Studying a stochastic gene regulatory network driving the germinal center B cell differentiation

Alexey Koshkin


Numéro National de Thèse: 2021LYSEN083

THESE DE DOCTORAT DE L'UNIVERSITE DE LYON
OPÉRÉE PAR

L'ECOLE NORMALE SUPÉRIEURE DE LYON

Ecole Doctorale N°340
Biologie Moléculaire Intégrative et Cellulaire
Spécialité de doctorat: Bio-informatique et Biologie des Systèmes
Discipline: Sciences de la vie et de la santé

Souteneu publiquement le 16/12/2021, par:

Alexey Koshkin

Studying a stochastic gene regulatory network driving the germinal center B cell differentiation

Étude d'un réseau de régulation génétique stochastique déterminant la différenciation des lymphocytes B dans un centre germinatif

---

Devant le jury composé de:

CRÉPIEUX Pascale	Directrice de recherche, INRAE	Rapporteure
MARTÍNEZ María Rodríguez	Directrice de recherche, IBM Zurich	Examinatrice
ROPERS Delphine	Chargée de recherche, INRIA Grenoble	Rapporteure
CRAUSTE Fabien	Directeur de recherche, Université de Paris	Co-Directeur de thèse
GANDRILLON Olivier	Directeur de recherche, LBMC, ENS Lyon	Directeur de thèse

Acknowledgments

This work was performed at ENS de Lyon "Laboratoire de Biologie et Modélisation de la Cellule" and Inria Dracula team "Multiscale modelling of cellular dynamics", as a part of the COSMIC Marie Curie. I would like to say thank you to my supervisors Fabien Crauste and Olivier Gandrillon for accepting me into the group, for making my secondments possible, even in times of world pandemic, for their constant guidance, scientific precision, motivation, and for sharing their knowledge. To Arnaud Bonnaffoux and Ulysse Herbach, for introducing me to the field of stochastic modelling. To Ronan, Geneviève, Elias, Camille, Margaux, Matteo, Gerard, and all members of "Systems Biology of Decision Making" for scientific discussions. Also thanks to Claire, Mostafa, and all members of Inria Dracula team. I also want to thank IN2P3 administrators and colleagues for always answering my questions and trying to solve all possible issues. I want to acknowledge my secondment supervisor María Rodríguez Martínez (at IBM Zurich) and Stefano Casola (at IFOM Milan) for receiving me in their labs, making me feel welcome and comfortable, for very interesting and productive meetings and discussions. I also want to thanks all the COSMIC supervisors and organizers of the courses, conferences, and meetings. I also want to say thank you to all COSMIC fellows, and especially to Aurélien, Dani and Rodrigo for hours of scientific and not-so-much discussions.

Aurélien we almost manage to make a kite! Rodrigo, thank you for creating such a positive spirit (know)! Dani, thank you for andaluz and for Bohemian Rhapsody.

It was a very long journey and I want to give all the credits and a lot of thanks to all friends I was lucky to make on this road. First I want to say thank you krav de Lyon and especially Stephane for receiving me in the team, for all the trust, for sharing good vibes, and for being a great friend! Thank you David for French lessons, for great French comedies and "you are not English?". Also thank you to Sérgio, Jean-Pierre, Nath, Philippe, Alexandra, Sebastian for your amazing hours and for your team spirit, KIDA! Thank you PT krav mates: Nuno, Rui, Quim for the great vibes and for an amazing time!

I want to say a big thank you to Shiny and Venkat for always having their door open and for delicious Indian meals, which I never can call correctly. Thank you Andreia for the great Greek community :) Also thanks to Dima and Kirill for our interesting and long talks. Thanks to Antone and Jairo for bringing a piece of Brazil with them.

Big thanks to my ZH mates Savina, Greta, Elly, Martina, Nadine, Pol, Angela for a great time! Thank you Miguel for sharing a part of PT and for explaining small, but important details in PT history! Thank you Ming for our endless chats, for a great year in woko, and for your friendship! Thank you, Sanja for the tasty cakes, kind advice and for amazing training! Thank you Chantal for our long philosophical talks, for introducing me to the art, showing that it can be fun, for beautiful hikes, and for cheering for me during tournaments. Thank you Lara for the great experience in BBB field, for sharing your positive energy, kindness and for trusting in me. Also, thanks to my dear BBB mates

- Fufu, Julia, Jay - Friday breakfast was indeed a good routine. Thank you Elena for showing beautiful gels and for bringing Kumasi to work.

Of course, many thanks to Alexandre, Anys, Ana, Margarida, Barbara, Tiago, Ricardo, Hugo, Madalena - you guys made the first lab experience unforgettable! Thank you Daniel Pais for showing what is metabolism, why it is so remarkable, and why they made a song about cookies! Thank you Daniela for being a great friend, independently of the distance, years, or languages (and for tasty bolos de azeitão). Of course thank you Ana Teixeira for introducing me to Systems Biology, for showing me that mathematics is not just a set of beautiful formulas, but has great potential in the future biology and medicine! Also, thank you Ana for giving me a hand, believing in me when I first entered the virus lab, for teaching me how to make multiple tasks at once and how to succeed in all of them. Great thanks to Patricia for listening to all my MFA talks, for trying to pretend that it is interesting, and for helping me pipetting when the wells were too small. Thank you Cláudia for showing the beauty of cardiostem (CMs do contract and make synchronize waves) and for sharing your knowledge with me. Thank you for your friendship, support, and laughs during all these years, obrigado! Special Spasibo to my dear mom and dear dad for always being a great example to follow, for all their trust, support, dedication, and wisdom. Thank you for believing in me and for showing me what is important in life and that nothing is impossible. Without you, I would never be who I am.

To my Family.

Abstract

Germinal Centres (GCs) are the histological structures dedicated to generation and selection of B cells that produce high-affinity antibodies. However, unexpected malfunctioning in the GCs can cause the appearance of different pathologies, including the malignant transformation of B cell. Understanding the Gene Regulatory Networks (GRNs) which orchestrate that response is therefore a critical task. GRNs describe how the genomic sequence encodes the regulation of expression sets of genes that are responsible for generation of developmental patterns and execution of the multiple states of differentiation. Inferring and evaluating GRNs from gene expression data is a long-standing and challenging task in systems biology. Novel technology allows us to measure mRNA levels in individual cells, which promise significant increase of GRNs precision, but will require relevant models. Our aim was to assess the use of a new stochastic executable model for GRNs made from coupled Piecewise Deterministic Markov Process (PDMP) to fit single cell transcriptomic data from the GCs B cells differentiation sequence. We showed that our PDMP model, which was build from the coupling of three transcription factors and two cell surface receptors, can qualitatively estimate the distributions of the mRNA at different stages of GC B cell differentiation. A partial quantitative agreement was obtained through systematic parameter tuning but a full quan-

titative agreement proved to be highly challenging. PDMP allows to evaluate the structure of the GRN, and in the future may lead to further understanding of the different types of dysfunctions of the regulatory mechanisms.

Résumé

Les centres germinatifs (CG) sont les structures histologiques dédiées à la génération et à la sélection des cellules B qui produisent des anticorps de haute affinité. Cependant, des dysfonctionnements inattendus dans les CGs peuvent provoquer l'apparition de différentes pathologies, y compris la transformation maligne de ces cellules.Comprendre les réseaux de régulation de gènes (RRG) qui orchestrent cette réponse est donc une tâche critique. Les RRG décrivent comment la séquence génomique code la régulation de l'expression des gènes qui sont responsables de la génération des multiples états de différenciation. L'inférence de ces RRG à partir des données d'expression des gènes est une tâche ancienne et difficile en biologie des systèmes. Les nouvelles technologies permettent de mesurer les niveaux d'ARNm dans des cellules individuelles, ce qui promet une augmentation significative de la précision des RRG, mais nécessite des modèles pertinents. L'objectif de la thèse consistait à évaluer l'utilisation d'un nouveau modèle stochastique de RRG construit à partir d'un couplage de processus de Markov déterministes par morceaux (PDMP) pour ajuster les données transcriptomiques de cellules individuelles provenant de la séquence de différenciation des cellules B de CGs chez l'homme.Nous avons montré que ce modèle PDMP, qui a été construit à partir du couplage de trois facteurs de transcription et de deux récepteurs de surface cellulaire, peut

estimer qualitativement les distributions de l'ARNm à différents stades de la différenciation des cellules B de CGs. Un accord quantitatif partiel a été obtenu par le réglage systématique des paramètres, mais un accord quantitatif complet s'est avéré particulièrement difficile. Notre modèle PDMP permet d'évaluer la structure du RRG et, à l'avenir, pourrait permettre de mieux comprendre les différents types de dysfonctionnements des mécanismes de régulation, responsables de l'apparition de pathologies.

Summary

Fast development of Single Cell (SC) technologies and increased availability of computational clusters urges the development of novel tools for analysis and interpretation of SC transcriptomic data. Our contribution shows that stochastic PDMP model (2.4)-(2.6), applied a Gene Regulatory Network (GRN) made of three key genes (BCL6, IRF4, BLIMP1), is capable to qualitatively simulate the distribution of B cells mRNA at GC and PB_PC stages of differentiation. We hypothesized that the stochastic nature of our model could simulate experimental SC data and could depict its natural variability. To our knowledge, simulation of GRNs for SC data is mainly presented by Boolean Networks, co-expression and ODE models. Only recently, stochastic models started to be used to evaluate the variability in the SC data [1]. This study contribute to the understanding of potential applicability of stochastic model for simulation of experimental SC data in the field of computational immunology. Due the universality of the models used, it can be further used for SC analysis in any biological field.

We have approached the simulation of GRN in a few consecutive steps. We started with the kinetic ODE model (2.1)-(2.3), applied to the BCL6-IRF4-BLIMP1 GRN described by Martinez et al. [2]. We rewrote the PDMP model in terms of the kinetic ODE model (2.1)-(2.3) and used it to estimate the ini-

tial guess for the parameter set (see Tables 2.2-2.5, version I). We have used the parameter set for which the kinetic ODE model (2.1)-(2.3) associates the first steady state with the GC stage and the second steady state with the PB_PC stage of B cell differentiation. Interestingly, the mathematical rewriting allowed our ODE reduced PDMP model (2.11) to simulate the existence of two steady states (see Figure 2.2). We noticed that the parameter $k_{on,init,IRF4}$ estimated based on the rewriting of the ODE reduced PDMP System (2.11) in terms of the kinetic ODE System (2.1)-(2.3) can preserve an existence of two steady states (see Figure 2.2). Selecting random value of $k_{on,init,IRF4}$ resulted in the ODE reduced PDMP System (2.11) predicting the existence of only one steady state (see Supplementary Figure S2). The results that ODE reduced PDMP can qualitatively recapitulate two steady states of B cell differentiation, gave us further confidence to proceed with the PDMP System (2.4)-(2.6) and further investigate if it can be used to simulate the distribution of B cell differentiation at GC and PB_PC stages at the single cell level.

In the second part of our work, we executed the PDMP model and evaluated how well it describes the SC data of GC B cells, derived from human spleen and tonsil [3]. Initially, we used the same parameter set as for ODE reduced PDMP model, and for that reason we were not expecting that it completely reproduce the behavior of the single cell data. As we suggested, the PDMP (2.4)-(2.6), generated mRNA levels for BCL6, IRF4 and BLIMP1, which did not reproduce the experimental data (see Supplementary Figure S3) neither at GC nor at PB_PC stages and further improvement was required.

To investigate model's variability, we evaluated effect of the stochasticity for the identical parameter set of the PDMP model (2.4)-(2.6). We had to anal-

yse if for an independent simulation of the PDMP model (2.4)-(2.6) with the same parameter set, the model-generated number of mRNA would be similar between each other. For that purpose we used the KD and simulated multiple runs of the PDMP model (2.4)-(2.6) with the parameter set (see Tables 2.2-2.5, version II). We selected for each of gene (BCL6, IRF4 and BLIMP1) and for each stage (GC and PB_PC) the model-generated distribution of normalized mRNA values with the highest KD between each other and have seen that they posses high similarity between each other: the shapes of distributions were overlapping and mean values were comparable (see Figure 2.4 and Table 2.7). This have showed that while the PDMP model (2.4)-(2.6) is stochastic, it is capable to simulate the distributions of mRNA values with good reproductivity between independent execution. This allowed us to proceed with the parameter tuning, with a goal to improve the quality of the fitting of the experimental SC mRNA data.

After we determined that the PDMP model (2.4)-(2.6) generates similar outputs for a given parameter set, we performed the parameter search to find a parameter set that would better describe the experimental data from Milpied et al. [3]. We firstly tried an automatized approach, and screened parameters using a grid (see Figure 2.5).

Next, we used the knowledge of the structure of the GRN network (see Figure 2.1) to perform a semi-manual and more precise tuning of the parameters. After defining key parameters, we were able to increase the number of IRF4 and BLIMP1 molecules at the PB_PC stage (see Figure 2.5) and obtained better fits of the data.

We showed that the PDMP model, based on the GRN of key regulators BCL6,

IRF4 and BLIMP1, can be used to simulate experimental SC mRNA data of GC B cell differentiation. Firstly, it shows achievements of the computational biology, which is capable to simulate the complex differentiation process, using the PDMP model and concept of GRN. Second it can be used to further understanding biological processes, occurring during the differentiation or for any biological event, caused by an application of short time stimuli (for instance, for immune response). By manipulating the GRN structure (adding the regulations and new genes), one can evaluate an effect of candidate genes on the system. Also, due to executable properties of the PDMP model (2.4)-(2.6), one can potentially study an effect of the different layers of the biological process (promoter activity, transcription and translation rates), using as a reference the experimental SC mRNA data.

To summarise, we obtained reasonable fitting of the previously published experimental SC mRNA data. Different strategies can be used to further improve the fitting. For instance to perform more massive semi-manual tuning strategy, where one can screen the multiple parameters responsible for the specific reaction or subGRN (responsible for BCL6 repression). Otherwise, improvement of the structure of the GRN also may increase the quality of the fitting. The structure of the network may lack of reactions, preventing the PDMP System (2.4)-(2.6) from generation of distributions identical to experimental ones. And the last possibility, is that the experimental SC dataset used for this study, could be more complete. SC data was collected from the human lymphoid organs, and then were classified as B cells at the GC and PB_PC stage, based on the cell sorting and pseudotime algorithm. Nevertheless, up to date, the dataset produced by Milpied et al. [3] is the most novel and complete

analysis of human B cell, at the moment of current study, which present the data on the single cell level. Future access to the alternative data, would be advantageous and would allow to confirm an applicability of the PDMP model (2.4)-(2.6) for simulation of different GC B cell fates.

List of Abbreviations

ABC-DLBCL activated B cell-like DLBCL. 11

APC antigen-presenting cells. 6

BCL6 B-cell lymphoma 6. 44

BCR B cell receptors. 4

BFU-E burst-forming unit–erythroid. 6

BLIMP1 PR domain zinc finger protein 1. 10

CCA canonical correlation analysis. 28

CLP common lymphoid progenitors. 5

CMP common myeloid progenitor. 6

DC Dendritic Cells. 15

DLBCL diffuse large B cell lymphoma. 10

DZ dark zone. 44

Eo-CFC eosinophil colony-forming cells. 6

FACS Fluorescence-activated cell sorting. 24

FL Follicular lymphoma. 12

G-CFC granulocyte colony-forming cells. 6

GCB-DLBCL GC B cell-like DLBCL. 11

GCs Germinal Centres. iv, 43

GM-CFC granulocyte macrophage colony-forming cells. 6

GRNs Gene Regulatory Networks. iv, 22

HSC Hematopoietic Stem Cells. 5

IgV immunoglobulin variable regions. 7

IRF4 Interferon regulatory factor 4. 44

KD Kantorovich distance. 55

KDE Kernel Density Estimate. 62

LZ light zone. 44

M-CFC macrophage colony-forming cells. 6

mast-CFC mast colony-forming cells. 6

MC Memory Cells. 18

OC-CFC osteoblastic colony-forming cells. 6

ODE Ordinary Differential Equations. 13

PB plasmablast stage. 11

PB_PC Plasmablast, Plasma Cell. 44

PC plasma cells. 11

PCA principal component analysis. 47

PDMP Piecewise Deterministic Markov Process. iv

PL Piecewise-linear. 31

pro B cell the earliest committed B cells. 5

SC single cell. 24

scRNA-seq single-cell RNA sequencing. 24

SHM somatic hypermutation. 4

T-fh cell T follicular helper cells. 7

TCR T cell receptors. 4

TFs transcription factors. 22

UMIs unique molecular identifiers. 25

Contents

List of Figures

List of Tables

Chapter 1

Introduction

During many centuries brightest minds of humanity have been trying to understand how our body confronts diseases and why some individuals overcome illness without severe symptoms, while others have significantly weaker tolerance and lower survival rate. One of the earliest observations was performed by ancient Greek physicians. They noticed that patients who already had survived the plague, had much higher odds to be asymptomatically reinfected [4]. Later on, first attempts to induce the immunity in a population have been performed in China in the fifteenth century. The method, called variolation, consisted in the introduction of the dry crust from smallpox pustule of a patient to an open wound of a healthy individual. There was a strong belief that variolation could prevent the appearance and spreading of the disease, however the balance between efficiency and risk was quite low [5].

Interestingly, History recognizes the English physician Edward Jenner as pioneer of systematic vaccination against smallpox. He observed that the milkmaids, who contacted with infected cows, had high tolerance against the human version of the disease. To test the hypothesis that cow virus promotes

the adaptation to a human version, he collected the cowpox pustule and introduced it to a eight-year-old boy. Fortunately, the virus inoculation did not cause the appearance of the disease and the boy had developed a resistance to smallpox. Although the ethical part of Jenner's experiment is excessively questionable, his method successfully spread in Europe, and had a tremendous effect on the fighting against smallpox epidemic [5, 6]. One hundred years later Louis Pasteur discovered that the old bacterium, causing fowl cholera in culture while injected to the chicken, creates the immunity against the stronger stripes of cholera. He suggested and lately proved that the attenuated strain with a weakened virulence can be used to create an immunity. Further, Pasteur called his strain a "vaccine" to honor the work of Jenner. In subsequent years, humanity developed a vast number of vaccines using the same principle of attenuation, including rubella, influenza, rotavirus, tuberculosis, etc. [7].

Despite the countless achievements in the understanding of immunology during last century, this field is still full of enigmas. Nowadays the development of various mathematical models, together with the increasing computational power, promises to shade light on the mechanisms of the immune system. We are reaching the moment when deeper understanding of biological processes would not be possible without strong collaboration and support from mathematical and computational sciences.

1.1 Principle and organisation of immune response

1.1.1 Innate and acquired immune responses - General picture

The human immune system is a very complex coordinated network, whose main goal is to eliminate pathogens and maintain tissue homeostasis.

This system consists of dozens immune cell types, which are distributed in the organism. Immune cells are mainly developed from immature precursors in primary lymphoid organs (the bone marrow, the thymus). In combination with a spread network of secondary lymphoid organs, such as spleen, lymph nodes and different types of mucosal tissues, the immune system allows the body to provide highly efficient protection [8].

The immune system historically was divided in innate and acquired immune responses. Despite the fact that both mechanisms support each others functions, they have different roles in the complex response against infections [9].

Innate immunity is a first line of defense. It includes external physical barriers to the environment (i.e. skin, mucosa, protective structures), humoral innate immunity (antimicrobal proteins, soluble factors, cytokines, peptides) and cells (intraepithelial T lymphocytes, nuocytes, myeloid phagocytic cells, innate lymphoid cells, phagocytic B cells, etc.). Physical and chemical barriers protect the organism from invasion of the microorganisms or viruses. They mainly consist in the epithelial layer of cells with tight junctions and proteins,

providing the mechanical prevention of the pathogen infiltration [10]. Humoral innate immunity consists of the variant complex molecules, distributed in the extracellular fluids, and their main property is to destroy the pathogens and prevent the spread of the infection. The cell response is normally caused by intraepithelial T lymphocytes, myeloid phagocytic cells, phagocytic B cells and non-specific cytotoxic cells. Although the innate immune response is immediate, it is limited to a certain range of pathogens it can recognise. This is mainly caused by the imbalance between low number of the genetically predetermined germ line-encoded receptors of the cells participating in innate immune response versus the highly variable antigenic structures of the different pathogens [11].

The adaptive immunity on the other hand consists of the two major types of lymphocytes: bone-marrow-derived B cells and the thymus-derived T cells. Those lineages have high capacity of specific recognition and respond to antigenic variants of the pathogens. During B and T lymphocyte development, they acquire the set of prototypic immunoglobulin variables gene segments, such as V (variable), D (diversity) and J (joining). Together they represent V(D)J regions of the B cell receptors (BCR) and T cell receptors (TCR) [12]. High diversity of the adaptive immune response is provided by the constant modifications of the V(D)J region of B cells via somatic hypermutation (SHM) and class switch recombination ()[13]. Such a mechanism allows the immune response to react and confront various antigens (i.e. molecules/molecular structures covering pathogens) with a high specificity and efficiency.

1.1.2 Development of B cells from multipotent stem cells

B-cells originate during haematopoiesis, the multistep process of blood cell production. Hematopoietic Stem Cells (HSC) are multipotent stem cells capable to generate all types of blood cells, such as erythrocytes, megakaryocytes, monocytes, eosinophils, neutrophils, dendritic cells, natural killer cells/innate lymphoid cells, T cells and B cells, using different sets of transcriptional programs (see Figure 1.1, [14, 15]).

The production of B cells starts in primary lymphoid organs: bone marrow and thymus. The main properties, essential for correct B cell functions, are acquired consequently at each stage of development. Differentiation starts with transition of HSC to the multipotent progenitor cells (pro B cell) and consequently follows through common lymphoid progenitor (CLP), pro B cell (the earliest committed B cells), early pre B cell, late pre B cell, and finally to mature B cell stages.

The expression of specific membrane proteins allows to characterize the stage of the B cell development. At each step of the differentiation, specific sets of immunoglobulin rearrangements, expressions of surface molecules and different transcription factors are determined [16]. B cells perform the final stages of their differentiation in the peripheral lymphoid organs (spleen, lymph nodes, Peyer's patches and tonsils) and then they migrate to the blood flow, getting distributed in the whole body being capable to recognise antigens [17, 18].

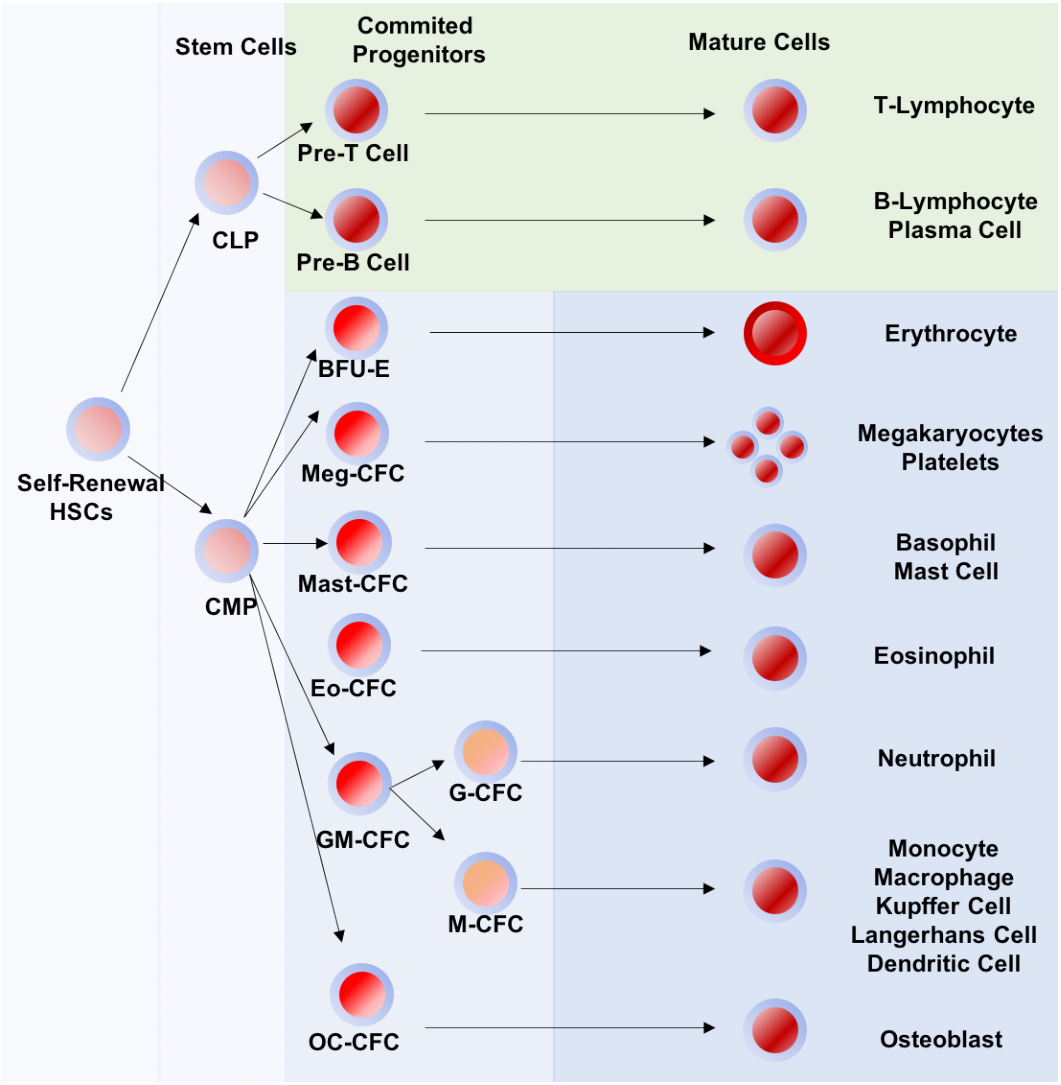


Figure 1.1. Schematic representation of the hematopoiesis process. All blood cells, including B lymphocytes, are produced from the successive differentiations of hematopoietic stem cells HSC. CLP - common lymphoid progenitors; CMP - common myeloid progenitor; BFU-E - burst-forming unit–erythroid; Meg-CFC - megakaryocytic colony-forming cell; mast-CFC - mast colony-forming cells; Eo-CFC - eosinophil colony-forming cells; GM-CFC - granulocyte macrophage colony-forming cells; G-CFC - granulocyte colony-forming cells; M-CFC - macrophage colony-forming cells; OC-CFC - osteoblastic colony-forming cells.

1.1.3 Germinal centres initialization

Apart from the blood, B cells mainly wait to encounter an antigen in peripheral lymphoid organs. As soon as a antigen-presenting cells APC assist in T-cells recognition of foreign antigen, antigen is delivered to B cells and adap-

tive immune response begins. Immediately, B cells start to form structures in peripheral lymphoid organs, called germinal centres (GCs). It was originally speculated that GCs are the main sites for the antigen-driven somatic hypermutation SHM of genes responsible for the immunoglobulin variable regions (IgV). This process causes the accumulation of the mutations in the IgV regions of the heavy and light chains of antibody [19, 20]. During SHMs, mutations tend to accumulate in complementarity determining regions () of the antibody V genes. Since SHM does not distinguish between favorable and unfavorable mutations additional selection process is required to generate the high-affinity antibodies for the humoral immune response and to remove the antibodies with a low affinity [19, 20].

Approximately one day after a naive B cell is activated by a foreign antigen, it migrates to the T-cell zone of the lymphoid tissue, where B cells encounter with the network of follicular dendritic cells (FDCs). At the second day, activated B cells form a connection with antigen-specific T cells to become fully activated. One day after, T follicular helper cells (T-fh cell) migrate to the follicle. In the same time the first defense mechanisms against pathogen start to form: some pairs of B and T cells move to specialized areas in the lymph nodes, where B cells differentiate into plasmablasts to secrete low affinity antibodies [21]. At the forth day, B cells migrate to the center of FDCs and start to proliferate, forming the early GC [22]. At days 5-6, the fast B cell proliferation increases the GC size. At day 7, the dark and light zones (DZ and LZ) form, which can be associated with the maturation of GC (see Figure 1.2) [23].

GC formation is controlled by the complex orchestra of different transcriptional pathways (see Figure 1.3). At GC initiation stage, from day 0 to day 1,

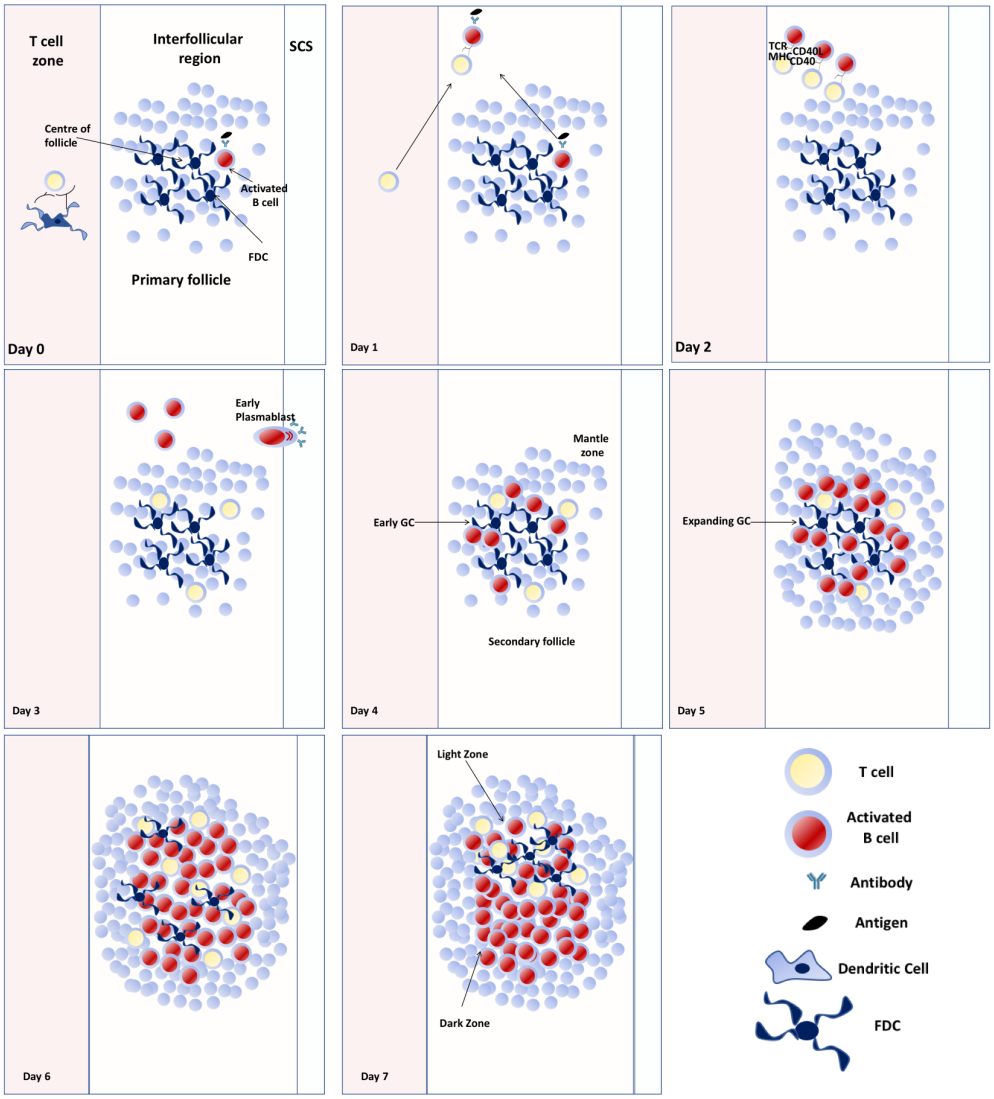


Figure 1.2. Germinal center initiation and formation in the lymph node, presented in 8 different stages. Pink area represents the T cell zone, white area corresponds to subcapsular sinus area . FDC - follicular dendritic cells. Adapted from [23].

after the B cell activation by antigen, a set of pathways responsible for correct GC formation and B cell proliferation (IRF4, NF-kB, OCA-B, MEF2C and MYC) is activated. Interferon regulatory factor 4 (IRF4) is crucial both for GC initialization and maturation (see Figure 1.3). Starting from day 2 early stage active pathways are modified, some genes are silenced, while others are upregulated, for instance: BCL6, MEF2B, IRF8. From the 5th day of the GC

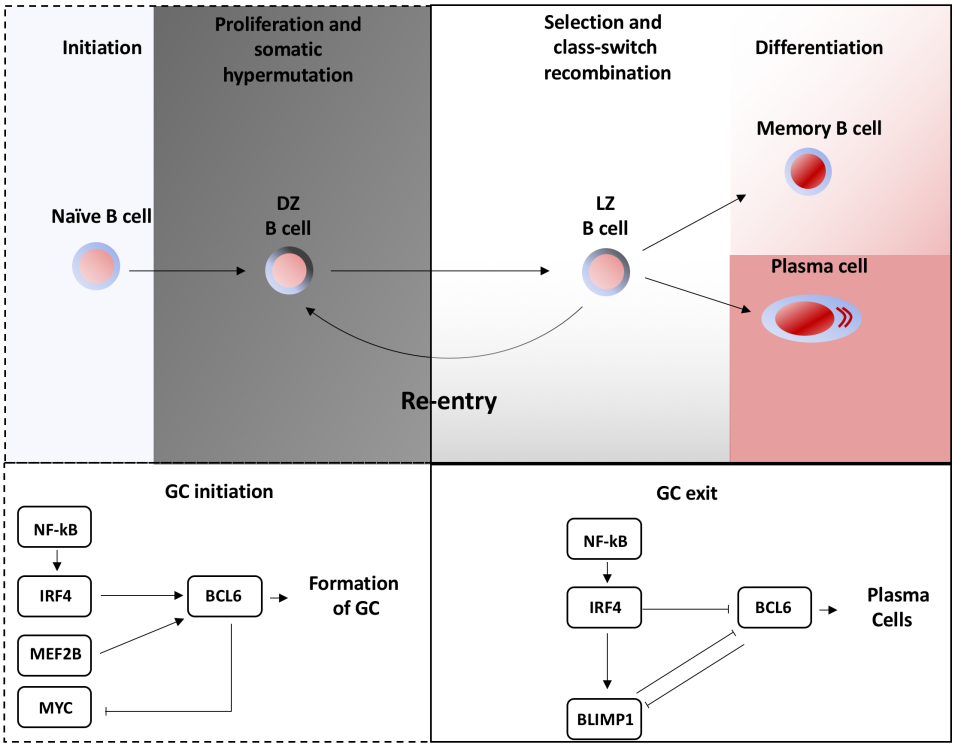


Figure 1.3. Different stages of B cell differentiation in a germinal center (top panels), combined with associated transcription pathways regulating GC initiation and GC exit (bottom panels). Adapted from [25].

formation, the EBF1, SPI-B, DOCK8 and BACH2 circuits are additionally activated to cause differentiation towards the Plasma B cells, and PAX pathway is activated to direct the cell differentiation to memory B cells (see Figure 1.3). When GC reaches the mature stage, one can distinguish between cells in the DZ area (called centrocytes) and LZ area (called centroblasts), using the immunophenotype and functional characterization [24].

B cells expansion, followed by SHM in the DZ, creates a screening list of candidates for further selection of B cells with the highest affinity for the antigen in the LZ. B cells with low affinity will be eliminated by apoptosis, while cells with the strongest affinity will re-enter the DZ, to further proliferation [24]. Selection of the B cell fate occurs due to differences in the transitional pathways

which are still to be elucidated.

1.1.4 Normal and pathological B cell responses

GC B cell development is controlled by an interconnected network of transcriptional pathways. The high complexity of such a system implies high risk of possible malfunctions. The set of cancer diseases, occurring at different stages of GC B cell development, is defined as lymphomas. Due to complexity and diversity of the field, one way of lymphoma classification can be based on stages of GC B cell differentiation at which the malignancy had occurred (see Figure 1.4).

At DZ stage, the continuous expression of MYC, combined with a number of small pathways deregulation, causes the appearance of mutated IgV sequences and its transcriptional signatures, which is called Burkitt lymphoma - aggressive and rapidly developing B non-Hodgkin's lymphoma (B-NHL) [27]. The most common B-NHL, diffuse large B cell lymphoma (DLBCL, 40% cases) generally divides in two types. First one occurs during early stage of GC B cell differentiation (in LZ of GC) it is GC B cell like DLBCL and it is caused by constant BCL6 expression (see Figure 1.4). BCL6 has a crucial role both during the B cell differentiation and lymphomagenesis by controlling the terminal differentiation of B cells [28]. BCL6 disturbance can be triggered by many regulatory mechanisms such as disruption of BCL6 autoregulation circuit or inactivation of CD40-induced IRF4 repression of BCL6 [29, 30].

A second type of DLBCL, activated B cell-like DLBCL (ABC-DLBCL), appears at the last stage, where inactivation of the master regulator of terminal B-cell differentiation, PR domain zinc finger protein 1 (BLIMP1), and activa-

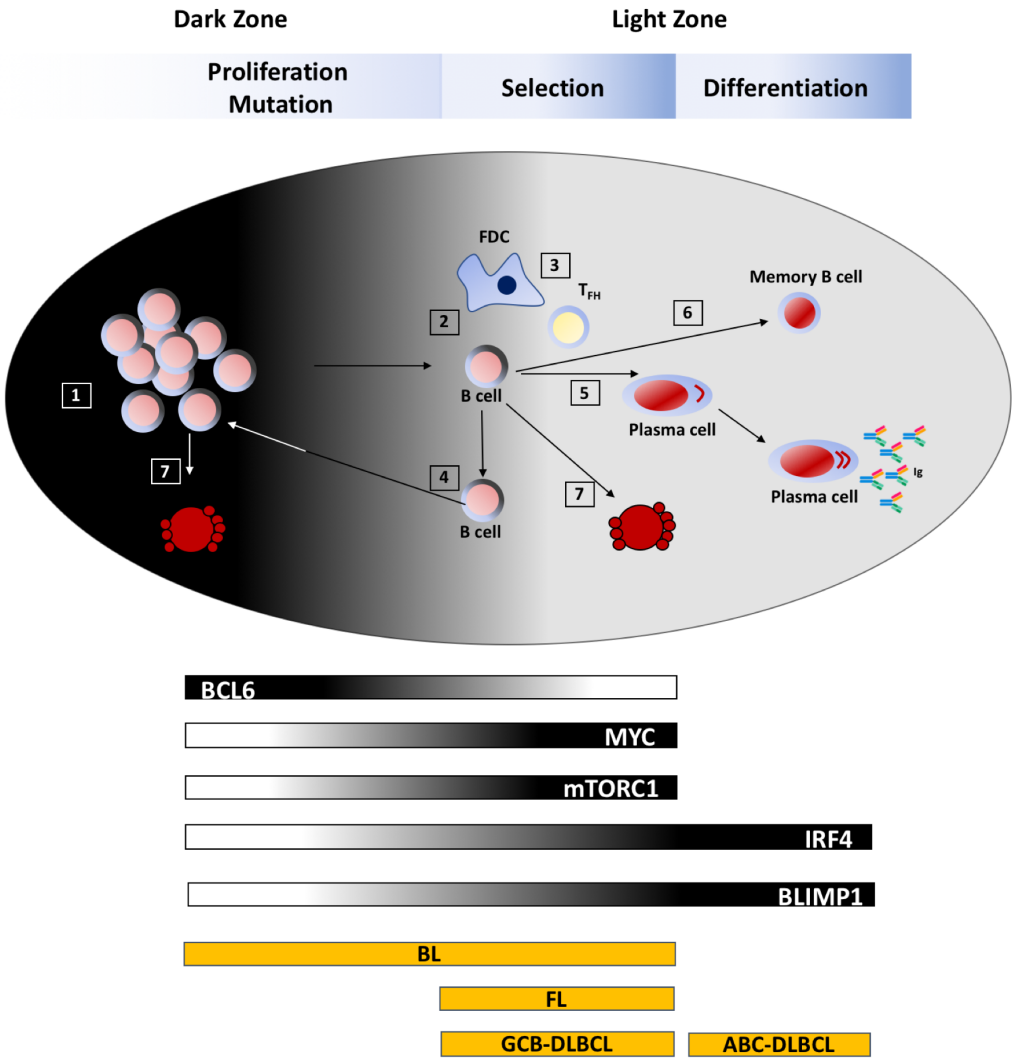


Figure 1.4. GC B cell fate maps in healthy and malignant conditions. 1). GC B cells proliferate and undergo SHM to introduce random mutation in their immunoglobulin gene in DZ. 2). B cells migrate from DZ to LZ. 3). GC cell selection in the LZ, as a function of their affinity for the antigen, via FDCs and T-fh. 4). B cells that successfully passed the positive selection are recycled back to the DZ for further mutation (via activation of MYC and mTORC1 pathways). 5). Alternative route towards plasma cells (PC), via plasmablast stage (PB). 6). B cell transformation towards memory B cell. 7). Cells, which have not been selected for either re-cycle, or PC, or PS, or had disadvantageous mutations undergo apoptosis. BL - Burkitt lymphoma. FL - Follicular lymphoma. DLBCL - diffuse large B cell lymphoma. GCB-DLBCL - GC B cell-like DLBCL. ABC-DLBCL - activated B cell-like DLBCL. Adapted from [26].

tion of NF-kB signalling pathway prevent cells from completing their differentiation processes [31, 25].

Follicle lymphoma is a malignant cancer, which develops and progress dur-

ing multiple stages of GC development. It starts with aberrant translocation affecting BCL2 during SHM at DZ and further evolves at later stages, specifically at cell selection phase [25].

It is important to mention that the number of genes and factors involved in the formation of different types of lymphomas are very extensive. We are mainly interested on the impact of three key factors (BCL6, IRF4 and BLIMP1) on the GC differentiation. This focus is explained by the key importance of those genes in a normal GC B cell development. Its aberrations can lead to different types of lymphoma, including follicular lymphoma (FL), diffuse large B-cell lymphoma germinal center B-like (GCB DLBCL) and activated B cell-like (ABC DLBCL) [26] .

1.2 Modelling of the Germinal Centre and B cell differentiation

In an era of big data, immunology faces the need to reconstruct and systematize the immune response in an integrative manner. Luckily, current stage of the mathematical and computational development is able to assist in such a challenge. Mathematical modelling enhances systematic elucidation of innate and adaptive immunity both separately and as an entire mechanism [32].

The main objectives of mathematical modelling can be described as theoretical understanding (validation of the model may be impossible due to technological limitations) and as establishment of a prediction of future biological behavior (that can be confirmed by biological experiment) [33]. With an increasing availability of computational power, many biological institutions are capable

to put more emphasis on the application of numerical methods and mathematical modelling [34].

1.2.1 Mathematical modelling in immunology

Mathematical immunology can assist on understanding in a spectrum of questions, including the elucidation of multiple molecular pathways responsible for activation, migration and death of immune cells (B cells, T cells, FDC, etc.), cancer-immune interactions and immune response on infections and many others [32, 35].

We can describe different modelling and simulation types, using abstract categories. Let's first discuss ones focusing on immune dynamics at the molecular level. Such models are mainly focused on simulation of TCR and BCR binding and diversity, responsible for adaptive immune response. Depending on the specific scientific question, researchers can use deterministic Ordinary Differential Equations (ODE), partial differential equations or stochastic equations. Many groups have applied those for deeper understanding of correct T and B cell activation. For instance, Chakraborty and Das [36] have performed a study of digital signalling and hysteresis in B and T lymphocytes during RAS activation, T cell sensitivity to antigen as a function of the positive and negative regulation and effect of spatial protein distribution on T cell activation.

Next are models generally applied for investigating the cell signalling and metabolism pathways, using systems of ODE or stochastic equations. Deregulation of such pathways (i.e. NF-kB, PI3K, IL-6, IFN-γ or glucose, glutathione, folate-mediated one carbon metabolism) causes a set of severe diseases [37]. For instance, Vodovotz et al. [38] have summarized the results of simulations

from different immune responses and inflammation events. Perley et al. [39] have presented a model capable to depicting the cross-talk between Erk, calcium, PKCθ and mTOR signaling pathway and its effect on the TCR. Modified ODE models fitted to the previously published data have suggested that the TCR signaling is regulated via SHP1 negative feedback (help to prevent hyperactivity) and CD45 (suppress possible overexpression). Perley et al. [39] also have shown that weak TCR and CD28 costimulation or reduction in CD45 activity, can be caused by elevated FOXP3 and reduced IL-2 signaling.

Going further, some models study the cell-level dynamics during immune response. Those allow to qualitatively and quantitatively estimate different cell-type balances and ratios, death and survival rates pre and post infection [40]. Cell-level dynamic models could be applied for studying either T lymphocyte dynamics (T cell turnover, T cell movement) or B cell turnover [40, 41, 42, 43]. Many models are specialized on simulating the immune response to a range of pathogens. For instance, Ankomah and Levin [44] have presented one which explores the differences in immune response during treatment of bacterial infection as a function of different doses of antibiotics. There is a range of publications describing the mathematical approach to elucidate different aspects of the immune response on the virus infection, including influenza virus [45, 46, 47], Epstein-Barr virus [48] and others [49]. Mochan et al. [50] have used an ODE model to analyse how pneumococcal bacteria populations in the lungs and blood affect mice's organisms and how the neutrophils are able to protect the host. For the analysis of the adaptive immune response, multiple mathematical models have been created to access T and B lymphocyte dynamics. Good example of stochastic model's applications have been presented by

Choo et al.[51], who mathematically evaluated memory CD8 T cells turnover's dynamics. Choo et al. [51] state that cell proliferation process has a stochastic nature. They also didn't detect direct dependency of memory CD8 T cells on the CD4f T cell population. The behavior of the Dendritic Cells DC dynamics was also successfully studied using mathematical modelling. As an example, we can refer the work of Celli et al. [52], who used a computational approach to evaluate the efficacy of T cell activation by DC cell in the lymph node. A variety of models, describing the cell-level dynamics, can be found in the literature and they open a huge opportunity for deeper understand structure of human immunity.

Finally, we will say few words about multiscale models, which are focused on the macro-scale interactions between different subtypes of immune cells [53]. More specifically, those models attempt to connect and evaluate causation of changes at the molecular level on the host organism. Multiscale models have recently been used to study cytotoxic T lymphocytes () differentiation [54], its role of immune system in wound healing [55] and the interaction with B cell during GC formation [56]. Currently, the biggest challenge is establishment and fitting of the model's parameter, due to high level of abstraction. For this reason multiscale models are normally used for qualitative, rather than quantitative purposes in immunology [32]. In the next subsection we would focus on a mathematical model, applied to GC formation.

1.2.2 Mathematical modelling of Germinal Centres

GCs are crucial structures of the adaptive immunity, which have been extensively studied during decades (see Section 1.1.3). In this subsection we will

discuss a set of mathematical models developed to assist in understanding of GC dynamics.

As we described in Section 1.1.3, SHM introduces random mutations to the variable regions of immunoglobulin, most of which are more likely reducing the affinity to the antigen. Oprea et al. [57] hypothesised that every 10 mutations there should be an intermediate selection step, which removes cells with disadvantageous mutations. Earlier, Kepler and Perelson [58] had suggested that according to control theory, B cells of GC should reenter to the re-cycle loop, approximately every 5-10 mutations. Such a mechanism should allow selection of B cells with highest affinity in a robust and iterative manner. In other words, every 5-10 mutations, B cells should pass selection process and only the best B cell candidates would follow the next iteration of SHMs. This hypothesis lined up relatively well with the compartmentalized structure of GC and led to a conjecture: SHM and proliferation occur in the DZ, while selection process occurs in the LZ [59]. Theoretical advantage of such a structure is obvious, it allows the future selection to be performed from the best candidate at each step. The percentage of positively selected B cells which re-enter to the DZ of the GC was estimated to be from 15% up to 70% [60, 61]. Quantitative modelling suggests that the reduction of B cells in GC are in a higher order of magnitude, compared to a normal apoptosis speed, which can be an evidence of positive feedback mechanism of B cell recycling. According to Meyer-Hermann et al. [62], probability of B cell recycling reaches 80%.

The mechanisms of B cells selection are still under discussion in the scientific community. At the moment it is hypothesized, that the B cell should bind to antigen-FDCs pair or to a soluble antigen [63, 64]. Mathematical models

could assist in understanding of such mechanism of selection and in hypothesis formation for future experiments. Currently the well accepted thesis was summarised by Siskind et al. [65], who suggested that selection can be caused by the amount of antigen paired to FDCs. Iber et al. [66] have succeeded to establish mathematical model based on the experimental data. Their model shows that soluble antibodies are accessible in the GC and as a consequence, increases the competition biding of antigen. Alternative hypothesis suggests that the concentration of antigen is not the critical factor for the B cell selection. Kecsmir and De Boer [67] have presented a work, which shows that the FDC-to-antigen pair is the mechanism which controls B cell selection. Additionally, Meyer-Hermann et al. [68, 69] proposed two possible scenarios: i) the B cell keeps trying to bind to FDCs until it either receives the apoptotic signal, or it is selected; and ii) B cell selection depends on the successful binding of B cell to Th-cell, i.e. the Th-cell plays the major role of selecting the B cell with higher affinity to the antigen.

Selection of appropriate model for each scientific question should be intensively discussed at all stages of experimental planing, from the experimental design, until the stage of data analysis and comparison with available public data.

In the next section we will cover the general state of the art of modelling of GC B cell differentiation in GCs .

1.2.3 ODE modelling of B cell differentiation

Martinez et al. [2] have performed a work on merging available models of B cell differentiation in GCs and on developing ODE model capable of simulating multiple types of GCs cells using an open-source published micro-array data.

Activated naive B cells during rounds of expansion and consecutive selection in GCs can have multiple exit fate from GC: either to cells producing antibodies to the specific antigen (Plasma cells, PCs) or to cells capable of storing a sequence against the antigen (memory B cells, MC) or will die via apoptosis. The exit from GCs is controlled by a small transcription network, which unites three key transcription factors: BCL6, IRF4 and BLIMP1. Briefly, BCL6 protein is responsible for formation of GCs , and its high level concentration is essential to maintenance of B cells in the centroblast stage. To urge the transition towards PCs and MCs, different mechanisms should be activated. Proteasome degradation of BCL6 via BCR [70] and T-cell-mediated stimulation through CD40 pathway are triggered for activation of NF-kB-mediated regulator of PCs development - IRF4. IRF4 represses BCL6 and also negatively regulates BLIMP1 [71]. BLIMP1 represses BCL6, which at the same time represses BLIMP1, forming a repression loop. The interconnection and relationship between the key three genes is visualised on Figure 1.5.

Martinez et al. [2] have described the kinetic of GRN orchestrating B cell differentiation in GCs , using a system of ODE. Each interaction between genes was modeled using a Hill function with cooperative coefficient 2. Martinez et al. [2] also assumed that each transcription factor has similar binding affinity, and is described as a dissociation constant k, with a maximum transcription rate σ. b, r, p are respectively the protein levels of BCL6, IRF4, BLIMP1, μ is the basal production rate of each protein, λ is the degradation rate. The model is given by the following equations:

$$\frac{dp}{dt} = \mu_p + \sigma_p \frac{k_b^2}{k_b^2 + b^2} + \sigma_p \frac{r^2}{k_r^2 + r^2} - \lambda_p p \tag{1.1}$$

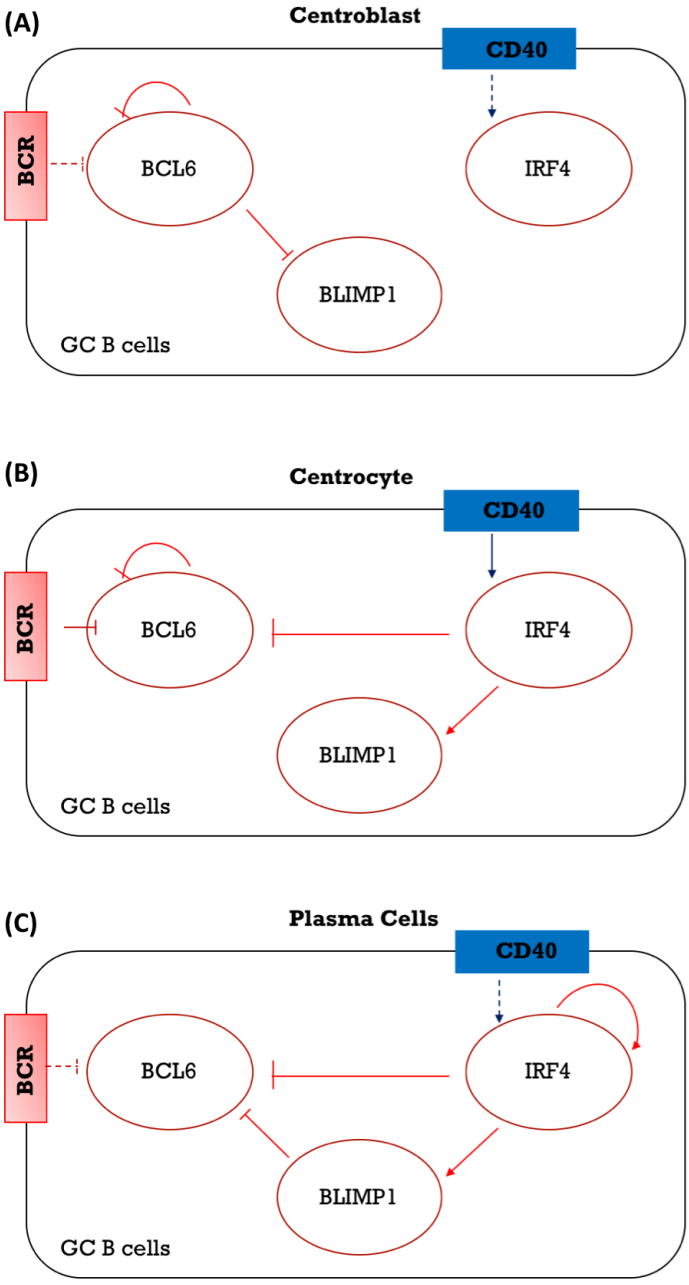


Figure 1.5. Time-dependent regulatory network of GC B cells. (A) Regulatory network at the centroblast stage. (B) At the centrocyte stage, the B cells compete for survival signals delivered by the BCRs and T cells, which lead to degradation of BCL6 protein and up-regulation of IRF4. (C) In the plasma cell stage, BLIMP1 and IRF4 are expressed and contribute to the transcriptional silencing of BCL6. The cell is kept in this stage by a positive autoloop on IRF4. Adapted from [2].

$$\frac{db}{dt} = \mu_b + \sigma_b \frac{k_p^2}{k_p^2 + p^2} * \frac{k_b^2}{k_b^2 + b^2} * \frac{k_p^2}{k_r^2 + r^2} - (\lambda_b + BCR)b \qquad (1.2)$$

$$\frac{dr}{dt} = \mu_r + \sigma_r \frac{r^2}{k_r^2 + r^2} - \lambda_r r \tag{1.3}$$

Systems (1.1)-(1.3) describes the effect of BCR and CD40 stimuli in a phenomenological form:

$$BCR = bcr_0 \frac{k_b^2}{k_b^2 + b^2} \tag{1.4}$$

and

$$CD40 = cd_0 \frac{k_b^2}{k_b^2 + b^2} \tag{1.5}$$

The authors fitted coefficients from the microarray gene expression dataset from GEO accession GSE12195 which allowed to reproduce the dynamics of normal GC B cell development from PC centroblasts towards plasma cells (see Figure 1.6). The model utilizes the hypothesis of bistability, suggesting that the B cell differentiation is irreversible, unless the inverse stimulus is applied. B cells in the centroblast stage have specific gene signatures (high BCL6, low IRF4 and BLIMP1) and as soon as the transition towards PC is finished, the gene expression pattern changes (low BCL6, high IRF4 and BLIMP1).

Martinez et al. [2] also have presented gene expression dynamics for different sets of gene deregulation of the three key GRN (which recapitulate majority of B cell lymphomas). Martinez et al. [2] have shown the importance of the elucidation of gene expression using an ODE model, applied to a GRN made of key genes. We have implemented a probabilistic model, based on the model of Martinez et al. [2] which will be described and discussed in Chapter 2.

In the following Section, we will present a state of the art of gene regulatory networks modelling.

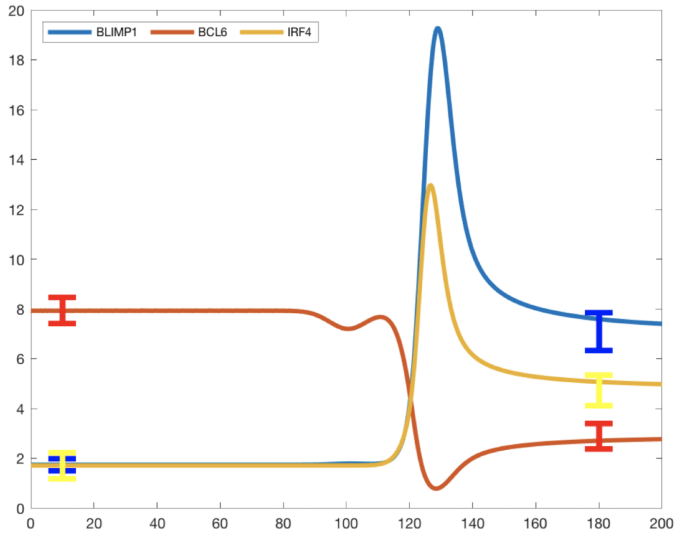


Figure 1.6. Hypothetical scenario for memory B-cell differentiation after BCR and CD40 stimulation at the GC. Simulations (lines) have been obtained with the model (1.1)-(1.3). Experimental data (bars represent mean +/- std values) GEO accession no. GSE12195 was used to fit the parameters for equations (1.1)-(1.3). Model successfully fits the experimental data and illustrates that in the first steady state the amount of BCL6 mRNA is high, while number of IRF4 and BLIMP1 mRNA molecules are low. At second steady state model confirms that the number of BCL6 mRNA molecules decreased, while the number of IRF4 and BLIMP1 increased.

1.3 Gene Regulatory Networks for comprehension of biological processes

A modern branch of research, "systems biology", mostly focused on the understanding and modelling of different parts of a living organism as a united system, rather than as a collection of independent biological features. Gene regulatory network (GRN) is a visualisation and a simulation concept, which unites a set of genes, molecules and regulations, interacting between each other directly or indirectly (through RNA and protein expression derivatives) [72]. Traditionally GRNs are represented as a directed or undirected graph, where each node corresponds either to a gene or a protein (see Figure 1.5). The edges

of the graph are considered to be molecular interactions (i.e. protein-DNA, protein-protein, indirect interaction between genes [73]).

GRNs play a key role in many biological processes, including cell differentiation, regulation of the metabolism, cell apoptosis, cell cycle, etc. For instance, each cell type and cell state can be defined by a specific set of transcription factors (TFs), which produce a specific gene expression profile, which determines a phenotype. GRN orchestrate the combination of TFs and their target genes.

In this section we will briefly discuss the applicability of GRNs concept in modern biology, today's achievements and future challenges.

1.3.1 Gene Regulatory Network inference: Aims and Methods

Intense development of the high-throughput technologies during the last two decades ended up as an explosive surge of accessible transcriptomic data, and as a consequence, lead to a higher distribution of GRN applications. Many researchers have used GRN modelling to focus on specific biological questions. For instance Basso et al. [74] showed that it is possible to reconstruct the GRN from the gene-expression profile of GC B cells, which allows to deeper characterise a set of changes occurring in B cells during different disease conditions. The authors also described a hierarchical scale-free behaviour of B cells, constructed MYC subnetwork and experimentally validated candidate MYC target [74]. Madhamshettiwar et al. [75] have applied GRN to elucidate the cross-regulation of angiogenesis-specific genes in a papillary ovarian adenocarcinoma dataset. They have defined 15 genes with distinct regulation in cancer

versus normal condition.

Generally, there are two main approaches for computational validation of GRNs . The first one is topological analysis of regulatory interactions of GRN, based on publicly available databases for proteins and genes (i.e. the Human Protein Database [76], IntAct [77], Biomolecular Interaction Network Database [78], Search Tool for the Retrieval of Interacting Genes/Proteins [79], for ChIP-seq [80]). This approach allows to evaluate different properties of a biological system, based on shape and structure of the GRN: the number of edges connected to each node, the connectivity score of each node, shortest pathways between key nodes, network motifs, etc. [81]. Alternatively, one can use different mathematical algorithms for GRNs inference based on experimental gene expression data. It allows to tune the topology of GRNs according to specific study cases and to obtain a set of optimal GRNs , which allows to fit the experimental data with the highest level of approximation [82] (see Section 1.3.2).

The final goal of GRN modelling is to simulate a network, which will enhance the comprehension of the observed phenomena. Up to now, it is still impossible to infer a unique GRN which would uniquely describes all interactions in an organism. For this reason, we should consider GRN being a "blueprint", rather than an exact representative of interconnections between all genes. Nevertheless, to our knowledge [82] there are approximately 20000 genes in Human, and advantage of GRN modelling approach is in its help to narrow the number of possible interactions and consequently enhance the effectiveness of hypothesis testing and future experimental design.

1.3.2 Bulk data versus single cell (SC) data

Nowadays, single-cell RNA sequencing (scRNA-seq) is one of the most frequently used SC technologies and can be generally grouped by SC isolation technique and capture methods for library preparation [83]. Pioneered method was developed by Tang et al. [84] back in 2009, and was lately followed by alternative and more advanced scRNA-seq methods [85].

The scRNA-seq methods are multi-steps protocols which can vary in a few aspects: cell isolation, cell lysis, reverse transcription, amplification, transcript coverage, strand specificity, UMI availability (see below). General outline of scRNA-seq experiment and data preparation can be presented ordered as: (i) isolation of SC; read alignment and expression quantification; (ii) quality control of the data; (iii) batch effect correction; (iv) scRNA-seq data normalization (v); imputation of scRNA-seq data; (vi) dimensional reduction; (vii) count table with expression value for each gene; (viii) data analysis and GRN inference (see Figure 1.7).

One of the first techniques of SC isolation, is a limiting dilution, which utilise pipette dilution to isolate cells. However this method has very low efficiency, due to the statistical distributions of cells during the sampling. Next popular method of cell isolation is based on the micromanipulation [86]. Even though it allow to select specific area of the biological object, it is time consuming and have few room for possible scalability. Recently, flow-activated cell sorting (FACS) start to gain its popularity due to multiple advantages, such as high throughput capacity, high purity of cell collection and accessibility of multiple markers as a selection criteria for sorting of different cell populations [87].

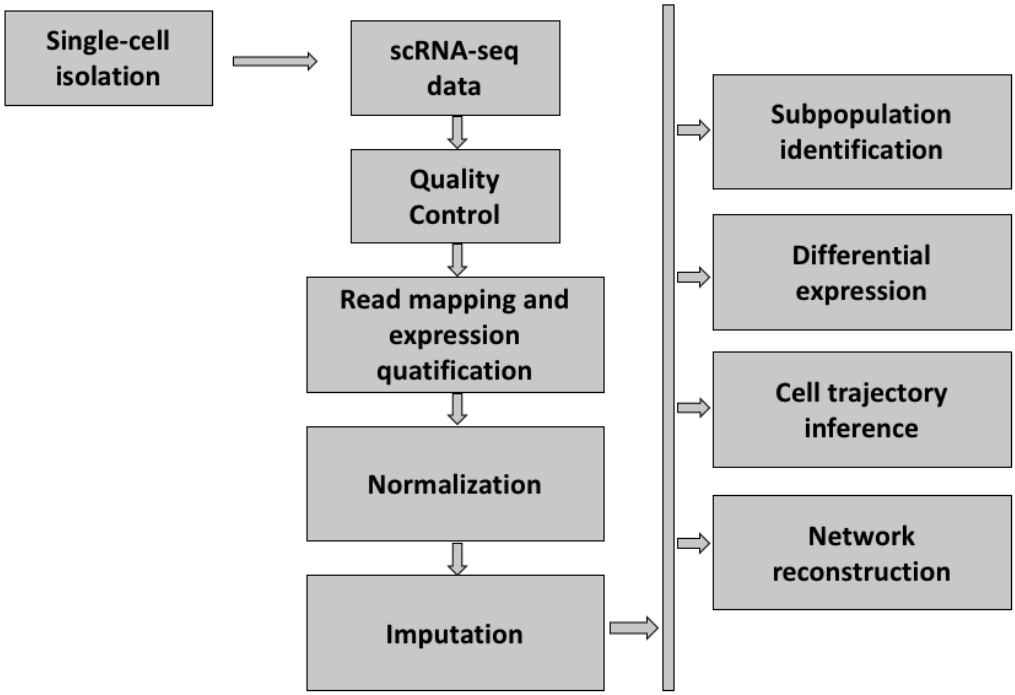


Figure 1.7. General pipeline of scRNA-seq experiment. Main steps of the scRNA-seq protocol are highlighted and adapted from [83].

Nowadays, the development of microfluidic technology allows its application in SC isolation. This technology allows to proceed and analyse the samples with reasonable number of cell number and accessible analysis cost [88, 89, 90]. Recently, the single cell isolation, via microdroplet-based microfluidic technology, starts to gain popularity [91]. One of the most developed commercial platforms, 10X Genomics, allows to perform high throughput analysis with a high efficiency [92, 93]. Next step in a SC analysis pipeline is a generation of scRNA-seq library, which should include the cell lysis, reverse transcription, second-strand synthesis and cDNA amplification. To increase the sequencing accuracy, researchers have added random 4-8bp sequences, known as unique molecular identifiers (UMIs) during the transcription stage. The usage of UMIs allows to remove PCR amplification bias, and increase the ac-

curacy by uniquely counting each mRNA molecule [94]. After the collection of the reads from scRNA-Seq, quality control step should be performed. At this stage, low-quality bases and adapter sequences should be removed. If the UMI barcoding is implemented, trimming should be performed and reads should then be aligned. Further, selection of the reads associated to exonic loci would be selected for the estimation of the gene expression matrix. Due to possible dropout or transient gene expression in SC, scRNA-seq tend to contain zero-inflated counts, which should be further normalized. There are different methods in the literature, aiming to improve the normalization step. Recently, set of different algorithms to perform normalization of scRNA-seq data, were summaries by Lytal et al. [95]. For instance, one of the normalization methods is implemented in R package's *NormalizeData* function [96]. It based on the division of the gene counts for each cell, followed by multiplying by the scalling factor and natural log transormation $log(\alpha + 1)$. There are different alternative methods of normalization, such as Single-Cell Reverse Transcription () [97], Bayesian Analysis of Single-CellSequencing Data () [98], Gamma Regression Model () [99], scran—Methods for Single-Cell RNA-SeqData Analysis [100], Robust Normalization of Single-Cell RNA-Seq Data () [101], Linear Model and Normality Based Normalizing Transformation Method (Linnorm) [102]. Existing methods have different effectiveness, different time of execution and selection of appropriate method should be done, based on accordingly to requirements of specific experiment [95]. After normalization and imputation, the final count table for each gene is created, and different data analysis can be applied, such as subpopulational identification, differentiating expression, cell trajectory, and GRN inference (see Figure 1.7). Due to the scope of the thesis,

we will mainly focus on the usage of count matrix, towards GRN inference, which we will discuss below.

First attempts to simulate GRN were performed based on bulk technologies, such as microarrays, RNA-seq, DHS-seqm, ATAC-seq and methylation-seq. One of the main characteristics of bulk dataset is that those only measure the average signal of the sample. Common methods of GRN inference from bulk data include simple correlation, regression, ODEs, mutual information, Gaussian graphical models and Bayesian approaches [103]. The bottleneck of a GRN inference from bulk data lies in the cellular heterogeneity and the need for averaging cell composition of the studied object [103]. With the development of novel SC technologies, scientists gained a tool to access the stochastic behavior of gene expression.

Unfortunately, standard algorithms of GRN inference applied to bulk data are not suitable for SC data. Recently, Chen and Mar [104] evaluated the quality of GRN inference algorithms developed for a bulk data, including GENIE3 [105], ARACNE [106] and CLR [107]. At the same time, the similarity between networks obtained by those methods did not present any significant overlap between each other. The study has shown that it is important to develop adapted algorithms for GRN inference based on SC data.

One of the possible reasons behind the insufficient quality of bulk methods for inference of SC data can be due to the unique structure of SC data. They are characterised by high sparsity, nonstandard distributions and increased data dimension [108, 1, 109].

For instance, one of the features of SC data (low amount of mRNA available for the scRNA-seq) implies that output data contains high amount of zero val-

ues. Importantly, those zero values can be a direct representation of the low number of mRNA molecules in a SC due to stochastic gene expression (biological zeros), rather than technical artefacts (by low sensibility of the method). Such property of SC data generates a range of complications for GRN inference. To overcome such issue algorithms were developed for imputation of nonzero values to zero-value datapoints. Current methods are MAGIC [110], scImpute [111], DrImpute [112], SAVER [113], ScUnif [114], PBLR [115], BISCUIT [116] and deepImpute [117]. Because its effectiveness vary, there is still no consensus in the scientific community on which algorithm is the more suitable for this role [118].

A second challenge is caused by high amount of zero values in the SC data, resulting in a nonstandard (skewed) data distribution. SC data rarely exhibit Guassian distributions and generally have a multi-modal shape. Such property violates the statistical assumptions of most of bulk inference algorithms, and consequently decreases their quality. Recent studies, performed by Pierson and Yau [119] and Rissso et al. [120], have resulted in the development of methods to analyse the zero-inflated SC data distributions (zero inflated factor analysis - ZIFA, and Zero-Inflated Negative Binomial-based Wanted Variation Extraction - ZINB-Wave).

A third difficulty is the incorporation of the data from multiple experiments into a unique set. Batch correction tools, traditionally applied for a bulk datasets, such as limma [121] and ComBat [122], have very limited performance in both simulated and real SC data [123]. Recently, significant progress on the batch integration was achieved and new algorithms were developed, such as canonical correlation analysis (CCA) [124], mnnCorrect [123], scmap

[125] and singleR [126] which are capable to improve the assembling of the data from the parallel experiments.

To dip further, a count table with expression values of each gene was generated. Different ways to infer GRN currently exist, to generate new knowledge of the network pattern. For instance, GRN can be used to simulate a specific set of TF, responsible for the state of the cell (static process), or to identify target TFs or master regulators for a transition of the cell from one state to another (dynamical process) [1]. There are different ways of performing GRN inference. One group of methods is focused on establishment of logic behind TFs combinations, responsible for the dynamical process of cell state transition from one to another and can be achieved by application of Boolean networks (Synthesis toolkit [127, 128] and BoolTraineR [129]). Others methods are focused on linking TFs to candidate target genes, with the aim to identifying specific master regulators, responsible for specific cell state. Table 1.1 summarizes different available tools for GRN inference data and specific features of each method. Due to high level of complexity of mathematical framework, we will not dive into details of each method. However we would describe highly promising inference algorithm, developed in our lab by previous group members (see Sections 1.3.4 and 1.3.5).

Model based on:	Acronym (platform):	Description:	
Co-expression	SINCERA (R)	Pipeline for analysis of scRNA-seq data which includes a prediction of key regulators for the differentially expressed genes.	Simulation of cell types/states
	ACTION (Matlab)	Pipeline for analysis of scRNA-seq data. Identifies cell types and cell-type-specific transcriptional network.	
	SCENIC (Python)	Infers trajectory and co-regulatory states.	
	nlnet (R)	Identifies gene modules based on distances. Sensitive and computationally efficient method for large matrices.	
Boolean Networks	SCNS toolkit (F#)	Builds a state-transition graph.	Simulation of dynamic processes
	BoolTraineR (R)	Infers network structure and Boolean rules, using information on trajectories through cell states.	
	SingCellNet (Matlab)	Infers regulatory circuits by integrating transcriptional patterns with the cell lineage tree.	
ODE	SCODE (R)	Simulates network dynamic. Can be used for no more than 100 TF.	
	InferenceSnapshot (C++ and Matlab)	Combines trajectory and co-expression information. Can be used for small (up to 6 genes) networks.	
PDMP	WASABI (Python)	Infers causal dynamical network from time-stamped SC data.	
Others	SINCERITIES (Matlab and R)	Infers GRNs from time-stamped SC data using regularized linear regression	
	Sinova (multiple scripts)	Pipeline for GRN inferrence using co-expression approach.	
	SCOUP (C)	Can be used for trajectory inference and identification of regulators and co-expression.	
	SCIMITAR (Python)	Infers trajectory and co-regulatory states.	
	AR1MA1-VBEM (Matlab)	Simulates Bayesian network (activation/inhibition) based on ordered cells data.	
	PIDC (Julia)	Can be sued to access co-expression based on multivariate information measures.	
	GENIE3	Tree-based method for GRN inference.	

Table 1.1. Summary of GRN inference tools from scRNA-seq, available in the literature [130, 131, 132, 133, 134, 129, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145]. Adapted from [1].

1.3.3 Piecewise Deterministic Markov Processes

The process of gene transcription has been under detailed study for a long time. Recently, different groups have published results, which testify that either in prokaryotes [146, 147, 148], or in yeast [149, 150, 151], or in higher eukaryotes [152, 153, 154], gene transcription is a discontinuous process and mainly follows the rules of bursting kinetics [155]. The stochastic nature of gene expression at single cell level was suggested to be partly responsible for the phenotype variability in organisms [156]. To visualize the transcription burst, a telegram model was introduced (see Figure 1.10), where "ON" is associated with a promoter being activated, "OFF" with promoter being in an inactivated state. Suter et al. [155] have tracked the transcription dynamics using the single-cell time-lapse bioluminescence imaging of mouse fibroblast, during the expression of a short-lived luciferase reporter gene. They have computed the switching rates between active and inactive state, k_{ON} and k_{OFF}, the stability of mRNA and protein, transcription and translation rates. Suter et al. [155] showed that there is a high gene-specificity of the bursting kinetic at a SC level, and that the gene expression is a discontinues process.

A class of models which combines deterministic dynamics and random jumps and allows to elucidate the stochastic behavior of the system is called Piecewise Deterministic Markov Processes (PDMP) [157]. Historically is was derived from Piecewise-linear (PL) Markov processes, mainly applied in studies of the queuing theory [158]. PL Markov processes were "static" and were developed to compute the expected stationary average values of certain quantities. Further development ended up with a generalization of the PL Markov model, i.e.

PDMP. It includes all possible variations of non-diffusion models and allows to formulate problems which can not be covered by PL theory. One of the reasons for popularity of the PDMP in biology is due to its continuous deterministic system nature combined with discrete random processes nature, what makes it a particularly good candidate for the simulation of the gene expression.
Recently Herbach et al. [159] proposed a model, which creates a base for GRN inference using PDMP. It was constructed from two-state model of gene expression [160] and was designed to study the dynamics of the system at the promoter, transcription and translation levels as a whole object (see Figure 1.8) and can be described by a system of equations:

$$\begin{cases} E(t) : 0 \xrightarrow{k_{ON}(P_1,...,P_n)} 1, 1 \xrightarrow{k_{OFF}(P_1,...,P_n)} 0 \\ M'(t) = s_0 E(t) - d_0 M(t) \\ P'(t) = s_1 M(t) - d_1 P(t) \end{cases} \tag{1.6}$$

In System (1.6), $k_{ON}(P_1, ..., P_n)$ is responsible for gene activation (the promoter switches from inactive to active state), $k_{OFF}(P_1, ..., P_n)$ for gene inactivation (the promoter switches from active to inactive) and both are dependent on the proteins $(P_1, ..., P_n)$. The parameter s_0 is a transcription rate, s_1 is a translation rate, d_0 is degradation rate of mRNA and d_1 is a protein degradation rate. $E(t)$, $M(t)$, $P(t)$ are the quantities of gene ("G", see Figure 1.9), mRNA ("M", see Figure 1.9) and protein ("P", see Figure 1.9) respectively at time t. The system of equations couples the stochastic dynamics of $E(t)$, with the rate equations for $M(t)$ and $P(t)$. The results of the simulation for model

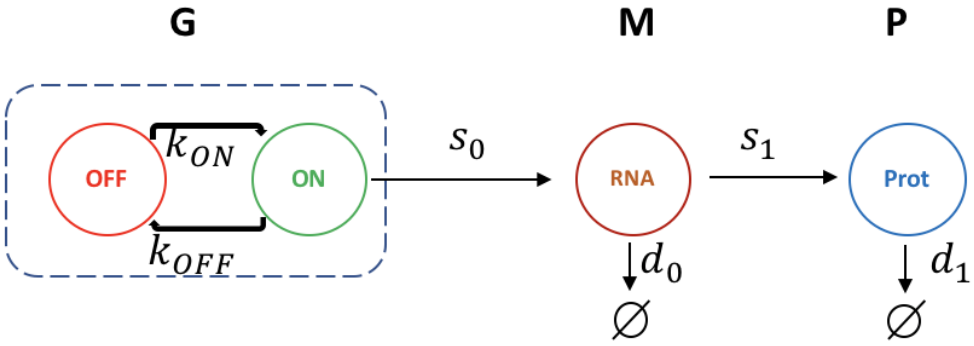


Figure 1.8. Schematic representation of the two-state model of gene expression. The promoter (G), mRNA (M), protein (P) are considered, s_0 is a transcription rate, s_1 is a translation rate, d_0 is degradation rate of mRNA and d_1 is a protein degradation rate, k_{ON} is responsible for gene activation (the promoter switches from inactive to active state), k_{OFF} for gene inactivation (the promoter switches from active to inactive), adapted from [159].

(1.6) illustrate that the PDMP model of gene expression can be used to study the stochastic behavior of the promoter activity and its effect on mRNA and protein time-dependent behaviour (see Figure 1.8).

After establishing a model for one gene, Herbach et al. [159] have expanded the system for multiple genes (see Figure 1.9). In such a construction, the promoter parameters $k_{ON,i}(P_1, ..., P_n)$, $k_{OFF,i}(P_1, ..., P_n)$ are functions of the proteins $P_1, ..., P_n$ i.e. the activity of gene i is mediated by protein levels. To describe the effect of the protein of gene i on gene j, Herbach et al. [159] have introduced the parameter $\theta_{i,j}$, which estimates the strength of activation/repression of the gene j by the protein of gene i (P_i).

To better represent the concept of coupled the PDMP model of i genes, we

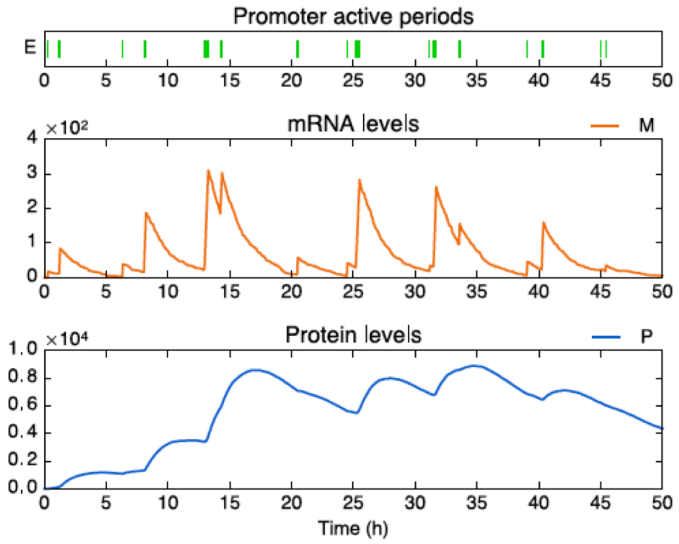


Figure 1.9. Illustrative simulation of the two-state model for a single gene. Top panel describes the activity of the promoter (green spike means the promoter is "ON", no bar means it's "OFF"). Middle panel presents the associated mRNA levels over time (h), while the bottom panel visualizes the corresponding effect on the protein level. From Herbach et al. [159].

would complement System (1.6) with the index i:

$$\begin{cases} E_i(t) : 0 \xrightarrow{k_{i,ON}(P_1,...,P_n)} 1, 1 \xrightarrow{k_{i,OFF}(P_1,...,P_n)} 0 \\ M_i'(t) = s_{0,i}E_i(t) - d_{0,i}M_i(t) \\ P_i'(t) = s_{1,i}M_i(t) - d_{1,i}P_i(t) \end{cases} \tag{1.7}$$

The correspondent interactions between genes of a coupled model are formed based on the assumptions that k_{ON} and k_{OFF} are described by following equations (1.8)-(1.9), and $s_{0,i}$ and $s_{0,j}$ performing the role of scaling constants. The estimation of k_{ON} can be formulated as:

$$k_{ON}(P_1, ..., P_n, Q) = \frac{k_{ON_{min},i} + k_{ON_{max},i}\beta_i\Phi_i(P_1, ..., P_n, Q)}{1 + \beta_i\Phi_i(P_1, ..., P_n, Q)} \tag{1.8}$$

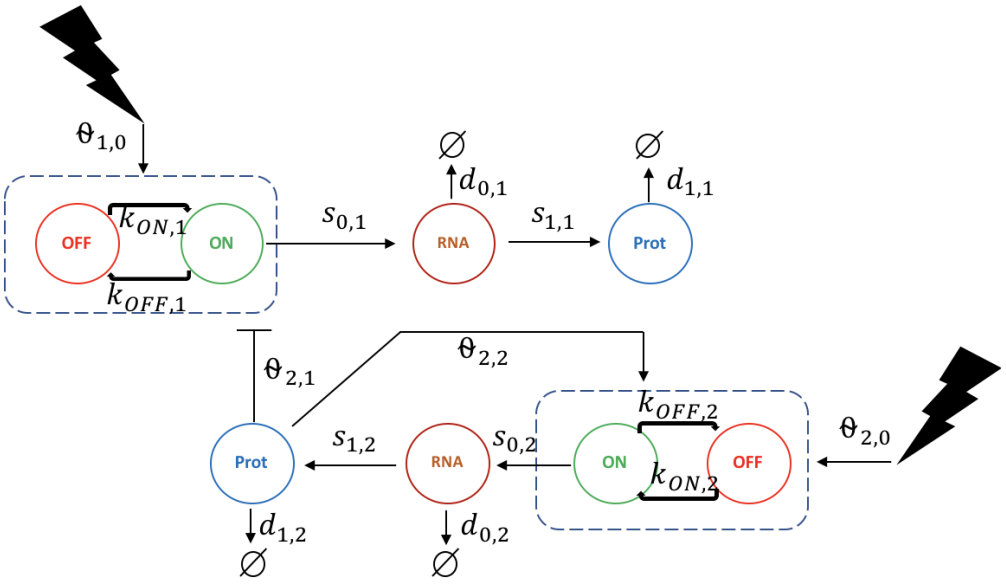


Figure 1.10. Schematic representation of coupled network of two interacting genes, 1 and 2. $k_{ON,i}$, $k_{OFF,i}$ are functions of the protein P_i and stimuli (black flashes). $\theta_{i,j}$, accounts for the influence (activation or repression) of gene i on gene j through its protein (P_j). $\theta_{1,0}$ represents the action of the stimulus on gene i. Parameters k_{ON}, k_{OFF}, s_0, s_1, d_0 and d_1 are the same than in Figure 1.9. Adapted from [159].

where

$$\Phi_i(P,Q) = \frac{1+\exp^{\theta_{i,0}} Q}{1+Q} \prod_{j=1}^{G} \frac{1+\exp^{\theta_{i,j}} (P_j/H_j)^{\gamma}}{1+(P_j/H_j)^{\gamma}} \tag{1.9}$$

and H_j is an interaction threshold of protein j, $\theta_{i,j}$ - an interaction parameter, estimated during the inference, β_i - a scaling parameter, $(P_1, ..., P_n)$- proteins, Q - a stimulus.

To summarise, Herbach et al. [159] have successfully developed a gene model in terms of the PDMP, which describes the relationship between promoter, mRNA and protein. The work performed by Herbach et al. [159] has defined a basis for a promising inference algorithm, which we will discuss in Section 1.3.5.

1.3.4 Iterative and executable framework for GRN Inference

Many different papers were recently published in the field of GRN inference, however most of them still do not present an optimal performance, giving room for search of alternative methods for GRN inference. One of the recent frameworks, WASABI (WAveS Analysis Based Inference), introduced by Bonnaffoux et al. [138] is an iterative algorithm, which relies on the model (1.7) and allows to infer executable GRN. WASABI takes advantage of growing popularity and accessibility of SC transcriptomic tools and uses time-sampled SC transcriptomic data which allow to track the dynamic of the systems, and infer network typologies with higher quality [138].

WASABI applies the concept of "waves" which considers that information provided by an external stimulus (i.e. any specific signal causing the perturbation of the system), affects genes though a cascade of network from previous to next gene. It represents the connection between the cause and its consequences, in which cause always occurs earlier (the stimulus) on the time line than the consequences (the effect of stimulus on the genes of the network are illustrated on Figure 1.12).

To visualise the relationship between signal over time, one can present the spread of the stimulus in the network (see Figure 1.12). As soon as stimulus is received by the node A (activation reaction), it starts to be detected at specific promoter wave time ($W_{prom,A,1}$) and increases the promoter activity of node A. Further the stimulus reaches the gene D (repressing reaction) what

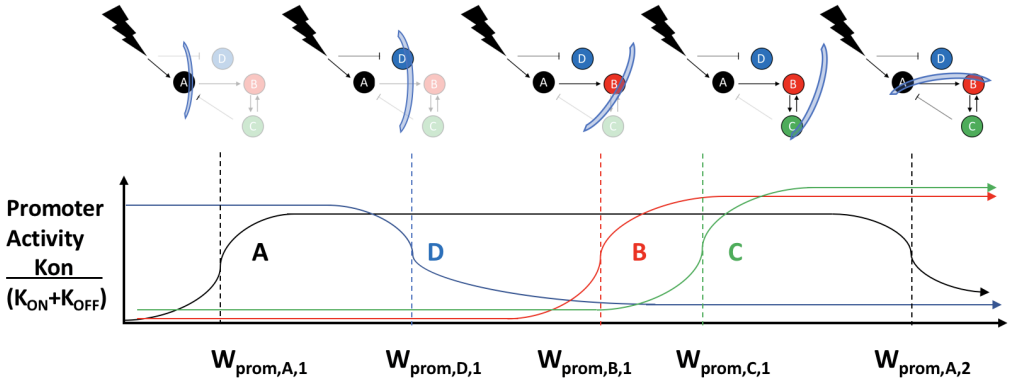


Figure 1.11. Distribution of the stimuli (black flash) and corresponding temporal promoter activity in the toy GRNs over time. Genes are presented by circles (gene A - black, gene B - red, gene C - green, gene D - blue). Repression reactions between genes are presented by T-ended lines, activation - by an arrow. Promoter wave times ($W_{prom,gene}$) correspond to the inflection point of gene promoter activity of the gene i, adapted from [138].

decreases the promoter activity of node D, which has higher promoter time ($W_{prom,D,1}$). Further, the effect of stimulus is transferred from the node A to the node B (corresponding to the time $W_{prom,B,1}$ increasing node B promoter activity, etc.). Being built on the PDMP model (see Section 1.1.3), WASABI is capable to simulate the number of mRNA and protein molecules of each node during the timeline [138].

When WASABI starts to proceed the input data, it first estimates the wave-time for each gene of the network and orders genes according to it. As soon as all candidate nodes are associated with specific wavetime, the main inference algorithm is initialised (see Figure 1.13). Briefly, at primary iteration, the first gene (the one with smallest value of wavetime, for instance, node A) forms the candidate network. At the next iteration, the next gene in an ordered list would be added to the initial gene node. For instance, at the second iteration, node D would be added no the network and all possible interactions for the current candidate network would be formed (between node A and D), and all

the simulated SC data would be collected. Next step is an estimation of the fit distances between simulated vs experimental SC data. Final step of the iteration is to select best set of GRN candidates which have the smallest fit distances and next iteration begins. The algorithm would be executed until the best set of GRNs would be estimated (see Figure 1.13).

The biggest advantage of the WASABI framework is that it incorporates time-stamped data, and allows to reconstruct the executable model. It helps to perform a set of *in-silico* simulations based on the previous generated GRNs and elucidate the candidate nodes responsible for a specific disease of interest. WASABI has a list of advantages compared to alternative inference methods. First, it allows to infer the different causalities from the time stamped data (including autoactivation and autorepression). Second, WASABI generates a range of GRN candidates, rather than a unique GRN and allows to deeper understand the network topology. Third, it has not restricted only to TFs but takes into account all types of biochemical reactions occurring in the network. Fourth, WASABI integrates the proteomic bulk data to the SC transcription data which further increase the understanding of the translation and post-translational regulation. Fifth, in WASABI is implemented a parallel computing strategy, which in combination with a wave concept, allows to significantly decrease the computational time, compared to a "brute force" approach [138].

As a practical example of an application of WASABI, Bonnaffoux et al. [138] presented results of *in-silico* validation and benchmarking of the algorithm for different "toy" networks. They applied WASABI to study the time stamped SC data (RT-qPCR) of primary chicken erythrocytes progenitor cells (T2EC)

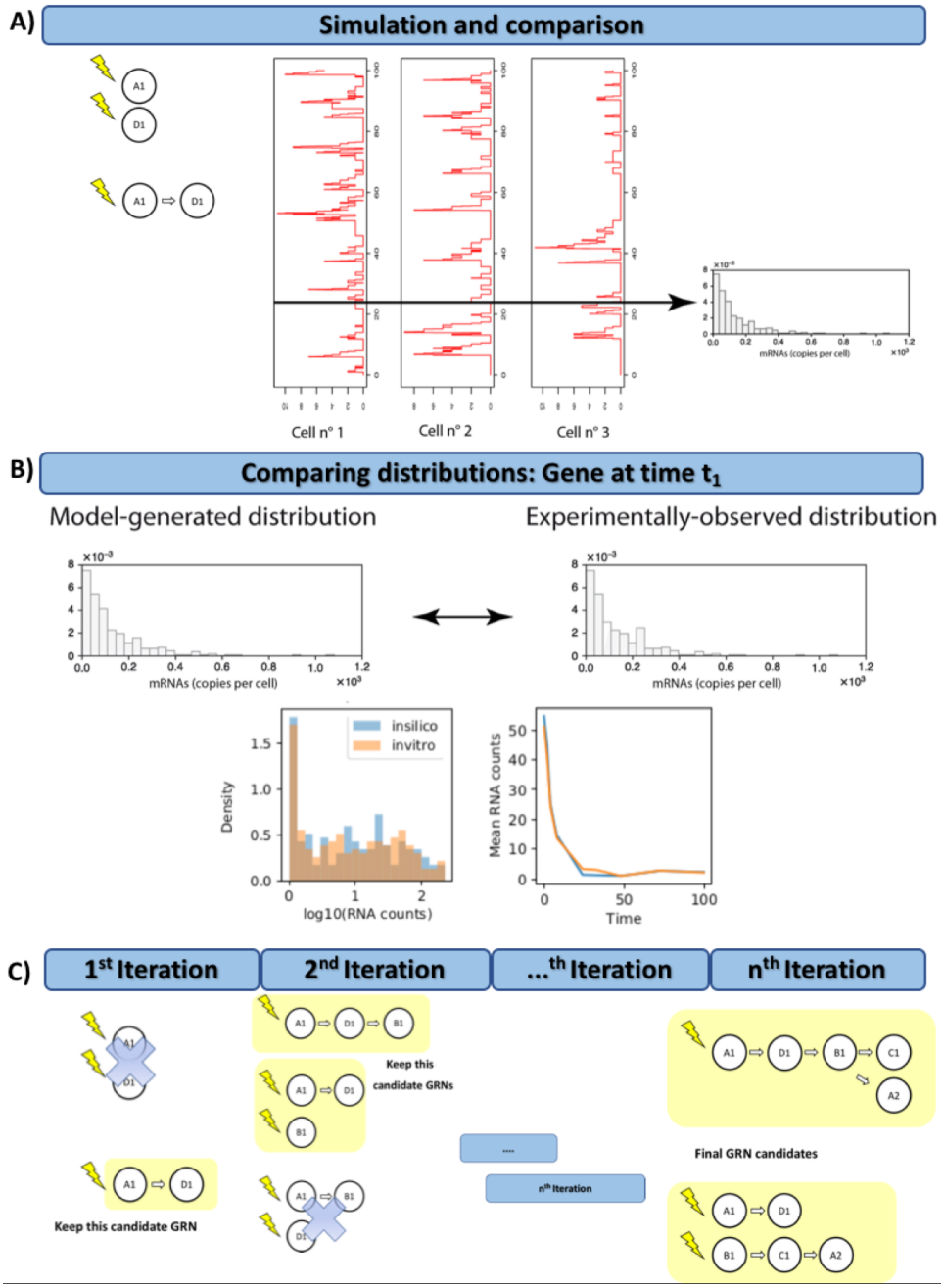


Figure 1.12. Visualization of each step of the iterative algorithm of GRN inference by WASABI. A) Simulation of the effect of the stimuli on the first candidate genes (A1, D1, and A1->D1). Panel shows the mRNA pattern for each simulated SC, and correspondent histogram of mRNA at specific timepoint (time of interest, depending on experimental setup). B) The comparison between model-generated distribution vs experimentally-observed distribution and selection of a set of GRNs with the lowest difference. C) At the n^{th} iteration the next candidate gene (accordingly to wavetime sorting) is added, and steps from A) and B) are performed again. Final selection of the set of the best candidate GRNs . Adapted from [138].

together with bulk proteomic data. WASABI has successfully inferred 364 GRN candidates. After analysis of the topology of the GRN candidates, the authors have found a list of interesting results. First, they found that the stimulus was a central regulator of GRN and most of the inferred early genes were inhibited by the stimuli. Second, none GRN candidates contained "hubs genes" affecting multiple genes in parallel. Third, the depth of the candidates was limited by 3 levels, due to the time of protein degradation (for more details, see Bonnaffoux et al. [138]).

To summarize, WASABI is a highly promising algorithm for inferring a set of candidate GRN, which allows to deeper understand the topology of network. Due to executable nature of the model, one can test different outcomes of the model, by tuning and testing different parameters of the PDMP, what potentially can assist in elucidation of specific properties of the gene-to-gene interconnections and assist in forming of hypothesis for future *in-vitro* validation. Due to the fact, that WASABI incorporates SC simulation for each GRN topology of interest, we would use the SC generation section to elucidate the candidate three-key GRNs of the human GC B cell (see Chapter 2).

1.4 Scope of current thesis

As we have discussed in current chapter, there is a constant interest to further understanding of an adaptive immune response and its critical element - GC B cell differentiation. In the next chapter, we will use coupled the PDMP (see Section 1.3.4) of the three-key gene GRN (see 1.2.3) and will apply it to human GC B cells SC RT-qPCR gene expression data (see Section 1.3.3).

We will present and discuss how one can fit the experimental SC data using coupled PDMP, what are advantages and limitations of our approach.

Chapter 2

Stochastic modelling of Gene Regulatory Network driving the B cell development in Germinal Centres

Alexey Koshkin[1,2], Ulysse Herbach[3], María Rodríguez Martínez[5], Fabien Crauste[4,*], Olivier Gandrillon[1,2,*]

1. Inria Dracula, Villeurbanne, France

2. Laboratory of Biology and Modelling of the Cell, Université de Lyon, ENS de Lyon, Université Claude Bernard, CNRS UMR 5239, INSERM U1210, 46 allée d'Italie, Site Jacques Monod, 69007 Lyon, France

3. Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France

4. Université de Paris, MAP5, CNRS, F-75006, France

5. IBM Research Zurich, Switzerland

* fabien.crauste@math.cnrs.fr * olivier.gandrillon@ens-lyon.fr

2.1 Abstract

Germinal centers (GCs) are the key histological structures of the adaptive immune system, responsible for development and selection of B cells, that produce the high-affinity antibodies against antigens. Due to their level of complexity, unexpected malfunctioning may lead to a range of pathologies, including various malignant formations. One promising way to improve understanding of the malignant transformation is to study the gene regulatory networks (GRNs). They are responsible for orchestration and regulation of gene sets in charge for the cell development and differentiation. Evaluation of the GRN structure and its inference from gene expression data is a challenging task in systems biology. Recent achievements in single cell (SC) transcriptomics allow the generation of SC gene expression data which can be used to sharpen the knowledge on GRN structure. In order to understand if GRN of three key gene regulators (BCL6, IRF4, BLIMP1), influenced by two external stimuli signals represented by cell surface receptors (BCR, CD40), can simulate the GC B cell differentiation, we applied a stochastic executable model, namely coupled Piecewise Deterministic Markov Process (PDMP), and evaluated if it can fit SC transcriptomic data from human lymphoid organ dataset. We showed that after parameter tuning, the PDMP qualitatively recapitulates the distributions of mRNA of germinal and plasmablast stages of B cell differentiation. Thus, the PDMP can assist in validation of GRN structure and, in the future, can be used to improve understanding of the different types of dysfunction of the regulatory mechanisms.

2.2 Introduction

Adaptive immune response is a complex mechanism, relying on B and T lymphocytes, which protects the organism against a range of pathogens. The crucial elements of adaptive immune response, the Germinal Centers (GC), are the structures in lymphoid organs where activated naive B cells are expanded (in a Dark Zone, DZ) and selected (in a Light Zone, LZ) and can have multiple exit fates, such as antibody producing cells (plasmablast and plasma cells, PB_PC), or cells which are capable of long term storage of the information related to the antigen (memory B cells, MC), or die via apoptosis [161, 162].

It is currently thought that B cell differentiation in GC is controlled by a small network of transcription factors (TF) constituted by B-cell lymphoma 6 (BCL6), interferon regulatory factor 4 (IRF4) and PR domain zinc finger protein 1 (BLIMP1) [2]. BCL6 is a TF controlling formation of GC, terminal differentiation of B cells and lymphomagenesis [163, 30]. BCL6 disturbance can be triggered by several mechanisms, including proteasome degradation by BCR, T-cell-mediated CD40-induced IRF4 repression of BCL6 or disruption of BCL6 autoregulation loop [29, 30]. In turn, IRF4 is a member of IRF family of TF and has critical functions for the termination of GC B cell differentiation, for immunoglobulin class switch recombination (CSR) and PC development [164]. Impairment of IRF4 expression is tightly connected with appearance of multiple malignancies [164]. In candidate GRN, IRF4 presence not only repress BCL6, but also activates BLIMP1 and is essential for the GC maturation and B cell differentiation to plasmablast. Last, but not least, BLIMP1 is a

TF, which regulates pathways responsible for B cell lineage (e.g., PAX5), for GC proliferation and metabolism (e.g., MYC) [165, 166]. BLIMP1 is involved in the induction of genes (e.g., XBP-1, ATF6, Ell2), facilitating the antibody synthesis [167, 168, 169]. In candidate GRN, BLIMP1 and BCL6 mutually repress each other.

Martinez et al. [2] developed a deterministic kinetic ODE model capable of simulating normal and malignant GC exits using a GRN based on these 3 transcription factors. For the normal differentiation of GC B cells towards PB_PC stage, the kinetic ODE model fits microarray data at two steady-states: the first one associated with the GC stage of B cell differentiation (with high levels of BCL6 and low levels of IRF4 and BLIMP1), and the second one associated with PB_PC stage (with low levels of BCL6 and high levels of IRF4 and BLIMP1).

Recently, multiple protocols which are capable of generating scRNA-seq data were developed and used to answer various questions in biology [83, 170]. At the same time, different groups showed that in eukaryotes gene transcription is a discontinuous process and follows the rules of bursting kinetics [152, 153, 154, 155]. Such results suggest that the stochastic nature of gene expression at the SC level can be partly responsible for the phenotype variation in living organisms [156]. Thus, by gaining access to a stochastic behavior of gene expression, SC may lead to further improvement of the understanding of the biological systems and their variability.

Nevertheless, the stochastic modelling of GRNs using SC gene expression data is still in its early stage [171, 172] and have never been studied for GC B cells. We have applied a class of stochastic models which combines deterministic dy-

namics and random jumps, called Piecewise Deterministic Markov Processes (PDMP) [157]. Herbach et al. [159] have constructed a two-state model of gene expression to study the system's dynamics at promoter, transcription and translation levels for a GRN structure of interest. We applied this model to the GRN of GC B cell differentiation based on three key genes, BCL6, IRF4 and BLIMP1 and simulated single B cell mRNA data [3]. We showed that the PDMP model used for BCL6-IRF4-BLIMP1 GRN can qualitatively simulate the SC mRNA patterns for normal B cell differentiation at GC and PB_PC stages.

2.3 Material, Methods and Models

2.3.1 Single cell data

For SC simulations we used the B cells dataset from human lymphoid organs published by Milpied et al. [3]. The authors have studied normal B cell subsets from germinal centers of human spleen and tonsil and have performed integrative SC analysis of gene expression. They used an adapted version of the integrative single cell analysis protocol [173]. Shortly, the authors prepared cells for flow cytometry cell sorting. Further in every 96-well plate the authors sorted three to six ten-cell samples of the same phenotype as a single cell. They performed multiplex qPCR analysis using the Biomark system (Fluidigm) with 96x96 microfluidic chips (fluidigm) and Taqmann assays (Thermofisher) [3]. They obtained results in the form of fixed fluorescence threshold to derive Ct values. We used Ct values to derive Expression threshold (Et) values:

Et= 30−Ct. When there was an unreliably low or undetected expression (Ct $>$ 30), Et was set to zero [3]. Using SC gene expression analysis of a panel of 91 preselected genes and pseudotime analysis (based on the cartesian coordinates of SC on the first and second principal components of the PCA), the authors separated GC DZ cells, GC LZ cells, memory cells and PB_PC cells.

Here we focused on three genes, BCL6, IRF4 and BLIMP1. We have selected the SC gene expression values for BCL6, IRF4 and BLIMP1 for GC DZ cells (317 SC) and for PB_PC (104 SC) (see Figure 2.5). The experimental dataset includes at the GC B cell stage 30 SC with zero BCL6 mRNA amount, 292 SC with zero IRF4 mRNA amount and 292 SC with zero BLIMP1 mRNA amount. At the same time for the experimental dataset representing the end of the B cell differentiation (the PB_PC SC), there were 25 SC with zero BCL6 mRNA amount, 79 SC with zero IRF4 and 5 SC with zero BLIMP1 mRNA amount.

2.3.2 Kinetic ODE model

Martinez et al. [2] derived an ODE model, which simulates B cell differentiation from mature GC cells towards PB_PC. Dynamics of each protein (BCL6, IRF4 and BLIMP1) are presented as a function of a production rate (μ), a degradation rate (λ), a dissociation constant (k) and a maximum transcription rate (σ). This system is given by (2.1)-(2.3), where p, b and r account for

proteins BLIMP1, BCL6 and IRF4, respectively:

$$\frac{dp}{dt} = \mu_p + \sigma_p \frac{k_b^2}{k_b^2 + b^2} + \sigma_p \frac{r^2}{k_r^2 + r^2} - \lambda_p p \tag{2.1}$$

$$\frac{db}{dt} = \mu_b + \sigma_b \frac{k_p^2}{k_p^2 + p^2} \frac{k_b^2}{k_b^2 + b^2} \frac{k_r^2}{k_r^2 + r^2} - (\lambda_b + BCR)b \tag{2.2}$$

$$\frac{dr}{dt} = \mu_r + \sigma_r \frac{r^2}{k_r^2 + r^2} + CD40 - \lambda_r r \tag{2.3}$$

In this model, CD40 and BCR act as stimuli on genes: BCR temporary represses BCL6 and CD40 temporary activates IRF4.

2.3.3 PDMP model

The coupled PDMP model, which describes the coupling between the genes i and j can be presented by the series of equations:

$$\begin{cases} E_i(t) : 0 \xrightarrow{k_{on,i}(P_1,...,P_n,Q_s)} 1, 1 \xrightarrow{k_{off,i}(P_1,...,P_n,Q_s)} 0 \\ M_i'(t) = s_{0,i} E_i(t) - d_{0,i} M_i(t) \\ P_i'(t) = s_{1,i} M_i(t) - d_{1,i} P_i(t) \end{cases} \tag{2.4}$$

where $E_i(t)$, $M_i(t)$ and $P_i(t)$ are the activation status of the promoter, the quantities of mRNA and proteins of node i respectively for $i \in \{BCL6, IRF4, BLIMP1\}$, protein $P_i \in \{P_1, ..., P_n\}$, Q_s for $s \in \{BCR, CD40\}$ is a stimuli. For each gene i, system (2.4) is defined by promoter state switching rates $k_{on,i}$ and $k_{off,i}$, by degradation rate of mRNA ($d_{0,i}$), protein degradation rate ($d_{1,i}$), transcription rate ($s_{0,i}$), translation rate ($s_{1,i}$) and the interaction parameters $\theta_{i,j}$. The interactions between genes are described based on the assumption

that $k_{on,i}(P_1, ..., P_n, Q_s)$ is a function of the proteins P_i and stimuli Q_s of the GRN given by:

$$k_{on,i}(P_1, ..., P_n, Q_s) = \frac{k_{on,i}^{min} + k_{on,i}^{max}\beta_i\Phi_i(P_1, ..., P_n, Q_s)}{1 + \beta_i\Phi_i(P_1, ..., P_n, Q_s)} \tag{2.5}$$

where

$$\Phi_i(P_1, ..., P_n, Q_s) = \prod_{s=BCR}^{s=CD40} \frac{1 + \exp^{\theta_{s,i}} Q_s}{1 + Q_s} \prod_{j=1}^{n=3} \frac{1 + \exp^{\theta_{j,i}}(P_j/H_{j,i})^\gamma}{1 + (P_j/H_{j,i})^\gamma} \tag{2.6}$$

$H_{j,i}$ represents an interaction threshold for the protein j on gene i (including stimuli BCR and CD40). $H_{j,i}$ also accounts for interactions between BCR and CD40 and the gene i (that is, with $j = BCR$ or $j = CD40$). $\theta_{j,i}$ the interaction parameter between the gene i and the protein j, β_i a scaling parameter. The effect of a stimuli Q_s on a gene i uses an interaction parameter $\theta_{s,i}$ [159].
For model (2.4)-(2.6), the promoter state evolution between time t and $t + dt$ is defined by Bernoulli distributed random variable [138]:

$$E(t + dt) = Bernoulli(pr(t)) \tag{2.7}$$

with probability $pr(t)$ depending on current $k_{on}(P_1, ..., P_n, Q_s)$ and $k_{off}(P_1, ..., P_n, Q_s)$ (for the sake of simplicity, they are denoted hereafter k_{on}, k_{off}), and promoter state at time t:

$$pr(t) = E(t)e^{-dt(k_{on}+k_{off})} + \frac{k_{on}}{k_{on} + k_{off}}\left(1 - e^{-dt(k_{on}+k_{off})}\right) \tag{2.8}$$

It follows that the mean value of the promoter equals, for every gene, $k_{on}/(k_{on}+k_{off})$. This will be used in Section 2.4.1 to reduce the PDMP model (2.4)-(2.6) to an ODE and compare it with the kinetic ODE model (2.1)-(2.3). Figure 2.1 illustrates the structure of System (2.4)-(2.6).

During the B cell differentiation in GC, B cells first receive BCR signal, through follicular dendritic cells (FDC) interaction, that represses BCL6 and then CD40, through T follicular helper (T-fh), activating IRF4. We have assumed that the BCR was acting on BCL6 from 0h until 25h, and CD40 was acting on IRF4 from 35h until 60h. Stimuli Q_s were implemented in three steps: first a linear increase ($t_{BCR} \in [0.5h; 1.5h]; t_{CD40} \in [35h; 36h]$), then a stable stimuli ($t_{BCR} \in [1.5h; 24h]; t_{CD40} \in [36h; 60h]$), finally a linear decrease ($t_{BCR} \in [24h; 25h]; t_{CD40} \in [60h; 61h]$) (see Supplementary Figure S1).

We let the system evolve in the GC stage for 500h so it can reach a so-called steady state before applying the stimuli. After the first stimulus (BCR) was applied, we further simulated the system behavior for an additional 500h. To summarize, the PDMP model (2.4)-(2.6) applied to BCL6-IRF4-BLIMP1 GRN is defined by 40 parameters, whose values are given in Tables 2.2 to 2.5.

2.3.4 Model execution on the computational center

All models were established as a part of the WASABI pipeline [138] and were implemented in Python 3.0 and PyCharm IDE. All computations have been executed using the computational center of IN2P3 (Villeurbanne/France).

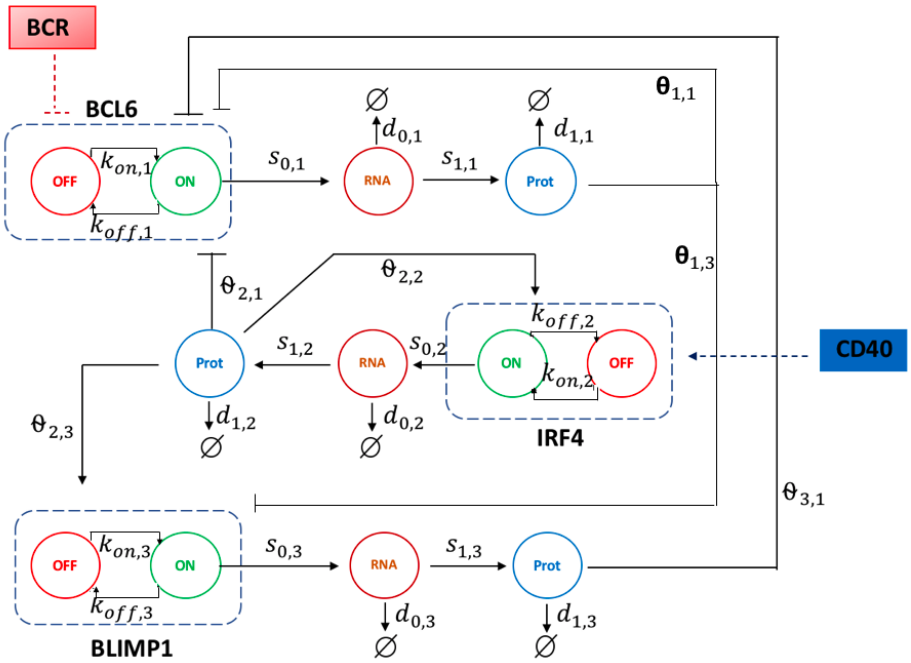


Figure 2.1. Three-genes PDMP model of the GC B-cell, consisting of BCL6, IRF4 and BLIMP1. A gene is represented by its promoter state (dashed rectangle), which can switch randomly from on to off (and vice versa), with rates $k_{on,i}$ ($k_{off,i}$). When promoter state is on, mRNA molecules are continuously produced at $s_{0,i}$ rate. Proteins are constantly translated from mRNA at $s_{1,i}$ rate. $d_{0,i}$, $d_{1,i}$ are degradation rates of mRNA and proteins. The interaction between a regulator gene j and a target gene i is defined by the dependence of $k_{on,i}$ ($k_{off,i}$) with respect to the protein level P_j of gene j and the interaction parameter between the gene i and j ($\theta_{i,j}$). IRF4 node has an autoactivation loop $\theta_{2,2}$. Additionally, two external stimuli, BCR and CD40, act on the GRN.

2.3.5 Tuning of the PDMP model

Parameters estimation for ODE reduced PDMP model

For an initial guess, we have chosen random values of the parameters for ODE reduced PDMP model (2.11), in the same order of magnitude as in Bonnaffoux et al. [138]. Further, we estimated the initial value of k_{on} for IRF4 ($k_{on,\,init,\,IRF4}$), based on the kinetic model from Martinez et al. [2]. Initial values of k_{on} for BCL6 and BLIMP1 were selected in the same order of magnitude as $k_{on,\,init,\,IRF4}$.

Estimation of the parameters for the PDMP model, by automatized

approach

After we have established the parameters for ODE reduced PDMP model (2.11), and have shown that (2.11) has two steady states, we used these values as initial guess for the PDMP model (2.4)-(2.6). Our goal was to further tune the parameters until the point when the PDMP model (2.4)-(2.6) fits the data described in Section 2.1.

We tested a possible effect of $H_{i,j}$ values, $\theta_{i,j}$ values and $k_{on,init}$ on the quality of the fitting. We tested the values of interaction threshold $H_{i,j}$ within a set $\{0.01, 0.1, 1\}$ for $i, j \in \{1, 2, 3\}$, $(i! = j)$ and interaction threshold $H_{i,j}$ within a set $\{0.0001, 0.001, 0.1, 1, 100\}$ for BCR repression stimuli on BCL6 node, CD40 activation stimuli on IRF4 and for $H_{2,2}$. We have also tested the values of interaction parameter $\theta_{i,j}$ by multiplying by a factor $f_\theta \in \{1, 5\}$ for $i, j \in \{1, 2, 3\}(i! = j)$ and interaction parameter $\theta_{i,j}$ by multiplying by the factor $f_\theta \in \{1, 10\}$ for BCR repression stimuli on BCL6 node, CD40 activation stimuli on IRF4 and for $\theta_{2,2}$ (IRF4 autoactivation loop). This generated approximately 8×10^6 combinations of parameters. During this automatized tuning procedure, we selected a parameter set that allows the system to provide the best fit of the experimental mRNA values for BCL6, IRF4 and BLIMP1 at the GC stage, based on a quality-of-fit criterion. This criterion was defined as a comparison between the average model-derived values (Υ) versus the average experimental data (Ω), with an objective function (OF) for set of genes $G = \{BCL6, IRF4, BLIMP1\}$ and stages $ST = \{GC, PB\_PC\}$

$$OF = \sum_{\delta'=1}^{|G|} \sum_{\delta''=1}^{|ST|} \left| \frac{\Omega_{\delta',\delta''} - \Upsilon_{\delta',\delta''}}{\Omega_{\delta',\delta''}} \right| \tag{2.9}$$

to minimize for parameter set (PS) of 40 parameters from Tables 2.2-2.5, the quality of fitting criterion is:

$$\min_{<PS>} OF \tag{2.10}$$

Semi-manual tuning of PDMP model

Further, we performed a semi-manual tuning of the parameters of the PDMP model to improve the quality of the fitting. We tested the values of the candidate parameters in an interval of interest, and fixed the rest of the parameters (each simulation was performed for 200 SC). Then we selected the values of the parameter of interest, which provided the best qualitative fitting (2.10) of the experimental SC data. All the ranges of tested values are summarised in Table 2.1.

Parameter	Definition	Tested values	Selected value
θ_{11}	Interaction parameter	$[-200; -10^{-2}]$	-0.2
θ_{21}	Interaction parameter	$[-200; -10^{-2}]$	-50
θ_{31}	Interaction parameter	$[-200; -10^{-2}]$	-0.5
θ_{22}	Interaction parameter	$[0.1; 200]$	11
θ_{13}	Interaction parameter	$[-200; -0.1]$	-1
θ_{23}	Interaction parameter	$[0.1; 200]$	50
$\theta_{BCR,1}$	Interaction parameter	$[0.1; 200]$	200
$\theta_{CD40,2}$	Interaction parameter	$[0.1; 200]$	10
$s_{0,BCL6}$	Transcription rate	$[0.1; 625]$	100
$s_{0,IRF4}$	Transcription rate	$[0.1; 625]$	2.1
$s_{0,BLIMP1}$	Transcription rate	$[0.1; 625]$	100
$d_{0,BCL6}$	Degradation rate of mRNA	$[10^{-3}; 10]$	0.05
$d_{0,IRF4}$	Degradation rate of mRNA	$[10^{-3}; 10]$	0.05
$d_{0,BLIMP1}$	Degradation rate of mRNA	$[10^{-3}; 10]$	0.007
$s_{1,BCL6}$	Translation rate	$[1; 1000]$	100
$s_{1,IRF4}$	Translation rate	$[1; 1000]$	160
$s_{1,BLIMP1}$	Translation rate	$[1; 1000]$	40
$d_{1,BCL6}$	Degradation rate of protein	$[0.1; 10]$	0.138
$d_{1,IRF4}$	Degradation rate of protein	$[0.1; 10]$	0.173
$d_{1,BLIMP1}$	Degradation rate of protein	$[0.1; 10]$	0.173
$k_{on,\,init,\,BCL6}$	Initial value of $k_{on,\,BCL6}$	$[10^{-5}; 10]$	0.15
$k_{on,\,init,\,IRF4}$	Initial value of $k_{on,\,IRF4}$	$[10^{-5}; 10]$	0.007
$k_{on,\,init,\,BLIMP1}$	Initial value of $k_{on,\,BLIMP1}$	$[10^{-5}; 10]$	0.001

Table 2.1. Parameters tested during the semi-manual tuning of the PDMP model.

2.3.6 Evaluation of model variability using Kantorovich Distance

To compare distributions and to evaluate model variability, we have used the Kantorovich distance (KD), which was defined by Baba et al. [174] and implemented to Python 3.0 by Bonnaffoux et al. [138]. Consider two discrete distributions p and q, defined on n bins of equal sizes, and denote t_i the center of the i-th bin. Then the Kantorovitch Distance (KD) between p and q is given by

$$KD = \sum_{i=1}^{n} \left| \sum_{j=1}^{i} p(t_j) - \sum_{j=1}^{i} q(t_j) \right|$$

We used KD to evaluate the variability of each SC mRNA generation by the PDMP model, by calculating the KD value for each gene for multiple (n=200) independent simulations.

2.4 Results

2.4.1 ODE reduced PDMP model

Application of GRN to study GC B cells differentiation is a recent approach to gain deeper understanding of the regulatory mechanisms controlling the GC B cell exit fates. In [2], Martinez et al. applied the kinetic ODE model (2.1)-(2.3) to the BCL6-IRF4-BLIMP1 GRN network of GC B cell differentiation and successfully simulated the GC B cell dynamics based on microarray data (GSE12195). Before application of the PDMP model (2.4)-(2.6), we de-

cided to test if the ODE reduced PDMP model (2.11) is able to have a similar dynamic compared to the previously studied ODE kinetic model. Since the kinetic ODE model (2.1)-(2.3) is a deterministic model, we need to simplify the PDMP model (2.4)-(2.6) to perform a comparison. To reduce the PDMP model (2.4)-(2.6) to an ODE and compare it with (2.1)-(2.3), we applied a simplifying assumption and substituted the stochastic process $E(t)$ by its mean value $\langle E(t)\rangle$:

$$\begin{cases} \langle E(t)\rangle = \dfrac{k_{on}(t)}{k_{on}(t)+k_{off}(t)} \\ \frac{dM}{dt} = s_0\langle E(t)\rangle - d_0 M(t) \\ \frac{dP}{dt} = s_1 M(t) - d_1 P(t) \end{cases} \tag{2.11}$$

Comparing mathematical formulas of systems (2.1)-(2.3) and (2.11) one can see that it is possible to establish an initial value of the promoter state $E(t)$ for IRF4 gene in system (2.11) which will correspond to GC differentiation stage (see Supplementary Material SM1). After rewriting System (2.11) in terms of System (2.1)-(2.3), we obtained the candidate value of $k_{on,\,init,\,IRF4} = 1.7 \times 10^{-3}$. Applying this value of $k_{on,\,init,\,IRF4}$, ODE reduced PDMP System (2.11) successfully simulated two steady states of IRF4, i.e. successfully recapitulates the qualitative dynamics (see Figure 2.2).

Before application of BCR and CD40 stimuli, the system is first at a steady state (simulating GC B cell stage) that corresponds to the low amount of IRF4 and BLIMP1 and a high amount of BCL6 mRNA molecules. After application of both stimuli, the system has transitioned to a second steady state that corresponds to a high number of IRF4 and BLIMP1 mRNA molecules and

Parameter	Version I, II, III
H_{12}	1
H_{32}	1
H_{33}	1
θ_{11}	-0.2
θ_{12}	0
θ_{32}	0
θ_{13}	-1
θ_{33}	0
$d_{0,BCL6}$	0.05
$d_{0,IRF4}$	0.05
$s_{1,BCL6}$	100
$s_{1,IRF4}$	160
$s_{1,BLIMP1}$	40
$d_{1,BCL6}$	0.138
$d_{1,IRF4}$	0.173
$d_{1,BLIMP1}$	0.173
$k_{off,\,init,\,BCL6}$	1
$k_{off,\,init,\,IRF4}$	1
$k_{off,\,init,\,BLIMP1}$	1

Table 2.2. Parameter set for the PDMP model (2.4)-(2.6) and ODE reduced PDMP (2.11). Version I - initial parameter set. Version II - parameter set obtained from the automatized approach. Version III - parameter set obtained after semi-manual tuning. Parameters are defined in the text.

Parameter	Version I	Version II	Version III
H_{11}	1	0.001	0.1
H_{13}	0.1	1	0.01
$H_{BCR,1}$	0.01	1	0.001
$H_{CD40,2}$	1	0.001	1
θ_{21}	-10	-100	-50
θ_{31}	-2	-20	-0.5
θ_{22}	8	5	11
$\theta_{BCR,1}$	-200	-20	-200
$\theta_{CD40,2}$	10	40	10
$s_{0,IRF4}$	2	1	2.1
$s_{0,BLIMP1}$	6.5	1	100

Table 2.3. Parameter set for the PDMP model (2.4)-(2.6) and ODE reduced PDMP (2.11) system, presented parameters are different between all versions. Version I - initial parameter set. Version II - parameter set obtained from the automatized approach. Version III - parameter set obtained after semi-manual tuning. Parameters are defined in the text.

Parameter	Version I, II	Version III
H_{22}	0.01	0.001
H_{23}	0.001	0.1
θ_{23}	40	50
$d_{0,BLIMP1}$	0.1733	0.007

Table 2.4. Parameter set for the PDMP model (2.4)-(2.6) and ODE reduced PDMP (2.11), presented parameters equal between version I and II. Version I - initial parameter set. Version II - parameter set obtained from the automatized approach. Version III - parameter set obtained after semi-manual tuning. Parameters are defined in the text.

Parameter	Version I	Version II, III
H_{21}	0.1	0.01
H_{31}	1	0.01
$s_{0,BCL6}$	6.5	100
$k_{on,\,init,\,BCL6}$	0.1	0.15
$k_{on,\,init,\,IRF4}$	0.0017	0.007
$k_{on,\,init,\,BLIMP1}$	0.1	0.001

Table 2.5. Parameter set for the PDMP model (2.4)-(2.6) and ODE reduced PDMP (2.11), presented parameters equal between version II and III. Version I - initial parameter set. Version II - parameter set obtained from the automatized approach. Version III - parameter set obtained after semi-manual tuning. Parameters are defined in the text.

low number of BCL6 mRNA molecules. However, for the current parameter set (see Tables 2.2-2.5, version I), model (2.11) underestimates the amount of IRF4 mRNA at both steady states (see Figure 2.2).

The dynamics of System (2.11) shows the existence of two steady-states for the parameter set from Tables 2.2-2.5, version I. Noticeably, if we test a random value of $k_{on,\,init,\,IRF4}$ in combination with the parameters from Tables 2.2-2.5, version I (see Supplementary Table S2), System (2.11) has only one steady-state (see Supplementary Figure S2). To our knowledge, there may be more than one set of parameter values associated with two steady states of the System (2.11).

We showed that for the parameter set from Tables 2.2-2.5, version I, the ODE reduced PDMP model (2.11) is capable to qualitatively recapitulate behavior of GC B cell differentiation GRN (see Figure 2.2). Next we wanted to understand

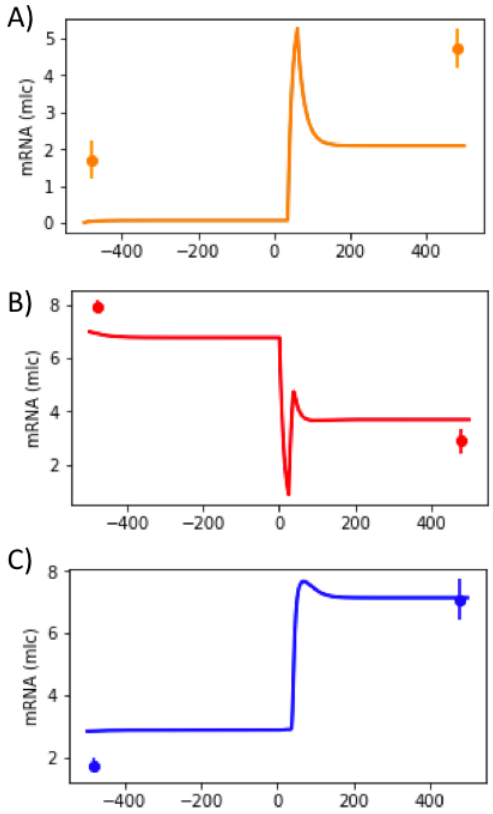


Figure 2.2. The behavior of the ODE reduced PDMP model for the nodes IRF4 (A), BCL6 (B) and BLIMP1 (C) of GRN (see Figure 2.1). BCR stimuli was applied from 0h until 25h and CD40 stimuli from 35h until 60h. The microarray gene expression dataset from GEO accession no. GSE12195 was used to estimate required parameters (see Equations (2.1)-(2.3) and Tables 2.2-2.5, version I) and are shown as dots.

if the PDMP system (2.4)-(2.6) can fit the experimental SC data (described in Section 2.3.1).

2.4.2 PDMP model applied to quantitative modelling of B cell differentiation

2.4.2.1 Accessing the variability of the PDMP model

Due to the stochastic nature of the PDMP model (2.4)-(2.6), we should first evaluate the variability of the model-generated SC data.

When one executes the PDMP model (2.4)-(2.6) multiple times for the same parameter set (Tables 2.2-2.5, version I) the resulting model-derived distributions are not exactly the same due to stochasticity of the model. We studied how strongly shapes of distributions of simulated SC mRNA molecules vary for the different executions of model (2.4)-(2.6).

For this purpose, we evaluated the level of variability of model (2.4)-(2.6) using the Kantorovich Distance (see Section 2.3.6). We performed 200 independent simulations of model (2.4)-(2.6) with a fixed parameter set (see Tables 2.2-2.5, version I). We estimated the KD between pairs of model outputs, and obtained a distribution of all KD that we call the model-to-model (m-t-m) distribution. The shape of m-t-m distributions were different between conditions and stages of differentiation. For instance, for BLIMP1, long tails were observed (see Figure 2.3). The reasonable question arises: what is the shape of distributions,

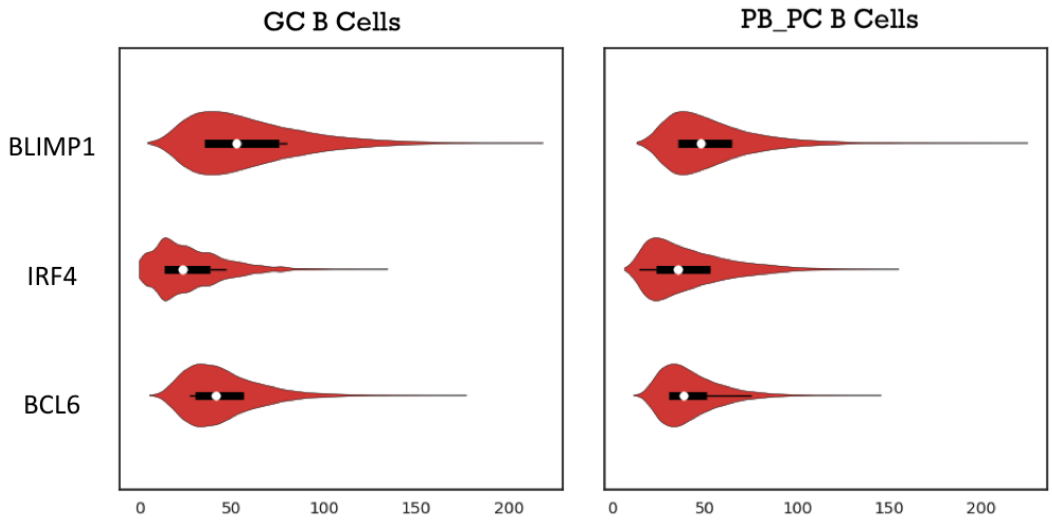


Figure 2.3. Model-to-model distributions for GC and PB_PC stages and the three genes, BCL6, IRF4, BLIMP1. Model (2.4)-(2.6) has been used, with parameter values from Tables 2.2-2.5 (version I). Results are displayed for both GC and PB_PC stages, and for the three genes. It shows the shape of distribution, median value, interquartile range and 1.5x interquartile range of the KD.

with the highest KD between each other, i.e. the distributions with the largest differences according to KD. To answer this question, we plotted distributions of the number of mRNA molecules for each node with the highest m-t-m KD at GC and PB_PC stages (see Figure 2.4). Qualitatively, we did not detect differences in the shapes of model-generated distributions with the highest KD between each other. For all 6 nodes (BCL6 GC, IRF4 GC, BLIMP1 GC, BCL6 PB_PC, IRF4 PB_PC, BLIMP1 PB_PC), the shapes of distributions were remarkably similar.

These results suggest that it may be sufficient to perform parameter tuning of the PDMP System (2.4)-(2.6), using only one simulation run for each unique parameter set.

2.4.2.2 Automatized approach

Then, we aimed to perform the search for the parameter set of the PDMP model (2.4)-(2.6) which would fit the experimental data. We first applied the strategy of straightforward parallel computation to find an optimal parameter set describing the experimental data [3].

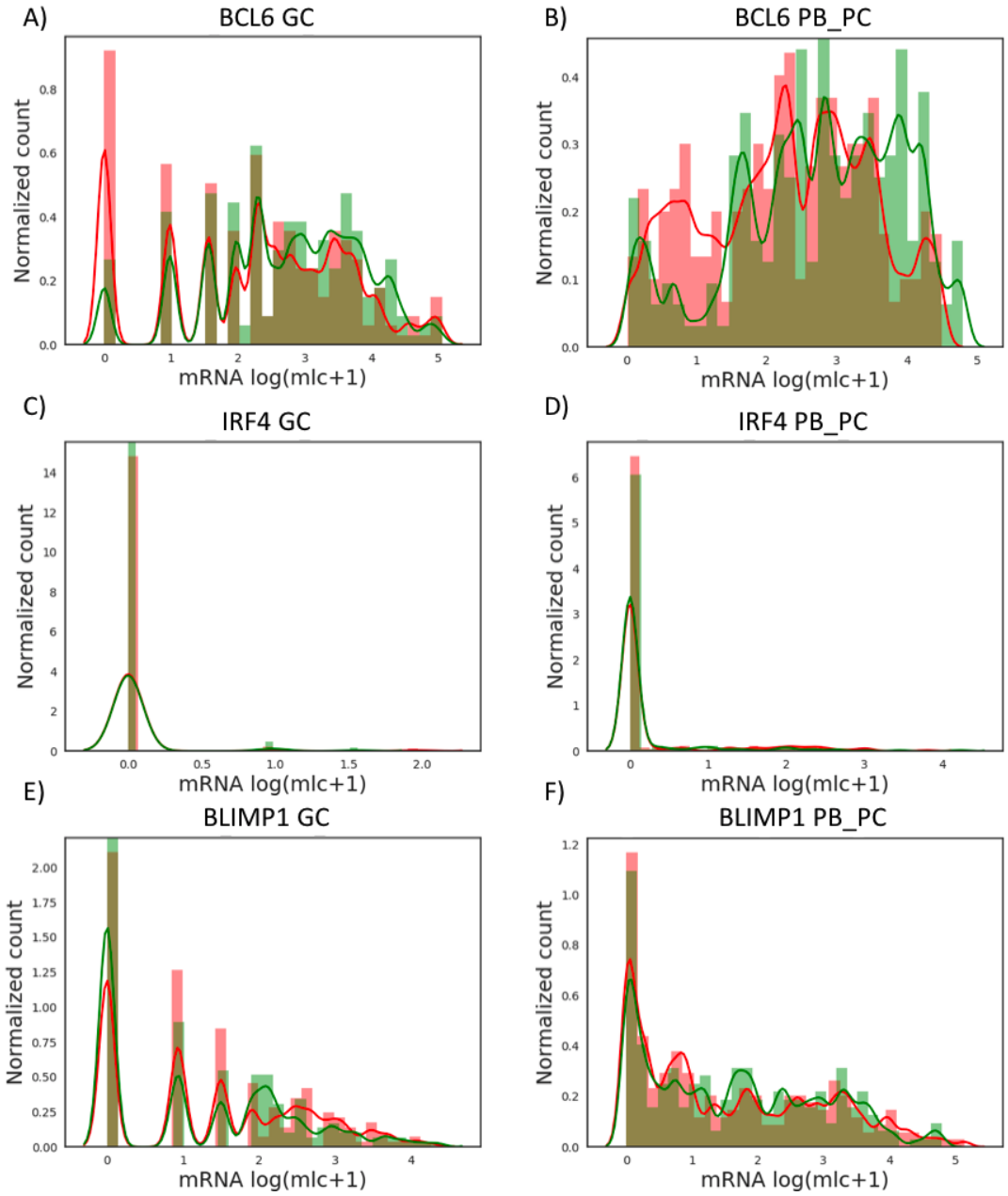


Figure 2.4. The Kernel Density Estimate (KDE) plot and histograms of two model-generated mRNA counts of BCL6, IRF4 and BLIMP1 at GC and PB_PC stages with highest KD. The subgraphs A, C, E represent log2 (molecule+1) normalized for BCL6, IRF4 and BLIMP1 compared between two models with highest KD between two of them at GC stage. The subgraphs B, D, F represent log (molecule+1) normalized SC with BCL6, IRF4 and BLIMP1 compared between two models with highest KD between two of them at PB_PC stage (Parameters from Tables 2.2-2.5, version I). Simulation of 200 SC were used based on the parameter set from ODE reduced System (2.4)-(2.6) (see Tables 2.2-2.5, version I)

We tested approximately 8×10^6 combinations of parameter sets (see Section 2.3.5), then selected the best candidate based on the quality of BCL6, IRF4 and BLIMP1 fitting at the GC stage (see equations (2.9)-(2.10)).

The PDMP System (2.4)-(2.6) estimates the number of mRNA molecules in a similar range of magnitude as in the experimental SC B cell dataset (see Supplementary Figure S4). However, the parameter set candidate (see Tables 2.2-2.5, version II) generates model derived mRNA distributions which have a sufficient overlap with experimental data for GC stage and insufficient overlap for PB_PC stage (see Supplementary Figure S4). Distributions of the number of mRNA molecules at PB_PC stage mostly underestimate the experimental SC data (see Supplementary Figure S4B, D and F).

The automatized approach helped us to establish a parameter set which allows the System (2.4)-(2.6) to correctly estimate the number of mRNA molecules for 3 out of 6 nodes. However, to perform more directed and sensitive tuning of the parameter set, we decided to further use a set of semi-automatized tests (described in Section 2.3.5).

2.4.2.3 Semi-manual approach

In order to understand which parameters should be semi-manually tested, it is important to focus on the GRN properties (see Figure 2.1) and which genes can be responsible for possible imbalances. Thanks to the structure of BCL6-IRF4-BLIMP1 GRN, where IRF4 activates BLIMP1 and represses BCL6, and where IRF4 also has an autoactivation loop, we can hypothesise that model (2.4)-(2.6) underestimates the experimental SC data due to low values of the parameters responsible for IRF4 autoactivation (θ_{22}), BCL6 repression by IRF4

(θ_{21}) and BLIMP1 activation by IRF4 (θ_{23}).

Indeed, if IRF4 autoactivation (θ_{22}) reaction is not efficient enough, there are no sufficient number of IRF4 molecules to affect BCL6 and BLIMP1 at PB_PC stage. Because IRF4 is only affected by an autoactivation loop, we started by modulating the parameter related to this reaction. During the preliminary tests we have noticed that this reaction is crucial for the cell state transition from GC towards PB_PC and that when θ_{22} and $s_{0,IRF4}$ have low absolute values then the system cannot reach PB_PC stage, even after application of the stimuli. It can be explained by the insufficient amount of IRF4 molecules produced during the simulation run (see Supplementary Figure S4C and S4D). On the other hand, when parameters θ_{22} and $s_{0,IRF4}$ have high values, model (2.4)-(2.6) transitions from GC towards PB_PC stage even before application of stimuli. This may occur due to the high activity of autoactivation reaction, which shifts the system to the second state, independently of the existence of the stimuli. After comparison of the PDMP System (2.4)-(2.6) outputs for a range of different θ_{22} and $s_{0,IRF4}$ values (described in Table 2.1), we selected the parameter set for which model (2.4)-(2.6) fits the experimental SC data at IRF4 both at GC and PB_PC stages. Such model-derived SC pattern is obtained using the values $(\theta_{22}; s_{0,IRF4}) = (11; 2.1)$. We additional performed simulations to improve the quality of the fitting of BLIMP1 and BCL6 distributions by tested parameters which are directly responsible for the balance between BLIMP1 and BCL6, such as interaction parameters (θ_{13}, θ_{31}, θ_{23}). We also tested parameters which can influence BCL6 and BLIMP1 indirectly, such as transcription rates ($s_{0,BCL6}$, $s_{0,BLIMP1}$), and degradation rates of mRNA ($d_{0,BCL6}$, $d_{0,IRF4}$, $d_{0,BLIMP1}$).

After comparison of the PDMP System (2.4)-(2.6) outputs, we selected the parameters which allow the model to have a qualitative fit of the experimental data for all nodes at GC and PB_PC stages (see Figure 2.5, and Tables 2.2-2.5, version III). For this tuned parameter set, we see that the PDMP model (2.4)-(2.6) can have good qualitative fitting of experimental data for all nodes. Results also show that for this parameter set (version III), the PDMP model (2.4)-(2.6) fits SC data at the GC stage for BCL6 (see Figure 2.5A). Model-derived distributions of BLIMP1 were capable of showing overlap with the experimental data at the PB_PC stage (see Figure 2.5F), but it overestimated the number of BLIMP1 mRNA molecules at the GC stage (see Figure 2.5E). Current parameter set (Tables 2.2-2.5, version III) has difficulties to correctly evaluate the number of zero values. The PDMP model (2.4)-(2.6) tends to overestimate the number of BCL6 mRNA molecules at PB_PC stage, as well as the number of IRF4 mRNA molecules at GC stage and number of BLIMP1 mRNA molecules at GC stage (see Figure 2.5). Nevertheless, this parameter set allowed the model to generate SC with the amount of mRNA in a similar level of magnitude as the experimentally assessed values.

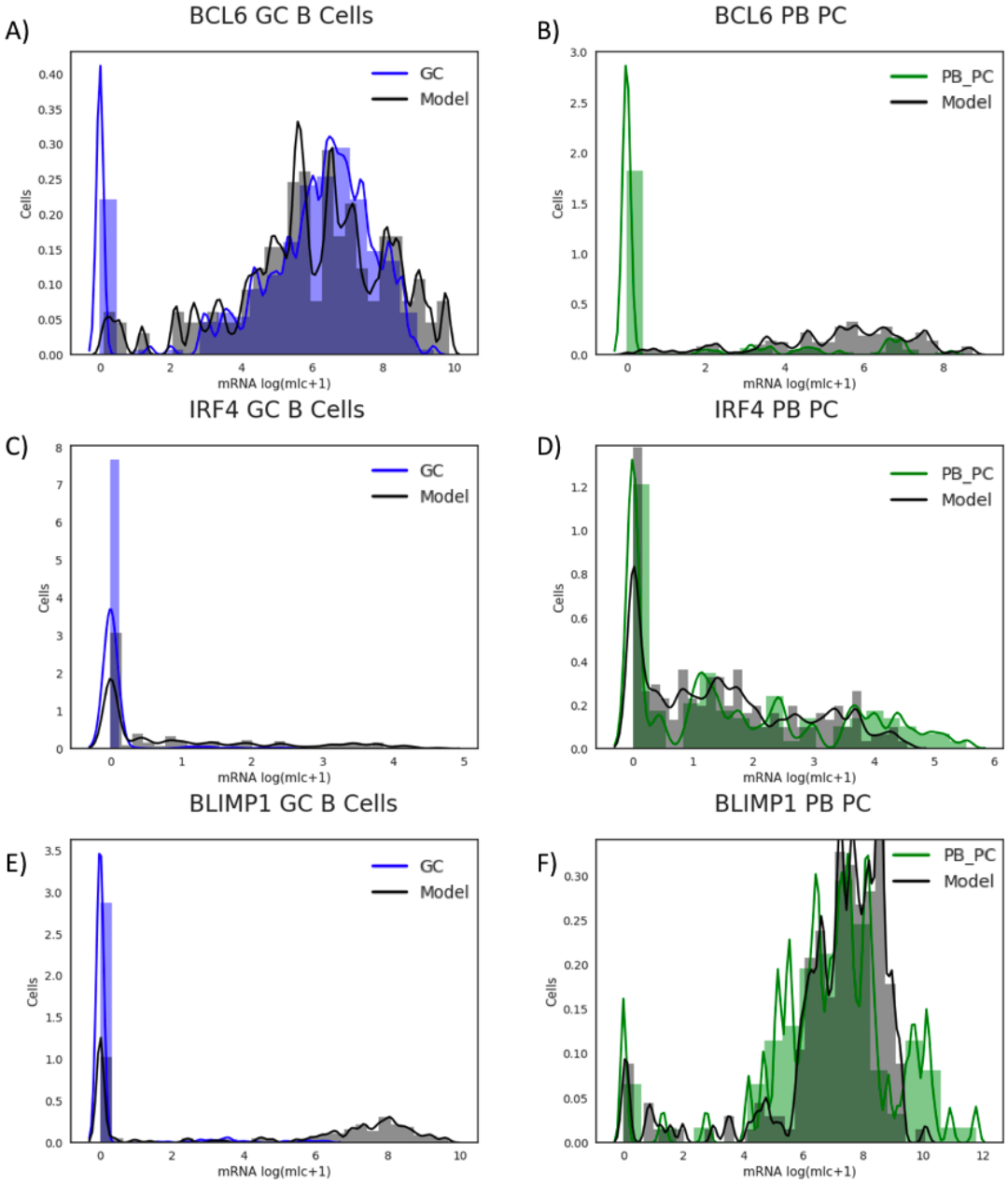


Figure 2.5. The Kernel Density Estimate (KDE) plot and histograms of model-generated and experimental mRNA counts of BCL6, IRF4, BLIMP1 at GC and PB_PC stages. The subgraphs A, C, E represent log (molecule+1) normalized SC with BCL6, IRF4 and BLIMP1 compared between the model estimations at GC stage (grey) vs the experimental data from GC B cells (blue). The subgraphs B, D, F represent log (molecule+1) normalized SC with BCL6, IRF4 and BLIMP1 compared between the model estimations at PB_PC stage (grey) vs the experimental data from PB_PC cells (green). Simulation of 200 SC were used based on the parameter set, selected after semi-automatized parameter screening (see Tables 2.2-2.5, version III). Performed based on dataset from Milpied et al. [3]

Mean Value	Model Estimated Value	Experimental Value
$\mu_{GC,\,BCL6}$	6.0	5.62
$\mu_{GC,\,IRF4}$	1.0	0.08
$\mu_{GC,\,BLIMIP1}$	4.69	0.34
$\mu_{PB\_PC,\,BCL6}$	5.26	1.25
$\mu_{PB\_PC,\,IRF4}$	1.49	1.68
$\mu_{PB\_PC,\,BLIMP1}$	7.03	6.91

Table 2.6. Mean values of simulations normalized as mRNA log(molecule+1) for the distribution generated by model (2.4)-(2.6) and the experimental data, see Figure 2.5. Based on the parameter set Tables 2.2-2.5, version III.

2.5 Discussion

In this work we first have rewritten an ODE reduced PDMP model (2.11) in terms of kinetic ODE model (2.1)-(2.3) and established an initial parameter value $k_{on,\,init,\,IRF4}$. We further have shown that for the parameter set (see Table 2.2-2.5, Version I), the ODE reduced PDMP model (2.11) simulates two steady states.

Secondly, we evaluated the effect of stochasticity of the multiple independent generations of the number of mRNA molecules by the PDMP model (2.4)-(2.6) and we confirmed that for the same parameter set there is no noticeable difference between each model-generated outputs for BCL6-IRF4-BLIMP1 GRNs (see Figure 2.4). These results allow to perform a combined parameter screening with a confidence that for each candidate parameter set, the algorithm needs to perform only one run of the PDMP model (2.4)-(2.6).

Lastly, we showed that the PDMP model (2.4)-(2.6) can simulate distributions of the number of mRNA molecules for BCL6, IRF4, BLIMP1 at GC and PB_PC stages with same order of magnitude than experimental data.

However, as a future scope of work, few strategies to improve the final param-

eter set (see Tables 2.2-2.5, version III) can be investigated.

At first, due to the fact that in BCL6-IRF4-BLIMP1 GRN, IRF4 depends only on its autoactivation reaction, we only have succeeded to directly rewrite ODE reduced PDMP model (2.11) in terms of kinetic model (2.1)-(2.3) and estimate the value of $k_{on,\,init,\,IRF4}$. It would be advantageous to additionally develop a process, which would allow to directly estimate the values of $k_{on,\,init,\,BCL6}$ and $k_{on,\,init,\,BLIMP1}$, using the same logic. However, because BLIMP1 depends on BLIMP1, IRF4 and BCL6 (see Equation (2.1)) and BCL6 depends on both IRF4 and BLIMP1 (see Equation (2.2)), the rewriting of system (2.4)-(2.6) in terms of (2.1)-(2.3) will require additional calculations and simplifications.

Secondly, we can evaluate an effect of mutual repression between BCL6 and BLIMP1 (see Figure 2.1) by performing more extensive parameter search. Current parameter set (see Tables 2.2-2.5, version III) makes the PDMP model (2.4)-(2.6) overestimates a number of mRNA molecules of BLIMP1 at GC stage. Increasing BCL6 repression of BLIMP1 could potentially decrease the quantity of BLIMP1 at the GC stage.

Third, it could be interesting to investigate the effect of the duration of the BCR and CD40 stimuli on the differentiation from GC B cells towards PB_PC. Multiscale modelling of GCs performed by Tejero et al. [56] showed that CD40 signalling in combination with asymmetric division of B cells results in switch from MC to PB. It would be good to evaluate a possible application of the PDMP model to study the effect of combined CD40 and BCR signalling with different intensity and duration at the SC level.

Additionally, one can evaluate the effects and inclusion of additional genes into the BCL6-IRF4-BLIMP1 GRNs and their impact on the quality of the fitting

of the data by the PDMP. One of the possible candidate to incorporation to GRN is paired box protein 5 (PAX5) which plays an important role in directing of lymphoid progenitors towards B cell development [175]. PAX5 positively regulates interferon regulatory factor 8 (IRF8) and BTB and CNC homologue 2 (BACH2) and which also positively regulates IRF4. In turn, IRF8 and BACH2 negatively regulates BLIMP1 at early stage of B cell differentiation. During further development, BLIMP1 starts to repress PAX5, consequently decreasing the expression of IRF8 and BACH2. The correct orchestration of PAX5-IRF8-BACH2 during the B cell differentiation is important for the successfully differentiation towards antigen producing cells (PB_PC), while its malfunction can cause aberration in GC B cell development [176].

CD40 stimulation of B cells initiates NF-κB signalling that is associated with cellular proliferation. In B cells, NF-κB activates IRF4, negatively regulates BACH2 what leads to positive regulation of BLIMP1 and consecutively repression of BCL6 [30, 177].

Another important transcription factor (TF) in GC development is MYC, which regulates B cell proliferation [178] and the DZ B cell phenotype [179]. MYC activates the histone methyltransferase enhancer of zeste homologue 2 (EZH2), which is responsible for the repression of IRF4 and BLIMP1 [180, 181, 182, 183].

The transcription factors mentioned above are present in SC RT-qPCR dataset from Milpied et al. [3] and could be used to extend BCL6-IRF4-BLIMP1 GRN. Inclusion of the additional TF may have both positive and negative effects on application of PDMP model (2.4)-(2.6). On the one side, it can increase the computational time and the number of parameters required for System (2.4)-

(2.6). On the other side, because inclusion of the TF can more precisely describe the biological system it could improve the quality of the fitting. However, any inclusion of new nodes to GRN should be carefully evaluated and only essential TF should be added. For instance, there is no advantageous in addition of TF that only have one downstream output. As example, MYC activates E2F Transcription Factor 1 (E2F1) and further activates EZH2. For this reason, incorporation of the chain MYC-E2F1-EZH2 should have similar outcome, as incorporation of simplified MYC-EZH2 reaction. This is expected, because in the modelling, intermediate elements of one-to-one redundant reactions can be omitted without significant changes in the quality of the simulations.

Overall, the parameter's tuning of a nonlinear model is not a trivial task and requires a combined approach. For instance, a modulation of any reaction in BCL6-IRF4-BLIMP1 GRN (for instance, activity of IRF4 activation of BLIMP1), causes a consecutive decrease of the number of BCL6 mRNA molecules. For this reason, the combination of the fully automatized and semi-automatized strategies is an optimal approach to improve the quality of the fitting of the experimental SC data of GC B cells. At first, the direct automatized screening allows to test the candidate space of parameter set in a general manner, and then the semi-automatized screening more precisely tune the PDMP model (2.4)-(2.6).

To further continue our study, we could also use scRNA-seq dataset from Milpied et. al [3]. The authors have produced scRNA-seq dataset from GC B cells and analysed the similarities between scRNA-seq and SC RT-qPCR dataset. Even though the gene-gene correlation levels were lower in scRNA-seq comparing to SC RT-qPCR, scRNA-seq analysis confirmed the observation

obtained by SC RT-qPCR [3]. From the stochastic modelling perspective, combining the data from SC RT-qPCR and scRNA-seq should give an additional information to improve evaluation of variability of the data and quality of the fitting.

To summarise, the PDMP model (2.4)-(2.6) is capable to qualitatively simulate and depict the stochasticity of the experimental SC gene expression data of human B cell at the GC and PB_PC stages of differentiation using the GRN of three-key genes BCL6-IRF4-BLIMP1. These results are encouraging, and suggest that our model may be used to test the different B cell exits from GC. Future steps may include testing of the PDMP model (2.4)-(2.6) on the alternative SC datasets [184, 185, 186] and investigate the malignant formations, and evaluate differences of the GRN comparing to the normal B cell differentiation from GC towards PB_PC.

2.6 Funding

This work was supported by a COSMIC (www.cosmic-h2020.eu) which has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no 765158.

2.7 Acknowledgments

We thank Arnaud Bonnaffoux and Matteo Bouvier for their help with the WASABI framework and their critical reading of the manuscript. We thank the computational center of IN2P3 (Villeurbanne/France), especially Gino Marchetti and Renaud Vernet. We also thank Aurelien Pélissier and Elias

Ventre for scientific discussions.

Chapter 3

Discussion and perspectives

SC PDMP system for modelling GRN of GC B cells

In our work we have used a novel stochastic PDMP model for a GRN of three key TF responsible for B cell differentiation (BCL6, IRF4, BLIMP1), applied to SC transcriptomic data from human lymphoid organs (spleen and tonsil) [3].

We first rewrote the currently available kinetic ODE model (2.1)-(2.3) [2] in terms of ODE reduced PDMP model (2.11) and estimated the value of $k_{on,\,init,\,IRF4}$ parameter. We have shown that for an estimated value of $k_{on,\,init,\,IRF4}$, ODE reduced PDMP model (2.11) successfully recapitulates two steady states associated with GC B cell differentiation (see Figure 2.2). We then focused on the full PDMP model dynamics. We have applied KD and estimated the variability of the PDMP model. We have shown that the PDMP model has qualitatively similar outputs and a unique run is sufficient for performing the parameter tuning (see Figure 2.4).

Finally, we performed a combination of automatized and semi-manual tuning, which allowed us to reproduce the stochastic variation of the SC data both at

the GC and PB_PC stages of differentiation (see Figure 2.5). Our results are promising and show that the PDMP model can be used to recapitulate the behavior of the GC B cell differentiation towards PB_PC and can potentially improve an understanding of the GC B cell development.

Sensitivity and identifiability analysis

Modelling of complex systems faces multiple uncertainties which complicate the estimation of the parameters and decrease the model's efficacy [187]. For instance, variation by orders of magnitude of some parameters of the model may not significantly influence the quality of the fitting of the experimental data [188]. Another challenge is establishing parameters that are functionally related to each other, i.e. interdependent parameters [189]. From the biological point of view, complex systems should possess robustness against external perturbation and internal noise, which in modelling language are represented by parameters with low sensitivity or insensitivity [188].

Tuning of the parameters for a high-dimensional stochastic model is not a trivial task. Our approach combines automatized and semi-manual tuning of parameters of the PDMP model and showed that the model can qualitatively reproduce the experimental SC data. However, sensitivity and identifiability analysis can be performed for better evaluation of the effect of each parameter on the PDMP model and potentially improve the tuning efficiency.

To understand which parameters are essential for the improvement of the data fitting by the PDMP model, we could perform an identifiability analysis [190, 191, 192]. Identifiability analysis is widely used for deterministic (ODE) models but is in an early adoption phase in stochastic modelling [193]. Browning et

al. [193] have presented an analysis of structural identifiability (i.e. parameters which can be estimated from unlimited noisy data) and practical identifiability (i.e. parameters can be estimated from the limited noisy experimental data). The authors used Markov chain Monte Carlo methods (MCMC) [194, 195] to study practical identifiability and DAISY [196] to study structural identifiability of various models, including birth-death process model, two-pool model, epidemic model, β-insulin-glucose circuit [193]. However, because PDMP model has higher complexity compared to the stochastic differential equation (SDE) model from Browning et al. [193], structural identifiability analysis may be hardly achieved for our model. We could rather perform a practical identifiability analysis which can be helpful to evaluate the reactions candidates to include in our BCL6-IRF4-BLIMP1 network, avoiding redundant nodes (exclusion of which, will not decrease the model's performance) [193]. On the other side, practical application (pros/cons) of the identifiability analysis for the PDMP model should be evaluated, taking into consideration the high computational costs of the MCMC method for stochastic models of GRN. A successful analysis of the PDMP model identifiability could allow us to decrease the number of parameters that should be tuned for our GRN (see Figure 2.1).

After determining which parameters of the models are identifiable and which are unidentifiable, one can perform an additional study of the parameters of the model, by using sensitivity analysis [188]. Sensitivity analysis allows evaluating which parameters are more relevant in the model, which parameters require additional tuning, and also assists in the evaluation of the robustness of the model [197].

There are multiple approaches for sensitivity analysis, such as Monte Carlo simulations, Fisher's information matrix with sensitivity measures, Girsanov transformation method and others [198, 199, 200, 201, 200, 202, 203]. Some authors use alternative algorithms while trying to adapt sensitivity methods for the stochastic models. For instance, Eric A. Sobie's group introduced parameter randomization followed by multivariable regression and showed that this method allows having precise predictions of biological data with low computational time [204, 205, 206].

However, due to the complexity of the PDMP model and due to the fast increase in degrees of freedom with addition of genes to the GRN, those strategies will be computationally expensive. For this reason, a detailed evaluation of the most appropriate algorithm, which will provide an optimal ratio time versus the effectiveness of sensitivity analysis, should be performed. For the PDMP model applied to BCL6-IRF4-BLIMP1 GRN, we can attempt to apply the sensitivity analysis suggested by Eric A. Sobie's group and compare it to a standard MCMC sensitivity analysis [204, 205, 206].

A deeper analysis of the relevance of each parameter of the PDMP model, accessed via identifiability and sensitivity analysis can be advantageous. This will allow the establishment of candidate parameters with the highest effect on the quality of the fitting of experimental data by the PDMP model.

Experimental SC dataset

In our work we have used a SC RT-PCR dataset describing GC B cell from human lymphoid organs at GC and PB_PC stages of differentiation [3]. To our current knowledge, it was one of the most recent and detailed SC charac-

terizations of human GC B cell differentiation.

The next step should be an application of the PDMP model to another SC dataset of the same biological system and an evaluation of how well the model simulates the experimental SC data obtained from an alternative technique (scRNA-seq). However, combining SC RT-PCR experimental data with scRNA-seq experimental data is not a trivial task, because different approaches are used to extract transcriptomic information from the cells [207, 208]. To our current knowledge, there is still no direct way to establish the common unit between scRNA-seq and SC RT-PCR. Richard et al. [209] have presented the relationship between the number of mRNA molecules of each gene in a single cell and the value of Ct obtained by SC RT-PCR. The number of mRNA molecules in a cell could be a candidate for the common unit between scRNA-seq and SC RT-PCR. However, current scRNA-seq protocols are only capable to estimate relative, rather than quantitative gene expression [210, 211, 212, 213, 214].

For this reason, instead of directly combining SC RT-PCR and scRNA-seq datasets, alternative strategies should be used. We should evaluate how different the parameter set of the PDMP model will be while fitting the experimental data from SC RT-PCR versus while fitting the experimental data from scRNA-seq. It is currently believed that scRNA-seq is highly correlated with the standard transcriptomic tools (SC RT-PCR) [215, 216]. For the GC B cell differentiation, Milpied et al. [3] also showed high similarity for the GC B cell dataset. The authors have collected scRNA-seq and SC RT-PCR datasets during the same experimental design and showed similar sources of heterogeneity between datasets. For this reason, we could expect that the PDMP model

would be able to fit each of the datasets with minor changes.

As a first approach to evaluate the performance of PDMP model between two experimental datasets, we can perform a combination of parameter's tuning (see Section 2.3.5) starting with the parameter set obtained from previous tuning (based on SC RT-PCR) and then tune the parameters set, until PDMP will be able to fit the scRNA-seq data. After this, one needs to evaluate which parameters are different between the parameter set obtained after SC RT-PCR tuning and the parameter set obtained after additional tuning of scRNA-seq dataset. As a second approach, one can perform independent parameter tuning starting with random values of the parameter values (with the same order of magnitude as Bonnaffoux et al. [138]) and tune the parameters of the model-based only on the scRNA-seq experimental data. Then one can compare the values of parameter sets for the PDMP model between each independent tuning procedure.

In any case, unless the study will be done, it is hard to predict which approach will be more advantageous. For this reason, both strategies should be explored and implemented.

Recently, Holmes et al. [186] group have performed 10x Genomics scRNA-seq for cells from human tonsils which were classified as DZ and LZ cells based on the expression of CXCR4 and CD83 markers. Both Holmes et al. [186] and Milpied et al. [3] have used 10X Genomics. For this reason it is important to evaluate how different will parameter values be when fitting data from Milpied et al [3] and data from Holmes et al [186]. Evaluating differences between parameter sets we could look for possible correlations between the differences in those experimental data sets. Because those datasets are taken from similar

biological systems we could expect the parameter sets for the PDMP to be similar too.

At last, in order to remove possible sample-related variation between different batches, we could perform the meta-analysis of scRNA-seq datasets from Holmes et al. [186] and Milpied et al. [3] and combine it in one dataset using the method for batch correction [217]. Further we could perform the fitting of the data by PDMP model and evaluate what will be the parameter set allowing the model to mimic the experimental data and how it will be different.

Simulation of GC B cells at DZ and at LZ

During the GC maturation, B cells transition from DZ to LZ and re-enter to DZ for further rounds of SHM. Milpied et al. [3], had performed pseudotime analysis of human GC B cells and separated GC B cells at DZ versus LZ. SC data showed detectable differences in gene expression patterns between DZ and LZ [3]. Because some authors suggest that DZ/LZ markers may represent the transitory state [186, 218], it will be important to evaluate how the stochastic modelling of GRN will be capable to detect the differences in GRN state between DZ and LZ GC B cells to further clarify that differences.

During B cells maturation in GC different mechanisms down-regulate the amount of BCL6. As currently known, those mechanisms include BCL6 degradation caused by antigen-triggered BCR activation, CD40 triggered up-regulation of IRF4 expression and consecutive BLIMP1 up-regulation [2].

During the inclusion of LZ data to our model, we should evaluate and test the time at which we should fit the experimental data for the LZ GC cells. For this, we could add the additional parameter which will correspond to the time

after the BCR stimuli when cells will have the LZ gene expression pattern.

To our current knowledge, the precise time for B cell migration from DZ to LZ in humans is not known. However, experiments in mice showed that after 4-6 hours, 50% of cells have migrated from DZ to LZ [219]. Later, Beltman et al. [220] constructed a model of GC B cells migration from DZ to LZ and estimated this time to be in the range of few hours. Based on these results, we could decrease the space for tuning this parameter in our simulations.

Effect of the stimuli on the model behaviour

During the differentiation from GC B cell towards PB_PC, cells receive BCR signal from the antigen (repressing BCL6) and T-cell-mediated stimulation through CD40 (activating IRF4). Koike et al.[221], Yam et al. [222] and Tejero et al. [56] suggest that the intensity and the duration of stimuli may be responsible for different B cell differentiation fates. Simulation of different stimuli duration and different intensities of the stimuli in the PDMP model could improve the quality of the fitting and the way the model represents the experimental data.

Modification and extension of BCL6-IRF4-BLIMP1 GRN

GC B cell development is a robust and well-controlled mechanism, which allows selecting the B cells producing antibodies with high affinity to the antigen. It is currently thought that GC B cell differentiation can be described by the three key regulatory transcription factors BCL6, IRF4 and BLIMP1 [2]. However, an extension of the BCL6-IRF4-BLIMP1 GRN with additional TFs, participating in GC B cell development, could be advantageous and could potentially

improve the fitting of the experimental dataset. The addition of the candidate TFs should allow the PDMP model to more precisely describe the biological process during the differentiation. It would also allow the PDMP model to have access to additional experimental SC information from the complementary genes to tune the parameters. Inclusion of the relevant TF candidates could allow better depicting the relationship between genes, which may be lost while using only three nodes BCL6-IRF4-BLIMP1 GRN. We have analysed possible candidate nodes, which may be relevant in the GC B cell differentiation and which are available in the experimental dataset from Milpied et al. [3].

First available candidate for incorporation to the BCL6-IRF4-BLIMP1 GRN is the PAX5-IRF8-BACH2 pathway (paired box protein 5, interferon regulatory factor 8, BTB and CNC homologue 2, see Figure 3.1A). PAX5 is a TF that plays an important role in directing the lymphoid progenitors towards B cell differentiation [223]. PAX5 positively regulates IRF8, BACH2, and IRF4. PAX5 also negatively regulates BLIMP1 via IRF8 and BACH2 [176]. At the same time, IRF4 also indirectly regulates PAX5 via BLIMP1 repression of PAX5 [176, 224].

One also could include NF-κB (factor nuclear kappa B), initiated by the CD40 stimulus (see Figure 3.1B). NF-κB is a TF that is associated with the proliferation of B cells, which positively regulates IRF4 and positively regulates BLIMP1 via BACH2. In turn, BACH2 is positively regulated by BCL6 [30, 177, 225].

Another possible candidate for extension of initial BCL6-IRF4-BLIMP1 GRN, is MYC-EZH2 (c-Myc, enhancer of zeste homologue 2, see Figure 3.1C). MYC

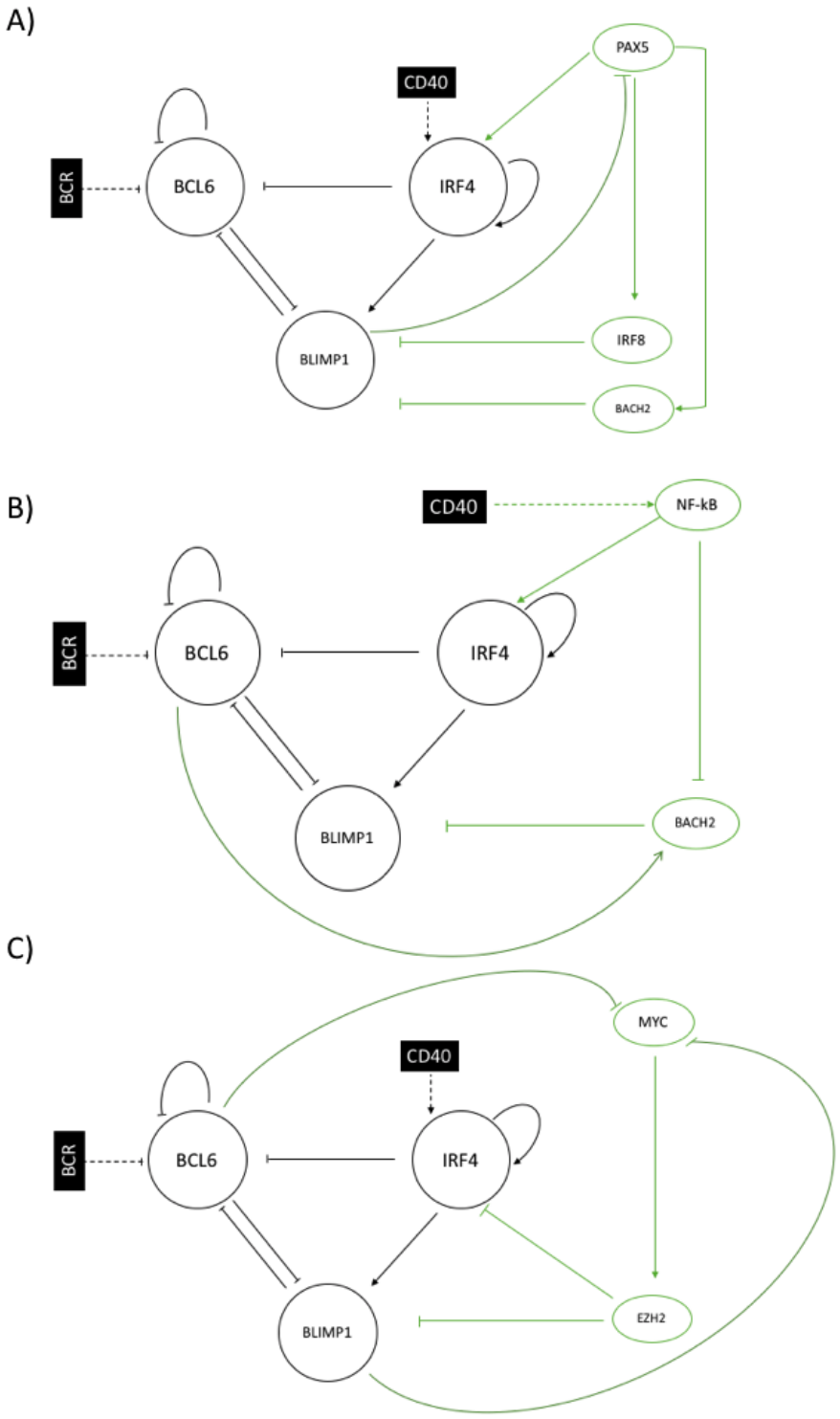


Figure 3.1. The candidate genes to include to BCL6-IRF4-BLIMP1 GRN. A) PAX5-IRF8-BACH2; B) NF-κB - BACH2; C) MYC-EZH2.

is an important TF regulating B cell proliferation and DZ B cell phenotype [178, 179]. MYC is also responsible for indirect activation of epigenetic regulation of GC B cell differentiation, via EZH2 and for negative regulation of IRF4 and BLIMP1. During the differentiation, MYC is repressed by BCL6 (before the formation of GC) and by BLIMP1 during differentiation towards plasma cells [226].

The addition of genes to the key BCL6-IRF4-BLIMP1 GRN may be performed in different ways. One of the possibilities is to add all the candidate TFs presented in the SC dataset and perform the sensitivity analysis and parameter tuning of the PDMP model. This approach is straightforward and will take into account all possible interconnections between candidate TFs. However, this approach may be computationally expensive for GRN with a complex structure.

Alternatively, we can add one branch at each step (for instance, PAX5-IRF8-BACH2), and then perform the sensitivity, identifiability analysis and parameter's tuning, and then add the next candidate branch. This strategy can create additional information regarding the way how the PDMP model fits experimental data for different GRN structures. Following this logic, when all the nodes will be added to the GRN, one would have a complete analysis of the model's robustness and its response to GRN modification.

Ideally, the inclusion of the genes to the network or modification of duration and intensity of the stimuli should be followed by a comparison of experiment with model-generated data and additionally, by sensitivity and identifiability analysis.

It is important to evaluate how the PDMP model's capacity to fit the ex-

perimental data will vary while including the regulators with unique regulation properties (i.e. MYC-EZH2) versus ones that have similar regulatory properties (i.e. PAX5-IRF8-BACH2 and NF-κB-BACH2). For instance, only MYC-EZH2 adds the repression effect on IRF4 which may be an important regulatory mechanism, which is not present in other reactions.

GRN inference using SC data

One of the alternative approaches to investigate, build, complement and modify GRN structure is to use a novel pipeline WASABI [138]. This algorithm allows to infer GRN using the experimental time-stamped SC transcriptomic and bulk proteomic data [138]. WASABI infer GRN by adding nodes "node-by-node" using as the main source information an experimental data, which allows diminishing an effect of predefined knowledge. WASABI can be helpful to evaluate which nodes and pathways may be added to BCL6-IRF4-BLIMP1 GRN to improve the representation of the experimental SC dataset.

However, to access all the advantages of the inference of GRN via WASABI, one will need an access to the time-stamped experimental data. Depending on the experimental design (cost versus effectiveness), it can be required up to 3-4 time points to follow GC B cell development (at each timepoint SC data should be collected and analysed). It is a very promising and challenging experimental design that should have been generated within the COSMIC consortium, and it should have allowed identifying which TFs are playing an essential role. Such a dataset also should have helped to establish what may be an optimal GRN regulating GC B cell differentiation and also help to analyse possible candidates TFs responsible for malignant GC B cell differentiation.

However, due to a pandemic outbreak, such a dataset has not been generated yet.

Application of SC PDMP model for the aberrant B cell differentiation

The advantage of GRN modelling of biological systems is that it allows to study normal conditions and compare those to different pathological cases. As was shown by Martinez et al. [2], kinetic ODE model of BCL6-IRF4-BLIMP1 is capable to simulate different malignant fates of GC B cell differentiation exits, including DLBCL, loss of BCL6 autoregulation, constitutive high expression of BCL6, synergistic loss of IRF4 and BLIMP1-mediated BCL6 silencing and reduced BLIMP1 stability.

One of the ways to use PDMP model is to analyse the parameter set tuned for the normal GC B cell differentiation towards PB_PC and try to simulate SC gene expression of the pathological condition by modulating the activity of the reaction of interest (for instance $\theta_{i,j}$). For instance, the most frequent genetic aberration in DLBCL affects BCL6 promoter region [2]. To simulate it, we can attempt to fit the experimental data, using the PDMP model. We can start with the parameter set tuned to the normal condition and then perform additional tuning of the parameters responsible for the BCL6 regulation (i.e. $\theta_{3,1}, \theta_{2,1}$ or $\theta_{1,1}$) until the model fits SC dataset from the pathological condition [3, 186, 185].

Alternatively, we can tune the parameter set directly to the malignant SC condition, and analyse which parameters of the PDMP model are different from the parameter set corresponding to the normal GC B cell development.

Those approaches may lead to a candidate parameter set that can be responsible for a specific pathological condition. After establishing the parameters that can be candidates for the specific pathological condition (for instance $\theta_{i,j}$), one could use the molecular biology tools to perform silencing/overexpression of the specific gene of the network *in-vitro* or *in-vivo* to validate the hypothesis obtained from the PDMP model parameter tuning (i.e. confirm the importance of candidate parameter in the biological system).

scRNA modelling in a multi-omic world and its place in systems biology

SC technologies have revolutionized the way biology can study multiscale processes orchestrating the normal cell development and its aberration, by providing multidimensional high-throughput data at different timepoints [227].
There are different strategies to extract and to structure the information from the high-dimensional data towards predictive models: artificial intelligence (AI) techniques and mechanistic models [228, 159].
AI unites unsupervised and supervised machine learning (ML) techniques. Unsupervised ML includes dimension reduction algorithms, latent methods and normalization procedures which are aimed at extracting the most relevant information from SC multi-dimensional data and presenting essential factors responsible for a specific biological behavior [229, 230]. Recently developed ML analysis of SC data based on topology methods showed that it is capable of recapitulating complex nonlinear regulatory processes in the biological system [231, 232, 233]. On the other hand, the supervised ML apply the concept of "training set" (i.e. currently known knowledge) to teach the model to

classify the unknown datasets which can be applied to automatic SC classification, annotation and regression [234, 235, 236, 237, 238]. However in order to improve understanding and establish the connections between the biological mechanisms and the inferred nonlinear properties obtained from the data the novel subfield, interpretable AI, is currently under development [227, 239]. Interpretable AI approach is facing the task to extract the relevant knowledge from the multiple ML models, applied to the dataset of interest [240]. As an example, Wang et al. [241] used interpretable AI to identify groups of genes responsible for distinct subcellular types.

The second class of the models currently used for the SC analysis is a class of mechanistic models that incorporates the carefully analysed and reviewed prior biological knowledge to the analysis of the novel experimental data [242, 243]. GRN is used to model the relationship between different TF and target genes. GRN allows studying the inter and intra cellular interactions on the different levels of the cell regulation, responsible for the cell development and for the cell fates. The advantage of GRN over the non supervised machine learning technique is that GRN combines together pre-existing knowledge and reasoning behind biological mechanisms with the experimental dataset [242]. Multiple algorithms for GRN inference from SC transcriptomic data have been recently developed [138, 244, 245, 246]. Nevertheless, because the mechanistic GRN models are built on the previously known rules and assumptions, it may limit the search of the new unknown regulations by adding pre-existing biases of known regulations. Recent algorithm WASABI (discussed above) infers the network based on the input SC transcriptomic and bulk proteomic and can overcome this limitation [138].

However, there is still room for further improvement in the extraction of useful information from complex datasets. Integration of the multi-omic data, which will combine the transcriptomic, spatial transcriptomic, epigenomic, proteomic, metabolomic and fluxomic still seems to be a hardly achievable goal [247, 248, 249, 250, 251]. It is common to analyse different omics separately and cross-validate the obtained conclusions, but complete multidimensional integration is not reached yet. Nevertheless, future analysis of multiscale models, will improve the quality of the studied biological system and its regulations in a complete, multilevel way and promises to open a new era in our understanding of the mechanisms of living organisms at normal and pathological cases [252, 253, 254, 255, 256, 257, 258].

Bibliography

[1] Mark WEJ Fiers, Liesbeth Minnoye, Sara Aibar, Carmen Bravo González-Blas, Zeynep Kalender Atak, and Stein Aerts. "Mapping gene regulatory networks from single-cell omics data". In: *Briefings in functional genomics* 17.4 (2018), pp. 246–254.

[2] Maria Rodriguez Martinez, Alberto Corradin, Ulf Klein, Mariano Javier Álvarez, Gianna M Toffolo, Barbara di Camillo, Andrea Califano, and Gustavo A Stolovitzky. "Quantitative modeling of the terminal differentiation of B cells and mechanisms of lymphomagenesis". In: *Proceedings of the National Academy of Sciences* 109.7 (2012), pp. 2672–2677.

[3] Pierre Milpied, Iñaki Cervera-Marzal, Marie-Laure Mollichella, Bruno Tesson, Gabriel Brisou, Alexandra Traverse-Glehen, Gilles Salles, Lionel Spinelli, and Bertrand Nadel. "Human germinal center transcriptional programs are de-synchronized in B cell lymphoma". In: *Nature immunology* 19.9 (2018), pp. 1013–1024.

[4] Peter Parham. "Innate immunity: the unsung heroes". In: *Nature* 423.6935 (2003), pp. 20–20.

[5] Theodoros Kyrkoudis, Gregory Tsoucalas, Vasileios Thomaidis, Ioannis Bakirtzis, Eleni Nalbandi, Alexandros Polychronidis, Aliki Fiska,

and A Polychronidis. “Vaccination of the ethnic Greeks (Rums) against smallpox in the Ottoman Empire: Emmanuel Timonis and Jacobus Pylarinos as precursors of Edward Jenner”. In: *ERC> YES MEDICAL JOURNAL* 43.1 (2020), pp. 100–106.

[6] Darren R Flower. *Bioinformatics for vaccinology*. John Wiley & Sons, 2008.

[7] Brian Greenwood. “The contribution of vaccination to global health: past, present and future”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1645 (2014), p. 20130433.

[8] Alexandra-Chloé Villani, Siranush Sarkizova, and Nir Hacohen. “Systems immunology: Learning the rules of the immune system”. In: *Annual review of immunology* 36 (2018), pp. 813–842.

[9] Thomas J Kindt, Richard A Goldsby, Barbara A Osborne, and Janis Kuby. *Kuby immunology*. Macmillan, 2007.

[10] Takeshi Matsui and Masayuki Amagai. “Dissecting the formation, structure and barrier function of the stratum corneum”. In: *International immunology* 27.6 (2015), pp. 269–280.

[11] Bani Preet Kaur and Elizabeth Secord. “Innate Immunity.” In: *Pediatric Clinics of North America* 66.5 (2019), pp. 905–911.

[12] Max D Cooper and Matthew N Alder. “The evolution of adaptive immune systems”. In: *Cell* 124.4 (2006), pp. 815–822.

[13] Xiying Chi, Yue Li, and Xiaoyan Qiu. “V (D) J recombination, somatic hypermutation and class switch recombination of immunoglobu-

lins: mechanism and regulation". In: *Immunology* 160.3 (2020), pp. 233–247.

[14] Elisa Laurenti and Berthold Göttgens. "From haematopoietic stem cells to complex differentiation landscapes". In: *Nature* 553.7689 (2018), pp. 418–426.

[15] Ashley P Ng and Warren S Alexander. "Haematopoietic stem cells: past, present and future". In: *Cell death discovery* 3.1 (2017), pp. 1–4.

[16] Harry W Schroeder Jr, Andreas Radbruch, and Claudia Berek. "B-cell development and differentiation". In: *Clinical Immunology*. Elsevier, 2019, pp. 107–118.

[17] Kenneth Murphy and Casey Weaver. *Janeway's immunobiology*. Garland science, 2016.

[18] Anja E Hauser and Uta E Höpken. "B cell localization and migration in health and disease". In: *Molecular biology of B cells*. Elsevier, 2015, pp. 187–214.

[19] IC MacLennan and David Gray. "Antigen-driven selection of virgin and memory B cells." In: *Immunological reviews* 91 (1986), pp. 61–85.

[20] Joshy Jacob, Garnett Kelsoe, Klaus Rajewsky, and Ursula Weiss. "Intraclonal generation of antibody mutants in germinal centres". In: *Nature* 354.6352 (1991), pp. 389–392.

[21] Joshy Jacob and Garnett Kelsoe. "In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl) acetyl. II. A common clonal origin for periarteriolar lymphoid sheath-associated foci and germinal cen-

ters." In: *The Journal of experimental medicine* 176.3 (1992), pp. 679–687.

[22] Facundo D Batista and Naomi E Harwood. "The who, how and where of antigen presentation to B cells". In: *Nature Reviews Immunology* 9.1 (2009), pp. 15–27.

[23] Nilushi S De Silva and Ulf Klein. "Dynamics of B cells in germinal centres". In: *Nature reviews immunology* 15.3 (2015), pp. 137–148.

[24] Gabriel D Victora and Michel C Nussenzweig. "Germinal centers". In: *Annual review of immunology* 30 (2012), pp. 429–457.

[25] Katia Basso and Riccardo Dalla-Favera. "Germinal centres and B cell lymphomagenesis". In: *Nature Reviews Immunology* 15.3 (2015), pp. 172–184.

[26] Coraline Mlynarczyk, Lorena Fontán, and Ari Melnick. "Germinal center-derived lymphomas: The darkest side of humoral immunity". In: *Immunological reviews* 288.1 (2019), pp. 214–239.

[27] Gabriel D Victora, David Dominguez-Sola, Antony B Holmes, Stephanie Deroubaix, Riccardo Dalla-Favera, and Michel C Nussenzweig. "Identification of human germinal center light and dark zone cells and their relationship to human B-cell lymphomas". In: *Blood* 120.11 (2012), pp. 2240–2248.

[28] Katia Basso and Riccardo Dalla-Favera. "Roles of BCL6 in normal and transformed germinal center B cells". In: *Immunological reviews* 247.1 (2012), pp. 172–183.

[29] Laura Pasqualucci, Anna Migliazza, Katia Basso, Jane Houldsworth, RSK Chaganti, and Riccardo Dalla-Favera. "Mutations of the BCL6 proto-oncogene disrupt its negative autoregulation in diffuse large B-cell lymphoma". In: *Blood, The Journal of the American Society of Hematology* 101.8 (2003), pp. 2914–2923.

[30] Masumichi Saito, Jie Gao, Katia Basso, Yukiko Kitagawa, Paula M Smith, Govind Bhagat, Alessandra Pernis, Laura Pasqualucci, and Riccardo Dalla-Favera. "A signaling pathway mediating downregulation of BCL6 in germinal center B cells is blocked by BCL6 gene alterations in B cell lymphoma". In: *Cancer cell* 12.3 (2007), pp. 280–292.

[31] José-Francisco Garcia, Giovanna Roncador, AI Sanz, L Maestre, E Lucas, S Montes-Moreno, R Fernandez Victoria, JL Martinez-Torrecuadrara, T Marafioti, DY Mason, et al. "PRDM1/BLIMP-1 expression in multiple B and T-cell lymphoma". In: *Haematologica* 91.4 (2006), pp. 467–474.

[32] Raluca Eftimie, Joseph J Gillard, and Doreen A Cantrell. "Mathematical models for immunology: current state of the art and future research directions". In: *Bulletin of mathematical biology* 78.10 (2016), pp. 2091–2134.

[33] Sarah M Andrew, Christopher TH Baker, and Gennady A Bocharov. "Rival approaches to mathematical modelling in immunology". In: *Journal of Computational and Applied Mathematics* 205.2 (2007), pp. 669–686.

[34] Ronald N Germain, Martin Meier-Schellersheim, Aleksandra Nita-Lazar, and Iain DC Fraser. "Systems biology in immunology: a computational modeling perspective". In: *Annual review of immunology* 29 (2011), pp. 527–585.

[35] Mario Castro, Grant Lythe, Carmen Molina-Parıs, and Ruy M Ribeiro. "Mathematics in modern immunology". In: *Interface focus* 6.2 (2016), p. 20150093.

[36] Arup K Chakraborty and Jayajit Das. "Pairing computation with experimentation: a powerful coupling for understanding T cell signalling". In: *Nature Reviews Immunology* 10.1 (2010), pp. 59–71.

[37] Boris N Kholodenko. "Cell-signalling dynamics in time and space". In: *Nature reviews Molecular cell biology* 7.3 (2006), pp. 165–176.

[38] Yoram Vodovotz, Gregory Constantine, Jonathan Rubin, Marie Csete, Eberhard O Voit, and Gary An. "Mechanistic simulations of inflammation: current state and future prospects". In: *Mathematical biosciences* 217.1 (2009), pp. 1–10.

[39] Jeffrey P Perley, Judith Mikolajczak, Gregery T Buzzard, Marietta L Harrison, and Ann E Rundell. "Resolving early signaling events in T-cell activation leading to IL-2 and FOXP3 transcription". In: *Processes* 2.4 (2014), pp. 867–900.

[40] Iren Bains, Rodolphe Thiébaut, Andrew J Yates, and Robin Callard. "Quantifying thymic export: combining models of naive T cell proliferation and TCR excision circle dynamics gives an explicit measure of

thymic output". In: *The Journal of Immunology* 183.7 (2009), pp. 4329–4336.

[41] Catherine Beauchemin, Narendra M Dixit, and Alan S Perelson. "Characterizing T cell movement within lymph nodes in the absence of antigen". In: *The Journal of Immunology* 178.9 (2007), pp. 5505–5512.

[42] Robin Callard and Phil Hodgkin. "Modeling T-and B-cell growth and differentiation". In: *Immunological reviews* 216.1 (2007), pp. 119–129.

[43] Edwin D Hawkins, Marian L Turner, Mark R Dowling, C Van Gend, and Philip D Hodgkin. "A model of immune regulation as a consequence of randomized lymphocyte division and death times". In: *Proceedings of the National Academy of Sciences* 104.12 (2007), pp. 5032–5037.

[44] Peter Ankomah and Bruce R Levin. "Exploring the collaboration between antibiotics and the immune response in the treatment of acute, self-limiting infections". In: *Proceedings of the National Academy of Sciences* 111.23 (2014), pp. 8331–8338.

[45] Fabien Crauste, Emmanuelle Terry, Isabelle Le Mercier, Julien Mafille, Sophie Djebali, Thibault Andrieu, Brigitte Mercier, Gaël Kaneko, C Arpin, Jacqueline Marvel, et al. "Predicting pathogen-specific CD8 T cell immune responses from a modeling approach". In: *Journal of theoretical biology* 374 (2015), pp. 66–82.

[46] Ha Youn Lee, David J Topham, Sung Yong Park, Joseph Hollenbaugh, John Treanor, Tim R Mosmann, Xia Jin, Brian M Ward, Hongyu Miao, Jeanne Holden-Wiltse, et al. "Simulation and prediction of the adaptive

immune response to influenza A virus infection". In: *Journal of virology* 83.14 (2009), pp. 7151–7165.

[47] Hongyu Miao, Joseph A Hollenbaugh, Martin S Zand, Jeanne Holden-Wiltse, Tim R Mosmann, Alan S Perelson, Hulin Wu, and David J Topham. "Quantifying the early immune response and adaptive immune response kinetics in mice infected with influenza A virus". In: *Journal of virology* 84.13 (2010), pp. 6687–6698.

[48] Giao T Huynh and Frederick R Adler. "Mathematical modelling the age dependence of Epstein–Barr virus associated infectious mononucleosis". In: *Mathematical medicine and biology: a journal of the IMA* 29.3 (2012), pp. 245–261.

[49] Shishi Luo, Michael Reed, Jonathan C Mattingly, and Katia Koelle. "The impact of host immune status on the within-host and population dynamics of antigenic immune escape". In: *Journal of The Royal Society Interface* 9.75 (2012), pp. 2603–2613.

[50] Ericka Mochan, David Swigon, G Bard Ermentrout, Sarah Lukens, and Gilles Clermont. "A mathematical model of intrahost pneumococcal pneumonia infection dynamics in murine strains". In: *Journal of theoretical biology* 353 (2014), pp. 44–54.

[51] Daniel K Choo, Kaja Murali-Krishna, Rustom Anita, and Rafi Ahmed. "Homeostatic turnover of virus-specific memory CD8 T cells occurs stochastically and is independent of CD4 T cell help". In: *The Journal of Immunology* 185.6 (2010), pp. 3436–3444.

[52] Susanna Celli, Mark Day, Andreas J Müller, Carmen Molina-Paris, Grant Lythe, and Philippe Bousso. "How many dendritic cells are required to initiate a T-cell response?" In: *Blood, The Journal of the American Society of Hematology* 120.19 (2012), pp. 3945–3948.

[53] Courtney L Davis and Frederick R Adler. "Mathematical models of memory CD8+ T-cell repertoire dynamics in response to viral infections". In: *Bulletin of mathematical biology* 75.3 (2013), pp. 491–522.

[54] Mikhail Kolev, Ana Markovska, and Adam Korpusik. "On a mathematical model of adaptive immune response to viral infection". In: *International Conference on Numerical Analysis and Its Applications*. Springer. 2012, pp. 355–362.

[55] Carlo Bianca and Julien Riposo. "Mimic therapeutic actions against keloid by thermostatted kinetic theory methods". In: *The European Physical Journal Plus* 130.8 (2015), pp. 1–14.

[56] Elena Merino Tejero, Danial Lashgari, Rodrigo Garcıa-Valiente, Xuefeng Gao, Fabien Crauste, Philippe A Robert, Michael Meyer-Hermann, Marıa Rodrıguez Martınez, S Marieke van Ham, Jeroen EJ Guikema, et al. "Multiscale Modeling of Germinal Center Recapitulates the Temporal Transition From Memory B Cells to Plasma Cells Differentiation as Regulated by Antigen Affinity-Based Tfh Cell Help". In: *Frontiers in immunology* 11 (2020).

[57] Mihaela Oprea, Erik Van Nimwegen, and Alan S Perelson. "Dynamics of one-pass germinal center models: implications for affinity maturation". In: *Bulletin of mathematical biology* 62.1 (2000), pp. 121–153.

[58] Thomas B Kepler and Alan S Perelson. “Cyclic re-entry of germinal center B cells and the efficiency of affinity maturation”. In: *Immunology today* 14.8 (1993), pp. 412–415.

[59] Ian CM MacLennan. “Germinal centers”. In: *Annual review of immunology* 12.1 (1994), pp. 117–139.

[60] Can Keşmir and Rob J De Boer. “A mathematical model on germinal center kinetics and termination”. In: *The Journal of Immunology* 163.5 (1999), pp. 2463–2469.

[61] Mihaela Oprea and Alan S Perelson. “Somatic mutation leads to efficient affinity maturation when centrocytes recycle back to centroblasts.” In: *The Journal of Immunology* 158.11 (1997), pp. 5155–5162.

[62] Michael Meyer-Hermann, Andreas Deutsch, and Michal Or-Guil. “Recycling probability and dynamical properties of germinal center reactions”. In: *Journal of Theoretical Biology* 210.3 (2001), pp. 265–285.

[63] Marie H Kosco-Vilbois. “Are follicular dendritic cells really good for nothing?” In: *Nature Reviews Immunology* 3.9 (2003), pp. 764–769.

[64] Ann M Haberman and Mark J Shlomchik. “Reassessing the function of immune-complex retention by follicular dendritic cells”. In: *Nature Reviews Immunology* 3.9 (2003), pp. 757–764.

[65] Gregory W Siskind and Baruj Benacerraf. “Cell selection by antigen in the immune response”. In: *Advances in immunology* 10 (1969), pp. 1–50.

[66] Dagmar Iber and Philip K Maini. “A mathematical model for germinal centre kinetics and affinity maturation”. In: *Journal of theoretical biology* 219.2 (2002), pp. 153–175.

[67] Can Keşmir and Rob J De Boer. "A spatial model of germinal center reactions: cellular adhesion based sorting of B cells results in efficient affinity maturation". In: *Journal of Theoretical Biology* 222.1 (2003), pp. 9–22.

[68] Michael E Meyer-Hermann, Philip K Maini, and Dagmar Iber. "An analysis of B cell selection mechanisms in germinal centers". In: *Mathematical medicine and biology: a journal of the IMA* 23.3 (2006), pp. 255–277.

[69] Michael Meyer–Hermann. "A concerted action of B cell selection mechanisms". In: *Advances in Complex Systems* 10.04 (2007), pp. 557–580.

[70] Huifeng Niu, H Ye Bihui, and Riccardo Dalla-Favera. "Antigen receptor signaling induces MAP kinase-mediated phosphorylation and degradation of the BCL-6 transcription factor". In: *Genes & development* 12.13 (1998), pp. 1953–1961.

[71] Roger Sciammas, AL Shaffer, Jonathan H Schatz, Hong Zhao, Louis M Staudt, and Harinder Singh. "Graded expression of interferon regulatory factor-4 coordinates isotype switching with plasma cell differentiation". In: *Immunity* 25.2 (2006), pp. 225–236.

[72] Enze Liu, Lang Li, and Lijun Cheng. "Gene Regulatory Network Review". In: (2019).

[73] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene Van Someren, and Reinhard Guthke. "Gene regulatory network inference: data integration in dynamic models—a review". In: *Biosystems* 96.1 (2009), pp. 86–103.

[74] Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. "Reverse engineering of regulatory networks in human B cells". In: *Nature genetics* 37.4 (2005), pp. 382–390.

[75] Piyush B Madhamshettiwar, Stefan R Maetschke, Melissa J Davis, Antonio Reverter, and Mark A Ragan. "Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets". In: *Genome medicine* 4.5 (2012), pp. 1–16.

[76] T tempspacetempspaceS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. "Human protein reference database—2009 update". In: *Nucleic acids research* 37.suppl_1 (2009), pp. D767–D772.

[77] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, et al. "The IntAct molecular interaction database in 2012". In: *Nucleic acids research* 40.D1 (2012), pp. D841–D846.

[78] Randall C Willis and Christopher WV Hogue. "Searching, viewing, and visualizing data in the Biomolecular Interaction Network Database (BIND)". In: *Current protocols in bioinformatics* 12.1 (2005), pp. 8–9.

[79] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. "STRING v10: protein–protein

interaction networks, integrated over the tree of life". In: *Nucleic acids research* 43.D1 (2015), pp. D447–D452.

[80] Dominic Schmidt, Michael D Wilson, Christiana Spyrou, Gordon D Brown, James Hadfield, and Duncan T Odom. "ChIP-seq: using high-throughput sequencing to discover protein–DNA interactions". In: *Methods* 48.3 (2009), pp. 240–248.

[81] Yuji Zhang. "Gene regulatory networks: Real data sources and their analysis". In: *Evolutionary Computation in Gene Regulatory Network Research* (2016), p. 49.

[82] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks". In: *Frontiers in cell and developmental biology* 2 (2014), p. 38.

[83] Geng Chen, Baitang Ning, and Tieliu Shi. "Single-cell RNA-Seq technologies and related computational data analysis". In: *Frontiers in genetics* 10 (2019), p. 317.

[84] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. "mRNA-Seq whole-transcriptome analysis of a single cell". In: *Nature methods* 6.5 (2009), pp. 377–382.

[85] Robert Roth, Soochi Kim, Jeesu Kim, and Siyeon Rhee. "Single-cell and spatial transcriptomics approaches of cardiovascular development and disease". In: *BMB reports* 53.8 (2020), p. 393.

[86] Fan Guo, Lin Li, Jingyun Li, Xinglong Wu, Boqiang Hu, Ping Zhu, Lu Wen, and Fuchou Tang. "Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells". In: *Cell research* 27.8 (2017), pp. 967–988.

[87] MH Julius, T Masuda, and LA Herzenberg. "Demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter". In: *Proceedings of the National Academy of Sciences* 69.7 (1972), pp. 1934–1938.

[88] George M Whitesides. "The origins and the future of microfluidics". In: *nature* 442.7101 (2006), pp. 368–373.

[89] Frederick K Balagaddé, Lingchong You, Carl L Hansen, Frances H Arnold, and Stephen R Quake. "Long-term monitoring of bacteria undergoing programmed population control in a microchemostat". In: *Science* 309.5731 (2005), pp. 137–140.

[90] Joshua S Marcus, W French Anderson, and Stephen R Quake. "Microfluidic single-cell mRNA isolation and analysis". In: *Analytical chemistry* 78.9 (2006), pp. 3084–3089.

[91] Todd Thorsen, Richard W Roberts, Frances H Arnold, and Stephen R Quake. "Dynamic pattern formation in a vesicle-generating microfluidic device". In: *Physical review letters* 86.18 (2001), p. 4163.

[92] Caixia Gao, Mingnan Zhang, and Lei Chen. "The comparison of two single-cell sequencing platforms: BD rhapsody and 10x genomics chromium". In: *Current Genomics* 21.8 (2020), pp. 602–609.

[93] Peter See, Josephine Lum, Jinmiao Chen, and Florent Ginhoux. "A single-cell sequencing guide for immunologists". In: *Frontiers in immunology* 9 (2018), p. 2425.

[94] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. "Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq". In: *Genome research* 21.7 (2011), pp. 1160–1167.

[95] Nicholas Lytal, Di Ran, and Lingling An. "Normalization methods on single-cell RNA-seq data: an empirical survey". In: *Frontiers in genetics* 11 (2020), p. 41.

[96] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. "Spatial reconstruction of single-cell gene expression data". In: *Nature biotechnology* 33.5 (2015), pp. 495–502.

[97] Shintaro Katayama, Virpi Töhönen, Sten Linnarsson, and Juha Kere. "SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization". In: *Bioinformatics* 29.22 (2013), pp. 2943–2945.

[98] Catalina A Vallejos, John C Marioni, and Sylvia Richardson. "BASiCS: Bayesian analysis of single-cell sequencing data". In: *PLoS Comput Biol* 11.6 (2015), e1004333.

[99] Bo Ding, Lina Zheng, Yun Zhu, Nan Li, Haiyang Jia, Rizi Ai, Andre Wildberg, and Wei Wang. "Normalization and noise reduction for single cell RNA-seq experiments". In: *Bioinformatics* 31.13 (2015), pp. 2225–2227.

[100] Aaron TL Lun, Karsten Bach, and John C Marioni. "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts". In: *Genome biology* 17.1 (2016), pp. 1–14.

[101] Rhonda Bacher, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski. "SCnorm: robust normalization of single-cell RNA-seq data". In: *Nature methods* 14.6 (2017), p. 584.

[102] Shun H Yip, Panwen Wang, Jean-Pierre A Kocher, Pak Chung Sham, and Junwen Wang. "Linnorm: improved statistical analysis for single cell RNA-seq expression data". In: *Nucleic acids research* 45.22 (2017), e179–e179.

[103] Montgomery Blencowe, Douglas Arneson, Jessica Ding, Yen-Wei Chen, Zara Saleem, and Xia Yang. "Network modeling of single-cell omics data: challenges, opportunities, and progresses". In: *Emerging topics in life sciences* 3.4 (2019), pp. 379–398.

[104] Shuonan Chen and Jessica C Mar. "Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data". In: *BMC bioinformatics* 19.1 (2018), pp. 1–21.

[105] Marco Scutari. "Learning Bayesian networks with the bnlearn R package". In: *arXiv preprint arXiv:0908.3817* (2009).

[106] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a

mammalian cellular context". In: *BMC bioinformatics*. Vol. 7. 1. Springer. 2006, pp. 1–15.

[107] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. "Inferring regulatory networks from expression data using tree-based methods". In: *PloS one* 5.9 (2010), pp. 1–10.

[108] Ann C Babtie and Michael PH Stumpf. "How to deal with parameters for whole-cell modelling". In: *Journal of The Royal Society Interface* 14.133 (2017), p. 20170237.

[109] Helena Todorov, Robrecht Cannoodt, Wouter Saelens, and Yvan Saeys. "Network inference from single-cell transcriptomic data". In: *Gene regulatory networks*. Springer, 2019, pp. 235–249.

[110] David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. "Recovering gene interactions from single-cell data using data diffusion". In: *Cell* 174.3 (2018), pp. 716–729.

[111] Wei Vivian Li and Jingyi Jessica Li. "An accurate and robust imputation method scImpute for single-cell RNA-seq data". In: *Nature communications* 9.1 (2018), pp. 1–9.

[112] Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa, and Daniel J Garry. "DrImpute: imputing dropout events in single cell RNA sequencing data". In: *BMC bioinformatics* 19.1 (2018), pp. 1–10.

[113] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy

R Zhang. “SAVER: gene expression recovery for single-cell RNA sequencing”. In: *Nature methods* 15.7 (2018), pp. 539–542.

[114] Lingxue Zhu, Jing Lei, Bernie Devlin, and Kathryn Roeder. “A unified statistical framework for single cell and bulk RNA sequencing data”. In: *The annals of applied statistics* 12.1 (2018), p. 609.

[115] Lihua Zhang and Shihua Zhang. “PBLR: an accurate single cell RNA-seq data imputation tool considering cell heterogeneity and prior expression level of dropouts”. In: *bioRxiv* (2018), p. 379883.

[116] Sandhya Prabhakaran, Elham Azizi, Ambrose Carr, and Dana Pe’er. “Dirichlet process mixture model for correcting technical variation in single-cell gene expression data”. In: *International Conference on Machine Learning*. PMLR. 2016, pp. 1070–1079.

[117] Cédric Arisdakessian, Olivier Poirion, Breck Yunits, Xun Zhu, and Lana X Garmire. “DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data”. In: *Genome biology* 20.1 (2019), pp. 1–14.

[118] Lihua Zhang and Shihua Zhang. “Comparison of computational methods for imputing single-cell RNA-sequencing data”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 17.2 (2018), pp. 376–389.

[119] Emma Pierson and Christopher Yau. “ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis”. In: *Genome biology* 16.1 (2015), pp. 1–10.

[120] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. "A general and flexible method for signal extraction from single-cell RNA-seq data". In: *Nature communications* 9.1 (2018), pp. 1–17.

[121] 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". In: *Nucleic acids research* 43.7 (2015), e47–e47.

[122] Yuqing Zhang, Giovanni Parmigiani, and W Evan Johnson. "ComBat-Seq: batch effect adjustment for RNA-Seq count data". In: *NAR genomics and bioinformatics* 2.3 (2020), lqaa078.

[123] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors". In: *Nature biotechnology* 36.5 (2018), pp. 421–427.

[124] Maren Büttner, Zhichao Miao, F Alexander Wolf, Sarah A Teichmann, and Fabian J Theis. "Assessment of batch-correction methods for scRNA-seq data with a new test metric". In: *BioRxiv* (2017), p. 200345.

[125] Vladimir Yu Kiselev and Martin Hemberg. "scmap-A tool for unsupervised projection of single cell RNA-seq data". In: *bioRxiv* (2017), p. 150292.

[126] Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al. "Reference-based analysis of lung single-cell sequencing

reveals a transitional profibrotic macrophage". In: *Nature immunology* 20.2 (2019), pp. 163–172.

[127] Laleh Haghverdi, Florian Buettner, and Fabian J Theis. "Diffusion maps for high-dimensional single-cell analysis of differentiation data". In: *Bioinformatics* 31.18 (2015), pp. 2989–2998.

[128] Victoria Moignard, Steven Woodhouse, Laleh Haghverdi, Andrew J Lilly, Yosuke Tanaka, Adam C Wilkinson, Florian Buettner, Iain C Macaulay, Wajid Jawaid, Evangelia Diamanti, et al. "Decoding the regulatory network of early blood development from single-cell gene expression measurements". In: *Nature biotechnology* 33.3 (2015), pp. 269–276.

[129] Chee Yee Lim, Huange Wang, Steven Woodhouse, Nir Piterman, Lorenz Wernisch, Jasmin Fisher, and Berthold Göttgens. "BTR: training asynchronous Boolean models using single-cell expression data". In: *BMC bioinformatics* 17.1 (2016), pp. 1–18.

[130] Minzhe Guo, Hui Wang, S Steven Potter, Jeffrey A Whitsett, and Yan Xu. "SINCERA: a pipeline for single-cell RNA-Seq profiling analysis". In: *PLoS computational biology* 11.11 (2015), e1004575.

[131] Philip A Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. "The nf-core framework for community-curated bioinformatics pipelines". In: *Nature biotechnology* 38.3 (2020), pp. 276–278.

[132] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, et al. "SCENIC: single-cell regulatory network inference and clustering". In: *Nature methods* 14.11 (2017), pp. 1083–1086.

[133] Haodong Liu, Peng Li, Mengyao Zhu, Xiaofei Wang, Jianwei Lu, and Tianwei Yu. "Nonlinear network reconstruction from gene expression data using marginal dependencies measured by DCOL". In: *PloS one* 11.7 (2016), e0158247.

[134] Steven Woodhouse, Nir Piterman, Christoph M Wintersteiger, Berthold Göttgens, and Jasmin Fisher. "SCNS: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data". In: *BMC systems biology* 12.1 (2018), pp. 1–7.

[135] Haifen Chen, Jing Guo, Shital K Mishra, Paul Robson, Mahesan Niranjan, and Jie Zheng. "Single-cell transcriptional analysis to uncover regulatory circuits driving cell fate decisions in early mouse development". In: *Bioinformatics* 31.7 (2015), pp. 1060–1066.

[136] Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru SH Ko, Shigeru BH Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. "SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation". In: *Bioinformatics* 33.15 (2017), pp. 2314–2321.

[137] Andrea Ocone, Laleh Haghverdi, Nikola S Mueller, and Fabian J Theis. "Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data". In: *Bioinformatics* 31.12 (2015), pp. i89–i96.

[138] Arnaud Bonnaffoux, Ulysse Herbach, Angélique Richard, Anissa Guillemin, Sandrine Gonin-Giraud, Pierre-Alexis Gros, and Olivier Gandrillon. "WASABI: a dynamic iterative framework for gene regulatory network inference". In: *BMC bioinformatics* 20.1 (2019), pp. 1–19.

[139] Nan Papili Gao, SM Minhaz Ud-Dean, Olivier Gandrillon, and Rudiyanto Gunawan. "SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles". In: *Bioinformatics* 34.2 (2018), pp. 258–266.

[140] Junxiang Li, Haofei Luo, Rui Wang, Jidong Lang, Siyu Zhu, Zhenming Zhang, Jianhuo Fang, Keke Qu, Yuting Lin, Haizhou Long, et al. "Systematic reconstruction of molecular cascades regulating GP development using single-cell RNA-seq". In: *Cell reports* 15.7 (2016), pp. 1467–1480.

[141] Hirotaka Matsumoto and Hisanori Kiryu. "SCOUP: a probabilistic model based on the Ornstein–Uhlenbeck process to analyze single-cell expression data during differentiation". In: *BMC bioinformatics* 17.1 (2016), pp. 1–16.

[142] Pablo Cordero and Joshua M Stuart. "Tracing co-regulatory network dynamics in noisy, single-cell transcriptome trajectories". In: *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*. World Scientific. 2017, pp. 576–587.

[143] Manuel Sanchez-Castillo, David Blanco, Isabel M Tienda-Luna, MC Carrion, and Yufei Huang. "A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data". In: *Bioinformatics* 34.6 (2018), pp. 964–970.

[144] Thalia E Chan, Michael PH Stumpf, and Ann C Babtie. "Gene regulatory network inference from single-cell data using multivariate information measures". In: *Cell systems* 5.3 (2017), pp. 251–267.

[145] Alexandre Irrthum, Louis Wehenkel, Pierre Geurts, et al. "Inferring regulatory networks from expression data using tree-based methods". In: *PloS one* 5.9 (2010), e12776.

[146] Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. "Real-time kinetics of gene activity in individual bacteria". In: *Cell* 123.6 (2005), pp. 1025–1036.

[147] Long Cai, Nir Friedman, and X Sunney Xie. "Stochastic protein expression in individual cells at the single molecule level". In: *Nature* 440.7082 (2006), pp. 358–362.

[148] Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. "Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells". In: *science* 329.5991 (2010), pp. 533–538.

[149] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. "Single-RNA counting reveals alternative modes of gene expression in yeast". In: *Nature structural & molecular biology* 15.12 (2008), p. 1263.

[150] Rui Zhen Tan and Alexander Van Oudenaarden. "Transcript counting in single cells reveals dynamics of rDNA transcription". In: *Molecular systems biology* 6.1 (2010), p. 358.

[151] Long Cai, Chiraj K Dalal, and Michael B Elowitz. "Frequency-modulated nuclear localization bursts coordinate gene regulation". In: *Nature* 455.7212 (2008), pp. 485–490.

[152] Arjun Raj, Scott A Rifkin, Erik Andersen, and Alexander Van Oudenaarden. "Variability in gene expression underlies incomplete penetrance". In: *Nature* 463.7283 (2010), pp. 913–918.

[153] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. "Stochastic mRNA synthesis in mammalian cells". In: *PLoS Biol* 4.10 (2006), e309.

[154] Jonathan R Chubb, Tatjana Trcek, Shailesh M Shenoy, and Robert H Singer. "Transcriptional pulsing of a developmental gene". In: *Current biology* 16.10 (2006), pp. 1018–1025.

[155] David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. "Mammalian genes are transcribed with widely different bursting kinetics". In: *Science* 332.6028 (2011), pp. 472–474.

[156] Arjun Raj and Alexander Van Oudenaarden. "Nature, nurture, or chance: stochastic gene expression and its consequences". In: *Cell* 135.2 (2008), pp. 216–226.

[157] Mark HA Davis. "Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.3 (1984), pp. 353–376.

[158] Domokos Vermes. “Optimal dynamic control of a useful class of randomly jumping processes”. In: (1980).

[159] Ulysse Herbach, Arnaud Bonnaffoux, Thibault Espinasse, and Olivier Gandrillon. “Inferring gene regulatory networks from single-cell data: a mechanistic approach”. In: *BMC systems biology* 11.1 (2017), pp. 1–15.

[160] Jonathan M Raser and Erin K O’Shea. “Control of stochasticity in eukaryotic gene expression”. In: *science* 304.5678 (2004), pp. 1811–1814.

[161] Luka Mesin, Jonatan Ersching, and Gabriel D Victora. “Germinal center B cell dynamics”. In: *Immunity* 45.3 (2016), pp. 471–482.

[162] Tomohiro Kurosaki, Kohei Kometani, and Wataru Ise. “Memory B cells”. In: *Nature Reviews Immunology* 15.3 (2015), pp. 149–159.

[163] Shuang Song and Patrick D Matthias. “The transcriptional regulation of germinal center formation”. In: *Frontiers in immunology* 9 (2018), p. 2026.

[164] Nilushi S De Silva, Giorgia Simonetti, Nicole Heise, and Ulf Klein. “The diverse roles of IRF4 in late germinal center B-cell differentiation”. In: *Immunological reviews* 247.1 (2012), pp. 73–92.

[165] Stephen L Nutt, Kirsten A Fairfax, and Axel Kallies. “BLIMP1 guides the fate of effector B and T cells”. In: *Nature Reviews Immunology* 7.12 (2007), pp. 923–927.

[166] Yi Lin, Kwok-kin Wong, and Kathryn Calame. “Repression of c-myc transcription by Blimp-1, an inducer of terminal B cell differentiation”. In: *Science* 276.5312 (1997), pp. 596–599.

[167] Daniel Radtke and Oliver Bannard. “Expression of the plasma cell transcriptional regulator Blimp-1 by dark zone germinal center B cells during periods of proliferation”. In: *Frontiers in immunology* 9 (2019), p. 3106.

[168] Julie Tellier, Wei Shi, Martina Minnich, Yang Liao, Simon Crawford, Gordon K Smyth, Axel Kallies, Meinrad Busslinger, and Stephen L Nutt. “Blimp-1 controls plasma cell function through the regulation of immunoglobulin secretion and the unfolded protein response”. In: *Nature immunology* 17.3 (2016), pp. 323–330.

[169] Martina Minnich, Hiromi Tagoh, Peter Bönelt, Elin Axelsson, Maria Fischer, Beatriz Cebolla, Alexander Tarakhovsky, Stephen L Nutt, Markus Jaritz, and Meinrad Busslinger. “Multifunctional role of the transcription factor Blimp-1 in coordinating plasma cell differentiation”. In: *Nature immunology* 17.3 (2016), pp. 331–343.

[170] Seitaro Nomura. “Single-cell genomics to understand disease pathogenesis”. In: *Journal of Human Genetics* 66.1 (2021), pp. 75–84.

[171] Cameron P Gallivan, Honglei Ren, and Elizabeth L Read. “Analysis of single-cell gene pair coexpression landscapes by stochastic kinetic modeling reveals gene-pair interactions in development”. In: *Frontiers in genetics* 10 (2020), p. 1387.

[172] Payam Dibaeinia and Saurabh Sinha. “Sergio: a single-cell expression simulator guided by gene regulatory networks”. In: *Cell Systems* 11.3 (2020), pp. 252–271.

[173] Louise J McHeyzer-Williams, Pierre J Milpied, Shinji L Okitsu, and Michael G McHeyzer-Williams. "Class-switched memory B cells remodel BCRs within secondary germinal centers". In: *Nature immunology* 16.3 (2015), pp. 296–305.

[174] Akinori Baba and Tamiki Komatsuzaki. "Construction of effective free energy landscape from single-molecule time series". In: *Proceedings of the National Academy of Sciences* 104.49 (2007), pp. 19297–19302.

[175] Jasna Medvedovic, Anja Ebert, Hiromi Tagoh, and Meinrad Busslinger. "Pax5: a master regulator of B cell development and leukemogenesis". In: *Advances in immunology* 111 (2011), pp. 179–206.

[176] Stephen L Nutt, Philip D Hodgkin, David M Tarlinton, and Lynn M Corcoran. "The generation of antibody-secreting plasma cells". In: *Nature Reviews Immunology* 15.3 (2015), pp. 160–171.

[177] Brian J Laidlaw and Jason G Cyster. "Transcriptional regulation of memory B cell differentiation". In: *Nature Reviews Immunology* 21.4 (2021), pp. 209–220.

[178] Dinis Pedro Calado, Yoshiteru Sasaki, Susana A Godinho, Alex Pellerin, Karl Köchert, Barry P Sleckman, Ignacio Moreno De Alborán, Martin Janz, Scott Rodig, and Klaus Rajewsky. "The cell-cycle regulator c-Myc is essential for the formation and maintenance of germinal centers". In: *Nature immunology* 13.11 (2012), pp. 1092–1100.

[179] Hirokazu Tanaka, Itaru Matsumura, Sachiko Ezoe, Yusuke Satoh, Toshiyuki Sakamaki, Chris Albanese, Takashi Machii, Richard G Pestell, and Yuzuru Kanakura. "E2F1 and c-Myc potentiate apoptosis through in-

hibition of NF-κB activity that facilitates MnSOD-mediated ROS elimination". In: *Molecular cell* 9.5 (2002), pp. 1017–1029.

[180] Wendy Béguelin, Martın A Rivas, Marıa T Calvo Fernández, Matt Teater, Alberto Purwada, David Redmond, Hao Shen, Matt F Challman, Olivier Elemento, Ankur Singh, et al. "EZH2 enables germinal centre formation through epigenetic silencing of CDKN1A and an Rb-E2F1 feedback loop". In: *Nature communications* 8.1 (2017), pp. 1–16.

[181] Frank M Raaphorst, Folkert J van Kemenade, Elly Fieret, Karien M Hamer, David PE Satijn, Arie P Otte, and Chris JLM Meijer. "Cutting edge: polycomb gene expression patterns reflect distinct B cell differentiation stages in human germinal centers". In: *The Journal of Immunology* 164.1 (2000), pp. 1–4.

[182] Irina Velichutina, Rita Shaknovich, Huimin Geng, Nathalie A Johnson, Randy D Gascoyne, Ari M Melnick, and Olivier Elemento. "EZH2-mediated epigenetic silencing in germinal center B cells contributes to proliferation and lymphomagenesis". In: *Blood, The Journal of the American Society of Hematology* 116.24 (2010), pp. 5247–5255.

[183] Laurie Herviou, Michel Jourdan, Anne-Marie Martinez, Giacomo Cavalli, and Jerome Moreaux. "EZH2 is overexpressed in transitional preplasmablasts and is involved in human plasma cell differentiation". In: *Leukemia* 33.8 (2019), pp. 2047–2060.

[184] Hamish W King, Nara Orban, John C Riches, Andrew J Clear, Gary Warnes, Sarah A Teichmann, and Louisa K James. "Single-cell analysis of human B cell maturation predicts how antibody class switch-

ing shapes selection dynamics". In: *Science Immunology* 6.56 (2021), eabe6291.

[185] Noemi Andor, Erin F Simonds, Debra K Czerwinski, Jiamin Chen, Susan M Grimes, Christina Wood-Bouwens, Grace XY Zheng, Matthew A Kubit, Stephanie Greer, William A Weiss, et al. "Single-cell RNA-Seq of follicular lymphoma reveals malignant B-cell types and coexpression of T-cell immune checkpoints". In: *Blood, The Journal of the American Society of Hematology* 133.10 (2019), pp. 1119–1129.

[186] Antony B Holmes, Clarissa Corinaldesi, Qiong Shen, Rahul Kumar, Nicolo Compagno, Zhong Wang, Mor Nitzan, Eli Grunstein, Laura Pasqualucci, Riccardo Dalla-Favera, et al. "Single-cell analysis of germinal-center B cells informs on lymphoma cell of origin and outcome". In: *Journal of Experimental Medicine* 217.10 (2020).

[187] Joep Vanlier, Christian A Tiemann, Peter AJ Hilbers, and Natal AW van Riel. "An integrated strategy for prediction uncertainty analysis". In: *Bioinformatics* 28.8 (2012), pp. 1130–1135.

[188] Ekaterina Myasnikova and Alexander Spirov. "Relative sensitivity analysis of the predictive properties of sloppy models". In: *Journal of bioinformatics and computational biology* 16.02 (2018), p. 1840008.

[189] Attila Gábor, Alejandro F Villaverde, and Julio R Banga. "Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems". In: *BMC systems biology* 11.1 (2017), pp. 1–16.

[190] Roland Brun, Peter Reichert, and Hans R Künsch. "Practical identifiability analysis of large environmental simulation models". In: *Water Resources Research* 37.4 (2001), pp. 1015–1030.

[191] Stefan Hengl, Clemens Kreutz, Jens Timmer, and Thomas Maiwald. "Data-based identifiability analysis of non-linear dynamical models". In: *bioinformatics* 23.19 (2007), pp. 2612–2618.

[192] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood". In: *Bioinformatics* 25.15 (2009), pp. 1923–1929.

[193] Alexander P Browning, David J Warne, Kevin Burrage, Ruth E Baker, and Matthew J Simpson. "Identifiability analysis for stochastic differential equation models in systems biology". In: *Journal of the Royal Society Interface* 17.173 (2020), p. 20200652.

[194] Keegan E Hines, Thomas R Middendorf, and Richard W Aldrich. "Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach". In: *Journal of General Physiology* 143.3 (2014), pp. 401–416.

[195] Ivo Siekmann, James Sneyd, and Edmund J Crampin. "MCMC can detect nonidentifiable models". In: *Biophysical journal* 103.11 (2012), pp. 2275–2286.

[196] Giuseppina Bellu, Maria Pia Saccomani, Stefania Audoly, and Leontina D'Angiò. "DAISY: A new software tool to test global identifiability of

biological and physiological systems". In: *Computer methods and programs in biomedicine* 88.1 (2007), pp. 52–61.

[197] Chiara Damiani, Alessandro Filisetti, Alex Graudenzi, and Paola Lecca. "Parameter sensitivity analysis of stochastic models: Application to catalytic reaction networks". In: *Computational biology and chemistry* 42 (2013), pp. 5–17.

[198] Michał Komorowski, Maria J Costa, David A Rand, and Michael PH Stumpf. "Sensitivity, robustness, and identifiability in stochastic chemical kinetics models". In: *Proceedings of the National Academy of Sciences* 108.21 (2011), pp. 8645–8650.

[199] Rudiyanto Gunawan, Yang Cao, Linda Petzold, and Francis J Doyle III. "Sensitivity analysis of discrete stochastic systems". In: *Biophysical journal* 88.4 (2005), pp. 2530–2540.

[200] Muruhan Rathinam, Patrick W Sheppard, and Mustafa Khammash. "Efficient computation of parameter sensitivities of discrete stochastic chemical reaction networks". In: *The Journal of chemical physics* 132.3 (2010), p. 034103.

[201] Sergey Plyasunov and Adam P Arkin. "Efficient stochastic sensitivity analysis of discrete event systems". In: *Journal of Computational Physics* 221.2 (2007), pp. 724–738.

[202] David F Anderson. "An efficient finite difference method for parameter sensitivities of continuous time Markov chains". In: *SIAM Journal on Numerical Analysis* 50.5 (2012), pp. 2237–2258.

[203] Vo Hong Thanh, Roberto Zunino, and Corrado Priami. "Efficient finite-difference method for computing sensitivities of biochemical reactions". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 474.2218 (2018), p. 20180303.

[204] Eric A Sobie. "Parameter sensitivity analysis in electrophysiological models using multivariable regression". In: *Biophysical journal* 96.4 (2009), pp. 1264–1274.

[205] Amrita X Sarkar and Eric A Sobie. "Quantification of repolarization reserve to understand interpatient variability in the response to proarrhythmic drugs: a computational analysis". In: *Heart Rhythm* 8.11 (2011), pp. 1749–1755.

[206] Young-Seon Lee, Ona Z Liu, Hyun Seok Hwang, Bjorn C Knollmann, and Eric A Sobie. "Parameter sensitivity analysis of stochastic models provides insights into cardiac calcium sparks". In: *Biophysical journal* 104.5 (2013), pp. 1142–1150.

[207] Tomer Kalisky, Paul Blainey, and Stephen R Quake. "Genomic analysis at the single-cell level". In: *Annual review of genetics* 45 (2011), pp. 431–445.

[208] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. "Single-cell RNA-seq: advances and future challenges". In: *Nucleic acids research* 42.14 (2014), pp. 8845–8860.

[209] Angélique Richard, Lois Boullu, Ulysse Herbach, Arnaud Bonnafoux, Valérie Morin, Elodie Vallin, Anissa Guillemin, Nan Papili Gao, Rudiyanto Gunawan, Jérémie Cosette, et al. "Single-cell-based analysis highlights

a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process". In: *PLoS biology* 14.12 (2016), e1002585.

[210] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. "Comparative analysis of single-cell RNA sequencing methods". In: *Molecular cell* 65.4 (2017), pp. 631–643.

[211] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Ines Hellmann, and Wolfgang Enard. "Quantitative single-cell transcriptomics". In: *Briefings in functional genomics* 17.4 (2018), pp. 220–232.

[212] Asif Adil, Vijay Kumar, Arif Tasleem Jan, and Mohammed Asger. "Single-cell transcriptomics: Current methods and challenges in data acquisition and analysis". In: *Frontiers in Neuroscience* 15 (2021), p. 398.

[213] Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. "Power analysis of single-cell RNA-sequencing experiments". In: *Nature methods* 14.4 (2017), pp. 381–387.

[214] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. "Normalization of RNA-seq data using factor analysis of control genes or samples". In: *Nature biotechnology* 32.9 (2014), pp. 896–902.

[215] Luis A Corchete, Elizabeta A Rojas, Diego Alonso-López, Javier De Las Rivas, Norma C Gutiérrez, and Francisco J Burguillo. "Systematic

comparison and assessment of RNA-seq procedures for gene expression quantitative analysis". In: *Scientific reports* 10.1 (2020), pp. 1–15.

[216] SEQC Consortium et al. "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium". In: *Nature biotechnology* 32.9 (2014), p. 903.

[217] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. "Fast, sensitive and accurate integration of single-cell data with Harmony". In: *Nature methods* 16.12 (2019), pp. 1289–1296.

[218] Katia Basso. "Biology of Germinal Center B Cells Relating to Lymphomagenesis". In: *HemaSphere* 5.6 (2021).

[219] Gabriel D Victora, Tanja A Schwickert, David R Fooksman, Alice O Kamphorst, Michael Meyer-Hermann, Michael L Dustin, and Michel C Nussenzweig. "Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter". In: *Cell* 143.4 (2010), pp. 592–605.

[220] Joost B Beltman, Christopher DC Allen, Jason G Cyster, and Rob J de Boer. "B cells within germinal centers migrate preferentially from dark to light zone". In: *Proceedings of the National Academy of Sciences* 108.21 (2011), pp. 8755–8760.

[221] Takuya Koike, Koshi Harada, Shu Horiuchi, and Daisuke Kitamura. "The quantity of CD40 signaling determines the differentiation of B

cells into functionally distinct memory cell subsets". In: *Elife* 8 (2019), e44245.

[222] Juan Carlos Yam-Puc, Lingling Zhang, Raul A Maqueda-Alfaro, Laura Garcia-Ibanez, Yang Zhang, Jessica Davies, Yotis A Senis, Michael Snaith, and Kai-Michael Toellner. "Enhanced BCR signaling inflicts early plasmablast and germinal center B cell death". In: *Iscience* 24.2 (2021), p. 102038.

[223] Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M Garske, Jae Hoon Sul, Kirsi H Pietiläinen, Päivi Pajukanta, and Eran Halperin. "Accurate estimation of cell composition in bulk expression through robust integration of single-cell information". In: *Nature communications* 11.1 (2020), pp. 1–11.

[224] Kuo-I Lin, Cristina Angelin-Duclos, Tracy C Kuo, and Kathryn Calame. "Blimp-1-dependent repression of Pax-5 is required for differentiation of B cells to immunoglobulin M-secreting plasma cells". In: *Molecular and cellular biology* 22.13 (2002), pp. 4771–4780.

[225] Yuichi Miura, Mizuho Morooka, Nicolas Sax, Rahul Roychoudhuri, Ari Itoh-Nakadai, Andrey Brydun, Ryo Funayama, Keiko Nakayama, Susumu Satomi, Mitsuyo Matsumoto, et al. "Bach2 promotes B cell receptor–induced proliferation of B lymphocytes and represses cyclin-dependent kinase inhibitors". In: *The Journal of Immunology* 200.8 (2018), pp. 2882–2893.

[226] German Ott, Andreas Rosenwald, and Elias Campo. "Understanding MYC-driven aggressive B-cell lymphomas: pathogenesis and classifica-

tion". In: *Blood, The Journal of the American Society of Hematology* 122.24 (2013), pp. 3884–3891.

[227] Genevieve L Stein-O'Brien, Michaela C Ainsile, and Elana J Fertig. "Forecasting cellular states: from descriptive to predictive biology via single cell multi-omics". In: *Current Opinion in Systems Biology* (2021).

[228] Genevieve L Stein-O'Brien, Raman Arora, Aedin C Culhane, Alexander V Favorov, Lana X Garmire, Casey S Greene, Loyal A Goff, Yifeng Li, Aloune Ngom, Michael F Ochs, et al. "Enter the matrix: factorization uncovers knowledge from omics". In: *Trends in Genetics* 34.10 (2018), pp. 790–805.

[229] Giovanni Palla and Enrico Ferrero. "Latent Factor Modeling of scRNA-Seq Data Uncovers Dysregulated Pathways in Autoimmune Disease Patients". In: *Iscience* 23.9 (2020), p. 101451.

[230] Shiquan Sun, Jiaqiang Zhu, Ying Ma, and Xiang Zhou. "Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis". In: *Genome biology* 20.1 (2019), pp. 1–21.

[231] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[232] Kevin R Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. "Visualizing structure and transitions in high-dimensional biological data". In: *Nature biotechnology* 37.12 (2019), pp. 1482–1492.

[233] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[234] Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel JT Reinders, and Ahmed Mahfouz. "A comparison of automatic cell identification methods for single-cell RNA sequencing data". In: *Genome biology* 20.1 (2019), pp. 1–19.

[235] Praneet Chatuverdi, Aaron Zorn, and Konrad Thorner. "ELeFHAnt: A supervised machine learning approach for label harmonization and annotation of single cell RNA-seq data". In: *bioRxiv* (2021).

[236] Liang Chen, Yuyao Zhai, Qiuyan He, Weinan Wang, and Minghua Deng. "Integrating deep supervised, self-supervised and unsupervised learning for single-cell RNA-seq clustering and annotation". In: *Genes* 11.7 (2020), p. 792.

[237] Félix Raimundo, Laetitia Papaxanthos, Céline Vallot, and Jean-Philippe Vert. "Machine learning for single cell genomics data analysis". In: *Current Opinion in Systems Biology* (2021).

[238] Abel Szkalisity, Filippo Piccinini, Attila Beleon, Tamas Balassa, Istvan Gergely Varga, Ede Migh, Csaba Molnar, Lassi Paavolainen, Sanna Timonen, Indranil Banerjee, et al. "Regression plane concept for analysing continuous cellular processes with machine learning". In: *Nature communications* 12.1 (2021), pp. 1–9.

[239] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. "Interpretable machine learning: Fundamental

principles and 10 grand challenges". In: *arXiv preprint arXiv:2103.11251* (2021).

[240] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. "Definitions, methods, and applications in interpretable machine learning". In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080.

[241] Lifei Wang, Rui Nie, Zeyang Yu, Ruyue Xin, Caihong Zheng, Zhang Zhang, Jiang Zhang, and Jun Cai. "An interpretable deep-learning architecture of capsule networks for identifying cell-type gene expression programs from single-cell RNA-sequencing data". In: *Nature Machine Intelligence* 2.11 (2020), pp. 693–703.

[242] Gabriel Torregrosa and Jordi Garcia-Ojalvo. "Mechanistic models of cell-fate transitions from single-cell data". In: *Current Opinion in Systems Biology* (2021).

[243] Kyle Akers and TM Murali. "Gene regulatory network inference in single cell biology". In: *Current Opinion in Systems Biology* (2021).

[244] Mirjana Efremova, Miquel Vento-Tormo, Sarah A Teichmann, and Roser Vento-Tormo. "CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes". In: *Nature protocols* 15.4 (2020), pp. 1484–1506.

[245] Hung Nguyen, Duc Tran, Bang Tran, Bahadir Pehlivan, and Tin Nguyen. "A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data". In: *Briefings in bioinformatics* 22.3 (2021), bbaa190.

[246] Robin Browaeys, Wouter Saelens, and Yvan Saeys. "NicheNet: modeling intercellular communication by linking ligands to target genes". In: *Nature methods* 17.2 (2020), pp. 159–162.

[247] Matthew A Kukurugya, Caroll M Mendonca, Mina Solhtalab, Rebecca A Wilkes, Theodore W Thannhauser, and Ludmilla Aristilde. "Multi-omics analysis unravels a segregated metabolic flux network that tunes co-utilization of sugar and aromatic carbons in Pseudomonas putida". In: *Journal of Biological Chemistry* 294.21 (2019), pp. 8464–8479.

[248] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. "Multi-omics data integration, interpretation, and its application". In: *Bioinformatics and biology insights* 14 (2020), p. 1177932219899051.

[249] Anupam Chowdhury and Costas D Maranas. "Personalized Kinetic Models for Predictive Healthcare". In: *Cell systems* 1.4 (2015), pp. 250–251.

[250] Fei He, Ettore Murabito, and Hans V Westerhoff. "Synthetic biology and regulatory networks: where metabolic systems biology meets control engineering". In: *Journal of The Royal Society Interface* 13.117 (2016), p. 20151046.

[251] Fei He and Michael PH Stumpf. "Quantifying dynamic regulation in metabolic pathways with nonparametric flux inference". In: *Biophysical journal* 116.10 (2019), pp. 2035–2046.

[252] Mansoor Saqi, Johann Pellet, Irina Roznovat, Alexander Mazein, Stéphane Ballereau, Bertrand De Meulder, and Charles Auffray. "Systems medicine:

the future of medical genomics, healthcare, and wellness". In: *Systems Medicine* (2016), pp. 43–60.

[253] Angélique Stéphanou, Eric Fanchon, Pasquale F Innominato, and Annabelle Ballesta. "Systems biology, systems medicine, systems pharmacology: the what and the why". In: *Acta biotheoretica* 66.4 (2018), pp. 345–365.

[254] Sebastian Schleidgen, Sandra Fernau, Henrike Fleischer, Christoph Schickhardt, Ann-Kristin Oßa, and Eva C Winkler. "Applying systems biology to biomedical research and health care: a précising definition of systems medicine". In: *BMC health services research* 17.1 (2017), pp. 1–16.

[255] Priya Tolani, Srishti Gupta, Kirti Yadav, Suruchi Aggarwal, and Amit Kumar Yadav. "Big data, integrative omics and network biology". In: (2021).

[256] Rui Chen, George I Mias, Jennifer Li-Pook-Than, Lihua Jiang, Hugo YK Lam, Rong Chen, Elana Miriami, Konrad J Karczewski, Manoj Hariharan, Frederick E Dewey, et al. "Personal omics profiling reveals dynamic molecular and medical phenotypes". In: *Cell* 148.6 (2012), pp. 1293–1307.

[257] Gregory L Szeto and Stacey D Finley. "Integrative approaches to cancer immunotherapy". In: *Trends in cancer* 5.7 (2019), pp. 400–410.

[258] Robert Clarke, John J Tyson, Ming Tan, William T Baumann, Lu Jin, Jianhua Xuan, and Yue Wang. "Systems biology: perspectives on multiscale modeling in research on endocrine-related cancers". In: *Endocrine-related cancer* 26.6 (2019), R345–R368.

Supplementary Material

SM1 Mathematical reduction of the PDMP model and estimation of $k_{on,init}$

From Martinez et al. [2], one can see that IRF4 is a crucial node and its autoactivation reaction is responsible for the bistability switch from the GC to PB_PC during B cell differentiation. At the same time, equation (2.3) which describes the dynamics of IRF4 only depends on IRF4. Based on those observations, we have decided to use IRF4 as connecting edge between models (2.4)-(2.6) and (2.1)-(2.3).

Starting from the stochastic PDMP model (2.4)-(2.6) written for IRF4, we reduced it to an ODE version by making a simplifying assumption. We substituted of the stochastic process $E(t)$ by its mean value $\langle E(t) \rangle$

$$\begin{cases} \frac{dM_{IRF4}}{dt} = s_{0,\,IRF4}\langle E_{IRF4}(t)\rangle - d_{0,\,IRF4}M_{IRF4}(t) \\ \frac{dP_{IRF4}}{dt} = s_{1,\,IRF4}M_{IRF4}(t) - d_{1,\,IRF4}P_{IRF4}(t) \end{cases} \tag{SF.1}$$

We are looking for the parameter set of System (SF.1) which will allow ODE reduced PDMP (2.11) to reproduce the same behavior than the kinetic ODE model (2.1)-(2.3) (i.e. two steady states). At steady state, left part of equa-

tions of System (SF.1) equal zero, namely, $\frac{dM}{dt} = 0$ and $\frac{dP}{dt} = 0$, and after simplification of the right part of equations of System (SF.1):

$$M_{IRF4}(t) = \frac{s_{0,\,IRF4}\langle E_{IRF4}(t)\rangle}{d_{0,\,IRF4}}$$

$$P_{IRF4}(t) = \frac{s_{1,\,IRF4}s_{0,\,IRF4}\langle E_{IRF4}(t)\rangle}{d_{0,\,IRF4}d_{1,\,IRF4}} \tag{SF.2}$$

Introducing the new variable c:

$$c = \frac{s_{1,\,IRF4}s_{0,\,IRF4}}{d_{1,\,IRF4}d_{0,\,IRF4}}$$

we can write (SF.2) as:

$$P_{IRF4}(t) = c\langle E_{IRF4}(t)\rangle \tag{SF.3}$$

In Martinez et al. [2], IRF4 behavior was described by equation (2.3), that is

$$\frac{dr}{dt} = \mu_r + \sigma_r\frac{r^2}{k_r^2 + r^2} + CD40 - \lambda_r r \tag{SF.4}$$

i.e. in our notation it can be written as:

$$\frac{dp_{IRF4}}{dt} = \mu_{IRF4} + \sigma_{IRF4}\frac{p_{IRF4}^2}{k_{IRF4}^2 + p_{IRF4}^2} + CD40 - \lambda_{IRF4}\,p_{IRF4} \tag{SF.5}$$

Assuming that CD40 = 0 at the beginning of the simulation, that equation (SF.5) is at the steady state, that λ_{IRF4} is a degradation rate of protein for

IRF4 and using (SF.3), we write (SF.4) written as:

$$\mu_{IRF4} + \sigma_{IRF4}\frac{c^2(\langle E_{IRF4}(t)\rangle)^2}{k_{IRF4}^2 + c^2(\langle E_{IRF4}(t)\rangle)^2} - cd_{1,\,IRF4}\langle E_{IRF4}(t)\rangle = 0 \qquad \text{(SF.6)}$$

Equation (SF.6) can be written as:

$$cd_{1,\,IRF4}\langle E_{IRF4}(t)\rangle = \mu_{IRF4} + \sigma_{IRF4}\frac{c^2(\langle E_{IRF4}(t)\rangle)^2}{k_{IRF4}^2 + c^2(\langle E_{IRF4}(t)\rangle)^2} \qquad \text{(SF.7)}$$

Solving equation (SF.7) in terms of $\langle E_{IRF4}(t)\rangle$ leads to:

$$c^3 d_{1,\,IRF4}(\langle E_{IRF4}(t)\rangle)^3 - (\langle E_{IRF4}(t)\rangle)^2(\mu_{IRF4}c^2 + \sigma_{IRF4}c^2) +$$

$$+cd_{1,\,IRF4}k_{IRF4}^2(\langle E_{IRF4}(t)\rangle) - \mu_{IRF4}k_{IRF4}^2 = 0$$

and can be simplified in the form:

$$a'(\langle E_{IRF4}(t)\rangle)^3 - b'(\langle E_{IRF4}(t)\rangle)^2 + c'(\langle E_{IRF4}(t)\rangle) + d' = 0 \qquad \text{(SF.8)}$$

where

$$\begin{cases} a' = c^3 d_{1,\,IRF4} \\ b' = \mu_{IRF4}c^2 + \sigma_{IRF4}c^2 \\ c' = cd_{1,\,IRF4}k_{IRF4}^2 \\ d' = -\mu_{IRF4}k_{IRF4}^2 \end{cases}$$

Because parameters a', b', c' are positive and d' is negative, there is at least one positive root to (SF.8). Further, we fitted the parameters $\mu_{IRF4}, \sigma_{IRF4}, k_{IRF4}$, applying fitting procedure from Martinez et al. [2] and using the experimental data accession no. GSE 12195 (see Tables 2.2-2.5), and we found that $E_{IRF4}(t_{init})$ value which would correspond to a bistable regime of system (2.1)-(2.3) is:

$$E_{IRF4}(t_{init}) = 1.7 \times 10^{-3}$$

We also know from Section 2.3.3 that:

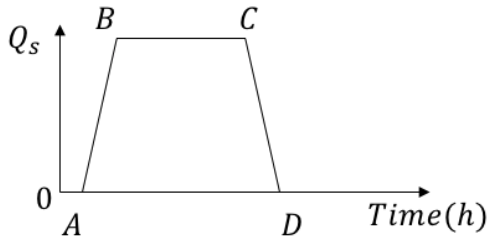
$$\langle E_{IRF4} \rangle = \frac{k_{on}}{k_{on} + k_{off}}$$

Further assuming that at initial time $t = t_{init}$, $k_{on} << k_{off}$ and $k_{on} = \alpha k_{off}$, with $\alpha << 1$, one can define $\alpha = \frac{E_{IRF4}}{1-E_{IRF4}} = 1.7 \times 10^{-3}$. Assuming $k_{off} \approx 1$ ($k_{off,\,init,\,IRF4} \approx 1$) allows to estimate the value of k_{on} for IRF4, which should keep ODE reduced PDMP model (2.11) in a two steady state regime:

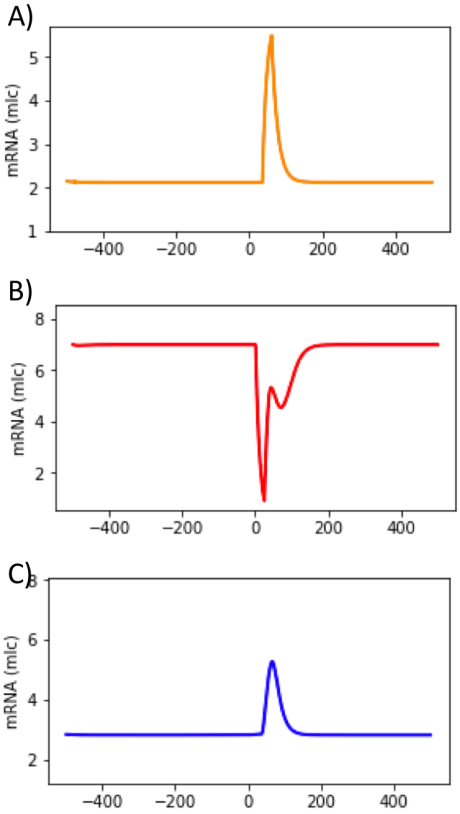
$$k_{on,\,IRF4} = 1.7 \times 10^{-3} \tag{SF.9}$$

Further, we called the value (SF.9), the initial value $k_{on,\,init}$ for IRF4.
Such procedure establish an initial guess for model (2.11) parameter set, based on the previous knowledge of ODE model (2.1)-(2.3), applied to experimental SC data of interest [3]. Thanks to the structure of the model (2.1)-(2.3), we are able to apply this strategy to estimate $k_{on,\,init,\,IRF4}$. Further, we assume that $k_{on,\,init,\,BCL6}$ and $k_{on,\,init,\,BLIMP1}$ values are in a similar range.

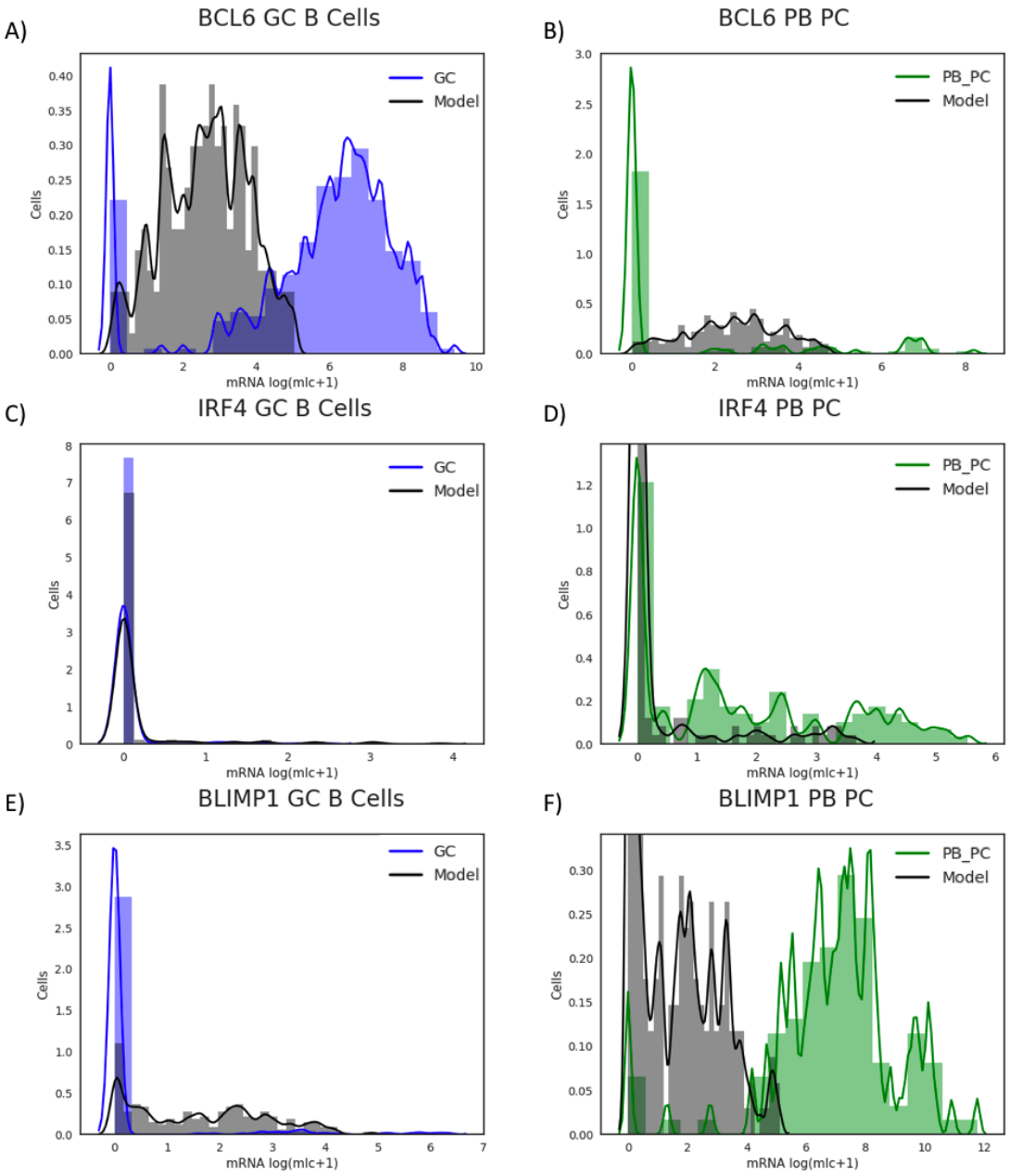
SM2 Supplementary Figures



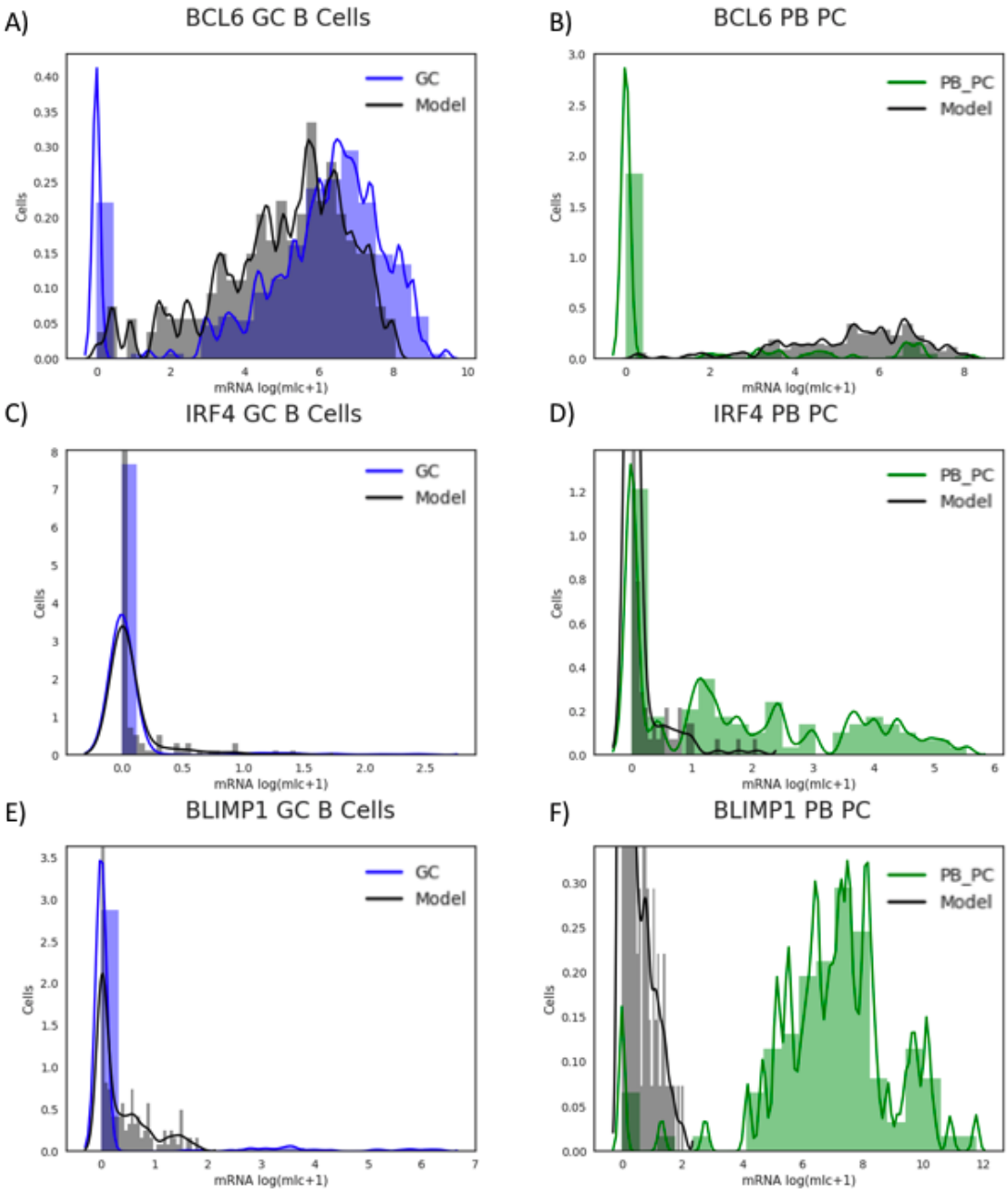
Supplementary Figure S1. The scheme of application of the stimuli Q_s, where $s \in \{BCR, CD40\}$. Stimuli Q_s were implemented in three steps: AB - linear increase ($t_{BCR} \in [0.5h; 1.5h]; t_{CD40} \in [35h; 36h]$), BC - stable stimuli ($t_{BCR} \in [1.5h; 24h]$; $t_{CD40} \in [36h; 60h]$), CD - linear decrease ($t_{BCR} \in [24h; 25h]; t_{CD40} \in [60h; 61h]$).



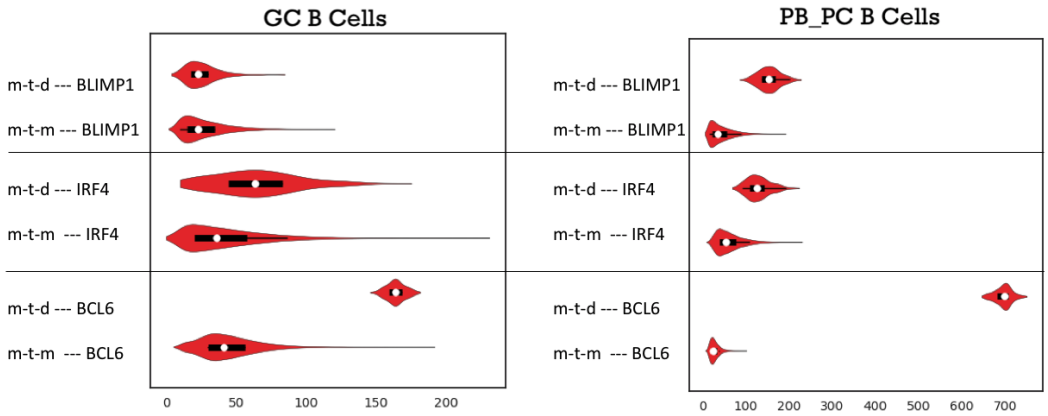
Supplementary Figure S2. Absence of bistability in the ODE reduce PDMP model (2.11), due to random parameter values. The average behavior of the ODE reduced PDMP model (2.11) for the nodes IRF4 (A), BCL6 (B) and BLIMP1 (C) of GRN (see Figure 2.1). BCR stimuli was applied from 0h until 25h and CD40 stimuli - from 35h until 60h. Parameters used for (2.11) are listed in Supplementary Table S3 accordingly to Bonnaffoux et al. [138] of $k_{on,\,init,\,IRF4} = 0.1$ (taken randomly in a same order of magnitude).



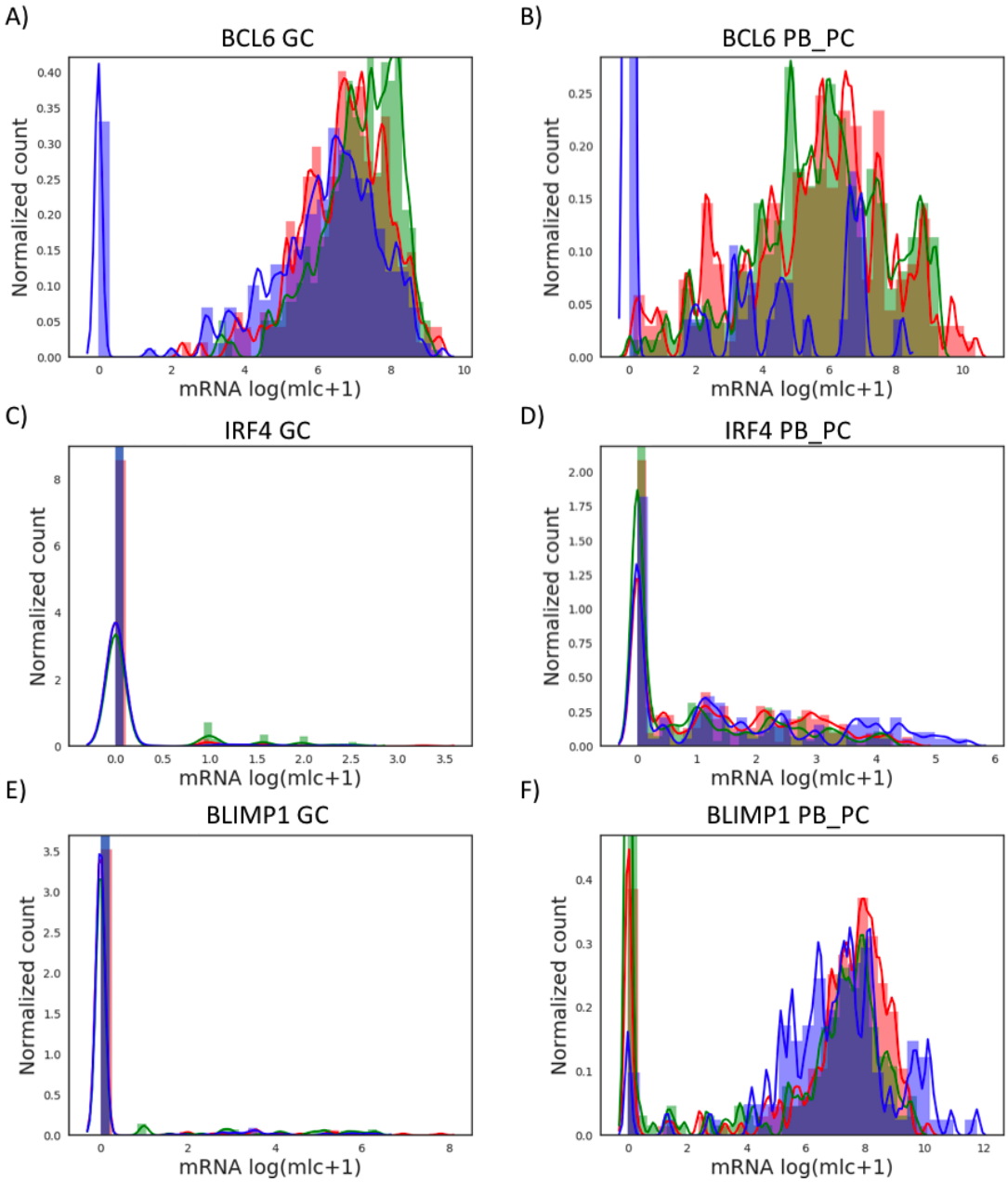
Supplementary Figure S3. The Kernel Density Estimate (KDE) plot and histograms of model-generated and experimental mRNA counts of BCL6, IRF4, BLIMP1 at GC and PB_PC stages. The subgraphs A, C, E represent log (molecule+1) normalized SC with BCL6, IRF4 and BLIMP1 compared between the model estimations at GC stage (grey) vs the experimental data from GC B cells (blue). The subgraphs B, D, F represent log (molecule+1) normalized SC with BCL6, IRF4 and BLIMP1 compared between the model estimations at PB_PC stage (grey) vs the experimental data from PB_PC cells (green). Simulation of 200 SC were used based on the parameter set from ODE reduced System (2.4)-(2.6) (see Tables 2.2-2.5, version I). Experimental SC dataset from Milpied et al. [3]



Supplementary Figure S4. The Kernel Density Estimate (KDE) plot and histograms of model-generated and experimental mRNA counts of BCL6, IRF4, BLIMP1 at GC and PB_PC stages. The subgraphs A, C, E represent log (molecule+1) normalized SC with BCL6, IRF4 and BLIMP1 compared between the model estimations at GC stage (grey) vs the experimental data from GC B cells (blue). The subgraphs B, D, F represent log (molecule+1) normalized SC with BCL6, IRF4 and BLIMP1 compared between the model estimations at PB_PC stage (grey) vs the experimental data from PB_PC cells (green). Simulation of 200 SC were used based on the parameter set, selected after automatized parameter screening strategy (see Tables 2.2-2.5, version II). Performed based on dataset from Milpied et al. [3]



Supplementary Figure S5. "Model-to-model and model-to-data distributions for GC and PB_PC stages and the three genes, BCL6, IRF4, BlIMP1. Graphs show the shape of distribution, median value, interquartile range and 1.5x interquartile range of the m-t-m and m-t-d distributions. Parameter set used is summarised in Tables 2.2-2.5, version III. Dataset from Milpied et al. [3]



Supplementary Figure S6. The Kernel Density and histograms of model-generated distributions with biggest and smallest KD between model-generated and experimental data distribution. The subgraphs A, C, E represent log (molecule+1) normalized SC with BCL6, IRF4 and BLIMP1 compared between the model with the lowest (red) and highest (green) KD from experimental data at GC stage (grey) vs the experimental data from GC B cells (blue). The subgraphs B, D, F represent log (molecule+1) normalized SC with BCL6, IRF4 and BLIMP1 compared between the model with the lowest (red) and highest (green) KD from experimental data estimations at PB_PC stage vs the experimental data from PB_PC cells (blue). GC - Germinal centre and PB_PC - Plasma Blast and Plasma cells. stage(node) min vs max vs exp data stands for 1) distribution with the lowest KD vs experimental data for a node 2) distribution with the highest HD vs experimental data for a node. Simulation of 200 SC were used based on the parameter set, selected after semi-automatized parameter screening (see Tables 2.2-2.5, version III). Performed based on dataset from Milpied et al. [3].

SM3 Supplementary Tables

Parameter	Value	Property
μ_{IRF4}	0.99	Production rate
σ_{IRF4}	10.16	Maximum transcription rate
k_{IRF4}	6.20	Dissociation constant

Supplementary Table S1. Parameter values of System (2.1)-(2.3) obtained after fitting the kinetic ODE model to microarray data accession no. GSE 12195.

Parameter	Value	Property
μ_{IRF4}	0.74	Production rate
σ_{IRF4}	18.5	Maximum transcription rate
k_{IRF4}	10.02	Dissociation constant

Supplementary Table S2. Parameter values of System (2.1)-(2.3) obtained by fitting the kinetic ODE model to SC data from Milpied et al. [3].

Parameter	Values
H_{11}	1
H_{21}	0.1
H_{31}	1
H_{12}	1
H_{22}	0.01
H_{32}	1
H_{13}	0.1
H_{23}	0.001
H_{33}	1
$H_{BCR,1}$	0.01
$H_{CD40,2}$	1
θ_{11}	-0.2
θ_{21}	-10
θ_{31}	-2
θ_{12}	0
θ_{22}	8
θ_{32}	0
θ_{13}	-1
θ_{23}	40
θ_{33}	0
$\theta_{BCR,1}$	-200
$\theta_{CD40,2}$	10
$s_{0,BCL6}$	6.5
$s_{0,IRF4}$	2
$s_{0,BLIMP1}$	6.5
$d_{0,BCL6}$	0.05
$d_{0,IRF4}$	0.05
$d_{0,BLIMP1}$	0.1733
$s_{1,BCL6}$	100
$s_{1,IRF4}$	160
$s_{1,BLIMP1}$	40
$d_{1,BCL6}$	0.138
$d_{1,IRF4}$	0.173
$d_{1,BLIMP1}$	0.173
$k_{on,init,BCL6}$	0.1
$k_{on,init,IRF4}$	0.1
$k_{on,init,BLIMP1}$	0.1
$k_{off,init,BCL6}$	1
$k_{off,init,IRF4}$	1
$k_{off,init,BLIMP1}$	1

Supplementary Table S3. Parameters of System (2.11) with values accordingly to Bonnaffoux et al. [138].

Mean Value	Model Estimated Value	Experimental Value
$\mu_{GC,\,BCL6}$	2.62	5.62
$\mu_{GC,\,IRF4}$	0.19	0.08
$\mu_{GC,\,BLIMP1}$	1.61	0.34
$\mu_{PB\_PC,\,BCL6}$	2.68	1.25
$\mu_{PB\_PC,\,IRF4}$	0.33	1.68
$\mu_{PB\_PC,\,BLIMP1}$	0.92	6.91

Supplementary Table S4. Mean values of model output generated by (2.4)-(2.6) and the experimental data, see Supplementary Figure S3. Based on the parameter set from Tables 2.2-2.5, version I.

Mean Value	Model Estimated Value	Experimental Value
$\mu_{GC,\,BCL6}$	5.20	5.62
$\mu_{GC,\,IRF4}$	0.08	0.08
$\mu_{GC,\,BLIMP1}$	0.35	0.34
$\mu_{PB\_PC,\,BCL6}$	5.31	1.25
$\mu_{PB\_PC,\,IRF4}$	0.1	1.68
$\mu_{PB\_PC,\,BLIMP1}$	0.3	6.91

Supplementary Table S5. Mean values of model output generated by (2.4)-(2.6) and the experimental data, see Supplementary Figure S4. Based on the parameter set from Tables 2.2-2.5, version II.

Mean Value	Model Value 1	Model Value 2	Experimental Values
$\mu_{GC,\,BCL6}$	6.73	7.15	5.62
$\mu_{GC,\,IRF4}$	0.09	0.22	0.08
$\mu_{GC,\,BLIMIP1}$	0.34	0.62	0.34
$\mu_{PB\_PC,\,BCL6}$	5.53	5.74	1.25
$\mu_{PB\_PC,\,IRF4}$	1.41	0.99	1.68
$\mu_{PB\_PC,\,BLIMP1}$	6.46	4.95	6.91

Supplementary Table S6. Mean values of model outputs, generated by System (2.4)-(2.6). Model estimated values of run 1 represents the distributions generated by the PDMP model (2.4)-(2.6) with minimum KD to experimental data. Model estimated values of run 2 represents the distributions generated by the PDMP model (2.4)-(2.6) with maximum KD to experimental data (see Supplementary Figure S6). Based on the parameter set Tables 2.2-2.5, version III.