Profiling the Human Dendritic Cell System by Multiplexed Gene Expression Analysis

Kile James Green

Thesis submitted to the Institute of Cellular Medicine, Newcastle University for the degree of Doctor of Philosophy

June 2018

Abstract

Dendritic cells (DCs) are rare immune cell populations that play a significant role in phagocytosis, antigen presentation and pathogen response. While extensive research has produced robust models of murine DC biology and development, the human dendritic cell system remains relatively unexplored due to the inherent difficulty in obtaining the required cell numbers for analysis. Furthermore, with the widespread expansion of gene expression technologies now available to researchers, much of the historical knowledge of human DC biology has begun to be revised and revisited with a fresh approach. In this thesis, flow cytometry, multiplexed hybridisation-based expression analysis and microarray experiments have been used in combination with cutting-edge single-cell transcriptomics to reveal the nature of mature dendritic cells subsets, their relation to monocytes and their developmental heterogeneity.
Initially, a previously published microarray dataset (GEO:GSE35457) was interrogated to identify robust mononuclear cell signatures applicable across experimental platforms and tissue origin to address the fickle nature of currently known surface markers of DC subtypes. This dataset was then used as a surrogate in a feasibility study to determine the efficacy of the immune-focused nCounter platform with validation of the *in-silico* results confirmed using the NanoString platform.

NanoString technology was implemented to investigate the correlation of *ex-vivo* and *in-vitro* generated DC populations, culminating in the discovery of a 'universal culture effect' influencing global expression in cultured cells. Removal of this signature revealed the underlying equivalence of the cells.

Finally, single-cell transcriptomics was employed to expose heterogeneity in pre-DCs, utilising the Illumina and NanoString-derived cell signatures to highlight the early lineage commitment of pre-DC cells.

By utilising multiple high-dimensional analysis platforms and covering *ex-vivo* blood and skin, as well as DCs generated from CD34+ bone marrow progenitor cells, novel insights into the functional roles, expression patterns and molecular signatures of DC subsets have been revealed.

Dedication

For my dear friend, Victor Tudorica.

Acknowledgements

My absolute gratitude is given to Matt Collin and Venetia Bigley who after taking me on as a Master's student in 2013, provided continued support throughout my position as a technician and PhD student in their group. I know that without their faith and input I would not be where I am now. Particular thanks are also given to Graham Smith, who after kindly stepping into the role of PhD supervisor early in my second year has truly been a fantastic mentor. His continued support and suggestions have been invaluable.

The Human Dendritic Cell group deserves great thanks for their support both in the lab and outside of it. Urszula Cytlak-Chaudhuri is mentioned multiple times in this thesis thanks to her tremendous contributions with cell culture and flow cytometry, alongside the other members of HuDC with FACS and flow. Rachel Queen was also invaluable, helping me understand single cell command-line coding. I would like to mention the continued friendship and support of Paul Milne, Rachel Dickinson and Sarah Pagan who have been there for every step of my PhD as well as members past and present that made the last 4 years thoroughly enjoyable.

Xiao Wang, Alison Tyson-Capper and Michael Drinnan along with the other members of the PGR committee have been wonderful, offering support, friendship and inspiration to me both as a PhD student and throughout my role as a school representative. Thank you all for being there throughout.

None of this work would have been possible without the help and contributions of the patients, volunteers and clinical staff who provided the research material as well as everyone at Bright Red who provided my funding throughout my technical post and majority of my PhD, followed by Muzlifah Haniffa who gave me my current position within the Haniffa Lab and has generously given me time and support to complete my PhD during the transition.

Finally, I would like to thank all of my family and friends who have helped me though this work, particularly Anastasia Resteu whose support, kindness and understanding during my write-up was fantastic.

Candidate Declaration

All of the work contained within this thesis is the sole work of the candidate, with the exception of the material detailed below:

Chapter 3:

Illumina expression data GSE35457 was extracted from online data resource NCBI-GEO repository by the candidate and analysed solely by the candidate. The dataset was originally published in Haniffa *et al*, Immunity 2012. Details of sample processing for this dataset can be found in the initial publication.

Illumina expression data GSE65128 was extracted from online data resource NCBI-GEO repository by the candidate and analysed solely by the candidate. The dataset was originally published in Lee *et al*, J Exp Med 2015. Details of sample processing for this dataset can be found in the initial publication.

Human blood samples for NanoString analysis were isolated and pre-processed by the candidate and members of the HuDC lab. Flow Cytometric analysis and FACS was performed in part by the candidate and various members of the Human Dendritic Cell Lab. All NanoString work-up and analysis, from RNA extraction or direct-cell lysis, RNA pre-processing, probe-binding, running the machine, collecting the data and subsequent analysis was performed solely by the candidate.

Chapter 4:

Cultured cells were supplied by Dr Urszula Cytlak-Chaudhuri. Bone marrow samples were pre-processed by both the candidate and members of the Human Dendritic Cell Lab.

Human blood samples for NanoString analysis in this chapter were isolated and pre-processed by the candidate and HuDC members. Flow Cytometric analysis and FACS was performed in part by the candidate and various members of the Human Dendritic Cell Lab. All NanoString work-up and analysis of blood and cultured cells, from RNA extraction or direct-cell lysis, RNA pre-processing, probe-binding, running the machine, collecting the data and subsequent analysis was performed solely by the candidate.

Chapter 5:

FACS was performed in part by the candidate and various members of the Human Dendritic Cell Lab.

Cultured cells were supplied by Dr Urszula Cytlak-Chaudhuri. Bone marrow samples were pre-processed by both the candidate and members of the Human Dendritic Cell Lab.

Sample pre-processing for single cell analysis was performed at Oxford Genomics.

Initial command-line pre-processing scripts for single-cell analysis were adapted from those generated by Dr Rachel Queen. All data processing and analysis was performed solely by the candidate.

All figures contained in this thesis are of original design by the candidate unless otherwise stated in the figure legends.

Table of Contents

List of Abbreviations

APCs	Antigen Presenting Cells
BM	Bone Marrow
CD	Cluster of Differentiation
cDC1	Classical Dendritic Cell 1 (CD141+)
cDC2	Classical Dendritic Cell 2 (CD1c+)
CDP	Common Dendritic Cell Progenitor
CMP	Counts Per Million
ComBat	Combining Batches of microarray data
D-PBS	Dulbecco's Phosphate-Buffered Saline
DCML	DC, Monocyte, B, NK Cell Deficiency
DCs	Dendritic Cells
DMSO	Dimethyl Sulphoxide
EDTA	Ethylenediaminetetraacetic Acid
ERCC	External RNA Controls Consortium
FACS	Fluorescence-Activated Cell Sorting
FCS	Fetal Calf Serum
FDR	False Discovery Rate
FF	Fresh Frozen
FFPE	Formalin-Fixed Paraffin-Embedded
FPKM	Fragments Per Kilobase Million
FSc	Forward-Scatter
G-CSF	Granulocyte colony stimulating factor
GSEA	Gene Set Enrichment Analysis
HSCs	Haematopoietic stem cells
iDCs	Interstitial Dendritic Cell
infDC	Inflammatory Dendritic Cell
KLR	Killer Cell Lectin Receptor
LC	Langerhans Cells
LDA	Linear Discriminant Analysis
LPS	Lipopolysaccharide
mDC	Myeloid Dendritic Cell
MDP	Monocyte/Macrophage, DC Precursor
MHC	Major Histocompatibility Complex
MLP	Multi-Lymphoid Progenitor
moDC	Monocyte-Derived Dendritic Cell
NK	Natural Killer (Cell)
NOD	nucleotide-binding oligomerisation
PB	Peripheral Blood
PBC	Primary Biliary Cirrhosis
PBMC	Peripheral Blood Mononuclear Cells
PBS	Phsophate Buffered Saline
PCA	Principal Component Analysis
pDC	Plasmacytoid Dendritic Cell

PID	Primary Immunodeficiency
RBC	Red Blood Cell
RLT	RNA Lysis Buffer
RNA	Ribonucleic Acid
RNAseq	RNA-Sequencing
RPMI	Roswell Park Memorial Institute (Media)
RT-PCR	Real Time Polymerase Chain Reaction
scRNAseq	Single-Cell RNA-Sequencing
SSc	Side-Scatter
t-SNE	t-Distributed Stochastic Neighbour Embedding
TiGER	Tissue-specific Gene Expression and Regulation database
TLR	Toll-Like Receptor
TMP	Transcripts Per Million
UMI	Unique Molecular Identifier
VST	Variance stabilizing normalisation

List of Tables

List of Figures

RESEARCH QUESTIONS AND RATIONALE

Each research chapter (chapters 3-5) contained in this thesis aimed to address a fundamental research question driving the project. Each of these encompassed three smaller topics that were to be addressed in the chapter. Below is a breakdown of each chapter's research questions and a summary of the scientific rationale behind posing the questions.

CHAPTER 3: CODESET DESIGN AND DIMENSIONALITY REDUCTION FOR DENDRITIC CELL SUBSET ANALYSIS

Primary research question:
Can RNA expression analysis be used to distinguish and identify human DC and monocyte subsets?

Sub-topic questions:

1. Can human blood dendritic cells and monocytes be classified by their RNA signatures?

2. Can a focused panel of immune genes be used to identify common blood dendritic cell and monocyte subsets?

3. How many genes are required to maintain dendritic cell and monocyte subset classification?

Rationale:

Dendritic cells are extremely potent, but exceedingly rare cell subsets. They are particularly difficult to isolate *ex-vivo* due to the fickle nature of surface markers and their susceptibility to environmental and mechanical stresses. Unfortunately, relatively few surface marker proteins can typically be used to identify these rare DC populations and expression of these markers can be perturbed under inflammatory conditions. Moving from surface protein expression to RNA transcriptome analysis provides a much wider spectrum of parameters for testing, including the investigation of intracellular markers such as transcription factors, detailed functional pathway analysis and cell cycle progression.

This chapter aimed to determine if RNA transcriptome profiling could uncover greater distinguishing features between closely related DC and monocyte populations from the blood and skin, whilst also providing a robust pipeline of analysis to take forwards into the later stages of the project, incorporating novel gene reduction methods and deconvolution of tissue-specific cell signatures. This analysis would reveal robust, biologically relevant DC and monocyte markers, capable of distinguishing these subsets from one another, whilst simultaneously allowing for cross-tissue comparisons of equivalent mononuclear cell subsets with the removal of an over-arching tissue-specific gene signature from the dataset, with minimal loss of data.

Machine learning techniques were implemented to produce and interrogate a novel gene signature capable of distinguishing DC and monocyte subsets, both within the dataset used for analysis, as well as in an independently generated dataset of similar mononuclear cell subsets. Generation of such a verified signature provided a backbone for confirmation of cell types in later chapters of the thesis, but could also be applied to external projects for identification of mononuclear cells by methods other than RNA profiling, particularly in cases where traditional markers may not be expressed or available for assessment for example, on cells grown in culture conditions, through cross-tissue comparisons or under inflammation.

Minimisation of the resulting gene signature was performed to determine the fundamental genes involved in DC and monocyte differentiation. While this process revealed potent discriminators of mononuclear cells, the process had a wider applicability within DC research by providing gene combinations that could be used in FACS or flow cytometry to distinguish there cells using far fewer antibodies than currently required, opening up spare channels for other antibodies to be used and reducing the costs of the technique. Along with analyzing the RNA transcriptome of DC and monocyte subsets by Illumina BeadArray technology to uncover potential subset markers, the similarity of these subsets was assessed on a focused immunocentric panel of RNA probes using NanoString technology.

Robust data acquisition is an integral part of modern transcriptome profiling and thus, minimizing potential sources of bias should be a major consideration in research methodology. Through this change in RNA analysis platform, the potential for widespread alterations in cell numbers, expression of markers and mechanical and environmental stresses related to the long purification, isolation, pre-processing and amplification stages necessary for microarray analysis would be drastically diminished. NanoString technology required no amplification and very few preparation steps, could be performed directly from crude cell lysates after FACS and performed in one day without the introduction of freeze-thaw cycles that can drastically degrade the available RNA, thereby mitigating the altering effects of sample preparation and analysis on the cells prior to data capture that would otherwise skew the results.

CHAPTER 4: IN-VITRO DENDRITIC CELL SUBSET CLASSIFICATION

Primary research question:
Can transcriptomic signatures aid in the identification and validation of cells generated *in-vitro*?

Sub-topic questions:

1. Can phenotypically equivalent human dendritic cells be generated *in-vitro*?

2. Are there any culture-specific expression patterns identifiable *in-vitro*?

3. Do *in-vitro* derived DCs share a similar transcriptome profile with primary human DCs?

Rationale:

Culture models are frequently incorporated into dendritic cell research due to analytical limitations associated primarily by the relative scarcity of DCs and their environmentally sensitive and potentially inconsistent expression patterns. These issues are of particular concern to researchers studying early DC development, where precursor populations and intermediary populations can be near impossible to obtain at quantities effective to research. This can profoundly restrict the assays and techniques that can be applied in the investigation of primary dendritic cell subsets. Furthermore, looking beyond basic research and into translational medicine, DCs generated *in-vitro* already offer potential therapeutic options in the treatment of cancers, but greater applicability of *in-vitro* DCs may be possible with improvements to the yield, consistency and correlation of these cells to their primary counterparts. For either application, the *in-vitro* derived DCs must be functionally identical to those taken directly from the blood, well beyond the basic appearance and surface marker expression patterns associated with each subset.

This chapter addresses the issue of comparability between blood and cultured dendritic cell and monocyte subsets through comparative RNA transcriptome analysis, uncovering the distinct and insightful transcriptional changes affecting *in-vitro* generated cell populations and providing a novel basis to identify and isolate these obscuring transcriptional changes to reveal the underlying conservation of each unique dendritic cell subset's phenotypic and functional features.

By flow cytometry and FACS, our culture system is capable of producing phenotypically similar DCs and monocytes from CD34+ bone marrow. In order to determine if the similarity runs more than ‘skin-deep’, NanoString RNA analysis was performed on the samples to reveal any potential transcriptional changes affected by the culture conditions. By unraveling the gene specific and functional or pathway-level alterations to expression in culture conditions, it may be possible to identify potential changes to the methodology and conditions to better recapitulate *bona-fide* DCs. Once any potential differences were identified, they would likely overshadow the unique DC subset signatures capable of matching the subsets across conditions. The final section of this chapter addressed the issue of extracting a backbone cell signature from a dataset with a dominating conditional signature through a novel process similar to that used to deconvolute skin and blood DC subsets in chapter 3. This approach could be applicable to a range of projects where conservation of expression is to be assessed across a range of conditions, particularly in instances where those conditions are likely to strongly influence global expression.

CHAPTER 5: A SINGLE-CELL APPROACH TO DENDRITIC CELL DIFFERENTIATION

Primary research question:
Can transcriptomic analysis identify DC lineage priming in progenitor cells and precursors?

Sub-topic questions:
1. Can single-cell RNA-sequencing be used to investigate dendritic cell precursors?

2. Are DC-precursors skewed towards mature DC signature expression at the single cell level?

3. Does the *in-vitro* development assay on pre-DC populations correlate with transcriptome-level expression patterns?

Rationale:

Chapter 5 incorporated cutting-edge single cell transcriptomic analysis to gain insight into DC development and lineage priming of progenitor cells. Recent publications have highlighted a shift in haematopoietic development research, with a focus on 'early priming' of haematopoietic stem cells, such that most early progenitor populations, including the granulocyte macrophage dendritic cell precursor (GMDP), macrophage DC progenitor (MDP), and common DC progenitor (CDP) populations are composed of heterogeneous, uni-primed, phenotypically similar cells with restricted mature cell potential, compared to the earlier linear bifurcation model of haematopoietic development, where progenitor cells types represented a single population of multi-potent cells. Determining if progenitor cells are early primed or multi-potent is a critical aspect of developmental research. Technological advances in single-cell sequencing have recently opened up this avenue of research. One of the major draws of single cell transcriptomics is its ability to reveal the diversity and heterogeneity of cells that would be obscured using bulk-level techniques such as qPCR, NanoString Technology, microarray analysis or conventional bulk-RNA sequencing.

Single cell transcriptomics is a continuously growing field and at the time of writing, was so novel that the extent of what could be explored using this technique was relatively unknown. The single cell analysis contained in this thesis represents one of the first uses of the technology at Newcastle University and was one of the first examples of its applicability to DC progenitor development. With this in mind, one of the first sub-questions to be addressed was whether adequate, good quality, viable cells could be captured by FACS and processed through a SmartSeq2 plate-based single cell pipeline to provide enough data for analysis.

By investigating the cell populations found within a conventional progenitor cell FACS gate, their heterogeneity could be revealed and attributed to familiar mature cell populations to determine if there was any intra-population variance suggestive of early cell priming, or if the cells shared a single un-primed, multi-potent phenotype.

Identifying potentially primed progenitor cells by transcriptome comparison to mature cell populations would not prove that these cells were destined to become their mature cell counterparts, and thus an *in-vitro* development assay was developed to isolate and grow transcriptionally primed progenitors to determine if their culture output recapitulated the transcriptional priming upon development and maturation. If primed progenitors could be isolated and grown in culture, they could be used to generate *bona-fide* DCs for research and medical applications in the future.

Chapter 1: INTRODUCTION

1.1 STEADY-STATE AND INFLAMMATORY MONONUCLEAR CELL SUBSETS

1.1.1 Peripheral Blood Dendritic Cells (DCs)

Dendritic cells (DCs) are irregularly shaped haematopoietic cells with branched dendrites. First identified by Ralph Steinman in 1973 as a previously unknown form of 'accessory cell', their unusual appearance and movements distinguished them from typical macrophages and their tree-like processes led Steinman to coin the term 'dendritic cell' (Rockefeller Institute, 2014).

While initially identified in spleen tissue, an expansion of dendritic cell research has led to their identification in all lymphoid and most non-lymphoid tissues, as well as the discovery of many immature DC subsets in the blood (Nairn, 2002). It has been established that peripheral blood mononuclear cells comprise of approximately 1% DCs and DC precursors (Nairn, 2002; Shurin and Salter, 2009).

Dendritic cells, described as 'sentinels' or professional 'antigen presenting cells' (APCs), have a number of characteristic functions [Figure 1.1]. These motile cells are particularly significant in T-cell mediated immunity and are active in both adaptive and innate immune responses, their prototypic function being to activate and prime naïve T-cells (Mak and Saunders, 2005; Mellman and Steinman, 2001; R N Germain and Margulies, 1993).

Dendritic cells are defined by both their typical locale and the presence or absence of phenotypic cell surface markers, distinguishable from other blood cell types by flow cytometry due to their lack of lineage-specific surface markers; CD3 on T-cells, CD14 on monocytes, CD19 and CD20 on B-cells, and CD56 on natural killer (NK) cells (Timmerman and Levy, 1999).

Surface marker expression on DCs can be highly variable, although this does not necessarily represent distinct subtypes, but may imply that DCs can undergo numerous developmental stages involving maturation and migration. DCs do express a number of molecules known to be involved in T-cell interactions including MHC class I and class II, adhesion molecules including CD11a, CD11b, CD11c and CD54, as well as co-stimulatory molecules such as CD40. While there is no known global, robust DC surface marker, CD80 and CD86 are classical co-stimulatory molecules expressed on the surface of DCs that are conventionally used as markers of these cell types. CD86 is used as a marker of early DC maturation and CD80 is expressed highly on mature DCs, but it is not exclusive to this lineage (Timmerman and Levy, 1999; Weider, 2003). CD83 is also mostly restricted to DC subsets, but is present on activated B-cells as well.

Experiments investigating the development pathways of DC subsets have indicated that four main groups exist; Langerhans cells (LC), myeloid DCs, monocyte-derived DCs and lymphoid DCs (Austyn, 1998). It is widely accepted that all well documented DC populations, with the exception of Langerhans cells and microglia are developed from bone-marrow derived, blood-borne precursors (Bigley et al., 2011; Collin et al., 2013). In the revised model of haematopoiesis entities such as the common dendritic cell progenitor (CDP) are transient, existing of phenotypically related cells with single potential. In this instance, lymphoid-primed multipotent progenitors are at the apex of all lymphoid and myeloid cell lineages (Collin and Bigley, 2018). Where dendritic cells are derived from different regions of the CD34+ progenitor compartment, by splitting on CD38 and CD45RA expression, they appear as transcriptionally homogeneous cells once developed.

Myeloid DCs are considered ‘classical’ or ‘conventional’ DCs (cDCs), originating from more myeloid-primed CD34+ progenitor cells and driven to become DCs in the presence of GM-CSF and TNF-α ± IL-4. These DCs, once matured, are known as interstitial DCs (iDCs). They have the functional capability of activating and priming naïve CD4+ and CD8+ T-cells and can induce naïve B-cell differentiation to plasma cells and activate naïve CD4+ and CD8+ T-cells.

Lymphoid lineage DCs originate from more lymphoid-primed CD34+ precursors and mature via IL-3 exposure into plasmacytoid DCs (pDCs). Such pDCs can produce IFN-α and are found in close proximity to T-cells in lymphoid tissue T-cell compartments (Weider, 2003).

Human myeloid DCs (mDCs) express CD11c, in contrast with lymphoid DCs, which are CD11c- (although murine pDCs are CD11c+). Human peripheral blood contains two 'classical' myeloid DC subtypes. These both express typical myeloid surface markers, including CD11b, CD11c, CD13 and CD33, but differ in their expression of CD1c and CD141 amongst other surface markers, with cDC1 cells having lower average expression of CD11b and CD11c than cDC2. The CD1c+ myeloid DC subset is also known as cDC2 and contributes to 0.6% of healthy peripheral blood mononuclear cells (PBMC), while the CD141+, cDC1 subset makes up <0.05% of the PBMC population (Sato and Fujita, 2007). Flow cytometry experiments conducted by the Human Dendritic Cell Laboratory typically show approximately 1,000 CD141+ cDC1s, 4,000 CD11c+ cDC2s and 5,000 CD123+ pDCs per milliliter of blood, equating to 1% of the typical $1x10^6$ total cells per ml of blood.

1.1.1.1 Circulating Pre-DCs

Advances in single cell sequencing and *in-vitro* culture of dendritic cells has uncovered potential heterogenic populations of early DC-restricted precursor cells with a phenotype comprised of CD34+ progenitor-like signals and a partial, or lowly expressed mature DC profile. Multiple research groups have reported pre-DC like cell types, although their exact relation to each other is not clearly understood.

One approach to unravel the pre-DC concept was undertaken by Breton *et al* and transcriptionally compared potential pre-cDCs as defined by a lack of DC expression markers, namely CD123 and CD303 for the exclusion of conventional mature pDCs, CD141 for the exclusion of cDC1s and CD1c for the exclusion of mature cDC2 cells (Breton et al., 2016). Although this population was observed to have expression patterns indicative of either a pre-cDC1 or pre-cDC2-like phenotype, collected cell numbers were low. Similar experiments using single-cell transcriptomics and subsequent bulk-profiling performed on all HLA-DR+, lineage-negative cells without the exclusion of cells with mature dendritic cell surface markers have since revealed multiple potential pre-DC populations including CD123+ subsets located within the conventional pDC flow cytometry gate, which were excluded in Breton *et al* (See et al., 2017; Villani et al., 2017).

This recent finding may have a drastic impact on the conclusions drawn on pDCs from previous studies and may explain prior observations of apparent cDC differentiation from pDCs *in-vitro*. Further discussion of recently published pre-DC subsets is given in Chapter 5.

1.1.1.2 Plasmacytoid Dendritic Cells (pDCs)

Plasmacytoid DCs are conventionally understood to be a unique lymphoid-lineage DC subset (Shortman and Liu, 2002), although certainty over their myeloid or lymphoid origin is still debated, with mixed culture results in murine plasmacytoid DCs (Sathe et al., 2013). These cells are defined by expression of surface markers including CD123, CD303 and CD304 (Collin et al., 2013). Preliminary computational experiments for this thesis using Illumina expression data (Haniffa et al., 2012) have uncovered a number of additional gene markers for pDCs including PTGDS, PACSIN1, and COBLL1 in relation to other PBMCs from healthy skin and blood.

Functional studies of pDCs have highlighted their ability to prime naïve CD4+ T-cells, including both Th1 and Th2 responses as well as inducing tolerance *in-vivo* (Ito et al., 2004), although this may be linked to the presence of a cDC precursor within typical pDC flow cytometry and FACS gating strategies (Villani et al., 2017). Plasmacytoid DCs are specialised to their role as sentinels under viral infection due to their specific TLR expression patterns, including TLR7 and TLR9, and high secretion of type-1 interferons compared to conventional DCs [Figure 1.2]. They do not express MHC class II to the same extent as cDCs and do no process antigens as efficiently as cDCs. In mice, they have been noted to provide a supporting role in antigen presentation by conventional DCs through their expression of CD40 ligand after activation via TLR9. CD40 ligand binds CD40 on cDCs and promotes IL-12 production, in turn inducing greater IFN-γ production in T-cells (Murphy and Weaver, 2016).

1.1.1.3 cDC1 Dendritic Cells (CD141+)

Conventional myeloid CD141+ dendritic cells, also described as classical DC1 (cDC1) are defined by their expression of surface thrombomodulin (CD141) and lower expression of CD11b and CD11c than CD1c+ DCs (Haniffa et al., 2012). Human cDC1s have a typical expression profile of CD141+, CLEC9A+, XCR1+, TLR3+, FLT3+, CD11b-low, and CD11c-low (Collin et al., 2011). Historically, CD141 positivity alone was used to differentiate cDC1 from cDC2 cells, particularly in flow cytometry experiments, however gene expression analysis of cDC and monocyte subsets have indicated RNA-level expression of CD141 is induced across multiple cell types and thus a selection of robust cDC1 markers should be used in combination to define the subset. They comprise of approximately 10% of human blood myeloid DCs and share homology with CD8+ lymph node DCs and CD103+ tissue DCs in mice (Bachem et al., 2010; Jongbloed et al., 2010).

CD141+ DCs have a greater capacity for cross-presentation than CD1c+ DCs in mouse, particularly cross-presentation of necrotic antigens, related to this cell types' expression of CLEC9A (Boltjes and van Wijk, 2014). Cross-presentation is discussed further in section 1.2.3.

1.1.1.4 cDC2 Dendritic Cells (CD1c+)

CD1c+ myeloid DCs contribute less than 1% of mononuclear cells (Dzionek et al., 2000). These cells are defined by expressing CD11c to a greater degree than CD141+ DCs as well as expressing CD11b, CD1c and SIRPA (Merad et al., 2013). Blood CD1c+ DCs are CD1a negative, while their tissue and lymph node equivalents are CD1a positive (Collin et al., 2011). Functional roles of CD1c+ DCs include Th2 induction with allergen response and Th17 induction in response to fungal infection and immune regulation (O'Keeffe et al., 2015).

1.1.1.5 Inflammatory DCs

Inflammatory DCs (infDCs) are a specialised subset of DCs that are primarily observed in the inflamed state and are believed to develop *in situ* from monocytes recruited under inflammation. Human inflammatory DCs express HLA-DR, FCER1, CD1a, CD11b, CD11c, CD14, CD172a and CD206, reflecting their hybrid monocyte-dendritic cell expression pattern (Segura et al., 2013; Segura and Amigorena, 2013).

Studies on human infDCs have largely been based on cells present in synovial fluid of rheumatoid arthritis patients or inflammatory ascites of cancer patients (Segura et al., 2013). Under these conditions, a population of myeloid CD16-CD1c+ cells with dendritic morphology and T-cell stimulatory capabilities has been observed, distinct from the more macrophage-like CD16+CD1c- population. Subsequent comparative transcriptomics have revealed distinction between infDCs and steady-state DCs and have instead linked infDCs to monocytes-derived DCs generated *in-vitro* based on their global gene expression patterns (Robbins et al., 2008).

Functional studies of infDCs have highlighted a potent T_H17 response when cultured with naïve CD4+ T-cells, which is not shared by inflammatory macrophages (Segura and Amigorena, 2013). T_H17 responses are critical in the pathogenesis of numerous autoimmune and inflammatory diseases and thus study of infDCs may provide insight into the management and development of these diseases, however their similarity to *in-vitro* generated moDCs may imply that for *in-vitro* experiments, moDCs are not an appropriate surrogate for the study of steady-state DCs.

1.1.2 Peripheral Blood Monocytes

Along with DCs and macrophages, monocytes are a major component of the mononuclear phagocyte system and were historically considered precursors of macrophages and DCs. Succeeding research initiated during the 21st century has indicated that both monocytes and macrophages are phenotypically, functionally and developmentally distinct from steady-state dendritic cells (Collin et al., 2013; Naik et al., 2013; Schraml et al., 2013).

Monocytes are highly heterogeneous. There are three well-established monocyte subsets in the blood that differ significantly in both phenotype and function (Auffray et al., 2009).

The major peripheral blood monocyte subset, accounting for 80-90% of all blood monocytes is a large CD14+,CD24+,CD16- group of highly phagocytic cells with low cytokine production. These cells can be distinguished by low CX3CR1 expression and high CCR2 expression and are deemed ‘classical monocytes’

A smaller CD16+ monocyte subset conversely expresses high CX3CR1, low CCR2 and have been shown, in the presence of lipopolysaccharide (LPS) stimulation, to produce TNF-α, IL-1β and IL-12 (Ziegler-Heitbrock, 2000). With Major histocompatibility complex (MHC) class II expression and presence of co-stimulatory antigens, CD16+ cells, termed ‘non-classical’, have been associated with acute inflammation and inflammatory response to infections (Belge et al., 2002), however it is now understood that CD16+ monocytes comprise of two groups, distinguishable by CD14 expression.

Multiple research groups (Collin et al., 2013; Grage-Griebenow et al., 2001) have suggested that it is in fact the CD14+CD16+ subset that exhibits inflammatory activity and cytokine secretion, aided in part by their expression of Fc receptors CD32 and CD64, while the function of the CD14^{low}CD16+ subset is relatively unknown, but suspected to perform a role similar to murine ‘patrolling’ monocytes.

1.1.3 Peripheral Blood Neutrophils

Neutrophils are generated from myeloid precursors in the bone marrow compartment under the control of granulocyte colony stimulating factor (G-CSF) (Borregaard, 2010). Neutrophils are the major component of circulating leukocytes and appear as large, granular cells with multi-lobed nucleus under microscopy (Kolaczkowska and Kubes, 2013). By flow cytometry, neutrophils can be distinguished by size, CD15, CD16 and CD68 positivity.

Neutrophils are capable of both intracellular and extracellular immune functions via phagocytosis, release of antibacterial proteins such as defensins or extracellular DNA-based traps (Häger et al., 2010).

1.1.4 Dermal Mononuclear Cells

Skin is the most readily accessible tissue available in relatively large amounts from patients and healthy volunteers from routine non-invasive surgical procedures, making it one of the main sources for human dendritic cell research after peripheral blood.

The dermis contains two major dendritic cell populations and a number of smaller, specialized subsets. The monocyte-macrophage-derived CD14+ population was initially described as a DC based on their MHC Class II expression and migratory capacity, but have subsequently been associated with a human and mouse blood monocyte-like signature by microarray analysis (McGovern et al., 2014). These cells function as helper follicular T-cell and regulatory T-cell inducers, but being monocyte-derived, have a poor ability to induce allogeneic T-cell proliferation (Klechevsky et al., 2008).

The largest population of dendritic cells in the skin is the CD1c+ DCs, believed to be a tissue equivalent to peripheral blood cDC2s (Haniffa et al., 2012). CD1c+ dermal DCs express CD40, CD80, CD83, and CD86, similar to their blood counterparts, but to a greater extent, as well as TLRs, lectins and other antigen processing proteins (Collin et al., 2013).

A small population of CD141+ DCs have also been identified in the dermis, believed to be functionally homologous to cross-presenting CD141+ DCs in the blood. Expressing XCR1, TLR3, CLEC9A and CADM1, CD141+ DCs exhibited typical cross-presentation markers above the levels of CD1c+ and CD14+ dermal populations (Haniffa et al., 2012).

Macrophages are large phagocytic cells capable of ingesting and degrading microorganisms though the innate immune response, but also play a role in adaptive immunity. MHC-II molecules and B7 are usually lowly expressed on steady-state macrophages, but are induced upon ingestion and recognition of microorganisms. These can be employed to present antigens after phagosome ingestion and degradation within the cell. Unlike DCs, dermal macrophages have a slow turnover rate and do not migrate out of tissue in cultures as DCs do. Similarly, they do not migrate when activated in the tissue, in contrast to DCs, which typically migrate to lymphoid tissue T-cell zones when activated. Instead, macrophages induce already primed T-cell responses locally, supporting effector functions and memory T-cells that get recruited to the site. Macrophages do not activate naïve T-cells in the same manner as dendritic cells, but instead present antigens to already primed T-cells to promote T-cell helper functions. Their inability to activate naïve T-cells is particularly important as macrophages continuously scavenge dead and apoptotic cells, and thus may present copious amounts of self-antigens, but are unlikely to induce an immune response related to this (Murphy and Weaver, 2016). In mouse, macrophages appear to have two developmental routes, a prenatally established population and a second population derived from blood LY6C+ monocytes that develops after birth (Malissen et al., 2014). They are specialised for wound healing and phagocytosis and can maintain their numbers through self-renewal (Ginhoux and Jung, 2014). Despite also being CD14+, distinguishing between CD14+ ‘DC’ cells and CD14+ macrophages in human can be achieved based on the their inherent autofluorescent properties related to melanin ingestion observed in the FITC channel and a high side-scatter profile by flow cytometry (Haniffa et al., 2009, 2015). Further comparisons between dendritic cells and macrophages are described in Figure 1.3.

1.2 DENDRITIC CELLS IN IMMUNITY AND TOLERANCE

Although they are rare in the steady-state, dendritic cells play a fundamental dual-role in their immune-mediated T-cell interactions, capable of inducing an immune response or tolerogenic capability depending on the activation state of an interacting DC and type of peptide presented (Heath and Carbone, 2001).

In immunity, cDCs act as sentinels to link the innate and adaptive immune systems though pathogen sensing and detection of peptides through antigen processing via MHC and subsequent activation of naïve T-lymphocytes (Mildner and Jung, 2014). Dendritic cells are capable of receptor-mediated endocytosis, phagocytosis and pinocytosis enabling them to process and present antigens and particulates efficiently (Mak and Saunders, 2005; Steinman, 1991). Their functional repertoire extends from antigen presentation to encompass aspects of innate immunity, T-cell activation, differentiation and regulation of immunotolerance (Akira et al., 2006).

Immature DCs, primarily located at sites with exposure to an external environment, act to capture, transport and present antigens to T-cells with a specific complementary surface complex (TCRs). Once activated in the presence of a pathogen and under the influence of cytokines, DCs mature to function as potent T-cell activators (Bluestone and Abbas, 2003; François Bach, 2003) [Figure 1.4].

Conventional DCs are most common in the skin, lungs and intestines, but also present in other major organs where they can serve as a first-line defense, primarily through the uptake of antigens by phagocytosis and presentation via MHC class II molecules. Dendritic cells can uptake and process antigens in a number of ways depending on the pathogen, route of presentation and intended response [Figure 1.5]. Extracellular bacteria, soluble antigens and viral particles can be engulfed through receptor-mediated phagocytosis or macropinocytosis for processing and presentation via MHC class II to CD4 T-cells.

Antigen production in the cytosol, typically as a result of viral infection, is understood to be the major route for delivering peptides to MHC class I molecules for CD8 T-cells. Cross-presentation of exogenous antigens can also occur through the endocytic pathway after phagocytic uptake for delivery to MHC class I and eventual presentation to CD8 T-cells [Figure 1.5] (Murphy and Weaver, 2016).

1.2.1 Pathogen Recognition

As major, potent stimulators of T-cells, dendritic cells orchestrate the immune response by both presenting antigens and releasing cytokine molecules to induce immune activation (Banchereau and Steinman, 1998; Lotze and Thomson, 2001). Antigens presented on specific surface-bound major histocompatibility complexes (MHCs) are recognised by specific subsets of T-lymphocytes (Ni and O'Neill, 1997). Cytosol-derived antigens, typically generated by intracellular viral or bacterial replication, bind to MHC class I surface molecules for presentation to CD8+ cytotoxic T-cells. Conversely, extracellular or vesicle-derived antigens are presented by MHC class II molecules to the other major T-lymphocyte subset, the CD4+ helper T-cells (Gargani, 2012; Mak and Saunders, 2005).

Presentation of an antigen by MHC class I or class II molecules is determined solely by the path that the antigen takes through the cell (Nairn, 2002; Rockefeller Institute, 2014). Between dendritic cells and these major T-cell subsets, both the intra-cellular and inter-cellular compartments can be monitored for pathogenic insults.

Of the pathogen recognition receptors, the Toll-like receptor (TLR) family has been the most extensively studied in mouse and human dendritic cells and differences in TLR expression may reflect specific subtype functions. Of the 10 known TLRs in humans, pDCs have been shown to express TLR1, TLR6, TLR7, TLR9 and TLR10, while cDC1s express TLR1, TLR3, TLR6, TLR8 and TLR10 and cDC2s express TLR1, TLR2, TLR4, TLR5, TLR6 and TLR8 (Hémont et al., 2013; Jin et al., 2014; Jongbloed et al., 2010). High expression of TLR3 and lack of TLR7 on cDC1 cells reflects TLR expression in the equivalent mouse CD8α+ DC (Edwards et al., 2003), although significant differences remain with mouse CD8α+ DCs expressing TLR4 and TLR9, while human cDC1s expresses TLR10, not present in mice and for which there are no known ligands.

C-type lectin receptors have also been linked to functional specialization of DC subsets in response to pathogens and vaccines. Of particular interest, CLEC9A on cDC1s, CLEC10A on cDC2s and BRCA2 on pDCs are now frequently used as classifiers of DC subsets in flow cytometry and FACS (Durand and Segura, 2015).

1.2.2 Cytokine Secretion

Cytokine secretion by DCs drives effector T-cell lineage via STAT activation pathways. The specific cocktail of cytokines is in turn determined through recognition of pathogen-associated molecular patterns by Toll-like Receptors, NOD-like receptors and other pattern recognition receptors on the surface of DCs. This process also drives activation of the DC and induces MHC and co-stimulatory expression.

Blood pDCs have long been noted to produce a strong type I interferon response, although more recently cDC1 cells have also been shown to be potent type I and type III interferon producers after TLR3 activation or Hepatitis-C interaction (Jongbloed et al., 2010). cDC2 cells are capable of cytokine production after stimulation with an array of TLR ligands, suggesting a less restrictive pattern of stimulation in cDC2 compared to cDC1 and a specialization for IL-12p70 secretion. Activated cDC1s produce primarily pro-inflammatory signals and may play a role in the recruitment of NK cells, monocytes and leukocytes, acting on the innate immune system (Hémont et al., 2013).

1.2.3 Cross-Presentation

In addition to MHC class II presentation for CD4+ T-cell activation, all blood and lymphoid DC subsets are able to engulf, process and present antigens to antigen specific CD8+ cytotoxic T-cells via MHC class I surface molecules along a cross-presentation pathway. This is one of the major functional methods to distinguish DCs from monocytes.

While lymph, spleen and tonsil cDCs are capable of cross-presenting antigen without TLR activation, blood cDCs under TLR-ligand stimulation also exhibit similar efficiency for cross presenting antigen, although as discussed in section 1.2.1, blood cDCs express differing TLRs and will therefore react differently depending on the specific stimuli detected (MacDonald et al., 2002; Mittag et al., 2011).

Examples of subset specific cross-presentation have been noted with cDC2s capable of inducing CD103 expression in CD8+ T cells for adherence to epithelial cells (Yu et al., 2013), while the association between mouse CD8α+ DC and cDC1 has promoted interest in cDC1 cross-presentation due to its expression of CLEC9A, which specifically recognises actin-binding cytoskeletal proteins bound to actin, which are exposed when a cell membrane is significantly damaged. Indeed increased cross-presentation of necrotic cell-derived antigen by cDC1 was noted in Crozat *et al* (Crozat et al., 2010).

Human pDCs cross-present soluble, viral, cell-associated antigens and antigen specific to pDC surface markers including CD40, CD205 and BDCA2 (Bachem et al., 2010; Heath and Carbone, 2001).

DC cross-presentation in the inflammatory setting and in tissue, with the exception of skin DCs, has not yet been conclusively addressed.

1.2.4 DC-Mediated Tolerance

Immune tolerance prevents auto-reactive immune components from inducing an immune response, leading to the development and progression of autoimmune disease.

A recent review highlighting the importance of dendritic cells in prevention of widespread auto-reactivity (Audiger et al., 2017) focused on transgenic mouse models with induced specific DC subset depletions (Birnberg et al., 2008) and relating them to clinical observations of autoimmune disease states including IRF8 mutations which frequently present as defects in dendritic cell development and reduced numbers of DCs in the blood and tissue, resulting in high infection rates and an increase in anergic, non-reactive T-cells (Bigley et al., 2017; Hambleton et al., 2011; Salem et al., 2014).

Far from being just naïve T-cell activators, DCs play an important role in T-cell thymic selection and have been implicated in T-cell peripheral toleration. DC involvement in thymic selection allows for the conservation of cellular materials by inducing apoptosis in non-functional and hyperfunctional CD4+/CD8+ phase T-cells (Hawiger et al., 2001; Steinman et al., 2003). These defunct cells can be stripped from the cellular population though a series of positive and negative selection processes that involve DC presentation of peptides to developing T-cells (Cannarile et al., 2004).

T-cells that cannot develop functional MHC-recognising surface receptors would be useless in their peripheral immune role and are thus destroyed in the thymus rather than being released into the circulating population. Various studies have indicated that T-cells failing positive selection in this manner can account for up to 80% of all generated T-cells (Hogquist et al., 2005; Shurin and Salter, 2009).

Negative T-cell selection filters out T-cells which display activation activity against self-peptides and MHC molecules (Brocker et al., 1997). In this selection step, DCs and other APCs will present self-peptides in an effort to root out overly sensitive thymocytes. These potentially auto-immunogenic T-cells may, if released from the thymus, cause extensive cell damage to the host. For this reason, it is imperative that such autoreactive cells are induced to undergo apoptosis before they can cause any cell damage in the periphery. Approximately 20% of the developing T-cells are destroyed for failing this selection step.

The thymocytes that pass the selection process are expected to respond most specifically to non-self peptides presented on self-MHCs. This is the exact combination that is most likely to warrant an immune response and may indicate the presence of viral, bacterial or antigen infection (Delves et al., 2011).

Those T-cells that ‘fall through the net’ of thymic selection can be neutralised in the periphery by dendritic cells and other APCs via a process known as T-cell tolerisation. This process affects cells that bind and respond to any peptide that is presented in the absence of a ‘danger signals’ including INFα, TNF, double-stranded or single-stranded (viral) RNA and a number of ‘heat shock proteins’ (Gallo and Gallucci, 2013). Without the induction of an innate response by such signals, DCs will not produce activating co-stimulatory molecules and an immune response to the self-peptide will not be initiated. In this manner, healthy tissue can induce tolerance to the self (Xing and Hogquist, 2012). In the steady-state, DCs express low levels of activation markers and are capable of presenting self-antigens to T-cells, resulting in T-cell anergy or deletion of the cell. This process is likely to occur frequently due to migration of tissue DCs to lymph organs and provides a mechanism of tolerance outside of the thymus selection process. This process is compounded by secretion of inhibitory cytokines including TGFβ by tolerogenic DCs expressing low levels of activation markers. TGFβ induces expression of T-cell Foxp3 and production of regulatory T-cells, which in turn maintain DCs in their inactivated state (Chen et al., 2003).

1.3 HAEMATOPOIESIS AND DC DEVELOPMENT

Haematopoiesis is the process through which blood cells (leukocytes, red cells and platelets) develop. In the adult, self-renewing, multipotent haematopoietic stem cells (HSCs) give rise to all cells of lymphoid and myeloid lineages. Modern studies of the haematopoietic process suggest that these pathways exhibit a degree of plasticity, rather than representing hierarchical differentiation through a series of bifurcations with successive loss of potential. Early progenitor cells exhibiting a capacity for both lymphoid and myeloid potential are termed multipotent, while successive downstream cell types are typically more lineage-restricted (Ceredig et al., 2009).

Due to the role of mature blood cells in immune response, circulation and inflammation, most HSC-derived cells are short-lived with a rapid cell turnover. This facilitates immediate, specialised defense to stimulus, but requires precise regulatory mechanisms to maintain appropriate populations. Haematopoietic cancers (ie leukaemia) are an accumulation of immature cells (blasts) due to genetic mutations and failure to properly differentiate, which result in a deficient, incapable immune system.

1.3.1 Early In-Vitro Models of Haematopoietic Development

Discovering how haematopoietic stem cells differentiate into their diverse mature cell types is fundamental to understanding the functional role that each cell type plays. By understanding the haematopoietic developmental pathway, manipulation of the process may offer therapeutic treatment potential in haematological disorders (Kawamoto et al., 2010).

The classical model of haematopoiesis, derived from early morphological studies suggested a model of rapid restriction of cell plasticity and lineage which fits well with the classical model of DC development (Kondo et al., 1997).

Plasmacytoid DCs (pDCs), with a characteristic appearance, surface receptor profile and interferon production capacity have long been recognised as lymphoid lineage cells. Such interferon producing cells were not found in *in vitro* culture experiments on CD34+ progenitor cells, supporting the concept of early lineage commitment. Further studies in the murine system have since identified myeloid precursor-derived pDCs, calling into question the accuracy of this classical model (Lai and Kondo, 2006). It is now generally accepted that the haematopoiesis model does not take into account the capacity of myeloid potential in lymphoid lineage cell precursors, specifically multi-lymphoid progenitors (MLPs), nor the capacity of CDP to produce both pDCs and conventional myeloid DCs (Doulatov et al., 2010).

Most research into human dendritic cell development and maturation has been conducted on cultured iDCs, moDCs or precursor DCs. Older *in-vitro* studies have led to a theory of multiple DC lineage pathways, although these are believed to exhibit some plasticity [Figure 1.6]. Whether DC maturation occurs in the same manner *in-vitro* as *in-vivo* is still debated (Shortman and Liu, 2002).

1.3.2 Recent Models of Haematopoietic Development

With the expansion of bulk sequencing, single-cell sequencing, culture techniques and cytometry, the working model of dendritic cell development is progressing rapidly.

The linear bifurcation model [Figure 1.7], in which there is a step-wise progression through multi-potent progenitor cell types with a dichotomous fate decision at each step was likely reinforced by the bulk-level analysis techniques frequently used in combination with fluorescent activated cell sorting. Isolation of previously defined multi-potent progenitor cell types by FACS, followed by population-level analysis would provide an overall expression profile for the cell type, leading researchers to conclude that multipotent progenitors exist in the periphery as a homogenous population of distinct cells with multi-potential. This notion has since been overhauled with the advent of single-cell technologies. Single-cell level analysis of previously described multi-potent progenitor cells types has revealed the true heterogenic nature of these cells, providing an alternative model in the form of early lineage priming [Figure 1.8]. In this model, CDP, MDP and other multi-potent cell populations represent a transection of multiple distinct, primed, uni-potent cells that share a largely related, transitionary phenotype (Notta et al., 2016; Paul et al., 2015). This feature extends back to the haematopoietic stem cell and progenitor compartment described as a cloud of early low-primed, undifferentiated, uni-potent cells with plasticity (Hamey and Göttgens, 2017). A major implication in this model posits no distinction between lymphoid and myeloid lineages in the origins of DCs, since all DCs are the product of a core lymphoid-myeloid pathway (Karamitros et al., 2017). The typical dichotomy of lymphoid and myeloid developmental pathways can instead be considered a spectrum, ranging from most myeloid-like monocytes, through cDC2, cDC1s and into lymphoid-like pDCs.

1.3.3 Transcription Factors Affecting DC Development

A number of transcription factors have been recognised for their involvement in DC progenitor commitment, specification and survival. These transcription factors generally either influence all DC subsets at the early progenitor stage, having a profound effect on multiple DC lineages, or regulate later-stage committed DCs (Satpathy et al., 2011) [Figure 1.9].

Transcription factor PU.1 has been identified as an essential factor for DC development. It is a nuclear ETS-domain containing protein that binds to a PU-box located near the promoter region of target genes, regulating their expression or influencing alternative splicing of target genes. This factor controls and regulates expression of multiple essential DC development genes, including Flt3, GM-CSFR and M-CSFR (Zhang et al., 1994). In mouse models with conditional deletion of PU.1, no DC subsets develop although defects in other myeloid cells were also noted, implicating PU.1 as an essential factor not only in DC development, but the wider haematopoietic compartment (Carotta et al., 2010). In human disease states, defects in the SPI1 gene controlling PU.1 protein formation have been linked to neutrophil-specific granule deficiency, as well as affecting type II interferon signaling and IL-4 signaling pathways.

Further mouse and human studies have described the importance of nuclear zinc-finger proteins in early DC development (Ng et al., 2009; Rathinam et al., 2005; Yoshida et al., 2010). Growth factor independent 1 transcriptional repressor (GFI1), is one such zinc-finger protein that works in combination with other cofactors to silence the promoters of target genes through induction of histone modification. The UniProt entry for GFI1 highlights its broad influence over blood cell development, affecting neutrophil differentiation, lymphoid proliferation and granulocyte development as well as TLR regulation, cell-cycle progression and T-cell receptor signaling.

IKAROS, another fundamental zinc-finger transcription factor, plays a profound role in the development of pDCs and to a lesser extent, other early progenitor cell types through its control over a range of receptors and transcription factors including FLT3, IL-7R and NOTCH1 (Ng et al., 2009). Interestingly, a lack of IKAROS in humans results in selective reduction of pDCs, but expansion of cDC1 (Bigley et al., 2017), although whether this observation is a direct result of IKAROS, or a compensatory mechanism spurred on by the lack of pDCs is yet to be established.

Along with broad-ranging transcription factors, other factors are fundamental to the development of specific cell-types and lineages. In pDC development, E2-2 is the major lineage-determining factor that is negatively regulated by ID2 (Spits et al., 2000, p. 3). Modulation of these factors can up-regulate or down-regulate pDC production, favouring pDC or cDC lineages (Satpathy et al., 2011). A competing mechanism between related interferon regulatory factor family members IRF4 and IRF8 determines lineage commitment along the cDC2 or cDC1 pathways, respectively. The competing and complimentary mechanisms of these factors influence the lineage potential in early DC development, priming progenitor cells towards a uni-potent potential and reinforcing maturation through transcriptional control of other factors.

1.4 CONSIDERATIONS IN DC RESEARCH

The field of dendritic cell research has been substantially expanded following extensive development of analytical techniques, however the fundamental qualities of dendritic cells and complexity of their development and activation still prove to be significant obstacles to the progress of human dendritic cell research.

1.4.1 Considerations in DC Transcriptome Profiling

As noted in section 1.1, dendritic cells are a rare cell type and even though they are readily available in two of the most non-invasive tissues, the skin and peripheral blood, their scarcity in these tissues makes their collection, isolation and pre-processing an arduous process. Cell attrition from collection to subsequent flow cytometry analysis or FAC sorting can result in very few viable cells reaching the end of the pre-processing stages. Furthermore, each of these pre-processing steps affects the fragile regulatory processes and survivability of the DCs that remain. The practice of tissue dissociation and disruption can rapidly alter the expression and protein profile of dendritic cell subsets. Separating biological features from technically induced artifacts may prove an impossible task.

Additionally, discrepancies between human and mouse DC development and immunobiology resulting in poor efficacy of otherwise promising targeted immunotherapies in humans, reveal issues with reliance upon extrapolation of data from mouse disease state and knock-out models into the human system (Haley, 2003; Mestas and Hughes, 2004; Monaco, 2003; Oehler and Bicknell, 2000; Sykes, 2001). Continued use of mouse models is unavoidable at this time as although multiple human DC deficiency states have been identified, providing possible avenues for in-depth analysis of fundamental *in-vivo* development, collecting adequate sample material with enough replicates for research is problematic.

To circumvent this problem, *in-vitro* models of human DC development are frequently relied upon to infer *in-vivo* biology, yet further intrinsic issues accompany this, namely the dependence on monocyte-derived dendritic cells or CD34-derived DCs to produce adequate cell numbers for transcriptomic or metabolic analysis. Neither of these cells' surrogates accurately recapitulates *bona-fide* DCs, typically expressing hangover-signatures from their myeloid-monocyte origins and inflammatory signals resulting from their forced activation. A method for the production of true DC equivalents *in-vitro* is an ongoing enigma as further highlighted in Chapter 4.

Whether experimenting on primary DCs or *in-vitro* populations, analysis pipelines typically begin with the sorting or flow cytometric analysis of cell subsets according to established cell surface markers, yet surface marker expression is fickle across tissues, populations, and within individuals. The historical lack of robust, positive cell type markers in human DCs is beginning to be addressed through single-cell and bulk level transcriptomics, but these techniques have also revealed flaws in the use of traditional DC markers such as CD123, CD141 and CD1c, which have been shown to have greater variance across cell types than previously believed and have since been combined with other markers identified by single cell transcriptomics for greater purity during isolation (Villani et al., 2017).

Transcriptional profiling of dendritic cells may be more reliable, provide a much wider coverage of global gene expression than flow cytometry or qPCR and provide information on transcription factor expression patterns.

1.4.2 Cytometric, Gene Expression and Sequencing Technologies

While dendritic cell research has been studied extensively over the past 30 years by traditional means, the advent of high-throughput expression analysis has opened a new avenue of cell research using large-scale arrays on small cell numbers, including single cell gene expression analysis.

1.4.2.1 Flow Cytometry

As dendritic cell subsets may differ by expression of surface antigens and markers, techniques involving the detection of these subset markers can be used to identify and separate out DC subsets from a heterogeneous mix.

Flow cytometry is a well-established and widely implemented technique for cellular analysis. The principle of flow cytometry involves the use of fluorescent dye-labeled antibodies or molecules that specifically bind to cellular components of interest. These dyes are designed to respond to certain wavelengths of light by emitting photons of a different wavelength that can, after passing through a variety of filters, be detected and measured (Watson, 2004).

Subpopulations of a heterogenous collection of cells can be easily identified and analysed by flow cytometry. Multi-parameter data is collected for thousands of cells per second, with each parameter providing an intensity score for each interrogated cell. This can be particularly important if two cell populations share similar surface markers, but with a varied density, for example expression of CD14 on CD14+ classical monocytes and CD14+16+ monocytes.

For flow cytometry experiments, a suspension of cells is directed into a narrow, single-cell stream that is interrogated by a laser light source [Figure 1.10]. Flow cytometers can rapidly detect and analyse hundreds of thousands of cells per minute. The scattered and emitted light from the fluorescently labeled cells is measured using a number of detectors, producing a potentially highly multi-dimensional dataset of the physical and fluorescent properties of each cell (Abcam, 2014; Rahman, 2009).

The development of flow cytometry technology has resulted in considerable increases in its complexity. The main benefit of this has been an increase in the number of parameters that can be interrogated at once. Simple flow cytometers may be have the capacity to measure three different fluorescent wavelengths, while the latest models offer over 20 parameters, allowing for wide and varied multi-parameter data acquisition and identification of rare cell subsets.

Two parameters commonly tested in flow cytometry are forward scatter (FSc) and side scatter (SSc). Forward scatter is the term for the light collected from the opposite side of the particle stream to the light source. Light scatter in the forward direction (up to 20° from the light source axis) provides an estimation of particle size. This parameter is particularly useful for filtering out living cells from small particles and debris. Side scatter (light detected 90° from the light source axis) indicates the level of granularity of a cell or particle. These two parameters can be used to distinguish neutrophils, monocytes and lymphocytes to a certain degree [Figure 1.11] (Rowley, 2015).

1.4.2.2 FACS-Based Cell Sorting

The first step in DC subset analysis generally involves an initial separation of the major subsets based on established and well-defined parameters. Fluorescence-activated cell sorting (FACS) is a cytometry-derived method for identification and isolation of cell subsets based on their physical attributes and presence of cell surface markers. Using the same, or similar parameters to regular flow cytometry, FACS allows the separation of subsets by their physical and fluorescent properties. For FACS, a thin, flowing, unicellular stream of suspended cells is vibrated in such a manner that small droplets of fluid are produced. These droplets are expected to contain a maximum of just one cell, meaning that cells can be interrogated individually. After passing through the fluorescence measuring apparatus, and prior to breaking into individual drops, an electrically charged ring places a charge on the partially formed droplet, which is then passed through electrostatic deflection plates, skewing the droplet's movement towards one of a number of collection vessels based on its charge (Wersto, 2014). [Figure 1.12].

When investigating homogeneous populations, FACS can quickly detect and display any non-uniformity through two-dimensional flow plots. Gating strategies to filter out dead cells, cell debris and cells expressing unusual or unexpected cell surface markers allow for the separation of 'purified', uniform cells (Rahman, 2009).

Despite the high speed of cell interrogation, sorting small populations of cells from a heterogeneous mix can still be highly time-consuming. For many experiments relating to gene expression, a relatively large quantity of cells is required. For major cell types the number of required cells can be obtained within a few minutes of FACS sorting, however for small, rare subsets (eg. CD141+ DCs), even high speed sorters may require many hours of sorting to collect a sufficient number of cells. Use of the machine for this period of time can be expensive and may also pose quality issues as the cells may die or degrade in the time taken to collect sufficient cell numbers. This time issue is further compounded if sterile cells are required for an experiment, where throughput can be even lower (Watson, 2004).

1.4.2.3 Illumina Microarray Assays

Illumina BeadArray Technology such as the HumanHT-12 v4 Expression BeadChip used in this project, provides researchers with the ability to perform extremely high-throughput RNA or DNA based gene expression analysis of over 47,000 probes derived from the National Center for Biotechnology Information (NCBI) Human RefSeq 38 database. This level of transcriptome coverage includes a number of splice variant genes, possible gene candidates and many well-characterised gene targets for a diverse range of applications and experiments.

Illumina BeadArray chips are composed of self-assembled 3μm silica beads randomly arranged in microwells with a spacing of ~5.7μm [Figure 1.13]. Each of these beads has attached to its surface hundreds of thousands of 79 nucleotides-long oligonucleotide sequences representing a particular Illumina probe that exhibits complementarity to mRNA sequences of interest.

Due to the random nature of Illumina BeadArray construction each array is unique and must be decoded during the quality control phase to determine adequate coverage of probes. To do this, each illumina probe sequence has a 27-nucleotide sequence that is used to map the location of each bead type on the array. Typically, each of the 47,000 bead types is present on an array around 30 times. The main benefit of random bead assignment is the elimination of spatially localised artifacts. Such issues can result in major expression flaws in other array-based gene expression technologies.

Reading of the Illumina BeadArray cartridge is performed by a confocal laser-scanning microscope, the BeadArray Reader (Illumina, 2011).

1.4.2.4 Affymetrix GeneChip Technology

Affymetrix GeneChips are an older and more established technology than Illumina BeadArray. This technology utilizes photolithography to generate an array on a quartz chip in a process similar to computer microchip construction. Manufacturing costs are high with this process but the resulting GeneChips attain comparable coverage to Illumina arrays (47,000 transcripts). Each GeneChip contains two sets of probes. One set, designated 'perfect match', is designed against the 3' end of a gene of interest. The second set of probes have a number of single nucleotide substitutions at the 13^{th} base. This 'mismatch probeset' acts to measure background hybridisation, data from which can be used as a background or negative quality control [Figure 1.14].

As the Affymetrix arrays are mapped out during creation, all probes of a particular type are localised to a designated region of the chip. This can become a major issue if the chip is damaged or soiled in any way, including presence of dust particles on the chip surface that may obscure intensity readings from a region of the chip. While region specific drop-outs are not a major problem with randomly generated Illumina arrays, region-specific drop-outs in Affymetrix experiments may potentially obstruct all probes for a specific gene present on the chip [Figure 1.15].

1.4.2.5 NanoString nCounter Assays

NanoString nCounter technology utilises an alternative approach to fluorescence-based expression compared to the intensity-based methods of Affymetrix and Illumina, as well as traditional methods such as Real Time Polymerase Chain Reactions (RT-PCR).

Rather than calculating an intensity score to determine relative expression levels, NanoString directly profiles individual molecules in a highly multiplexed reaction by assigning fluorescently labeled probes to genes of interest. In this manner, probes that have bound to a molecule of interest can be counted by a computerised optical lens. A chain of 7 fluorescent tags are bound to a biotin-containing oligo complementary to a 100bp region of the 3' end of a gene of interest. The order of these coloured tags act as a 'barcode' to indicate which gene the complementary oligo refers to [Figure 1.16]. Restrictions on the sequence of the four available fluorophore colours limits the number of genes that can be assessed in one reaction to approximately 800. This scope puts NanoString nCounter technology between RT-PCR and whole genome microarrays with respect to gene coverage.

Unlike other gene expression methods, NanoString does not require any pre-amplification or polymerisation of sample input material. As a result, issues with amplification bias (where even small deviations in molecule amplification in a heterogenous mix are amplified over a number of cycles, resulting in comparative over- or under-amplification of certain sequences and therefore a skewing of any true underlying expression differences) are non-existent. The random nature of probe binding to the detection surface of the nCounter cartridges eliminates region specific gene drop-outs, while the non-reliance on fluorophore signal intensity negates compensation based errors. Fully automated loading of the cartridge, pre-processing, digital detection and analysis minimizes user error, making NanoString assays reproducible, highly sensitive and extremely robust.

NanoString arrays are single tube reactions, which can be designed to simultaneously detect specific mRNA, miRNA and more recently, proteins of interest. This level of complexity can allow a researcher to amass a vast quantity of data relating to all stages of the gene expression process in a single experiment using a range of sample types including fresh, frozen, lysed, extracted or formalin-fixed, paraffin-embedded (FFPE) samples from a number of organisms.

1.4.2.6 Bulk-RNA Sequencing

RNA sequencing provides a complete, precise measuring of transcripts and isoforms far above the level of microarray-base platforms. By determining the complete population-level transcriptome, functional gene elements, transcriptional structure of gene, splicing patterns and post-transcriptional modifications can be interrogated and inferred (Wang et al., 2009). Rather than the probe-based approach of microarray technologies, Sequence-based methods of RNA-Seq are capable of determining the definite cDNA sequence, although the process is typically low-throughput. Tag-based methods are of higher throughput, but usually based on expensive Sanger sequencing technology.

As there are multiple competing technologies available to RNA-seq, associated costs have been reduced and methods can be tailored to suit a users needs. Some of the major variables are read depth, read length and single or paired-end reading. Read depth determines the number of reads per sample. Global gene expression comparison experiments may require lower read depth than those for the identification of low-expressing genes or novel transcripts. Read length is typically set between 30 base-pairs and 400 base-pairs, with 75bp being typical in many applications. Longer reads may be required to the investigation of novel isoforms, gene fusion events or identification of unknown transcripts. Single-end reads can be adequate for population comparison studies, although paired-end data can be used to detection of alternative splicing patterns and gene fusions.

A normal workflow for RNA-Seq involves library preparation, sequencing and analysis with most of the time allocation given to the subsequent analysis of the data.

1.4.2.7 Single-Cell Sequencing

Interest in ‘single cell’ sequencing (scRNAseq) has expanded rapidly since the start of this project. Initially prohibitively expensive, protocol development, competition and multiplexing techniques have made such experiments more accessible. A fundamental switch in the field from basic science over to large-scale bioinformatics and ‘big data’ has fuelled this progress.

In contrast to traditional ‘bulk’ RNASeq, scRNAseq allows for the detection and quantitation of mRNA from individual cells, opening avenues for in-depth heterogeneity studies, cellular development and population genomics. While variations in protocol are now commonplace, most single cell experiments follow a similar pipeline beginning with the isolation of live, viable single cells. This can be performed through flow cytometric sorting, droplet-based methods or micro-dissection (Haque et al., 2017). Once individual cells are captured, they can be lysed to release the mRNA required for capture using poly(T) sequence primers before priming and conversion to cDNA with a reverse transcriptase. cDNA can then be amplified and prepped with a library of nucleotide ‘barcodes’ and sequenced.

Different protocols deliver different outputs, which need to be considered when selecting a single cell platform and protocol. Microfluidics based protocols such as the Fluidigm ‘C1’ can offer full length transcripts for 1,000 to 10,000 cells with a read depth of 10^6 per cell, but require nanolitre volumes to perform with high sensitivity (Pollen et al., 2014). The plate-based Smart-seq2 protocol employed in chapter 5 of this thesis with some modifications can also produce full-length transcripts with equivalent throughput to C1, but work on a microlitre scale to allow for direct FAC sorting of cells into plates (Picelli et al., 2013). More recently, droplet-based protocols have emerged which allow for very high throughput using dedicated hardware to perform thousands of individual single-cell reactions with throughput up to 100,000 cells at a cost of decreased read-depth per cell (Macosko et al., 2015; Zilionis et al., 2017). These drop-seq protocols are typically cheaper than other methods but lack sensitivity as a result of lower read depth.

Whichever method is used for pre-processing, extensive bioinformatics are required for cleaning, processing and analyzing the resulting read counts. No ‘gold standards’ have yet emerged to govern quality control, data filtering and differential expression, but common features of scRNAseq publications include filtering on mitochondrial reads, non-binding spike-in External RNA Controls Consortium (ERCC) controls and total reads per cell (Lun et al., 2016).

1.5 BIOINFORMATICS AND STATISTICAL ANALYSIS

Bioinformatics has become a major part of scientific research in the last decade with wider accessibility to high-throughput, multi-parameter and highly multiplexed experimental techniques. With this, the demand for computationally and mathematically savvy scientists has also increased. Where previously data analysis could be performed by hand on paper or using common low-power desktop computers for qPCR, flow cytometry and imaging, the sheer bulk of data and computational demands of microarray, next-generation and single-cell sequencing data requires the use of dedicated high-powered computing clusters, development of intricate pre-processing and analysis pipelines and vast stores of digital memory storage.

1.5.1 Data Collection and Storage

Data are collected in various formats depending on platform and experiment types.

Typical output for flow cytometry and FACS experiments include an index for each cell, experimental meta-data, time, light scatter values for determination of a cells area, height and width, voltages, colour compensations and fluorescence intensity for each measured parameter. Containing data on over 15 parameters for 10^5-10^7 cells, single flow cytometry output files can be up to hundreds of megabytes in size.

In contrast, NanoString data output is rarely above 30 kilobytes in size for a single sample, with count files and metadata for a 12-sample, single cartridge experiment totaling less than half a megabyte. While the number of parameters in a nanostring experiment can be up to 800 gene targets, even the largest published experiments uploaded to NCBI contain around 2,000 samples (Ye et al., 2014), limiting the total experimental data to a manageable size.

While flow cytometry and NanoString data can typically be handled and stored on a desktop computer or shared storage space and transferred by email or USB-storage devices, output from NGS and single-cell experiments are far more difficult to handle. With millions of transcripts per sample and experiments commonly involving the analysis of 10^4 to 10^5 cells (Villani et al., 2017) up to 10^6 samples from 10x genomics ("10x Genomics," n.d.), bulk-sequencing data can amount to gigabytes, and single-cell RNA-sequencing can reach terabytes worth of storage, requiring the purchase and upkeep of dedicated high-memory, high-storage computer clusters and servers. Because of this, data storage and use of HPC facilities and expertise are an increasing expense for researchers.

1.5.2 Data Normalisation

The purpose of normalisation is to improve the reliability of experimental data by adjusting for sample variance, user error and technical variability. To separate biological variance from technical variance many technologies incorporate a known molecules or signals into the reaction chemistry to provide a stable background for comparison.

For flow cytometry, this could come in the form of freeze-dried fluorescent pellets of a known number added to each sample to provide a comparator for determining absolute counts of mononuclear cells in blood.

NanoString Technology uses positive spike-in control sequences of known concentrations, which are spiked into every codeset to adjust for variation attributable to minor differences in the amount of codeset added to each well of the cartridge. In sequencing experiments, ERCC-controls are commonly used to a similar effect. Expected to produce the same number of reads in for each sample, any variation in these counts can be attributable to technical variance and thus adjusted for (Arzalluz-Luque et al., 2017). In the absence of, or parallel to the use of, positive controls, 'housekeeping genes' – those genes believed to have equal expression across all samples can be used to develop a normalisation ratio applicable to the data.

Library size normalisation is frequently used in multiplexed RNAseq experiments to account for inherent variation of reads from each sample. Counts Per Million (CMP) is a common method of single cell library size normalisation that normalises data based on the number of reads for an individual gene and sample, divided by the total counts multiplied by 1,000,000. A variation on this is FPKM (Fragments per kilobase of exon per million reads mapped). This is usually a staple normalisation method of bulk RNA-seq experiments and integrated into many 'R'-based sequencing analysis packages, although it is steadily being replaced in favour of transcripts per million (TMP).

1.5.3 Statistical Analysis

Statistical analysis is performed at the gene level, sample population level and functional enrichment level for the robust distinction of cell types, developmental stages or disease states. The purpose of statistical testing is to back up conclusions with mathematical confidence that such a result would be repeatable, true and was not likely just the outcome of random chance or variation. Statistical analysis may also provide insights into cell development, signature genes and functional enrichment. Identification of 'statistically significant genes' does not necessarily equate to biological relevance and thus in gene expression studies, shared gene functions, pathway mapping and enrichment analyses are usually implemented to provide a biological interpretation of differential expression.

1.5.3.1 Significance Testing and Differential Expression

Significance testing aims to support or reject an assumption made about a population based on a sample taken from this population. Typically significance testing is broken down into one or more hypotheses. A null hypothesis (H_0) indicates that there are no differences between the test groups, while an alternative hypothesis (H_a) usually represents a statement that the experiment was developed to assess; that one group is distinct, on average, from another, beyond that expected by chance. If this verdict is reached, a researcher would reject the null hypothesis in favour of the alternative.

One of the most common analysis techniques in transcriptomics is the identification of differentially expressed genes. Genes related to specific disease states, risk factors and genotypes may offer molecular and biological insights into disease risk and expression variance (Sun et al., 2017).

For gene expression technologies, methods of differential expression differ widely based on the format and distribution of output data and the research question to be answered.

NanoString Technology, using a digital, multiplexed, PCR-like chemistry produces integer counts, which when normalised become non-integers. NanoString recommended use of a two-tailed student's t-test for analysis of differentially expressed genes, with the inclusion of Benjamini-Hochberg false discovery rate adjustment to account for multiple testing. Very recently, NanoString have shifted recommendations for some panels over to multivariate linear regression with Benjamini-Yekutieli false discovery rate adjustment, in a closer approximation of microarray-based analysis techniques, although most publications using NanoString refer to t-tests for differential expression.

Illumina BeadArray differential expression analysis is most commonly performed using the 'limma' package of 'R'. Very high probe numbers, combined with very few biological replicates poses a statistical problem shared by many microarray and sequencing techniques. Limma addresses this with the use of gene-wise linear models to estimate log-expression ratios between sample types, incorporating an empirical Bayes framework to estimate variances. The high dynamic range of microarray data typically results in very small p-values for differentially expressed genes, but the very high degree of multiple testing must be accounted for using false discovery rate adjustment, otherwise false positives in the data will be significant (Ritchie et al., 2015).

Bulk-RNAseq analysis for differential expression is frequently performed using 'DESeq2' package of 'R'. This relies on a generalized linear model and negative binomial distribution. Such algorithms have been shown to work quite well and have been validated using RT-PCR (Rapaport et al., 2013).

scRNAseq is still fluid in its accepted analysis methods although the consensus is that bulk-RNAseq methods are generally not appropriate for single cell work. Rather than defined groups, single cell experiments usually rely on unsupervised clustering to produce groups of similar cells, which can then be compared for variance. Sample numbers are higher than in bulk RNAseq, although drop-out rates are much higher. 'SCDE' package is one of the first single-cell differential expression packages utilizing a zero-inflated negative binomial model with Bayesian statistics to specifically account for these differences between single-cell and bulk-RNA sequencing (Kharchenko et al., 2014).

1.5.3.2 Functional Analysis and Geneset Enrichment

Functional analysis is generally performed after differential expression analysis to better understand and deconvolute the individual gene-level expression patterns that have been identified. By investigating the effect and functions of differentially expressed genes as a group, the nature and mechanism of a disease state or population may be revealed. In this method, the Gene Ontology initiative functional annotations are used for each gene and statistical analysis for over-representation of genes with particular functions is performed to determine the functions shared by more genes than statistically expected in the dataset (Ashburner et al., 2000).

The same over-representation tests can be applied to curated databases of genes associated with specific disease states, cancer types, cell types or other bin types. MsigDB (Liberzon et al., 2011) is one such resource for this, incorporating hallmark datasets, cell type genesets and other published genesets from thousands of studies.

1.5.3.3 Pathway Mapping

Similar in purpose to functional enrichment, pathway mapping uses over-representation of genes along curated genesets related to biological pathways. The expression patterns of genes along these pathways are usually mapped onto a diagrammatic representation of the pathway to identify potential target genes or 'choke points' in the pathway that may be targeted for follow up experiments, this may be particularly valuable for drug discovery experiments, but it rarely employed in other research areas. Pathway mapping may be used in disease progression studies to identify therapeutic targets to correct altered expression along a disease pathway, or in the case of cell type analysis, may be used to infer cell maturity or development.

1.5.3.4 Machine Learning, Hierarchical Clustering and Gene Reduction

Machine learning is an expanding field of bioinformatics and computer science. Within the sphere of data analysis, machine learning is used to predict patterns in the data using various algorithms, that would otherwise be too time consuming or convoluted to process manually. Dimensionality reduction is distinct from machine learning, but is usually incorporated into it. Dimension reduction aims to remove much of the collinearity within a dataset, shed unnecessary variables and streamlines the data for easier data interpretation, visualisation and better fitting of machine learning models.

Principal Component Analysis (PCA) is a common linear method of dimension reduction. This method aims to maximise variance in low dimensional space by mapping the data to eigen vectors (principal components) based on a correlation matrix of the data. The first principal components equate to the most data variance and thus, may be used to reconstruct most of the variance in the data, but with far fewer dimensions. The first two or three components are typically plotted in a scatterplot format to identify groups of closely related samples and interpret the variance between sample groups.

Linear discriminant analysis (LDA) is based on Fisher's linear discriminant and is used in machine learning to identify a linear combination of genes capable of separating multiple populations (Venables, 2002). When building a LDA model, a training dataset is used to extract the differences between populations. This same model can then be applied to 'unknown' samples to categorise them by similarity to known populations, outputting a confidence score for each assignment.

Hierarchical clustering is a frequently used method of clustering to produce a dendrogram display of the relationship of samples in a dataset. Hierarchical clustering differs from k-means clustering by its method of clustering, but also provides information on how closely each of the clusters are related to one another. In this thesis, 'agglomerative' clustering was performed, by which small clusters are merged into larger ones (essentially working from the bottom up on a dendrogram). This in in contrast to 'divisive' clustering methods, which splits larger clusters into smaller ones (working from the top down on a dendrogram).

The specific method of agglomerative hierarchical clustering used in this project was 'Ward's method'. This method aims to minimise the 'merging cost' of combining clusters by combining clusters with the smallest distance or sum of squares first. As with all hierarchical clustering, the sum of squares begins at zero as every point is assigned to its own cluster. Under Ward's method agglomeration, the 'closest' of these clusters are merged, with the centre of this new cluster used as the reference point for the next round of agglomeration. This continues until all clusters have been merged.

To read a hierarchical clustering dendrogram, such as that displayed in Figure 3.3, one should compare where and how samples are merged into clusters, and how resulting clusters are further merged together. The metric for assessing the difference between clusters is given along the Y-axis. 0 is the point at which every sample is contained within its own cluster. As these clusters are merged, the 'height' on the Y-axis increases in proportion to the difference between the clusters. In Figure 3.3, the individual samples cluster into their respective cell-specific clusters at a height between 50 and 100. This close relationship reflects the similarity of these samples. Agglomeration of the cell type clusters then occurs at a much greater height, reflecting greater differences between the cell type clusters. In Figure 3.3, the cDC1 and cDC2 subsets are merged at a height of approximately 170. These are therefore the two most closely related cell types in this dataset, closely followed by CD14 monocytes and CD16 monocytes. The monocyte populations and DC populations are merged last, at a height of 300, suggesting that these two populations are the most different clusters from one another.

Chapter 1 Figures & Tables

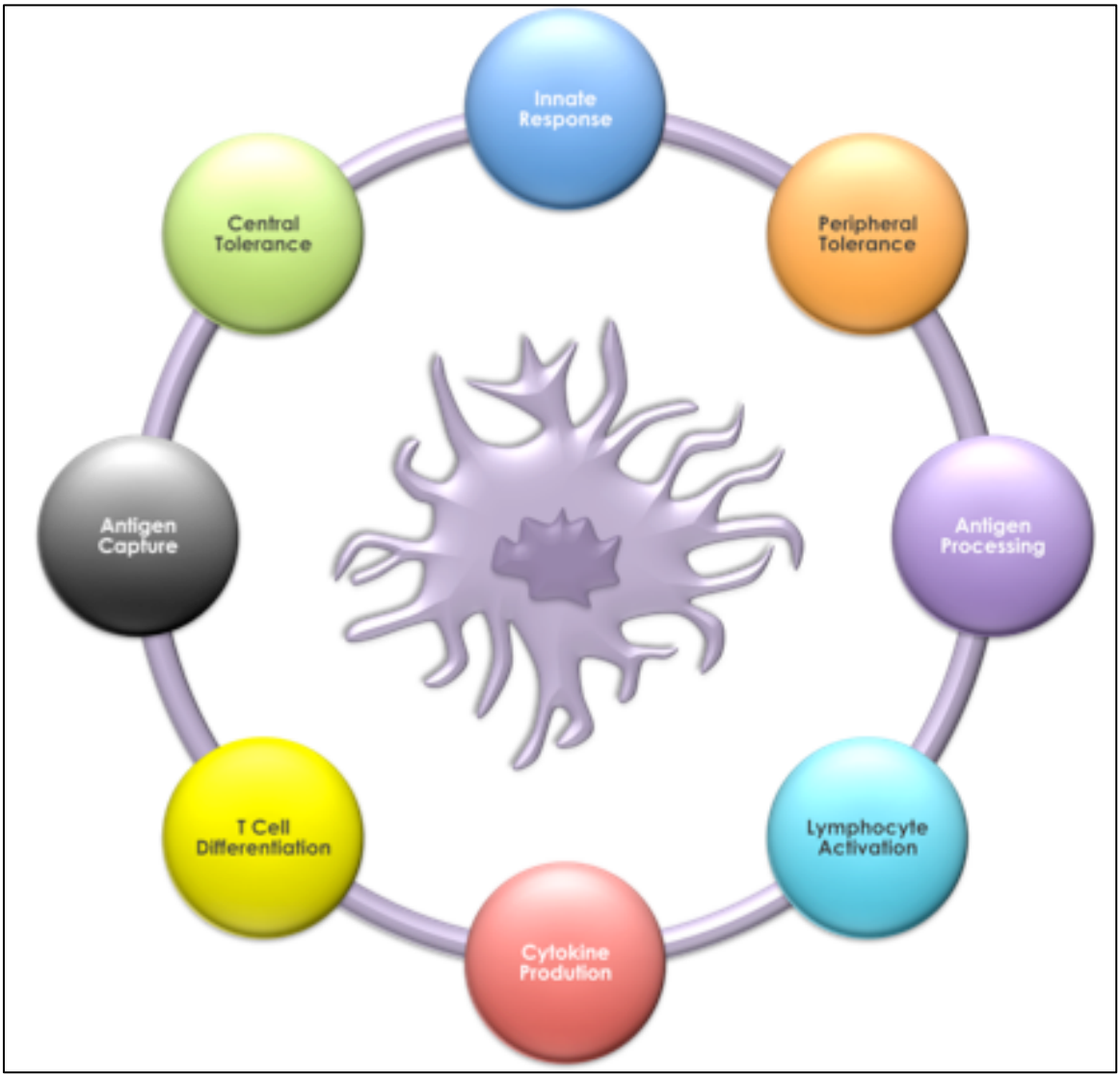


Figure 1.1: The functional roles of the dendritic cell

Adapted from Mak and Saunders (Mak and Saunders, 2005). This diagram highlights the broad spectrum of known functions associated with dendritic cells *in-vivo*. These essential immune cells process antigen, produce cytokines, activate lymphocytes and contribute to immune tolerance.

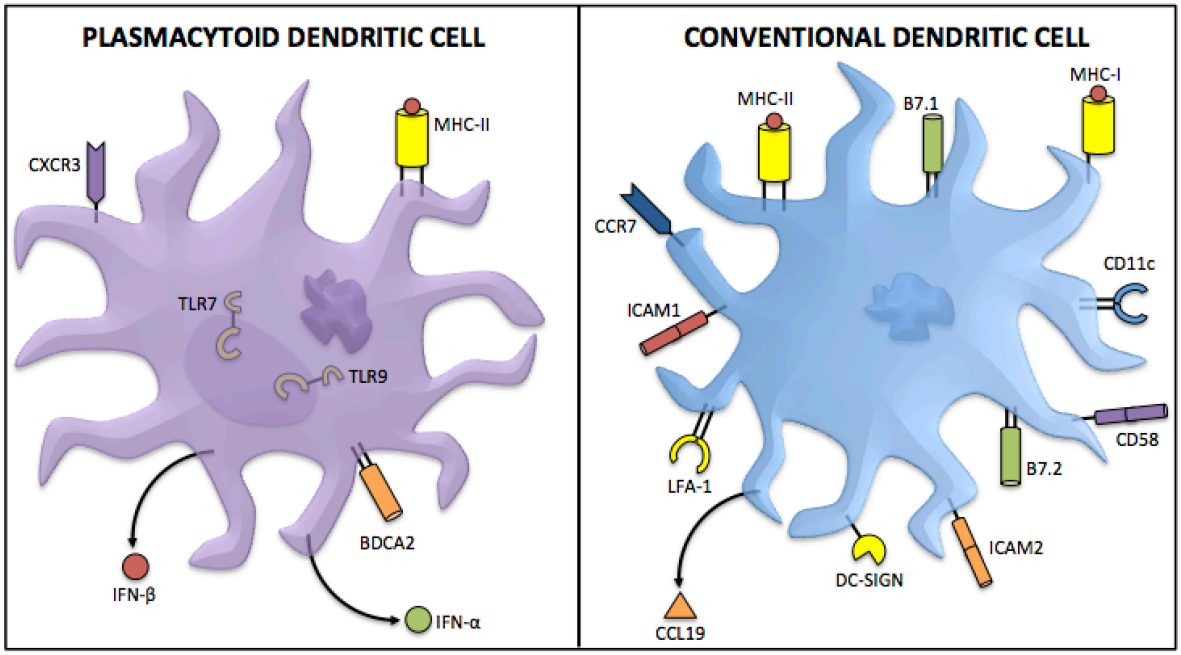


Figure 1.2: cDCs and pDCs have specific roles in immune response

Plasmacytoid DCs (pDCs) are primarily viral sentinels and secrete large amounts of class I interferons, particularly interferon-α. They do not have the same capacity for naïve T-cell priming, but do express TLR7 and TLR9 for sensing viral infections. Typical surface markers for the isolation of pDCs include CD123, CD303 and CD304.

Conventional DCs process antigens efficiently and once matured, express MHC molecules and co-stimulatory molecules to prime naïve T-cells. Mature cDCs can be identified by their expression of surface marker receptors, particularly CD141, CLEC9A and XCR1 on cDC1s and CD1c, CD11b and SIRPα on cDC2 cells. Immature dendritic cells lack many of these markers, but recognise pathogens through multiple Toll-like receptors. cDC1 cells express IFN-λ and IL12 and are particularly efficient as cross-presenting to CD8+ T-cells, particularly necrotic antigens, specialising in anti-tumour and anti-viral responses. cDC2 DCs are TNF-α, IL10 and IL12 producing cells and specialise in Th17 induction in response to fungal infection.

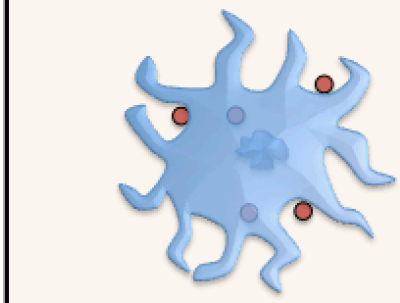
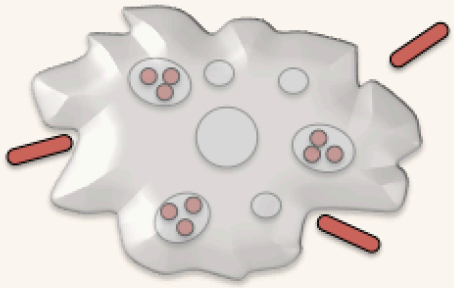
	DENDRITIC CELLS	MACROPHAGES
		
Antigen uptake	Macropinocytosis and phagocytosis by tissue resident dendritic cells	Macropinocytosis and phagocytosis
MHC expression	Low on tissue-resident dendritic cells High on lymphoid tissue dendritic cells	Inducible by bacteria and cytokines
Co-stimulation delivery	Inducible High on dendritic cells in lymphoid tissues	Inducible
Location	Ubiquitous throughout the body	Lymphoid tissue Connective tissue Body cavities
Effect	Results in activation of naïve T-cells	Results in activation of macrophages

Figure 1.3: Properties of dendritic cells and macrophages

Figure based on Murphy and Weaver, 2016. DCs and macrophages are two of the main cells involved in presentation of antigens to T-cells, along with B-cells (not shown). Dendritic cells are focused towards expansion and differentiation of naïve T-cells though the activation of naïve T-cells. Macrophages do not activate naïve T-cells, but instead present antigens to already primed T-cells specifically for the recruitment of effector T-cells that can release cytokines and enhance their own effector functions.

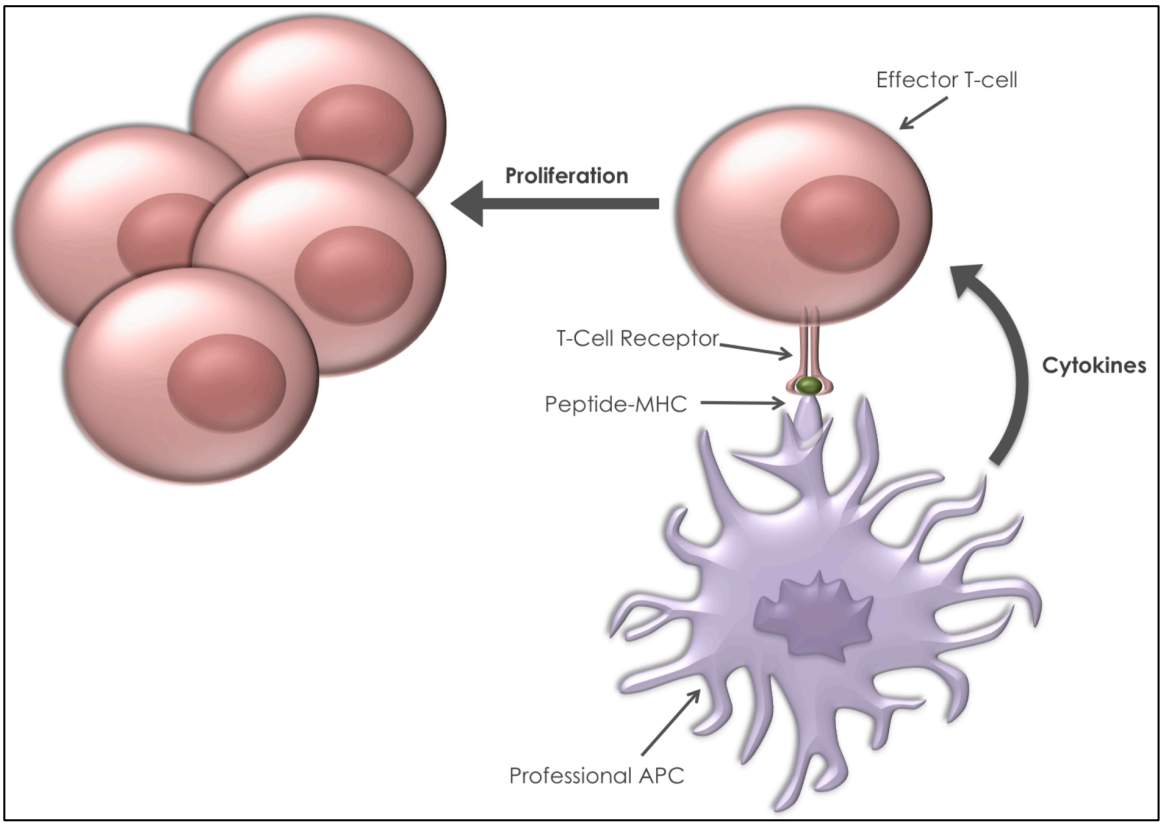


Figure 1.4: T-cell proliferation and antigen presentation under inflammatory conditions

Based on Bluestone and Abbas, *Nat Rev Immunology*, 2003. Effector T-cells (red) presented with foreign peptide by a professional antigen presenting cell (purple) in the presence of cytokine stimulation will undergo proliferation and induce an immune response.

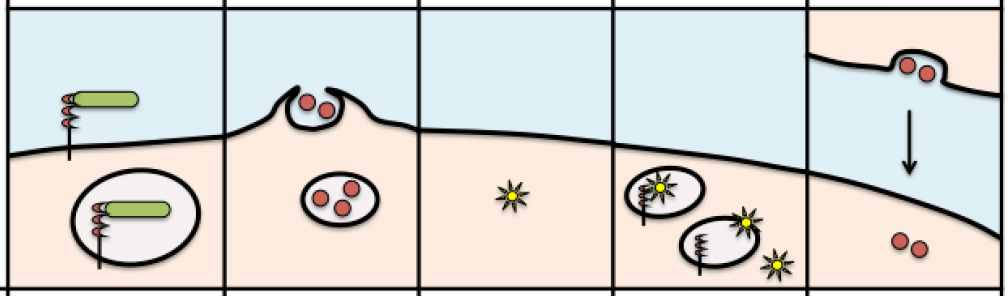
	Receptor-mediated phagocytosis	Macropinocytosis	Viral infection	Cross-presentation after phagocytic or macropinocytic uptake	Transfer from incoming dendritic cell to resident dendritic cell
					
Type of Pathogen	Extracellular bacteria	Extracellular bacteria, soluble antigens, viral particles	Viruses	Viruses	Viruses
MHC molecules loaded	MHC class II	MHC class II	MHC class I	MHC class I	MHC class I
Type of naïve T-cell activated	CD4 T-cells	CD4 T-cells	CD8 T-cells	CD8 T-cells	CD8 T-cells

Figure 1.5: Routes of antigen presentation and processing by dendritic cells

The major routes for delivering peptides to MHCII molecules for presentation to CD4 T-cells are via the uptake of antigens through receptor-mediated phagocytosis or by macropinocytosis. Antigen production in the cytosol, typically as a result of viral infection, is understood to be the major route for delivering peptides to MHC class I molecules for CD8 T-cells. Cross-presentation of exogenous antigens can also occur through the endocytic pathway after phagocytic uptake for delivery to MHC class I and eventual presentation to CD8 T-cells. Finally, research has shown that antigens can be transmitted directly from one dendritic cell to another, particularly for presentation to CD8 T-cells although the exact process for this is largely still not understood. (*Murphy and Weaver*, 2017.)

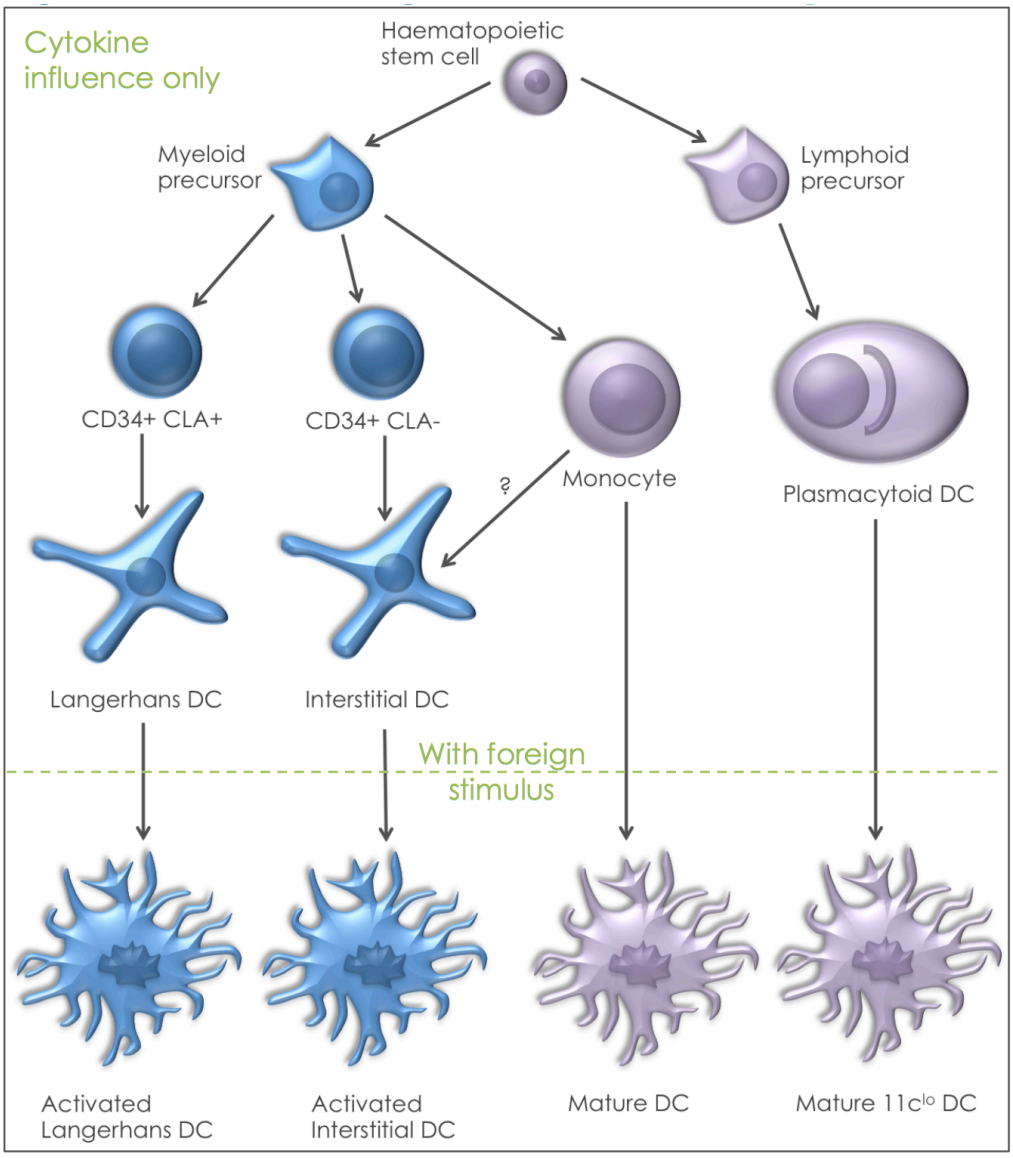


Figure 1.6: Dendritic cell lineage determined from *in vitro* experimentation (2002)

Based on the work of Shortman and Liu, *Nat Rev Immunology*, 2002 and others. Culture studies identified pDCs as lymphoid precursors, while monocytes were myeloid in origin. Some DC lineages were shown to develop into mature cells in the presence of cytokines, however, the terminal, activated stages of other lineages was reached only under the influence of antigen or foreign stimuli.

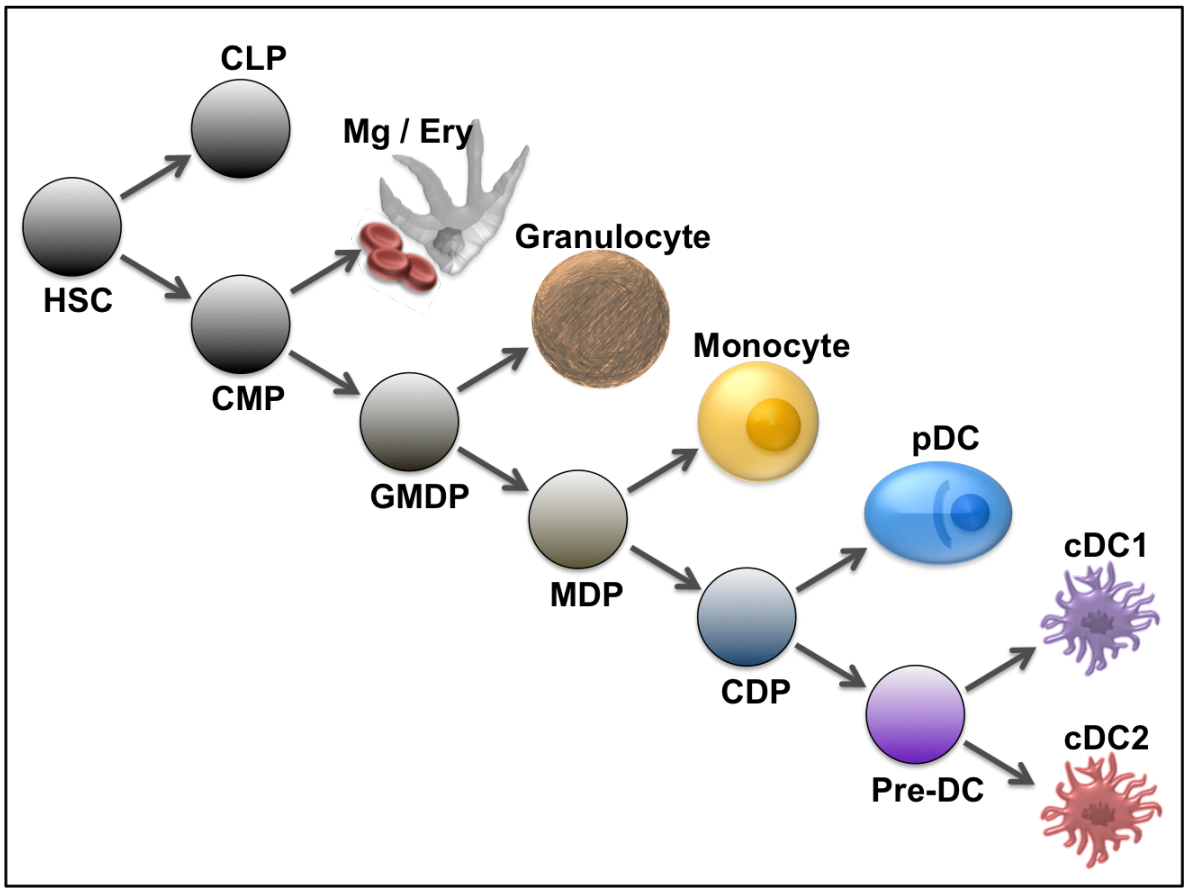


Figure 1.7: Linear bifurcation model of haematopoietic development

The linear bifurcation model postulates a step-wise progression through stable multi-potent progenitor cell types. Starting at the Haematopoietic stem cell (HSC), a myeloid or lymphoid fate decision pushes a cell towards a common lymphoid progenitor (CLP) type or a common myeloid progenitor (CMP). The CLP will develop into lymphoid cells through further bifurcation steps. The CMP branches into either megakaryocytes and erythrocyte (Mg / Ery) cell types, or into a more restricted granulocyte macrophage dendritic cell precursor (GMDP), which in turn produces a macrophage DC progenitor (MDP), followed by a common DC progenitor (CDP) and into a pre-DC stage with cDC1 or cDC2 potential. In this model, each stage has homogeneous potential.

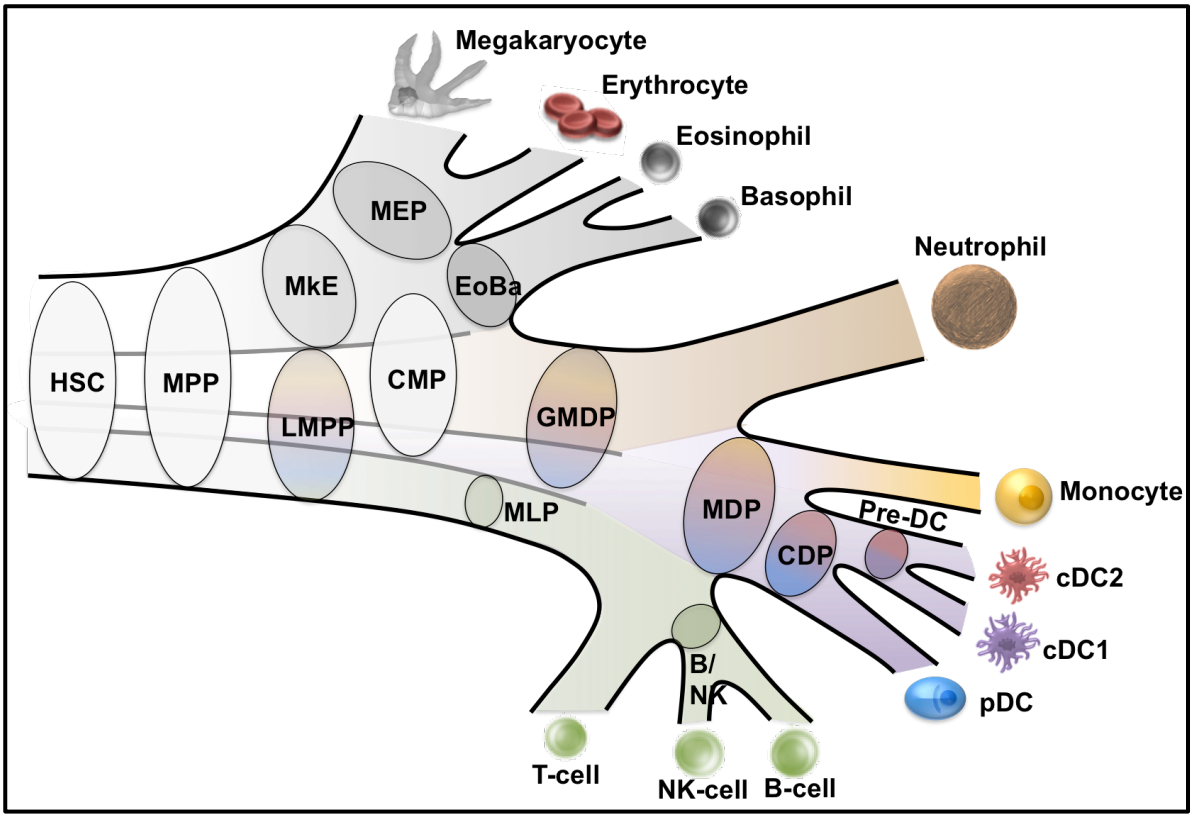


Figure 1.8: Early priming model of haematopoietic development

Adapted from Collin and Bigley, *Immunology*, 2017 (in press). Experimental data supports early lineage priming of cells at the early progenitor stage such that most populations contain cells with phenotypically related but uni-potential capacity. In this model, the lymphoid and myeloid pathways run in parallel from a mixed lymphoid primed multi-potent progenitor (LMPP), separated from megakaryocyte and erythroid potential (MkE). Subsequently, the granulocyte macrophage dendritic cell precursor (GMDP), macrophage DC progenitor (MDP), and common DC progenitor (CDP) populations are also composed of heterogenous, uni-primed, phenotypically similar cells comprising increasingly restricted mature cell potential.

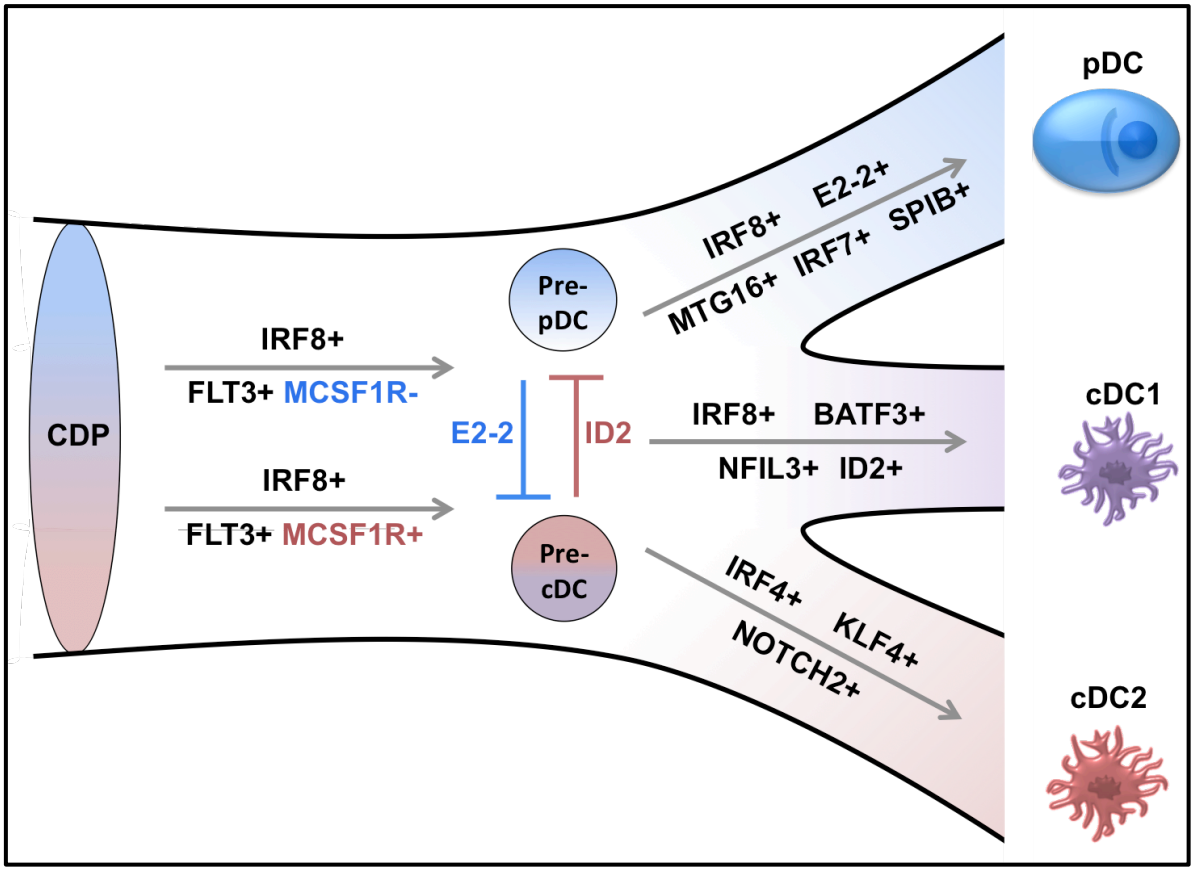


Figure 1.9: Major transcription factor requirements in dendritic cell development

Displayed are the major transcription factor requirements for DC lineage development. Acquisition or loss of one or more transcription factors can drive, reinforce or redirect lineage. E2-2 and ID2 act in opposing roles to push progenitors towards pDC or cDC potential.

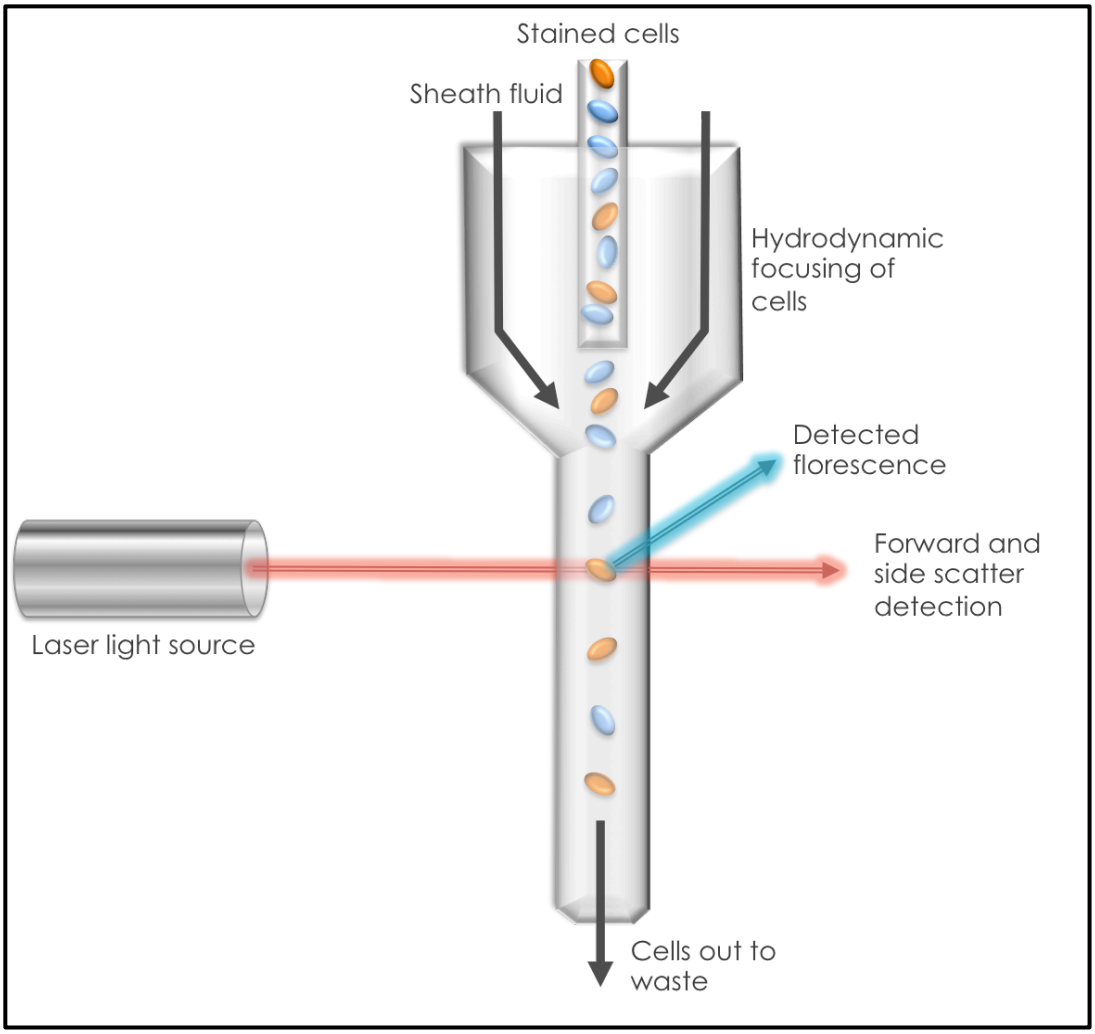


Figure 1.10: The Flow Cytometer

Flow cytometry induces hydrodynamic focusing of a cell suspension by using fast flowing sheath fluid to create a drag effect on the central chamber. This effect pulls the sample suspension into a narrow channel, forcing cells into a single file. Using a number of laser light sources of varying wavelengths individual cell characteristics can be identified, including cell size, granularity and the presence of pre- strained surface markers. These multi-dimensional parameters allow a user to identify groups of cells as well as any potential outliers at a rate of tens of thousands of cells per second.

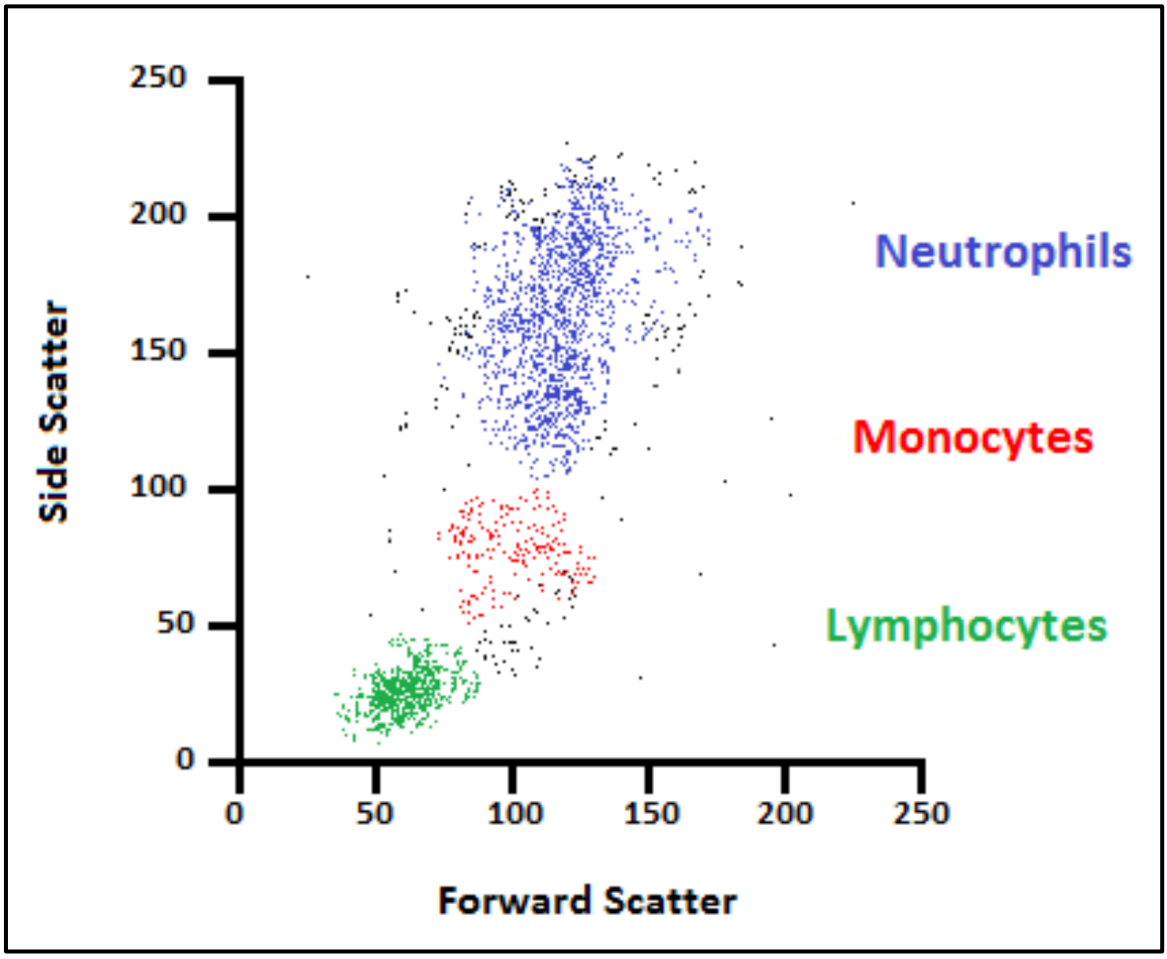


Figure 1.11: Classification of major blood cell types by flow cytometry using forward and side scatter

Adapted from Rowley, *Materials and Methods*, 2015. This flow cytometry plot of side scatter against forward scatter produces three distinct clusters identified as blood cells of three lineage types: large, highly granular neutrophils (blue); smaller, less granular monocytes (red) and very small lymphocytes (green). As flow cytometry interrogates cells individually, each cell can be displayed as its own point on the plot.

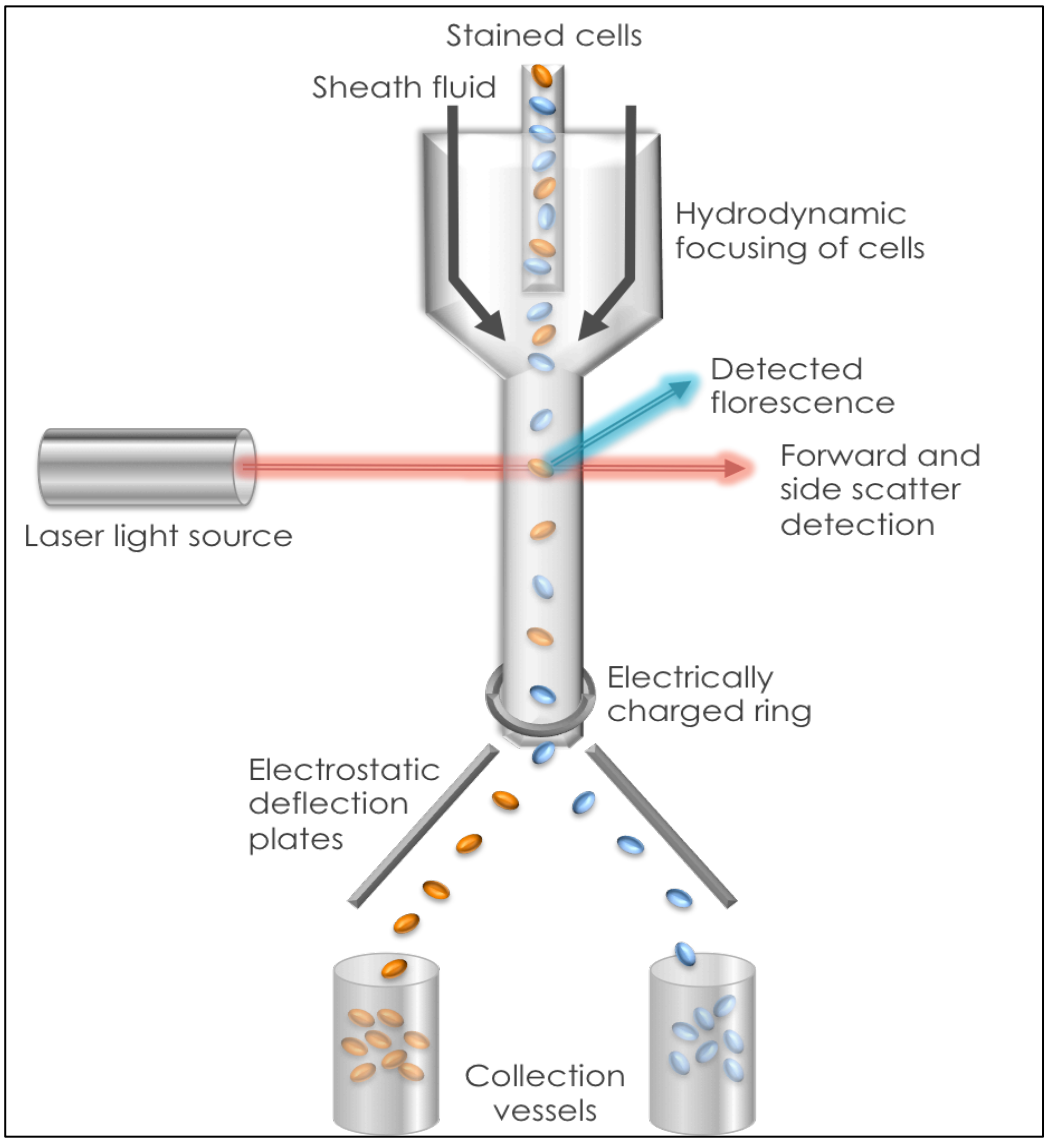


Figure 1.12: The Fluorescence-Activated Cell Sorter (FACS)

Using the same hydrodynamic principles as regular flow cytometry, FACS sorters use laser light and fluorescence to characterise individual cells in up to twenty dimensions. Once a cell has been analysed, an electrically charged ring induces polarisation of the cell, either positively or negatively. Using the principles of electrostatic attraction, deflection plates pull the cell and a small amount of its surrounding fluid towards one of a number of collection vessels based on the induced charge of the cell. By this mechanism, cells with a similar phenotype can be separated out and purified from a heterogeneous mix.

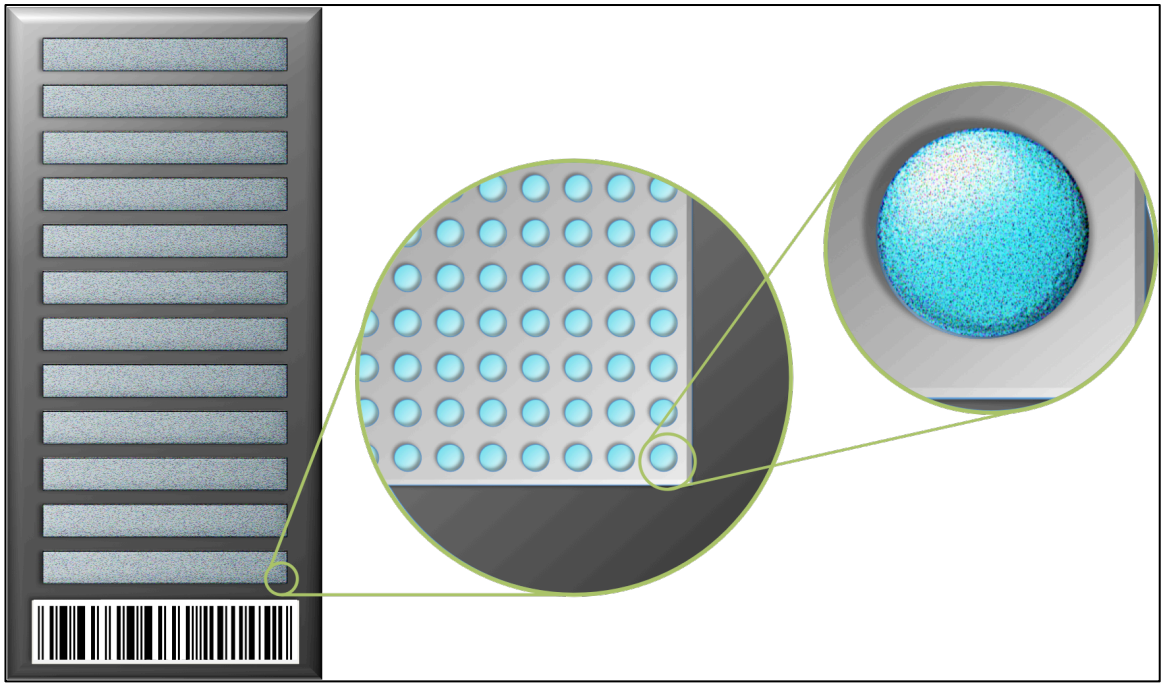


Figure 1.13: Schematic representation of Illumina BeadArray Technology

Illumina bead array technology uses silica beads arranged in a random fashion and loaded with hundreds of thousands of duplicate probes to analyse up to twelve samples at once. Complementary binding of gene targets to the probes induces a fluorescence response, which is detected and scored. More probe binding will induce a greater fluorescent response and thus relative gene expression levels can be determined. With over 47,000 bead types, Illumina BeadArray technology can identify and infer the relative expression of over 30,000 genes per sample.

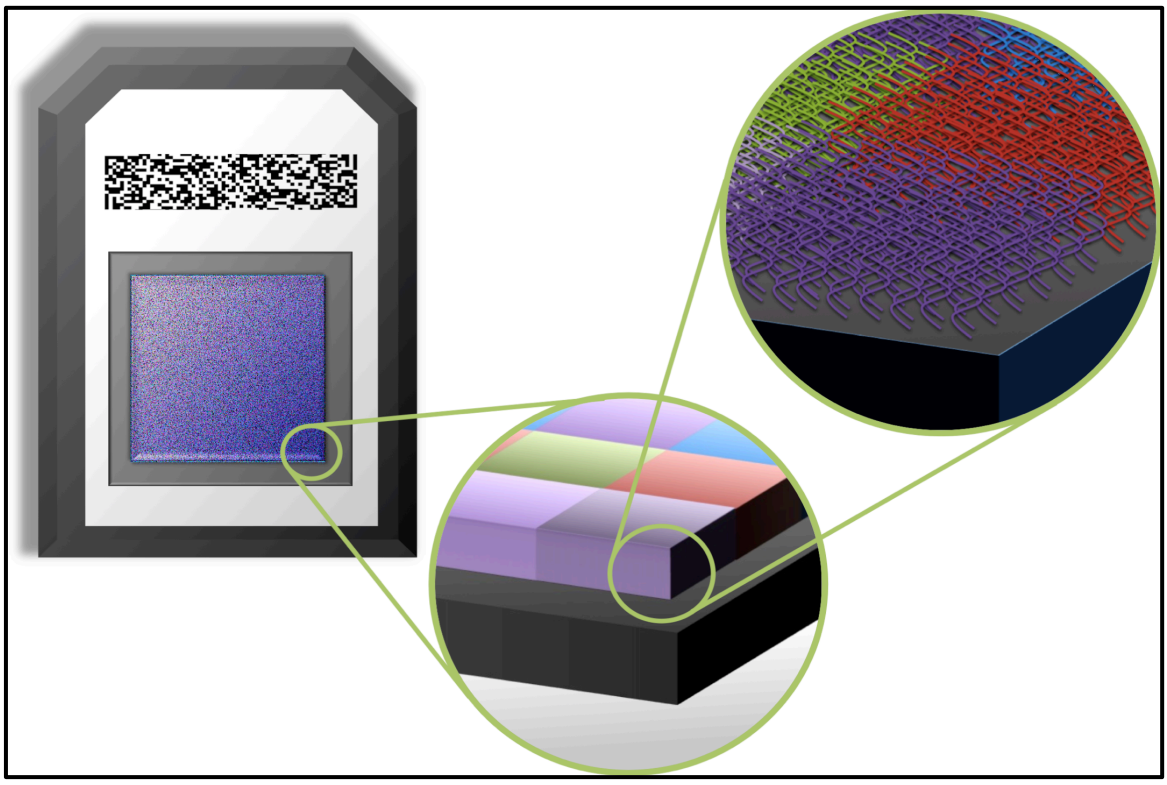


Figure 1.14: Schematic representation of Affymetrix GeneChip Technology

Affymetrix arrays use single-sample chips divided into over 400,000 squares. Each of these squares contains a unique DNA strand probe with over 1 million individual copies. Manufacturing of these plates using photolithography and hundreds of specific UV masking plates allows millions of probes to be synthesized at once, one base at a time. This process also means that each type of probe will be created only in one specific region.

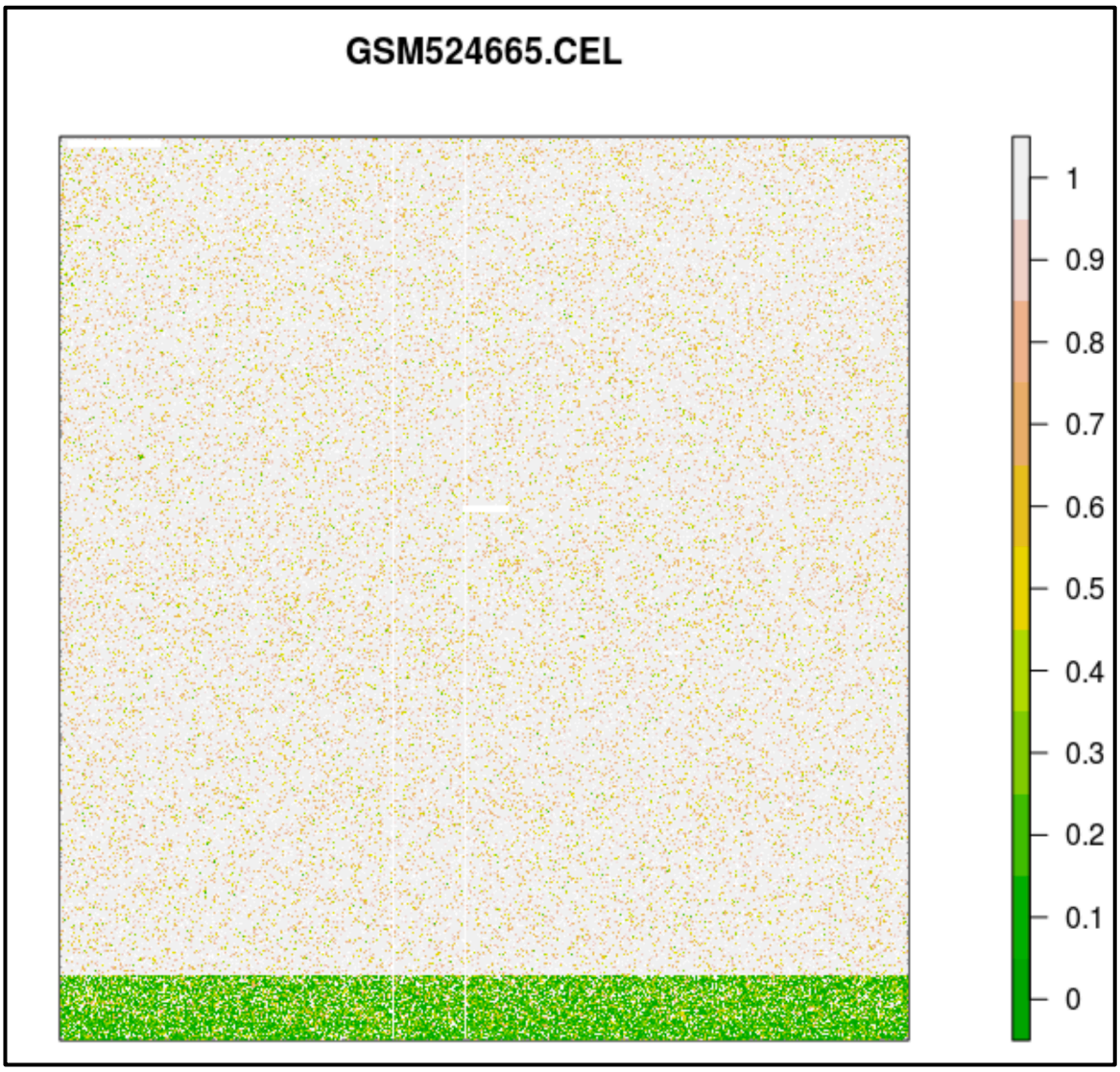


Figure 1.15: Affymetrix array CEL file analysis from publically available data, indicating region-specific anomaly (GSM524665)

This CEL file image taken from mouse Affymetrix data (GSM524665) highlights one of the major drawbacks of Affymetrix array design. A CEL file is essentially a false-colour map of an Affymetrix GeneChip based on the fluorescence intensity of each region. As shown in green on the figure, the lower section of this chip is showing an anomaly with extremely low or no probe detection. Such an issue might be the result of machine error or damage to the sensitive array plate. While other expression technologies could suffer a similar issue of region specific anomalies, with Affymetrix, region specific design of probe coverage will mean all probes for a given gene can be lost at once.

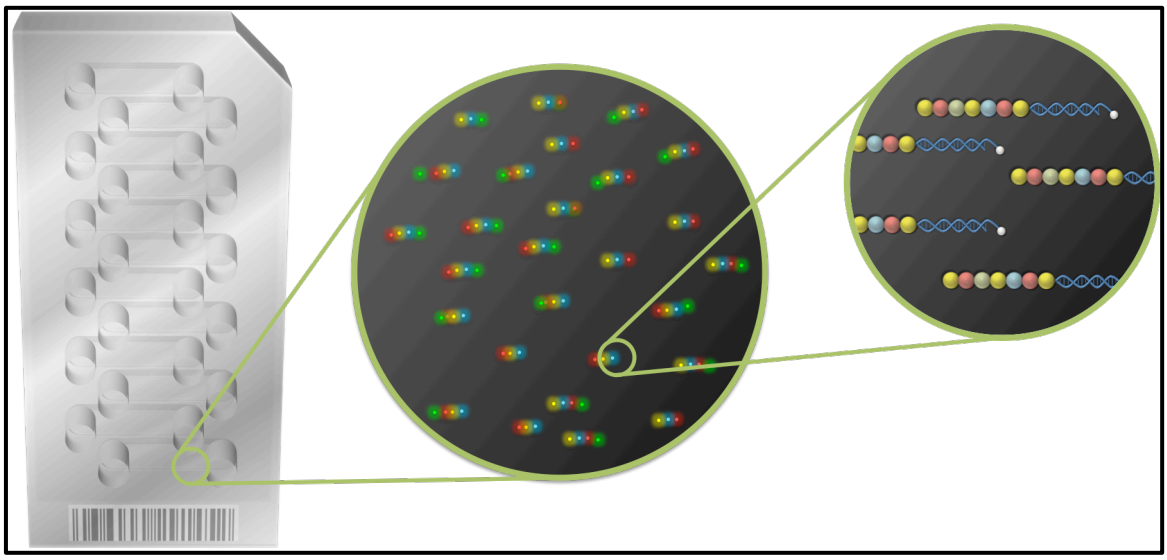


Figure 1.16: Schematic representation of NanoString nCounter Array

Fluorescence-based NanoString expression technology uses a chain of fluorescent tags as ‘barcodes’ for a specific mRNA transcript of interest. Over 800 genes can be assessed per sample in this manner for up to 12 samples per cartridge. As NanoString Technology does not rely on signal intensity or fluorescence scores, there is very little bias introduced. Instead, NanoString uses biotin molecules to bind captured mRNA hybridised-reporter probe strands to the streptavadin-coated surface of the cartridge. Once linearised by induction of an electrical gradient, the ‘barcodes’ can be optically detected by a computerised camera and microscope. From this, individual mRNA strands can be detected and identified as representing a specific gene. Each captured probe is counted to give the actual number of mRNA strands that were captured in each sample. This highly specific direct profiling and digital detection process produces large quantities of robust data in a short period of time.

Chapter 2: MATERIALS AND METHODS

2.1 ETHICAL STATEMENT AND APPROVALS

All healthy donor samples used in this thesis were obtained following written, informed consent with the approval of NRES Committee North East – Newcastle and North Tyneside (14/NE/1136 and 14/NE/1212). Plastic surgery skin was obtained from the Newcastle Biobank under 08/H0906/95+5. Illumina expression data was extracted from GEO repository with the accession number GSE35457. Research ethics for this published study were obtained from Singapore Singhealth and National Heath Care Group Research Ethics Committees and are discussed further in the original publication (Haniffa et al., 2012).

2.2 BUFFERS AND REAGENTS

2.2.1 Lymphoprep™ Solution

Lymphoprep™ solution (Axis-Shield Diagnostics Ltd.) is a density gradient medium used for the isolation of mononuclear cells from peripheral blood or bone marrow. It contains sodium diatrizoate (9.1% w/v) and polysaccharide (5.7% w/v) in water, resulting in a density of 1.077g/ml. Due to their greater density, erythrocytes and granulocytes pass through the medium and settle below it, while lower density mononuclear cells float on top.

2.2.2 Dulbecco's Phosphate-Buffered Saline (PBS)

PBS (ThermoFisher Scientific Inc.) is an isotonic salt solution usually containing sodium and potassium compounds. These buffers are non-toxic to most cells and are therefore commonly used for dilutions and cell-washing protocols

2.2.3 Flow Buffer

Flow buffer is used as a dilutant during flow cytometry protocols. The main component of flow buffer is D-PBS. The addition of 2% by volume fetal calf serum (FCS) introduces proteins necessary for the reduction of heterophilic (non-specific) antibody interference and maintains cell viability, while 0.004% by volume EDTA reduces cell clumping during the cytometric process.

2.2.4 Sort Buffer

Sort buffer is used as a dilutant during fluorescence-activated cell sorting (FACS) protocols. The main component of sort buffer is also D-PBS. 0.5% by volume fetal calf serum (FCS) introduces proteins necessary for the reduction of heterophilic (non-specific) antibody interference and maintains cell viability. The addition of 0.004% by volume EDTA reduces cell clumping during the sorting process.

2.2.5 Culture Mediums

RPMI-1640 is a liquid cell culture medium containing L-glutamine and sodium bicarbonate and commonly used to support lymphoblastoid cells or anchorage dependent cells.

RF-10 was used to collect cells from FACS for NanoString analysis in Chapter 3 and Chapter 4. It is composed of RPMI 1640 media with the addition of 10% FCS, 1% penicillin-streptomycin and 1% glutamine.

aMEM (Gibco™) is a culture medium solution based on Minimum Essential Medium with the inclusion of non-essential amino acids; sodium pyruvate, lipoic acid, ascorbic acid, biotin and vitamin B12 to support a wider range of cell types. aMEM was used in the culture of DCs from CD34+ bone marrow cells used in Chapter 4 and further supplemented with 1% penicillin/streptomycin (Sigma), 10% Fetal Calf Serum (Gibco), 20ng/ml GM-CSF (R&D systems), 100ng/ml Flt3-ligand (Immunotools) and 20ng/ml SCF (Immunotools).

2.2.6 NanoString nCounter™ Hybridisation Buffer

nCounter hybridisation buffer is a sodium chloride containing buffer medium used to stabilise NanoString nCounter codesets and target material during hybridisation.

2.2.7 NanoString nCounter Codesets

NanoString nCounter codesets consist of a probe pair designed against a target of interest. The reporter probe carries a string of fluorophores at its 5' end, the sequence of which determines the target gene. The reporter signal complexity carries four colours in six positions. Up to 900 individual reporter probes can be multiplexed and hybridised at once.

The capture probe is bound to a biotin molecule at its 3' end. During the hybridisation step, the capture and reporter probes bind to a 100bp target region of RNA, DNA or protein. The capture probe's biotin molecule binds it to the reading surface of a NanoString cartridge while the reporter probe's fluorophore sequence allows the hybridised material to be individually resolved and identified.

2.2.8 RNA Lysis Buffer (RLT)

RNA lysis buffer is a guanidium ISO thiocyanate-containing buffer used in the preservation and stabilization of RNA. It readily denatures protein and RNAses that would otherwise degrade RNA. RLT with the addition of 1% β-mercaptoethanol was used in the NanoString hybridisation protocol for Chapter 3 and Chapter 4. RNA lysis buffer consisting of RNAse free water, 0.2% triton X (Sigma) and 2U/µl RNAse inhibitor (Sigma) was used in the single cell hybridization protocol in Chapter 5.

2.3 PROTOCOLS FOR SAMPLE PROCESSING AND STORAGE

2.3.1 Bone Marrow Cell Isolation

Bone marrow was obtained from donations after hip replacement operations. The core of the bone was washed with RPMI though a cell filter and into a falcon tube. The bone marrow was then scraped from the bone using bone clippers before a second washing step. Unfiltered tissue was pulped in the cell filter and washed with RPMI.

Wash-through was diluted with PBS at a 1:4 ratio and layered over lymphoprep solution at a ratio of two-parts diluted sample to one-part lymphoprep. This was then centrifuged for 15 minutes at 800g to separate the blood constituents along a density gradient.

The cell layer was aspirated into a sterile falcon tube and topped up with PBS prior to two centrifugation-washing steps for 5 minutes at 500g.

Excess PBS was poured off the cell pellet, which was then resuspended in 1mL of PBS. 10µl of this suspension was used to count viable cells using a haemocytometer. The remaining cells were centrifuged for 5 minutes at 500g. The PBS was poured off and the cells were resuspended in sorting buffer for downstream staining and FACS as noted in sections 2.3.1 and 2.4.1.

2.3.2 Peripheral Blood Cell Isolation

Peripheral blood was obtained from consenting patients or healthy volunteers. The blood was diluted with PBS at a 1:2 ratio and layered over lymphoprep solution (STEMCELL™) at a ratio of two-parts diluted sample to one-part lymphoprep. This was then centrifuged for 15 minutes at 800g to separate the blood constituents along a density gradient.

The cell layer was aspirated into a sterile 50ml falcon tube and topped up with PBS prior to two centrifugation-washing steps for 5 minutes at 500g. Excess PBS was poured off the cell pellet, which was then resuspended in 1mL of PBS. 10µl of this suspension combined with 10µl of trypan blue (ThermoFisher) was used to count viable cells in a haemocytometer. The remaining cells were used in downstream steps including FACS sorting as detailed in 2.4. Cells surplus to requirement were resuspended in freezing solution, transferred into a cryovial and stored at -80°C.

2.3.3 Dermal Cell Isolation

Dermis was obtained from plastic surgery operations with written informed consent. The sample was first placed into a large petri dish with 5ml of PBS. Forceps and a scalpel were used to cut the skin into equal strips. Each strip was then pinned to a Teflon-coated corkboard and split-thickness sections taken with a sterile Webster skin knife with a size 8 guard. The split strips were floated on 10ml RPMI with 100µL dispase (1:100 ratio) and incubated at 37°C for 1 hour, after which the epidermal layer was peeled from the dermis using forceps. Both parts were washed in fresh RPMI to remove remaining dispase.

For digestion, dermis strips were placed in XVivo10 with the addition of Worthington's collagenase at a ratio of 1:150. Epidermal strips were placed in XVivo10 with the addition of Worthington's collagenase at a ratio of 1:200. Both were then incubated at 37°C overnight to digest. The final solutions were flushed through a 100 micron cell strainer and washed twice with PBS before use in FACS sorting experiments.

2.4 CELL CULTURE

2.4.1 Culture Conditions and Sorting

$CD34^+$ bone marrow progenitor cells were sorted by FACs to >98% purity and seeded at 3,000 cells per well onto a pre-seeded (>/=4hrs prior) feeder layer of OP9 stromal cells at 5,000 per well in 96 well U-bottomed plates. Cells were cultured in 200mL aMEM (Gibco™) with 1% penicillin/streptomycin (Sigma), 10% Fetal Calf Serum (Gibco), 20ng/ml GM-CSF (R&D systems), 100ng/ml Flt3-ligand (Immunotools) and 20ng/ml SCF (Immunotools). Half the volume of media (including cytokines) was replaced every seven days. At Day 21 cells were harvested on ice, passed through a 50mm filter, washed and stained for flow cytometric analysis or FACS.

2.5 FLOW CYTOMETRY AND SORTING

2.5.1 Cell Staining and Sorting

Cells were stained in aliquots of up to $1x10^7$ cells in 100µl of DPBS with 2% fetal calf serum and 0.4% EDTA. Dead cells were excluded by DAPI (Partec). Cells were sorted with a FACSAria III (BD Biosciences) running BD FACSDIVA™ 8.0 software, into eppendorfs with appropriate media as detailed in Chapter 4, section 4.2.2.

2.5.2 Sorting Strategy

Major DC and monocyte subsets were sorted according to the parameters displayed in and Table 2.1. Further chapter specific sorting strategies are displayed in the relevant chapters.

2.6 ILLUMINA BEAD ARRAY

Generation of the Illumina expression data used in this thesis was conducted previously in the lab and published in Haniffa *et al*, 2012 *'Human Tissues contain CD141hi cross-presenting dendritic cells with functional homology to mouse CD103+ nonlymphoid dendritic cells'.* The full dataset can be obtained from the NCBI GEO repository at *www.ncbi.nlm.nih.gov/geo/* under the accession number GSE35457.

2.6.1 Machine Specifications and Protocols

Illumina BeadArray readers were used for all Illumina BeadChip experiments, utilizing Illumina Human HT-12 v.4-0 Whole Genome gene expression arrays. Cell subsets used for transcriptomics analysis were purified and collected using FACS sorting. The cells were processed using Qiagen RNeasy Mini kits to extract RNA according to manufacturer guidelines. The extracted RNA was checked for quality and quantified using an Agilent Bioanalyzer. Total RNA samples were then amplified and subsequently biotinylated using Illumina TotalPrep RNA Ampification kits and finally processed according to standard BeadChip array protocols. A full methodology for this data is available in Haniffa *et al*, 2012.

2.6.2 Data Normalisation Protocol

Illumina BeadArray data were pre-processed using Illumina Genome Studio and were normalized via the *lumi* and *limma* packages in R, using the Loess method or variance stabilizing transformation (VST) and robust spline normalisation. This normalisation was re-performed for this thesis prior to further downstream analysis. Genes failing detection threshold QC (p-value <= 0.05) were removed from further analysis at the normalisation stage.

2.7 NANOSTRING NCOUNTER ANALYSIS PLATFORM

2.7.1 Machine Specifications

All NanoString experiments were performed on a 2nd Generation 'Flex' system, utilising a 'prep station' running software version 4.0.11.2 and 'high sensitivity' settings. The 'digital analyser' was set at 555 fields of view, running software version 3.0.1.4.

2.7.2 NanoString Codesets

NanoString experiments were performed using pre-built, focused nCounter gene expression panels created by NanoString.

NanoString Human Immunology_V2 mRNA codesets were used for most experiments with or without the addition of a 30 gene 'Panel+' custom codeset.

2.7.3 Gene Expression Hybridisation Protocol

Cell subsets used for NanoString analysis of Chapter 3 and Chapter 4 were purified and collected using FAC sorting, pelleted and stored in RNA lysis buffer (described in 2.1.8) at a concentration of 2,000 cells/µl. Cell lysates were directly hybridised to the NanoString CodeSets according to manufacturer guidelines for cell lysate protocol.

For the initial comparison to extracted RNA described in Chapter 3, figure 3.10, an equal number of cells were processed using Qiagen RNeasy Mini Kits or Micro Kits to extract RNA according to manufacturer guidelines. RNA content was quantified using an Agilent Bioanalyzer and normalised to a consistent concentration with the addition of nuclease-free water. Extracted RNA was hybridised to the NanoString Codeset according to manufacturer protocols.
For both the extracted and lysate protocol, the process involved involved the addition of 20µl of diluted NanoString reporter codeset to each well of a 12-well strip tube, followed by 5µl of RNA-content normalised sample material. Finally, 5µl of NanoString capture codeset was added to each well before the samples were hybridised for 18 hours at 65°C and moved into the NanoString Prep-Station for automated processing, followed by the NanoString Digital Analyser for counting.

2.7.4 Data Normalisation Protocol

Normalisation of NanoString data was performed using NanoString's freeware 'nSolver' version 2.6 with manufacturer standard normalisation procedures. Each pre-built human NanoString CodeSet includes eight 'negative control' probes that do not bind to any human RNA region. The geometric mean of the counts of these probes is removed from the count number of all endogenous probes to account for random background binding. Six scaled 'positive controls' were spiked into the codeset at known concentrations to control for any differences in the amount of codeset added to each well. Again, the geometric means of these counts were used to provide a normalisation factor. The POS_F control was also used for quality control proposes to determine if the lowest concentration of the positive controls can be seen above the level of background binding.

A final round of normalisation was performed against a number of well-known 'housekeeping' gene targets, including GAPDH, RPL19 and EEF1G. The geometric mean counts of these housekeeping genes were used to compute a normalisation factor that was then applied to all endogenous gene probe counts to account for inter-sample variation. After each normalisation factor was applied, the final count data was considered 'normalised'. Log2 transformation of the data was applied after normalisation to normalise the distribution of gene intensity values.

2.8 DATA ANALYSIS

2.8.1 Combining Batches (ComBat)

The issue of batch effects in the Illumina expression data was addressed through the implementation of the 'R' package 'SVA' with the 'ComBat' function. Firstly, a tab-delimited text file of normalised expression values was loaded into the 'R' statistical programme. A second file was loaded in, containing information on the dataset, including which samples were from each batch. The ComBat function adjusted the data to counter the effects of any variation linked to sample batches using an empirical Bayesian framework and returned a corrected copy of the dataset (Johnson et al., 2007).

Frequently, microarray experiments run across multiple array chips, the datasets from which then require combining for analysis. Combining datasets without adjusting for batch effects between cartridges would result in unaccounted for bias during analysis (Leek et al., 2012).

2.8.2 Statistical Testing

A number of context-dependent methods of statistical analysis techniques were performed in this thesis. In all cases, a p-value cut-off of 0.05 was used in combination with False Discovery Rate (FDR) adjustment of the p-values where necessary.

The novel approach to culture and tissue signature removal from a dataset applied in chapter 3 and chapter 4 was based on a standard t-test applied across two populations grouped by condition. In chapter 3, samples were grouped into skin-derived or blood-derived samples before a t-test was applied across these two groups. The novelty lay with the purpose of the test. Rather than identifying and focusing on the differential expressed genes, as would be typical in differential expression analysis, the genes identified as differentially expressed between skin and blood samples were removed from the dataset, it leave only those genes conversed across both tissues. This revealed the underlying cell type associations that were previously obscured by the stronger tissue-specific differences between the populations. In chapter 4, the same approach was implemented to remove culture-specific genes and effects from the dataset, again, revealing the similarity between the mononuclear cell subsets that was previously hidden behind a strong culture-specific signature. These approaches both opened up the project to a much deeper interrogation of the dataset than would otherwise be possible and exposed the underlying patterns within the data, through the condition-associated effects.

2.8.3 Principal Component Analysis (PCA)

PCA was performed using the 'stats' package of 'R' ("R Core Team," 2012) and incorporating 'ggbiplot' ("V.Q. Vu," 2011), 'ggplot2' (Wickham, 2009a) and 'RColorBrewer' ("E. Neuwirth," 2014) for visualisation and grouping.

2.8.4 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE was performed in 'R' using the 'Rtsne' package (Maaten and Hinton, 2008). A sample seed was set to ensure reproducible results over multiple iterations and perplexity was selected based on expected grouping of samples.

2.8.5 Gene Set Enrichment Analysis (GSEA) and BubbleGUM

Gene Set Enrichment Analysis was performed using the GSEA functions developed by the Broad Institute of MIT (Subramanian et al., 2005) incorporated into the BubbleGUM GUI interface developed at the Centre d'Immunologie de Marseille-Luminy (Spinelli et al., 2015). Gene signatures used in Chapter 3, 4 and 5 were developed according to this strategy with a False Discovery Rate (FDR) adjustment cut-off p-value of 0.05 and $\log_2$ fold change of 1.5.

2.8.6 Minimal Gene Reduction and Dimensionality Reduction

A number of different dimensionality reduction techniques were employed in this thesis, depending on the required output. These are discussed in greater details in the Chapter-specific introduction and method sections, but overviewed here.

T-SNE analysis was performed using 'Rtsne' implementation of Barnes-Hut t-Distributed Stochastic Neighbor Embedding (Maaten and Hinton, 2008) with 50 initial PCA dimensions and variable perplexity based on the dataset, commonly in the range of 3 to 10 with a maximum number of iterations set at 1,000.

K-means clustering is one of the simplest unsupervised learning algorithms available for dividing data into a pre-determined number of clusters and was performed with 1,000 iterations and a k value of 4, representing the 4 expected cell populations in the data. In this k-means clustering, 4 initial centroids are mapped onto the space represented by the samples being clustered. Each sample is then assigned to the nearest centroid to it. Where total within-cluster variation is minimized. Once all samples have been assigned to centroids, new centroids are calculated at the average position of each of the samples in that cluster, thereby separating the samples. This process is repeated until the centroid clusters no longer move through the iterations and thus all points in a cluster are homed in to one point.

The novel custom gene reduction method produced in this thesis was based on repeated k-means testing with random gene removal as displayed in Figure 3.24. The novelty of this approach stems form the combination of events in the pipeline. Utilising repeated k-means testing in this way has not been described in literature before. K-means testing was performed on all genes and the between-group sum-of-squares were recorded for each group. This provided an indication of the 'distance' or separation between cluster. One gene was randomly removed and the process was repeated. If the new (smaller) geneset provided greater between-cluster sum-of-squares, another random genes would be removed and re-checked. This would mean that the reduced dataset separated the clusters to a greater extent then the full dataset. If the new geneset did not provide a greater sum-of-squares, the gene would be replaced and another gene randomly removed and re-analysed as the removed gene was likely driving the cluster separations and may therefore be a cluster-specific marker gene. This process was repeated for 10,000 iterations or until no more genes could be removed without disrupting the cell type clustering provided by k-means testing. Final gene lists were produced at this stage and displayed using heatmaps and clustering functions in 'R'. The entire process was repeated 2,000 times to produce 2,000 final minimised gene lists that were capable of separating the cell subsets.

2.8.7 Functional and Pathway Mapping

Functional analysis was performed using the 'GOstats' package (Falcon and Gentleman, 2007) hypergeometric testing functions with an FDR adjusted p-value of 0.05 deemed to be significant and a requirement of over five genes to share a function for it to be considered.

2.8.8 Single Cell RNA-Seq Analysis Pipeline

Plate-based single cell analysis involved a specialised pipeline of pre-processing as outlined in Figure 5.1. Freshly isolated PBMC were index-sorted as single cells using a FACSAria III with a 100um nozzle into a 96 well V bottomed plate, each well containing 2µl RNA lysis buffer consisting of RNAse free water, 0.2% triton X (Sigma) and 2U/µl RNAse inhibitor (Sigma). The cell identification and sorting strategy used is shown in Chapter 5, Figure 5.2. Plates were spun at 500g and 4°C for 1 minute, then frozen on dry ice and stored at -80°C until processing at the Oxford Genomics Centre.

Reverse transcription, library prep and sequencing steps were all performed at Oxford. To adjust for technical variability inherent to the reverse transcription process, ERCC spike-in control sequences were incorporated into each sample well prior to reverse transcription according to a non-UMI adapted SMARTseq2 protocol (Picelli et al., 2013). A Nextera XT DNA Library Prep kits was used for library prep with reads generated on an Illumina HiSeq 4000. 2.5 million reads were produced in total.

Alignment of the read count data was performed against the Genome Reference Consortium's GRCh38.p5 assembly with additional ERCC92 control sequences added in from Thermofisher. Trimmomatic (Bolger et al., 2014) was used to trim poor-quality sequences using a threshold quality trailing of 20 or if the read was below 60 bases.

After trimming low-quality bases from the reads the remaining sequences were mapped to the GRCh38.p5 genome build using the STAR tool version 2.4.0j (Dobin et al., 2013). Samtools version 1.3 (Li et al., 2009) was implemented for conversion of the aligned reads from SAM to BAM format before Python-based 'HTSEQ' version 0.6.1 (Anders et al., 2015) was used for generating output reports and feature count tables based on unambiguous read alignment to a single gene's exons. Once count tables were produced for each sample, they were collated into a single matrix and imported into 'R' for further processing.

‘SCATER’ (McCarthy et al., 2017) was implemented for gene and cell filtering by total reads, gene expression and percentage of control reads as well as the visualisations of these filters. SCATER was further used for data normalisation and conversion to Counts per Millions as well as to investigate PCA and t-SNE grouping and sources of variance in the data. ‘RUVseq’ (Risso et al., 2014) was used for ERCC content normalisation, SC3 (Kiselev et al., 2016) for consensus clustering, M3Drop (Andrews and Hemberg, 2017) for identification of potential signature genes within pre-DC sub-populations and novel code to compare pre-DC sub-population expression patterns to mature peripheral blood dendritic cell and monocyte gene signatures, displayed using ‘ggplot2’ (Wickham, 2009a).

Chapter 2 Figures & Tables

Tissue	Subset	Gating Strategy/ Phenotype
Skin	CD14 DC	Live, CD45+, singlet, HLA DR+, AF-,CD14+
	CD1c DC	Live, CD45+, singlet, HLA DR+, AF-,CD14-, CD11c+, CD141-
	CD141 DC	Live, CD45+, singlet, HLA DR+, AF-,CD14-, CD11clow, CD141+
Blood	CD14+ mono	Live, CD45+, singlet, HLA DR+, Lin-, CD14+, CD16-
	CD16+ mono	Live, CD45+, singlet, HLA DR+, Lin-, CD14low, CD16+
	CD14+CD16+ mono	Live, CD45+, singlet, HLA DR+, Lin-, CD14+, CD16+
	pDC	Live, CD45+, singlet, HLA DR+, Lin-, CD14-, CD16-, CD123+
	CD1c DC	Live, CD45+, singlet, HLA DR+, Lin-, CD14-, CD16-, CD123-, CD11c+, CD1c+, CD141-
	CD141 DC	Live, CD45+, singlet, HLA DR+, Lin-, CD14-, CD16-, CD123-, CD11c+, CD1clow, CD141+

Table 2.1: Sorting Strategy for GSE35457 Illumina Expression Data

Human skin and blood samples were sorted according to these gating strategies and phenotypes for Illumina BeadArray analysis.

Chapter 3: CODESET DESIGN AND DIMENSIONALITY REDUCTION FOR DENDRITIC CELL SUBSET ANALYSIS

Primary research question:
Can RNA expression analysis be used to distinguish and identify human DC and monocyte subsets?

Sub-topic questions:
1. Can human blood dendritic cells and monocytes be classified by their RNA signatures?

2. Can a focused panel of immune genes be used to identify common blood dendritic cell and monocyte subsets?

3. How many genes are required to maintain dendritic cell and monocyte subset classification?

3.1 INTRODUCTION

3.1.1 Dendritic Cell and Monocyte Subsets

Dendritic cells and monocytes are heterogenous cell types with specialised subpopulations that play a significant role in phagocytosis, antigen presentation and pathogen response. The cell types of focus for this thesis are the three major DC populations and two main monocyte populations. The DCs are composed of two conventional DC subpopulations cDC1 and cDC2, as well as a specialised plasmacytoid DC subpopulation, pDCs.

Plasmacytoid DCs are an intriguing cell type lacking many conventional lineage markers by flow cytometry and a distinct cell gene expression profile. pDCs have been defined by CD123, CD303 and CD304 expression (Collin et al., 2013), although their unusual properties and phenotype suggests many more molecular identifiers of this population could distinguish it from other mononuclear cells. Historical functional studies have assigned T-cell priming, Th1 and Th2 responses and roles in tolerance to pDCs (Ito et al., 2004), although very recent papers and on-going single cell studies suggest that this population may be composed of its own unique subpopulations to which its observed diverse functionality may be attributable (Villani et al., 2017).

cDC2 DCs are broadly referred to as CD1c+ myeloid DCs. They are the most abundant myeloid DC in peripheral blood. Making up approximately 0.3-0.8% of all mononuclear cells, multiple studies on cDC2 cells over the past decade have highlighted the role of this subset as potent T-cell stimulators and chemokine producers (Collin et al., 2013; Dzionek et al., 2000; O'Keeffe et al., 2015). Surface expression markers of cDC2s include CD11c, CD1c and SIRPA. In the blood cDC2 cells are CD1a negative, but their tissue equivalents are not (Collin et al., 2011; Merad et al., 2013).

cDC1 DCs are a CD141+ myeloid cell found very infrequently in the peripheral blood, for this reason they are problematic to study or develop *in-vitro* cultures from. New techniques such as NanoString nCounter analysis, CyTOF and single cell sequencing are proving invaluable tools for the study of these rare cell types where FACS sorting provides only a few hundred or a few thousand cells per sample. They are known to be phenotypically similar to conventional CD1c+ myeloid DCs, but in the peripheral blood, do have some known surface marker differences including a lack of CD1c expression, lower CD11c expression and intermediate to high expression of CD141 by flow cytometry (Haniffa et al., 2012; MacDonald et al., 2002). Observations within the HuDC research group have also noted CD1c expression in cultured CD141+ cDC1s, mirroring that of their tissue-derived counterparts.

From the monocyte population, CD14+ classical monocytes and CD16+ non-classical monocytes are of interest in this thesis.

Classical CD14+ monocytes are by far the most abundant circulating monocyte in peripheral blood. This highly phagocytic, low cytokine producing cell type is characterised by surface markers including CD24, CD14, and CCR2. Non-classical CD16+ monocytes have some phenotypic differences to classical monocytes, being smaller in size with migratory capabilities, but are broadly similar by gene expression analysis (Bigley et al., 2011; Robbins et al., 2008). Converse to CD14+ monocytes, the CD16+ subtype expresses CX3CR1 to a high level and CCR2 lowly (Ziegler-Heitbrock, 2000).

3.1.2 Dimension Reduction

Dimensionality reduction is in essence any applicable method that can be implemented to reduce the number of variables in an experiment, whilst maintaining most of the data variability. The issue of data saturation and the need for dimensionality reduction techniques spans many disciplines and multiple methods have therefore been implemented to address these problems.

Principal component analysis (PCA) is one of the most common techniques that can be used to reduce the number of dimensions. PCA defines a linear subspace dimension that accounts for most of the variability in a dataset, using a linear combination of all of the variables. Each of these new dimensions are called ‘principal components’ (Hotelling, 1933). Components are ordered from the highest variability to the lowest with the first two components being displayed on a 2D plot. This methodology retains much of the high dimensional data structure, but loses the influence of the lower dimensions. The process itself does not reduce the number of variables, but the most variable resulting principal components can be interrogated with the exclusion of low variability or collinear variables to find genes that account for the most variance in the data. By extracting the genes weighted highly in the high variability principal components, PCA can be used as a feature selection technique. One drawback of this however, is that in some datasets, the subtle sources of data variance that are lost by this method may be biologically relevant.

t-Distributed Stochastic Neighbour Embedding (t-SNE) is a relatively new technique for dimension reduction developed by Laurens van der Maaten (Maaten and Hinton, 2008). This technique can be very well suited to large datasets such as gene expression and single cell RNAseq data as its non-linear dimensionality reduction method preserves the higher-order grouping of the dataset while reducing the dimensions through a probability distribution mapped to the high dimensional data and low dimensional data. The divergence between these maps is minimised so that the final 2D map reflects the arrangement and similarity between the samples in the higher dimensions (Maaten, 2014). Both PCA and t-SNE are primarily methods of visualisation, reliant upon unsupervised grouping of similar features to reveal patterns or variance within the data.

A novel method of dimension reduction was produced by the author with the aim of maintaining each individual genes expression data. This method incorporated unsupervised grouping methods and supervised classification. Based on repeated k-means testing and random gene removal, multiple loops of the code iteratively removed single genes from the dataset and re-analysed the remaining genes by k-means clustering. This procedure reduced the number of genes down to the minimum needed to provide adequate grouping of cell subsets.

3.2 MATERIALS AND METHODS

As this Chapter is concerned with the development and testing of new methodology, further details are provided in section 3.3 and 3.4 along with the general overview of common methods provided in Chapter 2.

3.2.1 Illumina Datasets

Following ethical approval by Newcastle and Singapore Ethics Committees, normal skin samples were taken from mammoplasty or reconstructive surgery patients and peripheral blood was obtained from healthy volunteers. Peripheral blood mononuclear cells were isolated by Ficoll density centrifugation and separated using FACS with a FACSAriaII [BD] to a purity of >91% according to the gating strategy displayed in *Haniffa et al*, 2012 and Figure 3.1. The number of replicates used for each subset is displayed in Table 3.1.

Qiagen RNeasy mini kits were used to isolate the sorted cell subsets and checked for RNA quality using a QIAxcel analyser. Three hundred nanograms of total RNA was amplified using Illumina TotalPrep RNA amplification kits and processed on an Illumina-HT12 V.4 [GEO platform: GPL10558] gene expression platform.

Output data was normalised using Loess method without background subtraction and uploaded to the NCBI data repository [GEO number: GSE35457].

3.2.2 GSE35457 Gating Strategy

Cells were gated by forward- and side-scatter area, with DAPI staining used to exclude dead cells. Forward scatter height and area were used to exclude doublets after which leukocytes were gated on CD45.

HLA-DR positivity identified antigen presenting cells within a lineage negative fraction (gated on CD3, CD19, CD20 and CD56 in FITC channel). CD14+, CD16- cells were labeled as CD14+ classical monocytes, with the corresponding CD14lo, CD16+ fraction was labeled as CD16+ monocytes.

From the CD14, CD16 double negative fraction, CD123+ pDCs were gated, with the CD123- fraction containing both cDC1s and cDC2s. CD11c+, CD1c+, CD141- cells were identified as cDC2s, while CD11clo, CD1clo/- CD141+ cells were classified as cDC1s. This strategy is displayed in figure 3.1 and table 3.2.

3.2.3 NanoString Analysis Protocol

All NanoString experiments used in this chapter followed the protocol outlined in Chapter 2, section 2.6. All samples were collected as lysates in RNA lysis buffer after FACS at a concentration of 2,000 cells/µl. 5µl of lysate was used in the hybridisation process, resulting in 10,000 cells of RNA present in each sample with the exception of specific correlation testing. The additional correlation testing presented in figure 3.9, sections of healthy donor skin were stored in formalin-fixed, paraffin-embedded sections or frozen directly after excision. RNA was extracted using RNeasy extraction kits (Qiagen) and normalised to 150ng of material based on Bioanalyzer 2100 (Agilent) before hybridisation to the NanoString codeset.

For the correlation testing of lysed and extracted material displayed in figure 3.10, pre-processing of the cell populations included using an RNeasy extraction kits (Qiagen) to extract RNA from 10,000 cells and direct lysis of 10,000 cells. The two samples were then processed on the NanoString nCounter Analysis System and tested for correlation.

3.3 RESULTS

For ease of reading and identification, the use of specific 'R'-based functions within this chapter have been denoted with '()' after the name of the function. Eg. lmFit() denotes the use of the 'lmFit' function from the 'Limma' package. In this case, lmFit() fits a linear model for each gene in a series of arrays. This mirrors the way in which these functions are applied in the 'R' environment.

Due to the largely technical nature of this chapter, details outlining the preparation, pre-processing, normalisation and analysis have been included in the results section below. The goal of these experiments were to test if common dendritic cell subsets could be analysed and distinguished from each other on a restricted gene panel of 730 genes. Once this was established, further analysis of the genes defining each subset was performed with the aim of reducing the number of genes required for DC identification to the minimum necessary, while also highlighting potential marker genes of each cell type that could be used in follow up experiments.

3.3.1 Illumina Expression Analysis for the Separation of Human Blood Dendritic Cells and Monocytes

The initial step in the analysis pipeline was to extract the human subsets from the GSE35457 dataset and process the data for subsequent analysis.

Using the 'Biobase' and 'GEOquery' (Davis and Meltzer, 2007) packages of 'R', the GPL10558 HT-12 Illumina dataset for GSE35457 was imported along with the associated metadata and subset names. Each sample was assigned to a group representing the subset from which the sample was derived. Initially this step was performed using which(). This resulted in 27 samples assigned to one of five different subset groups, numbered in a grouping vector, with the remaining samples excluded from analysis.

Once grouped, an expression was designed to log_2 transform the data and interrogate it for any anomalies that may have been caused by incorrect data extraction such as negative expression values or missing points. Log_2 transformed data is required for 'limma'-based analysis. Any false values were reassigned a value of 'NaN', which was then used as a filter for the data.

Using the previously described grouping vector as a factor, a contrast table was generated using lmFit(), and a contrast matrix using makeContrasts() from the 'stats' ("R Core Team," 2012) and 'limma' (Smyth, 2005) packages. This analysis involved sequential pair-wise comparisons between the subsets and made use of the eBayes() function of 'Limma' to apply an empirical Bayes statistical model to the contrast tables, from which a moderated t-statistic could be derived for each probe on the array [Figure 3.2].

'annotate' (Gentleman et al., 2004), annotation() and getGEO() functions were used to annotate the dataset using NCBI platform annotations. Merge() was employed to bind gene ontology (GO) functional data to each respective gene to allow for downstream gene set enrichment analysis (GSEA), pathway analysis and functional analysis of the dataset.

For p-value generation, the topTable() function of 'Limma' was implemented to extract the top-ranked genes from the eBayes linear model. This function incorporated a Benjamini-Hochberg method false-discovery-rate (FDR) adjustment for multiple comparisons using p.adjust().

Custom-designed volcano plots were generated to display the significantly differentially expressed genes between each subset [external file 2] and a table was produced for each [external file 1].

The number of differentially expressed genes (meeting the criteria of log_2 fold change of >1.5 and $p<0.05$ after FDR adjustment) between subsets ranged from 1,468 differentially expressed genes between CD1c+ cDC2 and CD141+ cDC1 and 4,313 differentially expressed genes between CD16+ Blood monocytes vs pDCs.

Legacy knowledge cell markers feature highly in these gene lists including CD14, C19orf59 and S100A9 in CD14+ classical monocytes, CX3CR1 in CD16+ non-classical monocytes, CLEC9A and BATF3 in cDC1s, CD1c, CD2 and FCER1A in cDC2s and PACSIN1 and ASIP in pDCs.

By hierarchical clustering using ward.D2 agglomeration and euclidean distance matrices [Figure 3.3], the full Illumina array was capable of grouping the human blood dendritic cells and monocyte populations by cell type exactly. Furthermore, upper branches of the clustering also follow expected trends with monocytes grouping separately to DCs and pDC cells branching from cDC2 and cDC1 cells.

As it was clear that the Illumina platform could separate and define common human mononuclear cells, revealing known cell surface markers as well as hundreds of other subtype-specific markers, NanoString Technologies nCounter platform was considered as an alternative multiplexing gene expression platform, providing faster throughput, cheaper material costs and an ability to use other material types such as FFPE or lysates, with enough probes to define human mononuclear blood cells.

3.3.2 In-Silico Testing of NanoString Panel Using Surrogate Illumina Expression Data

Once it was determined that differential expression could be detected between human dendritic cells and monocytes by Illumina beadarray technology, the dataset was interrogated further with an aim to transition the gene signatures over to a digital multiplexing system: NanoString Technologies nCounter Analysis Platform. This platform offered curated panels of 600+ directed gene targets, including an immune cell-focussed Immunology_V2 panel that appeared to have the greatest relevance to this work.

To estimate the capacity of the NanoString Immunology_V2 panel at separating human mature cell types before funds were committed to the project, human blood subsets from the GSE35457 dataset were imported into the 'R' environment using 'Biobase' (Gentleman et al., 2004) and 'GEOquery' (Davis and Meltzer, 2007) packages of 'R' along with associated metadata and subset data. Each sample was then assigned to a group representing the cell subset from which the sample was derived.

Once grouped, these data were $\log_2$ transformed in preparation for down-stream analysis and interrogated for anomalies such as negative expression values or missing data points. Any false values were reassigned a value of 'NA', and subsequently filtered from the dataset.

From this dataset, Illumina IDs were mapped to their respective Ensembl Gene IDs, whilst NanoString's Ref-Seq gene IDs were also converted to Ensembl gene IDs. At this point, merge() was implemented to combine the two platform libraries with shared common probes, resulting in a list of Illumina gene IDs that directly corresponded to the NanoString Immunology_V2 gene array panel. 543 of 579 Nanosting gene probes were correctly identified and annotated between the Illumina and NanoString platforms. The remaining probes did not have an equivalent corresponding Illumina gene probe and were omitted from the experiment ad a technical discrepancy between the two technologies.

3.3.2.1 Pilot testing of the NanoString array genes

From the 543-gene dataset, hierarchical clustering was performed from a Euclidean distance matrix calculation followed by 'ward' method agglomeration using 'gplots' ("R Core Team," 2012) [Figure 3.4] .The cluster diagram indicates that the 500+ immune-related genes may be capable of distinguishing blood DC and monocyte subsets, correctly grouping each sample to its respective cell type, however the higher-order clustering is somewhat unexpected with cDC2 cells grouping most closely with CD14+ classical monocytes rather than with pDCs or cDC1s.

The absence of CD1c amongst other major surface receptors on the Immunology panel may have exacerbated this problem. The conservation of gene expression DC and monocyte populations is relatively conserved, leaving only a handful of marker genes for reliable subset separation. Without markers such as CD1c (for the identification of cDC2 cells), gene-level subset separation may be difficult.

By principal component analysis (PCA) in Figure 3.5, the cell types appear well grouped. As PCA weights the genes with the highest variance above those that show lower variance, samples unlike each other are forced apart, while like-samples are drawn together. This suggests that pDCs are much more distinct in their expression than the other subtypes. Dendritic cells generally occupy the positive region of PC1, with monocytes located in the negative region, split by PC2. pDCs are also separated from the classical DCs by PC2.

From this clustering it was decided that the addition of a number of curated genes might result in greater separation of DC subsets from monocytes and provide additional expression information on genes of interest and more accurately represent the results found by NanoString. Table 3.3 displays 30 genes that were added to the Immunology_V2 gene list, taken from the initial Illimina expression data signatures, published papers and gene targets that would be useful for other experiments in the research group. These genes and their relationale for inclusion have been noted in the table legend.

The addition of these 30 genes had relatively little effect on the overall clustering patterns by hierarchical clustering and PCA. By hierarchical clustering displayed in Figure 3.6, cDC1 and pDC populations form one branch, with the cDC2 cluster falling closest to CD14+ monocytes. CD16+ monocytes form an off-shoot from this CD14+monocyte-cDC2 branch.

Although the layout is mirrored, by PCA [Figure 3.7] the addition of the 30 genes to the Immunology_V2 panel did not affect the arrangement of the cell type clusters. Variance explained by PC1 rose from 19.6% to 20.4% with the addition of the panel+ genes, while PC2 accounted for 11.5% from 11.2% of variability. The addition of these extra genes do therefore have a modest impact on amount of variability between blood subsets, but the overall structure remained the same with pDC occupying one region of the PCA plot alone, monocytes separated from dendritic cells by PC1 and classical and non-classical monocytes splitting by PC2.

t-SNE, t-distributed stochastic neighbour embedding was used as another form of data visualisation with the aim of better representing the multi-dimensional relationships of the samples in two dimensional space. Unlike PCA, which gives weighting to the genes based on their variance, this visualisation method is designed to maintain the higher-order variance of the data. The t-SNE plot [Figure 3.8] depicts each population as a relatively tight cluster, well distinguished from the other populations.

3.3.3 NanoString nCounter Analysis of Human Blood Mononuclear Subsets

Once it was determined that a restricted dataset could be used to define the human blood mononuclear subsets, the experiment was moved over to the NanoString platform.

Initial correlation tests were performed to determine which NanoString nCounter pre-processing protocols were to be used for the main experiment. In this initial pre-testing phase, FFPE and Fresh-Frozen (FF) material was taken from a single donor sample and analysed on the NanoString nCounter analysis platform using the Human PanCancer profiling panel, representing a pioneering use of this technology at Newcastle University for gene expression analysis. A correlation coefficient was produced across the data from the two samples and indicated an extremely high conservation of data despite the expected degradation of RNA under FFPE conditions with an R-value of 0.975 [Figure 3.9].

A second pre-test was performed on RNA extracted from 10,000 PMBCs and an equivalent number of cells placed directly into RNA lysis buffer at a concentration of 2,000 cells/µl displayed in figure 3.10. This provided the basis for the main experiments performed in chapter 3 and chapter 4 and indicated that lysed whole cells would be used in place of extracted RNA, with a correlation coefficient of 0.997.

For comparison of Illumina and NanoString platforms, five blood subsets relating to the five subsets used in the illumina expression study were sorted from the PMBCs of healthy donor individuals and ran on the NanoString nCounter Analysis platform.

3.3.3.1 NanoString nCounter analysis of human blood subsets

Figure 3.11 shows a hierarchical clustering diagram based on the NanoString dataset. Each subset has three replicates in this experiment, forming distinct clusters based on their subset. The initial branching on this diagram is between monocytes and dendritic cells, similar to the initial full Illumina data plot [Figure 3.3], rather than the minimised Illumina dataset based on the NanoString gene list [Figure 3.4]. Branching from the dendritic cell side of the plot is cDC2, suggesting a closer gene expression relationship to the monocyte subsets compared to the other dendritic cell subsets.

Although the *in silico* testing of the Nanostring gene list suggested that the cDC2 subset may cluster with the monocyte subsets, this was not the case when using the NanoString platform. As the two technologies operate using different methods and chemistry, with the output and dynamic range differing vastly, it was expected that some disparities such as this would be seen between the two experiments Additional panel+ genes aimed at distinguishing cDC2 cells may also have contributed to this positioning by hierarchical clustering.

By principal component analysis [Figure 3.12], similar grouping of the blood subsets was determined. Compared to the *in silico* experiment [Figure 3.7], PC1 accounted for 36.5% of total variance, while PC2 accounted for 19% explained variance. pDC and cDC1 subsets grouped closely and split from cDC2 and the monocyte populations by PC2. Along PC1, CD14+ monocytes appeared at the lower extreme of the plot, followed by CD16+ monocytes, with cDC2s showing high gene expression amongst PC1 genes.

By t-SNE [Figure 3.13], each subset is well defined and separated in t-SNE space. By the t-SNE1 variable, the three dendritic cell subsets were distant from the two monocyte subsets, while t-SNE2 separated the two monocyte subsets as well as separating pDCs from cDC1 and cDC2s. pDC samples occupied the lower quadrant alone with negative TSNE values. cDC1 and cDC2 groups were separated in t-SNE space with cDC2 samples displaying a greater positive value by t-SNE1 and a more negative value by t-SNE2 than cDC1 samples.

3.3.3.2 Correlation of gene signatures between Illumina and NanoString platforms

By producing curated lists of differentially expressed genes that are highly expressed on a single cell subset in comparison to the others, commonly described as ‘gene signatures’, for both the full Illumina microarray dataset and the NanoString nCounter dataset, correlation between the two platforms could be assessed.

For this, BubbleGUM and GeneSign were used to generate phenotype signatures for each dendritic cell and monocyte subset in a pairwise minimal mean ratio method. This robust, permutation based algorithm is able to calculate a p-value and FDR-adjusted p-value for each cell subset by comparing a reference sample type to each other sample type for every gene. This is computationally intensive, but provides the most robust and accepted differential expression testing method.

CD14+ classical monocytes on both NanoString and Illumina shared 13 gene expression signatures of the subset. CD14 is a well-cited typical surface marker protein used to classify CD14+ monocytes and S100A8 and S100A9 have been defined previously as highly-expressed on this subset (He et al., 2016).

For CD16+ non-classical monocytes, 6/15 of the NanoString signatures were also seen on the Illumina dataset, including CX3CR1, a marker frequently used to distinguish the subset by flow cytometry (Ancuta et al., 2009).

cDC1 subsets shared 12 signatures across both platforms. Of note, BTLA, CLEC9A, IDO1 and BATF3 were all signature genes for this subset that have been published on before (Breton et al., 2016).

The cDC2 subset only displayed three signature genes on the NanoString platform compared to the 20 genes by Illumina, however, given the restricted geneset of the NanoString platform and the presence of both CD1C and FCER1A as signature genes on both platforms for cDC2s, it appears to be a correctly assigned gene set (Minoda et al., 2017).

Finally, pDC subsets on NanoString shared the greatest number of signature genes with their Illumina equivalents at 20. These included PACSIN1, GZMB, and ASIP, which are all strong, unique markers for bulk pDC populations.

As there is consensus between both technologies for every subset, with many of the major published cell-defining gene expression signatures present on both arrays and assigned to the expected subset, NanoString Technology's multiplexed qPCR-hybridisation based platform provided a robust, yet cheaper and faster alternative to array-based gene expression experiments for this project.

3.3.4 Gene Reduction and Feature Extraction

In this thesis, gene reduction and feature extraction were investigated in various ways to achieve different goals. The first of these was the removal of genes indicated as tissue-specific in order to group cell subtypes together from both the skin and blood of healthy individuals. This tissue-specific gene reduction was performed in order to correlate skin and blood mononuclear cells that are believed to be related to each other. Without removal of tissue-specific genes, it would be expected that the major source of variation between samples would be related to their location in the body.

Feature extraction (the identification of genes enriched at p=<0.05 and 1.5 fold increased in a single subset after pair-wise comparison to each other subset) using BubbleGUM was implemented alongside a machine-learning algorithm 'Linear Discriminant Analysis' (LDA) to provide a robust gene signature for each human blood mononuclear cell investigated. An initial signature was designed through the integration of multiple Illumina expression array datasets and validated on an external dataset that made use of the same or similar gene subsets.

Gene reduction was also used to generate a minimised gene signature capable of defining each human blood subset with the fewest number of marker genes. The purpose of this was to allow for the study of other cell surface markers by flow cytometry, whilst maintaining information on the major blood mononuclear cells. As flow cytometry has an upper variable limit of approximately 20 targets, using as few of these as possible to initially identify dendritic cell and monocyte populations would provide more space on a flow cytometry panel for alternative markers such as cell cycle, markers, inflammatory markers or other surface proteins.

3.3.4.1 Tissue removal effect on gene expression data

The Illumina dataset GSE35457 contained a number of sorted skin subsets, which were anticipated to be skin equivalents of the blood mononuclear subsets as noted by shared major subset markers.

The addition of the skin samples to the analysis pipeline resulted in an initial split by hierarchical clustering of skin-derived samples and blood-derived components.

Figure 3.14 shows this split. Although each sample grouped depending on its subset, blood and skin equivalent cells do not group together. This observation was suspected to be the result of a tissue-specific signature so a t-test based method of feature reduction was used to remove genes associated with a combined, general 'tissue-specific signature'.

Figure 3.15 is the result of hierarchical clustering after the removal of genes deemed significantly differentially expressed ($p=<0.05$) between the grouped blood and culture samples. Specifically, blood-derived subsets with skin-derived equivalents were compared by a two-tailed t-test, incorporating CD14+ cells along with cDC1 and cDC2 cells. pDCs and CD16+monocytes were not included in the t-test, but were used in the hierarchical clustering to ensure that there were enough probes remaining in the dataset to correctly distinguish all mononuclear cell types.

From Figure 3.15 the first branch of the dendrogram separates the classical blood and skin-derived dendritic cell and pDCs from the CD14+ cells and CD16+ monocytes.

Further along the hierarchy, pDCs are separated from the classical dendritic cells, which are themselves split into cDC1 cells and cDC2 cells. Along the monocyte and CD14+ cell arm of the dendrogram, CD16+ cells branch off from the CD14+ blood monocytes and CD14+ skin cells. This pattern of similarity is conserved in Figure 3.16 through the visualisation of the data by t-SNE. By t-SNE plotting, each subset can be defined, with a cDC1 group, cDC2 group and CD14+ group consisting of both skin and blood equivalent cells. CD16+ monocytes and pDCs occupy distinct regions of the t-SNE space and do not share as much similarity to any of the other subsets.

From these results, it appeared that the 'tissue-effect' could be overcome by a general two-tailed t-test. This tissue effect may be generalised as removing a single signature differentially expressed between the combined skin subsets against their blood equivalents was enough to group all of the mononuclear cells by cell type, regardless of tissue type. Despite phenotypic differences related to the microenvironmental tissue niche, the cell type backbone remained across the tissues.

3.3.4.2 Collation multiple datasets and validation of normalisation

Two datasets from separate published illumina expression papers were combined using 'ComBat' from the 'SVA' package of R (Leek et al., 2012), with pDCs selected for normalization and validation of the batch effect adjustment due to their presence on both datasets and their distinctive cell signature [Table 3.5]. The observation of pDC distinction was derived from the initial data generated for this thesis in Figures 3.3 – 3.13.

In total, 41 samples (from 14 healthy individuals) were taken from the two datasets providing between 8 and 14 replicates for each mononuclear cell type to be studied. This larger dataset would provide greater statistical power for a cell type signature and encompass a broader variation of expression within each cell type.

After normalising and collating the two datasets into a single dataframe, initial hierarchical clustering of the entire data was performed to determine the accuracy of the ComBat function in combining the data. Figure 3.17 shows four distinct groups relating to the four DC and monocyte subsets. Initially all of the pDCs from both datasets branch off, followed by the CD14+ monocytes and then a split of the two cDC populations. There was no evidence by hierarchical clustering of any batch effects related to the use of two different datasets after ComBat correction.

By t-SNE, strong clustering of the subsets is observed. t-SNE1 and t-SNE2 variables separate each subset into a single quadrant and each group is very compact suggesting high correlation of gene expression within each subset [Figure 3.18].

3.3.4.3 Robust signature generation for cell type signatures

For the generation of a robust cell type signature with a good overall power and significance, the combined illumina expression dataset was interrogated using GeneSign (Spinelli et al., 2015) gene signature generation software. The mechanism for this signature generation involves extensive pairwise testing of every gene for each subset against each other subset as demonstrated in Figure 3.19. For each gene, minimal pairwise testing was performed with Benjamini-Hochberg false discovery adjustment (Benjamini and Hochberg, 1995) and a threshold cut-off of P-value <0.05 and $\log_2$ fold change of 1.5 (equating to a linear fold change of 3). Figure 3.19 highlights an excerpt from the GeneSign output for pDC signature genes, displaying PACSIN1, PTGDS and ASIP as some of the top differentially expressed genes, in-line with the initial illumina expression analysis and confirming their usefulness as marker genes. Table 3.6 further details the top 12 differentially expressed genes for each cell subset as well as the total number of signatures that GeneSign produced. Overall, 3,439 genes were identified as subset markers across the four cell subsets, meaning the expression of each gene was at least 1.5 log fold higher in one subset compared to any other subset after pair-wise comparison.

Within the pDC populations, GZMB, PACSIN1 and ASIP were amongst the most differentially expressed of the 1,425 genes identified by GeneSIgn as subset markers of pDCs. This was the largest of the cell signatures and reflects the distinctive features of pDC gene expression compared to the other mononuclear cells. These gene differences of pDCs compared to cDCs might be related to the lymphoid nature of pDC development compared to the myeloid-derived cDCs.

The CD14+ monocyte signature was 964 genes in length and included a number of common monocyte identifiers. S100A8, CD14 and C19orf59 were all in the top 12 monocyte markers, supporting their choice in the custom panel+ NanoString panel produced for section 3.3.2. As with the pDCs, the large number of signature genes for monocytes might be a reflection of their phenotypic differences to the cDCs and pDCs and earlier divergence from DC-like cells at the MDP stage of haematopoietic development.

The cDC subsets are more closely related to each other than the other subsets, producing smaller, but still distinct, gene signatures. cDC1 signatures included IRAK2 and MAPK13 amongst 481 other signatures, cDC2 signatures included a number of CD genes such as CD1c, CD1e and CD2 as well as previously identified cDC2 cell markers, FCER1A and CLEC10A. Some of the markers featured in the Panel+ custom codeset genelist for the purpose of distinguishing cDC2 cells from other cell types were identified by GeneSign, confirming their exclusive high expression pattern in cDC2 cells. In total, 567 cDC2 marker genes were identified using the GeneSign process.

3.3.4.4 Visualisation and machine-learning validation of cell type signatures

To visualise the capability of the 3,439 gene signature to group and identify each of the four mononuclear cell subsets, an integrated heatmap and hierarchical clustering diagram was developed using Euclidean distance and Ward method agglomeration and the Heatmap.2 function of 'gplots'. Figure 3.20 displays the gene signature as applied to the combined mononuclear cell subset data. Each of the cell types were well grouped and the robust GeneSign method of signature generation was observed as blocks of high expression for each of the subsets. From left to right on the heatmap there is a distinct block of genes enriched in the monocyte cluster, followed by a large block of genes expressed highly on pDCs. A small block of cDC1 genes was next, ending with the cDC2 signature genes on the right side of the heatmap. Reduction of the 47,000 illumina probes to the 3,439 gene signature still provided a strong distinction between the cell types. While visually, the gene list appeared robust enough for cell type clustering, machine learning was implemented as a non-biased validation method for the gene list.

A form of machine learning called 'linear discriminant analysis' (LDA) was used from the 'MASS' package on R (Venables, 2002). This form of pattern recognition was designed to find a combination of features within the dataset that maximise the separability of the sample classes. For this thesis, the technique was able to create new feature space, which then be used to infer or classify 'unknown' samples to a subset group based on similarity of gene expression.

The 41-sample dataset was split into randomly assigned groups of almost equal size (10-11 samples each). Linear discriminant analysis was used on all but one of these groups to 'train' the machine-learning algorithm. Once this was complete, the sample identifiers were removed from the final 'test' group, which had remained unseen by the machine. Based only on the expression profile, the machine would assign each test sample to a subset group, providing a confidence score for the assignment. The classification was then checked by comparing the assigned group to the sample identifier. The process was repeated so that each 10-11 sample group became the 'test' group and the other groups were used as a training set.

Figure 3.21 highlights the output from the LDA analysis. The machine learning algorithm assigned each cell subset sample with a cell type based on the training data provided. For each test group, the number of correctly assigned and incorrectly assigned samples were displayed, where 'correctly assigned' refers to instances where the LDA assigned the same cell type to the sample as expected and any 'incorrectly assigned' samples would be counted, if the LDA assignment did not match the cell type identifier. From the run shown in Figure 3.21, 10/10 samples were correctly predicted and none of the samples were incorrectly defined by the algorithm. After four iterations, all samples had been subject to testing, and the algorithm was found to be 100% accurate with its assignment. The table included in Figure 3.21 provides an accuracy score for each of the test samples in the first iteration of the experiment. Three CD14 monocyte samples (red) were correctly assigned as having a CD14+ monocyte signature with a confidence score ranging from 85.6% to 98.4%. The single cDC1 sample (blue) was assigned to the cDC1 group with 99.7% confidence. There were four cDC2 samples (light blue) in the first iteration test group, which were all correctly defined as cDC2s with three samples having a confidence score of over 99.3% and one with a lower score of 74.0%. This lower confidence may be a result of different gating strategies between the two illumina expression sets used in this analysis providing a slightly different cDC2 bulk signature in these samples. Finally, the two test pDC samples (black) were very confidently assigned as having a pDC signature by LDA with confidence scores of 99.5% and 100%.

The high prediction scores and 100% correct assignment of the full dataset was a strong indicator that the gene signature created by the author was strong enough to define the four mononuclear cell subsets used in this analysis in an un-biased manner. To further test the signature, a separate publicly available mononuclear cell Illumina dataset was downloaded from *Lee et al*, 2015 published paper 'Restricted dendritic cell and monocyte progenitors in human cord blood and bone marrow' under the repository code GSE65128. This dataset was restricted to subsets used in this section of the thesis and totaled 36 samples split into 9x cDC2, 7x CD14+monocytes, 10x pDCs and 10x cDC1 samples. The dataset and publication were both created without any involvement or collaboration from the HuDC group or Newcastle University, providing a distinct, independent dataset for cell signature testing. Once the data was downloaded, the 3,439 gene signature was applied to the dataset and the results of clustering and heatmap generation were displayed in Figure 3.22. In clear correlation with Figure 3.20, each subset was well defined and broad blocks of high gene expression could be observed mirroring the training dataset. The ability of this signature to function on previously unseen and unanalysed data was an indication of the robustness of the signature itself. A further display of the clustering of the GSE65128 dataset was included in Figure 3.23 as a t-SNE plot. Each cluster was relatively well grouped and defined, positioned in a separate region of t-SNE space.

3.3.4.5 Creation and testing of a novel gene reduction technique for subset classification

The success of the GeneSign gene signature experiment and subsequent validation highlighted the potential for an extremely minimal set of potent signature genes that could separate the mononuclear cell subsets effectively.

Starting from the GeneSign signature set, as each signature gene was highly expressed in only one cell type, it was concluded that a smaller geneset could probably be created whilst still maintaining the distinct gene clustering, this led to the development of a novel gene reduction method created by the author to sequentially narrow a geneset down to the minimal number of individual genes required for successful clustering. The major aim of this experiment was to retain sample information and expression values of each individual gene after gene reduction, unlike feature reduction by other reduction methods such as PCA or t-SNE that produce artificial new features to display a larger dimension dataset in smaller dimensions.

Figure 3.24 shows a basic schematic of the novel gene reduction function generated by the author for this thesis on the R programming platform. The basis of the technique was iterative pruning of a gene set, re-tested after every reduction to maintain the clustering of the original dataset but with fewer genes involved. The randomised gene reduction method was based on k-means testing. From the 3,439 gene signature displayed in Figure 3.20, the samples were clustered based on the between-group sum-of-squares variable of k-means clustering. This value is a measure of the difference between each group. A greater value equated to a greater separation of the subtype groups.

A random number generator was implemented to select and remove a single gene from the geneset and a new k-means analysis was performed on the remaining genes. A string of nested loops and functions tested the between-group sum-of-squares and if the new geneset maintained the correct grouping of the samples and provided a greater separation of the cell subset groups, the new genelist would replace the original genelist and the process would be repeated with another random gene removed. If at any point in the process a new genelist was not capable of maintaining cell type clustering or provided less separation of the clusters than its parent geneset, the randomly removed gene would be replaced and another random gene selected instead.

After fine-tuning, 20,000 attempts of gene removal were settled on as a limiter and 2,000 iterations of the entire process were implemented from 3,439 genes to the minimal number found to correctly assign cell subsets to individual clusters. In many instances the results of this minimisation produced minimised gene lists of two or three genes that could be used to accurately assign each mononuclear cell subset to its correct cell type cluster. Frequently the same genes were identified in these final gene lists despite the random nature of the minimisation method. The output of this analysis is displayed in PDF format [external file 3].

As a final refinement, this finding paved the way for an implementation of the technique for the purpose of identifying potential cell surface proteins that could be exploited as flow cytometry or FACS based antibody bound florophore marker targets. The GeneSign dataset was screened for genes with high confidence of cell-surface expressed proteins using data stripped and curated from the 'COMPARTMENTS' subcellular localisation index database. 612 genes remained and these were used as the parent dataset for a further round of gene minimisation.

From this analysis, a number of iterations produced two-gene signatures capable of correctly grouping the subset sample data by mononuclear cell type. One example of this is shown in Figure 3.25, the combination of FCER1A and SLC2A3. FCER1A an IgE receptor is associated with initiation of the allergic response. FCER1A was highly expressed on pDCs and cDC2 cells, but lowly expressed on CD14+ monocytes and cDC1 cells. SLC2A3 was expressed highly on cDC2 cells and CD14+ monocytes, but lowly expressed on both pDCs and cDC1 cells. This gene encodes a glucose transporter that can also mediate the uptake of various other monosaccharides across the cell membrane and may be expressed highly on CD14+ monocytes due to the high metabolism of these cells. Such minimized signature combinations could be a basis for FACS sorting of DC and monocyte populations, potentially reducing the need for multiple cell type marker genes in flow cytometry experiments and freeing up channels for new flow antibodies to be used to assess other cell qualities. Both FCER1A and SLC2A3 may warrant further testing as potential novel FACS DC and monocyte subset markers.

3.4 DISCUSSION

Gene expression profiling of common DC and monocyte cell types is integral to immunology research. With the development of microarray and multiplexed gene expression technologies offering full coverage or analysis of hundreds of gene targets at once, new avenues for DC and monocyte biology have been uncovered. This chapter focused on comparing and defining pDC cells, cDC1, cDC2 and CD14+ classical monocytes using gene expression and microarray technologies combined with a multitude of analysis software and programmes. Other subsets and tissue types were investigated for comparative purposes and gene signatures were developed to accurately and consistently group these common DC and monocyte cell types using a combination of well-established clustering and visualisation techniques with custom-designed novel methods of gene set reduction analysis.

3.4.1 Initial Gene Expression Profiling of Dendritic Cells and Monocytes

The GSE35437 Haniffa *et al* dataset (Haniffa et al., 2012) proved a strong basis for dendritic cell and monocyte gene expression analysis, composed of five cell types split into three main populations; monocytes, cDCs and pDCs. The samples and sorting were performed by the HuDC group in collaboration with SIgN, the intention being to define and contrast human and mouse skin and peripheral blood cells.

The initial differential expression analysis for GSE35457 samples provided an overview for potential gene signatures, whilst highlighting the individual differentially expressed genes between each cell subset. This analysis provided insights into cell type diversity and closeness as cell types that are more closely related in haematopoietic development would be likely to have fewer distinguishing features between them than cells that diverge early in haematopoietic development. This aspect was particularly important for cDC1 vs cDC2 subsets and CD16+ vs CD14+ monocytes, which were expected to be the mostly closely related subsets to each other.

Between cDC1 and cDC2 subsets, 1,468 Illumina probes were identified as differentially expressed in external file 1. A few of the most significant of these included cell surface markers CD1c, CLEC10A and CLEC9A which are used as flow cytometry markers for cell isolation. Identifying such markers is a validation of the Illumina gene expression analysis as these genes and their surface proteins have been identified as cell type markers in multiple publications (Breton et al., 2016; Jongbloed et al., 2010).

CD16+ monocytes and CD14+ monocytes had a total of 1,635 differentially expressed Illumina probes between them through the 'Limma'/'Lumi' analysis pipeline. The most differentially expressed of these targets included CCR2 (highly expressed on CD14+ monocytes), CX3CR1 (highly expressed on CD16+ monocytes), and CKB (also highly expressed on CD16+ monocytes). These targets were also supported in the literature as unique defining genes for the identification of CD14+ or CD16+ monocytes (Williams et al., 2014; Wong et al., 2011).

pDCs were expected to be the most distinct of the mononuclear cell subsets investigated and this conclusion appeared to be reflected in the number of differentially expressed genes between this subset and each of the other subsets. About 3,000 probes were indicated as differentially expressed between the pDC subset and each of the cDC subsets, with around 4,000 probes implicated between pDCs and the monocyte subsets. Common to each of these comparisons was PACSIN1. PACSIN1 is a potent regulator of TLR7 and TLR9, which in the case of pDC cells acts as an adaptor molecule enabling type I interferon production through TLR7 and TLR9 activation after microbial stimulation (Esashi et al., 2012, p. 1). Within the DC lineage, TLR7 and TLR9 are specific to pDCs, although other TLRs are expressed by the cDC subsets or displayed across the DC lineage cells (Jin et al., 2014; Jongbloed et al., 2010). TLR expression in DCs is investigated further in chapter 4 and figure 4.11. KCNA5 and MZB1 also featured highly as upregulated pDC genes in this analysis, mirroring observations in very recent publications (Loughland et al., 2017; Villani et al., 2017).

As well as individual comparisons between subsets, hierarchical clustering techniques were employed to visualise the entire dataset. Figure 3.3, a hierarchical clustering diagram based on Euclidean distance and Ward method agglomeration, visualises the relationships between the different cell subsets. This figure supported the conclusion that the two cDC subsets were developmentally closest as noted by the distance measure at the side of the diagram. The CD16+ and CD14+ monocytes are also closely related as reflected by the fewer differentially expressed genes between these subsets in the '1 vs 1' comparisons. The position and grouping of cell types in Figure 3.3 recapitulates traditional monocyte and DC development pathways in humans (Geissmann et al., 2010; Ginhoux and Jung, 2014). From the macrophage and dendritic cell precursor stage, cells develop through the common dendritic cell precursor, or common monocyte precursor. At this point, pre-DCs split from pre-monocyte cells, as in the first split in the dendrogram. Monocytes split from here into CD14+ populations or CD16+ populations, while along the DC lineage, pDCs are distinguishable from the cDC populations after the CDP stage. Later on in development, clear cDC1 and cDC2-like cells form (Schlitzer et al., 2015). In Figure 3.3, the DC populations and monocyte populations are the furthest distance apart, indicating least similarity. pDC cells next split from the population of cDCs, followed by a split of the remaining subsets at almost equal distances; the monocytes into CD14+ and CD16+ subsets and the DCs into cDC1 and cDC2 subsets.

3.4.2 The Basis for Changing Technology Platforms

At the time of writing, single cell RNAseq experiments were not yet available and technologies including Illumina BeadArray and Affymetrix GeneChip were particularly expensive for repeated experiments and testing over a longer period of time. Data analysis of such techniques was not as widespread and few analysis packages were in use. NanoString nCounter was a new alternative to microarrays providing multiplexed RNA expression assays incorporating focused panels of curated gene targets in the region of 200-800 probes. While cheaper pricing was of some benefit, the major draw of NanoString technology over microarrays was a minimal requirement of around 100ng of material, or 8,000-12,000 sorted, lysed cells. Given that DC populations represent less than 1% of total cells in the blood and cDC1 cells are <0.05% of the PBMC population, low cellular input was a major concern in this project. The NanoString nCounter could accommodate whole RNA directly from lysed sorted cells without need for amplification or library prep. In comparison, Illumina BeadArray required 300ng of material that would need to be amplified further, possibly introducing a degree of amplification bias and putting very rare cell types and subpopulations out of reach. Outside the scope of this thesis, but relevant to the change in technology, the nCounter platform was at the time of writing the only platform capable of reliable RNA analysis from FFPE material. The nCounter platform used 100bp regions for barcoding, meaning it would work effectively with extremely degraded RNA. RIN values of less than 2 were frequently found from FFPE extracted RNA, but the data output from the machine was relatively unaffected by this extent of degradation. This provided the research group and other internal users with a method of analysing historical clinical sections easily and simply.

With the nCounter platform, samples could be collected, isolated, sorted, counted and processed for NanoString analysis within a single day and could be done in-house by a single operator. The turn-around time from sample collection to data-collection was reduced from three weeks with Illumina GeneChip to three days with NanoString nCounter once the platform was set up in the HuDC laboratory.

Analysis and normalisation of NanoString data was expected to be simpler in comparison to Illumina, especially for a research group with no dedicated data analysis personnel, thanks to the nSolver analysis package. Ultimately this proved too basic for the applications in this thesis and so custom R code was developed for analysis purposes by the author. Over the last few years there has been huge investment into analysis techniques and more general use of Python and R has resulted in the development of multiple analysis and visualisation packages for microarrays and multiplexed digital RNA-PCR alternatives such as NanoString. Some such as Lumi and Limma have become common standards.

3.4.3 In-Silico Testing on Illumina and Comparison to NanoString Analysis

NanoString Technology utilised focused panels of probe targets specific to individual research interests. Unlike microarray platforms that typically cover the entire transcriptome and therefore every known human gene, NanoString panels used probes designed against only a portion of these. It was therefore imperative that initial i*n-silico* experiments were conducted to determine if panels of 700+ immune related genes were capable of separating out dendritic cell and monocyte populations of interest in a manner similar to the Illumina expression analysis data GSE35457.

The *in-silico* analysis of the NanoString Immunology_V2 panel using surrogate Illumina expression data provided swift insight into the capability of NanoString nCounter for this research.

By stripping out only the genes found on the NanoString panel from the Illumina expression dataset and displaying the data, the research group could get an estimated representation of how the NanoString machine would perform with common mononuclear cell subsets. Inherent differences in the two platforms were to be expected, as the technologies behind them are quite different, however the *in-silico* experiment indicated that the combination of gene probes in the smaller NanoString panel could define cell subsets. Figure 3.4 indicated that such a restricted genelist was capable of distinguishing each cell subset, although cell type relationships were not maintained as they were in Figure 3.3. As the focus of this thesis was in defining cell subsets and profiling in-vivo and in-vitro cells, maintaining this higher-order grouping was unnecessary.

Hierarchical clustering of the restricted NanoString geneset panel of 543 genes correctly assigned all samples into subset groups that were distinct enough from one another by Euclidean distance to be a convincing success. This was promising for future gene reduction methods as it appeared that from 47,000 Illumina probes and 23,000 gene targets, each cell subset could be well defined and replicated even when reduced down to just 543 gene on the NanoString platform. The positioning of the cDC2 subset amongst the monocytes was the major difference between the Illumina clustering and the NanoString gene *in-silico* Illumina data clustering, although given that the NanoString dataset is a comparatively small dataset curated around immune-related genes, changes such as this were anticipated. From an immune basis and functional standpoint cDC2 and cDC1 cells have their own individual roles to play in immunity so while they may be developmentally close overall by Illumina expression analysis, they are functionally different when observed using an immune-biased dataset. cDC1 cells exhibit potent Th1 responses and are known to be superior MHC class 1 cross-presenting cells for viral proteins including HIV-1 and Hepatitis-B as a result of their unique surface phenotype including particularly high expression of CLEC9A as noted in external file 1 (Castell-Rodríguez et al., 2017; Jongbloed et al., 2010; McGovern et al., 2015).

cDC2 cells express higher levels of CCR7 and a range of activation markers. They are more plastic than cDC1s in this respect and in tissues, respond to a range of TLR agonists with a variety of cytokines. Initiation of T-cell response and antigen detection have been noted in cDC2 publications (McGovern et al., 2015; McLellan et al., 1998; Yu et al., 2013).

By principal component analysis in Figure 3.5, the five cell subsets have relatively low internal variation despite being derived from different healthy donors. The groups are well positioned away from each other with the monocytes in the negative space of PC1 and the cDC1 and pDC subsets in the positive region of PC1. Interestingly, cDC2 is around the mid-point of PC1, located between the cDC1 cluster and CD14+ monocyte cluster. From Figure 3.3, cDC2 cells branched from cDC1 cells, while in Figure 3.4 displayed cDC2 cells branching from CD14+ monocytes. The PCA in Figure 3.4 partially recapitulates both aspects, pDC cells in all diagrams for this section were positioned far from other clusters, echoing their unique physiology and developmental pathways (Collin et al., 2013; Ito et al., 2004).

The addition of 30 genes to the NanoString Immunology_V2 panel was deemed necessary to provide additional gene expression information not covered by the original panel genelist. Many of the gene signatures identified in this chapter and discussed already in section 3.4 were discussed in publications arising around the time of this study. The full rationale for the 30 gene Panel+ selection is included in appendix B.

Future work was expected to include further analysis of cDC1 and cDC2 distinctions and thus it was felt that the addition of cDC distinguishing genes was necessary. CLEC10A and CLEC9A and in particular CD1c were important for this. Observing a relation between flow cytometry output and gene expression for such cell sorting target genes may have been investigated. It was thought that these additional genes might have also provided a stronger separation of cDC2s from the monocyte populations, thereby reproducing the full illumina expression analysis clustering pattern seen in Figure 3.6, this was not the case with the in-silico analysis by hierarchical clustering in Figure 3.7 and by PCA in Figure 3.7, this extended dataset still mirrored the original PCA in Figure 3.5, with only marginal increases in PC1 and PC2 explained variance, suggesting that the additional genes were not adding a significant push to the clustering, although their addition was still beneficial at the individual gene expression level for providing gene counts for CD1c amongst other research focus targets.

The first of the analysis figures presenting NanoString data on the NanoString platform began from Figure 3.9. Correlation testing of the NanoString platform proved highly successful with R-values of over 0.975 for both FFPE vs fresh frozen samples and extracted RNA vs cell lysates. The reason for such high correlation is likely due to the chemistry of the NanoString assay. As each capture and reporter probe is 50bp in length, the RNA integrity can be extremely low (in the region of 2.0 RIN as recorded by the Agilent Bioanalyzer 2100) and still be long enough to provide adequate binding area for the capture and reporter probe complexes to bind. The comparison of FFPE and fresh frozen material did exhibit less correlation for lowly expressed genes, but correlation was improved above a $\log_2$ expression value of 5 with an extremely correlative R-value of 0.975 overall. Correlation between lysed cells and extracted RNA was also extremely high, although it was measured on a smaller NanoString panel composed of ‘housekeeping genes’. This may have mitigated any correlation differences attributed to lowly expressed genes as the lowest $\log_2$ expression value was noted at 5, but displayed strong associations between the samples despite the different storage and pre-processing methods applied. The ease of sorting cells via FACS directly into lysates for immediate use on the NanoString platform saved a significant amount of processing time and consumables and was selected as the method of choice for the experiments detailed in chapter 3 and 4.

The first experimental data for the comparison of human blood mononuclear cells is shown in figure 3.11, with hierarchical clustering of the resulting dataset. Although *in-silico* testing of the geneset suggested cDC2 cells might have grouped with the monocyte subsets they did not. The basis of the two technologies is the most likely explanation for this observation. Illumina expression arrays use coated beads containing hundreds of thousands of labelled 79 nucleotide-long oligo-sequences for binding RNA targets. Fluorescence intensity is measured from these beads to determine the relatively level of expression by intensity. In contrast, NanoString uses individual molecular barcodes of 100 nucleotides in length for each RNA target that are subsequently read and recorded for a direct digital count of expression. Correlation between the two technologies is limited due to scaling, dynamic ranges and underlying differences in data distribution.

cDC2 samples in Figure 3.11 cluster with the remaining DC subsets, although they appear to branch off first, unlike in Figure 3.3, where pDC cells were first to branch off. The reasoning for this might be related to the Panel+ enrichment of cDC2 positive marker genes. These additional genes did not appear to provide much difference during the illumina *in-silico* testing phase, but may have had a greater influence upon transference to the NanoString system.

By PCA, the position of the cDC2 cells was again altered. The pDC and cDC1 subsets were grouped relatively close on the negative region of PC2, with the monocyte subsets on the positive region of PC2, however although distinguishable from the other subsets by PC1, the cDC2 cluster was also found on the monocyte side of PC2 suggesting that PC2 weighted genes were strongly separating cDC2 cells from the other DC subsets. Another feature of note in Figure 3.12 was the stretched, linear grouping of the cell type clusters, most notably observed in the CD16+ monocyte and CD14+ monocyte clusters. This was likely a technical issue relating to the normalisation process for NanoString data. Unlike normalised Illumina expression data that typically ranges from an intensity index of six to fifteen, NanoString data is based on counts and thus ranges from one to approximately 400,000 counts, with the upper limit determined by probe density on the streptavidin reading surface of the NanoString cartridge and depending on the amount of input RNA. As a result of a normalisation factor derived from the geometric mean of 20 housekeeping gene targets applied to each sample during the QC stages of analysis, a baseline count value is altered so that a sample with a normalisation factor of 2 will have all counts multiplied by that value. The linear orientation of the samples was likely a visual manifestation of this normalisation factor applied to the lower expressed genes in these samples and thereby altering the data in a constant manner. This is not a major issue for differential expression analysis, but as the genes driving PC1 appeared to separate out DC subsets from one another, these genes may have had relatively low counts in the monocyte subsets, making this baseline alteration apparent along PC1. Figure 3.13 reinforced this conclusion where the t-SNE algorithm did not weigh the genes as in a PCA analysis resulting in less of this bias. The t-SNE plot, appears to show strong separation of the cell subsets, most closely resembling the Illumina expression analysis visualisations. As a visual representation of the high dimensional expression patterns of the data, the t-SNE plot appears to recapitulate traditional expectations of gene expression. pDC samples occupy a very distinct region of the t-SNE plot, the monocytes occupy the positive region of t-SNE2 and the cDC subsets occupy the lower, positive region of t-SNE1, thereby splitting the data by major cell type; monocytes or DCs, then splitting the pDCs from the cDC subsets and finally separating the cDC subsets into cDC1 and cDC2 cells in a similar pattern to Figure 3.3 and Figure 3.5 from this thesis and approximated in

the Illumina expression analysis of Figure 3 in McGovern *et al* (McGovern et al., 2014).

3.4.4 Effect of Additional DC Skin Samples and Tissue Signature Removal

The GSE35457 dataset contained a number of skin-derived dendritic cell subsets believed to be equivalent to the blood DC subsets. Table 3.4 highlighted the similarity of surface markers between these cell types, which were subsequently used for flow cytometry and cell sorting of the populations for later Illumina GeneArray analysis. For the skin samples, autofluorescence was used as a measure to ensure any macrophages were removed from the end sort gates. Very granular cells such as macrophages typically display a degree of autofluorescence, which is exacerbated under macrophage activation. Apart from this, the majority of the gating strategy was maintained between the skin and blood equivalent cell types. CD11c and CD141 were considered fundamental to cDC identity with CD11c-/lo, CD141+ cells considered CD141+ cDC1 cells and CD11c+, CD141- cells qualifying as CD1c+ cDC2 cells in both skin and blood subsets.

Addition of the skin equivalent subsets was a simple process, but resulted in a strong defining split in the dataset by hierarchical clustering, as displayed in Figure 3.14. Blood subset clustering was maintained, identical to the earlier iteration in Figure 3.3 as expected, however all of the skin-derived subsets clustered along their own branch of the dendrogram, which was interpreted as a tissue-specific signature present in all cell types regardless of their DC or monocyte subtype. It was concluded that cells entering the tissue or peripheral blood must have conserved or partially conserved transcriptional changes related to their entry and exit across tissues. Within the skin branch of the dendrogram cell type relationships were maintained similar to the blood derived subsets with the CD14+ skin cells branching off first from the skin-derived cDC subset cluster. The cDC cluster subsequently grouping into cDC1 and cDC2 cells.

In an effort to address and overcome the issue of tissue type signatures overwhelming the subset clustering and obscuring any conclusions regarding the closeness of tissue derived equivalent cells, a simple yet effective method of gene reduction was devised. A twist on typical differential expression analysis, a two-tailed t-test was performed on grouped CD14+ cells, cDC1 and cDC2 subsets based on their tissue type producing a list of genes that were significantly differentially expressed between the blood samples and skin samples. These were expected to consist primarily of tissue-specific genes, and indeed, upon removal of these genes from the dataset, a second round of hierarchical clustering produced a dendrogram separating samples by their cell types rather than tissue type. This effect was confirmed by both hierarchical clustering using Ward agglomeration as well as t-SNE in Figures 3.15 and 3.16. Both plots resulted in a clear separation of monocytes, cDCs and pDCs. By removing the same signature from every sample irrespective of their cell type, both blood and skin equivalent cells were grouped together in a format reflecting the initial blood-only dendrogram from Figure 3.3. Importantly, blood cell types that did not have a skin equivalent and took no part in the differential expression testing for tissue signatures were still clustered an a biologically correct manner. CD16+ monocytes were branched off of the monocyte portion of the dendrogram above the combined blood and skin CD14+ cell cluster. Likewise, pDC samples branched from the DC clusters before the combined blood and skin cDC samples branched into cDC1 and cDC2 clusters appropriately.

Through this novel process of reversed-differential expression analysis, a constant tissue specific gene signature was identified as shared by all cell subsets and was subsequently isolated and removed from a large multi-dimensional dataset resulting in a recreation of the clustering and visualisations resulting from a single tissue type dataset.

The idea of a single ‘tissue’ signature is extremely valuable as it may infer that cells are transcriptionally and epigenomically altered in their gene expression based on their location. Outside of the scope of this thesis, but of major interest to DC and monocyte biology and development, such tissue type imprinting may act as an interrogable ‘road-map’ of cell movement and development stages, opening avenues in single cell sequencing, flow cytometry and *in-vivo* imagery for DC biology and haematopoietic cell differentiation networks. Similar processes have already been studied in the form of global regulatory elements in macrophages, monocytes and neutrophils (Lavin et al., 2014), yet with the advent of single cell sequencing and *in-vivo* cell tracking, this could be applied to individual cells to produce true developmental cell-tracking across tissues.

3.4.5 Producing a Dataset for Signature Generation

A major aim for this chapter was the generation of robust cell-specific gene signatures. To develop this, a large dataset of well-defined cell subsets was required. Using two separately generated Illumina expression sets as the basis for a signature provided a much greater statistical power for differential expression testing and expanded the repertoire of samples to include approximately twice the initial number of healthy individuals used in section 3.4.1, thereby incorporating and accounting for greater individual variation in each cell type population. Later analysis in Chapter 4 of primary blood and cultured cells highlighted the difference in expression of cells from different individuals deemed ‘identical’ by FACS gating and builds on the concept of environmental transcriptional imprinting that a cell population develops based on it’s environment, interactions with other cell types and contact with foreign materials, analogous to the transcriptional changes associated with a cell transition into tissues highlighted in section 3.4.4.

Combining the datasets required the implementation of a specialist normalisation technique ‘ComBat’. pDC samples were selected as the normalisation group as previous experimentation suggested that pDCs were the most distinct cell type and contained the least variation between individuals as a result of this. The normalisation was successful as displayed in Figure 3.17 with all samples grouping by subtype. The overall branching of the dendrogram differed from that shown in Figure 3.3, yet given the doubling of samples in the analysis, differences in the hierarchical clustering were expected. Importantly, t-SNE analysis suggested a very strong grouping of the samples by subtype and very low internal variation within each subtype, which was of most concern when initially combining the two Illumina datasets.

3.4.6 GeneSign Signature Generation and Visualisation

Once the outcome of normalisation was deemed complete, signature generation was performed. The fundamental difference between the initial testing performed prior to NanoString panel analysis and GeneSign implementation was that GeneSign tested each gene against every other subset at once, represented in Figure 3.19. This produced much smaller gene lists than the ‘1 vs 1’ individual testing, but resulted in far more robust signatures as every subset was interrogated at once, resulting in gene signatures that were only highly expressed in one subset. This procedure aimed to minimise the inclusion of genes that were equally highly expressed in more than one subset and thus would not be able to define a single cell type.

The number of differentially expressed genes for each subset reflected developmental relationships between the subsets similar to the initial differential expression testing outlined in Figure 3.2 and external file 1. The two cDC subsets are the most closely related developmentally and transcriptionally (McGovern et al., 2015) and this strongly correlate with the number of genes identified as signature genes for each subset. cDC subsets had the fewest signature genes, likely as a result of the transcriptional similarity between cDC1s and cDC2s. CD14+ monocytes had the second greatest number of signature genes assigned to them. As 16+ monocytes were not included in this analysis due to their absence from the 2015 Illumina GeneArray, shared monocyte signatures were likely also assigned to the CD14+ monocyte subset. Had CD16+ monocytes also been included in the analysis, the number of signature genes for CD14+ monocytes will have been reduced as a result of shared or similar expression of monocyte related genes between classical and non-classical monocytes resulting in such genes failing to meet the criteria of 1.5 fold expression in a single subset and p-value of less than 0.05 after false discovery rate adjustment. The pDC subset had the greatest number of assigned signature genes as this subset is developmentally and phenotypically distinct from other DCs and monocytes to a much greater extent than the other subsets analysed here (Castell-Rodríguez et al., 2017). Overall, expression patterns for signature genes were supported by literature and included a number of strongly defining gene targets frequently used for flow cytometric analysis including CD14, CD1c, CLEC10A and PACSIN1 as highlighted in Table 3.6. The signature gene pattern further reflects the dendrogram pattern shown in Figure 3.17, improving confidence in the resulting gene signature. Figure 3.20 was the result of initial visualisation of this analysis section, recapitulating the dendrogram from Figure 3.17, but emphasising the strong blocks of gene expression relating to each subset signature. This diagram was the inspiration behind further gene refinement discussed in section 3.4.8 as it suggested an aspect of redundancy was present in the signature list even upon reduction from full Illumina probe counts to 3,439 genes. Being highly expressed only in a single subset meant fewer genes would still be able to reproduce the same or similar cell type grouping but allow potential for further translation of future experiments into smaller and cheaper technologies or more refined NanoString Panels.

3.4.7 GeneSign Validation and Machine Learning

Promising results from initial visualisation of the DC/monocyte cell type signature list prompted further investigation. In order to remove the user-bias from interpretation of the robustness of the gene signature, a Linear Discriminant Analysis (LDA) supervised machine-learning algorithm was applied to the dataset. The requirement for such analysis was to ensure that the results produced were not artificial or incidental and therefore specific only to the exact dataset and samples used in the initial signature generation. Such a signature would have been inapplicable to any other samples or cell types and therefore inadequate as a DC and monocyte subtype signature list.

LDA from the 'MASS' package was employed in a cyclic fashion across 5 iterations so that every sample was used as part of both a 'training set' – used to define each population, and a 'testing set' – a subset of the samples from which the subset information was hidden from the machine, which was then tasked with assigning a cell type to the 'unknown' samples.

The results of this analysis were 41/41 correct assignments, proving in an unbiased fashion that the cell type gene signature developed in this chapter could accurately predict a sample's subset of origin based on the expression of the 3,439 signature genes with 100% accuracy.

As a final validation procedure, the gene list was applied to an external validation dataset GSE65128 from Lee et al 2015. This dataset was not worked on or previously seen by the Human Dendritic Cell Lab or other research group at Newcastle University and therefore was subject to differing methodology for cell sorting, sample preparation and subset isolation, however the samples were taken from comparable healthy donors and so while there may be technical variance, biological variance was not expected to be an issue. This dataset was selected as it contained Illumina expression data on the four major DC and monocyte subsets investigated in this thesis, although the underlying sorting strategy was not identical to that used in GSE35457. This final validation was the major authentication hurdle to overcome. Applying the signature to a previously unseen dataset challenged its applicability to other datasets, samples and technology. The combined heatmap and dendrogram produced in Figure 3.22 and accompanying t-SNE visualisation in Figure 3.23 were testament to the successful generation of a robust and reproducible GeneSign signature for the assignment and grouping of the major DC and monocyte populations. The heatmap, was almost identical to the heatmap produced from the GSE35457 dataset in Figure 3.20 despite GSE65128 data not being used in any step of the signature generation experiment.

3.4.8 Analysis of and Uses for Gene Minimisation Experiments

The robust DC and monocyte gene signature contained hundreds of genes for each cell subset. This provided the greatest base for testing reproducibility during the machine learning analysis and subsequent use on external datasets. This 3,439 gene signature was capable of distinguishing the cell subsets in a manner comparable to the full Illumina expression array, proving a smaller geneset could be used in future experiments, however visual assessment of Figure 3.20 and 3.22 and strong between sum-of-squares for each of the dendrogram branches suggested a high degree of redundancy in the genelist. In an effort to minimise the gene list to the fewest genes required to maintain the grouping of cell types produced by the full Illumina array and the GeneSign signature geneset, a novel method of gene reduction was developed for this thesis.

Many methods of dimensionality reduction such as PCA and t-SNE produce pseudo-variables across a plane in the data to account for variability or reconstruct the dataset in low-dimensional space, yet these pseudo-variables are a combination of true variables and thus do not have a biological relevance. A t-SNE variable or PCA weighting can't be interrogated directly after minimisation in a research setting or wet-lab experiment. To overcome this, a randomised gene-removal script was written in 'R' to repeatedly reduce a gene signature down to the minimial number of variables required to maintain the cell subset groupings produced from the initial parent dataset. The essential goal to this experiment was to produce a minimised gene list to remove redundancy from the signature gene list, but still retain the individual gene information for each of the remaining genes. Unlike a PCA variable, this script would not produce pseudo-variables to account for the variance in the data, but remove unnecessary genes that were not powering the cell type distinction. By writing this script, new cell markers could be investigated for potential as sorting or flow cytometry targets resulting in easier distinction of cell types from a mixed cell population. By reducing the number of markers required to define a cell and thereby freeing up spare channels in the cytometer, a researcher could additionally investigate other cell properties or marker genes in a single experiment.

The technique behind the gene reduction experiment was relatively simple. The full genelist and expression data to be reduced was uploaded and K-means testing was performed to determine the groups that the samples clustered into. This grouping was then used as a framework for all subsequent iterations of the reduction programme to ensure that final genelists would be able to define and group all of the same populations as the initial dataset. The between-group sum-of-squares (the 'distance' between each subset cluster) was recorded too.

Once the baseline grouping was established, a gene was selected through a random number generator to be removed from the dataset at which point the k-means grouping was re-tested against the initial dataset, if the same grouping was maintained, the new between-group sum-of-squares was tested for the new gene list and if the distances between the groups was increased, the new dataset replaced the full dataset and the process was repeated. If the new dataset failed any of the testing steps the removed gene was replaced and another selected at random. For such a randomised process to be successful, a larger number of iterations was required. After some testing, 10,000 iterations of the reduction process was deemed adequate to reach the gene number end-point. The entire process from full gene list to minimal gene list was produced 2,000 times with the output recorded for every iteration.

The random nature of the technique produced resulted in a number of different gene combinations that could successfully recreate the full dataset dendrogram and k-means clusters. These final gene lists ranged from 2 genes to approximately 15 genes in length with many of the same gene combinations reoccurring in the final gene lists. One of the most common combinations was FCER1A and SLC2A3 displayed in Figure 3.25, both markers exhibited plasma membrane bound protein expression and had antibodies available from major companies. Such gene combinations found by the gene reduction script may prove useful for flow cytometry experiments as four subsets could be reliably distinguished by just two markers. A combination of IRF8 and IRF4 markers could also perform this feat and has been noted in published articles as markers of terminal differentiation in CD11c+ cells (Bajaña et al., 2016, p. 4; Vander Lugt et al., 2014). This marker combination has been utilised for flow cytometry by the Human Dendritic Cell Laboratory.

A drawback of randomised gene minimisation appeared to stand out during analysis. The random nature of the process when combined with the high number of possible successful gene combinations resulted in many different gene combinations being produced. Only relatively rarely did the exact same gene combination appear as a final gene list, although some genes appeared frequently, suggesting they exhibited a strong distinguishing pattern of expression. This was interesting from a research standpoint and highlighted how simply DC and monocyte subsets could be defined, but did not provide much reproducibility. Further refinements to the code could have improved reproducibility if desired, including subsetting the data to cell surface markers if a user wished to find new flow cytomertry marker genes or cell cycle genes if the aim was to investigate cell cycle processes. Weighting the genes first by PCA and then taking only the most variable genes on to the minimisation step may also be an option, particularly if the initial dataset was very large or was not first pruned down to cell type signatures.

Another consideration when applying the information gained from this work into further experiments would be the differences in the data dynamic range. While not considered in the example used in this thesis, it is possible to find genes with variable expression across the cell subsets, allowing a single gene to group samples by cell type. This would be simple for Illumina expression data, which ranges from around a $\log_2$ value of six to sixteen, or NanoString counts which are direct counts typically ranging from zero to around 500,000, however less dynamic range is available in older techniques such as flow cytometry where a cell is labeled as ‘negative’ for expression from an intensity value of 10^3, but strongly ‘positive’ in expression upwards of 10^5. Such ‘golden’ genes with variable expression patterns across each cell subset by Illumina GeneArray would likely appear as a blur or smear by flow cytometry. For flow cytometry applications, targets with only very high or very low expression should be considered, such as FCER1A expression which was found to be high on pDC and cDC2 cells, but low on cDC1 and CD14+ monocytes in this thesis.

3.5 RESEARCH SUMMARY AND KEY POINTS FOR PROJECT PROGRESSION

Dendritic cells are extremely rare cells, despite their prominent location in peripheral blood and skin as well as ubiquitous expression throughout the body. They play a fundamental role in directing the immune response under viral, bacterial, fungal and malignant insults, making them extremely valuable in immunological research. Multiple DC subsets have been identified and shown to be extremely specialised to their role in immunity, from unique plasmacytoid DCs, specialised to release interferon-α in response to viral infection, to cDC1 DCs, capable of secreting interferon-λ and IL12 and highly capable of cross-presentation of necrotic antigens to CD8+ T-cells, and the more abundant cDC2 DCs which specialise in anti-fungal and anti-microbial response. However, their major identifying surface markers are extremely fickle in nature, making DC research by cell isolation a tedious and complicated process.

In this chapter, a pipeline for DC analysis via NanoString Technology was established and a novel method of deconvoluting cell signatures across tissues and culture conditions was developed for use in the subsequent chapters. RNA transcriptome profiling was implemented here to reveal further possible DC and monocyte specific genes that may aid researchers in distinguishing and isolating dendritic cells from peripheral blood, culture conditions (where currently used surface markers have shown to be unpredictable) and across tissues through the generation of a robust transcriptome-based mRNA gene signature. This signature resulted in the identification of known and novel DC subset marker gene targets, including S100A8 and C19orf59 on monocytes, PACSIN1 and GZMB on pDCs and FCER1A cDC2s. All of these genes encode surface membrane-bound proteins and are thus ripe for use in immunohistochemistry, flow cytometry and FACS analysis, indeed many of the 3,439 genes identified as markers of individual DC or monocytes subsets encoded surface proteins, reflecting their high capacity for cell signaling and antigen presentation and the importance of cell-to-cell communication in orchestrating the immune response.

In order to ensure that the signature produced from the illumina expression analysis was robust and diverse enough to be applicable to later chapters of this project, the signature was applied to an external dataset and firmly recapitulated the layout and hierarchy produced in the initial dataset. This validating experiment may indicate that this gene signature can be applied to any other DC transcriptome datasets, and may be of use to the wider immunology field, providing a template for cell identification. Such a signature would be extremely valuable in the field of single cell RNA-sequencing, where one of the major problems facing researchers is the annotation of the cells collected and processed by the sequencing platforms. The signature was used for a similar purpose in chapter 5, for the identification of mature DC signatures in precursor populations. Individual cells could be correlated to mature cell populations by their shared expression patterns.

In order to address the question of scope for the identification of monocytes and dendritic cells, an immune-based panel of markers were applied to cells from the same gating as the illumina expression dataset, but using the quicker, cheaper and more accurate NanoString nCounter platform. Using this platform would open up a greater array of samples for analysis, particularly due to the much lower RNA requirements for the nCounter platform compared to the Illumina BeadArray. This was an important consideration to this project as the cDC1 populations of interest was particularly rare, even amongst other DC sub-populations.

Analysis of the NanoString dataset revealed a strong correlation between the illumina and NanoString platforms, solidifying the research groups' switch over of platforms. Furthermore, this platform and analysis pipeline was then tested and applicable to the comparison of culture and blood-derived DCs and monocytes in chapter 4. After the highly correlative results from the NanoString platform were noted, it was evident that relatively few, strong markers of DCs and monocytes could potentially be used to distinguish the subsets efficiently. To investigate this, a novel method of geneset minimization was developed by the author and applied to the Illumina expression dataset, exposing 2-5 genes that could reliably identify the three DC subsets and monocyte subset of interest. One of these combinations, FCER1A and SLC2A3 were displayed in this thesis as en example, but others including a combination of IRF8 and IRF4 were also identified. By highlighting that four rare, developmentally related and phenotypically similar could be distinguished by the expression of just two markers was quite revealing and may prove useful in future work as a way to isolate these subsets by FACS with fewer antibodies, for the purpose of cost reductions over the 6-10 antibodies currently used to define these populations, or to open up fluorescent channels for the inclusion of other antibodies in the reaction mix to provide other important information about the cells or further distinguish sub-populations of interest.

Chapter 3 Figures & Tables

Human Mononuclear Subset	Number of Replicates
Blood CD14+ Monocytes	6
Blood CD16+ Monocytes	6
Blood 141+ DCs (cDC1)	5
Blood 1c+ DCs (cDC2)	6
Blood pDCs	4

Table 3.1: Human blood samples extracted from GSE35457 for identification of human mononuclear cell signature genes

Samples sorted from human blood for Haniffa *et al*, 2012 were used as well-defined monocyte and dendritic cell subsets for the generation of initial gene signatures and *in-silico* testing of NanoString Technology's Immunology_v2 human nCounter gene expression array panels.

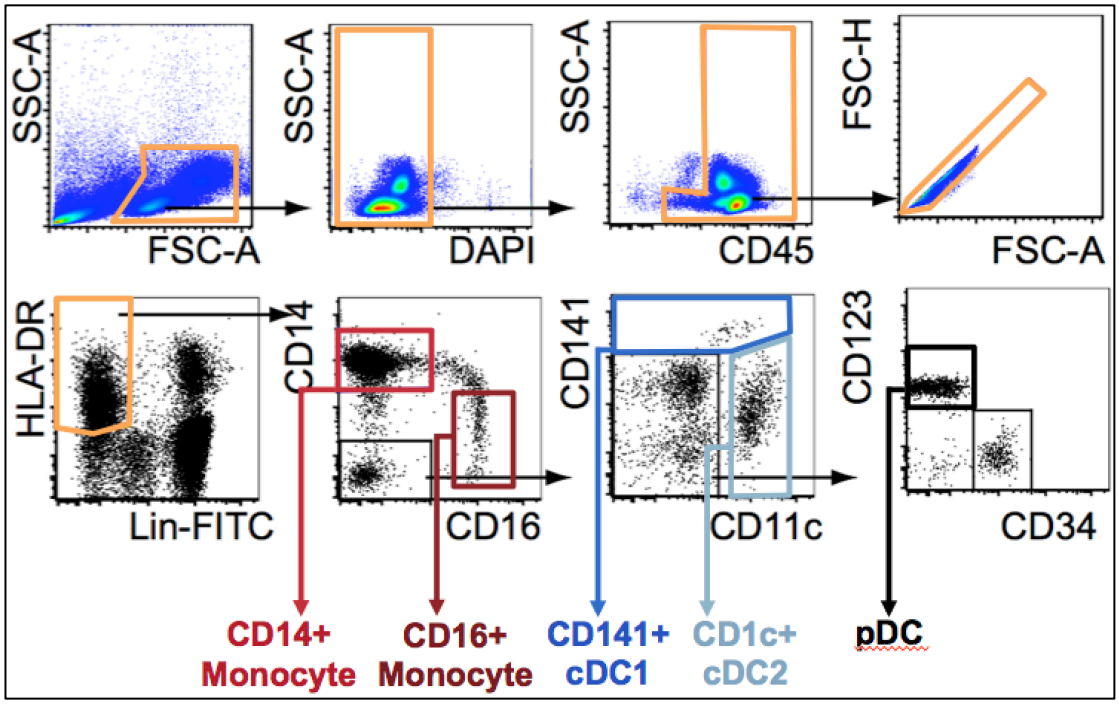


Figure 3.1: Gating strategy for GSE35457 human blood samples for Illumina expression data

Full gating strategy to identify human DC and monocyte subsets in the blood. Lineage makers were CD3, CD19, CD20 and CD56 all in FITC channel. Panels were arranged from upper-left to bottom-right with the following rationale:

1. Large cells (FSC+), 2. Live cells (DAPI-), 3. HSC-derived (CD45+),

4. Singlets (FSC-A x FSC-H), 5. MHCII-expressing (HLA-DR+), Lineage-,

6. CD14+ monocytes (CD14+,CD16-), CD16+ monocytes (CD14-, CD16+),

7. cDC1 (CD141+), cDC2 (CD11c+), 8. pDC (CD123+,CD34-)

Human Mononuclear Subset	Gating Strategy
Blood CD14+ Monocytes	Live CD45+ singlets HLA-DR+ Lin- CD14+ CD16-
Blood CD16+ Monocytes	Live CD45+ singlets HLA-DR+ Lin- CD14lo CD16+
Blood 141+ DCs (cDC1)	Live CD45+ singlets HLA-DR+ Lin- CD14- CD16- CD123- CD11clo CD1c-/lo CD141+
Blood 1c+ DCs (cDC2)	Live CD45+ singlets HLA-DR+ Lin- CD14+ CD16- CD123- CD11c+ CD1c+ CD141-
Blood pDCs	Live CD45+ singlets HLA-DR+ Lin- CD14+ CD16- CD123+

Table 3.2: Gating strategy for GSE35457 human blood samples for Illumina expression data

Full gating strategy to identify human DC and monocyte subsets in the blood. All subsets were classed as 'live' by DAPI staining, CD45+, single-cells, negative for lineage markers (CD3, CD19, CD20 and CD56 all in FITC) and HLA-DR+. Monocytes were split into classical and non-classical by CD16 and CD14, while cDC1 and cDC2 were split by CD14, CD11c, CD1c and CD141.

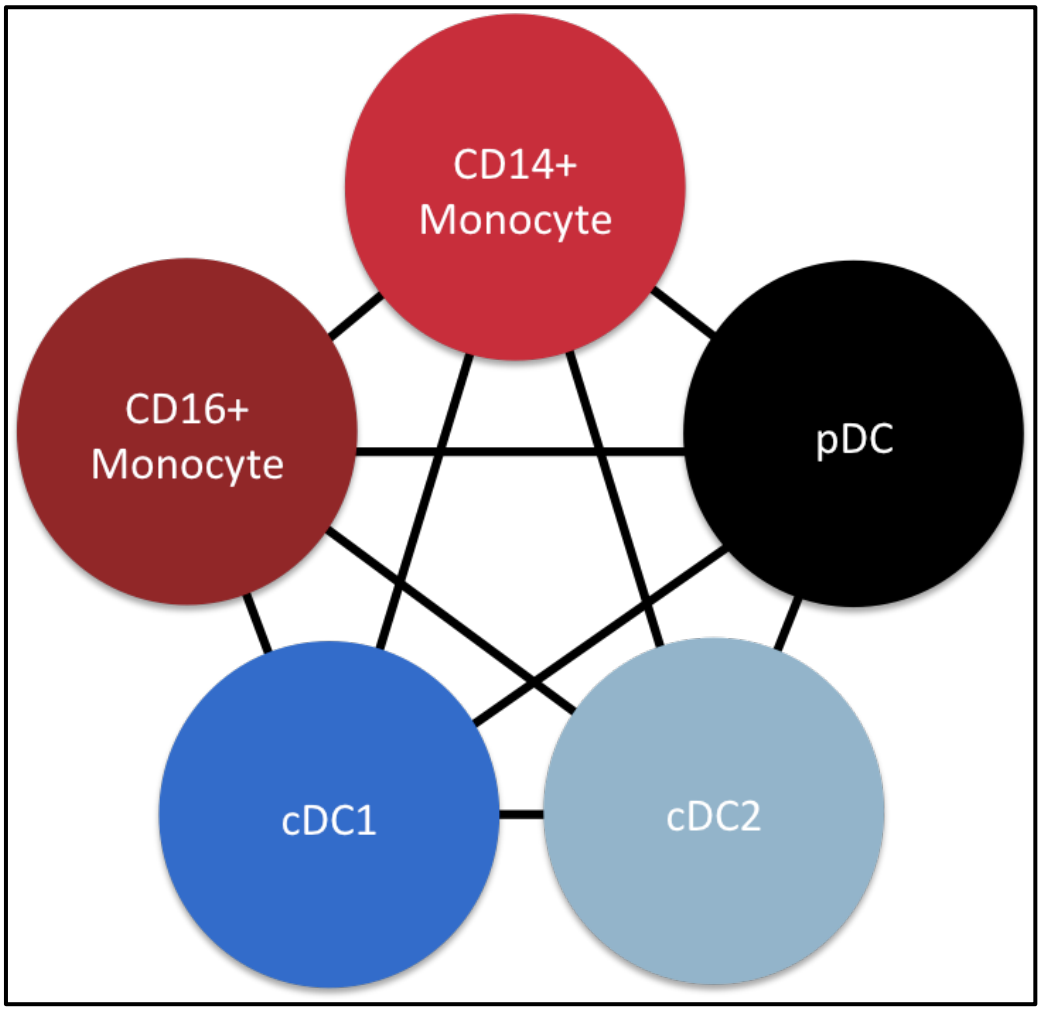


Figure 3.2: Schematic representation of the '1 vs 1' differential expression technique

This network schematic shows the comparison table layout for all iterations of the '1 vs 1' approach to subset analysis. In this case, each subset was compared against each individual subset in a pairwise manner before the next subset is compared against the remaining subsets for a total of 10 comparisons.

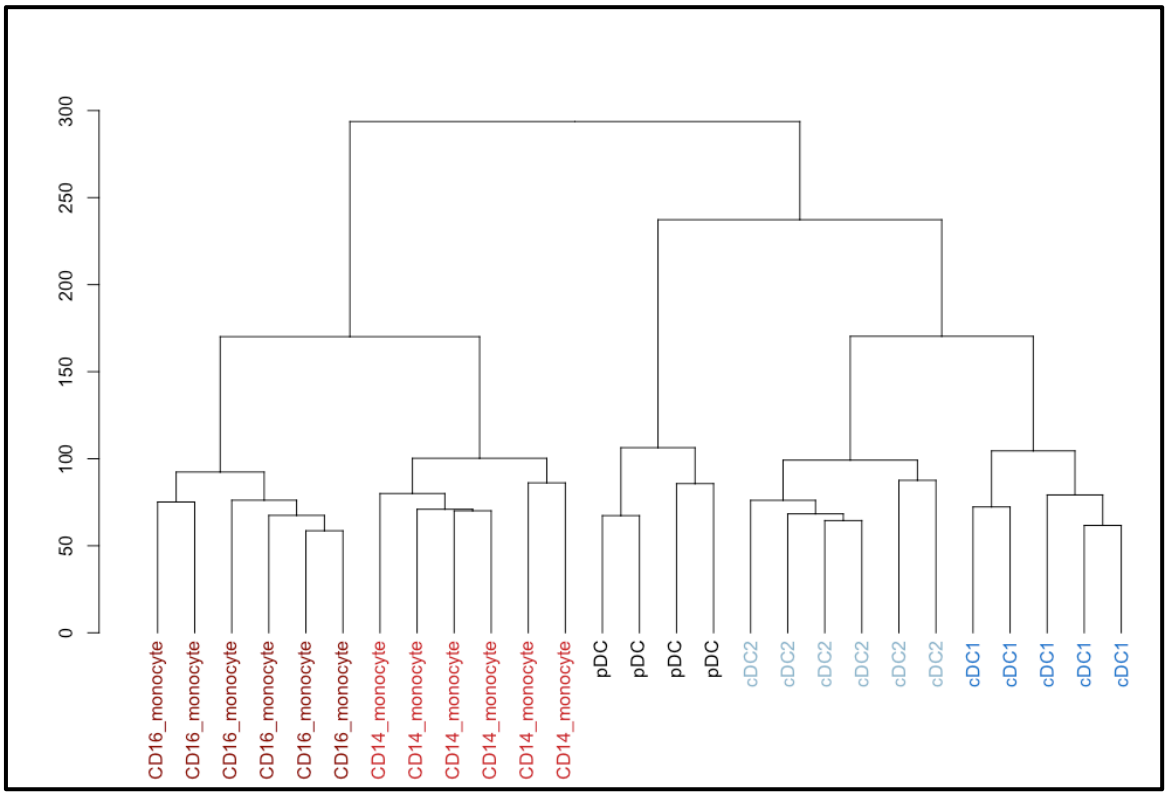


Figure 3.3: Hierarchical clustering of human blood mononuclear cells using Illumina expression data

Hierarchical clustering of the full Illumina array data for human blood subsets was capable of distinguishing each population and group them into their developmental cell types. Here, sample labels were assigned based on the FACS gating used to isolate the cells as displayed in Figure 3.1 and Table 3.2. The Y-axis in this figure represents 'height' (a measure of increasing dissimilarity), increasing 'height' suggests clusters are less similar to one another. The first branch of the dendrogram splits monocytes from dendritic cells. Further along the hierarchy, monocytes are split into their conventional (CD14+) and non-conventional (CD16+) major monocyte subsets, while the dendritic cell branch splits into plasmacytoid DCs (pDC) and classical DC types. In turn, the classical DC branch is split into CD1c+ cDC2 samples and CD141+ cDC1 samples.

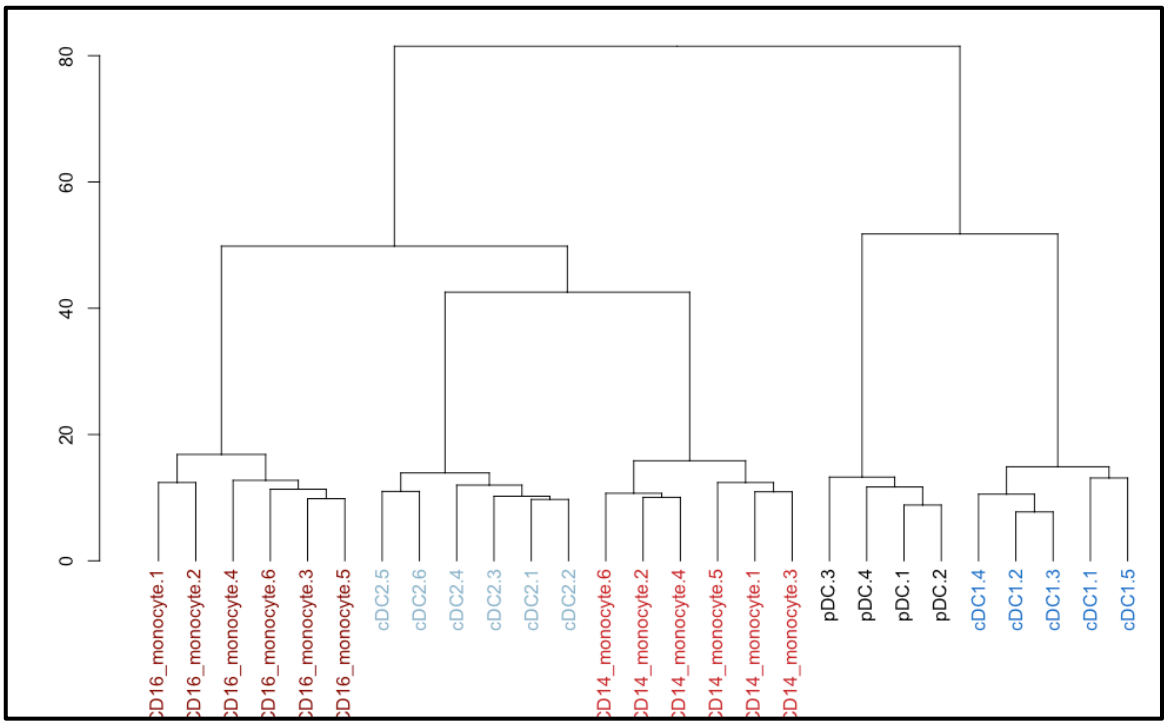


Figure 3.4: Hierarchical clustering of NanoString Immunology_V2 mapped human blood Illumina expression data

The 543 gene dataset of NanoString Immunology_V2 genes translated into the Illumina platform produced an *in-silico* representation of how the NanoString machine would likely perform at clustering the common monocyte and dendritic cell subsets. The Y-axis in this figure represents 'height' (a measure of increasing dissimilarity), increasing 'height' suggests clusters are less similar to one another. Here, sample labels were assigned based on the FACS gating used to isolate the cells as displayed in Figure 3.1 and Table 3.2. While each subset was grouped correctly, with all populations appearing grouped together, cDC2 samples were grouped between the CD16 monocyte cluster and CD14 monocyte cluster, rather than with the pDC and cDC1 clusters. This could be the result of the clustering method used, the restricted dataset not providing enough

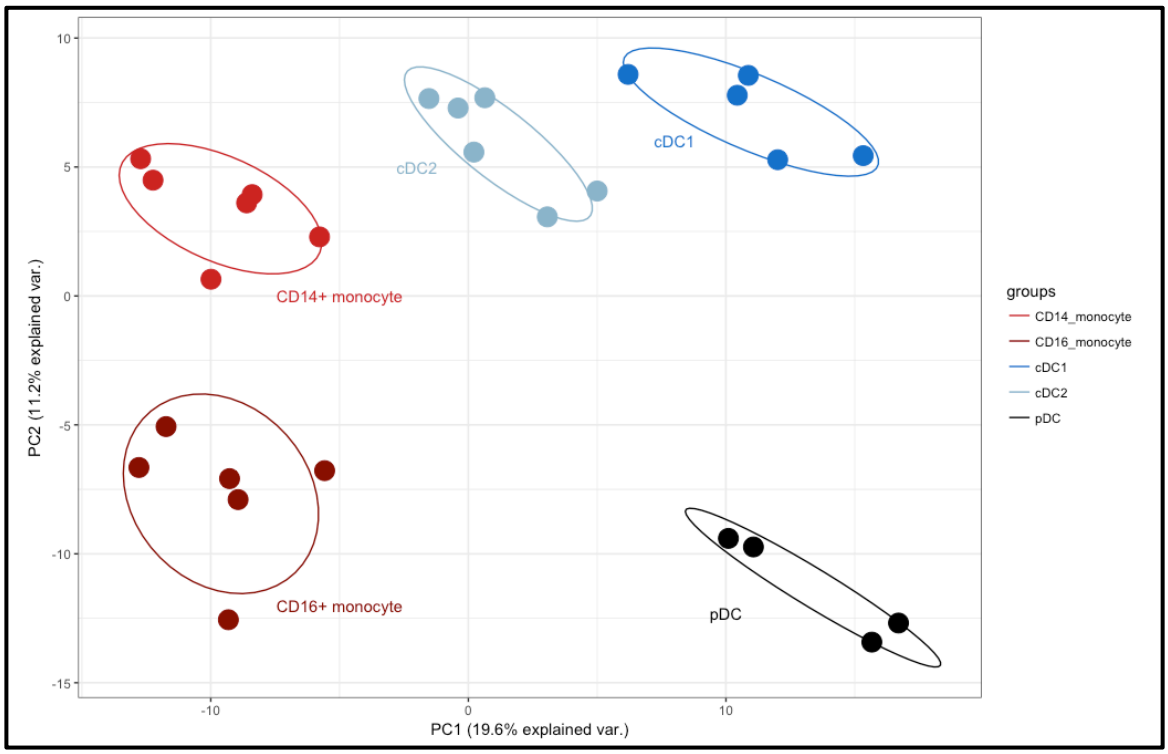


Figure 3.5: Principal component analysis of NanoString Immunology_V2 mapped human blood Illumina expression data

PCA analysis of the surrogate Illumina expression data resulted in five defined subset populations. PC1 accounts for 20% of sample variance and appears to split the DC populations from the monocyte populations. pDCs and cDC1 appear on the positive axis of PC1, with the monocyte populations on the negative region of PC1. cDC2 subset samples are found at the zero point. PC2 accounts for 11.2% variability and splits the two monocyte subsets, as well as splitting cDC1 from pDC samples.

Distance relationships appear as expected with pDC samples far from all other subsets and the cDC populations relatively close to each other.

ASIP	DAXX	MERTK
C19orf59	DBN1	Ki67
CCL17	F13A1	NDRG2
CD1c	FGD6	PACSIN1
CD207	FLT3	PPM1N
CLEC10A	GCSAM	PRAM1
CLEC9A	GGT5	S100A12
CLNK	LPAR2	TMEM14A
COBLL1	LYVE1	UPK3A
CXCL5	MAFF	ZBTB46

Table 3.3: Gene list for NanoString Panel+ custom codeset

This table lists the 30 genes that were selected as additional NanoString gene probes to add to the Immunology_V2 gene expression panel.

These genes contained a number of known subset marker genes, the full rationale for each gene can be found in Appendix B. Briefly, these additional genes included CD1c, ASIP, PACSIN1 and CLEC9A that were not present on the standard NanoString panel. Also included were genes related to cell state such as Ki67 and LYVE1. Furthermore, as the panel was designed to be applicable to projects outside the scope of this thesis, genes relevant to other work were also added including FLT3, CD207 and F13A1. A full rationale for the selection of these 30 genes and their functions are given in appendix B.

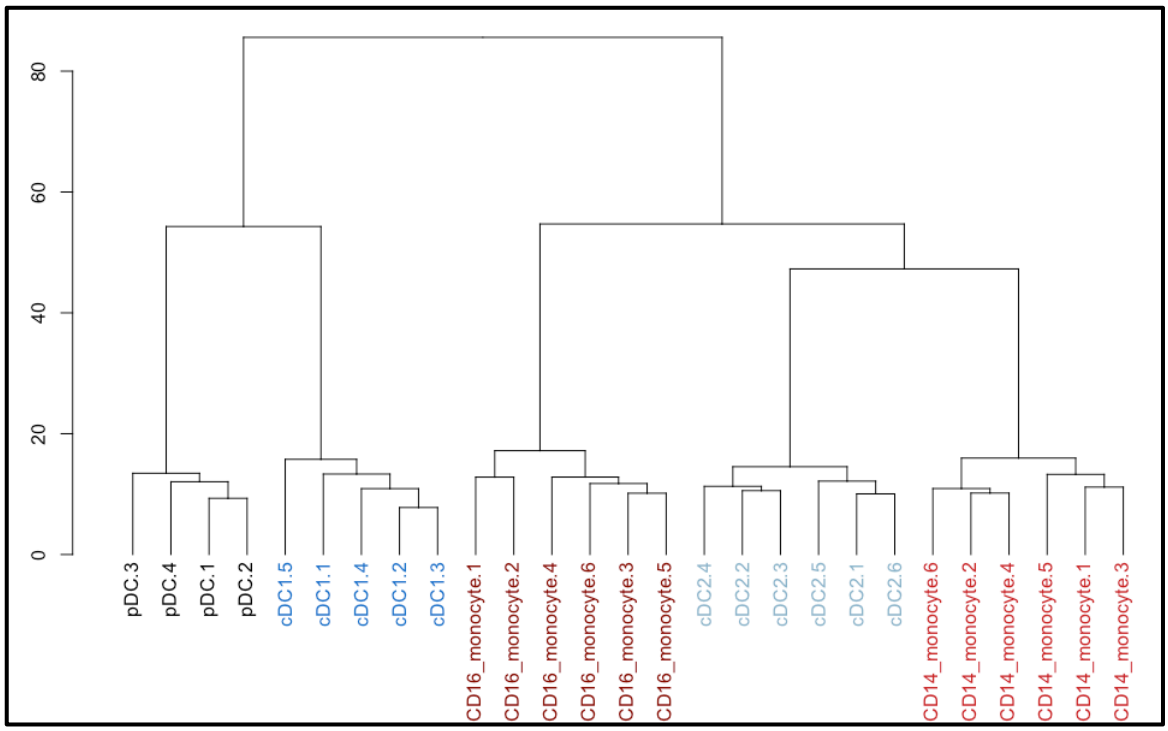


Figure 3.6: Hierarchical clustering of NanoString Immunology_V2 with Panel+ genes mapped Illumina expression data of human blood mononuclear subsets

The NanoString Immunology_V2 genes with additional Panel+ custom codeset included were translated into the Illumina platform IDs, which produced an i*n-silico* representation of how the NanoString machine would likely perform at clustering the common monocyte and dendritic cell subsets with the addition of custom probes. The Y-axis in this figure represents ‘height’ (a measure of increasing dissimilarity), increasing ‘height’ suggests clusters are less similar to one another. While each subset was grouped correctly, with all populations appearing grouped together, cDC2 samples were still grouped between the CD16 monocyte cluster and CD14 monocyte cluster, rather than with the pDC and cDC1 clusters as seen in Figure 3.4. While this was an unexpected result given the inclusion of a number of cDC-specific gene targets, 30 additional genes might not have contributed significantly to the overall hierarchical clustering. At the individual gene level, expression values for these subset marker genes were as expected.

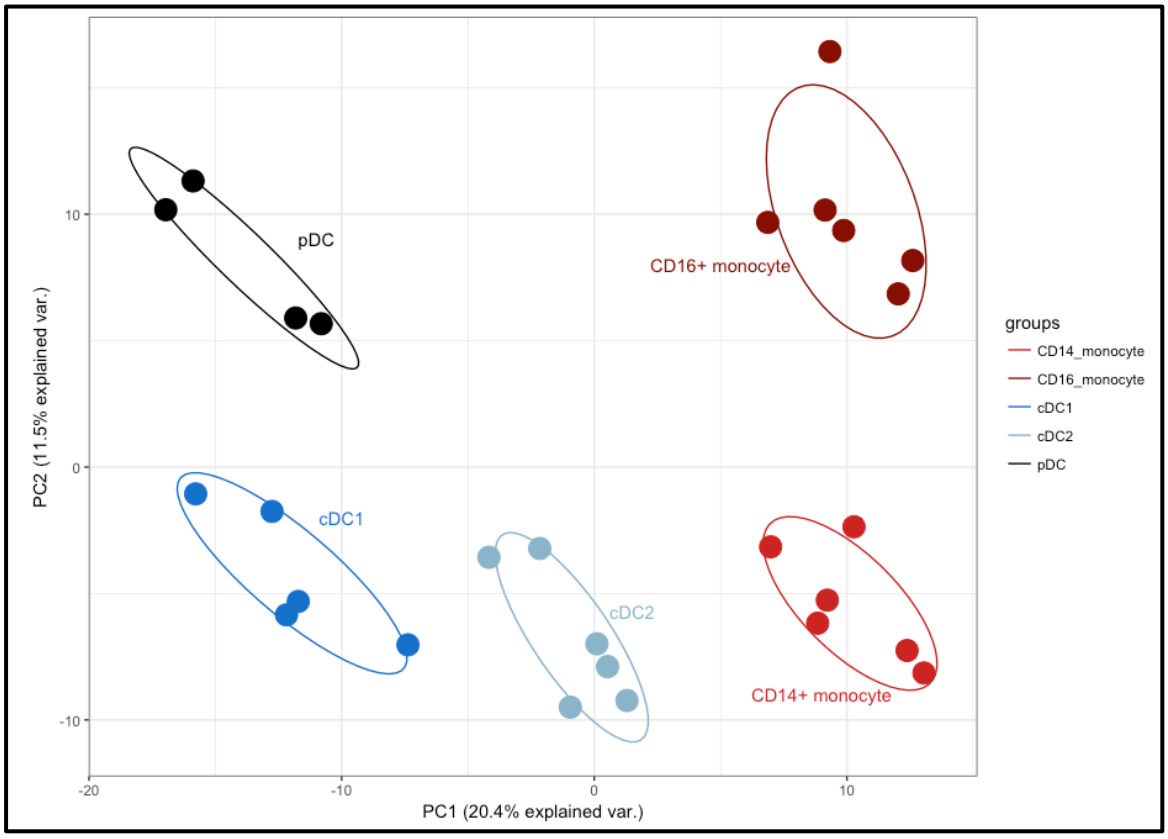


Figure 3.7: Principal component analysis of NanoString Immunology_V2 with Panel+ genes mapped human blood Illumina expression data

PCA analysis of the surrogate Illumina expression data with additional Panel+ genes resulted in five defined subset populations. PC1 accounts for 20.4% of sample variance and appears as Figure 3.5 to split the DC populations from the monocyte populations. pDCs and cDC1 appear on the negative region of PC1, with the monocyte populations on the positive region of PC1. cDC2 subset samples are found at the zero point as they were without the Panel+ genes. PC2 accounts for 11.5% variability and splits the two monocyte subsets, as well as splitting cDC1 from pDC samples.

Distance relationships appear as expected with pDC samples far from all other subsets and the cDC populations relatively close to each other. Very minor increases in PC1 and PC2 explained variance could be attributed to the inclusion of the additional panel+ gene probe although the overall look of the PCA is similar to the non-Panel+ PCA.

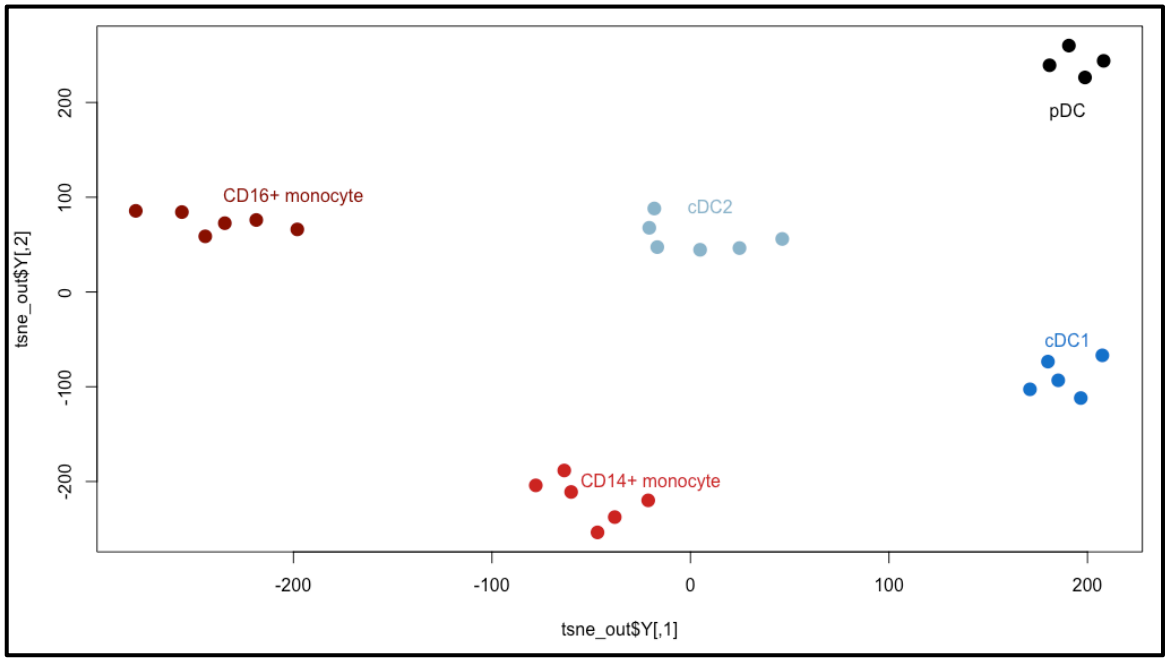


Figure 3.8: t-SNE analysis of NanoString Immunology_V2 with Panel+ genes mapped human blood Illumina expression data

T-SNE analysis of the surrogate Illumina expression data with additional Panel+ genes resulted in sufficient clustering of the five defined subset populations. In comparison to PCA, this t-SNE plot has a similar overall appearance although cDC2 samples are located close to the zero value along both the x- and y-axis. T-SNE variable 1 (x-axis) splits the DC populations from the monocyte populations. pDCs and cDC1 appear on the far right-hand side, cDC2 subset samples are found at the zero point with CD14+monocytes averaging a value of -50 and CD16+ monocytes found at -200. t-SNE 2 splits pDCs and cDC1 samples at around 200 and -100 respectively. CD16+ monocytes and cDC2 samples occupy the same region of t-SNE space on this axis, while CD14+ monocyte samples are at the extreme negative region of the plot.

Distance relationships are not proportionate in this type of analysis so although the subset appears at different regions in 2D t-SNE space, their ‘closeness’ to other subsets cannot be determined by this method.

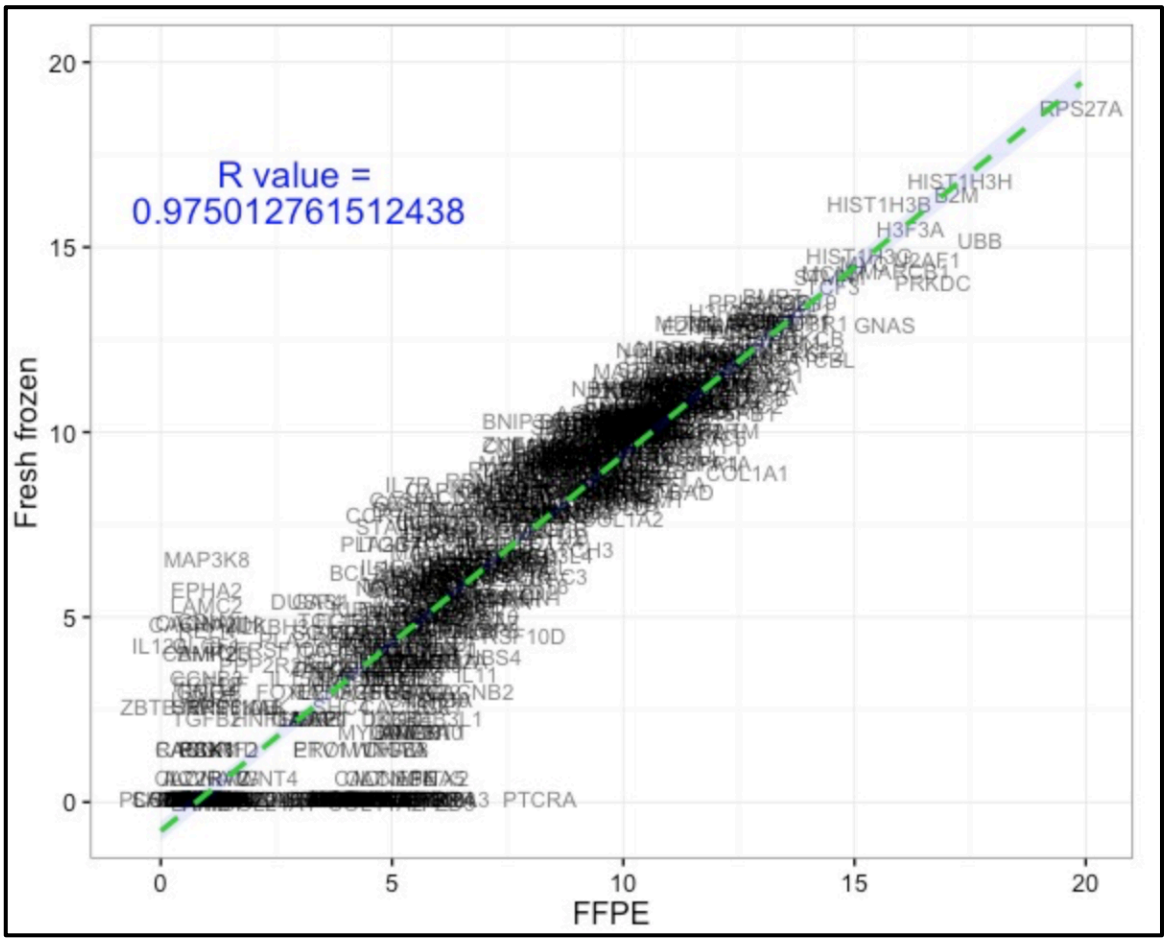


Figure 3.9: Correlation testing of Qiagen extracted mRNA from fresh frozen material and from FFPE material on the NanoString nCounter Analysis System

Correlation is highly conserved between fresh-frozen and Formalin-Fixed, Paraffin-Embedded material, despite the highly degraded nature of FFPE-based RNA.

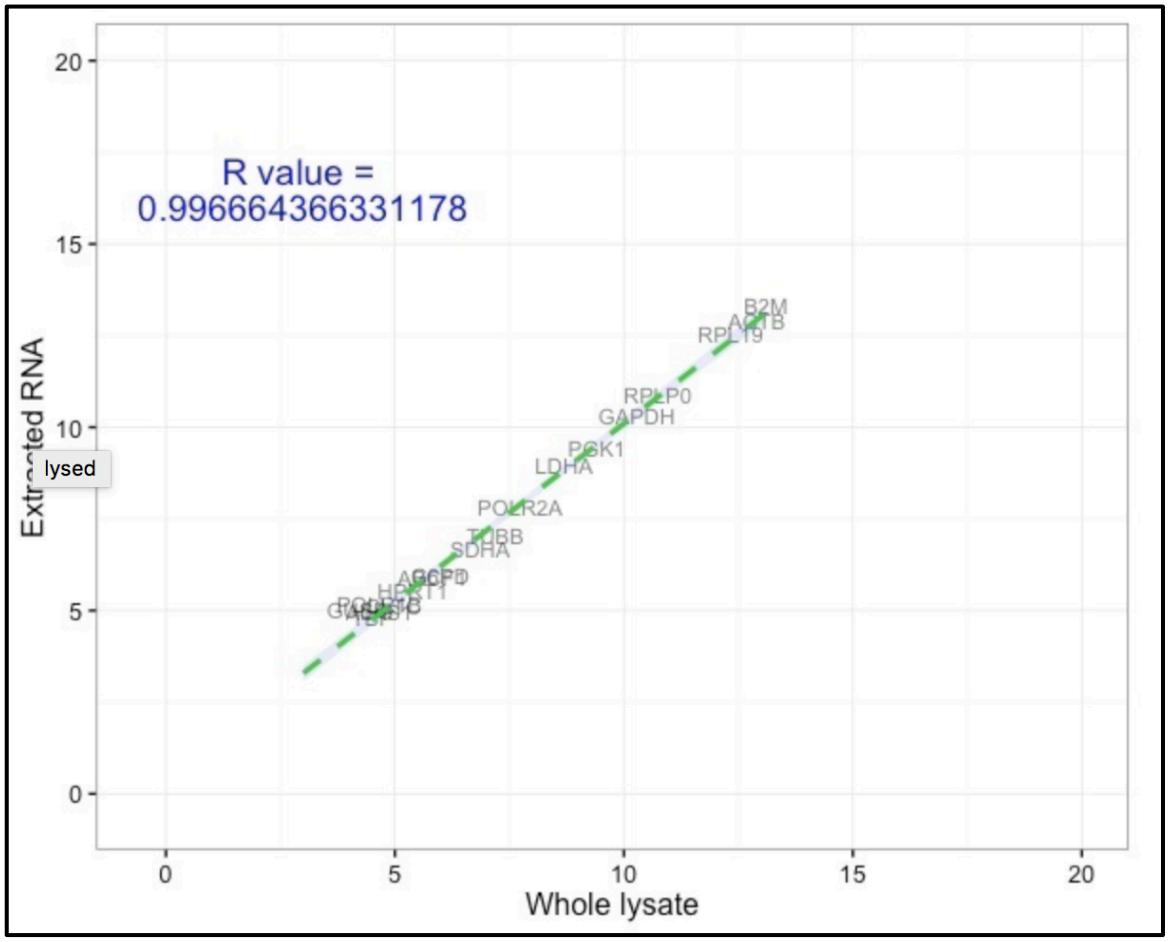


Figure 3.10: Correlation testing of Qiagen extracted mRNA against whole cell lysates on the NanoString nCounter Analysis System

Almost perfect correlation was noted between extracted and lysed, matched donor samples when using the NanoString Analysis system. This test provided the basis for the protocol used in chapter 3 and chapter 4 as lysing cells directly after FACS provided comparable data to those undergoing RNA extraction via RNeasy Kits (Qiagen).

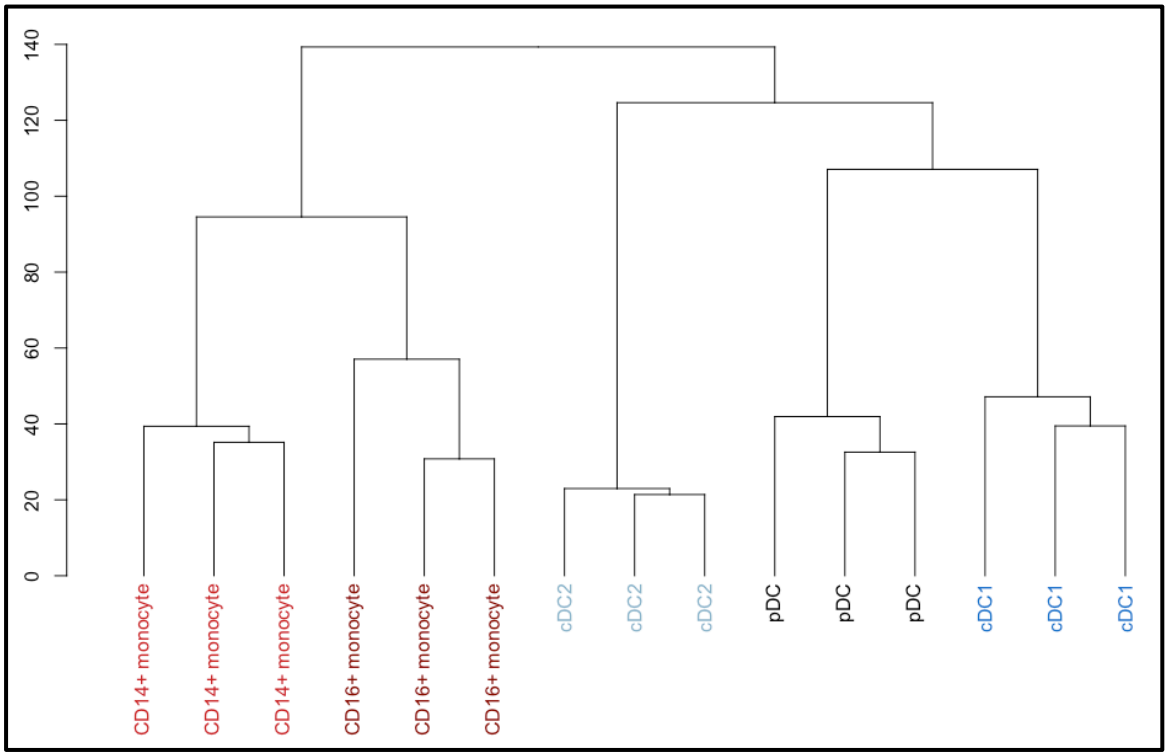


Figure 3.11: Hierarchical clustering of human blood mononuclear cells using NanoString Immunology_V2 with Panel+ genes and the NanoString nCounter platform

This hierarchical clustering dendrogram uses data generated on the NanoString nCounter Analysis platform with the Immunology_V2 codeset and the Panel+ additional probes from Table 3.3. The Y-axis in this figure represents 'height' (a measure of increasing dissimilarity), increasing 'height' suggests clusters are less similar to one another. The NanoString system was capable of providing robust data with enough genes present to separate each of the 5 mononuclear cell subsets. Unlike the *in-silico* experiment using surrogate Illumina expression data (Figure 3.6), the first branch of this dendrogram clearly separates the monocyte subsets from dendritic cell subsets. Each CD14+ monocyte sample grouped together, as did the CD16+ monocyte samples. Along the dendritic cell branch of the dendrogram, cDC2 cells branched off from the other DCs first, followed by a separation of pDC samples from cDC1 samples further down the tree. Differences in dendrogram layout between the Illumina-based experiment and this NanoString based work cause be the result of the individual probe targets and underlying differences in the sample preparation protocols for each technology being fundamentally different.

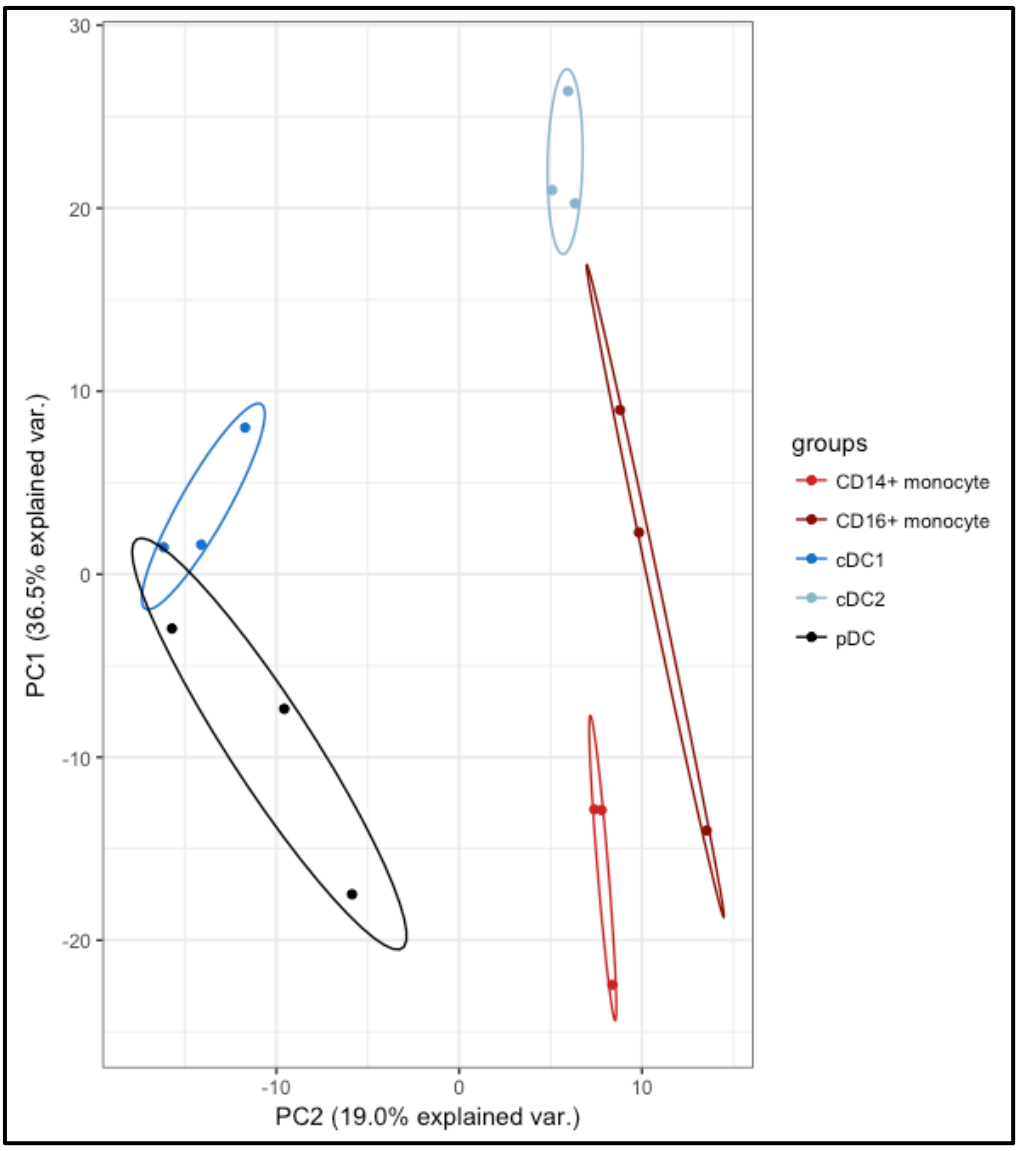


Figure 3.12: Principal component analysis of human blood mononuclear cell subsets on the NanoString nCounter Analysis platform using Immunology_V2 and Panel+ probesets

PCA analysis of the human mononuclear cell subsets resulted in grouping of each of the five defined subset populations. PC1 accounts for 36.5% of sample variance and defines separation of the three DC populations. cDC2 samples appear high on PC1 with a value of approximately 23, with cDC1 samples appearing around a value of 5. pDC samples are on the negative portion of the PC1 axis. While two of the CD16+ samples appear positive on PC1, the other sample is found near the CD14+ monocyte cluster in the negative region of PC1. PC2 accounts for 19% variance and splits cDC1 and pDC samples from the monocyte and cDC2 samples. Although grouped with the other DC samples by hierarchical clustering in Figure 3.9, cDC2 appears separate from cDC1 and pDC clusters by PCA.

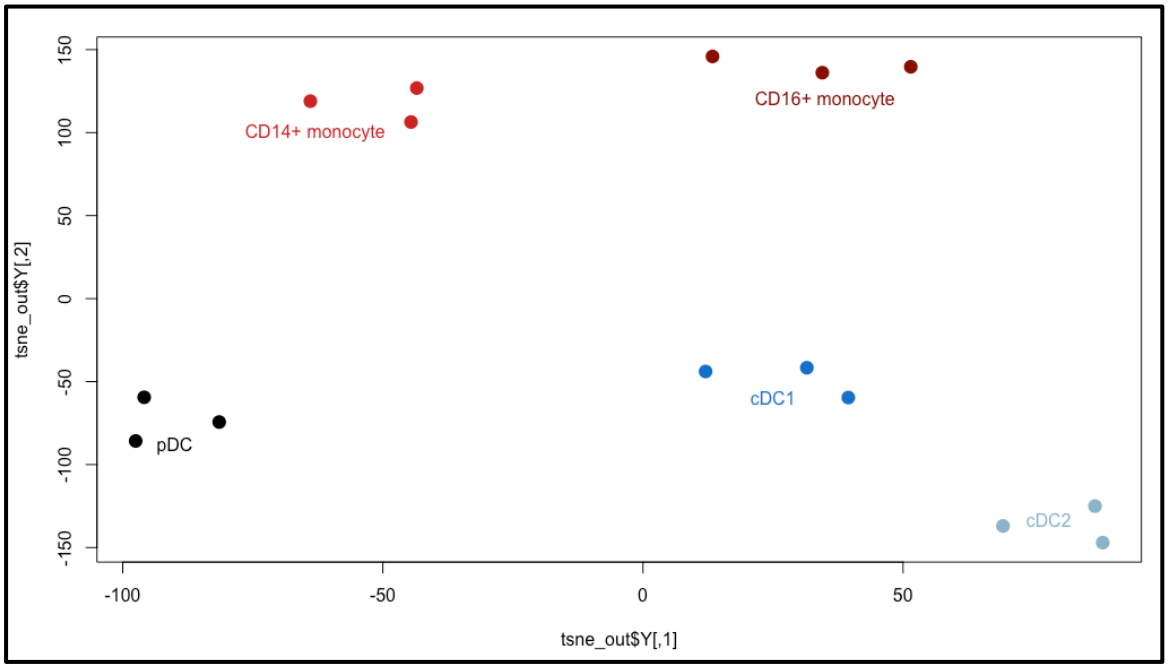


Figure 3.13: t-SNE analysis of human blood mononuclear cell subsets on the NanoString nCounter Analysis platform using Immunology_V2 and Panel+ probesets

t-SNE blood mononuclear cells by NanoString produced sufficient clustering of the five defined subset populations. In comparison to PCA, this t-SNE produced clear and distinct grouping of each subset. T-SNE variable 1 (x-axis) splits the two monocyte populations as well as splitting the pDC samples from the other DCs. cDC2 samples are found at the extreme positive end of the t-SNE1 range, cDC1s and CD16+ monocytes at the mid-range positive end and CD14+ monocytes and pDCs in the negative region. t-SNE2 splits the DC populations from the monocyte populations.

Distance relationships are not proportionate in this type of analysis so although the subset appears at different regions in 2D t-SNE space, their ‘closeness’ to other subsets cannot be determined by this method.

Human Mononuclear Subset	Gating Strategy
Blood CD14+ Monocytes	Live CD45+ singlets HLA-DR+ Lin- CD14+ CD16-
Blood CD16+ Monocytes	Live CD45+ singlets HLA-DR+ Lin- CD14lo CD16+
Blood 141+ DCs (cDC1)	Live CD45+ singlets HLA-DR+ Lin- CD14- CD16- CD123- CD11clo CD1c-/lo CD141+
Blood 1c+ DCs (cDC2)	Live CD45+ singlets HLA-DR+ Lin- CD14+ CD16- CD123- CD11c+ CD1c+ CD141-
Blood pDCs	Live CD45+ singlets HLA-DR+ Lin- CD14+ CD16- CD123+
Skin CD14+ DC	Live CD45+ singlets HLA-DR+ CD14+ Autoflorescence-
Skin 141+ DCs (cDC1)	Live CD45+ singlets HLA-DR+ CD14- Autoflorescence- CD11c lo CD141+
Skin 1c+ DCs (cDC2)	Live CD45+ singlets HLA-DR+ CD14- Autoflorescence- CD11c+ CD141-

Table 3.4: Sample list and gating strategy for blood and skin derived mononuclear cells for Illumina gene expression

Full gating strategy to identify human DC and monocyte subsets in the blood and their skin equivalents. All subsets were classed as 'live' by DAPI staining, CD45+, single-cells and HLA-DR+. Blood subsets were negative for lineage markers (CD3, CD19, CD20 and CD56 all in FITC). For the blood subsets, monocytes were split into classical and non-classical by CD16 and CD14, while cDC1 and cDC2 were split by CD14, CD11c, CD1c and CD141. For the skin samples, all subsets were negative by autoflorescence. Skin CD14+ DCs were CD14+, while skin cDC1 cells were CD14-, CD11c lo and CD141+. cDC2 skin cells were CD14-, CD11c+ and CD141-.

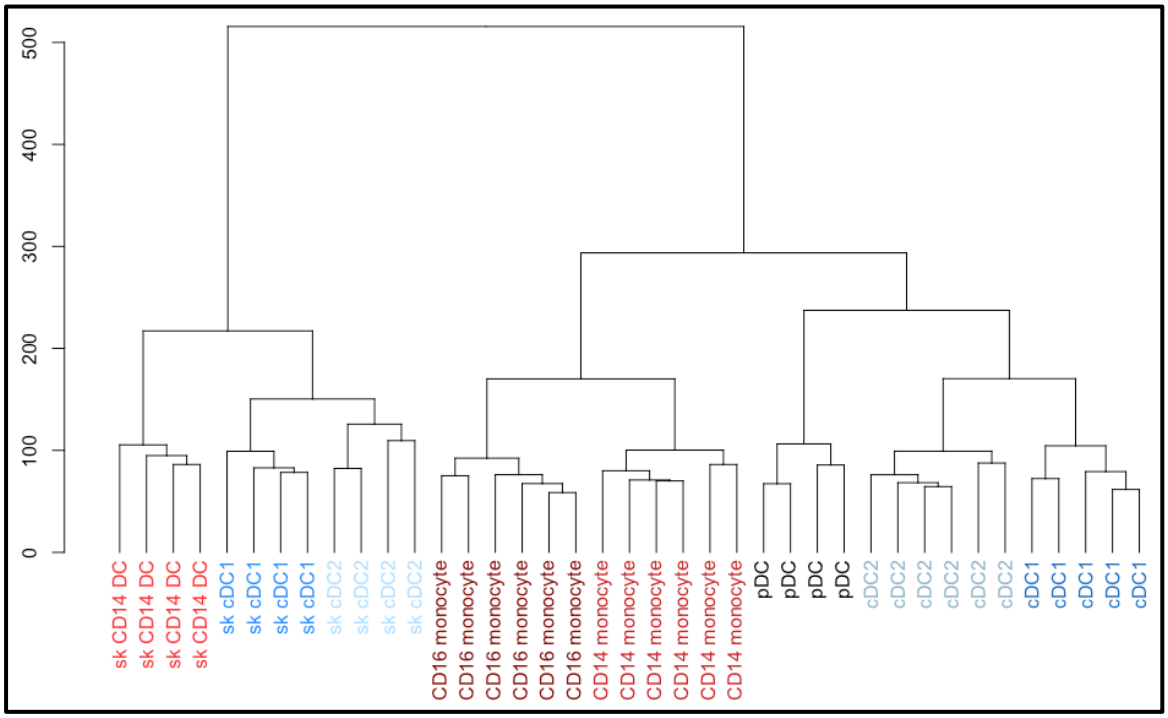


Figure 3.14: Hierarchical clustering of human blood mononuclear cells using Illumina expression data

Hierarchical clustering of the full Illumina array data for human blood and skin subsets was capable of distinguishing each population and group them into their developmental cell and tissue types. The Y-axis in this figure represents 'height' (a measure of increasing dissimilarity), increasing 'height' suggests clusters are less similar to one another. The first branch of the dendrogram splits blood-derived samples from skin-derived samples, possibly suggestive of a conserved tissue-specific signature. Further along the hierarchy on the skin-side, CD14 DCs branch from the cDCs first, followed by a split of the cDCs into cDC1s and cDC2s. On the blood-side of the dendrogram, monocytes are split into their conventional (CD14+) and non-conventional (CD16+) major monocyte subsets, while the blood dendritic cell branch splits into plasmacytoid DCs (pDC) and classical DC types. In turn, the classical DC branch is split into CD1c+ cDC2 samples and CD141+ cDC1 samples, mirroring the skin-equivalent section.

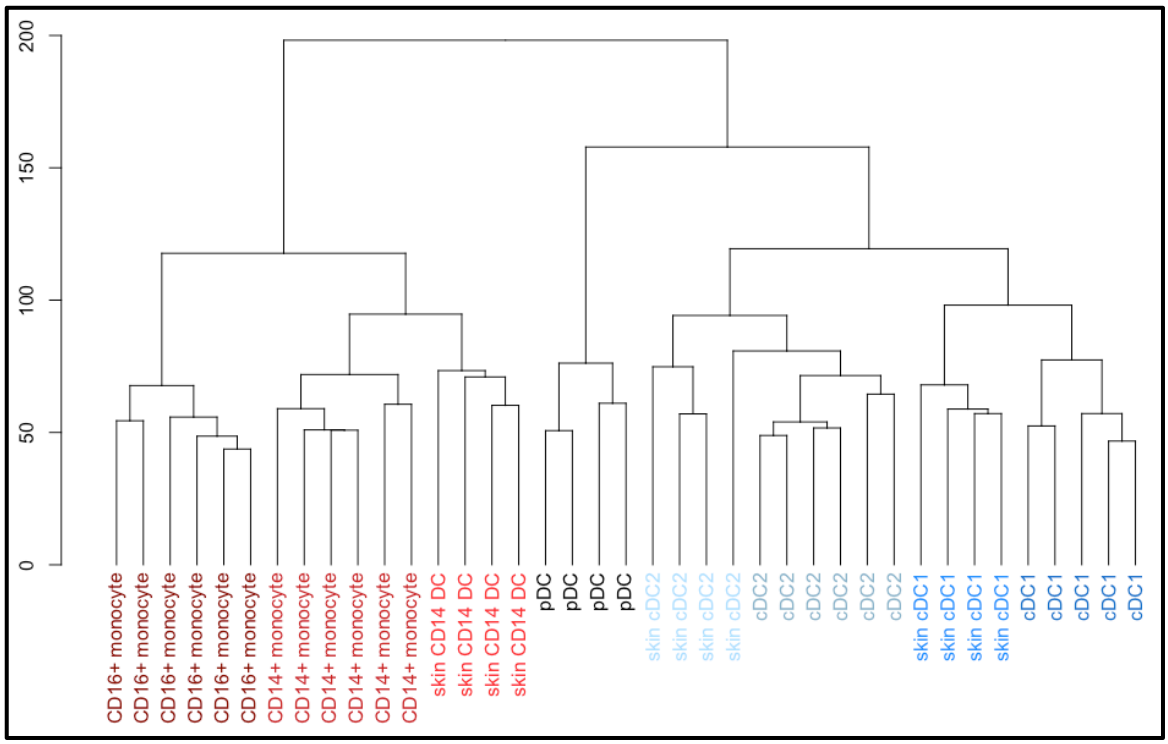


Figure 3.15: Hierarchical clustering of human blood and skin subset data Illumina BeadArray data with a conserved skin signature removed (P>0.05)

After performing a t-test to remove genes distinguishing the grouped skin samples from grouped blood samples, the hierarchical clustering diagram produced a tree splitting all of the five subsets regardless of their tissue of origin. The Y-axis in this figure represents ‘height’ (a measure of increasing dissimilarity), increasing ‘height’ suggests clusters are less similar to one another. The first branch of the dendrogram split the monocytes and CD14+ DC samples from the cDC and pDC samples. Along the monocyte and CD14+ cell branch, CD16+ non-classical monocytes were separated from the CD14+ monocytes and skin CD14 DCs. On the cDC and pDC branch, the pDCs branched from the skin and blood cDCs, which themselves further split into cDC1-like and cDC2-like groups. Each of the matched blood and skin subtypes would still be distinguished at the lowest branches of the dendrogram, but the major branching reflects cell types rather than tissue types.

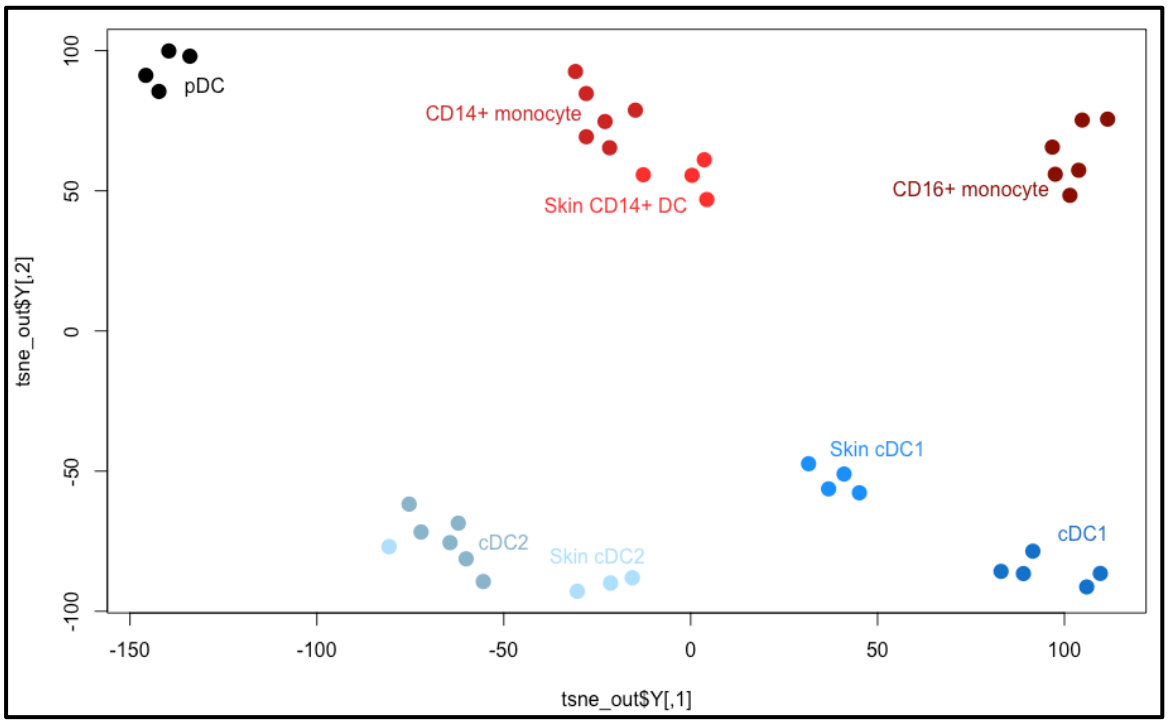


Figure 3.16: t-SNE analysis of human blood and skin subset data Illumina BeadArray data with a conserved skin signature removed (P>0.05)

After performing a t-test to remove genes distinguishing the grouped skin samples from grouped blood samples, this t-SNE diagram produced a plot splitting all of the five subsets regardless of their tissue of origin. T-SNE1 (x-axis) split pDCs at the extreme negative region from CD14+ cells at the zero-point and CD16+ monocytes at the positive region. cDC2 skin and blood cells were present at approximately t-SNE1 of -50, with skin cDC1s at 50 and blood cDC1 at 100. T-SNE2 (y-axis) split the blood and skin cDCs from the CD14+ cells, monocytes and pDCs.

CD14+ cells and cDC2 cells appeared to be more transcriptionally similar in the skin and blood than cDC1 cells. No skin-derived samples appeared near pDC samples or CD16+ monocyte samples, as they did not have skin equivalents.

2012 Illumina HT12 dataset	2015 Illumina HT12 dataset
6x Blood CD14+ Monocytes	4x Blood CD14+ Monocytes
4x Blood 141+ DCs (cDC1)	3x Blood 141+ DCs (cDC1)
6x Blood 1c+ DCs (cDC2)	8x Blood 1c+ DCs (cDC2)
4x Blood pDCs	6x Blood pDCs

Table 3.5: Sample list for combining Illumina expression data

This table lists the 41 samples from two different microarray experiments that were combined into one data frame for the purpose of generating robust cell signatures, using as many individual samples as possible, that could then be trimmed down using gene reduction techniques. The larger replicate numbers would also allow for the testing of machine learning algorithms. The two experiments were combined using the 'ComBat' function, part of the 'sva' package on R (Leek et al, 2007). ComBat adjusts for known batch effects using emperical Bayesian frameworks. For this analysis, pDCs were chosen as the subset to base the normalisation and validation on for data adjustment due to their strong signature that appears distinct from the other DC and monocyte subsets (Figures 3.3 - 3.11)

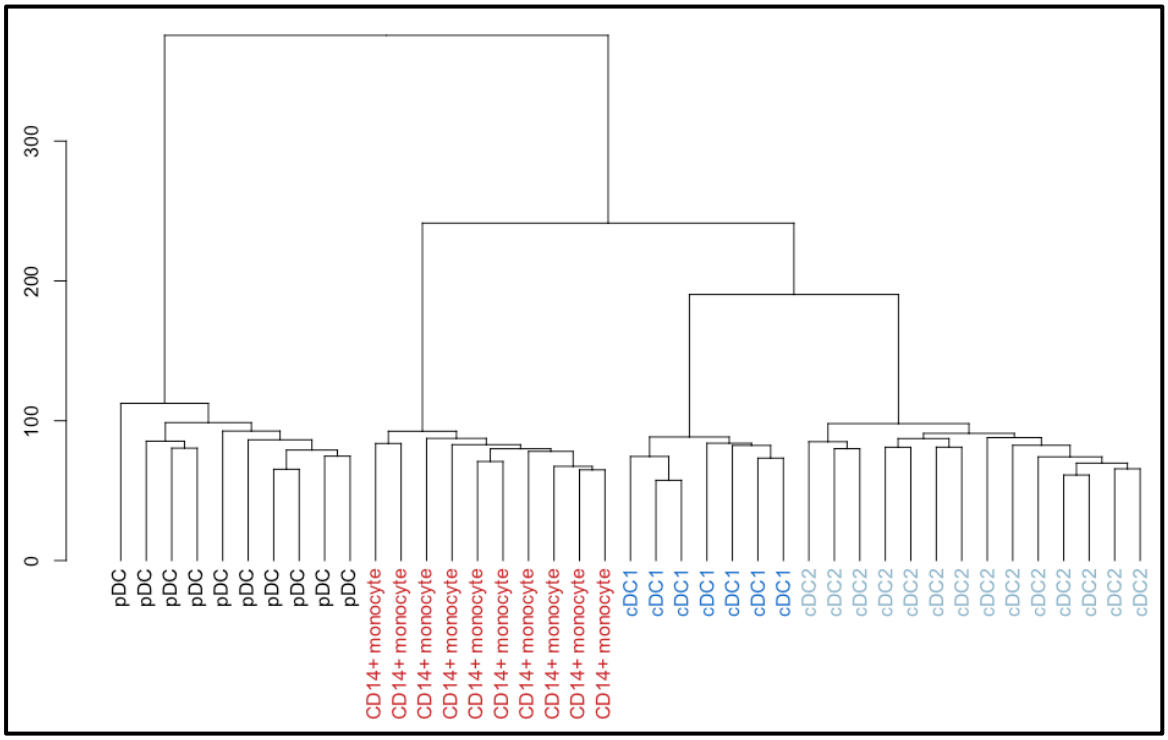


Figure 3.17: Hierarchical clustering of combined human blood mononuclear cell subsets from two independent Illumina microarray assays after normalisation using ComBat

This hierarchical clustering dendrogram uses the combined and normalised samples from two independent Illumina expression datasets (Figure 3.4). The Y-axis in this figure represents 'height' (a measure of increasing dissimilarity), increasing 'height' suggests clusters are less similar to one another. The diagram highlights the successful merging of the datasets as no batch effects appear to be present. pDCs are the first subset group to branch from the dendrogram typically showing a unique gene expression signature unlike the other cell types. Next to split off are the CD14+ monocytes, followed by a branching of the cDC subsets into cDC1 and cDC2 groups.

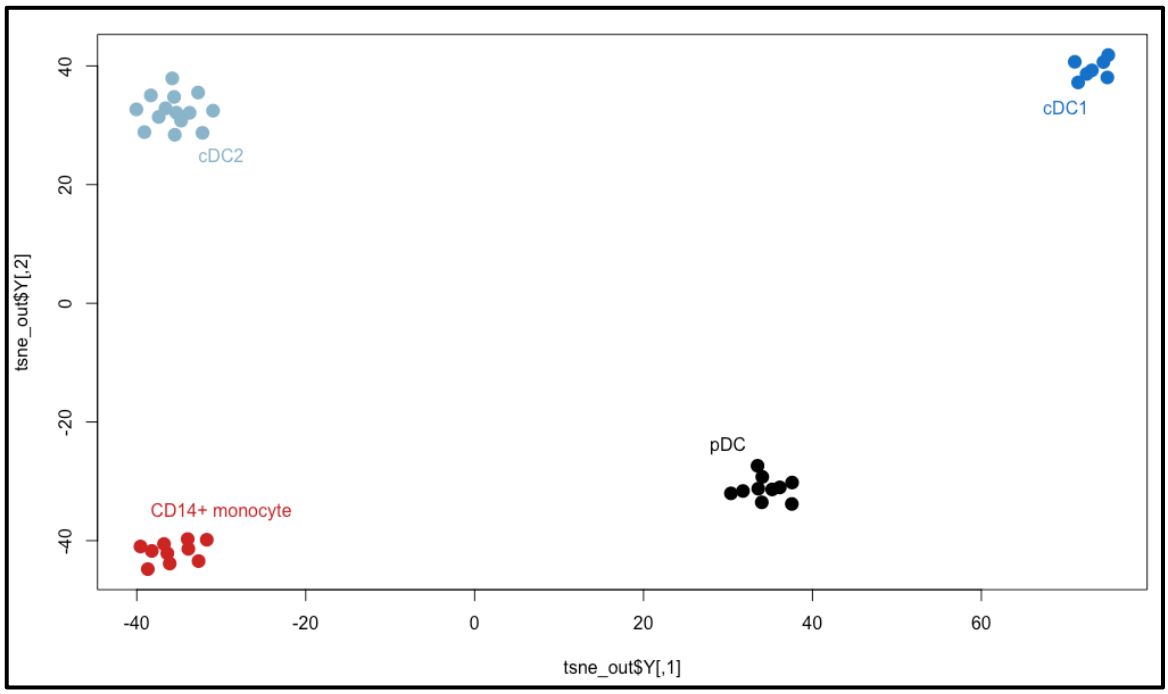


Figure 3.18: t-SNE analysis combined human blood mononuclear cell subsets from two independent Illumina microarray assays after normalisation using ComBat

t-SNE analysis of the combined Illumina expression data produced 4 distinct subset population clusers. Each subset occupies a single region of the plot. T-SNE variable 1 (x-axis) CD14+ monocytes and cDC2 cells from pDCs and cDC1 cells. t-SNE2 (y-axis) splits the cDC populations from the monocyte population and pDCs. Distance relationships are not proportionate in this type of analysis so although the subset appears at different regions in 2D t-SNE space, their ‘closeness’ to other subsets cannot be determined by this method.

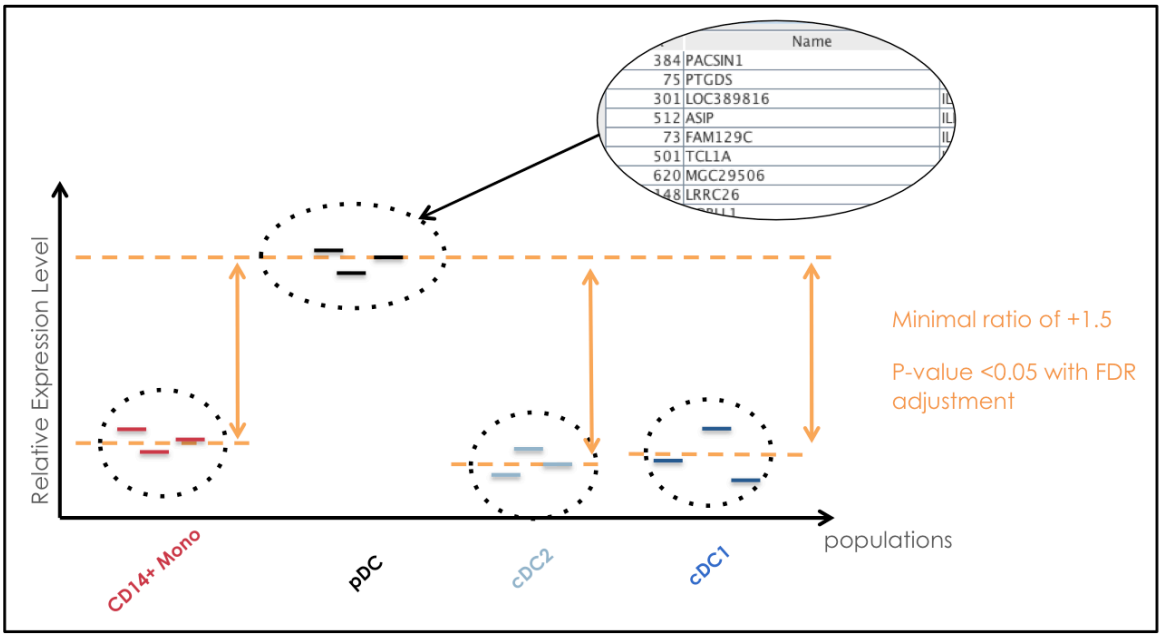


Figure 3.19: Schematic of GeneSign signature generation using combined Illumina expression datasets

GeneSign was used to generate cell type signatures using data from 41 blood mononuclear cell samples. To do this, relative expression levels were compared between each subset group for every Illumina probe in a pair-wise manner. Minimal pairwise testing with Benjamini-Hochberg false discovery rate adjustment was employed with cut-off values of P=<0.05 and a minimal ratio of 1.5 fold. In the representative example, genes that had 1.5 fold greater expression in pDCs and a p-value of <0.05 compared to all other individual subset groups was considered a pDC signature gene.

Mononuclear cell subset	CD14+ monocyte	cDC1	cDC2	pDC
Number of signature genes	964	483	567	1425
Top signature genes	SLC11A1 **S100A8** HK3 **CD14** LILRA3 CFD VCAN **C19orf59** S100A9 AQP9 SERPINA1 KLF2	**TMEM137** **IRAK2** LOC731682 ACVRL1 JARID1D **MAPK13** AGPS TANC2 MGC57346 PPY LOC643995 TSPAN33	**CD1C** **CD1E** **CLEC10A** IL1R2 MYO5C FBLN2 ENHO CACNA2D3 CD2 GPR44 RHOU **FCER1A**	**GZMB** **PACSIN1** PTGDS C20orf103 FAM129C LOC389816 **ASIP** MGC29506 LRRC26 EPHB1 SHD TCL1A

Table 3.6: Total gene signatures and top signature genes from GeneSign pair-wise analysis

GeneSign analysis produced 3,439 gene signatures across the four mononuclear cell subsets. All signature genes are 'positive' and therefore expressed most highly in their signature subset. The most signature genes (n=1,425) were attributed to pDCs, which are more distinct In phenotype and lineage than the other analysed subsets. 964 gene signatures were attributed to CD14+ monocytes, 483 to cDC1s and 567 to cDC2s. As pDCs appear the most distinct according to hierarchical clustering, PCA and t-SNE analysis this subset also has the highest number of signature genes. Similarly, as cDC1 and cDC2 are more closely related, they will have fewer signatory genes by comparison. Differences in underlying gating strategies for cDC1 and cDC2 subsets between the two Illumina array experiments may also be a cause for reduced signature gene numbers.

In bold text, major cell surface marker genes and genes encoding for well-established internal marker proteins are highlighted, including some of the genes involved in CD14 and CD1c expression.

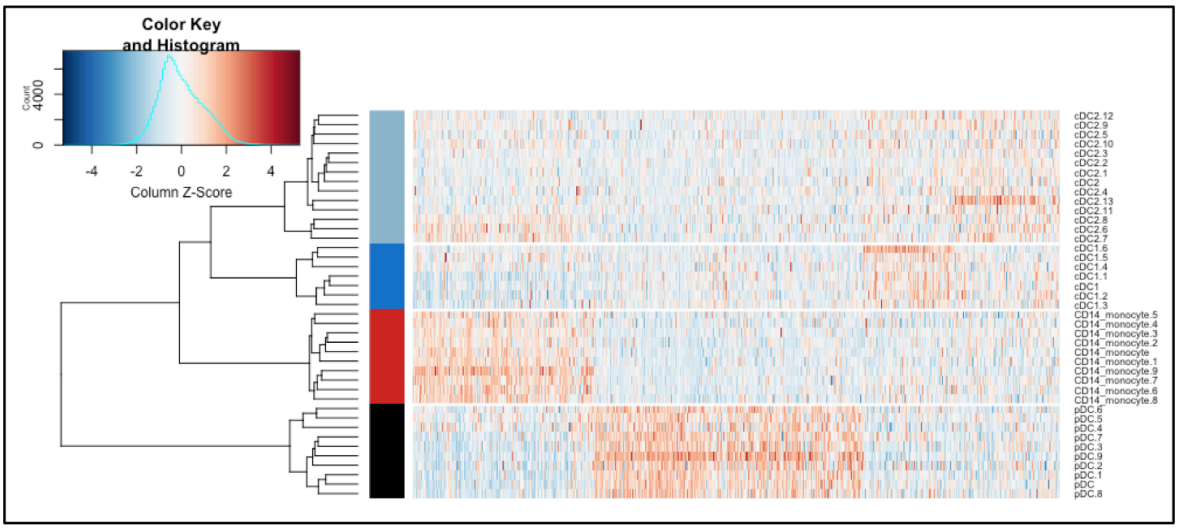
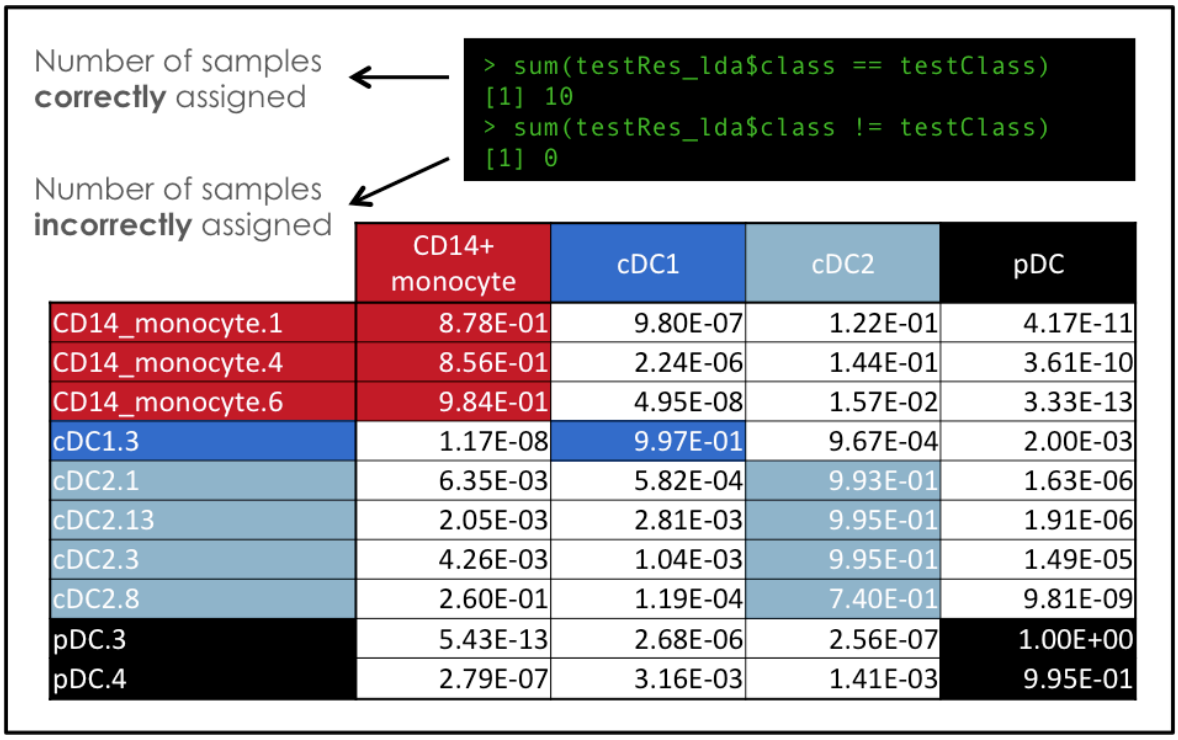


Figure 3.20: Heatmap of combined blood mononuclear subset data using GeneSign signatures

The 3,439 gene signature generated by GeneSign was plotted as a heatmap for all samples. Each row of the heatmap represents an individual sample, with the colour bars linked to the sample type. The colour bars and accompanying dendrogram show a clear grouping of the samples into their four subsets. cDC2 cells appear at the top of the heatmap, followed by cDC1 samples. These are the closest related by hierarchical clustering too. Monocytes feature below the cDC subsets with pDCs at the bottom of the diagram.

From the heatmap clear blocks of high gene expression levels can be seen, which correspond to each of the subset gene signatures.

Reducing the dataset from the full human transcriptome to 3,439 genes still provided enough distinction between the mononuclear cell subsets to correctly and robustly group them by subset.



	CD14+ monocyte	cDC1	cDC2	pDC
CD14_monocyte.1	8.78E-01	9.80E-07	1.22E-01	4.17E-11
CD14_monocyte.4	8.56E-01	2.24E-06	1.44E-01	3.61E-10
CD14_monocyte.6	9.84E-01	4.95E-08	1.57E-02	3.33E-13
cDC1.3	1.17E-08	9.97E-01	9.67E-04	2.00E-03
cDC2.1	6.35E-03	5.82E-04	9.93E-01	1.63E-06
cDC2.13	2.05E-03	2.81E-03	9.95E-01	1.91E-06
cDC2.3	4.26E-03	1.04E-03	9.95E-01	1.49E-05
cDC2.8	2.60E-01	1.19E-04	7.40E-01	9.81E-09
pDC.3	5.43E-13	2.68E-06	2.56E-07	1.00E+00
pDC.4	2.79E-07	3.16E-03	1.41E-03	9.95E-01

Figure 3.21: Linear Discriminant Analysis (LDA) machine learning output for sample subset classification based on GeneSign gene signatures

Linear discriminant analysis was performed on the 41 sample dataset to confirm if the GeneSign signatures were robust enough to define 'unknown' mononuclear cell subsets. For this analysis, The 41 sample dataset was split into randomly assigned groups. Linear discriminant analysis (LDA) was used on all but one of these groups to 'train' the machine. All sample subtype identifiers were removed from the final 'test' group. Based only on the expression profile, the machine would assign each test sample to a subset group, providing a confidence score for the assignment. Multiple iterations of this analysis was performed so that every sample was part of a training group and test group. Four iterations were performed with 100% total prediction accuracy. All unlabeled samples were correctly assigned to their subset of origin by the LDA predictor.

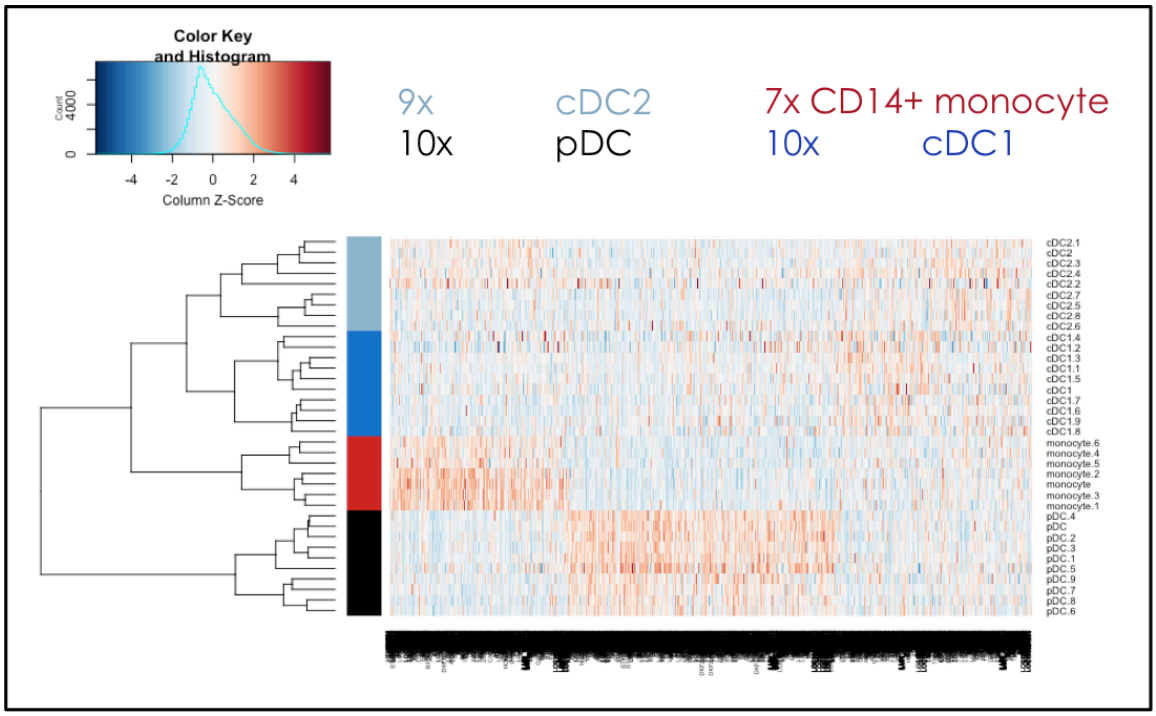


Figure 3.22: heatmap for sample subset classification based on GeneSign gene signatures using an external validation dataset GSE65128

Data from Lee *et al* 2015 paper 'Restricted dendritic cell and monocyte progenitor in human cord blood and bone marrow' (GSE65128) was used to validate the GeneSign signatures. By taking the 36 sample dataset and restricting the data to the 3,439 gene signatures, a clear correlation between this independent dataset and the one used for signature generation can be seen. As in figure 3.18, each subset is well defined, despite differences in the initial gating strategy for each dataset. cDC2 samples group at the top of the heatmap, with cDC1 cells below this in a separate group. From the cDC groups CD14+ monocytes branch, followed finally by the pDC group. The capability of the gene signatures to correctly group and define each sample into their subset groups is maintained even when applied to data that was not used in the initial signature generation experiment.

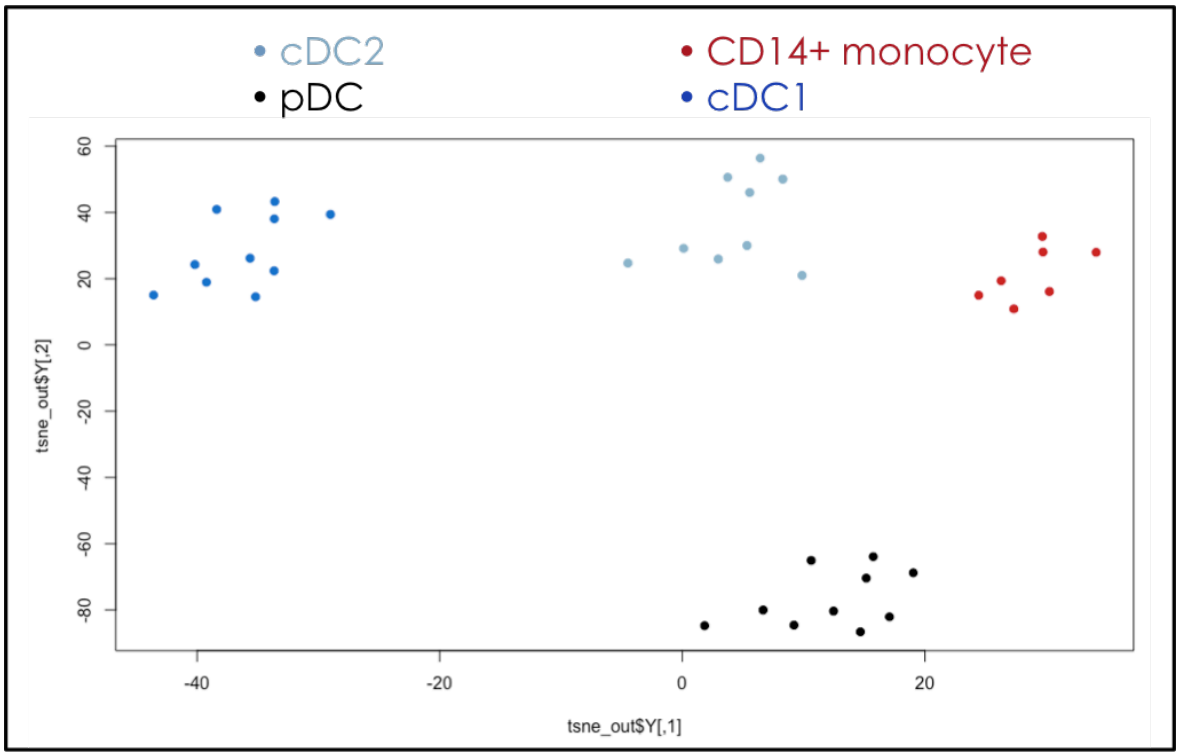


Figure 3.23: t-SNE output for sample subset classification based on GeneSign gene signatures using an external validation dataset GSE65128

Data from Lee *et al* 2015 paper 'Restricted dendritic cell and monocyte progenitor in human cord blood and bone marrow' (GSE65128) was used to validate the GeneSign signatures. By taking the 36 sample dataset and restricting the data to the 3,439 gene signatures, t-SNE analysis pulled the samples out into clusters based on their mononuclear cell type. cDC2 samples group at the top-centre of the t-SNE plot, with cDC1 cells forming a distinct group at the top-left of the plot. CD14+ monocytes branch are at the top-right of the t-SNE plot, with pDCs at the lower region. Each cluster is relatively tightly grouped and located apart from any other sample cluster, suggesting good separation and definition of each of the subsets.

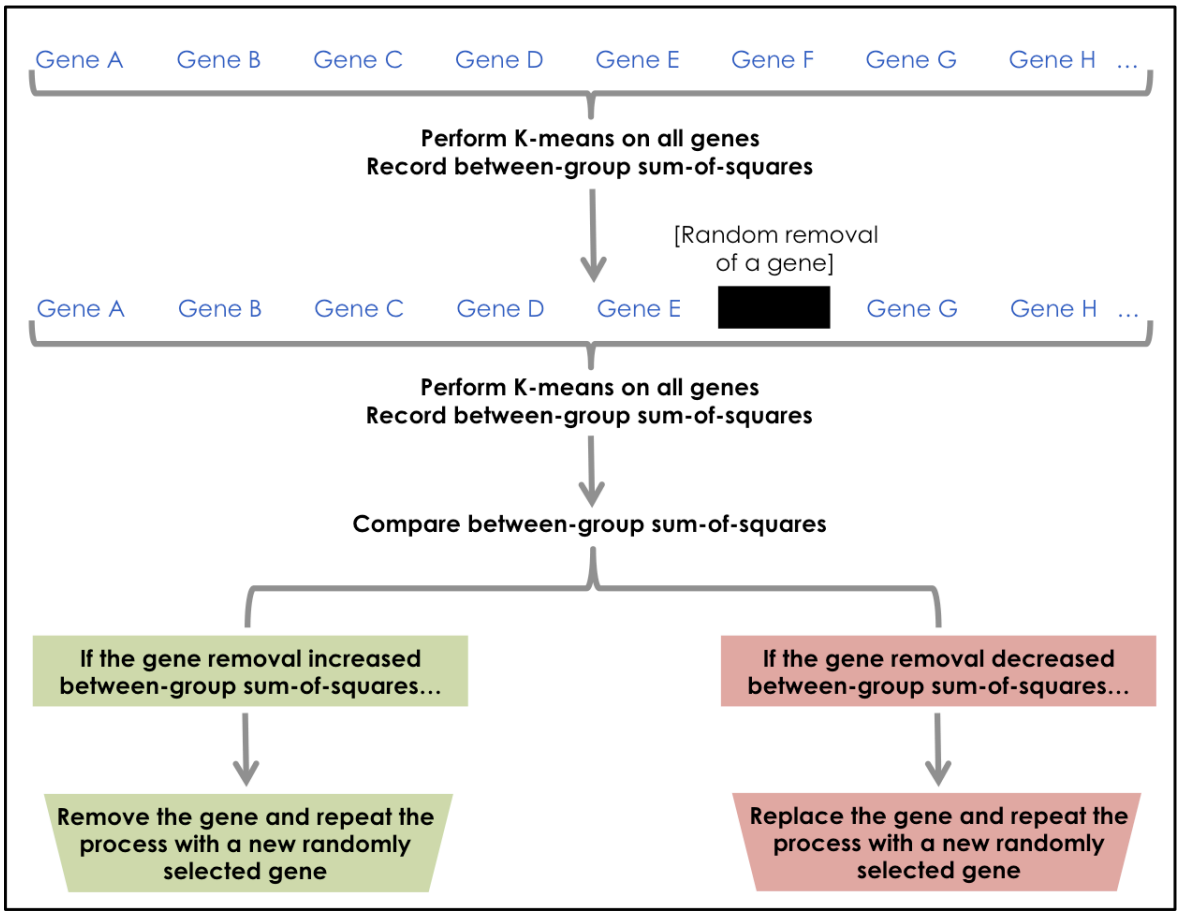


Figure 3.24: Gene reduction technique for minimal gene list subset classification

Signature genes were expressed highly in a single subset as seen in figure 3.18. As each signature gene is specific to a subset, multiple genes were unlikely to be necessary for accurate subset assignment. In order to test the minimum number of genes required to accurately group the four mononuclear cell subsets, a randomised gene reduction method was designed based on K-means testing to reduce the geneset to the minimum genes required for successful classification and grouping. K-means testing was performed on all genes and the between-group sum-of-squares were recorded for each group. One gene was randomly removed and the process repeated. If the new geneset provided grater between-cluster sum-of-squares, another random genes would be removed and re-checked. If the new geneset did not provide a greater sum-of-squares, the gene would be replaced and another gene randomly removed and re-analysed. This process was repeated until after 10,000 iterations, the final gene list was recorded and used to produce a heatmap and hierarchical clustering.

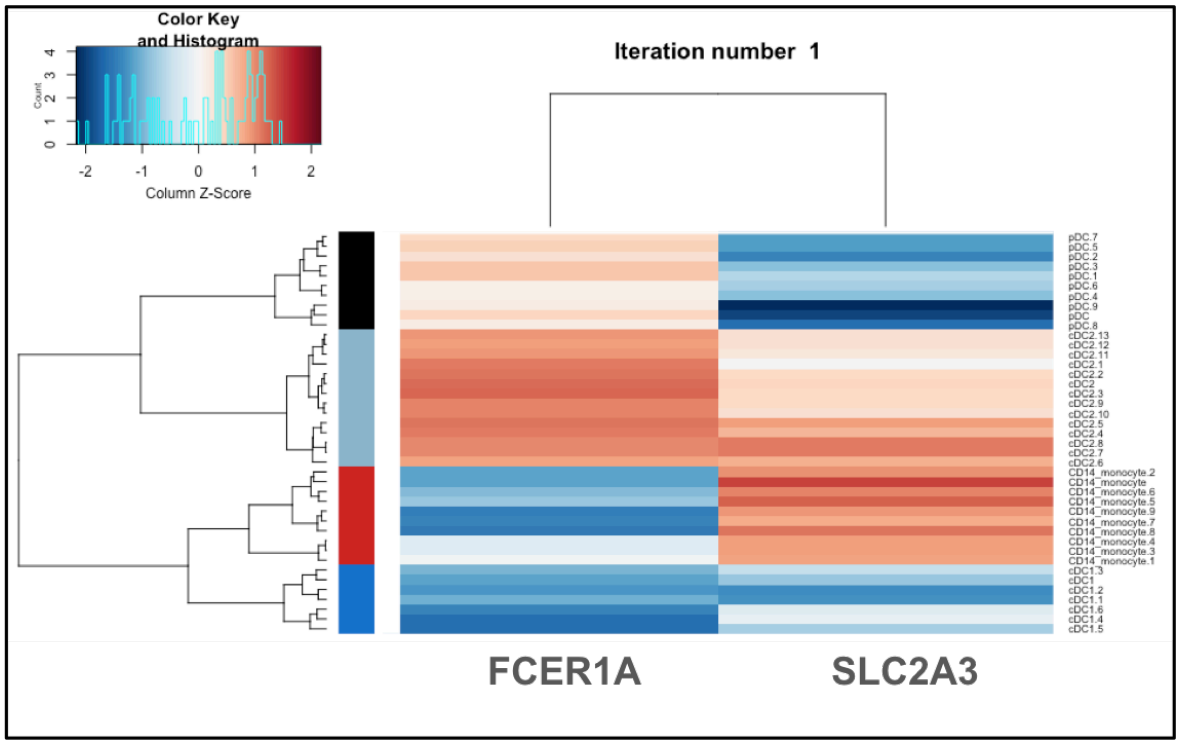


Figure 3.25: Example of surface marker gene expression after gene minimisation

From the 3,439 signature gene list, 612 genes were noted to have cell surface protein expression. This 612 gene list was used as the starting point for a new round of gene minimisation using the randomised K-means reduction technique created for this thesis. From this analysis, a number of iterations produced two-gene signatures capable of correctly grouping the subset sample data by mononuclear cell type. One example of this was the combination of FCER1A and SLC2A3. FCER1A was highly expressed on pDCs and cDC2 cells, but lowly expressed on CD14+ monocytes and cDC1 cells. SLC2A3 was expressed highly on cDC2 cells and CD14+ monocytes, but lowly expressed on both pDCs and cDC1 cells. This combination of marker genes could split these four subsets in a ++,+-,-+,-- expression pattern, potentially reducing the need for multiple cell type marker genes in flow cytometry experiments and freeing up channels for new flow antibodies to be used.

Chapter 4: IN-VITRO DENDRITIC CELL SUBSET CLASSIFICATION

Primary research question:
Can transcriptomic signatures aid in the identification and validation of cells generated *in-vitro*?

Sub-topic questions:

1. Can phenotypically equivalent human dendritic cells be generated *in-vitro*?

2. Are there any culture-specific expression patterns identifiable *in-vitro*?

3. Do *in-vitro* derived DCs share a similar transcriptome profile with primary human DCs?

4.1 INTRODUCTION

The multiple roles of dendritic cells in the innate and adaptive immune system, their ability to influence CD4+ and CD8+ T-cells and importance in antigen presenting, immunotherapy and autoimmune disease have led to significant scientific research interest in the production of high numbers of *bona-fide* DCs (Morse and Lyerly, 2002).

As described in previous chapters, dendritic cells and DC pre-cursors make up just 1% of peripheral blood mononuclear cells (Nairn, 2002). The myeloid component of this 1% is mostly comprised of cDC2 cells, with cDC1 cells approximately 10x less abundant. The scarcity of such cells makes research on primary DCs a challenging hurdle to overcome. Growing DCs, monocyte-derived DCs and monocytes in culture conditions provides researchers with the cell numbers required for RNA-seq, NanoString and other RNA, DNA or protein based techniques. Such *in-vivo* generated cells or *ex-vivo* monocyte-derived DCs can be generated in vast numbers after appropriate cytokine and growth factor stimulation and are phenotypically similar to primary DCs by flow cytometric analysis (Nair et al., 2012).

Generation of *in-vitro* cells provides the scalability required for effective functional and molecular studies examining response to stimulants, cytokine interactions and inflammatory processes. The development process can also be tailored through the use of transcription factors and cytokines to enrich the cellular output for specific cell subsets, offering potential avenues for cellular therapy.

Many publications rely on induced monocyte-derived DCs or cultured cells for analysis, typically drawing phenotypic or functional conclusions for primary cells based on these experiments. While similar by flow cytometry and sorting protocols, some surface marker expression changes are noted, including CD14+ expression on monocytes, or low MHC class II on CD11c+ DCs (Aldo et al., 2013; Mayuzumi et al., 2009). This may be due to the fickle nature of *ex-vivo* surface marker expression when applied for the identification of *in-vitro* populations in combination with the underlying differences between moDCs and *ex-vivo* DCs. MoDCs are closely related to inflammatory DCs and monocytes, which may impact the effectiveness of monocyte-derived DCs in dendritic cell-based therapies.

Because of these discrepancies, this chapter focuses on the relationship and correlation between primary DC and monocyte subsets isolated from blood and those generated under culture conditions from CD34+ bone marrow or peripheral blood progenitor cells using gene expression data generated on the NanoString platform. This platform provides over 500 immune-related gene targets, incorporating chemokines, cytokines, transcription factors and other functional molecules to generate a directed, but wide-scoping dataset for cell identification between *ex-vivo* and *in-vitro* generated cells.

4.1.1 Dendritic Cell and Monocyte Culture Research

In contrast to dendritic cells, monocytes are abundant white blood cells, easily identified and isolated by FACS and flow cytometry as Lin-HLA-DR+CD14+CD16+/- cells. *In-vitro*, such monocytes can be converted into monocyte-derived dendritic cells (moDCs) in the presence of GM-CSF and IL-4 (Autenrieth et al., 2015) providing adequate numbers for further research in cellular therapy, vaccination and immunotolerance.

While *in-vitro*-derived DCs can also be generated from bone marrow, cord blood and peripheral blood mononuclear cells (Zou and Tam, 2002) using a variety of stimulus, many studies make use of the standard GM-CSF and IL-4 protocol to derive monocyte-induced dendritic cells. The addition of IL-4 to GM-CSF containing monocyte cultures induces a much more DC-like phenotype over GM-CSF alone. This feature was identified over two decades ago, where under GM-CSF alone, resulting cells took on a macrophage-like phenotype, yet in combination with IL-4, CD14 expression was decreased and non-adherent DC-like cells developed, (Kiertscher and Roth, 1996). Despite some phenotypic differences, these generated DCs expressed the activation markers CD80, CD83 and CD86 Mature MoDCs exhibited a Th1 response, while CD34+ progenitor cell-derived DCs leant towards Th2 polarisation and IFNα and IFNβ production (Zou and Tam, 2002). Subsequent research has shown an earlier divergence between the DC and monocyte lineages that may infer that moDCs are phenotypically distant from bona-fide DCs outside of the standard cell activation and identification markers (Geissmann et al., 2008).

The work in this chapter explores the relationship between *ex-vivo* peripheral blood dendritic cell subsets and those derived *in-vitro* from CD34+ bone marrow cells under culture conditions developed by Dr Urszula Cytlak-Chaudhuri of the Human Dendritic Cell Lab, capable of generating DC subsets with phenotypic similarities to *ex-vivo* DCs, yet distinct from monocytes.

4.2 MATERIALS AND METHODS

4.2.1 Sample Collection and Isolation

All samples were taken from healthy volunteers after written consent under ethics contained in Chapter 2, section 2.1. PBMCs were isolated according to the protocol listed in Chapter 2, section 2.3.2. Bone marrow was obtained from patients undergoing hip replacement surgery and the cells isolated according to the protocol outlined in section 2.3.1.

4.2.2 Flow Cytometry and Sample Sorting

Bone marrow and peripheral blood mononuclear cells, or cells harvested from culture, were stained in aliquots of up to $6x10^6$cells/100µl. CD34+ bone marrow progenitors for culture and mononuclear cells from peripheral blood or generated in culture for gene expression analysis were enriched to >98% purity by FACS, using a FACSAria III (BD Biosciences) running BD FACSDIVA 8.0 software. DAPI was used for dead cell exclusion. CD34+ cells for culture were collected into culture medium (section 4.2.3) and cells for gene expression analysis were collected into RPMI before cell pelleting and resuspension in RNA lysis buffer (RLT buffer + 1% bME). Replicate DC/monocyte FACS purification experiments were performed by various members of the human DC lab.

Cells were stained with DAPI for dead cell exclusion, a lineage cocktail (CD3, CD19, CD20, CD56), CD45, HLA-DR, CD14, CD11c, CD141, CLEC9A, CD1c, CD123 and CD303/4.

4.2.3 Culture Conditions

FACS purified CD34+ bone marrow cells were seeded at 3,000 cells per well in 96 well U-bottomed plates containing pre-seeded 5,000 OP9 stromal cells per well. These were cultured for 21 days in 200mL 1% penicillin/streptomycin supplemented αMEM (Gibco™) with 10% Fetal Calf Serum (Gibco), 20ng/ml GM-CSF (R&D systems), 100ng/ml Flt3-ligand (immunotools) and 20ng/ml SCF (Immunotools). 50% of the media volume was removed and replaced weekly along with the cytokines. After 21 days, the culture products were harvested on ice and filtered through a 50mm filter, after which they were washed and stained for flow cytometry and FACS. Cell culture and subsequent FACS purification was performed by Urszula Cytlak-Chaudhuri. FACS gating on surface parameter expression was consistent between blood and cultured cells.

4.2.4 NanoString Panel+ Codeset

NanoString data was generated on the Immunology_V2 panel with Panel+ genes added as described in section 3.3.2.1. This panel contained immunologically relevant gene targets encompassing major classes of cytokines and their receptors, as well as chemokines and receptors, interferons and TNF-receptor superfamily targets. This panel was specifically designed to address allergy, immune response, autoimmunity and infectious disease response. Allowing sample input between 10-100ng of FFPE-derived RNA, total RNA or cell lysates provided an opportunity to study low frequency cell types including CD141+ cDC1s without the need for amplification.

The NanoString nCounter cell lysate protocol was selected following the preliminary experiments described in Chapter 3, section 3.3.3. This ensured 10,000 cells of each population, sorted by FACS, would be used in each NanoString assay for comparability without any disruption in expected gene counts compared to extracted RNA.

4.2.5 Significance Testing and Analysis

Significance testing for culture effect removal was performed by two-tailed t-test with p-values of ≤0.05 after Benjamini-Hochberg FDR adjustment deemed significant. Dendrograms were drawn based on Euclidean distance and Ward method agglomeration. PCA was performed using the 'stats' package of 'R' and visualised using 'ggplot2' and 'ggbiplot'.

t-SNE was performed using 'Rtsne' package with a fixed set seed for reproducibility. A perplexity of 3 was set for visualisation purposes.

GSEA was performed with FDR-adjusted p-value of ≤0.05 deemed significant after hypergeometric testing for functional Gene Ontology enrichment.

4.3 RESULTS

4.3.1 Surface phenotype of cells generated in culture

The four mononuclear cell subsets used in Chapter 3 were studied for this chapter, consisting of CD14+ monocytes, CD141+ cDC1s, CD1c+ cDC2s and CD123+ pDCs taken from healthy peripheral blood. These were compared to phenotypically similar cells (Figure 4.1) generated in culture for 21 days from CD34+ bone marrow. A minimum of two positive surface markers of each subset were used to identify cells by FACS.

The proportions of cells falling into each sort gate varied between the blood and their cultured counterparts when equivalent gates were applied.

From the CD14+ monocyte gate, 91% of blood cells fell into the CD14+ CD11c+ category, while the cultured output had 38% of the cells in the CD14+ category. For the cDC1s, blood cDC1s (Clec9A+ and CD141+) comprised of 1% of the cells from the CD14- gate, while the cultured equivalent gate contained just 0.1% of cells. The percentage of cells found in the pDC gate were also lower in the cultured subsets, down from 67% with CD123+, CD303/304+ expression, to 3.4%, with the majority of the cells falling lower in CD123 expression than the blood equivalents. Interestingly, cDC2 cells in the blood and culture were roughly comparable in terms of the percentage of cells falling into the CD1c+ CD11c+ gate, although increased CD1c expression appears commonplace under *in-vitro* conditions.

Of note, cells appearing in the cDC2 population plot from Figure 4.1 had a differing characteristic in CD1c and CD11c expression. Blood cells were generally low in CD1c and medium by CD11c, in contrast, cultured cells in the equivalent plot ranged from low to very high expression of CD1c and variable CD11c expression reflecting both the fickle nature of commonly used cell markers when applied to *in-vitro* generated cell populations, as well as the fact that cells captured in culture represent a developmental continuum ranging from the CD1c+CD11c- cells, believed to be in the early stages of development and the CD1c+CD11c+ later stage cells. Further technical and biological reasons for these observations are discussed in section 4.4.

4.3.2 Generation of gene expression dataset for peripheral blood and cultured cells

Three replicate samples of each selected mononuclear cell subset from blood and culture were analysed using the NanoString nCounter platform utilising the human Immunology_V2 panel and custom Panel+ add-on. As correlation between the samples was to be assessed and global differential expression was likely needed to address any culture-specific effects, an equal number of replicates were produced for each subset so that none were over-represented, which would ultimately skew the differential expression analysis downstream.

4.3.3 Comparison of primary and in vitro derived DCs and Monocytes

Figure 4.2 shows the dendrogram clustering of the dataset using Ward method agglomeration and Euclidean distance. There was no cross over between subsets at the level of the first two nodes; all eight cell subsets were distinct enough to form individual groups composed of their three replicate samples. The upper nodes of the dendrogram split the monocyte subsets from the DCs, followed by a split of the pDC cells from the bulk cDCs. The cell subsets grouped according to their phenotypic identity with the exception of the cDCs (cDC1 and cDC2), which clustered according to their origin.

By principal component analysis (Figure 4.3), distinction between cell types was more apparent. Again, each of the eight cell subsets were separable, but by PCA there was no cross-over between sample types as observed by hierarchical clustering. PC1 explained 24.4% variance and split cDC1, cDC2 and monocyte subsets, both from blood and from culture. PC2 split pDC subsets from the others and explained 19% of variance. Grouping by cell type was not obvious, however matched cultured and blood samples appeared to exhibit a similar expression pattern by PCA. Each *in-vitro* derived subset was located at a euclidean distance of -10 to -15 less thans its blood counterpart along PC1 and PC2. This was true for all four cell types and gave rise to the idea of a conserved 'culture signature' that was shared by all *inivitro* derived cells. This effect mirrored earlier observations regarding the 'tissue effect', initially identified in chapter 3, whereby heterogeous cells derived from tissue showed conserved expression of a unique tissue-specific gene set compared to peripheral blood counterparts.

As the PCA process applies a weighting to genes based on expression variance, it was anticipated that large changes in expression of certain conserved genes must occur as a result of differences in the blood and culture environment.

t-SNE analysis of the blood and cultured samples was performed in Figure 4.4, which recapitulated the observations noted for the hierarchical clustering in Figure 4.2. pDCs and their culture counterparts grouped closely together and occupied a distinct region of the t-SNE plot, separated from the other cell types along t-SNE1. The two monocyte groups clustered closely together at the zero-mark of t-SNE1 and t-SNE2, with the blood cDCs grouped at the positive region of t-SNE2 and the cultured cDCs in the negative region of t-SNE2.

4.3.4 Removal of Culture Effect

To determine if there was a conserved 'culture effect' on gene expression between the samples, a similar gene removal approach to that used in section 3.3.4 for skin and blood cells was employed. As displayed in Figure 4.1, phenotypically similar cells could be identified in blood and culture, however their transcriptome profile by NanoString analysis was altered, largely based on whether the sample was from primary blood or culture conditions.

The two-tailed t-test based method of feature reduction was used to remove genes differentially expressed between the culture-derived samples and the blood-derived samples, with p=<0.05 acting as a cut-off for significance.

230 genes were deemed significant. The result of their influence on sample groupings is displayed in Figure 4.5 in the form of a dendrogram. This demonstrates division of the samples into blood-derived or culture-derived groups. The *in-vivo* branch exhibited sample groupings discordant with previous results, with monocytes and cDC2 cells branching from pDCs and cDC1s. Along the *in-vitro* branch, monocytes split initially from the DCs, with pDCs then branching off from the cDC subsets, as observed previously in Figure 3.13. Importantly, the two-tailed t-test method of 'culture signature' generation proved capable of separating the two sample groups effectively.

4.3.5 Comparability and Functional Changes

Figure 4.6 shows PCA of the resulting 'culture effect' gene list as applied to the dataset. Again, strong dissimilarity was observed between the blood and cultured samples, as expected, with PC1 explained variance at 41.3%. One additional point of note is the tightness of the sample clusters. The PCA reveals greater variation in expression of 'culture signature' genes within blood-derived cells compared to *in-vitro* derived cells, consistent with the uniform *in-vitro* conditions.

To further investigate the transcriptomic differences associated with culture conditions, functional gene-set enrichment analysis was performed on the differentially expressed genes. Using a hypergeometric test for functional Gene Ontology enrichment with FDR-adjusted p-value of ≤0.05 deemed significant, the functional differences associated with blood or culture subsets was apparent., Gene expression in primary cells was enriched for immune-related genes, antigen response genes and genes involved in mixed cell type interactions. Functional enrichment in the cultured subsets included genes involved in cellular stress, proliferation, apoptosis and response to stimulants. The full list of functions and p-values are displayed in Figure 4.7. Ki67 was noted as one of the most differentially expressed genes between blood and cultured samples, reflecting the forced replicative conditions in culture.

The culture signature was enriched for chemokine ligands, with 10/17 chemokines present on the NanoString panel found to be differentially expressed between pooled blood and pooled culture subsets.

Although 24/50 CD antigens from the NanoString Immunology_V2 panel were differentially expressed including CD1a, CD40 and CD86 encompassing DC differentiation, activation and cellular response pathways, none of the CD surface markers used for FAC-sorting the cell subsets were differentially expressed at the transcriptomic level, suggesting correlation between protein expression at the surface and intercellular mRNA expression for these marker proteins.

Interleukins and their receptors probed by NanoString were differentially expressed as well as both ITGB1 and ITGB2. CD11a (ITGAL) expressed on all DC and monocytes and CD103 (ITGAE), which defines mouse DCs and is also present on human DCs were differentially expressed.

4.3.6 Grouping of Samples After Removal of the 'Culture Effect' Genes

Comparison of samples following the removal of the 230 genes comprising the 'culture signature' was undertaken (figures 4.8 to 4.10).

Hierarchical clustering of the remaining dataset grouped cells as identified by their phenotype, irrespective of origin (Figure 4.8). The first branch of the dendrogram split the two monocyte populations from the dendritic cell subsets, as observed in chapter 3, representing a greater developmental 'distance' between monocytes and the DC subsets. This was followed by a split of the pDC subsets from the cDC subsets and finally the cDC subsets split into cDC1 and cDC2 populations. Compared to Figure 4.2, where convoluting genes were still present in the dataset, the 'Culture Signature' adjusted dataset grouped the cultured cDC2 and primary cDC2 samples together, distinct from the blood and cultured cDC1s.

By PCA in Figure 4.9, samples were associated by their cell type, comparable to the relationships identified by illumina microarray gene expression analysis (figure 3.5). PC1 accounted for 27% variance, with PC2 accounting for 20%, representing extensive deconvolution of the geneset upon removal of the culture signature. pDCs separated out by PC2, forming close groups of cultured and primary cells. The positive region of PC2 contained the cDC populations, with cDC1 and cDC2 populations distinct by PC1. In the far-positive region of PC1 both monocyte populations were found. The monocyte populations overlapped, suggesting an almost identical transcriptome profile when displayed by PCA. In all cases, there was greater variance within the blood-derived cell groupings as observed previously in figure 4.3.

Figure 4.10 showed the t-SNE output of the data based on the first 50 principal components. The inclusion of these additional components had no effect on the overall distance mapping of the samples, indicating that PC1 and PC2 displayed in figure 4.9 were weighted towards variables contributing most genes defining cellular identity. The diagram displayed the higher order relationships of the samples, highlighting the underlying similarities of the blood and culture-derived subsets with the removal of a global ‘culture signature’ geneset.

t-SNE1 separated out the DC subsets, with both pDC populations at the negative region of t-SNE1, the cDC1 populations around the zero-point and the cDC2 populations in the positive region. T-SNE2 separated the monocyte populations from the DCs, but also split the cDC2 populations (found at +20 on t-SNE2) from the cDC1 and pDC populations (at -20 to -60 by t-SNE2). No populations grouped by their *in-vitro* or *in-vivo* status, suggesting that the culture-specific gene removal procedure was effective.

4.3.7 Conservation of DC signature genes in cultured cells

Upon removal of the culture-specific gene signature, the remaining genes were investigated to determine if key factors and critical developmental genes were present and conserved between the blood and cultures subsets. As described in Chapter 1, section 1.2, Toll-like receptor (TLR) molecules are integral to dendritic cell biology, reflecting the functional roles of each DC subset. Expression of TLRs and other functional molecules present in the NanoString panel are displayed in Figure 4.11. Expression of TLRs between the blood and culture equivalents are conserved with pDCs expressing TLR1, TLR7 and TLR9; cDC1s expressing TLR1, TLR3 and TLR8 and cDC2s expressing TLR1, TLR2, TLR8 and higher expression of TLR5 than cDC1 or pDC subsets. Both blood and cultured cDC1s express IRF8, CLEC9A and XCR1. Both cDC2 populations expressed CLEC10A, IRF4 and CX3CR1 and both pDC subsets expressed TCF4. Both the blood and their cultured equivalent populations expressed comparable HLA-DR at protein and mRNA-levels along with 10 other HLA molecules.

4.4 DISCUSSION

Bona fide dendritic cells and monocytes generated under culture conditions may provide new avenues for immune-based therapy. Furthermore, generating significant numbers of cells usually found infrequently in peripheral blood opens up greater research potential for microarray and RNA-sequencing, which typically require over 300ng of RNA material. To be effective as surrogates or viable treatment options, such *in-vitro* generated cells must be both functionally and transcriptionally comparable to their *in-vivo* counterparts.

In this section, the transcriptional effect of culture conditions on developing dendritic cells and their independence from monocytes was assessed to determine what differences, if any, would be present.

Phenotypically similar cells identified in culture and blood using FACS were processed using the NanoString nCounter platform with the resulting data highlighting the extent of this similarity beyond the surface markers used for cell sorting. Clustering of the data displayed cellular association, particularly between cDC subsets and revealed a 'universal' culture-related gene set present in all *in-vitro* derived cells, the removal of which allowed de-convolution of the primary and cultured cell subsets into comparable DC and monocyte subsets. Individual gene-level analysis revealed the cultured cDCs had an expression profile matching *ex-vivo* cDCs, rather than a monocyte-like profile typically observed in monocyte-derived dendritic cells.

4.4.1 DCs and Monocytes Exhibit Some Altered Gene Expression in Culture

The production of the cultured cells from bone-marrow derived CD34+ progenitors under the influence of GM-CSF, FLT-3 ligand and SCF was intended to generate cells analogous to primary blood DCs, independent of monocyte origin. Phenotypic and transcriptomic analysis was performed to determine if the resulting mature cell subsets were comparable to their blood-derived mature counterparts. DCs and monocytes phenotypically similar to primary cells, contained within equivalent FACS gates, were identified in culture. These cells were purified by FACs for further transcriptomic analysis using the NanoString platform.

Despite equivalent gating strategies, the pattern of cell expression was altered at the FACS level. The composition and counts of cells falling into each subset gate differs drastically in some subsets and appears globally altered in others.

CD14+ monocytes from the blood formed a strong cluster of cells with CD11c and CD14 expression, yet the cultured CD14+ monocytes comprised of a loose population of CD11c med/CD11c+ cells with greater variability in CD14 expression, likely the result of differing exposure to cytokines under culture conditions.

CD1c expression in CD14-, CD141-/CLEC9A- cells appeared to indicate a different developmental process in culture than in the blood. While blood cells in the cDC2 population plot produced two major populations; double-negative cells, or CD1c+, CD11c+ cDC2 cells, the cultured equivalents exhibited a greater variability in CD1c and CD11c expression, producing a smear from double-negative cells to CD11c-, CD1c+ cells and through to the CD11c+, CD1c+ cDC2 population. The cultured cDC2 population gate contained cells with a much greater expression of CD1c than their blood counterparts along with variable expression of CD11c, a common feature in cultured cells and reminiscent of primary cells in the skin.

Considering their origins however, the end-gate populations used for NanoString analysis were largely comparable in their surface marker gene expression, providing a positive baseline for comparability by NanoString analysis.

Global expression differences between blood and cultured cells appeared to differ between individual subsets, illustrated most clearly by hierarchical clustering and tSNE analysis. Using the NanoString immunology V2 panel, the unique identity of monocytes and pDC was distinguishable, regardless of origin. This effect may be a technical consequence of the restricted panel of genes used on the NanoString geneset, curated to cover major immune pathways and functions, although the geneset was designed for equal coverage of immune processes indicating the observation is a biological result attributable to the exclusive gene expression profile and functional specialisation of the monocyte and plasmacytoid lineages.

Classical monocytes, being CD14+, equipped with chemokine receptors and largely blood-borne are highly distinguishable in the steady-state from dendritic cells. Their development processes are understood to split them early in haematopoiesis from DCs at the GMP stage, after which their transcriptome and associated functions become divergent from CDP-derived pDC and cDC populations (Collin and Bigley, 2016), this feature is reflected in the NanoString data with monocytes separating from the other cell types throughout the analysis.

pDCs, as observed in chapter 3, expressed highly a number of genes including PACSIN1, ASIP, PTGSD and GZMB relating to their specific immune functions. Lacking typical myeloid antigens but retaining some lymphoid features, their distinction from the conventional cDCs is apparent by NanoString analysis (Collin et al., 2013).

4.4.2 Reviewing the Culture Effect Genes and Functional Disparity

Removal from the analysis of a gene-set consisting of 230 differentially expressed genes between pooled blood and culture cells, the ‘culture signature’, revealed equivalence between phenotypically similar primary and in vitro derived cells at the transcriptomic level.

Analysis of the culture signature composition by PCA showed the greatest variance was accounted for by cell origin. However, significant variance remained to distinguish subsets, particularly among *in-vitro* derived cells.

Here, PC1 accounted for sample origin, while PC2 appeared to distinguish the cultured, and to a lesser extent, the blood-derived cell subsets.

This observation suggests that genes capable of distinguishing the cell subsets were likely also removed during the process. The geneset could have been further tailored from the ‘culture signature’ plots to return any genes associated with variance between cell subsets, however this would have undermined the unsupervised, and statistically robust two-tailed t-test method through direct manipulation of the geneset. As the remaining genes had ample capacity for cell subset distinction, additional supervised curating of the ‘culture-signature’ was unnecessary. The variance that appeared to distinguish cell types could also have been driving the distinction between the culture and blood separation, for example in the case of KI67, high expression in the cultured cells compared to blood ensured it’s removal during the two-tailed t-test, but direct inspection of KI67 expression also revealed subset-level differences in expression amongst the cultured subsets, reflecting differential cellular turnover rates across the cell subsets.

Interestingly, the cultured clusters displayed on the PCA were tight and easily distinguishable, reflecting the rigorously controlled culture conditions producing cell populations with very homogeneous expression profiles. Conversely, the *ex-vivo* samples, with the exception of the cDC1 population displayed more disparate associations. This variability across the blood subsets may be explained as donor variability of both genetic and environmental origin. As human donors are 'outbred', they exhibit high genetic variance, much greater than that typically found in mouse studies or *in-vitro* assays. The genetic impact on gene expression can be extensive, even amongst healthy populations of similar ages and gender, which is why GWAS studies typically require thousands of donors to reach statistical significance. Environmental factors may have also influenced the global gene expression variance amongst *ex-vivo* samples as history, age, diet, ethnicity, gender and exercise were not controlled for, but can all influence the immune system and expression of target genes. This would not have been a factor in the cultured populations which were subjected to the same proliferation and stimulating signals at all stages of their development, producing populations of equally developed mature cells. Similarly, peripheral blood populations will have had varied interactions with other cells in the blood and tissues, influenced other cell types and been transcriptionally altered by these interactions resulting in more heterogeneous populations with regards to gene expression linked to cell-to-cell interactions, exposure to infections and allergens as well as cell cycle functions, while cultured subsets were enriched for a single cell type and were not exposed to intra-cellular interactions.

Functional analysis and gene enrichment highlighted genes defining the blood subsets shared functions in immunity, interactions with other cell types and exposure to foreign antigens which are expected to be a result of their role and interactions in circulating peripheral blood with other populations, tissues and foreign antigens. Those genes upregulated in the cultured samples shared functions related to proliferation, apoptosis, cellular stress and response to stimulants, reflecting the forced cellular expansion and division under sterile and controlled culture conditions. It is therefore likely that the most significant functions enriched for in each population may also account for the heterogeneity identified by PCA analysis of the culture signature.

Subsequent interrogation of gene groups found to be differentially expressed exposed some expected immune-related components and some unexpected probes that may warrant further investigation in future experiments.

One of the most significantly differentially expressed genes identified was Ki67, a typical proliferation marker gene. This was more highly expressed in cultured cells (linear 5-fold difference) and underpins the proliferative capacity of cultured populations. Similarly, epidermal growth receptor genes were widely up-regulated under the forced proliferation of culture.

The large impact on chemokine expression in culture may be explained by the difference between *in-vivo* and *in-vitro* cytokine levels. Cytokine stimulation in culture results in highly concentrated, continuous exposure to certain cytokines throughout the culture process, while completely excluding other cytokines typically found in circulating peripheral blood. Of the cytokines differentially expressed, 8/10 were up-regulated in culture conditions, further reinforcing this conclusion.

Further to this, as reflected in the initial FACS gate proportions, the cellular composition of the culture wells were not equal to those of the peripheral blood. This would mean exposure of cultured cells to other typical blood cell types would have been altered, particularly T-cells and NK cells that are frequently induced by antigen-presenting cells *in-vivo*. Interleukin and interleukin-receptor expression was widely up-regulated in blood compared to culture along with LILRA1, LILRA5 and LILRB2. These genes are generally associated with induction and interaction with T-cells and thus very low expression of interleukins in cultured cells may be explainable by the lack of T-cell interactions in these cells.

Although they encompass multiple functional groups and immune-mediated pathways, differences in CD antigens were an important consideration to the research group as these are typically the targets of fluorescent probes for flow cytometry and FACS experiments. Changes here would result in incorrectly gated cells and may lead to misinterpretation of results if these surface markers are used to define a population. CD34 featured in the list of differentially expressed genes. DC activation markers including CD1a, CD40, CD80 and CD86 were up-regulated in cultured cells. Again, this is likely to be developmentally induced activation for the purpose of producing matured cultured cells. CD1a expression is typically linked to dermal DCs, but in the presence of an adherent-cell feeder layer, it is possible that CD1a expression in the cultured blood DCs was influenced by the 'dermal-like' presence of this feeder cell layer, although increased CD1c expression is common amongst many *in-vitro* derived cells, reflecting a potential inflammatory DC-like potential.

While differences in gene expression were evident between the cultured and blood-derived cell samples, the genes driving these differences appeared to respond in an expected manner, with culture cells expressing proliferative and cytokine-like response signals, while blood subsets were largely expressing more interleukins and killer cell lectin receptor (KLR) genes suggesting interactions with other cell types in the peripheral blood environment.

4.4.3 Removal of the Culture Effect Highlights Underlying Cell Type Similarity

Removal of the ‘culture signature’ genes from the dataset had resolved the issue of samples grouping by their origin rather than their subset and suggests that the underlying transcriptome of blood and culture derived samples of the same subset were conversed through culture conditions. The major cell type markers, developmental and functional targets were similarly conserved. With the removal of the cell-cycle, proliferation and interaction genes found to be differentially expressed between pooled *in-vitro* and pooled *ex-vivo* samples, the remaining genes were capable of separating cell subsets into monocytes, pDCs, cDC1s and cDC2s, regardless of their growth conditions, furthermore, the grouping of the samples by hierarchical clustering reflects biological development and haematopoiesis with the monocyte and DC populations developmentally distinguishable. By PCA and t-SNE close grouping of the blood and culture equivalent populations were observed after the removal of the global ‘culture effect’ genes. This observation suggested that the transcriptomic difference between the blood and cultured samples was largely shared amongst each subset despite the differences in developmental pathways of each subset in the blood.

The PCA showed increased variance amongst the blood-derived samples compared to the culture equivalents, as noted during the gene removal step in 4.4.2. Again, this reinforces the conclusions drawn in section 4.2.2, that cultured cells were less heterogeneous than the genetically and environmentally distinct blood-derived equivalents taken from different healthy donors, however despite this variability, the overall expression pattern was conserved and cell types were clearly identifiable by their gene expression patterns after the ‘culture signature’ genes were omitted.

t-SNE visualisation of the high-dimensional data displayed each DC and monocyte subset in a separate quadrant of the plot, The blood and cultured pDCs formed the closest clusters although all four subsets still had some minor distinctions between their *ex-vivo* and *in-vitro* samples. These were typically global, minor expression differences, resulting in broadly similar cell types from the peripheral blood and culture.

Correlation of the developmentally important marker gene expression levels between blood and culture equivalent subtypes were conducive to the development of *bona-fide* DCs with cultured cDC subsets expressing major cDC development markers, rather than monocyte-associated markers typical of moDCs. This level of phenotypic and transcriptomic equivalence is a particularly important distinction if such cultured population are to be used for scalability in further cellular development assays, functional assays or in future clinical cellular therapy as a surrogate for *in-vivo* cDCs.

Although the NanoString panel lacked probes for some fundamental genes such as ID2, CD141 and TLR6, which are typical DC marker genes, enough were present to reinforce the verdict of genuine DC generation. CLEC9A is now used as a flow cytometry and FACS cell marker for cDC1 as it is more stable in expression than CD141 (Guilliams et al., 2016). Conservation of expression of CLEC9A, along with XCR1, IRF8, TLR1, TLR3 and TLR8 in both cultured and blood cDC1s was a positive indicator that both surface phenotype and transcriptomic profile of *in-vivo* cDC1s was recapitulated *in-vitro*. The same recapitulation was noted for CLEC10A, CX3CR1, IRF4, TLR1, TLR2, TLR4, TLR5, TLR8 with cDC2 cells and IRF7, IRF8, TCF4, TLR1, TLR7 and TLR9 on both pDC subsets. The combination of surface markers, phenotype markers, development genes and specialised functional components provides evidence that the culture system produced bona-fide DCs, distinct from the monocyte lineage. The ability to produce these cells in culture will open further avenues for research by providing the scalability required for functional assessment, developmental studies or for further clinical therapeutic use.

4.5 RESEARCH SUMMARY AND KEY POINTS FOR PROJECT PROGRESSION

Chapter 4 was focused on the comparison and correlation of DCs isolated from peripheral blood and their culture derived equivalents. As culture models are commonly implemented by DC biologists, immunologists and clinical scientists for investigating DC response to stimulants, DC development and for use in clinical immunotherapy, comparing the immune-transcriptome of cultured cells to *bona-fide* peripheral blood DCs was of paramount concern. It was clear from FACS and flow cytometric analysis of blood DCs and cultured cells that phenotypically equivalent cells could be developed *in-vitro*, but similarity by a dozen surface markers does not necessarily equate to the wider functional and developmental conservation of these cells through culture.

This chapter addressed the issue of comparability between blood and cultured dendritic cell and monocyte subsets through comparative RNA transcriptome analysis, uncovering a number of transcriptional changes affecting *in-vitro* generated cell populations and providing a novel basis to identify and remove these obscuring transcriptional changes to reveal the underlying conservation of each unique dendritic cell subset's phenotypic and functional features between the cells generated through the Human Dendritic Cell Lab culture system and primary peripheral blood DCs. The deconvolution of the culture signature from the dataset employed the same novel technique developed, tested and explored in chapter 3, but rather than comparing skin and blood equivalent cells, it compared blood cells to their cultured equivalents.

By exposing the fundamental cell subset specific signature from the dataset despite the dominating conditional signature, the transcriptional conservation of the generated cells was revealed to be significant. Cells grown in culture aligned well with their blood-borne equivalents. This was an extremely important finding as it suggests that the culture model employed in this thesis could recapitulate bona-fide DCs, well beyond cell-surface marker equivalence. This finding had a major impact on the research performed in the Human Dendritic Cell Lab, as it meant large quantities of DCs could be produced that would likely react under investigation in a comparable way to primary DCs. This would also have a wider research impact for projects where obtaining sufficient cell numbers for analysis is not feasible from human blood, or in the case of immunotherapy, would allow for the generation of vast quantities of immune-specific DC cells that are phenotypically, developmentally and functionally as capable as true blood DCs, in an easily reproducible manner.

Chapter 4 Figures & Tables

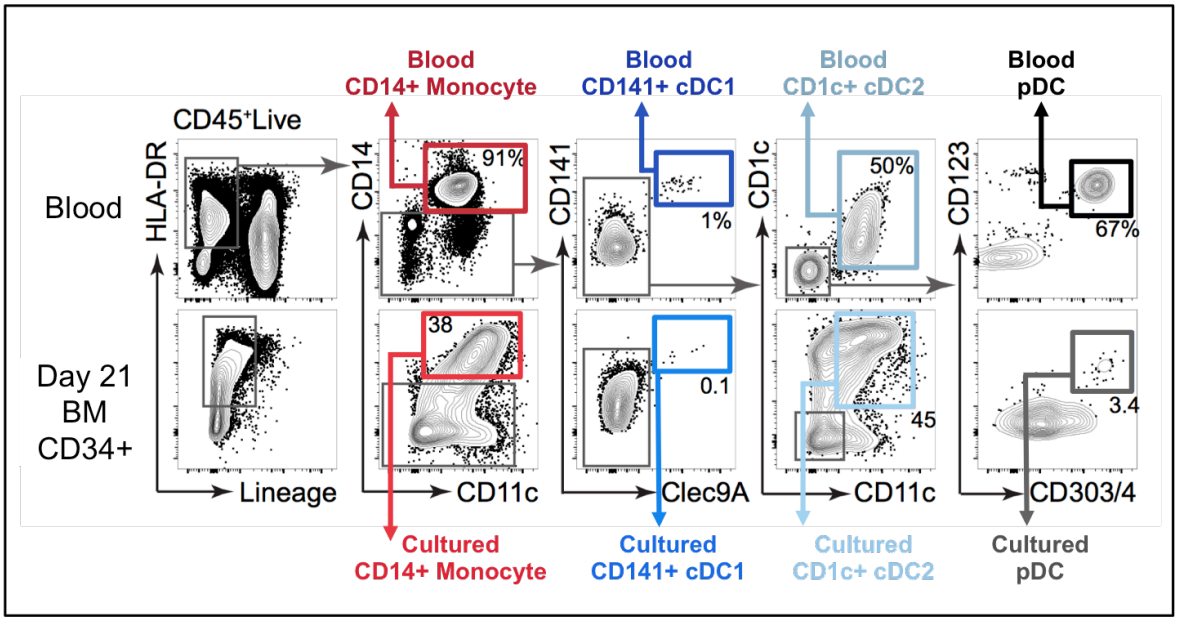


Figure 4.1: FACS Gating Strategy for Blood and Cultured Mononuclear Subsets

Gating was matched as closely as possible between peripheral blood subsets and their cultured equivalents. Classical monocytes were classed as HLA-DR+ CD14+ cells, while cDC1 cells were both CD141+ and CLEC9A+. cDC2 cells were defined by CD1c and CD11c positivity, while pDCs were defined by their CD123 and CD303/CD304 expression. The gating strategy relied on multiple defining surface markers to increase sample purity.

Panels were arranged from left to right with the following rationale (note: all cells used were also gated for: Live cells (DAPI-), HSC-derived (CD45+), singlets (FSC-A x FSC-H):

1. MHCII-expressing (HLA-DR+), Lineage- (CD3, CD19, CD20, CD56),
2. CD14+ monocytes (CD14+),
3. cDC1 (CD141+, Clec9A+),
4. cDC2 (CD11c+,CD1C+),
5. pDC (CD303/4+, CD123+)

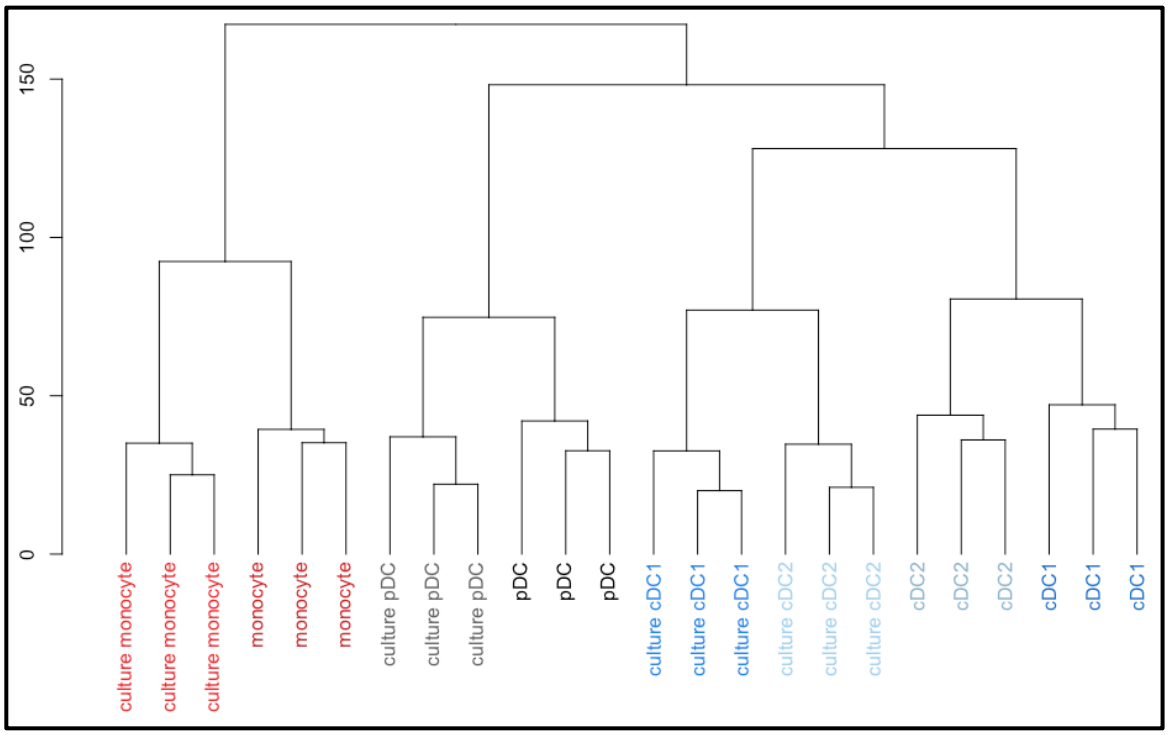


Figure 4.2: Dendrogram of blood and cultured samples using all NanoString Immunology_V2 and Panel+ genes

This hierarchical clustering dendrogram uses data generated on the NanoString nCounter Analysis platform with the Immunology_V2 codeset and the Panel+ additional probes for blood-derived and culture-derived mononuclear cell subsets. The Y-axis in this figure represents ‘height’ (a measure of increasing dissimilarity), increasing ‘height’ suggests clusters are less similar to one another. The first branch of this dendrogram splits both the cultured monocytes and blood derived monocytes from the other subsets. Along the dendritic cell branch of the dendrogram, pDC subsets branched off from the other DCs next. Both blood and cultuted pDCs appear here. The final cDC branches appear problematic as instead of grouping by cell subset, the cDCs are grouped by sample type. Both cultured cDC1 and cDC2 samples were grouped together, as were the blood equivalent subsets. It was suggested that this could be the result of a specific ‘culture signature’.

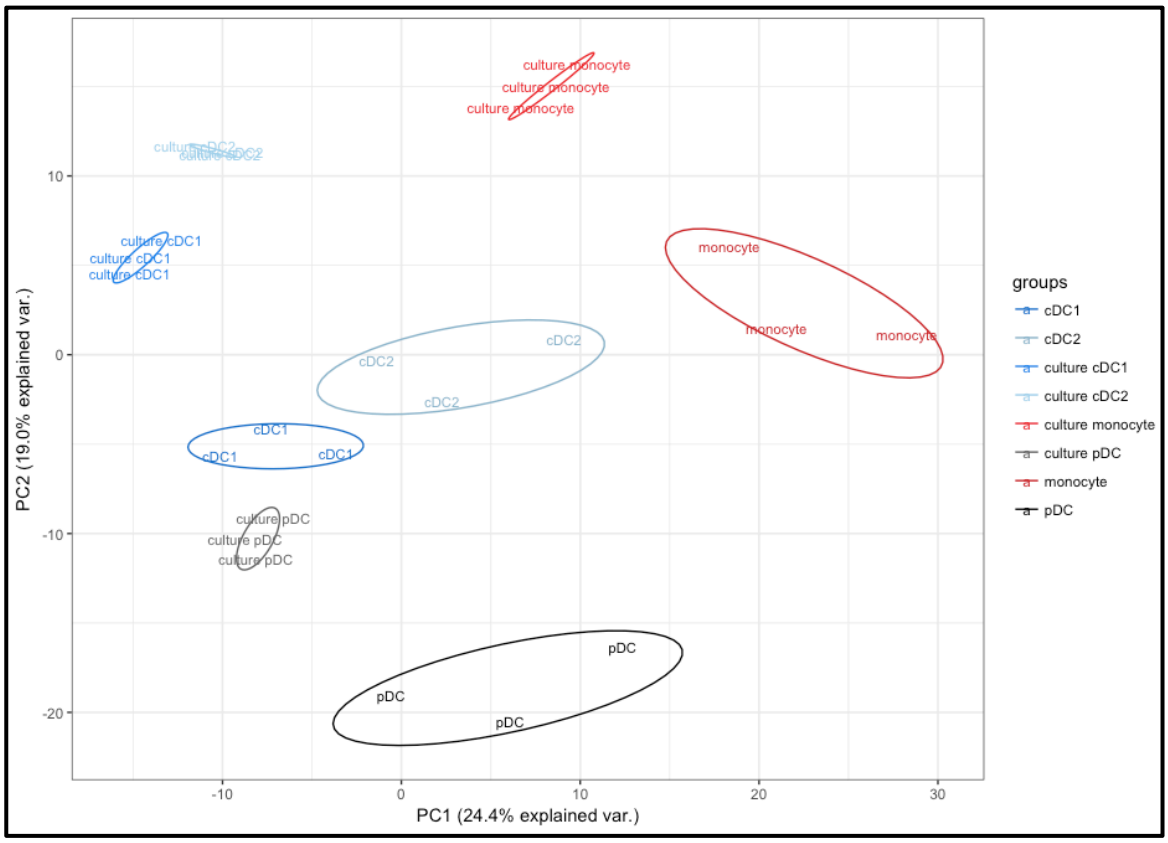


Figure 4.3: PCA of blood and cultured samples using all NanoString Immunology_V2 and Panel+ genes

Principal component analysis of the data generated on the NanoString nCounter Analysis platform with the Immunology_V2 codeset and the Panel+ additional probes for blood-derived and culture-derived mononuclear cell subsets show a similar shift in PCA rotation for cultured cells compared to their blood equivalents. PC1 accounts for 24.4% of sample variance and appears to loosely split monocytes from cDCs and pDC subsets. PC2 accounts for 19% variance and separates pDCs from the other subsets. Distance relationships of the blood subsets are similar to those of the *in-silico* experiment PCA (Figure 3.7) and the NanoString experiment PCA (Figure 3.10). The pattern of cultured cell subsets mirrors that of the blood-derived samples with cultured equivalents found in a position approximately 10-15 values less by PC1 and 10-15 values greater by PC2 than the blood samples. This pattern gave rise to an idea of a conserved ‘culture cell signature’.

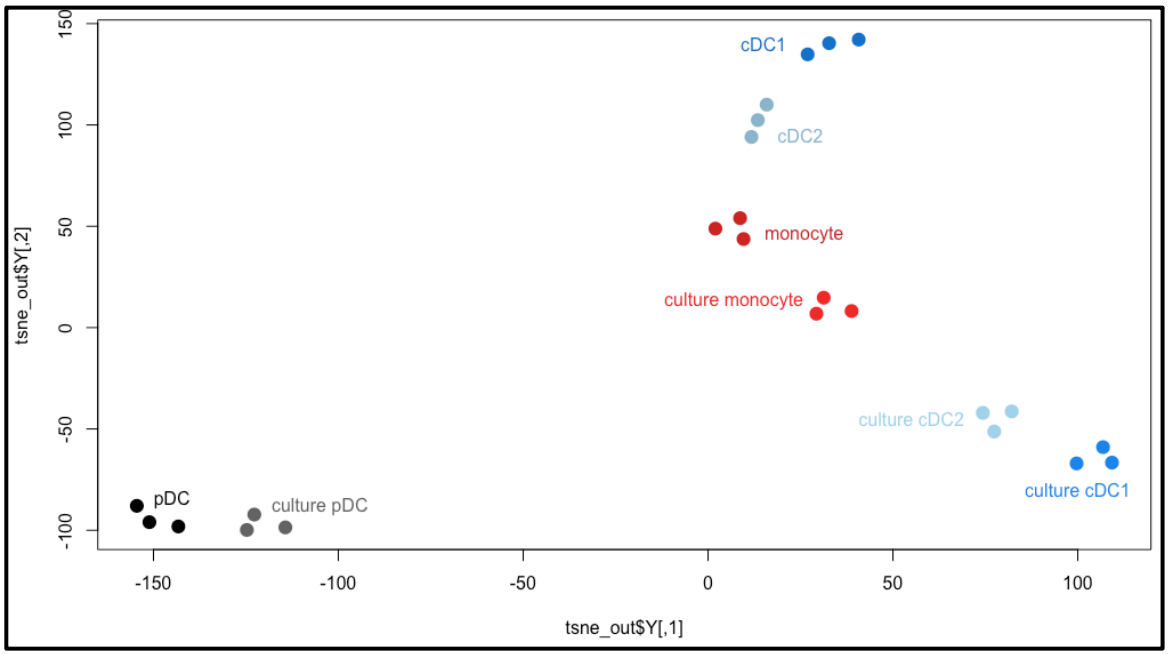


Figure 4.4: t-SNE plot of blood and cultured samples using all NanoString Immunology_V2 and Panel+ genes

t-SNE analysis of the blood and cultured samples using all NanoString Immunology_V2 and Panel+ genes produced a number of distinct subset population clusters. pDCs and cultured pDCs occupy the negative region of t-SNE1 and t-SNE2 space. Monocytes and cultured monocytes occupy the zero region of the same t-SNE space. As with the dendrogram (Figure 4.1), cDCs form a group separately from their cultured equivalents. cDC1 and cDC2 cells are positive on t-SNE2, while cultured cDC1 and cDC2 samples were negative along t-SNE2.

Cultured pDCs and cultured monocytes appeared to be closer to their *in-vivo* counterparts than the cDCs.

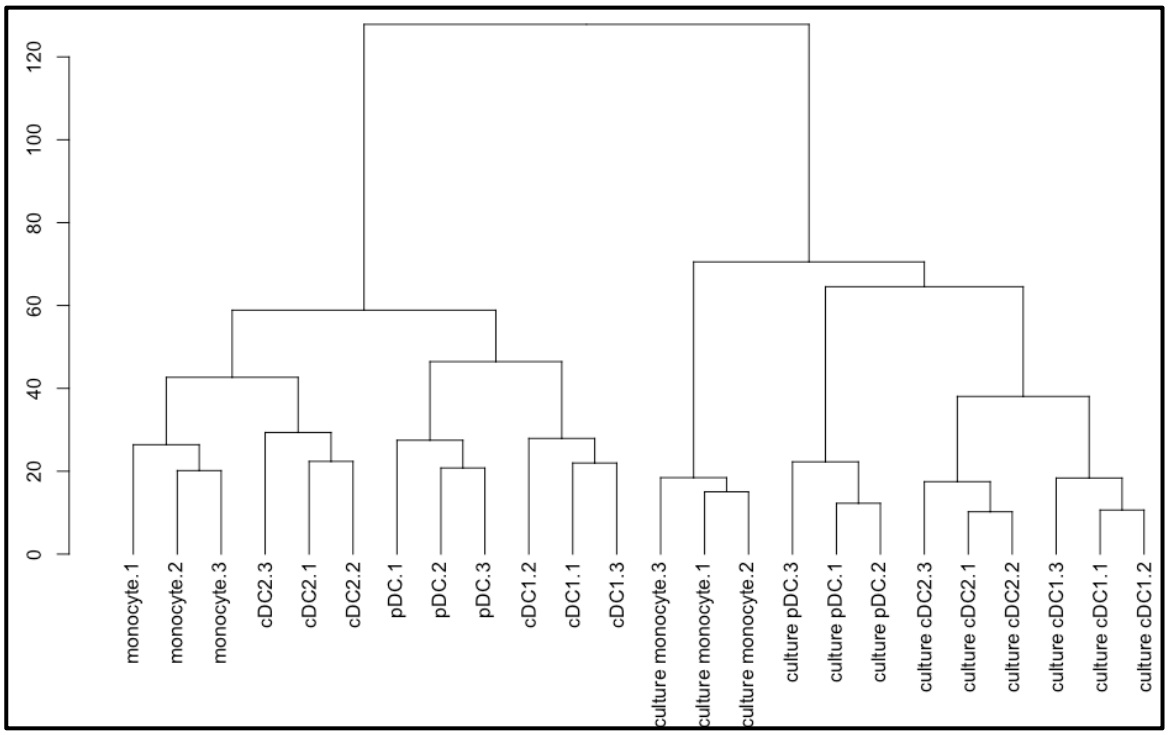


Figure 4.5: Dendrogram of blood and cultured samples using only the 'culture signature' geneset

This dendrogram was built based on a 230 gene signature produced as a result of a two-tailed t-test of all cultured samples against all blood samples. The Y-axis in this figure represents 'height' (a measure of increasing dissimilarity), increasing 'height' suggests clusters are less similar to one another. The resulting signature very strongly divided *in-vitro* from i*n-vivo* samples. Within this gene signature, the genes were capable of separating each individual subset, although the higher-order branching is not typical of gene expression analysis patterns. In the *in-vivo* branch of the dendrogram monocytes and cDC2 samples split initially from pDCs and cDC1s, while the *in-vitro* branch sees monocytes splitting away from the DCs, followed by pDC splitting from the cDC branch.

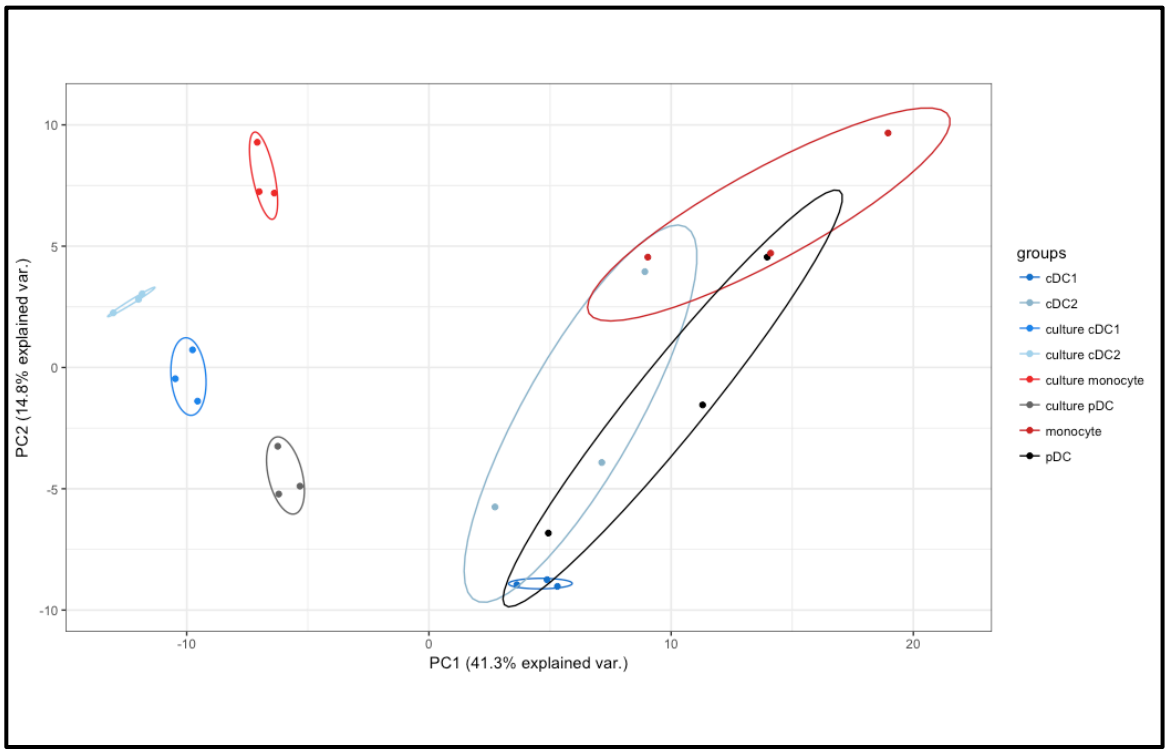


Figure 4.6: PCA of blood and cultured samples using only the 'culture signature' geneset

Principal component analysis of the 230 gene signature produced as a result of a two-tailed t-test of all cultured samples against all blood samples produced a plot separating cultured cells from their blood equivalents. PC1 accounts for 41.3% of sample variance and appears to strongly define *in-vitro* subsets from their *in-vivo* counterparts. PC2 accounts for 14.8% variance and separate the individual cultured cell types well. This 230 gene signature highlights the high degree of variation found in normal human samples compared to rigorously controlled culture conditions. The in-vitro subsets form tight, distinguished clusters, while the primary blood cells appear far more varied by PCA, resulting in loose, overlapping groups.

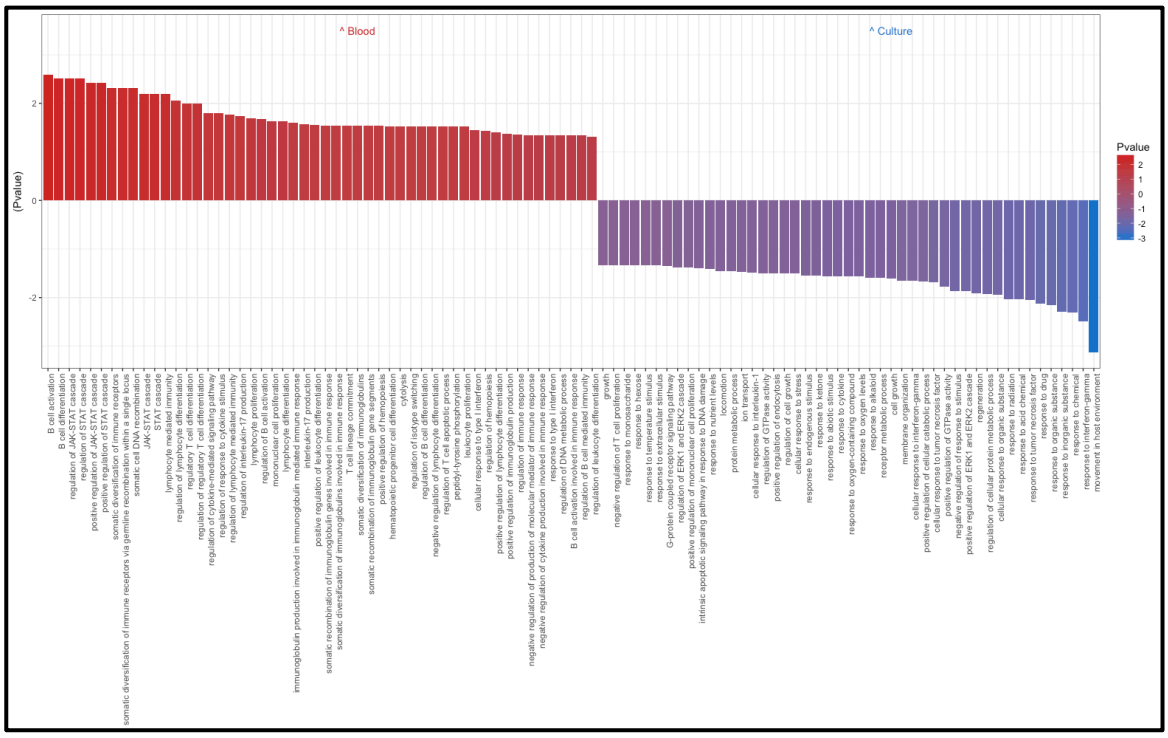


Figure 4.7: Functional enrichment diagram for the ‘culture signature’ geneset

Functional enrichment of the blood and culture differentially expressed genes produced a significant number of contributory functions specific to *in-vitro* or *in-vivo* derived cell types. For the blood subsets, many of these functions shared a theme of immunity, evidence of interactions and exposure to foreign antigens and other haematopoietic cell types as a result of their position and role in circulating peripheral blood. As for the cultured cell types, functions shared a theme of proliferation, apoptosis, cellular stress, cellular migration and response to interleukins and stimulants, reflecting their forced division and boosted cell cycle under culture conditions.

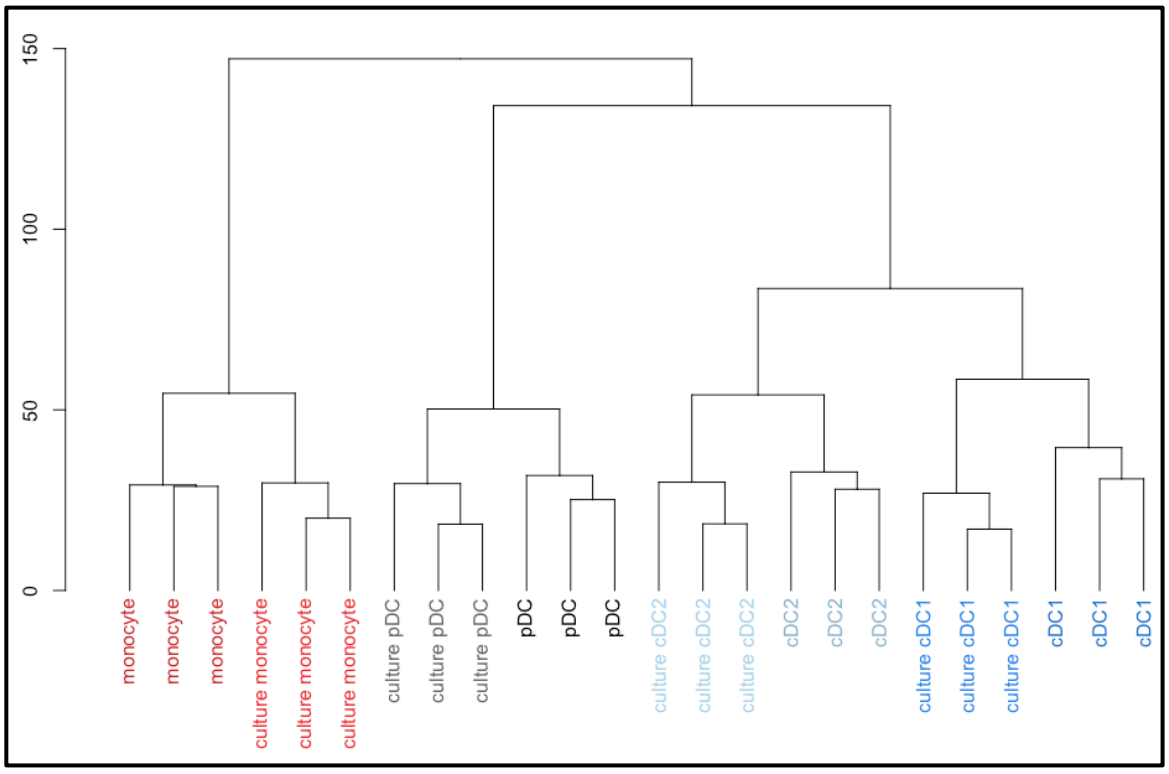


Figure 4.8: Dendrogram of blood and cultured samples after removal of a conserved 'culture signature' geneset

This hierarchical clustering dendrogram uses data generated on the NanoString nCounter Analysis platform with the Immunology_V2 codeset and the Panel+ additional probes for blood-derived and culture-derived mononuclear cell subsets. The Y-axis in this figure represents 'height' (a measure of increasing dissimilarity), increasing 'height' suggests clusters are less similar to one another. The first branch of this dendrogram splits both the cultured monocytes and blood-derived monocytes from the other subsets. Along the dendritic cell branch of the dendrogram, pDC subsets branched off from the other DCs next. Both blood and cultured pDCs appear here. In this figure, the final cDC branches are now resolved, with both cDC1 groups clustering together and both cDC2 subsets grouped together. Removal of 'culture signature' genes has aided the grouping of the cells by subset, rather than origin.

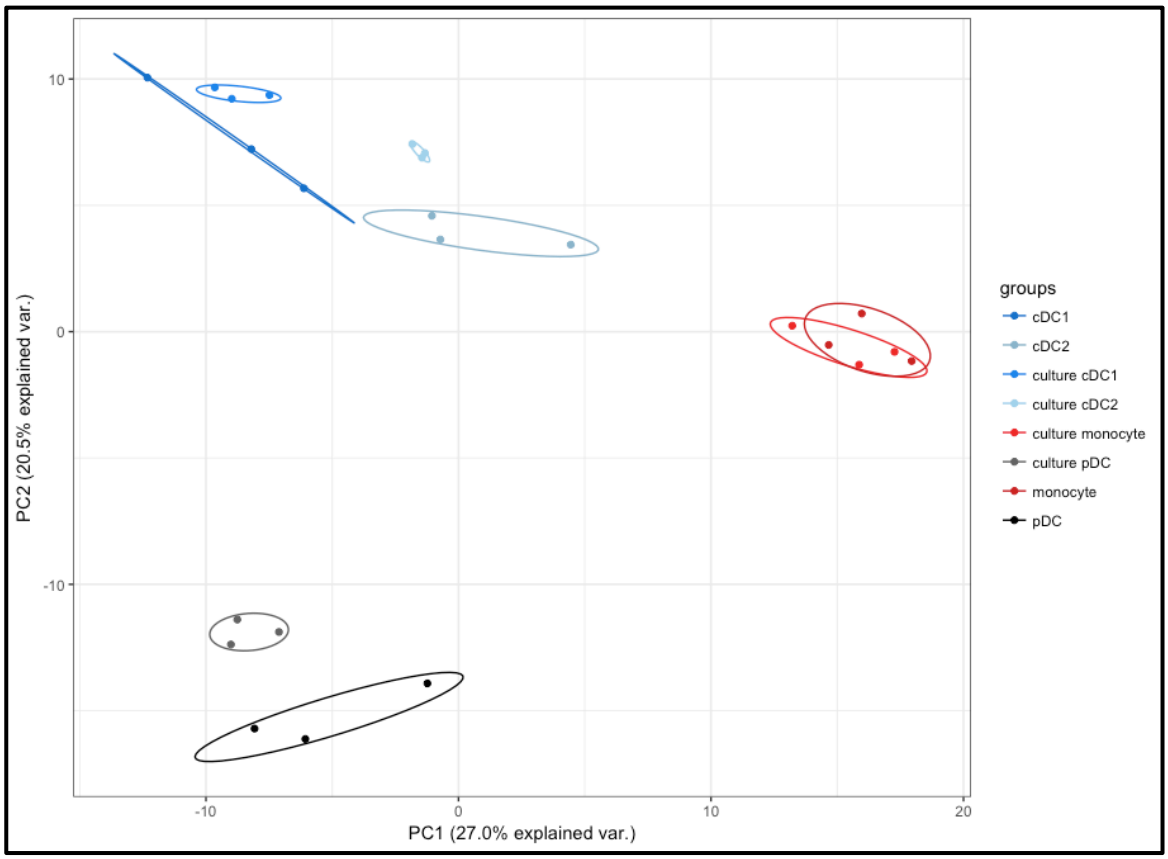


Figure 4.9: PCA of blood and cultured samples after removal of a conserved 'culture signature' geneset

Principal component analysis of the data generated on the NanoString nCounter Analysis platform with the Immunology_V2 codeset and the Panel+ additional probes for blood-derived and culture-derived mononuclear cell subsets produce four cell subset clusters after removal of a culture signature. PC1 accounts for 27% of sample variance and splits blood and cultured monocytes from cDCs and pDC subsets as well as pulling cDC1s from cDC2s. PC2 accounts for 20.5% variance and separates pDCs from the other subsets. Distance relationships of the blood cell subsets and their *in-vitro* counterparts highlights the beneficial effect of removing a general culture signature from the dataset. Each of the four major clusters is composed of both *in-vitro* and *in-vivo* generated cells. Monocyte subsets are overlapped, with the other culture subsets falling much closer to their blood equivalents.

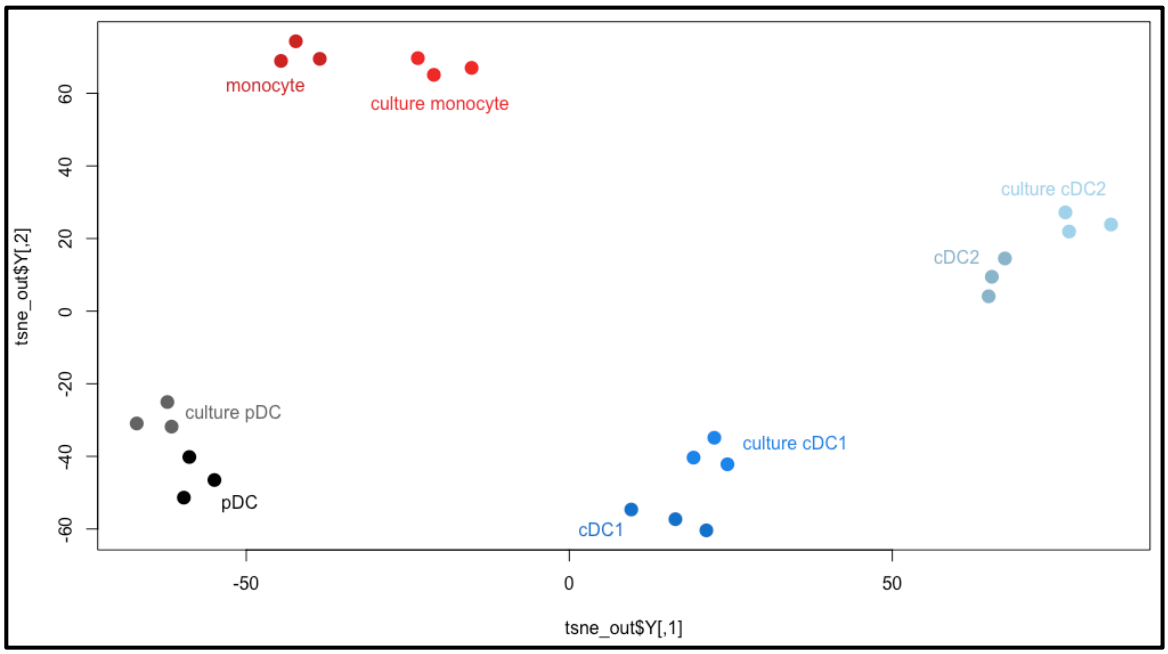


Figure 4.10: t-SNE analysis of blood and cultured samples after removal of a conserved 'culture signature' geneset

T-SNE analysis plot of the data generated on the NanoString nCounter Analysis platform with the Immunology_V2 codeset and the Panel+ additional probes for blood-derived and culture-derived mononuclear cell subsets produced four cell subset clusters after removal of a culture signature. t-SNE1 splits blood and cultured pDCs from cDC1s and cDC2 subsets, while t-SNE2 splits pDCs and cDC1s from cDC2 as well as monocytes. This pattern results in each subset being located in a single quadrant of the t-SNE space.

In contrast to Figure 4.3, the cDC1 and cDC2 subsets are clustered by cell type rather than their generation *in-vitro* or *ex-vivo*.

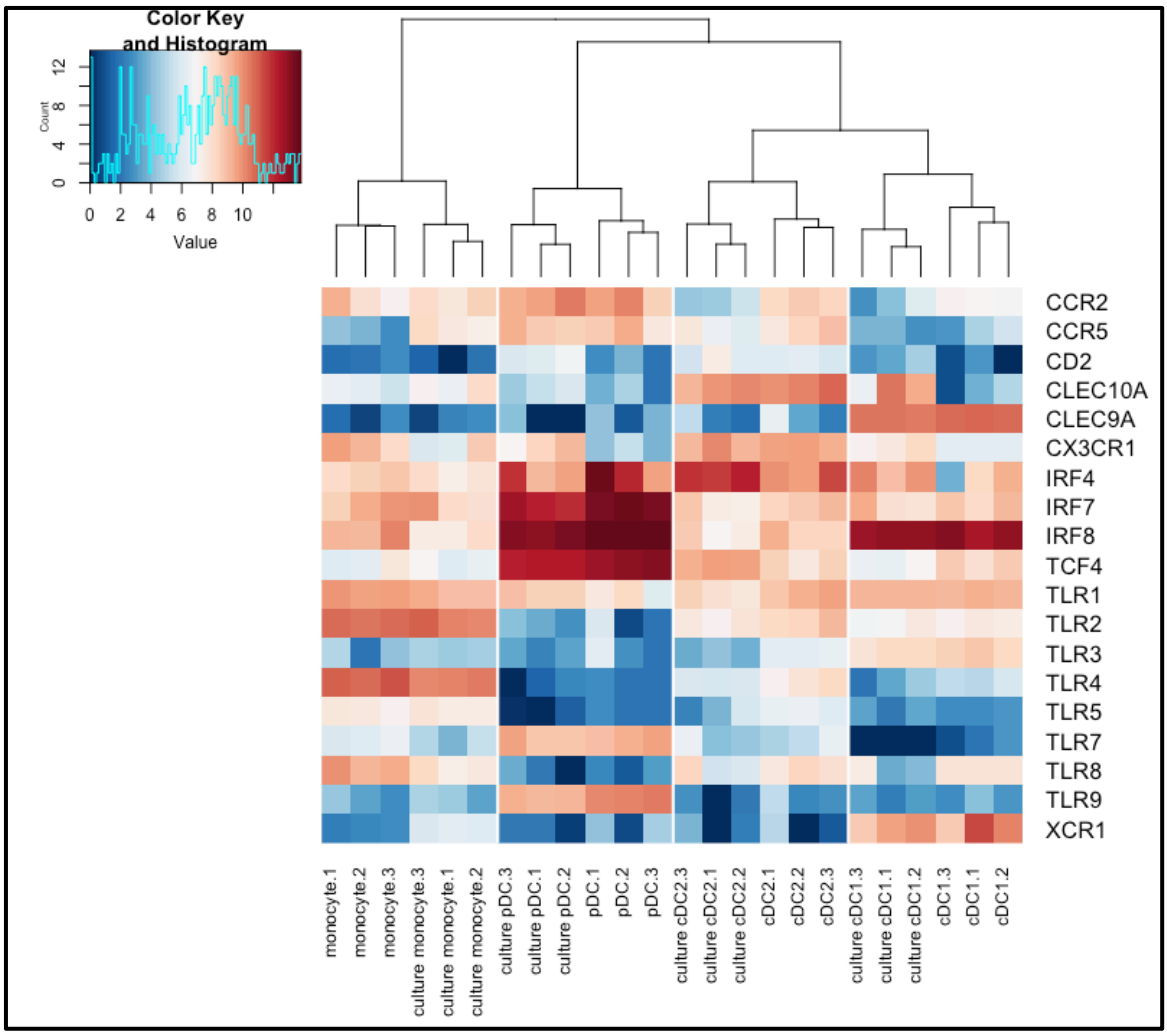


Figure 4.11: Heat-map of marker genes relevant to DC development for blood and cultured cells

TLRs, CCRs and IRF genes are displayed along with other important markers of mononuclear cell development for the blood and cultured subsets show conservation of these major signatures. Red signifies high expression for that marker, blue signifies low expression. High conversation of expression between the blood and cultured equivalent populations reflects a strong similarity between cell subsets under each condition.

Chapter 5: A SINGLE-CELL APPROACH TO DENDRITIC CELL DIFFERENTIATION

Primary research question:
Can transcriptomic analysis identify DC lineage priming in progenitor cells and precursors?

Sub-topic questions:
1. Can single-cell RNA-sequencing be used to investigate dendritic cell precursors?

2. Are DC-precursors skewed towards mature DC signature expression at the single cell level?

3. Does the *in-vitro* development assay on pre-DC populations correlate with transcriptome-level expression patterns?

5.1 INTRODUCTION

Transcriptome profiling of dendritic cells has largely been performed using bulk material, either for standard RNA-sequencing or microarray-based analysis, to provide a population level expression profile for major mononuclear cell populations. These technologies provide average gene expression values across a large cell population, useful for identifying global or population level gene expression differences between distinct cellular lineages or disease states. With the growth and increasing affordability of single cell transcriptomics in recent years, the true heterogeneity of dendritic cell biology has begun to be revealed. Using single cell technologies with dendritic cell populations has revealed much greater heterogeneity between cell types, culminating in the identification and isolation of multiple potential new DC and monocyte populations in human blood (Villani et al., 2017).

While culture studies have enabled researchers to speculate on maturation and development of dendritic cells from their pre-cursors, cell culture has inherent complications when compared directly to primary DCs. Murine work also share many of these same difficulties with debate is still ongoing in relation to the correlation between the human and mouse immune system and deconvolution of *in-vitro* derived conclusions that apply *in-vivo*.

In mouse, pre-DCs are traditionally considered to originate from a common monocyte and dendritic cell progenitor and in turn develop into pre-pDCs and pre-cDCs in the bone marrow (Lee et al., 2015). The work in this chapter, which represents one of the first single cell experiments at Newcastle University, explores the transcriptomic diversity at the single cell level within the peripheral blood pre-DC population, as identified by surface phenotype. Analysis focuses on single cell data manipulation, interpretation, visualisation and analysis techniques to generate a specially adapted pipeline. This is followed up with cell culture of pre-DCs to investigate their developmental heterogeneity *in-vitro*, supporting the early lineage-priming model of haematopoiesis described in chapter 1, section 1.3.2.

5.1.1 Advances in single-cell RNA sequencing

With some of the first publications arising around 2009, including expression of a single mouse blastomere (Tang et al., 2009), single cell sequencing is a very new technique in biological research. Initially prohibitively expensive, early publications featured relatively few cells, however, commercial competition, streamlining, multiplexing and protocol refinements have since driven the costs down to approximately $10 per cell (Han, 2015). With this accessibility, the number of single-cell sequencing publications has risen dramatically over the past three years. Broadly speaking, single cell protocols have evolved into two main strands. Plate-based protocols were first to be developed, offering low-throughput, high read-depth transcriptome sequencing, usually incorporating flow cytometry-based index-sorting of individual cells into wells of a 96-well plate. While this process is time-consuming and expensive for larger numbers of cells, users can refer back to the immunophenotype data produced during cytometry sorting to infer phenotypic properties of their collected cells and correlate surface-marker expression to transcriptome level counts. The greater read depth of plate-based techniques typically provides transcripts for up to 5,000 unique genes per cell. Such processes are suited for deconvolution of closely-related populations or investigation of heterogeneity in previously described ‘homogeneous’ populations (Papalexi and Satija, 2017). Alternatively, droplet-based microfluidic platforms provide multiplexing capability for high-throughput sequencing at the expense of number of transcripts per cell, producing around 1,000 mapped genes per cell. Cell multiplexing allows for hundreds of cells to share the same sequencing lane, yet remain distinguishable by unique molecular identifiers introduced to each cell. Such high-throughput processes are much cheaper on a per-cell basis than plate-based sequencing and are typically preferred for discovering rare cell types or investigating heterogeneous cell types where the lower sensitivity is not an issue.

Current large-scale single cell projects such as ‘The Human Cell Atlas’ have combined both approaches in a complementary manner, using droplet-based approaches to gain unbiased overall population insights and identify cell clusters of interest. Gene markers for these subsets will then be used to isolate and enrich for populations of interest before plate-based methods are employed for in-depth transcriptome analysis (Rozenblatt-Rosen et al., 2017). While this approach is expensive and data-intensive, it may yield great insights into the immune cell compartment across multiple human tissues.

The capacity to identify novel cell types and marker genes that can be used to isolate these same cell populations may be combined with culture techniques to generate large cell numbers required for functional and developmental investigation of extremely rare cell types where *ex-vivo* collection may not be viable and care must be taken when aligning populations identified by transcriptomics with previously defined cells.

5.1.2 The pre-DC concept

Dendritic cells are believed to originate from two distinct lineages in the classical model of haematopoiesis, arising from myeloid and lymphoid progenitor cells (Doulatov et al., 2010), although their observed mature phenotype and functions were conserved in mouse xenotransplantation models (Ishikawa et al., 2007) and more recent alternative human models of haematopoiesis largely undermine the significance of the myeloid and lymphoid pathways in DC biology (Hamey and Göttgens, 2017; Notta et al., 2016; Paul et al., 2015). The traditional interpretation of DC development from both common lymphoid and common myeloid progenitors is typically based on flow cytometry gating strategies and thus may be reliant upon relatively few marker genes and arbitrary gating points for cell sorting. With this basis, it is possible that early DC-primed progenitor cells may already be present across multiple sorting gates in the traditional model, therefore affecting the interpretation of lineage contribution assays associated with 'later stage' progenitor populations (Velten et al., 2017). Because of these shortfalls, the alternative approach to haematopoiesis has been adopted by multiple groups working in the field of single cell transcriptomics, revolving around transitional early lineage priming from HSCs in the absence of stable, discrete progenitor cell types (Lee et al., 2017; Schlitzer et al., 2015; Velten et al., 2017).

5.2 MATERIALS AND METHODS

5.2.1 Sample Collection and Isolation

All material was collected with written, informed consent for research purposes under ethical approval. Whole blood was diluted with PBS and layered over lymphoprep (Stemcell) density solution before centrifugation. The resulting peripheral blood mononuclear cell layer was aspirated off and washed before counting with a haemocytometer. Resulting cells were then stained for FACS. Cell isolation protocols for both peripheral blood and bone marrow are discussed in chapter 2, section 2.2.

5.2.2 FACS Sample Sorting

Single cells were isolated by FACS into 96-well plates containing 2uL of lysis buffer (RNAse-free water, 2uM RNAse inhibitor and 0.1% TritonX). Six wells contained mini-bulk (10 pooled cells) mature sample subsets, composed of CD14+ monocytes, pDCs (CD123+ CD303/4+ CD2-), cDC1s, BTLA+ cDC2s, BTLA- cDC2s, and CD34+ progenitor cells. Into each of the remaining wells HLA-DR+, Lineage- (CD3, CD19, CD20 and CD56), CD14-, CD123med/+ cells were sorted, omitting any CD303/4+.CD2- cells defining mature pDCs. A pre-processing overview is described in Figure 5.1 and sort gating is displayed in Figure.5.2. CD303/4+,CD2+ cells were collected as a single population with downstream gating was used to separate potential sub-populations of the sort gate, including early-pre-cDC2 (CD11c- CD5+), pre-cDC2 (CD11c+), CD123medium cells, pre-pDCs (CD303/4+ CD2+) and a mixed population of CD303/4- CD2+ cells termed ‘Tri-lineage’ cells.

5.2.3 Plate and Library Preparation

Reverse transcription, library preparation and sequencing were performed by the Oxford Genomics Centre. ERCC spike-ins were added to each well to adjust for technical variability. No blank wells were used for background subtraction. A modified SMARTseq2, full-length reverse transcription protocol was used in the absence of unique molecular identifier (UMI) sequences. The library prep utilized a Nextera XT DNA Library Prep Kit and paired-end reads were developed at 75base-pair length at 2.5 million read depth using the Illumina HiSeq 4000 platform.

SMARTseq2 was identified by the Enard group as having the greatest sensitivity, detecting the most number of genes compared to other RNA-Seq methods including CEL-seq, Drop-seq, SCRB-seq and Smart-seq (Ziegenhain et al., 2017).

5.2.4 Platform Selection

The Illumina HiSeq4000 was utliised as one of the newest HiSeq platforms available. It is capable of sequencing many standard libraries at a more efficient rate than the previous generation HiSeq2500. The HiSeq4000 flow cell contains billions of nanowells resulting in narrower feature variability.

5.2.5 Data Preparation and Matrix Building

From the Illumina HiSeq4000 machine, sample data was screened for read quality using Trimmomatic, a flexible trimming tool capable of handling paired-end data. Any technical errors were visualized based on the .BAM output. Low quality bases were trimmed from the reads prior to mapping of the reads to a reference genome using STAR aligner. In this instance GRCh38 with Gencode gene set volume 24 and additional ERCC library maps included. Once mapped and screened for mapping quality, reads for each gene and each cell were quantified using HTSEQ. From the HTSEQ report, gene count tables for each cell were collated into a single count matrix for further refinement and analysis in R.

5.2.6 Normalisation and Analysis

The ‘SCATER’ package (McCarthy et al., 2017) was used for expression QC, filtering out genes with very few counts, samples with few total reads and low features, as well as removal of cells with a high percentage of mitochondrial counts. Data visualisation was also performed using SCATER alongside variability analysis to determine the overall expression patterns and sources of sample variability. To reduce unwanted variability in the data, particularly in ERCC control counts, the RUVg function of the ‘RUVseq’ package (Risso et al., 2014) was implemented to adjust the data.

Initial differential expression analysis was performed on ‘R’ using ‘M3Drop’ (Andrews and Hemberg, 2017) and followed-up with ‘SC3’ (Kiselev et al., 2016) cluster analysis to identify cells with similar gene expression profiles.

Final correlations of pre-DCs to mature cell type signatures was performed using the signatures generated in Chapter 3 and visualized using ‘ggplot2’ (Wickham, 2009b). The final comparison to CD100+, CD34med cells incorporated data presented by Villani *et al* (Villani et al., 2017).

5.2.7 Culture Conditions and Staining

Cell culture and follow up was performed by Dr Urszula Cytlak-Chaudhuri from the Human Dendritic Cell Lab. Cells collected from the pre-DC gates displayed in figure 5.2 were seeded onto a pre-prepared (>/=4hrs prior) feeder layer of5,000 OP9 stromal cells per well in 96 well U-bottomed plates. Cells were cultured in 200mL aMEM (Gibco™) with 1% penicillin/streptomycin (Sigma), 10% Fetal Calf Serum (Gibco), 20ng/ml GM-CSF (R&D systems), 100ng/ml Flt3-ligand (Immunotools) and 20ng/ml SCF (Immunotools). Half the volume of media (including cytokines) was replaced every seven days. At Day 14 cells were harvested on ice, passed through a 50micron filter, washed and stained for flow cytometric analysis. Cells were stained in aliquots of up to $1x10^7$ cells in 100µl of DPBS with 2% fetal calf serum and 0.4% EDTA. Dead cells were excluded by DAPI (Partec). Cell numbers and barplots were produced using GraphPad Prism version 6.

5.3 RESULTS

5.3.1 Data Acquisition and Pre-Processing

Pre-processing of the scRNA-Seq data was performed in three stages. From the initial text-based FASTQ files produced by the sequencer, 'Trimmomatic-0.33' (Bolger et al., 2014) was used to remove up to 20 bases from the end of a read if the read was below the threshold quality (Q10 – indicating an error rate less than 0.1) and drop any read completely if the length of the read was below 60 bases in length. The remaining reads were then mapped to the reference genome GRCh38 version 25, augmented with ERCC control sequences. In total, 2.5 million reads were obtained for the Pre-DC plates. The aligner tool of choice for this analysis was the 'STAR' package (version 2.4.0j) (Dobin et al., 2013) implemented in Java. A low proportion of mapped reads in a cell indicated some contamination or other issue and resulted in the removal of the cell from analysis. Data for the number of uniquely mapped reads, unmapped reads and multi-mapped reads (aligning to more than one location on the genome) were recorded, however only uniquely mapped reads were used in the analysis. SAMtools-1.3 was used (Li et al., 2009) to convert the output '.sam' files into the compressed '.bam' file type before the Python implementation of HT-seq version 0.6.1 (Anders et al., 2015) was finally used in the pre-processing step for gene-level expression quantification. The output gene-level counts were collated into a single count matrix for data normalisation and analysis in the 'R' environment. This process is outlined in figure 5.1. From here they were annotated accorded to the 'Single Cell Analysis gates' they fell into in Figure 5.2. These cells equated to pre-DC populations displayed in Figure 5.3 as displayed over the 'early-priming' model of the haematopoietic lineage tree.

5.3.2 Quality Control of Pre-DC Single Cell RNA-Seq Data

Quality control steps are fundamental to accurate interpretation of single cell RNA-sequencing experiment data. After the count matrix was produced and incorporated into 'R', ERCC and mitochondrial genes were marked for later interrogation before the 'SCATER' package was implemented for initial quality control. Figure 5.4 and Figure 5.5 display the cut-off values (displayed as a red line) for total counts and total features, respectively. Samples with less than 25,000 total counts (reads) or 2,000 total features (genes) were considered 'failed' and removed from subsequent analysis. Such samples were likely damaged, burst during sorting, or did not correctly undergo library-prep. In these failed samples, ERCC control spike-in reads and mitochondrial genes are usually highly proportionally represented in the total counts. None of the samples failed the total counts QC, suggesting good read numbers per cell, however 18 of the 92 samples failed the total features QC.

A non-normally distributed data profile can be observed in Figure 5.4, with cells ranging up to 1.4 million total counts. Most samples exhibited between 200,000 and 800,000 total counts.

In Figure 5.5, after exclusion of the low feature cells from the dataset, a near normally distributed profile remained with most cells expressing between 3,000 and 5,000 total features. The six samples with the greatest number of features (between 5,000 and 10,000) were the mini-bulk cell samples. As these contained a pool of 10 cells, increased feature detection was expected here.

Percentage of mitochondrial genes detected for each cell sample were displayed in Figure 5.6. Mitochondrial gene percentages are frequently used as a quality score in single cell analysis. High percentages of mitochondrial genes are usually the result of cell degradation. As single cell analysis is a relatively new topic in research, no consensus has yet been reached in the single cell community for a definitive cut-off value for mitochondrial reads, but recent papers confirm that lower percentages of mitochondrial reads are an indication of greater quality (Bacher and Kendziorski, 2016; Ilicic et al., 2016). Commonly, cut-offs are placed between 10% and 25% depending on the experiment and focus. For this analysis, <15% was used as a cut-off resulting in the exclusion of two cells, one of which was undefined at the flow cytometry level, and the other of which was labeled as a CD5- 'Trilineage' cell by flow gating.

Figure 5.7 displays the ERCC spike-in read percentages for each cell. In similarity to mitochondrial read percentages, high ERCC percentage reads indicate the cell was damaged, degraded or underwent incomplete library prep. For this analysis, a 25% cut-off was used to filter the data, resulting in 15 of the 92 cells failing this QC step.

The final quality control results are displayed in table 5.1 as a summary of cell (5.1a) and gene (5.1b) filters. In summary, 21 cells were removed by the quality control filters for high mitochondrial gene or ERCC control reads, less than 2,000 mapped features or less than 25,000 total reads. The gene-level filters reduced the dataset from 60,675 possible features to 22,951 features with at least one count. This was further refined to remove any genes that were expressed in less than two cells, reducing the gene list to 14,412 genes.

5.3.3 QC Visualisation and Pre-Normalised Data Variance

Before determining the extent of normalisation, data quality and sample variance were visualised. Figure 5.8 highlights the correlation between the percentage of feature controls and the total number of features. Good quality cells are expected to have high total features and low percentage of feature controls. As indicated, the unknown/unclassified samples displayed in pink exhibited the worst correlation, with high feature controls and low total features; these were thus excluded from the normalisation and expression analysis stages. Conversely, the mature/mini-bulk samples displayed above 4,000 total features and less than 20% feature controls. The samples passing QC were those with the least correlation between feature controls and total features.

By displaying the top 50 most expressed features in the dataset, the main contributors to sample variance can be uncovered. Narrow distribution of the 50 top expressed genes, is suggestive of good transcriptome coverage. If the top 50 features accounted for most of the total counts and included many ERCC or mitochondrial genes, the experiment may have failed. Figure 5.9 displays the top 50 most expressed genes in the dataset, accounting for 27.6% of all counts; of these six appeared to be ERCC controls. This result suggests that future experiments would benefit from greater dilution of the ERCC spike-ins so they do not occupy as much sequencing real-estate. interestingly, the ERCC controls in the top 50 genes were highly variable between cells. This may have been an artifact from failed cells (where control reads are a typically high percentage of total counts) but may have indicated variance within the ERCC control counts. For this reason, 'RUVseq' normalisation was implemented based on ERCC counts to normalize the data in section 5.3.3.

Figure 5.10 and Figure 5.11 exemplify typical single-cell experiment features, principally, grouping of cells by total features. These figures highlight the need for content normalisation in these datasets. Figure 5.10 was produced as a 'first-glance' at the expression variability by cell type. Principal Component 1 accounted for 31% of the variance and appeared to correlate very strongly with the total number of features in the cells. The unknown samples in pink all expressed less than 2,000 total features and appeared as a tight cluster on the positive region of Component 1. The main bulk of pre-DC cells clustered loosely around the zero-point of Component 2 and spread along Component 1. Principal Component 2 accounted for 7% of the variance and clearly separated the mini-bulk samples from the pre-DC population, most strongly defining the CD34+ sample, along with the mature BTLA+ and BTLA- cDC2s and mature cDC1. Log_2 transformation of the data in Figure 5.11 provided greater separation of the cells, but still indicated total features were a major source of pre-normalised variance. This was further displayed by t-SNE in Figure 5.12. Evidently, without proper normalisation the data would not be interpretable. The defining feature of the pre-normalised dataset was the number of total features, rather than biological variance.

The final QC-step in the single cell pipeline incorporated the ‘plotQC’ function of ‘SCATER’ and displayed in Figure 5.13. The relative impact on explained variance in the dataset was revealed with a linear regression model plotted against each variable. Variables with a greater shift and higher density exhibit the greatest variance. Flow gating cell type designations are a major source of sample variance in this diagram, due in part to the difference in expression profile of the mature mini-bulk samples and the undefined low expressing ‘QC failure’ samples. ERCC contols (displayed in green) exhibited higher variance than total counts (in red) and thus ‘RUVg’ normalisation was implemented from the ‘RUVseq’ package to adjust for ERCC variance.

5.3.4 Normalisation and Variance Reduction

Normalisation of the data was performed in two stages. Initially, normalisation for library size was performed by converting the data into ‘counts per million’ using the ‘SCATER’ package. This adjusts the data to account for the differing reads detected in each cell. Without this normalisation, cells with higher features would likely also have the highest expression counts for many genes. ‘RUVg’ was used to further normalise the data by removing variation attributable to differences in the ERCC-control counts (Risso et al., 2014) using a factor-analysis based method developed on the ERCC controls and subsequently applied to each cell sample. Figure 5.14 highlights the correlation between PCA and total features, particularly for the mini-bulk samples and poor quality cells. With the principal components principally separating by feature counts, downstream analysis of the dataset would have been convoluted. After normalisation and conversion to counts per millions, variance attributable to ERCC spike-ins were modestly reduced as shown in Figure 5.15. Cell type annotations remained highly associated with variance suggesting cell type differential expression may have been detectable.

5.3.5 Global Expression Patterns of pre-DCs

Initial expression analysis of the normalised data required the use of principal component analysis and t-SNE techniques. The analysis depicted in figure 5.16, while not producing strong clusters due to the similarity of the cells analysed (all pre-DCs were taken from a single pre=DC sort gate) exhibits some detectable groupings. Both BTLA+ (red) and BTLA- (light green) mature cDC2 samples were grouped closely along the -10 region of PC1, with the majority of the pre-cDC2-like pre-DCs (brown) in the same region. Early pre-cDC2 cells (green) were also located along the lower region of the pre-cDC2 cells suggesting some correlation and progression from early-pre-cDC2, to pre-cDC2 to mature cDC2 cells.

Equally, the mature pDC sample (light purple) was located in the negative region of PC2 with the CD5- pre-pDC cells (in orange) again, suggesting some developmental relationship between these cell types.

Further visualisation using t-SNE in Figure 5.17 showed both BTLA+ (red) and BTLA- (light green) mature cDC2 cells were located in the same region of t-SNE space, surrounded by the majority of the pre-cDC2 cells (brown) and interspersed with early pre-cDC2 cells (green). A small cluster of cells in the positive region of tSNE1 and tSNE2 was formed around the CD34+ sample (pink) composed of pre-cDC2 and early-pre-cDC2 cells along with CD123med CD11c- and CD123med CD5 low cells, which may be the most immature of the pre-DC subsets.

The lower portion of the t-SNE diagram contained CD5- pre-pDCs and the pDC mini-bulk sample suggesting some degree of pDC-like features in a subset of pre-DCs.

5.3.6 Differential Expression Testing and Clustering

'M3Drop' was used for initial differential expression testing but proved unsuccessful due to the similarity of the pre-DC cells (Figure 5.18). This 'R' package was designed to overcome the issue of dropouts in single cell data with the assumption that the majority of dropouts occurred as a result of failure of reverse transcription.

Traditional differential expression packages used in bulk-RNAseq deal poorly with high-zero data. M3Drop's Michaelis-Menton-based modeling was designed to overcome this by modeling the pattern of dropouts in single cell sequencing. This dataset was composed of very similar, non-mature cells and had many dropouts and relatively low coverage, resulting in no genes being identified by M3Drop as highly variable. This was to be expected, as the isolated pre-DC cells were from the same FACS gate and expected to differ by only a small number of genes. To determine if these pre-DCs were skewed towards a particular mature-cell phenotype and de-convolute the dataset, the mature cell signature designed in Chapter 3 was applied across the 14,412 expressed features to produce a 647 DC/Monocyte gene sub-dataset.

SC3 (single cell consensus clustering) method was used to cluster the pre-DC and mini-bulk mature cells into 7 mini-clusters of similar expression shown in Figure 5.19. Most of the resulting clusters contained pre-cDC2 and early pre-cDC2-like cells, but a few clusters contained primarily other cell types. The most mature-like pre-cDC2s were found in cluster 1 with the mature cDC2 samples. Cluster 6 contained the CD34+ mini-bulk sample along with the majority of the CD123med cells. Cluster 5 contained the mature pDC mini-bulk sample and the majority of the pre-pDC samples, suggesting a pDC-like cluster was present In the pre-DC population. The remaining clusters 2-4 and cluster 7 were composed of pre-cDC2 and early pre-cDC2 cells along with naïve tri-lineage cells, suggestive of varying degrees of pre-cDC maturity across these clusters, some of which were more cDC2-like, such as cluster 2 and 3 and other more mixed clusters.

5.3.7 Comparison of Pre-DCs and Mature Cell Marker Expression

Further to the SC3 clustering using the mature cell markers defined in Chapter 3, boxplots were drawn to investigate expression differences for each cell type signature for all of the pre-DC cells. The number of genes expressed across each of the mature DC signatures were displayed in figures 5.20 – 5.22. In each case, a cell had to have at least 5 reads in each marker gene to be classed as a 'positive' marker.

Figure 5.20 displayed the expression of cDC1 signature genes across the pre-DC single cell dataset. Overall, no pre-DC subset was enriched for these mature cDC1 genes. The CD123medium cells (which were most enriched for cDC1 potential in culture) did not show enrichment of typical cDC1 markers, although the low range of markers was higher in the CD123medium population than any of the other pre-DC populations.

Figure 5.21 displays the enrichment for the cDC2 signature genes amongst the pre=DC subsets. Here, pre-cDC2 and early pre-cDC2 subsets expressed the most mature cDC2 signature genes, with median expression in the pre-cDC2 population at 46 and the early pre-cDC2 population at 44, with a number of high-expressing 'outliers' that were likely more mature pre-cDC2 cells. The CD123med cells expressed the lowest median number of cDC2 marker genes, followed by the pre-pDC subpopulation.

Figure 5.22 shows a strong positive pDC signature amongst the pre-pDC sub-population with 90 positive marker genes expressed. This was markedly higher than the medium expression of the other sub-populations for pDC signature genes. The pre-cDC2, early pre-cDC2 and tri-lineage cells, which were high expressers of cDC2 signature genes, expressed the fewest pDC signature genes, suggesting an inherent bias of these subsets towards maturation into cDC2-like cells, while the pre-pDC sub-population appeared to indicate a pDC developmental bias.

Although they are enriched for cDC1 potential in culture and the highest expressors of transcription factors known to be critical for cDC1 development in mouse, the CD123med, CD11c- cells did not express a mature cDC1-like signature at the single cell level. A subset of pre-DCs identified by *Villani et al* via single cell RNAseq identified a CD100hi, CD34med cell type that appeared to share similar expression patterns to the CD123med subpopulation (Villani et al., 2017). Of the 11 features identified as signature genes for the Villani subset, nine were expressed in the pre-DC dataset generated for this thesis and applied across the pre-DC subsets, displaying enrichment in the CD123med cells in Figure 5.23 with a median of 5.4 genes per cell compared to the median of 2 genes per cell in the other cell sub-populations.

5.3.8 In-vitro Development Assay for pre-DC Populations

To further support the discovery of 'primed' cells in the pre-DC gate, cells sorted according to the same cytometry gates were cultured on OP9 feeder cells with SCF, FLT3 and GM-CSF as described in section 5.2.7 by Dr Urszula Cytlak-Chaudhuri. After 14 days in culture, cell output was analysed by flow cytometry, with the lineage-, HLA-DR+ cells assigned to mature monocyte and DC populations. The proportion of mature monocyte and DC populations generated from pre-DC populations are displayed in Figure 5.24 and overlaid onto the early-priming model of haematopoiesis as pie-charts in Figure 5.25.

CD34+ progenitor cell culture resulted in a mixed population of cDCs, pDCs and monocyte cells, as would be expected from an early progenitor cell type.

Tri-lineage cells developed into an equal distribution of CD14+1c+ cells, cDC2 cells and pDCs, suggestive of a heterogeneous population of mixed-lineage cells in this gate.

The pDC population produced a majority of pDC cells after 14 days with 70% of the culture output falling into the pDC gate. Minor populations of CD1c+ cDC2 and CD14+1c+ cells were also present, although no CD14+1c- monocytes or CD141+ cDC1 cells were produced.

Early pre-cDC2 (collected from bone marrow) culture output was over 80% CD1c+ DCs, with a minor (2-3%) population of pDC and CD141+ DC cells, while peripheral blood pre-cDC2 cell output was a mix of cDC2 and CD14+CD1c+ cell types with a median output at 20% of each. Again, no monocytes were found.

The *in-vitro* production of cDC2 cells in tri-lineage, early pre-cDC2 and pre-cDC2 populations was reflected in the mixed cell clusters of SC3, which may have represented earlier, lowly primed pre-cDC2 cells.

CD123med culture output demonstrated that these cells were CD141+ cDC1 primed, with a median of 40% of the culture output from CD123med cells falling into the cDC1 gate.

5.4 DISCUSSION

Single cell analysis of dendritic cell populations has already uncovered greater heterogeneity than observed under bulk-RNA sequencing and flow cytometry alone (Villani et al., 2017). Single cell resolution of the whole transcriptome has allowed researchers to investigate new populations and sub-populations of immune cells with enough capacity and coverage to subsequently identify protein markers unique to those subsets, which can in turn be used to isolate, propagate and functionally analyse potentially novel immune cell types. Care must be taken to integrate populations identified through transcriptomics with those identified through previous phenotypic and functional analysis in order to validate any conclusions drawn.

By taking cells from a single pre-DC sort gate and sequencing their transcriptome, true heterogeneity within the gate could be assessed, which after quality control, normalisation and expression analysis, indicated sub-populations within the pre-DC sort gate had gene expression profiles skewed towards specific mature cell types including cDCs and pDCs. This skewing was confirmed during culture output analysis of the same pre-DC populations highlighting the heterogeneity of pre-DCs and committed nature of pre-DC sub-populations.

This single-cell approach to pre-DC transcriptomics was useful in demonstrating that lineage-specific enrichment in pre-DC subsets identified through *in-vitro* culture analysis were also identifiable at the transcriptomic level in individual cells. Furthermore, this methodology identified heterogeneity within the ‘trilineage gate’ that would not have been identifiable through bulk RNA-sequencing analysis of this gate. By scRNAseq, the ‘trilineage gate’ may represent a phenotypic artifact as lineage bias was identified by transcriptomics within individual cells.

5.4.1 Pre-Processing and Quality Control in Single-Cell RNA Sequencing is Dependent on the Research Question

In contrast to bulk RNA-Sequencing, 'gold standard' pre-processing and quality control pipelines have not yet been established within the single cell research community, due to it being a novel technique with a diverse range of applications. For this experiment, the Smart-Seq2 protocol was chosen for the initial sample processing as it provided high sensitivity and has recently been shown to result in detection of twice as many genes as the Drop-Seq method after controlling for cell number and sequencing depth (Ziegenhain et al., 2017). The Smart-Seq2 plate-based method was the most suited to the experimental question as the cells of interest could be isolated based on their cell surface markers and indexed so that paired information from FACS and single cell sequencing could be investigated in unison.

Initial QC of the cDNA reads included a read trimming step to ensure only high-quality reads were included in the mapping stage. A quality score of Q10 was selected as the cut off for the FASTQ files, which equates to an error probability of 0.1. This was combined with a minimum read length filter of 60 base-pairs. Over-trimming and inclusion of short read-length sequences has been shown to dramatically alter the RNA-Seq expression estimates, particularly in genes with a high GC content or low exon number (Williams et al., 2016). Read quality decreases towards the 3' end of a read and are more frequently trimmed, however smaller reads map to more regions of the genome, thus modest trimming at Q10 combined with a minimum length filter provided the optimum compromise between quality and read length (Conesa et al., 2016).

The STAR alignment tool was selected for alignment due to its performance speed when mapping to annotated genomes (Dobin et al., 2013). As discovery of novel transcripts and intron mapping were outside of the scope of this research, multi-mapped and short-sequence reads were not retained for the feature counting step performed using HTSeq (Anders et al., 2015). The main aim of the chapter, Pre-DC cell type identification, was reliant upon accurate gene-level annotation and so the most recent release of the human genome assembly available at the time of writing was used as the reference.

5.4.2 Normalisation of the Single-Cell Data Revealed Cellular Diversity Within the Pre-DC Gate

High levels of both technical and biological noise are inherent to single cell RNA-Sequencing. Combined with a reduced number of expressed features per sample compared to bulk sequencing, distinguishing technical noise from biological observations is challenging (Ramsköld et al., 2012).

The earliest stages of QC were designed to remove noise in the form of poor quality cells. Without removing these cells, clustering and visualisations of the data were overwhelmed by the number of features detected, rather than any biological gene-level differences. By removing these cells from further analysis biological information may be lost as some specialised cell types may have an overall low expression of genes, more frequently however, low expression is attributed to technical errors in sample preparation and sequencing. As the unfiltered data were observed to cluster based on total features, obscuring potential biological differences, removal of these cells was necessary.

Normalisation of data is another fundamental process in RNA-Seq analysis, but is often overlooked. With single-cell sequencing projects increasing in complexity, robust normalisation methods must account for many sources of variation that can otherwise impact read counts. The normalisation process itself has been shown to have a significant impact on the calling of differentially expressed genes (Bullard et al., 2010). With this in mind, RUV normalisation was selected as the most appropriate form of normalisation for this experiment, designed for single cell and bulk RNA-Seq analysis and tested against qRT-PCR, producing the closest correlation compared to other common normalisation methods including upper-quartile normalisation in two different external datasets (Risso et al., 2014).

The high variability within the ERCC spike-in controls and percentage of ERCC reads per cell was also noticed in other datasets by Risso *et al*. RUVg normalisation to reduce the ERCC variance, combined with log transformation of the data revealed the biological traits of each cell, and although the clustering was subtle across PCA and t-SNE space, in part due to the similarity of all cells within the pre-DC gate, some cell-type associations could be discerned according to the single cell flow gates downstream of the pre-DC gate.

5.4.3 Global Expression in Pre-DCs is Largely Conserved, But Deeper Analysis Suggests Early Skewing of Expression Patterns Towards Distinct Mature Dendritic Cell Populations

The combination of single cell RNA-sequencing, FACS and cell culture provided evidence of early priming of pre-DCs towards major pDC and cDC cell subsets from within the HLA-DR+, Lineage- (CD3, CD19, CD20, CD56), CD14-, CD123+, CD303/CD304 variable gate.

The analyses of the normalised pre-DC data were not overshadowed by the effect of differing 'total features' as the non-normalised dataset was. Initial visualisation of the normalised Pre-DC gate single cell data revealed subtle population heterogeneity, despite low overall variance within the samples. This low variance is common in single-cell experiments due to the high number of features, typically in the range of 10,000-20,000 genes, combined with a high degree of dropouts (zero-values) in each cell producing a consistent majority background signal and small number of variable genes. The same observation was noted upon analysis by M3Drop, the Michaelis-Menten-based single-cell feature selection module. No features were observed to have high variance after FDR adjustment by this method, although SC3 consensus clustering, utilizing variable PCA loadings and k-means clustering was able to group the cells into populations skewed towards cDC and pDC expression patterns. The mapping of cells by PCA and t-SNE methods highlighted that the majority of the pre-DC cells were pre-cDC2 and early pre-cDC2 cells signifying that most of the pre-DC samples were cDC2-like with cDC2 priming. In support of this, the pre-cDC2s and early pre-cDC2s clustered around the mature cDC2 mini-bulk populations.

By SC3 analysis, the distinction of pre-pDCs from cDC2s was made apparent. The majority of the pre-pDC cells clustered with the mature pDC bulk sample, suggesting a high degree of correlation between these cells. The genes defining this pDC-like cluster included SCAMP5, TMP2 and MZB1 all of which were assigned as pDC markers from the Illumina expression analysis in chapter 3. NCF2, SLC2A3 and CASP4 were identified amongst gene markers of cluster 1, which were noted as cDC2 markers in chapter 3, implying pre-DC priming towards individual cDC and pDC populations.

The lack of pre-cDC1 like cells with mature cDC1 marker expression may be attributable to their scarcity in peripheral blood populations as only 96 cells were sorted in total. It is possible that no late stage pre-cDC1 cells were collected because of this low cell number, alternatively the cell markers may not have been extensive enough to allow for the detection of low cDC1 signature expressing cells. Another possibility is that the pre-cDC1 populations do not fall into the pre-DC gate defined by the FACS gating strategy employed, although later analysis of *in-vitro* development suggests that pre-cDC1s were collected in the form of the immature (CD34med) CD123med, CD5- cells.

The mixed cell clusters defined by SC3 represented cells of various stages of development, mostly pre-cDC2 and early pre-cDC2 cells with greater or lesser expression of mature cDC2 signatures.

SC3 clustering of cells and individual analysis of the cluster markers provided evidence of pre-DC priming. This was further supported upon comparison of the pre-DC single cells to the full mature DC signatures generated in chapter 3. The observation of no cDC1-like cells by SC3 was recapitulated in the signature boxplots as each pre-DC subset expressed a largely similar number of mature cDC1 genes. This analysis provided a deeper interrogation of the gene expression of pre-DC populations as it avoided the possibility of heterogeneous cells captured in some of the SC3 clusters that would otherwise obscure marker gene selection in the mixed cell clusters.

cDC1 marker gene expression analysis did not indicate a particular pre-DC subset was enriched for the mature cDC1 signature as noted by the lack of cDC1 markers found by SC3 clustering. The inclusion of CD100+,CD34med signatures revealed the potential of CD123med cells to be early pre-cDC1s as defined by Villani *et al* (Villani et al., 2017). The Villani CD100hi population was shown to produce some CD1c+ cDCs along with CLEC9A+ cDCs through *in-vitro* differentiation assays and thus, both CD100hi, CD34med cells and the CD123med, CD11c- cells could have potential as very early pre-cDC1 precursors. Genes for this signature included ID2 and KIT, which are known to drive cDC1 development and maturation as displayed in figure 1.9, as well as CCR7, a migration signal found to be expressed highly on mature cDC1. The cDC2 signature was enriched for in the cDC2 and pre-cDC2 subsets as well as in the tri-lineage populations, suggesting cDC2 priming of cells may occur in all of these populations. This was also noted in SC3 analysis in the mixed cell clusters, which were composed mostly of these cell subsets. As in previous analyses in this thesis, the pDC signature appeared to be extremely strong in pre-pDC cells. By SC3, the pDC and pre-pDC cluster had a high stability index and expressed pDC genes that were not expressed in any of the other clusters. This strong bias may be explained by the exclusive interactions of transcription factors during lineage priming, driving cells towards a pDC or cDC cell type. IRF8 and E2.2 act to enforce and reinforce pDC lineage bias.

With the inclusion of *in-vitro* development data supplied by Dr Urszula Cytlak-Chaudhuri, conclusions drawn from the single cell analysis were supported with further substantial evidence. The data suggests that pre-DCs are a heterogeneous mix of committed progenitor cells, distinct from CD34+ progenitors and potentially, missing monocyte potential, although culture conditions were designed to discourage monocyte differentiation, and thus true monocyte potential could not be tested in this model. The pre-DC populations exhibited a varied capacity to produce DCs and were enriched towards one or more mature cell types. While pDC and cDC2 cells can be collected simply from peripheral blood in sufficient numbers to be used in subsequent research or potential clinical therapeutics, cDC1 cells are frequently noted to be too difficult to obtain in numbers conducive to research.

A median of 40% of the culture output from CD123med cells were cDC1s providing a possible basis for the generation of great numbers of cDC1s for research or cell therapy. This work has produced viable evidence for a source of *in-vitro* generated, *bona-fide* cDC1s from CD5-, CD123medium pre-DC cells, and shown that the DC lineage at the pre-DC stage is exclusive from monocyte development. In turn, work supports the production of steady-state DCs from sources other than monocytes as monocyte-derived DCs exhibit an expression pattern closer to inflammatory DCs rather than steady-state DCs and such moDCs may not be functionally and developmentally comparable. Despite low cell numbers and the inherent difficulty in analyzing such large-scale datasets, this single cell analysis has revealed the heterogeneity within the pre-DC sort gate that would not have been possible through bulk transcriptomics or narrower-scope analysis techniques. The combination of novel single-cell analysis, backed by traditional techniques in cell culture and flow cytometry produce a solid research pipeline that has become a common feature in the field of single-cell immunology and provided new insights in haematopoiesis, DC development and lineage commitment, as well as supporting the ‘early-priming’ model of haematopoiesis.

5.5 RESEARCH SUMMARY AND KEY POINTS FOR PROJECT PROGRESSION

Chapter 5 combined single-cell transcriptomics with knowledge gained from the previous two chapters to investigate the development and priming of early DC precursors. 'Early priming' models of haematopoiesis, infer that most early progenitor populations, including the granulocyte macrophage dendritic cell precursor (GMDP), macrophage DC progenitor (MDP), and common DC progenitor (CDP) populations are composed of heterogeneous, uni-primed, phenotypically similar cells with restricted mature cell potential. Determining if these progenitor cells were early primed or multi-potent was a critical aspect of the chapter and haematopoietic development research as a whole.

Utilizing plate-based single-cell sequencing revealed the diversity and heterogeneity within populations of precursor cells that would have been otherwise obscured using bulk population techniques such as qPCR, NanoString Technology, microarray analysis or conventional bulk-RNA sequencing, where 'average' expression of the population is observed, whether or not this average expression pattern is a true reflection of the population, or merely a chimeric expression pattern formed by the averaging of multiple, distinct cell types.

Initially, a FACS and SmartSeq2 based pipeline was developed to address whether adequate, good quality, viable cells could be captured by FACS and processed through a SmartSeq2 plate-based single cell pipeline to provide enough quality data for analysis. 71 samples passed initial analysis QC in total from the 96 cells sorted from the pre-DC FACS gate, expressing between 2,000 and 10,000 genes per cell.

By investigating the cell populations found within conventional progenitor cell FACS gates, their heterogeneity was revealed and expression was compared to mature cell populations to determine if there was any intra-population variance suggestive of early cell priming. Interestingly, some cells did express strong mature signatures based on the gene signature produced and refined in chapter 3, suggestive of early priming of precursor cells. This was an important initial step in uncovering the nature of early immunopoiesis and aligns well with the current priming model of haematopoiesis.

To confirm the initial findings, an *in-vitro* development assay was incorporated to isolate and grow transcriptionally primed progenitors to determine if their culture output recapitulated the transcriptional priming upon development and maturation. If primed progenitors could be isolated and grown in culture, they could be used to generate *bona-fide* DCs for research and medical applications in the future. This experiment provided supporting evidence for the early priming model, with pre-pDC populations forming pDCs, pre-cDC2 and early pre-cDC2 forming a majority population of cDC2s and the suspected pre-cDC1, CD123medium population producing cDC1s.

The ability to collect pre-DC populations and culture them into mature *bona-fide* DCs has great implications in research. Increasing the number of cells that can be collected and produced from blood is an important consideration. Scarcity of DCs in peripheral blood can make their study arduous, however from the comparison of blood and cultured cells in chapter 4, and the development of DCs from pre-DCs in chapter 5, the capacity for cDC collection and culture has been revealed for this culture model. Further impacts on the wider research community may be gleamed from the possibility that primed progenitors of various DC subsets could be isolated and grown in culture and used to generate *bona-fide* DCs for research and medical applications in the future.

Chapter 5 Figures & Tables

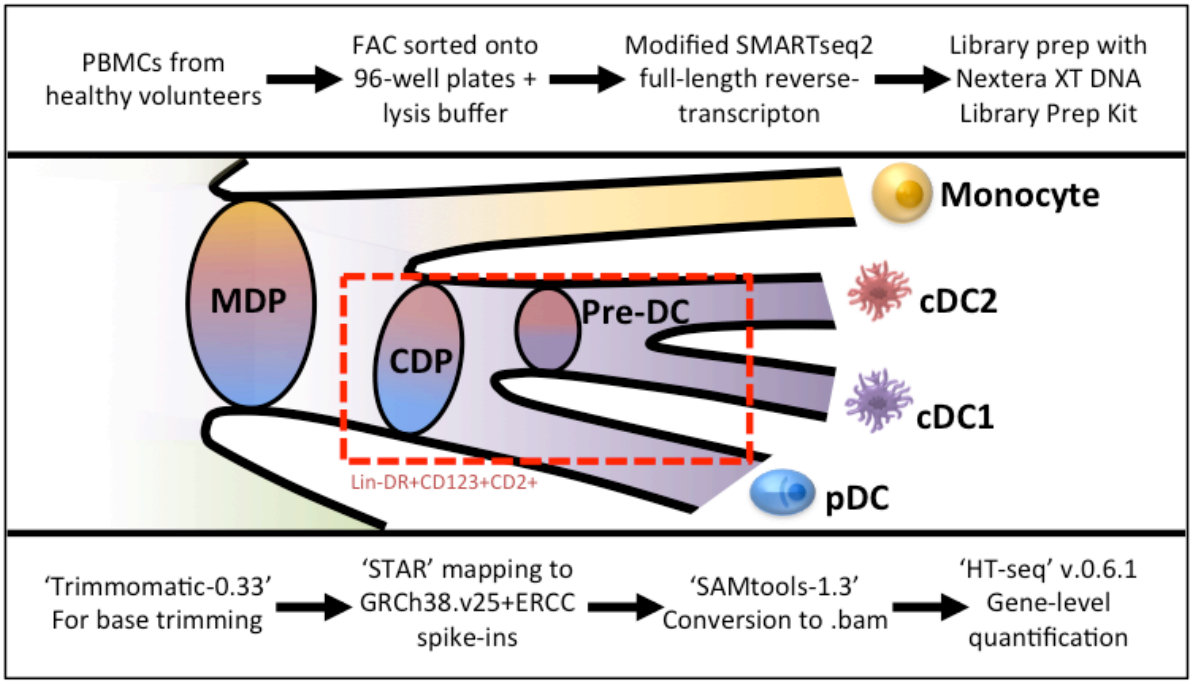


Figure 5.1: Outline for pre-DC single-cell RNA-sequencing and pre-processing

Peripheral blood mononuclear cells were sorted into 96-well plates. Cells were isolated from a Lineage-DR+ gate, predicted to cover the Common Dendritic Cell Progenitor (CDP) and pre-DC populations displayed in the figure. The cells were negative for conventional mature myeloid DC markers, CD1C and CD141, but expressed CD123 and CD2 by FACS. The full gating strategy for these cells is displayed in figure 5.2. After collection, reverse-transcription was performed according to a modified SMARTseq2 protocol and library prep was performed with Nextera XT DNA library prep kit. Sequencing was performed at Oxford Genomics on an Illumina HiSeq4000.

Trimmomatic-0.33 was used to trim poor-quality bases from reads before the reads were mapped to GRCh38.v25 genome construct with External RNA Control Consortium (ERCC) reads spiked-in. Pre-processing was finalised via gene-level quantification using HT-seq v.0.6.1.

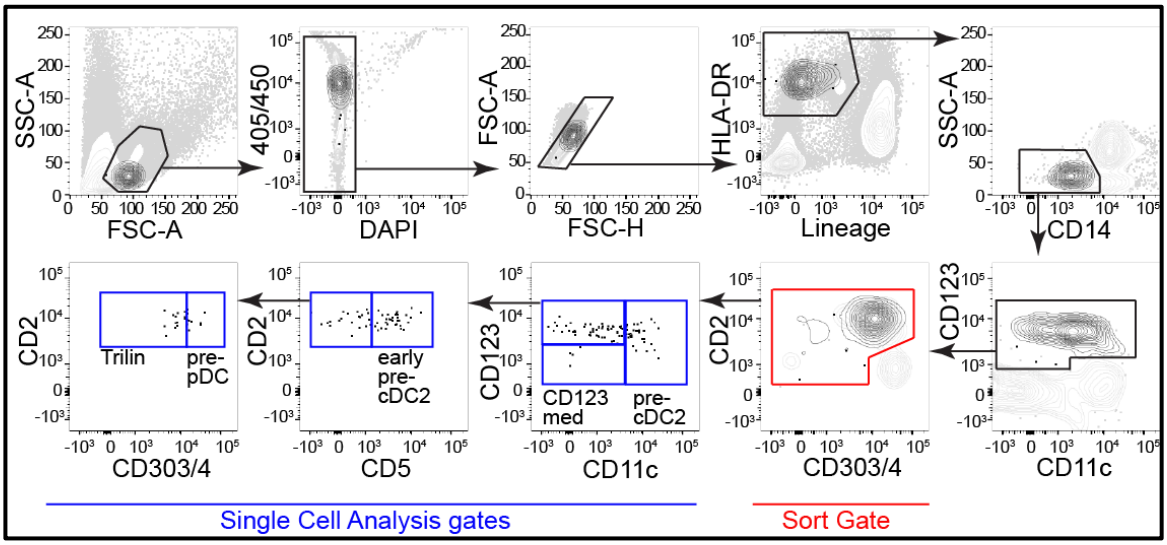


Figure 5.2: Sorting strategy for pre-DC single-cell RNA-sequencing

Single cells were collected by FACS for plate-based single-cell RNA-sequencing according to this strategy. Additionally, 6 mini-bulk (10 cells) samples were collected according to the gating strategy in Figure 2.1 for CD14+ monocytes, pDCs, cDC1, BTLA split cDC2 cells and CD34+ progenitors. For the pre-DC cells, DAPI staining was used to screen out dead or damaged cells before HLA-DR+, Lineage- (CD3, CD19, CD20 and CD56), CD14-, CD123+, CD303/304 variable cells were isolated for analysis by single-cell RNA-sequencing.

From the sort gate (indicated in red), further gates were defined as follows:

Pre-cDC2 (CD11c+),

CD123med (CD123 med),

Early pre-cDC2 (CD123+, CD5+),

Pre-pDC (CD123+, CD2+, CD5-, CD303/4+),

Trilineage (CD123+, CD2+, CD5-, CD303/4-)

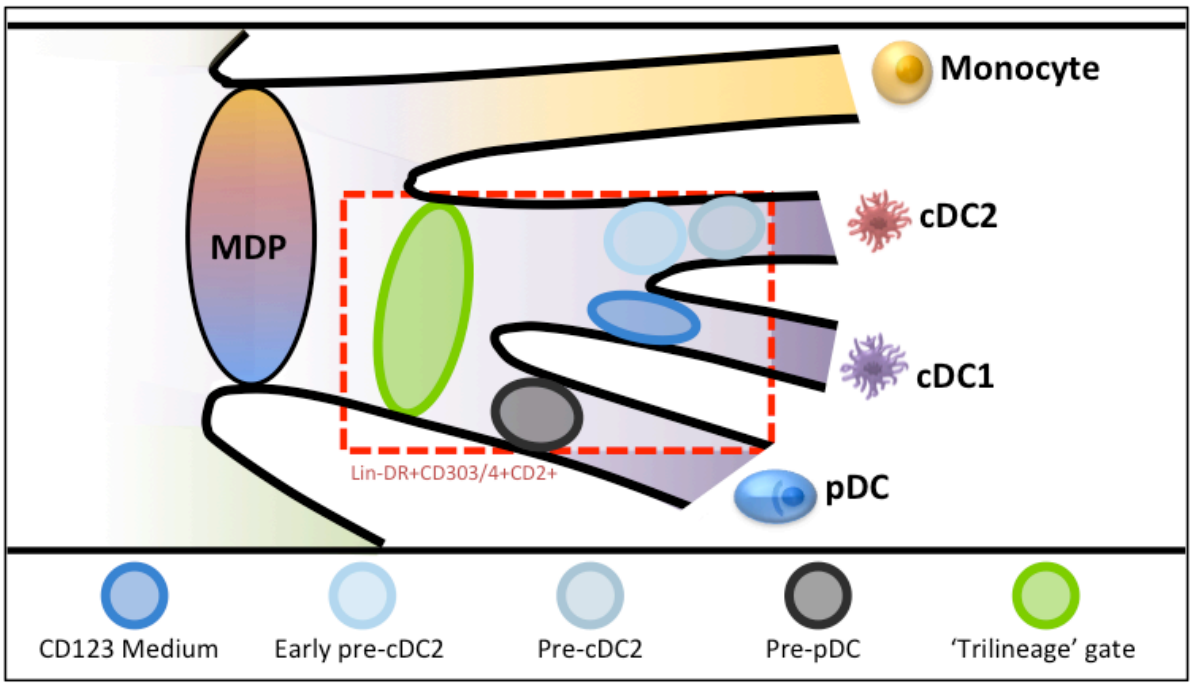


Figure 5.3: The single cell sorting gates equate to expected pre-DC populations

Pre-DC populations isolated by FACS were predicted to equate to populations falling into the highlighted regions of the haematopoietic lineage tree. The 'Trilineage' gate (green) overlaps closely with the conventional Common Dendritic Cell Progenitor (CDP) population, with the other sort gates containing suspected 'primed' pre-DC populations.

The pre-cDC2 population (grey-blue) was CD11c+, early pre-cDC2 (light-blue) was CD5+, CD2+, CD123med (predicted to be pre-cDC1) (dark-blue) was CD11c-, CD123med and the pre-pDC population was CD303/304+, CD123+, CD5low.

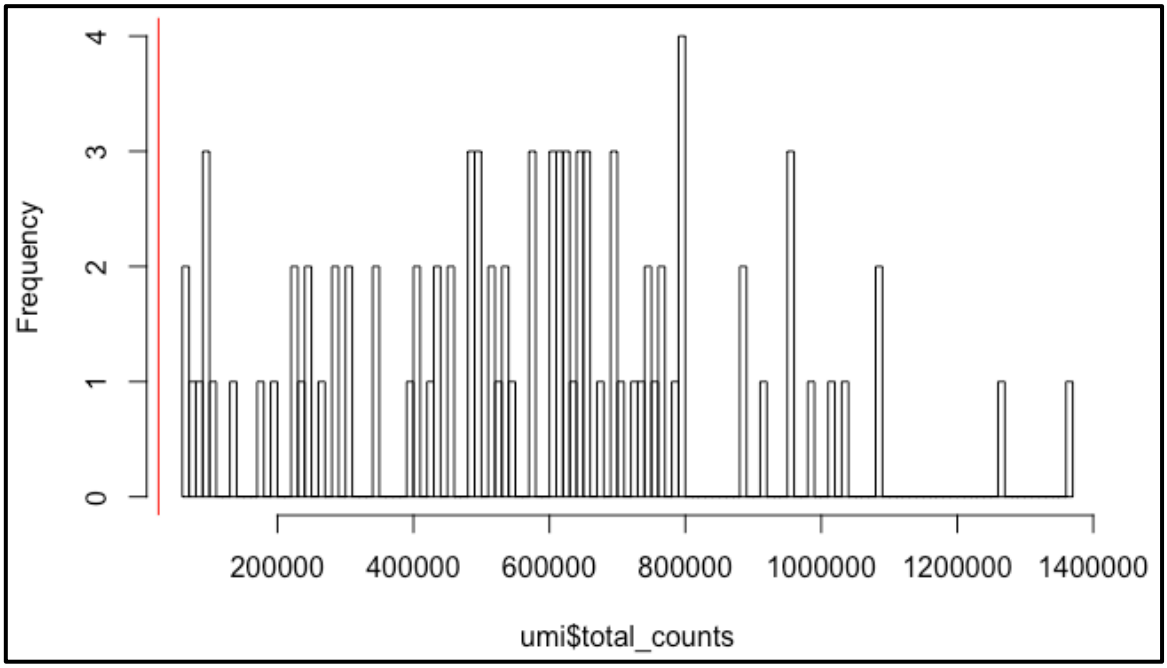


Figure 5.4: Histogram of total counts for pre-DC single cell analysis

As part of the single-cell quality control steps, any samples with less than 25,000 total reads were considered ‘failed’ and would be removed from further analysis. These cells were likely damaged, burst or did not undergo full library-prep during sample preparation. None of the samples failed this QC step. Those samples with the highest total counts were the six mature mini-bulk samples, which contained 10 cells in each well. Most of the single cells had between 200,000 and 800,000 total reads.

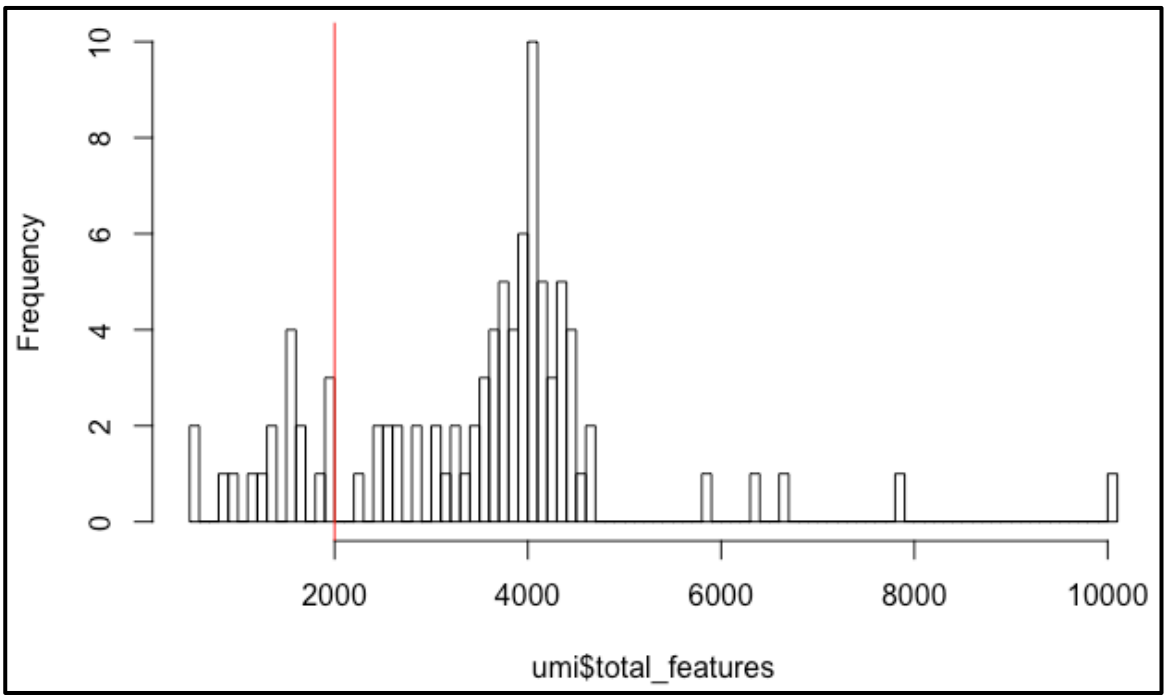


Figure 5.5: Histogram of total features (genes) for pre-DC single cell analysis

As part of the single-cell quality control steps, any cells expressing less than 2,000 total features were considered ‘failed’ and were removed from further analysis. These cells were likely damaged, burst or did not undergo full library-prep during sample preparation. Typically most of the samples with very few total features are composed mostly of mitochondrial or ERCC spike-in reads. 18 of the 92 samples failed this QC step. The samples with the greatest number of features were the mature mini-bulk cell samples, ranging from 4,000-10,000 unique genes expressed. For the single cells, approximately 4,000 was the average number of expressed features per cell.

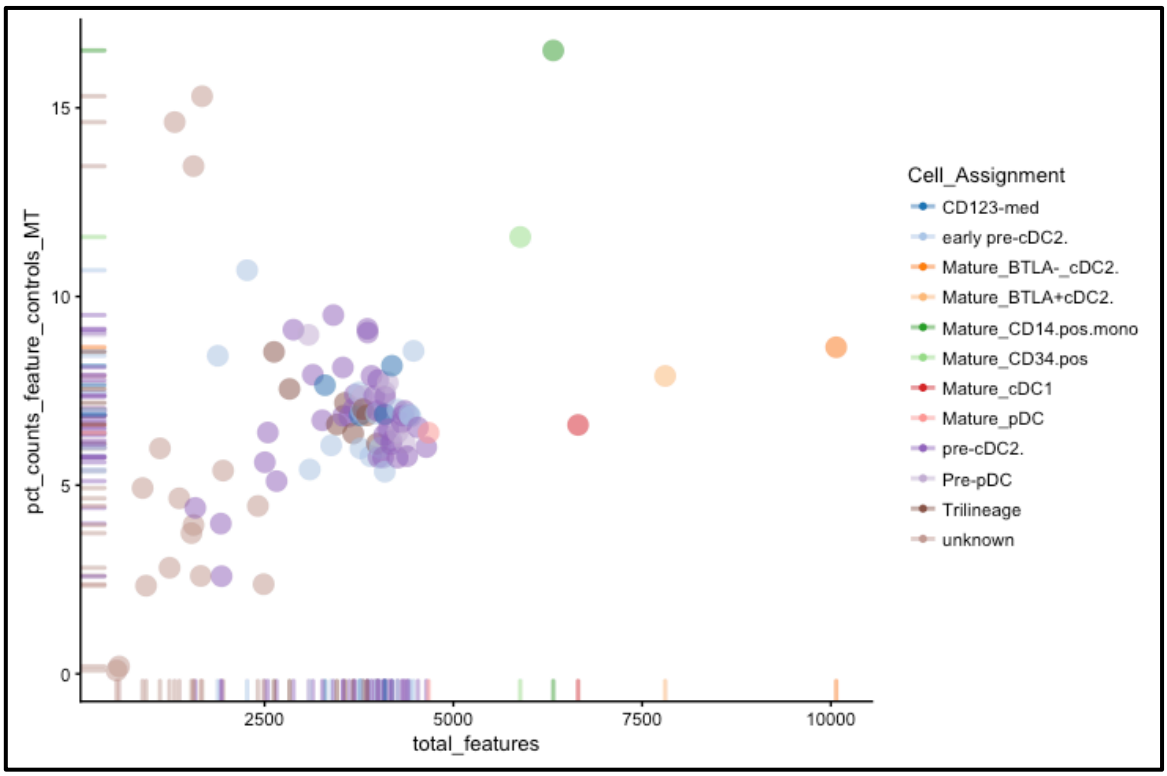


Figure 5.6: Mitochondrial gene percentage quality control analysis

Percentage of mitochondrial gene reads are used as a quality score in single cell analysis. High percentages of mitochondrial genes in the data is usually the result of cell degradation. No consensus has been reached in the single cell community with regards to a definitive cut-off value for mitochondrial reads, but lower percentages of mitochondrial reads are an indication of better data quality. For this analysis, a percentage of mitochondrial genes above 15% was used as a cell quality filter. 2 of the 92 cells failed this QC and were removed from analysis.

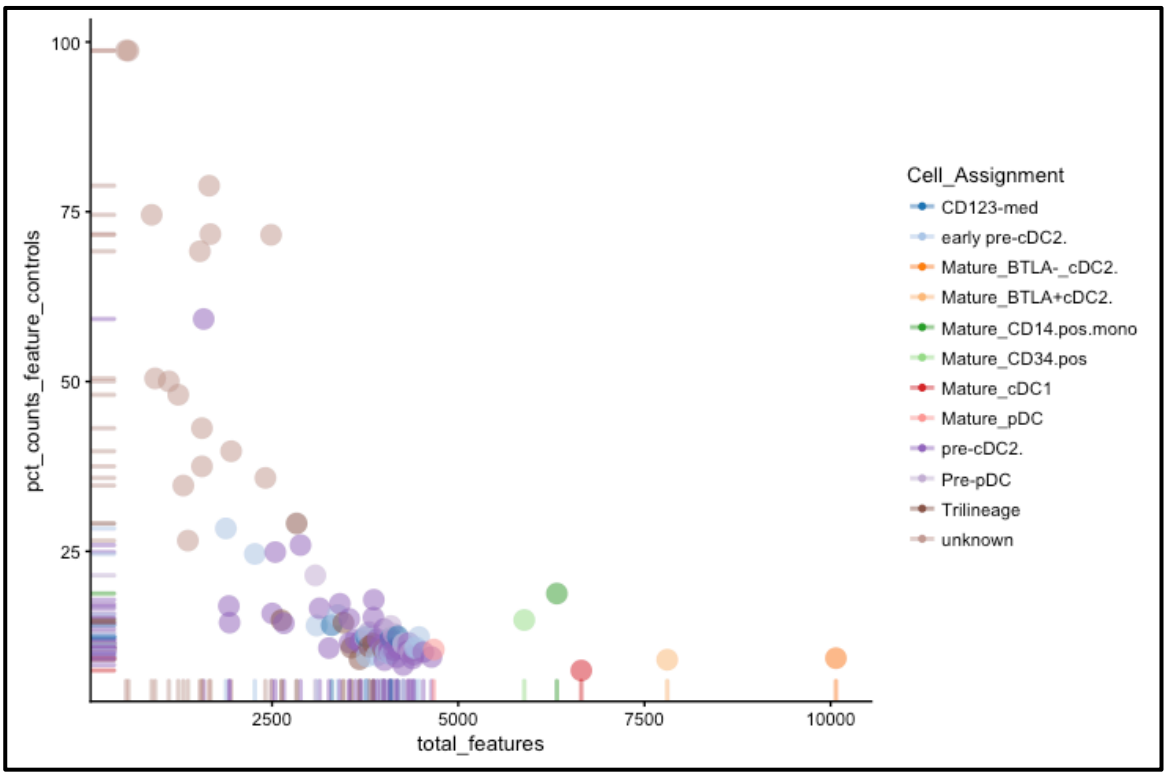


Figure 5.7: ERCC spike-in read percentage - quality control analysis

Percentage of ERCC reads are used as a quality score in single cell analysis. ERCC reads are spiked-in non-genomic targets that act as technical normalisation reads. High percentages of ERCC reads in the data is usually the result of cell degradation, missing cells or incomplete library prep. Lower percentages of ERCC reads are an indication of better data quality. For this analysis, a percentage of ERCC reads above 25% was used as a cell quality filter. 15 of the 92 cells failed this QC and were removed from analysis.

Table 5.1a: Quality control filters for cells

Number of cells removed by filters	Number of cells remaining for analysis
21	71

Table 5.1b: Quality control filters for genes

Number of genes removed by filters	Number of genes remaining for analysis
8,539	14,412

Table 5.1: Quality control filters for pre-DC single cell analysis

Quality control filters for this study involved sample filters and gene filters. Sample filters, summarised in Table 5.1a involved removing any cells with less than 25,000 total reads, less than 2,000 total mapped gene feature, more than 25% ERCC reads and greater than 15% mitochondrial gene reads. In total, 21 samples failed to pass this quality control, leaving 71 samples for analysis.

Feature filters were performed in two stages. Initially, all features with no counts in any cells were removed. This reduced the dataset from 60,675 features to 22,951. After normalisation, features were filtered again to remove any genes that were expressed in less than two cells, reducing 22,951 genes to 14,412 genes.

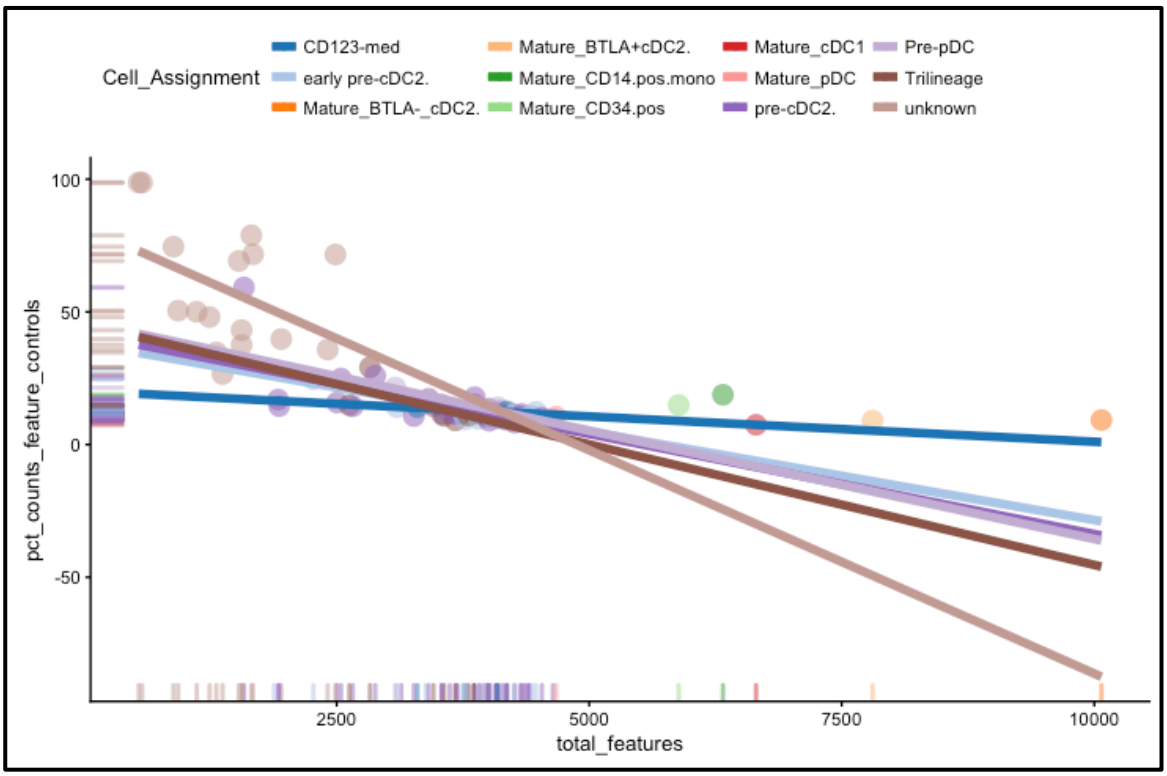


Figure 5.8: Correlation feature control plot for pre-DC dataset

For good quality cells relatively high total features are expected along with a low percentage of feature controls. In this dataset, many of the mini-bulk mature subsets have above 4,000 total features and less than 20% feature controls. Relatively horizontal lines indicate subsets with a low expression of feature controls independent of total features. The unknown/unclassified samples in this dataset have a very high expression of feature controls and relatively low expression of total features, which is the reason for their removal during QC. High percentage expression from feature controls and low total features is an indication that the sample was blank or failed during pre-processing.

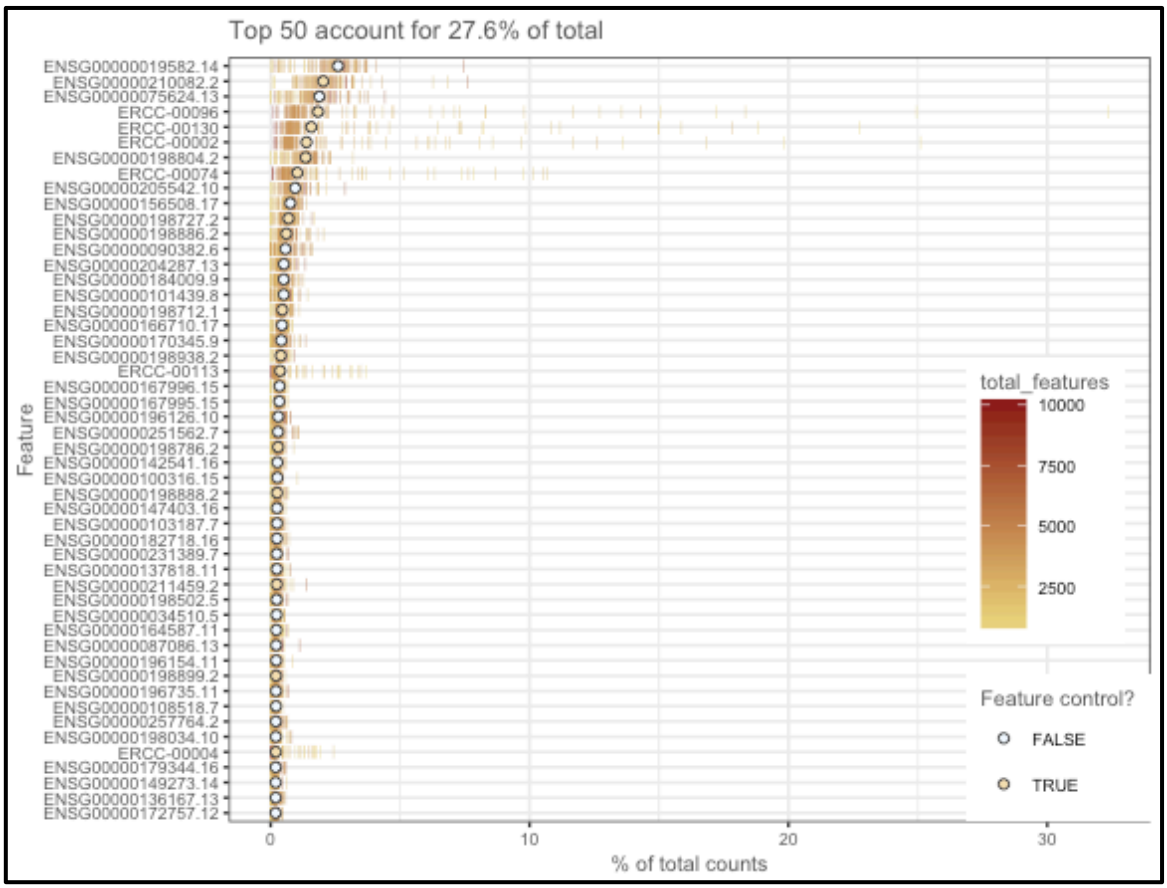


Figure 5.9: Top 50 most expressed features from pre-DC dataset reveals high expression of some feature controls

Distribution of the top 50 most expressed features is relatively narrow, indicating good coverage of the full transcriptome, however a number of ERCC controls appeared in the top 50 expressed features suggesting that future experiments may benefit from a greater dilution of the ERCC spike-ins. The ERCC features on this figure also display wider variation between samples than many endogenous genes, thus ‘ RUVseq’ normalisation based on ERCCs was used to normalise counts based on ERCC controls.

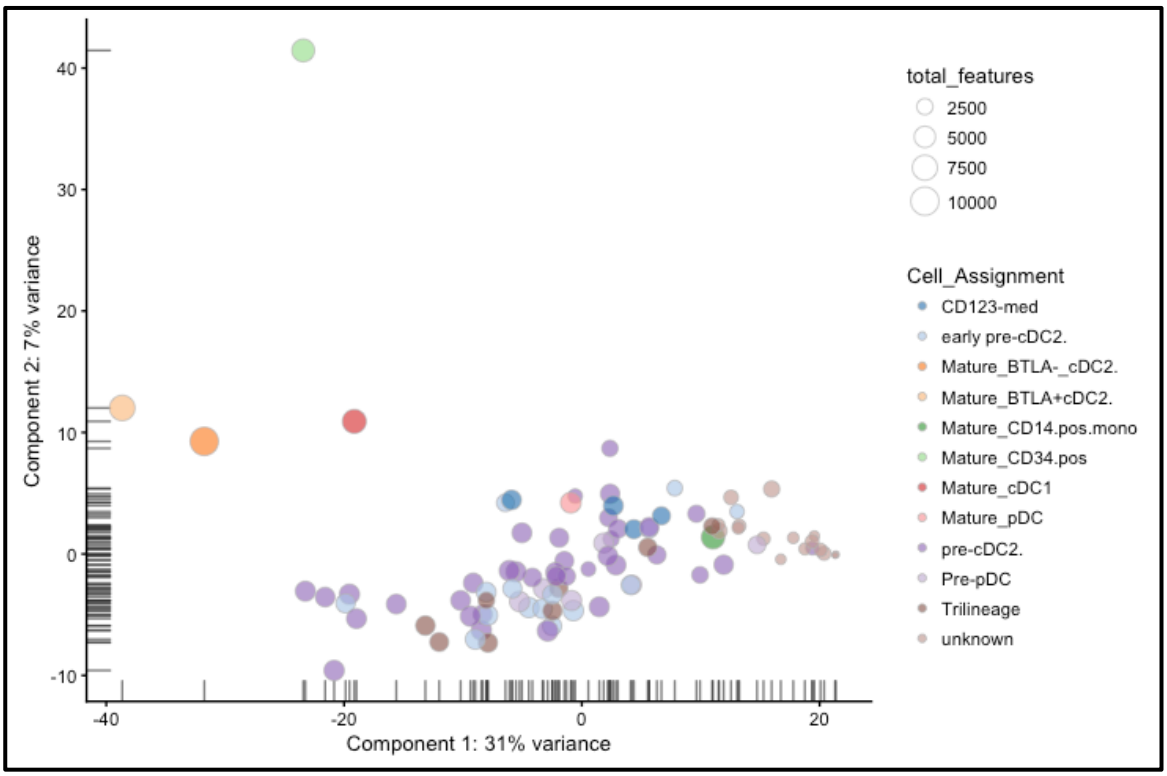


Figure 5.10: PCA of pre-log transformed single cell pre-DC data

This principal component plot was produced as a first-line look at the single cell expression data. Pre-normalisation and based on raw counts, this plot indicates total features as a major source of variance within the dataset. This is a common feature of single cell expression data.

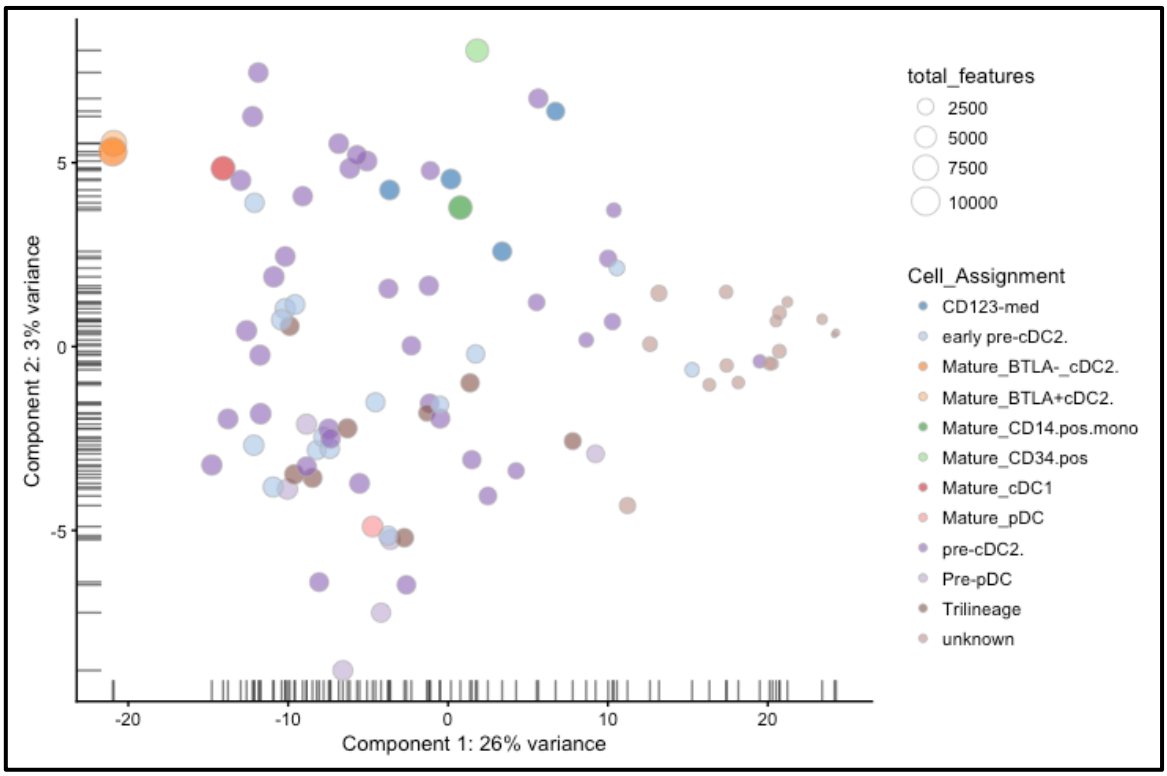


Figure 5.11: PCA of log_2 transformed single cell pre-DC data

Log transformation of the single cell expression data indicated total features were a major source of pre-normalised variation. Those samples that were later removed from the analysis at the end of the QC stages (seen here in pink as the 'unknown' group) all share the same region of PCA space and a small number of total features.

The mature/mini-bulk subsets appear on the opposite side of the plot by PC1 from the QC failure samples. These cells had the greatest number of total features and from figure 5.6, the lowest percent of ERCC controls.

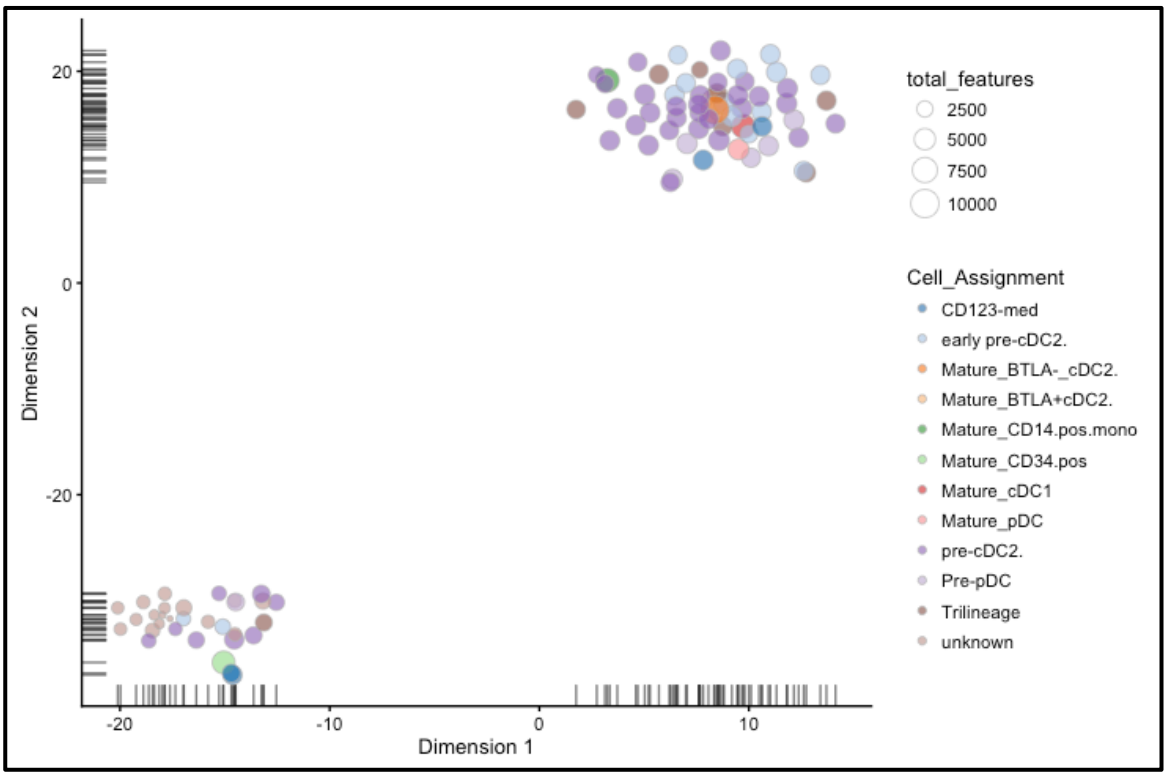


Figure 5.12: t-SNE analysis of pre-normalised pre-DC single cell data

Pre-normalised t-SNE produces two clear clusters. The main division between these clusters is the number of total features. Many of the cells that failed QC were located in the negative region of t-SNE space. From this plot, it is evident that without suitable normalisation the data would not be interpretable. The main defining feature of the dataset is the number of total features rather than biological variance.

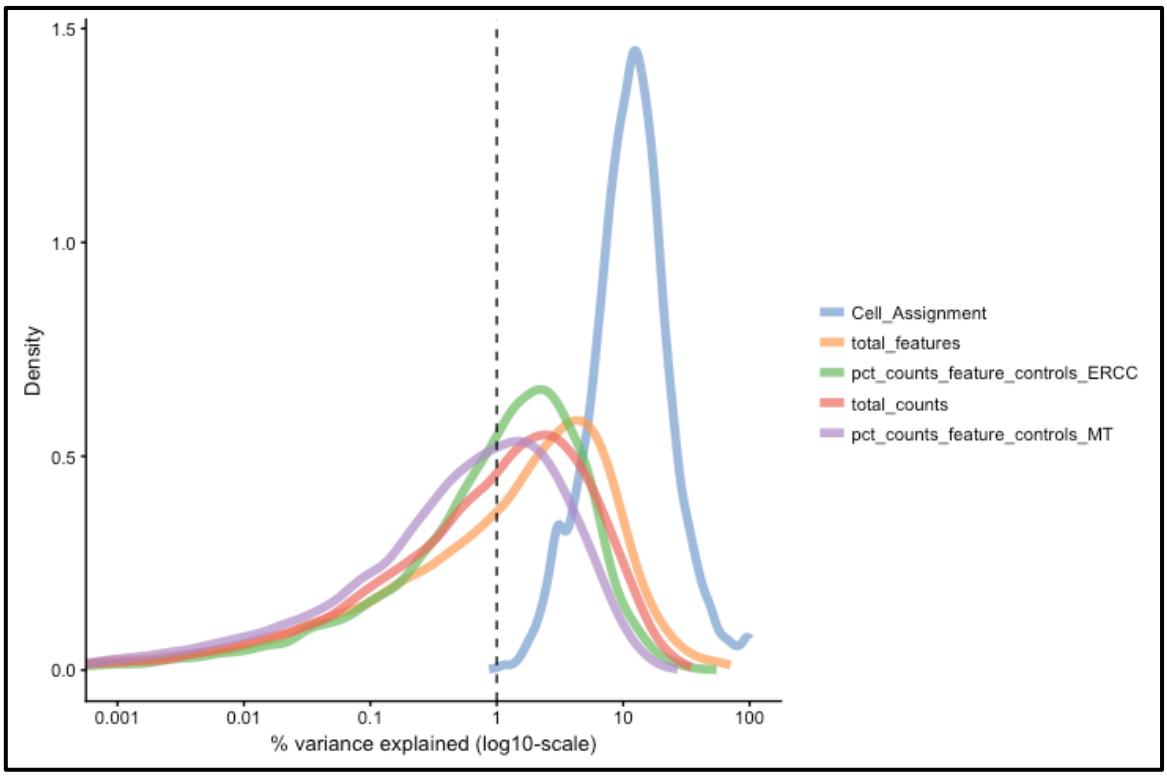


Figure 5.13: SCATER package variance plot for identification of sources of data variance

For quality scoring the experimental and technical variables in the dataset, 'SCATER' function 'plotQC()' was used. The relative importance to sample variance was plotted with a linear regression model plotted against each variable. Total features and total counts are known to cause issues with single cell data analysis. This was mitigated to some extent by converting the count data to counts per million. The SCATER package variance plot is read as higher peaks shifted to the right of the plot are variables that explain more of the data variance. The cell type designation from the flow sorting data clearly defines the most sample variance; due in part to the high expressed and feature counts of the mature mini-bulk samples. Even without these, the sample labels from the sort data that did not pass QC were left as 'unassigned', thereby grouping poor quality samples together. Of note, the percentage of ERCC controls (green line) accounts for the third greatest variance. As ERCC controls should be a control measure, 'RUVseq' function 'RUVg' was implemented to reduce ERCC variance and normalise the data to ERCC controls.

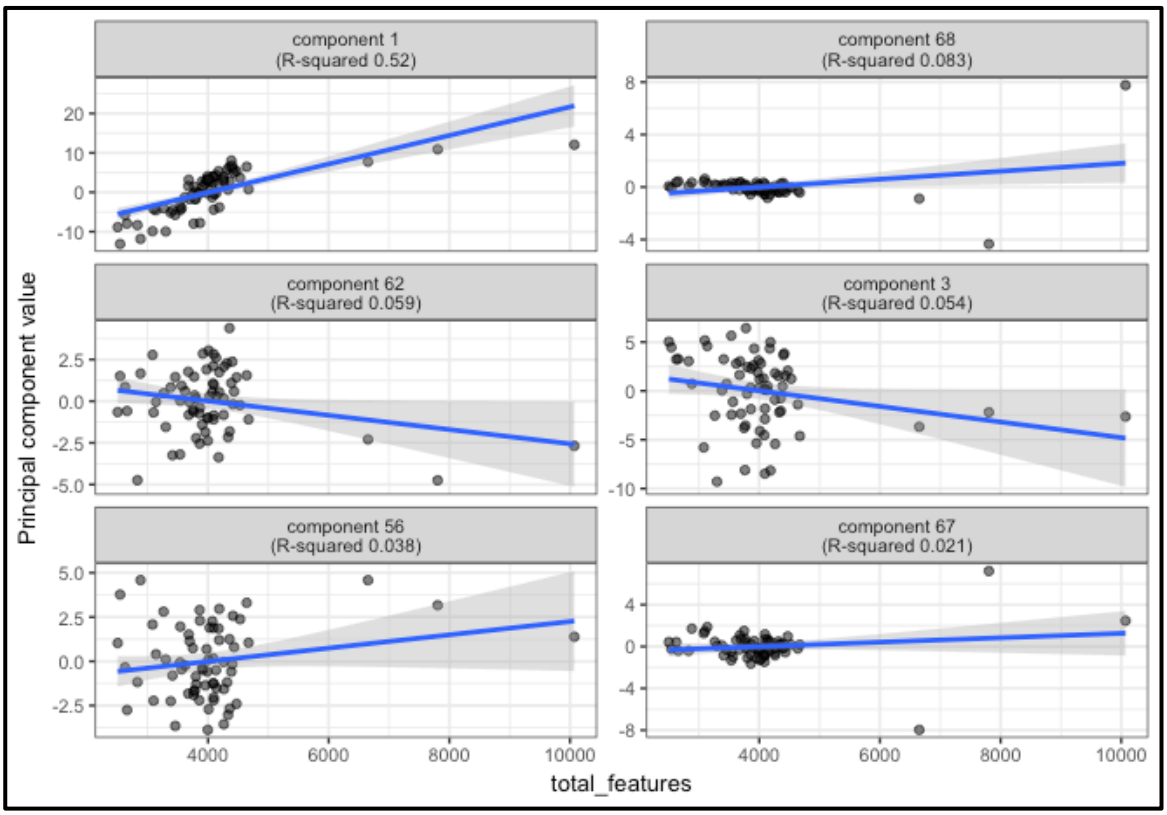


Figure 5.14: PCA breakdown analysis correlates PC1 with total features

Principal component 1 exhibits correlation with total features in each cell. This is partially due to the inclusion of the mini-bulk mature subsets containing 1,000 to 6,000 more features than the pre-DC cells. These samples are on the positive scale of PC1. Equally, cells with relatively low feature counts are located on the far negative scale by PC1. The majority of samples with approximately 4,000 features are found at the zero-point of PC1. Variance attributed to feature counts is a known artifact of single cell RNA sequencing.

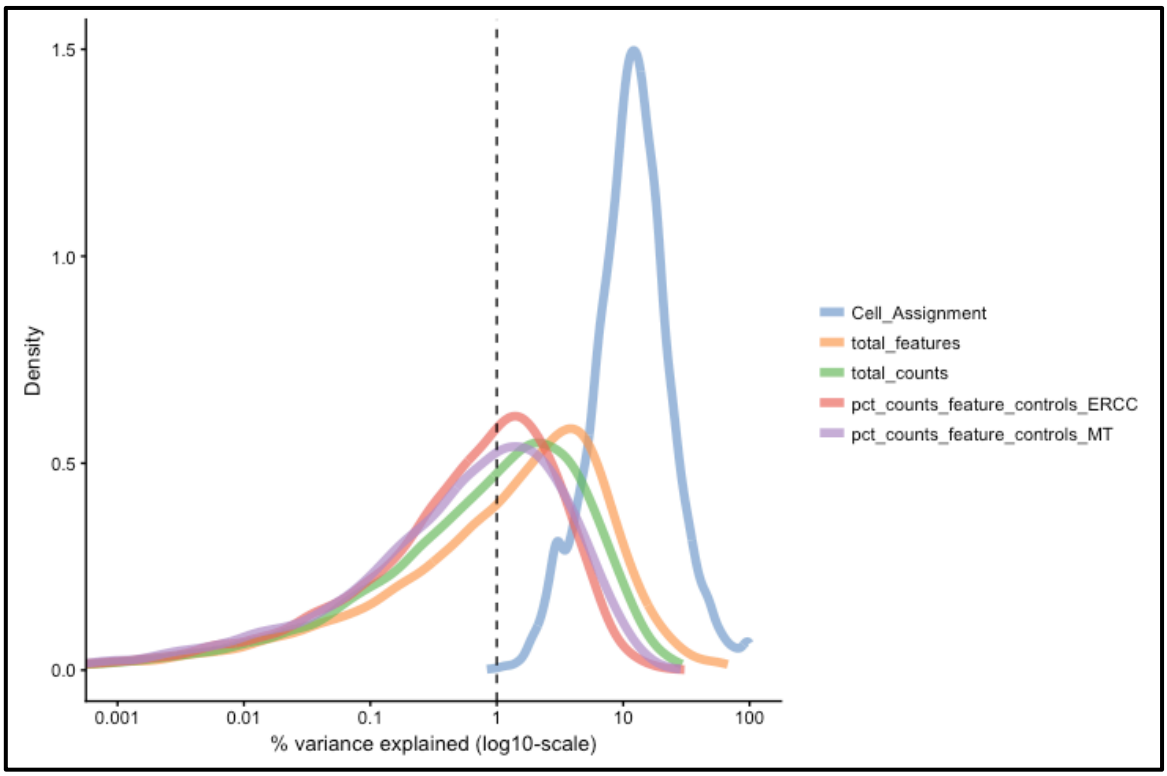


Figure 5.15: SCATER package variance plot for identification of sources of data variance after RUV normalisation for control probes

After RUVq ERCC normalisation and conversion to counts per million, variance attributable to ERCC controls has been modestly reduced to a level below that of total counts. There is still residual variance associated with both mitochondrial gene expression and ERCC spike-in variation after normalisation, but this level is acceptable and could not be further adjusted for. Cell type remains the major variable accounting for sample variance with total features being the next greatest source of variance.

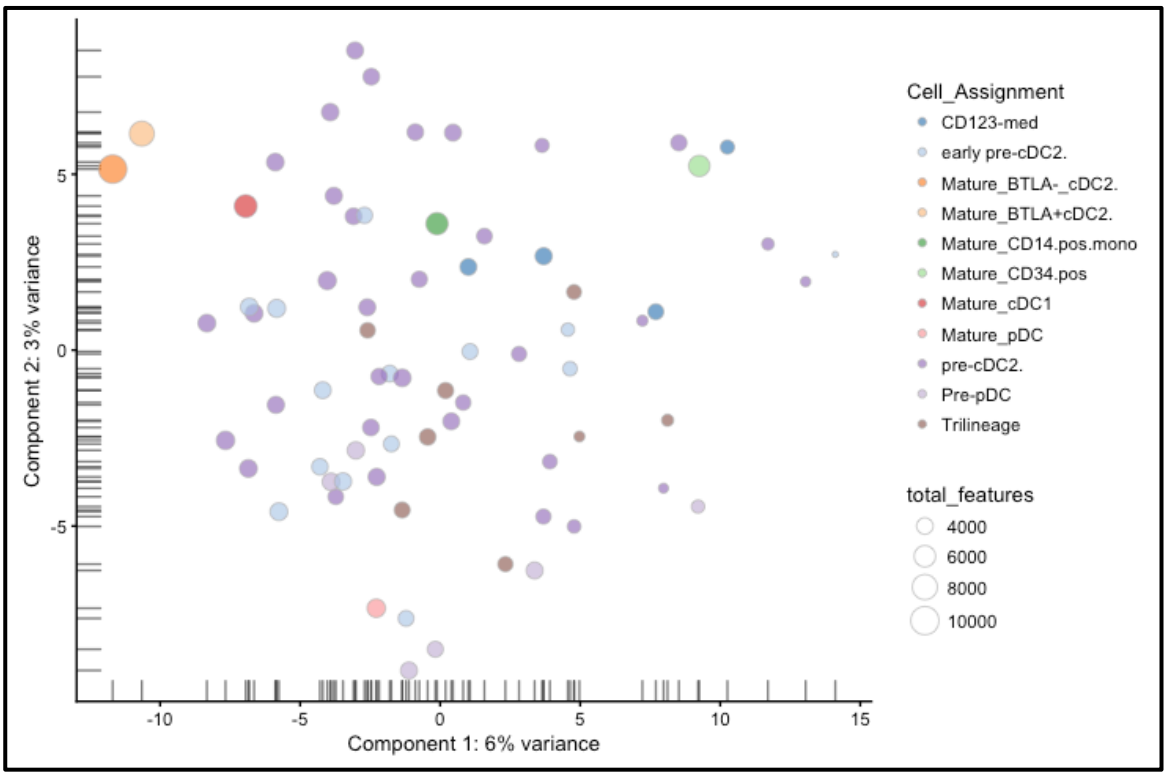


Figure 5.16: Principal component analysis of normalised pre-DC single cell data

After normalisation of the data, total features has less of an overarching effect on the PCA. While it does not appear to show strong cell type clusters due to similarity between all of the cell types analysed and feature counts in the region of 4,000 for most cells, some observations can be made. Both BTLA+ and BTLA- mature cDC2 cells are grouped together on the far-left of the plot with a loose cluster of pre-cDC2 cells in the same quadrant of the PCA. Early pre-cDC2 cells are generally located along the edge of the pre-cDC2 cluster.

The mature pDC cell is located in the negative region of PC2, with many of the CD5- pre-pDC cells nearby in the lower third of the plot. Trilineage cells are located loosely around the zero-point of PC1 and PC2, tailing towards pre-pDC cells. In the upper-right quarter of the plot in the positive region of PC1 and PC2 are the CD123med cells and the CD34+ mini-bulk sample.

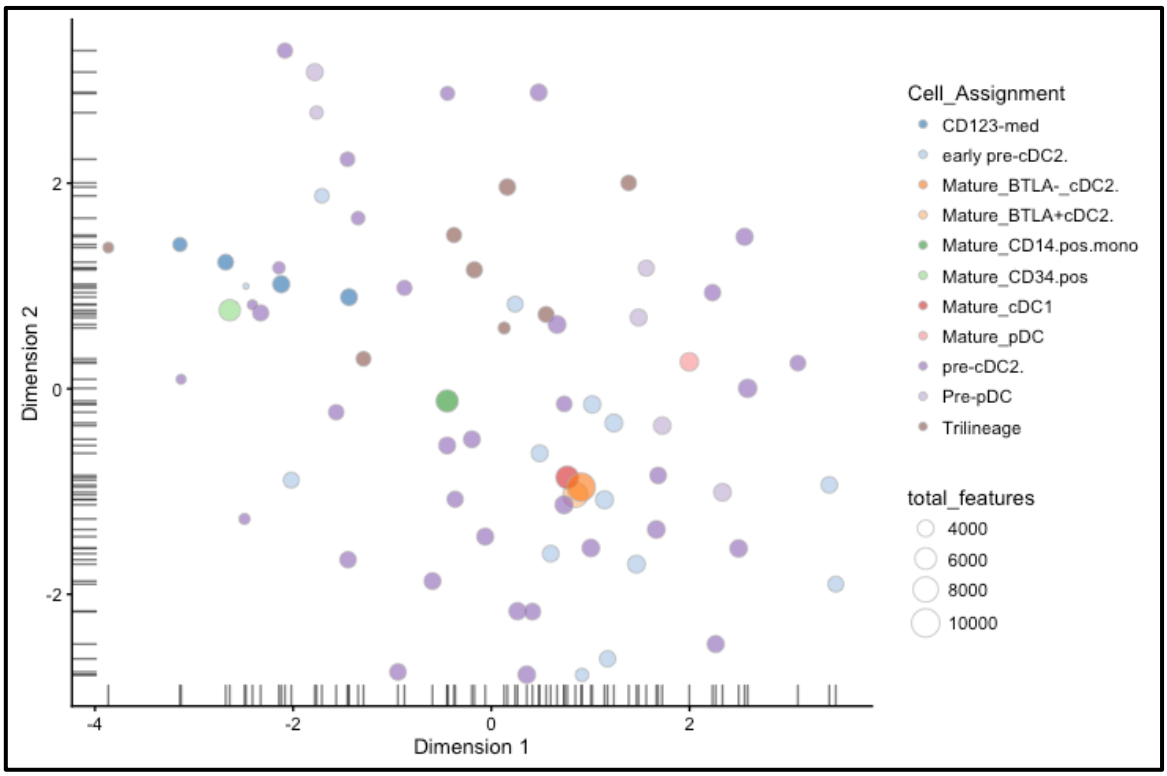


Figure 5.17: t-SNE analysis of normalised pre-DC single cell data

After normalisation of the data, some slight features can be observed in the data. As in figure 5.16, both BTLA+ and BTLA- mature cDC2 cells are located in the same region of t-SNE space, surrounded by the majority of the pre-cDC2 cells and interspersed with early pre-cDC2. In the positive region of t-SNE1 are the majority of pre-pDC cells and the mature mini-bulk pDC sample. The CD123-med cells appear to cluster around the CD34+ mini-bulk sample. The loose clustering of pre-DCs around the mini-bulk samples suggests a degree of similarity between immature pre-DC cells and one or more of the mature samples, although variation of individual genes for each pre-DC cell type may be high.

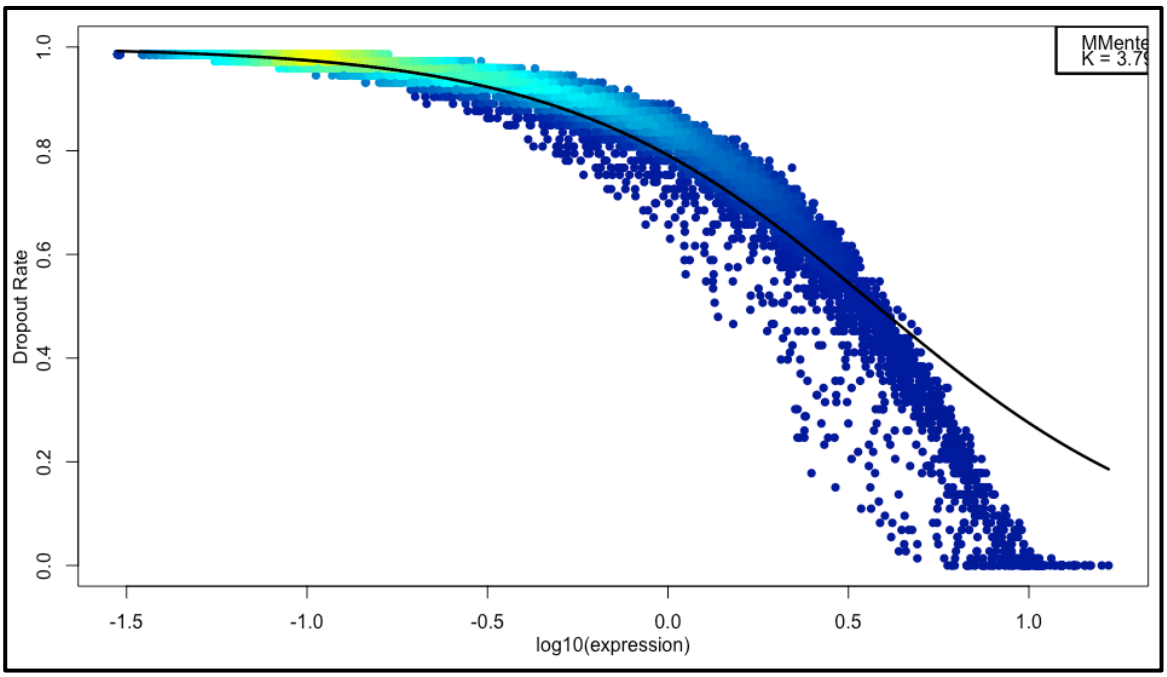


Figure 5.18: Differential gene expression based M3Drop package

M3Drop is a method of single cell RNAseq differential expression analysis that takes into account the inherent high number of dropouts in single cell expression data. Differential expression tools for bulk RNAseq are inappropriate in this instance as the zero expression values violate assumptions made in bulk RNAseq statistical models such as negative binomials. M3Drop is Michaelis-Menten based modeling of dropouts. Most dropouts occur as a result of failure to be reverse transcribed. In this figure, features shifted to the right of the Michaelis-Menton curve are identified as expressed at different levels in a subpopulation of the cells. This model was not able to provide any differential expression markers in the dataset, likely due to the samples being very similar.

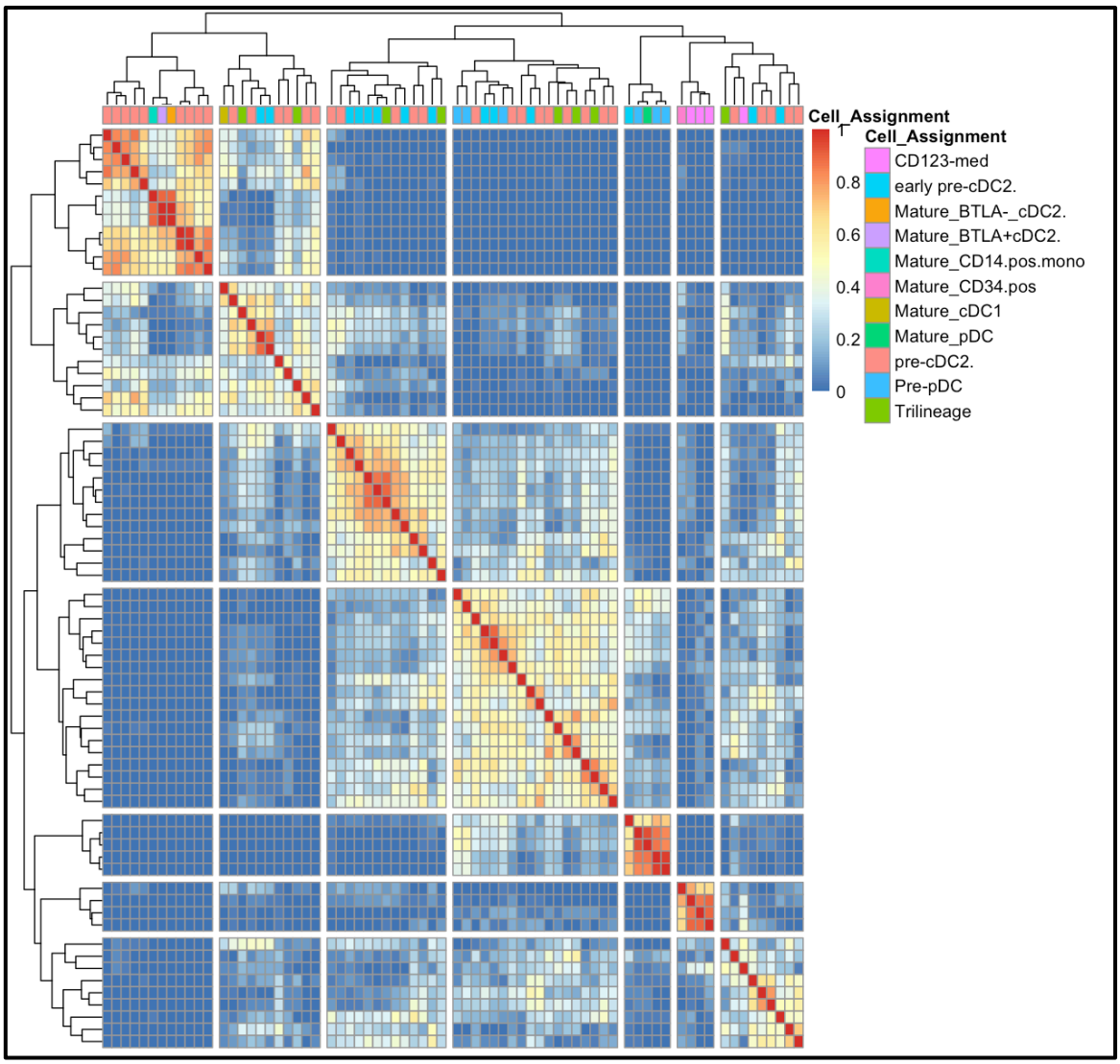


Figure 5.19: SC3 clustering of single cell data based on mature subset signatures

[Figure legend displayed on following page]

Figure 5.19: SC3 clustering of single cell data based on mature subset signatures

SC3 utilises PCA, dimensionality reduction and k-means clustering to identify patterns of similarity between cells. Additionally, the package performs consensus clustering to determine a confidence score for each defined cluster. In this figure, the 14,412 features remaining after QC were screened against the mature cell markers generated in chapter 3. 647 features were present in this dataset and were used to filter the genes for SC3. A k-value of seven was selected, producing seven clusters, three of which are of significant interest. Cluster 1, the left most cluster of the plot contains both the mature BTLA+ and BTLA- cDC2 cells as well as the most strongly cDC2-like pre-cDC2 cells. The stability index for this cluster was high at 0.78 indicating that the same cells repeatedly fell into the same cluster at 1,000 iterations. The second most stable cluster at stability index 0.60 was cluster 6, composed of the mini-bulk CD34+ sample and the majority of the CD123med cells. These all appear to have a relatively early-stage naïve expression signature. Cluster 5 is significant as it contains a majority of the pre-pDC cells as well as the mature mini-bulk pDC sample. This cluster was the third most stable cluster, with a stability index of 0.58. The presence of a single early pre-cDC2 sample in cluster 9 may have impacted this stability.

Clusters 2 to 4 and 7 were highly variable and composed of pre-cDC2 and early pre-cDC2 cells along with the trilineage cells and mature cDC1 sample. The stability index of these clusters were in the region of 0.20 to 0.40, indicating variability in cluster contents over iterations of the SC3 analysis.

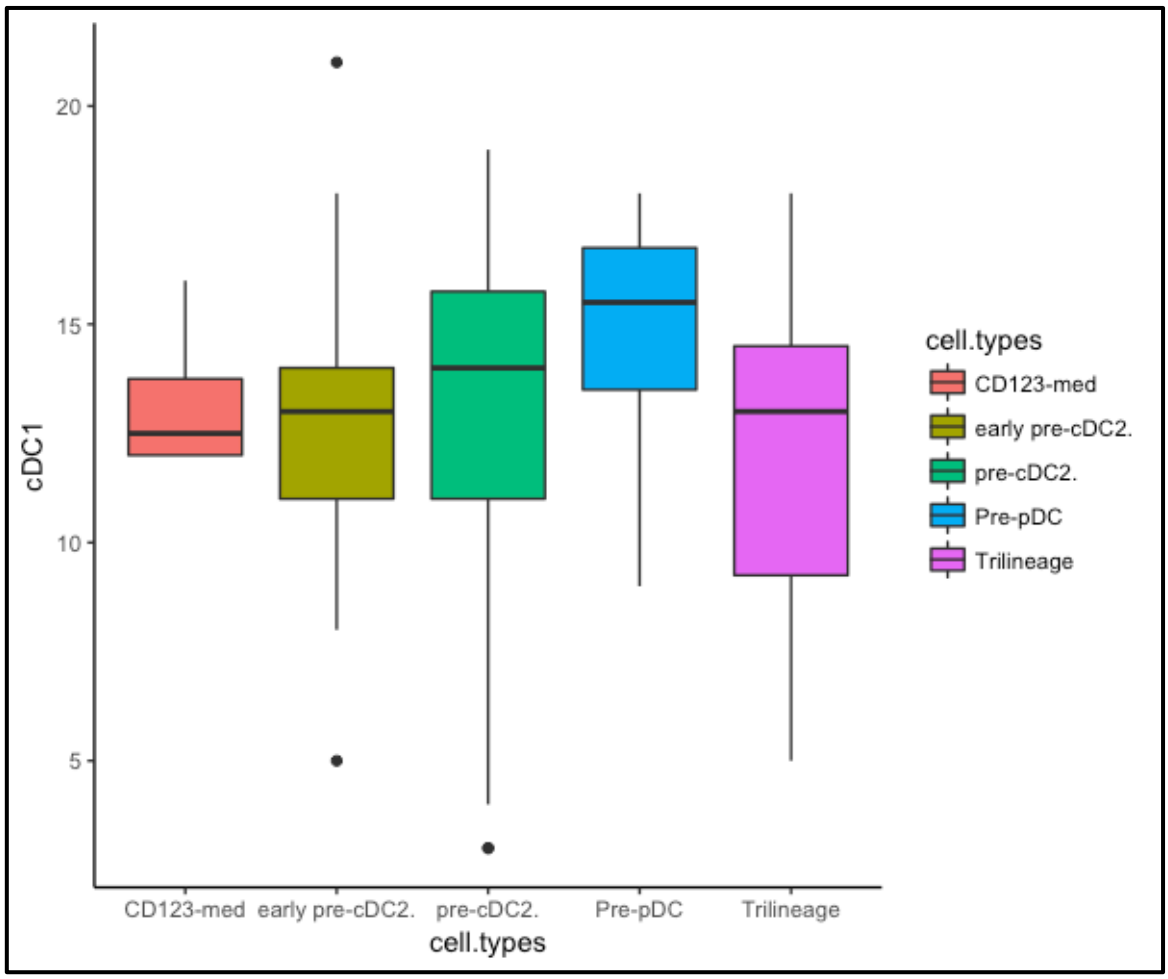


Figure 5.20: Expression of mature cDC1 signatures on single cell pre-DC data arranged by flow parameter grouping

By subsetting the genes to the cDC1 mature cell signatures derived from illumina expression analysis in chapter 3, each population of the pre-DCs were plotted alongside each other to determine if any of the expected subpopulations of pre-DCs were enriched for this mature cell type signature. This plot did not indicate a particular pre-DC subset was enriched for the mature cDC1 signature. CD123-med cells were subtly lower than the other cell populations in terms of cDC1 mature cell signature expression.

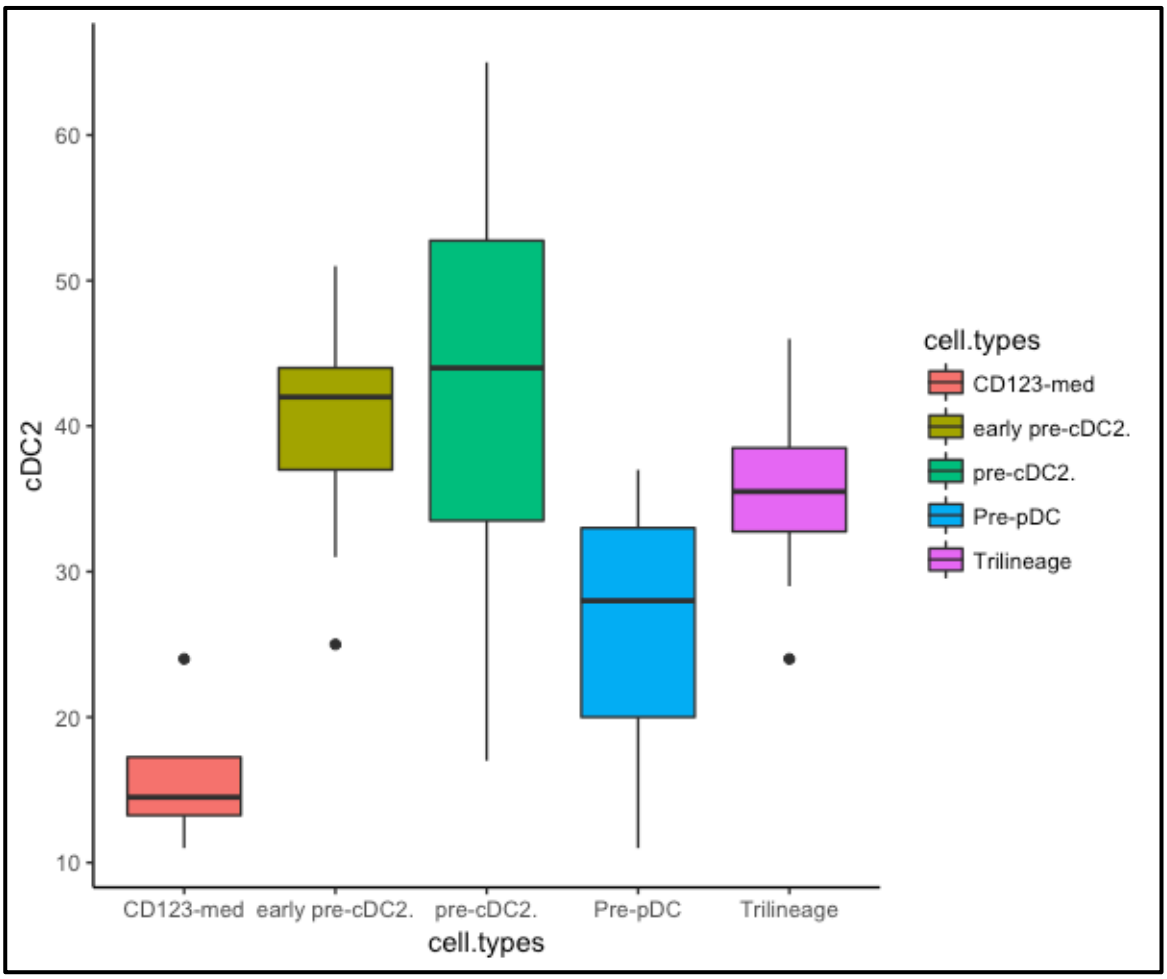


Figure 5.21: Expression of mature cDC2 signatures on single cell pre-DC data arranged by flow parameter grouping

By subsetting the genes to the cDC2 mature cell signatures derived from illumina expression analysis in chapter 3, each population of the pre-DCs were plotted alongside each other to determine if any of the expected subpopulations of pre-DCs were enriched for this mature cell type signature. The flow gating of the pre-DC populations uncovered a cluster of potential pre-cDC2 cells (green) that appear to express higher cDC2 signature markers than the other cell subpopulations. Median expression of cDC2 markers in the pre-cDC2 population was 46, with the early pre-cDC2 population (beige) at 44. The CD123-medium cells expressed the lowest median number of cDC2 signature genes at 15, closely followed by the pre-pDC population at 28. A number of the early pre-cDC2 cells expressed particularly high number of cDC2 marker features, between 50 and 70 each. This figure suggests that pre-cDC2 cells may be enriched for cDC2 mature cell signatures at the pre-DC stage.

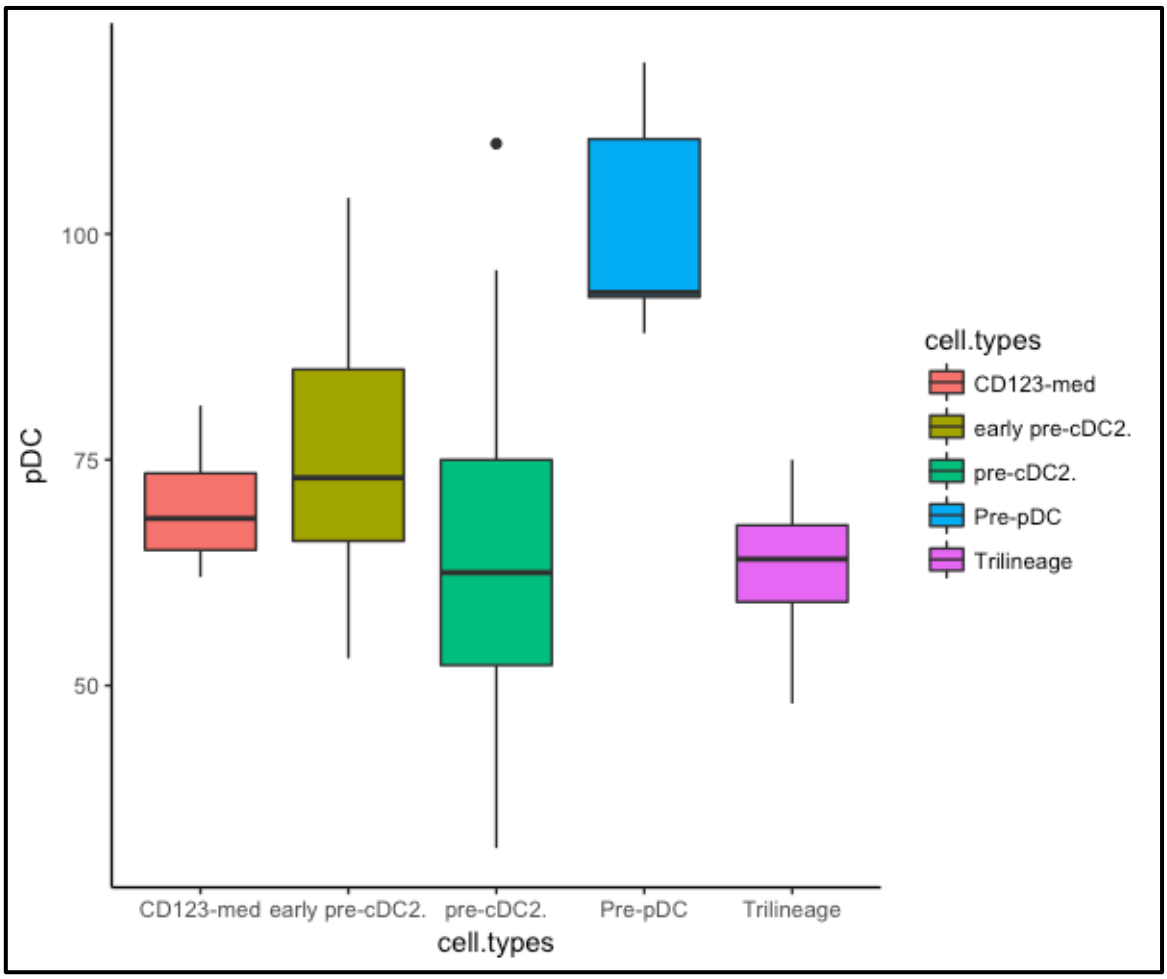


Figure 5.22: Expression of mature pDC signatures on single cell pre-DC data arranged by flow parameter grouping

By subsetting the genes to the pDC mature cell signatures derived from illumina expression analysis in chapter 3, each population of the pre-DCs were plotted alongside each other to determine if any of the expected subpopulations of pre-DCs were enriched for this mature cell type signature. In this figure, a strong bias towards expression of pDC markers in the pre-pDC (blue) population was observed. Median expression of around 90 pDC markers in the pre-pDC population was markedly higher than the other pre-DC populations. The lowest median expressers of pDC markers was noted in the pre-cDC2, early pre-cDC2 and trilineage cell populations. From this plot, it is evident that a pre-pDC subpopulation of pre-DCs are significantly enriched for mature pDC marker genes, skewing this subpopulation towards a pDC-like expression signature bias.

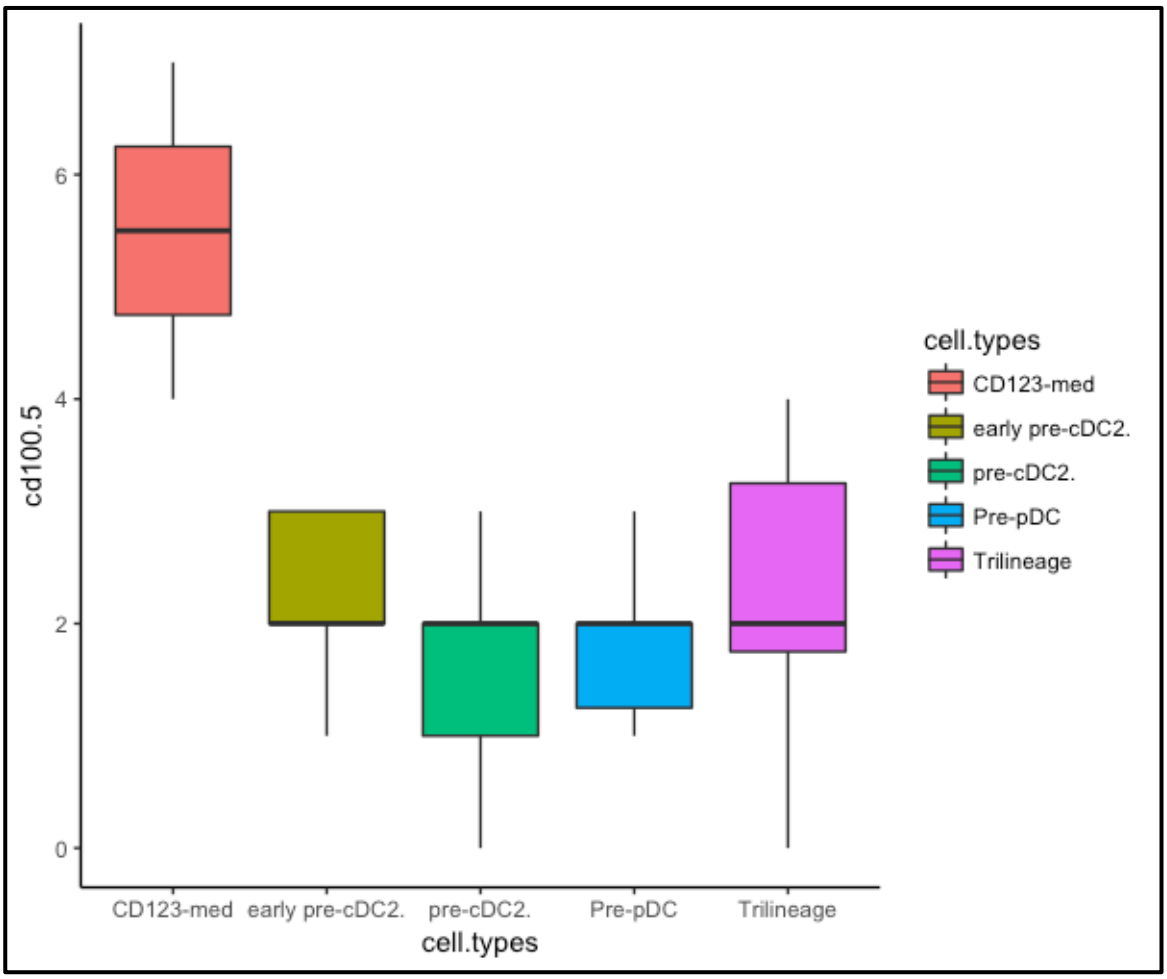


Figure 5.23: Expression of CD100+ CD34med signatures from *Villani et al* applied to single cell pre-DC data arranged by flow parameter grouping

Upon further investigation of the CD123medium expressing cells, it was suggested that although they leaned towards cDC1-like patterns by flow cytometry, they did not appear to express cDC1 mature markers at the single cell transcriptomic level. A subset of pre-DCs identified by *Villani et al* through single cell RNAseq implicated a CD100hi CD34med cell type that appeared similar to the CD123med subpopulation identified in this thesis. Of the 11 cell signatures for the CD100+ CD34med Villani population, nine of them were present in the 14,412 QC-passed feature list in this dataset. The expression of these genes were plotted for each of the pre-DC populations. The suspected equivalent CD123med subpopulation expressed the highest number of these markers with a median count of 5/9. The medians for the other subpopulations were approximately 2/9. Although the CD123med subpopulation was not enriched for cDC1 marker signatures in figure 5.20, they appeared to be enriched for markers highlighted by Villani *et al* in their identified CD100+ CD34med pre-DC population and may be an early cDC1 progenitor cell.

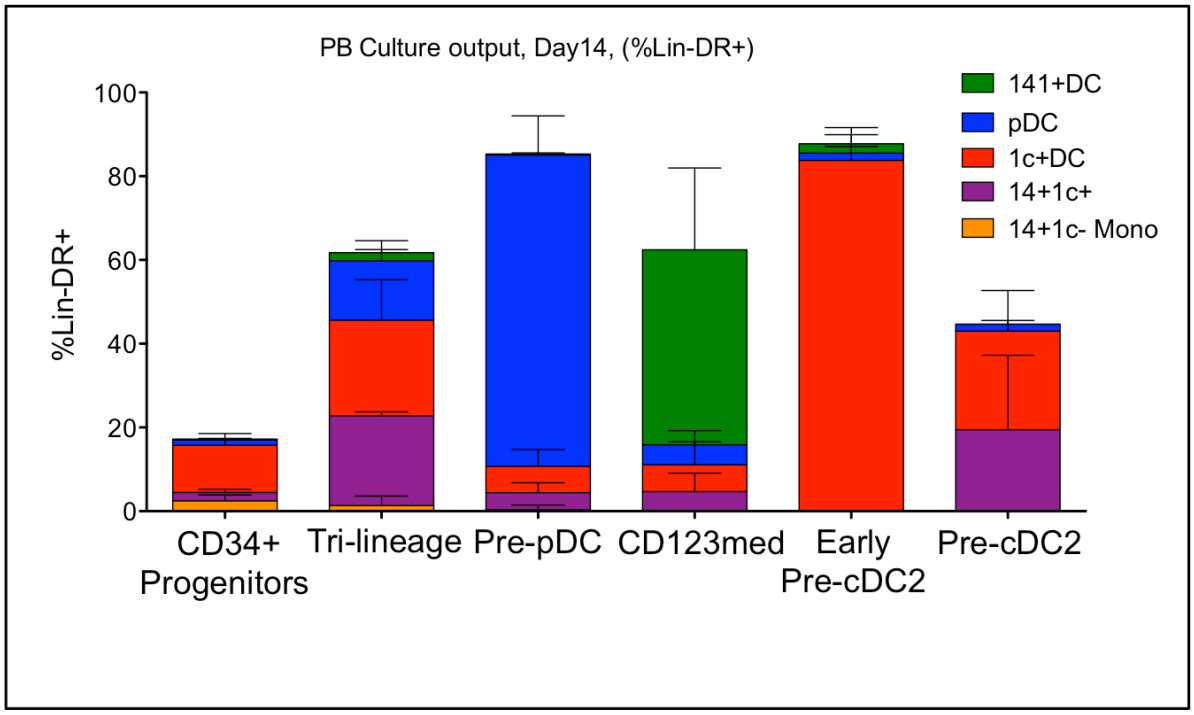


Figure 5.24: *In-vitro* Development Assay output from pre-DC subset populations and CD34+ progenitors shows cell type enrichment and early lineage commitment bias

Peripheral blood cells (and bone marrow cells for the early pre-cDC2) from each pre-DC sort gate were collected and cultured for 14 days with OP9 feeder cells with SCF, FLT3 and GM-CSF with the output sorted to determine if the pre-DC sort gate contained a heterogeneous mix of pre-committed cell types. Cells from the pre-pDC gate (CD123+,11c-,2+,303/304+) were enriched for pDC potential, pre-cDC2 and early pre-cDC2 cultures were highly enriched for cDC2 cells, while CD123med culture produced a majority cDC1 cells. The Tri-lineage culture appeared to be composed of a mixed, equal population of pDCs, cDC2s and CD14+1c+ cells. This data suggests that pre-DCs are a heterogeneous mix of committed progenitor cells, distinct from CD34+ progenitors and without monocyte potential. They exhibited a varied capacity to produce DCs and were enriched towards one or more mature cell types as noted during the single cell analysis of the populations.

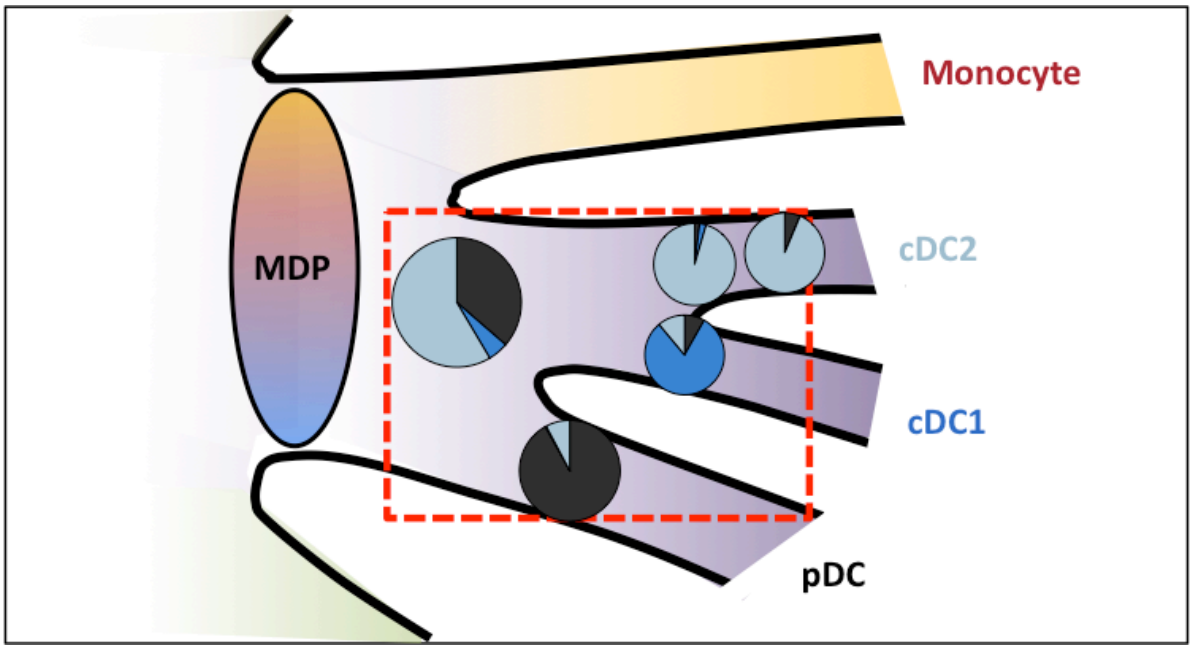


Figure 5.25: *In-vitro* development assay output from pre-DC subset gates and CD34+ progenitors show lineage-specific enrichment

This graphic plot highlights the findings from Figure 5.24 as pie-charts based on the percentages of cDC2, cDC1 and pDC cells from the *in-vitro* development assay output, correlated with where the population was expected to be along the haematopoietic lineage tree from Figure 5.1 and Figure 5.3.

The 'Trilineage' population, closely resembling the Common Dendritic Cell Precursor population was found to develop into all three DC lineages. The pre-cDC2 and early-pre-cDC2 populations were highly enriched for cDC2 cells through this assay. Similarly the pre-cDC1 population exhibited a strong mature cDC1-enriched output and the pre-pDC population produced a majority of pDC cells. The development assay successfully recapitulated the expected populations based on the 'priming' of the early pre-DC progenitors as displayed though single-cell analysis.

Chapter 6: OVERVIEW, DISCUSSION AND CONCLUSIONS

6.1 OVERVIEW OF TECHNIQUES USED

This thesis encompassed a range of novel and established expression techniques and bioinformatic analyses. Initial work was performed using publically available Illumina BeadArray expression data (GSE35457) to produce gene expression signatures for each mononuclear cell type of interest. Array-based techniques are useful for bulk-level analysis of the transcriptome and were used extensively prior to the advancement of RNA-Sequencing. In order to analyse the array data, an empirical Bayes statistic model was used to determine differential expression. The output correlated well with legacy knowledge of immune cell markers including CD14, C19orf59 and S100A9 in CD14+ classical monocytes, CX3CR1 in CD16+ non-classical monocytes, CLEC9A and BATF3 in cDC1s, CD1c, CD2 and FCER1A in cDC2s and PACSIN1 and ASIP in pDCs.

Once the initial analysis was complete, further mononuclear cell subsets extracted from skin were incorporated into the dataset and used to test a novel method of cross-tissue analysis designed by the author to deconvolute tissue-specific cellular expression differences between equivalent cell types. The basis of this analysis involved a two-tailed t-test between the pooled skin cell subsets and their peripheral blood equivalents in a process reversal of standard differential expression analysis. Here, rather than focusing on the differentially expressed genes, the differentially expressed genes were removed to reveal the extent of cellular equivalence in the absence of tissue-specific signatures that otherwise overshadowed the unsupervised clustering of the data. The successful clustering of blood and skin equivalent cells after applying the technique was used again in the comparison of blood and culture cell equivalents during Chapter 4.

Making use of the extensive 48,000 probe coverage offered by the Illumina microarray data and incorporating GeneSign (Spinelli et al., 2015) for signature generation, a list of positive gene signatures was produced to define peripheral blood monocytes, pDCs, cDC1 and cDC2 cell types. This signature was then validated in two ways, by machine learning using Linear Discriminant Analysis (LDA) to assign unknown samples a cell type in a blinded manner based on the samples gene expression profile and through the analysis of a comparable publically available dataset from another research group. Both processes confirmed the legitimacy of the cell type signatures, which were used as a basis for another novel analysis method created by the author for feature reduction based on repeated k-means testing and unsupervised clustering to determine a minimal gene list for monocycle and dendritic cell classification.

Using the full Illumina expression data as a surrogate, Illumina probe data related to the genes encoded for by the NanoString Immunology_V2 geneset were isolated and re-analysed as a new experiment. This provided a validated and robust dataset from an established technology with which to interrogate the efficacy of the newly developed NanoString nCounter analysis system available at Newcastle University. Results from hierarchical clustering and unsupervised analysis during the *in-silico* experimental stage, combined with initial correlation testing of RNA under diverse conditions including FFPE and fresh isolated RNA as well as whole cell lysate highlighted the benefits of the NanoString system for its ease of use, reproducibility, cost and directed profiling of major immune-related genes and pathways. Although competition within the field of RNA-sequencing has since driven costs down, at the time of writing, both transcriptome sequencing and microarray analysis was prohibitively expensive and thus the NanoString system proved a capable in-house alternative combining a high-throughput, probe-based array with capability for direct hybridisation to RNA from lysed cells without amplification.

Cells equivalent to those used in the surrogate dataset were collected, isolated and analysed using the nCounter platform and similarly visualised using hierarchical clustering, PCA and t-SNE. Consistency between the Illumina array, *in-silico* analysis and NanoString dataset reinforced the use of the NanoString system in later sections in order to build upon the blood mononuclear cell dataset with culture cell output.

DC and monocyte cells cultured from CD34+ bone marrow by Dr Urszula Cytlak-Chaudhuri were compared to the blood equivalent subsets on the NanoString platform using unsupervised clustering methods before the same novel method of cross-tissue analysis used in the comparison of blood and skin subsets was applied to deconvolute the blood and cultured subsets. Interrogation of the culture-specific genes was performed by functional enrichment analysis before the non-culture-specific genes were analysed, displaying a high degree of correlation between the *in-vitro* generated DC subsets and peripheral blood equivalents, extending beyond the cell surface markers used in initial FAC sorting and including known TLRs, interferon and cytokine receptors linked to developmental pathways inherent to each cell subset.

The final chapter of this thesis focused on the normalisation, analysis and comparison of pre-DC subsets, incorporating another gene expression technique in the form of advanced single-cell RNA-sequencing and further novel analysis techniques. Initial data handling and pre-processing required intricate command-line level pipelines combining Trimmomatic, STAR, SAMtools and HTSEQ packages (Anders et al., 2015; Bolger et al., 2014; Dobin et al., 2013; Li et al., 2009), prior to normalisation and analysis of the data in the 'R' environment using SCATER, RUVseq, M3Drop and SC3 (Andrews and Hemberg, 2017; Kiselev et al., 2016; McCarthy et al., 2017; Risso et al., 2014). Incorporating gene signature output generated from Chapter 3 as well as external datasets from Villani *et al* (Villani et al., 2017), pre-DC expression of mature cell signatures was explored, highlighting the heterogenic cell type bias as the pre-DC developmental stage. Culture developmental output of the heterogeneous pre-DC populations confirmed the developmental bias of pre-DCs towards a pDC, cDC1 or cDC2 lineage, exclusive of monocyte potential.

6.2 RESEARCH OUTPUT

The research output of this thesis has included the generation of cultured mononuclear cell subsets with the support of Dr Urszula Cytlak Chaudhuri from CD34+ bone marrow cells as well as the generation of bona-fide, non-inflammatory, non-monocyte-derived mature DC subsets from pre-DC populations in peripheral blood and bone marrow. Both flow cytometry and FACS of blood and cultured DCs and monocytes was produced with support from members of the Human Dendritic Cell Lab.

Microarray data was extracted from an online GEO repository and used to produce a validated gene signature of 3,439 genes relating to the four mononuclear cell subsets. Code relating to the normalisation and analysis of Illumina expression data was developed along with ComBat-based normalisation to incorporate multiple datasets into a single experiment. This pipeline has since been used for the generation of data and visualisations in Immunity publication, McGovern *et al*, 2014. Following this, a novel method of deconvoluting mixed-tissue datasets was developed and tested on two platforms, Illumina BeadArray and NanoString nCounter.

A further novel method of analysis was developed and implemented to reduce the developed signature to the minimal number of genes required to maintain clear distinction of the DC and monocyte subsets, proving successful and separating the four populations by their expression of only two genes.

A dataset of equivalent blood and cultured monocytes and DCs was developed using the NanoString platform alongside code for the normalisation, analysis and visualisation of NanoString data. This pipeline of analysis has since been used in multiple publications as highlighted in the Appendix.

A pipeline for pre-processing of Smart-Seq2 single-cell RNA-Sequencing data was incorporated with the support of Dr Rachel Queen and code for the normalisation, visualisation and analysis of single cell expression data was produced for this thesis. This work is on-going in the Human Dendritic Cell Lab and thus has not been incorporated into publication at this time.

The scope of this thesis has resulted in the production of analysis pipelines and data across NanoString, microarray and RNA-sequencing platforms, incorporating all of the major high-dimensional transcriptome profiling techniques currently available in research. The analytical skills derived from this work are applicable to current and future gene expression research and have been used in multiple first author and contributing author papers outside the scope of this thesis.

6.3 RESEARCH IMPACT

This research project covered a range of key questions in dendritic cell research and its applications though transcriptomics technology and bioinformatics has uncovered novel findings in the field as well as supporting currently accepted research.

The initial key research questions for this project focused on the identification and distinctions between dendritic cells and monocytes from peripheral blood. As discussed previously, while closely related and developmentally similar, each dendritic cell subset has a specialised function and role to play in immunity. Because of this, they also express a few specific cell surface receptors, CD antigens and co-stimulatory molecules to support these functions. Output from the Illumina expression dataset provided an interrogable resource of 47,000 probes for the identification of greater distinguishing features between these cell types. The full illumina expression dataset, including skin DC and monocyte equivalents as well as mouse counterpart data (unused in this thesis) is available at the NCBI Gene Expression Omnibus under GSE35457 and so can be used freely by other researchers wishing to investigate DC biology and transcriptomics in human and mouse.

As well as the whole resource being available for external use, the specific markers identified in this thesis may also aid in the future isolation and identification of DC subsets for analysis. Differential expression testing revealed a 3,439 gene signature of DC and monocyte subsets, which could be applied to multiple datasets, including other Illumina expression data (as observed with GSE65128, Chapter 3), applicable across tissues (Chapter 3), culture conditions (Chapter 4) and even applicable to single-cell sequencing (Chapter 5). This transcriptome-derived signature resulted in the identification of known and novel DC subset marker gene targets, including S100A8 and C19orf59 on monocytes, PACSIN1 and GZMB on pDCs and CD1c cDC2s. All of these genes encode surface membrane-bound proteins and are thus easily applicable for use in immunohistochemistry, flow cytometry and FACS analysis. Through the use of 'COMPARTMENTS' web resource developed by the Jensen Lab and Novo Nordisk Foundation Centre for Protein Research, many of the 3,439 genes identified in this thesis as markers of individual DC or monocytes subsets have been shown through curated literature, high-throughput screening and sequencing-based prediction to encode surface proteins, but also included some nuclear proteins and cytosolic proteins, which may be useful for DC identification in other techniques or conditions where expression of cell surface proteins may be perturbed or disrupted.

Not only has the data contained in this thesis expanded on current knowledge of DC and monocyte development and biology, but the analysis pipelines designed and produced by the author for this thesis have since been used in other projects and by other research groups for analysis of transcriptome data and microarrays. At the time of writing, NanoString Technology in particular, had no user-friendly analysis tools or established pipelines for analysis and so those contained in this thesis became the backbone pipeline for all NanoString assays performed by the Human Dendritic Cell Lab and their collaborators. Furthermore the tSNE app developed by the author and displayed in Appendix C, was also part of the analysis pipeline established by the author and is still used frequently by members of the HuDC Lab for quickly visualising high-dimensional flow cytometry data in a user-friendly environment.

The analysis code first used to deconvolute the skin and blood in Chapter 3, and used again in Chapter 4 to correlate blood and cultured DC subsets, can be applied to any other project where biological insights are overshadowed by a conditional expression pattern. Indeed, a similar approach was employed in McGovern *et al*, *Nature,* 2017 based on the same GSE35457 Illumina expression data and is in consideration for use in the Fetal Cell Atlas project to deconvolute DC and progenitor cell equivalents across fetal tissues.

Comparisons between the cultured subsets and blood equivalents have indicated that under culture conditions and with OP9 stromal feeder cells, phenotypically equivalent cells to DCs and monocytes found in peripheral blood were obtained after 21 days. Not only were the cells phenotypically equivalent, but their transcriptional conservation was also revealed to be significant after removal of condition-specific genes. The genes and functional pathways affected by culture conditions and those that were conserved are both important factors for research impact. In determining culture methods or the applicability of research findings collected using *in-vitro* generated cells to primary cells, knowledge of the genes and pathways likely to be deregulated is an important consideration. The typical plasticity of DCs combined with their ease of disruption when removed from their native environment may result in dramatic alterations in the predicted properties and functions of DCs. The pathways determined in this thesis as being altered under culture conditions were proliferative, stressed and apoptotic signals as well as response to stimulants, which is to be expected under culture conditions, but also involved cellular migration and exposure to foreign antigens which may need to be considered when studying cell motility or immune response signals in cultured DCs and monocytes.

The conservation of subset specific signatures between cultured and peripheral blood DCs and monocytes does suggests that the culture model employed in this thesis could recapitulate bona-fide DCs, well beyond cell-surface marker equivalence. This finding has already had a major impact on the research performed in the Human Dendritic Cell Lab, as it meant large quantities of DCs could be produced that would likely react under investigation in a comparable way to primary DCs. This may also have a wider research impact for projects where obtaining sufficient cell numbers for analysis is not feasible from human blood, or in the case of immunotherapy, would allow for the generation of vast quantities of immune-specific DC cells that are phenotypically, developmentally and functionally as capable as true blood DCs, but expanded in an easily reproducible manner.

The pre-DC priming experiment described in Chapter 5 supports the 'early lineage priming' model of haematopoiesis by revealing heterogeneity within precursor populations and a skewing of the precursor cells towards a specific mature cell profile. In confirmation of initial single-cell transcriptomics analysis, the *in-vitro* development assay was incorporated into the project and determined that the primed progenitor cells could be isolated and grown in culture to generate *bona-fide* DCs enriched for particular DC subsets base don the expression profile of the pre-DC precursor cells. The ability to collect pre-DC populations and culture them into mature *bona-fide* DCs enriched for particular subsets of interest has great implications in research. Scarcity of DCs, particularly CD141+ cDC1s in peripheral blood can make their study arduous, however from the comparison of blood and cultured cells in chapter 4, and the development of DCs from pre-DCs in chapter 5, the capacity and consistency of cDC collection and culture has been uncovered. Further impacts on the wider research community may be gleamed from the possibility that primed progenitors of various DC subsets could be isolated and grown in culture and used to generate *bona-fide* DCs for research and medical applications in the future.

6.4 LIMITATIONS OF THE PROJECT

This project incorporated multiple cell sequencing and staining technologies including flow cytometry, multiplexed hybridisation-based expression analysis and microarray experiments, alongside single-cell transcriptomics, culture systems and development assays to reveal the nature of mature dendritic cells subsets, their relation to monocytes and their developmental heterogeneity. Each of these techniques and methods had their own inherent limitations, some of which could be mitigated through experimental design, and some of which can only be accounted for afterwards.

One major technical constraint in this project, common to all of these applications, is their initial reliance upon flow cytometry and FACS for initial cell sorting and isolation. While FACS is well-established in dendritic cell research, its capacity stretches only to a small number of fluorophores marking only cell surface receptors. With the difficulty in collecting and isolating DCs and limited number of commonly used cell surface markers, the process of isolation is confounded. Adding additional fluorophores increases the spillover of signals, which can only be partially mitigated through compensation, which requires some advanced skill and knowledge of the fluorphores to balance effectively. Dual staining (where multiple markers are added to the same channel) was used to increase the number of markers that could be used, however this is still hindered by the software limitations of the sorter, limiting sorting to 8 sub-gates. Further gating of populations can't then be sorted. Furthermore, FACS relies on cell suspensions and thus samples isolated by FACS no longer retain their tissue architecture or cell-to-cell interactions. More technical limitations of flow technology is that the fluorescent intensity data has a typical dynamic data range from 10^{-1}-10^{5}, which is far below the range of NanoString or RNA-sequencing, making comparisons between the two difficult, further exacerbated by potential differences between protein and RNA expression in cells which are frequently quoted to correlate up to 60-80%.

While the costs of sorting are relatively low compared to microarray and sequencing costs, to collect cDC cell numbers sufficient for analysis took over 5 hours per sample and included a cocktail of 14 fluorescent markers, seriously increasing the associated costs.

As described in Chapter 1 and Chapter 3, NanoString Technology proved to be a reasonable, cheaper and faster alternative to microarray assays for this project's specific research requirements, however, when the experiments were initially performed, the novelty of the technology meant that there was very little published data available, and published methods for experimental procedures and analysis were not consistent. NanoString recommended analysis by two-tailed T-test, although published papers at the time did not always perform statistical tests on the data, relying solely on fold change estimates (Baugh *et al.*, 2011). Others performed a T-test, but did not adjust for multiple testing (Waisberg *et al.*, 2014). Such methods are not stringent enough for such high throughput assays and so a pipeline for analysis of NanoString data had to be developed for this project, incorporating a more robust differential expression analysis with correction for multiple testing, as no established analysis pipeline was available.

Preliminary NanoString experiments also highlighted the need for a modified standard operating procedure (SOP) from that published by NanoString, as the exact reagent quantities supplied by NanoString did not include additional volume to account for loss through pipette use or error and were not sufficient to supply all 12 wells of a cartridge with the required volumes. As part of this project, and to ensure such issues did not arise during the main stages of this investigation, a modified SOP was designed and tested that involved the addition of excess buffer to the reagents prior to the hybridisation step. To compensate for any potential dilution, the prep-station was set to a 'high sensitivity' setting which extended the cartridge washing steps to improve binding of captured molecules to the streptavidin surface of the cartridge and increase probe counts.

While analysis limitations were met sufficiently through design and protocol modification, the limited number of probes available on the NanoString panels did prove to be an issue. The 579 probe Immunology_v2 panel used in the NanoString analysis sections of Chapter 3 and Chapter 4 lacked some of the major defining genes for DC subsets including, most critically, CD1c. To address this limitation, an additional 30 gene probes were spiked in to the assay to provide count data on omitted probes, but even still, many genes associated with DC functions and development were not included in the panel. Additionally, the Immunology_V2 panel was a commercially available pre-constructed probe panel and so included many probes that were not relevant or of interest to this project, but could not be replaced or exchanged.

The SmartSeq2 single cell transcriptomics protocol excels in producing high quality, full-length transcriptome data, with the trade-off being the very low throughput due to the 96-well plate-based processing. The pre-DC experiment in Chapter 5 resulted in 71 viable cells after QC, which when compared to modern sequencing, is relatively few. However, the cells did express 4,000 to 10,000 unique genes each, which is far higher than comparative protocols such as the 10X chromium, which typically generates between 200 and 3,000 genes per cell. While cell numbers, and thus power, were relatively low, the high gene counts per cell were necessary for determining subtle differences in pre-DC populations with reduced risk of gene drop-outs affecting the output. Total reads per cell were projected to reach 1,000,000, but fell short of this, with a range of 200,000 to 900,000 total reads per cell. This may have caused some issues as for example, in the mature pDC spike-in, CD123 (the conventional surface marker for pDC identification) was missing. It is likely that other gene drop-outs may have affected the overall clustering of the data too, exacerbated by the fact that the mature samples were pooled samples of 10 cells each, thereby vastly increasing their average gene counts and total reads. This had an impact on the initial cell clustering, as ‘total number of genes’ proved to be a high source of data variance as indicated in Figure 5.13 and may have been driving some of the clustering algorithms to separate cells by total gene count.

Two of the major limitations affecting this project stemmed from the scarcity of samples available for research. Healthy volunteers largely provided the peripheral blood samples required for this project, typically up to a volume of 50-80ml. As CD141+ cDC1s are present at 0.1% of peripheral blood mononuclear cells (approximately 1,000 cells per ml of blood), after pre-processing, and enrichment via FACS (from which internal optimization experiments have indicated a 60-80% return on cells based on the cell counts given by the FlowJo software), successfully obtaining the required number of cells for analysis proved to be a difficult task. After RNA extraction and quantification, the typical yield would not reach the required number for analysis. This issue was partially addressed through the use of the NanoString nCounter protocol, which at the time of processing, prior to a change in the reaction chemistry of the NanoString assay, allowed for the use of direct cell lysates in the protocol. Using this method, cell lysates approximated at 2,000 cells per µL were collected directly from the FACS machine for NanoString analysis, reducing RNA loss through the extraction process. Unfortunately, after completion of this project, the lysate protocol was altered through a change in the chemistry of reagents provided by NanoString for the assay, limiting the volume of lysates to 1µL, which is not sufficient to adequately lyse the 10,000 cells used in the assay.

The difficultly in obtaining enough viable cDC1 cell samples resulted in a down sampling of other more abundant cell samples during the analysis stages. In order to ensure bias was not introduced into the comparative analysis of the NanoString dataset by the inclusion of more of some populations than others, only 3 samples of each cell type were used in the analysis for Chapter 3 and 4. This meant that some more abundant samples such as CD123+ pDCs and CD14+ monocytes were omitted from analysis. These samples have since been included in other research projects, but for direct comparison between the subsets, equal sample numbers were required, otherwise generated signatures and clustering would be highly skewed towards subsets with the greater number of samples.

The issue of sample collection was also a concern during the development assay and culture models, both of which relied on CD34+ bone marrow progenitor cells. Bone marrow was extracted from the femoral head of donors following hip replacement surgery and were relatively infrequent and in high demand. Because of this, a lot of time had to spent waiting for samples to arrive. A further implication of this method of bone marrow collection was that the majority of donors were between 60-80 years of age and thus, twice as old as the majority of the healthy blood donors. As the immune system changes with age and older patients are more frequently taking prescriptions for long-term health issues that may affect the immune system, growing cells from the marrow of elderly patients may introduce unforeseen changes in the development and proportions of DC populations and introduce errors when comparing the bone marrow of an older individual with blood from younger donors. For this reason, bone marrow from younger donors were used when possible.

Finally, while the use of previously generated Illumina expression data was extremely useful in this project, the reliance of this dataset restricted the possibilities for comparison as no additional samples could be added as the work progressed. The ComBat algorithm was used to merge already developed datasets in this case, but this relied on a conserved population between each dataset and fundamentally, the fact that both datasets were analysed on the same version of the microarray using the same reagents and protocol. Even so, it is unlikely that all batch effects were removed through this regression technique. Adding additional samples iteratively, as can be done relatively simply using NanoString, flow cytometry or 10X, is not easily performed using microarrays or SmartSeq2 protocols without major batch effects being introduced.

6.5 FUTURE RESEARCH VISION

As this project involved the generation of novel analysis techniques and pipelines, the development and testing of which took a vast amount of time, there remains a number of possible follow-up experiments that could be performed to further support the finding in this thesis or develop them further. At the time of writing, some of these experiments are already being performed by the Human Dendritic Cell Lab for publication or as part of other separate projects.

One of the major outcomes of this project was the production of analysis pipelines for NanoString, Illumina BeadArray and single cell sequencing experiments. These pipelines have great scope beyond this work and have already been implemented in part or in full for other projects and publications. To date, the NanoString analysis code created for this thesis has been implemented in a number of publications. Analysis of high-risk primary biliary cholangitis (PBC), where patients do not respond adequately to Ursodeoxycholic acid (UDCA) treatment after 1 year, was performed using the analysis pipeline in this thesis. This analysis revealed a distinct ‘high-risk’ gene signature with a strong senescence signal that could accurately stratify patients at initial disease presentation into potential responders and non-responders to UDCA, with a clear clinical utility in patient prognosis and indicator for second-line therapeutic intervention in these cases. Further follow up of this work is continuing with further refinement and validation of potential gene targets or proteins for clinical use. A similar project in the related condition, autoimmune hepatitis (AIH), also used this analysis code for investigating the possibility of patient stratification based on treatment responders and non responders using follow-up data from the UK-AIH consortium. Similarly, this project has moved into the validation stages for potential early-state patient stratification.

The same approach is currently in use for the stratification of cSCC patients to determine if there are tumour related signatures associated with risk of developing metastatic disease upon presentation of cSCC, some data from which has already been incorporated into Nature Immunology paper ‘Epithelial damage and tissue gd T cells promote a unique tumour-protective IgE response’ currently in print. The ability for NanoString Technology to use low-quality mRNA from FFPE material opens up a vast resource for clinical biology and patient stratification with thousands of biopsies from hundreds of disease states stored across Newcastle University Medical School ripe for mRNA analysis with years of patient follow-up data available to interested researchers.

Besides the analysis code and methodology produced in this thesis, the direct data output from this project can be utlised and further interrogated for more potential information. The DC and monocyte geneset and generated signatures can be used in future experiments comparing cell type specific perturbations linked to genetic diseases. Congenital and acquired cellular deficiencies such as dendritic cell, monocyte, B and NK lymphoid deficiency (DCML) presents as an almost complete depletion of HLA-DR+ Lineage- cells by flow cytometry, including loss of pDCs, cDCs and monocytes. Very few of these cell populations remain. By comparing these cells to healthy populations and the signatures generated in this project, we can determine the extent of purtubation and reveal important insight into the biology of DC and monocyte development through phenotypic and transcriptional studies of the remaining cells. Insights into monocytes and DCs have already been gleamed from the investigation of cell population perturbations in patients with IRF8 mutations, demonstrating the critical requirement of IRF8 in immunopoietic development and anti-mycobacterial immunity.

The generated genesets can also be used to determine the correlation between inflammatory DCs and steady state DC and monocyte subsets. For example, post transplant Graft versus Host Disease (GvHD) is a potentially fatal inflammatory immune complication of haematopoietic stem cell transplantation. On going projects in the Human Dendritic Cell Lab have already demonstrated that by 40 days post-transplant, dermal DCs are mainly donor derived and, in the absence of GvHD, phenotypically resemble steady state dermal DCs. Preliminary data suggest that there are significant phenotypic differences that occur in GvHD, such as a predominance of CD14+ dermal DCs. DC subsets from GvHD affected skin biopsies could therefore be FACS sorted and their gene expression profiles analysed by NanoString nCounter assay technology to assess their relationship to steady state and *in-vitro* derived DCs under the hypothesis that inflammatory DCs are more closely related to monocytes than steady-state DCs. A similar analysis could then be performed on DCs from other inflammatory skin disorders including psoriasis, eczema or drug eruptions to determine if a 'core' inflammatory gene expression signal can be distinguished, or if the underlying mechanisms of such conditions arise from different functional gene expression changes in affected tissue.

Such an investigation may be extended to include other major tissue types commonly affected by high-grade (III-IV) GvHD such as the lung and gut. GvHD of these tissues is associated with higher risk of mortality and long-term morbidity and thus uncovering the gene expression 'phenotype' of such diseases may promote or facilitate further translational research into high-grade GvHD clinical therapies.

The pre-DC single cell analysis to determine lineage bias and early priming of cells may be expanded upon by the incorporation of earlier and later cell populations. For this project, the cells of focus were mainly extracted from the traditional CDP gate and displayed early priming of both pDCs and cDCs. Using later, more developed cells in the analysis may provide a more defined indication of mature cell transcriptome. Combining this with earlier cell populations, tracing back to haematopoietic stem cells, will provide a complete dataset for lineage tracing through comparative expression or pseudotime analysis and reveal the exact stages of DC development and lineage priming. This work could then be moved into culture experiments with an aim to return a greater yield of mature cell types of interest through the development and expansion of primed pre-DC cells.

While the culture work contained in this thesis provided a strong basis for the generation of *bona-fide* DCs, beyond surface marker expression and phenotype, it also revealed a culture-specific gene signature that altered the expression patterns of the cell subsets to an extent. Further experiments aimed at determining the effect of different feeder layers; soluble mediators and stimulants on DC potential and correlation to steady-state peripheral blood DCs would be a further step towards creating true transcriptionally identical *bona-fide* DCs from culture or increasing the output of particular cell subsets. Work in the Human Dendritic Cell Lab has already begun to generate cDC1s from pre-DC subsets, with a future vision to take the knowledge into the clinic, where precursor cDC1 cells from patients could be expanded in culture and implanted for use in DC-based immunotherapy.

Appendix

A: EXTERNAL FILES

Large files related to the work discussed in this thesis are available with the electronic version of this thesis as:

External file 1

External file 2

External file 3

B: PANEL+ GENE SELECTION RATIONALE

The full list of Panel+ markers selection for inclusion in the NanoString analysis of Chapter 3 and 4 is listed below, including the gene function, rationale for selection and referenced journal article where applicable.

Gene name	Function	Rationale	Reference
ASIP	Novel marker. Involved hair pigmentation. May have some role in lipid metabolism	pDC marker identified from Chapter 3 Illumina expression analysis	Novel target.
C19orf59	Novel marker. Speculated to be involved in immune responses and mast cell differentiation.	Monocyte marker identified from Chapter 3 Illumina expression analysis	Novel target.
CCL17	Antimicrobial cytokine displaying chemotactic activity for T-cells. Binds to CCR4 and CCR8.	Expressed by stimulated DCs.	Stutte *et al*, *PNAS*, 2010
CD1C	Mediates presentation of lipid and glycolipid antigens from self or microbial origin to T-cells	Well publicised marker of cDC2 DCs	Robbins *et al*, *Genome Biology*, 2008

CD207	Internalisation of antigen into Langerhans-specific organelle, Birbeck granules. This provides a route for non-classical antigen-processing pathways.	For external project. Encodes Langerin. Expressed only in Langerhans cells.	Crozat *et al*, *Immunological Reviews*, 2010
CLEC10A	Diverse functions including cell adhesion, signalling, glycoprotein turnover and roles in inflammation and immunity.	Well-publicised marker of cDC2.	McGovern *et al*, *Immunity*, 2014
CLEC9A	Endocytic receptor for uptake and processing of material from dead cells. Mediates cross-presentation of dead-cell antigens	Well-publicised marker of cDC1.	McGovern *et al*, *Immunity*, 2014
CLNK	Plays a role in regulation of immunoreceptor signalling, including BCR and FCER1A signalling.	Found on mCD8 mouse DCs, expected to be equivalent to CD141+ BDCA3+ cDC1 DCs in humans	Robbins *et al*, *Genome Biology*, 2008
COLBLL1	Encodes actin regulator protein, important I reorganisation of the actin cytoskeleton.	Murine pDC specific marker.	Robbins *et al*, *Genome Biology*, 2008
CXCL5	Chemokine for the activation and recruitment of neutrophils. CXCL5 is proposed to bind CXCR2 to recruit neutrophils and promote angiogenesis.	Well-characterised marker of inflammation.	Koltsova *et al*, *Immunity*, 2010
DAXX	Regulates apoptosis and cell differentiation, as well as immune system response	cDC1 marker identified from Chapter 3 Illumina expression analysis	Torri *et al*, *PLOSone*, 2010
DBN1	Involved in cell migration, neuronal processes and plasticity of dendrites. Required for actin polymerisation and CXCR4 recruitment to immunological synapses.	Found on mCD8 mouse DCs, which are expected to be equivalent to CD141+ BDCA3+ cDC1 DCs in humans	Robbins *et al*, *Genome Biology*, 2008

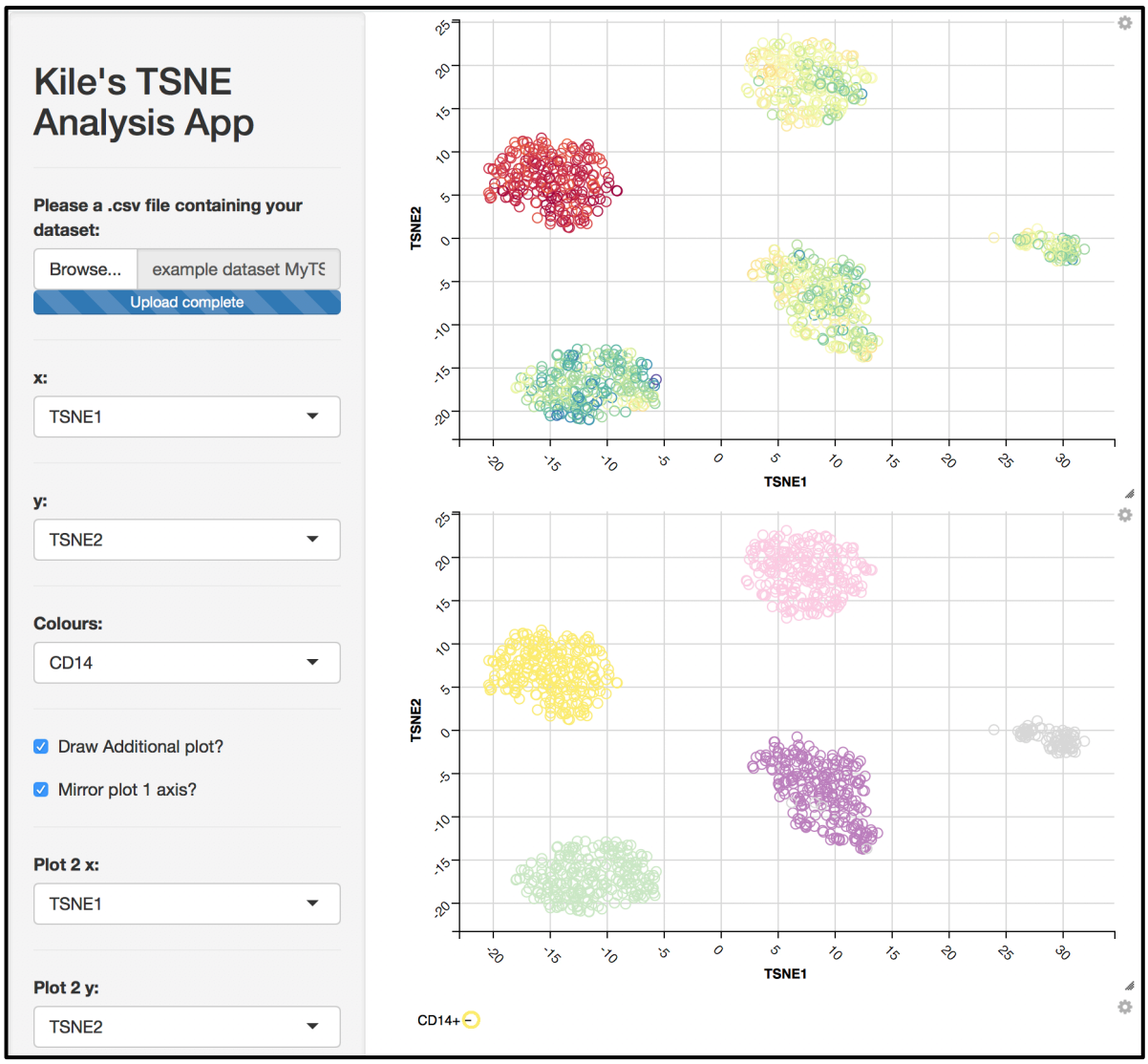
F13A1	Activated by thrombin, in DCs and macrophages it plays a role in the regulation of cell motility.	Noted expression on alternatively activated macrophages. Also included for external project.	Haniffa *et al*, *J Exp Med*, 2009
FGD6	May activate CDC42 and plays a role in regulating the actin cytoskeleton and cell shape.	Found on mCD8 mouse DCs, which are expected to be equivalent to CD141+ BDCA3+ cDC1 DCs in humans	Robbins *et al*, *Genome Biology*, 2008
FLT3	Class three-receptor tyrosine kinase that regulate haematopoiesis. Pathways include apoptosis, proliferation and differentiation of haematopoietic cells.	Found highly expressed on DC subsets. Also included for external project.	Robbins *et al*, *Genome Biology*, 2008
GCSAM	Novel marker. Involved in signal transduction and negatively regulates lymphocyte motility. Also a regulator of B-cell receptor signalling.	cDC1 marker identified from Chapter 3 Illumina expression analysis	Novel target.
GGT5	Encodes an enzyme capable of converting leukotriene C4 to leukotriene D4. Pro-inflammatory macrophages may upregulate this to modulate inflammatory processes.	Identified in dermal CD14+ cells, but not blood DCs or monocytes	Haniffa *et al*, *J Exp Med*, 2009
LPAR2	Functions as a lysophosphatidic acid receptor and induces Ca2+ mobilisation in response to LPA in cells.	Negatively regulates dendritic cell activation and inflammation	Emo *et al*, *J Immunol*, 2012
LYVE1	Has a role in autocrine regulation of cell growth. Mediates uptake of hyaluronan for catabolism within lymphatic cells or transport to lymph nodes for degradation.	Shown to identify dermal macrophages in situ.	Haniffa *et al*, *J Exp Med*, 2009

MAFF	May be involved in the cellular stress response. Encodes a transcription factor to enhance expression of oxytocin receptor gene.	Mouse cDC marker published in Robbins et al.	Robbins *et al*, *Genome Biology*, 2008
MERTK	Regulates cell survival, migration, differentiation and in the case of macrophages, in the phagocytosis of apoptotic cells.	Monocyte and macrophage marker in humans and mice.	McGovern *et al*, *Immunity*, 2014
KI67	The expression of Ki67 is strictly associated with cell proliferation. It is present in all active phases of the cell cycle, but absent in resting cells.	Well-publicised marker of cell proliferation.	Scholzen *et al*, *J Cell Physiol*, 2000
NDRG2	May be involved in dendritic cell and neuron differentiation. Contributes to the regulation of Wnt signalling pathway.	Expressed on pDCs and cDC2s, identified from Chapter 3 Illumina expression analysis	Novel target.
PACSIN1	May play an important role in pDC IFN-I production	Found on murine pDCs. Also identified from Chapter 3 Illumina expression analysis.	Robbins *et al*, *Genome Biology*, 2008
PPM1N	Novel marker. Unknown functions, but it encode a protein phosphatase. May be involved in magnesium ion binding. Paralog of PPM1A, which is a negative regulator of cell stress response.	CD16+ monocyte marker identified from Chapter 3 Illumina expression analysis	Novel target.
PRAM1	Novel marker. Expressed and regulated during normal myelopoiesis. May be involved in integrin signalling in neutrophils.	Monocyte marker identified from Chapter 3 Illumina expression analysis	Novel target.

S100A12	Plays a role in regulation of inflammatory processes and immune response. Stimulates innate immune cells, recruits leukocytes, promotes cytokine and chemokine production and regulates cell adhesion and migration.	CD14+ monocyte marker. Also identified from Chapter 3 Illumina expression analysis	Schmidl *et al*, *Blood*, 2014
TMEM14A	Novel marker. Regulates apoptosis signalling pathways via negative regulation of mitochondrial outer membrane permeabilisation.	cDC1 marker identified from Chapter 3 Illumina expression analysis	Novel target.
UPK3A	Novel marker. May play an important role in preventing bacterial adherence.	cDC marker identified from Chapter 3 Illumina expression analysis	Novel target.
ZBTB46	Zinc finger and BTB domain containing protein. DC transcription factor.	DC transcription factor and marker of pDC and cDC.	Haniffa *et al*, *J Exp Med*, 2009

C: T-SNE ANALYSIS APP

t-SNE visualisation provides high dimensional reconstruction of a dataset in a two-dimensional plot, aiding researchers in identifying patterns within the data. This technique is implemented in the 'R' environment and thus is not easily accessible to non-bioinformatically trained researchers. To enable members of the Human Dendritic Cell Lab to visualise their data, particularly flow cytometry, the author created an 'R'-based companion App for t-SNE analysis. The app processed .csv or .fcs files, performed t-SNE analysis on the data and presented the output with options for altering the axis, colours and display in a drop-down format. A smaller working example of this app is included as an electronic supplemental to this thesis with example output displayed on the following page.


Kile's TSNE Analysis App
Please a .csv file containing your dataset:
Browse...
example dataset MyTS
Upload complete
x:
TSNE1
y:
TSNE2
Colours:
CD14
Draw Additional plot?
Mirror plot 1 axis?
Plot 2 x:
TSNE1
Plot 2 y:
TSNE2
TSNE2
TSNE1
TSNE2
TSNE1
CD14+

D: PUBLICATIONS AND ABSTRACTS

Publications for Submission or Review

Green K, Hardie C.

'High-Risk Primary Biliary Cholangitis (PBC) has a Distinct Liver RNA Signature'

Green K, Barron E

'Self-perception in healthy aging'

Current Publications

McGovern N, Schlitzer A, Gunawan M, Jardine L, Shin A, Poyner E, **Green K**, Dickinson R, Wang XN, Low D, Best K, Covins S, Milne P, Pagan S, Aljefri K, Windebank M, Miranda-Saavedra D, Larbi A, Wasan PS, Duan K, Poidinger M, Bigley V, Ginhoux F, Collin M, Haniffa M.

'Human dermal CD14+ cells are a transient population of monocyte-derived macrophages.' *Immunity*, 2014.

Hardie C, **Green K**, Jopson L, Millar B, Innes B, Pagan S, Tiniakos D, Dyson J, Haniffa M, Bigley V, Jones DE, Brain J, Walker LJ.

‘Early Molecular Stratification of High-risk Primary Biliary Cholangitis.’ *EBioMedicine*, 2016.

Millar B, Wong LL, **Green K**, Resteu A, Kendrick S, Jones DE, Dyson J.

‘Autoimmune hepatitis patients with poor treatment response have a distinct liver transcriptome: implications for personalized therapy.’ *EASL LiverTree*, 2017

Green K, Pearce K, Sellar RS, Jardine L, Nicolson PL, Nagra S, Bigley V, Jackson G, Dickinson AM, Thomson K, Mackinnon S, Craddock C, Peggs KS, Collin M.

‘Impact of Alemtuzumab Scheduling on Graft-versus-Host Disease after Unrelated Donor Fludarabine and Melphalan Allografts.’ *Biology of Blood and Marrow Transplantation*, 2017.

Crossland RE, Norden J, Juric MK, **Green K**, Pearce KF, Lendrem C, Greinix HT, Dickinson AM.

‘Expression of Serum microRNAs is Altered During Acute Graft-versus-Host Disease’ *Frontiers in Immunology*, 2017

Bigley V, Maisuria S, Cytlak U, Jardine L, Care MA, **Green K**, Gunawan M, Milne P, Dickinson R, Wiscombe S, Parry D, Doffinger R, Laurence A, Fonseca C, Stoevesandt O, Gennery A, Cant A, Tooze R, Collin M.

‘Biallelic interferon regulatory factor 8 mutation: A complex immunodeficiency syndrome with dendritic cell deficiency, monocytopenia, and immune dysregulation’ *Journal of Allergy and Clinical Immunology*, 2017

Abstracts

Green K, Sellar R, Jardine L, Ward J, Ferguson P, Nicolson P, Pearce K, Bigley V, Jackson G, Nagra S, Dickinson AM, Thomson K, Mackinnon S, Craddock C, Collin MP, Peggs KS

‘Defining the Optimal Dose of Alemtuzumab in Unrelated Donor Reduced Intensity Allografts: A UK Retrospective Study.’ *In: 55th annual American Society of Hematology meeting.* 2013, New Orleans, USA.

Laverick O, Publicover A, Jardine L, **Green K**, Potter A, Jackson GH, Collin M.

‘Synergy of Unrelated Donor and Full Intensity Conditioning Breaks the Control of Graft Versus Host Disease By Alemtuzumab’ in: *56th ASH Annual Meeting and Exposition*. 2014, San Francisco, California: American Society of Hematology.

McGovern N, Schlitzer A, Gunawan M, Jardine L, Shin A, Poyner E, **Green K**, Dickinson R, Wang XN, Low D, Best K, Covins S, Milne P, Pagan S, Aljefri K, Windebank M, Miranda-Saavedra D, Wasan P, Kaibo D, Poidinger M, Bigley V, Ginhoux F, Collin M, Haniffa M.

‘Human tissue mononuclear phagocyte system revisited’ *In: British Society for Immunology Annual Congress*. 2014, Brighton, UK: Wiley-Blackwell Publishing Ltd.

Crossland RE, **Green K**, Bacon C, Rand V.

'Direct digital profiling of multiplexed mRNA expression from degraded formalin fixed paraffin embedded aggressive paediatric B-cell lymphoma tumour tissue'. *In: Fifth International Symposium on Childhood, Adolescent and Young Adult Non-Hodgkin Lymphoma*. 2015, Varese, Italy: Wiley.

Presentations

'Current usages of high-throughput screening techniques' *Faculty of Medical Sciences, Newcastle University*, 2013

'Optimising the dose of Alemtuzumab in matched unrelated donor reduced intensity haematopoietic stem cell transplants: a three centre UK study'. *American Society of Haematology Annual Meeting*, 2013

'A comparison of primary and cultured human DCs using digital multiplexed technology'. *Irish Association of Cancer Research Annual Meeting*. 2014

‘NanoString technology for detection of expression changes in primary and cultured dendritic cells’. *NanoString International User Group Meeting, Frankfurt*. 2015

‘Risk stratification for Cutaneous Squamous Cell Carcinoma: Using Machine Learning to Identify Patients at Increased Risk of Metastatic Disease’. *British Society for Investigative Dermatology*, 2017

References

10x Genomics [WWW Document], n.d. URL https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons (accessed 10.11.17).

Abcam, 2014. Flow Cytometry Guide.

Akira, S., Uematsu, S., Takeuchi, O., 2006. Pathogen Recognition and Innate Immunity. Cell 124, 783–801. https://doi.org/10.1016/j.cell.2006.02.015

Aldo, P.B., Craveiro, V., Guller, S., Mor, G., 2013. Effect of culture conditions on the phenotype of THP-1 monocyte cell line. Am. J. Reprod. Immunol. N. Y. N 1989 70, 80–86. https://doi.org/10.1111/aji.12129

Ancuta, P., Liu, K.-Y., Misra, V., Wacleche, V., Gosselin, A., Zhou, X., Gabuzda, D., 2009. Transcriptional profiling reveals developmental relationship and distinct biological functions of CD16+ and CD16- monocyte subsets. BMC Genomics 10, 403. https://doi.org/10.1186/1471-2164-10-403

Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169. https://doi.org/10.1093/bioinformatics/btu638

Andrews, T.S., Hemberg, M., 2017. Modelling dropouts for feature selection in scRNASeq experiments. bioRxiv 065094. https://doi.org/10.1101/065094

Arzalluz-Luque, Á., Devailly, G., Mantsoki, A., Joshi, A., 2017. Delineating biological and technical variance in single cell expression data. Int. J. Biochem. Cell Biol. 90, 161–166. https://doi.org/10.1016/j.biocel.2017.07.006

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29. https://doi.org/10.1038/75556

Audiger, C., Rahman, M.J., Yun, T.J., Tarbell, K.V., Lesage, S., 2017. The Importance of Dendritic Cells in Maintaining Immune Tolerance. J. Immunol. 198, 2223–2231. https://doi.org/10.4049/jimmunol.1601629

Auffray, C., Michael H. Sieweke, Geissmann, F., 2009. Blood Monocytes: Development, Heterogeneity, and Relationship with Dendritic Cells. Annu. Rev. Immunol. 27, 669–692. https://doi.org/10.1146/annurev.immunol.021908.132557

Austyn, J.M., 1998. Dendritic cells. : Current Opinion in Hematology. LWW 5, 3–15.

Autenrieth, S.E., Grimm, S., Rittig, S.M., Grünebach, F., Gouttefangeas, C., Bühring, H.-J., 2015. Profiling of primary peripheral blood- and monocyte-derived dendritic cells using monoclonal antibodies from the HLDA10 Workshop in Wollongong, Australia. Clin. Transl. Immunol. 4, e50. https://doi.org/10.1038/cti.2015.29

Bachem, A., Güttler, S., Hartung, E., Ebstein, F., Schaefer, M., Tannert, A., Salama, A., Movassaghi, K., Opitz, C., Mages, H.W., Henn, V., Kloetzel, P.-M., Gurka, S., Kroczek, R.A., 2010. Superior antigen cross-presentation and XCR1 expression define human CD11c+CD141+ cells as homologues of mouse CD8+ dendritic cells. J. Exp. Med. 207, 1273–1281. https://doi.org/10.1084/jem.20100348

Bacher, R., Kendziorski, C., 2016. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol. 17. https://doi.org/10.1186/s13059-016-0927-y

Bajaña, S., Turner, S., Paul, J., Ainsua-Enrich, E., Kovats, S., 2016. IRF4 and IRF8 Act in CD11c+ Cells To Regulate Terminal Differentiation of Lung Tissue Dendritic Cells. J. Immunol. 196, 1666–1677. https://doi.org/10.4049/jimmunol.1501870

Banchereau, J., Steinman, R.M., 1998. Dendritic cells and the control of immunity. Nature 392, 245–252.

Belge, K.-U., Dayyani, F., Horelt, A., Siedlar, M., Frankenberger, M., Frankenberger, B., Espevik, T., Ziegler-Heitbrock, L., 2002. The Proinflammatory CD14+CD16+DR++ Monocytes Are a Major Source of TNF. J. Immunol. 168, 3536–3542. https://doi.org/10.4049/jimmunol.168.7.3536

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300. https://doi.org/10.2307/2346101

Bigley, V., Haniffa, M., Doulatov, S., Wang, X.-N., Dickinson, R., McGovern, N., Jardine, L., Pagan, S., Dimmick, I., Chua, I., Wallis, J., Lordan, J., Morgan, C., Kumararatne, D.S., Doffinger, R., Burg, M. van der, Dongen, J. van, Cant, A., Dick, J.E., Hambleton, S., Collin, M., 2011. The human syndrome of dendritic cell, monocyte, B and NK lymphoid deficiency. J. Exp. Med. 208, 227–234. https://doi.org/10.1084/jem.20101459

Bigley, V., Maisuria, S., Cytlak, U., Jardine, L., Care, M.A., Green, K., Gunawan, M., Milne, P., Dickinson, R., Wiscombe, S., Parry, D., Doffinger, R., Laurence, A., Fonseca, C., Stoevesandt, O., Gennery, A., Cant, A., Tooze, R., Simpson, A.J., Hambleton, S., Savic, S., Doody, G., Collin, M., 2017. Biallelic interferon regulatory factor 8 mutation: A complex immunodeficiency syndrome with dendritic cell deficiency, monocytopenia, and immune dysregulation. J. Allergy Clin. Immunol. https://doi.org/10.1016/j.jaci.2017.08.044

Birnberg, T., Bar-On, L., Sapoznikov, A., Caton, M.L., Cervantes-Barragán, L., Makia, D., Krauthgamer, R., Brenner, O., Ludewig, B., Brockschnieder, D., Riethmacher, D., Reizis, B., Jung, S., 2008. Lack of Conventional Dendritic Cells Is Compatible with Normal Development and T Cell Homeostasis, but Causes Myeloid Proliferative Syndrome. Immunity 29, 986–997. https://doi.org/10.1016/j.immuni.2008.10.012

Bluestone, J.A., Abbas, A.K., 2003. Natural versus adaptive regulatory T cells. Nat. Rev. Immunol. 3, 253–257. https://doi.org/10.1038/nri1032

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boltjes, A., van Wijk, F., 2014. Human Dendritic Cell Functional Specialization in Steady-State and Inflammation. Front. Immunol. 5. https://doi.org/10.3389/fimmu.2014.00131

Borregaard, N., 2010. Neutrophils, from marrow to microbes. Immunity 33, 657–670. https://doi.org/10.1016/j.immuni.2010.11.011

Breton, G., Zheng, S., Valieris, R., Silva, I.T. da, Satija, R., Nussenzweig, M.C., 2016. Human dendritic cells (DCs) are derived from distinct circulating precursors that are precommitted to become CD1c+ or CD141+ DCs. J. Exp. Med. jem.20161135. https://doi.org/10.1084/jem.20161135

Brocker, T., Riedinger, M., Karjalainen, K., 1997. Targeted Expression of Major Histocompatibility Complex (MHC) Class II Molecules Demonstrates that Dendritic Cells Can Induce Negative but Not Positive Selection of Thymocytes In Vivo. J. Exp. Med. 185, 541–550. https://doi.org/10.1084/jem.185.3.541

Bullard, J.H., Purdom, E., Hansen, K.D., Dudoit, S., 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11, 94. https://doi.org/10.1186/1471-2105-11-94

Cannarile, M.A., Decanis, N., Meerwijk, J.P.M. van, Brocker, T., 2004. The Role of Dendritic Cells in Selection of Classical and Nonclassical CD8+ T Cells In Vivo. J. Immunol. 173, 4799–4805. https://doi.org/10.4049/jimmunol.173.8.4799

Carotta, S., Dakic, A., D'Amico, A., Pang, S.H.M., Greig, K.T., Nutt, S.L., Wu, L., 2010. The Transcription Factor PU.1 Controls Dendritic Cell Development and Flt3 Cytokine Receptor Expression in a Dose-Dependent Manner. Immunity 32, 628–641. https://doi.org/10.1016/j.immuni.2010.05.005

Castell-Rodríguez, A., Piñón-Zárate, G., Herrera-Enríquez, M., Jarquín-Yáñez, K., Medina-Solares, I., 2017. Dendritic Cells: Location, Function, and Clinical Implications. https://doi.org/10.5772/intechopen.68352

Ceredig, R., Rolink, A.G., Brown, G., 2009. Models of haematopoiesis: seeing the wood for the trees. Nat. Rev. Immunol. 9, 293–300. https://doi.org/10.1038/nri2525

Chen, W., Jin, W., Hardegen, N., Lei, K.-J., Li, L., Marinos, N., McGrady, G., Wahl, S.M., 2003. Conversion of peripheral CD4+CD25- naive T cells to CD4+CD25+ regulatory T cells by TGF-beta induction of transcription factor Foxp3. J. Exp. Med. 198, 1875–1886. https://doi.org/10.1084/jem.20030152

Collin, M., Bigley, V., 2018. Human dendritic cell subsets: an update. Immunology imm.12888.

Collin, M., Bigley, V., 2016. Monocyte, Macrophage, and Dendritic Cell Development: the Human Perspective. Microbiol. Spectr. 4. https://doi.org/10.1128/microbiolspec.MCHD-0015-2015

Collin, M., Bigley, V., Haniffa, M., Hambleton, S., 2011. Human dendritic cell deficiency: the missing ID? Nat. Rev. Immunol. 11, 575–583. https://doi.org/10.1038/nri3046

Collin, M., McGovern, N., Haniffa, M., 2013. Human dendritic cell subsets. Immunology 140, 22–30. https://doi.org/10.1111/imm.12117

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 17. https://doi.org/10.1186/s13059-016-0881-8

Crozat, K., Guiton, R., Contreras, V., Feuillet, V., Dutertre, C.-A., Ventre, E., Manh, T.-P.V., Baranek, T., Storset, A.K., Marvel, J., Boudinot, P., Hosmalin, A., Schwartz-Cornil, I., Dalod, M., 2010. The XC chemokine receptor 1 is a conserved selective marker of mammalian cells homologous to mouse CD8α+ dendritic cells. J. Exp. Med. 207, 1283–1292. https://doi.org/10.1084/jem.20100223

Davis, S., Meltzer, P.S., 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23, 1846–1847. https://doi.org/10.1093/bioinformatics/btm254

Delves, P.J., Martin, S.J., Burton, D.R., Roitt, I.M., 2011. Roitt's Essential Immunology. John Wiley & Sons.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Doulatov, S., Notta, F., Eppert, K., Nguyen, L.T., Ohashi, P.S., Dick, J.E., 2010. Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic

cells in early lymphoid development. Nat. Immunol. 11, 585–593. https://doi.org/10.1038/ni.1889

Durand, M., Segura, E., 2015. The Known Unknowns of the Human Dendritic Cell Network. Front. Immunol. 6. https://doi.org/10.3389/fimmu.2015.00129

Dzionek, A., Fuchs, A., Schmidt, P., Cremer, S., Zysk, M., Miltenyi, S., Buck, D.W., Schmitz, J., 2000. BDCA-2, BDCA-3, and BDCA-4: Three Markers for Distinct Subsets of Dendritic Cells in Human Peripheral Blood. J. Immunol. 165, 6037–6046. https://doi.org/10.4049/jimmunol.165.11.6037

Edwards, A.D., Diebold, S.S., Slack, E.M.C., Tomizawa, H., Hemmi, H., Kaisho, T., Akira, S., Sousa, C.R. e, 2003. Toll-like receptor expression in murine DC subsets: lack of TLR7 expression by CD8α+ DC correlates with unresponsiveness to imidazoquinolines. Eur. J. Immunol. 33, 827–833. https://doi.org/10.1002/eji.200323797

Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. http://CRAN.R-project.org/package=RColorBrewer, 2014.

Esashi, E., Bao, M., Wang, Y.-H., Cao, W., Liu, Y.-J., 2012. PACSIN1 regulates the TLR7/9-mediated type I interferon response in plasmacytoid dendritic cells. Eur. J. Immunol. 42, 573–579. https://doi.org/10.1002/eji.201142045

Falcon, S., Gentleman, R., 2007. Using GOstats to test gene lists for GO term association. Bioinformatics 23, 257–258. https://doi.org/10.1093/bioinformatics/btl567

François Bach, J., 2003. Regulatory T cells under scrutiny. Nat. Rev. Immunol. 3, 189–198. https://doi.org/10.1038/nri1026

Gallo, P.M., Gallucci, S., 2013. The Dendritic Cell Response to Classic, Emerging, and Homeostatic Danger Signals. Implications for Autoimmunity. Front. Immunol. 4. https://doi.org/10.3389/fimmu.2013.00138

Gargani, Y., 2012. Crash Course Haematology and Immunology. Elsevier Health Sciences.

Geissmann, F., Auffray, C., Palframan, R., Wirrig, C., Ciocca, A., Campisi, L., Narni-Mancinelli, E., Lauvau, G., 2008. Blood monocytes: distinct subsets, how they relate to dendritic cells, and their possible roles in the regulation of T-cell responses. Immunol. Cell Biol. 86, 398. https://doi.org/10.1038/icb.2008.19

Geissmann, F., Manz, M.G., Jung, S., Sieweke, M.H., Merad, M., Ley, K., 2010. Development of monocytes, macrophages, and dendritic cells. Science 327, 656–661. https://doi.org/10.1126/science.1178331

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J., 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80. https://doi.org/10.1186/gb-2004-5-10-r80

Ginhoux, F., Jung, S., 2014. Monocytes and macrophages: developmental pathways and tissue homeostasis. Nat. Rev. Immunol. 14, 392–404. https://doi.org/10.1038/nri3671

Grage-Griebenow, E., Flad, H.-D., Ernst, M., 2001. Heterogeneity of human peripheral blood monocyte subsets. J. Leukoc. Biol. 69, 11–20.

Guilliams, M., Dutertre, C.-A., Scott, C.L., McGovern, N., Sichien, D., Chakarov, S., Van Gassen, S., Chen, J., Poidinger, M., De Prijck, S., Tavernier, S.J., Low, I., Irac, S.E., Mattar, C.N., Sumatoh, H.R., Low, G.H.L., Chung, T.J.K., Chan, D.K.H., Tan, K.K., Hon, T.L.K., Fossum, E., Bogen, B., Choolani, M., Chan, J.K.Y., Larbi, A., Luche, H., Henri, S., Saeys, Y., Newell, E.W., Lambrecht, B.N., Malissen, B., Ginhoux, F., 2016.

Unsupervised High-Dimensional Analysis Aligns Dendritic Cells across Tissues and Species. Immunity 45, 669–684. https://doi.org/10.1016/j.immuni.2016.08.015

Häger, M., Cowland, J.B., Borregaard, N., 2010. Neutrophil granules in health and disease. J. Intern. Med. 268, 25–34. https://doi.org/10.1111/j.1365-2796.2010.02237.x

Haley, P.J., 2003. Species differences in the structure and function of the immune system. Toxicology 188, 49–71.

Hambleton, S., Salem, S., Bustamante, J., Bigley, V., Boisson-Dupuis, S., Azevedo, J., Fortin, A., Haniffa, M., Ceron-Gutierrez, L., Bacon, C., Menon, G., Trouillet, C., McDonald, D., Carey, P., Ginhoux, F., Alsina, L., Zumwalt, T.J., Kong, X., Kumararatne, D., Butler, K., Hubeau, M., Feinberg, J., Al-Muhsen, S., Cant, A., Abel, L., Chaussabel, D., Doffinger, R., Talesnik, E., Grumach, A., Duarte, A., Abarca, K., Moraes-Vasconcelos, D., Burk, D., Berghuis, A., Geissmann, F., Collin, M., Casanova, J.-L., Gros, P., 2011. Mutations in IRF8 and Human Dendritic Cell Immunodeficiency. N. Engl. J. Med. 365, 127–138. https://doi.org/10.1056/NEJMoa1100066

Hamey, F.K., Göttgens, B., 2017. Demystifying blood stem cell fates. Nat. Cell Biol. 19, 261. https://doi.org/10.1038/ncb3494

Han, A.P., 2015. Microwell Device for Single-Cell RNA Capture Shows Potential for High-Throughput Applications [WWW Document]. GenomeWeb. URL https://www.genomeweb.com/microarrays-multiplexing/microwell-device-single-cell-rna-capture-shows-potential-high-throughput (accessed 11.20.17).

Haniffa, M., Ginhoux, F., Wang, X.-N., Bigley, V., Abel, M., Dimmick, I., Bullock, S., Grisotto, M., Booth, T., Taub, P., Hilkens, C., Merad, M., Collin, M., 2009. Differential rates of replacement of human dermal dendritic cells and macrophages during hematopoietic stem cell transplantation. J. Exp. Med. 206, 371–385. https://doi.org/10.1084/jem.20081633

Haniffa, M., Gunawan, M., Jardine, L., 2015. Human skin dendritic cells in health and disease. J. Dermatol. Sci. 77, 85–92. https://doi.org/10.1016/j.jdermsci.2014.08.012

Haniffa, M., Shin, A., Bigley, V., McGovern, N., Teo, P., See, P., Wasan, P.S., Wang, X.-N., Malinarich, F., Malleret, B., Larbi, A., Tan, P., Zhao, H., Poidinger, M., Pagan, S., Cookson, S., Dickinson, R., Dimmick, I., Jarrett, R.F., Renia, L., Tam, J., Song, C., Connolly, J., Chan, J.K.Y., Gehring, A., Bertoletti, A., Collin, M., Ginhoux, F., 2012. Human Tissues Contain CD141hi Cross-Presenting Dendritic Cells with Functional Homology to Mouse CD103+ Nonlymphoid Dendritic Cells. Immunity 37, 60–73. https://doi.org/10.1016/j.immuni.2012.04.012

Haque, A., Engel, J., Teichmann, S.A., Lönnberg, T., 2017. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 9, 75. https://doi.org/10.1186/s13073-017-0467-4

Hawiger, D., Inaba, K., Dorsett, Y., Guo, M., Mahnke, K., Rivera, M., Ravetch, J.V., Steinman, R.M., Nussenzweig, M.C., 2001. Dendritic Cells Induce Peripheral T Cell Unresponsiveness under Steady State Conditions in Vivo. J. Exp. Med. 194, 769–780. https://doi.org/10.1084/jem.194.6.769

He, Z., Riva, M., Björk, P., Swärd, K., Mörgelin, M., Leanderson, T., Ivars, F., 2016. CD14 Is a Co-Receptor for TLR4 in the S100A9-Induced Pro-Inflammatory Response in Monocytes. PLOS ONE 11, e0156377. https://doi.org/10.1371/journal.pone.0156377

Heath, W.R., Carbone, F.R., 2001. Cross-presentation in viral immunity and self-tolerance. Nat. Rev. Immunol. 1, 126. https://doi.org/10.1038/35100512

Hémont, C., Neel, A., Heslan, M., Braudeau, C., Josien, R., 2013. Human blood mDC subsets exhibit distinct TLR repertoire and responsiveness. J. Leukoc. Biol. 93, 599–609. https://doi.org/10.1189/jlb.0912452

Hogquist, K.A., Baldwin, T.A., Jameson, S.C., 2005. Central tolerance: learning self-control in the thymus. Nat. Rev. Immunol. 5, 772–782. https://doi.org/10.1038/nri1707

Hotelling, H., 1933. Analysis of a Complex of Statistical Variables Into Principal Components. Warwick & York.

Ilicic, T., Kim, J.K., Kolodziejczyk, A.A., Bagger, F.O., McCarthy, D.J., Marioni, J.C., Teichmann, S.A., 2016. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 17, 29. https://doi.org/10.1186/s13059-016-0888-1

Illumina, 2011. http://www.illumina.com/documents/products/datasheets/datasheet_gene_exp_analysis.pdf - Google Search [WWW Document]. URL https://www.google.co.uk/?gfe_rd=cr&ei=HikjV_CrM6izgAaejLv4DQ#q=http:%2F%2Fwww.illumina.com%2Fdocuments%2Fproducts%2Fdatasheets%2Fdatasheet_gene_exp_analysis.pdf (accessed 4.29.16).

Ishikawa, F., Niiro, H., Iino, T., Yoshida, S., Saito, N., Onohara, S., Miyamoto, T., Minagawa, H., Fujii, S., Shultz, L.D., Harada, M., Akashi, K., 2007. The developmental program of human dendritic cells is operated independently of conventional myeloid and lymphoid pathways. Blood 110, 3591–3660. https://doi.org/10.1182/blood-2007-02-071613

Ito, T., Amakawa, R., Inaba, M., Hori, T., Ota, M., Nakamura, K., Takebayashi, M., Miyaji, M., Yoshimura, T., Inaba, K., Fukuhara, S., 2004. Plasmacytoid Dendritic Cells Regulate Th Cell Responses through OX40 Ligand and Type I IFNs. J. Immunol. 172, 4253–4259. https://doi.org/10.4049/jimmunol.172.7.4253

Jin, J.-O., Zhang, W., Du, J., Yu, Q., 2014. BDCA1-Positive Dendritic Cells (DCs) Represent a Unique Human Myeloid DC Subset That Induces Innate and Adaptive Immune Responses to Staphylococcus aureus Infection. Infect. Immun. 82, 4466–4476. https://doi.org/10.1128/IAI.01851-14

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127. https://doi.org/10.1093/biostatistics/kxj037

Jongbloed, S.L., Kassianos, A.J., McDonald, K.J., Clark, G.J., Ju, X., Angel, C.E., Chen, C.-J.J., Dunbar, P.R., Wadley, R.B., Jeet, V., Vulink, A.J.E., Hart, D.N.J., Radford, K.J., 2010. Human CD141+ (BDCA-3)+ dendritic cells (DCs) represent a unique myeloid DC subset that cross-presents necrotic cell antigens. J. Exp. Med. 207. https://doi.org/10.1084/jem.20092140

Karamitros, D., Stoilova, B., Aboukhalil, Z., Hamey, F., Reinisch, A., Samitsch, M., Quek, L., Otto, G., Repapi, E., Doondeea, J., Usukhbayar, B., Calvo, J., Taylor, S., Goardon, N., Six, E., Pflumio, F., Porcher, C., Majeti, R., Göttgens, B., Vyas, P., 2017. Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. Nat. Immunol. https://doi.org/10.1038/s41590-017-0001-2

Kawamoto, H., Wada, H., Katsura, Y., 2010. A revised scheme for developmental pathways of hematopoietic cells: the myeloid-based model. Int. Immunol. 22, 65–70. https://doi.org/10.1093/intimm/dxp125

Kharchenko, P.V., Silberstein, L., Scadden, D.T., 2014. Bayesian approach to single-cell differential expression analysis. Nat. Methods 11, 740–742. https://doi.org/10.1038/nmeth.2967

Kiertscher, S.M., Roth, M.D., 1996. Human CD14+ leukocytes acquire the phenotype and function of antigen-presenting dendritic cells when cultured in GM-CSF and IL-4. J. Leukoc. Biol. 59, 208–218.

Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., Hemberg, M., 2016. SC3 - consensus clustering of single-cell RNA-Seq data. bioRxiv 036558. https://doi.org/10.1101/036558

Klechevsky, E., Morita, R., Liu, M., Cao, Y., Coquery, S., Thompson-Snipes, L., Briere, F., Chaussabel, D., Zurawski, G., Palucka, A.K., Reiter, Y., Banchereau, J., Ueno, H., 2008. Functional Specializations of Human Epidermal Langerhans Cells and CD14+ Dermal Dendritic Cells. Immunity 29, 497–510. https://doi.org/10.1016/j.immuni.2008.07.013

Kolaczkowska, E., Kubes, P., 2013. Neutrophil recruitment and function in health and inflammation. Nat. Rev. Immunol. 13, 159–175. https://doi.org/10.1038/nri3399

Kondo, M., Weissman, I.L., Akashi, K., 1997. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. Cell 91, 661–672.

Lai, A.Y., Kondo, M., 2006. Asymmetrical lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors. J. Exp. Med. 203, 1867–1873. https://doi.org/10.1084/jem.20060697

Lavin, Y., Winter, D., Blecher-Gonen, R., David, E., Keren-Shaul, H., Merad, M., Jung, S., Amit, I., 2014. Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. Cell 159, 1312–1326. https://doi.org/10.1016/j.cell.2014.11.018

Lee, J., Breton, G., Oliveira, T.Y.K., Zhou, Y.J., Aljoufi, A., Puhr, S., Cameron, M.J., Sékaly, R.-P., Nussenzweig, M.C., Liu, K., 2015. Restricted dendritic cell and monocyte progenitors in human cord blood and bone marrow. J. Exp. Med. 212, 385–399. https://doi.org/10.1084/jem.20141442

Lee, J., Zhou, Y.J., Ma, W., Zhang, W., Aljoufi, A., Luh, T., Lucero, K., Liang, D., Thomsen, M., Bhagat, G., Shen, Y., Liu, K., 2017. Lineage specification of human dendritic cells is marked by IRF8 expression in hematopoietic stem cells and multipotent progenitors. Nat. Immunol. 18, 877–888. https://doi.org/10.1038/ni.3789

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., Storey, J.D., 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28, 882–883. https://doi.org/10.1093/bioinformatics/bts034

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., Mesirov, J.P., 2011. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

Lotze, M.T., Thomson, A.W., 2001. Dendritic Cells: Biology and Clinical Applications. Academic Press.

Loughland, J.R., Minigo, G., Sarovich, D.S., Field, M., Tipping, P.E., Oca, M.M. de, Piera, K.A., Amante, F.H., Barber, B.E., Grigg, M.J., William, T., Good, M.F., Doolan, D.L., Engwerda, C.R., Anstey, N.M., McCarthy, J.S., Woodberry, T., 2017. Plasmacytoid dendritic cells appear inactive during sub-microscopic Plasmodium falciparum

blood-stage infection, yet retain their ability to respond to TLR stimulation. Sci. Rep. 7, 2596. https://doi.org/10.1038/s41598-017-02096-2

Lun, A.T.L., McCarthy, D.J., Marioni, J.C., 2016. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research 5, 2122. https://doi.org/10.12688/f1000research.9501.2

Maaten, L. van der, 2014. Accelerating t-SNE using Tree-Based Algorithms. J. Mach. Learn. Res. 15, 3221–3245.

Maaten, L. van der, Hinton, G., 2008. Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

MacDonald, K.P.A., Munster, D.J., Clark, G.J., Dzionek, A., Schmitz, J., Hart, D.N.J., 2002. Characterization of human blood dendritic cell subsets. Blood 100, 4512–4520. https://doi.org/10.1182/blood-2001-11-0097

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., Trombetta, J.J., Weitz, D.A., Sanes, J.R., Shalek, A.K., Regev, A., McCarroll, S.A., 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Mak, T.W., Saunders, M.E., 2005. The Immune Response: Basic and Clinical Principles. Academic Press.

Malissen, B., Tamoutounour, S., Henri, S., 2014. The origins and functions of dendritic cells and macrophages in the skin. Nat. Rev. Immunol. 14, 417–428. https://doi.org/10.1038/nri3683

Mayuzumi, N., Matsushima, H., Takashima, A., 2009. IL-33 Promotes DC Development in BM Culture by Triggering GM-CSF Production. Eur. J. Immunol. 39, 3331–3342. https://doi.org/10.1002/eji.200939472

McCarthy, D.J., Campbell, K.R., Lun, A.T.L., Wills, Q.F., 2017. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinforma. Oxf. Engl. 33, 1179–1186. https://doi.org/10.1093/bioinformatics/btw777

McGovern, N., Chan, J.K.Y., Ginhoux, F., 2015. Dendritic cells in humans—from fetus to adult. Int. Immunol. 27, 65–72. https://doi.org/10.1093/intimm/dxu091

McGovern, N., Schlitzer, A., Gunawan, M., Jardine, L., Shin, A., Poyner, E., Green, K., Dickinson, R., Wang, X., Low, D., Best, K., Covins, S., Milne, P., Pagan, S., Aljefri, K., Windebank, M., Miranda-Saavedra, D., Larbi, A., Wasan, P.S., Duan, K., Poidinger, M., Bigley, V., Ginhoux, F., Collin, M., Haniffa, M., 2014. Human Dermal CD14+ Cells Are a Transient Population of Monocyte-Derived Macrophages. Immunity 41, 465–477. https://doi.org/10.1016/j.immuni.2014.08.006

McLellan, A.D., Heiser, A., Sorg, R.V., Fearnley, D.B., Hart, D.N., 1998. Dermal dendritic cells associated with T lymphocytes in normal human skin display an activated phenotype. J. Invest. Dermatol. 111, 841–849. https://doi.org/10.1046/j.1523-1747.1998.00375.x

Mellman, I., Steinman, R.M., 2001. Dendritic Cells: Specialized and Regulated Antigen Processing Machines. Cell 106, 255–258. https://doi.org/10.1016/S0092-8674(01)00449-4

Merad, M., Sathe, P., Helft, J., Miller, J., Mortha, A., 2013. The Dendritic Cell Lineage: Ontogeny and Function of Dendritic Cells and Their Subsets in the Steady State and the Inflamed Setting. Annu. Rev. Immunol. 31. https://doi.org/10.1146/annurev-immunol-020711-074950

Mestas, J., Hughes, C.C.W., 2004. Of Mice and Not Men: Differences between Mouse and Human Immunology. J. Immunol. 172, 2731–2738. https://doi.org/10.4049/jimmunol.172.5.2731

Mildner, A., Jung, S., 2014. Development and Function of Dendritic Cell Subsets. Immunity 40, 642–656. https://doi.org/10.1016/j.immuni.2014.04.016

Minoda, Y., Virshup, I., Leal Rojas, I., Haigh, O., Wong, Y., Miles, J.J., Wells, C.A., Radford, K.J., 2017. Human CD141+ Dendritic Cell and CD1c+ Dendritic Cell Undergo Concordant Early Genetic Programming after Activation in Humanized Mice In Vivo. Front. Immunol. 8. https://doi.org/10.3389/fimmu.2017.01419

Mittag, D., Proietto, A.I., Loudovaris, T., Mannering, S.I., Vremec, D., Shortman, K., Wu, L., Harrison, L.C., 2011. Human Dendritic Cell Subsets from Spleen and Blood Are Similar in Phenotype and Function but Modified by Donor Health Status. J. Immunol. 186, 6207–6217. https://doi.org/10.4049/jimmunol.1002632

Monaco, A.P., 2003. Chimerism in organ transplantation: conflicting experiments and clinical observations. Transplantation 75, 13S-16S. https://doi.org/10.1097/01.TP.0000067945.90241.F4

Morse, M.A., Lyerly, H.K., 2002. Isolation and Culture of Dendritic Cells, in: Human Cell Culture, Human Cell Culture. Springer, Dordrecht, pp. 171–191. https://doi.org/10.1007/0-306-46886-7_7

Murphy, K., Weaver, C., 2016. Janeway's Immunobiology, 9th edition. Garland Science.

Naik, S.H., Perié, L., Swart, E., Gerlach, C., van Rooij, N., de Boer, R.J., Schumacher, T.N., 2013. Diverse and heritable lineage imprinting of early haematopoietic progenitors. Nature 496, 229–232. https://doi.org/10.1038/nature12013

Nair, S., Archer, G.E., Tedder, T.F., 2012. ISOLATION AND GENERATION OF HUMAN DENDRITIC CELLS. Curr. Protoc. Immunol. Ed. John E Coligan Al 0 7, Unit7.32. https://doi.org/10.1002/0471142735.im0732s99

Nairn, R., 2002. Immunology for medical students. PA: Mosby, Philadelphia.

Ng, S.Y.-M., Yoshida, T., Zhang, J., Georgopoulos, K., 2009. Genome-wide lineage-specific transcriptional networks underscore Ikaros-dependent lymphoid priming in hematopoietic stem cells. Immunity 30, 493–507. https://doi.org/10.1016/j.immuni.2009.01.014

Ni, K., O'Neill, H.C., 1997. The role of dendritic cells in T cell activation. Immunol. Cell Biol. 75, 223–230.

Notta, F., Zandi, S., Takayama, N., Dobson, S., Gan, O.I., Wilson, G., Kaufmann, K.B., McLeod, J., Laurenti, E., Dunant, C.F., McPherson, J.D., Stein, L.D., Dror, Y., Dick, J.E., 2016. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science 351, aab2116. https://doi.org/10.1126/science.aab2116

Oehler, M.K., Bicknell, R., 2000. The promise of anti-angiogenic cancer therapy. Br. J. Cancer 82, 749–752. https://doi.org/10.1054/bjoc.1999.0991

O'Keeffe, M., Mok, W.H., Radford, K.J., 2015. Human dendritic cell subsets and function in health and disease. Cell. Mol. Life Sci. 72, 4309–4325. https://doi.org/10.1007/s00018-015-2005-0

Papalexi, E., Satija, R., 2017. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat. Rev. Immunol.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., David, E., Cohen, N., Lauridsen, F.K.B., Haas, S., Schlitzer, A., Mildner, A., Ginhoux, F., Jung, S., Trumpp, A., Porse, B.T., Tanay, A.,

Amit, I., 2015. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell 163, 1663–1677. https://doi.org/10.1016/j.cell.2015.11.013

Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., Sandberg, R., 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods 10, 1096–1098. https://doi.org/10.1038/nmeth.2639

Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P., Ramalingam, N., Sun, G., Thu, M., Norris, M., Lebofsky, R., Toppani, D., Kemp Ii, D.W., Wong, M., Clerkson, B., Jones, B.N., Wu, S., Knutsson, L., Alvarado, B., Wang, J., Weaver, L.S., May, A.P., Jones, R.C., Unger, M.A., Kriegstein, A.R., West, J.A.A., 2014. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat. Biotechnol. 32, 1053–1058. https://doi.org/10.1038/nbt.2967

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, 2012.

R N Germain, Margulies, and D.H., 1993. The Biochemistry and Cell Biology of Antigen Processing and Presentation. Annu. Rev. Immunol. 11, 403–450. https://doi.org/10.1146/annurev.iy.11.040193.002155

Rahman, M., 2009. Introduction to Flow Cytometry.

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., Schroth, G.P., Sandberg, R., 2012. Full-Length mRNA-Seq from single cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. 30, 777–782. https://doi.org/10.1038/nbt.2282

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., Betel, D., 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 14, 3158. https://doi.org/10.1186/gb-2013-14-9-r95

Rathinam, C., Geffers, R., Yücel, R., Buer, J., Welte, K., Möröy, T., Klein, C., 2005. The transcriptional repressor Gfi1 controls STAT3-dependent dendritic cell development and function. Immunity 22, 717–728. https://doi.org/10.1016/j.immuni.2005.04.007

Risso, D., Ngai, J., Speed, T.P., Dudoit, S., 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. 32. https://doi.org/10.1038/nbt.2931

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47–e47. https://doi.org/10.1093/nar/gkv007

Robbins, S.H., Walzer, T., Dembélé, D., Thibault, C., Defays, A., Bessou, G., Xu, H., Vivier, E., Sellars, M., Pierre, P., Sharp, F.R., Chan, S., Kastner, P., Dalod, M., 2008. Novel insights into the relationships between dendritic cell subsets in human and mouse revealed by genome-wide expression profiling. Genome Biol. 9, R17. https://doi.org/10.1186/gb-2008-9-1-r17

Rockefeller Institute, 2014. Introduction to Dendritic Cells.

Rowley, T., 2015. Flow Cytometry - A Survey and the Basics. Mater. Methods.

Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A., Teichmann, S.A., 2017. The Human Cell Atlas: from vision to reality. Nat. News 550, 451. https://doi.org/10.1038/550451a

Salem, S., Langlais, D., Lefebvre, F., Bourque, G., Bigley, V., Haniffa, M., Casanova, J.-L., Burk, D., Berghuis, A., Butler, K.M., Leahy, T.R., Hambleton, S., Gros, P., 2014. Functional characterization of the human dendritic cell immunodeficiency associated with the IRF8K108E mutation. Blood 124, 1894–1904. https://doi.org/10.1182/blood-2014-04-570879

Sathe, P., Vremec, D., Wu, L., Corcoran, L., Shortman, K., 2013. Convergent differentiation: myeloid and lymphoid pathways to murine plasmacytoid dendritic cells. Blood 121, 11–19. https://doi.org/10.1182/blood-2012-02-413336

Sato, K., Fujita, S., 2007. Dendritic Cells-Nature and Classification. Allergol. Int. 56, 183–191. https://doi.org/10.2332/allergolint.R-06-139

Satpathy, A., Murphy, K.M., Wumesh, K., 2011. Transcription Factor Networks in Dendritic Cell Development. Semin. Immunol. 23, 388–397. https://doi.org/10.1016/j.smim.2011.08.009

Schlitzer, A., Sivakamasundari, V., Chen, J., Sumatoh, H.R.B., Schreuder, J., Lum, J., Malleret, B., Zhang, S., Larbi, A., Zolezzi, F., Renia, L., Poidinger, M., Naik, S., Newell, E.W., Robson, P., Ginhoux, F., 2015. Identification of cDC1- and cDC2-committed DC progenitors reveals early lineage priming at the common DC progenitor stage in the bone marrow. Nat. Immunol. 16, 718–728. https://doi.org/10.1038/ni.3200

Schraml, B.U., van Blijswijk, J., Zelenay, S., Whitney, P.G., Filby, A., Acton, S.E., Rogers, N.C., Moncaut, N., Carvajal, J.J., Reis e Sousa, C., 2013. Genetic tracing via DNGR-1 expression history defines dendritic cells as a hematopoietic lineage. Cell 154, 843–858. https://doi.org/10.1016/j.cell.2013.07.014

See, P., Dutertre, C.-A., Chen, J., Günther, P., McGovern, N., Irac, S.E., Gunawan, M., Beyer, M., Händler, K., Duan, K., Sumatoh, H.R.B., Ruffin, N., Jouve, M., Gea-Mallorquí, E., Hennekam, R.C.M., Lim, T., Yip, C.C., Wen, M., Malleret, B., Low, I., Shadan, N.B., Fen, C.F.S., Tay, A., Lum, J., Zolezzi, F., Larbi, A., Poidinger, M., Chan, J.K.Y., Chen, Q., Renia, L., Haniffa, M., Benaroch, P., Schlitzer, A., Schultze, J.L., Newell, E.W., Ginhoux, F., 2017. Mapping the human DC lineage through the integration of high-dimensional techniques. Science eaag3009. https://doi.org/10.1126/science.aag3009

Segura, E., Amigorena, S., 2013. Inflammatory dendritic cells in mice and humans. Trends Immunol. 34, 440–445. https://doi.org/10.1016/j.it.2013.06.001

Segura, E., Touzot, M., Bohineust, A., Cappuccio, A., Chiocchia, G., Hosmalin, A., Dalod, M., Soumelis, V., Amigorena, S., 2013. Human Inflammatory Dendritic Cells Induce Th17 Cell Differentiation. Immunity 38, 336–348. https://doi.org/10.1016/j.immuni.2012.10.018

Shortman, K., Liu, Y.-J., 2002. Mouse and human dendritic cell subtypes. Nat. Rev. Immunol. 2, 151–161. https://doi.org/10.1038/nri746

Shurin, M.R., Salter, R.D., 2009. Dendritic Cells in Cancer. Springer Science & Business Media.

Smyth, G.K., 2005. limma: Linear Models for Microarray Data, in: Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A., Dudoit, S. (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Statistics for Biology and Health. Springer New York, pp. 397–420. https://doi.org/10.1007/0-387-29362-0_23

Spinelli, L., Carpentier, S., Montañana Sanchis, F., Dalod, M., Vu Manh, T.-P., 2015. BubbleGUM: automatic extraction of phenotype molecular signatures and comprehensive visualization of multiple Gene Set Enrichment Analyses. BMC Genomics 16, 814. https://doi.org/10.1186/s12864-015-2012-4

Spits, H., Couwenberg, F., Bakker, A.Q., Weijer, K., Uittenbogaart, C.H., 2000. Id2 and Id3 Inhibit Development of Cd34+ Stem Cells into Predendritic Cell (Pre-Dc)2 but Not into Pre-Dc1. J. Exp. Med. 192, 1775–1784.

Steinman, R.M., 1991. The Dendritic Cell System and its Role in Immunogenicity. Annu. Rev. Immunol. 9, 271–296. https://doi.org/10.1146/annurev.iy.09.040191.001415

Steinman, R.M., Hawiger, D., Nussenzweig, M.C., 2003. Tolerogenic Dendritic Cells. Annu. Rev. Immunol. 21, 685–711. https://doi.org/10.1146/annurev.immunol.21.120601.141040

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550.

Sun, S., Hood, M., Scott, L., Peng, Q., Mukherjee, S., Tung, J., Zhou, X., 2017. Differential expression analysis for RNAseq using Poisson mixed models. Nucleic Acids Res. 45, e106–e106. https://doi.org/10.1093/nar/gkx204

Sykes, M., 2001. Mixed chimerism and transplant tolerance. Immunity 14, 417–424.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K., Surani, M.A., 2009. mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods 6, 377–382. https://doi.org/10.1038/nmeth.1315

Timmerman, Levy, 1999. Dendritic Cell Vaccines for Cancer Immunotherapy. Annu. Rev. Med. 50, 507–529. https://doi.org/10.1146/annurev.med.50.1.507

Vander Lugt, B., Khan, A.A., Hackney, J.A., Agrawal, S., Lesch, J., Zhou, M., Lee, W.P., Park, S., Xu, M., DeVoss, J., Spooner, C.J., Chalouni, C., Delamarre, L., Mellman, I., Singh, H., 2014. Transcriptional programming of dendritic cells for enhanced MHC class II antigen presentation. Nat. Immunol. 15, 161–167. https://doi.org/10.1038/ni.2795

Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., Boch, T., Hofmann, W.-K., Ho, A.D., Huber, W., Trumpp, A., Essers, M.A.G., Steinmetz, L.M., 2017. Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. 19, 271–281. https://doi.org/10.1038/ncb3493

Venables, W.N., 2002. Modern Applied Statistics with S | W.N. Venables | Springer.

Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., Jardine, L., Dixon, D., Stephenson, E., Nilsson, E., Grundberg, I., McDonald, D., Filby, A., Li, W., De Jager, P.L., Rozenblatt-Rosen, O., Lane, A.A., Haniffa, M., Regev, A., Hacohen, N., 2017. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science 356. https://doi.org/10.1126/science.aah4573

Vincent Q. Vu (2011). ggbiplot: A ggplot2 based biplot. R package version 0.55. http://github.com/vqv/ggbiplot, 2011.

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63. https://doi.org/10.1038/nrg2484

Watson, J.V., 2004. Introduction to Flow Cytometry. Cambridge University Press.

Weider, E., 2003. Dendritic Cells: A Basic Review.

Wersto, R., 2014. Applications of Flow Cytometry.

Wickham, H., 2009a. ggplot2: Elegant Graphics for Data Analysis. Springer Science & Business Media.

Wickham, H., 2009b. Ggplot2: Elegant Graphics for Data Analysis, 2nd ed. Springer Publishing Company, Incorporated.

Williams, C.R., Baccarella, A., Parrish, J.Z., Kim, C.C., 2016. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics 17. https://doi.org/10.1186/s12859-016-0956-2

Williams, D.W., Byrd, D., Rubin, L.H., Anastos, K., Morgello, S., Berman, J.W., 2014. CCR2 on CD14+CD16+ monocytes is a biomarker of HIV-associated neurocognitive disorders. Neurol. Neuroimmunol. Neuroinflammation 1. https://doi.org/10.1212/NXI.0000000000000036

Wong, K.L., Tai, J.J.-Y., Wong, W.-C., Han, H., Sem, X., Yeap, W.-H., Kourilsky, P., Wong, S.-C., 2011. Gene expression profiling reveals the defining features of the classical, intermediate, and nonclassical human monocyte subsets. Blood 118, e16–e31.

Xing, Y., Hogquist, K.A., 2012. T-Cell Tolerance: Central and Peripheral. Cold Spring Harb. Perspect. Biol. 4, a006957. https://doi.org/10.1101/cshperspect.a006957

Ye, C.J., Feng, T., Kwon, H.-K., Raj, T., Wilson, M.T., Asinovski, N., McCabe, C., Lee, M.H., Frohlich, I., Paik, H., Zaitlen, N., Hacohen, N., Stranger, B., De Jager, P., Mathis, D., Regev, A., Benoist, C., 2014. Intersection of population variation and autoimmunity genetics in human T cell activation. Science 345, 1254665. https://doi.org/10.1126/science.1254665

Yoshida, T., Ng, S.Y.-M., Georgopoulos, K., 2010. Awakening lineage potential by Ikaros-mediated transcriptional priming. Curr. Opin. Immunol. 22. https://doi.org/10.1016/j.coi.2010.02.011

Yu, C.I., Becker, C., Wang, Y., Marches, F., Helft, J., Leboeuf, M., Anguiano, E., Pourpe, S., Goller, K., Pascual, V., Banchereau, J., Merad, M., Palucka, K., 2013. Human CD1c+ dendritic cells drive the differentiation of CD103+ CD8+ mucosal effector T cells via the cytokine TGF-β. Immunity 38, 818–830. https://doi.org/10.1016/j.immuni.2013.03.004

Zhang, D.E., Hetherington, C.J., Chen, H.M., Tenen, D.G., 1994. The macrophage transcription factor PU.1 directs tissue-specific expression of the macrophage colony-stimulating factor receptor. Mol. Cell. Biol. 14, 373–381.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., Enard, W., 2017. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol. Cell 65, 631-643.e4. https://doi.org/10.1016/j.molcel.2017.01.023

Ziegler-Heitbrock, H.W., 2000. Definition of human blood monocytes. J. Leukoc. Biol. 67, 603–606.

Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A.M., Mazutis, L., 2017. Single-cell barcoding and sequencing using droplet microfluidics. Nat. Protoc. 12, 44–73. https://doi.org/10.1038/nprot.2016.154

Zou, G.M., Tam, Y.K., 2002. Cytokines in the generation and maturation of dendritic cells: recent advances. Eur. Cytokine Netw. 13, 186–199.