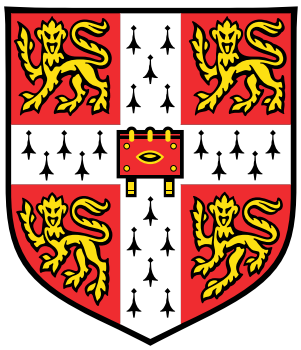
Normalisation and Clustering Methods Applied to Association Studies in Type 1 Diabetes



Nikolas Pontikos

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Homerton College January 2015

Pour ma petite Maman chérie, à qui je dois tout ...

Για Ηλία, Σοφία, Στέφανος και η Ναταλία ...

Zu meinem unerschrockene Bruder, Theo ...

Til min åndelige far, Constantine ...

To Laurent Cullinan who left before his time ... I only hope this dedication brings some joy to his family ...

Declaration

This thesis is submitted as part requirement for the PhD Degree in "Medical Genetics" at Cambridge University. It is substantially the result of my own work except where explicitly stated otherwise. The report may be freely copied and distributed provided the source is explicitly acknowledged.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text It does not exceed the prescribed word limit for the relevant Degree Committee.

Nikolas Pontikos

January 2015

Acknowledgements

I am very grateful for having been offered the great opportunity to undertake this PhD and to learn so much about statistics, genetics, and science in general. I am certain that these three years of intellectual development have truly reshaped the way I think and see the world and will undoubtedly bear great positive influence on my future life.

First and foremost, my eternal gratitude to my Mother for having devoted so much time to my upbringing and to have kindled my enthusiasm for science and adventure from a very young age. Although she is sadly no longer among us, she is the reason why I decided to dedicate my computer science and programming skills to scientific and medical research.

Secondly, I wish to thank all the great people I have had the privilege to know and work with during this academic adventure. In particular, my super-hero supervisor, Chris Wallace, for her efficiency, advice, support, and encouragement when I was feeling down, and her pragmatism, guidance, and helping me channel my enthusiasm productively when it was overflowing. I'd also like to thank my advisor, Anna Petrunkina Harrison, for providing much encouragement and support, and putting me in touch with the right people in flow cytometry. John Todd's intense presence and input have been incredibly valuable and I would like to thank him for his relentless drive, insight, experience and pushing me to publish.

Having no formal training in statistics, I was privileged enough to be surrounded by smart people throughout my PhD from whom I have learned a great deal. Notably, Xin

Yang, Hui Guo, Nick Cooper, Marina Evangelou and Mary Fortune spring to mind, with whom I have had many interesting lunchtime conversations. Mary Fortune's mathematical prowess is only rivalled by her baking skills. Xin Yang also makes delicious chinese buns, a skill that will no doubt continue to serve her well.

I also learned a lotl about biology thanks to discussions with Marcin Pekalski, Tony Cutler, Xaquin Dopico Castro, Ricardo Fereirra and Charlie Bell. They were also very important in making me feel welcome to the lab (perhaps even a little too welcome at times). I wish them all the best in their future endeavours and I am sure science or some other mysterious force will bring us back together.

The first chapter, I owe in great part to my second supervisor Anna Petunkina-Harrison for the introduction to flow cytometry and putting me in touch with the flow cytometry community. I should also mention that my first year examiner, Lorenz Wernisch, gave me some very constructive feedback on my first year report and food for thought.

The second chapter of my thesis would not have been possible without the previous work and extensive email communications with Vincent Plagnol, Calliope Dendrou and Linda Wicker. Notably, I would like to give special thanks to Vincent for offering me the opportunity of continuing to do what I love at University College London for the next two years. I am sure we will do some great work together.

The third chapter was enabled by Tony Cutler's data set, but also benefitted greatly from Marcin Pekalski's and Charlie Bell's detailed knowledge of immune cell types and flow cytometric artefacts.

The fourth chapter, which lead to a publication in BMC Genomics, was a continuation of the important work done by Helen Schuilenberg, Debbie Smyth, James Traherne and Jyothi, on characterising KIR copy number variation. This body of work is also proof to the testimony of the kind of great work can be done in a collaborative environment.

The success of this project rests entirely on the supervision of Chris Wallace, to which I am very grateful for the time and effort invested in the publication. I also received very meticulous feedback from John Todd and Jo Howson which greatly facilitated the submission procees.

The fifth and final chapter, was greatly aided by discussions with office discussion with Chris Wallace and Mary Fortune.

However, no PhD would be complete without a little bit of extra-curricular adventures which surely should be the object of another thesis. My thanks to my team mates Sriyal Mendis and Gary Low for helping me out on my autorickshaw Indian chapter. Their driving skills kept me alive and unscathed during our 4000km autorickshaw adventure across India, in spite of my unfortunate cowllision in Ongole .

Thanks to Ben Silva-Weatherley, Steve Sawiak, and the rest of the team at RNA Crossfit for making sure I didn't completely neglect my body through the daily PhD grind. I'd like to think I tried to stay true to the humanist ideal of a healthy mind in a healthy body "mens sana in corpore sans"

I would also like thank to all the friends, both professionals and students, I have made during these five fun years at Cambridge, first at the European Bioinformatics Institute, then at the University and Addenbrookes. Notably, my good friends, Shawn O'Donell and Pallawi Sinha, who have often fed me and generally been very supportive and patient with my poor time-keeping (something I'm still working one). Thanks also to all housemates of 386 Cherry Hinton Road, past and present, my home of five years. Thank you Cambridge for being such a friendly, unique, quircky and curious place, one which I have grown very fond of.

Lastly and most importantly, my everlasting gratitude to my Father and my Brother, inspirations in their own right, for their unconditional love, incessant support and for listening.

Abstract

Genetic association studies have discovered many variants which influence type 1 diabetes (T1D) risk and further correlate with quantitative cell-type specific phenotypes. However, disease associated differences can be small, and large numbers of samples are required to overcome the heterogeneity that exists between humans. Novel high-throughput biotechnologies measure large number of samples but technical or within-batch variation, may undermine reproducibility of measurements.

In my thesis, I analyse two types of these datasets, central to the study of T1D. The first is generated by flow cytometry, a biotechnology utilising light scatter and fluorescently stained markers to discriminate different cell types. Unfortunately, flow cytometry can be prone to batch effects since blood samples are often collected, prepared and analysed at different times and by different operators. I consider several normalisation techniques to address these issues, using external or within sample controls. The other major challenge in flow cytometry data analysis is that of identifying different cell types. While this is essentially a clustering problem, currently the most widely applied method is a manual approach which can be inefficient and biased. I investigate ways this process can be automated by fitting mixture models to emulate the manual process. I show that, in the absence of manual gates, data-driven approaches using regression trees can be applied to detect new cell subsets, not targeted by manual gating, that respond to IL-2 in an in vitro stimulation experiment.

The second type of dataset is generated by qPCR and genotyping arrays, which

are applied to DNA from T1D cases and controls to determine whether copy number variation in two Killer Immunoglobulin-like Receptors (KIRs) genes associates with T1D. I apply normalisation to correct for batch effects between qPCR plates and clustering using mixture models to identify copy number groups. Supervised clustering is then used to correlate qPCR copy number with SNP data, allowing for association testing in a twenty-fold larger sample size than ever previously considered for KIR genes.

Finally, I conclude with what I have learned from applying these methods and how these may be further developed, with special attention to flow cytometry where these remain under utilised. In particular, I discuss how normalisation and clustering relate, and how prior knowledge, when available, can be incorporated into the clustering scheme.

Contents

Contributors xxiv

Genes xxviii

SNPs xxix

Proteins xxx

1 Introduction **1**

1.1 Motivation 1

1.2 Biology of type 1 diabetes 3

1.2.1 Aetiology and diagnosis 3

1.2.2 Genetics of type 1 diabetes 5

1.2.3 The immune cell mechanisms 6

1.3 Studying the immune system with flow cytometry 7

1.4 Normalisation 17

1.5 Manual clustering 19

1.6 Automatic methods for identifying clusters 20

1.6.1 Model-based methods 20

1.6.2 Model-free methods 22

1.6.3 Estimating the optimal number of clusters from the data 28

2 The influence of gating on repeatability and effect size estimation **31**

2.1 Background 31

2.2 Univariate clustering of bead data 40
2.3 Univariate gating on CD25: defining a CD25$^+$ threshold on naive cells . . 45
2.4 Univariate gating on CD45RA: fitting two-component mixtures on non-regulatory T cells 49
2.4.1 Using the mixing proportions of the mixture model 49
2.4.2 Emulating manual gating by picking a threshold 51
2.5 Association tests 60
2.6 Discussion 64

3 Methods to assess cell response to ex-vivo stimulation in flow cytometry 67

3.1 Background 67
3.2 Cell Phenotypes Identified by Manual Analysis 72
3.3 Reproducibility of pSTAT5 response within an individual 82
3.3.1 Normalisation approaches 82
3.3.2 Repeatability 89
3.3.3 Association of pSTAT5 response with type 1 diabetes 92
3.4 Response in the whole sample 97
3.4.1 Visualising response in whole sample with Spanning-tree Progression of Density-normalised Events (SPADE) 99
3.4.2 Visualising response in whole sample with recursive partitioning . 115
3.4.3 Alternative visualisation using partial least squares 136
3.4.4 Identification of low-dose sensitive cells by recursively applying a two component mixture model on pSTAT5 145
3.5 Discussion 153
3.5.1 Association of pSTAT5 with T1D 153
3.5.2 Methods of identifying dose-responsive cell populations 154

4 KIR3DL1/KIR3DS1 copy number variation in type 1 diabetes 164

4.1 Background . . . 165

4.1.1 Killer immunoglobulin receptors and their interaction with human leukocyte antigen molecules . . . 165

4.1.2 *KIR3DL1* and *KIR3DS1*: two strong candidates for T1D association 166

4.2 Samples and genotyping assays . . . 169

4.2.1 Samples . . . 169

4.2.2 HLA and SNP Genotyping . . . 169

4.2.3 qPCR experimental protocol . . . 170

4.3 Data Analysis . . . 171

4.3.1 Quality control and normalisation of qPCR data . . . 171

4.3.2 Bivariate clustering: copy number calling in qPCR data . . . 173

4.3.3 k-nearest neigbour (KNN) classification: copy number imputation into the SNP data . . . 178

4.4 Association testing of *KIR3DL1/KIR3DS1* copy number with T1D . . . 184

4.5 Discussion . . . 189

4.5.1 Previous association studies of KIR genes with T1D . . . 189

4.5.2 My approach . . . 194

4.5.3 Future work . . . 195

5 Discussion 199

5.1 Important practical considerations with computational clustering methods . . . 199

5.2 Dealing with noise by using prior knowledge . . . 202

5.3 Identifying outliers . . . 205

5.4 Prioritising normalisation or clustering . . . 207

5.5 Influence of normalisation and clustering on statistical association test . 212

5.6 The future 215
5.7 Summary 219

Appendix A Flow markers **223**

List of Figures

1.1 **Flow cytometer diagram.** 13

1.2 **Leaking of signal from FITC fluorochrome into PE detector.** . . 13

1.3 **The same data encoded in FCS 2.0 (a) and FCS 3.0 (b).** 14

1.4 **Log transform inflates the variance around zero.** 14

1.5 **Effect of different values of w parameter of the Logicle transformation on the data distribution.** 15

1.6 **Normalisation by peak alignment.** 18

1.7 **Effect of shrinking variance in a two compnent Gaussian mixture model (GMM) on the decision boundary.** 27

1.8 **Variance ratio criterion used to estimate the number of clusters with K-means.** 30

2.1 **Number of samples analysed per day.** 37

2.2 **Manual gating strategy followed by Dendrou et al (2009b).** . . 38

2.3 **Effect of time on CD25 MFI in memory cells.** 39

2.4 **Linear regression of bead APC MEF against the APC MFI as defined in Table 2.5.** 42

2.5 **Comparison of bead population MFI using manual and `flowBeads` gating.** 43

2.6 **Bead normalisation partially corrects for long term time effect in CD25 MFI of the memory cell population.** 44

2.7 **Position of the CD25$^+$ gate over duration of experiment.** 46

2.8 **Mean square difference (MSD) of the position of the manual gate with that of `beads.thresh`.** 47

2.9 **Agreement with manual of percentage of CD25$^+$ naive cell phenotype.** 48

2.10 **Repeatability of percentage of naive CD25$^+$ with manual (black) and `beads.thresh` (red).** 48

2.11 **Agreement with manual of the percent memory cell phenotype obtained with `mm` (a) and `spmmm` (b).** 50

2.12 **Repeatability of the percent memory cell phenotype with manual (black), `mm` (red) and `spmm` (blue).** 51

2.13 **Example on individual d of the two approaches, `pct.thresh` (a and c) and `post.thresh` (b and d), of selecting a threshold.** . . . 54

2.14 **Mean square difference (MSD) of the position of the manual gate with that of `pct.thresh` (a) and `post.thresh` (b).** 55

2.15 **Samples for which the 99 maximum posterior probability is not reached.** 56

2.16 **Agreement of memory CD25 MEF (a and b) and percentage of memory cells (c and d), obtained from `pct.thresh` and `post.thresh` with manual.** 57

2.17 **Influence of threshold for `pct.thresh` and `post.thresh` on distribution of memory CD25 MEF and percent memory cell phenotypes.** 58

2.18 **Repeatability of the memory cells phenotypes, CD25 MEF (a) and percentage (b), obtained from manual gating (black), `pct.thresh` (red) and `post.thresh` (blue).** 59

2.19 **Effect of rs2104286 on percent memory gated by manual (black), `post.thresh` (green) and `pct.thresh` (red).** 62

3.1 **Schematic of major IL-2 signaling pathway (taken from Liao et al (2013)).** 68

3.2 **Number of cases and controls analysed per day.** 71

3.3 **Gates applied across doses.** 72

3.4 **Distribution pSTAT5 in the manually gated cell subsets in the sample manually gated in Figure 3.3.** 73

3.5 **pSTAT5 distribution in an individual on visit one (black) and visit two (red) clearly shows that pSTAT5 distribution is not stable across days in the four cell subsets.** 74

3.6 **The percent of pSTAT5$^+$ cells increases with proleukin dose in memory Teffs, but the measured response is not consistently repeatable (f, g).** 76

3.7 **The percent of pSTAT5$^+$ cells increases with proleukin dose in memory tregs.** 77

3.8 **The percent pSTAT5$^+$ cells increases with proleukin dose in naive Teffs.** 78

3.9 **The percent pSTAT5$^+$ cells increases with proleukin dose in naive tregs.** 79

3.10 **Repeatability in six individuals** 80

3.11 **Association test of percent pSTAT5$^+$ in four cell subsets.** . . . 81

3.12 **Variation in sample MFI is not captured by variation in bead MFI.** 83

3.13 **pSTAT5 MFI (black), background subtracted pSTAT5 MFI (red)** 84

3.14 **pSTAT5 intensity across the four proleukin doses, before (a) and after (b) per-cell baseline pSTAT5 subtraction in the ungated sample.** 86

3.15 **Repeatability of pSTAT5 MFI in the nearest-neighbour joined (black), nearest-neighbour background subtracted (red).** 87

3.16 **Repeatability of percent pSTAT5$^+$ in the individual samples (black), nearest-neighbour joined (red) and nearest-neighbour joined samples baseline corrected (blue).** 88

3.17 **Repeatability of pSTAT5 MFI measured as Pearson correlation squared (r^2) per dose per cell type.** 90

3.18 **Repeatability of the percent of pSTAT5$^+$ as Pearson correlation squared (r^2) per dose per cell type.** 91

3.19 **Association test of pSTAT5 MFI with T1D.** 93

3.20 **Association test of pSTAT5 MFI, after nearest-neighbour normalisation, with T1D.** 94

3.21 **Association test of percent pSTAT5$^+$ with T1D.** 95

3.22 **Association test of percent pSTAT5$^+$, after nearest-neighbour normalisation, with T1D.** 96

3.23 **Gates applied across doses.** 97

3.24 **minimum spanning tree (MST) generated by applying Spanning-tree Progression of Density-normalised Events (SPADE) on lymphocytes. MST nodes are coloured by pSTAT5 mean fluorescence intensity (MFI).** 103

3.25 **(a) Mapping of cell types defined by manual gates, memory Teffs (black), memory Tregs (red), naive Teffs (green) and naive Tregs (blue), to the MST obtained in Figure 3.24. (b) Manually identified subset of cells (purple) which respond to 1000 units but lie far from the other manually gated cell subsets.** 104

3.26 **A 1000 unit responsive cell subset (purple) within lymphocytes is identified which is not assigned to any manual gate.** 105

3.27 **The pSTAT5 MFI dose-response of the different cell subsets within lymphocytes.** 106

3.28 **pSTAT5 MFI coloured MST generated by applying SPADE on cells which fall outside of the lymphocyte gate.** 109

3.29 **The three cell subsets, blue, purple and pink, manually identified in the MST of Figure 3.28 are mapped back to scatter coordinates. The 10 unit responsive groups, blue and purple points in (a), and the 1000 units responsive group, pink points in (b), generally tend to lie close to the lymphocyte cluster (black ellipse), but a potential secondary scatter cluster of 1000 unit responsive cells (delineated by pink polygon) are worthy of further investigation.** 110

3.30 **Pink cells in relation to lymphocyte cluster on all markers.** . . 111

3.31 **Dose-response** 112

3.32 **The progression of the marker MFI along the horizontal coordinate of the MST nodes in lymphocytes (a) and non-lymphocytes (b).** 114

3.33 **Sample recursively partitioned into 128 bins on side and forward scatter.** 116

3.34 **Each sample is recursively partitioned using 128 bins.** 117

3.35 **pSTAT5 response in sample recursively partitioned into 128 bins on side and forward scatter.** 120

3.36 **MST built using on the 1024 bins obtained from recursive partitioning on the lymphocytes core markers.** 121

3.37 **A 10 unit responsive cell subset (pink) and a 1000 unit responsive cell subset (purple) within lymphocytes are identified which are not assigned to any manual gate.** 122

3.38 **Dose response** 123

3.39 **MST built on the 1024 bins obtained from rpart on non-lymphocytes, a cluster of cells stands out from the rest which shows pSTAT5 response at 1000 units.** 124

3.40 **A 1000 unit responsive cell subset (purple) is identified which does not belong to the manually define lymphocytes population (black).** 125

3.41 **A 1000 unit responsive cell subset (purple) is identified which does not belong to the manually define lymphocytes population (black).** 126

3.42 **A 1000 unit responsive cell subset (purple) is identified which does not belong to the manually define lymphocytes population (black).** 127

3.43 **classification and regression tree (CART) of 1000 unit response against side and forward scatter identifies three subsets.** 132

3.44 **The recursive partitioning tree obtained for the pSTAT5 at 0.1 units (a), 10 units (b) and 1000 units (c) in the lymphocytes which do not belong to any manually identified cell subset.** . . 133

3.45 **The recursive partitioning tree obtained for the pSTAT5 at 0.1 units (a), 10 units (b) and 1000 units (c) in the non-lymphocyte cells.** 134

3.46 **Partition tree obtained in two different sample from running CART on all core markers against pSTAT5 response at 1000U.** 135

3.47 **First two components of partial least squares (PLS) projection. Clusters newly identified using PLS in relation to known manually gated ones within lymphocytes.** 138

3.48 **The newly identified cluster using PLS in Figure 3.47 (purple) in relation to known manually gated ones.** 139

3.49 **Dose response.** 140

3.50 **First two components of PLS projection. Clusters newly identified using PLS in relation to known manually gated ones within non-lymphocytes.** 142

3.51 **First two components of PLS projection. Clusters newly identified using PLS in relation to known manually gated ones within non-lymphocytes.** 143

3.52 **Dose response.** 144

3.53 **Recursive partitioning of pSTAT5 response into low (red) and high (green) populations to identify cells responsive to the lowest dose of proleukin.** 146

3.54 **Clusters identified from recursive partitioning.** 147
3.55 **Clusters identified from recursive partitioning.** 148
3.56 **Recursive partitioning of pSTAT5 response into low (red) and high (green) populations to identify cells responsive to the lowest dose of proleukin.** 150
3.57 **Clusters identified from recursive partitioning.** 151
3.58 **Dose response** 152
3.59 **Dose response in lymphocytes.** 158
3.60 **Identified cell subset in lymphocytes.** 159
3.61 **Dose response in non-lymphocytes.** 160
3.62 **Identified cell subset in non-lymphocytes.** 161
3.63 **Identified cell subset consensus in non-lymphocytes.** 162

4.1 ***KIR3DS1*** **and** ***KIR3DL1*** **ΔCt values for cases (red) and controls (blue) per qPCR plate.** 173
4.2 **Copy number calling of** ***KIR3DL1/KIR3DS1*** **from qPCR ΔCt.** 175
4.3 **Post-QC cases (red) and controls (blue) are plotted separately for each qPCR plate.** 176
4.4 **Repeatability of qPCR assay.** 177
4.5 **Overlay of ImmunoChip and qPCR samples for R and θ at SNP rs592645.** 178
4.6 **Signal plots of ImmunoChip SNPs which fall in the** ***KIR3DL1*** **region.** 180
4.7 **Leave-one-out crossvalidation error rate for k-nearest neighbour (KNN) prediction.** 181
4.8 **Error rate of k-nearest neighbour prediction from R and θ of SNP rs592645 in random subset of samples.** 183

4.9 **Flow chart summarising the key steps involved in the KNN imputation of *KIR3DL1/KIR3DS1* copy number in SNP data (R and θ signal) from qPCR copy number predictions obtained from GMM clustering.** 198

List of Tables

1.1 **Spillover matrix of the fluorochomes used by Dendrou et al (2009b) obtained using single colour beads.** 10

2.1 **Number of days till second visit for recalled individuals.** 32

2.2 **Distribution of subjects in study, by genotype, age and sex.** . . 33

2.3 **The fluorochrome-antibody panels with six markers used in the IL2RA dataset.** 33

2.4 **Repeatability and significance of effects of percentage of naive CD25$^+$ CD25 MEF and percentage of memory cell phenotypes.** 34

2.5 **FluoroSpheres from DakoCytomation.** 41

2.6 **Genotype, age and sex effect sizes on percentage of CD25$^+$ cells.** 61

2.7 **Memory CD25 MEF effect sizes.** 62

2.8 **Memory percentage effect sizes.** 63

3.1 **Ten individuals recalled between 98 and 168 days later to assess stability of the cell phenotypes.** 70

3.2 **Proleukin stimulation assay antibody-fluorochrome panels.** . . . 71

3.3 **Cell phenotypes in lymphocytes.** 158

3.4 **Cell phenotypes, non-lymphocytes.** 160

4.1 **Grouping of HLA alleles by HLA-Bw4 epitope.** 168

4.2 **Classification of subjects in study by HLA epitope (as defined in Table 4.1).** 169
4.3 **The qPCR probes and primers.** 171
4.4 **ImmunoChip SNPs which fall in *KIR3DL1*.** 182
4.5 **Association test of *KIR3DS1-KIR3DL1* copy number with T1D.** 186
4.6 **Association test of *KIR3DS1-KIR3DL1* copy number with T1D, conditional on HLA-Bw4.** 187
4.7 **Case-only χ^2 test for interaction between *KIR3DS1-KIR3DL1* copy number and HLA-Bw4, across the ten multiply imputed qPCR and SNP datasets.** 188
4.8 **Proportion of cases to controls in all known killer immunoglobulin-like receptor (KIR) studies in T1D.** 191
4.9 **Case-control ratio in all known KIR studies in T1D.** 192

Contributors

Calliope Dendrou 45

Charlie Bell 119

Deborah Smyth 170

James Traherne 195

Linda Wicker 51

Marcin Pekalski 98, 149

Tony Cutler . 69, 72, 74, 75, 97, 149, 153

Abbreviations

1958BC British 1958 Birth Cohort 169

AIC Akaike Information Criterion 29

AML actute myeloid leukemia 115

ANN approximate-nearest-neighbour 85, 128, 200

ANOVA analysis of variance 29, 184

BIC Bayesian Information Criterion 29

CART classification and regression tree 128, 129, 132, 145

CN copy number 198

D-GAP Diabetes - Genes, Autoimmunity and Prevention 69

EM expectation maximisation 22, 49, 64, 174

FCS Flow Cytometry Standard 10

FlowCAP Flow Cytometry Critical Assessment of Population Identification Methods 20

GMM Gaussian mixture model 20, 21, 23, 27, 29, 145, 198

GRID Genetic Resource Investigating Diabetes 169

GWAS genome-wide association study 1, 5, 166

HIP Human Immunology Project 16

HLA human leukocyte antigen 5, 165, 189

HWE Hardy-Weinberg equilibrium 190, 196

IL-2 interleukin-2 ..... 4, 7

ITIM immune tyrosine-based inhibitory motif 165

KIR killer immunoglobulin-like receptor 165, 166, 191, 192

KNN k-nearest neigbour 178, 181, 201, 202

LD linkage disequilibrium 194, 196

LOOCV leave-one-out cross-validation 179, 198, 202

MARS multivariate adaptive regression splines 131

MDS multidimensional scaling 99, 156

MEF molecules of equivalent fluorochrome 35, 40

MFI mean fluorescence intensity 35, 40, 74, 103, 213

MISE mean integrated square error 66

MSD mean square difference 65

MST minimum spanning tree ... 100, 101, 103, 104, 109, 110, 113–115, 118, 128, 130, 136, 155, 156

NK natural killer 102, 165

PBMC peripheral blood mononuclear cells 98, 153, 157

PCA principal component analysis 99, 113, 136, 169

PLS partial least squares 113, 136–139, 141–143

PRIM Patient Rule Induction Method 131

qPCR quantitative Polymerase Chain Reaction 1, 3, 170, 171

RF random forests 196

SNP single nucleotide polymorphism 1–3, 5, 31, 166, 189

SPADE Spanning-tree Progression of Density-normalised Events ... 99–101, 103, 107, 109, 115, 128, 215

T1D type 1 diabetes 3–5, 67, 68, 70, 165, 189, 192

Treg regulatory T cell 68

Genes

CTLA4 5

FOXP3 68

HLA-A 168
HLA-B 168
HLA-DQ2 193
HLA-DQ8 193
HLA-DQB1 6
HLA-DRB1 6

IFIH1 5
IL2RA 5–7, 31, 67–69, 75
INS 5

KIR2DL1 193
KIR2DL2 ..... 189, 190, 193
KIR2DL3 ..... 193, 194
KIR2DL4 195
*KIR2DL4*005* 195
KIR2DL5 ..... 189, 194
KIR2DS1 194
KIR2DS2 ..... 189, 193
KIR2DS4 194
KIR2DS5 194
KIR3DL1 .. 164, 166, 167, 170–188, 194, 195, 198, 202
KIR3DS1 .. 164, 166, 167, 170–188, 194, 195, 198, 202

PTPN2 ..... 68, 69, 75
PTPN22 5

STAT6 ..... 170, 171

SNPs

rs11594656 . . . . . 32, 61–63

rs12722495 . . . . . 31, 34, 61–63, 67, 75

rs1893217 68

rs2104286 . . . . . 32, 34, 60–63

rs2756923 . . . . . 189, 190

rs45450798 75

rs478582 76

rs592645 . . . . . 179–183, 194, 195, 198

Proteins

CD122 98

CD127 33

CD132 98

CD25 . . 31–34, 39, 44–46, 48, 61, 64, 68, 73, 98

CD4 . . . . . 33, 68

CD45RA 34, 35, 51, 53, 64

CD8 33

FOXP3 33

IL-2 . . . . . 31, 68

IL2RA 68, 223

KIR3DL1 167

STAT5 68

Chapter 1

Introduction

1.1 Motivation

Over the last twenty years, the size and dimensionality of biological datasets has increased at a tremendous rate, giving rise to very large data matrices. There is growing interest in developing computational methods for identifying patterns or trends within these datasets. However, the analysis of such large datasets is not without difficulties for both practical and theoretical reasons. One of the major challenges is dealing with noise due to differences in instrument configuration, experimental protocol or method of analysis. This complicates the extraction and comparison of biologically useful information across datasets. Another challenge is identifying rare subsets and outliers. Rare subsets may be biologically significant whereas outliers tend to be noisy samples which can skew association statistics. Throughout my thesis I will investigate how to address these issues using normalisation and clustering techniques. In particular, I will focus on normalisation and clustering of cell-level parameters, acquired with flow cytometry, as well as genetic data, acquired from quantitative Polymerase Chain Reaction (qPCR) and single nucleotide polymorphism (SNP) arrays. SNP arrays have been at the core of genome-wide association study (GWAS) over the last ten years. They concurrently

probe several hundred thousand of SNPs within an individual and the technology has been parallelised so that a single experiment can assay thousands of individuals. Each SNP probe is fluorescently labelled so that at a given genomic locus, the intensity values can be clustered across individuals, after normalisation. Genotypes at a given loci can be called by individual, depending on the cluster in the pooled dataset to which the individual is assigned. Common genotypes form larger clusters whereas rarer genotypes form smaller clusters. An equally influential technology which has brought cell biology into the sphere of "big biology", is flow cytometry, a high-throughput technique for measuring up to twenty cell parameters in millions of cells. In flow cytometry, clustering is performed to group similar cell measurements to identify cell types within a sample. Typically, the relative size of these clusters and their means are of biological interest and are compared between samples after normalisation. While clustering applied to SNP calling can be fully automated, clustering in flow cytometry is still reliant on manual inspection. One reason is that clustering of cell types is more uncertain than clustering of genotypes because the number of cell types is unknown and cells are often in intermediate states between cell types. In genetics, in the absence of copy number variation, the expected number of clusters at a given loci is typically known since there are only a finite number of possible genotypes at a SNP.

The choice of normalisation and clustering methods can have an important impact on reproducibility and association statistics (Plagnol et al, 2007). The influence of clustering on reproducibility and association testing will be the focus of Chapter 2, where I revisit a large, long-running flow cytometry experiment designed to measure genotype-phenotype correlation in hundreds of individuals. The cell type clustering was initially conducted manually by drawing gates to delineate populations of cells. I assess the influence of clustering method on association testing, when part of the gating is replaced by computational thresholding and clustering methods.

In Chapter 3, I analyse another flow cytometry dataset, this time an *ex vivo* stimulation dataset, primarily generated to assess whether there are differences in stimulation response in certain cell types, between type 1 diabetics and matched controls. I will once more consider normalisation for the purpose of improving the reproducibility of the cell phenotypes and hence the power of association testing. Furthermore, I apply computational methods to discover new clusters not identified by manual gating which are responsive to stimulation.

In Chapter 4, I apply normalisation and clustering to genetic copy number genotyping of qPCR and SNP array data where no prior manual analysis has been done. Instead, I use prior information in terms of population frequencies obtained from previous studies to guide the clustering of qPCR data. The clusters applied in qPCR are used to identify the SNP patterns which are predictive of copy number, using supervised clustering.

Finally, I will conclude with what I have learned about normalisation and clustering in general, and the specifics of their application to these datasets. I will discuss how these methods could be further applied and refined, especially with regards to flow cytometry, where they are not yet as commonly used as in genetics.

1.2 Biology of type 1 diabetes

Since all the datasets I have analysed in my thesis relate to type 1 diabetes (T1D), I will first give some background on what we know about the disease and the technologies we are using to gain new insight.

1.2.1 Aetiology and diagnosis

T1D (OMIM:222100), also known as insulin dependent diabetes mellitus (diabetes - διαβήτης, a passer through, and mellitus - μέλι, honey), is a disease reported as early

as 1500 BC (Poretsky, 2010). It holds its name from the characteristic symptom of excessive discharge of high-glucose urine (glycosuria or hyperglycemia-induced osmotic diuresis polyuria). It has since been established that this symptom is the consequence of persistently high levels of glucose (hyperglycemia) in the blood due to an insufficiency in insulin, the hormone responsible for glucose regulation. Long term high-glucose levels lead to dehydration, drowsiness, cardio-vascular complications, increased chances of morbidity and death. If left untreated T1D is a debilitating and life-threatening disease. From post-mortem analysis of pancreatic samples and animal models, it is widely accepted that the cause of the insulin deficiency in T1D is an autoimmune reaction in which insulin and insulin-producing β-cells of the pancreatic islets are progressively destroyed primarily through auto-reactive T cells (Todd, 2010).

In the last 50 years, the number of cases of T1D worldwide has increased and is predicted to continue increasing in the next decade, affecting mainly children under the age of 5 (Patterson et al, 2009). The World Health Organization reported that in August 2011 around 34 million people worldwide were diagnosed with T1D. At present there is no cure for T1D. The only existing treatment is the regular intravenous administration of exogenous insulin. Pre-symptomatic detection of T1D relies on testing for presence of auto-antibodies against insulin and its precursors. Early detection of T1D allows a better understanding of how the disease progresses and how we can develop therapies to delay its onset, reduce the symptoms and hopefully in the future, cure the disease. One such therapy currently undergoing clinical trials in our lab attempts to restore immune tolerance to pancreatic β-cells with low-dose interleukin-2 (IL-2), in newly diagnosed T1D patients (`http://www.clinical-trials-type1-diabetes.com/`).

1.2.2 Genetics of type 1 diabetes

Patterns of familial clustering suggest that a portion of T1D risk is inherited. A measure of heritability used by geneticists is the sibling recurrence risk λ_s, which is defined as the ratio of the probability that a sibling of an affected individual has the disease over the probability of a random individual in the population having the disease. For T1D, λ_s has been estimated to be close to 15 (Risch, 1987), although Clayton (2009) suggested that this might be an overestimate and that λ_s is more likely to lie between 5 and 8.9.

Regardless of the exact estimate of λ_s, some genetic predisposition to T1D is indisputable and researchers have long been interested in identifying likely causal variants in our genetic code that might lead to some insights into the mechanism of the disease. Linkage studies based on the recombination of multiallelic genetic markers in families affected by T1D, first mapped a genetic risk factor to the human leukocyte antigen (HLA) region on chromosome 6 (Singal and Blajchman, 1973; Cudworth and Woodrow, 1974; Nerup et al, 1974). As insulin is a target of the autoimmune response in T1D, the insulin gene (*INS*), on chromosome 11, was tested as a strong candidate region and was also found to associate with the disease (Bell et al, 1984; Permutt et al, 1984). In 1994, Davies et al, using a linkage map of 290 marker loci in 96 sibling pairs, confirmed the association with the HLA and insulin gene regions, and further reported a number of new chromosome regions showing some potential evidence of linkage to T1D. The study confirmed that T1D is a polygenic disease and that furthermore, there were unlikely to exist other loci with as strong an effect as HLA.

More recently, GWAS using high density SNP arrays, have confirmed strong association of T1D within the HLA (chromosome 6p21) and *INS* (chromosome 11p15) loci, and reported 50 other loci, including regions near *CTLA4* (chromosome 2q33), *PTPN22* (chromosome 1p13), *IL2RA* (chromosome 10p15) and *IFIH1* (chromosome 2q24) (Smyth et al, 2006; Nejentsev et al, 2007; Wellcome Trust Case Control Consortium, 2007; Bar-

rett et al, 2009). Within the HLA region, the strongest effect comes from the HLA class II loci, *HLA-DRB1* and *HLA-DQB1*, but, there is evidence for additional independent effects from the HLA class I loci, involving HLA-A and HLA-B alleles (Nejentsev et al, 2007; Howson et al, 2009). A comprehensive and updated list of all T1D associated loci found so far is maintained on the T1DBase website (`www.t1dbase.org`).

1.2.3 The immune cell mechanisms

Many of the reported T1D-associated genetic variants are located in proximity to genes and regions with known immune function such as HLA and *IL2RA*, which code for receptors found at the surface of immune cells.

Immune cells are white blood cells formally known as leukocytes, that function in the lymph nodes and other lymphoid tissues but can also be found at lower concentrations in the peripheral blood as they circulate throughout the body. They include lymphocytes, monocytes and granulocytes, and within these subsets, there exists a huge diversity in terms of size, gene expression and function. It is this diversity that enables the versatility of the immune system in neutralising all kinds of pathogens (innate immunity), its ability to distinguish them from endogenous cells (self tolerance), and its capacity to adapt to better counter future infections (adaptive immunity). An important type of leukocyte in the adaptive immune response are T lymphocytes also known as T cells. After having undergone central selection in the thymus, T cells in the peripheral blood have an affinity for foreign antigens but are tolerant to self. Initally these cells are in a naive state (naive T cells) until presented with an antigen, at which point they mainly differentiate into effector T cells, capable of mounting an immediate response, but also into longer-lived memory T cells, capable of mounting a stronger and faster response in the future thus resulting in long lasting immunity against this pathogen (acquired immunity). In order to moderate the scale of the immune response and preserve self-tolerance, some T cells

also have a regulatory function on the immune response mediated using small signalling molecules known as cytokines (for example interleukin-2 (IL-2)). These regulatory T cells are important in preventing autoimmunity and hence are the object of thorough study in T1D.

Some insight into the aetiology of T1D may be gained by seeing how T1D-associated variants correlate with quantitative cell phenotypes such as, ratios of different cell types or mean expression of surface proteins. For example, Dendrou et al (2009b) showed that protective T1D risk variants in proximity of the protein coding gene *IL2RA*, correlate with increased mean expression of CD25 on the surface of memory T cells. $CD4^+$ memory cells with higher CD25 levels are likely more responsive to IL-2 and TCR-mediated activation, which in turn leads to increased production of IL-2, suggesting a mechanism by which self-tolerance may be boosted.

1.3 Studying the immune system with flow cytometry

The established method for measuring immune cell phenotypes is flow cytometry. By labelling cells with fluorescent probes conjugated to antibodies, it is possible to distinguish a wealth of distinct cell subsets which concomitantly express specific molecules. Flow cytometry allows us to identify and quantify different types of cells, through individual cell measurements.

The flow cytometer Fluorescence intensity is measured accurately using photosensitive detectors, normally a photomultiplier tube (PMT), which turn light into an analogue (current or voltage) or digital (photon counting) electronic signal which is translated into a digital number indicating the intensity of the fluorescence (Shapiro, 2003; Snow, 2004). For a fluorochrome to emit fluorescent light, it needs to have absorbed high energy light

of a given wavelength from an illumination source, usually from a laser, which it can then release at a lower energy, longer wavelength, resulting in a so-called Stokes shift. The wavelength spectrum at which a given fluorochrome most efficiently absorbs and emits light and Stokes shift are known and depend on the physico-chemical properties of that molecule. To enable optical illumination, separation and collection of various fluorochromes with different emission and excitation spectra, a flow cytometer is usually equipped with several lasers which emit at different wavelengths and specially configured optical mirror, filters and photosensitive detectors which are sensitive to light at distinct frequency ranges (Shapiro, 2003).

Sample staining When staining a sample, fluorochromes are conjugated with antibodies with an affinity for the target polypeptide we wish to quantify. The target can be external, for example a cell receptor, or internal, for example a transcription factor or a cytokine. If the target is internal, the cells have to undergo permeabilisation that can deteriorate the general quality of the staining. Fluorochromes should be selected to minimise overlapping of their emission spectra. Spectral overlap, also known as spillover, leads to a convoluted signal reaching the detectors. Antibodies are also a potential source of noise, since both primary and secondary antibodies may bind to more than one target. Antibodies differing in the constant regions of the heavy and light chains, known as isotypes, or non-immune sera, can be used to control non-specific staining and/or reduce non-specific binding by blocking secondary targets.

Running a sample on the flow cytometer Once a solution of fluorescently labeled cells is fed to the flow cytometer, the sample is delivered to the flow cell after hydrodynamic focusing (Figure 1.1). In the flow cell, the cells are filed up individually so they cross a laser beam one by one (Shapiro, 2003). As a cell crosses the laser beam some light is scattered and some is absorbed. The detected scattered light is used to

provide an estimate of the size and granularity of the cell. Light scattered in the forward direction (diffracted light) is correlated with the size of the cell, whereas light scattered sideways (refracted light) is correlated with the complexity of the cellular structure. The absorbed light is later emitted as heat and fluorescent light. The intensity of the scattered and fluorescently emitted light measured by the detector thus provides quantitative information about the correlates of size and granularity, and the presence of certain fluorescently-marked molecules for each cell. When examining leukocytes using only the physical properties provided by the scattered light intensity, it is possible to distinguish lymphocytes from monocytes and more granular neutrophils. Combining this information with the fluorescent intensities it is possible to further distinguish between different types of lymphocytes which have in common certain cellular receptors or transcription factors.

Fluorescent crosstalk As we delve deeper into the lymphocyte subsets more fluorochromes are needed to further distinguish between different classes (Perfetto et al, 2004). However when adding more and more fluorochromes, overlap of emission spectra becomes unavoidable (Roederer, 2001). This implies that the intensity signal measured in one detector is in fact a mixture of signals from other flurochromes which spillover across detectors (Figure 1.2). The deconvolution of this signal is a process known as compensation. The matrix solution is known as the spillover matrix and is usually a square matrix with as many rows as there are fluorochromes and columns as there are detectors. To calculate the spillover matrix, single coloured beads are used. The pairwise contribution of a fluorochrome to a non-specific channel is then summarised as a compensation matrix (Table 1.1). By subtracting the spillover values from the mixed intensity one can then recover the original intensity. This compensation step is usually performed after all the data from an experiment has been collected before commencing analysis.

Signal \ Detector	PMT 1	PMT 2	PMT 3	PMT 4	PMT 5	PMT 6
Alexa-488	1	0	0	0.16	0	0
PE-Cy7	0	1	0	0.1	0.2	0
APC	0	0	1	0	0.3	0
PE	0.2	0.1	0	1	0	0
Alexa-700	0	0.1	0.2	0	1	0
Pacific Blue	0	0	0	0	0	1

Table 1.1. Spillover matrix of the fluorochomes used by Dendrou et al (2009b) obtained using single colour beads. Each entry is the percentage of the emitted fluorochrome signal (row) picked up by a detector (column). The rows represent the fluorochomes and the columns are the PMT detectors. Each detector is tuned to capture the intensity of a single fluorochrome (diagonal entries). Spillover occurs when certain fluorochromes are detectable by more than one detector (non-zero terms off the diagonal). Notice that there is non-negligeable spillover (30 %) of APC into PMT 5, the detector meant for Alexa-700.

Compensation to account for fluorescence crosstalk is just one of the intricacies of flow data. Other intricacies are data format and the choice of the data transformation, which can both have an important impact on the analysis.

Flow cytometry data format The data format determines the range and the precision of the data stored. The objective of the Flow Cytometry Standard (FCS) is to define a unified file format for flow data that allows files created by one type of acquisition hardware and software to be analyzed by any other type. The first FCS format for data files was FCS 1.0 (Murphy and Chused, 1984). The standard was later updated in 1990 as FCS 2.0 (Dean et al, 1990) and again in 1997 as FCS 3.0 (Seamer et al, 1997). FCS 2.0 and FCS 3.0 are the current two main competing standards. FCS 2.0 is a logarithmically compressed format which does not allow negative intensities. Instead negative values reported by the instrument are arbitrarily assigned the minimum value. This leads to what is described as the log artefact: a pile up of intensities on the axes for low intensity values. FCS 2.0 data are integers in the range 1 to 10000 (4 decades).

FCS 3.0 on the other hand is closer to the raw data, covers a greater range and allows for negative values. FCS 3.0 leaves more flexibility to the choice of transform. FCS 3.0 are floating point numbers in the range -211 to 262143 (8 decades) FCS 2.0 requires practically no post processing except for a log transform. FCS 3.0 requires more careful thought as it leaves to us the compensation and the choice of a suitable transformation. For low intensity fluorescence, FCS 3.0 is the preferred format as the truncation at zero of intensity values for FCS 2.0 can lead to loss of information (Figure 1.3).

Data transformation for display and analysis As fluorescence intensity tends to scale multiplicatively, intensity data needs to be linearised for the purpose of visualisation and clustering. Clustering algorithms based on variance (average distance to the mean) perform poorly on skewed data, and in general humans are more comfortable dealing with data on a linear scale. Given FCS 2.0 data is strictly positive, a simple $\log_{10}$ transform is usually applied to linearise the data. As FCS 3.0 allows for negative values a different transform is required. Transforms that are closer to linear near zero are preferred, since the logarithmic transform is distorting (Figure 1.4) for low intensity values (Durbin et al, 2002; Herzenberg et al, 2006). Some appropriate transformations for FCS 3.0 are the Generalized Arcsinh, the Logicle transform, the LinLog and the Generalized BoxCox (Bagwell, 2005; Parks et al, 2006; Finak et al, 2010). Given the data, parameters for these transformations can be estimated using maximum likelihood assuming a multivariate Gaussian distribution of the data (Finak et al, 2010). However, as illustrated by Herzenberg et al (2006), care needs to be taken as the transforms can introduce spurious peaks in the intensity distribution around zero. The Logicle transform as defined by Parks et al (2006), is the most widely used transform for FCS 3.0 data and is the one I used in this thesis. It takes as input the w parameter which influences the linearization width around zero in asymptotic decades. The influence of the w parameter on the shape of the transformed intensity distribution is illustrated in

Figure 1.5.

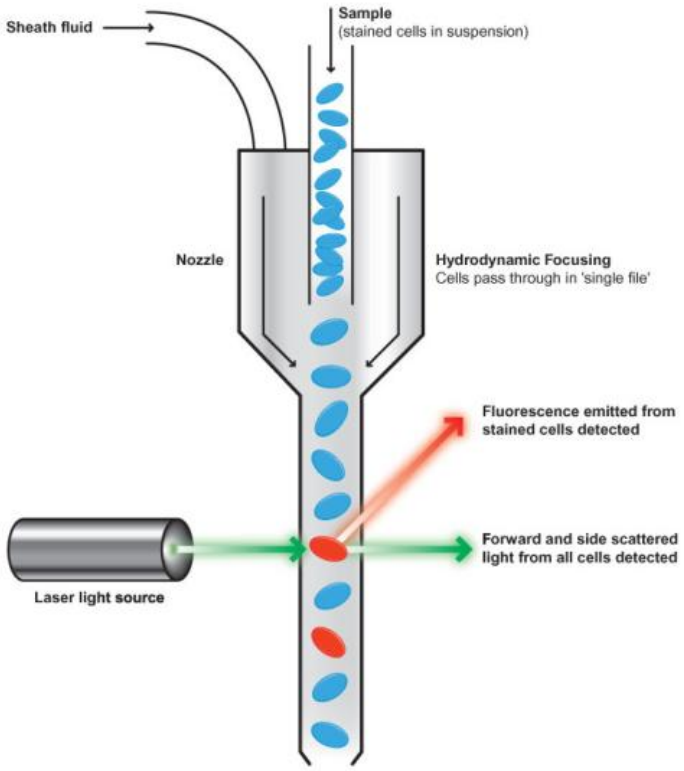


Figure 1.1. Flow cytometer diagram. Source `www.abcam.com`

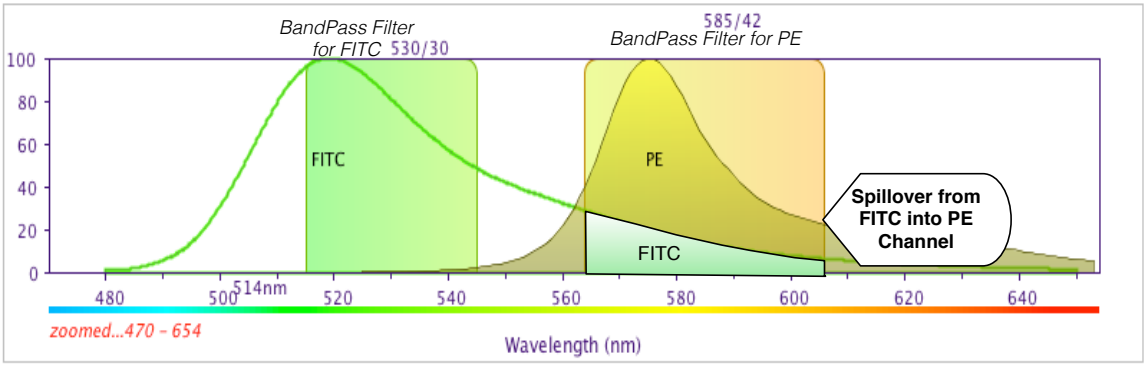
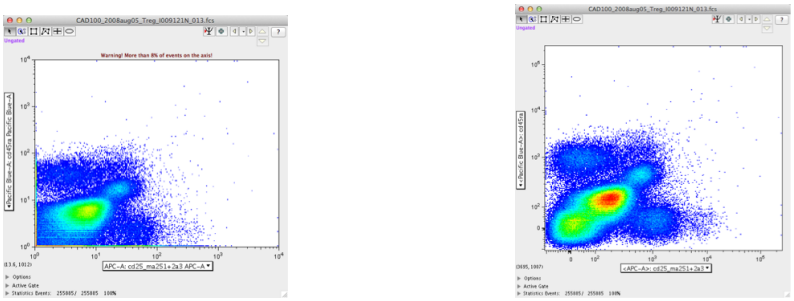


Figure 1.2. Leaking of signal from FITC fluorochrome into PE detector.
Created using: `http://www.bdbiosciences.com/research/multicolor/spectrum_viewer/`



(a) FCS 2.0 **(b)** FCS 3.0

Figure 1.3. The same data encoded in FCS 2.0 (a) and FCS 3.0 (b). In FCS 2.0 (a), truncation at zero leads to loss of low intensity clusters as compared to FCS 3.0 (b).

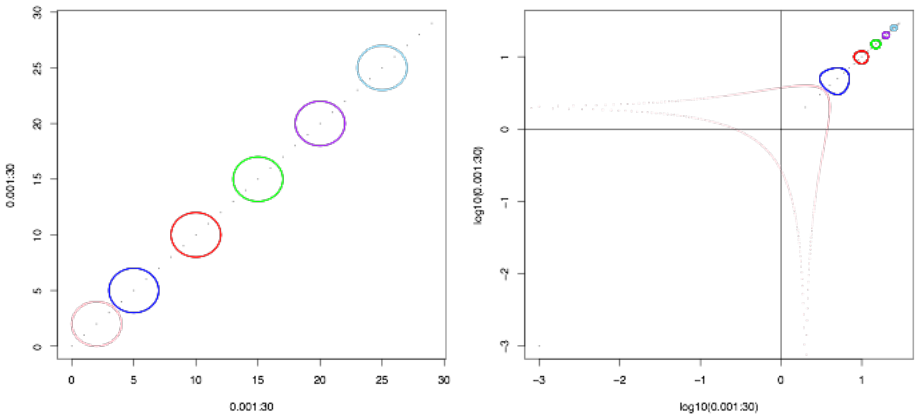
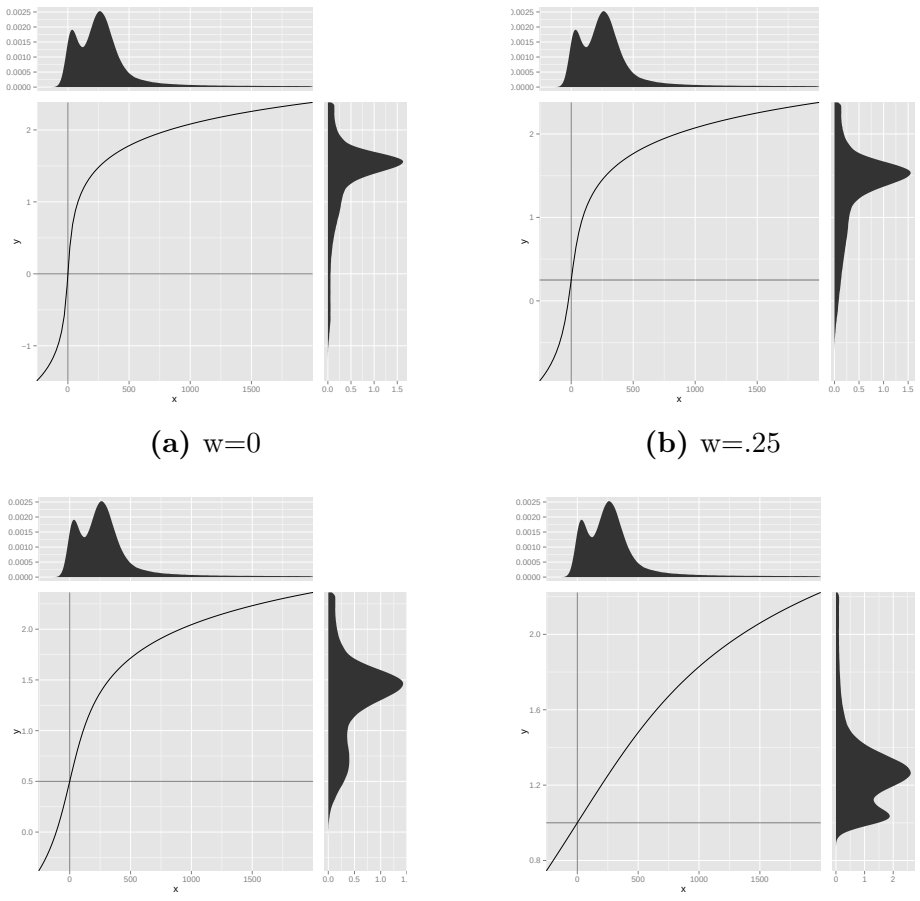


Figure 1.4. Log transform inflates the variance around zero. Depicted on the left are circles of equal diameter viewed on a linear scale. These circles represent spherical two-dimensional Gaussian distributions, When a logarithm transform is applied, the shape of the circles are distorted and they no longer have the same area. Close to zero, the area of the circles increases and their shape is distorted. This would translate into an increase of the covariance for a Gaussian distribution.



(a) w=0 **(b)** w=.25

(c) w=.5 **(d)** w=1

Figure 1.5. Effect of different values of w parameter of the Logicle transformation on the data distribution. The Logicle transform maps the data distribution defined on x to the data distribution defined on y. The transform approximates a linear transform near zero and a log transform elsewhere. The w parameter maps the zero of the transform to w.

Noise in flow cytometry Beside these data format and transformation intricacies, there are many sources of noise in flow cytometry that complicate sample comparison and impact reproducibility:

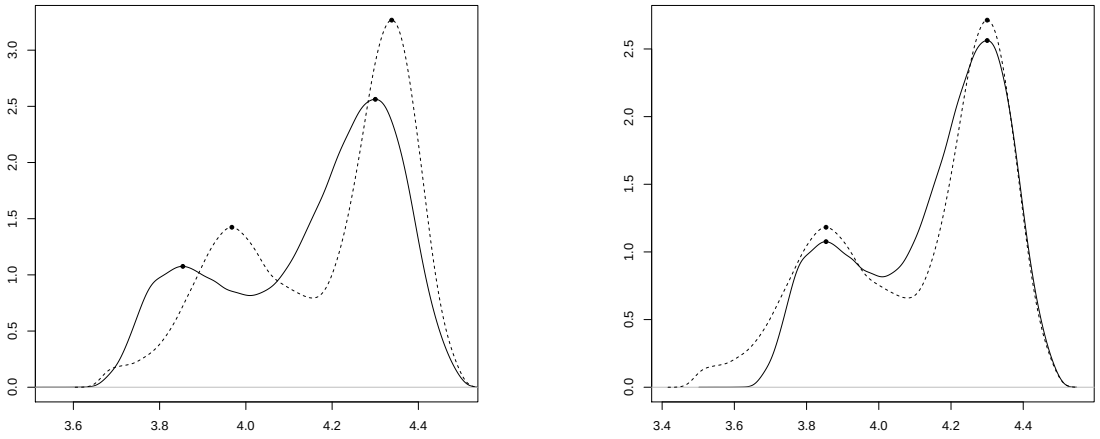
- **Sample processing and storage conditions noise.** Certain surface markers are more fragile than others and may be shed when cells are frozen for storage. Also certain chemical treatments like permeabilisation can affect the quality of the staining.
- **Noise associated with the staining of the sample.** The qualitative and quantitative choices for selecting antibodies and fluorochromes influence the quality of the staining. Antibodies have a tendency to be sticky and can bind to other targets in an erratic manner leading to spurious signal. The level of non-specific binding can be assessed with isotype antibodies.
- **Noise linked to the instrument.** The reliability of the lasers and detectors may decrease with time. Fluorescent beads can be used to detect and correct these variations.
- **Noise due to the flow operator.** Sometimes the operator might decide to not collect all the events and for example apply a cutoff on the side and forward scatter.

All these sources of noise contribute to different patterns of staining and concentrations of debris which may lead to spurious cell populations or skew the analysis, complicating the analysis across samples and laboratories. Some of these issues are being addressed by the Human Immunology Project (HIP) consortium standardisation efforts (Maecker et al, 2012), which aims to enhance reproducibility of flow analysis results across laboratories, through the use of lyoplates, and agreement on experimental protocols and instrument configurations. However, data analysis techniques such as normalisation are still necessary to deal with residual noise.

1.4 Normalisation

The purpose of normalisation is to remove unwanted experimental variation to make data comparable even when the samples are collected on different days, processed with different protocols or instrumental configurations. Nonetheless, distinguishing between unwanted and biological variation necessitates some prior knowledge about the datasets, either in the form of global distributional assumptions or in the form of local features which exist in a predictable relationship across samples. In microarray gene expression datasets, for example, one distributional assumption is that the majority of genes are not differentially expressed between similar samples, hence the expected log ratio of gene expression in two samples should be centered on zero (Smyth and Speed, 2003; Bolstad et al, 2003). It is also possible to use reference points to normalise across samples by using local features of the distribution or objects with known properties, such as beads in flow cytometry or reference probes in microarrays. In microarray, since the number of data points is constant and the distributions are unimodal across samples, normalisation methods like quantile normalisation perform well (Bolstad et al, 2003). However, in flow cytometry, this type of normalisation is not appropriate because samples contain different number of events and the distributions are typically multimodal, as commonly found in datasets containing mixtures of groups. While in theory, the locations of these modes or peaks of the density function should remain fairly stable across samples provided experimental parameters are kept constant, in practise there is often variation attributed to factors that are beyond our control such as long-term instrument decalibration. On the other hand, the height of the peaks, the relative frequencies of the cell populations, are expected to change since they are are sensitive to sampling variation. These observations motivate a normalisation method which aligns the peaks of the distributions so that cell populations are centered in a similar location across samples even when their relative proportions changes. The implementation of this normalisation method then

depends on the technique used to identify and match the peaks across samples. One method of identifying peaks of the density function is with a sliding window approach. The sliding window records the point with the highest density estimate in the current window and returns a list of highest density points of which the top K may be chosen. This is one of the approaches implemented in the R BioConductor package flowStats (Hahne et al, 2013). Figure 1.6 illustrates this method on real flow data where two common groups stand out and are reasonably well separated. Unfortunately, peaks are not always consistently identifiable across samples. In these cases, it may be preferable to only identify the most distinguishable subset of peaks, those representative of the most common groups, in order to do the alignment.



(a) identification of peaks **(b)** peak alignment

Figure 1.6. Normalisation by peak alignment. Distributions of the same marker in two different flow cytometry samples. The two peaks of each distribution are identified in (a) using a sliding window of size 40 on the density function and aligned in (b) using a linear transform.

1.5 Manual clustering

Once the data has been made comparable thanks to normalisation, the next task is to identify clusters, which in the case of flow cytometry are groups of cells that share some similar properties, which can be matched across samples.

Given a one, two or three dimensional projection of the data, clusters can in some cases, be identified by eye. When the properties of the sought population are known, a manual or supervised (semi automatic) approach can be used.

The manual method, known as manual gating, is a step-by-step method where we consider and plot two channels at a time and delineate a region, called a gate, such that cells which lie outside the gate are filtered out. The result is that a population is defined as an intersection of multiple one or two dimensional gates. This process scales poorly when we increase the number of parameters (fluorochromes in flow cytometry) and samples. As only the pairwise correlation can be assessed, identification of higher-dimensional clusters can be compromised. The ordering in which the gates are drawn can affect the final clustering solution.

Manual gating introduces further technical variation since the position of the gates on the same data can differ between gaters (Maecker et al, 2010). It suffers from strong bias as it tends to force data to fit a model (the gater's expectation). Finally, manual gating is not practical when an exhaustive enumeration of all identifiable cell populations is required (Siebert et al, 2010; Aghaeepour et al, 2012) especially as the number of cellular markers increases. For this, unsupervised computational methods, which do not rely on visualisation, are essential.

1.6 Automatic methods for identifying clusters

Automatic flow data analysis methods have been reviewed by Bashashati and Brinkman (2009), Lugli et al (2010), and, more recently, by Aghaeepour et al (2013). They are benchmarked annually by the Flow Cytometry Critical Assessment of Population Identification Methods (FlowCAP) group and broadly fall in two camps: unsupervised methods which have have unlabelled data and supervised methods which require manual training by giving approximate starting gates.

Clustering methods which make explicit assumptions about the shapes of populations are model-based or parametric. Methods which do not are said to be model-free or non-parametric, although the latter can be limiting cases of parametric models. Here I will mostly review unsupervised methods where training data is not provided, and focus on those which require specification of only some parameters such as the expected number of clusters.

1.6.1 Model-based methods

Model-based methods stipulate that flow data can be explained by a mixture of multivariate distributions where each distribution is representative of a different type of cell. A useful property of these methods is that they can assign a probability of population membership to each cell which can be exploited in downstream statistical analysis to account for uncertainty in the clustering. The first and simplest of these methods applied to flow cytometry data (Chan et al, 2008), assumed cell populations could be represented by a Gaussian mixture model (GMM) of K multivariate Gaussians:

$$p(x_i) = \sum_{k=1}^{K} \tau_k \frac{1}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma}_k|}} e^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k{}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)}; \quad \sum_{k=1}^{K} \tau_k = 1$$

where x_i are the coordinates of the ith cell in a data set of size N. The parameters,

$\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and τ_k, correspond to the cluster mean, covariance and weight, respectively, and k indexes the clusters from 1 to K. Assuming the data are identically and independently distributed, we can attempt to estimate the parameter θ, representing the sets of (μ, Σ, τ), which maximises the joint probability, or likelihood function, of observing the data given the model:

$$\mathcal{L}(\theta|X) = \prod_{i=1}^{N} p(x_i|\theta).$$

As products of small numbers are hard to deal with analytically and are numerically unstable, this is more commonly done by maximising the logarithm of the likelihood:

$$\ln \mathcal{L}(\theta|X) = \sum_{i=1}^{N} \ln p(x_i|\theta).$$

In order to find a global optimum of the likelihood function, algorithms proceed in an iterative fashion to explore the parameter space. The stopping criterion is reached upon convergence of the likelihood function or equivalently of the parameter updates. However local optimums in the likelihood function can also lead to convergence. There are also regions of the parameter space which need to be avoided. For example, the likelihood function can be made arbitrarily large if the variance of one of the clusters is allowed to shrink to zero. To safeguard from these situations, some guidance can be provided by picking sensible starting conditions or by setting hard boundaries on the parameter space. Another softer approach is to weight parameter updates with a distribution. This approach is also called regularisation. Regularisation can be achieved using a prior probability density function on the parameters as implemented in the R package mclust (Chris Fraley and Scrucca, 2012) and the R BioConductor package flowClust (Lo et al, 2009).

One drawback of the GMM is that it tends to overestimate the number of multivariate Gaussians which best models the data since outliers which are in the tails of

the distributions are explained by new low mixture distributions, increasing the number of reported cell populations. To account for this, Flowclust replaces Gaussians by t-distributions which have more weight in the tails (Lo et al, 2008). Even so, t-distributions are symmetric and so cannot model skewed populations commonly found in stimulation experiments or more generally when cells are in a transitional state from one cell type to another. Pyne et al (2009) addressed this issue with FLAME (Flow Analysis with Automated Multivariate Estimation) by employing skewed t-distributions instead. Yet a remaining issue is that these distributions are convex by nature and so a concave population which can arise in transitional cell populations undergoing progressive change on more than one marker may only be represented by the merging of several convex populations. This merging step can be accomplished using FlowMerge (Finak et al, 2009). By adding more parameters to these distributions we can make model-based methods more flexible, but this comes at the price of reducing the degrees of freedom and having to estimate more parameters which can potentially be computationally expensive and risks overfitting. For certain parameters such as mean and covariance, closed-form solutions exist or can efficiently be estimated with an expectation maximisation (EM) algorithm (Dempster et al, 1977), but others, such as the skewness factor and degrees of freedom of the t-distributions, may need to be estimated numerically using computationally expensive iterative methods.

1.6.2 Model-free methods

Model-free methods state no explicit assumptions about the shape of populations but instead attempt to minimise some loss function such as the total within-cluster sum of squares. These methods are sometimes qualified as non-parametric because there are no explicit parameters to the model, although the parameter estimation is usually implicit. They typically use local estimates of density or distance to identify clusters of points.

Perhaps the oldest and most popular of the model-free methods, due to its simplicity and speed, is the K-means algorithm (MacQueen, 1967). K-means attempts to minimise the total within-cluster sum of squares as its objective function:

$$SS_w = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in S_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)^2; \quad \boldsymbol{\mu}_k = \mathrm{E}(x_i | x_i \in S_k) \tag{1.1}$$

where each point, x_i, is assigned to exactly one of the clusters $S_1, S_2, \cdots S_k$.

The algorithm starts by picking K random points which are initial guesses as to where the cluster means lie. Then for each iteration:

- Each point x_i is assigned to the cluster S_k of the closest current cluster mean μ_k.
- Based on this new cluster assignment, the new cluster mean of each cluster is computed.

The algorithm terminates when no cluster mean changes, or equivalently when the cluster assignment does not change.

However, it can be shown that certain so called non-parametric methods turn out to be limiting cases of parametric methods. For example K-means, is a limiting case of spherical GMMs when the covariance tends to zero because maximising the log-likelihood of a multivariate Gaussian is asymptotically equivalent to minimising the residual sum of squares (Figure 1.7).

There are numerous model-free clustering methods which rely on density estimation:

- Density Based Merging (DBM) uses a region growing density approach so that points of high density are linked (Walther et al, 2009).

- Flow Clustering without K (FLOCK) relies on grid based density estimation in higher dimensions (Qian et al, 2010).

- Misty Mountain takes decreasing cross sections of the two dimensional density histogram so that clusters are progressively merged as the threshold decreases

(Sugár and Sealfon, 2010).

- CurvHDR (negative Curvature and High Density Region) identifies regions of significant curvature in the multivariate density function, by using a bandwidth rather than the number of bins, but cannot be applied beyond four dimensions (Naumann et al, 2010).

All these density based approaches require the number of bins or the bandwidth to be specified by the user or estimated from the data according to some heuristic. Care needs to be taken when selecting an appropriate bin or bandwidth to avoid under or over smoothing of the data. For example, smaller numbers of bins or equivalently larger bandwidths, lead to overestimation of the density in low density regions and underestimation in high density regions. However, as the number of dimensions increases, data tend to become sparser and density based clustering becomes inefficient because the lower bound on the density estimation error increases with the dimensionality. Therefore in high-dimensional settings, distance based clustering becomes more appropriate and fast methods like K-means, which do not rely on the computation of the entire distance matrix, are popular. For example, flowMeans is an extension to K-means for flow cytometry (Aghaeepour et al, 2010). Unfortunately, K-means is known to be very sensitive to starting conditions and is also not robust to outliers, so small changes in the data can lead to very different clustering solutions. These shortcomings are addressed by a related but slower algorithm: K-medoids. Instead of using the cluster means, K-medoids updates the cluster medoids. The medoid is defined as the point of the cluster which minimises the overall distance to all other points belonging to that cluster. The objective function of K-medoids is therefore:

$$\sum_{k=1}^{K} \sum_{\mathbf{x}_i \in S_k} (\mathbf{x}_i - \boldsymbol{M}_k)^2; \quad \boldsymbol{M}_k = \text{Medoid}(x_i | x_i \in S_k)$$

Since the medoids can only be points of the dataset, the complete distance matrix need only be calculated once at the onset. The algorithm starts by picking K starting points which belong to the set of data points. These represent the initial guess as to where the cluster medoids lie. Then for each iteration:

- Each point x_i is assigned to the cluster S_k of the current closest cluster medoid M_k.
- Based on this new cluster assignment, the new cluster medoid of each cluster is selected.

The algorithm terminates when no cluster medoid changes. The algorithm is reasonably fast but can be slower for larger datasets due to the first step of calculating the complete distance matrix. For sufficiently large N, the size of the distance matrix may be prohibitive and too large for memory. Therefore this version of the algorithm may not scale with large N as well as K-means. This performance issue can be ameliorated by clustering subsets of the data and combining the results. Points which were not selected in the subsampling can be assigned to the closest cluster. This is the approach implemented by the `R` function `clara` (Clustering Large Application) in the `R` package cluster (Maechler et al, 2014).

If subsamping is used only once to reduce the number of datapoints then caution is required, since uniform downsampling runs the risk of discarding smaller clusters. One workaround is to account for local density so that most of the downsampling occurs in regions of high-density, thus conserving low density regions. Once the downsampling has sufficiently reduced the number of points so that distance matrix may be computed, methods such as, spectral clustering which uses spectral graph theory to determine where to partition the network as implemented in Sampling followed by Spectral Clustering (SAMSpectral, Zare et al (2010)) or more conventional hierarchical clustering as done by Spanning-tree Progression Analysis of Density-normalized Events (SPADE, Qiu et al

(2011)), may be applied.

The advantage of these non-parametric methods is that they are usually fast and flexible and often provide some heuristics for estimating the parameters. Unfortunately, this seeming flexibility can come at the price of poor interpretability or generalisation to other datasets since there is no explicit model and the parameters tend to be dataset specific. Furthermore, these parameters do not necessarily have an intuitive interpretation or any biological relevance, which makes setting their value hard to justify.

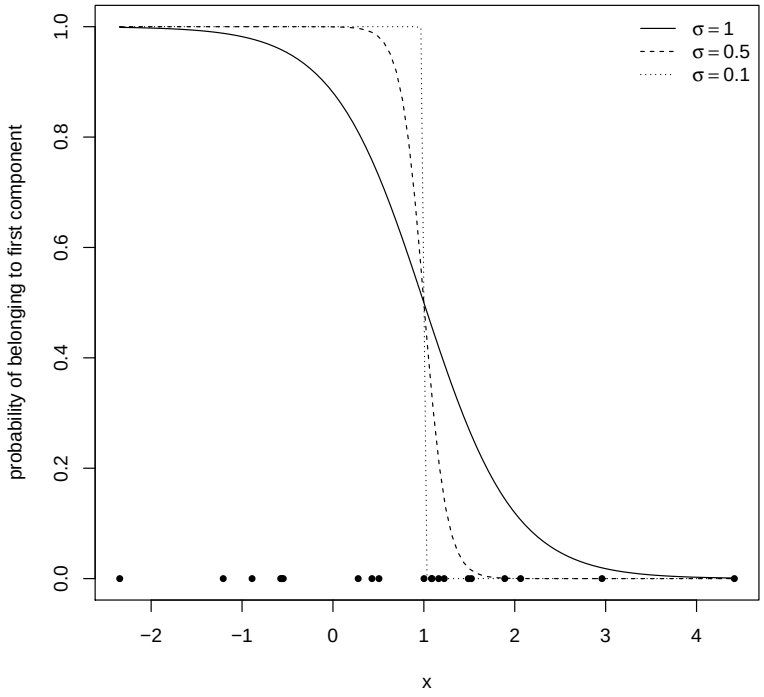


Figure 1.7. Effect of shrinking variance in a two compnent GMM on the decision boundary. In this two component univariate GMM, the y axis represents the posterior probability of being assigned to the first component and the x axis represent the value of the data points. As the variance of the two Gaussian components shrinks, the soft decision boundary approaches the hard K-means decision boundary, so that we go from a soft probabilistic assignment to a hard binary assignment.

1.6.3 Estimating the optimal number of clusters from the data

The methods described above normally require the expected number of clusters as an input parameter. Although there are many suggested heuristics, estimating an optimal number of clusters from the data remains an open problem in machine learning and statistics.

Generally these methods seek to maximise utility by reaching a compromise between model complexity, number of parameters in the model, and accuracy, in supervised clustering, or some other metric like variance explained, in unsupervised clustering.

One way of estimating the expected number of clusters, is to find modes in the data: regions of significantly high density (Duong et al, 2008; Jing et al, 2009). An alternative brute force approach is to cluster for increasing numbers of clusters and to pick the cluster value which optimises some criterion, such as the Gap statistic (Tibshirani et al, 2001) or the variance-ratio criterion (Calinski and Harabasz, 1974). These clustering metrics are usually derived from the ratio of the between-cluster sum-of-squares (variance explained by the model) to the within-cluster sum-of-squares (remaining variance). For example, the variance-ratio criterion as defined by Calinski and Harabasz (1974), for k clusters defined on a dataset of size n:

$$\text{VRC} = \frac{SS_b/(k-1)}{SS_w/(n-k)}$$

where SS_w is the total within-sum-of-squares (Equation (1.1)) and SS_b, the total between-sum-of-squares, is defined as:

$$SS_b = SS_t - SS_w$$

where SS_t is the total sum-of-squares:

$$SS_t = \sum_{i=i}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^2; \quad \boldsymbol{\mu} = \mathrm{E}(x_i)$$

Calinski and Harabasz note that the variance-ratio criterion relates to to the analysis of variance (ANOVA) F-test. Figure 1.8 illustrates that the performance of this metric in predicting the number of clusters is very much dependent on the inherent noise in the data.

In model-based methods, another approach is to maximise the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), which are both functions of the log likelihood penalised by the number of clusters and free parameters. The BIC is used by the `R` package mclust (Chris Fraley and Scrucca, 2012) to find the optimal number of parameters when fitting a GMM.

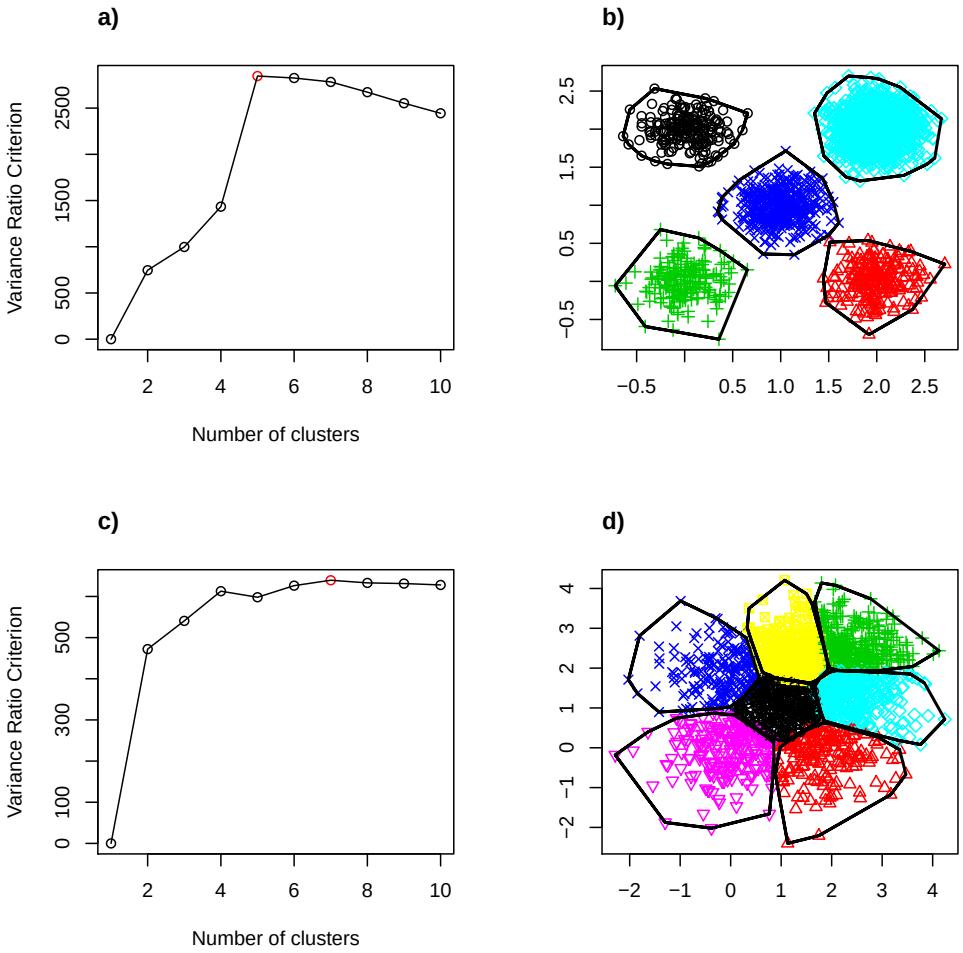


Figure 1.8. Variance ratio criterion used to estimate the number of clusters with K-means. Two simulated datasets are used to illustrate the variance criterion ratio on K-means. The first dataset (a and b) is simulated from a five component mixture of bivariate spherical Gaussians with covariance of 0.05, the second (c and d) is simulated from the same model but with the covariance scaled 10-fold to 0.5. While on the first dataset the optimal number of clusters according to the variance ratio criterion is correctly estimated as 5 (a), on the second dataset it is overestimated as 7 due to the increased noise (c). b and d depict the clustering resulting from K-means with $K = 5$ and $K = 7$ respectively.

Chapter 2

The influence of gating on repeatability and effect size estimation

2.1 Background

At least two SNPs that associate with T1D are located in the chromosome 10p15 region containing *IL2RA*, the interleukin-2 receptor subunit alpha gene (Lowe et al, 2007). *IL2RA* codes for CD25, the high affinity binding alpha chain of the trimeric IL-2 cytokine receptor. CD25 is found at varying quantities on the surface of numerous T lymphocyte subsets such as naive, memory and regulatory cells. CD25 is also upregulated upon activation in lymphocytes and monocytes, and is known to play a key role in immunoregulation and immune responsiveness (Brusko et al, 2009; Boyman and Sprent, 2012). SNPs in the *IL2RA* region have also been associated with other immune mediated diseases including multiple sclerosis (Beecham et al, 2013) and rheumatoid arthritis (Stahl et al, 2010).

To better study the downstream implications of three *IL2RA* SNPs, namely rs12722495,

rs2104286 and rs11594656, on CD25 expressing T lymphocyte subsets, Dendrou et al analysed blood samples obtained from healthy donors from the CambridgeBioresource[1], selected by genotype at these SNPs, and matched by sex and age. The experiment consisted of a total of 180 individuals, fifteen of which were recalled for a second sample (Table 2.1). The distribution by age (20 to 50 years old) and sex was split evenly across genotype groups (Table 2.2).

Individual	pch	number of days between visits
1	a	196
2	b	225
3	c	217
4	d	197
5	e	161
6	f	153
7	g	133
8	h	133
9	i	117
10	j	112
11	k	112
12	l	116
13	m	119
14	n	98
15	o	79

Table 2.1. Number of days till second visit for recalled individuals. Fifteen individuals recalled between 79 and 225 days later. pch is the plotting character used to refer to these individuals in plots later in this chapter.

After lysis of the red blood cells, the samples were stained with the antibody panel specified in Table 2.3. The running time of the whole experiment was seven months over which samples were analysed on 51 days, between one and six samples per day (Figure 2.1).

The cell phenotypes studied by Dendrou et al (2009b), were obtained using manual gating with the FlowJo software[2]. Manual gating follows the current state of knowledge

[1] `www.cambridgebioresource.org.uk`
[2] `www.flowjo.com`

rs12722495	rs2104286	rs11594656	F	M	mean age
AA	AA	AA	18	10	38.9
AA	AA	AT	12	12	38.2
AA	AA	TT	31	32	39
AA	AG	TT	10	6	37.7
AA	GG	TT	12	9	40.8
AG	AG	TT	12	10	41.5
GG	GG	TT	9	12	38.4

Table 2.2. Distribution of subjects in study, by genotype, age and sex.

Fluorochrome	Antibody target
Alexa-488	CD127
PE-Cy7	HLADR
APC	CD25
PE	CD101
Alexa-700	CD4
Pacific Blue	CD45RA

Table 2.3. The fluorochrome-antibody panels with six markers used in the IL2RA dataset.

of immune cell lineages and the gating strategy followed by Dendrou et al (2009b) is described in Figure 2.2. Lymphocytes are distinguishable from more granular and larger cell types based on forward and side scatter (Figure 2.2a). The lymphocytes include, B cells and T cells, and the latter population includes cells expressing CD8 or CD4. Within the lymphocytes, the subset expressing CD4 are defined as T lymphocytes (Figure 2.2b). The $CD4^+$ T lymphocyte subset can be further divided into regulatory and non-regulatory cells. Regulatory cells represent a low-frequency subset which has the highest CD25 expression compared to other resting cells, and which expresses no or very low level of CD127. Regulatory T cells can be defined more precisely by the intracellular FOXP3 transcription factor, which is constitutively expressed only in regulatory T cells. Non-regulatory T cells represent a larger proportion of T lymphocytes, they express more CD127 and less CD25 than regulatory T cells (Figure 2.2c). Non-regulatory T cells can be further divided into naive and memory subsets (Figure 2.2d). Upon anti-

gen presentation, naive cells are activated and differentiate into effector cells, some of which further differentiate into memory cells while the remainder die. As part of the transition process from naive to memory, the cell surface protein CD45RA is lost so that consequently naive cells have higher CD45RA expression than memory cells. A further difference between these subsets is that memory cells tend to have a higher CD25 expression than naive cells. Since CD25 expression on the naive cells is low, with only a subset of the cells expressing substantial levels of the molecule, Dendrou et al (2009b); Pekalski et al (2013) define a threshold above which naive cells are deemed positive for CD25.

Following this manual gating strategy, two T cell phenotypes, percent of $CD25^+$ naive cells over total naive cell count and normalised fluorescence intensity of CD25 on memory cells, were found to be associated with rs2104286 and rs12722495 respectively. The percent of $CD25^+$ naive cells and percent of memory cells, were found to be associated with age and marginally associated with sex. The repeatability was also tested thanks to the 15 recalled individuals. Repeatability is an important factor to take into consideration, because reduced within-individual variation increases the power to detect between-individual variation. The repeatability and association results, as reported by Dendrou et al (2009b), are summarised in Table 2.4.

$CD4^+$ T Cell Subset	Phenotype	Repeatability (r^2)	Genetic Effect	Age Effect	Sex Effect
$CD25^+$ Naive	Percentage	0.669	$\downarrow$ rs2104286 $P = 4.25 \times 10^{-6}$	$\uparrow$ $P = 2.22 \times 10^{-9}$	M < F $P = 0.005$
Memory	CD25 MEF	0.997	$\uparrow$ rs12722495 $P = 1.16 \times 10^{-10}$	None	None
	Percentage	0.862	None	$\uparrow$ $P = 8.97 \times 10^{-5}$	None

Table 2.4. Repeatability and significance of effects of percentage of naive $CD25^+$ CD25 MEF and percentage of memory cell phenotypes. Subset of results from Dendrou et al (2009b) for cell populations under re-analysis in this chapter. r^2 is the Pearson correlation squared.

Besides the genetic and environmental factors driving variation in these cell phenotypes, there are two important sources of technical variation which need to be controlled for as they can have some bearing on the repeatability and association statistics: instrument variation over time and the subjectivity of manual gating.

Over the seven month period of the experiment during which samples were collected and analysed, instrument variation is detectable in the CD25 MFI of the memory cell population (Figure 2.3). The influence of this time effect on repeatability was first discussed by Dendrou et al (2009a). To correct for this, Dendrou et al (2009b) run fluorescent six-peak beads daily in order to define a normalisation transform from MFI to molecules of equivalent fluorochrome (MEF). Concretely, the transformation is defined by a linear regression of the observed MFI of the six, manually gated, bead populations, against their expected MEF, as specified by the bead manufacturer. In the first section of the chapter, I will show that this process can be automated by computationally gating beads using my `R` package flowBeads (Pontikos, 2013).

Manual gating, besides being laborious on large datasets, is an inherently subjective task as it relies on the opinion of the gater and can lead to between gater variation (Maecker et al, 2005). Although there is considerable interest in standardising and automating the gating process (Aghaeepour et al, 2013), some customisation is often necessary since some gating steps remains experiment and context specific. For example, when deciding on the position of the $CD25^+$ gate on the naive cell subset (Figure 2.2d), the manual gater may rely on fluorescent bead information but also on external experimental information, such as an isotype control. Nonetheless, I will show, in the second section of this chapter, that bead data is sufficient in order to develop a computational method which can closely emulate the $CD25^+$ manual gating.

In the third section of this chapter, I will look at automating the univariate gating on CD45RA in order to identify memory cells. From this gate, we obtain the percentage

of memory cells and the CD25 MFI of memory cells. As the CD45RA distribution is typically bimodal, the manual gating of CD45RA into negative (memory) and positive (naive) subsets, translates well into a clustering problem, which I will attempt to solve by fitting a two-component mixture model.

For both the CD25 and CD45RA gating, I will assess their performance in terms of their repeatability. Finally, in the fourth section of this chapter, I will test the association of the cell phenoytpes, percentage of $CD25^+$ naive cells, percentage of memory cells and CD25 MEF of memory cells, obtained using these computational methods, in order to assess the influence of these automatic gating approaches on effect size estimation.

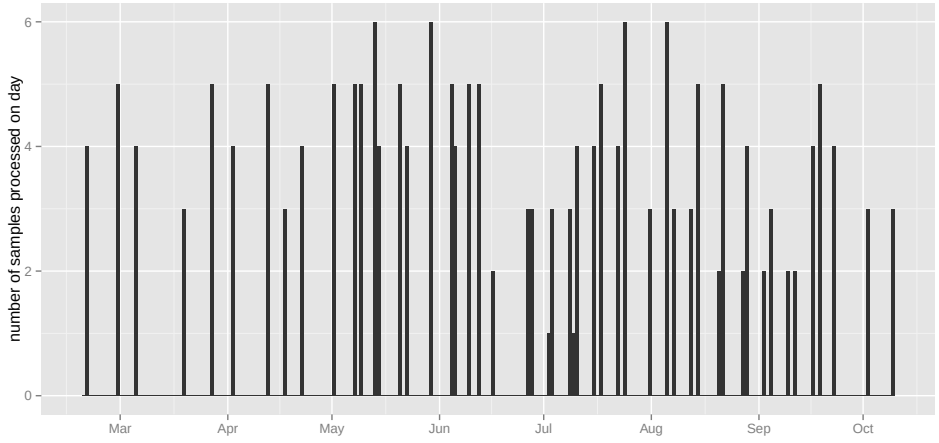
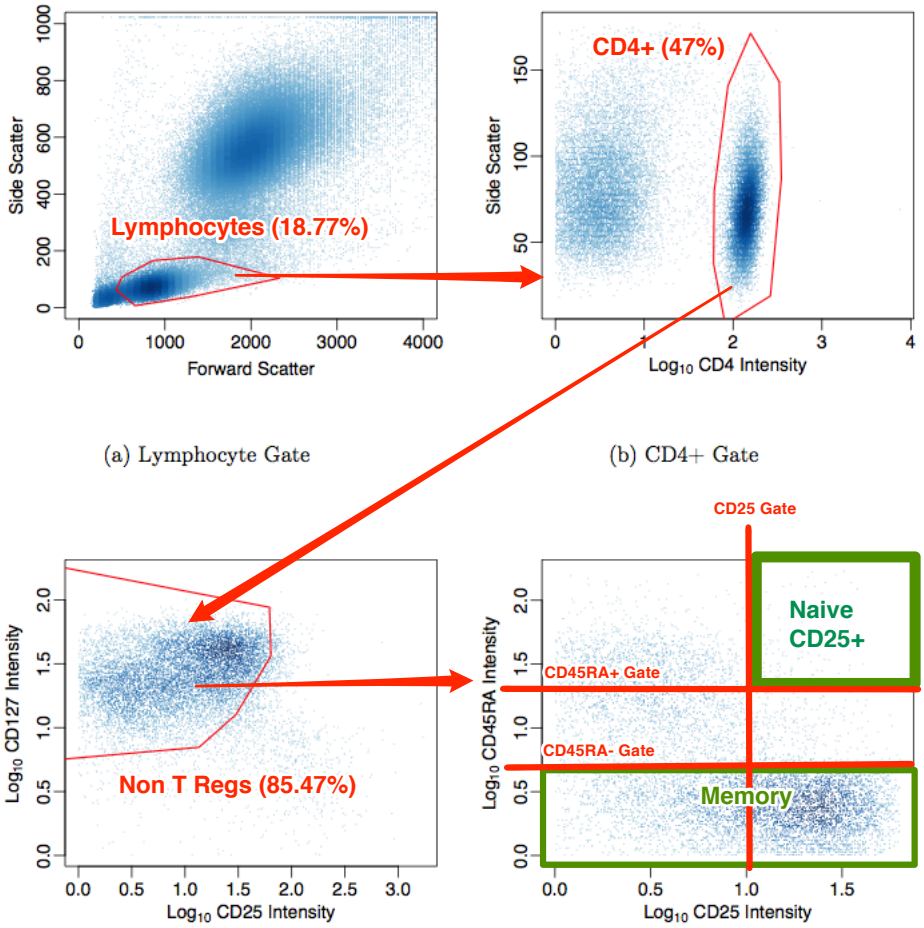


Figure 2.1. Number of samples analysed per day. A total of 195 (180 + 15 repeats) samples were analysed over seven months (from March to October). During that period, samples were analysed on 51 days, with between one and six samples analysed each day.



(a) Lymphocyte Gate (b) CD4+ Gate

(c) Non T Regs Gate (d) Naive/Memory Gate

Figure 2.2. Manual gating strategy followed by Dendrou et al (2009b). Manual gating strategy to extract memory T cells and CD25$^+$ naive T cells (green boxes). Note that the CD45RA gates exclude cells which are considered to be neither memory nor naive. Our automated gating replaces the final stage of the manual gating on CD25 and CD45RA.

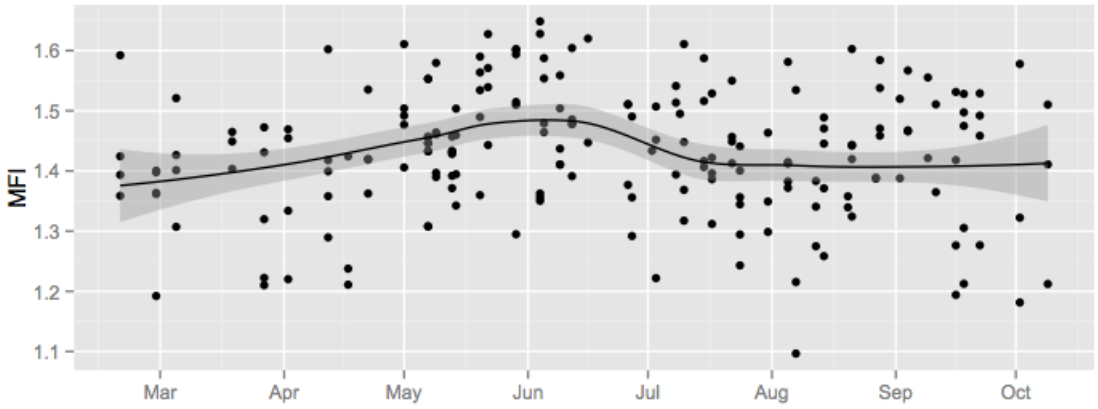


Figure 2.3. Effect of time on CD25 MFI in memory cells. CD25 MFI of memory cell population (manually gated) over time of experiment. The black line represents the loess regression line. Samples analysed after July tend to have a lower CD25 MFI than those analysed before.

2.2 Univariate clustering of bead data

In flow cytometry, a method of normalising fluorescence intensity to account for instrument variation, is to convert the MFI measured on a population to MEF (Schwartz et al, 1996; Dendrou et al, 2009a). In order to apply this conversion, specially designed beads of known and (assumed) constant fluorescence defined in terms of MEF, are used as a reference. The MEF property of these beads is deemed stable whereas the MFI of the bead population is dependent on the instrument and varies over time.

The beads used here are specially manufactured so that they belong to six distinct populations of increasing MEF as shown in Table 2.5. Following the bead manufacturer's guidelines, plotting the $\log_{10}(MEF)$ of these six bead populations against the corresponding calculated $\log_{10}(MFI)$ from the gated bead populations, we fit the linear regression:

$$\log_{10}(\text{MEF}) = \beta \times \log_{10}(\text{MFI}) + \alpha \tag{2.1}$$

The MEF is in fact a power transform of the MFI (only defined for strictly positive MFI values):

$$\text{MEF} = 10^{\alpha} \times \text{MFI}^{\beta}$$

The original MEF transform used by Dendrou et al (2009a) assumes that $\beta = 1$, although I relax that assumption.

In calculating the slope β and the intercept α parameters of the linear model, only the five brighest bead populations are used because the MEF of the blank beads is not specified by the manufacturer. However, as we will see in the next section, the blank beads can be used to define a threshold for positivity.

Typically bead data are gated manually. Here, in order to obtain the parameters of

the MEF transform, I will use an automatic process to gate the beads.

Since all beads are manufactured to be of identical dimensions, we expect a single cluster in the scatter channels: the singlet bead population. Events which lie away from the singlet population are deemed to be beads clumped together or debris and so are discarded. Filtering of singlets can be achieved by fitting a bivariate normal distribution on forward and side scatter and only keeping points within the 95th percentile. Having gated the singlets, I subset the data and proceed to gate on the fluorescence channels to identify the six bead populations. Given that the number of bead populations is known, that the bead signal is sufficiently clear and that the number of events is small (in the order to 10,000), I use the K-medoids algorithm. The solution has been implemented in the `R` package flowBeads (Pontikos, 2013), available on BioConductor. Automatic gating shows near perfect agreement with manual gating (Figure 2.5). Applying the bead normalisation to the memory CD25 MFI from Figure 2.3, we improve on the repeatability of that cell phenotype from $r^2 = 0.972$ to $r^2 = 0.985$, where r^2 is the Pearson correlation squared (Figure 2.6).

Population	FITC	RPE	REP-Cy5	**APC**	PE-Texas Red
1	B	B	B	**B**	B
2	2,500	1,500	750	**4,100**	552
3	6,500	4,400	2,100	**10,300**	2,014
4	19,000	14,500	6900	**25,500**	6,975
5	55,000	43,800	22,100	**67,300**	20,685
6	150,000	131,200	77,100	**139,100**	71,888

Table 2.5. FluoroSpheres from DakoCytomation. The Molecules of Equivalent Fluorochromes (MEF) values for the six bead populations as provided by the manufacturer. B denote the blank beads which by design contain no fluorochrome. Of the six fluorochromes contained by each bead only APC is used in the experiment.

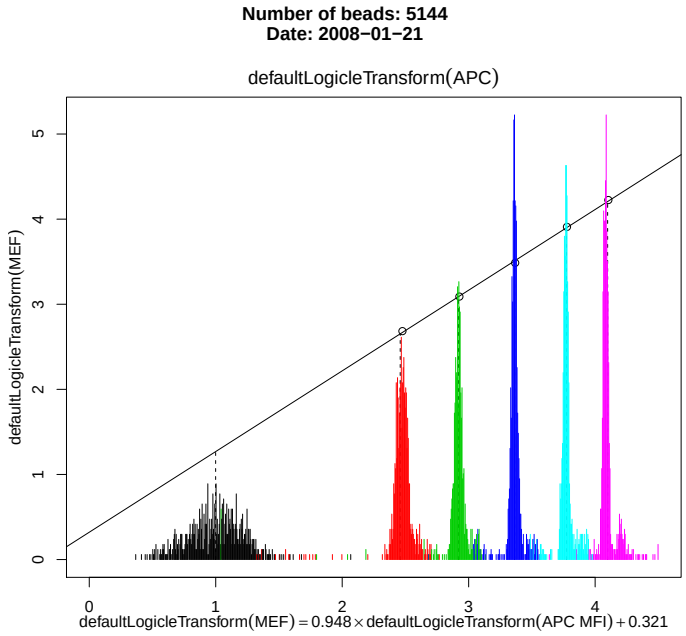


Figure 2.4. Linear regression of bead APC MEF against the APC MFI as defined in Table 2.5. The six peaks represent the six bead populations. These types of plots are generated automatically by the R package flowBeads (Pontikos, 2013).

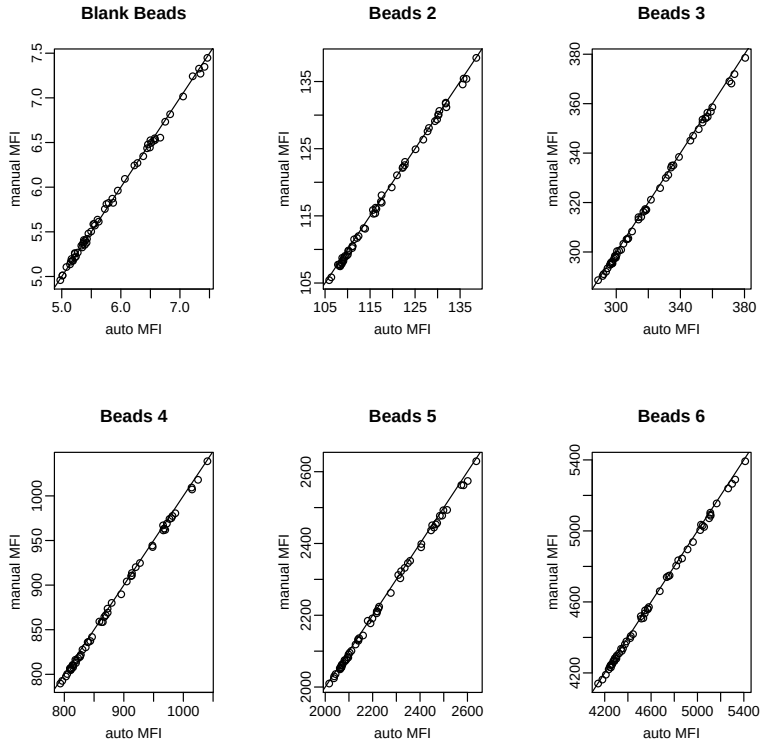
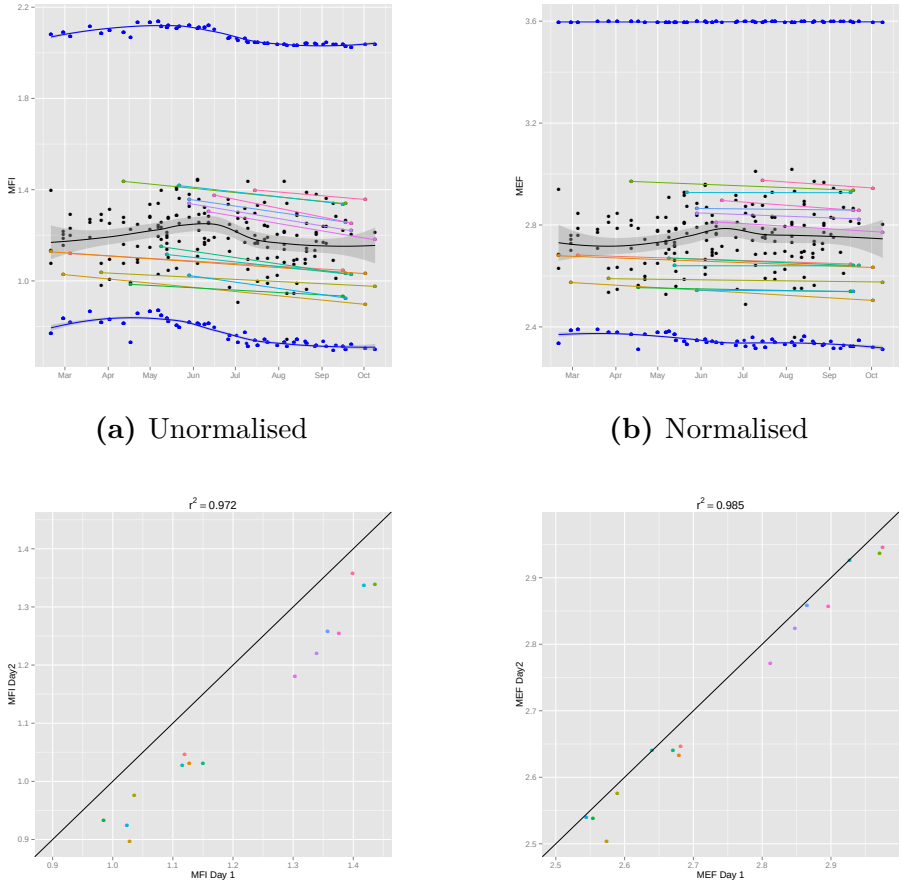


Figure 2.5. Comparison of bead population MFI using manual and `flowBeads` gating. There is good agreement of the APC MFIs of the six bead populations identified with manual and using the automatic approach.



(a) Unormalised

(b) Normalised

(c) Unormalised: $r^2 = 0.972$

(d) Normalised: $r^2 = 0.985$

Figure 2.6. Bead normalisation partially corrects for long term time effect in CD25 MFI of the memory cell population. In **(a)** and **(b)**, the blue points represent the CD25 MFIs of the two lowest bead populations, in black the CD25 MFIs of the memory cell populations. The dashed blue lines represent the overall mean of each the two bead populations. A loess regression is fitted to the MFIs of the beads and memory cells to illustrate the MFI variation over days. The points joined by lines are memory cell CD25 MFIs from the 15 recalled individuals (Table 2.1). The bead normalisation transform Equation (2.1) improves the repeatability of the MFI in recalled individuals from $r^2 = 0.972$ **(c)** to $r^2 = 0.985$ **(d)**.

2.3 Univariate gating on CD25: defining a CD25$^+$ threshold on naive cells

The approach adopted by the manual method is to define a threshold above which cells are considered positive for CD25. According to Dendrou et al (2009b) (and Calliope Dendrou personal communication), the CD25$^+$ threshold is set manually using an ad-hoc process based on an isotype control, bead data and ultimately a judgement call by the manual gater. An isotype control is a sample stained with the same fluorochrome (APC) but conjugated to a non-specific antibody not designed to target the marker we are interested in quantifying. It is used as a technique for assessing background APC fluorescence not resulting from CD25 binding.

This manual approach to setting the threshold, leads to a different gate position per sample per day (Figure 2.7). We notice that on some days there is greater variability in the positions of the gates. Also, the gate position moves down with time, reflecting the same downwards time-trend in the position of the gates observed in Figure 2.3.

Drawbacks of the manual approach are its lack of consistency and its reliance on isotype controls. The gating criterion is subject to human judgement and so may not be consistent across samples. Isotype controls are costly since part of the sample and fluorochromes is consumed for control purposes, consequently they are not always analysed. Also they are not necessarily an accurate measure of background fluorescence since they are also a source of noise linked to differences in the constitution of the control sample, the behaviour of the staining and other sources of technical variation (O'Gorman and Thomas, 1999; Maecker and Trotter, 2006).

I wished to improve on this process by using a more consistent and economical approach, using only beads, which I called `beads.thresh`. Instead of using isotype data, my working hypothesis, was that blank beads would constitute a more stable reference,

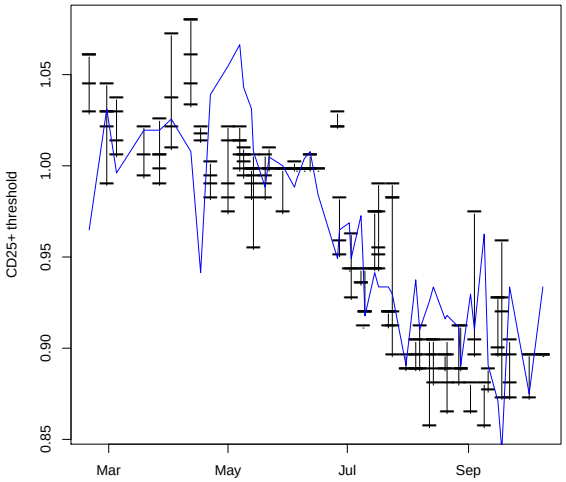


Figure 2.7. Position of the CD25$^+$ gate over duration of experiment. The black horizontal dashes are the positions of the manual CD25$^+$ gates for all 196 samples over the time course of the experiment (51 days). The vertical black lines represent the days and so define the range of the manual gate positions on a given day. The blue line represents our automatic CD25$^+$ gate which corresponds to the 86^{th} percentile of the blank bead population.

which could be used to define an APC-CD25 threshold. To find a suitable bead-derived threshold, I first gated the blank beads using my `R` package flowBeads (Pontikos, 2013), then I searched for the APC percentile of the blank bead population which best agreed with the manual gate (Figure 2.8). I found that in this dataset, the 86^{th} percentile of the blank bead population, best matched the manual gate position.

Hence the CD25$^+$ threshold defined by my approach, `beads.thresh`, is set as the 86^{th} percentile of the automatically gated blank bead population on that day. As we only have one bead set per day, we have a single fixed CD25 gate for all samples on that day (Figure 2.3).

The `beads.thresh` method for setting CD25 thresholds shows improved repeatability of the percentage of CD25$^+$ naive T cell phenotype over manual (Figure 2.10).

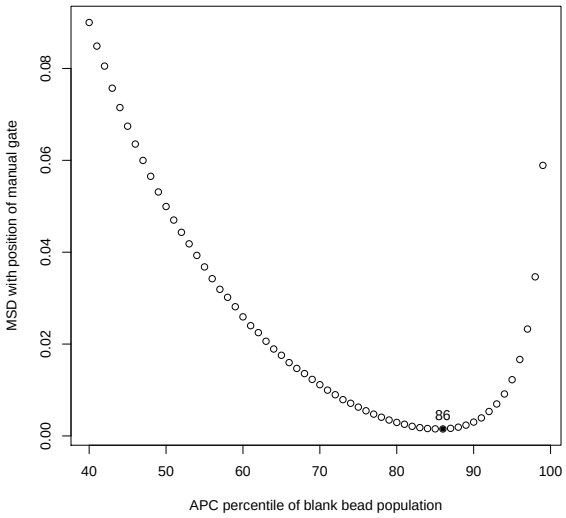


Figure 2.8. Mean square difference (MSD) of the position of the manual gate with that of `beads.thresh`. On the x axis, the APC-CD25 percentiles of the blank bead population from 40 to 99. On the y axis, the mean squared difference between the position of the manual gate and that of the bead-derived gate for that percentile threshold. The 86^{th} percentile yields the lowest mean squared difference hence the best agreement with the manual gating. The automatic threshold selection method (`beads.thresh`) is therefore defined as the APC 86^{th} percentile of the blank beads population.

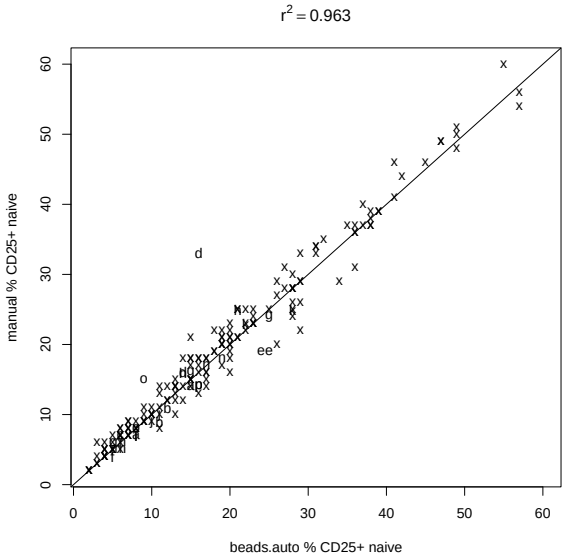


Figure 2.9. Agreement with manual of percentage of CD25$^+$ naive cell phenotype. Except for individual d, the agreement of `beads.thresh` with manual for percentage of CD25$^+$ naive cells is very good. r^2 is the Pearson correlation squared.

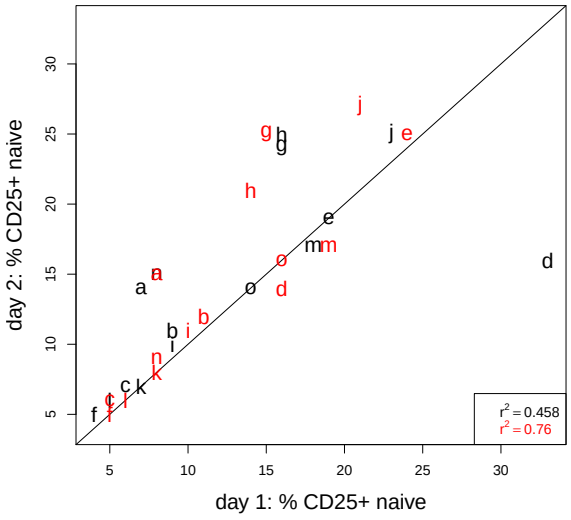


Figure 2.10. Repeatability of percentage of naive CD25$^+$ with manual (black) and `beads.thresh` (red). Repeatability of the percentage of naive cells which are CD25$^+$ from day one to day two. The overall repeatability of this cell phenotype was better with `beads.thresh` (red) than the manual (black). Letters are used to identify indidividuals, see Table 2.1. The Pearson correlation squared is $r^2 = 0.458$ for manual and $r^2 = 0.76$ for `beads.thresh`.

2.4 Univariate gating on CD45RA: fitting two-component mixtures on non-regulatory T cells

Non-regulatory $CD4^+$ T cells appear bimodal with respect to CD45RA expression, because this marker is lost upon activation of naive cells ($CD45RA^+$) to memory cells ($CD45RA^-$). In this section, we will model the CD45RA distribution by fitting a two component mixture model. Although we model both populations, we will only gate the memory ($CD45RA^-$) cell population, which corresponds to the first component, since the naive ($CD45RA^+$) cell population, the second component, is not a terminal gate as it is further divided into CD25 negative and positive subsets (see previous section).

In order to model the bimodal CD45RA distribution, I use the `R` function `normalmixEM` function in the `R` package mixtools (Young et al, 2009a), which provides an implementation of the EM algorithm (Dempster et al, 1977). The parameters, mean, variance and component weight, of the two-component Gaussian mixture model, are first initialised by the K-medoids algorithm. The parameter estimates are then obtained by running the EM algorithm until convergence. I call this method `mm`.

2.4.1 Using the mixing proportions of the mixture model

Since we are fitting a mixture model, instead of emulating manual gating by picking a threshold, I will first try a more statistically intuitive approach, using the mixing proportions obtained from `mm`. Additionally to the `mm` approach, I will also apply a more flexible mixture model of semi-parametric symmetric distributions (`R` function `spEMsymloc`) again from the `R` package mixtools (Young et al, 2009a), which I will call `spmm`. Semiparametric symmetric distributions are kernel density estimates centered around a location parameter.

Comparing the percent of memory cell obtained using `mm` and `spmm` to those ob-

tained using manual (Figure 2.11), we see that although there is agreement between the methods, the automatic methods tend to underestimate the percentage of memory cells. Also with regards to repeatability, `mm` and `spmm`, yield worse repeatability than manual (Figure 2.12).

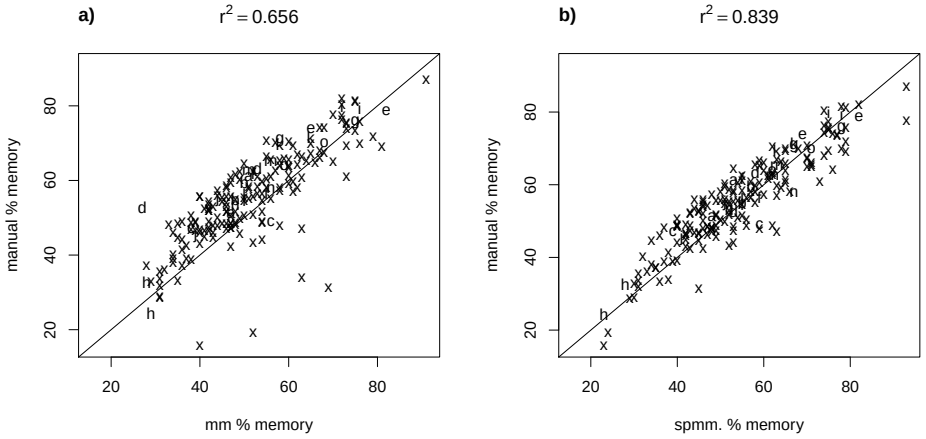


Figure 2.11. Agreement with manual of the percent memory cell phenotype obtained with `mm` (a) and `spmmm` (b). The more flexible model, `spmm`, tends to agree better with percent memory cell phenotype returned by manual, than `mm`. Again we can see in (a), that `mm` underestimates the percentage of memory cells in individual d compared to manual. r^2 is the Pearson correlation squared.

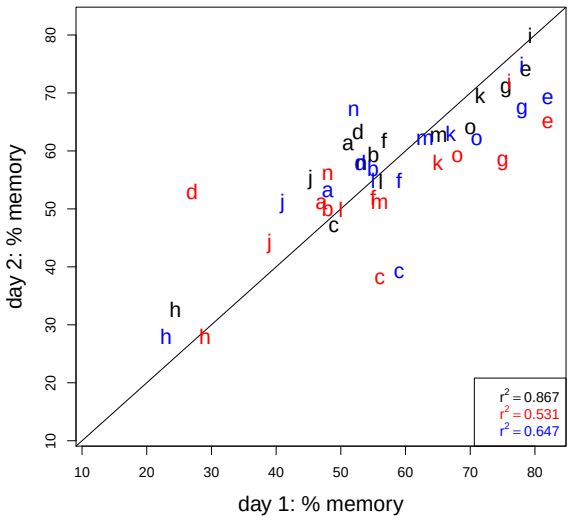


Figure 2.12. Repeatability of the percent memory cell phenotype with manual (black), `mm` (red) and `spmm` (blue). While manual still shows the best repeatability, the repeatability of the automatic methods is quite encouraging, given these have no knowledge of the manual gates and are based purely data driven. Also as seen using the thresholding methods, individual d is a clear outlier when gated with the `mm` method.

2.4.2 Emulating manual gating by picking a threshold

Although the previous approach makes sense from a statistical perspective, it does not exclude the transitional cell population which lie between the memory and naive cells. Instead in the manual gating, the memory cell subset is defined by a threshold on the bimodal CD45RA distribution, below which cells are regarded as CD45RA$^-$. Here, I attempt to emulate manual gating by defining a threshold on the fitted two-component mixture model, `mm`. I consider two approaches of selecting a threshold, `pct.thresh` and `post.thresh`, both which are illustrated in Figure 2.13.

The first method, `pct.thresh`, closely replicates the manual CD45RA$^-$ gating procedure, as explained to me by Linda Wicker. In this approach, only the shape of the first left-most, component of the mixture model defines the position of the CD45RA$^-$ memory gate. In order to delineate the memory population, we first identify the first

peak of the bimodal CD45RA distribution, which should correspond to the peak of the first component, after the two-component mixture has been fitted. Then, following the CD45RA density curve from the peak towards the left-hand-side, we record the CD45RA value after which the density curve drops below a certain given threshold. This CD45RA value is then mirrored to the right-hand-side of the peak in order to define the CD45RA-threshold. This technique is in fact equivalent to selecting a fixed percentile threshold for the first component to gate consistently across all samples.

The second method, `post.thresh`, considers the density ratio of both components in order to decide where to draw a threshold. Formally, `post.thresh` selects a threshold on the posterior probability of belonging to the first component, the memory population, across all samples. At a given point, the posterior probability of belonging to the first component is defined as the ratio of the density of the first component, over that of the total density. Concretely, given a two-component mixture model where, f_1 is the density of the first component and f_2 the density of the second, and a posterior threshold of p, then a point x is assigned to component 1 provided that:

$$f_1(x) \geq f_2(x)\frac{p}{1-p}$$

For example, if the posterior probability threshold was $p = 95\%$, then for x to be assigned to the first component, $f_1(x)$ would need to be 19 times larger than $f_2(x)$.

Given these two thresholding approaches, I wish to select a threshold for `pct.thresh` and for `post.thresh`, which most closely matches the manual gating. To this purpose, I use the method described in the previous section (Figure 2.8), to find the threshold which minimises the mean square difference with the manual gate position. Applying this method, I find that the optimal threshold is the 88^{th} percentile for `pct.thresh`, and 89% for `post.thresh` (Figure 2.14). Also, I notice that for `post.thresh`, in certain samples, the posterior probability of belonging to the first component does not exceed

89% (Figure 2.15). This is why in Figure 2.14, we do not obtain points beyond a threshold of 89, because gates are missing for certain samples. This can be due to poor model fit (Figure 2.15a) or too much overlap between the memory and naive cell populations.

Using either approach, there is good agreement between the CD25 MEF values obtained for the memory population when gated using either the automated or the manual approach (Figure 2.16). This is to be expected as this cell phenotype is not very sensitive to the position of the CD45RA gate (Figure 2.17). Hence, for this phenotype, this translates to similar repeatability to that obtained with manual gating (Figure 2.18). On the other hand, the repeatability of the percentage of memory cells is very sensitive to the gate position.

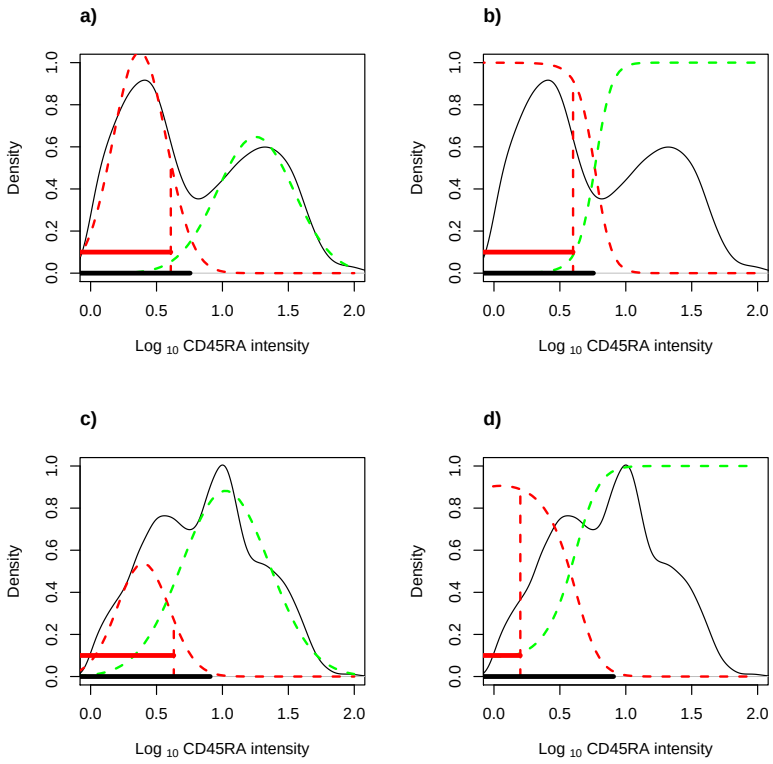


Figure 2.13. Example on individual d of the two approaches, `pct.thresh` (a and c) and `post.thresh` (b and d), of selecting a threshold. Individual d was chosen to illustrate `pct.thresh` and `post.thresh`, because the CD45RA distribution takes on a very different shape on day one (a and b) compared to day two (c and d). In (a) and (c), the `pct.thresh` method, places the gate at the 88th percentile of the first component. In (b) and (d), the `post.thresh` method, places the gate at the largest CD45RA value where the posterior of the first component reaches 89 percent. This poses a problem for `post.thresh` in (d) because the overlap of the components is such that the posterior is only reached close to zero which yields a much smaller gate and consequently a lower percent of memory cells (Figure 2.16d). On the other hand, while the two-component distribution is not a good fit to the data, this is less of an issue for `pct.thresh`, as can be seen in (c).

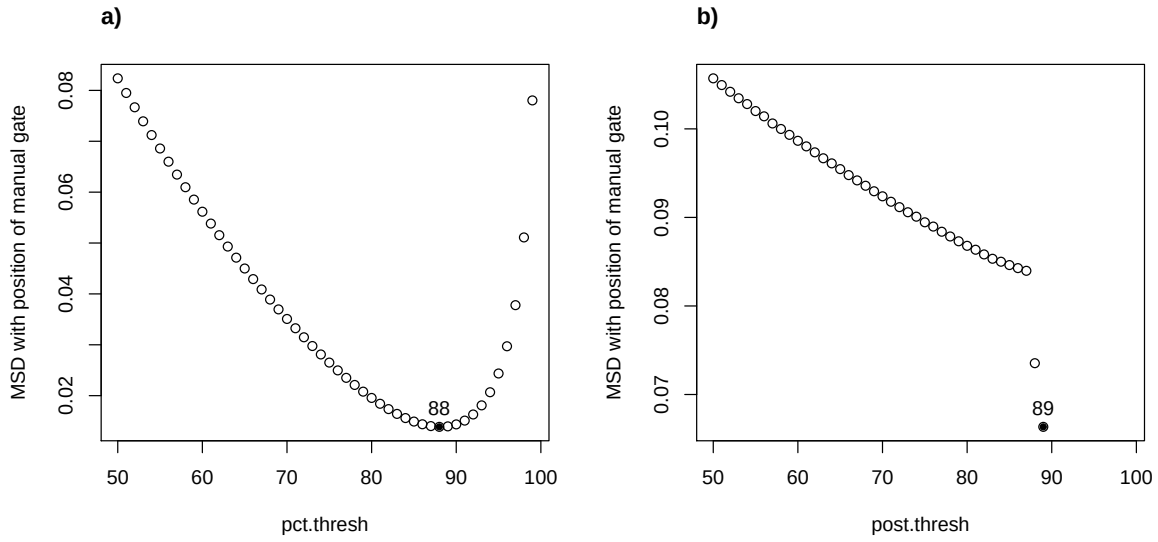


Figure 2.14. Mean square difference (MSD) of the position of the manual gate with that of `pct.thresh` (a) and `post.thresh` (b). The threshold which minimises the MSD is 88 for `pct.thresh` (a) and 89 for `post.thresh`. At that threshold, the `pct.thresh` (a) gate position matches better the manual than `post.thresh` (b). For `post.thresh`, the MSD is not defined for threshold larger than 89, because there are samples for which the posterior probability does not reach 89 percent (Figure 2.15e).

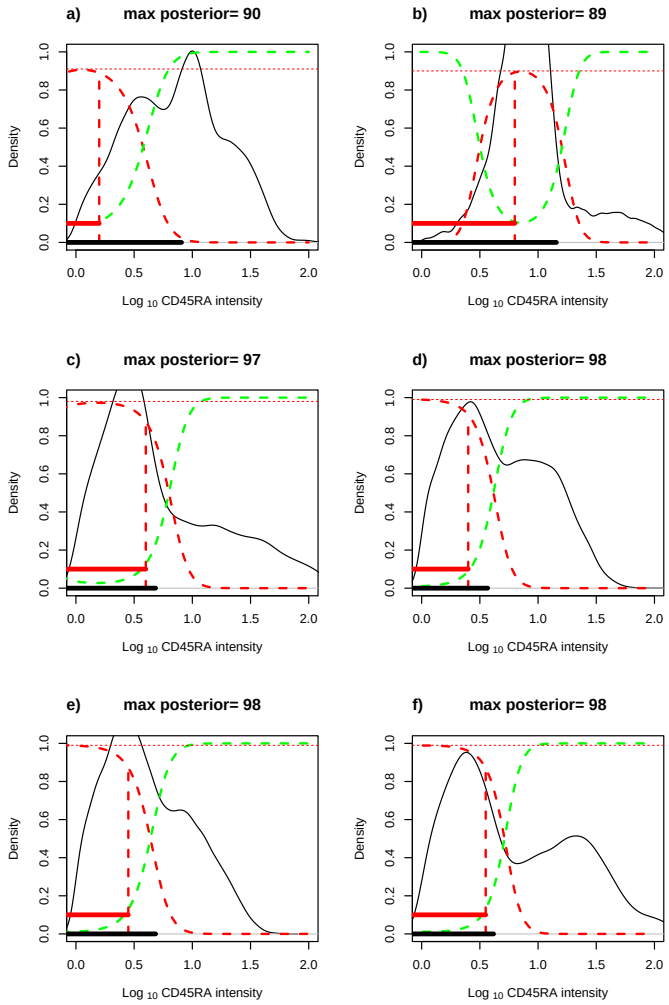


Figure 2.15. Samples for which the 99 maximum posterior probability is not reached. In black the manual gate. In red the `post.thresh` gate drawn at 89. The posterior probability does not reach 99 percent in these six samples. In a) this is because of poor model fit. In the others b), c) and d), this is due to the mixing of the two distributions. In b), the non-uniform decreasing posterior function, can be explained by the green distribution, component 2, being much wider than the red distribution.

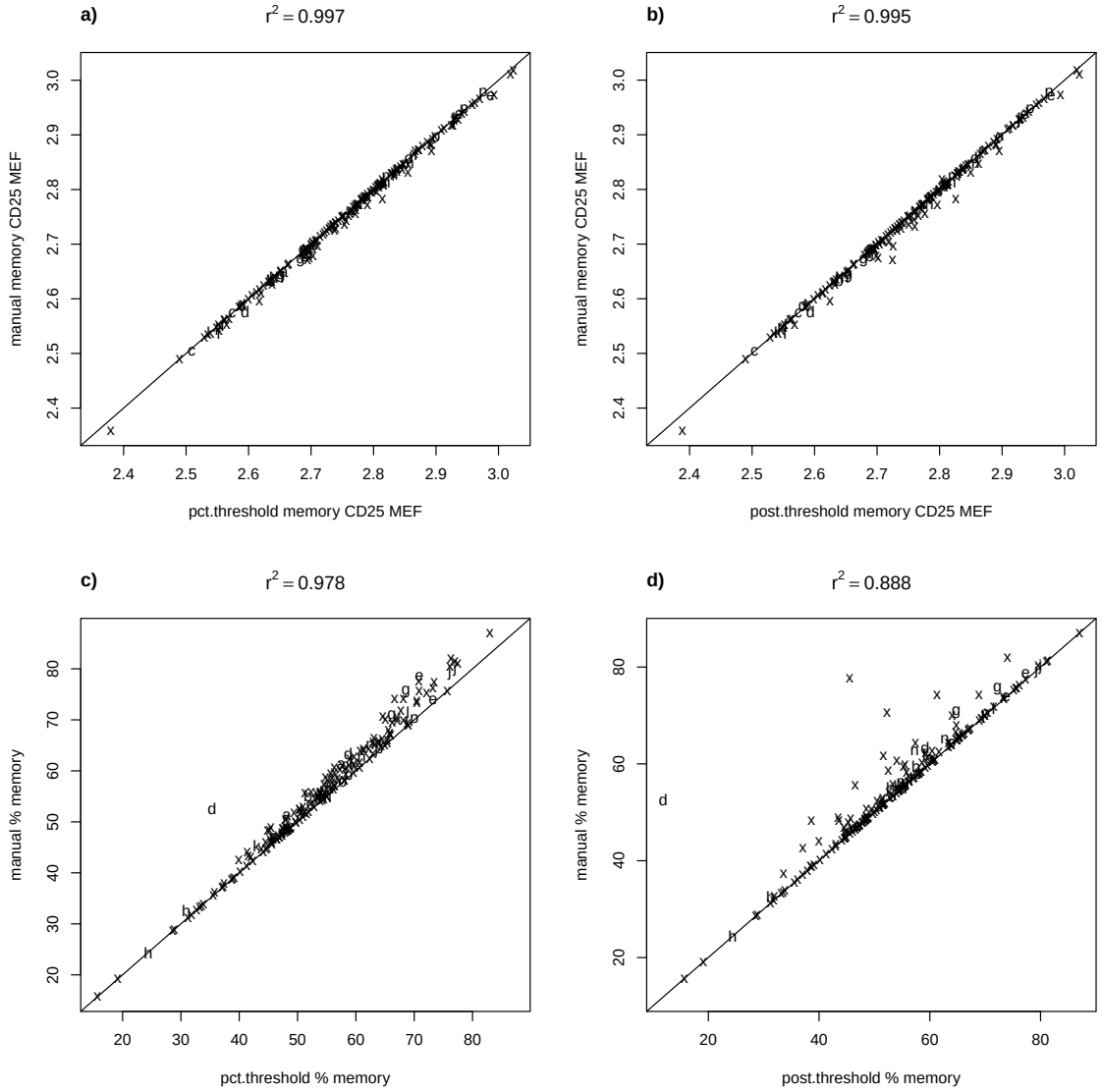


Figure 2.16. Agreement of memory CD25 MEF (a and b) and percentage of memory cells (c and d), obtained from `pct.thresh` and `post.thresh` with manual. The agreement of memory CD25 MEF is very close to manual (a and b) while the automatic methods tend to yield smaller memory cell percentages (c and d).

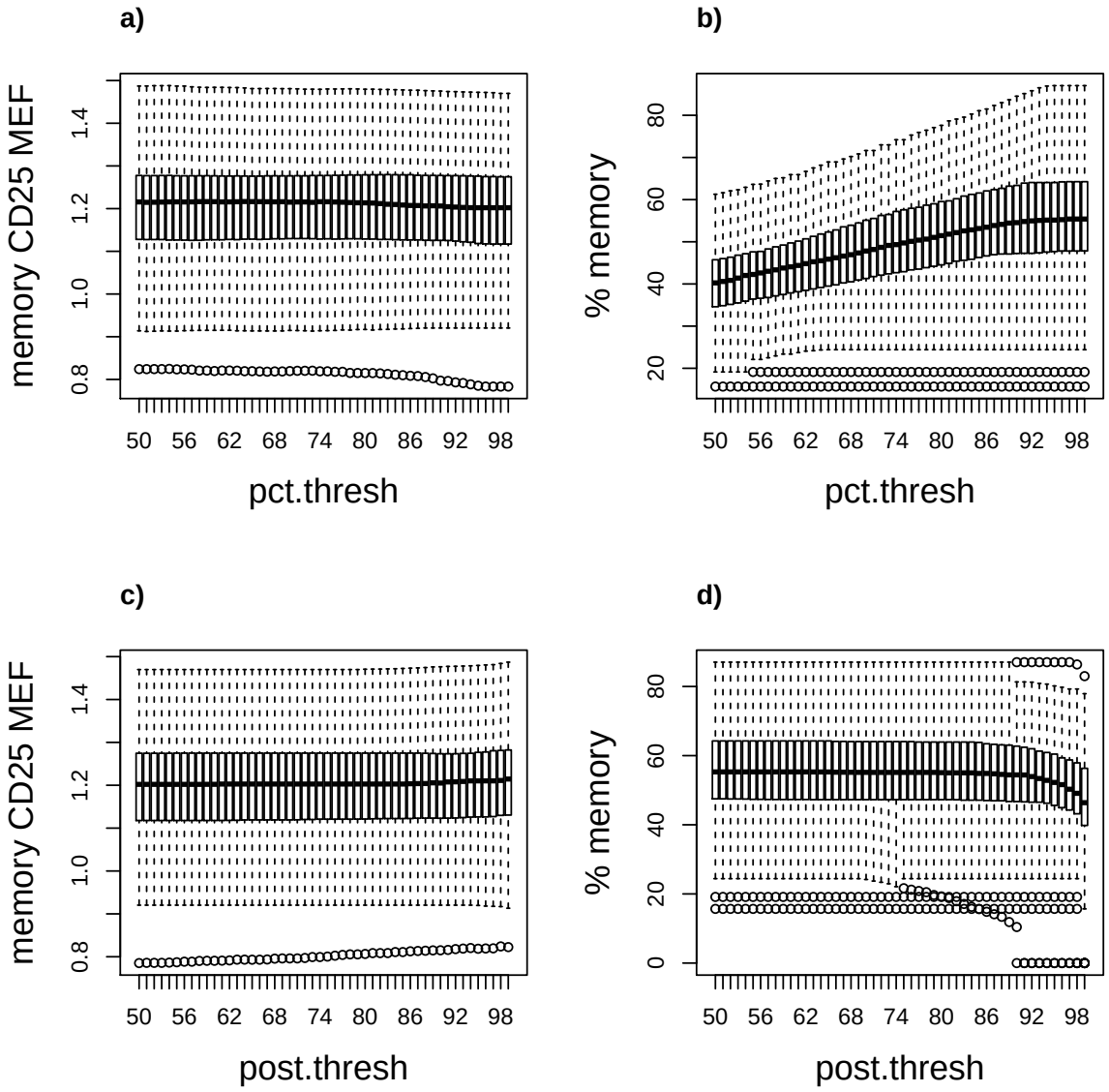


Figure 2.17. Influence of threshold for `pct.thresh` and `post.thresh` on distribution of memory CD25 MEF and percent memory cell phenotypes. Memory CD25 MEF is not sensitive to position of CD45RA gate (a and c) whereas percent memory is (b and d). On the other hand, the percent memory phenotype is more sensitive in particular when using the `pct.thresh` (c) method as opposed to the `post.thresh` (d).

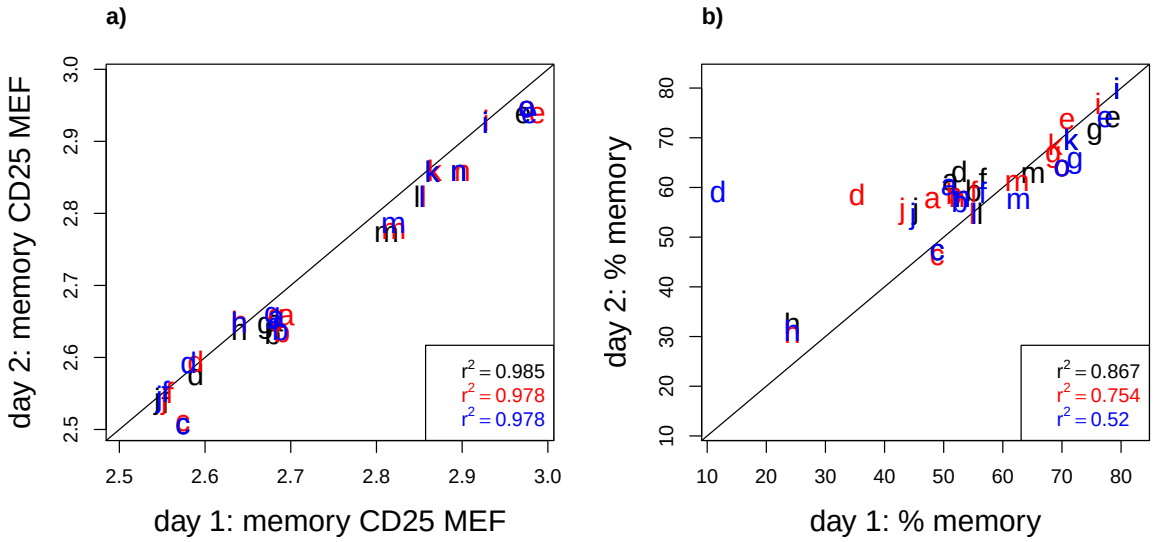


Figure 2.18. Repeatability of the memory cells phenotypes, CD25 MEF (a) and percentage (b), obtained from manual gating (black), `pct.thresh` (red) and `post.thresh` (blue). While the repeatability is very close for CD25 MEF ($r^2 = 0.985$ for manual, $r^2 = 0.978$ for `pct.thresh` and $r^2 = 0.978$ for `post.thresh`) it varies considerably for the percentage of memory cells ($r^2 = 0.867$ for manual, $r^2 = 0.754$ for `pct.thresh` and $r^2 = 0.52$ for `post.thresh`). Individual d is a clear outlier when gated with the `post.thresh` method (b)

2.5 Association tests

Having obtained cell phenotypes by different gating strategies, I would like to assess how these influence our association test statistics. Since our dataset contains 15 repeated cell phenotypes from recalled individuals, I accounted for those in my association testing by applying a linear mixed effects model with random intercept to allow for per individual effect. To that purpose I used the `R` function `lme` from the `R` package `nlme` (Pinheiro et al, 2014). Each covariate, genotype, age and sex, was tested separately, with an additive recessive model assumed for the SNP effect.

Overall, the association test with the percent $CD25^+$ naive cells phenotype yields similar effect sizes to manual (Table 2.6). A significant age and rs2104286 effect are reported using both manual and `beads.thresh` gating. However, the significance of the rs2104286 effect found with `beads.thresh`, is an order of magnitude less (10^{-4}) than with manual (10^{-5}). On the other hand, `beads.thresh` adds some evidence to the suggested association by Dendrou et al (2009b) of a sex effect on percentage of $CD25^+$ naive cells, whereby males have a lower percentage of naive $CD25^+$ than females, although the effect remains marginal.

Regarding the percentage memory cell phenotype, an age effect is also detected using automatic methods, `post.thresh` and `pct.thresh`, however the signifcance of the association is an order of magnitude less with `post.thresh` (pvalue 10^{-2}) than with manual and `pct.thresh` (pvalue 10^{-3}). This could be due to greater noise in the measurement as suggested by lower repeatability. Also, noteworthy, is that a marginally significant rs2104286 effect (pvalue=0.042322) is reported with `post.thresh`, which is not found with the other methods. However, on closer inspection, the association appears to be driven by the outlying sample from individual d (Figure 2.19).

For the memory CD25 MEF cell phenotype, the association results between manual and automatic are virtually identical (Table 2.7), which is to be expected, given this cell

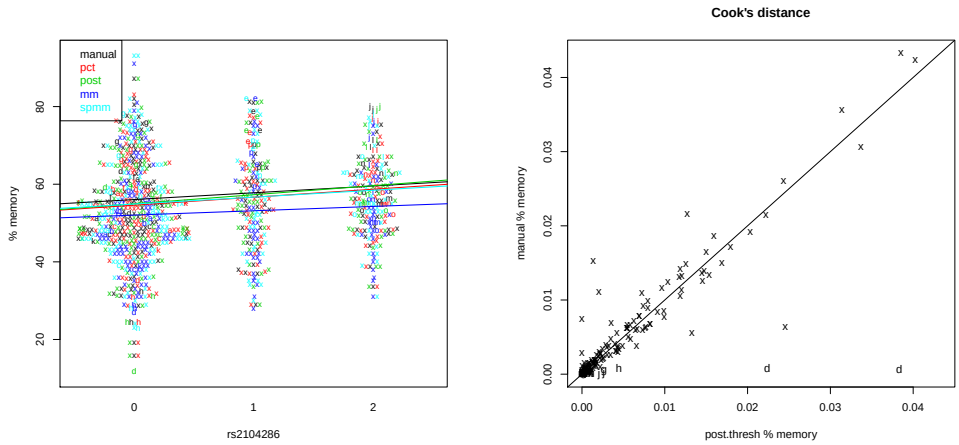
phenotype is largely unchanged by the CD45RA gate position (Figure 2.17).

rs12722495	effect	95%CI	p-value
manual	-2.479	[-5.422;0.463]	0.098145
beads.thresh	-1.509	[-4.453;1.435]	0.31326
rs2104286	effect	95%CI	p-value
manual	-4.714	[-6.894;-2.534]	3.2017e-05
beads.thresh	-4.39	[-6.569;-2.212]	0.0001014
rs11594656	effect	95%CI	p-value
manual	-1.459	[-3.924;1.006]	0.2443
beads.thresh	-1.328	[-3.774;1.118]	0.28531
Age	effect	95%CI	p-value
manual	0.475	[0.286;0.664]	1.6584e-06
beads.thresh	0.457	[0.269;0.645]	3.514e-06
Sex	effect	95%CI	p-value
manual	-4.216	[-7.856;-0.575]	0.023475
beads.thresh	-4.327	[-7.936;-0.718]	0.019046

Table 2.6. Genotype, age and sex effect sizes on percentage of CD25$^+$ cells. Effect of rs12722495, rs2104286, rs11594656, age and sex, on the percentage of CD25$^+$ naive cells.

rs12722495	effect	95%CI	p-value
manual	0.062	[0.036;0.088]	4.7176e-06
pct.thresh	0.06	[0.034;0.086]	9.0361e-06
post.thresh	0.061	[0.035;0.087]	6.822e-06
rs2104286	effect	95%CI	p-value
manual	0.014	[-0.007;0.035]	0.18022
pct.thresh	0.014	[-0.007;0.035]	0.19525
post.thresh	0.014	[-0.007;0.035]	0.18634
rs11594656	effect	95%CI	p-value
manual	0.012	[-0.011;0.035]	0.29945
pct.thresh	0.011	[-0.012;0.034]	0.34754
post.thresh	0.011	[-0.012;0.033]	0.35147
Age	effect	95%CI	p-value
manual	0.001	[-0.001;0.003]	0.43937
pct.thresh	0.001	[-0.001;0.003]	0.44201
post.thresh	0.001	[-0.001;0.003]	0.38071
Sex	effect	95%CI	p-value
manual	0	[-0.034;0.034]	0.9983
pct.thresh	0.001	[-0.033;0.035]	0.95584
post.thresh	0.002	[-0.032;0.036]	0.89534

Table 2.7. Memory CD25 MEF effect sizes. Effect of rs12722495, rs2104286, rs11594656, sex and age on memory CD25 MEF.



(a) **(b)**

Figure 2.19. Effect of rs2104286 on percent memory gated by manual (black), `post.thresh` (green) and `pct.thresh` (red). In Table 2.8, marginally significant association is detected with rs2104286. This is due to the leverage of individual d which stands out as an outlier (b).

rs12722495	effect	95%CI	p-value
manual	1.736	[-1.246;4.718]	0.25209
pct.thresh	2.326	[-0.408;5.06]	0.094952
post.thresh	2.624	[-0.26;5.508]	0.074263
mm	1.131	[-1.84;4.101]	0.45366
spmm	1.957	[-1.301;5.214]	0.23746
rs2104286	effect	95%CI	p-value
manual	1.75	[-0.544;4.045]	0.13399
pct.thresh	2.051	[-0.063;4.165]	0.057164
post.thresh	2.337	[0.082;4.593]	0.042322
mm	1.118	[-1.194;3.431]	0.34129
spmm	1.775	[-0.744;4.293]	0.16611
rs11594656	effect	95%CI	p-value
manual	-0.018	[-2.514;2.479]	0.98884
pct.thresh	0.269	[-2.048;2.585]	0.81927
post.thresh	0.402	[-2.092;2.896]	0.75075
mm	1.253	[-1.265;3.771]	0.32737
spmm	1.272	[-1.466;4.01]	0.36044
Age	effect	95%CI	p-value
manual	0.387	[0.192;0.582]	0.000127
pct.thresh	0.35	[0.169;0.531]	0.00018575
post.thresh	0.316	[0.119;0.513]	0.0018033
mm	0.431	[0.236;0.625]	2.0985e-05
spmm	0.534	[0.326;0.743]	1.0325e-06
Sex	effect	95%CI	p-value
manual	3.485	[-0.198;7.168]	0.063526
pct.thresh	2.826	[-0.594;6.246]	0.10471
post.thresh	2.811	[-0.86;6.481]	0.1325
mm	2.793	[-0.925;6.51]	0.13998
spmm	3.361	[-0.687;7.408]	0.10308

Table 2.8. Memory percentage effect sizes. Effect of rs12722495, rs2104286, rs11594656, sex and age on memory cell percentage.

2.6 Discussion

In this chapter, I have shown that bead data is readily gated by automatic methods and that the results are comparable to manual gating. Automatic gating of bead data is fast and automates other related tasks such as MFI to MEF transformation, and threshold selection.

Gating of biological data is more difficult as we have little prior knowledge of the sample we are analysing and the data is far more noisy. So far, I have developed two automatic univariate gating strategies:

- a bead defined threshold method on CD25 to identify $CD25^+$ naive cells
- a two component mixture model on CD45RA to identify memory cells ($CD45RA^-$)

My CD25 univariate gating method (`beads.thresh`) relies on defining a threshold based on automatically gated bead data. The value of the threshold is selected as the percentile of the blank bead population which mimimises the mean squared difference with manual gate positions. The percentage of naive $CD25^+$ cells phenotype identified with my approach showed better repeatability than manual (Figure 2.10). My approach defines one threshold for all samples gated on the same day, whereas the manual approach, relies on isotype controls and allows for different thresholds per day. Isotype controls should theoretically be an estimate of background but have been criticised for being an extra source of noise (O'Gorman and Thomas, 1999; Maecker and Trotter, 2006).

My CD45RA univariate gating method fits a specific model to the data: a mixture of two univariate distributions. The parameters of the model are estimated using an EM algorithm (Dempster et al, 1977) initialised with K-medoids.

In a first instance, I used the parameters estimated by the two-component mixture model. Specifically, I used the weight parameter of the first component as the percentage of memory cells phenotype. Although this seemed a sensible approach from the statistical

perspective of fitting a two-component mixture model, it does not match the biological perspective that transitional cells should be excluded.

Therefore, in the second instance, I attempted to emulate manual gating by defining a threshold. I tried two approaches of defining a threshold, `pct.thresh` which thresholds on the percentile of the first component of the fitted mixture model, and `post.thresh` which thresholds instead on the posterior probability of the first component. As with the `beads.thresh`, the value of the threshold is selected as the value which mimimises the mean square difference (MSD) with manual gate positions.

Two benchmarks were used to evaluate my univariate gating strategies: repeatability and comparison of the effects sizes obtained by Dendrou et al (2009b) using manual.

Repeatability is an independent measure which does not require comparison to other gating methods (such as manual). Unfortunately, given that in our data set only 15 samples are repeated, it is difficult to evaluate methods on such a small sample size. Moreover, good repeatability does not necessarily imply that the gating is unbiased but rather that the gating is consistent. Hence repeatability, needs to be complemented with some metric, in the form of manual gating or some prior biological knowledge, to assess whether the computed cell phenotypes are in a sensible range.

I have shown that the difference in the identification of cell phenotypes by different gating methods can influence the effect size estimates in association studies. In particular, outliers can have an important influence in relation to their leverage as seen in Figure 2.19. For example when testing association with age, outlier cell phenotypes from younger or older individuals have more leverage than ones closer to the mean. When testing for association with genotype, outlier cell phenotypes from rarer variants have more leverage than one from common variants.

Hence, if we are to deploy automatic gating techniques more generally, detection of outliers is crucial, to avoid false positive associations. In particular, we require outlier

detection metrics which do not only rely on the availability of repeated samples or manual gates. Already, we have seen that looking at the maximum posterior probability in a sample can give us some insight (Figure 2.15). Another metric of evaluating how well a model fits the data could be a cost function like the mean integrated square error (MISE).

When outliers have been detected, we may want to exclude or down-weight them, or extend the gating method to account for these. One simple way of modifying the method, could be to use the gate positions in non-outlier samples to influence that in outliers. This could motivate borrowing information from other samples, using for example a hierarchical Bayesian framework as was recently developed by Cron et al (2013), designed specifically to deal with rare cell populations which might not consistently be detectable across all samples.

However, one may argue that this approach, conceals rather than addresses the underlying problem of poor model fit. For example, as we see from the trimodal distribution in Figure 2.13, it may be more appropriate to fit a three component instead of a two component mixture model on this sample.

Chapter 3

Methods to assess cell response to ex-vivo stimulation in flow cytometry

In the previous chapter, I looked at normalisation of the MFI using beads and replicating two univariate gating steps with thresholding on CD25 or with a mixture of two components on CD45RA. Here on a different dataset on which only preliminary manual analysis has been done, I will also consider normalisation methods and replicating all steps of the manual gating with an automatic procedure. Furthermore, I will also consider approaches of discovering biologically relevant subsets not reported by the manual gating.

3.1 Background

Motivation Genomewide association studies have implicated the IL-2 signalling as an important aetiological pathway associated with the development of T1D. As seen in Chapter 2, the protective T1D associated *IL2RA* variant at rs12722495 in healthy

individuals predicts an increase in expression of CD25, the alpha chain of the trimeric IL-2 receptor, on memory CD4$^+$ T lymphocytes (Dendrou and Wicker, 2008; Dendrou et al, 2009b). Garg et al (2012) found that regulatory and memory CD4$^+$ T cells in healthy carriers of T1D risk associated *IL2RA* variants, also exhibit decreased sensitivity to IL-2 in terms of decreased MFI levels of phosphorylated signal transducer and activator of transcription 5 protein (pSTAT5), STAT5 dimerises or tetramerises on phosphorylation and acts as a transcription factor (Figure 3.1), which also induces the transcription of *FOXP3*, a transcription factor characteristic of regulatory T cells (Tregs).

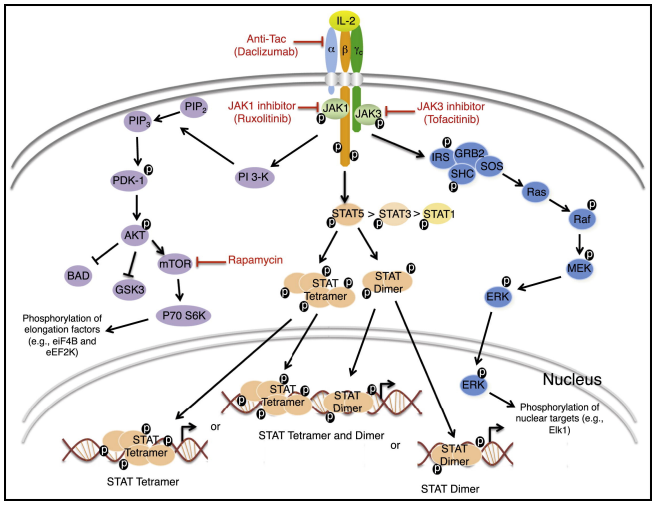


Figure 3.1. Schematic of major IL-2 signaling pathway (taken from Liao et al (2013)). CD25 is IL2RA, the α subunit of the trimeric IL-2 receptor. The α chain is the highest binding affinity receptor of the three chains. STAT5 is phosphorylated to pSTAT5 and acts as a transcription factor.

Long et al (2011) have reported that a T1D associated variant at rs1893217 of the protein tyrosine phosphatase N2 gene (*PTPN2*), a negative regulator of the IL-2 pathway, also correlates with lower STAT5 phosphorylation. Furthermore, it is suspected that IL-2 production might be diminished in T1D since disease associated *IL2RA* variants correlate with reduced CD25 levels and reduced IL-2 production on activated CD69$^+$ CD4$^+$ memory T cells after antigen stimulation (Dendrou et al, 2009b). Long-term reduced sensitivity to IL-2 also correlates with diminished maintenance of FOXP3 ex-

pression in the $CD4^+$ $CD25^+$ regulatory T-cells of type 1 diabetic subjects (Long et al, 2010). Hence, these findings appear to consolidate the hypothesis that type 1 diabetics tend to have a reduced ability to respond to IL-2 in part due to genetic defects in *PTPN2*, *IL2RA* and possibly other gene variants involved in the IL-2 signalling pathway (Long et al, 2010, 2011, 2012). These results are of great clinical relevance to us, since our lab is currently conducting an adaptive study of IL-2 dose on regulatory T cells in type 1 diabetes (`http://www.clinical-trials-type1-diabetes.com/`), and to all researchers interested in low-dose IL-2 therapy to autoimmune diseases (Koreth et al, 2011; Saadoun et al, 2011).

However some concerns have been raised with these studies. One concern was that the Tregs discrimination was not particulary thorough. Long et al (2010) define Tregs based only on two markers $CD4^+$ $CD25^+$ whereas these cells are usually also defined on CD127 and FOXP3. Another a notable omission by Long et al (2010) was the lack of repeated samples to assess the within-individual variance or reliability of the assay. Hence, Tony Cutler, in our lab, set to find if he could replicate some of these findings in an independent cohort using a more refined gating stategy by including the FOXP3 regulatory T cell marker, as well as NK cell markers, CD3, CD8, and CD56, to discover whether other potential cell subsets are also sensitive to IL-2 (Table 3.2).

Samples and panels He selected 22 long-standing diabetics (6 males and 16 females, mean age 29) and 28 controls (mean age 27) from the Cambridge Bioresource, as well as 30 newly diagnosed (20 males and 9 females, mean age 11.7) and 15 unaffected siblings (5 males and 12 females, mean age 12.3) from the Diabetes - Genes, Autoimmunity and Prevention (D-GAP) resource. To guard from false positive association and non-reproducible results, it is important to ascertain the repeatability of these cell phenotypes before conducting any form of association study. In order to test the repeatability, ten individuals were recalled for a second blood sample (Table 3.1). Hence a total of 52 cases

and 43 controls, of which, 10 individuals (5 cases and 5 controls) from the Cambridge Bioresource were recalled to assess reproducibility of the phenotypes.

Individual	status	pch	number of days between visits
1	control	a	98
2	case	b	140
3	control	c	167
4	control	d	98
5	case	e	167
6	case	f	112
7	control	g	112
8	case	h	98
9	case	i	120
10	control	j	140

Table 3.1. Ten individuals recalled between 98 and 168 days later to assess stability of the cell phenotypes. pch is the plotting character used to refer to these individuals in plots later in this chapter.

Blood samples were prepared and analysed by flow cytometry on day of collection. Each sample was split into four aliquots of 500 µl. The first aliquot was left unstimulated. The remaining three were stimulated ex-vivo for 30 minutes at four 100-fold increasing doses of proleukin (a polymer of IL-2) at 0.1, 10 and 1000 $\mathrm{U\,ml^{-1}}$ respectively. After a set stimulation time, the samples were fixed, permeabilised and stained, with different panels (Table 3.2), on a set of core markers, not expected to be affected by short-term proleukin stimulation, CD4, CD25, CD45RA and FOXP3. These were used to delineate different cell types, and the functional marker, pSTAT5, a signalling protein of the IL-2 pathway phosphorylated on IL-2 stimulation (Figure 3.1), was used to measure IL-2 response. Samples were analysed individually with flow cytometry, with all T1D samples except for two run in parallel with healthy controls to account for batch effects (Figure 3.2). Samples were also matched for age and sex to allow for paired analysis.

Flurochrome	T/NK cell panel	CD4 T cell panel	CD4/naive cell panel	NK cell panel
Alexa Fluor 488	pSTAT5	pSTAT5	pSTAT5	pSTAT5
Alexa Fluor 700	CD4	CD4	CD4	CD4
APC	CD25	CD25	CD25	CD25
Pacific Blue	CD56	CD45RA	CD45RA	CD45RA
PE YG	FOXP3	FOXP3	FOXP3	FOXP3
PE-Cy7 YG	CD45RA		CD31	CD56
PerCP Cy5-5	CD3			
Qdot 605	CD8			
Number of samples	10	95	12	66

Table 3.2. Proleukin stimulation assay antibody-fluorochrome panels. The fluorochrome-antibody panels used in IL-2 stimulation. The panel used on the majority of samples was the CD4 T cell panel, used to disriminate effector and regulatory naive and memory T cells. The T/NK cell panel, which contains the most markers, was only ran on a subset of 26 samples.

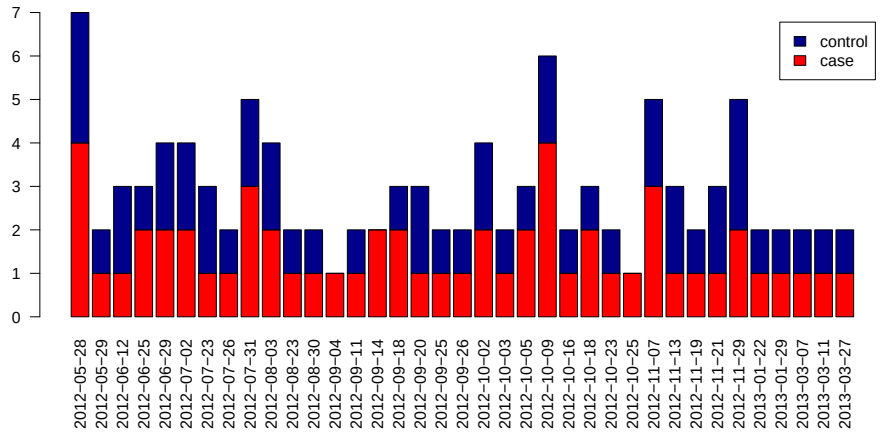


Figure 3.2. Number of cases and controls analysed per day. To account for batch effects in the case-control association testing, Tony Cutler aimed to analyse, when possible, at least one healthy control sample and one type 1 diabetic sample per day. On two days, 2012-09-04 and 2012-10-25, no matching controls were ran.

3.2 Cell Phenotypes Identified by Manual Analysis

In the preliminary manual analysis using FlowJo, Tony Cutler, gated four $CD4^+$ lymphocyte subsets (Figure 3.3):

- memory effector T cells (Teffs)
- memory regulatory T cells (Tregs)
- naive Teffs
- naive Tregs

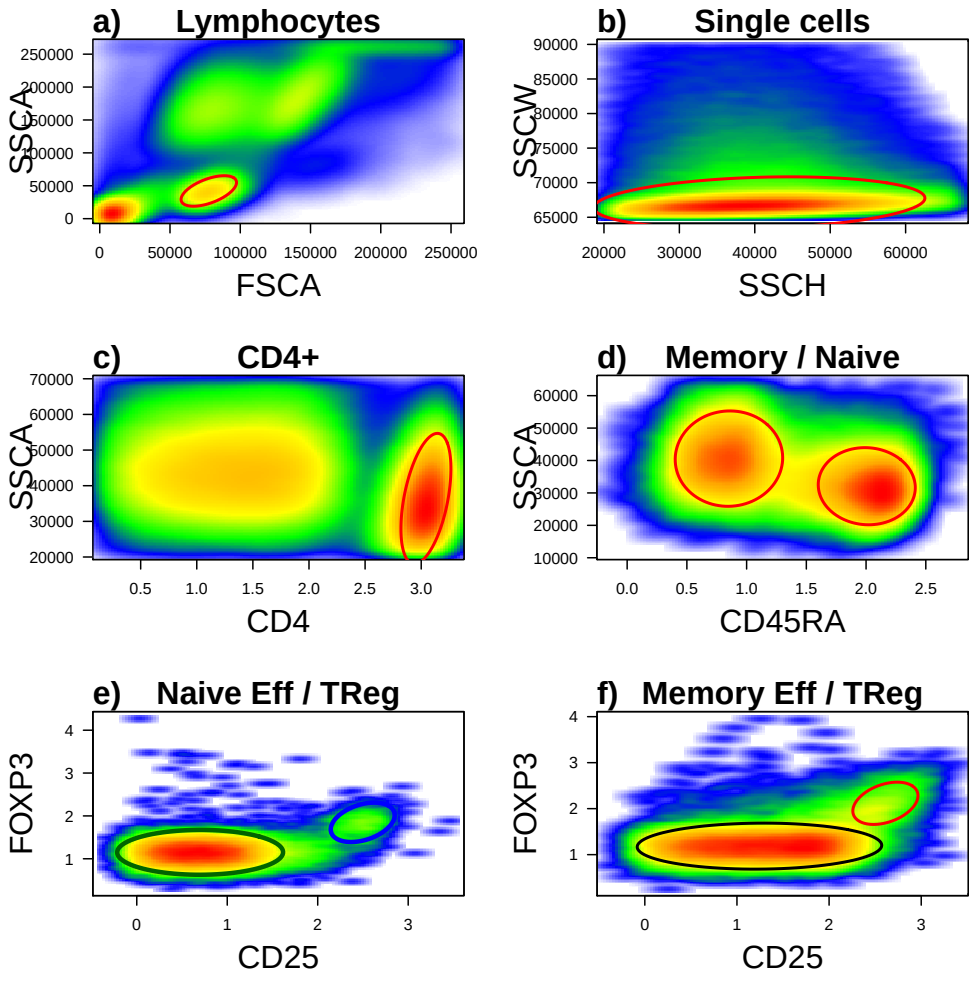


Figure 3.3. Gates applied across doses. Manual gating conducted using FlowJo by Tony Cutler to identify naive Teffs (green ellipse) and Tregs (blue ellipse) (e) and memory Teffs (black ellipse) and Tregs (red ellipse) (f).

Within each lymphocyte subset, the pSTAT5 distribution was measured, in each of the four samples stimulated at an increasing proleukin dose (Figure 3.4). As expected, the pSTAT5 distribution shifts progressively right for higher doses of proleukin, as more STAT5 is phosphorylated. Of the four subsets, the most sensitive cells to proleukin are the smaller memory and naive Treg subsets (Figure 3.4 b and d), then the memory Teffs (Figure 3.4 a) and finally the naive Teffs (Figure 3.4 c). This correlates with the level of CD25 expressed by these cells.

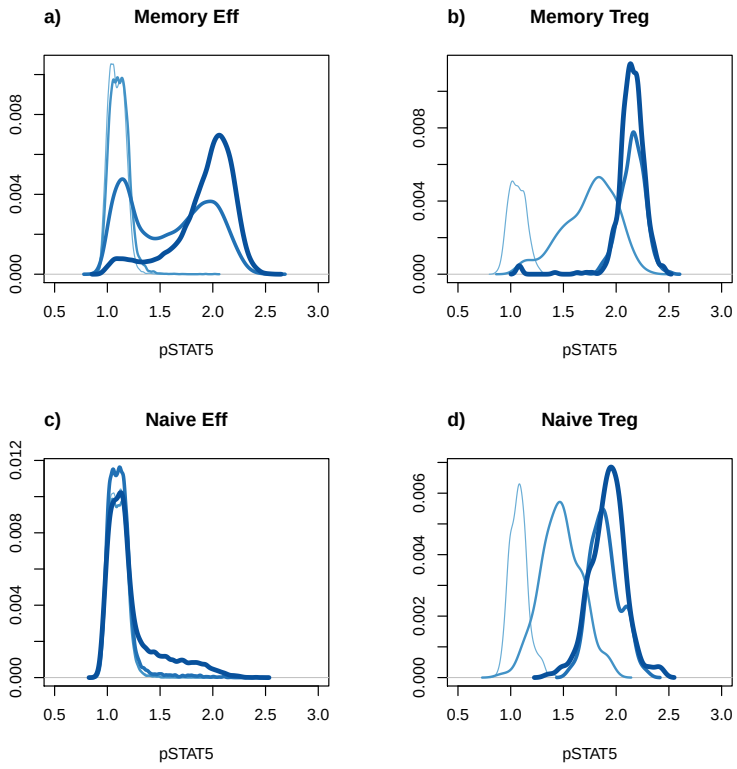


Figure 3.4. Distribution pSTAT5 in the manually gated cell subsets in the sample manually gated in Figure 3.3. The thickness of the lines is representative of the four increasing doses of proleukin (0, 0.1, 10 and 1000 units). The dose-response is most striking in the smaller Treg subsets with higher CD25 (b and d). The dashed vertical line represents the 99th percentile of the pSTAT5 distribution in the resting sample, which is used to define the pSTAT5$^+$ threshold.

Next Tony Cutler, assessed the repeatability of the pSTAT5 response by measuring the pSTAT5 MFI for each individual subset of lymphocytes tested. However, this cell phenotype was poorly reproducible, since the location of the peaks was not stable across days as illustrated by the sample shown in Figure 3.5. This motivated Tony Cutler to

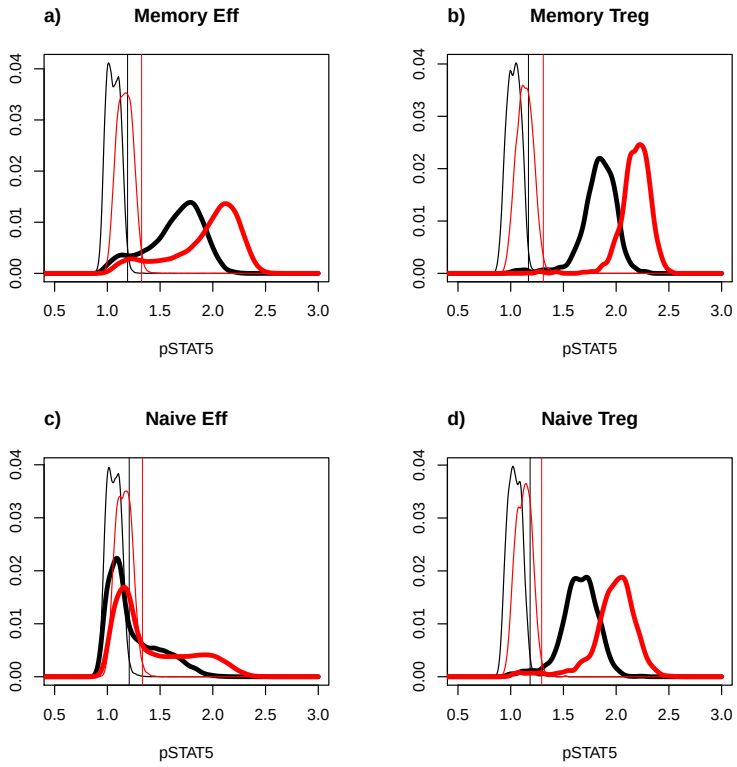


Figure 3.5. pSTAT5 distribution in an individual on visit one (black) and visit two (red) clearly shows that pSTAT5 distribution is not stable across days in the four cell subsets. In black, the pSTAT distribution on visit one and in red, on visit two. The thinner lines are from the resting sample whereas the thicker lines are from the sample stimulated at 1000 units. The vertical lines represent the pSTAT5$^+$ threshold set at the 99^{th} percentile of the pSTAT5 distribution in the resting sample.

instead use a threshold approach to define the ratio of cells which are pSTAT5$^+$. An idea similar to that applied in Chapter 2 to define naive cells as CD25$^+$. However, here an internal threshold was used, namely the threshold was defined to be the 99th percentile

of the pSTAT5 distribution in the resting cell subset per sample. Thus each each sample and cell subset had its own pSTAT5$^+$ threshold. He presented his results in six of the ten repeated individuals for the four stimulated cell subsets, memory Teffs (Figure 3.6), memory Tregs (Figure 3.7), naive Teffs (Figure 3.8) and naive Tregs (Figure 3.9). For memory and naive Tregs (Figures 3.7 and 3.9), at the highest 1000 units dose, practically all memory and naive Tregs are pSTAT5$^+$. However, these cell populations show a significant response already at the lowest 0.1 units dose of proleukin. Based on this observation, Tony Cutler selected the memory and naive Treg pSTAT5 cell phenotype to be the percentage of pSTAT5$^+$ cells at the 0.1 units dose. On the other hand, since memory Teffs are less reponsive, 10 units was chosen as the representative dose. For naive Teffs, the least reponsive of the four cell subsets considered, the repeatability was assessed at the 1000 unit dose. Tony assessed the repeatability with the coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(x_{i1} - x_{i2})^2}{\sum_{i=1}^{N}(x_{i1} - \overline{x_1})^2}$$

where x_{i1} is the phenotype of i^{th} individual on the first day and x_{i2} is the phenotype of the i^{th} individual on the second day. The coefficient of determination can take negative values if the correlation between x_{i1} and x_{i2} is very low. Contrary to the Pearson correlation, used in the previous chapter, the coefficient of determination is sensitive to linear transforms. Even when using the percent of pSTAT5$^+$ cell phenotype, the overall reproducibility across the four cell subsets was still poor, with the more sensitive and smaller cell subsets, naive and memory Tregs showing the worst correlation ($R^2 = -0.16$ and $R^2 = -0.82$), memory Teffs showing slightly better correlation ($R^2 = 0.021$) and finally the less responsive naive Teffs showing good correlation ($R^2 = 0.7728$) (Figure 3.10). Tony then went on to test association with type 1 diabetes using a two-tailed paired t-test of 20 cases matched with 20 controls analysed on the same day (Figure 3.11). He also tested for association with *IL2RA* SNP rs12722495, and the *PTPN2* SNPs rs45450798

and rs478582 (plots not shown). No significant association was detected either with disease nor with genotype.

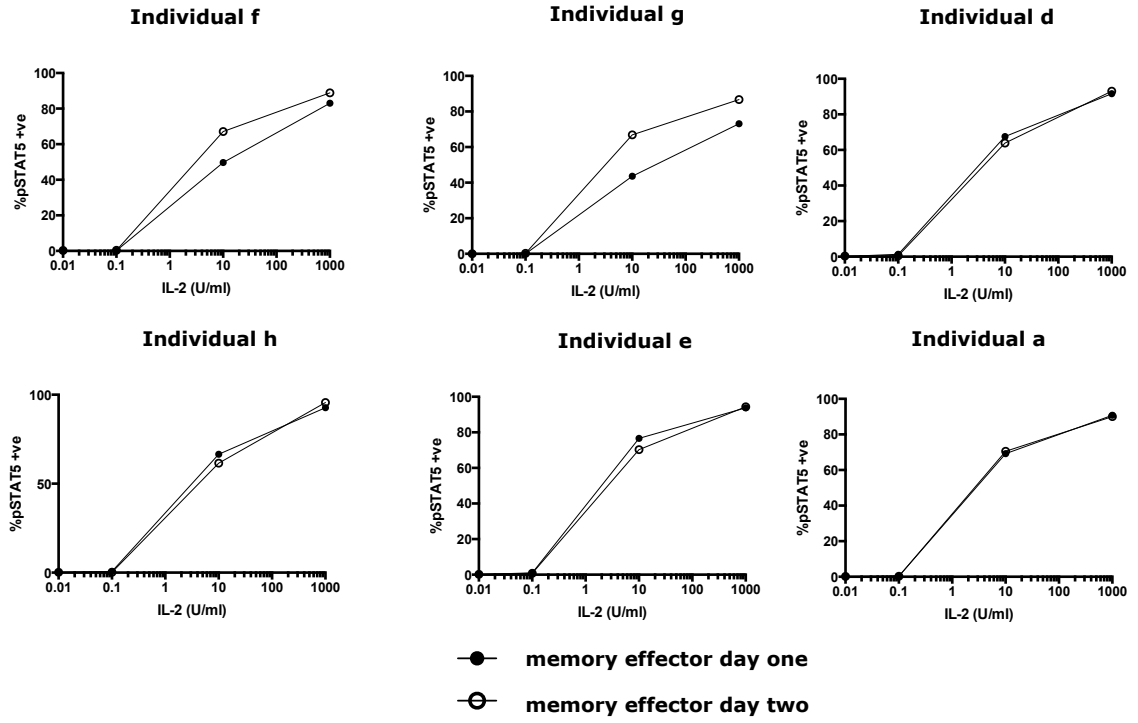


Figure 3.6. The percent of pSTAT5$^+$ cells increases with proleukin dose in memory Teffs, but the measured response is not consistently repeatable (f, g). Plot produced by Tony Cutler.

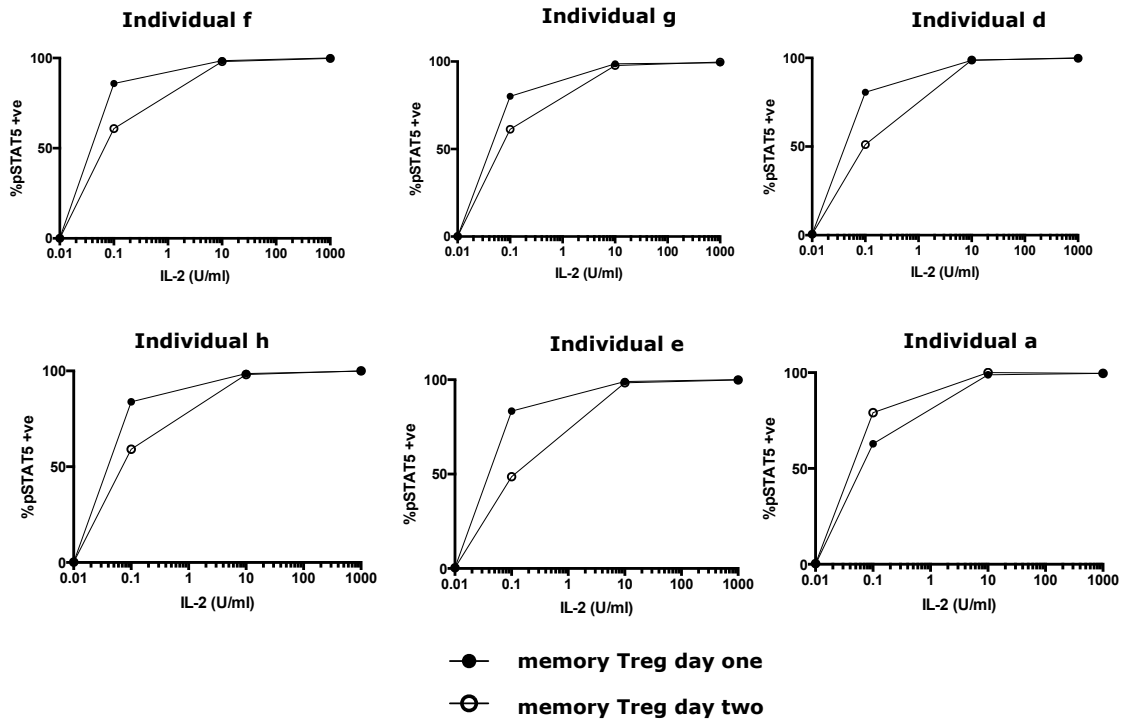


Figure 3.7. The percent of pSTAT5$^+$ cells increases with proleukin dose in memory tregs. Plot produced by Tony Cutler. While at the highest proleukin dose of 10 and 1000 units, all memory tregs are consistently pSTAT5$^+$, there is more discrepancy at the low dose of 0.1 units.

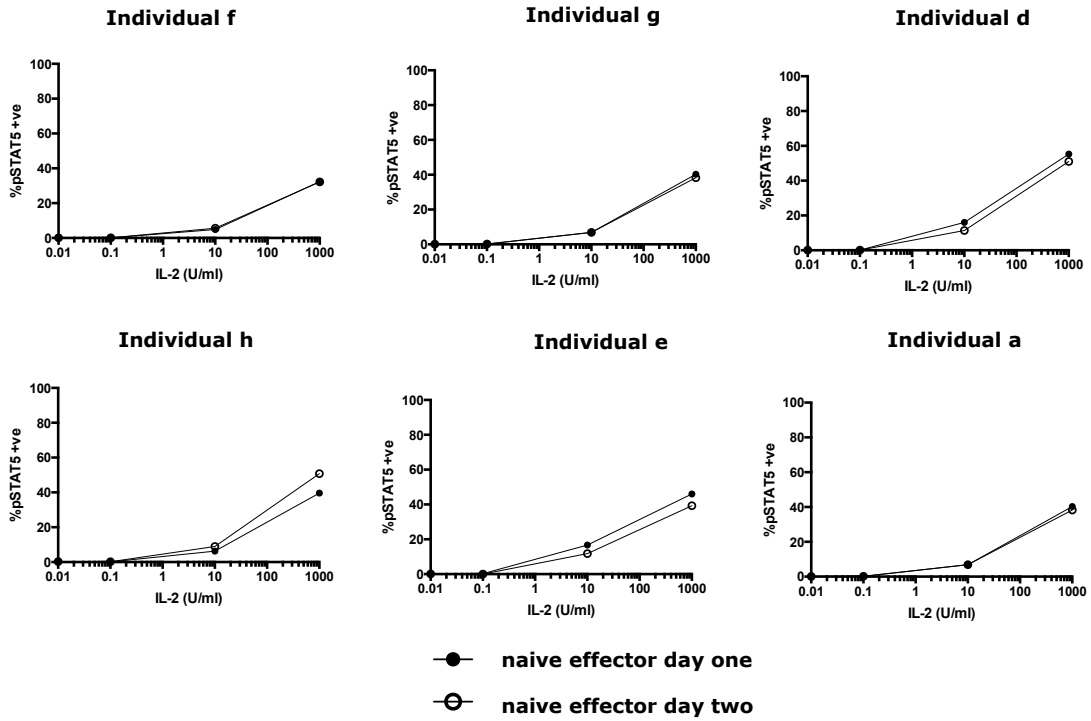


Figure 3.8. The percent pSTAT5$^+$ cells increases with proleukin dose in naive Teffs. Plot produced by Tony Cutler. Only 40 % of the naive effector cells are pSTAT5$^+$ even at the highest 1000 unit proleukin dose.

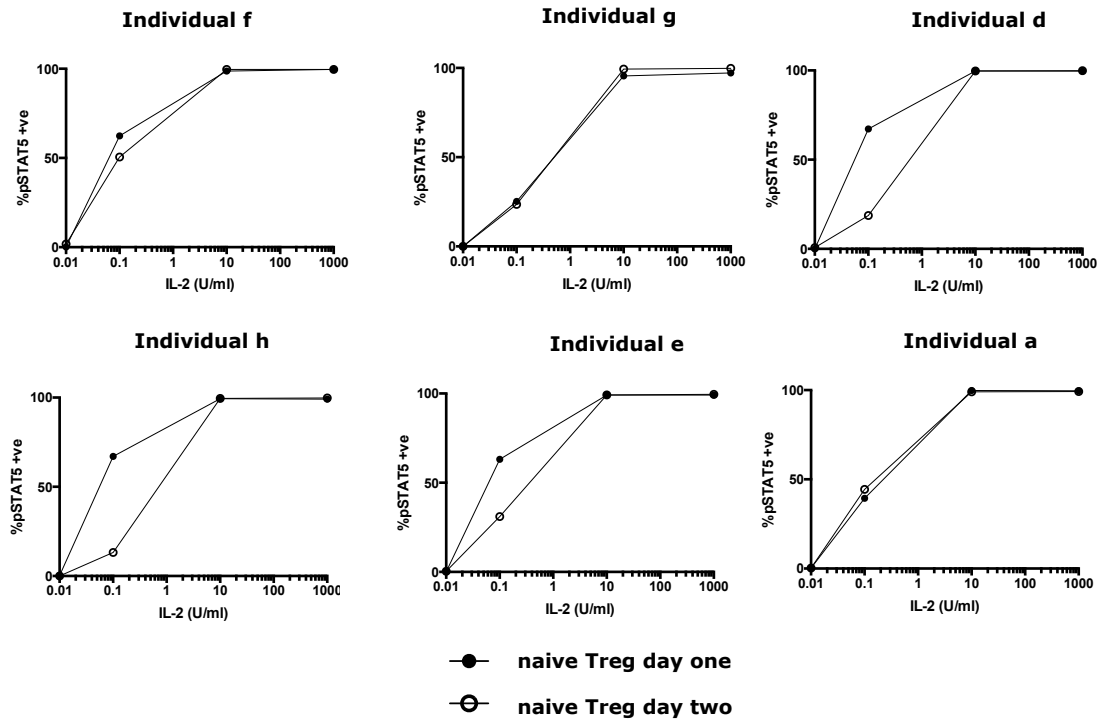


Figure 3.9. The percent pSTAT5$^+$ cells increases with proleukin dose in naive tregs. Plot produced by Tony Cutler. While at the highest proleukin doses of 10 and 1000 units, all naive tregs are consistently pSTAT5$^+$, there is more discrepancy at the low dose of 0.1 units.

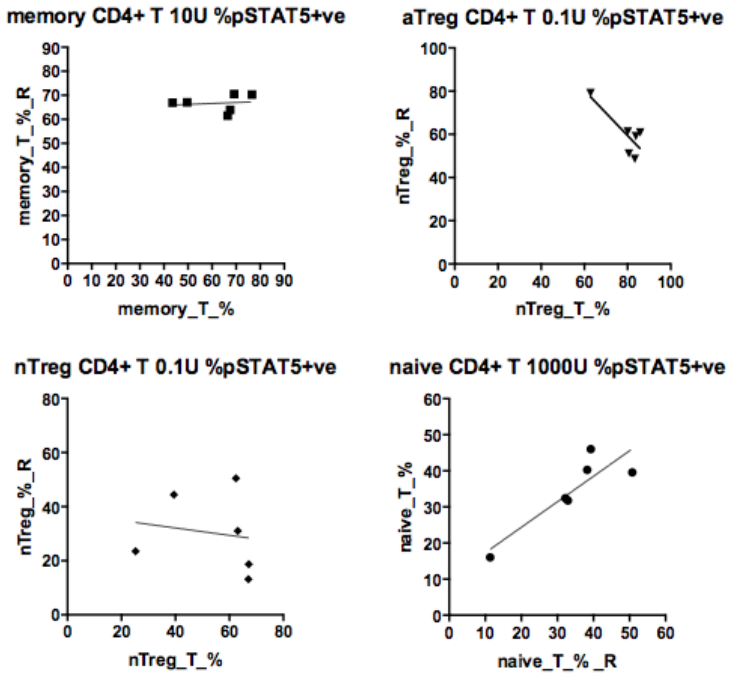


Figure 3.10. Repeatability in six individuals Plot produced by Tony Cutler. The repeatability of the percent of cells which are pSTAT5$^+$ is assessed in effector memory and naive at the 10U and 1000U dose respectively, and in memory and naive Tregs at the 0.1U dose. While the repeatability in the naive effector subset was good ($R^2 = 0.7728$), the repeatability in the other cell subsets is poor (memory Teffs $R^2 = 0.021$, naive tregs $R^2 = -0.16$ and memory tregs $R^2 = -0.82$).

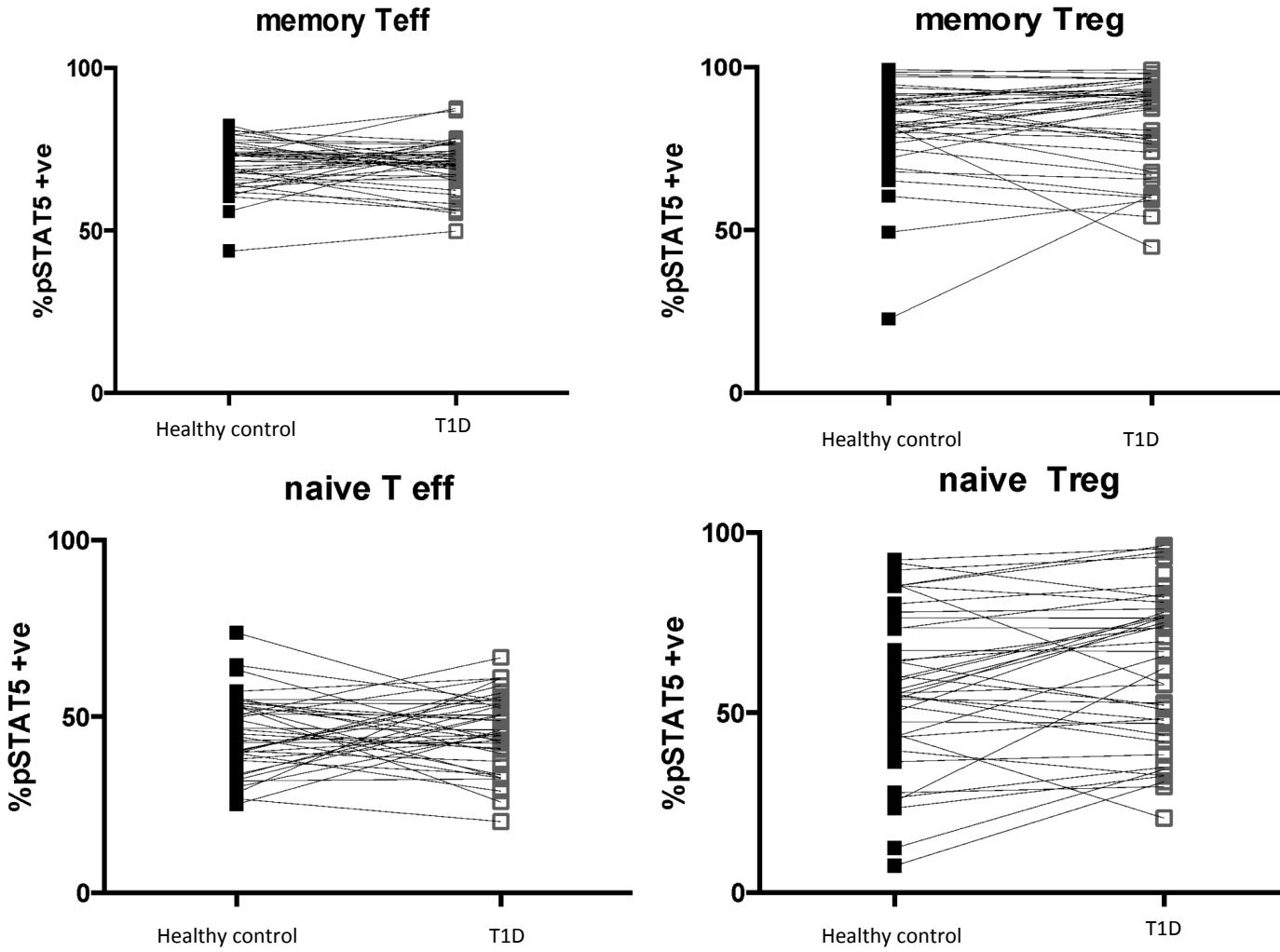


Figure 3.11. Association test of percent pSTAT5$^+$ in four cell subsets. Plot produced by Tony Cutler. The association with T1D of the percent of cells which are pSTAT5$^+$ is assessed in effector memory and naive at the 10U and 1000U dose respectively, and in memory and naive Tregs at the 0.1U dose. The association test is a two-tailed paired-t-test on 20 cases paired with 20 controls analysed on the same day (40 out of the available 96 individuals). No significant association is detected.

3.3 Reproducibility of pSTAT5 response within an individual

Tony's preliminary results in the six repeated individuals suggested that the pSTAT5 MFI and percent pSTAT5$^+$ were poorly reproducible cell phenotypes using the methodology and approach described, which consequently would give us little power to detect an association with disease status or genetics. This motivated me to see if I could improve the repeatability of these cell phenotypes using normalisation. I evaluated the reproducibility of these cell phenotypes with different normalisation approaches.

3.3.1 Normalisation approaches

Here I describe the methods I considered.

Bead normalisation In Chapter 2, I used beads to correct for day to day variation in the CD25 channel. However for the stimulation data, beads in the Alexa-488 channel, the fluorochrome conjugated to pSTAT5 (Table 3.2), did not adequately capture the short term variation in pSTAT5 (Figure 3.12).

Correcting for baseline MFI One observation which can be drawn from Figure 3.5 is that the MFI of the pSTAT5 distribution in the resting sample is different across days. If a cell population had a higher resting pSTAT5 MFI due only to day to day variability, then one might expect that the pSTAT5 MFI in the stimulated population would also be higher. I first attempted to account for the difference in resting pSTAT5 by taking the ratio of the MFI of the stimulated populations over that of the MFI of the resting population, or equivalently by subtracting the log transformed MFIs. However, this did not appear to improve the repeatability significantly but instead just reduced the MFI by the same factor on both days (Figure 3.13). This suggests that the day to day

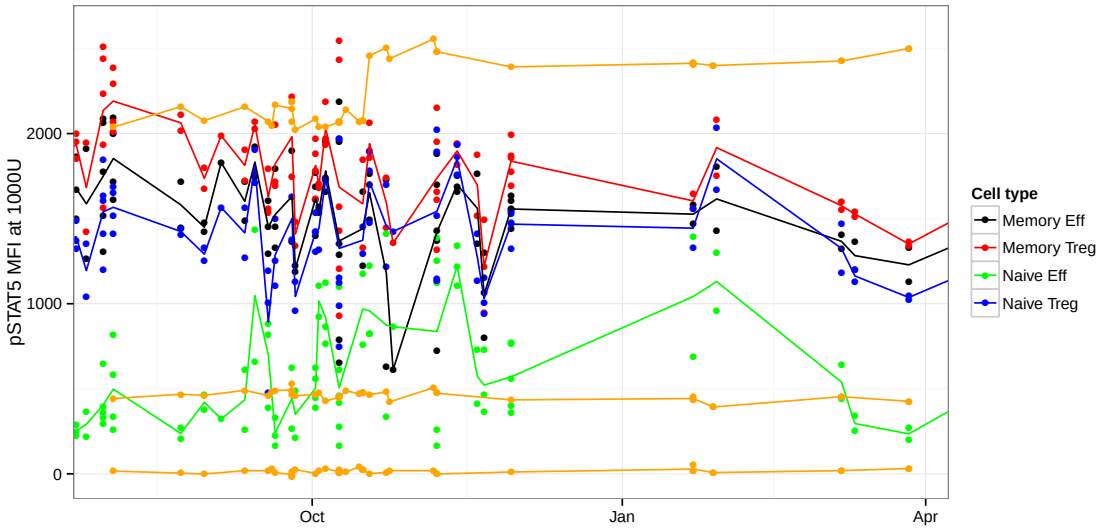


Figure 3.12. Variation in sample MFI is not captured by variation in bead MFI. For the purpose of MFI normalisation, the fluorescence intensity of six-peak flow cytometry beads was measured in the Alexa 488 channel, the fluorochrome conjugated to the pSTAT5 marker. However, as illustrated by the loess lines, the beads are not appropriate for pSTAT5 normalisation, since the MFI of the three dimmest populations of beads (orange) does not capture the pSTAT5 MFI time variation in the four cell subsets. The pSTAT5 MFIs are obtained from samples stimulated at 1000 units of proleukin.

variation in the resting sample MFI does not capture the variation in the stimulated population .

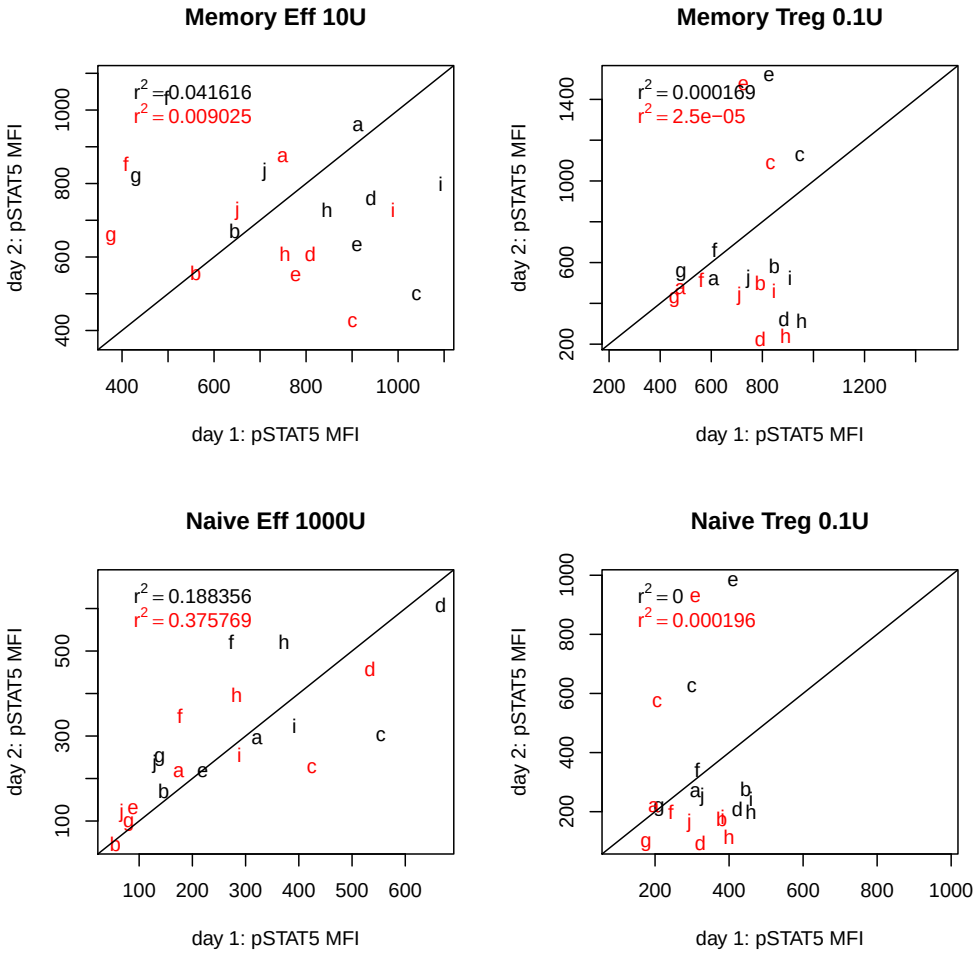


Figure 3.13. pSTAT5 MFI (black), background subtracted pSTAT5 MFI (red) Correcting for the MFI in the resting sample does not appear to improve the repeatability significantly but instead just reducing the MFI by the same factor on both days. This suggests that the day to day variation in the resting sample MFI does not explain the variation in the stimulated population.

Nearest-neighbour joining One concern is that the difference in population cell counts across samples stimulated at different doses, may influence the accuracy of the MFI estimate. Another concern is that since the pSTAT5 distribution is often bimodal in the cell populations considered, substracting the MFI may not be ideal. Instead, a more correct approach would be to subtract the pSTAT5 fluorescence intensity for each resting cell. One way of emulating this is to match each cell to its closest neighbour in the unstimulated sample. This was accomplished by joining samples on their core markers using the approximate-nearest-neighbour (ANN) to the resting sample (Jones et al, 2011) as implemented in the `R` package RANN (Arya et al, 2013). This created a dataset of the same number of cells as the resting dataset, but where each cell now had a total of four functional pSTAT5 markers, one for each stimulation dose. At each cell it is now possible to assess the difference in pSTAT5 response between resting and stimulated states. This is important because cells do not all have the same resting level of pSTAT5 (Figure 3.14). This approach presents a number of advantages. Firstly, only the sample to which the other samples are joined needs to be gated. Secondly, since the pSTAT5 response is relative to the baseline, it should be more robust to variation between days and consequently, more reproducible than pSTAT5 fluorescence intensity. Thirdly, since we have response at the cell-level, we can apply methods to do multivariate regression of pSTAT5 from core markers. This could help identify cells which would have been missed from only examining core markers. The ANN approach is valid without normalisation since the distributions of the core markers align across doses. However, using this nearest-neighbour joining method, the repeatability of the cell phenotypes pSTAT5 MFI (Figure 3.15) and the percent of $pSTAT5^+$ (Figure 3.16) are not substantially improved.

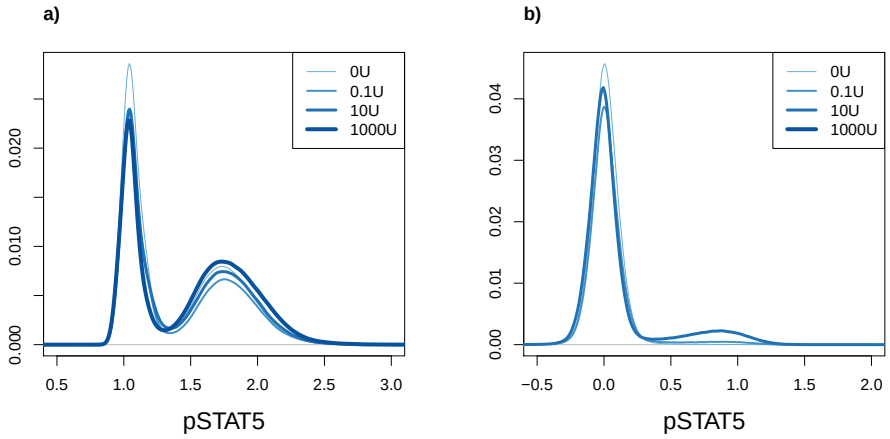


Figure 3.14. pSTAT5 intensity across the four proleukin doses, before (a) and after (b) per-cell baseline pSTAT5 subtraction in the ungated sample. An important proportion of cells are already saturated for pSTAT5 (high baseline) in the resting sample (a). Correcting for the per cell pSTAT5 baseline, shows the true proportion of cells which responds to proleukin within this sample (b).

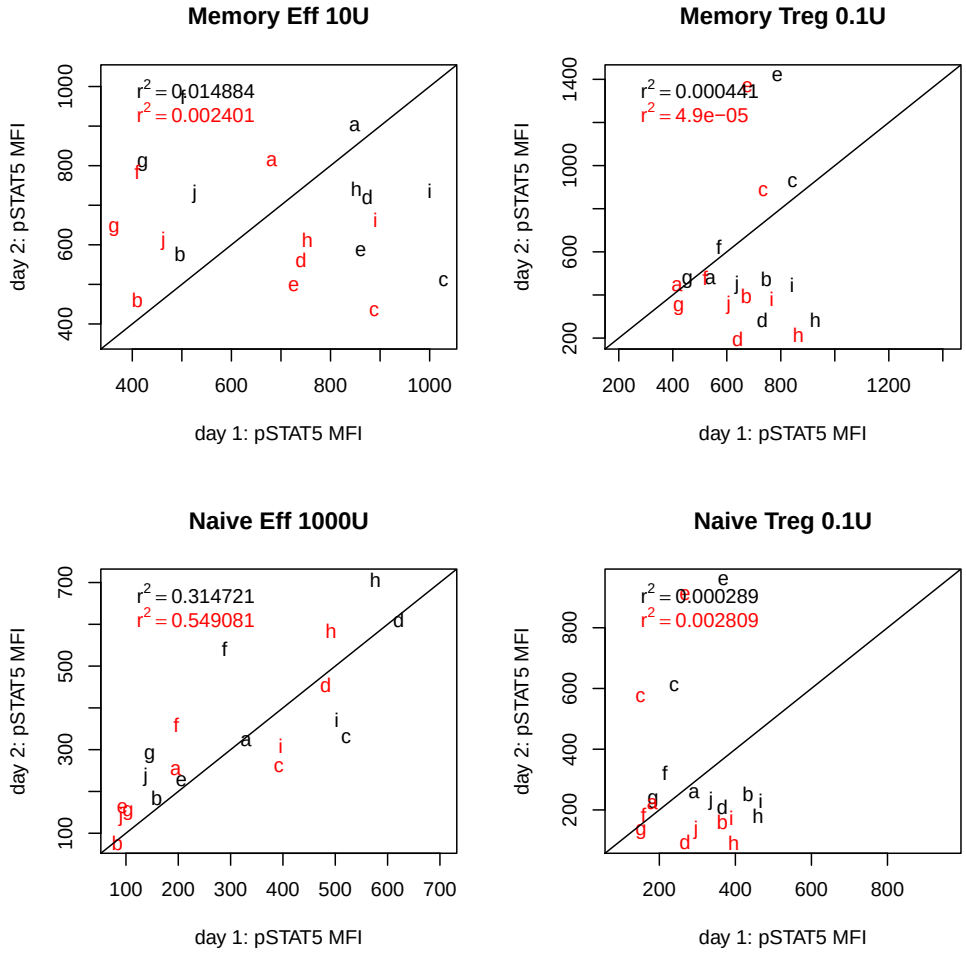


Figure 3.15. Repeatability of pSTAT5 MFI in the nearest-neighbour joined (black), nearest-neighbour background subtracted (red). Background subtraction does no substantially improve the repeatability of the pSTAT5 MFI phenotype.

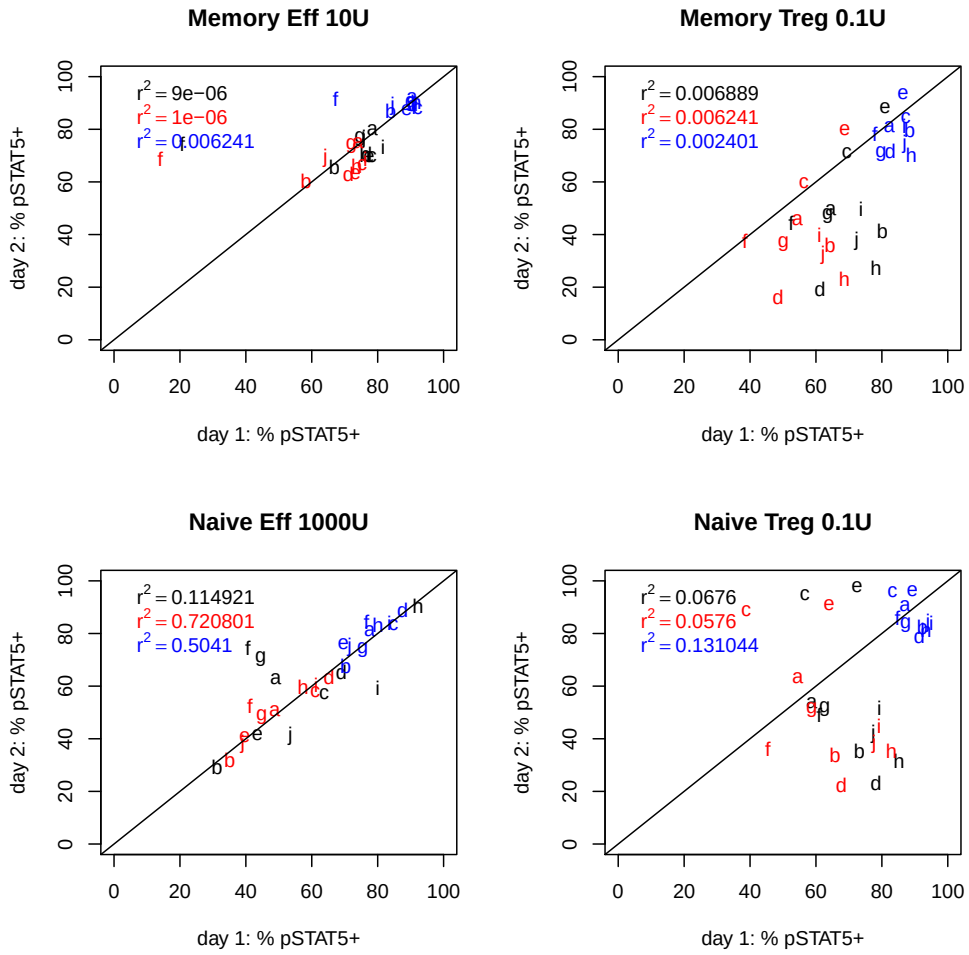


Figure 3.16. Repeatability of percent pSTAT5$^+$ in the individual samples (black), nearest-neighbour joined (red) and nearest-neighbour joined samples baseline corrected (blue).

3.3.2 Repeatability

For all ten repeated samples, the Pearson correlation squared r^2 of the MFI was assessed at the four increasing doses, in the four cell subsets, for the raw (Figure 3.17a), baseline corrected (Figure 3.17b), nearest-neighbour joined (Figure 3.17c) and nearest-neighbour baseline corrected (Figure 3.17d). The repeatability of the percent pSTAT5$^+$ across doses gives a different pattern depending on the cell type but at the highest dose it appears that the least responsive naive effector T cells (Teffs) yield the best repeatability (Figure 3.18). Suprisingly, the memory Tregs have the highest repeatability at the 0.1 unit dose. The repeatability of the naive Treg phenotype is poor at all doses. The NN joining does not significantly improve the repeatability (Figure 3.15). No normalisation approach stands out as improving the repeatability.

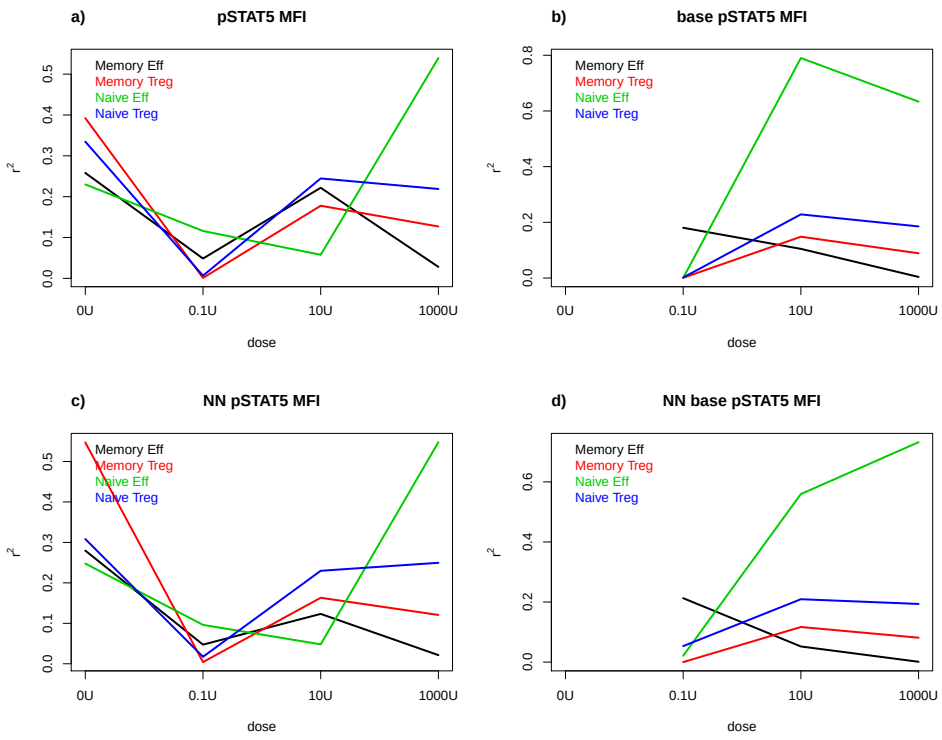


Figure 3.17. Repeatability of pSTAT5 MFI measured as Pearson correlation squared (r^2) per dose per cell type. On subtracting the baseline, the repeatability of the naive effector subset is improved but not in the other cell subsets.

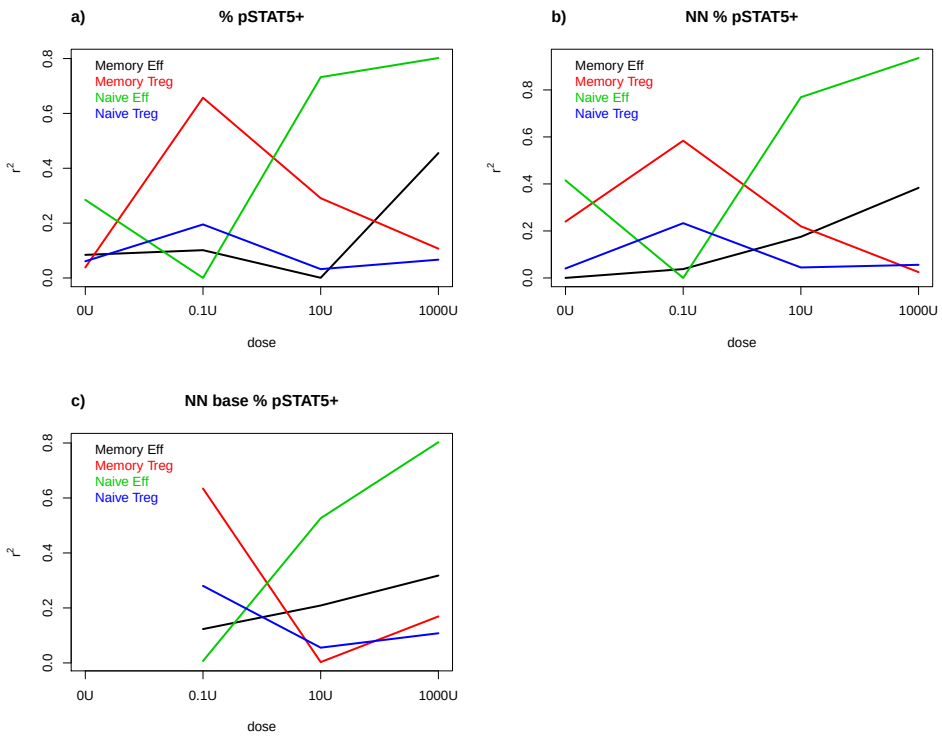


Figure 3.18. Repeatability of the percent of pSTAT5$^+$ as Pearson correlation squared (r^2) per dose per cell type. Only the pSTAT5 in the naive effector subset stimulated at 100U shows good repeatability across all normalisation methods.

3.3.3 Association of pSTAT5 response with type 1 diabetes

I tested the association with T1D at each dose as well as for the area under the curve. I accounted for repeated individuals and day of analysis by including them as random effects in a linear mixed effects model as applied in Chapter 2. No T1D association was detected with the pSTAT5 MFI (Figures 3.19 and 3.20) nor with the percent pSTAT5$^+$ (Figures 3.21 and 3.22) cell phenotypes in the four cell subsets considered.

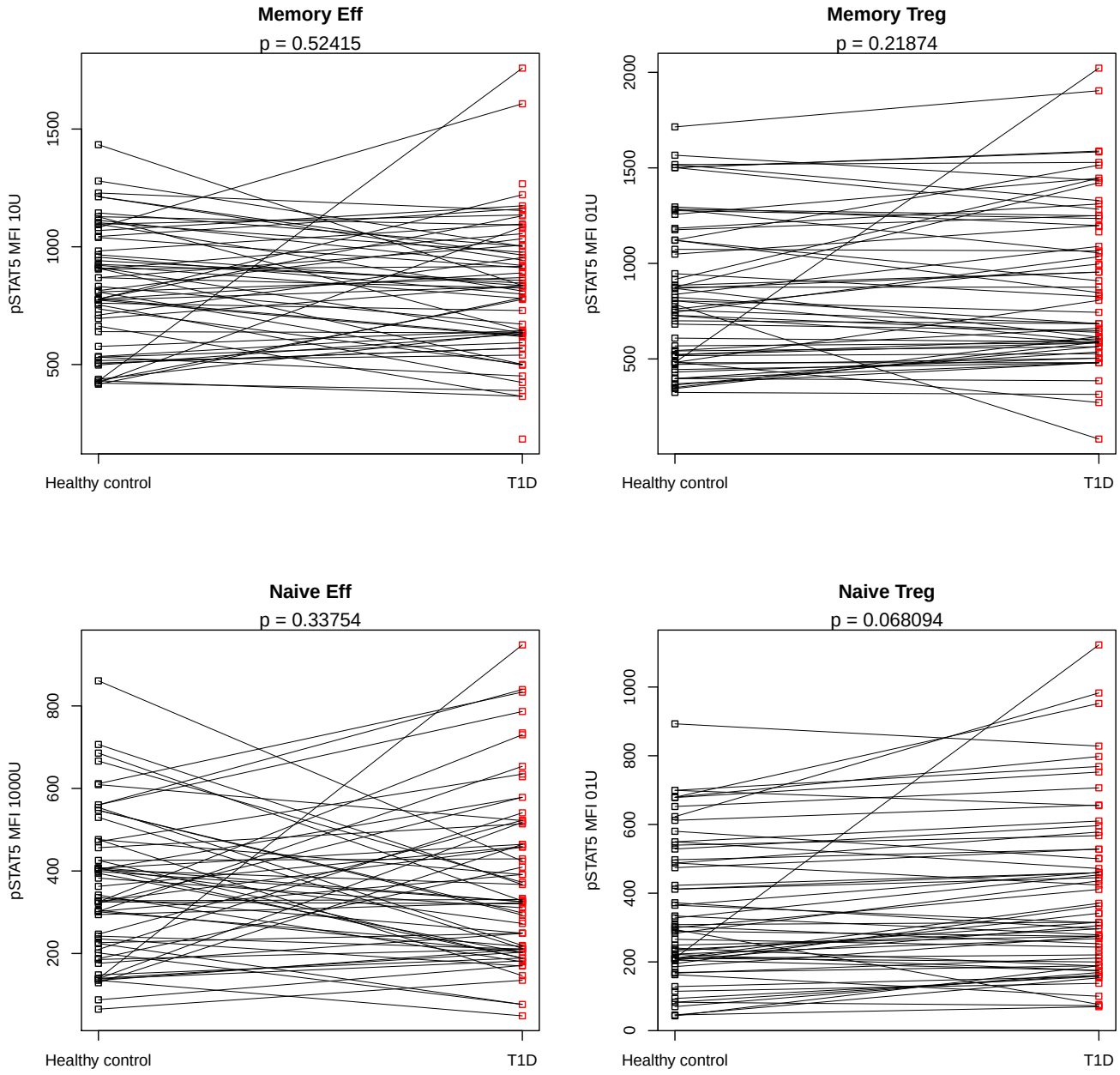


Figure 3.19. Association test of pSTAT5 MFI with T1D. Samples are paired by day of analysis.

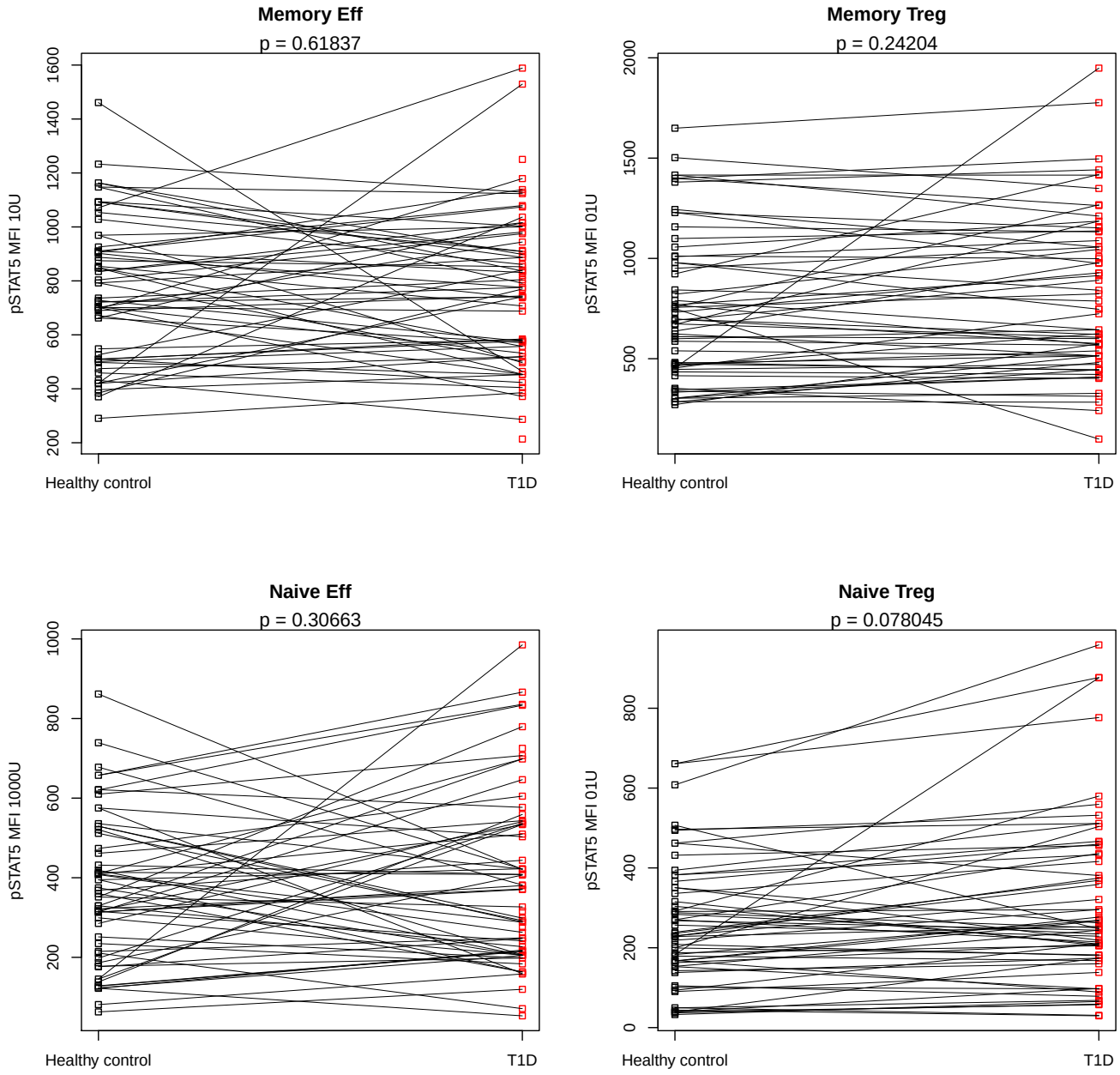


Figure 3.20. Association test of pSTAT5 MFI, after nearest-neighbour normalisation, with T1D.

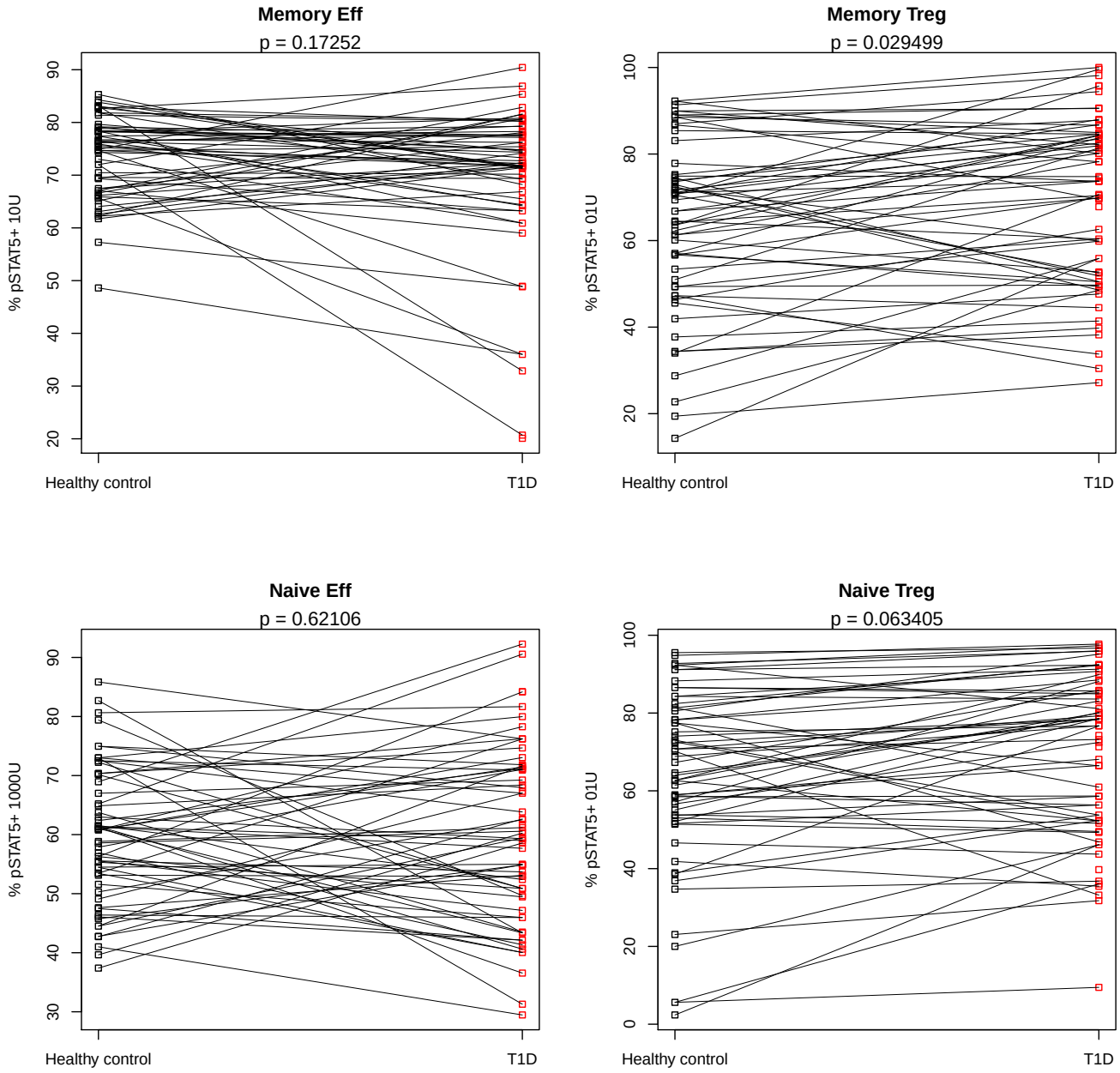


Figure 3.21. Association test of percent pSTAT5$^+$ with T1D.

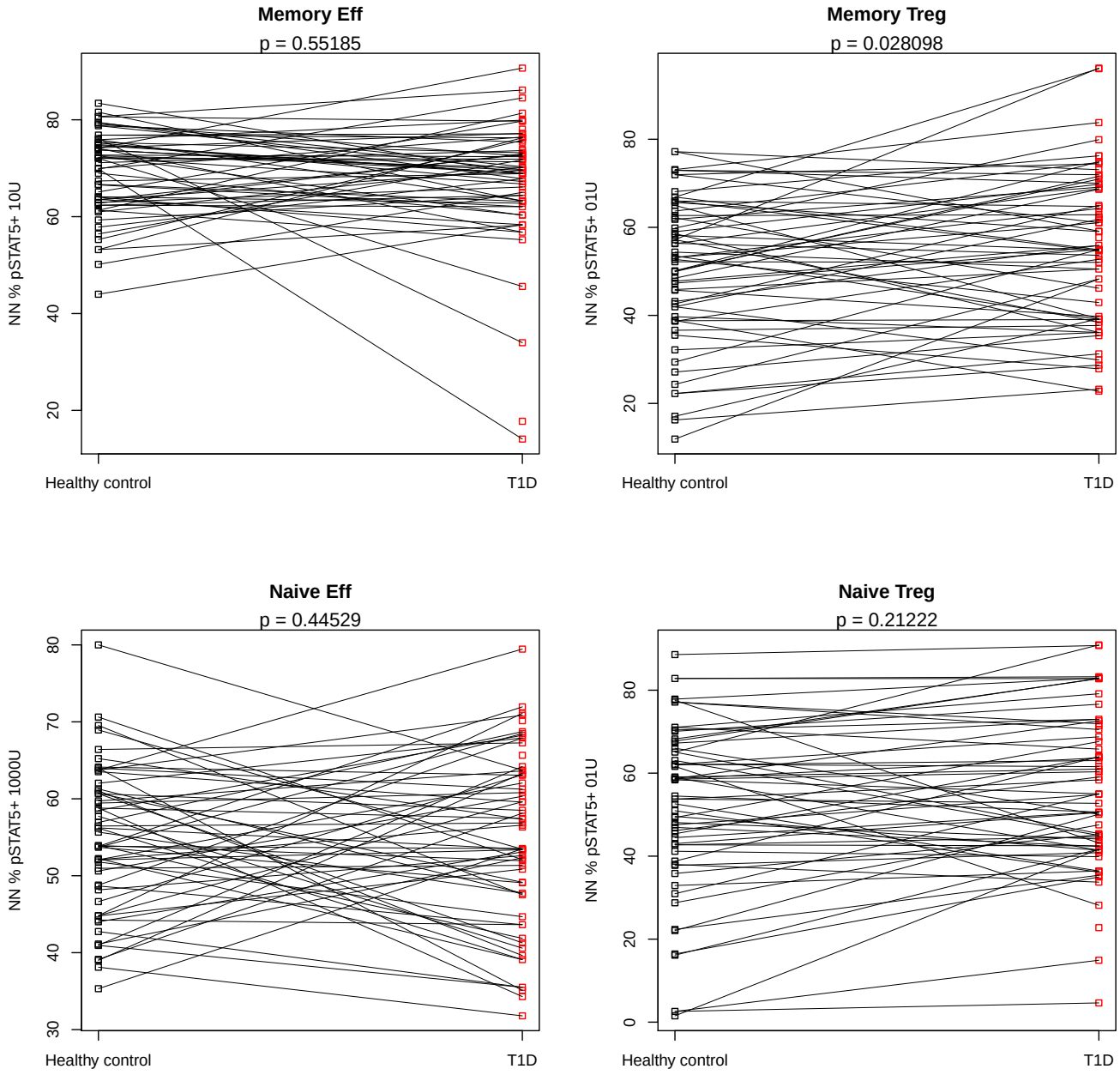


Figure 3.22. Association test of percent pSTAT5$^+$, after nearest-neighbour normalisation, with T1D.

3.4 Response in the whole sample

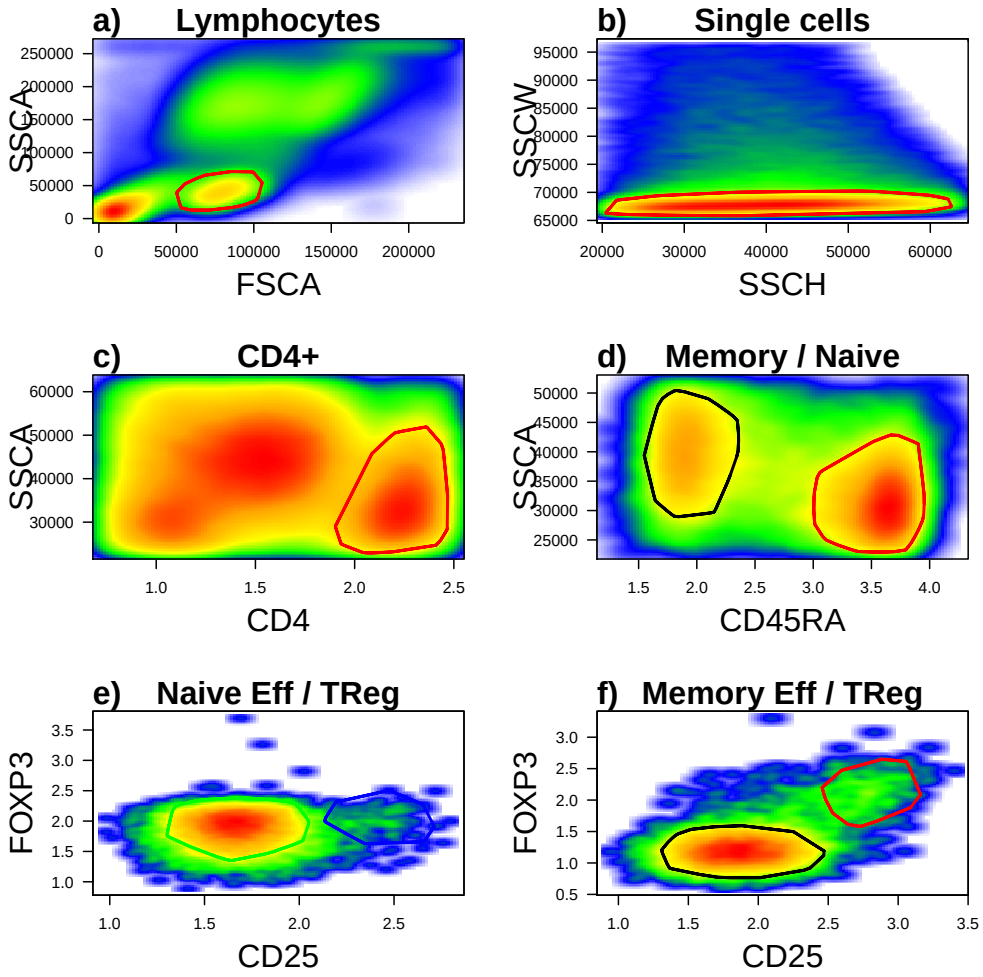


Figure 3.23. Gates applied across doses. Manual gating conducted using FlowJo by Tony Cutler to identify naive Teffs (green ellipse) and Tregs (blue ellipse) (e) and memory Teffs (black ellipse) and Tregs (red ellipse) (f).

Using the normalisation methods described in the previous section, I was unable to significantly improve the reproducibility of this assay. This implies that this dataset, as it stands, is not sufficiently reproducible to allow for successful association testing. However, it still contains useful biological information, and may be repurposed to answer another important question: beside the cells we know of, are there other cells which

respond to proleukin within the peripheral blood mononuclear cells (PBMC)?

This question is extremely relevant for DILT1D and other clinical trials of IL-2 which have mostly focused on lymphocytes. Biologists know that any cell which carries high levels of any of the trimeric components of the IL-2 receptor, alpha (CD25), beta (CD122) or gamma (CD132), should respond to IL-2, but due to low affinity of the CD122 antibody and the limitation on the possible number of fluorochromes per tube, these cannot all be included as part of every flow experiment. This study of *ex vivo* stimulated whole blood offers the opportunity to identify other, potentially new, cell subsets also responsive to IL-2. In order to increase my chances of characterising these subsets, I only analysed samples stained with the most comprehensive marker panel we had available, the one containing the additional fluorochromes for CD3, CD8 and CD56 (Table 3.2). Unfortunately, the staining using this panel was of poor quality in many samples, partly due to the permeabilisation protocol and to the larger number of markers used, which caused certain marker stains, such as CD56, to not work well across all batches. This made between-batch analysis infeasible so instead, I focused on the analysis of a single batch for which the staining had worked. Assessing the staining quality of a sample requires prior knowledge, and so Marcin Pekalski, an experienced flow cytometrist, assisted me in finding a batch in which the staining matched the expectation (Figure 3.23).

When assessing the dose-response to stimulation in a flow cytometry sample, the classic approach is to first gate cell populations in each sample based on their core markers, then to assess the response of the functional marker in the gated subset. Obviously, this approach is not exhaustive and consequently may miss other dose-responsive cell populations which are not included in the gating strategy. Here I first explored approaches to visualise the pSTAT5 dose-response in the whole sample in order to spot other potential dose-responsive cell subsets. I also considered more automated methods which use the pSTAT5 response to guide the identification of these cell subsets. One of the challenges

faced in identifying new cell types is how to reconcile the two different kinds of flow cytometric data, scatter and fluorescence, since these represent quite different physical properties and are measured on different scales, linear and logarithmic respectively. Usually scatter gating takes precedence over fluorescence gating, and the standard is to first use scatter to discriminate live cells from debris. In the next section, I followed this protocol, by first gating on side and forward scatter in order to distinguish lymphocytes from non-lymphocytes. I then conducted separate analysis on the lymphocyte subset using the fluorescent markers only, and to non-lymphocytes, using both fluorescent and scatter markers.

3.4.1 Visualising response in whole sample with Spanning-tree Progression of Density-normalised Events (SPADE)

Visualisation is a fundamental tool for exploring high dimensional datasets. In this instance, I am interested in visualising the pSTAT5 response in the whole sample as a function of proleukin dose and a total of nine core markers, seven fluorescent markers and two scatter markers). Dimensionality reduction methods can provide a two-dimensional representation of a higher dimensional data set from a distance matrix. These methods are particularly suited for datasets with less data points but more dimensions than in flow cytometry, as generated by mass cytometry technologies such as CytoF. In mass cytometry datasets, more emphasis is given on uncovering cell lineages and progressions rather than discrete cell populations which share marker properties. Most of these methods, like multidimensional scaling (MDS), require computation of the complete pairwise distance matrix, but some like principal component analysis (PCA) can use the covariance matrix instead to identify the components which accounts for most of the variation. However, information is lost when considering only the first two components of a linear projection. Therefore, there is considerable interest in developing methods

that capture both the local and global structure of the data, so that points which lie close in higher-dimensional space tend to lie close in two-dimensional space. In flow cytometry, one such method is Spanning-tree Progression of Density-normalised Events (SPADE) which gives a minimum spanning tree (MST) representation, where each node in the tree represent a multi-dimensional cluster in flow marker space (Qiu et al, 2011). The MST is defined as the shortest path that connects all points in a network. The computation of the MST necessitates the complete distance matrix which is not feasible for most ungated flow cytometry samples. Hence SPADE first needs to reduce the number of data points by making use of downsampling and clustering. In order for all existing regions of the marker space to be equally represented in the reduced dataset, the density is normalised across the sample. At each data point, the multivariate density is estimated, then the number of points is reduced by preferentially removing points with high local density while preserving lower density ones. Once the number of points has been reduced to some target number or by some factor, in each individual sample, the samples are pooled and agglomerative clustering is applied. The distance matrix is then calculated on the clusters and they are joined as nodes by the minimum spanning tree (MST) algorithm for the purpose of visualisation. The points from each sample which were discarded in the density normalisation step, are then added back and assigned to their closest node in the tree. Hence the structure of the tree is the same across all samples but the size of the tree nodes is dependent on the number of data points assigned to each node per sample. The tree nodes can then be coloured according to the intensity of a functional marker, for example here pSTAT5, which was not used in its construction. Two steps of the algorithm require user specified parameters, the parameter that defines the downsampling, which can either be a target number of data points of a factor, and parameter that defines the number of desired clusters in the agglomerative clustering step.

Lymphocytes

I first ran SPADE on the manually gated lymphocyte subset (excluding doublets), in the resting and stimulated samples from an individual. The algorithm was run on the core surface markers, CD25, CD3, CD4, CD45RA, CD56, CD8 and FOXP3, which are expected to be stable across within-batch stimulation doses. The number of events in each sample was reduced by a factor of 90 percent. The desired number of cluster in the agglomerative step was set to 1000. The layout of the resulting MST was determined by the `R` function `SPADE.layout.arch` which aims to orientate the longest branch of an MST along an arch with shorter offshoot branches hanging below. The nodes in the tree were then coloured according to the fold increase in median pSTAT5 compared to the same node in the resting sample (Figure 3.24). The initial pSTAT5 response to the lowest 0.1 units dose proleukin, clusters in two regions of the tree, as seen in Figure 3.24b. As the stimulation dose is increased to 10 units, the level of the response increases in these two regions, and there are signs of response in further adjacent nodes of the tree (Figure 3.24c). Finally at the highest dose of 1000 units, the majority of the nodes show some level of response (Figure 3.24d).

In order to discover where the responsive cells lie on the tree in relation to the cells identified using manual gates, I mapped the cells labelled by manual gating as memory and naive, both Teffs and Tregs, cells onto their assigned tree nodes (Figure 3.25a). I found that the manually identified cell types tend to appear in neighbouring tree nodes but certain appeared in other locations of the tree. Also, memory Teffs and Tregs lie closer to each other than naive Teffs and Tregs (Figure 3.25a). From visual inspection, the pattern of pSTAT5 response in the MST corresponds to the locations of the cell types with the memory and naive Tregs showing the first signs of response at 0.1 units, memory Teffs starting to show activation at 10 units, followed by naive Teffs at 1000 units.

In Figure 3.25b, a number of dose-responsive tree nodes which lie far on the main branch to the other studied subsets, were selected and the cells they contain were projected back to marker space, I could visualise where these cells lied in relation to the known subsets (Figure 3.26). These cells constitute approximately one percent of the cells in the lymphocyte subset. As depicted in Figure 3.26, some properties of these cells which distinguishes them from naive and memory subsets are that they are high for CD56, $CD3^{-}$, $CD4^{-}$, $CD8^{+}$ and express moderate levels of CD25. Since those cells are $CD8^{+}$ $CD56^{+}$ they are natural killer (NK) like cells, and may have cytotoxic properties. These cells constitute 1.14 percent of all lymphocytes. However they are only stimulated at a 1000 units of proleukin so unlikely to be influential at the low doses used in DILT1D.

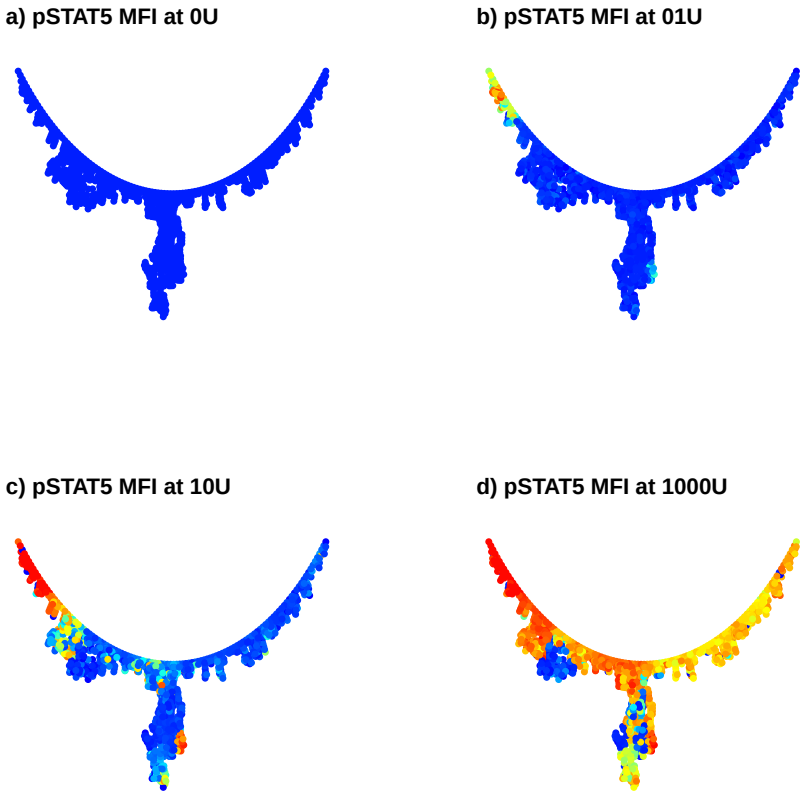


Figure 3.24. MST generated by applying SPADE on lymphocytes. MST nodes are coloured by pSTAT5 MFI. The MST was constructed from running SPADE on the core surface markers, CD25, CD3, CD4, CD45RA, CD56, CD8 and FOXP3, in the manually gated lymphocyte subset (after double exclusion), pooled across the four stimulation doses in a sample from one individual. The required number of clusters in the agglomerative clustering step was set to k=1000. The colouring of the nodes from dark blue to bright red follows the pSTAT5 MFI fold increase. In samples where the proleukin dose is increased, more nodes in the tree are illuminated since the pSTAT5 MFI increases in various cell subsets. The size of the tree nodes are proportional to the number of cells in the data file which are assigned to that node.

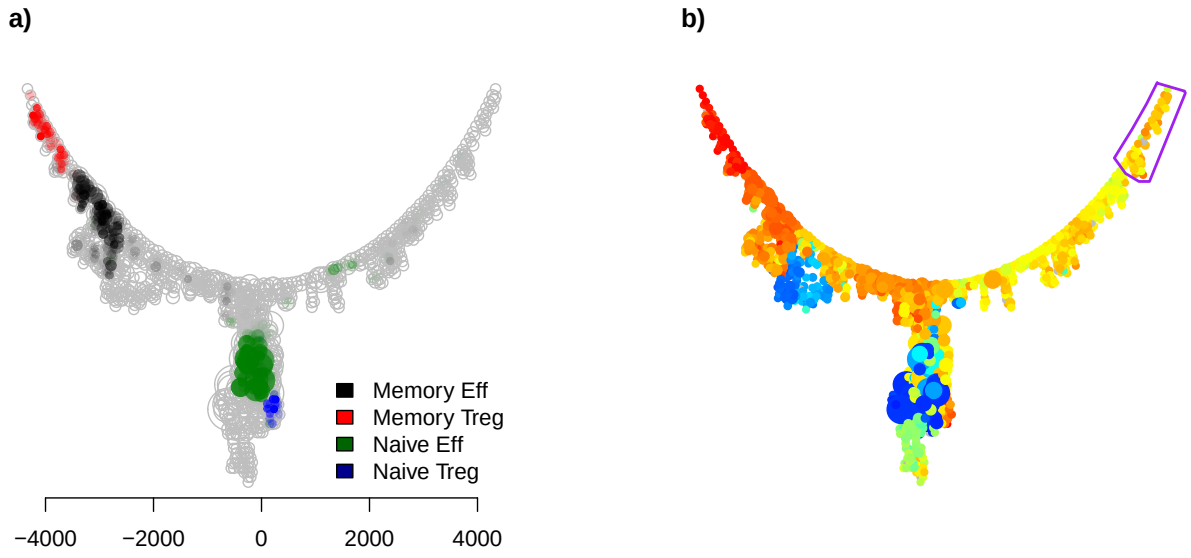


Figure 3.25. (a) Mapping of cell types defined by manual gates, memory Teffs (black), memory Tregs (red), naive Teffs (green) and naive Tregs (blue), to the MST obtained in Figure 3.24. (b) Manually identified subset of cells (purple) which respond to 1000 units but lie far from the other manually gated cell subsets. (a) The different manually gated cell types do not always segregate to different branches but can be spread across the tree. For example, naive Teffs appear in different regions of the tree. Furthermore, certain nodes of the tree can contain a mixture of cell types which complicates the interpretation. In order to guard against this, the number of clusters in the agglomerative clustering needs to be set to a high number. (b) Manual identification of a 1000 units responsive subset of cells (blue line) which lies far from the manually gated cell populations. The MST was generated on the lymphocytes stimulated at 1000 units and coloured by the pSTAT5 MFI fold increase.

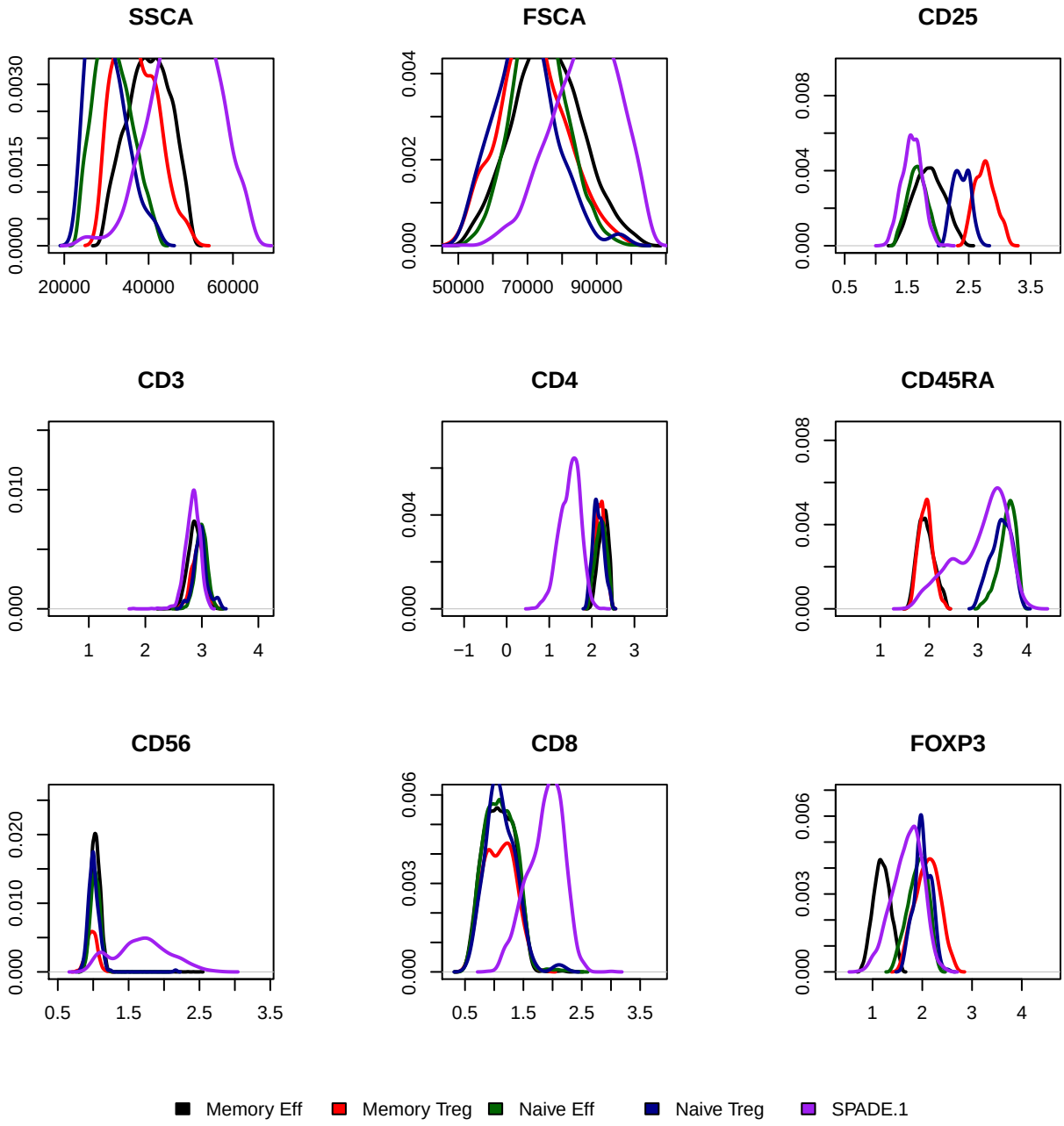


Figure 3.26. A 1000 unit responsive cell subset (purple) within lymphocytes is identified which is not assigned to any manual gate. The subset of cells, delineated in purple, manually identified in Figure 3.25, is distinct from the manually gated cell subsets memory Teffs, memory Tregs, naive Teffs and naive Tregs. Its discriminating features are that it is $CD4^{-}$, high for CD8 and CD56, while expressing moderate levels of CD25.

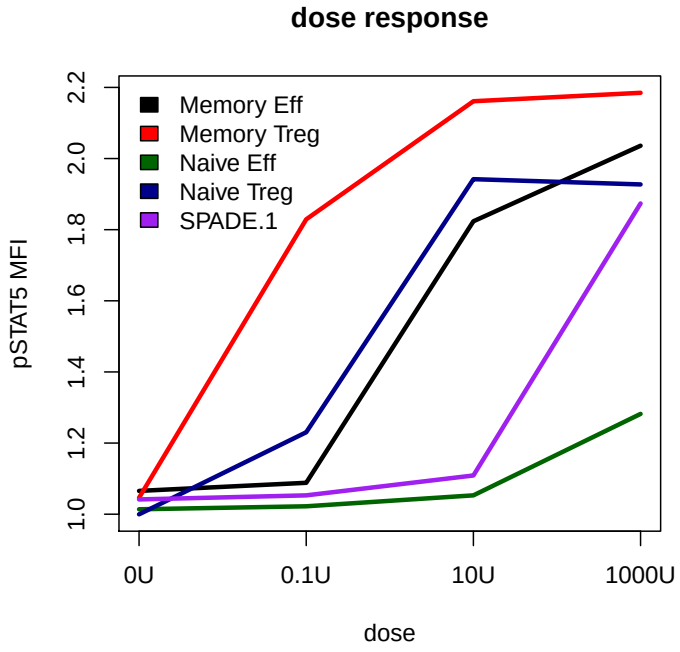


Figure 3.27. The pSTAT5 MFI dose-response of the different cell subsets within lymphocytes. In purple, the newly identified cell subsets `SPADE.1` in Figure 3.25 only shows response at 1000 units of proleukin.

Non-lymphocytes

In order to verify whether I could detect other cell subsets besides lymphocytes, which respond to proleukin, I reran SPADE on the same dataset, this time excluding cells lying within the lymphocyte scatter gate, but including forward and side scatter as markers in the clustering. While at 0.1 units little response was seen (Figure 3.28b), two small clusters showed response at 10 units (blue and purple) and a much larger cluster was responsive at 1000 units (pink) (Figure 3.28c and d).

I manually selected these groups of nodes and projected them back to forward and side scatter space to see where they lied in relation to the lymphocyte cluster (Figure 3.29). I found that these three groups cluster around the lymphocyte scatter gate which suggested that there are no detectable clusters of cells which fall within the other major scatter clusters which constitute physically larger and more granular cells such as monocytes or granulocytes (Figure 3.29). Furthermore the cells responsive to the lower 10 unit dose of proleukin (in light blue and purple) cluster closer to the lymphocyte scatter (Figure 3.29a) than the large 1000 unit responsive subset of cells (in pink) (Figure 3.29b), suggesting that the former are more likely to be lymphocytes which were not included in the manual gate. While the larger population (pink) also tends to aggregate around the lymphocyte scatter it further appears to aggregate in another potential, less well defined, scatter cluster, delineated by the pink polygon in Figure 3.29b.

Following further investigation of this subset of cells on other core markers in Figure 3.30, and after filtering of doublets on side scatter width and height, while they appeared mostly $CD3^+$, $CD4^+$ and $CD56^-$ therefore likely to be T cells or NKT cells, they also contained a small fraction of $CD3^-$ cells, hence likely to be a heterogeneous subset that may contain some monocytes.

As they are bigger on forward and side scatter than lymphocytes, they could be bigger T cell blasts, and as they are mostly $CD45RA^+$, possibly activated T cells. However,

they may well result from a technical artefact of the fixation protocol. Further markers, possibly monoycte markers such as CD19, are needed to better define this cell type and ascertain its clinical relevance.

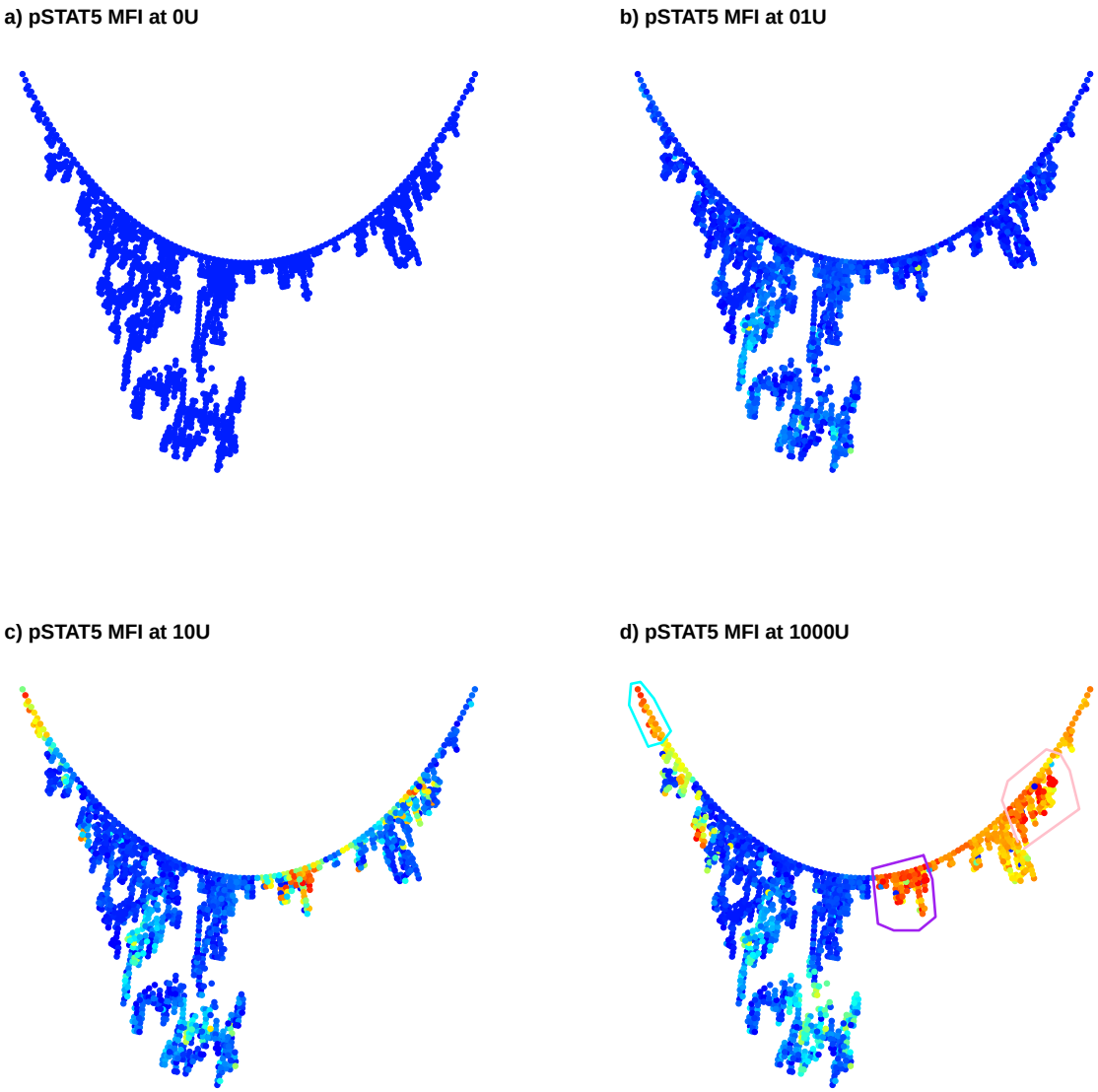


Figure 3.28. pSTAT5 MFI coloured MST generated by applying SPADE on cells which fall outside of the lymphocyte gate. Three subsets of cells are manually identified which show response to proleukin at 10 units (blue and purple), and 1000 units (pink).

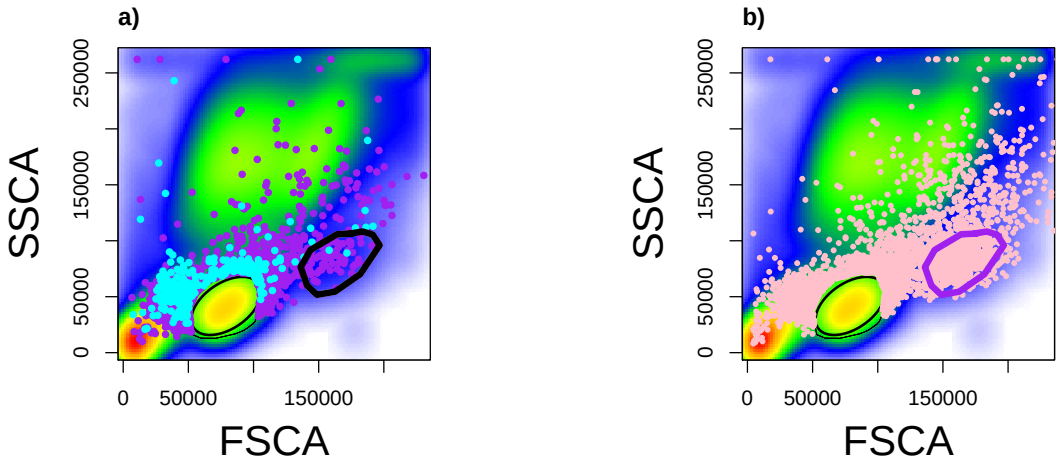


Figure 3.29. The three cell subsets, blue, purple and pink, manually identified in the MST of Figure 3.28 are mapped back to scatter coordinates. The 10 unit responsive groups, blue and purple points in (a), and the 1000 units responsive group, pink points in (b), generally tend to lie close to the lymphocyte cluster (black ellipse), but a potential secondary scatter cluster of 1000 unit responsive cells (delineated by pink polygon) are worthy of further investigation. In order to visualise where the points lie in relation to the rest of the sample they are are overlayed on top of a sample in which lymphocytes are present. While the 10 unit responsive cells cluster mostly around the lymphocyte gate, the 1000 unit responsive cells also appear to cluster within a secondary scatter cluster (pink polygon).

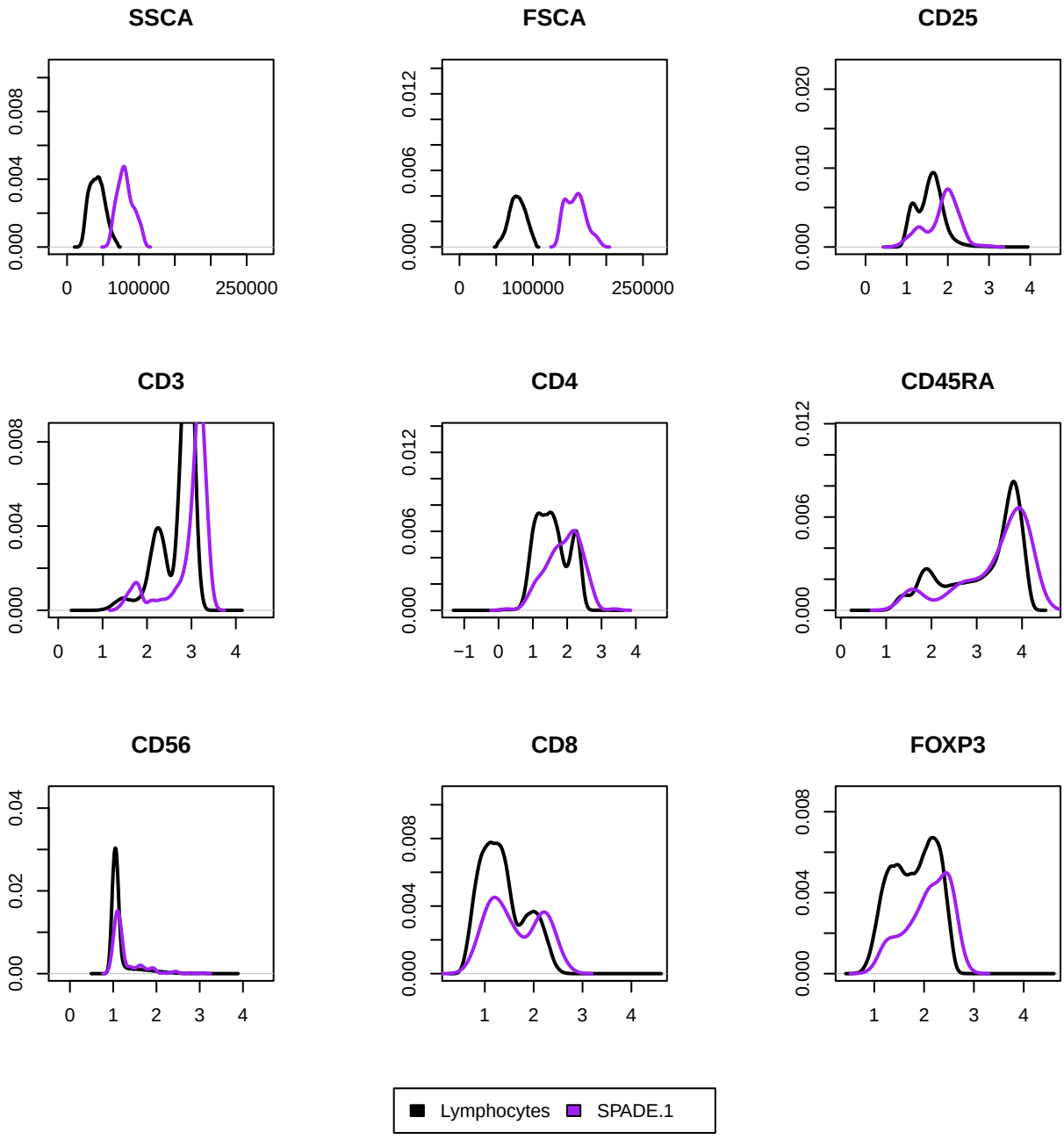


Figure 3.30. Pink cells in relation to lymphocyte cluster on all markers. After filtering of doublets on the the side scatter height, the pink cluster defined on side and forward scatter in Figure 3.29 is displayed on the other core markers. The cluster appears to be quite heterogeneous but contains predominantly $CD3^+$, $CD4^+$, $CD45^+$ and high for CD25.

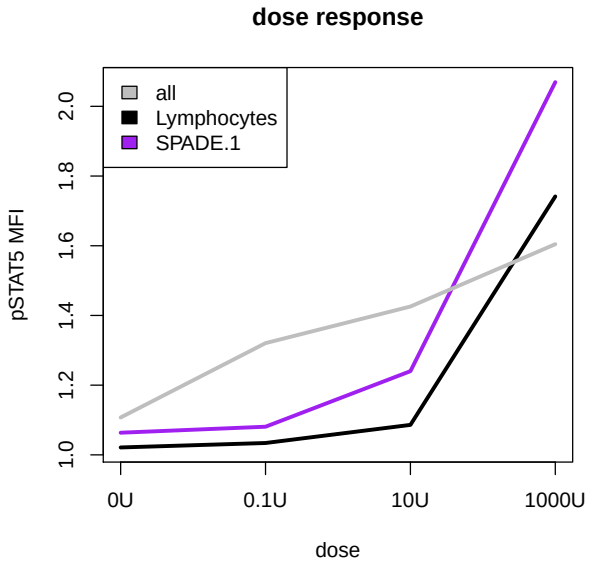


Figure 3.31. Dose-response The dashed line represents the whole sample. The black line is the lymphocyte population and the purple line is the new subset identified in Figure 3.29.

Discussion

Visualisation of cell types One general criticism is that the MST representation of the data, while visually appealing, is hard to interpret. The mapping of the manual gates to the MST is not necessarily intuitive, since cell types defined by the manual gates were spread across several nodes and branches of the tree. There was also overlap with different manually gated cell types (Figure 3.25). This can make it difficult for biologists to interpret the MST. The MST requires some annotation in order to understand where the different cell types lie. An obvious visualisation is to colour the tree according to each marker individually. However this approach is clearly not practical for a large number of markers, nor does it yield an overview of the relationship between the various markers. It is also over reliant on coding information as colour patterns which tends to be harder represent and assimilate than spatial information. Instead, a more useful alternative is to plot the tree node coordinates against the core marker node MFI, as illustrated in Figure 3.32. This approach not only provides some insight into the marker progression, at least along the main branch of the tree where the different cell types lie, but also into the relationship between the markers in the sample. However, using the absolute node coordinate is misleading since it is relative to the layout algorithm, hence this approach is better applied to following individual branch paths. Potentially, this approach could be repeated along each branch in the MST to identify the different types of cells progression in the sample. This would rely on the definition of a root node, an idea which needs to be explored further.

Sometimes from the MST it may be difficult to delineate the cluster of cells which are responsive as they do not cluster in distinct sections of the tree. Instead one could resort to PCA or PLS using pSTAT5 as the response variable.

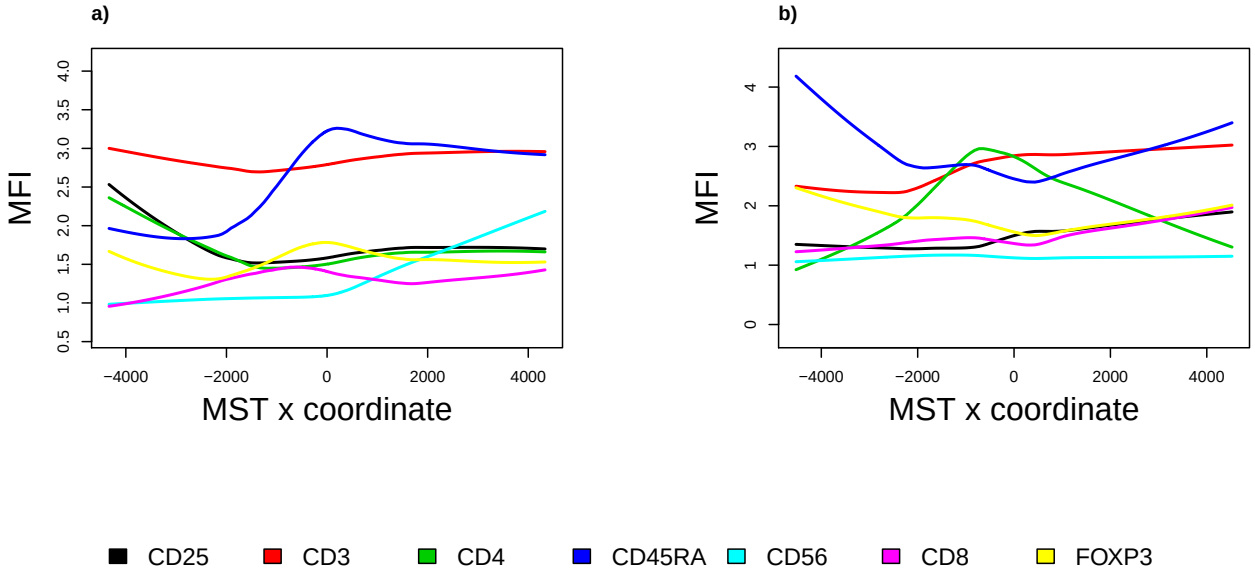


Figure 3.32. The progression of the marker MFI along the horizontal coordinate of the MST nodes in lymphocytes (a) and non-lymphocytes (b). The lowess smoothed progression of the marker MFI along the horizontal coordinate of the MST. For the MST constructed on the lymphocytes (a), the markers which show the clearest progression are markers, CD45RA CD56 which increase from left to right, and markers CD4 and CD25 which decrease. For the MST constructed on the non-lymphocytes (b), the progression of the markers is not monotonic.

3.4.2 Visualising response in whole sample with recursive partitioning

The SPADE approach first clustered on the core markers across batches, then built the MST from these, allowing for visual identification of clusters with a high pSTAT5 response. The clustering was preceded by a downsampling step which aimed to make the density more uniform across the sample so that all points had an equal probability of being represented as part of a distinct cluster. An alternative to clustering across samples, is to use recursive partitioning instead to split the core marker space into bins containing roughly the same number of events across samples within the same batch. This can be achieved by recursively splitting on the median of each marker, so that at each split, half of the dataset is assigned to each branch. The process is applied recursively to each bin until a minimum bin size or maximum number of recursive steps is reached. Typically, the order in which the markers are selected is guided by picking the marker with the largest variance or range at each split. Since each bin contains approximately the same proportion of events, this implies that the binning is finer in regions of high density and coarser in regions of low density. Conceptually, this is another approach of reducing the number of events while preserving lower density regions, provided the number of bins is sufficiently large, similar to the method of downsampling. Recursive partitioning was first introduced to flow cytometry by Roederer et al (2001), under the name of "probabilty binning", as a means of translating a multivariate distribution into a univariate one, in order to test statistical significant differences in event counts between individual bins or whole samples. The algorithm was later implemented in the `R` BioConductor package flowFP (Holyst and Rogers, 2009) as "flow cytometric fingerprinting" and has been applied to discriminating cell subsets which differ significantly in proportion between healthy controls and actute myeloid leukemia (AML) patients (Rogers et al, 2008; Rogers and Holyst, 2009).

Binary recursive partitioning on core markers For the purpose of visualisation, I first illustrate the recursive partitioning algorithm on side and forward scatter using 128 bins (Figure 3.33). The binning was defined by pooling all four samples on side and forward scatter. Since each bin should contain approximately $\frac{1}{128^{th}}$ of the events, finer binning is applied to higher density regions and coarser binning to sparser regions. Applying the same binning across the four samples, the relative proportion of events

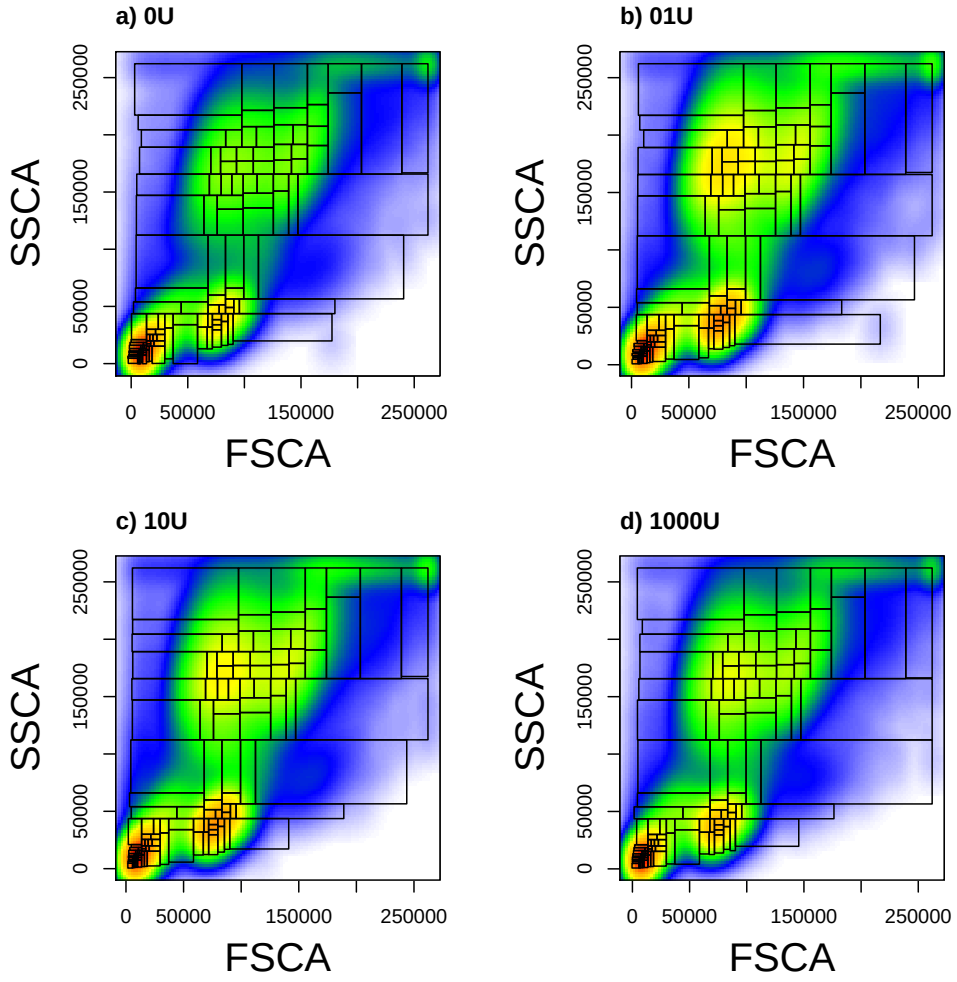


Figure 3.33. Sample recursively partitioned into 128 bins on side and forward scatter. The binning is determined in the resting sample (a) and then the same binning is applied across all samples (b, c and d). While each each bin contains the same number of events in the resting sample (a), this does not necessarily hold in the other samples (b, c and d).

assigned to each bin varies between samples (Figure 3.34). If the number of bins is increased each bin represents a smaller proportion of the sample so consequently the variations between samples should also become smaller. The pSTAT5 response on side

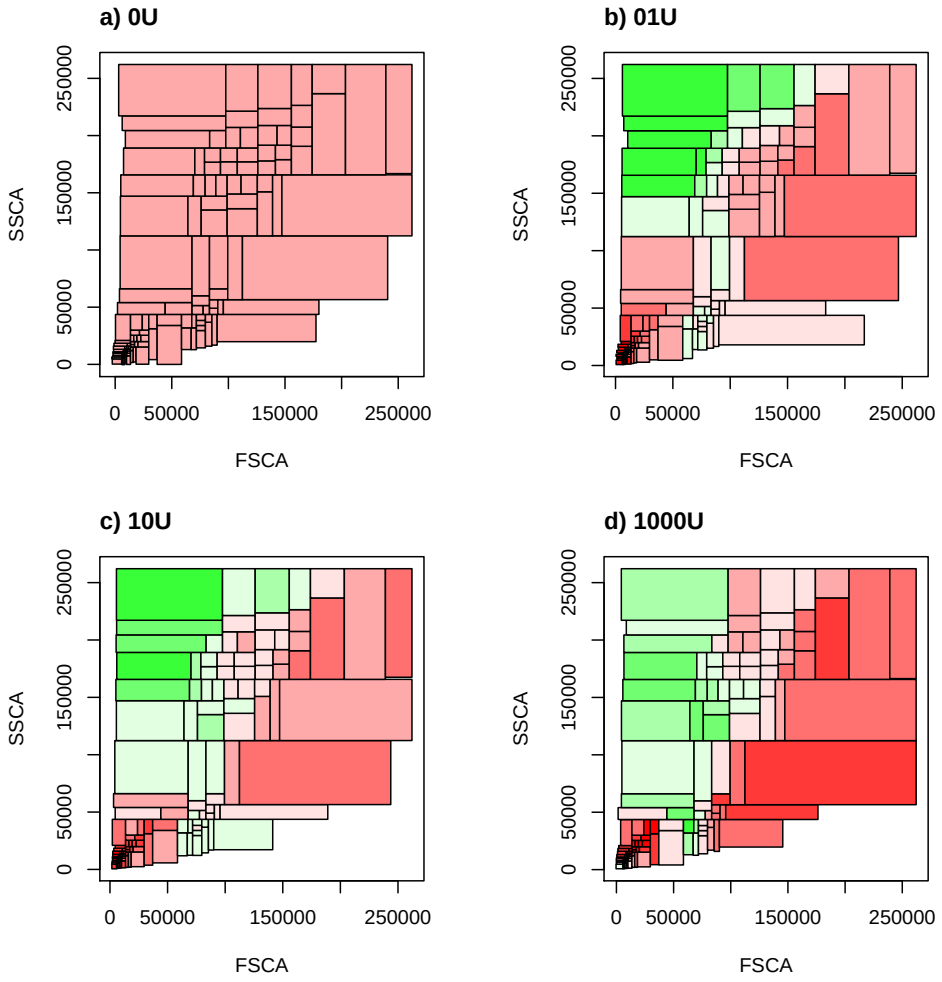


Figure 3.34. Each sample is recursively partitioned using 128 bins. The colour indicates wherever the proportion increases (green) or decreases (red) relative to the mean proportion of each bin across all samples.

and forward scatter is visualised in Figure 3.35 by colouring each bin by its median pSTAT5 response. No pSTAT5 is visible in any of the 128 bins at 0.1 or 10 units (Figure 3.35b and c) which suggest that the proportion of 0.1 unit and 10 unit responsive cells is too small within each bin to influence the pSTAT5 median. However at 1000

units, bins which overlap with the lymphocyte cluster show clear response as well as the bin which overlaps with the uncharacterised cells delineated in pink in Figure 3.29b.

Lymphocytes Extending recursive partitioning to all core markers, I aimed to identify dose-responsive cells not assigned to any manual gate within the lymphocyte subset. The number of bins was increased to 1024 and the recursive partitioning was ran in the core markers CD25, CD3, CD4, CD45RA, CD56, CD8 and FOXP3, on the lymphocyte subset, after excluding doublets. I also excluded cells which were assigned to the manually gating subsets, both memory and naive, Teffs and Tregs, in order to focus on potentially unidentified cell subsets. Since for any two dimensional projection of the data, many bins overlap, I used the same MST visualisation as described in the previous section, where each node this time represents the core marker median of one of the 1024 bins (Figure 3.36). Using the MST display, I was able to identify a cluster of cells which responded to 1000 units. From the MST, I visually identified two responsive cell subsets, a 10 unit responsive one (in pink) and a 1000 unit responsive cluster (delineated in purple). Selecting the tree nodes manually and projecting the corresponding bins back to marker space, I plotted these two cell subsets in relation to the manually gated subsets, memory Teffs (black), memory Tregs (red), naive Teffs (green) and naive Tregs (blue) (Figure 3.37). The cell subset delineated in pink constitutes around 4 percent of the lymphocytes. This cell subset contains moderate levels of CD25 similar to a memory Teffs and is $CD3^+$, $CD4^+$, $CD8^-$, $CD45RA^-$ and $FOXP3^-$. It is likely to represent transitional cell population between memory Teffs and Tregs which was not included in the manual gating. On the other hand, the cell subset delineated in purple appeared to be $CD3^-$, $CD4^+$, $CD8^-$ and high for CD56. These cells were also low in CD25, with the same level of expression as naive Teffs, this explains there limited response at lower doses. They constitute around 1.5 percent of the total lymphocyte population within this sample. These CD56 bright cells include $CD3^-$ cells so could belong to the cell

subset currently under investigation by Charlie Bell using RNAseq, which are known to express high levels of CD122 (the beta chain of the IL2 receptor).

Non-lymphocytes Recursive partitioning was next applied to non-lymphocytes, in order to detect potentially new responsive subsets. As side and forward scatter need to be included in the recursive partitioning of non-lymphocytes, I scaled the scatter so as to have a similar range to the fluorescent markers. I again used 1024 bins although this number could have been increased because we are dealing with larger number of cell. The recursive partitioning was ran on the same markers but this time also included the side and forward scatter. Once more, the MST was used to visualise the response across the whole sample (Figure 3.39). At 0.1 units, no clear response is visible (Figure 3.39b) but at 10 units (Figure 3.39c) certain nodes show moderate response within a branch of the MST which later become part of cluster (delineated in purple) which shows strong response at 1000 units (Figure 3.39d). Projecting the cells contained within this cluster back to marker space (Figure 3.40), they seem to lie close to the lymphocyte cluster (black) on side and forward scatter and also appear to contain both $CD4^{-}$ and $CD4^{+}$ cells. They also overlap on side and forward scatter with the cluster identified in Figure 3.29.

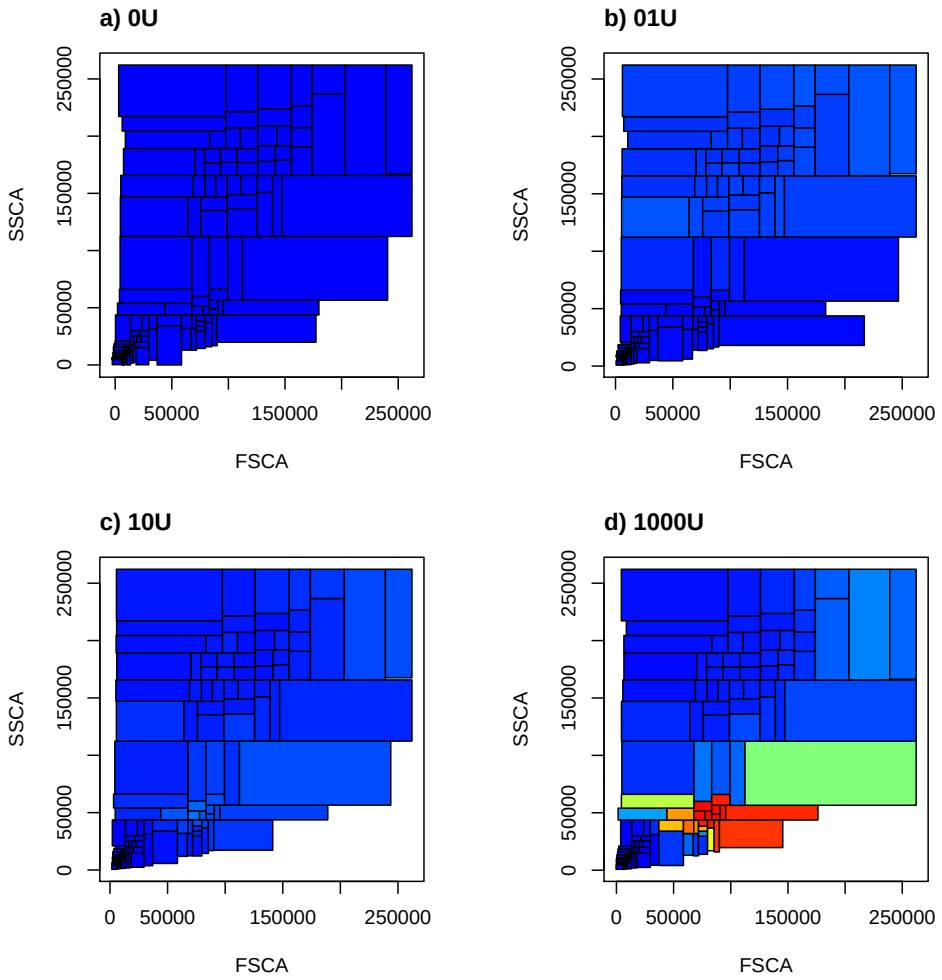


Figure 3.35. pSTAT5 response in sample recursively partitioned into 128 bins on side and forward scatter. No pSTAT5 response is visible in any of the bins at 0.1 (b) or 10 (c) units. Hoever at 1000 units (d), the bins which overlap with the location of the lymphocytes on side and foward scatter show strong response. Also the bin which coincides with the location of the cluster in Figure 3.29 shows some moderate response confirming that there is likely to be some responsive cells lying within that cluster.

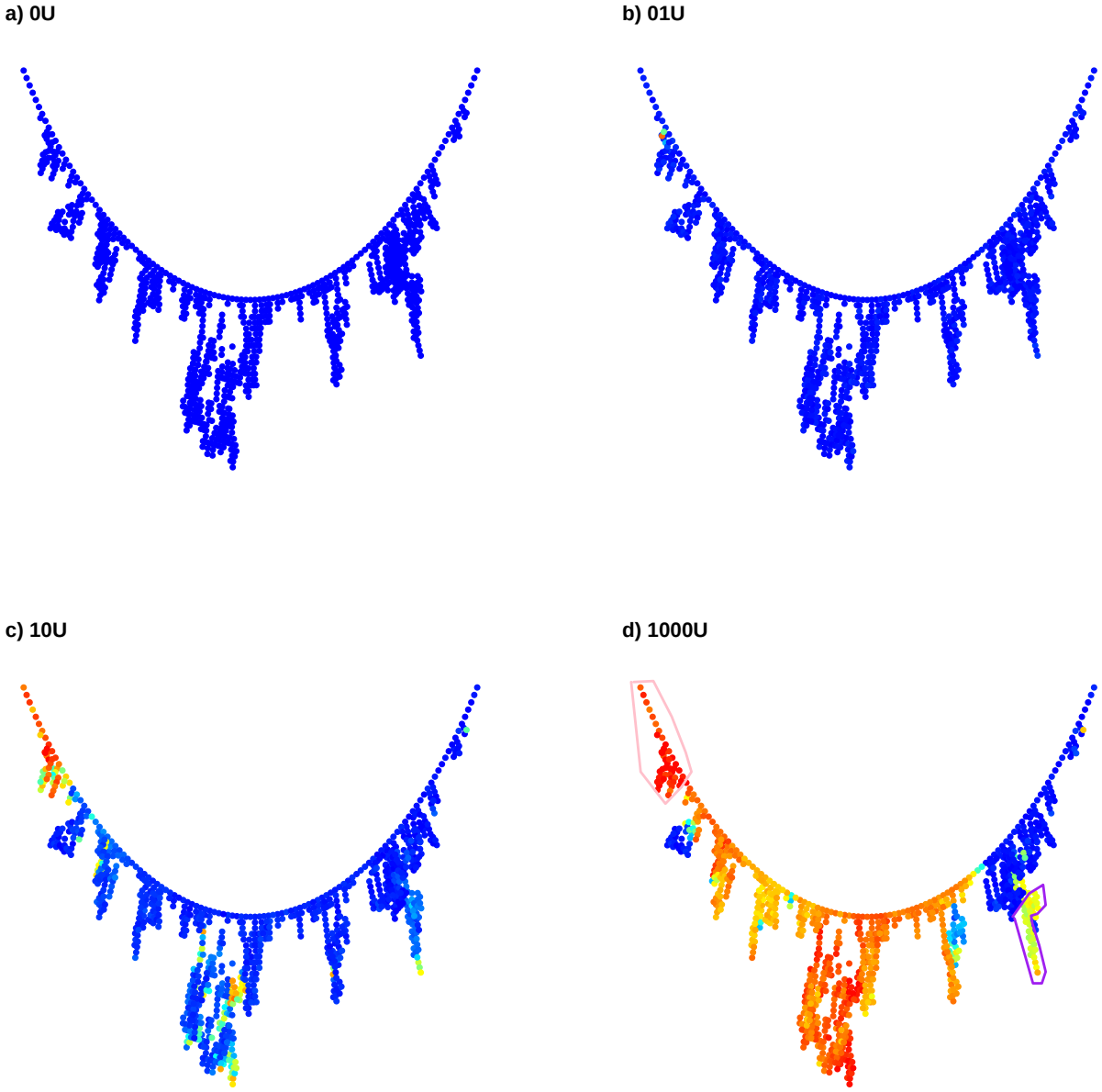


Figure 3.36. MST built using on the 1024 bins obtained from recursive partitioning on the lymphocytes core markers. A cluster of cells (delineated in purple) lies far from the rest of the cells which shows pSTAT5 response at 1000 units.

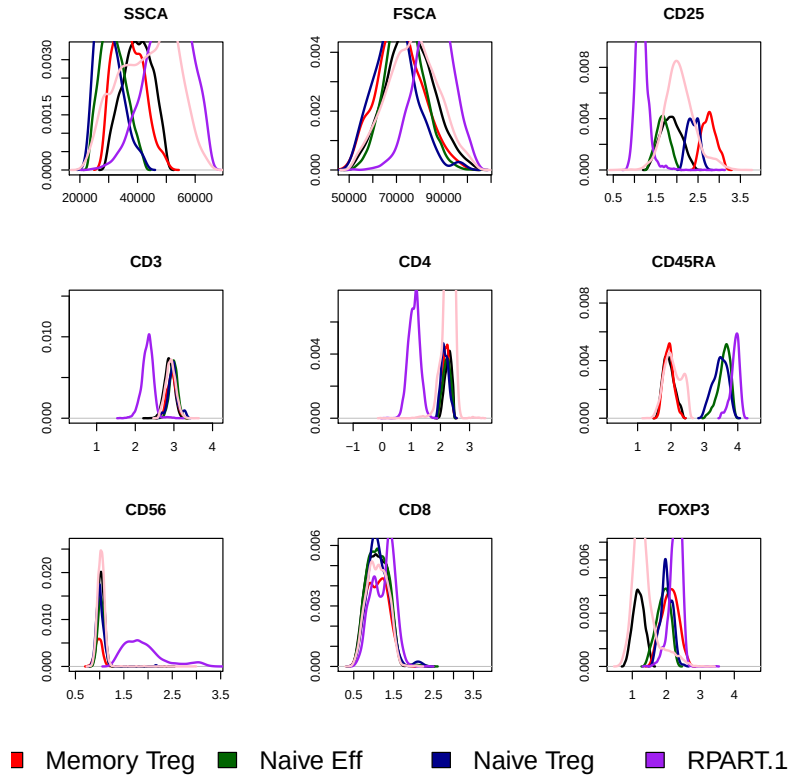


Figure 3.37. A 10 unit responsive cell subset (pink) and a 1000 unit responsive cell subset (purple) within lymphocytes are identified which are not assigned to any manual gate. Two subset of cells, a 10 unit responsive subset delineated in pink and a 1000 unit responsive subset delineated in purple, manually identified in the MST of Figure 3.36, which are distinct from the manually gated cell subsets memory Teffs (black), memory Tregs (red), naive Teffs (green) and naive Tregs (dark blue), are pojected back to core marker space. The pink cell subset overlaps with the manually identified cell subsets Its discriminating features are that it is $CD3^-$, $CD4^-$ and high in CD56, while expressing low levels of CD25.

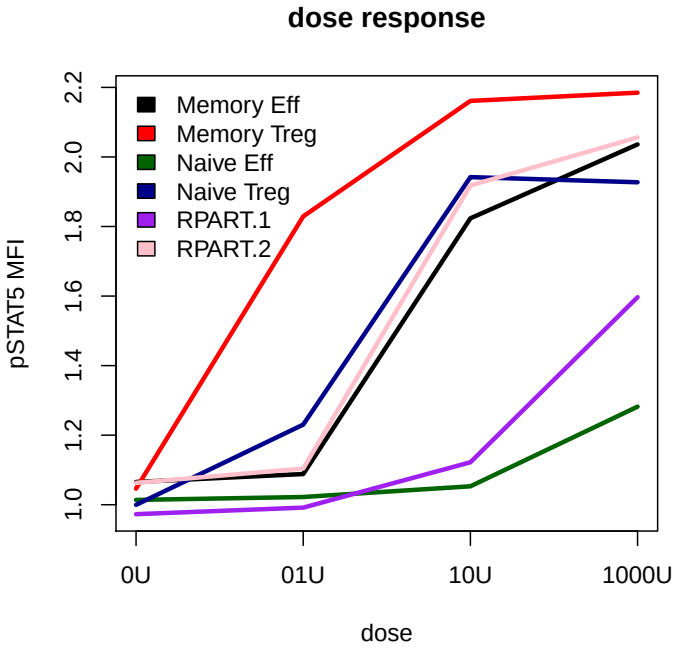


Figure 3.38. Dose response

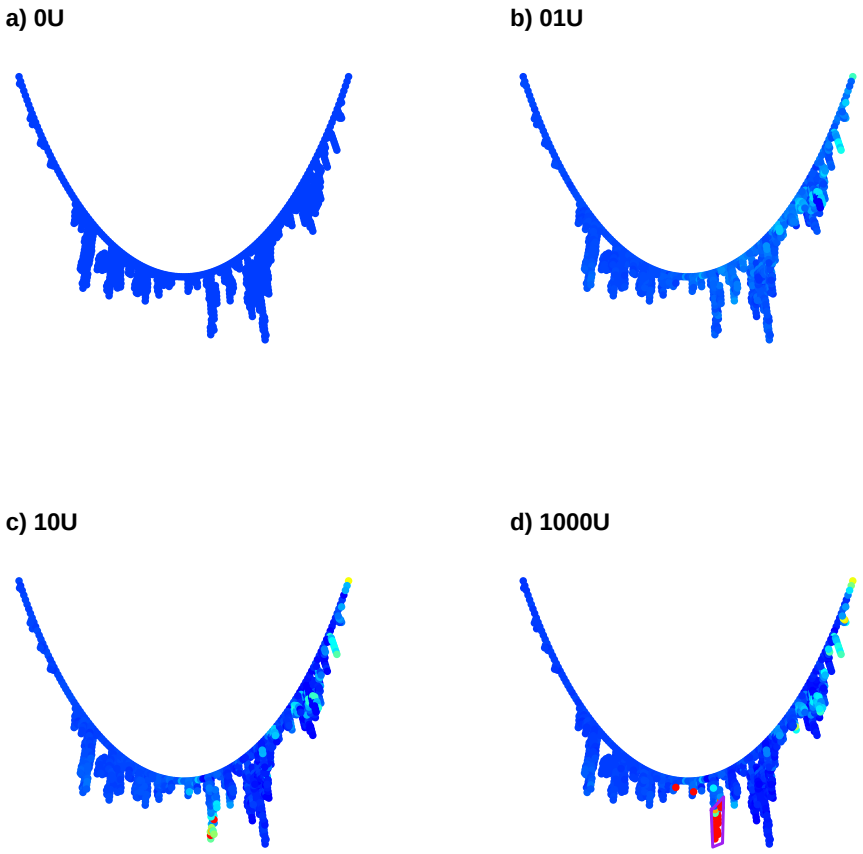


Figure 3.39. MST built on the 1024 bins obtained from rpart on non-lymphocytes, a cluster of cells stands out from the rest which shows pSTAT5 response at 1000 units.

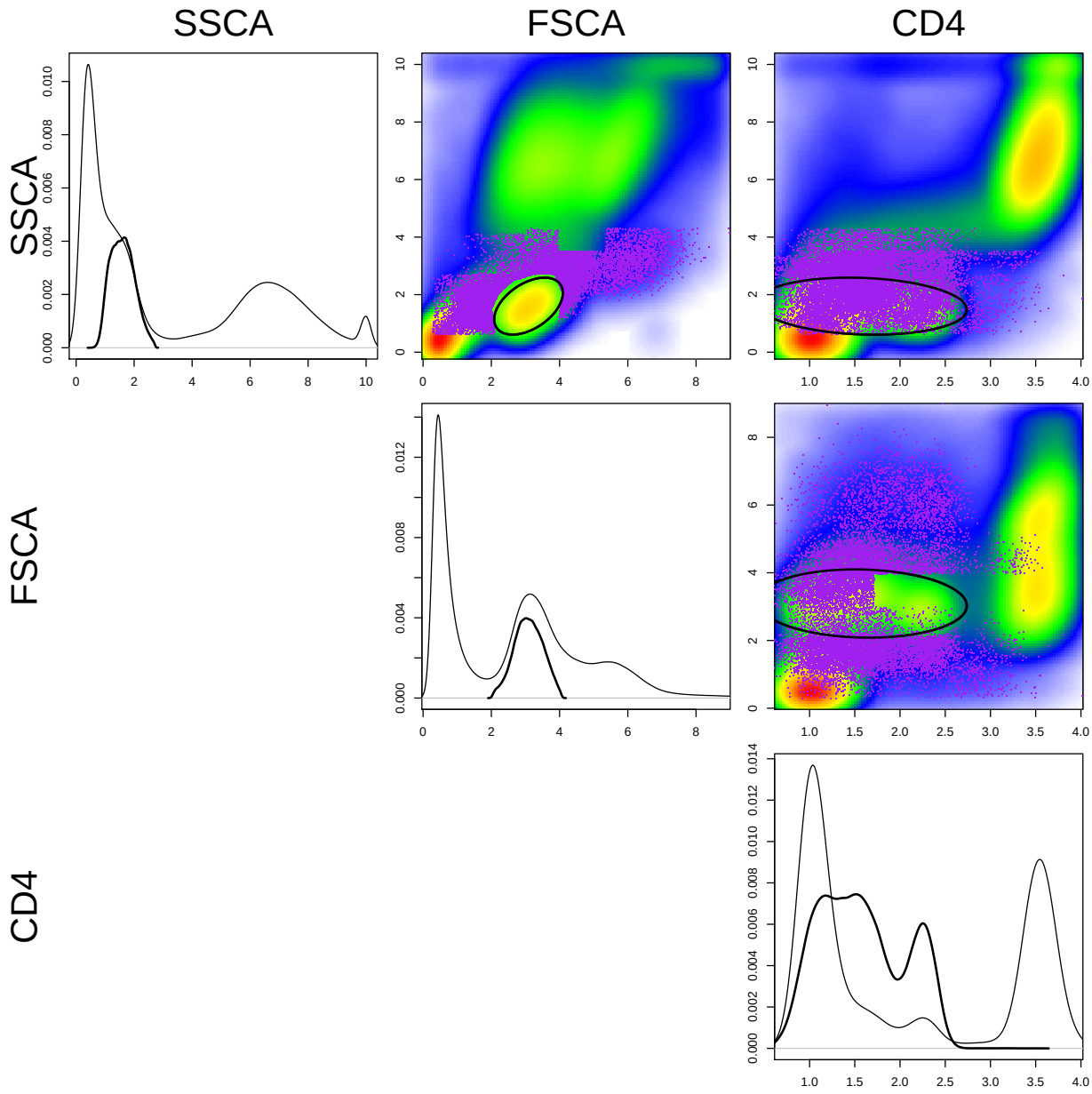


Figure 3.40. A 1000 unit responsive cell subset (purple) is identified which does not belong to the manually define lymphocytes population (black). In order to visualise where the points lie in relation to the rest of the sample they are are overlaid on top of a sample in which lymphocytes are present. The subset of cells, delineated in purple, was manually identified in Figure 3.39.

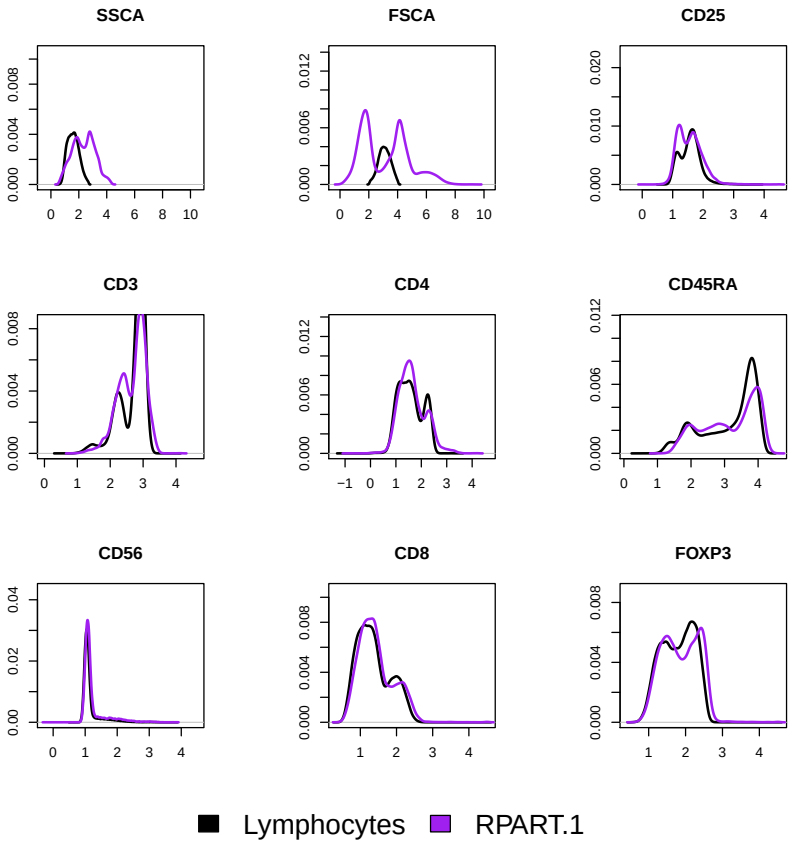


Figure 3.41. A 1000 unit responsive cell subset (purple) is identified which does not belong to the manually define lymphocytes population (black).

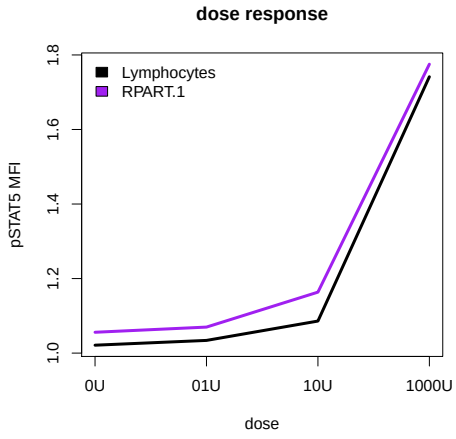


Figure 3.42. A 1000 unit responsive cell subset (purple) is identified which does not belong to the manually define lymphocytes population (black).

Binary recursive partitioning using regression tree on pSTAT5 The SPADE and recursive partitioning methods described in the previous sections have proceeded by first reducing the number of events through clustering or binning on the core markers across stimulation doses, then visually identifying clusters or bins which show pSTAT5 response using a two-dimensional MST projection of the reduced dataset. The clusters or bins are then projected back to core marker space to examine their MFI and relative size.

However since the objective is to identify dose responsive cell subsets, a logical extension to the methods described previously would be to include the pSTAT5 response in the clustering of these datasets. One way this can be achieved is to build on the recursive partitioning approach from the previous section, by applying the classification and regression tree (CART) method to the pSTAT5 response. Instead of using the variance of the core markers, the CART uses the variance of the pSTAT5 response to guide the recursive partitioning. This approach however requires the pSTAT5 response to be known at each point in the dataset. Such a dataset can be constructed by using the ANN algorithm to join samples from the same batch on their core markers as was explained in Section 3.3.1. In fact, the ANN algorithm uses the recursive partitioning technique covered in the previous section, on the core markers to build a K dimensional tree (KD-tree) data structure. A KD-tree serves as an indexing data structure, allowing faster retrieval of data points based on their coordinates by refining the search to the bin within which the point lies. This indexing is exploited to efficiently find the approximate nearest neighbour between datasets.

The CART, as implemented in the `R` package `tree` (Ripley, 2014), proceeds by considering each core marker coordinate as a potential splitting point. The splitting point which minimises the sum of the within branch variance of the response variable, is selected and the data is split between the left and the right branch. Note that contrary

to the recursive partitioning scheme defined in the previous section, since the split point does not usually correspond with the median, the tree is not balanced. This splitting is applied recursively until some minimum leaf node size is reached or the reduction in variance from splitting reaches some threshold (the default is 0.01 of the total variance). A leaf node represents a partition obtained by applying the cuts defined along the branch. In order to reduce the number of partitions, the tree can be pruned to minimise the cost-complexity for a desired number of leaf nodes. On the same ungated sample as was used in the previous section, partitioning only on side and forward scatter using pSTAT5 response at 1000 units and pruning the tree to the best three subsets, I obtained the clustering in Figure 3.43. This confirms that based on forward and side scatter alone, the lymphocyte cluster is the most responsive cluster to proleukin. However, if the pSTAT5 of the sample stimulated at 0.1 or even 10 units is used as the response, then the CART algorithm does not consistently partition the data, since few cells within the lymphocytes respond at these doses of proleukin and so the reduction in the total variance is not sufficient to justify a split.

Lymphocytes I first ran CART on lymphocytes, excluding all manually gated $CD4^+$ cells, in order to see if any dose-responsive non-manually gated cell subsets were identified. The CART was ran without pruning on the ANN joined dataset using the pSTAT5 response at each of the stimulation doses. In Figure 3.44 are the trees obtained from 0.1 (a), 10 (b) and 1000 units (c).

At 0.1 units, only two subsets can be distinguished based on CD25. At 10 units, four clusters are distinguishable based on CD45RA, CD56, CD25 and CD4. However, at 1000 units, only three subsets are discernible. This may be because there is more homogeneity in the response as an important proportion of the lymphocyte subset will have reached saturation at that dose. Furthermore CD25 does not feature in the regression tree, which may be because once the response is saturated, CD25 adds little predictive value. Since

these regression trees include only a few of the available markers, their utility is rather limited in identifying cell subsets. Nonetheless, some information can be extracted. For example, the inclusion of the CD56 marker at 10 and 1000 units suggests that it becomes a relevant predictor of the pSTAT5 response. In particular at 10 units, where the CD45RA$^-$ CD25$^+$ subset shows the strongest response, the CD45RA$^+$ CD56^{hi} subset shows the second strongest response. At 1000 units, the highest response is in the CD3$^+$ subset, but the CD3$^-$ CD56$^+$ subset is the second highest. This points to the same CD3$^-$ CD56$^+$ subset that was identified from the MST in Figure 3.37 at 1000 units of stimulation.

Non-lymphocytes Next, I repeated the analysis, including side and forward scatter, on non-lymphocytes (Figure 3.45). Side and forward scatter were scaled so as to have a comparable variance to the other parameters. At 1000 units, the strongest response comes from the subset with low side scatter, high CD25 and high CD3. This subset includes lymphocytes and the cells identified in Figure 3.40. Since the cells are high in CD3, they are also likely to include T cells which would overlap with the pink and purple cell subsets defined in Figure 3.28.

Discussion A drawback of recursive partitioning techniques is that they are prone to overfitting and consequently very sensitive to batch effects. For example, when the same algorithm is applied to the lymphocyte subset in another sample, the returned partitioning is very different (Figure 3.46). Another issue with the regression tree approach is that it doesn't exploit the bimodality of the pSTAT5 distribution at higher doses of proleukin. This motivates the next approach, which instead of using the variance of the pSTAT5 response in the partitioning, fits a bimodal distribution.

There are possible extensions to the regression trees methods. In the "Elements of Statistical Learning" (Hastie et al, 2009), extensions of this algorithm are mentioned

which allow for a linear combination of more than one marker at each split, or for more than one split at each level. multivariate adaptive regression splines (MARS) for example uses multiple additive regression splines as a generalisation of stepwise linear regression or a modification of the CART method to improve its performance in the regression setting. MARS foregoes the tree structure. Instead of approximating each bin by the pSTAT5 mean, the pSTAT5 function is piecewise linear with knots at core markers points. Unfortunately I found the R implementation to be too slow, making it impractical to run on this dataset. Another extension is the Patient Rule Induction Method (PRIM) also known as the "bump hunting" algorithm introduced by Friedman and Fisher (1999). PRIM searches for a bounding boxes in the marker space in which the average response is high. No binary constraint is placed which can make the individual rules simpler. The main box construction method in PRIM works from the top down, starting with a box containing all the data. The box is compressed along one face by a small amount, and the observations then falling outside the box are peeled off. The face chosen for compression is the one resulting in the largest box mean, after the compression is performed. Then the process is repeated, stopping when the box contains some minimum amount of points. After the top-down sequence is computed, PRIM reverses the process expanding along any edge, if such an expansion increases the box mean. Since the top-down procedure is greedy at each step such an expansion is often possible. The result of these steps is a sequence of boxes, with different number of observations per box. An advantage of PRIM over CART is its patience. Because of its binary splits, CART fragments the data quite quickly whereas PRIM is more progressive. This however comes at the cost of performance and I found the R implementation of this method to be too slow to run on my dataset. How well these variants perform depends on the data but generally they tend to over-partition data so that the bins are uneven or the tree is unbalanced.

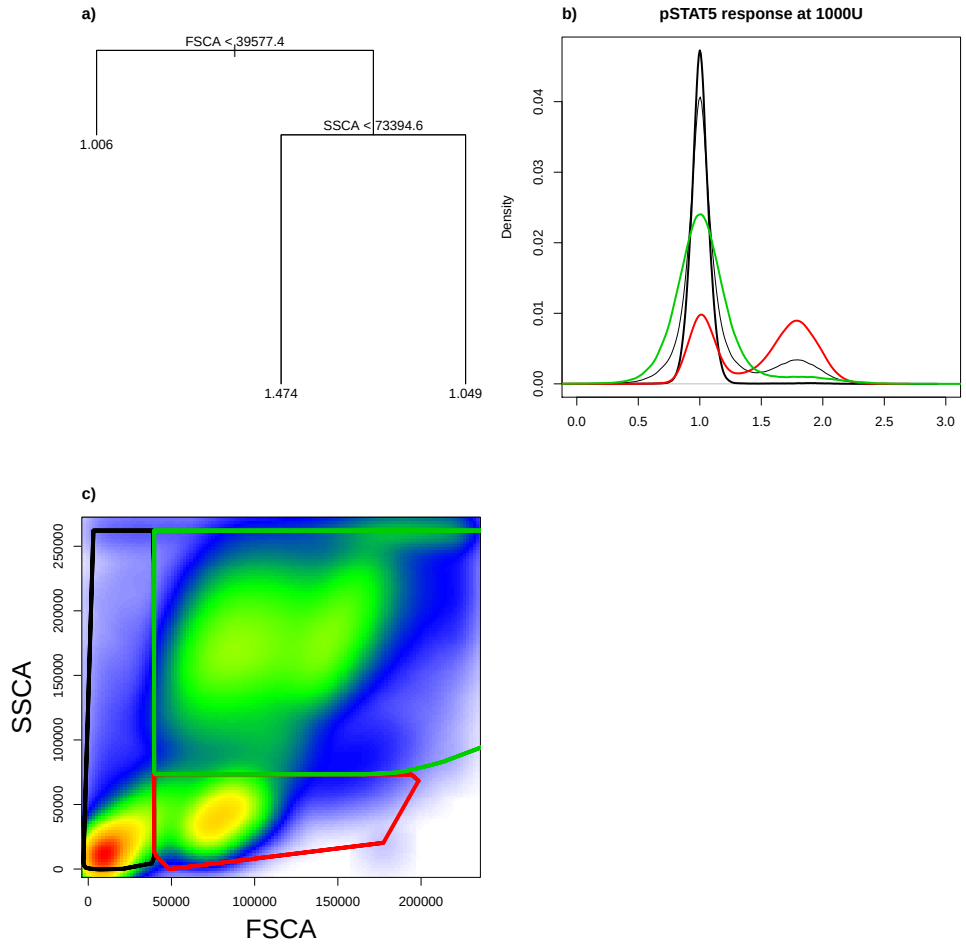


Figure 3.43. CART of 1000 unit response against side and forward scatter identifies three subsets. The regression tree obtained from recursive partitioning of side (SSCA) and forward (FSCA) scatter against the pSTAT5 response at 1000U, after pruning on the best three subsets (a). The values at the terminal nodes are the expected pSTAT5 response within each subset. According to this regression tree, most of the pSTAT5 response comes from the lymphocyte cluster (red) whereas the black and green clusters are less responsive (b).

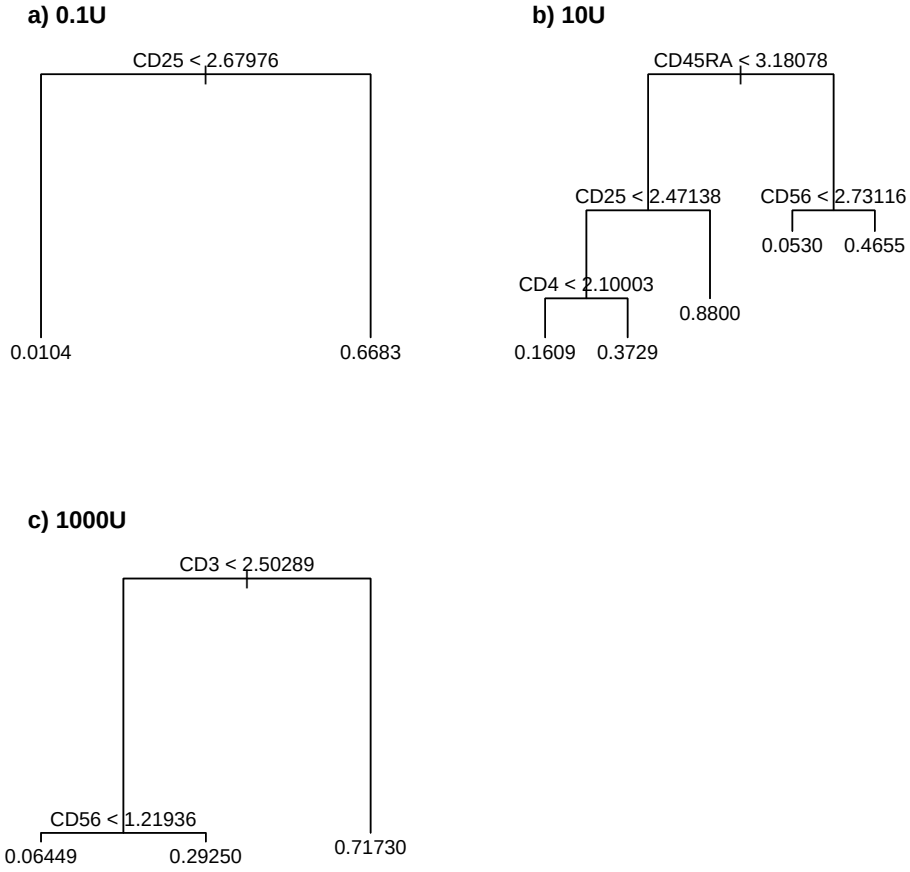


Figure 3.44. The recursive partitioning tree obtained for the pSTAT5 at 0.1 units (a), 10 units (b) and 1000 units (c) in the lymphocytes which do not belong to any manually identified cell subset. Each non-leaf node of the tree represents a split point where the dataset is partitioned along the left or the right branch according to the inequality. The numbers at the bottom of each leaf represent the mean pSTAT5 response within that partition of the data. The markers which are selected as split points differ depending on the doses. The height of each branch is proportional to the reduction in variance which results from that split.

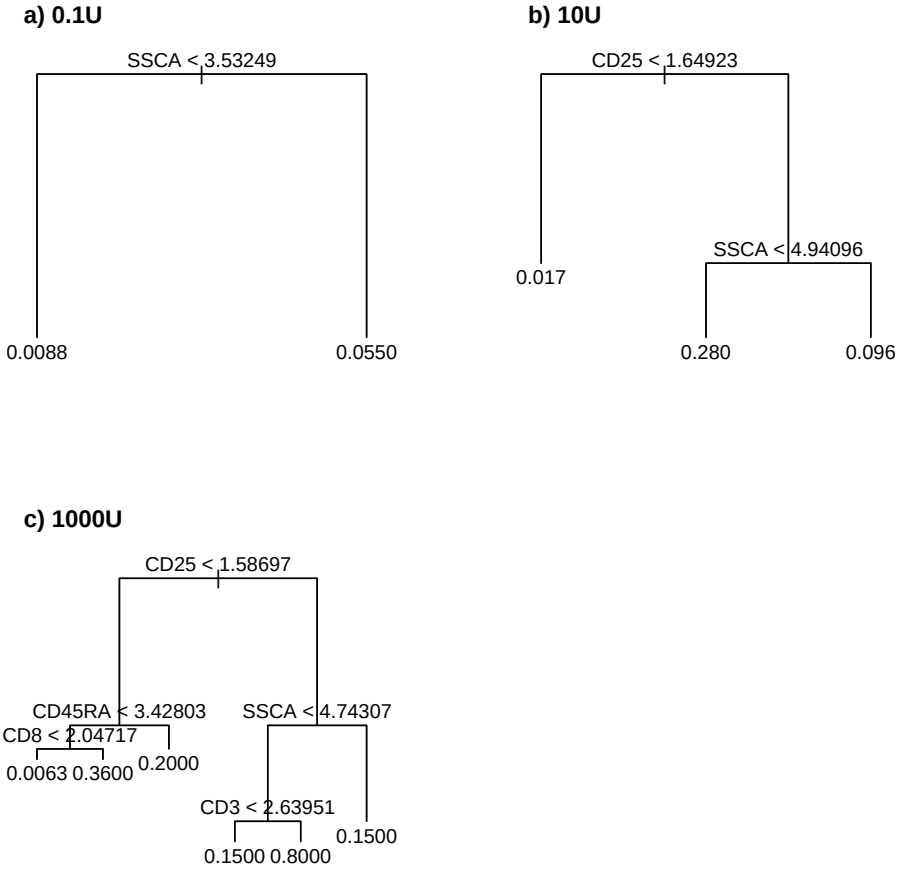


Figure 3.45. The recursive partitioning tree obtained for the pSTAT5 at 0.1 units (a), 10 units (b) and 1000 units (c) in the non-lymphocyte cells. The numbers at the bottom of the tree represent the mean pSTAT5 response within that partition of the data. The height of each branch is proportional to the reduction in variance which results from that split.

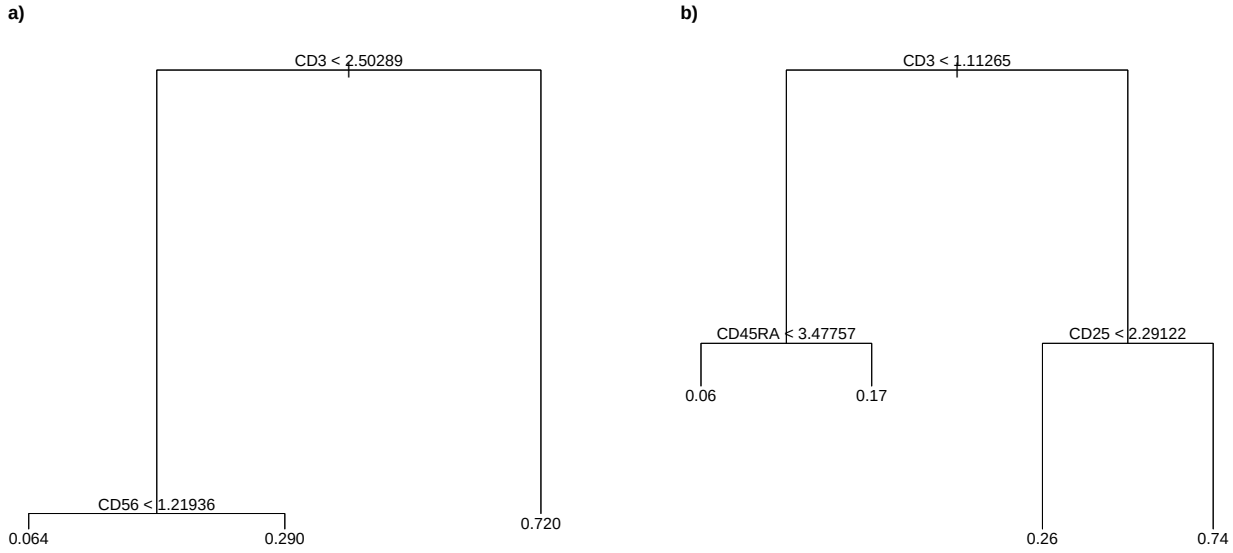


Figure 3.46. Partition tree obtained in two different sample from running CART on all core markers against pSTAT5 response at 1000U. While both trees agree on the initial partitioning on CD25 and further partitioning on CD45RA in the first branch, the partitioning in the second branch is done on different markers and so not comparable.

3.4.3 Alternative visualisation using partial least squares

I have only explored the MST which is a non-linear representation of a multivariate dataset. No information about the pSTAT5 response was captured in the layout of the tree but instead I used colour to visually identify clusters which respond to pSTAT5. An alternative may be to include the pSTAT5 in the multidimensional scaling using a technique known as partial least squares (PLS). PLS is based on the well known PCA method which uses single value decomposition of the covariance matrix in order to identify orthogonal components which more succinctly represent the variation in the sample. However while in PCA all variables are treated the same, in PLS, variables can be specified as response variable Y or predictors X. The general underlying model of multivariate PLS is then:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\top} + \boldsymbol{E}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^{\top} + \boldsymbol{F}$$

where $\boldsymbol{X}$ is an n $\times$ m matrix of predictors and $\mathbf{Y}$ is an n $\times$ p matrix of responses. $\boldsymbol{T}$ and $\boldsymbol{U}$ are n $\times$ l orthogonal matrices that are, respectively, projections of $\boldsymbol{X}$ and projections of $\boldsymbol{Y}$. The matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ are, respectively, m $\times$ l and p $\times$ l orthogonal loading matrices. The matrices $\boldsymbol{E}$ and $\boldsymbol{F}$ are the error terms, assumed to be independent and identically distributed random normal variables. The decompositions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are made so as to maximise the covariance between $\boldsymbol{T}$ and $\boldsymbol{U}$.

I applied partial least squares regression, fitted with the kernel algorithm using the `R` function `plsr` function provided in the `R` package `pls` (Mevik et al, 2013), to both the lymphocytes and non-lymphocytes.

Lymphocytes I ran the PLS regression of the pSTAT5 response at each dose, within the manually gated lymphocyte subset (Figure 3.47). I found that manually gated subsets generally project to distinct clusters. However the naive Teffs and naive Tregs

greatly overlap and that may be due to spurious correlation created by spillover between CD45RA and FOXP3, which makes the naive Teffs look abnormally high in FOXP3, as is apparent in the FOXP3 channel in Figure 3.48, Based on the PLS projections at 1000U, three additional clusters, delineated in purple, pink and light-blue, were manually identified (Figure 3.48). These clusters were then plotted across all doses. As can be appreciated the known cell populations tend to be better distinguishable at 0.1U and 10U, while the three new clusters are better separated at 1000U. This is likely the consequence of the pSTAT5 response being more distinguishable for the known subsets at 0.1U and 10U, while at 1000U, the response is saturated and so undistinguishable. Plotting the dose response in Figure 3.49 shows that of the three new subsets, only the purple subset is responsive at 1000U while the pink and light-blue subset conserve baseline pSTAT5 even at the highest dose. The light-blue and pink subsets are therefore not of interest here and so only the purple subset is considered for further study. In Figure 3.48, the purple subset is plotted in relation to the know subset on all core markers. Its main distinguishing features are that it is $CD8^+$ and $CD4^-$. It is also $CD45RA^+$ indicating that it could include naive CD8 T cells. It constitutes a total of 16 percent of the lymphocytes in this sample.

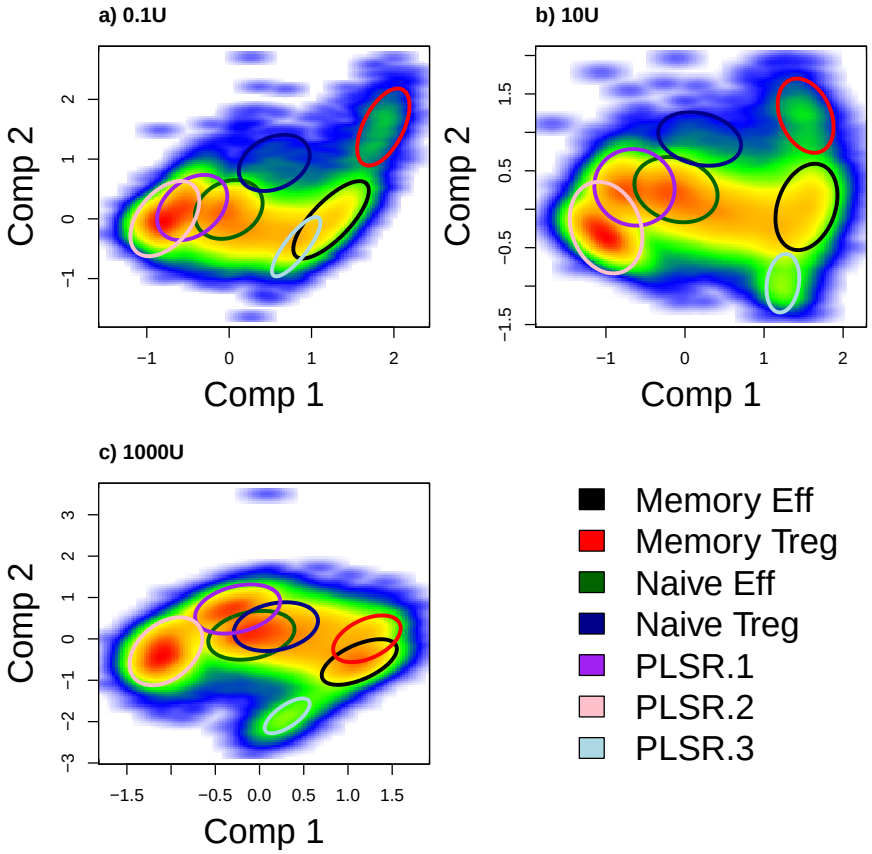


Figure 3.47. First two components of PLS projection. Clusters newly identified using PLS in relation to known manually gated ones within lymphocytes. The known manually gated cell populations are naive Teffs (dark green), naive Tregs (dark blue), memory Teffs (black) and memory Tregs (red). Three other clusters have been identified manually in light blue, pink and purple.

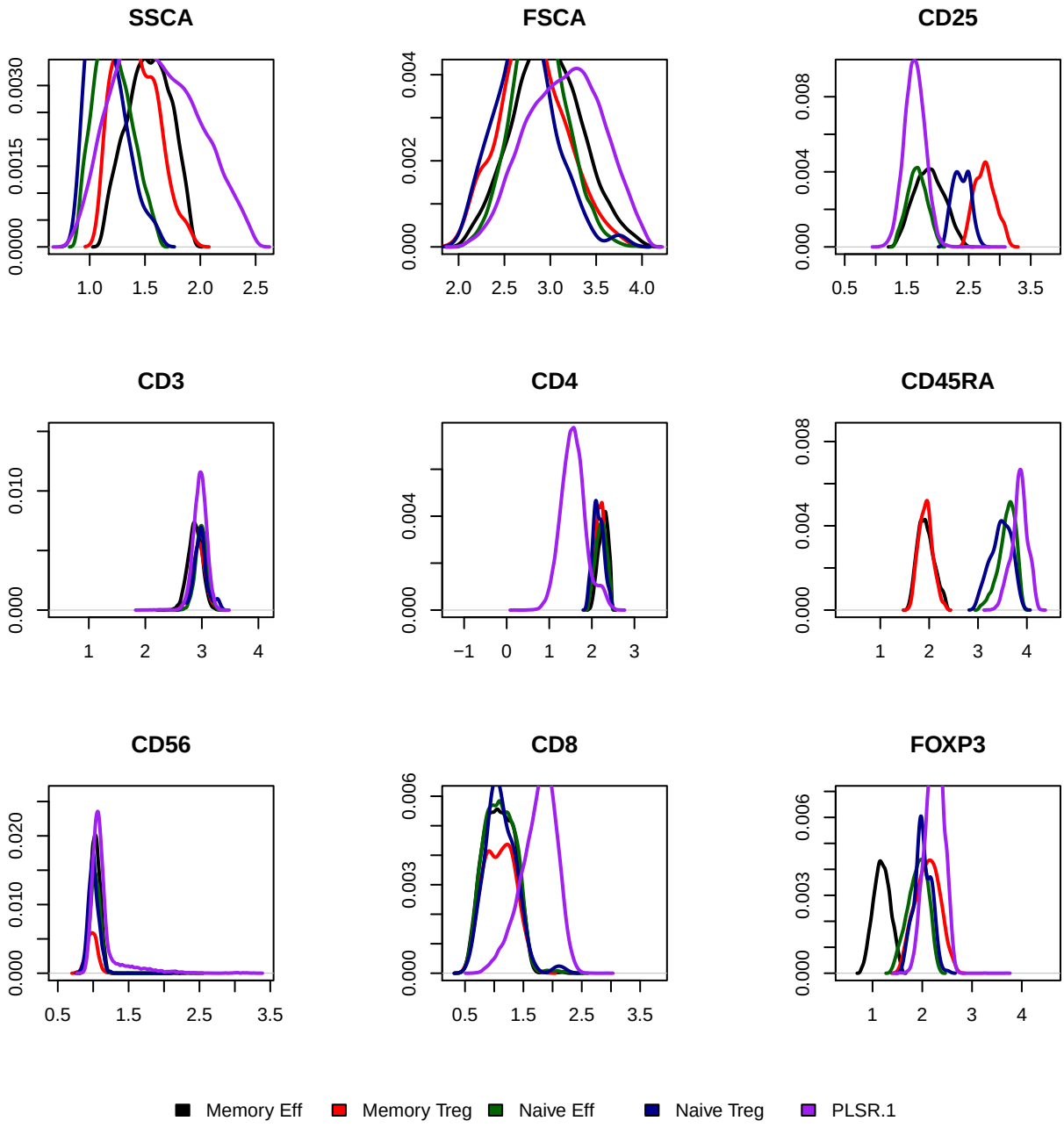


Figure 3.48. The newly identified cluster using PLS in Figure 3.47 (purple) in relation to known manually gated ones. I only included the cluster which shows response to proleukin. A distinctive property of the purple cluster is that it is high for CD8 and low for CD4. It constitutes approximately 16 percent of the lymphocyte cells in this sample.

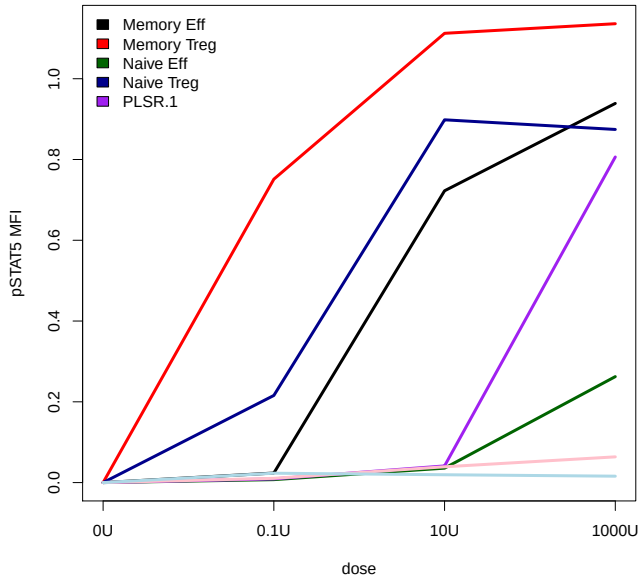


Figure 3.49. Dose response. The MFI of the pSTAT5 at each dose is shown, in the known cell subsets, memory Teffs, memory Tregs, naive Teffs and naive Tregs, as well as the newly identified cell subsets, light-blue, pink and purple (Figure 3.47). Of the newly identified cells subsets, only the purple subset shows signs of response. The pink and light-blue cluster do not respond even at the highest dose of 1000U.

Non-lymphocytes I repeated the same analysis on the non-lymphocytes, this time including side and forward scatter as predictors in PLS. From the first two components of the PLS projections, I visually identified three distinct subsets in purple, light-blue and pink, in addition to the known lymphocyte cluster in black, which were clearly discernible across all three stimulation doses. I further identified two less discernible cell populations, a subset of the lymphocytes (orange) and a low-density cluster most visible at 10U (yellow). As can be seen in Figure 3.52, of the newly identified cell subsets, only the orange and the yellow show response from 10U. While this is to be expected of the orange cluster as it constitutes a subset of the lymphocytes, the yellow subset when plotted along with the lymphocytes in Figure 3.51, has higher side and forward scatter which suggests it constitutes an artefact or possibly another type of dose-responsive T cell, since it is $CD3^{+}$. The yellow cell subset however makes up a very small proportion of the whole sample at less that 1 percent, compared to the lymphocytes at 16 percent.

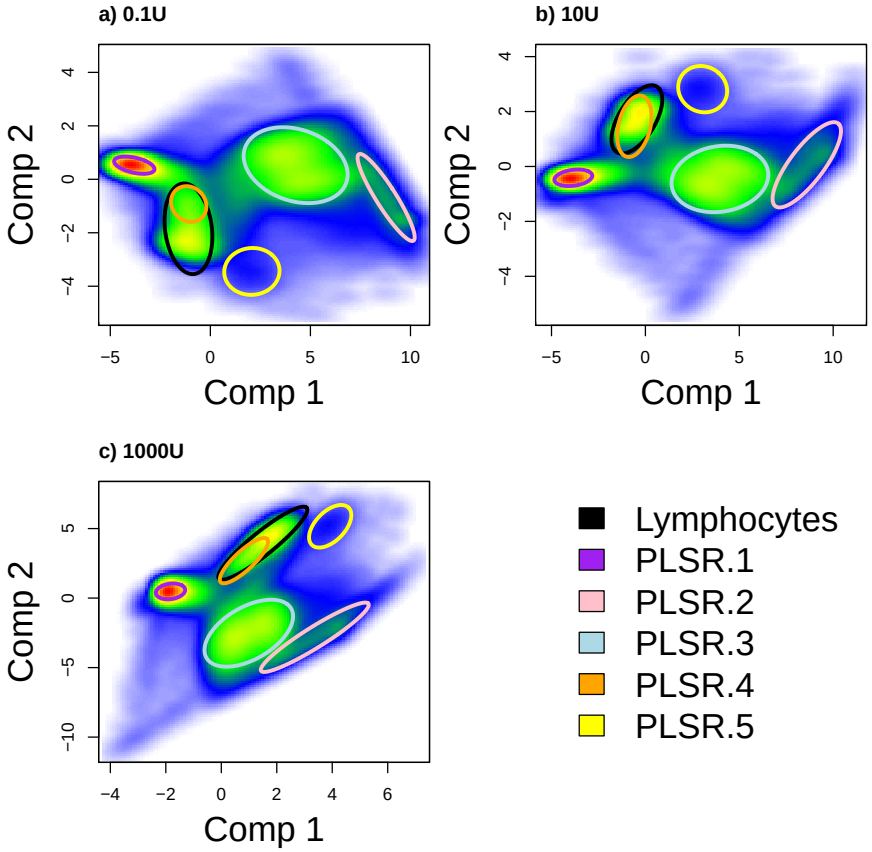


Figure 3.50. First two components of PLS projection. Clusters newly identified using PLS in relation to known manually gated ones within non-lymphocytes. In black, the previously manually identified lymphocytes, and

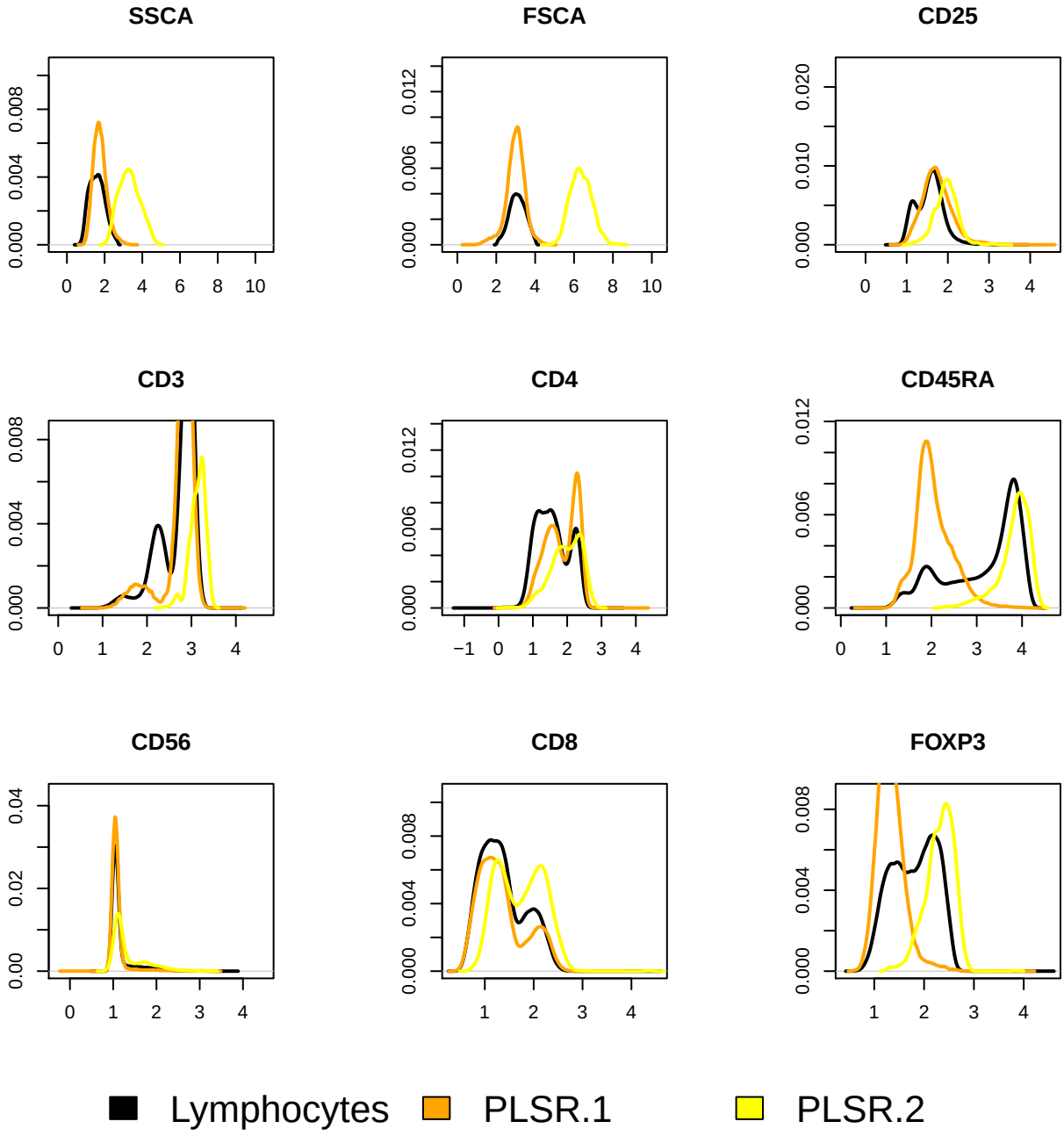


Figure 3.51. **First two components of PLS projection. Clusters newly identified using PLS in relation to known manually gated ones within non-lymphocytes.**

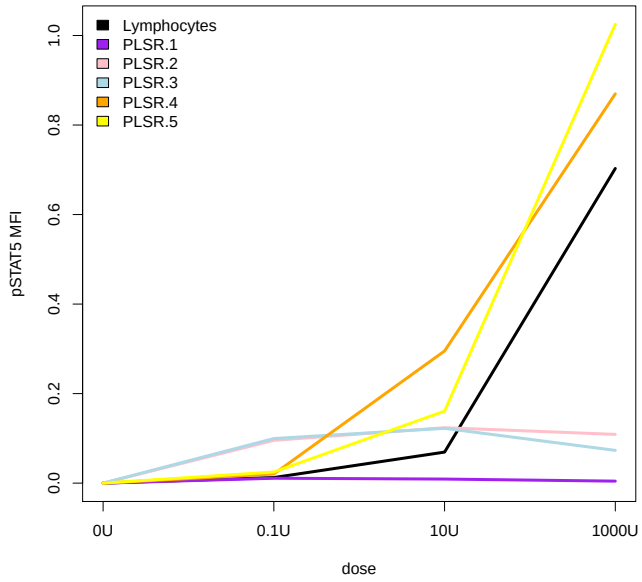


Figure 3.52. Dose response. Of the five newly identified cells (pink, light-blue, purple, yellow and orange) and lymphocytes (black), only the lymphocytes, yellow and orange subsets show response at 10U.

3.4.4 Identification of low-dose sensitive cells by recursively applying a two component mixture model on pSTAT5

The CART approach, described previously, seeks the core marker split point which minimises the deviance of the response variable. This approach successfully discriminates, based on side and forward scatter, the lymphocytes as the most responsive cluster when stimulated at the highest 1000 unit dose of proleukin. Unfortunately, it is not sufficiently sensitive to detect the small proportion of cells which are responsive to lower doses of proleukin.

In order to address this issue, I developed an approach based on the theory that by recursively splitting cells into responding and non-responding subsets, at decreasing doses of proleukin, it should be possible to identify cells which respond to the lowest proleukin dose. The algorithm, as illustrated in Figure 3.53, proceeds by first dividing cells as low responders and high responders on pSTAT5 response at 1000U by fitting a two-component GMM. The responder population (in green) is then further divided into low and high subsets by fitting the GMM on the pSTAT5 response at 10U. This process is then repeated in the pSTAT5 response stimulated at the lowest doses of 0.1 units. Cells which consistently appear in the high group are the most sensitive. This hierarchical approach draws some similarity to the recursive partitioning using CART except that the splitting decision depends only on applying a two-component GMM to the pSTAT5 distribution rather than selecting a core marker value on which to do the split. Also, at each step, the pSTAT5 at a lower dose is considered to discover cells which respond to the lowest dose of proleukin. The process is entirely driven by the bimodality of the pSTAT5 distribution. The clustering on the core markers is only applied right at the end to identify subsets of cells.

Lymphocytes I applied this algorithm within the lymphocyte subset. However, this time I kept the gated cells so as to improve the two-component GMM fit. The manually gated cells were only removed at the end once the subsets were identified.

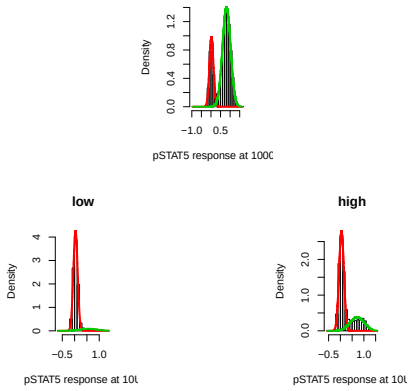


Figure 3.53. Recursive partitioning of pSTAT5 response into low (red) and high (green) populations to identify cells responsive to the lowest dose of proleukin. In the ungated sample the majority of cells are none responsive even at the highest dose. Cells are divided as low responders and high responders on pSTAT5 response (i.e baseline subtracted) at 1000U within responders further divide on low/high on pSTAT5 response at 10U repeat on pSTAT5 response at 0.1U cluster ones which are consistently high, these are the most sensitive cell populations

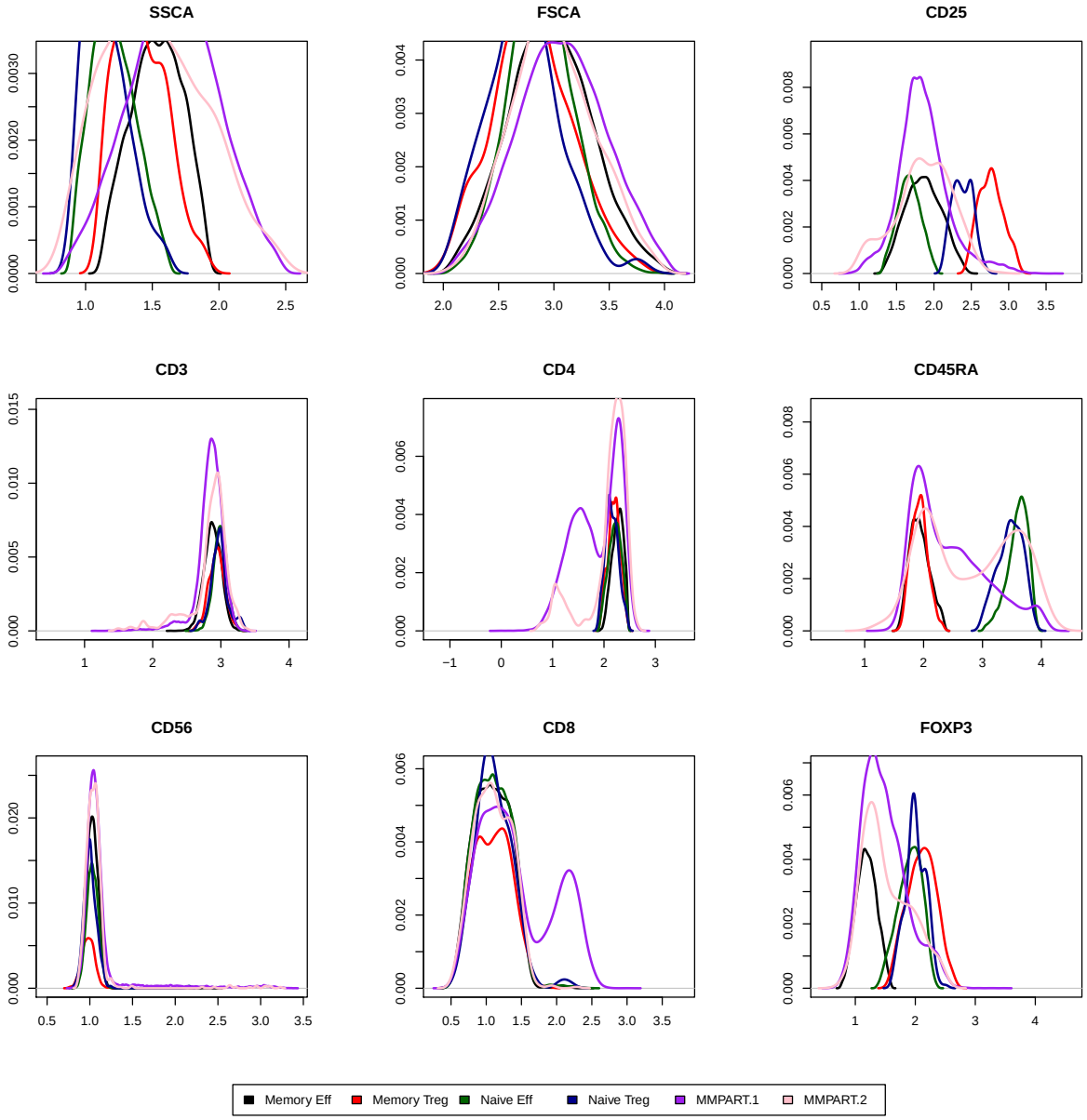


Figure 3.54. Clusters identified from recursive partitioning.

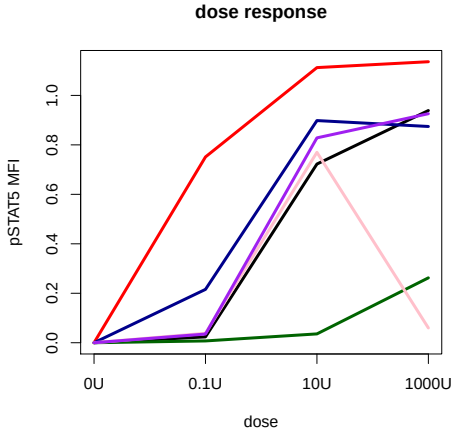


Figure 3.55. Clusters identified from recursive partitioning.

Non-lymphocytes I also applied the algorithm on the non-lymphocyte subset. Once again I left the lymphocytes in until the end to improve the model fit.

I find that, while many of the cells identified using this approach fall within the CD4 gate, certain highly-sensitive cells cluster in other subsets Figure 3.57. Excluding doublets on the basis of side scatter width, and examining the remainder on other channels these cells appear to be monocytes (from discussion with Marcin Pekalski and Tony Cutler), although additional markers would be required to better characterise these cells. Importantly, these cells would have been missed by manual gating since they are not lymphocytes. This approach could potentially be extended to identify cells which respond to low doses of proleukin.

Successive univariate clustering on response is not an obvious approach to multivariate data analysis but proves to be useful in identifying potentially interesting cells. One drawback of this approach is that since the scatter and core markers don't influence the gating, some cleaning of the reported cells is required to eliminate debris and doublets.

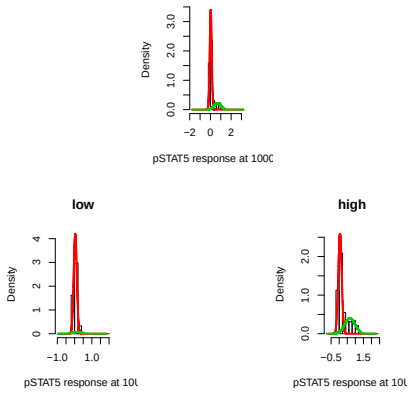


Figure 3.56. Recursive partitioning of pSTAT5 response into low (red) and high (green) populations to identify cells responsive to the lowest dose of proleukin.

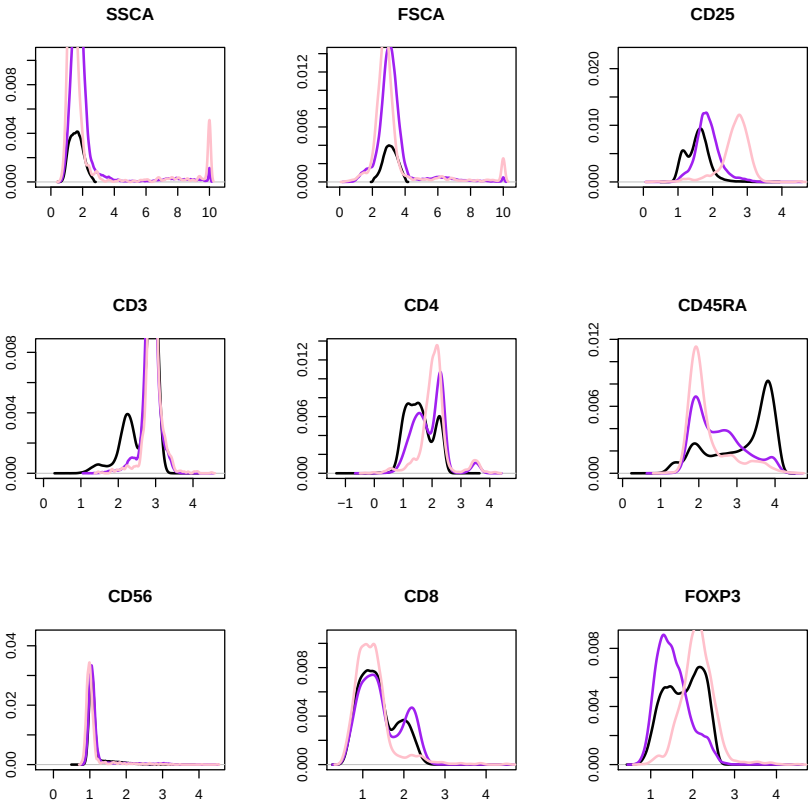


Figure 3.57. Clusters identified from recursive partitioning.

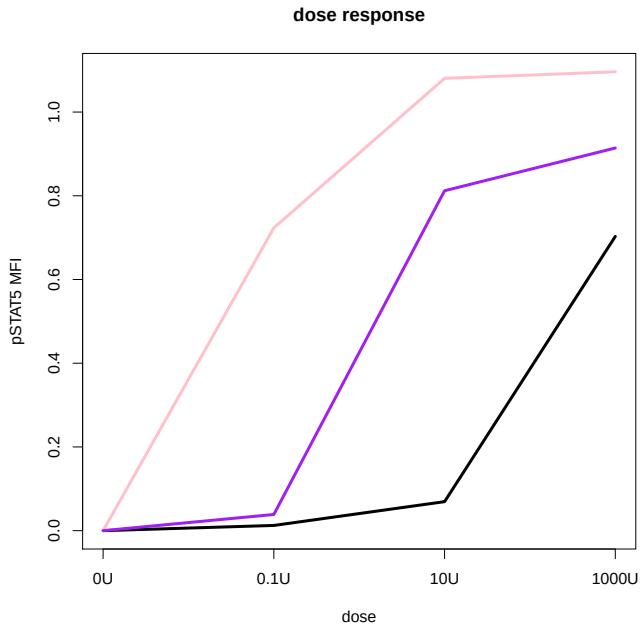


Figure 3.58. Dose response

3.5 Discussion

3.5.1 Association of pSTAT5 with T1D

Comparison with previous studies The Long et al study was in a relatively small number of individuals 66 cases and 125 controls and the reproducibility of the phenotypes was not assessed. Tony Cutler, on the other hand, found that the response of the fluorescence intensity of pSTAT5 to stimulation was poorly reproducible in the various cell subsets examined. I attempted to improve on the reproducibility of the fluorescence intensity by using single-cell background correction and peak normalisation. One reason why the pSTAT5 MFI is not reproducible in memory effector cells is because the pSTAT5 distribution is not unimodal. In naive and memory regulatory T cells, the pSTAT5 distribution is unimodal but the peak shift on stimulation is not reproducible. Tony Cutler claims that this is because the titration is too fine and noisy so that there is a lot of uncontrolled variation in the proleukin doses. This motivated counting the percent of cells whose pSTAT5 fluorescence is greater than the 99th percentile of the pSTAT5 distribution in the unstimulated sample. Since this phenotype was found to be slightly more reproducible, it was used to test the association with T1D in the four cell subsets, memory, naive, Teffs and Tregs.

Also of concern is that the Long et al dose is too high for the studied cell subsets. Our dose range is from 0, 0.1, 10 and 1000 U/ml, whereas theirs was 100 U/ml. We found that pSTAT5 Tregs are maximally stimulated by 10 U/ml and near maximum at 0.1 U/ml, which is in contrast to Long et al in which maximum stimulation was not achieved at 100 U/ml. On possible explanation is that the Long et al used frozen PBMC, while we used fresh blood. Dendrou et al (2009b) showed that the repeatability of a cell phenotype can be compromised with frozen samples and this difference could explain the difference in response in the two studies.

Normalisation of pSTAT5 Several normalisation methods were attempted to make the pSTAT5 dose-response phenotype more reproducible. However none were able to substantially improve the repeatability. I can only conclude that the noise in this dataset is substantial and not systematic, which makes normalisation very challenging. Unsuprisingly given the small dataset and the poor repeatability, no significant association was detected with dose-response and disease status. The conclusion is that this assay is not sufficiently reproducible for this sort of analyses.

3.5.2 Methods of identifying dose-responsive cell populations

I have attempted here a total of five different methods of identifying dose-responsive cell populations visually or semi-automatically:

- SPADE
- RPART
- CART
- PLSR
- MMPART

The first two methods, density normalisation with SPADE and recursive partitioning with RPART, used only the core markers in order to pool across samples. The other two methods, PLSR and CART, included the response variable pSTAT5 alongside with the core markers, and the final method, MMPART, used only information about the response variable.

I have also tried the two main approaches of combining data across samples by either pooling or by joining. Given the large number of events per sample in flow cytometry, pooling necessitates a way of reducing the number of events, and I have looked at two

ways of achieving this, first with density-dependent downsampling and then with binning using recursive partitioning. Both methods aim to achieve a uniform sampling of the core marker space, the first by thinning the data, the second by dividing the space into regions containing approximately the same number of points. The regions were then used instead of the points in downstream analysis. On the other hand, the joining did not aim to reduce the number of events but instead to normalise the number of events across samples. The joining was implemented using the nearest neighbour approximation in each sample to obtain a sample containing as many events as in the base sample.

From the pooled data, I was able to calculate multivariate dimensional scaled representations such the MST. Colour coding the MST by pSTAT5 response I was then able to identify new dose-responsive clusters of cells within lymphocytes and also within the general cell population which were ignored by the manual gating. One drawback I found with considering core markers only was that each cluster and especially clusters in regions of high density, may contain both dose-responsive and none responsive cells, which obfuscates the identification of specific cells types.

From the joined data, I explored methods which use the pSTAT5 response to guide the binary recursive partitioning of the core marker space using regression trees. Applying this approach to side and forward scatter, the lymphocyte population was consistently identified as the most responsive cell population in all samples stimulated at 1000 units. However, I found that applying this method on all markers across samples would yield different recursive partitioning schemes on different markers. An alternative to regression trees which avoids overfitting is random forests but needs reduction of the number of splits to be interpretable.

I found recursive partitioning to be a useful non-parametric method of exploring a dataset, in line with the tree like approach of studying flow cytometry datasets. Unfortunately these methods are very sensitive to even small changes in the data and suffer

from the biases of tree like data structures in which errors propagate.

The third approach I attempted was applied with the objective of identifying highly sensitive dose-responsive cells. These are cells which would respond at the lowest dose of proleukin. By recursively splitting the bimodal pSTAT5 distribution in a sample at decreasing doses of proleukin, the hope was to identify cells which respond to low doses of proleukin. Using this method certain a subset of highly responsive cells was identified which lie very close to the lymphocyte gate based on side and forward scatter. The disadvantage of this approach however is since the core markers do not feature in the identification of cell subsets, there is no guarantee that the identified dose-responsive cells represent a homogeneous subset. This is why it was necessary to pool across samples in order to

The methods described in this chapter can be applied to identifying dose-responsive cells in stimulation experiments. Although spade suggests the MST as a visualisation tool, it is not necessarily the most representative representation of the data since established cell types do not necessarily cluster in the branches. Furthermore it can be misleading since the layout is arbitrary and the branching is not always meaningful. SPADE contains an element of stochasticity in its downsampling so that running spade twice on the same data does not give the same tree. In my opinion, the true value of spade as applied to automatic gating or exploration of the dose-response in these datasets, lies in the downsampling and agglomerative clustering steps which allow for probing of the entire marker space. Furthermore, the raw data used to create the MST visualisation can been used to represent the data in a number of ways using established MDS methods which rely on the distance matrix computation.

Another advantage of the spade approach over the other two described methods is that it does not rely on joining samples on nearest-neighbour. Since pooling effectively averages the signal across samples it is less sensitive to debris. Nearest neighbour joining

can run into trouble if certain subsets are not present across all samples, as this can lead to mapping to the wrong subset or certain subsets not being represented. Nearest neighbour is advised if the total number of events varies greatly between samples. Also if the location of cell subsets differs significantly then this can also lead to joining to the wrong subsets. One way of assessing whether samples are significantly different is to use probability binning on in the context of spade to identify clusters whose proportion varies greatly between samples. Generally when pooling within batches this does not seem to occur, although Figure 3.33 illustrates how sensitive binning is to small changes in the underlying data distribution which often result from noise, caused by debris for example, rather than being biological significant. Also when the number of bins is increased it becomes much harder to detect statistically significant changes, since we lose degrees of freedom.

Using these methods in these PBMC samples, it is pretty conclusive that the strongest dose-response comes from within the lymphocyte subset. Although other cell types which carry IL-2 receptors are candidates, the majority of the response appears to lie within the lymphocyte subset.

Lymphocytes The markers and frequencies of these cell subsets identified with these methods are summarised in Table 3.3. Some cell overlapped with more than one method, I merged these together into a consensus cluster. I present in Figure 3.60, the cell populations which were identified by the different methods, on all core markers, and their response in Figure 3.59.

	FSCA	SSCA	CD25	CD3	CD4	CD45RA	CD56	CD8	FOXP3	freq
Memory Eff	2.96	1.54	1.86	2.88	2.27	1.92	1.03	1.08	1.19	5.03
Memory Treg	2.8	1.39	2.76	2.95	2.2	1.93	0.98	1.1	2.11	0.22
Naive Eff	2.87	1.2	1.67	2.99	2.22	3.62	1.03	1.09	1.92	9.76
Naive Treg	2.71	1.13	2.38	2.98	2.14	3.48	1	1.09	1.99	0.13
SPADE.1	3.41	1.88	1.59	2.85	1.51	3.19	1.69	1.92	1.73	1.14
RPART.1	3.35	1.95	1.16	2.33	1.1	3.94	1.82	1.29	2.31	3.36
RPART.2	3.02	1.71	2.04	2.92	2.31	2.04	1.03	1.09	1.27	4.05
PLSR.1	3.19	1.61	1.62	2.96	1.53	3.86	1.08	1.82	2.27	15.08
MMPART.1	3.01	1.58	1.89	2.88	2.18	2.77	1.04	1.09	1.39	0.32
MMPART.2	3.11	1.72	1.8	2.86	1.68	2.49	1.07	1.42	1.5	12.69
consensus	3.1	1.77	1.93	2.88	2.16	2.22	1.06	1.23	1.53	4.12

Table 3.3. Cell phenotypes in lymphocytes.

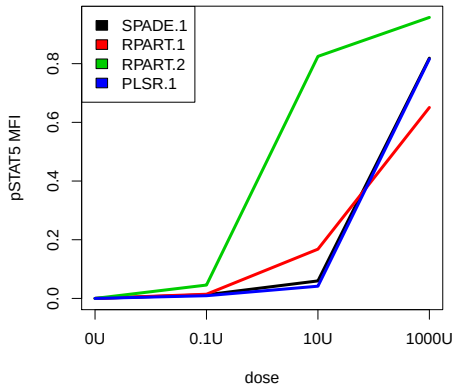


Figure 3.59. Dose response in lymphocytes.

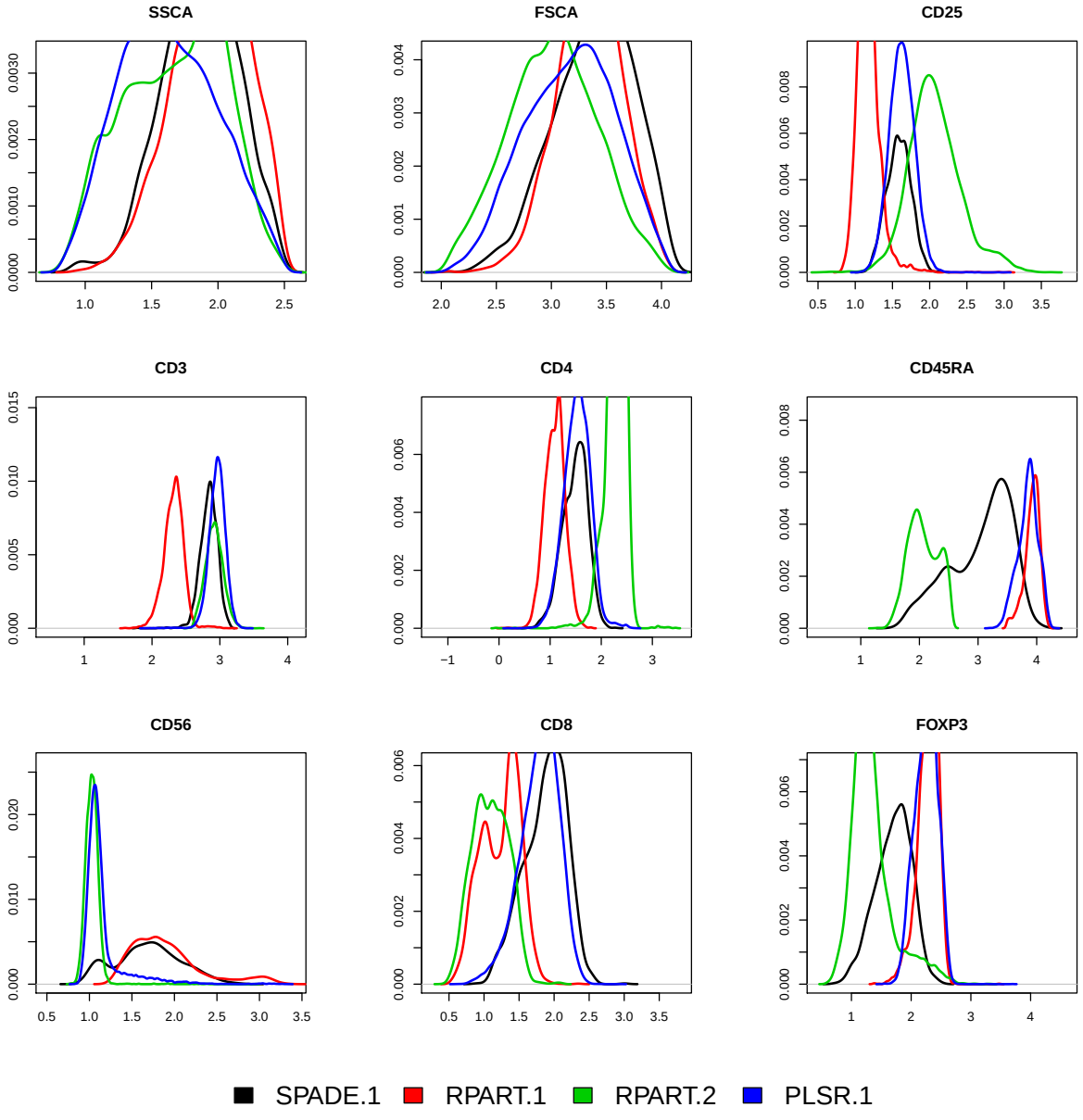


Figure 3.60. Identified cell subset in lymphocytes.

Non lymphocytes Outside of the lymphocytes, no strong response was visible in any of the clusters based on side and forward scatter. However a smaller cluster was consistently identified with larger side scatter than the lymphocytes. This was found to be a heterogeneous cluster containing approximately x percent of the events in the sample.

	FSCA	SSCA	CD25	CD3	CD4	CD45RA	CD56	CD8	FOXP3	freq
Lymphocytes	3.09	1.59	1.58	2.84	1.53	3.46	1.07	1.29	1.83	16.97
SPADE.1	6.1	3.25	1.9	3.1	1.97	3.79	1.15	1.53	2.23	0.21
RPART.1	3.24	2.36	1.53	2.77	1.57	3.33	1.1	1.35	1.86	1.66
PLSR.1	6.33	3.29	1.97	3.16	2.01	3.91	1.17	1.75	2.35	0.14
MMPART.1	4.62	6.61	2.4	3.1	2.35	2.74	1.15	1.51	2.23	0.03
MMPART.2	4.07	2.74	1.84	3	2.21	2.76	1.13	1.47	1.7	0.57
consensus	4.13	2.47	1.87	2.97	1.93	3.08	1.12	1.43	1.86	0.5

Table 3.4. Cell phenotypes, non-lymphocytes.

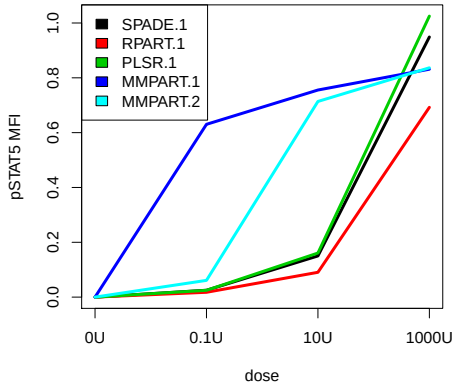


Figure 3.61. Dose response in non-lymphocytes.

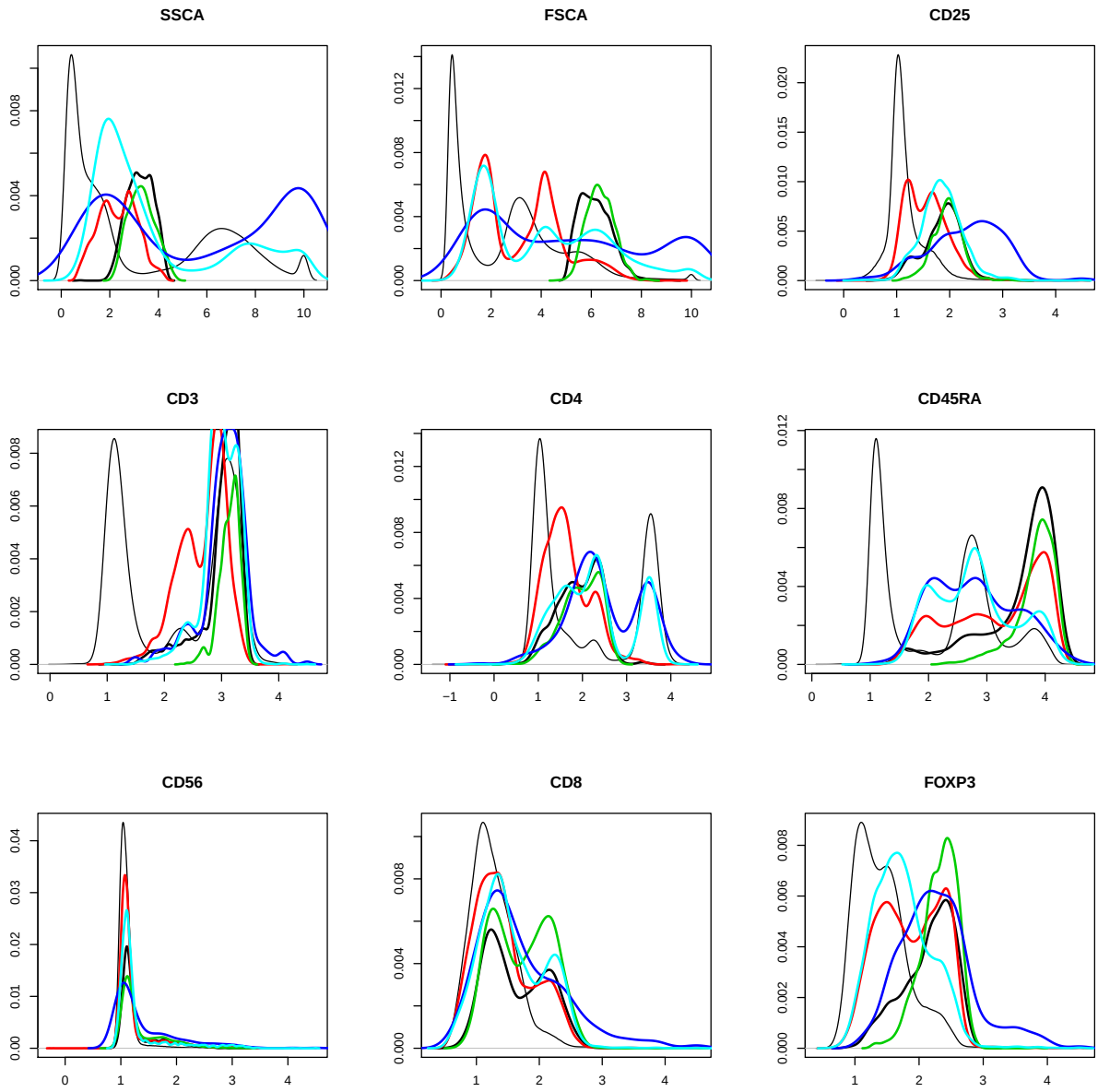


Figure 3.62. Identified cell subset in non-lymphocytes.

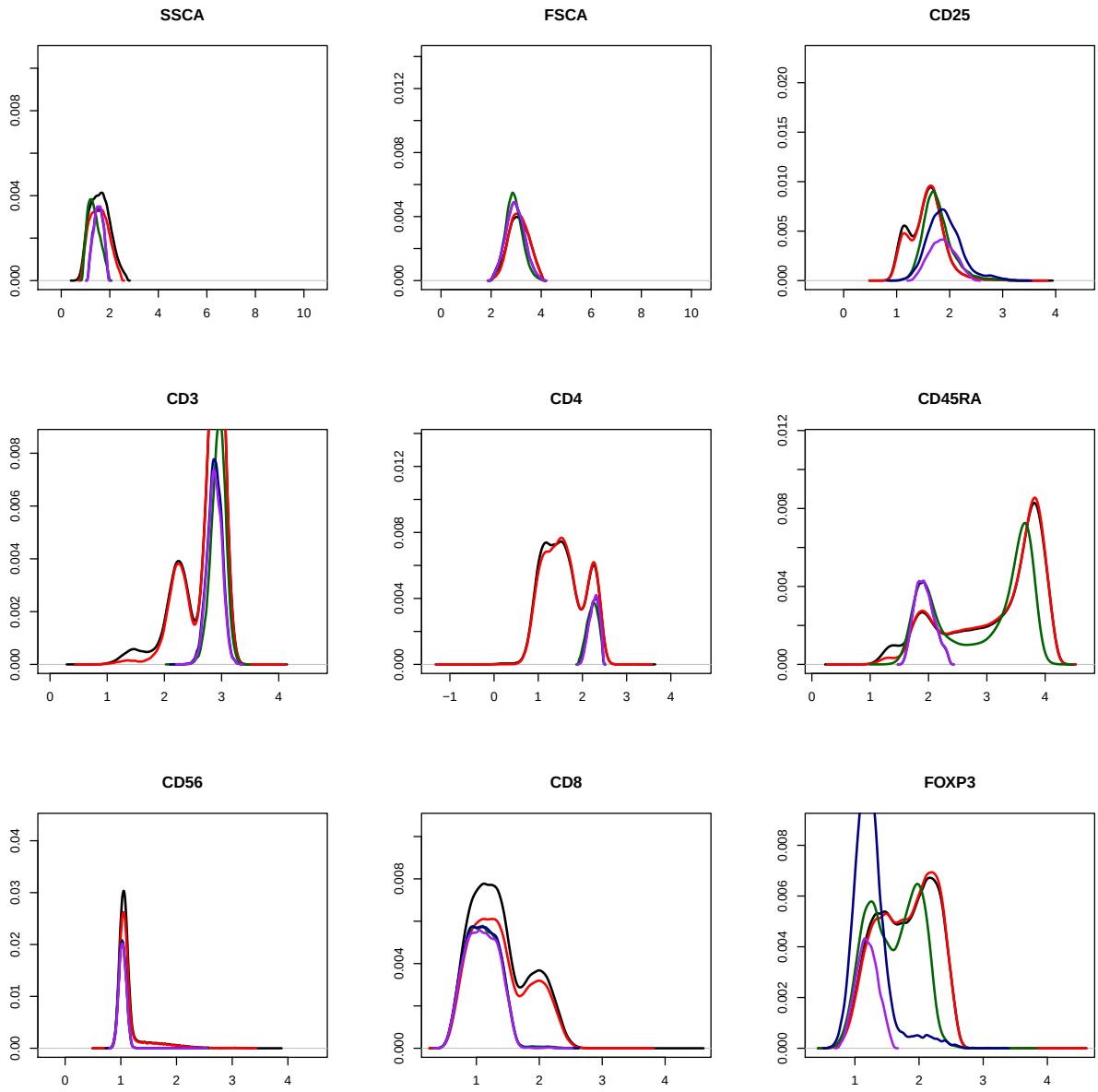


Figure 3.63. Identified cell subset consensus in non-lymphocytes.

Correlation with CD25, CD122, CD132 At lower doses, 0.1 and 10 units, of proleukin, CD25 explains a significant proportion of the pSTAT5 response, however at 1000 units the response is saturated in cells with high CD25 and so the other trimeric receptors, CD122 and CD132 become more important.

Spillover causes spurious correlation of CD45RA and FOXP3 The fluorescent dye used to tag CD45RA, PE Cy7 YG, has a significant spillover into the channel used to measure PE YG. This explains the spurious correlation between CD45RA positive and FOXP3 positive cells making it impossible to discriminate the $CD45RA^+$ subsets, naive Teffs from naive Tregs, based on FOXP3.

Chapter 4

KIR3DL1/KIR3DS1 copy number variation in type 1 diabetes

In Chapter 2 and Chapter 3, I have identified and analysed cell populations in flow cytometry using normalisation and clustering methods. Of the clustering methods I have applied, mixture-model clustering proved to be particularly useful in dealing with noise. As discussed, one of the benefits of mixture-model approach are the posterior probabilities which can be used in downstream statistical analysis. So far, I have not made full use of this feature in the association tests, partly because the number of cells is large and the fraction of cells which lie clearly within one cluster is an adequate measure for association testing. However, as one deals with smaller datasets, the uncertainty in clustering can have an important impact on the association statistics. In this chapter, I will apply normalisation and mixture model clustering to a much smaller genetic dataset, and account for the clustering uncertainty in testing association disease.

4.1 Background

4.1.1 Killer immunoglobulin receptors and their interaction with human leukocyte antigen molecules

The KIR region, a 150 kb cluster of 17 identified genes located within the 1 Mb Leukocyte Receptor Complex on chr19q13.4, is an interesting candidate region in HLA-associated autoimmune diseases such as T1D, due to the interaction between KIR and HLA molecules. KIRs are transmembrane glycoproteins, expressed by NK cells and subsets of T cells, which bind to the peptide presenting HLA class I molecules on the surface of target cells.

It is thought that the interaction of these two loci plays an important part in immunity and, as a result, these regions have co-evolved (Parham and Moffett, 2013), leading to much diversity in the allelic frequency of HLA and KIR genes between populations. However, while the polymorphism of the HLA region is primarily due to allelic diversity, the alleles of the KIR region also vary in copy number. The copy number variation of these genes is thought to correlate with the level of expression of KIRs and to bear some influence on disease outcome. KIRs are named according to their number of extracellular immunoglobulin domains (2D and 3D) and whether their cytoplasmic tail is short (S) or long (L). Generally, KIR proteins with the long cytoplasmic domain transduce inhibitory signals upon ligand binding via an immune tyrosine-based inhibitory motif (ITIM), whereas KIRs with the short cytoplasmic domain, do not contain the ITIM motif and instead transduce an activating signal upon ligand binding. The fate of the target cell then depends on the composite signal generated by the combination of inhibiting/activating KIRs in the presence of their HLA class 1 ligands (Bashirova et al, 2006). The longer KIR genes tend to have greater allelic diversity, whereas the shorter KIRs tend to vary more in copy number. The polymorphic and highly homologous na-

ture of these genes leads to very extensive haplotype and copy number diversity in the KIR region (Jiang et al, 2012).

Despite the important biological function of KIRs, no GWAS hits have been reported in the KIR region. This could well be due to the shortcomings of GWAS in detecting trait-associated sequence polymorphism in more complex, poorly mapped regions of the genome. The technology primarily used in GWAS is the SNP array. SNP arrays assay the polymorphism in single nucleotides positions across the genome by the means of SNP probes of typically 20 base pairs in length. Depending on the region of the genome and the array used, the SNP probe coverage varies greatly. In certain regions, the SNP probe coverage is insufficient to capture the underlying genetic complexity. Also, SNP probes are template based, they are designed based on reference sequences. They are not designed to discover new sequences, only the distribution of known alleles. Consequently, SNP probes targeting regions which are more polymorphic than anticipated, such as regions of allelic specific copy number variation like KIR, may lead to signals which cannot be clustered into the expected three genotypes (e.g AA, AT, TT) of bi-allelic SNPs. Instead the signal returned by these probes can return a variable number of clusters which requires more careful analysis using flexible genotype calling algorithms (Kumasaka et al, 2011). Additionally, KIR is poorly mapped in the human reference genome (build36/hg18) and does not contain all KIR genes. Thus, KIR has been mostly overlooked by GWAS, which makes it worthy of further investigation and characterisation.

4.1.2 *KIR3DL1* and *KIR3DS1*: two strong candidates for T1D association

Two genes in the KIR complex, *KIR3DL1* and *KIR3DS1*, are particularly interesting candidates for T1D association due to their interaction with T1D-associated HLA class

I molecules. The KIR3DL1 protein is known to interact with the HLA class I allotypes that contain the HLA-Bw4 serological epitope (Gumperz et al, 1997; Vivian et al, 2011), whereas the protein encoded by *KIR3DS1*, which shares 97 % sequence similarity to *KIR3DL1*, is thought to bind the more restrictive HLA-Bw4-80I epitope subset (Martin et al, 2007).

The grouping of HLA-A and HLA-B alleles according to HLA-Bw4 serological epitope (Martin et al, 2002) is given in Table 4.1 and includes several HLA class I alleles that are associated with T1D risk after conditioning on the major HLA class II effects (Nejentsev et al, 2007; Howson et al, 2009).

Copy number variation in the *KIR3DS1* gene is thought to be implicated in viral diseases, such as HIV-1 (Martin et al, 2002; Pelak et al, 2011), and certain autoimmune diseases, but there is no substantial evidence of association with T1D (Körner and Altfeld, 2012). However, studies to date have been small, and evidence for its association has not yet been addressed in large, well powered studies.

Epitope		Residues (77-83)	HLA-B	HLA-A
HLA-Bw4	80I	NLR**I**ALR	B*1516 B*1517 B*1524 B*2702 B*3801 B*4901 B*5101 B*5108 B*5201 B*5301 B*5302 B*5701 B*5702 B*5801	A*2301 A*2402 A*2403 A*2407 A*2501 A*3201
HLA-Bw4	80T	DLR**T**LLR SLR**T**LLR NLR**T**ALR	B*1302 B*2701 B*2704 B*2705 B*3701 B*3802 B*4402 B*4403 B*4404 B*4405 B*4414 B*4417 B*4429 B*4435 B*4701	
HLA-Bw6		SLR**N**LRG	B*702 B*703 B*705 B*706 B*708 B*710 B*716 B*726 B*801 B*1401 B*1402 B*1501 B*1503 B*1504 B*1505 B*1507 B*1508 B*1509 B*1510 B*1514 B*1515 B*1518 B*1539 B*1801 B*3501 B*3502 B*3503 B*3508 B*3901 B*3906 B*3928 B*4001 B*4002 B*4006 B*4011 B*4023 B*4101 B*4102 B*4202 B*4501 B*4601 B*4801 B*5001 B*5002 B*5501 B*5601	

Table 4.1. Grouping of HLA alleles by HLA-Bw4 epitope. *HLA-A* and *HLA-B* alleles which carry the serological epitope HLA-Bw4 can be further subdivided as HLA-Bw4-80I or HLA-Bw4-80T, depending on whether the amino acid at position 80 in the heavy alpha chain of the HLA class I protein is an isoleucine (I) or a threonine (T) (Gumperz et al, 1997; Martin et al, 2002).

4.2 Samples and genotyping assays

4.2.1 Samples

Our study involved 12,106 individuals: 6,744 cases (age at diagnosis less than 17 years) from the Genetic Resource Investigating Diabetes (GRID) cohort, and 5,362 controls from the British 1958 Birth Cohort (1958BC). All subjects were of white European ancestry (as confirmed by PCA of earlier GWAS data in these samples (Barrett et al, 2009)) with written informed consent and Ethics Committee/Institutional Review Board approval. The DNA for the cases and controls was prepared using the same protocols in Cambridge and in Bristol respectively, and all samples were cell-line derived.

4.2.2 HLA and SNP Genotyping

HLA Epitope	Cases	Controls	Total
N/A	3822 (11)	2681 (70)	6503 (81)
HLA-Bw6	1175 (308)	753 (199)	1928 (507)
HLA-Bw4-80T	651 (162)	754 (174)	1405 (336)
HLA-Bw4-80I	1096 (266)	1174 (284)	2270 (550)
HLA total	2922 (736)	2681 (657)	5603 (1393)

Table 4.2. Classification of subjects in study by HLA epitope (as defined in Table 4.1). In parentheses, number of subjects analysed with qPCR post QC. No HLA typing was done for the N/A category. The HLA epitopes are defined in Table 4.1. An individual is assigned to an HLA epitope group if he is a carrier of at least one allele of that group. So that each individual only belongs to a single HLA epitope group, the assignment priority is first HLA-Bw4-80I, then HLA-Bw4-80T and finally HLA-Bw6 allele if no HLA-Bw4 alleles were found.

HLA genotypes were available on a subset of 5,603 individuals, 2,922 cases and 2,681 controls. HLA-A and HLA-B genes were typed at four-digit allele resolution using Dynal RELI SSO assays (Invitrogen, Paisley, U.K.) (Table 4.2). The epitope classification of HLA-A and HLA-B alleles is given in Table 4.1.

All 12,106 samples were genotyped using the ImmunoChip SNP array, according to the manufacturer's protocol, and processed at the University of Virginia in Charlottesville, USA. ImmunoChip is a custom Illumina 200K Infinium high-density SNP array (Nikula et al, 2005), which contains 100 SNPs in the LILR complex, 30 of which fall in the 14 kb *KIR3DL1* region (Table 4.4).

4.2.3 qPCR experimental protocol

Jiang et al (2012) have designed qPCR assays to study copy number variations in KIR, which have led to the discovery of many rare haplotypes. In collaboration with Jiang et al (2012), Deborah Smyth developed multiplexed qPCR 384-well assays, designed to determine copy numbers in most known alleles of *KIR3DL1* and *KIR3DS1*. The gene *STAT6*, known to always be present in two copies, was used as a reference. The forward/reverse primers and probe sequences for *KIR3DL1*, *KIR3DS1* and *STAT6* are summarised in Table 4.3.

Nonetheless, qPCR assays remain expensive (£12 per sample) and labour intensive compared to SNP arrays, and thus qPCR was only performed on a subset of 1629 samples, 816 cases and 813 controls by Deborah Smyth.

The qPCR plaform used was the LightCycler 480 Real-Time PCR Instrument. For each qPCR reaction, 2 µl of DNA at 5 ng µl^{-1} were used with 5 µl of Quantifast Multiplex PCR mastermix (0.25 µl primer mix, 0.045 µl probe mix and 4.705 µl of water). qPCR conditions were 95 °C for 5 min, followed by 40 cycles at 95 °C for 15 s and 66 °C for 50 s. Data was collected at 66 °C. The samples were tagged with three different dyes, Fam for *KIR3DS1*, Cy5 for *KIR3DL1* and DFO for *STAT6*, and amplified on eighteen 384-well plates. On all plates, samples were replicated across four wells. So that each plate contained a maximum of 96 samples. Four calibrator samples of known *KIR3DL1*/*KIR3DS1* copy number and one water sample were included on all but one

plate. Cases and controls were distributed evenly across all plates. Four plates were analysed in duplicate.

Gene	Oligos	Sequence (5'-3')
KIR3DS1	Forward Primer	CATCGGTTCCATGATGCG
	Reverse Primer	GGGAGCTGACAACTGATAGG
	Probe	AACAGAACCGTAGCATCTGTAGGTCCCT
KIR3DL1	Forward Primer	CACAGTTGGATCACTGCGT
	Reverse Primer	CCGTGTACAAGATGGTATCTGTA
	Probe	CCCTTCTCAGAGGCCCAAGACAC
STAT6	Forward Primer	CCAGATGCCTACCATGGTG
	Reverse Primer	CCATCTGCACAGACCACTCC
	Probe	CTGATTCCTCCATGAGCATGCAGCTT

Table 4.3. The qPCR probes and primers. The qPCR probes and primers used in our assay, these were originally designed by Jiang et al (2012).

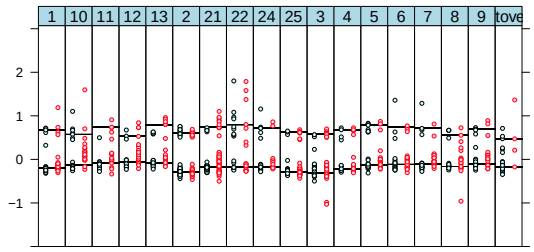
4.3 Data Analysis

4.3.1 Quality control and normalisation of qPCR data

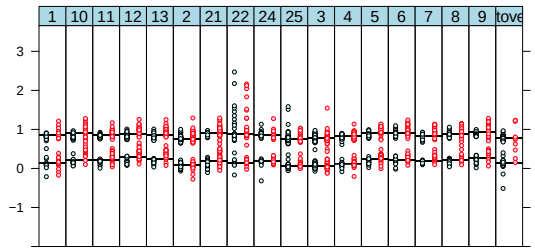
The experiment files exported from the LightCycler gave us the crossingpoint (Ct) value for each dye-DNA conjugate. By subtracting from the Ct value of the reference dye-DNA conjugate, DFO-STAT6, I obtained the baseline relative ΔCt value for Fam-KIR3DL1 and Cy5-KIR3DS1. Since *STAT6* is known to have two copies, negative values of ΔCt should indicate two copies or less, and positive values, two copies or more. However, due to qPCR differences in efficiency this threshold does not necessarily hold in practice as shown in Figure 4.1, which is why it is more correct to cluster when calling copy number. As part of the quality control (QC), I excluded 64 samples that did not yield a DFO-STAT6 Ct reading in all four well replicates. All remaining samples were summarised by the ΔCt median of the four well replicates.

The individual distributions of *KIR3DS1* and *KIR3DL1* ΔCt differed between plates

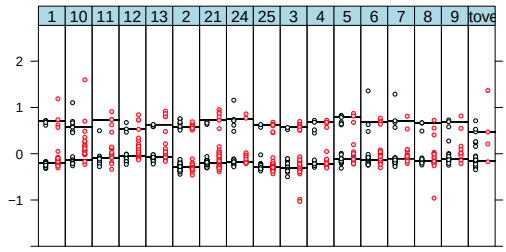
(Figure 4.1.a.b) which prevented clustering all samples together. Visual inspection of the data distributions by plate led us to drop plate 22 because it appeared excessively noisy (Figure 4.1.a.b). To normalise the ΔCt values across the remaining plates, I first applied the k-medoids algorithm within plates for *KIR3DL1* and *KIR3DS1* separately to identify the location of the most distinguishable copy number groups, one and two copies, then normalised across plates by a linear transformation so that the median ΔCt of the two groups were aligned across all seventeen plates. Samples repeated across different plates were summarised by the median of their repeated value. Following QC, 1474 unique individuals, 747 cases and 727 controls, were available for analysis.



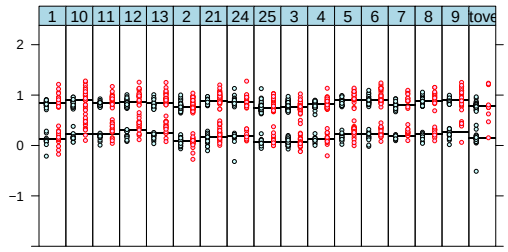
(a) Pre-QC *KIR3DS1* ΔCt



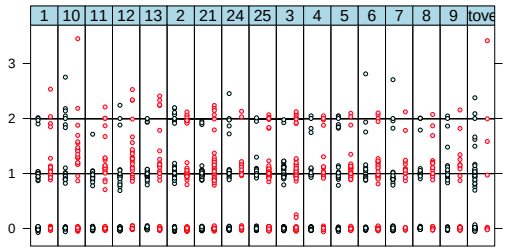
(b) Pre-QC *KIR3DL1* ΔCt



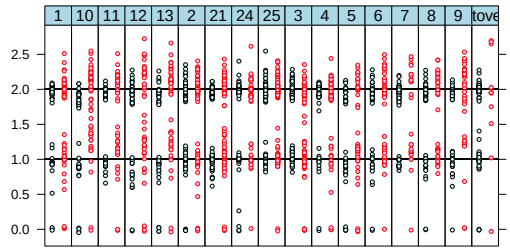
(c) Post-QC *KIR3DS1* ΔCt



(d) Post-QC *KIR3DL1* ΔCt



(e) Peak normalised *KIR3DS1* ΔCt



(f) Peak normalised *KIR3DL1* ΔCt

Figure 4.1. *KIR3DS1* and *KIR3DL1* ΔCt values for cases (red) and controls (blue) per qPCR plate. Plate 22 stands out as the noisiest for both *KIR3DL1* (a) and *KIR3DS1* (b), and so is subsequently dropped as part of the QC (c and d). Negative ΔCt are not displayed for pre and post QC so as to better visualise the one and two copy number groups. Normalisation consists a linear transform which maps the medians of the one and two copy groups from each plate to 1 and 2 (e and f). After normalisation, negative ΔCts values are assigned to zero.

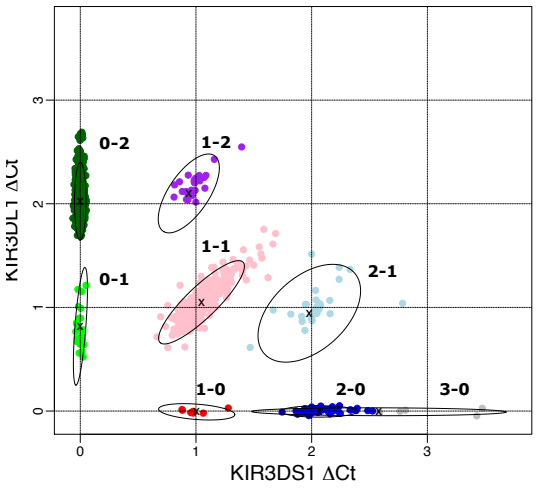
4.3.2 Bivariate clustering: copy number calling in qPCR data

Samples which yielded one or less Ct reading for Fam-KIR3DL1 or Cy5-KIR3DS1, but all four Ct readings for the reference DFO-STAT6, were assumed to contain zero copies

of *KIR3DL1* or *KIR3DS1*. For the remainder of the samples, I called copy number groups by fitting a mixture of bivariate Gaussian distributions to the two dimensional normalised ΔCt values, allowing for eight *KIR3DS1*/*KIR3DL1* copy number groups: three common groups of two copy numbers (0-2, 1-1, 2-0) and five rarer groups of lower or higher copy numbers (**??**). The mixture was fitted using an EM algorithm (Young et al, 2009b) with initial parameters calculated from the clusters returned by k-means with centers set to the eight expected locations of the copy number groups. After fitting the mixture model each sample was assigned a posterior probability of belonging to each of the eight copy number groups which allows for uncertainty in copy number calling. These posterior probabilities were taken into account in downstream statistical analysis via multiple imputation.

Raw median ΔCt distributions varied across plates which prevented simple visual copy number assignment (Figure 4.1). After normalisation, samples repeated across different plates showed good reproducibility (Figure 4.4) and two dimensional clustering enabled 1474 samples to be confidently assigned to a single copy number group, including all samples with known copy number which were assigned to the correct cluster.

Jointly clustering on *KIR3DL1* and *KIR3DS1*, has the advantage of exploiting the correlation between the ΔCt values. For example, this can be seen in plate 10, where noisy cases (Figure 4.1.f) are difficult to assign as one or two copies based solely on their *KIR3DL1* ΔCt, but are much more clearly distinguishable when I also consider their *KIR3DS1* ΔCt value (Figure 4.3).



Copy Number	count (percentage)		
KIR3DS1-KIR3DL1	cases	controls	total
0–2	444 (59.44)	446 (61.35)	890 (60.38)
1–1	229 (30.66)	207 (28.47)	436 (29.58)
2–0	26 (3.48)	28 (3.85)	54 (3.66)
2–1	15 (2.01)	16 (2.20)	31 (2.10)
1–2	13 (1.74)	14 (1.93)	27 (1.83)
0–1	13 (1.74)	11 (1.51)	24 (1.63)
1–0	4 (0.54)	3 (0.41)	7 (0.47)
3–0	3 (0.40)	2 (0.28)	5 (0.34)
KIR3DL1	cases	controls	total
2	457 (61.18)	460 (63.27)	917 (62.21)
1	257 (34.40)	234 (32.19)	491 (33.31)
0	33 (4.42)	33 (4.54)	66 (4.48)
KIR3DS1	cases	controls	total
0	457 (61.18)	457 (62.86)	914 (62.01)
1	246 (32.93)	224 (30.81)	470 (31.89)
2	41 (5.49)	44 (6.05)	85 (5.77)
3	3 (0.40)	2 (0.28)	5 (0.34)
	747 (100)	727 (100)	1,474 (100)

Figure 4.2. Copy number calling of *KIR3DL1/KIR3DS1* from qPCR ΔCt. On the left, the median normalised ΔCt values for *KIR3DS1* and *KIR3DL1* are shown with the results of clustering into the eight copy number groups coloured according to the group with the highest posterior probability. The three most common copy number groups are the ones with a total copy number of two: *KIR3DL1* 0-2 (dark green), *KIR3DL1/KIR3DS1* 1-1 (pink) and *KIR3DS1* 2-0 (dark blue). The ellipses delimit the 95^{th} percentile. On the right, the counts of the most probable copy number state are shown for cases and controls.

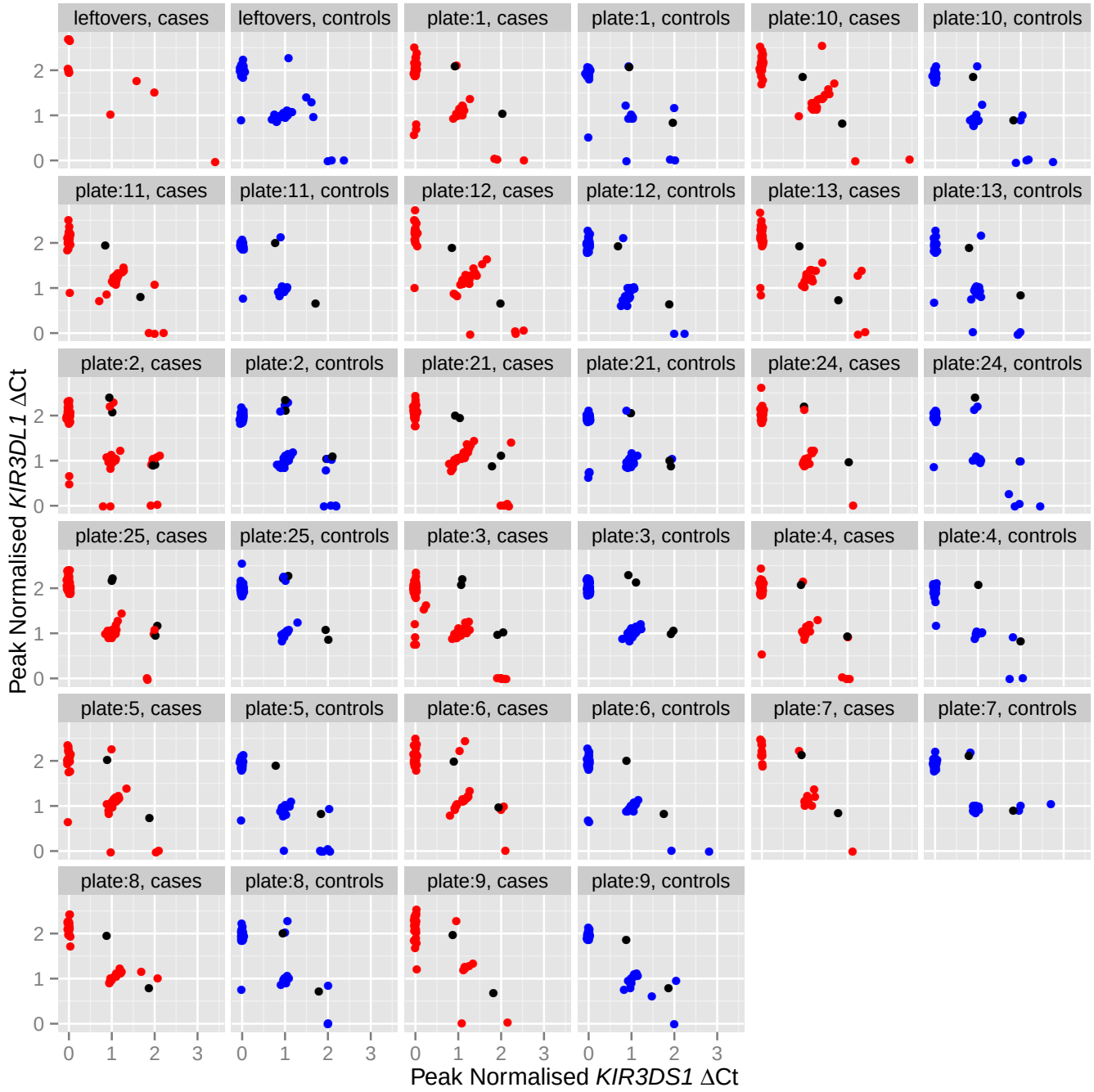
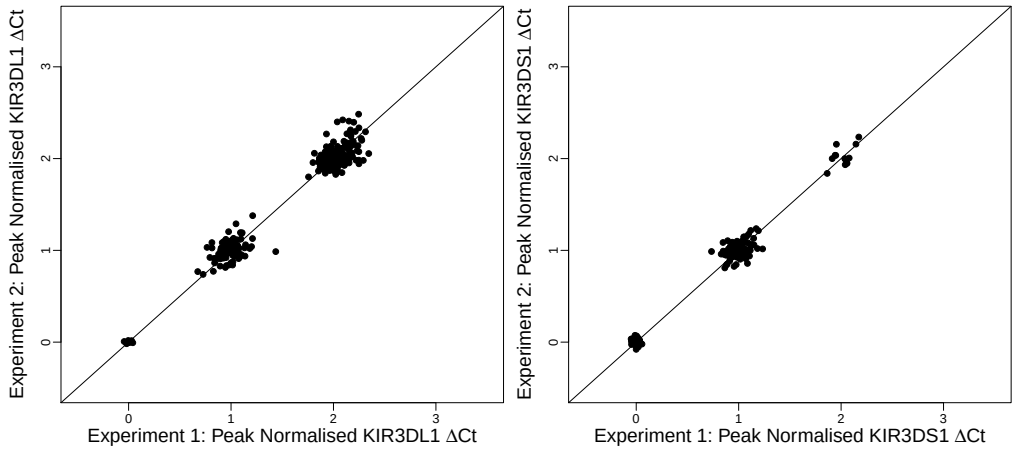


Figure 4.3. Post-QC cases (red) and controls (blue) are plotted separately for each qPCR plate. The samples with known *KIR3DL1*/*KIR3DS1* copy number are plotted in black. We can see that there is a larger spread in cases than in controls which is especially clear in the 1-1 copy number group. Also, it is apparent that the ΔCt of *KIR3DL1* and *KIR3DS1* are correlated in the 1-1, 2-1 and 1-2 groups. We exploited this correlation in the copy number calling by doing bivariate clustering.



(a) Repeatability of *KIR3DL1* ΔCt post normalisation and QC ($r^2 = 0.961$).

(b) Repeatability of *KIR3DS1* ΔCt post normalisation and QC ($r^2 = 0.99$).

Figure 4.4. Repeatability of qPCR assay. In order to assess the reliability of the qPCR assay 310 samples were re-analysed. We found very high reproducibility of the ΔCt values ($r^2 > 0.96$) confirming the reliability of our qPCR assay. r^2 is the Pearson correlation squared.

4.3.3 KNN classification: copy number imputation into the SNP data

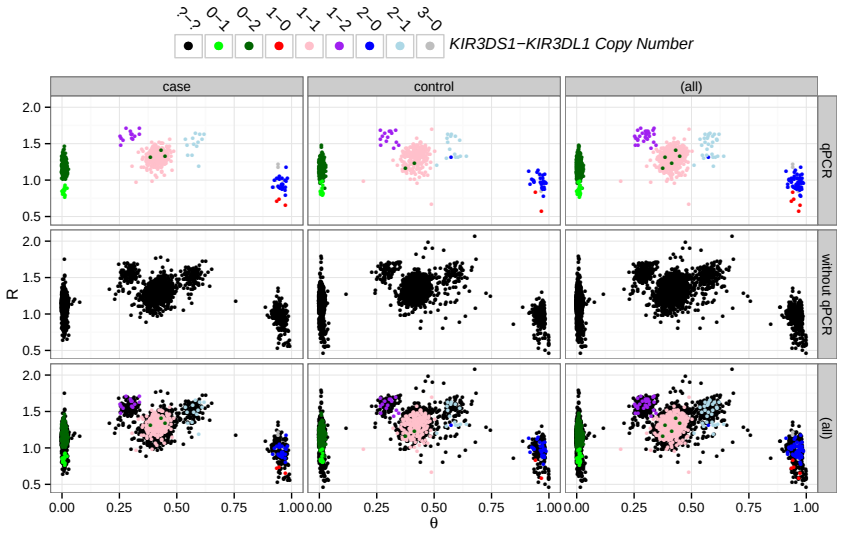


Figure 4.5. Overlay of ImmunoChip and qPCR samples for R and θ at SNP rs592645. Samples are coloured by the most likely *KIR3DS1-KIR3DL1* copy number group according to the qPCR analysis (see Figure 4.2). The first and second row split the samples on the availability of qPCR data, and the third row is the overlay of the samples from the first and second row. The first and second column split the samples by case-control status and the third column is the overlay of the samples from the first and second column.

We extended our sample size by using the subset of samples common between the qPCR and SNP datasets, 747 cases and 727 controls, to train a k-nearest neighbour (knn) classifier to predict *KIR3DL1*/*KIR3DS1* copy number using the R and θ signals from ImmunoChip SNPs.

Each of 30 SNPs lying within the *KIR3DL1* (since *KIR3DS1* is not on the reference genome) region were assessed for association with either *KIR3DL1* or *KIR3DS1* copy number in individual linear regression of copy number against R and θ (Table 4.4). SNP signals, R and θ, showed good association with copy numbers of *KIR3DL1* and of *KIR3DS1* for 19 of 30 SNPs in the *KIR3DL1* region (Table 4.4). The best example

is shown in Figure 4.5, in which seven clusters for SNP rs592645 can be discerned that correspond closely with qPCR derived *KIR3DL1/KIR3DS1* copy numbers. This is also visually apparent in Figure 4.6 where SNP rs592645 shows the best clustering of copy number out of those 30 SNPs. Figure 4.5 also illustrates a number of important points about using SNP signals for imputation. First, θ corresponds to the ratio of copies of *KIR3DL1* to *KIR3DS1*, while R corresponds to the total copy number. Second, some clusters overlap; without the qPCR data, the number of clusters and their boundaries would be difficult to define, particularly along the R axis. Finally, the clusters are in slightly different positions in cases and controls, reflecting the known sensitivity of genotyping chips to subtle differences in DNA preparation and storage conditions. This has two implications: probabilistic clustering of the SNP data alone is likely to be poor in the combined sample, while unsupervised clustering of cases and controls separately when clusters are not clearly separated risks increasing type 1 error rates (Plagnol et al, 2007). Instead, I used the qPCR copy numbers as training data to perform supervised clustering of the SNP signals.

We first explored the validity of our imputation approach by means of leave-one-out cross-validation (LOOCV) in the samples with qPCR data. We examined using all nineteen predictive SNPs, or various subsets, and found optimal knn imputation was achieved with the single most predictive SNP, rs592645 with $k = 8$, which minimised the mean LOOCV error rate to 2.0 % across ten multiply imputed qPCR datasets (Figure 4.7).

We also explored the effect of varying the size of the training data set by setting KIR gene copy numbers to missing for a randomly chosen subset of samples and imputing them in the remaining samples. We suggest that only 295 samples are required to achieve LOOCV error rates $< 5\,\%$ and 590 for error rates $< 2.5\,\%$ (Figure 4.8).

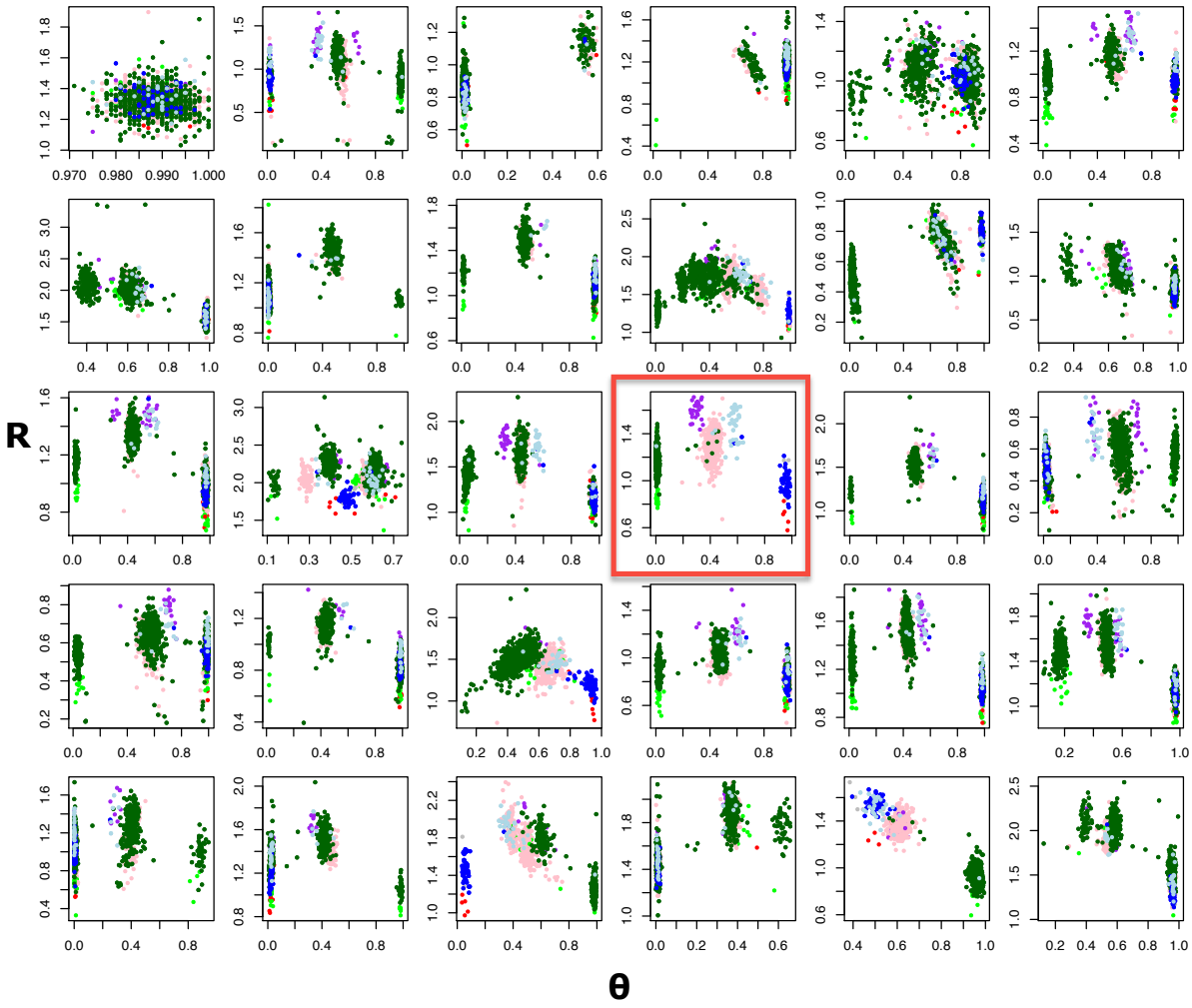


Figure 4.6. Signal plots of ImmunoChip SNPs which fall in the *KIR3DL1* region. Each of the 30 ImmunoChip SNPs from Table 4.4, coloured by *KIR3DL1*/*KIR3DS1* copy number (see Figure 4.5 for legend). SNP rs592645 (red square) shows the best clustering by copy number.

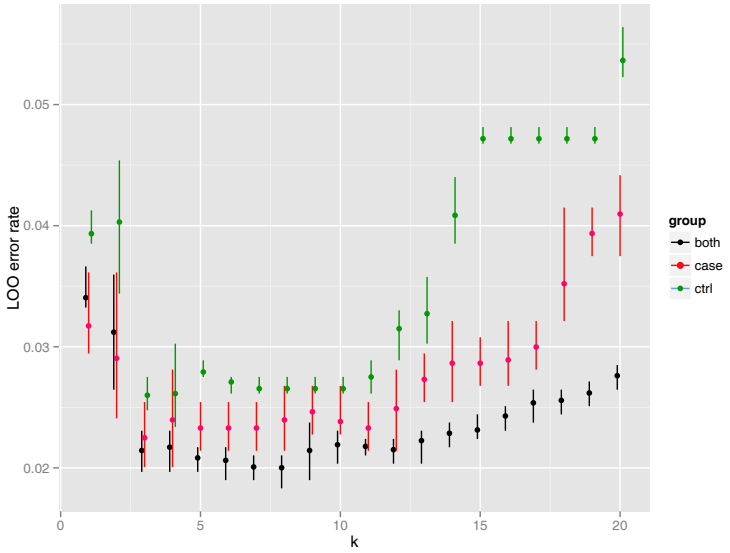


Figure 4.7. Leave-one-out crossvalidation error rate for k-nearest neighbour (KNN) prediction. Leave-one-out cross validation error rates obtained from KNN prediction of *KIR3DL1/KIR3DS1* copy numbers from the R and θ signals of SNP rs592645. Each point shows the proportion of samples for which the knn predicted copy number did not match the qPCR call, averaged over ten multiply imputed qPCR call datasets (using the posterior probabilities from **??**). Error bars show the minimum and maximum error rates over the ten multiply imputed datasets. Knn was run in parallel for cases only, controls only and on all samples together. The minimum error rate is achieved for $k = 8$ when the prediction uses both cases and controls.

Name	Position	SNP	GenCall QC	p-value θ	p-value R
seq-rs597598	60007252	[A/G]	ok	3.19E-03	7.81E-01
seq-rs598452	60007428	[A/G]	ok	6.53E-01	2.62E-01
seq-t1d-19-60007809-C-G	60007809	[G/C]	ok	3.64E-02	6.27E-06
seq-rs55761930	60008141	[T/C]	ok	6.33E-01	6.12E-01
seq-rs10500318	60012591	[A/G]	ok	7.59E-11	1.31E-13
seq-rs592645	60012739	[A/T]	ok	8.85E-01	3.38E-09
seq-rs604077	60013208	[A/G]	ok	4.82E-03	1.20E-01
seq-rs604999	60013409	[A/G]	ok	1.77E-15	9.99E-04
seq-t1d-19-60014013-A-C	60014013	[T/G]	lowcallrate	8.74E-01	3.15E-08
rs3865507	60014188	[T/G]	ok	8.62E-03	6.93E-17
seq-rs3865510	60016051	[A/C]	ok	2.23E-10	2.04E-10
seq-rs648689	60016286	[A/G]	ok	2.31E-01	1.03E-02
seq-rs649216	60016447	[T/C]	ok	2.85E-02	1.04E-13
rs581623	60018551	[A/G]	ok	3.76E-02	2.06E-13
seq-rs4806568	60022568	[A/G]	lowcallrate	1.44E-20	2.93E-01
seq-rs674268	60024002	[T/C]	lowcallrate	1.43E-02	2.90E-01
rs12461010	60024413	[A/G]	ok	4.72E-01	1.72E-01
seq-rs2295805	60028513	[T/C]	lowcallrate	9.55E-08	8.40E-04
seq-rs12976350	60030391	[T/C]	lowcallrate	1.70E-05	5.07E-01
seq-t1d-19-60034052-C-T	60034052	[A/G]	hwe	3.27E-02	4.07E-01
rs4806585	60038236	[T/G]	hwe	2.20E-11	2.42E-02
seq-rs62122181	60039178	[T/C]	lowcallrate	2.40E-13	2.26E-01
rs10422740	60052298	[T/C]	monomorph	7.78E-01	8.49E-02
rs640345	60054671	[A/G]	ok	3.61E-07	6.83E-02
seq-t1d-19-60054973-T-C	60054973	[A/G]	ok	2.92E-01	2.28E-04
seq-t1d-19-60056605-A-T	60056605	[A/T]	ok	3.99E-01	1.48E-16
seq-t1d-19-60056721-C-T	60056721	[A/G]	ok	9.02E-01	2.04E-09
seq-rs10407958	60063974	[T/A]	ok	1.06E-02	5.45E-10
seq-rs1654644	60065174	[T/G]	ok	7.94E-14	5.21E-12
rs3826878	60069023	[A/G]	ok	2.63E-05	3.55E-06

Table 4.4. ImmunoChip SNPs which fall in *KIR3DL1*. The 30 ImmunoChip SNPs which fall in the *KIR3DL1* region according to build36/hg18, nineteen of which are significantly associated with *KIR3DL1/KIR3DS1* copy number (highlighted in blue). *KIR3DS1* is missing from build36/h18. SNP rs592645 which has shown to be highly predictive of *KIR3DL1/KIR3DS1* copy number is highlighted in light red. The QC column reports the GenCall quality control diagnosis: ok, low call rate, failure to meet Hardy Weinberg equilibrium (hwe) or monomorphic SNP.

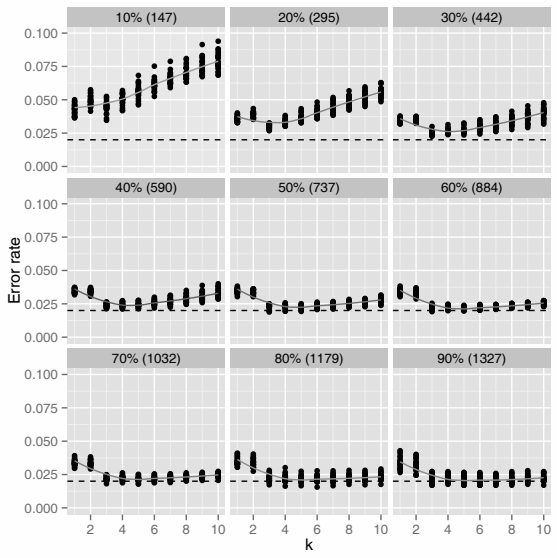


Figure 4.8. Error rate of k-nearest neighbour prediction from R and θ of SNP rs592645 in random subset of samples. Each panel shows the LOOCV error rates of *KIR3DL1/KIR3DS1* copy number prediction from R and θ of rs592645 in the remaining unlabeled samples when using a different size subset of the training data. The percentage of the complete training data set and the size of the subset is given in the title of each panel. Each point represents the LOOCV error rate averaged over ten multiply imputed qPCR call datasets (using the posterior probabilities from **??**). Smoothing lines show the average over 25 independent random subsets of training data. The black dashed line represent the observed error rate in the complete sample. As the size of the training dataset increases the error rate becomes less sensitive to the choice of the parameter k. Only 295 samples are required to achieve LOOCV error rates $< 5\%$ and 590 for error rates $< 2.5\%$.

4.4 Association testing of *KIR3DL1/KIR3DS1* copy number with T1D

In calling copy number states from qPCR data, rare amplifications of three or more copies are harder to classify with certainty than more common copy number states such as zero, one, or two copies. This is mainly because copy numbers higher than two are rare but also because the ΔCt difference between successive higher copy numbers shrinks logarithmically. However dropping samples which can not be classified with certainty can lead to bias. A less biased approach is to allow for uncertainty by using posterior probabilities in the association tests (Plagnol et al, 2007).

We tested for association of T1D with the predicted copy numbers from the qPCR and SNP datasets using logistic regression. We allowed for uncertainty in the copy number call when estimating individual odds ratios by using ten multiply imputed datasets (Cordell, 2006), and averaging results over those using the `R` package `mitools` (Lumley, 2012).

We allowed for statistical interaction with HLA-Bw4 by repeating the association test in the subsets of carriers of the target ligand HLA-Bw4 epitopes, HLA-Bw4 for *KIR3DL1* and the putative ligand HLA-Bw4-80I for *KIR3DS1*. We directly tested for interaction with a more powerful case-only χ^2 test (Yang et al, 1999; Cordell, 2009).

Finally, I attempted to correlate *KIR3DL1/KIR3DS1* copy number with T1D status. We performed an ANOVA χ^2_7 test on the logistic model using the most likely copy number group, which yielded a p-value of 0.9776 in the qPCR and 0.1739 in the SNP dataset, thus showing no overall association. We also used multiple imputation to assess the effect of individual copy number groups while allowing for the uncertainty in copy number calling. We found no significant evidence for association, in the qPCR data (747 cases and 727 controls) nor in the extended SNP data (6744 cases and 5362 controls)

(Table 4.5). By expanding to these large samples, which would be infeasible to genotype directly with qPCR, I am able to exclude odds ratios outside of the range [.92; 1.08] for the common copy number groups with 95 % certainty.

We also repeated the association tests in the subset of individuals carriers of the HLA-Bw4 epitope, and again detected no significant association (Table 4.6). A disadvantage of subsetting by HLA-Bw4 is that power is lost by greatly reducing the sample size. A more powerful test for interaction between unlinked genes is a case-only test (Yang et al, 1999). If there were an interaction effect between *KIR3DL1/KIR3DS1* and HLA-Bw4 then this should be detectable as a difference in *KIR3DL1/KIR3DS1* copy number frequencies across HLA-Bw4 strata in the cases. However, I found no significant difference in either the qPCR or SNP data sets, before or after reducing the degrees of freedom by collapsing the KIR copy number to present/absent to increase power (Table 4.7).

a)

	qPCR					SNP				
KIR3DS1-KIR3DL1	case:control	total	OR	95%CI	p-value	case:control	total	OR	95%CI	p-value
0-2	444:446	890	1.00			4094:3222	7316	1		
1-1	229:207	436	1.11	0.88-1.40	0.3673	2050:1628	3678	0.99	0.92-1.07	0.8349
2-0	26:28	54	0.92	0.52-1.61	0.7713	229:225	454	0.79	0.65-0.96	0.0193
2-1	15:16	31	0.94	0.46-1.93	0.8695	121:101	222	0.92	0.7-1.2	0.5246
1-2	13:14	27	0.93	0.43-2.01	0.8587	98:74	172	1.04	0.77-1.42	0.7822
0-1	13:11	24	1.19	0.53-2.68	0.6794	116:77	193	1.19	0.89-1.59	0.2535
1-0	4:3	7	1.34	0.30-6.02	0.7031	25:21	46	0.94	0.52-1.68	0.8255
3-0	3:2	5	1.52	0.27-8.62	0.6369	11:14	25	0.74	0.3-1.82	0.518
Overall	747:727	1474			0.9842	6744:5362	12106			0.3552

b)

	qPCR					SNP				
KIR3DL1	case:control	total	OR	95%CI	p-value	case:control	total	OR	95%CI	p-value
2	457:460	917	1.00			4192:3296	7488	1		
1	257:234	491	1.11	0.89-1.38	0.3702	2287:1806	4093	0.99	0.92-1.07	0.8883
0	33:33	66	1.01	0.61-1.66	0.9795	265:260	525	0.8	0.67-0.96	0.0151
Overall	747:727	1474			0.6651	6744:5362	12106			0.0506

c)

	qPCR					SNP				
KIR3DS1	case:control	total	OR	95%CI	p-value	case:control	total	OR	95%CI	p-value
0	457:457	914	1.00			4210:3299	7509	1		
1	246:224	470	1.10	0.88-1.37	0.4096	2173:1723	3896	0.99	0.91-1.07	0.7785
2	41:44	85	0.94	0.60-1.47	0.7787	350:326	676	0.83	0.71-0.97	0.0212
3	3:2	5	1.24	0.21-7.28	0.8084	11:14	25	0.74	0.3-1.82	0.5119
Overall	747:727	1474			0.8044	6744:5362	12106			0.1494

Table 4.5. Association test of *KIR3DS1-KIR3DL1* copy number with T1D. Association with T1D tested for the joint *KIR3DS1-KIR3DL1* **(a)**, marginal *KIR3DL1* **(b)** and *KIR3DS1* **(c)** copy number group. No evidence of a significant, joint or marginal, effect detected in the qPCR dataset, 747 cases and 727 controls, nor in the SNP dataset, 6744 cases and 5362 controls. Case-control counts shown are derived from the most likely copy number assignment across the 10 multiply imputed qPCR and SNP datasets. Statistical inference for association is derived from the multiply imputed datasets using the `R` package mitools (Lumley, 2012). The last row of each table contains the pooled p-value for each association test using the `R` package mice (van Buuren and Groothuis-Oudshoorn, 2011).

a) HLA-Bw4 subset			qPCR					SNP		
KIR3DS1-KIR3DL1	case:control	total	OR	95%CI	p-value	case:control	total	OR	95%CI	p-value
0-2	259:286	545	1.00			1025:1156	2181	1.00		
1-1	123:128	251	1.06	0.79-1.43	0.6976	556:583	1139	1.08	0.93-1.24	0.3194
2-0	16:15	31	1.22	0.58-2.57	0.5985	61:87	148	0.79	0.56-1.11	0.1733
2-1	7:13	20	0.59	0.23-1.51	0.2754	32:40	72	0.90	0.56-1.45	0.6695
1-2	8:8	16	1.10	0.41-2.98	0.8450	27:32	59	0.95	0.57-1.60	0.8513
0-1	10:7	17	1.58	0.59-4.20	0.3621	36:26	62	1.56	0.94-2.60	0.0876
1-0	2:1	3	2.21	0.20-24.50	0.5187	7:3	10	2.63	0.68-10.19	0.1614
3-0	3:0	3				3:1	4	3.38	0.35-32.51	0.2910
	428:458	886				1747:1928	3675			

b) HLA-Bw4 subset			qPCR					SNP		
KIR3DL1	case:control	total	OR	95%CI	p-value	case:control	total	OR	95%CI	p-value
2	267:294	561	1.00			1052:1188	2240	1.00		
1	140:148	288	1.04	0.78-1.38	0.7787	624:649	1273	1.09	0.95-1.25	0.2414
0	21:16	37	1.45	0.74-2.83	0.2822	71:91	162	0.88	0.64-1.21	0.4399
	428:458	886				1747:1928	3675			

c) HLA-Bw4-80I subset			qPCR					SNP		
KIR3DS1	case:control	total	OR	95%CI	p-value	case:control	total	OR	95%CI	p-value
0	159:187	346	1.00			650:734	1384	1.00		
1	93:83	176	1.32	0.92-1.90	0.1370	384:365	749	1.19	0.99-1.42	0.0578
2	12:14	26	1.01	0.45-2.24	0.9842	61:75	136	0.92	0.64-1.31	0.6376
3	2:0	2				1:0	1			
	266:284	550				1096:1174	2270			

Table 4.6. Association test of *KIR3DS1-KIR3DL1* copy number with T1D, conditional on HLA-Bw4. In order to test whether *KIR3DL1/KIR3DS1* is associated with T1D risk conditional on the presence of the respective the HLA-Bw4 epitope, association with T1D is tested in the subset of individuals carriers of an HLA-Bw4 epitope for the joint *KIR3DS1/KIR3DL1* **(a)** and marginal *KIR3DL1* **(b)** copy number groups and, also tested in the subset of individuals carriers of HLA-Bw4-80I for the marginal *KIR3DS1* **(c)** copy number group.

a)

		qPCR		SNP	
		HLA-Bw4−	HLA-Bw4+	HLA-Bw4−	HLA-Bw4+
KIR3DS1−	KIR3DL1+	183	269	739	1063
KIR3DS1+	KIR3DL1−	12	21	40	71
KIR3DS1+	KIR3DL1+	113	138	396	613
		p-value = 0.4094		p-value = 0.4235	

b)

	qPCR		SNP	
	HLA-Bw4−	HLA-Bw4+	HLA-Bw4−	HLA-Bw4+
KIR3DL1−	12	21	40	71
KIR3DL1+	296	407	1135	1676
	p-value = 0.5144		p-value = 0.3609	

c)

	qPCR		SNP	
	HLA-Bw4-80I−	HLA-Bw4-80I+	HLA-Bw4-80I−	HLA-Bw4-80I+
KIR3DS1−	293	159	1153	649
KIR3DS1+	159	107	673	447
	p-value = 0.4922		p-value = 0.0353	

Table 4.7. Case-only χ^2 test for interaction between *KIR3DS1-KIR3DL1* copy number and HLA-Bw4, across the ten multiply imputed qPCR and SNP datasets. Counts in each contingency table are derived from the most likely copy number assignment across the multiply imputed datasets. To reduce the degrees of freedom and improve power, I summarise copy numbers higher or equal to one by presence (+) and zero by absence (−). The pooled p-value of each χ^2 test, across the multiply imputed datasets, is given in the last row of each contingency table. We find no significant association with HLA-Bw4, within cases, in either the joint **(a)** or the marginal **(b)(c)** *KIR3DS1-KIR3DL1* distributions.

4.5 Discussion

4.5.1 Previous association studies of KIR genes with T1D

So far, case-control studies using PCR in different ethnicities have looked at whether the presence or absence of KIR genes but not the copy number are associated with T1D (van der Slik et al, 2003, 2007; Nikitina-Zake et al, 2004; Santin et al, 2006; Middleton et al, 2006; Park et al, 2006; Mogami et al, 2007; Shastry et al, 2008; Jobim et al, 2010; Zhi et al, 2011). These, however, represent an incomplete version of the KIR genotype because, as shown by Jiang et al (2012), a considerable portion of the diversity in the KIR haplotypes arises from copy number variation. Although presence/absence might have a stronger effect than copy number variation.

From the studies I know of, as summarised in Tables 4.8 and 4.9, only two have reported individual KIR genes to be associated with T1D independently of HLA. First Nikitina-Zake et al (2004), reported that *KIR2DS2*/*KIR2DL2* were both more frequent in cases ($n = 98$) than in controls ($n = 100$) in the Latvian population. Then Park et al (2006), reported that *KIR2DS2* and *KIR2DL5* were both significantly associated in the South Korean population, but that the *KIR2DS2* was instead less present in cases than in controls. They also found that *KIR2DL5* was significantly more frequent in cases than in controls. In an independent study, Ramos-Lopez et al (2009) attempted to confirm the association of *KIR2DL2* in German and Belgian families, by a transmission test of rs2756923, a SNP in exon 8 of the *KIR2DL2* gene. They found that there was over-transmission of the G allele of rs2756923 in T1D.

However, a number of issues surrounding these studies cast some doubt on the results. Firstly, as pointed out by Middleton et al (2006), the difference in frequency between *KIR2DS2* and *KIR2DL2*, two genes which are normally in high linkage disequilibrium (Single et al, 2007), is suspiciously large in both the Latvian study, 53% vs 81%, and in

the South Korean study, 20% vs 46% (Table 4.8). Secondly, in the Ramos-Lopez et al (2009) German/Belgian study, rs2756923 is not in Hardy-Weinberg equilibrium (HWE). Both these issues are possibly linked to genotyping errors due to differences in primer sequences. Thirdly, rs2756923 has since disappeared from the current genome build, which leads us to think that rs2756923 may not in fact tag *KIR2DL2* or at least not all isoforms of that gene. Finally, this KIR association has not been replicated in other populations including Dutch (van der Slik et al, 2003), Finnish (Middleton et al, 2006), Basque (Santin et al, 2006), Japanese (Mogami et al, 2007), South Brazilian (Jobim et al, 2010) and Chinese Han (Zhi et al, 2011) (Tables 4.8 and 4.9).

	Study	Pop	cases	controls
1	van der Slik et al (2003)	Dutch	149	207
2	Nikitina-Zake et al (2004)	Latvian	98	100
3	Middleton et al (2006)	Finnish	137	101
4	Santin et al (2006)	Basque	76	71
5	Park et al (2006)	South Korean	139	132
6	Mogami et al (2007)	Japanese	204	240
7	van der Slik et al (2007)	Dutch	275	215
8	Shastry et al (2008)	Latvian	98	70
9	Ramos-Lopez et al (2009)	Belgian	394	401
10	Ramos-Lopez et al (2009)	German	380	315
11	Jobim et al (2010)	South Brazilian	248	250
12	Zhi et al (2011)	Chinese Han	259	262
13	Mehers et al (2011)	British	551	168
14	Pontikos et al (2014)	British	6744	5362

Study	2DL1	2DL2	2DL3	2DL4	2DL5	2DS1	2DS2
1	94.6:97.6	55.7:48.4	91.9:92.3		50.3:46.9	36.2:35.7	55.7:47.8
2	95:98	81:32	86:91	98:100	65:55	43:27	53:25
3	97.1:100	35:41.6	94.9:96		46:55.4	43.1:48.5	38.7:41.6
4	97:98	52:62	93:95		49:66	48:54	52:63
5	99.3:100	46:34.8	98.6:98.5	97.8:97.7	42.4:84.1	33.8:43.9	20.1:47
6	98.8:98.5	15.4:13.7	98.8:99	100:100	40.8:34.8	40.8:35.8	15.4:13.7
7						41.1:35.8	53.5:48.8
8		82.65:32			66.32:55	43.87:27	54.08:25
11	95.6:97.6	49.2:54.4	87.9:86.4	99.2:100	56:49.6	46.4:36.4	52.8:53.6
12	93.82:96.56	28.19:32.44	98.46:99.62		40.15:42.37	37.84:37.4	28.96:30.92
13	97.9:100	52.8:53.6	94.2:90.5		54.1:56.5	42.9:41.7	52.8:53.6
14							

Study	2DS3	2DS4	2DS5	3DL1	3DL2	3DL3	3DS1
1	24.8:27.1	40.9:42	32.9:27.1	96:96.1			38.9:33.3
2	35:19	94:92	29:22	92:94	98:100	98:100	40:27
3	18.2:23.8	92.7:94.1		92.7:93.1			40.1:49.5
4	24:25	80:85	35:43	89:90			53:63
5	10.1:9.8	96.4:96.2	22.3:33.3	96.4:96.2	97.8:98.5	96.4:99.2	36:37.1
6	9.6:9.8	87.1:85.3	34.6:30.8	99.6:100	99.6:99.5	100:100	44.1:36.8
7							40.4:34
8	35.71:19						40.81:27
11	33.9:33.2	85.2:95.2	37.1:34	95.2:97.6	100:100	100:100	47.6:42.4
12	11.58:12.98	92.66:93.13	27.03:27.1	93.05:95.42			36.29:35.5
13	30.5:33.3	95.7:94.6	34.5:34.5	96.4:95.2		100:100	44.9:44
14				96.07:95.15			37.6:38.47

Table 4.8. Proportion of cases to controls in all known KIR studies in T1D. Study 14 is the study presented in this chapter. Table cells highlighted in gray are the ones which report a significant association.

	Study	Pop	cases	controls	2DL1	2DL2	2DL3	2DL4	2DL5	2DS1	2DS2	2DS3	2DS4	2DS5	3DL1	3DL2	3DL3	3DS1
1	van der Slik et al (2003)	Dutch	149	207	0.97	1.15	1		1.07	1.01	1.17	0.92	0.97	1.21	1			1.17
2	Nikitina-Zake et al (2004)	Latvian	98	100	0.97	2.53	0.95	0.98	1.18	1.59	2.12	1.84	1.02	1.32	0.98	0.98	0.98	1.48
3	Middleton et al (2006)	Finnish	137	101	0.97	0.84	0.99		0.83	0.89	0.93	0.76	0.99		1			0.81
4	Santin et al (2006)	Basque	76	71	0.99	0.84	0.98		0.74	0.89	0.83	0.96	0.94	0.81	0.99			0.84
5	Park et al (2006)	South Korean	139	132	0.99	1.32	1	1	0.5	0.77	0.43	1.03	1	0.67	1	0.99	0.97	0.97
6	Mogami et al (2007)	Japanese	204	240	1	1.12	1	1	1.17	1.14	1.12	0.98	1.02	1.12	1	1	1	1.2
7	van der Slik et al (2007)	Dutch	275	215						1.15	1.1							1.19
8	Shastry et al (2008)	Latvian	98	70		2.58			1.21	1.62	2.16	1.88						1.51
9	Jobim et al (2010)	South Brazilian	248	250	0.98	0.9	1.02	0.99	1.13	1.27	0.99	1.02	0.89	1.09	0.98	1	1	1.12
10	Zhi et al (2011)	Chinese Han	259	262	0.97	0.87	0.99		0.95	1.01	0.94	0.89	0.99	1	0.98			1.02
11	Mehers et al (2011)	British	551	168	0.98	0.99	1.04		0.96	1.03	0.99	0.92	1.01	1	1.01		1	1.02
12	Pontikos et al (2014)	British	6744	5362											1.01			0.98

Table 4.9. Case-control ratio in all known KIR studies in T1D. KIR studies in T1D. Study 14 is the study presented in this chapter. The case-control ratio is given per KIR gene. Table cells highlighted in gray are the ones which report a significant association.

Nonetheless, some of these KIR studies do report conditional association when they conduct subset analysis by grouping by age, HLA genotype or by grouping into activating or inhibiting composite KIR-HLA genotypes (Carrington et al, 2005; van der Slik et al, 2007). For example, van der Slik et al (2003) report association with *KIR2DS2* in the HLA-C1, *HLA-DQ2/HLA-DQ8* (high risk) subset of the Dutch cohort. Jobim et al (2010) report association with *KIR2DL1* in the HLA-C2 subset of the South Brazilian cohort. In the Chinese Han cohort, Zhi et al (2011) report association with *KIR2DL3* in the HLA-C1 subset. In the Japanese cohort, Mogami et al (2007) find association in the adult-onset diabetes subset (age of onset older than 35 years) after assignment into three KIR-HLA activation groups as defined by Carrington et al (2005). Mehers et al (2011) find association with *KIR2DS2/KIR2DL2* and *KIR2DL3* in the early-onset (less than 5 years old), HLA-C1 subset of the UK cohort.

Of concern in these analyses is that, as the starting samples are small (no more than 300 individuals), further subsetting and testing for multiple hypotheses (presence/absence of up to seventeen KIR genes) is likely to lead to false positives (Wittes, 2009).

Also since the HLA region is known to be associated with T1D it is difficult to tell whether the KIR-HLA interaction is significant, independently of HLA. In fact, these studies only control for HLA Class II and have not checked whether the effect is actually driven by other HLA Class I risk factors. Furthermore, it is unclear whether the established biological interaction between KIR and HLA should translate into the statistical KIR-HLA interaction claimed in those studies. As HLA-C is significantly associated with T1D before controlling for HLA Class II and HLA-B (Nejentsev et al, 2007; Howson et al, 2009), careful interaction analysis such as case-only tests (Yang et al, 1999; Cordell, 2009) are required to assess whether there is a significant epistatic KIR-HLA effect or if the reported associations to T1D are only driven by HLA-C or some other latent HLA risk factor.

4.5.2 My approach

As discussed, regions with great allelic and copy number variation are difficult to properly assess using genome-wide SNP arrays. While these arrays are typically cost effective ways to assay common genetic variation, very polymorphic regions can make the design of SNP probes difficult or impossible, which has contributed to low SNP coverage in the KIR region on most kinds of SNP arrays. The SNPs that do exist on arrays, such as ImmunoChip, are often discarded during the QC phase of any GWAS because they do not exhibit the expected three clusters. In contrast, assaying individual genes by other methods can prove expensive. For example, the qPCR assays used here to target *KIR3DL1* and *KIR3DS1* cost £12 per sample.

Our hybrid approach, the key steps of which are summarised in **??**, allowed us to perform the largest study (twenty-fold) of *KIR3DL1/KIR3DS1* copy number in T1D to date. In 12,106 samples, I found no association of *KIR3DS1-KIR3DL1* copy number with T1D, alone or conditional on presence of the HLA-Bw4 epitope. Our results suggest that, despite the association of certain HLA-A and HLA-B alleles with T1D and the established biological interaction between HLA-Bw4 and *KIR3DL1*, copy number variation in *KIR3DL1/KIR3DS1* is unlikely to have a significant effect on the risk of developing T1D.

Other KIR genes that are in high linkage disequilibrium (LD) with *KIR3DL1* and *KIR3DS1* are also unlikely to be associated. According to the Allele Frequency Net database (Gonzalez-Galarza et al, 2011), these include *KIR2DS4* (97 %) and *KIR2DL3* (86 %), for *KIR3DL1* and, *KIR2DL5* (81 %), *KIR2DS5* (84 %) and *KIR2DS1* (92 %), for *KIR3DS1* (`http://www.allelefrequencies.net/kir6010a.asp`). Thus, copy number variation in *KIR3DL1/KIR3DS1* or neighbouring genes is unlikely to be an important risk factor in T1D.

In order to better understand why rs592645 is the best available SNP for predicting

copy number variation in *KIR3DL1/KIR3DS1*, I used BLAT (Kent, 2002) to match the probe sequences of rs592645 on ImmunoChip against the allelic sequences of all KIR genes available from the Immuno Polymorphism Database (Robinson et al, 2010). Interestingly, I found that the SNP probes do not target *KIR3DL1/KIR3DS1* but instead bind uniquely to *KIR2DL4*, a neighbouring framework gene. Examining the *KIR2DL4* alleles matched by the rs592645 probes, I discovered, thanks to James Traherne, that the SNP probes are in fact picking up copy number variation of *KIR2DL4*005*, an allele of *KIR2DL4* that undergoes copy number variation along with *KIR3DL1/KIR3DS1* (Gómez-Lozano et al, 2005). This explains the small but persistent misclassification error rate of 2 % since our imputation is based on linkage disequilibrium between rs592645 and *KIR3DL1/KIR3DS1* rather than on perfect discrimination between our target genes.

4.5.3 Future work

I expect that, eventually, fully sequenced KIR haplotypes will be available in a large number of individuals. However long reads will be required for correct assembly, because of great sequence similarity in this region. According to our collaborator James Traherne, sequencing would require amplification of the the polymorphic exons using locus-specific primers and sequencing of the barcoded products. This would necessitate read lengths of at least 300 base pairs to span the exons. The cost of this using Roche 454 has been estimated to be of approximately £30 per sample.

Until KIR sequencing becomes sufficiently cheap, hybrid methods combining allele typing techniques such as qPCR in a subset of samples, with SNP typing in a larger cohort, are likely to remain the most cost-effective approach for large scale analysis. This method could also be applied to other allele typing techniques such as pyrosequencing (Norman et al, 2009) or sequence-specific oligo hybridisation (Martin et al, 2007). Another alternative to using raw SNP signal would be to use tagging SNP as has been done

in HLA imputation by Leslie et al (2008) and Dilthey et al (2013) at Oxford University. Here, instead of correlating copy number with raw SNP signals like in my approach, copy number is correlated with the genotype of flanking SNPs. However, from a recent seminar I attended on the 1st of November 2014, the speaker, the same Stephen Leslie of Leslie et al (2008) now at the Murdoch Childrens Research Institute in Australia, stated that the tagging approach, based on a Hidden Markov Model using positional information, that has successfully been applied to HLA imputation, performs poorly for KIR imputation. He explained that this is likely due to the unreliable positional information and unknown patterns of SNP LD in the region. Instead, Stephen Leslie and Damjan Vukcevic, found that training a random forests (RF) algorithm, which does not use positional information, on 300 SNPs taken on either side of the KIR region, performs better in the prediction of common KIR gene copy numbers. However, performance remains poor for rarer copy numbers or KIR alleles for which there is insufficient training data available. Furthermore, because of the high degree of homology between certain KIR genes, the ambiguity in their definition can lead to mislabelling of the training data in the reference panel. However Stephen Leslie did mention that perhaps the approach I have take of using signal intensities instead of genotype calls could improve on performance, since the genotype calls are often highly unreliable, most failing HWE. They intend to submit a manuscript in the coming months.

As discussed, hybrid qPCR/SNP approaches are particularly suited to the large case-control cohorts genotyped on ImmunoChip, since the platform contains numerous SNPs in the KIR region (Nikula et al, 2005). In fact in our dataset, I have observed other SNPs in KIR with more than three clusters which may correlate with copy number of other KIR genes and, given the availability of qPCR data, could be imputable in a similar manner. One possible improvement to my method, in order to achieve better prediction rate at smaller samples sizes (as suggested in Figure 4.8), would be to preferentially select

samples to qPCR from smaller SNP clouds, since these are more likely to correlate with rarer copy number groups (for example the 3-0 group in Figure 4.5).

I would recommend that this approach be adopted in KIR association studies which have so far been hindered by small sample sizes (Table 4.9). But it could also be applied more generally to other genes in non-genotypable chromosome regions of similar common copy number variation and sequence complexity as KIR, in order to leverage existing SNP datasets.

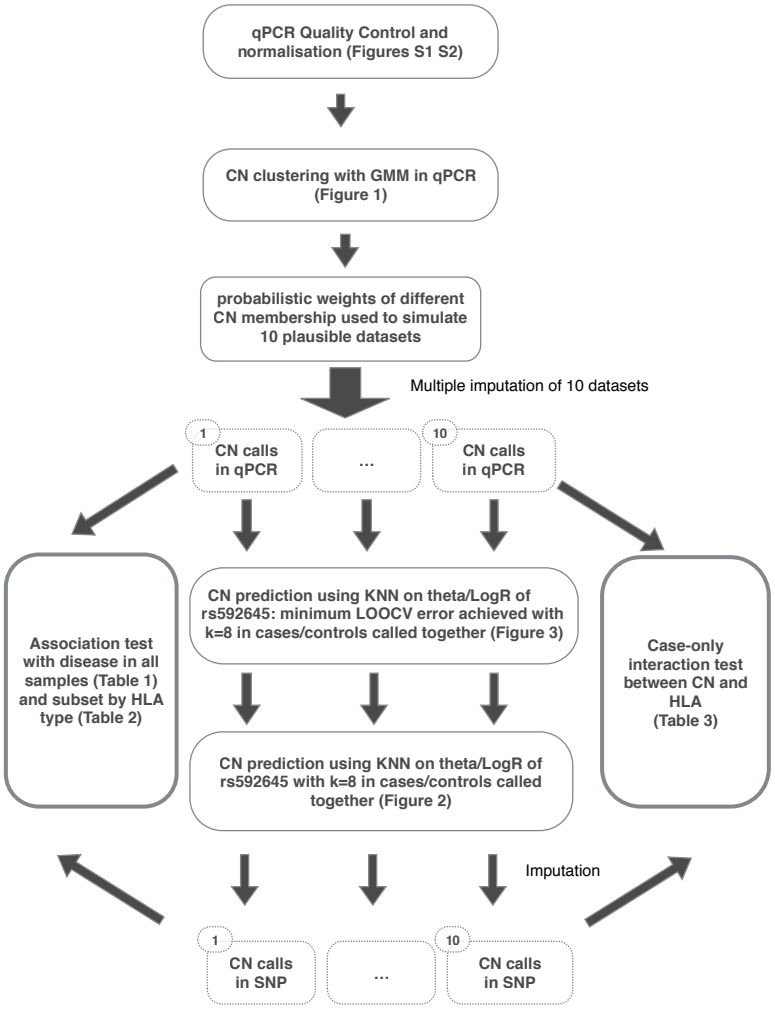


Figure 4.9. Flow chart summarising the key steps involved in the KNN imputation of *KIR3DL1/KIR3DS1* copy number in SNP data (R and θ signal) from qPCR copy number predictions obtained from GMM clustering. GMMs clustering of the qPCR data assigns to each sample a posterior probability of belonging to each copy number (CN) group. Using these probabilities I can allow for the uncertainty of the CN calling when testing for association with disease by using multiple imputation. Multiple imputation involves the simulation of datasets (in this case, ten) from the probabilities returned by the GMM. We find that the SNP rs592645 and k=8 minimises the LOOCV error rate. Association tests are conducted on each imputed dataset and inference combined using methods in the `R` package mice (van Buuren and Groothuis-Oudshoorn, 2011).

Chapter 5

Discussion

5.1 Important practical considerations with computational clustering methods

Throughout my thesis I have considered methods of computationally analysing flow cytometry and genetic datasets. The key motivation was that these methods would be more efficient, consistent, objective and better at dealing with uncertainty, than manual analysis of the data. In the case of flow data, automating gating can also provide the added benefit of formalising manual gating, which remains to this day a very subjective process. The hope was that these methods would improve on the reproducibility of results, thus enabling more powerful association testing. Assessing how well computational methods perform is a matter of debate and it is safe to say very much data-dependent. However, I will attempt here to give an overview of the important practical and theoretical considerations of these methods, in the context of my thesis.

Memory usage and running time Two crucial practical properties are their memory usage and running time. While the applications that have been studied in this thesis don't require real-time analysis, there are still some practical limitations on the amount

of memory and compute time demands. It is a well known fact in computer science that memory can often be traded for running time. For example, methods which rely on global knowledge of the data, such as computation of the complete pairwise distance matrix, have a large initial memory footprint (data matrix of size $\frac{N\times(N-1)}{2}$), but only require one pass of the data. A solution to the clustering solution is reached within just a few computational steps, by for instance, selecting a distance cut threshold on the dendrogram generated from hierarchical clustering to define K clusters. On the other hand, methods like K-means necessitate a much smaller memory footprint since they only compute the distance of every point to the K cluster means (data matrix of size $N \times K$), which is usually a much smaller number than the total number of points, but on the other hand, several updates of the matrix are required until the cluster centers are fixed. The choice of which method to apply is very much data dependent. In ungated flow cytometry, data matrices contain over a million rows, so the complete pairwise distance matrix is too large to fit in memory, making these type of methods impractical. I found that distance matrix computation was only feasible in subsets of the order of $10,000$, like for example the CD4+ lymphocyte subset, or data downsampled using the SPADE or binning in Chapter 3. The data downsampling as performed by SPADE, relies on first estimating the local density at each point and then preferentially thinning the data in regions of high density to even out the density across the whole sample. The local density estimation step is computationally intensive as it needs to consider the distance from a single point to all other points in the sample to find the number of neighbours. I found this step could be greatly sped up using ANN which uses the K dimensional trees (KD tree) lookup method. KD-trees are a space-partitioning data structure for organising points in a k-dimensional space which makes for an efficient way of storing a high-dimensional dataset to lookup proximal datapoints. It can also be used for approximation to reduce the number of datapoints which need to be considered.

This approach can also be applied to mixture models (McLachlan and Peel, 2004), but I haven't had the opportunity of trying this.

Consistency and accuracy Consistency can be defined as how often does a method return a similar result. This obviously depends on how much the data has changed. If a method is consistent, one would expect that small perturbations to the data should lead to small changes in the clustering outcome. However, algorithms which rely on initialisation using random starting positions like K-means may reach different clustering solutions, even on the same data! When running K-means I therefore had to pick the initial clusters or run it with multiple restarts. Over all runs of K-means the solution which returns the lowest within-sum-of-squares is picked. The accuracy of a method is defined based on how frequently the methods assigns the correct label. Hence it relies on the existence of a test dataset, typically labelled using manual analysis or some other method. In Chapter 2, accuracy was assessed by comparing the cluster proportions and means with those obtained from manual gating. In Chapter 3, the accuracy was determined in terms of RSS of the pSTAT5 response. In Chapter 4, I used qPCR labelled data to assess the prediction accuracy of the KNN classifier. However, labelled data may not always be available especially in the case of flow cytometry data. Also, in the context of flow cytometry, even when labelled data is available, this approach may not always be ideal, as it is merely comparing the relative agreement between methods rather than the objective truth. A sometimes more useful alternative is instead to assess the prediction accuracy with clinical outcome, case-control status or genotype. This is usually implicitly measured when doing an association test. The consistency-accuracy tradeoff is analogous to the variance-bias tradeoff in statistics. Consistency and accuracy are typically measured on simulated data.

Interpretability While a method might be accurate and consistent, it may be difficult to interpret the results, much like a black box. Making a model more flexible by adding parameters can obfuscate the relationship between the input parameters and the clustering output. Random forests and neural networks are example of methods from which it is difficult to extract an interpretable model to justify the result. This is an issue because as part of the iterative process of knowledge discovery, its important to understand what combination of features make objects distinguishable. Sometimes the additional information which is available from previous studies can be used to guide the clustering.

5.2 Dealing with noise by using prior knowledge

Having gone through some of the practical aspects I had to consider when applying computational clustering methods, I will now address some of the more theoretical challenges. Perhaps the biggest challenge in clustering of biological data which greatly complicates consistent identification of clusters across multiple samples is noise. Noise can be due to batch effects, for example staining discrepancies in flow cytometry, insufficient number of parameters to distinguish clusters, or because of sampling variation for clusters containing few elements.

In cases when certain characteristics of the data are known, these can greatly assist in the clustering especially with regards to identifying smaller groups. For example, in Chapter 4, I used prior information about the copy number of the *KIR3DL1/KIR3DS1* genes, obtained from qPCR, in order to identify clusters in the SNP dataset. The KNN was used to classify unlabelled points (without qPCR data) from the vote of their K nearest labelled neighbours. The optimal value of K was selected by minimsing the LOOCV. Another source of prior information in labelling the samples with qPCR, would have been to use expected copy number group frequencies obtained from previous studies

Jiang et al (2012) as prior group frequencies in the clustering of the copy number groups in the qPCR data.

In the flow data, the manual gates contribute prior information to guide the automatic gating. They yield information about the expected relative frequency of the different types of cells and their relative marker expression. Their absolute marker expression is generally not readily comparable across samples and requires normalisation. The established manual gating method involves obtaining gate coordinates from drawing gates in FlowJo and applying these same gate coordinates across samples. Manual gates can also provide initial starting parameters to a clustering algorithm, by calculating the mean and the covariance of the gated populations. Using an EM style algorithm, the parameters of the ellipse can then be refined to better fit the data. This is the approach taken by X-Cyt (Hu et al, 2013) which uses manual gates as templates. Also FlowJo has a magnetic gate feature which is described as "gate is as moved to accomodate the maximum number of events".

I have also tried this approach. I have let the mean of the ellipse be influenced by the data while the covariance was set as fixed. The Mahalanobis distance then allows us to go from a mean and covariance of an ellipse to a classification by defining a threshold above which points are excluded from the ellipse.

If the manual gating is done in several samples then they can be combined to define the priors in the mixture model, so to guide the parameter estimation, as is done in flowClust. If several samples have been manually gated then these can be incorporated in the definition of the prior on the parameters of the mixture model.

One outstanding problem remains with supervised clustering however, how to account for unlabelled data, or in the context of flow cytometry, nuisance clusters which are not part of the gating? While in manual gating these points are discarded and not explained by the clustering, automatic methods need to consider all points in the

dataset and consequently these points influence the gating. One solution is to define a background cluster but this is unlikely to work for multimodal distributions. Another solution could be to allow for the creation of a large number of clusters to account for all the background clusters which are not part of the study.

Supervised clustering can be really beneficial in trying to identify rare subsets. This is because rare subsets, by definition, may not always be visible in all samples.

Advantages and disadvantages to a tree structure to gating Some advantages of using a tree structure for gating are:

- fast, points can be eliminated at each step reducing the number of points that needs to be considered at each clustering step.
- Agrees with the established nomenclature facilitating comparison to manual gating and interpretability. Methods which use a binary structure are easy to interpret as they conform to the manual gating strategy which provides rules on how to obtain the leaf subsets.
- Reduces ambiguity.

Some disadvantages of using a tree structure for gating are:

- Bias, imposes an ordering on the markers and thus an an importance to the markers. It may be more ambiguous markers are gated on first. The RF gives us some idea of the importance of the markers.

  However their construction can be very sensitive to the underlying data and errors propagate. Errors may be mitigated by introducing conditional probabilities at each step. This becomes more of an issue as more subsets are defined. If no restriction are appied the rules can become over-complicated.

5.3 Identifying outliers

Outliers have an important influence on the mean of the data. They lie far from cluster means, usually in a sparsely inhabited region of space. Thus clustering methods and statistical tests which rely on mean or covariance estimation are sensitive to outliers. Certain outliers may be clearly identifiable because their values lie at the extreme of the instrument range, far from the rest of the data. Others which lie within a more plausible range of values are harder to identify. Also in higher dimensions, data points can be outliers without being outliers in any single dimension. Generally in order to detect outliers, large sample sizes are required. Reducing the dimensionality makes outliers easier to identify.

Flow cytometry datasets contain millions of events and many events are discarded as part of the gating process. Some of them are real biological clusters which are ignored because they are not part of the cell populations under study: for example the first gate drawn is usually on the lymphocytes whereas the monocytes and granulocytes are often ignored. This this in part because they are no markers included in the experiment defined on these cell populations which would allow any further division. Other events are discarded because they are outliers. These outliers occur commonly in flow cytometry datasets caused by debris and cells clumping together. They can be found at the extremes of the value range on a given channel, but they may also be found with intermediate values and these are harder to spot and require large samples sizes to confidently call as outliers. Model-based methods can account for these by defining a background mixture with a covariance defined on the entire sample which essentially mops up all points which are not assigned with high posterior weight to any of the known components. Mclust allows to define a noise component for that purpose. On the other hand, flowClust has an outliyingness parameter which is inversely proportional to the Mahalanobis distance from a point to a cluster. Another method which does not rely on

a background component is to exclude low density points from belonging to any of the components. A metric for detecting outliers is the Mahalanobis distance, large distances imply a point is far from a cluster center:

$$(\boldsymbol{x_i} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu})$$

where $\boldsymbol{x_i}$ is the point, and $\boldsymbol{\mu}$ is the cluster mean and $\boldsymbol{\Sigma}$ is the cluster covariance.

One issue however is that the outliers influences and greatly inflate the covariance matrix which is used in calculating the Mahalanobis distance. To address this, R packages like robustbase, use leave-one-out methods to identify outliers which have high leverage on the covariance estimation. Another outlier metric, commonly used in linear regression is Cook's distance, which returns the leverage of every data point i on the estimation of $\hat{Y}$ when the point is left out:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j \ - \hat{Y}_{j(i)})^2}{p \ \text{MSE}}$$

In genetics, outliers may be harder to spot because of smaller sample sizes which also implies that we are also generally more reluctant to exclude samples. In Chapter 4, a whole qPCR plate was excluded because the ΔCt distribution didn't align well even after normalisation. Later I found however that qPCR samples may possibly be scored independently of where they lie in relation to other samples by considering their background Ct median and spread. This score could have been used to downweight these samples in the mixture model clustering which was used to assign them to copy number groups.

As well as techniques for visualising multiple dimensions. Tools are being developed for visualisation of summary statistics in mutliple samples in order to spot outliers. Boxplots are the typical visualisation tool for viewing univariate intensity data but in

flow cytometry where we typically have multivariate data, SPICE presents an overview using pie charts of multiple samples (Roederer et al, 2011).

They are however obvious outliers and less obvious outliers, which is why I believe a confidence score is appropriate in calling outliers. While certain outliers can be confidently called independently because their values are at the extreme of the range of the instrument, others are less clear and should not be discarded completely because in light of new data they may be assigned to existing clusters or to new clusters.

In particular distinguishing between rare subsets and outliers is hard as illustrated in Chapter 4, where the 3-0 copy number group is rare and could have been regarded as outliers had not we not had strong prior evidence supporting its existence from previous studies. However having prior evidence supporting the existence of a cluster is not sufficient to identify the position of the cluster. Priors usually determine expected proportions but give little information about the exact position and shape of the cluster, which are usually experiment specific. This can be due to the stability of the biological sample but also the stability of the instrumentation and sample preparation. In genotyping calling a genotype based on only a few samples is sometimes possible, since DNA is a very stable molecule and SNP arrays are highly standardised (Di et al, 2005; Giannoulatou et al, 2008). In flow cytometry, the MFI of cell populations is generally not directly comparable across staining panels but instead clusters can be defined in relation to one another.

5.4 Prioritising normalisation or clustering

Normalisation is a two-part process, the first part involves matching across samples, the second part involves transforming the data so that the data is directly comparable across samples.

In certain situations, the first part of normalisation is similar to clustering and can

rely on related algorithms. For example normalisation by peak-alignment can use clustering of univariate data to identify peaks (with known k). This is the method I used in Chapter 2 when gating bead data and the transform I defined to align the peaks of the bead data, was later applied to the biological data to do bead-normalisation. When K is unknown a sliding window approach can be used to estimate the number of modes peaks. The highest peaks can then be selected across samples in the hope that these are the same across samples. Here normalisation is comparable to mode or bump hunting on the density function. The number of peaks returned is influenced by the span of the sliding window. A large window span will tend to oversmooth the data, leading to fewer peaks while a smaller window will call more peaks, but increases the chances of picking up spurious peaks. Hence the number of clusters is controlled by the window span parameter, although the exact relationship between the parameter and the number of clusters is data dependent.

While normalisation is typically applied to univariate data whereas clustering is applied to multivariate clustering. Meta-clustering which involves matching clusters across samples can also be considered to be a form of normalisation/clustering.

The question really boils down to whether, when we have sufficient data, should we normalise between samples or cluster each sample separately?

In genotyping, between sample normalisation is usually applied to enable data pooling, since we have a large number of markers and too few replicate probes to do within sample clustering, Clustering is then applied on the pooled data to call genotypes. In genotyping, normalisation is faciliated by having the same number of markers per sample. Features are directly comparable. On the other hand, in flow cytometry, there is much variation in the count of cells per sample.

In flow cytometry, we have much larger number of cell than markers per sample, so clustering tends to happen within a single sample in order to identify cell populations,

which are then meta-clustered across samples. Normalisation can be done after the clustering to match and align the MFIs across samples. However, while this transform is appropriate for bead data which is not expected to change, it is not obviously not appropriate if the MFI is expected to change with genetic differences or ex-vivo stimulation. It can nonetheless be useful as a means to match populations across samples when the relative cell proportion is the parameter under study. Normalisation can also be done before the clustering. On one hand in can facilitate the clustering so that the gate positions need not be moved betwen samples as is suggested in Hahne et al (2009). Or on the other it can allow for pooling of samples, which can be useful in order to identify rarer cell populations, to improve the signal-to-noise-ratio or to define a transform which can be used to align the data. However from my experience I have found that normalisation of flow cytometry is actually a harder problem than clustering because of the level of noise inherent in these data. Univariate clustering in flow does not make use of the full information available. In my thesis, I have used normalisation both before and after clustering. In Chapter 4, I have have applied normalisation to align copy number peaks in the ΔCt of qPCR plates before the clustering. On flow data in Chapter 2, I was able to improve the repeatability of CD25 MFI by correcting long-term fluctuations thanks to bead normalisation. In Chapter 3, unfortunately the variation in pSTAT5 MFI was not adequately captured by beads, so instead I attempted various other normalisation approaches. Since pSTAT First I attempted to correct for background

Normalising after clustering: good for matching cluster across samples In the case when the centre of mean of the gate has moved, after a few E iterations of the EM algorithm for example, it possible to realign the gates on the data. The advantage of this approach is that it doesn't require all the data to be moved. But normalisation is still required in the meta-clustering step to compare properties of the clusters. Normalisation is effectively analogous to meta-clustering since we are attempting to match cell

populations across samples. Normalisation facilitates matching clusters across datasets but in doing so can remove meaningful biological variation.

The samples are related and we wish to exploit this structure without forcing data to be perfectly comparable. A less biased approach is to allow for different cluster location and shapes across datasets and instead incorporate the clustering in some sort of hierarchical framework. Normalisation or meta-clustering is a necessary step when data is pooled across studies.

Normalising before clustering: good for pooling to find identify rare populations If applied before the clustering it enables data pooling to identify clusters across all samples. Pooling may improve the coefficient of variation of populations but also facilitates matching clusters across samples, a step known as meta-clustering. Pooling is crucial for identifying rare groups which are difficult to identify within one sample. If the alignment isn't perfect, the clustering results can be refined across all samples.

Theoretically, if flow cytometry data was perfectly aligned across days then gates should not be moved, unless they are dependent on some internal control. For example a threshold gate on a positive subset may be defined in relation to another population of cells.

As discussed in Chapter 2, the cell phenotypes usually measure, the mean intensity on a particular marker, or the percentage of cells can be more or less sensitive to the position of the gate, depending on the shape of the distribution.

Conclusion Whilst normalisation is generally a necessary process for pooling data or comparing across samples, over correction can be detrimental to the analysis. Normalisation is meant to be a simpler preliminary step to simply the clustering, however, in flow cytometry, because of batch effects and unequal number of events per sample, normalisation can be as challenging as performing the gating.

However, when they are insufficient data points, pooling is necessary. Clustering sometimes relies on pooled samples when they are not sufficient data points to form clusters such as in the case of rare subsets like 3-0 in KIR (Chapter 4) or Tregs in flow cytometry data. In this situation some form of normalisation is usually necessary. Unfortunately, multivariate normalisation is not always trivial and relies on defining linear transforms.

Some issues with normalisation Normalisation removes unwanted experimental variation to make data comparable even when collected on different days, processed with different protocols or analysed with different instruments. However distinguishing between what is unwanted and what is meaningful variation relies ultimately on a bias-prone judgement call.

As we have seen, different normalisation approaches make different assumptions about what is unwanted variation and the shape of the data. The actual choice of normalisation depends on the characteristics of the data we wish to compare.

The Bayesian question of relative weight of prior vs data is very relevant to gating. The position of the gates are the priors. Is it better to have an absolute definition/threshold or should the gates be relative to the data?

One issue with normalisation based on peak alignment is that the noise variation is greater than the biological variation between samples, so much so that in certain samples, the peaks are only identifiable after preliminary gating of lymphocytes.

We have only considered the univariate density function here but identification can be extended to the multivariate case.

Normalisation of multivariate distributions When applying a linear transform to multivariate data the correlation between the dimensions is preserved since the multiplicative factors cancel out. However the covariance changes. Applying a non-linear

transform also changes the correlation. If the data is binned using flowBin or clustered using SPADE then a transform could be a mapping like Earth Moving to make both distributions of event proportions identifical across samples.

Combining transforms with clustering Choosing an optimal transform in flow data is not trivial. Transforms tend to be channel specific and sometimes even sometimes sample specific. The FlowClust solution proposes estimating the Box-Cox parameter as part of the clustering using a numeric optimiser. However the parameters need to be transformed back, for allowing for a different transform per sample. This is an issue if comparing intensity data but obviously if comparing cell ratios then the transform is of little consequence. Nonetheless the transform can have an important effect on the clustering result.

Clustering is an iterative method of discovery not a fully automated process. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties. Clustering can also be viewed as a latent variable problem where the cluster labels are considered to be the missing data.

5.5 Influence of normalisation and clustering on statistical association test

Ultimately the goal of normalisation and clustering is to allow for association testing of genetics with biological traits. In complex biological traits, a significant proportion of the variation is explained by the cumulation of a large number of small genetic effects.

The most studied factor which influences the statistical power to detect such effects is sample size, and we have witnessed ever-increasing sample sizes which have greatly benefitted from dropping genotyping and sequencing costs. However, a less explored and harder to define factor is the accuracy with which traits are measured. For example

in T1D, depending on the stage at which the disease is diagnosed, the phenotype can be quite different. If the disease is diagnosed earlier then the associated phenotype may be recurrent auto-antibody positivity or if diagnosed late, high blood glucose. In most diseases, case heterogeneity is a confounder and one which needs addressing by collection of additional clinical covariates during the recruitment of cases. Since these disease status phenotypes are hard to define accurately and consistently, this has motivated looking at better defined intermediate phenotypes (also known as endophenotypes), such as the CD25 cell phenotypes of Chapter 2, or the pSTAT5 exvivo response cell phenotype of Chapter 3, which might show stronger correlation with genotype as they are closer to the gene expression. Cell phenotypes such as relative cell frequency over its parent population or expression of surface markers can in theory be measured accurately with flow cytometry, but there are still factors which influence the accuracy at which these can be measured. In particular in Chapter 2 of this thesis, I have concentrated on how clustering can influence the relative frequency and MFIs of the cell populations under study. Other known factors are, staining batch effects as seen in Chapter 3, or long-term instrument sensitivity as seen in Chapter 2, which influence the cell population MFIs, or sampling variation on each bleed which influences the cell frequency.

Using recalled individuals, I have assessed the repeatability of the cell phenotypes.

If the repeatability is good then the within-individual variance is small which should increase our power to detect small differences.

Consistency is an important factor in identifying clusters. Decomposing the variance into within sample and between variation

Subset analysis is also very important because of a phenomenon known as Simpson's paradox whereby a trend that is visible in a group may not hold in all subsets of the group and may even be reversed in certain subsets of that group.

When a hierarchical gating approach is taken, errors at the top of the hierarchy

propagate down the tree. This complicates an automated approach to step by step gating since errors which occur in earlier steps of the gating may not be recovered from.

In the same way wrongly gating samples can lead to outliers which can create false positive association due to their high leverage, fixed gating can lead to all samples have the same profile.

Reproducibility in flow cytometry is challenging. While in genetic data, the number of probes across arrays is constant and identifiable, flow cytometry data can contain very different number of events between samples. Also distinguishing staining noise from actual differences in cells biology requires a certain level of prior knowledge which is difficult to implement programmatically. Subsequently, a sample from the same individual analysed on different days can have a very different profile. Biologists tend to have strong opinions on which markers are stable and objectively it does appear that certain stains are much more stable than others. As seen in Chapter 2, this within-individual variation greatly compromises statistical power in detecting between-individual effects.

In long-running experiments when samples are analysed over a period of months, some noise may be attributed to variation in instrument sensitivity. In Chapter 2, we showed we can account for this using beads. However beads do not capture variation on shorter time scale or due to batch effects. Staining and in particular intra-cellular staining, of internal markers such as STAT5 and FOXP3, is subject to batch and day variation.

Another approach is to align the peaks of the univariate distribution in the whole sample as was done for qPCR ΔCt values in Chapter 4. However, in the flow cytometry samples we studied, we found that the peaks cannot always be identified reliably and choosing the right window-size parameter for peak finding algorithms is channel specific and not trivial. Also mismatching of peaks in alignment is more detrimental to repeatability than not any doing normalisation.

5.6 The future

The methods discussed in this thesis suggests that there are many scenarios in which flow cytometry analysis can be automated. There are however a number of outstanding challenges, some technical, some more philosophical, in applying these methods.

Seeing is believing Statisticians, biologists, we all like to visualise our data. We rely on visual inspection as a quality control check, learning about properties of the data like for example if the data is skewed or symmetrically distributed, looking for patterns, confirming clustering results, spotting outliers. While visualisation works well for up to three dimensional data, information is lost when higher dimensional datasets are decomposed into a series of two-dimensional projections. Clusters which exist in higher dimensions do not necessarily map to clusters in two dimensions. This has motivated research into how to visualise high-dimensional data in two-dimensions with minimal loss of information. In Chapter 3, I presented one these approaches, SPADE which relies on a network visualisation of a dataset using a minimum spanning tree. There are others which use more probabilistic approaches such stochastic neighbour embedding. While I agree that visualising high-dimensional data using network representations can be insightful, beyond a certain number of nodes, network visualisations quickly become uninformative. Certainly in the case of flow cytometry some clustering or data smoothing is required to reduce the number of points.

Experiment vs analysis: a communication problem Analysing flow cytometry data has brought to light the many issues surrounding the generation and analysis of data, and more generally with the division of labour. Perhaps the greatest obstacle to standardising analysis is our inherent dislike of rules and standards which we may find superfluous. However, simple things like naming conventions waste precious man hours for people analysing the data. Part of the solution is to involve the biologists so that

they appreciate the implications, another part of the solution is to encourage automation of the more tedious tasks.

On the other hand, perhaps us statisticians need to have an understanding of: the underlying technology, the purpose of the experiment. For example, in flow cytometry, a large number of events are debris of no biological interest. Similarly some patterns may be simply staining artefacts or from sample prepartion (permeabilisation).

A great deal of time was also spent on seeing if clustering could be fully automated. One approach is to select large K and then merge clusters together with flowMerge. My approach was generally to select a K that gave consistent clustering results across samples.

While I am sure these observations are not only specific to flow cytometry, they are more apparent perhaps in flow given how much freedom flow cytometry offers both in terms of generation and analysis.

Incomplete experiments and small sample sizes The majority of experiments undertaken in flow cytometry are pilot experiments or tubes ran to test and optimise panels. Pilot experiments are often implemented with varying degrees of thoroughness. Normalisation beads or controls are not consistently which complicates analysis. While the FCS files are saved and analysed by the person who generated the data, the naming and documentation is incomplete making it hard to automate the analysis of these data. Since the FCS file does not contain sufficient metadata to understand the context of the experiment, the name of the FCS file is typically used to map the sample back to the donor in order to retrieve covariates such as disease status, age, sex or genotype. A lot of my work unfortunately has involved dealing with typos and inconsistencies in these file names. Also channel names as given in the FCS files are not always consistent across experiments. When the FCS file do contain metadata it does not always match the naming of the file. For example the data as strored inside the FCS file did not

match the date given in the filename. In particular, FCS files do not contain sample identification information, which makes matching back to genotypes cumbersome and error-prone.

Typically flow cytometry experiments do not have large number of samples because in most labs, sample preparation and running tubes on the flow cytometer are manual operations. However there are labs where these processes have been automated with robotics and consequently may run thousands of samples a day. Automatic methods are more pervasive in those labs since manual analysis is no longer a viable option. Presently most flow experiments contain too few samples to do well-powered association testing.

Stability of markers Staining in flow cytometry is notoriously noisy. Even when using the same fluorochrome-antibody panel and the same PMT voltage, the shape and location of clusters is not stable. The treatment of the cells can also lead very different scatter patterns (example Tony vs Marcin). While the scatter channels are very noisy because of debris, for a given panel and experimental protocol, the location of the clusters should not move much on the scatter channels because the morphological attributes of the cells should not be dependent on staining titration.

Flowjo interface with R FlowJo is the main tool used by immunologists for identifying groups of cells in flow cytometry. The unit of work in FlowJo is the workspace in which FCS files are first loaded and then gated. The workspace also saves the cell populations statistics which need to be updated when the gates move.

Unfortunately, parsing FlowJo workspaces in order to extract manual gates is not straighforward. Although there are several BioConductor packages designed to import and parse flowJo workspaces, flowUtils gatingML, flowJo, flowWorkspace, I have found the R/FlowJo interface to not be very reliable, although the flowWorkspace was able to parse it returned the statistics calculated in FlowJo were wrong. This is probably why

Vincent Plagnol developed his own XML parser to extract gates from flowWorkspace files but this approach is time-consuming as it requires in knowledge of the FlowJo XML schema which frequently changes on each new release of FlowJo. Furthermore the gate coordinates in FlowJo are often imprecise. In fact I found that just loading an FCS file into FlowJo and reexporting it changes the data! In the end I found the best solution was to export CLR files which are simply the classification results from FlowJo. Unfortunately, instead exporting these files in a compressed memory-efficient format binary format, FlowJo, in its unfathomable wisdom, exports them in text which results in very large files and makes exporting of all clustering results impractical. Hence I have often resorted to exporting only a few CLR files from which I can estimate the gate coordinates. One thing however that lacks from the CLR format is to retrieve the gate coordinates is the dimensions in which the gate is defined. One method of approximating gate coordinate is to calculate the mean and covariance of a CLR cluster and to use the Mahalanobis distance, hence approximating the cluster with an ellipse. However, it helps to know in which dimensions the gate was defined manually.

Another solution to including manual gates in R without relying on their coordinates in FlowJo is to draw polygons in R and use the `R` function `in.polygon` to extract points in the polygon.

All that said, there is definitely a gap in the market for a new piece of software which reconciles manual, supervised and unsupervised flow cytometry analysis, as well as provides further visualisation techniques.

Mass cytometry Time of flight cytometry (CyTOF) is a biotechnology which combines mass spectometry with cytometry. The throughput is not as high as fluorescence flow cytometry but the number of markers which can measure up to 34 markers. Also no side and forward scatter information is given. It cannot be used for sorting as the cells are destroyed when measured. Since it does not report side and forward scatter

live/dead markers tend to be used instead to spot debris.

5.7 Summary

Larger datasets have allowed us to see finer biological variation both in genotypes and cell subsets than previously possible. However sometimes taking a different view of the same dataset, by doing a different type of experiment, like qPCR, or adding parameters, for example additional markers in flow cytometry, can help uncover patterns which might not have been visible even at larger sample sizes. For example in Chapter 4, qPCR allowed us to discover a SNP predictive of KIR3DL1/3DS1 copy number. Furthermore, even without doing futher experimens, analytical methods such as unsupervised clustering can reveal previously unnown features. For example in Chapter 3, unspervised clustering algorithms analysing pSTAT5 response at different doses of proleukin uncovered responsive subset of cells we previously ignored.

Although large datasets can support methods with larger number of parameters as they are less prone to overfitting, it is easy to fall into the trap of applying over complex methods to account for all the intricacies of the data, when in practice, simpler methods may perform as well and are much faster and easy to implement. Simple methods can also be combined to reach a consensus and this popular machine learning approach known as boosting may increase performance at the expense of interpretability.

Interpretability is perhaps one of the more important issues in order to encourage biologists to use these methods. While mixture modelling approaches are conceptually close to manual gating, probabilistic populations are not intuitive to biologists, so their true power cannot be fully exploited. Biologists enjoy manual gating becomes it gives them the freedom to draw somewhat arbitrary exclusive gates. This freedom however comes at the cost of exacerbating the disagreement in standards and definitions in immunology. While discrepancies in gate positions are unlikely to make much of an impact

on the MFI and relative proportion of common cell populations, they can make a big difference on rarer cell populations such as regulatory cells.

In order to encourage the use of these methods, biologist need to be lured in by incrementally habituating them to these tools. Completely removing the two dimensional visualisation for example might be discouraging. Also the cellular hierarchical view is so deeply engrained in the minds of certain biologists that presenting them the clustering results in a different order would perturb them. Although attractive, this hierarchical model misses populations and imposes a directionality in cell lineages which may not always be correct. In the same way that phylogenies change as new animal species are discovered, the cell lineage model should be influenced by the discovery of new cell subsets.

A first step is to use the manual gates but to allow them to move with the data.

While in my opinion automated clustering need to be applied more widely in flow cytometry data analysis, hence it is important that we continue developing these methods because biologists will need to resort to fully automated method once the number of parameters or number of samples becomes unmanageable. However, I recognise that these methods require some level of expertise and decent visualisation to guide (and reassure?) the user. Although, over-reliance on visualisation can mislead the analysis of high-dimensional because clustering is always projected back to two dimensions or linear combinations of dimensions. The automatic gating of flow cytometry community is strong with a lot of contibutions to BioConductor and the GenePattern web interface from the Broad Institute. In particular two labs, Raphael Gottardo at the Fred Hutchinson Cancer Research Center in the USA and Ryan Brinkman at the Terry Fox Laboratory in Canada, are central in developing auto gating software and bring together the automatic gating flow cytometry community as part of the FlowCAP challenge every year.

In Stanford, Gary Nolan lab are using mass cytometry to analyse cell heterogeneity in cancer. These high-dimensional datasets require visualisation and Dana Pe'er group at Columbia University have devised various tools to do so.

I have also identified regions of flow cytometry which need more work such as normalisation and selection an optimal transform. Both of these can be included as part of the clustering step. Logarithmic transform greatly influence the identification of lower intensity populations which overlap into the negative range. The wrong transform can introduce splits giving rise to spurious cell populations. This is why applying a straightforward arcinsh transform like is done in spade is wrong. In FlowJo, the transform is selected visually, given the knowledge of what cell populations to expect. The only existing automated methods of optimally selecting a transform that I am aware of are flowTrans and flowClust. FlowTrans assumes an underlying Gaussian distribution and used maximum likelihood to estimate the optimal transform parameter. But there is still the question of which transform function to apply? FlowClust applies a Box-Cox transform for which the lambda parameter is estimated as part of the ML estimation.

There is also a growing need for non-proprietary software which integrates well with the manual analysis. The openCyto BioConductor package currently being developed by Gottardo The automated methods need to complement the manual methods for now, so that the change from manual to automated happens gradually. On the other hand by continuously benchmarking automated analysis against manual, we are not exploiting the true power of automated algorithms which is to teach us new biology or to put back into question our hiearchical view of immunology. As an example I ran flowClust unsupervised with a large number of clusters and then picked the cluster which gave the best association with each SNP. This is the idea which was explored with flowMeans. The issue however is the metaclustering step of matching clusters across samples is not trivial especially if there is a lot of noise between samples.

Sophisticated the methods are no replacement for good data. Being able to make this judgment call between good and bad data necessitates understanding of the experimental context. This prior knowledge is acquired through having seen a large number of samples and hence not encoded into an automated method.

Processing in larger batches or perhaps reducing the human element in flow cytometry is a first step towards automation. However as the number of samples grows so will the need for computational methods and the gap between the biological of computational way of thinking will become less striking.

On a sufficiently large dataset methods all methods tend to be equivalent. As the datasets are growing larger, in a Bayesian setting the data will have a much stronger influence than the prior.

In conclusion, fully unsupervised methods cannot be expected to deal with the level of noise possible in flow cytometry experiments, beyond a certain threshold of uncertainty, a sample is of little value. Poor staining can make populations difficult to distinguish or can even make populations disappear. Spillover introduces artificial marker correlation and can increase or decrease the fluorescence intensity of cell populations. With an in depth understanding of the patterns of noise, it is possible to develop more targeted approaches such as Adiyct but may run the danger of including outliers. Ultimately deciding when or not to include an outlier sample is down to a judgement call from the person who generated the data, so in that regard some subjectivity can persist which can introduce bias. As we've seen in Chapter 2, the 1D sequential top down gating strategy can easily be coded up as an algorithm using mixture models or bead-derived thresholds. However poor staining or instrumental configuration can lead to unexpected distributions. In Chapter 3, the staining is even worse

Appendix A

Flow markers

CD3 CD3 marks all T cells

CD8 Cytotoxic T cells. Killer cells.

CD4 A protein found on a subset of T lymphocytes. Helper cells. Response orchestrator.

CD31 Largely present on naive CD4 T cells, is lost on maturation of naive cell after leaving the thymus.

CD45RA The protein isoform lost on activation of naive $CD4^+$ and $CD8^+$ T cells. It can be used to distinguish CD45RA high naive cells from CD45RA low memory cells.

CD127 The alpha chain of the IL-7 receptor. The IL-7 receptor is expressed on various cell types, including naive and memory T cells, and usually expressed at higher levels on T effector and regulatory T cells.

CD25 Better known as the IL2RA, the alpha chain of the heterotrimeric IL-2 receptor. High affinity binding of IL-2 requires all three chains of the receptor.

CD122 The beta chain of the IL-2 receptor.

CD132 The gamma chain of the IL-2 receptor.

CD56 NK cell marker.

CD19 Found on the surface of B-cells. It is expressed on follicular dendritic cells and B cells. It is a lineage marker which is lost on maturation to plasma cells.

CD69 A protein induced by the activation of T lymphocytes and Natural Killer cells. It is involved in lymphocyte proliferation and functions as a signal-transmitting receptor in lymphocytes.

Bibliography

Aghaeepour N, Chattopadhyay PK, Ganesan A, O'Neill K, Zare H, Jalali A, Hoos HH, Roederer M, and Brinkman RR. 2012. Early immunologic correlates of HIV protection can be identified from computational analysis of complex multivariate T-cell flow cytometry assays. *Bioinformatics* **28**: 1009–1016.

Aghaeepour N, Finak G, FlowCAP Consortium, Dougall D, Khodabakhshi AH, Mah P, Obermoser G, Spidlen J, Taylor I, Wuensch SA, et al. 2013. Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods* **10**: 228–238.

Aghaeepour N, Nikolic R, Hoos HH, and Brinkman RR. 2010. Rapid cell population identification in flow cytometry data. *Cytometry Part A* **79A**: 6–13.

Arya S, Mount D, Kemp SE, and Jefferis G. 2013. *RANN: Fast Nearest Neighbour Search (wraps Arya and Mount's ANN library)*. R package version 2.3.0.

Bagwell C. 2005. HyperLog - A flexible log-like transform for negative, zero, and positive valued data. *Cytometry Part A* **64A**: 34–42.

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, et al. 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics* **41**: 703–707.

Bashashati A and Brinkman RR. 2009. A Survey of Flow Cytometry Data Analysis Methods. *Advances in Bioinformatics* **2009**: 1–19.

Bashirova AA, Martin MP, McVicar DW, and Carrington M. 2006. The killer immunoglobulin-like receptor gene cluster: tuning the genome for defense. *Annual Review of Genomics and Human Genetics* **7**: 277–300.

Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, Cotsapas C, Shah TS, Spencer C, Booth D, Goris A, et al. 2013. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics* **45**: 1353–1360.

Bell GI, Horita S, and Karam JH. 1984. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* **33**: 176–183.

Bolstad BM, Irizarry RA, Astrand M, and Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**: 185–193.

Boyman O and Sprent J. 2012. The role of interleukin-2 during homeostasis and activation of the immune system. *Nature Reviews Immunology* **12**: 180–190.

Brusko TM, Wasserfall CH, Hulme MA, Cabrera R, Schatz D, and Atkinson MA. 2009. Influence of membrane CD25 stability on T lymphocyte activity: implications for immunoregulation. *PloS one* **4**: e7980.

Calinski T and Harabasz J. 1974. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods* **3**: 1–27.

Carrington M, Wang S, Martin MP, Gao X, Schiffman M, Cheng J, Herrero R, Rodriguez AC, Kurman R, Mortel R, et al. 2005. Hierarchy of resistance to cervical neoplasia mediated by combinations of killer immunoglobulin-like receptor and human leukocyte antigen loci. *The Journal of Experimental Medicine* **201**: 1069–1075.

Chan C, Feng F, Ottinger J, Foster D, West M, and Kepler TB. 2008. Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry Part A* **73A**: 693–701.

Chris Fraley Adrian E Raftery TBM and Scrucca L. 2012. mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597, Department of Statistics, University of Washington.

Clayton DG. 2009. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genetics* **5**: e1000540.

Cordell HJ. 2006. Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. *Genetic epidemiology* **30**: 259–275.

Cordell HJ. 2009. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics* **10**: 392–404.

Cron A, Gouttefangeas C, Frelinger J, Lin L, Singh SK, Britten CM, Welters MJP, van der Burg SH, West M, and Chan C. 2013. Hierarchical Modeling for Rare Event Detection and Cell Subset Alignment across Flow Cytometry Samples. *Plos Computational Biology* **9**: e1003130.

Cudworth AG and Woodrow JC. 1974. Letter: HL-A antigens and diabetes mellitus. *The Lancet* **2**: 1153.

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, CORDELL HJ, Pritchard LE, Reed PW, Gough SC, Jenkins SC, and Palmer SM. 1994. A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* **371**: 130–136.

Dean PN, Bagwell CB, Lindmo T, Murphy RF, and Salzman GC. 1990. Data file standard for flow cytometry. *Cytometry Part A* **11**: 323–332.

Dempster AP, Laird NM, and Rubin DB. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B-Methodological* **39**: 1–38.

Dendrou CA, Fung E, Esposito L, Todd JA, Wicker LS, and Plagnol V. 2009a. Fluorescence Intensity Normalisation: Correcting for Time Effects in Large-Scale Flow Cytometric Analysis. *Advances in Bioinformatics* **2009**: 1–6.

Dendrou CA, Plagnol V, Fung E, Yang JHM, Downes K, Cooper JD, Nutland S, Coleman G, Himsworth M, Hardy M, et al. 2009b. Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nature Genetics* **41**: 1011–1015.

Dendrou CA and Wicker LS. 2008. The IL-2/CD25 pathway determines susceptibility to T1D in humans and NOD mice. *Journal of clinical immunology* **28**: 685–696.

Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, and Yang G. 2005. Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics (Oxford, England)* **21**: 1958–1963.

Dilthey A, Leslie S, Moutsianas L, Shen J, Cox C, Nelson MR, and McVean G. 2013. Multi-Population Classical HLA Type Imputation. *Plos Computational Biology* **9**: e1002877.

Duong T, Cowling A, Koch I, and Wand MP. 2008. Feature significance for multivariate kernel density estimation. *Computational Statistics & Data Analysis* **52**: 4225–4242.

Durbin BP, Hardin JS, Hawkins DM, and Rocke DM. 2002. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics (Oxford, England)* **18 Suppl 1**: S105–10.

Finak G, Bashashati A, Brinkman R, and Gottardo R. 2009. Merging Mixture Components for Cell Population Identification in Flow Cytometry. *Advances in Bioinformatics* **2009**: 1–12.

Finak G, Perez JM, Weng A, and Gottardo R. 2010. Optimizing transformations for automated, high throughput analysis of flow cytometry data. *BMC Bioinformatics* **11**: 546.

Friedman JH and Fisher NI. 1999. Bump hunting in high-dimensional data. *Statistics and Computing* **9**: 123–143.

Garg G, Tyler JR, Yang JHM, Cutler AJ, Downes K, Pekalski M, Bell GL, Nutland S, Peakman M, Todd JA, et al. 2012. Type 1 diabetes-associated IL2RA variation lowers IL-2 signaling and contributes to diminished CD4+CD25+ regulatory T cell function. *The Journal of Immunology* **188**: 4644–4653.

Giannoulatou E, Yau C, Colella S, Ragoussis J, and Holmes CC. 2008. GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. *. . . .*

Gómez-Lozano N, Estefanía E, Williams F, Halfpenny I, Middleton D, Solís R, and Vilches C. 2005. The silent KIR3DP1 gene (CD158c) is transcribed and might encode a secreted receptor in a minority of humans, in whom the KIR3DP1, KIR2DL4 and KIR3DL1/KIR3DS1 genes are duplicated. *European journal of immunology* **35**: 16–24.

Gonzalez-Galarza FF, Christmas S, Middleton D, and Jones AR. 2011. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic acids research* **39**: D913–9.

Gumperz JE, Barber LD, Valiante NM, Percival L, Phillips JH, Lanier LL, and Parham P. 1997. Conserved and variable residues within the Bw4 motif of HLA-B make separable contributions to recognition by the NKB1 killer cell-inhibitory receptor. *Journal of immunology (Baltimore, Md. : 1950)* **158**: 5237–5241.

Hahne F, Gopalakrishnan N, Khodabakhshi AH, Wong CJ, and Lee K. 2013. *flowStats: Statistical methods for the analysis of flow cytometry data*. R package version 3.20.3.

Hahne F, Khodabakhshi AH, Bashashati A, Wong CJ, Gascoyne RD, Weng AP, Seyfert-Margolis V, Bourcier K, Asare A, Lumley T, et al. 2009. Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A* **9999A**: NA–NA.

Hastie T, Tibshirani R, and Friedman J. 2009. *The Elements of Statistical Learning*. Data Mining, Inference, and Prediction, Second Edition. Springer.

Herzenberg LA, Tung J, Moore WA, Herzenberg LA, and Parks DR. 2006. Interpreting flow cytometry data: a guide for the perplexed. *Nature immunology* **7**: 681–685.

Holyst H and Rogers W. 2009. *flowFP: Fingerprinting for Flow Cytometry*. R package version 1.18.0.

Howson JMM, Walker NM, Clayton D, Todd JA, and Diabetes Genetics Consortium. 2009. Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A. *Diabetes, Obesity and Metabolism* **11**: 31–45.

Hu X, Kim H, Brennan PJ, Han B, Baecher-Allan CM, De Jager PL, Brenner MB, and Raychaudhuri S. 2013. Application of user-guided automated cytometric data analysis to large-scale immunoprofiling of invariant natural killer T cells. *Proc Natl Acad Sci USA* **110**: 19030–19035.

Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, Moffatt MF, Cookson WO, Trowsdale J, and Traherne JA. 2012. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. *Genome Research* **22**: 1845–1854.

Jing J, KOCH I, and Naito K. 2009. Polynomial histograms for multivariate density and mode estimation. *Scandinavian Journal of Statistics* .

Jobim M, Chagastelles P, Salim PH, Portela P, Wilson TJ, Curti AG, Jobim MR, João DA, Nardi NB, Tschiedel B, et al. 2010. Association of killer cell immunoglobulin-like receptors and human leukocyte antigen–C genotypes in South Brazilian with type 1 diabetes. *Human Immunology* **71**: 799–803.

Jones PW, Osipov A, and Rokhlin V. 2011. Randomized approximate nearest neighbors algorithm. *Proc Natl Acad Sci USA* **108**: 15679–15686.

Kent WJ. 2002. BLAT–the BLAST-like alignment tool. *Genome Research* **12**: 656–664.

Koreth J, Matsuoka Ki, Kim HT, McDonough SM, Bindra B, Alyea III EP, Armand P, Cutler C, Ho VT, Treister NS, et al. 2011. Interleukin-2 and Regulatory T Cells in Graft-versus-Host Disease. *New England Journal of Medicine* **365**: 2055–2066.

Körner C and Altfeld M. 2012. Role of KIR3DS1 in human diseases. *Frontiers in Immunology* **3**.

Kumasaka N, Fujisawa H, Hosono N, Okada Y, Takahashi A, Nakamura Y, Kubo M, and Kamatani N. 2011. PlatinumCNV: A Bayesian Gaussian mixture model for genotyping copy number polymorphisms using SNP array signal intensity data. *Genetic epidemiology* **35**: 831–844.

Leslie S, Donnelly P, and McVean G. 2008. A statistical method for predicting classical HLA alleles from SNP data. *American Journal of Human Genetics* **82**: 48–56.

Liao W, Lin JX, and Leonard WJ. 2013. Interleukin-2 at the crossroads of effector responses, tolerance, and immunotherapy. *Immunity* **38**: 13–25.

Lo K, Brinkman RR, and Gottardo R. 2008. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A* **73**: 321–332.

Lo K, Hahne F, Brinkman RR, and Gottardo R. 2009. flowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics* **10**: 145.

Long SA, Cerosaletti K, Bollyky PL, Tatum M, Shilling H, Zhang S, Zhang ZY, Pihoker C, Sanda S, Greenbaum C, et al. 2010. Defects in IL-2R Signaling Contribute to Diminished Maintenance of FOXP3 Expression in CD4+CD25+ Regulatory T-Cells of Type 1 Diabetic Subjects. *Diabetes* **59**: 407–415.

Long SA, Cerosaletti K, Wan JY, Ho JC, Tatum M, Wei S, Shilling HG, and Buckner JH. 2011. An autoimmune-associated variant in PTPN2 reveals an impairment of IL-2R signaling in CD4(+) T cells. *Genes and immunity* **12**: 116–125.

Long SA, Rieck M, Sanda S, Bollyky JB, Samuels PL, Goland R, Ahmann A, Rabinovitch A, Aggarwal S, Phippard D, et al. 2012. Rapamycin/IL-2 combination therapy in patients with type 1 diabetes augments Tregs yet transiently impairs β-cell function. *Diabetes* **61**: 2340–2348.

Lowe CE, Cooper JD, Brusko T, Walker NM, Smyth DJ, Bailey R, Bourget K, Plagnol V, Field S, Atkinson M, et al. 2007. Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nature Genetics* **39**: 1074–1082.

Lugli E, Roederer M, and Cossarizza A. 2010. Data analysis in flow cytometry: The future just started. *Cytometry Part A* **77A**: 705–713.

Lumley T. 2012. *mitools: Tools for multiple imputation of missing data.* R package version 2.1.

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pp. Vol. I: Statistics, pp. 281–297. Univ. California Press, Berkeley, Calif.

Maechler M, Rousseeuw P, Struyf A, Hubert M, and Hornik K. 2014. *cluster: Cluster Analysis Basics and Extensions.* R package version 1.15.2 — For new features, see the 'Changelog' file (in the package source).

Maecker HT, McCoy JP, FOCIS Human Immunophenotyping Consortium, Amos M, Elliott J, Gaigalas A, Wang L, Aranda R, Banchereau J, Boshoff C, et al. 2010. A model for harmonizing flow cytometry in clinical trials. *Nature immunology* **11**: 975–978.

Maecker HT, McCoy JP, and Nussenblatt R. 2012. Standardizing immunophenotyping for the Human Immunology Project. *Nature Reviews Immunology* **12**: 191–200.

Maecker HT, Rinfret A, D'Souza P, Darden J, Roig E, Landry C, Hayes P, Birungi J, Anzala O, Garcia M, et al. 2005. Standardization of cytokine flow cytometry assays. *BMC Immunology* **6**: 13.

Maecker HT and Trotter J. 2006. Flow cytometry controls, instrument setup, and the determination of positivity. *Cytometry Part A* **69**: 1037–1042.

Martin MP, Gao X, Lee JH, Nelson GW, Detels R, Goedert JJ, Buchbinder S, Hoots K, Vlahov D, Trowsdale J, et al. 2002. Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nature Genetics* **31**: 429–434.

Martin MP, Qi Y, Gao X, Yamada E, Martin JN, Pereyra F, Colombo S, Brown EE, Shupert WL, Phair J, et al. 2007. Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. *Nature Genetics* **39**: 733–740.

McLachlan G and Peel D. 2004. *Finite Mixture Models.* John Wiley & Sons.

Mehers KL, Long AE, van der Slik AR, Aitken RJ, Nathwani V, Wong FS, Bain S, Gill G, Roep BO, Bingley PJ, et al. 2011. An increased frequency of NK cell receptor and HLA-C group 1 combinations in early-onset type 1 diabetes. *Diabetologia* **54**: 3062–3070.

Mevik BH, Wehrens R, and Liland KH. 2013. *pls: Partial Least Squares and Principal Component regression.* R package version 2.4-3.

Middleton D, Halfpenny I, Meenagh A, Williams F, Sivula J, and Tuomilehto-Wolf E. 2006. Investigation of KIR Gene Frequencies in Type 1 Diabetes Mellitus. *Human Immunology* **67**: 986–990.

Mogami S, Hasegawa G, Nakayama I, Asano M, Hosoda H, Kadono M, Fukui M, Kitagawa Y, Nakano K, Ohta M, et al. 2007. Killer cell immunoglobulin-like receptor genotypes in Japanese patients with type 1 diabetes. *Tissue Antigens* **70**: 506–510.

Murphy RF and Chused TM. 1984. A proposal for a flow cytometric data file standard. *Cytometry Part A* **5**: 553–555.

Naumann U, Luta G, and Wand MP. 2010. The curvHDR method for gating flow cytometry samples. *BMC Bioinformatics* **11**: 44.

Nejentsev S, Howson JMM, Walker NM, Szeszko J, Field SF, Stevens HE, Reynolds P, Hardy M, King E, Masters J, et al. 2007. Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature* **450**: 887–892.

Nerup J, Platz P, Andersen OO, Christy M, Lyngsoe J, Poulsen JE, Ryder LP, Nielsen LS, Thomsen M, and Svejgaard A. 1974. HL-A antigens and diabetes mellitus. *The Lancet* **2**: 864–866.

Nikitina-Zake L, Rajalingham R, Rumba I, and Sanjeevi CB. 2004. Killer cell immunoglobulin-like receptor genes in Latvian patients with type 1 diabetes mellitus and healthy controls. *Annals of the New York Academy of Sciences* **1037**: 161–169.

Nikula T, West A, Katajamaa M, Lonnberg T, Sara R, Aittokallio T, Nevalainen O, and Lahesmaa R. 2005. A human ImmunoChip cDNA microarray provides a comprehensive tool to study immune responses. *Journal of immunological methods* **303**: 122–134.

Norman PJ, Abi-Rached L, Gendzekhadze K, Hammond JA, Moesta AK, Sharma D, Graef T, McQueen KL, Guethlein LA, Carrington CVF, et al. 2009. Meiotic recombination generates rich diversity in NK cell receptor genes, alleles, and haplotypes. *Genome Research* **19**: 757–769.

O'Gorman MRG and Thomas J. 1999. Isotype controls—time to let go? *Cytometry Part A* **38**: 78–80.

Parham P and Moffett A. 2013. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nature Reviews Immunology* **13**: 133–144.

Park Y, Choi H, Park H, Park S, Yoo EK, Kim D, and Sanjeevi CB. 2006. Predominance of the group A killer Ig-like receptor haplotypes in Korean patients with T1D. *Annals of the New York Academy of Sciences* **1079**: 240–250.

Parks DR, Roederer M, and Moore WA. 2006. A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. *Cytometry Part A* **69**: 541–551.

Patterson CC, Dahlquist GG, Gyürüs E, Green A, and Soltész G. 2009. Incidence trends for childhood type 1 diabetes in Europe during 1989–2003 and predicted new cases 2005–20: a multicentre prospective registration study. *The Lancet* **373**: 2027–2033.

Pekalski ML, Ferreira RC, Coulson RMR, Cutler AJ, Guo H, Smyth DJ, Downes K, Dendrou CA, Castro Dopico X, Esposito L, et al. 2013. Postthymic expansion in human CD4 naive T cells defined by expression of functional high-affinity IL-2 receptors. *The Journal of Immunology* **190**: 2554–2566.

Pelak K, Need AC, Fellay J, Shianna KV, Feng S, Urban TJ, Ge D, De Luca A, Martinez-Picado J, Wolinsky SM, et al. 2011. Copy Number Variation of KIR Genes Influences HIV-1 Control. *PLOS Biology* **9**: e1001208.

Perfetto SP, Chattopadhyay PK, and Roederer M. 2004. Innovation: Seventeen-colour flow cytometry: unravelling the immune system. *Nature Reviews Immunology* **4**: 648–655.

Permutt MA, Chirgwin J, Rotwein P, and Giddings S. 1984. Insulin gene structure and function: a review of studies using recombinant DNA methodology. *Diabetes Care* **7**: 386–394.

Pinheiro J, Bates D, DebRoy S, Sarkar D, and R Core Team. 2014. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-117.

Plagnol V, Cooper JD, Todd JA, and Clayton DG. 2007. A Method to Address Differential Bias in Genotyping in Large-Scale Association Studies. *PLoS Genetics* **3**: e74.

Pontikos N. 2013. *flowBeads: Analysis of flow bead data*. R package version 1.0.0.

Pontikos N, Smyth DJ, Schuilenburg H, Howson JMM, Walker NM, Burren OS, Guo H, Onengut-Gumuscu S, Chen WM, Concannon P, et al. 2014. A hybrid qPCR/SNP array approach allows cost efficient assessment of KIR gene copy numbers in large samples. *BMC Genomics* **15**: 274.

Poretsky L. 2010. *Principles of Diabetes Mellitus*. Springer.

Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, et al. 2009. Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences* **106**: 8519–8524.

Qian Y, Wei C, Eun-Hyung Lee F, Campbell J, Halliley J, Lee JA, Cai J, Kong YM, Sadat E, Thomson E, et al. 2010. Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. *Cytometry Part A* **78B**: S69–S82.

Qiu P, Simonds EF, Bendall SC, Gibbs KD, Bruggner RV, Linderman MD, Sachs K, Nolan GP, and Plevritis SK. 2011. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature Publishing Group* **29**: 886–891.

Ramos-Lopez E, Scholten F, Aminkeng F, Wild C, Kalhes H, Seidl C, Tonn T, Van der Auwera B, and Badenhoop K. 2009. Association of KIR2DL2 polymorphism rs2756923 with type 1 diabetes and preliminary evidence for lack of inhibition through HLA-C1 ligand binding. *Tissue Antigens* **73**: 599–603.

Ripley B. 2014. *tree: Classification and regression trees*. R package version 1.0-35.

Risch N. 1987. Assessing the role of HLA-linked and unlinked determinants of disease. *American Journal of Human Genetics* **40**: 1–14.

Robinson J, Mistry K, McWilliam H, Lopez R, and Marsh SGE. 2010. IPD–the Immuno Polymorphism Database. *Nucleic acids research* **38**: D863–9.

Roederer M. 2001. Spectral compensation for flow cytometry: visualization artifacts, limitations, and caveats. *Cytometry Part A* **45**: 194–205.

Roederer M, Moore W, Treister A, Hardy RR, and Herzenberg LA. 2001. Probability binning comparison: a metric for quantitating multivariate distribution differences. *Cytometry Part A* **45**: 47–55.

Roederer M, Nozzi JL, and Nason MC. 2011. SPICE: exploration and analysis of post-cytometric complex multivariate datasets. *Cytometry Part A* **79**: 167–174.

Rogers WT and Holyst HA. 2009. FlowFP: A Bioconductor Package for Fingerprinting Flow Cytometric Data. *Advances in Bioinformatics* p. 193947.

Rogers WT, Moser AR, Holyst HA, Bantly A, Mohler ER, Scangas G, and Moore JS. 2008. Cytometric fingerprinting: Quantitative characterization of multivariate distributions . *Cytometry Part A* **73**: 430–441.

Saadoun D, Rosenzwajg M, Joly F, Six A, Carrat F, Thibault V, Sene D, Cacoub P, and Klatzmann D. 2011. Regulatory T-Cell Responses to Low-Dose Interleukin-2 in HCV-Induced Vasculitis. *New England Journal of Medicine* **365**: 2067–2077.

Santin I, de Nanclares GP, Calvo B, Gaafar A, Castaño L, and Bilbao JR. 2006. Killer Cell Immunoglobulin-Like Receptor (KIR) Genes in the Basque Population: Association Study of KIR Gene Contents With Type 1 Diabetes Mellitus. *Human Immunology* **67**: 118–124.

Schwartz A, Repollet E, Vogt R, and Gratama J. 1996. Standardizing flow cytometry: Construction of a standardized fluorescence calibration plot using matching spectral calibrators. *Cytometry Part A* **26**: 22–31.

Seamer LC, Bagwell CB, Barden L, Redelman D, Salzman GC, Wood J, and Murphy RF. 1997. Proposed new data file standard for flow cytometry, version FCS 3.0. *Cytometry Part A* **28**: 118–122.

Shapiro HM. 2003. *Practical flow cytometry.* John Wiley and Sons.

Shastry A, Sedimbi SK, Rajalingam R, Nikitina-Zake L, Rumba I, Wigzell H, and Sanjeevi CB. 2008. Combination of KIR 2DL2 and HLA-C1 (Asn 80) confers susceptibility to type 1 diabetes in Latvians. *International Journal of Immunogenetics* **35**: 439–446.

Siebert JC, Wang L, Haley DP, Romer A, Zheng B, Munsil W, Gregory KW, and Walker EB. 2010. Exhaustive expansion: A novel technique for analyzing complex data generated by higher-order polychromatic flow cytometry experiments. *Journal of Translational Medicine* **8**: –.

Singal DP and Blajchman MA. 1973. Histocompatibility (HL-A) Antigens, Lymphocytotoxic Antibodies and Tissue Antibodies in Patients with Diabetes-Mellitus. *Diabetes* **22**: 429–432.

Single RM, Martin MP, Gao X, Meyer D, Yeager M, Kidd JR, Kidd KK, and Carrington M. 2007. Global diversity and evidence for coevolution of KIR and HLA. *Nature Genetics* **39**: 1114–1119.

van der Slik AR, Alizadeh BZ, Koeleman B, Roep BO, and Giphart MJ. 2007. Modelling KIR–HLA genotype disparities in type 1 diabetes. *Tissue Antigens* **69**: 101–105.

van der Slik AR, Koeleman BPC, Verduijn W, Bruining GJ, Roep BO, and Giphart MJ. 2003. KIR in Type 1 Diabetes: Disparate Distribution of Activating and Inhibitory Natural Killer Cell Receptors in Patients Versus HLA-Matched Control Subjects. *Diabetes* **52**: 2639–2642.

Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tîrgovişte C, Widmer B, Dunger DB, et al. 2006. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nature Genetics* **38**: 617–619.

Smyth GK and Speed T. 2003. Normalization of cDNA microarray data. *Methods (San Diego, Calif.)* **31**: 265–273.

Snow C. 2004. Flow cytometer electronics. *Cytometry Part A* **57**: 63–69.

Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FAS, Zhernakova A, Hinks A, et al. 2010. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics* **42**: 508–514.

Sugár IP and Sealfon SC. 2010. Misty Mountain clustering: application to fast unsupervised flow cytometry gating. *BMC Bioinformatics* **11**: 502.

Tibshirani R, Walther G, and Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**: 411–423.

Todd JA. 2010. Etiology of type 1 diabetes. *Immunity* **32**: 457–467.

van Buuren S and Groothuis-Oudshoorn K. 2011. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software* **45**: 1–67.

Vivian JP, Duncan RC, Berry R, O'Connor GM, Reid HH, Beddoe T, Gras S, Saunders PM, Olshina MA, Widjaja JML, et al. 2011. Killer cell immunoglobulin-like receptor 3DL1-mediated recognition of human leukocyte antigen B. *Nature* **479**: 401–405.

Walther G, Zimmerman N, Moore W, Parks D, Meehan S, Belitskaya I, Pan J, and Herzenberg L. 2009. Automatic Clustering of Flow Cytometry Data with Density-Based Merging. *Advances in Bioinformatics* **2009**: 1–7.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661–678.

Wittes J. 2009. On looking at subgroups. *Circulation* **119**: 912–915.

Yang Q, Khoury MJ, Sun F, and Flanders WD. 1999. Case-only design to measure gene-gene interaction. *Epidemiology (Cambridge, Mass.)* **10**: 167–170.

Young D, Hunter D, Chauveau D, and Benaglia T. 2009a. mixtools: An R Package for Analyzing Mixture Models. *Journal of Statistical Software* **32**.

Young D, Hunter D, Chauveau D, and Benaglia T. 2009b. mixtools: An R Package for Analyzing Mixture Models. *Journal of Statistical Software* **32**.

Zare H, Shooshtari P, Gupta A, and Brinkman RR. 2010. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics* **11**: 403.

Zhi D, Sun C, Sedimbi SK, Luo F, Shen S, and Sanjeevi CB. 2011. Killer cell immunoglobulin-like receptor along with HLA-C ligand genes are associated with type 1 diabetes in Chinese Han population. *Diabetes/Metabolism Research and Reviews* **27**: 872–877.