

https://dspace.org/

Bioinformatic pipelines for transcriptome analyses : understanding gene expression in black South Africans with systemic sclerosis

Mpangase, Phelelani Thokozani

2019

https://hdl.handle.net/10539/29858

Mpangase, Phelelani Thokozani, Bioinformatic pipelines for transcriptome analyses: understanding gene expression in black South Africans with systemic sclerosis, University of the Witwatersrand, Johannesburg, <http://hdl.handle.net/10539/29858>

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



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Bioinformatic pipelines for transcriptome analyses: Understanding gene expression in black South Africans with systemic sclerosis

Phelelani Thokozani Mpangase

A Thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy.

Johannesburg, 2019

Declaration

I Phelelani Thokozani Mpangase declare that this Thesis is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



Phelelani T. Mpangase

on 07 June 2019 at Parktown, Johannesburg

This thesis is specially dedicated to the memory of my late grandmother, Thombile Silvina "MaMdwengula" Mbele, who passed away during the course of this PhD. You may have passed on from this world, but your love, wisdom and guidance will forever live in our hearts.

To my parents, Aletta and Mike Mpangase, and my siblings, Siyabonga, Zamangwane, MaSisi and MaBhuti Mpangase. I hope the news of this PhD shock you as much as it did me.

To my partner Sisonke Ngantweni and my son Butho Mpangase, thank you for being beside me throughout this journey. Cubsy-Wobsy, I hope one day you understand why I could not sleep at night or take you out to ride your bike.

To all of my family and friends, thank you for your support, encouragement and always believing in me.

Acknowledgements

I would like to express my utmost gratitude to the following people and institutional bodies who have contributed towards the completion of this PhD:

- My supervisors Prof. Michèle Ramsay and Prof. Scott Hazelhurst. Thank you for having seen potential in me and motivating me to do this work. Thank you for your patience, guidance and always being there throughout this journey.

- Dr. Jacqueline Frost and Prof. M. Tikly for kindly providing the RNA-seq data used in this study and having the patience to answer all my questions.

- Shaun Aron and Freedom Mukomana at the Sydney Brenner Institute for Molecular Bioscience (SBIMB) Bioinformatics Office. Thank you for all your advice, support and tolerance throughout this PhD, and thank you for not getting the swear jar, otherwise I would be bankrupt by now.

- Yusuf Ismail, Jocelyn Geyenga and Mmatshepo Taunyane at the SBIMB Projects Office. Your assistance with all the administration stuff was greatly appreciated.

- My colleagues and friends at SBIMB, especially Carl Chen, Natalie Smyth, Romuald Boua, Hleli Mthimkhulu, Mahtaab Hayat, Cassandra Soo, Juddy Mamabolo, Busisiwe Mthembu, Evans Mathebula, Jenny Mathew and Jorge Da Rocha. Thank you for all your kind words of encouragement and willingness to lend an ear.

- Dr. Julia Ponomarenko, Sarah Bonnin, Dr. Luca Cozzuto, Anna Vlasova and Paolo Di Tommaso from the Centre for Genomic Regulation (CRG) for hosting me and helping me look at my research from a different perspective.

- Dr. Thandiswa Ngcungcu for kindly providing her RNA-seq data on albinism for testing the workflows.

- The University of the Witwatersrand and the National Research Foundation (NRF) Thuthuka Funding Instrument (grant no. 99206) for providing the funding which made it possible to complete this work.

Abstract

The rate of raw sequence production through Next-Generation Sequencing (NGS) has been growing exponentially due to improved technology and reduced costs. This has enabled researchers to answer many biological questions through "multi-omics" data analyses. Even though such data promises new insights into how biological systems function and understanding many disease mechanisms, computational analyses performed on such large datasets comes with its pitfalls. In many cases, analyses of raw sequencing data can be computationally intensive and involves a combination of many bioinformatic applications, which often require different file formats in-between the analyses. Bioinformatic and computational pipelines can overcome these issues and the tedious repetitive tasks associated with the analyses of sequencing data, facilitate reproducibility of results and sharing of workflows for common analyses.

The aim of this study was to develop robust portable and reproducible bioinformatic pipelines for the automation of RNA sequencing (RNA-seq) data analyses. Using `Nextflow` as a workflow management system and `Singularity` for application containerisation, two bioinformatic workflows have been developed: `rnaSeqCount` (https://github.com/phelelani/nf-rnaSeqCount) for mapping raw RNA-seq reads to a reference genome and quantifying abundance of identified genomic features for differential gene expression analyses, and `rnaSeqMetagen` (https://github.com/phelelani/nf-rnaSeqMetagen) for performing metagenomic analyses on RNA-seq data. The RNA-seq data of black South African patients affected with systemic sclerosis (SSc) and unaffected individuals from the study by Frost *et al.* (2018) was used to illustrate the value of the workflows. SSc is a rare autoimmune disorder in which abnormalities in the vascular and immune systems result in the fibrosis of the connective tissue, skin and internal organs.

The RNA-seq data validated the usefulness of the workflows and provided biological insights into SSc in black South African populations through differential gene expression, pathway and metagenomic analyses. A number of genes were down-regulated in the affected skin of SSc patients and supported findings from other studies. These genes play potential roles in the identified down-regulated pathways associated with SSc, including "toll-like receptor" and "chemokine signaling" pathways. Metagenomic analyses revealed taxonomic classification of the *de novo* assembled unmapped reads, where more than one species belonging to *Arthrobacter*, *Bacillus*, *Brachybacterium*, *Dietzia* and *Pseudarthrobacter* genera were present in the SSc patients but not in the unaffected individuals. Bioinformatic and computational pipelines for RNA-seq data analysis, from QC to sequence alignment and comparative analyses, will reduce analysis time, and increase accuracy and reproducibility of findings to promote transcriptome and meta-analysis research.

Research Outputs

Conference proceedings

Presentation Details	Conference details
Phelelani T. Mpangase, Scott Hazelhurst, Michele Ramsay **Transcriptome analysis and genetic variation in systemic sclerosis (Poster)**	Research Day (1 September 2016) *University of the Witwatersrand, Johannesburg, South Africa*
Phelelani T. Mpangase, Scott Hazelhurst, Michele Ramsay **Transcriptome analysis and genetic variation in systemic sclerosis (Poster)**	SASBi/SAGS Conference 2016 (20 - 23 September 2016) *Nelson Mandela School of Medicine, Durban, South Africa*
Shakuntala Baichoo, Yassine Souilmi, Sumir Panji, Gerrit Botha, Ayton Meintjes, Scott Hazelhurst, Hocine Bendou, Eugene de Beste, Phelelani T. Mpangase, Oussema Souiai, Mustafa Alghali, Long Yi, Brian D. OConnor, Michael Crusoe, Don Armstrong, Shaun Aron, Fourie Joubert, Azza E. Ahmed, Mamana Mbiyavanga, Peter van Heusden, Lerato E. Magosi, Jennie Zermeno, Liudmila Sergeevna Mainzer, Faisal M. Fadlelmola, C. Victor Jongeneel and Nicola Mulder **Developing reproducible and portable bioinformatics workflows for African genomics research (Poster)**	26th ISMB 2018 Conference (6 - 10 July 2018) *Chicago, United States of America*
Phelelani T. Mpangase, Scott Hazelhurst, Michele Ramsay **rnaSeqMetagen: A portable and reproducible Nextflow pipeline for Metagenomic analysis of high throughput RNA-seq data (Oral)**	2nd SASBi-SC Student Symposium 2018 (15 October 2018) *Golden gate, Free State. South Africa*
Phelelani T. Mpangase, Scott Hazelhurst, Michele Ramsay **rnaSeqMetagen: A portable and reproducible Nextflow pipeline for Metagenomic analysis of high throughput RNA-seq data (Poster)**	2018 SASBi/SAGS Conference (16 - 18 October 2018) *Golden gate, Free State. South Africa*

Publications

The publications listed below are related to the work presented in this thesis and were produced during the course of this study:

Ahmed, A. E., Mpangase, P. T., Panji, S., Baichoo, S., Botha, G., Fadlelmola, F. M., Hazelhurst, S., Van Heusden, P., Jongeneel, C. V., Joubert, F., Mainzer, L. S., Meintjes, A., Armstrong, D., Crusoe, M. R., O'connor, B. D., Souilmi, Y., Alghali, M., Aron, S., Bendou, H., De Beste, E., Mbiyavanga, M., Souiai, O., Yi, L., Zermeno, J. and Mulder, N. (2018) Organizing and running bioinformatics hackathons within Africa: The H3ABioNet cloud computing experience *AAS Open Research* **1**, 9 ISSN 2515-9321.

Baichoo, S., Souilmi, Y., Panji, S., Botha, G., Meintjes, A., Hazelhurst, S., Bendou, H., de Beste, E., Mpangase, P. T., Souiai, O., Alghali, M., Yi, L., O'Connor, B. D., Crusoe, M., Armstrong, D., Aron, S., Joubert, F., Ahmed, A. E., Mbiyavanga, M., van Heusden, P., Magosi, L. E., Zermeno, J., Mainzer, L. S., Fadlelmola, F. M., Jongeneel, C. V. and Mulder, N. (2018) Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics *BMC Bioinformatics* **19**, 1, 457 ISSN 1471-2105.

Table of Contents

Declaration i
Dedication ii
Acknowledgements iii
Abstract iv
Research Outputs v
Conference proceedings v
Publications v
Table of Contents vi
List of Figures vii
List of Tables viii
List of Abbreviations xiii

Chapter 1: Introduction and Literature Review **1**
1.1 Reproducible computational workflows 1
1.1.1 Workflow management systems 3
1.1.2 Software containerisation & portability 5
1.1.3 Workflow scaling 7
1.1.4 Workflow documentation & sharing 9
1.2 Introduction to Systemic Sclerosis 10
1.3 Classification of SSc 11
1.4 Autoantibodies in SSc 13
1.4.1 Anti-centromere antibodies 14
1.4.2 Anti-topoisomerase I antibodies 15
1.4.3 Anti-RNA polymerase I, II and III antibodies 16
1.4.4 Anti-fibrillarin antibodies 16
1.4.5 Anti-Th/To antibodies 17
1.5 Pathology, aetiology and pathogenesis of SSc 17
1.5.1 Vascular dysregulation 18
1.5.1.1 Microcirculation impairment 18
1.5.1.2 Endothelial dysfunction 19
1.5.1.3 Impaired angiogenesis and vasculogenesis 20
1.5.1.4 Platelet activation 20
1.5.2 Immune activation 20
1.5.2.1 Innate immunity in SSc 21
1.5.2.2 Interferon signature in SSc 23

1.5.2.3 Adaptive immunity in SSc . . . 25
1.5.3 Fibrosis . . . 25
1.5.4 Pathogenic mechanism of SSc . . . 26
1.6 Environmental risk factors in SSc . . . 27
1.6.1 Crystalline silica . . . 27
1.6.2 Pathogens and infectious agents . . . 28
1.7 Genetic risk factors in SSc . . . 28
1.7.1 Twin Studies . . . 29
1.7.2 Candidate gene studies . . . 30
1.7.3 Genome wide association studies . . . 30
1.7.4 Gene expression studies . . . 30
1.8 SSc in African Populations . . . 31
1.9 Rationale and Motivation for the Study . . . 32
1.10 Study Aims and Objectives . . . 32
1.11 Limitations of this study . . . 33

Chapter 2: RNA-seq Data and Pre-processing 35
2.1 Introduction . . . 35
2.1.1 Quality control of RNA-seq data . . . 37
2.2 Analyses . . . 38
2.2.1 Initial QC using FastQC . . . 38
2.2.2 Removal of artefacts using Trimmomatic . . . 39
2.2.3 Merging of duplicated/resequenced samples . . . 40
2.3 Results . . . 40
2.4 Discussion . . . 40

Chapter 3: Developing a Pipeline for Gene/Transcript Identification and Quantification 42
3.1 Introduction . . . 42
3.1.1 Gene/Transcript identification . . . 43
3.1.1.1 Mapping to reference genome . . . 43
3.1.1.2 Mapping to reference transcriptome . . . 44
3.1.1.3 Transcriptome reconstruction . . . 45
3.1.2 Gene/Transcript quantification . . . 47
3.1.2.1 Raw read counts . . . 47
3.1.2.2 Normalised read counts . . . 48
3.1.2.3 Raw versus normalised read counts . . . 48
3.2 Analyses . . . 49
3.2.1 Mapping to reference genome . . . 49
3.2.2 Read assignment to genomic features . . . 50
3.2.3 Read-count matrices and QC . . . 51

3.2.4 Pipeline for gene-level feature counting of RNA-seq data 51
3.2.4.1 Implementing the workflow in `Nextflow` 52
3.2.4.2 `Singularity` images for applications 53
3.2.4.3 GitHub repository for the pipeline 53
3.3 Results 53
3.3.1 `rnaSeqCount`: A portable and reproducible `Nextflow` pipeline for gene-level feature counting of RNA-seq data 54
3.3.1.1 Obtaining the pipeline 54
3.3.1.2 Obtaining `Singularity` images and indexing the reference genome 55
3.3.1.3 Executing the `rnaSeqCount` workflow 56
3.3.1.4 Results produced by the `rnaSeqCount` pipeline 56
3.3.2 Using `rnaSeqCount` to generate raw read counts for the SSc data . 57
3.3.2.1 CPU usage 57
3.3.2.2 Memory usage 58
3.3.2.3 Execution time 59
3.3.2.4 QC plots produced by `MultiQC` 59
3.4 Discussion 61

Chapter 4: Differential Expression Analysis **64**
4.1 Introduction 64
4.1.1 Differential expression using RNA-seq data 65
4.1.1.1 Normalisation methods: RPKM, FPKM and TMM 65
4.1.1.2 Statistical modelling of gene expression 66
4.1.1.3 Testing for differential gene expression 66
4.1.2 Enrichment of genes 67
4.1.3 Pathway analyses 68
4.2 Analyses 68
4.2.1 Pre-processing of read-count data 69
4.2.1.1 Power and sample size analyses 71
4.2.1.2 Preparing comparison sets for expression analysis 72
4.2.1.3 Assessing sample-sample similarities 73
4.2.2 Running differential expression analysis with `DESeq2` 73
4.2.3 Prioritisation of significant genes with `UpSet` 74
4.2.4 Gene-set enrichment with `enrichR` 76
4.2.5 Pathway analysis with `gage` and `pathview` 76
4.3 Results 76
4.3.1 Power and sample size analysis 76
4.3.2 Data pre-processing and differential expression analysis 78
4.3.3 Sample-sample similarities 78
4.3.4 Visualisation of differential expression results 80

4.3.5 Visualisation and clustering of significant genes . . . 81
4.3.6 Prioritisation . . . 86
4.3.7 Gene-set enrichment with `enrichR` . . . 86
4.3.8 Pathway analysis with `gage` and `pathview` . . . 90
4.4 Discussion . . . 94
4.4.1 Down-regulated genes in black South African patients with SSc . . . 95

Chapter 5: Developing a Pipeline for Metagenomics Analysis 99
5.1 Introduction . . . 99
5.1.1 Host-read filtering . . . 100
5.1.2 Classifying exogenous reads . . . 100
5.2 Analyses . . . 101
5.2.1 Pipeline for metagenomic analysis of RNA-seq data . . . 102
5.2.1.1 Implementing the workflow in `Nextflow` . . . 102
5.2.1.2 Singularity images for applications . . . 103
5.2.1.3 GitHub repository for the pipeline . . . 104
5.3 Results . . . 104
5.3.1 `rnaSeqMetagen`: A portable and reproducible `Nextflow` pipeline for metagenomic analysis of RNA-seq data . . . 104
5.3.1.1 Obtaining the pipeline . . . 104
5.3.1.2 Obtaining Singularity images and generating the Kraken2 database . . . 106
5.3.1.3 Executing the `rnaSeqMetagen` workflow . . . 107
5.3.1.4 Results produced by the `rnaSeqMetagen` pipeline . . . 107
5.3.2 Using `rnaSeqMetagen` for metagenomic analysis of the SSc data . . . 108
5.3.2.1 CPU usage . . . 108
5.3.2.2 Memory usage . . . 108
5.3.2.3 Execution time . . . 109
5.3.2.4 QC plots produced by `rnaSeqMetagen` . . . 109
5.3.3 Exploring possible pathogens in SSc patients using `UpSet` and `krona` charts . . . 110
5.4 Discussion . . . 112
5.4.1 Clinical relevance of identified genera . . . 115

Chapter 6: Concluding Discussion 117
6.1 Reproducible workflows for RNA-seq data analyses . . . 118
6.2 Differential expression & pathway analysis . . . 118
6.3 Metagenomic analysis . . . 119
6.4 Study limitations and future work . . . 119
6.5 Conclusion . . . 120

References 121

Appendices 134

Appendix A: Ethics clearance certificate 135

Appendix B: QC plots for RNA-seq Data Pre-processing 136

Appendix C: `Singularity` Recipes: `rnaSeqCount` Workflow 137
C.1 `STAR` container . . . 137
C.2 `htseq-count` container . . . 137
C.3 `featureCounts` container . . . 137
C.4 `MultiQC` container . . . 138

Appendix D: Genes Discarded in the Differential Expression Analysis 139

Appendix E: Differential Expression Analysis - PCA Plots 140

Appendix F: Differential Expression Analysis - Volcano Plots 141

Appendix G: Differential Expression Analysis - Heatmaps 142

Appendix H: Prioritised Genes 145

Appendix I: Significantly Differentially Expression Genes 148

Appendix J: Pathways Identified using `gage` 156

Appendix K: Pathways Constructed with `pathview` 161

Appendix L: `R` Session Information and Packages 166

Appendix M: Singularity recipes: `rnaSeqMetagen` workflow 169
M.1 `trinity` container . . . 169
M.2 `kraken` container . . . 170
M.3 `upset` container . . . 171

Appendix N: Microbial Taxonomies Identified using `UpSet` 172

List of Figures

1.1 Challenges and solutions to developing reproducible workflows 3
1.2 Comparison of the native and dockerised versions of the pipelines on different computational platforms. 5
1.3 Comparison of the virtual machine and container architecture 6
1.4 Standard `Singularity` workflow for building production images 8
1.5 Summary of resources and best practices for reproducible workflows 10
1.6 Proposed pathogenesis of SSc 18
1.7 TLR signalling through MyD88 and TRIF adaptor proteins 24
1.8 Proposed pathogenic mechanism of SSc 27
1.9 Flow diagram summarising the overall structure of the thesis 33

3.1 An overview of `kallisto`'s "pseudoalignment" algorithm 45
3.2 Two main approaches to transcriptome reconstruction 46
3.3 Overall summary of the `rnaSeqCount` workflow 55
3.4 Summary report and metadata for `rnaSeqCount` pipeline execution 58
3.5 % CPU usage by each process of the `rnaSeqCount` pipeline 58
3.6 % Memory usage by each process of the `rnaSeqCount` pipeline 59
3.7 % Requested time usage by each process of the `rnaSeqCount`pipeline . . . 59
3.8 QC summary of the mapping and quantification of the `rnaSeqCount` pipeline produced by `MultiQC` 60
3.9 Summary report and metadata for `rnaSeqCount` pipeline execution on AWS 62

4.1 `UpSet` queries for gene prioritisation 75
4.2 Power calculations using `PROPER` for the different sample sizes 77
4.3 PCA of the comparison sets with the affected forearms included 79
4.4 Volcano plots for the differential expression results and filtering for comparisons sets with forearms included 80
4.5 Heatmap showing gene clustering according to gene expression signals in the ALL comparison 82
4.6 Heatmap showing gene clustering according to gene expression signals in the CASE.ARM comparison 83
4.7 Heatmap showing gene clustering according to gene expression signals in the SEVERE.ARM comparison 84
4.8 Heatmap showing gene clustering according to gene expression signals in the MILD.ARM comparison 85
4.9 `UpSetR` plot for gene prioritisation 87

4.10 Significance plots summarising gene-set enrichment of the 77 prioritised genes 89
4.11 Toll-like receptor pathway (CASE.ARM) 93
4.12 Chemokine signaling pathway (CASE.ARM) 93
4.13 Wnt signaling pathway (ALL) 94

5.1 Overall summary of the `rnaSeqMetagen` workflow 105
5.2 Summary report and metadata for `rnaSeqMetagen` pipeline execution . . . 108
5.3 % CPU usage by each process of the `rnaSeqMetagen` pipeline 109
5.4 % Memory usage by each process of the `rnaSeqMetagen` pipeline 109
5.5 % Requested time usage by each process of the `rnaSeqMetagen`pipeline . . 110
5.6 Alignment of reads to the reference genome by `STAR` in the `rnaSeqMetagen` pipeline 110
5.7 `UpSet` query for identifying microbial taxonomies shared between SSc patient's forearm and/or back samples 112
5.8 `krona` charts for the *Arthrobacter* genus classification 114

B.1 QC plots for the RNA-seq data used in this study 136

E.1 PCA of the within individual comparison sets and affected backs included . 140

F.1 Volcano plots for the differential expression results and filtering for comparisons sets with backs included 141

G.1 Heatmap showing gene clustering according to gene expression signals in the CASE.BACK comparison 142
G.2 Heatmap showing gene clustering according to gene expression signals in the CASE comparison 143
G.3 Heatmap showing gene clustering according to gene expression signals in the SEVERE.BACK comparison 143
G.4 Heatmap showing gene clustering according to gene expression signals in the MILD.BACK comparison 144
G.5 Heatmap showing gene clustering according to gene expression signals in the SEVERE.MILD comparison 144

K.1 Phagosome pathway (CASE.ARM) 161
K.2 Osteoclast differentiation pathway (CASE.ARM) 162
K.3 Natural killer cell mediated cytotoxicity pathway (CASE.ARM) 162
K.4 MAPK signaling pathway (ALL) 163
K.5 NOD-like receptor pathway (CASE.ARM) 164
K.6 Apoptosis pathway (CASE.ARM) 165

List of Tables

1.1 Classification of scleroderma-related disorders . . . 12
1.2 Comparison of the scleroderma-related disorders . . . 13
1.3 SSc anti-nuclear antibody frequency and clinical correlations . . . 15
1.4 TLRs implicated in SSc and their respective ligands and origins . . . 22

2.1 Summary of the RNA-seq sequencing and clinical data . . . 36
2.2 Summary of the RNA-seq data before and after pre-processing . . . 41

4.1 Different sets of comparisons carried out in this study . . . 71
4.2 Experimental design formulas used in the nine expression comparisons . . . 72
4.3 Genes returned by different $\log_2$FC cutoff using a fixed FDR of 0.01 . . . 74
4.4 Summary of the power calculations performed with `PROPER` . . . 77
4.5 Summary of the pre-processing steps carried out . . . 79
4.6 Summary of pathways identified by `gage` and constructed using `pathview` . 91
4.7 Down-regulated genes associated with SSc in the Open Targets Platform . 96

D.1 Genes excluded from differential expression analysis . . . 139

H.1 List of genes prioritised using `UpSet` . . . 145

I.1 Differentially expressed genes in the ALL comparison set. . . . 148
I.2 Differentially expressed genes in the CASE.ARM comparison set. . . . 149
I.3 Differentially expressed genes in the CASE.BACK comparison set. . . . 150
I.4 Differentially expressed genes in the CASE comparison set. . . . 151
I.5 Differentially expressed genes in the SEVERE.ARM comparison set. . . . 152
I.6 Differentially expressed genes in the SEVERE.BACK comparison set. . . . 153
I.7 Differentially expressed genes in the MILD.ARM comparison set. . . . 153
I.8 Differentially expressed genes in the MILD.BACK comparison set. . . . 155
I.9 Differentially expressed genes in the SEVERE.MILD comparison set. . . . 155

J.1 Down-regulated pathways in the ALL comparison identified by `gage` . . . 156
J.2 Up-regulated pathways in the CASE.ARM comparison identified by `gage` . 156
J.3 Down-regulated pathways in the CASE.ARM comparison identified by `gage` 157
J.4 Down-regulated pathways in the CASE.BACK comparison identified by `gage` 157
J.5 Up-regulated pathways in the CASE comparison identified by `gage` . . . 158
J.6 Down-regulated pathways in the SEVERE.ARM comparison identified by `gage` . . . 158

J.7 Up-regulated pathways in the MILD.ARM comparison identified by gage . 159
J.8 Down-regulated pathways in the MILD.ARM comparison identified by gage 159
J.9 Down-regulated pathways in the MILD.BACK comparison identified by gage 159
J.10 Up-regulated pathways in the SEVERE.MILD comparison identified by gage 160

N.1 Microbial taxonomies shared between the SSc patients only 172

List of Abbreviations

ACA	Anti-centromere autoantibodies
ALBIA	Addressable laser bead immunoassays
AMI	Amazon Machine Image
ANA	Anti-nuclear antibodies
AP-1	Activation protein 1
AQP9	Aquaporin 9
ATA	Anti-topoisomerase autoantibodies
AWS	Amazon Web Services
BAM	Binary Alignment Map
BCL2	B-cell lymphoma 2
BCL2A1	BCL2 related protein A1
CCL2	C-C motif chemokine ligand 2
CCL20	C-C motif chemokine ligand 20
CCN2	Connective tissue growth factor
CD69	CD69 molecule
cDNA	Complementary DNA
CENP	Centromeric proteins
CKLF	Chemokine-like factor
CLI	Command line interface
CMTM2	CKLF like MARVEL trans-membrane domain containing 2
CMV	Cytomegalovirus
CPU	Central processing unit
CREST	Calcinosis, Raynauds phenomenon, esophageal dysmotility, sclerodactyly, telangectasia
CRG	Center for Genomic Regulation
CSF3	Colony stimulating factor 3
CTD	Connective tissue disease
CTGF	Connective tissue growth factor
CXCL1	CXC motif chemokine ligand 1
CXCL3	CXC motif chemokine ligand 3
CXCL5	CXC motif chemokine ligand 5
CXCL6	CXC motif chemokine ligand 6
CXCL8	C-X-C motif chemokine ligand 8

CXCR1	CXC motif chemokine receptor 1
CXCR2	CXC motif chemokine receptor 2
DAMP	Damage-associated molecular patterns
DID	Double immunodiffusion
DSL	Domain-specific language
dsRNA	Double-stranded RNA
dsSSc	Diffuse cutaneous SSc
DZ	Dizygotic twins
EBV	EpsteinBarr virus
EC2	Elastic Compute Cloud
ECDS	En coupe de sabre
ECM	Extra-cellular matrix
ELISA	Enzyme linked immunoassay
ELR	Glu-Leu-Arg
ENA	Epithelial-derived neutrophil-activating peptide
ET-1	Endothelin-1
FCGR3B	Fc fragment of IgG receptor IIIb
FD	False discoveries
FDC	False discovery cost
FDR	False discovery rate
FFAR2	Free fatty acid receptor 2
FOS	Fos protooncogene, AP1 transcription factor subunit
FPKM	Fragments per kilobase of transcript per million reads
GCP	granulocyte chemotactic protein 2
GFF	General Feature Format
GLM	Generalised linear model
GRO	Growth-regulated oncogene
GUI	Graphical user interface
GWAS	Genome-wide association studies
HBA1	Hemoglobin subunit alpha 1
HLA	Human leukocyte antigen
HMGB-1	High mobility group box 1
HMP	Human Microbiome Project
HPC	High-performance computing
HREC	Human Research Ethics Committee
HSC	Hematopoietic stem cells
HSP	Heat shock protein

HTS	High-throughput sequencing
IB	Immunoblotting
IFN	Interferon
IgG	Immunoglobulin G
IIF	Indirect immunofluorescence
IL1A	Interleukin 1 α
IL1B	Interleukin 1 β
IP	Immunoprecipitation
IRF	Interferon response factor
LCA	Lowest common ancestor
lcSSc	Limited cutaneous SSc
LFC	Log fold change
LIA	Line immunoassay
LPS	Lipopolysaccharide
LS	Linear scleroderma
LSF	Platform Load Sharing Facility
LUCAT1	Lung cancer associated transcript 1
MAMP	Microbe-associated molecular patterns
MAPK	Mitogen-activated protein kinase
MCP-1	Monocyte chemoattractant protein-1
MCTD	Mixed connective tissue disease
MHC	Major histocompatibility complex
miRNA	microRNA
MMP	Maximal mappable prefix
MMP1	Matrix metallopeptidase 1
MMPs	Matrix metalloproteinases
mRNA	Messenger RNA
MRP	Mitochondrial RNA processing
MRSS	Modified Rodnan skin score LoS
MZ	Monozygotic twins
NF-κB	Nuclear factor κB
NGS	Next generation sequencing
NK	Natural killer cells
NO	Nitric oxide
NOD	Nucleotide-binding and oligomerisation domain
OS	Operating system
PAH	Pulmonary arterial hypertension

PAMP	Pathogen-associated molecular patterns
PDGF	Platelet-derived growth factor
PDGF-R	Platelet-derived growth factor
PF4	Platelet factor 4
PHTN	Pulmonary arterial hypertension
PI3	Peptidase inhibitor 3
PPBP	Pro-platelet basic protein
PROK2	Prokineticin 2
PRR	Pattern recognition receptor
QC	Quality control
RA	Rheumatoid arthritis
RIN	RNA integrity number
RNA-seq	RNA sequencing
RNAP	RNA polymerase
RNAse	Ribonuclease
RNAse MRP	Ribonuclease for mitochondrial RNA processing
ROS	Reactive oxygen species
RP	Raynauds phenomenon
RPKM	Reads per kilobase of transcript per million reads
rRNA	Ribosomal RNA
S100A12	S100 calcium binding protein A12
S100A8	S100 calcium binding protein A8
S100A9	S100 calcium binding protein A9
SA	Suffix array
SAM	Sequence Alignment Map
Scl-70	Topoisomerase 1
SGE	Sun Grid Engine
siRNA	Small interfering RNA
SLE	Systemic lupus erythematosus
SLURM	Simple Linux Utility for Resource Management
SMAD7	Mothers against decapentaplegic homolog 7
SNP	Single nucleotide polymorphism
SRC	Scleroderma renal crisis
SS	Sample size
SSc	Systemic Sclerosis
ssRNA	Single-stranded RNA
STAR	Spliced transcripts alignment to a reference

STAT4	Signal transducer and activator of transcription 4
T-DBG	Transcriptome de Bruijn graph
TD	True discoveries
TGF-β	Transforming growth factor beta
TLR	Toll-like receptor
TNF	Tumour necrosis factor
topo 1	Topoisomerase 1
TREM1	Triggering receptor expressed on myeloid cells 1
TRIF	TIR domain-containing adaptor inducing IFN-β
U3RNP	U3-ribonucleoprotein
UCT	University of Cape Town
VCS	Version control systems
VEGF	Vascular endothelial growth factor
vWF	von Willebrand factor
Wits	University of the Witwatersrand

Chapter 1

Introduction and Literature Review

There is a gap in the availability of robust bioinformatic and computational methods that overcome the tedious repetitive tasks associated with the analyses of raw sequencing data that is being produced by researchers. This study aims to develop efficient computational workflows that aid with the automation of analyses and biological interpretation of sequencing data. To illustrate the utility of the workflows, this study uses RNA sequencing (RNA-seq) data of black South African patients affected with the systemic sclerosis (SSc) disease and unaffected individuals from the study by Frost *et al.* (2018) to demonstrate the value of the workflows produced in this study. The RNA-seq data used in this study does not only illustrate the value of the workflows, but was also used to seek biological insights into differential gene expression and metagenomic analysis of SSc in black South African populations, and to contribute towards the current understanding of the disease.

In this chapter, the background and literature review of the core aspects that form the basis of this study will be presented. Since this study focuses on designing computational pipelines for analysing transcriptome data from patients affected with SSc, the key concepts of designing computational pipelines (Section 1.1) will first be discussed. The introduction to SSc (Section 1.2) will then be presented, followed by the an in-depth discussion of the work that has been done by other researchers in an effort to understand the disease in terms of classification (Section 1.3), the autoantibodies involved (Section 1.4) as well as the disease pathogenesis (Section 1.5). I will then discuss the studies that have been done to understand the environmental (Section 1.6) and genetic (Section 1.7) risk factors associated with SSc. Finally, I will discuss the current understanding of SSc in African populations (Section 1.8). The rationale and motivation of this study is presented in Section 1.9, the aims and objectives in Section 1.10, and finally the limitations of this study in Section 1.11.

1.1 Reproducible computational workflows

With the increase in the rate at which raw sequencing data is produced due to improved technology and reduced cost of Next-Generation Sequencing (NGS), researchers in the field of bioinformatics and computational biology are able to perform "multi-omics" data analyses to answer many biological questions (Fan *et al.*, 2014; Kluge and Friedel, 2018; Schulz *et al.*, 2016). However, analysis of such large datasets comes with a number of challenges, especially when it comes to sharing data analysis methods with the scientific community and being able to reproduce consistent results using the same data across

different computational platforms (Boettiger, 2015; Di Tommaso *et al.*, 2017; Kurtzer *et al.*, 2017).

When performing computational analyses of NGS data, often different tools are required at each processing step of the analysis. For example, Haas *et al.* (2013) describe a procedure for assembling RNA-seq data, quantifying expression levels for transcripts and identifying differentially expressed transcripts between samples. This protocol requires a number of applications in order to be executed successfully: `Trinity` version trinityrnaseq_r2013-02-25 (`http://trinityrnaseq.sourceforge.net`), `bowtie` version 0.12.9 (`http://bowtie-bio.sourceforge.net`), `samtools` version 0.1.18 (`http://sourceforge.net/projects/samtools/files/samtools/`), `R` version 2.15 (`http://www.r-project.org`), `NCBI-Blast+` version 2.2.27 (`ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/`) (Haas *et al.*, 2013).

To a bioinformaticist, computational biologist or someone who is familiar with the `UNIX` environment, installing these applications and running this protocol described by Haas *et al.* (2013) would be a straight-forward procedure. However, to a novice, this would be a difficult task. Not being an administrator also significantly complicates installation of applications. Another challenge in performing such a protocol procedure would be having to re-do the analysis, either multiple times whilst changing certain parameters, or performing the analysis using more than one dataset. In this case, simply executing the protocol commands on a command line interface (CLI) would not suffice. Custom scripts would have to be created in order to compile and order the multiple commands needed to execute the protocol procedure repeatedly or on multiple datasets (Piccolo and Frampton, 2016). Another option would be to implement "workflow management systems" to construct a "workflow" of the multiple analyses steps, handle input/output files between applications and also automate the analysis (Di Tommaso *et al.*, 2017; Piccolo and Frampton, 2016; Schulz *et al.*, 2016).

Another challenge that the scientific community faces in performing multi-step analysis that requires different pieces of software at each analysis step is software dependencies and libraries (Kurtzer *et al.*, 2017). Many bioinformatics tools are built from sources, and thus there will be a complexity of dependencies and libraries between the softwares needed to perform the analyses (Schulz *et al.*, 2016). In addition to software and library dependency, there is also a computational environment or an operating system (OS) dependency. Installation of different application softwares on different OSs requires different configuration steps, and some applications are only designed to be executed on a specific environment of a specific OS (Kurtzer *et al.*, 2017; Piccolo and Frampton, 2016). A solution to software and OS dependency is to use virtual machines or software package managers (containers) (Boettiger, 2015; Kurtzer *et al.*, 2017; Piccolo and Frampton, 2016; Schulz *et al.*, 2016).

When big datasets are being analysed, personal desktop machines and laptops are not an option. In most cases, bioinformatic analyses will require a significant amount of computing power, memory and will sometimes need to be processed simultaneously (in parallel) in order to reduce the amount of time needed to perform each task (Di Tommaso *et al.*, 2017; Kurtzer *et al.*, 2017). These analyses have to be perform on high-performance computing (HPC) clusters available in most research institutes or cloud-computing services which offer significantly high computing resources that can meet the requirements of intensive bioinformatic analyses (Kurtzer *et al.*, 2017). This "scaling up" of bioinformatic analyses to cloud environments and HPC clusters is further enhanced by a combination of workflow management system and containerisation of software; making bioinformatic analyses workflows "portable" across different computing platforms (Boettiger, 2015; Kurtzer *et al.*, 2017; Piccolo and Frampton, 2016; Schulz *et al.*, 2016).

This combination also overcomes the limitation of software installation on HPC clusters and cloud-services as sometimes the users do not have privileges to install softwares and their dependencies (Kurtzer *et al.*, 2017). Furthermore, coupling the combination workflow management systems, software containers and HPC with proper documentation and storing code using version control systems (VCS) creates portable workflows that can be shared amongst the scientific community and ensures reproducibility across different platforms (Di Tommaso *et al.*, 2017; Kurtzer *et al.*, 2017; Perkel, 2016; Piccolo and Frampton, 2016). Figure 1.1 summarises the challenges faced by the bioinformatic and computational biology community when analysing large datasets, and solutions to these challenges. In the sections that follow (Sections 1.1.1 to 1.1.4), good practices and techniques/tools that can be used to facilitate reproducibility of bioinformatic workflows are discussed in detail.

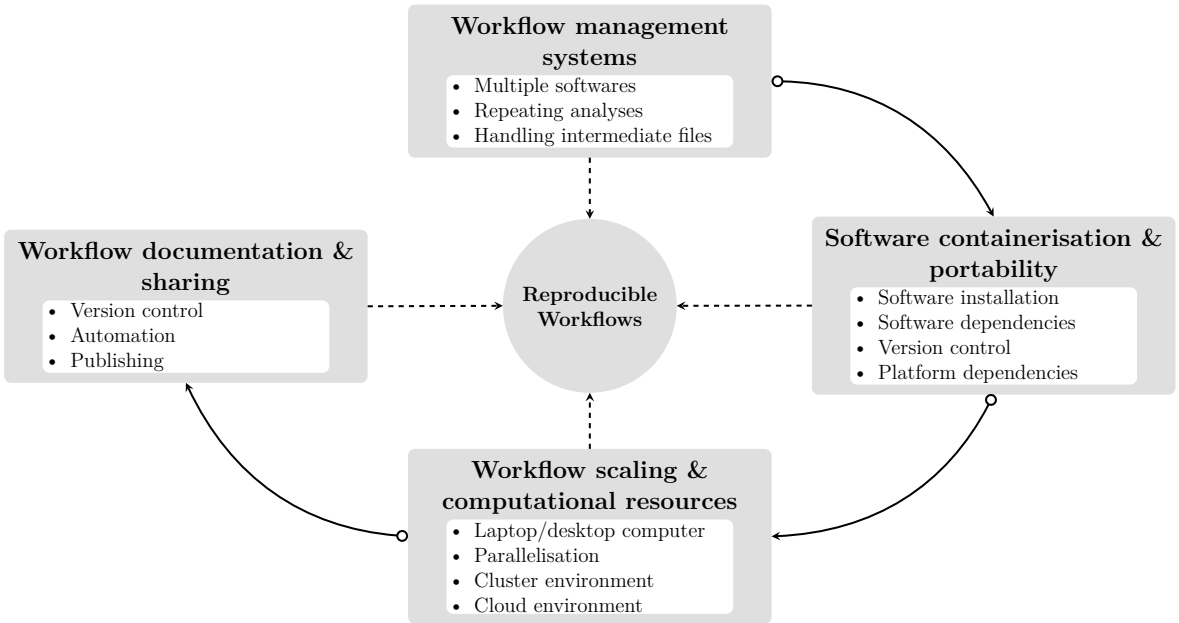


Figure 1.1: Flow diagram summarising the challenges and solutions to developing reproducible workflows.

1.1.1 Workflow management systems

When it comes to multi-step computational analyses of biological data on a large scale, workflow management systems are an essential component (Piccolo and Frampton, 2016; Schulz *et al.*, 2016). Amongst the number of workflow management systems available to the bioinformatics community, Galaxy (`https://galaxyproject.org/`) is perhaps one of the most popular and most used tools available, especially in the analysis of NGS data (Afgan *et al.*, 2016; Piccolo and Frampton, 2016). Galaxy is a web-based data analysis platform that allows users to conduct their computational analyses of NGS data without the use of the CLI, which is an advantage to novice bioinformatics users. The main focus of Galaxy is to enable data-driven research through making data analyses accessible, reproducible and communicable to other researchers (Afgan *et al.*, 2016). Galaxy users have a number of options for carrying out their analyses, i.e., using the free public server (`http://usegalaxy.org/`) that can be accessed by anyone, installing Galaxy locally or on a cloud-based infrastructure, or using the Galaxy servers installed in some institutions.

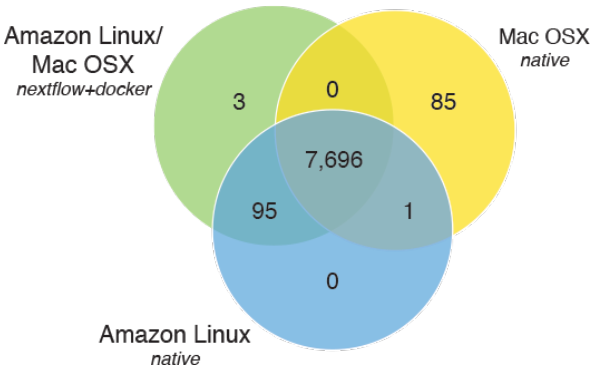
Another workflow management tool that is gaining significant popularity in the bioinformatics community is the `Nextflow` (`https://www.nextflow.io/`) tool developed by Di Tommaso *et al.* (2017). `Nextflow` is a Groovy-based domain-specific language (DSL) specifically designed for bioinformaticists with a vast programming knowledge to solve many of the challenges of the inability to reproduce data analysis. Some of these challenges are due to computational platform variations, software and database management, complexity of pipelines, intermediate file handling and lack of good practice (Di Tommaso *et al.*, 2017). Unlike Galaxy, `Nextflow` does not use a graphical user interface (GUI), and thus reduces the burden of having to re-implement third-party pipelines and tools using the GUI; `Nextflow` can easily be used to adapt different pipelines in any scripting language (Di Tommaso *et al.*, 2017).

`Nextflow` has a number of features that promote workflow reproducibility and portability. It supports `Docker` (`http://docker.io/`) and `Singularity` (`https://www.sylabs.io/`), the two most used containerisation softwares in the bioinformatics community; it integrates/supports the popular VCS GitHub (`http://github.com/`) for sharing of code, and version management; and it allows for scaling of computational workflows on HPC and cloud systems by providing out of the box scheduler support for Sun Grid Engine (SGE), PBS/Torque, Platform Load Sharing Facility (LSF), Simple Linux Utility for Resource Management (SLURM), HTCondor and Amazon Web Services (AWS) (Di Tommaso *et al.*, 2017).

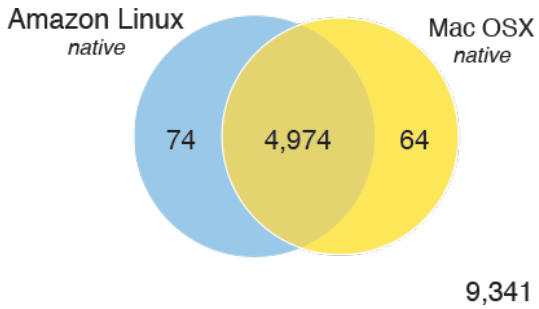
To demonstrate the reproducibility and portability of `Nextflow`, Di Tommaso *et al.* (2017) performed two studies: (1) to predict genome annotations for the *Leishmania infantum* using the Sanger Companion pipeline (Steinbiss *et al.*, 2016); and (2) to perform gene expression quantification and differential expression using `kallisto` (`http://pachterlab.`

github.io/kallisto/) and `sleuth` (https://pachterlab.github.io/sleuth/) (Bray *et al.*, 2016). In both these studies, Di Tommaso *et al.* (2017) performed the analyses on Amazon Linux and Mac OSX environments, without (native version of the pipeline) or with the combination of `Nextflow` and `Docker` (dockerised version of the pipeline). The results (Figure 1.2) of these analyses showed that when using `Docker`, `Nextflow` is able to completely remove computational variations caused by the instability of OSs, thus making workflows completely reproducible (Di Tommaso *et al.*, 2017). In both workflows, there were no differences in the results of the dockerised version of the pipelines (Figure 1.2a - green; and Figure 1.2c). However, differences were observed in the native versions of the pipelines executed on the different computing environments (Figure 1.2a - yellow/blue; and Figure 1.2b).

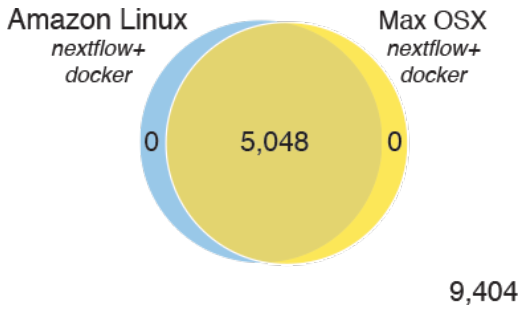
The metagenomics pipeline, YAMP (Yet Another Metagenomics Pipeline), developed by Visconti *et al.* (2018) for processing raw shotgun sequencing data also demonstrates the reproducibility and portability of `Nextflow`. YAMP is a ready-to-use workflow implemented in `Nextflow`, and uses `Singularity` and `Docker` containers for its analyses tools. YAMP facilitates repeatability and reproducibility through the integration of BitBucket (https://bitbucket.org/), GitHub (https://github.com/) and GitLab (https://about.gitlab.com/) code repositories, which allow for tracking of both code



(a) Gene annotation prediction using the dockerised and native versions of the Sanger Companion pipeline.



(b) Quantification of transcripts and differential expression using `kallisto` and `sleuth` using the native version of the pipeline.



(c) Quantification of transcripts and differential expression using `kallisto` and `sleuth` deployed with dockerised version of the pipeline.

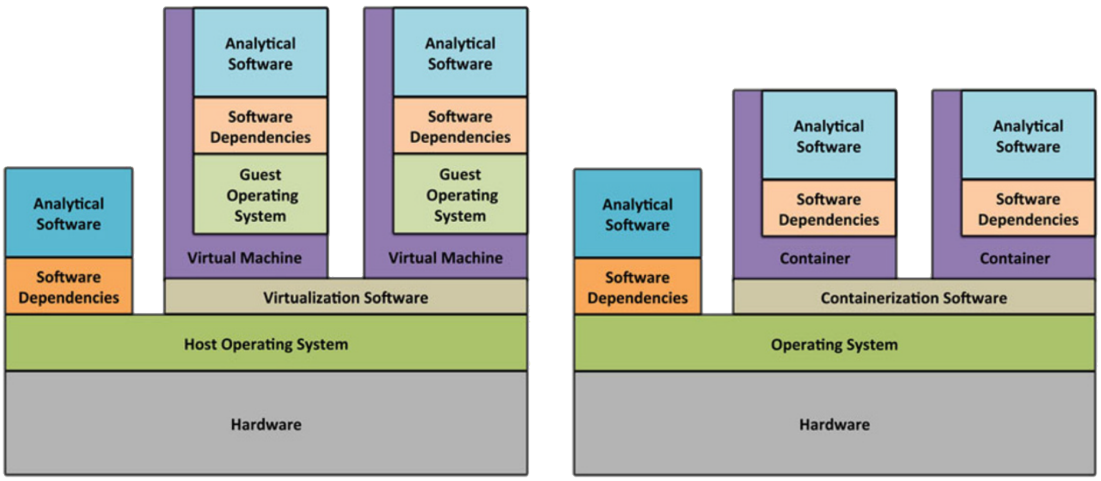
Figure 1.2: Comparison of native and dockerised version of the pipelines on different computational platforms. *Note:* Adapted from Di Tommaso *et al.* (2017)

and software versions. Visconti *et al.* (2018) evaluated their workflow using two simulated datasets as well as 18 real-world cases of randomly selected samples from the phase III of the Human Microbiome Project (HMP). They found that their workflow, YAMP, was able to reproduce results that are consistent with the original findings, illustrating that `Nextflow` workflows coupled with software containers (`Singularity`/`Docker`) are powerful solutions for computational portability and reproducibility, as well as facilitating collaborative work.

1.1.2 Software containerisation & portability

Before containerisation software came into light, to overcome the problem of software and OS dependencies, virtual machines (like the popular Oracle's VirtualBox [`https://www.virtualbox.org/`]) were used. These "guest" OSs would be installed on "host" machines and encapsulated the required software, libraries and data needed to perform certain analysis (Kurtzer *et al.*, 2017). However, to emulate an OS requires significantly high computational resources, which has a great impact on the performance of the host machine (Figure 1.3a). There is also a challenge of sharing virtual machines, since they are typically large in size (Kurtzer *et al.*, 2017). Lightweight virtualisation softwares were required to overcome these problems. `Docker` (`https://www.docker.com/`) was one of the solutions (Boettiger, 2015).

`Docker` provides a lightweight platform for building, testing and sharing Linux-based software containers by specifying instructions inside a "`Docker`" file, which can be shared amongst researchers for building an image with the specified softwares (Boettiger, 2015; Piccolo and Frampton, 2016). These lightweight containers could be used on host machines without the computational cost that virtual machines have (Figure 1.3b). `Docker` shares/utilises the same Linux kernel as the host, thus their performance is far supe-



(a) Virtual machine architecture.

(b) Software container architecture.

Figure 1.3: Comparison of the virtual machine and container architecture. (a) Virtual machines ("guest") runs alongside the main ("host") OS. Encapsulated software is executed in the virtual environment, independent of the host OS. (b) Encapsulates software is executed in the context of the main OS. *Note:* Adapted from Piccolo and Frampton (2016)

rior than that of the virtual machines. More than one `Docker` image can be executed on a single host machine, with each image addressing a specific software of the analysis (Boettiger, 2015; Piccolo and Frampton, 2016). However, even though `Docker` provides solutions to software dependencies and portability, it is associated with high security risks on HPC environments (Kurtzer *et al.*, 2017). When `Docker` executes an image, the processes inside the image are executed as subprocesses of a root owned by `Docker`; thus it is possible to "fool" `Docker` into granting a user privileges to access the host file system (Kurtzer *et al.*, 2017). Another containerisation platform similar to `Docker` that may be used as an alternative is `Singularity` (`https://www.sylabs.io/`).

Kurtzer *et al.* (2017) define the term "mobility of compute" as an essential ingredient in designing reproducible workflows. *"Mobility of compute is defined as the ability to define, create, and maintain a workflow locally while remaining confident that the workflow can be executed on different hosts, Linux operating systems, and/or cloud service providers."* (Kurtzer *et al.*, 2017). This basically means that for workflows to be portable, the workflow itself (scripts), its components (required applications and libraries) and base OS/environment must be developed in such a way that they can be executed reliably on any Linux-based environment without any issues.

`Singularity` achieves this by using an image format that is supported across different versions of the C library and kernels, and gives users the ability to encapsulate their workflow, all required applications, dependencies and base OS environment into a single file that can be locked, copied, shared and archived (Kurtzer *et al.*, 2017). Such image files have standard Linux/UNIX file permission and cannot be modified (not even by the host OS), thus can be used with confidence that nothing within the image has changed. If changes have to be made to the container, the image has to be rebuilt from scratch with the necessary changes before it can be shared. Figure 1.4 summarises the overall workflow steps for building `Singularity` images. Unlike `Docker`, `Singularity` overcomes the security risks by giving users inside the container (at runtime) the same permissions they have on the host environment (Kurtzer *et al.*, 2017). This means that if a user wants to execute applications inside a `Singularity` image as root, they must first execute the image as root on the host environment. For these reasons, `Singularity` is quickly gaining popularity in the bioinformatics community, especially for performing analysis on a large scale.

1.1.3 Workflow scaling

One of the many challenges in performing analysis on the massive data produced by NGS platforms is computational power. Typical servers with a single multi-core central processing unit (CPU) cannot offer the amount of computational capacity that is needed to process and store such large amounts of data; scalable technologies/infrastructure that can be used to perform analysis in parallel and on a large scale are required (Muir *et al.*,

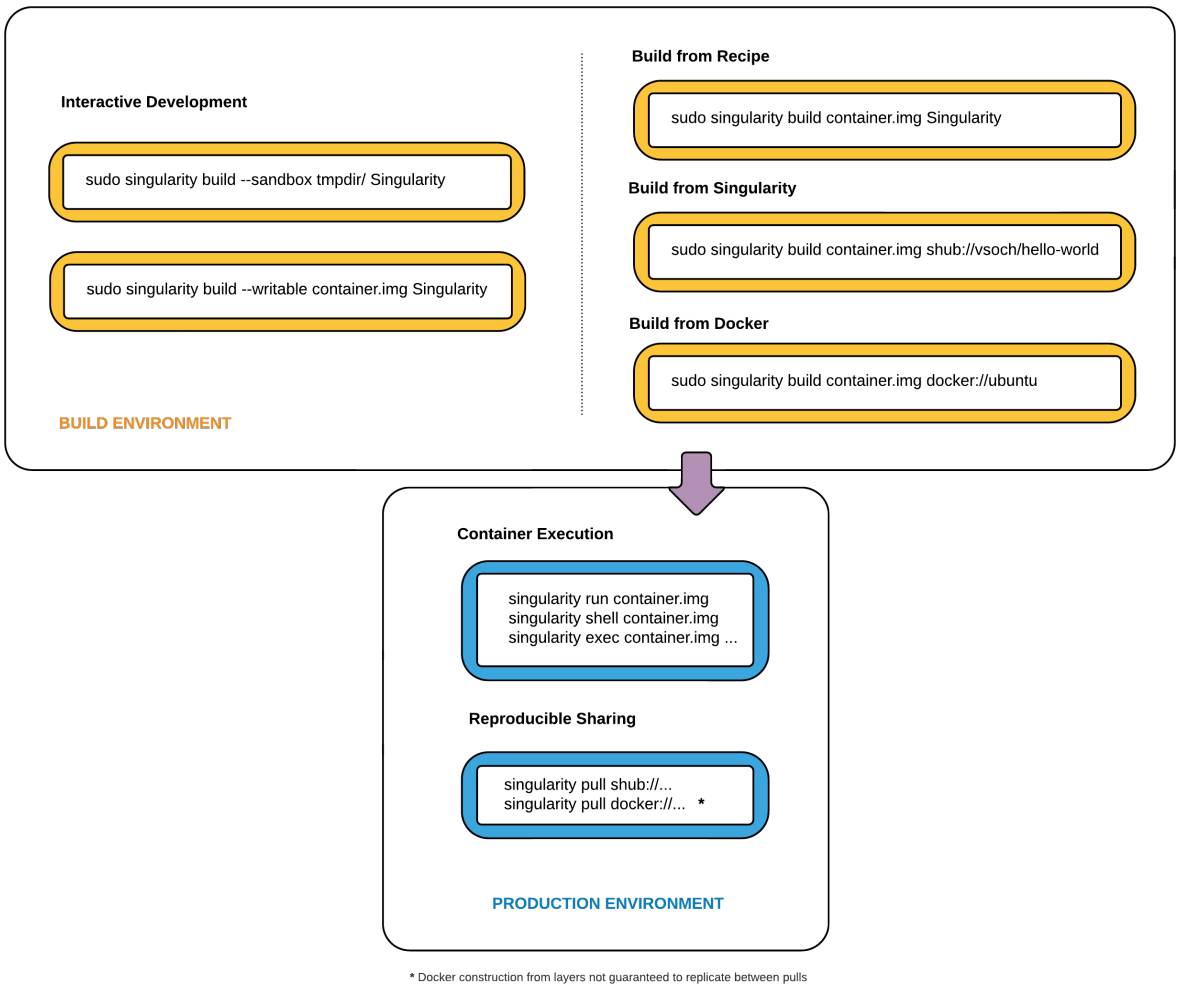


Figure 1.4: Standard `Singularity` workflow for building production images. "Build environment" (left): User can build images interactively, from recipes, `Singularity` or `Docker`. "Production environment" (right): Once the container images have been built and tested, they can then be shared and executed on different OS without the need for root privileges. *Note:* Adapted from https://www.sylabs.io/guides/2.6/user-guide/singularity_flow.html?highlight=workflow

.

2016; Schulz *et al.*, 2016). Computational scaling can be achieved by using distributed systems, whereby several computers/servers are connected to each other within a network to serve a common purpose (Agarwal and Owzar, 2014). Distributed systems can be classified into two: (1) compute clusters, where computers are connected to the same administrative domain, usually a local area network (LAN); and (2) computer grid, where computers are connected across different networks and administrative domains (Agarwal and Owzar, 2014).

Compute cluster and computer grid architectures (also referred to as HPC clusters) solve the computational capacity problem in complex multi-step analysis workflows by dividing and distributing tasks amongst the computers in the collective and executing them in parallel (Agarwal and Owzar, 2014; Muir *et al.*, 2016). This "parallelisation" of tasks in a workflow significantly improves execution times and speed. Each connected computer on the distributed system is referred to as a "node". Nodes share resources, i.e., storage and memory, amongst each other; however, their CPUs are autonomous (Agarwal and Owzar,

2014). The sharing of resources between nodes requires coordinated communication in order to maintain the integrity and consistency of the input/output data. Communication between nodes can be through shared storage, memory or an underlying system management software (Agarwal and Owzar, 2014; Muir *et al.*, 2016). Some HPC clusters do not share any memory or storage; they rely on system management software, or batch queuing systems, to distribute tasks amongst nodes, allocate required resources and communicate processes between nodes.

HPC infrastructure does have its pitfalls; hosting and maintenance of HCP clusters, including management of hardware and softwares installations, come at a practicality and financial cost (Agarwal and Owzar, 2014). The advent of cloud computing has seen many researchers and research institutes resorting to cloud computing as a platform for conducting their research. Google Cloud (`https://cloud.google.com/`), Microsoft Azure (`https://azure.microsoft.com`) IBM Cloud (`https://www.ibm.com/cloud/`) and the earlier mentioned AWS (`https://aws.amazon.com/`), are some of the companies that offer infrastructure, platform and software as a service to researchers to conduct their research on a scalable virtual environment based on their computing needs. Even the most popular bioinformatics sequence alignment tool, NCBI's BLAST (Basic Local Alignment Search Tool, `https://blast.ncbi.nlm.nih.gov/`), has been adapted to the cloud environment on the AWS `https://ncbi.github.io/blast-cloud/` for high-volume BLAST searches that are not feasible on the public NCBI site or local installations (Agarwala *et al.*, 2018). Users wishing to use this service can simply create or login to their AWS accounts, setup BLAST Amazon machine image (AMI) and perform their BLAST searches.

1.1.4 Workflow documentation & sharing

A workflow may be developed on a workflow management system that supports containerisation software and scaling to HPC cluster and cloud infrastructure; however, without proper documentation, archiving and version control, it cannot be fully reproducible and portable. A major challenge in the computational research community is having to keep track of the changes made in programming code and being able to reproduce the results, especially when collaborating with other researchers (Blischak *et al.*, 2016; Piccolo and Frampton, 2016; Wilson *et al.*, 2014). Subtle changes to a program can have unintended results, and thus when code is shared, the exact script used to produce results, along with the specific versions of applications used, parameters and computational resources/infrastructure need to be documented before distributing a piece of code or workflow (Sandve *et al.*, 2013). To fully support workflow reproducibility and portability, VCS like `Git` (`https://git-scm.com/`) and GitHub (`https://github.com/`) can be used to document and track changes throughout development of a workflow, allowing other researchers to contribute to your code (Blischak *et al.*, 2016).

Although this may seem like a trivial part of a reproducible workflow development, version

control is perhaps the most important and pivotal component of the process that programmers must practice more often. Figure 1.5 summarises the best practices discussed in this section that researchers should apply to their research and reproducible workflow development. As shown in the figure, VCSs (GitHub in particular) are central to the process. Workflow development starts at a local institutional server or desktop machine, where researchers can use a `Nextflow` to orchestrate a working pipeline using small testing datasets and applications. `Singularity` images can then be created for each piece of software needed to execute the workflow. The instructions to create the `Singularity` images must be saved as recipes along with the code scripts for the workflow. A GitHub repository can be created to keep track of the changes of all the scripting files and pushed to GitHub, or pulled using local machine using Git.

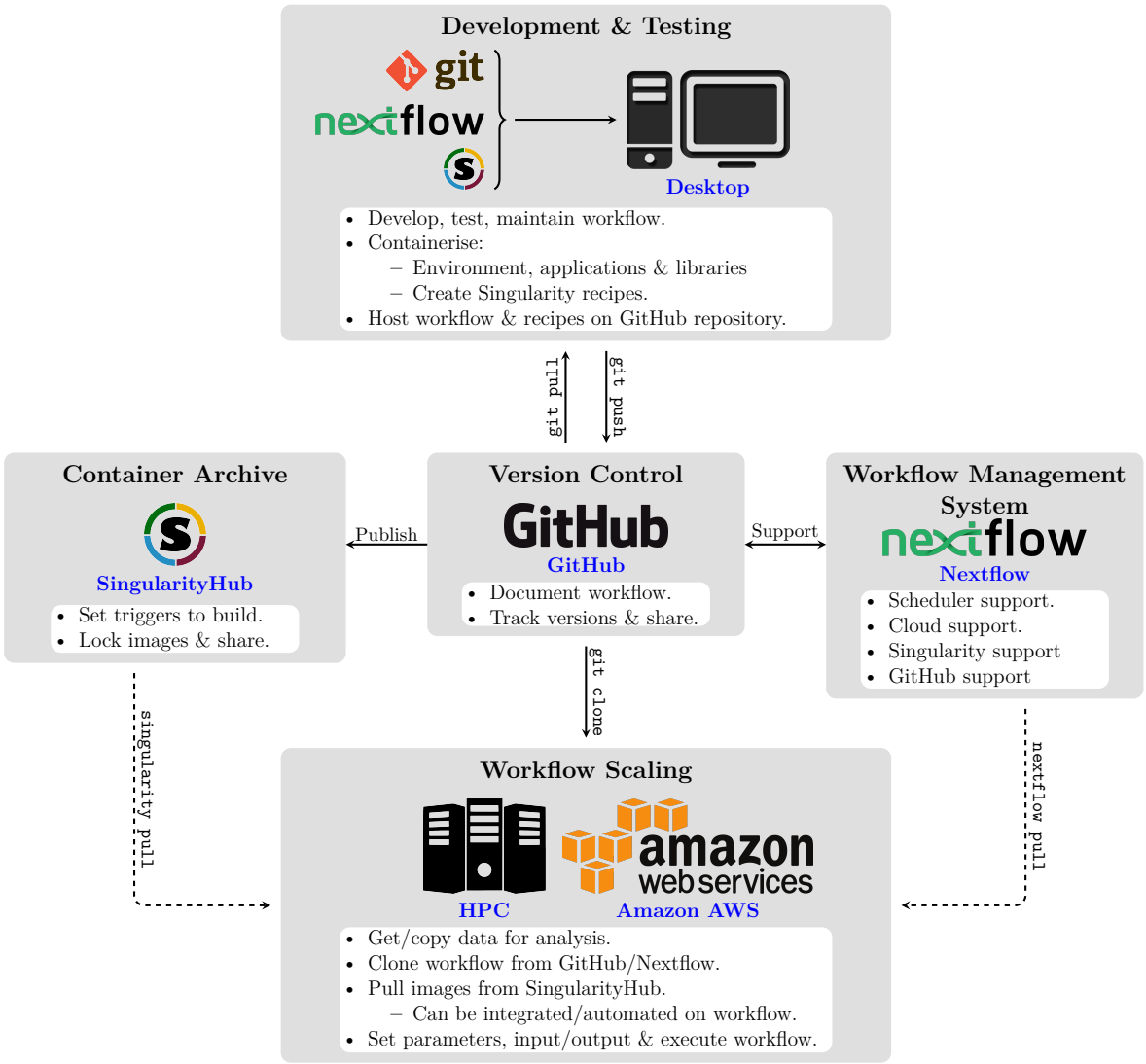


Figure 1.5: Flow chart summarising resources and best practices for development, maintenance, sharing and publishing of reproducible and portable workflows. Development of reproducible workflows start on individual desktop machines using `Nextflow`, `Singularity` and Git. A workflow repository can be created on GitHub to track version changes. SingularityHub can be used to create and archive images triggered by a GitHub push. The workflow can be cloned on HPC or cloud-services for analyses on a larger scale.

An account/repository can also be created on the SingularityHub to create and archive the `Singularity` images for the workflow. The images can be synchronised with the changes made to the `Singularity` image recipe files that are in the GitHub repository by creating triggers - everytime a recipe file is changed and pushed to GitHub, an image for that recipe is updated on the SingularityHub repository. The workflow can then be scaled to HPC clusters and cloud infrastructure by cloning from GitHub and using the added support that `Nextflow` provides. Images needed to scale the workflow can be pulled directly from SingularityHub, or configured for automatic download in the workflow configuration files. The workflow can then be published, shared and used for collaborative work with other researchers.

In the following sections of this chapter, I introduce the disease example that will be used to illustrate the utility of the workflows designed in this study. Systemic sclerosis was chosen because of the availability of the RNA-seq data from black South African patients affected with the disease and unaffected controls, and the need to better understand the pathogenesis of this devastating disease.

1.2 Introduction to Systemic Sclerosis

Systemic sclerosis (SSc), also known as scleroderma, is rare multi-system connective tissue disorder. It is an autoimmune disease in which abnormalities in the vascular and immune systems result in the fibrosis of the connective tissue, skin and internal organs (Broen *et al.*, 2014; Fonseca *et al.*, 2007; Gabrielli *et al.*, 2009; Katsumoto *et al.*, 2011). The mechanism by which the three main hallmarks in SSc (vasculopathy, immune dysfunction and fibrosis) are interlinked is poorly understood. This is mainly due to the heterogeneity of clinical manifestations presented by patient populations and lack of animal models to accurately reproduce the phenotypes observed in SSc (Katsumoto *et al.*, 2011; van Bon *et al.*, 2014).

SSc occurs worldwide, with a varying geographical prevalence of 50-300 per million and incidence of 2.3-22.8 per million individuals per year (van Laar *et al.*, 2014). Females are more commonly affected by SSc than males (ratio ~8:2), with an age of onset between 45 and 65 years old (Hachulla and Launay, 2011). In different populations, the prevalence and manifestations of SSc vary with occupation and ethnicity (Tager and Tikly, 1999). In America, European-Americans seem to have a lower risk of developing SSc compared to African-Americans whose ancestries are largely of west African origin. In Africa, the disease is rare and very few cases have been reported. In southern Africa, studies have shown that the risk of developing SSc is rare for both black and white populations, but more common among gold miners exposed to silica (Cowie and Dansey, 1990; Pudifin *et al.*, 1991; Silber, 1983). Although a number of genetic susceptibility factors and environmental triggers for SSc have been identified, their contribution to the onset and progression of the disease is still unclear (Korman *et al.*, 2008; Tsuchiya *et al.*, 2009; Zhou *et al.*, 2009).

There is no cure for SSc; however, some features of the disease can be treated through available therapies (Denton, 2015). The progression of SSc in affected individuals is chronic and associated with heterogeneous manifestations that can be lethal or reduce the quality of life. Compared to other immune-mediated rheumatic diseases (disease that affect connective/supportive structures of the body, mostly characterised by inflammation), SSc has a high morbidity and mortality rate (Denton and Khanna, 2017; Katsumoto *et al.*, 2011). The clinical symptoms presented by SSc patients at an early stage can be varied and are common in the general population. This can lead to a delay in the diagnosis (or even lead to misdiagnosis) in general clinics, and ultimately result in improper treatment/management of the disease (Denton and Khanna, 2017; Hachulla and Launay, 2011).

1.3 Classification of SSc

SSc is one of the many scleroderma-related heterogeneous disorders, which are all related by the presence of thickened, sclerotic skin lesions. The nomenclature and characterisation of scleroderma disease can be quite confusing and has changed in recent years. According to Careta and Romiti (2015), these disorders can be broadly characterised into two main forms: localised scleroderma (LoS) and systemic sclerosis (SSc) (Table 1.1). In adult LoS, the term "morphea" is often preferred when referring to the localised form of scleroderma, whilst in paediatric LoS, the term is kept the same (Fett, 2013). Referring to LoS as "morphea" in adults reduces miscommunication as patients and doctors often link "scleroderma" with SSc. LoS (or morphea) can cause significant morbidity, but does not affect mortality. SSc on the other hand has the highest morbidity and mortality of all immune-related connective tissue diseases (Fett, 2013).

Table 1.1: Classification of scleroderma-related disorders

Disorder	Description
Localised Scleroderma	
Linear scleroderma	Linear, band-like, sclerotic lesions which often follow a dermatome. Problematic when occurring over a joint line
Morphea	Localised area of skin involvement, but can become generalised
En coup de sabre	Linear SSc occurring on the face or scalp. Underlying bone and brain tissue may be involved
Systemic Scleroderma	
Limited cutaneous SSc	Skin thickening limited distally to elbows and knees, and may involve the neck and face. PHTN*may occur
Diffuse cutaneous SSc	Diffuse skin thickening, including the trunk, extremities and face. Internal organ involvement is common
CREST†syndrome	A limited form of SSc with prominent calcinosis, RP§ esophageal dysmotility, sclerodactyly and telangectasia. PHTN*can occur
Overlap SSc	SSc disease manifestations in coexisting with other rheumatic disease: SLE,¶ myositis, or RA‡
SSc sine scleroderma	RP and other clinical and serological manifestations of SSc, but lacking the skin changes

* PHTN: pulmonary arterial hypertension.
† CREST: calcinosis, Raynauds phenomenon, esophageal dysmotility, sclerodactyly, telangectasia.
‡ RA: rheumatoid arthritis.
§ RP: Raynauds phenomenon.
¶ SLE: systemic lupus erythematosus.
Note: Adapted from Meier *et al.* (2013)

The presentation of the scleroderma spectrum diseases vary from involvement of a small area of skin to diffuse skin and internal organs (Table 1.1). The localised form of scleroderma is mostly confined to the skin and/or underlying tissue (Careta and Romiti, 2015). It can be further subdivided into linear scleroderma (LS), morphea and "*en coupe de sabre* (ECDS). LS is most common in children and causes skin and subcutaneous tissue abnormalities that have a dermatomal (skin area supplied by a single nerve) distribution (Fett, 2013). Morphea is generally limited to skin, but can extend to affect muscle and bone tissue. According to Fett (2013), morphea can present as circumscribed (few circles on the trunk or limbs), generalised (many circles on the trunk and limbs), linear (lines of involvement on the limbs or head), mixed (combination of circumscribed and linear or generalised and linear) or pansclerotic (involvement of all of the skin). ECDS is a rare form of LS affecting the head and neck, characterised by band-like sclerotic lesions involving the frontoparietal regions of the forehead and scalp (Anderson *et al.*, 2018; Duman and Ekinci, 2018; Meier *et al.*, 2013).

The systemic form of scleroderma (SSc) is mainly characterised by cutaneous sclerosis and involvement of internal organs, especially the lungs, oesophagus and the vascular system (Careta and Romiti, 2015). SSc can be divided into four subcategories, i.e., limited cutaneous SSc (lcSSc), diffuse cutaneous SSc (dcSSc), CREST (calcinosis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, telangectasia) syndrome, overlap SSc and SSc sine scleroderma as shown in Table 1.1 (Meier *et al.*, 2013). In lcSSc, there are limited skin changes which are often restricted to the hands, arms and face. In dcSSc, there is thickening of a large area of the skin, and it also affects one or more internal organs. The CREST syndrome is considered to be a subset of lcSSc; it cannot be assigned to only one subgroup of patients with SSc and does not present with significant internal organ involvement (Gabrielli *et al.*, 2009; Hachulla and Launay, 2011; Katsumoto *et al.*, 2011).

The overlap SSc/syndrome subtype of SSc is a combination of one or more connective tissue diseases. These may include SSc, Sjögren's syndrome, dermatomyositis/polymyositis, rheumatoid arthritis or systemic lupus erythematosus (Balbir-Gurman and Braun-Moscovici, 2011). SSc sine scleroderma is a rare variant of SSc that characterised by the internal organ(s) and immunological manifestation, but lack the clinically detectable proximal skin thickening/fibrosis (sclerodactyly) involvement (Diab *et al.*, 2014; Marangoni *et al.*, 2013). Table 1.2 compares the characteristics and differences between LoS and SSc. Both LoS and SSc have fibrosis of the skin in common. In LoS, there are patches or linear distribution of skin thickening, whereas in SSc there is sclerodactyly and proximal skin thickening. The Raynaud's phenomenon is not observed in LoS, but present in SSc. Raynaud's phenomenon is a disorder that results in skin discoloration of the extremities (fingers and/or toes) due to lack of blood flow. The decrease in blood flow is caused by abnormal spasm of the blood vessels (vasospasm) as a result of exposure to temperature changes or emotional stimuli (Katsumoto *et al.*, 2011).

Table 1.2: Comparison of the scleroderma-related disorders

Feature	Localised Scleroderma	Systemic Sclerosis
Skin findings	Patches or linear distribution of thickened skin	Sclerodactyly ± proximal skin thickening
Raynauds phenomenon	Absent	Present
Digital ischaemic changes	Absent	Usually present (digital pitting scars or ulcers, loss of fingerpad substance)
Internal organ disease	Absent	Present
Antinuclear antibody	Positive in ≥50% of cases	Positive in ≥85% of cases
Scleroderma-specific auto-antibodies*	Negative	Positive in 60% of cases
Biopsy – histologic findings	Dermal fibrosis	Dermal fibrosis

* Scleroderma-specific antibodies include antibodies to centromere, topoisomerase-1 (Scl-70), and RNA polymerase III.
Note: Adapted from Ferri (2016)

Digital ischaemia can also be observed in SSc but not in LoS. These ischaemic events can be very painful and may often lead to tissue loss, which can have a significant effect on the quality of life (McMahan and Wigley, 2010). Internal organ involvement can be seen in SSc, but not in LoS. When it comes to antibodies, antinuclear antibodies are present in ≥50% and ≥85% of the cases in LoS and SSc respectively. Scleroderma specific auto-antibodies, i.e., anti-centromere, anti-topoisomerase-1 (Scl-70), and anti-RNA polymerase III, are only positive on 60% of the cases in SSc and absent in LoS. The remainder of the thesis will focus on the dcSSc and lcSSc subtypes of the SSc disorders.

1.4 Autoantibodies in SSc

Accurate classification of SSc subsets is crucial for research in understanding the molecular pathogenesis of the disease. It ensures that patients with similar features, which can be compared across studies, are properly recruited. It also increases accuracy of diagnostic measures and the development of appropriate screening and treatment programs for patients with SSc (Hachulla and Launay, 2011). The initial criteria for classification of SSc were published by Masi (1980). However, as these were based on clinical and chest X-ray observations, 10-20% of patients were misdiagnosed as they did not meet these criteria.

In 1988, LeRoy *et al.* (1988) proposed a new classification schema of SSc with two subsets: (1) ***diffuse cutaneous (dcSSc)***, characterised by widespread and fast progressing thickening of the skin and early internal organ involvement; (2) ***limited cutaneous (lcSSc)***, in which skin involvement is limited to hands, arms and face, and may be associated with less severe internal organ involvement at a later stage. The two SSc subgroups could further be differentiated by the presence of specific autoantibodies; lcSSc was associated with anti-centromere autoantibodies (ACA) and dcSSc was associated with anti-topoisomerase autoantibodies (ATA) (LeRoy *et al.*, 1988).

Since then, there has been a number of SSc-specific autoantibodies that have been identified, each with unique clinical associations, disease severity and specific clinical manifestation (Koenig *et al.*, 2008; Senécal *et al.*, 2005; Steen, 2005; van den Hoogen *et al.*, 2013). The presence of serum autoantibodies, also known as anti-nuclear antibodies (ANA),

against intracellular antigens in SSc is one of the hallmarks of the disease (Denton, 2015; Katsumoto *et al.*, 2011; Kayser and Fritzler, 2015). These serum autoantibodies are found in 95% of patients with SSc. ANAs are useful biomarkers for classifying patients into SSc subsets. They are also useful in providing an accurate early diagnostic and prognostic information for patients (Choi and Fritzler, 2016; Villalta *et al.*, 2012). However, there is no known evidence of the role that these molecules play in the disease pathogenesis (Katsumoto *et al.*, 2011; Kayser and Fritzler, 2015; Steen, 2005).

The major ANAs in SSc, which are significant diagnostic markers, include: (1) anti-centromere (ACA), (2) anti-topoisomerase I (ATA), (3) anti-RNA polymerase III (anti-RNAP), (4) anti-fibrillarin and (5) anti-Th/To antibodies (Katsumoto *et al.*, 2011; Kayser and Fritzler, 2015; Mehra *et al.*, 2013; Steen, 2005; van den Hoogen *et al.*, 2013). These ANAs are considered to be mutually exclusive and do not change from one antibody to another in the course of the disease (Denton and Khanna, 2017; Senécal *et al.*, 2005; Villalta *et al.*, 2012). This allows early differentiation of SSc patients as well as development of an approach to the disease management. Table 1.3 summarises the frequency, disease subtype, clinical associations and prognosis of SSc-specific ANAs as well as other autoantibodies that are less specific to SSc, but found in overlap SSc and other immune-mediated rheumatic diseases. The most common SSc-specific ANAs associated with the dcSSc and lcSSc subtypes are discussed in Sections 1.4.1 to 1.4.5 below.

Table 1.3: SSc anti-nuclear antibody frequency and clinical correlations

Autoantibody	% Freq. in SSc	Disease subtype	Clinical associations	Prognosis
Anti-centromere	20-38	lcSSc	Pulmonary arterial hypertension	Better prognosis
Anti-topoisomerase I	15-42	dcSSc	Pulmonary fibrosis; Heart involvement	Worse prognosis
Anti-RNA polymerase III	5-31	dcSSc	Renal crisis; Tendon friction rubs, synovitis, myositis, joint contractures	Increased mortality
Anti-U3RNP (fibrillarin)	4-10	dcSSc	Renal crisis and cardiac involvement	Poor prognosis especially in African-Americans
Anti-Th/To	1-13	lcSSc	Pulmonary fibrosis and renal crisis	Poor prognosis
Anti-U11/U12 RNP*	3.2	-	Raynauds phenomenon; Gastrointestinal involvement; Lung fibrosis	Increased mortality
Anti-U1-RNP*	2-14	lcSSc	Raynauds phenomenon, puffy fingers, arthritis, myositis, overlap syndrome (i.e., MCTD†)	Better prognosis
Anti-PM-Scl	4-11	Overlap with polymyositis lcSSc	Raynauds phenomenon, arthritis, myositis, pulmonary involvement, calcinosis, and sicca symptoms	Better prognosis
Anti-Ku	2-4	-	Myositis, arthritis, and joint contractures	-
Anti-hUBF (NOR 90)	<5	lcSSc	Mild internal organ involvement	Better prognosis
Anti-Ro52/TRIM21‡	15-20	Association with other autoimmune diseases	Older age onset, pulmonary fibrosis	-

* RNP: ribonucleoprotein
† MCTD: mixed connective tissue disease
‡ TRIM: tripartite motif
Note: Adapted from Kayser and Fritzler (2015)

1.4.1 Anti-centromere antibodies

Anti-centromere antibodies (ACA) are mostly associated with lcSSc and/or CREST syndrome, and have been shown to precede the clinical manifestation of the disease (Hudson *et al.*, 2012). Patients with ACA have also been reported to be significantly prone to developing pulmonary hypertension (Table 1.3). Traditionally, ACAs have been identified through the indirect immunofluorescence (IIF) cell based-assays, which identify autoantibodies that bind to centromeric proteins (CENP) autoantigens. CENPs are crucial intracellular proteins, which are involved in the assembly of the kinetochore protein complex during cell division. The kinetochore is a multi-protein complex consisting of a number of proteins, including a few belonging to the CENP autoantigen family (Fritzler *et al.*, 2011; Perosa *et al.*, 2016). These include the 17kDa CENP-A, 80kDa CENP-B, 140kDa CENP-C (made up of three functional domains), 50kDa CENP-D, 312kDa CENP-E, 400kDa CENP-F, 95kDa CENP-G and the 38kDa CENP-O.

Even though there are several autoantigens in the CENP family of proteins, most studies in SSc have directed their focus on anti-CENP-A and anti-CENP-B autoantibodies as a result of their detection using the IIF in SSc sera (Hudson *et al.*, 2012). Anti-CENP-A and anti-CENP-B are the most commonly detected autoantibodies in patients with SSc and are typically associated with the lcSSc subtype of SSc (Favoino *et al.*, 2013). Compared to other SSc-related autoantibodies, anti-CENP autoantibodies are better predictors of SSc; and are typically detected in SSc patients with a frequency ranging from 20-40% and a specificity of more than 90%. However, the frequency and specificity of both anti-CENP-A and anti-CENP-B varies amongst different ethnicities (Mehra *et al.*, 2013; Perosa *et al.*, 2016; Steen, 2005; Wielosz *et al.*, 2014).

1.4.2 Anti-topoisomerase I antibodies

In 1979, Douvas *et al.* (1979) identified an autoantibody in sera from five unrelated SSc patients, anti-Scl-70, which reacted with a 70kDa immunoglobulin G (IgG) proteins using the immunoblotting (IB) technique. However, the study done by Shero *et al.* (1986) revealed that the 70kDa Scl-70 autoantigen was a misnomer breakdown product of a 100kDa full-length protein topoisomerase 1 (topo 1) (Kayser and Fritzler, 2015; Mahler *et al.*, 2010; Mehra *et al.*, 2013). Although the terms anti-Scl-70 and anti-topo 1 are still used interchangeably, anti-topo 1 antibody (ATA) is a more preferred and accurate term. Topo 1 is an ATP-dependent enzyme that plays an essential role in mitosis, transcription, replication, recombination and chromosome condensation. Its function is to relieve stress on DNA, thus removing the coils that form in DNA allowing the DNA strands to pass through each other allowing cellular processes on DNA to progress (Wang, 2002).

A number of assays can be used to detect ATA from SSc sera with varying specificity. These include IB, double immunodiffusion (DID), enzyme linked immunoassay (ELISA), line immunoassay (LIA), addressable laser bead immunoassays (ALBIA) or immunopre-

cipitation (IP) (Mehra *et al.*, 2013). ATA are highly associated with the dcSSc subtype of the SSc disorder, even though associations with other immune-mediated rheumatic and lcSSc have been reported, and can be detected from SSc patients with a frequency range of 15-42% and specificity of 90-100% (Kayser and Fritzler, 2015; Mehra *et al.*, 2013). ATA is associated with high mortality and is a poor predictor of SSc. The clinical manifestations of ATA include lung fibrosis, cardiac involvement and renal crisis (Denton, 2015; Favoino *et al.*, 2013; Fonseca *et al.*, 2007; Silver and Silver, 2015).

1.4.3 Anti-RNA polymerase I, II and III antibodies

RNA polymerase (RNAP) antibodies were first reported by Stetler *et al.* (1987, 1982), through their radioimmunoassay findings in screening for antibodies in sera of patients with immune-mediated rheumatic diseases (Sjögren's syndrome, systemic lupus erythematosus [SLE], rheumatoid arthritis [RA] and mixed connective tissue disease [MCTD]) that bind to a purified RNAP I antigen. It wasn't until 1987 that autoantibodies against RNAP 1 were identified in 4% of the 208 sera of patients with SSc Reimer *et al.* (1987). Since then, reports of anti-RNAP I, II and III antibodies have been reported to be associated with the dcSSc subtype of the SSc disorder (Kuwana *et al.*, 1993).

Anti-RNAP I and III are considered to be significantly specific for SSc and are typically found to exist side-by-side. However, anti-RNAP II can also be found in other immune-mediated rheumatic disease, including the SSc overlap disorder and SLE, and are thus not considered specific to SSc (Kayser and Fritzler, 2015; Mehra *et al.*, 2013). The frequency of anti-RNAP I and III ranges from 5-31%, however, there are variations of based on geographical differences (Sobanski *et al.*, 2014; Steen, 2005). These variations may implicate significant roles played by genetic and environmental factors in SSc. Clinical manifestations of anti-RNAP 1 and III in SSc patients typically include renal crisis and increased risk of joint contractures, myositis, tendon friction rubs and synovitis (Kayser and Fritzler, 2015).

1.4.4 Anti-fibrillarin antibodies

Fibrillarin is a 34kDa protein of the small nucleolar U3-ribonucleoprotein (U3RNP) macromolecular complex (Kayser and Fritzler, 2015; Lischwe *et al.*, 1985; Mehra *et al.*, 2013). Anti-fibrillarin (or anti-U3RNP) antibodies were first identified by Lischwe *et al.* (1985) using immunostaining techniques to screen 16 SSc patients sera for SSc-specific antigens. Anti-U3RNP have a frequency of 4-10%, are associated with the dcSSc subtype of SSc and are disjoint with ACA, anti-RNAP and anti-topo 1 (Hamaguchi, 2010). Compared to Asian and Caucasian SSc patients, anti-U3RNP occurs more frequently in African-American patients and their clinical manifestations include cardiac and renal involvement. However, clinical manifestations of anti-U3RNP vary with ethnicity (Hamaguchi, 2010; Steen, 2005).

1.4.5 Anti-Th/To antibodies

Anti-Th/To antibodies were first reported by Okano *et al.* (1992), identified through immunoprecipitation techniques on sera from SSc patients. These antibodies are reactive against the subunits of the ribonuclease (RNAse) mitochondrial RNA processing (MRP) and RNAse P complexes. Both RNAse MRP and RNAse P have a number of subunits in common, contain an RNA component that has similar secondary structure and function as site-directed endonucleases (Van Eenennaam *et al.*, 2002). Anti-Th/To are present in SSc patients at a frequency of 1-13% and are reported to be specific to lcSSc subtype of SSc. Clinical manifestations of anti-Th/To include renal involvement and pulmonary fibrosis, thus making them poor predictors of SSc (Hamaguchi, 2010; Kayser and Fritzler, 2015; Mehra *et al.*, 2013; Van Eenennaam *et al.*, 2002).

1.5 Pathology, aetiology and pathogenesis of SSc

To date, the aetiology and pathogenesis of SSc are poorly understood. However, a combination of genetic predisposition and environmental triggers are believed to trigger vascular and immune abnormalities, which ultimately leads to the fibrosis of tissue and internal organs as a manifestation of SSc (Fett, 2013; Katsumoto *et al.*, 2011; Stern and Denton, 2015). It is not clear how these three hallmarks of SSc are mechanistically linked or related, which of them is the most important and how the interactions between them causes the disease to develop (Murdaca *et al.*, 2016). The genetic predisposition and environmental triggers that lead to the development of the disease are also poorly understood. Figure 1.6 shows a proposed schematic overview of how the three pathways involved in the development and progression of SSc (vasculopathy, autoimmune dysregulation and fibrosis) interplay in the pathogenesis of the disease.

Vasculopathy in SSc involves the fibrosis of the endothelial layer of the small blood vessels, due to endothelial dysfunction, and vasospasms (Raynaud's phenomenon) brought about by temperature changes and/or stress. In larger blood vessels, vasculopathy presents as scleroderma renal crisis (SRC) or pulmonary arterial hypertension (PAH/PHTN) (Elhai *et al.*, 2015; Katsumoto *et al.*, 2011; Tejera Segura and Ferraz-Amaro, 2015). Aberrations in the immunologic pathways in SSc leads to the activation of lymphocytes, which in turn leads to an overproduction of growth factors, cytokines and autoantibodies, and the dysregulation of the innate immune system where macrophages and lymphocytes attack the affected tissue (Hua-Huy and Dinh-Xuan, 2015).

Fibrosis in SSc can lead to thickening of the skin and involvement of internal organs. Pulmonary fibrosis, cardiac dysfunction are the most common manifestations of internal organ involvement in SSc, with lung involvement being the common cause of death (Hua-Huy and Dinh-Xuan, 2015; Katsumoto *et al.*, 2011). In the following sections (Sections 1.5.1 to 1.5.3), the contribution of the components of the pathways involved with SSc are

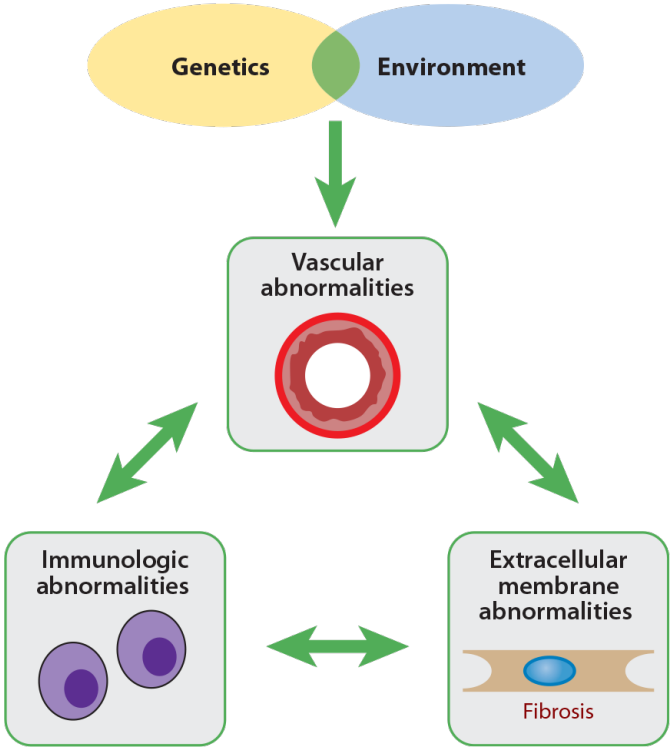


Figure 1.6: Proposed pathogenesis of SSc. A combination of genetic and environmental factors trigger abnormalities in the vascular, immunologic and fibrotic pathways. *Note:* From Katsumoto *et al.* (2011)

discussed, followed by the complex pathogenesis mechanism proposed for SSc in Section 1.5.4.

1.5.1 Vascular dysregulation

In SSc, vasculopathy is of central importance to the pathophysiology of the disease from its early onset to late complications (Kavian and Batteux, 2015). Vasculopathy can be severe and its presentation may even lead to amputation since it is irreversible in many cases (Tejera Segura and Ferraz-Amaro, 2015). Due to its high morbidity and mortality, vasculopathy in SSc can be particularly challenging due to lack of effective treatment options. The exact stimulus that triggers vasculopathy in SSc is currently not known. However, it is believed that infectious agents such as pathogens, nitrogen oxide-related free radicals, cytotoxic T-lymphocytes or anti-endothelial cell antibodies could be involved (Gabrielli *et al.*, 2009; Katsumoto *et al.*, 2011; Stern and Denton, 2015; Tejera Segura and Ferraz-Amaro, 2015). Regardless of what the vasculopathy triggers may be in SSc, a series of cascade events that alter the normal functioning of the vascular system are activated, which eventually lead to fibrosis. These include impairment of the blood circulation in the small blood vessels (microcirculation), dysfunction of the endothelial cells, impaired angiogenesis and vasculogenesis, platelet activation and maintained vasoconstriction (Tejera Segura and Ferraz-Amaro, 2015).

1.5.1.1 Microcirculation impairment

The earliest manifestations of SSc are those that result from the abnormalities in the microcirculatory system (Chora *et al.*, 2015). These include large, dilated and malformed capillaries due to endothelial cell injury and vascular remodelling (Rabquer and Koch, 2012). The damages to the small blood vessels evolve progressively from the early stages to the late stages of SSc with a variety of manifestations. Vascular damage consists of thickening of the basement membrane, increased size of the endothelial cells of the capillary wall, formation of gaps due to loss of intracellular junctions, vacuolization of the cytoplasm in the endothelial cells and loss of plasmalemma vesicles (Gabrielli *et al.*, 2009; Tejera Segura and Ferraz-Amaro, 2015). As SSc progresses, the tissues where there are damaged small blood vessels becomes more avascular, which eventually leads to loss of microcirculation.

Raynaud's phenomenon is the earliest typical clinical manifestation of impaired microcirculation in SSc observed in more than 90% of patients (Rabquer and Koch, 2012; Tejera Segura and Ferraz-Amaro, 2015). Prolonged vasospasm of the digital arteries due to Raynaud's phenomenon causes further injury to the microvasculature, eventually leading to the formation of digital ulcers (Chora *et al.*, 2015; Kavian and Batteux, 2015; Rabquer and Koch, 2012). Digital ulcers are a considerable burden to patients with SSc, and are associated with extremely high morbidity and are difficult to treat. They often lead to loss of tissue and amputation in severe cases since they can become sites of secondary infections (Kavian and Batteux, 2015; Strange and Nash, 2009). PAH and renal disease are more severe manifestations of SSc caused by abnormalities in the microvascular system, and are the main contributors of increased mortality in SSc (Rabquer and Koch, 2012). Parallel to abnormalities in the microvascular system, macrovascular abnormalities are also seen in patients with SSc, where there is a significantly high proliferation and accumulation of endothelial, proteoglycans and smooth muscle cells in the arteries and arterioles. This proliferative vasculopathy eventually leads to intimal thickening and blockage of the blood vessel, which puts SSc patients at a high risk of developing atherosclerosis (Kavian and Batteux, 2015; Rabquer and Koch, 2012; Tejera Segura and Ferraz-Amaro, 2015).

1.5.1.2 Endothelial dysfunction

Normal blood flow, anti-thrombosis, blood coagulation, enzymatic breakdown of fibrin and trans-endothelial migration of leukocytes are all processes regulated by the endothelium (Katsumoto *et al.*, 2011). Injury to the endothelium can lead to an imbalance in the production of vasoactive molecules; there is an aberrant upregulation of vasoconstrictors, such as endothelin-1 (ET-1), and down-regulation of vasodilators, such as nitric oxide (NO) and prostacyclin. The presence of high levels of ET-1, along with the von Willebrand factor (vWF), in SSc patients are indications of the active form of SSc and dysfunction of the endothelial cells (Kavian and Batteux, 2015). The increased levels of ET-1 and vWF cause an aggregation of platelets/thrombocytes and constriction of the blood

vessels, which ultimately leads to formation of intravascular thrombosis as the platelets continuously adhere and fibrin accumulates (Tejera Segura and Ferraz-Amaro, 2015). Endothelial dysfunction could also arise due to apoptosis of the endothelial cells. Endothelial cell apoptosis is thought to be triggered as a direct result of viral infection or indirectly through recognition of infected cells by cytotoxic T-lymphocytes (Kavian and Batteux, 2015). The indirect apoptosis of endothelial cells could activate innate immunity; whereby in the viral-infected endothelial cells, the viral epitopes cross-react with endothelial cells. This leads to the production of anti-endothelial cell antibodies, and eventually systemic endothelial cell apoptosis through recognition by cytotoxic T-lymphocytes (Katsumoto *et al.*, 2011; Kavian and Batteux, 2015).

1.5.1.3 Impaired angiogenesis and vasculogenesis

New blood vessels are formed from pre-existing blood vessels through a process known as angiogenesis. They can also be formed through vasculogenesis, whereby new blood vessels are formed from hemangioblasts or vascular stem cells (Del Papa and Pignataro, 2018; Tejera Segura and Ferraz-Amaro, 2015). Both angiogenesis and vasculogenesis processes are dysregulated in patients with SSc; thus, the abnormalities that occur in the vascular system cannot be reversed or compensated for. The formation of new blood vessels through the highly regulated process of angiogenesis relies on the activation, proliferation and migration of the endothelial cells. This process is initiated by pro-angiogenic mediators, which promote the degradation of the extracellular matrix through the release of proteolytic enzymes including matrix metalloproteinases (MMPs) (Katsumoto *et al.*, 2011; Kavian and Batteux, 2015; Rabquer and Koch, 2012; Tejera Segura and Ferraz-Amaro, 2015).

In SSc, inadequate blood supply to the affected tissue induces pro-angiogenic factors, including vascular endothelial growth factor (VEGF), platelet-derived growth factor (PDGF), transforming growth factor beta (TGF-β), ET-1 and the monocyte chemoattractant protein-1 (MCP-1). However, even though there is an upregulation of these pro-angiogenic factors in the affected tissues of SSc patients, there is still diminished angiogenesis (Asano and Sato, 2015; Katsumoto *et al.*, 2011; Rabquer and Koch, 2012; Tejera Segura and Ferraz-Amaro, 2015). Anti-angiogenic factors have also been reported to be highly expressed in SSc patients, including angiostatin, platelet factor 4 (PF4), endostatin and thrombospondin. This suggests that there may be a combination of a number of mechanisms between pro- and anti-angiogenic factors that favour impaired angiogenesis in SSc patients (Katsumoto *et al.*, 2011; Rabquer and Koch, 2012). In contrast to angiogenesis, the process of formation of new blood vessels through vasculogenesis from stem cells in SSc is still unclear. Even though the progenitor stem cells migrate to the site of vascular damage in SSc, they seem to have reduced ability to differentiate and form new endothelial cells to replace the damaged cells in the vascular walls (Rabquer and Koch, 2012; Tejera Segura and Ferraz-Amaro, 2015).

1.5.1.4 Platelet activation

As described in Section 1.5.1.2, the imbalanced production of vasoactive agents due to endothelial injury eventually leads to the constriction of blood vessels and formation of intravascular thrombosis. In addition to the imbalance of the production of vasoactive agents in SSc patients, there is also an activation of platelets which play a significant role in vasculopathy (Kavian and Batteux, 2015). In the presence of ET-1, collagen and serotonin, the platelets are chronically activated and contribute towards the hypercoagulation observed in SSc. The activated platelets release inflammatory mediators, chemokines, cytokines and growth factors, which contribute to the manifestations of SSc (Asano and Sato, 2015; Kavian and Batteux, 2015).

1.5.2 Immune activation

In SSc, the dysregulation of both adaptive and innate immune systems have been observed and are associated with inflammation. In adaptive immune system, the autoantibodies produced by B lymphocyte and autoreactive T-lymphocytes play a crucial role in the pathogenesis of SSc (Fuschiotti, 2018; Laurent *et al.*, 2018). Tissues of SSc patients in both late and early stages of the disease have also been found to contain high levels of inflammatory cytokines, which are produced by the cells of the innate immune system (Dowson *et al.*, 2017). However, the role played by both adaptive and innate immune systems in the initiation and/or maintenance of SSc remains poorly understood. In SSc, the presence of markers of innate immune system activation in both early and late stages of the disease suggests that there is a critical role played by the innate immune system in the initiation of the disease as compared to the adaptive immune system (O'Reilly, 2014).

1.5.2.1 Innate immunity in SSc

The innate immune system is the body's first response against foreign/native unrecognised agents (microbes and pathogens) as well as mechanical/chemical damage to the body. The response to these foreign agents and damage to cells and tissues in the body is accomplished through their recognition by the pattern recognition receptors (PRRs) found on the cells of the immune system (O'Reilly, 2014). PRRs (Toll-like receptors [TLRs] in particular) recognise molecular signatures that are either pathogen-associated molecular patterns (PAMPs), microbe-associated molecular patterns (MAMPs), or damage-associated molecular patterns (DAMPs) (Dowson *et al.*, 2017; Varga, 2017). TLRs are a class of protein immune receptors that are expressed both intracellularly in endocytic vesicles and on membranes of the cells of the immune system, including macrophages, dendritic cells, natural killer (NK) cells, B- and T-lymphocytes. This enables TLRs to uniquely respond to specific ligands in an appropriate manner (Dowson *et al.*, 2017). The proper functioning of the innate immune system depends on the 13 different TLRs on the cells of the immune systems, i.e., TLR1, TLR2, TLR3, TLR4, TLR5, TLR6, TLR7, TLR8, TLR9, TLR10, TLR11, TLR12, and TLR13. However, in humans, TLR11, TLR12, and

TLR13 are not expressed. TLRs are not antigen-specific, but rather recognise patterns based on their nature, i.e., PAMPs, MAMPs and DAMPs (Pattanaik *et al.*, 2015).

Recognition of these molecular signatures by the TLRs causes a signalling cascade which results in the production of a number of pro-inflammatory cytokines, along with chemokines and interferons (IFNs), which help recruit other immune cells to the site of infection in order to clear out foreign agents (Dowson *et al.*, 2017; O'Reilly, 2014). In general, TLRs recognise PAMPs and MAMPs. However, under unfavourable conditions, TLRs can also recognise DAMPs which are released from endogenous cells as an alarm signal due to cellular stress or damage. In SSc, it has been established that there is a dysregulation of the innate immune system and TLR signalling in the early stages of the disease, and these abnormalities can be linked to the activation of the adaptive immunity and tissue fibrosis. The evidence of the innate immune system dysregulation arises from the histological observations of reduced capillaries and perivascular accumulation of inflammatory cells in the biopsies of skin lesions from SSc patients, which precedes the persistent fibrosis in SSc patients (O'Reilly, 2014; Varga, 2017). The inflammatory cells observed include macrophages, monocytes, T- and B-lymphocytes, and mast cells (Katsumoto *et al.*, 2011).

A number of TLRs have been implicated in SSc (Table 1.4 and Figure 1.7), i.e., TLR2 and TLR4 expressed on the cellular surfaces of cells, where they recognise different types of PAMPs in the extracellular environment, as well as TLR3, TLR7, TLR8 and TLR9 expressed intracellularly in the endoplasmic reticulum and endosomal membranes, where they recognise different types of viral and bacterial nucleic acids (Dowson *et al.*, 2017; Pattanaik *et al.*, 2015). TLR2 recognises most bacterial and parasitic ligands, including peptidoglycan and glycolipids. TLR4 specifically recognises bacterial endotoxin lipopolysaccharides. TLR3, TLR7 and TLR8 specifically recognise single-stranded and double-stranded viral RNAs. TLR9 recognises unmethylated CpG motifs in DNA, which is more common in microbial genomes as compared to mammalian genomes (Dowson *et al.*, 2017; Fullard and O'Reilly, 2015; O'Reilly, 2014; Pattanaik *et al.*, 2015).

The recognition of these different types of ligands by the TLRs results in the activation of various signalling pathways, leading to the activation of transcription factors. These transcription factors stimulate the expression of genes that encode for cytokines, enzymes and other proteins whose roles are essential in inflammation, antimicrobial and antiviral responses. Nuclear factor κB (NF-κB), activation protein 1 (AP-1) and interferon response factors 3 & 7 (IRF3 and IRF7) are the major transcription factors activated by the TLR signals (Abbas *et al.*, 2015, 2018). NF-κB and AP-1 are responsible for stimulating the expression of genes that encode for molecules that play major roles in inflammatory responses, including inflammatory cytokines (e.g. tumour necrosis factor [TNF] and interleukin 1 [IL-1]), chemokines (e.g. C-C motif chemokine ligand 2 [CCL2] and C-X-C motif chemokine ligand 8 [CXCL8]) and endothelial adhesion molecules. IRF-3

Table 1.4: TLRs implicated in SSc and their respective ligands and origins

TLR	Ligand	Ligand origin
TLR2	Lipoproteins	Various pathogens
	Zymosan	Fungi
	HMGB-1*	Endogenous
	Glycolipids	*Treponema maltophilum*
	Peptidoglycan	Gram-positive bacteria
	Serum amyloid	Endogenous
	Snapin A	Endogenous
TLR3	dsRNA†	Viruses
	Poly(I:C)	Synthetic
TLR4	LPS‡	Gram-negative bacteria
	Hyaluronan fragments	Endogenous
	HMGB-1*	Endogenous
	HSP§ -20, -60, -70 and -96	Endogenous
	Fibrinogen	Endogenous
	Fibronectin	Endogenous
	Tenascin C	Endogenous
	Surfactant protein-A	Endogenous
	Amyloid β fibril	Endogenous
TLR7	ssRNA¶	Viruses
	ssRNA¶ (immune complexes)	Endogenous
	Resiquimod	Synthetic
	siRNA#	Synthetic
TLR8	Resiquimod	Synthetic
	ssRNA¶ (viral)	Viruses
	ssRNA¶ (immune complexes)	Endogenous
TLR9	Unmethylated CpG DNA	Viral and bacterial
	DNA	Viruses
	DNA (immune complexes)	Endogenous

* HMGB-1: High mobility group box 1
† dsRNA: Double-stranded RNA
‡ LPS: Lipopolysaccharide
§ HSP: Heat shock protein
¶ ssRNA: Single-stranded RNA
# siRNA: Small interfering RNA
Note: Adapted from O'Reilly (2014)

and IRF-7 are responsible for stimulating genes that produce the antiviral cytokines, type I IFNs (IFN-α, IFN-β and IFN-ω), which play crucial roles in antiviral responses of the innate immune system (Abbas *et al.*, 2015, 2018).

There are unique and shared downstream consequences of TLR signalling resulting in inflammation and antiviral responses (Figure 1.7). These are accomplished through various signalling intermediates in combination with adaptor proteins myD88 and TRIF (TIR domain-containing adaptor inducing IFN-β) (Abbas *et al.*, 2018). The outer cell membrane TLRs (TLR1, TLR2, TLR5, TLR6 and TLR4) all signal through MyD88, leading to the activation of NF-κB and AP-1 which induces an inflammatory response. In addition to signalling through myD88, TLR4 is also able to signal through TRIF, leading to the activation of IFRs (IFR-3 and IFR-7) and the production of type I IFNs responsible for antiviral responses. Thus, TLR4 is able to induce both inflammatory and antiviral responses. TLR3 only signals through TRIF, and is therefore only able to activate IRFs which stimulates type 1 IFNs leading to an antiviral response. TLR7 and TLR9 are able to induce both types of responses through a MyD88-dependent, TRIF-independent signalling that induces NF-κB and IRFs. Thus, like TLR4, TLR7 and TLR9 are able to induce both inflammatory and antiviral responses (Abbas *et al.*, 2018).

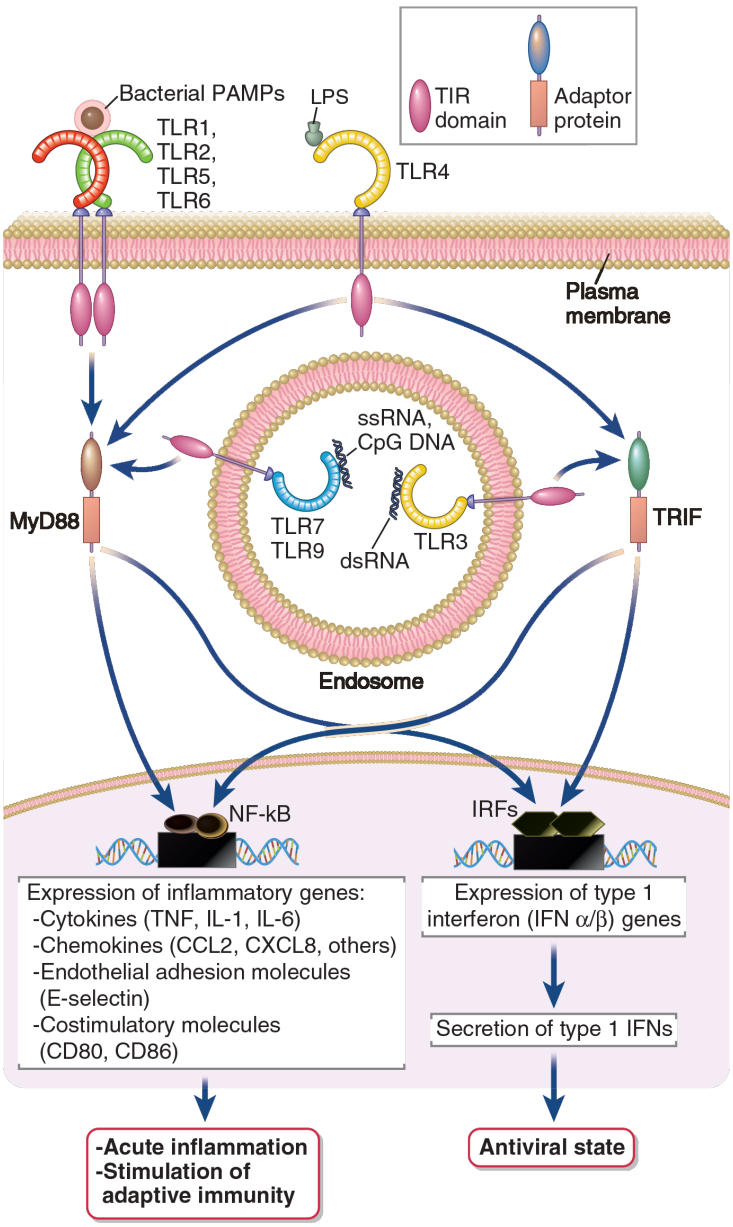


Figure 1.7: TLR signalling through MyD88 and TRIF adaptor proteins. *Note:* From Abbas *et al.* (2018)

Even though the molecular composition of DAMPs and PAMPs share no similarities and are a diverse set of molecules, they all trigger inflammatory and antiviral responses through interacting with TLRs and activation of the NF-κB, AP-1 and IRF transcription factors (Dowson *et al.*, 2017). It is speculated that in the early stages of SSc, DAMPs are released due to the persistent injury that occurs to the endothelium, and that the abnormalities in the TLR signalling somehow contribute to the onset and progression of the disease. TLRs ligands that have been described for SSc are of both microbial and endogenous origin (Table 1.4). Dowson *et al.* (2017) hypothesise that the fibrosis observed in SSc is mainly due to the compromised resolution of inflammation and wound healing that results from the initial immune response to the release of mediators caused by the initial damage to the epithelial cells of the vascular system. The immune system's failure to regulate normal wound healing and tissue repair-processes manifests in chronic scar

formation (Katsumoto *et al.*, 2011).

1.5.2.2 Interferon signature in SSc

As discussed in the previous section (Section 1.5.2.1), the stimulation of TLRs involved in SSc, through adaptor protein TRIF signalling, activates the IFRs transcription factors which induce production of type I IFNs (IFN-α and IFN-β). In SSc, there is an evidence of a prominent IFN signature, which can be directly linked to the activation of TLRs by either PAMPs and/or DAMPs (Katsumoto *et al.*, 2011; Pattanaik *et al.*, 2015; Varga, 2017). IFNs are multifunctional cytokines and early mediators of the innate immune system that play crucial roles in the resistance to viruses. They have the capacity to directly and indirectly influence the adaptive immune responses through dendritic cells, lymphocytes and NK cells (Katsumoto *et al.*, 2011). The IFN signature observed in SSc is similar to the IFN signature first described in SLE, where the active stages of the SLE disease display a significant upregulation of IFN-responsive genes, suggesting a common underlying disease mechanism.

As with SLE, the serum of SLE patients contains circulating immune complexes of DNA, RNA and antibodies to nucleic acids (autoantibodies), all of which serve as endogenous ligands for TLRs that promote type I IFN secretion by dendritic cells, monocytes and macrophages (Katsumoto *et al.*, 2011; Pattanaik *et al.*, 2015; Varga, 2017). This ultimately leads the production of pro-inflammatory cytokines and chemokines, presentation of antigens and stimulation of the adaptive immune response. The prominent IFN signature observed in SSc is due to the potent response of dendritic cells to TLR stimulations, which produces large amounts of type I IFNs (Katsumoto *et al.*, 2011). The IFN signature in SSc is also highly correlated with the presence of ATA and anti-U1-RNP (SSc autoantibodies discussed in Chapter 1.4). However, the exact role played type I IFNs in the pathogenesis of SSc is still poorly understood and intensive research into understanding its role is needed.

1.5.2.3 Adaptive immunity in SSc

The innate and adaptive immune systems are not mutually exclusive immune systems; there is a continuous interplay between the two systems. Dendritic cells play a key role in linking the innate immune responses to the adaptive immune responses through the TLR-mediated recruitment of antigens from PAMPs and DAMPs. These antigens are processed through TLR-mediated signalling and presented to T-lymphocytes, along with cytokines and chemokines that stimulate T lymphocyte differentiation (Pattanaik *et al.*, 2015). In SSc patients, adaptive immune responses manifest as serum autoantibodies (discussed in Chapter 1.4) with varying antigen specificities and their direct role in the pathogenesis of the disease is inconclusive (Varga, 2017). The levels of the autoantibodies in SSc patients (ATA in particular) are thought to vary with the disease severity and correlate with the skin and lung fibrosis. It is hypothesised that autoantibodies are generated when

self-antigens (e.g. DNA topoisomerase 1) undergo proteolytic cleavage in the presence of reactive oxygen species (ROS), thus exposing epitopes that compromise immune tolerance. Another mechanism suggests that endogenous DAMPs containing nucleic acids interact with TLRs, which activates B-lymphocytes to produce autoantibodies (Varga, 2017). The exact mechanism remains unclear, and whether autoantibodies are produced in the early stages before fibrosis, or as a result of fibrosis, still remains to be established.

1.5.3 Fibrosis

The connective tissue is made up of fibroblasts, myofibroblasts (a modified fibroblast exhibiting features of smooth muscle cells) and infiltrating cells embedded in an extra-cellular matrix (ECM) composed of structural proteins. These structural proteins include collagen, fibronectin, elastin, microfibrils as well as adhesive proteins (Varga, 2017). The ECM compartment serves as a source for growth factors (e.g. TGF-β) and other matricellular proteins (e.g. connective tissue growth factor [CTGF/CCN2]) responsible for controlling the differentiation, function and survival of mesenchymal cells (Varga and Trojanowska, 2008). Excessive overproduction and accumulation of ECM proteins, coordinated by the fibroblasts and myofibroblasts in response to cytokines and chemokines produced by immune responses during tissue damage, results in fibrosis, which is characterised by the replacement of normal tissue with dense/stiff connective tissue. Fibroblasts play crucial roles in tissue contraction and remodelling during wound healing (Katsumoto *et al.*, 2011; Varga, 2017). In SSc patients, fibrosis is a pathologic hallmark of the disease, with widespread manifestations involving the skin, multiple internal organs (including lungs and heart), tendons and ligaments.

Under normal physiological conditions and appropriate extra-cellular cues, fibroblasts (or their progenitor cells) are stimulated to produce ECM matricellular proteins, growth factors, signalling molecules and proteolytic enzymes; express surface receptors for these growth factors and signalling molecules; adhere to and contract connective tissue; and undergo trans-differentiation into myofibroblasts. All these properties (biosynthetic, pro-inflammatory, contractile and adhesive) allow fibroblast to successfully mediate effective wound healing (Katsumoto *et al.*, 2011; Varga, 2017; Varga and Trojanowska, 2008). The pathologic fibrosis seen in SSc patients is due to the failure to regulate normal tissue tissue-repair process, characterised by uncontrolled activation of fibroblasts which results in excessive ECM accumulation and tissue remodelling. A major molecular determinant implicated in normal wound healing and tissue repair, as well as the pathologic fibrosis seen in SSc patients is the cytokine TGF-β (Desbois and Cacoub, 2016; Varga, 2017). SSc patients have been found to have elevated levels of TGF-β in lung and skin tissues. Platelet derived growth factor (PDGF-R) and CTGF have also been implicated to play roles in the late stages of fibrosis in the disease. As with other two hallmarks of SSc discussed in Section 1.5.1 and 1.5.2, the exact mechanism that regulates the excessive ECM accumulation resulting in fibrosis is also unknown.

1.5.4 Pathogenic mechanism of SSc

The previous sections discussed the current knowledge on the three hallmarks of SSc, i.e., vasculopathy (Section 1.5.1), immune dysregulation (Section 1.5.2) and fibrosis (Section 1.5.3), how they are related to each other and their contribution to the onset and progression of the disease. However, even thought the mechanisms of these three hallmarks are poorly understood, this section attempts to summarise the current knowledge of the pathogenic mechanism of SSc (Figure 1.8), as proposed by Katsumoto *et al.* (2011) and based on the discussion of the previous sections, in four steps. *Step 1:* The disease is initiated, in a genetically predisposed individual, at a vascular level, by either infectious agents/pathogens, viral infections or autoantibodies against the epithelial cells. This causes damage to the endothelial cells and apoptosis, which ultimately leads to leaks in the small blood vessels. In this step, the genetic and environmental factors play a role at initiating the disease.

Step 2: The impaired angiogenesis and vasculogenesis causes an imbalance in the vasoactive molecules, whereby vasoconstrictors are up-regulated and vasodilators are down-regulated. This cause aggregation of platelets/thrombocytes and blood vessel constriction causing the formation of intravascular thrombosis. *Step 3:* DAMPs and/or PAMPs activate the innate immune responses through TLRs leading to the release of NF-κB and IRF transcription factors, that initiate inflammation and antiviral responses, respectively. There is also an increase production of type I IFN that stimulates the adaptive immune responses and production of autoantibodies through lymphocytes. *Step 4:* The fibroblasts in the ECM are recruited and stimulated to differentiate into myofibroblast by the cytokines and chemokines produced by the innate responses. Dysregulation of TGF-β signalling in the fibroblast and myofibroblasts causes a failure to control normal tissue repair process. This leads to excessive ECM accumulation and tissue remodelling, which ultimately cause the pathologic fibrosis observed in SSc patients.

1.6 Environmental risk factors in SSc

A number of environmental triggers have been implicated in SSc, which are though to be responsible for initiating the events that eventually lead to the development of the disease in a genetically predisposed individual. These include infectious agents (e.g. pathogens and viruses) environmental toxins, organic solvents, chemical compounds (e.g, silica exposure) and certain drugs (Arron *et al.*, 2014; Marie and Gehanno, 2015; Murdaca *et al.*, 2016). These environmental triggers remain elusive despite the health burden SSc causes on a global scale. In this section, the current understanding of the environmental factors and how they are associated with SSc are discussed.

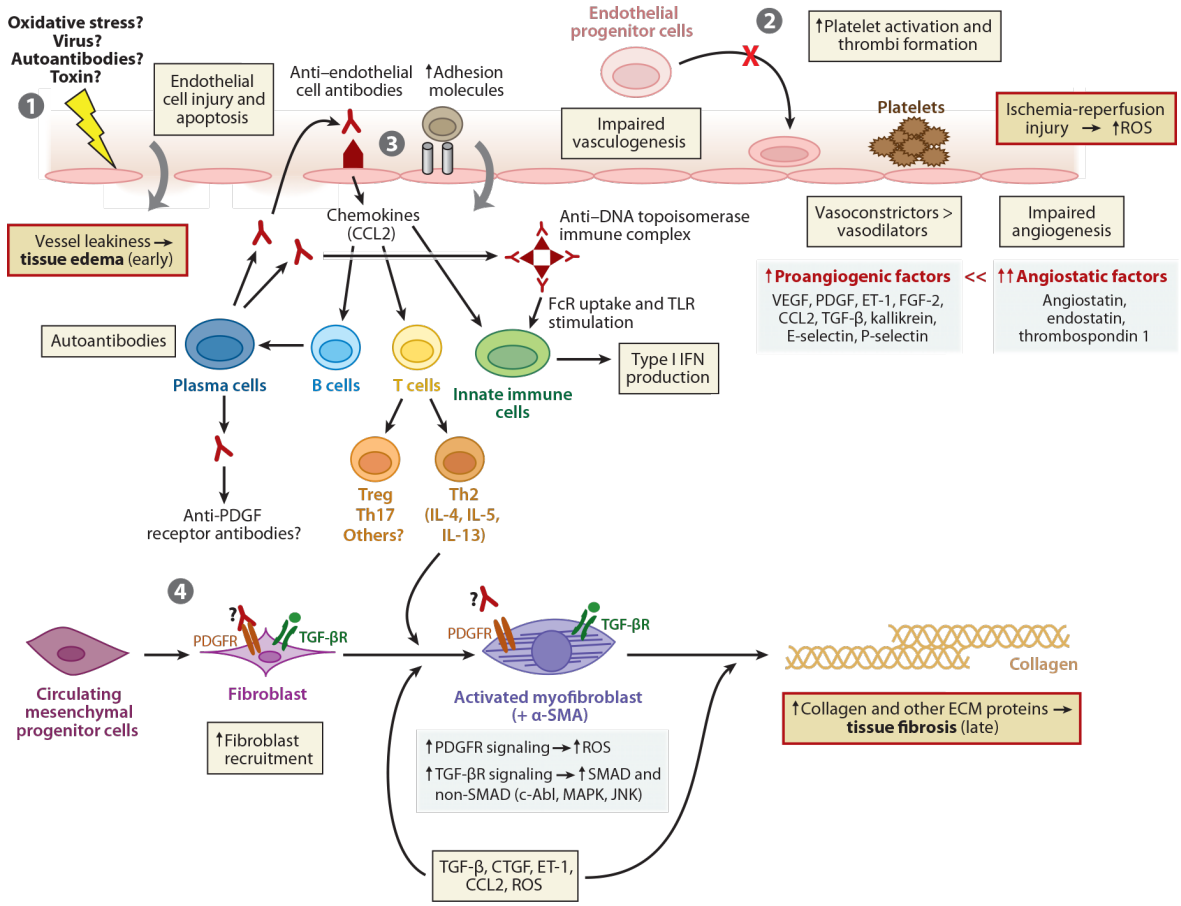


Figure 1.8: Proposed pathogenic mechanism of SSc. The mechanism can be summarised into four steps, i.e., initiation, vasculopathy, immune system dysregulation and fibrosis. *Note:* From Katsumoto *et al.* (2011)

1.6.1 Crystalline silica

The exposure to crystalline silica has been linked to the environmental risk of developing a number of rheumatic autoimmune diseases, including RA, SLE and SSc (Costenbader *et al.*, 2012; Marie and Gehanno, 2015). Exposure to silica seems to be more of an occupational exposure as the particulate silica is release in uranium, gold mining, asbestos and also during rock drilling and most scouring powder factory works. This occupational exposure to crystalline silica, in relation to SSc, seems to be more elevated in males as compared to females, possibly due to the fact that males are more commonly exposed (mostly males work in silica exposed environments) or there must be some biological sex determining factor that makes males more predisposed to develop SSc due to crystalline silica exposure (Costenbader *et al.*, 2012). It is suggested that because silica is a strong T lymphocyte adjuvant, genetically predisposed individuals exposed to crystalline silica develop dysregulated immune responses that trigger tissue damage and pathogenic fibrosis (Marie and Gehanno, 2015).

1.6.2 Pathogens and infectious agents

The existence of an IFN signature in SSc patients that can be linked to the activation of the innate immune system (as discussed in Section 1.5.2.1) validates the suspicion that infec-

tious agents are responsible for triggering SSc. Studies have linked a number of pathogens with SSc, including EpsteinBarr virus (EBV, also known as *Human gammaherpesvirus 4*), *Toxoplasma gondii*, cytomegalovirus (CMV), *Rhodotorula glutinis*, endogenous retroviruses, *Helicobacter pylori* and Chlamydia (*Chlamydia trachomatis*) (Arron *et al.*, 2014; Grossman *et al.*, 2011; Katsumoto *et al.*, 2011; Marie and Gehanno, 2015; Murdaca *et al.*, 2016). However, the insufficient evidence of these findings cannot prove that infectious agents and pathogens play a significant role in the onset and progression of SSc. The presence of these pathogens and infectious agents in SSc could also be a result of opportunistic infection caused by a compromised immune system. A number of hypotheses have been proposed, which link pathogens and infectious agent infections with the onset of SSc. These include: (1) molecular mimicry; (2) endothelial cell damage; and (3) super-antigens (Grossman *et al.*, 2011; Radić, 2014).

The basis of the mechanism of molecular mimicry is the homology that exist between viral/microbial peptides and autoantigens, and it explains the pathogenicity of antibodies against viral/microbial proteins in SSc. Autoreactive T-lymphocytes could be activated by the viral/microbial peptides that are structurally similar to autoantigens, which can induce endothelial cell apoptosis (Grossman *et al.*, 2011). Super-antigens are proteins that are derived either exogenously by bacteria or expressed endogenously by the host organism, and are thought to stimulate intermediate T-lymphocytes. It has also been shown that B-lymphocytes responses could potentially bind to microbial super-antigens to the surface class II major histocompatibility complex (MHC) molecules and become targets of T-helper lymphocytes (Radić, 2014). Damage to the endothelial cells could be initiated by pathogen infections, which could be followed by cell abnormalities and necrosis and stimulation of immune responses (Grossman *et al.*, 2011; Varga, 2017).

1.7 Genetic risk factors in SSc

As with most complex autoimmune diseases, the aetiology of SSc is poorly understood. Even though it is widely accepted that environmental factors play a role in the onset and progression of SSc, as discussed in the previous section (Section 1.6), the knowledge of environmental risks factors in SSc is not sufficient to draw valid conclusions. Findings from twin and family studies, and studies in different ethnic groups, support a significant relevance of genetic risk factors in the development and progression of SSc (Radstake *et al.*, 2010). In this section, the current studies aimed at understanding the genetic risk factors associated with the onset and progression of SSc are discussed.

1.7.1 Twin Studies

SSc is a complex disease, thus it does not follow the simple Mendelian rules of inheritance as seen in monogenic disorders; there is an interplay between multiple genes and/or the environment, i.e., the susceptibility alleles will collude with the environment to increase

the risk of developing a disease (Allanore *et al.*, 2010). Twin studies are a powerful approach in complex, multi-factorial trait diseases as they allow for the distinction of whether both genetic and environmental factors contribute towards the development and progression of a disease (Bogdanos *et al.*, 2012). This can be quantified through the disease concordance between between monozygotic (MZ) and dizygotic (DZ) twins. The assumptions behind the disease concordance between MZ and DZ twins is that MZ twins share 100% of their DNA, thus if the phenotype concordance of the DZ twins is lower than that of MZ twins, this would indicate that heritability or genetic factors are important in the disease, whereas lower phenotype concordance of MZ compared to DZ would be suggestive of non-genetic factors playing an important role (Bogdanos *et al.*, 2012). Twin studies are limited by small sample sizes and the fact that twins often have a very similar environment when reared together.

A number of twin studies have been carried out in complex autoimmune disease, including SSc, to identify whether SSc is influences by genetic factors and to what extent they play a role. Feghali-Bostwick *et al.* (2003) examined the concordance of SSc in 42 pairs of twins (24 MZ and 18 DZ), in which 17 sets had dcSSc, 18 sets had lcSSc, 5 had dcSSc/lcSSc overlap with another connective tissue disease (CTD) and 2 sets were unconfirmed. They found that of the 42 pairs, 2 female pairs were concordant; 1 DZ pair (both with lcSSc) and 1 MZ pair (1 lcSSc and 1 with dcSSc). The concordance rate for this study was similar, i.e., 4.2% for MZ twins and 5.6% for DZ twins, with an overall concordance rate of 4.7%. These results suggest that the triggers for SSc are likely have a strong environmental contribution (Feghali-Bostwick *et al.*, 2003).

Zhou *et al.* (2005) performed a microarray study from RNA extracted from cultured dermal fibrosis from 10 discordant MZ and 5 discordant DZ twins and 5 normal healthy controls of similar ages. In analysing the gene expression profiles, Zhou *et al.* (2005) found that there was a 40-50% concordance in MZ twins discordant for SSc, but no concordance for gene expression in DZ discordant for SSc. These findings suggest that the unaffected MZ twins have a ∼40-50% chance of developing SSc as MZ twins share 100% of their DNA material. This suggests that there is a rather strong genetic predisposition to SSc at gene expression level. The variability of the concordance rates reported between the MZ and DZ twin pose a challenge for SSc twin studies.

1.7.2 Candidate gene studies

A number of candidate-gene studies were carried out on genes and single nucleotide polymorphisms (SNPs). These genes and SNPs were selected based on their involvement with other autoimmune diseases. A case-control study by Tsuchiya *et al.* (2009) was carried out on 282 patients with SSc and 590 controls to examine the association of a SNP (rs7574865) found on the signal transducer and activator of transcription 4 (*STAT4*) gene with SSc. STAT4 is a transcription factor that induces IL-12, IL-23 and type 1 IFN cytokine signals

in T-lymphocytes and monocytes. The transduction of these signals results in type 1 and type 17 T-helper differentiation, monocyte activation and IFN-γ production (Korman *et al.*, 2008; Tsuchiya *et al.*, 2009). The study by Tsuchiya *et al.* (2009) showed that the *STAT4* gene is associated with SSc and that it is a common susceptibility gene in the Caucasian and Asian populations.

1.7.3 Genome wide association studies

Genome-wide association studies (GWAS) have also been used to identify a number of SNPs involved in the susceptibility to SSc. By performing GWAS on 137 patients with SSc and 564 controls, Zhou *et al.* (2009) have just showed that SNPs near the human leukocyte antigen (HLA) regions of chromosome 6, i.e., *HLA-DPB1* and *HLA-DPB2*, were strongly associated with SSc susceptibility. Another GWAS study by Radstake *et al.* (2010) identified a new SSc susceptibility locus, *CD247*, in their investigation involving 2296 SSc patients and 5171 healthy individuals. This study further confirmed the role of the *HLA* region and the *STAT4* gene in the genetic predisposition to SSc, which are loci known to be associated with risk in other autoimmune conditions.

1.7.4 Gene expression studies

Despite the intensive research into SSc and the identification of genes and cell types that are more likely to be involved in SSc, the aetiology and pathogenesis of the disease remains unclear, especially in the black South Africans. One approach to examining the aetiology of SSc is to utilise RNA-seq data to study gene expression in tissue samples from SSc patients (both affected and unaffected tissue samples) and compare it to healthy skin samples from control individuals. In living organisms, the DNA contained in the nucleus of cells encodes information that determines the properties and functions of each cell. Through the process of "gene expression", this information in the DNA is accessed and transcribed into RNA molecules, which in turn are translated into proteins that carry out functions in the cell and determine their properties (Finotello and Di Camillo, 2015). Therefore, the RNA molecules, or transcriptome, that are transcribed in a particular disease condition, and at a particular time, can reveal the underlying mechanism of the disease when compared to normal conditions. However, regulation of gene expression is complex and involves many different mechanisms including epigenetic modifications. RNA expression does not necessarily indicate how much protein is present in the cell, but it is one step in the process of understanding the disease and the mechanism leading to disease.

High-throughput RNA-seq technology has enabled the study of the function of the genome, revealing information on gene expression in specific cells and tissues, at defined time points, at a remarkable scale and speed in a cost-effective manner (Haas *et al.*, 2013; Mele *et al.*, 2015; Trapnell *et al.*, 2012). Using RNA-seq data (mRNA and sRNA), we are able to identify transcripts, quantify their expression and perform differential gene/transcript

expression analysis. To perform differential gene/transcript expression, using RNA-seq data, the RNA molecules have to be first obtained from tissue samples (both normal and diseased), fragmented and reverse-transcribed into complementary DNA (cDNA). The cDNA is then amplified and subjected to high-throughput sequencing (HTS) on NGS platforms to produce billions of short read sequences (Costa-Silva *et al.*, 2017; Finotello and Di Camillo, 2015).

These short reads are then mapped to reference genomes/transcriptomes to identify the genes that the reads belong to. "Read counts" are then generated by counting the number of reads aligning to each gene, which gives a measure of the level of expression for a particular gene under the conditions being studied. Finally, the read counts are normalised across different samples and conditions by applying statistical models and differentially expressed genes are identified (Costa-Silva *et al.*, 2017; Finotello and Di Camillo, 2015). The genes identified are annotated and analysed for their biological relevance. RNA-seq data could be used to discover new SSc-associated genes and transcripts, study transcript structure (alternative splicing) and identify allelic information (SNPs) in a single assay (Grabherr *et al.*, 2011; Haas *et al.*, 2013; Trapnell *et al.*, 2012).

Expression profiling studies have identified differences in micro RNA (miRNA) levels in skin and fibroblasts cultured *ex vivo* from SSc patients and healthy controls. miRNA are small non-coding RNA molecules of 15 to 25 nucleotides in length. They play major roles in the regulation of gene expression at the transcriptional and/or post-transcriptional level by binding to the 3' untranslated regions of target messenger RNAs (mRNAs), thus inhibiting their translation or inducing destabilisation and degradation. Expression profiling studies have also identified that the differentially expressed miRNAs have a role in SSc by altering the gene expression of fibrosis-related genes and their protein products (Li *et al.*, 2012; Maurer *et al.*, 2010; Zhu *et al.*, 2013). In their study, Zhu *et al.* (2013) demonstrated that TGF-β plays a fundamental role in SSc by regulating the miRNA mR-21 and fibrosis-related genes. They found that upregulation of mR-21 was induced by TGF-β, which in turn promoted TGF-β-regulated fibrogenic activation of skin fibroblasts by targeting SMAD7 (mothers against decapentaplegic homolog 7).

1.8 SSc in African Populations

Cases of SSc in black African populations are rarely reported, and when they are reported, they are usually single-case reports (Adelowo and Oguntona, 2009). Most of the cases of SSc reported in SSc in Africa show higher incidences of the disease in South African black populations. Cowie and Dansey (1990) reported 24 cases of black South Africans with SSc, all of whom were goldminers with pulmonary complications. In another survey study by Tager and Tikly (1999), 63 cases of black South Africans were reported with a clinical and laboratory manifestations of the disease, including Raynaud's phenomenon, myositis, pulmonary fibrosis, pulmonary hypertension, renal dysfunction and abnormal

lung function. In Nigeria, Adelowo and Oguntona (2009) reported 14 cases, most of which were females and had the dcSSc form of the disease.

1.9 Rationale and Motivation for the Study

Recent years have seen the rate at which sequencing data produced by the improved NGS sequencing technology and reduced costs grow exponentially. This has allowed researchers to perform "multi-omic" data analyses, that are now becoming routine, to answer many biological questions. However, there are challenges to performing analyses of the large datasets produced by the NGS sequencing technology. In most cases, analysis of NGS data requires many applications, which require input/output to be efficiently coordinated between them. These analyses are also computationally intensive, and most often require resources that are not available in most research institutes. Computational pipelines are a solution to these challenges and can be used to automate the repetitive tasks associated with day-to-day NGS data analysis. They also allow analysis to be scaled to different computational platforms, and also shared with other researchers.

The availability of RNA-seq data from black South African patients affected with SSc and healthy individuals from the study by Frost *et al.* (2018) presented an opportunity develop robust computational pipelines in an effort to bridge the gap between repetitive (and most often complicated) bioinformatic analyses and the large datasets produced by NGS sequencing technologies. Two bioinformatic analyses were of importance in this study, i.e., differential gene expression and metagenomic analyses, and these were used in conjunction with the RNA-seq data to demonstrate the value of the workflows as well as to obtain biological insights into SSc in order to contribute towards the current understanding of SSc, especially in black South African populations.

1.10 Study Aims and Objectives

This project aims to develop pipelines to analyse data to contribute to the understanding of the molecular aetiology and genetic variation of SSc through various comparative bioinformatic analyses of transcriptome data from black South African SSc patients and healthy unaffected individuals. The main objectives of this study are as follows:

Objective 1: To develop a novel pipeline for mapping raw mRNA reads to reference genomes and quantifying transcripts using the RNA-seq data.

Objective 2: To identify significantly differentially expressed transcripts/genes and non-coding RNAs and examine them for novel disease association.

Objective 3: To identify pathways that influence the onset and severity of SSc and to identify potential biomarkers.

Objective 4: To develop a novel pipeline for metagenomic analyses and use it to identify

potential microorganisms/pathogens associated with SSc.

The overall structure of this thesis is summarised in Figure 1.9. Chapter 2 presents the RNA-seq data used in this study and the pre-processing methods applied. Chapter 3 addresses objective 1; the development of a reproducible `Nextflow` pipeline for mapping raw RNA-seq reads to reference genomes and quantification of transcripts using the pre-processed data. Chapter 4 addresses objectives 2 and 3, the identification of differentially expressed genes and pathways using the data generated from the pipeline designed in Chapter 3. Chapter 5 addresses objective 4; the metagenomic analyses of the pre-processed RNA-seq data and the development of a reproducible `Nextflow` pipeline for metagenomic analyses. Finally, Chapter 6 discusses the significance of the findings from this study by bringing together the results from the study aims and objectives.

1.11 Limitations of this study

The study presented here is unique in the sense that it is the first study to use RNA-seq data generated from black South Africans in an effort to understand the aetiology of SSc through differential expression and metagenomic analyses. However, as the RNA-seq data were generated for a different study (presented in Chapter 2), by Frost (2016), to achieve a different goal, I had no control over the study design, selection of participants (including the sex, age and number of participants in the case and control groups), sampling of tissue as well as the library preparation method used.

The objective of this study was to develop pipelines to aid in the analyses of RNA-seq data. To illustrate the value of these pipelines, the SSc data was used, in addition

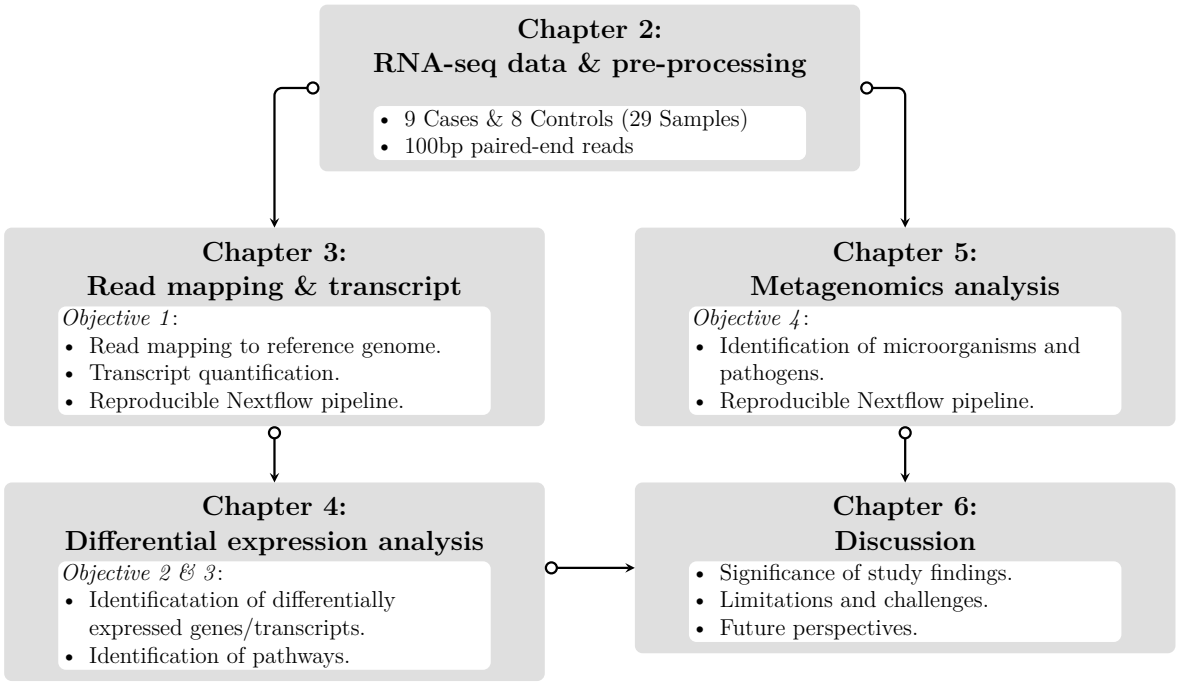


Figure 1.9: Flow diagram summarising the overall structure of the thesis.

to seeking biological insights from the data despite the limitations in the study design. These limitations of the study are highlighted below.

- ***Sample size*** The small sample size (9 cases and 8 controls) for the available data limit the power of the study to detect robust associations with small effects of differential expression. RNA-seq data can be used for a wide variety of analyses, which makes experimental design very complex. This in turn complicates the statistical power calculation. The major factors that influence the statistical power of detecting differential expression in RNA-seq data include the effect size, within-group variation, average sequencing count levels, natural variation of gene expression, type I error control (P-value) and distribution of differential expression (Wu *et al.*, 2015). These factors were taken into account in this study by performing a power calculation to confirm if the sample size is adequate to detect differentially expressed genes.

- ***RNA-seq library preparation*** Library preparation to generate these data was based on RNA selected for the presence of the polyadenylated (poly-A) tail. This step ensures that the ribosomal RNA (rRNA) and other small RNA molecules do not get sequenced along with the mRNA (Conesa *et al.*, 2016). This posed a problem for Objective 4 that is a metagenomic analysis since bacterial mRNA does not contain poly-A tails.

- ***Cell types in skin biopsies*** The skin biopsies from the different sites (forearm, back and breast) may have tissue specific differences in gene expression and each biopsy is a combination of different cell types (e.g. keratinocytes, melanocytes, blood capilary epithelium, etc.). Expression of genes in these tissues is expected to be somewhat different even though they are all skin biopsies.

- ***Unmatched age and sex*** The participants (cases and controls) in the study include both males and females, with differing ages. These differences limit the ability of the study to adequately address potential sex and age-related differences.

Chapter 2

RNA-seq Data and Pre-processing

Bioinformatic analyses on raw sequencing data require pre-processing to remove any artefacts and low quality reads that may arise due to sequencing errors and sequence library preparations. This step is crucial in all analyses as presence of artefacts in the data may cause sub-optimal results in downstream analyses. In this chapter, the RNA-seq data from black South African SSc patients and healthy individuals are introduced, including the participants' demographic and clinical information. The steps that were taken in order to prepare the data for bioinformatic analyses performed in this study are discussed.

2.1 Introduction

The transcriptome data for this study was generated from a previous PhD research study by Frost (2016). The study was conducted with ethics approval of the Human Research Ethics Committee (HREC [Medical]) of the University of the Witwatersrand (certificate number M120512). Briefly, nine consenting black South African patients with SSc were recruited from the Chris Hani Baragwanath Hospital. The patients met the classification criteria for SSc and had less than five years disease duration. From each of the nine SSc patients, two 4mm punch skin biopsies were taken: one from the forearm (affected skin) and another from the back (unaffected skin). Control samples were obtained from eight consenting healthy black South African individuals recruited from the Helen Joseph Hospital. The control individuals were undergoing reconstructive plastic surgery. From each of the eight healthy controls, one 4mm skin biopsy was taken from the breast (four individuals) and forearm (four individuals). The skin biopsies from both the patients and controls were then placed in 15ml tubes with RNAlater and stored at -20°C in preparation for RNA extraction, which was carried out within six months of sample collection at the Centre for Genomic Regulation (CRG) in Barcelona, Spain (Frost, 2016).

The RNA was extracted from the tissue (after homogenisation) and purified using the Qiagen's RNeasy miRNA kit. The quality and quantity of the RNA for each sample was assessed: first by determining the absorbance ratios at 260nm and 280nm, then by determining the RNA integrity number (RIN) scores to assess the quality using a Bioanalyser. Samples were accepted for library preparation if they had an absorbance ratio of approximately ~2 (260/280) and a RIN score above 7 (RIN scores range between 1 and 10, where 1 is the lowest quality and 10 is highest quality). The Illumina TruSeq Stranded mRNA preparation protocol was used to prepare mRNA sequencing libraries. 100bp paired-end reads were produced for each sample using the Illumina's HiSeq sequencing

platform (Frost, 2016). Ethical clearance (certificate number M1710104) was obtained from the HREC (Medical) of the University of the Witwatersrand (Wits) to conduct a sub-study using this data (Appendix A). The summary for the mRNA sequencing data and clinical information for the samples is shown on Table 2.1.

In Table 2.1, the patient samples are labelled "P" (skin biopsies taken from the diseased forearm) and "B" (skin biopsies taken from the unaffected back). Each patient has two samples, e.g., patient 1 has a forearm sample P1 and a back sample B1. For the controls, the samples are labelled "C" (skin biopsies from the forearm) and "R" (skin biopsies from the breast). Table 2.1 also shows that the ratio of males to females in the patient group is 1:8, whilst that of the healthy control individuals is 3:5. However, for the control individual R4, the sample was discarded due to the inability to amplify cDNA, which brings the ratio of males to females to 3:4. The age of the SSc patients ranges from 48 to

Table 2.1: Summary of the RNA-seq sequencing and clinical data

	Sample	# of Reads	Sex	Age	Clinical Information					
					Year of onset	Disease duration (months)	MRSS	Lung Disease	Myositis	Anti-Scl 70
SSc Patients*	P1	65 069 646	F	48	2007	60	13	-	-	-
	B1	45 438 576								
	P2	67 917 276	F	54	2009	36	34	True	True	True
	B2	52 382 324								
	P3	44 047 332	F	48	2010	36	4	False	False	False
	B3	67 477 960								
	P4	64 344 432	F	64	2007	60	9	False	True	False
	B4	62 788 060								
	P5	64 299 320	F	49	2010	36	16	True	False	False
	B5	55 418 906								
	P6	43 815 618	F	55	2010	36	32	True	True	True
	B6	58 289 942								
	P7	84 023 154	F	58	2011	24	5	False	False	False
	B7	56 555 144								
	P8	66 143 818	F	46	2012	12	25	True	False	True
	B8	63 706 182								
	P9	56 968 454	M	64	2012	12	22	False	False	False
	B9	58 639 158								
Controls†	C1	43 143 926	F	29	-	-	-	-	-	-
	C2	67 608 242	M	33	-	-	-	-	-	-
	R3	61 565 754	F	40	-	-	-	-	-	-
	R4§	-	F	35	-	-	-	-	-	-
	R5	66 871 166	F	38	-	-	-	-	-	-
	C6	58 409 290	M	31	-	-	-	-	-	-
	C7	72 226 420	M	34	-	-	-	-	-	-
	R8	58 219 818	F	30	-	-	-	-	-	-
Reps	B1‡	36 268 612	F	48	-	-	-	-	-	-
	P3‡	38 760 700	F	48	-	-	-	-	-	-
	P6‡	37 409 158	F	55	-	-	-	-	-	-
	C1‡	32 313 602	F	29	-	-	-	-	-	-
Total:		1 650 121 990								

* Skin biopsy taken from the diseased forearms ("P") and unaffected back ("B") of patients.
† Skin biopsy taken from the forearms ("C") and breasts ("R") of controls.
‡ Sample resequenced due to low number of reads.
§ Sample discarded due to inability to amplify cDNA.
year of onset: year the disease started; ***disease duration***: duration at the disease at the time of sample collection; ***MRSS (Modified Rodnan skin score)***: assessment of the severity of subcutaneous skin thickening, which ranges from 0 (normal) to 3 (severe) at each of the 17 sites to yield a maximum score of 51; ***lung disease***: evidence of lung fibrosis from chest radiographs or high resolution computed tomography of the lungs; ***myositis***: muscle weakness (muscle degeneration or inflammation); and ***anti-Scl-70***: antibody reactivity against topoisomerase I.

64 years, whilst that of the healthy control individuals ranges from 29 to 40 years.

Clinical information for the patients was obtained from Frost (2016) based on the relevance to this study (Table 2.1). The clinical information includes: (1) ***year of onset***, year the disease started; (2) ***disease duration***, duration at the disease at the time of sample collection; (3) ***MRSS (Modified Rodnan skin score)***, assessment of the severity of subcutaneous skin thickening, which ranges from 0 (normal) to 3 (severe) at each of the 17 sites to yield a maximum score of 51; (4) ***lung disease***, evidence of lung fibrosis from chest radiographs or high resolution computed tomography of the lungs; (5) ***myositis***, muscle weakness (muscle degeneration or inflammation); and (6) ***anti-Scl-70***, antibody reactivity against topoisomerase I (Frost, 2016). At the time of sample collection, all the patients were within five years of disease duration: two patients were at 12 months, one patient at 24 months, four patients at 36 months and two patients at 60 months. Six patients had an MRSS score below 25 and three had a score of 25 or above. Out of the nine patients, eight had information collected for lung disease, myositis and anti-Scl-70; four were positive for lung disease, three were positive for myositis and three were positive of anti-Scl-70. In addition to the clinical information, Table 2.1 also shows that the samples for B1, P3, P6 and C1 were resequenced due to low numbers of reads.

2.1.1 Quality control of RNA-seq data

As with all NGS data analysis, the first step in analysing RNA-seq data is to perform quality control (QC) checks on the sequencing data. This is to ensure that the data does not contain low quality reads and technical sequences (adapters). Poor quality reads and adapters in the sequencing data are a result of errors during the sequencing process (Bolger *et al.*, 2014; Conesa *et al.*, 2016). If these reads are not removed, they can result in poor mapping to the reference genome and sub-optimal downstream analyses. To remove the poor quality reads and adapters, they must first be identified. This can be accomplished by utilising the `FastQC` (`http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/`) program (Andrews, 2010).

The `FastQC` program contains different modules, each of which identifies potential problems in the raw sequencing data. The supported files that can be passed onto `FastQC` include FASTQ files, Sequence Alignment Map (SAM) files and Binary Alignment Map (BAM) files. Sequencing data can be passed to the `FastQC` program and a report is generated by each one of the modules. The `FastQC` report generated for each raw sequencing file contains information on: (1) basic statistics, (2) per base sequence quality, (3) per sequence quality scores, (4) per base sequence content, (5) per sequence GC content, (6) per base N content, (7) sequence length distribution, (8) duplicate sequences, (9) overrepresented sequences, (10) adapter content, (11) k-mer content and (12) per tile sequence quality. Each `FastQC` report will have either a green tick (normal/pass), an orange exclamation mark (abnormal) or a red cross (unusual/fail) next to each of the modules. This

gives users more of an overview as to which aspects of the data need attention. These can then be examined in order to make informed decisions as to how to proceed with QC of the sequencing data.

Once the regions of poor quality and contamination in the data have been identified, these must be removed or trimmed. One useful and popular tool for removing artefacts in sequencing data caused by sequencing errors is `Trimmomatic` (Bolger *et al.*, 2014). `Trimmomatic` is an efficient and flexible pre-processing tool that can handle pair-end and single-end Illumina NGS data. The main filtering and algorithms in `Trimmomatic` are: (1) identification and removal of technical sequences, and (2) filtering based on the quality of the bases in the reads. The advantage `Trimmomatic` has over other pre-processing tools is that when using the paired-end mode, the corresponding read pairs are maintained and the information in the read pairs assists in identifying technical sequences that may have been introduced during the preparation of sequencing libraries (Bolger *et al.*, 2014). After removal of technical sequences and poor quality bases in the data with `Trimmomatic`, `FastQC` can be used again to check if the data needs further pre-processing. When the data no longer needs further pre-processing, it can be then used for bioinformatic analyses.

In the following Analyses section (Section 2.2), the initial steps undertaken to prepare the RNA-seq data for the bioinformatic analyses carried out in this study are described. First, we look at the quality of the initial RNA-seq data using `FastQC` (Section 2.2.1), then describe methods used to pre-process poor quality reads and sequencing artefacts identified using `Trimmomatic` (Section 2.2.2), and finally the merging of duplicate samples (Section 2.2.3).

2.2 Analyses

This section gives a description of the pre-processing of the RNA-seq data used in this study before any analyses could be performed on the data. This was to ensure that all artefacts and sequencing errors were excluded in the data. All the analyses were carried out on the Wits Computer Cluster, through the UNIX CLI, where the data was stored. The applications used were either already installed on the cluster or installed locally.

2.2.1 Initial QC using FastQC

To assess the quality of the initial RNA-seq data obtained from Frost (2016), `FastQC` (Version 0.11.3) was used. The analysis were carried out on the Wits Cluster using the `BASH` script shown below. The script analysed all 58 FASTQ files (29 samples each with two FASTQ files, one forward reads and one reverse reads) and created an `HTML` report for each FASTQ file. To better analyse all 58 reports generated by `FastQC`, the `MultiQC` (Version 1.5, `http://multiqc.info/`) tool was used to aggregate all the reports and summarise the results into a single `HTML` report (Ewels *et al.*, 2016).

```
#!/bin/bash

#PBS -N FASTQC
#PBS -q WitsLong
#PBS -l nodes=1:ppn=11,walltime=24:00:00,mem=40G
#PBS -o /spaces/phelelani/ssc_data/data/logs/output.log
#PBS -e /spaces/phelelani/ssc_data/data/logs/error.log

DATA_DIR=/spaces/phelelani/ssc_data/data

cd $DATA_DIR

# Cases + Control QC before trimming
/home/phelelani/applications/FastQC/fastqc \
    $(ls *.fastq.gz) \
    --threads 10 \
    --outdir qc \
    --noextract
```

2.2.2 Removal of artefacts using Trimmomatic

Poor quality reads and technical sequences were filtered out using `Trimmomatic` (Version 0.32) (Bolger *et al.*, 2014). Since the data for this study was paired-end, the paired-end mode in `Trimmomatic` was used. Three filtering steps were used in the following order: (1) `ILLUMINACLIP`, (2) `TRAILING` and (3) `MINLEN` (`BASH` script shown below). First, `ILLUMINACLIP` option was used to remove any Illumina-specific technical sequences in the data. `Trimmomatic` provides a list of technical sequences used by Illumina for GAII machines (TruSeq2) and HiSeq/MiSeq machines (TruSeq3). These sequences can be passed to the `ILLUMINACLIP` option for both paired-end and single-end reads. Since the data for this study is paired-end and was produced on Illumina's HiSeq platform, the TruSeq3 sequences were used to remove technical sequences.

```
#!/bin/bash

#PBS -N TRIMMOMATIC
#PBS -q WitsLong
#PBS -l nodes=1:ppn=12,walltime=24:00:00,mem=120gb
#PBS -o /spaces/phelelani/ssc_data/data_trimmed/logs/output.log
#PBS -e /spaces/phelelani/ssc_data/data_trimmed/logs/error.log

DATA_DIR=/spaces/phelelani/ssc_data/data
OUT_DIR=/spaces/phelelani/ssc_data/data_trimmed
SOFTS=/opt/exp_soft/bioinf

cd $DATA_DIR

for the_read in $(ls *R1*)
do
    name_in=$the_read
    name_out=CLEAN_$(sed -n s@_R1_@_@p <<< "$the_read")

    java -jar $SOFTS/trinity/trinity-plugins/Trimmomatic-0.32/trimmomatic.jar PE \
         -threads 12 -phred33 \
         -trimlog $OUT_DIR/logs/report_${the_read:2:18}.log \
         -basein $name_in \
         -baseout $OUT_DIR/$name_out \
         ILLUMINACLIP:$OUT_DIR/TruSeq3-PE-2.fa:2:30:10:8:true \
         TRAILING:28 \
         MINLEN:40
done
```

The `TRAILING` option followed after the removal of technical sequences. Generally, the

quality of the bases drops towards the 3′ end of sequencing reads due to the side effects of the Illumina's sequencing system (Conesa *et al.*, 2016). To improve the quality of mapping in downstream analyses, these bases should be trimmed off the reads. The `TRAILING` option removed any low quality bases towards the 3′ end of the reads that fall below a quality threshold of 28. The final option, `MINLEN`, was used to remove any reads that were $\leq$ 40bp after trimming. The quality of the resulting reads was assessed using `FastQC`, and the reports summarised using `MultiQC` for better interpretation.

2.2.3 Merging of duplicated/resequenced samples

The samples (B1, P3, P6 and C1) that were resequenced due to low number of reads in the first sequencing assay were merged with their corresponding original samples (Table 2.1) using a simple "`zcat`" command in `BASH` for concatenating files.

2.3 Results

The QC results for the RNA-seq data used in this study are shown in Appendix B and Table 2.2. When assessing the quality of the initial data, all 29 samples (58 FASTQ files) passed the QC with an average per base sequence quality for the reads above 30 (Appendix B.1b). The samples also passed the per sequence GC content (Appendix B.1c), with an exception of sample B9. When looking at the presence of technical sequences, all the samples contained adapter sequences (Appendix B.1a). After removing artefacts in the data using `Trimmomatic`, no adapter contamination was present when assessing the data using `FastQC`. The reads passed the average per base sequence quality (Appendix B.1d) and per sequence GC content (Appendix B.1e), with an exception to samples B2 and B9, which both failed the per sequence GC content. After merging the four samples that were replicated, there was a total of 25 samples (50 FASTQ files). The merged samples (B1, P3, P6 and C1) passed the average base quality (Appendix B.1f) and GC content (Appendix B.1g). Table 2.2 shows that only a small fraction of reads was lost during the pre-processing step; only 1.52% (24 981 130) of the 1 650 121 990 initial reads were removed from the data in the pre-processing step.

2.4 Discussion

Raw data QC and pre-processing are crucial steps in all NGS data analysis, and are sometimes overlooked. If data is not rid of all artefacts that arise from sequencing errors, this may lead to problems in the downstream analyses and may also lead to sub-optimal results. The most useful attributes to check for in raw sequencing data are the sequence base quality, GC content and the presence of technical sequences in the raw reads (Bolger *et al.*, 2014; Guo *et al.*, 2014). `FastQC` is a widely and commonly used tool that identifies and creates visual reports for analysing such features of the raw data. Once analysed and irregularities have been identified, the challenge comes in having to exclude the irregularities whilst leaving the valid sequence data behind. This challenge is further heightened

Table 2.2: Summary of the RNA-seq data before and after pre-processing

Sample		# of reads	# of reads after each pre-processing step			
			after trimming	reads removed	% reads removed	after merging
SSc Patients*	P1	65 069 646	63 691 766	1 377 880	2.12	-
	B1	45 438 576	45 038 056	400 520	0.88	80 890 192
	P2	67 917 276	66 814 078	1 103 198	1.62	-
	B2	52 382 324	51 738 962	643 362	1.23	-
	P3	44 047 332	43 665 950	381 382	0.87	81 989 430
	B3	67 477 960	65 096 064	2 381 896	3.53	-
	P4	64 344 432	63 600 160	744 272	1.16	-
	B4	62 788 060	61 835 508	952 552	1.52	-
	P5	64 299 320	63 248 176	1 051 144	1.63	-
	B5	55 418 906	54 791 828	627 078	1.13	-
	P6	43 815 618	43 428 656	386 962	0.88	80 372 030
	B6	58 289 942	57 595 008	694 934	1.19	-
	P7	84 023 154	82 317 704	1 705 450	2.03	-
	B7	56 555 144	55 773 292	781 852	1.38	-
	P8	66 143 818	65 067 670	1 076 148	1.63	-
	B8	63 706 182	62 795 740	910 442	1.43	-
	P9	56 968 454	55 846 868	1 121 586	1.97	-
	B9	58 639 158	57 685 510	953 648	1.63	-
Controls†	C1	43 143 926	42 705 270	438 656	1.02	74 609 832
	C2	67 608 242	66 775 204	833 038	1.23	-
	R3	61 565 754	60 618 148	947 606	1.54	-
	R5	66 871 166	65 988 702	882 464	1.32	-
	C6	58 409 290	57 648 674	760 616	1.30	-
	C7	72 226 420	70 931 848	1 294 572	1.79	-
	R8	58 219 818	57418466	801352	1.38	-
Reps	B1‡	36 268 612	35 852 136	416 476	1.15	-
	P3‡	38 760 700	38 323 480	437 220	1.13	-
	P6‡	37 409 158	36 943 374	465 784	1.25	-
	C1‡	32 313 602	31 904 562	409 040	1.27	-
Total:		1 650 121 990	1 625 140 860	24 981 130	1.52	

* Skin biopsy taken from the diseased forearms ("P") and unaffected back ("B") of patients.
† Skin biopsy taken from the forearms ("C") and breasts ("R") of controls.
‡ Sample resequenced due to low number of reads.

when dealing with paired-end data, where complementing reads (forward and reverse) are in separate FASTQ files. Some downstream tools use the positional relationship of the reads for analysis; so the order and length of the sequences in both FASTQ files must be maintained during pre-processing steps (Bolger *et al.*, 2014).

The results shown in Appendix B and Table 2.2 show that the pre-processing steps for the initial data used in this study was successful. Only a small fraction of the data (1.25%) was removed in the data pre-processing step; whilst the remaining reads had no technical sequences, a high per base sequence quality above 30 and a per sequence GC that is concordant across the samples. However, even though the data QC and pre-processing steps were a success, the unevenly-matched patient and control data (in terms of age, sex and site of tissue collection - Table 2.1) still posed a challenge in the downstream analysis. These challenges are further discussed in Chapter 3.

Chapter 3

Developing a Pipeline for Gene/Transcript Identification and Quantification

In Chapter 2, the RNA-seq data used in this study was introduced, along with the steps taken to perform QC in order to prepare the data for the different types of bioinformatic analyses performed in this study. This chapter addresses the first objective of this study: ***to develop a novel pipeline for mapping raw mRNA reads to reference genomes and quantifying genes/transcripts using the RNA-seq data.*** The `Nextflow` workflow developed here aims to simplify the process of gene/transcript identification and quantification of RNA-seq data in preparation for differential expression analysis. The workflow is portable, scalable and reproducible, and can be applied to other RNA-seq experiments.

3.1 Introduction

Genome sequencing is still a relatively costly venture, mainly due to the genome size. However, since only a fraction of a genome is transcribed, the set of the transcribed RNA molecules (transcriptome) reflects the current state of the cell (or group of cells) at a given condition and time. The analysis of the transcribed RNA molecules can reveal the aetiology and the underlying pathological mechanism of a disease can contribute to revealing the genes and transcripts that are over or under expressed (Finotello and Di Camillo, 2015). Although a useful approach, it does not always reflect the relative amount of protein produced for each gene in that group of cells and should therefore be interpreted with caution. RNA-seq provides a quick and cost effective way of obtaining large amounts of transcriptome data. Such data allow us to identify transcribed genes, study transcripts and exon structures, discover new disease-associated genes, measure transcript and gene expression, study allelic information and identify splice variants for genes (Grabherr *et al.*, 2011; Haas *et al.*, 2013; Li *et al.*, 2014; Trapnell *et al.*, 2013, 2012).

A typical RNA-seq analysis involves three major steps: (1) identifying a set of genes and/or transcripts from the hundreds of millions of short ($\sim$ 36-125 bases) RNA-seq reads produced from the sequencing experiment; (2) quantifying the abundance of the genes/-transcript through mapping to a reference genome/transcriptome; and (3) performing differential expression analysis (Conesa *et al.*, 2016). The first two steps of RNA-seq data analysis, which are essentially steps for preparing data for differential gene expression, are discussed in Sections 3.1.1 (gene/transcript identification) and 3.1.2 (gene/transcript

quantification).

3.1.1 Gene/Transcript identification

After obtaining the RNA-seq data from sequencing and performing QC to remove poor quality reads and contaminants, the next step in RNA-seq data analysis is to identify genes/transcripts from the raw reads. Identification of genes/transcripts is a process that typically involves mapping reads to either genome or transcriptome reference sequences. Both methods have their own advantages and disadvantages, and the choice of mapping to either one of them depends on the biological question being answered (Conesa *et al.*, 2016). For example, mapping to a reference genome allows for the identification of novel genes/transcripts and requires a gapped (splice aware) aligner. In contrast, if no novel gene/transcripts discovery is needed for the analysis, an ungapped aligner can be used to map the reads to a reference transcriptome (Dobin *et al.*, 2013). However, for some experiments, the reference genome or transcriptome may not be present. In such cases, a reference transcriptome would have to be constructed from the experimental RNA-seq data through assembly, then the constructed reference transcriptome can be used to map the reads and quantify expression (Conesa *et al.*, 2016; Grabherr *et al.*, 2011; Haas *et al.*, 2013). The concepts behind these three methods, along with the tools available for each method, are discussed in the following sections.

3.1.1.1 Mapping to reference genome

When performing RNA-seq analysis for an organism for which a complete and annotated reference genome is available, the RNA-seq reads can be mapped to the reference genome to identify genes/transcripts that are present in the data. However, for RNA-seq, read mapping is challenging given that the RNA-seq reads are sequenced from mature mRNA, which has been processed through RNA splicing to remove introns (Dobin *et al.*, 2013). This means that in order to properly align reads to the reference genome (which has intronic regions for genes), the reads must be placed correctly across introns and the exon-intron boundaries (splice junctions) accurately identified (Conesa *et al.*, 2016; Dobin *et al.*, 2013; Engström *et al.*, 2013). The presence of multiple copies of the same or related genomic sequences further complicates the mapping process for RNA-seq as reads cannot be uniquely mapped to a single locus. The most commonly used splice aware aligner, `STAR` (Dobin *et al.*, 2013), implements a sophisticated algorithm that overcomes these alignment issues in RNA-seq.

The `STAR` (Spliced Transcripts Alignment to a Reference) alignment program was designed for overcome the problem of RNA-seq aligners that were designed as extensions of DNA aligners, which either aligned reads to a database of splice junctions or aligned split-read portions contiguously to a reference genome (Dobin *et al.*, 2013). `STAR` aligns non-contiguous reads directly to a reference genome. Its algorithm consists of two major steps: (1) seeding step and (2) clustering/stitching/scoring step. In the seeding step, `STAR`

implements a sequential search for a maximal mappable prefix (MMP). In this method, the algorithm finds the longest sub-string of the read matching one or more sub-strings on the reference genome, starting from the first base of the read. If the read contains a splice junction, it cannot map contiguously to the reference genome, and thus the read will be assigned to a donor splice cite (Dobin *et al.*, 2013).

An MMP search is then repeated (sequentially) for the unmapped portion of the read and then assigned to an acceptor splice site. This sequential MMP search, implemented through uncompressed suffix arrays (SAs), makes the `STAR` algorithm extremely fast. The splice junctions are identified in a single alignment pass without any prior knowledge of the splice site. In addition to identifying splice junctions, `STAR` is also able to identify multiple mismatches and indels, thus allowing for alignments to the reference genome to contain mismatches (Dobin *et al.*, 2013).

In the second step of the `STAR` algorithm, the seeds that were aligned to the reference genome in the first step are clustered together by proximity to a set of selected "anchor" seeds in order to build the alignments of an entire read sequence. Anchor seeds are those seeds that do not align to multiple genomic loci on the reference genome (Dobin *et al.*, 2013). All the seeds that are in proximity to the anchor seed, and are within a user-defined genomic window, are stitched together using a frugal dynamic programming algorithm that allows for any number of mismatches, but only one indel (insertion or deletion gap). The stitching of the seeds uses a local alignment scoring scheme which allows for ranking and quantitative assessment of the alignments. The local alignment scoring scheme is user-defined and assesses for matches, mismatches, indels (gaps) and splice junction gaps (Dobin *et al.*, 2013). The stitched seeds with the highest scores are selected as the best alignment for the read.

3.1.1.2 Mapping to reference transcriptome

When it comes to mapping reads to a reference transcriptome, the issue of multi-mapped reads is further aggravated as reads that would map uniquely to genomic regions, due to the presence of intronic regions, map equally well to the different gene isoforms in the reference transcriptome that share exons (Conesa *et al.*, 2016). Recently, Bray *et al.* (2016) developed `kallisto`, a "pseudoalignment" method for aligning reads to a reference transcriptome. This pseudoalignment methods focus on trying to identify transcripts from the reference transcriptome from which the reads could have originated. This method does not use a direct alignment method to identify where exactly the reads originated from on the reference, as with the `STAR` aligner described by (Dobin *et al.*, 2013), but rather it uses a hash table of k-mers constructed from reads together with a transcriptome de Bruijn graph (T-DBG) constructed from k-mers present on the transcriptome (Bray *et al.*, 2016).

The basis of the model used by `kallisto` is that when performing the transcript quantification step, accuracy does not depend on where inside the transcript the read may

have originated, but rather which transcript could the read have potentially come from (Conesa *et al.*, 2016). Figure 3.1 shows an overview of how the `kallisto`'s pseudoalignment algorithm works using an example read (black) and three overlapping transcripts (pink, blue and green) shown as exons (Figure 3.1a). First, the transcripts are converted into a T-DBG, with each node/vertex (empty circle) representing a k-mer in the T-DBG and associated with a transcript(s) referred to as a "k-compatibility class" (Figure 3.1b). The hashed k-mers of the read (black circles) are then mapped back to the T-DBG (Figure 3.1c) to identify the k-compatibility classes that the read is associated with. Finally, the algorithm uses implements "skipping" (black dashed lines, Figure 3.1d), whereby the k-mers that share the same k-compatibility class (redundant k-mers) are ignored, and then finally taking the intersection of all k-compatibility classes to determine the transcript(s) the read is associated with (pink and blue in this case). The algorithm implemented by `kallisto` is highly efficient and reduces computational times as compared to other direct RNA-seq aligners. It also performs read quantification simultaneously as it does the pseudoalignment. However, one major limitation of `kallisto` is that it can only be used to map to a reference transcriptome (Conesa *et al.*, 2016).

3.1.1.3 Transcriptome reconstruction

Transcriptome reconstruction is a difficult and computationally intensive task. This is because of three main reasons: (1) some transcripts are only represented by a fraction of the reads, (2) it is difficult to identify mature transcripts since the reads originate from both mature mRNA and incompletely spliced precursor RNA, (3) it is difficult to identify which read was produced by which of the isoforms from a single gene as reads are short

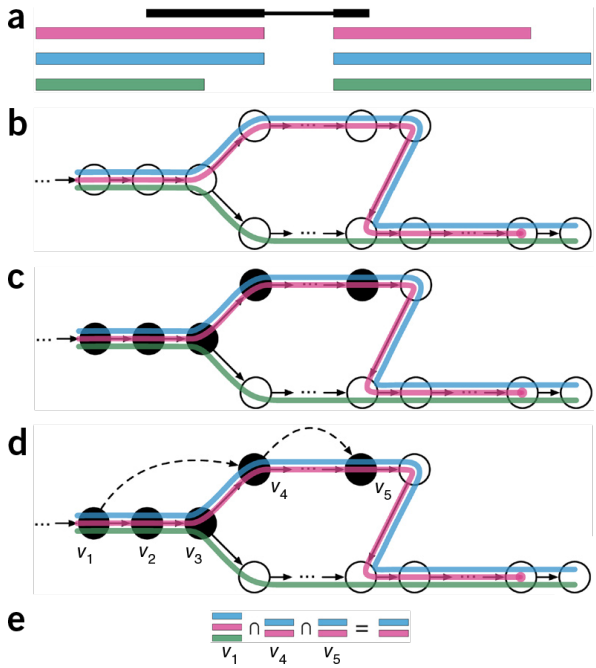


Figure 3.1: An overview of `kallisto`'s "pseudoalignment" algorithm. Read - black line; transcripts - pink, blue and green lines; k-compatibility classes - empty circle; hashed read k-mers - black circles. *Note:* From Bray *et al.* (2016)

(Garber *et al.*, 2011). A number of computational tools have been developed to address these problems. These can be classified as genome-guided and genome-independent (*de novo*) methods (Figure 3.2). Genome-guided approaches rely on the presence of a reference genome. They work by first aligning the reads to the reference genome then assembling the reads that overlap into transcripts (Trapnell *et al.*, 2012).

`TopHat` and `Cufflinks` are the two most popular tools that are used in conjunction with each other for the reconstruction of transcriptomes using a reference genome (Trapnell *et al.*, 2012, 2010). TopHat first uses `Bowtie` to align the reads to the reference genome. `Bowtie` is one of the most efficient alignment programs which stores the reference genome sequence in its FM index data structure and allows rapid searches against it. However, on its own, it cannot align reads that span introns as it does not allow large gaps between the reads and the genome sequence. `TopHat` overcomes this problem by breaking up the reads that Bowtie cannot align into smaller pieces, which can align to the reference genome when processed independently. Cufflinks then assembles the reads that align to the reference genome into individual transcripts.

The *de novo* approaches, however, do not rely on the presence of a reference genome to reconstruct transcriptomes. They use the reads to directly reconstruct transcripts. Trinity is one such tool (Grabherr *et al.*, 2011; Haas *et al.*, 2013). The Trinity assembly pipeline is made up of three sequential modules: Inchworm, Chrysalis and Butterfly. In the first

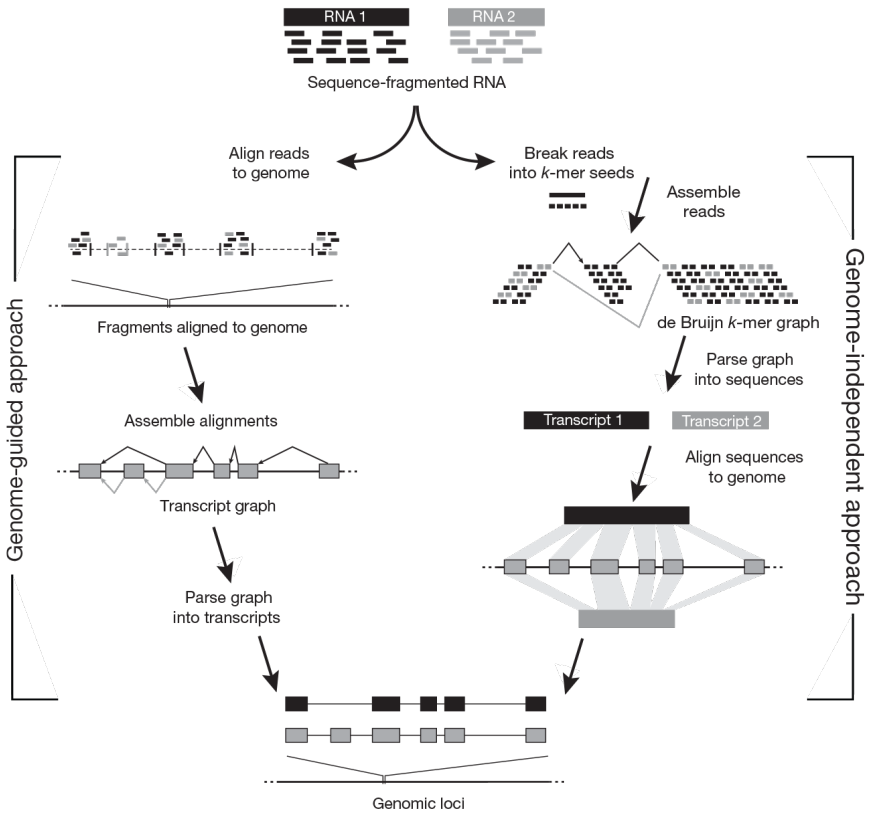


Figure 3.2: The two main approaches to transcriptome reconstruction. *Note:* From Garber *et al.* (2011).

step, Inchworm extracts k-mers from the RNA-seq reads and creates a dictionary. From the dictionary, the most frequent k-mers are selected and used to seed a contig assembly using a greedy (k-1)-mer extension in each direction until the sequence cannot extend any further, then removes the k-mer from the dictionary after use and reports the linear contig. Once all the k-mers have been used up in the dictionary, Chrysalis uses raw reads to cluster related Inchworm contigs into connected components and then constructs a de Bruijin graph for each component. The de Bruijin graphs are subsequently processed by Butterfly in parallel to construct full-length transcripts (Grabherr *et al.*, 2011; Haas *et al.*, 2013). Once the reference transcriptome has been constructed, it can then be used to align the reads in order to identify transcripts.

3.1.2 Gene/Transcript quantification

As mentioned earlier, one of the main goals of transcriptome studies is to identify differentially expressed genes/transcripts in different tissues and under different conditions in order to identify genes that play a major role in a certain phenotype. Unlike microarray studies, the abundance of each gene/transcript unit is measured by the number of RNA-seq reads that map to a transcript (Rapaport *et al.*, 2013). After identifying a set of genes/transcripts present in the read data, through read mapping to either available genome/transcriptome or assembled transcriptome, the abundance of these genes/transcripts have to be quantified in order to perform differential expression analysis (Conesa *et al.*, 2016). The abundance of each gene/transcript is thus greatly affected by sequencing depth and expression levels of other transcripts. It should be noted that when performing read alignments to a genome, the outcome is a set genes to which the reads originated from. In contrast, when performing read alignments to a reference transcriptome, the outcome is a set of transcripts to which the reads belong. Genes and transcripts have been used interchangeably, however, they denote features identified by mapping reads to reference genome and transcriptome, respectively.

A number of tools for quantifying the abundance of genes/transcripts are available, including HTSeq (Anders *et al.*, 2015), featureCounts (Liao *et al.*, 2014) and Cufflinks (Trapnell *et al.*, 2010), and these take as input the aligned reads (output from read mapping to either reference genome or transcriptome) in SAM or BAM file format together with an annotation file in GTF (General Feature Format) format and produce a simple count matrix $\mathbf{N}$ of $n \times m$, where N_{ij} is the number of reads assigned to gene/transcript i in sequencing experiment j (Rapaport *et al.*, 2013). This read count matrix can be used by a number of differential expression detection methods, including `Cuffdiff` (Trapnell *et al.*, 2013), `edgeR` (Robinson *et al.*, 2010), `RSEM` (Li and Dewey, 2011) and `DESeq2` (Anders and Huber, 2010; Love *et al.*, 2014). These differential expression software tools and the different statistical algorithms they employ will be discussed in Chapter 4. The read count matrices produced from read alignments can be broadly classified into two categories, i.e., raw read and normalised read counts.

3.1.2.1 Raw read counts

Generally, counting reads that map to genomic locations (features) is a pretty simple and straight forward method. However, because of intronic regions that exist in the reference genome, decisions need to be made as to how to treat reads falling into these intronic regions (as with read mapping). Also, decisions on how to treat reads falling beyond the annotated features and multiple features have to be made (Anders *et al.*, 2013). When it comes to performing differential gene expression analyses, reads that cannot be uniquely assigned to a single feature due to multi-mapping being part of overlapping features, such reads should be discarded. This is because multi-mapped reads would cause genes to appear differentially expressed if reads mapping to one gene are also counted on other genes due to ambiguous assignment (Anders *et al.*, 2013). `htseq-count`, distributed as a stand-alone Python script in the `HTSeq` package for high-throughput sequencing (HTS) data processing and analysis (Anders *et al.*, 2015), and `featureCounts` (Liao *et al.*, 2014) are the two most popular used programs for producing gene-level raw read counts. Both `htseq-count` and `featureCounts` provide users with options to discard reads that map to multiple features depending on the type of analysis being carried out.

3.1.2.2 Normalised read counts

There are other challenges to gene/transcript abundance quantification using RNA-seq data in addition to multi-mapping reads. These include: (1) during library construction, cDNA fragments are selected based on size causing longer transcripts to produce more reads as compared to shorter transcripts present at the same abundance in a sample, (2) sequencing runs of the same library may produce different volumes of sequencing reads causing variations in the number of fragments mapped across samples (Garber *et al.*, 2011; Trapnell *et al.*, 2012). These two main sources of variability have to be taken into account when using RNA-seq data in order to correctly estimate the expression level of each gene/transcript and to compare the expression level across runs. This is achieved by the fragments per kilobase of transcript per million mapped reads (FPKM, also known as RPKM in single-ended sequencing) metric (Finotello and Di Camillo, 2015; Trapnell *et al.*, 2010).

The FPKM metric normalises the transcripts read count by both their length and total number of mapped reads in a sample, ensuring that the transcript and gene expression levels can be compared across samples. The process of estimating the expression levels of a transcript using RNA-seq is further complicated by isoforms arising from an alternatively spliced gene in a sample (Finotello and Di Camillo, 2015; Trapnell *et al.*, 2013, 2012, 2010). Reads that map to an alternatively spliced gene will also map to the constitutive isoforms and shared exons, thus complicating the process of uniquely assigning reads to each isoform and reducing accuracy when estimating expression levels. Abundance estimation programs, such as `Cufflinks` and `Cuffdiff`, implement more sophisticated algorithms and statistical models that model the sequencing process and determine the

isoform abundance estimates that best explain the observed reads in that experiment with maximum likelihood (Trapnell *et al.*, 2013, 2010). The expression level of a gene with alternatively spliced transcripts will then be the sum of the expression levels of all splice variants for that gene since the FPKM is directly proportional to abundance.

3.1.2.3 Raw versus normalised read counts

Choosing between which read count matrix to use for differential expression analysis depends on the type of analysis being carried out. However, raw read counts have an advantage over normalised read counts, and a number of tools have been developed to work with raw read counts. These tools, such as `DESeq2` (Love *et al.*, 2014), implement normalisation methods within the program and no prior normalisation is needed. The different normalisation methods for raw read counts will be discussed in more detail in Section 4.1.1.1 of Chapter 4.

3.2 Analyses

All analyses were performed through the UNIX CLI on the Wits Computing cluster using the `PBS` scheduler to manage the required resources (amount memory, time and central processing units [CPUs]). Applications used were either already installed on the cluster or installed locally.

3.2.1 Mapping to reference genome

To perform read alignment, the `STAR` alignment program (version 2.5.3a) was used. First, the compressed chromosomes (chromosome 1-22, X, Y and mitochondrial) of the human reference genome (GRCh38 release 94) were downloaded from Ensembl's (https://www.ensembl.org/) file transfer protocol (FTP) site, along with the annotation file in GTF format, then indexes generated using `STAR` and `Bowtie` to prepare the reference genome for alignment using the script below.

```
#!/bin/bash

BASE_FTP=ftp://ftp.ensembl.org/pub/release-94

## Get Chromosome 1-22,X,Y and MT DNA - UNMASKED!
lftp -e "mirror -n -c -r -I Homo_sapiens.GRCh38.dna.chromosome.* \
    --use-pget-n=20 $BASE_FTP/fasta/homo_sapiens/dna/ .; bye"
lftp -e "pget -n20 $BASE_FTP/fasta/homo_sapiens/dna/CHECKSUMS; bye"

## Get Annotations in GFF format!
lftp -e "pget -n20 $BASE_FTP/gtf/homo_sapiens/Homo_sapiens.GRCh38.94.chr.gtf.gz; bye"
lftp -e "pget -n20 $BASE_FTP/gtf/homo_sapiens/CHECKSUMS; bye"

## Extract ALL
gunzip *.gz

## Rename GTF file
mv Homo_sapiens.GRCh38.94.chr.gtf genes.gtf

## Combine chromosome into single file and sort them
cat `ls Homo_sapiens.GRCh38.dna.* | sort -t . -k 5 -V` > genome.fa
sed -i 's/^\(.*\) dna:.*$/\1/' genome.fa

## Generate STAR indexes
```

```
STAR --runThreadN 12 \
     --runMode genomeGenerate --genomeDir . \
     --genomeFastaFiles genome.fa \
     --sjdbGTFfile genes.gtf --sjdbOverhang 99

## Generate Bowtie indexes
bowtie2-build --threads 12 genome.fa genome
```

Once the reference genome was prepared and the indexes built, the alignment of each of the 25 samples (paired-end reads for each sample) was performed using 50GB of memory, 12 CPUs and 24hrs execution time. This was accomplished using the `BASH` script below, submitted to the `PBS` scheduler on the cluster.

```
#!/bin/bash

#PBS -N STAR
#PBS -q WitsLong
#PBS -l nodes=1:ppn=12,walltime=24:00:00,mem=50gb
#PBS -o /spaces/phelelani/ssc_data/results/logs/output.log
#PBS -e /spaces/phelelani/ssc_data/results/logs/error.log

WORK_DIR=/spaces/phelelani/ssc_data/data_trimmed/compressed
OUT_DIR=/spaces/phelelani/ssc_data/results

cd $WORK_DIR

## Loop through the samples
for the_read in $(ls *R1*)
do
    ## Get forward and reverse reads
    read_1=$the_read
    read_2=$(sed -n s@_R1_@_R2_@p <<< "$the_read")
    prefix=${the_read:0:12}

    ## Perform alignment using STAR
    STAR --runMode alignReads \
         --genomeDir $HOME/scleroderma_analysis/indexes \
         --readFilesCommand gunzip -c --readFilesIn $read_1 $read_2 \
         --runThreadN 12 --outSAMtype BAM Unsorted \
         --outFileNamePrefix $OUT_DIR/$prefix
done
```

3.2.2 Read assignment to genomic features

The resulting BAM files from the alignment step were then used to assign reads to genomic features at gene-level in order to quantify the expression of the identified genes. Both `featureCounts` (version 1.6.0) and `htseq-count` (version 0.9.1) were used for this step in order to compare which of the two programs performs better in terms of assigning reads to features as well as the amount of time it takes for each program to execute. Because `htseq-count` is single threaded (can only run on a single CPU), 1 CPU, 8GB of memory and 24hrs of execution time were requested as shown on the `BASH` script below.

```
#!/bin/bash

#PBS -N HTSEQCOUNT
#PBS -q WitsLong
#PBS -l nodes=1:ppn=1,walltime=24:00:00,mem=8gb
#PBS -o /spaces/phelelani/ssc_data/results/htseqcount/logs/output.log
#PBS -e /spaces/phelelani/ssc_data/results/htseqcount/logs/error.log

WORK_DIR=spaces/phelelani/ssc_data/results
```

```
OUT_DIR=/spaces/phelelani/ssc_data/results/htseqcount

cd $WORK_DIR

find . -iname *.bam -exec bash -c \
     'htseq-count -f bam \
            -r name \
            -i gene_id \
            -a 10 \
            -s reverse \
            -m union \
            -t exon \
            $1 $HOME/scleroderma_analysis/indexes/genes.gtf > $OUT_DIR/${1:0:12}.txt'
       - {} \;
```

For `featureCounts`, 12 CPUs, 8GB of memory and 24hrs of execution time were requested since `featureCounts` is multi-threaded. The script below was used to perform gene-level quantification of features identified.

```
#!/bin/bash

#PBS -N FEATURECOUNTS
#PBS -q WitsLong
#PBS -l nodes=1:ppn=12,walltime=24:00:00,mem=8gb
#PBS -o /spaces/phelelani/ssc_data/results/featurecount/logs/output.log
#PBS -e /spaces/phelelani/ssc_data/results/featurecount/logs/error.log

WORK_DIR=/spaces/phelelani/ssc_data/results
OUT_DIR=/spaces/phelelani/ssc_data/results/featurecounts

cd $WORK_DIR

featureCounts -p -B -C -P -J -s 2 \
              -G $HOME/index/genome.fa -J \
              -t exon \
              -d 40 \
              -g gene_id \
              -a $HOME/scleroderma_analysis/indexes/genes.gtf \
              -T 12 \
              -o $OUT_DIR/gene_counts.txt \
              `find . -iname *.bam -printf "%p "`
```

3.2.3 Read-count matrices and QC

The read count matrices produced by `featureCounts` and `htseq-count` are different. `featureCounts` performs gene-level expression for all samples (25 samples in this case) at the same time, and the result is a matrix with rows as genes and columns as sample, and each cell represents the raw count for a gene in a particular sample. `htseq-count`, however, performs the analysis for an individual sample at a time and produces separate matrices for each sample. Tools like `DESeq2` (Love *et al.*, 2014) can handle both types of inputs produced by these tools.

To assess the quality of read mapping and quantification, the `MultiQC` (Ewels *et al.*, 2016) program was executed in the results directory where the analysis took place. `MultiQC` (version v1.5) is able to automatically search for analysis logs and results in a given directory and compiles an HTML report that summarises the different results produced by the different bioinformatic tools. It supports results produced by `STAR`, `featureCounts` and `htseq-count` used in these analyses to perform read mapping and gene-level quan-

tification of features.

3.2.4 Pipeline for gene-level feature counting of RNA-seq data

The main aim of this chapter is to address the first objective of this study (Section 1.10), which is to develop a novel pipeline for mapping reads to a reference genome and quantifying the abundance of the identified genomic features at gene-level. As seen in Sections 3.2.1 and 3.2.2, performing these type of analyses can be pretty daunting and having to keep track of the input, output and intermediate files in between the analysis can be a difficult task. Every script submitted to the cluster for analysis must have inputs and outputs explicitly declared in every script. The purpose of the pipeline developed here was to make the task of producing raw read counts for performing differential expression analysis easier, especially for other researchers with little or no knowledge of bioinformatics, who wish to perform such analysis. The pipeline also needed to be portable and reproducible in order to allow scaling to different computational platforms when large or small datasets are being analysed. `Nextflow` and `Singularity` were used to implement the workflow defined in Sections 3.2.1 and 3.2.2 into a portable and reproducible pipeline as discussed in the sections that follow.

3.2.4.1 Implementing the workflow in `Nextflow`

The advantages of `Nextflow` (discussed in Section 1.1.1) as a workflow management system are that it can handle input and output as channels between processes and reduce the need of having to create intermediate directories to store intermediate results. Variables can also be declared dynamically with no need to explicitly name files, and only the output that is required can be saved to files in each analysis step. The first step in designing the workflow was to create a channel that takes RNA-seq read pairs in FASTQ format, and store them into a channel, so that they can be accessed by applications wrapped as processes on the workflow. Other pipeline inputs, i.e. indexed reference genome and an annotation file, that are required for the analysis are also stored in channels and used by the different processes in the pipeline. The overall workflow of the pipeline, from obtaining the read pairs to producing the raw read counts and QC in `Nextflow` can be summarised as follows:

1. ***Collect read pairs, reference genome and annotation file:*** The pipeline takes as input a path (directory) where the RNA-seq data is, i.e., the pairs of FASTQ file for the samples, and stores them in a channel. It also collects the location of the reference genome (including indexes) and the annotation file.
2. `runSTAR`: Each of the read pairs in the channel created above are accessed by the `runSTAR` process and aligned to the reference genome, producing BAM files that are then stored into channels for processing by `htseq-count` and `featureCounts`. The statistic files are also stored into channels for later processing by `MultiQC`. All other files are discarded.

3. `runHTSeqCount`: The `runHTSeqCount` process accesses the BAM files from the `runSTAR` process and performs quantification of features identified in the BAM file. This is done for each sample as `htseq-count` can only process one BAM file at a time. When this process completes for all the samples, the locations of all the output matrices are stored in a file, which is then used by an intermediate process to combine all the matrices into one.
   (a) `runCleanHTSeqCounts`: The file with the locations for all the matrices is used as input for this process to combine the matrices produced by the `runHTSeqCount` process into a single matrix. All other unnecessary information in the matrix is discarded.
4. `runFeatureCounts`: This process is executed in parallel to `runHTSeqCount` process. The BAM files produced by the `runSTAR` process are accessed by this process and gene feature quantification is performed using `featureCounts`. The resulting read count matrix is used by an intermediate process to remove unnecessary information in the matrix.
   (a) `runCleanFeatureCounts`: All unnecessary information in the matrix produced by `runFeatureCounts` process is removed.
5. `runMultiQC`: Finally, the `MultiQC` program is then used to summarise the results from `runSTAR`, `runHTSeqCount` and `runFeatureCounts` processes. The QC information is stored in an HTML file that can be used to access the quality of mapping and gene quantification steps.

It should be noted that the pipeline runs with no need for user intervention. Users only need to specify the location of the input files required to run the pipeline and the output of where the results should be stored.

3.2.4.2 `Singularity` images for applications

To facilitate reproducibility and portability of the pipeline, `Singularity` container images were created for each of the processes in the pipeline. Each `Singularity` image contains the necessary software required by each process to run. This removes the need to install all the software tools used for these analysis. The `Singularity` containers were created as recipes, which can be used to build the images using `Singularity`'s `build` command. Appendix C lists the `Singularity` recipes used to package softwares needed to run the `runSTAR` (Appendix C.1), `runHTSeqCount` (Appendix C.2), `runFeatureCounts` (Appendix C.3) and `runMultiQC` (Appendix C.4) processes.

3.2.4.3 GitHub repository for the pipeline

A GitHub repository was created to keep track of the changes to the pipeline, as well as to documented and share the pipeline with other researchers interested in using the workflow for their analysis. This workflow is available via https://github.com/phelelani/nf-rnaSeqCount. The `Singularity` recipes are also included in the GitHub repository

of the pipeline, and are linked to SingularityHub (https://www.singularity-hub.org/collections/770) where their images are hosted. Any changes made to the recipes are updated in SingularityHub and the images rebuilt on the server. The images can be downloaded directly using `Singularity`'s `pull` command or via a script packaged with this pipeline.

3.3 Results

The workflow has been successfully implemented in `Nextflow` and `Singularity`, and can be executed on any UNIX-based OS with `Nextflow` and `Singularity` installed. It is available on GitHub (https://github.com/phelelani/nf-rnaSeqCount) and all the `Singularity` images with softwares required for running the pipeline are hosted on SingularityHub (https://www.singularity-hub.org/collections/770). In this section, the details of the workflow, including using the workflow, data preparation and results produced, will be presented.

3.3.1 `rnaSeqCount`: A portable and reproducible `Nextflow` pipeline for gene-level feature counting of RNA-seq data

As mentioned before, the pipeline designed here is available for use on GitHub, referred to as "`rnaSeqCount`", mainly because it is a workflow that uses RNA-seq data to produce raw read "counts" for performing differential expression analysis. Figure 3.3 summarises the overall workflow of the pipeline and how the different processes are "stitched" together to produce a matrix of raw read counts. In the following sections, the usability of the pipeline will be discussed.

3.3.1.1 Obtaining the pipeline

The pipeline depends on `Nextflow` and `Singularity` to run. These two softwares must be installed in order for the pipeline to be executed. The `rnaSeqCount` pipeline can be obtained using either the `Git` or `Nextflow` commands as shown below.

```
## Using git command
git clone https://github.com/phelelani/nf-rnaSeqCount.git

## Using nextflow command
nextflow pull phelelani/nf-rnaSeqCount
nextflow pull https://github.com/phelelani/nf-rnaSeqCount.git
nextflow clone phelelani/nf-rnaSeqCount <target-dir>
```

Once the GitHub repository of the pipeline has been downloaded, the contents of the `rnaSeqCount` folder will consist of files and folders as listed below. Only the two configuration files, `main.config` and `prepare_data.config`, need to be edited in order to run the script (Figure 3.3). The rest of the files and folders do not need to be edited, especially the `nextflow.config` as it contains all of the necessary instructions for performing analysis on different computational platforms.

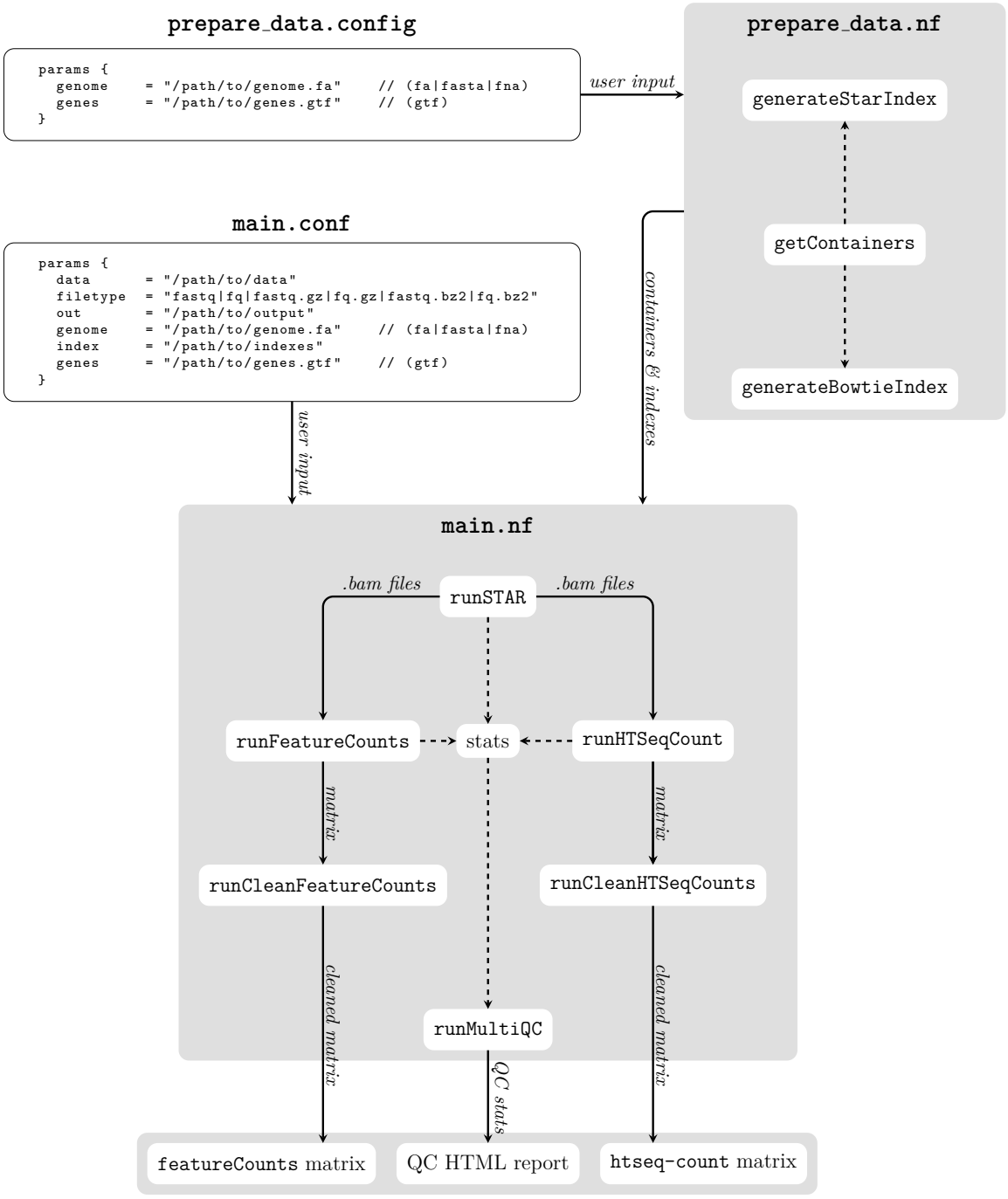


Figure 3.3: Overall summary of the `rnaSeqCount` workflow. The `rnaSeqCount` pipeline works in two stages: (1) **Genome indexing:** The `prepare data.config` need to be provided the location of the reference genome and annotation file. The `prepare data.nf` can then be executed in order to download singularity images and perform genome indexing using `STAR` and `Bowtie`. (2) **Pipeline execution:** Once the reference genome has been indexed, the location of the FASTQ files, output directory, FASTQ file type and location of the reference genome and annotation file can be provided in the `main.config` and the `main.nf` executed to perform the analysis on the RNA-seq data.

```
nf-rnaSeqCount
|-containers            # (folder) Singularity recipes and location for images
|   |--Singularity.featureCounts
|   |--Singularity.htseqCount
|   |--Singularity.multiQC
|   |--Singularity.star
|   |--Singularity.trinity
|--templates            # (folder) Location of extra scripts for performing analysis
```

```
|   |--clean_featureCounts.sh
|   |--clean_htseqCounts.sh
|--README.md            # Pipeline documentation
|--main.config          # Configuration file for the main Nextflow script (user input)
|--main.nf              # Main Nextflow script for running the pipeline
|--nextflow.config      # Pipeline configuration file
|--nf-rnaSeqCount.png   # Summary of the pipeline
|--prepare_data.config # Configuration file for preparing genome indexes (user input)
|--prepare_data.nf      # Main Nextflow script for preparing genome indexes
```

3.3.1.2 Obtaining `Singularity` images and indexing the reference genome

The first step to perform once the pipeline has been cloned from GitHub is to download the `Singularity` images from SingularityHub. This step is crucial as all processes in the pipeline depend on the applications that are packaged in these images. To accomplish this, first the location of the reference genome FASTA file and the annotation GTF file must be provided in the `prepare_data.config` (Figure 3.3). Once that has been done, the `prepare_data.nf` can be executed using one of the three options for `--mode`, i.e., `getContainers`, `generateStarIndex` or `generateBowtieIndex`, depending on whether the `Singularity` images or indexes are already available or not.

In addition to running the pipeline locally, the `rnaSeqCount` pipeline also supports the PBS and SLURM job schedulers on HPCs, and this information can be passed to the `-profile` option of `Nextflow` when executing the workflow. Available options for this step are `slurmPrepare` (for SLURM scheduler) and `pbsPrepare` (for PBS scheduler). Assuming that the images required and genome indexes are not available, data preparation can proceed in the following manner using `Nextflow` on a HPC with PBS as a scheduler.

```
## Download Singularity images
nextflow run prepareData.nf --mode getContainers -profile pbsPrepare

## Generate STAR indexes
nextflow run prepareData.nf --mode generateStarIndex -profile pbsPrepare

## Generate Bowtie2 indexes
nextflow run prepareData.nf --mode generateBowtieIndex -profile pbsPrepare
```

3.3.1.3 Executing the `rnaSeqCount` workflow

Once the `Singularity` images and genome indexes are available, the main `rnaSeqCount` workflow can be executed. First, the FASTQ files location, FASTQ file type (one of `fastq`, `fq`, `fastq.gz`, `fq.gz`, `fastq.bz2` or `fq.bz2`), output folder, reference genome (`.fa`, `.fasta` or `.fna` file extensions), reference genome indexes and annotation file (`gtf` file extension) must be provided in the `main.config` file. As with running the `prepare_data.nf`, executing `main.nf` can also be used with either SLURM (`slurm`) or PBS (`pbs`) `-profile` option. An example of running the `rnaSeqCount`'s `main.nf` script using a SLURM scheduler is as follows.

```
## Execute rnaSeqCount workflow using a SLURM scheduler
nextflow run main.nf -profile slurm
```

3.3.1.4 Results produced by the rnaSeqCount pipeline

In the output directory that is specified in the `main.config` file for the `rnaSeqCount` workflow, a number of folders can be found:

- n number of folders corresponding to each of the samples that were processed by the pipeline. These folders contain general statistics on mapping using `STAR`.
- ***featureCounts*** folder containing read counts matrix (`gene_counts_final.txt`) for `htseq-count`. This file can be used for differential expression analysis.
- ***htseqCounts*** folder containing read counts matrix (`gene_counts_final.txt`) for `featureCounts`. This file can be used for differential expression analysis.
- ***report_QC*** folder containing `MultiQC` QC reports in HTML format. This file can be used to assess the quality of read mapping and gene quantification.
- ***report_workflow*** folder containing pipeline execution reports. These files can be used to trace the execution of the pipeline and check other metadata in order to assign resources correctly to the processes.

Documentation on how to use the `rnaSeqCount`workflow is also available via the GitHub repository (https://github.com/phelelani/nf-rnaSeqCount).

3.3.2 Using rnaSeqCount to generate raw read counts for the SSc data

The RNA-seq data for patients with SSc introduced in Chapter 2 were used to produce read count matrices for differential expression (discussed in Chapter 3) using the `rnaSeqCount` pipeline. Initial testing of the `rnaSeqCount`workflow was performed on the Wits Computing cluster using the PBS scheduler. However, in the middle of 2018, the scheduler was changed to SLURM. It is worth mentioning that even with the changes to the cluster, these changes did not have much effect on the pipeline as only a few lines were added to the `nextflow.config` file to accommodate these changes. This also proves the portability and scalability of the pipeline to run on different computing environments. In total, 25 samples were processed by the `rnaSeqCount` pipeline. Figure 3.4 shows the output summary that is generated by `Nextflow` in HTML format when executing the workflow.

In addition to the summary and metadata produced, `Nextflow` also includes interactive charts produced by HighCharts (https://www.highcharts.com/) to summarise resources (CPU, memory and execution times) used to execute the workflow. These can be useful for monitoring and requesting resources efficiently when performing these analyses using the `rnaSeqCount` pipeline.

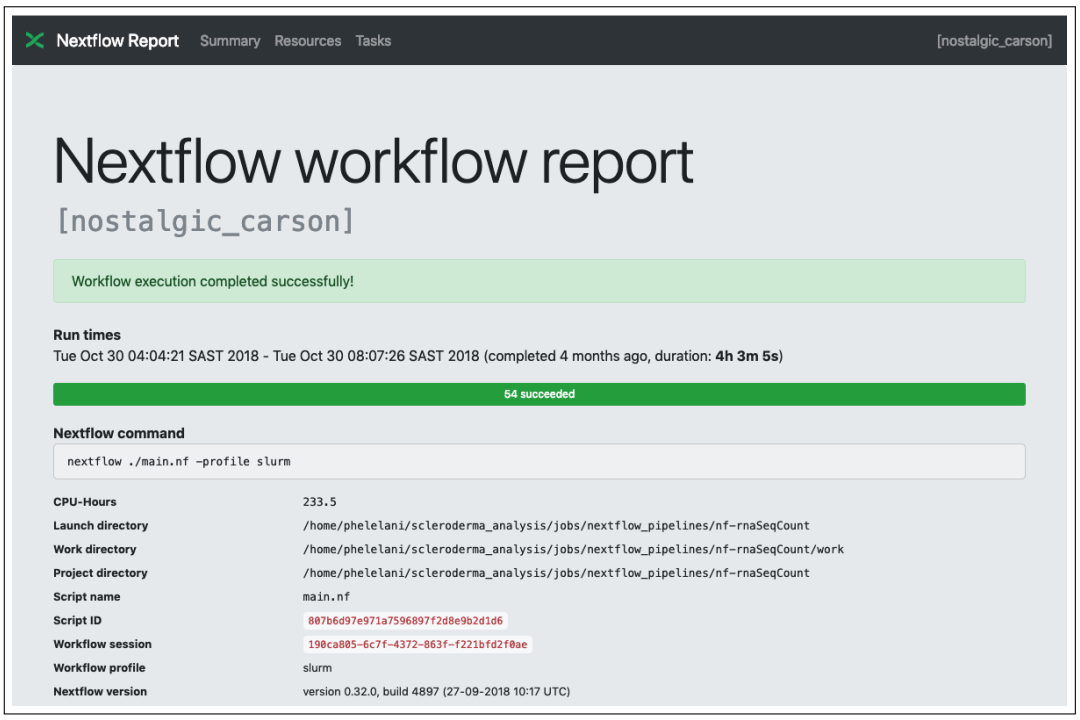


Figure 3.4: Summary report and metadata for `rnaSeqCount` pipeline execution. Overall execution time and metadata for performing analysis on all 25 RNA-seq samples from the SSc patients using `rnaSeqCount` pipeline

3.3.2.1 CPU usage

Figure 3.5 summarises the CPU usage by each process in the workflow in terms of the amount of CPU requested. It can be seen that all processes utilised the requested CPU efficiently, except for the `runHTSeqCount` process and `runStar` processes, which used more than 100% of the 1 and 13 CPUs requested, respectively. This could cause potential issues when these processes are executed when the cluster is under a lot of pressure, as it could cause the jobs to be terminated. The amount of CPU requested by the two processes could be increased to prevent premature termination.

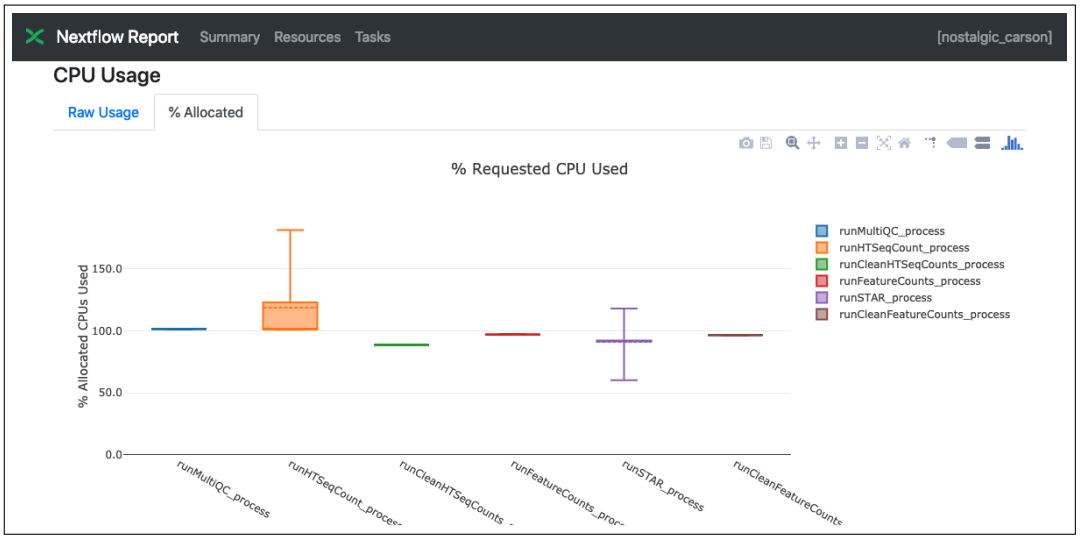


Figure 3.5: % CPU usage by each process of the `rnaSeqCount` pipeline.

3.3.2.2 Memory usage

The % memory usage by the processes in the `rnaSeqCount` pipeline is summarised in Figure 3.6. All processes in the workflow utilised the requested memory efficiently and were well within their limits (all processes used less than 70% of allocated memory). The memory requested for these processes could be reduced by at most 30% for each process to allow fair usage of memory resources on the cluster.

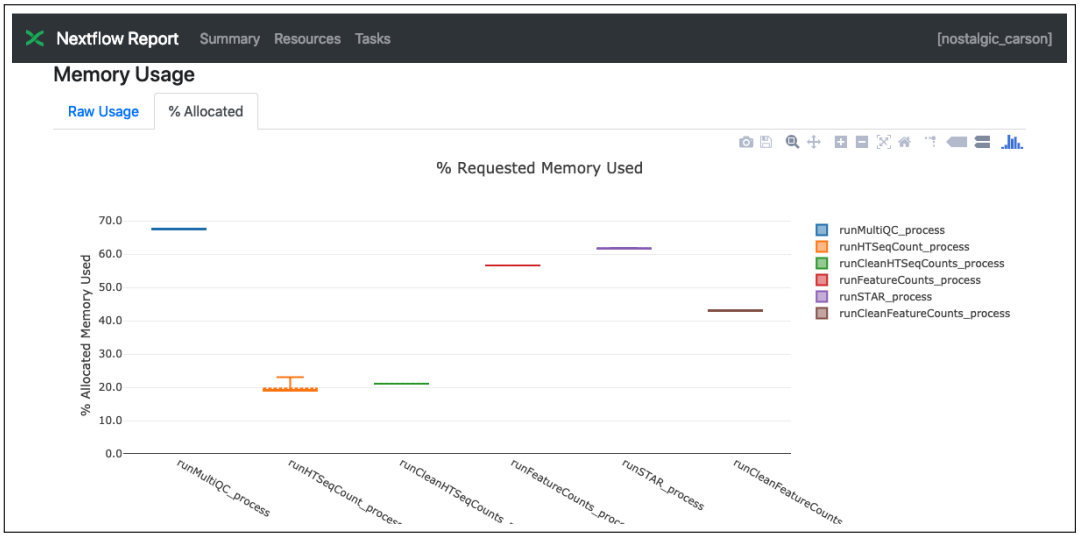


Figure 3.6: % Memory usage by each process of the `rnaSeqCount` pipeline.

3.3.2.3 Execution time

The usage of the requested time to run each process on the workflow was less than 10% (Figure 3.7). This also shows that all processes were within their limits in terms of time requested to process each sample. These times can also be reduced by at most 90% for fair usage on the cluster, and also to allows less queue times when the jobs are submitted to the job scheduler.

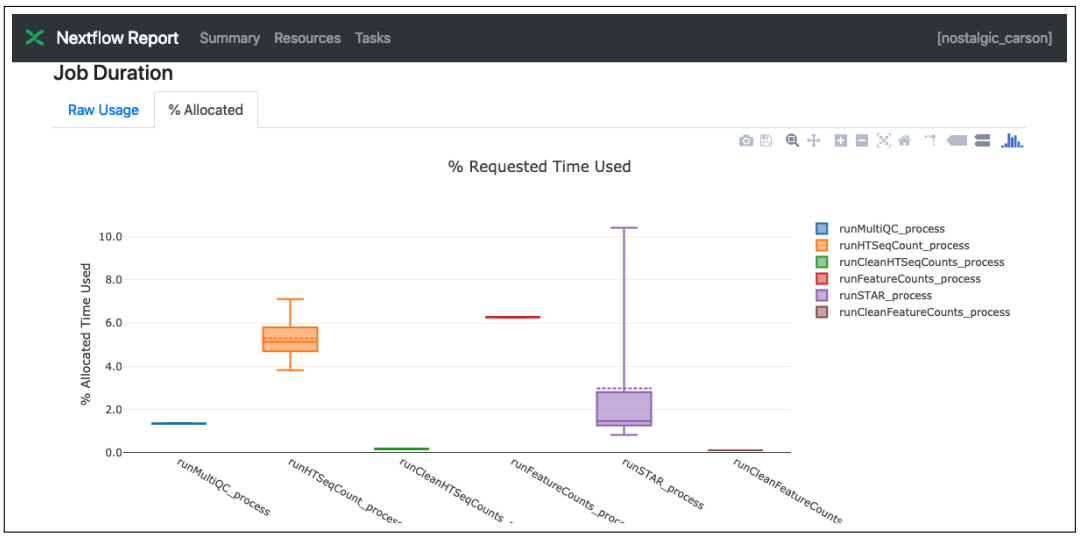
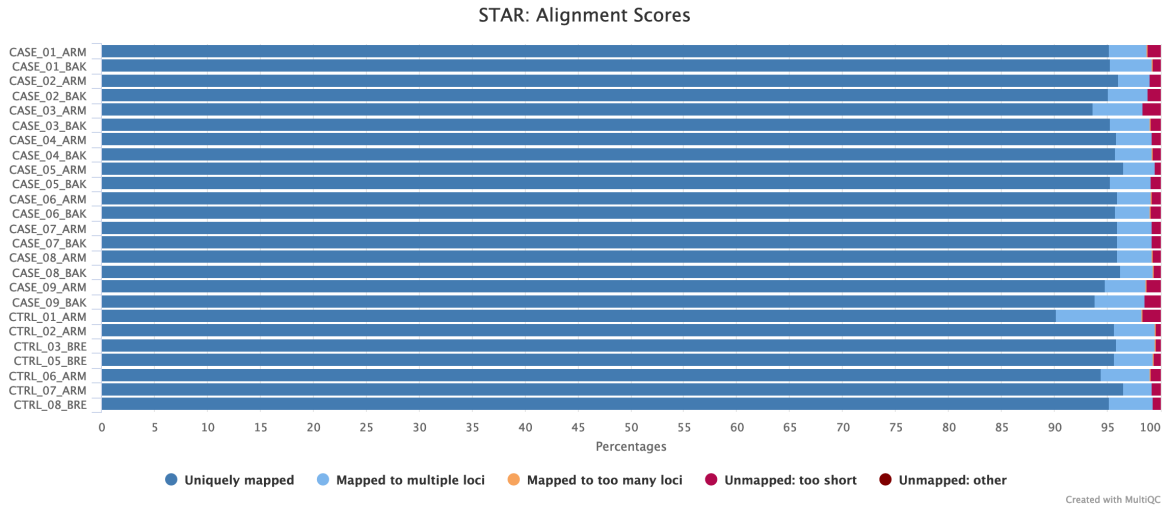


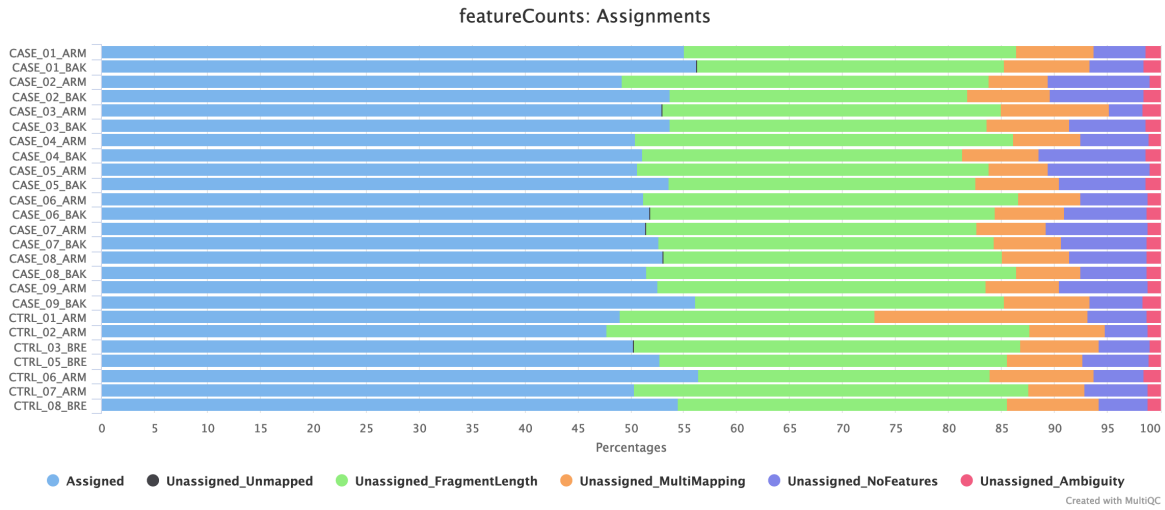
Figure 3.7: % Requested time usage by each process of the `rnaSeqCount`pipeline.

3.3.2.4 QC plots produced by `MultiQC`

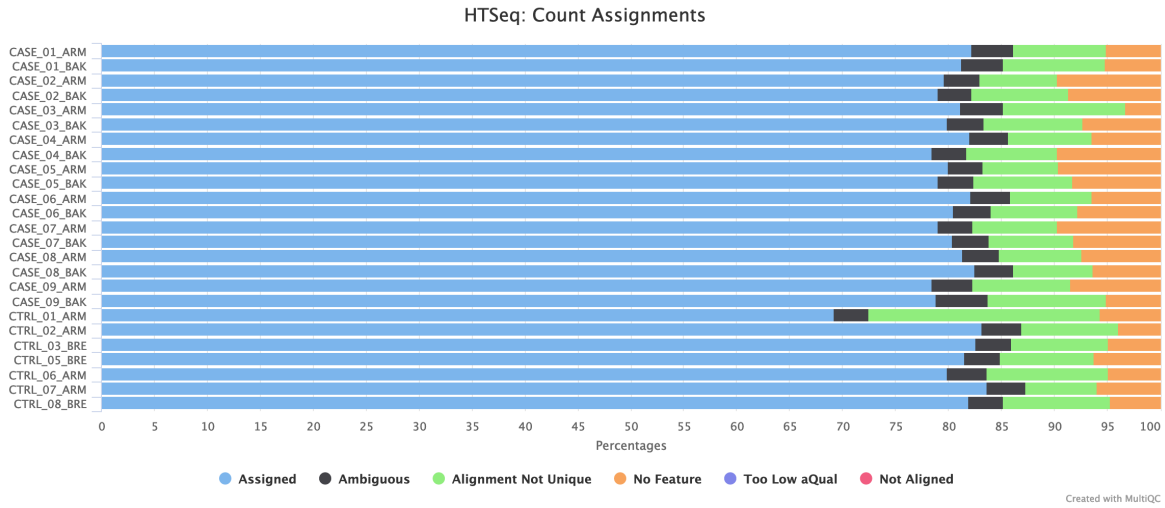
Figure 3.8 summarises the QC produced by `MultiQC` in the workflow for mapping of RNA-seq reads of the 25 SSc samples to the reference genome using `STAR` (Figure 3.8a)



(a) Alignment of reads to the reference genome by `STAR`.



(b) Assignment of reads to features by `featureCounts`.



(c) Assignment of reads to features by `htseq-count`.

Figure 3.8: QC summary of the mapping and quantification of the `rnaSeqCount` pipeline produced by `MultiQC`.

and assignment of reads to the identified features using `featureCounts` (Figure 3.8b) and `htseq-count` (Figure 3.8c). There were more than ∼90% reads uniquely aligned by `STAR` to the reference genome, showing that the alignment step was successful at assigning reads uniquely to genes on the reference genome sequence. In comparison, `htseq-count` produced better results at identifying features than `featureCounts`; approximately 50% and 80% of unique reads across all samples were assigned to features by `featureCounts` and `htseq-count`, respectively.

3.4 Discussion

This chapter was aimed at addressing the first objective of this study, which was to design a workflow/pipeline for mapping raw RNA-seq reads to the reference genome and quantifying the abundance of the identified features (genes and/or transcripts) to produce a matrix $\mathbf{N}$ of $n \times m$, where N_{ij} is the number of reads assigned to gene/transcript i in sequencing experiment j. This matrix is the input for a majority of differential expression tools that take raw read counts (before normalisation) and use different statistical algorithms to calculate differences in the levels of expression of genes/transcripts. In Section 1.1 of Chapter 1, the main requirement for a highly efficient workflow were discussed, and these include reproducibility (capability of the workflow to reproduce the results under different computational resources), portability (capability of using the pipeline on different computational platforms) and scalability (being able to execute the pipeline on desktop machines, cloud or HPC environments).

The pipeline presented in this chapter, `rnaSeqCount` (https://github.com/phelelani/nf-rnaSeqCount), meets these requirements for an efficient workflow. `rnaSeqCount` is designed on `Nextflow` and all its application dependencies are packaged in `Singularity` images. This means that the pipeline can be run on any machine, from desktop to HPC, with both `Nextflow` and `Singularity` installed. `Nextflow` supports a wide variety of job schedulers, and the `rnaSeqCount` pipeline comes packaged with support for PBS and SLURM scheduler support. Advanced users can add their own scheduler support using the `nextflow.config` file. The workflow also comes with detailed documentation on GitHub, where users interested in using this workflow for their analysis can obtain it. The `Singularity` images hosted on SingularityHub (https://www.singularity-hub.org/collections/770) ensure that users who cannot build the images from the `Singularity` recipes that are packaged with the workflow can download the required images from the server.

Execution of the `rnaSeqCount` pipeline is pretty straight forward, and required users to only edit the `prepare_data.config` (for downloading the required `Singularity` images and indexing the reference genome) and `main.config` (for execution of the main `rnaSeqCount`'s workflow) files with the necessary input files/directories, i.e., reference genome FASTA file and its indexes, annotation GTF file,input FASTQ files location, in-

put FASTQ file type and the output directory. A help menu can be accessed by executing the `main.nf` file with the `--help` option as follows:

```
nextflow run main.nf --help
```

In addition to all the analysis and testing performed on the Wits Computing cluster using SLURM and PBS, the workflow also has been successfully tested on the University of Cape Town (UCT) eResearch HPC (http://hpc.uct.ac.za/) and as the Amazons AWS (https://aws.amazon.com/) using the SSc data for this study. On the UCTs eResearch HPC, which has SLURM as the job scheduler, the same computing requirements as with the Wits Computing cluster were used. However, there was some incompatibility with one of the `Singularity` images (`MultiQC` image) as all the images in this study were built on `Singularity` version 2.6, whilst the UCT eResearch HPC had `Singularity` version 3 installed. For now, the workflow images only support `Singularity` version 2.6.

For the AWS execution of the workflow, the `Nextflow` supported Amazon Machine Image (AMI), `ami-4b7daa32`, was used to deploy an Amazon Elastic Block Store (EBS) of 1000GB using the Elastic Compute Cloud (EC2), `m4.10xlarge`, with 40 virtual CPUs and 160GB of memory. All AWS analysis were performed on the European (Ireland) region since the `Nextflow` AMI was only available for this region. The workflow was executed using the standard computing environment (no job scheduler) on the EC2. The summary report for executing the `rnaSeqCount` workflow on the AWS is shown in Figure 3.9. Even though execution of the workflow on the Wits Computing cluster was parallelised, as com-

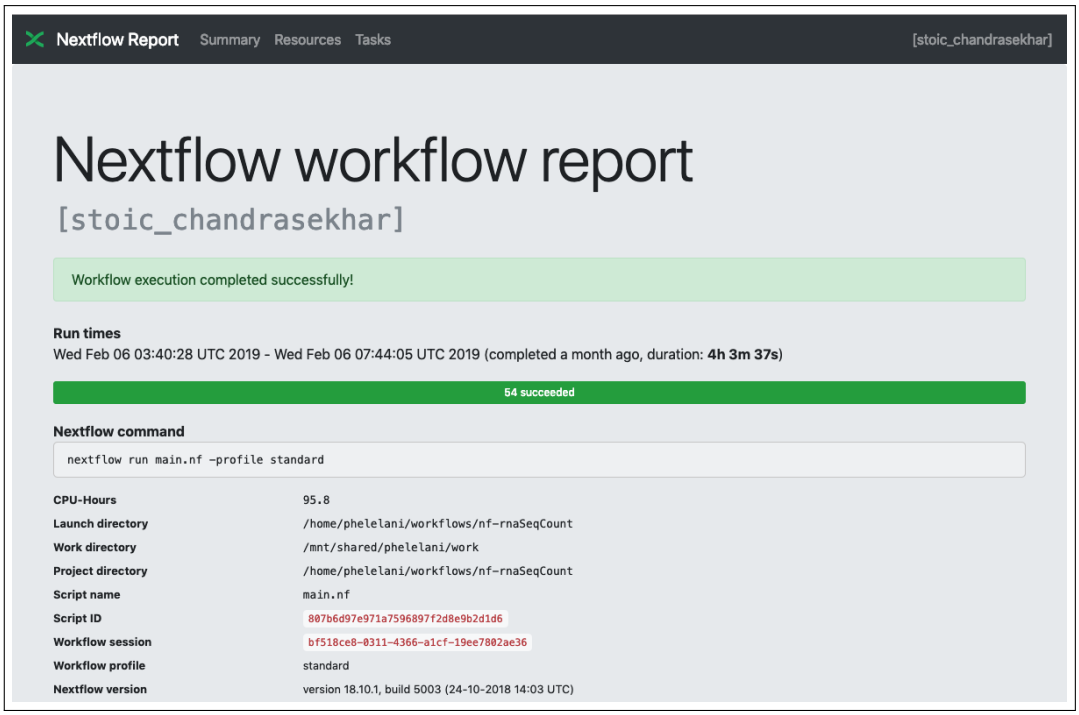


Figure 3.9: Summary report and metadata for `rnaSeqCount` pipeline execution on AWS. Overall execution time and metadata for performing analysis on all 25 RNA-seq samples from the SSc patients using `rnaSeqCount` pipeline on the Amazon AWS

pared to the native execution of the pipeline on the Amazon AWS, the execution times are almost the similar (Figure 3.4 and 3.9), with less CPU hours used on the Amazon AWS. Estimating the cost of running the analysis on the AWS, the `m4.10xlarge` cost $2.22 per hour (https://aws.amazon.com/ec2/pricing/on-demand/), the standard general purpose solid-state drive (SSD) costs $0.11 per GB-month (https://aws.amazon.com/ebs/pricing/). Given that the analysis took approximately 4 hours, the total approximated cost for running the `rnaSeqCount` pipeline on the SSc data was:

$$\left(\frac{\$2.2}{hour} \times 4\ hours\right) + \left(\frac{\$0.11/GB}{month} \times 1000GB \times \frac{4\ hours}{740\ hours}\right) = \$9.39$$

The `rnaSeqCount` pipeline was also tested on RNA-seq data for albinism (data consisted of 24 samples), which was kindly provided by Dr. Thandiswa Ngcungcu for the purpose of testing the pipeline. The pipeline was able to successfully run on this data and produce a matrix for differential expression analysis. In conclusion, this chapter has presented a reproducible, portable and scalable pipeline, `rnaSeqMetagen`, that is able to identify and quantify genomic features given a reference genome, gene annotation file and RNA-seq raw data. The `STAR` alignment program used for the alignment of reads to the reference genome proved to be an accurate tool for this purpose since it was able to uniquely map more than 95% of the reads in each sample to the reference genome. The inclusion of both `htseq-count` and `featureCounts` presented a way to compare the two tools at quantifying genomic features, with `htseq-count` proving to be superior than `featureCounts`. This workflow is freely available to other researchers wishing to perform similar analyses.

Chapter 4

Differential Expression Analysis

In Chapter 3, the initial steps for preparing RNA-seq data for performing differential expression, i.e., producing read counts from raw reads, were introduced, along with the `rnaSeqCount` pipeline that takes RNA-seq reads, map them to the reference genome and quantify the expression of the identified features. This chapter addresses the second and third objectives of this study: ***to identify significantly differentially expressed transcripts/genes and non-coding RNAs and examine them for novel disease association*** and ***to identify pathways that influence the onset and severity of SSc and to identify potential biomarkers***, respectively. The differential expression and pathway analyses performed in this chapter take, as input, the raw read counts that are produced by the `rnaSeqCount` pipeline. A workflow in `R` was developed for these differential expression and pathway analyses. However, this workflow is specific to the RNA-seq data and samples used, and cannot be applied directly to other studies as there are a number of analyses and parameters that are data-dependent. The workflow, however, can be modified to suit other differential expression and pathway analyses.

4.1 Introduction

As discussed in Section 1.7.4, a transcriptome is a collection of all the RNA molecules transcribed from genes in a particular cell (or group of cells). Their relative abundance in a cell represents the level of expression of the genes from which they are transcribed from for a particular tissue, developmental stage or condition of an organism. Even though the RNA molecules are not the final product of the process of gene expression (transcription and translation), studying these molecules and their relative abundance can reveal important aspects about the state of the cell, tissues and conditions being investigated (Costa-Silva *et al.*, 2017; Finotello and Di Camillo, 2015).

Earlier methods for performing gene expression profiling and differential gene expression relied on hybridisation-based approaches, such as microarrays. These methods used array probes, which consisted of known sequence(s) being investigated from a particular region of a genome. By washing samples being investigated over the array probes, the fluorescent labelled cDNA molecules in the sample bind to the probes containing a complementary sequence, producing an image that enables quantification of the specific cDNA molecules corresponding to a gene (Finotello and Di Camillo, 2015). However, even though these methods were widely used in transcriptomics and are of relatively low cost, they are limited by their reliance on prior knowledge of the genome and genes being studied in

order to design the probes. Obtaining and studying transcripts on a larger scale is not possible using such methods (Costa-Silva *et al.*, 2017; Finotello and Di Camillo, 2015).

The advent of RNA-seq on NGS platforms has revolutionised transcriptomics as a method for studying gene expression and is now being implemented widely in place of microarrays. The ability to identify known and novel transcripts, sequencing and quantification of all transcripts in a sample are the main advantages of RNA-seq over microarrays (Costa-Silva *et al.*, 2017; Finotello and Di Camillo, 2015; Grabherr *et al.*, 2011; Haas *et al.*, 2013; Trapnell *et al.*, 2012)

4.1.1 Differential expression using RNA-seq data

After obtaining the read counts from the RNA-seq data through mapping (Chapter 3), these can be used to evaluate how the expression levels vary across different conditions. The read counts represent the expression levels of genes/transcripts across the samples or conditions being investigated. However, as discussed in Section 3.1.2, the expression levels of each gene/transcript in the samples/conditions is limited by the sequencing depth and dependent on the expression of other genes/transcripts. Library size selection and sequencing of the same sample on multiple lanes also bring technical variability to the levels of gene expression quantified (Rapaport *et al.*, 2013). A number of statistical algorithms have been developed for performing differential expression, whilst taking the challenges of raw read counts into account, either before or during differential gene expression analysis.

In the matrix produced for read counts (matrix $\mathbf{N}$ of $n \times m$, where N_{ij} is the number of reads assigned to gene/transcript i in sequencing experiment j as discussed in Section 3.1.2), the reads that are counted for gene i are not a direct measure of the gene's expression. They are an estimated measure of the expression which can be represented by $N_{ij} \propto l_{ij}\mu_{ij}$, where l_{ij} is the expected length of the gene and μ_{ij} is the expected expression of the gene (Rapaport *et al.*, 2013). Due to this length bias, raw read counts alone are not sufficient nor accurate enough to perform differential expression of genes/-transcripts between conditions. This is because the detection of differential expression amongst shorter genes is also reduced by the lack of coverage. The lower the number of read counts for a gene, the lower the power of the statistical test, therefore low abundance transcripts are more difficult to assess in determining the significance of the differential expression. In this section, the three components for analysing differential expression, i.e., read count normalisation, estimation of parameters for the statistical model and testing for differential expression, will be discussed.

4.1.1.1 Normalisation methods: RPKM, FPKM and TMM

In Section 3.1.2, the FPKM metric implemented in `Cufflinks` and `Cuffdiff` for normalising read counts was introduced (Trapnell *et al.*, 2010). The FPKM metric was introduced to accommodate paired-end reads, and is an extension of the RPKM (reads per kilobase

per million mapped reads) method for quantifying transcript levels. The RPKM measure of read abundance promotes comparison of transcript levels both within and between samples by normalising for transcript length and total read count. This is a reflection of the total molar concentration of a transcript in a sample (Mortazavi *et al.*, 2008). The RPKM method of normalisation of transcripts is implemented in the `ENRAGE` (Enhanced Read Analysis of Gene Expression) software (`http://woldlab.caltech.edu/rnaseq`).

Another read count normalisation method implemented in the differential expression software, `edgeR`, is the trimmed means of M values (TMM) (Robinson *et al.*, 2010; Robinson and Oshlack, 2010). This method was aimed at normalising read counts by taking into account the different library compositions between samples. The TMM normalisation method first excludes the genes that have both high average read counts and large differences in expression levels. These are the 30% of genes between the samples being compared which are characterised by the most highest log fold changes (LFC), i.e., "M-values", based on the fact that LFC from genes with high read counts exhibit low variance on the logarithm scale. A scaling/normalisation factor is then computed from the remaining subset of genes and used to correct for differences in library sizes between the samples being compared (Finotello and Di Camillo, 2015; Rapaport *et al.*, 2013; Robinson and Oshlack, 2010).

4.1.1.2 Statistical modelling of gene expression

The RNA-seq tools for identifying differential expression can be grouped into two categories, i.e., parametric and non-parametric. The parametric methods for detecting differential expression first capture information about the count data (usually after normalisation), then assume a specific model to describe the underlying distribution of the data. A statistical test is then performed to identify genes whose expression between the conditions being compared is greater than the variance predicted by the model. The majority of the parametric methods implement models based on the Poisson and Negative Binomial (NB) distributions (Costa-Silva *et al.*, 2017; Finotello and Di Camillo, 2015).

In contrast, non-parametric methods do not force a specific model to fit the data, but rather capture all information on the distribution of the data by taking into account that the data cannot be defined using a specific set of parameters. Thus, with this method, the amount of information on the distribution of the data is directly proportional to the size of the data. The non-parametric methods therefore do not perform well for RNA-seq data due the small number of replicates that are usually available for RNA-seq studies; parametric methods are preferred for RNA-seq analyses (Finotello and Di Camillo, 2015). A number of well established tools for differential expression detection implement parametric methods, including `DESeq2` and `Cuffdiff`, and these tools first estimate model parameters from the RNA-seq data, then use a statistical test to detect differentially expressed genes (Love *et al.*, 2014; Trapnell *et al.*, 2013, 2012).

4.1.1.3 Testing for differential gene expression

Following parameter estimation for a particular statistical model used by the different tools, a statistical test for significance in differential expression of a gene between two conditions is performed. Amongst the most used, and well established methods of differential expression tools used, `DESeq2` (Love *et al.*, 2014) and `Cuffdiff` (Trapnell *et al.*, 2013, 2012) are the most commonly used due to their precision, accuracy and sensitivity (Costa-Silva *et al.*, 2017). `Cuffdiff`, which is included in `Cufflinks` as a separate program, calculates the gene/transcript expression in two or more samples (different conditions) and tests the statistical significance observed between them (Trapnell *et al.*, 2013, 2012). `Cufflinks` allows the user to provide multiple biological replicates in each condition. The replicates are essential in identifying how read counts differ for each gene/transcript across replicates and uses the variance estimates to calculate the significance of the observed changes in expression. Replicates are also useful in controlling for batch effects such as variance in sample preparation. It then reports the change in expression for each gene and transcript together with the statistical significance scores (fold changes in $\log_2$ scale and P-values) of the changes. `Cuffdiff` can also identify alternatively spliced genes, genes that are differentially regulated via promoter switching, group genes with the same transcription start site (TSS) and calculate the expression levels of a TSS group. However, `Cuffdiff` does not use raw read counts as input, but rather takes input produced by `TopHat` and `Cufflinks` for performing differential expression analysis.

`DESeq2` (Love *et al.*, 2014) is perhaps one of the most well-established and comprehensive tools for performing differential expression on RNA-seq data. `DESeq2`, extension of its predecessor `DESeq` (Anders and Huber, 2010), uses shrinkage to calculate dispersion and fold changes for a more quantitative approach to the analysis of differential expression of RNA-seq data. Its main features include visualisation, gene ranking, hypothesis testing below and above a specified absolute $\log_2$FC threshold, regularised log transformations for quality assessment and clustering of overdispersed count data. When performing differential expression analysis, `DESeq2` first uses the input read counts to build a model by fitting each gene to a generalised linear model (GLM), then uses an empirical Bayes shrinkage procedure to estimate dispersion and fold-change. Once the GLMs have been fitted for each gene, the model coefficients are tested for significance and reported as standard errors for each shrunken LFC. A Wald test is then used to test for significance of genes, and the genes whose P-values pass the filtering step are adjusted for multiple testing using the Benjamini and Hochberg procedure (Costa-Silva *et al.*, 2017; Love *et al.*, 2014). The shrinkage estimators in `DESeq2` improve the reproducibility of the results, allowing `DESeq2` to provide consistent performance for small and large datasets.

4.1.2 Enrichment of genes

Once the differentially expressed genes and transcripts have been identified, there is the challenge of extracting biological meaning relating to the study. This is typically achieved

by characterising the genes and transcripts in order to identify the roles they play in a given phenotype or condition. One approach is to group them into smaller subsets of related genes or transcripts using public resources or "knowledge bases", such as the Gene Ontology (GO) database, to identify genes that participate in the same/similar pathways. The GO database provides consistent gene or gene products descriptors (ontologies) and standardised classification for sequences and their features (Gene Ontology Consortium, 2004). By annotating the genes and transcripts with GO terms, we can identify genes that are involved in the same *molecular processes*, *biological process* or *cellular compartment*, allowing us to identify and focus on the pathways that differ between phenotypes. A more comprehensive tool, `enrichR` (Kuleshov *et al.*, 2016), that is not only limited to GO, but provides a wide variety of annotation libraries is also available. The dataset libraries for annotation in `enrichR` include GO, KEGG (Kyoto Encyclopedia of Genes and Genomes), Panther, GeneSigDB and many other database libraries. At the time of writing this thesis, 143 libraries were available (`http://amp.pharm.mssm.edu/Enrichr/#stats`) for annotation. `enrichR` can be used to annotate a set of genes online (`http://amp.pharm.mssm.edu/Enrichr/`) or using the `R` package (`https://cran.r-project.org/web/packages/enrichR/vignettes/enrichR.html`).

4.1.3 Pathway analyses

Pathway analysis provides insight into the underlying biology of the differentially expressed genes and transcripts; it goes beyond the simple GO term annotation. There are more advanced knowledge bases which provide gene-specific functional data, which includes information of where and how gene products interact with each other. These databases include KEGG (`https://www.genome.jp/kegg/`), Protein Information Resource (PIR; `https://proteininformationresource.org/`), Ensembl (`https://www.ensembl.org/index.html`), UniProt/Swiss-Prot (`https://www.uniprot.org/`) and Reactome (`https://reactome.org/`) (Khatri *et al.*, 2012). Even though these databases provide a plethora of functional data for a given gene, they are not designed to explore the biological knowledge associated with hundreds of genes in a coordinated manner. Tools such as the Database for Annotation, Visualisation and Integrated Discovery (DAVID; Dennis Jr *et al.*, 2003), Ingenuity Pathway Analysis (IPA; Jiménez-Marín *et al.*, 2009) and the Topology-based Pathway Analysis of microarray and RNA-Seq Data (`ToPASeq`; Ihnatova and Budinska, 2015) were designed to integrate the biological information with a set of genes in order to aid with the interpretation of the data.

For pathway analysis methods that go hand-in-hand with differential expression data, gene set analysis (GSA) methods such as GAGE (Generally Applicable Gene-set Enrichment) are available (Luo *et al.*, 2009). GAGE is able to perform GSE for pathway analysis by taking the genes that are significantly differentially expressed in RNA-seq data, along with their $\log_2$FC and use the information to identify pathways that are either up- or down-regulated by these genes. This information produced by GAGE can be coupled with

Pathview (Luo and Brouwer, 2013) to integrate the pathway information with pathway graphs downloaded from KEGG in order to visualise how the genes interact with other components in the pathways they are involved in.

4.2 Analyses

The read counts matrices produced by the Nextflow `rnaSeqCount` pipeline (described in Chapter 3) were used for the differential expression analysis described in this chapter. The analyses presented here are based on the read counts produced by the `htseq--count` program since it had a higher percentage of uniquely mapped reads across all samples as compared to the `featureCounts` program (Chapter 3, Figure 3.8). The differential expression analyses were carried out on the open-source `R` (version 3.5.1; https://www.r-project.org/) software for statistical computing and graphics, using the `DESeq2` (version 1.22.1; http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html; Love *et al.*, 2014) package available for `R`. Gene enrichment was carried out using the `enrichR` (version 1; https://cran.r-project.org/web/packages/enrichR/vignettes/enrichR.html; Kuleshov *et al.*, 2016) package. Pathway analyses were carried out using `gage` (version 2.32.0; Luo *et al.*, 2009 and `pathview` (version 1.22.0; Luo and Brouwer, 2013), with pathway graphs downloaded from KEGG. All other `R` packages used in these analyses are listed in Appendix L. The `R` script used for processing all the analysis in this chapter is available on GitHub (https://github.com/phelelani/transcriptomics/).

4.2.1 Pre-processing of read-count data

The read count matrix (`the.data`) was loaded onto `R` and the samples renamed with more meaningful names to represent the condition and site the samples were obtained from (as in Table 2.1). For patients, samples were named "P" and "B" for forearm and back tissue samples respectively, whereas for controls, samples were named "C" and "R" for forearm and breast tissue samples respectively. The numbers represent an individual in each of the patient and control groups. Once the names were assigned, demographic (sex and age) and clinical information was assigned to the individuals using a separate table (`the.coldata`). Both the read count matrix (`the.data`) and the sample information (`the.coldata`) are datasets required by the `DESeq2` program to perform differential expression analysis.

Before differential expression analysis could be carried out, pre-filtering of the samples and genes was carried out. Samples from patient 1 (P1 and B1) were removed from the dataset as there was no clinical information available. Samples from patient 9 (P9 and B9) were removed from the dataset in order to have females only in the patient group. Sample C1 was also removed from the control group in order to have samples coming from male forearms only in the control. Removal of samples P9, B9 and C1 was to reduce potential sex bias in the patient and control groups. Samples P9 and B9 are from the

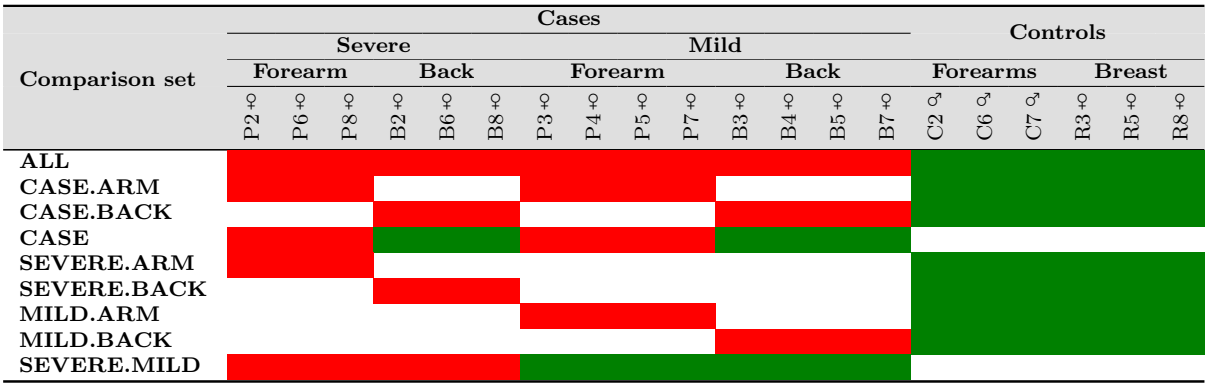
only male individual in the cases, whilst sample C1 is the only female forearm sample in the control group. Removing these samples leave only females in the cases group and only forearm samples from males in the control group. After removing these samples from the dataset, genes that had no reads mapped to them were excluded from the read counts matrix. Genes with low counts were also excluded from the dataset by removing genes where there are less than three samples with a read count of more than five. Only genes where there were three or more samples with a read count of more than 5 were kept.

The remaining genes were annotated with full gene names (description), gene biotype, chromosome location and HUGO Gene Nomenclature Committee (HGNC) symbols using the `biomaRt` (version 2.38.0; http://bioconductor.org/packages/release/bioc/html/biomaRt.html; Durinck *et al.*, 2009) package in `R`. The chromosome location in the annotation was used to remove the genes mapped to the Y chromosome in the dataset in order to further remove sex bias before performing differential expression analysis. X chromosome genes were not removed due to dosage compensation in females who have one inactivated X chromosome, gene expression of the majority of genes on the X chromosome are expected to be equivalent. Since the control group was composed of males (three forearm tissue samples) and females (three breast tissue samples), an extra step was carried out to remove sex and tissue site bias in the control group in order to use both forearm and breast samples together as a reference condition for differential expression comparisons against the patient group. In this "sanitisation" step, pre-filtering differential expression was carried out between the males and females in the control group, using the females (back) as normal "normal" reference condition and males as "affected" individuals. A FDR of 0.05 (5%) and an absolute $\log_2$FC threshold of 1.1 was used. The genes that were found to be differentially expressed between these two groups were excluded in the analysis under the assumption that they represent the differences between tissue site (forearm and breast) as well as sex differences (male and female).

Once all the genes with low read counts and those that may have an impact in tissue and sex bias were removed, the dataset was prepared for differential expression analysis. To ensure sensitivity in identifying differentially expressed genes between patients with SSc and unaffected individuals, a number of differential expression comparisons were carried out, which take into account the condition, source of tissue sample as well as severity of the disease in the affected individuals. These comparisons are summarised in Table 4.1. The first comparison, **"ALL"**, compares expression of genes between all patients (P2, B2, P3, B3, P4, B4, P5, B5, P6, B6, P7, B7, P8 and B8) against all controls (C2, C6, C7, R3, R5 and R8). The second comparison, **"CASE.ARM"**, compares gene expression between forearm samples (P2, P3, P4, P5, P6, P7 and P8) of patients against all control samples.

The third comparison, **"CASE.BACK"**, compares gene expression between back sam-

Table 4.1: Different sets of comparisons carried out in this study



Samples highlighted in red represent affected individuals and those in green represent unaffected individuals or the reference condition to compare against. For the "within-group" comparisons, CASE and SEVERE.MILD, the reference condition to compare against is represented in green.

ples (B2, B3, B4, B5, B6, B7 and B8) of patients with all control samples. The fourth comparison, **"CASE"**, is a within-group comparison of forearm and back samples in the patient group. The fifth comparison, **"SEVERE.ARM"**, compares the expression of genes in the forearm samples (P2, P6 and P8) of patients with the severe form of SSc against all controls. The **"SEVERE.BACK"** comparison compares the back samples (B2, B6 and B8) of patients with the severe form of the disease with all controls. The seventh comparison, **"MILD.ARM"** compares forearm samples (P3, P4, P5 and P7) in patients with the mild form of SSc with all controls. The eighth comparison, **"MILD.BACK"** compares gene expression in back samples (B3, B4, B5 and B7) of patients with the mild form of the disease with all controls. The last comparison, **"SEVERE.MILD"**, is the second within-group gene expression comparison between patients with the severe form of the disease against patients with the mild form of the disease. For the remainder of this thesis, these different types of comparisons will be referred to with their short set names as shown in Table 4.1.

4.2.1.1 Power and sample size analyses

Since the data used in this study was produced from a different study, and I had no control in selecting the number of samples, sex and age per group (as discussed in Section 1.11), power assessment for identifying differentially expressed genes using this dataset was calculated based the different comparisons and the number of individuals in each group. The `PROPER` method described by Wu *et al.* (2015) was used in `R` for power calculations. The *"Cheung"* data was used to create 20 simulations for performing differential expression using a $|\log_2 FC|$ of 2, 5% differentially expressed genes (of 25000 genes), FDR of 0.05 (5%) and the number of replicates corresponding to each comparison (Table 4.1). Since the `DESeq2` package was used in the analysis, the option method for detecting differentially expressed genes was set to "`DESeq2`". Power calculations in `R` using `PROPER` were performed as shown below.

```
## PROPER POWER CALCULATIONS
# Simulation options
```

```
sim.opts.Cheung <- RNAseq.SimOptions.2grp(ngenes = 25000,
                                          p.DE=0.05, lOD="cheung",
                                          lBaselineExpr="cheung",
                                          lfc=2)

# Perform simulations
simres = runSims(Nreps = c(3, 4, 6, 7, 14),
                 Nreps2 = c(6, 6, 8, 6, 6),
                 sim.opts=sim.opts.Cheung,
                 DEmethod="DESeq2", nsims=20)

# Calculate powers
powers = comparePower(simres, alpha.type="fdr", alpha.nominal=0.05,
                      stratify.by="expr", strata.filtered=1,
                      target.by = 'lfc')
```

4.2.1.2 Preparing comparison sets for expression analysis

To prepare the data in each of the nine comparisons for differential expression analysis, the read counts matrix of genes (`the.data`) with its associated information on samples (`the.coldata`) were used. For each comparison, `the.data` and `the.coldata` were subsetted to contain only the samples for that particular comparisons as summarised in Table 4.1. The samples remaining after subsetting in the `the.data` and `the.coldata` for each comparison were then passed to the `DESeq2`'s `DESeqDataSetFromMatrix` function along with a formula that describes the experimental design for each comparison. Since the individuals, conditions and sites differ between the nine comparisons, the design formulas were also different to take into account these differences (Table 4.2).

For the ALL comparison, the design formula used takes into account the condition and tissue site differences, as well as the within-individual condition effect between the forearm and back samples in the cases coming from the same individual. In the CASES.ARM, CASE.BACK, SEVERE.ARM, SEVERE.BACK, MILD.ARM and MILD.BACK comparisons, the experimental design formula used only takes into account the condition and sample site variations; there are no within-individual effects in these comparisons. However, the two within-group comparisons, CASE and SEVERE.MILD, have no condition effect

Table 4.2: Experimental design formulas used in each of the nine expression comparisons

Comparison set	Experiment design formula
ALL	`~ site + site:ind.n + site:condition`
CASE.ARM	`~ site + condition`
CASE.BACK	`~ site + condition`
CASE	`~ ind.n + site`
SEVERE.ARM	`~ site + condition`
SEVERE.BACK	`~ site + condition`
MILD.ARM	`~ site + condition`
MILD.BACK	`~ site + condition`
SEVERE.MILD	`~ site + site:anti.scl.n + site:anti.scl`

`condition`: case or control (reference condition to compare against for the within-group comparisons).
`site`: forearm or back.
`ind.n`: individuals nested within the patient group (forearm/back).
`anti.scl`: individuals nested within the patient group (severe/mild).

since they are comparing the affected individuals against each other. The only variations taken into account are the site differences, severity of the disease (in the SEVERE.MILD comparison) and individual effects. A final gene filtering step was then carried out on each of the comparison datasets in the `DESeqDataSetFromMatrix` to remove genes with low normalised read counts. In this filtering step, the normalised read counts were obtained through the `estimateSizeFactors` function of `DESeq2`. Genes with fewer than three samples with a normalised gene count of more than five were discarded. Genes that had three or more samples with with a normalised gene count of more than five were kept. These datasets were then used for differential gene expression.

4.2.1.3 Assessing sample-sample similarities

Before performing differential expression analysis, the similarities and differences between samples were assessed through principal component analysis (PCA) in order to determine whether they fit the experimental designs. However, the normalised counts could not be used directly to perform PCA as the genes with the highest counts would contribute the largest variance on the PCA (Love *et al.*, 2014). The commonly used method of taking the logarithms of the normalised counts and adding a value (pseudocount) of 1 could also not be used. This is because logarithms of small counts amplify the variance, thus genes with low counts would contribute the largest variance on the PCA. `DESeq2` overcomes these issues by using a "shrinkage" approach in its `rlog` function which stabilises the variance across the mean. This method takes logarithm-transformed values of genes with high counts, and shrinks together the values of the genes with low counts in different samples (Love *et al.*, 2014). The `rlog` function of `DESeq2` was used to transform the normalised read counts in the different comparisons, and the transformed data was used for PCA.

4.2.2 Running differential expression analysis with `DESeq2`

Differential expression in each of the comparisons was carried out using the `DESeq` function of `DESeq2`, which takes as input the dataset in the `DESeqDataSetFromMatrix`. The `results` function of `DESeq2` was then used to get the significantly differentiated genes from the results of the `DESeq` function. A significance threshold (FDR) and $\log_2$FC threshold were passed to the `results` function. To determine the best $\log_2$FC threshold to use across the comparison, differing values of $\log_2$FC threshold were tested across the comparison at a FDR of 0.01 (1%) to identify the number of significant genes returned by changing the $\log_2$FC from 0 - 2 in increments of 0.1. These results are shown in Table 4.3. A $\log_2$FC threshold of 2 was selected across the comparisons, except for the CASES and SEVERE.MILD comparisons, where a $\log_2$FC threshold cutoff of 1.1 was used. To visualise the differentially expressed genes (up- and down-regulated), outliers and significant genes that met both the $\log_2$FC and FDR thresholds, a volcano plot was generated for each comparison. A volcano plot is a scatter plot of genes with the absolute $\log_2$FC on the x-axis and -$\log_{10}$(P-value) on the y-axis.

Table 4.3: Genes returned by different $\log_2$FC cutoff using a fixed FDR of 0.01

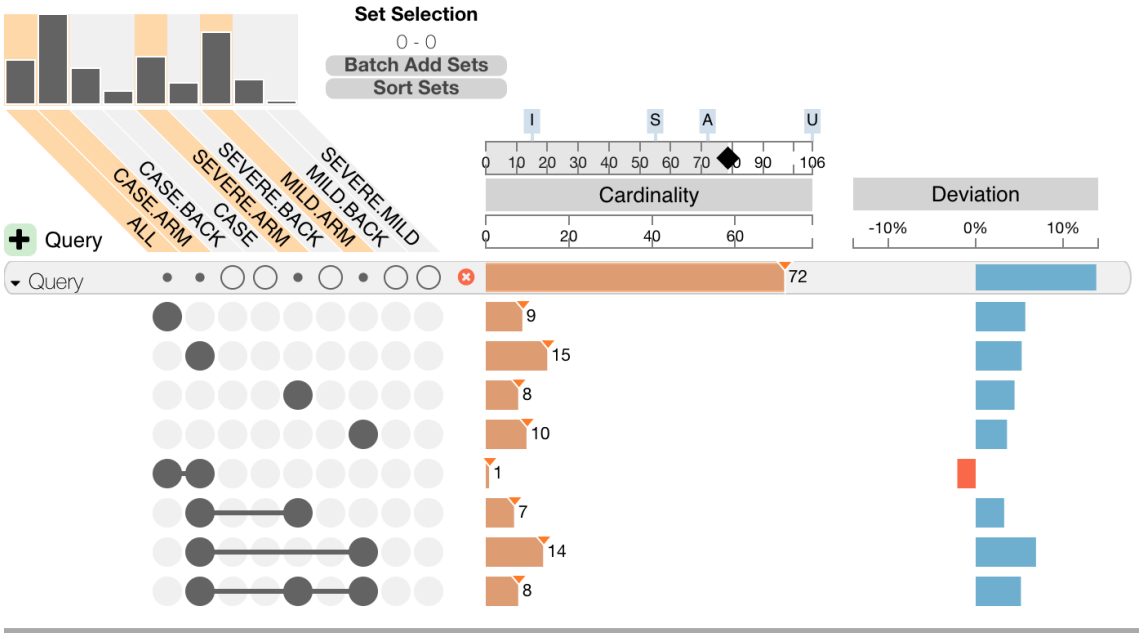
FC Cutoff	# of singnificant genes in each comparison								
	ALL	CASE ARM	CASE BACK	CASE	SEVERE ARM	SEVERE BACK	MILD ARM	MILD BACK	SEVERE MILD
0.00	290	2276	601	246	782	134	1407	322	9
0.10	245	1735	423	142	605	114	1095	255	7
0.20	210	1344	310	87	484	86	865	213	5
0.30	179	1041	238	60	371	70	712	183	4
0.40	143	817	210	43	310	60	574	146	3
0.50	128	661	165	33	268	52	479	115	3
0.60	109	529	139	26	227	46	401	98	3
0.70	100	442	120	19	195	42	344	90	2
0.80	92	370	97	15	173	41	292	81	2
0.90	83	308	79	14	148	34	249	70	2
1.00	75	261	71	11	128	31	203	66	2
1.10	67	212	61	8	111	28	179	57	2
1.20	60	189	53	5	94	27	151	46	2
1.30	58	170	48	5	85	25	129	44	1
1.40	51	148	45	5	70	22	114	41	1
1.50	42	128	40	3	62	22	98	28	1
1.60	38	112	38	3	52	19	82	26	1
1.70	33	95	36	3	43	18	69	25	1
1.80	31	85	32	2	37	17	60	20	1
1.90	27	64	26	2	32	15	48	16	1
2.00	27	55	22	1	29	13	44	15	1

To visualise the significantly differentiated genes across the samples in all nine comparisons, heatmaps were created. Genes that have been implicated in SSc as drug targets were downloaded from the Open Targets Platform (`https://www.targetvalidation.org/`) to determine if any of these genes were found to be differentially expressed in these analyses. The Open Targets Platform is a database for the identification and prioritisation of potential drug targets for diseases (Koscielny *et al.*, 2017). 870 genes were found to be associated targets with SSc. These genes were highlighted on the heatmaps for each comparison.

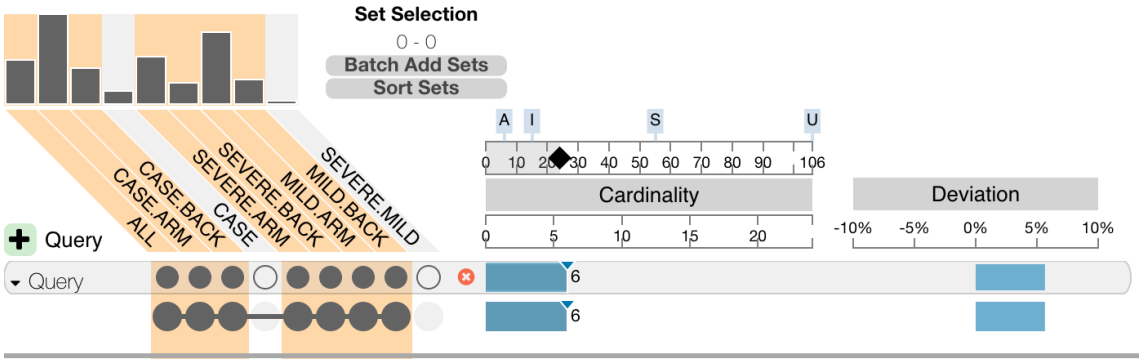
4.2.3 Prioritisation of significant genes with `UpSet`

The `UpSet` visualisation program (Conway *et al.*, 2017; Lex *et al.*, 2014) was used for the prioritisation of significantly differentially expressed genes between the nine different comparison. `UpSet` is a tool for quantitative analysis of large datasets, the intersections of their elements as well as aggregates of intersections, which would be rather be difficult, if not impossible, to analyse using common methods such as Venn diagrams. In an `UpSet` plot (Figure 4.1), there are three views for analysing dataset, i.e., *sets* (for intersections, unions and aggregates of elements), *cardinality* (number of elements in an intersection) and *deviation* (disproportionality to expected cardinality). Under the sets view, the columns of the matrix represents the sets (or dataset) and rows represent intersections, which are equivalent to Venn diagrams. If a set has an element participating in an intersection, the matrix cell is filled (●). If a set has no element participating in the intersection, the cell in the matrix is not filled (●). Intersections that are exclusive (filled circles) are connected by lines and cross-over excluded sets in each row. Queries can be created to select sets that either share of have unique elements in them.

The genes that were found to be common and/or unique between the most important comparisons were selected. The important comparisons were those that had the affected forearm of the patients, i.e., ALL, CASE.ARM, SEVERE.ARM and SEVERE.MILD, against the controls. All genes that were significant in these four sets, and not found in the within-group comparison sets (CASE and SEVERE.MILD) and comparisons with back samples (CASE.BACK, SEVERE.BACK and MILD.BACK) were selected (Figure 4.1a). The within-group comparison sets were excluded from the prioritisation because they are comparisons of affected individuals; genes that show significant differential expression are a result of individual differences and not a result of disease. The comparison sets with back samples were excluded because they were taken from sites that had no disease phenotype.



(a) `UpSet` query for identifying genes that are shared (intersects) and/or unique between the ALL, CASE.ARM, SEVERE.ARM and MILD.ARM comparison sets.



(b) `UpSet` query for identifying genes that are shared (intersect) amongst all the comparisons, except the CASE and SEVERE.ARM within-group comparisons.

Figure 4.1: `UpSet` queries for gene prioritisation. Comparison sets with genes participating in the queries are have filled circles (●) in the matrix. Comparison sets with no genes participating in the queries have empty circles in the matrix (●). Sets with genes that are shared between them are connected by lines.

To increase the sensitivity in identifying genes that might be associated with SSc, the `UpSet` visualisation program was again used to identify genes that are shared between all the comparison sets except the CASE and SEVERE.MILD sets (Figure 4.1b). This set was included under the assumption that genes associated with SSc might not only be expressed in the affected sites (forearms), but there might be some expression of these genes on other unaffected tissues (back) as, well which is consistent with the expression in the affected tissue.

4.2.4 Gene-set enrichment with `enrichR`

Gene-set enrichment was carried out using the `enrichR` (version 1.0) package in `R`. The databases for GO terms (Molecular Function 2018, Biological Process 2018 and Cellular Component 2018), pathways (WikiPathways 2016, KEGG 2016, BioCarta 2016, Panther 2016 and Reactome 2016) and disease (OMIM Disease, OMIM Expanded, Jensen DISEASES, Jensen COMPARTMENTS and Jensen TISSUES) were selected from the 130 databases offered by the `enrichR` package. Enrichment was carried out for significant genes in each of the nine different comparison sets, all of the nine sets combined, as well as for the genes that were prioritised (Section 4.2.3). The enrichment terms for genes-sets were selected based on an adjusted P-value of 0.05 (5%).

4.2.5 Pathway analysis with `gage` and `pathview`

Pathway analyses were carried out using `gage` (version 2.32.0) and `pathview` (version 1.22.0). First, for each comparison set, the differential expression results were sorted by the adjusted P-value and the genes mapped to ENTREZ IDs. The $\log_2$FCs for the genes were then extracted and used for pathway analysis using `gage` and KEGG human pathway sets (signalling and metabolic pathways). The most significant up- and down-regulated pathways (q-value of less than 0.05) identified by `gage` were then selected and passed onto `pathview`, which downloaded pathway graphs from KEGG and used the $\log_2$FC information from `gage` to draw the pathways for visualisation. Tables for the up- and down-regulated pathways identified for each comparison set, along with the statistics, were also saved.

4.3 Results

All the analyses carried out for differential expression analysis were carried out in `R`, since all the packages for performing the analysis are available for `R`. The `R` software is a very powerful platform for statistical analysis and visualisation of large datasets. This coupled with the wide variety of RNA-seq packages available for analysis on this platform make `R` a powerful tool for performing differential expression analysis. The workflow script used in performing the analysis, visualisation of data and producing the results presented here is available on GitHub (https://github.com/phelelani/transcriptomics).

4.3.1 Power and sample size analysis

The results for power assessment of the data are shown in Table 4.4 and Figure 4.2. Table 4.4 summarises the related information over the different count strata (average number of reads per gene in a sample) in the different sample sizes (SS). SS=3,6 corresponds to the SEVERE.ARM and SEVERE.BACK comparisons; SS=4,6 corresponds to the MILD.ARM and MILD.BACK comparisons; SS=6,8 corresponds to the SEVERE.MILD comparison; SS=7,6 corresponds to the CASE.ARM and CASE.BACK comparisons; SS=7,7 corresponds to the CASE comparison; and SS=14,6 corresponds to the ALL comparison. The average power for detecting differentially expressed at a FDR of 0.05 and $|\log_2 FC|$ of 2 for different sample sizes corresponding to the comparison ranges from 59% - 68%.

The number of true discoveries (TD) increases with sample size, and the number of false discoveries (FD) decreases with the sample size. However, even though the summarised marginal power for detecting differentially expressed genes is below 70%, when looking at the individual sample sizes in Figure 4.2, the power increases with the average number of reads (count strata) for each gene in each sample size to over 80%. Overall, with the number of samples available, this study is highly powered to perform differential expression.

4.3.2 Data pre-processing and differential expression analysis

The initial read counts matrix from `htseq-count` used for the differential expression analysis consisted of 58 676 genes. This is the total number of genes that are found in the annotation file during mapping and gene quantification, some of which have no reads mapped to. After the first filtering step of removing genes with no reads mapped, 39 318 genes remained (Table 4.5). The second filtering step, where genes with less than 3 samples with read counts more than 5 were removed, resulted in 25 021 genes remaining in the dataset. After annotating the 25 021 genes with chromosomal location and removing genes mapping to the Y chromosome (Appendix D), 25 000 genes were left in the dataset. After performing differential expression between the males and females in the control group, 4 genes were found to be differentially expressed (Appendix D); 24 996

Table 4.4: Summary of the power calculations performed with `PROPER`.

SS1	SS2	Nominal FDR	Actual FDR	Marginal power	Avg # of TD*	Avg # of FD†	FDC‡
3	6	0.05	0.18	0.59	161.7	35.55	-
4	6	0.05	0.16	0.63	172.8	33.75	-
6	8	0.05	0.12	0.68	191.4	25.8	-
7	6	0.05	0.12	0.66	187.35	26	-
14	6	0.05	0.093	0.68	198.4	20.45	-
7	7	0.05	0.11	0.68	192	23.9	-

* TD: True discoveries.
† FD: False discoveries.
‡ FDC: False discovery cost (not calculated for this data).

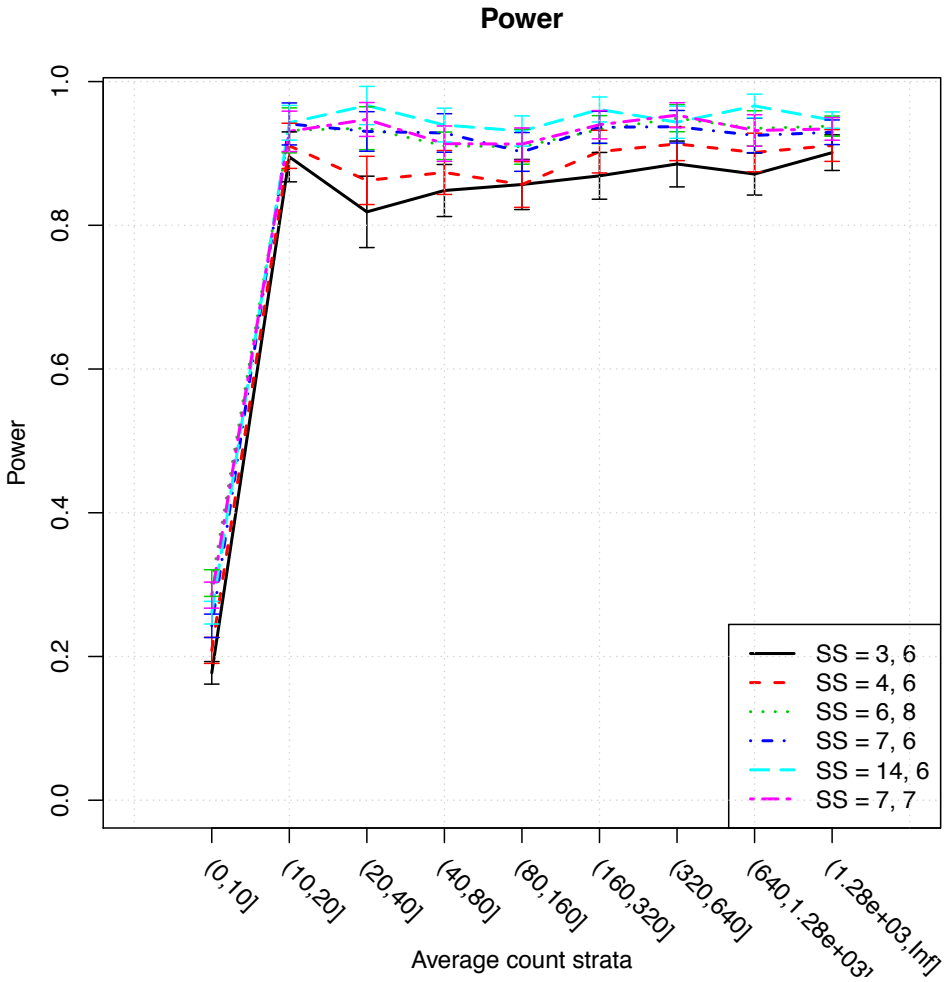


Figure 4.2: Power calculations using PROPER for the different sample sizes. Power for detecting differentially expressed genes for each sample size at different count strata (average number of reads). *SS=3,6*: SEVERE.ARM and SEVERE.BACK; *SS=4,6*: MILD.ARM and MILD.BACK; *SS=6,8*: SEVERE.MILD; *SS=7,6* CASE.ARM and CASE.BACK; *SS=7,7*: CASE; and *SS=14,6* comparison. Calculations were performed for 25 000 genes at 0.05 FDR, LFC of 2 and 5% genes to be differentially expressed.

genes remained after removing these genes from the dataset. Table 4.5 summarises the the number of genes remaining after each filtering step.

The remaining 24 996 genes were used in each of the nine comparisons to perform differential expression. Before performing differential expression analysis, further filtering was applied to the datasets in each comparison using the design formulas and removing genes with less than 3 samples with normalised read counts more than 5. The ALL, CASE.ARM, CASE.BACK, CASE, SEVERE.ARM, SEVERE.BACK, MILD.ARM, MILD.BACK and SEVERE.MILD had 24 368, 23 224, 22 983, 23 664, 21 941, 21 673, 22 404, 22 290 and 23 664, respectively, after this final filtering step (Table 4.5). Differential gene expression was performed with these genes in each of the comparison sets. After applying the FDR and $\log_2$FC to the differential expression results, the ALL, CASE.ARM, CASE.BACK, CASE, SEVERE.ARM, SEVERE.BACK, MILD.ARM, MILD.BACK and SEVERE.MILD comparison sets had 27, 55, 22, 8, 29, 13, 44, 15 and 2 significantly expressed genes, respectively (Table 4.5). Tables I.1 - I.9 in Appendix I list the significantly

Table 4.5: Summary of the pre-processing steps carried out

Set	# of genes at each filtering step						
	Genes	Step 1*	Step 2†	Step 3‡	Step 4§	Step 5¶	Step 6‡
ALL	58676	39318	25021	25000	24996	24368	27
CASE.ARM	58676	39318	25021	25000	24996	23224	55
CASE.BACK	58676	39318	25021	25000	24996	22983	22
CASE	58676	39318	25021	25000	24996	23664	8
SEVERE.ARM	58676	39318	25021	25000	24996	21941	29
SEVERE.BACK	58676	39318	25021	25000	24996	21673	13
MILD.ARM	58676	39318	25021	25000	24996	22404	44
MILD.BACK	58676	39318	25021	25000	24996	22290	15
SEVERE.MILD	58676	39318	25021	25000	24996	23664	2

* Step 1: Removing genes where no reads mapped.
† Step 2: Removing genes with less than 3 samples with read counts less than 5.
‡ Step 3: Removing genes on the Y chromosome.
§ Step 4: Removing genes differentially expressed between controls.
¶ Step 5: Removing genes with less than 3 samples with normalised read counts less than 5 and applying design formula for each comparison set.
|| Step 6: Applying FRD and Log2FC after differential expression.

differentially expressed genes for each of the nine comparisons, along with the gene name (Description), chromosome (Chr), the average of the normalised counts of the gene taken over all samples (baseMean), fold change between the groups being compared ($\log_2$FC), standard error of the $\log_2$FC estimate (lfcSE), Wald statistic (stat), Wald test *P*-value (*P*-value) and the Benjamini-Hochberg adjusted *P*-value (padj).

4.3.3 Sample-sample similarities

When looking at the PCA for the different comparisons (Figure 4.3 and Appendix E), we can see a clear clustering of the SSc patients separate from the controls. However, in almost all of the PCAc, sample C7 seems to be an outlier as it does not form part of the clusters with the other control samples. Even though this sample C7 showed this type of behaviour, it was not excluded from the analysis as the reduced sample size would undermine the differential expression analysis. For the comparisons involving the affected arms, approximately 40% of the variance can be explained by the first principal component, whilst the second principal component can explain approximately 20% of the variance.

4.3.4 Visualisation of differential expression results

To demonstrate the effects of performing filtering using $\log_2$FC and FDR cutoffs to get significantly differentially expressed genes, volcano plots were plotted for the unfiltered results in order to view the significant genes identified in the context of the unfiltered genes. Volcano plots represent the $|\log_2\text{FC}|$ on the x-axis and $-\log_{10}$(*P*-value) on the y-axis, thus, because the *P*-values of the genes are log transformed, the genes with the smallest *P*-value (significant) have high values on the y-axis, and those with a high $|\log_2\text{FC}|$ appear on the far ends of the x-axis. Using the combination of the $|\log_2\text{FC}|$ and $-\log_{10}$(*P*-value), the most significantly differentially expressed genes will be scattered towards the upper corners of the plot. Figure 4.4 and Appendix F show the visual representation of the significant genes identified (in green) in the context of the unfiltered

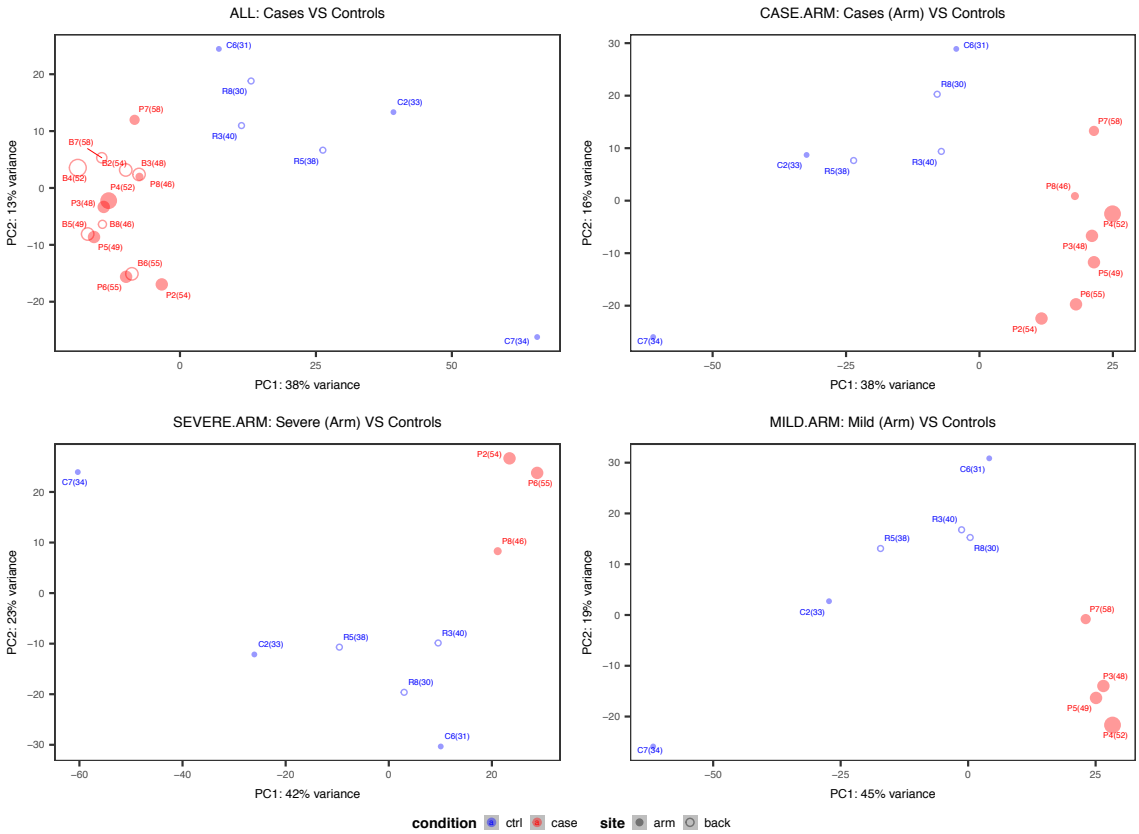


Figure 4.3: PCA of the comparison sets with the affected forearms included. ***Red:*** affected individuals; ***blue:*** control individuals; ***open circle:*** back sample (or breast in the case of controls); ***solid circle:*** forearm sample; ***numbers in brackets:*** age of the individuals.

differential expression results in a form of volcano plots.

To produce these volcano plots, the differential expression data ($\log_2$FC and P-values), before applying filters, were used. To visualise the $\log_2$FC cutoffs in the plots, vertical lines were drawn for the $|\log_2\text{FC}|$ cutoffs. For the significance cutoff, i.e. the $-\log_{10}()$ transformed P-values, the maximum P-value corresponding to the $\log_2$FC and FDR specified for the filtering applied on each of the different comparisons was calculated, and the $-\log_{10}()$ transformation of the P-value drawn as a horizontal line on each of the volcano plots. In Figure 4.4 and Appendix F, the genes in gray are those that did not pass both the $|\log_2\text{FC}|$ and significance thresholds used in that particular comparison. The genes in red are those that met the FDR, but did not pass the $|\log_2\text{FC}|$.

The genes in orange are those that met the $|\log_2\text{FC}|$ threshold, but did not pass the significance cutoff. The genes in blue are those that pass both the significance and $|\log_2\text{FC}|$. The genes in green are those that were found to be significant when the $\log_2$FC and FDR filters were applied to the unfiltered data. It is worth noting that the method used by `DESeq2` to perform filtering is superior compared to other methods, for example, sorting by P-value and selecting the top n number of genes as being significant. If the filtering step was not applied to the differential expression results, the genes in blue would have been thought to be significant, yet they are false positives.

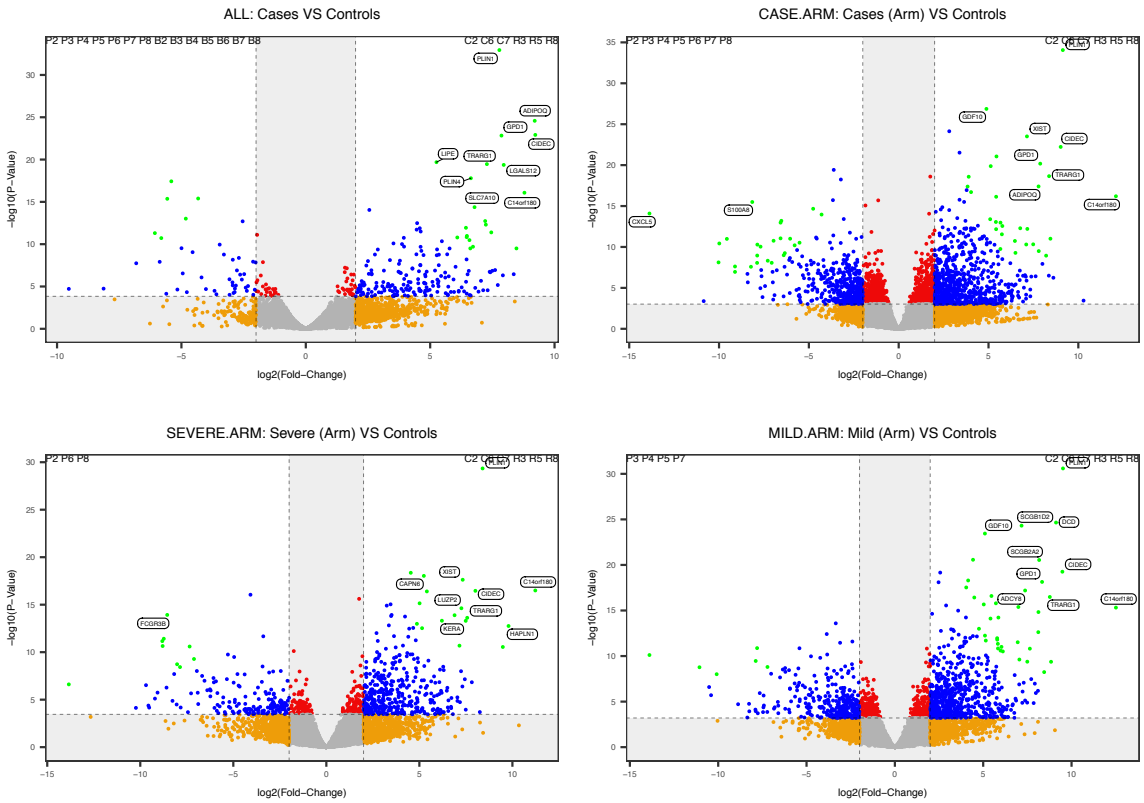


Figure 4.4: Volcano plots for the differential expression results and filtering for comparisons sets with forearms included. ***Green:*** significant genes after filtering; ***gray:*** genes that did not meet the $|\log_2 FC|$ and significance threshold; ***red:*** genes that met the significance threshold only; ***orange:*** genes that met the $|\log_2 FC|$ threshold only; and ***blue:*** genes that met both $|\log_2 FC|$ and significance thresholds. The names of the top ten significant genes are shown and the samples being compared (cases and controls) in each comparison are listed on the top corners of the plots.

4.3.5 Visualisation and clustering of significant genes

To visualise the significant genes, their levels/signals of expression and how they cluster together in terms of these expression signals, heatmaps were used plotted for each comparison. The heatmaps also show how the samples (cases and controls) from used in the comparisons cluster according to condition, sample site, sex, presence of lung fibrosis, presence of muscle weakness, reaction to anti-Scl-70 antibodies based on the expression levels of the significant genes identified. Figures 4.5 - 4.7 and Appendices G.1 - G.5 show the heatmaps for each of the nine comparisons performed in these analysis.

For the ALL comparison with 27 genes found to be significantly differentially expressed (Figure 4.5), three clusters of genes can be seen: cluster 1 starting from *HAND2* to *FOS* genes (6 genes), which seems to be up-regulated in the controls and down-regulated in the cases; cluster 2 starting from *SPX* to *PLIN4* genes (16 genes), which seems to be down-regulated in the cases and up-regulated in the controls; and the cluster 3 starting from *IGLV7* to *IGKV3* genes (5 genes), which is down-regulated in the controls and up-regulated in the cases. In total, 6 genes are down-regulated (when compared the normal condition) and 21 genes are up-regulated in the ALL comparison, as also confirmed by the volcano plot (ALL comparison plot in Figure 4.4). The clustering of the samples

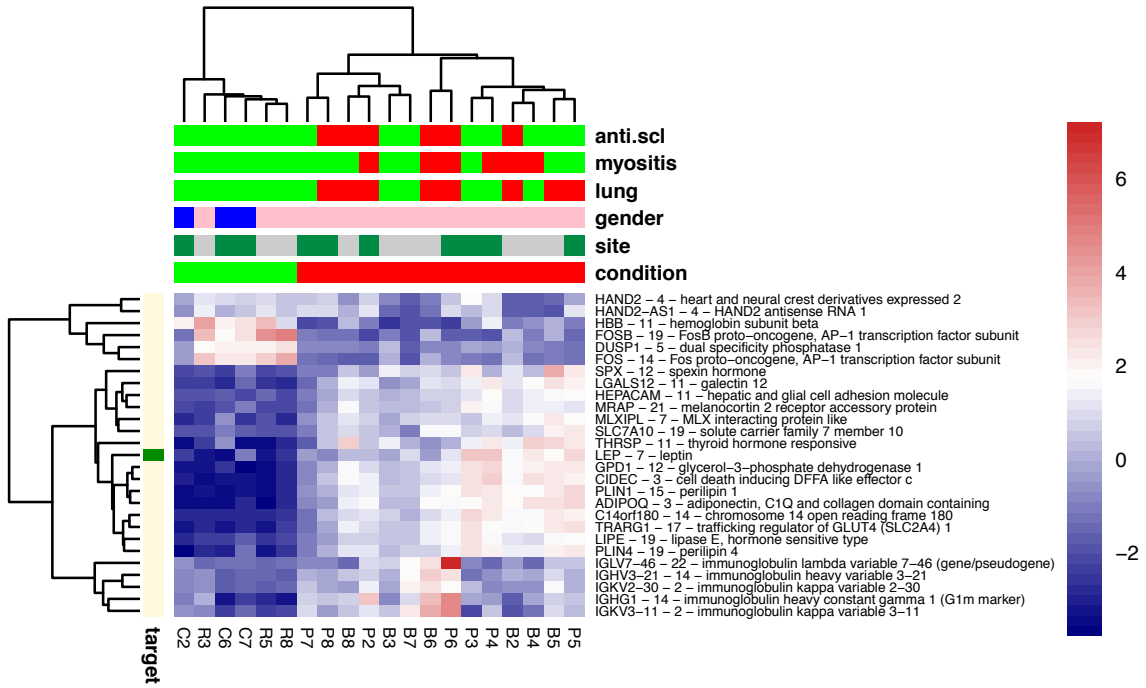


Figure 4.5: Heatmap showing gene clustering according to gene expression signals in the ALL comparison. The color for each cell in the matrix represents gene expression signal from down-regulated (navy), to no change (white), to up-regulated (red). ***condition:*** red = case, green = control; ***site:*** green = forearm, gray = back/breast; ***sex:*** blue = male, pink = female; ***lung:*** red = evidence of lung fibrosis, green = no lung fibrosis; ***myositis:*** red = muscle weakness/inflammation present, green = no muscle weakness/inflammation; ***anti.scl:*** red = positive anti-Scl-70 test, green = negative anti-Scl-70; ***target:*** green = gene on the Open Targets Platform for SSc, yellow = gene not found on the Open Targets Platform for SSc.

corresponds to the conditions, i.e., all control samples are to the left of the heatmap and all case samples are to the right of the heatmap. However, there is no clear clustering of the samples within each condition in terms of the sex and site, as well as other clinical features of the samples. Some samples do, however, cluster together but not fully, in a sense that the back samples in the cases are mixed with the forearms and the same is seen with the controls. One would expect (in the patients) that either the samples from the same individual would be seen together, or all the back samples would cluster separately from the forearm samples. Only one gene, *LEP* (leptin) was found to be a possible target for SSc.

In the CASE.ARM comparison, 55 genes were found to be significantly differentially expressed, of which 30 were up-regulated and 25 were down-regulated. There are only two clusters of genes in the CASE.ARM comparison (Figure 4.6): cluster 1 starting from *COMP* to *NAT8L* genes and is composed of 30 up-regulated genes; and cluster 2 starting from *MMP1* to *CXCL1* genes and is made up of 25 down-regulated genes. Interestingly, the cluster 2 of the down-regulated genes is mainly composed of genes that have been identified as possible targets for SSc in the Open Targets Platform, with only just one gene, *COMP*, identified as a target for the up-regulated cluster.

When looking of the composition of genes in the cluster, cluster 1 seems to have a mixed

variety of genes whilst cluster 2 seems to be mainly composed of transcription factors, interleukines and chemokines (Figure 4.6). The clustering of the samples in the CASE.ARM comparison does not seem to follow the conditions being compared, but rather they cluster according to the tissue site of the samples. All the forearm samples (green) are to the left of the heatmap, whilst the back/breast samples are to the right of the heatmap (gray). It is also worth noting that when looking at the samples clustering in terms of condition, all the case forearm samples are clustered together, but the control samples are "flanking" the case cluster on either side.

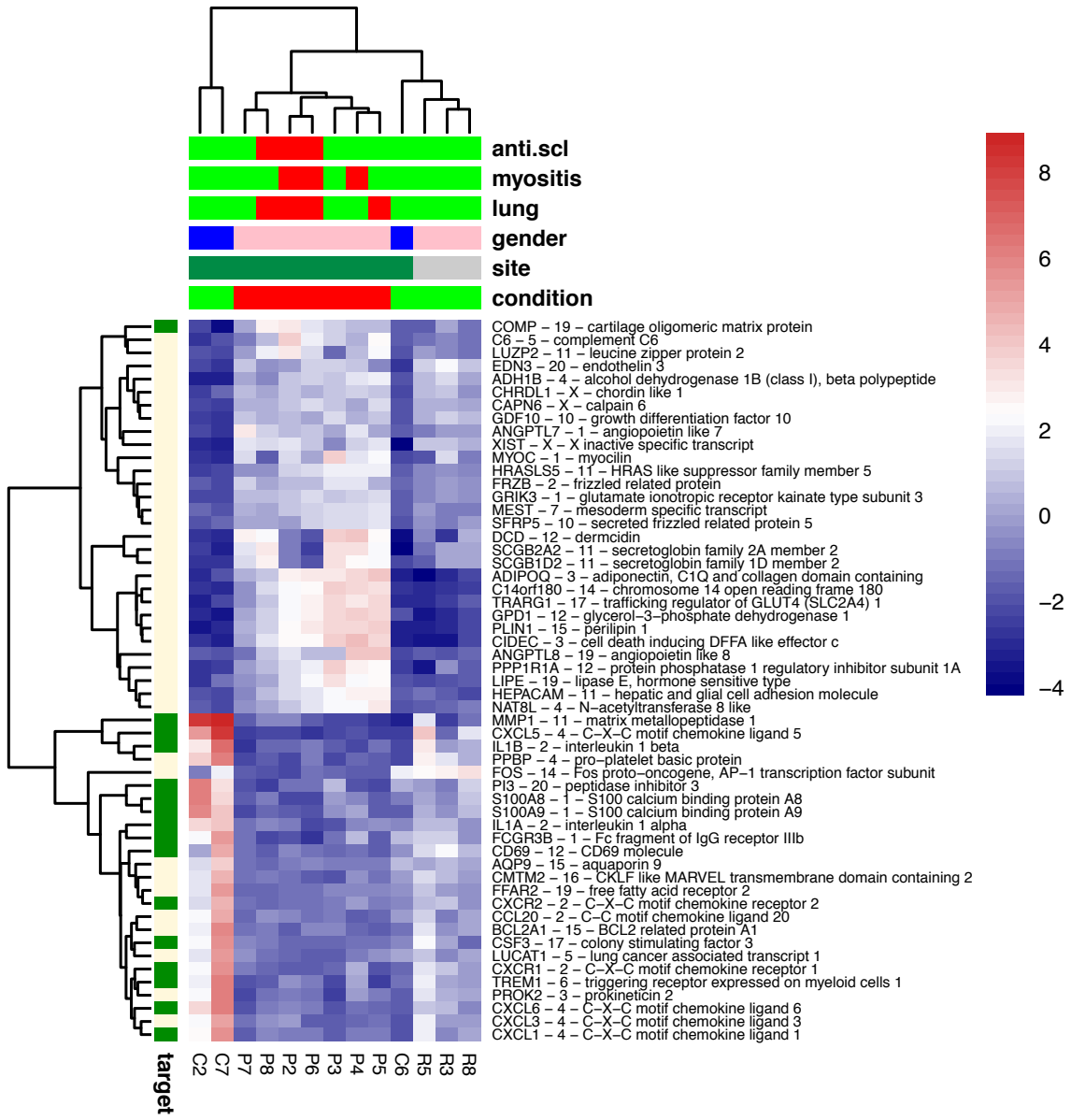


Figure 4.6: Heatmap showing gene clustering according to gene expression signals in the CASE.ARM comparison. The color for each cell in the matrix represents gene expression signal from down-regulated (navy), to no change (white), to up-regulated (red). ***condition:*** red = case, green = control; ***site:*** green = forearm, gray = back/breast; ***sex:*** blue = male, pink = female; ***lung:*** red = evidence of lung fibrosis, green = no lung fibrosis; ***myositis:*** red = muscle weakness/inflammation present, green = no muscle weakness/inflammation; ***anti.scl:*** red = positive anti-Scl-70 test, green = negative anti-Scl-70; ***target:*** green = gene on the Open Targets Platform for SSc, yellow = gene not found on the Open Targets Platform for SSc.

The SEVERE.ARM comparison resulted in 29 significantly differentially expressed genes, of which 20 were up-regulated and 9 were down-regulated. Figure 4.7 shows the clustering of the genes and their expression signals in their respective samples. Two clusters of genes are also visible in the heatmap for this comparison: cluster 1 starting from *CILP2* to *TRARG1* composed of 20 up-regulated genes; cluster 2 starting from *HBA1* to *S100A9* composed of 9 down-regulated genes. Almost all of the down-regulated genes were identified as targets for SSc. The clustering of the samples in this comparison follows the sample conditions, presence of lung fibrosis, reactivity to anti-Scl-70 as well as presence of muscle weakness and inflammation. All the patient samples in this comparison had the severe form of SSc, which may explain the distinct clustering of samples according to condition and clinical features.

44 genes were identified as significantly differentially expressed in the MILD.ARM com-

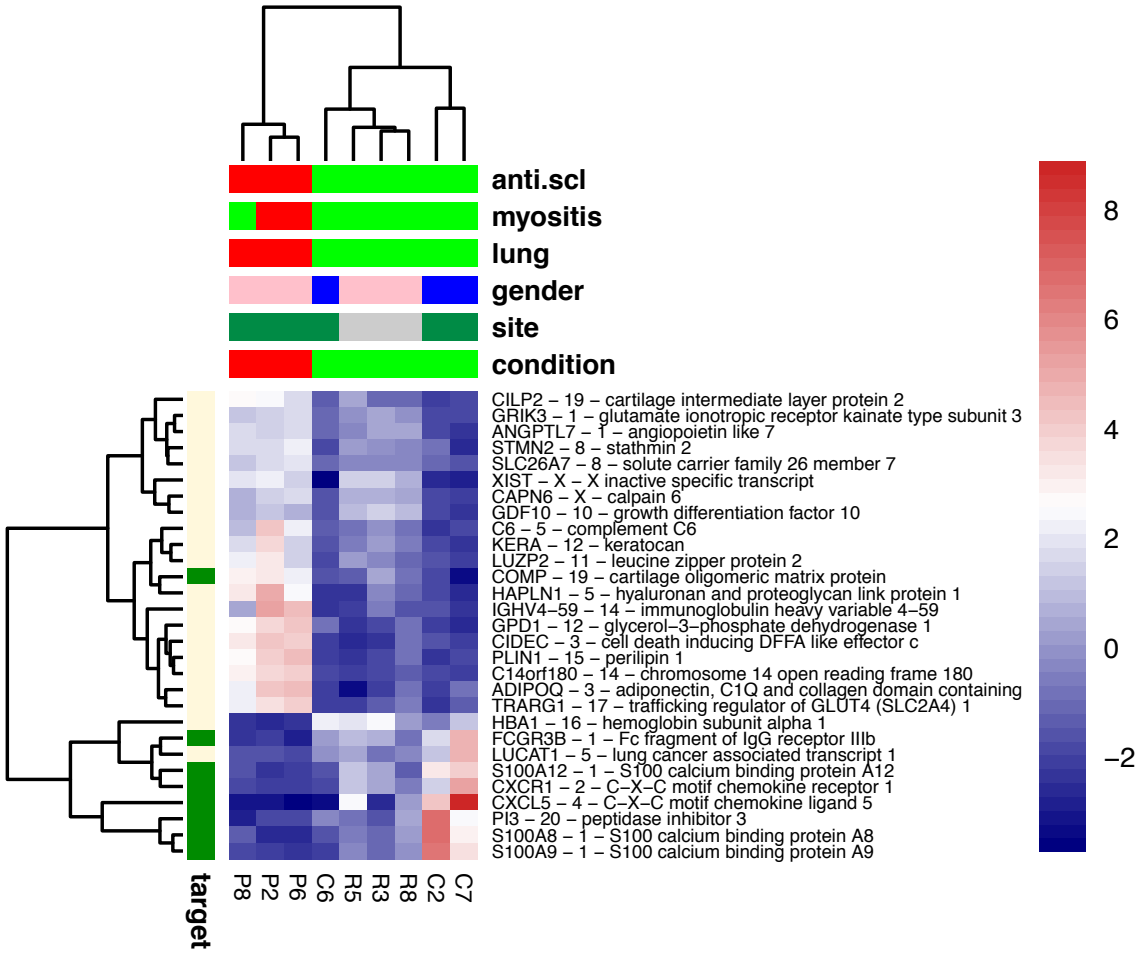


Figure 4.7: Heatmap showing gene clustering according to gene expression signals in the SEVERE.ARM comparison. The color for each cell in the matrix represents gene expression signal from down-regulated (navy), to no change (white), to up-regulated (red). ***condition:*** red = case, green = control; ***site:*** green = forearm, gray = back/breast; ***sex:*** blue = male, pink = female; ***lung:*** red = evidence of lung fibrosis, green = no lung fibrosis; ***myositis:*** red = muscle weakness/inflammation present, green = no muscle weakness/inflammation; ***anti.scl:*** red = positive anti-Scl-70 test, green = negative anti-Scl-70; ***target:*** green = gene on the Open Targets Platform for SSc, yellow = gene not found on the Open Targets Platform for SSc

.

parison; 38 of these genes were up-regulated and 6 were down-regulated. The clustering of the genes on the heatmap of this comparison also seems to follow the direction of the regulation of the genes (Figure 4.8): cluster 1 starting from *CXCL5* to *IL1B* composed of 6 down-regulated genes; and cluster 2 starting from *XIST* to *PPP1R1A* composed of 38 up-regulated genes. 5 of the 6 down-regulated genes in cluster 1 were identified as targets for SSc in the Open Targets Platform. The clustering of the samples in the heatmap follows the condition of the samples.

Although the other comparisons involving either the back samples (CASE.BACK, SEVERE.BACK and MILD.BACK) are for the within individual comparisons (CASE and

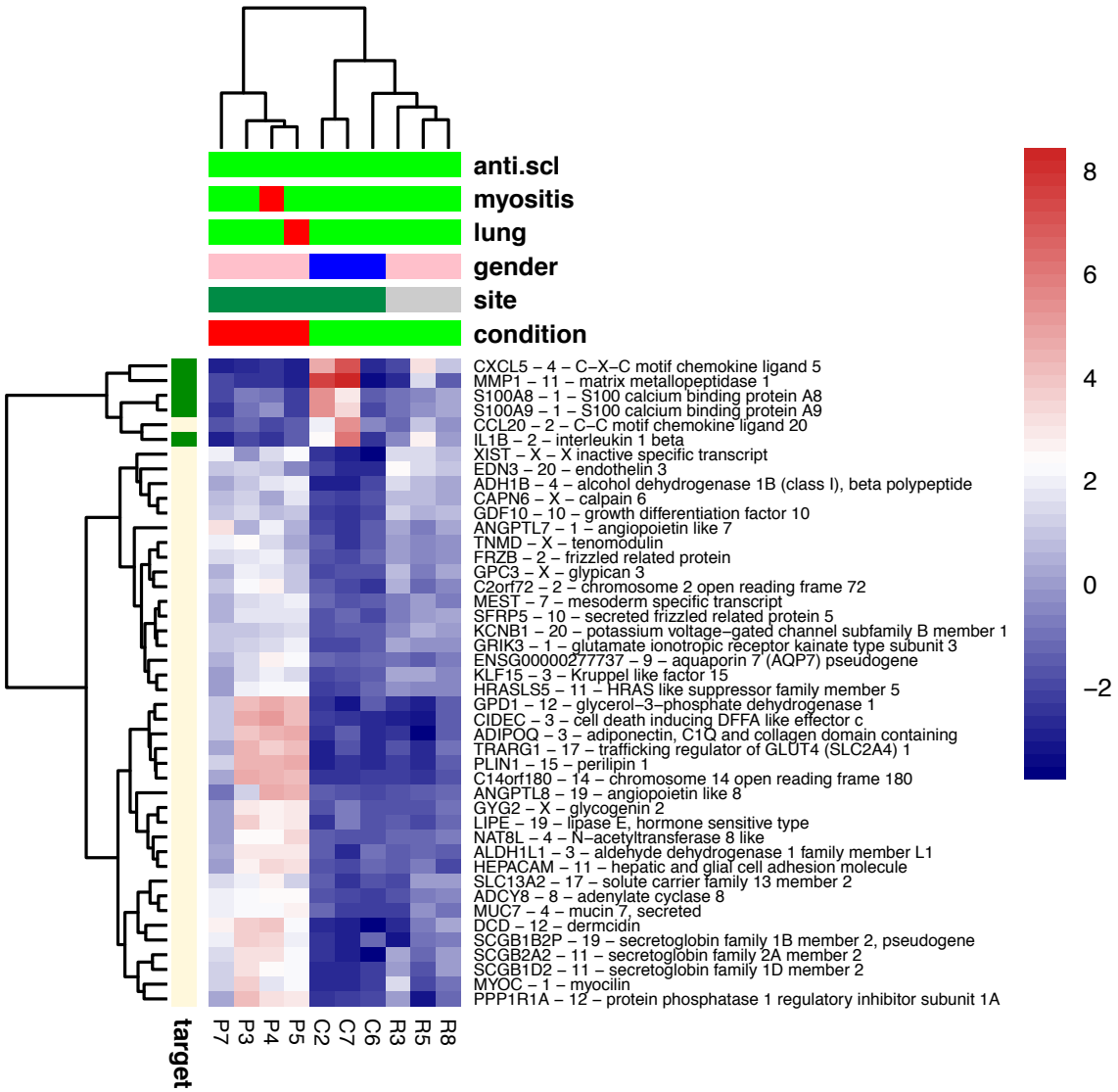


Figure 4.8: Heatmap showing gene clustering according to gene expression signals in the MILD.ARM comparison. The color for each cell in the matrix represents gene expression signal from down-regulated (navy), to no change (white), to up-regulated (red). ***condition:*** red = case, green = control; ***site:*** green = forearm, gray = back/breast; ***sex:*** blue = male, pink = female; ***lung:*** red = evidence of lung fibrosis, green = no lung fibrosis; ***myositis:*** red = muscle weakness/inflammation present, green = no muscle weakness/inflammation; ***anti.scl:*** red = positive anti-Scl-70 test, green = negative anti-Scl-70; ***target:*** green = gene on the Open Targets Platform for SSc, yellow = gene not found on the Open Targets Platform for SSc.

SEVERE.MILD) will not be covered in detail in these results, it is worth mentioning the number of differentially expressed genes found to be significant and how many were up- and down-regulated. For the CASE.BACK comparison (Appendix G.1), 22 genes were found to be significant, of which 21 were up-regulated and 1 was down-regulated. 13 significant genes were identified for the SEVERE.BACK, all of which were up-regulated (Appendix G.3). 15 genes were significant for the MILD.BACK comparison, 13 were up-regulated and 2 were down-regulated (Appendix G.4). The CASE within individual comparison had 8 significant genes, 4 were up-regulated and the other 4 were down-regulated (Appendix G.2). The SEVERE.MILD only had 2 significant genes, 1 up- and 1 down-regulated (Appendix G.5).

4.3.6 Prioritisation

The `R` version of `UpSet`, `UpSetR`, was used to obtain the genes that were prioritised in order to perform analysis (gene enrichment and visualisation) of the genes. The same concepts for filtering explained in the methods section (Section 4.2.3) using the web version of `UpSet` were applied to the `R` version of `UpSet`, `UpSetR`. In total, 78 genes were identified using the prioritisation method. However, when performing the gene-set enrichment with `enrichR` (Section 4.3.7), 1 gene was removed as it did not have an HGNC symbol required for the analysis. Figure 4.9 summarises the prioritisation of the genes from all the different comparisons based on uniqueness and commonality.

Figure 4.9 (top panel) shows that there were 9 genes (brown bar) that are shared between all the comparisons except the within individual comparisons (CASE and SEVERE.MILD). There were, in total, 30 genes (green bars) shared amongst the comparisons involving affected arm samples (ALL, CASE.ARM, SEVERE.ARM and MILD.ARM) and not found in the within group comparisons (CASE and SEVERE.MILD) and comparisons with the back samples (CASE.BACK, SEVERE.BACK and MILD.BACK). The unique genes in the comparisons with affected forearms, i.e., ALL (9, blue bar), CASE.ARM (15, red bar), SEVERE.ARM (8, orange bar) and MILD.ARM (10, purple bar), were also added to the list of prioritisation as these are not shared with the within-individual and back comparisons.

To visualise the behaviour of the selected 77 genes (Appendix H) in the different comparisons, additional volcano sub-plots were added to the `UpSetR` plot. These sub-plots used the data from differential expression analysis (P-value and $\log_2$FC values) to plot significance and highlight the selected 77 genes in the volcano plot for that comparison using the colours of the intersection bars. The volcano plots plotted with the 77 genes highlighted shows that the genes act differently in each comparison. Some comparisons do not have all the genes selected due to the expression of the genes and different samples used in each comparison. For the comparison with the affected forearms (ALL, CASE.ARM, SEVERE.ARM and MILD.ARM), the genes seem to behave in a similar manner, suggesting

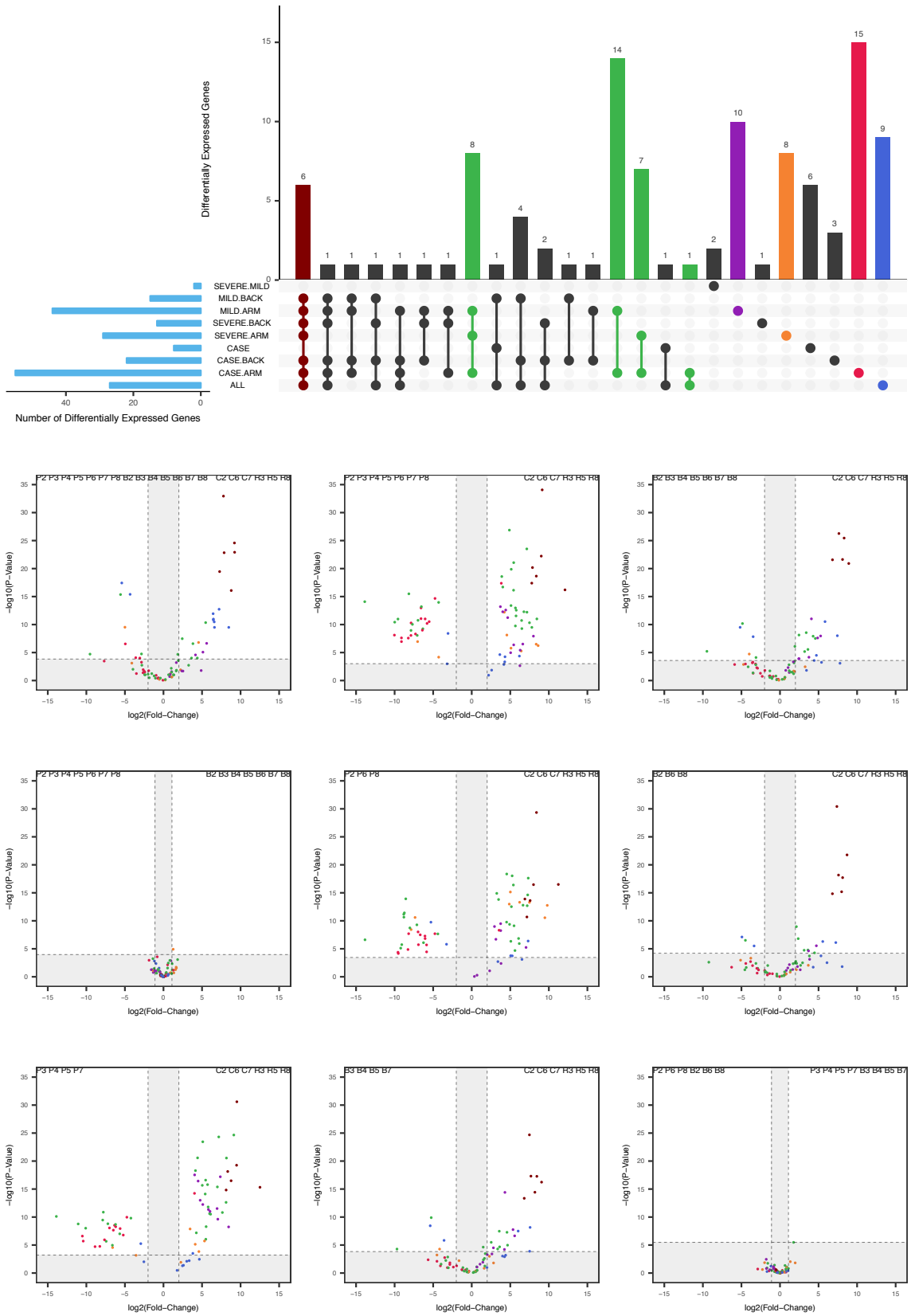


Figure 4.9: `UpSetR` plot for gene prioritisation. ***Top panel:*** `UpSet` plot displaying the intersection (number of shared genes) between the nine different comparisons. *Brown:* genes shared by all comparisons except CASE and SEVERE.MILD; *green:* genes shared by ALL, CASE.ARM, SEVERE.ARM and MILD.ARM comparisons only; *blue:* genes unique to ALL; *red:* genes unique to CASE.ARM; *orange:* genes unique to SEVERE.ARM; and *purple:* genes unique to MILD.ARM. ***bottom panel:*** volcano plots highlighting the selected genes in the context of each comparison.

their role in the disease.

4.3.7 Gene-set enrichment with `enrichR`

Gene-set enrichment was carried out on the significant genes (as a list) from each of the nine comparisons separately, as well as the 77 genes selected from the prioritisation step (Section 4.3.6). Even though a number of databases were used for the annotation of the significant genes, and annotation was done for the 10 separate (each of the nine comparisons and the prioritised list) lists of genes, not all results for the annotation will be presented; only the annotation for the prioritisation list will be presented as it represents the highly significant genes from the different comparisons.

The databases that were selected for annotation of the 77 prioritised genes were the GO databases ("GO Biological Process 2018", "GO Molecular Function 2018" and "GO Cellular Component 2018") and three pathway databases ("WikiPathways 2016", "KEGG 2016" and "Reactome 2016"). The adjusted *P*-values returned with the list of terms for the gene annotations were used to select on the significant annotations with adjusted *P*-value less than 0.05. Plots of the significant terms were created by plotting the number of genes annotated with the term on the y-axis, and the -$\log_{10}$(*P*-value) for the term (Figure 4.10). The top 5 terms for GO biological process (Figure 4.10) include "inflammatory response (GO:0006954)" (13 genes: *CXCL6*; *CCL20*; *CXCL1*; *PPBP*; *FOS*; *CXCL3*; *CXCL5*; *IL1A*; *IL1B*; *CXCR2*; *PROK2*; *S100A9*; *S100A8*), "neutrophil chemotaxis (GO:0030593)" (8 genes: *CXCL6*; *EDN3*; *CCL20*; *CXCR2*; *S100A12*; *CXCL3*; *S100A9*; *S100A8*), "granulocyte chemotaxis" (8 genes: *CXCL6*; *EDN3*; *CCL20*; *CXCR2*; *S100A12*; *CXCL3*; *S100A9*; *S100A8*), "neutrophil migration (GO:1990266)" (8 genes: *CXCL6*; *EDN3*; *CCL20*; *CXCR2*; *S100A12*; *CXCL3*; *S100A9*; *S100A8*), and "regulation of leukocyte chemotaxis (GO:0002688)" (6 genes: *CXCL6*; *EDN3*; *CXCL1*; *PPBP*; *CXCL3*; *CXCL5*).

The top 5 GO molecular function terms (Figure 4.10) include "cytokine activity (GO:0005125)" (11 genes: *GDF10*; *IL1A*; *CXCL6*; *CSF3*; *CCL20*; *IL1B*; *ADIPOQ*; *CXCL1*; *PPBP*; *CXCL3*; *CXCL5*), "CXCR chemokine receptor binding (GO:0045236)" (5 genes: *CXCL6*; *CXCL1*; *PPBP*; *CXCL3*; *CXCL5*), "chemokine activity (GO:0008009)" (6 genes: *CXCL6*; *CCL20*; *CXCL1*; *PPBP*; *CXCL3*; *CXCL5*), "chemokine receptor binding (GO:0042379)" (6 genes: *CXCL6*; *CCL20*; *CXCL1*; *PPBP*; *CXCL3*; *CXCL5*) and "RAGE receptor binding (GO:0050786)" (3 genes: *S100A12*; *S100A9*; *S100A8*). The GO cellular component terms (Figure 4.10) include "cytoplasmic vesicle lumen (GO:0060205)" (5 genes: *HBB*; *S100A12*; *HBA1*; *S100A9*; *S100A8*), "tertiary granule lumen (GO:1904724)" (3 genes: *HBB*; *CXCL1*; *PPBP*) and "endocytic vesicle lumen (GO:0071682)" (3 genes: *HBB*; *HBA1*).

For the pathways, the top 5 pathways identified from WikiPathways (Figure 4.10) in-

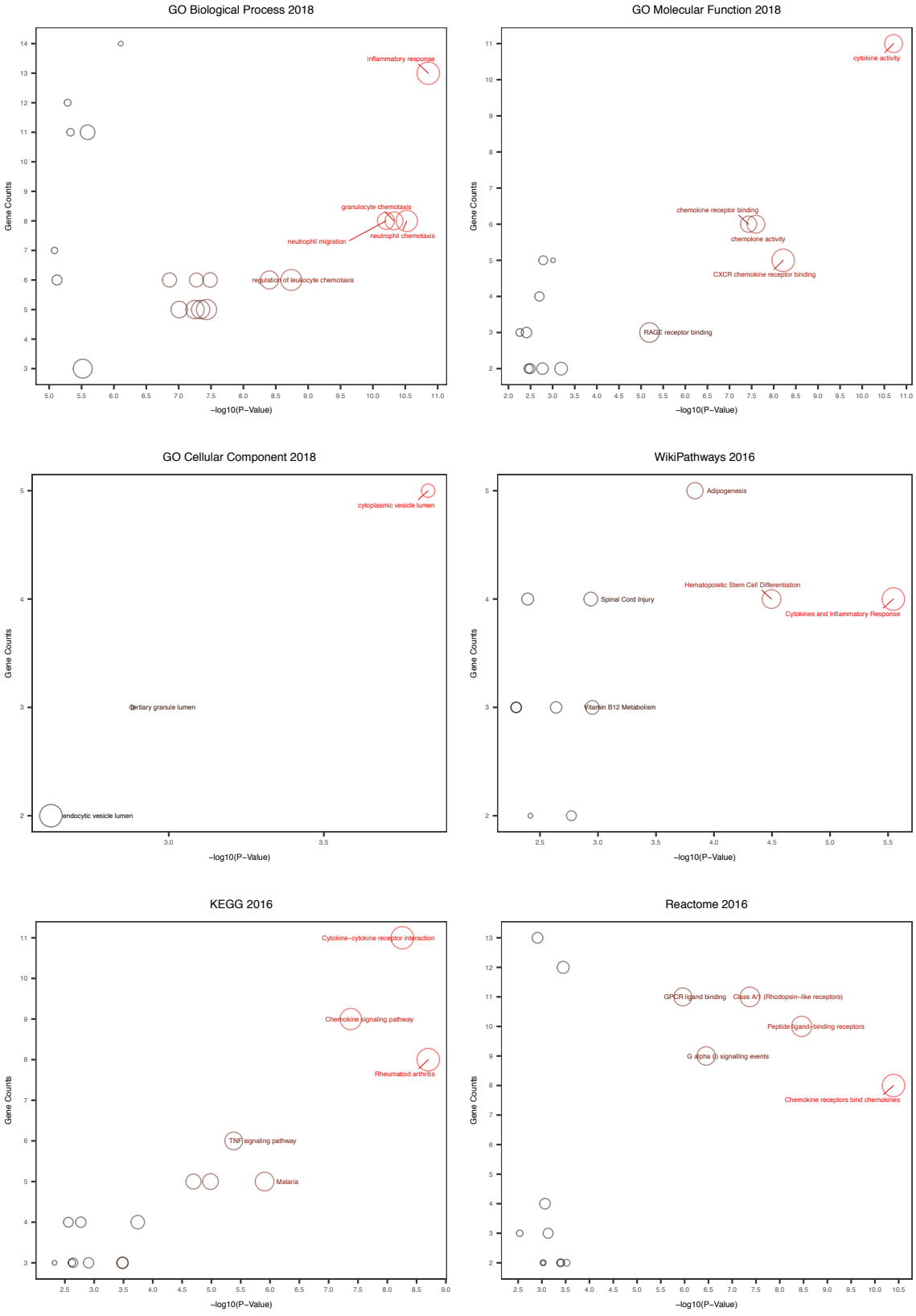


Figure 4.10: Significance plots summarising gene-set enrichment of the 77 prioritised genes. The enrichment was done using `enrichR` using the GO databases, KEGG, WikiPathways and Reactome. The plots show the number of genes on the y-axis and the -$\log_{10}$(P-value). Color of the terms in the in the plot represent significance, from black (less significant) to red (highly significant). The size of the circle represents the number of genes for each term. Only the top 5 terms names are shown.

clude "Cytokines and Inflammatory Response" (4 genes: *IL1A*; *CSF3*; *IL1B*; *CXCL1*), "Hematopoietic Stem Cell Differentiation" (4 genes: *IL1A*; *CSF3*; *IL1B*; *FOS*), "Adipoge-

nesis" (7 genes: *GDF10*; *LIPE*; *FRZB*; *LEP*; *ADIPOQ*; *PLIN1*; *KLF15*) ,"Vitamin B12 Metabolism" (3 genes: *IL1B*; *HBB*; *HBA1*) and "Spinal Cord Injury" (5 genes: *IL1A*; *IL1B*; *LEP*; *CXCL1*; *FOS*). WikiPathways also identified pathways from the mouse (*Mus musculus*), but these were ignored as the focus of this study is on human.

For KEGG annotation (Figure 4.10), "Cytokine-cytokine receptor interaction" (11 genes: *IL1A*; *CXCL6*; *CSF3*; *CXCR1*; *CCL20*; *IL1B*; *LEP*; *CXCR2*; *CXCL1*; *PPBP*; *CXCL3*; *CXCL5*), "Rheumatoid arthritis" (8 genes: *IL1A*; *CXCL6*; *CCL20*; *MMP1*; *IL1B*; *CXCL1*; *FOS*; *CXCL5*), "Chemokine signaling pathway" (9 genes: *CXCL6*; *CXCR1*; *CCL20*; *CXCR2*; *CXCL1*; *PPBP*; *CXCL3*; *ADCY8*; *CXCL5*), "TNF signaling pathway" (6 genes: *CCL20*; *IL1B*; *CXCL1*; *FOS*; *CXCL3*; *CXCL5*) and "Malaria" (5 genes: *COMP*; *CSF3*; *IL1B*; *HBB*; *HBA1*).

The top 6 pathway terms for Reactome (Figure 4.10) include "Chemokine receptors bind chemokines" (8 genes: *CXCL6*; *CXCR1*; *CCL20*; *CXCR2*; *CXCL1*; *PPBP*; *CXCL3*; *CXCL5*), "Peptide ligand-binding receptors" (10 genes: *CXCL6*; *GAL*; *CXCR1*; *EDN3*; *CCL20*; *CXCR2*; *PROK2*; *CXCL1*; *PPBP*; *CXCL3*; *CXCL5*), "Class A/1 (Rhodopsin-like receptors)" (11 genes: *CXCL6*; *GAL*; *CXCR1*; *EDN3*; *CCL20*; *CXCR2*; *PROK2*; *CXCL1*; *FFAR2*; *PPBP*; *CXCL3*; *CXCL5*), "G alpha (i) signalling events" (9 genes: *CXCL6*; *GAL*; *CXCR1*; *CCL20*; *CXCR2*; *CXCL1*; *PPBP*; *CXCL3*; *ADCY8*; *CXCL5*) and "GPCR ligand binding" (11 genes: *CXCL6*; *GAL*; *CXCR1*; *EDN3*; *CCL20*; *CXCR2*; *PROK2*; *CXCL1*; *FFAR2*; *PPBP*; *CXCL3*; *CXCL5*).

It is noteworthy that the GO terms and pathway annotations contain terms that are associated with the disease, i.e., inflammation, signaling, chemokine activity, cytokine activity and receptor binding, all of which are terms that are associated with SSc and have been used to describe the pathogenicity of the disease as discussed in Section 1.5 of Chapter 1.

4.3.8 Pathway analysis with `gage` and `pathview`

The differential expression data for each of the nine comparisons was used in `gage` in conjunction with `pathview` to construct pathways that are affected by the up- and down-regulated genes in the analysis. A q-value (adjusted P-value) of less than 0.05 was used to filter significant pathways that are either up- or down-regulated by the genes. In total, 39 pathways were identified (Table 4.6) using `gage` and `pathview`. These pathways could be found in one or more of the comparisons, and were either up-regulated (blue dots) or down-regulated (red dots) depending on the gene set used for constructing the pathways. Using `pathview`, a total of 117 pathways were constructed (Appendices J.1 - J.10).

Given the knowledge gathered on SSc (Chapter 1), a number of pathways identified here can be directly and/or indirectly linked to the disease, even though some are general pathways involved in normal functioning of cellular processes. These general pathways in-

Table 4.6: Summary of pathways identified by gage and constructed using `pathview`

ID	Pathway Name	ALL	CASE.ARM	CASE.BACK	CASE	SEVERE.ARM	SEVERE.BACK	MILD.ARM	MILD.BACK	SEVERE.MILD
hsa04145	Phagosome	●	●	●	●	●		●	●	●
hsa04380	Osteoclast differentiation	●	●	●		●	●	●		●
hsa04650	Natural killer cell mediated cytotoxicity	●	●	●		●		●	●	
hsa04621	NOD-like receptor signaling pathway	●	●	●		●		●	●	
hsa04110	Cell cycle	●	●	●		●		●	●	
hsa04062	Chemokine signaling pathway	●	●	●		●		●	●	
hsa03010	Ribosome	●	●	●		●		●	●	
hsa04141	Protein processing in endoplasmic reticulum	●	●	●				●	●	
hsa03040	Spliceosome	●	●	●				●	●	
hsa03013	RNA transport	●	●	●				●	●	
hsa04662	B cell receptor signaling pathway	●	●	●					●	
hsa04660	T cell receptor signaling pathway	●		●					●	
hsa04620	Toll-like receptor signaling pathway		●			●		●		
hsa04360	Axon guidance	●		●					●	
hsa04310	Wnt signaling pathway	●		●					●	
hsa04144	Endocytosis	●		●					●	
hsa04120	Ubiquitin mediated proteolysis	●		●					●	
hsa04010	MAPK signaling pathway	●		●					●	
hsa03050	Proteasome		●			●		●		
hsa03008	Ribosome biogenesis in eukaryotes	●	●					●		
hsa04972	Pancreatic secretion		●					●		
hsa04970	Salivary secretion		●					●		
hsa04810	Regulation of actin cytoskeleton				●				●	
hsa04640	Hematopoietic cell lineage		●						●	
hsa04623	Cytosolic DNA-sensing pathway		●					●		
hsa04210	Apoptosis	●	●							
hsa00240	Pyrimidine metabolism		●					●		
hsa04666	Fc gamma R-mediated phagocytosis								●	
hsa04610	Complement and coagulation cascades									●
hsa04520	Adherens junction								●	
hsa04514	Cell adhesion molecules (CAMs)									●
hsa04512	ECM-receptor interaction									●
hsa04510	Focal adhesion									●
hsa04350	TGF-beta signaling pathway				●					
hsa04115	p53 signaling pathway							●		
hsa04114	Oocyte meiosis	●								
hsa04012	ErbB signaling pathway	●								
hsa03018	RNA degradation	●								
hsa00190	Oxidative phosphorylation					●				

● Down-regulated pathways.
● Up-regulated pathways.

clude “Cell cycle”, “Ribosome”, “Protein processing in endoplasmic reticulum”, “Spliceosome”, “RNA transport”, “Axon guidance”, “Ubiquitin mediated proteolysis”, “Proteasome”, “Ribosome biogenesis in eukaryotes”, “Pancreatic secretion”, “Salivary secretion”, “Pyrimidine metabolism”, “Oocyte meiosis”, “RNA degradation” and “Oxidative phosphorylation”. However, even though these pathways are general, they may be affected by other pathways or pathway products associated with SSc, thus their investigation of association to the disease will not be excluded.

When looking at the list of pathways identified in Table 4.6, there seems to be a relationship between the identified pathways, source of tissue from which the samples were obtained (forearm and backs/breast) and the severity of the disease (mild and severe) in the samples in the comparison groups. For example, “Phagosome”, “Osteoclast differenti-

ation", "Natural killer cell mediated cytotoxicity", "NOD-like receptor signaling pathway" and "Chemokine signaling pathway" pathways seem to be affected in the majority of the comparisons except in the within-group comparisons, where they are either up-regulated (compared to being down-regulated in other comparison sets) or not present. All these pathways can be linked to the innate immunity that is dysregulated in SSc patients and is also persistent in the disease, i.e., there is evidence of dysregulated innate immunity in the early and late stages of SSc (as discussed in Chapter 1.5.2.1).

Phagosomes (Appendix K.1) are vesicles that form around the particles (mostly PAMPs and MAMPs) that have been engulfed by phagocytes (including macrophages and dendritic cells implicated in the dysregulated innate immunity in SSc) in the process of phagocytosis. As seen in Appendix K.1, the "Phagosome" pathway can be directly linked to three other pathways identified (Table 4.6), i.e., "Fc γ R-mediated phagocytosis", and the two phagocytosis promoting pathways, "Complement and coagulation cascades" and "Toll-like receptor signaling pathway" (Figure 4.11). The "Osteoclast differentiation" pathway is (Appendix K.2) responsible for production of osteoclast (large multi-nucleate cells responsible for bone matrix degradation) from mature macrophages/monocytes. This pathway can also be linked directly to the "Hematopoietic cell lineage" pathway (also identified in these analysis) that is responsible for the production of macrophages and monocytes from hematopoietic stem cells (HSC).

The "Natural killer cell mediated cytotoxicity" pathway (Appendix K.3) plays a crucial role in the innate immunity by determining the fate of damaged host cells as well as cells that are infected with microorganisms, like viruses. NK cells interact with targeted stressed, tumour or infected cells through antigen presentation and TLRs. This interaction induces a cascade of intracellular signaling pathways in the NK cell, including "MAPK signaling pathway" (identified in these analysis), which ultimately results in the movement and exocytosis of cytotoxic granules from the NK cell into the targeted cell, thus causing apoptosis. The "NOD-like receptor signaling" pathway (Appendix K.5) also plays a crucial role in the innate immune system. Nucleotide-binding and oligomerisation domain (NOD)-like receptors are intra-cellular PRR that are activated by "inflammasomes" in the cytoplasm cells of the innate immune system, and just like TLRs (Figure 4.11), they also recognise PAMPs and DAMPs (Pattanaik *et al.*, 2015). Activation of NOD-like receptors induces NF-κB and mitogen-activated protein kinase (MAPK) signaling pathways (also identified in these analyses; Appendix K.4), which are responsible for the production of pro-inflammatory cytokines.

The "Toll-like receptor signaling" pathway (Figure 4.11) associated with the recognition of PAMPs and MAMPs (discussed in Section 1.5.2) was also identified in the pathway analysis. The pathway, like the "NOD-like receptor signaling" plays a crucial role in the recognition of foreign agents and particles from damaged cells (epithelial cells in SSc)

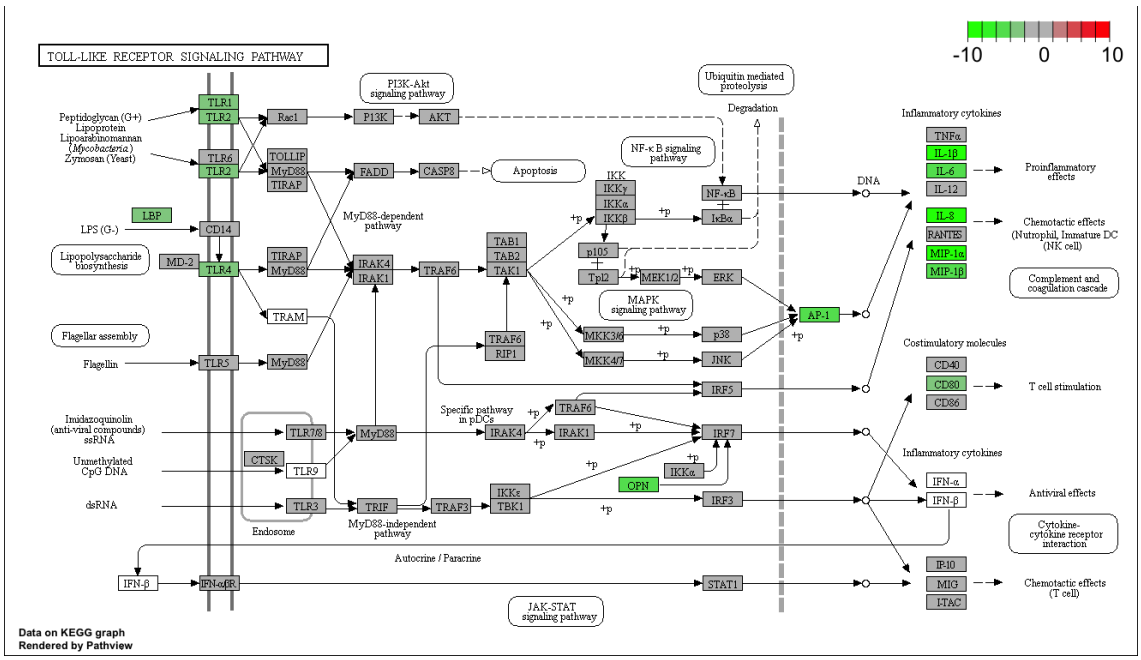


Figure 4.11: Toll-like receptor pathway (CASE.ARM) identified by gage and constructed with `pathview`. *Green:* down-regulated genes; *red:* up-regulated genes.

that are associated with the disease. This pathway was identified for all the comparisons with affected forearms (CASE.ARM, SEVERE.ARM and MILD.ARM). The presence of this pathway implicates a continuous activation of the TLRs, thus activation of the innate

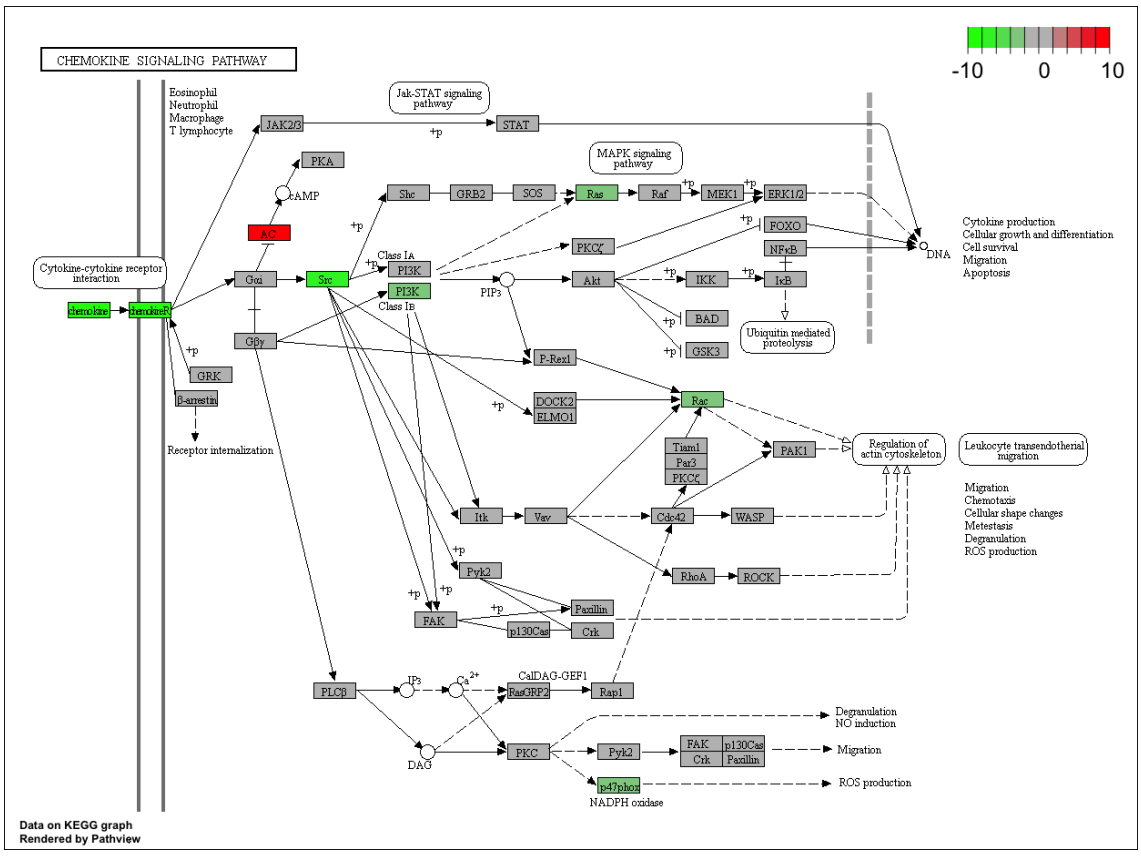


Figure 4.12: Chemokine signaling pathway (CASE.ARM) identified by gage and constructed with `pathview`. *Green:* down-regulated genes; *red:* up-regulated genes.

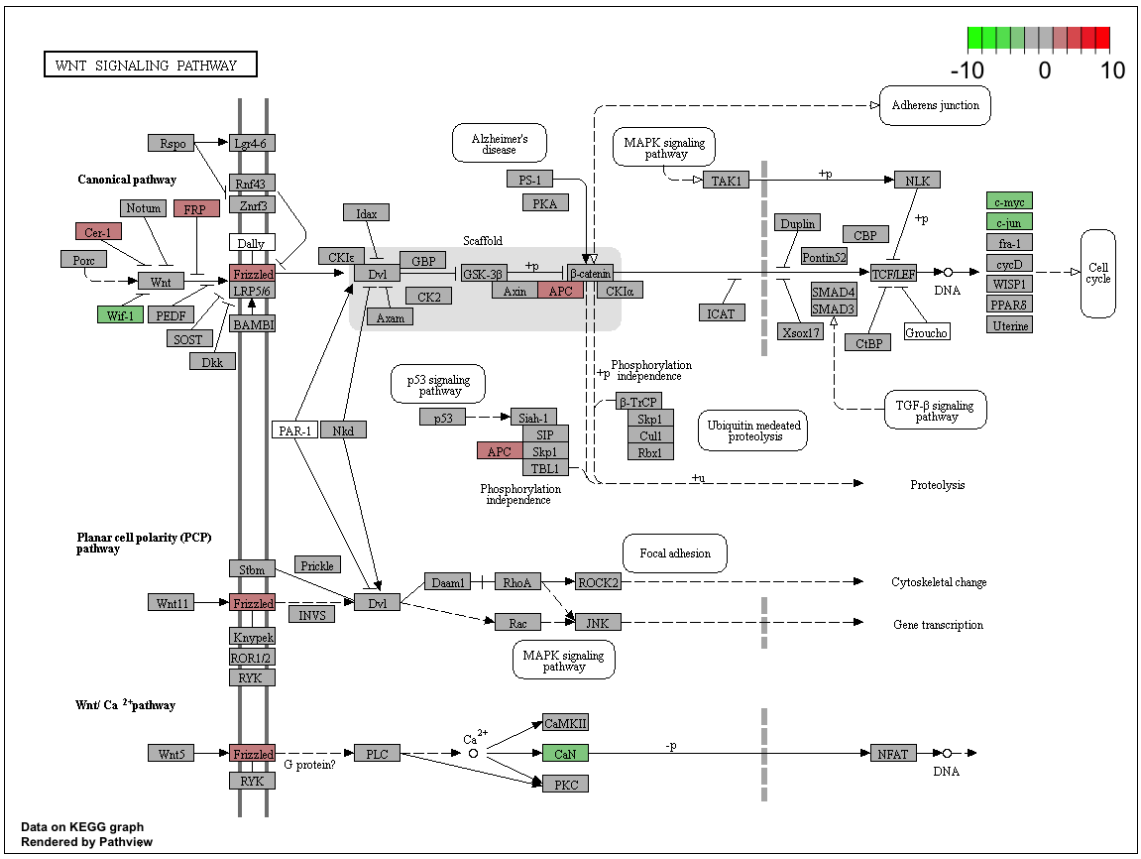


Figure 4.13: Wnt signaling pathway (ALL) identified by gage and constructed with `pathview`. *Green:* down-regulated genes; *red:* up-regulated genes.

immunity, through the course of the disease and the maintenance of the disease phenotype in both mild and severe forms of the disease.

The KEGG pathway annotations using `enrichR` (Figure 4.10) revealed "Chemokine signaling pathway" as one of the terms that was significant for the pathway annotation of the list of 77 prioritised genes. In the `gage` and `pathview` pathway analysis, this pathway (Figure 4.12) was also identified, which further implicates the role of chemokine production by macrophages and T-lymphocytes in the inflammatory responses observed in SSc. Dysregulation of the "Wnt signaling" pathway, also identified in these analyses (Table 4.6 and Figure 4.13) in the ALL, CASE.BACK and MILD.BACK comparisons, has been implicated in SSc, including the study by Frost (2016) for which the data used in this study was initially generated for the purpose of investigating the role played by genes involved in this pathway (Frost *et al.*, 2018).

4.4 Discussion

In this chapter, I addressed the second and third objectives the study, i.e., to perform differential expression and pathway analyses on the RNA-seq data from the SSc patients using the read counts produced from Chapter 3. Given the fact that there was no control over the RNA-seq data used in this study, since it was produced from a different study to answer specific questions relating to that study, steps were taken to address the limi-

tations (discussed in Section 1.11) of using the data for this study. The genes on the Y chromosome were excluded from this study as they represent sex biases in the samples (unbalance sex between the cases and controls). The 21 genes found on the Y chromosome were removed from the analysis (Appendix D).

The "sanitisation" method used in the initial steps of the expression analysis attempted to remove further sex and tissue site biases from the data by first performing differential expression between the samples from the controls (males against females) in order to exclude the genes that are expressed differentially between the two groups. With this method, 4 genes were identified and excluded from the data (Appendix D). In total, 25 genes were excluded from the analysis due to possible sex and tissue site bias they would bring in the analysis. The remaining data was then used after removing P1, B1, P9, B9 and C1 samples from the data.

To increase sensitivity, the data were partitioned into nine different comparisons (Table 4.1) based on the clinical features of the samples, and for each comparison, a different formula (Table 4.2) was used to take into account the condition, site and individual differences that exist between the case and control groups in the comparison sets. When performing PCA analysis (Figure 4.3 and Appendix E), there was a clear clustering of samples from affected individuals, and clustering of control individuals, with an exception of individual C7 from the controls group. However, because of limited sample size, this individual could not be removed from the analysis. Performing differential expression analysis on each of the nine comparisons using an $|\log_2 FC|$ of 2 (1.1 for CASE and SEVERE.ARM) and a FDR of 0.01 revealed interesting patterns of gene expression.

When looking at differential gene expression between the forearms of cases and controls (CASE.ARM, SEVERE.ARM, MILD.ARM), the clusters of down-regulated genes in these comparison sets are mostly composed of genes that have been associated with SSc in the Open Targets Platform database (Figures 4.6, 4.7 and 4.8). The merged genes found in these clusters are listed in Table 4.7. Even though this chapter of the study presents a wealth of information in terms of differentially expressed genes as well as pathways identified to be associated with SSc based on the expression data, the remainder of this discussion will focus on this list of genes (Table 4.7) for the following reasons: (1) the forearm comparisons (CASE.ARM, SEVERE.ARM and MILD.ARM) represent comparisons of the affected areas in the SSc individuals with the controls; (2) some of the genes in the list have been associated with annotation terms related to clinical features of SSc (Section 4.3.7); (3) the genes (most) have been found to be associated with SSc in the Open Targets Platform; (4) These genes are not found to be significantly differentially expressed in any other comparison group (with the exception of *FOS* found in ALL comparison), nor do they show similar expression patterns as seen in CASE.ARM, SEVERE.ARM and MILD.ARM; and (5) the majority (if not all) of pathways identified by

Table 4.7: List of 27 down-regulated genes associated with SSc in the Open Targets Platform. The 27 genes (merged) are down-regulated in all the affected forearm comparisons (CASE.ARM, SEVERE.ARM and MILD.ARM).

Gene ID	Chrom	Gene name
AQP9	15	Aquaporin 9
BCL2A1	15	BCL2 related protein A1
CCL20	2	C-C motif chemokine ligand 20
CD69	12	CD69 molecule
CMTM2	16	CKLF like MARVEL trans-membrane domain containing 2
CSF3	17	Colony stimulating factor 3
CXCL1	4	CXC motif chemokine ligand 1
CXCL3	4	CXC motif chemokine ligand 3
CXCL5	4	CXC motif chemokine ligand 5
CXCL6	4	CXC motif chemokine ligand 6
CXCR1	2	CXC motif chemokine receptor 1
CXCR2	2	CXC motif chemokine receptor 2
FCGR3B	1	Fc fragment of IgG receptor IIIb
FFAR2	19	Free fatty acid receptor 2
FOS	14	Fos protooncogene, AP1 transcription factor subunit
HBA1	16	Hemoglobin subunit alpha 1
IL1A	2	Interleukin 1 alpha
IL1B	2	Interleukin 1 beta
LUCAT1	5	Lung cancer associated transcript 1
MMP1	11	Matrix metallopeptidase 1
PI3	20	Peptidase inhibitor 3
PPBP	4	Proplatelet basic protein
PROK2	3	Prokineticin 2
S100A12	1	S100 calcium binding protein A12
S100A8	1	S100 calcium binding protein A8
S100A9	1	S100 calcium binding protein A9
TREM1	6	Triggering receptor expressed on myeloid cells 1

`gage` and `pathview` (Table 4.6) using the differential expression data are down-regulated for CASE.ARM, SEVERE.ARM and MILD.ARM comparison sets. The genes that are down-regulated in SSc patients and their association with SSc are discussed in Section 4.4.1 below.

4.4.1 Down-regulated genes in black South African patients with SSc

The *AQP9* gene encodes a aquaglyceroporins protein, aquaporin 9, which is expressed in the skin fibroblasts and is responsible for water and glycerol transmission in the processes of hydration and tissue regeneration. The study by Yousefi *et al.* (2017) examined the expression of *AQP1*, *AQP3*, and *AQP9* mRNA in dermal fibroblasts of SSc patients (n=20) and healthy fibroblasts (n=20). Their findings revealed that the `AQP3` gene was down-regulated in the in SSc fibroblasts whilst *AQP1* and *AQP9* had similar expression levels in fibroblasts from both healthy and diseased individuals. They concluded that the dryness, dysregulated tissue healing and fibrotic lesions may be linked to the down-regulation of *AQP3*. In this study, the *AQP9* gene is also down-regulated, thus could also be linked to the findings by (Yousefi *et al.*, 2017). However, no pathway was associated with the down-regulation of the *AQP9* gene in this study.

The BCL2A1 protein is a member of the B-cell lymphoma 2 (BCL2) proteins, whose main functions are to regulate cell death (Vogler, 2012). Regulation of the *BCL2A1* gene is through NF-κB signaling, in which increased levels of its expression results in pro-

survival function to prevent cell death. In the pathway analysis carried out in this study, *BCL2A1* (A1) was found to be down-regulated in the "Apoptosis" pathway through NF-κB signaling (Appendix K.6). Down-regulation of this gene implies that its pro-survival functions are lost, and cannot prevent cell death. The study by Noble *et al.* (1999) investigated the underlying mechanism that prevented endothelial cell death by the blood monocytes. In their study, they identified that blood monocytes interact with endothelial cells during an inflammatory response and provide survival signals to the endothelial cells by promoting up-regulation of *BCL2A1* to protect it from cell death. Thus, down-regulation of the *BCL2A1* observed in this study could be linked to the release of DAMPs from epithelial cells (down-regulation of *BCL2A1* cannot protect epithelial cells from cell death) that initiate (or maintain) autoimmunity observed throughout the course of SSc.

The *CD69* gene encodes a CD69 trans-membrane C-Type lectin protein, which is an activation marker expressed in regulatory T-lymphocytes (Tregs), HSCs and other cells of the immune system (Asano, 2018). Tregs are the main players in the maintenance of immune self-tolerance and prevent autoimmunity; and their presence in healthy individuals does not confer development of autoimmune disease. In the early stages of SSc, the activated Tregs, bearing the CD69 receptors which regulate T-lymphocyte interaction with other cells, affect the activation of fibroblasts in SSc. The studies by Radstake *et al.* (2009) and Kalogerou (2005) revealed that, even though *CD69* is highly expressed in the early stages of the disease, over time, the expression of *CD69* on the surface of Tregs is decreased with the disease duration. They also showed that the diminished functioning of Tregs correlates with down-regulated expression of *CD69* and *TGF-β*. In the study presented in this thesis, the disease duration of the patients with SSc is, on average, 34.3 months; and this could perhaps explain the down-regulation of the *CD69* gene.

Chemokines are small chemoattractant cytokines belonging to three families, i.e., CC, CXC, or CX_3C, based on the number of amino acids between the first two cysteine that form disulfide bonds in their structure. Chemokines (and their respective receptors) are very similar in structure. However, their specific receptor binding, regulation and expression allows them to play roles in a wide variety of cellular processes including leukocyte trafficking and hematopoiesis (Gartzke and Lange, 2002). A number of chemokine genes belonging to the CC and CXC families were identified to be down-regulated in this study. These include *CXCL1* (also known as growth-regulated oncogene α [*GROα*]), *CXCL3* (also known as *GROγ*), *CXCL5* (also known as epithelial-derived neutrophil-activating peptide 78 [*ENA-78*]) and *CXCL6* (also known as granulocyte chemotactic protein 2 [*GCP-2*]) (Table 4.7). Chemokine receptors for these chemokines were identified to be down-regulated, i.e., *CXCR1* (also known as IL-8Rα) recognises CXCL6; and *CXCR2* (IL-8Rβ) recognises CXCL1, CXCL3, CXCL3 and CXCL6 (Gartzke and Lange, 2002).

The CXC family of chemokines has been implicated in angiogenic activity in SSc (dis-

cussed in Section 1.5.1.3), depending on the presence or absence of the ELR motif (Glu-Leu-Arg) (Tsou *et al.*, 2016). These three amino acids are highly conserved and appear to be responsible for ligand/receptor interactions. For example, PF4 and IL-8 both belong to the CXC family of chemokines; however, due to the presence of the ELR motif in IL-8 and absence in PF4, IL-8 has pro-angiogenic properties, whilst PF4 has anti-angiogenic effects (Marriott *et al.*, 2002). The CXC chemokines down-regulated in this study could implicate possible mechanism of dysregulated/impaired angiogenesis in SSc. In the study by Brabcová and L. Kolesár, E. Thorburn (2014), they noticed that epithelial cells produced high levels of CXC chemokines in response to IL-1β. The down-regulation of *IL-1β* also seen in this study could be the cause of the down-regulated CXC chemokines identified.

The CC chemokine, CCL20, is a selective chemoattractant that functions in the recruitment of immature dendritic cells, naive B-lymphocytes as well as effector/memory T-lymphocytes to the sites of inflammation. Its role in SSc is not completely understood; however, it has been found to be in diseases such as RA and Crohn's disease (Di Sabatino *et al.*, 2016). Tao *et al.* (2011) studied the expression and distribution of *CCL20* together with its receptor, CCR6, in patients with early SSc skin lesions. They observed that *CCL20/CCR6* levels were highly up-regulated in the lesional skin of SSc patients as compared to normal skin tissue, even thought there were some variations amongst the patients. They also found that two cytokines, TNF-α and IL-1β, were responsible for significantly increasing/regulating the expression of *CCL20*. In this study, both *IL-1β* (*IL1B*, Table 4.7) and *CCL20* were found to be down-regulated in SSc patients. However, down-regulation of both *CCL20* and *IL-1β* might be dependent on the stage of the disease as these genes are down-regulated in patients with the mild form of SSc (MILD.ARM, Figure 4.8) and not in patients with the severe form of SSc (SEVERE.ARM, Figure 4.7).

The findings presented here on the down-regulation of genes implicated in SSc (from the Open Targets Platform) seem to correlate with findings from other studies; and the disease duration seems to play a major role in the down-regulation of these genes. However, as SSc is a very complex disease and still poorly understood, the underlying cause of the down-regulation of these genes could not be identified in the differential expression results presented here. Other factors that may have affected the results are the unmatched samples and small sample sizes of the data used in this study. In future, it would be beneficial to carry out a study with a larger sample size and use skin tissues from the same body sites in order to carry out the differential expression and pathway analysis.

Chapter 5

Developing a Pipeline for Metagenomics Analysis

In Section 1.6.2, I discussed pathogens and infectious agents as possible triggers of SSc based on the interferon (IFN) signature that can be linked with the innate immune system. This study also set out to investigate possible pathogens that may be associated with SSc in black South Africans using metagenomic analyses of the RNA-seq data. This chapter addresses the fourth and final objective of this study: ***to develop a novel pipeline for metagenomic analyses and use it to identify potential microorganisms/pathogens associated with SSc.*** The `Nextflow` workflow for metagenomic analyses presented in this chapter is portable, reproducible and scalable, and produces interactive HTML charts that can be used to identify unique/shared taxonomy between affected individuals and controls using raw RNA-seq data.

5.1 Introduction

The term "microbiome" refers to the genetic material of all microorganisms (including bacteria, viruses, archaea, protozoans and fungi) present within the human body or environment. The term microbiome is often confused with the term "microbiota", which is the microbial population present at particular sites of the human body or habitat (Bakhtiar *et al.*, 2013; Huttenhower *et al.*, 2012; Martín *et al.*, 2014; Methé *et al.*, 2012; Rawat *et al.*, 2014). Human microbiome studies are amongst the fastest growing areas of research. This is because the human microbiome forms a critical part of the human physiology and imbalances in the microbiota have been reported to be associated with many diseases (Huttenhower *et al.*, 2012). Microbial profiling of clinical samples is typically carried out by sequencing variable regions of the 16S ribosomal RNA (rRNA) gene. However, the use of a single marker gene has its limitations as species identification depends on the extent of diversification of the variable 16S regions. These limitations include difficulties in assessing operational taxonomy units (OTUs), assessing diversity and limited resolution of the 16S rRNA gene among closely related species (Poretsky *et al.*, 2014; Zhang *et al.*, 2015).

As discussed in Section 1.7.4, RNA-seq provides a cost effective way of obtaining large amounts of transcriptome data, providing us with a window of opportunity to study and understand host-associated microbial communities. Metagenomics is a rapidly growing field that allows for the study of microbial genomes within diverse environmental sam-

ples. The key research area in metagenomics is the identification of microbial sequences within a host genomic background, which may represent potential pathogens associated with a disease (Dimon *et al.*, 2013; Rawat *et al.*, 2014). A number of sequence-driven metagenomics tools have been developed which make use of sequencing data. These tools include MetaGeniE (Rawat *et al.*, 2014), Pipeline for Analysis of RNA-Seq Exogenous Sequences (PARSES), PathSeq, MGAviewer, MetaSAMS and Integrated Metagenomics Sequence Analysis (IMSA; Dimon *et al.*, 2013).

The principle behind these metagenomics tools is that given a set of reads, the host reads can be "subtracted" from the read set through alignment to a reference genome, thereby leaving behind exogenous microbial reads. These can be further characterised by alignment to a comprehensive, non-redundant and well annotated sequence database like RefSeq. The IMSA pipeline has already been applied in an SSc study by Arron *et al.* (2014). In their study, RNA-seq data obtained from the forearm skin biopsies of patients with early dcSSc and healthy individuals were analysed to quantify the nonhuman reads in each sample. The metagenomic analyses revealed that there was an overrepresentation of the fungus *Rhodotorula glutinis* in the patient samples as compared to healthy individuals, suggesting that it might play a role in triggering the inflammatory response found in SSc.

5.1.1 Host-read filtering

The first challenge in metagenomic analyses is identifying reads in the RNA-seq data that do not belong to the host (organism being studied). This is a crucial step in the analysis as failure to remove host reads can lead to host reads being used in the downstream analyses, including taxonomic classification, causing noise in identifying relevant taxonomies in the analysis (Visconti *et al.*, 2018). The IMSA pipeline described by Dimon *et al.* (2013) uses an "action file" that guides the filtering of host reads from the data. In this action file, users can specify `Bowtie`, `blat` or `blastn` for filtering host reads, along with filtering parameters at each filtering step, in any order and as many times as desired until the host reads have been filtered.

The issue with the method, however, is that it uses the well-known `BLAST` program to align reads to the reference genome (Acland *et al.*, 2014). Although `BLAST` is a widely used tool for sequence alignment, performing alignment on millions of short sequencing reads can take a significant amount of time; thus such a method can be very slow when performing metagenomics to filter out host reads (Wood and Salzberg, 2014). Tools like `STAR`, already implemented in this study for aligning short reads to the reference genome, are possible alternatives to this alignment issue. As seen in the mapping results in Section 3.3.2 (Figure 3.8), `STAR` was able to uniquely align more than 95% of the reads to the reference genome, whilst ∼4% were mapped to multiple regions, leaving behind less than ∼1% of the reads as unmapped. This small fraction of unmapped reads represents reads belonging to possible microorganisms and could be taxonomically classified to identify

their origins.

5.1.2 Classifying exogenous reads

In metagenomic analysis, taxonomic classification of sequencing reads is accomplished through querying a comprehensive database of "non-redundant" sequences, like RefSeq (`https://www.ncbi.nlm.nih.gov/refseq/`) and `BLAST`'s `nt` database (non-redundant nucleotide sequences database), which also requires the use of the `BLAST` program (Dimon *et al.*, 2013). The alignment issue in the initial filtering step is seen at taxonomic classification as well. Although the alignment of the millions of short sequencing reads for taxonomic classification can be overcome by performing *de novo* assembly with tools such as `trinity` (discussed in Section 3.1.1.3) to produce longer contigs, the classification databases are still relatively large for performing alignment with `BLAST`. To overcome these alignment issues, tools like `kraken2` were developed for taxonomic classification of sequencing reads (Wood and Salzberg, 2014).

`kraken2` is a sequence classification tool that functions by creating a database smaller than the comprehensive RefSeq or `nt`, which allows for much faster classification of reads as compared to searching a full collection of all genomes. What makes `kraken2` successful at classifying reads according to taxonomy is the exact alignment of k-mers implemented in its algorithm. This algorithm, together with the pre-computed database makes the exact matching of k-mer extremely efficient (Wood and Salzberg, 2014). The `kraken2`'s database is created by first selecting a set of microbial sequence libraries. `Jellyfish`'s multi-threaded k-mer counter is then used to create a database of unique 31-*mers* in the microbial sequence libraries, and instead of storing counts of the 31-*mers*, taxonomic IDs of the k-mers lowest common ancestor (LCA) values are stored. Each sequence in the library is then processed, one at a time, and its taxon number is used to set the stored LCA values of all k-mers in the sequence.

To classify sequences using the database created, the `kraken2`'s classification algorithm collects all k-mers for that sequence and queries each k-mer against the database. For all matching k-mers in the database, the algorithm uses the resulting set of LCA taxa to determine the most high scoring LCA and assign it to the sequence being classified (Wood and Salzberg, 2014). All other sequences whose k-mers do not have hits in the database are left unclassified. Another feature of `kraken2` that makes it efficient in classifying sequences with high accuracy and maximal speed is that it is able to hold its entire database into the computer's memory (database $\sim$70GB). This greatly reduces paging faults, whereby the computation speeds are reduced due to data being taken form storage to physical memory for processing.

5.2 Analyses

After realising the pitfalls of implementing the analyses manually on the cluster based on the initial steps in designing the `rnaSeqCount` pipeline (Chapter 3), the metagenomics analyses described here were implemented directly on the Wits Computing cluster using `Nextflow`. The following sections describe the implementation of the metagenomics pipeline in `Nextflow`.

5.2.1 Pipeline for metagenomic analysis of RNA-seq data

The pipeline for metagenomic analysis described here depends on two major software, `Nextflow` and `Singularity`. These are required to execute the workflow on different computational platforms. The workflow takes FASTQ files as input and also relies on the standard `kraken2` database (human, viral, archaeal and bacterial sequences) for taxonomic classification. Generating this database is computationally intensive and takes a long time. The following commands can be used to create a standard `kraken2` database using 200GB of memory and 7 CPUs, which take ∼20 hours to create. On completion, the size of the database is ∼35GB, and contains `hash.k2d` (the minimizer to taxon mappings), `opts.k2d` (information about the options used to build the database) and `taxo.k2d` (taxonomy information used to build the database) files required for classification.

```
## Build database
kraken2-build --standard --threads 6 --db kraken2_std

## Cleanup unnecessary files
kraken2-build --cleanup --db kraken2_std
```

5.2.1.1 Implementing the workflow in `Nextflow`

1. ***Collect read pairs, reference genome and `kraken2` database:*** The first step of the pipeline is to collect input and store it into channels. The location of the RNA-seq data, where the read pairs are stored, is used to create an input channel for the paired-end reads in FASTQ format. The reference genome as well as the indexes are also taken as input for the pipeline, as well as the location for the `kraken2` database for taxonomic classification and stored into channels used by the pipeline processes.
2. `runSTAR`: The first process, `runSTAR`, aligns the FASTQ files to the reference genome and removes the host (human) reads from the analysis. The reads that are not mapped to the reference genome are then collected as unmapped FASTQ files into two channels, one for assembling using `trinity` and the other for direct classification using `kraken2`. Assembling unmapped reads using `trinity` was to compare whether taxonomic classification of reads after assembling them into longer sequences is superior/beneficial than classifying raw short unmapped reads.
   (a) `runMultiQC`: The mapped reads (BAM files) reports are passed onto this process for QC using `MultiQC` in order to examine how many of the host reads

aligned to the reference genome, thus how many of the reads were excluded from the metagenomic analysis.

3. `runKrakenClassifyReads`: One of the channels containing unmapped reads from the `runSTAR` process is used as an input for this process to directly classify the reads using `kraken2`. A channel is created with the output, which is then used by the `runKronareport` process to create `krona` charts to visualise the classification.
4. `runTrinityAssemble`: This process is for assembling unmapped exogenous reads. One of the channels containing the unmapped reads in FASTQ format from the `runSTAR` process is used to assemble reads into longer contigs/sequences using `trinity`. This step was to identify whether the short unmapped reads belong to a larger sequence, which may arise from an organism, since it could be possible that sequences from microorganisms residing in the host skin tissue could be sequenced along with the host reads. The resulting FASTA file with a list of assembled sequences was passed on to the `runKrakenClassifyFasta` process for taxonomic classification using `kraken2`.
5. `runKrakenClassifyFasta`: Once the reads have been assembled into FASTA sequences using `trinity`, taxonomic classification is carried out on them using `kraken2`. The results of the classfication are stored in a channel and used in the `runKronareport` process to create `krona` charts for visualising the taxonomic classifications.
6. `runKronareport`: This process creates interactive `krona` charts for visualising the taxonomic classification from the `runKrakenClassifyFasta` and `runKrakenClassifyReads` processes. The results are HTML files which are stored in the output directory for each sample processed. The results are also used by the `runPrepareMatrixData` process to create a matrix to be used by `UpSet` in order to determine the taxonomies that are unique and shared between the samples from the cases and the controls.
7. `runPrepareMatrixData` and `runCreateMatrix`: Output from the `runKronareport` is used to generate a matrix to be used by the `UpSet` program for interactive exploration of taxonomic species that are shared between the samples. In the `runPrepareMatrixData` process, the list of all the unique taxonomy IDs are obtained for each sample from the `krona` reports generated. The `runCreateMatrix` then uses these lists of taxonomies from all samples to create a combination matrix, where samples are columns and taxonomy IDs are rows. Each cell in the matrix has either a "1" or "0", depending on whether a taxonomy ID is found in the sample or not, respectively. An extra column is added at the end with the taxonomy name corresponding to the taxonomy IDs. This matrix is provided with the `UpSet` application that is created in the results from executing this pipeline. The `UpSet` application opens up in a browser and the user is able to create custom queries in order to select which samples they are interested in, and whether they are interested

in unique or shared taxonomies between the samples.

5.2.1.2 Singularity images for applications

The reproducibility and portability of this pipeline is facilitated by the `Singularity` images created for all applications used by the workflow. Appendix M lists all the `Singularity` recipes that are needed to run the workflow for metagenomic analysis on RNA-seq data. The `runStar` and `runMultiQC` processes use the same `Singularity` image for `STAR` (Appendix C.1) for `MultiQC` (Appendix C.4) as the `rnaSeqCount` workflow, respectively. The `runKrakenClassifyReads`, `runKrakenClassifyFasta` and `runKronareport` all depend on the `kraken2` image (Appendix M.2). The `runTrinityAssemble` depends on the `trinity` image (Appendix M.1). The `runPrepareMatrixData` and `runCreateMatrix` both depend on the `UpSet` image (Appendix M.3).

5.2.1.3 GitHub repository for the pipeline

To keep track, share and document the workflow described here, a GitHub repository was created (https://github.com/phelelani/nf-rnaSeqMetagen). The `Singularity` recipes are included in the GitHub repository for the workflow and are linked to the SingularityHub (https://www.singularity-hub.org/collections/728) where they are hosted, so that any changes that are made on the recipes are updated on the SingularityHub server.

5.3 Results

The workflow for metagenomic analyses of RNA-seq data has been successfully developed in `Nextflow` and `Singularity`, and can be executed on any UNIX-based OS with `Nextflow` and `Singularity` installed. The workflow is available on GitHub (https://github.com/phelelani/nf-rnaSeqMetagen) and the `Singularity` images it depends on for execution can be found on the SingularityHub server (https://www.singularity-hub.org/collections/728). This section discusses the details of the metagenomic analysis pipeline developed, including obtaining, executing and interpreting the results produced by the workflow.

5.3.1 `rnaSeqMetagen`: A portable and reproducible `Nextflow` pipeline for metagenomic analysis of RNA-seq data

The pipeline developed here, "`rnaSeqMetagen`" (using RNA-seq data for "metagenomics" analysis), is available on GitHub and can be used by other researchers to perform metagenomic analysis using RNA-seq data. Figure 5.1 summarises the overall workflow and how the processes in the workflow handle input and output data in order to produce the results.

5.3.1.1 Obtaining the pipeline

Just like the `rnaSeqCount`, the `rnaSeqMetagen` workflow depends on `Nextflow` and `Singularity` to be executed. The commands below can be used to obtain the pipeline.

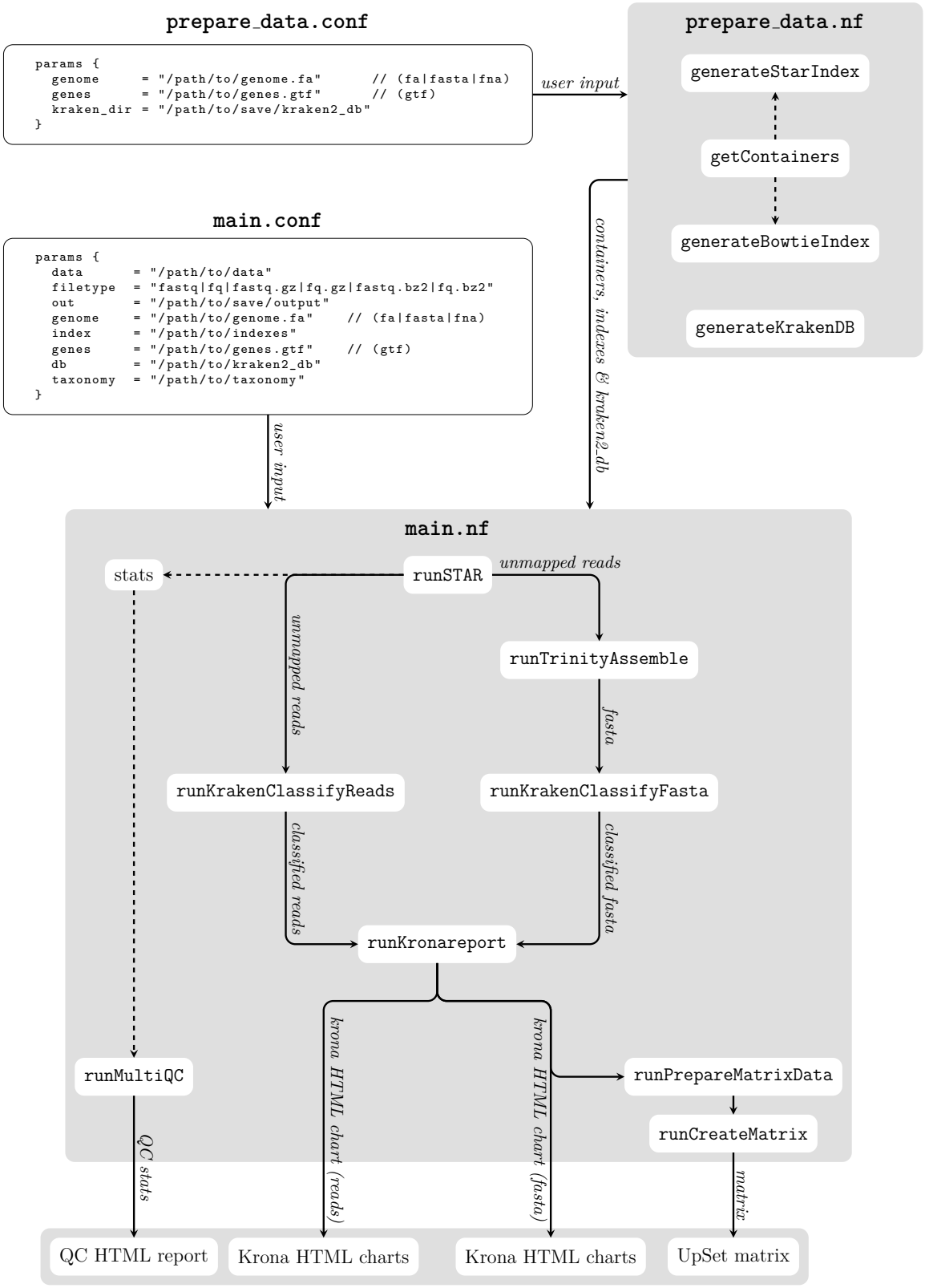


Figure 5.1: Overall summary of the `rnaSeqMetagen` workflow. The `rnaSeqMetagen` pipeline works in two stages: (1) **Genome indexing and `kraken2` database creation:** The `prepare data.config` needs to be provided with the location of the reference genome as well as the location where the pre-computed `kraken2` database will be stored. The `prepare data.nf` can then be executed in order to download singularity images, perform genome indexing and create the `kraken2` database. (2) **Pipeline execution:** Once the reference genome has been indexed, the location of the reference genome (and indexes), FASTQ files (and file type), output directory and `kraken2` database location can be provided in the `main.config` and the `main.nf` executed to perform the analysis on the RNA-seq data.

```
## Using git command
git clone https://github.com/phelelani/nf-rnaSeqMetagen.git

## Using nextflow command
nextflow pull phelelani/nf-rnaSeqMetagen
nextflow pull https://github.com/phelelani/nf-rnaSeqMetagen.git
nextflow clone phelelani/nf-rnaSeqMetagen <target-dir>
```

The contents of the downloaded pipeline are shown below. The user needs only to edit the `main.config` and the `prepare_data.config` with the necessary input in order to execute the pipeline. The `nextflow.config` file does not need to be edited, especially by inexperienced users as it contains crucial instructions for pipeline execution as well as different profiles for running the pipeline on different computational platforms.

```
nf-rnaSeqCount
|-containers            # (folder) Singularity recipes and location for images
|  |--Singularity.kraken2
|  |--Singularity.multiQC
|  |--Singularity.star
|  |--Singularity.trinity
|  |--Singularity.upset
|--templates            # (folder) Location of extra scripts for performing analysis
|  |--create_matrix.R
|  |--get_taxons.sh
|--README.md            # Pipeline documentation
|--main.config          # Configuration file for the main Nextflow script (user input)
|--main.nf              # Main Nextflow script for running the pipeline
|--nextflow.config      # Pipeline configuration file
|--nf-rnaSeqMetagen.png# Summary of the pipeline
|--prepare_data.config # Configuration file for preparing genome indexes (user input)
|--prepare_data.nf      # Main Nextflow script for preparing genome indexes
```

5.3.1.2 Obtaining Singularity images and generating the Kraken2 database

Once the `rnaSeqMetagen` pipeline has been obtained, the first step is to download the `Singularity` images from SingularityHub, prepare the reference genome (create index) for host read filtering and download and create the `kraken2` standard database for taxonomic classification. This first step can simply be accomplished by editing the `prepare_data.config` with the location of the reference genome FASTA file and the desired output location for the `kraken2` standard database in order to execute the `prepare_data.nf` script which will perform the initial steps.

The `prepare_data.nf` can then be executed with the one of the options for `--mode`, i.e., `getContainers` to download the `Singularity` images; `generateStarIndex` to generate `STAR` reference genome indexes; `generateBowtieIndex` to generate `Bowtie` indexes for the reference genome; and `generateKrakenDB` to download and create the `kraken2` standard database. In addition to the `--mode` option for executing the `prepare_data.nf`, the `-profile` option can also be passed with either `prepare` (for local execution with no job scheduler), `slurmPrepare` (for SLURM scheduler) or `pbsPrepare` (for PBS scheduler)

options, depending on the scheduler used by the computational platform. On an HPC cluster with a PBS scheduler, the data preparation can proceed as follows:

```
## Download Singularity images
nextflow run prepareData.nf --mode getContainers -profile pbsPrepare

## Generate STAR indexes
nextflow run prepareData.nf --mode generateStarIndex -profile pbsPrepare

## Generate Bowtie2 indexes
nextflow run prepareData.nf --mode generateBowtieIndex -profile pbsPrepare

## Generate Bowtie2 indexes
nextflow run prepareData.nf --mode generateKrakenDB -profile pbsPrepare
```

5.3.1.3 Executing the `rnaSeqMetagen` workflow

To execute the `rnaSeqMetagen` workflow, the main script for metagenomic analysis (`main.nf`), the `main.config` script must be provided with the following: the RNA-seq data (FASTQ files) location, the FASTQ file type (one of `fastq`, `fq`, `fastq.gz`, `fq.gz`, `fastq.bz2` or `fq.bz2`), the output folder, reference genome (`.fa`, `.fasta` or `.fna` file extensions), reference genome indexes and the location of the `kraken2` database created (`prepare.nf`). The `-profile` option can be passed when executing the `main.nt` script with either `standard` (for local execution with no scheduler), `pbs` (for PBS schedulers) or `slurm` (for SLURM job schedulers) depending on the job scheduler used on the computational platform. An example for executing the `rnaSeqMetagen` pipeline on an HPC with a SLURM scheduler is as follows:

```
## Execute rnaSeqMetagen workflow using a SLURM scheduler
nextflow run main.nf -profile slurm
```

5.3.1.4 Results produced by the `rnaSeqMetagen` pipeline

After executing the `rnaSeqMetagen` workflow, a number of folders can be found inside the output directory specified by the user:

- *n* number of folders, each corresponding to the sample name. Inside these folders are the `krona` HTML report for both reads and FASTA sequences for taxonomy classification; classified and unclassified FASTA sequences (from `runKrakenClassifyFasta`) and FASTQ files (from `runKrakenClassifyReads`); `.kron` files from `krona`; and mapping statistics produced by `STAR`.
- ***report_QC*** folder containing `MultiQCQC` reports in HTML format. This file can be used to assess the quality of read mapping to the reference genome to find out how many of the host reads were excluded from the analysis.
- ***report_workflow*** folder containing pipeline execution reports. These files can be used to trace the execution of the pipeline and check other metadata in order to assign resources correctly to the processes.

- ***upset*** folder containing the `UpSet python` application and the matrix created in the analysis inside the `data/nf-rnaSeqMetagen`. The application is started inside the folder using `python` and accessed on the browser.

5.3.2 Using `rnaSeqMetagen` for metagenomic analysis of the SSc data

The `rnaSeqMetagen` was used to perform metagenomic analysis on the RNA-seq data of the 25 samples described in Chapter 2. The analyses were performed on the Wits Computing cluster using the SLURM scheduler. Figure 5.2 shows the summary HTML report generated by `Nextflow` from executing the workflow on the 25 samples. The overall execution time for the pipeline on all the samples was 07:54:50 hours. The execution time also includes the time it took for each job to be queued on the cluster. Other interactive HighCharts (https://www.highcharts.com/) reports produced by `Nextflow` on execution of the `rnaSeqMetagen` pipeline for CPU usage, memory usage and execution time are discussed below.

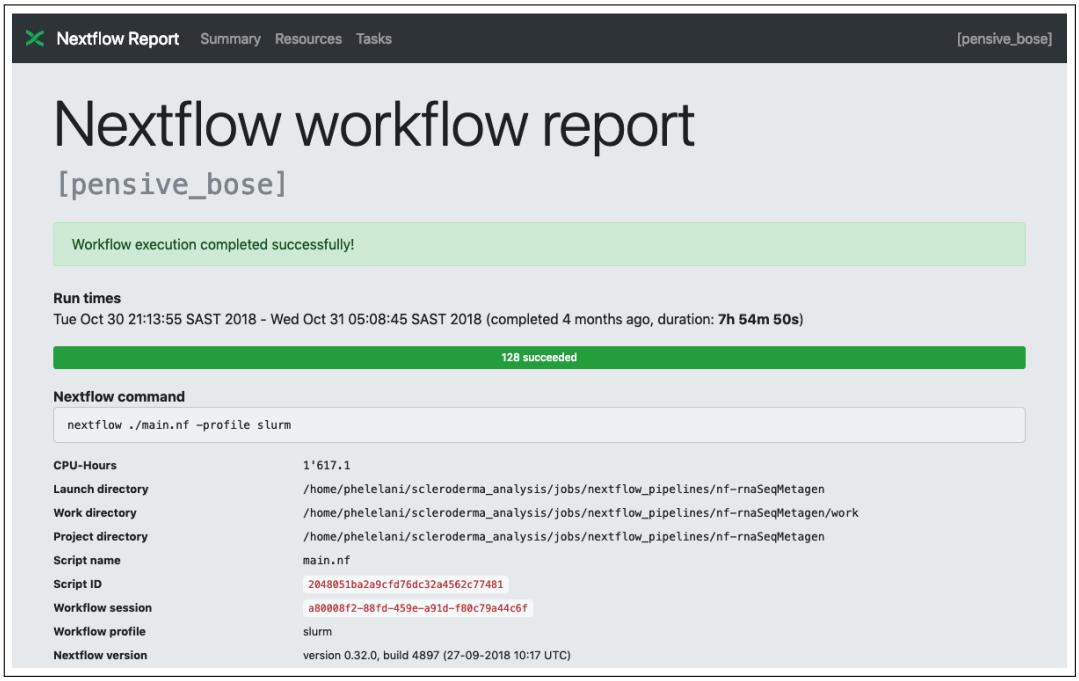


Figure 5.2: Summary report and metadata for `rnaSeqMetagen` pipeline execution. Overall execution time and metadata for performing analysis on all 25 RNA-seq samples from the SSc patients using `rnaSeqMetagen` pipeline

5.3.2.1 CPU usage

The % CPU usage for each of the processes in the `rnaSeqMetagen` is summarised in Figure 5.3. All the processes in the pipeline used at most 100% of the CPUs allocated, with an exception of the `runStar` process, where some of the samples used more than the allocated CPUs. 13 CPUs were allocated for the `runStar`.

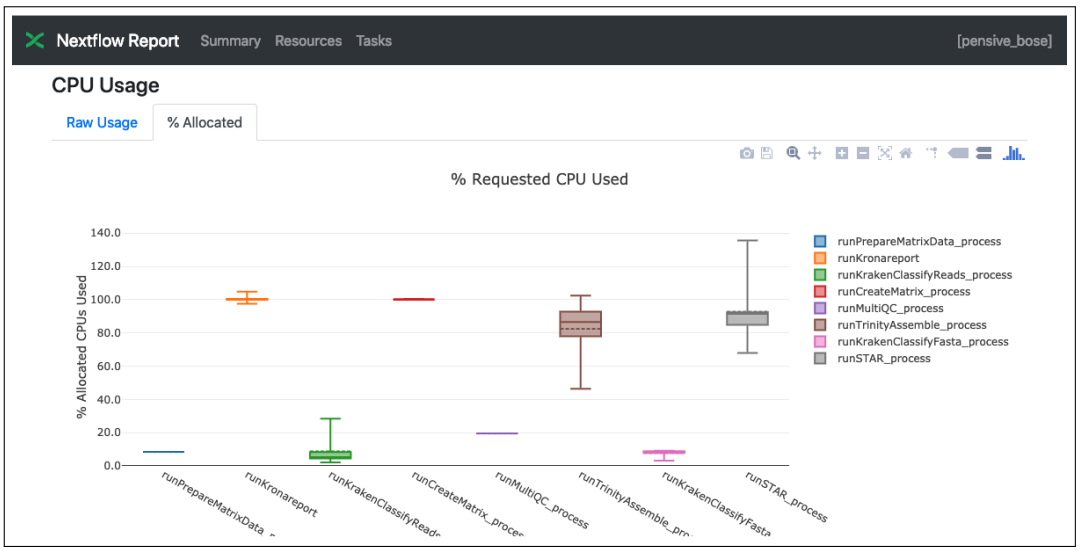


Figure 5.3: % CPU usage by each process of the `rnaSeqMetagen` pipeline.

5.3.2.2 Memory usage

Figure 5.4 summarises the percentage memory usage by each process in the `rnaSeqMetagen` pipeline. All the processes in the workflow utilised memory within the limit requested. Not more than 70% of memory requested for each process was used. However, the amount of memory reported is the virtual memory; the amount actual memory used is more. The requested memory could be reduced for jobs to spend less time on the job queue.

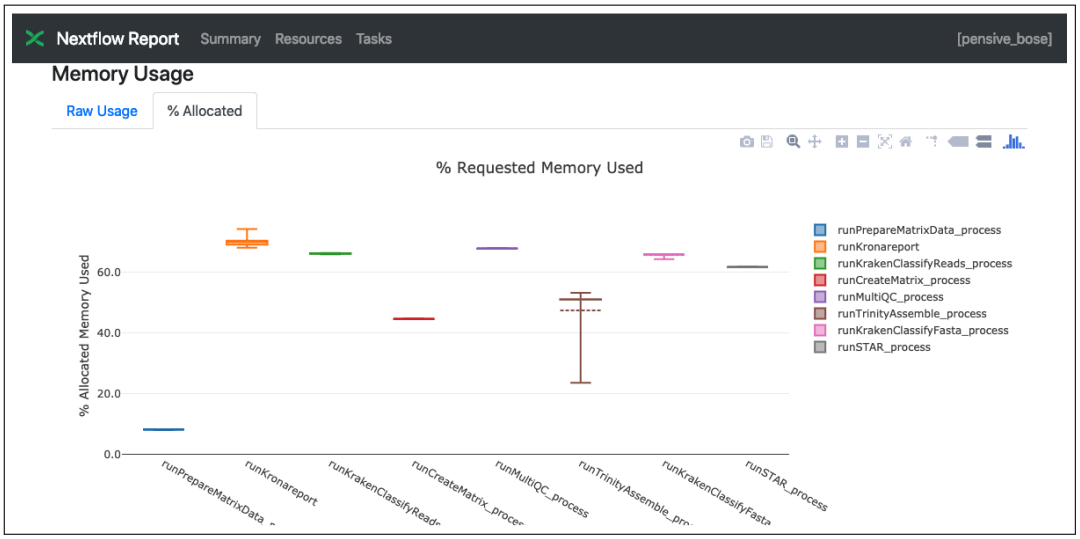


Figure 5.4: % Memory usage by each process of the `rnaSeqMetagen` pipeline.

5.3.2.3 Execution time

The execution time for each process in the `rnaSeqMetagen` workflow is summarised in Figure 5.5. Again, as with other resources requested, the amount or time used by each process (%) is way less than the time requested. All the processes took less than 30% to execute the required computation on each sample.

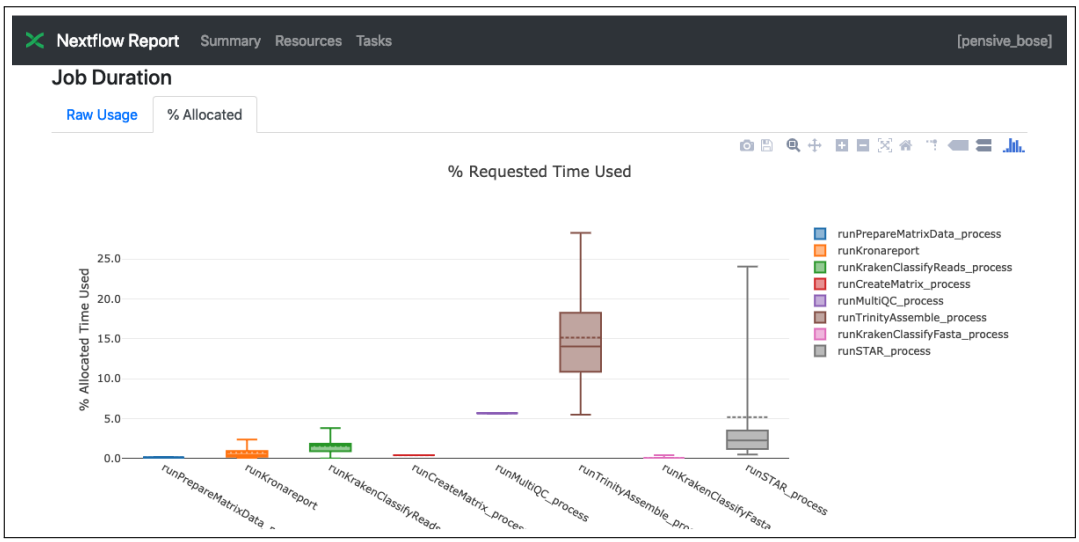


Figure 5.5: % Requested time usage by each process of the `rnaSeqMetagen`pipeline.

5.3.2.4 QC plots produced by `rnaSeqMetagen`

The `MultiQC` QC plot produced by the `rnaSeqMetagen` pipeline reveals that more than 95% overall of the reads from all the samples uniquely aligned to the reference genome, and ∼3% - 4% were mapped to multiple loci. This means that less than 1% of the reads were unmapped, thus exogenous, in all of the samples. These are the reads that could possibly arise from microorganism residing in the skin tissues of the affected SSc patients. It is worth noting that the number of unmapped reads in the patients is higher, on average, compared to the control group. These unmapped reads were further analysed for alignment with the genomes of pathogens that could be associated with the disease being studied here.

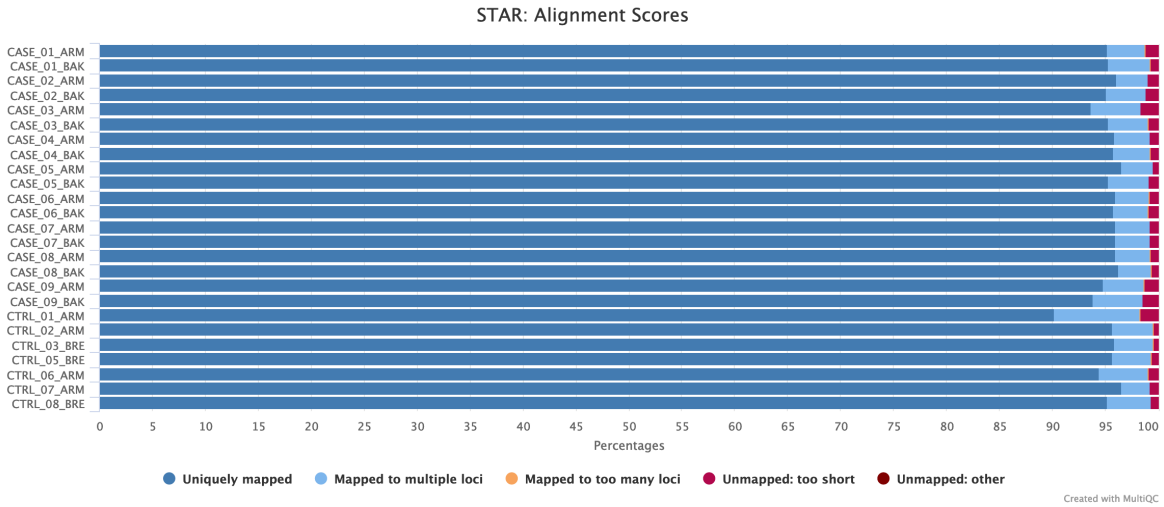


Figure 5.6: Alignment of reads to the reference genome by `STAR` in the `rnaSeqMetagen` pipeline

5.3.3 Exploring possible pathogens in SSc patients using `UpSet` and `krona` charts

The main aim of developing the `rnaSeqMetagen` pipeline presented in this chapter was to aid in the metagenomic analysis of RNA-seq data. Once all the analyses have been performed by the pipeline, the results can be used to explore the taxonomic classifications of the unclassified reads and assembled contigs. One tool that is of particular importance in the exploration of the taxonomic classification is the `UpSet` visualisation tool included in the results, which is accompanied by the matrix of taxonomic classifications produced by the pipeline. The Web version of the `UpSet` visualisation program, deployed locally using `python` can be used to identify taxonomies that are unique/shared between the samples. The results folder of the `rnaSeqMetagen` contains a GitHub repository of `UpSet`, with the matrix inside the `data` folder of `UpSet`. To launch the `UpSet` program, users have to navigate to the `upset` folder inside the `rnaSeqMetagen` results folder and execute the following command:

```
## Start UpSet program
python -m SimpleHTTPServer
```

The application can be accessed on any Web browser (http://localhost:8000/), and then loading the JSON file (`data/nf-rnaSeqMetagen/nf-rnaSeqMetagen.json`) using the "**Load Data**" button. Users can then select the data sets and add custom queries to explore the taxonomic classifications between the samples they are interested in. For this study, the taxonomies that are shared between the SSc patients (at least two or more forearm and/or back samples), but not found in any of the control samples were identified. The same samples that were excluded for differential expression (P1, B1, P9, B9 and C1 as discussed in Section 4.2.1) were also excluded in these analyses.

Figure 5.7 shows the results from querying `UpSet` for taxonomies shared between the SSc patient samples but not found in the control samples. A detailed list of the taxonomies or species names that were found in Figure 5.7 can be found in Appendix N. When looking at Figure 5.7, the bottom five shared taxonomies have the highest degree of intersections between the cases. The highest degree of intersection has nine samples (P2, B2, B3, P4, B4, P5, P6, B6 and P8) and the taxonomy corresponding to this intersection is the *Brachybacterium* genus.

The second highest degree of intersection between the patients has eight samples (P2, B2, B3, P4, B4, P5, B6 and P8) and corresponds to two bacterial species, *Brachybacterium saurashtrense* and *Dietzia sp. oral taxon 368*. Two intersections have seven samples in them; the intersection with samples P2, B2, P4, B4, P5, B6 and P8 corresponding to the *Dietzia* genus and *Arthrobacter sp. QXT-31* species; and the intersection with samples P2, B2, B3, P4, B4, P5 and P8 which has the species *Pseudarthrobacter chlorophenolicus A6*,

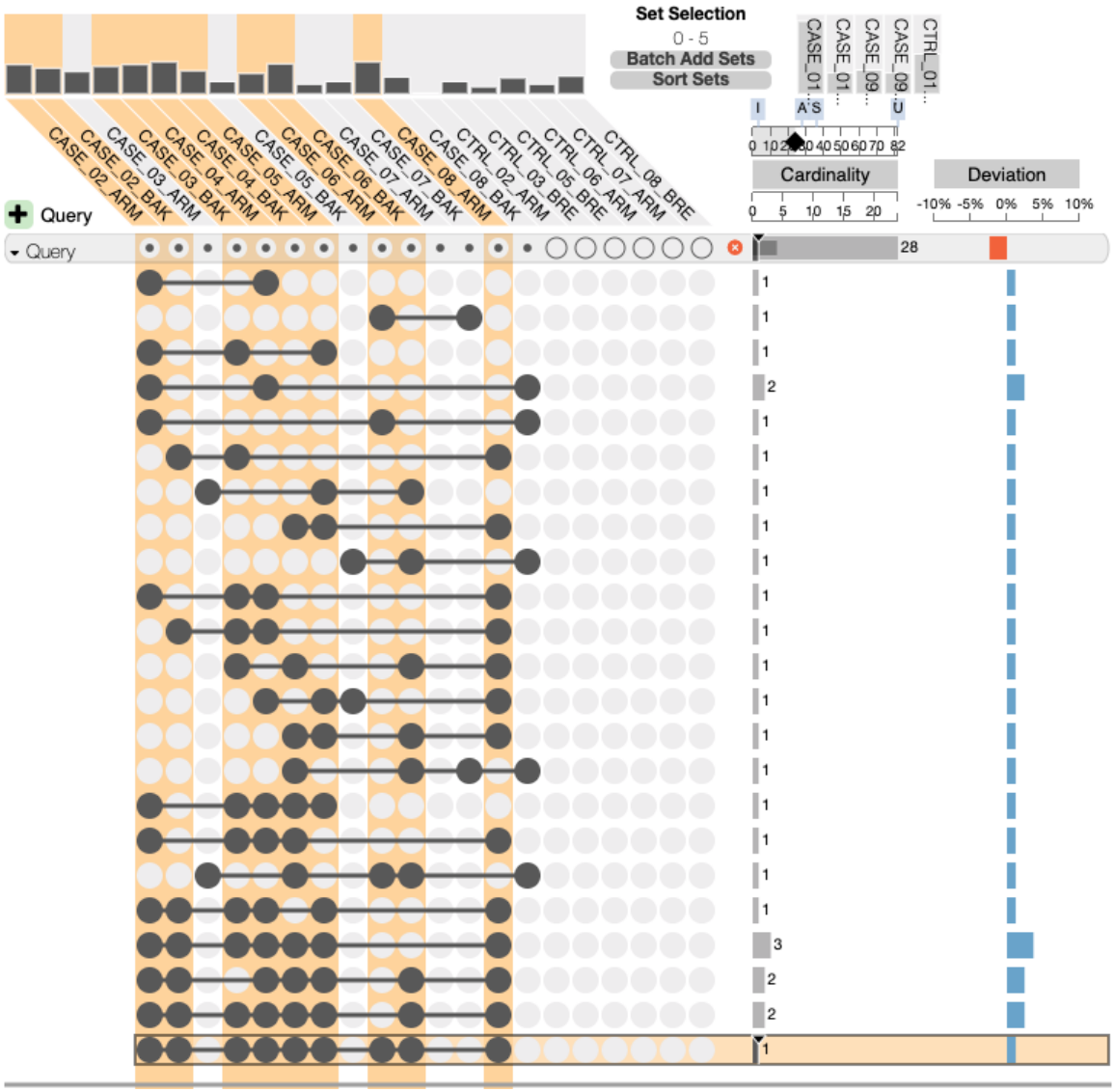


Figure 5.7: `UpSet` query for identifying microbial taxonomies shared between SSc patient forearm and/or back samples. Columns represents the samples and rows represent taxonomies. Samples with no taxonomies shared with other samples have unfilled cells (●). Samples that have shared taxonomies have filled cells (●) connected by lines. Columns highlighted in orange are all the cases samples that share the *Brachybacterium* genus.

Pseudarthrobacter phenanthrenivorans Sphe3 and *Brachybacterium sp. P6-10-X1* shared between them. *Dietzia lutea* species is shared between the samples P2, B2, B3, P4, P5 and P8. Appendix N gives a detailed list of the species that are common between the samples from the affected patients.

In the list of the pathogens that are shared between the SSc patient samples but not found in the controls (Appendix N, shaded in gray), there was more than one species belonging to the *Arthrobacter*, *Bacillus*, *Brachybacterium*, *Dietzia* and *Pseudarthrobacter* genera. It is possible that these species (or some other species belonging to the different genera) could play roles in the initiation of the disease through infection, or they are opportunistic pathogens as a result of compromised immunity in the disease. It is of interest to investigate the possible link that might exist between the onset and progression

of SSc and these species (and genera) of pathogens.

5.4 Discussion

Studying microbial genomes within HTS data using metagenomic analyses is a rapidly growing field. The application of this kind of study on RNA-seq data allows for the identification of possible pathogens that might be associated with certain diseases, especially those pathogens that take advantage of the compromised host. The aim of this chapter was to perform metagenomic analysis on RNA-seq data from the SSc patients with the aid of a novel metagenomic pipeline. The `rnaSeqMetagen` (https://github.com/phelelani/nf-rnaSeqMetagen) pipeline designed in this study has met the requirements for the three main features of an efficient pipeline, i.e., reproducibility, scalability and portability. The pipeline is implemented in `Nextflow` and all the applications needed to execute the workflow are containerised in `Singularity`. Users wishing to use the workflow can download it from the GitHub repository, which also includes detailed documentation on using the pipeline. All images for the `rnaSeqMetagen` workflow are hosted on SingularityHub (https://www.singularity-hub.org/collections/728) and can be downloaded directly, or using the `prepare_data.nf` script packaged with the workflow.

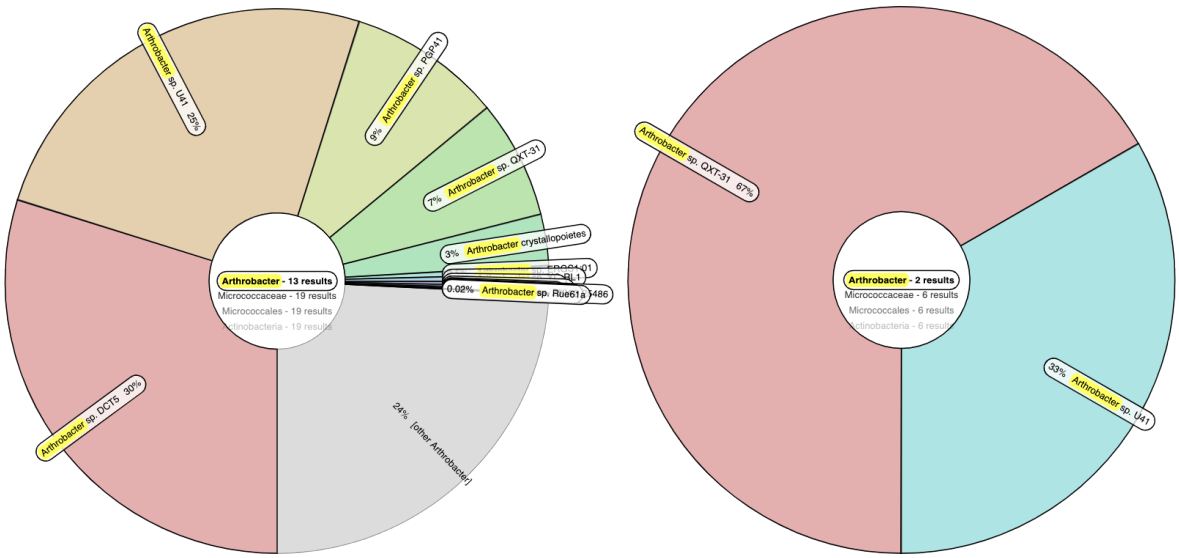
In addition to testing and implementing the `rnaSeqMetagen` pipeline on the Wits cluster using the PBS and SLURM schedulers, it has also been tested on different computing environments, including the UCT eResearch HPC and Amazon AWS. However, due to the `Singularity` version 3 installed on the UCT eResearch HPC, the pipeline could not be properly executed as it depends on version 2.6 of `Singularity`. Testing the `rnaSeqMetagen` workflow on the Amazon AWS using the same EC2 instance as the one used for the `rnaSeqCount` pipeline (using the `Nextflow ami-4b7daa32` AMI, EBS with 1000GB of storage and the `m4.10xlarge` EC2 instance with 40 virtual CPUs and 160GB of memory) was also not successful. The reason that the `rnaSeqMetagen` was not tested successfully on the Amazon AWS was because of the issue of parallelisation and cost, especially since it has computationally intensive processes (assembly of unmapped reads using `trinity`, which requires approximately 150GB of memory). When the `trinity` assembly process is executed, no other process can run as it uses up all the memory, thus only one sample can be run at a time. A single `trinity` assembly process took close to 250 minutes, and for 25 samples to have to each perform such analysis would be very costly. Future improvements for this issue on the Amazon AWS would be to include `awsbatch` support, which allows more instances to be initiated when computational requirements are high.

However, even though there were challenges encountered in testing the `rnaSeqMetagen` pipeline on different computing environments, the pipeline was able to serve its purpose, i.e., performing metagenomic analysis with the purpose of identifying possible microorganisms that might be associated with SSc. The `rnaSeqMetagen` pipeline attempts to

addresses three challenges in the metagenomic analyses: (1) excluding host reads from the RNA-seq data; (2) taxonomic classification of the remaining unmapped/exogenous reads; and (3) identifying which taxonomies/pathogens are relevant to the disease being studied here. The issue of excluding host reads from the RNA-seq data was accomplished through the `STAR` aligner (Dobin *et al.*, 2013). As seen in Chapter 3, `STAR` was able to uniquely map more than ∼95% of the reads in each sample to the human reference genome. The same was also seen in the `rnaSeqMetagen` pipeline. More than ∼95% of the reads were uniquely mapped, whilst ∼4% were mapped to multiple regions in the genome, leaving less than ∼1% of the reads being unmapped (Figure 5.6). This small fraction of unmapped reads represents possible reads belonging to either pathogens or contaminants.

The second challenge of classifying reads on taxonomy was accomplished through `kraken2` and its database of archaeal, viral, bacterial and human sequences (Wood and Salzberg, 2014). Even though `kraken2` is able to classify short raw reads, an optimal method would be to first assemble the reads into longer contigs. This method does not only improve the accuracy of `kraken2`'s classification algorithm, but it also tries to use all the sequences from the unmapped reads to reconstruct, "*de novo*", the possible pathogenic sequences that gave rise to the unmapped reads. The unmapped reads are assembled in the `rnaSeqMetagen` pipeline using `trinity`, then the resulting longer sequences were assigned to taxonomies using `kraken2`. This method of first assembling reads before classification greatly reduces noise and increases specificity of classification as there would be fewer sequences to classify after assembly compared to the hundreds of thousands of short unmapped reads being directly classified. This can be seen, for example, in the P2 sample when looking at the *Arthrobacter* genus to see which species are identified in both reads and assembled sequences (Figure 5.8). There are 13 species identified from classifying reads, compared to only 2 species of *Arthrobacter* when classifying assembled sequences. This is because `kraken2` matches all the portions (reads) to the database, and because of shared sequences within species, all these pieces match to different species of the same genus, thus increasing noise. Assembling the reads reduces this noise and increases specificity of classification.

The last issue that `rnaSeqMetagen` addresses is the method for identifying which of the taxonomies in the classifications are relevant to the disease. This is accomplished through the `UpSet` tool. A combination matrix of all the taxonomies (rows) and samples (columns) is created, and each cell in the matrix either has a "1" or "0" to represent whether that taxonomy is found in that sample or not, respectively. This matrix is created by the `rnaSeqMetagen` pipeline and can be loaded directly onto the `UpSet` visualisation application found in the results as explained in Section 5.3.3. Users can interact with the `UpSet` visualisation tool and create custom queries that lets them identify taxonomies that are unique or shared in the different samples they are interested in. For this study, 34 species (and genera) were identified, and five of the genera (*Arthrobacter*, *Bacillus*,



(a) P2 reads classification `krona` chart for the *Arthrobacter* genus.

(b) P2 assembled sequences classification `krona` chart for the *Arthrobacter* genus.

Figure 5.8: `krona` charts for the *Arthrobacter* genus classification.

Brachybacterium, *Dietzia* and *Pseudarthrobacter*) had more than one species that had a high degree of sharedness between the samples of SSc patients, but not found in the controls. Section 5.4.1 below discusses the clinical relevance of the identified genera to SSc.

5.4.1 Clinical relevance of identified genera

There have been some reports of clinical relevance of species belonging to the five genera identified in the SSc patients samples in this study. Funke *et al.* (1998) reported different species of the *Arthrobacter* genus isolated from patients with urinary tract infections, chronic otorrhea (ear infection), vaginal swab, blood culture, leg wound and deep tissue infection in the upper leg. Another case of *Arthrobacter* species infection was reported by Wang *et al.* (2005), where they isolated *Arthrobacter scleromae* from a swollen scleromata (back and hip) of a patient with dermatosis. Some species from the *Dietzia* genus have been implicated to be of clinical relevance as they were isolated from patients with acute infections. These include *Dietzia cinnamea*, *Dietzia maris* and *Dietzia papillomatosis* (Koerner *et al.*, 2009). *Dietzia cinnamea* was isolated from a patient with infection who had a bone marrow transplant. *Dietzia maris* was isolated from blood cultures of an immunocompromised patient suffering from septic shock, another patient who had a hip infection as well as from an aortic wall of a patient suffering from aortitis. *Dietzia papillomatosis* was isolated from a patient with reticulated papillomatosis (CRP), a rare skin disorder characterised by scaly papules with red to brown-grey discoloration (Koerner *et al.*, 2009).

The study by Bank (2016) examined the role of *Bacillus subtilis* in pathogenic pulmonary fibrosis using mice, following infection. In this study, mice were repeatedly exposed to *Bacillus subtilis*, and they observed that the mice had a significant increased collagen

deposition in the lungs after exposure. This suggested a possible role of the pathogen in inducing immune responses leading to the pulmonary fibrosis observed. Species of *Brachybacterium* genus have been reported in human clinical samples. The study by Tamai *et al.* (2018) isolated a *Brachybacterium* species from blood that was a causative agent of blood-stream infection. Mages *et al.* (2008) reported 38 different strains of *Brachybacterium* isolated from blood cultures (12 strains), wounds (12 strains), urine samples (8 strains), sterile tissue (5 strains) and respiratory tract (1 strain). Another study by Martel *et al.* (2008) reported *Brachybacterium* species isolated from dermatitis and organ lesions of agamid lizards.

A literature search revealed no clinical association of the *Pseudarthrobacter* genus with SSc. However, Ben Fekih *et al.* (2018) recently reported isolating *Pseudarthrobacter* sp. AG30 from soil of a gold-copper mine in China. A number of studies have associated gold-mining and exposure to silica as an environmental risk factors to developing SSc (Cowie and Dansey, 1990; Tager and Tikly, 1999). Identification of the *Pseudarthrobacter* genus amongst the SSc patients, as well as the report of isolation of its species in a gold-copper mine could, perhaps, imply a possible link between *Pseudarthrobacter* species and the onset/progression of SSc. The presence of the five main genera identified in this study, and their clinical relevance in human infections and diseases, provides a potential role of these bacterial species (and their sequences identified) in the dysregulated immunity and fibrosis in SSc.

In conclusion, this chapter has presented a novel `rnaSeqMetagen` pipeline for performing metagenomic studies on RNA-seq data. The pipeline meets the three main features of efficient workflows in the sense that it is reproducible, portable and scalable to different types of computing platforms. The `rnaSeqMetagen` pipeline was used in this study to identify possible pathogens associated with the disease, and has revealed a number of bacterial species belonging to five genera (*Arthrobacter*, *Bacillus*, *Brachybacterium*, *Dietzia* and *Pseudarthrobacter*) to be of clinical relevance and highly common in the SSc patients, and not found in the control group. What is mostly interesting about the identified species and genera shared between the SSc patients is that it was through per sample *de novo* sequence reconstruction of the unmapped reads and classification using `kraken2`, independent of other samples. The `rnaSeqMetagen` pipeline was able to reconstruct sequences in each sample, and at the same time identify the assembled sequences as belonging to common taxonomies between the SSc patients. It would be interesting to investigate the sequences classified as belonging to the genera mentioned, perform functional annotation and evaluate their possible roles in SSc.

Chapter 6

Concluding Discussion

The studies presented in this thesis were aimed at developing bioinformatic workflows to aid in the analyses of RNA-seq data and to contribute to the limited understanding of the molecular aetiology of SSc by performing comparative bioinformatic analyses of transcriptomic/RNA-seq data produced from black South African patients affected with SSc and healthy individuals. The four main objectives (Section 1.10) that were set out guided this study in terms of the choices of bioinformatic analyses to be performed and the choices of tools to be developed and used in the exploratory analysis of the RNA-seq data.

The data for this study were produced for another study by Frost (2016), where the gene expression in the Wnt signaling pathway was investigated (Frost *et al.*, 2018). Since the data were generated for another study, I had no contribution to the study design, which was flawed. Ideally a larger sample of sex and body site matched cases and controls would have provided more power to the study. However, steps were taken to overcome these limitations. The study itself is unique in the sense that no other study of this kind has been performed in South Africa, i.e., it is the first study that uses bioinformatic analyses on RNA-seq data produced from black South Africans affected with SSc in an effort to understand gene expression profiles in the disease.

Chapter 1 of this study addressed the current knowledge of SSc as well as the major concepts of workflow design since the study was aimed at developing workflows to ease the analysis of RNA-seq data for this study as well as for other researchers wishing to perform similar studies. Chapter 2 addressed the data used in this study, i.e., the demographic information of SSc patients and control individuals from which the data was collected and the pre-processing steps taken to ensure that the data was "clean" before it could be used for downstream analyses.

Chapter 3 addressed the first workflow of the study, which takes raw RNA-seq reads, aligns them to a reference genome to identify genes/transcripts and produces a matrix of features and their abundance that is used for differential expression analysis. Chapter 4 addressed the differential expression analysis using the results from the first workflow of the study. Chapter 5 addressed the designing of a novel workflow for metagenomic analysis, which takes raw RNA-seq data, filters out host reads and classifies the remaining unmapped reads as they are or after *de novo* assembly. In this chapter, the main findings of this study, limitations and future perspectives will be discussed.

6.1 Reproducible workflows for RNA-seq data analyses

Two novel workflows, `rnaSeqCount` and `rnaSeqMetagen`, were developed in this study using `Nextflow` and `Singularity` in order to facilitate RNA-seq data analysis for this study as well as to be used by other researchers wishing to carry out analyses similar to this study. Both workflows are reproducible, portable and scalable; all of which are properties required for an efficient workflow. The pipelines are available on GitHub (`rnaSeqCount`: https://github.com/phelelani/nf-rnaSeqCount; `rnaSeqMetagen`: https://github.com/phelelani/nf-rnaSeqMetagen), where they have been extensively documented.

Users wishing to use the workflows can do so by cloning the repositories onto their computational platform (desktop, HPC or cloud) with UNIX-based OS. The availability of `Singularity` images on SingularityHub for executing both `rnaSeqCount` (https://www.singularity-hub.org/collections/770) and `rnaSeqMetagen` (https://www.singularity-hub.org/collections/728) eliminates the need for manual installation of applications used by the tools. Although these workflows have proven to be useful in this study, future improvements could include adding support for the latest version of `Singularity` as well as the "`awsbatch`" support for performing parallelised job submission on the Amazon EC2 machine.

6.2 Differential expression & pathway analysis

The `rnaSeqCount` pipeline developed here provided a relatively quick and efficient way to obtain a matrix of read counts that can be used for differential expression analysis. The workflow could produce the matrix for all 25 samples from the SSc data in ∼3 hours. A workflow was created in `R` (https://github.com/phelelani/transcriptomics) specifically for performing differential expression analysis of the 25 samples from the SSc RNA-seq data. Further steps were taken to reduce bias in the data, including removing genes on the Y-chromosome and performing initial differential expression analysis on the control group (males vs females) in order to remove genes that might have a negative effect on the analysis.

Samples P1 and B1 were excluded from the analysis as there was no clinical information. Samples P9, B9 and C1 were also excluded in order to keep balance of male to female ratio between the patient and control groups. A total of nine comparisons were made in order to determine the differential expression of genes based on all the samples, disease severity (mild and severe), tissue sites (forearms and backs) as well as the within-individual effects. The nine different comparisons performed for the differential expression analyses also demonstrate different ways in which the RNA-seq data could be used to identify significantly differentially expressed genes. This method evaluates all possible comparisons between the cases and controls, and is not limited to a single comparison of looking at

differences between cases and controls.

The main focus of the differential expression was on the affected forearms; the other comparisons were used to increase confidence in the genes identified between the affected forearms and the controls. Using a $|\log_2 FC|$ of 2 and a FDR of 0.01, 77 genes were identified as being significantly differentially expressed in the affected forearms in these analyses, 27 of which showed down-regulation in clusters with genes identified in the Open Targets Platform to be associated with SSc. Some of these genes supported findings from other studies done on SSc, and their down-regulation seems to be correlated with the disease duration. However, the fact that the cases and controls were not well matched for age, sex and biopsy site, could also contribute to the differential expression results obtained. The pathways identified here using `gage` and `pathview` were of significance to SSc and also show down-regulation of the genes playing roles in the pathways.

6.3 Metagenomic analysis

The metagenomic analysis using the `rnaSeqMetagen` workflow developed here revealed that there were more than one species belonging to five genera in the patients that were not present in the control group. These were *Arthrobacter*, *Bacillus*, *Brachybacterium*, *Dietzia* and *Pseudarthrobacter*. These genera are of clinical significance as they have been associated with diseases and infections in humans. However, further investigation needs to be done on the sequences that produced the taxonomic classifications of these genera. The inclusion of the `UpSet` tool with the results from running the `rnaSeqMetagen` makes it a powerful tool in trying to discern which organisms are shared or unique in the samples being analysed.

6.4 Study limitations and future work

Even though the workflows developed in this study have proved to be of great value to the analyses of RNA-seq data in terms of differential expression, pathway and metagenomic analyses, the workflows do have their limitations. Both `rnaSeqCount` and `rnaSeqMetagen` were tested on the Wits Computing cluster, UCT eResearch HPC and Amazon AWS. The schedulers and job queues currently available for the pipeline only work on these computational platforms, however, advanced users can modify the `nextflow.config` file to suit their computational environment. The workflows also currently support version 2.6 of `Singularity`. These computational limitations will be addressed in future improvements of the workflows by adding options for users to specify the job schedulers and queues available on the computational platforms withouth having to edit the `nextflow.config` file. A support for `Singularity` version 3 will also be added to the workflows.

The `rnaSeqMetagen` workflow currently has no statistical measure of significance for the identified taxonomies that are shared/unique in the samples being analysed, i.e., it currently creates a combination matrix using all the unique taxonomies identified in each

sample, and marks them as present or not. The combination matrix is used in `UpSet` for creating custom queries and exploring taxonomies between the samples. I would like to address this limitation in future improvements of the `rnaSeqMetagen` workflow by incorporation a quantitative measure of significance for the identified taxonomies or pathogens. This statistical measure would also allow for the identification of taxonomies that are common/shared between the cases and controls, but have more relevance in the patient group in terms of disease association. I would also like to further improve the workflow by adding a method for obtaining all the *de novo* assembled sequences from each sample, perform functional annotation and identify possible functions of the sequences in relation to the disease being studied.

The `rnaSeqCount` workflow can be further extended to include differential expression and pathway analyses in the pipeline, in addition to producing raw read counts. Although this can be challenging, especially since performing differential expression analysis is an interactive process requiring addition of different parameters in different steps, the first step would be to extend the analyses only a single pairwise comparison (cases vs. controls) of differential expression and pathway analyses, which would require minimal user input. Users would only need to specify the samples for each of the two groups being compared, the desired $|\log_2 FC|$ and FDR needed to perform differential expression analysis. The pathway analysis part of the workflow would the take the output from the expression analysis and identify pathways affected by the differentially expressed genes identified.

6.5 Conclusion

This study used a combination of bioinformatic analyses to explore the RNA-seq data from black South African patients with SSc and healthy individuals. Novel workflows were developed that facilitated the differential expression and metagenomic analyses performed. However, the application of the workflows was limited in that the test data was from a small sample, and cases and controls were poorly matched for age, sex and site of tissue where the samples were collected. The strength of the study is that two independent workflows have been developed that can be applied to any RNA-seq data where the objective is to examine differential gene expression or the analysis of non-host sequences to explore potential pathogens in groups of samples compared to controls. The workflows are well documented and are available on GitHub and facilitate reproducibility, portability and scalability of RNA-seq data analyses.

References

Abbas, A. K., Lichtman, A. H. and Pillai, S. (2015) Innate Immunity: The early defense against infections in Abbas, A. K., Lichtman, A. H. and Pillai, S., editors, *Basic Immunology: Functions and Disorders of the Immune System* chapter Chapter 2 5th edition.

Abbas, A. K., Lichtman, A. H. and Pillai, S. (2018) Innate Immunity in Abbas, A. K., Lichtman, A. H. and Pillai, S., editors, *Cellular and molecular immunology* chapter Chapter 4, 57–95 Elsevier 9th edition ISBN 9780323523233.

Acland, A., Agarwala, R., Barrett, T. *et al.* (2014) Database resources of the National Center for Biotechnology Information *Nucleic Acids Research* **42**, D1, D7–D17 ISSN 0305-1048.

Adelowo, O. O. and Oguntona, S. (2009) Scleroderma (systemic sclerosis) among Nigerians. *Clinical rheumatology* **28**, 9, 1121–1125 ISSN 14349949.

Afgan, E., Baker, D., van den Beek, M. *et al.* (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update *Nucleic Acids Research* **44**, W1, W3–W10 ISSN 0305-1048.

Agarwal, P. and Owzar, K. (2014) Next generation distributed computing for cancer research *Cancer Informatics* **2014**, 97–109 ISSN 11769351.

Agarwala, R., Barrett, T., Beck, J. *et al.* (2018) Database resources of the National Center for Biotechnology Information *Nucleic Acids Research* **46**, D1, D8–D13 ISSN 0305-1048.

Allanore, Y., Dieude, P. and Boileau, C. (2010) Genetic background of systemic sclerosis: autoimmune genes take centre stage *Rheumatology* **49**, 2, 203–210 ISSN 1462-0324.

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data *Genome Biology* **11**, 10, R106 ISSN 1465-6906.

Anders, S., McCarthy, D. J., Chen, Y. *et al.* (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor *Nature Protocols* **8**, 9, 1765–1786 ISSN 17542189.

Anders, S., Pyl, P. T. and Huber, W. (2015) HTSeq-A Python framework to work with high-throughput sequencing data *Bioinformatics* **31**, 2, 166–169 ISSN 14602059.

Anderson, L. E., Treat, J. R., Licht, D. J. *et al.* (2018) Remission of seizures with immunosuppressive therapy in Parry-Romberg syndrome and en coup de sabre linear scleroderma: Case report and brief review of the literature *Pediatric Dermatology* **35**, 6, e363–e365 ISSN 07368046.

Andrews, S. (2010) FastQC: A quality control tool for high throughput sequence data.

Arron, S. T., Dimon, M. T., Li, Z. *et al.* (2014) High Rhodotorula Sequences in Skin Transcriptome of Patients with Diffuse Systemic Sclerosis *Journal of Investigative Dermatology* **134**, 8, 2138–2145 ISSN 0022202X.

Asano, Y. (2018) Systemic sclerosis *The Journal of Dermatology* **45**, 2, 128–138 ISSN 03852407.

Asano, Y. and Sato, S. (2015) Vasculopathy in scleroderma *Seminars in Immunopathology* **37**, 5, 489–500 ISSN 1863-2297.

Bakhtiar, S. M., LeBlanc, J. G., Salvucci, E. *et al.* (2013) Implications of the human microbiome in inflammatory bowel diseases *FEMS Microbiology Letters* **342**, 1, 10–17 ISSN 03781097.

Balbir-Gurman, A. and Braun-Moscovici, Y. (2011) Scleroderma overlap syndrome. *The Israel Medical Association journal : IMAJ* **13**, 1, 14–20 ISSN 1565-1088.

Bank, I. (2016) The Role of $\gamma\delta$ T Cells in Fibrotic Diseases *Rambam Maimonides Medical Journal* **7**, 4, e0029.

Ben Fekih, I., Ma, Y., Herzberg, M. *et al.* (2018) Draft Genome Sequence of Pseudarthrobacter sp. Strain AG30, Isolated from a Gold and Copper Mine in China *Microbiology Resource Announcements* **7**, 17, 4–5 ISSN 2576-098X.

Blischak, J. D., Davenport, E. R. and Wilson, G. (2016) A Quick Introduction to Version Control with Git and GitHub *PLOS Computational Biology* **12**, 1, e1004668 ISSN 1553-7358.

Boettiger, C. (2015) An introduction to Docker for reproducible research *ACM SIGOPS Operating Systems Review* **49**, 1, 71–79 ISSN 01635980.

Bogdanos, D. P., Smyk, D. S., Rigopoulou, E. I. *et al.* (2012) Twin studies in autoimmune disease: Genetics, gender and environment *Journal of Autoimmunity* **38**, 2-3, J156–J169 ISSN 08968411.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data *Bioinformatics* **30**, 15, 2114–2120 ISSN 1460-2059.

Brabcová, E. and L. Kolesár, E. Thorburn, I. S. (2014) Chemokines Induced in Human Respiratory Epithelial Cells by IL-1 Family of Cytokines *Folia Biol (Praha)* **60**, 180 – 186.

Bray, N. L., Pimentel, H., Melsted, P. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification *Nature Biotechnology* **34**, 5, 525–527 ISSN 1087-0156.

Broen, J. C. A., Radstake, T. R. D. J. and Rossato, M. (2014) The role of genetics and epigenetics in the pathogenesis of systemic sclerosis *Nature Reviews Rheumatology* **10**, 11, 671–681 ISSN 1759-4790.

Careta, M. F. and Romiti, R. (2015) Localized scleroderma: clinical spectrum and therapeutic update *Anais Brasileiros de Dermatologia* **90**, 1, 62–73 ISSN 0365-0596.

Choi, M. Y. and Fritzler, M. J. (2016) Progress in understanding the diagnostic and pathogenic role of autoantibodies associated with systemic sclerosis *Current Opinion in Rheumatology* **28**, 6, 586–594 ISSN 1040-8711.

Chora, I., Guiducci, S., Manetti, M. *et al.* (2015) Vascular biomarkers and correlation with peripheral vasculopathy in systemic sclerosis *Autoimmunity Reviews* **14**, 4, 314–322 ISSN 15689972.

Conesa, A., Madrigal, P., Tarazona, S. *et al.* (2016) A survey of best practices for RNA-seq data analysis *Genome Biology* **17**, 1, 13 ISSN 1474-760X.

Conway, J. R., Lex, A. and Gehlenborg, N. (2017) UpSetR: an R package for the visualization of intersecting sets and their properties *Bioinformatics* **33**, 18, 2938–2940 ISSN 1367-4803.

Costa-Silva, J., Domingues, D. and Lopes, F. M. (2017) RNA-Seq differential expression analysis: An extended review and a software tool *PLOS ONE* **12**, 12, e0190152 ISSN 1932-6203.

Costenbader, K. H., Gay, S., Alarcón-Riquelme, M. E. *et al.* (2012) Genes, epigenetic regulation and environmental factors: Which is the most relevant in developing autoimmune diseases? *Autoimmunity Reviews* **11**, 8, 604–609 ISSN 15689972.

Cowie, R. L. and Dansey, R. D. (1990) Features of systemic sclerosis (scleroderma) in South African goldminers *South African Medical Journal* **77**, 8, 400–402 ISSN 02569574.

Del Papa, N. and Pignataro, F. (2018) The Role of Endothelial Progenitors in the Repair of Vascular Damage in Systemic Sclerosis *Frontiers in Immunology* **9**, JUN, 1–10 ISSN 1664-3224.

Dennis Jr, G., Sherman, B. T., Hosack, D. A. *et al.* (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery *Genome Biology* **4**, 5, P3 ISSN 1465-6906.

Denton, C. P. (2015) Advances in pathogenesis and treatment of systemic sclerosis *Clinical Medicine* **15**, Suppl_6, s58–s63 ISSN 1470-2118.

Denton, C. P. and Khanna, D. (2017) Systemic sclerosis *The Lancet* **390**, 10103, 1685–1699 ISSN 01406736.

Desbois, A. C. and Cacoub, P. (2016) Systemic sclerosis: An update in 2016 *Autoimmunity Reviews* **15**, 5, 417–426 ISSN 15689972.

Di Sabatino, A., Monteleone, G., Laudisi, F. *et al.* (2016) CCL20 Is Negatively Regulated by TGF-β1 in Intestinal Epithelial Cells and Reduced in Crohn's Disease Patients With a Successful Response to Mongersen, a Smad7 Antisense Oligonucleotide *Journal of Crohn's and Colitis* jjw191 ISSN 1873-9946.

Di Tommaso, P., Chatzou, M., Floden, E. W. *et al.* (2017) Nextflow enables reproducible computational workflows *Nature Biotechnology* **35**, 4, 316–319 ISSN 1087-0156.

Diab, S., Dostrovsky, N., Hudson, M. *et al.* (2014) Systemic Sclerosis Sine Scleroderma: A Multicenter Study of 1417 Subjects *The Journal of Rheumatology* **41**, 11, 2179–2185 ISSN 0315-162X.

Dimon, M. T., Wood, H. M., Rabbitts, P. H. *et al.* (2013) IMSA: Integrated Metagenomic Sequence Analysis for Identification of Exogenous Reads in a Host Genomic Background *PLoS ONE* **8**, 5, e64546 ISSN 1932-6203.

Dobin, A., Davis, C. A., Schlesinger, F. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner *Bioinformatics* **29**, 1, 15–21 ISSN 1460-2059.

Douvas, A. S., Achten, M. and Tan, E. M. (1979) Identification of a nuclear protein (Scl-70) as a unique target of human antinuclear antibodies in scleroderma *The Journal of Biological Chemistry* **254**, 20, 10514–10522.

Dowson, C., Simpson, N., Duffy, L. *et al.* (2017) Innate Immunity in Systemic Sclerosis *Current Rheumatology Reports* **19**, 1 ISSN 15346307.

Duman, I. E. and Ekinci, G. (2018) Neuroimaging and clinical findings in a case of linear scleroderma en coup de sabre *Radiology Case Reports* **13**, 3, 545–548 ISSN 19300433.

Durinck, S., Spellman, P. T., Birney, E. *et al.* (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt *Nature Protocols* **4**, 8, 1184–1191 ISSN 1754-2189.

Elhai, M., Avouac, J., Kahan, A. *et al.* (2015) Systemic sclerosis: Recent insights *Joint Bone Spine* **82**, 3, 148–153 ISSN 1297319X.

Engström, P. G., Steijger, T., Sipos, B. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data *Nature Methods* **10**, 12, 1185–1191 ISSN 1548-7091.

Ewels, P., Magnusson, M., Lundin, S. *et al.* (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report *Bioinformatics* **32**, 19, 3047–3048 ISSN 1367-4803.

Fan, J., Han, F. and Liu, H. (2014) Challenges of Big Data analysis *National Science Review* **1**, 2, 293–314 ISSN 2053-714X.

Favoino, E., Digiglio, L., Cuomo, G. *et al.* (2013) Autoantibodies Recognizing the Amino Terminal 1-17 Segment of CENP-A Display Unique Specificities in Systemic Sclerosis *PLoS ONE* **8**, 4, e61453 ISSN 1932-6203.

Feghali-Bostwick, C., Medsger, T. A. and Wright, T. M. (2003) Analysis of systemic sclerosis in twins reveals low concordance for disease and high concordance for the presence of antinuclear antibodies *Arthritis & Rheumatism* **48**, 7, 1956–1963 ISSN 0004-3591.

Ferri, F. F. (2016) Diseases and Disorders: Scleroderma (Systemic Sclerosis) in *Ferri's Clinical Advisor 2016 - 5 Books in 1* chapter Section I, 1106–1108e2 Elsevier, Inc., Philadelphia, PA ISBN 978-0-323-28047-1.

Fett, N. (2013) Scleroderma: Nomenclature, etiology, pathogenesis, prognosis, and treatments: Facts and controversies *Clinics in Dermatology* **31**, 4, 432–437 ISSN 0738081X.

Finotello, F. and Di Camillo, B. (2015) Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis *Briefings in Functional Genomics* **14**, 2, 130–142 ISSN 2041-2649.

Fonseca, C., Lindahl, G. E., Ponticos, M. *et al.* (2007) A Polymorphism in the CTGF Promoter Region Associated with Systemic Sclerosis *New England Journal of Medicine* **357**, 12, 1210–1220 ISSN 0028-4793.

Fritzler, M. J., Rattner, J. B., Luft, L. M. *et al.* (2011) Historical perspectives on the discovery and elucidation of autoantibodies to centromere proteins (CENP) and the emerging importance of antibodies to CENP-F *Autoimmunity Reviews* **10**, 4, 194–200 ISSN 15689972.

Frost, J., Estivill, X., Ramsay, M. *et al.* (2018) Dysregulation of the Wnt signaling pathway in South African patients with diffuse systemic sclerosis *Clinical Rheumatology* ISSN 0770-3198.

Frost, J. M. (2016) *The Wnt signalling pathway in systemic sclerosis* (PhD Thesis) University of the Witwatersrand.

Fullard, N. and O'Reilly, S. (2015) Role of innate immune system in systemic sclerosis *Seminars in Immunopathology* **37**, 5, 511–517 ISSN 18632300.

Funke, G., Pagano-Niederer, M., Sjödén, B. *et al.* (1998) Characteristics of Arthrobacter cumminsii, the most frequently encountered Arthrobacter species in human clinical specimens *Journal of Clinical Microbiology* **36**, 6, 1539–1543 ISSN 00951137.

Fuschiotti, P. (2018) T cells and cytokines in systemic sclerosis *Current Opinion in Rheumatology* **30**, 6, 1 ISSN 1040-8711.

Gabrielli, A., Avvedimento, E. V. and Krieg, T. (2009) Scleroderma *New England Journal of Medicine* **360**, 19, 1989–2003 ISSN 0028-4793.

Garber, M., Grabherr, M. G., Guttman, M. *et al.* (2011) Computational methods for transcriptome annotation and quantification using RNA-seq *Nature Methods* **8**, 6, 469–477 ISSN 1548-7091.

Gartzke, J. and Lange, K. (2002) Cellular target of weak magnetic fields: ionic conduction along actin filaments of microvilli *American Journal of Physiology-Cell Physiology* **283**, 5, C1333–C1346 ISSN 0363-6143.

Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource *Nucleic Acids Research* **32**, 90001, 258D–261 ISSN 1362-4962.

Grabherr, M. G., Haas, B. J., Yassour, M. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome *Nature Biotechnology* **29**, 7, 644–652 ISSN 1087-0156.

Grossman, C., Dovrish, Z., Shoenfeld, Y. *et al.* (2011) Do infections facilitate the emergence of systemic sclerosis? *Autoimmunity Reviews* **10**, 5, 244–247 ISSN 15689972.

Guo, Y., Ye, F., Sheng, Q. *et al.* (2014) Three-stage quality control strategies for DNA re-sequencing data *Briefings in Bioinformatics* **15**, 6, 879–889 ISSN 1467-5463.

Haas, B. J., Papanicolaou, A., Yassour, M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis *Nature Protocols* **8**, 8, 1494–1512 ISSN 1754-2189.

Hachulla, E. and Launay, D. (2011) Diagnosis and Classification of Systemic Sclerosis *Clinical Reviews in Allergy & Immunology* **40**, 2, 78–83 ISSN 1080-0549.

Hamaguchi, Y. (2010) Autoantibody profiles in systemic sclerosis: Predictive value for clinical evaluation and prognosis *The Journal of Dermatology* **37**, 1, 42–53 ISSN 03852407.

Hua-Huy, T. and Dinh-Xuan, A. (2015) Cellular and molecular mechanisms in the pathophysiology of systemic sclerosis *Pathologie Biologie* **63**, 2, 61–68 ISSN 03698114.

Hudson, M., Mahler, M., Pope, J. *et al.* (2012) Clinical correlates of CENP-A and CENP-B antibodies in a large cohort of patients with systemic sclerosis. *The Journal of rheumatology* **39**, 4, 787–94 ISSN 0315-162X.

Huttenhower, C., Gevers, D., Knight, R. *et al.* (2012) Structure, function and diversity of the healthy human microbiome *Nature* **486**, 7402, 207–214 ISSN 0028-0836.

Ihnatova, I. and Budinska, E. (2015) ToPASeq: an R package for topology-based pathway analysis of microarray and RNA-Seq data *BMC Bioinformatics* **16**, 1, 350 ISSN 1471-2105.

Jiménez-Marín, Á., Collado-Romero, M., Ramirez-Boo, M. *et al.* (2009) Biological pathway analysis by ArrayUnlock and Ingenuity Pathway Analysis *BMC Proceedings* **3**, Suppl 4, S6 ISSN 1753-6561.

Kalogerou, A. (2005) Early T cell activation in the skin from patients with systemic sclerosis *Annals of the Rheumatic Diseases* **64**, 8, 1233–1235 ISSN 0003-4967.

Katsumoto, T. R., Whitfield, M. L. and Connolly, M. K. (2011) The Pathogenesis of Systemic Sclerosis *Annual Review of Pathology: Mechanisms of Disease* **6**, 1, 509–537 ISSN 1553-4006.

Kavian, N. and Batteux, F. (2015) Macro- and microvascular disease in systemic sclerosis *Vascular Pharmacology* **71**, 16–23 ISSN 15371891.

Kayser, C. and Fritzler, M. J. (2015) Autoantibodies in Systemic Sclerosis: Unanswered Questions *Frontiers in Immunology* **6**, MAR, 2–7 ISSN 1664-3224.

Khatri, P., Sirota, M. and Butte, A. J. (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges *PLoS Computational Biology* **8**, 2, e1002375 ISSN 1553-7358.

Kluge, M. and Friedel, C. C. (2018) Watchdog a workflow management system for the distributed analysis of large-scale experimental data *BMC Bioinformatics* **19**, 1, 97 ISSN 1471-2105.

Koenig, M., Joyal, F., Fritzler, M. J. *et al.* (2008) Autoantibodies and microvascular damage are independent predictive factors for the progression of Raynaud's phenomenon to systemic sclerosis: A twenty-year prospective study of 586 patients, with validation of proposed criteria for early systemic sclerosi *Arthritis & Rheumatism* **58**, 12, 3902–3912 ISSN 00043591.

Koerner, R. J., Goodfellow, M. and Jones, A. L. (2009) The genus Dietzia: A new home for some known and emerging opportunist pathogens *FEMS Immunology and Medical Microbiology* **55**, 3, 296–305 ISSN 09288244.

Korman, B. D., Kastner, D. L., Gregersen, P. K. *et al.* (2008) STAT4: Genetics, mechanisms, and implications for autoimmunity *Current Allergy and Asthma Reports* **8**, 5, 398–403 ISSN 1529-7322.

Koscielny, G., An, P., Carvalho-Silva, D. *et al.* (2017) Open Targets: a platform for therapeutic target identification and validation *Nucleic Acids Research* **45**, D1, D985–D994 ISSN 0305-1048.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update *Nucleic Acids Research* **44**, W1, W90–W97 ISSN 0305-1048.

Kurtzer, G. M., Sochat, V. and Bauer, M. W. (2017) Singularity: Scientific containers for mobility of compute *PLOS ONE* **12**, 5, e0177459 ISSN 1932-6203.

Kuwana, M., Kaburaki, J., Mimori, T. *et al.* (1993) Autoantibody reactive with three classes of RNA polymerases in sera from patients with systemic sclerosis. *Journal of Clinical Investigation* **91**, 4, 1399–1404 ISSN 0021-9738.

Laurent, P., Sisirak, V., Lazaro, E. *et al.* (2018) Innate Immunity in Systemic Sclerosis Fibrosis: Recent Advances *Frontiers in Immunology* **9**, JUL ISSN 1664-3224.

LeRoy, C. E., Black, C., Fleischmajer, R. *et al.* (1988) Scleroderma (systemic sclerosis): Classification, subsets and pathogenesis *Journal of Rheumatology* **15**, 2, 202–205 ISSN 0315162X.

Lex, A., Gehlenborg, N., Strobelt, H. *et al.* (2014) UpSet: Visualization of Intersecting Sets *IEEE Transactions on Visualization and Computer Graphics* **20**, 12, 1983–1992 ISSN 1077-2626.

Li, B. and Dewey, C. N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome *BMC Bioinformatics* **12**, 1, 323 ISSN 1471-2105.

Li, B., Fillmore, N., Bai, Y. *et al.* (2014) Evaluation of de novo transcriptome assemblies from RNA-Seq data *Genome Biology* **15**, 12, 553 ISSN 1474-760X.

Li, H., Yang, R., Fan, X. *et al.* (2012) MicroRNA array analysis of microRNAs related to systemic scleroderma *Rheumatology International* **32**, 2, 307–313 ISSN 0172-8172.

Liao, Y., Smyth, G. K. and Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features *Bioinformatics* **30**, 7, 923–930 ISSN 1367-4803.

Lischwe, M. A., Ochs, R. L., Reddy, R. *et al.* (1985) Purification and partial characterization of a nucleolar scleroderma antigen (Mr = 34,000; pI, 8.5) rich in NG,NG-dimethylarginine *Journal of Biological Chemistry* **260**, 26, 14304–14310.

Love, M. I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 *Genome Biology* **15**, 12, 550 ISSN 1474-760X.

Luo, W. and Brouwer, C. (2013) Pathview: An R/Bioconductor package for pathway-based data integration and visualization *Bioinformatics* **29**, 14, 1830–1831 ISSN 13674803.

Luo, W., Friedman, M. S., Shedden, K. *et al.* (2009) GAGE: generally applicable gene set enrichment for pathway analysis *BMC Bioinformatics* **10**, 1, 161 ISSN 1471-2105.

Mages, I. S., Frodl, R., Bernard, K. A. *et al.* (2008) Identities of Arthrobacter spp. and Arthrobacter-like bacteria encountered in human clinical specimens *Journal of Clinical Microbiology* **46**, 9, 2980–2986 ISSN 00951137.

Mahler, M., Silverman, E. D., Schulte-Pelkum, J. *et al.* (2010) Anti-Scl-70 (topo-I) antibodies in SLE: Myth or reality? *Autoimmunity Reviews* **9**, 11, 756–760 ISSN 15689972.

Marangoni, R. G., Rocha, L. F., Del Rio, A. P. T. *et al.* (2013) Systemic sclerosis sine scleroderma: distinct features in a large Brazilian cohort *Rheumatology* **52**, 8, 1520–1524 ISSN 1462-0324.

Marie, I. and Gehanno, J.-F. (2015) Environmental risk factors of systemic sclerosis *Seminars in Immunopathology* **37**, 5, 463–473 ISSN 1863-2297.

Marriott, D., Chan, S.-Y., Kunkel, S. L. *et al.* (2002) The Functional Role of the ELR Motif in CXC Chemokine-mediated Angiogenesis *Journal of Biological Chemistry* **270**, 45, 27348–27357 ISSN 0021-9258.

Martel, A., Pasmans, F., Hellebuyck, T. *et al.* (2008) Devriesea agamarum gen. nov., sp. nov., a novel actinobacterium associated with dermatitis and septicaemia in agamid lizards *International Journal of Systematic and Evolutionary Microbiology* **58**, 9, 2206–2209 ISSN 14665026.

Martín, R., Miquel, S., Langella, P. *et al.* (2014) The role of metagenomics in understanding the human microbiome in health and disease *Virulence* **5**, 3, 413–423 ISSN 2150-5594.

Masi, A. T. (1980) Preliminary criteria for the classification of systemic sclerosis (scleroderma) *Arthritis & Rheumatism* **23**, 5, 581–590 ISSN 00043591.

Maurer, B., Stanczyk, J., Jüngel, A. *et al.* (2010) MicroRNA-29, a key regulator of collagen expression in systemic sclerosis *Arthritis & Rheumatism* **62**, 6, 1733–1743 ISSN 00043591.

McMahan, Z. H. and Wigley, F. M. (2010) Raynaud's phenomenon and digital ischemia: a practical approach to risk stratification, diagnosis and management *International Journal of Clinical Rheumatology* **5**, 3, 355–370 ISSN 1758-4272.

Mehra, S., Walker, J., Patterson, K. *et al.* (2013) Autoantibodies in systemic sclerosis *Autoimmunity Reviews* **12**, 3, 340–354 ISSN 15689972.

Meier, S., Erickson, A. R., Mikuls, T. R. *et al.* (2013) Connective Tissue Diseases in Mikuls, T. R., Cannella, A. C., Moore, G. F. *et al.*, editors, *Rheumatology: A Color Handbook* chapter Chapter 5, 113–150 Taylor & Francis Group, LLC, Boca Raton, FL ISBN 13: 978-1-84076-634-9 (eBook - PDF).

Mele, M., Ferreira, P. G., Reverter, F. *et al.* (2015) The human transcriptome across tissues and individuals *Science* **348**, 6235, 660–665 ISSN 0036-8075.

Methé, B. A., Nelson, K. E., Pop, M. *et al.* (2012) A framework for human microbiome research *Nature* **486**, 7402, 215–221 ISSN 0028-0836.

Mortazavi, A., Williams, B. A., McCue, K. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq *Nature Methods* **5**, 7, 621–628 ISSN 1548-7091.

Muir, P., Li, S., Lou, S. *et al.* (2016) The real cost of sequencing: Scaling computation to keep pace with data generation *Genome Biology* **17**, 1, 1–9 ISSN 1474760X.

Murdaca, G., Contatore, M., Gulli, R. *et al.* (2016) Genetic factors and systemic sclerosis *Autoimmunity Reviews* **15**, 5, 427–432 ISSN 15689972.

Noble, K. E., Wickremasinghe, R. G., DeCornet, C. *et al.* (1999) Monocytes stimulate expression of the Bcl-2 family member, A1, in endothelial cells and confer protection against apoptosis *JImmunol* **162**, 0022-1767 (Print), 1376–1383 ISSN 0022-1767.

Okano, Y., Steen, V. D. and Medsger, T. A. (1992) Autoantibody to U3 Nucleolar Ribonucleoprotein (Fibrillarin) In Patients with Systemic Sclerosis *Arthritis & Rheumatism* **35**, 1, 95–100 ISSN 00043591.

O'Reilly, S. (2014) Innate immunity in systemic sclerosis pathogenesis *Clinical Science* **126**, 5, 329–337 ISSN 0143-5221.

Pattanaik, D., Brown, M., Postlethwaite, B. C. *et al.* (2015) Pathogenesis of systemic sclerosis *Frontiers in Immunology* **6**, 272, 755–759 ISSN 0315162X.

Perkel, J. (2016) Democratic databases: science on GitHub *Nature* **538**, 7623, 127–128 ISSN 0028-0836.

Perosa, F., Prete, M., Di Lernia, G. *et al.* (2016) Anti-centromere protein A antibodies in systemic sclerosis: Significance and origin *Autoimmunity Reviews* **15**, 1, 102–109 ISSN 15689972.

Piccolo, S. R. and Frampton, M. B. (2016) Tools and techniques for computational reproducibility *GigaScience* **5**, 1, 30 ISSN 2047-217X.

Poretsky, R., Rodriguez-R, L. M., Luo, C. *et al.* (2014) Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics *PLoS ONE* **9**, 4, e93827 ISSN 1932-6203.

Pudifin, D. J., Dinnematin, H. and Duursma, J. (1991) Antinuclear antibodies in systemic sclerosis. Clinical and ethnic associations *South African Medical Journal* **80**, 9, 438–440 ISSN 02569574.

Rabquer, B. J. and Koch, A. E. (2012) Angiogenesis and Vasculopathy in Systemic Sclerosis: Evolving Concepts *Current Rheumatology Reports* **14**, 1, 56–63 ISSN 1523-3774.

Radić, M. (2014) Role of Helicobacter pylori infection in autoimmune systemic rheumatic diseases *World Journal of Gastroenterology* **20**, 36, 12839 ISSN 1007-9327.

Radstake, T. R., van Bon, L., Broen, J. *et al.* (2009) Increased frequency and compromised function of T regulatory cells in systemic sclerosis (SSc) is related to a diminished CD69 and TGFβ expression *PLoS ONE* **4**, 6 ISSN 19326203.

Radstake, T. R. D. J., Gorlova, O., Rueda, B. *et al.* (2010) Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus *Nature Genetics* **42**, 5, 426–429 ISSN 1061-4036.

Rapaport, F., Khanin, R., Liang, Y. *et al.* (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data *Genome Biology* **14**, 9, R95 ISSN 1465-6906.

Rawat, A., Engelthaler, D. M., Driebe, E. M. *et al.* (2014) MetaGeniE: Characterizing Human Clinical Samples Using Deep Metagenomic Sequencing *PLoS ONE* **9**, 11, e110915 ISSN 1932-6203.

Reimer, G., Rose, K. M., Scheer, U. *et al.* (1987) Autoantibody to RNA polymerase I in scleroderma sera. *Journal of Clinical Investigation* **79**, 1, 65–72 ISSN 0021-9738.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data *Bioinformatics* **26**, 1, 139–140 ISSN 1367-4803.

Robinson, M. D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* **11**, 3, R25 ISSN 1465-6906.

Sandve, G. K., Nekrutenko, A., Taylor, J. *et al.* (2013) Ten Simple Rules for Reproducible Computational Research *PLoS Computational Biology* **9**, 10, e1003285 ISSN 1553-7358.

Schulz, W., Durant, T., Siddon, A. *et al.* (2016) Use of application containers and workflows for genomic data analysis *Journal of Pathology Informatics* **7**, 1, 53 ISSN 2153-3539.

Senécal, J.-L., Hénault, J. and Raymond, Y. (2005) The pathogenic role of autoantibodies to nuclear autoantigens in systemic sclerosis (scleroderma). *The Journal of rheumatology* **32**, 9, 1643–9 ISSN 0315-162X.

Shero, J., Bordwell, B., Rothfield, N. *et al.* (1986) High titers of autoantibodies to topoisomerase I (Scl-70) in sera from scleroderma patients *Science* **231**, 4739, 737–740 ISSN 0036-8075.

Silber, W. (1983) The prevalence, course and management of some benign oesophageal diseases in the Black population. The Groote Schuur Hospital experience. *South African Medical Journal* **63**, 957–959 ISSN 00382469.

Silver, K. C. and Silver, R. M. (2015) Management of Systemic-Sclerosis-Associated Interstitial Lung Disease *Rheumatic Disease Clinics of North America* **41**, 3, 439–457 ISSN 0889857X.

Sobanski, V., Dauchet, L., Lefèvre, G. *et al.* (2014) Prevalence of Anti-RNA Polymerase III Antibodies in Systemic Sclerosis: New Data From a French Cohort and a Systematic Review and Meta-Analysis *Arthritis & Rheumatology* **66**, 2, 407–417 ISSN 23265191.

Steen, V. D. (2005) Autoantibodies in Systemic Sclerosis *Seminars in Arthritis and Rheumatism* **35**, 1, 35–42 ISSN 00490172.

Steinbiss, S., Silva-Franco, F., Brunk, B. *et al.* (2016) Companion : a web server for annotation and analysis of parasite genomes *Nucleic Acids Research* **44**, W1, W29–W34 ISSN 0305-1048.

Stern, E. P. and Denton, C. P. (2015) The Pathogenesis of Systemic Sclerosis *Rheumatic Disease Clinics of North America* **41**, 3, 367–382 ISSN 0889857X.

Stetler, D. A., Reichlin, M., Berlinis, C. M. *et al.* (1987) Autoantibodies against RNA polymerase I in scleroderma and Sjögren's syndrome scera *Biomedical and Biophysical Research Communications* **144**, 3, 1296–1302.

Stetler, D. A., Rose, K. M., Wenger, M. E. *et al.* (1982) Antibodies to distinct polypeptides of RNA polymerase I in sera from patients with rheumatic autoimmune disease. *Proceedings of the National Academy of Sciences* **79**, 23, 7499–7503 ISSN 0027-8424.

Strange, G. and Nash, P. (2009) The manifestations of vasculopathy in systemic sclerosis and its evidence-based therapy. *International journal of rheumatic diseases* **12**, 3, 192–206 ISSN 1756-185X.

Tager, R. E. and Tikly, M. (1999) Clinical and laboratory manifestations of systemic sclerosis (scleroderma) in Black South Africans. *Rheumatology* **38**, 5, 397–400 ISSN 1462-0324.

Tamai, K., Akashi, Y., Yoshimoto, Y. *et al.* (2018) First case of a bloodstream infection caused by the genus Brachybacterium *Journal of Infection and Chemotherapy* **24**, 12, 998–1003 ISSN 14377780.

Tao, J., Li, L., Tan, Z. *et al.* (2011) Up-regulation of CC chemokine ligand 20 and its receptor CCR6 in the lesional skin of early systemic sclerosis *European Journal of Dermatology* **21**, 5, 731–736 ISSN 11671122.

Tejera Segura, B. and Ferraz-Amaro, I. (2015) [Large vessels vasculopathy in systemic sclerosis]. *Medicina clinica* **145**, 11, 488–92 ISSN 1578-8989.

Trapnell, C., Hendrickson, D. G., Sauvageau, M. *et al.* (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq *Nature Biotechnology* **31**, 1, 46–53 ISSN 1087-0156.

Trapnell, C., Roberts, A., Goff, L. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks *Nature Protocols* **7**, 3, 562–578 ISSN 1754-2189.

Trapnell, C., Williams, B. A., Pertea, G. *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 ISSN 1087-0156.

Tsou, P. S., Rabquer, B. J., Ohara, R. A. *et al.* (2016) Scleroderma dermal microvascular endothelial cells exhibit defective response to pro-angiogenic chemokines *Rheumatology (United Kingdom)* **55**, 4, 745–754 ISSN 14620332.

Tsuchiya, N., Kawasaki, A., Hasegawa, M. *et al.* (2009) Association of STAT4 polymorphism with systemic sclerosis in a Japanese population *Annals of the Rheumatic Diseases* **68**, 8, 1375–1376 ISSN 0003-4967.

van Bon, L., Affandi, A. J., Broen, J. *et al.* (2014) Proteome-wide Analysis and CXCL4 as a Biomarker in Systemic Sclerosis *New England Journal of Medicine* **370**, 5, 433–443 ISSN 0028-4793.

van den Hoogen, F., Khanna, D., Fransen, J. *et al.* (2013) 2013 Classification Criteria for Systemic Sclerosis: An American College of Rheumatology/European League Against Rheumatism Collaborative Initiative *Arthritis & Rheumatism* **65**, 11, 2737–2747 ISSN 00043591.

Van Eenennaam, H., Vogelzangs, J. H. P., Lugtenberg, D. *et al.* (2002) Identity of the RNase MRP- and RNase P-associated Th/To autoantigen *Arthritis & Rheumatism* **46**, 12, 3266–3272 ISSN 0004-3591.

van Laar, J. M., Farge, D., Sont, J. K. *et al.* (2014) Autologous Hematopoietic Stem Cell Transplantation vs Intravenous Pulse Cyclophosphamide in Diffuse Cutaneous Systemic Sclerosis *JAMA* **311**, 24, 2490 ISSN 0098-7484.

Varga, J. (2017) Etiology and Pathogenesis of Scleroderma in Firestein, G. S., Budd, R. C., Gabriel, S. E. *et al.*, editors, *Kelley and Firestein's Textbook of Rheumatology* chapter Chapter 83, 1400 – 1423.e3 Elsevier tenth edit edition ISBN 978-0-323-31696-5.

Varga, J. A. and Trojanowska, M. (2008) Fibrosis in Systemic Sclerosis *Rheumatic Disease Clinics of North America* **34**, 1, 115–143 ISSN 0889857X.

Villalta, D., Imbastaro, T., Di Giovanni, S. *et al.* (2012) Diagnostic accuracy and predictive value of extended autoantibody profile in systemic sclerosis *Autoimmunity Reviews* **12**, 2, 114–120 ISSN 15689972.

Visconti, A., Martin, T. C. and Falchi, M. (2018) YAMP: a containerized workflow enabling reproducibility in metagenomics research *GigaScience* **7**, 7, 1–9 ISSN 2047-217X.

Vogler, M. (2012) BCL2A1: The underdog in the BCL2 family *Cell Death and Differentiation* **19**, 1, 67–74 ISSN 13509047.

Wang, J. C. (2002) Cellular roles of DNA topoisomerases: a molecular perspective *Nature Reviews Molecular Cell Biology* **3**, 6, 430–440 ISSN 1471-0072.

Wang, L., You, M., Zhao, N. *et al.* (2005) Arthrobacter scleromae sp. nov. Isolated from Human Clinical Specimens *Journal of Clinical Microbiology* **43**, 3, 1451–1455 ISSN 0095-1137.

Wielosz, E., Dryglewska, M. and Majdan, M. (2014) Serological profile of patients with systemic sclerosis. *Postepy higieny i medycyny doswiadczalnej (Online)* **68**, 987–91 ISSN 1732-2693.

Wilson, G., Aruliah, D. A., Brown, C. T. *et al.* (2014) Best Practices for Scientific Computing *PLoS Biology* **12**, 1, e1001745 ISSN 1545-7885.

Wood, D. E. and Salzberg, S. L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments *Genome Biology* **15**, 3, R46 ISSN 1465-6906.

Wu, H., Wang, C. and Wu, Z. (2015) PROPER: comprehensive power evaluation for differential expression using RNA-seq *Bioinformatics* **31**, 2, 233–241 ISSN 1460-2059.

Yousefi, B., Mahmoudi, M., Sarafnejad, A. *et al.* (2017) Downregulation of Aquaporin3 in Systemic Sclerosis Dermal Fibroblasts. *Iranian journal of allergy, asthma, and immunology* **16**, 3, 228–234 ISSN 1735-1502.

Zhang, C., Cleveland, K., Schnoll-Sussman, F. *et al.* (2015) Identification of low abundance microbiome in clinical samples using whole genome sequencing *Genome Biology* **16**, 1, 265 ISSN 1474-760X.

Zhou, X., Lee, J. E., Arnett, F. C. *et al.* (2009) HLAâ€“DPB1 and DPB2 are genetic loci for systemic sclerosis: A genome-wide association study in Koreans with replication in North Americans *Arthritis & Rheumatism* **60**, 12, 3807–3814 ISSN 00043591.

Zhou, X., Tan, F. K., Xiong, M. *et al.* (2005) Monozygotic twins clinically discordant for scleroderma show concordance for fibroblast gene expression profiles *Arthritis & Rheumatism* **52**, 10, 3305–3314 ISSN 0004-3591.

Zhu, H., Luo, H., Li, Y. *et al.* (2013) MicroRNA-21 in Scleroderma Fibrosis and its Function in TGF-β- Regulated Fibrosis-Related Genes Expression *Journal of Clinical Immunology* **33**, 6, 1100–1109 ISSN 0271-9142.

Appendices

Appendix A:
Ethics clearance certificate

R14/49 Mr Phelelani Thokozani Mpangase et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M1710104

NAME: (Principal Investigator)	Mr Phelelani Thokozani Mpangase et al
DEPARTMENT:	Human Genetics University of the Witwatersrand Sydney Brenner Institute for Molecular Bioscience
PROJECT TITLE:	Bioinformatic Analyses of Transcriptone and Genetic Variation in Black South Africans with Systemic Sclerosis
DATE CONSIDERED:	Adhoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-Study under Primary Study (M120512)
SUPERVISOR:	Prof Michelle Ramsay and Prof Scott Hazelhurst
APPROVED BY:	Professor P. Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	06/12/2017

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

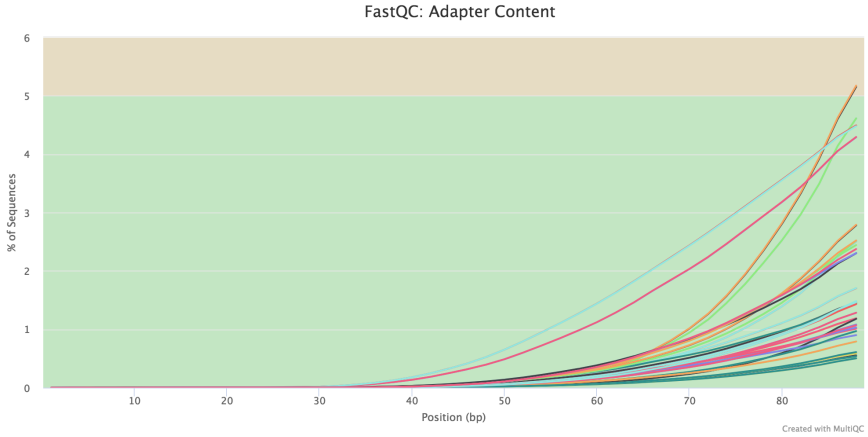
To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand. I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. in this case, the study was initially review November and will therefore be due in the month of November each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

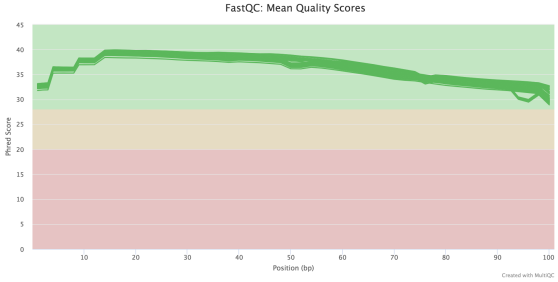
Date 07 DECEMBER 2017

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

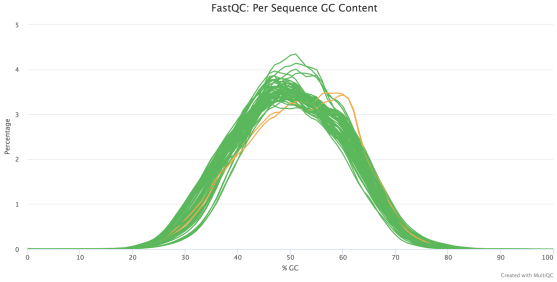
Appendix B:
QC plots for RNA-seq Data Pre-processing



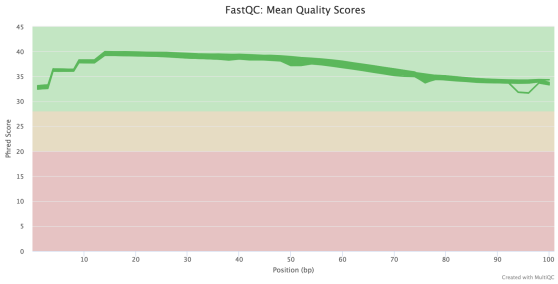
(a) Technical sequence contamination in the initial the RNA-seq data.



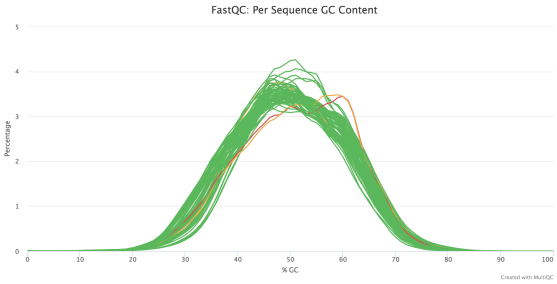
(b) Read quality for the initial RNA-seq data.



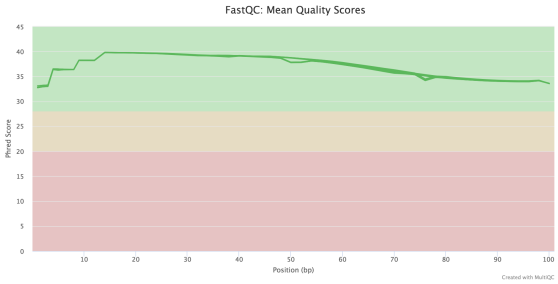
(c) GC content of the original RNA-seq data.



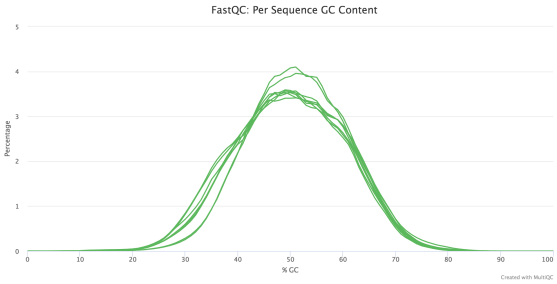
(d) Read quality after trimming.



(e) GC content of the RNA-seq data after trimming.



(f) Read quality for merged duplicate samples.



(g) GC content for merged duplicate samples.

Figure B.1: QC plots for the RNA-seq data used in this study. This figure shows the QC plots for the RNA-seq data before trimming (a-c), after trimming (d-e) and after merging the duplicated samples (f-g). The plots were generated using `MultiQC` (Version 1.5)

Appendix C: Singularity Container Recipes for the rnaSeqCount Workflow

C.1 STAR container

```
Bootstrap:shub
From:singularityhub/ubuntu

%labels
Maintainer Phelelani.Mpangase@wits.ac.za

%post
#### Updates and essentials
apt-get update
apt-get install -y build-essential
apt-get install -y wget git zlib1g-dev unzip

## Install STAR Aligner
## From Source: https://github.com/alexdobin/STAR
cd /opt \
    && wget https://github.com/alexdobin/STAR/archive/2.5.3a.tar.gz \
    && tar -vxf 2.5.3a.tar.gz \
    && make STAR -C STAR-2.5.3a/source \
    && rm /opt/2.5.3a.tar.gz

%environment
## Add paths to environment
export PATH=/opt/STAR-2.5.3a/source:$PATH
```

C.2 htseq-count container

```
Bootstrap:shub
From:singularityhub/ubuntu

%labels
Maintainer Phelelani.Mpangase@wits.ac.za

%post
## Updates and essentials
apt-get update
apt-get install -y build-essential
apt-get install -y wget python python-pip

## Install HTSeq using PIP
pip install -U pip
pip install HTSeq

%environment
export PYTHONPATH=/usr/local/lib/python2.7/dist-packages
```

C.3 featureCounts container

```
Bootstrap:shub
From:singularityhub/ubuntu

%labels
Maintainer Phelelani.Mpangase@wits.ac.za

%post
## Updates and essentials
apt-get update
apt-get install -y build-essential
apt-get install -y wget

## Install Subread (featureCounts)
```

```
cd /opt \
    && wget --no-check-certificate -c https://sourceforge.net/projects/subread/files/
    ↪ subread-1.6.0/subread-1.6.0-Linux-x86_64.tar.gz \
    && tar -vxf subread-1.6.0-Linux-x86_64.tar.gz \
    && rm /opt/subread-1.6.0-Linux-x86_64.tar.gz

%environment
## Add paths to environment
export PATH=/opt/subread-1.6.0-Linux-x86_64/bin:$PATH
```

C.4 MultiQC container

```
Bootstrap:shub
From:singularityhub/ubuntu

%labels
Maintainer Phelelani.Mpangase@wits.ac.za

%post
## Updates and essentials
apt-get update
apt-get install -y build-essential
apt-get install -y wget git python python-dev python-pip

## Install HTSeq using PIP
pip install -U pip
pip install multiqc

%environment
export PYTHONPATH=/usr/local/lib/python2.7/dist-packages
```

Appendix D:
Genes Discarded in the Differential Expression Analysis

Table D.1: **Genes excluded from differential expression analysis.**

Ensembl ID	Description	Biotype	Chrom	HGNC
Gene on the Y chromosome:				
ENSG00000012817	lysine demethylase 5D	protein coding	Y	KDM5D
ENSG00000067048	DEAD-box helicase 3 Y-linked	protein coding	Y	DDX3Y
ENSG00000067646	zinc finger protein Y-linked	protein coding	Y	ZFY
ENSG00000099725	protein kinase Y-linked (pseudogene)	transcribed unprocessed pseudogene	Y	PRKY
ENSG00000114374	ubiquitin specific peptidase 9 Y-linked	protein coding	Y	USP9Y
ENSG00000129824	ribosomal protein S4 Y-linked 1	protein coding	Y	RPS4Y1
ENSG00000131002	taxilin gamma pseudogene, Y-linked	transcribed unprocessed pseudogene	Y	TXLNGY
ENSG00000154620	thymosin beta 4 Y-linked	protein coding	Y	TMSB4Y
ENSG00000165246	neuroligin 4 Y-linked	protein coding	Y	NLGN4Y
ENSG00000176728	testis-specific transcript, Y-linked 14	lincRNA	Y	TTTY14
ENSG00000183878	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	protein coding	Y	UTY
ENSG00000198692	eukaryotic translation initiation factor 1A Y-linked	protein coding	Y	EIF1AY
ENSG00000215414	proteasome subunit alpha 6 pseudogene 1	processed pseudogene	Y	PSMA6P1
ENSG00000233864	testis-specific transcript, Y-linked 15	lincRNA	Y	TTTY15
ENSG00000235001	eukaryotic translation initiation factor 4A1 pseudogene 2	processed pseudogene	Y	EIF4A1P2
ENSG00000235649	matrix remodeling associated 5 Y-linked (pseudogene)	unprocessed pseudogene	Y	MXRA5Y
ENSG00000237917	poly(ADP-ribose) polymerase family member 4 pseudogene 1	unprocessed pseudogene	Y	PARP4P1
ENSG00000260197	novel transcript	lincRNA	Y	
ENSG00000278212	MAFF interacting protein (pseudogene)	unprocessed pseudogene	Y	
ENSG00000278847	novel transcript	lincRNA	Y	
ENSG00000279274	ribosomal protein L41 (RPL41) pseudogene	processed pseudogene	Y	
Genes significantly expressed between controls:				
ENSG00000113196	heart and neural crest derivatives expressed 1	protein coding	5	HAND1
ENSG00000128714	homeobox D13	protein coding	2	HOXD13
ENSG00000217236	Sp9 transcription factor	protein coding	2	SP9
ENSG00000250654		transcribed unprocessed pseudogene	12	

Appendix E:
Differential Expression Analysis - PCA Plots

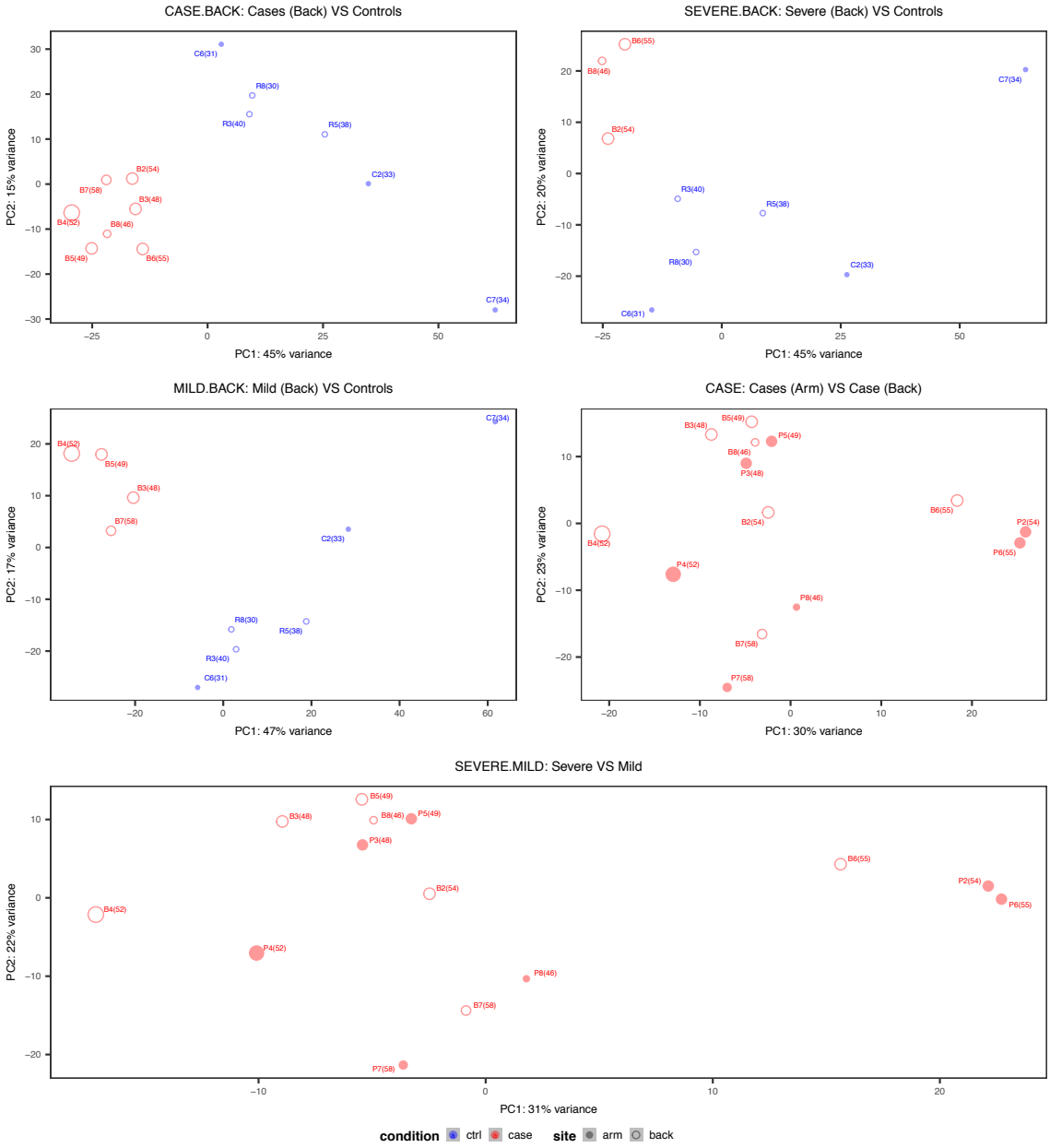


Figure E.1: PCA of the within individual comparison sets and affected backs included. ***Red:*** affected individuals; ***blue:*** control individuals; ***open circle:*** back sample (or breast in the case of controls); ***solid circle:*** forearm sample; ***numbers in brackets:*** age of the individuals.

Appendix F: Differential Expression Analysis - Volcano Plots

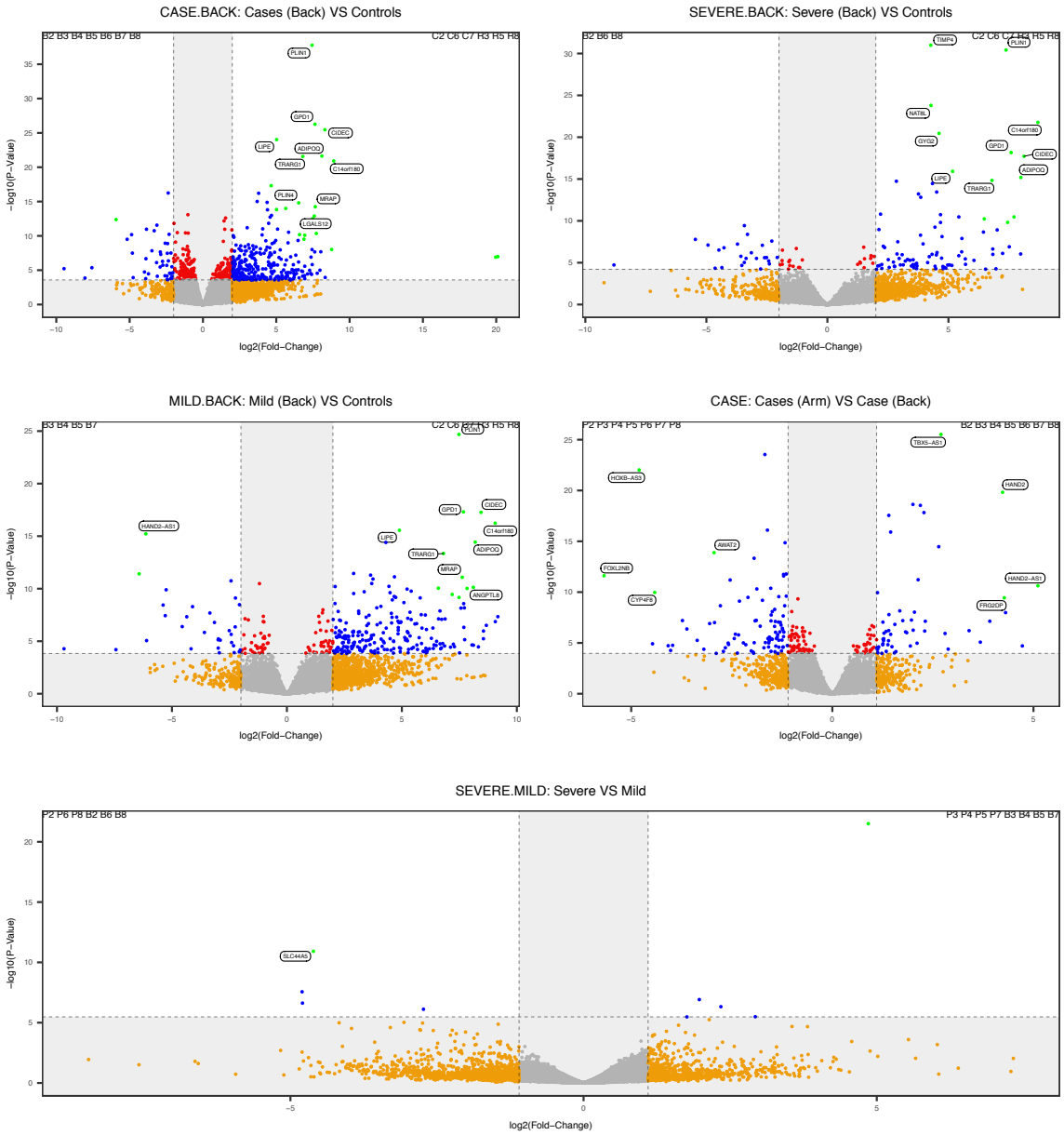


Figure F.1: Volcano plots for the differential expression results and filtering for comparisons sets with backs included. ***Green:*** significant genes after filtering; ***gray:*** genes that did not meet the $|\log_2 FC|$ and significance threshold; ***red:*** genes that met the significance threshold only; ***orange:*** genes that met the $|\log_2 FC|$ threshold only; and ***blue:*** genes that met both $|\log_2 FC|$ and significance thresholds. The names of the top ten significant genes are shown and the samples being compared (cases and controls) in each comparison are listed on the top corners of the plots.

Appendix G:
Differential Expression Analysis - Heatmaps

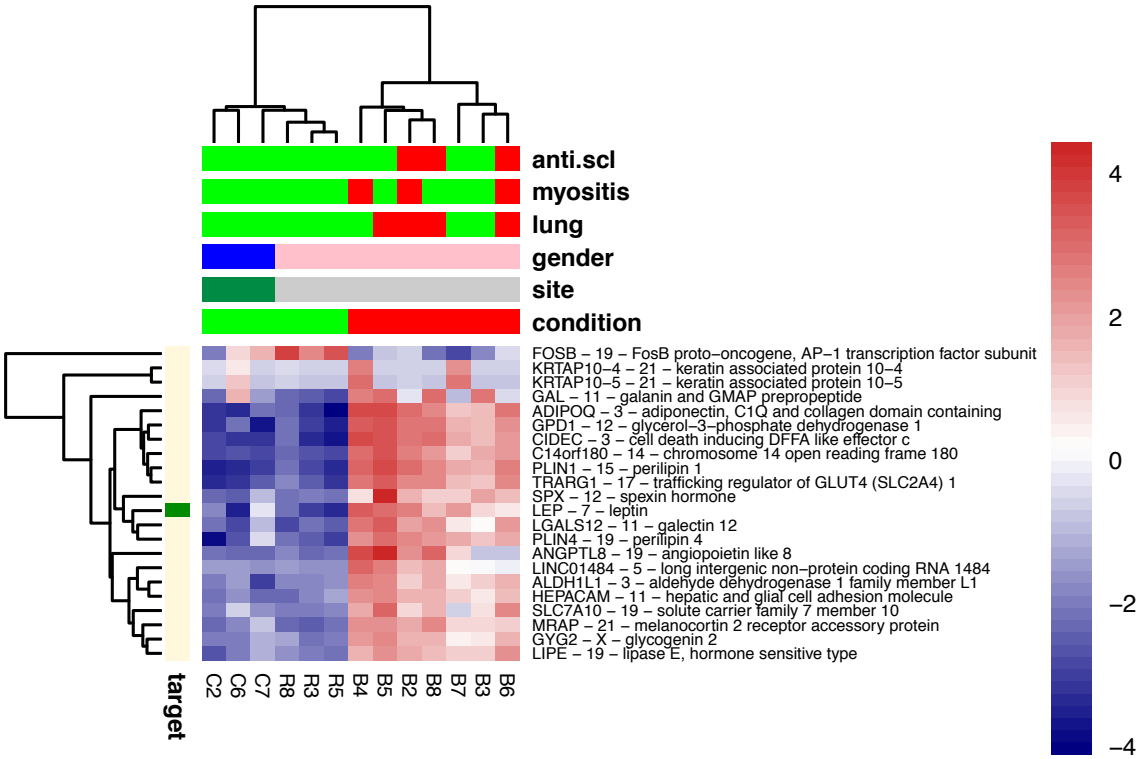


Figure G.1: Heatmap showing gene clustering according to gene expression signals in the CASE.BACK comparison. The color for each cell in the matrix represents gene expression signal from down-regulated (navy), to no change (white), to up-regulated (red). ***condition:*** red = case, green = control; ***site:*** green = forearm, gray = back/breast; ***gender:*** blue = male, pink = female; ***lung:*** red = evidence of lung fibrosis, green = no lung fibrosis; ***myositis:*** red = muscle weakness/inflammation present, green = no muscle weakness/inflammation; ***anti.scl:*** red = positive anti-Scl-70 test, green = negative anti-Scl-70; ***target:*** green = gene on the Open Targets Platform for SSc, yellow = gene not found on the Open Targets Platform for SSc.

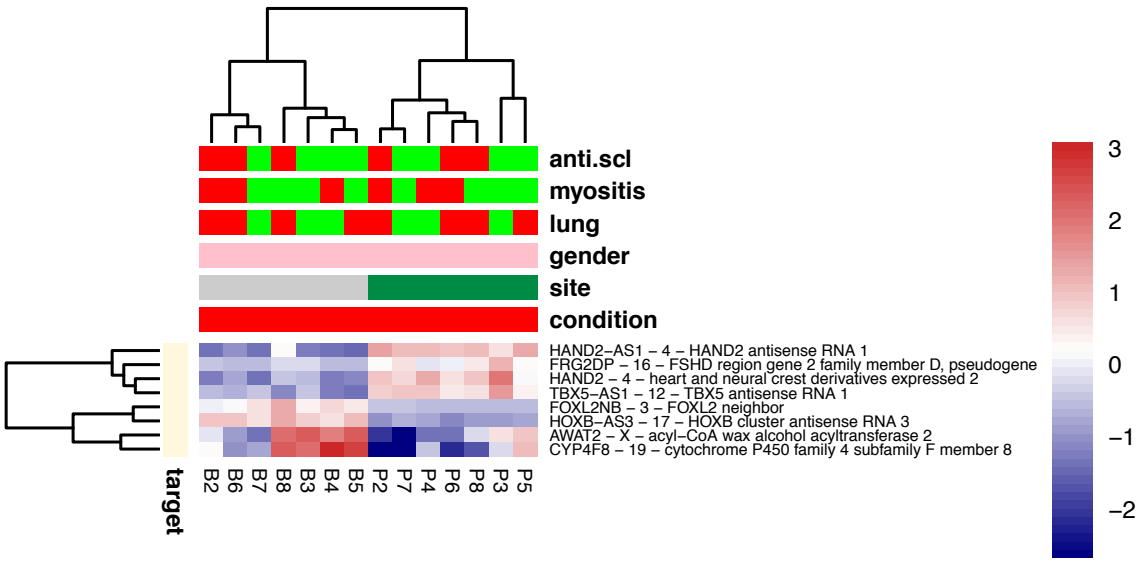


Figure G.2: Heatmap showing gene clustering according to gene expression signals in the CASE comparison. The color for each cell in the matrix represents gene expression signal from down-regulated (navy), to no change (white), to up-regulated (red). ***condition:*** red = case, green = control; ***site:*** green = forearm, gray = back/breast; ***gender:*** blue = male, pink = female; ***lung:*** red = evidence of lung fibrosis, green = no lung fibrosis; ***myositis:*** red = muscle weakness/inflammation present, green = no muscle weakness/inflammation; ***anti.scl:*** red = positive anti-Scl-70 test, green = negative anti-Scl-70; ***target:*** green = gene on the Open Targets Platform for SSc, yellow = gene not found on the Open Targets Platform for SSc.

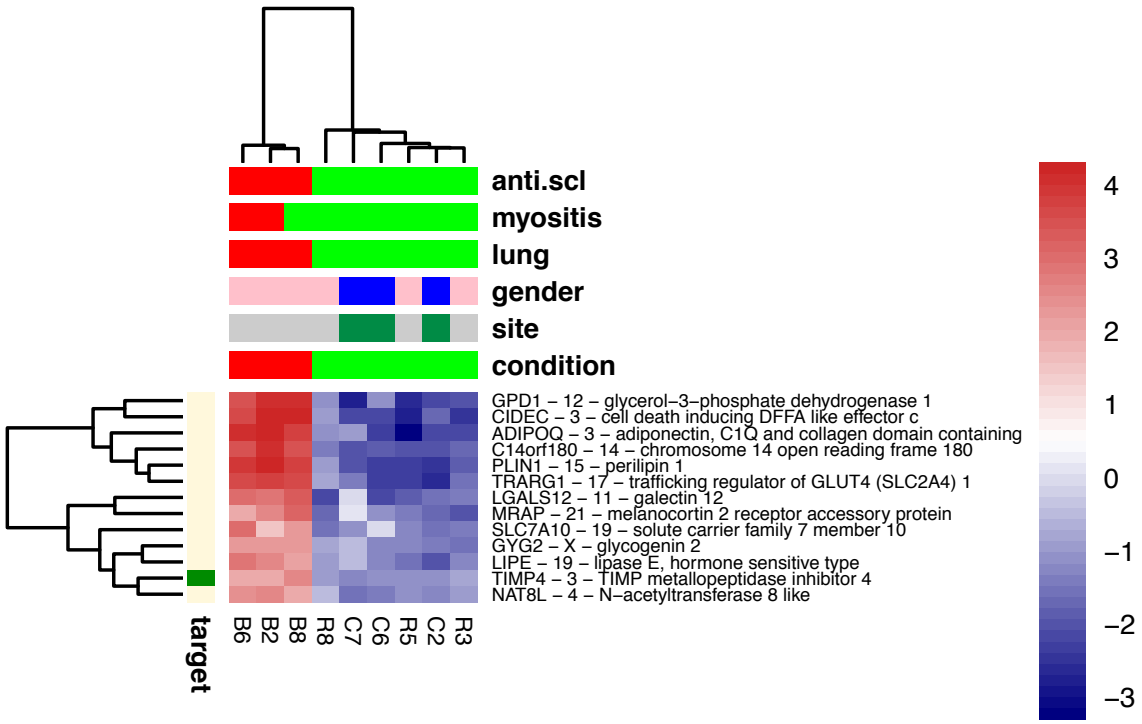


Figure G.3: Heatmap showing gene clustering according to gene expression signals in the SEVERE.BACK comparison. The color for each cell in the matrix represents gene expression signal from down-regulated (navy), to no change (white), to up-regulated (red). ***condition:*** red = case, green = control; ***site:*** green = forearm, gray = back/breast; ***gender:*** blue = male, pink = female; ***lung:*** red = evidence of lung fibrosis, green = no lung fibrosis; ***myositis:*** red = muscle weakness/inflammation present, green = no muscle weakness/inflammation; ***anti.scl:*** red = positive anti-Scl-70 test, green = negative anti-Scl-70; ***target:*** green = gene on the Open Targets Platform for SSc, yellow = gene not found on the Open Targets Platform for SSc.

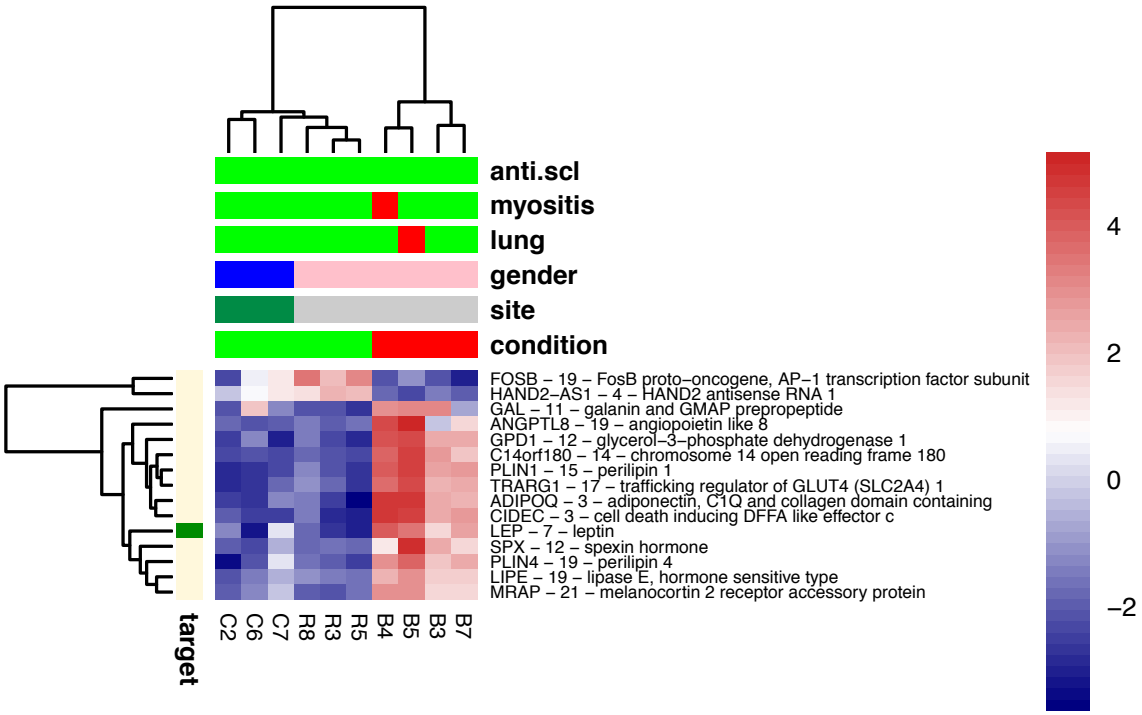


Figure G.4: Heatmap showing gene clustering according to gene expression signals in the MILD.BACK comparison. The color for each cell in the matrix represents gene expression signal from down-regulated (navy), to no change (white), to up-regulated (red). ***condition:*** red = case, green = control; ***site:*** green = forearm, gray = back/breast; ***gender:*** blue = male, pink = female; ***lung:*** red = evidence of lung fibrosis, green = no lung fibrosis; ***myositis:*** red = muscle weakness/inflammation present, green = no muscle weakness/inflammation; ***anti.scl:*** red = positive anti-Scl-70 test, green = negative anti-Scl-70; ***target:*** green = gene on the Open Targets Platform for SSc, yellow = gene not found on the Open Targets Platform for SSc.

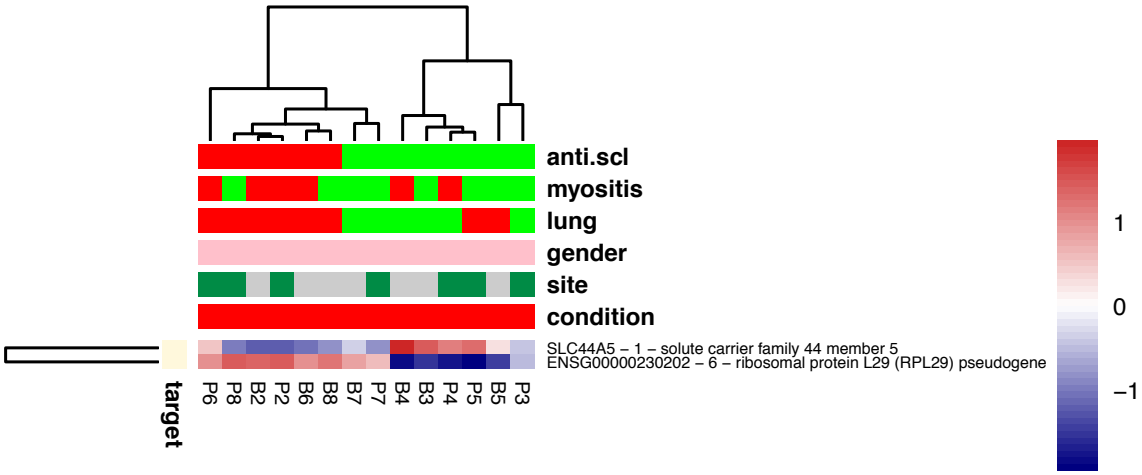


Figure G.5: Heatmap showing gene clustering according to gene expression signals in the SEVERE.MILD comparison. The color for each cell in the matrix represents gene expression signal from down-regulated (navy), to no change (white), to up-regulated (red). ***condition:*** red = case, green = control; ***site:*** green = forearm, gray = back/breast; ***gender:*** blue = male, pink = female; ***lung:*** red = evidence of lung fibrosis, green = no lung fibrosis; ***myositis:*** red = muscle weakness/inflammation present, green = no muscle weakness/inflammation; ***anti.scl:*** red = positive anti-Scl-70 test, green = negative anti-Scl-70; ***target:*** green = gene on the Open Targets Platform for SSc, yellow = gene not found on the Open Targets Platform for SSc.

Appendix H: Prioritised Genes

Table H.1: List of genes prioritised using `UpSet`

Ensembl ID	Description	Biotype	Chrom	HGNC
ENSG00000000005	tenomodulin	protein coding	X	TNMD
ENSG00000007216	solute carrier family 13 member 2	protein coding	17	SLC13A2
ENSG00000009950	MLX interacting protein like	protein coding	7	MLXIPL
ENSG00000034971	myocilin	protein coding	1	MYOC
ENSG00000039537	complement C6	protein coding	5	C6
ENSG00000077274	calpain 6	protein coding	X	CAPN6
ENSG00000101938	chordin like 1	protein coding	X	CHRDL1
ENSG00000103569	aquaporin 9	protein coding	15	AQP9
ENSG00000104435	stathmin 2	protein coding	8	STMN2
ENSG00000105664	cartilage oligomeric matrix protein	protein coding	19	COMP
ENSG00000106484	mesoderm specific transcript	protein coding	7	MEST
ENSG00000108342	colony stimulating factor 3	protein coding	17	CSF3
ENSG00000110484	secretoglobin family 2A member 2	protein coding	11	SCGB2A2
ENSG00000110848	CD69 molecule	protein coding	12	CD69
ENSG00000115008	interleukin 1 alpha	protein coding	2	IL1A
ENSG00000115009	C-C motif chemokine ligand 20	protein coding	2	CCL20
ENSG00000120057	secreted frizzled related protein 5	protein coding	10	SFRP5
ENSG00000120129	dual specificity phosphatase 1	protein coding	5	DUSP1
ENSG00000124102	peptidase inhibitor 3	protein coding	20	PI3
ENSG00000124205	endothelin 3	protein coding	20	EDN3
ENSG00000124731	triggering receptor expressed on myeloid cells 1	protein coding	6	TREM1
ENSG00000124875	C-X-C motif chemokine ligand 6	protein coding	4	CXCL6
ENSG00000124935	secretoglobin family 1D member 2	protein coding	11	SCGB1D2
ENSG00000125538	interleukin 1 beta	protein coding	2	IL1B
ENSG00000126262	free fatty acid receptor 2	protein coding	19	FFAR2
ENSG00000135447	protein phosphatase 1 regulatory inhibitor subunit 1A	protein coding	12	PPP1R1A
ENSG00000139330	keratocan	protein coding	12	KERA
ENSG00000140379	BCL2 related protein A1	protein coding	15	BCL2A1
ENSG00000140932	CKLF like MARVEL transmembrane domain containing 2	protein coding	16	CMTM2

Table H.1 continues on next page...

Table H.1 continued...

Ensembl ID	Description	Biotype	Chrom	HGNC
ENSG00000143546	S100 calcium binding protein A8	protein coding	1	S100A8
ENSG00000145681	hyaluronan and proteoglycan link protein 1	protein coding	5	HAPLN1
ENSG00000147257	glypican 3	protein coding	X	GPC3
ENSG00000147606	solute carrier family 26 member 7	protein coding	8	SLC26A7
ENSG00000151365	thyroid hormone responsive	protein coding	11	THRSP
ENSG00000155897	adenylate cyclase 8	protein coding	8	ADCY8
ENSG00000158445	potassium voltage-gated channel subfamily B member 1	protein coding	20	KCNB1
ENSG00000160161	cartilage intermediate layer protein 2	protein coding	19	CILP2
ENSG00000161634	dermcidin	protein coding	12	DCD
ENSG00000162747	Fc fragment of IgG receptor IIIb	protein coding	1	FCGR3B
ENSG00000162998	frizzled related protein	protein coding	2	FRZB
ENSG00000163220	S100 calcium binding protein A9	protein coding	1	S100A9
ENSG00000163221	S100 calcium binding protein A12	protein coding	1	S100A12
ENSG00000163421	prokineticin 2	protein coding	3	PROK2
ENSG00000163464	C-X-C motif chemokine receptor 1	protein coding	2	CXCR1
ENSG00000163734	C-X-C motif chemokine ligand 3	protein coding	4	CXCL3
ENSG00000163735	C-X-C motif chemokine ligand 5	protein coding	4	CXCL5
ENSG00000163736	pro-platelet basic protein	protein coding	4	PPBP
ENSG00000163739	C-X-C motif chemokine ligand 1	protein coding	4	CXCL1
ENSG00000163873	glutamate ionotropic receptor kainate type subunit 3	protein coding	1	GRIK3
ENSG00000163884	Kruppel like factor 15	protein coding	3	KLF15
ENSG00000166819	perilipin 1	protein coding	15	PLIN1
ENSG00000167588	glycerol-3-phosphate dehydrogenase 1	protein coding	12	GPD1
ENSG00000168004	HRAS like suppressor family member 5	protein coding	11	HRASLS5
ENSG00000170345	Fos proto-oncogene, AP-1 transcription factor subunit	protein coding	14	FOS
ENSG00000171195	mucin 7, secreted	protein coding	4	MUC7
ENSG00000171819	angiopoietin like 7	protein coding	1	ANGPTL7
ENSG00000180871	C-X-C motif chemokine receptor 2	protein coding	2	CXCR2
ENSG00000181092	adiponectin, C1Q and collagen domain containing	protein coding	3	ADIPOQ
ENSG00000184601	chromosome 14 open reading frame 180	protein coding	14	C14orf180
ENSG00000184811	trafficking regulator of GLUT4 (SLC2A4) 1	protein coding	17	TRARG1
ENSG00000187288	cell death inducing DFFA like effector c	protein coding	3	CIDEC
ENSG00000187398	leucine zipper protein 2	protein coding	11	LUZP2
ENSG00000196611	matrix metallopeptidase 1	protein coding	11	MMP1

Table H.1 continues on next page...

Table H.1 continued...

Ensembl ID	Description	Biotype	Chrom	HGNC
ENSG00000196616	alcohol dehydrogenase 1B (class I), beta polypeptide	protein coding	4	ADH1B
ENSG00000204128	chromosome 2 open reading frame 72	protein coding	2	C2orf72
ENSG00000206172	hemoglobin subunit alpha 1	protein coding	16	HBA1
ENSG00000211649	immunoglobulin lambda variable 7-46 (gene/pseudogene)	IG V gene	22	IGLV7-46
ENSG00000211896	immunoglobulin heavy constant gamma 1 (G1m marker)	IG C gene	14	IGHG1
ENSG00000211947	immunoglobulin heavy variable 3-21	IG V gene	14	IGHV3-21
ENSG00000224373	immunoglobulin heavy variable 4-59	IG V gene	14	IGHV4-59
ENSG00000229807	X inactive specific transcript	lincRNA	X	XIST
ENSG00000241351	immunoglobulin kappa variable 3-11	IG V gene	2	IGKV3-11
ENSG00000243238	immunoglobulin kappa variable 2-30	IG V gene	2	IGKV2-30
ENSG00000244734	hemoglobin subunit beta	protein coding	11	HBB
ENSG00000248323	lung cancer associated transcript 1	antisense	5	LUCAT1
ENSG00000266524	growth differentiation factor 10	protein coding	10	GDF10
ENSG00000268751	secretoglobin family 1B member 2, pseudogene	lincRNA	19	SCGB1B2P

Appendix I:
Significantly Differentially Expression Genes

Table I.1: Significantly differentially expressed genes in the ALL comparison set, where FDR$\leq$ 0.01 (1%) & $|Log_2FC| \geq 2$.

GeneID	Description	Chr	baseMean	log2FC	lfcSE	stat	pvalue	padj
PLIN1	perilipin 1	15	8533.97	7.78	0.64	8.99	2.58e-19	5.31e-15
ADIPOQ	adiponectin, C1Q and collagen domain containing	3	12362.85	9.20	0.89	8.14	4.12e-16	4.24e-12
CIDEC	cell death inducing DFFA like effector c	3	3106.79	9.23	0.92	7.85	4.17e-15	2.86e-11
GPD1	glycerol-3-phosphate dehydrogenase 1	12	7309.27	7.87	0.79	7.46	8.63e-14	4.44e-10
LGALS12	galectin 12	11	471.57	7.96	0.87	6.87	6.26e-12	2.58e-08
TRARG1	trafficking regulator of GLUT4 (SLC2A4) 1	17	1565.84	7.28	0.79	6.68	2.46e-11	8.46e-08
C14orf180	chromosome 14 open reading frame 180	14	587.45	8.79	1.06	6.43	1.26e-10	3.70e-07
PLIN4	perilipin 4	19	12065.99	6.64	0.76	6.14	8.49e-10	2.19e-06
LIPE	lipase E, hormone sensitive type	19	2444.74	5.25	0.57	5.73	9.77e-09	2.23e-05
SLC7A10	solute carrier family 7 member 10	19	152.30	6.78	0.86	5.53	3.12e-08	6.42e-05
HBB	hemoglobin subunit beta	11	15429.56	-5.41	0.62	-5.47	4.41e-08	8.25e-05
IGKV3-11	immunoglobulin kappa variable 3-11	2	473.11	7.21	0.98	5.32	1.04e-07	1.79e-04
MRAP	melanocortin 2 receptor accessory protein	21	206.39	7.24	1.00	5.23	1.69e-07	2.68e-04
FOS	Fos proto-oncogene, AP-1 transcription factor subunit	14	13320.79	-5.57	0.69	-5.21	1.89e-07	2.78e-04
LEP	leptin	7	2407.31	7.46	1.08	5.08	3.87e-07	5.31e-04
MLXIPL	MLX interacting protein like	7	633.70	6.44	0.91	4.90	9.48e-07	1.22e-03
IGLV7-46	immunoglobulin lambda variable 7-46 (gene/pseudogene)	22	96.01	8.47	1.35	4.81	1.55e-06	1.87e-03
THRSP	thyroid hormone responsive	11	4455.89	6.48	0.95	4.70	2.61e-06	2.98e-03
IGHG1	immunoglobulin heavy constant gamma 1 (G1m marker)	14	14647.15	6.47	0.96	4.66	3.11e-06	3.37e-03
HAND2-AS1	HAND2 antisense RNA 1	4	119.90	-6.07	0.88	-4.63	3.64e-06	3.75e-03
IGKV2-30	immunoglobulin kappa variable 2-30	2	92.17	6.58	0.99	4.62	3.92e-06	3.84e-03
HEPACAM	hepatic and glial cell adhesion molecule	11	153.70	6.09	0.90	4.52	6.04e-06	5.65e-03
SPX	spexin hormone	12	279.36	6.72	1.06	4.47	7.86e-06	7.04e-03
FOSB	FosB proto-oncogene, AP-1 transcription factor subunit	19	3606.41	-5.81	0.87	-4.40	1.07e-05	9.21e-03
IGHV3-21	immunoglobulin heavy variable 3-21	14	101.79	6.61	1.05	4.39	1.14e-05	9.40e-03
DUSP1	dual specificity phosphatase 1	5	8617.51	-4.33	0.53	-4.38	1.20e-05	9.48e-03
HAND2	heart and neural crest derivatives expressed 2	4	102.87	-4.82	0.65	-4.36	1.31e-05	9.96e-03

Table I.2: Significantly differentially expressed genes in the CASE.ARM comparison set, where FDR≤ 0.01 (1%) & $|Log_2FC| \geq 2$.

GeneID	Description	Chr	baseMean	log2FC	lfcSE	stat	pvalue	padj
PLIN1	perilipin 1	15	7562.39	9.14	0.74	9.61	7.23e-22	1.08e-17
CIDEC	cell death inducing DFFA like effector c	3	2649.76	9.02	0.91	7.68	1.63e-14	1.21e-10
XIST	X inactive specific transcript	X	7127.21	7.14	0.70	7.31	2.63e-13	1.30e-09
GPD1	glycerol-3-phosphate dehydrogenase 1	12	6211.17	7.88	0.84	7.00	2.58e-12	9.03e-09
C14orf180	chromosome 14 open reading frame 180	14	518.64	12.09	1.45	6.98	3.03e-12	9.03e-09
TRARG1	trafficking regulator of GLUT4 (SLC2A4) 1	17	1261.06	8.37	0.93	6.85	7.31e-12	1.81e-08
CXCL5	C-X-C motif chemokine ligand 5	4	6307.93	-13.87	1.79	-6.65	3.02e-11	6.42e-08
ADIPOQ	adiponectin, C1Q and collagen domain containing	3	9674.24	7.79	0.90	6.45	1.12e-10	2.09e-07
GDF10	growth differentiation factor 10	10	281.28	4.88	0.45	6.43	1.29e-10	2.14e-07
S100A8	S100 calcium binding protein A8	1	29269.50	-8.15	1.00	-6.16	7.20e-10	1.07e-06
GRIK3	glutamate ionotropic receptor kainate type subunit 3	1	105.56	5.44	0.57	6.07	1.30e-09	1.76e-06
CAPN6	calpain 6	X	162.46	5.12	0.55	5.67	1.41e-08	1.74e-05
IL1B	interleukin 1 beta	2	4629.94	-9.56	1.41	-5.38	7.46e-08	8.54e-05
MMP1	matrix metallopeptidase 1	11	9518.18	-9.99	1.51	-5.29	1.22e-07	1.30e-04
ADH1B	alcohol dehydrogenase 1B (class I), beta polypeptide	4	10976.16	5.43	0.65	5.27	1.39e-07	1.38e-04
DCD	dermcidin	12	17657.60	8.45	1.24	5.20	2.02e-07	1.77e-04
LUCAT1	lung cancer associated transcript 1	5	584.73	-6.53	0.87	-5.20	1.95e-07	1.77e-04
PPP1R1A	protein phosphatase 1 regulatory inhibitor subunit 1A	12	1069.66	7.16	1.00	5.19	2.14e-07	1.77e-04
FFAR2	free fatty acid receptor 2	19	879.69	-6.58	0.88	-5.17	2.31e-07	1.81e-04
SCGB1D2	secretoglobin family 1D member 2	11	1350.39	6.77	0.94	5.09	3.66e-07	2.72e-04
BCL2A1	BCL2 related protein A1	15	2302.65	-7.87	1.20	-4.90	9.72e-07	6.89e-04
MYOC	myocilin	1	400.05	7.49	1.14	4.82	1.43e-06	9.64e-04
CCL20	C-C motif chemokine ligand 20	2	110.47	-6.97	1.04	-4.79	1.63e-06	9.69e-04
EDN3	endothelin 3	20	225.78	5.67	0.76	4.81	1.51e-06	9.69e-04
S100A9	S100 calcium binding protein A9	1	50489.23	-7.69	1.19	-4.79	1.63e-06	9.69e-04
ANGPTL7	angiopoietin like 7	1	244.63	5.77	0.79	4.77	1.80e-06	1.03e-03
CMTM2	CKLF like MARVEL transmembrane domain containing 2	16	94.19	-6.58	0.96	-4.75	2.05e-06	1.13e-03
HEPACAM	hepatic and glial cell adhesion molecule	11	124.72	7.10	1.08	4.72	2.35e-06	1.25e-03
NAT8L	N-acetyltransferase 8 like	4	519.25	5.41	0.73	4.70	2.59e-06	1.33e-03
SCGB2A2	secretoglobin family 2A member 2	11	5430.99	7.80	1.24	4.68	2.93e-06	1.45e-03
C6	complement C6	5	317.06	6.47	0.96	4.64	3.47e-06	1.67e-03
PPBP	pro-platelet basic protein	4	742.45	-10.03	1.74	-4.62	3.76e-06	1.75e-03
AQP9	aquaporin 9	15	3325.67	-4.76	0.60	-4.60	4.23e-06	1.85e-03

Table I.2 continues on next page...

Table I.2 continued...

GeneID	Description	Chr	baseMean	log2FC	lfcSE	stat	pvalue	padj
ANGPTL8	angiopoietin like 8	19	187.25	8.21	1.35	4.60	4.16e-06	1.85e-03
HRASLS5	HRAS like suppressor family member 5	11	277.08	5.09	0.67	4.59	4.41e-06	1.87e-03
COMP	cartilage oligomeric matrix protein	19	6836.22	5.75	0.82	4.55	5.26e-06	2.12e-03
CD69	CD69 molecule	12	1112.63	-6.04	0.89	-4.56	5.23e-06	2.12e-03
PI3	peptidase inhibitor 3	20	3376.05	-7.88	1.29	-4.54	5.50e-06	2.15e-03
LIPE	lipase E, hormone sensitive type	19	2300.32	5.42	0.75	4.54	5.74e-06	2.19e-03
MEST	mesoderm specific transcript	7	1504.64	3.90	0.43	4.37	1.22e-05	4.55e-03
CXCL6	C-X-C motif chemokine ligand 6	4	376.53	-9.11	1.63	-4.35	1.34e-05	4.85e-03
LUZP2	leucine zipper protein 2	11	124.69	6.50	1.05	4.30	1.71e-05	6.05e-03
TREM1	triggering receptor expressed on myeloid cells 1	6	1666.50	-7.88	1.37	-4.29	1.76e-05	6.10e-03
IL1A	interleukin 1 alpha	2	149.20	-5.79	0.89	-4.28	1.87e-05	6.25e-03
FRZB	frizzled related protein	2	654.57	4.03	0.47	4.28	1.89e-05	6.25e-03
FCGR3B	Fc fragment of IgG receptor IIIb	1	1823.63	-6.34	1.02	-4.24	2.20e-05	6.98e-03
CXCL1	C-X-C motif chemokine ligand 1	4	1080.87	-7.29	1.25	-4.25	2.16e-05	6.98e-03
CXCR2	C-X-C motif chemokine receptor 2	2	1004.16	-5.52	0.83	-4.24	2.26e-05	7.00e-03
PROK2	prokineticin 2	3	835.65	-8.22	1.48	-4.21	2.54e-05	7.72e-03
CXCL3	C-X-C motif chemokine ligand 3	4	344.36	-6.40	1.05	-4.20	2.69e-05	8.01e-03
CHRDL1	chordin like 1	X	4999.76	3.84	0.44	4.15	3.32e-05	9.67e-03
CSF3	colony stimulating factor 3	17	153.29	-9.11	1.72	-4.14	3.44e-05	9.83e-03
SFRP5	secreted frizzled related protein 5	10	58.14	5.66	0.89	4.13	3.68e-05	9.97e-03
CXCR1	C-X-C motif chemokine receptor 1	2	989.41	-7.07	1.23	-4.13	3.67e-05	9.97e-03
FOS	Fos proto-oncogene, AP-1 transcription factor subunit	14	19540.12	-4.29	0.56	-4.13	3.67e-05	9.97e-03

Table I.3: Significantly differentially expressed genes in the CASE.BACK comparison set, where FDR$\leq$ 0.01 (1%) & $|\text{Log}_2\text{FC}| \geq$2.

GeneID	Description	Chr	baseMean	log2FC	lfcSE	stat	pvalue	padj
PLIN1	perilipin 1	15	5205.15	7.44	0.57	9.49	2.36e-21	5.34e-17
CIDEC	cell death inducing DFFA like effector c	3	1993.59	8.32	0.79	8.04	9.10e-16	1.03e-11
GPD1	glycerol-3-phosphate dehydrogenase 1	12	4716.53	7.64	0.71	7.94	1.98e-15	1.49e-11
C14orf180	chromosome 14 open reading frame 180	14	358.89	8.92	0.93	7.41	1.24e-13	7.00e-10
ADIPOQ	adiponectin, C1Q and collagen domain containing	3	8742.47	8.11	0.83	7.33	2.36e-13	1.07e-09
TRARG1	trafficking regulator of GLUT4 (SLC2A4) 1	17	1073.01	6.81	0.70	6.86	6.94e-12	2.61e-08
LIPE	lipase E, hormone sensitive type	19	1397.21	5.03	0.49	6.18	6.25e-10	2.02e-06

Table I.3 continues on next page...

Table I.3 continued...

GeneID	Description	Chr	baseMean	log2FC	lfcSE	stat	pvalue	padj
MRAP	melanocortin 2 receptor accessory protein	21	146.37	7.66	0.98	5.77	7.79e-09	2.20e-05
PLIN4	perilipin 4	19	7150.49	6.53	0.82	5.53	3.19e-08	8.00e-05
LGALS12	galectin 12	11	344.89	7.59	1.02	5.46	4.84e-08	1.09e-04
LEP	leptin	7	1088.80	7.49	1.03	5.34	9.26e-08	1.90e-04
HEPACAM	hepatic and glial cell adhesion molecule	11	105.51	5.65	0.73	5.00	5.80e-07	1.09e-03
GYG2	glycogenin 2	X	1264.91	4.65	0.54	4.94	7.99e-07	1.39e-03
ANGPTL8	angiopoietin like 8	19	312.93	7.73	1.17	4.88	1.04e-06	1.68e-03
FOSB	FosB proto-oncogene, AP-1 transcription factor subunit	19	5109.59	-5.92	0.82	-4.80	1.58e-06	2.34e-03
KRTAP10-4	keratin associated protein 10-4	21	11.29	20.11	3.78	4.79	1.65e-06	2.34e-03
KRTAP10-5	keratin associated protein 10-5	21	17.14	19.96	3.78	4.75	2.00e-06	2.66e-03
SPX	spexin hormone	12	268.19	6.94	1.07	4.63	3.72e-06	4.46e-03
ALDH1L1	aldehyde dehydrogenase 1 family member L1	3	338.90	5.02	0.65	4.62	3.75e-06	4.46e-03
SLC7A10	solute carrier family 7 member 10	19	104.66	6.59	1.01	4.55	5.41e-06	6.12e-03
GAL	galanin and GMAP prepropeptide	11	930.60	6.89	1.09	4.47	7.90e-06	8.51e-03
LINC01484	long intergenic non-protein coding RNA 1484	5	43.79	8.78	1.53	4.43	9.46e-06	9.72e-03

Table I.4: Significantly differentially expressed genes in the CASE comparison set, where FDR$\leq$ 0.01 (1%) & $|\text{Log}_2\text{FC}| \geq$1.1.

GeneID	Description	Chr	baseMean	log2FC	lfcSE	stat	pvalue	padj
HOXB-AS3	HOXB cluster antisense RNA 3	17	55.52	-4.80	0.49	-7.57	3.78e-14	8.94e-10
HAND2	heart and neural crest derivatives expressed 2	4	91.44	4.23	0.46	6.88	6.10e-12	7.22e-08
TBX5-AS1	TBX5 antisense RNA 1	12	192.57	2.70	0.25	6.28	3.38e-10	2.67e-06
FOXL2NB	FOXL2 neighbor	3	20.89	-5.68	0.81	-5.65	1.61e-08	9.52e-05
HAND2-AS1	HAND2 antisense RNA 1	4	111.09	5.11	0.76	5.24	1.58e-07	7.46e-04
AWAT2	acyl-CoA wax alcohol acyltransferase 2	X	711.31	-2.94	0.38	-4.82	1.42e-06	4.81e-03
CYP4F8	cytochrome P450 family 4 subfamily F member 8	19	864.26	-4.42	0.68	-4.85	1.25e-06	4.81e-03
FRG2DP	FSHD region gene 2 family member D, pseudogene	16	19.54	4.27	0.68	4.66	3.22e-06	9.53e-03

Table I.5: Significantly differentially expressed genes in the SEVERE.ARM comparison set, where FDR≤ 0.01 (1%) & $|Log_2FC| \geq 2$.

GeneID	Description	Chr	baseMean	log2FC	lfcSE	stat	pvalue	padj
PLIN1	perilipin 1	15	2668.23	8.40	0.74	8.68	3.95e-18	8.55e-14
C14orf180	chromosome 14 open reading frame 180	14	166.80	11.23	1.33	6.94	4.02e-12	4.35e-08
CIDEC	cell death inducing DFFA like effector c	3	776.12	8.01	0.95	6.33	2.53e-10	1.37e-06
XIST	X inactive specific transcript	X	5793.42	7.33	0.84	6.35	2.10e-10	1.37e-06
FCGR3B	Fc fragment of IgG receptor IIIb	1	2431.33	-8.55	1.11	-5.91	3.38e-09	1.46e-05
HAPLN1	hyaluronan and proteoglycan link protein 1	5	450.73	9.79	1.33	5.86	4.57e-09	1.65e-05
LUZP2	leucine zipper protein 2	11	125.07	7.26	0.92	5.74	9.57e-09	2.96e-05
TRARG1	trafficking regulator of GLUT4 (SLC2A4) 1	17	430.27	7.57	0.99	5.61	2.00e-08	5.41e-05
KERA	keratocan	12	176.50	7.49	0.99	5.52	3.38e-08	8.13e-05
CAPN6	calpain 6	X	127.87	5.24	0.59	5.47	4.52e-08	8.88e-05
GPD1	glycerol-3-phosphate dehydrogenase 1	12	1849.02	6.89	0.89	5.47	4.49e-08	8.88e-05
S100A9	S100 calcium binding protein A9	1	68248.85	-8.74	1.26	-5.35	8.62e-08	1.55e-04
C6	complement C6	5	308.53	7.19	0.98	5.30	1.13e-07	1.68e-04
PI3	peptidase inhibitor 3	20	4568.31	-8.82	1.29	-5.30	1.13e-07	1.68e-04
GRIK3	glutamate ionotropic receptor kainate type subunit 3	1	68.82	5.40	0.64	5.30	1.17e-07	1.68e-04
IGHV4-59	immunoglobulin heavy variable 4-59	14	453.99	9.49	1.43	5.25	1.50e-07	2.03e-04
COMP	cartilage oligomeric matrix protein	19	7569.08	6.65	0.90	5.16	2.49e-07	2.99e-04
S100A8	S100 calcium binding protein A8	1	39676.03	-8.80	1.32	-5.17	2.36e-07	2.99e-04
CILP2	cartilage intermediate layer protein 2	19	186.35	6.21	0.82	5.11	3.24e-07	3.69e-04
GDF10	growth differentiation factor 10	10	209.00	4.54	0.51	4.99	5.94e-07	6.42e-04
SLC26A7	solute carrier family 26 member 7	8	57.01	5.01	0.62	4.85	1.24e-06	1.22e-03
HBA1	hemoglobin subunit alpha 1	16	204.37	-7.35	1.10	-4.86	1.20e-06	1.22e-03
ADIPOQ	adiponectin, C1Q and collagen domain containing	3	3684.68	7.16	1.07	4.83	1.38e-06	1.29e-03
CXCR1	C-X-C motif chemokine receptor 1	2	1334.36	-8.02	1.33	-4.51	6.46e-06	5.82e-03
ANGPTL7	angiopoietin like 7	1	111.55	5.15	0.71	4.46	8.21e-06	6.83e-03
LUCAT1	lung cancer associated transcript 1	5	784.94	-7.11	1.14	-4.47	7.97e-06	6.83e-03
CXCL5	C-X-C motif chemokine ligand 5	4	8612.38	-13.84	2.68	-4.42	1.01e-05	8.05e-03
S100A12	S100 calcium binding protein A12	1	282.02	-7.86	1.33	-4.40	1.10e-05	8.46e-03
STMN2	stathmin 2	8	1316.00	4.87	0.65	4.38	1.18e-05	8.80e-03

Table I.6: Significantly differentially expressed genes in the SEVERE.BACK comparison set, where FDR≤ 0.01 (1%) & $|Log_2FC| \geq 2$.

GeneID	Description	Chr	baseMean	log2FC	lfcSE	stat	pvalue	padj
PLIN1	perilipin 1	15	2985.12	7.38	0.64	8.46	2.66e-17	5.67e-13
C14orf180	chromosome 14 open reading frame 180	14	182.03	8.70	0.89	7.51	5.72e-14	6.11e-10
CIDEC	cell death inducing DFFA like effector c	3	1041.50	8.12	0.93	6.60	4.02e-11	2.86e-07
GPD1	glycerol-3-phosphate dehydrogenase 1	12	2727.71	7.59	0.85	6.54	6.19e-11	3.31e-07
TIMP4	TIMP metallopeptidase inhibitor 4	3	312.51	4.27	0.36	6.23	4.59e-10	1.96e-06
ADIPOQ	adiponectin, C1Q and collagen domain containing	3	4826.74	7.99	0.99	6.06	1.38e-09	4.92e-06
TRARG1	trafficking regulator of GLUT4 (SLC2A4) 1	17	639.92	6.80	0.85	5.63	1.78e-08	5.43e-05
NAT8L	N-acetyltransferase 8 like	4	277.23	4.28	0.42	5.44	5.21e-08	1.39e-04
GYG2	glycogenin 2	X	769.74	4.61	0.49	5.35	8.77e-08	2.08e-04
LIPE	lipase E, hormone sensitive type	19	952.52	5.17	0.62	5.08	3.74e-07	7.99e-04
MRAP	melanocortin 2 receptor accessory protein	21	92.13	7.71	1.16	4.91	9.17e-07	1.78e-03
LGALS12	galectin 12	11	189.29	7.45	1.16	4.68	2.81e-06	5.01e-03
SLC7A10	solute carrier family 7 member 10	19	59.88	6.48	0.99	4.53	5.93e-06	9.75e-03

Table I.7: Significantly differentially expressed genes in the MILD.ARM comparison set, where FDR≤ 0.01 (1%) & $|Log_2FC| \geq 2$.

GeneID	Description	Chr	baseMean	log2FC	lfcSE	stat	pvalue	padj
PLIN1	perilipin 1	15	7000.81	9.51	0.82	9.19	3.78e-20	8.33e-16
DCD	dermcidin	12	20064.47	9.12	0.88	8.13	4.36e-16	4.80e-12
SCGB1D2	secretoglobin family 1D member 2	11	1292.03	7.17	0.69	7.45	9.14e-14	6.71e-10
CIDEC	cell death inducing DFFA like effector c	3	2595.68	9.48	1.04	7.22	5.12e-13	2.82e-09
SCGB2A2	secretoglobin family 2A member 2	11	5064.66	8.17	0.86	7.15	8.74e-13	3.85e-09
C14orf180	chromosome 14 open reading frame 180	14	493.50	12.51	1.54	6.82	9.16e-12	3.36e-08
GPD1	glycerol-3-phosphate dehydrogenase 1	12	6068.93	8.33	0.94	6.74	1.57e-11	4.94e-08
TRARG1	trafficking regulator of GLUT4 (SLC2A4) 1	17	1183.83	8.77	1.04	6.51	7.40e-11	2.04e-07
ADCY8	adenylate cyclase 8	8	129.20	7.37	0.85	6.29	3.26e-10	7.98e-07
GDF10	growth differentiation factor 10	10	269.03	5.10	0.50	6.16	7.16e-10	1.58e-06
ADIPOQ	adiponectin, C1Q and collagen domain containing	3	8722.99	8.13	1.02	6.01	1.81e-09	3.62e-06
XIST	X inactive specific transcript	X	5452.26	6.99	0.86	5.81	6.35e-09	1.16e-05
CXCL5	C-X-C motif chemokine ligand 5	4	7858.08	-13.89	2.14	-5.57	2.56e-08	4.33e-05
MYOC	myocilin	1	443.95	8.12	1.11	5.52	3.38e-08	5.32e-05
GRIK3	glutamate ionotropic receptor kainate type subunit 3	1	82.45	5.47	0.65	5.37	7.95e-08	1.12e-04
ADH1B	alcohol dehydrogenase 1B (class I), beta polypeptide	4	10659.44	5.72	0.69	5.36	8.17e-08	1.12e-04

Table I.7 continues on next page...

Table I.7 continued...

GeneID	Description	Chr	baseMean	log2FC	lfcSE	stat	pvalue	padj
FRZB	frizzled related protein	2	637.83	4.42	0.47	5.19	2.10e-07	2.72e-04
NAT8L	N-acetyltransferase 8 like	4	507.98	5.86	0.75	5.14	2.74e-07	3.35e-04
S100A8	S100 calcium binding protein A8	1	36198.08	-7.80	1.15	-5.03	4.92e-07	5.71e-04
PPP1R1A	protein phosphatase 1 regulatory inhibitor subunit 1A	12	1066.43	7.64	1.13	4.98	6.41e-07	7.04e-04
SCGB1B2P	secretoglobin family 1B member 2, pseudogene	19	1182.69	6.95	1.00	4.97	6.72e-07	7.04e-04
CAPN6	calpain 6	X	128.12	5.03	0.61	4.95	7.48e-07	7.49e-04
MMP1	matrix metallopeptidase 1	11	11814.22	-11.06	1.84	-4.93	8.03e-07	7.69e-04
HRASLS5	HRAS like suppressor family member 5	11	253.24	5.41	0.70	4.90	9.58e-07	8.79e-04
ANGPTL8	angiopoietin like 8	19	206.65	8.84	1.42	4.83	1.35e-06	1.19e-03
CCL20	C-C motif chemokine ligand 20	2	135.93	-7.88	1.25	-4.68	2.83e-06	2.39e-03
GPC3	glypican 3	X	1109.54	4.49	0.53	4.67	2.96e-06	2.42e-03
LIPE	lipase E, hormone sensitive type	19	2126.31	5.79	0.81	4.66	3.18e-06	2.50e-03
EDN3	endothelin 3	20	230.51	5.79	0.82	4.63	3.70e-06	2.81e-03
MEST	mesoderm specific transcript	7	1336.54	4.15	0.47	4.61	3.94e-06	2.89e-03
IL1B	interleukin 1 beta	2	5759.60	-10.09	1.76	-4.60	4.28e-06	3.04e-03
HEPACAM	hepatic and glial cell adhesion molecule	11	117.65	7.49	1.20	4.58	4.63e-06	3.19e-03
C2orf72	chromosome 2 open reading frame 72	2	140.67	6.03	0.89	4.55	5.31e-06	3.54e-03
SLC13A2	solute carrier family 13 member 2	17	167.65	5.83	0.84	4.55	5.49e-06	3.56e-03
MUC7	mucin 7, secreted	4	125.57	7.05	1.11	4.54	5.75e-06	3.62e-03
ANGPTL7	angiopoietin like 7	1	227.98	6.11	0.92	4.47	7.88e-06	4.82e-03
SFRP5	secreted frizzled related protein 5	10	53.94	5.98	0.90	4.44	8.92e-06	5.13e-03
ALDH1L1	aldehyde dehydrogenase 1 family member L1	3	357.33	5.84	0.86	4.44	9.08e-06	5.13e-03
ENSG00000277737	aquaporin 7 (AQP7) pseudogene	9	30.12	8.45	1.45	4.45	8.77e-06	5.13e-03
KCNB1	potassium voltage-gated channel subfamily B member 1	20	194.95	4.04	0.46	4.40	1.09e-05	6.00e-03
TNMD	tenomodulin	X	185.26	5.10	0.71	4.38	1.18e-05	6.37e-03
S100A9	S100 calcium binding protein A9	1	62455.59	-7.21	1.19	-4.36	1.29e-05	6.75e-03
KLF15	Kruppel like factor 15	3	169.04	4.76	0.64	4.31	1.60e-05	8.19e-03
GYG2	glycogenin 2	X	1605.88	5.15	0.73	4.30	1.70e-05	8.52e-03

Table I.8: Significantly differentially expressed genes in the MILD.BACK comparison set, where FDR≤ 0.01 (1%) & $|Log_2FC| \geq 2$.

GeneID	Description	Chr	baseMean	log2FC	lfcSE	stat	pvalue	padj
PLIN1	perilipin 1	15	3970.71	7.48	0.72	7.63	2.30e-14	5.06e-10
CIDEC	cell death inducing DFFA like effector c	3	1615.53	8.44	0.98	6.60	4.18e-11	4.59e-07
C14orf180	chromosome 14 open reading frame 180	14	294.85	9.06	1.08	6.52	6.98e-11	5.11e-07
GPD1	glycerol-3-phosphate dehydrogenase 1	12	3586.32	7.68	0.89	6.40	1.55e-10	8.50e-07
ADIPOQ	adiponectin, C1Q and collagen domain containing	3	6849.68	8.19	1.04	5.94	2.79e-09	1.22e-05
HAND2-AS1	HAND2 antisense RNA 1	4	87.84	-6.13	0.76	-5.45	4.98e-08	1.82e-04
TRARG1	trafficking regulator of GLUT4 (SLC2A4) 1	17	798.30	6.81	0.90	5.33	9.87e-08	3.09e-04
MRAP	melanocortin 2 receptor accessory protein	21	107.02	7.63	1.12	5.04	4.60e-07	1.26e-03
ANGPTL8	angiopoietin like 8	19	300.10	8.11	1.24	4.91	9.30e-07	2.27e-03
LIPE	lipase E, hormone sensitive type	19	967.51	4.89	0.60	4.84	1.31e-06	2.83e-03
LEP	leptin	7	1032.10	7.83	1.21	4.82	1.42e-06	2.83e-03
FOSB	FosB proto-oncogene, AP-1 transcription factor subunit	19	6479.82	-6.43	0.93	-4.78	1.74e-06	3.18e-03
GAL	galanin and GMAP prepropeptide	11	878.24	7.19	1.15	4.53	5.95e-06	9.29e-03
SPX	spexin hormone	12	289.89	7.48	1.21	4.52	6.13e-06	9.29e-03
PLIN4	perilipin 4	19	5571.87	6.59	1.02	4.51	6.35e-06	9.29e-03

Table I.9: Significantly differentially expressed genes in the SEVERE.MILD comparison set, where FDR≤ 0.01 (1%) & $|Log_2FC| \geq 1.1$.

GeneID	Description	Chr	baseMean	log2FC	lfcSE	stat	pvalue	padj
ENSG00000230202	ribosomal protein L29 (RPL29) pseudogene	6	1267.19	4.86	0.50	7.50	6.40e-14	1.51e-09
SLC44A5	solute carrier family 44 member 5	1	195.40	-4.61	0.68	-5.16	2.46e-07	2.91e-03

Appendix J:
Pathways Identified using gage

Table J.1: Down-regulated pathways in the ALL comparison identified by gage.

Pathway name	p.geomean	stat.mean	p.val	q.val	set.size	exp1
hsa03010 Ribosome	3.25e-08	-5.84e+00	3.25e-08	5.26e-06	85	3.25e-08
hsa04141 Protein processing in endoplasmic reticulum	9.74e-06	-4.37e+00	9.74e-06	7.89e-04	160	9.74e-06
hsa04120 Ubiquitin mediated proteolysis	2.28e-05	-4.18e+00	2.28e-05	1.23e-03	133	2.28e-05
hsa03040 Spliceosome	3.68e-05	-4.08e+00	3.68e-05	1.41e-03	127	3.68e-05
hsa03013 RNA transport	4.37e-05	-4.02e+00	4.37e-05	1.41e-03	141	4.37e-05
hsa04110 Cell cycle	6.09e-05	-3.94e+00	6.09e-05	1.64e-03	121	6.09e-05
hsa04062 Chemokine signaling pathway	2.69e-04	-3.50e+00	2.69e-04	6.22e-03	175	2.69e-04
hsa04144 Endocytosis	4.81e-04	-3.33e+00	4.81e-04	9.73e-03	199	4.81e-04
hsa04660 T cell receptor signaling pathway	6.92e-04	-3.25e+00	6.92e-04	1.25e-02	104	6.92e-04
hsa04010 MAPK signaling pathway	7.99e-04	-3.18e+00	7.99e-04	1.26e-02	247	7.99e-04
hsa04380 Osteoclast differentiation	8.57e-04	-3.17e+00	8.57e-04	1.26e-02	124	8.57e-04
hsa04621 NOD-like receptor signaling pathway	2.20e-03	-2.91e+00	2.20e-03	2.97e-02	58	2.20e-03
hsa04650 Natural killer cell mediated cytotoxicity	2.54e-03	-2.84e+00	2.54e-03	3.08e-02	109	2.54e-03
hsa03018 RNA degradation	2.67e-03	-2.86e+00	2.67e-03	3.08e-02	68	2.67e-03
hsa04310 Wnt signaling pathway	3.41e-03	-2.73e+00	3.41e-03	3.47e-02	142	3.41e-03
hsa03008 Ribosome biogenesis in eukaryotes	3.43e-03	-2.77e+00	3.43e-03	3.47e-02	71	3.43e-03
hsa04360 Axon guidance	3.95e-03	-2.68e+00	3.95e-03	3.76e-02	125	3.95e-03
hsa04114 Oocyte meiosis	4.38e-03	-2.65e+00	4.38e-03	3.95e-02	104	4.38e-03
hsa04012 ErbB signaling pathway	5.30e-03	-2.59e+00	5.30e-03	4.26e-02	85	5.30e-03
hsa04662 B cell receptor signaling pathway	5.42e-03	-2.58e+00	5.42e-03	4.26e-02	73	5.42e-03
hsa04210 Apoptosis	5.52e-03	-2.58e+00	5.52e-03	4.26e-02	83	5.52e-03
hsa04145 Phagosome	6.45e-03	-2.50e+00	6.45e-03	4.75e-02	142	6.45e-03

Table J.2: Up-regulated pathways in the CASE.ARM comparison identified by gage.

Pathway name	p.geomean	stat.mean	p.val	q.val	set.size	exp1
hsa04970 Salivary secretion	3.60e-04	3.46e+00	3.60e-04	2.96e-02	73	3.60e-04
hsa04972 Pancreatic secretion	3.65e-04	3.45e+00	3.65e-04	2.96e-02	77	3.65e-04

Table J.3: Down-regulated pathways in the CASE.ARM comparison identified by gage.

Pathway name	p.geomean	stat.mean	p.val	q.val	set.size	exp1
hsa03010 Ribosome	1.39e-06	-5.00e+00	1.39e-06	2.24e-04	85	1.39e-06
hsa04380 Osteoclast differentiation	1.68e-05	-4.23e+00	1.68e-05	1.36e-03	124	1.68e-05
hsa04110 Cell cycle	2.70e-05	-4.12e+00	2.70e-05	1.46e-03	121	2.70e-05
hsa04621 NOD-like receptor signaling pathway	7.04e-05	-3.97e+00	7.04e-05	2.85e-03	58	7.04e-05
hsa04062 Chemokine signaling pathway	1.03e-04	-3.76e+00	1.03e-04	2.96e-03	174	1.03e-04
hsa04650 Natural killer cell mediated cytotoxicity	1.10e-04	-3.76e+00	1.10e-04	2.96e-03	109	1.10e-04
hsa04620 Toll-like receptor signaling pathway	2.20e-04	-3.60e+00	2.20e-04	4.76e-03	84	2.20e-04
hsa03040 Spliceosome	2.35e-04	-3.57e+00	2.35e-04	4.76e-03	127	2.35e-04
hsa03013 RNA transport	3.70e-04	-3.44e+00	3.70e-04	6.65e-03	141	3.70e-04
hsa04141 Protein processing in endoplasmic reticulum	5.46e-04	-3.31e+00	5.46e-04	8.84e-03	159	5.46e-04
hsa03050 Proteasome	7.19e-04	-3.39e+00	7.19e-04	1.06e-02	42	7.19e-04
hsa04145 Phagosome	8.15e-04	-3.18e+00	8.15e-04	1.10e-02	142	8.15e-04
hsa03008 Ribosome biogenesis in eukaryotes	1.74e-03	-3.01e+00	1.74e-03	2.17e-02	71	1.74e-03
hsa04623 Cytosolic DNA-sensing pathway	2.32e-03	-2.92e+00	2.32e-03	2.68e-02	40	2.32e-03
hsa00240 Pyrimidine metabolism	4.01e-03	-2.68e+00	4.01e-03	4.34e-02	96	4.01e-03
hsa04210 Apoptosis	4.58e-03	-2.64e+00	4.58e-03	4.42e-02	83	4.58e-03
hsa04640 Hematopoietic cell lineage	4.64e-03	-2.64e+00	4.64e-03	4.42e-02	79	4.64e-03
hsa04662 B cell receptor signaling pathway	5.52e-03	-2.58e+00	5.52e-03	4.97e-02	73	5.52e-03

Table J.4: Down-regulated pathways in the CASE.BACK comparison identified by gage.

Pathway name	p.geomean	stat.mean	p.val	q.val	set.size	exp1
hsa03010 Ribosome	1.50e-07	-5.47e+00	1.50e-07	2.42e-05	85	1.50e-07
hsa04141 Protein processing in endoplasmic reticulum	3.37e-05	-4.07e+00	3.37e-05	2.73e-03	159	3.37e-05
hsa04120 Ubiquitin mediated proteolysis	1.31e-04	-3.73e+00	1.31e-04	5.78e-03	133	1.31e-04
hsa04062 Chemokine signaling pathway	1.78e-04	-3.61e+00	1.78e-04	5.78e-03	172	1.78e-04
hsa04110 Cell cycle	2.06e-04	-3.61e+00	2.06e-04	5.78e-03	121	2.06e-04
hsa03013 RNA transport	2.14e-04	-3.59e+00	2.14e-04	5.78e-03	141	2.14e-04
hsa03040 Spliceosome	2.64e-04	-3.54e+00	2.64e-04	6.10e-03	127	2.64e-04
hsa04144 Endocytosis	5.17e-04	-3.31e+00	5.17e-04	9.57e-03	197	5.17e-04
hsa04010 MAPK signaling pathway	5.32e-04	-3.29e+00	5.32e-04	9.57e-03	244	5.32e-04
hsa04380 Osteoclast differentiation	6.02e-04	-3.28e+00	6.02e-04	9.76e-03	124	6.02e-04
hsa04660 T cell receptor signaling pathway	9.07e-04	-3.17e+00	9.07e-04	1.34e-02	103	9.07e-04

Table J.4 continues on next page...

Table J.4 continued...

Pathway name	p.geomean	stat.mean	p.val	q.val	set.size	exp1
hsa04621 NOD-like receptor signaling pathway	1.29e-03	-3.09e+00	1.29e-03	1.74e-02	58	1.29e-03
hsa04360 Axon guidance	1.62e-03	-2.98e+00	1.62e-03	2.02e-02	124	1.62e-03
hsa04310 Wnt signaling pathway	2.29e-03	-2.86e+00	2.29e-03	2.65e-02	141	2.29e-03
hsa04650 Natural killer cell mediated cytotoxicity	2.74e-03	-2.81e+00	2.74e-03	2.95e-02	107	2.74e-03
hsa04662 B cell receptor signaling pathway	3.33e-03	-2.76e+00	3.33e-03	3.37e-02	73	3.33e-03
hsa04145 Phagosome	4.45e-03	-2.64e+00	4.45e-03	4.24e-02	142	4.45e-03

Table J.5: Up-regulated pathways in the CASE comparison identified by gage.

Pathway name	p.geomean	stat.mean	p.val	q.val	set.size	exp1
hsa04350 TGF-beta signaling pathway	3.24e-04	3.49e+00	3.24e-04	3.18e-02	80	3.24e-04
hsa04145 Phagosome	3.93e-04	3.40e+00	3.93e-04	3.18e-02	142	3.93e-04
hsa04810 Regulation of actin cytoskeleton	7.21e-04	3.21e+00	7.21e-04	3.89e-02	196	7.21e-04

Table J.6: Down-regulated pathways in the SEVERE.ARM comparison identified by gage.

Pathway name	p.geomean	stat.mean	p.val	q.val	set.size	exp1
hsa03010 Ribosome	2.49e-06	-4.85e+00	2.49e-06	4.01e-04	84	2.49e-06
hsa04062 Chemokine signaling pathway	1.69e-04	-3.63e+00	1.69e-04	8.66e-03	171	1.69e-04
hsa04380 Osteoclast differentiation	1.71e-04	-3.63e+00	1.71e-04	8.66e-03	123	1.71e-04
hsa00190 Oxidative phosphorylation	2.49e-04	-3.55e+00	2.49e-04	8.66e-03	126	2.49e-04
hsa04621 NOD-like receptor signaling pathway	2.69e-04	-3.59e+00	2.69e-04	8.66e-03	58	2.69e-04
hsa04650 Natural killer cell mediated cytotoxicity	7.30e-04	-3.23e+00	7.30e-04	1.96e-02	106	7.30e-04
hsa04110 Cell cycle	1.01e-03	-3.13e+00	1.01e-03	2.21e-02	121	1.01e-03
hsa04620 Toll-like receptor signaling pathway	1.10e-03	-3.12e+00	1.10e-03	2.21e-02	84	1.10e-03
hsa03050 Proteasome	1.68e-03	-3.10e+00	1.68e-03	3.00e-02	41	1.68e-03
hsa04145 Phagosome	2.43e-03	-2.84e+00	2.43e-03	3.92e-02	140	2.43e-03

Table J.7: Up-regulated pathways in the MILD.ARM comparison identified by gage.

Pathway name	p.geomean	stat.mean	p.val	q.val	set.size	exp1
hsa04972 Pancreatic secretion	1.28e-04	3.75e+00	1.28e-04	1.04e-02	76	1.28e-04
hsa04970 Salivary secretion	1.28e-04	3.76e+00	1.28e-04	1.04e-02	72	1.28e-04

Table J.8: Down-regulated pathways in the MILD.ARM comparison identified by gage.

Pathway name	p.geomean	stat.mean	p.val	q.val	set.size	exp1
hsa03010 Ribosome	1.03e-05	-4.49e+00	1.03e-05	1.19e-03	85	1.03e-05
hsa04110 Cell cycle	1.98e-05	-4.20e+00	1.98e-05	1.19e-03	121	1.98e-05
hsa04380 Osteoclast differentiation	2.20e-05	-4.16e+00	2.20e-05	1.19e-03	123	2.20e-05
hsa04621 NOD-like receptor signaling pathway	7.15e-05	-3.98e+00	7.15e-05	2.90e-03	58	7.15e-05
hsa04650 Natural killer cell mediated cytotoxicity	2.64e-04	-3.52e+00	2.64e-04	7.42e-03	105	2.64e-04
hsa03040 Spliceosome	2.77e-04	-3.52e+00	2.77e-04	7.42e-03	127	2.77e-04
hsa04620 Toll-like receptor signaling pathway	3.21e-04	-3.49e+00	3.21e-04	7.42e-03	84	3.21e-04
hsa04062 Chemokine signaling pathway	4.00e-04	-3.39e+00	4.00e-04	8.10e-03	173	4.00e-04
hsa03013 RNA transport	4.86e-04	-3.35e+00	4.86e-04	8.75e-03	141	4.86e-04
hsa04141 Protein processing in endoplasmic reticulum	1.30e-03	-3.04e+00	1.30e-03	2.11e-02	159	1.30e-03
hsa04145 Phagosome	1.49e-03	-3.00e+00	1.49e-03	2.19e-02	141	1.49e-03
hsa03050 Proteasome	1.65e-03	-3.10e+00	1.65e-03	2.23e-02	41	1.65e-03
hsa03008 Ribosome biogenesis in eukaryotes	1.80e-03	-2.99e+00	1.80e-03	2.24e-02	71	1.80e-03
hsa04623 Cytosolic DNA-sensing pathway	2.11e-03	-2.95e+00	2.11e-03	2.44e-02	40	2.11e-03
hsa04115 p53 signaling pathway	4.20e-03	-2.68e+00	4.20e-03	4.53e-02	68	4.20e-03
hsa00240 Pyrimidine metabolism	4.77e-03	-2.62e+00	4.77e-03	4.83e-02	95	4.77e-03

Table J.9: Down-regulated pathways in the MILD.BACK comparison identified by gage.

Pathway name	p.geomean	stat.mean	p.val	q.val	set.size	exp1
hsa03010 Ribosome	1.93e-07	-5.41e+00	1.93e-07	3.13e-05	85	1.93e-07
hsa04380 Osteoclast differentiation	2.61e-05	-4.12e+00	2.61e-05	1.84e-03	123	2.61e-05
hsa04141 Protein processing in endoplasmic reticulum	3.41e-05	-4.06e+00	3.41e-05	1.84e-03	159	3.41e-05
hsa04062 Chemokine signaling pathway	2.11e-04	-3.57e+00	2.11e-04	8.54e-03	172	2.11e-04
hsa04120 Ubiquitin mediated proteolysis	2.75e-04	-3.52e+00	2.75e-04	8.79e-03	133	2.75e-04
hsa04010 MAPK signaling pathway	3.47e-04	-3.42e+00	3.47e-04	8.79e-03	243	3.47e-04

Table J.9 continues on next page...

Table J.9 continued...

Pathway name	p.geomean	stat.mean	p.val	q.val	set.size	exp1
hsa04145 Phagosome	3.80e-04	-3.41e+00	3.80e-04	8.79e-03	142	3.80e-04
hsa04110 Cell cycle	4.53e-04	-3.38e+00	4.53e-04	9.18e-03	121	4.53e-04
hsa04144 Endocytosis	6.26e-04	-3.26e+00	6.26e-04	1.01e-02	197	6.26e-04
hsa03013 RNA transport	6.59e-04	-3.27e+00	6.59e-04	1.01e-02	141	6.59e-04
hsa03040 Spliceosome	6.88e-04	-3.26e+00	6.88e-04	1.01e-02	127	6.88e-04
hsa04660 T cell receptor signaling pathway	8.75e-04	-3.18e+00	8.75e-04	1.10e-02	102	8.75e-04
hsa04310 Wnt signaling pathway	9.32e-04	-3.14e+00	9.32e-04	1.10e-02	140	9.32e-04
hsa04360 Axon guidance	1.01e-03	-3.13e+00	1.01e-03	1.10e-02	124	1.01e-03
hsa04621 NOD-like receptor signaling pathway	1.02e-03	-3.16e+00	1.02e-03	1.10e-02	58	1.02e-03
hsa04810 Regulation of actin cytoskeleton	1.62e-03	-2.97e+00	1.62e-03	1.64e-02	191	1.62e-03
hsa04650 Natural killer cell mediated cytotoxicity	2.18e-03	-2.89e+00	2.18e-03	2.07e-02	106	2.18e-03
hsa04662 B cell receptor signaling pathway	2.49e-03	-2.86e+00	2.49e-03	2.24e-02	72	2.49e-03
hsa04666 Fc gamma R-mediated phagocytosis	3.69e-03	-2.72e+00	3.69e-03	3.15e-02	92	3.69e-03
hsa04520 Adherens junction	4.96e-03	-2.63e+00	4.96e-03	4.02e-02	72	4.96e-03
hsa04640 Hematopoietic cell lineage	6.23e-03	-2.53e+00	6.23e-03	4.81e-02	78	6.23e-03

Table J.10: Up-regulated pathways in the SEVERE.MILD comparison identified by gage.

Pathway name	p.geomean	stat.mean	p.val	q.val	set.size	exp1
hsa04510 Focal adhesion	1.18e-05	4.29e+00	1.18e-05	1.91e-03	196	1.18e-05
hsa04380 Osteoclast differentiation	3.11e-05	4.09e+00	3.11e-05	2.11e-03	124	3.11e-05
hsa04512 ECM-receptor interaction	3.90e-05	4.06e+00	3.90e-05	2.11e-03	83	3.90e-05
hsa04145 Phagosome	4.60e-04	3.35e+00	4.60e-04	1.86e-02	142	4.60e-04
hsa04610 Complement and coagulation cascades	7.11e-04	3.29e+00	7.11e-04	2.18e-02	53	7.11e-04
hsa04514 Cell adhesion molecules (CAMs)	8.06e-04	3.19e+00	8.06e-04	2.18e-02	128	8.06e-04

Appendix K:
Pathways Constructed with `pathview`

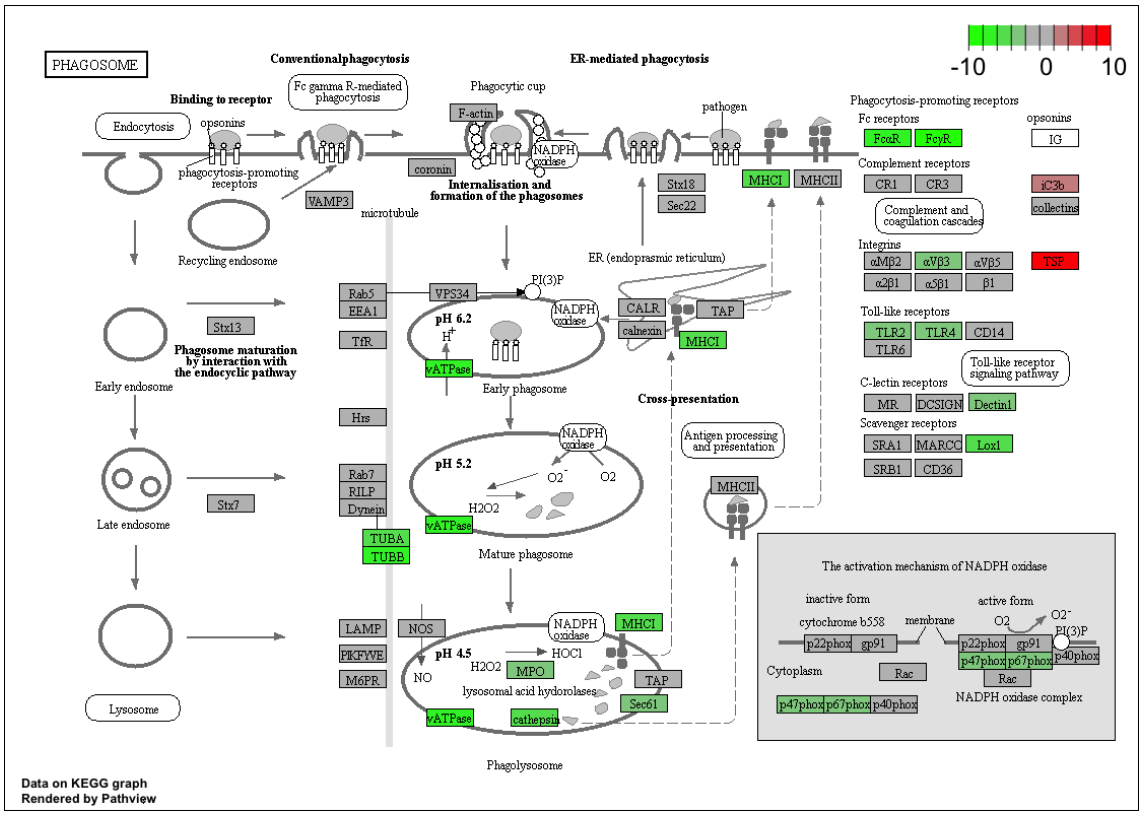


Figure K.1: Phagosome pathway (CASE.ARM) identified by gage and constructed with `pathview`. *Green:* down-regulated genes; *red:* up-regulated genes.

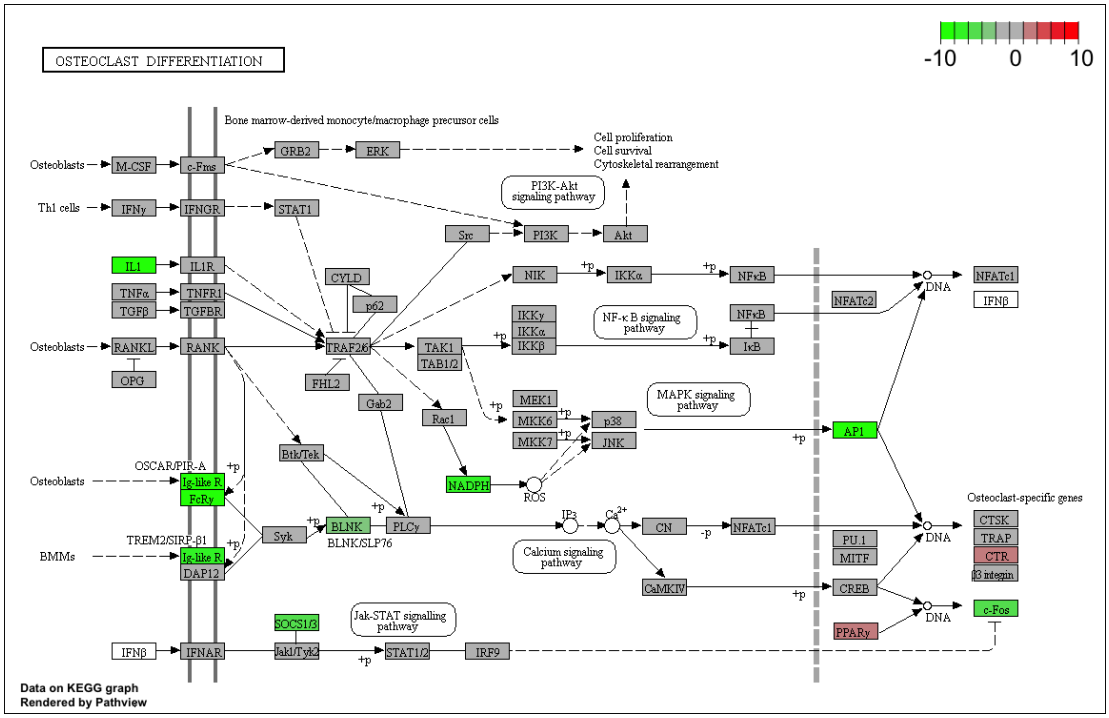


Figure K.2: Osteoclast differentiation pathway (CASE.ARM) identified by gage and constructed with `pathview`. *Green:* down-regulated genes; *red:* up-regulated genes.

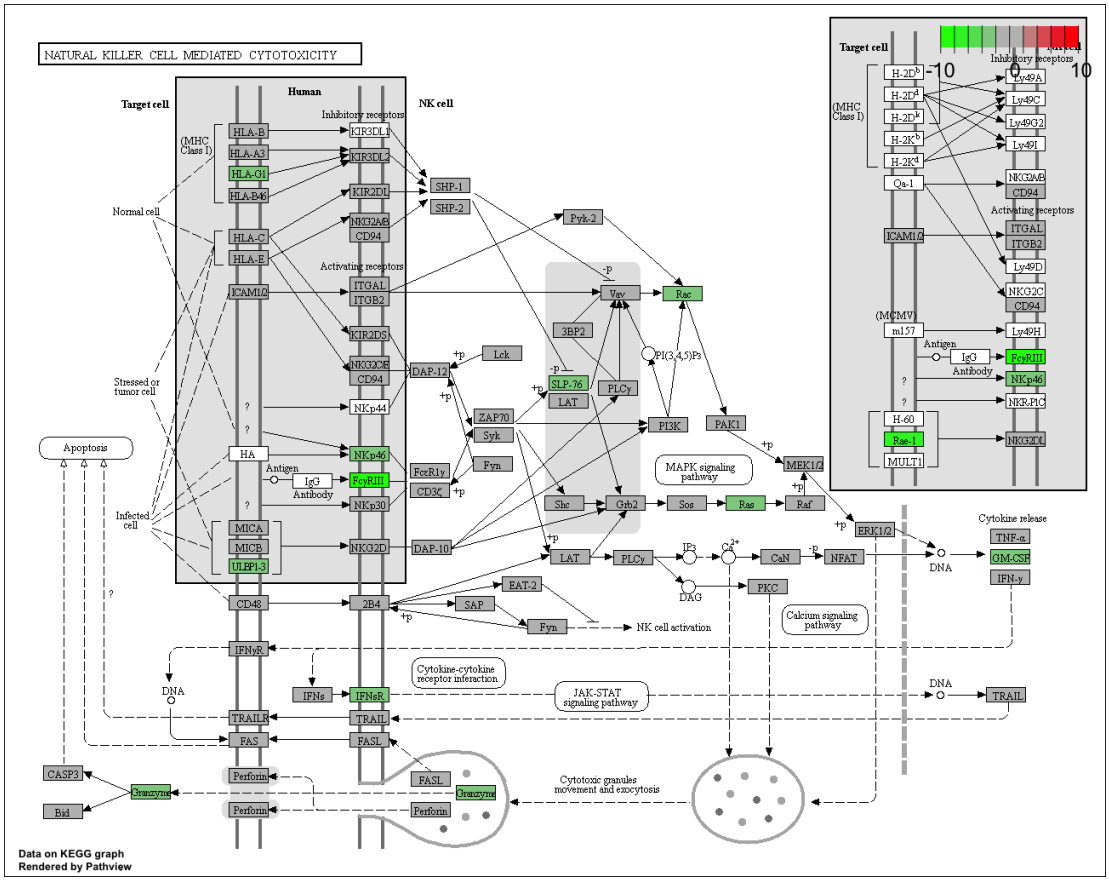


Figure K.3: Natural killer cell mediated cytotoxicity pathway (CASE.ARM) identified by gage and constructed with `pathview`. *Green:* down-regulated genes; *red:* up-regulated genes.

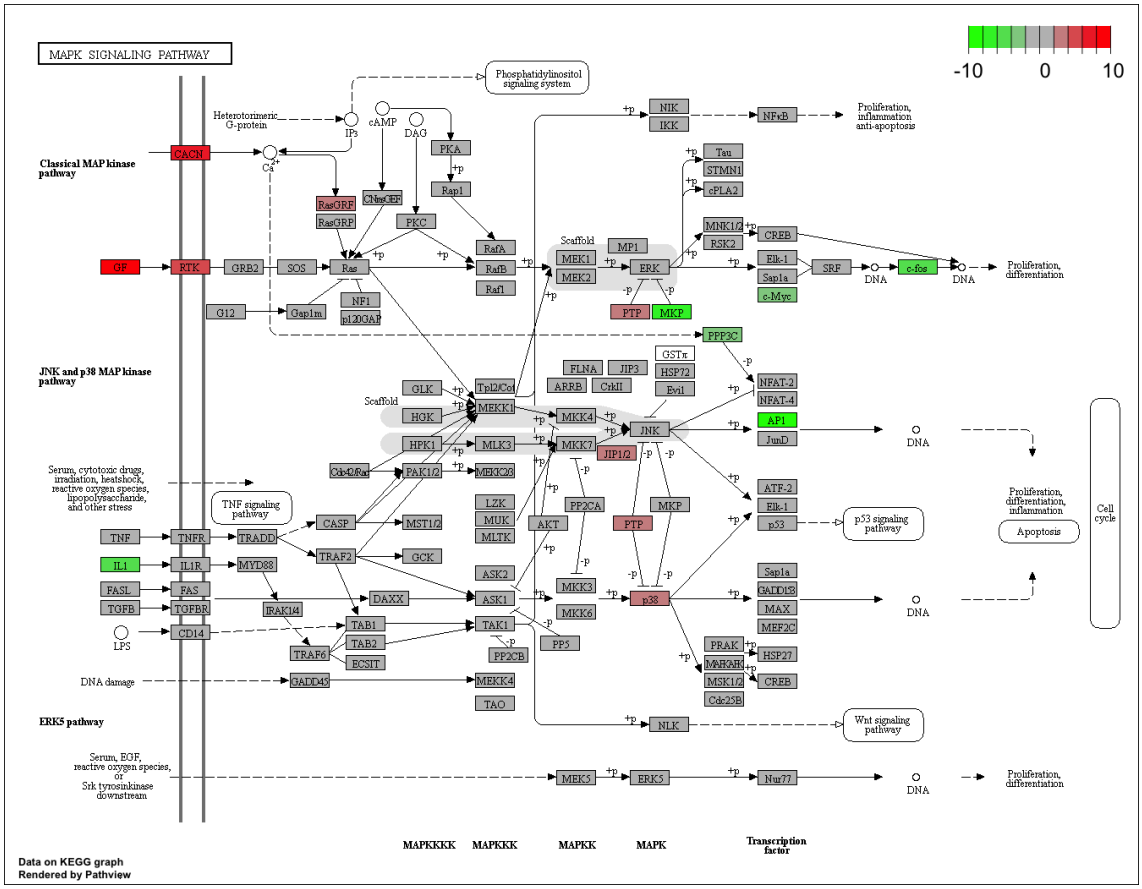


Figure K.4: MAPK signaling pathway (ALL) identified by `gage` and constructed with `pathview`. *Green:* down-regulated genes; *red:* up-regulated genes.

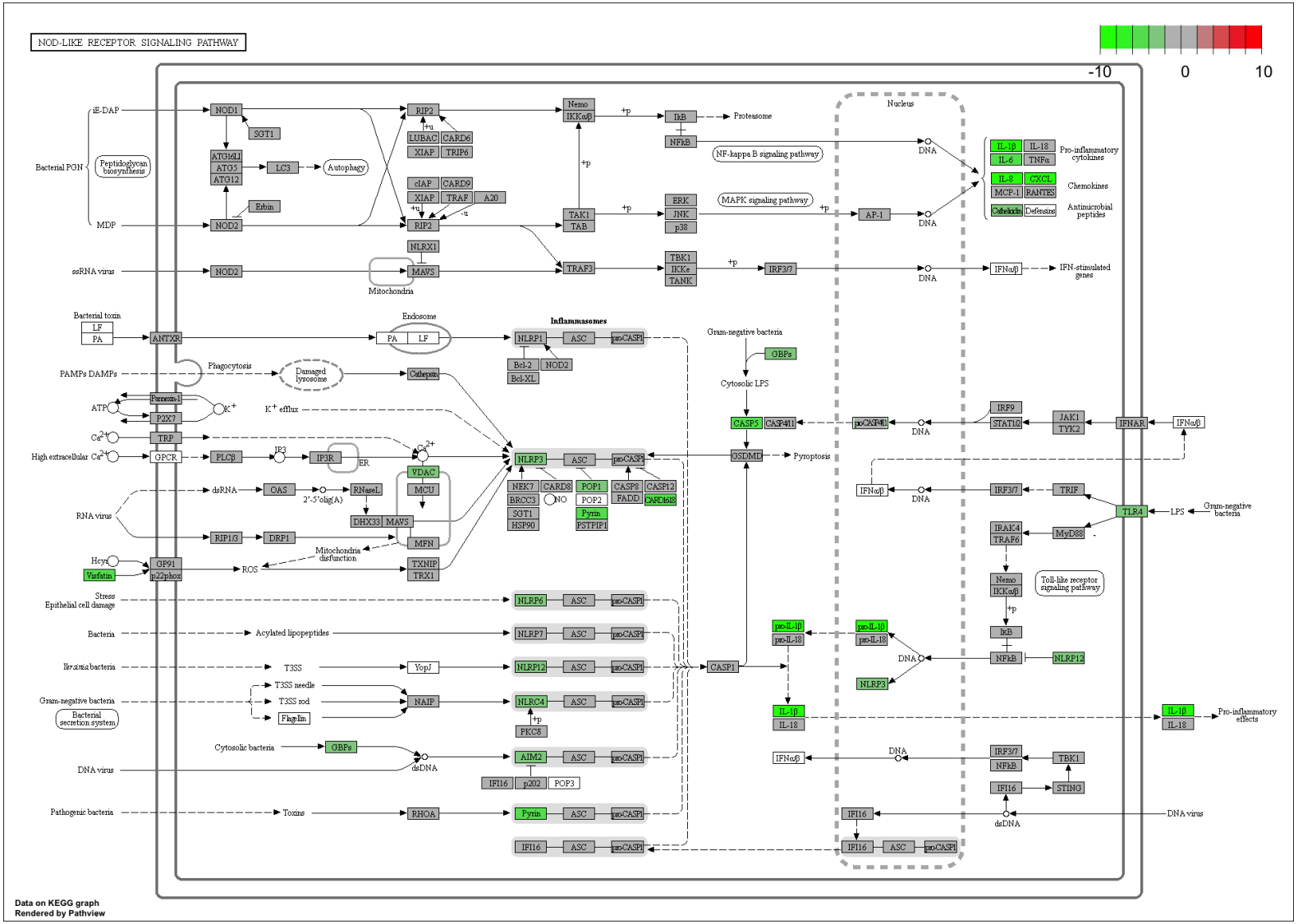


Figure K.5: NOD-like receptor pathway (CASE.ARM) identified by gage and constructed with `pathview`. *Green:* down-regulated genes; *red:* up-regulated genes.

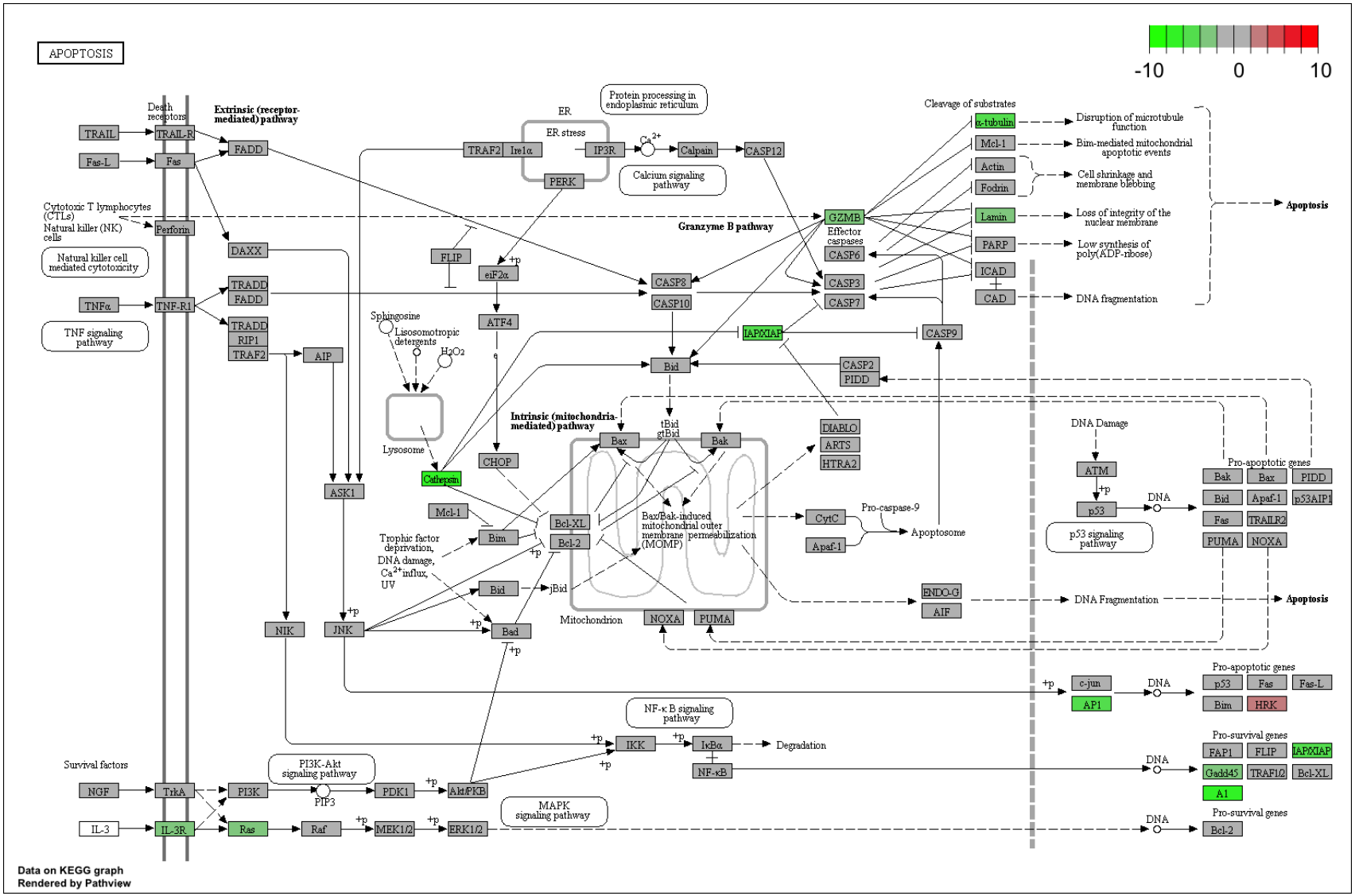


Figure K.6: Apoptosis pathway (CASE.ARM) identified by gage and constructed with `pathview`. *Green:* down-regulated genes; *red:* up-regulated genes.

Appendix L:
R Session Information and Packages

```
## Session info -------------------------------------------------------------
 setting  value
 version  R version 3.5.1 (2018-07-02)
 os       macOS  10.14.3
 system   x86_64, darwin18.0.0
 ui       unknown
 language (EN)
 collate  en_US.UTF-8
 ctype    en_US.UTF-8
 tz       Africa/Johannesburg
 date     2019-02-26

## Packages -----------------------------------------------------------------
 package              * version    date       lib source
 acepack                1.4.1      2016-10-29 [2] CRAN (R 3.5.1)
 annotate               1.60.0     2018-10-30 [2] Bioconductor
 AnnotationDbi        * 1.44.0     2018-10-30 [2] Bioconductor
 assertthat             0.2.0      2017-04-11 [2] CRAN (R 3.5.1)
 backports              1.1.2      2017-12-13 [2] CRAN (R 3.5.1)
 base64enc              0.1-3      2015-07-28 [2] CRAN (R 3.5.1)
 bindr                  0.1.1      2018-03-13 [2] CRAN (R 3.5.1)
 bindrcpp               0.2.2      2018-03-29 [2] CRAN (R 3.5.1)
 Biobase              * 2.42.0     2018-10-30 [2] Bioconductor
 BiocGenerics         * 0.28.0     2018-10-30 [2] Bioconductor
 BiocParallel         * 1.16.0     2018-10-30 [2] Bioconductor
 biomaRt              * 2.38.0     2018-10-30 [2] Bioconductor
 Biostrings             2.50.1     2018-11-06 [2] Bioconductor
 bit                    1.1-14     2018-05-29 [2] CRAN (R 3.5.1)
 bit64                  0.9-7      2017-05-08 [2] CRAN (R 3.5.1)
 bitops                 1.0-6      2013-08-17 [2] CRAN (R 3.5.1)
 blob                   1.1.1      2018-03-25 [2] CRAN (R 3.5.1)
 caTools                1.17.1.1   2018-07-20 [2] CRAN (R 3.5.1)
 checkmate              1.8.5      2017-10-24 [2] CRAN (R 3.5.1)
 cli                    1.0.1      2018-09-25 [2] CRAN (R 3.5.1)
 clipr                  0.5.0      2019-01-11 [1] CRAN (R 3.5.1)
 cluster                2.0.7-1    2018-04-13 [1] CRAN (R 3.5.1)
 colorspace             1.3-2      2016-12-14 [2] CRAN (R 3.5.1)
 crayon                 1.3.4      2017-09-16 [2] CRAN (R 3.5.1)
 data.table             1.11.8     2018-09-30 [2] CRAN (R 3.5.1)
 DBI                    1.0.0      2018-05-02 [2] CRAN (R 3.5.1)
 DelayedArray         * 0.8.0      2018-10-30 [2] Bioconductor
 DESeq2               * 1.22.1     2018-11-05 [2] Bioconductor
 digest                 0.6.18     2018-10-10 [2] CRAN (R 3.5.1)
 dplyr                * 0.7.8      2018-11-10 [2] CRAN (R 3.5.1)
 enrichR              * 1.0        2017-04-02 [2] CRAN (R 3.5.1)
 evaluate               0.12       2018-10-09 [2] CRAN (R 3.5.1)
 foreign                0.8-70     2017-11-28 [1] CRAN (R 3.5.1)
 Formula                1.2-3      2018-05-03 [2] CRAN (R 3.5.1)
 gage                 * 2.32.0     2018-10-30 [2] Bioconductor
 gageData             * 2.20.0     2018-11-01 [2] Bioconductor
 gdata                  2.18.0     2017-06-06 [2] CRAN (R 3.5.1)
 genefilter           * 1.64.0     2018-10-30 [2] Bioconductor
 geneplotter            1.60.0     2018-10-30 [2] Bioconductor
 GenomeInfoDb         * 1.18.0     2018-10-30 [2] Bioconductor
 GenomeInfoDbData       1.2.0      2018-11-03 [2] Bioconductor
 GenomicRanges        * 1.34.0     2018-10-30 [2] Bioconductor
 ggplot2              * 3.1.0      2018-10-25 [2] CRAN (R 3.5.1)
```

```
ggrepel              * 0.8.0      2018-05-09 [2] CRAN (R 3.5.1)
glue                   1.3.0      2018-07-17 [2] CRAN (R 3.5.1)
gplots               * 3.0.1      2016-03-30 [2] CRAN (R 3.5.1)
graph                  1.60.0     2018-10-30 [2] Bioconductor
gridExtra            * 2.3        2017-09-09 [2] CRAN (R 3.5.1)
gtable               * 0.2.0      2016-02-26 [2] CRAN (R 3.5.1)
gtools                 3.8.1      2018-06-26 [2] CRAN (R 3.5.1)
Hmisc                  4.1-1      2018-01-03 [2] CRAN (R 3.5.1)
hms                    0.4.2      2018-03-10 [2] CRAN (R 3.5.1)
htmlTable              1.12       2018-05-26 [2] CRAN (R 3.5.1)
htmltools              0.3.6      2017-04-28 [2] CRAN (R 3.5.1)
htmlwidgets            1.3        2018-09-30 [2] CRAN (R 3.5.1)
httr                   1.3.1      2017-08-20 [2] CRAN (R 3.5.1)
IRanges              * 2.16.0     2018-10-30 [2] Bioconductor
kableExtra           * 0.9.0      2018-05-21 [2] CRAN (R 3.5.1)
KEGGgraph              1.42.0     2018-10-30 [2] Bioconductor
KEGGREST               1.22.0     2018-10-30 [2] Bioconductor
KernSmooth             2.23-15    2015-06-29 [1] CRAN (R 3.5.1)
knitr                * 1.20       2018-02-20 [2] CRAN (R 3.5.1)
lattice              * 0.20-35    2017-03-25 [1] CRAN (R 3.5.1)
latticeExtra           0.6-28     2016-02-09 [2] CRAN (R 3.5.1)
lazyeval               0.2.1      2017-10-29 [2] CRAN (R 3.5.1)
locfit                 1.5-9.1    2013-04-20 [2] CRAN (R 3.5.1)
magrittr               1.5        2014-11-22 [2] CRAN (R 3.5.1)
Matrix                 1.2-14     2018-04-13 [1] CRAN (R 3.5.1)
matrixStats          * 0.54.0     2018-07-23 [2] CRAN (R 3.5.1)
memoise                1.1.0      2017-04-21 [2] CRAN (R 3.5.1)
munsell                0.5.0      2018-06-12 [2] CRAN (R 3.5.1)
nnet                   7.3-12     2016-02-02 [1] CRAN (R 3.5.1)
org.Hs.eg.db         * 3.7.0      2018-11-04 [2] Bioconductor
pathview             * 1.22.0     2018-10-30 [2] Bioconductor
pheatmap             * 1.0.10     2018-05-19 [2] CRAN (R 3.5.1)
pillar                 1.3.0      2018-07-14 [2] CRAN (R 3.5.1)
pkgconfig              2.0.2      2018-08-16 [2] CRAN (R 3.5.1)
plyr                   1.8.4      2016-06-08 [2] CRAN (R 3.5.1)
png                    0.1-7      2013-12-03 [2] CRAN (R 3.5.1)
PoiClaClu            * 1.0.2      2013-12-02 [2] CRAN (R 3.5.1)
prettyunits            1.0.2      2015-07-13 [2] CRAN (R 3.5.1)
progress               1.2.0      2018-06-14 [2] CRAN (R 3.5.1)
purrr                  0.2.5      2018-05-29 [2] CRAN (R 3.5.1)
R6                     2.3.0      2018-10-04 [2] CRAN (R 3.5.1)
RColorBrewer         * 1.1-2      2014-12-07 [2] CRAN (R 3.5.1)
Rcpp                   1.0.0      2018-11-07 [2] CRAN (R 3.5.1)
RCurl                  1.95-4.11  2018-07-15 [2] CRAN (R 3.5.1)
readr                  1.1.1      2017-05-16 [2] CRAN (R 3.5.1)
Rgraphviz              2.26.0     2018-10-30 [2] Bioconductor
rjson                  0.2.20     2018-06-08 [2] CRAN (R 3.5.1)
rlang                  0.3.0.1    2018-10-25 [2] CRAN (R 3.5.1)
rmarkdown              1.10       2018-06-11 [2] CRAN (R 3.5.1)
rpart                  4.1-13     2018-02-23 [1] CRAN (R 3.5.1)
rprojroot              1.3-2      2018-01-03 [2] CRAN (R 3.5.1)
RSQLite                2.1.1      2018-05-06 [2] CRAN (R 3.5.1)
rstudioapi             0.8        2018-10-02 [2] CRAN (R 3.5.1)
rvest                  0.3.2      2016-06-17 [2] CRAN (R 3.5.1)
S4Vectors            * 0.20.1     2018-11-09 [2] Bioconductor
scales                 1.0.0      2018-08-09 [2] CRAN (R 3.5.1)
sessioninfo            1.1.1.9000 2019-02-26 [1] Github (r-lib/sessioninfo@ac8fcc1)
stringi                1.2.4      2018-07-20 [2] CRAN (R 3.5.1)
stringr              * 1.3.1      2018-05-10 [2] CRAN (R 3.5.1)
SummarizedExperiment * 1.12.0     2018-10-30 [2] Bioconductor
survival               2.42-3     2018-04-16 [1] CRAN (R 3.5.1)
tibble                 1.4.2      2018-01-22 [2] CRAN (R 3.5.1)
```

```
 tidyselect             0.2.5      2018-10-11 [2] CRAN (R 3.5.1)
 UpSetR               * 1.3.3      2017-03-21 [2] CRAN (R 3.5.1)
 viridisLite            0.3.0      2018-02-01 [2] CRAN (R 3.5.1)
 withr                  2.1.2      2018-03-15 [2] CRAN (R 3.5.1)
 XML                    3.98-1.16  2018-08-19 [2] CRAN (R 3.5.1)
 xml2                 * 1.2.0      2018-01-24 [2] CRAN (R 3.5.1)
 xtable               * 1.8-3      2018-08-29 [2] CRAN (R 3.5.1)
 XVector                0.22.0     2018-10-30 [2] Bioconductor
 zlibbioc               1.28.0     2018-10-30 [2] Bioconductor

[1] /usr/local/Cellar/r/3.5.1/lib/R/library
[2] /usr/local/lib/R/3.5/site-library
```

Appendix M: Singularity Container Recipes for the `rnaSeqMetagen` Workflow

M.1 `trinity` container

```
Bootstrap: shub
From: singularityhub/ubuntu

%labels
    Maintainer Phelelani.Mpangase@wits.ac.za

%post
## Updates and essentials
apt-get update
apt-get install -y build-essential
apt-get install -y wget unzip curl python python-dev python-pip rsync
apt-get install -y libncurses5-dev libncursesw5-dev libbz2-dev liblzma-dev libtbb-dev

## Java
apt-get install -y software-properties-common debconf-utils
add-apt-repository -y ppa:webupd8team/java
apt-get update
echo "oracle-java8-installer shared/accepted-oracle-license-v1-1 select true" |
    ↪ debconf-set-selections
apt-get install -y zlib1g-dev oracle-java8-installer

## Update PIP
pip install -U pip

## Install Jellyfish
cd /opt \
    && wget https://github.com/gmarcais/Jellyfish/releases/download/v2.2.9/jellyfish
    ↪ -2.2.9.tar.gz \
    && tar -vxf jellyfish-2.2.9.tar.gz \
    && cd jellyfish-2.2.9 \
    && ./configure --prefix=/opt/jellyfish-2.2.9 \
    && make \
    && make install \
    && ldconfig \
    && rm /opt/jellyfish-2.2.9.tar.gz

## Install samtools
cd /opt \
    && wget --no-check-certificate https://github.com/samtools/samtools/releases/
    ↪ download/1.7/samtools-1.7.tar.bz2 \
    && tar -vxf samtools-1.7.tar.bz2 \
    && cd samtools-1.7 \
    && ./configure \
    && make \
    && make install \
    && rm /opt/samtools-1.7.tar.bz2

## Install Bowtie2
cd /opt \
    && wget -O bowtie2.zip --no-check-certificate https://sourceforge.net/projects/
    ↪ bowtie-bio/files/bowtie2/2.3.4.1/bowtie2-2.3.4.1-source.zip \
    && unzip bowtie2.zip \
    && cd bowtie2-2.3.4.1 \
    && make \
    && rm /opt/bowtie2.zip

## Install Salmon
cd /opt \
    && wget -c https://github.com/COMBINE-lab/salmon/releases/download/v0.9.1/Salmon
    ↪ -0.9.1_linux_x86_64.tar.gz \
    && tar -vxf Salmon-0.9.1_linux_x86_64.tar.gz \
    && rm /opt/Salmon-0.9.1_linux_x86_64.tar.gz

## Install Trinity
cd /opt \
```

```
    && wget -c https://github.com/trinityrnaseq/trinityrnaseq/archive/Trinity-v2.6.5.tar.gz \
    && tar -vxf Trinity-v2.6.5.tar.gz \
    && cd trinityrnaseq-Trinity-v2.6.5 \
    && make \
    && make plugins \
    && make install \
    && rm /opt/Trinity-v2.6.5.tar.gz

%environment
## Add paths to  environment
export PATH=/opt/jellyfish-2.2.9/bin:$PATH
export LD_LIBRARY_PATH=/opt/jellyfish-2.2.9/lib
export PATH=/opt/samtools-1.7/bin:$PATH
export PATH=/opt/bowtie2-2.3.4.1:$PATH
export PATH=/opt/Salmon-latest_linux_x86_64/bin:$PATH
export PATH=/opt/trinityrnaseq-Trinity-v2.6.5:$PATH
export JAVA_HOME=/usr/lib/jvm/java-8-oracle
export PYTHONPATH=/usr/local/lib/python2.7/dist-packages
```

M.2 kraken container

```
Bootstrap:shub
From:singularityhub/ubuntu

%labels
Maintainer Phelelani.Mpangase@wits.ac.za

%post
## Updates and essentials
apt-get update
apt-get install -y build-essential
apt-get install -y wget git curl lftp

## Install Jellyfish
cd /opt \
    && wget http://www.cbcb.umd.edu/software/jellyfish/jellyfish-1.1.11.tar.gz \
    && tar -vxf jellyfish-1.1.11.tar.gz \
    && cd jellyfish-1.1.11 \
    && ./configure \
    && make \
    && make install \
    && rm /opt/jellyfish-1.1.11.tar.gz

## Install Kraken | wget doesn't work fine. Replaced with lftp
cd /opt \
    && git clone https://github.com/DerrickWood/kraken2.git \
    && cd kraken2 \
    && ./install_kraken2.sh bin

### Install Krona
cd /opt \
    && wget https://github.com/marbl/Krona/releases/download/v2.7/KronaTools-2.7.tar \
    && tar -vxf KronaTools-2.7.tar \
    && cd KronaTools-2.7 \
    && ./install.pl \
    && ./updateTaxonomy.sh \
    && rm /opt/KronaTools-2.7.tar

## Instal NCBI BLAST+
cd /opt \
    && wget ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.7.1/ncbi-blast-2.7.1+-x64-linux.tar.gz \
    && tar -vxf ncbi-blast-2.7.1+-x64-linux.tar.gz \
    && rm /opt/ncbi-blast-2.7.1+-x64-linux.tar.gz

%environment
export PATH=/opt/kraken2/bin:$PATH
export LD_LIBRARY_PATH=/usr/local/lib:$LD_LIBRARY_PATH
export PATH=/opt/ncbi-blast-2.7.1+/bin:$PATH
```

M.3 `upset` container

```
Bootstrap:shub
From:singularityhub/ubuntu

%labels
Maintainer Phelelani.Mpangase@wits.ac.za

%post
## Updates and essentials
apt-get update
apt-get install -y build-essential
apt-get install -y wget git curl
apt-get install -y apt-transport-https

## Add R source and install R
echo "deb https://cloud.r-project.org/bin/linux/ubuntu trusty/" >> /etc/apt/sources.list
apt-get update
apt-key update
apt-get install -y --force-yes r-base

## Install R packages
R -e 'install.packages("stringr", repos="http://cloud.r-project.org/", dependencies=TRUE)'

## Install UpSet
cd /opt \
    && git clone https://github.com/VCG/upset.git
```

Appendix N:
Microbial Taxonomies Identified using `UpSet`

Table N.1: Microbial taxonomies shared between the SSc patients only.

Taxon	Taxon name (or species)	Sets	P2	B2	P3	B3	P4	B4	P5	B5	P6	B6	P7	B7	P8	B8
642	Aeromonas	6						•				•		•		•
28211	Alphaproteobacteria	5									•			•		
361177	Altererythrobacter	5								•		•				•
1663	Arthrobacter	4										•				
2211210	Arthrobacter sp. DCT5	5						•	•						•	
1357915	Arthrobacter sp. QXT-31	10	•	•			•	•	•			•			•	
1849032	Arthrobacter sp. U41	6	•			•	•								•	
1386	Bacillus	8					•		•	•					•	
1396	Bacillus cereus	5	•								•					•
526969	Bacillus cereus m1550	5	•				•									•
859143	Bacillus kochii	3	•				•									
43668	Brachybacterium	13	•	•		•	•	•	•		•	•			•	
446465	Brachybacterium faecium DSM 4810	6		•		•	•								•	
556288	Brachybacterium saurashtrense	11	•	•		•	•	•	•			•			•	
1903186	Brachybacterium sp. P6-10-X1	10	•	•		•	•	•	•						•	
2017484	Brachybacterium sp. VM2412	7				•		•				•			•	
2017485	Brachybacterium sp. VR2415	7	•			•	•	•	•							
85007	Corynebacteriales	4	•			•				•						
37914	Dietzia	11	•	•			•	•	•			•			•	
546160	Dietzia lutea	8	•	•		•	•		•						•	
712270	Dietzia sp. oral taxon 368	11	•	•		•	•	•	•			•			•	
237	Flavobacterium	7			•			•			•	•				•
2053	Gordonia	3					•									
32067	Leptotrichia	2										•				
475	Moraxella	7						•	•			•			•	
2065379	Paracoccus sp. CBA4604	7	•			•	•	•							•	
296591	Polaromonas sp. JS666	2									•					
1742993	Pseudarthrobacter	5		•		•									•	
452863	Pseudarthrobacter chlorophenolicus A6	10	•	•		•	•	•	•						•	

Table N.1 continues on next page...

Table N.1 continued...

Taxon	Taxon name (or species)	Sets	P2	B2	P3	B3	P4	B4	P5	B5	P6	B6	P7	B7	P8	B8
930171	Pseudarthrobacter phenanthrenivorans Sphe3	10	•	•		•	•	•	•						•	
364197	Pseudomonas pohangensis	3									•					
1063	Rhodobacter sphaeroides	2									•					
1915078	Thioclava nitratireducens	5			•				•			•				
32033	Xanthomonadaceae	5	•				•									•