



NOVA

IMS

Information
Management
School

MAAA

Mestrado em Métodos Analíticos Avançados

Master Program in Advanced Analytics

**Data Science Methods Applied to the Study of
The Signature of Regulatory CD4 T Cells in the
Human Thymus and its Modulation by the
Chromatin Landscape**

Susana Maria Santos do Paço

Dissertation presented as partial requirement for obtaining
the Master's degree in Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2021

Title: Data Science Methods Applied to the Study of The Signature of Regulatory CD4 T Cells in the Human Thymus and its Modulation by the Chromatin Landscape

Student Susana Maria Santos do Paço

MAA



NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**DATA SCIENCE METHODS APPLIED TO THE STUDY OF THE SIGNATURE OF
REGULATORY CD4 T CELLS IN THE HUMAN THYMUS AND ITS MODULATION
BY THE CHROMATIN LANDSCAPE**

by

Susana Paço

Dissertation presented as partial requirement for obtaining the Master's Degree in
Advanced Analytics

Co Advisor: Mauro Castelli

Co Advisor: Alexandre Raposo

November 2021

Dedication

Over all, I dedicate this little piece of work to the original authors of me, my parents, Clotilde and João. Through doubts and success they never stop believing in me.

Mauro and Alexandre, thank you for always expecting more of me. It's because of you that I've surpassed my own expectations.

To my dear friends, Guilherme, Ivo, Luís, Tiago, David, Filipa, Lara, Ana Sofia and Gabriela, I love you with all my heart, thank you for enduring me during this challenge. I'm better because of all of you.

This dissertation is also in loving memory of my grandparents, José and Fidélia, who taught me the values of hard work, grit and compassion for others. I hope they are proud of their granddaughter.

Acknowledgments

This work was supported by:

GenomePT project (POCI-01-0145-FEDER-022184), supported by COMPETE 2020 - Operational Programme for Competitiveness and Internationalisation (POCI), Lisboa Portugal Regional Operational Programme (Lisboa2020), Algarve Portugal Regional Operational Programme (CRESC Algarve2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF), and by Fundação para a Ciência e a Tecnologia (FCT).

Abstract

Thymic-derived Regulatory T cells (tTregs) play a central role in maintaining immune homeostasis by suppressing pro-inflammatory activity of conventional T cells (tTconvs). Disruption of tTreg development and/or function is at the origin of many pathologies, from allergies and autoimmunity to chronic inflammation and cancer. To understand tTreg development it is necessary to characterise tTreg genes and uncover the regulation of their expression.

This dissertation aims to contribute to the characterisation of regulatory CD4 T cells in the human thymus and the regulation of their development by exploring the relationship between differences in transcription factor binding to chromatin and changes in gene expression (differential gene expression). To do this, I analysed vast amounts of epigenomic and transcriptomic data produced by Next-Generation Sequencing, respectively, ATAC-seq and RNA-seq, generated from human tTregs and tTconvs using computational biology and data science methodologies.

In this dissertation I will discuss 3 steps of this project where Data Science played an important role: The discovery of a linear relationship between transcription factor accessibility to chromatin and associated gene expression in tTregs; the systematization and standardization of a gene set enrichment analysis protocol (GSEA) to detect signatures of activated biological pathways in ranked datasets of differential gene expression; and the development of systematised k-means clustering of Transcription Factor Binding Sites (TFBS), with heatmap visualisation, to discover relationships between the TFBS landscape and gene expression profile of tTregs.

Keywords: Immunology, Human CD4+ T cells, Genomics, Next Generation Sequencing, Data Science, K means

Contents

Dedication	i
Abstract	i
List of Figures	vii
List of Tables	viii
Abbreviations	viii
1 Introduction	1
1.1 The Immune System and Acquired Immunity	1
1.2 T Cell Development in Humans and Their Role in Immune Regulation	3
1.3 Objectives	4
1.4 Dissertation Organization	5
2 Theoretical Background	6
2.1 Clinical Immunology	6
2.2 Acquired Immunity and T cell Development	7
2.2.1 Regulatory T Cell Development and its Importance	8
2.3 Computational Immunology	11
2.3.1 Genetics and Genomics in Computational Immunology	12
3 Methodology	14
3.1 Multiomics Data: Extraction of thymic T Cell Data and Pre Preparation	14
3.1.1 Cell Sorting and Selection	14
3.1.2 RNA_{seq} and Differential Expression	14
3.1.3 $ATAC_{seq}$ and Differential Chromatin Accessibility	16
3.1.4 Digital Genomic Footprinting and Transcription Factor Binding analysis	16
3.2 Differential Expression vs Accessibility of The Chromatin	18
3.3 Standardization of Gene Set Enrichment Analysis	19
3.4 Clustering TFBS/Gene Binding Patterns in tTreg/tTconv Cells	24
4 Results	29
4.1 Analysing Gene Expression vs Differential Chromatin Accessibility	29
4.2 Gene Set Enrichment Analysis - Standardizing the Algorithm	33

4.3	Clustering of Digital Footprinting Analysis Results	37
4.3.1	Bound Thymic T regs - UP regulated Genes	38
4.3.2	Bound Thymic T regs - DOWN regulated Genes	40
4.3.3	Bound Thymic T regs - DEGs regulated Genes	41
4.3.4	Conclusions	43
5	Discussion and Conclusions	48
5.1	Analysing Gene Expression vs Differential Chromatin Accessibility	48
5.2	Gene Set Enrichment Analysis - Standardizing the Algorithm	49
5.3	Clustering of Digital Footprinting Analysis	49
6	Limitations And Recommendations For Future Works	51
	Bibliography	51

List of Figures

- 1.1 Distribution of the major organs of the Immune System in a Human(image created with the aid of BioRender.com) 1
- 1.2 CD4 Regulatory T cells are the scale keeping immune homeostasis in balance. 2
- 1.3 The location of the Thymus Gland (image created with the aid of BioRender.com) 3
- 1.4 Thymus T cell development, (image created with the aid of BioRender.com) . 3

- 2.1 Major Cell Groups of the Blood Lymphocyte. Acquired Immunity is mediated by the branches of the Lymphoid B and T cell precursors . Image from <https://www.genome.gov/genetics-glossary/Lymphocyte> 7
- 2.2 Initial stages of T cell Development of T cells in the Bone Marrow. Image created with the help of www.biorender.com 9
- 2.3 Development of T cells in the Thymus. Image retrieved from Germain (2002) 9
- 2.4 Cross-disciplinary efforts have allowed considerable advances in human medical research, from the clinic (a) to technology (b) to bioinformatics (c) and the laboratory (d), it's the collaboration of all that moves Computational Immunology forward Davis et al. (2017)) 11
- 2.5 Laboratory and Computational Techniques used to study different parts of the immune system. In special interest is the **D** group where we see Repertoire sequencing being used to study CD4+ T cells. Image retrieved from Davis et al. (2017)) 12

- 3.1 Strategy for sorting tTregs and tTconvs from human thymuses collected during routine corrective paediatric cardiac surgery. Mature CD4 single-positive thymic Tregs (tTregs) and their conventional counterparts (tTconvs) were sorted using CD25 and CD127. 15
- 3.2 Representative profiles of raw *RNA_{seq}* gene expression of emblematic genes of tTregs (*FOXP3* and *CTLA4*) and tTconvs (*IL7RA* and *CD40LG*) paired with the Accessibility to Chromatin Data (*ATAC_{seq}* data) within their genomic domains. Top Row indicates their location in their respective chromosome. "Regions of Open Chromatin" row indicates detection of regions with significant *ATAC_{seq}* signal enrichment, tTreg signal is depicted in red, tTconv signal in blue. For the gene row, black depicts sense direction, blue depicts antisense direction 15

3.3	Tobias, the package used for Digital Genomic Footprinting, which can be found in https://github.molgen.mpg.de/pages/loosolab/www/software/TOBIAS/	17
3.4	The TOBIAS framework, ScoreBigWig is represented by Differential Binsing Analysis.	17
3.5	Raw Data to extract Differential Chromatin Accessibility and Gene Expression info from	18
3.6	Data cleaned for input in data visualization of Gene Expression vs Differential Chromatin Accessibility	18
3.7	Differential Gene Expression in x and Differential Chromatin Accessibility in y	19
3.8	Diagram of the major stages of Gene Set Enrichment Analysis	20
3.9	The FGSEA algorithm is depicted in the image. Image from Sergushichev (2016)	20
3.10	The collections existent in the mSigDB databaset. They can be found at https://www.gsea-msigdb.org/gsea/msigdb/index.jsp	21
3.11	R function created to run the fgsea protocol in a standardized fashion. The full function can be found in https://github.com/theinsilicobiology/fgsea_msigDB_Thymus_paper/blob/main/Functions/FunctionsForGSEA.R	22
3.12	Function that runs the function Figure 3.11 iterating over the whole mSigDB database. The full function can be found in https://github.com/theinsilicobiology/fgsea_msigDB_Thymus_paper/blob/main/Functions/fgseaMsigDb.R	23
3.13	CSV output of Figure 3.11.	23
3.14	Heatmap to observe the results of FGSEA (with data originated in Figure 3.13) with the aid of heatmap.2() from the package <i>gplots</i> https://github.com/talgalili/gplots . The function that generates this heatmap is found in https://github.com/theinsilicobiology/fgsea_msigDB_Thymus_paper/blob/main/Functions/generateHeatmapCSVfgsea.R	24
3.15	Raw Data extracted from analysis from the TOBIAS framework described in subsection 3.1.4	25
3.16	Example of a small portion of a Matrix ready to be used as an input for clustering analysis. Genes are in rows, TFBS are in columns	25
3.17	<i>ComplexHeatmap</i> package logo Gu et al. (2016).	25
3.18	Function to extract relevant data to execute the heatmap. This example is for Treg_score. The full code for this task can be seen in https://github.com/theinsilicobiology/Kmeans_TOBIAS_CD4Thymus_paper/blob/main/Functions/ExtractInfoFromDataset.R	26
3.19	Function that executes the elbow and silhouette methods for the extracted data to determine the ideal number of clusters - k. It outputs the graphs for both rows and columns. The full function can be found in https://github.com/theinsilicobiology/Kmeans_TOBIAS_CD4Thymus_paper/blob/main/Functions/AssessNumClusters.R	27

3.20	Function that calculates the mode per column or per row (according to scaling) and sets the colours of the heatmap according to it. Scale from blue to green in <code>treg</code> and <code>tconv</code> score and green/black/orange for <code>diffbinding</code> . The full functions can be found in https://github.com/theinsilicobiology/Kmeans_TOBIAS_CD4Thymus_paper/blob/main/Functions/ColoursHeatmap.R	27
3.21	Function to extract cluster information after a heatmap is created. It generates 3 CSVs, one for gene clusters, one for TFBS clusters and one with the information combined. The functions can be found in https://github.com/theinsilicobiology/Kmeans_TOBIAS_CD4Thymus_paper/blob/main/Functions/extractClusterInformation.R	28
3.22	Function to assess which genes exist in a specific subset of gene expression but in the correspondent TOBIAS output data. The full function can be found in https://github.com/theinsilicobiology/Kmeans_TOBIAS_CD4Thymus_paper/blob/main/Functions/notintable.R	28
4.1	Bubble Plot of Differential Gene Expression in x and Differential Chromatin Accessibility in y . To assure one point per gene, the DCA was assumed to be the mean of all ROCs for each gene. The number of ROCs is stored as the size of each bubble giving us an idea on how many each gene possesses.	30
4.2	The same graph as Figure 4.1 but with an Added Linear Regression line between x and y	31
4.3	Initial Linear Regression results for the Gene Expression vs DCA combo.	31
4.4	Breusch-Pagan test results of the linear regression executed in Figure 4.3 using the function <code>bptest()</code> from the <code>lmtest</code> package. As the p -value is <0.05 , heteroscedasticity is in fact, present.	32
4.5	Robust regression of the same variables as in Figure 4.3. The R-squared did indeed improve to 0.2308	33
4.6	GSEA results for the full Gene Expression dataset in comparison with the Hallmark Collection from mSigDB. The Rows are the pathways with Enrichment, in the columns are the genes identified in common between our data and the hallmark collections, the value in the heatmap corresponds to the NES calculated during the FGSEA protocol	36
4.7	GSEA results for the Gene Expression dataset with a cut-off for significance in p -value in comparison with the Hallmark Collection from mSigDB. The Rows are the pathways with Enrichment, in the columns are the genes identified in common between our data and the hallmark collections, the value in the heatmap corresponds to the NES calculated during the FGSEA protocol	36
4.8	Enrichment plots for the pathways in Figure 4.7 where the p -value is significant.	37
4.9	Types of Heatmaps Created	38
4.10	Heatmap for Clustering of <code>Treg_score</code> data for the Bound Up regulated subset	39
4.11	Heatmap for Clustering of <code>Tconv_score</code> data for the Bound Up regulated subset	40
4.12	Heatmap for Clustering of <code>DiffBinding</code> data for the Bound Up regulated subset	41

4.13 Heatmap for Clustering of Treg_score data for the Bound Down regulated subset	42
4.14 Heatmap for Clustering of Tconv_score data for the Bound Down regulated subset	43
4.15 Heatmap for Clustering of DiffBinding data for the Bound Down regulated subset	44
4.16 Heatmap for Clustering of Treg_score data for the Bound DEGs subset	45
4.17 Heatmap for Clustering of Tconv_score data for the Bound DEGs subset . . .	45
4.18 Heatmap for Clustering of DiffBinding data for the Bound DEGs subset . . .	46
4.19 Heatmap for Clustering of DiffBinding data for the Bound UP regulated subset - Annotated for to show the GRMs	46
4.20 Heatmap for Clustering of DiffBinding data for the Bound Down regulated subset - Annotated for to show the GRMs	47

List of Tables

- 4.1 ROCs main characteristics in the data 29
- 4.2 Genes main characteristics in the data 29
- 4.3 Distribution of the Genes per Gene Biotype 30
- 4.4 Subset of the results table of the GSEA applied to our gene expression data with the Hallmark mSigDB as a comparison. 34
- 4.5 Subset of the results table of the GSEA applied to our gene expression data (with a previous cutoff applied to those which p -value for the expression was significant) with the Hallmark mSigDB as a comparison. 35

Chapter 1

Introduction

1.1 The Immune System and Acquired Immunity

Immunity is the ability of an organism to resist damage from foreign substances such as microorganisms, harmful chemicals and internal threats such as cancer. The immune system as a whole is divided by a network of cells, molecules and organs that can be found all over the body as we can see in 1.1 .

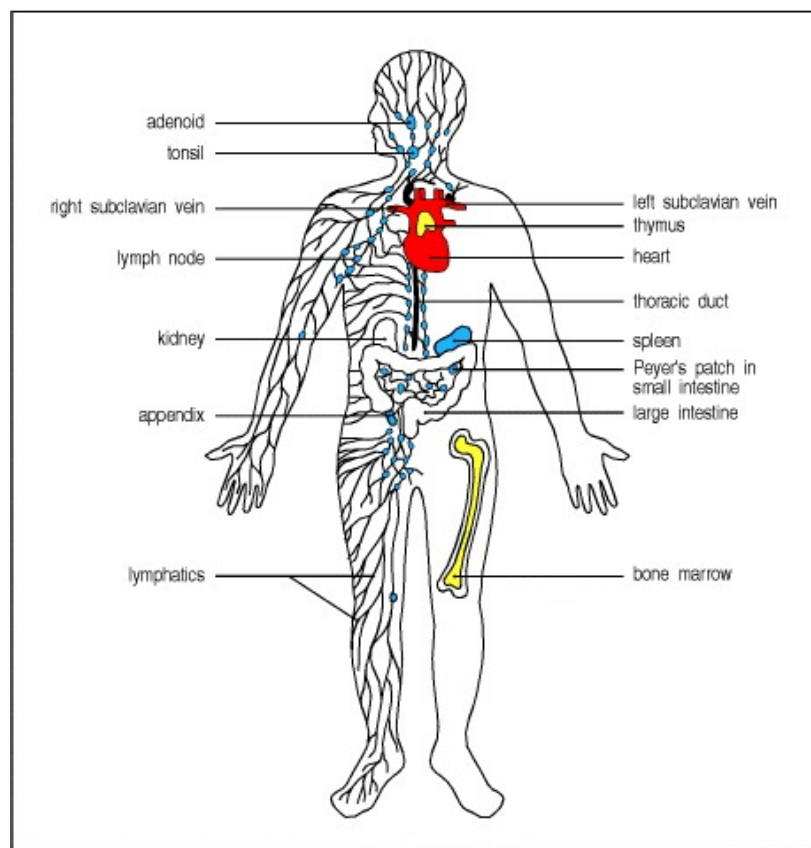


Figure 1.1: Distribution of the major organs of the Immune System in a Human(image created with the aid of BioRender.com)

It can be divided into two types: innate immunity and adaptive immunity. In innate immunity, the body detects and reacts to threats the same way with each exposure

to them, while with adaptive immunity the body recognizes the threat and improves the response to it with each exposure to it. Tate and Seeley (2009).

Adaptive immunity, also known as acquired immunity, is characterized by its ability to learn and develop which each exposure to a pathogen, through creating immunological memory specific to each one after exposure to them. This mechanism is kept working through a fine equilibrium mechanism, known as **immune homeostasis**. Through this mechanism, the body learns to identify what belongs to it (self) from what is foreign (non-self), ignoring the first but reacting to the second. A lack of balance in self/non self recognition and the organism can end up in auto-immunity, where it attacks itself or in immunodeficiency, where it's unable to sufficiently respond to a pathogen Arosa et al. (2012).

In the balance of this mechanism lies a great number of immune related diseases, mostly by shifting the balance of immune homeostasis towards one side. Allergies, Lupus and Rheumathoid Arthritis are examples of a shift towards auto immunity, HIV, primary immunodeficiencies or the usage of immunosuppressors in transplant patients are examples of a shift towards immunodeficiency Lee and Lee (2018), Sakaguchi et al. (2020) Godinho-Santos et al. (2020).

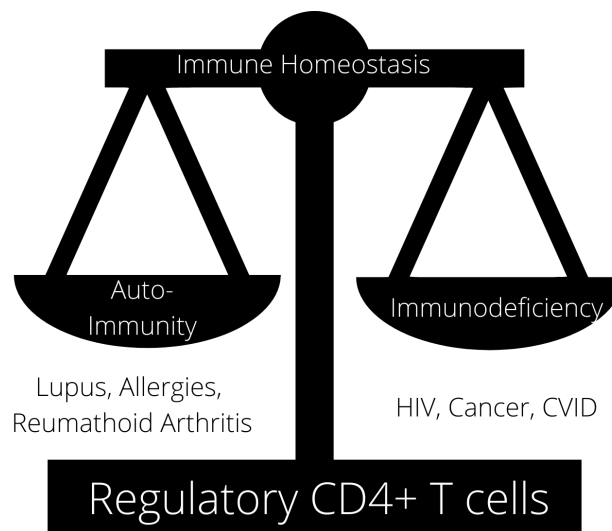


Figure 1.2: CD4 Regulatory T cells are the scale keeping immune homeostasis in balance.

Regulatory T cells, also known as regulatory CD4+ T cells (Treg), play an important role in this mechanism, actively suppressing the immune system, preventing auto-immunity. Their importance is such that dysregulation of their functioning, either through genetics or through acquired form through a virus or bacteria, can lead into serious diseases such as diabetes, allergies or associated with higher propensity to certain cancers.

Studying the development of CD4 t cells in humans thus becomes a crucial point into both understanding the mechanism of immune homeostasis and provide new clinical insights into the illnesses dependent on the malfunctioning of this group of cells.

1.2 T Cell Development in Humans and Their Role in Immune Regulation

T cells begin their development as haematopoietic precursors in the bone marrow, travelling to the thymus through the blood stream, the organ located beneath the sternum, in the upper front part of the chest as seen in fig 1.3. It's in this gland that they complete their development and earn their name as T cells Arosa et al. (2012).

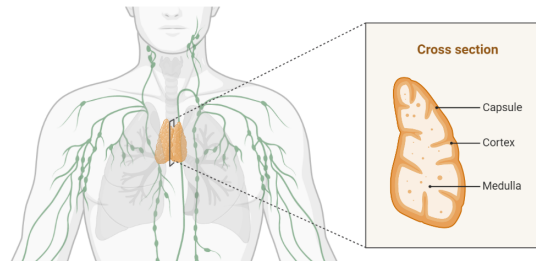


Figure 1.3: The location of the Thymus Gland (image created with the aid of BioRender.com)

Inside the thymus, the haematopoietic precursors become thymocytes and initiate their last stage of maturation. This stage occurs mostly in the cortex (which you can see in the diagram 1.4) and is composed by two parts:

- **Positive selection** - where each thymocytes gains an antigen (protein compound that react to a substance or pathogen) and those who are able to produce a suitable reaction with the major histocompatibility complex (MHC) survive move towards the next stage;
- **Negative selection** - where the thymocytes are exposed to self-antigens (antigens that react against the organisms cells) and those that react die by apoptosis.

A thymocyte that successfully completes these two selections matures as a T cell and can exit the thymus. We'll dwell deeper into this maturation in chapter 2.

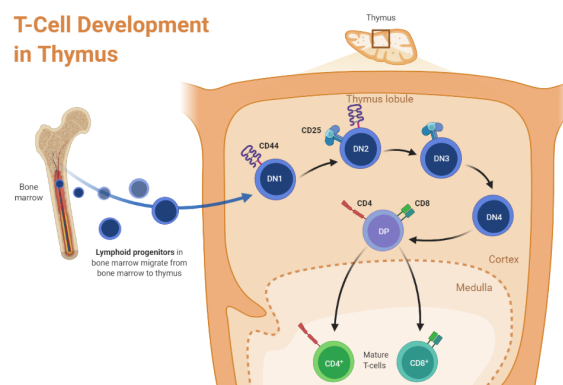


Figure 1.4: Thymus T cell development, (image created with the aid of BioRender.com)

The maturation of the thymocytes is thus a crucial point in the development of the **immunotolerance** mechanism and important to be studied in both healthy subjects and patients with diseases associated to this subset of cells.

A good pathway to uncover this development process is to use Next Generation Sequencing (NGS) techniques and uncover new sources of data about these cells such the Genetic, Genomic and Transcriptomic data and study patterns in them to understand their biological significance. In a **multi-omics** approach, where data from several approaches (genomics, transcriptomics, proteomics,...) is gathered and analysed, a better overview of the cell development can be achieved.

Although there are some studies in *tTregs* NGS data with *Mus musculus* such as Hu et al. (2018), an extensive study of thymic CD4+ T cells in humans didn't exist until the project in which this dissertation is integrated existed.

1.3 Objectives

This dissertation is integrated in a project being developed at "AESousaLab" at the Instituto de Medicina Molecular (IMM-FMUL). This project, named "Decoding Genotype-Phenotype correlation in Immune Complex Disorders through the Gene Regulatory Landscape of CD4 T Cells" aims to decode the development stages of CD4 T cells in humans through the usage of NGS data and computational biology techniques, which are an essential cell group to the maintenance of immune tolerance mechanisms. The main goals of the project go as follows:

1. Generate the Gene Regulatory Landscape (GRL) of Human CD4 T Cells;
2. Apply the GRL to uncover Genotype/Phenotype correlation in Complex Immune Disorders.

The dissertation integrates in the first stage of the project and where several tasks required the usage of computational techniques in the data science sphere. The dissertation will highlight 3 points, which techniques were used in each and the results obtained.

The 3 points of focus for this dissertation are:

1. Using Linear Regression Modelling accounting for Heteroscedasticity to model the relationship between gene expression and transcription factor accessibility in *CD4 + TCells*;
2. Standardization of the Gene Set Enrichment Analysis, its application to gene expression data in *CD4 + TCells* using the whole MSigDb Database;
3. Clustering binding data of *CD4 + TCells* to uncover patterns between genes and Transcription Factor Binding Sites(TFBS).

The work developed under the dissertation resulted in a paper being submitted in a near future from the AESousaLab at IMM. The current provisional name for the paper is "Differential Binding uncovers key transcriptional modules defining regulatory T-cell identity in the human thymus".

1.4 Dissertation Organization

This dissertation is organized as follows:

- This first chapter presents the problem being studied in a general view, as well as why it should be studied with the help of data science;
- The second chapter describes the theoretical biological and technological background behind the study of CD4+ T regulatory cell development;
- Chapter 3 is dedicated to the Methodology behind the dissertation, discussing in depth the techniques used in the 3 major points discussed;
- Chapter 4 presents and discusses the numerical and graphical results obtained while discussing the meaning of those results within the biological reality;
- The final chapters summarize the main conclusions obtained in this work and provide some ideas for future work.

Chapter 2

Theoretical Background

In this section we'll discuss the Theoretical Background behind this project.

We'll start first by understanding the major area of study in which the project is inserted, Clinical Immunology in section 2.1, then in section section 2.2 we'll discuss how acquired immunity develops and the importance of studying t cell development. Then in section section 2.3 we'll discuss what Computational Immunology is and how it can address the problems underlying the project.

2.1 Clinical Immunology

Clinical immunology is the study in depth of disease caused by disorders of the immune system (such as failure, aberrant actions and malignant growth of the cellular elements of the system) and the mechanisms subjacent to those disorders. It can also study diseases from other systems such as diabetes, where immune reactions can play a part in the pathology and clinical features of the diseaseChapel et al..

The diseases studied by clinical immunology usually fall within 3 categories:

- **Immunodeficiency**, in which part of the immune system fails to create an adequate response (in this group we have diseases such as chronic granulomatous disease and primary immune diseases);
- **Auto-immunity** in which the immune system attacks it's own host cells (in this group with have diseases such as Lupus, rheumatoid arthritis and Hashimoto's disease);
- **Various hypersensitivities** in which the immune system responds inappropriately to otherwise harmless compounds (in which we find asthma and other allergies).

Clinical immunology began as sub-speciality of Internal Medicine or Paediatrics but soon became a research area on its own. It also studies acquired immunodeficiencies such as AIDS and ways to prevent the immune system to destroy allografts (transplant rejection).

Within research environments, clinical immunology focus both on the mechanisms of immunology related diseases as well as the biological processes underlying them. Most of these teams are multidisciplinary, aggregating doctors, biochemists, microbiologists, molecular biologists, computational biologists and others. They conduct basic, translational and clinical aimed at understanding and treating these complex diseases.

2.2 Acquired Immunity and T cell Development

One of the major areas of study within clinical immunology are the mechanism underlying acquired immunity.

Acquired immunity, in a simple definition Tate and Seeley (2009), is the subsystem of one's immune system that develops over the person's lifetime.

It includes both humoral immunity and cell mediated immunity components both used in destroying non self entities in the organism, namely pathogens or cancer.

The main characteristic of acquired immunity that distinguishes it from innate immunity is the specificity of its response allied to capabilities to memorize previous attacks of a pathogen, i.e., it builds a immune response specific to an attack and memorizes that response so it can trigger it faster and more effectively if such attack is repeated Tate and Seeley (2009). It's fine mesh of cells and molecules and their communications that are mostly mediated by two major groups of cells that together constitute the lymphocytes, B cells and T cells Figure 2.1.

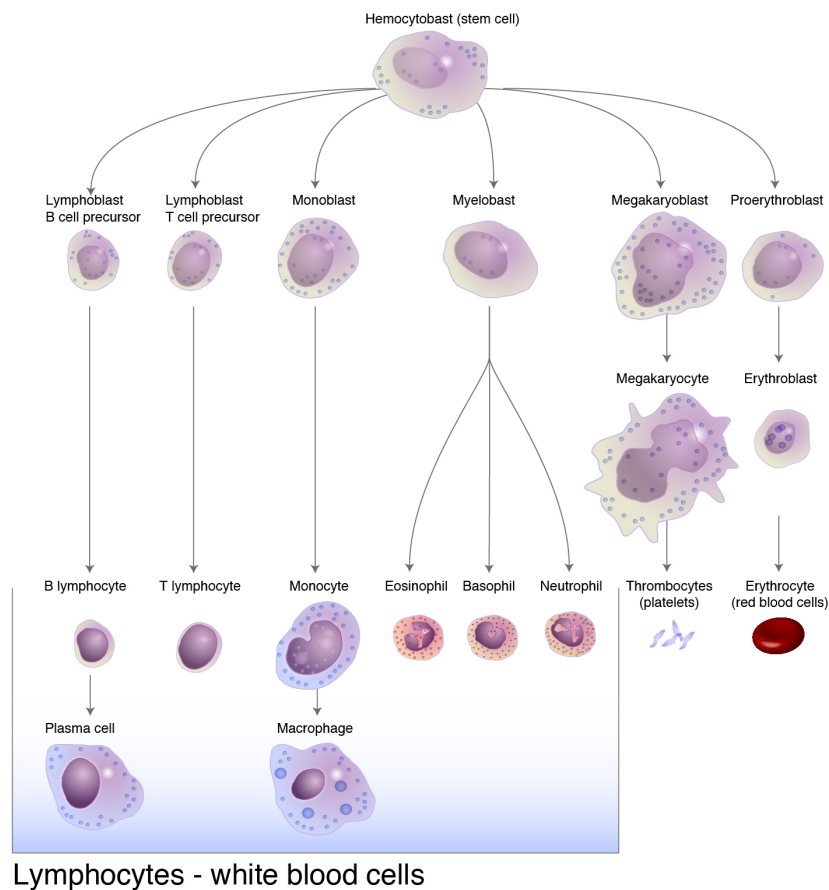


Figure 2.1: Major Cell Groups of the Blood Lymphocyte. Acquired Immunity is mediated by the branches of the Lymphoid B and T cell precursors . Image from <https://www.genome.gov/genetics-glossary/Lymphocyte>

The normal population values for the blood lymphocyte in a human Edgar (2011)

is comprised by:

- 70-90% T cells;
- 5-10% B cells;
- 1-10% Natural Killer (NK) cells

All of them are derived from lymphoid stem cells in the bone marrow Germain (2002). Lymphocyte subsets are defined by their expression of surface markers named CD antigens (CD for Cluster of Differentiation).

- T cells are classified by CD3+
 - T helper cells are classified by CD3+CD4+;
 - Cytotoxic T cells are classified by CD3+CD8+;
- B cells are classified by CD19+;
- NK cells are classified by CD16+CD56+

T cells as a whole are distinguished by the presence of the TCR receptor and are mostly known as a whole regulators of the immune response and responsible for collecting specific immune responses in antigens.

This project aims to study a specific subgroup of T helper cells, known as Regulatory T cells, that plays a crucial task of regulating the immune response due to their suppressive behaviour.

2.2.1 Regulatory T Cell Development and its Importance

To fully understand the importance of this subgroup of CD4+ T cells it's important to understand its development.

Regulatory T Cells express at the surface the biomarkers *CD4*, *FOXP3* and *CD25*. Due to conventional T cells also expressing *CD4* and *CD25*, makes this subgroup specially difficult to study Singh et al. (2013)Hori et al. (2017).

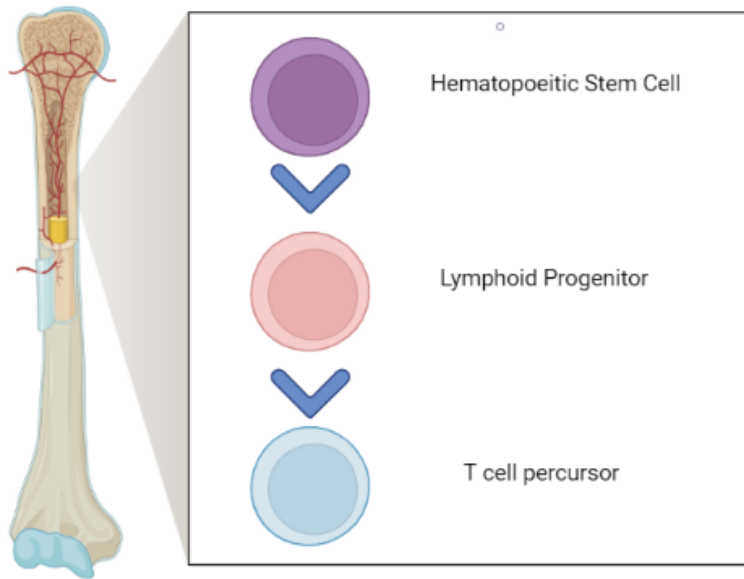
The main function of this subgroup is to suppress immune response of other cells. The suppressive function of this group is crucial to act as a "self check" built into the immune system to prevent excessive reactions, balancing the inflammatory and anti-inflammatory response.

The development of T regulatory cells starts in the bone marrow where the haematopoietic pluripotent stem cells transforms into lymphoid progenitor and in place the lymphoid progenitor transforms itself into T cell precursor (the cell lineage that gives rise to all T cells) Figure 2.2.

The T cell precursors migrate to the thymus for their second stage of development. In the thymus they undergo their second stage of development Silva et al. (2017).

Committed lymphoid progenitors arise in the bone marrow and migrate to the thymus Figure 2.3.

- Early committed T cells lack expression of T-cell receptor (TCR), CD4 and CD8, and are named double negative (DN; no CD4 or CD8) thymocytes;



Created in BioRender.com

Figure 2.2: Initial stages of T cell Development of T cells in the Bone Marrow. Image created with the help of www.biorender.com

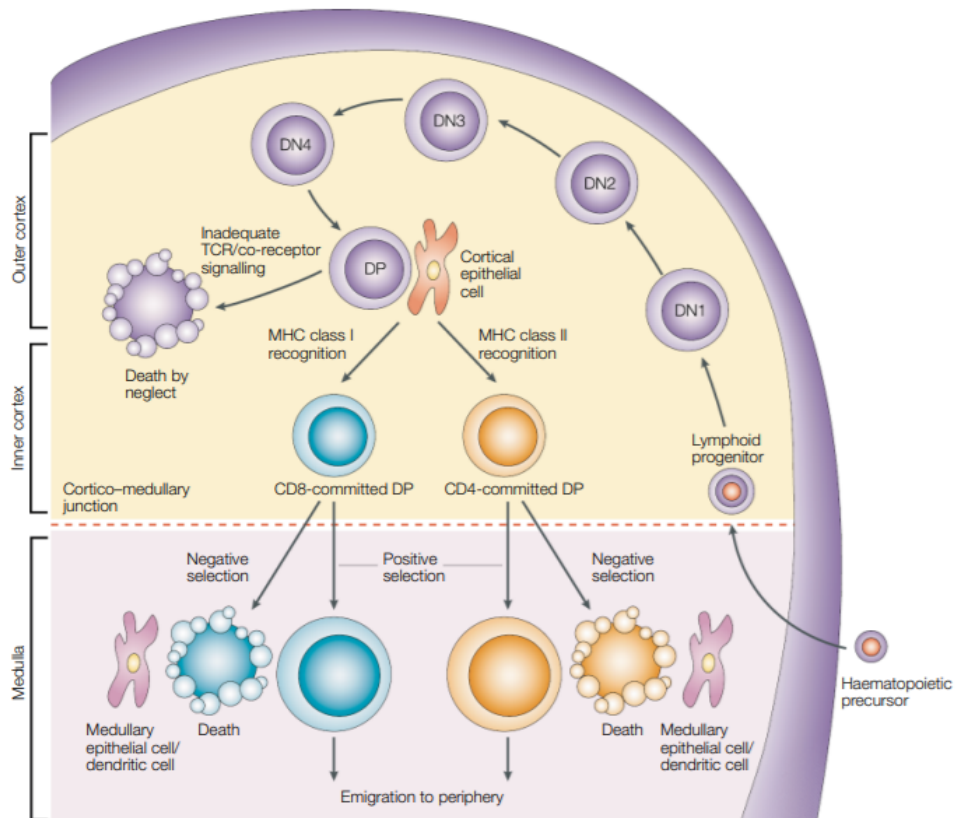


Figure 2.3: Development of T cells in the Thymus. Image retrieved from Germain (2002)

- DN thymocytes can then be further subdivided into four stages of differentiation (DN1, $CD44 + CD25-$; DN2, $CD44 + CD25+$; DN3, $CD44 - CD25+$; and DN4, $CD44 - CD25-$);
- As cells go through the DN2 to DN4 stages, they start expressing the pre-TCR, which is composed of the non-rearranging pre-T α chain and a rearranged TCR β chain;
- Successful pre TCR expression leads to substantial cell proliferation during the DN4 to double positive (DP) transition and replacement of the pre TCR α chain with a newly rearranged TCR α -chain, which yields a complete $\alpha\beta$ TCR;
- The $\alpha\beta$ -TCR + $CD4 + CD8+$ (DP) thymocytes then interact with cortical epithelial cells of the thymus that express a high density of MHC class I and class II molecules associated with self-peptides;
- The fate of the DP thymocytes depends on signalling that is mediated by interaction of the TCR with these self-peptide-MHC ligands:
 - Too little signalling results in delayed apoptosis (death by neglect);
 - Too much signalling can promote acute apoptosis (negative selection);
- The appropriate, intermediate level of TCR signalling initiates effective maturation (positive selection);
- Thymocytes that express TCRs that bind self-peptide-MHC-class-I complexes become $CD8 + Tcells$, whereas those that express TCRs that bind self-peptide-MHC-class-II ligands become $CD4 + Tcells$;
- These cells are then ready for export from the medulla to peripheral lymphoid sites. SP, single positive.

A small subset of the $CD4+$ T cells goes then to express *FOXP3* and constitutes the subset thymic T regulatory cells. Íris Caramalho et al. (2015) After this they move into the periphery and until they come in contact with the antigen, they stay N ave T cells. During a person's lifetime, a reservoir of N ave and Memory T regulatory cells is maintained by the organism to keep immune regulation in check.

As this subset of cells is associated with regulation of immunity, there's a considerable association between them and several immune system pathologies Kond elkova et al. (2010). It has been associated to Complex Variable Immunodeficiencies Silva et al. (2019), HIV Godinho-Santos et al. (2020) and in tumour progression.

This turns the study of their development into a crucial task. Due to the difficulties in clearly isolating this cell subset, a lot of questions are yet to be answered, namely the cell development changes they undergo.

This project aims to combine the data acquisition power of today's genetic and genomic techniques with the data science power of computational methods to understand the various "-omics" levels (genome, epigenome, transcriptome,...) of $CD4+$ T reg cells and unveil a bit more the intricate cell "ballet" that creates this subset, with hopes that it will lead to potential treatment targets for the pathologies mentioned.

2.3 Computational Immunology

Computational immunology (or systems immunology) involves the development and application of bioinformatics methods, mathematical models and statistical techniques for the study of immune system biology. The field's main aim is to convert immunological data into computational problems, solving them using mathematical and computational approaches and then convert the results into immunologically meaningful interpretations.

Its applications span from cancer informatics, allergies, infectious diseases and host responses and Immune system function.

Although the beginnings of this area can be traced back to a century ago to the very first theoretical models of malaria Ross (1916) it experienced a boom in the 90s and 2000's during the tech boom with the first systematic immunology related databases Petrovsky and Brusic (2002) and a second boom in the 2010's due to the onset of Next Generation Sequencing Techniques and the increasing accessibility of the techniques Davis et al. (2017).

The area studies at all levels of clinical immunology and its success depends on this contribution from the clinic to the laboratory Figure 2.4 where data acquisition, data processing and information systems techniques are crucial.

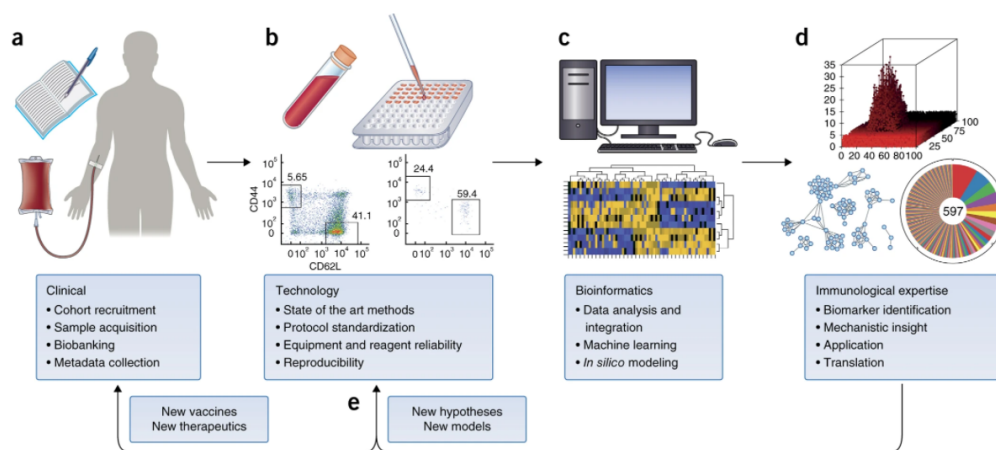


Figure 2.4: Cross-disciplinary efforts have allowed considerable advances in human medical research, from the clinic (a) to technology (b) to bioinformatics (c) and the laboratory (d), it's the collaboration of all that moves Computational Immunology forward Davis et al. (2017))

Computational Immunology encompasses many areas such as imaging, clinical data, allergy studies and mathematical modelling. In this project we will discuss the crossing between Genetics and Genomics techniques to study Immune system function and development and the multitude of techniques used in this area which can be seen in Figure 2.5.

By crossing multiple techniques which address various aspects of the cell such as genome, gene expression, protein expression and others hopefully the fine tuned orchestra that's happening inside CD4⁺ Treg cells will be unveiled.

This project uses RNAseq to establish gene expression and ATAC-seq to establish Chromatin accessibility (one of the most important measures of the epigenome) to start this task.

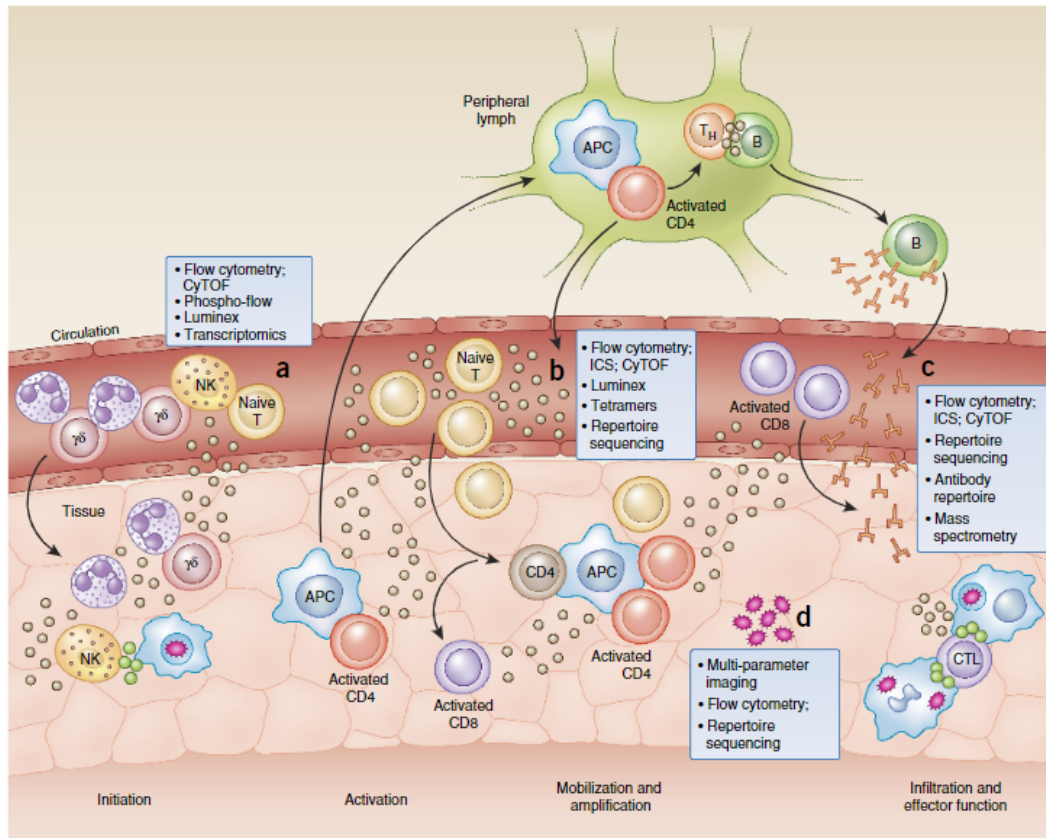


Figure 2.5: Laboratory and Computational Techniques used to study different parts of the immune system. In special interest is the **D** group where we see Repertoire sequencing being used to study CD4+ T cells. Image retrieved from Davis et al. (2017))

2.3.1 Genetics and Genomics in Computational Immunology

Immunology relies a lot on cell lines and animal models, for obvious ethical reasons. However translating discoveries to our own human immunology reveals itself to be quite hard Dheilly et al. (2014) as cell to cell interaction is crucial for immune function and comparing our own immune system with the one from model species doesn't always translates.

Next Generation Sequencing techniques might be the answer to this problem. Consistently acquiring genomic, transcriptomic and epigenomic data at an unprecedented scale at affordable rates allows us to have an overview of the events in each cell type being study and determine crucial mechanisms of regulation by doing comparative analysis of NGS data between cell types or stages of development.

The onset of Next Generation Sequencing (NGS) Techniques in 2004 commercially Slatko et al. (2018) which allows for sequencing efforts that are more accessible and more precise, has brought a new push in the usage of genetics and genomics efforts to study cell development.

It's also known as Massive Parallel Sequencing as the common protocol goes as follows

1. DNA sequencing libraries are generated by clonal amplification by PCR in vitro.;

2. The DNA is sequenced by synthesis, such that the DNA sequence is determined by the addition of nucleotides to the complementary strand rather than through chain-termination chemistry;
3. the spatially segregated, amplified DNA templates are sequenced simultaneously in a massively parallel fashion without the requirement for a physical separation step.

This methodology allows for a broad range of studies aiming at studying different components of genetics and genomics, varying just the molecule and protocol studied Slatko et al. (2018): **Whole Genome Sequencing** and **Whole Exome Sequencing** target sequencing of the genome, **RNAseq** targets gene expression, **ATACseq** targets opening of the chromatin and so on...

Targetting multiple dimensions of the genetic-genomic landscape by applying different techniques then allows to collect multiple dimensions of the same cell group being studied.

These techniques produce high amounts of data, with a high variance (as variance is inherent between organisms) and often without a big amount of replicates making them ideal candidates for analysis with data science techniques.

Chapter 3

Methodology

In chapter 2 we described the biological and computational backgrounds behind the study of T cell development. In this chapter we discuss the methodology and techniques used in this project.

First, we discuss in a overview the techniques and methods used to obtain and clean this data in section 3.1 in order to understand the origins of this data.

Then we discuss the statistics behind the discovery of the existence of a linear correlation between Differential Chromatin Accessibility (DCA) and Gene Expression in tTregs in section 3.2.

Next we discuss the methodology used behind the standardization of Gene Set Enrichment Analysis(GSEA) to run with the full msigDb database on section section 3.3

Finally, at section 3.4 we discuss in depth the protocol created that originated the clustering analysis of the digital footprinting results.

3.1 Multiomics Data: Extraction of thymic T Cell Data and Pre Preparation

This project assumed a multiomics approach that involved extracting mainly gene expression (RNA_{seq}) and chromatin accessibility data ($ATAC_{seq}$) allowing us a diverse overview of the tTreg cell development.

3.1.1 Cell Sorting and Selection

Biological replicates were extracted from CD4 single-positive thymocytes, isolated from thymuses obtained after paediatric cardiac surgery of three different individuals. Mature thymocytes were sorted (as seen in Figure 3.1) and purified as $TCR\alpha\beta_{++}$, $CD4_{++}$, $CD8_{-} CD27_{+}$. Cells were further purified into conventional (tTconv) and regulatory (tTreg), defining tTregs as $CD25_{+}$ and $CD127_{low}$ as seen in Figure 3.2.

3.1.2 RNA_{seq} and Differential Expression

RNA samples were extracted from tTregs and tTconvs, as explained in subsection 3.1.1. Libraries were built by BGI, selecting for polyadenylated RNA after depleting ribosomal fraction and then sequenced by high-throughput parallel sequencing (RNA_{seq}) in

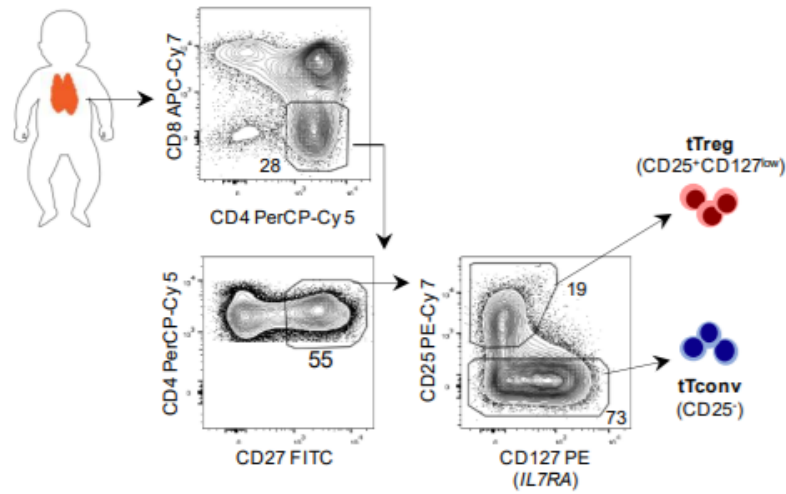


Figure 3.1: Strategy for sorting tTregs and tTconvs from human thymuses collected during routine corrective paediatric cardiac surgery. Mature CD4 single-positive thymic Tregs (tTregs) and their conventional counterparts (tTconvs) were sorted using CD25 and CD127.

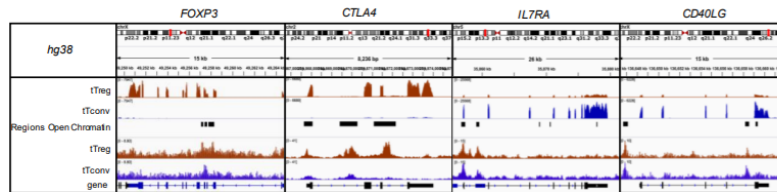


Figure 3.2: Representative profiles of raw RNA_{seq} gene expression of emblematic genes of tTregs (*FOXP3* and *CTLA4*) and tTconvs (*IL7RA* and *CD40LG*) paired with the Accessibility to Chromatin Data ($ATAC_{seq}$ data) within their genomic domains. Top Row indicates their location in their respective chromosome. "Regions of Open Chromatin" row indicates detection of regions with significant $ATAC_{seq}$ signal enrichment, tTreg signal is depicted in red, tTconv signal in blue. For the gene row, black depicts sense direction, blue depicts antisense direction

a Illumina $Hiseq^{4000}$ sequencer. Raw sequencing data was processed and analysed with appropriate tools, such as *samtools* Danecek et al. (2021) using the High-Performing Computer cluster iMM-LOBO, with quality control of reads made with FastQC Andrews (2010). The resulting ca. 200 million paired-end reads per biological replicate (PE100) were uniquely mapped and annotated to the human genome (hg38) with "TopHat" Kim et al. (2013) and transcript expression was quantified with R package "HTSeq" (Count Per Million, CPM), with exclusion of genes with less than 1 CPM in more than 2 libraries. Before determining the Differential Expression between tTregs and tTconvs with R package "edgeR", all libraries were scaled by Trimmed Mean of M-values (TMM) normalisation and corrected for heterogeneity of samples specific to contrast matrix with weighted scaling based on voom-limma (R package "limma"). Finally, we fitted multiple linear models by lmFit ("limma").

Conversion between annotations was made with “biomaRt”.

Differential Gene Expression threshold set between tTregs and tTconvs at $\log_2FC > \pm 2$, with FDR < 0.05.

3.1.3 *ATAC_{seq}* and Differential Chromatin Accessibility

ATAC_{seq} was performed following the Omni-ATAC protocol Corces et al. (2017) with minor modifications. Three biological replicate samples per cell type were extracted from three distinct healthy thymuses, in same conditions and as described in subsection 3.1.1. 5×10^4 sorted tTreg or tTconv cells were lysed for 3 minutes on ice, in 50 μ L of ATAC-Resuspension Buffer (10mM Tris-HCl pH 7.4, 10mM NaCl, 3mM MgCl₂) containing 0.1% NP40, 0.1% Tween-20, and 0.01% Digitonin. *tn5* tagmentation was performed using TDE1 Enzyme and Buffer TD (Illumina) at 37°C for 30 minutes, shaking at 1000rpm. After purification with a MinElute PCR Purification Kit (Qiagen), samples were amplified with NEBNext High Fidelity 2x PCR Master Mix (New England Biolabs) with index adapters from Buenrostro et al. (2015).

Final PCR reaction was then purified with a MinElute PCR Purification Kit followed by size-selection (150bp-1000bp using Ampure XP beads (Beckman Coulter). Sequencing was performed using a MGISEQ-2000 (BGI-Shenzhen, China), yielding a total sequencing depth between 200 and 600 million PE50 reads.

To identify the Regions of Open Chromatin (ROCs) and determine Differential Chromatin Accessibility raw sequencing read quality was assessed for quality using FastQC. Reads were uniquely mapped to hg38 using Bowtie2 Langmead and Salzberg (2012) and adapted for peak calling by MACS2 Zhang et al. (2008) using in-house pipeline, namely by converting to appropriate formats and correcting *tn5* shift. MACS2 command with the following parameters:

```
macs2 callpeak -t ${bam} -f BAMPE -g hs -q 0.05 --nomodel \
--extsize 200 --shift -100 -n ${bam} --outdir PEAKS
```

Peaks from all samples were merged to create the total landscape of Regions of Open Chromatin and we used PeakAnalyzer Salmon-Divon et al. (2010) to annotate these peaks to Nearest TSS using GTF annotation for hg38. To determine chromatin accessibility and its variation between tTregs and tTconvs (Differential Chromatin Accessibility), we used the same tools, method, normalisations and rescaling of *ATAC_{seq}* sequence libraries as for *RNA_{seq}* libraries, with the Peak_ID of each Region of Open Chromatin as the anchor for signal computation.

3.1.4 Digital Genomic Footprinting and Transcription Factor Binding analysis

For Digital Genomic Footprinting, transcription factor motifs within ROCs were identified using the Positional Weight Matrices (PWMs) in the JASPAR Core database Fornes et al. (2020) Khan et al. (2018). We selected 639 motif profiles matching “Homo Sapiens species” + “Latest Version”.

We used the TOBIAS framework 0.12.6 Bentsen et al. (2020) 3.3 to perform read bias correction of the list of ROCs using ATACCorrect, calculation of continuous footprint scores

TOBIAS - Transcription factor Occupancy prediction By Investigation of ATAC-seq Signal

pypi v0.12.11
downloads 217/month
install with bioconda
Maintained? yes
Publication NatComm

Introduction

ATAC-seq (Assay for Transposase-Accessible Chromatin using high-throughput sequencing) is a sequencing assay for investigating genome-wide chromatin accessibility. The assay applies a Tn5 Transposase to insert sequencing adapters into accessible chromatin, enabling mapping of regulatory regions across the genome. Additionally, the local distribution of Tn5 insertions contains information about transcription factor binding due to the visible depletion of insertions around sites bound by protein - known as *footprints*.

TOBIAS is a collection of command-line bioinformatics tools for performing footprinting analysis on ATAC-seq data, and includes:

- Correction of Tn5 insertion bias
- Calculation of footprint scores within regulatory regions
- Estimation of bound/unbound transcription factor binding sites
- Visualization of footprints within and across different conditions

For information on each tool, please see the [wiki](#).




Figure 3.3: Tobias, the package used for Digital Genomic Footprinting, which can be found in <https://github.molgen.mpg.de/pages/loosolab/www/software/TOBIAS/>

across accessible chromatin regions with ScoreBigWig (which can be seen in the framework depicted in 3.4), followed by classification as bound/unbound (p-value < 0.01) state for transcription factor binding sites (TFBS) across both cell populations and calculation of differential binding as the fold-change between the footprint scores of the two cell types. The differential binding scores and p-values between tTregs and tConvs are represented as a volcano plot and were obtained using the BinDetect module. TFs with $-\log_{10}(\text{p-value})$ above the 95% quantile or differential binding scores smaller/larger than the 5% and 95% quantiles (top 5% in each direction) are colored and shown with labels.

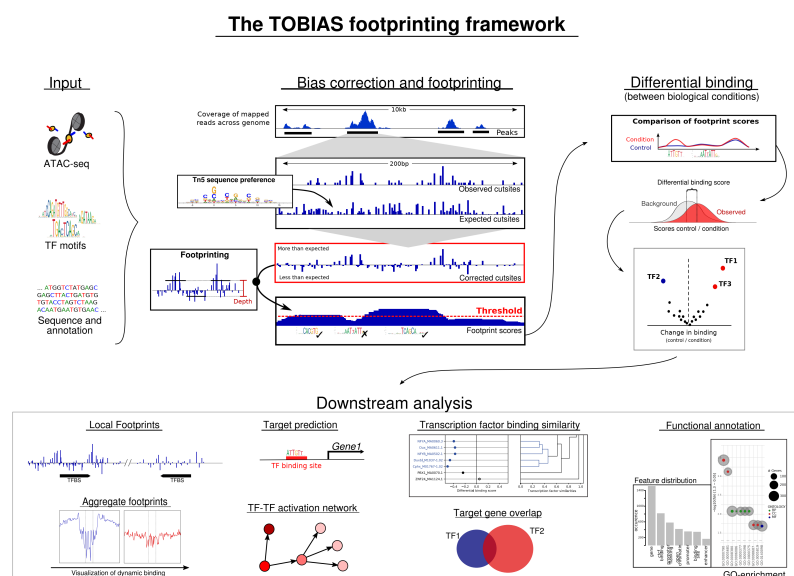


Figure 3.4: The TOBIAS framework, ScoreBigWig is represented by Differential Binding Analysis.

Aggregate footprints were created by aligning the genomic signals on the 200bp region surrounding the binding sites, with the aggregate signal being the mean of the score

on each bp.

After Digital Genomic Footprinting Analysis, we've obtained new data points, namely **treg_score** The footprinting score within treg cells for a specific TFBS, **tconv_score** The footprinting score within tconv cells for a specific TFBS and **treg_tconv_log2fc** (further called diffbinding) which is the log2 fold change between the footprinting scores of treg and tconv cells, telling us whether the TFBS was predicted to be more or less bound between the cells (positive equals more bound to treg, negative equals more bound to tconv).

3.2 Differential Expression vs Accessibility of The Chromatin

From the *RNAseq* data extracted as described in subsection 3.1.2 and the *ATACseq* data extracted as described in subsection 3.1.3 we can obtain a dataset that pairs Gene Expression values of a specific gene with ROCs associated to said gene identified by differential chromatin accessibility (DCA).

To assess if any relationship between chromatin accessibility (DCA) and gene expression (\log_2FC), the two dimensions were plotted with the help of ggplot2 Wickham (2009), R's most well known data visualization library.

	DCA	FC	hgnc_symbol	gene_biotype	Peak_ID
1	-0.091298284	389.64903	PCDH7	protein_coding	Peak_116441
2	0.188584102	389.64903	PCDH7	protein_coding	Peak_116450
3	-0.230985613	389.64903	PCDH7	protein_coding	Peak_116433
4	0.159802621	389.64903	PCDH7	protein_coding	Peak_116442
5	0.007503275	389.64903	PCDH7	protein_coding	Peak_116438
6	0.059005701	389.64903	PCDH7	protein_coding	Peak_116431
7	-0.442502506	389.64903	PCDH7	protein_coding	Peak_116447
8	-0.240613697	389.64903	PCDH7	protein_coding	Peak_116427
9	0.023739178	389.64903	PCDH7	protein_coding	Peak_116440

Figure 3.5: Raw Data to extract Differential Chromatin Accessibility and Gene Expression info from

As each gene has 1 or more ROCs (with some having more than 50 ROCs) the values for Differential Chromatin accessibility were reduced to the mean by gene and the number of ROCs kept to ease the visualization.

Before the visualization the data looked as follows Figure 3.6.

	hgnc_symbol	gene_biotype	logFCA	FCA	mean	meanslinear	median	sdev	counts	FactorFCA	FCAtype
4944	PCDH7	protein_coding	8.606031	389.649027	-0.013832646	1.0111920	0.022293719	0.303999828	30	1058	postFCA
3972	LRRRC32	protein_coding	7.303357	157.953650	0.644276286	1.5708758	0.650544918	0.177967331	3	1057	postFCA
1579	CPE	protein_coding	7.120521	139.152309	-0.222687759	0.8873526	-0.292128837	0.395819006	15	1056	postFCA
902	BTNL8	protein_coding	7.004244	128.377105	1.188753874	2.4559492	0.822632711	0.660629599	3	1055	postFCA
3257	IL1R1	protein_coding	6.947287	123.407605	-0.304861593	0.8130078	-0.303771149	0.163966216	3	1054	postFCA
6574	STAC	protein_coding	6.835038	114.169861	-0.119854957	0.9314349	-0.186557542	0.228039488	14	1053	postFCA
5932	RVR1	protein_coding	6.689645	103.224737	0.563846522	1.7753655	0.517918772	0.916884522	6	1052	postFCA
2476	FAT3	protein_coding	6.637009	99.526502	-0.063053542	0.9761793	-0.045994225	0.318211831	5	1051	postFCA
3469	KCN53	protein_coding	6.530584	92.448854	-0.130686934	0.9236887	-0.120296739	0.229463713	10	1050	postFCA
332	AKR1GCP	unprocessed_pseudogene	6.362011	82.151298	0.625738965	1.6315677	0.298066650	0.532132890	5	1049	postFCA
3272	IL2RA	protein_coding	6.327809	80.326796	0.253148347	1.2032908	0.216252574	0.211403641	8	1048	postFCA

Figure 3.6: Data cleaned for input in data visualization of Gene Expression vs Differential Chromatin Accessibility

A first attempt was at plotting the data as a bubble plot with the X_{axis} , the Fold Change values, were set as factor and set to ascending order as seen in 3.7. From the beginning it was decided to keep **down regulated genes in tTregs in blue** (to set as examples for tTconv cells) and **up regulated genes in red**. This colour scheme was kept in all visualizations for ease of reading and interpretation.

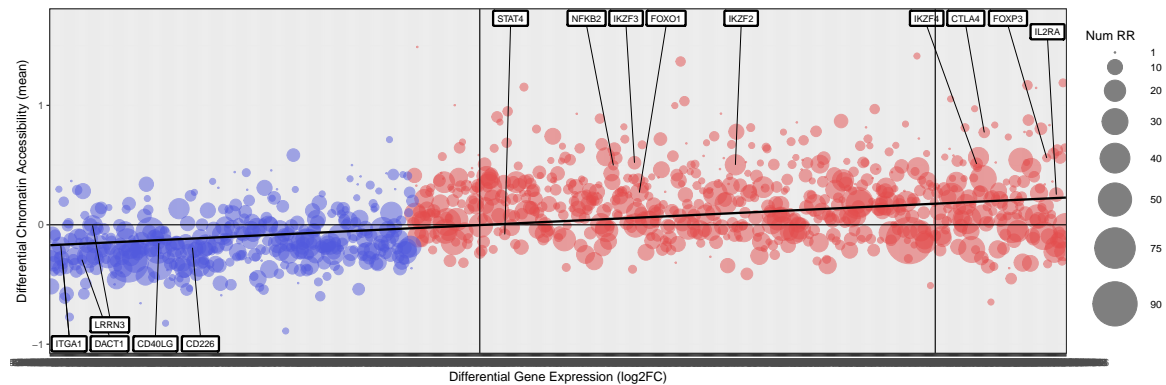


Figure 3.7: Differential Gene Expression in x and Differential Chromatin Accessibility in y

The figure 3.7 revealed that a conversion of Differential Gene Expression to log was warranted in order to clearly separate between down regulated genes in tTreg (in blue) and up regulated genes in tTreg (in red). This visualization also gave a first hint that a linear regression might exist between these 2 variables.

Fold Change was therefore transformed from linear to logarithmic and Linear Regression was calculated between gene expression and Differential Chromatin Accessibility. After a visual assessment, it was verified that this was a case of heteroscedasticity with the Breusch-Pagan Test and *lmrob()* from the *robustbase* package Maechler et al. (2021) was used to obtain the linear regressions values accounting for the existence of heteroscedasticity.

3.3 Standardization of Gene Set Enrichment Analysis

To explore the gene ontology of the data we possessed we have explored a few algorithms that provide gene ontology information such a Gene Ontology Project enrichment analysis Mi et al. (2019) and the Camera algorithm Wu and Smyth (2012) but in the end settled for the Gene Set Enrichment Analysis Algorithm (GSEA) Subramanian et al. (2005).

The code for this part of the project can be found in https://github.com/theinsilicobiology/fgsea_msigDB_Thymus_paper.

The basic Gene Set Enrichment Analysis algorithm should go roughly as follows according to its original proposal in Subramanian et al. (2005) and depicted in Figure 3.8.

1. Calculate the Enrichment Score (ES) that represents the amount to which the genes the given set are over-represented at either the top or the bottom of the list. This score is a Kolmogorov-Smirnov like statistic;
2. Estimate the statistical significance of the ES. This calculation is achieved through a phenotypic based permutation test in order to produce a null distribution for the ES.

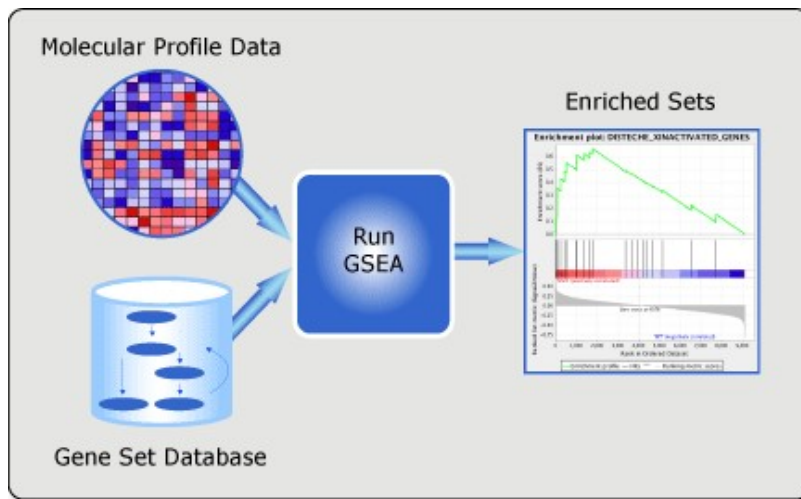


Figure 3.8: Diagram of the major stages of Gene Set Enrichment Analysis

The p -value is calculated in comparison with the null distribution;

3. Adjust for multiple hypothesis testing for when a large number of gene sets are being analysed at one time. The enrichment scores for each set are normalized and a false discovery rate is calculated

As the standard GSEA is slow to compute and not very sensitive when using small gene sets a variation of the algorithm, named Fast Gene Set Enrichment Analysis (fgsea) Sergushichev (2016) (<https://github.com/ctlab/fgsea/>) was chosen for the task. This variant of the algorithm is faster than the original, efficiently reusing one sample multiple times. This demonstrates the possibility of doing thousands of permutations in a small amount of time, leading to accurate p -values. It also allows the application of standard FDR correction procedures.

The algorithm goes as described in image Figure 3.9

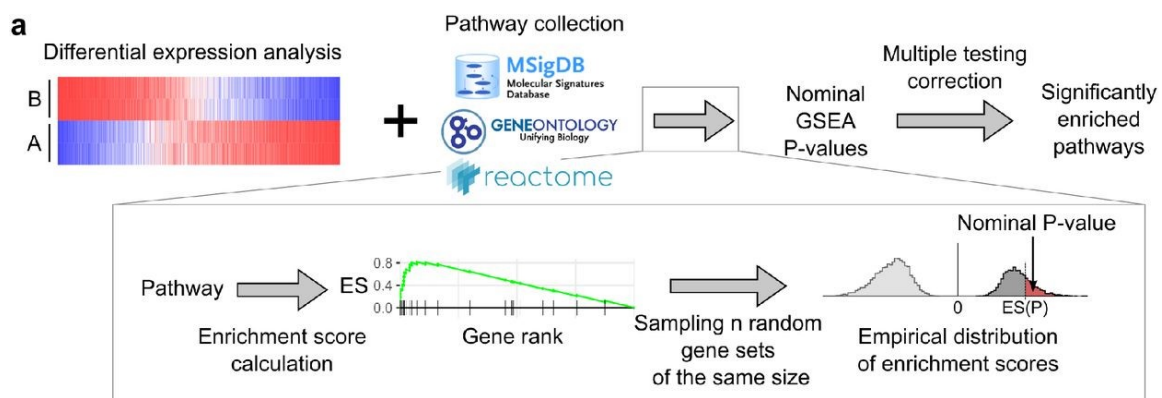


Figure 3.9: The FGSEA algorithm is depicted in the image. Image from Sergushichev (2016)

With the algorithm chosen a library of annotated gene sets to compare our own data was required. After a few tests, the mSigDB, a molecular signature database Subramanian et al. (2005), maintained by the same team that created the original GSEA algo-

rithm became the most appropriate choice. This database, which can be found in <https://www.gsea-msigdb.org/gsea/msigdb/index.jsp>, provides us with a variety of curated datasets that associate gene sets to certain phenotypes such as cancer, immunology, regulatory target genes or cell type signature gene sets. These collections come from various sources such as Ensembl BioMart, biomedical literature, BioCarta or KEGG (you can see the origins and details of each collection in https://www.gsea-msigdb.org/gsea/msigdb/collection_details.jsp).

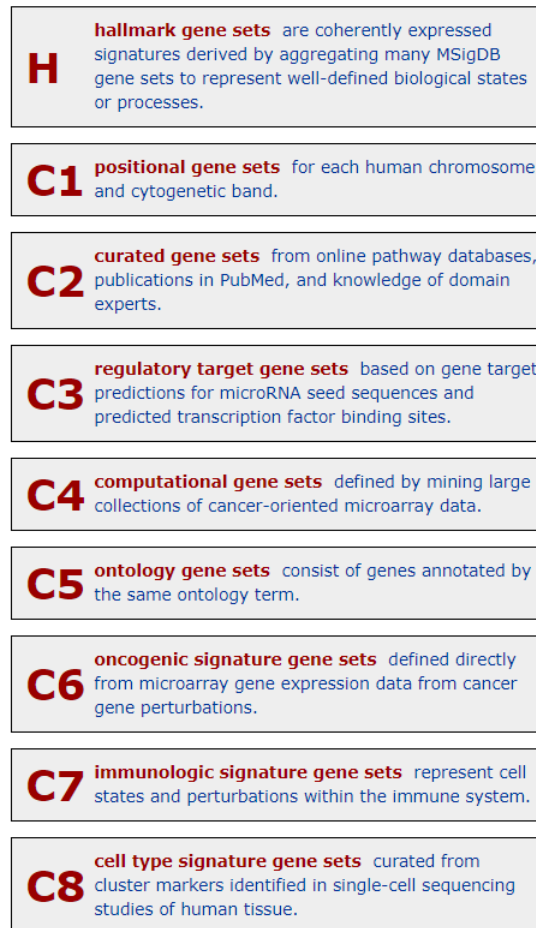


Figure 3.10: The collections existent in the mSigDB databaset. They can be found at <https://www.gsea-msigdb.org/gsea/msigdb/index.jsp>

To assure that no relevant results are forgotten, a standardized protocol to execute the FGSEA algorithm in all the datasets of the mSigDB database became important. Figure 3.10 An R project to execute this task so thus become crucial.

First, an function to standardize the execution of the fgsea protocol was developed. 3 outputs were chosen for this function, a table that systematizes Enrichment Scores, *p*-values and *leadingEdge* (genes in common between our input and a geneset from the mSigDB), a bar plot of the Normalized Enrichment Score for gene sets with a significant *p*-value and a sticks/barcode plot for the most enriched gene sets on both ends Figure 3.11.

With this function working reliably and without errors, a second function that runs the previous one in the whole mSigDB database was created Figure 3.12. To ease updates of the function, it was followed the order provided by the mSigDB website <https://www.gsea-msigdb.org/gsea/msigdb/index.jsp>

```

#function to execute GSEA
createGSEA<-function(statsLab, paths, genetabled, transition){
  library(stats)
  library(fgsea)

  library(tidyverse)
  library(dplyr)

  fgseaRes <- fgsea(pathways=paths, stats=statsLab,minSize=15,maxSize=500)

  #tidy it
  fgseaResTidy <- fgseaRes %>%
    as_tibble() %>%
    arrange(desc(NES))

  bardata<-subset(fgseaResTidy,fgseaResTidy$padj<=0.05)

  #barplot
  barname<-paste(transition,"/",genetabled,"/",genetabled,transition,"barplot.pdf", sep="")
  ggplot(bardata, aes(reorder(pathway, NES), NES)) +
    geom_col(aes(fill=(padj<0.05))) +
    coord_flip() +
    labs(x="Pathways", y="Normalized Enrichment Score",
         title= "NES from GSEA") +
    theme_minimal()
  ggsave(filename = barname, width = 20, height = 20)

  #dev.off()

  #dim(subset(fgseaResTidy,fgseaResTidy$padj<0.05))

  #sticks_plot
  topPathwaysUp <- fgseaRes[ES > 0, ][head(order(padj), n=10), pathway]
  topPathwaysDown <- fgseaRes[ES < 0, ][head(order(padj), n=10), pathway]
  topPathways <- c(topPathwaysUp, rev(topPathwaysDown))

  stickname<-paste(transition,"/",genetabled,"/",genetabled, transition,"stickstop10.pdf", sep="")

  pdf(file = stickname,h=10,w=12)
  plotGseaTable(paths[topPathways], statsLab, fgseaRes,gseaParam = 0.5)
  dev.off()

  return(fgseaResTidy)
}

```

Figure 3.11: R function created to run the fgsea protocol in a standardized fashion. The full function can be found in https://github.com/theinsilicobiology/fgsea_msigDB_Thymus_paper/blob/main/Functions/FunctionsForGSEA.R.

[//www.gsea-msigdb.org/gsea/msigdb/index.jsp](http://www.gsea-msigdb.org/gsea/msigdb/index.jsp) so it can be quickly updated when the database itself has updates.

Within each iteration of the function one of the collections from mSigDB is uploaded and the fgsea is calculated. Finally, a csv containing the results, the barplot with significant NES and the sticks plot with top enriched genes sets on both ends are produced. All outputs are arranged in folders thus organizing the outputs.

Finally, by observing the table output Figure 3.13 of the FGSEA algorithm we can observe 3 interesting columns. The *NES* column giving us the gene sets which have a more relevant enrichment score towards of data, the *padj* gives us which ones are significant and the *leadingEdge* where we get the genes in common in between our own data and each respective gene set.

Creating a visual way to observe this in interesting outputs of data/mSigDB collection was paramount. The final decision became a heatmap with genes in columns, gene sets in rows and the *NES* value as the value in the heatmap. Ordering this heatmap by *NES* we can then observe which gene sets are more enriched in our data and which gene sets share enrichment areas with our own data as seen in the example in page 24.

```

runGSEAonTest<-function(stats, nametransition){

  source("Functions/FunctionsForGSEA.R")

  library(fgsea)
  library(tidyverse)
  library(dplyr)

  toString(nametransition)

  # -----Tests -----

  # -- ALL Gene Libraries

  AllGenes<- gmtPathways("MSigDb/msigdb.v7.4.symbols.gmt")
  AllGenestable<- createGSEA(stats,AllGenes,"AllGenes", nametransition)
  #namingcsv
  genetabledused<- "AllGenes"
  #name csv
  tablename<-paste(nametransition,"/",genetabledused,"/",genetabledused,nametransition,"table.csv", sep="")

  #generate csv
  #AllGenestable <- subset(AllGenestable, padj<0.05)
  tibble_with_lists_to_csv(AllGenestable, tablename)

  # -- C1 positional

  C1pos<- gmtPathways("MSigDb/c1.all.v7.4.symbols.gmt")
  C1postable<- createGSEA(stats,C1pos,"C1pos", nametransition)
  #namingcsv
  genetabledused<- "C1pos"
  #name csv
  tablename<-paste(nametransition,"/",genetabledused,"/",genetabledused,nametransition,"table.csv", sep="")
  #generate csv
  #C1postable <- subset(C1postable, padj<0.05)
  tibble_with_lists_to_csv(C1postable, tablename)

  # -- Hallmark

  Hallmark<- gmtPathways("MSigDb/h.all.v7.4.symbols.gmt")
  Hallmarktable<- createGSEA(stats,Hallmark,"Hallmark", nametransition)
  #namingcsv
  genetabledused<- "Hallmark"
  #name csv
  tablename<-paste(nametransition,"/",genetabledused,"/",genetabledused,nametransition,"table.csv", sep="")
}

```

Figure 3.12: Function that runs the function Figure 3.11 iterating over the whole mSigDB database. The full function can be found in https://github.com/theinsilicobiology/fgsea_msigDB_Thymus_paper/blob/main/Functions/fgseaMsigDb.R.

pathway	pval	padj	logZerr	ES	NES	size	leadingEdge
(character)	(double)	(double)	(double)	(double)	(double)	(double)	(character)
HALLMARK_IL2_STAT5_SIGNALING	0.0006092495	0.01340349	0.47727082	0.4292047	1.9013415	66	IL2RA,IL1RL1,TNFRSF6,TNFRSF18,TNFRSF9,CSF1,TNFRSF4,CT...
HALLMARK_ESTROGEN_RESPONSE_LATE	0.0056646550	0.03870231	0.40701792	0.4988080	1.7745561	23	CPE,RAB31,PERP,LSR,CAV1,GFBR3,IGFBP4,FABP5,BATF,TMPR...
HALLMARK_INFLAMMATORY_RESPONSE	0.0045536778	0.03870231	0.40701792	0.4261441	1.7339229	41	PCDH7,IL1R1,TNFRSF9,EBI3,CSF1,CCL22,ICAM1,PTGER2,IL15...
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.0493119266	0.21697248	0.21654284	0.3651996	1.5168085	46	PTGS2,TNFRSF9,CSF1,ICAM1,CD83,MAP3K8,IL15RA,DUSP4,T...
HALLMARK_KRAS_SIGNALING_DN	0.0773333333	0.28355556	0.16470647	0.4616461	1.4836360	16	RYR1,CCR8,SPTBN2,FGFR3,TENT5C,CLDN16
HALLMARK_MYOGENESIS	0.1264080100	0.34920635	0.13649044	0.3684340	1.3405600	26	RYR1,LAMA2,FST,ACTN2,CASQ1,ADAM12,PLXNB2
HALLMARK_INTERFERON_GAMMA_RESPONSE	0.1286407767	0.34920635	0.13284630	0.3350748	1.3223376	37	PTGS2,IRF5,ICAM1,XCL1,IL15RA,CSF2R8,SECTM1,FGL2,IL2RB...
HALLMARK_ALLOGRAFT_REJECTION	0.1428571429	0.34920635	0.12384217	0.3238170	1.3073112	40	IL2RA,CSF1,CCL22,ICAM1,CD79A,LYN,TLR2,NCF4,PRF1,HDAC...
HALLMARK_KRAS_SIGNALING_UP	0.1753086420	0.38567901	0.11237852	0.3371282	1.2688735	30	CPE,RELN,PTGS2,HDAC9,BIRC3,TNFRSF1B,ARG1,PRDM1,MA...
HALLMARK_HYPOXIA	0.2172774869	0.39122040	0.10244941	0.3665518	1.2493633	20	HMOX1,HS3ST1,FBP1,CAV1,SDC4,S100A4,BHLHE40,GCNT2...
HALLMARK_APOPTOSIS	0.2197368421	0.39122040	0.10208011	0.3743343	1.2369854	17	HMOX1,LMNA,CAV1,HGF,PRF1,BIRC3,F2R,FAS,PMIAIP1,GAD...
HALLMARK_COMPLEMENT	0.2311756935	0.39122040	0.09923333	0.3642384	1.2284871	19	TMPRSS6,FN1,ACTN2,LYN,F5
HALLMARK_XENOBIOTIC_METABOLISM	0.2878980892	0.42225053	0.08455574	0.3218201	1.1490363	24	IL1R1,HMOX1,FBP1,AKR1C2,IGFBP4,ARG1,GCNT2,IRF8,FAS,A...
HALLMARK_P53_PATHWAY	0.5757961783	0.78401551	0.05029481	0.2566262	0.9162661	24	HMOX1,PERP,PLXNB2,S100A4,VDR,PTPN14,SESN1,RG516,F2...
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	0.6058301648	0.78401551	0.04773424	0.2516386	0.8952279	23	FN1,LAMA2,ADAM12,SDC4,GFBR4,FBN1
HALLMARK_APICAL_JUNCTION	0.7349869452	0.85103752	0.04024776	0.2316302	0.7975116	21	ACTG2,ICAM1,ACTN2,NEGR1,FBN1,TRAF1,LAYN,MDK
HALLMARK_ESTROGEN_RESPONSE_EARLY	0.7783505155	0.85618557	0.03699325	0.2179917	0.7643995	22	RAB31,IGFBP4,INPP5F,TMPRSS3,BHLHE40,NRIP1,KCNK3
HALLMARK_MTORC1_SIGNALING	0.8956406869	0.93829024	0.03149289	0.1854876	0.6256042	19	NIBAN1,TBK1,FGL2,BHLHE40,SQLE,GCLC,TFRIC,M6PR,SHMT2...
HALLMARK_MITOTIC_SPINDLE	1.0000000000	1.00000000	0.02527128	0.1013348	0.3488895	21	ARAP3,PREX1,CDC42BP4
HALLMARK_HEME_METABOLISM	0.6480000000	0.79200000	0.10395847	-0.2080587	-0.8591162	18	ADD2,ACSL6,DMTN

Figure 3.13: CSV output of Figure 3.11.

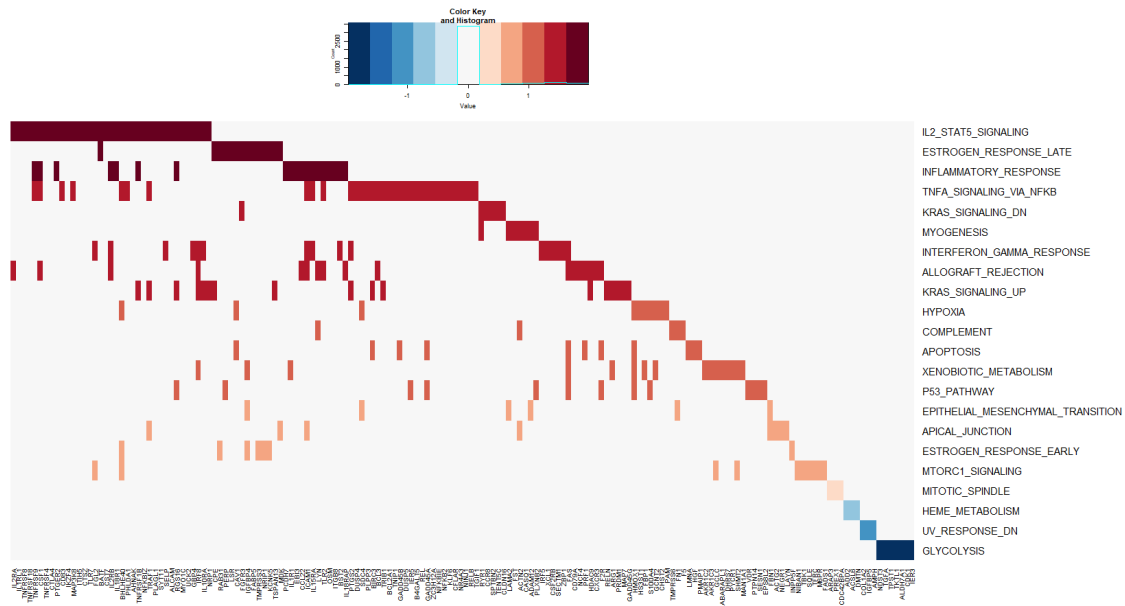


Figure 3.14: Heatmap to observe the results of FGSEA (with data originated in Figure 3.13) with the aid of heatmap.2() from the package *gplots* <https://github.com/talgalili/gplots>. The function that generates this heatmap is found in https://github.com/theinsilicobiology/fgsea_msigDB_Thymus_paper/blob/main/Functions/generateHeatmapCSVfgsea.R

The system created thus allows to create a standardized method that provides us with a way to analyse gene ontology of a gene expression dataset over a well maintained and varied collection of annotated gene sets and also provides us with easy to interpret results, visual when necessary.

3.4 Clustering TFBS/Gene Binding Patterns in tTreg/tTconv Cells

From the digital genomic footprinting analysis executed with the TOBIAS framework and described in subsection 3.1.4 we've obtained 3 new variables: *treg_score*, *tconv_score* and *diffbinding*.

The code for this part of the project can be found in https://github.com/theinsilicobiology/Kmeans_TOBIAS_CD4Thymus_paper.

These new variables lead us to a new question. Are there patterns in the relationship between tTreg signature genes and their respective Transcription Factor Binding Sites? To answer this question, clustering became the solution.

The data was obtained raw from TOBIAS in the form shown in Figure 3.15. Preprocessing was required to extract one of the variables (*treg_score*, *tconv_score* or *diffbinding*) for all TFBS and gene combinations. As in some situations more than one value can be found for the same combination, the mean was taken for those cases.

The final data form before clustering becomes a matrix where genes are rows and TFBS are columns such as the example in Figure 3.16.

From here *ComplexHeatmap*, a package in R Gu et al. (2016) Figure 3.17 becomes the tool of choice as it allows both the execution of a simple kmeans clustering and the

peak_start	peak_end	gene_name	peak_score	peak_strand	gene_id	peak_id	treg_score	tconv_score	treg_bound	tconv_bound	treg_tconv_log2fc
14968908	14969451	KAZN	-	-	ENSG00000189337	Peak_904	0.16125	0.17320	0	0	-0.06266
15792139	15792532	RPL12P14	-	-	ENSG00000224321	Peak_953	0.28356	0.27981	0	0	0.01389
28189123	28189900	PTAFR	-	-	ENSG00000169403	Peak_1609	0.95339	0.92284	1	1	0.04214
31770821	31771370	ADGRB2	-	-	ENSG00000121753	Peak_1979	0.30752	0.28896	0	0	0.06593
31771461	31771584	ADGRB2	-	-	ENSG00000121753	Peak_1980	0.29132	0.25550	0	0	0.13561
40665282	40666029	RIMS3	-	-	ENSG00000117016	Peak_2434	0.58255	0.28689	1	0	0.80639
77888483	77889269	NEXN	-	-	ENSG00000162614	Peak_4065	0.11757	0.15940	0	0	-0.24549
84720585	84720704	SSX2IP	-	-	ENSG00000117155	Peak_4445	0.20779	0.17362	0	0	0.16525
87292901	87293149	LMO4	-	-	ENSG00000143013	Peak_4597	0.20476	0.11178	0	0	0.50912
89064781	89065609	GBP1	-	-	ENSG00000117228	Peak_4687	0.18834	0.21501	0	0	-0.12435
100413110	100413662	CDC14A	-	-	ENSG00000079335	Peak_5366	0.08435	0.24286	0	0	-0.86730
100445476	100446450	CDC14A	-	-	ENSG00000079335	Peak_5371	0.39669	0.22593	0	0	0.59594
101235344	101237156	-	-	-	ENSG00000225938	Peak_5413	0.36707	0.24915	0	0	0.41167
111215820	111216551	CH3L2	-	-	ENSG00000064886	Peak_6034	1.28283	0.98765	1	1	0.34418
111225708	111226468	CH3L2	-	-	ENSG00000064886	Peak_6039	0.69289	1.05312	1	1	-0.53588
207077716	207078210	PKFB2	-	-	ENSG00000123836	Peak_12435	0.44442	0.40656	0	0	0.10244

Figure 3.15: Raw Data extracted from analysis from the TOBIAS framework described in subsection 3.1.4

	AR_MA0007.2	ATF2_MA1632.1	ATF4_MA0833.2	ATF6_MA1466.1	ATF7_MA0834.1	BACH1_MA1633.1	BACH2_MA1101.2
ENSG00000001036	0.00000000	-0.50639000	0.00000000	0.00000000	0.00000000	0.07874500	0.00000000
ENSG00000001084	0.04947000	-0.06489000	-0.10846000	0.26336000	0.00000000	0.34450000	0.34450000
ENSG00000001561	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
ENSG00000002587	0.00000000	0.00000000	-0.09697000	-0.26163000	0.00000000	0.00363000	0.00000000
ENSG00000003056	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
ENSG00000003147	-0.24920000	-0.14673000	0.00000000	0.06087000	0.00000000	-0.12838000	0.00000000
ENSG00000003400	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	-0.15763000
ENSG00000003402	0.51128000	0.00000000	0.00000000	-0.25246000	0.00000000	-0.34245500	0.00000000
ENSG00000004139	0.00000000	0.00000000	0.00000000	-0.19667000	0.00000000	0.00000000	0.00000000
ENSG00000004660	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.06258000	0.00000000
ENSG00000004866	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.43714000

Figure 3.16: Example of a small portion of a Matrix ready to be used as an input for clustering analysis. Genes are in rows, TFBS are in columns

elaboration of a heatmap for observation of said clustering.

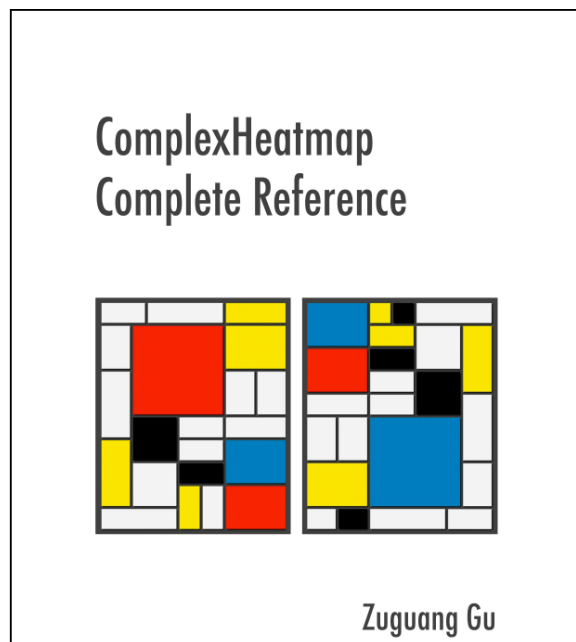


Figure 3.17: *ComplexHeatmap* package logo Gu et al. (2016)

To assure reproducibility the protocol was set as follows:

1. Extract relevant data from the raw data (pairs of TFBS/gene and their respective treg_score, tconv_score or diffbinding);
2. Extract expression data for the genes in rows
3. Convert the data into matrix form and calculate averages when pairs of gene/TFBS have more than 1 value;
4. Estimate the ideal number of clusters by calculating it through the silhouette and elbow methods (and estimating the best between both);
5. Scale the matrix by rows, by column and keep a matrix with no scaling for reference;
6. Calculate the colour scale for the heatmap according to the values of the matrix after scaling;
7. Create the heatmap with 2 k-means (one of columns, one for rows) with the k calculated in 4 and add a bar plot for columns and rows with the expression of each gene;
8. Extract cluster information for each pair of Gene/TFBS in each variation (column scaling, row scaling and no scaling).
9. Analyse results.

As the protocol is quite extensive and reproducibility is paramount, functions were created to automatize and standardize steps.

A first function was created to automatize extraction of the final matrix from the raw data and the row and column data for the bar plots Figure 3.18. 3 variants of the function were created to extract either treg_score, tconv_score or diffbinding from the data.

```
# ---- Function to Extract Heatmap Data - TregScore
getMatrixData_TregScore<-function(data){
  #necessary packages
  library (plyr)
  library(readr)
  library(tidyr)
  library(tibble)
  library(tidyverse)
  library(magrittr)

  #All of the Genes
  genesTreg <- read.table("Data/ExpressionData/TregvsTconv_Thy_DEGnoco.txt", sep = "\t", header = TRUE )

  #____pre-processing

  #correct ensemble
  data$ensembl_gene_id<-data$gene_id

  #remove rows where score is 0
  data=subset(data,data$treg_score!=0)

  #normalize names
  if(!"TFBS_motif" %in% colnames(data))
  {
```

Figure 3.18: Function to extract relevant data to execute the heatmap. This example is for Treg_score. The full code for this task can be seen in https://github.com/theinsilicobiology/Kmeans_TOBIAS_CD4Thymus_paper/blob/main/Functions/ExtractInfoFromDataset.R.

A second function was created to automatize the elbow and silhouette methods. It was set for both rows and columns (by transposing the matrix) allowing for us to retrieve

the ideal k for both k-means in one function. The outputs are the respective graphs for both methods Figure 3.19.

```

AssessNumClusters<-function(datamatrix){
  library(factoextra)

  # ----- Cluster by Rows
  data_widematrixRows<-scale(data_wide, center = TRUE)
  data_widematrixRows<-as.data.frame(data_widematrixRows)

  #max number clusters
  kmax<-nrow(data_widematrixRows)-1

  if(kmax>10){kmax=10}

  # Elbow method
  ElbowGenes<-fviz_nbclust(data_widematrixRows, kmeans, iter.max=1000, method = "wss", k.max = kmax) +
  labs(subtitle = "Elbow method - Genes")
  ElbowGenes

  # Silhouette method
  SilhouetteGenes<-fviz_nbclust(data_widematrixRows, kmeans, iter.max=1000, method = "silhouette", k.max = kmax)+
  labs(subtitle = "silhouette method - Genes")
  SilhouetteGenes
}

```

Figure 3.19: Function that executes the elbow and silhouette methods for the extracted data to determine the ideal number of clusters - k. It outputs the graphs for both rows and columns. The full function can be found in https://github.com/theinsilicobiology/Kmeans_TOBIAS_CD4Thymus_paper/blob/main/Functions/AssessNumClusters.R

After executing scaling according to rows or columns, a function was created to colour the heatmap and recentre it on 0 as seen in Figure 3.20. This function was adapted to diffbinding, `treg_score` and `tconv_score`.

```

#-----column scaling
getcolours_diffBindingHeatmap_colScaling<-function(datamatrix){
  Modes<-vanillaICE::colModes(datamatrix)
  Minimum<-min(Modes)
  Maximum<-max(Modes)

  library(circlize)
  col_fun = colorRamp2(c(min(datamatrix)-.001, Minimum-0.1, Minimum-0.01, Minimum, Maximum, Maximum+0.01, Maximum+0.1, max(datamatrix)+.001), c("#4fff2e", "#189b00", "#106a00", "black", "black", "#b36200", "#e67e00", "#ffa333"), space = "xyz") #
  return(col_fun)
}

```

Figure 3.20: Function that calculates the mode per column or per row (according to scaling) and sets the colours of the heatmap according to it. Scale from blue to green in `treg` and `tconv` score and green/black/orange for diffbinding. The full functions can be found in https://github.com/theinsilicobiology/Kmeans_TOBIAS_CD4Thymus_paper/blob/main/Functions/ColoursHeatmap.R

Finally, after executing the heatmap, two more functions are run. The first Figure 3.21 extracts the info about the created clusters and creates csvs ready to analyse. The second compares the gene expression dataset from the same subset with the TOBIAS data and checks which genes from the gene expression dataset do not exist in the TOBIAS data of the same subset Figure 3.22.

```

TregScore
extractClusterInformation_TregScore<-function(HM){
  #_____ cluster Genes

  r.dend <- row_dend(HM) #If needed, extract row dendrogram
  rcl.list <- row_order(HM) #Extract clusters (output is a list)

  lapply(rcl.list, function(x) length(x)) #check/confirm size gene clusters

  library(magrittr) # needed to load the pipe function '%>%'

  clu_df <- lapply(names(rcl.list), function(i){
    r=rownames(data_wideHeatmap)
    out <- data.frame(ensembl_gene_id = r[rcl.list[[i]]],
                     clusterGene = paste0(i),
                     stringsAsFactors = FALSE)

    return(out)
  }) %>% #pipe (forward) the output 'out' to the function rbind to create 'clu_df'
  do.call(rbind, .)

  clustersGenes<-unique(data.frame(clu_df))

  #_____

  #_____ cluster TFBS

```

Figure 3.21: Function to extract cluster information after a heatmap is created. It generates 3 CSVs, one for gene clusters, one for TFBS clusters and one with the information combined. The functions can be found in https://github.com/theinsilicobiology/Kmeans_TOBIAS_CD4Thymus_paper/blob/main/Functions/extractClusterInformation.R.

```

notintable<-function(AllClusters){
  #deg_datasets
  genesTreg <- read.table("Data/ExpressionData/TregvsTconv_Thy_DEGnoco.txt", sep = "\t", header = TRUE )

  genes<- genesTreg$hgnc_symbol
  genes<-toupper(genes)
  genes<-unique(genes)

  genesSet<-AllClusters$hgnc_symbol
  genesSet<-toupper(genesSet)
  genesSet<-unique(genesSet)

  genes0s<-setdiff(genes, genesSet)
  return(genes0s)
}

```

Figure 3.22: Function to assess which genes exist in a specific subset of gene expression but in the correspondent TOBIAS output data. The full function can be found in https://github.com/theinsilicobiology/Kmeans_TOBIAS_CD4Thymus_paper/blob/main/Functions/notintable.R.

Chapter 4

Results

In this chapter the results obtained during this project will be presented.

In section section 4.1 we'll discuss the final visualizations obtained regarding the existence of a linear correlation between DCA and gene expression that differentiates tTRegs from tTConvs. In section section 4.2 the final results from the standardization of the Gene Set Enrichment Analysis protocol to run with the full mSigDB in our data will be discussed. Finally in the section section 4.3 we'll discuss the final results and discoveries from the clustering procedure applied in the digital footprinting analysis data obtained from the TOBIAS framework.

4.1 Analysing Gene Expression vs Differential Chromatin Accessibility

To check how we've arrived at this results please consult section section 3.2 where you can see the full protocol followed for the following results.

From the analysis of the dataframe we can already conclude a few things, namely regarding ROCs Table 4.1 and genes Table 4.2. We can also conclude that biggest majority of our genes in the dataset are protein coding Table 4.3 vouching for the importance of this data to study tTreg regulation.

Question	Answer
How many ROCs do we have in total?	7520
How many ROCs have a positive DCA?	3593
How many ROCs have a negative DCA?	3927
Which Gene has more ROCs associated? How many?	CSMD1, has 99 ROCs

Table 4.1: ROCs main characteristics in the data

Question	Answer
How many genes do we have in total?	1058
How many genes have a positive logFC?	590
How many genes have a negative logFC?	378

Table 4.2: Genes main characteristics in the data

Gene Biotype	N° Genes
lncRNA	52
processed_pseudogene	17
protein_coding	979
TR_V_gene	1
transcribed_unitary_pseudogene	1
transcribed_unprocessed_pseudogene	6
unprocessed_pseudogene	2

Table 4.3: Distribution of the Genes per Gene Biotype

The bubble plot with Gene Expression in x and Differential Chromatin Accessibility in y Figure 4.1 reveals the possibility of existent linear correlation between both variables. At a first glance, that linear relationship seems to exist.

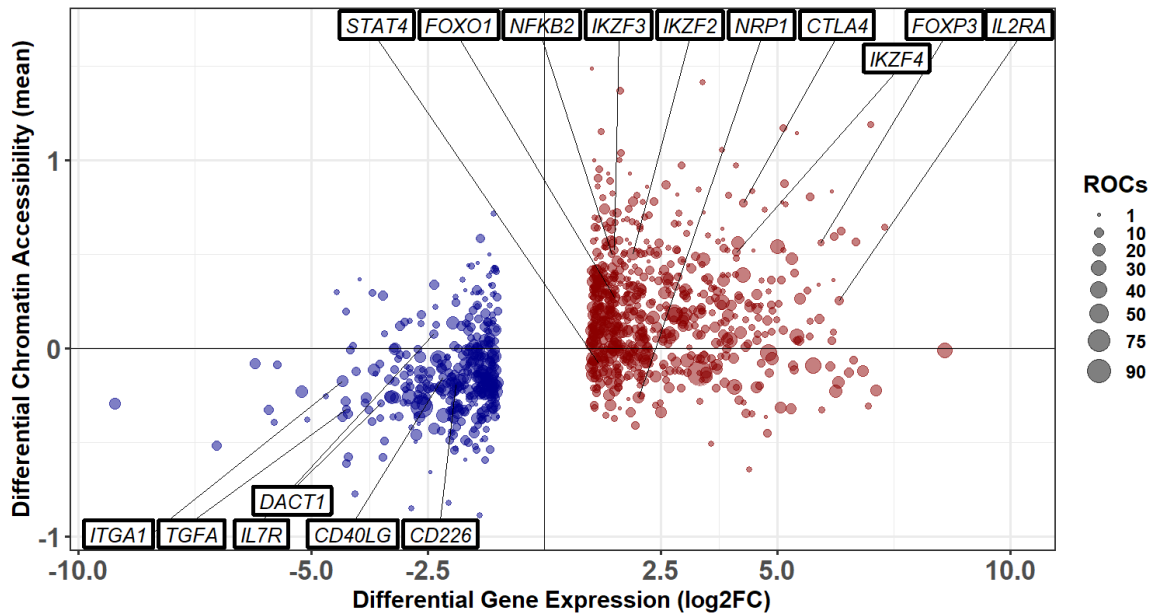


Figure 4.1: Bubble Plot of Differential Gene Expression in x and Differential Chromatin Accessibility in y . To assure one point per gene, the DCA was assumed to be the mean of all ROCs for each gene. The number of ROCs is stored as the size of each bubble giving us an idea on how many each gene possesses.

To assure that the calculation of the Linear Regression of this data is not affected by the restraints of the ggplot2 protocol and to allow full control of the regression, we've opted to calculate separately from the graph and then gather the two in one visualization. At first glance the correlation seems positive and consistent as seen in Figure 4.2.

However if we look at the report from the $lm()$ function from R the values observed are not the best as you can see in Figure 4.3. We can observe a significant p value but the R-Squared doesn't seem to give evidence of a strong correlation. Yet this is not the final form of this linear regression.

By observing the image we can faintly observe the visual aspect of heteroscedasticity as the points in the graph seem to progressively decrease in variance the further they

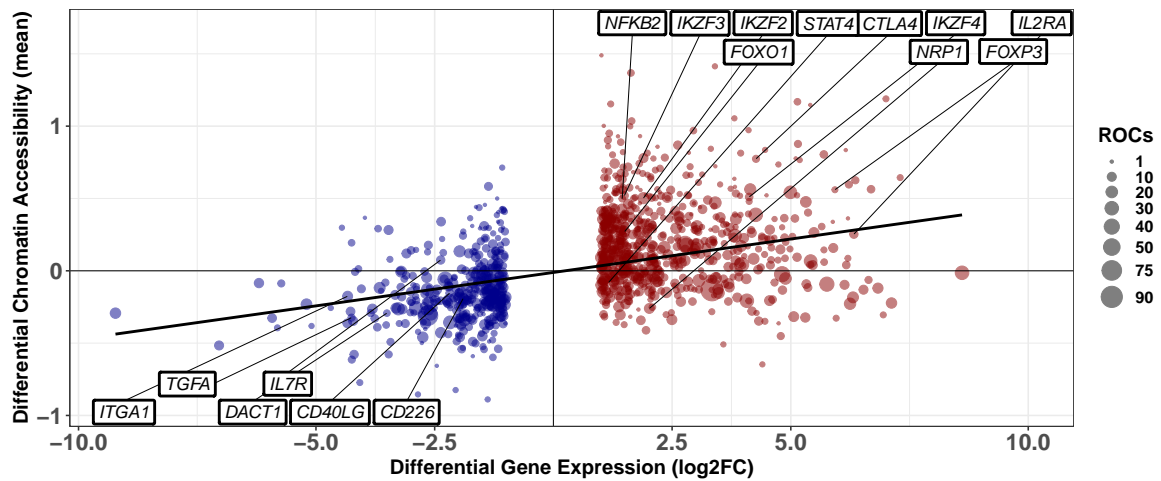


Figure 4.2: The same graph as Figure 4.1 but with an Added Linear Regression line between x and y

```

Call:
lm(formula = DCA ~ logFCA, data = temp)

Residuals:
    Min       1Q   Median       3Q      Max
-1.53928 -0.31710 -0.01637  0.27054  2.53884

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.011417   0.005863  -1.947  0.0516 .
logFCA       0.046222   0.002133  21.673 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4827 on 7518 degrees of freedom
Multiple R-squared:  0.05881, Adjusted R-squared:  0.05868
F-statistic: 469.7 on 1 and 7518 DF, p-value: < 2.2e-16

```

Figure 4.3: Initial Linear Regression results for the Gene Expression vs DCA combo.

are from the origin. So a validation of the existence of heteroscedasticity is in place. The Breusch-Pagan test was chosen and was executed with the aid of the function `bptest()` from the package `lmtest` Achim and Torsten (2002).

The Breusch-Pagan tests for the existence of heteroscedasticity in a linear regression by using the following null and alternative hypotheses:

- **Null Hypothesis (H0):** Homoscedasticity is present (the residuals are distributed with equal variance);
- **Alternative Hypothesis (HA):** Heteroscedasticity is present (the residuals are not distributed with equal variance)

If the p -value of the test is less than some significance level (we assume 0.05) then

we reject the null hypothesis and conclude that heteroscedasticity is present in the regression model.

The protocol for the Breusch-Pagan test goes as follows:

1. Fit the regression model (the model uses the function $lm()$);
2. Calculate the squared residuals of the model;
3. Fit a new regression model, using the squared residuals as the response values;
4. Calculate the Chi-Square test statistic X^2 as $n \cdot R^2_{\text{new}}$ where: n is the total number of observations and R^2_{new} : The R-squared of the new regression model that used the squared residuals as the response values

If the p -value that corresponds to this Chi-Square test statistic with p (the number of predictors) degrees of freedom is less than some significance level (we use 0.05) then reject the null hypothesis and conclude that heteroscedasticity is present.

The Breusch-Pagan Test can be executed simply in R with the `bptest()` function over the original linear regression and, if the p value is <0.05 then heteroscedasticity is indeed, present. By observing Figure 4.4 we can thus validate the existence of heteroscedasticity in this visualization

```
studentized Breusch-Pagan test
data: temp.lm
BP = 85.226, df = 1, p-value < 2.2e-16
```

Figure 4.4: Breusch-Pagan test results of the linear regression executed in Figure 4.3 using the function `bptest()` from the `lmtest` package. As the p -value is <0.05 , heteroscedasticity is in fact, present.

With heteroscedasticity validated we can then move to assess it and modify the original linear regression to deal with the difference in variance in this data. In R we can solve this simply by using the function `lmrob()` from the package `robustbase` Todorov and Filzmoser (2010) which computes a robust regression version of the original linear regression. As we can see in Figure 4.5, the R-Squared evolved towards a relevant 0.2308.

With this we end up with evidence of a significant positive correlation between Differential Chromatin Accessibility with the existence of heteroscedasticity and an R-squared of 0.2308.

The final equation is thus approximately,

$$DCA = \log FC * 0.061922 + 0.019303$$

Which lead us to conclude that Up regulated DEGs are more frequently associated to regions of chromatin where mean accessibility is increased in tTregs (3,593 “open” ROCs), when compared to tTconv. On the other hand, Down regulated DEGs show a stronger association with regulatory regions with an associated decreased accessibility in tTregs (3,927 “closed” ROCs),

```

Call:
lmrob(formula = mean ~ logFCA, data = tempcoruniqu1)
  --> method = "MM"
Residuals:
    Min       1Q   Median       3Q      Max
-0.93880 -0.17461 -0.01646  0.19381  1.43573

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.019303   0.008707   2.217   0.0268 *
logFCA       0.061922   0.004051  15.284  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.2674
Multiple R-squared:  0.2308,    Adjusted R-squared:  0.2301
Convergence in 12 IRWLS iterations

```

Figure 4.5: Robust regression of the same variables as in Figure 4.3. The R-squared did indeed improve to 0.2308

4.2 Gene Set Enrichment Analysis - Standardizing the Algorithm

From analyzing the results of crossing our data with the Hallmark collection of mSigDB we can already observe some interesting results.

At first we attempted to perform the GSEA on the full gene expression dataset and we find already interesting results in the Hallmark collection test. We can observe in table Table 4.4 that a few interesting pathways are significantly enriched in this data, namely *HALLMARK_IL2_STAT5_SIGNALING*, *HALLMARK_IL6_JAK_STAT3_SIGNALING*, *HALLMARK_INFLAMMATORY_RESPONSE*, *HALLMARK_TNFA_SIGNALING_VIA_NFKB*, *HALLMARK_INTERFERON_GAMMA_RESPONSE*, *HALLMARK_E2F_TARGETS* and *HALLMARK_WNT_BETA_CATENIN_SIGNALING*, all pathways related to thymic t cell metabolism. However, as the input hasn't been restricted for significance, these results might be dubious.

Repeating the test for the subset of our gene expression data restricted for significance is important to validate these results. We can observe in Table 4.5 that the results turn a lot more simplified.

By observing Table 4.5 we can observe a group of enriched pathways namely the first 4 (*HALLMARK_IL2_STAT5_SIGNALING*, *HALLMARK_ESTROGEN_RESPONSE_LATE*, *HALLMARK_TNFA_SIGNALING_VIA_NFKB* and *HALLMARK_INFLAMMATORY_RESPONSE*) and the last one (*HALLMARK_GLYCOLYSIS*) on the table.

From comparing both tables we can see improved enriched pathways in the table with cut-off than the whole table and more potential for explainability of the results. One question then arose, which genes exist in common with our data in these enriched pathways and are there any genes in common between pathways?

We took the results from these 2 tests and constructed the heatmap described in section 3.3 and we can see the results in Figure 4.6 for the dataset without cut-off and Figure 4.7 for the database with cut-off.

By observing at the heatmap that is a result of the GSEA of our full gene expression

Pathway	Padj	NES
HALLMARK_IL2_STAT5_SIGNALING	0.000000037	2.0393
HALLMARK_IL6_JAK_STAT3_SIGNALING	0.000020558	2.0215
HALLMARK_INFLAMMATORY_RESPONSE	0.00000623	1.9688
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.000001987	1.9169
HALLMARK_INTERFERON_GAMMA_RESPONSE	0.000047582	1.7822
HALLMARK_INTERFERON_ALPHA_RESPONSE	0.005222246	1.6692
HALLMARK_XENOBIOTIC_METABOLISM	0.003937029	1.6612
HALLMARK_KRAS_SIGNALING_DN	0.014707853	1.6357
HALLMARK_ALLOGRAFT_REJECTION	0.005222246	1.6115
HALLMARK_ESTROGEN_RESPONSE_LATE	0.014707853	1.5867
HALLMARK_HYPOXIA	0.014707853	1.5383
HALLMARK_KRAS_SIGNALING_UP	0.027959701	1.5131
HALLMARK_ANGIOGENESIS	0.121411483	1.5016
HALLMARK_MYOGENESIS	0.052956034	1.4639
HALLMARK_BILE_ACID_METABOLISM	0.112518519	1.4349
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.1225	1.4218
HALLMARK_APICAL_SURFACE	0.184429477	1.4043
HALLMARK_APOPTOSIS	0.062672849	1.3762
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	0.184429478	1.3192
HALLMARK_P53_PATHWAY	0.282982249	1.2081
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	0.522956623	1.1418
HALLMARK_UV_RESPONSE_UP	0.494093921	1.1184
HALLMARK_COAGULATION	0.557723578	1.101
HALLMARK_APICAL_JUNCTION	0.557723578	1.0763
HALLMARK_MTORC1_SIGNALING	0.557723578	1.0696
HALLMARK_COMPLEMENT	0.569405523	1.0548
HALLMARK_PROTEIN_SECRETION	0.604008992	1.0359
HALLMARK_ESTROGEN_RESPONSE_EARLY	0.608129729	1.0157
HALLMARK_HEDGEHOG_SIGNALING	0.608129729	1.0122
HALLMARK_UV_RESPONSE_DN	0.608129729	1.0097
HALLMARK_ADIPOGENESIS	0.636283797	0.9931
HALLMARK_TGF_BETA_SIGNALING	0.658373171	0.9592
HALLMARK_PEROXISOME	0.762645914	0.8913
HALLMARK_HEME_METABOLISM	0.837714058	0.8729
HALLMARK_FATTY_ACID_METABOLISM	0.91178119	0.8141
HALLMARK_OXIDATIVE_PHOSPHORYLATION	1	0.6123
HALLMARK_PI3K_AKT_MTOR_SIGNALING	1	0.5856
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	1	0.4229
HALLMARK_DNA_REPAIR	1	-0.7182
HALLMARK_MYC_TARGETS_V2	0.9195061728	-0.8117
HALLMARK_MYC_TARGETS_V1	0.9117811905	-0.8998
HALLMARK_MITOTIC_SPINDLE	0.6976821192	-0.9679
HALLMARK_ANDROGEN_RESPONSE	0.5587396849	-1.0375
HALLMARK_SPERMATOGENESIS	0.5662735849	-1.05
HALLMARK_GLYCOLYSIS	0.1686251834	-1.2076
HALLMARK_NOTCH_SIGNALING	0.1819462228	-1.4198
HALLMARK_G2M_CHECKPOINT	0.0052222456	-1.5029
HALLMARK_E2F_TARGETS	0.0052222456	-1.5189
HALLMARK_WNT_BETA_CATENIN_SIGNALING	0.0364379885	-1.6649

Table 4.4: Subset of the results table of the GSEA applied to our gene expression data with the Hallmark mSigDB as a comparison.

Pathway	Padj	NES
HALLMARK_IL2_STAT5_SIGNALING	0.0134	1.9013
HALLMARK_ESTROGEN_RESPONSE_LATE	0.0387	1.7745
HALLMARK_INFLAMMATORY_RESPONSE	0.0387	1.7339
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.2169	1.5168
HALLMARK_KRAS_SIGNALING_DN	0.2836	1.4836
HALLMARK_MYOGENESIS	0.3492	1.3406
HALLMARK_INTERFERON_GAMMA_RESPONSE	0.3492	1.3223
HALLMARK_ALLOGRAFT_REJECTION	0.3492	1.3073
HALLMARK_KRAS_SIGNALING_UP	0.3856	1.2688
HALLMARK_HYPOXIA	0.3912	1.2494
HALLMARK_APOPTOSIS	0.3912	1.2369
HALLMARK_COMPLEMENT	0.3912	1.2285
HALLMARK_XENOBIOTIC_METABOLISM	0.4226	1.1490
HALLMARK_P53_PATHWAY	0.784	0.9163
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	0.784	0.8952
HALLMARK_APICAL_JUNCTION	0.851	0.7975
HALLMARK_ESTROGEN_RESPONSE_EARLY	0.8562	0.7644
HALLMARK_MTORC1_SIGNALING	0.9383	0.6256
HALLMARK_MITOTIC_SPINDLE	1	0.3489
HALLMARK_HEME_METABOLISM	0.792	-0.8591
HALLMARK_UV_RESPONSE_DN	0.4223	-1.1485
HALLMARK_GLYCOLYSIS	0.0387	-1.9263

Table 4.5: Subset of the results table of the GSEA applied to our gene expression data (with a previous cutoff applied to those which p -value for the expression was significant) with the Hallmark mSigDB as a comparison.

dataset we can observe that the most enriched pathways are also those with more genes in common with other pathways. By observing the genes in the leadingEdge, we see some expected genes, namely IL2RA, IL2RB, IL10RA, CTLA4, DUSP4, IKZF4, BATF, IRF8, NFKB1, NFKB2, REL, RELB, NFKBIE, BHLHE40, KLF6, NR4A3, BCL3 and BCL2A1 Hayatsu et al. (2017) that were identified as present in the metabolism of the development of thymic Tregs.

The evidence of this group of genes is existent in Figure 4.6 but their enrichment is more evident in Figure 4.7

As enrichment and explainability have to go hand in hand in computational biology we've decided to move forward with the results from the gene expression data with the cut-off for significance as it gives meaningful, interpretable information, while also assuring that we are within significant statistics.

We progressed to analyze which kind of enrichment we were finding in these gene sets with our data, and to understand if the enrichment was mostly on top of the rank, bottom or a mix, the best approach is an enrichment plot

We thus create an enrichment plot for the pathways with a significant p -value

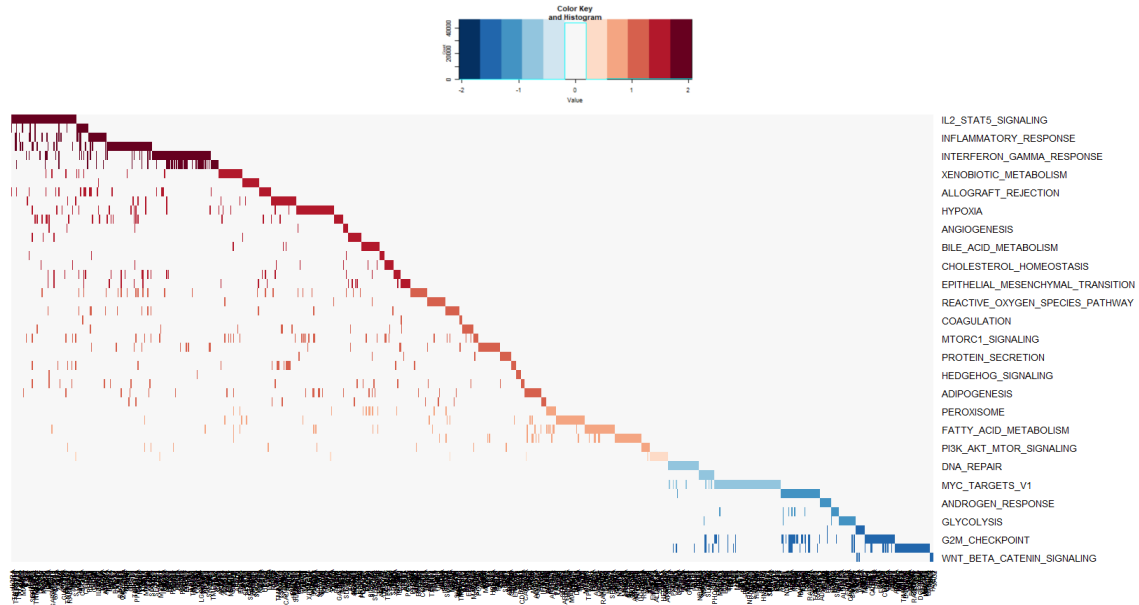


Figure 4.6: GSEA results for the full Gene Expression dataset in comparison with the Hallmark Collection from mSigDB. The Rows are the pathways with Enrichment, in the columns are the genes identified in common between our data and the hallmark collections, the value in the heatmap corresponds to the NES calculated during the FGSEA protocol

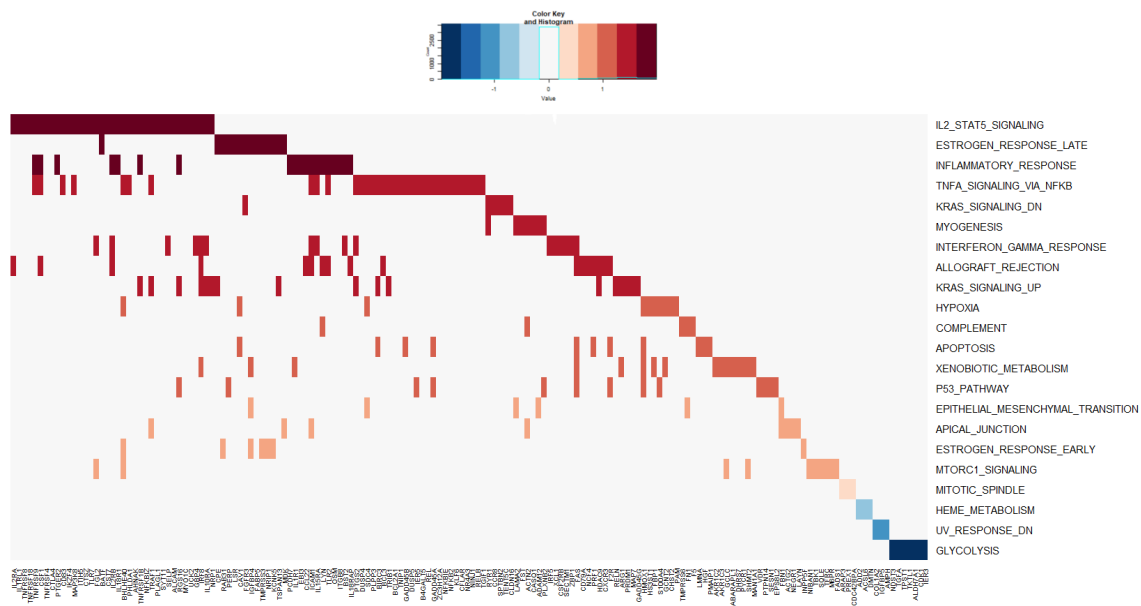


Figure 4.7: GSEA results for the Gene Expression dataset with a cut-off for significance in p -value in comparison with the Hallmark Collection from msigDB. The Rows are the pathways with Enrichment, in the columns are the genes identified in common between our data and the hallmark collections, the value in the heatmap corresponds to the NES calculated during the FGSEA protocol

value we can understand the dimension of this enrichment as in Figure 4.8 where can observe a higher number of genes (sticks in the barcode) in *HALLMARK_IL2_STAT5_SIG-*

NALING. We can also observe that most of the enrichment comes from the top of the gene sets. The least enriched pathway, *HALLMARK_GLYCOLYSIS*, only appearing as significant in the table with the previous cut-off, we find the smallest gene set in common being *NDST3, TGFA, TPST1, TKTL1, ALDH7A1, CDK1* and *IER3*.

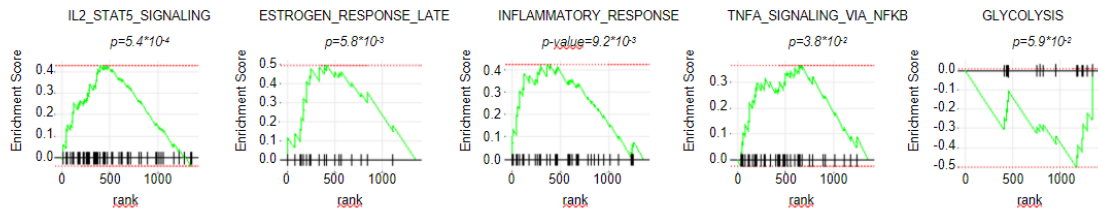


Figure 4.8: Enrichment plots for the pathways in Figure 4.7 where the p -value is significant.

To sum up, analysing the GSEA results for the cross between our data and the Hallmark collection from the mSigDB we can then conclude that:

- We find important signalling molecules such as *IL2RA*, *IL2RB*, *IL10RA*, and *CTLA4* in the case of *HALLMARK_IL2_STAT5_SIGNALING*, or *DUSP4*, and *IL15RA* both in *HALLMARK_TNFA_SIGNALING_VIA_NFKB* and *HALLMARK_INFLAMMATORY_RESPONSE*;
- The overlap of the significantly enriched signatures includes several transcription factors of relevance in the Tcell development context, namely *IKZF4* (*Eos*), *BATF*, *IRF8*, the NFKB2 pathway inhibitor *NFKBIZ* (*HALLMARK_IL2_STAT5_SIGNALING*; *NFKB2*, *REL*, *RELB*, their inhibitor *NFKBIE*, *BHLHE40*, *KLF6*, *NR4A3*, *BCL3* and *BCL2A1* (*HALLMARK_TNFA_SIGNALING_VIA_NFKB*).

4.3 Clustering of Digital Footprinting Analysis Results

With the protocol set as described in section 3.4 to configure and create the heatmaps, it was decided to set up a set of experiments creating subsets of the full dataset as it can be seen on diagram in 4.9.

First a division according to *Tregbound=1* defined as the combination of gene/TFBS that is bound in *tTreg*, and *ALL*, defined as all Data extracted from the TOBIAS analysis.

Within these a group of experiments was set:

- **DEGS**, defined as the list of Differentially Expressed Genes calculated during the Data Preparation stages,
- **UP** as the up regulated DEG genes in *tTregs*,
- **DOWN** as the down regulated DEG genes in *tTregs*,
- **NOCO** as the gene data without any set cutoffs,
- **NOCOUP** as the NOCO genes up regulated in *tTregs* and

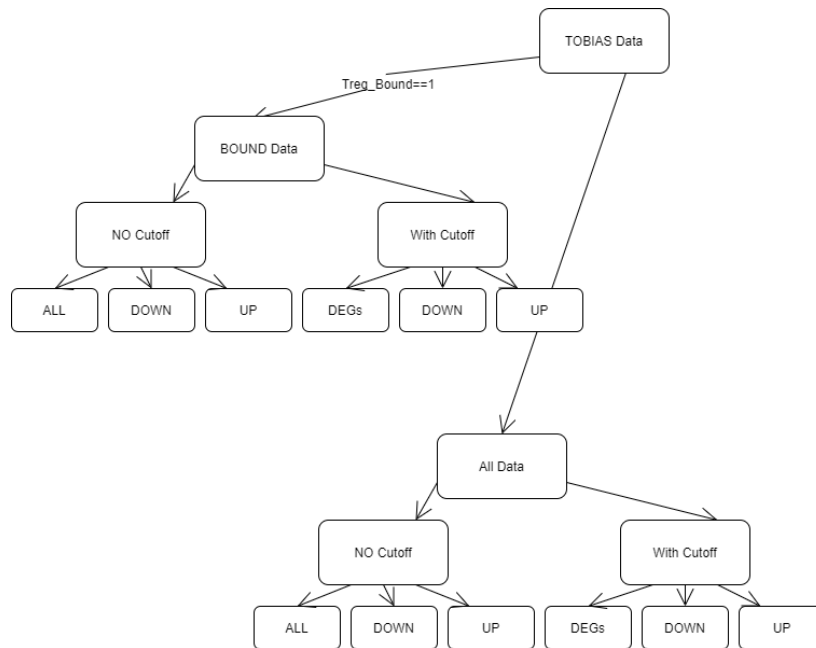


Figure 4.9: Types of Heatmaps Created

- **NOCODOWN** as the NOCO genes down regulated in *tTregs*.

For these experiments 3 sets of heatmaps were constructed, **ColScaling** were the matrix input was scaled by row putting emphasis on TFBS regulation across the genes, **RowScaling** were the matrix input was scaled by row putting emphasis on Gene Expression across the TFBS and **NoScaling** to use as reference.

For further analysis, as the patterns of gene expression crucial to *tTreg* development became the most preponderant question, the choice fell on analysing the results from the row scaling heatmaps and the tests were `treg_bound==1`, so pairs of TFBS/gene that were determined to be bound to thymic *tregs* by TOBIAS.

When analyzing the heatmaps, the intersection of a cluster from the kmeans of the genes and a cluster from the kmeans of the columns that has a distinct colour will be defined as a **Gene Regulatory Module (GRM)**.

We'll dwelve into discovering which transcription factors are associated with the discovered GRM's in Bound UP regulated genes subset, Bound Down Regulated Genes Subset and Bound DEGs genes Subset

4.3.1 Bound Thymic T regs - UP regulated Genes

First we'll analyse the results of for the bound thymic *tregs*, in this case for up regulated genes.

Beginning by analysing the `Treg_score` Figure 4.10 heatmap we can see 3 major GRM's:

- **Row 4, Column 1** - We find in this GRM, transcription factors such as *BACH1*, *BACH2*, *BATF*, *FOS*, *FOSL2*, *JUNB* and *MAFK*, we'll see this cluster often, and realize that all of them belong to the **Activator Protein 1 Family (AP-1)**, a group of transcription factors that regular cellular processes in response to stimuli;

- **Row 1, Column 2** - We find in this GRM, transcription factors such as *ETV5*, *IKZF1*, *ETS1*, *ELK3*, *FLI1*, *ERF*, *ETV6*, *ELK1*, *ETV1*, *ELF4*, *ELF2*, *ETS2*, *ETV3*, *ELF1*, *ELK3* and *ZBTB7A*, this group of **ETS/ETV/ELF TF's** will appear often;
- **Row 5, Column 5** - We find in this GRM, transcription factors such a *SP2*, *KLF9*, *KLF4*, *SP3*, *KLF3*, *KLF10*, *KLF6*, *KLF11*, *KLF16*, *KLF5*, *SP4*, *KLF2* and *SP1*, we'll see this cluster often, and realize that most of them belong to the **KLF/SP family** which are C2H2 zinc-finger containing transcription factors split into two groups based on the structure at the N-terminus, a group of transcription factors that regular cellular processes in response to stimuli.

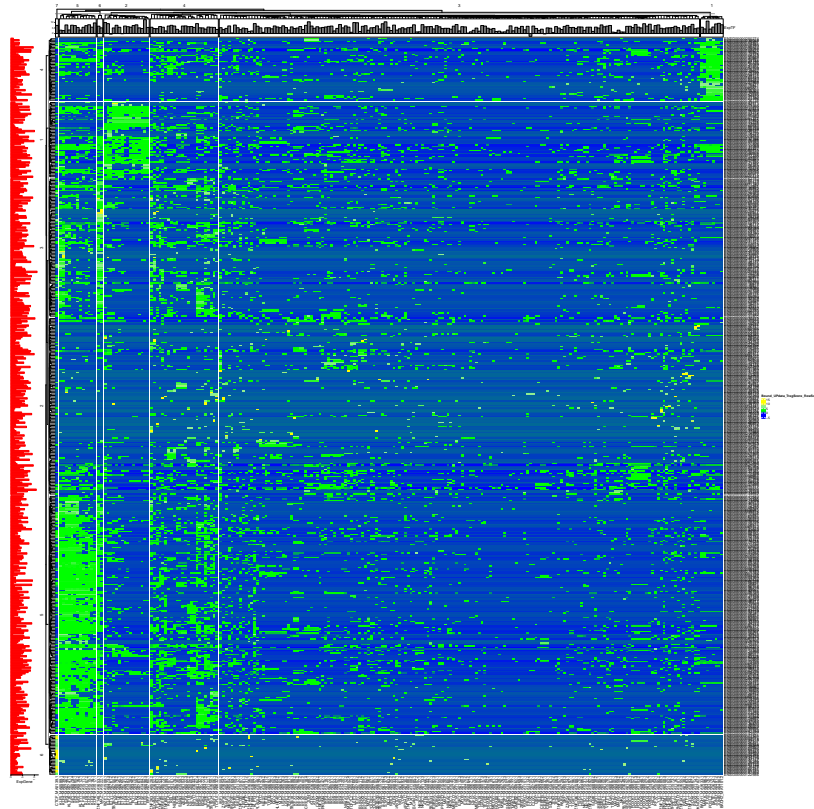


Figure 4.10: Heatmap for Clustering of Treg_score data for the Bound Up regulated subset

Next, analysing the Tconv_score Figure 4.11 heatmap we can see 2 major GRMs:

- **Row 4, Column 2** - It's a big cluster but here we find interesting TF's such as *FOXP3*, *MAF*, *REL*, *RUNX1*, *RUNX2* and *TBX21*;
- **Row 3, Column 1** - we can find transcription factors here such as *BACH1*, *BACH2*, *BATF*, *MAFK*, *FOS*, *FOSL2*, *JUND* and *JUNB*, the AP1 group.

Finally , analyzing the DiffBinding Figure 4.12 heatmap we can see 3 major GRM's:

- **Row 5, Column 7** . here we find above all TF's from the KLF family acting as repressors such as *KLF10*, *KLF11*, *KLF16*, *KLF2*, *KLF3*, *KLF4*, *KLF5*, *KLF6*, *KLF9* and TF's from the SP family such as *SP1*, *SP2*, *SP3* and *SP4*, the KLF/SP family of genes;

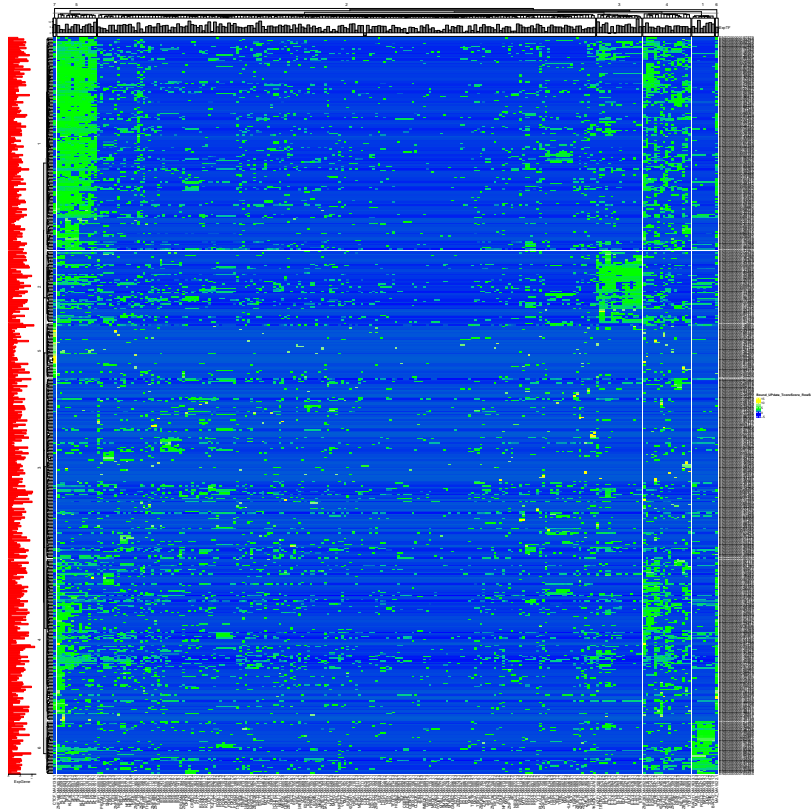


Figure 4.11: Heatmap for Clustering of Tconv_score data for the Bound Up regulated subset

- **Row 6, Column 5** . here we find the elements of the AP1 group such as *BACH1*, *BACH2*, *BATF*, *MAFK*, *FOS*, *FOSL2*, *JUND* and *JUNB*;
- **Row 4, Column 7** - here we find above the same KLFSP family protein as in the GRM of Row 5 Column 7, yet here they work mostly as activators.

4.3.2 Bound Thymic T regs - DOWN regulated Genes

In this section we'll analyse the results for the subset for Bound Down Regulated Genes.

Beginning by analyzing the Treg_score Figure 4.13 heatmap we can one major GRMs:

- **Row 1, Column 4** - here we find above all TF's from the KLF/SP family, *ZNF148* and *EGR1*

Next, analyzing the Tconv_score Figure 4.14 heatmap we can see one GRM:

- **Row 6, Column 4** - we find quite a bit of proteins from the KLF family (*KLF2*, *KLF3*, *KLF4*, *KLF6* and *KLF9*), some from the SP family (*SP1*, *SP2* and *SP4*), *EGR1*, *MAZ* and *ZNF148*, thus the KL/SP family.

Finally , analyzing the DiffBinding Figure 4.15 heatmap we can see 3 major GRMs:

- **Row 6, Column 5** . we find in this GRM TF's acting as activators from the KLF/SP family such as *KLF5* and *SP1*. We also find *EGR1*.

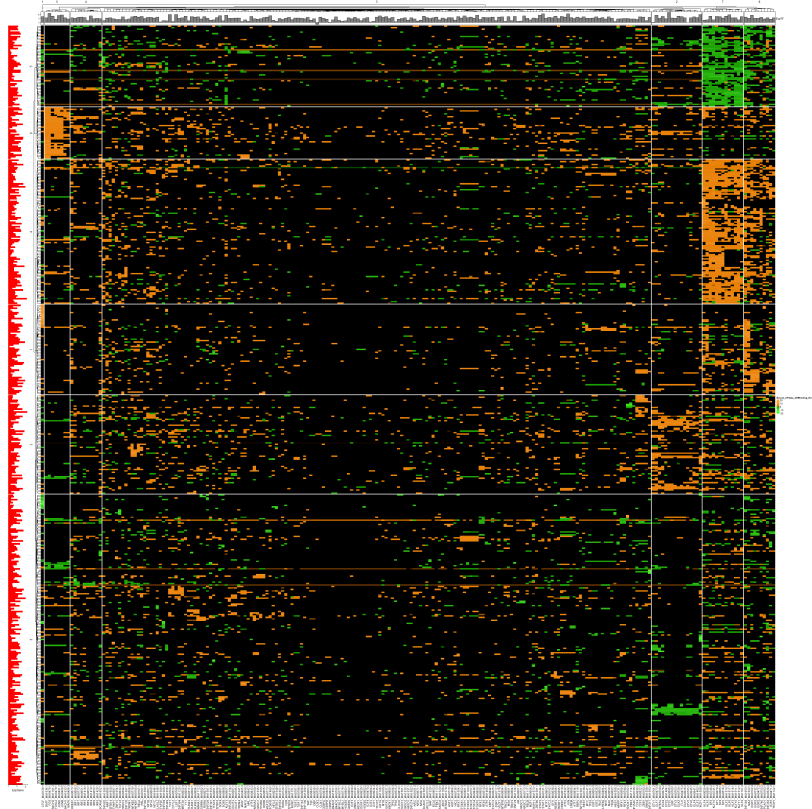


Figure 4.12: Heatmap for Clustering of DiffBinding data for the Bound Up regulated subset

- **Row 5, Column 3** - we find in this GRM, TF's such as *FLI1*, *ELF4*, *ETS2*, *ELK4*, *ETV5*, *ERF*, *ZBTB7A*, *ELK1*, *ELK3*, *ETV3*, *ETS1*, *ZKSCAN5*, *ELF2* and *ETV6*, being here the ETS/ETV/ELF TF's;
- **Row 1, Column 5** - we find in this GRM TF's acting as repressors from the KLF/SP family such as *KLF5* and *SP1*. We also find *EGR1*.

4.3.3 Bound Thymic T regs - DEGs regulated Genes

In this section, we'll analyse the results from the subset of Bound DEGs genes. This subset is a sum of the Bound Up and Bound Down genes.

Beginning by analyzing the Treg_score Figure 4.16 heatmap we can see 3 major GRMs:

- **Row 3, Column 3** - Here we find again the AP1 family, finding *FOS*, *FOSL2*, *BACH1*, *BACH2*, *BATF*, *MAFK*, *JUNB*, *JUND* and *CTCF*, thus the AP-1 family GRM;
- **Row 2, Column 2** - Here we find the TF's *ELF4*, *ETV5*, *FRF*, *FLI1*, *ETS2*, *ELF1*, *ELF1*, *ETS1*, *ETV3*, *ETV6*, *ELK1*, *ETV6*, *ELK1*, *IKZF1*, *ELF2*, *ELK4*, *ETV1*, *ELK3* and *ZBTB7A*, being this group part of the ETS/ETV/ELF TF's GRM;
- **Row 1, Column 5** - in this GRM we find once more the KLFSP family with *SP1*, *SP2*, *SP3*, *SP4*, *KLF3*, *KLF4*, *KLF5*, *KLF6*, *KLF9*, *KLF10*, *KLF11*, *KLF16* and *ZNF148*, being this group part of the KLF/SP family GRM.

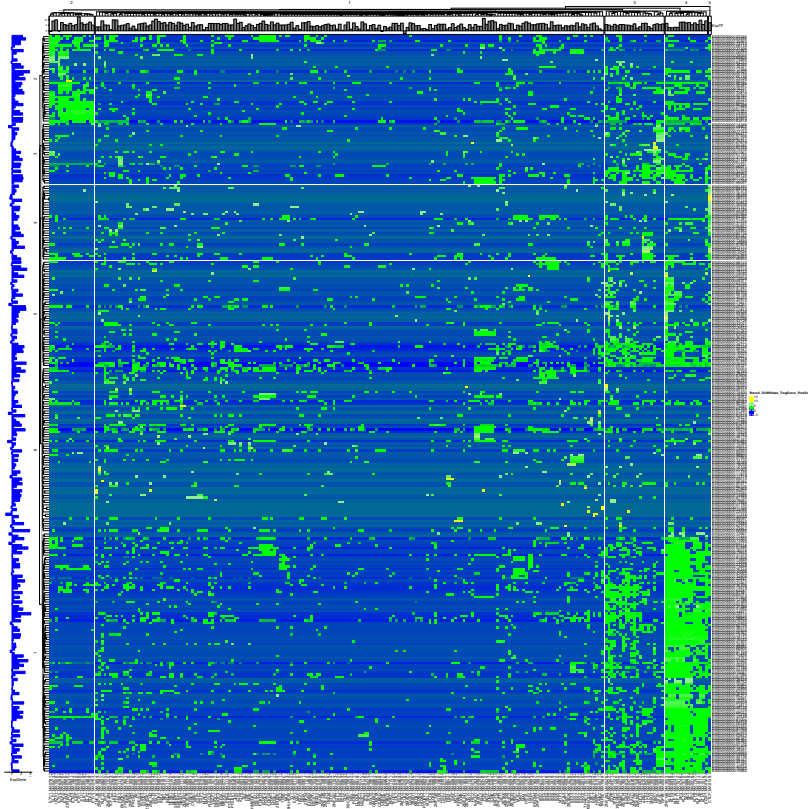


Figure 4.13: Heatmap for Clustering of Treg_score data for the Bound Down regulated subset

Next, analyzing the Tconv_score Figure 4.17 heatmap we can see 3 major GRMs:

- **Row 2, Column 3** - in this GRM we find *ERF, ELF1, ELF2, ELF4, ETS1, ETS2, ETV1, ETV3, ETV5, ELK1* and *FLI1*;
- **Row 1, Column 5** - in this GRM we find once more the KLFSP family with *SP1, SP2, SP3, SP4, KLF3, KLF4, KLF5, KLF6, KLF9, KLF10, KLF11, KLF16* and *ZNF148*
- **Row 6, Column 1** - we find the AP-1 family in this GRM namely *BATF, FOSL2, FOS, BACH1, MAFK, JUND, BACH2* and *JUNB*.

Finally, analyzing the DiffBinding Figure 4.15 heatmap we can see 3 major GRMs:

- **Row 6, Column 2** - we find *ELK1, ETV3, ETV5, ZBTB7A, IKZF1, ETS2, FLI1, ETS1, ELK4, ELK3, ERF, ELF4, ETV1, ELF1, ZKSCAN5, ELF2* and *ETV6*, being this group part of the ETS/ETV/ELF TF's family GRM;
- **Row 7, Column 5** - we see once again the KLFSP family activating genes in tTregs with *SP2, KLF16, KLF10, SP4, KLF9, KLF6, KLF5, SP1, KLF3, KLF4, SP3, KLF11* and *KLF2* being this group part of the KLF/SP family GRM;
- **Row 1, Column 5** - the same as before of the KLF/SP family, this time repressing genes in tTreg.

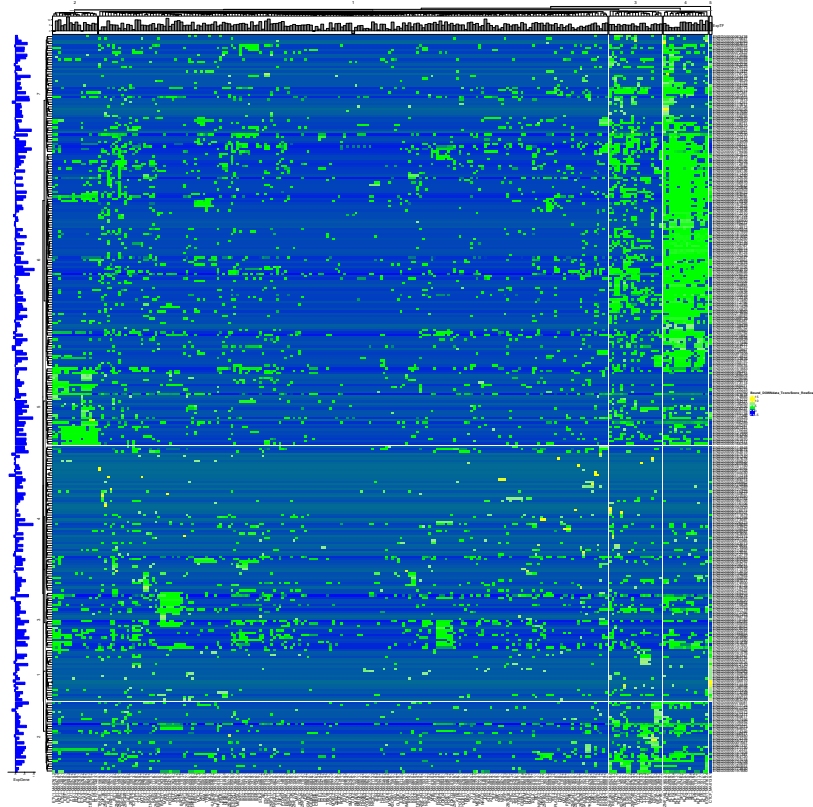


Figure 4.14: Heatmap for Clustering of Tconv_score data for the Bound Down regulated subset

4.3.4 Conclusions

From the sections subsection 4.3.1, subsection 4.3.2 and subsection 4.3.3 we can see that some GRM's are common between them.

As our main interest during the project is to understand tTreg development in contrast with tTconv development , we can reduce our discoveries mostly to the diffbinding heatmaps for Bound Up-Regulated and Bound Down-Regulated subsets.

We find 3 major groups of genes in notorious Gene Regulatory Modules all over the tests: the **AP1 family**, the **KLF/SP family** and the **ETS/ETV/ELF family**.

The **AP1 family** GRM is constituted by proteins that form heterodimers or homodimers and bind to the DNA Katagiri et al. (2021). It comprises 4 sub-families Jun (c-Jun, JunB, JunD), c-Fos (c-Fos, FosB, Fra1, Fra2), musculoaponeurotic fibrosarcoma (Maf; c-Maf, MafB, and MafA. Mafg/f/k, Nrl), and activating transcription f actor (ATF; ATF2, LRF1/ATF3, B ATF, JDP1, JDP2). AP-1 has pleiotropic effects and plays a central role in various aspects of the immune system, such as T cell activation, Th cell differentiation, T cell anergy, and fatigue . We find this group in **Row 6, Column 5** of Figure 4.19.

The **KLF/SP family** GRM of transcription factors are C2H2 zinc-finger containing transcription factors split into two groups based on the structure at the N-terminus Hart et al. (2012). We can find them with repressor effect in tTregs in **Row 5 Column 7** of Figure 4.19 and activator effect in **Row 4 Column 7**. On Figure 4.20 we find the same group acting as activators of transcription in **Row 6 Column 5** and repressors in **Row 1 Column 5**.

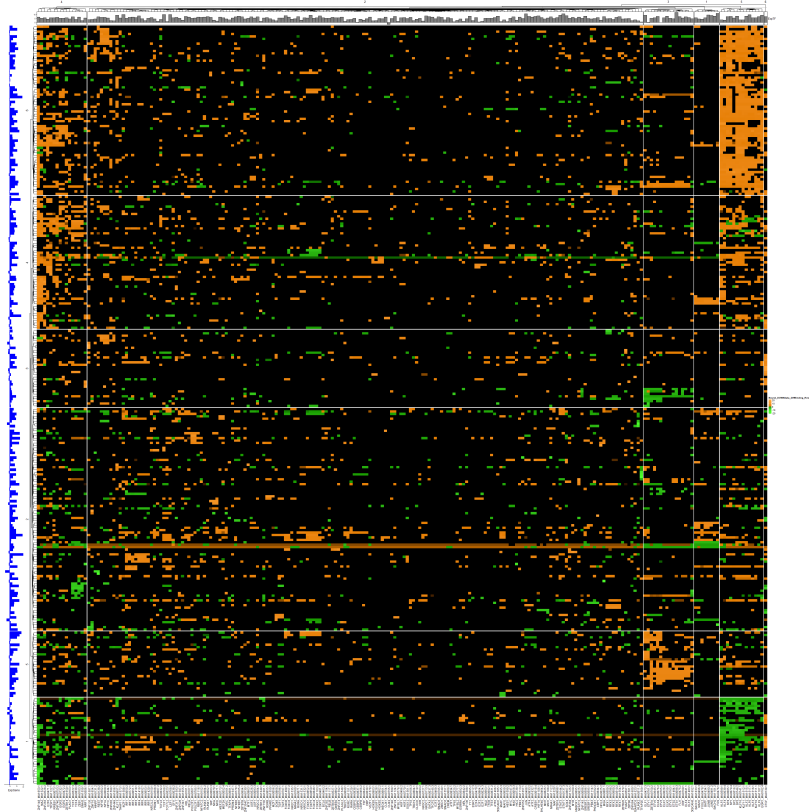


Figure 4.15: Heatmap for Clustering of DiffBinding data for the Bound Down regulated subset

Finally the **ETS/ETV/ELF family** GRM, which possesses Ets domain, which is shared by all ETS proteins, specifically recognizes DNA sequences that contain a GGAA/T core element. ETS group proteins are involved in multiple biological processes such as hematopoiesis, angiogenesis, or tumor progression. They are also associated to B and T cell development Mouly et al. (2010). We find the ETS/ETV/ELF family in Figure 4.20 in **Row 5 Column 3**.

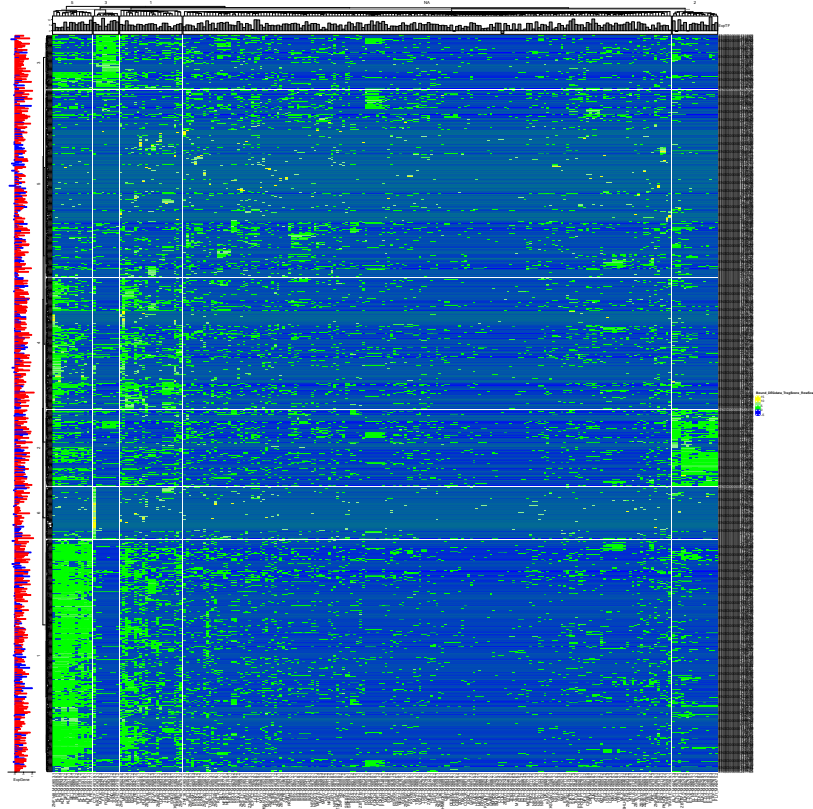


Figure 4.16: Heatmap for Clustering of Treg_score data for the Bound DEGs subset

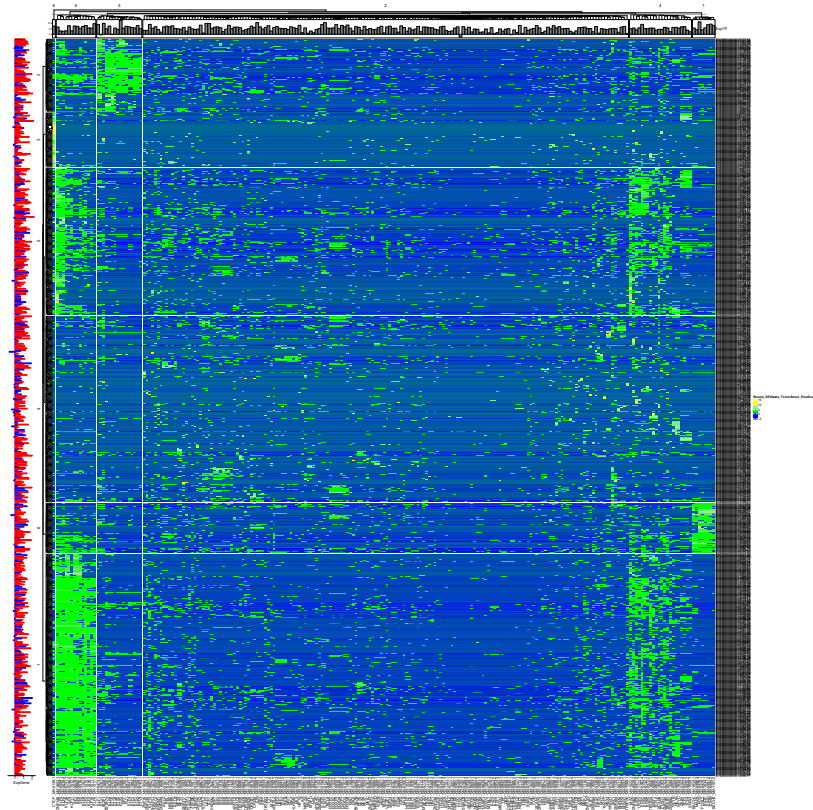


Figure 4.17: Heatmap for Clustering of Tconv_score data for the Bound DEGs subset

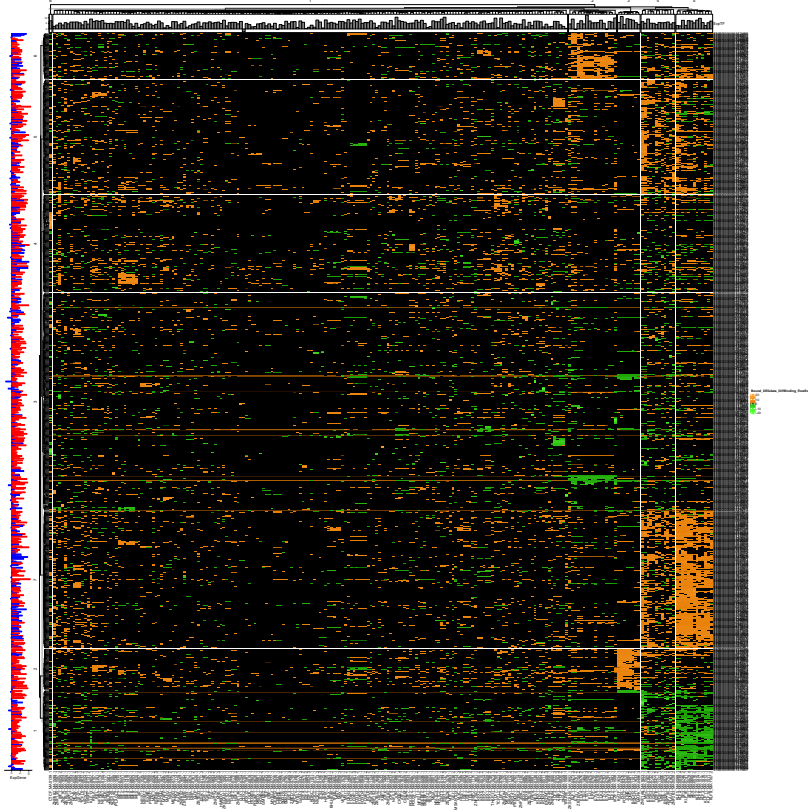


Figure 4.18: Heatmap for Clustering of DiffBinding data for the Bound DEGs subset

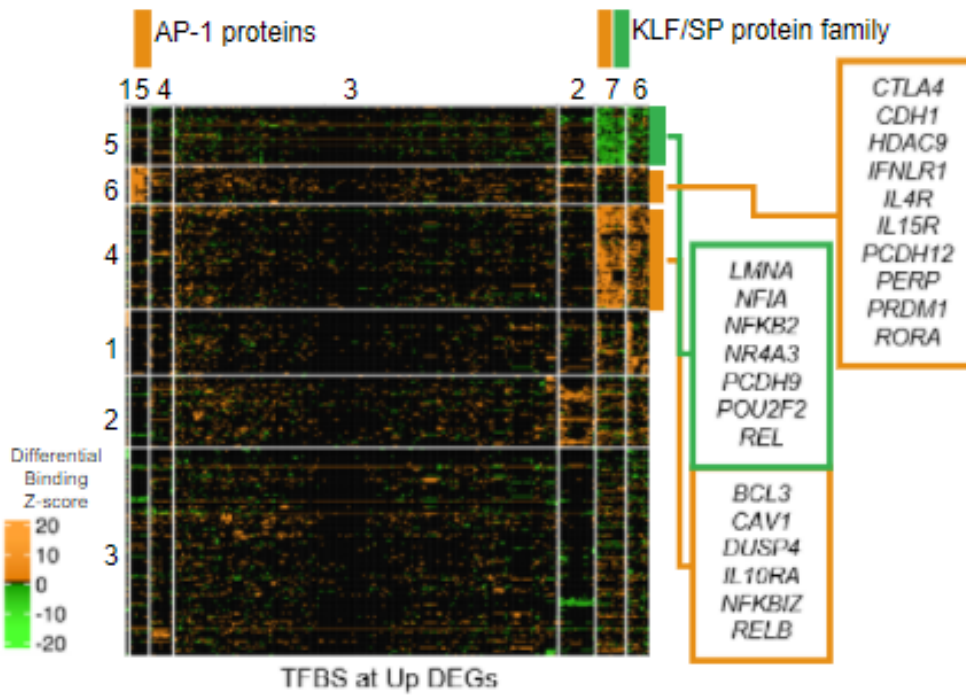


Figure 4.19: Heatmap for Clustering of DiffBinding data for the Bound UP regulated subset - Annotated for to show the GRMs

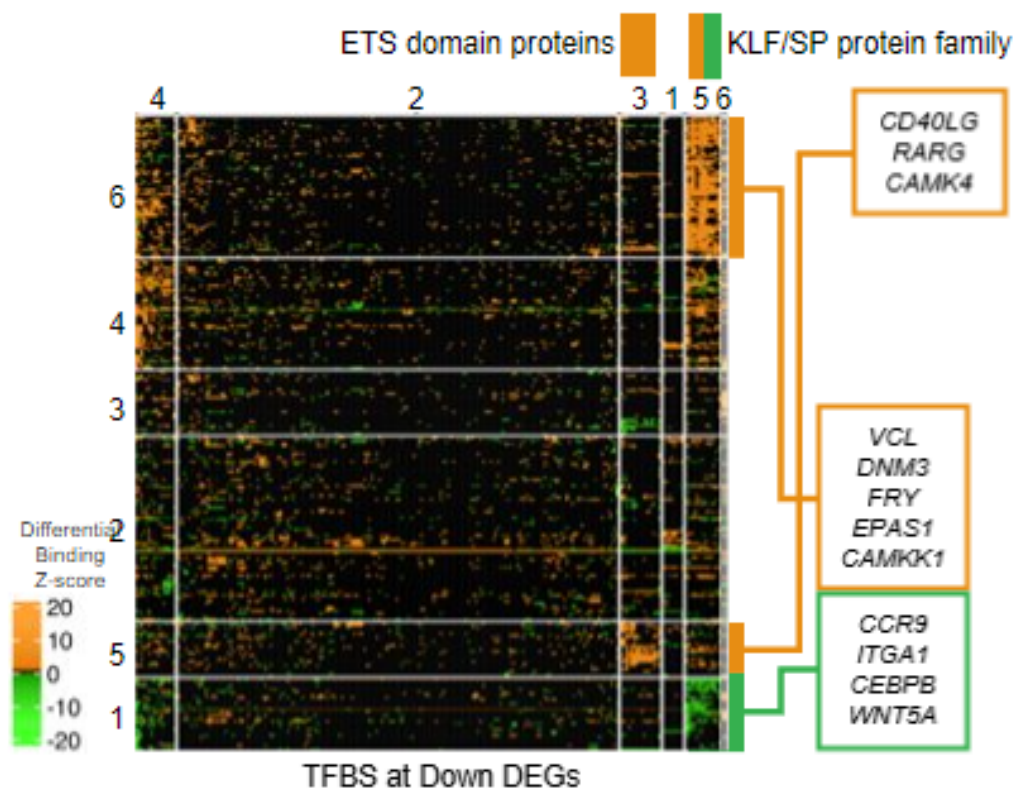


Figure 4.20: Heatmap for Clustering of DiffBinding data for the Bound Down regulated subset - Annotated for to show the GRMs

Chapter 5

Discussion and Conclusions

In this chapter we'll discuss the main findings of this dissertation and the conclusions to which we can reach

We'll divide this section into the sections of the results to make the reading easier. Final discussion and conclusion on the Gene Expression vs Differential Chromatin Accessibility is in section 5.1, for the Standardization of the Fast Pre Ranked Gene Set Enrichment Analysis go to section 5.2 and finally for the final analysis of the clustering of the Digital Footprinting Analysis results go to section 5.3.

5.1 Analysing Gene Expression vs Differential Chromatin Accessibility

While the linear regression found might be small (with a slope of just 0.061922) the significance of the finding is a lot bigger than it seems. Even with a dataset with just 3 replicates per cell type and stage, a significant R-squared of 0.2308 (which is significant in datasets with high variability such as genomic datasets) is obtained.

We can conclude that in healthy conditions, the chromatin is more open in T regulatory cells than in T conventional cells, leading us to observe that epigenetics plays a big role in the development of CD4+ T cells.

From the image Figure 4.2 we can also observe that typical genetic markers of the tTreg lineage such as *FOXP3*, *STAT4* and *IL2RA* can be found in the tTreg side with the most open chromatin, while markers for tTconv such as *TGFA* and *IL7R* are found on the tTconv side. These markers have been all added after the bubble plot and linear regression were calculated, to uncover where these gene markers were in the plot.

The protocol for this kind of analysis needs to be standardized as similar analysis are hard to find and compare. The discovery and validation of possible correlations between chromatin accessibility and gene expression require a stable protocol for the units used with each variable, the type of plots performed and how correlations are calculated. Heteroscedasticity is also common in such datasets and valid correlations might be ignored due to lack of statistical knowledge.

A next step should be repeating the same protocol for other stages of t cell development and with datasets reflecting particular immune illnesses in order to discover significant differences with this regression and if disease can significantly alter this balance.

5.2 Gene Set Enrichment Analysis - Standardizing the Algorithm

The necessity of a standardized protocol for the application of the Gene Set Enrichment Analysis became evident during this project.

This kind of algorithms allows us to measure how represented is our data in annotated datasets associated to metabolic pathways or disease. It's demonstrable more reliable than simple gene ontology Mi et al. (2019) as it associates both a ranked list and also the gene expression values.

By standardizing the procedure and having it run against the whole MSigDB we can assure at least that, for this thorough library, significant results are not ignored and that research bias does not exist.

This analysis produced the most relevant results in the Hallmark subset of MSigDBas it can be seen described in section 4.2. While some results appear with the overall dataset and the immunity subset, the most relevant are found in Hallmark. Some of the results were expected such as the enrichment in *HALLMARK_IL2_STAT5_SIGNALING* and *HALLMARK_TNFA_SIGNALING_VIA_NFKB*, some results were surprisingly interesting such as *HALLMARK_ESTROGEN_RESPONSE_LATE* and *HALLMARK_GLYCOSIS*.

It's important to reinforce the importance of the finding of significant signalling molecules such as *IL2RA*, *IL2RB*, *IL10RA*, and *CTLA4* in the case of *HALLMARK_IL2_STAT5_SIGNALING*, or *DUSP4*, and *IL15RA* both in *HALLMARK_TNFA_SIGNALING_VIA_NFKB* and *HALLMARK_INFLAMMATORY_RESPONSE*.

It's also crucial that we are finding in the overlap of the significantly enriched signatures, several transcription factors of relevance in the T cell development context, namely *IKZF4* (*Eos*), *BATF*, *IRF8*, the NFKB2 pathway inhibitor *NFKBIZ* (*HALLMARK_IL2_STAT5_SIGNALING*; *NFKB2*, *REL*, *RELB*, their inhibitor *NFKBIE*, *BHLHE40*, *KLF6*, *NR4A3*, *BCL3* and *BCL2A1* (*HALLMARK_TNFA_SIGNALING_VIA_NFKB*).

Most of these conclusions developed after the heatmap visualization was created, revealing to be a powerful way to analyse results from gene set enrichment analysis. Heatmaps such as the one created in Figure 4.7 allows us to check for how much of the dataset in question is represented in the curated database and checking for overlaps between significantly enriched pathways. The heatmap should be added to the Gene Set Enrichment Analysis protocol in a standard analysis due to its usefulness

5.3 Clustering of Digital Footprinting Analysis

Finally we reach the analysis of the results obtained in the Clustering of Digital Footprinting Analysis.

This method revealed itself to be quite reliable in unearthing patterns existent in the dataset. As a new method to analyse this kind of data the protocol still requires some cleaning but it demonstrates potential. It demonstrates that 2 K means clustering algorithms on a dataset considering the variable that unites two distinct components of the system (such as genes and transcription factor binding sites in this case) is a useful method to discover the patterns in those components of the system separately and also crossed patterns between both when observed in the heatmap.

The most interesting results come especially from the differential binding heatmaps in the Up Regulated Genes subset Figure 4.19 and Down Regulated Genes subset Figure 4.20, where we can observe which gene/TFBS combinations are more relevant towards tTreg development and tTconv development.

In the Up Regulated Genes subset Figure 4.19 we find mostly 2 promoted groups in tTregs: **Row 6, Col 5** and **Row 5, Col 3** where we find the AP1 family of proteins. AP-1 has pleiotropic effects and plays a crucial task in the T cell family, being identified as playing a role in T cell activation, Th cell differentiation and T cell anergy. We also find the **ETS/ETV/ELF** family, which possesses the Ets domain. It is identified in B and T cell development and in biological processes such as haematopoiesis and tumour progression.

In the Down Regulated Genes subset Figure 4.20 we find the KLS/SP family with repressor effects **Row 5, Column 7** and interestingly find the same group in the Up regulated genes with activator effect in **Row 4, Column 7** and repressor effect in **Row 1, Column 5**. Studying the dual effect of this subgroup should be a goal in future projects.

This methodology has thus proved itself useful in the discovery of interesting regulation patterns as the consistency of the results over the several tests performed and the association of the patterns discovered to T reg development shows.

Further developing this analysis protocol to generalize it to accept new datasets without an issue and improving the colour scheme algorithm will hopefully turn this methodology into a useful comparison method between cell development stages and healthy/illness associated datasets as the heatmaps have an ability to become fingerprints of the regulation patterns with the dataset.

Chapter 6

Limitations And Recommendations For Future Works

This work has successfully used data science techniques to improve the discovery of important regulatory pathways in the development of CD4+ T cells. The discoveries have allowed the laboratory to uncover new regulatory pathways for this subset of cells and new methods for the analysis of CD4+ T reg cells were developed.

The most significant limitations were found in the limited dataset available (only 3 replicates) which can interfere with the validity of the results obtained. Nonetheless it has set a good foundation for more data science to be incorporated in such projects. Next logical steps in this project are:

- Transform the standardization of the Fast Pre-ranked Gene Set Enrichment Analysis with the MSigDb into an R package, further distributing the standardization and easing the using of this technique by others;
- Compare the results of the Gene Expression vs Differential Chromatin Accessibility in CD4+ T Cells for healthy subjects with the same protocol for patients with Immune Diseases in search for significant changes;
- Extend the analysis performed to bigger datasets (more replicates) and/or to disease specific datasets to compare with this healthy subjects dataset;
- Extend the data to incorporate more genetic, genomic, proteomic, metabolomic and clinical data, broadening the analysis potential of such data and increasing the depth of the multi-omic approach to this analysis;
- Incorporate the findings of this project in a future classification algorithm for Complex Variable Immunodeficiencies (with cured data for healthy subjects and clinical immunology clinical cases) that could analyse new clinical immunology cases in a hospital setting and help predict outcomes of the case from the first data obtained from the patient, thus saving critical time for the patient to reach a full diagnosis.

Bibliography

- Achim, Z. and Torsten, H. (2002). Diagnostic checking in regression relationships. *R News*, 2:7–10.
- Andrews, S. (2010). Babraham bioinformatics - fastqc a quality control tool for high throughput sequence data.
- Arosa, F. A., Cardoso, E. M., and Pacheco, F. C. (2012). *Fundamentos de Imunologia*. LIDEL, 2nd edition.
- Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., Fust, A., Preussner, J., Kuenne, C., Braun, T., Kim, J., and Looso, M. (2020). Atac-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nature Communications* 2020 11:1, 11:1–11.
- Buenrostro, J., Wu, B., Chang, H., and Greenleaf, W. (2015). Atac-seq: A method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, 109:21.29.1.
- Chapel, H., Haeney, M., Misbah, S. A., and Snowden, N. Essentials of clinical immunology. page 365.
- Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B., Kathiria, A., Cho, S. W., Mumbach, M. R., Carter, A. C., Kasowski, M., Orloff, L. A., Risca, V. I., Kundaje, A., Khavari, P. A., Montine, T. J., Greenleaf, W. J., and Chang, H. Y. (2017). An improved atac-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods* 2017 14:10, 14:959–962.
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., and Li, H. (2021). Twelve years of samtools and bcftools. *GigaScience*, 10.
- Davis, M. M., Tato, C. M., and Furman, D. (2017). Systems immunology: just getting started. *Nature Immunology* 2017 18:7, 18:725–732.
- Dheilly, N. M., Adema, C., Raftos, D. A., Gourbal, B., Grunau, C., and Pasquier, L. D. (2014). No more non-model species: The promise of next generation sequencing for comparative immunology. *Developmental & Comparative Immunology*, 45:56–66.
- Edgar, J. D. M. (2011). Clinical immunology. *The Ulster Medical Journal*, 80:5.

- Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., and Mathelier, A. (2020). Jaspas 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 48:D87–D92.
- Germain, R. N. (2002). T-cell development and the cd4–cd8 lineage decision. *Nature Reviews Immunology* 2:5, 2:309–322.
- Godinho-Santos, A., Foxall, R. B., Antão, A. V., Tavares, B., Ferreira, T., Serra-Caetano, A., Matoso, P., and Sousa, A. E. (2020). Follicular helper t cells are major human immunodeficiency virus-2 reservoirs and support productive infection. *The Journal of infectious diseases*, 221:122–126.
- Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32:2847–2849.
- Hart, G. T., Hogquist, K. A., and Jameson, S. C. (2012). Kruppel-like factors in lymphocyte biology. *Journal of Immunology (Baltimore, Md. : 1950)*, 188:521.
- Hayatsu, N., Miyao, T., Tachibana, M., Murakami, R., Kimura, A., Kato, T., Kawakami, E., Endo, T. A., Setoguchi, R., Watarai, H., Nishikawa, T., Yasuda, T., Yoshida, H., and Hori, S. (2017). Analyses of a mutant foxp3 allele reveal batf as a critical transcription factor in the differentiation and accumulation of tissue regulatory t cells. *Immunity*, 47:268–283.e9.
- Hori, S., Nomura, T., and Sakaguchi, S. (2017). Control of regulatory t cell development by the transcription factor foxp3. *Journal of Immunology*, 198:981–985.
- Hu, G., Cui, K., Fang, D., Hirose, S., Wang, X., Wangsa, D., Jin, W., Ried, T., Liu, P., Zhu, J., Rothenberg, E. V., and Zhao, K. (2018). Transformation of accessible chromatin and 3d nucleome underlies lineage commitment of early t cells. *Immunity*, 48:227–242.e8.
- Katagiri, T., Kameda, H., Nakano, H., and Yamazaki, S. (2021). Regulation of t cell differentiation by the ap-1 transcription factor junb. <https://doi.org/10.1080/25785826.2021.1872838>, 44:197–203.
- Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., Baranasic, D., Arenillas, D. J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W. W., Parcy, F., and Mathelier, A. (2018). Jaspas 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, 46:D1284–D1284.
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* 2013 14:4, 14:1–13.
- Kondělková, K., Vokurková, D., Krejsek, J., Borská, L., Fiala, Z., and Ctirad, A. (2010). Regulatory t cells (treg) and their roles in immune system with respect to immunopathological disorders. *Acta medica (Hradec Kralove)*, 53:73–77.

- Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods* 2012 9:4, 9:357–359.
- Lee, W. and Lee, G. R. (2018). Transcriptional regulation and development of regulatory t cells. *Experimental and Molecular Medicine* 2018 50:3, 50:e456–e456.
- Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., and di Palma, M. A. (2021). robustbase: Basic robust statistics. R package version 0.93-9.
- Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P. D. (2019). Panther version 14: more genomes, a new panther go-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, 47:D419–D426.
- Mouly, E., Chemin, K., Nguyen, H. V., Chopin, M., Mesnard, L., de Moraes, M. L., Burlen-defranoux, O., Bandeira, A., and Bories, J.-C. (2010). The ets-1 transcription factor controls the development and function of natural regulatory t cells. *The Journal of Experimental Medicine*, 207:2113.
- Petrovsky, N. and Brusica, V. (2002). Computational immunology: The coming of age. *Immunology and Cell Biology*, 80:248–254.
- Íris Caramalho, Nunes-Cabaço, H., Foxall, R. B., and Sousa, A. E. (2015). Regulatory t-cell development in the human thymus. *Frontiers in Immunology*, 6:395.
- Ross, B. L.-C. S. R. (1916). An application of the theory of probabilities to the study of a priori pathometry.—part i. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 92:204–230.
- Sakaguchi, S., Mikami, N., Wing, J. B., Tanaka, A., Ichiyama, K., and Ohkura, N. (2020). Regulatory t cells and human disease. <https://doi.org/10.1146/annurev-immunol-042718-041717>, 38:541–566.
- Salmon-Divon, M., Dvinge, H., Tammoja, K., and Bertone, P. (2010). Peakalyzer: Genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics* 2010 11:1, 11:1–12.
- Sergushichev, A. A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*, page 060012.
- Silva, S. L., Albuquerque, A. S., Matoso, P., de Muijder, B. C., Cheynier, R., Ligeiro, D., Abecasis, M., Anjos, R., Barata, J. T., Victorino, R. M., and Sousa, A. E. (2017). Il-7-induced proliferation of human naive cd4 t-cells relies on continued thymic activity. *Frontiers in Immunology*, 8.
- Silva, S. L., Fonseca, M., Pereira, M. L., Silva, S. P., Barbosa, R. R., Serra-Caetano, A., Blanco, E., Rosmaninho, P., Pérez-Andrés, M., Sousa, A. B., Raposo, A. A., Gama-Carvalho, M., Victorino, R. M., Hammarstrom, L., and Sousa, A. E. (2019). Monozygotic twins concordant for common variable immunodeficiency: Strikingly similar clinical and immune profile associated with a polygenic burden. *Frontiers in Immunology*, 10.

- Singh, B., Schwartz, J. A., Sandrock, C., Bellemore, S. M., and Nikoopour, E. (2013). Modulation of autoimmune diseases by interleukin (il)-17 producing regulatory t helper (th17) cells. *The Indian Journal of Medical Research*, 138:591.
- Slatko, B. E., Gardner, A. F., and Ausubel, F. M. (2018). Overview of next generation sequencing technologies. *Current protocols in molecular biology*, 122:e59.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102:15545–15550.
- Tate, P. and Seeley, R. R. (2009). *Seeley's Principles of Anatomy and Physiology*. McGraw-Hill.
- Todorov, V. and Filzmoser, P. (2010). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32:1–47.
- Wickham, H. (2009). *Ggplot2 : elegant graphics for data analysis*. Springer.
- Wu, D. and Smyth, G. K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40:e133.
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of chip-seq (macs). *Genome Biology* 2008 9:9, 9:1–9.