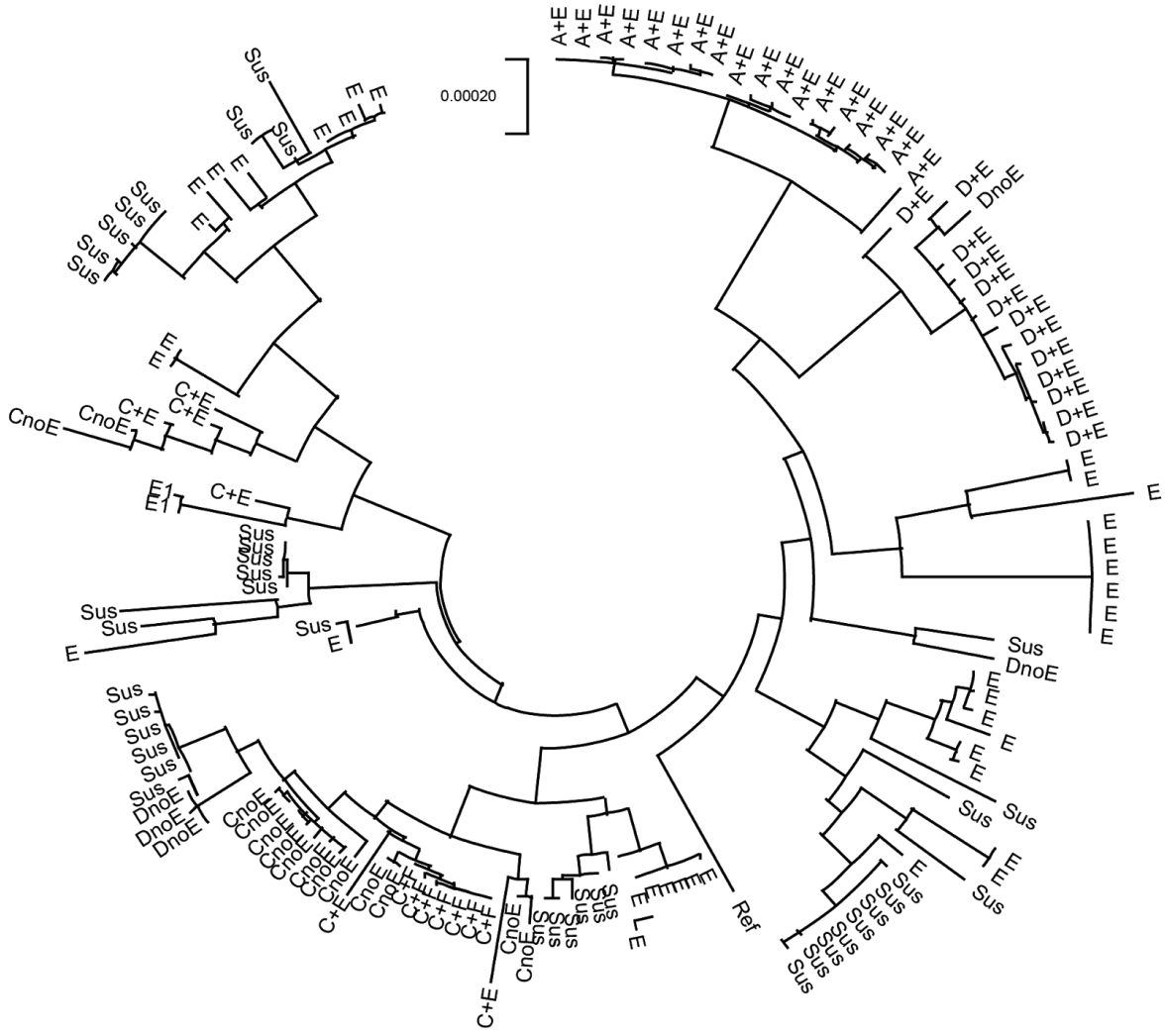


Figure 7.1. Phylogenetic analysis of the *IL2RA* region using Mega-X[111]. From the 1000 Genomes Project[105], 138 haplotypes of European ancestries were used to compute the phylogenetic analysis using Mega-X. Haplotypes located on the same branch are more closely related than haplotypes located on different branches of the phylogenetic tree.

7.2.2 Development of method to visualise haplotype similarity

The advantage of developing my own method to evaluate the similarity of donors carrying the same haplotype and between groups of individuals with different haplotypes is that I can make it specific to my requirements. The most important of which is to allow the

comparison of each SNP within the region, rather than a method that calculates haplotype divergence by averaging over a number of SNPs, along with a method to visualise the data, preferably with reference to genomic position. The UCSC genome browser[232] allows users to upload custom tracks of data in bed format and these can be layered on top of reference tracks, such as gene content, DNA conservation between species and epigenetic signals from the ENCODE[41,182] and Roadmap Epigenomics Program[40]. The data set used for this analysis is from the 1000 Genomes Project, which contains 5008 phased haplotypes from five different continental populations, each of which contain between four and seven regional populations (Chapter 2, Fig. 2.5).

I have selected individuals who carry all the protective alleles from each group of SNPs (A, C, D, E and F) and the susceptible haplotype is defined as carrying none of the A, C, D, E or F SNPs. Section 2.3.10 of the methods describes the method to calculate how similar donors with the same haplotype are to each other, as well as similarity between haplotypes. To ensure this method is robust I have looked at two genes which are associated with numerous autoimmune diseases, which are *PTPN22* and *TYK2*.

7.2.3 *PTPN22* as a proof of principal

PTPN22 encodes the protein tyrosine phosphatase LYP, a negative regulator of T- and B-cell antigen receptor signalling. A non-synonymous variant within *PTPN22* (rs2476601, R620W) has been associated with many autoimmune diseases[233]. In all studies the same variant (rs2476601) was identified, suggesting a simple one SNP model explains the association. This was a good setting to see how my haplotype visualisation would perform. I would expect that all haplotypes carrying the rs2476601 susceptibility

allele (A) to show a similar pattern to each other, and different from those carrying the protective (G) allele.

The analysis showed that indeed the haplotypes with the rs2476601 susceptible allele (A) were very similar to each other, with only 3 SNPs out of 5005 that were segregating (Fig. 7.2). In contrast, the 911 haplotypes carrying the rs2476601 G allele displayed a much more heterogeneous population of alleles with many more alleles segregating. However, upon closer examination I observed that within the bulk population of 911 haplotypes carrying the rs2476601 G allele were at least 4 different sub-haplotypes (Fig. 7.2). I compared the four sub-haplotypes to the one haplotype carrying the rs2476601 A allele and found that one showed remarkable similarity with only two SNPs (rs2476601 and rs6679677) having different alleles, whereas the other three rs2476601 G haplotypes were very different (Fig. 7.2). This analysis shows that the rs2476601 A allele probably arose once, as all donors carry the same combination of alleles. Comparison of these two haplotypes is like a 'natural' CRISPR experiment, with very few alleles differing apart from rs2476601, and if compared in the setting of functional experiments might reduce any background affects that other variants segregating within the population might introduce.

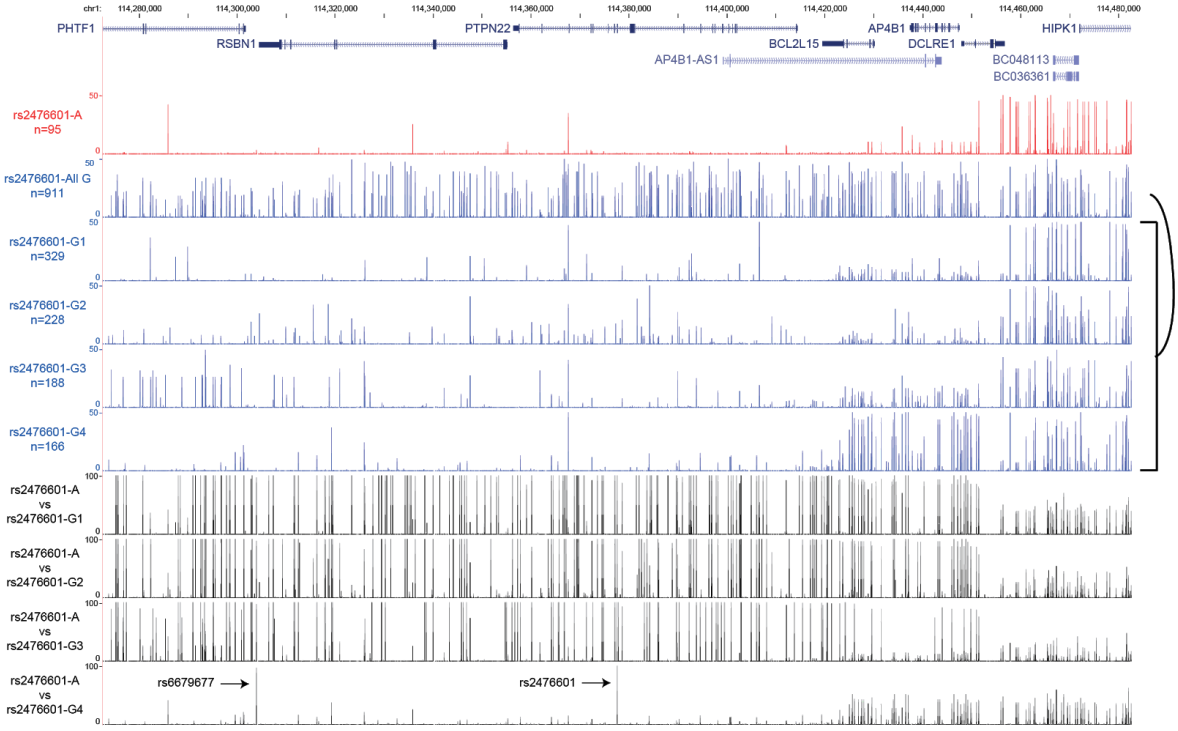


Figure 7.2. Visualisation of *PTPN22* haplotypes using 1006 phased haplotypes of European ancestries from the 1000 Genomes Project[105]. Individuals were selected based on rs2476601 genotype with 95 carrying the risk allele (A, line 1) allele and 911 the G allele (line 2). The 911 haplotypes were further divided into four haplotypes (lines 3-6). A score of zero means that at each allele all haplotypes have the same major allele, whereas the maximum score of 50 means there is an equal proportion of major and minor alleles present. Each of the four rs2476601 G haplotypes were compared to the haplotypes carrying the rs2476601 risk A allele (lines 7-10), where a score of 0 means the haplotypes being compared carry the same allele, whereas the maximum score of 100 means that the haplotypes being compared carry different alleles; one of the rs2476601 G haplotypes (line 6) showed almost identical distribution of alleles as the rs2476601 A haplotype (differences between line 1 and 6 displayed on line 10).

To provide further evidence that the SNP rs2476601 was responsible for the susceptibility to T1D, a haplotype analysis was performed by Jamie Inshaw (DIL) on a dataset of 6,734 T1D cases and 9,618 controls, using SNPs that define the *PTPN22* haplotypes in Fig. 7.3. Only the haplotype that carries the minor (A) allele of rs2476601 shows susceptibility to

T1D, and the haplotype named G4, which differs from the susceptible haplotype by two SNPs (rs2476601 and rs6679677) has no difference in T1D susceptibility to any of the other haplotypes carrying the major (G) allele at rs2476601 (Fig. 7.3). Therefore, the only consistent difference of alleles between the G1 and A haplotypes are the two SNPs rs2476601 and rs6679677, and one of these must be the variant that increases susceptibility to T1D.

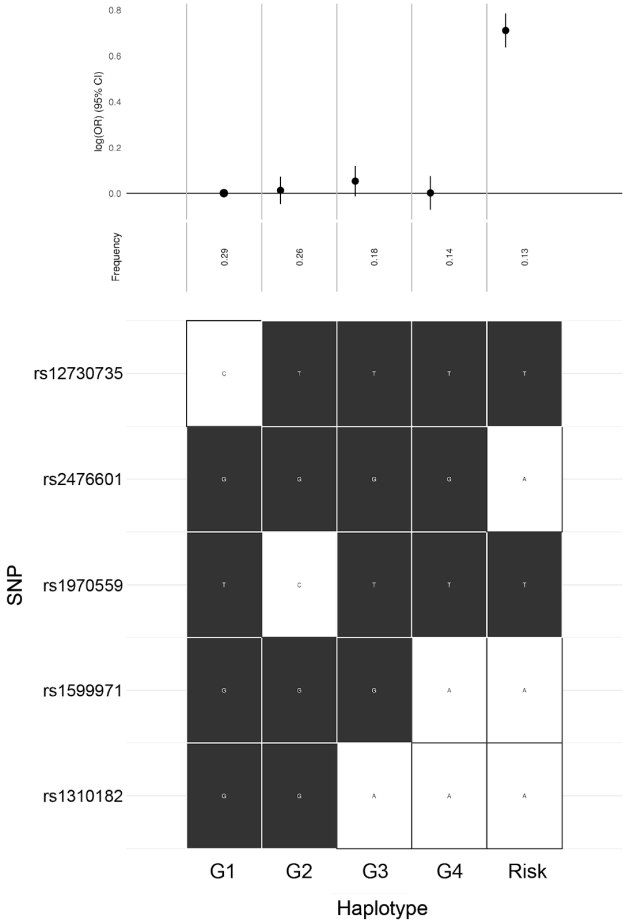


Figure 7.3. Haplotype analysis of T1D susceptibility in the *PTPN22* region. This analysis was performed by Jamie Inshaw (DIL). The most common haplotype is set as the reference haplotype and has a log OR of 0 and all haplotypes are compared to it. A log OR greater than 0 indicates susceptibility to T1D. Haplotype G1-G4 carry the major G allele at rs2476601, whereas the 'Risk' haplotype carries the minor A allele. The black filled boxes represent the major allele and the white boxes the minor allele.

This haplotype visualisation of *PTPN22* was useful, but the association at *PTPN22* was identified as the same variant for all autoimmune diseases. The next step was to investigate a more complex genetic association, for which I chose the gene *TYK2*.

7.2.3 *TYK2* as a proof of principal

The second gene region I investigated was *TYK2*, which is also associated with multiple autoimmune diseases, however the genetic association is more complex than *PTPTN22*, with three independent signals being identified[38]. I used the three disease associated variants from the *TYK2* region, rs12720356, rs34536443 and rs9797854[REF38] to divide the 1006 haplotypes of European acestries from the 1000 Genomes Project and visualise how similar donors with the same haplotype were to each other (Fig.7.4). Both haplotypes marked by the SNPs rs12720356 and rs34536443 showed a similar pattern of many segregating alleles across the region, apart from a 200 Kb window centered on *TYK2*, in which the degree of segregation decreased to very few alleles. However, when these two haplotypes were compared they showed the inverse pattern with the area outside of *TYK2* being more similar between haplotypes, and the region surrounding *TYK2* being more different. Even though these are both missense variants of TYK2, the two haplotypes are associated with a different array of autoimmune diseases, and only rs345536443 impairs the function of TYK2, and this analysis shows the two haplotypes are divergent at *TYK2* and must have arisen many generations ago. The haplotypes defined by the third SNP in the *TYK2* region, rs9797854, shows a very different pattern of alleles, with no lack of segregation at the *TYK2* gene, but at a more distal gene, ILF3, which makes sense as the most associated SNP is here.

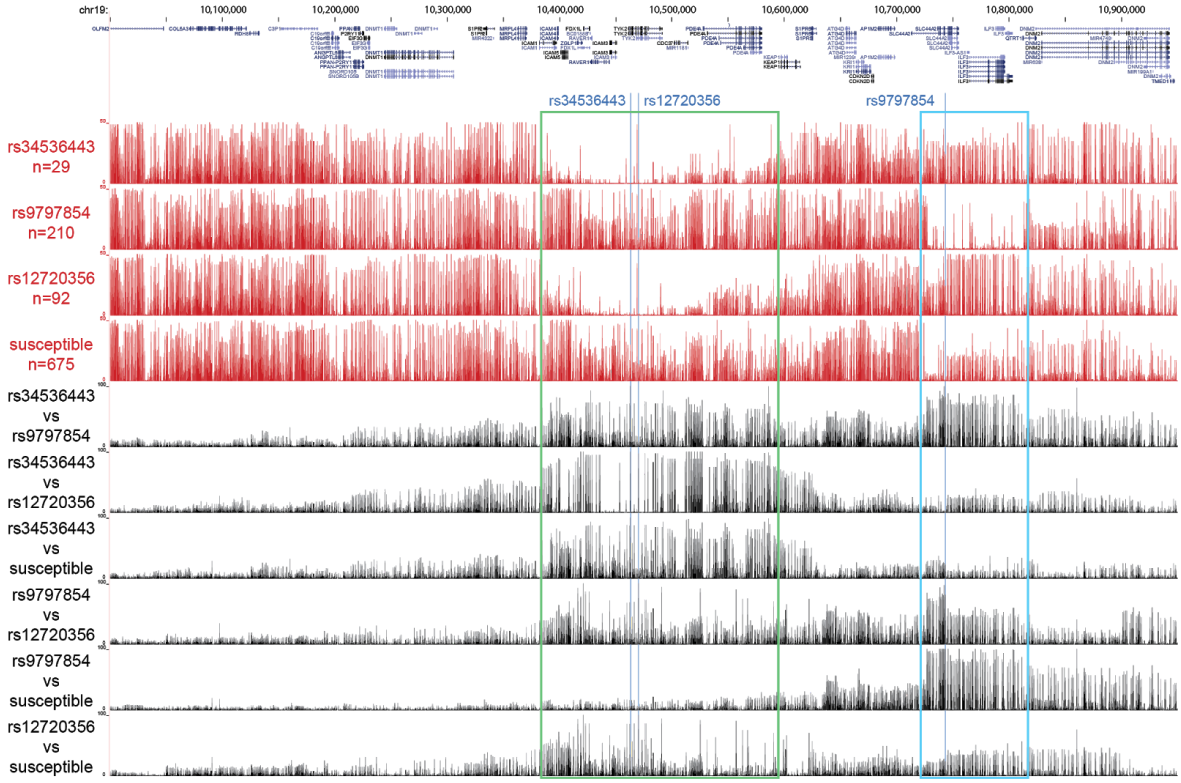


Figure 7.4. Visualisation of *TYK2* haplotypes using 1006 phased haplotypes of European ancestries from the 1000 Genomes Project[105]. Haplotypes were selected as carrying the minor allele at either rs12720356, rs34536443 or rs9797854[REF38] or as having no minor alleles at these three variants (susceptible). Within haplotype comparisons (red lines), a score of zero means that at each SNPall individuals with the same haplotypes have the same allele, whereas the maximum score of 50 means there is an equal proportion of major and minor alleles present. The four haplotypes were then compared to each other, and a score of 0 means the haplotypes being compared carry the same allele, whereas the maximum score of 100 means that the haplotypes being compared carry different alleles. The green box highlights a region around *TYK2*, where the haplotypes defined by rs34536443 and rs12720356 have similar alleles within this region, but are different when the four haplotypes are compared to one another. The blue box highlights the region around rs9797854 where the haplotype defined by the minor allele of rs9797854 is homozygous, and carries a different set of alleles to all other haplotypes being compared. 8905 SNPs were used in this analysis, 21,633 SNPs that are not polymorphic in the European population were excluded.

7.3 The *IL2RA* region

The haplotype visualisations of *PTPN22* and *TYK2* presented in Figure 7.2 and 7.4 validate the method I developed to visualise haplotype similarity. The method has higher resolution than that of the phylogenetic tree presented in Figure 7.1. Therefore this method of haplotype visualisation has been applied to the *IL2RA* A, C, D, E and F haplotypes, which has been discussed in Chapter 5 (Fig. 5.3). However this data used *IL2RA* haplotypes of European ancestries selected from the 1000 Genomes Project, as these are the ancestries on which the T1D, MS and AITD GWAS were performed.

Currently the vast majority of GWAS studies have been performed on people of European origin, and there has not been a large population GWAS performed on T1D, MS and AITD patients and controls of non-European ancestry. However, there is an increasing realisation that genetic analysis of divergent populations can improve the detection and mapping of complex traits[28,29]. Therefore I wondered if visualisation of the *IL2RA* haplotypes and allele frequencies from different world populations would provide useful information that could aid the identification of the likely causal variant for each haplotype.

7.3.1 The *IL2RA* A haplotype

The diversity seen in the different *IL2RA* haplotypes (Fig. 5.3 and 5.4) is fascinating, and my work has focused on individuals with European ancestries. To assess how the *IL2RA* haplotypes differ at a global population level, I have used the 1000 Genomes Project data from the whole genome sequencing of 2,504 individuals from five world regions with all

variants being identified[105]. From this resource I calculated the frequency of the A haplotype and determined that its maximal frequency is in the European population (5.17%) and it was not seen in any of the 504 people from East Asia (Fig. 7.5).

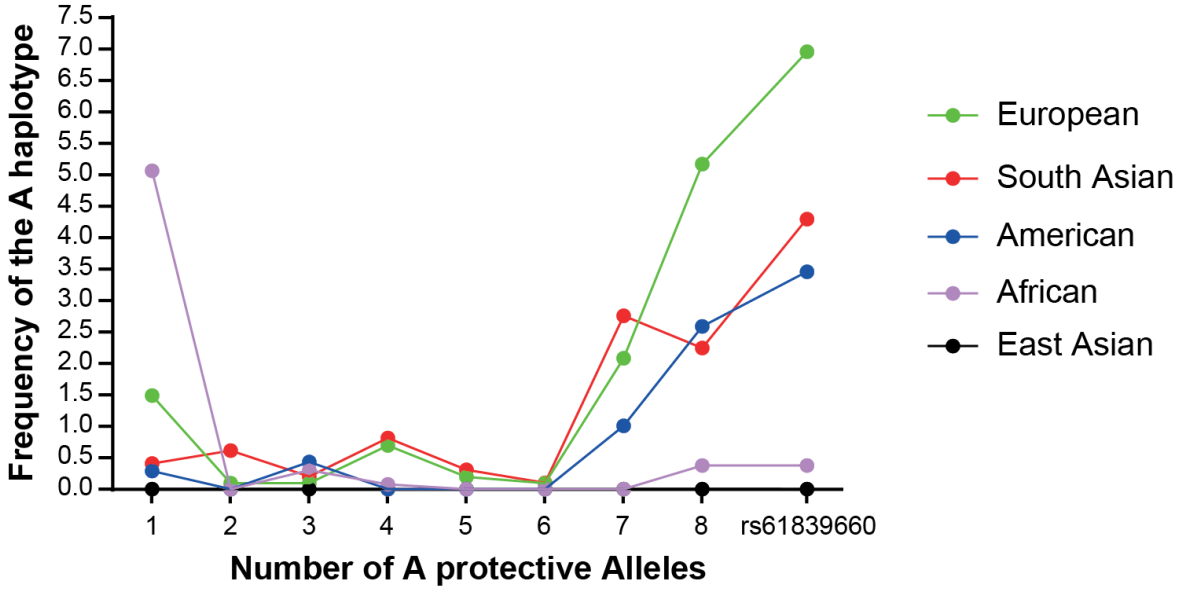


Figure 7.5. The frequency of the A haplotype in different continental populations. The 1000 Genomes Project genotype data were analysed to calculate the frequency of any 1, 2, 3, 4, 5, 6, 7 or 8 A haplotype SNPs (the full A haplotype contains 8 SNPs) in five world populations, with approximately 500 people per group. I have also shown the frequency of rs61839660, which I have proposed as the most likely causal variant of the A haplotype.

The A haplotype is not always found complete, but can be found as partial haplotypes in which only one or a few of the protective A haplotype alleles are found. Interestingly, the frequency that any one of the A protective alleles is found is highest in the African population (5.07%), but only 1.49% in the European population. Conversely, the full A 8 SNP haplotype is found at only 0.38% in the African population compared to 5.17% in

Europeans, suggesting that the full A haplotype was fixed relatively recently in evolutionary history, after humans had migrated out of Africa (Fig. 7.5).

The low frequency of the full A haplotype (0.38%) within the African population could be explained through admixture as the individuals with the full A haplotype are from the African Ancestry in Southwest US and African Caribbean in Barbados subpopulations. Additionally, the A haplotype is not the ancestral allele found in our closest primate relatives, suggesting the appearance of the protective haplotype in the European population must have been established and maintained through positive and or balancing selection. Only when the rs61839660 T allele appears on the A haplotype, does it become selected and the frequency of this haplotype increases, reinforcing the notion that it is the SNP rs61839660 that drives the protection associated with the A haplotype.

7.3.2 The *IL2RA* C haplotype

As discussed in Chapter 5.4 (Fig. 5.20), the *IL2RA* C haplotype can be found in combination with the solCD25 SNP (rs12722497) as well as the *IL2RA* E and F haplotypes, which appear to have arisen from recombination of these features onto the C haplotype as they are located at the proximal and distal ends of the *IL2RA* C haplotype, respectively. To simplify the analysis I have analysed the frequency of the *IL2RA* C haplotype regardless of the status of the solCD25 SNP (rs12722497), E and F haplotypes, as these will be discussed separately.

A group of three SNPs were identified by stochastic search as defining the *IL2RA* C haplotype, which is common within the European (29%), American (26%) and South Asian 20%) populations, whereas in the African (2.3%) and East Asian (3%) populations is much rarer (Fig. 7.6). The frequency of partial C haplotypes is low, with virtually no examples of 2 C SNPs, but in the African and South Asian population the presence of one C SNP is seen (Fig. 7.6). Interestingly, in both cases the one C SNP that is found alone is rs11597367, which in Chapter 5.4.2 I proposed as being the most likely causal variant for the C haplotype. This can be seen in Fig. 7.6 as the frequency of rs11597367 in the African and South Asian populations is higher than that of the three C SNP haplotype, whereas the frequency of rs11597367 and the three C SNP haplotype is the same in the European, American and South Asian populations.

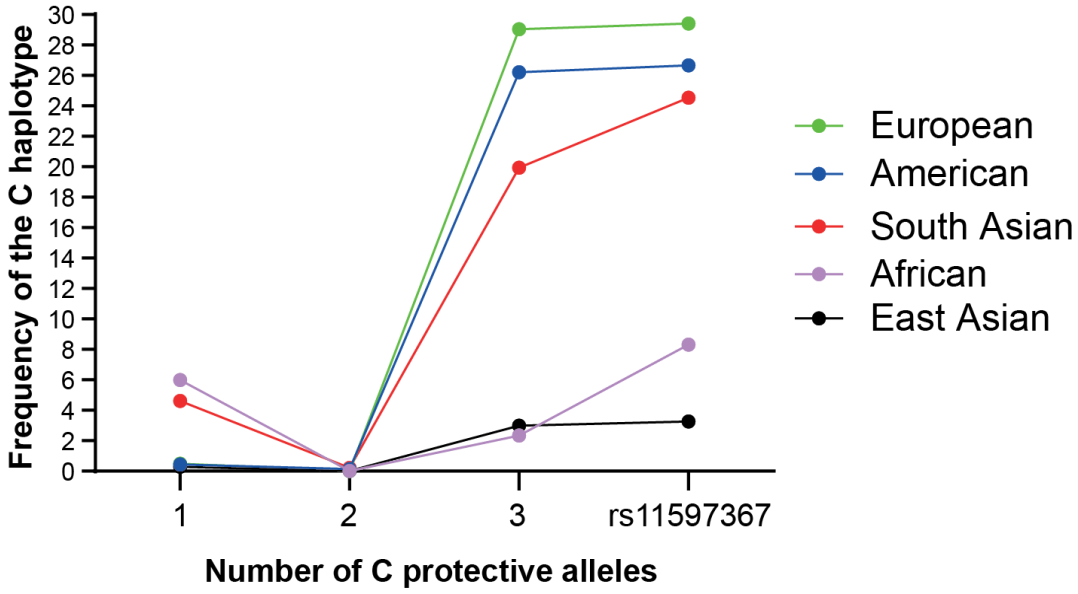


Figure 7.6. The frequency of the C haplotype in different continental populations from the 1000 Genomes Project[105]. Genotype data were analysed to calculate the frequency of any 1, 2 or 3 C haplotype SNPs (the full C haplotype contains 3 SNPs) in five world populations, with approximately 500 people per group. I have proposed that rs11597367 is the most likely causal variant for the C haplotype, therefore the frequency of this SNP is also displayed.

Could the C SNP, rs11597367, have been selected and maintained within divergent world populations as either a single SNP or as part of the wider *IL2RA* C haplotype, or has the single C SNP arisen through a recombination event of the full C haplotype? It is difficult to assess the question of selection, but I could compare the allele distribution of *IL2RA* haplotypes defined by just rs11597367 versus the full *IL2RA* C haplotype, and this would reveal if a recombination event had occurred, as there would be a break in the similarity of the allele frequency between the two haplotypes. As the single C haplotype occurs at a similar frequency in both the African and South Asian populations, I analysed both and compared to the European population.

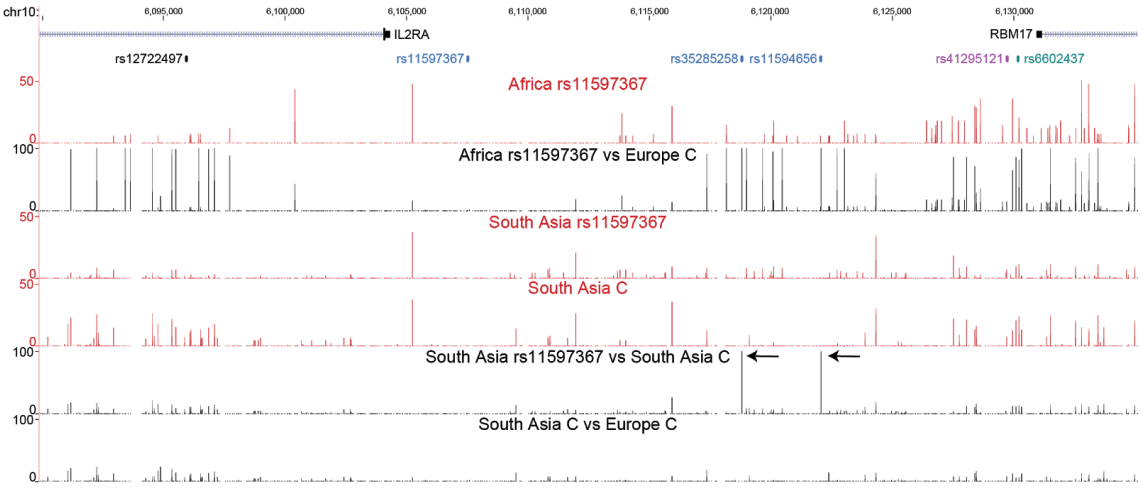


Figure 7.7. Comparison of *IL2RA* C haplotypes selected from the African, South Asian and European populations of the 1000 Genomes Project[105]. The red tracks display similarity of haplotypes from the same continental population and selected as either carrying just the C SNP rs11597367, or as carrying all three C SNPs; the minimum score of zero means all individuals carry the same allele at that particular SNP, whereas the maximum score of 50 means there is an equal proportion of the two alleles. The black tracks display the difference in haplotype allele frequency between the two haplotypes being compared; the minimum score of zero means both haplotypes carry the same allele, whereas the maximum score of 100 means the two haplotypes being compared carry the opposite alleles. The SNPs shown in blue are the three C SNPs, and the other SNPs show the soluble, E and one of the F SNPs.

From the haplotype visualisation presented in Fig. 7.7, it is clear that the single C SNP, rs11597367, has been defining the C haplotype for a long time as the distribution of alleles is very similar between the one C SNP and that of the full C haplotype in the African, South Asian and European population. It appears that the C SNP, rs11597367, is the original variant that defines the C haplotype, and then more recently in evolutionary time, the other two C SNPs, rs35285258 and rs11594656, have appeared on the C haplotype. This is particularly clear when comparing haplotypes from South Asia as the only two SNPs that differ between the one C haplotype and the full C haplotype is the two other group C SNPs (black arrows Fig. 7.7). Probably through a bottleneck event the three C haplotype has become the dominant haplotype that we see at higher frequencies in the American, European and South Asian population. There is no evidence that a recombination event caused the loss of the two C SNPs, but rather the single C haplotype appears to be the more ancient, and therefore, for the rs11597367 allele to be maintained within the population provides evidence that it has some selective advantage. A recent study has estimated the age of human genetic variants[197] (Genealogical Estimation of Variant Age, GEVA) at 45 million SNPs and the data is available in the Atlas of Variant Age (https://human.genome.dating). This resource confirms my observations that rs11597367 is the more ancient variant, estimated to have arisen 4,818 generations ago, whereas the other two SNPs, rs35285258 and rs11594656 are 'younger' and arose 2,075 and 3,519 generations ago, respectively.

7.3.3 The *IL2RA* D haplotype

The *IL2RA* D haplotype, which is protective for MS, was identified by stochastic search and defined by a group of 7 SNPs in high LD. The most common form of the *IL2RA* D

haplotype has the SNP rs2104286 (called B, identified by stepwise regression as the lead variant in the region for MS[REF]), located at the proximal end, and the *IL2RA* E SNP at the distal end in the European population (Fig. 7.8). As the haplotype visualisation on the *IL2RA* A and C haplotypes has yielded interesting results, I performed the same analysis looking at the frequency of the D haplotype in the five continetal populations from the 1000 Genomes Project[105], as well as the frequency of any single D allele or the frequency of rs62626317 that I proposed as the most likely causal variant for the D haplotype (Fig. 7.8).

The frequency of the full D haplotype that is defined by the 7 SNPs is highest in Europe (16%), South Asia (14%) and East Asia (13%), but lower in America (6%) and Africa (1%) (Fig. 7.8). This is the frequency regardless of the status of the *IL2RA* B and E SNPs that are located at the proximal and distal ends, respectively, of the *IL2RA* D haplotype. When the frequency of the haplotype is considered in the context of the presence of the *IL2RA* B and E SNPs, it is considerably higher than the *IL2RA* D haplotype that is missing the B or E SNP or both, suggesting that over time recombination has occurred. This pattern is seen in the American, European, East Asian and South Asian population. In the African population the frequency of the *IL2RA* D haplotype is low (1%), and very few examples of the *IL2RA* D haplotype without the B or E SNP are seen in the 1000 Genomes Project[105].

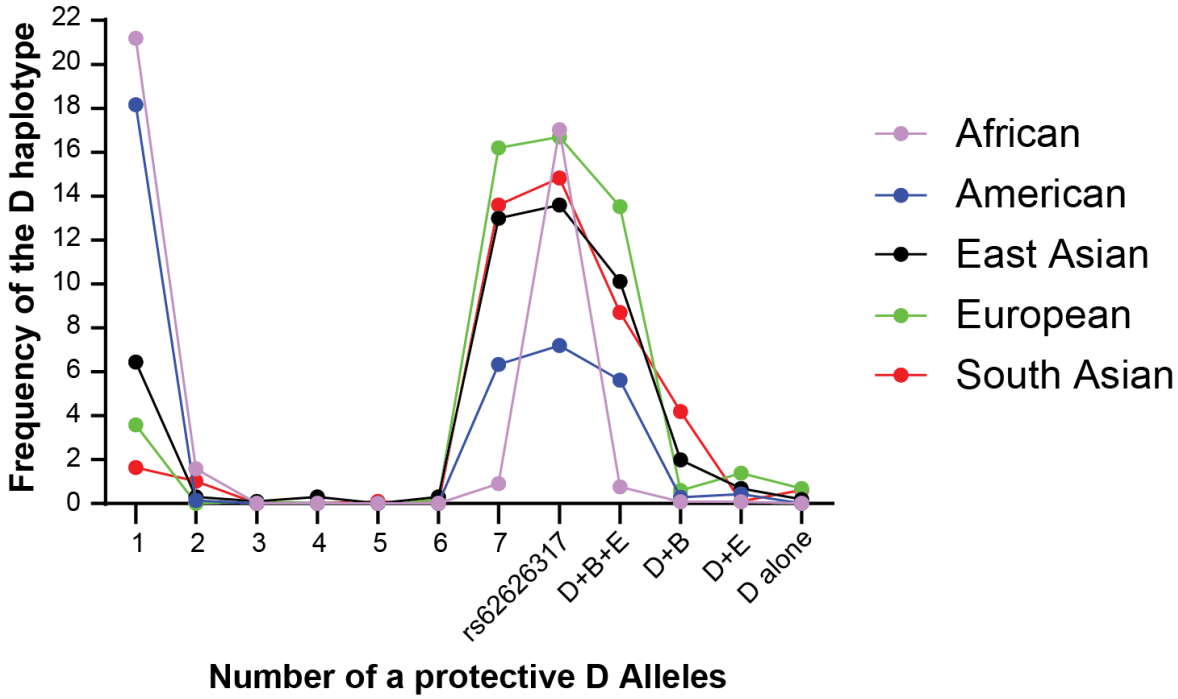


Figure 7.8. The frequency of the *IL2RA* D haplotype in different continental populations from the 1000 Genomes Project[105]. Genotype data were analysed to calculate the frequency of any 1, 2, 3, 4, 5, 6 or 7 D haplotype SNPs (the full D haplotype contains 7 SNPs) in five world populations, with approximately 500 people per group. The D haplotype is found primarily together with E (and the previously identified MS SNP, B), as well as some rare examples of D with and without combinations of B and E. I have proposed that rs62626317 is the most likely causal variant for the D haplotype, therefore the frequency of this SNP is also displayed.

Interestingly, like the A and C haplotypes, the presence of partial D haplotypes that contain between 2 and 6 of the D protective alleles are extremely rare. However, the presence of one D SNP is common in both the African and American population, however, it is not the same single D SNP which dominates, with rs62626317 being found most common in Africa and rs56382813 in America. I have performed a similar analysis as done for the C

haplotype, and looked to see if any of the one D SNP haplotypes seen in Africa, bear any similarity to that of the full D haplotype seen in Europe.

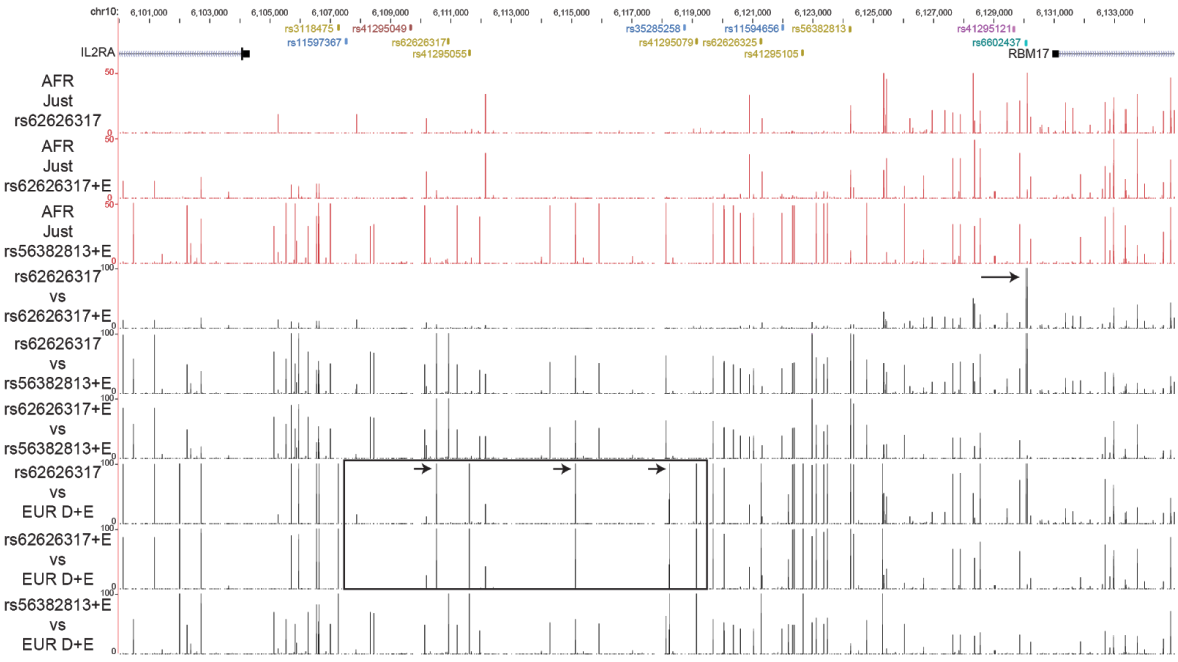


Figure 7.9. Comparison of single D SNPs in Africa to Europe D+E, using data from the 1000 Genomes Project. In Africa (AFR), a haplotype containing either the D SNP, rs62626317 or rs56382813 is common (Just rs62626317 n=105, just rs62626317+E n=91, just rs56382813+E n=65) and the frequency of other SNPs on these haplotype was calculated (red) and compared to each other (black). The haplotypes containing just rs62626317, with and without the E allele are remarkably similar to each other, apart from at the E allele (black arrow). The haplotypes containing rs62626317 and rs56382813 are very different from each other, whereas the rs62626317 containing haplotype is more similar to the complete D haplotype seen in Europe, with only three variants differing across a 12 Kb region (black box and black arrows).

Initial analysis revealed that the single D SNP haplotype defined by rs62626317 in Africa occurred half the time in combination with the E SNP. Therefore I chose to look at how similar donors carrying these haplotypes were to themselves and also to each other (Fig.

7.9). The seven D haplotype SNPs span a region of 18 Kb, and for both the rs62626317 haplotype, with and without E, show a homogeneous distribution of alleles, and when these haplotypes are compared the only difference is at the E SNP and a second closely located SNP. The D haplotype defined by the single SNP rs56382813 carries the E haplotype, but many of the SNPs are segregating, suggesting there are at least two different haplotypes contained within the definition of rs56382813 alone. Comparing this haplotype to the rs6262317 haplotype shows that many of the alleles are different, but a proportion of the rs56382813 haplotype does share alleles, this is shown by the difference track having scores of 70, rather than 100 if different alleles were present in each haplotype. With a larger dataset, it might be possible to split the rs56382813 haplotype and identify a group which shares more homology.

Next I looked to see how similar the rs62626317 haplotype was to the full D haplotype that is common in Europe, and is the rs62626317 haplotype seen in Africa the first building block of the D haplotype. Across an 11 Kb region, there are only three non-D SNPs seen in addition to three D SNPs that are not present in the Euorpean population. This suggests that indeed the rs62626317 haplotype in Africa is the predecessor of the D haplotype seen in Europe.

There are 12 examples of the full D haplotype seen in the African population from the 1000 Genomes Project, ten of these are from either the African Caribbean or Americans of

African Ancestry, suggesting these could have resulted from admixture as the D haplotype is more common in the American population.

7.3.4 The *IL2RA* E haplotype

The *IL2RA* E haplotype, defined by rs6602437 is different to the *IL2RA* A, C, D and F haplotypes in a number of ways. The first is that stochastic search identified only a single SNP, rs6602437, that defines the association with T1D, whereas the A, C, D and F haplotypes are groups of between 3 and 13 SNPs. This implies that rs6602437 is the causal variant for the E haplotype. Secondly, out of all the SNPs identified by stochastic search in the *IL2RA* A, C, D, E and F SNP groups (a total of 32 SNPs) only in the case of the E SNP, rs6602437, is the protective allele also the ancestral allele. This means that the protective T allele is evolutionarily very old and is shared with most mammalian species that have been sequenced[234]. Subsequently, the frequency of the E haplotype is much higher than all of the other *IL2RA* protective haplotypes and global frequency for the E haplotype is between 37-61% (Fig. 7.10).

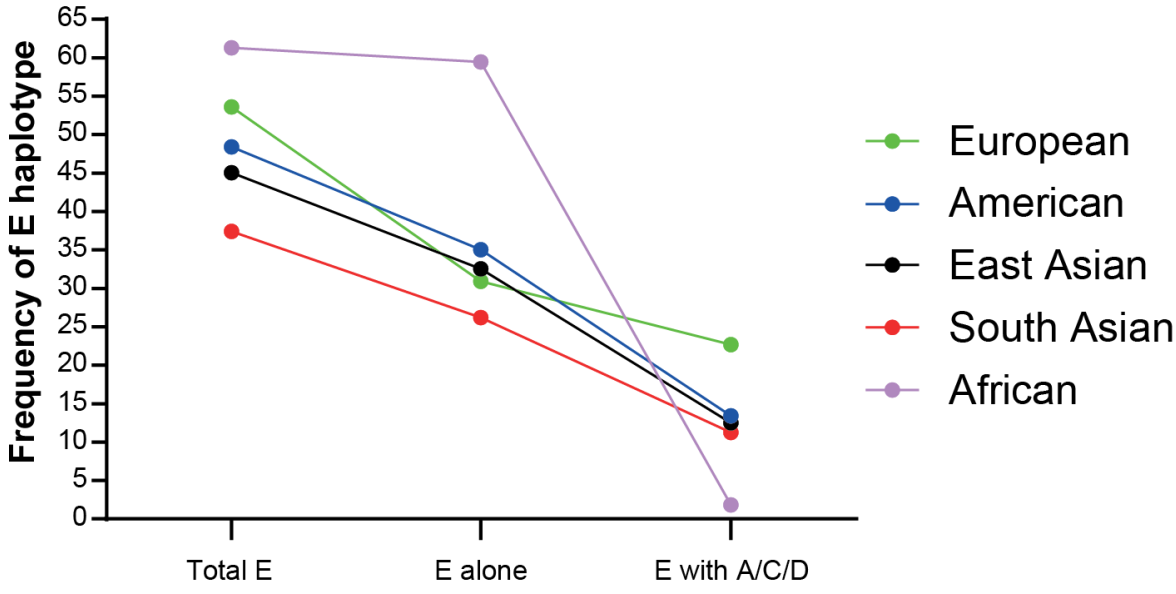


Figure 7.10. The frequency of the *IL2RA* E haplotype in different continental populations from the 1000 Genomes Project[105]. A single SNP, rs6602437, defines the E haplotype, which can be found in isolation or together with the A, C or D *IL2RA* haplotypes. Therefore the frequency of E is displayed regardless of the status of the *IL2RA* A, C and D haplotypes (Total E), when no *IL2RA* A, C or D haplotypes are present (E alone) or when E is found with the *IL2RA* A, C or D haplotypes (E with A/C/D).

The E SNP can be found alone or on the majority of *IL2RA* A and D haplotypes and approximately 10% of *IL2RA* C haplotypes. The E SNP is located distal to these haplotypes, and I presume through recombination has allowed for the E haplotype recombine between haplotypes. The frequency of the E haplotype alone is much higher in Africa, than the rest of the world, where the frequency of the A, C and D haplotypes is higher. As mentioned earlier in Chapter 5.7, the E SNP is located in a LINE repeat, and these have been associated with recombination hotspots, which might explain why we see the E SNP on so many of the *IL2RA* haplotypes.

I have already compared how similar donors carrying the E haplotype are to each other, as well as the A+E and D+E haplotype (Fig. 5.4), and have shown that the E haplotype is not a homogeneous collection of alleles, but rather is a composite of several haplotypes (probably seven common haplotype). However, in this chapter I am interested in how similar the region immediately surrounding the E SNP is between donors carrying the E SNP (that includes E, A+E and D+E) to haplotypes that carry the minor allele at E (susceptible and C haplotypes). In Fig. 7.11 I have centred the analysis of a 20 Kb region around the E SNP, and can see that the *IL2RA* A and D haplotypes carry the same sequence in a 1 Kb window surrounding the E SNP, which is very similar to that of the E haplotype. Whereas the C and susceptible haplotypes share a similar set of alleles that differ to that of the E containing haplotypes (Fig. 7.11). Interestingly, outside of this 1 Kb region, there is 1 Kb either side that contains no variation, which aligns with many transcription factor binding sites, suggesting this region is essential for gene regulation and any variation may be deleterious. The distal side of the E SNP also contains the promoter for the *RBM17* gene (Fig. 7.11).

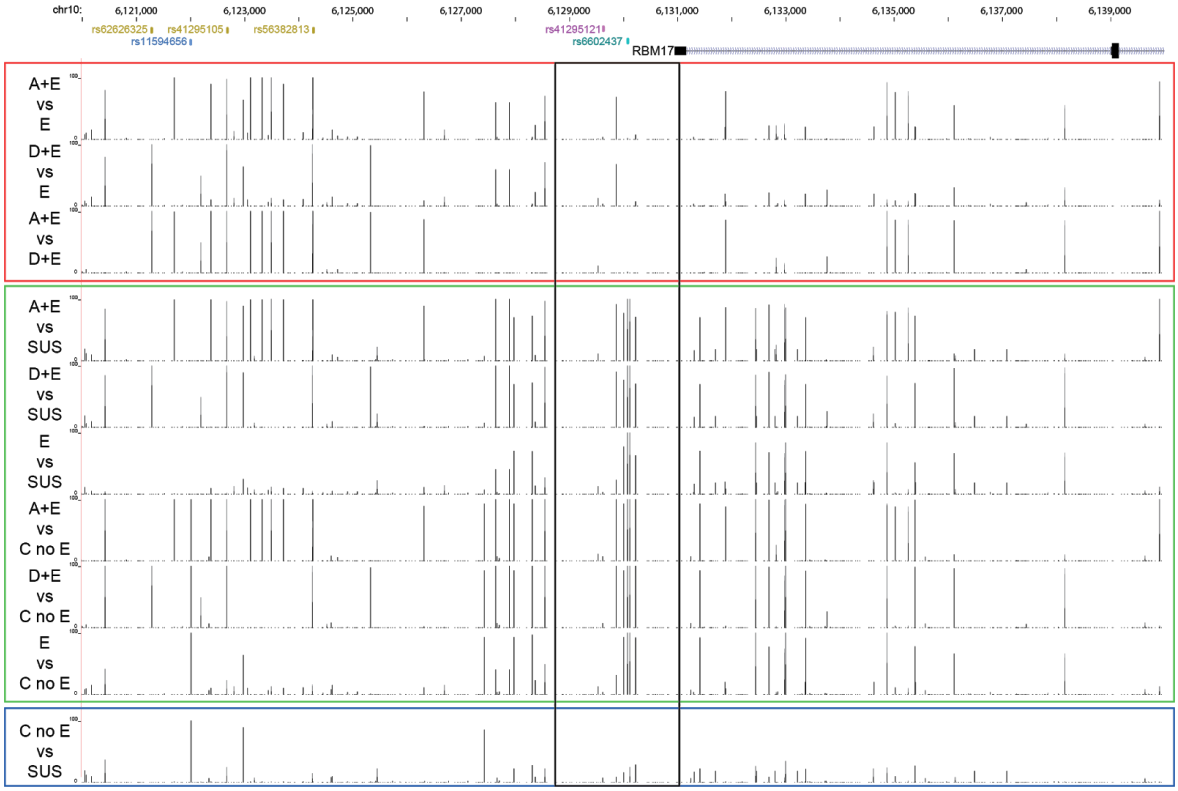


Figure 7.11. Comparison of the haplotypes that do (A, D and E) and do not (C and susceptible) carry the E SNP. The red box contains the comparison between haplotypes that have the E SNP. The green box compares haplotypes with the E SNP to those that do not have the E SNP. The blue box compares the C haplotype to the susceptible, neither of which carry the E SNP. The black box highlights the region around the E SNP where the haplotypes containing the E variant share a similar pattern on alleles, whereas the haplotypes that do not carry the E SNP share the same alleles, but these are different to those of the haplotypes carrying the E SNP. This analysis uses selected European haplotypes from the 1000 Genomes Project.

7.3.5 The *IL2RA* F haplotype

The F haplotype is the rarest of the *IL2RA* haplotypes I have studied (1.2%) and is located at the distal end of the *IL2RA* C haplotype in the European population. The stochastic search using GWAS data for T1D from the British population identified a group of 13 SNPs in high LD. As the *IL2RA* F haplotype is rare and there is limited recombination

between these SNPs, it has not been possible to identify which of the SNPs is the most likely causal variant.

Looking at the frequency of the *IL2RA* F haplotype in the 1000 Genomes Project[105] shows that the 13 SNP *IL2RA* F haplotype is only seen in the European population. In contrast, virtually no F alleles are seen in the African population, whereas there is a low frequency of partial F haplotypes seen in the American and East Asian population. The distribution of F alleles in the South Asian population is strikingly different from the rest of the world population, with the frequency of 2, 4, 5 and 6 F alleles, each being greater than 3.5% each, but no examples of the full F haplotype are seen (Fig. 7.12).

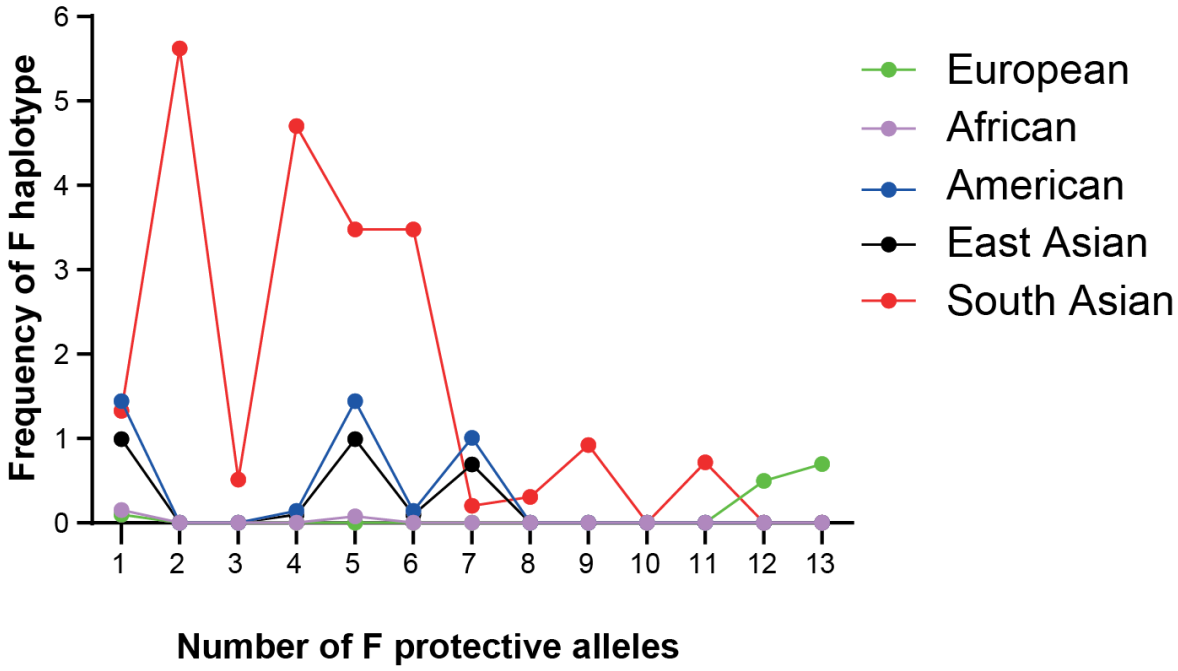


Figure 7.12. The frequency of the *IL2RA* F haplotype in different continental populations from the 1000 Genomes Project[105]. Genotype data were analysed to calculate the frequency of any 1, 2, 3….11, 12, or 13 F haplotype SNPs (the full F haplotype contains 13 SNPs) in five world populations, with approximately 500 people per group.

There are certain combinations of F alleles that are enriched in the South Asian population, units of 2, 4 or 5 F SNPs being most common. These units could either be the result of recombination breaking apart the haplotype or bringing the smaller haplotype units together to make longer F haplotypes (Table 7.1). As two of the F SNPs are unique to the European population (rs142414821 and rs76585197) in the 1000 Genomes Project[105], I would suggest that the partial F haplotypes seen in the South Asian population are not the result of recombination, but rather represent the early stages in the development of the F haplotype. In the large GNOMAD database of whole genome sequencing from 31,416 individuals[224], it is clear that these two variants have arisen most recently in time, as even within the European population there is enrichment of these two variants within the Northern and Southern Europeans compared to the Finnish and Estonian population (1% vs 0.1%).

Table 7.1. Distribution of *IL2RA* F alleles in the South Asian population, using genotype information from the 1000 Genomes Project[105]. Each combination of *IL2RA* group F SNP is listed with the right hand columns displaying the total number of F minor alleles and the number of occurrences. For each SNP, the major allele is shown in green and the minor allele is shown in red.

rs41295121	rs41295159	rs80250062	rs78938174	rs142414821	rs189193613	rs74765647	rs146135969	rs117678801	rs78815004	rs189961326	rs76585197	rs75252425	No. of F alleles	Count
0	1	0	0	0	0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	1	0	0	0	0	0	0	1	4
1	0	0	0	0	0	0	0	0	0	0	0	0	1	8
0	0	0	0	0	0	0	1	1	0	0	0	0	2	1
0	0	1	1	0	0	0	0	0	0	0	0	0	2	8
0	0	0	0	0	1	1	0	0	0	0	0	0	2	10
1	1	0	0	0	0	0	0	0	0	0	0	0	2	36
0	0	0	0	0	0	0	0	0	1	1	0	1	3	1
0	0	0	1	0	1	1	0	0	0	0	0	0	3	1
0	0	1	1	0	1	0	0	0	0	0	0	0	3	1
1	1	0	0	0	0	1	0	0	0	0	0	0	3	1
0	1	1	1	0	0	0	0	0	0	0	0	0	3	1
0	0	0	0	0	0	0	1	1	1	1	0	0	4	1
1	1	0	0	0	1	1	0	0	0	0	0	0	4	12
0	0	1	1	0	1	1	0	0	0	0	0	0	4	33
0	1	1	1	0	1	1	0	0	0	0	0	0	5	1
0	0	0	0	0	0	0	1	1	1	1	0	1	5	34
0	0	0	0	0	0	1	1	1	1	1	0	1	6	2
1	1	1	1	0	1	1	0	0	0	0	0	0	6	32
1	1	0	0	0	0	0	1	1	1	1	0	1	7	1
0	0	0	1	0	1	1	1	1	1	1	0	1	8	1
0	0	1	1	0	1	1	1	1	1	1	0	0	8	1
1	1	0	0	0	0	1	1	1	1	1	0	1	8	1
1	1	1	1	0	0	0	1	1	1	1	0	1	9	1
0	0	1	1	0	1	1	1	1	1	1	0	1	9	8
1	1	1	1	0	1	1	1	1	1	1	0	1	11	7

If there was GWAS data available for the South Asian population, the causal variant for the *IL2RA* F haplotype would be easier to identify as the LD structure between the F SNPs is broken, whereas in the European population, the 13 SNPs are inherited as a single block. As discussed in Chapter 5.8 and 6.4, the phenotype associated with the F haplotype has been identified, therefore, if individuals with the partial F haplotypes shown in Table 7.1 could be identified and studied to see if they display the decreased expression of CD25 on B cells associated with the *IL2RA* F haplotype, then the most likely causal variant for the F haplotype would be revealed.

7.4 Discussion

In this chapter I have described the method I have developed to visualise haplotypes selected by SNPs discovered through analysis of GWAS data. My method has the advantage over traditional phylogenetic trees in that you are seeing differences between haplotypes at the resolution of individual SNPs together with spatial information as the visualisation is mapped onto the human reference sequence. I have used *PTPN22* and *TYK2* regions that are associated with many autoimmune diseases to validate my method, and show the pattern of alleles segregating with the identified haplotypes was not restricted to the *IL2RA* region. For *PTPN22*, the SNP rs2476601 has been identified as being the most disease associated variant, therefore a simple association to visualise, whereas *TYK2* is a more complicated situation, with three disease associations that offer varying degrees of protection from autoimmune disease.

The idea of using divergent populations to aid the understanding of genetic association is not a new concept, and was used to map HLA-DQ in T1D[235]. However, in the last few years, there has been a growing interest in the analysis of both divergent populations and in understanding our ancestry and how this has shaped our genetic makeup. As noted in a recent editorial[236] "Diversity matters" and "will be an essential catalyst of research success". To this end I have used the 1000 Genomes Project phased genetic data to assess the ancestry of the *IL2RA* haplotypes, and show it is possible to trace the evolution of these haplotypes, and show certain haplotypes become selected once a certain protective allele appears. However, as larger and more diverse populations are sequenced, this could refine our understanding of how these alleles arose, and as Africa was the origin of the modern human population, a recent study of 44 indigenous populations will be a tremendous resource to mine[237].

The scientific community making data public and freely accessible has made a huge contribution to my thesis, and new information is being published continually. Whilst writing this discussion, a study that estimates the age of over 45 million genomic variants[197] was published (https://human.genome.dating) and I only had time to briefly include its use (Chapter 7.3.2), but such a method would be an alternate way to assess the order at which SNPs appeared in evolutionary time. A second study, used the 1000 Genomes project data to assess genome-wide genealogy[238] and identified multiple regions under strong positive selection and I would like to see if there is any overlap with these and the *IL2RA* region, or any of the regions associated with autoimmune disease.

From this chapter I have learnt a great deal about the *IL2RA* region, with insights into which alleles might have been subjected to selection. My studies highlight the importance of using all of the available resources when trying to gather evidence to determine which of the candidate SNPs are the most likely causal variants.

Chapter 8
General conclusions

At first glance, my thesis contained two quite distinct projects focused on the role of the IL-2 pathway in T1D: the first was the development and application of an assay to measure the *FOXP3* TSDR in Tregs, a cell type exquisitely dependent on IL-2 signaling and essential for regulating autoimmune effector cells in which the IL-2RA/CD25 molecule has an essential role, and the second was the analysis of phenotypes correlating with the genetic association of *IL2RA* in T1D, MS and AITD. However, there were themes shared between both projects that formed some of the key messages learnt from my research.

The method I developed to measure the methylation of the TSDR and allele-specific expression was pivotal to this thesis and during assay development and testing, it became apparent that understanding the qualities and limitations of these assays was key, and a substantial amount of time was spent in development. Going forward, I will need to perform more experiments to replicate the finding that the small overestimation of demethylation of the *FOXP3* TSDR in samples from females was caused by a bias in the DNA polymerase used, and if the results can be corrected via the use of a standard curve or an alternative DNA polymerase. However, this bias does not take away from the observation that the Treg population is not homogeneous when assessing TSDR demethylation. Using the combination of the FOXP3 and HELIOS transcription factors, it was possible to identify cells that contained a fully demethylated TSDR, which were all cells that co-expressed FOXP3 and HELIOS. Cells expressing FOXP3 but not HELIOS or

even cells lacking FOXP3 expression but located within the surface Treg gate displayed a mixed TSDR methylation profile, meaning there is a mixture of cells, those that contain a fully methylated TSDR and those with a demethylated TSDR. Cells that were $FOXP3^{-}$, but fully demethylated at the TSDR, did not produce cytokines, suggesting they are more similar to $FOXP3^{+}$ $HELIOS^{+}$ Tregs, and may represent a population that has downregulated FOXP3, and that I plan to follow up. As an intracellular stain is required to purify cells by FOXP3 and HELIOS expression, there are limited options for follow up, and the lab is currently analysing ATAC-Seq data from these purified populations to see which have a similar chromatin landscape. Ideally, a single cell ATAC-Seq experiment in which cells are purified based on the expression of FOXP3, HELIOS and cytokine production would resolve the relationship of these populations.

The IL-2 pathway is a major focus of the lab and we are working to fine-map the genetic association of the *IL2* and *IL2RA* regions, as well as conducting clinical trials to assess if administration of ultra-low-dose IL-2 will increase Tregs and preserve residual insulin production relative to placebo controls. The measurement of the *FOXP3* TSDR was used to show that the cells that have expanded in response to exogenous IL-2 are *bona fide* Tregs, as they possess a fully demethylated TSDR, and are not a consequence of activated T effector cells transiently upregulating FOXP3.

Understanding the genetics of the *IL2RA* region has been a challenge, and has involved many members of the lab, from the initial discovery of the association of *IL2RA* with T1D through a candidate gene approach 15 years ago[168]. Collaboration with Dr Chris Wallace has been instrumental in dissecting the association, and in particular the use of haplotype

analysis to explain why SNPs identified through traditional step-wise analysis of the GWAS data did not agree with the variants identified through stochastic search, but were actually jointly tagging two or more independent associations. Where there are multiple SNPs within a region with similar effect size and direction, then the stochastic search method more accurately identifies the variants that are most associated with disease in a primary way, rather than being tags for a number of haplotypes. We propose that the combination of stochastic search and haplotype association analysis is critical before embarking on genotype-to-phenotype studies. Even with the correct disease-associated variants identified, another lesson learnt from the *IL2RA* region, is that not all phenotypes are associated with disease, with phenotypes relating to the expression of CD25 on B cells and sCD25 illustrating the point. I have successfully identified *IL2RA* mRNA and in some cases CD25 protein expression phenotypes for the *IL2RA* C and F haplotypes as well as confirming and extending the previously identified phenotypes for the A and D haplotypes. A phenotype for the E haplotype was not identified; this haplotype has the weakest protective effect for T1D, and therefore the phenotype may be beyond the sensitivity of the assays used, or in a cell type or stimulation condition not tested, or indeed could alter expression of another gene not assessed.

Overall I can conclude, that not all Tregs are Tregs if a demethylated *FOXP3* TSDR is the 'gold standard' to define a Treg. There are TSDR methylated cells which are $FOXP3^+$ and TSDR demethylated cells that are $FOXP3^-$. Although these cells represent a minority of Tregs in the periphery, it highlights that the Treg population is heterogeneous. With clinical trials being conducted to either boost Treg frequency *in vivo*, or to infuse *ex vivo* expanded Tregs, a better characterisation of these populations and their interrelationship

should be a priority. Second, the study of expression phenotypes correlated with disease gene association is complex and in regions with more than one causal variant, such as *IL2RA*, the focus of my thesis, I have shown that a stochastic search method for fine mapping is essential for the interpretation of the mRNA and protein expression data. The methods used to define the *IL2RA* region are applicable to the fine mapping of any region and the design of expression studies. The methods developed also have the potential to point to causal variants as detailed in the thesis. Nevertheless, as technologies develop and methods such as single-cell eQTL and ATAC-Seq, and high-throughput CRISPR screens in primary cells become more commonplace, this should facilitate unambiguous identification of causal variants. Such screens together with the utilisation of the wealth of publicly available epigenetic, transcriptomic and human DNA sequence resources will enable the identification of causal variants and potential mechanisms that lead to the development of autoimmune disease, with the ultimate aim of identifying therapeutic targets.

Appendix 1: Publications with data from my thesis

- **Chromosome contacts in activated T cells identify autoimmune disease candidate genes.**
  Oliver S. Burren, Arcadio Rubio García, Biola-Maria Javierre, Daniel B. Rainbow, Jonathan Cairns, Nicholas J. Cooper, John J. Lambourne, Ellen Schofield, Xaquin Castro Dopico, Ricardo C. Ferreira, Richard Coulson, Frances Burden, Sophia P. Rowlston, Kate Downes, Steven W. Wingett, Mattia Frontini, Willem H. Ouwehand, Peter Fraser, Mikhail Spivakov, John A. Todd, Linda S. Wicker, Antony J. Cutler and Chris Wallace. Genome Biology. 2017.

- **Regulatory T Cell Responses in Participants with Type 1 Diabetes after a Single Dose of Interleukin-2: A Non-Randomised, Open Label, Adaptive Dose-Finding Trial.**
  Todd JA, Evangelou M, Cutler AJ, Pekalski ML, Walker NM, Stevens HE, Porter L, Smyth DJ, Rainbow DB, Ferreira RC, Esposito L, Hunter KM, Loudon K, Irons K, Yang JH, Bell CJ, Schuilenburg H, Heywood J, Challis B, Neupane S, Clarke P, Coleman G, Dawson S, Goymer D, Anselmiova K, Kennet J, Brown J, Caddy SL, Lu J, Greatorex J, Goodfellow I, Wallace C, Tree TI, Evans M, Mander AP, Bond S, Wicker LS, Waldron-Lynch F. PLoS Med. 2016.

- **Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases.**
  Asimit JL, Rainbow DB, Fortune MD, Grinberg NF, Wicker LS, Wallace C. Nat Commun. 2019.

- **Cells with Treg-specific *FOXP3* demethylation but low CD25 are prevalent in autoimmunity.**
  Ferreira RC, Simons HZ, Thompson WS, Rainbow DB, Yang X, Cutler AJ, Oliveira J, Castro Dopico X, Smyth DJ, Savinykh N, Mashar M, Vyse TJ, Dunger DB, Baxendale H, Chandra A, Wallace C, Todd JA, Wicker LS, Pekalski ML. J Autoimmun. 2017.

- **Chronic immune activation in systemic lupus erythematosus and the autoimmune PTPN22 Trp620 risk allele drive the expansion of FOXP3+ regulatory T cells and PD-1 expression.**
  Ricardo Ferreira, Xaquin Castro Dopico, Joao Oliveira, Daniel Rainbow, Jennie Hsiu Mien Yang, Sarah Todd, Mhairi McNeill, Maristella Steri, Valeria Orru, Edoardo Fiorillo, Daniel Crouch, Marcin Pekalski, Francesco Cucca, Timothy Ian Martin Tree, Timothy Vyse, Linda Wicker, John Todd. Front Immunol. 2019; 10: 2606.

- **Human IL-6RhiTIGIT- CD4+CD127lowCD25+ T cells display potent in vitro suppressive capacity and a distinct Th17 profile.**
  Ferreira RC, Rainbow DB, Rubio García A, Pekalski ML, Porter L, Oliveira JJ, Waldron-Lynch F, Wicker LS, Todd JA. Clin Immunol. 2017.

Preprint publication

- **A rare *IL2RA* haplotype identifies SNP rs61839660 as causal for autoimmunity**
  Daniel B. Rainbow, Marcin Pekalski, Antony J. Cutler, Oliver Burren, Neil Walker, John A. Todd, Chris Wallace, Linda S. Wicker. bioRxiv 108126; doi: https://doi.org/10.1101/108126.

Not included in my thesis

- **A Novel *Pkhd1* Mutation Interacts with the Nonobese Diabetic Genetic Background To Cause Autoimmune Cholangitis.**
  Huang W, Rainbow DB, Wu Y, Adams D, Shivakumar P, Kottyan L, Karns R, Aronow B, Bezerra J, Gershwin ME, Peterson LB, Wicker LS, Ridgway WM. J Immunol. 2018.

- **Genetic and functional data identifying *Cd101* as a type 1 diabetes (T1D) susceptibility gene in nonobese diabetic (NOD) mice.**
  Mattner J, Mohammed JP, Fusakio ME, Giessler C, Hackstein CP, Opoka R, Wrage M, Schey R, Clark J, Fraser HI, Rainbow DB, Wicker LS. PLoS Genet. 2019

- **The plasma biomarker soluble SIGLEC-1 is associated with the type I interferon transcriptional signature, ethnic background and renal disease in systemic lupus erythematosus.**
  Oliveira JJ, Karrar S, Rainbow DB, Pinder CL, Clarke P, Rubio García A, Al-Assar O, Burling K, Morris S, Stratton R, Vyse TJ, Wicker LS, Todd JA, Ferreira RC. Arthritis Res Ther. 2018.

- **Neonatal and adult recent thymic emigrants produce IL-8 and express complement receptors CR1 and CR2.**
  Pekalski ML, García AR, Ferreira RC, Rainbow DB, Smyth DJ, Mashar M, Brady J, Savinykh N, Dopico XC, Mahmood S, Duley S, Stevens HE, Walker NM, Cutler AJ, Waldron-Lynch F, Dunger DB, Shannon-Lowe C, Coles AJ, Jones JL, Wallace C, Todd JA, Wicker LS. JCI Insight. 2017

- **In-depth immunophenotyping data of IL-6R on the human peripheral regulatory T cell (Treg) compartment.**
  Ferreira RC, Rainbow DB, Rubio García A, Pekalski ML, Porter L, Oliveira JJ, Waldron-Lynch F, Wicker LS, Todd JA. Data Brief. 2017.

In preparation

- **Using the *FOXP3* TSDR to assess Treg heterogeneity**
  Essentially Chapter 4 of my thesis

- **The genetic association of *IL2RA* with autoimmune diseases**
  Essentially Chapter 5 of my thesis

- **Variants that alter the expression of *IL2RA* not associated with disease**
  Essentially Chapter 6 of my thesis

- **Haplotype visualisation**
  Essentially Chapter 7 of my thesis

References

1. Rainbow, D. B., Esposito, L., Howlett, S. K., Hunter, K. M., Todd, J. A., Peterson, L. B. & Wicker, L. S. Commonality in the genetic control of type 1 diabetes in humans and NOD mice: Variants of genes in the IL-2 pathway are associated with autoimmune diabetes in both species. *Biochem. Soc. Trans.* **36,** 312–315 (2008).
2. Wallace, C., Cutler, A. J., Pontikos, N., Pekalski, M. L., Burren, O. S., Cooper, J. D., García, A. R., Ferreira, R. C., Guo, H., Walker, N. M., Smyth, D. J., Rich, S. S., Onengut-Gumuscu, S., Sawcer, S. J., Ban, M., Richardson, S., Todd, J. A. & Wicker, L. S. Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping. *PLoS Genet.* **11,** e1005272 (2015).
3. Report reveals the rising rates of autoimmune conditions | British Society for Immunology. at <https://www.immunology.org/news/report-reveals-the-rising-rates-autoimmune-conditions>
4. Wang, L., Wang, F.-S. & Gershwin, M. E. Human autoimmune diseases: a comprehensive update. *J. Intern. Med.* **278,** 369–395 (2015).
5. Bogdanos, D. P., Smyk, D. S., Rigopoulou, E. I., Mytilinaiou, M. G., Heneghan, M. A., Selmi, C. & Eric Gershwin, M. Twin studies in autoimmune disease: Genetics, gender and environment. *J. Autoimmun.* **38,** J156–J169 (2012).
6. Cudworth, A. G. & Woodrow, J. C. HL-A System and Diabetes Mellitus. *Diabetes* **24,** 345–349 (1975).
7. Singal, D. P. & Blajchman, M. A. Histocompatibility (HL-A) Antigens, Lymphocytotoxic Antibodies and Tissue Antibodies in Patients with Diabetes

Mellitus. *Diabetes* **22,** 429–432 (1973).

8. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).

9. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467 (1977).

10. Mittal, K. K., Mickey, M. R., Singal, D. P. & Terasaki, P. I. Serotyping for homotransplantation XVIII. Refinement of microdroplet lymphocyte cytotoxicity test. *Transplantation* **6,** 913 (1968).

11. Gough, S. C. L. & Simmonds, M. J. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Curr. Genomics* **8,** 453–465 (2007).

12. Noble, J. A. & Valdes, A. M. Genetics of the HLA Region in the Prediction of Type 1 Diabetes. *Curr. Diab. Rep.* **11,** 533–542 (2011).

13. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13,** 101 (2011).

14. Huang, H., Fang, M., Jostins, L., Umićević Mirkov, M., Boucher, G., Anderson, C. A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., D'Amato, M., Deffontaine, V., Dmitrieva, J., Docampo, E., Elansary, M., Farh, K. K.-H., Franke, A., Gori, A.-S., Goyette, P., Halfvarson, J., Haritunians, T., Knight, J., Lawrance, I. C., Lees, C. W., Louis, E., Mariman, R., Meuwissen, T., Mni, M., Momozawa, Y., Parkes, M., Spain, S. L., Théâtre, E., Trynka, G., Satsangi, J., van Sommeren, S., Vermeire, S., Xavier, R. J., International Inflammatory Bowel Disease Genetics Consortium, Weersma, R. K., Duerr, R. H., Mathew, C. G., Rioux, J. D., McGovern, D. P. B., Cho, J. H., Georges, M., Daly, M. J. & Barrett, J. C. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547,** 173–178 (2017).

15. Pociot, F. & Lernmark, Å. Genetic risk factors for type 1 diabetes. *The Lancet* **387,** 2331–2339 (2016).

16. Sirota, M., Schaub, M. A., Batzoglou, S., Robinson, W. H. & Butte, A. J. Autoimmune disease classification by inverse association with SNP alleles. *PLoS Genet.* **5,** e1000792 (2009).

17. Onengut-Gumuscu, S., Chen, W.-M., Burren, O., Cooper, N. J., Quinlan, A. R., Mychaleckyj, J. C., Farber, E., Bonnie, J. K., Szpak, M., Schofield, E., Achuthan, P., Guo, H., Fortune, M. D., Stevens, H., Walker, N. M., Ward, L. D., Kundaje, A., Kellis, M., Daly, M. J., Barrett, J. C., Cooper, J. D., Deloukas, P., Type 1 Diabetes Genetics Consortium, Todd, J. A., Wallace, C., Concannon, P. & Rich, S. S. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47,** 381–386 (2015).

18. Parkes, M., Barrett, J. C., Prescott, N. J., Tremelling, M., Anderson, C. A., Fisher, S. A., Roberts, R. G., Nimmo, E. R., Cummings, F. R., Soars, D., Drummond, H., Lees, C. W., Khawaja, S. A., Bagnall, R., Burke, D. A., Todhunter, C. E., Ahmad, T., Onnie, C. M., McArdle, W., Strachan, D., Bethel, G., Bryan, C., Lewis, C. M., Deloukas, P., Forbes, A., Sanderson, J., Jewell, D. P., Satsangi, J., Mansfield, J. C., Cardon, L. & Mathew, C. G. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.* **39,** 830–832 (2007).

19. Rioux, J. D., Xavier, R. J., Taylor, K. D., Silverberg, M. S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M. M., Datta, L. W., Shugart, Y. Y., Griffiths, A. M., Targan, S. R., Ippoliti, A. F., Bernard, E.-J., Mei, L., Nicolae, D. L., Regueiro, M., Schumm, L. P., Steinhart, A. H., Rotter, J. I., Duerr, R. H., Cho, J. H., Daly, M. J. &

Brant, S. R. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* **39,** 596–604 (2007).

20. Chauhan, S., Mandell, M. A. & Deretic, V. IRGM governs the core autophagy machinery to conduct antimicrobial defense. *Mol. Cell* **58,** 507–521 (2015).
21. Jansen, P. R., Watanabe, K., Stringer, S., Skene, N., Bryois, J., Hammerschlag, A. R., Leeuw, C. A. de, Benjamins, J. S., Muñoz-Manchado, A. B., Nagel, M., Savage, J. E., Tiemeier, H., White, T., Tung, J. Y., Hinds, D. A., Vacic, V., Wang, X., Sullivan, P. F., Sluis, S. van der, Polderman, T. J. C., Smit, A. B., Hjerling-Leffler, J., Someren, E. J. W. V. & Posthuma, D. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat. Genet.* **51,** 394 (2019).
22. Lane, J. M., Liang, J., Vlasac, I., Anderson, S. G., Bechtold, D. A., Bowden, J., Emsley, R., Gill, S., Little, M. A., Luik, A. I., Loudon, A., Scheer, F. A. J. L., Purcell, S. M., Kyle, S. D., Lawlor, D. A., Zhu, X., Redline, S., Ray, D. W., Rutter, M. K. & Saxena, R. Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. *Nat. Genet.* **49,** 274–281 (2017).
23. Hammerschlag, A. R., Stringer, S., de Leeuw, C. A., Sniekers, S., Taskesen, E., Watanabe, K., Blanken, T. F., Dekker, K., Te Lindert, B. H. W., Wassing, R., Jonsdottir, I., Thorleifsson, G., Stefansson, H., Gislason, T., Berger, K., Schormair, B., Wellmann, J., Winkelmann, J., Stefansson, K., Oexle, K., Van Someren, E. J. W. & Posthuma, D. Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits. *Nat. Genet.* **49,** 1584–1592 (2017).

24. Consortium*†, I. M. S. G. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365,** eaav7188 (2019).

25. Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A. & O'Brien, S. J. Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* **11,** 724 (2010).

26. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322,** 881–888 (2008).

27. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* **24,** R111–R119 (2015).

28. Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., Belbin, G. M., Bien, S. A., Cheng, I., Cullina, S., Hodonsky, C. J., Hu, Y., Huckins, L. M., Jeff, J., Justice, A. E., Kocarnik, J. M., Lim, U., Lin, B. M., Lu, Y., Nelson, S. C., Park, S.-S. L., Poisner, H., Preuss, M. H., Richard, M. A., Schurmann, C., Setiawan, V. W., Sockell, A., Vahi, K., Verbanck, M., Vishnu, A., Walker, R. W., Young, K. L., Zubair, N., Acuña-Alonso, V., Ambite, J. L., Barnes, K. C., Boerwinkle, E., Bottinger, E. P., Bustamante, C. D., Caberto, C., Canizales-Quinteros, S., Conomos, M. P., Deelman, E., Do, R., Doheny, K., Fernández-Rhodes, L., Fornage, M., Hailu, B., Heiss, G., Henn, B. M., Hindorff, L. A., Jackson, R. D., Laurie, C. A., Laurie, C. C., Li, Y., Lin, D.-Y., Moreno-Estrada, A., Nadkarni, G., Norman, P. J., Pooler, L. C., Reiner, A. P., Romm, J., Sabatti, C., Sandoval, K., Sheng, X., Stahl, E. A., Stram, D. O., Thornton, T. A., Wassel, C. L., Wilkens, L. R., Winkler, C. A., Yoneyama, S., Buyske, S., Haiman, C. A., Kooperberg, C., Le Marchand, L., Loos, R. J. F., Matise, T. C., North, K. E., Peters, U., Kenny, E. E. & Carlson, C. S. Genetic analyses of diverse

populations improves discovery for complex traits. *Nature* (2019). doi:10.1038/s41586-019-1310-4

29. Prohaska, A., Racimo, F., Schork, A. J., Sikora, M., Stern, A. J., Ilardo, M., Allentoft, M. E., Folkersen, L., Buil, A., Moreno-Mayar, J. V., Korneliussen, T., Geschwind, D., Ingason, A., Werge, T., Nielsen, R. & Willerslev, E. Human Disease Variation in the Light of Population Genomics. *Cell* **177,** 115–131 (2019).
30. List of countries by incidence of Type 1 diabetes ages 0 to 14. *Diabetes UK* at <https://www.diabetes.org.uk/about_us/news_landing_page/uk-has-worlds-5th-highest-rate-of-type-1-diabetes-in-children/list-of-countries-by-incidence-of-type-1-diabetes-ages-0-to-14>
31. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11,** 459–463 (2010).
32. Sul, J. H., Martin, L. S. & Eskin, E. Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genet.* **14,** e1007309 (2018).
33. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **102,** 717–730 (2018).
34. Benner, C., Spencer, C. C. A., Havulinna, A. S., Salomaa, V., Ripatti, S. & Pirinen, M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32,** 1493–1501 (2016).
35. MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., Adams, D. R., Altman, R. B., Antonarakis, S. E., Ashley, E. A., Barrett, J. C., Biesecker, L. G., Conrad, D. F., Cooper, G. M., Cox, N. J., Daly, M. J., Gerstein, M. B., Goldstein, D. B., Hirschhorn, J. N., Leal, S. M., Pennacchio, L. A.,

Stamatoyannopoulos, J. A., Sunyaev, S. R., Valle, D., Voight, B. F., Winckler, W. & Gunter, C. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508,** 469–476 (2014).

36. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19,** 491–504 (2018).

37. Dendrou, C. A., Plagnol, V., Fung, E., Yang, J. H. M., Downes, K., Cooper, J. D., Nutland, S., Coleman, G., Himsworth, M., Hardy, M., Burren, O., Healy, B., Walker, N. M., Koch, K., Ouwehand, W. H., Bradley, J. R., Wareham, N. J., Todd, J. A. & Wicker, L. S. Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nat. Genet.* **41,** 1011–1015 (2009).

38. Dendrou, C. A., Cortes, A., Shipman, L., Evans, H. G., Attfield, K. E., Jostins, L., Barber, T., Kaur, G., Kuttikkatte, S. B., Leach, O. A., Desel, C., Faergeman, S. L., Cheeseman, J., Neville, M. J., Sawcer, S., Compston, A., Johnson, A. R., Everett, C., Bell, J. I., Karpe, F., Ultsch, M., Eigenbrot, C., McVean, G. & Fugger, L. Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity. *Sci. Transl. Med.* **8,** 363ra149-363ra149 (2016).

39. Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., Hafler, D. A. & Bernstein, B. E. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).

40. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J.,

Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Yaping Liu, Coarfa, C., Alan Harris, R., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., David Hawkins, R., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Scott Hansen, R., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., Jager, P. L. D., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T. & Kellis, M. Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

41. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

42. Soskic, B., Cano-Gamez, E., Smyth, D. J., Rowan, W. C., Nakic, N., Esparza-Gordillo, J., Bossini-Castillo, L., Tough, D. F., Larminie, C. G. C., Bronson, P. G., Willé, D. & Trynka, G. Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nat. Genet.* 1–8 (2019). doi:10.1038/s41588-019-0493-9

43. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B.,

Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., Pervjakova, N., Alvaes, I., Fave, M.-J., Agbessi, M., Christiansen, M., Jansen, R., Seppälä, I., Tong, L., Teumer, A., Schramm, K., Hemani, G., Verlouw, J., Yaghootkar, H., Sönmez, R., Brown, A., Kukushkina, V., Kalnapenkis, A., Rüeger, S., Porcu, E., Kronberg-Guzman, J., Kettunen, J., Powell, J., Lee, B., Zhang, F., Arindrarto, W., Beutner, F., Consortium, B., Brugge, H., Consortium, i2QTL, Dmitreva, J., Elansary, M., Fairfax, B. P., Georges, M., Heijmans, B. T., Kähönen, M., Kim, Y., Knight, J. C., Kovacs, P., Krohn, K., Li, S., Loeffler, M., Marigorta, U. M., Mei, H., Momozawa, Y., Müller-Nurasyid, M., Nauck, M., Nivard, M., Penninx, B., Pritchard, J., Raitakari, O., Rotzchke, O., Slagboom, E. P., Stehouwer, C. D. A., Stumvoll, M., Sullivan, P., Hoen, P. A. C. 't, Thiery, J., Tönjes, A., Dongen, J. van, Iterson, M. van, Veldink, J., Völker, U., Wijmenga, C., Swertz, M., Andiappan, A., Montgomery, G. W., Ripatti, S., Perola, M., Kutalik, Z., Dermitzakis, E., Bergmann, S., Frayling, T., Meurs, J. van, Prokisch, H., Ahsan, H., Pierce, B., Lehtimäki, T., Boomsma, D., Psaty, B. M., Gharib, S. A., Awadalla, P., Milani, L., Ouwehand, W., Downes, K., Stegle, O., Battle, A., Yang, J., Visscher, P. M., Scholz, M., Gibson, G., Esko, T. & Franke, L. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 447367 (2018). doi:10.1101/447367

44. The GTEx Consortium atlas of genetic regulatory effects across human tissues | bioRxiv. at <https://www.biorxiv.org/content/10.1101/787903v1>

45. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50,** 1–14 (2018).

46. Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke,

B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J. T., Theis, F. J., Uhlen, M., Van, A. O., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R. & Yosef, N. The Human Cell Atlas. *eLife* **6,** (2017).

47. Montoro, D. T., Haber, A. L., Biton, M., Vinarsky, V., Lin, B., Birket, S. E., Yuan, F., Chen, S., Leung, H. M., Villoria, J., Rogel, N., Burgin, G., Tsankov, A. M., Waghray, A., Slyper, M., Waldman, J., Nguyen, L., Dionne, D., Rozenblatt-Rosen, O., Tata, P. R., Mou, H., Shivaraju, M., Bihler, H., Mense, M., Tearney, G. J., Rowe, S. M., Engelhardt, J. F., Regev, A. & Rajagopal, J. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560,** 319–324 (2018).

48. Plasschaert, L. W., Žilionis, R., Choo-Wing, R., Savova, V., Knehr, J., Roma, G., Klein, A. M. & Jaffe, A. B. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560,** 377–381 (2018).

49. van der Wijst, M. G. P., de Vries, D. H., Groot, H. E., Trynka, G., Hon, C.-C., Nawijn, M. C., Idaghdour, Y., van der Harst, P., Ye, C. J., Powell, J., Theis, F. J., Mahfouz, A., Heinig, M. & Franke, L. Single-cell eQTLGen Consortium: a personalized understanding of disease. *ArXiv190912550 Q-Bio* (2019). at <http://arxiv.org/abs/1909.12550>

50. Todd, J. A. Etiology of type 1 diabetes. *Immunity* **32,** 457–467 (2010).

51. Yamanouchi, J., Rainbow, D., Serra, P., Howlett, S., Hunter, K., Garner, V. E. S., Gonzalez-Munoz, A., Clark, J., Veijola, R., Cubbon, R., Chen, S.-L., Rosa, R., Cumiskey, A. M., Serreze, D. V., Gregory, S., Rogers, J., Lyons, P. A., Healy, B., Smink, L. J., Todd, J. A., Peterson, L. B., Wicker, L. S. & Santamaria, P. Interleukin-2 gene variation impairs regulatory T cell function and causes autoimmunity. *Nat. Genet.* **39,** 329–337 (2007).

52. Waldron-Lynch, F., Kareclas, P., Irons, K., Walker, N. M., Mander, A., Wicker, L. S., Todd, J. A. & Bond, S. Rationale and study design of the Adaptive study of IL-2 dose on regulatory T cells in type 1 diabetes (DILT1D): a non-randomised, open label, adaptive dose finding trial. *BMJ Open* **4,** e005559 (2014).

53. Todd, J. A., Evangelou, M., Cutler, A. J., Pekalski, M. L., Walker, N. M., Stevens, H. E., Porter, L., Smyth, D. J., Rainbow, D. B., Ferreira, R. C., Esposito, L., Hunter, K. M. D., Loudon, K., Irons, K., Yang, J. H., Bell, C. J. M., Schuilenburg, H., Heywood, J., Challis, B., Neupane, S., Clarke, P., Coleman, G., Dawson, S., Goymer, D., Anselmiova, K., Kennet, J., Brown, J., Caddy, S. L., Lu, J., Greatorex, J., Goodfellow, I., Wallace, C., Tree, T. I., Evans, M., Mander, A. P., Bond, S., Wicker, L. S. & Waldron-Lynch, F. Regulatory T Cell Responses in Participants with Type 1 Diabetes after a Single Dose of Interleukin-2: A Non-Randomised, Open Label, Adaptive Dose-Finding Trial. *PLOS Med.* **13,** e1002139 (2016).

54. Seelig, E., Howlett, J., Porter, L., Truman, L., Heywood, J., Kennet, J., Arbon, E. L., Anselmiova, K., Walker, N. M., Atkar, R., Pekalski, M. L., Rytina, E., Evans, M., Wicker, L. S., Todd, J. A., Mander, A. P., Bond, S. & Waldron-Lynch, F. The DILfrequency study is an adaptive trial to identify optimal IL-2 dosing in patients with type 1 diabetes. *JCI Insight* **3,** (2018).

55. Truman, L. A., Pekalski, M. L., Kareclas, P., Evangelou, M., Walker, N. M., Howlett, J., Mander, A. P., Kennet, J., Wicker, L. S., Bond, S., Todd, J. A. & Waldron-Lynch, F. Protocol of the adaptive study of IL-2 dose frequency on regulatory T cells in type 1 diabetes (DILfrequency): a mechanistic, non-randomised, repeat dose, open-label, response-adaptive study. *BMJ Open* **5,** e009799 (2015).

56. Nelson, B. H., Lord, J. D. & Greenberg, P. D. Cytoplasmic domains of the interleukin-2 receptor beta and gamma chains mediate the signal for T-cell proliferation. *Nature* **369,** 333–336 (1994).

57. Nakamura, Y., Russell, S. M., Mess, S. A., Friedmann, M., Erdos, M., Francois, C., Jacques, Y., Adelstein, S. & Leonard, W. J. Heterodimerization of the IL-2 receptor beta- and gamma-chain cytoplasmic domains is required for signalling. *Nature* **369,** 330–333 (1994).

58. Liao, W., Lin, J.-X. & Leonard, W. J. IL-2 Family Cytokines: New Insights into the Complex Roles of IL-2 as a Broad Regulator of T helper Cell Differentiation. *Curr. Opin. Immunol.* **23,** 598–604 (2011).

59. Naranjo-Gómez, M., Oliva, H., Climent, N., Fernández, M. A., Ruiz-Riol, M., Bofill, M., Gatell, J. M., Gallart, T., Pujol-Borrell, R. & Borràs, F. E. Expression and function of the IL-2 receptor in activated human plasmacytoid dendritic cells. *Eur. J. Immunol.* **37,** 1764–1772 (2007).

60. Wuest, S. C., Edwan, J. H., Martin, J. F., Han, S., Perry, J. S. A., Cartagena, C. M., Matsuura, E., Maric, D., Waldmann, T. A. & Bielekova, B. A role for interleukin-2 trans -presentation in dendritic cell–mediated T cell activation in humans, as revealed by daclizumab therapy. *Nat. Med.* **17,** 604–609 (2011).

61. Liao, W., Lin, J.-X. & Leonard, W. J. Interleukin-2 at the Crossroads of Effector

Responses, Tolerance, and Immunotherapy. *Immunity* **38,** 13–25 (2013).

62. Ross, S. H. & Cantrell, D. A. Signaling and Function of Interleukin-2 in T Lymphocytes. *Annu. Rev. Immunol.* **36,** 411–433 (2018).

63. Owen, D. L., Mahmud, S. A., Vang, K. B., Kelly, R. M., Blazar, B. R., Smith, K. A. & Farrar, M. A. Identification of Cellular Sources of IL-2 Needed for Regulatory T Cell Development and Homeostasis. *J. Immunol. Baltim. Md 1950* **200,** 3926–3933 (2018).

64. Cheng, G., Yu, A., Dee, M. J. & Malek, T. R. IL-2R signaling is essential for functional maturation of regulatory T cells during thymic development. *J. Immunol. Baltim. Md 1950* **190,** 1567–1575 (2013).

65. Toomer, K. H., Lui, J. B., Altman, N. H., Ban, Y., Chen, X. & Malek, T. R. Essential and non-overlapping IL-2Rα-dependent processes for thymic development and peripheral homeostasis of regulatory T cells. *Nat. Commun.* **10,** 1–16 (2019).

66. Fontenot, J. D., Gavin, M. A. & Rudensky, A. Y. Foxp3 programs the development and function of CD4+CD25+ regulatory T cells. *Nat. Immunol.* **4,** 330–336 (2003).

67. Hori, S., Nomura, T. & Sakaguchi, S. Control of Regulatory T Cell Development by the Transcription Factor Foxp3. *Science* **299,** 1057–1061 (2003).

68. Khattri, R., Cox, T., Yasayko, S.-A. & Ramsdell, F. An essential role for Scurfin in CD4 + CD25 + T regulatory cells. *Nat. Immunol.* **4,** 337 (2003).

69. Bacchetta, R., Passerini, L., Gambineri, E., Dai, M., Allan, S. E., Perroni, L., Dagna-Bricarelli, F., Sartirana, C., Matthes-Martin, S., Lawitschka, A., Azzari, C., Ziegler, S. F., Levings, M. K. & Roncarolo, M. G. Defective regulatory and effector T cell functions in patients with FOXP3 mutations. *J. Clin. Invest.* **116,** 1713–1722 (2006).

70. Miyara, M., Gorochov, G., Ehrenstein, M., Musset, L., Sakaguchi, S. & Amoura, Z. Human FoxP3+ regulatory T cells in systemic autoimmune diseases. *Autoimmun. Rev.* **10,** 744–755 (2011).

71. Grant, C. R., Liberal, R., Mieli-Vergani, G., Vergani, D. & Longhi, M. S. Regulatory T-cells in autoimmune diseases: Challenges, controversies and—yet—unanswered questions. *Autoimmun. Rev.* **14,** 105–116 (2015).

72. Lindley, S., Dayan, C. M., Bishop, A., Roep, B. O., Peakman, M. & Tree, T. I. M. Defective suppressor function in CD4(+)CD25(+) T-cells from patients with type 1 diabetes. *Diabetes* **54,** 92–99 (2005).

73. Viglietta, V., Baecher-Allan, C., Weiner, H. L. & Hafler, D. A. Loss of functional suppression by CD4+CD25+ regulatory T cells in patients with multiple sclerosis. *J. Exp. Med.* **199,** 971–979 (2004).

74. Mao, C., Wang, S., Xiao, Y., Xu, J., Jiang, Q., Jin, M., Jiang, X., Guo, H., Ning, G. & Zhang, Y. Impairment of regulatory capacity of CD4+CD25+ regulatory T cells mediated by dendritic cell polarization and hyperthyroidism in Graves' disease. *J. Immunol. Baltim. Md 1950* **186,** 4734–4743 (2011).

75. Qureshi, O. S., Zheng, Y., Nakamura, K., Attridge, K., Manzotti, C., Schmidt, E. M., Baker, J., Jeffery, L. E., Kaur, S., Briggs, Z., Hou, T. Z., Futter, C. E., Anderson, G., Walker, L. S. K. & Sansom, D. M. Trans-Endocytosis of CD80 and CD86: A Molecular Basis for the Cell-Extrinsic Function of CTLA-4. *Science* **332,** 600–603 (2011).

76. Zhu, C., Anderson, A. C., Schubart, A., Xiong, H., Imitola, J., Khoury, S. J., Zheng, X. X., Strom, T. B. & Kuchroo, V. K. The Tim-3 ligand galectin-9 negatively regulates T helper type 1 immunity. *Nat. Immunol.* **6,** 1245–1252 (2005).

77. Wei, X., Zhang, J., Gu, Q., Huang, M., Zhang, W., Guo, J. & Zhou, X. Reciprocal Expression of IL-35 and IL-10 Defines Two Distinct Effector Treg Subsets that Are Required for Maintenance of Immune Tolerance. *Cell Rep.* **21,** 1853–1869 (2017).

78. Konkel, J. E., Zhang, D., Zanvit, P., Chia, C., Zangarle-Murray, T., Jin, W., Wang, S. & Chen, W. Transforming Growth Factor-β Signaling in Regulatory T Cells Controls T Helper-17 Cells and Tissue-Specific Immune Responses. *Immunity* **46,** 660–674 (2017).

79. Mandapathil, M., Hilldorfer, B., Szczepanski, M. J., Czystowska, M., Szajnik, M., Ren, J., Lang, S., Jackson, E. K., Gorelik, E. & Whiteside, T. L. Generation and Accumulation of Immunosuppressive Adenosine by Human CD4+CD25highFOXP3+ Regulatory T Cells. *J. Biol. Chem.* **285,** 7176–7186 (2010).

80. Kumar, P., Saini, S., Khan, S., Surendra Lele, S. & Prabhakar, B. S. Restoring self-tolerance in autoimmune diseases by enhancing regulatory T-cells. *Cell. Immunol.* **339,** 41–49 (2019).

81. Hull, C. M., Peakman, M. & Tree, T. I. M. Regulatory T cell dysfunction in type 1 diabetes: what's broken and how can we fix it? *Diabetologia* **60,** 1839–1850 (2017).

82. Baron, U., Floess, S., Wieczorek, G., Baumann, K., Grützkau, A., Dong, J., Thiel, A., Boeld, T. J., Hoffmann, P., Edinger, M., Türbachova, I., Hamann, A., Olek, S. & Huehn, J. DNA demethylation in the human FOXP3 locus discriminates regulatory T cells from activated FOXP3(+) conventional T cells. *Eur. J. Immunol.* **37,** 2378–2389 (2007).

83. Floess, S., Freyer, J., Siewert, C., Baron, U., Olek, S., Polansky, J., Schlawe, K., Chang, H.-D., Bopp, T., Schmitt, E., Klein-Hessling, S., Serfling, E., Hamann, A. & Huehn, J. Epigenetic control of the foxp3 locus in regulatory T cells. *PLoS Biol.* **5,**

e38 (2007).

84. Miyara, M., Yoshioka, Y., Kitoh, A., Shima, T., Wing, K., Niwa, A., Parizot, C., Taflin, C., Heike, T., Valeyre, D., Mathian, A., Nakahata, T., Yamaguchi, T., Nomura, T., Ono, M., Amoura, Z., Gorochov, G. & Sakaguchi, S. Functional Delineation and Differentiation Dynamics of Human CD4+ T Cells Expressing the FoxP3 Transcription Factor. *Immunity* **30,** 899–911 (2009).

85. Toker, A., Engelbert, D., Garg, G., Polansky, J. K., Floess, S., Miyao, T., Baron, U., Düber, S., Geffers, R., Giehr, P., Schallenberg, S., Kretschmer, K., Olek, S., Walter, J., Weiss, S., Hori, S., Hamann, A. & Huehn, J. Active demethylation of the Foxp3 locus leads to the generation of stable regulatory T cells within the thymus. *J. Immunol. Baltim. Md 1950* **190,** 3180–3188 (2013).

86. Stockis, J., Fink, W., François, V., Connerotte, T., de Smet, C., Knoops, L., van der Bruggen, P., Boon, T., Coulie, P. G. & Lucas, S. Comparison of stable human Treg and Th clones by transcriptional profiling. *Eur. J. Immunol.* **39,** 869–882 (2009).

87. Wieczorek, G., Asemissen, A., Model, F., Turbachova, I., Floess, S., Liebenberg, V., Baron, U., Stauch, D., Kotsch, K., Pratschke, J., Hamann, A., Loddenkemper, C., Stein, H., Volk, H. D., Hoffmüller, U., Grützkau, A., Mustea, A., Huehn, J., Scheibenbogen, C. & Olek, S. Quantitative DNA methylation analysis of FOXP3 as a new method for counting regulatory T cells in peripheral blood and solid tissue. *Cancer Res.* **69,** 599–608 (2009).

88. Zheng, Y., Josefowicz, S., Chaudhry, A., Peng, X. P., Forbush, K. & Rudensky, A. Y. Role of conserved non-coding DNA elements in the *Foxp3* gene in regulatory T-cell fate. *Nature* **463,** 808–812 (2010).

89. Lewin, J., Schmitt, A. O., Adorján, P., Hildmann, T. & Piepenbrock, C. Quantitative

DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates. *Bioinforma. Oxf. Engl.* **20,** 3005–3012 (2004).

90. Spreafico, R., Rossetti, M., van den Broek, T., Jansen, N. J. G., Zhang, H., Moshref, M., Prakken, B., van Loosdregt, J., van Wijk, F. & Albani, S. A sensitive protocol for FOXP3 epigenetic analysis in scarce human samples. *Eur. J. Immunol.* **44,** 3141–3143 (2014).

91. Rainbow, D. B., Yang, X., Burren, O., Pekalski, M. L., Smyth, D. J., Klarqvist, M. D. R., Penkett, C. J., Brugger, K., Martin, H., Todd, J. A., Wallace, C. & Wicker, L. S. Epigenetic analysis of regulatory T cells using multiplex bisulfite sequencing. *Eur. J. Immunol.* **45,** 3200–3203 (2015).

92. Pekalski, M. L., Ferreira, R. C., Coulson, R. M. R., Cutler, A. J., Guo, H., Smyth, D. J., Downes, K., Dendrou, C. A., Castro Dopico, X., Esposito, L., Coleman, G., Stevens, H. E., Nutland, S., Walker, N. M., Guy, C., Dunger, D. B., Wallace, C., Tree, T. I. M., Todd, J. A. & Wicker, L. S. Postthymic expansion in human CD4 naive T cells defined by expression of functional high-affinity IL-2 receptors. *J. Immunol. Baltim. Md 1950* **190,** 2554–2566 (2013).

93. Ohkura, N., Hamaguchi, M., Morikawa, H., Sugimura, K., Tanaka, A., Ito, Y., Osaki, M., Tanaka, Y., Yamashita, R., Nakano, N., Huehn, J., Fehling, H. J., Sparwasser, T., Nakai, K. & Sakaguchi, S. T Cell Receptor Stimulation-Induced Epigenetic Changes and Foxp3 Expression Are Independent and Complementary Events Required for Treg Cell Development. *Immunity* **37,** 785–799 (2012).

94. Mazzoni, A., Santarlasci, V., Maggi, L., Capone, M., Rossi, M. C., Querci, V., Palma, R. D., Chang, H.-D., Thiel, A., Cimaz, R., Liotta, F., Cosmi, L., Maggi, E., Radbruch, A., Romagnani, S., Dong, J. & Annunziato, F. Demethylation of the RORC2 and

IL17A in Human CD4+ T Lymphocytes Defines Th17 Origin of Nonclassic Th1 Cells. *J. Immunol.* **194,** 3116–3126 (2015).

95. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30,** 2114–2120 (2014).
96. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).
97. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinforma. Oxf. Engl.* **27,** 2957–2963 (2011).
98. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
99. Ferreira, R. C., Rainbow, D. B., Rubio García, A., Pekalski, M. L., Porter, L., Oliveira, J. J., Waldron-Lynch, F., Wicker, L. S. & Todd, J. A. Human IL-6RhiTIGIT-CD4+CD127lowCD25+ T cells display potent in vitro suppressive capacity and a distinct Th17 profile. *Clin. Immunol. Orlando Fla* **179,** 25–39 (2017).
100. Ferreira, R. C., Rainbow, D. B., Rubio García, A., Pekalski, M. L., Porter, L., Oliveira, J. J., Waldron-Lynch, F., Wicker, L. S. & Todd, J. A. In-depth immunophenotyping data of IL-6R on the human peripheral regulatory T cell (Treg) compartment. *Data Brief* **12,** 676–691 (2017).
101. Ferreira, R. C., Simons, H. Z., Thompson, W. S., Rainbow, D. B., Yang, X., Cutler, A. J., Oliveira, J., Castro Dopico, X., Smyth, D. J., Savinykh, N., Mashar, M., Vyse, T. J., Dunger, D. B., Baxendale, H., Chandra, A., Wallace, C., Todd, J. A., Wicker, L. S. & Pekalski, M. L. Cells with Treg-specific FOXP3 demethylation but low CD25 are prevalent in autoimmunity. *J. Autoimmun.* **84,** 75–86 (2017).
102. Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. & Lee, J. J.

Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4,** 7 (2015).

103. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J. & Sham, P. C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

104. Asimit, J. L., Rainbow, D. B., Fortune, M. D., Grinberg, N. F., Wicker, L. S. & Wallace, C. Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases. *Nat. Commun.* **10,** 3216 (2019).

105. 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A. & Abecasis, G. R. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

106. Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Mu, X. J., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu,

F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., 1000 Genomes Project Consortium, Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E. & Korbel, J. O. An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).

107. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinforma. Oxf. Engl.* **23,** 1289–1291 (2007).
108. Kõressaar, T., Lepamets, M., Kaplinski, L., Raime, K., Andreson, R. & Remm, M. Primer3_masker: integrating masking of template sequence with primer design software. *Bioinforma. Oxf. Engl.* **34,** 1937–1938 (2018).
109. Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. & Rozen, S. G. Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40,** e115 (2012).
110. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A. & Drummond, A. J. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Comput. Biol.* **10,** e1003537 (2014).
111. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35,** 1547–1549 (2018).
112. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9,** 179–181 (2011).
113. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).

114. Wang, J., Ioan-Facsinay, A., van der Voort, E. I. H., Huizinga, T. W. J. & Toes, R. E. M. Transient expression of FOXP3 in human activated nonregulatory CD4+ T cells. *Eur. J. Immunol.* **37,** 129–138 (2007).

115. Pillai, V., Ortega, S. B., Wang, C. K. & Karandikar, N. J. Transient regulatory T-cells: a state attained by all activated human T-cells. *Clin. Immunol. Orlando Fla* **123,** 18–29 (2007).

116. Polansky, J. K., Schreiber, L., Thelemann, C., Ludwig, L., Krüger, M., Baumgrass, R., Cording, S., Floess, S., Hamann, A. & Huehn, J. Methylation matters: binding of Ets-1 to the demethylated Foxp3 gene contributes to the stabilization of Foxp3 expression in regulatory T cells. *J. Mol. Med.* **88,** 1029–1040 (2010).

117. Schreiber, L., Pietzsch, B., Floess, S., Farah, C., Jänsch, L., Schmitz, I. & Huehn, J. The Treg-Specific Demethylated Region Stabilizes Foxp3 Expression Independently of NF-κB Signaling. *PLOS ONE* **9,** e88318 (2014).

118. Burchill, M. A., Yang, J., Vogtenhuber, C., Blazar, B. R. & Farrar, M. A. IL-2 receptor beta-dependent STAT5 activation is required for the development of Foxp3+ regulatory T cells. *J. Immunol. Baltim. Md 1950* **178,** 280–290 (2007).

119. Murawski, M. R., Litherland, S. A., Clare-Salzler, M. J. & Davoodi-Semiromi, A. Upregulation of Foxp3 expression in mouse and human Treg is IL-2/STAT5 dependent: implications for the NOD STAT5B mutation in diabetes pathogenesis. *Ann. N. Y. Acad. Sci.* **1079,** 198–204 (2006).

120. Yao, Z., Kanno, Y., Kerenyi, M., Stephens, G., Durant, L., Watford, W. T., Laurence, A., Robinson, G. W., Shevach, E. M., Moriggl, R., Hennighausen, L., Wu, C. & O'Shea, J. J. Nonredundant roles for Stat5a/b in directly regulating Foxp3. *Blood* **109,** 4368–4375 (2007).

121. Samstein, R. M., Arvey, A., Josefowicz, S. Z., Peng, X., Reynolds, A., Sandstrom, R., Neph, S., Sabo, P., Kim, J. M., Liao, W., O. Li, M., Leslie, C., Stamatoyannopoulos, J. A. & Rudensky, A. Y. Foxp3 exploits a preexistent enhancer landscape for regulatory T cell lineage specification. *Cell* **151,** 153–166 (2012).

122. Schmidt, A., Eriksson, M., Shang, M.-M., Weyd, H. & Tegnér, J. Comparative Analysis of Protocols to Induce Human CD4+Foxp3+ Regulatory T Cells by Combinations of IL-2, TGF-beta, Retinoic Acid, Rapamycin and Butyrate. *PloS One* **11,** e0148474 (2016).

123. Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434,** 400–404 (2005).

124. Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187,** 226–232 (1975).

125. Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.* **14,** 9–25 (1975).

126. Bestor, T. H. & Ingram, V. M. Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. *Proc. Natl. Acad. Sci.* **80,** 5559–5563 (1983).

127. Gerdes, J., Schwab, U., Lemke, H. & Stein, H. Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation. *Int. J. Cancer* **31,** 13–20 (1983).

128. Luijk, R., Wu, H., Ward-Caviness, C. K., Hannon, E., Carnero-Montoro, E., Min, J. L., Mandaviya, P., Müller-Nurasyid, M., Mei, H., Maarel, S. M. van der, Relton, C., Mill, J., Waldenberger, M., Bell, J. T., Jansen, R., Zhernakova, A., Franke, L., Hoen, P. A. C. ‘t, Boomsma, D. I., Duijn, C. M. van, Greevenbroek, M. M. J. van, Veldink,

J. H., Wijmenga, C., Meurs, J. van, Daxinger, L., Slagboom, P. E., Zwet, E. W. van & Heijmans, B. T. Autosomal genetic variation is associated with DNA methylation in regions variably escaping X-chromosome inactivation. *Nat. Commun.* **9,** 3738 (2018).

129. Cotton, A. M., Price, E. M., Jones, M. J., Balaton, B. P., Kobor, M. S. & Brown, C. J. Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum. Mol. Genet.* **24,** 1528–1539 (2015).
130. Yang, F., Babak, T., Shendure, J. & Disteche, C. M. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res.* **20,** 614–622 (2010).
131. Garieri, M., Stamoulis, G., Blanc, X., Falconnet, E., Ribaux, P., Borel, C., Santoni, F. & Antonarakis, S. E. Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts. *Proc. Natl. Acad. Sci.* **115,** 13015–13020 (2018).
132. Tukiainen, T., Villani, A.-C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., Cummings, B. B., Castel, S. E., Karczewski, K. J., Aguet, F., Byrnes, A., GTEx Consortium, Lappalainen, T., Aviv Regev, Ardlie, K. G., Hacohen, N. & MacArthur, D. G. Landscape of X chromosome inactivation across human tissues. *Nature* **550,** 244–248 (2017).
133. Schmiedel, B. J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A. G., White, B. M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J. A., McVicker, G., Seumois, G., Rao, A., Kronenberg, M., Peters, B. & Vijayanand, P. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175,** 1701-1715.e16 (2018).
134. Dheda, K., Huggett, J. F., Bustin, S. A., Johnson, M. A., Rook, G. & Zumla, A.

Validation of housekeeping genes for normalizing RNA expression in real-time PCR. *BioTechniques* **37,** 112–119 (2004).

135. Moskalev, E. A., Zavgorodnij, M. G., Majorova, S. P., Vorobjev, I. A., Jandaghi, P., Bure, I. V. & Hoheisel, J. D. Correction of PCR-bias in quantitative DNA methylation studies by means of cubic polynomial regression. *Nucleic Acids Res.* **39,** e77 (2011).
136. TOMMASINI, A., FERRARI, S., MORATTO, D., BADOLATO, R., BONIOTTO, M., PIRULLI, D., NOTARANGELO, L. D. & ANDOLINA, M. X-chromosome inactivation analysis in a female carrier of FOXP3 mutation. *Clin. Exp. Immunol.* **130,** 127–130 (2002).
137. Wang, J., Syrett, C. M., Kramer, M. C., Basu, A., Atchison, M. L. & Anguera, M. C. Unusual maintenance of X chromosome inactivation predisposes female lymphocytes for increased expression from the inactive X. *Proc. Natl. Acad. Sci.* **113,** E2029–E2038 (2016).
138. Souyris, M., Cenac, C., Azar, P., Daviaud, D., Canivet, A., Grunenwald, S., Pienkowski, C., Chaumeil, J., Mejía, J. E. & Guéry, J.-C. TLR7 escapes X chromosome inactivation in immune cells. *Sci. Immunol.* **3,** (2018).
139. Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R. V., Branco, M. R. & Reik, W. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol.* **19,** 33 (2018).
140. Rand, A. C., Jain, M., Eizenga, J. M., Musselman-Brown, A., Olsen, H. E., Akeson, M. & Paten, B. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14,** 411–413 (2017).

141. Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J. & Timp, W. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14,** 407–410 (2017).

142. Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J. & Turner, S. W. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7,** 461–465 (2010).

143. Thornton, A. M., Korty, P. E., Tran, D. Q., Wohlfert, E. A., Murray, P. E., Belkaid, Y. & Shevach, E. M. Expression of Helios, an Ikaros transcription factor family member, differentiates thymic-derived from peripherally induced Foxp3+ T regulatory cells. *J. Immunol. Baltim. Md 1950* **184,** 3433–3441 (2010).

144. Verhagen, J. & Wraith, D. C. Comment on 'Expression of Helios, an Ikaros transcription factor family member, differentiates thymic-derived from peripherally induced Foxp3+ T regulatory cells'. *J. Immunol. Baltim. Md 1950* **185,** 7129; author reply 7130 (2010).

145. Akimova, T., Beier, U. H., Wang, L., Levine, M. H. & Hancock, W. W. Helios Expression Is a Marker of T Cell Activation and Proliferation. *PLOS ONE* **6,** e24226 (2011).

146. Gottschalk, R. A., Corse, E. & Allison, J. P. Expression of Helios in peripherally induced Foxp3+ regulatory T cells. *J. Immunol. Baltim. Md 1950* **188,** 976–980 (2012).

147. Weissler, K. A., Garcia, V., Kropf, E., Aitken, M., Bedoya, F., Wolf, A. I., Erikson, J. & Caton, A. J. Distinct Modes of Antigen Presentation Promote the Formation, Differentiation, and Activity of Foxp3+ Regulatory T Cells In Vivo. *J. Immunol.* **194,** 3784–3797 (2015).

148. Roncarolo, M. G., Gregori, S., Bacchetta, R., Battaglia, M. & Gagliani, N. The Biology of T Regulatory Type 1 Cells and Their Therapeutic Application in Immune-Mediated Diseases. *Immunity* **49,** 1004–1019 (2018).

149. Romano, M., Fanelli, G., Albany, C. J., Giganti, G. & Lombardi, G. Past, Present, and Future of Regulatory T Cell Therapy in Transplantation and Autoimmunity. *Front. Immunol.* **10,** (2019).

150. Tran, D. Q., Ramsey, H. & Shevach, E. M. Induction of FOXP3 expression in naive human CD4+FOXP3− T cells by T-cell receptor stimulation is transforming growth factor-β–dependent but does not confer a regulatory phenotype. *Blood* **110,** 2983–2990 (2007).

151. Zhang, B., Zhang, X., Tang, F. L., Zhu, L. P., Liu, Y. & Lipsky, P. E. Clinical significance of increased CD4+CD25−Foxp3+ T cells in patients with new-onset systemic lupus erythematosus. *Ann. Rheum. Dis.* **67,** 1037–1040 (2008).

152. Venigalla, R. K. C., Tretter, T., Krienke, S., Max, R., Eckstein, V., Blank, N., Fiehn, C., Ho, A. D. & Lorenz, H.-M. Reduced CD4+,CD25− T cell sensitivity to the suppressive function of CD4+,CD25high,CD127−/low regulatory T cells in patients with active systemic lupus erythematosus. *Arthritis Rheum.* **58,** 2120–2130 (2008).

153. Bonelli, M., Savitskaya, A., Steiner, C.-W., Rath, E., Smolen, J. S. & Scheinecker, C. Phenotypic and Functional Analysis of CD4+CD25−Foxp3+ T Cells in Patients with Systemic Lupus Erythematosus. *J. Immunol.* **182,** 1689–1695 (2009).

154. Suen, J.-L., Li, H.-T., Jong, Y.-J., Chiang, B.-L. & Yen, J.-H. Altered homeostasis of CD4+ FoxP3+ regulatory T-cell subpopulations in systemic lupus erythematosus. *Immunology* **127,** 196–205 (2009).

155. Fransson, M., Burman, J., Lindqvist, C., Atterby, C., Fagius, J. & Loskog, A. T

regulatory cells lacking CD25 are increased in MS during relapse. *Autoimmunity* **43,** 590–597 (2010).

156. de Paz, B., Prado, C., Alperi-López, M., Ballina-García, F. J., Rodriguez-Carrio, J., López, P. & Suárez, A. Effects of glucocorticoid treatment on CD25−FOXP3+ population and cytokine-producing cells in rheumatoid arthritis. *Rheumatology* **51,** 1198–1207 (2012).

157. Hori, S. Lineage stability and phenotypic plasticity of $Foxp3^+$ regulatory T cells. *Immunol. Rev.* **259,** 159–172 (2014).

158. Bin Dhuban, K., Kornete, M., S Mason, E. & Piccirillo, C. A. Functional dynamics of $Foxp3^+$ regulatory T cells in mice and humans. *Immunol. Rev.* **259,** 140–158 (2014).

159. Delacher, M., Imbusch, C. D., Weichenhan, D., Breiling, A., Hotz-Wagenblatt, A., Träger, U., Hofer, A. C., Kägebein, D., Wang, Q., Frauhammer, F., Mallm, J. P., Bauer, K., Herrmann, C., Lang, P. A., Brors, B., Plass, C. & Feuerer, M. Genome-wide DNA-methylation landscape defines specialization of regulatory T cells in tissues. *Nat. Immunol.* **18,** 1160–1172 (2017).

160. Thornton, A. M., Lu, J., Korty, P. E., Kim, Y. C., Martens, C., Sun, P. D. & Shevach, E. M. Helios+ and Helios− Treg subpopulations are phenotypically and functionally distinct and express dissimilar TCR repertoires. *Eur. J. Immunol.* **49,** 398–412 (2019).

161. Russell, J. N., Clements, J. E. & Gama, L. Quantitation of Gene Expression in Formaldehyde-Fixed and Fluorescence-Activated Sorted Cells. *PLoS ONE* **8,** (2013).

162. Xu, J., Nuno, K., Litzenburger, U. M., Qi, Y., Corces, M. R., Majeti, R. & Chang, H. Y. Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *eLife* **8,** e45105 (2019).

163. Saadoun, D., Rosenzwajg, M., Joly, F., Six, A., Carrat, F., Thibault, V., Sene, D.,

Cacoub, P. & Klatzmann, D. Regulatory T-Cell Responses to Low-Dose Interleukin-2 in HCV-Induced Vasculitis. *http://dx.doi.org/10.1056/NEJMoa1105143* (2011). doi:10.1056/NEJMoa1105143

164. Koreth, J., Matsuoka, K., Kim, H. T., McDonough, S. M., Bindra, B., Alyea, E. P., Armand, P., Cutler, C., Ho, V. T., Treister, N. S., Bienfang, D. C., Prasad, S., Tzachanis, D., Joyce, R. M., Avigan, D. E., Antin, J. H., Ritz, J. & Soiffer, R. J. Interleukin-2 and Regulatory T Cells in Graft-versus-Host Disease. *N. Engl. J. Med.* **365,** 2055–2066 (2011).
165. Bluestone, J. A., Buckner, J. H., Fitch, M., Gitelman, S. E., Gupta, S., Hellerstein, M. K., Herold, K. C., Lares, A., Lee, M. R., Li, K., Liu, W., Long, S. A., Masiello, L. M., Nguyen, V., Putnam, A. L., Rieck, M., Sayre, P. & Tang, Q. Type 1 diabetes immunotherapy using polyclonal regulatory T cells. *Sci. Transl. Med.* **7,** 315ra189 (2015).
166. Hippen, K. L., Merkel, S. C., Schirm, D. K., Nelson, C., Tennis, N. C., Riley, J. L., June, C. H., Miller, J. S., Wagner, J. E. & Blazar, B. R. Generation and Large-Scale Expansion of Human Inducible Regulatory T Cells That Suppress Graft-Versus-Host Disease. *Am. J. Transplant.* **11,** 1148–1157 (2011).
167. Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L. A., Cunningham, F. & Parkinson, H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47,** D1005–D1012 (2019).

168. Vella, A., Cooper, J. D., Lowe, C. E., Walker, N., Nutland, S., Widmer, B., Jones, R., Ring, S. M., McArdle, W., Pembrey, M. E., Strachan, D. P., Dunger, D. B., Twells, R. C. J., Clayton, D. G. & Todd, J. A. Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **76,** 773–779 (2005).

169. Lowe, C. E., Cooper, J. D., Brusko, T., Walker, N. M., Smyth, D. J., Bailey, R., Bourget, K., Plagnol, V., Field, S., Atkinson, M., Clayton, D. G., Wicker, L. S. & Todd, J. A. Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat. Genet.* **39,** 1074–1082 (2007).

170. Maier, L. M., Lowe, C. E., Cooper, J., Downes, K., Anderson, D. E., Severson, C., Clark, P. M., Healy, B., Walker, N., Aubin, C., Oksenberg, J. R., Hauser, S. L., Compston, A., Sawcer, S., International Multiple Sclerosis Genetics Consortium, De Jager, P. L., Wicker, L. S., Todd, J. A. & Hafler, D. A. IL2RA genetic heterogeneity in multiple sclerosis and type 1 diabetes susceptibility and soluble interleukin-2 receptor production. *PLoS Genet.* **5,** e1000322 (2009).

171. International Multiple Sclerosis Genetics Consortium (IMSGC), Beecham, A. H., Patsopoulos, N. A., Xifara, D. K., Davis, M. F., Kemppinen, A., Cotsapas, C., Shah, T. S., Spencer, C., Booth, D., Goris, A., Oturai, A., Saarela, J., Fontaine, B., Hemmer, B., Martin, C., Zipp, F., D'Alfonso, S., Martinelli-Boneschi, F., Taylor, B., Harbo, H. F., Kockum, I., Hillert, J., Olsson, T., Ban, M., Oksenberg, J. R., Hintzen, R., Barcellos, L. F., Wellcome Trust Case Control Consortium 2 (WTCCC2), International IBD Genetics Consortium (IIBDGC), Agliardi, C., Alfredsson, L., Alizadeh, M., Anderson, C., Andrews, R., Søndergaard, H. B., Baker, A., Band, G.,

Baranzini, S. E., Barizzone, N., Barrett, J., Bellenguez, C., Bergamaschi, L., Bernardinelli, L., Berthele, A., Biberacher, V., Binder, T. M. C., Blackburn, H., Bomfim, I. L., Brambilla, P., Broadley, S., Brochet, B., Brundin, L., Buck, D., Butzkueven, H., Caillier, S. J., Camu, W., Carpentier, W., Cavalla, P., Celius, E. G., Coman, I., Comi, G., Corrado, L., Cosemans, L., Cournu-Rebeix, I., Cree, B. A. C., Cusi, D., Damotte, V., Defer, G., Delgado, S. R., Deloukas, P., di Sapio, A., Dilthey, A. T., Donnelly, P., Dubois, B., Duddy, M., Edkins, S., Elovaara, I., Esposito, F., Evangelou, N., Fiddes, B., Field, J., Franke, A., Freeman, C., Frohlich, I. Y., Galimberti, D., Gieger, C., Gourraud, P.-A., Graetz, C., Graham, A., Grummel, V., Guaschino, C., Hadjixenofontos, A., Hakonarson, H., Halfpenny, C., Hall, G., Hall, P., Hamsten, A., Harley, J., Harrower, T., Hawkins, C., Hellenthal, G., Hillier, C., Hobart, J., Hoshi, M., Hunt, S. E., Jagodic, M., Jelčić, I., Jochim, A., Kendall, B., Kermode, A., Kilpatrick, T., Koivisto, K., Konidari, I., Korn, T., Kronsbein, H., Langford, C., Larsson, M., Lathrop, M., Lebrun-Frenay, C., Lechner-Scott, J., Lee, M. H., Leone, M. A., Leppä, V., Liberatore, G., Lie, B. A., Lill, C. M., Lindén, M., Link, J., Luessi, F., Lycke, J., Macciardi, F., Männistö, S., Manrique, C. P., Martin, R., Martinelli, V., Mason, D., Mazibrada, G., McCabe, C., Mero, I.-L., Mescheriakova, J., Moutsianas, L., Myhr, K.-M., Nagels, G., Nicholas, R., Nilsson, P., Piehl, F., Pirinen, M., Price, S. E., Quach, H., Reunanen, M., Robberecht, W., Robertson, N. P., Rodegher, M., Rog, D., Salvetti, M., Schnetz-Boutaud, N. C., Sellebjerg, F., Selter, R. C., Schaefer, C., Shaunak, S., Shen, L., Shields, S., Siffrin, V., Slee, M., Sorensen, P. S., Sorosina, M., Sospedra, M., Spurkland, A., Strange, A., Sundqvist, E., Thijs, V., Thorpe, J., Ticca, A., Tienari, P., van Duijn, C., Visser, E. M., Vucic, S., Westerlind, H., Wiley, J. S., Wilkins, A., Wilson, J. F., Winkelmann, J.,

Zajicek, J., Zindler, E., Haines, J. L., Pericak-Vance, M. A., Ivinson, A. J., Stewart, G., Hafler, D., Hauser, S. L., Compston, A., McVean, G., De Jager, P., Sawcer, S. J. & McCauley, J. L. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45,** 1353–1360 (2013).

172. International Multiple Sclerosis Genetics Consortium, Hafler, D. A., Compston, A., Sawcer, S., Lander, E. S., Daly, M. J., De Jager, P. L., de Bakker, P. I. W., Gabriel, S. B., Mirel, D. B., Ivinson, A. J., Pericak-Vance, M. A., Gregory, S. G., Rioux, J. D., McCauley, J. L., Haines, J. L., Barcellos, L. F., Cree, B., Oksenberg, J. R. & Hauser, S. L. Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* **357,** 851–862 (2007).
173. Cooper, J. D., Simmonds, M. J., Walker, N. M., Burren, O., Brand, O. J., Guo, H., Wallace, C., Stevens, H., Coleman, G., Wellcome Trust Case Control Consortium, Franklyn, J. A., Todd, J. A. & Gough, S. C. L. Seven newly identified loci for autoimmune thyroid disease. *Hum. Mol. Genet.* **21,** 5202–5208 (2012).
174. Simeonov, D. R., Gowen, B. G., Boontanrart, M., Roth, T. L., Gagnon, J. D., Mumbach, M. R., Satpathy, A. T., Lee, Y., Bray, N. L., Chan, A. Y., Lituiev, D. S., Nguyen, M. L., Gate, R. E., Subramaniam, M., Li, Z., Woo, J. M., Mitros, T., Ray, G. J., Curie, G. L., Naddaf, N., Chu, J. S., Ma, H., Boyer, E., Van Gool, F., Huang, H., Liu, R., Tobin, V. R., Schumann, K., Daly, M. J., Farh, K. K., Ansel, K. M., Ye, C. J., Greenleaf, W. J., Anderson, M. S., Bluestone, J. A., Chang, H. Y., Corn, J. E. & Marson, A. Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549,** 111–115 (2017).
175. Garg, G., Tyler, J. R., Yang, J. H. M., Cutler, A. J., Downes, K., Pekalski, M., Bell, G. L., Nutland, S., Peakman, M., Todd, J. A., Wicker, L. S. & Tree, T. I. M. Type 1

Diabetes-Associated IL2RA Variation Lowers IL-2 Signaling and Contributes to Diminished CD4+CD25+ Regulatory T Cell Function. *J. Immunol.* **188,** 4644–4653 (2012).

176. Orrù, V., Steri, M., Sole, G., Sidore, C., Virdis, F., Dei, M., Lai, S., Zoledziewska, M., Busonero, F., Mulas, A., Floris, M., Mentzen, W. I., Urru, S. A. M., Olla, S., Marongiu, M., Piras, M. G., Lobina, M., Maschio, A., Pitzalis, M., Urru, M. F., Marcelli, M., Cusano, R., Deidda, F., Serra, V., Oppo, M., Pilu, R., Reinier, F., Berutti, R., Pireddu, L., Zara, I., Porcu, E., Kwong, A., Brennan, C., Tarrier, B., Lyons, R., Kang, H. M., Uzzau, S., Atzeni, R., Valentini, M., Firinu, D., Leoni, L., Rotta, G., Naitza, S., Angius, A., Congia, M., Whalen, M. B., Jones, C. M., Schlessinger, D., Abecasis, G. R., Fiorillo, E., Sanna, S. & Cucca, F. Genetic Variants Regulating Immune Cell Levels in Health and Disease. *Cell* **155,** 242–256 (2013).
177. Zhernakova, D. V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.-J., Bonder, M. J., van Rooij, J., Verkerk, M., Jhamai, P. M., Moed, M., Kielbasa, S. M., Bot, J., Nooren, I., Pool, R., van Dongen, J., Hottenga, J. J., Stehouwer, C. D. A., van der Kallen, C. J. H., Schalkwijk, C. G., Zhernakova, A., Li, Y., Tigchelaar, E. F., de Klein, N., Beekman, M., Deelen, J., van Heemst, D., van den Berg, L. H., Hofman, A., Uitterlinden, A. G., van Greevenbroek, M. M. J., Veldink, J. H., Boomsma, D. I., van Duijn, C. M., Wijmenga, C., Slagboom, P. E., Swertz, M. A., Isaacs, A., van Meurs, J. B. J., Jansen, R., Heijmans, B. T., 't Hoen, P. A. C. & Franke, L. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49,** 139–145 (2017).
178. Bonder, M. J., Luijk, R., Zhernakova, D. V., Moed, M., Deelen, P., Vermaat, M., van

Iterson, M., van Dijk, F., van Galen, M., Bot, J., Slieker, R. C., Jhamai, P. M., Verbiest, M., Suchiman, H. E. D., Verkerk, M., van der Breggen, R., van Rooij, J., Lakenberg, N., Arindrarto, W., Kielbasa, S. M., Jonkers, I., van 't Hof, P., Nooren, I., Beekman, M., Deelen, J., van Heemst, D., Zhernakova, A., Tigchelaar, E. F., Swertz, M. A., Hofman, A., Uitterlinden, A. G., Pool, R., van Dongen, J., Hottenga, J. J., Stehouwer, C. D. A., van der Kallen, C. J. H., Schalkwijk, C. G., van den Berg, L. H., van Zwet, E. W., Mei, H., Li, Y., Lemire, M., Hudson, T. J., the BIOS Consortium, Slagboom, P. E., Wijmenga, C., Veldink, J. H., van Greevenbroek, M. M. J., van Duijn, C. M., Boomsma, D. I., Isaacs, A., Jansen, R., van Meurs, J. B. J., 't Hoen, P. A. C., Franke, L. & Heijmans, B. T. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49,** 131–138 (2017).

179. Ye, C. J., Feng, T., Kwon, H.-K., Raj, T., Wilson, M. T., Asinovski, N., McCabe, C., Lee, M. H., Frohlich, I., Paik, H., Zaitlen, N., Hacohen, N., Stranger, B., De Jager, P., Mathis, D., Regev, A. & Benoist, C. Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345,** 1254665 (2014).

180. Huang, J., Ellinghaus, D., Franke, A., Howie, B. & Li, Y. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur. J. Hum. Genet. EJHG* **20,** 801–805 (2012).

181. Hinks, A., Cobb, J., Marion, M. C., Prahalad, S., Sudman, M., Bowes, J., Martin, P., Comeau, M. E., Sajuthi, S., Andrews, R., Brown, M., Chen, W.-M., Concannon, P., Deloukas, P., Edkins, S., Eyre, S., Gaffney, P. M., Guthery, S. L., Guthridge, J. M., Hunt, S. E., James, J. A., Keddache, M., Moser, K. L., Nigrovic, P. A., Onengut-Gumuscu, S., Onslow, M. L., Rosé, C. D., Rich, S. S., Steel, K. J. A., Wakeland, E. K., Wallace, C. A., Wedderburn, L. R., Woo, P., Boston Children's JIA

Registry, British Society of Paediatric and Adolescent Rheumatology (BSPAR) Study Group, Childhood Arthritis Prospective Study (CAPS), Childhood Arthritis Response to Medication Study (CHARMS), German Society for Pediatric Rheumatology (GKJR), JIA Gene Expression Study, NIAMS JIA Genetic Registry, Treat Study, United Kingdom Juvenile Idiopathic Arthritis Genetics Consortium (UKJIAGC), Bohnsack, J. F., Haas, J. P., Glass, D. N., Langefeld, C. D., Thomson, W. & Thompson, S. D. Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat. Genet.* **45,** 664–669 (2013).

182. Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y. & Cherry, J. M. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46,** D794–D801 (2018).

183. Soderquest, K., Hertweck, A., Giambartolomei, C., Henderson, S., Mohamed, R., Goldberg, R., Perucha, E., Franke, L., Herrero, J., Plagnol, V., Jenner, R. G. & Lord, G. M. Genetic variants alter T-bet binding and gene expression in mucosal inflammatory disease. *PLOS Genet.* **13,** e1006587 (2017).

184. Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M. & Snyder, M. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22,** 1790–1797 (2012).

185. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked

variants. *Nucleic Acids Res.* **40,** D930–D934 (2012).

186. Li, P., Mitra, S., Spolski, R., Oh, J., Liao, W., Tang, Z., Mo, F., Li, X., West, E. E., Gromer, D., Lin, J.-X., Liu, C., Ruan, Y. & Leonard, W. J. STAT5-mediated chromatin interactions in superenhancers activate IL-2 highly inducible genes: Functional dissection of the Il2ra gene locus. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 12111–12119 (2017).

187. Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. & Young, R. A. Transcriptional super-enhancers connected to cell identity and disease. *Cell* **155,** (2013).

188. Korunes, K. L. & Noor, M. A. F. Gene conversion and linkage: effects on genome evolution and speciation. *Mol. Ecol.* **26,** 351–364 (2017).

189. Kamanaka, M., Rainbow, D., Schuster-Gossler, K., Eynon, E. E., Chervonsky, A. V., Wicker, L. S. & Flavell, R. A. Amino acid polymorphisms altering the glycosylation of IL-2 do not protect from type 1 diabetes in the NOD mouse. *Proc. Natl. Acad. Sci.* **106,** 11236–11240 (2009).

190. Rainbow, D. B., Pekalski, M., Cutler, A. J., Burren, O., Walker, N., Todd, J. A., Wallace, C. & Wicker, L. S. A rare IL2RA haplotype identifies SNP rs61839660 as causal for autoimmunity. *bioRxiv* 108126 (2017). doi:10.1101/108126

191. Burren, O. S., Rubio García, A., Javierre, B.-M., Rainbow, D. B., Cairns, J., Cooper, N. J., Lambourne, J. J., Schofield, E., Castro Dopico, X., Ferreira, R. C., Coulson, R., Burden, F., Rowlston, S. P., Downes, K., Wingett, S. W., Frontini, M., Ouwehand, W. H., Fraser, P., Spivakov, M., Todd, J. A., Wicker, L. S., Cutler, A. J. & Wallace, C. Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biol.* **18,** 165 (2017).

192. Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., Cairns, J., Wingett, S. W., Várnai, C., Thiecke, M. J., Burden, F., Farrow, S., Cutler, A. J., Rehnström, K., Downes, K., Grassi, L., Kostadima, M., Freire-Pritchett, P., Wang, F., Martens, J. H., Kim, B., Sharifi, N., Janssen-Megens, E. M., Yaspo, M.-L., Linser, M., Kovacsovics, A., Clarke, L., Richardson, D., Datta, A., Flicek, P., Stunnenberg, H. G., Todd, J. A., Zerbino, D. R., Stegle, O., Ouwehand, W. H., Frontini, M., Wallace, C., Spivakov, M. & Fraser, P. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167,** 1369-1384.e19 (2016).

193. Ahola-Olli, A. V., Würtz, P., Havulinna, A. S., Aalto, K., Pitkänen, N., Lehtimäki, T., Kähönen, M., Lyytikäinen, L.-P., Raitoharju, E., Seppälä, I., Sarin, A.-P., Ripatti, S., Palotie, A., Perola, M., Viikari, J. S., Jalkanen, S., Maksimow, M., Salomaa, V., Salmi, M., Kettunen, J. & Raitakari, O. T. Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *Am. J. Hum. Genet.* **100,** 40–50 (2017).

194. Calderon, D., Nguyen, M. L. T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J. V., Knowles, D. A., Gao, Z., Blaeschke, F., Parent, A. V., Burt, T. D., Anderson, M. S., Criswell, L. A., Greenleaf, W. J., Marson, A. & Pritchard, J. K. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.* 1–12 (2019). doi:10.1038/s41588-019-0505-9

195. Zhou, D., Li, Z., Yu, D., Wan, L., Zhu, Y., Lai, M. & Zhang, D. Polymorphisms involving gain or loss of CpG sites are significantly enriched in trait-associated SNPs. *Oncotarget* **6,** 39995–40004 (2015).

196. Roth, T. L., Puig-Saus, C., Yu, R., Shifrut, E., Carnevale, J., Li, P. J., Hiatt, J., Saco, J., Krystofinski, P., Li, H., Tobin, V., Nguyen, D. N., Lee, M. R., Putnam, A. L., Ferris, A. L., Chen, J. W., Schickel, J.-N., Pellerin, L., Carmody, D., Alkorta-Aranburu, G., Del Gaudio, D., Matsumoto, H., Morell, M., Mao, Y., Cho, M., Quadros, R. M., Gurumurthy, C. B., Smith, B., Haugwitz, M., Hughes, S. H., Weissman, J. S., Schumann, K., Esensten, J. H., May, A. P., Ashworth, A., Kupfer, G. M., Greeley, S. A. W., Bacchetta, R., Meffre, E., Roncarolo, M. G., Romberg, N., Herold, K. C., Ribas, A., Leonetti, M. D. & Marson, A. Reprogramming human T cell function and specificity with non-viral genome targeting. *Nature* **559,** 405–409 (2018).

197. Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *bioRxiv* 416610 (2019). doi:10.1101/416610

198. McVean, G. What drives recombination hotspots to repeat DNA in humans? *Philos. Trans. R. Soc. B Biol. Sci.* **365,** 1213–1218 (2010).

199. Fast Computation and Applications of Genome Mappability. at <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0030377>

200. Wang, Y. J., Schug, J., Won, K. J., Liu, C., Naji, A., Avrahami, D., Golson, M. L. & Kaestner, K. H. Single-Cell Transcriptomics of the Human Endocrine Pancreas. *Diabetes* **65,** 3028–3038 (2016).

201. Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., Kycia, I., Robson, P. & Stitzel, M. L. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **27,** 208–222 (2017).

202. Duhen, T., Duhen, R., Lanzavecchia, A., Sallusto, F. & Campbell, D. J. Functionally

distinct subsets of human FOXP3+ Treg cells that phenotypically mirror effector Th cells. *Blood* **119,** 4430–4440 (2012).

203. Bielekova, B., Catalfamo, M., Reichert-Scrivner, S., Packer, A., Cerna, M., Waldmann, T. A., McFarland, H., Henkart, P. A. & Martin, R. Regulatory CD56(bright) natural killer cells mediate immunomodulatory effects of IL-2Ralpha-targeted therapy (daclizumab) in multiple sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 5941–5946 (2006).

204. Martin, J. F., Perry, J. S. A., Jakhete, N. R., Wang, X. & Bielekova, B. An IL-2 Paradox: Blocking CD25 on T Cells Induces IL-2–Driven Activation of CD56bright NK Cells. *J. Immunol.* **185,** 1311–1320 (2010).

205. den Braber, I., Mugwagwa, T., Vrisekoop, N., Westera, L., Mögling, R., Bregje de Boer, A., Willems, N., Schrijver, E. H. R., Spierenburg, G., Gaiser, K., Mul, E., Otto, S. A., Ruiter, A. F. C., Ackermans, M. T., Miedema, F., Borghans, J. A. M., de Boer, R. J. & Tesselaar, K. Maintenance of Peripheral Naive T Cells Is Sustained by Thymus Output in Mice but Not Humans. *Immunity* **36,** 288–297 (2012).

206. Brisslert, M., Bokarewa, M., Larsson, P., Wing, K., Collins, L. V. & Tarkowski, A. Phenotypic and functional characterization of human CD25+ B cells. *Immunology* **117,** 548–557 (2006).

207. Amu, S., Tarkowski, A., Dörner, T., Bokarewa, M. & Brisslert, M. The Human Immunomodulatory CD25+ B Cell Population belongs to the Memory B Cell Pool. *Scand. J. Immunol.* **66,** 77–86 (2007).

208. Lund, F. E. Cytokine-producing B lymphocytes—key regulators of immunity. *Curr. Opin. Immunol.* **20,** 332–338 (2008).

209. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F. & Parkinson, H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).

210. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* (2019). doi:10.1038/s41588-019-0404-0

211. Zhu, Z., Lee, P. H., Chaffin, M. D., Chung, W., Loh, P.-R., Lu, Q., Christiani, D. C. & Liang, L. A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. *Nat. Genet.* **50,** 857–864 (2018).

212. Johansson, Å., Rask-Andersen, M., Karlsson, T. & Ek, W. E. Genome-wide association analysis of 350,000 Caucasians from the UK Biobank identifies novel loci for asthma, hay fever and eczema. *bioRxiv* 195933 (2019). doi:10.1101/195933

213. Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., Essers, J., Mitrovic, M., Ning, K., Cleynen, I., Theatre, E., Spain, S. L., Raychaudhuri, S., Goyette, P., Wei, Z., Abraham, C., Achkar, J.-P., Ahmad, T., Amininejad, L., Ananthakrishnan, A. N., Andersen, V., Andrews, J. M., Baidoo, L., Balschun, T., Bampton, P. A., Bitton, A., Boucher, G., Brand, S., Büning, C., Cohain, A., Cichon, S., D'Amato, M., De Jong, D., Devaney, K. L., Dubinsky, M., Edwards, C., Ellinghaus, D., Ferguson, L. R., Franchimont, D., Fransen, K., Gearry, R., Georges, M., Gieger, C., Glas, J., Haritunians, T., Hart, A., Hawkey, C., Hedl, M., Hu, X., Karlsen, T. H., Kupcinskas, L., Kugathasan, S., Latiano, A., Laukens, D., Lawrance, I. C., Lees, C. W., Louis, E.,

Mahy, G., Mansfield, J., Morgan, A. R., Mowat, C., Newman, W., Palmieri, O., Ponsioen, C. Y., Potocnik, U., Prescott, N. J., Regueiro, M., Rotter, J. I., Russell, R. K., Sanderson, J. D., Sans, M., Satsangi, J., Schreiber, S., Simms, L. A., Sventoraityte, J., Targan, S. R., Taylor, K. D., Tremelling, M., Verspaget, H. W., De Vos, M., Wijmenga, C., Wilson, D. C., Winkelmann, J., Xavier, R. J., Zeissig, S., Zhang, B., Zhang, C. K., Zhao, H., International IBD Genetics Consortium (IIBDGC), Silverberg, M. S., Annese, V., Hakonarson, H., Brant, S. R., Radford-Smith, G., Mathew, C. G., Rioux, J. D., Schadt, E. E., Daly, M. J., Franke, A., Parkes, M., Vermeire, S., Barrett, J. C. & Cho, J. H. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).

214. Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P., Zhernakova, A., Stahl, E., Viatte, S., McAllister, K., Amos, C. I., Padyukov, L., Toes, R. E. M., Huizinga, T. W. J., Wijmenga, C., Trynka, G., Franke, L., Westra, H.-J., Alfredsson, L., Hu, X., Sandor, C., de Bakker, P. I. W., Davila, S., Khor, C. C., Heng, K. K., Andrews, R., Edkins, S., Hunt, S. E., Langford, C., Symmons, D., Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate, Wellcome Trust Case Control Consortium, Concannon, P., Onengut-Gumuscu, S., Rich, S. S., Deloukas, P., Gonzalez-Gay, M. A., Rodriguez-Rodriguez, L., Ärlsetig, L., Martin, J., Rantapää-Dahlqvist, S., Plenge, R. M., Raychaudhuri, S., Klareskog, L., Gregersen, P. K. & Worthington, J. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44,** 1336–1340 (2012).

215. Petukhova, L., Duvic, M., Hordinsky, M., Norris, D., Price, V., Shimomura, Y., Kim, H., Singh, P., Lee, A., Chen, W. V., Meyer, K. C., Paus, R., Jahoda, C. A. B., Amos,

C. I., Gregersen, P. K. & Christiano, A. M. Genome-wide association study in alopecia areata implicates both innate and adaptive immunity. *Nature* **466,** 113–117 (2010).

216. Rubin, L. A., Kurman, C. C., Fritz, M. E., Biddison, W. E., Boutin, B., Yarchoan, R. & Nelson, D. L. Soluble interleukin 2 receptors are released from activated human lymphoid cells in vitro. *J. Immunol.* **135,** 3172–3177 (1985).

217. De Paiva, C. S., Yoon, K.-C., Pangelinan, S. B., Pham, S., Puthenparambil, L. M., Chuang, E. Y., Farley, W. J., Stern, M. E., Li, D.-Q. & Pflugfelder, S. C. Cleavage of functional IL-2 receptor alpha chain (CD25) from murine corneal and conjunctival epithelia by MMP-9. *J. Inflamm. Lond. Engl.* **6,** 31 (2009).

218. Rubin, L. A., Galli, F., Greene, W. C., Nelson, D. L. & Jay, G. The molecular basis for the generation of the human soluble interleukin 2 receptor. *Cytokine* **2,** 330–336 (1990).

219. Downes, K., Marcovecchio, M. L., Clarke, P., Cooper, J. D., Ferreira, R. C., Howson, J. M. M., Jolley, J., Nutland, S., Stevens, H. E., Walker, N. M., Wallace, C., Dunger, D. B. & Todd, J. A. Plasma concentrations of soluble IL-2 receptor α (CD25) are increased in type 1 diabetes and associated with reduced C-peptide levels in young patients. *Diabetologia* **57,** 366–372 (2014).

220. Hayden, A., Lin, M., Park, S., Pudek, M., Schneider, M., Jordan, M. B., Mattman, A. & Chen, L. Y. C. Soluble interleukin-2 receptor is a sensitive diagnostic test in adult HLH. *Blood Adv.* **1,** 2529–2534 (2017).

221. Kniep, E. M., Strelow, I. & Lohmann-Matthes, M. L. The monocyte interleukin-2 receptor light chain: production of cell-associated and soluble interleukin-2 receptor by monocytes. *Immunology* **75,** 299–304 (1992).

222. Holter, W., Goldman, C. K., Casabo, L., Nelson, D. L., Greene, W. C. & Waldmann, T. A. Expression of functional IL 2 receptors by lipopolysaccharide and interferon-gamma stimulated human monocytes. *J. Immunol.* **138,** 2917–2922 (1987).

223. Schwartz, A. M., Demin, D. E., Vorontsov, I. E., Kasyanov, A. S., Putlyaeva, L. V., Tatosyan, K. A., Kulakovskiy, I. V. & Kuprash, D. V. Multiple single nucleotide polymorphisms in the first intron of the IL2RA gene affect transcription factor binding and enhancer activity. *Gene* **602,** 50–56 (2017).

224. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., Consortium, T. G. A. D., Neale, B. M., Daly, M. J. & MacArthur, D. G. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019). doi:10.1101/531210

225. Ferreira, M. A. R., Mathur, R., Vonk, J. M., Szwajda, A., Brumpton, B., Granell, R., Brew, B. K., Ullemar, V., Lu, Y., Jiang, Y., Magnusson, P. K. E., Karlsson, R., Hinds, D. A., Paternoster, L., Koppelman, G. H. & Almqvist, C. Genetic Architectures of

Childhood- and Adult-Onset Asthma Are Partly Distinct. *Am. J. Hum. Genet.* **104,** 665–684 (2019).

226. Tian, C., Hromatka, B. S., Kiefer, A. K., Eriksson, N., Noble, S. M., Tung, J. Y. & Hinds, D. A. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun.* **8,** 1–13 (2017).

227. Durda, P., Sabourin, J., Lange, E. M., Nalls, M. A., Mychaleckyj, J. C., Jenny, N. S., Li, J., Walston, J., Harris, T. B., Psaty, B. M., Valdar, W., Liu, Y., Cushman, M., Reiner, A. P., Tracy, R. P. & Lange, L. A. Plasma Levels of Soluble Interleukin-2 Receptor α: Associations With Clinical Cardiovascular Events and Genome-Wide Association Scan. *Arterioscler. Thromb. Vasc. Biol.* **35,** 2246–2253 (2015).

228. Wang, X., He, L., Goggin, S. M., Saadat, A., Wang, L., Sinnott-Armstrong, N., Claussnitzer, M. & Kellis, M. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.* **9,** 1–15 (2018).

229. Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., Shetty, A. C., Blackwell, T. W., Wong, Q., Aguet, F., Albert, C., Alonso, A., Ardlie, K. G., Aslibekyan, S., Auer, P. L., Barnard, J., Barr, R. G., Becker, L. C., Beer, R. L., Benjamin, E. J., Bielak, L. F., Blangero, J., Boehnke, M., Bowden, D. W., Brody, J. A., Burchard, E. G., Cade, B. E., Casella, J. F., Chalazan, B., Chen, Y.-D. I., Cho, M. H., Choi, S. H., Chung, M. K., Clish, C. B., Correa, A., Curran, J. E., Custer, B., Darbar, D., Daya, M., Andrade, M. de, DeMeo, D. L.,

Dutcher, S. K., Ellinor, P. T., Emery, L. S., Fatkin, D., Forer, L., Fornage, M., Franceschini, N., Fuchsberger, C., Fullerton, S. M., Germer, S., Gladwin, M. T., Gottlieb, D. J., Guo, X., Hall, M. E., He, J., Heard-Costa, N. L., Heckbert, S. R., Irvin, M. R., Johnsen, J. M., Johnson, A. D., Kardia, S. L. R., Kelly, T., Kelly, S., Kenny, E. E., Kiel, D. P., Klemmer, R., Konkle, B. A., Kooperberg, C., Köttgen, A., Lange, L. A., Lasky-Su, J., Levy, D., Lin, X., Lin, K.-H., Liu, C., Loos, R. J. F., Garman, L., Gerszten, R., Lubitz, S. A., Lunetta, K. L., Mak, A. C. Y., Manichaikul, A., Manning, A. K., Mathias, R. A., McManus, D. D., McGarvey, S. T., Meigs, J. B., Meyers, D. A., Mikulla, J. L., Minear, M. A., Mitchell, B., Mohanty, S., Montasser, M. E., Montgomery, C., Morrison, A. C., Murabito, J. M., Natale, A., Natarajan, P., Nelson, S. C., North, K. E., O'Connell, J. R., Palmer, N. D., Pankratz, N., Peloso, G. M., Peyser, P. A., Post, W. S., Psaty, B. M., Rao, D. C., Redline, S., Reiner, A. P., Roden, D., Rotter, J. I., Ruczinski, I., Sarnowski, C., Schoenherr, S., Seo, J.-S., Seshadri, S., Sheehan, V. A., Shoemaker, M. B., Smith, A. V., Smith, N. L., Smith, J. A., Sotoodehnia, N., Stilp, A. M., Tang, W., Taylor, K. D., Telen, M., Thornton, T. A., Tracy, R. P., Berg, D. J. V. D., Vasan, R. S., Viaud-Martinez, K. A., Vrieze, S., Weeks, D. E., Weir, B. S., Weiss, S. T., Weng, L.-C., Willer, C. J., Zhang, Y., Zhao, X., Arnett, D. K., Ashley-Koch, A. E., Barnes, K. C., Boerwinkle, E., Gabriel, S., Gibbs, R., Rice, K. M., Rich, S. S., Silverman, E., Qasba, P., Gan, W., Trans-Omics for Precision Medicine (TOPMed) Program, Topm. P. G. W. G., Papanicolaou, G. J., Nickerson, D. A., Browning, S. R., Zody, M. C., Zöllner, S., Wilson, J. G., Cupples, L. A., Laurie, C. C., Jaquish, C. E., Hernandez, R. D., O'Connor, T. D. & Abecasis, G. R. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv* 563866 (2019). doi:10.1101/563866

230. The “All of Us” Research Program. *N. Engl. J. Med.* **381,** 668–676 (2019).

231. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. *Distrib. Author* (2005).

232. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, and D. The Human Genome Browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

233. Zheng, J., Ibrahim, S., Petersen, F. & Yu, X. Meta-analysis reveals an association of PTPN22 C1858T with autoimmune diseases, which depends on the localization of the affected tissue. *Genes Immun.* **13,** 641–652 (2012).

234. Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A. & Flicek, P. Ensembl 2018. *Nucleic Acids Res.* **46,** D754–D761 (2018).

235. Todd, J. A., Mijovic, C., Fletcher, J., Jenkins, D., Bradwell, A. R. & Barnett, A. H. Identification of susceptibility loci for insulin-dependent diabetes mellitus by trans-racial gene mapping. *Nature* **338,** 587 (1989).

236. Diversity matters. *Nat. Rev. Genet.* **20,** 495–495 (2019).

237. Fan, S., Kelly, D. E., Beltrame, M. H., Hansen, M. E. B., Mallick, S., Ranciaro, A.,

Hirbo, J., Thompson, S., Beggs, W., Nyambo, T., Omar, S. A., Meskel, D. W., Belay, G., Froment, A., Patterson, N., Reich, D. & Tishkoff, S. A. African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol.* **20,** 82 (2019).

238. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51,** 1321–1329 (2019).