Università degli Studi di Napoli *Federico II*

DOTTORATO DI RICERCA IN

FISICA

Ciclo XXIX

Coordinatore: prof. Salvatore Capozziello

CD4+ T celll differentiation explored through Markov process

Settore Scientifico Disciplinare FIS/02

Dottorando
Andrea Piccolo

Tutore
Prof. Mario Nicodemi

Anni 2014/2017

Se qualcuno ti dirà che l'universo è una macchina, pensa com'è bello far finta almeno per un istante che non lo sia.

(Luther Blisset, Lettera al figlio sull'utilità della scuola)

Contents

Contents

Chapter 1

Introduction

Cell transformations and changes in their states impact a number of biological processes, having significant implications, for instance, in medical research. Examples of this range from stem-cell reprogramming, differentiation and carcinogenesis. These are complex processes characterized by alterations in the expression of thousands of genes and proteins in a coordinated way. Although much progress has been made in understanding these phenomena, their accurate and more complete comprehension is still in its scientific infancy.

In this thesis I discuss a project in collaboration with Professor Sarah Teichmann's group (MRC Laboratory of Molecular Biology and EMBL-EBI, Cambridge). We have been working on a cell transformation process of naïve cells into $Th2$ cells (also in this case the data has been purposely derived in Teichmann's lab). Here we investigate population data on cells at different stages of differentiation (FACS data) which contain a wealth of information on the microscopic nature of the events leading naïve cells becoming $Th2$ cells.

Many differentiation processes occur hand-in-hand with a change in cell cycle status:

this can be cell cycle arrest, as in the monocyte to macrophage transition [35], cell cycle entry, as for the pre-adipocyte to adipocyte differentiation [2], and entry and subsequent cell division, as in T helper (Th) cell differentiation [3].

Th cell differentiation is the process where naïve $CD4+$ T cells transition to effector lymphocytes and is central to mammalian adaptive immunity. After antigen stimulation of the T-cell receptor in the preence of specific cytokines, naïve Th cells start dividing rapidly to reach a differentiated state, with the best understood being $Th1$, $Th2$, $Th17$ and $pTregs$ [4]. So far, several master regulators have been identified (e.g. $Gata3$ for $Th2$, $T-bet$ for $Th1$, $Rorgt$ for $Th17$ and $Foxp3$ for $pTregs$) [4] and there is considerable insight into their regulatory networks [5].

While much is known in $CD8+$ (killer) T cells [6], the expansion of $CD4+$ (helper) T cells during an infection is less well understood at the cellular and molecular levels. How does the coupling between differentiation and the cell cycle occur in $CD4+$ T cells? Are the two processes indipendent and orthogonal, as suggested by Duffy and Hodgkin [7], or linked through molecules and hence intertwined [8]? Does differentiation occur in a gradual manner as suggested by many studies, including a recent single-cell analysis of lungh epithelial development [9], or in a cooperative switch-like manner?

Here, we use a new approach to tackle these questions, which is to extract biologically intermediate states of differentiation from a single chronological time point. By sorting out separate cell populations from a single cell culture of asynchronized, dividing cells, we aimed to reduce the biological variability in cytokine exposure, confluence, etc. With this approach we minimize the biological noise in our data and focus entirely on the processes of cell division and differentiation.

We used in-depth transcriptome profiling coupled with bioinformatics data analysis to identify three major cell states during Th2 differentiation. By counting cells in each cell generation using flow cytometry, we modelled the rates of death, division

and differentiation using a discrete time Markov branching process. This model gives information about which kind of transitions are most likely to happen at single cell level; for example, it can say if an activated cell (still not differentiated) is most likely to give rise to differentiated cells through a duplication or in a direct way. This revealed a higher cell division rate for differentiated cells compared with proliferating, activated cells. We validate those finding by DNA staining and by single-cell live imaging of $Th2$ cells. These in vitro data supported the idea of a fine-tuned relationship between cell cycle speed and differentiation status in $CD4$+T cells.

Finally, we related our findings from an ex vivo cell culture model of $Th2$ differentiation to single-cell transcriptomes of $Th1$ cells from a mouse model of malaria infection. The in vivo cytokine secreting $Th1$ cells also cycle more quickly than in vivo activated cells, showing the universal relevance of our results to primary activation of T cells. This implies that an acceleration of effector $CD4+$ T cell expansion upon differentiation is part of the immune system's machanism of pathogen clearance during primary activation. This thesis has been written on the basis of the article [1] written together with Valentina Proserpio at EMBL European Bioinformatics Institute (EBI) of Cambridge.

Chapter 2

Cell transformations: basic concepts

When we talk about cell transformation we refer to those processes in which a cell transforms from a type to another. In this thesis we focus on differentiation, which is a kind of transformation process in which a stem cell differentiate into a specific cell. However, there are many kind of cell transformations such as reprogramming and carcinogenesis. A key role in such transformations is played by the cell potency that is involved, especially, in differentiation and reprogramming. Here, we discuss briefly cell potency, differentiation, reprogramming and carcinogenesis in order to give general information to better understand the biology behind this thesis which is based on more mathematical and statistical issue rather than biological.

Cell Potency Cell potency is a general term which describes a stem cell's ability to differentiate into different cell types. The more cell types a stem cell can differentiate into, the greater its potency. Potency is also described as the gene activation potential within a cell which like a continuum begins with totipotency to designate a cell with the most differentiation potential, pluripotency, multipotency, oligopotency and finally unipotency. Potency is taken from the Latin term "potens" which means "having power".

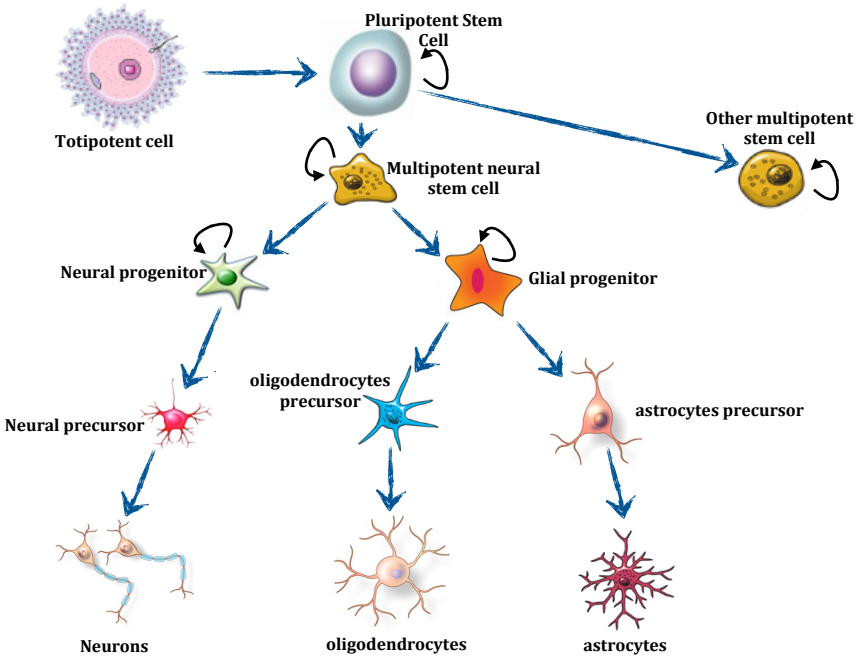


Figure 2.1 – Stem cell hierarchy and multipotency of neural stem cells: This schematic illustration depicts the generation of differentiated cell types from multipotent neural stem cells. Different specialized multipotent stem cells are produced from pluripotent embryonic stem cells. Multipotent neural stem cells produce the three lineages: neurons, oligodendrocytes, and astrocytes. In this schematic illustration, the curved arrows represent self-renewing ability.

In Figure 2.1 there is an example showing the generation of differentiated cell types from multipotent neural stem cells.

Totipotency is the ability of a single cell to divide and produce all of the differentiated cells in an organism, and example totipotent cells are spores and zygotes. In the spectrum of cell potency, totipotency represents the cell with the greatest differentiation potential. Toti comes from the Latin totus which means ”entirely.” It is possible for a fully differentiated cell to return to a state of totipotency. This conversion to totipotency is complex, not fully understood and the subject of recent research. Research in 2011 has shown that cells may differentiate not into a fully totipotent cell, but instead into a

"complex cellular variation" of totipotency [36].

In cell biology, **pluripotency** (from the Latin plurimus, meaning very many, and potens, meaning having power) refers to a stem cell that has the potential to differentiate into any of the three germ layers: endoderm (interior stomach lining, gastrointestinal tract, the lungs), mesoderm (muscle, bone, blood, urogenital), or ectoderm (epidermal tissues and nervous system).

Induced pluripotent stem cells, commonly abbreviated as iPS cells or iPSCs are a type of pluripotent stem cell artificially derived from a non-pluripotent cell, typically an adult somatic cell, by inducing a "forced" expression of certain genes and transcription factors. These transcription factors play a key role in determining the state of these cells and also highlights the fact that these somatic cells do preserve the same genetic information as early embryonic cells. Due to their great similarity to ESCs, iPSCs have been of great interest to the medical and research community. iPSCs could potentially have the same therapeutic implications and applications as ESCs but without the controversial use of embryos in the process, a topic of great bioethical debate. In fact, the induced pluripotency of somatic cells into undifferentiated iPS cells was originally hailed as the end of the controversial use of embryonic stem cells. However, iPSCs were found to be potentially tumorigenic, and, despite advances, were never approved for clinical stage research [37].

Multipotency describes progenitor cells which have the gene activation potential to differentiate into multiple, but limited cell types. For example, a multipotent blood stem cell is a hematopoietic cell — and this cell type can itself differentiate into several types of blood cell types like lymphocytes, monocytes, neutrophils, etc., but cannot differentiate into brain cells, bone cells or other non-blood cell types.

In biology, **oligopotency** is the ability of progenitor cells to differentiate into a few cell types. It is a degree of potency. Examples of oligopotent stem cells are the lymphoid

or myeloid stem cells. A lymphoid cell specifically, can give rise to various blood cells such as B and T cells, however, not to a different blood cell type like a red blood cell. Examples of progenitor cells are vascular stem cells that have the capacity to become both endothelial or smooth muscle cells.

In cell biology, a **unipotent** cell is the concept that one stem cell has the capacity to differentiate into only one cell type. It is currently unclear if true unipotent stem cells exist. Hepatoblasts, which differentiate in hepatocytes (which constitutes most of the liver) and cholangiocytes (epithelial cells of the bile duct), are bipotent. A close synonym for unipotent cell is precursor cell.

Differentiation In developmental biology, cellular differentiation is the process by which a less specialized cell becomes a more specialized cell type. Differentiation occurs numerous times during the development of a multicellular organism as the organism changes from a simple zygote to a complex system of tissues and cell types. Differentiation is a common process in adults as well: adult stem cells divide and create fully differentiated daughter cells during tissue repair and during normal cell turnover. Differentiation dramatically changes a cell's size, shape, membrane potential, metabolic activity, and responsiveness to signals. These changes are largely due to highly controlled modifications in gene expression. With a few exceptions, cellular differentiation almost never involves a change in the DNA sequence itself. Thus, different cells can have very different physical characteristics despite having the same genome.

Reprogramming Cellular reprogramming describes the process where a fully differentiated, specialized cell type is induced to transform into a different cell type that they would not otherwise become under normal physiological conditions. Cellular reprogramming has been achieved using a variety of methods, including somatic cell nuclear

transfer, cell-cell fusion and, most recently, through the introduction of four transcription factors. Most scientists have focused on reprogramming somatic cells into pluripotent stem cells, but recently some researchers have begun to focus on reprogramming somatic cells into multipotent stem cells, which have a more restricted developmental potential and are closer to the cell population the researcher ultimately wants to engineer.

Carcinogenesis Carcinogenesis or oncogenesis or tumorigenesis is literally the creation of cancer. It is a process by which normal cells are transformed into cancer cells. It is characterized by a progression of changes at the cellular, genetic and epigenetic level that ultimately reprogram a cell to undergo uncontrolled cell division, thus forming a malignant mass.

Naïve T cell

A naïve T cell, also called T lymphocyte, type of leukocyte (white blood cell) that is an essential part of the immune system. T cells are one of two primary types of lymphocytes - B cells being the second type - that determine the specificity of immune response to antigens (foreign substances) in the body. T cells originate in the bone marrow and mature in the thymus. In the thymus, T cells multiply and differentiate into helper, regulatory, or cytotoxic T cells or become memory T cells. They are then sent to peripheral tissues or circulate in the blood or lymphatic system.

T helper cells (Th cells) assist other white blood cells in immunologic processes, including maturation of B cells into plasma cells and memory B cells, and activation of cytotoxic T cells and macrophages. These cells are also known as CD4+ T cells (cluster of differentiation 4) because they express the CD4 glycoprotein on their surfaces. Helper T cells become activated when they are presented with peptide antigens by MHC class

II molecules, which are expressed on the surface of antigen-presenting cells (APCs). Once activated, they divide rapidly and secrete small proteins called cytokines that regulate or assist in the active immune response. In particular, IL-13 is cytokine secreted by many cell types, but especially T helper type 2 (Th2) cells. IL-13 has effects on immune cells that are similar to those of the closely related cytokine IL-4. These cells can differentiate into one of several subtypes, including Th1, Th2, Th3, Th17, Th9, or TFh, which secrete different cytokines to facilitate different types of immune responses.

Cytotoxic T cells (TC cells, CTLs, T-killer cells, killer T cells) destroy virus-infected cells and tumor cells, and are also implicated in transplant rejection. These cells are also known as CD8+ T cells since they express the CD8 glycoprotein at their surfaces. These cells recognize their targets by binding to antigen associated with MHC class I molecules, which are present on the surface of all nucleated cells. Through IL-10, adenosine, and other molecules secreted by regulatory T cells, the CD8+ cells can be inactivated to an anergic state, which prevents autoimmune diseases.

Regulatory T cells (suppressor T cells) are crucial for the maintenance of immunological tolerance. Their major role is to shut down T cell-mediated immunity toward the end of an immune reaction and to suppress autoreactive T cells that escaped the process of negative selection in the thymus. Suppressor T cells along with Helper T cells can collectively be called Regulatory T cells due to their regulatory functions.

Memory T cells are a subset of infection- as well as potentially cancer-fighting T cellsthat have previously encountered and responded to their cognate antigen; thus, the term antigen-experienced T cell is often applied. Such T cells can recognize foreign invaders, such as bacteria or viruses, as well as cancer cells. Memory T cells have become "experienced" by having encountered antigen during a prior infection, encounter with cancer, or previous vaccination. At a second encounter with the invader, memory T cells can reproduce to mount a faster and stronger immune response than the first time

the immune system responded to the invader.

RNA sequencing technology

RNA-Seq (RNA sequencing) technology allows you to discover and profile the transcriptome in any organism. RNA-seq (Wang 2009) is rapidly replacing gene expression microarrays in many labs. RNA-seq lets you quantify, discover and profile RNAs. For this technique, mRNA (and other RNAs) are first converted to cDNA. The cDNA is then used as the input for a next-generation sequencing library preparation. Specifically, RNA-Seq facilitates the ability to look at alternative gene spliced transcripts, post-transcriptional modifications, gene fusion, mutations/SNPs and changes in gene expression over time, or differences in gene expression in different groups or treatments. In addition to mRNA transcripts, RNA-Seq can look at different populations of RNA to include total RNA, small RNA, such as miRNA, tRNA, and ribosomal profiling. Single cell sequencing examines the sequence information from individual cells with optimized next generation sequencing (NGS) technologies, providing a higher resolution of cellular differences and a better understanding of the function of an individual cell in the context of its microenvironment.

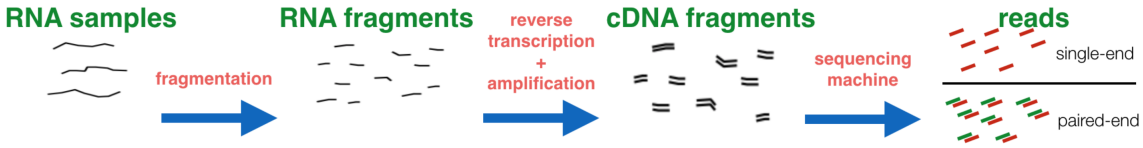


Figure 2.2 – Principal steps for RNA sequencing technology.

Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation (Figure 2.2). Sequencing adaptors are subsequently added to each cDNA fragment and a short sequence is obtained from

each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome and used to generate a base-resolution expression profile for each gene. Three principal units of measure are used for gene expression in RNA-seq experiments: RPKM (Reads Per Kilobase Million), FPKM (Fragments Per Kilobase Million) and TMP (Transcripts Per Kilobase Million).

FACS

In biotechnology, flow cytometry is a laser- or impedance-based, biophysical technology employed in cell counting, cell sorting, biomarker detection and protein engineering, by suspending cells in a stream of fluid and passing them through an electronic detection apparatus. Fluorescence-activated cell sorting (FACS) is a specialized type of flow cytometry. It provides a method for sorting a heterogeneous mixture of biological cells into two or more containers, one cell at a time, based upon the specific light scattering and fluorescent characteristics of each cell. It is a useful scientific instrument as it provides fast, objective and quantitative recording of fluorescent signals from individual cells as well as physical separation of cells of particular interest. In our work staining precursor naïve cells with CellTrace Violet dye allowed us to discriminate, count and profile cells that have undergone different numbers of cell divisions and primary cells derived from Il13-eGFP homozygous reporter mice allowed us to identify differentiated Th2 cells.

Chapter 3

Th2 differentiation: our first results

3.1 Th2 differentiation resolved by cell generation.

In Vivo After antigen stimulation of the T-cell receptor [10], naïve $CD4+$ T cells start dividing quickly and some cells initiate expression of cytokines, which is the hallmark of differentiation effector cells. To probe this process in vivo, we isolated and sequenced $CD3+/CD4+/CD62L-$ single cells from spleen and both mediastinal and mesenteric lymph nodes of *Nippostrongylus brasilensis* (Nb)-infected mice 5 days post-infection (Figure 3.1). In order to remove cells with a poor quality library a quality analysis control was performed and we retained data from 78 cells. In order to separate the fast cycling from the slow cycling ones, we clustered them according to the expression of cell cycle genes (Figure 3.2). We ranked the cells according to the expression of aggregate G2/M genes as measure of "cell cycle scores", thus reflecting the speed of the cell cycle [1]. We observed that the cells expressing higher amounts of G2/M genes were also significantly enriched in interleukin4 (IL4) (p-value = 0.008, Fisher's exact test). In order to verify that those G2/M high cells were proliferating faster and were enriched in IL4

expression, we looked at the expression level of proliferating marker genes [11, 12]. The cells enriched in those genes also expressed significantly higher amounts of IL4 (p-value = 0.001, Fisher's exact test), confirming that cytokine-producing cells are cycling faster. Based on this observation, we proceeded to study the link between cell cycle speed and differentiation in Th2 cells in more details in an in-vitro cell culture system.

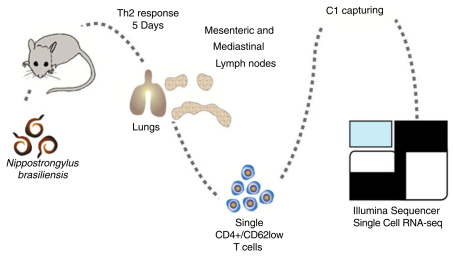


Figure 3.1 – Overview of the experiment. CD3+/CD4+/CD62L- T helper cells were isolated from lungs, mesenteric and mediastinal lymph nodes of Nb-infected mice on day 5 post-infection. After cell capturing and cDNA generation with the C1 system, samples were sequenced with an Illumina Hi-seq Sequencer.

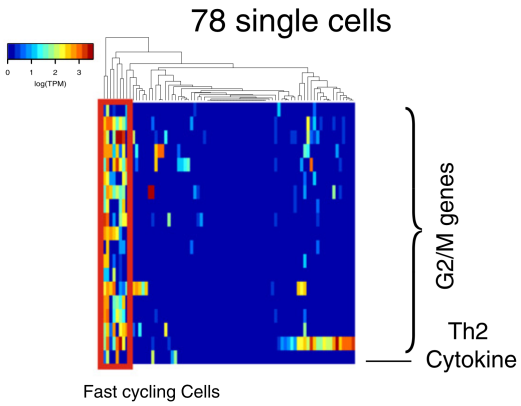


Figure 3.2 – Seventy-eight single cells were clustered according to the expression of G2/M genes (logTPM) as a measure of cell cycle speed. TPM transcripts per millions. Three cells expressing IL4 clustered within the group of cells expressing high levels of G2/M genes (in the red box) (p value = 0.008, Fisher's exact test).

3.1. Th2 differentiation resolved by cell generation.

In Vitro - Flow Citometry As a marker of differentiation we employed IL13 instead of IL4, as its expression is less susceptible to changes in IL4 concetration in the medium. Staining precursor naïve cells with CellTrace Violet dye allowed us to discriminate cells that have undergone different numbers of cell divisions (Figure 3.3). Primary cells derived from IL13-eGFP homozygous reporter mice allowed us to identify differentiated Th2 cells [11]. Using this system, we observe, consistent with previosly published data for other cytokines [8, 12], that the proportion of differentiated cells (with flourescent IL13+ reporter expression) increases linearly in each consecutive generation (Figure 3.4). In prevous reports, cytokine producing cells have been detected only from the third generation onwards [8], while we detected these cells in earlier generation already. This is probably due to our use of a green flourescent protein (GFP) reporter for the endogenous cytokine, instead of the traditional staining with flourescent antibody.

To dissect whether sudden or gradual changes in cell state occur during Th2 differentiation, we performed a transcriptome-wide characterization of cells that had undergone different number of mitotic divisions after 3.5 days of activation. From this single time point, we sorted and carried out mRNA-sequencing (mRNA-seq) of three non overlapping population of cells that were not expressing GFP and had, respectively, not divided (generation 0 negative G0N), divided twice (G2N) or divided four times (G4N). We also profiled a fourth population of cells that had divided four times and was positive for GFP (G4P; Figure 3.5).

Hierarchical clustering of these datasets indicates that there are three major states. A G0N cluster is clearly separate from the other population, indicating a major difference between cells taht have not yet undergone mitosis and other cells that have entered cell cycle. The G4P cluster is more distant from the other two dividing populations, indicating that te expression of one single marker of differentiation (IL13) occurs concomitantly with global changes to expression profile of growing lymphocytes. In contrast, the

3.1. Th2 differentiation resolved by cell generation.

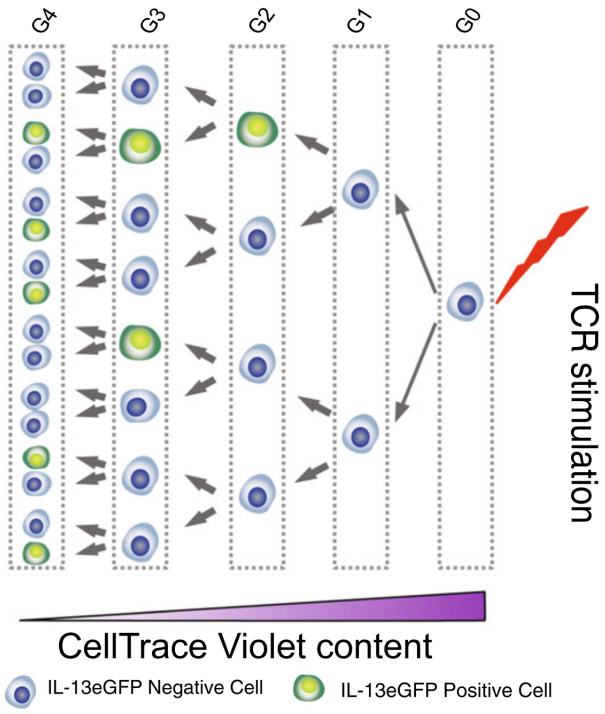


Figure 3.3 – Schematic representation of the division/differentiation process from a naïve cell to fully differentiated Th2 cells. The CellTrace Violet content is roughly equally distributed between daughter cells after each mitotic division. Cells expressing the Il13-eGFP Th2 differentiation marker are shown in green. TCR T-cell receptor.

3.1. Th2 differentiation resolved by cell generation.

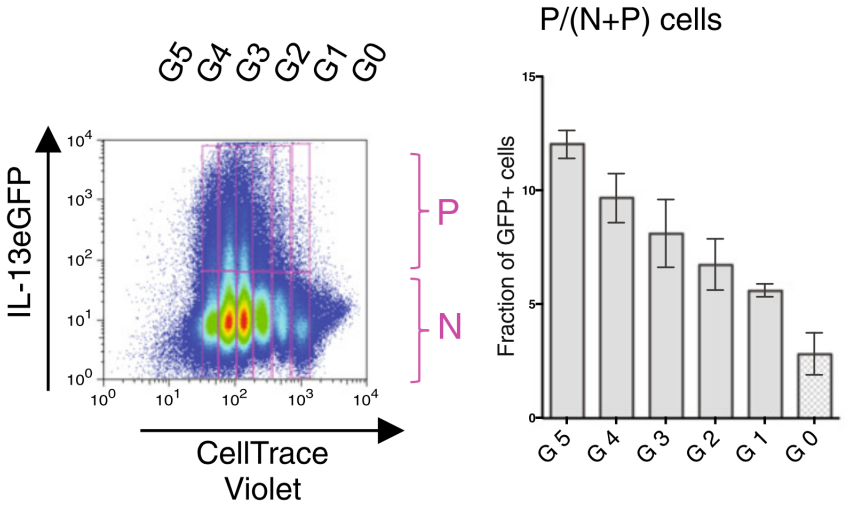


Figure 3.4 – Flow cytometry plot of CellTrace Violet versus Il13-eGFP differentiated Th2 cells at day 3.5. Consecutive generations (from G0 to G5) are visualized as pink gates. The upper gates are IL13-positive cells (P), and the lower gates contain IL13-negative cells (N). Ratio of GFP-P cells to the total number of cells per generation (average and standard deviation of three biological replicates).

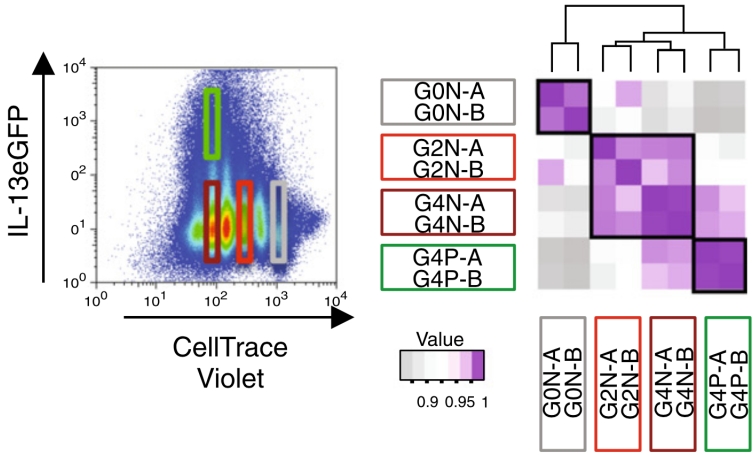


Figure 3.5 – Cells in the gates highlighted were sorted by FACS and profiled by mRNA-sequencing. Hierarchical clustering of the distance matrices between RNA expression profiles

G2N and G4N cluster together, sharing similar expression profiles. These results are supported and validated by Quantitative PCR (qPCR) and flow cytometry of individual genes and proteins [1].

3.2 RNA-seq analysis indicates three major states

Deep transcriptomic analysis reveals three discrete cell states during Th2 differentiation. While some group of genes increase or decrease apparently continuosly across the four RNA-seq data sets from G0N to G2N, G4N and G4P, there are also groups of genes that have non-monotonic patterns of expression (Figure 3.6). Therefore, it is unclear whther the differentiation is occurring through a single gradual progression or via discrete intermediate states.

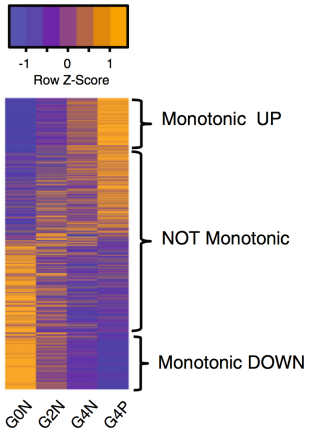


Figure 3.6 – Heatmap of all ˜14,000 protein-coding genes (rows) per generation (columns). At the top and bottom, genes with a monotonic increase/decrease are shown. In the middle, genes are ranked according to distance from G4P and G0N.

We analyzed differentially expressed genes (DEGs) between subgroups and found roughly 1500 DEGs between G0N and G2N and between G4N and G4P but only 170 between G2N and G4N (Figure 3.7). Gene Ontology (GO) enrichment analysis (Figure

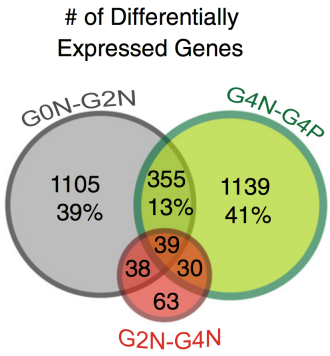


Figure 3.7 – Number and percentage of differentially expressed genes between samples.

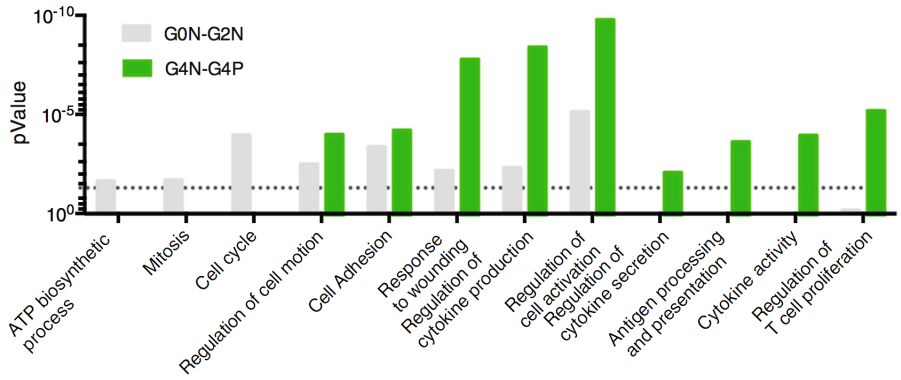


Figure 3.8 – Gene Ontology enrichment analysis was performed on differentially expressed genes between G0N and G2N (gray bars) and G4N and G4P (green bars). The threshold p-value of 0.05 is shown as a dotted line.

3.8) showed that G0N-G2N DEGs are enriched in "ATP biosynthetic process", "Mitosys" and "Cell Cycle". The majority of these genes (70 to 85%) are up-regulated in G2N versus G0N, providing further confirmation that G0N cells are not actively proliferating. At the same time, the high expression of the activation marker Cd69 [13] and the levels of L-selectin (Cd62l, Sell) and Cd44 [14] in G0N cells and the increase in size of some of these cells indicate that they have been partially activated so are no longer naïve cells.

Our GO analysis of DEGs between G4N and G4P indicated that the terms "Regulation of cytokine secretion", "Cytokine activity" and "Regulation of T cell proliferation"

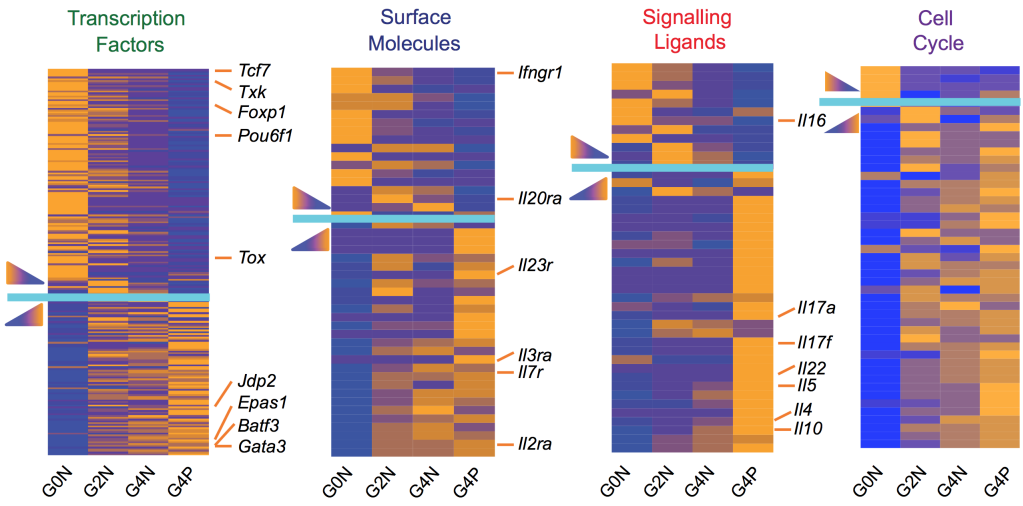


Figure 3.9 – expression changes of genes belonging to three important categories: signaling ligands (SLs), surface molecules (SM), transcription factor (TF) and cell cycle genes.

represent the main categories of genes that are specifically differentially expressed together with IL13 transcription. This means that the expression of IL13 coincides with the expression of other genes important for Th2 function (IL3, IL4, IL5, IL6, [1]).

We also analysed the expression changes of genes belonging to four important categories: signaling ligands (SLs), surface molecules (SM), transcription factor (TF) and cell cycle genes (Figure 3.9).

Among the upregulated TFs we found some genes for which a role in Th2 differentiation has already been demonstrated (Gata3, Batf3, Epas1, [15]) and some where their function still remains to be to elucidated. In the SM group we observed induction of IL2ra and IL7r, which are known to be involved in lymphocyte differentiation [16]. Moreover, the strong downregulation of Ifngr1 observed in conjunction with cell activation is consistent with previous reports [17]. Overall, the vast majority ($\sim 75\%$) of the differentially expressed SLs are upregulated from one generation to the next.

Within the upregualted genes, we calculated the average Z-score across conditions for

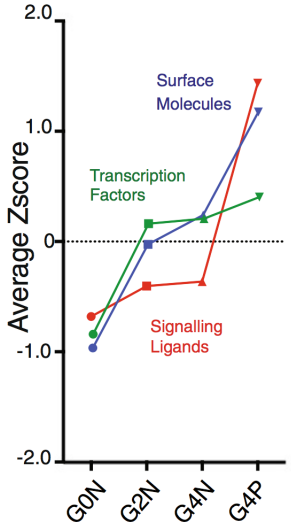


Figure 3.10 – Average Z-scores for upregulated genes belonging to different functional categories calculated from heatmaps in Figure 3.9

each of the four populations (Figure 3.10). TFs and SMs are promptly upregulated soon after entering the cell cycle (G2N) and no further increase is detected after further cell division. SMs show a second prominent increase in their expression in cells when IL13 is also expressed. No important effect of entering the cell cycle is visible on cytokine expression levels and only the production of IL13 correlates with the expression of all SLs. Importantly, all the Th2-specific cytokines follow the same pattern of expression, continuing to be lowly expressed after entering the cell cycle and only undergoing a sharp boost from G4N to G4P.

Finally, the expression level of cell cycle genes is strongly upregulated from G0N to G2N, as expected by definition for these two subpopulations. More interesting, we observed a second sharp increase in the expression of cell cycle genes in G4P cells, emphasizing the concomitant upregulation in cell cycle and differentiation genes, as we already observed from ex vivo Th2 single-cell RNA-seq. Finally, the expression of

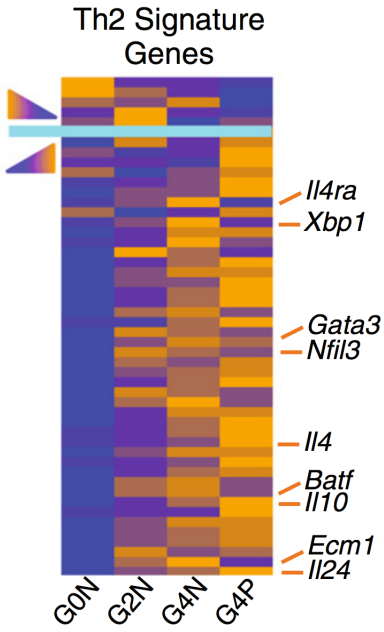


Figure 3.11 – Heatmap of Th2 signature gene expression across generations.

Th2 singnature increases from G0N to the consecutive negative generations and further increases from G4N to G4P (Figure 3.11). This group of genes includes most of the genes with a role in Th2 specification (Ecm1, IL24, Batf, IL10, Nfil3, Gata3, IL4ra) [18, 19, 20].

In summary, these data suggest that the G0N to G2N/G4N transition represents the exit from cell cycle arrest and entry into the proliferative cell state. The difference between G4N and G4P must result from a second major switch, which represent differentiation to the Th2 effector state with expression of the characteristic cytokine, including IL13. Together with expression of these cytokines, there is a parallel further increase in the expression of cell cycle genes.

The combination of the above results leads us to characterize Th2 cell differentiation as consisting of three major transcriptionally distinct states, which we name state A

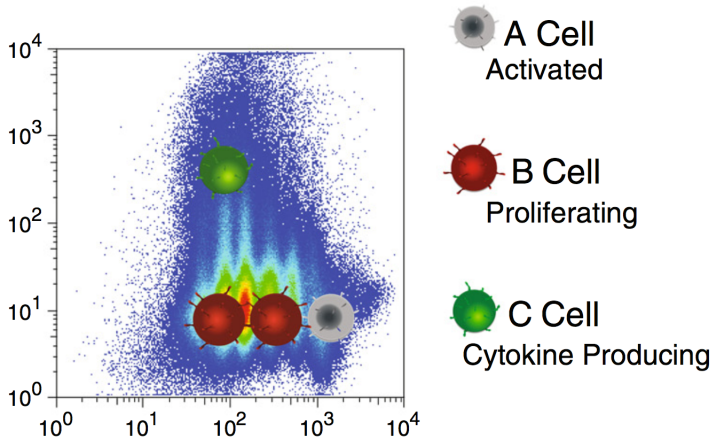


Figure 3.12 – A three-state differentiation model in which G0N cells are named A cells ("Activated" cells), G2N and G4N cells are named B cells ("Proliferating" cells) and G4P cells are named C cells ("Citokine-producing" cells)

(activated cells that correspond to G0N), B (proliferating cells that correspond to both G2N and G4P) and C (cytokine expressing cells that correspond to G4P) (Figure 3.12).

Validation of the three state model at single-cell resolution Our description of three cell states during Th2 differentiation comes from population mRNA-seq data. Therefore, we aimed to verify our hypothesis at single-cell resolution by performing high-throughput single-cell qPCR analysis with dozens of genes in parallel in 46 cells from each population. We obtained a good overall correlation between the RNA-seq data and the average of the single-cell qPCR results (Figure 3.13 and 3.14).

Based on these data, we aimed to assign each cell to one of the three specific states we identified. We employed principal component analysis (PCA) and, to quantify the separation, a linear support vector classifier (SVC) was trained using "one-hot" labels (e.g. is it G4N or not) for each of the conditions for the first two component values (Figure 3.15). What we observed is that the SVC vector is able to distinguish G0N and G4P from other cells with good accuracies (score are 0.83 and 0.72, respectively).

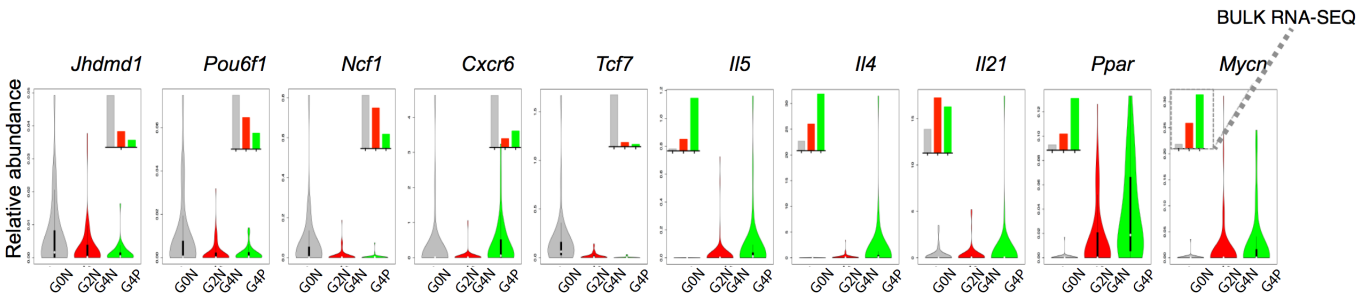


Figure 3.13 – Additional analysis confirmed the existence of three discrete states at the single cell level. Violin plots represent the distribution of selected genes in single cells. The insets show bulk RNA-seq results.

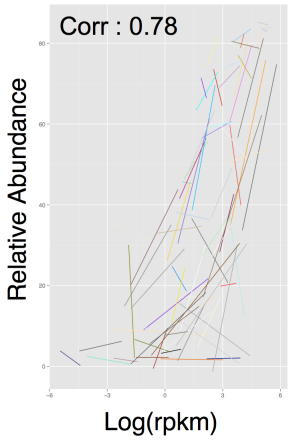


Figure 3.14 – Additional analysis confirmed the existence of three discrete states at the single cell level. Correlation between bulk data (x-axis) and single-cell qPCR (100-ct mean; y-axis). Each gene is represented by a line linking its values in G0N and G4P

Conversely, it fails to distinguish G2N and G4N states from the other cells and also the mixture of G2N and G4N from the rest of the cells. This analysis support the existence of three states represented by G0N, G4P and a mixture of G2N and G4N.

In order to probe the transcriptional regulation of these states on a single-cell basis, we focused on 11 highly expressed transcription factors (Epas1, Myb, Mycn, Jhdm1d, Pou6f1, Pparh, Tcf7, Txk, Zc3h12c, Zfp36, Hlx) and discretized them into "on" or "off" states in each cell. This yelded 124 unique binary states and 117 of these can be connected by single-gene changes to yeld a state graph as in Moingnard et al. [21] (Figure 3.16)

From this analysis we could observe that cells of the same state share similar TF organization as they cluster close to each other, underlining how TFs could act as master regulators of cell fate. We could also verify that the differentiation from A (in grey, Figure 3.16) to C (in green) requires the transition through at least on B cell (in red), further confirming the intermediate nature of this cell state.

Taken together, these data confirm the concept of a three-state model of differentiation during Th2 primary activation. We also quantified the homogeneity of cells belonging to each of the three states as the average of Spearman cell-to-cell correlation ($p - value < 0.05$) within each state (Figure 3.17).

We observed an increase in the correlation across cells belonging to state C with respect to states A and B. This suggest that A and B cells are flexible and heterogenous after primary activation, both before and after enter the cell cycle. In contrast, the cell that are more differentiated are more similar to each other, representing a more homogeneous population. These results, in agreement with data from Arsenio et al. on CD8+ T cells [22], support the concept of a commitment toward a more specific state in concert with the expression of the specific cytokines.

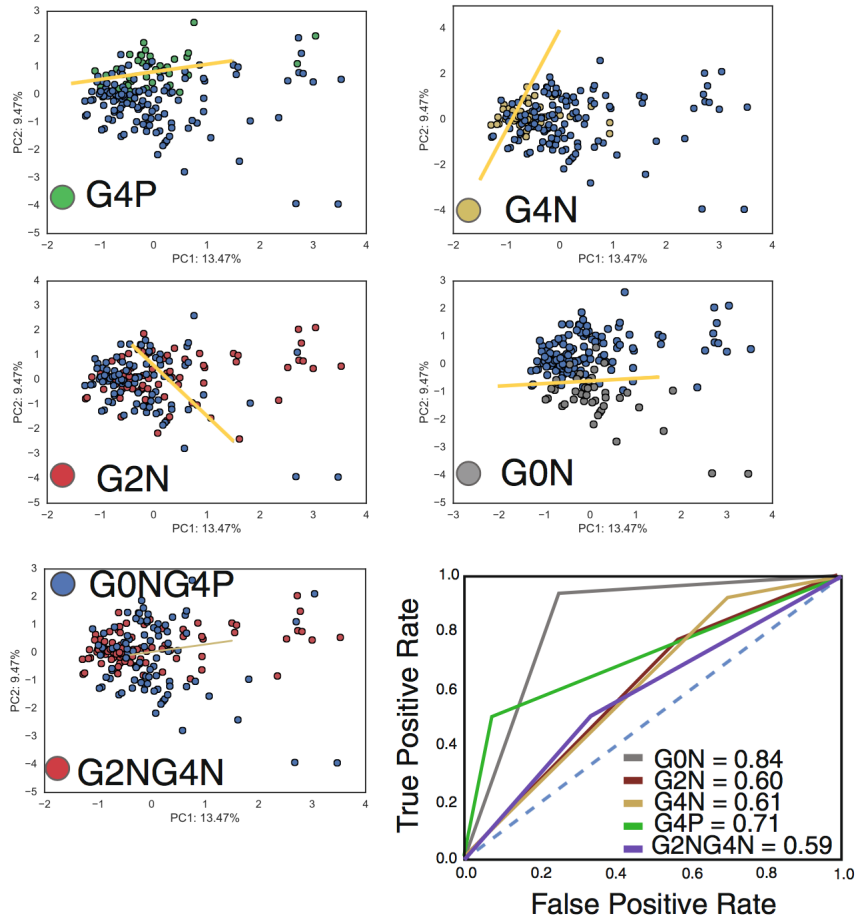


Figure 3.15 – Linear principal component analysis (PCA) with a linear support vector classifier (SVC the yellow line) trained with "one-hot" labels (e.g. is it G4N or not) for each of the conditions and the first two principal component values were used to separate each of the generation and G2N/G4N cells from the other cells (in blue).

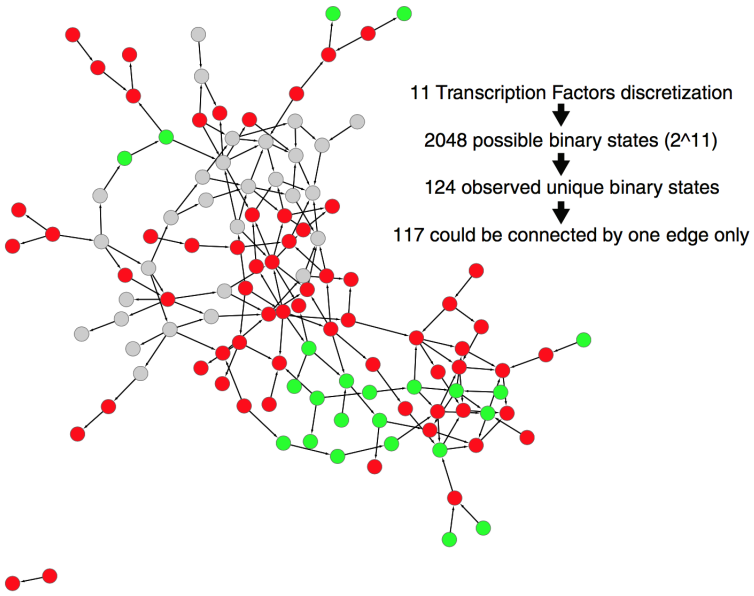


Figure 3.16 – State graph of 117 connected binary cell states for 11 transcription factors, constructed using the SCNS toolkit. Each edge represents the change in expression of a single gene. Grey circles are G0N cells, red circles are G2N/G4N cells and green circles represent G4P cells

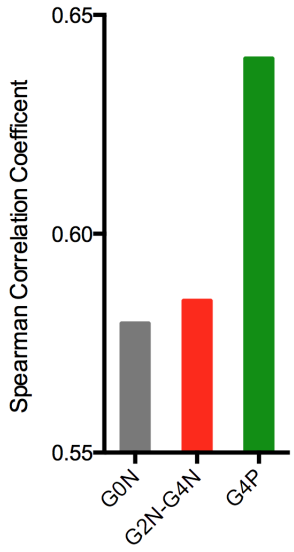


Figure 3.17 – Average of Spearman correlation values between any two different cells with p value <0.05 is reported for the indicated population.

Chapter 4

ABC Model

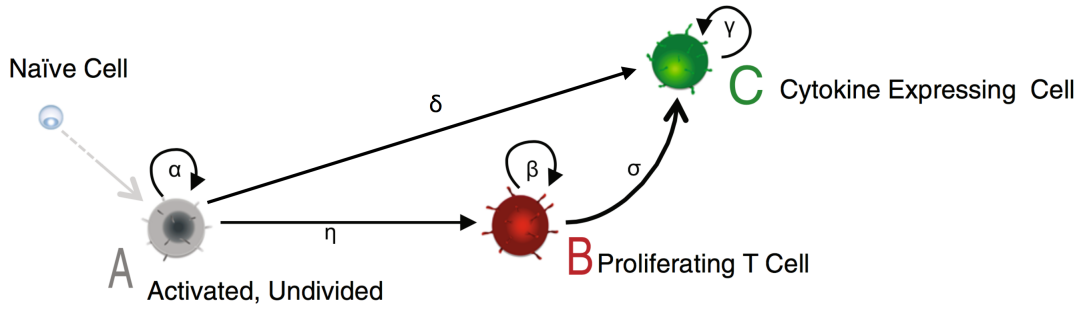


Figure 4.1 – The model predicts T-cell behaviour at the cellular level. Overview of the differentiation process that converts a naïve cell into a fully differentiated Th2 cell. Each naïve cell goes through three different states: state A (Undivided), state B (Proliferating) and state C (Cytokine expressing).

We model the cell dynamics as a stochastic Markov process which includes three states, named state A, B and C (Figure 4.1 and Figure 4.2). For sake of mathematical simplicity, we considered a discrete time branching process, whose time step unit is named Δt. We suppose that, from a time step to the next, an A cell can die (rate A_d), stay the same (rate A_i), divide giving rise to two type B cells (rate A_s), or produce an asymmetric

division in an B and a C cell (rate A_a). A B cell can die (rate B_d), stay the same (rate B_i), duplicate (rate B_s), give rise to a type C cell (rate B_t), or divide asymmetrically in an B and a C cell (rate B_a). Similarly, C cells can die, stay the same, divide (rates C_d, C_i, C_s). A number of other state changes could be considered. For simplicity we only include the above ones which are motivated on biological grounds. In our FACS data at $day3.5$ the last populated subgroup is $G6N$, which points out that six division steps must have occurred since $day0$. From that we derive $\Delta t = 14h$, a value compatible with the known duration of the cell cycle.

The ABC model can be analytically fully solved. The recursive relations of the average number of A, B and C cells, $A(t)$, $B(t)$ and $C(t)$ can be derived from the Master Equation of the Markov process:

$$\begin{pmatrix} A(t+1) \\ B(t+1) \\ C(t+1) \end{pmatrix} = \begin{pmatrix} \alpha & 0 & 0 \\ \eta & \beta & 0 \\ \delta & \sigma & \gamma \end{pmatrix} \begin{pmatrix} A(t) \\ B(t) \\ C(t) \end{pmatrix} \tag{4.1}$$

where

$$\begin{aligned} \alpha &= A_i \\ \beta &= 2B_s + B_i + B_a \\ \gamma &= 2C_s + C_i \\ \eta &= 2A_s + A_a \\ \sigma &= B_t + B_a \\ \delta &= A_a \end{aligned} \tag{4.2}$$

The eigenvalues α, β, γ of the above matrix give the longer time scale growth rates of the A, B and C populations respectively. The number composition of the different subpopulations, $G0N$, $G1N$, ..., $G6N$ can be analogously derived at any time, t, as a

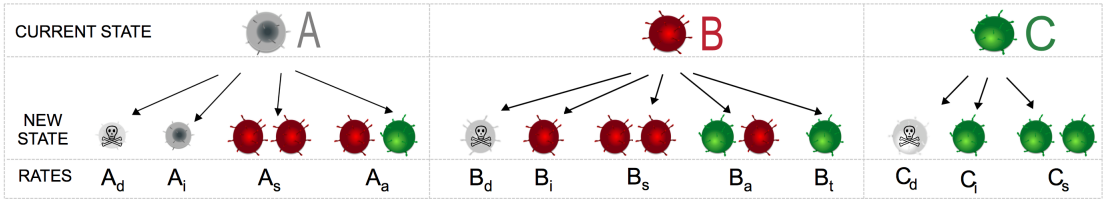


Figure 4.2 – The model with examples of state-specific cell transitions and their corresponding probabilities. Transition probabilities are labelled as follows: d death, i stay identical, s symmetric division, a asymmetric division t transdifferentiation.

function of the microscopic parameters of the model.

EAT model equation Populations at time-step N can be achieved by the following equations:

$$\begin{pmatrix} A_0^N \\ B_1^N \\ C_1^N \\ B_2^N \\ C_2^N \\ B_3^N \\ C_3^N \\ B_4^N \\ C_4^N \\ B_5^N \\ C_5^N \\ B_6^N \\ C_6^N \end{pmatrix} = \begin{pmatrix} A_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2A_s + A_a & A_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A_a & B_t & T_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2A_s + B_a & 0 & A_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & B_a & 2T_s & B_t & T_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2A_s + B_a & 0 & A_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & B_a & 2T_s & B_t & T_i & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2A_s + B_a & 0 & A_i & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & B_a & 2T_s & B_t & T_i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2A_s + B_a & 0 & A_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_a & 2T_s & B_t & T_i & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2A_s + B_a & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_a & 2T_s & 0 & 0 \end{pmatrix}^N \begin{pmatrix} A_0^0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{4.3}$$

Dynamic solution The Network which links the three states A, B, C can be summarized in Figure 4.1 and Figure 4.2

The recursive equations for the average number of A, B and C cells at time step t, $A(t)$, $B(t)$ and $C(t)$ are:

$$A(t+1) = (A_1)A(t) = \alpha A(t);$$

$$B(t+1) = (2A_s + A_a)\,A(t) + (B_i + 2B_s + B_a)\,B(t) = \eta A(t) + \beta B(t);$$

$C(t+1) = (A_a)A(t) + (B_t + B_a)\,B(t) + (C_i + 2C_s)\,C(t) = \delta A(t) + \sigma B(t) + \gamma C(t);$

That is

$$\begin{pmatrix} A(t) \\ B(t) \\ C(t) \end{pmatrix} = \begin{pmatrix} \alpha & 0 & 0 \\ \eta & \beta & 0 \\ \delta & \sigma & \gamma \end{pmatrix} \begin{pmatrix} A(t-1) \\ B(t-1) \\ C(t-1) \end{pmatrix} = \ldots = \begin{pmatrix} \alpha & 0 & 0 \\ \eta & \beta & 0 \\ \delta & \sigma & \gamma \end{pmatrix}^N \begin{pmatrix} A(0) \\ B(0) \\ C(0) \end{pmatrix} \tag{4.4}$$

Its eigenvalues are α, β, γ with eigenvectors:

$$|\alpha> = \begin{pmatrix} 1 \\ \frac{\eta}{\alpha-\beta} \\ \frac{\alpha\delta-\beta\delta+\eta\sigma}{(\alpha-\beta)(\alpha-\gamma)} \end{pmatrix} \tag{4.5}$$

$$|\beta> = \begin{pmatrix} 0 \\ 1 \\ \frac{\sigma}{\gamma-\beta} \end{pmatrix} \tag{4.6}$$

$$|\gamma> = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{4.7}$$

We can thus decompose the vector (A, B, C) in this way

$$\begin{pmatrix} A(0) \\ B(0) \\ C(0) \end{pmatrix} = a|\alpha> + b|\beta> + c|\gamma> \tag{4.8}$$

where

$$a = A(0) \tag{4.9}$$

$$b = B(0) - \frac{\eta}{\alpha - \beta} A(0) \tag{4.10}$$

$$c = C(0) + \frac{\sigma}{\gamma - \beta} B(0) + \frac{\beta\delta - \gamma\delta - \eta\sigma}{(\alpha - \gamma)(\gamma - \beta)} A(0) \tag{4.11}$$

thus

$$\begin{pmatrix} A(N) \\ B(N) \\ C(N) \end{pmatrix} = \alpha^N a|\alpha> + \beta^N b|\beta> + \gamma^N c|\gamma> \tag{4.12}$$

The solution is:

$A(N) = A(0)\alpha^N$

$B(N) = \beta^N B(0) + \frac{\eta\left(\alpha^N - \beta^N\right)}{\alpha - \beta} A(0)$

$C(N) = \gamma^N C(0) + \left(-\frac{\sigma\gamma^{N+1}}{(\alpha-\gamma)(\gamma-\beta)} + \frac{\alpha\sigma\gamma^N}{(\alpha-\gamma)(\gamma-\beta)} - \frac{\sigma\beta^N}{\gamma-\beta}\right) B(0) +$

$+ \left(\frac{\delta\alpha^{N+1}}{(\alpha-\beta)(\alpha-\gamma)} - \frac{\beta\delta\alpha^N}{(\alpha-\beta)(\alpha-\gamma)} + \frac{\eta\sigma\alpha^N}{(\alpha-\beta)(\alpha-\gamma)} + \frac{\eta\sigma\beta^N}{(\alpha-\beta)(\gamma-\beta)} - \frac{\delta\gamma^{N+1}}{(\alpha-\gamma)(\gamma-\beta)} + \frac{\beta\delta\gamma^N}{(\alpha-\gamma)(\gamma-\beta)} - \frac{\eta\sigma\gamma^N}{(\alpha-\gamma)(\gamma-\beta)}\right) A(0)$

Chapter 5

Results

5.1 Single cell fate: mathematical modelling of three cell states quantifies the link between acceleration of proliferation and differentiation

To further dissect our three state hypothesis and verify and quantify the existance of a difference in the proliferation rate of different cells, we investigated the cellular (as opposed to molecular) events underlying cell differentiation across such states. We exploited flow cytometry data at day 3.5 of differentiation to discriminate between different models of cell differentiation. We considered a simple schematic, mathematical model of the behaviour of individual cells and their transformation dynamics across the three states A, B and C (Figure 5.1).

5.1. Single cell fate: mathematical modelling of three cell states quantifies the link between acceleration of proliferation and differentiation

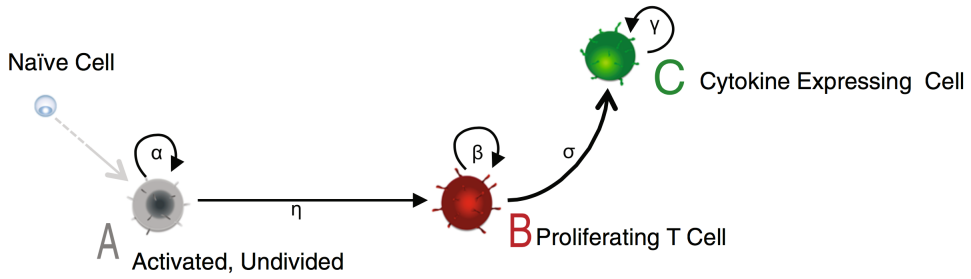
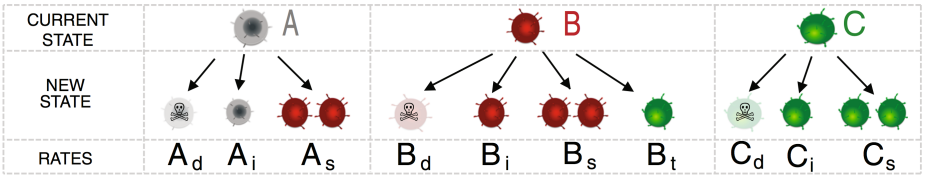


Figure 5.1 – Overview of the differentiation process that converts a naïve cell into a fully differentiated Th2 cell. Each naïve cell goes through three different states: state A (Undivided), state B (Proliferating) and state C (Cytokine expressing).

In our model, an activated naïve cell become an A cell that can then divide and give rise to B cells, wich in turn can transdifferentiate into effector C cells. We assume that each cell can stochastically divide, die or differentiate into anther state at given cell state-specific rates described by a Markov process (Figure 5.2).



A_d	A_i	A_s
0.24	0.70	0.06

B_d	B_i	B_s	B_t
0.0	0.67	0.31	0.02

C_d	C_i	C_s
0.0	0.22	0.78

Figure 5.2 – The model with examples of state-specific cell transitions and their corresponding probabilities. Transition probabilities are labelled as follows: d death, i stay identical, s symmetric division, t transdifferentiation. In the table, best fits of the model transition probabilities (expressed as probability per 14 h) from flow cytometry data at day 3.5 are reported. Data are representative of three independent mice

The death rate of an A cell in state A is named A_d; and A_i is the rate at which the A cell remains identical. Since upon activation an A cell can start dividing, we consider the transition where an A cell divides symmetrically into two type of B cells (rate A_s) .

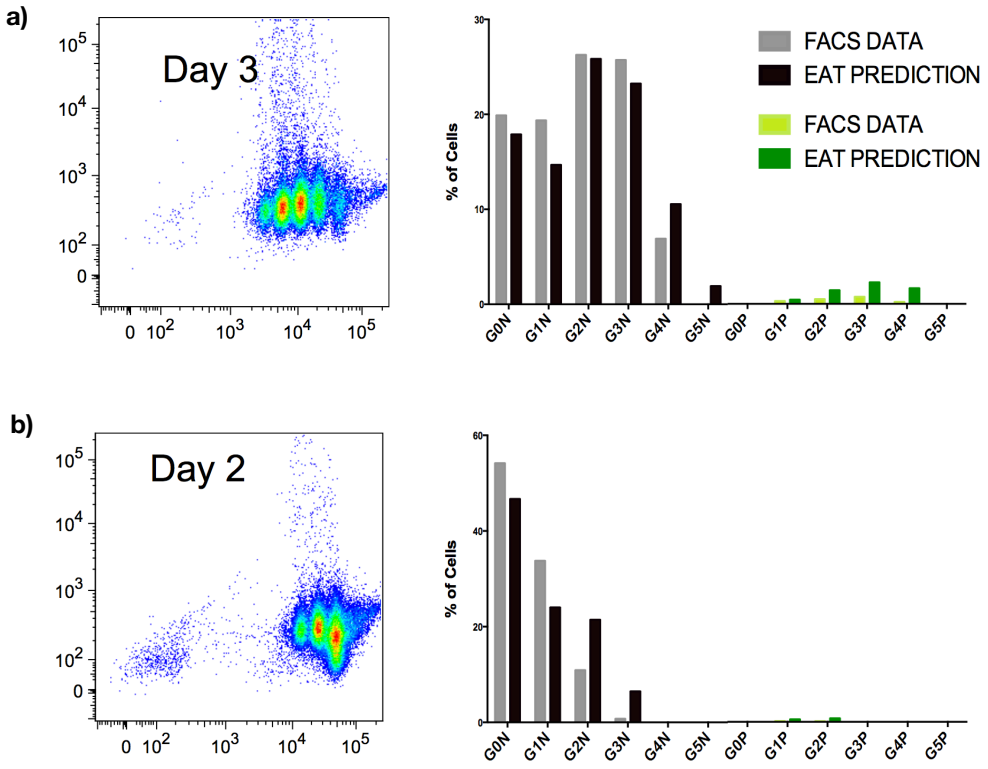


Figure 5.3 – Population dynamics of the three states over a 3day (panel a) and 2day (panel b) period as predicted by the model.

Analogously, a B cell can die (rate B_d), stay the same (rate B_i), duplicate (rate B_s) or transdifferentiate into a type C cell (rate B_t). T cells, similarly, can die, stay the same and divide (rates C_d, C_i and C_s).

The corresponding transition probability are shown in Figure 5.2 (lower panel). The model, with the same parameters, also accuratelly models two additional flourescence-activated cell sorting (FACS) datasets collected indipendently at day 2 and day 3 of differentiation (Figure 5.3).

We also calculated AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria) parameters for two, three and four state models to verify whther our three

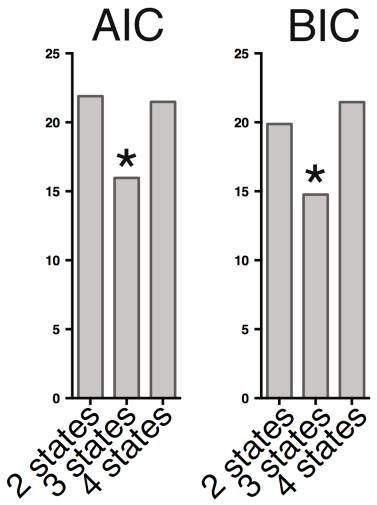


Figure 5.4 – AIC and BIC for two-, three- and four-state models (the asterisks indicate the minimum values).

state model has the higher performance. AIC and BIC analysis supports the idea that in vitro Th2 primary activation is best described by three states (Figure 5.4).

Importantly, the day 3.5 fit appears to be robust because indipendent fits of day 2 and day 3 data return fitting parameters very close to those of the day 3.5 fit (Figure 5.5).

Starting from the static picture of the system at day 3.5, our model predicts the detailed dynamics of the population composition (Figure 5.6).

By solving the master equation of the model, the average number of A, B and C cells, as well as the composition of the subpopulation, can be derived at any time (Figure 5.7)

By fitting FACS data at day 3.5, the values of the single cell transition probabilities can be determined (Figure 5.2). The model fits our FACS data well in terms of the composition of the different cell subpopulation at day 3.5 (Figure 5.7).

With the parameters returned by the day 3.5 FACS data fit, the model predicts a twofold faster proliferation of C cells with respect to B cells, as expected from gene

5.1. Single cell fate: mathematical modelling of three cell states quantifies the link between acceleration of proliferation and differentiation

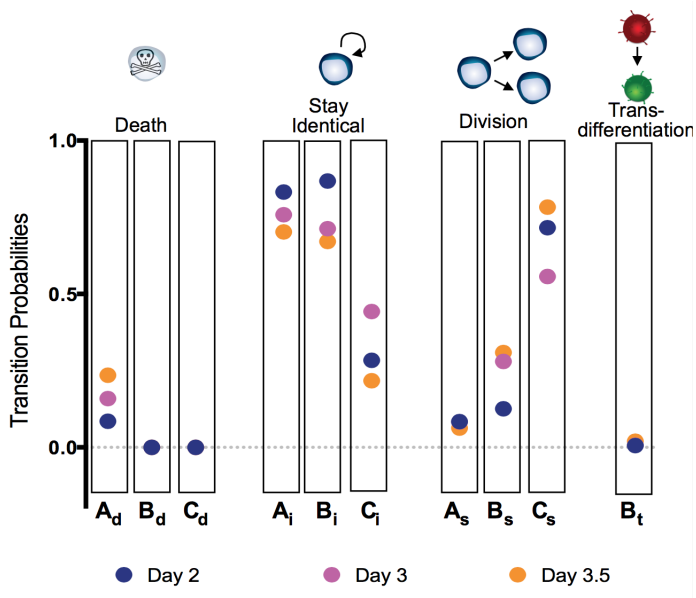


Figure 5.5 – Comparison of the parameters extracted from fits at days 2, 3 and 3.5 and comparison of the parameters extracted from flow cytometry data at days 2, 3 and 3.5.

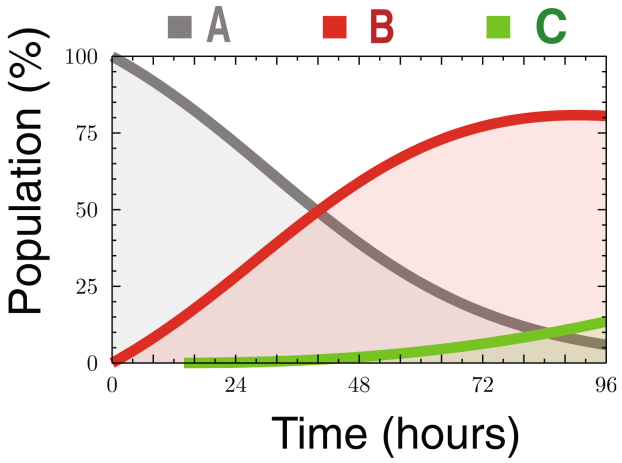


Figure 5.6 – Population dynamics of the three states over a 4-day period as predicted by the model.

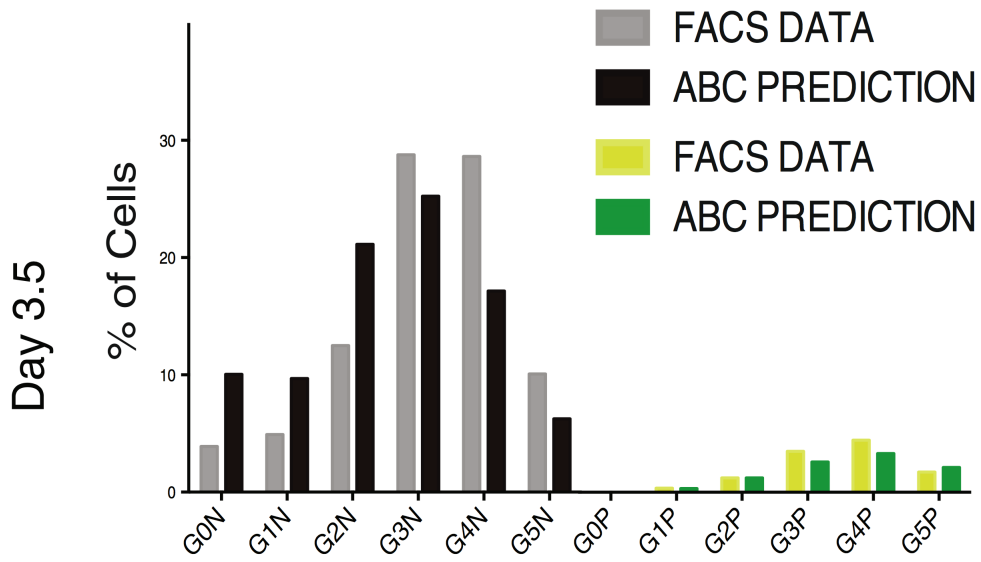


Figure 5.7 – Cell subpopulations in the flow data at day 3.5 and the model prediction with parameters extracted at day 3.5.

expression profiles of cell cycle genes mentioned above. Interestingly, the differentiation rates of A and B cells (given by A_s and B_t, respectively) are approximately one order of magnitude smaller than the growth rates of the populations of the three states α, β and γ(Figure 5.8).

It is worth noting that the predicted death rates of B and C cells are very small ($B_d \approx 0$, $C_d \approx 0$) during the first 3.5 days of differentiation. We were able to validate the difference of the death rate of A cells with respect to the other two states in indipendent experiments. Please note that apoptotic cells, measured as the sub-G1 DNA peak by Hoechst staining by flow cytometry, were only present in G0N cells, and completely absent in G1N, G2N and G3N and GFP-Positive cells (Figure 5.9).

To verify the higher division rate of differentiated cells, we also compared the cell cycle distribution of Hoechst-stained IL13-positive (C) and -negative cells (B) (Figure 5.10). Using the G2-M/G1 ratio as an indicator of the proportion of cycling cells, we

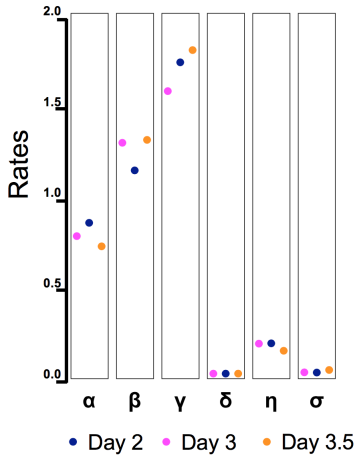


Figure 5.8 – Comparison of the parameters extracted from fits at days 2, 3 and 3.5 and comparison of the parameters extracted from flow cytometry data at days 2, 3 and 3.5.

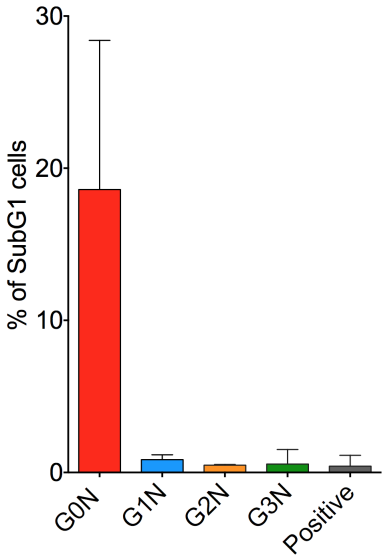


Figure 5.9 – Apoptotic cells, measured as the sub-G1 DNA peak by Hoechst staining by flow cytometry, were only present in G0N cells, and completely absent in G1N, G2N and G3N and GFP-Positive cells

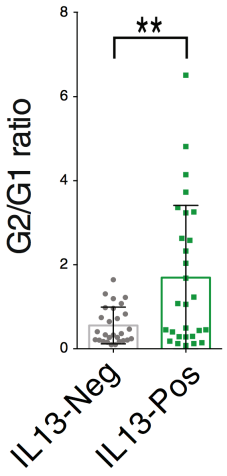


Figure 5.10 – The ratio between proportions of cells in G2 versus G1 is used as a measure of cell cycle speed when comparing both positive and negative cells within each individual generation (experiments are representative of four independent mice. Error bars indicate Standard Deviation, p value <0.01).

observed that cells expressing IL13 are cycling faster than IL13-negative cells in the same generation confirming our model prediction.

Also, from the transcriptional point of view, cell cycle genes are highly up-regualated in G4P compared with G4N cells, further suggesting an increase in cell cycle speed co-occurring with cytokine expression (Figure 5.11).

To give more quantitative estimation of the cell cycle lenght in G4P versus G4N cells, we employed an automated imaging system to image single lymphocyties over 20-40 h period. A MATLAB program, developed in house, was employed to extract data from frames and the division time for G4P and G4N were computed to be 12.5 ± 4.2 and 18.7 ± 3.5 h, respectively (Figure 5.12).

These further experiments not only confirm the acceleration of cell division that occurs concomitantly with Th2 differentiation but also precisely quantify the difference in cell cycle length of Th cells during primary activation.

5.1. Single cell fate: mathematical modelling of three cell states quantifies the link between acceleration of proliferation and differentiation

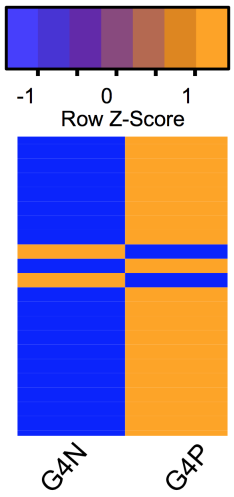


Figure 5.11 – cell cycle genes are highly up-regualated in G4P compared with G4N cells

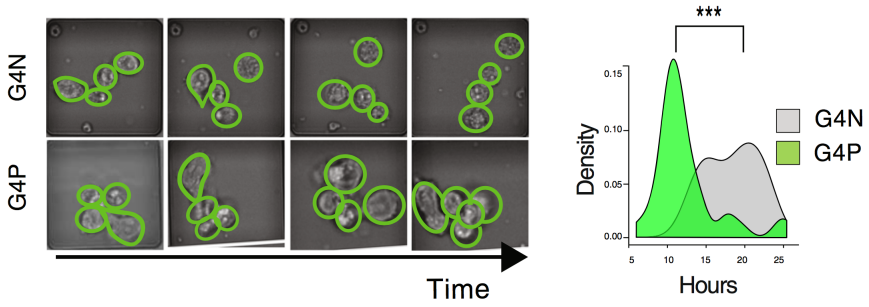


Figure 5.12 – Live imaging of G4P versus G4N cells. Representative pictures from the live imaging time course experiment of G4N (top) and G4P (bottom) cells. Distribution of time of first division for G4N (gray) and G4P (green) cells (p-value <0.001)

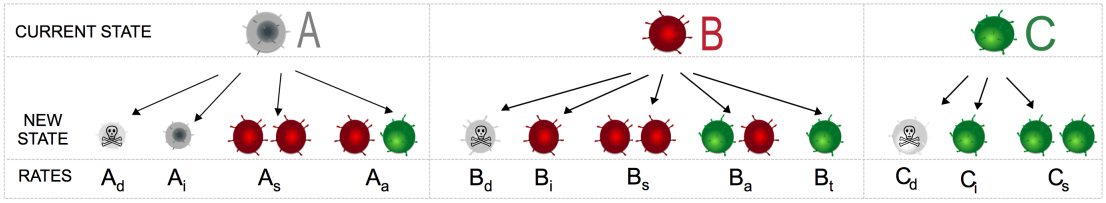


Figure 5.13 – Asymmetric division considered in the As and OA models.

5.2 Asymmetric division and robustness of the model

We also tested an extended version of the model including asymmetric divisions (named the "As" model) in which we added the possibility of A and B cells dividing asymetrically and giving rise to a B cell and a C cell (rate A_a and B_a, Figure 5.13).

The As model does fit day 3.5 FACS data but it returns very low asymmetric transition rates in most of the different fits. This suggests that asymmetric transitions are extremely rare and, in fact, can be considered negligible with respect to symmetric ones. These results are supported by the fact that when we considered a model (OA model) in which C cells can derive only by asymmetric division of A or B cells ($A_s = B_t = 0$),we obtain asymmetric transition parameters close to 0 ($B_a \sim 0$) and that C cells remain a very small fraction of the population even at long times, as $\gamma < \beta$ (Figure 5.14).

We evaluated the AIC and BIC and found that both the AIC and BIC minima correspond best to the original model (Figure 5.15), i.e. the without asymmetric division. Taken together, these results suggest that asymmetric transition do not substantially contribute to Th2 differentiation.

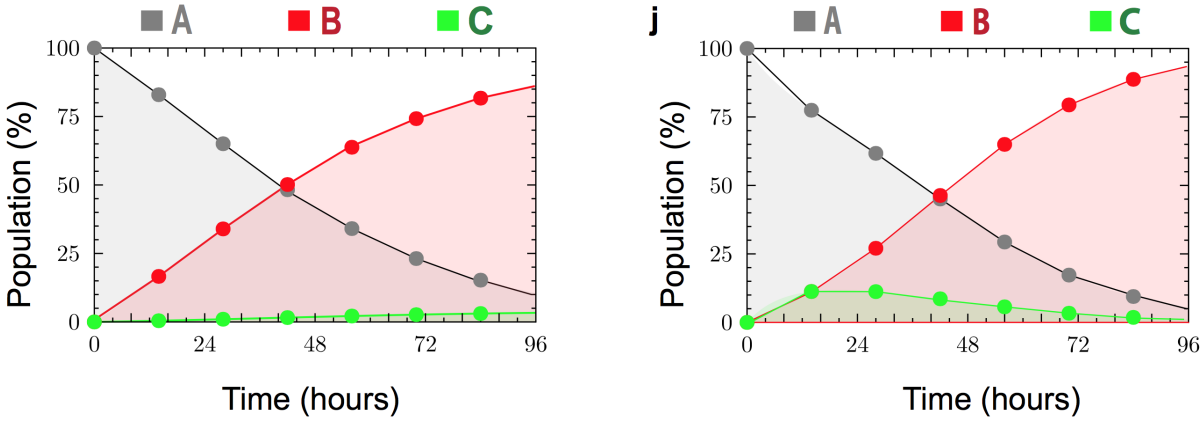


Figure 5.14 – The As (i) and OA (j) model- predicted dynamics of the population fractions of the three states over a 4-day period.

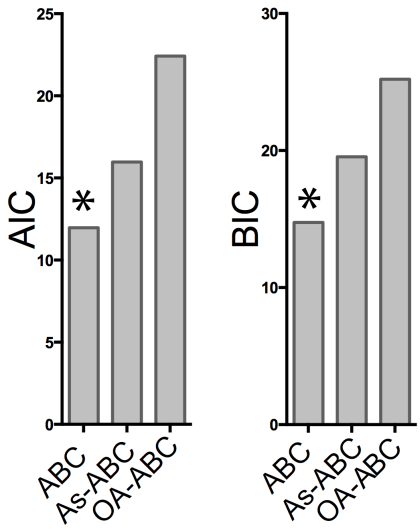


Figure 5.15 – AIC and BIC evaluated for day 3.5 across different models.

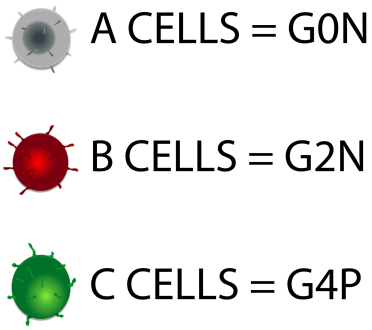


Figure 5.16 – To each state we assigned a specific expression profile as visualized here.

5.3 Validating the A, B and C cell states and parameters by expression profiling

Finally, we validated our model predictions with a dual experimental and computational approach. By combining the population dynamics predictions from the model with an RNA-seq data time-course, we aimed to link the cellular identity and the molecular characteristics of these cells.

First, we assigned a defined expression profile to each of the states: the G0N expression profile to A cells, the G2N profile to B cells and the G4P profile with C cells (Figure 5.16). Next, our model allow us to estimate the proportion of cells in each of the three states at different time points during Th2 differentiation (Figure 5.17). Based on the expression profile of each of the states, we are able to predict ensamble transcriptomic profiles at different time points.

To verify the accuracy of our predictions, we performed a time-course mRNA-seq experiment (6, 12, 24, 48 and 84 h post-activation) in the same culture conditions used before (Figure 5.17). We analyzed the expression profiles of all the genes at all time points and plotted the predicted versus the measured log(RPKM) values (Figure 5.18).

5.3. Validating the A, B and C cell states and parameters by expression profiling

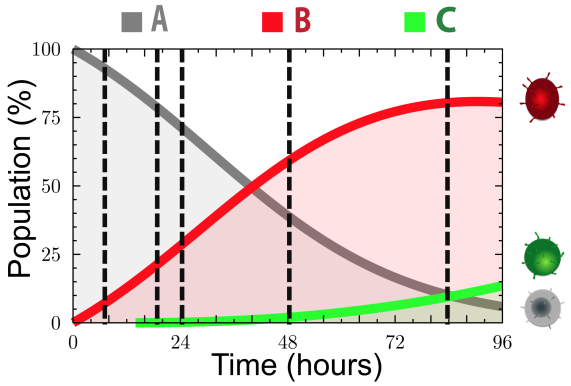


Figure 5.17 – Experimental setup of the time-course expression profiling. At the time points visualized total cells have been collected and total mRNA sequenced.

The correlation coefficient between the two was high ($r \sim 0.83$) and discrepancies were mainly at low expression levels, as expected.

Then we calculated the correlation between the single generation datasets, the model prediction and the new time-course data (Figure 5.19). As expected, the correlation of G0N decreases along consecutive time points while the correlation of G2N, G4N and G4P rises over time. Reassuringly, our prediction consistently have the highest correlation coefficient with the observed data over the whole time coirse (0.83 on average). Only the last point (84 h) correlates better with G4P expression data.

Globally, we observed that $\sim 43\%$ of total genes have a correlation $r > 0.5$; if we consider only DEGs, about $\sim 60\%$ of them have correlation $r > 0.5$ (Figure 5.20).

To further verify our prediction, we classified genes as negative and positive signature, i.e. genes that are overexpressed in one state only (positive) and genes that are downregulated in one state only (Figure 5.21).

5.3. Validating the A, B and C cell states and parameters by expression profiling

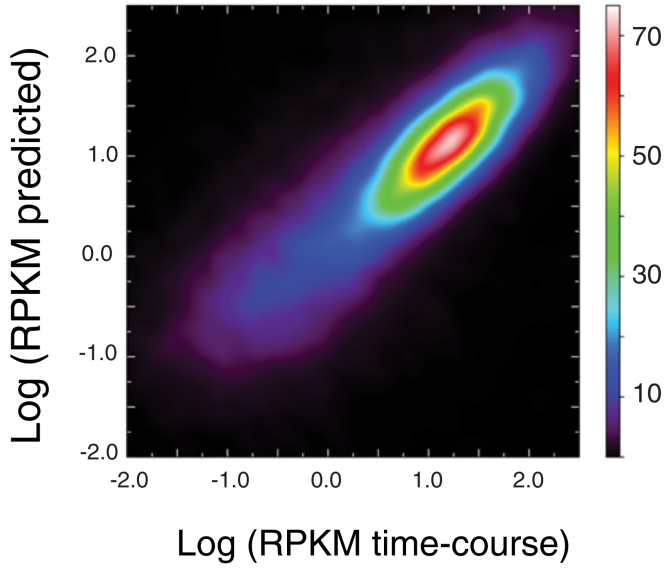


Figure 5.18 – Model validation in vitro and in vivo. a Cell culture expression time course: correlations of gene expression levels (logRPKM) between the model prediction and measured data at 6, 18, 24, 48 and 84 h. The colour scale represents the density of transcripts as a percentage of all expressed genes.

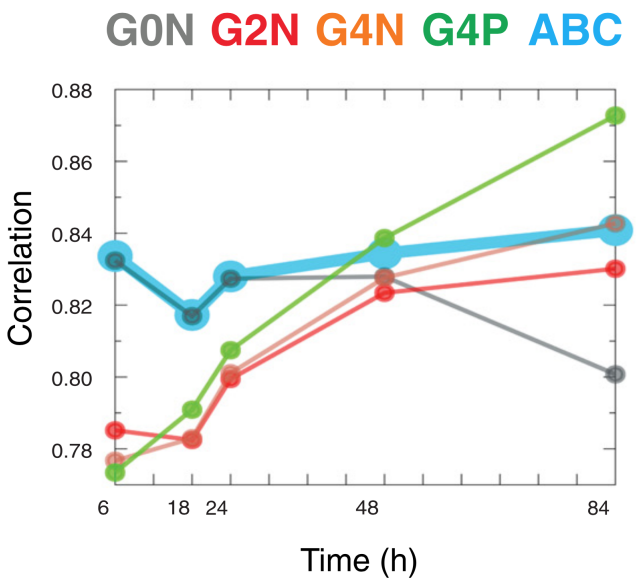


Figure 5.19 – Pearson correlation between model-predicted expression values (thick light blue line) and generation profiles (thin lines, grey for G0N, red for G2N, orange for G4N and green for G4P).

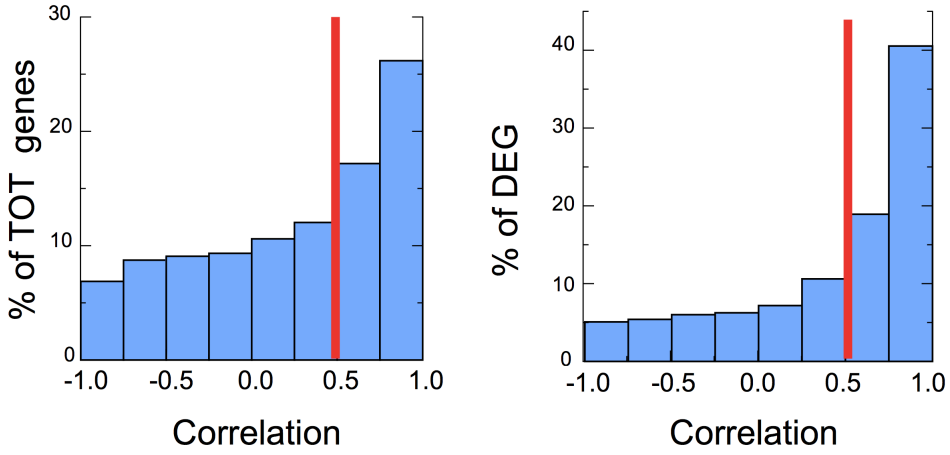


Figure 5.20 – Pearson correlation distribution between measured and predicted values for all genes and DEGs only.

5.3. Validating the A, B and C cell states and parameters by expression profiling

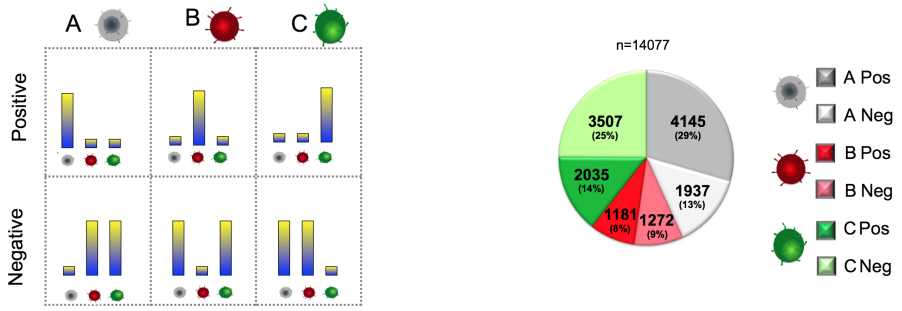


Figure 5.21 – Identification of positive and negative signature genes as specifically ON or OFF genes in that particular state and their relative abundances.

To minimize the noise, we considered only top 30% of the significant positive signature genes, as those genes should be most representative of each particular cell state. We compared our prediction (Figure 5.22) with measured data. Inspection of the data indicated that the trend was consistent for each of the three statess, with A-positive signature decresing over time, while B and C genes levels increase.

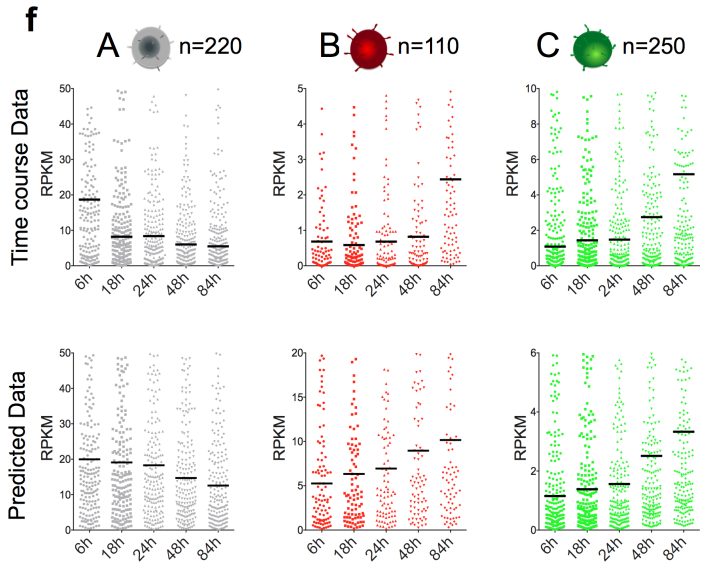


Figure 5.22 – Time-course data and the predicted expression levels of the 30 % top positive genes. The median is shown as a black solid bar.

5.3. Validating the A, B and C cell states and parameters by expression profiling

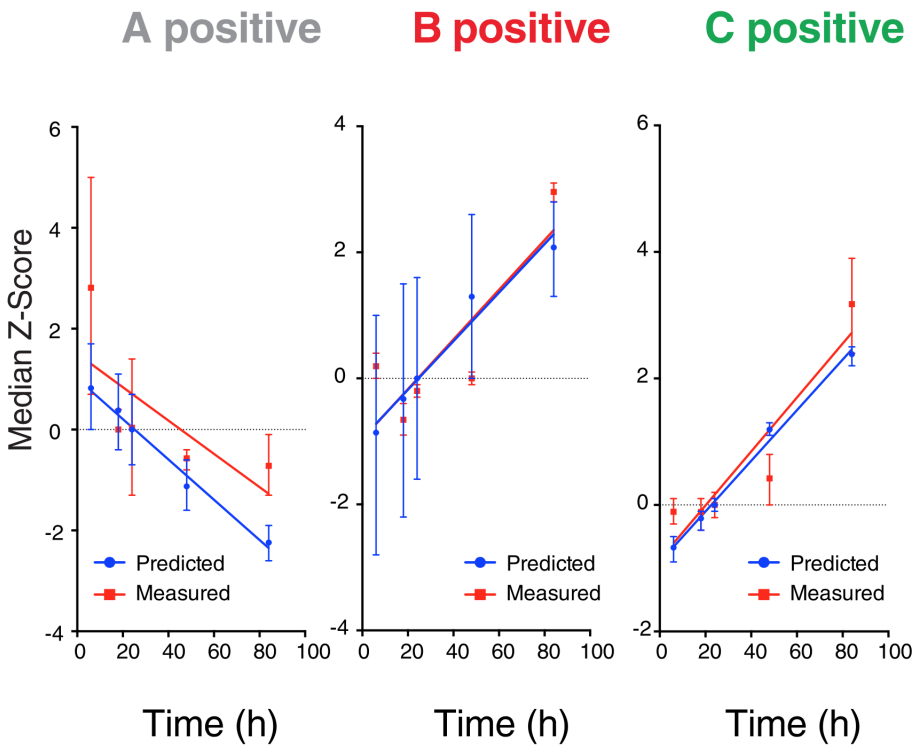


Figure 5.23 – For each of the states, the linear regression of the median values (Z-score normalized) is visualized as a red line for the time-course data and as a blue line for the predicted ones. Errors were calculated using data from different RNA-seq replicates.

Linear regression on the median expression (Figure 5.23) for the predicted and the experimentally determined data behave similarly, suggesting that our prediction fit well for all three states across all time points. Overall, the excellent agreement between the predicted and the measured RPKM values (Figure 5.18, 5.19 and 5.23) shows that our model is accurate in terms of transcriptomic changes in Th2 differentiation.

These results confirm not only the three cell states during Th2 differentiation but also the robustness of the cellular parameters inferred by the model.

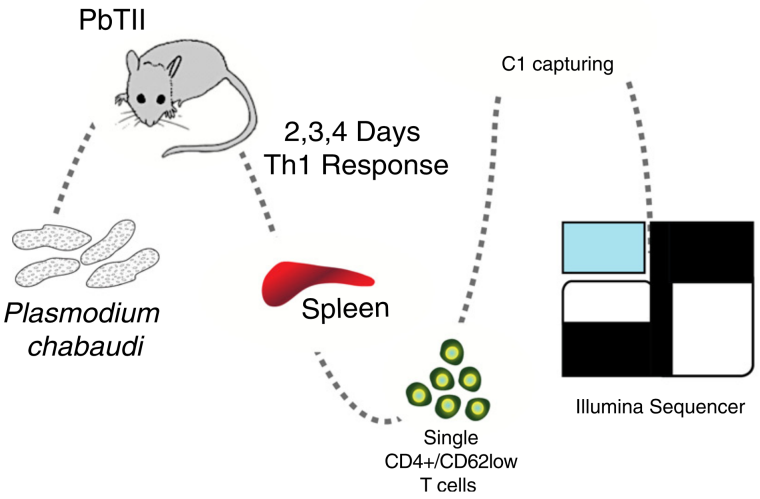


Figure 5.24 – Overview of splenic Th cells isolated from PcAS PbTII-infected mice at 2, 3 and 4 days post-infection.

5.4 Single-cell RNA-seq links CD4+ T-cell division rates to differentiation state in an in vivo Th1 infection model

To verify the link between cell cycle speed and differentiation rate in vivo and to ask if the model can be extended from Th2 to Th1 differentiation, we studied the CD4+ T-cell response against *Plasmodium chabaudi* AS (PcAS). Antigen-specific PbTII CD4+ T cells (CD45.1) were transferred into wild-type CD45.2 recipients and recovered from spleens at day 2, 3 and 4 post-infection (Figure 5.24)

As a measure of differentiation status inferred from the single-cell RNA-seq data, we developed a differentiation score based on the expression of "Th1 differentiation signature genes" [23]. We used aggregated G2/M gene expression levels across 26 genes as a "cell cycle score" reflecting division rate (analogous to [24]). Both of them are reported as

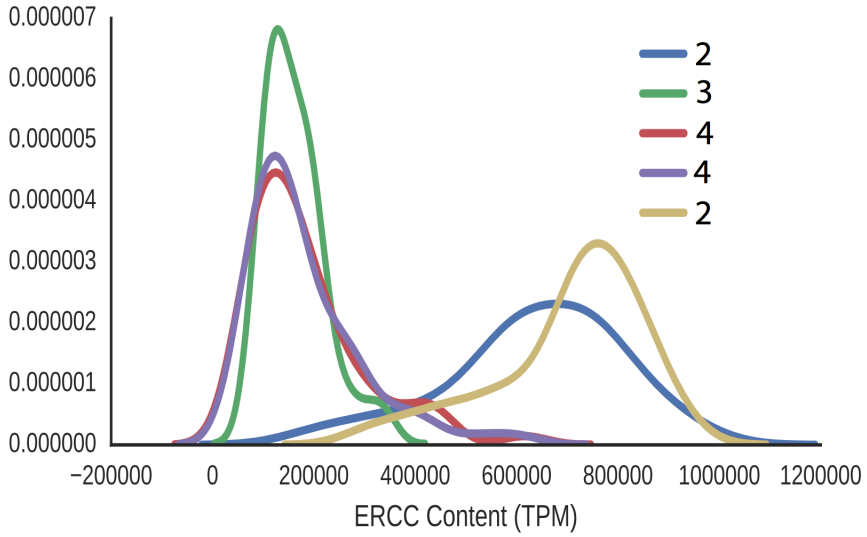


Figure 5.25 – Distribution of ERCC reads and total reads across samples.

TPM (transcript per million). We excluded that difference between cells in the different time points are confounded by batch effect with multiple analysis. First, the ERCC content across the two replicates of the same day shows no major differences (Figure 5.25). We can only observe a difference in the ERCC distribution across different days, but this variation can be due to change in cell size from day 2 to day 3 and 4 rather than ERCC degradation. In fact, the reduction in ERCC content from day 2 to days 3 and 4 can be described to changes in cell sizes and total mRNA content due to cell cycle activation (rather than ERCC degradation, which is an alternative explanation). Cells at day 2 have a diameter of roughly 6 μM, while at day 3 and 4 they can reach up to 10 μM of diameter. The corresponding cell volume can then vary by up to a factor of 5 with the result of a significant dilution of the ERCC molecules.

Moreover the number of reads is similar across all samples and PCA of total genes and ERCC clearly shows that replicates of the same day are similar to each others [1].

These single-cell transcriptomics data and bioinformatics analysis supports an increase

5.4. Single-cell RNA-seq links CD4+ T-cell division rates to differentiation state in an in vivo Th1 infection model

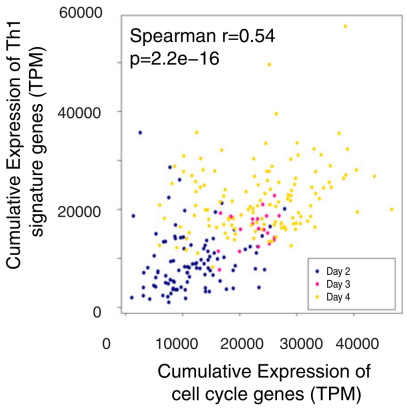


Figure 5.26 – Cumulative expression levels of cell cycle-associated genes (Cyclebase) and Th1 signature genes [23] in single CD4+ T cells as TPM (transcripts per millions). Cells from different days are plotted with different colours. The p value was calculated using the cor.test function in R, based on the Spearman's product moment correlation coefficient.

in proliferation in more mature, differentiated cells in vivo, as highlighted by the correlation between Th1 differentiatuion and cell cycle score (Figure 5.26).

Furthermore, similarly to our approach with Th2 cells, we focused only on the specific expression of *Ifng* as the main readout of Th1 cell differentiation.We first categorized cells in two subgroups: *Ifng*-low and *Ifng*-high cells (Figure 5.27).

When analysing the cumulative expression of cell cycle genes in these two subpopulation, we again observed that *Ifng*-high c ells were expressing significantly higher amounts of cell cycle genes compared with *Ifng*-low cells. This analysys, togheter with our previous results, confirms that our observation of a strong link between differentiation and cell cycle speed is conserved in different subtypes of T-h cells.

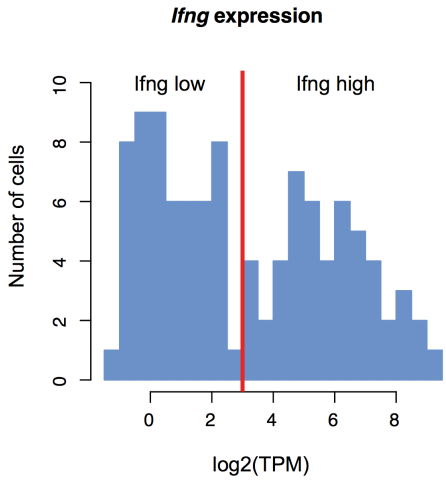


Figure 5.27 – Ifng expression across single cells. On this base we sorted cells in two subgroups: *Ifng*-low and *Ifng*-high cells

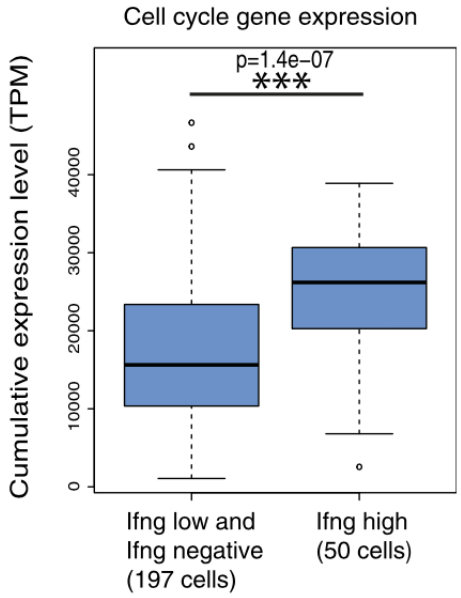


Figure 5.28 – The cumulative expression of 251 cell cycle-associated genes as TPM (Cyclebase database) in Ifng high cells and in cells expressing low levels of or no Ifng. The p value was calculated using a Wilcoxon rank sum test

5.4. Single-cell RNA-seq links CD4+ T-cell division rates to differentiation state in an in vivo Th1 infection model

Chapter 6

Summary and Conclusion

Analysing single cell transcriptomes of CD4+ T cells in a mouse model of a type 2 immune response, we noticed an inter-dependency between cell cycle rate and cytokine expression. Starting from this observation, we link a cell's history and state, obtained by flow cytometry analysis, with transcriptome-wide expression profiles. Thus, we infer specific probabilities of cell death, division and differentiation and associate these outcomes with transcriptomic profiles. What is unique about our strategy is that we sort cells from a single time point, thus reducing variability in confounding factors and allowing the identification of genes with essential function in lymphocyte differentiation (Figure 6.1).

This strategy can be applied to other cell types in which the cell cycle and differentiation are known to be inter-dependent (e.g. Th1, Th17 CD4+ cells, CD8+ T cells, intestinal epithelial cells, skin cells). Using this approach, we first employ traditional bulk RNA-seq to profile the gene expression dynamics among different generations of differentiated and undifferentiated Th2 cells. A deep analysis revealed three discrete cell states during the differentiation process from naïve cells to effector Th2 cells. We

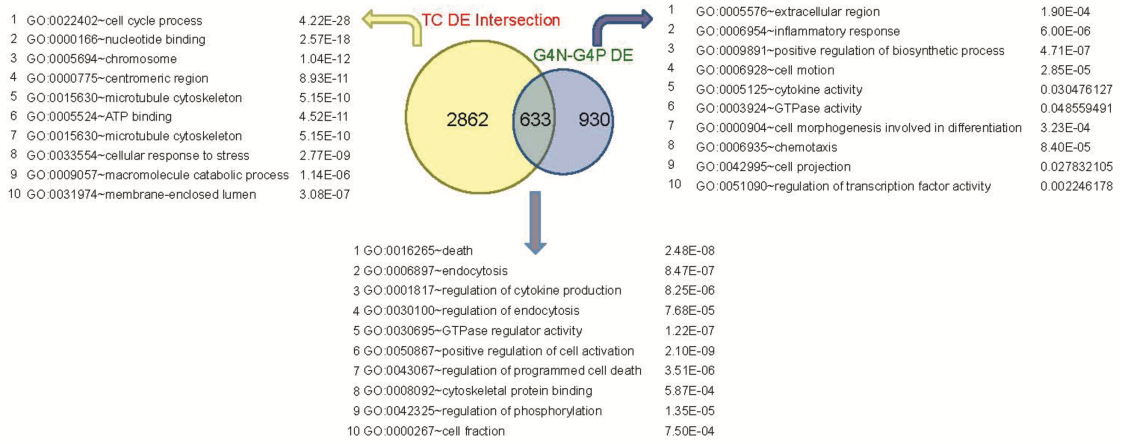


Figure 6.1 – GO analysis comparing the standard time-course approach and a generation-based approach. GO enrichment analysis of DEGs found only with a classic time-course-profiling experiment (in yellow, on the left) compared with genes found only with a generation- based approach (in blue, on the right). GO enrichment of the common genes is reported at the bottom.

named these three states A (activated, undivided) cells, which correspond to cells that have not entered the cell cycle (G0N), B (proliferating) cells, which are cycling but not differentiated (G2N and G4N), and C (Th2-like), which are cells that are cycling and expressing the differentiation marker (G4P). We validated the existence of these distinct states at the single-cell level by high-throughput single-cell qPCR, which supported a clear separation of G0N and G4P cells from the other cells and a single state across G2N and G4N cells. These gene expression analyses characterize the three cell states at the molecular level. We extracted the cellular properties of death, division and differentiation rates in each of the three states using a Markov process to model flow cytometry data. This confirmed a surprising new finding: that cell division rate increases in mature differentiated cells compared with activated cells as suggested by cell cycle gene expression patterns. Our findings at both the molecular and cellular levels are summarized in Figure 6.2.

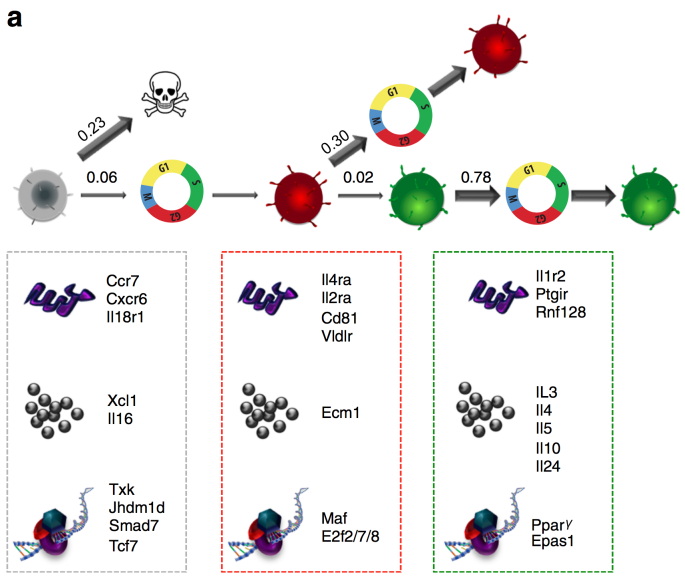


Figure 6.2

Cells in state A have a high probability of dying but can eventually enter the cell cycle and give rise to cells in state B. During this process, the expression level of some TFs specifically expressed in A cells is downregulated and the expression of new TFs (i.e. C-Maf [25]) is switched on. Amongst the A state-specific TFs are TFs with a known role in T-cell differentiation (e.g. Txk [26], Smad7 [27]), along with several regulators not previously recognized in the context of Th cell differentiation (e.g. Jhdm1d, Tcf7). For example, Xcl1 is known to act as a negative regulator of human CD4+ T-cell activation and to co-stimulate the apoptosis of human CD4+ T cells [28]. In our data, Xcl1 is downregulated after entering the cell cycle, allowing type 2 CK production to take place and providing molecular insight into the higher death rate of A cells. Further cell cycle repressors highly expressed in the A state are Ccr7, known to inhibit T-cell proliferation [28], and Il18r1, known to inhibit Th2 development [30]. Concurrently, the A-to-B transition exhibits an increase in expression of Il4ra, Il1r2, Ptgir, Rnf128

and Itk, in agreement with the literature [26, 30]. State C is characterized by the upregulation of several already known as well as previously uncharacterized transcription factors (i.e. Epas1, Pparg and Jhdm1d) in the context of Th2 differentiation. This new transcriptional network might contribute to the maintenance of a more defined and stable Th2 differentiated state, as observed by the higher cell-to-cell homogeneity of C cells compared with the other two states. Globally, our data confirm many molecular studies in the literature and at the same time provide a more comprehensive and quantitative definition of gene regulation during Th2 differentiation from a new perspective. This high-resolution data set will provide a rich resource for further molecular mechanistic studies of T-h cell differentiation. Finally, our mathematical model highlights important cellular characteristics of the three states. First, that undivided A cells have a higher death rate compared with activated cells. Second, that differentiated C cells have a higher division rate compared with B cells and that asymmetric divisions do not appear to contribute to the C state population. Live imaging of single lymphocytes not only confirmed our prediction but also gave a more robust quantification in terms of the difference in cell cycle length between cytokine -positive and -negative cells. Moreover, despite being extracted from both in vivo and in vitro differentiated Th2 cells, our findings on the link between cell cycle speed and differentiation status are conserved in in vivo Th1 differentiating cells. In particular, we used single-cell RNA-seq data from a model of CD4+ T cells purified from PcAS PbTII-infected mice at 2, 3 and 4 days post-infection. Amongst the ex vivo cycling cells, we confirm a correlation between cell cycle speed and differentiation. This finding supports a new concept during Th cell primary activation: not only is cell cycle entry a *sine qua non* for differentiation but a greater level of differentiation is associated with faster cycling cells; not only in Th2 cells but universally during primary activation of different subtypes of T lymphocytes. Our findings suggest that cytokine-secreting mature T cells comprise the fastest cycling

cell state as they are the most important cells for clearing an infection. Hence, the mechanism connecting cell division speed and cytokine secretion must have evolved for this purpose and may go awry in scenarios such as cytokine storms or, for instance, the severe immunopathology that can be associated with malaria. Insights into this process may be important to identify engineering strategies for T-cell therapy and targets for immunomodulation.

Bibliography

[1] Valentina Proserpio, Andrea Piccolo, Mario Nicodemi and Sarah Teichmann et al. Single-cell analysis of CD4+ T-cell differentiation reveals three major cell states and progressive acceleration of proliferation. Genome Biology. (2016) 17:103 DOI 10.1186/s13059-016-0957-5

[2] Juan G, Li X, Darzynkiewicz Z. Phosphorylation of retinoblastoma protein assayed in individual HL-60 cells during their proliferation and differentiation. Exp Cell Res. 1998;244:83–92.

[3] De Boer RJ, Ganusov VV, Milutinović D, Hodgkin PD, Perelson AS. Estimating lymphocyte division and death rates from CFSE data. Bull Math Biol. 2006; 68:1011–31.

[4] Zhu J, Yamane H, Paul WE. Differentiation of effector CD4 T cell populations. Annu Rev Immunol. 2010;28:445–89.

[5] Mariani L, Löhning M, Radbruch A, Höfer T. Transcriptional control networks of cell differentiation: insights from helper T lymphocytes. Prog Biophys Mol Biol. 2004;86:45–76.

[6] Yoon H, Kim TS, Braciale TJ. The cell cycle time of CD8+ T cells responding in vivo is controlled by the type of antigenic stimulus. PLoS One. 2010;5, e15423.

[7] Duffy KR, Hodgkin PD. Intracellular competition for fates in the immune system. Trends Cell Biol. 2012;22:457–64.

[8] Bird JJ, Brown DR, Mullen AC, Moskowitz NH, Mahowald MA, Sider JR, et al. Helper T cell differentiation is controlled by the cell cycle. Immunity. 1998;9: 229–37.

[9] Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 2014;509:371–5.

[10] Gay D, Maddon P, Sekaly R, Talle MA, Godfrey M, Long E, et al. Functional interaction between human T-cell protein CD4 and the major histocompatibility complex HLA-DR antigen. Nature. 1987;328:626–9.

[11] Neill DR, Wong SH, Bellosi A, Flynn RJ, Daly M, Langford TK, et al. Nuocytes represent a new innate effector leukocyte that mediates type-2 immunity. Nature. 2010;464:1367–70.

[12] Gett AV, Hodgkin PD. Cell division regulates the T cell cytokine repertoire, revealing a mechanism underlying immune class regulation. Proc Natl Acad Sci U S A. 1998;95:9488–93.

[13] Ziegler SF, Ramsdell F, Alderson MR. The activation antigen CD69. Stem Cells. 1994;12:456–65.

[14] Dailey MO. Expression of T lymphocyte adhesion molecules: regulation during antigen-induced T cell activation and differentiation. Crit Rev Immunol. 1998;18:153–84.

[15] Sasaki T, Onodera A, Hosokawa H, Watanabe Y, Horiuchi S, Yamashita J, et al. Genome-wide gene expression profiling revealed a critical role for GATA3 in the maintenance of the Th2 cell identity. PLoS One. 2013;8, e66468.

[16] Peschon JJ, Morrissey PJ, Grabstein KH, Ramsdell FJ, Maraskovsky E, Gliniak BC, et al. Early lymphocyte expansion is severely impaired in interleukin 7 receptor-deficient mice. J Exp Med. 1994;180:1955–60.

[17] Skrenta H, Yang Y, Pestka S, Fathman CG. Ligand-independent down- regulation of IFN-gamma receptor 1 following TCR engagement. J Immunol. 2000;164:3506–11.

[18] Betz BC, Jordan-Williams KL, Wang C, Kang SG, Liao J, Logan MR, et al. Batf coordinates multiple aspects of B and T cell function required for normal antibody responses. J Exp Med. 2010;207:933–42.

[19] Kosaka S, Tamauchi H, Terashima M, Maruyama H, Habu S, Kitasato H. IL-10 controls Th2-type cytokine production and eosinophil infiltration in a mouse model of allergic airway inflammation. Immunobiology. 2011;216: 811–20.

[20] Nelms K, Keegan AD, Zamorano J, Ryan JJ, Paul WE. The IL-4 receptor: signaling mechanisms and biologic functions. Annu Rev Immunol. 1999;17:701–38.

[21] Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol. 2015;33:269–76.

[22] Arsenio J, Kakaradov B, Metz PJ, Kim SH, Yeo GW, Chang JT. Early specification of CD8+ T lymphocyte fates during adaptive immunity revealed by single-cell gene-expression analyses. Nat Immunol. 2014;15:365–72.

[23] Stubbington MJ, Mahata B, Svensson V, Deonarine A, Nissen JK, Betz AG, et al. An atlas of mouse CD4(+) T cell transcriptomes. Biol Direct. 2015;10:14.

[24] Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344:1396–401.

[25] Ho IC, Hodge MR, Rooney JW, Glimcher LH. The proto-oncogene c-maf is responsible for tissue-specific expression of interleukin-4. Cell. 1996;85:973–83.

[26] Sahu N, Venegas AM, Jankovic D, Mitzner W, Gomez-Rodriguez J, Cannons JL, et al. Selective expression rather than specific function of Txk and Itk regulate Th1 and Th2 responses. J Immunol. 2008;181:6125–31.

[27] Yan X, Liu Z, Chen Y. Regulation of TGF-beta signaling by Smad7. Acta Biochim Biophys Sin Shanghai. 2009;41:263–72.

[28] Ziegler E, Oberbarnscheidt M, Bulfone-Paus S, Förster R, Kunzendorf U, Krautwald S. CCR7 signaling inhibits T cell proliferation. J Immunol. 2007; 179:6485–93.

[29] Yu Q, Chang HC, Ahyi AN, Kaplan MH. Transcription factor-dependent chromatin remodeling of Il18r1 during Th1 and Th2 differentiation. J Immunol. 2008;181:3346–52.

[30] Horiuchi S, Onodera A, Hosokawa H, Watanabe Y, Tanaka T, Sugano S, et al. Genome-wide analysis reveals unique regulation of transcription of Th2- specific genes by GATA3. J Immunol. 2011;186:6378–89.

[31] Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26:873–81.

[32] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.

[33] Kolodziejczyk AA, Kim JK, Tsang JC, Ilicic T, Henriksson J, Natarajan KN, et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell. 2015;17:471–85.

[34] Mehdi T, Bashardoost N, Ahmadi M. Kernel smoothing for ROC curve and estimation for thyroid stimulating hormone. Int J Public Health Res Special Issue. 2011;239–42.

[35] Juan G, Li X, Darzynkiewicz Z. Phosphorylation of retinoblastoma protein assayed in individual HL-60 cells during their proliferation and differentiation. Exp Cell Res. 1998;244:83–92.

[36] K Sugimoto, SP Gordon, and EM Meyerowitz. Regeneration in plants and animals: dedifferentiation, transdifferentiation, or just differentiation? Trends Cell Biol., 2011.

[37] Monya Baker. Adult cells reprogrammed to pluripotency, without tumors. Nature Reports Stem Cells, 2007.