Computational analysis of isoform level regulation in T helper cells

António Maria Forsado Martins Gonçalves

Instituto Superior Técnico, Lisboa, Portugal

October 2019

Abstract

T and B lymphocytes are the main mediators of adaptive immunity, responsible for highly effective immune responses to pathogens using specific antibodies. T follicular helper lymphocytes (Tfh cells) play a major role in high affinity antibody production through its interaction with B cells in the germinal center. T helper lymphocytes (Th cells) are another subset of T lymphocytes that attract other immune cells at the site of infection and can further be classified into different subsets, Th1 and Th2, based on the elicited immune response. The advances of high-throughput RNA sequencing technology has allowed us the accurate quantification of transcripts, while the increase in computational power makes differential transcript usage (DTU) analysis between different conditions possible. This type of analysis enables the detection of differential transcript expression (DTE), isoform switch and alternative splicing events. In thesis we have conducted a DTU analysis of Th and Tfh cells under different immune response types. We observe the presence of DTE and isoform switches is distributed equally and in the same relative percentage in both cases. 16% of the genes that showed DTU were also found in a differential gene expression (DGE) analysis. The transcripts obtained in the analysis were found to be involved in regulation mechanisms due to their transcript type and differences in the unstraled regions (UTRs) of the transcript. This work allowed to conclude that there is an important level of regulation by differential transcript usage that would not be found in a simple differential gene expression analysis.

Keywords: Tfh cells, Th cells, differential transcript usage, isoform switch, alternative splicing

1. Introduction

A great part of diseases nowadays are due to or partly affected by the disregulation of the immune system. The immune system has evolved to protect us from infectious agents and dangerous toxins known as pathogens. With the development of the field of immunology and discoveries on how the immune system works it was possible to achieve an amazing increase in better public health and overall life conditions through the creation of vaccines and antibiotics. This is easily observed if we look at the sharp reduction of diseases like tetanus, tuberculosis, mumps and hepatitis B since the introduction of vaccination plans. Some diseases such as polio, diphtheria, measles and rubella were even erradicated, or nearly erradicated, with the introduction of these plans.

1.1. Immunology

Immunity can be defined as the body's ability to fight against the infectious agents and pathogens and their negative effects on the human body. An important mechanism of the immune response is adaptive immunity.

Adaptive immunity is mediated by lymphocytes and these can be grouped into B lymphocytes (or B cells) and T lymphocytes (or T cells). Most lymphocytes are present in the blood and circulate through our body to detect antigens. When naive lymphocytes encounter antigens, they get activated and become effector lymphocytes. This step is called priming and it leads to differentiation of naive inexperienced lymphocytes to differentiated specialized lymphocytes. This process takes a few days to weeks and it leads to the development of the previously mentioned immunological memory.[1]

When Naive T cells become activated, they differentiate either into cytotoxic T lymphocytes (CTLs), which are CD8+ cells and have the role of killing any infected cell they may find, or into helper T cells, which are CD4+ cells and function by secreting cytokines that will attract other cells of the immune system, such as macrophages, neutrophils and eosinophils, to the site of infection thus helping in the destruction of the pathogens present in such case.[2]

The most studied CD4+ helper T cells fall into two categories, Th1 and Th2, according to the type

of cytokine being produced in the immune response. Th1 cells are characterized by interferon gamma (IFN-γ) secretion and are normally associated with bacterial/viral infections. Th1 release of IFN-γ promotes the secretion of IgG2a. Th1 cells are driven by a transcription factor named T-bet. Th2 cells are defined by secretion of IL-4, IL-5 and IL-13, and are important to protect against parasitic infections. Th2 cells are known to induce the secretion of antibodies of type IgE and IgG1, this last one only in mice. Th2 is driven by the transcription factors Gata-3.[3]

There is also another category of specialized effector CD4+ cells that is developmentally different from the previous two: Th17 cells which express interleukin seventeen (IL-17) and promotes responses towards extracellular pathogens, specifically with the attraction of neutrophils as a way to remove certain bacteria and fungi.[4, 5]

A special subset of T cells characterized by expression of Foxp3+ are known as T regulatory cells (Tregs) due to their regulatory function. These cells are known to regulate the activity and proliferation of auto-reactive T cells. [6] But none of these subsets were found to be capable of interacting with B cells.[7] Further information related to the subset of CD4+ T cells that interacts with B cells will be discussed ahead.

A big difference between T and B cells is that B cells do not develop in the thymus and in fact, their maturation occurs in the bone marrow in mammals or bursa in birds. Reason why they are called B lymphocytes or B cells, from bursa.[8]

Naive B cells unlike T cells are not activated by antigen presenting cells (APCs). Their B cell receptor(BCR) recognizes the antigen without presentation and when this happens, they differentiate either into memory B cells or plasma cells capable of producing large amounts of specific antibodies. The reaction leading to antibody production here described takes place in a anatomical structure between the cortex and the paracortex of the lymphatic nodes and spleen: the germinal center.

The germinal center is a transient phisiological structure within secondary lymphoid organs that results from the rapid expansion of B cells. This structure is the site where B cells go through the processess of clonal expansion, somatic hypermutation and affinity maturation, in order to become the cells that produce highly specific antibodies.

The activation of a B cell is done by exposition to antigen and by a special subset of CD4+ T cells called follicular helper T cells or Tfh cells. This activation happens when the B cell and the helper T cell meet in the outer edge of the follicle of the lymphoid organ and that's where Tfh cells get their name from, because they will be localized in the follicle.[9]

The phenotype of Tfh cells is characterized by expression of CXCR5, ICOS, PD-1, BCL-6 and production of IL-21.[10] Tfh differentiation starts by activation with an antigen as presented by antigen presenting cells (APCs) on the T cell zone of secondary lymphoid tissue. Then the T cell migrates to the outer edge of the follicle of the lymphoid organ, where it interacts with B cells. Tfh cells can help B cells to differentiate in the germinal center into antibody producing plasma cells, by giving survival signals to the antigen specific B cells. Without Tfh cells humoral responses do not occur.[7]

1.2. Biology of alternative splicing

CD4+ T cells are regulated by gene expression and the mechanisms outline previously. However, very poor information or none at all is known about particular transcript expression, more specifically the difference of transcript expression of a given gene. In this thesis, I explored the impact of differential transcript usage (DTU) regulation of Tfh and Th cell subsets as well as the immune responses type 1 and 2.

The relevance of mentioning alternative splicing (AS) within the context of this thesis has to do with the introduction of the idea that the change in transcript expression has its causes in processes of alternative splicing and by looking at transcript expression, some of these mechanisms will be observable and interesting to pinpoint or find. Specifically, alternative splicing might be involved in different processes or mechanisms that explain differences between Th and Tfh cells or differences between infection type 1 and 2.

A layer of special interest in biology of eukaryotes is the processing of pre-mRNA into mRNA in the post-transcriptional RNA processiong, which results in alternative splicing (AS). Its importance comes from the fact that despite the existence of only one given "template" of DNA, when a gene is transcribed it can lead to many different transcripts or isoforms version of the same DNA sequence through the exclusion of introns from the pre-mRNA, which in turn may lead to functional consequences such as a slight different protein being produced which will not take part in the normal pathway. This is part of the reason why higher organisms are more complex.

Previously thought to be junk DNA, the major regions involved in regulation of a given gene are the 5' and 3' untranslated regions or UTRs.[11] The UTR regions are of great importance from the regulation purposes point of view because they define the process of transcription but there is still much that is currently being researched about them.

Alternation of 5' and 3' UTR location is one of

three main mechanisms that are facilitated in alternative splicing (AS). Since the location of 5'UTR and 3'UTR carries important information for the regulation and facilitation of the complex spatial and temporal gene expression, any difference in the location of these regions in two transcripts of the same gene will translate in differences in regulation of gene expression.

The other two mechanisms of AS are exon skipping and intron retention. Exon skipping is when an exon is skipped meaning in transcript transcription which results in a different transcript. An intron is an alternative spliced transcript believed to contain an intronic sequence relative to another protein coding transcript of the same gene. The function of intron retention, or retention of the introns, is to reduce transcript expression of a transcript that is needed less or not at all in a particular physiology of a given cell, acting therefore by tunning the transcription expression.[12]

There are still the cases of lncRNA and nonsense-mediated decay. Long non coding RNA or lncRNA is a transcript whose length exceeds 200 nucleotides but is not translated into protein and is most likely present in the regulation of mRNA expression or some other cellular mechanism. Nonsense-mediated-decay or NMD is a pathway whose main function is to reduce errors in transcription by deleting mRNA transcripts that contain premature stop codons. When a transcript is said to be a NMD it means that it is one of these transcripts with a premature stop codon.

1.3. Sequencing technologies

In 1991, appeared a technique that allowed the characterization and quantification of the RNA in a cell (transcriptome) using Sanger sequencing and biochemical techniques that made use of the reverse transcriptase enzyme. This was called RNA-seq, short for RNA sequencing.[13]

Fast foward a few years and with the development of high-throughput sequencing techniques like Illumina HiSeq® 2000 System, as well as increase in computational power, the field of transcriptomic studies, where the questions of main interest are the characterization of the transcripts, the determination of their coding region and the quantification of gene expression between different biological conditions, really growed. A simple RNA sequencing experiment consists in collecting the sample of interest, extracting its total RNA, creating the cDNA library, sequencing, mapping and then data analysis.

In the end, a transcriptomic study that makes use of RNA-seq data to compare two or more conditions, will make use of high-throughput sequencing of the c-DNA reads from different samples in different conditions in order to accomplish it.

1.4. Alignment and Quantification

Alignment to a reference genome or transcriptome is a crucial step in turning raw data into something measurable, because it will allow for the performance of a quantification.

There are two types of alignment: a spliced alignment against the genome or an unspliced alignment against the transcriptome.

There are two ways to obtain gene level quantification: one is the already mentioned direct count of overlapped fragments for each gene and the other is to perform such quantification at transcript level, counting isoforms, followed by an assignment of the results to genes.[14]

Salmon is an algorithm that uses pseudo-alignment to perform a lightweight quantification of the transcripts.[15] Because of the use of pseudo alignment, the computational requirements of the algorithm decreases substantially and allows for a much faster quantification.

1.5. Differential Gene Expression, Differential Transcript Expression and Differential Transcript Usage details

Differential expression analysis or DEA can be divided into different categories such as: Differential Gene Expression or DGE, where the quantification is the number of counts for a particular gene and the goal is to find which genes show upregulated or downregulated in between experimental conditions; Differential Transcript Expression or DTE, where the quantification is the number of counts for a particular transcript and the goal is to find which transcripts show upregulated or downregulated in between experimental conditions; Differential Transcript Usage or DTU, where both the number of reads per gene and per transcript are taken into consideration and an assessment of how the transcript is used between the different conditions, is made.

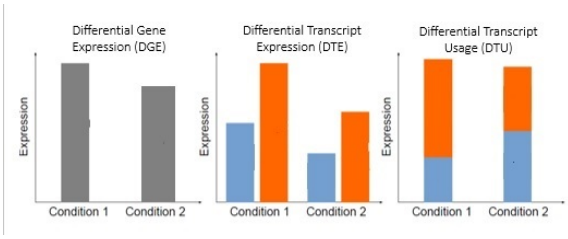

Figure 1: Graphical representation of DGE, DTE and DTU.

It is standard practice in research laboratories to do a DGE experiment when a transcriptomic experiment is proposed. The number of DTU analysis is far inferior (almost nonexistent) when compared

to the number of DGE analysis because scientists are not being aware that such methods even exist and also because when RNA-seq first appeared, its technical capabilities were not as precise as they are today.

Figure 1 is a graphical representation of the types of DEA analysis and while it only considers two conditions and two transcript, its implications can be carried to a higher number of conditions or transcript, the relevant aspect is the idea behind such DEA.

When comparing DTE to DGE, the analysis is indeed the same except that while one is at gene-level the other is at transcript level however it is observed that DTE leads to a lot of false positives due to the fact that a transcript, being smaller in size than a gene, will have more positive but incorrect matches across the whole genome than a gene.[16] Moreover, DTE will point to a change in transcript but it won't provide information regarding if this change follows gene expression or if it implies other transcripts' expression decrease, so the realization of DTE will provide less information than just DGE.

When comparing DTE to DTU, a huge difference can be observed in the goal of the analysis. While in DTE the comparison of the transcripts is made to its counter part in the other conditions, in DTU the comparison is made not transcript against transcript but how the transcript is expressed within a given gene. Also since at least one isoform must change expression for it to be considered DTU, DTE is implied in DTU.

DTU is a different method of transcriptomic study that serves as a complement to a DGE study, and both combined provide us with the best picture of transcriptomic data. The information obtained by performing only one type of analysis, either DGE or DTU, separately of each other will always fall short of the information obtained where both analysis are conducted in a way where they are complementar to each other. This complementary of analysis comes from the ability that DTU has to detect not only the cases that one might find in a DGE study, but also the ones which normally would be overlooked.

This allows for the conclusion that the best approach to a proper analysis of RNA-seq experiments is to perform both a DGE and a DTU analysis and combine its results for gathering the most meaningful information and it will be what is going to be done in this thesis. The purpose of a DTU analysis is to provide information that would not be otherwise obtained in a DGE analysis. This information can be seen as events or cases of transcript usage that wouldn't be captured in DGE.

After considering which cases might be interesting to find as a result of a DTU analysis, we came to the conclusion that there would be two major situations or cases: Case 1, DTE impacting Gene expression, when there is a difference in one particular transcript's expression which completely changes the gene expression in between conditions. With the performance of a DTU, the transcript of importance in this case can be identified, without the need for a parallel DTE analysis. This level of transcript precision would not be possible in a DGE analysis; and Case 2, Isoform Switch, a phenomenon where different transcripts belonging to the same gene are expressed completely differently in between conditions,and its expression changes for the other, something which a regular DGE experiment would never be able to notice.

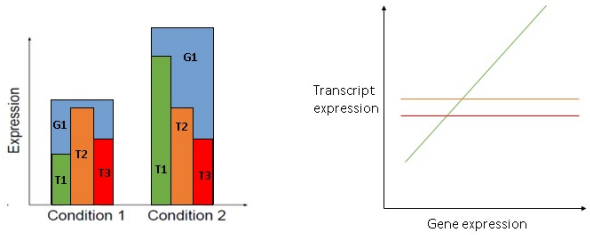

Figure 2: Case 1 - Differential transcript expression.

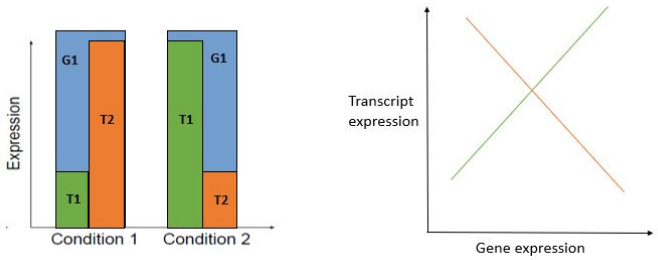

Figure 3: Case 2 - Isoform Switch.

1.6. Differential expression analysis and statistics

A usual pipeline of differential gene expression follows the following steps: alignment, quantification, filtration, normalization, specification of the statistical model and estimation of model parameters, statistical inference on the relevant parameters and adjustment for multiple testing.

After the quantification step, we have a matrix of counts where the rows are the features of interest, such as genes, transcripts or exons, and the columns are the samples used for the experiment, samples which relate to the different conditions being tested, to perform our statistics and other procedures on. This count matrix is going to be filtered based on the parameters that will allow for the performance of the most correct analysis.

In differential expression analyses, gene expression is rarely considered at the level of raw counts since libraries sequenced at a greater depth will result in higher counts. Therefore the step of count normalization is necessary to make accurate comparisons of gene expression between samples. The counts of mapped reads for each gene is proportional to the expression of RNA (interesting) in addition to many other factors (not interesting). Normalization is the process of scaling raw count values to account for the "interesting" factors. This way, the expression levels are more comparable between samples and more meaningful informations maybe retrieved.

The statistical model, in the case of a DGE analysis, is considered to be a negative binomial distribution and the model parameters to be estimated based on the count matrix are the average of feature expression and the dispersion per feature.

Statistical inference involves declaring or stating a null hypothesis that there is no differential expression between conditions and then testing this against the hypothesis where this statement is false. It envolves comparing a full model against a null model where one or more parameters were removed. The result is a list of p-values, where the smaller the value, higher the significance. Multiples testing controls for the possibility of existing false positives values.

1.7. Current methods

Limma[17] fits a linear model to each row of the data. The main aspect of limma is that it analyses the data as whole and not as separate comparisons, enabling the sharing of information between samples. Limma was designed for microarray data and not for RNA-seq data, however through the use of precision weights it enables RNA-seq count data to be used in linear models. Limma is extended to test for differential splicing events when exon or transcript-level expression data is available.

EdgeR[18] models count data using the NB distribution and employs an empirical Bayes method to moderate the degree of overdispersion between genes. An empirical Bayes procedure is used to shrink the dispersions towards a consensus value, effectively borrowing information between genes and differential expression is assessed for each gene using a Fisher's exact test but adapted for overdispersed data. EdgeR also provides an extension to test for DTU.

DESeq2[19] starts with a count matrix that will be modelled with a NB distribution and employs an Empirical Bayes shrinkage for dispersion estimation and fits GLMs to each gene. Then, a Wald test is performed to see if each model coefficient differs significantly from zero. The obtained Pvalues are then corrected by the Benjamini and Hochberg procedure. DESeq2 is only used for a DGE analysis and not in a DTU analysis.

DEXSeq[20] was designed to test for differential exon usage (DEU). It uses GLMs to model read counts and the counts are assumed to follow a NB distribution with mean and dispersion as the main parameters. The testing part is done by a likelihood ratio test which tests against the null hypothesis that none of the conditions influences exon usage. If positive, it can safely be concluded that this gene is affected by alternative isoform regulation but not much else.

SUPPA2[21] is a method which instead of performing statistics on transcript quantification, it identifies differential splicing events at the junction level that are used to reveal DTU.

RATs[22] identifies DTU independently at both the gene and transcript levels by detecting changes in proportions.

DRIMSeq is a tool that was designed to model specifically gene-level and transcript-level isoform expression in order to perform a DTU analysis. For this reason it will be the tool used in this thesis and extensive details will be further explained and characterized.

2. Objective

While there is a clear distinction between different types of Th cells, there is no consensus regarding classification of different subsets of Tfh cells.[7] Is it possible that distinct Tfh subsets provide cues to B cells regarding which type of antibody to produce based on their subtype.

Many transcriptomic studies have shown different profiles between Th and Tfh cells, based on gene expression studies. It has also been shown that alternatively spliced products play a significant role in defining the transcriptome and functioning of these cells.[23] While the gene expression studies between Tfh and Th and between putative Tfh1 and putative Tfh2 are on-going in the lab, my thesis work focused on dissecting the underlying differential transcript usage (DTU), with subcategories as Isoform Switch and Differential transcript expression, that may also shape the transcriptome of these cells. This project explores particularly the different isoforms used and their classification that may define the underlying regulation of these cells.

The main aims of the project are: to identify the best strategy for carrying out a DTU analysis; implementation of DRIMSeq package to design a DTU pipeline; and downstream analysis involving classification of different isoforms and their impact on gene expression.

3. Implementation

3.1. Experimental design for transcriptome data

In the laboratory where this master thesis was conducted, an experiment was designed and performed in order to obtain immune cells of interest and to use them in the construction of a proper dataset of RNA-seq: two inducers of a type 1 immune response, CpG and nCpG, as well as, an inducer of a type 2 immune response, IFA, were used to immunize mice without these mice cells differentiated, therefore only type 1 or type 2 pathway dependent on the that responded to each infection type.

3.2. Tximport

During the sequencing of the samples, fastq files are created per sample. The Salmon algorithm of pseudo-alignment described in subsection 1.4, was used to perform the alignment of the fastq files so that the count matrix can be generated ahead in the analysis.

All steps and computations that were performed in this analysis were conducted in R version 3.5.3.

An R function called tximport was used to create a table or matrix of counts which has features such as genes or transcripts as rows and the samples obtained in the experiment as columns.

Since a DTU analysis is being performed, the counts were imported at transcript level and were scaled using the median transcript length among the isoforms of the given gene and then scaled again using the library size. From the resulting counts matrix, the transcripts that weren't responsible for coding of proteins were removed using the protein coding file. Transcripts that belonged to ribossomal genes or to H2 genes were also removed.

3.3. DTU method

The modelling of univariate data through the negative binomial (NB) distribution has been described in subsection 1.5. However, in order to properly perform a DTU analysis, we take into consideration that, since the different isoforms are a consequence of alternative splicing, gene expression is a multivariate expression of these isoforms. This creates a situation of modelling multivariate data which requires an extension of the NB distribution to multidimensional space which is known as Dirichlet-multinomial (DM) distribuition. This creates a way to account for the intrinsic dependency between quantification values of the isoforms. Having this statistical framework in mind, researchers developed a R package called DRIMSeq, as in Dirichlet-multinomial sequencing, to model the alternative usage of transcript isoforms from RNA-seq count data and tested it on real datasets reaching the conclusion that it performs well for transcript counts.[24]

3.4. Normalization and Filtration

Normalization was previously done in the tximport command when selecting for dtuScaledTPM, where the counts are scaled using the median transcript length among isoforms of a given gene and then to library size. The choice of dtuScaledTPM has to do with the fact that this option was purposely built for DTU analysis and also because within a given gene the values are all scaled by the same median transcript length leading to a better assessment of proportions later on.

When doing a DTU analysis, there is a need to filter both at gene level and at transcript level. Following that thought, a function dmFilter in DRIMSeq allows for a more efficient filtration according to four parameters.

In the case of the dataset used, it was considered that the filtering parameters should be that: a given gene had to be expressed in all samples to be considered for analysis, a given transcript had to be expressed in at least half of the samples, the gene expression for each sample should be at least of 7 read counts and the transcript expression for each sample should be at least 4 read counts.

The dataset before filtration was characterized as having 21807 genes which possessed 91156 transcripts. After filtration, the new dataset had 3404 genes which possessed 14596 transcripts.

3.5. Data exploration

Our data exploration was conducted at gene-level to find gene-level information about the data. Using the gene names of the filtered dataset, the count matrix was reduced. Normalization was the next procedure and for this it was chosen to perform the base 2 logarithm of 0.5 plus the counts per million normalized count matrix: $log_2(cpm(y)+0.5)$, where y is the count matrix.

A tool called svaseq from SVA package was used to remove batch effect and other underlying noise to highlight the biological question of interest.[25]

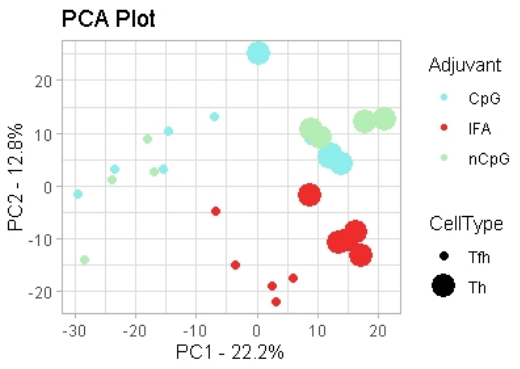

Figure 4: PCA plot.

The normalized data was then cleaned of the un-

wanted variables and the chosen method of data exploration, PCA in our case, was employed. This analysis led to the discovery of 2 major principal components within our data. It can be observed in Figure 4 that the first and second principal components account for, respectively, the type of cell and type of immune response (Adjuvant).

3.6. Precision Calculation, Proportion estimates, Differential Testing and StageR

After filtration of the dataset to 3404 genes, we proceeded to calculate the precision parameter. The precision parameter is related to the dispersion parameter by $dispersion = 1/(1 + precision)$.

In order to correctly calculate the mentioned precisions, these have to be calculated according to a model of known variables. Using the information from the PCA, an additive linear model that consisted in cell type, immune response type and batch effects was created: $\sim CellType + ImmuneResponse + BatchEffects$.

The function dmPrecision in DRIMSeq performs these calculations and also finds the value of the common precision among 10% of the considered genes, which is then used to find values of genewise precision and mean expression of each gene.

After Precision calculation, the same linear model described before is used for fitting and calculating the regression coefficients and the isoform proportions for each gene and per sample in the step of Proportions estimation, through the use of the function dmFit in DRIMSeq. The regression coefficients are defined by the design matrix built from the same model.

The testing step is done using the dmTest function in DRIMSeq and it tests for differences in isoform usage between variables of the linear model: Cell Type or Immune Response. It does so by creating a null design matrix from the previous design matrix without one of the variables in the linear model, estimating model parameters on this null design matrix and then performs gene-level and isoform-level likelihood ratio tests between the null design matrix and the full design matrix. This step was done twice, one for cell type and another for immune response. The results were a list of adjusted P values where it can be seen if a specific isoform is used or not among conditions, by having in mind that a p-value bellow 0.05 corresponds to a gene or transcript that is most certainly different between conditions.

There is the necessity to apply a correction to the adjusted p values obtained because DRIMSeq has been shown to generate poor sensitivity and perform better after stage-wise correction.[22] The StageR package [26] was developed with this end in mind and performs an effective filtering by doing a screening stage and a confirmation stage.

4. Results

4.1. Classification of DTU cases

The end result of the pipeline is a list of genes, its transcripts and the controlled adjusted p-values for both. This list consists in 369 genes and 556 transcripts for the case where the testing variable of interest was cell type, and 169 genes followed by 254 transcripts for immune response. The immediate interest after this outcome was to known which genes represented the interesting DTU cases mentioned in section 1.5.

The result, which can be observed in figures 7 and 8, was in 205 cases of DTE (56%) and 164 cases of isoform switch (44%) for the testing variable of cell type and in 99 cases of DTE (59%) and 70 cases of isoform switch (41%). It is interesting to see that DTE and Isoform switch showed similar spread between cell type and immune response variables and that DTE and Isoform switch are distributed equally among genes showing DTU.

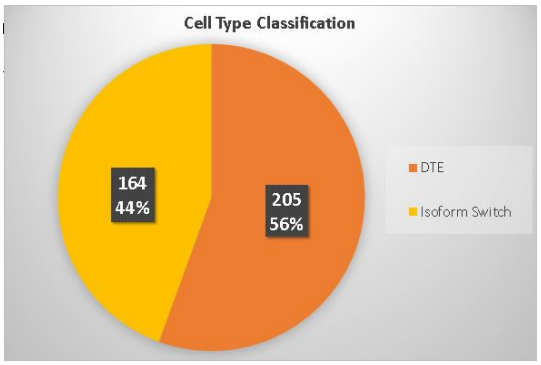

Figure 5: Classification of DTU cases in Cell Type.

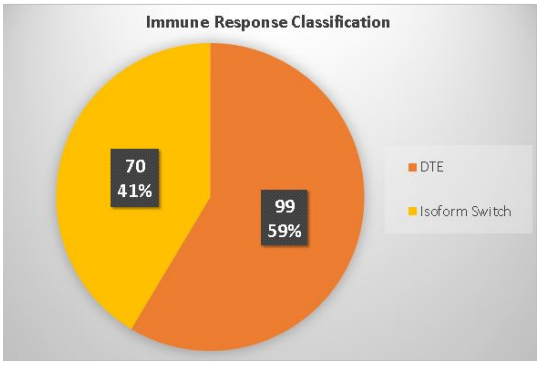

Figure 6: Classification of DTU cases in Immune Response.

When comparing the genes that appear to be differentially expressed in between cell types with the genes that appeared in this DTU analysis, it was observed that 12% of 369 DTU genes belonged to the cell type condition were also found in the differentially expressed genes (Figures 9 and 10).

Interestingly, the finding that only 12% of genes showing DTU impact DGE is different from what has been reported previously [23]. This may partly be due to our closely related cell subsets while in the study its a classification of broad cell subsets.

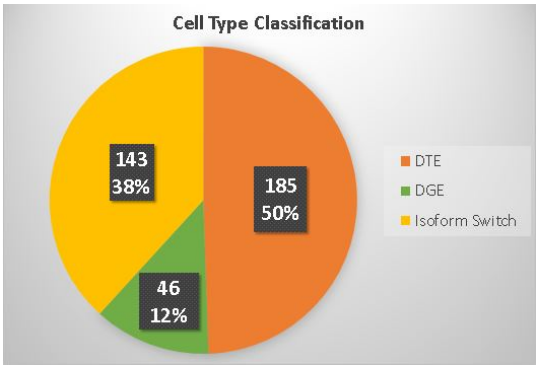

Figure 7: Classification of DTU cases and DGE in Cell Type.

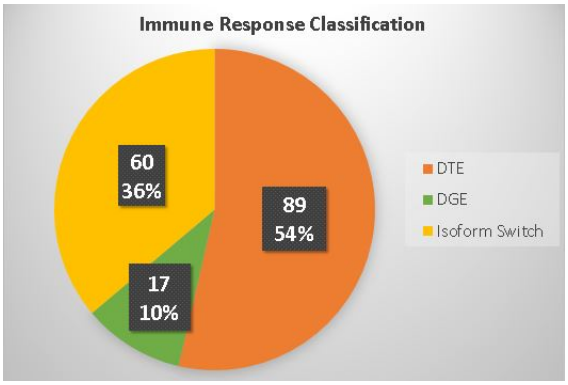

Figure 8: Classification of DTU cases and DGE in Immune Response.

Isoform switch is an event characteristic of cancer cells [27] and the indication that it also happens in great number in Th and Tfh cells, which are immune cells and therefore thought to be in steady state conditions, is something interesting and that would be not obvious at first glance.

4.2. Alternative Splicing cases classification in Isoform Switched Genes

We wanted to know, from within the genes that were previously identified as cases of isoform switch, how were they characterised in terms of alternative splicing information and transcript biotype. Gathering alternative splicing and transcript type information in this context meant finding if there is a change in the 5 prime or 3 prime untranslated region of the transcripts between them or if the transcripts are of the type retained intron or lncRNA.

It was found that of the genes that belonged to the isoform switch case in the cell type condition 9% are alternative 5'UTR, 24% are alternative 3'UTR or alternative polyAdenylation, 15% are different exon transcription, 21% are retained introns, 14% are lncRNA and 18% are Not Classificable (Figure 11).

It was found that of the genes that belonged to the isoform switch case in the immune response condition, 11% are alternative 5'UTR, 24% are alternative 3'UTR or alternative polyAdenylation, 16% are different exon transcription, 20% are retained introns, 16% are lncRNA and 13% are Not Classificable (Figure 12) .

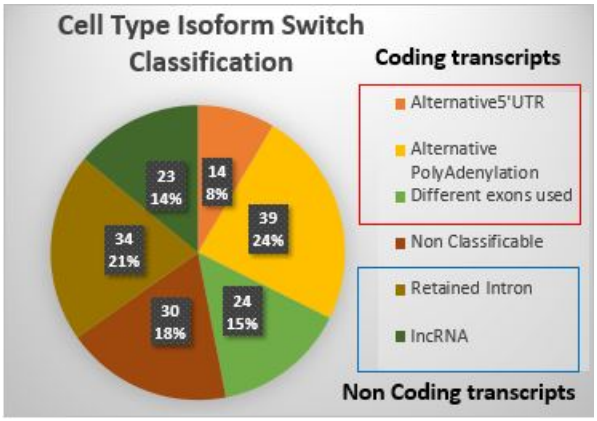

Figure 9: Classification of AS cases in genes showing isoform switch in CellType.

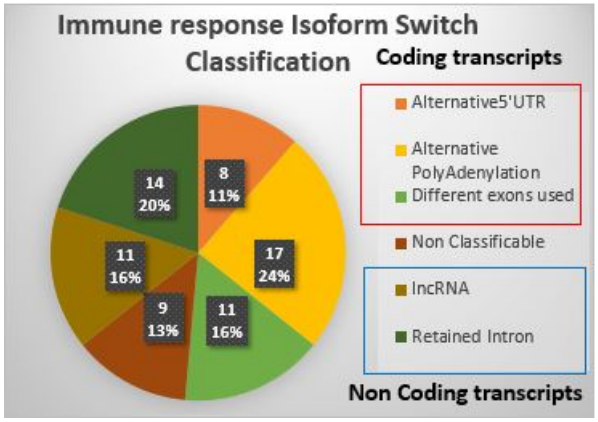

Figure 10: Classification of AS cases in genes showing isoform switch in Immune Response.

This high number of differences found by the DTU analysis and within the transcripts that take place in isoform switch points to the idea that the usage of isoforms in a given gene is responsible for great biological functional impact and is extremely evident in regulatory mechanisms.

This results are the first that show that regulation between Tfh and Th occurs at transcript level. The already mentioned study [23] was done between

very different types of immune cells, so the result to be expected is to exist a lot of difference in terms of DIU, and is different from this one because this one was done in cells that are very similar in essence since they are subtypes of T helper cells, so almost no difference would be expected due to their lineage proximity but that is not what is observed. Here we are comparing cells that are so much more similar to each other that makes us question what kind of regulation does exist, and it is being found that there are these differences in 3'UTR and 5'UTR location in the transcripts of the genes. This itself suggests that tunning is going on and despite what is happening, it is not necessarily at gene level but at transcript level.

4.3. Transcript type classification in DTE Genes
We also wanted to know what was the transcript type of the transcripts of genes showing DTE. It was found that, once more, both conditions showed the same distribution and that over 50% of the transcripts were a protein coding, 17% were lncRNA, moreless 20% were retained introns and the remaining part was nonsense mediated decay (Figures 13 and 14).

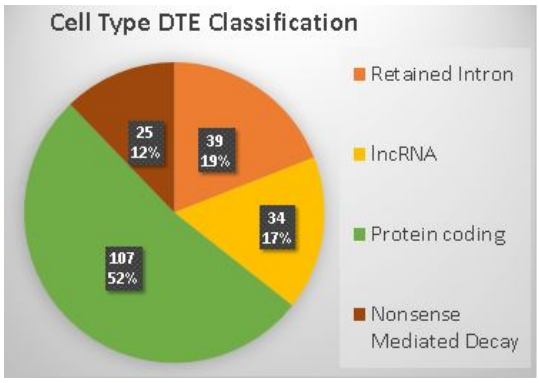

Figure 11: Classification of genes showing DTE in CellType.

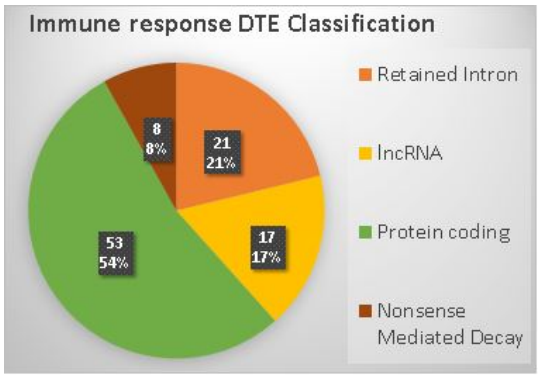

Figure 12: Classification of genes showing DTE in Immune Response.

5. Conclusions

In this thesis, a computational analysis of differential isoform usage was performed using transcriptomic data from immune cells, particularly T helper and T follicular helper cells. We investigated isoform usage for different cell subset categories using DRIMSeq which takes into consideration the gene abundance to correctly model isoform expression. Through this analysis we identified an equal spread of DTE and Isoform Switch cases in our data, with only a small percentage of these cases impacting expression at gene level.

We further categorized these isoforms based on their biotype. In case of genes showing isoform switch, we observe one-third of transcripts changing from coding to non-coding isoforms, while approximately half of isoforms switched between coding transcripts. On further exploration, we found that majority of these isoforms showed either alternate 5' or 3' UTR. Together these results indicate towards differential regulation of these genes between Tfh and Th cells as well as under different immune responses.

Overall, this thesis highlighted the potential significance of regulation of transcript selection, likely to have biological impact, while being often neglected.

Acknowledgements

I would like to thank my supervisor from iMM, Dr. Luis Graça, my supervisor from IST, Professor Susana Vinga, my PhD supervisor, Saumya Kumar, all the colleagues from LGraca Lab at iMM, my friends and my family for all the support and motivation.

References

[1] Kenneth Murphy and Casey Weaver. *Janeway İmmunobiology*. 2017.

[2] Mikaël Ebbo, Adeline Crinier, Frédéric Vély, and Eric Vivier. Innate lymphoid cells: Major players in inflammatory diseases, 11 2017.

[3] Tetsuya Sasaki, Atsushi Onodera, and Toshinori Nakayama. Genome-Wide Gene Expression Profiling Revealed a Critical Role for GATA3 in the Maintenance of the Th2 Cell Identity. *PLoS ONE*, 8(6), 6 2013.

[4] Hai Qi. T follicular helper cells in space-time. *Nature Reviews Immunology*, 16(10):612–625, 2016.

[5] Ivaylo I. Ivanov, Liang Zhou, and Dan R. Littman. Transcriptional regulation of Th17 cell differentiation, 12 2007.

[6] Gajendra M. Jogdand, Suchitra Mohanty, and Satish Devadas. Regulators of Tfh cell differentiation, 2016.

[7] Carola G. Vinuesa, Michelle A. Linterman, Di Yu, and Ian C.M. MacLennan. Follicular Helper T Cells. *Annual Review of Immunology*, 34(1):335–368, 2016.

[8] D. Ribatti, E. Crivellato, and A. Vacca. The contribution of Bruce Glick to the definition of the role played by the bursa of Fabricius in the development of the B cell lineage, 7 2006.

[9] Francis Coffey, Boris Alabyev, and Tim Manser. Initial clonal expansion of germinal center B cells takes place at the perimeter of follicles. *Immunity*, 30(4):599–609, 4 2009.

[10] Nilushi S. De Silva and Ulf Klein. Dynamics of B cells in germinal centres. *Nature Reviews Immunology*, 15(3):137–148, 2015.

[11] Alicia A. Bicknell, Can Cenik, Hon N. Chua, Frederick P. Roth, and Melissa J. Moore. Introns in UTRs: Why we should stop ignoring them. *BioEssays*, 34(12):1025–1034, 12 2012.

[12] Ulrich Braunschweig, Nuno L. Barbosa-Morais, and Qun Pan. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research*, 24(11):1774–1786, 11 2014.

[13] Zhong Wang, Mark Gerstein, and Michael Snyder. Nihms229948. *Nature Reviews Genetics*, 10(1):57–63, 2010.

[14] Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, 2009.

[15] Rob Patro, Geet Duggal, Michael Love, Rafael Irizarry, and Carl Kingsford. Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. *bioRxiv*, page 021592, 2015.

[16] Charlotte Soneson, Michael I. Love, and Mark D. Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 2016.

[17] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47, 2015.

[18] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 11 2009.

[19] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 12 2014.

[20] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017, 10 2012.

[21] Juan L. Trincado, Juan C. Entizne, Gerald Hysenaj, Babita Singh, Miha Skalic, David J. Elliott, and Eduardo Eyras. SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, 19(1), 3 2018.

[22] Kimon Froussios, Kira Mourão, Gordon Simpson, Geoff Barton, and Nicholas Schurch. Relative abundance of transcripts (RATs): Identifying differential isoform abundance from RNA-seq. *F1000Research*, 8, 2019.

[23] A. Ergun, G. Doran, and J. C. Costello. Differential splicing across immune system lineages. *Proceedings of the National Academy of Sciences*, 110(35):14324–14329, 2013.

[24] Malgorzata Nowicka and Mark D. Robinson. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, 2016.

[25] Jeffrey T. Leek. Svaseq: Removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, 42(21):e161, 2014.

[26] Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 107(21):9546–9551, 5 2010.

[27] Kristoffer Vitting-Seerup and Albin Sandelin. The landscape of isoform switches in human cancers. *Molecular Cancer Research*, 15(9):1206–1220, 2017.