

https://dspace.org/

Transcriptional profiling of CD4+T cells derived from HIV-1 elite controllers

Stansell, Alexander Colin

2020
https://hdl.handle.net/10539/31056

Downloaded from a DSpace repository. https://dspace.org/

Transcriptional Profiling of $CD4^+$ T Cells Derived From HIV-1 Elite Controllers



Alexander Colin Stansell

937996

Dissertation

Submitted to the Faculty of Science, University of the Witwatersrand

in fulfilment of the requirements for the degree of

Master of Science

Supervisor: Dr Nikki Gentle

April 2020

DECLARATION

I, Alexander Colin Stansell (937996), am a student registered for the degree of Master of Science (MSc) by research in the academic year 2020. I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.
- I have not submitted this work before for any other degree or examination at this or any other University.
- The information used in the Dissertation has not been obtained by me while employed by, or working under the aegis of, any person or organisation other than the University.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature:  **Date:** 15 April 2020

ABSTRACT

HIV-1 is still a substantial cause of morbidity and mortality worldwide. It infects $CD4^+$ T cell subsets, where it establishes a latent reservoir of virus that cannot be cleared by host immune responses. In most individuals, progressive infection leads to immune collapse and the development of opportunistic pathogens and cancers that cause mortality. Within all HIV-1 patients a small subset known as elite controllers establish stringent immune control over HIV-1 without the need for antiretroviral therapy. Transcriptional differences in $CD4^+$ T cell subsets within the context of elite control have yet to be fully described. This study used RNA sequencing data to transcriptionally profile Naive (T_N), Central Memory (T_{CM}) and Effector Memory (T_{EM}) $CD4^+$ T cell subsets derived from HIV-1 elite controllers. It aimed to identify differences in immune control mechanisms between T cell subsets. This project used stringent quality processing, normalisation, alignment and quantification methods to ascertain gene expression level, and identify transcriptional differences in T cell subsets and co-expressed gene networks. Three distinct methodologies for identifying co-expressed genes were compared. Findings showed similarities between T_{CM} and T_{EM} subsets, that likely reflect commonalities in their differentiation states, as well as between T_N and T_{CM} subsets that likely reflect commonalities in function. T_N and T_{EM} subsets were found to have the greatest number of differences with respect to their transcriptional profiles; which upon further investigation, were found to reflect differences in viral transcription, differentiation and cellular activation.

ACKNOWLEDGEMENTS

I would like to convey my sincere appreciation and thanks to my supervisor, Dr Nikki Gentle, whose scientific rigour, encouragement and support have been invaluable during the course of this research.

I would like to thank the University for funding my studies.

I would like to thank Colin Stansell for his help with proofreading the final document.

Finally, I would like to thank my family and friends.

RESEARCH OUTPUTS

Oral Presentations

Stansell, AC., and Gentle, N. (2019), Unsupervised Learning to Uncover Gene Co-expression Networks. Molecular Biosciences Research Thrust Annual Research Day. University of the Witwatersrand Johannesburg

Stansell, AC., and Gentle, N. (2019), Unsupervised Learning Techniques for Identifying Gene Co-expression Networks Within The Context Of HIV-1 Elite Control. Cross Faculty Postgraduate Symposium. University of the Witwatersrand, Johannesburg South Africa

TABLE OF CONTENTS

DECLARATION ii

ABSTRACT iii

ACKNOWLEDGEMENTS iv

RESEARCH OUTPUTS v

TABLE OF CONTENTS vi

LIST OF SYMBOLS x

LIST OF ABBREVIATIONS x

LIST OF FIGURES xiii

LIST OF TABLES xvii

1 Introduction 1

1.1 The HIV-1 Lifecycle 1

1.2 HIV-1 Disease Progression 3

1.3 Viral Latency 5

1.4 Outlook For An HIV-1 Cure 5

1.5 HIV-1 Elite Controllers - A Natural Model For A Functional Cure 7

1.6 Known Mechanisms of HIV-1 Elite Control 8

1.6.1 Innate Immune Response 8

1.6.2 Adaptive Immunity 9

1.7 $CD4^+$ T Cell Subsets 11

1.7.1 Naive $CD4^+$ T Cells 14

1.7.2 Effector $CD4^+$ T Cells 15

1.7.3 Central Memory and Effector Memory $CD4^+$ T cells 15

1.7.4 $CD4^+$ T Cell Subsets in HIV-1 infection 16

1.8 Transcriptional Profiling of Functionally Distinct Cell Subtypes 17

1.9 Transcriptomics and RNA Sequencing 18

1.10 Network Analysis Algorithms 19
1.10.1 Clustering in the Context of Co-expression Analysis 21
1.10.2 Unsupervised Machine Learning For Co-expression Analysis 21
1.11 Aims and Objectives 22
2 Materials and Methods 23
2.1 Description of Dataset 23
2.2 Quality Control of Raw Sequencing Data 25
2.3 Pseudo-alignment and Quantification of Gene Expression 25
2.3.1 Pseudo-alignment to a Reference Transcriptome 25
2.3.2 Quantification of Transcript Level Abundance Estimates 26
2.4 Normalisation to Remove Unwanted Technical and Biological Variation 30
2.5 Detection and Removal of Outliers 32
2.5.1 Low Variance and Low Count Gene Removal 32
2.6 Differential Gene Expression Analysis 33
2.6.1 DESeq2 Algorithm 34
2.6.2 DESeq2 Implementation 37
2.7 Co-expression Network Analysis 38
2.7.1 Self Organising Map 39
2.7.2 K-means Clustering 40
2.7.3 Weighted Gene Co-expression Analysis Algorithm and Implementation 41
2.8 Co-expression Method Evaluation 44
2.8.1 Selection of Genes and Comparison Design 44
2.8.2 GO Enrichment 45
2.8.3 Comparison Workflow 45
2.9 T_{EM} vs T_N $CD4^+$ Cell Subset Comparison 45
2.9.1 Hub Gene Identification 46

2.9.2 Visualisation of WGCNA Networks 46
2.9.3 Identification of Notable GO Terms 46
2.10 Code Availability 47
3 Results 48
3.1 Quality Control of Raw Sequencing Data 48
3.2 Quantification of Transcript and Gene Level Abundance Estimates 48
3.2.1 Summary of Abundance Quantification 48
3.2.2 Sequencing Consistency Across Individuals and Cell Subsets 49
3.3 Batch Effect Identification 50
3.3.1 Principal Components Analysis Showed No Discernable Batch Effects 50
3.4 Outlier Detection and Removal 52
3.4.1 Removal of Outliers and Selection of the Final Data Set 52
3.5 Normalisation 53
3.5.1 Comparison of VST and rlog Transformed Counts 53
3.5.2 Heatmaps Show Sample Clustering 54
3.6 Differential Gene Expression Analysis 56
3.6.1 DESeq2 Quality Analysis 56
3.6.2 Identification of Differentially Expressed Genes 56
3.6.3 Overlap of Differentially Expressed Genes Across Comparisons 56
3.7 Evaluation of Methods For Identifying Co-expressed Genes 59
3.7.1 WGCNA 59
3.7.2 K-means Clustering 59
3.7.3 Self Organising Map 59
3.8 Gene Ontology Enrichment Method Comparison 61
3.9 WGCNA was the Most Effective Co-expression Analysis Method 61
3.10 The $CD4^+$ T_N vs T_{EM} Comparison 65

3.10.1 Differences in Viral Transcription across $CD4^+$ T_{EM} and T_N Cell Subsets 65

3.11 Differences in Immune Cell Differentiation Mechanisms across $CD4^+$ T_{EM} and T_N Cell Subsets 68

3.12 Differences In Immune Activation Levels Across $CD4^+$ T_{EM} and T_N Cell Subsets 71

4 Discussion 74

4.1 Quality Control and Sample Normalisation Was Effective 74

4.2 WGCNA was the Most Effective Method for Identifying Co-expressed Genes 75

4.3 Broad Transcriptional Differences Distinguish $CD4^+$ T Cell Subsets 76

4.4 Differences Between T_{EM} and T_N Subsets May Reflect HIV-1 Elite Control 78

4.5 Greater Levels of Viral Transcription in T_{EM} Compared to T_N $CD4^+$ Subsets 78

4.5.1 RIPK1 and RBPJ Facilitate Cellular Differentiation and Viral Replication 79

4.5.2 Differences in Immune Response Mechanisms Across $CD4^+$ T_{EM} and T_N Subsets 80

4.6 Conclusions 80

4.7 Limitations 81

4.8 Future Work 81

5 Bibliography 83

6 Appendix A - Supplementary Results 100

7 Appendix B - Pseudocode 106

LIST OF SYMBOLS

α	Alpha
β	Beta
Δ	Delta
γ	Gamma
λ	Lamda

LIST OF ABBREVIATIONS

A	Adenine
ADCC	Antibody Dependent Cell Mediated Cytotoxicity
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ARC	AIDS-related complex
ART	Antiretroviral Therapy
BH	Benjamini-Hochberg
BMU	Best Matching Unit
C	Cytosine
CA	Capsid
$CD4^+$	Cluster of Differentiation Four Positive
CDC	Centers for Disease Control and Prevention
cDNA	Complementary Deoxyribonucleic Acid
CNET	Plot Depicting genes and biological concepts
CSV	Comma Separated Value
DE	Differential Expression
DNA	Deoxyribonucleic Acid
ENA	European Nucleotide Archive
env	Envelope
FACA	Fluorescence-Assisted Clonal Amplification
FACS	Fluorescence-Activated Cell Sorting
FYB1	FYN Binding Protein 1
G	Guanine

GLM	Generalised Linear Model
GO	Gene Ontology
HAC	Hierarchical Agglomerative Clustering
HDF	Hierarchical Data Format
HIV-1	Human Immunodeficiency Virus Type 1
HLA	Human Leukocyte Antigen
IFNAR1	Interferon-alpha/beta receptor alpha chain 1
IN	Integrase
ISG	Interferon Stimulated Gene
LTR	Long Terminal Repeat
MA	Matrix
MA	Log Ratio Mean Average Plot
MHC	Major Histocompatibility Complex
NC	Nucleocapsid
NGS	Next Generation Sequencing
PBMC	Peripheral Blood Mononuclear Cell
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PCR	Polymerase Chain Reaction
pDCs	Plasmacytoid Dendritic Cells
QC	Quality Control
RBPJ	Recombination signal binding protein
RIPK1	Receptor Interacting Protein Kinase 1
rlog	Regularised Logarithm
RNA	Ribonucleic Acid
RPS25	Ribosomal Protein S25
RT	Reverse Transcriptase
SOM	Self Organising Map
ssRNA	Single Stranded Ribonucleic Acid
STARD7	StAR-Related Lipid Transfer Domain Protein 7
STRING	Search Tool for the Retrieval of Interacting Genes
T	Thymine

T_{CM}	Central Memory $CD4^+$ T Cell
TCR	T cell Receptor
T_E	Effector $CD4^+$ T Cell
T_{EM}	Effector Memory $CD4^+$ T Cell
Th_1	T Effector Helper Cell 1
Th_{17}	T Effector Helper Cell 17
Th_2	T Effector Helper Cell
T_M	T Memory Cell
T_N	Naive $CD4^+$ T Cell
TOM	Topological Overlap Measure
T_{TM}	Transitional Memory $CD4^+$ T Cell
VST	Variance Stabilising Transformation
WGCNA	Weighted Gene Co-expression Network Analysis
WHO	World Health Organisation

LIST OF FIGURES

Figure 1.1:	Schematic diagram depicting HIV-1 viral entry and integration.	2
Figure 1.2:	Changes in CD4+ T cell counts and HIV-1 viral load during the course of HIV-1 infection.	3
Figure 1.3:	CD4+ T cell differentiation.	14
Figure 1.4:	Co-expression network analysis.	20
Figure 2.1:	Sequence de Bruijn graph construction for generating a kallisto transcript index for further sequence alignment.	27
Figure 2.2:	Mapping of reads to a transcriptome de Bruijn graph, as implemented in kallisto.	29
Figure 2.3	Dispersion estimates shrinkage in DESeq2.	35
Figure 2.4:	MA plots demonstrating log fold change shrinkage undergone by DESeq2 for low count genes.	36
Figure 3.1:	Proportion of kallisto 0.46.0 (Bray et al., 2016) aligned reads per individual.	49
Figure 3.2:	Proportion of kallisto 0.46.0 (Bray et al., 2016) pseudo-aligned reads per CD4+ T cell subset.	49
Figure 3.3:	Principal components analysis with sleuth show no discernible batch effects.	51

Figure 3.4:	Euclidean distance hierarchical clustering dendrogram showing clustering of samples.	52
Figure 3.5:	Mean - standard deviation plots showing differences in gene variation for raw, variance stabilised and rlog normalised counts.	53
Figure 3.6:	Hierarchical clustering heatmaps for raw and VST normalised counts.	55
Figure 3.7:	PCA from DESeq2 differentially expressed genes showing first (x-axis) and second (y-axis) principal components.	57
Figure 3.8:	DESeq2 differentially expressed genes for three CD4+ T cell subset comparisons.	58
Figure 3.9:	Number of differentially upregulated and downregulated genes identified by DESeq2 for CD4+ T cell subset comparisons (TEM vs TCM, TN vs TCM and TN vs TEM).	58
Figure 3.10:	Comparison of three co-expression approaches.	60
Figure 3.11:	Box and whisker plot comparing co-expression methodology using gene ratio scores.	62
Figure 3.12:	Figure 3.12: Self Organising Map heat maps for CD4+ TEM (top), TN (middle) and TCM (bottom) subsets for all eight individuals analysed in this study.	63
Figure 3.13:	Weighted Gene Network Analysis showing (from top) heatmaps of similarity matrices, scale-free topology, mean connectivity, heatmaps and adjacency matrices, and cluster dendrograms for CD4+ T cell subset comparisons.	64
Figure 3.14:	Weighted gene co-expression network for a group of DESeq2 differentially expressed genes identified in the WGCNA RoyalBlue module for the TN vs TEM CD4+ T cell subset comparison.	65

Figure 3.15:	ClusterProfiler Gene Ontology Enrichment Dot Plot for the RoyalBlue module for the CD4+ TN vs TEM comparison showing counts of genes found in significantly BH padj < 0.05 enriched sets.	66
Figure 3.16:	RoyalBlue module CNET plot showing the genes involved in GO enrichment pathways differentially expressed between CD4+ TN and TEM subsets.	67
Figure 3.17:	ClusterProfiler gene ontology enrichment dot plot for the SkyBlue3 module showing counts of genes found in significantly Benjamini-Hochberg (BH) padj < 0.05 enriched sets.	68
Figure 3.18:	Skyblue3 module CNET plot showing the genes involved in a given enrichment pathway.	69
Figure 3.19:	Weighted gene co-expression network from genes within the SkyBlue3 and light green modules from an TN vs TEM CD4+ T cell comparison.	70
Figure 3.20:	ClusterProfiler Gene Ontology Enrichment Bar Plot for the Salmon module showing counts of genes (x-axis) found in significantly BH padj < 0.05 enriched sets.	71
Figure 3.21:	Salmon module showing a network of genes from TN vs TEM CD4+ T Cell subset comparison.	72
Figure 3.22:	Salmon module CNET plot showing the genes involved in a given enrichment pathway.	73
Figure S1:	Sample distribution plots.	100
Figure S2:	Sample heatmap with hierarchical clustering of rlog Normalised Counts.	101
Figure S3:	M (log ratio plotted on y-axis) A (mean average plotted on x-axis) plots for DESeq2 (Love et al. 2014).	102

Figure S4:	DESeq2 dispersion estimate plots for different CD4+ T cell subsets.	102
Figure S5:	Mean distance between neurons and genes with higher-dimensional datasets.	103
Figure S6:	Self Organising Map - heat maps for $CD4^+$ T_{EM} (top), T_N (middle) and T_{CM} (bottom) subsets for all eight individuals analysed in the final cohort.	104
Figure S7:	Visualisation of hub genes identified with WGCNA and their respective known interactions from STRING.	105
Figure S8:	Ribosomal proteins found to be differentially co-expressed across the CD4+ TN vs TEM comparison and visualised in STRING.	105

LIST OF TABLES

Table 1.1:	Receptor and cytokine profiles of $CD4^+$ T Subsets.	13
Table 2.1:	Summary of the phenotypic characteristics of the 14 HIV-1 ECs included in this study.	23
Table 2.2:	Table 2.2: Summary of the cell surface markers used to define the $CD4^+$ T Cell subsets included in this study.	24

1 Introduction

Despite much effort over the past 30 years, Human Immunodeficiency Virus Type 1 (HIV-1) is still a substantial cause of morbidity and mortality worldwide. Recent estimates indicate that 32 million people have died from HIV-1 related causes and a further 37.9 million are currently infected (W H O, 2018). HIV-1 predominantly replicates in and destroys $CD4^+$ T cells. These cells are critical to immune function and their depletion leads to the development of opportunistic infections and cancers. Acquired Immunodeficiency Syndrome (AIDS) is clinically defined when $CD4^+$ T cell counts decrease to less than 200 cells/mm^3 of blood plasma. Without treatment, most HIV-1 patients progress to AIDS and will die as a result of immune collapse. Despite advancements in antiretroviral therapy (ART), there is still currently no widely available cure for HIV-1 infection. It is likely that only a vaccine or curative therapy will bring about the end of global HIV-1 related mortality.

1.1 The HIV-1 Lifecycle

HIV-1 is a lentivirus, and thus integrates its single-stranded RNA (ssRNA) genome into the genome of its host. During infection, HIV-1 binds to the CD4 receptor through its envelope glycoprotein gp120 (Figure 1.1; (Blumenthal et al., 2012) and one of its co-receptors, either chemokine receptor five (CCR5) or C-X-C chemokine receptor type 4 (CXCR4) (Nozza et al., 2012). Upon binding to co-receptors, the virus is inserted into the cellular plasma membrane (Blumenthal et al., 2012) and enters the cytoplasm of the cell, where the contents of the virion are released into the cytoplasm (Chauhan and Khandkar, 2015).

For integration into the host genome to proceed, HIV-1 reverse transcriptase (RT) must transcribe the ssRNA genome into double-stranded complementary deoxyribonucleic acid (cDNA). This cDNA is then transported into the nucleus through nucleopores and complexes with HIV-1 integrase to form a pre-integration complex (PIC); (Seitz, 2016). When infected cells are activated the Long Term Repeat (LTR) from HIV-1 acts as an attachment site of cellular DNA dependent RNA polymerase and transcription factors which cause the synthesis of viral messenger RNA (mRNA); (Pan et al., 2013). mRNA is then translated into the structural and regulatory proteins needed to build new virions. HIV-1 virions are then assembled and bud out of cells where they may repeat the process in additional cells. (Rosenblum et al., 2016).

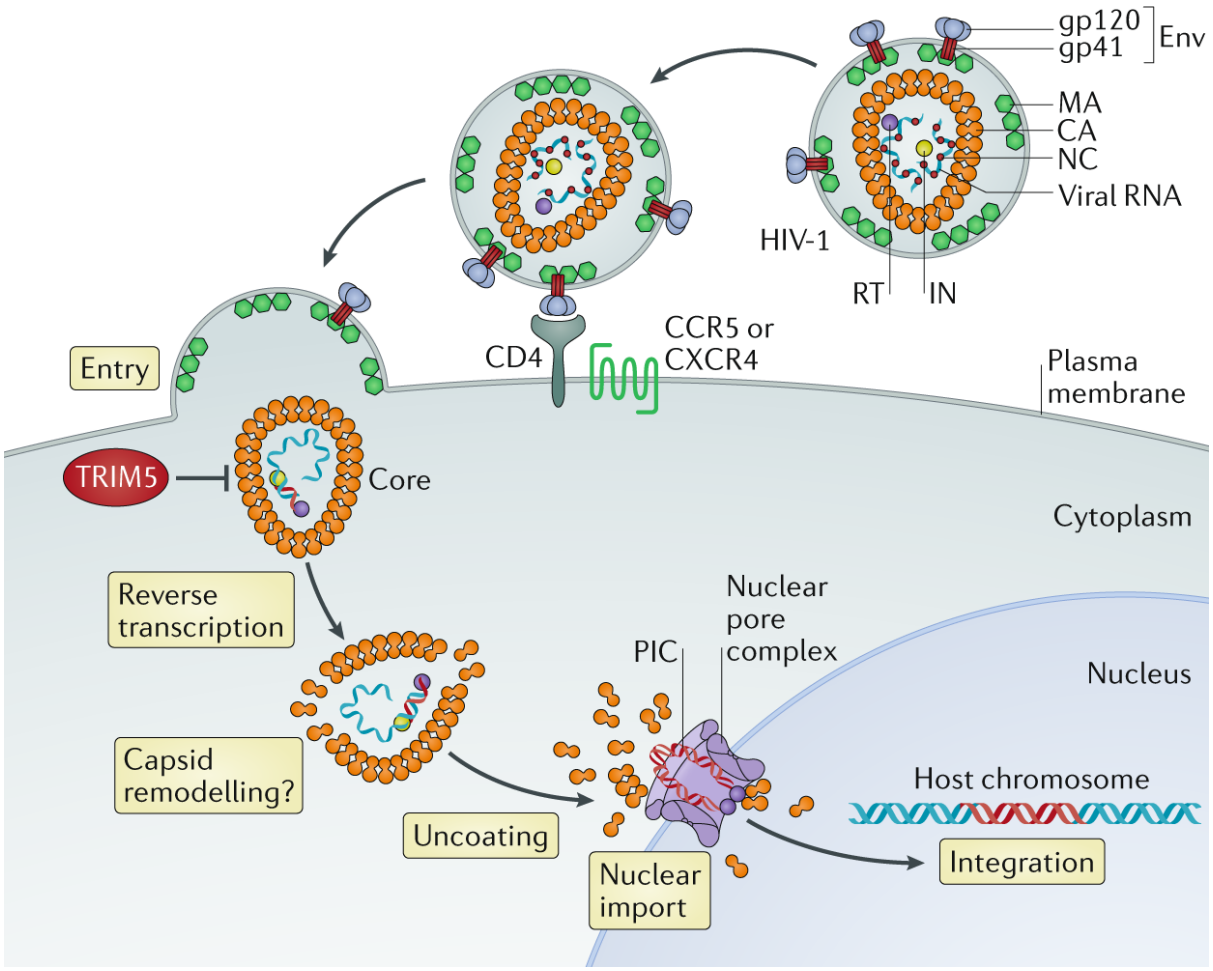


Figure 1.1 Schematic diagram depicting HIV-1 viral entry and integration. HIV-1 viral envelope (env) proteins gp120 and gp41 bind to host cell CD4 receptors and co-receptors (CCR5 or CXCR4). HIV-1 virions contain Matrix (MA), Capsid (CA), Nucleocapsid (NC), Reverse Transcriptase (RT) and Integrase (NC) viral proteins. HIV-1 enters the cell where host transcription factors such as TRIM5 may intervene and inactivate the viral capsid preventing further integration into the host cell. In the absence of this, reverse transcription causes viral ssRNA to be transcribed to double-stranded cDNA. This forms the Pre-Integration Complex (PIC), which may enter the nucleus through the nuclear pore complex. The viral genome may then be integrated into the host genome. Figure from (Ganser-Pornillos and Pornillos, 2019).

1.2 HIV-1 Disease Progression

Upon initial acute infection, HIV-1 replication is exponential, resulting in a peak in HIV-1 viral load and a dramatic decline in $CD4^+$ T cell counts (Figure 1.2) ; (Noel et al., 2016). Once $CD4^+$ T cell counts decrease beyond <300/µl more severe symptoms may be observed. Opportunistic infections are then likely to develop, leading to fatality. The World Health Organisation (WHO) and the Centres for Disease Control and Prevention (CDC) divide HIV-1 disease progression into different stages depending on $CD4^+$ count and symptoms (Brettle et al., 1993). Stage A is asymptomatic, with CD4 cell counts greater than 500/mm^3. Stage B leads to mild symptoms, with CD4 counts between 200 and 499/mm^3. Stage C marks AIDS onset, with severe symptoms and CD4 counts less than 200/mm^3 (Seitz, 2016).

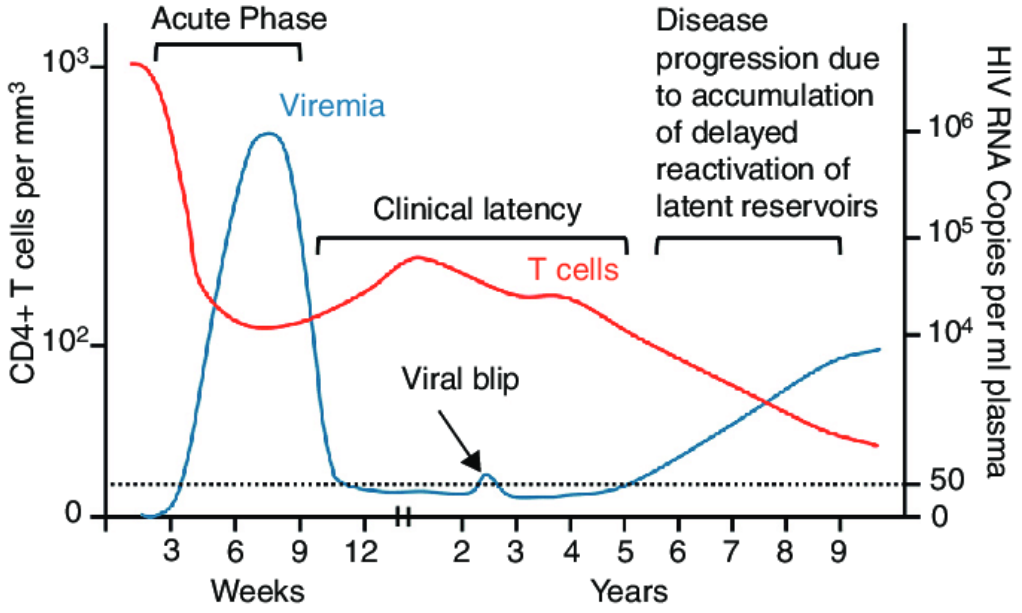


Figure 1.2: Changes in $CD4^+$ T cell counts and HIV-1 viral load during the course of HIV-1 infection. Initially, acute infection causes an exponential increase in viraemia, where T cell activation promotes viral replication. Subsequent immune control of the virus leads to the establishment of a viral set point and a period of asymptomatic clinical latency. During this chronic phase, $CD4^+$ T cell counts continue to decline. This may last for years until the reactivation of latent viral reservoirs causes progression to AIDS. Figure taken from (Selinger and Katze, 2013).

During clinical latency (Figure 1.2), initially depleted T cells are replaced with proliferating naive $CD4^+$ T cells (T_N) that originate in the thymus and then differentiate into effector T cells (Vidya et al., 2017). Effector cells increase the immune response against the virus, which helps reduce viral load to a quasi-homeostatic set point. This is then marked by a period of asymptomatic chronic infection. During this phase effector cells differentiate into memory subsets and harbour latent virus without actively replicating it (Cooper et al., 2013). Should patients remain untreated, latent infection continues for a median of 10 years (Morgan et al., 2002; The UK register of HIV seroconverters, 1996). This is until a resurgence of virions in latent infected cells cause $CD4^+$ T cell depletion and progression to AIDS.

During chronic infection, $CD4^+$ T cells continue to be infected and destroyed (Figure 1.2). Direct infection and subsequent death of $CD4^+$ T cells is one of the main explanations for their depletion, however $CD4^+$ T cell counts also decline as a result of chronic activation (Yates et al., 2007) and inflammation (Liu et al., 1998). Depletion by activation occurs when immune response against HIV-1 causes up-regulation of $CD4^+$ activation markers such as CD25 or HLA-DR, which promote viral replication (Biancotto et al., 2008; Février et al., 2011). Levels of immune activation have been shown to be a better indicator of disease progression than levels of HIV-1 viral RNA (Deeks et al., 2004; Giorgi et al., 1999). Continuous activation of $CD4^+$ T cells leads to induced cell death, as a result of replicating virus or via apoptosis and pyroptosis (Yates et al., 2007).

Further, constant activation may cause $CD4^+$ T cell exhaustion causing diminishing effectiveness of $CD4^+$ T cells against the virus. HIV-1 induced chronic immune activation is marked by high levels of proinflammatory cytokines and chemokines. These include type I Interferon (IFN), Interleukin 6 (IL-6), Transforming growth factor beta (TGFβ), IL-8, IL-1α, IL-1β and Regulated on Activation Normal T cell Expressed and Secreted (RANTES); (Valdez and Lederman, 1997). This pro-inflammatory environment and mass release of cytokines characterise acute and chronic HIV-1 infection (Vijayan et al., 2017). Persistent immune activation and high turnover of $CD4^+$ T effector cells prevent IL-2 based differentiation to memory cell subsets (Younes et al., 2003). This causes further proliferation of effector cells leading to their destruction.

$CD4^+$ T cell depletion may also be caused by the cell death processes pyroptosis and apoptosis (Doitsh et al., 2014). Apoptosis is a form of controlled cell death mediated by caspase-3 as a result of cellular identification of HIV-1. Apoptosis, in contrast to pyroptosis, normally leads

to the death of only the cell undergoing the process. Pyroptosis, driven by caspase-1 (Doitsh et al., 2010), is an inflammatory form of cell death and cells undergo release of all cellular contents including inflammatory cytokines. Such cytokines initiate a cascade of cells that undergo further pyroptosis (Doitsh et al., 2014), leading to $CD4^+$ T cell depletion and progression to AIDS.

1.3 Viral Latency

One of the main barriers that prevents host immune responses from effectively clearing HIV-1 infection is a group of cells, including $CD4^+$ T cells, that harbour the virus without actively replicating it. Host immune responses and ART cannot target these dormant viral reservoirs, which persist despite immune activation and treatment. Cells harbouring viruses are known as latently infected and it is the resurgence of virions in these cells that is known to cause HIV-1 progression to AIDS (Figure 1.2). Elimination of viral reservoirs, therefore, poses a barrier to treating and potentially curing HIV-1 infection. Latency may develop either prior to integration into the host genome or post-integration.

Pre-integration HIV-1 latency occurs when the virus cannot enter the genome as a result of incomplete or impaired RT or Integrase activity (Pierson et al., 2002). Post-integration latency, where HIV-1 is blocked at the translational level is a more common occurrence (Van Lint et al., 2013). It poses a greater threat to disease progression, as proviral DNA can be replicated when the cell undergoes division resulting in the expansion of a latently infected cell (Pan et al., 2013). Post-integration latency may be affected by several factors including the site of integration into the host genome and transcriptional silencing of integrated proviruses (Colin and Van Lint, 2009). Post-integration latency can be propagated through cellular replication in which reservoirs of HIV-1 may be established facilitating HIV-1 persistence.

1.4 Outlook For An HIV-1 Cure

There has been extensive effort to identify methods for treating and curing HIV-1. Much work has gone towards identifying a sterilising cure. Such a cure would rid a patient of all HIV-1 virions and prevent further infection from the virus. Unfortunately, the challenges of both highly variable HIV-1 viral proteins and latently infected HIV-1 cells have slowed progress. The most widely investigated approach has been that of the “shock and kill” strategy (Deeks, 2012). This aims to reactivate HIV-1 in latent cells and destroy it with ART. There has been success in the use of Histone deacetylase inhibitors (HDAC) to reactivate latently infected virus

(Margolis, 2011). HDAC combined with ART has however been ineffective in reducing the number of latently infected cells. A study in which Panobinostat, an HDAC, was used, no substantial decrease in latently infected cells following ART was observed (Rasmussen et al., 2014). This has highlighted the challenge of viral latency in identifying a sterilising cure.

An alternative approach would be the development of a preventative vaccine. Work towards a preventative vaccine has also been slow. The recent landmark HVTN 702 trial (Laher et al., 2020) conducted on 5407 HIV-negative volunteers across South Africa was aborted in January 2020 after finding the vaccine did not prevent HIV infection. The vaccine had consisted of two experimental approaches that had shown promise in a previous, RV144 trial in Thailand (Karasavvas et al., 2012). The HVTN 702 trial consisted of a canarypox vector-based vaccine called ALVAC-HIV and a two-component gp120 protein subunit with an adjuvant to boost host immune response. Findings showed that 129 HIV infections had occurred among vaccine recipients and 123 within the placebo control group (Laher et al., 2020). This led to the conclusion that the vaccine had no protective benefits against HIV-1.

There have also been failures in adenovirus vector-based approaches such as the HVTN 502 STEP HIV vaccine trial on 3000 individuals, in which a greater proportion of HIV cases were found in the vaccine treated group (Iaccino et al., 2008). Again, this has highlighted that an approach that aims to remove all viruses or prevent infection entirely is extremely challenging. Subsequently there has been a shift towards the goal of a functional cure in which prolonged asymptomatic latency is achieved without the need for ART. This approach would not remove all viruses from an infected individual but would confer control over infection to prevent AIDS and HIV-1 related symptoms.

To date, two patients have been cured of established HIV-1 infection. The ‘Berlin patient’ (Hütter et al., 2009) and more recently the ‘London patient’ (Gupta et al., 2020). Interestingly, in both cases it seems that a functional cure has resulted. Patients had received allogeneic stem-cell transplantation of progenitor cells that contained a 32 base pair deletion in the CCR5 allele (CCR5 Δ32) (Gupta et al., 2020; Hütter et al., 2009). This allele provides resistance against HIV co-receptor-mediated attachment to $CD4^+$ T cells (Gupta et al., 2019; Stephens et al., 1998). In the case of both patients, stem-cell transplantation was used to treat underlying cancer. The Berlin patient was treated for myeloid leukemia (Hütter et al., 2009), whereas the London patient received donor cells for treatment of stage IVb refractory Hodgkin Lymphoma (Gupta et al., 2020). The Berlin patient was originally thought to have received a sterilising

cure, however upon antibody analysis, certain HIV-1 proteins in serum were found. This indicated that a functional cure could be in place (Burbelo et al., 2014). Similarly, HIV-1 viral proteins were detected within the London patient, indicating trace levels of virions. In both cases, it is likely that even with very low numbers of latently infected cells or trace levels of HIV-1, host immunity prevents its resurgence. Therefore it is likely that both patients are an extreme functional cure example. Transplantations due to their risk and expense are unfortunately infeasible for the millions of HIV-1 infected individuals (Gupta et al., 2020; Hütter et al., 2009)). The establishment of a functional curative model has however given a goal for future therapies.

1.5 HIV-1 Elite Controllers - A Natural Model For A Functional Cure

Within the pool of HIV-1 infected patients, a small proportion of individuals spontaneously control HIV-1 replication (Sáez-Cirión et al., 2007). A further subset of this group of HIV-1 controllers maintain especially low viral copies of HIV-1 RNA over a period of more than 10 years without ART (Chakrabarti and Simon, 2010). These individuals are known as HIV-1 elite controllers (ECs), and have extremely low rates of disease progression (Blankson, 2010; Stafford et al., 2017). ECs make up less than 1 % of all HIV-1 infected patients (Autran et al., 2011) and are part of a much larger group of individuals known as Long Term Non-Progressors (LTNPs). LTNPS are individuals who progress slowly to AIDS for a number of reasons, including defective viruses, beneficial host genetics and immune control. ECs differ from other LTNPs in that they have been shown to maintain a level of immune control over the virus throughout infection and their mechanisms are not a result of defective virus (Bailey et al., 2008; Blankson et al., 2007; Miura et al., 2008). HIV-1 ECs potentially offer a natural model for a functional cure and their control mechanisms may be useful in the development of therapeutics and vaccines.

For control of HIV-1 to exist there are three main options: (1) control can be a result of defective viruses that cannot replicate, (2) it could be a result of beneficial host mutations such as CCR5Δ32, or (3) it could ensue as a result of stringent immune mechanisms that suppress the virus. The latter seems to be the case for ECs. When HIV-1 sequences from EC plasma and peripheral blood monocytes (PBMCs) were examined, no consistent gene deletions or signatures that would indicate lesser viral replication were found (Mens et al., 2010; Miura et al., 2008; O'Connell et al., 2010). This indicated that defective replication is not the only reason for the lower viral loads observed in ECs.

Furthermore, host genetics with favourable mutations in CCR5Δ32 and beneficial HLA class I alleles also do not fully explain elite control. Heterozygous CCR5Δ32 exists in approximately 30 % of all Caucasian EC individuals (Blankson, 2010), however it does not prevent all HIV-1 replication. The heterozygous CCR5 allele reduces entry to $CD4^+$ T cells, but it has not explained how cells maintain such low levels of virus. Furthermore, beneficial protective HLA Class I alleles also do not seem to fully explain the EC phenotype. HLA-B57, HLA-B27, HLA-B14 and HLA-B51 have been shown to offer effective protection from HIV-1 related disease and control over the virus (Dalmasso et al., 2008; International HIV Controllers Study et al., 2010). However, when one of the most beneficial alleles (HLA-B57) was examined, it did fully protect against disease evolution (Dalmasso et al., 2008). Further there is evidence that not all ECs have beneficial alleles. This indicates that there are likely other factors that allow for viral suppression.

1.6 Known Mechanisms of HIV-1 Elite Control

There is much evidence to suggest that HIV-1 elite control exists as a result of host immune responses to the virus. ECs are a heterogeneous group, and it is likely that there are multiple different means by which viral control is achieved. These mechanisms have been shown to exist as a result of both innate immunity and cell-mediated adaptive immune responses. Mechanisms predominantly exist in $CD4^+$ and $CD8^+$ T cell response to the virus. These mechanisms of immune control likely offer a natural model for a functional cure to HIV-1 and their study is valuable for potential therapeutic or vaccine development.

1.6.1 Innate Immune Response

The primary non-specific response to any pathogen is described as innate immunity. Studying primary response in HIV-1 ECs can give insight as to how they control the virus in the absence of therapy. HIV-1 ECs have a very low viral set point, which is likely caused by a stringent response to the virus during acute infection (Krishnan et al., 2014). Innate responses to the virus include recruitment of immune cells to the site of infection, and activation of cells to support the transition to cell-mediated adaptive immune responses (Sharma et al., 2020). Cells that play an important role in the innate immune response include dendritic and natural killer cells. Unlike patients who progress to AIDS, innate cell populations are preserved in ECs and maintain their functionality.

Natural killer cells play a key role against HIV-1. HIV-1 infection has been shown to cause the down-regulation of HLA-A and HLA-B proteins on the surface of infected cells (Cohen et al., 1999). This has been shown to trigger NK activation against the virus. Killer immunoglobulin-like receptors (KIRs) control the activation of natural killer cells against viruses. ECs may have a greater diversity of KIRs and possess certain KIR alleles that have been associated with slower HIV-1 disease progression. For instance, a study of patients that had both the KIR3DS1 allele and HLA-Bw480I had very slow rates of AIDS progression (Martin et al., 2002). Beneficial HLA alleles by themselves provide slower disease progression, but KIR3DS1 was shown to provide additional stringency in viral set point and slow disease progression (Martin et al., 2007).

A specific subset of dendritic cells called plasmacytoid dendritic cells (pDCs) play an important role against HIV-1. Endocytosis of HIV-1 is the primary activation mechanism of pDCs.(Bochud et al., 2007). Activated pDCs enable swift immune responses to HIV-1 (Blankson, 2010) through cytokine secretion. This includes interferon-alpha (IFNα), interleukin 12 and tumour necrosis factor-alpha (TNFα) which interfere with the replication cycle of HIV-1. High levels of IFNα cause T cell apoptosis and may facilitate viral suppression during acute infection. In progressor patients, pDCs are depleted (Soumelis et al., 2001) and may reach an exhausted state (Hosmalin and Lebon, 2006). EC pDCs show lower levels of depletion and maintain functionality by producing high levels of proinflammatory cytokines throughout infection (Blankson, 2010). Such cytokines cause further cellular recruitment and immune suppression of HIV-1.

1.6.2 Adaptive Immunity

Following a primary innate response, cell-mediated adaptive immunity allows for specialised targeting of pathogens such as HIV-1. It plays an important role during the course of chronic infection. ECs are thought to maintain a stringent adaptive immune response to HIV-1, which allows for continuous viral suppression during clinical latency.

1.6.2.1 Humoral Immune Mechanisms

A key component of the adaptive immune response is the production of antibodies by B plasma cells to counteract pathogens. B cells produce broadly neutralising antibodies (bNAbs), which can bind to a variety of strains of HIV. bNAbs can bind to envelop proteins of HIV-1, preventing entry to $CD4^+$ T cells (Blankson, 2010). This is important during acute infection, as

once binding of HIV-1 to cells occurs, latency makes eradication of the virus by the immune system or ART impossible. ECs have been shown to have lower levels of bNAbs, but with higher avidity allowing for effective neutralisation capacity (Laeyendecker et al., 2008). Despite these findings, only a small fraction of ECs are able to produce bNAbs against HIV-1 (González et al., 2018; Scheid et al., 2009). This suggests that production of bNAbs is not one of the main mechanisms for HIV-1 elite control, even though it may confer viral suppression in some individuals.

1.6.2.2 $CD8^+$ T cell Response

$CD8^+$ cytotoxic T lymphocytes (CTLs) are responsible for the direct killing of HIV-1 infected cells. Cells continually present intracellular peptides on HLA class I molecules, which $CD8^+$ T cells can then bind to. In the event that peptides are recognised as foreign, the infected cells are then killed by CTLs. This process is activated and modulated by $CD4^+$ T cells, dendritic cells, B cells, and through Antibody Dependent Cell Mediated Cytotoxicity (ADCC). $CD8^+$ T cells play an important role in suppressing virus within HIV-1 ECs. Control mechanisms within both $CD8^+$ and $CD4^+$ T cells are marked by a polyfunctional cytokine based response to HIV-1, whereas patients who rapidly progress to AIDs (HIV-1 progressors) exhibit mono or bifunctional responses (Douek et al., 2002; Ferre et al., 2009; Miura et al., 2008; Potter et al., 2007). Furthermore, both $CD8^+$ and $CD4^+$ T cells show strong proliferative responses to stimulus by HIV-1 (Dyer et al., 2008; Migueles et al., 2002), replicating quickly to suppress the virus. $CD8^+$ T cells have a suppressive effect on HIV-1 replication (Sáez-Cirión et al., 2007), and have been shown to have a strong cytolytic response to HIV-1 (Migueles et al., 2008).

A major component of $CD8^+$ T cell control mechanisms are the HLA receptors that present peptides. Overrepresentation of HLA-B57 and/or HLA-B27 have been associated with low viral loads within ECs (Migueles et al., 2002; Miura et al., 2008; Sajadi et al., 2009), although their suppression activity is not substantial enough to fully explain the EC phenotype (Autran et al., 2011). Viral escape of HLA alleles is common and escape mutations have been shown to exist during primary infection with HLA-B57 (Goonetilleke et al., 2009). Furthermore, many ECs have low frequencies of HIV-specific $CD8^+$ T cell responses and have been shown to have poor HIV-1 specific suppressive activity (Sáez-Cirión et al., 2009). It is likely that although $CD8^+$ T cells contribute to the EC phenotype they do not fully explain viral control.

1.6.2.3 $CD4^+$ T cells

$CD4^+$ T cells play an integral role in the adaptive immune response. $CD4^+$ T cells are responsible for regulating the immune response and activating both B cells and CTLs, as well as maintaining immunological memory. HIV-1 ECs maintain a normal or slightly diminished $CD4^+$ T cell population with relatively low levels of circulating viral DNA (Sáez-Cirión, Pancino, et al., 2007). Upon infection by HIV-1, EC T cell populations secrete large amounts of interleukin 2 (IL-2). IL-2 is one of the key cytokines required for the differentiation of $CD4^+$ T cells to exhibit effector function (Chakrabarti and Simon, 2010). IL2 leads to the rapid proliferation of T cell subsets, improving elite control immune response against HIV-1 and conferring control. Regulatory T (T_{reg}) cells, identified by the expression of the transcription factor FOXP3, play a role in modulating immune activation. A related transcription factor FOXO3 has been shown to play a role in the survival of $CD4^+$ T cells in ECs (van Grevenynghe et al., 2008). Mutations in FOXO3a have been shown to allow $CD4^+$ cells to show resistance to apoptosis (van Grevenynghe et al., 2008).

1.7 $CD4^+$ T Cell Subsets

$CD4^+$ T cells have been successfully classified into different subsets depending on immunological function, cytokine profile and cell surface receptors (Table 1.1; Raphael et al., 2015). The main groupings of subsets include Naive $CD4^+$ T cells (T_N), Effector $CD4^+$ T cells (T_E) and Memory $CD4^+$ T cells (T_M) subsets. T_N cells retain stem cell-like function and can undergo differentiation into all other subsets. T_E cell subsets offer direct support to CTL's and B plasma cells, while maintaining a proinflammatory environment. T_M subsets allow for the re-activation of a specific immune response should an individual be exposed to a pathogen again (Rosenblum et al., 2016). Each subset consists of further divisions. T_E cells include subsets that respond to different sets of pathogens including: T helper one (Th_1), two (Th_2) and 17 (Th_{17}) cells. T_M cells include Effector Memory (T_{EM}), Transitional memory (T_{TM}) and Central Memory (T_{CM}) subsets. T_N cells may also differentiate into regulatory $CD4^+$ (T_{reg}) cells, which modulate the immune response to pathogens, and follicular helper $CD4^+$ (T_{fh}) subsets that support humoral immune responses.

During an immune response, a $CD4^+$ T_N cell will differentiate when a dendritic cell presents an antigen to it via HLA Class II (Figure 1.3). T_N subsets then differentiate into T_E cells including, Th1, Th2 and Th17. The type of resulting T_E cell is dependent on the cytokines secreted (Table 1.1), allowing for pathogen-specific differentiation of cell subsets. T_E cells

immediately migrate to the site of infection to facilitate the elimination of the pathogen, through the activation of B cells and CTLs. Other T_N subsets differentiate into T_{reg}, follicular helper (T_{fh}), and central memory (T_{CM}) precursor subsets. T_E cells are short-lived and last the length of the primary immune response (Golubovskaya and Wu, 2016; Rosenblum et al., 2016). A small subset of T_E cells may then differentiate into T_{EM} cells (Figure 1.3). T_{EM} and T_{CM} persist in the absence of pathogenic antigens, and allow for immune reactivation and rapid cellular proliferation should they be re-exposed to an antigen (Golubovskaya and Wu, 2016).

Table 1.1: Receptor and Cytokine Profiles of $CD4^+$ T Subsets

$CD4^+$ T Cell Subset	Receptors	Cytokines Secreted	Cytokines Required for Differentiation from Naive $CD4^+$ T Cells
Naive T cells [λ]	$CD45RA^+$, $CCR7^+$, $CD62L^+$, $CD127^+$, $CD132^+$, $CD25^-$, CD44-, $CD69^-$, $CD45RO^-$, HLA-DR^-	Unable to produce cytokines	-
Effector (T_h1, T_h2, T_h17) [Δ]	$CD25^+$, $CD45RA^{+/-}$, $CD45RO^{+/-}$, CD127	IFN-γ , TNF, IL-4, IL-5, IL-13, IL17, IL 19, IL-21, IL-22, IL-25, IL-26	IL-12, IFN-γ, IL4, IL1, IL-6, IL-23, TGF-β
Central Memory [Δ]	$CD25^+$, $CD45RA^-$, $CD45RO^+$, $CD127^+$, $CCR7^+$, $CD62L^+$	No unique cytokines, dependent on antigen exposure	No unique cytokines, dependent on antigen exposure
Effector Memory [Δ]	$CD25^-$, $CD45RA^-$, $CD45RO^+$, $CD127^+$	No unique cytokines dependent on antigen exposure	No unique cytokines dependent on antigen exposure
Regulatory [Δ]	$CD25^{+/-}$, $CD45RA^-$, $CD45RO^+$, CTLA-4^+, $CD127^-$	IL-10, TGF- β	TGF-β, IL-2
Follicular Helper [Δ]	No unique receptors	IL-21	IL-6, IL-21

Δ - (Golubovskaya and Wu, 2016)

λ - (Md et al., 2018)

1.7.1 Naive $CD4^+$ T Cells

T_N cells differentiate in the thymus. T_N cells circulate frequently through blood and lymph nodes to improve the chances of encountering their appropriate antigen. T_N cells enter lymph nodes through regions called high endothelial venules (HEV). Within lymph nodes, T_N cells can encounter dendritic cell networks. Should they not encounter an antigen, they are drained through the thoracic duct, to re-circulate in blood (Punt, 2013). T_N cells do not secrete proinflammatory cytokines (Table 1.1) and have no effector immune function (Md et al., 2018). Upon activation by a relevant antigen, T_N differentiate into all other $CD4^+$ T cell subsets.

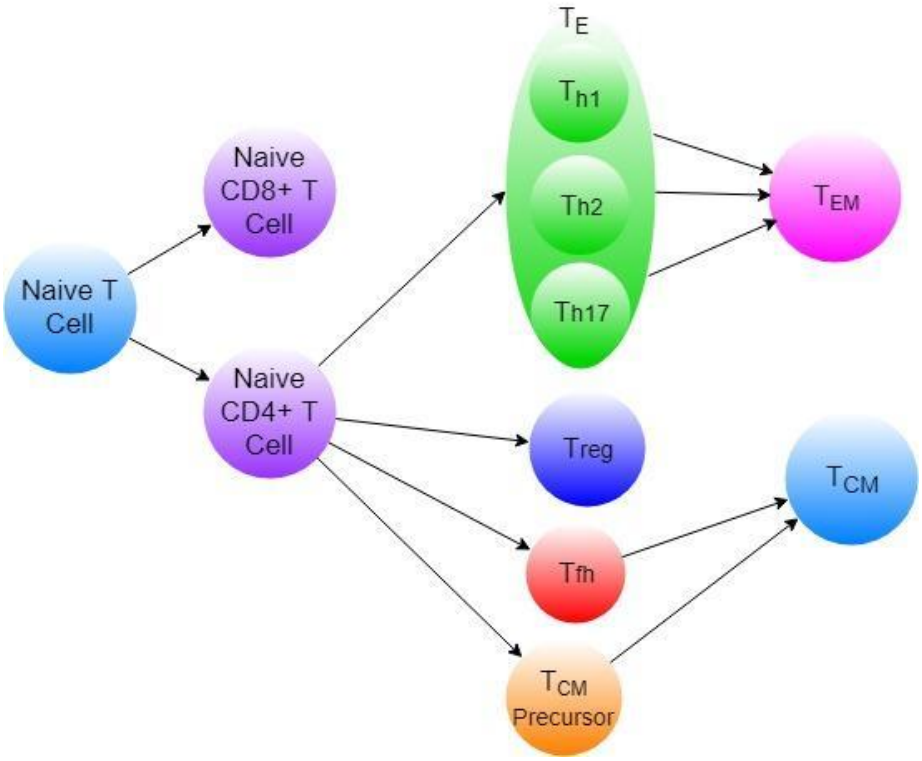


Figure 1.3: $CD4^+$ T cell differentiation. Differentiation from a naive $CD4^+$ T cell (T_N; shown in light blue) to a group of three effector T cells (T_E; shown in green). T_E cells mature into effector memory T cells (T_{EM}; shown in pink). T_N cells also differentiate into T regulatory cells (T_{reg} shown in dark blue), T follicular helper cells (T_{FH} shown in red), and T central memory precursors (shown in orange). T_{CM} and T_{FH} cells differentiate into mature central memory cells (T_{CM}; shown in light blue). T_{reg} cells regulate the expression of cytokines in T_E, T_{FH} cells, T_{EM} cells and T_{CM} cells.

1.7.2 Effector $CD4^+$ T Cells

$CD4^+$ T cell differentiation into T_E cells is complex and includes several distinct T_E subsets. Each subset contains a distinct cytokine profile, which is responsible for specialised immunological responses against different pathogens. All three subsets have a degree of plasticity and there is some crossover in their function and cytokine profile (Rosenblum et al., 2016). T_E cells bind directly to B cells, causing their differentiation into plasma cells, which allow for the generation of antibodies.

T_E cells produce the pro-inflammatory cytokine, interferon-gamma (IFN-γ), causing increased expression of toll-like receptors (TLRs) and antigen presentation. (Raphael et al., 2015). Increased presence of TLRs allows for increased recognition of pathogenic antigens and positive regulation of the immune response. Increased antigen presentation allows for increased efficacy of CTLs. This allows for destruction of infected cells (Raphael et al., 2015). Interleukin-4 (IL-4) and interleukin-5 (IL-5) are also produced by T_E cells (Table 1.1). IL-4 is a survival factor for T lymphocytes. It also promotes B cell and antibody differentiation, to activate and regulate a further immune response (Howard and Paul, 1982). IL-5 stimulates antibody production (Nakajima et al., 1992). These cytokines cause further differentiation of T_N cells into T_E cells, enabling an inflammatory cycle. This cycle enables the elimination of pathogens and infected cells.

1.7.3 Central Memory and Effector Memory $CD4^+$ T cells

Expansion and further differentiation of certain T_E cells allows for the development of immunological memory against pathogens. Memory cell subsets allow for protection in peripheral tissues against reinfection with a specific pathogen and will restart immune responses should they be re-exposed (Sallusto et al., 2004). Protective memory is provided by T_{EM} subsets, which migrate to peripheral tissues and can display immediate effector function. Slower reactive memory is facilitated by T_{CM} cells, which confer no immediate effector function but can differentiate into cells that act against specific pathogens. Central memory cells move to the secondary lymphoid organs and can then differentiate into T_E cells upon antigen stimulation (Sallusto et al., 2004).

Complementary to their functional differences, T_{CM} and T_{EM} cell subsets exhibit distinctions in their cell surface receptors (Table 1.1). T_{EM} subsets display homing receptors, allowing for migration towards infected tissues, whilst T_{CM} have chemokine receptors specific to secondary

lymphoid organs (Brinkman et al., 2013; Sallusto et al., 1999; Willinger et al., 2005). T_{CM} cells express CCR7 and CD62L, which are also expressed by T_N cells, enabling them to manoeuvre through HEV (Brinkman et al., 2013; Campbell et al., 1998; Sallusto et al., 1999; Willinger et al., 2005). In comparison to T_{EM} subsets, T_{CM} subsets have a greater sensitivity to antigen activation. T_{EM} cells can produce effector cytokines such as IFN-γ, IL-4, and IL-5, whilst T_{CM} subsets predominantly produce IL-2, which causes further T cell differentiation. Once differentiated, T_{CM} can produce IFN-γ or IL-4, which act against pathogens (Sallusto et al., 2004). T_{EM} cells can maintain a level of their previous effector function and produce cytokines that retain such functionality (Messi et al., 2003).

1.7.4 $CD4^+$ T Cell Subsets in HIV-1 infection

$CD4^+$ T cells have been shown to differ in their responses to HIV-1 infection and maintain different levels of intracellular virions during latent infection. T_N cells have been shown to consistently have lower levels of HIV-1 DNA than memory cell subsets (Chomont et al., 2009). Furthermore, there is much evidence to suggest that memory T cells harbour the majority of HIV-1 during latent infection (Bosque et al., 2011; Kulpa et al., 2019; Lee and Lichterfeld, 2017). To achieve latent infection, HIV-1 can either infect T cells which transition to a memory phenotype, or in much rarer cases, directly target resting T_M subsets (Agosto and Henderson, 2018).

Active viral replication and turnover of cells seem to occur in T_E and T_N subsets, where cells are rapidly infected and destroyed. This is predominantly the case during acute infection. During chronic infection, T_{CM} cells have been shown to serve as the predominant site for HIV-1 persistence (Chomont et al., 2009). This may be attributed to the greater half-life of T_{CM} cells compared with other T cell subsets (Kovacs et al., 2005). T_{CM} cells also show increased ability to proliferate as a result of IL-7 stimulus, which could cause clonal expansion of latently infected cells and explain HIV-1 persistence (Vandergeeten et al., 2013). In addition, a group of post-treatment controller patients showed low levels of HIV-1 infected T_{CM} cells, which may explain their lower levels of virus (Sáez-Cirión et al., 2013). Furthermore, experiments replicating in vitro T_{CM} cell expansion have shown that HIV-1 infected T_{CM} proliferation does not cause viral reactivation or production of viral antigens. This suggests that antiviral activity cannot target latently infected central memory cells and may explain how they continue to harbour the virus (Bosque et al., 2011).

In contrast, some studies suggest that T_{EM} cells have the largest inducible HIV-1 reservoir (Kulpa et al., 2019). There is also evidence that their activation leads to the highest reversal of latency (Novis et al., 2013). Furthermore, the differentiation of T_{CM} cell subsets to develop T_{EM} like function has been associated with HIV-1 reactivation (Kulpa et al., 2019). This suggests that these cell subsets play an important role in modulating the degree of latency. This serves as evidence that T_{CM} and T_{EM} subsets may play a key role in understanding how HIV-1 latency is reactivated and how disease progression may occur. ECs may display lesser ability for the differentiation of T_{EM} and T_{CM} subsets, preventing HIV-1 progression.

The polarisation of T_N cells into different subsets is facilitated by certain transcription factors that are also able to bind or affect the HIV-1 promoter sequence. This likely affects the formation of latency and may vary the degree to which HIV-1 is transcribed in the different cell subsets. Transcription factors GATA3 Yang and Engel, 1993), BCL6 (Baron et al., 1997) and FoxP3 have all displayed a binding ability to HIV-1 (Holmes et al., 2007). These mechanisms have the potential to reactivate latent infection or cause differences in how the virus is integrated into the cellular genome and transcribed during chronic infection.

The finding that certain cell subsets retain a greater degree of latent HIV-1 is complex and under debate. A recent study found that there was no preferential enrichment or differences in viral gene expression across T_N and different memory subsets (Kwon et al., 2020). This complicates the view that these cell types may be able to be targeted differently or perhaps maintain varying levels of latency. Therefore, there is still a need for further study of these $CD4^+$ cell subsets in the context of HIV-1 elite control.

1.8 Transcriptional Profiling of Functionally Distinct Cell Subtypes

Understanding the differences in cell subsets is valuable as it allows for the distinction of functional characteristics that separate them. Within the context of HIV-1 ECs, establishing differences in $CD4^+$ T cell subsets may discern the function of different subsets in response to the virus. Certain $CD4^+$ T cell subsets are known to harbour varying levels of latent virus, which may be of interest in establishing how the virus is suppressed in different cell subsets and the basis for differences in patients' viral loads. Moreover, differences specific to HIV-1 elite control may exist that differ from baseline differences between cell types. Such differences could improve understanding of how HIV-1 affects host transcription and perhaps elucidate more mechanisms of how ECs maintain immune control.

$CD4^+$ T cell subsets carry varying levels of HIV-1 replication, which offers an opportunity to study the effects of highly infected cells vs relatively non-infected subsets. T_{CM} and T_{EM} cells are known to have high levels of latently infected HIV-1 compared to T_N cells, which are known to have lower levels during chronic infection. Studying differences between these subsets may facilitate an understanding of the development of control over the virus. If there are indeed substantial differences in how cells harbour and replicate viruses this would have therapeutic and vaccine implications. Should the virus persist or be suppressed due to different mechanisms in different cell subtypes, this could be exploited in a treatment context. In addition, a single subset of $CD4^+$ T cells may be responsible for most of the immune suppression needed to maintain an EC phenotype. Perhaps a model for a functional cure can be based on one such subset.

1.9 Transcriptomics and RNA Sequencing

Cell subset-specific differences can be studied through the use of RNA sequencing-based transcriptomics. Next-Generation RNA sequencing (NGS) capabilities have accelerated the study of transcriptomics. Sequencing costs have decreased and the barriers to generating datasets from groups of rare individuals such as ECs have diminished. This has enabled the expressional profiling of many individuals at a genome-wide level. The resulting surplus of datasets has resulted in the development of complex computational tools to successfully process large amounts of data to generate meaningful results. This has led to progress in the development of tools to identify differences in gene expression across conditions.

RNA sequencing (RNA-seq) experiments use the reverse transcription of cellular RNA to cDNA for sequencing with NGS technologies. Current sequencing implementations predominantly produce short reads of 100-150 base pairs in length. These reads are often paired-end - meaning the cDNA fragments generated during sequencing library preparation have been sequenced in both the forward and reverse directions, with a known distance between reads. Longer read technologies are now available, however efficient, low error and cost-effective sequencing still rely heavily on short read fragments. The length of fragments presents a challenge from a computational perspective, as the transcripts to which reads belong have to be identified. For this sequence mapping technologies can efficiently find the position of a read on an existing transcript and quantification algorithms can determine the number of reads mapping to each gene. There are challenges to accurately achieve this, such as the varying ease to which sequencing can occur at different points of the genome, differences in gene length and

differences that occur when comparing samples from different sequencing runs. Modern quality processing, outlier detection and normalisation techniques can effectively combat these challenges leaving reads that are comparable. To compare differences in gene expression, well established advanced differential expression analysis algorithms can be used. Inferring causal relationships between genes or identifying genes that are similar in their expression value is however still a challenge (Serin et al., 2016), and new approaches are needed.

1.10 Network Analysis Algorithms

Network analysis has been used extensively to identify commonalities in gene expression patterns (Li et al., 2018). It predominantly combines clustering and similarity calculations to identify genes that have similar expression values. Co-expression network analysis can be divided into three typical stages: similarity matrix construction, network generation and module detection.

Pairwise relationships between genes are first determined through either correlation coefficients, or through information already known about the genes and their interaction (known as mutual information). There are varying approaches to calculating similarity, however correlation coefficients are perhaps the most commonly used. In the case of correlation values, pairwise coefficients (usually Pearson or Spearman) are calculated for all genes undergoing analysis and used to form a large matrix. (Figure 1.4) ; (van Dam et al., 2018). Correlation values are signed, where values range from -1 (negative correlation) to +1 (positive correlation). However in order for further computation to be performed, these values are scaled between 0 and 1 (Mason et al., 2009).

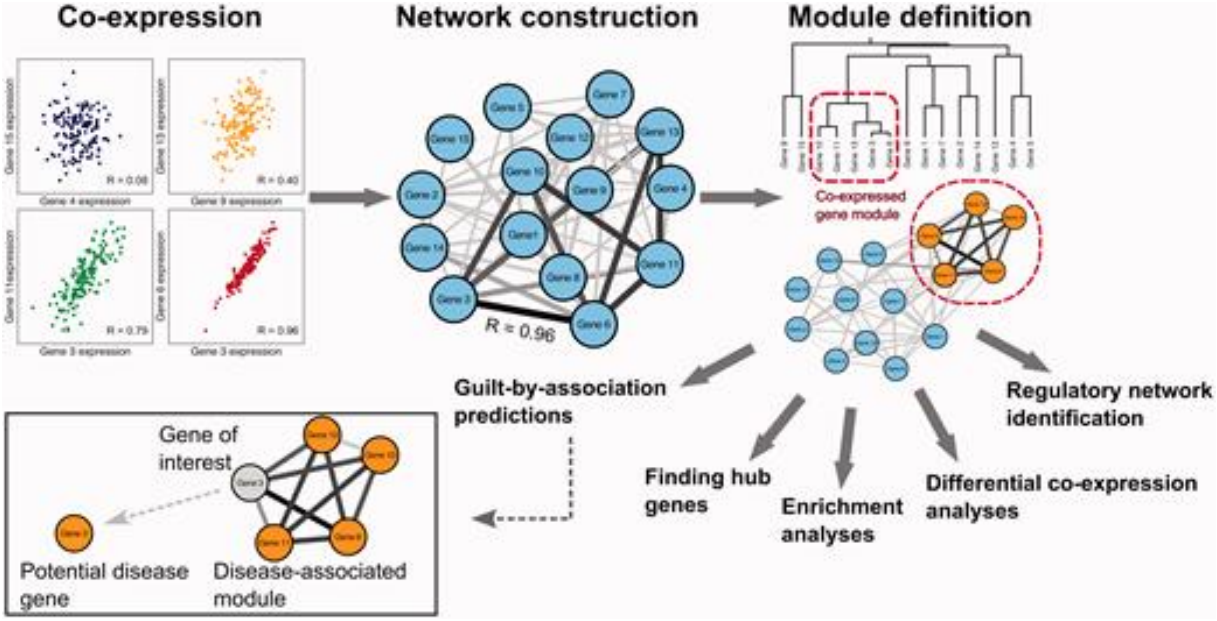


Figure 1.4: Co-expression network analysis. Co-expression: RNA sequence counts are used to compare expression values for thousands of genes. These are then used to calculate correlation coefficients for every gene pair under analysis. Network construction: Processed correlation coefficient values are used to define gene co-expression networks, where the coefficient indicates the strength of co-expression. Module definition: A measure of similarity, combined with hierarchical clustering, separates genes into groupings known as modules. Modules may then be mapped to groups of genes with known function. Genes with unknown function may then be associated with particular biological processes as a result of belonging to a network with a known function. Figure taken from van Dam et al. (2018)

Networks of genes are constructed with nodes representing genes and edges representing the strength of the co-expression calculation (Albert and Barabási, 2002). These networks can either be weighted, in which all genes are connected to each other with co-expression values ranging from 0 - 1, or unweighted, in which there is a binary relationship between each gene (either coexpressed or not). Weighted networks are preferred as they offer more information than un-weighted networks (Zhang and Horvath, 2005). Modules are defined using a distance metric calculated between modules allowing for grouping of different genes as co-expressed. Hierarchical agglomerative clustering (HAC) facilitates this process where similar expression values are grouped together.

1.10.1 Clustering in the Context of Co-expression Analysis

Clustering, an unsupervised learning technique, is the central part of any co-expression analysis workflow. Clustering forms partitions within expression datasets where groupings of genes ideally have similar levels of expression and are dissimilar in expression values to other clusters (Oyelade et al., 2016). There is a consensus that no single clustering algorithm performs best on any given gene expression clustering problem (Pirim et al., 2012). Furthermore, there is no single best criterion for forming an ideal cluster, as clusters can be of arbitrary shapes, sizes and types in multidimensional space (Jain and Dubes, 1988). Therefore methods for validating the use of a particular clustering algorithm when performing co-expression analysis are needed. Clustering can be evaluated based on either internal statistical properties or external information that validates a partition's accuracy (D'haeseleer, 2005). Internal properties are not always reliable in predicting the accuracy of a method for a particular problem. Ultimately the most effective method of comparison is to see how well the clustering algorithm works on a given task (Gibbons and Roth, 2002). Therefore preference is given to external validation that represents what an ideal cluster should conform to within a biological context.

1.10.2 Unsupervised Machine Learning For Co-expression Analysis

Unsupervised machine learning is a type of computational analysis that allows for the identification of commonalities in datasets. At its simplest level it involves clustering of items that have inherent similarities. This capability has been used for co-expression analysis in which certain clustering techniques such as HAC, Principal Components Analysis (PCA) and k-means clustering have been able to effectively group genes with similarities in expression value. More advanced Artificial Neural Network (ANN) approaches have been used to a much lesser extent. ANN approaches offer powerful approaches to cluster groups of objects and can handle much larger datasets than traditional strategies (Chang and Chen, 2015). A component of this study investigated the use of one such ANN approach and compared them with more established co-expression methodologies.

1.11 Aims and Objectives

Identification of mechanisms of HIV-1 elite control is key to developing a model of a functional cure for the virus. Immune mechanisms that control viral replication are likely active at varying levels in $CD4^+$ T cell subsets. Differences in mechanisms likely reflect nuances in the viral latency across subsets and may highlight differences in how subsets harbour the virus. To better understand the nature of these differences, this study aimed to examine the transcriptional profiles of T_N, T_{CM}, and T_{EM} cell subsets within the context of HIV-1 elite control, using computational approaches. To accomplish this dataset previously generated was used (Boritz et al., 2016). This data was from an existing study that had not yet investigated host transcriptomics to an extensive level. This aim was accomplished using four specific objectives:

1. To quantify gene expression at the transcript and gene level within $T_{CM,}$ T_{EM} and T_N cell subsets.

2. To identify genes that are differentially expressed between $T_{CM,}$ T_{EM} and T_N cell subsets.

3. To identify the best methodological approach for identifying co-expressed gene modules within $T_{CM,}$ T_{EM} and T_N cell subsets.

4. To identify co-expressed gene modules within $T_{CM,}$ T_{EM} and T_N cell subsets.

2 Materials and Methods

2.1 Description of Dataset

In order to describe the transcriptional landscape associated with natural control of HIV-1 infection, RNA sequencing (RNA-seq) data from 14 treatment-naive HIV-1 elite controllers (ECs; Table 2.1) were obtained from the European Nucleotide Archive (https://www.ebi.ac.uk/ena). Individuals in this study were defined as ECs based on having maintained plasma HIV-1 RNA levels of <1000 copies/ml during chronic infection, in the absence of antiretroviral therapy (ART), over the course of several years (range: 4 - 30 years); (Boritz et al., 2016).

Table 2.1: Phenotypic characteristics of the 14 HIV-1 ECs included in this study.

Individual	**Sex[1]**	**Date of Diagnosis**	**Sample Collection Date**	**CD4 Count (cells/mm^3)**	**Plasma HIV RNA (copies/ml)**
VRC200_912	M	1989	4/2011	604	442
SCOPE_1475	M	1992	6/2010	1216	66
SCOPE_1492	M	2004	6/2010	746	486
SCOPE_1495	M	1992	4/2014	293	525
SCOPE_1548	F	1991	5/2010	846	118
SCOPE_1541	M	1996	1/2014	933	988
SCOPE_1349	M	2005	2/2014	620	< 40[2]
SCOPE_1788	M	2000	3/2014	503	41
SCOPE_1270	M	1986	4/2014	578	< 40[2]
VRC200_907	M	2007	4/2011	772	854
LIR_01	M	1984	4/2014	934	< 40[2]
LIR_02	F	2005	6/2014	551	267
LIR_03	M	1994	7/2014	494	159
LIR_04	M	2003	11/2014	530	132

[1] Sex is listed as either male (M) or female (F)

[2] Viral load was recorded as below the limit of detection (<40 RNA copies/ml)

These data were available from a previous study (accession number PRJNA326115; Boritz et al., 2016). It included data from each of three $CD4^+$ T cell subsets: $CD4^+$ Naive T cells (T_N), $CD4^+$ Effector Memory T cells (T_{EM}), and $CD4^+$ Central Memory T cells (T_{CM}). These were derived from each of the 14 individuals; obtained by Fluorescence Activated Cell Sorting (FACS) from peripheral blood mononuclear cells (PBMCs), based on the expression of appropriate cell surface markers (Table 2.2). Messenger RNA sequencing libraries were prepared and sequenced using the Illumina HiSeq 2000, to obtain an average of 14869356.77 paired-end 2 x 75 base pair (bp) paired end reads.

Table 2.2: Summary of the cell surface markers used to define the $CD4^+$ T Cell subsets included in this study.

$CD4^+$ T Cell Subset	**Receptor Profile**
Naive (T_N)	CD27+CD45RO- CD8- CD3+
Effector Memory (T_{EM})	CD45RO+/- CD27- CD8- CD3+
Central Memory (T_{CM})	CCR7+ CD27+ CD45RO+ CD8- CD3+

2.2 Quality Control of Raw Sequencing Data

Before any analyses could be performed, it was first necessary to ensure that only high-quality RNA-seq data were included in this study. FastQC (Andrews, 2018) is a widely used tool, designed to screen the quality of sequence read files (in FASTQ format). FastQC generates a summary report for each FASTQ file, detailing the per-base sequence quality, per base sequence content, sequence duplication levels, overrepresented sequence levels and adaptor content for that file. Assessment of these quality metrics allows files containing poor sequence data to be identified and removed from further analyses. FastQC v0.11.8 was used to assess the quality of the sequence data included in this study. The FastQC reports from each sample were then aggregated into a single report, using MultiQC (Ewels et al., 2016), in order to allow for simultaneous visualisation and comparison of all sequencing files.

2.3 Pseudo-alignment and Quantification of Gene Expression

2.3.1 Pseudo-alignment to a Reference Transcriptome

In order to perform transcriptional profiling of $CD4^+$ T cell subsets, alignment of sequencing reads and quantification of gene expression is necessary. Alignment of RNA-seq reads can be performed using either traditional alignment approaches, such as those implemented in STAR (Dobin et al., 2013) and TopHat2 (Kim et al., 2013), or through the use of newer pseudo-alignment approaches, such as those implemented in kallisto (Bray et al., 2016) and salmon (Patro et al., 2017). In contrast with traditional approaches; pseudo alignment approaches allow gene expression to be quantified more quickly, using reduced computational resources, but with similar accuracy to traditional alignments (Bray et al., 2016).

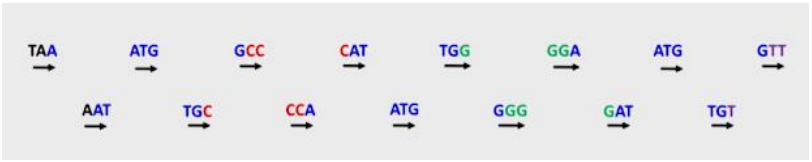
The pseudo-alignment and quantification tool, kallisto v0.46.0 (Pimentel et al., 2016), was chosen to align the available sequencing reads to a reference transcriptome (GRCh38.p12). Kallisto generates an index from a known reference transcriptome, against which it maps and quantifies sequencing reads. This index consists of a de Bruijn graph representation of the transcriptome, against which transcript-derived k-mers are matched (Figure 2.1 A; Bray et al., 2016). To form a transcriptome index graph kallisto breaks up the known transcriptome into k-mers. For example, the case of a three-base pair transcript, ‘ATG’, kallisto breaks this into the 2-mers, ‘AT’ and ‘TG’. These ‘AT’ or ‘TG’ sequences are matched with other ‘AT’ and ‘TG’ edges from different k-mers to form a graph (Figure 2.2B). Transcripts are represented by kallisto as Eulerian paths through de Bruijn’s graph such that a set of all paths which together

represents the transcriptome (Bray et al., 2016; (Figure 2.1D). Eulerian paths are defined to have the same number of inputs and output connections (balanced graph) and by going through each k-mer once. By creating this graph kallisto has an index which describes all the transcripts that occur within the transcriptome. Due to the reduction of the graph, the vertices of the graph may contain multisets (instances where more than one path or transcript travels through vertices of the graph). These multisets are described by the kallisto package as k-compatibility classes. K-compatibility classes are used during the mapping and quantification process in the context of paired-end sequence data the distance between reads is known (Compeau and Pevzner, 2014). This additional information reduces the de Bruijn graph further reducing the number of different options a read might map to (Figure 2.1E).

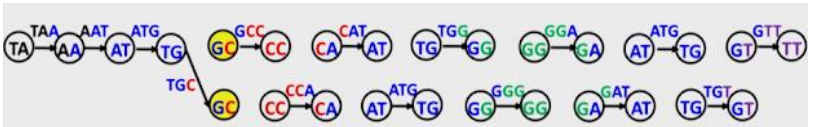
2.3.2 Quantification of Transcript Level Abundance Estimates

Should a mapping read have a high number of intersecting multisets (k-compatibility class) for a particular Eulerian path (transcript) through the graph, it is likely that read maps to that particular transcript (Figure 2.2C). This is determined by hashing the constituent k-compatibility classes of k-mers that together make up the read using an expectation maximisation algorithm (maximum likelihood). Therefore should a read have a high number of intersections with a particular path through the graph it is likely it maps to that read (Figure 2.2C). To reduce computational burden kallisto skips certain k-mers (Bray et al., 2016). This means that only certain constituent k-mers of a read have to be mapped (Figure 2.2D and 2.2E). The resulting intersection of occurrences through a particular path can be quantified and hence the abundance of a particular transcript determined.

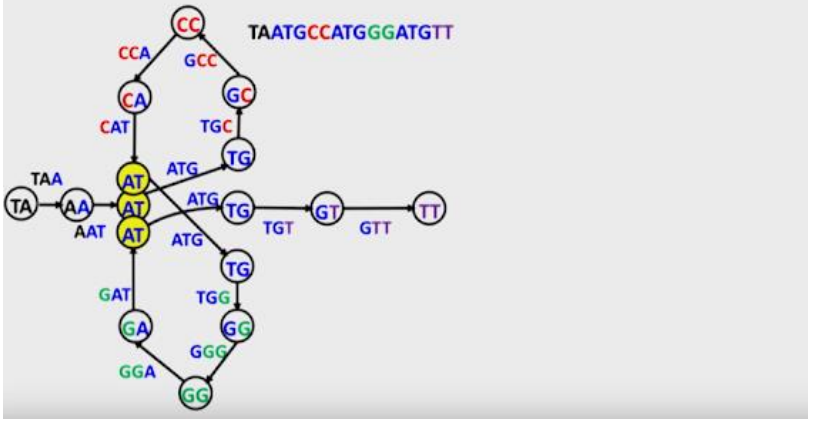
A



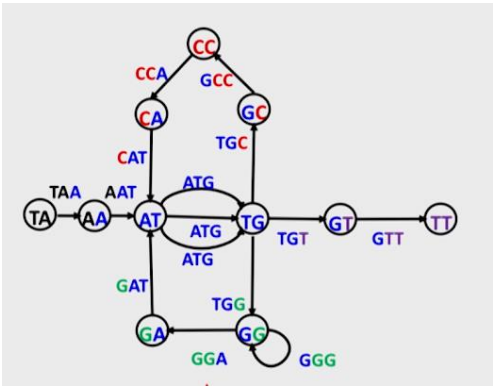
B



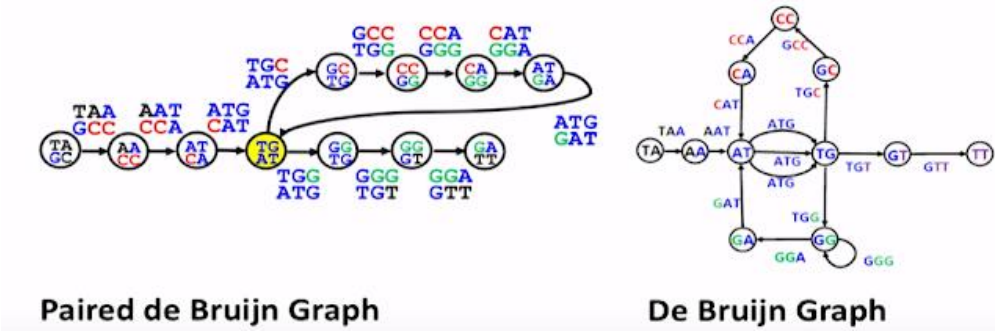
C



D



E



F

(TAATGCCATGGGATGTT)

Figure 2.1: Sequence de Bruijn graph construction for generating a kallisto transcript index for further sequence alignment. (A): To form an index de Bruijn graph, known transcripts from the transcriptome are broken into k-mers (in this case three base pairs in length). (B): Reads are broken into edges such that edges can be matched to generate a graph.

(C): Where multiple instances of the same k-mer occur glueing simplifies the graph reducing redundancy (D): Final De Bruijn graph where a Eulerian path (path that goes through each k-mer once) describes a transcript within the transcriptome. (E): Paired-end sequence reads give information as to the distance between k-mers. This allows for further simplification of the graph resulting in fewer potential Eulerian paths and a more accurate representation of the transcriptome. (F): a likely transcript that exists within the graph. Figure adapted from (Compeau and Pevzner, 2014).

Kallisto performs this process exceptionally efficiently, which allows time for statistical bootstrapping to be undertaken. Bootstrapping is a form of statistical resampling where analyses can be repeated using replacement sampling in order to determine the uncertainty of each abundance quantification (Bray et al., 2016). This gives an indication of the accuracy of each abundance quantification, that can then be used to ensure the best alignment and quantification is selected. This project used 100 bootstrap estimates (the recommended number in kallisto documentation).

As kallisto reports expression estimates at the transcript level, it was necessary to aggregate these transcript abundances to the gene level using the R package, tximport (Soneson et al., 2015). Gene level abundance estimates allowed for a broader profile of the cell subsets to be studied and was necessary for further differential expression analysis with the R package DESEq2 (Love et al., 2014). Tximport uses a transcript to gene mapping file to determine which transcripts form particular genes. It then corrects for differences in gene length (Trapnell et al., 2010) to accurately quantify gene-level abundance estimates.

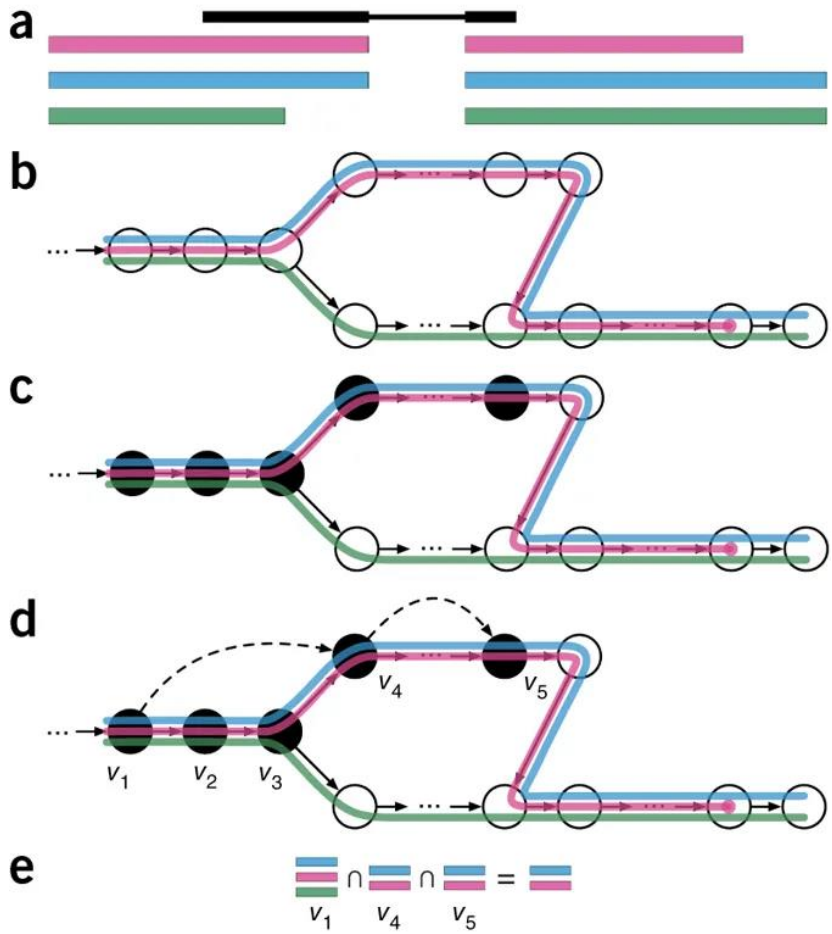


Figure 2.2: Mapping of reads to a transcriptome de Bruijn graph, as implemented in kallisto. (a): A read is shown in black with three potential transcripts which are options for it to map to (transcripts are coloured pink blue and green) that it may belong to. Spaces show regions between k-mers. (b): Nodes shown as circles represent k-mers that form a coloured de Bruijn graph with transcripts (coloured pink blue and green) as Eulerian paths through the graph. (c): k-mers are hashed to determine the k-compatibility class of a read for a particular transcript. Should it have a high number of multisets (k-compatibility classes) for a read it is likely that it belongs to that read. (d): To reduce computational burden, redundant reads can be skipped. Skipping is shown in dashed lines (e): The k-compatibility class of a read for a given transcript is determined by taking the k-compatibility of its constituent k-mers. Figure taken from (Bray et al., 2016).

2.4 Normalisation to Remove Unwanted Technical and Biological Variation

Normalisation is necessary in RNA sequencing experiments to mitigate unwanted variation caused by technical factors, which may result in the over or under-estimation of gene expression estimates. This needed to be accomplished while still maintaining the biological variation inherent in the sample of interest. Certain levels of biological variation caused by unique characteristics of individuals would be adjusted for, at a later step in network and differential expression analysis. Therefore, there was a need for extensive normalisation of the raw gene-level abundance estimates produced from tximport (Soneson et al., 2015). Technical variation can be induced by differences in sequence library size (Mortazavi, et al., 2008); library preparation techniques (Bullard et al., 2010); differences in sequencing depth (Ekblom et al., 2014); gene length (Mortazavi et al., 2008); sequence composition (Jiang et al., 2011), mapping biases (Jiang et al., 2011; Marioni et al., 2008) and the introduction of sequencing batch effects (Nygaard et al., 2016).

Sequence depth refers to the number of times a region of a transcript is sequenced. Differences in the level to which a particular region is sequenced are one of the most common sources of technical error in RNA sequence experiments. To mitigate this problem, this project employed a size factor median ratio normalisation approach. This was implemented for all network and co-expression analysis. Differential expression analysis did not require this normalisation as DESeq2 undertakes this within its own model (see Section 3.7.3.1).

Size factors render samples with different sequence depths comparable (Anders and Huber, 2010). These factors adjust counts relative to the sequencing depth, allowing for a comparable count value. In the case of a matrix of n x m counts, k_{ij}, where i indexes the genes and j indexes the samples, the size factor is calculated as the median of the ratio of k_{ij} to those of a pseudo reference (Equation 1). This pseudo reference is calculated by taking the geometric mean across all samples in a comparison. This size factor estimation is described in Equation 1 (Anders and Huber, 2010). Each sequence read value is then adjusted by a sample size factor resulting in a mechanism to reduce difference in the sequence depths across individuals or cell types.

$$\hat{s}_j = \underset{i}{median} \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv}\right)^{\frac{1}{m}}} \quad (1)$$

Another challenge is that raw count sequence data is heteroscedastic. This refers to the dependence of variance on the mean. Meaning that genes with a high mean level of expression (high count values) across samples, have a greater degree of variation than genes with lower expression. This is a challenge when comparing samples as it is more likely that higher expression genes will be detected as having differences (due to these carrying more variation with greater absolute differences in counts). A log transformation can be implemented to reduce this problem. This can be calculated as log(counts + 1). In this calculation, the pseudo count of 1 avoids undefined log(0) values, replacing them with log(1) = 0. This basic log transformation reduces variation in counts at higher levels, reducing heteroscedasticity. It does however cause another problem where the variance in count values close to zero are amplified. Regularised logarithm transformations (rlog) and Variance Stabilised Transformations (VST) resolve this problem by shrinking low gene counts to the gene average across samples, reducing variance. Therefore both VST and rlog transformations reduce the problem of heteroscedasticity by reducing variation in read counts.

This project implemented size factor, rlog and VST normalisations using the R package DESeq2 (Love et al., 2014). Functions for running rlog and VST first calculate size factors normalising for differences in read-depth across samples. rlog or VST transformations then result in approximately homoscedastic data, which was ideal for network and co-expression analyses. Both methods have the added advantage in transforming discrete count data to continuous values, which are required by co-expression analysis and unsupervised machine learning algorithms (Hughitt, 2016).

VST and rlog normalisation are both known to be effective in reducing variation in gene values, however there are nuances in their efficacy on different datasets. To identify the best approach for the dataset used in this study, a comparison between methods was undertaken. This involved generating several quality plots representing the degree to which each transformation reduced variation. Plots included mean-variance figures, sample distributions and sample heat maps all of which were generated using ggplot2 (Wickham, 2011). On this basis, variance-stabilised counts were used for all co-expression analysis conducted in this project.

Batch effects are a form of technical variation that arises as a result of groups of samples being processed together (Johnson et al., 2007). When samples are prepared and sequenced in the same batch, there is a greater likelihood that there will be technical similarities amongst samples. These batch effects can be induced by different experiment times, reagents, handlers

and sequence runs (Goh et al., 2017). It was therefore important to identify potential sources of batch effects and remove any variance caused by them.

To identify batch effects in quantified read data, principal components analysis (PCA) using the R package, sleuth (Pimentel et al., 2017) was undertaken. PCA uses orthogonal transformations to simplify higher-dimensional datasets for clustering. It does this by breaking variance up into individual principal components. The variance of multiple principal components may then be plotted. PCA analysis allowed for clustering and labelling on the basis of sample date, individual and cell subset. To generate PCA plots, metadata from the individuals that passed quality control was processed and manipulated into a sample-to-covariate file. This file contained the location of kallisto abundance.h5 files for each sample and sample metadata. The columns for different metadata conditions were read in as factors and numbers as numerics for processing with sleuth. Principal components analysis plots were generated and labelled based on all metadata conditions. These were analysed for batch effects.

2.5 Detection and Removal of Outliers

Following normalisation, sequenced data needed to be screened for potential outliers. Outliers can be as a result of biological variation, in which data from an individual's cell subsets greatly vary from the rest of the dataset. Differential and co-expression network analysis algorithms are very sensitive to intersample variation (Langfelder and Horvath, 2008) and require a level of homogeneity to form accurate results. Therefore, removing outliers and reducing variation was necessary to ensure accurate transcriptional profiling of the $CD4^+$ T cell subsets.

To identify outliers, samples were clustered using a hierarchical agglomerative method with the R package, fastcluster (Müllner, 2013). Hierarchical clustering used euclidean dissimilarity measures to determine the distances between the expression profiles of different samples. Clustering was visualised using a dendrogram with the R package, ggplot2 (Wickham, 2011). Outliers were identified and removed using the CutdynamicTree function. Results were filtered so as to maintain three cell types of interest per individual.

2.5.1 Low Variance and Low Count Gene Removal

Low variance and low count genes cause noise in both co-expression and differential expression analysis. DESeq2 accomplishes this in its own model (see Section 2.6.1), but this processing was needed for network and co-expression analysis. Initially, low count (genes with zero counts) and low variance genes (as defined by WGCNA's goodSampleGenes function)

were removed from the dataset using WGCNA (Langfelder and Horvath, 2008). However, in the end, this was not found to be necessary as ultimately only differentially expressed genes were used to form final co-expression results.

2.6 Differential Gene Expression Analysis

HIV-1 control mechanisms are likely active at different levels across $CD4^+$ T cell subsets, as they probably vary depending on the age of the cell, exposure to HIV-1 and immune function. Therefore a methodology to understand differences between cell subsets was central to this study. Hence a robust statistical approach was needed to accurately identify genes that were differentially expressed. Differential expression analysis offered a strategy for identifying these genes, and was undertaken in this project. Its analysis accounts for differences in sequence coverage, sequence depth, and errors that occur when testing thousands of genes simultaneously. This made it an ideal approach for transcriptionally profiling and identifying differences amongst $CD4^+$ T cell subsets.

There are many differential expression analysis algorithms available for RNA sequence data. A method that could identify differentially expressed (DE) genes in opposition to DE transcripts was favoured. Therefore transcript based methods such as sleuth (Pimentel et al., 2017) were excluded from selection. Furthermore, an algorithm designed specifically for discrete RNA sequence read counts was favoured in opposition to methods originally designed to process continuous microarray data. This meant the exclusion of methods such as limma, which was initially designed for microarray experiments (Ritchie et al., 2015).

This left two main options for a DE analysis - DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010). Both of these methods have been shown to be more sensitive and precise than other options (Love et al., 2014) such as voom (Law et al., 2014), SAMseq (Li and Tibshirani, 2013), EBSeq (Leng et al., 2013) and DESeq (Anders and Huber, 2010). DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010) employ methodologies and are based on a negative binomial distribution. However, DESeq2 (Love et al., 2014) has two advantages over edgeR: (1) it uses a log fold change shrinkage step to account for noise in low mean count data, and (2) it has an efficient automated process of independent filtering for tests that are unlikely to result in DE genes. Based on this reasoning, DESeq2 (Love et al., 2014), was chosen as the differential expression analysis method for this project.

2.6.1 DESeq2 Algorithm

DESeq2 (Love et al., 2014) is an R Bioconductor package. Its algorithm consists of six core steps: normalisation, model fit, dispersion estimation, log fold change shrinkage, correction for multiple hypothesis testing and independent filtering.

DESeq2's normalisation procedure aims to mitigate differences in library composition and library size. Library compositional differences occur when outlier genes with very high expression levels bias the relative expression levels of other genes. Differences in sequence depth are as a result of sequence libraries containing varying numbers of reads. To control for these differences, DESeq2 applies a size factor-based normalisation (see Section 2.4). Size factors adjust counts to a relative pseudocount that accounts for the unique sequence depth and composition characteristics of a particular sample. DESeq2's model assumes that a matrix of counts, k_{ij}, follows a negative binomial distribution with mean, μ_{ij}, and dispersion (a parameter related to the variance of count values), α_i. The mean of counts is calculated as $\mu_{ij} = s_{ij} q_{ij}$ where a quantity, q_{ij}, is proportional to the number of sequence reads and s_{ij} defines a scaling size factor (Equation 1; Section 2.4).

Following normalisation, DESeq2 fits a generalised linear model (GLM) with a log link function (Equation 2). Generalised models refer to the use of a non-normal distribution (in this case, the negative binomial). The use of a log link function exponentiates linear predictors to accommodate for such non-normality. The DESeq2 GLM describes an outcome value, q_{ij}, indicative of a gene's relative expression for the comparison of interest (in this study, two different cell subsets). The value, q_{ij} (Equation 2), is calculated as the sum of the product of the design matrix, x_{jr}, and model coefficients, β_{ir}. Coefficients, β_{ir}, indicate the expression strength of a gene as a result of count values across samples. The experiment design formula describes the different cell subsets and individuals to be compared. Through this process, the GLM determines the effect of the condition of interest (cell subset) on the level of expression, while minimising the certain β coefficient expression values attributable to intra-individual differences (described in the design formula). The log link function produces a $\log_2$ fold change value, representing the degree of change in expression as a result of the condition of interest (cell subset).

$$\log_2 q_{ij} = \sum_r x_{jr} \beta_{ir} \qquad (2)$$

DESeq2 uses a dispersion parameter, α_i (Equation 3), to describe variability between different genes, such that the variance of a matrix of counts, K_{ij} , is calculated as:

$$Var\ K_{ij} = \mu_{ij} + \alpha_i \mu_{ij}^2 \qquad (3)$$

Heteroscedastic data results in high counts causing greater variation than low counts. Further in the case of small replicate sets (most RNA sequence experiments with fewer than 50 replicates) the dispersion parameter tends to be high, and can cause a lot of noise in a model. Differences in dispersion across different mean count values therefore need to be adjusted for. DESeq2 accommodates for variation and noise by estimating and shrinking dispersion values. First DESeq2 calculates an initial gene-wise dispersion estimate using a maximum likelihood algorithm (Figure 2.3A). It then fits a mean dispersion smooth curve to the data, modelling the average dispersion values across genes. This process of dispersion estimation shares information from other genes to all other estimates. Through this, DESeq2 assumes genes with similar expression values will have similar dispersion. Estimates are then adjusted using an empirical Bayes approach to shrink values to the smooth line reducing noise in the experiment (Figure 2.3B). The level of shrinkage depends on how close the estimates are to the trend line and the mean count value. Some estimates are calculated as outliers (Figure 2.3B) and they do not undergo shrinkage as they would cause false positives during hypothesis testing. Outliers are removed during the process of independent filtering.

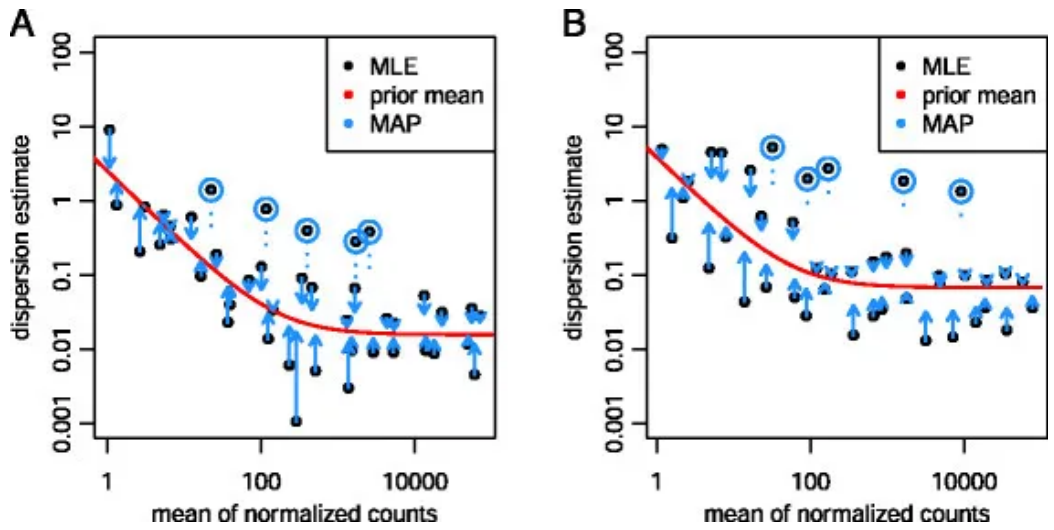


Figure 2.3 Dispersion estimates shrinkage, as performed by DESeq2. (A): Initial gene-wise estimates calculated with a maximum likelihood algorithm are shown in black, a red trend line plots the mean dispersion relationship and then blue lines indicate shrinkage towards the trend line. Outliers with very high dispersion are pictured as circled blue dots that do not undergo shrinkage and are adjusted later in DESeq2's filtering step. Figure adapted from (Anders and Huber, 2010).

DESeq2 then undergoes log fold change shrinkage for low count genes (Figure 2.4A). Errors or variation in low counts can easily cause large log fold changes to be identified. Whereas with higher counts values the same errors have less of an effect. DESeq2 reduces these errors by shrinking fold changes with low mean count values towards 0 (Figure 2.4B). This effectively removes log fold change scores susceptible to error.

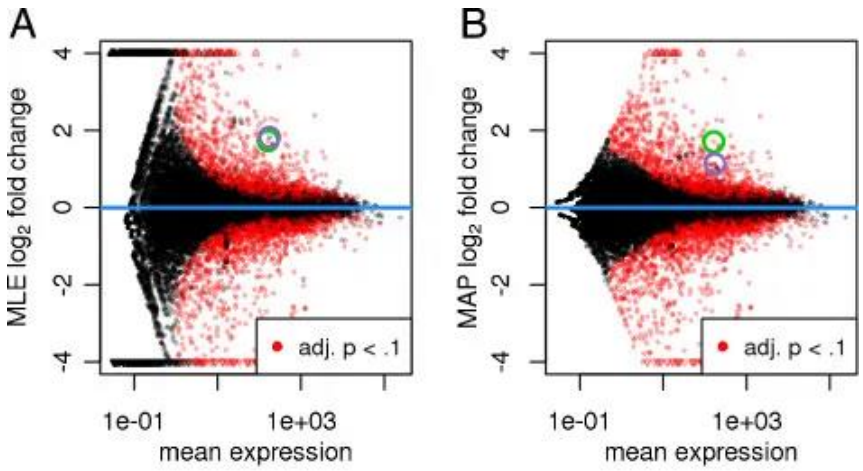


Figure 2.4: MA plots demonstrating log fold change shrinkage performed by DESeq2 for low count genes. A - Log fold changes from a sample experiment without shrinkage showing high log fold changes for low mean expression changes. B - With shrinkage mean expression values near zero are less likely to have high log fold changes. Circled values show the change of a value that has undergone shrinkage (purple) in B compared to a value in green that does not undergo shrinkage. Figure adapted from (Anders and Huber, 2010).

DESeq2 uses a Wald hypothesis test (Wald, 1945) to determine whether model coefficients for a given gene differ significantly from zero (indicating differential expression). To do this, shrunken log fold change estimates are divided by the standard error to calculate a z statistic. This z statistic can then be evaluated as to whether a gene is significantly differentially expressed to a given confidence interval, p. P-values are then adjusted for multiple testing using the Benjamini Hochberg (BH) method (Benjamini et al., 2009).

Multiple testing adjustment accounts for the misidentification of false positives when comparing thousands of genes simultaneously. To reduce this, error values can be adjusted through different techniques. One of the most effective methods is the Benjamini Hochberg (BH) rank-based approach (Benjamini et al., 2009). To calculate BH adjusted p-values, p-values are ordered from smallest to largest, multiplied by the total number of tests, m, and divided by the positional rank, j, of each test. Any p-values exceeding 1 are made to equal 1. This is represented in the following equation:

$$p_{(i)}^{BH} = min\left\{ min_{j>1}\left\{ \frac{mp_{(j)}}{j} \right\}, 1 \right\} \tag{4}$$

The BH method is stringent and deduces the number of false positives identified. Due to BH's calculation being affected by the number of independent tests performed, removing tests that are unlikely to yield DE genes reduces this stringency. DESEq2 does this by removing genes that have 0 expression values and genes with low relative expression value to each other. This process is known as independent filtering. It is independent as the removal of gene tests is not influenced by the sample or group paired genes belong to. Hence, DESeq2's algorithm effectively accommodates for RNA sequence read variance and noise to produce reliable tests that discern differences in gene expression.

2.6.2 DESeq2 Implementation

To run DESeq2, a list of condition pairs were generated to assign each cell subset comparison (T_{EM} vs T_{CM}, T_{EM} vs T_N, and T_N vs T_{CM}). Abundance hdf5 file types relevant to a particular comparison and metadata pertaining to the individual and cell subset were then selected. Hdf5 files were imported using tximport (Soneson et al., 2015) to maintain gene-level abundance estimates. A DESeq2 object (dds) was then created with all information relevant to a particular comparison. DESeq2 was then run, producing a data frame with Wald test results for every gene. This included the gene identifier, base mean, DESeq2 $\log_2$ fold change, z statistic, significance value and Benjamini Hochberg adjusted p-value. Lists of genes produced were filtered for adjusted p values < 0.05 and were saved to CSV format for further analysis. Quality control plots were generated for every DESeq2 comparison using R packages, ggplot2 (Wickham, 2011) and DESeq2 (Love et al., 2014). This included dispersion plots, histograms of p-values, distribution plots of rlog transformations, sample distributions, heatmaps, MA plots, and volcano and PCA plots.

2.7 Co-expression Network Analysis

Once DE genes were identified, there was a need to understand how genes interacted with one another in a given cell subset. This would allow the comparison of different functions of groups of genes and how they interact with one another. Co-expression network analysis is a method that allows for the study of these interactions (Amar et al., 2013). This approach is based on the likelihood that genes with correlations in their expression patterns represent likely co-activated or co-regulated gene sets (Langfelder and Horvath, 2008). This may not necessarily be a causal relationship between genes, but gives inferences as to where potential interactions occur. Co-regulated genes are likely involved in similar functional processes, which can be annotated. Combined with differential expression analysis, groups of genes that are different and function together can be identified. This was invaluable in the process of identifying differences in the expression patterns of cell subsets and identifying potential HIV-1 control-related mechanisms.

Weighted Gene Co-expression Network Analysis is the most widely used co-expression approach (WGCNA) (Langfelder and Horvath, 2008). It is thought to be one of the most robust methods for identifying co-expressed genes (Li et al., 2018), and as a result was selected for use in this project. In addition to WGCNA, this project explored the use of unsupervised machine learning methods for identifying co-expressed gene sets. These approaches allow for clustering of high dimensional datasets and are highly effective at identifying similarities in datasets. This offered the invaluable potential for identifying genes with similar expression patterns. Therefore two methods were selected for comparison against WGCNA. These were a Self Organising Map (SOM) artificial neural network overlaid with Hierarchical Agglomerative Clustering (HAC) and k - means clustering.

A SOM was selected as it is effective at reducing higher-dimensional datasets for clustering of thousands of data points (genes) with multiple attributes (samples). Furthermore, SOM can generate heat maps that represent the clustering of thousands of data points. This was useful in visualising genome-wide expression values of an entire cell subset. SOM is highly efficient at clustering and has potential for use at identifying commonalities in expression patterns. To bolster the efficacy of each SOM, HAC was applied as another layer on each map. HAC has been proven to be effective in co-expression analysis studies and is also used in WGCNA (Langfelder and Horvath, 2008), and therefore combining it with SOM was thought to improve the effectiveness of gene clustering.

K-means was selected as it is one of the most commonly used benchmarking clustering methods (Pirim et al., 2012). Furthermore, it has been shown to perform faster and result in better biologically enriched clusters than many other methods (Richards et al, 2008). On this basis, it was selected for comparison with SOM and WGCNA.

2.7.1 Self Organising Map

Self-organising maps are based loosely on how a human brain works. This involves a set of neurons being activated in response to a set of input patterns. When exposed to a similar set of patterns, the same neurons are again activated. This approach allows for a method for grouping similar items as they map to the same neurons. This is effective in clustering datasets that have multiple attributes for a given data point. Such an approach was ideal for clustering genes with multiple expression values (biological replicates) that could then be clustered to identify co-expressed gene sets.

To model gene expression, genes with multiple data points for every sample expression value were exposed to a set of neurons. Neurons contained weight vectors in a vector space with randomised starting values. Training occurred when the euclidean distance between all neurons and a specific gene was calculated. The neuron that was closest to the gene (in higher dimensional space) was labelled as the best matching unit (BMU). The weight vector of this neuron was then updated to become closer to the expression values of the gene. Neurons near the BMU also update thus pulling the entire map of all neurons towards the gene data point in the vector space. This process was repeated until an entire SOM was associated with every gene. This was performed iteratively, until an optimum distance between neurons and genes was reached.

To control the degree to which neuron weight vectors are updated as a parameter, alpha (the learning rate) is changed. A decreasing alpha rate is preferred as initially large distance changes between neurons and genes are needed and later only small changes to find the optimum map are needed. Epochs refer to the number of times neurons are updated and can be controlled. To identify when neurons appear to be the closest to the genes in the vector space, the mean distance between neurons and data points can be calculated. This is plotted against the number of iterations and can be used to identify when the distance is no longer decreasing (Figure S5). At this point, it is likely that the map has reached its optimum. Following SOM generation, HAC was run over neurons to cluster them further into broader groupings. This enabled a larger scale for gene co-expression analysis.

The R package, Kohonen (Wehrens and Kruisselbrink, 2018), was used to run self-organising maps across cell subsets. Two SOM's were trained, one for a method comparison with WGCNA and k-means, and a genome-wide SOM using all 30 157 genes for all 24 samples to visualise genome-wide expression.

2.7.1.1 SOM for Comparison of Methods

The VST count matrix was subsetted to include only DE genes for the comparisons T_{CM} vs T_{EM}, T_{CM} vs T_N, and T_N vs T_{EM}. Counts were manipulated into a training matrix and metadata objects for the SOM. It was decided to use a 25 x 25 hexagonal grid for SOM comparison, where 625 neurons (nodes) were used for clustering. This selection was arbitrary, but suited heatmap generation in a readable format across multiple cell subsets. For the comparison, SOM object genes were trained with an alpha value (learning rate) decreasing from 0.1 to 0.01 (the recommended metric for a 25 x 25 neuron map). The number of iterations and the mean distance to closest units were plotted to identify when an optimum map was reached. This was identified after 2000 epochs. HAC was then run on top of the map with the R package, fastcluster (Müllner, 2013). Within-group sum of squares plots were used to determine the optimum number of hierarchical clusters to be used. SOM heatmaps were made to visualise differences for the method comparison and a table was produced with the gene cluster assignments.

2.7.1.2 SOM for Genome-Wide Expression

For a genome-wide expression visualisation, a SOM was trained with all genes, for all samples. VST counts were used to train the model, with learning rate decreasing from 0.1 to 0.01. This model ran for 5000 iterations before the mean distance between units plateaued. SOM heatmaps were plotted and coloured on the basis of sample type as well as node distance quality plots, neighbour distance plots and code spread plots. These plots were generated with the R packages, Kohonen (Wehrens and Kruisselbrink, 2018) and ggplot2 (Wickham, 2011), and the colour palettes used for plotting were generated with rgdal and rgeos (R Core Team, 2018).

2.7.2 K-means Clustering

K-means clustering requires specification of the number of clusters to be used in partitioning data (k). To determine the optimum number of clusters within the sum of squares plots were created. This allowed for testing of between 1 and 15 potential clusters. This plot gives an indication as to the optimum number to partition the data. Once a value for k was selected, k-

means randomly generates centroids (centre values) and assigns data points to clusters. This is analogous to a SOM except in 2-dimensional space. It then modifies the location of centroids based on the sum of squared distance between data points to the closest centroid (Pirim et al., 2012). This is continually updated until an optimum for centroids is reached and data points are assigned to each centroid.

Clustering using k-means, with multiple attributes per gene data point, required additional dimension reduction in the dataset. This was achieved for this project by averaging all expression values for a given gene to allow clustering at a uni-dimensional level. The R package, stats (R Core Team, 2018), was used to run k-means clustering. Within the sum of square plots were generated for every comparison. K-means then clustered genes using a maximum number of 50 iterations, for 50 randomly sampled sets.

2.7.3 Weighted Gene Co-expression Analysis Algorithm and Implementation

The WGCNA algorithm is complex, consisting of several substeps: similarity matrix construction, adjacency matrix construction and clustering for module detection. WGCNA was used to produce gene co-expression cluster assignments to be compared with the other two methods.

To identify genes that are co-expressed, WGCNA requires an adjacency matrix, a_{ij}, that encodes gene co-expression networks with nodes as genes and edges as connection strength between genes. Such a matrix contains entries in [0, 1], where a_{ij}denotes the connection strength between nodes i and j (Langfelder and Horvath, 2008). To produce an adjacency matrix a similarity matrix, s_{ij} , is first created.

There are different similarity operations that can be used to generate a similarity matrix. Conventionally similarity (s_{ij}) is calculated as the absolute value of either a Pearson or Spearman correlation coefficient between two nodes, x_i and, where$s_{ij} = |cor(x_i, x_j)|$ (Langfelder and Horvath, 2008). This had the downside of not incorporating the distance between expression values, so this study opted for an alternate implementation described by (Hughitt, 2016). This approach applied equal weighting to both Pearson correlation and Euclidean distance values, resulting in a value between [-1, 1] (Hughitt, 2016). Values close to 1 were associated with positive correlation and values close to -1 with negative correlation. This is detailed in following R function equation:

$$S = sign(cor(X)) \times \frac{|cor(X)| + (1 - \frac{log(dist(X) + 1)}{max(log(dist(X)+1))}}{2} \quad (5)$$

The similarity S is calculated from a variance stabilised counts matrix, X, with samples as columns and genes as rows. $cor()$ calculates the pearson correlation coefficient matrix and $dist()$ returns a euclidean distance matrix. $max()$ returns the maximum value, while $sign()$preserves the sign of the Pearson correlation coefficient. Similarity matrices were visualised as heatmaps for all comparisons.

Adjacency matrices were formed for all comparisons that were used to describe co-expression relationships between all genes. Adjacency matrix construction can either be binary (unweighted network) or continuous (weighted). This project opted for a weighted approach, where continuous values are calculated that demonstrate the level of co-expression. To calculate adjacency, this project used a power transformation as described by Zhang and Horvath (2005). This raises similarity calculations to a power, reducing small similarity values and increasing large correlations, thereby acting to filter out genes with low correlation values and leaving genes that are likely co-expressed with high correlation values. To calculate adjacency similarity, values from [-1, 1] were first scaled to [0, 1] and then raised to a soft threshold power, γ (with $\gamma > 1$). This is described in the following equation (Zhang and Horvath, 2005):

$$a_{ij} = \left(\frac{1}{2}(1 + s_{ij})\right)^{\gamma} \quad (6)$$

To decide the power (γ)or soft threshold, powers from 1 to 20 were run through the pickSoftThreshold WGCNA function (Langfelder and Horvath, 2008). This enabled two metrics to be plotted and studied: scale-free topology fit and mean connectivity. Scale-free topology is a property that most correlation networks adhere to (Zhang and Horvath, 2005), and is found to exist within biological systems (Barabási and Oltvai, 2004). Hence should a network have a high scale-free topology fit (close to 1), it is likely that it resembles the structure of an accurate biological network. Scale free topology refers to a property of networks that maintain centrality, i.e. there is an exponential relationship between degree of connectivity of a node and its frequency.

Scale free topology is represented by a power law, where the probability that a node is connected with another node - k is $p(k) \sim k^{-\gamma}$, i.e. there is an exponential relationship

between degree of connectivity of a node and the frequency of its occurrence. As scale free topology approaches 1, mean connectivity of all nodes decreases. Hence when choosing a threshold one needs to select a power that calculates a high scale-free topology, but does not lose all connectivity (the ability to detect co-expression).

Modules are clusters of highly interconnected genes (Langfelder and Horvath, 2008). To identify modules, a definition as to what separates them is needed. The most widely used method is the topological overlap similarity measure (Ravasz et al., 2002). It was selected for use in WGCNA (Langfelder and Horvath, 2008) as it results in biologically meaningful modules (Zhang and Horvath, 2005). Ravasz et al. (2002) describes the topological overlap measure (TOM) a similarity calculation that is effective in discerning modules. The TOM of two genes, i and j, is defined as w_{ij} (Equation 6), where l_{ij} (Equation 7) describes the number of nodes to which both i and j are connected, and k_{ij} (Equation 8) describes the number of direct connections to other nodes (calculated as the sum of all connected genes - a_{ij}). This is demonstrated in the equations from (Zhang and Horvath, 2005) as follows:

$$w_{ij} = \frac{l_{ij} + a_{ij}}{min\{k_i, k_j\} + 1 - a_{ij}} \tag{6}$$

$$l_{ij} = \sum_u a_{iu} a_{uj} \tag{7}$$

$$k_{ij} = \sum_{j=1}^{n} a_{ij} \tag{8}$$

This calculation is a similarity value (Rousseeuw and Kaufman, 1990). To convert it to a dissimilarity measure for gene clustering, it is subtracted from 1 (Equation 9). This is then formed into a matrix containing all dissimilarity values, for all gene pairs compared in the co-expression experiment.

$$d_{ij}^{w} = 1 - w_{ij} \tag{9}$$

The dissimilarity matrix can then undergo average linkage hierarchical clustering, in order to group networks of genes into modules. This was performed using the fastcluster function, hclust (Müllner, 2013). To separate a hierarchical clustering dendrogram into clusters of gene modules, the dynamic branch cutting function, cutDynamicTree (Langfelder et al., 2008) was

used with a minimum module size of 30 genes. This uses a dynamic algorithm that separates clusters based on their shape within a dendrogram (Langfelder et al., 2008).

Separated modules were assigned labels and hex and Red Green Blue (RGB) colour codes were assigned for later visualisation. Coloured dendrograms were plotted and analysed, in order to determine the optimum minimum number of genes per module. Dendrograms also served as a quality assessment of module detection. The R package, Homo.sapiens (R Team, 2015), was implemented to add known gene metadata to the co-expression data frame. This included gene names, symbols, and chromosome locations of genes to be used in further analyses. Genes were ordered by the $\log_2$ fold change value produced by DESeq2, and a further column was added, mapping continuous colouring to these values. WGCNA was then run through an additional script to convert files to graphml format for later visualisation.

2.8 Co-expression Method Evaluation

2.8.1 Selection of Genes and Comparison Design

To make an effective comparison of the three co-expression methods, a selection of genes was needed. This study used DE genes identified from the comparisons between $CD4^+$ T cell subsets. DE genes were clustered through each co-expression method and the cluster/module assignments for every gene were annotated across the three methods. These could then be compared using an additional workflow in R.

To compare the different approaches, known biological information was used to validate how well a particular method identified genes as co-expressed. This was based on the premise that if a method identified a cluster of genes as co-expressed, it was likely that there was already some evidence confirming known co-expression of those genes. Therefore, validation of the ability of a method to identify co-expressed genes could be undertaken, using an external database with known interactions between genes. To quantify a metric for comparison, the ratio of genes identified as co-expressed could be compared to those in a known pathway or interaction. The mean of these ratios for all groups of co-expressed genes could be used to compare the different methods. A suitable database for this purpose was the Gene Ontology (GO) Resource (The Gene Ontology Consortium, 2019), and the use of GO enrichment could be used to determine ratios for method comparison.

2.8.2 GO Enrichment

The most comprehensive database of known gene functional interactions is the GO resource (The Gene Ontology Consortium, 2019). GO terms annotate groups of genes through three main classifications: biological processes, molecular function and cellular location. GO enrichment determines whether the occurrence of a specific term appears at a significantly higher incidence in the query dataset compared to a baseline genome known as a universe. This is achieved where each GO term is assigned a statistical enrichment score and should this score be greater in the query set than the genome, the term is said to be GO enriched. GO enrichment provided an ideal framework for comparing the co-expression methods. Should a group of genes be co-expressed, it was likely that there was an existing GO term that already defines known interactions or co-expression relationships between them. Even if this was not the case for all genes it was likely it was the case for most genes identified as co-expressed. Hence the mean of the gene ratios from multiple clusters for a particular method could be calculated and compared.

2.8.3 Comparison Workflow

To compare methods, cluster assignment data frames - produced from each co-expression method for the three cell subset comparisons - were imported into R. Gene identifiers from a particular module were then isolated and converted from Ensembl to Entrez identifiers using the R package, biomaRt (Durinck et al., 2009). GO enrichment was then undertaken using the R package, clusterProfiler (Yu et al., 2012). Enriched terms were then written to CSV files. This meant all clusters/modules underwent GO enrichment for all three cell comparisons, for all three methods. Gene ratio enrichment scores were then extracted from all clusters, and box and whisker plots plotted for comparison using ggplot2 (Wickham, 2011). Distributions of ratios were compared using the Analysis of Variance (ANOVA) (Kaufmann and Schering, 2014) test. Results showed that WGCNA was the best approach for identifying differentially co-expressed genes and therefore it was selected for further analysis of the differences between $CD4^+$ T cell subsets.

2.9 T_{EM} vs T_N $CD4^+$ Cell Subset Comparison

When analysing differential expression analysis results, it became clear that the T_{EM} vs T_N $CD4^+$ T cell comparison had the greatest number of DE genes. This was interesting as memory cells are known to harbour latent infected virus, while T_N cells tend not to. Activation of T_{EM}

cells is also known to spur replication of HIV-1 (Kulpa et al., 2019). On this basis, the comparison was selected for study in greater depth using the method that was identified as the most effective, WGCNA (Langfelder and Horvath, 2008).

2.9.1 Hub Gene Identification

Within the T_{EM} vs T_N $CD4^+$ T cell comparison hub genes were identified using WGCNA. These are genes that represent a high level of centrality and are the most connected within a given WGCNA module. They also are highly connected with other modules and in a biological context, represent genes that likely influence expression or function of other genes. To identify hub genes, eigene values were calculated with WGCNA hub gene identification function.

2.9.2 Visualisation of WGCNA Networks

All network graphml files for the cell subset comparisons were visualised using Cytoscape 3.7.1 (Shannon et al., 2003). Cytoscape was a useful tool for visualising networks of co-expressed genes. Networks were coloured by modules and log_2 fold change scores, and edges (connections between genes) were coloured by weight scores produced from WGCNA. Results were exported to png format for analysis.

2.9.3 Identification of Notable GO Terms

Within the $CD4^+$ T_{EM} vs T_N comparison, lists of GO enriched terms identified from each coexpressed module using clusterProfiler (Yu et al. 2012) were aggregated into a single data frame. This enabled analysis of all GO terms associated with differentially expressed genes for the cell subset comparison. GO enrichment lists, hub genes and visualised networks were screened to identify potentially interesting modules of co-expressed genes. Modules of interest underwent further screening for high levels of connectivity (as visualised in Cytoscape (Shannon et al. 2003), the presence of more than 10 genes within a module, and high clusterProfiler gene ratio scores. Modules that passed this three-point screening step were selected for further visualisation. For this purpose, clusterProfiler (Yu et al. 2012) CNET plots, heatmaps and GO term bar plots were produced. Cytoscape networks were coloured by log2 fold change per module, and edges coloured by WGCNA weight values. Hub genes were imported into STRING v11 (Szklarczyk et al., 2019) (a protein interaction resource) for visualisation of known gene interactions and analysis.

2.10 Code Availability

Access to all code used in this study may be found at: https://github.com/alecstansell/rna-seq. Pseudo-code describing the core scripts used in this study may be found in Appendix B.

2.11 Package URL's

FastQC	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Kallisto	https://pachterlab.github.io/kallisto/
DESeq2	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
ClusterProfiler	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
STRING	https://string-db.org/
Kohonen	https://cran.r-project.org/web/packages/kohonen/index.html
WGCNA	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
R	https://www.r-project.org/about.html

3 Results

3.1 Quality Control of Raw Sequencing Data

Assessment of the RNA-seq data obtained from each of the 14 HIV-1 elite controllers (ECs) included in this study using FastQC (Andrews, 2018) and MultiQC (Ewels et al., 2016), allowed poor quality sequencing files to be identified and removed prior to performing any further analyses. From an initial set of 60 sequencing files, 32 files were flagged by FastQC. Reasons for this included: low mean phred scores, high adaptor content, an excess of reads shorter than 20 base pairs (bp), high sequence duplication levels and the presence of overrepresented sequences. For 26 of these files, poor quality sequencing data was removed through an iterative process of de-duplication of samples with sequence duplication levels greater than 60%, removal of reads with phred scores less than 20, trimming of adaptor sequences with Cutadapt v2.7 (Martin, 2011) and the removal of reads shorter than 20 bp.

Six samples, from four individuals (namely LIR_02, LIR_04, SCOPE_1492 and SCOPE_1475), had sequence data that failed to meet the quality control thresholds, even after processing. Sequencing data obtained from LIR_02 had very low per base sequence content, high levels of sequence duplication and poor GC content. The data obtained from LIR_04 had very high levels of sequence duplication, while those obtained from SCOPE_1492 had poor per base sequence content and those obtained from SCOPE_1475 had high levels of overrepresented sequences. As a result, these individuals were excluded from all further analyses.

3.2 Quantification of Transcript and Gene Level Abundance Estimates

3.2.1 Summary of Abundance Quantification

Following quality control and processing, all gzipped fastq files underwent quantification and alignment. A total of 446 080 703 reads across all 30 samples were mapped against 187 626 transcripts. Of these reads, 292 191 328 were successfully aligned and quantified. An average of 65.41% of reads was successfully pseudo-aligned to the reference transcriptome (Figure 3.1) and 18.96% were uniquely aligned. The R package tximport (Soneson et al., 2015) then summarised transcript counts to the gene level, leading to the quantification of reads to a total of 30 157 genes.

3.2.2 Sequencing Consistency Across Individuals and Cell Subsets

The percentage of reads pseudo-aligned across individuals was consistent with minimal variation across individuals (Figure 3.1). The mean proportion of pseudo-aligned across cell subsets was also consistent (Figure 3.2). This highlighted that kallisto (Bray et al., 2016) results were likely of high quality and there was minimal bias introduced as a result of mapping and quantification.

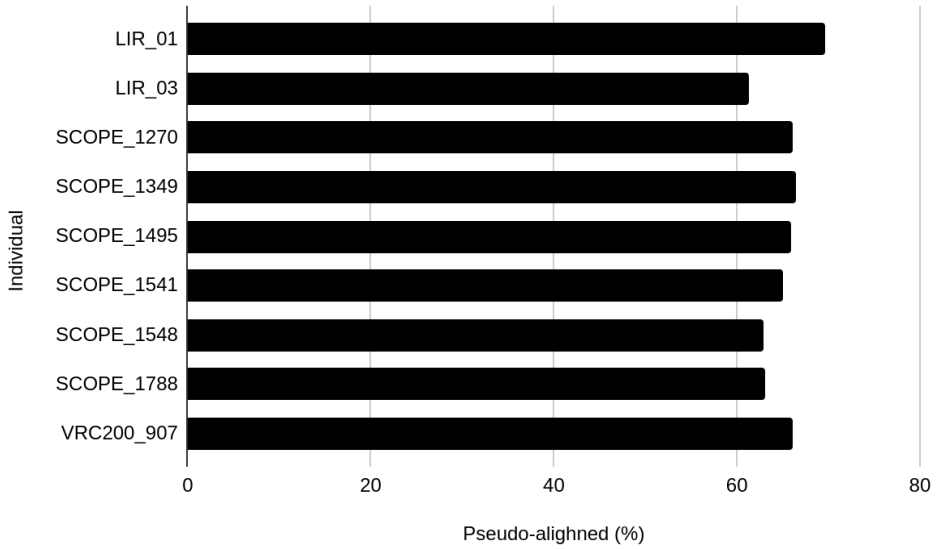


Figure 3.1: Proportion of pseudo-aligned reads per individual. showing the percentage of reads pseudo-aligned across the final set of individuals analysed in this study.

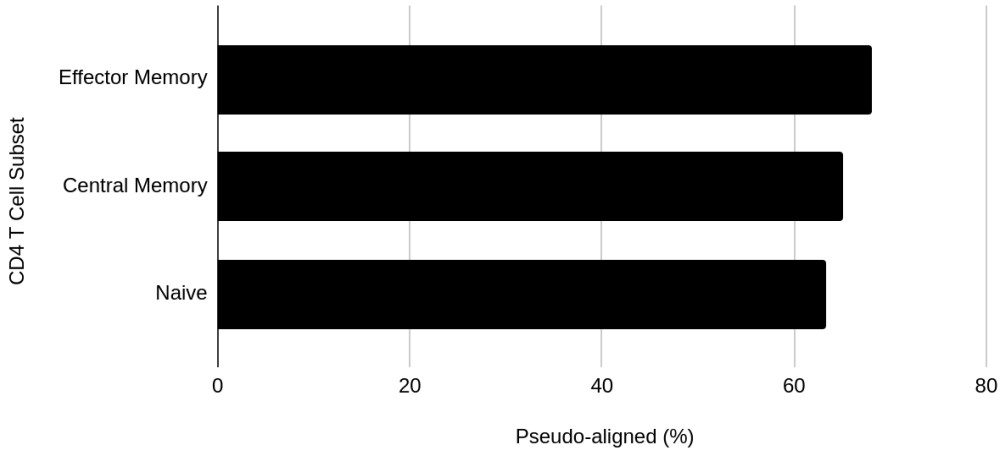


Figure 3.2: Proportion of pseudo-aligned reads per $CD4^+$ T cell subset. Showing the percentage of reads pseudo-aligned compared to all available reads per cell subset.

3.3 Batch Effect Identification

3.3.1 Principal Components Analysis Showed No Discernable Batch Effects

Batch effect biases were analysed using sleuth's Principal Component Analysis (PCA). PCA plots showed no apparent clustering for individuals (Figure 3.3A) or sample date (Figure 3.34B). Samples showed some clustering on the basis of cell subset (Figure 3.3C). Therefore it appeared that cells from the same subsets clustered more closely than different subsets within the same individual. This indicated that it was unlikely batch effects existed within the data.

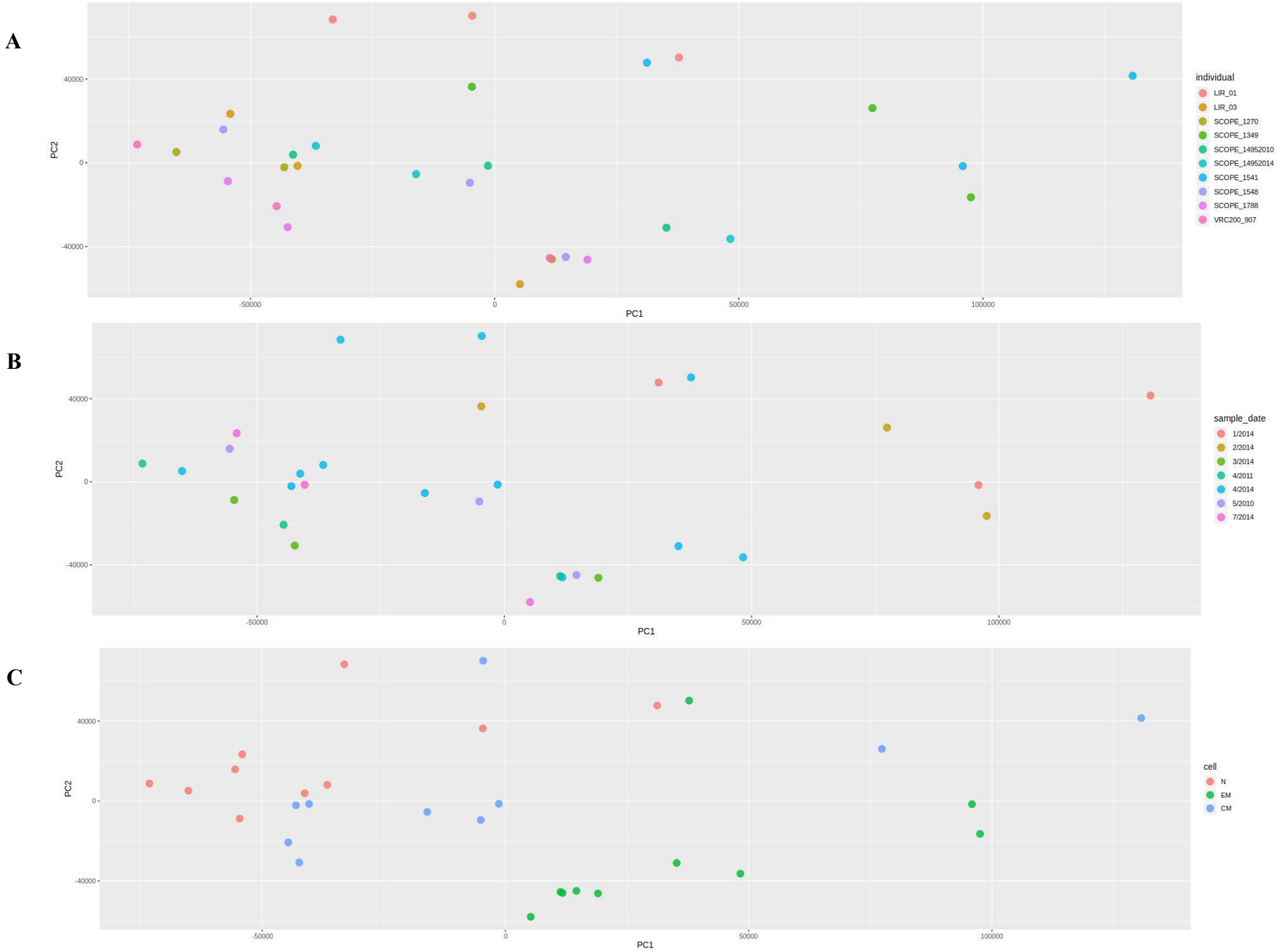


Figure 3.3: Principal components analysis with sleuth show no discernible batch effects. (A): PCA clustering coloured by individuals shows minimal to no clustering. Individuals are represented by different colours with three cell subsets per individual. (B): PCA clustering coloured by sample collection date does not exhibit clustering. Sample collection date for cell sample subsets are shown in different colours (C): PCA clustering coloured by cell type. Different cell types are coloured with N, EM and CM representing Naïve, Effector Memory and Central Memory $CD4^+$ T cells.

3.4 Outlier Detection and Removal

3.4.1 Removal of Outliers and Selection of the Final Data Set

Euclidean distance hierarchical clustering produced a sample dendrogram for outlier analysis. Two samples, LIR_01-PBMC_EM and VRC200_907-PBMC_N, were identified as outliers from the sample dendrogram due to large dendrogram height differences (Figure 3.4). Both samples clustered separately to the rest of these samples. Outliers were likely to negatively impact differential and co-expression analyses which require intra-sample variance to be minimised (Langfelder and Horvath, 2008). As a result, the sample distance tree was cut at a dendrogram height of 110 000 (Figure 3.4). To maintain consistency, all datasets from the individuals LIR_01-PBMC and VRC200_907-PBMC were excluded from further analyses. This resulted in a total of 24 samples (eight individuals across three cell types), with quantified read data for 30 157 genes available for all further analyses.

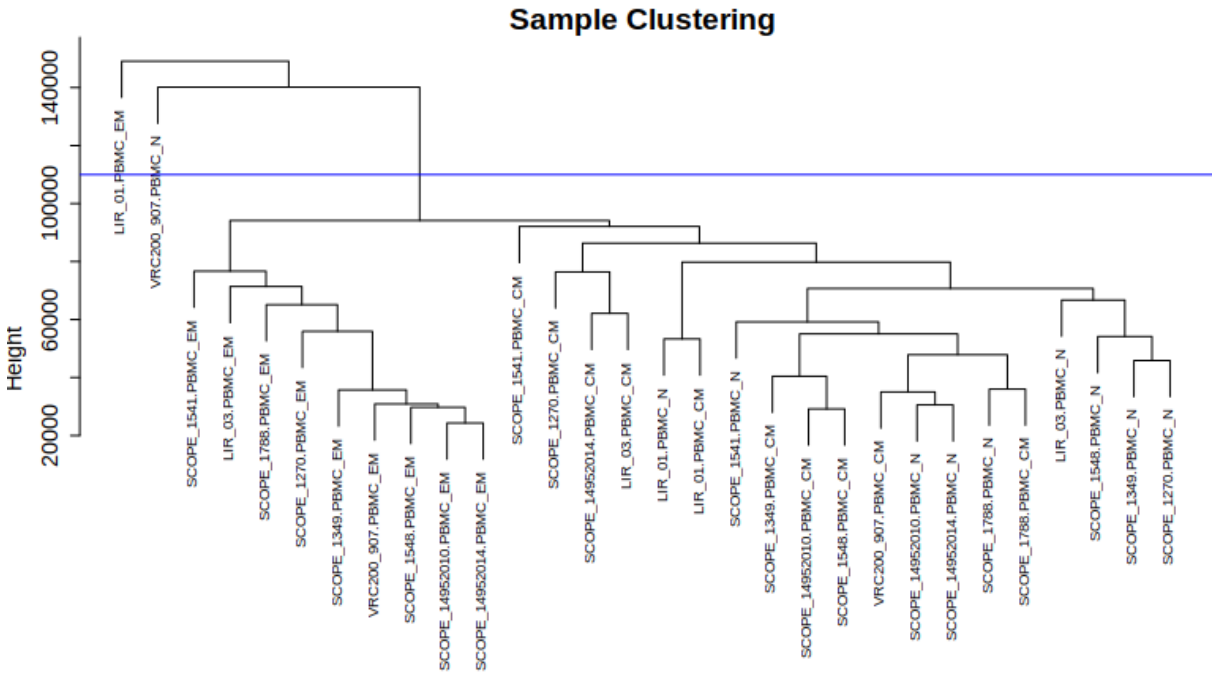


Figure 3.4: Euclidean distance hierarchical clustering dendrogram showing clustering of samples. Height values indicate differences in the similarity of samples. Two samples, LIR_01-PBMC and VRC200_907-PBMC, were identified as outliers. A line drawn at the sample distance height of 110 000 was used to select and remove outlier samples.

3.5 Normalisation

3.5.1 Comparison of VST and rlog Transformed Counts

Raw counts showed high levels of heteroscedasticity - where variance varied across mean count values. Two normalisation approaches were compared: Variance Stabilised (VST) and Regularised Logarithm Transformations (rlog). VST and rlog counts both reduced substantial variation that existed in the raw count data set (Figure 3.5). VST reduced the most variation from raw counts (Figure 3.5 and Figure S1) and therefore was selected to transform data for all further co-expression analysis.

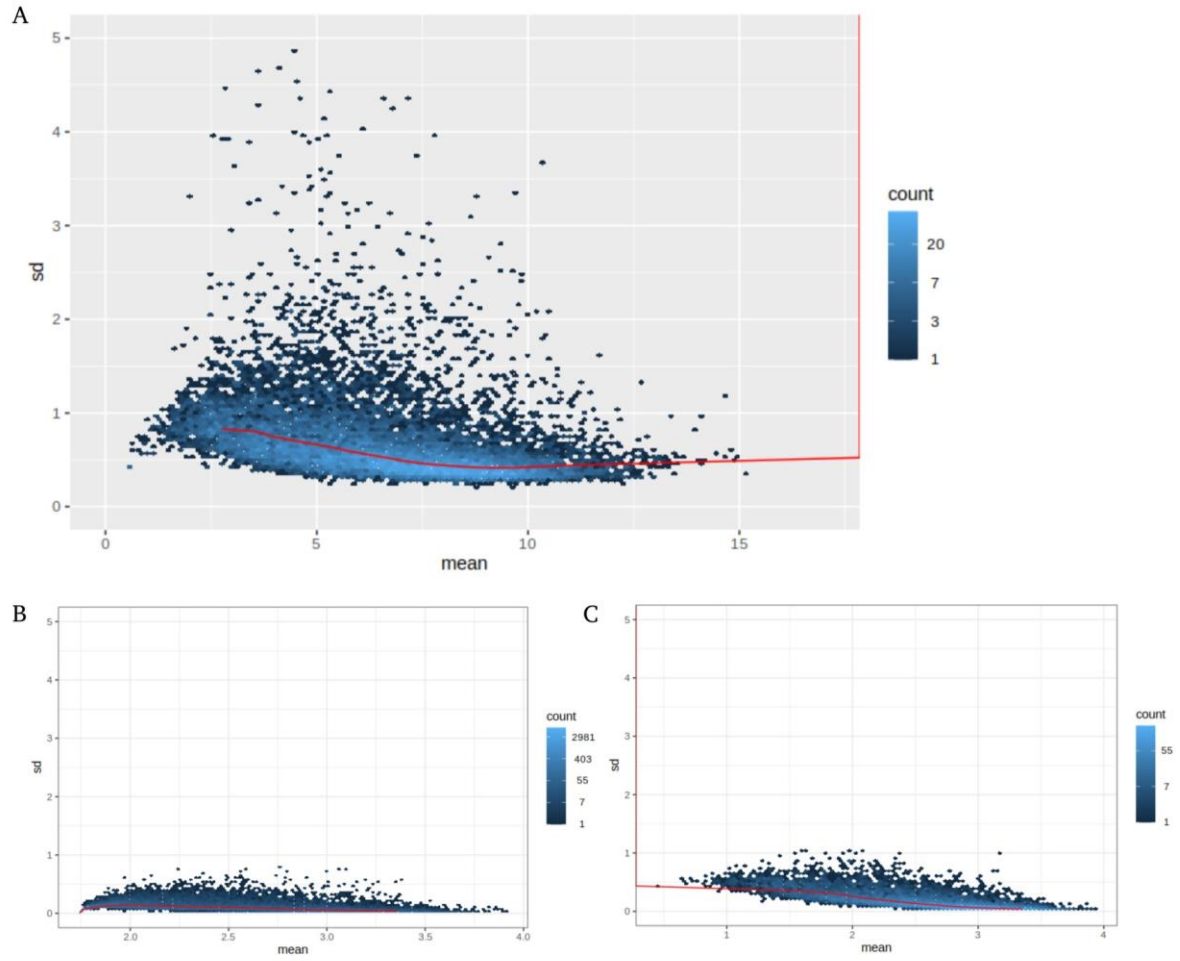


Figure 3.5: Mean - standard deviation plots showing differences in raw, VST and rlog counts. (A): Raw counts (B): Variance Stabilised Counts and (C): rlog Normalised Counts. Red lines indicate a curve plotted to model the mean standard deviation relationship. For each gene counts of reads are visualised in blue.

3.5.2 Heatmaps Show Sample Clustering

Heatmap analysis of samples with raw and VST normalised counts both showed hierarchical clustering on the basis of cell subsets (Figure 3.6). Evaluation of the Pearson correlation coefficient (PCC) value distributions (Figure 3.6) showed a high level of positive correlation between all samples. Raw counts had a mode PCC value of 0.9, rlog normalised counts had a value of 0.994 (Figure S2) and VST counts had a value of 0.97 (Figure 3.6). WGCNA (Langfelder and Horvath, 2008) documentation recommends using VST normalisation for processing due to the reduction in noise caused by high variation. On this basis, and based on the evidence from mean - standard deviation plots, heatmaps and distribution plots; VST counts were selected for use in all further co-expression analysis.

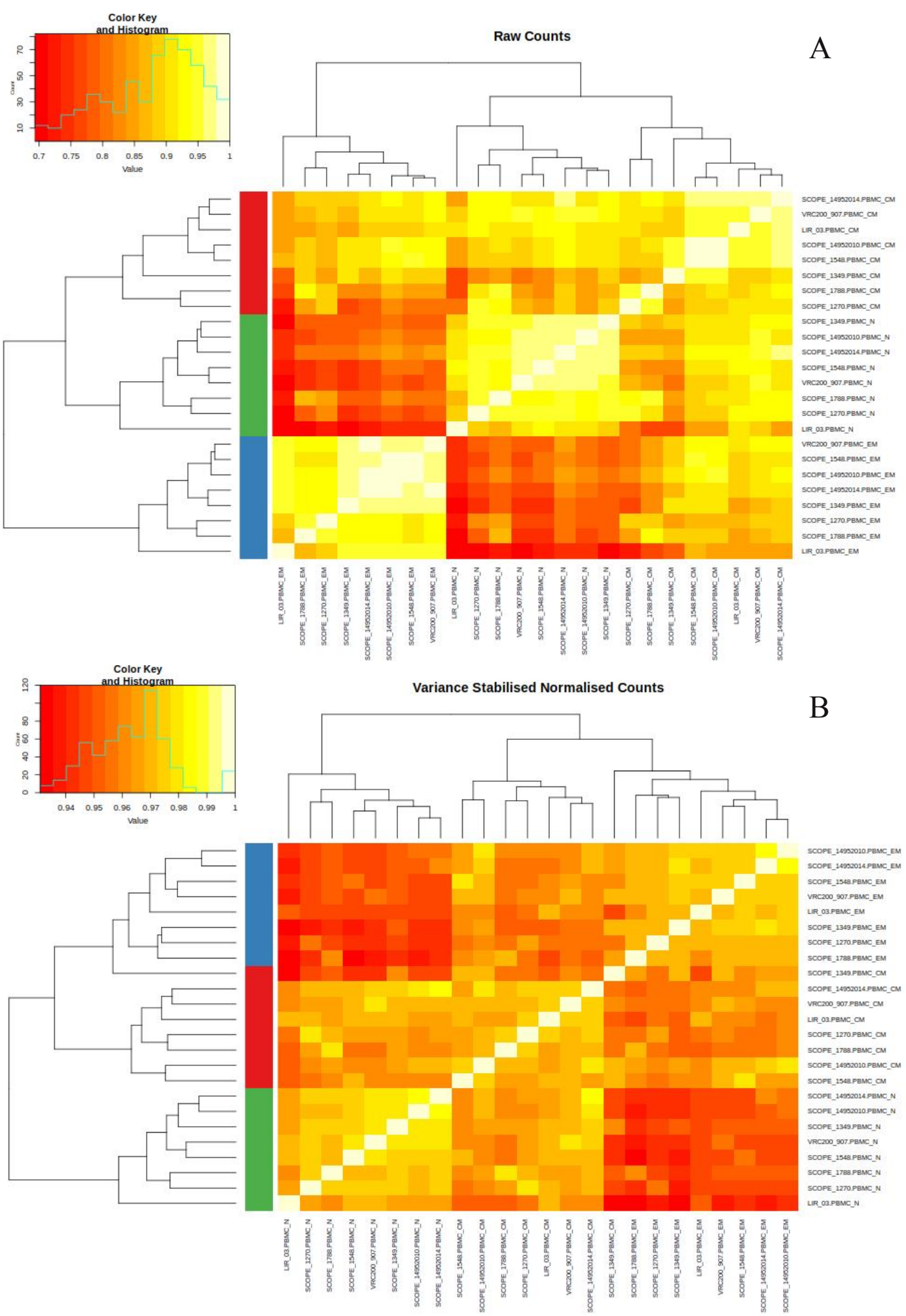


Figure 3.6: Hierarchical clustering heatmaps for raw and VST normalised counts. (A): Raw Counts and (B): Variance Stabilised Counts. Left: red green and blue colours indicate the cell type samples belong to. Colours depict sample PCC's. Values close to 1 (high correlation) are shown in light colours and closer to 0 (low correlation) as darker colours. Heatmaps show the differences in raw and normalised counts.

3.6 Differential Gene Expression Analysis

3.6.1 DESeq2 Quality Analysis

Quality plots from the DESeq2 experiment showed that the analysis had been run effectively. Mean estimate dispersion plots (Figure S3) showed DESeq2 had effectively shrunk gene estimates to reduce intra-sample variance. DESeq2 M (log ratio) A (mean average) plots showed symmetry in the numbers of up and down-regulation of genes (Figure S4). Log score shrinkage towards zero also was evident (Figure S4), with tests likely to result in errors removed with independent filtering. This highlighted consistency in the experiment, with no bias towards up or down-regulation of genes. Pairwise PCA showed pairwise clustering of the cell subsets, highlighting their differences (Figure 3.7). Hence Quality plots and PCA analysis indicated that the identification of differentially expressed genes was effective.

3.6.2 Identification of Differentially Expressed Genes

Differential gene expression analysis identified a total of 4550 differentially expressed (DE) genes across all $CD4^+$ T cell subsets. The Effector Memory (T_{EM}) vs Naive (T_N) comparison had the highest number of DE genes with 2760 genes, followed by the T_{EM} vs Central Memory (T_{CM}) comparison with 1018 genes, and the T_{CM} vs T_N comparison with 772 genes (Figure 3.8; Figure 3.9). Pairwise PCA clustering supported these findings, where the most variation in clustering was found between the T_{EM} vs T_N subsets followed by T_{EM} vs T_{CM} and T_{CM} vs N (Figure 3.7).

3.6.3 Overlap of Differentially Expressed Genes Across Comparisons

Out of all differentially expressed genes identified, only 179 were common to all three comparisons. This illustrated that differences between cell subsets were not as a result of the varying expression of the same set of genes. 117 genes were unique to the EM vs T_{CM} comparison, 93 were unique to the T_N vs T_{CM} comparison and 1413 were unique to the T_N vs EM comparison (Figure 3.9).

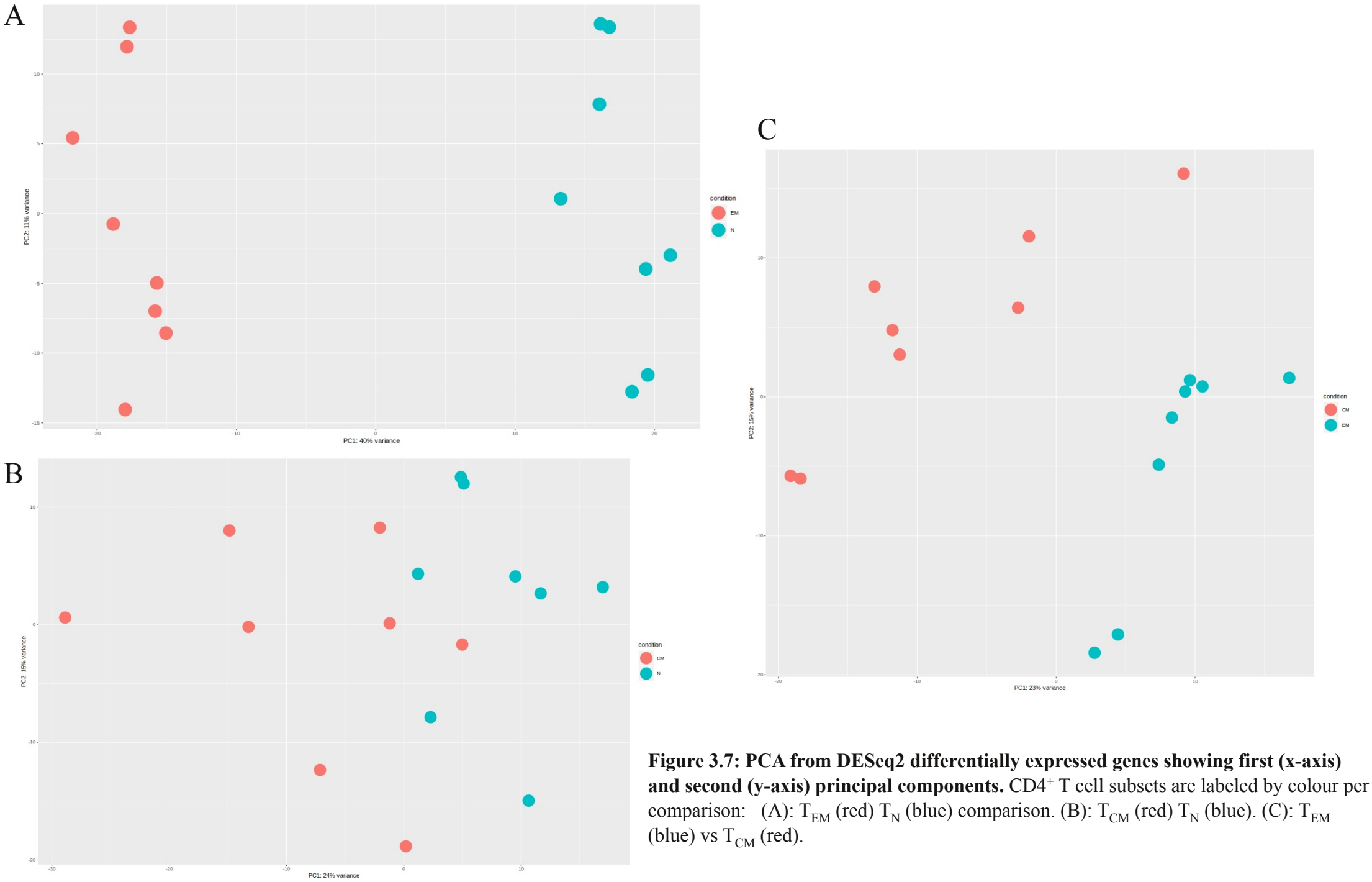


Figure 3.7: PCA from DESeq2 differentially expressed genes showing first (x-axis) and second (y-axis) principal components. $CD4^+$ T cell subsets are labeled by colour per comparison: (A): T_{EM} (red) T_N (blue) comparison. (B): T_{CM} (red) T_N (blue). (C): T_{EM} (blue) vs T_{CM} (red).

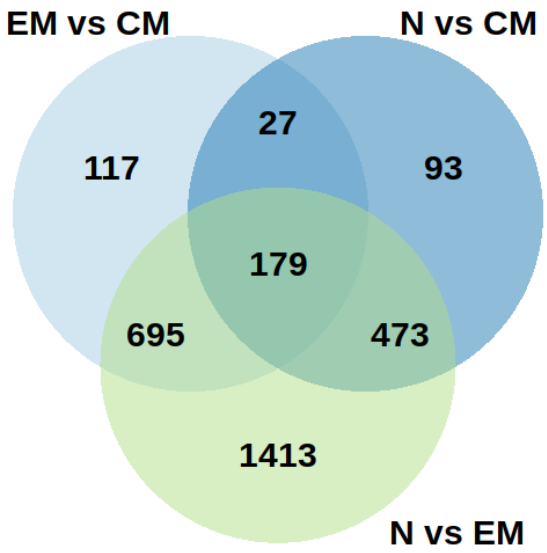


Figure 3.8: Differentially expressed genes identified across all three pairwise CD4⁺ T cell subset comparisons. The diagram shows T_{EM} vs T_{CM}, T_N vs T_{CM}, and T_N vs T_{EM} comparisons, highlighting the number of overlapping genes between the sets.

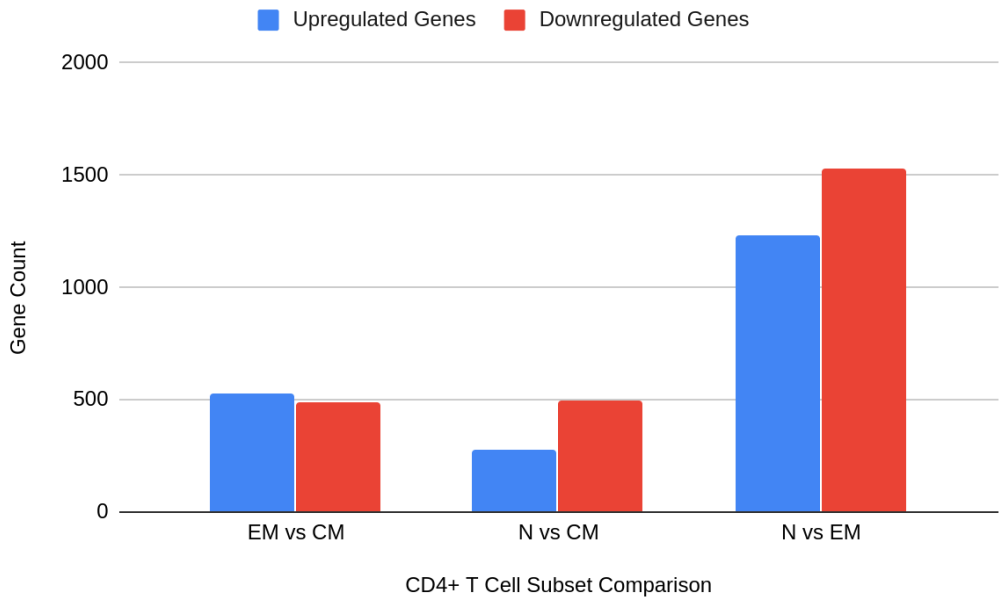


Figure 3.9: Number of differentially expressed genes identified across all three pairwise for CD4⁺ T cell subset comparisons (T_{EM} vs T_{CM}, T_N vs T_{CM}, and T_N vs T_{EM}). Upregulated genes are shown in blue (for the first cell type against the second), while downregulated genes are shown in red.

3.7 Evaluation of Methods For Identifying Co-expressed Genes

It was important to identify the best co-expression methodology applicable to understanding the transcriptional differences across cell types. Two unsupervised machine learning methods (SOM and k-means clustering) were compared against Weighted Gene Co-expression Network Analysis (WGCNA), using differentially expressed genes identified with DESeq2 for three cell subset comparisons.

3.7.1 WGCNA

Within WGCNA, quality control plots were produced to evaluate the effectiveness of analyses. This included heatmaps of similarity and adjacency matrices, scale-free independence plots and mean connectivity plots (Figure 3.11). Analysis of these indicated distinct clustering of genes in quality control heatmaps, good scale-free topology fit and high mean connectivity. This showed that the WGCNA results for the comparison were of high quality and were satisfactory for a comparison with the other two methods. WGCNA produced cluster dendrograms for visualisation of modules of co-expressed gene sets (Figures 3.10 and 3.11). WGCNA was then used to produce a table of cluster/module assignments for each gene. This was inputted into a pipeline for method comparison.

3.7.2 K-means Clustering

Prior to clustering with k-means, a within the sum of squares plot determined the ideal number of clusters (k) to be five. A visualisation of the clustering was produced showing distinct groupings with each cell comparison (Figure 3.10A). As with WGCNA, a table was produced containing assignments of the clusters to all differentially expressed genes for methodology comparison.

3.7.3 Self Organising Map

A self-organising map was run for 2000 iterations, until the mean distance between nodes plateaued for all differentially expressed genes. Expressed genes were grouped together by decreasing the distance between weight vectors and the genes in higher-dimensional space. These clustered genes were then mapped back to nodes on a two-dimensional graph. Graphs with 25 x 25 nodes were generated using the R package, Kohonen (Wehrens and Kruisselbrink, 2018), and coloured based on the density of genes mapping to each node for a particular sample. These heatmaps contained neurons showing scaled variance stabilised counts of groups of

genes that mapped to each neuron (Figure 3.10 C; Figure 3.12). This figure was overlaid with lines indicating further Hierarchical Agglomerative Clustering (HAC) assignment. Gene lists were produced that reflected both the SOM and HAC cluster assignments, to be used in a comparison with the two other methods.

SOM heatmaps were also used to visualise the expression profile of all 30 157 genes for the entire cohort (as opposed to just DE genes in the method comparison). For this a map was trained using over 5000 iterations (Figure S5), for a period of 2466.83 seconds, with a distance between units decreasing from 0.0020 to 0.008, where further iterations lead to a plateau around 0.008. Twenty-four heatmap plots (1 per sample) were produced to evaluate differences in gene expression patterns across cell subtypes. SOMs produced heatmaps for each cell sample for comparison (Figure S6).

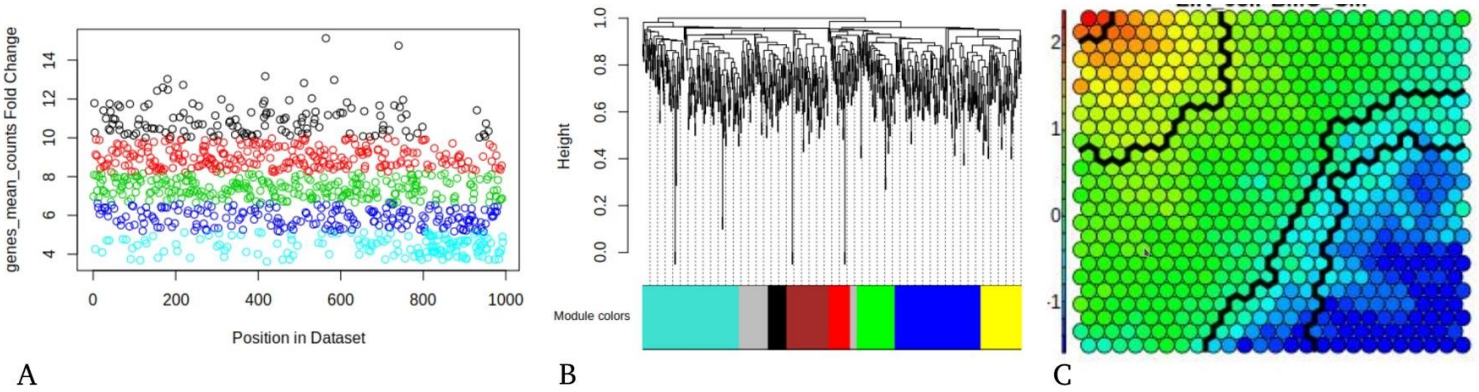


Figure 3.10: Comparison of three co-expression approaches. (A): K-means clustering with mean counts clustered by colour. (B): Weighted Gene Co-expression Network Analysis showing a dendrogram with identified modules. Each module represents a set of co-expressed genes. (C): Self-organising map heatmap with colours representing a scaled value that represents the underlying groups of genes variance stabilised count values. Lines show hierarchical clustering run on top of the self-organising map.

3.8 Gene Ontology Enrichment Method Comparison

Genes identified as co-expressed by WGCNA, SOM and k-means underwent Gene Ontology (GO) enrichment analysis with clusterProfiler (Yu et al., 2012) for their respective clusters or modules. This occurred for all three methods, for DE genes across all three cell types. GO terms were produced per module or cluster, and these were grouped together into a single data frame. GO enrichment gene ratio scores were plotted for each method, across the different methods for each cell type, and grouped for a final figure with all gene ratios per method (Figure 3.11). ANOVA (Analysis of Variance) statistical testing revealed that the distributions of gene ratios across the three methods were significantly different, with $p < 2.2 \times 10^{-16}$.

3.9 WGCNA was the Most Effective Co-expression Analysis Method

WGCNA (Figure 3.11) was identified as having the highest mean gene ratio of 8.92 %, followed by k-means with 6.87 %, and then self-organising maps with 6.35 %. The distributions of gene ratios for the respective methods were identified as significantly different using an ANOVA. On this basis, it was determined that WGCNA was the best method for co-expression network analysis for this dataset. Further analysis of differences between cell subsets was therefore conducted with WGCNA.

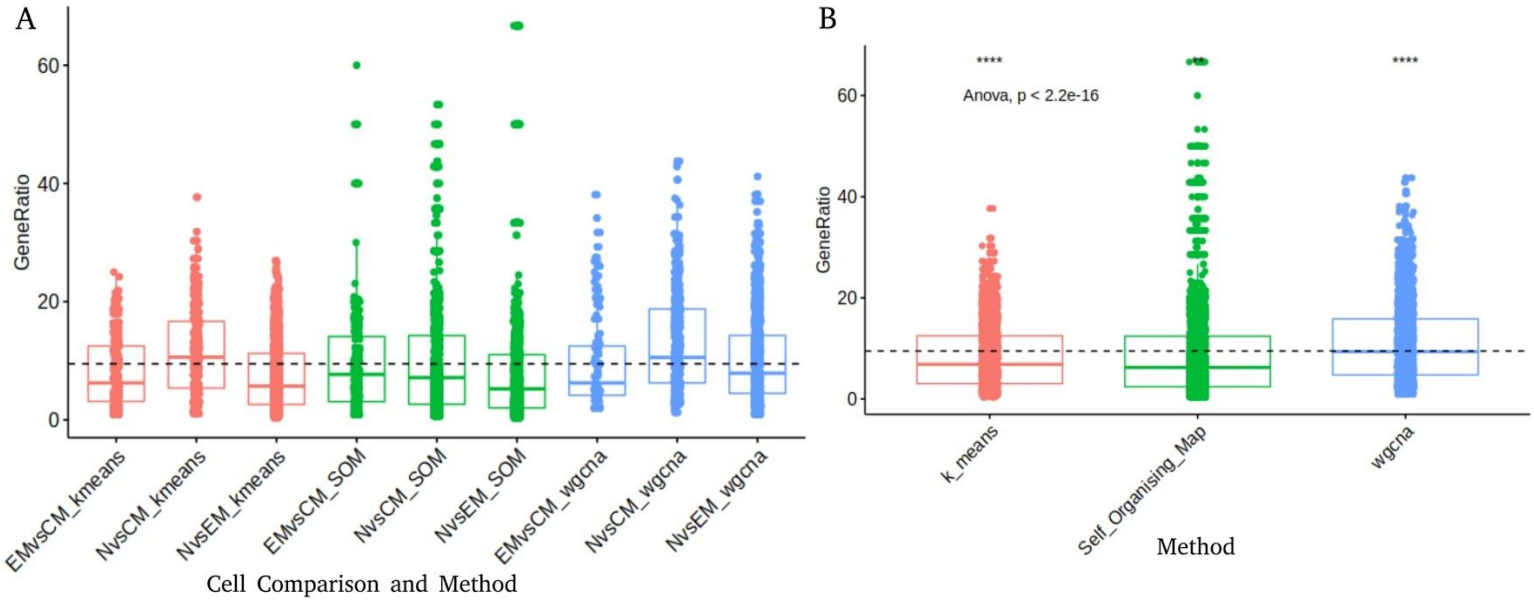


Figure 3.11: Box and whisker plot comparing Co-expression methodology using gene ratio scores. Within each module identified as differentially co-expressed GO term gene ratio scores are plotted. Colours indicate the different methods: k-means clustering (red), self-organising map (green) and WGCNA (blue). (A): Cell type comparisons across the three different methods. (B): Results from all cell type comparisons for a given method. ANOVA statistics show results are significantly different from one another. EM refers to Effector Memory Cells $CD4^+$ T cells, CM to Central Memory Cells $CD4^+$ T cells and N to Naïve $CD4^+$ T cells.

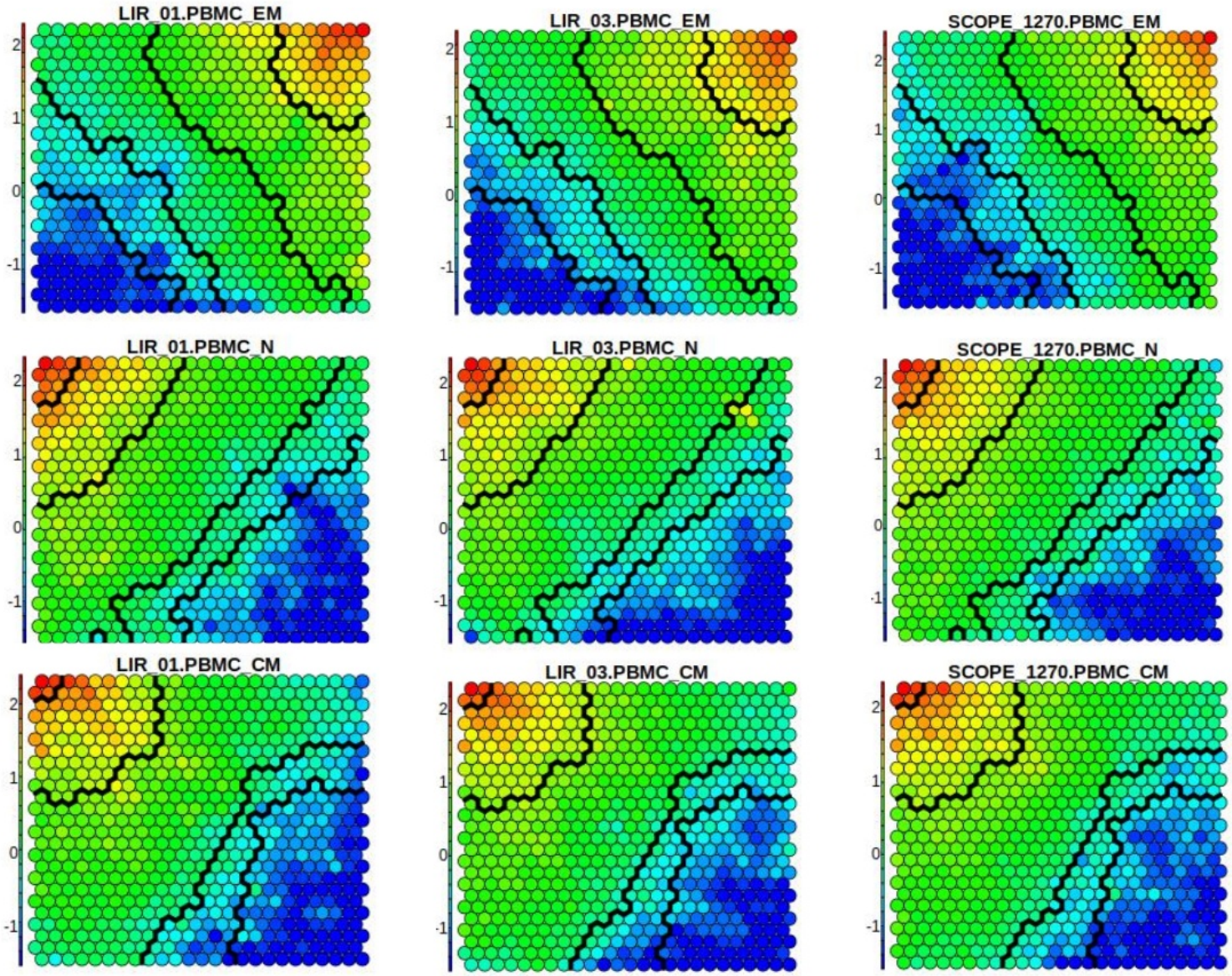


Figure 3.12: Self Organising Map heat maps for CD4$^+$ T_{EM} (top), T_N (middle) and T_{CM} (bottom) CD4$^+$ samples. Each plot represents the expression profile of differentially expressed genes identified per sample. Colours represent scaled variance stabilised counts of multiple genes that map to a particular node. Black lines represent hierarchical agglomerative clusters applied ontop of the map.

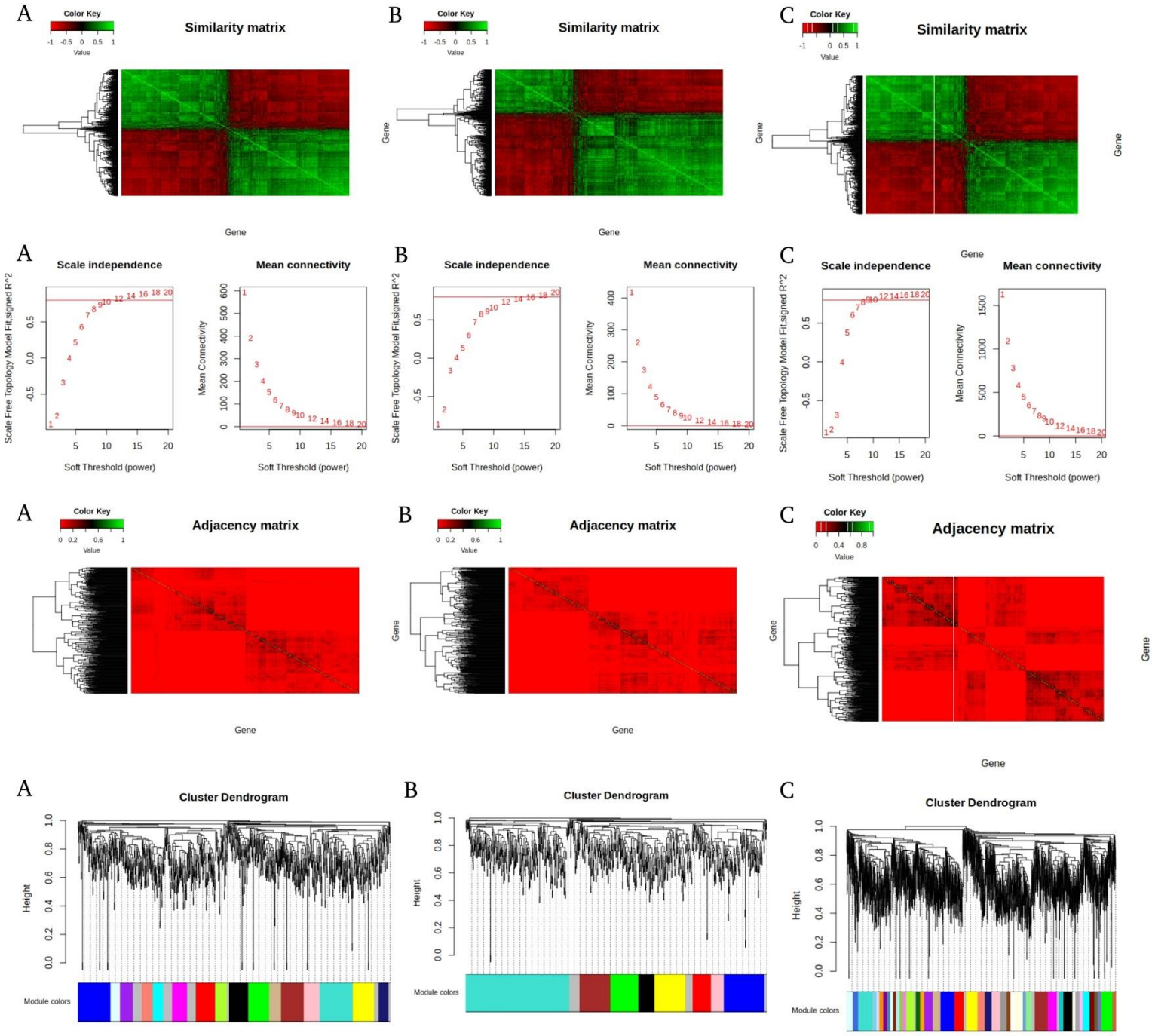


Figure 3.13: Weighted Gene Network Analysis showing (from top) heatmaps of similarity matrices, scale-free topology, mean connectivity, heatmaps and adjacency matrices, and cluster dendrograms for CD4$^+$ T cell subset comparisons. For each comparison an initial correlation similarity matrix is calculated, then scale free independence plots allow for determination of the power to raise each correlation value to. A resulting adjacency matrix is generated which is then clustered to form individual modules of genes. Labels refer to each comparison (A): T_{CM} vs T_{EM}, (B): T_{CM} vs T_N and (C): T_{EM} vs T_N.

3.10 The CD4$^+$ T_N vs T_{EM} Comparison

Further inspection of the CD4$^+$ T_N vs T_{EM} comparison identified three differentially co-expressed modules that showed potential to reflect differences in HIV-1 control mechanisms active across cell subsets. These were the Royalblue (Figure 3.14), Skyblue3 (Figure 3.19) and Salmon (Figure 3.21) modules. ClusterProfiler (Yu et al. 2012) identified GO terms that were significantly enriched in these modules. These were used to discern differences in the cell types.

3.10.1 Differences in Viral Transcription across CD4$^+$ T_{EM} and T_N Cell Subsets

The RoyalBlue WGCNA module identified GO terms involved in viral transcription and viral processes that were differentially co-expressed across T_{EM} and T_N cells. One notable change downregulated in T_N cells compared to T_{EM} cells were the Ribosomal proteins RPS25, RPSL24 and RPS27 (Figure 3.14; Figure S8). These genes were identified as being expressed at a higher level in T_{EM} than T_N cells. A GO Dot plot (Figure 3.15) demonstrated the diversity of GO enriched terms for the module (Figure 3.15). A CNET plot (Figure 3.16) demonstrated the involvement of these proteins in viral process and viral transcriptional GO terms (Figure 3.16).

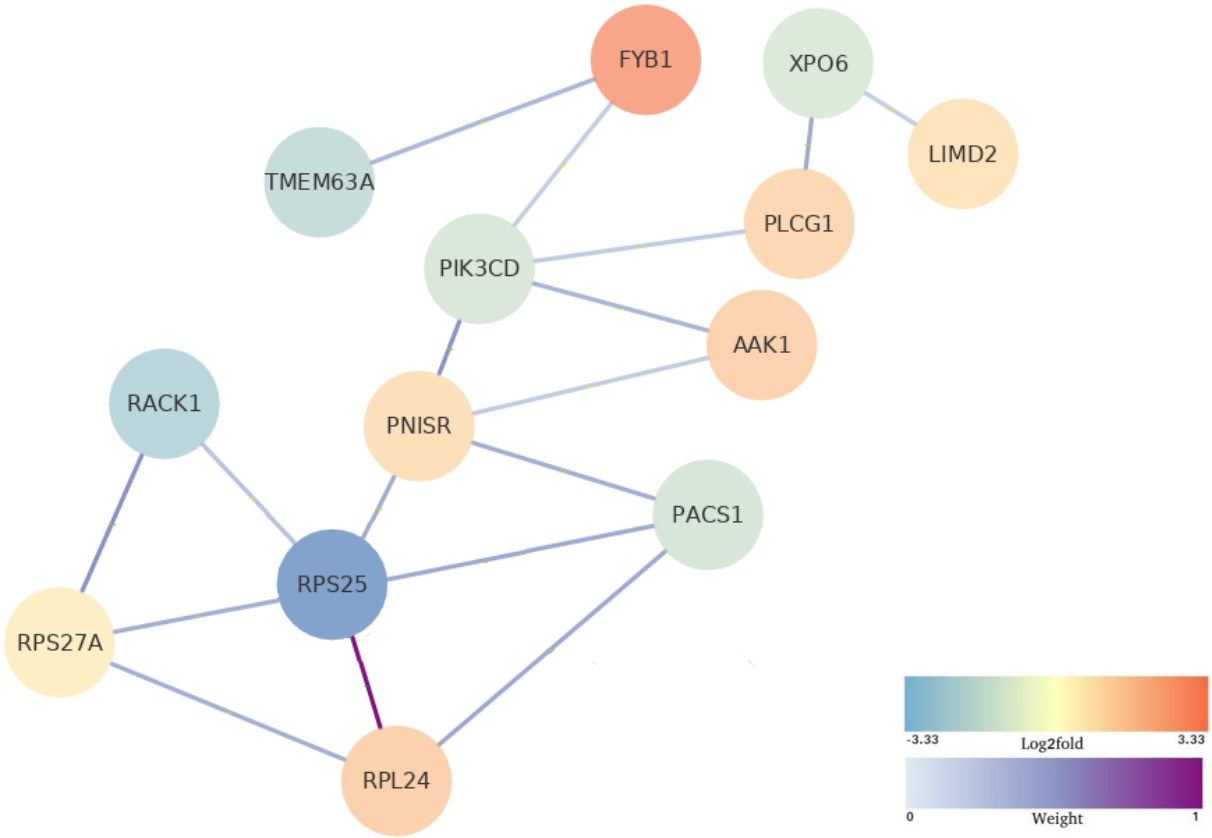


Figure 3.14: Weighted gene co-expression network for the WGCNA RoyalBlue module produced from the T_N vs T_{EM} CD4$^+$ T cell subset comparison. Nodes are coloured by

DESeq2 log_2 fold change scores, where positive values indicate up-regulation in T_N vs T_{EM}, and negative values indicate down-regulation in T_N vs T_{EM}. Edges were coloured by WGCNA weight values, representing the strength of co-expression between genes. One notable gene RPS25 shows greater expression in T_{EM} than T_N with strong co-expression relationship to RPL24.

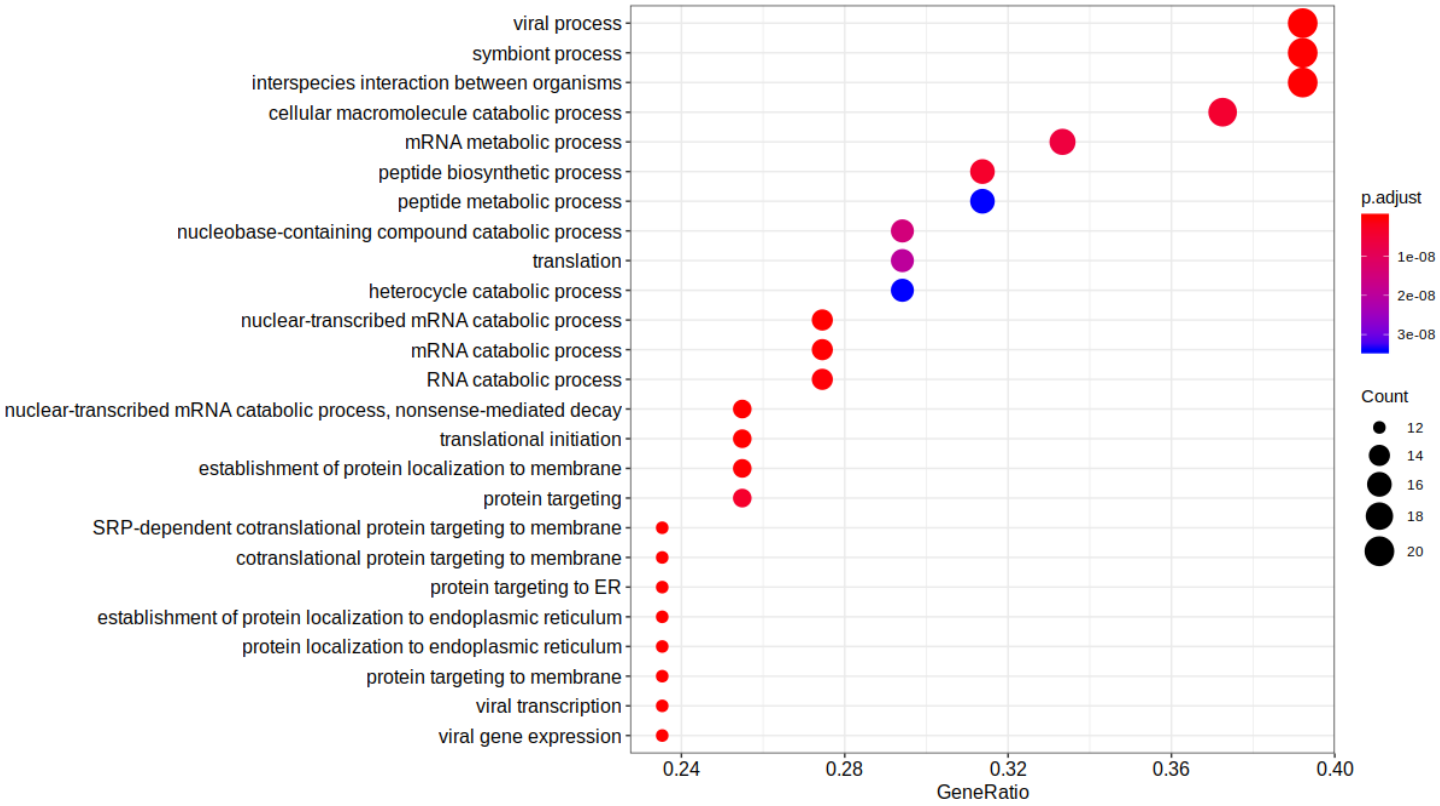


Figure 3.15: GO enrichment dot plot for the RoyalBlue module for the $CD4^+$ T_N vs T_{EM} comparison, showing counts of genes found in different significantly enriched GO terms. Adjusted p-values are coloured and counts are represented by the size of each circle. Results showed the diversity of viral related processes including symbiont, translation, RNA and viral transcriptional GO terms.

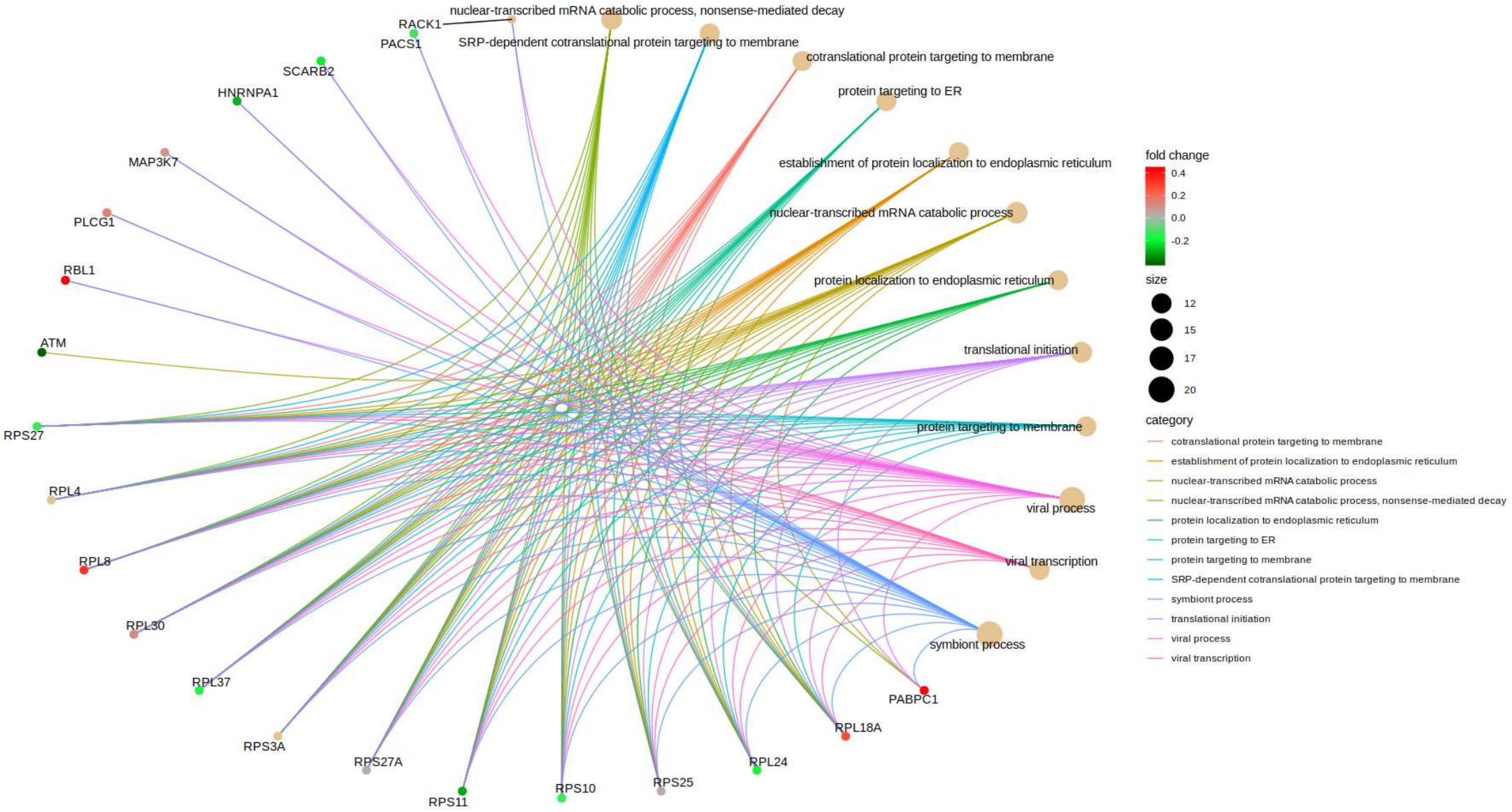


Figure 3.16: RoyalBlue module CNET plot showing the genes involved in GO enrichment pathways differentially expressed between CD4$^+$ T_N and T_{EM} subsets. Coloured nodes show DESeq2 log2 fold change values from DESeq2 with red indicating positive log fold changes and green indicating negative log fold changes (Love, 2014) differential expression analysis. Positive log2 fold changes represent upregulated sets in T_N cells whilst negative values represent downregulation in T_N vs T_{EM}.

3.11 Differences in Immune Cell Differentiation Mechanisms across CD4$^+$ T_{EM} and T_N Cell Subsets

Within the Skyblue3 module differentially expressed GO terms relating to immune cell activation were found. These included five genes mapping to myeloid leukocyte differentiation, and three genes involved in macrophage differentiation (Figure 3.17). Receptor-interacting serine/threonine-protein kinase 1 (RIPK1) and Recombination Signal Binding Protein For Immunoglobulin Kappa J Region (RBPJ) were two genes involved in myeloid leukocyte differentiation that were downregulated in T_N vs T_{EM} cells (Figure 3.19). RIPK1 is also involved in macrophage differentiation and was investigated further (Figure 3.18). Hub gene analysis (Figure S7) indicated that RIPK1 and RBPJ were connected to the central gene SNX5. This was in turn identified to be co-expressed with MAPK genes, MAP4 and MAPK1 in the light green modules (Figure 19B).

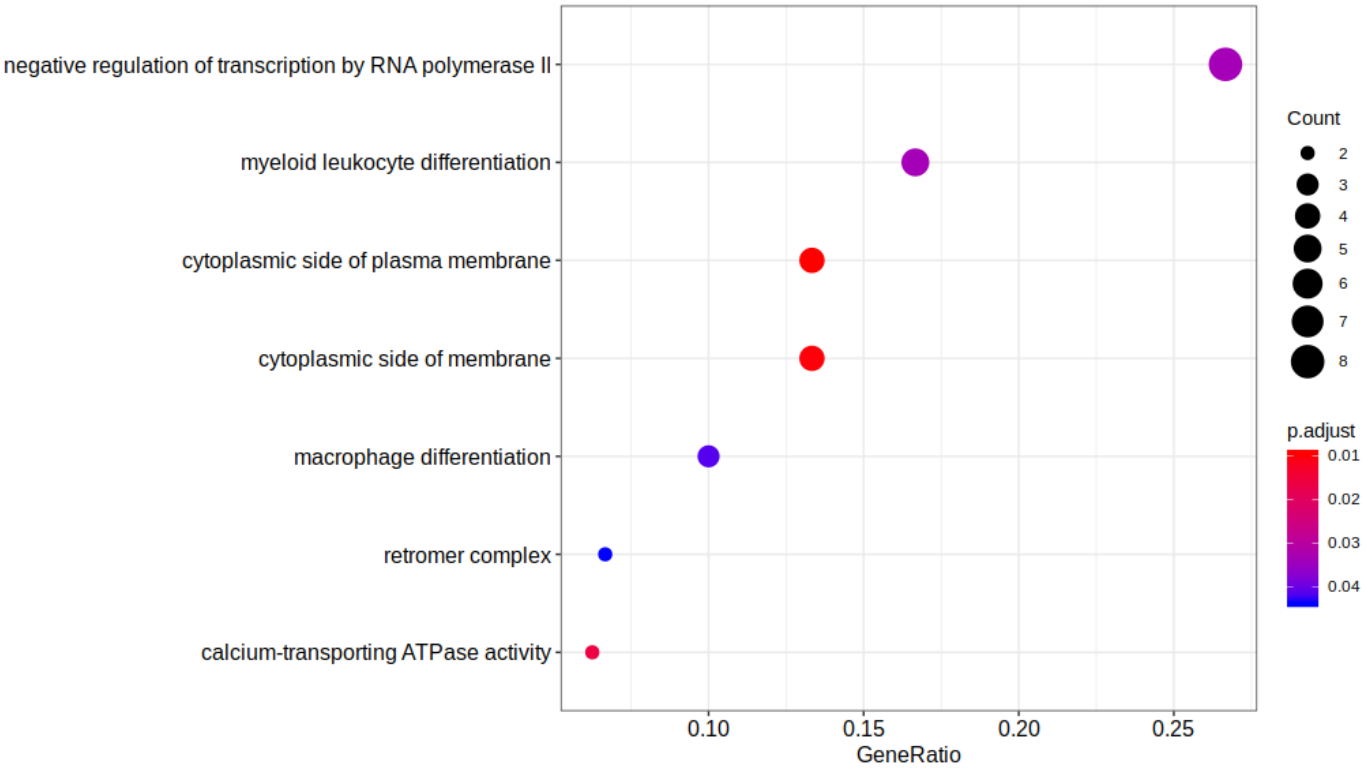


Figure 3.17: GO enrichment dot plot for the SkyBlue3 module showing counts of genes found in significantly enriched sets. Adjusted p-values are coloured and counts represented by the size of each circle.

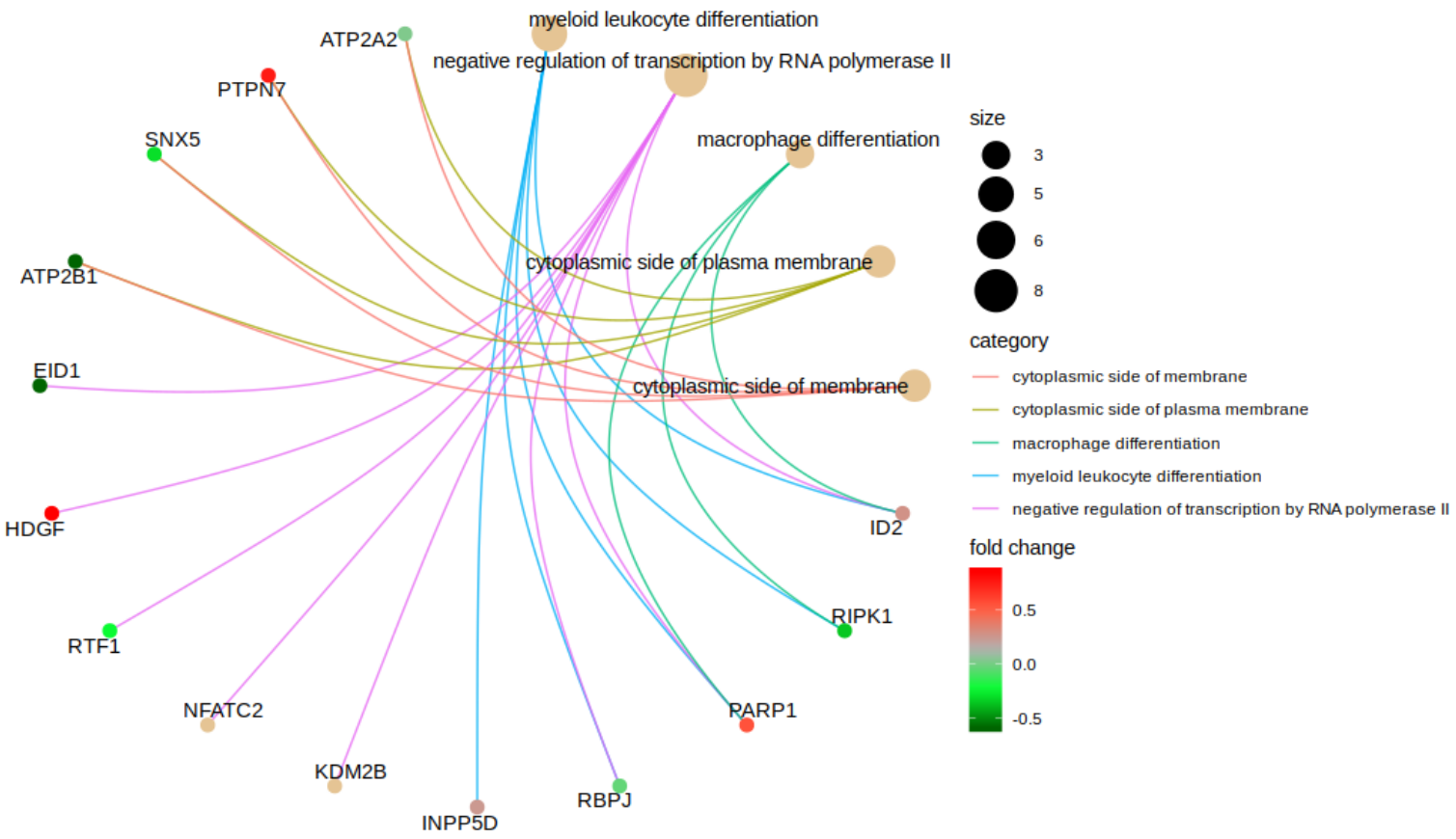


Figure 3.18: CNET plot showing the genes within the Skyblue3 module involved in a given enrichment pathway. Coloured nodes show DESeq2 $\log_2$ fold change values between $CD4^+$ T_N and T_{EM} cells, with positive $\log_2$ fold changes representing upregulated sets in T_N cells.

A

B

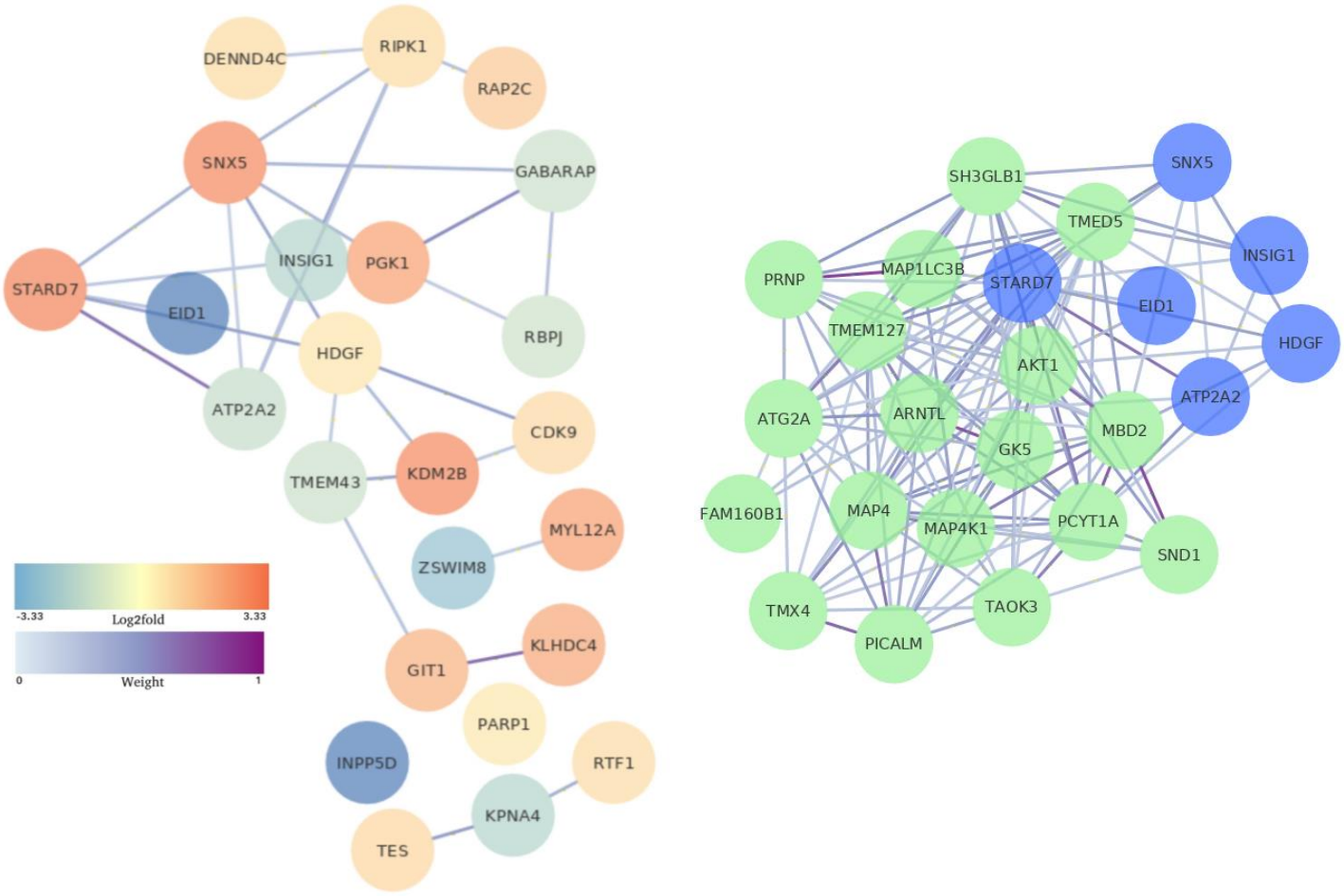


Figure 3.19: Weighted gene co-expression network from genes within the SkyBlue3 and light green modules from a T_N vs T_{EM} CD4$^+$ T cell comparison. (A): Nodes are coloured by DESeq2 $\log_2$ fold change scores with edges coloured by WGCNA weight values representing the strength of co-expression between genes. (B): Overlap of the Skyblue3 (blue) lightgreen (green) module showing high connectivity. SNX5 links to a group of MAP Kinase Genes MAP4, AKT1 and MAP4K of the two modules edges are coloured by weight values.

3.12 Differences In Immune Activation Levels Across CD4$^+$ T_{EM} and T_N Cell Subsets

Genes associated with immune activation were found to be differentially expressed across the T_N vs T_{EM} comparison. GO terms including immune response, leukocyte activation and phagocytosis were overrepresented (Figure 3.20). Within the immune response GO term, 29 genes were differentially co-expressed (Figure 3.20; Figure 3.21). These included the Cell Division Control Protein 42 (CDC42), Protein-disulfide isomerase A3 (PDIA3) and Insulin-like growth factor 2 Receptor (IGF2R). CNET analysis showed all three mapped to the immune response GO term (Figure 3.22).

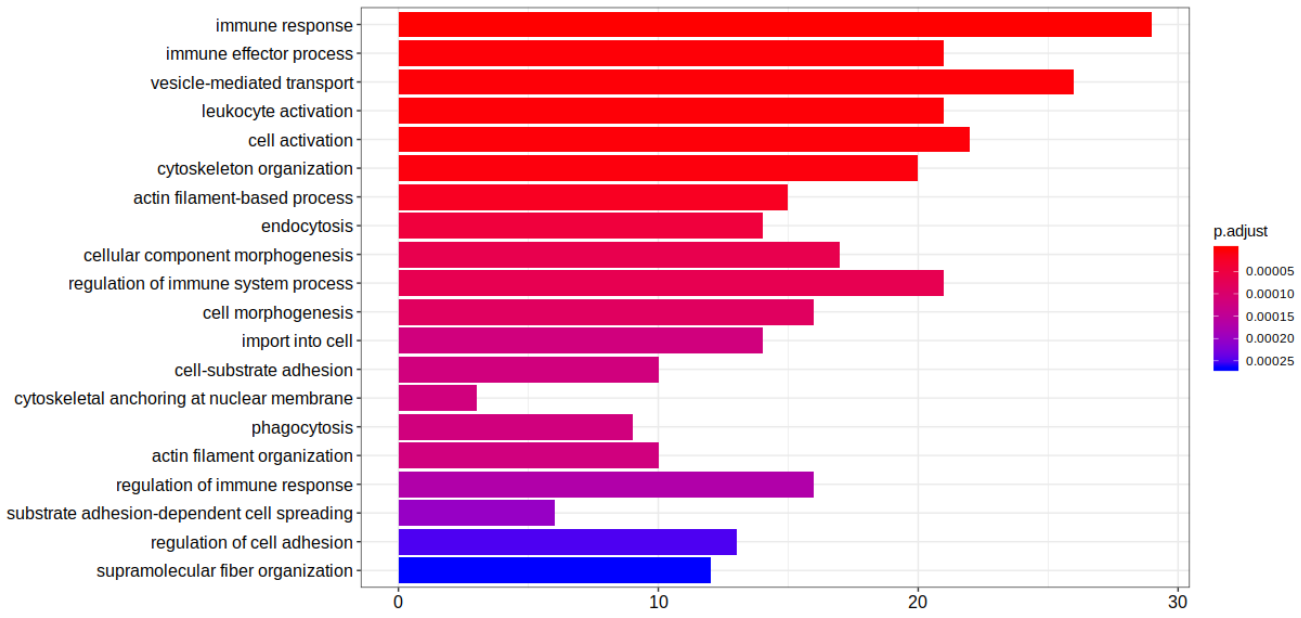


Figure 3.20: GO enrichment bar plot for the Salmon module showing counts of genes (x-axis) found in significantly enriched sets. Adjusted p-values are coloured.

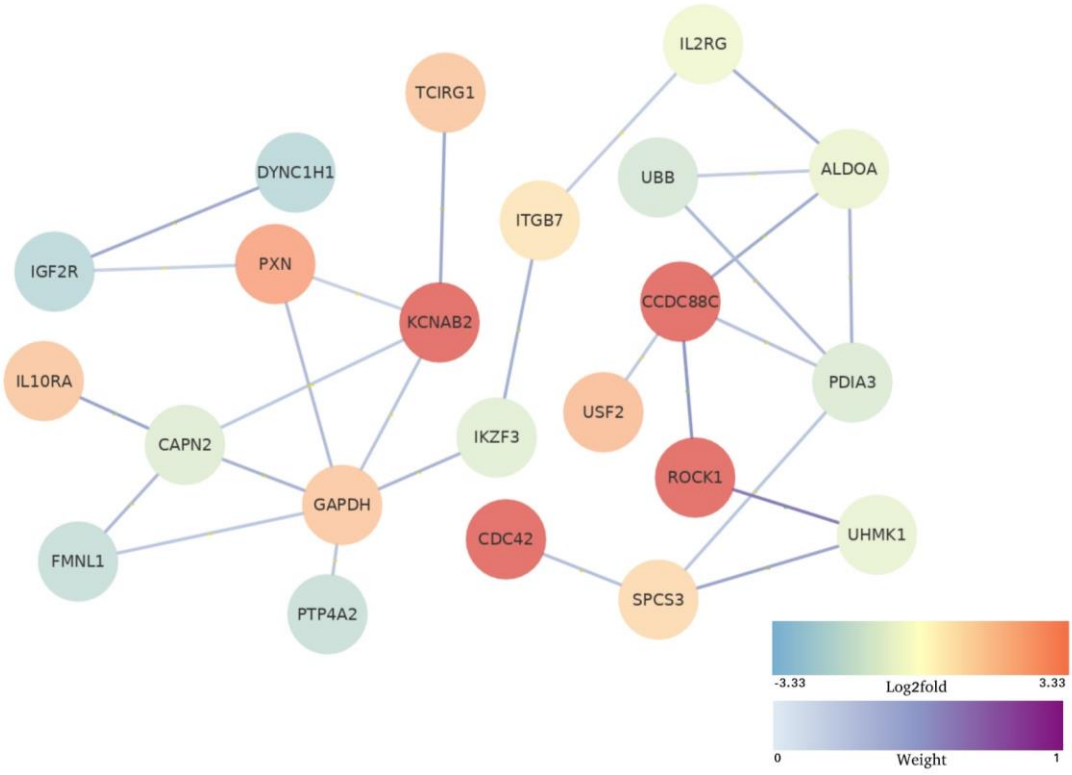


Figure 3.21: Weighted gene co-expression network for the WGCNA Salmon module produced from the T_N vs T_{EM} $CD4^+$ T cell subset comparison. Networks produced in Cytoscape (Shannon et al. 2003) with nodes are coloured by DESeq2 (Love et al. 2014) $\log_2$ fold change scores, where positive values indicate up-regulation in T_N vs T_{EM}, and negative values indicate down-regulation in T_N vs T_{EM}. Edges were coloured by WGCNA (Langfelder and Horvath 2008) weight values, representing the strength of co-expression between genes.

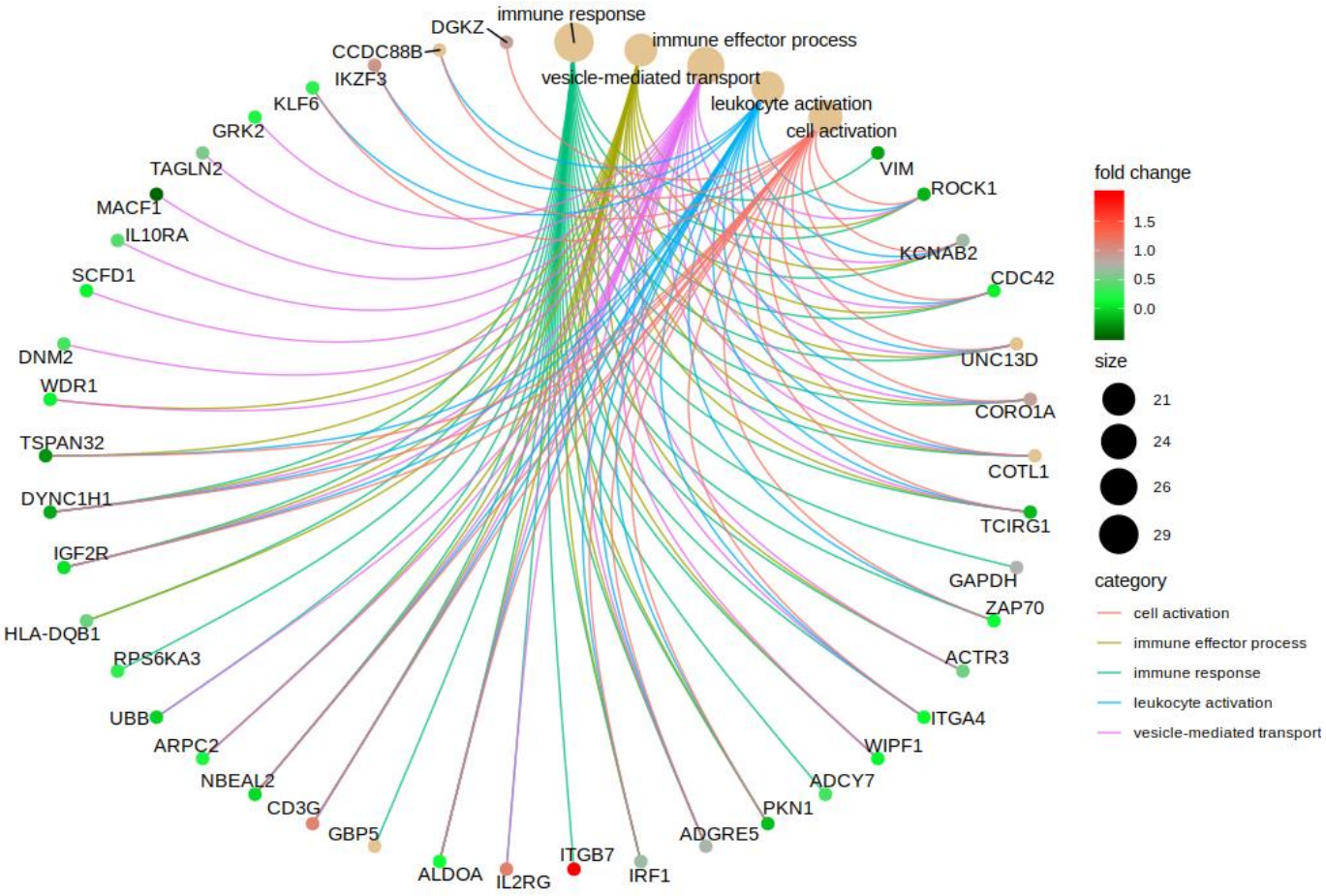


Figure 3.22: CNET plot showing the genes within the Salmon module involved in their respective enrichment pathways. Coloured nodes show DESeq2 log_2 fold change values between T_N and T_{EM} $CD4^+$ cells. Positive log_2 fold changes representing upregulated sets in T_N cells. Negative log_2 fold changes representing downregulated gene sets in T_N cells. Plot highlights the genes involved in immune response and activation GO terms including CDC42, PDIA3 and IGF2R.

4 Discussion

The recent finding of a second patient cleared of HIV-1 in March 2020 has again highlighted that a model for a cure exists (Gupta et al., 2020; Hütter et al., 2009). It is likely that this model is a functional one in which trace levels of HIV-1 persist despite viral replication being eliminated (Gupta et al., 2020; Hütter et al., 2009). HIV-1 elite controllers (ECs) maintain stringent control of the virus, in the absence of antiretroviral treatment (ART), for over 10 years, and thus represent an effective model for a functional cure. This study transcriptionally profiled Naive (T_N), Effector Memory (T_{EM}), and Central Memory (T_{CM}) $CD4^+$ T cell subsets within the context of HIV-1 elite control (EC). It compared different normalisation and co-expression methodologies to profile $CD4^+$ T cell subsets and identified transcriptional differences between these subsets that may reflect nuances in latent infection. Here, the quality processing, method comparison, broad transcriptional differences and findings between T_{EM} and T_N cells are discussed.

4.1 Quality Control and Sample Normalisation Was Effective

Quality reporting using FastQC (Andrews, 2018) and improvement with Cutadapt (Martin, 2011) was effective. Where data quality was irretrievable, samples were removed allowing for only the highest quality data to be used for further analyses. Kallisto alignment was consistent across different individuals and cell types, indicating a successful and unbiased alignment step. Batch effect analysis revealed that no such effects were apparent, and samples did not need further processing to remove them. Outlier detection identified two samples that clustered together, separate from the remainder of the cohort. It is likely the reasons for their differences were as a result of biological variation as there were no apparent differences in sample collection date, CD4 count, HIV-1 viral loads or duration of infection. Outlier samples were removed to reduce intra individual biases that would affect co-expression and differential analysis. This included the other samples from these individuals to ensure only complete datasets were studied. In summary, stringent quality screening allowed for confidence that all processing and analysis would reflect the biological differences in the cell subsets and would not be as a result of technical errors.

Normalisation was necessary to ensure the removal of differences in library size and composition. Size-factor based normalisation effectively removed biases for both differential and co-expression analysis. Following this step, normalisation with two different steps, namely

Variance Stabilising Transformation (VST) and Regularised Logarithm (rlog) normalisation. These both successfully reduced heteroscedasticity in the data. VST and rlog both effectively reduced variance, allowing for consistency in dispersion across different mean values. This meant that differences in the levels of gene expression would not bias the likelihood of identifying genes as co-expressed. When mean standard deviation plots and distributions were analysed, VST was shown to reduce slightly more variance than rlog did. On this basis, it was selected for use in further co-expression network analysis. DESeq2 Differential Expression (DE) analysis successfully used shrinkage dispersion to solve the problem of heteroscedasticity, resulting in consistency of dispersion estimates across mean values. Overall, the normalisation procedures employed reduced unwanted noise and variance, improving the quality of co-expression and DE analyses.

For both co-expression and differential expression analysis, quality reporting during experimentation checked the effectiveness of all experimentation. DESeq2 MA and mean dispersion plots showed no biases for upregulated and downregulated gene sets, indicating an effective experiment. Weighted Gene Co-expression Network Analysis (WGCNA) dendrogram visualisation showed distinct bands of colours, suggesting that the groupings of genes into modules was effective. This together with the similarity heatmaps, adjacency heat maps and careful selection of mean connectivity and scale-free topology scores indicated that WGCNA results were of a high quality. The sum of square plots ensured that the right number of clusters was chosen for k-means clustering, and k-means visualisation showed distinct bands of clusters, highlighting effective clustering. Self-organising maps (SOM) reached a plateau when analysing mean distance to closest units, indicating that an optimum distance between neurons and genes was reached. Therefore DESeq2, WGCNA, SOM and k-means quality evaluation showed all methods had been run effectively.

4.2 WGCNA was the Most Effective Method for Identifying Co-expressed Genes

Using gene ratio scores identified with GO Enrichment, the efficacy of SOM, WGCNA and k-means as methodological approaches for identifying co-expressed genes was evaluated. Differentially expressed genes had been clustered with the different methods and then underwent GO enrichment. Gene ratios from each enriched cluster established the effectiveness of methodologies to identify co-expressed genes. Findings showed that WGCNA had the highest mean gene ratio, followed by k-means and then SOM which had similar ratios. On this basis, WGCNA was shown to be the best co-expression approach for the study. This may be

as a result of the wide use of WGCNA for most co-expression analyses (Botía et al., 2017), meaning many GO terms have been established using this method - causing a bias towards WGCNA. Furthermore, WGCNA seemed to break clusters up into smaller modules than k-means clustering and SOM. Identification of smaller clusters may have resulted in more specific gene sets causing higher ratio scores in WGCNA.

It is, however, also likely that WGCNA may still be the best and most accurate methodology, as it combines many effective techniques together. These include Pearson correlation coefficients, topological overlap similarity calculations, parameter adjustment to optimise scale-free topology (a measure of the accuracy of a biological network) and clustering (HAC) (Langfelder and Horvath, 2008). Both k-means and SOM provided only clustering of gene expression values. Perhaps combined, their use will lead to greater efficacy. An existing study had shown when k-means was implemented within WGCNA, it improved the biological features identified (Botía et al., 2017). Perhaps using SOM as the clustering step of WGCNA may strengthen results. Further improvements of these methods are needed however, and on the basis of the GO ratios, WGCNA was selected for further co-expression analysis.

SOM heatmaps offered a useful visualisation of the complete transcriptional profile of all samples and showed some similarities with PCA and DESeq2 findings (although differences were likely caused by normalisation technique differences). SOM like PCA, dendrogram, and differential expression analysis, identified the T_{EM} vs T_N comparison to have the greatest differences. This has highlighted it as a useful method for successfully visualising the genome-wide transcriptional differences of a particular cell subset. Further integration of SOM with other techniques could likely yield better co-expression methodologies.

4.3 Broad Transcriptional Differences Distinguish $CD4^+$ T Cell Subsets

Upon analysis of sample heatmaps, sample dendrograms, and SOM heatmaps, it was clear that cell subsets clustered more closely than different cells within the same individual. This indicated that there were core transcriptional similarities within cell subsets that superseded the differences that occur between individuals. This was the case for all cell subsets. Analysis of Pearson correlation coefficients for the samples showed that there was also a very high degree of correlation between subsets. This was expected as all three $CD4^+$ T cell subsets are known to be phenotypically very similar. The degree of the similarities and differences between subsets could then be examined.

When analysed, SOM heatmaps across all 24 samples showed that T_{CM} and T_{EM} subsets seemed to have the greatest transcriptional similarities. This was not, however, supported by PCA, differential expression analysis and the raw count heatmap dendrogram. DE analysis also did not support SOM and VST heatmap results, with T_{CM} vs T_N having the fewest differentially expressed genes, followed by T_{EM} vs T_{CM} and vs T_{EM} and T_N. All methods showed T_{EM} and T_N had the greatest number of differences. Discrepancies in which cell subsets were the most similar are likely a result of differences in the DE and co-expression normalisation procedures. Both methods employed a size-factor based normalisation, however further processing for DESeq2 was using DESeq2's own model (Love et al., 2014), while VST was used for co-expression analysis. It is likely that T_{CM} vs T_N and T_{EM} vs T_{CM} are both very similar and hence cluster closely together. Normalisation procedure likely biases the similarities slightly in either direction causing the discrepancy.

T_{CM} and T_N subsets are known to have some similarity in function. Therefore, the findings suggesting their transcriptional similarity are explainable. T_{CM} and T_N both express CCR7 and CD62L, which enable them to manoeuvre through secondary lymphoid organs and High Endothelial Venules (HEV). They are similar in that they both have little to no effector function, and both can differentiate to produce effector subsets (Federica Sallusto et al., 2004). The manner in which they undergo differentiation is however, different - T_{CM} cells maintain memory function while differentiated whilst T_N cells do not. Unlike T_N, T_{CM} produce high amounts of IL2, which facilitates further T cell differentiation. T_{EM} are CCR7 negative and have chemokine receptors specific to homing towards inflamed tissues. T_{EM} cells are also the most differentiated subsets and it makes sense that T_{EM} and T_N would have a higher number of differences than T_{EM} vs T_{CM} and T_{CM} vs T_N. Findings supporting these differences in subsets would therefore make sense, as they reflect the underlying functional differences of these cell subsets.

The challenge for this study was attempting to establish whether subset differences were heightened as a result of HIV-1 infection. One notable observation from the study that produced the dataset used in this project (Boritz et al., 2016) was that T_{EM} cells were found to have the highest levels of HIV-1. In contrast, T_N cells had the lowest levels of HIV-1. This had potential to explain the high number of differences between these cell subsets. Memory $CD4^+$ T cell subsets are known to harbour latent infected virus and a high copy number of HIV-1 DNA is expected (Zerbato et al., 2016). T_{CM} cells also had relatively high levels of HIV-1 DNA, possibly explaining some of the similarity between T_{CM} and T_{EM} expression profile. While this

study cannot conclusively determine whether differences are induced by elite control of HIV-1, it seems likely that high levels of HIV-1 DNA in T_{EM} cells could affect their expression profile. The comparison of T_{EM} cells to T_N cells could therefore elucidate differences that occur as a result of high HIV-1 viral load. On this basis, the T_{EM} vs T_N comparison was selected for further study using WGCNA (Langfelder and Horvath 2008).

4.4 Differences Between T_{EM} and T_N Subsets May Reflect HIV-1 Elite Control

During chronic latent infection, HIV-1 is known to reside in reservoirs of $CD4^+$ T memory cells (Bosque et al., 2011; Kulpa et al., 2019; Lee and Lichterfeld, 2017). T_{EM} cells were known to have the highest levels of HIV-1 DNA of all subsets in this study (Boritz et al., 2016). T_{EM} cells maintain effector function, and it was likely that these subsets had a degree of immune control over the virus. When comparing T_{EM} to cells with the lowest levels of HIV-1 DNA (T_N; (Boritz et al., 2016), it was observed that these cells had the potential to reveal mechanisms that contribute towards controlling viral replication. T_N cells could therefore be used as a baseline control group for how lesser infected subsets function. Findings showed differences in viral transcription, immune differentiation and immune activation mechanisms.

4.5 Greater Levels of Viral Transcription in T_{EM} Compared to T_N $CD4^+$ Subsets

GO terms associated with viral transcription were found to be over-represented within genes identified as co-expressed across T_{EM} and T_N $CD4^+$ cell subsets. Further analysis revealed that several ribosomal proteins were expressed at a greater level in T_{EM} subsets than T_N. One notable example was ribosomal protein S25 (RPS25), which was expressed at a much higher level in T_{EM} than T_N cells, and was co-expressed with other ribosomal proteins including RPS27A and RPSL24. The HIV-1 internal ribosome entry site (IRES), which drives important Gag HIV-1 viral protein replication, is dependent on RPS25 and likely other ribosomal proteins (Carvajal et al., 2016). IRES causes recruitment of the 40S ribosomal unit which can initiate HIV-1 transcription. High levels of RPS25 are therefore associated with high levels of viral replication. This supports the findings that T_{EM} had higher levels of HIV-1 than T_N cells (Boritz et al., 2016). Furthermore, it indicates that within the context of T_{EM} cells, viral replication is likely upregulated, causing increased transcription of HIV-1. Such HIV-1 replication may allow for clearance of the virus from latent reservoirs by the immune system or allow for its continuous immune suppression. Latent reservoirs of HIV-1 in ECs are known to be much smaller than in progressor patients (Blankson et al., 2007), and there is evidence that stringent cytotoxic T lymphocyte (CTL) responses prevent the seeding of latent reservoirs (Migueles

and Connors, 2015). Perhaps stimulation of HIV-1 replication enables continued immune response to the virus and prevents large viral reservoirs developing in ECs.

4.5.1 RIPK1 and RBPJ Facilitate Cellular Differentiation and Viral Replication

GO terms associated with myeloid leukocyte and macrophage differentiation were overrepresented in genes differentially expressed across the $CD4^+$ T_N vs T_{EM} cell subsets. Upon further investigation, RIPK1 (Receptor Interacting Protein Kinase 1), which is associated with both GO terms, was found to be expressed at a greater level in T_{EM} cells than T_N (downregulated in T_N vs T_{EM}). RIPK1 is a key regulator of innate immune signalling pathways (Lalaoui et al., 2020), and its up-regulation indicates activation and differentiation of immune pathways within these EC patients.

Interestingly, RIPK1 is usually cleaved by HIV-1 protease during chronic infection (Wagner et al., 2015). Such cleavage is thought to counter host defence mechanisms against HIV-1, as it inhibits the RIPK1/RIPK3 complex and resulting Nuclear Factor kappa light chain enhancer (NF-kB) transcription factor activation. NF-kB includes a family of transcription factors that are involved in leukocyte and macrophage differentiation (Wagner et al., 2015). It was notable that this gene was downregulated within T_{EM} (when compared to a baseline of T_N cells), which had the greatest levels of HIV-1 (Boritz et al., 2016). This suggests that perhaps EC patients maintain a mechanism that prevents such cleavage allowing for continued immune cell differentiation and viral suppression.

RBPJ was also expressed at a greater level in T_{EM} vs T_N $CD4^+$ subsets. RBPJ is known to be involved in ERK MAP Kinase signalling, facilitating the activation and differentiation of immune cell subsets. This was supported by RBPJ mapping to GO enriched pathways for macrophage differentiation and myeloid leukocyte differentiation within the Skyblue3 module. MAPK, ERK1 and ERK2 facilitated pathways are known to link cytokine signals to cause activation of HIV-1 through stimulation of AP-1 and NF-κB (Yang et al., 1999). These in turn cause reactivation of latently infected virus through the HIV-1 long terminal repeat (LTR). Up-regulation of RBPJ causes greater NF-κB and associated immune cell differentiation. This may also cause expression of the virus in T_{EM} cells, reducing the latent reservoir and promoting HIV-1 control.

The above findings are interesting as it means potential activation of the virus can also be associated with activation of immune differentiation mechanisms. Perhaps ECs have greater

levels of leukocyte differentiation that can then target replication virus to reduce viral reservoirs. Furthermore, RIPK1 and RBPJ were found to be co-expressed with SNX5 (Sorting nexin-5) a central gene in the Skyblue3 module. SNX5 was found to be likely co-regulated or co-expressed with members of the MAP kinase signalling pathway, MAP4, MAPK1 and AKT1. Therefore, it seems that T_{EM} subsets support the activation of MAPK pathways which promote many cell processes including leukocyte and macrophage differentiation.

4.5.2 Differences in Immune Response Mechanisms Across $CD4^+$ T_{EM} and T_N Subsets

Immune response, immune effector process, and phagocytosis GO terms were all overrepresented in genes identified as differentially expressed between the T_{EM} and T_N subsets. Within the Salmon module, 29 genes were identified as co-expressed that were involved in immune response. This supports significant levels of immune functional differences between T_{EM} and T_N $CD4^+$ subsets. A notable difference between the two subsets was that T_N cells had high levels of CDC42, which has been shown to facilitate HIV-1 cell-cell transmission in dendritic cells (Nikolic et al., 2011). CDC42 likely supports this process in other cell subsets. Downregulation of this gene within T_{EM} cells may slow HIV-1 transmission to other $CD4^+$ memory subsets, reducing viral seeding in other subsets. Other genes known to facilitate HIV-1 replication were found to be upregulated in T_{EM} subsets, notably Protein-disulfide isomerase A3 (PDIA3) and Insulin-like growth factor 2 Receptor (IGF2R). PDIA3 has been shown to play a predominant role in HIV-1 viral entry allowing for its replication. IGF2R is an IFNgamma inducible protein that is also known to facilitate intracellular HIV-1 replication (Suh et al., 2010). These support findings of increased HIV-1 replication within $CD4^+$ T_{EM} cell types.

4.6 Conclusions

This study examined the transcriptional profile of $CD4^+$ T cell subsets within a context of HIV-1 elite control. It investigated the differences amongst cell subsets and identified differences that potentially allow for HIV-1 suppression in ECs. Broad transcriptional findings showed that T_{CM} and T_N cells and T_{EM} and T_{CM} subsets had a greater number of similarities than T_{EM} and T_N subsets. This was attributable to similar functioning of T_{CM} and T_{EM} as memory subsets and of T_{CM} and T_N as cells able to navigate HEV and undergo further differentiation. The T_{EM} vs T_N comparison had the greatest number of transcriptional differences, with 2760 differentially expressed genes that was supported by SOM, PCA and sample dendrogram analysis. T_{EM} subsets had the greatest level of viral DNA and it was likely that EC mechanisms

allowed for suppression of viruses in these subsets. GO terms associated with viral transcription, immune differentiation and immune activation were found to be overrepresented across the subsets. This included differences in groups of ribosomal proteins known to be involved in HIV-1 viral replication; proteins RIPK1 and RPBJ involved in differentiation of and activation of immune cells and immune activation genes CDC42, PDIA3 and IGF2R which facilitate virus replication. This highlighted that EC T_{EM} cells underwent high amounts of viral replication which could have the potential to reduce residual latency. Activation of the immune response and differentiation of immune cells may cause a response to the virus that suppresses whilst preventing build-up of latent reservoirs. In summary, the transcriptional profile of CD4+ T cell subsets showed differences in their mechanisms which likely reflect the functional differences in subsets and differences in HIV-1 infection levels.

4.7 Limitations

One of the main limitations of this study was that ECs represent a heterogeneous set of individuals, which are likely to have multiple different mechanisms to suppress HIV-1. After extensive quality processing, a final cohort of eight individuals was selected from a total of 14 for study. With a smaller sample size, diverse mechanisms of EC were likely to be lost. A smaller set of individuals is unlikely to represent the majority of differences in $CD4^+$ T cell subsets within ECs. Quality processing did however ensure that the results from these individuals were of the highest quality.

A further challenge in this study was the inability to determine whether the transcriptional differences observed reflected EC mechanisms, or where a consequence of differing levels of viral replication between subsets. Transcriptional differences may have been induced by baseline HIV-1 infection or normal functional cell type differences. To investigate these, further crossover work with datasets derived from patients on ART, non-controller HIV-1 infected individuals and uninfected patients is needed. This would help establish to what extent mechanisms are induced by an elite control, HIV-1 or functional context.

4.8 Future Work

Advances in sequencing and deep machine learning models offer opportunities in further study of ECs. New artificial neural networks that use more extensive layering of neurons, known as deep learning models can outperform most clustering techniques. Deep learning models have already shown promise in co-expression analysis (Ahn et al., 2019; Svensson et al., 2020). One

such technique is autoencoders, which have recently been used in a co-expression context (Ahn et al., 2019; Svensson et al., 2020). The use of more advanced machine learning models to identify causal relationships between genes will be beneficial in understanding gene sets from individuals such as HIV-1 ECs.

Furthermore, progress in gating and sequencing technologies now allow for single-cell RNA sequencing. That is when the transcriptome of individual cells can be defined. This gives a much higher resolution picture of the functioning of cell subsets and could be applied to EC data for further progress in identifying the EC phenotype. To summarise, future work would ideally incorporate the latest techniques of single cell RNA sequencing with advanced deep learning models on a large sample subset of ECs in sub-Saharan Africa. This could include a study with patients who progress to AIDS, normal uninfected individuals and ECs.

5 Bibliography

Agosto, L. M., & Henderson, A. J. (2018). CD4+ T cell subsets and pathways to HIV latency. *AIDS Research and Human Retroviruses*, *34*(9), 780–789.

Ahn, H., Son, S., & Kim, S. (2019). Deepfunnet: deep learning for gene functional similarity network construction. *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 1–8). Presented at the 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), IEEE.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97.

Amar, D., Safer, H., & Shamir, R. (2013). Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Computational Biology*, *9*(3), e1002955.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106.

Andrews, S. (2018). *FastQC: a quality control tool for high throughput sequence data.* . Computer software, Babraham Bioinformatics. Retrieved April 9, 2020, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Autran, B., Descours, B., Avettand-Fenoel, V., & Rouzioux, C. (2011). Elite controllers as a model of functional cure. *Current Opinion in HIV and AIDS*, *6*(3), 181–187.

Bailey, J. R., O'Connell, K., Yang, H.-C., Han, Y., Xu, J., Jilek, B., Williams, T. M., et al. (2008). Transmission of human immunodeficiency virus type 1 from a patient who developed AIDS to an elite suppressor. *Journal of Virology*, *82*(15), 7395–7410.

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews. Genetics*, *5*(2), 101–113.

Baron, B. W., Desai, M., Baber, L. J., Paras, L., Zhang, Q., Sadhu, A., Duguay, S., et al. (1997). BCL6 can repress transcription from the human immunodeficiency virus type I promoter/enhancer region. *Genes, Chromosomes & Cancer*, *19*(1), 14–21.

Benjamini, Y., Heller, R., & Yekutieli, D. (2009). Selective inference in complex research. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering*

Sciences, *367*(1906), 4255–4271.

Biancotto, A., Iglehart, S. J., Vanpouille, C., Condack, C. E., Lisco, A., Ruecker, E., Hirsch, I., et al. (2008). HIV-1 induced activation of CD4+ T cells creates new targets for HIV-1 infection in human lymphoid tissue ex vivo. *Blood*, *111*(2), 699–704.

Blankson, J. N. (2010). Effector mechanisms in HIV-1 infected elite controllers: highly active immune responses? *Antiviral Research*, *85*(1), 295–302.

Blankson, J. N., Bailey, J. R., Thayil, S., Yang, H.-C., Lassen, K., Lai, J., Gandhi, S. K., et al. (2007). Isolation and characterization of replication-competent human immunodeficiency virus type 1 from a subset of elite suppressors. *Journal of Virology*, *81*(5), 2508–2518.

Blumenthal, R., Durell, S., & Viard, M. (2012). HIV entry and envelope glycoprotein-mediated fusion. *The Journal of Biological Chemistry*, *287*(49), 40841–40849.

Bochud, P.-Y., Hersberger, M., Taffé, P., Bochud, M., Stein, C. M., Rodrigues, S. D., Calandra, T., et al. (2007). Polymorphisms in Toll-like receptor 9 influence the clinical course of HIV-1 infection. *AIDS*, *21*(4), 441–446.

Boritz, E. A., Darko, S., Swaszek, L., Wolf, G., Wells, D., Wu, X., Henry, A. R., et al. (2016). Multiple Origins of Virus Persistence during Natural Control of HIV Infection. *Cell*, *166*(4), 1004–1015.

Bosque, A., Famiglietti, M., Weyrich, A. S., Goulston, C., & Planelles, V. (2011). Homeostatic proliferation fails to efficiently reactivate HIV-1 latently infected central memory CD4+ T cells. *PLoS Pathogens*, *7*(10), e1002288.

Botía, J. A., Vandrovcova, J., Forabosco, P., Guelfi, S., D'Sa, K., United Kingdom Brain Expression Consortium, Hardy, J., et al. (2017). An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks. *BMC Systems Biology*, *11*(1), 47.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527.

Brettle, R. P., Gore, S. M., Bird, A. G., & McNeil, A. J. (1993). Clinical and epidemiological implications of the Centers for Disease Control/World Health Organization

reclassification of AIDS cases. *AIDS*, *7*(4), 531–539.

Brinkman, C. C., Peske, J. D., & Engelhard, V. H. (2013). Peripheral tissue homing receptor control of naïve, effector, and memory CD8 T cell localization in lymphoid and non-lymphoid tissues. *Frontiers in immunology*, *4*, 241.

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.

Burbelo, P. D., Bayat, A., Rhodes, C. S., Hoh, R., Martin, J. N., Fromentin, R., Chomont, N., et al. (2014). HIV antibody characterization as a method to quantify reservoir size during curative interventions. *The Journal of Infectious Diseases*, *209*(10), 1613–1617.

Campbell, J. J., Bowman, E. P., Murphy, K., Youngman, K. R., Siani, M. A., Thompson, D. A., Wu, L., et al. (1998). 6-C-kine (SLC), a lymphocyte adhesion-triggering chemokine expressed by high endothelium, is an agonist for the MIP-3beta receptor CCR7. *The Journal of Cell Biology*, *141*(4), 1053–1059.

Carvajal, F., Vallejos, M., Walters, B., Contreras, N., Hertz, M. I., Olivares, E., Cáceres, C. J., et al. (2016). Structural domains within the HIV-1 mRNA and the ribosomal protein S25 influence cap-independent translation initiation. *The FEBS Journal*, *283*(13), 2508–2527.

Chakrabarti, L. A., & Simon, V. (2010). Immune mechanisms of HIV control. *Current Opinion in Immunology*, *22*(4), 488–496.

Chang, C.-C., & Chen, S.-H. (2015). A comparative analysis on artificial neural network-based two-stage clustering. *Cogent Engineering*, *2*(1).

Chauhan, A., & Khandkar, M. (2015). Endocytosis of human immunodeficiency virus 1 (HIV-1) in astrocytes: a fiery path to its destination. *Microbial Pathogenesis*, *78*, 1–6.

Chomont, N., El-Far, M., Ancuta, P., Trautmann, L., Procopio, F. A., Yassine-Diab, B., Boucher, G., et al. (2009). HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nature Medicine*, *15*(8), 893–900.

Cohen, G. B., Gandhi, R. T., Davis, D. M., Mandelboim, O., Chen, B. K., Strominger, J. L., & Baltimore, D. (1999). The selective downregulation of class I major histocompatibility

complex proteins by HIV-1 protects HIV-infected cells from NK cells. *Immunity*, *10*(6), 661–671.

Colin, L., & Van Lint, C. (2009). Molecular control of HIV-1 postintegration latency: implications for the development of new therapeutic strategies. *Retrovirology*, *6*, 111.

Compeau, P., & Pevzner, P. (2014, May 19). *How Do We Assemble Genomes? (Part 10/12)*. Online Presenation presented at the Bioinformatics Algorithms: An Active Learning Approach.

Cooper, A., García, M., Petrovas, C., Yamamoto, T., Koup, R. A., & Nabel, G. J. (2013). HIV-1 causes CD4 cell death through DNA-dependent protein kinase during viral integration. *Nature*, *498*(7454), 376–379.

Dalmasso, C., Carpentier, W., Meyer, L., Rouzioux, C., Goujard, C., Chaix, M.-L., Lambotte, O., et al. (2008). Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS Genome-Wide Association 01 study. *Plos One*, *3*(12), e3907.

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Briefings in Bioinformatics*, *19*(4), 575–592.

Deeks, S. G. (2012). HIV: Shock and kill. *Nature*, *487*(7408), 439–440.

Deeks, S. G., Kitchen, C. M. R., Liu, L., Guo, H., Gascon, R., Narváez, A. B., Hunt, P., et al. (2004). Immune activation set point during early HIV infection predicts subsequent CD4+ T-cell changes independent of viral load. *Blood*, *104*(4), 942–947.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21.

Doitsh, G., Cavrois, M., Lassen, K. G., Zepeda, O., Yang, Z., Santiago, M. L., Hebbeler, A. M., et al. (2010). Abortive HIV infection mediates CD4 T cell depletion and inflammation in human lymphoid tissue. *Cell*, *143*(5), 789–801.

Doitsh, G., Galloway, N. L. K., Geng, X., Yang, Z., Monroe, K. M., Zepeda, O., Hunt, P. W., et al. (2014). Cell death by pyroptosis drives CD4 T-cell depletion in HIV-1 infection. *Nature*, *505*(7484), 509–514.

Douek, D. C., Brenchley, J. M., Betts, M. R., Ambrozak, D. R., Hill, B. J., Okamoto, Y., Casazza, J. P., et al. (2002). HIV preferentially infects HIV-specific CD4+ T cells. *Nature*, *417*(6884), 95–98.

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*(8), 1184–1191.

Dyer, W. B., Zaunders, J. J., Yuan, F. F., Wang, B., Learmont, J. C., Geczy, A. F., Saksena, N. K., et al. (2008). Mechanisms of HIV non-progression; robust and sustained CD4+ T-cell proliferative responses to p24 antigen correlate with control of viraemia and lack of disease progression after long-term transfusion-acquired HIV-1 infection. *Retrovirology*, *5*, 112.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology*, *23*(12), 1499–1501.

Ekblom, R., Smeds, L., & Ellegren, H. (2014). Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. *BMC Genomics*, *15*, 467.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048.

Ferre, A. L., Hunt, P. W., Critchfield, J. W., Young, D. H., Morris, M. M., Garcia, J. C., Pollard, R. B., et al. (2009). Mucosal immune responses to HIV-1 in elite controllers: a potential correlate of immune control. *Blood*, *113*(17), 3978–3989.

Février, M., Dorgham, K., & Rebollo, A. (2011). CD4+ T cell depletion in human immunodeficiency virus (HIV) infection: role of apoptosis. *Viruses*, *3*(5), 586–612.

Ganser-Pornillos, B. K., & Pornillos, O. (2019). Restriction of HIV-1 and other retroviruses by TRIM5. *Nature Reviews. Microbiology*, *17*(9), 546–556.

Gibbons, F. D., & Roth, F. P. (2002). Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Research*, *12*(10), 1574–1581.

Giorgi, J. V., Hultin, L. E., McKeating, J. A., Johnson, T. D., Owens, B., Jacobson, L. P., Shih, R., et al. (1999). Shorter survival in advanced human immunodeficiency virus type

1 infection is more closely associated with T lymphocyte activation than with plasma virus burden or virus chemokine coreceptor usage. *The Journal of Infectious Diseases*, *179*(4), 859–870.

Goh, W. W. B., Wang, W., & Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends in Biotechnology*, *35*(6), 498–507.

Golubovskaya, V., & Wu, L. (2016). Different Subsets of T Cells, Memory, Effector Functions, and CAR-T Immunotherapy. *Cancers*, *8*(3).

González, N., McKee, K., Lynch, R. M., Georgiev, I. S., Jimenez, L., Grau, E., Yuste, E., et al. (2018). Characterization of broadly neutralizing antibody responses to HIV-1 in a cohort of long term non-progressors. *Plos One*, *13*(3), e0193773.

Goonetilleke, N., Liu, M. K. P., Salazar-Gonzalez, J. F., Ferrari, G., Giorgi, E., Ganusov, V. V., Keele, B. F., et al. (2009). The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *The Journal of Experimental Medicine*, *206*(6), 1253–1272.

van Grevenynghe, J., Procopio, F. A., He, Z., Chomont, N., Riou, C., Zhang, Y., Gimmig, S., et al. (2008). Transcription factor FOXO3a controls the persistence of memory CD4(+) T cells during HIV infection. *Nature Medicine*, *14*(3), 266–274.

Gupta, Ravindra K, Abdul-Jawad, S., McCoy, L. E., Mok, H. P., Peppa, D., Salgado, M., Martinez-Picado, J., et al. (2019). HIV-1 remission following CCR5Δ32/Δ32 haematopoietic stem-cell transplantation. *Nature*, *568*(7751), 244–248.

Gupta, Ravindra Kumar, Peppa, D., Hill, A. L., Gálvez, C., Salgado, M., Pace, M., McCoy, L. E., et al. (2020). Evidence for HIV-1 cure after CCR5Δ32/Δ32 allogeneic haemopoietic stem-cell transplantation 30 months post-analytical treatment interruption: a case report. *The Lancet. HIV*, 7(5), E340-E347.

Holmes, D., Knudsen, G., Mackey-Cushman, S., & Su, L. (2007). FoxP3 enhances HIV-1 gene expression by modulating NFkappaB occupancy at the long terminal repeat in human T cells. *The Journal of Biological Chemistry*, *282*(22), 15973–15980.

Hosmalin, A., & Lebon, P. (2006). Type I interferon production in HIV-infected patients. *Journal of Leukocyte Biology*, *80*(5), 984–993.

Howard, M., & Paul, W. E. (1982). Interleukins for B lymphocytes. *Lymphokine research*, *1*(1), 1–4.

Hughitt, K. (2016, June 15). *Co-expression network analysis using RNA-Seq data*. Workshop presented at the DC ISCB Bioinformatics Genomics and Computational Biology. Retrieved July 8, 2019, from https://github.com/iscb-dc-rsg/2016-summer-workshop/tree/master/3B-Hughitt-RNASeq-Coex-Network-Analysis/tutorial

Hütter, G., Nowak, D., Mossner, M., Ganepola, S., Müssig, A., Allers, K., Schneider, T., et al. (2009). Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *The New England Journal of Medicine*, *360*(7), 692–698.

Iaccino, E., Schiavone, M., Fiume, G., Quinto, I., & Scala, G. (2008). The aftermath of the Merck's HIV vaccine trial. *Retrovirology*, *5*, 56.

International HIV Controllers Study, Pereyra, F., Jia, X., McLaren, P. J., Telenti, A., de Bakker, P. I. W., Walker, B. D., et al. (2010). The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science*, *330*(6010), 1551–1557.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. *Englewood Cliffs: Prentice Hall*, *1988*.

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., et al. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, *21*(9), 1543–1551.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127.

Karasavvas, N., Billings, E., Rao, M., Williams, C., Zolla-Pazner, S., Bailer, R. T., Koup, R. A., et al. (2012). The Thai Phase III HIV Type 1 Vaccine trial (RV144) regimen induces antibodies that target conserved regions within the V2 loop of gp120. *AIDS Research and Human Retroviruses*, *28*(11), 1444–1457.

Kaufmann, J., & Schering, A. G. (2014). Analysis of variance ANOVA. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J. L. Teugels (Eds.), *Wiley statsref: statistics reference online*. Chichester, UK: John Wiley & Sons, Ltd.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013).

TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36.

Kovacs, J. A., Lempicki, R. A., Sidorov, I. A., Adelsberger, J. W., Sereti, I., Sachau, W., Kelly, G., et al. (2005). Induction of prolonged survival of CD4+ T lymphocytes by intermittent IL-2 therapy in HIV-infected patients. *The Journal of Clinical Investigation*, *115*(8), 2139–2148.

Krishnan, S., Wilson, E. M. P., Sheikh, V., Rupert, A., Mendoza, D., Yang, J., Lempicki, R., et al. (2014). Evidence for innate immune system activation in HIV type 1-infected elite controllers. *The Journal of Infectious Diseases*, *209*(6), 931–939.

Kulpa, D. A., Talla, A., Brehm, J. H., Ribeiro, S. P., Yuan, S., Bebin-Blackwell, A.-G., Miller, M., et al. (2019). Differentiation into an Effector Memory Phenotype Potentiates HIV-1 Latency Reversal in CD4+ T Cells. *Journal of Virology*, *93*(24).

Kwon, K. J., Timmons, A. E., Sengupta, S., Simonetti, F. R., Zhang, H., Hoh, R., Deeks, S. G., et al. (2020). Different human resting memory CD4+ T cell subsets show similar low inducibility of latent HIV-1 proviruses. *Science Translational Medicine*, *12*(528).

Laeyendecker, O., Rothman, R. E., Henson, C., Horne, B. J., Ketlogetswe, K. S., Kraus, C. K., Shahan, J., et al. (2008). The effect of viral suppression on cross-sectional incidence testing in the Johns Hopkins Hospital emergency department. *Journal of Acquired Immune Deficiency Syndromes (1999)*, *48*(2), 211–215.

Laher, F., Moodie, Z., Cohen, K. W., Grunenberg, N., Bekker, L.-G., Allen, M., Frahm, N., et al. (2020). Safety and immune responses after a 12-month booster in healthy HIV-uninfected adults in HVTN 100 in South Africa: A randomized double-blind placebo-controlled trial of ALVAC-HIV (vCP2438) and bivalent subtype C gp120/MF59 vaccines. *PLoS Medicine*, *17*(2), e1003038.

Lalaoui, N., Boyden, S. E., Oda, H., Wood, G. M., Stone, D. L., Chau, D., Liu, L., et al. (2020). Mutations that prevent caspase cleavage of RIPK1 cause autoinflammatory disease. *Nature*, *577*(7788), 103–108.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559.

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, *24*(5), 719–720.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29.

Lee, G. Q., & Lichterfeld, M. (2017). Diversity of HIV-1 reservoirs in CD4+ T-cell subpopulations. *Current Opinion in HIV and AIDS*, *11*(4), 383–387.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M. G., Haag, J. D., et al. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, *29*(8), 1035–1043.

Liu, Z., Cumberland, W. G., Hultin, L. E., Kaplan, A. H., Detels, R., & Giorgi, J. V. (1998). CD8+ T-lymphocyte activation in HIV-1 disease reflects an aspect of pathogenesis distinct from viral burden and immunodeficiency. *Journal of acquired immune deficiency syndromes and human retrovirology : official publication of the International Retrovirology Association*, *18*(4), 332–340.

Li, Jianqiang, Zhou, D., Qiu, W., Shi, Y., Yang, J.-J., Chen, S., Wang, Q., et al. (2018). Application of Weighted Gene Co-expression Network Analysis for Data from Paired Design. *Scientific Reports*, *8*(1), 622.

Li, Jun, & Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, *22*(5), 519–536.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Margolis, D. M. (2011). Histone deacetylase inhibitors and HIV latency. *Current Opinion in HIV and AIDS*, *6*(1), 25–29.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, *18*(9), 1509–1517.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10.

Martin, M. P., Gao, X., Lee, J.-H., Nelson, G. W., Detels, R., Goedert, J. J., Buchbinder, S., et al. (2002). Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nature Genetics*, *31*(4), 429–434.

Martin, M. P., Qi, Y., Gao, X., Yamada, E., Martin, J. N., Pereyra, F., Colombo, S., et al. (2007). Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. *Nature Genetics*, *39*(6), 733–740.

Mason, M. J., Fan, G., Plath, K., Zhou, Q., & Horvath, S. (2009). Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics*, *10*, 327.

Md, R. R. R., Thomas A Fleisher Md Faaaai Facaai, William T. Shearer Md Phd, Schroeder, H., Anthony J. Frew Md Frcp, & Cornelia M. Weyand Md Phd. (2018). *Clinical Immunology: Principles And Practice* (5th ed., p. 1392). Elsevier.

Mens, H., Kearney, M., Wiegand, A., Shao, W., Schønning, K., Gerstoft, J., Obel, N., et al. (2010). HIV-1 continues to replicate and evolve in patients with natural control of HIV infection. *Journal of Virology*, *84*(24), 12971–12981.

Messi, M., Giacchetto, I., Nagata, K., Lanzavecchia, A., Natoli, G., & Sallusto, F. (2003). Memory and flexibility of cytokine gene expression as separable properties of human T(H)1 and T(H)2 lymphocytes. *Nature Immunology*, *4*(1), 78–86.

Migueles, S. A., & Connors, M. (2015). Success and failure of the cellular immune response against HIV-1. *Nature Immunology*, *16*(6), 563–570.

Migueles, S. A., Laborico, A. C., Shupert, W. L., Sabbaghian, M. S., Rabin, R., Hallahan, C. W., Van Baarle, D., et al. (2002). HIV-specific CD8+ T cell proliferation is coupled to perforin expression and is maintained in nonprogressors. *Nature Immunology*, *3*(11), 1061–1068.

Migueles, S. A., Osborne, C. M., Royce, C., Compton, A. A., Joshi, R. P., Weeks, K. A., Rood, J. E., et al. (2008). Lytic granule loading of CD8+ T cells is required for HIV-infected cell elimination associated with immune control. *Immunity*, *29*(6), 1009–1021.

Miura, T., Brockman, M. A., Brumme, C. J., Brumme, Z. L., Carlson, J. M., Pereyra, F., Trocha, A., et al. (2008). Genetic characterization of human immunodeficiency virus

type 1 in elite controllers: lack of gross genetic defects or common amino acid changes. *Journal of Virology*, *82*(17), 8422–8430.

Morgan, D., Mahe, C., Mayanja, B., Okongo, J. M., Lubega, R., & Whitworth, J. A. G. (2002). HIV-1 infection in rural Africa: is there a difference in median time to AIDS and survival compared with that in industrialized countries? *AIDS*, *16*(4), 597–603.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628.

Müllner, D. (2013). fastcluster : Fast Hierarchical, Agglomerative Clustering Routines for *R* and *Python*. *Journal of statistical software*, *53*(9), 1–18.

Nakajima, H., Iwamoto, I., Tomoe, S., Matsumura, R., Tomioka, H., Takatsu, K., & Yoshida, S. (1992). CD4+ T-lymphocytes and interleukin-5 mediate antigen-induced eosinophil infiltration into the mouse trachea. *The American Review of Respiratory Disease*, *146*(2), 374–377.

Nikolic, D. S., Lehmann, M., Felts, R., Garcia, E., Blanchet, F. P., Subramaniam, S., & Piguet, V. (2011). HIV-1 activates Cdc42 and induces membrane extensions in immature dendritic cells to facilitate cell-to-cell virus propagation. *Blood*, *118*(18), 4841–4852.

Noel, N., Peña, R., David, A., Avettand-Fenoel, V., Erkizia, I., Jimenez, E., Lecuroux, C., et al. (2016). Long-Term Spontaneous Control of HIV-1 Is Related to Low Frequency of Infected Cells and Inefficient Viral Reactivation. *Journal of Virology*, *90*(13), 6148–6158.

Novis, C. L., Archin, N. M., Buzon, M. J., Verdin, E., Round, J. L., Lichterfeld, M., Margolis, D. M., et al. (2013). Reactivation of latent HIV-1 in central memory $CD4^+$ T cells through TLR-1/2 stimulation. *Retrovirology*, *10*, 119.

Nozza, S., Canducci, F., Galli, L., Cozzi-Lepri, A., Capobianchi, M. R., Ceresola, E. R., Narciso, P., et al. (2012). Viral tropism by geno2pheno as a tool for predicting CD4 decrease in HIV-1-infected naive patients with high CD4 counts. *The Journal of Antimicrobial Chemotherapy*, *67*(5), 1224–1227.

Nygaard, V., Rødland, E. A., & Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses.

Biostatistics, *17*(1), 29–39.

Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M., et al. (2016). Clustering algorithms: their application to gene expression data. *Bioinformatics and biology insights*, *10*, 237–253.

O'Connell, K. A., Brennan, T. P., Bailey, J. R., Ray, S. C., Siliciano, R. F., & Blankson, J. N. (2010). Control of HIV-1 in elite suppressors despite ongoing replication and evolution in plasma virus. *Journal of Virology*, *84*(14), 7018–7028.

Pan, X., Baldauf, H.-M., Keppler, O. T., & Fackler, O. T. (2013). Restrictions to HIV-1 replication in resting CD4+ T lymphocytes. *Cell Research*, *23*(7), 876–885.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419.

Pierson, T. C., Kieffer, T. L., Ruff, C. T., Buck, C., Gange, S. J., & Siliciano, R. F. (2002). Intrinsic stability of episomal circles formed during human immunodeficiency virus type 1 replication. *Journal of Virology*, *76*(8), 4138–4144.

Pimentel, H., Bray, N. L., Puente, S., Melsted, P., & Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, *14*(7), 687–690.

Pirim, H., Ekşioğlu, B., Perkins, A., & Yüceer, C. (2012). Clustering of High Throughput Gene Expression Data. *Computers & operations research*, *39*(12), 3046–3061.

Potter, S. J., Lacabaratz, C., Lambotte, O., Perez-Patrigeon, S., Vingert, B., Sinet, M., Colle, J.-H., et al. (2007). Preserved central memory and activated effector memory CD4+ T-cell subsets in human immunodeficiency virus controllers: an ANRS EP36 study. *Journal of Virology*, *81*(24), 13904–13915.

Punt, J. (2013). *Cancer Immunotherapy*. (G. C. Prendergast & E. M. Jaffee, Eds.). Elsevier.

Raphael, I., Nalawade, S., Eagar, T. N., & Forsthuber, T. G. (2015). T cell subsets and their signature cytokines in autoimmune and inflammatory diseases. *Cytokine*, *74*(1), 5–17.

Rasmussen, T. A., Tolstrup, M., Brinkmann, C. R., Olesen, R., Erikstrup, C., Solomon, A., Winckelmann, A., et al. (2014). Panobinostat, a histone deacetylase inhibitor, for latent-

virus reactivation in HIV-infected patients on suppressive antiretroviral therapy: a phase 1/2, single group, clinical trial. *The lancet. HIV*, *1*(1), e13-21.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, *297*(5586), 1551–1555.

Richards, A. L., Holmans, P., O'Donovan, M. C., Owen, M. J., & Jones, L. (2008). A comparison of four clustering methods for brain expression microarray data. *BMC Bioinformatics*, *9*, 490.

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 Research*, *43*(7), e47.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Rosenblum, M. D., Way, S. S., & Abbas, A. K. (2016). Regulatory T cell memory. *Nature Reviews. Immunology*, *16*(2), 90–101.

Rousseeuw, P. J., & Kaufman, L. (1990). Finding groups in data. *Hoboken: Wiley Online Library*.

R Core Team. (2018). *R: A language and environment for statistical computing. R Foundation for Statistical Computing* . Computer software, Vienna, Austria: R Project. Retrieved January 8, 2020, from https://www.R-project.org/

Sáez-Cirión, A., Bacchus, C., Hocqueloux, L., Avettand-Fenoel, V., Girault, I., Lecuroux, C., Potard, V., et al. (2013). Post-treatment HIV-1 controllers with a long-term virological remission after the interruption of early initiated antiretroviral therapy ANRS VISCONTI Study. *PLoS Pathogens*, *9*(3), e1003211.

Sáez-Cirión, A., Lacabaratz, C., Lambotte, O., Versmisse, P., Urrutia, A., Boufassa, F., Barré-Sinoussi, F., et al. (2007). HIV controllers exhibit potent CD8 T cell capacity to suppress HIV infection ex vivo and peculiar cytotoxic T lymphocyte activation phenotype. *Proceedings of the National Academy of Sciences of the United States of*

America, *104*(16), 6776–6781.

Sáez-Cirión, A., Pancino, G., Sinet, M., Venet, A., Lambotte, O., & ANRS EP36 HIV CONTROLLERS study group. (2007). HIV controllers: how do they tame the virus? *Trends in Immunology*, *28*(12), 532–540.

Sáez-Cirión, A., Sinet, M., Shin, S. Y., Urrutia, A., Versmisse, P., Lacabaratz, C., Boufassa, F., et al. (2009). Heterogeneity in HIV suppression by CD8 T cells from HIV controllers: association with Gag-specific CD8 T cell responses. *Journal of Immunology*, *182*(12), 7828–7837.

Sajadi, M. M., Constantine, N. T., Mann, D. L., Charurat, M., Dadzan, E., Kadlecik, P., & Redfield, R. R. (2009). Epidemiologic characteristics and natural history of HIV-1 natural viral suppressors. *Journal of Acquired Immune Deficiency Syndromes (1999)*, *50*(4), 403–408.

Sallusto, F, Lenig, D., Förster, R., Lipp, M., & Lanzavecchia, A. (1999). Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature*, *401*(6754), 708–712.

Sallusto, Federica, Geginat, J., & Lanzavecchia, A. (2004). Central memory and effector memory T cell subsets: function, generation, and maintenance. *Annual Review of Immunology*, *22*, 745–763.

Scheid, J. F., Mouquet, H., Feldhahn, N., Seaman, M. S., Velinzon, K., Pietzsch, J., Ott, R. G., et al. (2009). Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature*, *458*(7238), 636–640.

Seitz, R. (2016). Human immunodeficiency virus (HIV). *Transfusion medicine and hemotherapy : offizielles Organ der Deutschen Gesellschaft fur Transfusionsmedizin und Immunhamatologie*, *43*(3), 203–222.

Selinger, C., & Katze, M. G. (2013). Mathematical models of viral latency. *Current opinion in virology*, *3*(4), 402–407.

Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M., & Ligterink, W. (2016). Learning from Co-expression Networks: Possibilities and Challenges. *Frontiers in plant science*, *7*, 444.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., et al.

(2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504.

Sharma, V., Bryant, C., Montero, M., Creegan, M., Slike, B., Krebs, S. J., Ratto-Kim, S., et al. (2020). Monocyte and CD4+ T cell antiviral and innate responses associated with HIV-1 inflammation and cognitive impairment. *AIDS*, 34(9), 1289-1301.

Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. [version 2; peer review: 2 approved]. *F1000Research*, *4*, 1521.

Soumelis, V., Scott, I., Gheyas, F., Bouhour, D., Cozon, G., Cotte, L., Huang, L., et al. (2001). Depletion of circulating natural type 1 interferon-producing cells in HIV-infected AIDS patients. *Blood*, *98*(4), 906–912.

Stafford, K. A., Rikhtegaran Tehrani, Z., Saadat, S., Ebadi, M., Redfield, R. R., & Sajadi, M. M. (2017). Long-term follow-up of elite controllers: Higher risk of complications with HCV coinfection, no association with HIV disease progression. *Medicine*, *96*(26), e7348.

Stephens, J. C., Reich, D. E., Goldstein, D. B., Shin, H. D., Smith, M. W., Carrington, M., Winkler, C., et al. (1998). Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *American Journal of Human Genetics*, *62*(6), 1507–1515.

Suh, H.-S., Cosenza-Nashat, M., Choi, N., Zhao, M.-L., Li, J., Pollard, J. W., Jirtle, R. L., et al. (2010). Insulin-like growth factor 2 receptor is an IFNgamma-inducible microglial protein that facilitates intracellular HIV replication: implications for HIV-induced neurocognitive disorders. *The American Journal of Pathology*, *177*(5), 2446–2458.

Svensson, V., Gayoso, A., Yosef, N., & Pachter, L. (2020). Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics, 36(11), 3418-3421*.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613.

Team, B. C. (2015). Homo. sapiens: Annotation package for the Homo. sapiens object. *R*

package version, *1*(1).

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338.

The UK register of HIV seroconverters. (1996). The UK register of HIV seroconverters: methods and analytical issues. UK register of HIV seroconverters (UKRHS) Steering Committee. *Epidemiology and Infection*, *117*(2), 305–312.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515.

Valdez, H., & Lederman, M. M. (1997). Cytokines and cytokine therapies in HIV infection. *AIDS clinical review*, 187–228.

Vandergeeten, C., Fromentin, R., DaFonseca, S., Lawani, M. B., Sereti, I., Lederman, M. M., Ramgopal, M., et al. (2013). Interleukin-7 promotes HIV persistence during antiretroviral therapy. *Blood*, *121*(21), 4321–4329.

Van Lint, C., Bouchat, S., & Marcello, A. (2013). HIV-1 transcription and latency: an update. *Retrovirology*, *10*, 67.

Vidya Vijayan, K. K., Karthigeyan, K. P., Tripathi, S. P., & Hanna, L. E. (2017). Pathophysiology of CD4+ T-Cell Depletion in HIV-1 and HIV-2 Infections. *Frontiers in immunology*, *8*, 580.

Wagner, R. N., Reed, J. C., & Chanda, S. K. (2015). HIV-1 protease cleaves the serine-threonine kinases RIPK1 and RIPK2. *Retrovirology*, *12*, 74.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*(2), 117–186.

Wehrens, R., & Kruisselbrink, J. (2018). Flexible Self-Organising Maps in kohonen 3.0. *Journal of Statis*.

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(2), 180–185.

Willinger, T., Freeman, T., Hasegawa, H., McMichael, A. J., & Callan, M. F. C. (2005). Molecular signatures distinguish human central memory from effector memory CD8 T cell subsets. *Journal of Immunology*, *175*(9), 5895–5903.

W H O. (2018). *Global Health Observatory (GHO) data.* World Health Organsiation. Retrieved November 13, 2020, from http://www.who.int/gho/hiv/en/

Yang, X., Chen, Y., & Gabuzda, D. (1999). ERK MAP kinase links cytokine signals to activation of latent HIV-1 infection by stimulating a cooperative interaction of AP-1 and NF-kappaB. *The Journal of Biological Chemistry*, *274*(39), 27981–27988.

Yang, Z., & Engel, J. D. (1993). Human T cell transcription factor GATA-3 stimulates HIV-1 expression. *Nucleic Acids Research*, *21*(12), 2831–2836.

Yates, A., Stark, J., Klein, N., Antia, R., & Callard, R. (2007). Understanding the slow depletion of memory CD4+ T cells in HIV infection. *PLoS Medicine*, *4*(5), e177.

Younes, S.-A., Yassine-Diab, B., Dumont, A. R., Boulassel, M.-R., Grossman, Z., Routy, J.-P., & Sekaly, R.-P. (2003). HIV-1 viremia prevents the establishment of interleukin 2-producing HIV-specific memory CD4+ T cells endowed with proliferative capacity. *The Journal of Experimental Medicine*, *198*(12), 1909–1922.

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology*, *16*(5), 284–287.

Zerbato, J. M., Serrao, E., Lenzi, G., Kim, B., Ambrose, Z., Watkins, S. C., Engelman, A. N., et al. (2016). Establishment and Reversal of HIV-1 Latency in Naive and Central Memory CD4+ T Cells In Vitro. *Journal of Virology*, *90*(18), 8059–8073.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*, Article17.

6 Appendix A - Supplementary Results

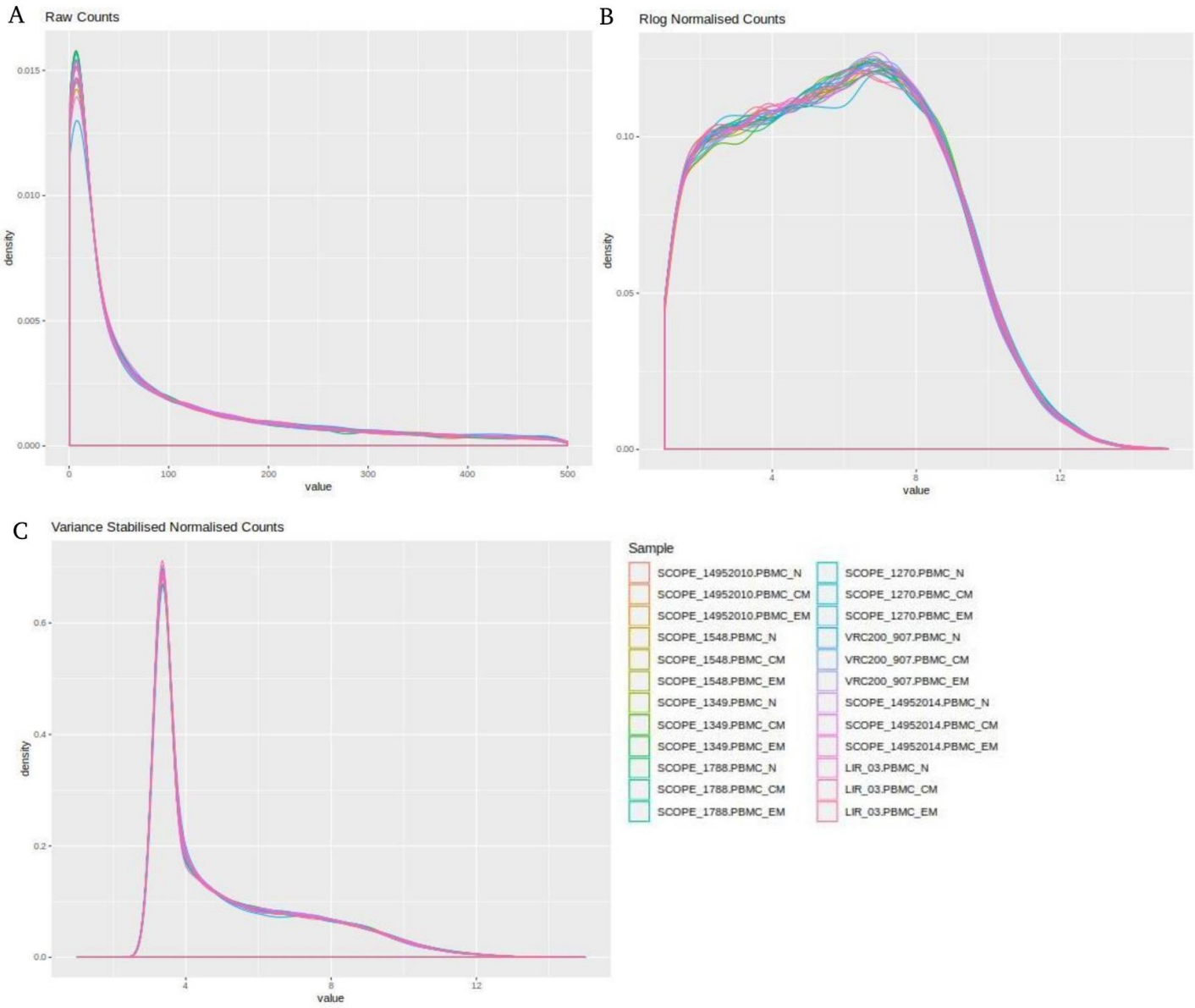


Figure S1: Sample distribution plots (A): Raw Counts (B): rlog Normalised Counts and (C): Variance Stabilised Counts. Colours represent the 24 different samples examined. Distributions highlight the effects of normalisation procedures on datasets. Raw counts maintained the highest spread of data whilst variance stabilised counts reduced most variation. Size factor and rlog counts show a high degree of variability in the different sample distributions.

.

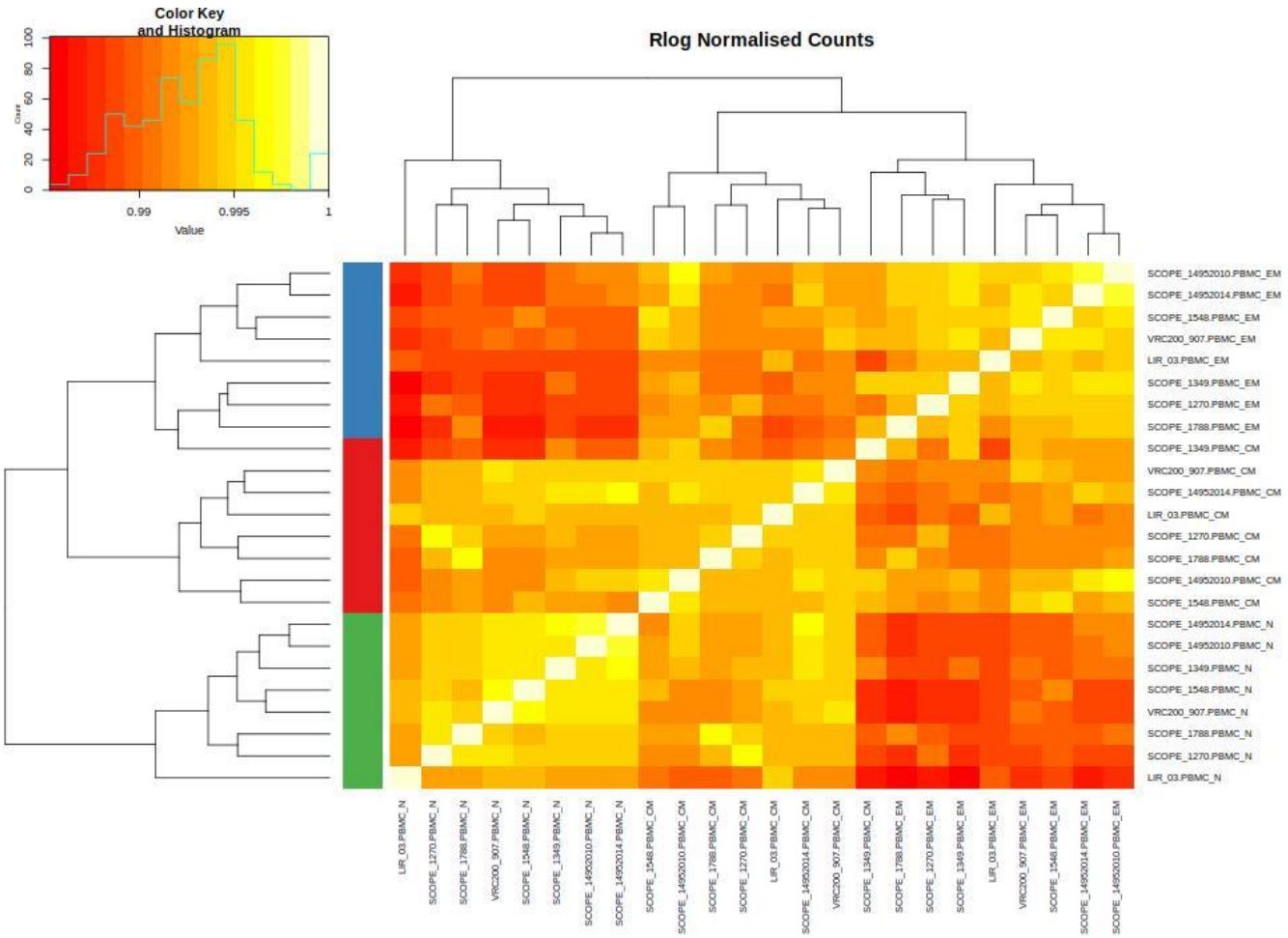


Figure S2: Sample heatmap with hierarchical clustering of rlog Normalised Counts. Colours depict sample PCC's. Values close to 1 (high correlation) are shown in light colours and closer to 0 as darker colours.

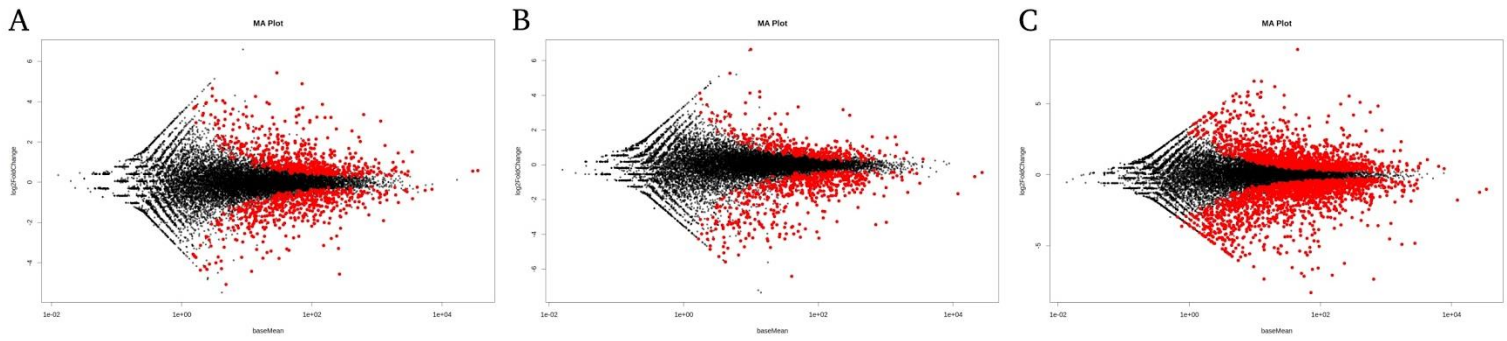


Figure S3: M (log ratio plotted on y-axis) A (mean average plotted on x-axis) plots for DESeq2 (Love et al., 2014). Show similarities in the number of up and downregulated genes across (A) CD4+ T CM vs EM, (B) CD4+ T CM vs N and (C) CD4+ T EM vs N. Red values are significantly differentially expressed genes.

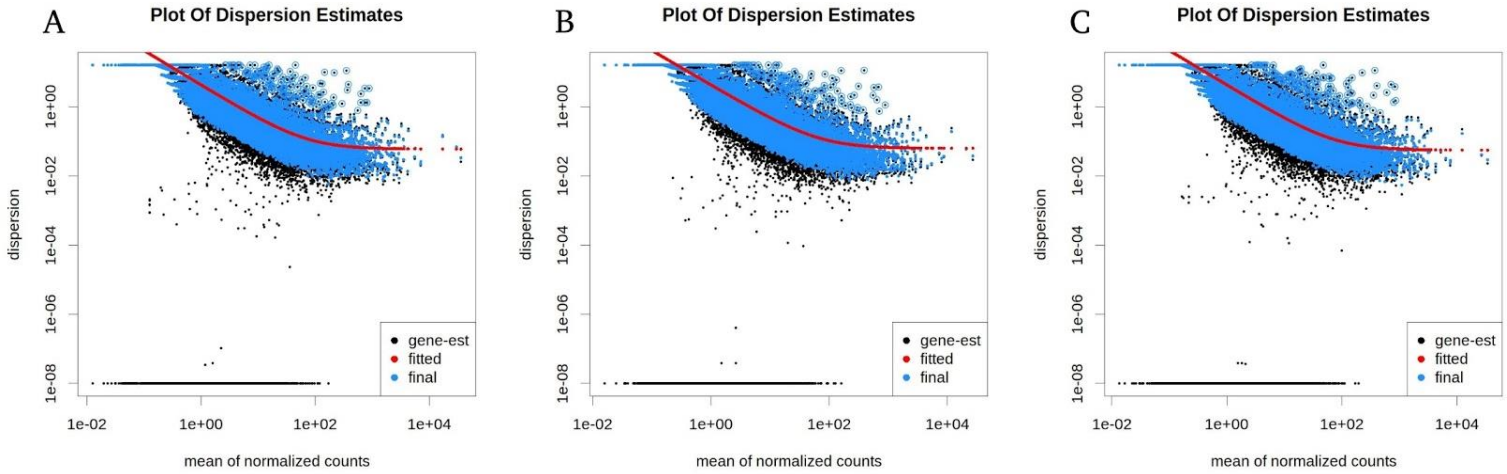


Figure S4: DESeq2 Dispersion estimate plots for different CD4$^+$ T cell subsets (A): CD4$^+$ T CM vs EM, (B) CD4+ T CM vs N and (C) CD4+ T EM vs N. Black dots show initial maximum likelihood gene wise dispersion estimates, a red curve matches the dispersion mean relationship, blue dots show estimates after shrinkage estimation and circled dots show dispersion outliers that do not undergo shrinkage.

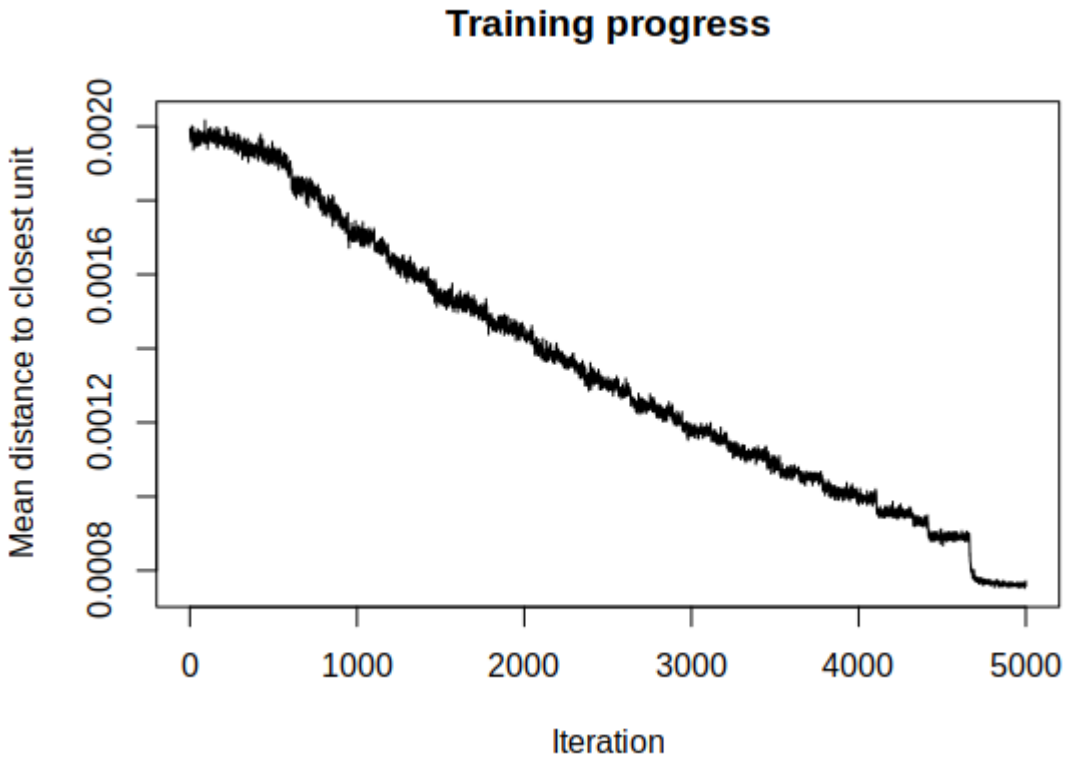
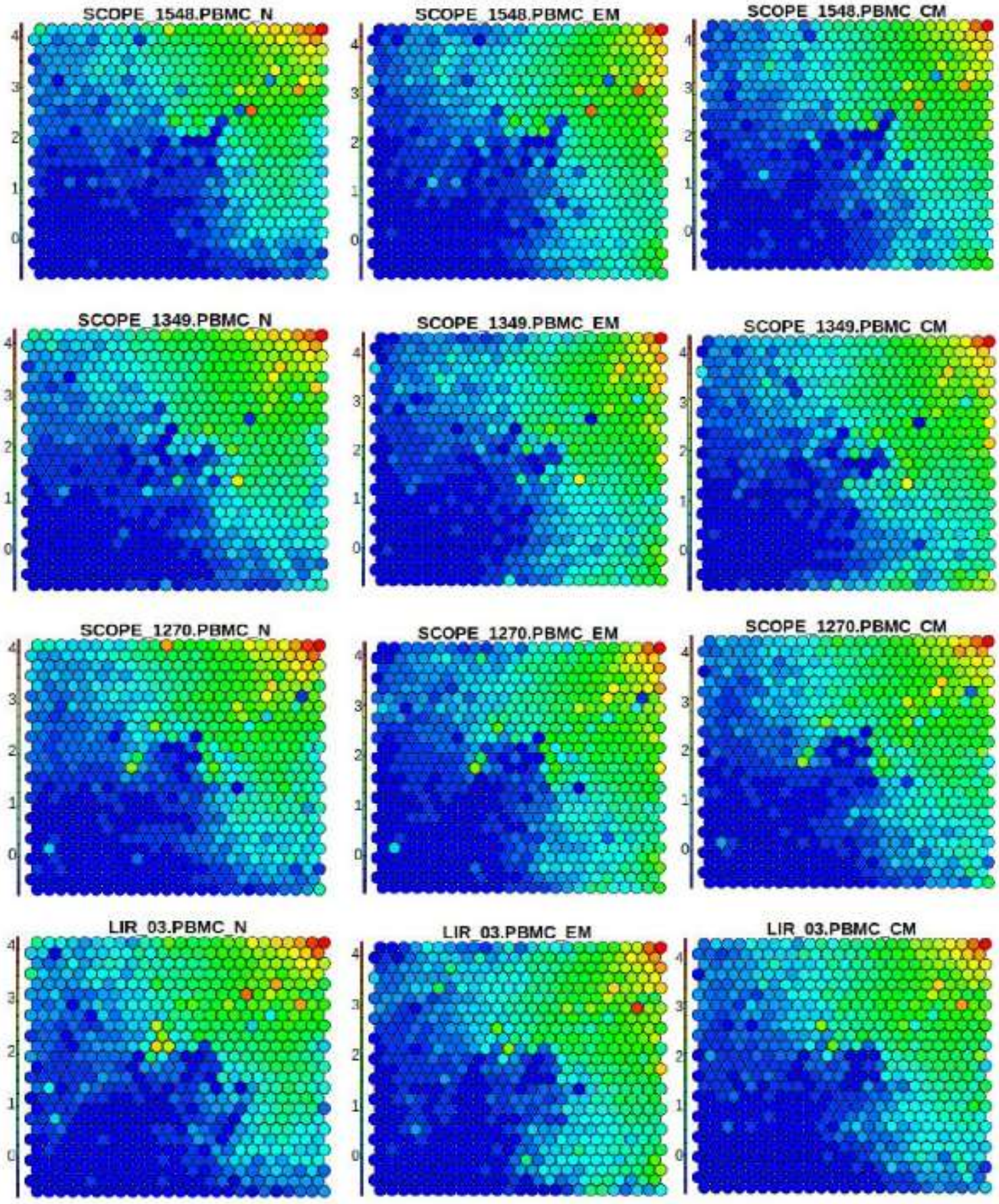


Figure S5: Mean distance between neurons and genes with higher-dimensional datasets. Mean distance between units reaches a plateau 5000 indicating that the minimum distance between nodes and genes has been reached


SCOPE_1548.PBMC_N
SCOPE_1548.PBMC_EM
SCOPE_1548.PBMC_CM
SCOPE_1349.PBMC_N
SCOPE_1349.PBMC_EM
SCOPE_1349.PBMC_CM
SCOPE_1270.PBMC_N
SCOPE_1270.PBMC_EM
SCOPE_1270.PBMC_CM
LIR_03.PBMC_N
LIR_03.PBMC_EM
LIR_03.PBMC_CM

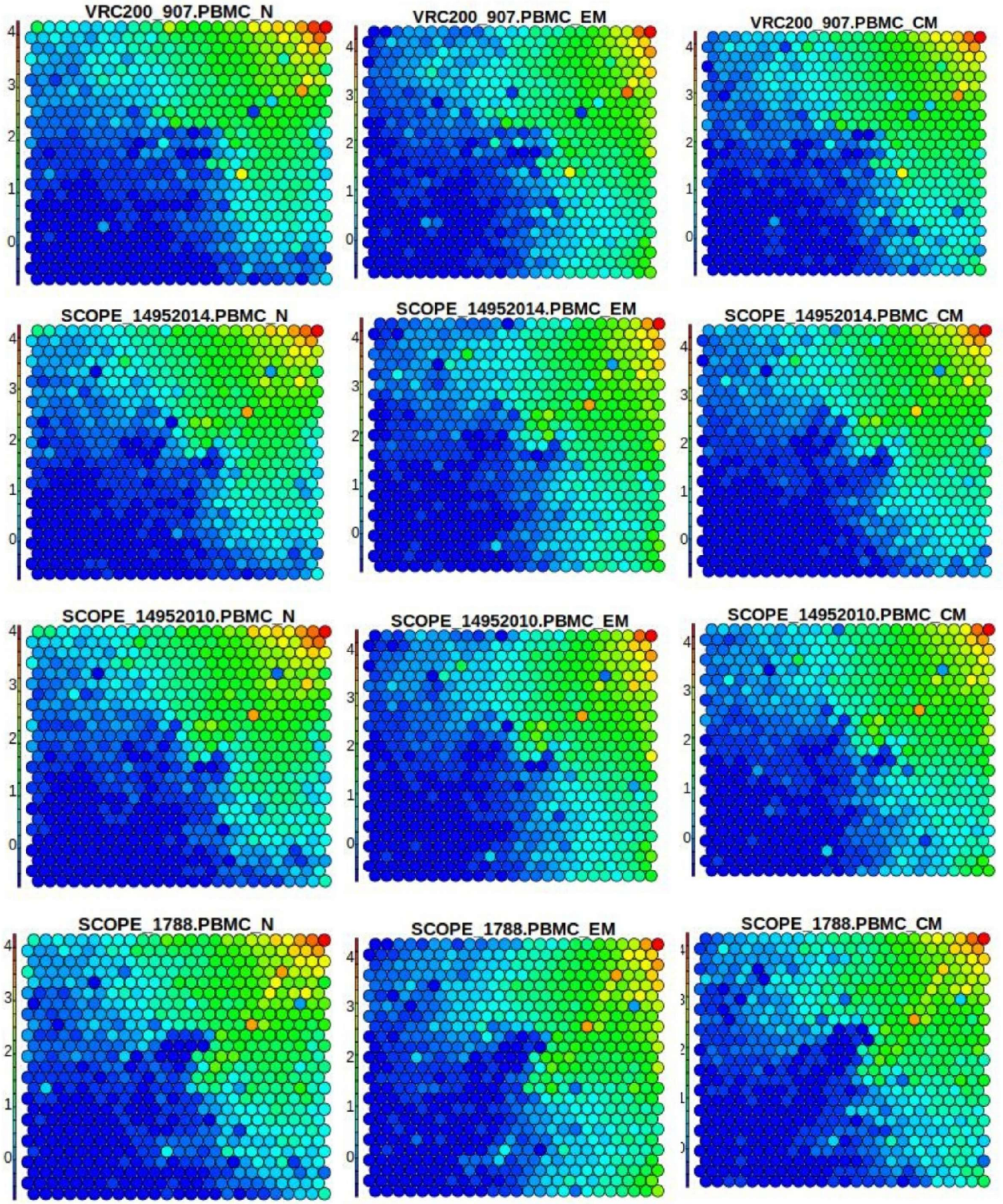


Figure S6: Self Organising Map - heat maps for CD4$^+$ T_{EM} (top), T_N (middle) and T_{CM} (bottom) subsets for all eight individuals analysed in the final cohort. Each plot represents the expression profile of 30 157 genes per cell sample for an individual. Colours represent scaled variance stabilised counts (between 0 and 4) of multiple genes that map to a particular node. Similarities in topology across all samples indicate transcriptional similarities in the same CD4$^+$ cell type. T_N and T_{CM} have visually similar maps indicating similarities in expression. T_N and T_{EM} have differences in topology which may be attributable to their differences.

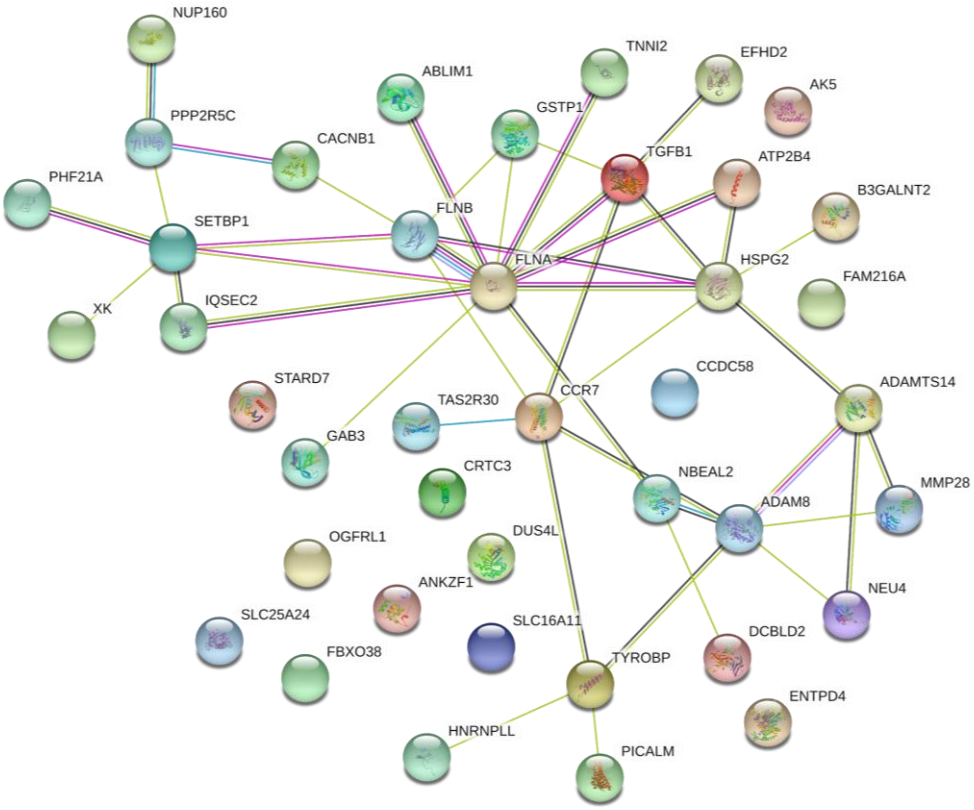


Figure S7: Visualisation of hub genes identified with WGCNA and their respective known interactions from STRING. Hub genes identified by WGCNA visualised using STRING with a high confidence level of 0.700, no more than 10 interactions for the first shell and no interactions for the second shell.

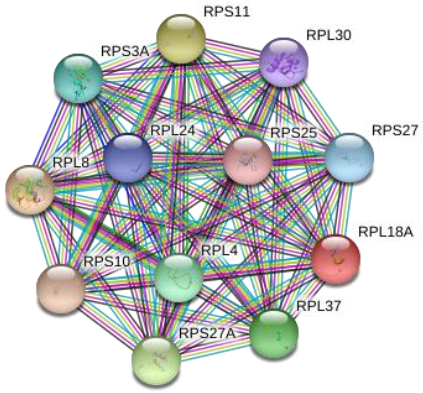


Figure S8: Ribosomal Proteins found to be differentially co-expressed across the CD4$^+$ T_N vs T_{EM} comparison and visualised in STRING. The majority of these genes were found to be expressed at a greater level in T_{EM} cells compared to T_N subsets. Genes visualised using STRING with a high confidence level of 0.700, no more than 10 interactions for the first shell and no interactions for the second shell.

7 Appendix B - Pseudocode

Pseudocode included in this section summarises the key blocks of code used in this study.

Workflow:

kalisto_script.sh	alignment of and quantification of reads
differential_expression.R	differential expression analysis
normalisation.R	normalisation for network analyses
coexpression_wgcna.R	co-expression analysis with WGCNA
self_Organising_Map.R	co-expression analysis with SOM
k_means_clustering.R	co-expression analysis with k-means
coexpression_method_comparison.R	co-expression method comparison

kallisto_script.sh

This script allowed **for** pseudo-alignment reference indexing **and** quantification

get BiomaRt transcripts file

build index.idx file **for** alignment

for (each pair of fastq files per sample) {
    quantify transcript level abundance
    bootstrap estimates = 100
}

```
differential_expression.R


This script identified differentially expressed genes with DESeq2.

packages <- "ggplot2", "knitr", "DeSeq2", "reshape2", "dplyr",
"RColorBrewer", "Homo.sapiens","RColorBrewer"
load packages

import {
       meta_data
       hdf5_files #from kallisto
       trancript_to_gene #from biomaRt
}

function create_tximport_object(meta_data,  hdf5_file,
trancript_to_gene){
    create tximport objects #gene level abundance quantification
    return(tximport_object)
}

function run_deseq(tximport_object, comparison){
    remove low variance and low count genes
    fit deseq2 model
    filter by Benjamini Hochburg adjusted p < 0.05
    order by log2FoldChange
    return(deseq_object)
}


comparisons <- (CD4+ EM vs CM; CD4+ EM vs N; CD4+ CM vs N)

for(comparisons) {

    select samples_for_comparison
```

```
    create_tximport_object()
    run_deseq()
    plot MA Dispersion and PCA plots
    write significantly differentially expressed genes to file

}
```

normalisation.R

```
This script normalised data for library differences and
heteroscedasticity and removed outliers.

packages <- "fastcluster", "DESeq2"
load(packages)

import{
       raw_counts
       meta_data
}

function remove_outliers(raw_counts, meta_data){
    calculate euclidean distance matrix
    hierarchical clustering
    visualise dendrogram
    option to cut tree to remove outliers
    cut dynamic tree
    return(meta_data_samples_outliers_removed)

}

function normalise_it(raw_counts, meta_data_samples_removed){

    convert counts to integers

    if(all_genes_unique = FALSE){
      stop_code

    }

  create DESeq2_object #with no design formula
```

```
  sizefactor_counts <- generate_sizesizeFactors(DESeq2_object)
  variance_stabilised_counts <- generate_VST(DESeq2_object)
  rlog_counts <- rlogTransformation(DESeq2_object)
  log2_counts <- log2(input_counts + 1)

  all_counts <- list(sizefactor_counts, variance_stabilised_counts,
rlog_counts, log2_counts, raw_counts, input_meta)

  return(all_counts)


}
normalised_counts <- normalise_it(raw_counts, meta_data_samples_removed)
```

```
coexpression_wgcna.R
```

```
This script underwent co-expression analysis with WGCNA.

packages <- "gplots", "tidyverse", "riskyr", "ggplot2",
"knitr","reshape2", "dplyr", "RColorBrewer", "WGCNA",
"Homo.sapiens","RColorBrewer"
load packages

import{
       normalised_counts
       meta_data
       differential expressed gene lists
}
```

```
get{
  annotation metadata for genes #using biomaRt
}


source {
    heatmap_script.R
    distribution_plot_script.R
    plot_scale_independence.R
    export_network_to_graphml.R

}

for("variance_stabilised_counts", "rlog_counts", "raw_counts",
"size_factor_counts"){
  plot_sample_heatmap()
  plot_distribution()
  save to png format
}

counts_input <- normalsied_counts$variance_stabilised_counts

for comparison

comparisons <- (CD4+ EM vs CM; CD4+ EM vs N; CD4+ CM vs N)

for(comparisons) {

    select samples_for_comparison
    select differentially expressed genes for comparison

    create similarity_matrix
      plot similarity_matrix
    create adjacency_matrix
      plot adjacency_matrix
```

```
    plot_scale_independence_mean_connectivity()
    select soft threshold value

    calculate Topological Overlap Measure
    form distance matrix
    hierarchical cluster distance matrix
      plot cluster Dendrogram

  identify hub genes
      output hub gene list

    assign logFC scores to genes
    select hard threshold value

    export_network_to_graphml()
    write WGCNA cluster assignments to file

}
```

```
Self_Organising_Map.R


This script underwent co-expression analysis with SOM.

packages <- "kohonen", "dummies", "ggplot2", "rgdal", "rgdal", "sp",
"reshape2", "maptools", "rgeos"

for (packages) {
    load packages

}

import{
       normalised_counts
       meta_data
       differential expressed gene lists
}

all_genes <- normalised_counts$variance_stabilised_counts


function train_som(normalised_counts,  rlen,  alpha, grid size){

    train self organising map
    run hierarchical clustering within sum of squares plot
    run hierarchical clustering over map
    write gene cluster assignments to file

}

comparisons <- (CD4+ EM vs CM; CD4+ EM vs N; CD4+ CM vs N)

for(comparisons){
  select samples for comparison
```

```
  select differentially expressed genes

  define SOM som_grid
  define rlen #the number of model iterations
  define alpha #the learning rate
  train_som()
    plot distances between unit
    for(all samples) {
      plot heatmaps per sample
  }

}

train_som(all_genes)
plot sample heatmaps for all samples using all genes
```

```
K_means_clustering.R

This script underwent co-expression analysis with k-means clustering.

import{
       normalised_counts
       differential expressed gene lists
}

function K_means_clustering(normalised_counts, differential expressed
list){

  calculate mean of counts per gene
```

```
  for(i in 1:15){
    calculate within sum of squares for i clusters
    plot within sum of squares
  }
  select optimum number of clusters
  select number of iterations = 50
  select number of resamples = 50
  run k-means model
  plot clusters
  write cluster assignments to file

}

comparisons <- (CD4+ EM vs CM; CD4+ EM vs N; CD4+ CM vs N)
for(comparisons){
  select samples for comparisons
  select differentially expressed genes
  K_means_clustering()
}
```

```
coexpression_method_comparison.R


This script compared co-expression analysis approaches.

packages <- "tidyverse", "clusterProfiler", "org.Hs.eg.db", "biomaRt",
"tidyr", "forcats", "ggplot2", "praise"


load(packages)

source functions {
    getmodule.R   #gets a specific gene module or cluster within a method
    get_entrez_function.R #gets the right entrez ids needed for
clusterProfiler
    clusterit.R #clusterProfiler script undergoes GO enrichment and
calculates gene ratio scores for a given module / cluster
    almalgamate_GO.R #puts all GO terms identified into a single dataframe
for analysis
}

import{
       SOM_cluster_assighnments
       WGCNA_cluster_assighnments
       k-means_cluster_assignment
}


list_GO <- NULL #initialise object to store GO terms

function get_gene_ratios(gene_list_with_cluster_assighnment){

  for(module in range[1:number(modules)]){

    genes_in_module <- getmodule(gene_list, module)

    genes_entrez <- getentrezids(genes_in_module)
```

```
    remove duplicate ids

    GO_term <-  run_cluster_profiler(genes_entrez)

    aggregate all GO terms for a method

    get gene ratios

    return(gene_ratios)

    }


}

methods <- (SOM_cluster_assighnments; WGCNA_cluster_assighnments; k-
means_cluster_assignment)

for(methods){

  get_gene_ratios(methods)
  amalgamate gene ratios for all methods

}

plot box and whisker with all gene_ratios
run ANOVA hypothesis test
per method return mean(gene_ratios)
```