Interdisciplinary investigation of the balance between T cell subsets throughout life

Stephanie J. Lax, MSc.

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

Supervisors: Dr. Andrew Jackson, Dr. Ian Spendlove and Professor Uwe Aickelin

2013
To Matthew
Abstract

To understand how immune balance is affected in immunosenescence and discover opportunities for correction with immunotherapy, a cross-sectional study of pro- and anti-inflammatory CD4⁺ T cell subsets in the peripheral blood of donors of different ages was conducted. A whole blood assay using extra- and intracellular flow cytometry was designed to enumerate anti-inflammatory natural CD127lowCD25highFoxp3⁺ T regulatory cells (nTreg) and inducible IL-10⁺IFNγ⁻ T regulatory cells (iTreg) versus proinflammatory T helper 17 cells (Th17) and T helper 1 cells (Th1), both IL-10⁺ and IL-10⁻. The frequency of these cell types was also linked to functional measurements of IL-17, IL-10 and IFNγ in whole blood supernatants after an overnight stimulation. We tested the hypothesis that the balance between these subsets changes with older age. This work shows for the first time that iTreg increase relative to nTreg and Th17, but not Th1, with healthy ageing in man. We also identify that the previously reported increases in IL-10 levels relate to changes in the iTreg population. However, they did not increase relative to the Th1 subset.

During this project, scope for improvement in multicolour flow cytometry data analysis was identified, in order to minimise subjectivity and maximise efficiency of nTreg enumeration. The hypothesis that automating nTreg analysis with the SamSPECTRAL algorithm is superior to baseline k-means clustering and traditional manual gating was investigated. SamSPECTRAL was qualitatively better than k-means in clustering nTreg from flow cytometry data containing overlapping, non-spherical clusters with different densities, and was more objective than traditional manual gating. However, across 90 data files, an optimal solution was not always achieved, and statistical measures of cluster validity did not support the visual evidence that SamSPECTRAL better captured the natural structure of the data. A novel extension of SamSPECTRAL to include an automated elliptical gating step allowed for
comparison of test and control datasets to correct nTreg frequency measurements for false positive events. As manual inspection of each solution was required, however, the ability to entirely automate nTreg analysis was prevented.

We hope that this work will encourage further collaboration between the disciplines of immunology and computer science to advance the study of cancer and ageing.
Acknowledgements

This PhD experience has been hugely rewarding because of the interesting, supportive and enthusiastic people I have been privileged to work with. Firstly, my thanks go to Andrew Jackson, Poulam Patel, Ian Spendlove, Julie Greensmith and Uwe Aickelin for their supervision, mentoring and generous sharing of their breadth of experience.

I am profoundly grateful to the volunteers who donated time, travel expense, blood and personal information for this research. Also, I am grateful to the clinical research fellows who took the samples, whilst juggling their own busy projects. Specifically, I would like to thank Helen Angell, Hester Franks, Karen Robinson, Katie Cook, Darren Ketley and Adrian Robins who gave advice and reagents whilst developing the whole blood flow cytometry assay. I thank Anna Malecka, Qunwei Wang and Ruhcha Sutavani for providing pre-stimulated CD4\(^+\) lymphocytes, helping out with ELISAs and stimulations and for generally being a brilliant team to work with. At IMA, I also thank Daniele Soria, Jenna Reps and Chris Roadknight for answering my questions on data mining and simulation.

I have been particularly appreciative of the opportunities I have had to meet like-minded and fascinating people through meetings and conferences such as ICARIS, especially Kieran Alden, Mark Read, Richard Alun Williams and Jonathan Timmis. Most recently, I would like to thank Surajit Ray, Ryan Brinkman, Habil Zare, Mehrnoush Malek and Simon Crouch for always taking the time to reply to my emails, the benefit of their knowledge in data analysis, their willingness to help and their good humour. I look forward to further collaboration and friendship.

Lastly, although my words will not express what their unfailing belief and encouragement has meant to me, I thank my family. I thank Matt in particular, for putting up with the day-to-day life of living with a PhD student and still marrying me in the middle of it all.
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>nTreg</td>
<td>natural T regulatory cell</td>
<td>page 2</td>
</tr>
<tr>
<td>iTreg</td>
<td>induced T regulatory cell</td>
<td>page 2</td>
</tr>
<tr>
<td>Th17</td>
<td>T helper 17 cell</td>
<td>page 2</td>
</tr>
<tr>
<td>Th1</td>
<td>T helper 1 cell</td>
<td>page 2</td>
</tr>
<tr>
<td>DC</td>
<td>dendritic cell</td>
<td>page 4</td>
</tr>
<tr>
<td>MHC</td>
<td>major histocompatibility complex</td>
<td>page 4</td>
</tr>
<tr>
<td>TCR</td>
<td>T cell receptor</td>
<td>page 5</td>
</tr>
<tr>
<td>CTLA-4</td>
<td>cytotoxic T lymphocyte antigen 4</td>
<td>page 8</td>
</tr>
<tr>
<td>ROR</td>
<td>RAR-related orphan receptor</td>
<td>page 12</td>
</tr>
<tr>
<td>Treg</td>
<td>generic CD25+ T regulatory cell</td>
<td>page 12</td>
</tr>
<tr>
<td>OX-LDL</td>
<td>oxidised low-density lipoprotein</td>
<td>page 12</td>
</tr>
<tr>
<td>RANK-L</td>
<td>receptor activator of nuclear factor kappa-B ligand</td>
<td>page 12</td>
</tr>
<tr>
<td>CMV</td>
<td>cytomegalovirus</td>
<td>page 13</td>
</tr>
<tr>
<td>MDSCs</td>
<td>myeloid derived suppressor cells</td>
<td>page 13</td>
</tr>
<tr>
<td>IRP</td>
<td>immune risk profile</td>
<td>page 14</td>
</tr>
<tr>
<td>PHA</td>
<td>phytohaemagglutinin</td>
<td>page 25</td>
</tr>
<tr>
<td>poly(I:C)</td>
<td>polynosinic-polycytidylic acid</td>
<td>page 25</td>
</tr>
<tr>
<td>PBMCs</td>
<td>peripheral blood mononuclear cells</td>
<td>page 25</td>
</tr>
<tr>
<td>PMA</td>
<td>phorbol 12-myristate 13-acetate</td>
<td>page 25</td>
</tr>
<tr>
<td>CART</td>
<td>classification and regression trees</td>
<td>page 32</td>
</tr>
<tr>
<td>ANN</td>
<td>artificial neural networks</td>
<td>page 33</td>
</tr>
<tr>
<td>SVM</td>
<td>support vector machine</td>
<td>page 33</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
<td>page 34</td>
</tr>
<tr>
<td>FLAME</td>
<td>flow analysis with automated multivariate estimation</td>
<td>page 38</td>
</tr>
<tr>
<td>SPADE</td>
<td>spanning-tree progression analysis of density-normalized events</td>
<td>page 40</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>------------</td>
</tr>
<tr>
<td>APS</td>
<td>automatic population separator</td>
<td>page 43</td>
</tr>
<tr>
<td>SPICE</td>
<td>simplified presentation of incredibly complex evaluations</td>
<td>page 48</td>
</tr>
<tr>
<td>ASW</td>
<td>average silhouette width</td>
<td>page 53</td>
</tr>
<tr>
<td>FlowCAP</td>
<td>flow cytometry: critical assessment of population identification methods</td>
<td>page 55</td>
</tr>
<tr>
<td>ELISA</td>
<td>enzyme-linked immunosorbent assay</td>
<td>page 60</td>
</tr>
<tr>
<td>FACS</td>
<td>fluorescence-activated cell sorting</td>
<td>page 61</td>
</tr>
<tr>
<td>FMO</td>
<td>fluorescence minus one</td>
<td>page 61</td>
</tr>
<tr>
<td>CV</td>
<td>coefficient of variation</td>
<td>page 69</td>
</tr>
<tr>
<td>IQR</td>
<td>interquartile range</td>
<td>page 73</td>
</tr>
<tr>
<td>MFI</td>
<td>median fluorescence intensity</td>
<td>page 89</td>
</tr>
<tr>
<td>KS</td>
<td>Kolmogorov-Smirnov normality test</td>
<td>page 97</td>
</tr>
<tr>
<td>SW</td>
<td>Shapiro-Wilk normality test</td>
<td>page 97</td>
</tr>
<tr>
<td>EBV</td>
<td>Epstein-Barr virus</td>
<td>page 136</td>
</tr>
<tr>
<td>slanDCs</td>
<td>human 6-sulfo LacNAc dendritic cells</td>
<td>page 193</td>
</tr>
</tbody>
</table>
# Contents

Abstract i

Acknowledgements iii

List of abbreviations iv

Contents vi

## 1 Introduction

1.1 Hypothesis ............................................. 1
1.2 Aims .................................................... 1
1.3 Thesis organisation ................................. 2
1.4 Ageing and the human immune system ................. 3
  1.4.1 Ageing .............................................. 3
  1.4.2 Adaptive immunity ................................. 4
  1.4.3 CD4$^+$ T cell subsets and related cytokines ....... 8
1.5 Immunosenescence ................................... 13
  1.5.1 Key features ....................................... 13
  1.5.2 Subject selection in ageing research ............... 14
  1.5.3 The importance of context ......................... 15
  1.5.4 nTreg ............................................. 16
  1.5.5 iTreg, Th1 and Th17 ............................... 25
1.6 Rationale for studying multiple T cell subsets with age 27
1.7 Computational methods ............................... 28
  1.7.1 Introduction ....................................... 28
  1.7.2 Simulation ......................................... 28
  1.7.3 Data mining ....................................... 32
1.8 Flow cytometry data analysis .......................... 35
1.9 Specialised flow cytometry data clustering algorithms . 37
3 Development and validation of the whole blood assay

3.1 Introduction .......................................................... 83
3.2 Time course ELISAs .................................................. 83
3.3 Staining panel development and optimisation .................. 87
  3.3.1 Fix/perm buffer selection ........................................ 87
  3.3.2 Scatter characteristics of leukocyte lineages ................. 87
  3.3.3 Antibody titrations .............................................. 89
  3.3.4 RORγ(t) staining ................................................ 89
  3.3.5 Helios staining .................................................. 90
  3.3.6 Staining verification ............................................ 92
3.4 Non-parametric statistics for ageing study .................... 95
3.5 Typical staining examples ......................................... 95
3.6 ELISA error assessment and plate comparison ................. 102

4 Balance between T cell subsets with age

4.1 Introduction .......................................................... 109
4.2 Donor characteristics .............................................. 110
4.3 Increase in iTreg with age ......................................... 110
4.4 Decrease in Th17 cells with age .................................. 114
4.5 Th17:iTreg balance decreases with age ......................... 114
4.6 The source of secreted IL-10 is independent of Th1 cells .... 117
4.7 Increase in Th1 relative to Th17 cells with age ............... 117
4.8 Absolute number measurements with age ....................... 120
4.9 Correlation between flow cytometry and ELISA measurements 123
4.10 Age-stratified correlation analysis ............................... 123
4.11 Using a supervised learning technique to identify confounding factors .................................................. 128
4.12 Discussion .................................................................. 134

5 Investigation of SamSPECTRAL for automated nTreg enumeration

5.1 Introduction .......................................................... 138
5.2 Inadequacy of manual 2-dimensional gating for nTreg flow cytometry data analysis ........................................ 138
  5.2.1 Traditional gating strategies for nTreg interpreted differently lead to different results ................................. 138
  5.2.2 Poor reproducibility of manual gating even with a stringent rule-based strategy ......................................... 139
5.3 Cluster validity indices ............................................... 144
5.4 Comparison and optimisation of $k$-means and SamSPECTRAL performance on synthetic data .......................... 148
  5.4.1 SamSPECTRAL parameter optimisation for synthetic datasets ......................................................... 159
5.5 Comparison and optimisation of $k$-means and SamSPECTRAL performance on nTreg data ......................... 166
  5.5.1 SamSPECTRAL parameter optimisation for nTreg datasets ............................................................... 170
  5.5.2 Stability of SamSPECTRAL on nTreg data ................. 175
5.6 Comparison between SamSPECTRAL and manual gating . 177
5.7 Drawing hard partitions around SamSPECTRAL-generated clusters for multisample comparison .................. 179
5.8 Does SamSPECTRAL give a different conclusion to nTreg frequency in peripheral blood with age? ............ 179
5.9 Discussion ......................................................... 181

6 Discussion and future work 186
  6.1 General evaluation ............................................ 186
    6.1.1 Critique of ageing study ................................ 186
    6.1.2 Critique of SamSPECTRAL investigation ............. 189
    6.1.3 Comparison with other recent work ..................... 189
  6.2 Future directions ............................................ 192
    6.2.1 Improving the whole blood assay for routine immunophenotyping ............................................ 192
    6.2.2 Constrained spectral clustering ......................... 194
    6.2.3 Simulating the ageing immune system .................. 195
  6.3 Summary ....................................................... 196

Bibliography 198

R code vignette 227

List of Figures 246

List of Tables 250

Index 251
Chapter 1

Introduction

1.1 Hypothesis

There is a need to better understand how humans age in order to maximise healthy life expectancy and make provisions for caring for the elderly within an ageing population. Of particular interest is designing effective cancer immunotherapies for older people, as cancer is generally considered an age-related disease. However, the immune system, the very machinery immunotherapies attempt to manipulate in cancer treatment, is also thought to decline with age. The laboratory work reported in this thesis investigated blood samples from donors of different ages to test the hypothesis that changes in the balance between pro- and anti-inflammatory CD4$^+$ T cell subsets occur that may have implications on the way cancer immunotherapies should be designed.

During this work, scope for improvement using the computational technique of clustering in multicolour flow cytometry data analysis was identified, in order to minimise subjectivity and error, and to maximise efficiency. This thesis represents an interdisciplinary collaboration between cancer immunologists and computer scientists.

1.2 Aims

The aims of this work were both immunological and computational. Relating to the immunological, the aim was to conduct a cross-sectional study of pro- and anti-inflammatory CD4$^+$ T cell subsets in the peripheral blood of donors of different ages. Therefore a whole blood assay was
designed to enumerate anti-inflammatory natural T regulatory cells (nTreg, CD127\textsuperscript{low} CD25\textsuperscript{high} Foxp3\textsuperscript{+}) and inducible T regulatory cells (iTreg, IL-10\textsuperscript{+} IFN\textgamma\textsuperscript{−}) versus proinflammatory T helper 17 cells (Th17, IL-17A\textsuperscript{+}) and T helper 1 cells (Th1, IFN\textgamma\textsuperscript{+}, both IL-10\textsuperscript{+} and IL-10\textsuperscript{−}). The frequency of these cell types was also linked to functional measurements of IL-17, IL-10 and IFN\textgamma in whole blood supernatants after an overnight stimulation. Although each of these subsets had previously been investigated individually, changes to the ratios between these subsets is a more crucial indication of immune function in older age [197]. Thus we aimed to generate a novel dataset containing multiple measurements of these subsets from the same blood samples.

Relating to the computational, the aim was to identify a limitation of current routine laboratory practice that might be improved using computational methods. A particular limitation was found in the manual analysis of flow cytometry data, which is laborious, error-prone and poorly reproducible [257, 23]. nTreg raw flow cytometry data, especially, is very difficult to analyse objectively, and there is debate in the field of immunology about how best to enumerate nTreg [109, 184, 72, 258, 226, 105, 165, 72]. Within the CD4\textsuperscript{+} T cell compartment, the natural nTreg and non-nTreg populations are diffuse, non-spherical and overlapping, presenting challenges for both manual and computational approaches [128]. This work therefore aimed to find a method that could automate nTreg enumeration and give a more robust measurement of their frequency.

In future work we aim to increase the frequency of favourable clustering with the SamSPECTRAL algorithm, and to expand it for multisample comparison and more general use on flow cytometry datasets. We hope that this work will interest users of flow cytometry in the “wet-lab” environment and encourage deeper analysis of multicolour flow cytometry data than can be achieved with manual 2-dimensional gating.

1.3 Thesis organisation

The remainder of Chapter 1 expands on the motivation outlined and provides an introduction to ageing and immunosenescence. The best characterised changes to the immune system with older age are highlighted and the literature on T cell subsets throughout the lifetime is appraised, substantiating the argument for further experimental
investigation of T cell subsets in donors of different ages. This proceeds into a detailed literature review of currently available algorithms for automating flow cytometry data analysis.

Chapter 2 presents the materials and methods used in the whole blood flow cytometry assay for T cell subsets, and the results of developmental experiments are presented in Chapter 3. Chapter 4 then shows the results of a cross-sectional study of CD4$^+$ T cell subsets in donors of different ages. The need for improvement in nTreg data analysis is shown in Chapter 5 through a comparison of results drawn through different interpretations of the same manual gating strategy and measurements of inter- and intra-observer variation in nTreg enumeration. Then, a flow cytometry data-specific clustering algorithm, SamSPECTRAL, is compared to a basic benchmarking algorithm, k-means, for the ability to automatically enumerate nTreg. Preliminary data is shown to illustrate the feasibility of extending SamSPECTRAL for multisample comparison.

Chapter 6 then contains general discussions and evaluations of both the in vitro and in silico investigations outlined in this thesis. We compare our findings with the work of other groups and suggest how the whole blood assay might be improved for future work immunophenotyping blood samples from cancer patients. The scope for using simulation in the study of human ageing is also addressed.

1.4 Ageing and the human immune system

1.4.1 Ageing

In developed countries, the population aged over 85 has undergone a marked expansion in the last few decades and most children born in the UK post-2000 will live to be centenarians [59]. However, it has been argued that average life-expectancy has risen without a sufficient increase in disease-free life expectancy [15]. Clearly there are profound health, social and economic consequences of population maturation and thus research to “compress morbidity” [95] in older ages and maximise healthy life expectancy is key.

Although specific genetic polymorphisms have been associated with longevity, it is thought that a host of genetic and other systemic factors may accumulate over time and contribute small and overlapping alterations that predispose to longer or shorter life [273]. Of particular interest are the changes to the immune system which contribute to health
or disease in the elderly; the phenomenon of immunosenescence. For example, a combination of cytokine and chemokine alterations as opposed to a single change has been associated with atrophy in the aged brain [25]. When the immune system functions optimally, it protects the host from infection, clears dying or aberrant host cells and is tightly controlled to minimise bystander damage to normal host cells resulting from these processes. In older age, the regulation of these mechanisms may become relaxed leading to inadequate protection, failure to control tumour cell growth and increased inflammation and autoimmunity. The rest of this section describes general aspects of immunity, and we discuss why CD4$^+$ T cell subsets were the main focus of this work.

1.4.2 Adaptive immunity

“Immunity” is a collective name for the processes that protect an organism from illness. It includes barriers to the environment such as skin and mucous membranes, as well as intricate cellular and humoral cascades which control disease transversing or beginning within those barriers [140]. The internal mechanisms are loosely categorised into the innate and adaptive immune systems [35]. The innate immune system acts quickly against common molecular cues and thus prevents the majority of illnesses, and the adaptive immune system mobilises relatively slowly but delivers a highly specific attack on sources of illness that evade the innate immune response [35]. The adaptive immune system also has memory functionality; it will mount a specific and rapid immune response upon a second encounter with the same pathology [35].

The focus of this thesis is to test the hypothesis that CD4$^+$ T lymphocyte subpopulations change with age. Figure 1.1 shows CD4$^+$ T cell subsets at the centre of the innate and adaptive antitumour immune responses and is adapted from [163] with additional information from several other sources [237, 66, 195, 240, 152, 94, 283]. These cells recognise preprocessed peptide fragments of pathogen or tumour markers presented in major histocompatibility complexes (MHC) on antigen-presenting cells, such as dendritic cells (DC). Depending on the context of their stimulation, they coordinate the immune response, partly by secreting cytokines to inflame or mediate downstream processes [247]. A distinction is made between the Th2 subset, which chiefly controls the humoral immune system, and the Th1, Th17 and regulatory subsets which influence the cellular components. As tumour cells are altered host cells, unique or
aberrant intracellular proteins can generate novel peptide fragments for presentation in MHC class I and therefore recognition by CD8+ T cells [140]. However, tumour infiltration of regulatory immune cells associated with poor prognosis may prevent effective antitumour immunity [161, 215]. Figure 1.2 is an expansion of Figure 1.1 showing the relationship between and cytokines required for the differentiation of CD4+ T lymphocyte subsets [136, 251, 152, 278, 277, 272, 139]. CD4+ T cells are of particular interest in disease processes, as their reaction to an insult can divert the resulting immune response to either fight or ignore the trigger [197].

T cell precursors originate in the bone marrow and undergo development in the thymus [190]. Once they have undergone T cell receptor (TCR) gene rearrangement, acquire a low CD4 and CD8 expression and the majority commit to an α:β lineage, they exist for 3-4 days before apoptosing unless they receive survival signals, corresponding to TCR ligation and MAP kinase signalling [253]. In the thymic cortex, the immature T cell is exposed to cortical epithelial cells expressing self peptides within MHC molecules and is positively selected if its TCR weakly interacts with them [190]. In a process of negative selection, immature T cells with TCRs that associate too strongly with peptide:MHC complexes on antigen-presenting cells die by apoptosis [253]. However, some immature T cells with higher affinity for self peptide:MHC survive and give rise to the nTreg population [151]. Other T cells surviving both positive and negative selection differentiate to CD4+ or CD8+ lineages and enter the periphery. Th17, Th1 and iTreg all develop from the resulting CD4+ naïve precursors, but under different cytokine conditions, which will be discussed in more detail [74].

The challenge for vaccine therapies is often to break host tolerance [103]. Several immunotherapy strategies have considered how to polarise the CD4+ T cell response in order to maximise efficacy. For example, adoptive DC transfer immunotherapeutic strategies under development aim to promote antitumour T cell responses whilst inhibiting nTreg. We have recently demonstrated the ability to alter the balance of IL-12 and IL-10 secretion by myeloid DC favourably in vitro by selective inhibition of p38 alongside TLR stimulation [92]. In addition, we have shown an ability to prime Th17 responses by treating myeloid DC with the PI3 kinase inhibitor wortmannin, encouraging IL-23 secretion without an increase in IL-12p70 [278].
Figure 1.1: Schematic of CD4+ T cell subsets at the centre of the innate and adaptive antitumour immune responses.
Figure 1.2: *Schematic of the relationship between CD4+ T cell subsets.*
The selection of adjuvants administered alongside active and passive immunotherapies should also be considered. In a murine model where CD4+ T cells from young and old animals were adoptively transferred into a CD4 knockout prior to pigeon cytochrome c immunisation, using TNF-α, IL-1 and IL-6 as a combined adjuvant recovered the responsiveness of the aged CD4+ T cells [180]. Gravekamp also suggests injecting IL-7 into the tumour site where possible to aid recruitment of T cells [107]. Such a strategy may preferentially recruit effector T cells as nTreg express lower levels of the IL-7 receptor [72, 165].

Cytotoxic lymphodepletion has also been shown to favourably tip the balance between regulatory and effector T cell subsets in murine models [17]. Following radiotherapy, resistant regulatory cells prevented adoptively transferred naïve T cells from mounting an effective immune response [17]. However, upon Treg depletion, tumour-specific immune responses were effectively primed [17]. Also, the human monoclonal anti-CTLA-4 antibody ipilimumab, used primarily for the treatment of metastatic melanoma, acts by preventing CTLA-4-mediated attenuation of T cell activation [161]. CTLA-4 is also constitutively expressed on nTreg [241]. In combination with GM-CSF it has also been shown in the B16/BL6 murine system to have an affect on the balance between regulatory and effector T cell subsets, markedly increasing the intratumoural ratio of effector CD4+ and CD8+ T cells to Tregs [215]. The activity of CD8+ T cells was also increased, correlating with tumour rejection [215]. In human studies, using ipilimumab in combination therapies has been shown to give better overall survival [161]. However, a number of adverse events relating to manipulation of the immune system can occur, including rashes, hepatitis, pancreatitis, lymphadenopathy, neuropathies, nephritis and colitis, which may correlate with antitumour response [280]. This supports an argument that maintenance of immune balance is crucial and that therapeutic modulation requires careful design, careful dosing and vigilant monitoring.

1.4.3 CD4+ T cell subsets and related cytokines

As they have an important role in the cellular and more general antitumour immune response, the anti-inflammatory T cell subsets this study focused on were nTreg and iTrég. The proinflammatory cells considered were the Th1 and Th17 populations. These subsets will be introduced in the rest of this section.
nTreg

nTreg make up 1-5% of human CD4+ lymphocytes [109]. nTreg are a heterogeneous subset and it is debated which are the most appropriate molecular markers to identify them [66]. Extracellularly, nTreg are classically defined as CD4+CD25^{high} [109]. CD25 is the α chain of the IL-2 receptor expressed on T cells [141], although it has been shown that some T cells expressing low levels of CD25 possess regulatory functionality and only upregulate CD25 upon activation [66]. Intracellularly, Foxp3 is currently the best functional marker for nTreg as it contributes to the control of CD25 and IL-2 gene expression [258, 226], but is also transiently expressed in activated T cells without suppressive function [226]. Conversely, it is not a reliable marker of iTreg which are also suppressive [225]. CD25 and Foxp3 are, however, the best available markers.

CD127 expression has recently been shown to be inversely correlated with Foxp3 expression and suppressive function in T helper cells, thus arguably, the absence of this surface marker indicates a population enriched for T regulatory cells [165, 72]. CD127 is the IL-7 receptor and IL-7 is considered to be an important survival signal [165, 211]. IL-7 is thought to synergise with IL-2 and IL-15 to encourage effector T cells to become long-lived memory cells [211]. A study in patients with rheumatoid arthritis, whose disease was well controlled, demonstrated that the amount of circulating IL-7 correlated with frequency of CD4+CD25^{high} nTreg [149]. CD127 expression has also been shown to clearly separate CD4+CD25^{high}Foxp3+ from CD4+CD25^{high}Foxp3− populations [184] and has not yet been used to measure nTreg frequency in donors of different ages. The robust and reproducible analysis of nTreg raw flow cytometry data still remains a challenging task, and we will discuss our efforts to advance this field in Chapter 5.

In particular circumstances in vivo, it is uncertain how nTreg exert their function [66]. nTreg can kill both CD4+ and CD8+ T cells via the perforin and granzyme mechanism in vitro and inhibit conventional T cells by CTLA-4 expression [66]. By cognate interaction, nTreg tolerise other T cells [66], and nTreg also consume large amounts of the T cell growth factor IL-2, outcompeting their non-regulatory counterparts [195, 240]. A variety of other potential functions are also outlined in a comprehensive review by Schmetterer and colleagues [241].

nTreg cells prevent cell-mediated autoimmunity in a number of animal models [271]. Autoimmune diseases, chronic inflammatory diseases and
cancers are associated with nTreg as they temper autoimmune and pathogen-related immune function [109, 66, 204]. These conditions are associated with older age [109]. Therefore, we hypothesise that nTreg frequency and function changes with age. Later in this chapter the detailed evidence relating to nTreg ageing is discussed.

iTreg and Th1

Many immune cell subsets secrete the anti-inflammatory cytokine IL-10 which suppresses various processes in order to dampen unwanted immune responses [237]. We have previously shown that naïve CD4+ T cells may be differentiated to secrete very large quantities of IL-10 depending on the context of their stimulation [41]. The balance between cytokines is important, for example the relative secretion of IL-10 and IFNγ by Th1 cells determines whether an immune response clears infections or tumours or allows them to persist [197]. In patients admitted to hospital with *Staphylococcus aureus* bacteraemia, high serum IL-10 is associated with mortality [227]. In patients admitted to hospital with pneumococcal infections, a prolonged increase in serum IL-10 is associated with patient age, suggesting inefficient clearance of the pathogen in older donors relative to young [37]. Interestingly, IL-10-deficient mice have improved survival from influenza, a pathogen causing high morbidity and mortality in the elderly, but are more likely to develop cardiac and vascular problems with older age [276, 224, 246]. However, too much IL-10 in the damaged vasculature of older people may impede healing processes, including cell proliferation [71]. This suggests that IL-10 has important impacts on disease progression in the elderly and it is necessary to maintain IL-10 levels within a narrow range. In agreement with this, a study of IL-10 genotypes showed that longevity was positively associated with genotypes corresponding to intermediate production as opposed to low or high [194]. In an independent clinical study, a low IL-10 secreting genotype was more frequent in patients diagnosed with acute myocardial infarction [160]. However, some centenarians have high levels of IL-10 and still resist various pathologies, including tumour development [232].

IFNγ, conversely, coordinates a variety of immune processes to encourage protection against bacteria, viruses and tumours, increasing both MHC class I and II presentation, cellular cytotoxicity by CD8+ T cells and NK cells, further proinflammatory cytokine secretion in a positive feedback loop and restricting Th2 development [243, 223]. Whilst
IFN-γ promotes antitumour immunity, it may also be directly responsible for immunoediting and evolution of a tumour, resulting in tumour escape [243]. Thus the frequency of iTreg (CD4+IFN-γ−IL-10+), Th1 cells (CD4+IFN-γ+, either IL-10− or IL-10+) and IL-10 secretion by whole blood cultures were also measured in this immunosenescence study of the balance between nTreg and Th17 cells.

Th17

In contrast to nTreg and iTreg, Th17 cells are proinflammatory, and amplify immune responses. They recruit and aggregate neutrophils and macrophages, and encourage other cells, including cells of the epithelium, endothelium and fibroblasts, to secrete inflammatory mediators [157]. Any disruption to the homeostatic balance that exists between nTreg and Th17 cells in favour of Th17 cells is now thought to contribute to inflammation and autoimmunity [162]. Secretion of IL-17 causes the abovementioned stromal cells to produce IL-6, TNF-α, IL-8 and GM-CSF [152]. These exacerbate the inflammatory response further by recruiting other T cells, dendritic cells and neutrophils [152]. Therefore, Th17 cells represent an important bridge between adaptive and innate arms of the immune system, and impairment of their function has been shown to result in increased susceptibility to some infectious diseases in animal models [63]. They also have both positive and detrimental influences on cancer development and progression [36, 145]. For example, IL-17 encourages tumour rejection, and we have developed mechanisms to control Th17 differentiation by manipulating IL-23 secretion by dendritic cells [278, 277]. However, our group also showed previously that IL-17 promotes the invasiveness of breast cancer cells in vitro [298]. IL-17 stimulates colorectal cancer cells to secrete VEGF, promoting angiogenesis [164] and also increases expression of MMP-9, implicated in metastasis [244, 58, 285]. In addition, alterations in Th17 cell phenotype and function may contribute to an increased risk of infectious disease in the elderly, which represents a significant proportion of disease burden in older people [152].

Balance

A balance exists between regulatory T cell populations and Th17 cells. Firstly, their differentiation pathways are related. IL-6 promotes differentiation to a Th17 phenotype, inhibiting the promotion of iTreg differentiation by TGF-β [139]. Their signature transcription factors are
also known to interact; the signature for Th17s, RORγ(t), has been shown to bind to the Foxp3 promoter region inhibiting its expression [39]. Foxp3 has also been shown to bind to RORc2 (the human orthologue of RORγ(t)) directly, inhibiting its function [39]. Also, a high proportion of Th17 cells are known to have previously expressed Foxp3 [297], which provides evidence for their plasticity and a likelihood that many of these T cell subsets are not lineage committed. This supports the notion that they can be adapted to their environment and self-maintain the balance of immune function and regulation to some extent.

An example disease of disruption to the balance between regulatory T cell subsets and Th17 cells is acute coronary syndrome. It has been demonstrated that patients with acute coronary syndrome have more Th17 cells, RORγ(t) transcription factor expression and serum IL-17 with a corresponding decrease in CD25+ T regulatory cells (Treg), Foxp3 and IL-10 [157]. Oxidised low-density lipoprotein (OX-LDL) is involved in initiation and maintenance of atherosclerosis in a number of ways and appears to have some control over Treg balance with Th17 cells [157]. OX-LDL levels positively correlated with Th17 cell numbers and negatively correlated with Treg numbers, and OX-LDL actively reduced the number of Treg and increased Th17s [157]. Independently, high plasma IL-10 was associated with lower coronary artery calcification [93]. However, high plasma IL-10 has also been shown to be positively associated with cardiovascular disease risk [282]. Whether any change in balance between Treg and Th17 cells causes the disease or is a consequence of the disease process is unknown. Conditions such as rheumatoid arthritis, inflammatory bowel disease and multiple sclerosis may also be associated with disruption of the Treg and Th17 cell balance [275]. Animal studies have investigated IL-17 expression and found it to be increased in the coronary arteries of older animals compared to younger ones [69]. The shift to a proinflammatory phenotype occurs in the coronary arteries with advancing age [65]. Therefore a study of T cells throughout the lifetime is useful to other age-related diseases with an immune component.

Amongst other factors, IL-17 upregulates RANK-L (receptor activator of nuclear factor kappa-B ligand) involved in bone remodelling and osteoclast activation, therefore is implicated in osteoporosis [178]. However, it has been shown that IFNγ may prevent this [178], therefore considering the relationship between the proinflammatory Th1 and Th17 subtypes is also important.

Therapeutic options have been proposed to control Th17 differentiation,
such as retinoic acid to inhibit and dioxin to promote Th17 differentiation [139]. The anti-IL-6 receptor antibody therapy tocilizumab used for autoimmune conditions such as Crohn’s disease and rheumatoid arthritis has been shown to give clinical benefit, but it is not known whether the inhibition of Th17 differentiation is how it makes this contribution [139]. An important consideration here is how such therapies might affect the balance between Treg and Th17 cells, as both cells types are required for healthy ageing and it is the maintenance of their homeostasis that is important.

1.5 Immunosenescence

1.5.1 Key features

Ageing affects almost all arms of the innate and adaptive immune system and the best-characterised change is a decrease in naïve T cells relative to memory T cells, largely due to thymus involution post adolescence [34]. The lifetime dynamics of T cells are also thought to undergo changes, for example with altered rates of proliferation, apoptosis and activation-induced cell death [34]. Decreased T cell IL-2 production and receptor expression has also been noted [192].

There is also evidence in support of a change in inflammatory status, with cytokine profiles changing from Th1 to Th2, with decreased levels of IL-2 and increased levels of proinflammatory cytokines such as IL-6 and TNF-α [114]. Collectively these changes have been termed “inflamm-ageing”. In a murine adoptive transfer model, an altered chemokine milieu has been associated with impaired T helper cell trafficking to the spleen and a reduction in the ability of DC to prime naïve responses [153]. Clonal expansions of CD8+ T cells towards particular cytomegalovirus (CMV) epitopes have also been shown, meaning that the immune system is less able to mount adequate responses to new and diverse challenges [68]. Intracellular changes have been observed with altered responses of T cells to stimulus and alterations in the PI3 kinase signalling pathway across a number of cell types [9, 249]. In addition, an increase in myeloid derived suppressor cells (MDSCs) may be observed in both immunosenescence and the tumour microenvironment, contributing to a lack of T cell responsiveness in older age [107, 76]. All of these changes impact on the efficacy of cancer immunotherapies [107].
An immune risk profile (IRP) has been proposed to be a simple test to identify older people with a dysfunctional immune system in order to make therapeutic decisions [81, 34, 255]. The IRP includes low levels of B cells with increased levels of CD8+, CD28− and CD57+ cells, a poor T cell proliferative response to ConA, low CD4 frequency relative to high CD8 and CMV seropositivity [255]. It has been argued that TNF-α and IL-6 may also be important predictive markers to include as they are independent predictors of mortality in the elderly [34]. IL-6 also crucially regulates the balance between Th17 cells and Treg [139]. If an individual has an IRP, it remains to be established whether an immunocompetent phenotype can be recovered therapeutically.

1.5.2 Subject selection in ageing research

It is difficult to characterise associations between immune cell number and function with age due to individual variation between subjects, the complexity of the immune system and interrelationships between immune cells [198, 144]. Differing conclusions between ageing studies may result from discrepancies in subject selection criteria [150]. An example benchmark is the SENIEUR protocol, a set of stringent guidelines for subject selection originally proposed in 1984 [158] and revised recently [55]. The protocol includes reference ranges for serum, haematology and urinalysis parameters as well as clinical and pharmacological guidelines. It has shaped the course of ageing research in the subsequent years by minimising as many variables as possible to allow for robust experimental investigation [49]. However, most of our aged population is excluded on the basis of these criteria [49], and thus the final selectable population is not representative of normal. For example, it has been noted that centenarians are more resilient to cancer [45]. Therefore, arguably, excluding subjects with comorbidities associated with age is missing the point as differences in the immune systems of SENIEUR donors with age do not directly relate to clinically relevant age-related pathologies [49].

Conversely, many studies do not express any inclusion criteria other than “healthy”. This is not enough, and preferred strategies document their inclusion/exclusion criteria so that the reader can decide if they are robust [150, 78]. For example, one strategy documented the exclusion of subjects with relevant acute and chronic inflammatory conditions or taking certain medications, and then used the SENIEUR criteria to distinguish a “healthy” group from an “unhealthy” group [270]. Indeed, it
has been postulated that the SENIEUR protocol was only ever intended to be for definition of a reference population [205]. This study opted for an inclusive approach, only excluding volunteers suffering with acute inflammatory pathologies, and documenting relevant demographic and medical information using a self-reported questionnaire. This will be explained in detail in Chapter 2. Adequate sample sizes are necessary for testing hypotheses [77], and in Chapter 2 a power calculation is shown. Also, it is best to investigate the immune system in a population with a continuous spectrum of ages to obtain richer data on the dynamics of any changes throughout life, rather than just comparing a young and old group. The study by Faria and colleagues discussed later in this chapter is a case in point [78].

1.5.3 The importance of context

Immunosenescence factors should not be considered independently as it is their combination that contributes to a general decline in immune function with particular clinical manifestations as well as a higher mortality rate [15]. Factors secondary to direct immune function are also relevant, for example thymic involution is influenced by endocrine hormones including adrenal, growth and sex hormones [34, 15]. Sex steroids are thought to contribute by acting on androgen receptors in the stroma [15]. IL-7 is also thought to instigate thymopoiesis and has been shown to decline with age [15]. In older patients who experience traumatic injury, cortisol is less effectively opposed by dehydroepiandrosterone sulphate, which enhances immune function and counterbalances the inhibitory effects of cortisol, and therefore immunosuppression is promoted [15]. A recent study also investigated how alterations in circadian rhythm were associated with inflammation in the context of shift work [47].

Some additional molecular mechanisms have been associated with ageing, such as increased DNA damage and increased oxidative stress [123, 112]. Telomere shortening is also implicated, but not causally linked to clinical disease outcomes [245]. There are also several examples of functional genetic polymorphisms associated with ageing, such as the APOE2 allele (apolipoprotein E) linked to cardiovascular and Alzheimer’s disease [245]. Therefore it is crucial to understand individual changes to the immune system in a wider context.

An ideal study of human ageing would be a longitudinal investigation of a large population of individuals, taking regular measurements throughout
their lives, enabling particular events that promote an immunocompetent phenotype to be elucidated. Such studies are prohibitively expensive, difficult to manage and it would be impossible to maintain consistency in experimental technique over an entire human lifetime. Even the renowned Swedish studies such as OCTA, NONA and HEXA, which are longitudinal by design, can only follow individuals for a few years, and resort to comparing old and young groups [81, 255, 256]. In Strindhall’s 2007 paper for example, subjects were followed for 6 years, with 4 measurements taken every 2 years [255]. Having initially recruited 138 subjects without exclusion criteria, all four measurements were available for only 31 [255]. Measurements were taken to characterise the IRP, and analysis of the longitudinal data revealed no changes in T cell subsets, with donors exceeding 100 years of age still not displaying an IRP, demonstrating “successful ageing”, rather than a “normal” pattern of immunosenescence [255].

Arguably, owing to the change in lifestyle with each passing generation, it is more meaningful to characterise the immune balance in today’s younger and older individuals in order to provide the best health and social care, and to repeat similar cross-sectional studies for each generation.

The following subsections describe some of the literature concerning the balance between pro- and anti-inflammatory T cell subsets, addressing the hypothesis that their balance changes with advancing age. The majority of the literature reviewed was examined following an initial MEDLINE search and updates have been added throughout the study period.

1.5.4 nTreg

Mechanisms have been proposed in support of the hypothesis that number and function of nTreg changes with age [66]. Firstly, thymic involution from childhood into adulthood produces lower levels of naïve CD4+ and CD8+ T cells [66], and the production of new nTreg may also decline as a result. Therefore, like the conventional T cell pool, the peripheral nTreg pool must maintain homeostasis by inducing peripheral nTreg proliferation [191, 20]. In addition, the peripheral nTreg pool is unlikely to be able to divide and proliferate indefinitely, as nTreg possess shorter telomeres than their non-regulatory counterparts [66]. If the nTreg pool loses functionality faster than the conventional T cell pool, bodily tissues will have less protection from autoreactive T cells and thus increased levels of autoimmune conditions and inflammatory states may arise [66].
correlates with the increased prevalence of autoimmunity and chronic inflammatory disease seen clinically in older age [34].

The lack of nT reg output from the thymus would correspond to a growing proportion of nT reg with shorter telomeres or a “memory” phenotype. This could be ascertained by looking at CD45RA and CD45RO expression. Interestingly, it has been noted that CD45RO expression correlates with upregulation of CD25, so expansion of the CD4+CD25high population may be in fact expansion of the memory population relative to naïve, and not an increase in nT reg numbers at all [66]. It may reflect a population of autoreactive anergic T cells which become exhausted after chronic antigen exposure and take on a weakly suppressive role [66].

Another consequence of chronic antigen exposure might be restriction of the nT reg repertoire [66]. nT reg are thought to have a repertoire as diverse as non-regulatory T cells in the early decades of life. Potentially, nT reg might be “obsessed” with a particular antigen much like the conventional T cell pool’s apparent obsession with particular CMV antigens with advancing age [202]. Ultimately, this may also contribute to an inability of nT reg to adequately protect against autoimmunity [66]. It is not clearly understood how and to what extent these mechanisms might contribute to immune function in advancing age.

As a substantial body of information exists addressing how nT reg change with advancing age, this section presents a structured review of the major primary studies. Similar reviews of other immune cell subsets excluded subjects not using the SENIEUR protocol for selection of experimental subjects [88]. However, as discussed previously, the SENIEUR protocol selects for an unnaturally healthy elderly population. Thus, the criteria have been listed for all of the studies that have been reviewed to allow for comparison and studies were not rejected on this basis. Table 1 outlines the subjects investigated in these studies so their approaches might be compared and contrasted. Tables 2, 3 and 4 then sequentially address how the number, phenotype and function of nT reg was found to change with age according to each study.

Table 1.1 demonstrates the variation in how studies of nT reg throughout life are conducted, making it difficult to draw general conclusions. Of 11 studies, two included donors satisfying the SENIEUR protocol. 4 of the studies gave no specific inclusion and exclusion criteria. The remaining 5 studies documented a variety of parameters that were considered in subject selection. The studies by Faria in 2008 [78], Baltcheva in 2010 [20]
and Yan in 2010 [288] include the greatest spread of ages and also consider a continuous spectrum of ages, making for interesting studies of the lifetime dynamics of nTreg. The most common phenotypic definition of nTreg used in these studies is CD4⁺CD25^{high}, although it also considered Foxp3 as a marker. It could be argued that these 3 molecules should be the minimum requirements for enumerating nTreg. Interestingly, Miyara and colleagues in 2009 also included CD45RA in their staining panel, as they argue it provides distinction between resting and activated nTreg and prevents the inclusion of Foxp3⁺ non-nTreg cells in the enumeration.
<table>
<thead>
<tr>
<th>Study</th>
<th>Subjects</th>
<th>Inclusion and exclusion criteria</th>
<th>nTreg markers for enumeration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gregg 2005 [109]</td>
<td>44 (21-93); 20-30 n=14, 30-40 n=5, 40-50 n=5, 50-60 n=5, 60+ n=15</td>
<td>Healthy volunteers; no specific criteria</td>
<td>CD4⁺, CD25⁺⁺⁺</td>
</tr>
<tr>
<td>Hwang 2009 [127]</td>
<td>29 (&lt;40), 32 (&gt;65)</td>
<td>Volunteers excluded if taking immunosuppressive drugs, had an autoimmune or infectious disease, malignancy, diabetes or asthma</td>
<td>CD4⁺, CD25⁺⁺⁺, Foxp3⁺</td>
</tr>
<tr>
<td>Tsaknaridis 2003 [271]</td>
<td>27 (22-60)</td>
<td>Healthy volunteers; no specific criteria</td>
<td>CD4⁺, CD25⁺⁺⁺</td>
</tr>
<tr>
<td>Valmori 2005 [272]</td>
<td>7 cord, 47 adult (ages not given, 20-70 years)</td>
<td>Healthy volunteers; no specific criteria</td>
<td>CD4⁺, CD25⁺⁺⁺</td>
</tr>
<tr>
<td>Trzonkowski 2006 [270]</td>
<td>31 healthy and 36 non-healthy (19-40), 30 healthy and 49 non-healthy (65-99)</td>
<td>No current infection, no acute autoimmunity, 5 years after cancer remission, no vaccination in 6 months, no ischaemic/atherosclerotic event in 6 months, SENIEUR protocol to differentiate healthy from non-healthy in young and old, medical histories given for non-SENIEUR group</td>
<td>CD4⁺, CD25⁺⁺⁺</td>
</tr>
<tr>
<td>Rosenkranz 2007 [228]</td>
<td>Considered healthy cohorts: 38 (23-40), 33 (51-87)</td>
<td>No autoimmunity, no current respiratory infection, not on immunomodulatory drugs/corticosteroids</td>
<td>CD4⁺, Foxp3⁺⁺⁺, CD25⁺⁺⁺ for function</td>
</tr>
</tbody>
</table>

Continued on next page.
<table>
<thead>
<tr>
<th>Study</th>
<th>Subjects</th>
<th>Inclusion and exclusion criteria</th>
<th>nTreg markers for enumeration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lages 2008 [146]</td>
<td>16 (&lt;30), 16 (&gt;70)</td>
<td>No more than 2 morbidities, no chronic infection nor immunosuppressive medication</td>
<td>CD4⁺, CD25⁺, Foxp3⁺, CD127⁻</td>
</tr>
<tr>
<td>Faria 2008 [78]</td>
<td>232 (0-86)</td>
<td>Healthy volunteers; no infection, no acute or chronic inflammation, no autoimmune disease, no heart disease, no anaemia, no undernourishment, no leucopenia, no mood disorders, no neurodegenerative disease, no neoplasias nor use of hormones and drugs. Included a SENIEUR group</td>
<td>CD4⁺, CD25&lt;sup&gt;high&lt;/sup&gt;</td>
</tr>
<tr>
<td>Baltcheva 2010 [20]</td>
<td>60 males (20-81), 60 females (19-78), 7 cord blood samples</td>
<td>Healthy volunteers; no specific criteria</td>
<td>CD4⁺, CD25&lt;sup&gt;high&lt;/sup&gt;</td>
</tr>
<tr>
<td>Miyara 2009 [184]</td>
<td>29 (18-40), 12 (79-90)</td>
<td>Volunteers had no acute or chronic infectious disease, thrombosis or cancer</td>
<td>CD4⁺, CD25&lt;sup&gt;high&lt;/sup&gt;, Foxp3&lt;sup&gt;high/low&lt;/sup&gt;, CD45RA⁺⁻</td>
</tr>
<tr>
<td>Yan 2010 [288]</td>
<td>37 males, 43 females (20&gt;80)</td>
<td>Volunteers had no acute illness, no medication other than for hypertension and no serious prior illness</td>
<td>CD4⁺, CD25&lt;sup&gt;high&lt;/sup&gt;, Foxp3⁺</td>
</tr>
</tbody>
</table>

Table 1.1: Comparison of nTreg ageing studies for subject selection and nTreg markers.
<table>
<thead>
<tr>
<th>Study</th>
<th>Number of nTreg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gregg 2005 [109]</td>
<td>A 2.4 fold increase in the number of CD4&lt;sup&gt;+&lt;/sup&gt;CD25&lt;sup&gt;high&lt;/sup&gt; nTreg between the youngest cohort and the oldest</td>
</tr>
<tr>
<td>Hwang 2009 [127]</td>
<td>Comparable in young and older donors</td>
</tr>
<tr>
<td>Tsaknaridis 2003 [271]</td>
<td>Not investigated; unclear if normal function is retained by a compensatory increase in nTreg</td>
</tr>
<tr>
<td>Valmori 2005 [272]</td>
<td>No significant correlation between nTreg and age</td>
</tr>
<tr>
<td>Trzonkowski 2006 [270]</td>
<td>Increased significantly with age and in the non-SENIEUR older group</td>
</tr>
<tr>
<td>Rosenkranz 2007 [228]</td>
<td>Significant increase by 1.5% (p&lt;0.05) in CD4&lt;sup&gt;+&lt;/sup&gt;Foxp3&lt;sup&gt;+&lt;/sup&gt; cells and CD4&lt;sup&gt;+&lt;/sup&gt;CD25&lt;sup&gt;high&lt;/sup&gt; in older age, with a significant increase in absolute number of CD4&lt;sup&gt;+&lt;/sup&gt;CD25&lt;sup&gt;high&lt;/sup&gt; in the older group</td>
</tr>
<tr>
<td>Lages 2008 [146]</td>
<td>Significant increase by 1.5% (p&lt;0.05) in CD4&lt;sup&gt;+&lt;/sup&gt;Foxp3&lt;sup&gt;+&lt;/sup&gt; cells in older age group relative to CD4&lt;sup&gt;+&lt;/sup&gt;Foxp3&lt;sup&gt;-&lt;/sup&gt;, significant increase in CD4&lt;sup&gt;+&lt;/sup&gt;CD25&lt;sup&gt;+&lt;/sup&gt;CD127&lt;sup&gt;-&lt;/sup&gt;, non-significant increase in CD4&lt;sup&gt;+&lt;/sup&gt;CD25&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>Faria 2008 [78]</td>
<td>Oscillating number of CD4&lt;sup&gt;+&lt;/sup&gt;CD25&lt;sup&gt;high&lt;/sup&gt; cells with age: highest in 11-18 year group and in 61-75</td>
</tr>
<tr>
<td>Baltcheva 2010 [20]</td>
<td>Comparable in donors throughout the age range</td>
</tr>
<tr>
<td>Miyara 2009 [184]</td>
<td>CD45RA&lt;sup&gt;+&lt;/sup&gt;Foxp3&lt;sup&gt;low&lt;/sup&gt; resting nTreg declined with age, CD45RA&lt;sup&gt;-&lt;/sup&gt;Foxp3&lt;sup&gt;high&lt;/sup&gt; activated nTreg increased</td>
</tr>
<tr>
<td>Yan 2010 [288]</td>
<td>Comparable in donors throughout the age range</td>
</tr>
</tbody>
</table>

Table 1.2: Comparison of conclusions for number of nTreg throughout life.
Table 1.2 shows that 5 of the studies reviewed concluded that the number of nTreg in peripheral blood increases with increasing age. Although none of the remaining studies demonstrated the opposite, 4 found the number of nTreg to be comparable in young and old subjects. One of the two studies investigating a continuous spectrum of ages found an oscillating number of nTreg throughout the lifetime, emphasising the importance of not simply comparing a single young to a single old group. Interestingly, one study found the number of nTreg to be higher both in the older group and in the non-SENIEUR older group, implying that disruption to the homeostasis of nTreg throughout life may be associated with disease in the elderly. Therefore, it remains to be established whether or not the nTreg population is constant throughout life.

As shown in Table 1.3, the studies reviewed make similar conclusions regarding the phenotype of nTreg throughout life. All studies examining the dynamics of naïve and memory nTreg by CD45RO or CD45RA expression found a decrease in naïve cells (CD45RO− or CD45RA+) and an increase in antigen experienced cells (CD45RO+ or CD45RA−) with age. One study examined CD45RB and found more CD45RB+nTreg in the frailer groups and more CD45RB−nTreg with age. In addition, one study found the number of nTreg expressing CXCR4 to be decreased in the older donors relative to the young. This is a chemokine receptor involved in transport of T cells to secondary lymphoid tissue [146]. Downregulation of this molecule with age could amount to a decreased ability of nTreg to enter such tissue and thus prevent normal function.

Very few studies investigated the function of nTreg (Table 1.4). Some of the studies found comparable ability of nTreg from young and older donors to suppress conventional T cell proliferation and cytokine secretion. Some studies found that nTreg from older donors were less able to suppress T cell proliferation, but these observations were non-significant at the 5% level. The number of donors for which the functional assays are performed is generally very low in comparison to the whole study population. There is need to conduct a more in depth study of nTreg function in order to draw meaningful conclusions.
<table>
<thead>
<tr>
<th>Study</th>
<th>Phenotype of nTreg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gregg 2005 [109]</td>
<td>Consistent CD4, CD25, CTLA-4 and activation marker expression between 11 young (21-35) and 11 elderly (63-78). Significantly more CD45RO⁺ cells in the older cohort (95.51%) compared to the younger cohort (83.3%), p&lt;0.05</td>
</tr>
<tr>
<td>Hwang 2009 [127]</td>
<td>Significantly less CD4⁺Foxp3⁺CD45RA⁺ cells in the elderly group (mean 8.13 ± standard deviation 0.987) compared to the young (3.49 ± 0.846), p = 0.002. Otherwise comparable phenotype (CTLA-4, CCR7, CCR4, CCR5, CXCR3, IL-15Ra, IL-2/15Rb, IL-7Ra)</td>
</tr>
<tr>
<td>Tsaknaridis 2003 [271]</td>
<td>Not investigated with respect to age</td>
</tr>
<tr>
<td>Valmori 2005 [272]</td>
<td>Significant inverse correlation between naïve (CD25⁺ CCR7⁺ CD62L⁺ CTLA-4⁺ Foxp3⁺ CD45RA⁺ CD45RO⁻) nTreg and age (p&lt;0.0001)</td>
</tr>
<tr>
<td>Trzonkowski 2006 [270]</td>
<td>More CD45RB⁻ nTreg in the frailer groups, more CD45RB⁻ nTreg with age</td>
</tr>
<tr>
<td>Rosenkranz 2007 [228]</td>
<td>Higher CD45RO expression in the older cohort (9% increase, p&lt;0.05)</td>
</tr>
<tr>
<td>Lages 2008 [146]</td>
<td>CD3, CD4, CD25, CD69, CCR5, CCR7, integrin-α4, integrin-β7, C27, integrin-β1, GITR, TGFβRII, CD127, CD45RA, CXCR4, CTLA-4, PD-1, granzyme-A and granzyme-B were investigated. The only statistically significant observation was the number of CXCR4⁺ nTreg, mean 28.2 ± standard deviation 3.7 in 9 young and 15.8 ± 2.0 in 11 old</td>
</tr>
<tr>
<td>Faria 2008 [78]</td>
<td>Not investigated</td>
</tr>
<tr>
<td>Baltcheva 2010 [20]</td>
<td>The ratio between CD45RO⁻ and CD45RO⁺ Treg was inverted in adulthood</td>
</tr>
<tr>
<td>Miyara 2009 [184]</td>
<td>CD45RA⁺ subsets decreased in frequency with age</td>
</tr>
<tr>
<td>Yan 2010 [288]</td>
<td>Not investigated</td>
</tr>
</tbody>
</table>

Table 1.3: Comparison of conclusions for phenotype of nTreg throughout life.
<table>
<thead>
<tr>
<th>Study</th>
<th>Function of nTreg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gregg 2005 [109]</td>
<td>In 3 young and 3 old, consistent suppression of T effector proliferation and IFNγ secretion</td>
</tr>
<tr>
<td>Hwang 2009 [127]</td>
<td>Comparable suppression of proliferation and cytokine secretion between 5 young and 8 older donors. IL-10 production by conventional T cells suppressed more in 9 older donors compared to 8 young</td>
</tr>
<tr>
<td>Tsaknaridis 2003 [271]</td>
<td>Suppression of conventional T cell proliferation declined with age but no indication of significance. Suppression was overcome by IL-2. Antibody blocking experiments demonstrated roles of key molecules in suppressive function: CTLA-4, GITR, IL-10, IL-17, CD103, TGFβ. Transwell assays determined need for cognate interaction</td>
</tr>
<tr>
<td>Valmori 2005 [272]</td>
<td>Not investigated with respect to age</td>
</tr>
<tr>
<td>Trzonkowski 2006 [270]</td>
<td>In 23 healthy and 16 non-healthy young versus 30 healthy and 54 non-healthy older, number of nTreg increased with age and presence of disease was correlated with poorer CD8+ and NK cell function, a decrease in serum IL-2 and in IL-2+ cell frequency</td>
</tr>
<tr>
<td>Rosenkranz 2007 [228]</td>
<td>Non-significant decrease in suppression of proliferation in the old relative to the young</td>
</tr>
<tr>
<td>Lages 2008 [146]</td>
<td>nTreg depletion increased proliferation of the remaining CD4+ lymphocytes in 4 of seven elderly donors, but failed in 3</td>
</tr>
<tr>
<td>Faria 2008 [78]</td>
<td>Not investigated</td>
</tr>
<tr>
<td>Baltcheva 2010 [20]</td>
<td>Not investigated with respect to age</td>
</tr>
<tr>
<td>Miyara 2009 [184]</td>
<td>Not investigated</td>
</tr>
<tr>
<td>Yan 2010 [288]</td>
<td>Not investigated</td>
</tr>
</tbody>
</table>

Table 1.4: *Comparison of conclusions for function of nTreg throughout life.*
1.5.5 iTreg, Th1 and Th17

For iTreg and Th1 cells, the literature is more heterogeneous. Although some high sensitivity assays can measure spontaneous cytokine expression in CD4$^+$ lymphocytes, to elicit a recall cytokine response a variety of stimulation conditions can be used. Variation in the conclusions drawn from these studies may be explained by different recruitment strategies and stimulation methods [98]. For example in studies using phorbol 12-myristate 13-acetate (PMA) and ionomycin stimulation, an increased proportion of CD4$^+$ T lymphocytes that were IFN$\gamma^+$ was observed in the older participants relative to the young [233, 231]. However, in a study using anti-CD3 stimulation of lymphocytes, a decline in IFN$\gamma^+$ production was observed [159]. In a study using both, a decline was observed [132]. When stimulating peripheral blood mononuclear cells (PBMCs) with Staphylococcus enterotoxin B, increased IL-10 production was shown in older donors relative to the young [50]. IL-10 production by PBMCs also increased in the older donors in response to phytohaemagglutinin (PHA) stimulation, with a corresponding decline in IL-12p40 [48]. However, both IL-10 and IFN$\gamma$ production from PBMCs stimulated with PHA and polyinosinic-polycytidylic acid (poly(I:C)) did not change with donor age [287]. Where plasma measurements were used to determine whether the immune system changes with age, such as a study by Forsyth and colleagues investigating 138 older donors recruited through the NONA study versus 18 younger healthy donors, IL-10 did not change [87]. In another exceptionally large study of 1,411 individuals, plasma IL-10 did not change with age [254].

In a study of 349 pairs of twins, a moderate correlation was observed in serum IL-10 measured using a high sensitivity bead-based assay, indicating a genetic component [238]. This degree of heritability was constant in twins of different ages, contrasting with other cytokine measurements where environmental factors played a larger role in older age [238]. IL-10 was also significantly associated with body mass index [238]. Interestingly, in a study of 249 donors with a mean age of 67.5 tested in the mid 1990s and once again 10 years later, a decline in serum IL-10 was observed with age [21]. This was accompanied by an increase in proinflammatory cytokines and did not relate to CMV seropositivity [21]. IL-10 has also been measured in a number of clinical studies. For example, age did not affect the serum IL-10 levels in patients presenting
with community-acquired pneumonia, however no comparison was made with healthy controls [134]. In a study of stress and ageing, no change in CD4+IL-2+IFNγ+ cell frequency was observed, but an increase in CD4+IL-10+ cells was observed in stressed individuals [104]. Furthermore, older individuals suffering with stress had a more pronounced increase than the younger individuals also suffering with stress [104]. However, in a study of IL-10 levels in frail and non-frail older people, IL-10 was not shown to be associated with frailty [154].

Of particular interest are studies investigating vaccine responses in young and older donors, with respect to IL-10 and IFNγ. PHA-stimulated whole blood IFNγ abundance before and after DT-Polio-Typhim vaccination showed that elderly donors had a marked decrease in Th1 [290]. Similarly, influenza vaccination also failed to elicit IFNγ production in elderly subjects, with no change in IL-10 [97].

In two previous studies of the balance between IL-10 and IFNγ, IL-10 secretion was shown to increase with age using whole blood stimulation assays, corresponding to decreased IFNγ secretion [40, 37].

Contrasting with the previous subsets described, there is little known about how Th17 cells behave throughout the lifetime, but it is likely that they are susceptible to the mechanisms of dysfunction and dysregulation with age as outlined previously for other T cell subsets. To the best of our knowledge, a causal mechanism linking aged Th17 cells to inflammatory pathologies has not yet been found. A single recent paper by Lee and colleagues (2011) has directly addressed the number and function of Th17 cells throughout the lifetime, comparing naïve and memory responses [152]. This work has concluded that there are age-related changes to IL-17-producing cells, particularly concerning maintenance of immune memory with a preservation of function of naïve cells recently differentiated to a Th17 phenotype [152].

38 donors aged ≥ 65 years and 41 donors aged ≤ 40 were studied [152]. The subjects were apparently healthy, taking no known immunosuppressive drugs or suffering from autoimmune, infectious, or malignant conditions, nor diabetes or asthma [152]. PBMCs were sorted by flow cytometry to CD4+ naïve cells as CD45RA+CCR7+ and memory as CD45RA−CCR7+/− [152]. The naïve cells were then stimulated with CD3 and CD28 over 7 days in the presence of IL-1β, IL-6, IL-23 and TGF-β [152]. The memory cells were given IL-1β and IL-23 [152]. In addition, the authors stimulated freshly separated PBMCs from their young and old donors with PMA and ionomycin for 4 hours to measure
the presence of IL-17-producing cells [152].

Their findings included a small decrease in the frequency of Th17 cells in the CD4\(^+\) memory population in elderly donors relative to young [152]. This contradicts a study of IFN\(\gamma\) and IL-17A production in response to pneumococcal antigens, which found no correlation between T cell immunity and age [242]. Lee and colleagues also reported an increased tendency to differentiation to a Th17 cell phenotype from the elderly CD4\(^+\) naïve T cells relative to young [152]. This had been demonstrated previously in mice [120, 126]. Elderly CD4\(^+\) memory cells produced less IL-17 than young, but there was increased IL-17 production by elderly CD4\(^+\) lymphocytes that were initially naïve relative to those from young donors [152]. The changes observed were subtle and the proportions of CD4\(^+\) naïve and memory cells that were of Th17 phenotype were minute [152]. Thus, it is necessary to establish the clinical relevance of these changes, and how these alterations combine with other age-related changes to immune system components and other tissues. It could be argued that, owing to the immune-amplifying and inflammatory effects of Th17 cells, apparently insignificant changes could amount to large-scale alterations in immune function downstream.

1.6 Rationale for studying multiple T cell subsets with age

This review has so far summarised key changes to the immune system with older age and established that the literature concerning T cell subsets is contradictory. Owing to the heterogeneity of the literature there is still a need to establish, by laboratory experimentation, whether certain age-related changes actually occur. Age-related diseases may involve disruption of the balance between T cell subsets and thus it is important to conduct research into the relationship between these cells in older people.

There is a growing body of evidence to suggest that ageing affects many aspects of the immune system and that this is associated with age-related clinical manifestations. However, studies to date have been largely limited to single leukocyte subsets or individual serum cytokines [109, 152, 217, 150, 27, 182], whereas it is likely that relative changes between leukocyte populations contribute to clinical manifestations of immune ageing. Therefore, we investigated changes in overall balance between cell populations as these ultimately translate into immune
function. The healthy immune system carefully controls the magnitude and duration of immune responses by balancing the activation of effective immunity with regulation [85]. We postulate that the important mediators of balance are Th17 and Th1 cells that amplify immune responses, versus nTreg and iTreg.

One study of circulating cytokines with age measured our three cytokines of interest, finding serum IL-10 to be positively correlated with age, IL-17 to be negatively correlated and IFN\(\gamma\) to remain unchanged [11]. However, this study only included 73 participants and did not couple their ELISA measurements to flow cytometry enumeration by proportion of CD4\(^+\) lymphocytes and absolute number [11]. In Chapter 2 our method for investigating these subsets in donors of different ages is presented in more detail and in Chapter 4 the corresponding findings are documented.

1.7 Computational methods

1.7.1 Introduction

Increasingly, there is a need to better integrate expertise from different disciplines to make progress with complex clinical and translational research [29]. In particular, the field of biomedical informatics can contribute a portfolio of computational techniques to help get the most out of “wet-laboratory” data in exploratory or hypothesis-driven contexts [29]. However, meaningful collaborations between mathematicians, computer scientists and immunologists are still rare.

This section introduces the concepts of simulation, defined as “methods and applications to mimic the behaviour of a real system” [135], and data mining, defined as “the extraction of hidden predictive information from large databases” [111]. We discuss how these techniques can and should be used more routinely by experimental immunologists to create and test research hypotheses relevant to both fields [73].

1.7.2 Simulation

A computer, by its most basic definition, takes a variety of inputs, processes them, and returns some sort of output. The immune system also takes a variety of stimuli, processes them via a network of cellular interactions and signalling pathways, then executes an appropriate course of action. It is the scale and complex nature of the immune system’s
overlapping layers of functionality and regulation that lends it to mathematical and computational methods [138]. Very few processes in the immune response are coordinated by one cell, or one discrete pathway of signalling or cytokine secretion, yet a reductionist approach to laboratory experimentation is the norm [138]. In vitro experimentation is useful for looking at individual interactions, but is far removed from the whole picture, and in vivo experimentation is useful to look at the whole picture, but is less able to answer specific questions [138]. Computational analysis can manage complexity and by feeding in what is known experimentally about a system, it is possible to derive theories as to the missing pieces when looking at a broad overview [138]. It can help researchers to ask the questions “is this a true representation of the biological system?” or “is this a likely conclusion given the physical restraints on the system?”. For example, recent investigatory modelling has questioned the role of the thymus in the developing T cell pool [19]. Models may also capture aspects of immunity that are unexpected, dynamic and non-linear [53].

So far, the majority of the work simulating the immune system has used ordinary differential equations, delay differential equations, partial differential equations, stochastic differential equations and agent-based modelling [138]. Simulations can be static or dynamic, continuous or discrete, deterministic or stochastic, according to the needs of the user [135].

For example, numerous systems of ordinary differential equations of varying complexity already exist which deterministically model cancer interacting with the immune system [73]. Populations are considered collectively as “number of tumour cells” or “number of immune cells”, and changes to these populations over time are determined by the initial size of population, multiplied by functions incorporating rates of proliferation and rates of death [73]. These models have been used to investigate and predict tumour aggression [57] and tumour growth rates with age [281].

Agent-based modelling, a stochastic method, is useful for modelling cellular interactions [263]. Agents representing cells are assigned size, shape, rules and rates of change dependent on their function, and their behaviours are simulated to predict how they will interact in an unknown setting in order to address a biological hypothesis [263]. Agent-based models can generate emergent phenomena, described as “the global behaviours or patterns that arise through ‘self-organisation’ and that could not otherwise have been characterised a priori” [53]. In the context of the immune system, this means you cannot learn about all functions of
cells of the immune system by looking at them individually, you have to look at them in the context of their interactions. Agent-based models have been used to shed light on the complex process of T cell recognition, for example [46].

The scope of the model and the nature of the interactions involved dictates the appropriate modelling approach [138]. For example, agent-based models are difficult to parameter fit and analyse statistically, and usually require a prohibitively large amount of processing power [138, 82]. However, they are considered best able to mimic the stochastic nature of the immune system and allow for individual entities to be represented, in contrast to system dynamics modelling, which only represents entities collectively [82].

Contrasting with the quantitative methods described, a qualitative model was recently built employing logical formalism to predict how mammalian CD4+ T cells differentiate according to the environmental stimuli they receive, in particular the cytokine milieu at the time of stimulation [193]. The signalling pathways and transcription factor interactions were simulated and steady states of Th1, Th2, Th17 and T regulatory cells were reached, along with transient intermediates. Such a model could be tested and refined by mimicking these experiments in vitro, and in considering alternative costimulatory molecules.

Considering more T cell models, in other previous work a system of ordinary differential equations was written based on the hypothesis that the immune system responds to uncontrolled tumour growth in two ways: some tumours expand Tregs whereas others do not [156, 155]. This model is sometimes referred to as the cross-regulation model as the dynamics of Treg depend on their interactions with antigen-presenting cells and other T cell subsets [43]. It was argued that fast growing, poorly immunogenic and T cell effector function-resistant tumours predispose to development of a Treg-independent response [155]. In contrast, tumours which are slow-growing, highly immunogenic and T cell effector function-sensitive were argued to predispose to development of a Treg-dependent response [155]. What was not known was how these developing tumours respond to different therapeutic strategies [155]. The authors applied simulated treatment options to their model of tumour growth to predict their efficacy. With knowledge of the Treg status of an individual it was possible to predict the most appropriate therapeutic option using this model [155]. By consideration of the timing of treatment application relative to initiation of tumour progression, appropriate dosages and use of
combination therapies, it was theoretically possible to induce tumour rejection and long lasting immunity [155]. Importantly, it was also theoretically possible to accelerate tumour progression by intervening at the wrong time with the wrong treatment [155].

A recent paper elaborated on the scope of the cross-regulation model using differential and delay-differential equations and more explicitly modelled the transit of immune cells between lymph, blood, and tumour compartments [67]. Using melanoma immunisation as a treatment, a detailed analysis of hypothetical dosing and site of injection regimens was conducted, as well as a parameter sensitivity analysis to find out which patient characteristics could reduce efficacy [67].

Of particular interest to our study of ageing and T cell subsets was Baltcheva’s model of Treg dynamics throughout the lifetime [20]. It comprehensively incorporates the functional dynamics of Treg in homeostasis and during an acute immune response, and characterises the changing precursor and mature Treg populations throughout the human lifetime. The original model was based on numbers of CD4+CD25+CD45RO− (precursor) and CD4+CD25+CD45RO+ (mature) populations in 119 peripheral blood samples of donors aged 19 to 81 [20]. Ordinary differential equations describe the dynamics of the abovementioned cells, and stochastic processes control the frequency, duration and antigen-specific nature of primary and secondary immune responses on the different cell compartments [20]. We have also modelled one of Baltcheva’s scenarios in order to test the hypothesis that the original equations could be implemented in AnyLogic [85]. It was shown that the simulation mimics the key feature of inversion of precursor and memory cells in early adulthood [85].

Even though some types of simulation method boast an ability to be stochastic and adequately deal with complexity, simulation is always limited by the quality of the input, and the assumptions that have to be made owing to experimental noise in the data on which it is based and lack of conceptual knowledge [135]. If something is difficult for the programmer to understand, it is difficult to simulate. Also, when models scale up in complexity, they become more difficult to debug, harder to analyse, and it is more difficult to approximate values of parameters [73]. Parameters are often estimated because of the unreliability of the published evidence available [73]. Keeping in mind these caveats, simulation remains useful as it is easy, cheap, fast and mistakes do not matter as much as in a wet laboratory context [135]. Therefore,
exploratory simulation work may be valuable in generating the next hypotheses to test experimentally. In Chapter 6 we discuss how the ageing dataset presented in this thesis may be used as a basis for a deterministic model of T cell immunity throughout the lifetime.

1.7.3 Data mining

Data mining is essential for finding patterns within a dataset that are not immediately apparent to the observer, either because of the contribution of multiple attributes and complex relationships or because of the sheer volume of information. Data mining techniques can be categorised into supervised and unsupervised approaches.

Supervised learning

Supervised learning is performed when each data point is known to belong to a particular class [128]. The aim is to learn the important characteristics, from a training set of attributes measured, that best model which class each object belongs to, so that class labels can be predicted for unlabelled objects [142]. Also, much can be learned by looking at the models built about the important interactions within the dataset.

Previously, three supervised methods (a C4.5 tree classifier, a multilayer perceptron and a naive Bayes classifier) were compared and used to validate a categorisation of breast cancer pathologies to enhance clinical decision making [250].

Various different categories of supervised learning algorithms exist. Logic-based methods include decision tree algorithms and rule-based classifiers [142]. Decision tree algorithms are particularly popular because of the simplicity of their interpretation [143]. By recursive partitioning, they split the source dataset into subsets based on an attribute value, which stops when the subset at a node has the same value as the target variable, or when further splitting no longer improves the prediction [143, 167]. Then, the next attribute is considered, building a tree structure, until all of the training data is classified with a minimum of classification error [267]. Classification trees are appropriate for ordinal data, whereas regression trees are appropriate for continuous data [167].

A particular decision tree building algorithm, CART (Classification and Regression Trees), commonly uses the Gini index as a splitting rule [167, 267]. In Chapter 4, an implementation of CART was used to explore
the relationship between T helper cell measurements and age, and to identify potential confounding factors. Important reasons for selecting CART are that it is a fast, non-parametric algorithm, which does not require for variables to be selected in advance by the user and handles noisy data well [267]. However, CART does not exploit the multidimensionality of the data to its full potential as it divides by only one attribute at a time [267]. Also, the decision trees may not be very stable when small changes are made to the input data [267].

Another category of supervised learning algorithms includes single or multilayered perceptron-based techniques, also known as artificial neural networks (ANN) [142]. ANN encode the principles of a neuron in the computation at a network’s nodes [100]. Nodes receive inputs, become “activated” when the signals are strong enough, and in turn produce outputs, which might activate further nodes [100]. Inputs are the dataset of interest and an initial weighting that is optimised until a pre-determined set of outputs is achieved, i.e. a classification [100]. Backpropagation is an example algorithm that is often used, where signals are sent forwards and errors are sent backwards until the network learns the rules and can no longer improve the error [100]. Such an approach was recently used to identify cases of prostate cancer from pre-clustered flow cytometry datasets [268].

Other supervised methods include statistical learning algorithms, which differ as they consider each object as having a probability of belonging to each class, rather than a definitive class label [142]. Also, support vector machines (SVMs) map the dataset into a higher dimensional space such that a hyperplane can be fitted, separating the data into classes [38].

Selecting an algorithm for a given problem is a challenge as no single approach is universally most appropriate [44]. However, it could be argued that one should start with the simplest algorithm for purpose and increase the complexity of the algorithm if necessary, hence the selection of CART for the analysis in Chapter 4.

Unsupervised learning

Often in raw biological datasets, such as those generated by flow cytometry, class labels are unavailable. Unsupervised learning, in contrast to supervised, is performed without prior knowledge of any class label assigned to each data point and is therefore a more challenging task [128]. Clustering, an unsupervised method, is entirely data driven. The emphasis
is to discover a natural structure in the data and consequently separate out groups within the data [128]. It is classically best suited to data exploration as opposed to specific hypothesis testing. Clustering has become the main computational technique exploited in this work and later in this chapter we will discuss several examples of clustering algorithms used to automate flow cytometry data analysis [130, 220, 295, 212, 84, 83, 218]. However their use in biomedical science is not limited to raw flow cytometry data; a combination of several clustering algorithms was used to ascertain if breast cancers could be divided into clinical phenotypes in order to manage patient treatment effectively [108]. Clustering algorithms and their use in biomedical research, from genomic sequence and image analyses to the mining of medical texts, have been extensively reviewed [286].

Another form of unsupervised learning is blind signal separation, where individual signals need to be separated from a mixture and only the mixture is available [42]. An example of this might be extracting individual voices from a recording of a whole room full of people [26]. Principal component analysis (PCA) and independent component analysis are example techniques for feature extraction in this context [42].

Clustering is easiest when natural groups are tightly packed together and separated from one another [128]. Where the natural structure of the data takes this form, a plethora of different clustering algorithms will perform well. However, when natural groups are diffuse, non-spherical and overlapping, the task is more difficult [128], and more attention needs to be paid to the choice of algorithm. The aim of the clustering task also impacts on the choice of algorithm; different clustering algorithms will allow for the data structure to be understood in different ways.

Clustering algorithms can be hierarchical and partitional. K-means is the most widely used partitional clustering algorithm due to its simplicity and ease of interpretation [173]. It seeks to minimise the Euclidean distance between the datapoints and the nearest cluster centroid by iteratively moving the centroid locations and measuring the sum of the squared distances until the minimum is found [173]. This results in “hard assignment” of data points into one of the clusters [128]. Fuzzy c-means extends k-means by “soft assignment” of a membership for each cluster to each datapoint, depending on the distance from each cluster centroid [128]. However, both these approaches are sensitive to random number seed, and seek to find spheroidal and well-separated clusters. In Chapter 5 we show preliminary data using k-means, and compare and contrast its
performance with SamSPECTRAL, an algorithm optimised for flow cytometry data clustering problems.

The main criticism of clustering is that a clustering algorithm will generate some kind of solution whether meaningful or random [248]. Without prior knowledge of the class label it is a challenge to argue whether clustering has worked for a particular problem or not. Also, many clustering algorithms, such as $k$-means, are sensitive to initial conditions. It is typical to conduct multiple runs with different initial conditions and to select the best output [129], or summarise all of the runs. Often, in the field of data mining, the resultant classification is validated at least in part by the domain expert [3, 54]. However, there exist statistical measures of cluster validity that can be computed which indicate how representative a partition is, given the underlying data structure, such as the average silhouette width, entropy and Pearson’s $\Gamma$ statistic [113]. These will be evaluated in Chapter 5.

For precise enumeration it is not clear which algorithms are best for which problems. It is also prudent to note that, as with supervised approaches, there is no perfect clustering algorithm [129]; it is impossible to satisfy all conflicting requirements, and therefore alternatives to what is commercially available may need to be found for certain problems. Later in this chapter this review is extended to demonstrate a need for clustering algorithms to automate flow cytometry analysis for objective enumeration of nTreg.

1.8 Flow cytometry data analysis

Flow cytometry is a powerful tool for qualitatively and quantitatively determining the characteristics of individual cells within a population of interest, thus essential to enumerate T cell subsets and to calculate ratios between them. It involves a sample being processed to a single cell suspension and passed through a laser beam. If, prior to running the sample, fluorescent dye-conjugated antibodies specific for molecules of interest are added, the molecules present both on the surface and within these cells can be detected. The higher the fluorescence intensity, the more abundant the molecule of interest is on a particular cell. The current standard of data analysis in multicolour flow cytometry is to manually define a hierarchy of 2-dimensional gates in order to enumerate or measure the fluorescence intensity of a given subpopulation [173, 28, 6]. This is
time consuming, error-prone, subjective and poorly reproducible [257, 23, 214]. The multidimensional nature of the data is also ignored with a 2-dimensional approach [23], and it may be the case that the most interesting and accurate data only emerges in more than two dimensions. There has been recent interest in developing automated methods that are more objective, more accurate and overcome some of the initial teething problems in applying generic algorithms to flow cytometry datasets, although analysis tools have not advanced as quickly as the hardware and consumables otherwise required for flow cytometry assays [130].

Put simply, flow cytometry data is a collection of points in multidimensional space where the fluorescence intensity for each channel is defined by a set of coordinates, therefore, it lends itself to analysis by clustering techniques [257]. However, the main challenges are in clustering rare, overlapping and adjacent populations, and when two clusters of very different densities lie next to one another [295].

In order for laboratory scientists to use automated techniques, other attributes of clustering processes should also be considered, such as transparency; the clustering concept and its software implementation must be intuitive to a biology-biased audience. Obviously, the clustering algorithm must result in a division of the data into biologically relevant clusters in order for the process to be meaningful, in the case of our nTreg data minimal contamination of T helper cells with an “effector” phenotype in the “nTreg” cluster would be optimal, without spurious splitting or aggregation of the clusters of interest. Ideally, there would be minimal data preprocessing requirements. Perhaps most importantly, the ability for multiple samples to be compared is essential for any hypothesis testing. Hence, an automated elliptical gating method was developed to follow the clustering analysis in this work.

In the rest of this section we review the current state-of-the-art in automated flow cytometry analysis and discuss which approaches may be most appropriate for nTreg data. We highlight three studies which have already looked at nTreg-specific markers, and in Chapter 5 we build on these previous investigations by applying an appropriate algorithm to datasets stained according to a more recently defined and specific phenotypic definition of nTreg.
1.9 Specialised flow cytometry data clustering algorithms

1.9.1 Introduction

Whilst technological advancement in flow cytometry has improved its speed, efficiency, precision and the number of parameters one can measure at at time, the ability to analyse the data produced has not kept pace [5, 166, 23]. In recent years, several novel platforms have been developed to augment and improve flow cytometry data analysis. Four main groups can be described; tools that enable automated data exploration and population detection such as clustering algorithms, tools that extend this to enable automated gating, tools that arrange multidimensional data visually in novel ways to permit better manual gating and tools that take manually gated populations to automate multisample comparison and facilitate deeper interpretation. Some algorithms can bridge these groups, automating a larger portion of the analysis pipeline connecting raw data to publication-ready figures.

A comprehensive review has been presented previously by Bashashati and Brinkman [23]. In the next sections, the particularly interesting features of key algorithms from each major group are discussed and their major strengths and weaknesses are highlighted. Where an algorithm has previously been used to analyse nTreg data is clearly indicated. The selection of SamSPECTRAL for further investigation is then justified and a detailed description is given.

1.9.2 Algorithms for data exploration and population detection

In general, the current best methods for automated analysis of multiparameter flow cytometry data generally fall into three camps; those which fit complex models to the data, those which transform and refine the data such that complexity of the clustering problem is reduced and a simple algorithm such as $k$-means can be used, and hierarchical clustering methods which consider pairwise relationships between every data point up to population level. Those which fit models can be used to infer various properties about the dataset, detect deviations from a particular model and estimate how each population contributes to overlapping sections. Those which transform or refine the data may lead to reduction of raw data integrity, but are less computationally expensive [212]. Hierarchical
clustering algorithms provide a rich and deep interpretation of the data, but are less useful for flow cytometry data analysis unless coupled to a sampling algorithm to reduce data resolution. In this section we discuss some examples of these algorithms and the value of preserving the human perspective.

FLAME (FLow analysis with Automated Multivariate Estimation) is a model-based approach which fits multivariate skew t-distributions using Expectation Maximisation to flow cytometry data in order to detect populations of interest [212]. By minimising a weighted ratio of average intracluster to intercluster Mahalanobis distances, the number of clusters present within the data is determined, and post-clustering processing allows for inappropriately split clusters to be recombined [212]. Although Gaussian model fitting has also been applied to flow cytometry data [23], arguably flow cytometry data tends not to be symmetrical and therefore these approaches can generate arbitrary partitions as a result [212, 216]. For example, FLAME modelled a lymphoblastic cell line stained for HLA-DQ and CD95 more accurately than a Gaussian mixture modelling approach using MCLUST [212]. FLAME is a genuine multivariate approach, therefore superior to other model fitting methods which consider only one parameter at a time [212]. FLAME’s most valuable characteristic is in preserving the integrity of the data [212], however it is inappropriate for large datasets and those with more than 5 attributes due to its computational inefficiency [295]. One example of CD4$^+$, CD25$^\text{high}$, Foxp3$^+$ nTreg was successfully clustered from peripheral blood mononuclear cells by Pyne and colleagues using FLAME [212].

A related algorithm is flowClust, a mixture modelling algorithm that allows for multivariate t-distributions to be fitted to the data, and normalisation of the data using a Box-Cox transform [166, 83]. flowMerge, as in the merging step of FLAME, is an extension of flowClust, allowing for clusters drawn from the same population to be merged and allowing non-spherical populations to be clustered [8, 83].

SamSPECTRAL is an example of a data reduction approach based on spectral clustering [295]. This is usually also computationally expensive, so SamSPECTRAL was designed to firstly reduce the dimensionality of the dataset by a novel sampling method [295]. Approaches to sampling large datasets include Nyström’s method, where a uniform sample is extracted [295]. However, this approach may fail to represent low density clusters, and therefore Zare and colleagues opted for a “faithful sampling” approach, selecting 1500-3000 “community representative” data points.
uniformly over the solution space containing raw data and weighting them according to the density of events, thereby balancing computational feasibility and data resolution [295].

As a spectral clustering algorithm, SamSPECTRAL avoids the assumptions parametric approaches make that the data can be approximated to a particular distribution and therefore the controversy about selecting the appropriate distribution [295]. Being a non-parametric approach it is less sensitive to outliers and noise [295]. As with FLAME and flowMerge, clusters which appear to belong to the same population can be detected and merged automatically if the ratio of inter-cluster similarity to intra-cluster similarity is sufficiently large [295]. Such a merging step is essential for the detection of irregularly shaped populations. In another example, flowMeans is an adapted $k$-means algorithm for the analysis of flow cytometry datasets [8]. It improves on the traditional $k$-means algorithm by using change point detection to identify single populations, modelling them with multiple clusters [8].

Hierarchical clustering is a bottom-up technique, clustering data from single events to find the hierarchy of subpopulations present within a dataset without user specification of expected cluster number [84]. As it considers event-level relationships, hierarchical clustering allows for more sensitive detection of very small populations and discrimination between populations [84]. Overlapping populations should be easily identifiable and it is particularly suited to data exploration, and the user has only to define the level of the dendrogram that is of most interest for cluster enumeration, or this can be automatically computed [84]. In the paper by Fišer and colleagues, a novel technique was used to compute intercluster distance, smoothly transiting from Euclidean to Mahalanobis distance measurement depending on the number of events within the cluster, which is better suited for non-spherical data [84]. Hierarchical clustering using Mahalanobis distance successfully detected the main cell lineages within whole blood from scatter data, and also identified some meaningful subsets of B cells [84]. The current limitation of the algorithm in [84] is its inability to be used on $10^5$-$10^7$ datapoints, necessary for our raw nTreg datasets. However, once processing capabilities improve, this would be a very promising approach. Fišer and colleagues suggest using a $10^4$-sized dataset as a training set, then a support vector machine to classify the full dataset [84]. However, it could be argued this directly contradicts the purpose of a bottom-up approach priorising each data point, if the first requirement is to reduce the dataset to 1-10% of its original size.
Another algorithm based on hierarchical clustering is SPADE (Spanning-tree Progression Analysis of Density-normalized Events) [214]. The paper by Qui and colleagues introducing SPADE provides a novel change of emphasis to consider sample heterogeneity and relatedness, as opposed to the classical clustering paradigm of maximising separation. To quote: “Because these algorithms strive to define maximally different clusters, they often miss the underlying continuity of phenotypes (progression) that is inherent in cellular differentiation” [214]. This may be particularly relevant for nTreg flow cytometry data which is characterised by molecules expressed as a gradation. SPADE has been used to generate the well characterised hierarchy of phenotypes of mouse bone marrow haematopoetic cells from flow cytometry data, then used to explore 31-parameter mass cytometry data characterising human bone marrow haematopoetic cells [214]. By observing that the shape of the trees generated only changed when parameters that were significantly different to the other markers within the experiment were added or subtracted, redundant parameters were identified [214].

As hierarchical clustering methods are classically limited because of their high computational cost, SPADE, like SamSPECTRAL, includes an elegant, density-dependent sampling step which ensures rare populations are represented [214]. A sample of 30,000 events is taken [214], an order of magnitude larger than the sample size SamSPECTRAL selects. Agglomerative, or hierarchical clustering is then conducted, resulting in the generation of a minimum spanning tree representing multidimensional data as a 2-dimensional plot, using the Fruchterman-Reingold algorithm [214]. SPADE does not require prior knowledge of the cellular hierarchy and can therefore be used as a data exploration approach [214]. The final stage is up-sampling to cluster each individual event [214]. SPADE, however, requires 3 parameters to be defined by the user; the outlier density, the target density and the desired number of clusters [214]. Therefore, prior knowledge of the biology and its implications on how the clustering algorithm will function is necessary.

1.9.3 Algorithms for automated gating

A simple elliptical gate can be fitted to 2-dimensional flow cytometry data using the flowCore package if the data has not been transformed [75]. However, such a gating method alone is limited in defining populations of interest. Usually, basic gating procedures are coupled to a clustering
algorithm to detect populations within a dataset. For example, flowType [5, 4] extends flowMeans clustering, using it to generate a partition of events into positive and negative, then translating single-dimensional partitions into multiple dimensions. Therefore, more populations than the user originally intended can be analysed and the scope for finding novel populations that vary from sample to sample is expanded [5].

To the best of our knowledge, only one automated gating algorithm has been used to enumerate nTreg. Jeffries and colleagues used breakpoint regression and a fuzzy clustering Gath-Geva algorithm to enumerate CD4\(^+\)Foxp3\(^+\) cells in the peripheral blood of 22 HIV-infected donors [130]. Heavy image preprocessing was performed to remove isolated events and events from object boundaries and to improve curve smoothness [130]. The CD4 population was successfully divided into Foxp3\(^+\) and Foxp3\(^-\) clusters, despite the small relative size of the Foxp3\(^+\) cluster [130].

However, for many of these algorithms it is unclear whether the same gates generated are able to be translated on to negative control datasets for comparison. In 2001, Roederer and colleagues developed a multisample, multiparametric automated gating method based on probability binning [221, 220, 219]. When one attribute is considered, the histogram for a relevant control sample is divided into bins containing equal numbers of events [221]. These bins are then applied to the test sample and a normalised $\chi^2$-squared statistic is calculated for the difference between the contents of each bin [221]. When looking at more than one attribute, the algorithm picks the parameter with the highest variance, then divides the data in half around about the median [220]. The algorithm then repeats the halving step about the median of the parameter with the highest variance in the resultant two data subsets [220]. The limitations of this are that large, homogeneous populations may be more likely to be halved, and the algorithm may bring out deviations in this population, ignoring smaller outlying populations [220]. This method will also become less sensitive to subtle changes in fluorescence intensity the more channels that are being considered at any one time [220], likely to be a problem for our nTreg data. For the sake of computational efficiency, this seems to ignore the rich individual data. Gating then occurs by adding all of the highly differential bins together in order to enumerate one cluster from another [219]. It is possible using this technique, therefore, to enumerate statistically significant populations from the control distribution, or to say if two samples are statistically similar or different [221]. The difficulty is assigning a threshold for
significant difference [219]. Previous attempts to cluster univariate data by histogram subtraction were often also inappropriately sensitive [171].

Using the Multisample Distancing Platform (multivariate) in FlowJo, an implementation of frequency difference gating, we compared our test sample with the F-Foxp3 control. Only bins including a positive difference were included in the gates generated; the first limitation of this platform for nTreg analysis is that we enumerate events with low CD127, but high Foxp3 and CD25 expression. Whilst one can select bins with a negative difference and multiple parameters can be considered, the combination of parameters must be either all positive or all negative. In three experiments, two varying the $\chi^2$-squared threshold, the correct phenotype was not separated from the rest of the population in any instance. An example of unsuccessful nTreg clustering at two $\chi^2$-squared gating thresholds is shown in Figure 1.3. Using the isotype control as a benchmark, a high number of false positives are enumerated due to the increase in autofluorescence of the negative events.
Figure 1.3: Backgating analysis of FlowJo’s “Population Comparison” tool (multivariate) on an example test nTreg dataset at two chi squared lower gating cut-off values.

Whole blood stained for CD4, CD25, CD127 and Foxp3 was acquired by flow cytometry. Samples were compensated and CD4$^+$ lymphocytes were gated before running Population Comparison to enumerate nTreg. a) Unsuccessfully clustered nTreg (black) shown backgated onto previously gated CD4$^+$ lymphocytes (grey) on a CD25/Foxp3 plot. b) Unsuccessfully clustered nTreg backgated onto CD4$^+$ lymphocytes on a CD25/CD127 plot. c) Unsuccessfully clustered nTreg backgated onto CD4$^+$ lymphocytes on a CD25/Foxp3 plot. d) Unsuccessfully clustered nTreg backgated onto CD4$^+$ lymphocytes on a CD25/CD127 plot. a-b) Lower Gating Cut-Off Value: 0.01899, % Cells Gated = 73.74. c-d) Lower Gating Cut-Off Value: 0.5, % Cells Gated = 3.457.

1.9.4 Algorithms for visualisation of multidimensional data

This section discusses examples of novel visualisation algorithms that have recently been developed to better facilitate manual gating of multidimensional data. Firstly, Kaluza$^\text{®}$ v1.2 (Beckman Coulter) contains a radar plot function allowing for visualisation of $n$-channel data, however the ability to gate or cluster what is observed is limited to manual polygonal gating (Figure 1.4). 

Infinicyt$^\text{TM}$ v1.6 (Cytognos) also supports 3-dimensional data visualisation with an ability to draw a freehand gate around a population of interest (Figure 1.5a) and its multicolour 3D density plot and automatic population separator (APS) functions (Figures 1.5b and c respectively) are also interesting techniques for visualising multiparametric data, $n$.
parameters at a time. The APS uses PCA to determine the best presentation of the data such that distinct populations can emerge. However, the ability for the user to gate on emergent populations is even more limited; freehand gates have to be drawn on 2D plots following their observation through the multicolour 3D density plot or APS. APS plots are particularly difficult to interpret as their axes do not correspond to a single antigen.

The tool viSNE has very recently been released as a flow and mass cytometry data visualisation platform which maps high dimensional data into 2 dimensions [12]. viSNE plots are similar to scatter plots, and the pairwise distances between datapoints in all dimensions are used to position each event [12]. viSNE is based on a t-Distributed Stochastic Neighbour Embedding algorithm where a 2D projection of the high dimensional data is selected for optimal separation between populations [12]. Unlike a traditional 2D gating method, it is possible to observe all parameters on this single map [12]. Major lineages were easily separated [12]. Interestingly, viSNE seems well equipped to observe cells which lack a particular lineage marker, e.g. have downregulated it for some reason, but otherwise resemble the cells of a particular population [12]. Therefore, it avoids the problem of traditional gating misclassifying events, and like SPADE, captures the heterogeneity of populations of interest [12]. However, viSNE is limited to datasets of approximately 30,000 events by its high computational cost [12], and this is insufficient for nTreg data.

These novel methods of data visualisation are immediately appealing to the flow cytometry user because of their integration into established commercial software platforms, and similarity to conventional manual gating. However, the subjective and laborious aspects of manual gating remain, made more confusing by the low dimensional representation of multidimensional data. Also, inevitably, information is still lost by mapping multidimensional data into two dimensions [12].
Whole blood stained for CD4 was acquired by flow cytometry. Compensated, ungated events were viewed using the radar plot with forward scatter (FS) on the $x$-axis, side scatter (SS) on the $y$-axis and FL5 on the $z$-axis. FL5 corresponds to CD4 expression. A polygonal gate was manually drawn to enumerate CD4$^+$ lymphocytes.

Figure 1.4: *Kaluza®* radar plot for 3-dimensional CD4$^+$ lymphocyte gating.
Whole blood stained for CD4, CD25, CD127 and Foxp3 was acquired by flow cytometry. Samples were compensated and CD4⁺ lymphocytes were gated before investigating the Infinicyt™ platforms. a) Example freehand gating of nTreg on a 3-dimensional dot plot. b) Multicolour 3D density plot of nTreg data. FL1 corresponds to Foxp3 expression, FL2 corresponds to CD127 and FL4 corresponds to CD25. c) APS used to visualise nTreg data.
1.9.5 Algorithms for multisample comparison and presentation

Some of the algorithms already described can be used for multisample comparison. For example, using FLAME, the modes of each population for each sample within an experiment can be determined and metaclustered [212]. This, however, has not been demonstrated on nTreg data [212]. Were our nTreg measurements always acquired on the same machine with daily calibration to always read the same signal intensity, this might have been a very useful function for comparing nTreg numbers in donors of different ages. However, even experiments performed on different days within the same study will include various sources of experimental error and bias, therefore limiting the FLAME’s use for metaclustering.

Fiser and colleagues also suggested using their flow cytometry-adapted hierarchical clustering algorithm combined with support vector machine classification to automate analysis of new samples using previously clustered results as training datasets [84]. Again, this would not be appropriate for the nTreg ageing data presented in this thesis as the flow cytometer was not routinely calibrated [84]. Similarly, flowScape templates can be manually constructed using test datasets of the specific type of interest to enable automated gating of subsequent data files [216]. However, whilst these approaches permit the comparison of different data files, the variation between files is not actually considered in the clustering process. Frequency difference gating, as previously described, actually uses a control dataset in the gating process of the test dataset.

Lugli and colleagues believe that it is necessary to reassess the habit of manual gating and to consider maturing automated methods, which integrate the expertise of the operator [171]. Lugli and colleagues published a PCA-based automated comparison algorithm for comparing manually gated populations across multiple data files [170]. Although this algorithm does not directly deal with the automatic gating of individual events within flow cytometry datasets, it allows for automated classification of subjects according to specified populations [170]. To the best of our knowledge this is the only paper where donors of different ages were automatically compared. PBMCs were stained for CD3, CD4, CD8, CD45RA, CD95, CCR7, CD127 and CD38 [170]. Following manual gating combined with automated Boolean combination to identify all possible phenotypes, a 2-dimensional matrix was generated, with each column as an individual subject and each row as a phenotype [170]. PCA was then applied to reduce the dimensionality of the dataset by identifying new
variables (so called principal components), which represented combinations of variables that described the majority of the variation between samples [170]. Hierarchical clustering then highlighted differences between young donors and centenarians, with middle aged donors having characteristics of both extreme age groups [170]. Specifically, the memory T cell population of younger donors had more CD127\(^+\)CD95\(^-\) T cells, whereas older donors had more CD127\(^-\)CD95\(^+\) T cells [170]. A similar automated analysis of pregated populations between samples was conducted by Larsen and colleagues[147].

RchyOptimyx is another algorithm designed to optimise a gating strategy for investigating populations according to their clinical relevance [7]. Having either manually or automatically gated a population, the best cellular hierarchy is identified using dynamic programming [7]. It can also be used to refine a staining panel by identifying redundant parameters and therefore it is most useful following a large-scale exploratory study using a large panel of reagents [7]. One could criticise this “chicken-and-egg” style of approach; data generated from conducting a large study with multiple samples and high reagent costs is necessary to inform the staining panel, reduce reagent use and optimise the gating strategy after the work has been done. As RchyOptimyx was developed for use with the flowType automated gating algorithm it currently requires that cells can be partitioned as positive or negative, and that the flow cytometer is routinely calibrated [7]. For nTreg analysis, the method of choice needs to be robust to markers expressed as a gradation.

viSNE is also specialised for multisample comparison, and the Jenson-Shannon divergence was used to compare maps of three healthy donors, finding negligible differences [12]. However, the viSNE map of AML patient samples showed unique, deformed shapes compared to healthy donors [12].

Simpler still is the analysis tool SPICE (Simplified Presentation of Incredibly Complex Evaluations), a novel platform optimised for analysis of the antigen-specific immune response [196, 171]. Again, it relies upon a manual gating strategy, but simplifies the comparison of multiple, heterogeneous data files [171]. It is possible to analyse trends in any combination of pre-gated populations with respect to a variable of choice, for example vaccine strategy [171]. Betts and colleagues showed that there was a high frequency of polyfunctional HIV-specific CD8 T cells in patients with non-progressing HIV using the SPICE pipeline [30, 171].
1.10 Selection of SamSPECTRAL

1.10.1 Justification

Although we are not the first to investigate automated analysis of nTreg flow cytometry data [130, 212, 216], we address the need for an algorithm that is adequate for the current most specific nTreg phenotype: CD4$^+$, CD127$^{\text{low}}$, CD25$^{\text{high}}$, Foxp3$^+$, a case study presenting challenges typical of flow cytometry data. The pre-gated CD4$^+$ lymphocyte datasets are large (from 23,888 to 179,265 rows) and multiparametric (3 attributes: CD127, CD25, Foxp3). The population of nTreg overlaps significantly with the non-nTreg cluster with respect to all 3 attributes, the populations are non-spherical and the data is noisy. Therefore, not all of the algorithms discussed would be able to improve nTreg analysis.

The inadequacy of manual gating for nTreg is demonstrated at the beginning of Chapter 5, therefore algorithms requiring initial manual gating as an input would limit the potential for improvement. The aim was to objectivise and automate the gating step, but also to explore the data and learn the boundaries of the nTreg population with respect to CD25, CD127 and Foxp3. Therefore, the choice of algorithm was restricted to clustering algorithms for data exploration and population detection.

As the populations within nTreg data are overlapping, non-spherical and the population of interest is of low frequency, the selected algorithm needed to be robust for use on data with these characteristics. A spectral clustering algorithm such as SamSPECTRAL was most appealing being non-parametric and therefore less sensitive to outliers, noise or unusual cluster shapes [295]. However, as the nTreg population is rare, like Pyne and colleagues [212] we were initially opposed to using a data reduction technique, necessary as classical spectral clustering is too computationally expensive for flow cytometry data [295]. SamSPECTRAL was a clear choice as its faithful sampling step preserves density information. Other features about SamSPECTRAL were beneficial, such as its ability to automatically merge clusters drawn from the same populations if the ratio of inter-cluster similarity to intra-cluster similarity is sufficiently large [295].

As well as being conceptually appropriate, the ability of SamSPECTRAL above other algorithms to cluster data similar to our nTreg datasets has been previously demonstrated. SamSPECTRAL correctly identified overlapping populations in FSC/SSC/telomere length
datasets, and alongside FLAME, correctly clustered a graft versus host disease dataset characterised by non-elliptical shaped populations [295]. In a stem cell enumeration problem, where low density populations of interest were located next to dense ones, SamSPECTRAL successfully detected the low density population whereas FLAME often merged them into adjacent clusters [295]. The stem cell enumeration problem was also an example of rare population detection (<2% datapoints), and SamSPECTRAL successfully clustered 79% of the data files correctly; FLAME only achieved 32% [295]. The failed attempts by SamSPECTRAL were on data files where the stem cell population was less than 0.15% [295]. Therefore, we selected SamSPECTRAL for further investigation on nTreg datasets.

However, as these results were presented in the paper introducing SamSPECTRAL, it could be argued that the authors finely tune their use of SamSPECTRAL and do not apply the same attentiveness to parameter selection for FLAME, but there is insufficient evidence in the publication that this is the case. In the comparative FlowCAP study discussed later in this chapter, where each developer was given the same dataset and could optimise the use of their own algorithm, SamSPECTRAL still performed favourably [291, 6]. Qui and colleagues, the developers of SPADE, also describe SamSPECTRAL as one of the only algorithms with a mechanism for detecting rare subpopulations [214]. Our study contributes an independent evaluation of SamSPECTRAL’s performance on nTreg data.

However, the SamSPECTRAL method itself does not contain an automated gating step and does not allow for multisample comparison without extension. Therefore, we added an elliptical gating step to permit standardisation of nTreg enumeration against false positives estimated through a comparison with control datasets. This method is explained in Chapter 2 and the results are presented in Chapter 5.

1.10.2 Detailed description

As previously described, the main desirable feature about SamSPECTRAL is its faithful sampling step which reduces the number of features used for spectral clustering. Once this has occurred, a graph is built with each community representative as a vertex [295]. The edges between each representative are weighted according to the density of data points surrounding them and the calculated similarity between communities [295]. The similarity between two points is inversely proportional to the squared Euclidean distance between them, and depends on the $\sigma$.
parameter ($\sigma$) that determines how sensitive similarity is to changes in squared Euclidean distance, $D^2(p_i, p_j)$ [295]. To compute the similarity between two vertices $i$ and $j$, the following formula is used [295]:

$$s_{i,j} = e^{-\frac{D^2(p_i, p_j)}{2\sigma^2}}$$

The similarity $S$ between two communities is the sum of the pairwise similarities between all individuals from each community [295]. The following is used to calculate the similarity between community $c$, of which $i$ is a member, and $c'$, of which $j$ is a member [295]:

$$S_{c,c'} = \sum_{i \in c} \sum_{j \in c'} s_{i,j}$$

The adjacency matrix, $A$, is then normalised according to the following, where $D$ is the diagonal matrix containing the sum of the weights adjacent to each vertex [295]:

$$\tilde{A} = D^{-1/2} A D^{-1/2}$$

The eigenspace is then computed to find the eigenvalues, $V_i$, and eigenvectors, $\lambda_i$ [295]:

$$\tilde{A} \vec{V}_i = \lambda_i \vec{V}_i$$

It is a normalised “community representative by $k$” matrix of eigenvectors that is clustered by $k$-means [295], with number of clusters estimated as the horizontal coordinate of the knee point of the eigenvalues curve or fixed to 15 (example given in Figure 2.9). A post-clustering processing stage is included to merge clusters thought to belong to the same population, controlled by varying the separation.factor parameter [295]. Specifically, the ratio of between cluster similarity to within cluster similarity is computed, and if this is greater than the separation.factor, the two clusters are merged and the list of clusters is updated. Finally, as each raw data point is “registered” to a community, it is possible to extract a class label for each datapoint [295].
1.11 Determining data clusterability and cluster validity

1.11.1 Introduction

Clustering algorithms, when applied to a dataset, will return a separation of the data into the number of clusters specified, regardless of whether a resultant clustering structure is a true representation of the underlying data structure [248]. This encouraged us to consider whether we could objectively decide whether any given dataset was clusterable or not.

A test for uniformity in multidimensional data was previously outlined by Smith and colleagues [248]. The test involved analysis of the dataset to determine its shape within the feature space [248]. Then a uniform dataset of the same size was synthesised, bound by the derived data shape, or "convex hull" [248]. A comparison of the synthetic versus real data was then conducted by first pooling both samples, drawing a minimal spanning tree with the data points as nodes, and computing a Friedman-Rafsky statistic [248]. If the statistic fell below a threshold, the null hypothesis of a uniform distribution was rejected [248].

Assuming an attribute is normally distributed, it would be appropriate to use multivariate algorithms based on traditional tests of skew or kurtosis, such as KS or SW [261]. FLAME, for example, is a skew t-mixture model fitting method [212]. Perhaps an approach for testing for uniformity in multidimensional data could be adapted for use with more complicated distributions, such as skew t-distributions. This, however, is beyond the scope of this thesis.

Finding an appropriate criterion for testing whether a dataset is clusterable relies on whether the data can be likened to a particular distribution. As experts cannot come to an agreement as to which distribution best describes flow cytometric data [212, 295, 257], such a criterion is unlikely to be generally accepted.

There are caveats with these methods; firstly, how the sampling window is defined directly impacts on the performance of the test [248]. A sampling window which is too large may contain a lot of white space surrounding the data, and view the data as a single cluster when it is in fact uniform. Also, if a sampling window is so small it only contains a few datapoints it may be considered uniform when it comes from one of two larger clusters. At a level which is smaller still, containing 3 datapoints, these may be viewed as 3 separate clusters. Secondly, the more attributes are considered, the less likely the null hypothesis is to be rejected [248].
This relates to an increase in the mean of the test statistic, in turn related to the convex hull being under-representative of the ideal sampling window [248].

Visual inspection has also been used to assess data clusterability, for example Zare and colleagues justified the difficulty their algorithm had with some particular data files by highlighting that the size of the rare population was smaller in those instances [295].

As well as objectively defining whether the data is clusterable or not, it is also ideal to determine objectively whether a clustering algorithm has generated a solution that reflects the underlying data. There are three categories of techniques to assess cluster validity: external, internal and relative [262, 113]. External criteria compare a clustering structure to a previously set partition [113]; this could mean a comparison with manual gating, or a comparison with a completely random partition [113]. Internal criteria relate a clustering structure to features within the data, for example a proximity matrix [113]. Relative criteria involve comparing a clustering structure from an algorithm with a set of parameter values to a clustering structure generated by the same algorithm with different parameter settings, and can therefore be used to select optimal parameters for an algorithm [113].

The following subsections define some commonly used cluster validity indices which we evaluate for our synthetic and nTreg data in Chapter 5.

1.11.2 Average silhouette width

The average silhouette width (ASW) is an attractive measure of validity as its value does not depend on the clustering algorithm used to obtain a given partition of data [229]. It is used by FLAME in its metaclustering step as a relative measure to determine the optimum number of clusters [212]. A cluster’s silhouette is based on its compactness and distance from other clusters, and is outlined below according to Rousseeuw [229].

If $A$ is one cluster and $C$ is another:

- $a(i) = \text{average dissimilarity between one point } (i) \text{ and all other points within } A$
- $d(i, C) = \text{average dissimilarity between } (i) \text{ and all points within } C$
- $b(i) = \text{minimum } d(i, C)$

The silhouette width is then calculated as (taken from [229]):

\[ \text{silhouette width} = \frac{a(i) - b(i)}{\max(a(i), b(i))} \]

53
\[ s(i) = \begin{cases} 
1 - a(i)/b(i) & \text{if } a(i) < b(i), \\
0 & \text{if } a(i) = b(i), \\
b(i)/a(i) - 1 & \text{if } a(i) > b(i). 
\end{cases} \]

Therefore, \(-1 \leq s(i) \leq 1\).

When \( s(i) \) is close to 1, \( a(i) \) is much smaller than \( b(i) \) and therefore we deduce that the clustering is strong [229]. If \( s(i) \) is near 0, \( a(i) \) and \( b(i) \) are equal, and therefore \( i \) could belong to either \( A \) or \( C \) [229]. If \( s(i) \) is near -1, an object has been misclassified and should actually be in cluster \( C \) [229].

### 1.11.3 Hubert’s \( \Gamma \) statistic

Hubert’s \( \Gamma \) statistic is another common index that can be used as an external or internal validity measure. Here we use the normalised version as an internal measure. If \( M \) is the total number of pairs, and \((i, j)\) are the elements of \( X \), the proximity matrix, and \( Y \), the matrix of distances between allocated cluster centroids, the normalised Hubert’s \( \Gamma \) statistic is as follows, taken from Halkidi and colleagues [113]:

\[
\Gamma = \left[ \frac{(1/M) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X(i,j) - \mu_x)(Y(i,j) - \mu_y)}{\sigma_x \sigma_y} \right].
\]

\( \mu_x \) and \( \mu_y \) are the means of the \( X \) and \( Y \) matrices. \( \sigma_x \) and \( \sigma_y \) are the standard deviations of the \( X \) and \( Y \) matrices. The higher the value, the stronger the similarity between clustering partitions [113].

### 1.11.4 Entropy

Another common measure of cluster validity is entropy, where the lower the value of the index, the less random and more reliable the clustering partition is for that dataset. The entropy definition below is from Meilä 2007 [181]:

\[
H(X) = - \sum_{k=1}^{K} \sum_{v \in A_j} P(k) \log P(k)
\]

where
\[ P(k) = \frac{n_k}{n} \]

and where there are \( K \) clusters and \( n \) datapoints.

### 1.11.5 Calinski and Harabasz index

Finally, the Calinski and Harabasz Index is a relatively simple calculation based on the ratio of inter- to intra-cluster distance, taken from Charrad and colleagues [52]:

\[ CH(k) = \frac{B/(k - 1)}{W/(n - k)} \]

Where \( B \) is the sum of the squared distances between clusters, \( W \) is the sum of the squared distances within the clusters, \( k \) is the number of clusters and \( n \) is the number of points [52]. The higher the Calinski and Harabasz Index, the better a clustering partition is.

### 1.11.6 Clustering validation previously used in flow cytometry

In a recent paper, Aghaeepour and colleagues demonstrated the use of an external validity measure, the “clustering F-measure”, comparing several clustering algorithms to the manual gating results from a number of experts on the Flow Cytometry: Critical Assessment of Population Identification Methods (FlowCAP) challenges [6]. Firstly, the recall (Re) and precision (Pr) are calculated. Re, which resembles a classical sensitivity test, divides the number of events the algorithm correctly assigned to a particular cluster by the number of events assigned to that cluster by a manual method [6]. Pr, which resembles a classical positive predictive value test, divides the number of events the algorithm correctly assigned to a particular cluster by the total number of events the algorithm assigned to that cluster [6]. This method therefore assumes perfect manual gating. The equations below are taken directly from [6].

\[ Re = \frac{TP}{TP + FN} \quad Pr = \frac{TP}{TP + FP} \]

The F-measure is a function of the above:
\[ F = \frac{2 \times Pr \times Re}{Pr + Re} \]

This calculation is extended to compare the complete set of gated populations \( C = \{c_1, c_2, ..., c_n\} \) with the complete clustering structure \( K = \{k_1, k_2, ..., k_m\} \). The matrix of comparisons is \( M = [a_{ij}], i \in [1, n] \) and \( j \in [1, m] \). \( Pr(c_i, k_j) = \frac{a_{ij}}{|k_j|} \) and \( Re(c_i, k_j) = \frac{a_{ij}}{|c_i|} \), where \( |c_i| \) is the number of elements in \( c_i \). The F-measure for comparing clusters is therefore:

\[ F(c_i, k_j) = \frac{2 \times Pr(c_i, k_j) \times Re(c_i, k_j)}{Pr(c_i, k_j) + Re(c_i, k_j)} \]

This must be compared for every combination of clusters, and the largest F-measure \((\text{max})\) normalised by \( k_j \) is the reported measurement. This “total” F-measure is therefore:

\[ F(C, K) = \sum_{c_i \in C} \frac{c_i}{N} \max_{k_j \in K} \{F(c_i, k_j)\} \]

1.12 Discussion

The previous section began by demonstrating a need to improve nTreg flow cytometry data analysis and led to a literature review of currently available computational methods to cluster, improve the visualisation of, and compare between multiparametric flow cytometry datasets. The breadth of currently available techniques and their main limitations has been demonstrated, and the selection of SamSPECTRAL for automating the analysis of a challenging example of real flow cytometry data has been justified. Towards the end of Chapters 5 and 6 we make recommendations for how to prioritise the development of new techniques and encourage their uptake in routine laboratory flow cytometry data analysis.

Clustering itself, however, can also be viewed as a subjective process, as different clustering algorithms will cluster the same dataset differently [129], thus the selection of an algorithm that outperforms another is a subjective process even if the division of the data by the algorithm is objective. Some commonly used cluster validity indices have been discussed in this chapter. Many studies verify their assumptions about
whether their algorithm has clustered correctly or not by visually inspecting a subset of their experiments [13]. It has been argued that expert opinion and what is known about the immunology is a more robust way of determining cluster number, more so than currently available algorithms to optimise cluster number [130]. Often, in the field of data mining, the resultant classification is validated at least in part by the domain expert [3, 54]. Although clustering tendency “is in the eye of the beholder” [128], it is important to compare visual inspection with objective measurements in order to quantify improvement.

Many authors validate their findings against solutions derived from expert manual gating, the process they argue is subjective and error prone [214, 6, 12, 84]. However, others avoid this validation strategy and offer more qualitative comparisons [216]. Usually, in clinical trials of pharmaceuticals, a novel agent is compared to the gold standard method of treatment, but the aim is to demonstrate a significant difference from the benchmark, not similarity to the benchmark. In contrast with clinical trials, the challenge for validating clustering algorithms is in demonstrating a significant improvement, as the endpoints - the class labels - are unknown. Arguably, developing better cluster validity indices to demonstrate that clustering represents the underlying data more truthfully than the manual method is the best way to advance the field. However, developing such methods is outside the scope of this thesis.

A disadvantage of using internal measures of cluster validity is that they compare a clustering partition to a full distance matrix [113], impossible to compute on large flow cytometry datasets. However, methods such as distcritmulti available within the fpc in R, separate datasets into a specified number of subsets, calculate the distance matrix and subsequently validity indices for each subset, then report the average validity index of these subsets [121]. Although analysing cluster validity in large datasets is a significant research area in itself [119, 118, 110], Havens and colleagues, advise that in most cases a uniform sampling method is most appropriate [119]. As each cluster validity index performs slightly differently, the selection of a validity index becomes subjective and therefore adds bias. Also, as with any statistic, the results of clusterability and cluster validity statistics are more convincing when the data is nicely clusterable [248].

This review demonstrates that the field of bioinformatics is already working with the field of flow cytometry to develop automated, objective analysis methods, enabling experts to identify features of multiparametric
flow cytometry data [171]. There are now a number of tools in the flow cytometry analysis armoury to suit the needs and computational expertise of the user. However, it is not immediately clear which is best for the testing of certain hypotheses, and whether the process of analysis can be fully automated.
Chapter 2

Materials and methods

2.1 Selection of healthy donors

In accordance with University of Nottingham Medical School Research Ethics Committee approval, donors were recruited through a poster campaign and by presentations given to large groups at Nottingham University Hospitals and in the wider community. Donors were excluded if they: were under 18 years of age, had previously had an adverse reaction to blood being taken, had recently given a large amount of blood, suffered from a chronic infection such as Hepatitis, HIV/AIDS or TB, were anaemic, had a long-term inflammatory condition, were currently taking anti-inflammatory medication, had an infection or vaccination in the previous 2 weeks, had suffered a heart attack or stroke in the previous 2 weeks, had been diagnosed with or received treatment for cancer in the previous 5 years or were currently taking illegal drugs or under the influence of alcohol.

Informed consent was obtained and donors completed a questionnaire accounting for demographic variables including age, sex, ethnicity and occupation. If retired, donors reported their previous occupation. The questionnaire also asked donors to self report their relevant medical history, including long-term medical conditions (e.g. asthma, allergies, diabetes, hypertension, heart failure), current medication, past history of cancer, chronic infections or inflammatory conditions and smoking habits. Ethnicity and occupation were categorised using guidelines from The UK Office for National Statistics [86].

This research used a single whole blood sample (10ml) per donor
withdrawn from a vein in the cubital fossa into one BD Vacutainer\textsuperscript{TM} with heparin.

### 2.2 Whole blood culture

We observed in preliminary experiments that resting T helper cells in whole blood did not secrete cytokines, but that stimulation increased the number of Foxp3\textsuperscript{+} events. The resting Foxp3\textsuperscript{+} T cells are classical nTreg, therefore we stained unstimulated sample for our nTreg markers of interest, and stimulated sample for our measurements of cytokine-secreting T helper cell numbers. A routinely used PMA and ionomycin stimulation was used to provoke T helper cells to secrete their signature cytokines \[152\]. PMA activates protein kinase C-mediated signal transduction, whilst ionomycin raises intracellular calcium \[80, 208\].

500\(\mu\)l RPMI alone was added to wells of a 48-well flat-bottomed culture plate for nTreg staining. PMA (100ng/ml) was added to wells intended for intracellular cytokine staining. Ionomycin (2ug/ml) was also added to these wells for stimulation together with brefeldin A according to manufacturer’s instructions (BD Golgiplug\textsuperscript{TM}). 500\(\mu\)l whole blood was mixed into all wells and the plate was incubated overnight in a 37°C incubator, 5% CO\textsubscript{2}. After 20 hours, 500\(\mu\)l of the supernatant from conditions without brefeldin A was aliquoted into 96-well plates and stored at -20°C prior to measurement of cytokine secretion by enzyme-linked immunosorbent assay (ELISA) with R&D Systems\textsuperscript{®} Human DuoSet kits. The remaining whole blood sample was used for extra- and intracellular staining.

### 2.3 Flow cytometry

#### 2.3.1 Staining method and antibody panel

Antibodies were added to the samples at the concentrations shown in Table 2.2, in the combinations shown in Table 2.1, and mixed by vortexing before incubating for 15 minutes in darkness at room temperature. 100\(\mu\)l CAL-LYSE\textsuperscript{TM} solution (Invitrogen) was added per 100\(\mu\)l whole blood before incubating for a further 10 minutes. 1ml distilled water per 100\(\mu\)l whole blood was added to quench the red blood cell lysis and the sample was rested for 5 minutes. Samples were washed by centrifugation at 180xg
for 10 minutes. The supernatant was discarded and the cells resuspended in 1ml BioLegend 1X Foxp3 Perm Buffer per 100µl whole blood and incubated for 15 minutes. Following further washing, cells were resuspended in BioLegend 1X Foxp3 Perm Buffer (volumes shown in brackets in Table 2.1) in order to be stained with intracellular antibodies.

Cells were incubated at room temperature in darkness for 30 minutes before a final wash. The supernatant was discarded and the cells resuspended in 700µl FACS buffer (1% foetal calf serum, 0.1% sodium azide in PBS). Flow-Count\textsuperscript{TM} Fluorospheres (Beckman Coulter) were thoroughly vortexed in their container and 25µl (≈25000 Fluorospheres) were added to test samples (Figure 2.1). Single stained controls were performed for each fluorochrome on BD\textsuperscript{TM} Compensation Particles (Anti-Mouse Ig, Positive/Negative Control (FBS) Compensation Particles Set, BD Biosciences). Unstained cells and single-stained controls were used to set voltages and for pre- and post-acquisition compensation [177]. “Fluorescence Minus One” (FMO) controls were required in order to objectively set gates on the populations of interest as detailed in Table 2.1 [177]. FMOs are subsequently referred to as the antigen that is not stained for, prefixed with “F-“, such as “F-Foxp3” is without anti-Foxp3 antibody.

In the following chapter developmental work supporting this protocol design is presented.

2.3.2 Measurements of absolute number

Each CD4\textsuperscript{+} T cell subset was measured as a percentage of total CD4\textsuperscript{+} lymphocytes and an absolute frequency (cells/µl whole blood). The latter involved adding a known number of Flow-Count\textsuperscript{TM} Fluorospheres to each sample and measuring the proportion of these that were acquired with the sample. Figure 2.1B shows box gating of events corresponding to FlowCount Fluorospheres\textsuperscript{TM}. The following equations were used to standardise the absolute number measurements generated using the proportion of events with specific staining and isotype control false positive proportion:
<table>
<thead>
<tr>
<th>Description</th>
<th>Beads/cells</th>
<th>FL1</th>
<th>FL2</th>
<th>FL3</th>
<th>FL4</th>
<th>FL5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Unstained</td>
<td>100µl Stim</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 AF488 control</td>
<td>Beads</td>
<td>AF488 α-Foxp3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 PE control</td>
<td>Beads</td>
<td></td>
<td>PE α-CD86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 ECD control</td>
<td>Beads</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ECD α-CD4</td>
</tr>
<tr>
<td>5 PE-Cy5 control</td>
<td>Beads</td>
<td></td>
<td></td>
<td></td>
<td>PE-Cy5 α-CD8</td>
<td></td>
</tr>
<tr>
<td>6 AF647 control</td>
<td>Beads</td>
<td></td>
<td></td>
<td></td>
<td>AF647 α-IFNγ</td>
<td></td>
</tr>
<tr>
<td>7 PE-Cy7 control</td>
<td>Beads</td>
<td></td>
<td></td>
<td></td>
<td>PE-Cy7 α-CD4</td>
<td></td>
</tr>
<tr>
<td>8 F-Foxp3</td>
<td>300(300)µl unstim</td>
<td>AF488 mIgG1</td>
<td>PE α-CD127</td>
<td>PE-Cy5 α-CD25</td>
<td>PE-Cy7 α-CD4</td>
<td></td>
</tr>
<tr>
<td>9 F-CD25</td>
<td>300(300)µl unstim</td>
<td>AF488 α-Foxp3</td>
<td>PE α-CD127</td>
<td>PE-Cy5 mIgG1</td>
<td>PE-Cy7 α-CD4</td>
<td></td>
</tr>
<tr>
<td>10 F-CD127</td>
<td>300(300)µl unstim</td>
<td>AF488 α-Foxp3</td>
<td>PE mIgG1</td>
<td>PE-Cy5 α-CD25</td>
<td>PE-Cy7 α-CD4</td>
<td></td>
</tr>
<tr>
<td>11 CD4, CD25, Foxp3, CD127, beads</td>
<td>300(300)µl unstim</td>
<td>AF488 α-Foxp3</td>
<td>PE α-CD127</td>
<td>PE-Cy5 α-CD25</td>
<td>PE-Cy7 α-CD4</td>
<td></td>
</tr>
<tr>
<td>12 F-IL17</td>
<td>400(100)µl stim</td>
<td></td>
<td>PE mIgG1</td>
<td></td>
<td>PE-Cy7 α-CD4</td>
<td></td>
</tr>
<tr>
<td>13 CD4, IL-17, beads</td>
<td>400(100)µl stim</td>
<td></td>
<td>PE α-IL-17</td>
<td></td>
<td>PE-Cy7 α-CD4</td>
<td></td>
</tr>
<tr>
<td>14 F-IL-10</td>
<td>600(100)µl stim</td>
<td></td>
<td>PE rIgG1</td>
<td>AF647 α-IFNγ</td>
<td>PE-Cy7 α-CD4</td>
<td></td>
</tr>
<tr>
<td>15 F-IFNγ</td>
<td>600(100)µl stim</td>
<td></td>
<td>PE α-IL-10</td>
<td>AF647 mIgG1</td>
<td>PE-Cy7 α-CD4</td>
<td></td>
</tr>
<tr>
<td>16 CD4, IL-10, IFNγ, beads</td>
<td>600(100)µl stim</td>
<td></td>
<td>PE α-IL10</td>
<td>AF647 α-IFNγ</td>
<td>PE-Cy7 α-CD4</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: Staining panel.
Absolute Number of Specific Events =

\[ \text{Total}_t \times \frac{(\text{Positive}_t)}{\text{Proportion of Sample Counted}} - \frac{(\text{Positive}_c)}{\text{Total}_c} \]

where \( \text{Positive}_t \) represents positive events in the test dataset, \( \text{Positive}_c \) represents positive events in the control, \( \text{Total}_t \) represents total events in the test dataset, \( \text{Total}_c \) represents total events in the control, and

\[ \text{Proportion of Sample Counted} = \frac{\text{Count Beads Counted}}{\text{Count Beads Added}} \]

Figure 2.1: Absolute numbers of CD4\(^+\) lymphocytes (/µl whole blood) in donors of different ages.
Figure 2.1: Stimulated whole blood stained for CD4 with a known volume of FlowCount™ Fluorospheres added was acquired by flow cytometry. a) Compensated CD4+ events gated from total whole blood lymphocytes (see gating strategy later in this chapter). b) FlowCount™ Fluorospheres gated on the uppermost axis where events show green on a pseudocolour plot with forward scatter along the X axis and any fluorescence channel on the Y axis. c) Absolute numbers of CD4+ lymphocytes (cells/µl) shown against donor age (n=90, r= 0.0798, p=0.455).

2.3.3 Acquisition and analysis

Samples were acquired in a CYTOMICS™ FC500 flow cytometer with CXP software (Beckman Coulter). 1-2x10^4 events were acquired from unstained and single-stained control tubes. 5.2x10^5 ± 8.6x10^4 events were acquired for unstimulated test samples and 4.1x10^5 ± 1.3x10^5 were acquired for stimulated test samples. Post-acquisition analysis was performed using FlowJo.

Gating Strategies

Stringent gating strategies were designed to enumerate our CD4+ T cell populations of interest. FMOs were essential controls for these flow cytometry assays where the boundaries between positivity and negativity were ambiguous in overlapping or small populations [201, 176]. However, conventional FMOs do not give an account of how non-specific Fc-portion binding affects the number of events measured [176]. Therefore fluorochrome-, manufacturer- and concentration-matched isotype controls were used. Details of all antibodies can be found in Table 2.2. The volume given in brackets in the third column refers to the volume of BioLegend 1X Foxp3 Perm Buffer the cells were resuspended in prior to adding antibodies to intracellular antigens. Tubes 11, 13 and 16 were the test panel of antibodies to enumerate CD4+CD127loCD25highFoxp3+, CD4+IL17+ and CD4+IFNγ+IL10+ events in whole blood respectively. Tubes 8-10 were FMOs for tube 11, tube 12 was an FMO for tube 13 and tubes 14 and 15 were FMOs for tube 16. A typical pre-acquisition compensation matrix is shown in Figure 2.3. The samples were acquired slightly undercompensated, then FlowJo autocompensation was used to finalise compensation post-acquisition for both the AF647 and PE-Cy5 panels separately. Owing to machine drift it was not possible for the exact same conditions to be repeated across all assays; all voltage and compensation settings were checked and adjusted if necessary for each experiment. Complete gating strategies are given in Figures 2.2, 2.3 and 2.4.

64
<table>
<thead>
<tr>
<th>Specificity</th>
<th>Fluorochrome</th>
<th>Clone</th>
<th>Supplier</th>
<th>Working concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4</td>
<td>PE-Cy7</td>
<td>RPA-T4</td>
<td>BD</td>
<td>0.5 µg/ml</td>
</tr>
<tr>
<td>CD25</td>
<td>PE-Cy5</td>
<td>BC96</td>
<td>BioLegend</td>
<td>0.3125 µg/ml</td>
</tr>
<tr>
<td>Mouse IgG1</td>
<td>PE-Cy5</td>
<td>MOPC-21</td>
<td>BioLegend</td>
<td>0.3125 µg/ml</td>
</tr>
<tr>
<td>CD127</td>
<td>PE</td>
<td>eBioRDR5</td>
<td>eBioscience</td>
<td>0.6 µg/ml</td>
</tr>
<tr>
<td>IL-17</td>
<td>PE</td>
<td>eBio64DEC17</td>
<td>eBioscience</td>
<td>0.625 µg/ml</td>
</tr>
<tr>
<td>Mouse IgG1</td>
<td>PE</td>
<td>P3.6.2.1</td>
<td>eBioscience</td>
<td>0.6 µg/ml matched to α-CD127, 0.625 µg/ml matched to α-IL-17</td>
</tr>
<tr>
<td>Foxp3</td>
<td>AF488</td>
<td>259D</td>
<td>BioLegend</td>
<td>1 µg/ml</td>
</tr>
<tr>
<td>Mouse IgG1</td>
<td>AF488</td>
<td>MOPC-21</td>
<td>BioLegend</td>
<td>1 µg/ml</td>
</tr>
<tr>
<td>IL-10</td>
<td>PE</td>
<td>JES3-9D7</td>
<td>BD</td>
<td>0.6 µg/ml</td>
</tr>
<tr>
<td>Rat IgG1</td>
<td>PE</td>
<td>R3-34</td>
<td>BD</td>
<td>0.6 µg/ml</td>
</tr>
<tr>
<td>IFNγ</td>
<td>AF647</td>
<td>B27</td>
<td>BD</td>
<td>0.15 µg/ml</td>
</tr>
<tr>
<td>Mouse IgG1</td>
<td>AF647</td>
<td>MOPC-21</td>
<td>BD</td>
<td>0.15 µg/ml</td>
</tr>
</tbody>
</table>

Table 2.2: List of monoclonal antibodies used.

<table>
<thead>
<tr>
<th>FL1</th>
<th>FL2</th>
<th>FL3</th>
<th>FL4</th>
<th>FL5</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FL2</td>
<td>12</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FL3</td>
<td>4</td>
<td>49.2</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>FL4</td>
<td>0</td>
<td>7</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>FL5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2.3: Typical pre-acquisition compensation required.
Figure 2.2: Gating strategy to enumerate $CD_4^+ CD_{127}^{low} CD_{25}^{high} Foxp3^+$ events in whole blood.

Unstimulated whole blood samples stained for CD4, CD25, CD127 and Foxp3 or isotype controls were acquired by flow cytometry. a) An F-Foxp3 whole blood control viewed on a dot plot with forward scatter along the X axis and side scatter on the Y axis to demonstrate gating of a representative population of lymphocytes. b) Out of the lymphocytes gated, those which are CD4$^-$ were selected. c) The CD4$^-$ events were viewed on a 5% contour plot with AF488 mIgG1 on the X axis and anti-CD25 on the Y axis in order to set the CD25$^{high}$ gate. d) An F-CD25 whole blood control viewed on a dot plot with side scatter along the X axis and CD4 positivity on the Y axis to demonstrate gating of CD4$^+$ lymphocytes. e) The gates from c) are pasted on to this F-CD25 control to adjust the vertical gate to the last complete contour of the Foxp3$^-$ population. f) An F-CD127 whole blood control previously gated on CD4$^+$. This control is used to objectively set a threshold for CD127$^{high/low}$ on the last complete contour of the isotype negative population shown. g) The gate from f) is then pasted on to the test sample. h) The CD127$^{low}$ events are viewed as a pseudocolour plot with $\alpha$-Foxp3 on the X axis and $\alpha$-CD25 on the Y axis. The gates for CD25$^{high/low}$ and Foxp3$^+/-$ are pasted from part e). This enumerated that 2.97% of CD4$^+$ lymphocytes were CD127$^{low}$CD25$^{high}$Foxp3$^+$ in this representative donor.
Whole blood samples stimulated for 20 hours with PMA and ionomycin before staining for CD4 and IL-17A or an mIgG1 isotype control were acquired by flow cytometry. a) Both test (top row) and control (bottom row) viewed on a 2% contour plot with forward scatter along the X axis and side scatter on the Y axis to demonstrate gating of a representative population of lymphocytes. b) Out of the lymphocytes gated, those which were double negative were selected (the double positive population was highly autofluorescent in all channels even in an unstained control, hence it was gated out and assumed to be contaminating cell debris or granulocytes). c) The double negative events were viewed on a contour plot with α-IL-17A or mIgG1 on the X axis and α-CD4 on the Y axis in order to gate on CD4⁺IL-17A⁺. d) Test and control viewed as pseudocolour plots to illustrate CD4⁺IL-17A⁺ event enumeration. There were 1.05% IL-17A⁺ of CD4⁺ lymphocytes in this representative donor.
Figure 2.4: *Gating strategy to enumerate Th1, iTreg and IL-10*+ *Th1 cells in whole blood.*
Figure 2.4: Whole blood samples stimulated for 20 hours with PMA and ionomycin before staining for CD4, IFN$\gamma^+$ and IL-10 or isotype controls were acquired by flow cytometry. a) Test (right column) and FMO controls (left and middle columns for F-IL-10 and F-IFN$\gamma$ respectively) viewed on 2% contour plots with forward scatter along the X axis and side scatter on the Y axis to demonstrate gating of a representative population of lymphocytes. b) Out of the lymphocytes gated, those which were double negative were selected as in the previous figure. c) The double negative events were viewed on a contour plot with $\alpha$-CD4 on the Y axis in order to gate on CD4$^+$. d) The test CD4$^+$ events were viewed on a 2% contour plot with $\alpha$-IL-10 on the X axis and $\alpha$-IFN$\gamma$ on the Y axis in order to gate on CD4$^+\text{IFN}\gamma^+\text{IL-10}^-$, CD4$^+\text{IFN}\gamma^-\text{IL-10}^+$ and CD4$^+\text{IFN}\gamma^+\text{IL-10}^-$. e) Test and FMO controls viewed as pseudocolour plots to illustrate CD4$^+\text{IFN}\gamma^+\text{IL-10}^-$ (Th1), CD4$^+\text{IFN}\gamma^-\text{IL-10}^+$ (iTreg), and CD4$^+\text{IFN}\gamma^+\text{IL-10}^+$ (IL-10$^+$ Th1) event enumeration.

2.4 Validity measures

2.4.1 Extent of CD4 downregulation

The CD4$^+$ population contained cytokine positive cells after 20 hours stimulation and CD4 was downregulated as previously documented [80]. However, in preliminary assays, CD4 downregulation did not always occur, as shown in Figure 2.5. This was interpreted as an indicator of incomplete stimulation, and the stimulation technique was modified by creating stocks of stimulation medium, rather than adding PMA and ionomycin to each well. Following this, partial CD4 downregulation was not observed.

A threshold for CD4 downregulation was defined to enable rejection of assays should poor stimulation occur. Given that most completely stimulated samples had a percentage coefficient of variation (CV) of 50 or less in FL5 (PE-Cy7 anti-CD4) and all partially stimulated samples had a CV of 70 or more, the threshold was set to a CV of equal to 70 in FL5. Figure 2.6 shows that secretion of IFN$\gamma$ for 5 donors that partially downregulated CD4 was less than in 5 donors that completely downregulated CD4. The CV in FL5 was larger in the partially CD4-downregulating groups and the frequency of cytokine-secreting cells was less in the partially downregulating samples.
Figure 2.5: *Examples of complete versus incomplete CD4 downregulation.*

Whole blood samples stimulated for 20 hours with PMA and ionomycin before staining for CD4. 2% contour plots showing 3 samples with complete CD4 downregulation and 3 samples with partial downregulation.
Whole blood samples stimulated for 20 hours with PMA and ionomycin before staining for CD4 and IL-17A and taking supernatant samples for IL-17 ELISA. a) CV of CD4 expression shown for 3 partially downregulating and 3 completely downregulating donors, p=0.1. b) Frequency of IL-17A+ cells as percentage of CD4+ Lymphocytes shown for 3 partially downregulating and 3 completely downregulating donors, p=0.1. c) Secretion of IFNγ shown for 5 partially downregulating donors and 5 complete donors, p=0.0317. Median and interquartile range shown as columns with error bars. P values generated from the Mann Whitney U test.
2.4.2 Events acquired per small population of interest

In order to acquire enough sample for our observations to be statistically robust, it was ensured that the mean number of events for each small population of interest was above 100 for a CV of <10% as previously shown [186]. The mean number of events acquired are shown in Table 2.4. Adjustment for false positives indicated by the isotype controls was not included, even though the proportion of false positives in the isotype control was subtracted from the proportion of specific positives in the test sample for enumeration of each subset. Also, the mean percentage CV for each measurement calculated from 3 donors stained in triplicate is shown in Table 2.5.

<table>
<thead>
<tr>
<th>Population description</th>
<th>Mean events</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL-17+</td>
<td>1172</td>
<td>967</td>
</tr>
<tr>
<td>IFNγ+IL10−</td>
<td>10093</td>
<td>7438</td>
</tr>
<tr>
<td>IFNγ−IL-10+</td>
<td>243</td>
<td>180</td>
</tr>
<tr>
<td>IFNγ+IL10+</td>
<td>203</td>
<td>158</td>
</tr>
<tr>
<td>CD127lowCD25++Foxp3+</td>
<td>1408</td>
<td>782</td>
</tr>
</tbody>
</table>

Table 2.4: Events per small population of interest.

<table>
<thead>
<tr>
<th>Population Description</th>
<th>Percentage CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4+ of Unstimulated Blood</td>
<td>1.41</td>
</tr>
<tr>
<td>CD4+ of Stimulated Blood</td>
<td>1.87</td>
</tr>
<tr>
<td>IL-17+</td>
<td>10.42</td>
</tr>
<tr>
<td>IFNγ+IL10−</td>
<td>2.64</td>
</tr>
<tr>
<td>IFNγ−IL-10+</td>
<td>7.36</td>
</tr>
<tr>
<td>IFNγ+IL10+</td>
<td>9.18</td>
</tr>
<tr>
<td>CD127lowCD25++Foxp3+</td>
<td>8.34</td>
</tr>
</tbody>
</table>

Table 2.5: Mean percentage CVs from 3 donors stained in triplicate.
2.5 ELISA

ELISA was used to determine if the ability of T cells to secrete the cytokines IL-10, IL-17 and IFN\(\gamma\) changed throughout life. R&D Systems® Human DuoSet kits were used according to manufacturer’s instructions. Plates were read in a FLUOstar OPTIMA (BMG LABTECH) measuring absorbance at 450nm. On each plate a duplicate of standard cytokine concentrations was included. Using Microsoft Excel, the optical densities measured for the standard samples were plotted on a scatter plot and a linear or quadratic curve was fitted to the data. Using the equation of the curve, the concentrations of the unknown samples were interpolated.

2.6 Statistical analyses

Non-parametric statistical tests were used unless otherwise stated, such as the Spearman’s Rank test of correlation, the Mann Whitney test for comparing two groups, and the median and interquartile range (IQR) for summary statistics. Trendlines were fitted using linear regression simply to draw attention to the direction of association. These occasionally appeared curved when the data was shown on a log scale. GraphPad Prism was used unless otherwise stated (version 6.02 for Windows, GraphPad Software, San Diego California USA, www.graphpad.com). Statistically significant observations are indicated according to the level of significance: *\(p \leq 0.05\), **\(p \leq 0.01\), ***\(p \leq 0.001\), ****\(p \leq 0.0001\).

A sample size calculation was performed to ensure this study design would be of sufficient power to detect any Spearman’s Rank Correlation Coefficient \((r)\) of \(\geq 0.3\). This is a conventional medium effect size seen in a variety of correlation studies, which tends to be appreciable with the naked eye [62]. The convention of accepting a 0.05 probability of a Type I error \((\alpha)\) was also adhered to [61]. As the Spearman’s Rank Correlation Coefficient is known to be less powerful that its parametric equivalents, it was specified that the sample size should be large enough to give a higher power \((1-\beta)\) of 0.88, rather than the standard 0.8 [51]. A minimum sample size of 86 participants was estimated using the previously defined values of \(r\), \(\alpha\) and \(1-\beta\) and an open access sample size estimator for correlation [51].

Owing to sample volume, incomplete stimulation and early experimental optimisation, not all measurements are available for each of the 112 donors. Individual figures indicate how many donors contributed to the data presented.
2.7 Supervised learning to identify confounding factors in ageing study

A well characterised decision tree algorithm was selected to classify the ageing dataset and identify confounding factors, CART [14]. The implementation used was SimpleCAR T in the software package Waikato Environment for Knowledge Analysis (WEKA) 3.6.6 using leave-one-out cross-validation. SimpleCAR T was used in a recent publication to identify an indicator of metabolic syndrome [222].

A .csv file of flow cytometry and ELISA measurements was compiled. Measurements of % CD4$^+$ that were IL-17A$^+$, IFN$\gamma^+$IL-10$^+$ and CD127$^{-}\text{low}$CD25$^{\text{high}}$Foxp3$^+$ were included, whereas % CD4$^+$ that were IFN$\gamma^+$IL-10$^+$ measurements were excluded on the basis that they were too highly dependant on other measurements. ELISA measurements of IL-17, IL-10 and IFN$\gamma$ were also added alongside age where, initially, donors were divided into youngest and oldest as determined by the median age. Missing values were replaced with the median value.

2.8 Clustering analyses

2.8.1 Data preparation

Synthetic datasets

In order to figure out how each algorithm would perform on our nTreg data we dissected Zare and colleagues’ synthetic dataset and regenerated synthetic experiments separately varying the distance between two clusters, the density ratio between two clusters, and signal to noise ratio using a single 2-dimensional Gaussian against noise (Figure 2.7). We expected that $k$-means and SamSPECTRAL would cluster optimally when the distance between clusters was large, when the density ratio between clusters was small, and when the signal:noise ratio was large. However, we were interested to see how SamSPECTRAL behaved as the clusters moved closer together, the density ratio increased and the signal:noise ratio reduced, to see if SamSPECTRAL would be appropriate for use with our nTreg flow cytometry datasets.
Figure 2.7: Synthetic datasets generated to dissect algorithm performance.

a) Experiment varying the distance between two clusters. b) Experiment varying the density ratio between two clusters. c) Experiment varying signal:noise ratio. The code to generate these synthetic datasets can be found in the vignette in the appendix.
Real datasets

Compensated, gated CD4$^+$ lymphocytes were exported to an .fcs file using FlowJo for nTreg enumeration. There were 90 test data files with a panel of FMO control datasets for each. In R, the flowCore package of data structures was used to import data from these .fcs files, to extract the channels of interest (corresponding to CD127, CD25 and Foxp3) and to common log-transform the data [183, 75]. Data was randomised, fixing the seed to 1 unless otherwise stated.

For CD4$^+$ lymphocyte enumeration from total events, post-acquisition compensation was performed using FlowJo before exporting the whole dataset as a .fcs file. Channels were selected in R and only the fluorescence channel data were common log-transformed. Again, data was randomised, fixing the seed to 1 unless otherwise stated. Where clustering analysis is compared to manual gating, the results were generated using the quadrant gating strategies outlined previously.

2.8.2 $k$-means clustering

Benchmark $k$-means clustering was conducted in R on the aforementioned datasets, for a maximum of 100 iterations and with the cluster number as specified in the relevant sections of this manuscript.

2.8.3 SamSPECTRAL clustering

SamSPECTRAL (1.14) was run with $\text{sigma} = 200$ and $\text{separation.factor} = 2$ to be in the recommended range for flow cytometry data [294, 292] unless otherwise stated. As the parameters $m$, $\text{community.weakness.threshold}$, $\text{precision}$ and $\text{maximum.number.of.clusters}$ should not need to be changed, we used the recommended settings (3000, 1, 6 and 30, respectively) [294].

Examples of SamSPECTRAL on a simple synthetic dataset (fixing the number of clusters to 2) and for nTreg enumeration (cluster number not specified) are shown in Figures 2.8 and 2.9. In the synthetic example, two populations were created. One contained 30 datapoints from a 2-dimensional Gaussian distribution with $x$-mean=1, $y$-mean=4. The second contained 10 from a distribution with $x$-mean=2, $y$-mean=6. For both populations, $\sigma_{xx} = 0.08$, $\sigma_{yy} = 0.9$ and $\sigma_{xy} = \sigma_{yx} = 0$.
Figure 2.8: SamSPECTRAL method on synthetic data.
Figure 2.9: *SamSPECTRAL* method on nTreg data.
Figure 2.8: A simple synthetic dataset comprising 30 datapoints from a Gaussian distribution with $x$-mean=1 and $y$-mean=4 and 10 datapoints drawn from a distribution with $x$-mean=2 and $y$-mean=6 to illustrate the SamSPECTRAL algorithm components. a) Synthetic data scaled to between 0 and 1 in each dimension. A larger cluster can be seen in the bottom left and a smaller cluster in the top right. b) The $x$ parameter of each datapoint viewed against the datapoint’s index; the data was not randomised for clarity; events from the largest cluster have lower indices and events from the smaller cluster have larger indices. c) The $y$ parameter of each datapoint viewed against the datapoint’s index. d) The similarity graph of the synthetic dataset, organising the data to minimise the distance between datapoints of the same population, showing each event (vertex) grouped with other similar vertices by an edge. Two distinct groups are produced, with event 33 failing to be joined to either group. The functions `graph.adjacency` and `plot.igraph` from the `igraph` package were used to generate this part [64]. e) The corresponding eigenvalues curve. f) The eigenvalues curve with `line1` coloured in blue, `line2` coloured in red, `last.value` coloured in green, `number.of.clusters` coloured in orange. g) Clustering of the eigenvectors matrix with the indices corresponding to the largest population coloured in black and the smallest in red. h) Successful clustering with the largest population coloured in black and the smallest in red. Data point 33 originally located between the two clusters failed to be classified and therefore is not visible in this final part. The following parameter values were used: \( \sigma = 200 \) and \( \text{separation.factor} = 2 \).

Figure 2.9: Whole blood compensated and gated on CD4$^+$ lymphocytes and otherwise stained for CD25, CD127 and Foxp3 to illustrate the SamSPECTRAL algorithm components for nTreg enumeration. a) nTreg CD127 and CD25 expression data scaled to between 0 and 1 in each dimension. b) Community representatives selected through faithful sampling. c) The eigenvalues curve with `line1` coloured in blue, `line2` coloured in red, `last.value` coloured in green, `number.of.clusters` coloured in orange. d) Clustering of the eigenvectors matrix. e) nTreg CD127 and CD25 expression data coloured according to the initial clustering. f) nTreg CD127 and CD25 expression data coloured following cluster merging. In this case the merging step is necessary to achieve optimal clustering. The following parameter values were used: \( \sigma = 200 \) and \( \text{separation.factor} = 2 \).

2.8.4 Determining cluster validity

Result categorisation

Visual inspection of the SamSPECTRAL results informed a numerical categorisation of the results for each data file into 1 = “aggregated” (where SamSPECTRAL failed to separate nTreg from non-nTreg), 2 = “split” (where SamSPECTRAL broke the data into apparently random clusters), 3 = “split nTreg” (where the Treg population was split across more than one cluster), 4 = “split non-nTreg” (where the non-nTreg population was split across more than one cluster), 5 = “underestimate” (where the nTreg
cluster looked to be an underestimate of the nTreg population) and 6 = “optimal” (where optimal clustering was achieved). Examples are shown in Figure 2.10.

Figure 2.10: Example SamSPECTRAL solutions.

6 examples of compensated and gated CD4$^+$ lymphocytes stained for CD25, CD127 and Foxp3 analysed by SamSPECTRAL. a) An example “aggregated” solution. b) An example “split” solution. c) An example “split nTreg” solution. d) An example “split non-nTreg” solution. e) An example “underestimated” solution. f) An example “optimal” solution.

distcritmultiple

Owing to the large size of flow cytometry data files, it was not possible to calculate a distance matrix on an entire dataset, so the distcritmulti method within the fpc package was adapted to approximate the four cluster validity indices reviewed in Chapter 5 [121]. We refer to this adaptation as distcritmultiple. distcritmultiple separates the data into subsets, calculates the distance matrix and validity indices for each subset, then reports the average validity index of the subsetted data. distcritmulti previously only allowed the computation of average silhouette width and Pearson’s $\Gamma$ statistic, whereas cluster.stats, the method it inherits, is capable of generating all four measurements.
Following an initial comparison of cluster validity indices, ASW was reported as it does not depend on the clustering algorithm used to generate the partition [229]. The original \texttt{distcritmulti} method was therefore used to calculate the distance matrix on 30 subsets, unless otherwise stated [121]. We have reported the mean ASW across the subsets to quantify cluster validity in our experiments and to provide a benchmark from which to measure improvement.

It should be noted that when applying SamSPECTRAL at low specified values of $k$, the following error for the calculation of ASW was often observed:

\texttt{Error in sc$\text{clus.avg.widths}$ : $\$ operator is invalid for atomic vectors}

This resulted when the majority of datapoints belonged to one single cluster and some clusters were not represented in all subsets. When these errors manifested, we considered only the manual inspection of cluster output (always “aggregated” in these instances) and censored the ASW measurement.

2.8.5 Fitting minimum volume spanning ellipses

SamSPECTRAL does not currently permit the comparison of multiple datasets within the same experiment. Therefore, we fitted minimum volume spanning ellipses to our SamSPECTRAL-generated clusters to provide a partition we could apply to other datasets. For simplicity, we began by focusing on the 2 largest clusters and their Foxp3 and CD25 expression. From the \texttt{car} package in R [89], we used an adaptation of \texttt{dataEllipse()} to fit the minimum volume spanning ellipses containing at least 90\% of the data points in each cluster, and \texttt{ellipse()} to plot the two ellipses fitted on scatter plots of the test and control data. Then the number of data points falling within each ellipse was calculated. $(x_0, y_0)$ represented the centre of the ellipse, and $a$ and $b$ were the lengths of the major and minor semiaxes, respectively. These were derived from the covariance matrix of the fitted ellipse and assumed to be a ratio. A data point $(x, y)$ fell within the ellipse if it satisfied the following basic condition:
\[
\frac{(x - x_0)^2}{a^2} + \frac{(y - y_0)^2}{b^2} < r^2
\]

This was modified to account for ellipse rotation:

\[
\frac{(u)^2}{a^2} + \frac{(v)^2}{b^2} < r^2
\]

where

\[
u = \cos(\Theta)(x - x_0) + \sin(\Theta)(y - y_0)
\]

and

\[
v = -\sin(\Theta)(x - x_0) + \cos(\Theta)(y - y_0)
\]

\(\Theta\) was found by taking the eigenvectors of the covariance or shape matrix, and calculating arctan of the eigenvectors corresponding to the \((a, a)\) over \((b, a)\) positions in the matrix.
Chapter 3

Development and validation of the whole blood assay

3.1 Introduction

For this investigation a whole blood assay to measure the T cell subsets of interest needed to be optimised for use in our laboratory. Firstly, in human whole blood, the secretion kinetics of IL-17 and IL-10 in response to PMA and ionomycin needed to be established. Secondly, a multicolour flow cytometry assay needed to be designed for the current nTreg phenotype of interest, and intracellular staining for cytokines and transcription factors needed to be explored and validated. Then, the experimental variation in both flow cytometry and ELISA protocols also needed to be quantified to ensure it was within reasonable limits. This chapter outlines the developmental experiments conducted prior to running the resulting whole blood assay on donors of different ages, the results of which can be found in Chapter 4.

3.2 Time course ELISAs

The optimal incubation length for whole blood stimulation with PMA and ionomycin to detect IL-17, IL-10 and IFN-γ was deduced by conducting time course ELISAs. IFN-γ was the earliest cytokine detected in the supernatant; within 4 hours almost all the IFN-γ was secreted and it did not degrade before 20 hours (Figure 3.1). IL-10 and IL-17 in the supernatant appeared later. IL-17 was not detectable until 10-12 hours
and the highest level of IL-17 secretion was observed at 20 hours (Figure 3.2). IL-10 was not detectable until 12-16 hours and the highest level of IL-10 secretion was observed at 20 hours (Figure 3.3). As with IL-17, the level of IL-10 detected in the supernatants after 20 hours was much less across all three donors. An incubation length of 20 hours was selected for optimal staining of cytokine-secreting lymphocytes, as a longer incubation with brefeldin A would lead to pronounced alterations in cell function and integrity.

Figure 3.1: Time course of IFNγ secretion after stimulation.

Isolated CD4+ cells already differentiated from naïve were restimulated with PMA and ionomycin. Supernatants were collected every 4 hours and IFNγ measured by ELISA in singlicate for one donor as the secretion kinetics of IFNγ is well characterised [200]. Filled bars indicate that the original stimulation was performed using CD3 and CD55 with recombinant IL-1, IL-6 and IL-23 added to the culture. Empty bars indicate that the original stimulation was performed using CD3 and CD55 alone.
Figure 3.2: *Time course of IL-17A secretion by whole blood cultures after stimulation.*

Whole blood was stimulated with PMA and ionomycin over 48 hours prior to measuring IL-17 by ELISA. Single supernatants were collected every two hours from 3 donors. Filled bars indicate the abundance of IL-17A in the stimulated wells, whereas empty bars indicate the abundance in the unstimulated wells.
Figure 3.3: *Time course of IL-10 secretion by whole blood cultures after stimulation.*

Whole blood was stimulated with PMA and ionomycin over 48 hours prior to measuring IL-10 by ELISA. Single supernatants were collected every two hours from 3 donors. Again, filled bars indicate the abundance of IL-10 in the stimulated wells, whereas empty bars indicate the abundance in the unstimulated wells.
3.3 Staining panel development and optimisation

3.3.1 Fix/perm buffer selection

A recent paper documented some of the characteristic associations between Foxp3 clone, fluorochrome and fixative and permeabilising buffers used to enumerate Foxp3 positive events by flow cytometry [148]. Of the antibodies studied, the 259D/C7, PCH101, 236A/E7 and 206D antibodies yielded a mean of 6.9, 5.1, 4.7 and 3.7% of CD4⁺ events, respectively, although there was variation depending on the fix/perm buffers used [148]. Non-specific binding to Foxp3⁻ events was observed relative to isotype controls with the 259D/C7 and PCH101 clones [148]. 259D/C7 and 236A/E7 were both shown to give high Foxp3 staining when combined with the BioLegend buffers [148], thus we used 259D with the BioLegend Fix/Perm Buffer Kit.

3.3.2 Scatter characteristics of leukocyte lineages

Permeabilising reagents can profoundly alter the morphology of whole blood leukocytes and the correct setting of forward and side scatter voltages enables gating on these parameters. Therefore, a whole blood flow cytometry experiment staining the relevant lineages was conducted (Figure 3.4). The CD4⁺, CD8⁺, and CD19⁺ populations were all co-located in a small, non-granular population relative to the CD16⁺ and CD14⁺ populations. The lymphocytes were clearly separable from a much smaller “debris” population and the larger monocyte and granulocyte populations. The monocyte and granulocyte populations were not separable based on scatter characteristics alone. Following stimulation the populations became less distinctive (data not shown), although as a CD4 stain was always included in the staining panel, it was possible to set optimal forward and side scatter voltages for gating the lymphocyte population in each experiment.
Figure 3.4: Scatter characteristics of leukocyte lineages in permeabilised whole blood.

Whole blood was stained for the major leukocyte lineages prior to lysis, fixation, permeabilisation and acquisition of the sample by flow cytometry. a) Whole blood events were first gated to include all leukocytes and ignore debris. This gate was applied to all other samples in this figure. b) PE-Cy5 CD4$^+$ events in whole blood. c) PE-Cy5 CD4$^+$ events in whole blood compared to total leukocytes using backgating. d) PE-Cy5 CD8$^+$ events in whole blood. e) PE-Cy5 CD8$^+$ events in whole blood compared to total leukocytes. f) PE-Cy5 CD19$^+$ events in whole blood. g) PE-Cy5 CD19$^+$ events in whole blood compared to total leukocytes. h) FITC CD16$^+$ events in whole blood. i) FITC CD16$^+$ events in whole blood compared to total leukocytes. j) FITC CD14$^+$ events in whole blood. k) FITC CD14$^+$ events in whole blood compared to total leukocytes.
3.3.3 Antibody titrations

Titration experiments on each antibody were performed to ensure optimal sensitivity in distinguishing positive and negative populations with a minimal increase in background fluorescence. An example Foxp3 titration assay is shown in Figure 3.5. Consequently, 0.1µg of anti-Foxp3 antibody was used per 100µl whole blood as this was the smallest amount of antibody necessary for a Foxp3+ population to emerge. Using a higher concentration only raised the median light fluorescence (MFI) of the negative population, i.e. increased background fluorescence, without an increase in the number of Foxp3+ events.

![Figure 3.5: Optimisation of AF488-conjugated anti-Foxp3 on whole blood.](image)

Whole blood samples were stained with anti-CD4 and different concentrations of anti-Foxp3 antibody prior to acquisition by flow cytometry. This figure shows increasing concentrations of anti-Foxp3 antibody from top left to bottom right as noted below each part.

3.3.4 ROR\(\gamma(t)\) staining

The signature transcription factor of Th17 cells is ROR\(\gamma(t)\), or RORc2 which is the human orthologue [152, 39]. RORc2 is from the retinoic acid receptor-related orphan receptor C family [39]. The use of RORc as a
marker was believed to be particularly important to identify Th17 cells but also its expression relative to Foxp3 may have given valuable insight into how the balance between Th17 and nTregs behaves throughout life.

The RORγ(t) antibody available (PE-conjugated, AFKJS-9, eBioscience) at the time of designing this staining panel was tested for binding to the major lineage markers (Figures 3.6 and 3.7). Not only did it fail to bind the CD4+, CD8+, CD19+, CD16+ or CD14+ populations, but it bound a large proportion of events that were lineage negative. The nature of the RORγ(t)+ events in whole blood remains to be established. Consistently, the scatter characteristics of RORγ(t)+ events were indistinct, suggesting there may be some binding to dead cells. This would be consistent with the theory that RORγ has a function in regulating apoptosis, previously characterised in murine thymocytes [259]. In any case, in the previous study by Lee and colleagues, of the cells producing IL-17, the fluorescence intensity of RORc was not shown to be significantly increased in young donors relative to old [152].

![Figure 3.6: Anti-RORγ(t) and anti-CD4 staining.](image)

Whole blood samples were stained with anti-CD4 and different concentrations of anti-RORγ(t) antibody prior to acquisition by flow cytometry. The figure shows increasing concentration of anti-RORγ(t) from left to right. Quadrant gates set on concentration and brand matched isotype controls (PE-conjugated Rat IgG2a, eBR2a).

### 3.3.5 Helios staining

In the initial stages of designing this work, a novel biomarker of nTregs had been proposed by Thornton and colleagues, named Helios, a member of the Ikaros family of transcription factors [264]. Early evidence demonstrated 100% of CD4+Foxp3+ mouse thymocytes expressing Helios, and 70% of peripheral lymphoid tissue CD4+Foxp3+ cells expressing
Figure 3.7: \( \text{ROR} \gamma(t)^+ \) events in whole blood were not CD8\(^+\), CD19\(^+\), CD16\(^+\) or CD14\(^+\).

Whole blood samples were stained for major leukocyte lineages and with anti-ROR\( \gamma \)(t) antibody prior to acquisition by flow cytometry. The left-most column shows PE rlgG2a along the X axis and the lineage marker of interest on the Y axis. The central column shows the lineage marker on the Y axis and PE anti-ROR\( \gamma \)(t) on the X axis. The right-most column shows the events falling in the ROR\( \gamma \)(t)\(^+\) gate backgated on to the scatter plot for total events. a), b) and c) demonstrate that ROR\( \gamma \)(t)\(^+\) events are not CD8\(^+\). d), e) and f) demonstrate that ROR\( \gamma \)(t)\(^+\) events are not CD19\(^+\). g), h) and i) demonstrate that ROR\( \gamma \)(t)\(^+\) events are not CD16\(^+\). j), k) and l) demonstrate that ROR\( \gamma \)(t)\(^+\) events are not CD14\(^+\).
Helios in mice and humans [264]. Again, in both mouse and human, Helios was not expressed in \textit{in vitro} induced CD4$^+$Foxp3$^+$ T lymphocytes [264], suggesting it was a good biomarker for nTreg. It was also shown that Helios bound to the Foxp3 promoter, suggesting that Helios may regulate Foxp3 expression and nTreg function [101]. However, following the publication of this work, another group induced expression of Helios by \textit{in vitro} stimulation in the presence of antigen presenting cells [274]. Thornton and colleagues stimulated with anti-CD3 and anti-CD28 antibodies [265]. It was also recently observed in induced nTreg by Gottschalk and colleagues [105].

Helios was briefly investigated in this study as a signature transcription factor of nTreg (Figure 3.8) using a PE-conjugated anti-Helios antibody, clone 22F6 from Biolegend at 0.03$\mu$g. Helios expression was associated with high CD25 and Foxp3 expression, however, it was not possible to design an objective gating strategy that made sensible use of contour plots and did not marginalise FMO controls. Taken together with the published evidence, Helios staining was not included in the whole blood T cell assay.

Since the design of this study, Himmel and colleagues showed that Helios$^-$ cells can be found within a naïve CD45RA$^+$, CD31$^+$, CCR7$^+$, CD62L$^+$, Foxp3$^+$ T reg population [122]. A comment following this paper claimed that such a population was absent from cord blood, and therefore Helios is indeed a marker of nTreg [16]. Himmel and colleagues responded by challenging the phenotypic definition of “ naïve”, and summarise their own findings of a very small, but distinct population of Helios$^-$Foxp3$^+$ cells in cord blood [174]. Thus, whether Helios is a marker of nTreg still represents a controversial topic.

3.3.6 Staining verification

To verify the final staining panel, a Figure presented by Liu and colleagues was replicated [165]. Foxp3 expression was shown to be higher in CD127$^{low}$ and CD25$^{high}$ T helper cells (Figure 3.9).
Figure 3.8: Foxp3$^+$ Helios$^+$ events in whole blood were CD25$^+$.

Whole blood was stained for CD4, CD25, FoXP3 and Helios before acquisition by flow cytometry. Gates were set on FMO controls with concentration and brand-matched isotype controls (PE-conjugated Armenian Hamster IgG, HTK888). a) Whole blood viewed on a dot plot with forward scatter along the X axis and side scatter on the Y axis to demonstrate gating of lymphocytes, b) out of the lymphocytes gated, those which were CD4$^-$ were selected, c) the CD4$^-$ events were used to set the CD25$^{high}$ gate on a histogram, d) whole blood viewed on a dot plot with side scatter along the X axis and PE-Cy7 CD4 positivity on the Y axis to demonstrate gating of CD4$^+$ lymphocytes, e) the test sample shown as a 5% contour plot enumerating Foxp3$^+$ Helios$^+$ events, f) a pseudocolour dot plot of e), g) a histogram demonstrating the CD25 staining of the box gates in e) according to the legend.
Figure 3.9: Expression of Foxp3 in different CD25 and CD127 high and low subgroups.
Figure 3.9: Whole blood was stained for CD4, CD25, Foxp3 and CD127 and acquired by flow cytometry. a) CD4$^+$ lymphocytes gated as in previous figures and shown with CD127 expression on the X axis and CD25 expression on the Y axis in both a 2% contour and pseudocolour plot to illustrate quadrant and box gating of CD127$^{high/low}$. b) Foxp3 expression of CD127$^{low}$CD25$^{high}$ events. An arbitrary gate set on an isotype control for comparison between tubes gave 82.9% of events being Foxp3$^+$. c) Foxp3 expression of CD127$^{low}$CD25$^{high/lo}$ events. Here, only 22.3% of events were Foxp3$^+$. d) A comparison of b) and c). e) CD4$^+$ lymphocytes gated as in previous figures and shown with CD127 expression on the X axis and CD4 expression on the Y axis in both a 2% contour and pseudocolour plot to illustrate quadrant gating of CD4$^+$CD127$^{high/low}$. f) CD127$^{lo}$ events shown with CD25 expression on the X axis and CD4 expression on the Y axis. 44.7% of CD127$^{low}$ events were also CD25$^{high}$. g) CD127$^{high}$ events shown with CD25 expression on the X axis and CD4 expression on the Y axis. Only 15.3% of CD127$^{high}$ events were also CD25$^{high}$. h) The CD127$^{low}$CD25$^{low}$ events gated in f) are shown. 13.9% were positive for Foxp3 given the same gating as in b), c) and d). i) The CD127$^{low}$CD25$^{high}$ events gated in f) are shown. Here 75.9% were positive for Foxp3. j) The CD127$^{high}$CD25$^{low}$ events gated in g) are shown. 3.05% were Foxp3$^+$. k) The CD127$^{high}$CD25$^{high}$ events gated in g) are shown. 44.9% were Foxp3$^+$.

3.4 Non-parametric statistics for ageing study

Most flow cytometry and ELISA measurements were drawn from positively skewed distributions as shown in Figures 3.10 and 3.11. Therefore, non-parametric statistical tests were favoured.

3.5 Typical staining examples

The next few pages show representative stains for each CD4$^+$ T cell subset and demonstrate the level of consistency within experiments. In each of the Figures 3.12, 3.13, 3.14 and 3.15, examples of young, middle-aged and older donors are shown. The right-most donors in each plot were selected because of their unusually high proportions of each cell type.
Figure 3.10: Histograms to demonstrate flow cytometry data positive skew.
Figure 3.10: Frequencies of each T cell subset as measured by whole blood flow cytometry were collated and the resulting data were tested for normality and presented as frequency histograms. P values derived from Kolmogorov-Smirnov (KS) and Shapiro-Wilk (SW) tests are given for each measurement. a) KS: p<0.0001, SW: p<0.0001. b) KS: p=0.0006, SW: p<0.0001. c) KS: p>0.1, SW: p=0.0313. d) KS: p>0.1, SW: p=0.0132. e) KS: p>0.1, SW: p=0.154. f) KS: p=0.0036, SW: p<0.0001. g) KS: p>0.1, SW: p=0.0213. h) KS: p=0.0008, SW: p<0.0001.

Figure 3.11: Histograms to demonstrate ELISA data positive skew.

Abundance of each cytokine as measured by whole blood ELISA were collated and the resulting data were tested for normality and presented as frequency histograms. Again, P values were derived from KS and SW tests are given. a) KS: p<0.0001, SW: p<0.0001. b) KS: p<0.0001, SW: p<0.0001. c) KS: p<0.0001, SW: p<0.0001.
Figure 3.12: nTreg staining examples.

Four samples of unstimulated whole blood were stained for CD4, CD25, CD127 and Foxp3 for nTreg enumeration. The top row shows the initial selection of CD127$^{low}$ events. The bottom row shows CD25/Foxp3 quadrant gating. The lower horizontal gate was set on the uppermost 5% contour on the CD4$^-$CD25$^+$ events and the % of nTreg reported relates to this version of the gating strategy. The additional horizontal gate was set using a 2% contour plot and a comparison between these styles of gating is shown in Chapter 5. The % nTreg events within the CD4$^+$ lymphocyte population following subtraction of false positives was 2.39 for D383401, 1.05 for D416256, 0.552 for D481436 and 3.54 for D451431.
Figure 3.13: **IL-10 staining examples.**

Four samples of stimulated whole blood were stained for CD4, IFN-γ and IL-10 for iTreg enumeration. The top row shows IFN-γ/IgG1 staining from the F-IL-10 FMO control and the bottom row shows IFN-γ/IL-10 staining from the test panel. The % IL-10^+ IFN-γ^- events within the CD4^+ lymphocyte population following subtraction of false positives was 0.319 for D383401, 0.252 for D416256, 0.187 for D481436 and 0.950 for D292075.
Figure 3.14: IL-17 staining examples.

Four samples of stimulated whole blood were stained for CD4 and IL-17A for Th17 cell enumeration. The top row shows CD4/mIgG1 staining from the F-IL-17 FMO control and the bottom row shows CD4/IL-17 staining from the test panel. The % IL-17⁺ events within the CD4⁺ lymphocyte population following subtraction of false positives was 2.24 for D383401, 0.553 for D416256, 1.29 for D481436 and 5.87 for D174219.
Figure 3.15: IFNγ staining examples.

Four samples of stimulated whole blood were stained for CD4, IFNγ and IL-10 for Th1 cell enumeration. The top row shows mIgG1/IL-10 staining from the F-IFNγ FMO control and the bottom row shows IFNγ/IL-10 staining from the test panel. The % IL-10−IFNγ+ events within the CD4+ lymphocyte population following subtraction of false positives was 13.6 for D383401, 6.36 for D416256, 7.68 for D481436 and 35.7 for D174219. Interestingly, D174219 had both extremely high % IL-10−IFNγ+ events and % IL-17+ events.
3.6 ELISA error assessment and plate comparison

Figures 3.16, 3.18 and 3.20 show donor ELISA results derived from each plate (shown with respect to age in the next chapter), and indicate the standard deviation between triplicate measurements with error bars. Standard curves for all assays are shown in Figures 3.17, 3.19 and 3.21, colour coded according to the legends in the corresponding bar charts. ELISAs conducted most recently often measured higher values as they contained replicates of supernatants previously assayed at concentrations exceeding the standard.

The standard deviations are negligible for the duplicate standards. However, for the whole blood triplicates, individual measurements differ more (mean standard deviation in IL-17 triplicates was 313 pg/ml, IL-10 was 67.5 pg/ml and IFN\(\gamma\) was 1513 pg/ml). This can be attributed to the stochasticity inherent in the whole blood overnight stimulation otherwise equally high variation would be observed in the standards. The high standard deviation for IFN\(\gamma\) was to be expected as the cytokine is more abundant in whole blood and therefore the sample was diluted further (up to 1 in 100). Owing to high variation in the whole blood stimulation, experimental triplicates were performed for each donor, and the mean cytokine abundance was reported in Chapter 4.
Whole blood samples from each donor were stimulated with PMA and ionomycin for 20 hours prior to measuring IFN-γ in the supernatant by ELISA. The height of the bars represents the means with error bars representing standard deviations of triplicates, except for duplicates indicated in red. 107 donors are shown.
Figure 3.17: *Standard curves of known IFNγ concentrations for each ELISA.*

Points and error bars represent means and standard deviations of duplicates. Each standard curve is colour coordinated according to the corresponding experiment presented in the previous figure.
Figure 3.18: Whole blood IL-10 abundance for each donor.

Whole blood samples from each donor were stimulated with PMA and ionomycin for 20 hours prior to measuring IL-10 in the supernatant by ELISA. The height of the bars represents the means with error bars representing standard deviations of triplicates, except for duplicates indicated in red. 107 donors are shown.
Figure 3.19: *Standard curves of known IL-10 concentrations for each ELISA.*

Points and error bars represent means and standard deviations of duplicates. Each standard curve is colour coordinated according to the corresponding experiment presented in the previous figure.
Figure 3.20: Whole blood IL-17 abundance for each donor.

Whole blood samples from each donor were stimulated with PMA and ionomycin for 20 hours prior to measuring IL-17 in the supernatant by ELISA. The height of the bars represents the means with error bars representing standard deviations of triplicates, except for duplicates indicated in red. When samples exceeded the standard curve but there was insufficient supernatant to repeat the experiment, error bars are coloured blue. 107 donors are shown.
Figure 3.21: *Standard curves of known IL-17 concentrations for each ELISA.*

Again, points and error bars represent means and standard deviations of duplicates, except for a singlicate indicated in red. Each standard curve is colour coordinated according to the corresponding experiment presented in the previous figure.
Chapter 4

Balance between T cell subsets with age

4.1 Introduction

Using the flow cytometry assay outlined and validated in Chapters 2 and 3, we examined the frequency and function of CD4$^{+}$CD127$^{low}$CD25$^{high}$Foxp3$^{+}$ nTreg, CD4$^{+}$IFN$\gamma^{-}$IL-10$^{+}$ iTreg, CD4$^{+}$IL-17$^{+}$Th17 cells and CD4$^{+}$IFN$\gamma^{+}$ Th1 cells (both IL-10$^{+}$ and IL-10$^{-}$). The secretion of IL-10, IL-17 and IFN$\gamma$ by whole blood cell cultures was also measured to determine if this correlated with the intracellular phenotypes. We adopted a whole blood approach to allow analysis of all the populations of interest in a context resembling in vivo, with minimum sample touching and potential for cell loss [117].

Our data shows that the frequency of iTreg as a percentage of CD4$^{+}$ T lymphocytes increases in donors of different ages, without a compensatory increase in proinflammatory cells. Indeed, categorised analysis showed that the proportion of Th17 cells decreased in the older population. This data was substantiated by the secretion of greater amounts of IL-10 in aged individuals. Although there was no reduction in IL-17 secretion by whole blood cultures, the abundance of IL-10 increased with age relative to IL-17. Rather surprisingly, we found no relationship between age and frequency of nTreg or IL-10-secreting Th1 cells. Overall, this data supports the concept of an age-related tendency towards anti-inflammatory immune regulation in the periphery, rather than a proinflammatory environment.
4.2 Donor characteristics

112 healthy donors were recruited into this investigation of immune system ageing. Tables 4.1 and 4.2 show the distribution of age, sex, ethnicity and occupation in the study population alongside self reported relevant medical history. White British was coded “WB”, White Other was “WO”, Indian “I”, Asian “AO”, Chinese “C”, and Other “O”. Occupation was categorised according to SOC2010 [86]. “UE”, “L2”, “L3” and “L4” refer to unemployed, Level 2, 3 and 4 respectively. ‘LT” refers to long term medical conditions and “Hx” refers to history of cancer, chronic inflammation and smoking.

4.3 Increase in iTreg with age

The hypothesis that the balance between pro- and anti-inflammatory CD4$^+$ T cell subsets changes with age was tested. To this end we analysed CD4$^+$CD127$^{low}$CD25$^{high}$Foxp3$^+$ nTreg and CD4$^+$IFN$\gamma^{-}$IL-10$^+$ iTreg regulatory populations and inflammatory Th17 cells by staining for CD4$^+$IL-17$. As the phenotype of human nTreg is controversial owing to the lack of unique markers in man, we used a panel of biomarkers. Unstimulated blood samples were stained for CD4, CD127, CD25 and Foxp3 and these constituted 1.87% of total CD4$^+$ lymphocytes (1.35-2.62).

There was no correlation between the number of peripherally circulating nTreg and age (p=0.297, r=-0.111, Figure 4.1a).

In addition to nTreg, changes in the IL-10 producing iTreg subtype were investigated. The median percentage of iTreg was 0.284 (0.162-0.442) of CD4$^+$ T lymphocytes. Upon activation there was a positive correlation between iTreg and age (p=0.0035, r=0.303, Figure 4.1c). This finding was substantiated by the secretion of IL-10. As shown in Figure 4.1e, IL-10 secretion by whole blood cultures increased with donors of increasing age (p=0.0091, r=0.251). Whilst there was no correlation between iTreg and nTreg (p=0.940, r=0.00879, Figure 4.1g), there was a significant decrease in the ratio of nTreg to iTreg with advancing age, as the number of iTreg increased without a corresponding increase in nTreg (p=0.0123, r=−0.267, Figure 4.1h).
<table>
<thead>
<tr>
<th>Age Group</th>
<th>Number Recruited</th>
<th>Male: Female</th>
<th>Ethnicity</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>22</td>
<td>9:13</td>
<td>WB=8, WO=2, I=4, AO=5, C=2, O=1</td>
<td>L2=1, L3=3, L4=2, Student=16</td>
</tr>
<tr>
<td>30-39</td>
<td>17</td>
<td>5:12</td>
<td>WB=10, WO=2, I=2, AO=1, C=1, O=1</td>
<td>L2=2, L3=4, L4=7, Student=4</td>
</tr>
<tr>
<td>40-49</td>
<td>16</td>
<td>9:7</td>
<td>WB=13, WO=2, O=1</td>
<td>L2=3, L3=6, L4=5, Student=2</td>
</tr>
<tr>
<td>50-59</td>
<td>18</td>
<td>6:12</td>
<td>WB=17, I=1</td>
<td>L2=5, L3=9, L4=3, Student=1</td>
</tr>
<tr>
<td>60-69</td>
<td>22</td>
<td>7:15</td>
<td>WB=21, WO=1</td>
<td>L2=8, L3=4, L4=10</td>
</tr>
<tr>
<td>70-79</td>
<td>14</td>
<td>4:10</td>
<td>WB=14</td>
<td>L2=2, L3=4, L4=8</td>
</tr>
<tr>
<td>80-89</td>
<td>3</td>
<td>0:3</td>
<td>WB=3</td>
<td>UE=1, L2=1, L3=1</td>
</tr>
</tbody>
</table>

Table 4.1: Demographic details of sample population.
<table>
<thead>
<tr>
<th>Age Group</th>
<th>LT Medical Conditions</th>
<th>Current Medication</th>
<th>Hx Cancer</th>
<th>Hx Chronic Inflammation</th>
<th>Hx Smoking</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>None=19, Single=2, More than one=1</td>
<td>None=20, Single=2</td>
<td>None=22</td>
<td>None=21, None=21, Mild/Epis=1 Current=1</td>
<td></td>
</tr>
<tr>
<td>30-39</td>
<td>None=11, Single=4, More than one=2</td>
<td>None=8, Single=7, More than one=2</td>
<td>None=16, Astrocystoma=1</td>
<td>None=16, None=16, Mild/Epis=1 Current=1</td>
<td></td>
</tr>
<tr>
<td>40-49</td>
<td>None=14, Single=1, More than one=1</td>
<td>None=15, More than one=1</td>
<td>None=16</td>
<td>None=16, None=14, Current=2</td>
<td></td>
</tr>
<tr>
<td>50-59</td>
<td>None=12, Single=4, More than one=2</td>
<td>None=10, Single=3, More than one=5</td>
<td>None=18</td>
<td>None=18, None=18</td>
<td></td>
</tr>
<tr>
<td>60-69</td>
<td>None=8, Single=7, More than one=7</td>
<td>None=9, Single=4, More than one=9</td>
<td>None=20, None=20, Mild/Epis=2</td>
<td>None=22</td>
<td></td>
</tr>
<tr>
<td>70-79</td>
<td>None=4, Single=8, More than one=2</td>
<td>None=4, Single=5, More than one=5</td>
<td>None=14</td>
<td>None=13, None=11, Mild/Epis=1 Previous=2, Current=1</td>
<td></td>
</tr>
<tr>
<td>80-89</td>
<td>Single=1, More than one=2</td>
<td>None=1, Single=1, More than one=5</td>
<td>None=2, None=2, Breast=1, None=2, Mild/Epis=1 Previous=1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Medical details of sample population.
Unstimulated whole blood from donors of different ages was stained for CD4, CD25, CD127 and Foxp3 for nTreg enumeration. Then, whole blood stimulated with PMA and ionomycin for 20 hours was stained for CD4, IFNγ and IL-10 for iTreg enumeration and supernatants were collected for IL-10 ELISA. a) % CD4\(^+\) T cells that are nTreg in donors of different ages (n=90). b) Subgroup analysis of nTreg frequency between younger (n\(_y\)) and older donors (n\(_o\)): n\(_y\)=27 [median (IQR): 1.954 (1.57-2.48)], n\(_o\)=26 [1.73 (1.08-2.53)]. c) % CD4\(^+\) T cells that are iTreg in donors of different ages (n=91). d) Subgroup analysis of iTreg frequency: n\(_y\)=28 [median (IQR): 0.188 (0.088-0.339)], n\(_o\)=26 [0.344 (0.243-0.475)]. e) IL-10 secretion in donors of different ages (n=107). f) Subgroup analysis of IL-10 secretion: n\(_y\)=37 [median (IQR): 160 (2.90-526)], n\(_o\)=26 [305 (123-477)]. g) % CD4\(^+\) T cells that are iTreg plotted against % CD4\(^+\) T cells that are nTreg (n=87, r=0.00879). h) Ratio of % CD4\(^+\) that are nTreg to % CD4\(^+\) that are iTreg in donors of different ages (n=87).
4.4 Decrease in Th17 cells with age

To directly address the hypothesis that the balance between pro- and anti-inflammatory CD4\(^+\) T cells changes with age, we simultaneously measured the abundance and function of Th17 cells. Th17 cells comprised 1.32% of CD4\(^+\) T lymphocytes on average (0.68-2.42). In contrast to iTreg, whilst there was no correlation between Th17 cells and age (p=0.320, r=-0.0979, Figure 4.2a), a subgroup analysis comparing younger (\(\leq 40\) years) and older (\(\geq 65\)) donors showed a significant decrease in the number of Th17 cells with age (a decrease of 0.294%, p=0.0159, Figure 4.2b). However, there was no correlation between IL-17 secretion by whole blood cultures and age (p=0.949, r=-0.00631, Figure 4.2c). This was further supported by the subgroup analysis (p=0.456, Figure 4.2d).

4.5 Th17:iTreg balance decreases with age

To determine whether the balance between Th17 and T regulatory subsets in individuals changed with age, ratios between T cell subsets were calculated. Although there was no change in the ratio of Th17 to nTreg with age (p=0.741, r=-0.0360, Figure 4.3a), the ratio of Th17 to iTreg decreased in older donors (p=0.0144, r=-0.257, Figure 4.3c). This was attributed to the increase in iTreg without a corresponding increase in Th17 cells. This would suggest that suppression of the immune response, rather than exacerbated inflammation, occurs with age. Changes in secreted cytokines agreed with the flow cytometry data, as the ratio of IL-17 to IL-10 from whole blood cultures decreased with age, owing to the relative increase in IL-10 (p<0.0001, r=-0.461, Figure 4.3e).

When comparing different measurements from the same individual there was no correlation between the proportion of Th17 and nTreg (p=0.844, r=-0.0214, Figure 4.2e). However, the frequency of Th17 and iTreg cells positively correlated with one another (p<0.0001, r=0.600, Figure 4.2f). Strong positive correlation was also observed between the abundance of IL-17 and IL-10 secreted by whole blood cultures (p<0.0001, r=0.748, Figure 4.2g).
Figure 4.2: Th17 cells do not change with age.

Whole blood from donors of different ages stimulated with PMA and ionomycin for 20 hours was stained for CD4 and IL-17 for Th17 enumeration and supernatants were collected for IL-17 ELISA. These results were compared with the previous nTreg and iTreg measurements. a) % CD4$^+$ T cells that are Th17 cells in donors of different ages (n=105). b) Subgroup analysis of Th17 cell frequency between younger ($n_\text{y}$) and older donors ($n_\text{o}$): $n_\text{y} = 35$ [median (IQR): 1.11 (0.642-1.56)], $n_\text{o} = 26$ [1.4 (0.956-2.47)]. c) IL-17 measurements in donors of different ages (n=107). d) IL-17 secretion: $n_\text{y} = 37$ [median (IQR): 388 (72.9-174)], $n_\text{o} = 26$ [336 (89.0-539)]. e) % CD4$^+$ T cells that are Th17 cells plotted against % CD4$^+$ T cells that are nTreg (n=87). f) % CD4$^+$ T cells that are Th17 cells plotted against % CD4$^+$ T cells that are iTreg (n=90). g) IL-17 ELISA measurements plotted against IL-10 (n=107).
Figure 4.3: Th17:iTreg balance decreases with age.

Further comparison of whole blood measurements of nTreg, iTreg and Th17 cells. a) Ratio of % CD4+ that are Th17 cells to % CD4+ that are nTreg in donors of different ages ($n=87$). b) Subgroup analysis of the ratio of % CD4+ that are Th17 cells to % CD4+ that are nTreg between younger ($n_y=24$ [median (IQR)]: [0.931 (0.512-1.35) $n_o=26$ [0.792 (0.562-1.14)]]). c) Ratio of % CD4+ that are Th17 cells to % CD4+ that are iTreg in donors of different ages ($n=90$). d) Subgroup analysis of the ratio of % CD4+ that are Th17 cells to % CD4+ that are iTreg: $n_y=27$ [median (IQR): [2.02 (1.60-2.64)], $n_o=26$ [1.62 (1.22-1.84)]. e) Ratio of IL-17 secretion to IL-10 ($n=107$). f) Ratio of IL-17 to IL-10 secretion: $n_y=37$ [median (IQR): [6.70 (2.30-10.6)], $n_o=26$ [0.981 (0.274-2.54)].

116
4.6 The source of secreted IL-10 is independent of Th1 cells

Th1 cells can produce IL-10 [236], therefore we examined the IFN-γ⁺IL-10⁺ population and compared it to single IL-10⁺ iTreg and single IFNγ⁺ Th1 cells. As shown previously for IL-17 and IL-10, IFNγ and IL-10 measurements positively correlated across the flow cytometry and ELISA datasets (p<0.0001, r=0.577 and p<0.0001, r=0.551 respectively, Figure 4.4a and b). There was no change in the ratio of Th1 and iTreg (r=-0.0225, p=0.833, Figure 4.4c), nor in the ratio of IFNγ and IL-10 secretion in donors of different ages (p=0.208, r=-0.123, Figure 4.4d).

The median ratio of iTreg to IL-10⁺ Th1 cells was 1.09 (1.02-1.17), and this increased in the older donors (p=0.0005, r=0.358, Figure 4.4e). This was supported by the observation that the ratio of IFNγ⁺IL-10⁻ Th1 cells to IFNγ⁺IL-10⁺ Th1 cells did not change with age (p=0.580, r=0.0588, Figure 4.4f). The frequency of IFNγ⁺IL-10⁻ and IFNγ⁺IL-10⁺ Th1 cells in peripheral blood also did not change (data not shown).

4.7 Increase in Th1 relative to Th17 cells with age

It was previously shown that Th1 cells (CD4⁺IL-17⁻IFNγ⁺) increased relative to Th17 cells (CD4⁺IL-17⁺IFNγ⁻) with increasing age, owing to a decrease in Th17 cell frequency [152]. Therefore, we compared our measurements of CD4⁺ IL-17⁺ Th17 cell frequency with CD4⁺ IFNγ⁺ Th1 cell frequency. Unlike Th17 cells, which were shown to change with age upon subgroup analysis, Th1 cells did not change with age on correlation or subgroup analysis (p=0.670 and p=0.760, Figures 4.5a and b). However, the ratio between Th1 and Th17 cells did change with age, with a relative increase in Th1 cells (p=0.0254, r=0.234, Figure 4.5f). The ratio of IFNγ to IL-17 present in the whole blood supernatants also increased with age (p=0.0028, r=0.287, Figure 4.5g), as IFNγ was increased in the supernatants of older donors (p=0.0008, r=0.318, Figure 4.5c). IFNγ and IL-17 measurements also positively correlated across the flow cytometry and ELISA datasets (p<0.0001, r=0.683 and p<0.0001, r=0.490 respectively, Figure 4.5d and 4.5e).
Figure 4.4: The increase in IL-10 is independent of age-related induction of IL-10 expression by Th1 cells.

Whole blood from donors of different ages stimulated with PMA and ionomycin for 20 hours was stained for CD4, IFNγ and IL-10 for iTreg versus Th1 measurements and supernatants were collected for IL-10 and IFNγ ELISA. a) % CD4+ T cells that are Th1 plotted against % CD4+ T cells that are iTreg (n=91). b) IFNγ ELISA measurements plotted against IL-10 (n=107). c) Ratio of % CD4+ that are Th1 cells to % CD4+ that are iTreg in donors of different ages (n=91). d) Ratio of IFNγ secretion to IL-10 in donors of different ages (n=107). e) Ratio of % CD4+ that are iTreg to % CD4+ that are IL-10+ Th1 cells in donors of different ages (n=91). f) Ratio of % CD4+ that are Th1 cells to % CD4+ that are IL-10+ Th1 cells in donors of different ages (n=91).
Further comparison of whole blood measurements of Th1 and Th17 cells. a) % CD4+ T cells that are Th1 in donors of different ages (n=92). b) Subgroup analysis of Th1 frequency between younger (n_y) and older donors (n_o): n_y=28 [median (IQR): [13.1 (9.55-18.8)], n_o=26 [12.9 (9.00-18.9)]. c) IFNγ measurements in donors of different ages (n=107). d) % CD4+ T cells that are Th17 cells plotted against % CD4+ T cells that are Th1 (n=91). e) IL-17 secretion plotted against IFNγ secretion (n=107). f) Ratio of % CD4+ T cells that are Th1 cells to % CD4+ that are Th17 cells in donors of different ages (n=91). g) Ratio of IFNγ to IL-17 secretion in donors of different ages (n=107).
4.8 Absolute number measurements with age

Whilst we were primarily interested in whether the relative proportions of pro- and anti-inflammatory subsets within the CD4$^+$ lymphocyte compartment changed with age, we also calculated measurements of absolute number of each subset (Figures 4.6 and 4.7). The absolute frequencies showed the same trends as the percentage data, but in the case of subgroup analysis of Th17 frequency, did not give the same statistically significant difference.

Absolute measurements of peripheral blood CD4$^+$ lymphocytes were taken from 90 healthy donors (Figure 2.1). There was no correlation between the frequency of CD4$^+$ lymphocytes and age ($p=0.455$, $r=0.0798$).
Figure 4.6: Absolute numbers of anti-inflammatory CD4$^+$ lymphocytes subsets (/µl whole blood) in donors of different ages.

Whole blood samples stained for anti-inflammatory T cell subset markers with a known volume of FlowCount$^{TM}$ Fluorospheres added were acquired by flow cytometry. a) Absolute numbers of nTreg (/µl) are shown against donor age (n=89). b) Subgroup analysis of absolute nTreg number between younger (n$_y$) and older donors (n$_o$): n$_y$=26 [median (IQR): 10.3 (5.94-14.7)], n$_o$=26 [8.79 (5.75-15.6)]. c) Absolute numbers of iTreg (/µl) are shown against donor age (n=88). d) Subgroup analysis of absolute iTreg number between younger (n$_y$) and older donors (n$_o$): n$_y$=27 [median (IQR): 0.616 (0.238-0.772)], n$_o$=25 [0.906 (0.591-1.27)]. e) Absolute numbers of IL-10$^+$ Th1 cells (/µl) are shown against donor age (n=88). f) Subgroup analysis of absolute IL-10$^+$ Th1 number between younger (n$_y$) and older donors (n$_o$): n$_y$=27 [median (IQR): 0.397 (0.115-0.731)], n$_o$=25 [0.503 (0.305-0.777)].
Figure 4.7: Absolute numbers of proinflammatory CD4\(^+\) lymphocytes subsets (/\(\mu l\) whole blood) in donors of different ages.

Whole blood samples stained for proinflammatory T cell subset markers with a known volume of FlowCount\textsuperscript{TM} Fluorospheres added were acquired by flow cytometry. a) Absolute numbers of Th17 cells (/\(\mu l\)) are shown against donor age (n=89). b) Subgroup analysis of absolute Th17 cell number between younger (n\(_y\)) and older donors (n\(_o\)):

- n\(_y\)=26 [median (IQR)]: [4.75 (2.82-7.16)], n\(_o\)=25 [3.37 (2.22-6.01)].

- n\(_y\)=27 [median (IQR)]: [28.1 (17.4-37.7)], n\(_o\)=25 [39.2 (22.8-61.3)].
4.9 Correlation between flow cytometry and ELISA measurements

As the ELISA measurements complemented the flow cytometry data presented, the relationship between function and phenotype in each individual donor was compared. In Figure 4.8, T cell subsets as both a proportion of total CD4$^+$ lymphocytes and an absolute measurement per µl whole blood are shown in relation to cytokine abundance in supernatant. Strong positive correlation is shown in each instance.

4.10 Age-stratified correlation analysis

Lee and colleagues previously showed that the correlation between % CD4$^+$ lymphocytes that were IFN$\gamma$+ and those that were IL-17$^+$ was stronger in younger donors ($\leq$40) than in older donors ($\geq$65) [152]. Given the strong correlation between cytokine measurements we observed across both ELISA and flow cytometry experiments, we investigated if this was the case within our study population, and included an additional comparison with the donors falling between the ages of 40 and 65.

As shown for the ELISA data in Figure 4.9a, there was no systematic change in the correlation between IFN$\gamma$ and IL-17 secretion across the three age groups. However, for the correlation between IFN$\gamma$ and IL-10, and IL-17 and IL-10 shown in Figure 4.9b and c respectively, the correlation coefficient decreases with age, the p value increases with age, but the number of donors also decreases, so there is a systematic decrease in power that may confound these observations. No trends were observed in the strength of correlation with age in the flow cytometry datasets, as shown in Figures 4.10 and 4.11.
Figure 4.8: Correlation between flow cytometry phenotype and ELISA functional cytokine measurements.

Whole blood flow cytometry measurements of each T cell subset were compared with corresponding functional cytokine ELISA measurements. a) Total % CD4+ cells that were IFNγ+ plotted against IFNγ secretion (n=95). b) Absolute number of total CD4+IFNγ+ cells per µl whole blood plotted against IFNγ secretion (n=89). c) Total % CD4+ cells that were IL-10+ plotted against IL-10 secretion (n=94). d) Absolute number of total CD4+IL-10+ cells per µl whole blood plotted against IL-10 secretion (n=88). e) Total % CD4+ cells that were IL-17A+ plotted against IL-17 secretion (n=105). f) Absolute number of total CD4+IL-17A+ cells per µl whole blood plotted against IL-17 secretion (n=89).
Figure 4.9: *Age stratified ELISA data correlation analysis.*

The correlation between whole blood cytokine measurements was compared between different age groups. a) Correlation between IL-17 and IFN-γ abundance in young (n=46), middle (n=35) and older age groups (n=26). b) Correlation between IL-10 and IFN-γ abundance in young (n=46), middle (n=35) and older age groups (n=26). c) Correlation between IL-10 and IL-17 abundance in young (n=46), middle (n=35) and older age groups (n=26).
The correlation between whole blood flow cytometry measurements was compared between different age groups. 
a) Correlation between IL-17+ and IFNγ+ T cell frequencies in young (n=34), middle (n=31) and older age groups (n=26). 
b) Correlation between IL-10+ and IFNγ+ T cell frequencies in young (n=35), middle (n=30) and older age groups (n=26). 
c) Correlation between IFNγ+ and IFNγ+IL-10+ T cell frequencies in young (n=35), middle (n=30) and older age groups (n=26).
Figure 4.11: Age stratified flow cytometry data correlation analysis (part 2).

The correlation between whole blood flow cytometry measurements was further compared between different age groups. 

a) Correlation between IL-17+ and IL-10+ T cell frequencies in young (n=34), middle (n=30) and older age groups (n=26). 

b) Correlation between IL-17+ and IFNγ+IL-10+ T cell frequencies in young (n=34), middle (n=30) and older age groups (n=26). 

c) Correlation between IL-17+ and IL-10+ T cell frequencies in young (n=35), middle (n=30) and older age groups (n=26).
4.11 Using a supervised learning technique to identify confounding factors

A mid-term analysis was conducted after the first 73 donors were recruited to identify potential confounding factors affecting our ageing study using SimpleCART. The decision tree generated using the entire ageing dataset is shown in Figure 4.12. The algorithm placed heavy importance on IL-10 secretion and \% CD127^{low}CD25^{high}Foxp3^{+} within CD4^{+} cells to predict whether donors fell into the older or younger half of the sample population.

In order to improve performance and to remove weaker data, SimpleCART was next applied to a restricted dataset with 8 younger and 8 older donors removed due to high proportions of missing values. Again, the algorithm placed heavy importance on IL-10 secretion and \% CD127^{low}CD25^{high}Foxp3^{+} within CD4^{+} cells to predict age, but the tree contained less decisions, as shown in Figure 4.13, and was more accurately able to predict young from old.
SimpleCART implemented for WEKA was used to predict donor age group from a .csv file comprising the entire dataset of CD4$^+$ T cell subset measurements. Correctly classified instances were 57.5342% whereas incorrectly classified instances were 42.4658%. SimpleCART has derived from the dataset that low IL-10 secretion identifies young from old, then given high IL-10 secretion, low % CD127$^{low}$CD25$^{high}$Foxp3$^+$ cells within total CD4$^+$ cells identifies old from young. Given both of the above, IL-10 secretion <137.116 pg/ml and >248.679 identifies young from old.
SimpleCART was used again to predict donor age group from the reduced dataset of CD4\(^+\) T cell subset measurements. Correctly classified instances were then 66.6667\%, with incorrectly classified instances at 33.3333\%. SimpleCART has derived from the dataset that low % CD127\(^{low}\)CD25\(^{high}\)Foxp3\(^+\) cells within total CD4\(^+\) cells identifies old from young, lending higher weight to the association it learned before. Given high % CD127\(^{low}\)CD25\(^{high}\)Foxp3\(^+\), IL-10 secretion <188.071 pg/ml subsequently identifies young from old.
The SimpleCART work shown previously gave particular importance to IL-10 ELISA data as an attribute on which age could be predicted. Owing to some of the measurements being particularly low and nearer the low end of ELISA sensitivity, this prompted some concern that a time-dependency might be present within the IL-10 ELISA data. Early on, one plate in particular gave very low measurements, and also included several younger donors. We have conducted an assessment for confounding factors presented in Figures 4.14, 4.15 and 4.16.

Our first comparison was to look at our frequency and functional measurements against the chronological assay number instead of age, as shown in Figure 4.14. % CD4+ T cells that were Th17 cells negatively correlated with assay number, whereas % CD4+ T cells that were nTreg and IFNγ secretion positively correlated with assay number.

We also assessed whether our frequency and functional measurements were associated with time of day the blood sample was taken, as shown in Figure 4.15. % CD4+ T cells that were IL-10+ and IFNγ+IL10+ positively correlated with the time the blood sample was taken as well as IL-10 and IFNγ secretion.

In light of finding correlation with chronological assay number and the time samples were taken, we plotted age of donor against both of these variables, as shown in Figure 4.16. Correlation was observed between chronological assay number and age of donor, which may account for its appearance as a confounding factor. Time of blood taking was also positively correlated with age of donor. In a previous study of the IL-10 secretion in samples from young and old donors, three measurements were made throughout the year showing that IL-10 decreased from October through to June, approximately the same months that recruitment occurred for this study [239]. This supports the notion that IL-10 increase with age is a real phenomenon and not a consequence of seasonal variation.

Since the WEKA mid-term experiment, repeated analysis and further volunteer recruitment was done. However, the correlation analyses contain all the most recent data.
Figure 4.14: Assessment of correlation of measurements with assay number.

Figure 4.14: Correlation of whole blood flow cytometry and ELISA measurements of CD4+ T cell subsets with chronological assay number. a) % CD4+ T cells that are IL-17A+ plotted against chronological assay number (n=105). b) % CD4+ T cells that are IFNγ+ against assay number (n=92). c) % CD4+ T cells that are IL-10+ against assay number (n=91). d) % CD4+ T cells that are IFNγ+IL-10+ against assay number (n=91). e) % CD4+ T cells that are CD127lowCD25highFoxp3+ against assay number (n=90). f) Secretion of IFNγ into whole blood culture supernatant plotted against assay number (n=107). g) Secretion of IL-17 against assay number (n=107). h) Secretion of IL-10 against assay number (n=107).
Figure 4.15: Assessment of correlation of measurements with time blood sample extracted.

Figure 4.15: Correlation of whole blood flow cytometry and ELISA measurements of CD4+ T cell subsets with the time the blood sample was extracted. a) % CD4+ T cells that are IL-17A+ plotted against time (n=105). b) % CD4+ T cells that are IFNγ+ against time (n=92). c) % CD4+ T cells that are IL-10+ against time (n=91). d) % CD4+ T cells that are IFNγ+IL-10+ against time (n=91). e) % CD4+ T cells that are CD127lowCD25highFoxp3+ against time (n=90). f) Secretion of IFNγ into whole blood culture supernatant plotted against time (n=107). g) Secretion of IL-17 against time (n=107). h) Secretion of IL-10 against time (n=107).
4.12 Discussion

We examined alterations in Treg:Th17 balance by measuring the abundance and function of these cell types with regards to healthy ageing. This represents an important advance, as previous work was restricted to only the nTreg or the Th17 population. This was therefore limited in interpretation because of the homeostatic relationship between these cells [109, 228, 127, 272, 78]. We examined a spectrum of ages (20-82 years) and observed no change in the number of nTreg throughout the lifetime. However, unlike other studies, we also enumerated Th17 cells and iTreg and looked at the ratio between nTreg and these T cell subsets. We found that the number of nTreg was preserved relative to Th17 cells, but decreased relative to iTreg with advancing age. This was owing to an increase in the number of iTreg without a corresponding change in nTreg. Such increases are not thought to be solely resulting from a change in the ratio between naïve and memory populations [132]. This indicates that the ageing immune system is characterised by increased regulation of responses by an induced regulatory T cell population.

As the ratio between IFNγ and IL-10 had previously been shown to decrease with age by Bruunsgaard and colleagues [37], we were surprised by our finding of no correlation. In the previous study, PHA was used as a stimulant over a 24 hour period; it is possible that if we had conducted

![Figure 4.16: Assessment of correlation of between age of donor with assay number and time blood sample extracted.](image)

a) Chronological assay number plotted against donor age (n=112). b) Time blood sample was extracted against donor age (n=112).
another 4 hours of stimulation that IL-10 secretion may have continued to increase, and eventually we may have achieved the same conclusion.

Cakman and colleagues employed a similar whole blood assay to show that IL-10 secretion increased with age [217, 40]. We built on this by combining cytokine measurements in whole blood cultures alongside intracellular IL-10 staining of CD4+ T lymphocytes, in addition to other T cell subsets. Although dendritic cells, B cells and macrophages may have also contributed to IL-10 release [236], we showed that an increased frequency of iTreg reflected the increased IL-10 secretion.

It was observed that individuals that secreted the greatest levels of IL-10 also secreted larger amounts of IL-17 and IFNγ. The frequencies of Th17 and iTreg strongly correlated and there was also a correlation between the secretion of IL-17 and IL-10. In addition, IFNγ and IL-10 measurements positively correlated across both the flow cytometry and ELISA datasets. Positive correlation between cytokine secreting cells was previously shown in the tumour microenvironment [145]. Taken together, this suggests that donors were high or low cytokine secretors, and our data suggested this phenomenon was not age-dependent. This would not have been apparent from a study looking at single cytokines. The reason for this is unclear, although there is evidence of age-related loss of CpG methylation in other cytokine gene promoter regions such as tumour necrosis factor, leading to increased gene expression [106]. Epigenetic control of gene expression has been shown to play a role in regulation of both IFNγ and IL-17 expression [70, 60]. However, currently there is no evidence that this is the case for IL-10 [70].

As we took multiple measurements from each donor blood sample, we considered applying multiplicity corrections to our statistics. There is a debate as to whether methods such as Bonferroni corrections are necessary and in which circumstances they should be used. We did not apply a multiplicity correction as, for our hypotheses of interest, they are likely to simply exchange increased type I error (α, likelihood of reporting a false positive result) for increased type II error (likelihood of reporting a false negative) [207, 2]. Were we testing a global hypothesis, such as “is there an association between HLA types and susceptibility to influenza” by measuring multiple HLA types for example, then we would consider it more appropriate to use multiplicity corrections [207]. Multiplicity corrections also assume that each of the tests are independent from one other [2]; our hypothesis is to compare variables that we think are related from the same individual and therefore it may not be appropriate to use
We acknowledge that p<0.05 remains a statistical likelihood even if no real association is present. However, we consider the reliability of each p value as we see it [207]. A p value of <0.0001 is highly significant with or without multiplicity correction, whereas a p value of 0.04 is unlikely to be robust to multiplicity correction. We are very confident, for example, in the observation that the frequency of IL-17^+ T helper cells is strongly correlated with the frequency of IL-10^+, where the Spearman’s Rank test gives p<0.0001. However, the observation that nTreg:iTreg ratio decreases with age is only supported by p<0.0122. Although this meets our current criterion of statistical significance, it may not be robust if we were to apply a multiplicity correction.

We considered that an increase in the frequency of iTreg with age may reflect an effect of CMV or Epstein-Barr virus (EBV) infection, as some viruses secrete large quantities of IL-10 [68]. Viral IL-10 is indistinguishable from the human cytokine using the JES3-9D7 anti-IL-10 monoclonal antibody used for staining. However, this is unlikely as the human IL-10 ELISA kit used (DY217B) does not substantially react with CMV or EBV viral IL-10 derivatives (0% reactivity with recombinant CMV-derived IL-10, and 0.6% cross-reactivity with recombinant EBV-derived IL-10 [260]).

To date, few groups have studied human Th17 cells in relation to immunosenescence. The study by Lee and colleagues, who conducted the same subgroup analysis of Th17 measurements, observed a decrease in Th17 frequency in the memory population in elderly donors [152]. We employed comparable inclusion and exclusion criteria for donor recruitment. However, we investigated Th17 ageing in greater depth by looking at a continuous spectrum of ages and found that the association between Th17 cells and age was not retained in a correlation analysis. Furthermore, the ratio of Th17 cells to iTreg decreased in older donors, as iTreg increased without a corresponding increase in Th17. This suggests that suppression of the immune response occurs in older age groups, rather than exacerbated inflammation as was previously thought [34].

In conclusion, we examined the balance between proinflammatory Th17 cells and anti-inflammatory nTreg and iTreg by both cell enumeration and functional cytokine production within donors from 20-82 years of age. The frequency of iTreg as a percentage of CD4^+ T lymphocytes and the abundance of IL-10 secreted by whole blood cell cultures increased with age. There was no change in nTreg or the IL-10^+ Th1 subset. Increases in
IL-10 were not compensated for by increases in IL-17 and there was even evidence that the frequency of IL-17$^+$ CD4 cells decreased in the older age group. We continue with a more general evaluation of these laboratory findings and discuss how the whole blood assay might be extended and generally improved in Chapter 6.
Chapter 5

Investigation of SamSPECTRAL for automated nTreg enumeration

5.1 Introduction

During the ageing study the gating of nTreg events from raw flow cytometry data was difficult and time consuming owing to the overlapping nature of the nTreg and non-nTreg populations with respect to CD25, CD127 and Foxp3 expression. This chapter aimed to convince the reader, firstly, of a need to automate nTreg analysis by highlighting the sources of error in the interpretation and execution of a stringent, rule-based manual gating strategy. Then, the SamSPECTRAL algorithm was compared to a benchmark clustering algorithm and manual gating. The difficulties of implementing and validating an automated analysis pipeline are also discussed.

5.2 Inadequacy of manual 2-dimensional gating for nTreg flow cytometry data analysis

5.2.1 Traditional gating strategies for nTreg interpreted differently lead to different results

The previously described staining panel and gating strategy for enumerating nTreg was designed to reflect current literature [165, 72, 258, 226, 18, 172, 124]. However, using sequential 2-dimensional
gating in FlowJo, the fine details of implementing this strategy were left open to interpretation. Figure 5.1 shows results from 4 different versions of the nTreg gating strategy. Parts a) and b) show measurements of nTreg without considering CD127 expression. These correspond to the black gates in Figure 5.2. Parts c) and d) do specifically measure CD127low cells. These correspond to the purple gates in Figure 5.2. Parts a) and c) show measurements of nTreg where CD25high expression was set on the CD4− population viewed on a 5% contour plot, whereas in b) and d) the gates were set on a more specific 2% contour plot. This demonstrates that, although statistical significance was only achieved with one measurement, differences could be observed in the variance of the data and the slope and direction of the linear regression line depending on how the details of the gating strategy were defined. We reported part c), as it included the CD127low expression which more specifically identified nTreg, with a 5% contour plot used to define CD25high expression, to more sensitively capture nTreg. However, this process revealed the potential for bias to be introduced as an experimenter could easily select the method generating a desired solution.

5.2.2 Poor reproducibility of manual gating even with a stringent rule-based strategy

Another cause for concern was the lack of reproducibility of manual gating strategies. It had previously been reported that the coefficient of variation in manual gating strategies can be as high as 40% [23]. We performed a comparison of coefficient of variation between a CD4+ lymphocyte enumeration from total whole blood events and an enumeration of CD127lowCD25highFoxp3+nTreg from CD4+ lymphocytes to capture the inter- and intra-experimenter error between simple and complex manual analyses (Figure 5.3). We conclude that even with a heavily rule-based gating strategy, using contour plots, a high degree of variation can be observed when analysing nTreg.

Three sets of test and FMO control data files were selected for triplicate analysis by one expert and independent validation by a further two experts. The repeated analyses by the single expert were conducted one month apart to prevent a remembering of gating placement. Although the first expert briefed the second two on the gating strategy used, the others did not see the gating provided by the first, and the first was not present.
when the analysis was being performed by the others. Thus, an independent validation of the specific gating strategy was performed.

Inspection of individual expert gatings does not provide an obvious source of error. However, CD4 lymphocyte enumeration is understandably more robust as the population of interest is better separated from the rest of the events. Also, the CD4 lymphocyte gate is the first in a gating hierarchy. nTreg, however, are enumerated following a sequence of four gating decisions, all of which introduce gating error. The error is also large relative to the small size of the nTreg population; donors 1 and 2 in Figure 5.3 have a lower frequency of nTreg and therefore a larger standard deviation than donor 3.

Figure 5.1: Different measurements of nTreg give different conclusions about whether their frequency changes with age.

Whole blood from donors of different ages was stained for CD4, CD25, CD127 and Foxp3. a) % CD4+ T cells that are CD25^{high(5%)}Foxp3+ in donors of different ages (n=90). b) % CD4+ T cells that are CD25^{higher(2%)}Foxp3+ in donors of different ages (n=90). c) % CD4+ T cells that are CD127^{low}CD25^{high(5%)}Foxp3+ in donors of different ages (n=90). d) % CD4+ T cells that are CD127^{low}CD25^{higher(2%)}Foxp3+ in donors of different ages (n=90).
Whole blood from donors of different ages was stained for CD4, CD25, CD127 and Foxp3.  

a) An F-Foxp3 whole blood control viewed on a dot plot with forward scatter along the X axis and side scatter on the Y axis to demonstrate gating of a representative population of lymphocytes.  
b) Out of the lymphocytes gated, those which are CD4$^-$ were selected.  
c) The CD4$^-$ events were viewed on a 5% contour plot with mlgG1 on the X axis and anti-CD25 on the Y axis in order to set the CD25$^+$ gate. The second set of quadrant gates were set on a 2% contour plot, not shown.  
d) An F-CD25 whole blood control viewed on a dot plot with side scatter along the X axis and CD4 positivity on the Y axis to demonstrate gating of CD4$^+$ lymphocytes.  
e) The gates from c) are pasted onto this F-CD25 control to adjust the vertical gate to the last complete contour of the Foxp3$^-$ population.  
f) The previous gates shown pasted on to the sample stained with the test panel of antibodies.  
g) An F-CD127 whole blood control previously gated on CD4$^+$. This control is used to objectively set a threshold for CD127$^{high/low}$ on the last complete contour of the isotype negative population shown.  
h) The gates from g) are then pasted on to the test sample.  
i) The CD127$^{low}$ events are viewed as a pseudocolour plot with Foxp3 on the X axis and CD25 on the Y axis. The gates for CD25$^{high/low}$ and Foxp3$^{+/-}$ are pasted from part f).
Figure 5.3: Intra- and inter-experimenter variation in manual gating analysis.

Whole blood from donors of different ages was stained for CD4, CD25, CD127 and Foxp3. a) The percentage of total events gated as CD4\(^+\) lymphocytes measured three times by the same investigator (average %CV = 0.42). b) The percentage of CD4\(^+\) lymphocytes gated as CD127\(^{low}\)CD25\(^{high}\)Foxp3\(^+\) Treg measured three times by the same investigator (average %CV = 29.2). c) The percentage of total events gated as CD4\(^+\) lymphocytes measured by three different investigators (average %CV = 1.57). d) The percentage of CD4\(^+\) lymphocytes gated as CD127\(^{low}\)CD25\(^{high}\)Foxp3\(^+\) Treg measured by three different investigators (average %CV = 37.76). The height of each bar represents the mean and the error bars the standard deviation. The p values reported related to a Kruskal-Wallis test for significance between multiple groups with the Dunn comparison of each pair of groups.
Stringent gating rules are not routinely demonstrated in published flow cytometry analyses of nTreg and in such studies we anticipate that the inter- and intra-experiment error is greater still. Aghaeepour and colleagues recently commented on the differing opinions of experts as to how flow cytometry data should be manually gated, and where discrepancies were too large, removed data files from their analyses [7]. Therefore, a method of analysis that is both more reproducible and less prone to observer bias is needed.

Qualitatively, a quadrant gating strategy could be also considered inappropriate for nTreg gating as the population thought to be nTreg-enriched is cut relative to CD25 expression (Figure 5.2). Arguably, any gating strategy to enumerate nTreg balances robustness and reproducibility with accuracy, and sensitivity with specificity, to a certain extent. Therefore, we have investigated automated analysis of nTreg datasets.

Of several novel algorithms for use with flow cytometry data reviewed in Chapter 1, the modified spectral clustering algorithm SamSPECTRAL was compared to elementary $k$-means. Spectral clustering is a well-defined and well-used non-parametric clustering method and is robust to the presence of outliers, noise and variation in shape of clusters [293]. Owing to current limitations in computational power, it is not possible to use conventional spectral clustering for large data sets, therefore SamSPECTRAL was developed for specific application to flow cytometry [293]. Its implementation in R and compatibility with the flowCore package permits adaptation for a particular problem or dataset and coupling to additional packages for data visualisation and calculation of cluster validity indices.

This results chapter proceeds initially with a comparison of various cluster validity indices and a demonstration that SamSPECTRAL’s inability to create a clustering solution with a prespecified number of clusters complicates their interpretation. A combination of ASW and manual inspection was subsequently used, therefore at present the analysis pipeline of nTreg datasets using either algorithm cannot be fully automated.

Then the following hypotheses were tested:

1. Does SamSPECTRAL perform better than $k$-means on synthetic datasets?

2. Does SamSPECTRAL perform better than $k$-means on nTreg datasets?
3. Is SamSPECTRAL clustering superior to manual gating of nTreg data?

Parameter optimisation was conducted to optimise the use of SamSPECTRAL on synthetic and nTreg datasets individually. Lastly, an automatic elliptical gating step was implemented to allow for comparison between test and control datasets, and the resulting analysis pipeline was used to assess whether nTreg change with age.

5.3 Cluster validity indices

Correlation between validity indices

Firstly, the correlation between four cluster validity indices was assessed to determine if one could be used as a substitute for them all. As shown in Figure 5.4, ASW had strong correlation with Pearson’s $\Gamma$ and entropy, but not with the Calinski and Harabasz Index. Therefore ASW measurements were subsequently presented where appropriate.

Can ASW discern acceptable solutions?

Then, the ability of ASW to determine optimal clustering was assessed. SamSPECTRAL was run once on each of 90 datasets, the resulting solutions were categorised by manual inspection and ASW was calculated. As shown in Figure 5.6, ASW was sensitive to changes between optimal and split non-nTreg, however did not differ significantly for underestimate, split nTreg, split or aggregated solutions. The standard deviations of the ASW also showed an ability to differentiate between groups.

However, SamSPECTRAL is programmed to amalgamate clusters it believes to be drawn from the same population. Figure 5.5 below shows that, unlike $k$-means, SamSPECTRAL did not adhere to the specified value of $k$. This is a beneficial feature for data exploration. However, it complicates the interpretation of experiments using ASW to optimise parameter settings and compare methods. For consistency, the results of subsequent experiments specifying cluster number are shown against the specified value of $k$.

Halkidi and colleagues suggest that if an algorithm does not take a specific number of clusters as an argument, the range of values for which the number of resultant clusters remains constant should be considered
Compensated, gated CD4$^+$ lymphocyte events otherwise stained for CD25, CD127 and Foxp3 were clustered by SamSPECTRAL. a) Correlation between ASW and Pearson’s $\Gamma$ statistic. b) Correlation between ASW and entropy. c) Correlation between ASW and the Calinski and Harabasz index. A preliminary single run on each nTreg data file with normal.sigma = 200 and separation.factor = 2.

[113]. This both indicates the number of clusters present naturally in the data and the optimal parameter settings as the middle value in this range [113]. This strategy was considered in the subsequent parameter optimisation experiments presented.
Figure 5.5: Specified versus resultant cluster number following SamSPECTRAL clustering of nTreg data.

Compensated, gated CD4$^+$ lymphocyte events otherwise stained for CD25, CD127 and Foxp3 from 5 donors were clustered by SamSPECTRAL with different specified values of $k$ (cluster number).

Figure 5.6: ASW on different SamSPECTRAL-clustered nTreg solution types.

Compensated, gated CD4$^+$ lymphocyte events otherwise stained for CD25, CD127 and Foxp3 from all donors were clustered by SamSPECTRAL prior to calculating ASW. a) The mean ASW measured across all 30 subsets. b) The standard deviation (sd) of ASW measured across all 30 subsets.
Evaluation of manual result categorisation

As ASW was incapable of differentiating optimal solutions from all categories, manual inspection remained a requirement. The agreement in solution categorisation between three independent investigators was evaluated across all 90 nTreg flow cytometry datasets. SamSPECTRAL was used to enumerate nTreg and the resultant biaxial plots were examined in triplicate by three experts. Again, the first expert briefed the second two on the categorisation rationale used and the experts were blinded to each others’ categorisation. All three experts were also asked to specify the percentage attributed to the nTreg cluster. Figure 5.7 demonstrates that independent experts had qualitatively similar approaches to categorisation of SamSPECTRAL nTreg clustering attempts. 73.9% of data files were categorised the same way by all three experts. In 92.4% of cases, all three experts return the same percentage of CD4⁺ lymphocytes that were nTreg.

Figure 5.7: Categorisation of SamSPECTRAL nTreg clustering solutions by three independent experts.

Compensated, gated CD4⁺ lymphocyte events otherwise stained for CD25, CD127 and Foxp3 from all donors were clustered by SamSPECTRAL. The single solutions were categorised following manual inspection of the scatter plots generated.
5.4 Comparison and optimisation of \( k \)-means and SamSPECTRAL performance on synthetic data

To test the hypothesis that SamSPECTRAL was better than \( k \)-means, their ability to cluster synthetic data, which deliberately mimicked nTreg/non-nTreg data characteristics, was assessed. In the original SamSPECTRAL work, the authors demonstrated that the algorithm works effectively on a synthetic dataset consisting of five populations randomly drawn from two dimensional Gaussian distributions against a background of random noise [295, 291]. Qui and colleagues also showed a synthetic experiment of 2-dimensional data to illustrate how SPADE works for flow cytometry-like datasets [214]. Later, they added noise to their data, illustrating that SPADE was robust to a small amount. In this study, the dataset from Zare and colleagues was dissected and separate experiments were regenerated varying the distance between two clusters, the density ratio between two clusters, and signal to noise ratio using a single 2-dimensional Gaussian against noise. The aims were also to test if \( k \)-means and SamSPECTRAL tolerated small distances between clusters, large inter-cluster density ratios and signal:noise ratios, and to understand their behaviour. For the distance variation experiment, the mean \( x \)-value (\( \mu \)) is reported for each cluster. For density, the ratio of the largest to smallest cluster relating to the density of each Gaussian is given, and the the signal:noise ratio is reported for the experiment varying noise.

On \( k \)-means analysis of the synthetic datasets varying distance as shown in Figure 5.8a, there was a peak in ASW at \( k = 2 \), the correct number of clusters, highest when the distance between clusters was largest. ASW then decreased rapidly as \( k \) increases. Plotting this data with distance along the \( x \)-axis (Figure 5.8c) demonstrates there was little discrimination in ASW as the distance decreased, therefore \( k \)-means is not sensitive to changes in distance between clusters. Example clustering given in Figure 5.11a shows that \( k \)-means only correctly identified the two clusters at \( \mu = 1.9 \).

For all distances, allowing SamSPECTRAL to decide on cluster number gave poor validity clustering with respect to ASW (Figure 5.8b, Specified \# Clusters = “NA”). Far too many clusters are generated as SamSPECTRAL defaults to generating 15 clusters if the kneepoint is found to be less than 15. Again, the highest ASW was at \( k = 2 \). Figure 5.8d demonstrates that for \( k = 2 \), ASW decreased more slowly as the
distance decreased than for \( k \)-means. A larger separation was observed between curves corresponding to \( k = 2 \) and 3 than for the other numbers of clusters. Example clustering given in Figure 5.12a clearly shows that SamSPECTRAL correctly distinguished the two clusters when they were well separated and more closely approximated the two clusters when they overlapped.

In the experiment varying density ratio between clusters, the highest ASW for \( k \)-means was at \( k = 2 \) (Figure 5.9a), the correct number of clusters, and when the density ratio was smaller (Figure 5.9c). As shown in Figure 5.9b, for all densities the ASW for SamSPECTRAL clustering was highest at \( k = 2 \). Interestingly, at \( k = 4 \), ASW was low for most density ratios. At \( k = 2 \) and 3, the ASW for SamSPECTRAL clustering was higher for more density ratios than \( k \)-means (Figure 5.9d). An example of this experiment with clusters according to \( k = 2 \) is shown in Figures 5.11b and 5.12b for \( k \)-means and SamSPECTRAL respectively. In each part, \( k \)-means inappropriately cuts one cluster or the other, whereas SamSPECTRAL performs qualitatively better clustering with respect to the underlying data.

To summarise the results of the distance and density experiments, SamSPECTRAL produced qualitatively superior solutions than \( k \)-means. However, it was expected that measurements of ASW would support manual observation. Whilst some key features were identified, such as \( k = 2 \) being the optimal cluster number, in many cases \( k \)-means achieved a very similar ASW to SamSPECTRAL.

There is no perfect clustering algorithm as it is impossible to satisfy all conflicting requirements [129]. Arguably, the same is true for cluster validity indices [210]. Still, silhouette measurements are highest for compact clusters with good separation between them [229]. This is a limitation for our nTreg flow cytometry datasets which demonstrate neither of these characteristics. As we can never know the true structure of our data and therefore be certain of cluster validity, we are left with the “best under these circumstances” [210]. In the case of flow cytometry data analysis, a good deal of work still needs to be done to achieve the “best possible” indication of cluster validity given the computational state of the art.

When the ratio between signal and noise was varied, \( k \)-means gave a constant measurement of ASW for different cluster numbers (Figure 5.10a). This suggests that, again, \( k \)-means partitions according to probability and not the natural data structure, generating a “football”-like
pattern as shown in Figure 5.11c. Generally, SamSPECTRAL also did not
cluster the signal:noise experiment well (Figure 5.10b). However, negative
values of ASW for the “Uniform Noise” part of the experiment suggest
that the combination of ASW measurement and SamSPECTRAL
clustering indicates when a partition has been generated randomly with
respect to the underlying data. The “No Noise” part also gave low ASW
for SamSPECTRAL, illustrating that the splitting of a single population
into 2 or more clusters is reflected in low cluster validity. Also, Figure
5.10d shows that ASW peaks for SamSPECTRAL clustering when the
noise is low for several numbers of clusters, contrasting with \( k \)-means.
There was a downward trend in ASW as cluster number increased,
reflecting an inability of SamSPECTRAL to distinguish signal from noise;
the higher the value of \( k \), the more the signal cluster was separated into
one of the peripheral clusters. At the highest values of \( k \), the signal cluster
would also be partitioned into roughly equal clusters. An example of this
experiment with clusters according to \( k = 14 \) is shown in Figure 5.12c.
The SamSPECTRAL clustering solution was qualitatively similar to the
\( k \)-means solution apart from where cluster merging could be observed
(Figure 5.12c, 35000 Events:5000 Noise). Interestingly, spatial separation
of events within one cluster was observed on rare occasions with
SamSPECTRAL as shown in Figure 5.13.
Figure 5.8: *k*-means versus SamSPECTRAL on distance data.

a) ASW on *k*-means analysis at varying numbers of clusters. b) ASW on SamSPECTRAL analysis at varying numbers of clusters. c) ASW on *k*-means analysis varying the distance between clusters. d) ASW on SamSPECTRAL analysis varying the distance between clusters.
Figure 5.9: \textit{k}-means versus SamSPECTRAL on density data. 

a) ASW on \textit{k}-means analysis at varying numbers of clusters. b) ASW on SamSPECTRAL analysis at varying numbers of clusters. c) ASW on \textit{k}-means analysis varying the density between clusters. d) ASW on SamSPECTRAL analysis varying the density between clusters.
Figure 5.10: *k*-means versus SamSPECTRAL on signal:noise data.

a) ASW on *k*-means analysis at varying numbers of clusters.  b) ASW on SamSPECTRAL analysis at varying numbers of clusters.  c) ASW on *k*-means analysis varying signal:noise.  d) ASW on SamSPECTRAL analysis varying signal:noise.
Figure 5.11: Example k-means clustering of synthetic datasets.

a-c) Experiments varying distance between clusters, density ratio between clusters and signal:noise ratio, respectively.
Figure 5.12: *Example SamSPECTRAL clustering of synthetic datasets.*

a-c) Experiments varying distance between clusters, density ratio between clusters and signal:noise ratio, respectively.
Figure 5.13: Example of spatial separation of events within one cluster.

SamSPECTRAL clustering of signal:noise synthetic dataset with cluster number set to 6 and seed set to 7.

As synthetic datasets were used, we were able to further compare $k$-means and SamSPECTRAL to the perfect cluster assignment using the F-measure. After generating the synthetic datasets, each datapoint was labelled according to the population it was drawn from. After running $k$-means and SamSPECTRAL, the resulting cluster assignments were compared to the original class labels.

Firstly, for the distance and density experiments, generally values of the F-measure were higher relative to ASW (Figure 5.14). There was also generally a larger difference in cluster validity, with SamSPECTRAL achieving higher F-values than $k$-means. Also, there with smoother transitions between high and low F-values than ASW as the distance between clusters decreased and the density ratio increased with a less prominent nTreg-like cluster. In the distance experiment, the F-measure for SamSPECTRAL-generated partitions remained high until it began to decrease at $mu = 1, 1.75$, whereas a rapid decrease in F-values occurred for $k$-means at $mu = 1, 5$. In the distance experiment, F-measures for SamSPECTRAL were close to 1 at all density ratios, whereas F-values for $k$-means decreased from 20000:20000.
In the noise experiment, the F-measures were generally lower than the ASW. However, at low signal:noise ratios, peaks in the F-measures could be seen for both $k$-means and SamSPECTRAL, which were not reflected in measurements of ASW. Also, SamSPECTRAL achieved considerably higher F-values for low signal:noise ratios than $k$-means.

Experiments using the F-measure, therefore, provided consistent and stronger statistical evidence that SamSPECTRAL was superior to $k$-means in clustering synthetic data mimicking typical characteristics of flow cytometry data. It also generated measurements for all of the experiments shown, even where aggregated solutions occur. However, it was not possible to use the F-measure on the real datasets owing to lack of knowledge of the perfect cluster assignment.
Figure 5.14: Comparison of ASW and F-measure in synthetic experiments.

ASW and F-measure were used to quantify cluster validity from partitions generated by k-means versus SamSPECTRAL on synthetic datasets. $k = 2$ for the distance and density experiments, $k = 14$ for the signal:noise experiments. a) ASW on experiments varying distance between clusters. b) F-measure on experiments varying distance between clusters. c) ASW on experiments varying density ratio between clusters. d) F-measure on experiments varying density ratio between clusters. e) ASW on experiments varying signal:noise ratio between clusters. f) F-measure on experiments varying signal:noise ratio between clusters. Measurements are means of the solutions generated across 10 replicated runs iterating the random number seed according to run number.
5.4.1 SamSPECTRAL parameter optimisation for synthetic datasets

As SamSPECTRAL gave qualitatively better clustering than $k$-means and that it clustered favourably with respect to distance and density variation, it was explored further through parameter optimisation experiments. In this context, the parameters referred to are those used to calibrate SamSPECTRAL rather than those within the datasets themselves. As the majority of parameters should not need to be altered for flow cytometry data [294], only the `normal.sigma` and `separation.factor` were investigated. `normal.sigma` is a scaling parameter that determines how sensitive similarity is to changes in squared Euclidean distance [295].

When `normal.sigma` is too small, populations of interest are inappropriately split into multiple subclusters. When `normal.sigma` is too large, populations of interest can be inappropriately merged. The other parameter of interest is `separation.factor`, the threshold ratio of between cluster similarity (similarity between data points of separate clusters) to within cluster similarity (similarity between data points of the same cluster). If the similarity between two clusters is large and the `separation.factor` parameter is too low inappropriate cluster aggregation will occur as the clusters are considered to be part of the same population. When `separation.factor` is too high, patchy and spurious population splitting can be observed [295], as SamSPECTRAL relies on cluster merging of a larger number of component clusters to achieve irregular cluster shapes. Therefore, an investigation to find the best parameter values for each dataset was conducted.

nTreg-like configurations of the synthetic datasets previously described were used. As the distance and density experiments were very similar, we used a single dataset combining features of both. The largest cluster was formed of 37500 events compared to a smaller cluster of 2500. The clusters were overlapping with $x$-means of 1 and 1.75. For the signal:noise dataset, the central cluster was formed of 35000 events with uniform background noise of 5000.

SamSPECTRAL was run with varying `normal.sigma` and `separation.factor` values. The number of clusters was specified to achieve a more sensitive measurement of ASW. For the density experiment the number of clusters was set to 2. For the signal:noise experiment the number of clusters was set to 14. 10 runs were conducted for each parameter setting, iterating the random number seed each time. The
mean of each measurement across all replicates is presented.

The ASW measurements from each parameter variation experiment are shown in Figure 5.15, alongside presentations of mean category number and mean resultant number of clusters for each parameter configuration. For both density and signal:noise datasets, ASW plateaued when separation.factor was between 2 and 8. This also corresponded with the largest mean category number, providing a useful indication of the range of appropriate values. In the normal.sigma experiment, a plateau was also observed in mean category number between the values of 100 and above for the density dataset, but not for signal:noise. A stable plateau was not observed in the measurements of mean cluster number for both normal.sigma and separation.factor. Therefore, we cannot apply the logic of Halkidi and colleagues of selecting parameters from the range at which the cluster number remains constant [113]. Example results are shown in Figures 5.16 and 5.17.

In Figure 5.16, although the nTreg-like cluster is often contaminated with datapoints from the non-nTreg cluster, the best clustering was achieved at higher values of sigma and normal.sigma and separation.factor. If separation.factor was 1 or less, aggregated solutions occurred. For the signal:noise experiment (Figure 5.17), considering the plotted data, the signal cluster was routinely arbitrarily split into multiple subclusters. Therefore, these parameters were re-optimised on nTreg data.

Taken together, these data indicate that separation.factor must be 2 or more for optimal performance on the density and signal:noise datasets. Values less than 2 lead to inappropriate cluster merging. For normal.sigma, the optimal values are less obvious owing to the lack fluctuation in the values of ASW. However, for the density dataset, clustering measured by mean category number was best at normal.sigma values of 100 or above, as there was less contamination of the smaller cluster with data points from the larger cluster.
Figure 5.15: *SamSPECTRAL parameter optimisation experiments on synthetic data fixing* $k$.

Synthetic datasets were analysed by SamSPECTRAL with varying parameter settings. 
a) ASW when varying `normal.sigma`. b) ASW when varying `separation.factor`. c) Standard deviation (sd) of ASW when varying `normal.sigma`. d) sd of ASW when varying `separation.factor`. e) Mean solution category when varying `normal.sigma`. f) Mean solution category when varying `separation.factor`. g) Mean number of clusters when varying `normal.sigma`. h) Mean number of clusters when varying `separation.factor`. Measurements are reported as means across 10 runs, where the random number seed was iterated.
Figure 5.16: Example SamSPECTRAL parameter optimisation on density synthetic datasets.

a) `normal.sigma` optimisation on density datasets. b) Corresponding eigenvalue curves. c) `separation.factor` optimisation on density dataset.
Figure 5.17: Example SamSPECTRAL parameter optimisation on signal:noise synthetic datasets.

a) `normal.sigma` optimisation on signal:noise datasets. b) Corresponding eigenvalue curves. c) `separation.factor` optimisation on signal:noise dataset.
As the F-measure had provided strong evidence of SamSPECTRAL clustering being superior to \( k \)-means clustering, consistent with manual inspection, it was also used to predict optimal parameter values on synthetic datasets.

Importantly, F-measures decreased for the noise experiment as they increased for density experiment (Figure 5.18). As visual inspection of the clustering solutions showed more favourable clustering for the density experiments than the noise experiments, the density experimental data was considered more helpful. This was supported by the higher F-measures observed in the density experiments. For values of \( \text{sigma} \) of 50 or greater, the F-values tended towards one. Similarly, for \text{separation.factor} \) values 2 or greater, the F-values tended towards one. Although parameter optimisation was conducted separately on nTreg data, this suggested that appropriate ranges for acceptable parameters were 50-1000 for \( \text{sigma} \) and 2-8 for \text{separation.factor}. 

164
Figure 5.18: *F-measures for parameter optimisation in synthetic datasets.*

The F-measure was used to quantify cluster validity from partitions generated using SamSPECTRAL with different parameter values on synthetic datasets to find ranges of optimal parameter settings. \( k = 2 \) for the density experiments, \( k = 14 \) for the signal:noise experiments. a) F-measures generated for SamSPECTRAL with different values of \( \sigma \). b) F-measures generated for SamSPECTRAL with different values of `separation.factor`. Measurements are means of the solutions generated across 10 replicated runs iterating the random number seed according to run number.
5.5 Comparison and optimisation of $k$-means and SamSPECTRAL performance on nTreg data

To further test the hypothesis that SamSPECTRAL is superior to $k$-means, they were then compared on the nTreg datasets comprising compensated, pre-gated CD4$^+$ lymphocytes. The experiments of fixing the cluster number and systematically changing parameter values were repeated. Firstly, Figure 5.19a shows that ASW is higher for lower numbers of clusters using $k$-means. Figure 5.19b then shows that $k$-means does not meaningfully cluster the nTreg from the non-nTreg population at any value of $k$; it simply generates roughly equal partitions. For example, ASW is highest at $k = 2$, yet manual inspection of the data clearly shows that $k$-means failed to separate the two populations of interest.

SamSPECTRAL was then applied to the nTreg datasets 10 times and the mean measurements of the acceptable solutions were reported. As shown in Figure 5.20a, proportions of nTreg could only be drawn in runs where $k > 9$. Although no optimal solutions were reached for D406318, reasonable consistency was observed in the $k$-specified and non-specified runs (marked NA), even though the cluster number specified was often not the resultant cluster number. Figure 5.20b demonstrates that when an appropriate number of clusters is selected, SamSPECTRAL generates qualitatively better solutions than $k$-means.

As with $k$-means, Figure 5.21a shows that the ASW of SamSPECTRAL clustering generally either decreases or stays constant with increasing cluster number, depending on the donor. The mean ASW reported was of the acceptable runs, excluding the aggregated and split category solutions. This again contradicts visual inspection which only indicates optimal clusterings at high values of $k$ as shown for donor D383401 in Figure 5.20. This demonstrates that ASW is an inappropriate indicator of optimal clustering and therefore optimal cluster number in this experiment. There was an instance with donor D469375 where setting $k$ to 14 was the only scenario where SamSPECTRAL clustered optimally, whereas “split non-nTreg” solutions were generated for $9 < k < 14$ and where SamSPECTRAL selected its own optimum number of clusters. However, given that the resultant number of clusters was 10, and when 10 clusters was specified SamSPECTRAL did not give an optimal solution, arguably this did not result from specifying $k$. 

Compensated, gated CD4$^+$ lymphocytes otherwise stained for CD25, CD127 and Foxp3 were analysed by $k$-means for different specified numbers of clusters. a) Specified cluster number against average silhouette width showing that $k$-means gives better clustering at lower values of $k$. The median of 10 runs is reported, with random number seed corresponding to run number. b) D383401: a representative example of $k$-means clustering for all specified values of $k$). Although the nTreg cluster is often distinct, the “football” pattern of non-nTreg event clustering suggests that the nTreg cluster is generated by probability.
Figure 5.20: Example SamSPECTRAL clustering of nTreg datasets at different specified values of $k$ (1).

Compensated, gated CD4$^+$ lymphocytes otherwise stained for CD25, CD127 and Foxp3 were analysed by SamSPECTRAL for different specified numbers of clusters. a) Mean frequency of nTreg measured from optimal SamSPECTRAL-generated solutions for different specified values of $k$. b) D383401: a representative example of visually-judged optimal solutions occurring at the highest values of $k$.
Figure 5.21: Example SamSPECTRAL clustering of nTreg datasets at different specified values of \( k \) (2).

Again, datasets comprising compensated, gated CD4\(^+\) lymphocytes otherwise stained for CD25, CD127 and Foxp3 were analysed by SamSPECTRAL for different specified numbers of clusters. a) Mean ASW from acceptable SamSPECTRAL-generated solutions for different specified values of \( k \). b) D469375: a representative example of visually-judged “split”/“optimal” solutions occurring at the highest values of \( k \).
5 randomly selected donors have often been used where it was not feasible to analyse all available datasets. This test dataset contained 3 donors which most often were clustered optimally, one that was often underestimated, and one where the non-nTreg cluster was often split into multiple clusters. It could be argued that a more targeted approach to selecting test files would have been more appropriate and that the method used may have added some experimental artefacts.

5.5.1 SamSPECTRAL parameter optimisation for nTreg datasets

Parameter optimisation experiments were initially attempted on our synthetic datasets for objectivity. However, we were unconvinced that the parameter values for the synthetic datasets would be appropriate for nTreg datasets. Therefore, the \texttt{normal.sigma} and \texttt{separation.factor} parameters were varied on the 5 randomly selected nTreg datasets. 10 runs were conducted, iterating the random number seed with each run. Figure 5.22a shows that \texttt{normal.sigma} value of 200 may give the best ASW for the most donors. This also corresponds to the peak in the mean category number and mean number of optimal solutions, as shown in Figures 5.22e and g. Similarly, for most donors, the peak in ASW occurs when \texttt{separation.factor} is 1 or 2, also corresponding with high mean category number and mean frequency of optimal solutions (Figure 5.22b, f and h respectively). This experiment indicates that the optimum parameter values may vary for each donor.

Again, a plateau in mean cluster number was not reached in either the \texttt{normal.sigma} or \texttt{separation.factor} parameter variation experiments. However, taking into account manual inspection of the data it is inappropriate to investigate higher parameter values. For example, in Figure 5.23a, the only example of optimal clustering occurs at \texttt{normal.sigma} = 200. At values lower than this, inappropriate cluster splitting was observed, yet for values higher than this, split non-nTreg and underestimated solutions were generated.

In the \texttt{separation.factor} variation experiment, the only examples of optimal clustering were seen at \texttt{separation.factor} = 2 or 4. Values less than this gave inappropriate cluster aggregation, and values higher than this gave inappropriate cluster splitting.

In summary, and in agreement with the conclusions from the synthetic experiments, \texttt{separation.factor} = 2 is appropriate for nTreg datasets.
Taking into account ASW and manual inspection of optimal clustering, the most appropriate value for `normal.sigma` is 200.
Figure 5.22: SamSPECTRAL parameter optimisation experiments on nTreg data.
Figure 5.22: Datasets comprising compensated, gated CD4$^+$ lymphocytes otherwise stained for CD25, CD127 and Foxp3 were analysed by SamSPECTRAL with varying parameter settings.  

a) ASW when varying \texttt{normal.sigma}.  
b) ASW when varying \texttt{separation.factor}.  
c) Standard deviation (sd) of ASW when varying \texttt{normal.sigma}.  
d) sd of ASW when varying \texttt{separation.factor}.  
e) Mean solution category when varying \texttt{normal.sigma}.  
f) Mean solution category when varying \texttt{separation.factor}.  
g) Mean number of optimal solutions when varying \texttt{normal.sigma}.  
h) Mean number of optimal solutions when varying \texttt{separation.factor}.  
i) Mean number of clusters when varying \texttt{normal.sigma}.  
j) Mean number of clusters when varying \texttt{separation.factor}.  

Measurements are reported as means across 10 runs, where the random number seed was iterated.
Datasets comprising compensated, gated CD4\(^+\) lymphocytes otherwise stained for CD25, CD127 and Foxp3 were analysed by SamSPECTRAL with varying parameter settings. a) normal.sigma optimisation on an nTreg dataset. b) Corresponding eigenvalue curves. c) separation.factor optimisation on an nTreg dataset.
5.5.2 Stability of SamSPECTRAL on nTreg data

Proceeding with \texttt{normal.sigma} = 200 and \texttt{separation.factor} = 2, SamSPECTRAL’s sensitivity to changes in data order and random number seed was investigated. The 5 randomly-selected data files were analysed by SamSPECTRAL 100 times. A \% of nTreg was reported for each run with an optimal solution. The number of optimal solutions was also reported. Few runs gave the same solution for each donor, and even for donors where the solution was often optimal, this was not preserved over all 100 runs.

Figure 5.24a shows the categories of solutions generated upon multiple runs on the same datasets. In each case, the orange segments represent the optimal solutions. D406318, for example, gave very few optimal solutions. D383401, however, gave an optimal solution in 96\% of runs. Figure 5.24b then shows the means and error bars of standard deviation for each donor in the analysis and reports only the percentage CV for the “optimal” data. Some donors have very large \% CVs, often corresponding to a small number of nTreg or a small frequency of optimal solutions. This reinforces the notion that assessing cluster validity by manual inspection or statistical methods is essential to ensure the reporting of optimal solutions and exclusion of other results.

Increasing the number of runs increases the likelihood of obtaining an optimal solution for each data file. However, it is not feasible to manually inspect 100 solutions for each file and to categorise them fairly. Therefore, in the subsequent nTreg enumeration across all 90 data files, only 10 runs were performed.
Datasets comprising compensated, gated CD4+ lymphocytes otherwise stained for CD25, CD127 and Foxp3 were analysed by SamSPECTRAL over 100 runs. a) Solution categories generated for 5 randomly selected datasets. b) Mean % nTreg from optimal runs within the same donors. The coefficients of variance for each donor were as follows: 18.3 for D491368, 37.5 for D406318, 3.88 for D469375, 6.85 for D817339 and 2.18 for D383401.
5.6 Comparison between SamSPECTRAL and manual gating

The value of using a clustering algorithm as opposed to manual gating has already been discussed. However, it was important to ascertain the agreement between SamSPECTRAL and manual gating in nTreg enumeration to test the hypothesis that SamSPECTRAL is as good as manual gating. For the five randomly selected donors, the FlowJo-gated quadrants were exported as separate .fcs files which were labelled and recombined in R. Following SamSPECTRAL clustering the quadrant labels versus the cluster assignments were compared using the F-measure outlined in Chapter 1 [6].

10 runs were conducted and the mean (± standard deviation) F-measures of the optimal solutions were 0.971 ± 0.00184 for D491368, 0 for D406318 as no optimal solutions were achieved in 10 runs, 0.962 ± 0.0179 for D469375, 0.937 ± 0.00346 for D817339 and 0.956 ± 0.0103 for D383401. These are surprisingly high given the discrepancy visible upon inspection of Figure 5.25, implying that the F-measure is an inappropriate measure of cluster validity. In addition, values such as 0.989, 0.940 and 0.985 were achieved for aggregated, split nTreg and underestimated failures, respectively. Split category solutions did generate lower values such as 0.751, and split non-nTreg solutions achieved F-statistics ranging from 0.657 to 0.951.

An example plot of manual gating compared to optimal SamSPECTRAL clustering is shown in Figure 5.25. It is clear that SamSPECTRAL gives a qualitatively superior solution than manual gating, encapsulating the whole CD127\textsuperscript{low}, CD25\textsuperscript{high} and Foxp3\textsuperscript{+} population, even though the F-measure suggests the two strategies are statistically similar.
Figure 5.25: *D383401: Manual gating compared to SamSPECTRAL clustering.*

Whole blood compensated, gated CD4$^+$ lymphocytes otherwise stained for CD25, CD127 and Foxp3 were analysed by SamSPECTRAL and compared to the original manual gating strategy. a)-b) Quadrant gating to enumerate nTreg, shown in red. c)-d) SamSPECTRAL clustering to enumerate nTreg, shown in red.
5.7 Drawing hard partitions around SamSPECTRAL-generated clusters for multisample comparison

A limitation of SamSPECTRAL clustering is that it does not consider negative control data as was routinely the case in the manual analyses. Previously, the FMOs were used to standardise our manually-generated measurements by subtracting false positive events. Quality assessment should remain an important feature within automated analysis techniques [23]. Therefore, drawing elliptical partitions around SamSPECTRAL-generated clusters that could allow for comparison with negative controls was investigated.

5.8 Does SamSPECTRAL give a different conclusion to nTreg frequency in peripheral blood with age?

Minimum volume spanning ellipses were fitted to the two largest SamSPECTRAL-derived clusters on a sample of nTreg datasets, the largest being the non-nTreg cluster, and the second largest usually being the nTreg cluster (Figure 5.26a). These elliptical “gates” were copied on to the scatter plots of the corresponding FMO controls in Figure 5.26b-c. The number of cells from the test and F-Foxp3 control datasets falling into the elliptical partitions was calculated, in this instance excluding the overlapping partition. The proportion of false positive events falling into the nTreg ellipse was calculated using the FMO and subtracted from the proportion calculated from the test dataset. This pipeline, interrupted by a manual checking stage, was applied to all 90 donors stained for nTreg markers, to observe if SamSPECTRAL provided a different conclusion to whether nTreg frequency changed with age.

To enumerate nTreg in all 90 donors of different ages, SamSPECTRAL was run 10 times on each data file, fixing the random number seed from 1 to 10. When 10 runs were performed at least one “optimal” solution for each file was generated for 86.7% of cases (ascertained by manual inspection). For one additional file successfully clustered by SamSPECTRAL, ellipses could not be fitted to the nTreg and non-nTreg clusters for correction for false positives for any of 10 runs, owing to large
Figure 5.26: Ellipses fitted to SamSPECTRAL-clustered data shown on all test and control samples.

Compensated, gated CD4\(^+\) lymphocytes stained for CD25, CD127 and Foxp3 were analysed by SamSPECTRAL and compared to FMO control datasets using gates of minimum volume spanning ellipses. a) Two ellipses fitted to the largest two clusters on the SamSPECTRAL-clustered nTreg datasets. b-d) The same two ellipses shown superimposed on to the corresponding F-Foxp3, F-CD25 and F-CD127 controls. All parts correspond to a single donor.
additional noise clusters. Therefore, only 85.6% of cases were automatically gated.

To improve the likelihood of obtaining an optimal solution for a given dataset more runs could be conducted with different random number seeds. However, this is currently limited by the requirement for manual inspection of each solution. A future analysis pipeline would need to be entirely unsupervised and to allow for replicates to be performed until an optimal solution is generated.

The mean number of optimal solutions for each data file was 6.42/10, with a mean %CV = 2.76 (not including zero values). This is significantly lower than for the manual gating method. Across the optimal solutions, the median percentage of nTreg in CD4+ lymphocytes was 5.89 (IQR: 5.03-6.64). This is significantly more sensitive than the manual gating method, which gave a median of 1.87 (1.35-2.62). The average number of clusters drawn by SamSPECTRAL from each dataset was 8 (7-9).

As shown in Figure 5.27, there was no correlation between nTreg enumerated using SamSPECTRAL with donor age. Comparing this with previous analyses, the conclusion about whether nTreg change in frequency with age does not differ. This validates the use of both manual gating and SamSPECTRAL for this purpose. However, it is notable that the percentage of nTreg changes markedly between the two methods.

5.9 Discussion

This chapter primarily compared the flow cytometry data-specific clustering algorithm SamSPECTRAL to benchmark k-means, and has shown it to be qualitatively superior for enumeration of nTreg. In synthetic experiments, SamSPECTRAL was able to cluster data when distance and density ratio of clusters was varied, but was less able to distinguish signal from noise. Then, SamSPECTRAL generated optimal solutions for the majority of our raw nTreg datasets.

However, we would have liked to see statistical proof of optimal solutions being achieved relative to suboptimal, and k-means and manual gating being inferior to SamSPECTRAL. Particularly with respect to the nTreg data, ASW was not a useful indicator of cluster validity. Using SamSPECTRAL on nTreg data, measurement of ASW is around 0.3, a moderately strong clustering. However, the most difficult failure to distinguish using both ASW and the clustering F-measure was the
Whole blood compensated, gated CD4+ lymphocytes stained for CD25, CD127 and Foxp3 were used to enumerate nTreg. The original manual gating method was compared to SamSPECTRAL clustering standardised for false positives using FMO control datasets and elliptical gated. a) Manual nTreg measurements included for comparison (n=90, r=-0.111, p=0.297). b) Optimal solutions following successful SamSPECTRAL clustering, elliptical gating and subtraction of false positive events from the corresponding FMO control datasets (n=77, r=-0.101, p=0.382). Points represent means of the optimal solutions generated from 10 repeats.

“aggregated” subtype where nTreg were not separated from non-nTreg. Having performed an initial correlation analysis of several cluster validity indices on SamSPECTRAL-clustered nTreg data, we reported ASW as it is not influenced by the clustering algorithm that originally generated a partition [229]. Although a single algorithm was originally selected, different validity indices may detect different aspects of a clustering partition, and thus not be universally inappropriate, as is the case with the choice of clustering algorithm. With retrospect, all four clustering indices could have been reported for each experiment rather than selecting ASW early on. Such an approach was used by Farrow and Masland to cluster visual channels in the mouse retina [79].

However, visual inspection of the clustering solution clearly showed that SamSPECTRAL best reflected the natural structure of the data and could indicate optimal solutions and parameter values. Thus, it was not possible to move to an entirely automated and objective method for analysing nTreg data, but using SamSPECTRAL in this way contributed an unbiased enumeration requiring only a manual checking stage. Such pipelines often build in interruptions for manual checking, such as flowType [5].
Although we were concerned that the percentage of differing opinions for categorisation of each data file was high, the three experts tended to conclude on very similar numbers of nTreg for each data file. A criticism of manual observation with the current implementation is that random graph colouring makes some clusters easier to separate than others. In a few cases, two clusters are coloured in the same colour and the expert is led to categorise the solution as a result of the percentages shown for each colour, not the clustering of the dot plot shown. In future implementations it will be necessary to standardise the colours used for each dataset. Another point of criticism of this analysis is that it takes so long to manually categorise each plot for each donor, that it cannot be guaranteed that the same stringency is carried out for each graph. Additional aspects such as light in the room and screen resolution will also impact if the analysis takes place in different locations, which may often be necessary.

Usually there is strong agreement between experts when manually determining the boundaries between large cell populations, but for small populations there is often disagreement [6]. In the FlowCAP study, F-measures were calculated between expert manual clusterings, and where the measure was 0.8 or lower, i.e. for data files where there was difficulty assigning clusters, the data files were ignored [6]. The authors thought it inappropriate to exclude outlying analyses as each set of gates represented an expert opinion on where the gates should lie [6]. This demonstrates a need to resolve the error inherent in manual gating strategies by developing automated methods [6].

Another approach to validate our clustering would be to use a supervised classification algorithm [250]. This method was used previously to validate a clustering of breast cancer pathologies to enhance clinical decision making [250]. A “valid clustering” would be achieved if a classification algorithm would be able to learn the rules behind the clustering in order to achieve an acceptable level of congruence [250]. The resultant classes could then be compared with the cluster assignment using a contingency table, and the proportion of agreeing solutions can be derived. A criticism of this approach is that the clustering and classification algorithms are used on the same data. However, if cross-validation is inherent in the classification implemented, this is arguably not the case. One might also run a variety of different classifiers to see if they have agreement.

SamSPECTRAL is hosted on the GenePattern server, and is therefore available to experimentalists who do not have programming expertise [252]. Modules from basic file conversions to automated gating and
multisample comparison can be linked into a customised analysis pipeline [252]. However, the stability of SamSPECTRAL with respect to random number seed, whilst encouraging, makes multiple runs for each donor a requirement. We have also shown that nTreg is sensitive to parameter variation, as are most algorithms [113]. To use such an algorithm, even via GenePattern, is fundamentally flawed if the parameters are suboptimal [113]. Therefore, it is inappropriate to perform a single run of SamSPECTRAL using the GenePattern platform [252]. The experimentalist needs both an understanding of the algorithm (arguably only achieved through programmatic and statistical knowledge) and to perform parameter optimisation experiments, programmatically in R or via tedious manual iteration using GenePattern. Unfortunately, in order for other laboratory experimentalists to take up SamSPECTRAL analysis routinely, they are most likely to use the GenePattern implementation for a single run, than to program R to perform multiple runs and collate the results.

Bashashati and colleagues previously proposed the following steps for data analysis: 1) Quality assessment, 2) Normalisation, 3) Outlier removal, 4) Automated gating, 5) Cluster labelling, 6) Feature extraction and 7) Interpretation [23]. Depending on the population of interest, these are more or less appropriate. For our nTreg data, quality assessment was initially performed through visual inspection and manual gating, comparing test and isotype control datasets. Normalisation and outlier removal were not performed as we prioritised maintaining the integrity of the raw data. We would be interested to see if normalisation improves SamSPECTRAL analyses in future. SamSPECTRAL was then used for automating nTreg analysis, with automated elliptical gating added. Cluster labelling and interpretation could not be automated owing to the complexity of the whole blood nTreg enumeration problem and could only be performed manually. Cluster labelling often requires a comparison of subpopulations across multiple data files [252], and is likely to be complicated by lack of machine calibration. Feature extraction is more relevant to exploratory data analysis as opposed to enumeration of a known subtype of interest. Were we not to use SamSPECTRAL in future, GenePattern does include a module for optimal cluster number estimation. It would also be useful to include similar modules for parameter optimisation analyses that would prevent inappropriate use of these algorithms by experimentalists without programming knowledge.

We have begun developing a strategy for comparison of test and
negative controls by drawing hard, elliptical partitions around SamSPECTRAL-generated clusters which can be translated from the test to control datasets. This could be further extended with the computation of a normalised $\chi^2$ statistic, as validated in [221, 220, 219], to discern statistically significant differences in the contents of elliptical partitions between test and control datasets. The F-Foxp3 and F-CD25 FMO controls would be expected to give statistically significant results, and the F-CD127 FMO controls would be statistically similar to the test stain. Owing to the small number of partitions defined, it will be necessary to define a high $\chi^2$ threshold to balance sensitivity with specificity for achieving statistically significant results in the case of our nTreg data. The original proponents of the frequency difference gating approach advocated that this parameter should be optimised for certain types of data analyses [219]. In future we would like to fit hyperellipsoidal or arbitrary-shaped multidimensional gates to better encapsulate nTreg clusters and to generalise to other multidimensional flow cytometry datasets, but their mathematical complexity currently places them beyond the scope of this work.

An important limitation of the current ellipse method is that the ellipses are fitted to 90% of the data. This is necessary to encapsulate the majority of the population of interest, to exclude outliers and to reduce the contribution of noise. However, curve smoothening and noise reduction was not performed prior to running SamSPECTRAL to ensure that the nTreg cluster was not compromised. The ellipse gating method may allow for consideration of isotype controls and subtraction of false positives, but still allows for some data to be lost in the process.

Although the main focus of this work is in automating nTreg enumeration, automated approaches to flow cytometry analysis are generalisable to other multiparametric staining panels and mass cytometry data analysis, both for data exploration and in testing particular hypotheses [28]. Several similar algorithms have been proposed in recent years and have been shown to perform well on large, multiparametric flow cytometry datasets, but very few have been independently validated by experimentalists. It is also unclear which algorithms perform best on which types of flow cytometry experiment. However, specific applications may still require specialist assay design and therefore specialist analysis techniques [218]. We further evaluate this work in Chapter 6.
6.1 General evaluation

The hypothesis that the balance between peripherally circulating pro- and anti-inflammatory T cell subsets changes throughout the lifetime was investigated in a cross-sectional study of 112 donors of different ages. A novel dataset of T cell subsets throughout the lifetime was presented, and the knowledge that an IL-10-producing induced regulatory T cell population had an increased frequency relative to Th17 cells and nTreg in older donors was contributed. Then, the performance of SamSPECTRAL on our nTreg dataset was compared to basic $k$-means clustering and manual gating and SamSPECTRAL was shown to be qualitatively but not quantitatively superior. A novel method of fitting elliptical gates to SamSPECTRAL-generated clusters then permitted comparison between test and control datasets within the same experiment.

In the rest of this chapter the work presented in this thesis is evaluated, compared to recent work by other groups, and avenues for further improvement and extension in future work are identified.

6.1.1 Critique of ageing study

Despite this area of immunology representing significant complexity arising from overlapping functionality and compensatory mechanisms, our data showed that the number of iTreg significantly, but moderately correlated with age ($p=0.0035$ and $r=-0.303$). Given that IL-10 has local effects on almost all cells of the immune system and is able to induce further IL-10 secretion in a positive feedback loop [237], we propose that even a modest
correlation with age might contribute to clinically significant alterations in disease processes and therapeutic efficacy. For example, IL-10 is a limiting factor in the effectiveness of antitumour immunisation [197]. Unfortunately, IL-10 seems to be the by-product of many current immunotherapy strategies in cancer and infectious disease [175, 213]. Antagonising IL-10, for example, with neutralising antibodies to the IL-10 receptor, may be a useful adjuvant in a number of immunisations targeting tumour antigens or in broader use against infectious diseases. This approach has recently been shown to improve clinical and immunological responses to BCG vaccination against *M. tuberculosis* infection in mice [209]. Our data suggests that a similar approach of dampening the effect of IL-10 may be especially important when designing therapeutic strategies for the elderly. However, therapeutic manipulation of IL-10, though promising, ought to be approached with care as cytokine imbalance can result in a variety of problems. Infections may persist if there is a relative abundance of regulatory cytokines, or inappropriate damage to the host can occur with an abundance of proinflammatory cytokines [276]. Though IL-10 antagonists have benefits as adjuvants, the lack of IL-10 and sufficient regulation comes with its own problems. Autoimmune conditions have been shown to correlate with altered IL-10-mediated regulation, for example, ulcerative colitis in both mice and humans [237, 91]. Therapeutic correction of IL-10 abundance with agents designed to stimulate or block is likely to be limited in its ability to maintain a state of optimum balance.

In future work it would be crucial to ascertain the source of the increase in iTreg and to conduct a deeper analysis of their function in donors of different ages. Also, in order to demonstrate a causal link between immune parameters and age-related disease, it would be interesting to compare a cohort with a certain diagnosis with SENIEUR healthy and “normal” healthy control populations to try to determine specific immune system changes relevant to that particular pathology, as suggested by Castle and colleagues [49]. A similar strategy was used by Rosenkranz and colleagues in their study of regulatory T cells in neurodegenerative diseases [228].

The whole blood assay for enumerating cytokine-secreting cells relies on the assumption that PMA and ionomycin stimulation initiates a recall response in which pre-differentiated CD4⁺ T cells secrete their signature panel of cytokines. However, it has previously been shown that repeated TCR triggering causes increased IL-10 production by Th1 cells [236]. Although we found that the IL-10⁺ Th1 population does not increase with age, it is possible that enhanced TCR triggering may cause other T helper
subsets to secrete IL-10. Firstly, it is possible that a strong mitogenic stimulus such as PMA and ionomycin may actually cause an increase in IL-10 secretion by cells that ordinarily would not have been programmed to. However, we minimise this effect by applying consistent stimulation conditions to each blood sample and comparing a large number of donors. Secondly, this repeated TCR triggering may arise from infection with persistent herpes viruses such as cytomegalovirus (CMV) or Epstein-Barr virus (EBV), known to markedly change the size of leukocyte populations [56]. In which case, we hypothesise that there may be a causal link between an inflated IL-10-secreting population and CMV seropositivity known to be associated with ageing [203].

Relating to our assumption that we effectively mimick a recall response, we also assume that each T cell subtype retains its phenotypic integrity and is committed by the same mechanism. The cytokine secretion during the memory response is controlled by gene expression, which is, in turn, regulated in part by epigenetic modifications [70]. A recent paper showed that only approximately 10% of isolated IL-10^+IFNγ^- T cells retained the ability to produce IL-10 following 1 week of expansion in culture and restimulation with PMA and ionomycin [70]. In comparison, over 90% of IFNγ^+ continued to express IFNγ with the same conditions [70]. Hypomethylation of the IFNγ gene promotor region was observed in IFNγ^+ memory T cells, and was not observed in IFNγ^- memory T cells [70]. In contrast, little difference was observed in the methylation status of the IL-10 gene between IL-10^+ and IL-10^- memory T cells, suggesting that memory T cells may not retain the tendency to produce IL-10 as they may for other cytokines such as IFNγ [70]. This may explain why such low numbers of CD4^+ T cells secrete IL-10 on stimulation with PMA and ionomycin. Th17 cells have been shown to maintain their characteristic cytokine profiles [60]. ChIP was used to show that naïve, Th1 and Th1/17 cells had enrichment of activating histone modifications (H3K4me3) at the IFNG promoter, whereas this was decreased in conventional Th17 cells in favour of more H3K4me3 at the IL17A promoter [60].

It has previously been shown that lifespan is different for men and women, and that T cells subsets shown to change with age (CD8^+αβ^+ and CD3^+CD45RA^-CCR7^- effector memory populations) changed more steeply in men [288]. Were we to recruit more donors, it would be interesting to perform a deeper analysis to ascertain if the balance between regulatory and inflammatory T cell subjects changes according to medical and demographic variables.
6.1.2 Critique of SamSPECTRAL investigation

The developers of SPADE argue that the quality of results from automated analysis depends on the raw data; they use the example of staining for the right markers [214]. However, the quality of the data going in can vary according to antibody clone, batch, fluorochrome, quality of experimental technique, etc, in more subtle ways than just the selection of the right markers, which they do not acknowledge. We developed a high quality flow cytometry assay for nTreg using a sensitive α-Foxp3 clone and buffer pair and titrating antibodies for maximum sensitivity, minimum background and cost-effectiveness. However, it is clear from our comparison of nTreg clustering with a simple CD4 enumeration problem that the nature of the raw data influences the likelihood of achieving adequate automation. Even with an improved clustering algorithm, it still may be difficult to identify the correct nTreg and non-nTreg clusters in all cases as the problem remains that if a population is small, the algorithm assumes it is noise and may attempt to cluster wrongly [23].

Although we have shown SamSPECTRAL to cluster our nTreg datasets favourably with high frequency, the barrier to its more widespread use by laboratory immunologists is its implementation. The fact that Bakker and colleagues wrote a flow cytometry data-specific clustering algorithm based on $k$-means in Pascal back in 1993, and yet we still prefer manual analysis 20 years on, is evidence for this. Whilst R developers can exploit open resources to improve clustering algorithms and create bespoke analysis pipelines at the code level, it will be necessary to develop more user-friendly software platforms to advance the field as a whole. Although some commercial tools are available, such as the “cluster” and “multi-sample comparison” function in FlowJo [1], prohibitively technical or opaque descriptions of algorithm function and poor graphical user interfaces which do not allow automated replicates or systematic parameter sensitivity analyses restrict their use. We have shown that such replication and parameter optimisation is required to maximise the efficiency of SamSPECTRAL on our nTreg data files.

6.1.3 Comparison with other recent work

An alternative whole blood assay design

A recent paper by Hardy and colleagues outlined a whole blood nTreg enumeration assay staining for CD4, CD25, CD127 and Foxp3 [116].
Although the exact antibodies and fixation/permeabilization buffers used differ, their staining is qualitatively similar to ours [116].

Hardy and colleagues opted to stain in TruCount\textsuperscript{TM} Tubes (BD) as an alternative to adding FlowCount Fluorospheres for measuring absolute frequencies, which are standard FACS tubes containing a freeze-dried pellet that dissolves during sample preparation, releasing a known number of fluorescent beads [31]. Certainly for protocols not requiring a wash step stain, they allow for a more reliable absolute cell enumeration as cell loss is avoided and pipetting FlowCount Fluorospheres manually involves a degree of error. However, as several wash steps are necessary for intracellular staining, cell loss would undoubtedly have occurred. Therefore, a more precise measure of beads did not translate to a better quality measure of absolute number.

Hardy and colleagues also assessed inter-assay variation in nTreg enumeration by taking 3 samples from 3 donors, each 7 days apart and the CVs were between 8.3 and 34.5% for absolute counts, and 8.1 to 19.6% in the proportion measurements [116]. For CD4\textsuperscript{+} lymphocytes, the CVs were 2.4 to 17.8% [116]. Arguably, however, this is not a test of inter-assay variation but a test of donor variation; at least it is impossible to attribute the error to the contribution of each source. The CVs of intra-assay variation across 5 replicate stains in 5 donors ranged from 2.6 to 8.5% [116]. In our experiments, the mean CV of intra-assay variation across 3 donors was 8.34%.

The authors make reference to the use of CD25 and CD127 FMO controls [116]. Manually-drawn tight polygonal gates are the preferred method of distinction, with no justification for their specific shape [116]. Assuming that an optimal plot has been chosen by the authors, one has to question whether the gates drawn are robust to all donor variation across all assays, in terms of isotypic binding, cell death, population frequency as well as sources of experimental error [116]. We would argue the case for a more stringent, rule-based manual gating strategy as outlined in Chapter 2 of this thesis, or preferably, an automated algorithm such as SamSPECTRAL for more objective analysis of these datasets.

**An alternative staining panel for nTreg enumeration**

A recent pivotal paper by Miyara and colleagues has advocated an alternative nTreg staining panel that would allow for discrimination between three T cell populations by differential expression of CD45RA.
and Foxp3 [184]. CD45RA$^+$Foxp3$^{low}$ events are thought to correspond to resting nTreg, CD45RA$^-$Foxp3$^{high}$ are thought to be activated nTreg, and CD45RA$^-$Foxp3$^{low}$ [184] are thought to be cytokine-secreting but not regulatory [184]. This strategy has been advocated by other reputable groups [24, 32].

Miyara and colleagues proposed that CD25 and CD45RA may be better markers for fluorescence-activated cell sorting (FACS) than CD127 and CD25 [184]. However, they used CD127 positivity to confirm inability of their CD45RA$^-$Foxp3$^{low}$ fraction to suppress [184]. This validates our use of CD127 in preference to CD45RA. It would be interesting in future to conduct some validation work comparing our enumeration strategies with theirs. We would argue that both strategies are likely to allow for non-Treg contamination of cultures sorted on the basis of these molecular markers.

We would argue that some experiments conducted by Miyara and colleagues are flawed, such as their use of PMA/ionomycin stimulation prior to cytokine and Foxp3 staining. Additional Foxp3 staining was performed on unstimulated cultures, which we have shown to measure very different numbers of Foxp3$^+$ events. Two different Foxp3 clones were also used in this paper which we have observed are clearly different in their sensitivity for binding Foxp3 (data not shown). Furthermore, it was unclear how Miyara and colleagues decided the positioning of various gates on CD45RA$^+/-$, Foxp3$^{high/low}$ and CD25$^{high/+low/-}$ populations making it very difficult to repeat their experiments meaningfully in our own work.

Alternative clustering algorithms for nTreg enumeration

Since we began our work on SamSPECTRAL, a paper by Pyne and colleagues introduced the flowScape algorithm, arguing that the human perspective integral to manual gating methods should be preserved in computational approaches to flow cytometry data analysis [216]. flowScape makes the subjective gating processes a human performs objective by mapping the dataset using Modal Expectation Maximisation, and then the modal structures are arranged in a hierarchy from the bottom up [216]. Ridgeline analysis is then used to locate peaks and troughs in the data and to interpret how closely they are connected from the top down, arguably as a human does visually [216]. flowScape was shown to correctly identify lymphocytes from forward and side scatter data at 2 timepoints from a single donor from the GvHD dataset, where flowCore failed, and SamSPECTRAL only succeeded on one of the two
data files. It would be interesting to see how flowScape performs on the nTreg datasets. In addition, the novel clinical criteria-driven population selection approaches also reviewed in Chapter 1 are more streamlined in characterising the most relevant populations from a thorough mine of the raw data. In future studies of cancer and ageing, these algorithms ought to be more frequently used.

6.2 Future directions

6.2.1 Improving the whole blood assay for routine immunophenotyping

As part of the ageing study, an assessment of confounding factors was conducted to see if the changes observed could be attributed to a variable other than age. As the age of donors correlated strongly with both chronological assay number and the time the sample was extracted, it was not possible to identify if either of these were confounding factors. Seasonal variation, circadian variation and reagent degradation over time remain known phenomena [185, 137]. Therefore in future studies an approach to recruitment that does not systematically relate to the underlying biology will require careful thought and planning.

Many flow cytometry-based assays now include viability stains routinely [33, 112, 189]. Owing to the nature of the whole blood assay and its short term culture, we estimate that the proportion of dead events is minimal in both the stimulated and unstimulated conditions, prior to lysis. However, even this small amount is a concern when investigating rare populations, as dead cells are able to bind antibodies non-specifically [206]. If this had occurred within our experiments we would have corrected for it by subtracting false positive events as indicated by the isotype-FMO controls we ran with each experiment.

In order to minimise the noise present within the gated CD4$^+$ lymphocyte population in future experiments, we will add a CD14 stain to allow us to gate out monocytes which express low levels of CD4 [141]. Also, in the stimulated conditions, we would also add CD3 and CD8 stains to assist with gating CD4-downregulated T cell subsets.

In addition to healthy ageing, altered Treg:Th17 balance may occur in many different cancers, including gastric and uterine cervical cancer [179, 296, 145]. Whether this is beneficial or not remains a point of
controversy, and warrants further investigation. For example, a relative increase in tumour-infiltrating Th17 cells has been shown to be associated with good prognosis in ovarian cancer [188], but detrimental in colorectal cancer [269]. In addition to the direct effects of tumour and metastasis on surrounding tissues, approximately half of patients with cancer experience cachexia, a systemic pathology negatively associated with decreased survival and response to therapy [266]. It is characterised by altered metabolism of carbohydrates, lipids and proteins in a cytokine-driven process that is incompletely understood [266]. An increase in proinflammatory cytokines such as IL-6 is associated, [22, 187], and IL-6 is known to influence development and function of regulatory T cells and their balance with inflammatory subsets [96]. We hypothesise that the balance between T cell subsets may be altered in patients with cachexia, may drive the disease process and may prevent efficacy of therapeutic intervention. Therefore, in the next paragraphs, we highlight how the whole blood assay may be improved for use in future clinical studies of cancer and cachexia.

This work has concentrated on specific natural and induced CD4+ regulatory T cell subtypes. However, a number of other regulatory T cells have been proposed, for example NKT-cells, CD8+CD25+ thymocytes, CD8+CD28- T cells, γδ T cells, and TGF-β secreting Th3 cells which will be worth investigating in future work [66]. Recent literature has shown that other non T-cell subsets may influence and complicate the ability to maintain Treg:Th17 balance in disease processes. MDSCs are currently being investigated for their role in suppressing antitumour immune responses and reducing the efficacy of immunotherapies [199, 235, 234]. MDSCs are in turn thought to be induced by proinflammatory signals [199], illustrating that they are another core component in the dynamics of maintaining immune homeostasis. Also, although this thesis has focused on CD4+ T cells which coordinate the immune response, how they become differentiated from naïve precursors to a particular mature phenotype depends both on the context of their stimulation and subsequent exposure to polarising signals. For example, a subset of bone marrow dendritic cells differentiated in an IL-10-rich environment have been shown to convert CD4+CD25lowFoxp3- into CD4+CD25highFoxp3+ upon in vitro coculture and in a mouse model [125]. Depending on the context of DC activation, DC can become more immunogenic or tolerogenic [230]. Conversely, the proinflammatory Human 6-sulfo LacNAc dendritic cell population of the myeloid lineage (slanDC) has been shown to drive Th1 and Th17
responses [115].

It would be beneficial to integrate assays for these novel populations into the current panel for T cell subsets. Accompanying this new panel design, we aim to develop automated analysis pipelines using SamSPECTRAL as the main clustering algorithm, customised for each population of interest.

6.2.2 Constrained spectral clustering

Taking forward the opinion of Pyne and colleagues that the human perspective integral to manual gating methods ought to be preserved in computational approaches to flow cytometry data analysis [216], we intend to investigate constrained clustering. Such an approach can be termed “semi-supervised” [128]. The algorithm will learn the features of a dataset that we think are important, without us directly manipulating the solution [284, 289]. We hypothesise that influencing SamSPECTRAL with user-specified constraints will increase the frequency of optimal clustering.

Constrained versions of many popular clustering algorithms such as $k$-means and hierarchical clustering already exist [279]. These incorporate user-specified must- and cannot-links; examples of datapoints which belong in the same cluster or different clusters, respectively [279, 133]. They fill the niche for problems where some expert knowledge is available, yet the class labels for each data point are not [131]. There has been some interest in implementing constrained spectral clustering algorithms [279, 133, 131, 169], but how best to integrate these constraints is unclear [279]. Either the input similarity matrix, or the resultant eigenspace can be altered [279]. By building on SamSPECTRAL in this way the majority of its benefits would be preserved, such as faithful sampling and modified spectral clustering [279].

To ensure that adding constraints does not reduce the quality of SamSPECTRAL-generated solutions, the F-measure used to compare algorithms within the FlowCAP initiative would be an optimal measure of cluster validity [6]. Demonstrating an improvement in clustering validity, however, would remain a problem. It is possible that such a method will perform poorly owing to the small number of constrained pairs relative to the total number of pairs in the dataset [169]. An approach by Lu and Carreira-Perpiñán is to increase the influence of the constraints by affinity propagation [169]. Should we find our basic implementation performs poorly we could look at implementing such a more complex method.
6.2.3 Simulating the ageing immune system

With only a handful of T cell immunophenotyping studies in donors of different ages, little insight has been obtained as to the reasons for observed changes or homeostasis in T cell populations. Considering the unchanging frequency of nTreg with age we have observed and the evidence that nTreg are generated in the thymus, which is known to involute from adolescence, the lifetime dynamics of nTreg cells are very interesting [66, 168]. Methods of achieving homeostasis in nTreg have been previously examined using computer simulation [20, 85].

It has been argued that the trend of proposing a single hypothesis and testing it is stepping aside, making way for high-throughput experiments combined with data mining and simulation techniques to generate a broad and comprehensive view of system function [99]. A cyclical model of the management of such an interdisciplinary project involves data mining of laboratory data, simulation of the knowledge of a system quantified using the results, then systematic hypothesis generation prior to further laboratory testing [99]. This thesis describes the initial generation of a laboratory dataset with an example of data mining to improve the quality of analysis compared to traditional manual techniques, and to identify potential confounding factors. We aim to build a simulation of how the immune system changes throughout life, integrating all of our T cell subset data to consider how the balance between pro- and anti-inflammatory subsets changes. Sensitivity analyses will be performed to elucidate the most important parameters for control of their dynamics throughout life. Our preliminary work based on Baltcheva’s model of regulatory T cell homeostasis outlined in [85] has not fitted the scope of this thesis and we aim to complete the cycle in future work.

For the abstraction of how T cell subsets change with age, differential equation based mathematical models are likely to be most appropriate given the cell count data that has been collected [99]. However, in order to make meaningful progress, more detailed mechanisms need to be elucidated. This poses significant challenges experimentally. To simulate (with ordinary differential equations) the change in numbers of cells over time requires rates of association and dissociation of molecules or cells depending on the relevant level of abstraction [99]. Making things more complicated, stochastic spatially resolved simulations would require individual particle data, data on cell motility and their interactions, from multicolour time-lapse microscopy, for example [99]. Specific quantitative
data is required, such as molecules in a given volume, whether they have any biochemical modification, their affinities and rates of transformation, cell proliferation and death rates, etc [99]. It is also very difficult to simulate a phenomenon without complete knowledge of the mechanisms involved and predictable outputs must be then testable in the laboratory [99]. It will remain a challenge to ask the right questions to best optimise the technology we have available.

Frequently, simulation results are not equal to the results of *in vitro* experimentation. This is due to the abstraction and simplification required to produce a model, and the incomplete knowledge of the biological field on which a model is likely to be based [10]. A systematic approach to validation has been recently proposed [10, 102]. Argument-driven validation techniques such as goal-structuring notation may provide a comprehensive framework with which to assess all aspects of the modelling process [10, 102]. The model is broken down into steps which are checked against all the relevant evidence supporting that step, with an aim of increasing confidence in parts of the model that are well supported and identifying parts of the model that need to receive more attention [10].

The process of simulation alone might allow researchers to address assumptions and allow for systematic generation of hypotheses [85]. Also, hypotheses which are difficult to test in the laboratory might be testable with a simulation [85]. Simulating the dynamics of T cell subsets in parallel will allow us to make predictions about the maintenance of their balance throughout life, would also allow for extreme parameter values to be tested and may indicate a maximum length of time for homeostasis to be maintained.

### 6.3 Summary

In summary, this thesis has contributed a validated whole blood flow cytometry assay to characterise peripherally circulating CD4\(^+\) T cell subsets using extra- and intracellular staining. This was used to show, for the first time, that iTreg increase relative to nTreg and Th17 with healthy ageing. The source of increased iTreg and the implications of their increased frequency for disease processes and immunotherapies remain to be established.

This work also included an evaluation of the performance of SamSPECTRAL on nTreg datasets generated during the ageing study.
We conclude that SamSPECTRAL is much more capable than $k$-means in analysing flow cytometry data, and argue that SamSPECTRAL clustering is superior to manual analysis in quickly producing more objective and accurate results. However, work still needs to be done to automate the SamSPECTRAL analysis pipeline, improve the frequency of obtaining an optimal solution, demonstrate mathematical validity of clustering partitions generated and include multi-sample comparison.
Bibliography


[100] Carlos Gershenson. *Artificial Neural Networks for Beginners*. [cited at p. 33]


Fixation/Permeabilization Buffer Combinations in Identifying CD4+CD25+Foxp3+ Regulatory T Cells. *Cytometry A*, 77A(2). [cited at p. 87]


212


[cited at p. 10]

[cited at p. 5, 8]

[cited at p. 11]


S. C. Pino, B. O’Sullivan-Murphy, E. A. Lidstone, T. B. Thornley, A. Jurczyk, F. Urano, D. L. Greiner, J. P. Mordes, A. A. Rossini, and


220


R code vignette

This section outlines example R code required to replicate the experiments presented in this thesis.

Example 1: Loading pre-installed packages required for automated nTreg data analysis

```r
library(flowCore)
library(SamSPECTRAL)
library(cluster)
library(car)
library(stats)
library(fpc)
library(gtools)
```

Example 2: Importing data from a .csv file containing donor ID in column 2, test filename in column 4 and channel for PE-Cy5 detection in column 5. Two FC500 flow cytometers were used in this analysis which recorded PE-Cy5 and ECD fluorescence in different columns of the .fcs file. Vectors are created which are referred to in subsequent loops

```r
master<-read.csv("MasterSheetFCS.csv",header=TRUE,sep=" ")
masterID<-master[,2]
mastertestfile<-master[,4]
testfl4channel<-master[,5]
```
Example 3: To record algorithm runtime, line one is added at the start of the loop, and line two at the end

```r
initime<-Sys.time()
endtime<-Sys.time()-initime
```

Example 4: Looping through the above variables to find which is the FL3 channel to be excluded for each dataset, produce the file path and open each test file

```r
for(d in 1:92){
donorID<-masterID[d]
testfile<-mastertestfile[d]
testchannel<-testfl4channel[d]

testfilepath<-paste("AllFCS/",testfile,sep="")
test<-read.FCS(system.file(testfilepath, package="flowCore"))
...
}
```

Example 5: Removing all superfluous scatter and fluorescence channel information

```r
size<-ncol(test)
if (size==12){
test<-test[,c(12)]
}
if(testchannel==11){
testfl3channel<-10
}else{
testfl3channel<-11
}
restrictedtest<-test[,c(1,2,3,4,5,6,7,testfl3channel)]
```
Example 6: Log-transforming and randomising the data prior to running SamSPECTRAL. Adapted from the flowCore package vignette [75]

data.pointstest<-restrictedtest@exprs
fulltest<-log10(data.pointstest)
lengthtest<-length(fulltest)/3
set.seed(1)
randomisedtest<- fulltest[order(runif(nrow(fulltest)))]

Example 7: Running SamSPECTRAL on our nTreg datasets and collecting cluster frequencies. Adapted from the SamSPECTRAL package vignette [294]

P<-SamSPECTRAL(randomisedtest,dimensions=c(1,2,3),normal.sigma=200,separation.factor = 2)
frequency<-c()
minimum.frequency<-0.01
frequency.large<-c()
labels<-as.character(unique(P))
for(label in labels){
  if(!is.na(label)){
    frequency[label]<-length(which(P==label))/length(P)
    if(frequency[label] > minimum.frequency)
      frequency.large[label]<-frequency[label]
  }
}
Example 8: Calculating the maximum number of clusters generated, combining the data and clustered matrix, identifying unclustered data, ordering the clusters from largest to smallest clusters and looping through the clusters to collect information about individual clusters

```r
maxP <- max(P, na.rm = TRUE)
clustered <- matrix(cbind(randomisedtest, P$labels), nrow(randomisedtest))

clusteredtest <- na.omit(clustered)
lengthNA <- length(clustered[,4]) - length(clusteredtest[,4])

resclusterrsize <- numeric(maxP)
mfifl1 <- numeric(maxP)
mfifl2 <- numeric(maxP)
mfifl4 <- numeric(maxP)

for (j in 1:maxP) {
  clusterlocation <- order[j]
  resclusterrsize[j] <- frequency[clusterlocation]
  cluster <- subset(clustered, clustered[,4] == clusteralabel[clusterlocation])
  mfifl1[j] <- median(cluster[,1])
  mfifl2[j] <- median(cluster[,2])
  mfifl4[j] <- median(cluster[,3])
}

resdfrm <- data.frame(j, resclusterrsize, mfifl1, mfifl2, mfifl4)
```
Example 9: Adapting the dataEllipse method from the car+ package [90] for fitting and reproducing minimum volume spanning ellipses

```r
adapted.dataEllipse <- function (x, y, groups, group.labels = group.levels, ellipse.label, weights, log = "", levels = c(0.5, 0.95), center.pch = 19, center.cex = 1.5, draw = TRUE, plot.points = draw, add = !plot.points, segments = 51, robust = FALSE, xlab = deparse(substitute(x)), ylab = deparse(substitute(y)), col = if (missing(groups)) palette()[1:2] else palette()[1:length(group.levels)], pch = if (missing(groups)) 1 else seq(group.levels), lwd = 2, fill = FALSE, fill.alpha = 0.3, grid = TRUE, labels, id.method = "mahal", id.n = if (id.method[1] == "identify") Inf else 0, id.cex = 1, id.col = if (missing(groups)) palette()[1] else palette()[1:length(groups)], ...) {
  label.ellipse <- function(ellipse, label, col, ...) {
    if (cor(ellipse)[1, 2] >= 0) {
      index <- which.max(ellipse[, 2])
      x <- ellipse[index, 1] + 0.5 * strwidth(label)
      y <- ellipse[index, 2] + 0.5 * strheight("A")
      adj <- c(1, 0)
    } else {
      index <- which.min(ellipse[, 2])
      x <- ellipse[index, 1] - 0.5 * strwidth(label)
      y <- ellipse[index, 2] - 0.5 * strheight("A")
      adj <- c(0, 1)
    }
    text(x, y, label, adj = adj, col = col, ...)
  }
  if (missing(y)) {
    if (is.matrix(x) && ncol(x) == 2) {
      if (missing(xlab))
        xlab <- colnames(x)[1]
      if (missing(ylab))
        ylab <- colnames(x)[2]
      y <- x[, 2]
      x <- x[, 1]
    }
  }
}
```

231
else stop("x and y must be vectors, or x must be a 2 column matrix")
}
else if (!(is.vector(x) && is.vector(y) &&
    length(x) == length(y)))
  stop("x and y must be vectors of the same length")
if (missing(weights))
  weights <- rep(1, length(x))
if (length(weights) != length(x))
  stop("weights must be of the same length as x and y")
if (!missing(groups)) {
  xlab
  ylab
  if (!is.factor(groups))
    stop("groups must be a factor")
  if (!(length(groups) == length(x)))
    stop("groups, x, and y must all be of the same length")
  if (missing(labels))
    labels <- seq(length(x))
  valid <- complete.cases(x, y, groups)
  x <- x[valid]
  y <- y[valid]
  weights <- weights[valid]
  groups <- groups[valid]
  labels <- labels[valid]
  group.levels <- levels(groups)
  result <- vector(length(group.levels), mode = "list")
  names(result) <- group.levels
  if (draw) {
    if (!add) {
      plot(x, y, type = "n", xlab = xlab, ylab = ylab,
      ...
      if (grid) {
        grid(lty = 1, equilogs = FALSE)
        box()
      }
    }
  }
}
for (lev in 1:length(group.levels)) {
    level <- group.levels[lev]
    sel <- groups == level
    result[[lev]] <- dataEllipsoid(x[sel], y[sel],
        weights = weights[sel], log = log, levels = levels,
        center.pch = center.pch, center.cex = center.cex,
        draw = draw, plot.points = plot.points,
        add = TRUE, segments = segments, robust = robust,
        col = rep(col[lev], 2), pch = pch[lev], lwd = lwd,
        fill = fill, fill.alpha = fill.alpha,
        labels = labels[sel], id.method = id.method,
        id.n = id.n, id.cex = id.cex, id.col = col[lev],
        ellipse.label = group.labels[lev],
        ...)
}
return(invisible(result))
}
if (length(col) == 1)
    col <- rep(col, 2)
if (draw) {
    if (!add) {
        plot(x, y, type = "n", xlab = xlab, ylab = ylab,
        ...) 
        if (grid) {
            grid(lty = 1, equilogs = FALSE)
            box()
        }
    }
    if (plot.points)
        points(x, y, col = col[1], pch = pch[1], ...)
}
dfn <- 2
dfd <- length(x) - 1
if (robust) {
    use <- weights > 0
    v <- cov.trob(cbind(x[use], y[use]), wt = weights[use])
    shape <- v$cov
    center <- v$center
} else {
    v <- cov.wt(cbind(x, y), wt = weights)
    shape <- v$cov
    center <- v$center
}
result <- vector("list", length = length(levels))
names(result) <- levels
for (i in seq(along = levels)) {
  level <- levels[i]
  radius <- sqrt(dfn * qf(level, dfn, dfd))
  result[[i]] <- ellipse(center, shape, radius, 
    log = log, center.pch = center.pch, 
    center.cex = center.cex, segments = segments, 
    col = col[2], lwd = lwd, fill = fill, 
    fill.alpha = fill.alpha, draw = draw, ...)
  if (!missing(ellipse.label)) {
    lab <- if (length(ellipse.label) < i) 
      ellipse.label[1] 
    else ellipse.label[i]
    label.ellipse(result[[i]], lab, col[2], ...)
  }
}
if (missing(labels))
  labels <- seq(length(x))
showLabels(x, y, labels = labels, id.method = id.method, 
  id.n = id.n, id.cex = id.cex, id.col = id.col)
invisible(if (length(levels) == 1) result[[1]] else result)

out <- list(shape=shape, radius=radius, center=center, 
  alpha=fill.alpha, center.cex = center.cex, segments=segments)
out
}
Example 10: Applying the adapted dataEllipse function to fit and transfer minimum volume spanning ellipses to the largest two clusters, harvesting the data and ellipse parameters in order to reproduce ellipses on FMO controls and calculate how many events fall within them, creating a vector of control partitions

```r
i<-1
clusterlocation<-order[i]
resclustersize[i]<- frequency[clusterlocation]
cluster1<-subset(clustered, clustered[,4]== clusterlabel[clusterlocation])
sizecluster1<-length(cluster1)/4
x<-cluster1[,c(1)]
y<-cluster1[,c(3)]
ellipse1<-adapteddataEllipse(x,y,levels=0.9,
pch='.', xlab="Foxp3",ylab="CD25", xlim=c(0, 4), ylim=c(0,4))
i<-2
clusterlocation<-order[i]
resclustersize[i]<- frequency[clusterlocation]
cluster2<-subset(clustered, clustered[,4]== clusterlabel[clusterlocation])
sizecluster2<-length(cluster2)/4
x<-cluster2[,c(1)]
y<-cluster2[,c(3)]
ellipse2<-adapteddataEllipse(x,y,levels=0.9,pch='.’,
xlab="Foxp3",ylab="CD25", xlim=c(0, 4), ylim=c(0,4))

datasize<-length(clustered)/4
foxp3<- clustered[,1]
cd25<- clustered[,3]
diagshape1<-diag(ellipse1$shape)
diagshape2<-diag(ellipse2$shape)
a21<- diagshape1[1]
a22<- diagshape2[1]
c21<- diagshape1[2]
c22<- diagshape2[2]
```
r1 <- ellipse1$radius^2
r2 <- ellipse2$radius^2
x01 <- ellipse1$center[1]
z01 <- ellipse1$center[2]
x02 <- ellipse2$center[1]
z02 <- ellipse2$center[2]
inside1 <- numeric(datasize)
inside2 <- numeric(datasize)

eigens1 <- eigen(ellipse1$shape)
evs1 <- sqrt(eigens1$values)
evecs1 <- eigens1$vectors
alpha1 <- atan(evecs1[, 1][1] / evecs1[, 2][1])

eigens2 <- eigen(ellipse2$shape)
evs2 <- sqrt(eigens2$values)
evecs2 <- eigens2$vectors
alpha2 <- atan(evecs2[, 1][1] / evecs2[, 2][1])

for (p in 1:datasize)
{x <- foxp3[p]
z <- cd25[p]

u <- cos(alpha1)*(x-x01) + sin(alpha1)*(z-z01)
v <- -sin(alpha1)*(x-x01) + cos(alpha1)*(z-z01)
in1 <- (u)^2/(a21) + (v)^2/(c21)
if (in1 < r1)
{inside1[p] <- 1
} else{
inside1[p] <- 0
}
}
propinside1 <- sum(inside1)/datasize
propinside1
for(q in 1:datasize){
x<-foxp3[q]
z<-cd25[q]
  u <- cos(alpha2)*(x-x02) + sin(alpha2)*(z-z02)
v <- -sin(alpha2)*(x-x02) + cos(alpha2)*(z-z02)
in2<-(u)^2/(a22)+(v)^2/(c22)
  if(in2<r2){
in2[q]<-1
  }else{
in2[q]<-0
  }
}
propinside2<-sum(inside2)/datasize
propinside2

fullCD4fFoxp3datasize<-length(fullCD4fFoxp3)/3
fullCD4fFoxp3foxp3<- fullCD4fFoxp3[,1]
fullCD4fFoxp3cd25<- fullCD4fFoxp3[,3]
fullCD4fFoxp3inside1<-numeric(fullCD4fFoxp3datasize)
fullCD4fFoxp3inside2<-numeric(fullCD4fFoxp3datasize)

for(p in 1: fullCD4fFoxp3datasize){
x<- fullCD4fFoxp3foxp3[p]
z<- fullCD4fFoxp3cd25[p]
  u <- cos(alpha1)*(x-x01) + sin(alpha1)*(z-z01)
v <- -sin(alpha1)*(x-x01) + cos(alpha1)*(z-z01)
in1<-(u)^2/(a21)+(v)^2/(c21)
  if(in1<r1){
fullCD4fFoxp3inside1[p]<-1
  }else{
fullCD4fFoxp3inside1[p]<-0
  }
}
fullCD4fFoxp3propinside1<-sum(fullCD4fFoxp3inside1)/
fullCD4fFoxp3datasize
fullCD4fFoxp3propinside1
for(q in 1:fullCD4fFoxp3datasize)
  x<- fullCD4fFoxp3foxp3[q]
  z<- fullCD4fFoxp3cd25[q]
  u <- cos(alpha2)*(x-x02) + sin(alpha2)*(z-z02)
  v <- -sin(alpha2)*(x-x02) + cos(alpha2)*(z-z02)
  in2 <- (u)^2/(a22) + (v)^2/(c22)
  if(in2< r2)
    fullCD4fFoxp3inside2[q]<-1
  else
    fullCD4fFoxp3inside2[q]<-0
}

fullCD4fFoxp3propinside2<-sum(fullCD4fFoxp3inside2)/fullCD4fFoxp3datasize
fullCD4fFoxp3inside2

testboth<-numeric(datasize)
testcluster1<-numeric(datasize)
testcluster2<-numeric(datasize)
testneither<-numeric(datasize)

for(r in 1:datasize)
  if((inside1[r]==1) & (inside2[r]!=1)) testcluster1[r]<-1
  if((inside1[r]!=1) & (inside2[r]==1)) testcluster2[r]<-1
  if((inside1[r]==1) & (inside2[r]==1)) testboth[r]<-1
  if((inside1[r]!=1) & (inside2[r]!=1)) testneither[r]<-1

sum(testcluster1)
sum(testcluster2)
sum(testboth)
sum(testneither)

controlboth<-numeric(fullCD4fFoxp3datasize)
controlcluster1<-numeric(fullCD4fFoxp3datasize)
controlcluster2<-numeric(fullCD4fFoxp3datasize)
controlneither<-numeric(fullCD4fFoxp3datasize)
for(r in 1: fullCD4fFoxp3datasize){
  if((fullCD4fFoxp3inside1[r]==1) & (fullCD4fFoxp3inside2[r]!=1))
    controlcluster1[r]<-1
  if((fullCD4fFoxp3inside1[r]!=1) & (fullCD4fFoxp3inside2[r]==1))
    controlcluster2[r]<-1
  if((fullCD4fFoxp3inside1[r]==1) & (fullCD4fFoxp3inside2[r]==1))
    controlboth[r]<-1
  if((fullCD4fFoxp3inside1[r]!=1) & (fullCD4fFoxp3inside2[r]!=1))
    controlneither[r]<-1
}
sum(controlcluster1)
sum(controlcluster2)
sum(controlboth)
sum(controlneither)

categorisedc1 <- matrix(cbind(clustered, testcluster1), nrow(clustered))
categorisedc2 <- matrix(cbind(clustered, testcluster2), nrow(clustered))
categorisedboth <- matrix(cbind(clustered, testboth), nrow(clustered))
categorisedneither <- matrix(cbind(clustered, testneither), nrow(clustered))
datac1<-subset(categorisedc1, categorisedc1[,5]==1)
datac2<-subset(categorisedc2, categorisedc2[,5]==1)
databoth<-subset(categorisedboth, categorisedboth[,5]==1)
dataneither<-subset(categorisedneither, categorisedneither[,5]==1)

239
controlcategorisedc1 <- matrix(cbind(fullCD4fFoxp3, controlcluster1), nrow(fullCD4fFoxp3))
controlcategorisedc2 <- matrix(cbind(fullCD4fFoxp3, controlcluster2), nrow(fullCD4fFoxp3))
controlcategorisedboth <- matrix(cbind(fullCD4fFoxp3, controlboth), nrow(fullCD4fFoxp3))
controlcategorisedneither <- matrix(cbind(fullCD4fFoxp3, controlneither), nrow(fullCD4fFoxp3))

controldatac1<-subset(controlcategorisedc1, controlcategorisedc1[,4]==1)
controldatac2<-subset(controlcategorisedc2, controlcategorisedc2[,4]==1)
controldataboth<-subset(controlcategorisedboth, controlcategorisedboth[,4]==1)
controldataneither<-subset(controlcategorisedneither, controlcategorisedneither[,4]==1)

testresults<-c(sum(testcluster1),sum(testcluster2), sum(testboth),sum(testneither))
controlresults<-c(sum(controlcluster1),sum(controlcluster2), sum(controlboth),sum(controlneither))
Example 11: Adapting the distcritmulti method from the fpc+ package [121] to determine cluster validity using a variety of common indices

distcritmultiple<-function(x, clustering, part = NULL, ns = 10, fun = "dist", metric = "euclidean", count = FALSE, seed = NULL, ...)
{
  if (!is.null(seed))
    set.seed(seed)
  n <- length(clustering)
  if (is.null(part)) {
    pn1 <- n%/%ns
    pn2 <- n%%ns
    part <- rep(pn1, ns)
    part[ns] <- part[ns] + pn2
  }
  np <- sum(part)
  ns <- length(part)
  n <- length(clustering)
  npsam <- sample(n, np)
  cp <- cumsum(part)
  ss <- list()
  ss[[1]] <- npsam[1:cp[1]]
  asw <- numeric(0)
  pg <- numeric(0)
  e <- numeric(0)
  ch <- numeric(0)
  for (i in 2:ns) ss[[i]] <- npsam[(cp[i - 1] + 1):cp[i]]
  for (i in 1:ns) {
    if (count)
      cat("Subset ", i, "\n")
    if (fun == "dist")
      dx <- dist(x[ss[[i]], ], method = metric, ...)
    else dx <- daisy(x[ss[[i]], ], metric = metric, ...)
  }
}
validitySS <- cluster.stats(dx, clustering[ss[[i]]],
   alt.clustering=NULL, silhouette = TRUE, G2 = FALSE,
   G3 = FALSE, wgap = TRUE, sepindex = TRUE, sepprob = 0.1,
   sepwithnoise = TRUE, compareonly = FALSE, aggregateonly = FALSE)

asw[i] <- validitySS$avg.silwidth
pg[i] <- validitySS$pearsongamma
e[i] <- validitySS$entropy
ch[i] <- validitySS$ch

}  

aasw <- sum(part*asw)/np
sdasw <- sd(asw)
apg <- sum(part*pg)/np
sdpdg <- sd(pg)
ae <- sum(part*e)/np
sde <- sd(e)
ach <- sum(part*ch)/np
sdch <- sd(ch)

out <- list(a.silwidth = aasw, sd.silwidth = sdasw,
   a.pearsongamma = apg, sd.pearsongamma = sdpdg,
   a.entropy = ae, sd.entropy = sde, a.ch = ach, sd.ch = sdch,
   subsets = ss)

out
Example 12: Writing the clustering results to a .csv file and plotting a scatter plot

```r
resfile = paste("20130409",donorID,"SamSPECTRALres.csv")
write.table(resdf, resfile, sep="", append = TRUE,
row.names=FALSE, col.names = FALSE)

plotname<-paste("SamSPECTRAL Validity (",d,"),",donorID,".tiff",
sep="")
mypath=file.path("C:\", "Users", "mrxsf",
"SamSPECTRALautofigures", plotname)
mytitle = paste("SamSPECTRAL Output (",d,"): ",donorID)
tiff(file=mypath, width = 2000, height = 1000, pointsize = 36,
compression = "lzw")
par(mfrow=c(1,2))
plot(randomisedtest[,2], randomisedtest[,3],pch='.',
main = mytitle, xlab="CD127", ylab="CD25", col= P,
xlim=c(0, 4), ylim=c(0,4))
legend(x="topleft",as.character(round(frequency.large,3)),
col=names(frequency.large),pch=19)
plot(randomisedtest[,1], randomisedtest[,3],pch='.',
xlab="Foxp3", ylab="CD25", col= P,
xlim=c(0, 4), ylim=c(0,4))
legend(x="topleft",as.character(round(frequency.large,3)),
col=names(frequency.large),pch=19)
dev.off()
```

Example 13: Calculating cluster validity indices on a small subset of the data clustered and writing the contents of the general and cluster-specific validity variables to a .csv file. distcritmulti()+ can be substituted with distcritmultiple()+, as previously described, for a greater variety of validity indices

```r
validityidx<-distcritmulti(clusteredtest[,1:2], clusteredtest[,3],
part=NULL, ns=30, criterion="asw", fun="dist", metric="euclidean",
count=FALSE, seed=1)
```
Example 14: Writing summary variables to .csv files

dfrm <- data.frame(donorID, s, maxP, lengthtest, lengthNA, endtime, validityidx$crit.overall, validityidx$crit.sd)
write.table(dfrm, file = "20130807SamSPECTRALresult.csv", sep = ",", append = TRUE, row.names = FALSE, col.names = FALSE)

Example 15: Code to reproduce synthetic data for varying the distance between clusters. The first element of the $\mu$ vector for $\text{gaussian.2}$ is varied

gaussian.1 <- mvrnorm(n=40000, mu=c(1,4), Sigma=matrix(c(0.08,0,0,0.9),2,2))
gaussian.2 <- mvrnorm(n=2000, mu=c(1.75,6), Sigma=matrix(c(0.07,0,0,0.3),2,2))
fulltest <- rbind(gaussian.1, gaussian.2)
set.seed(1)
synthetic.data <- fulltest[order(runif(nrow(fulltest))),]

Example 16: Code to reproduce synthetic data varying the density ratio between clusters. The $n$ terms for both Gaussian distributions is varied

gaussian.1 <- mvrnorm(n=37500, mu=c(1,4), Sigma=matrix(c(0.08,0,0,0.9),2,2))
gaussian.2 <- mvrnorm(n=2500, mu=c(1.75,6), Sigma=matrix(c(0.07,0,0,0.3),2,2))
fulltest <- rbind(gaussian.1, gaussian.2)
set.seed(1)
synthetic.data <- fulltest[order(runif(nrow(fulltest))),]
Example 17: Code to reproduce synthetic data varying the signal:noise ratio between clusters. The uniform.num and the n of the second gaussian.big reference

```r
gaussian.big <- mvrnorm(n=40000, mu=c(5,5),
Sigma=matrix(c(2,0,0,2),2,2))
gaussian.small.1 <- mvrnorm(n=300, mu=c(9,10),
Sigma=matrix(c(0.08,0,0,0.08),2,2))
gaussian.small.2 <- mvrnorm(n=300, mu=c(1,10),
Sigma=matrix(c(0.07,0,0,0.07),2,2))
gaussian.small.3 <- mvrnorm(n=300, mu=c(2,0),
Sigma=matrix(c(0.5,0,0,0.5),2,2))
gaussian.small.4 <- mvrnorm(n=300, mu=c(10,1),
Sigma=matrix(c(0.1,0,0,0.1),2,2))
gaussian.all <- rbind(gaussian.big,gaussian.small.1,gaussian.small.2,
gaussian.small.3,gaussian.small.4)
xlim=c(min(gaussian.all[,1]),max(gaussian.all[,1]))
ylim=c(min(gaussian.all[,2]),max(gaussian.all[,2]))

uniform.num <- 5000
x.uniform <- (runif(uniform.num)*(xlim[2]-xlim[1]))+xlim[1]
y.uniform <- (runif(uniform.num)*(ylim[2]-ylim[1]))+ylim[1]
unifform.noise <- cbind(x.uniform, y.uniform)
gaussian.big <- mvrnorm(n=35000, mu=c(5,5),
Sigma=matrix(c(2,0,0,2),2,2))
fulltest <- rbind(gaussian.big, unifform.noise)
```

245
List of Figures

1.1 Schematic of CD4$^+$ T cell subsets at the centre of the innate and adaptive antitumour immune responses. ................. 6
1.2 Schematic of the relationship between CD4$^+$ T cell subsets. ...... 7
1.3 Backgating analysis of FlowJo’s “Population Comparison” tool (multivariate) on an example test nTreg dataset at two chi squared lower gating cut-off values. ........................................ 43
1.4 Kaluza® radar plot for 3-dimensional CD4$^+$ lymphocyte gating. 45
1.5 Infinicyt™ multidimensional data visualisation. ................. 46
2.1 Absolute numbers of CD4$^+$ lymphocytes (/µl whole blood) in donors of different ages. ................................. 63
2.2 Gating strategy to enumerate CD4$^+$ CD127$^{low}$ CD25$^{high}$ Foxp3$^+$ events in whole blood. ............................. 66
2.3 Gating strategy to enumerate Th17 cells in whole blood. ....... 67
2.4 Gating strategy to enumerate Th1, iTreg and IL-10$^+$ Th1 cells in whole blood. .............................................. 68
2.5 Examples of complete versus incomplete CD4 downregulation. 70
2.6 Analysis of samples with partial versus complete CD4 downregulation. ....................................................... 71
2.7 Synthetic datasets generated to dissect algorithm performance. 75
2.8 SamSPECTRAL method on synthetic data. ......................... 77
2.9 SamSPECTRAL method on nTreg data. ............................. 78
2.10 Example SamSPECTRAL solutions. ................................ 80
3.1 Time course of IFN$\gamma$ secretion after stimulation. ............. 84
3.2 Time course of IL-17A secretion by whole blood cultures after stimulation. ..................................................... 85
3.3 Time course of IL-10 secretion by whole blood cultures after stimulation. ........................................... 86
3.4 Scatter characteristics of leukocyte lineages in permeabilised whole blood. ........................................ 88
3.5 Optimisation of AF488-conjugated anti-Foxp3 on whole blood. ......................................................... 89
3.6 Anti-RORγ(t) and anti-CD4 staining. ........................................................................................................... 90
3.7 RORγ(t)+ events in whole blood were not CD8+, CD19+, CD16+ or CD14+. ........................................... 91
3.8 Foxp3+Helios+ events in whole blood were CD25+. .................................................................................. 93
3.9 Expression of Foxp3 in different CD25 and CD127 high and low subgroups. ........................................ 94
3.10 Histograms to demonstrate flow cytometry data positive skew. ........................................................... 96
3.11 Histograms to demonstrate ELISA data positive skew. .......................................................................... 97
3.12 nTreg staining examples. .......................................................................................................................... 98
3.13 IL-10 staining examples. ........................................................................................................................... 99
3.14 IL-17 staining examples. ............................................................................................................................ 100
3.15 IFNγ staining examples. ............................................................................................................................ 101
3.16 Whole blood IFNγ abundance for each donor. ......................................................................................... 103
3.17 Standard curves of known IFNγ concentrations for each ELISA. ......................................................... 104
3.18 Whole blood IL-10 abundance for each donor. ......................................................................................... 105
3.19 Standard curves of known IL-10 concentrations for each ELISA. ....................................................... 106
3.20 Whole blood IL-17 abundance for each donor. ......................................................................................... 107
3.21 Standard curves of known IL-17 concentrations for each ELISA. ....................................................... 108
4.1 Alterations in T regulatory cell populations with age. ................. ......... ............................... 113
4.2 Th17 cells do not change with age. ............................................................................................................ 115
4.3 Th17:nTreg balance decreases with age. ................................................................................................. 116
4.4 The increase in IL-10 is independent of age-related induction of IL-10 expression by Th1 cells. .......... 118
4.5 Alterations in proinflammatory T cell populations with age. .................................................................... 119
4.6 Absolute numbers of anti-inflammatory CD4+ lymphocytes subsets (/µl whole blood) in donors of different ages. ............................................................................................................ 121
4.7 Absolute numbers of proinflammatory CD4+ lymphocytes subsets (/µl whole blood) in donors of different ages. ............................................................................................................ 122
4.8 Correlation between flow cytometry phenotype and ELISA functional cytokine measurements. ........ 124
4.9 Age stratified ELISA data correlation analysis. ..................................................................................... 125
4.10 Age stratified flow cytometry data correlation analysis. ......................................................................... 126
4.11 Age stratified flow cytometry data correlation analysis (part 2). .......................................................... 127
4.12 SimpleCART with 73-fold cross-validation on the full dataset of 73 donors. .................................................. 129
4.13 SimpleCART with 57-fold cross-validation on the part dataset of 57 donors. .................................................. 130
4.14 Assessment of correlation of measurements with assay number. 132
4.15 Assessment of correlation of measurements with time blood sample extracted. .................................................. 133
4.16 Assessment of correlation of between age of donor with assay number and time blood sample extracted. .................. 134

5.1 Different measurements of nTreg give different conclusions about whether their frequency changes with age. ................. 140
5.2 Gating strategy to enumerate CD4\textsuperscript{+}CD127\textsuperscript{low}CD25\textsuperscript{high}Foxp3\textsuperscript{+} events in whole blood. .................................................. 141
5.3 Intra- and inter-experimenter variation in manual gating analysis. 142
5.4 Correlation between cluster validity indices on nTreg data. ..... 145
5.5 Specified versus resultant cluster number following SamSPECTRAL clustering of nTreg data. .................................................. 146
5.6 ASW on different SamSPECTRAL-clustered nTreg solution types. 146
5.7 Categorisation of SamSPECTRAL nTreg clustering solutions by three independent experts. .................................................. 147
5.8 k-means versus SamSPECTRAL on distance data. ................. 151
5.9 k-means versus SamSPECTRAL on density data. .................. 152
5.10 k-means versus SamSPECTRAL on signal:noise data. .......... 153
5.11 Example k-means clustering of synthetic datasets. ............... 154
5.12 Example SamSPECTRAL clustering of synthetic datasets. ..... 155
5.13 Example of spatial separation of events within one cluster. .... 156
5.14 Comparison of ASW and F-measure in synthetic experiments. .. 158
5.15 SamSPECTRAL parameter optimisation experiments on synthetic data fixing k. .................................................. 161
5.16 Example SamSPECTRAL parameter optimisation on density synthetic datasets. .................................................. 162
5.17 Example SamSPECTRAL parameter optimisation on signal:noise synthetic datasets. .................................................. 163
5.18 F-measures for parameter optimisation in synthetic datasets. .. 165
5.19 Example k-means clustering of nTreg datasets at different specified values of k. .................................................. 167
5.20 Example SamSPECTRAL clustering of nTreg datasets at different specified values of k (1). ................................. 168
5.21 Example SamSPECTRAL clustering of nTreg datasets at different specified values of $k$ (2). 169
5.22 SamSPECTRAL parameter optimisation experiments on nTreg data. 172
5.23 D491368: example parameter optimisation on an nTreg dataset. 174
5.24 The variation in nTreg frequency from between “optimal” SamSPECTRAL-generated solutions. 176
5.25 D383401: Manual gating compared to SamSPECTRAL clustering. 178
5.26 Ellipses fitted to SamSPECTRAL-clustered data shown on all test and control samples. 180
5.27 nTreg frequency with age determined by SamSPECTRAL clustering with elliptical gating. 182
List of Tables

1.1 Comparison of nTreg ageing studies for subject selection and nTreg markers. ............................................. 20
1.2 Comparison of conclusions for number of nTreg throughout life. 21
1.3 Comparison of conclusions for phenotype of nTreg throughout life. 23
1.4 Comparison of conclusions for function of nTreg throughout life. 24

2.1 Staining panel. ............................................................. 62
2.2 List of monoclonal antibodies used. ............................... 65
2.3 Typical pre-acquisition compensation required. ................. 65
2.4 Events per small population of interest. .......................... 72
2.5 Mean percentage CVs from 3 donors stained in triplicate. .... 72

4.1 Demographic details of sample population. ....................... 111
4.2 Medical details of sample population. ........................... 112
Index

k-means, 34, 76, 143, 148
Abbreviations, v
Absolute frequency, 61, 120
Abstract, i
Acknowledgements, iii
Adaptive immunity, 4
Ageing, 3
Agent-based modelling, 29
Aims of thesis, 1
Antibody titrations, 89
Artificial neural networks, 33
Assumptions in laboratory experimentation, 187
Automated data exploration, 37
Automated gating, 40
Automated nTreg analysis, 38, 41
Average silhouette width, 53
Balance, 11, 114
Baltcheva's regulatory T cell model, 31
Bibliography, 198
Blind signal separation, 34
Cachexia, 193
Calinski and Harabasz index, 55
CART, 32
CD127, 9, 92
CD25, 9
CD4 downregulation, 69
Cluster validity indices, 35, 52, 79, 144
Clustering, 33
Clustering F-measure, 55, 156, 164, 177
Clustering result categorisation, 79
Computational methods, 28
Confounding factors, 74, 128
Constrained spectral clustering, 194
Contents, vii
Correlation analyses, 123
Data clusterability, 52
Data mining, 32
Dedication, ii
distcritmulti, 57, 81
distcritmultiple, 80
Donor characteristics, 110
Donor demographics, 59
Donor medical histories, 59
ELISA, 73
Entropy, 54
Error report, 81
Ethics approval, 59
Exclusion criteria, 59
Fitting minimum volume spanning ellipses, 81, 179
Fix/Perm buffer selection, 87
FLAME, 38
Flow cytometry, 35
Flow cytometry acquisition, 64
Flow cytometry analysis, 64
Flow cytometry data clustering, 37
Flow cytometry staining panel, 190
FlowCAP, 50, 55
flowClust, 38
flowCore, 40, 76
flowMeans, 39
flowMerge, 38
flowScape, 191
flowType, 41
Foxp3, 9, 92
Future directions, 192
Fuzzy c-means, 34

Gating strategies, 64
GenePattern, 184

Helios, 90
Hierarchical clustering, 39
Hubert’s $\Gamma$ statistic, 54
Hypothesis, 1

IFN$\gamma$, 10, 25, 117
IL-10, 10, 117, 186
IL-17, 11, 117
Immunosenescence, 13
Immunotherapies, 5
Improving nTreg analysis, 138
Infinicyt$^{\text{TM}}$ automatic population separator, 43
iTreg, 10, 25, 102, 110

Kaluza® radar plot, 43

List of figures, 246
List of tables, 250

Manual flow cytometry data analysis, 35
Manual result categorisation, 147
MDSCs, 13, 193
Mean events acquired, 72
Multi-sample comparison, 179
multisample comparison, 47

nTreg, 9, 16, 97, 110
nTreg datasets, 76

Ordinary differential equations, 29, 30
Organisation of thesis, 2
Other regulatory subsets, 193

Parameter optimisation, 159
PMA and ionomycin, 25, 60, 187
Principal component analysis, 34, 47
Probability binning, 41

Qualitative modelling, 30
Questionnaire, 59

R, 76

R code vignette, 227
ROR$\gamma(t)$, 89
SamSPECTRAL, 38, 76, 143, 148, 189
Scatter characteristics of leukocyte lineages, 87
SimpleCART, 74, 128
Simulation, 28, 195
slanDCs, 193
SPADE, 40
SPICE, 48

Stability analysis, 175
Staining panel, 60
Statistical analyses, 73
Subject selection, 14, 59
Summary, 196
Supervised learning, 32, 74, 128
Support vector machines, 33
Synthetic datasets, 74

T cell ontogeny, 5
Th1, 10, 25, 102, 117
Th17, 11, 26, 89, 102, 114
Time course ELISAs, 83
Tolerogenic DCs, 193
Typical staining examples, 95

Unsupervised learning, 33
viSNE, 44
Visualisation methods, 43
Whole blood assay, 60, 189
Whole blood assay development, 83