UNIVERSITY OF TURIN

DOCTORAL THESIS

Development of bioinformatics algorithms for signatures discovery in biological systems

Author:
Dr. Greta ROMANO

Supervisor:
Dr. Francesca CORDERO

XXXII cycle
Doctor of Complex Systems for Life Sciences
Department of Computer Science

2016-2019

"*There is always a new ATG.*"

UNIVERSITY OF TURIN

Abstract

Doctoral School in Life and Health Sciences
XXXII cycle
Department of Computer Science

Doctor of Complex Systems for Life Sciences

Development of bioinformatics algorithms for signatures discovery in biological systems

by Dr. Greta ROMANO

Cells belonging to living organisms are composed of different levels of organization and through the intersection of these levels the cells give rise to a variety of cellular phenotypes. The levels of organization comprise different molecules and compounds starting from the DNA, i.e. genome level, the epigenetic modification, i.e. the epigenome level, to the RNA, i.e. transcriptome level, until the proteins, i.e. proteome level. The genome level is defined as the complete set of DNA of an organism, including all of its genes. Each genome contains all the information to build and maintain the organism (Lodish et al., 2000). The genome level is supported by the epigenome that comprises all the chemical compounds added to the genome as a way to regulate the activity (expression) of all the genes. The actors of the epigenome are not part of the DNA sequence, but are on or attached to the DNA. Mechanisms of epigenetic activity include DNA methylation, DNA-protein interactions, chromatin accessibility, histone modifications and more. Epigenetic modifications persist as cells divide and in some cases can be inherited through the generations (Lodish et al., 2000).
The transcriptome is defined as the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition (Li, Lim, and Ling, 2019). Since the transcriptome includes all messenger RNA (mRNA) in the cell, it reflects the genes that are being actively expressed at any given time in an entire organism or in a specific cell type. However, the transcriptome also includes other RNAs that do not fall into this category for their non-coding nature.
Finally, the proteome is the entire set of proteins that can be expressed by a genome, cell, tissue, or organism at a certain time. The proteome is many-fold larger than the genome since a single gene (composed of many exons) can generate hundreds of different protein molecules by alternative splicing and post-translational modifications (Garrels, 2001). Thus, the whole biological architecture makes the cell a complex system and hard to analyse in its entirety. However, in the era of deep sequencing techniques (i.e. Next Generation Sequencing (NGS)) significant advancements in elucidating the organization of cells have been achieved. In particular, the latest research are now increasingly focused to investigate in deep and precise way the nature and behaviour of the different cell types. Indeed, sensitive and accurate sequencing methods permit to analyse each level of organization and derive *signatures* (e.g. DNA, RNA, epigenetic and proteome signatures). Then, these *signatures* can be used to classify groups of cells both in physiological and pathological contexts as cancer. The signature of a cell or group of cells can be defined from all the different levels of organization. A *DNA signature* can be produced consequently to a genetic event occurring in the DNA that finally differentiate the cell (or group of cells) from all the others in the organism. An example of *DNA signature* in physiological context can be the recombination events occurring in the two helices generating a DNA signature. Indeed, DNA can be subject to homologous recombination events that result in the assortment of DNA pieces between chromosome pairs, without altering the arrangement of genes within the genome (Lodish et al., 2000). Other types of recombinations have specific roles in controlling gene expression in specific cell types or they may play an evolutionary role by contributing to genetic diversity. One example in human body of recombination event leading to signature specificity is the site-specific rearrangement that occurs during the development of the immune system, which recognizes foreign substances and provides protection against infectious agents. During the early stages of T and B lymphocytes maturation (i.e. part of the immune system actors), rearrangements in the V(D)J gene segments occur as a mechanism of genetic recombination in developing lymphocytes. The process results in a diverse repertoire of antibodies/immunoglobulins (Igs), B-cell receptors (BCRs) and T cell receptors (TCRs) belonging to B cells and T cells, respectively. However, gene rearrangements events characterizing lymphocytes can be also a signature of malignant conditions as Leukemia and Lymphoma (Dongen et al., 2015; Kotrova et al., 2015; Campana, 2010). Consequently to this event, the numerous immune repertoire of lymphocytes gives rise to *unique fingerprint-like signatures* that are different in healthy lymphocytes

cells (polyclonal), but constant in Leukemia and Lymphoma cells population (monoclonal). These populations retain the gene rearrangements as characteristic of a dominant tumour population (Beccuti et al., 2017a). In this context, DNA-based NGS approaches exploit the rearranged DNA signature of lymphocytes to identify and monitor the tumor population during treatment of patients (Beccuti et al., 2017a).
Beyond the genome, the discovery of the epigenome level shades the light on new regulatory mechanisms inside the cells and also on the putative association of epigenetic modifications and diseases. Indeed, in the original sense of the definition, epigenetics referred to all molecular pathways modulating the expression of a genotype into a particular phenotype. Even though epigenetic changes are regular and natural occurrences, they can also be influenced by several factors including age, environment, lifestyle, and disease state (Dupont, Armant, and Brenner, 2009). An example of *epigenetic signature* can be found during the biological event of gametogenesis, that is the process of division and differentiation of diploid or haploid precursor cells to form mature haploid gametes. In this process, genes undergo to the event of genomic imprinting that is an epigenetic event causing maternal and paternal alleles to be differentially expressed. These imprinted genes are epigenetically modified and maintain the established epigenetic signatures after fertilization, causing parental-specific gene expression (Moreno-Romero et al., 2019). Epigenetic modifications also play a significant role in determining the fate of stem cells and directing the differentiation into multiple lineages (Di Tizio et al., 2018). Among these, DNA methylation and histone modifications produce a unique signature that contribute to lineage-specific differentiation (Di Tizio et al., 2018). However, several studies (Urdinguio, Sanchez-Mut, and Esteller, 2009; Zhu et al., 2019) also focused their attention on the effect of mutations in epigenetic genes (e.g. encoding histone modifying enzymes or components of DNA methylation machinery) that seems to be associated with rare neurological conditions. The result of these alterations leads to a wide-scale disruption in epigenetic patterns across the genome, creating a distinct and clear *epi-signature*. Concerning transcriptome, it is possible to define a *RNA signature* as the set of expressed genes indicative of the presence of a given cell or population. The signature contains a uniquely characteristic pattern of gene expression that occurs as a result of physiologic or altered biological process (Itadani, Mizuarai, and Kotani, 2008).
Even before sequencing era, RNA signatures discovery was a powerful method to investigate both physiological and pathological condition of the cells in biological and medical research. For example, several studies used microarrays technique to generate gene signatures associated with precise transcriptomic characteristics of normal cells (Merienne et al., 2019; Chen et al., 2018) as well as cancer cells (Liu et al., 2008; Ng et al., 2014). The study of Liu et al., 2008 regards a novel approach based on microarray to derive gene expression signatures for cancer prognosis in the context of known biological pathways. In detail, the authors integrated gene expression profiling data and well-known pathway information to develop pathway specific gene expression signatures for breast cancer prognosis.
Finally, regarding the proteome level, several studies (Ricci et al., 2019; Pimienta et al., 2019; Ferreira et al., 2015) show how the presence both quantitative and qualitative of a specific set of proteins can be used as a signature indicative of pathological events as sepsis, renal cystis , diabetes (Ricci et al., 2019; Pimienta et al., 2019) or healthy tissue (Ferreira et al., 2015).

Nowadays thanks to more sensitive and accurate NGS techniques, the number of available signatures is increased allowing to obtain exhaustive coverage of the existing biological and pathological processes as cancer disease (Cantini et al., 2017). However, the results and the interpretation of new gene signatures is far from be fully exploited due to the poor consensus gene signatures describing molecular mechanisms and lack of computational approaches for signature analysis (Liu et al., 2008).

This thesis project is mainly based on the development of bioinformatics techniques to

discover signatures belonging to different cell types exploring both physiological and disease context, starting from NGS data. The project is divided in three main branches that converge on the characterisation of the immune system, essential for living organisms. The first branch is focused on the characterisation of a DNA signature from lymphocytes involved in a malignant condition (i.e. Lymphoma disease); the second branch is focused on the identification of RNA signatures in differentiating immune cells, analyzed with single-cell RNA sequencing technique (scRNA-seq). The third part of the thesis is focused on the exploitation of mathematical models to find a disease trend signature in patients affected by Multiple Sclerosis (MS).

The first part of the thesis is focused on the characterisation of a DNA signature in B-cell Lymphoma tumour.

Background Thanks to the advent of NGS technologies, genome studies revealed extensive intratumour heterogeneity as essential determinant of tumour progression, diagnosis and treatment. High levels of tumour heterogeneity affect several cancer types and may contribute to the treatment failure, by initiating phenotypic diversity and enabling more aggressive and drug resistant clones (Greaves and Maley, 2012). Among liquid cancers, B-cell Lymphomas are a group of blood cancers highly heterogeneous and following the clonal evolution model (Nowell, 1976). The most common molecular marker of clonality in this type of tumour is the Immunoglobulin Heavy Chain (IGH) gene rearrangement, occurring in the B-cells lineage (Velden et al., 2007). In particular, the patient-specific IGH rearrangement of the major B-cell tumour clone (i.e. first clone detected as tumoral at diagnosis) can be considered as the signature of the patient disease and it is extensively used to monitor a clinic parameter, the Minimal Residual Disease (MRD) during and after treatment. MRD represents the small number of tumour cells from the bone marrow that persist in the patient during or after a treatment (Ferrero et al., 2011; Velden et al., 2007; Dongen et al., 2015).

State-of-art method The gold-standard technique used for MRD detection in Lymphoma is Real-Time Quantitative Polymerase Chain Reaction (RQ-PCR) that stratifies patients in prognostic groups based on relapse risk. However, some limitations currently hamper a widespread RQ-PCR-MRD use for decision-making. The main disadvantage regards the lack of one specific marker and consequently a multi-target approach is necessary for identifying the cancer clonal landscape of Lymphoma patients.
Since its recently introduction, NGS technology showed the ability to overpass some limitations of the standard laboratory technique in MRD monitoring in terms of sensitivity and specificity. More importantly, NGS approach allows a full repertoire analysis thought multiclones detection at diagnosis and it gives the opportunity to monitor the clones during several follow-up samples (Ferrero et al., 2011; Kotrova et al., 2017; Ladetto et al., 2014). However, this issue is strictly associated with an appropriate computational analysis of the complex data obtained by NGS.

Aim of the work To solve this issue, we developed an innovative and easy-to-use bioinformatics suite, called HashClone, that simultaneously provides clonality assessment and MRD monitoring over time in patients affected by Lymphoma, starting from NGS data (Beccuti et al., 2017a; Romano et al., 2018). HashClone is based on a parallel implementation of four C++ applications for the data processing based on the analysis of all set of sample reads simultaneously and returns the corresponding set of tumour clones aligned with respect to the IGH rearrangement sequences, associated with their frequency in all input samples. HashClone is also provided of a graphic interface that allows the inspection of the tumour clones and their associated features in all the samples under study. Finally, HashClone is integrated

in a Docker container, a platform that allows to easily install and run the application packaged with all its dependencies and libraries.

Results The performances of HashClone were tested in the context of a particular rare type of Lymphoma, Mantle Cell Lymphoma (MCL), with the aims to detect the major B-cell tumour clone and monitor it over time during clinical follow ups. We performed a clonality analysis with HashClone algorithm starting from NGS data of a cohort composed of 28 patients affected by MCL enrolled in Fondazione Italiana Linfomi (FIL). The clonality analysis performed by HashClone was aimed to investigate differences among patients in terms of MRD clones and follow their trend during the entire therapy protocol. Moreover, the sensibility, sensitivity and flexibility performances of HashClone to identify clones with respect to the gold-standard technique RQ-PCR, were analyzed. HashClone overpassed the limitations of RQ-PCR technique, revealing its potential application in the clinic to establish clonal evolution of Lymphoma patients. To assess the accuracy of HashClone in the identification of the major B-cell clone and MRD monitoring, we compared its performance with respect to the results obtained by the state-of-the-art tool, ViDJil. Also in this case the performance of HashClone outperformed the ViDJil results.

The second part of the thesis is focused on the characterisation of RNA signatures in differentiating immune cells analyzed by single-cell RNA sequencing technique (scRNA-seq).

Background In the last century, the advent of scRNA-seq technology allows the researchers to analyze in deepest way the transcriptome of indivual cells. scRNA-seq is a powerful experimental technique ables to examine the sequence information from single cells with optimized NGS technologies, highlighting cell subpopulations specific signatures. Single cell analysis is particularly crucial to understand the functional differences existing among cells within a tissue. Indeed, individual cells of the same phenotype are commonly viewed as identical functional units of a tissue. However, single cells sequencing results suggest the presence of a complex organisation of heterogeneous cell states, producing together system-level functionality (Buettner et al., 2015).

State-of-art single-cell sequencing experiments result in a big amount of data that need to be analyzed in order to correctly cluster together cells belonging to the same type. Thus, from the advent of scRNA-seq a large variety of algorithms and tools for data analysis were developed: Seurat (Butler et al., 2018), SINCERA (Guo et al., 2015), Scanpy (Wolf, Angerer, and Theis, 2018). However, few of them provide an entire computational workflow equipped with all the necessary steps for single cell data and achieve at the same time functional (i.e. information about data and the utilized tools are saved in terms of meta-data) and computational reproducibility (i.e. real image of the computation environment used to generate the data is stored) through a user-friendly environment.

Aim of the work In this context, we propose a two-goals work:
(i) development of a new computational pipeline, named rCASC: reproducible Classification Analysis of Single Cell Sequencing data (Alessandri et al., 2019). (ii) development of rMLSC: reproducible Machine Learning Analysis of Single Cell Sequencing data, a computational module based on a machine learning approach to improve the interpretation of the results derived from rCASC data analysis, recovering a robust set of gene signatures for each cluster. rCASC is a modular scRNA-seq analysis workflow providing data analysis

tools from counts generation to cells subpopulation identification and exploiting docker containerisation to achieve computational reproducibility in the data analysis. The steps include: quality control, mapping, quantification, normalization, clustering and identification of Differentially Expressed Genes (DEGs). In particular, rCASC provides pre-processing tools to remove low quality cells and/or specific bias (e.g. cell cycle). Subpopulations discovery can be achieved using different clustering techniques based on different distance metrics. Quality of clusters is then estimated through a new metric namely Cell Stability Score (CSS), which describes the stability of a cell in a cluster as consequence of a perturbation induced by removing a random set of cells from the overall cells population (Alessandri et al., 2019). After rCASC analysis, rMLSC module is applied to the results with the goal to recover a clusters-specific gene-signature. The method includes several steps: pre-processing of rCASC results to create an appropriate input for machine learning tool; processing of the data using Random Forest classification algorithm (implemented in Weka tool (Witten et al., 2016)); post-processing analysis of the results with estimation of entropy measure and generation of the sets of clusters-specific gene-signature and visualisation.

Results rCASC and rMLSC were used to analyze a single-cell dataset from Pace et al., 2018 composed of about 10.000 immune cells. Data were firstly analyzed with rCASC for clusters derivation. Then, rMLSC was applied on these results using Random Forest classification algorithms and retrieving a set of genes signatures.
The results of our analysis allowed more suitable characterisation of the cell clusters, improved data interpretation and extend the information described in Pace et al., 2018, suggesting the presence of further enriched subsets of cells.

The third part of the thesis is focused on the characterisation of Multiple Sclerosis (MS) exploiting mathematical models to find a disease trend signature (Pernice et al., 2019b).

Background MS is an immune-mediated inflammatory disease of the central nervous system, characterized by alternate episodes of symptom exacerbation (relapses) with periods of disease stability (remission). Several treatments were proposed to contrast the MS progression with Daclizumab therapy, exhibiting promising results. However, its efficacy is often accompanied by serious side effects. Therefore, Daclizumab has been withdrawn from the market worldwide (Trapp and Nave, 2008). More studies are needed to understand these effects as well as to explain why women affected by MS seem to improve when they become pregnant and during the pregnancy period.

Aim of the work To improve the understanding of pathological disease as Multiple Sclerosis, we proposed in Pernice et al., 2019b a new methodology based on Coloured Petri Nets and sensitivity analysis for modelling and analyzing complex biological systems, showing that it reproduces the disease dynamics and the expected behaviours of MS.

Results To show the effectiveness of such methodology, a model of MS has been constructed and studied. Two different scenarios of MS were thus considered. In the former scenario the effect of the DAC administration is investigated, while in the latter MS was studied in pregnant women due to the association of pregnancy with fewer relapses in RRMS and reduced activity of the disease. Our results are in line with what comes from biological knowledge and clinical observations. The results of the simulations give us the opportunity to identify the key parameters involving in the modulation of the effect of the DAC therapy (Pernice et al., 2019b). In this work we provide a promising application of a computational

framework based on Coloured Petri Nets and sensitivity analysis to perform *in silico* experiments helping to improve the understanding of the MS disease, possibly giving some indications that may ameliorate the clinical management.

In conclusion, in this thesis project we deal with the development of several bioinformatics algorithms and tools that allow to identify signatures at different levels of biological organisation. In the first part, we focused our attention on DNA signature developing a tool, HashClone, able to analyze the complex tumour clonal pattern of Lymphoma patients starting from the derivation of VDJ gene signatures from tumour lymphocytes. We applied HashClone clonality analysis on 28 patients affected by MCL and proved that our tool improve and surpass the state-of-art tool and the standard technique. In the second part, we proposed a new pipeline for scRNA-seq data analysis composed of rCASC and rMLSC to retrieve clusters-specific gene-signatures and improve data interpretation. We tested rCASC and rMLSC on single cell dataset from Pace et al., 2018 characterized by differentiating immune cells. Finally, in the last part of the thesis, a new methodology based on mathematical models was used to model the complex biological scenario of MS disease in two different conditions: patients treated with DAC therapy and pregnant women affected by MS. We showed how MS shows a particular disease trend under the effect of the DAC therapy and how the simulation of pregnancy in MS patients mimics the resetting of the immune system.

Acknowledgements

First, I express my sincere gratitude to Professor Francesca Cordero and Professor Marco Beccuti for allowing me to conduct this research under their auspices. I am especially grateful for their confidence and the freedom they gave me to do this work. As a thesis supervisor, Professor Francesca Cordero supported me in all stages of this work. She is the initiator of this project and she always gave me constant encouragement and advice, despite her busy agenda. Without a coherent and illuminating instruction, this thesis would not have reached its present form. I would like to acknowledge also Professor Marco Beccuti who offered me valuable suggestions for the entire project. During my PhD project, he spent much time to teach me coding and provide me relevant advice. The presence of Professor Cordero and Beccuti was fundamental during my project and I will never thank them enough for what they have done for me. I extend my thanks to all the members of the q-Bio Group for the constant daily support and I wish them the best career in science. I wish to express my gratitude to Simone Ferrero and Elisa Genuardi from the Department of Molecular Biotechnology and Health Sciences, who helped me to learn molecular biology techniques in the laboratory and supported the entire project. I extend my sincere thanks to the members of the Italian Institute for Genomic Medicine, Raffaele Calogero and Luigia Pace, and all those who contributed directly or indirectly to the dissertation. Without the support of all members of my family, I would never finish this thesis and I would never find the courage to overcome all these difficulties during this work. My thanks go to my parents for their confidence and their love during all these years. I would especially like to express my gratitude to my sister Sara. Our geographical distance prevented me from participating in important phases of your life but you have always been there when I needed and I will always be there if you need. I hope my work can represent a small source of inspiration for you in realizing your dreams. I also extend my thanks to Laura Zavattini, who has always supported me and helped me overcoming the life difficulties which allowed the development of a wonderful friendship. If I came to the end of this path, I owe it above all to her. I would like to extend my warmest thanks to my dear boyfriend, Mattia Chiesa. If this work has sometimes prevented us from sharing important moments of life, know that I never stopped thinking about you.

Esprimo la mia sincera gratitudine alla professoressa Francesca Cordero e al professor Marco Beccuti per avermi permesso di condurre questa ricerca sotto la loro guida. Sono particolarmente grata per la loro fiducia e la libertà che mi hanno dato nel fare questo lavoro. Come supervisore della tesi, la professoressa Francesca Cordero mi ha supportato in tutte le fasi di questo lavoro. È la promotrice di questo progetto e mi ha sempre dato costante incoraggiamento e consulenza, nonostante i suoi mille impegni. Senza un'istruzione coerente e illuminante, questa tesi non avrebbe raggiunto la sua forma attuale. Vorrei ringraziare anche il professor Marco Beccuti che mi ha offerto preziosi suggerimenti durante l'intero progetto. In questi tre anni, ha trascorso molto tempo ad insegnarmi la programmazione e mi ha fornito validi consigli. La presenza della professoressa Cordero e del professor Beccuti è stata fondamentale durante il mio progetto e non li ringrazierò mai abbastanza per tutto quello che hanno fatto per me. Ringrazio tutti i membri del q-Bio group per il costante supporto quotidiano e auguro loro la migliore carriera scientifica. Desidero esprimere la mia gratitudine a Simone Ferrero ed Elisa Genuardi del Dipartimento di Biotecnologie Molecolari e Scienze della Salute, per avermi insegnato le tecniche di biologia molecolare nel loro laboratorio e per aver supportato l'intero progetto. Ringrazio sinceramente i membri dell'Italian Institute for Genomic Medicine, Raffaele Calogero e Luigia Pace, e tutti coloro che hanno contribuito direttamente o indirettamente alla tesi. Senza il supporto di tutti i membri della mia famiglia, non avrei mai finito questa tesi e non avrei mai trovato il coraggio di superare tutte le difficoltà durante questo lavoro. I miei ringraziamenti vanno ai miei genitori per la loro fiducia e

il loro amore durante tutti questi anni. In particolare vorrei esprimere la mia gratitudine a mia sorella, Sara. La nostra distanza geografica mi ha impedito di partecipare a fasi importanti della tua vita ma tu sei sempre stata lì quando ne avevo bisogno e io ci sarò sempre se tu ne avrai. Spero che il mio lavoro possa rappresentare una piccola fonte di ispirazione per te nella realizzazione dei tuoi sogni. Ringrazio anche e soprattutto la mia amica Laura Zavattini per avermi supportato e aiutato a superare alcune importanti difficoltà che hanno permesso lo svilupparsi di una splendida amicizia. Se sono arrivata alla fine di questo percorso, lo devo soprattutto a lei. Vorrei estendere i miei più sentiti ringraziamenti al mio caro ragazzo, Mattia Chiesa. Se questo lavoro a volte ci ha impedito di condividere momenti importanti della vita, sappi che non ho mai smesso di pensare a te.

Contents

List of Figures

List of Tables

To my Family

Chapter 1

Signatures in biological systems

1.1 Background

The levels of organisation existing in the cells (i.e. the genome, epigenome, transcriptome and proteome) intersect among them and give rise to a variety of cellular phenotypes, even though the cells of one living organism are composed of nearly identical genotypes. All together, this biological architecture make the cell a complex system hard to analyze in its entirety. However, it is possible to derive from all these levels of organisation a ***signature*** **whose elements and characteristics can help in classifying groups of cells, both in physiological and pathological context**. In the era of deep sequencing techniques, i.e. Next Generation Sequencing (NGS), significant advancements in elucidating the architecture and the organisation of cells through **signatures discovery** have been achieved and the latest research are now increasingly focused to investigate in deep and precise way how these signatures are composed and linked to the different nature and behaviour of cell types. Thanks to sensitive and accurate NGS techniques, the number of available signatures from the genome, the epigenome and more is increased, allowing to obtain exhaustive coverage of the existing biological and pathological processes as hematopoiesis or cancer (Cantini et al., 2017). However, the results and the interpretation of signatures, especially in cancer, is far from be fully exploited due to the poor consensus gene signatures describing molecular mechanisms and lack of computational approaches for signatures analysis (Cantini et al., 2017). In the next sections, an overview of the signatures in biological systems and the methodologies to retrieve them will be enhanced.

1.2 Genome signature

A *DNA signature* can be produced consequently to a modification event of the DNA that finally differentiate a cell, group of cells or an organism from all the others. Thus, it is possible to derive a DNA signature both at organisms and cells level. Regarding humans,the genome is full of DNA signatures thanks to the fact that the DNA is not identical among individuals and the genetic difference is estimated around 0.1%, on average (Hartman, Garvik, and Hartwell, 2001). This means that about one base pair out of every 1,000 will be different between any two individuals (Hartman, Garvik, and Hartwell, 2001). Causes of differences between individuals include independent assortment, the exchange of genes (crossing over and recombination) during reproduction (through meiosis) and various mutational events. The most common genetic differences in the human genome are single base-pair differences called single-nucleotide polymorphisms (SNPs). SNPs occur, on average, about every 1,000 bases when two different haploid genomes are compared (Hartman, Garvik, and Hartwell, 2001). Other types of polymorphisms (e.g. differences in copy number, insertions, deletions, duplications, and rearrangements) also occur, but much less frequently. Almost all human genetic variations have no adaptive significance. Some variations (e.g. a neutral mutation) alter

the amino acid sequence of the resulting protein but produce no detectable change in its function. Other variations (e.g. a silent mutation) do not even change the amino acid sequence (Hartman, Garvik, and Hartwell, 2001). All together, the totality of the genetic differences of an individual can be considered as a *unique signature individual-specific*. Regarding signatures at the cell level, DNA often undergoes to recombinational events of the two helices. Indeed, DNA can be subject to homologous recombination events that result in the reassortment of genes between chromosome pairs without altering the arrangement of genes within the genome (Lodish et al., 2000). In contrast, other types of recombination have a specific role in controlling gene expression in specific cell types while others may play an evolutionary role by contributing to genetic diversity. One example in human body of recombinational event leading to signature specificity is the site-specific rearrangement that occur during the development of the immune system, which recognises foreign substances and provides protection against infectious agents. During the early stages of T and B lymphocytes maturation (i.e. part of the immune system actors), a rearrangement in the V(D)J gene segments occurs as a mechanism of genetic recombination in developing lymphocytes. The process results in a diverse repertoire of antibodies/immunoglobulins (Igs), B-cell receptors (BCRs) and T cell receptors (TCRs) found on B cells and T cells, respectively. However, the gene rearrangement event characterising lymphocytes is not unique to physiological condition but also of malignant condition as Leukemia and Lymphoma (Dongen et al., 2015; Kotrova et al., 2015; Campana, 2010). The immune repertoire of lymphocytes obtained gives rise to unique fingerprint-like signature sequences that are different in each healthy lymphocytes cell (polyclonal), but constant in Leukemia and Lymphoma cells population (monoclonal). Thus, these rearrangements are characteristic of a dominant tumour population (Beccuti et al., 2017a). In this context, DNA-based sequencing approaches exploit the rearranged DNA signatures to monitor the tumour population during the patients treatment (Beccuti et al., 2017a).
Another example of DNA signature in pathological context can be found in all the diseases with complete penetrance. In genetics, penetrance refers to the proportion of people with a particular genetic change (such as a mutation in a specific gene) who exhibit signs and symptoms of a genetic disorder (Shawky, 2014). A complete penetrance indicates that individuals carrying the disease-causing mutation have clinical symptoms of the disease. Thus, in the complete penetrance case, geneticists assume that the mutation in a specific gene is the leading cause of the disease and can be considered as a signature of the disease event (e.g. Huntington corea, retinoblastoma, familial adenomatous polyposis and multiple endocrine neoplasia) (Shawky, 2014; Karp, 2009).

1.3 Epigenome signature

The epigenome comprises all the chemical compounds added to the genome, without alterations in the DNA sequence, as a way to regulate the activity (expression) of all the genes (Karp, 2009). The actors of this level of organisation are not part of the DNA sequence, but are on or attached to DNA. Mechanisms of epigenetic activity include DNA methylation, DNA-protein interactions, histone modifications and more. DNA methylation process is associated with important processes as genomic imprinting, X-chromosome inactivation, suppression of repetitive elements, and regulation of cell specific gene expression. DNA-protein interaction is a mechanism of the cell to control various essential cellular processes (e.g. replication, transcription, recombination, DNA repair). Histone modifications are modification of histone proteins (i.e. proteins that package and order the DNA into structural units called nucleosomes) which includes methylation, phosphorylation, acetylation, ubiquitylation, and sumoylation. Epigenetic modifications remain as cells divide and in some cases can be inherited through the generations. (Lodish et al., 2000). Beyond the genome,

the discovery of the epigenome level led to the identification of new regulatory mechanisms inside the cells and also to the putative association of epigenetic modifications and diseases. Indeed, in the original sense of the definition, epigenetics referred to all molecular pathways modulating the expression of a genotype into a particular phenotype. Even though epigenetic change is a regular and natural occurrence, it can also be influenced by several factors including age, environment, lifestyle, and disease state (Dupont, Armant, and Brenner, 2009). An example of *epigenetic signature* can be found during the biological event of gametogenesis (i.e. process of cell division and differentiation of diploid or haploid precursor cells to form mature haploid gametes). In this process, genes undergo to the event of genomic imprinting that is an epigenetic event causing maternal and paternal alleles to be differentially expressed. These imprinted genes are epigenetically modified and maintain the established epigenetic signatures after fertilisation, causing parental-specific gene expression (Moreno-Romero et al., 2019).
Epigenetic modifications also play a significant role in determining the fate of stem cells and directing the differentiation into multiple lineages (Di Tizio et al., 2018). In particular DNA methylation and histone modifications have a unique signature that contribute to lineage-specific differentiation (Di Tizio et al., 2018). However, several studies (Urdinguio, Sanchez-Mut, and Esteller, 2009; Zhu et al., 2019) also focused their attention on the effect of mutations in epigenetic genes (e.g. encoding histone modifying enzymes or components of DNA methylation machinery) that seems to be associated with rare neurological conditions. The result of these alterations leads to a wide-scale disruption in epigenetic patterns across the genome, creating a distinct and clear *epi-signature*.

1.4 Transcriptome signature

The product of gene expression is the transcriptome that is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition (Li, Lim, and Ling, 2019). The primary division of transcripts is between coding RNA and non-coding RNA. Coding RNA transcripts (mRNA) reflect the genes that are being actively expressed at any given time in an entire organism or in specific cell types. The expression of different genes allows cells to differentiate and perform different functions. The non-coding part of the transcriptome is composed of RNA molecules that are not translated into a protein (e.g. microRNA, ribosomal RNA (rRNA), transfer RNA (tRNA)) (Brown, 2002). It is possible to define a *RNA signature* as the set of expressed genes indicative of the presence of a given cell, population or phenotype. The signature contains a uniquely characteristic pattern of gene expression that occurs as a result of physiologic/altered biological process or pathogenic medical condition (Itadani, Mizuarai, and Kotani, 2008). RNA signatures can be identified also with non-coding RNAs as microRNA or tRNA whose presence is indicative of on-going biological processes.
The phenotypes that may theoretically be defined by a gene expression signature, range from those that could predict the survival or prognosis of an individual with a disease (i.e. *prognostic*)(Oldenhuis et al., 2008), those that are used to differentiate between different subtypes of a disease (*diagnostic*), to those that predict activation of a particular pathway (*predictive*). In general, a *prognostic RNA signature* is likely to describe the outcome of a biological process or disease (Oldenhuis et al., 2008). Indeed, the classification of a biological phenotype or medical condition based on a specific RNA signature or multiple signatures can be a useful prognostic biomarker for the associated phenotype or condition. Several studies have been conducted focusing on *prognostic RNA signatures* identification to improving the diagnostic methods and the therapeutic courses. For example, Liu et al., 2007 generated a 186-gene RNA signature, which was evaluated for its association with overall survival and

metastasis-free survival in patients with breast cancer. Moreover the signature was used to stratify patients with high-risk early breast cancer into prognostic categories (good or poor) (Liu et al., 2007). For the same type of tumour, also Kwan et al., 2018 identifies a 17-genes cancer-related transcripts signature of circulating tumour cell as a measurement to inform therapy in breast cancer. However, prognostic signatures were also found in hepatocellular carcinoma (Hoshida et al., 2013), leukaemia (Verhaak et al., 2005) and are continually being developed for other types of cancers and disorders as well (Oldenhuis et al., 2008).
A *diagnostic gene signature* can be considered a signature that distinguishes phenotypically similar medical conditions with several thresholds of severity (Nguyen, Welty, and Cooperberg, 2015). In Wouters et al., 2009 the authors analyzed a recurrent mutation of CEBPA gene in acute myeloid leukemia and found that, even though most AMLs exhibit two mutations of the gene, also cases with single mutation can occur. Thus, CEBPA gene with double and single mutation give rise to two gene expression profiles. AML patients with double mutation were associated with a favourable outcome with respect to patients with single mutation (Wouters et al., 2009).
Finally, a *predictive gene signature* can predict the effect of treatment in patients that exhibit a particular disease phenotype (Oldenhuis et al., 2008). *Predictive gene signatures* can be employed in the personalized therapeutic intervention through identification of novel therapeutic targets as well as the most qualified subjects that will receive optimal benefit from the specific treatment (Pujol et al., 2015). In Grob et al., 2017, the authors evaluated RNA signature profiling as a way to predict survival in the metastatic cutaneous melanoma. Forty-eight patients were enrolled and tested as positive or negative for previously identified RNA signature associated with the clinical outcome (32 patients positives, 15 negatives). Median overall survival was 11.4 months for patients positive to the RNA signature and 5.3 months for patients negative to the RNA signature. Thus, this study evaluates RNA signature profiling as predictive signature of survival (Grob et al., 2017).

1.5 Technologies for genes signature discovery

In the era of deep sequencing techniques, significant advancements in elucidating the architecture and the organisation of cells have been achieved and the latest research are now increasingly focused to investigate in deep and precise way nature and behaviour of the different cell types. Indeed, sensitive and accurate sequencing methods allow the researchers to analyze each level of organisation and derive *signatures* (e.g. DNA, RNA, epigenetic signatures) that can be used to classify groups of cells both in physiological and pathological contexts as cancer.

Sequencing and its manifest disciplines, known as genomics, transcriptomics and more, is the most tip technique in *signature discovery* field. Sequencing is the result of a combination of molecular biology with nucleotide chemistry born in the early 1970s with the publication of the first method for DNA sequencing by Dr. Frederick Sanger (Sanger, Nicklen, and Coulson, 1977; Mardis, 2013). Sanger methods has paved the way for Next Generation Sequencing (NGS) techniques that could be applied to genome sequencing for DNA signature, transcriptome profiling (RNA-Seq) for RNA signature, DNA-protein interactions (ChIP-sequencing) and more in general epigenetic modifications (ELISA, ChIP-on-chip, ChIP qPCR) for the epigenomic signature.

The critical difference between Sanger sequencing and NGS is the sequencing volume. While the Sanger method only sequences a single DNA fragment at time, NGS techniques can analyze millions of fragments in a massive parallel way. Nowadays, NGS methods permit to sequence every type of molecule in the cells as DNA, RNA as well as proteins. Nowadays,

thanks to these more sensitive and accurate NGS techniques, the number of available *signatures*, relative to the several level of cell organisation (i.e. genome, transcriptome, ecc), is increased allowing to obtain exhaustive coverage of the existing biological and pathological processes as cancer disease (Cantini et al., 2017). Since this thesis is focused on DNA and RNA signatures, in the next section a brief explanation regarding NGS methods for DNA and RNA sequencing is presented.

DNA sequencing Depending from the fragment length, high-throughput sequencing can require to pre-process DNA before sequencing. For longer targets such as chromosomes, common approaches consist of cutting (with restriction enzymes) or shearing (with mechanical forces) large DNA fragments into shorter fragments. Then, a library of fragments carrying synthetic DNAs adapters added covalently to each fragment, is constructed. These adapters are sequences specific to each platform that can be used to amplify the library fragments during specific steps of the process (Figure 1.1 A). For example, in the Illumina/Solexa library protocol, the library fragments are amplified in situ on a solid surface, either a bead or a flat glass microfluidic channel with adapter sequences complementary to those on the library fragments. Amplification is required to provide sufficient signal from each of the DNA sequencing reaction steps that determine the sequencing data for that library fragment (Figure 1.1 B). The scale and throughput of NGS are often referred to as massively parallel, which is an appropriate descriptor for the process that follows fragment amplification and yield sequencing data (Mardis, 2013) (Figure 1.1 C). After sequencing, ad-hoc computational pipelines are exploited for data analysis (Figure 1.1 D)
The approaches at the bases of DNA sequencing for signatures discovery are: Whole-Genome Sequencing (WGS), Whole-Exome Sequencing (WES) and Targeted Sequencing Panels. Whole-Genome sequencing determines the order of the nucleotides in the entire genome, while Whole-Exome Sequencing is an alternative approach to sequence only the exome (i.e. exons, or coding regions, of the genes which are then transcribed and translated). Exomes compose only about 2% of the whole genome and can be sequenced at a much greater depth (number of times a given nucleotide is sequenced) for lower cost. The last method exploits targeted or hot-spot sequencing panel with a selection of specific genes that are known to harbor mutations and contribute to pathogenesis of disease (Karp, 2009). All together, these types of sequencing allow the researchers to look for DNA signatures as genetic variations and aberrations (e.g. single nucleotide variants, deletions, insertions and copy number variants) in the whole genome or in a gene-specific manner.

RNA sequencing RNA sequencing reveals the presence and quantity of RNA in a pool of cells, analyzing the continuous changes in the transcriptome of the cells. Specifically, it allows to look at alternative spliced transcripts, post-transcriptional modifications, gene fusions, mutations, SNPs and changes in gene expression (Ozsolak and Milos, 2011). In addition to mRNA transcripts, RNA sequencing can look at the identification of small RNA, such as miRNA, tRNA, and ribosomal profiling. Thus, beyond quantifying gene expression, the data generated by RNA sequencing facilitate the discovery of novel transcripts, identification of alternatively spliced genes and detection of allele-specific expression (Kukurba and Montgomery, 2015). Unfortunately, conventional RNA-Seq studies do not capture the transcriptomic composition of individual cells. Indeed, considerable evidence indicates that single cells are heterogeneous, and the cell-to-cell variability in gene expression is ubiquitous, even within phenotypically homogeneous cell population (Huang, 2009). The transcriptome of a single cell is highly dynamic, reflecting its functionality and responses to changing stimuli. Furthermore, genes that show mutually exclusive expression in individual cells may be

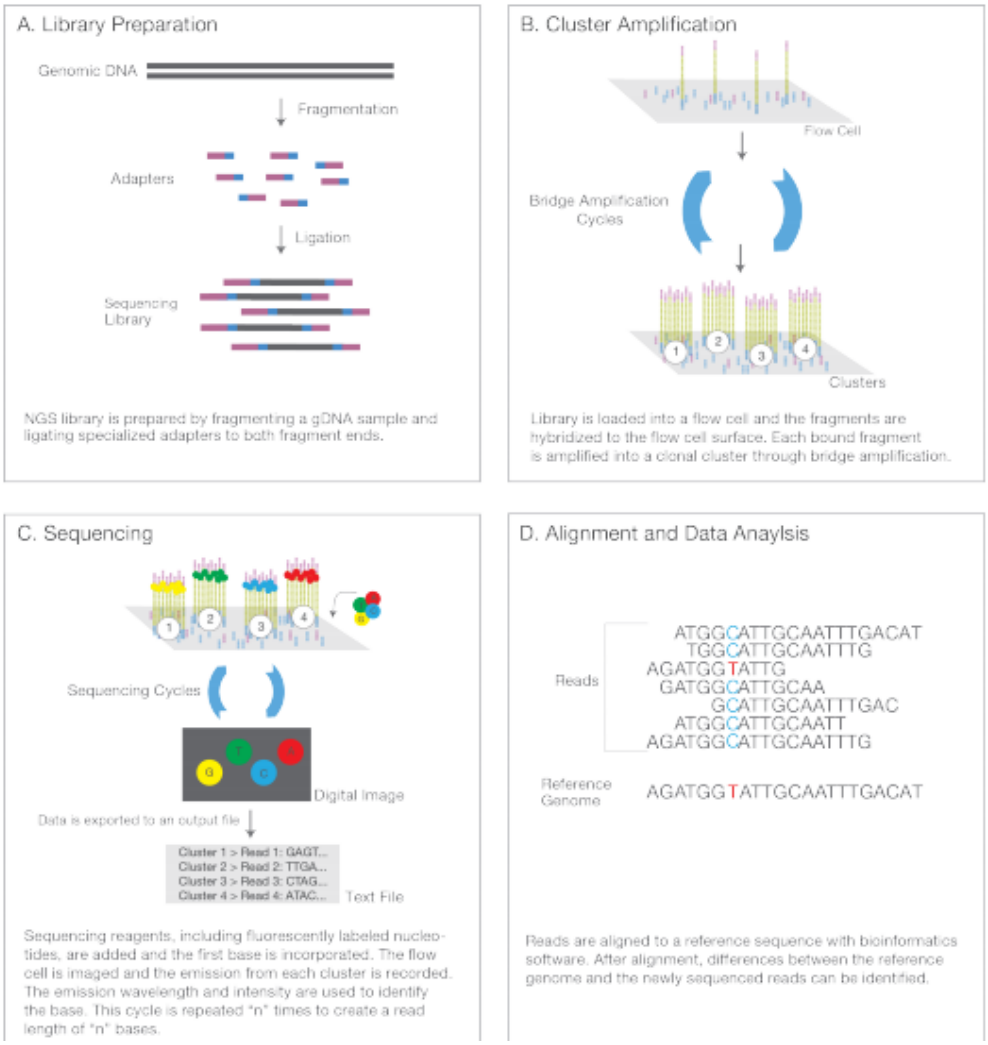


FIGURE 1.1: **Next-Generation Sequencing Chemistry Overview for Illumina/Solexa platform.** The workflow includes four steps: (A) library preparation, (B) cluster generation,(C) sequencing, and (D) alignment and data analysis.(Figure from Illumina website)

actually observed as genes showing co-expression in expression analyses of bulk cell populations (Kukurba and Montgomery, 2015). In this context, further improvements were made with the introduction of single cell RNA sequencing (scRNA-seq). Indeed, scRNA-seq technique allows the researchers to analyze in deepest way the transcriptome of indivual cells, providing a higher resolution of cellular differences. This can uncover the existence of rare cell types within a cell population that may never have been seen before (Kukurba and Montgomery, 2015). In 1990, Iscove and colleagues (Brady, Barbara, and Iscove, 1990) followed by James Eberwine et al. (Eberwine et al., 1992) in 1992, revealed an entire transcriptome at single-cell level using exponential amplification by PCR and linear amplification by in vitro transcription, respectively. However, only in 2009 thanks to the work of Tang et al., 2009, it was published the first description of single-cell transcriptome analysis based on an NGS platform describing the characterisation of cells from early developmental stages. All these studies brought to an explosion of interest in obtaining high-resolution views of single-cell heterogeneous content on a global scale.

scRNA-seq technique impose the single-cell isolation as first step to obtain transcriptome information from an individual cell. Several methods are available for this task and they differentiate for the collection and isolation technique of the cells (Hwang, Lee, and Bang, 2018). Limiting dilution method (Figure 1.2,a) is a commonly used technique in which pipettes are used to isolate individual cells by dilution. The micromanipulation (Figure 1.2,b) is a method used to retrieve cells from early embryos or uncultivated microorganisms, and microscope-guided capillary pipettes have been utilised to extract single cells from a suspension. The most commonly used strategy is the flow-activated cell sorting (FACS, Figure 1.2,c) for isolating many purified single cells that express a very low level of surface markers and enables sorting of distinct populations. Alternatively, negative selection is possible for unstained populations. The potential limitations of these techniques include the requirement for large starting volumes and the need for monoclonal antibodies to target proteins of interest. Another method is the laser capture microdissection (Figure 1.2,d) where a laser system aided by a computer system is used to isolate cells from solid samples. Finally, microfluidic technology (Figure 1.2,e) for single-cell isolation allows low sample consumption and low analysis cost together with a precise fluid control and a substantially reduced risk of external contamination. Notably, the rapid expansion of microfluidic technology in recent years has transformed the research capabilities of both basic scientists and clinicians (Goldman et al., 2019). Applications of this technology include long-term analysis of single bacterial cells in a microfluidic bioreactor and the quantification of single-cell gene expression profiles in a highly parallel manner (Goldman et al., 2019). Another promising method in microfluidic technology for single-cell isolation is microdroplet-based microfluidics, which comprises aqueous droplets monodispersion in a continuous oil phase (Goldman et al., 2019). The lower volume required by this system enables the manipulation and screening of thousands to millions of cells at a reduced cost. Consequently, this high-throughput processing method enables analysis of rare cell types in a sufficiently heterogeneous biological space (Goldman et al., 2019). Finally, a system to isolate and enumerate rare circulating tumour cells (CTCs) in patient blood samples (Figure 1.2,f). Following the isolation of single cells, libraries are typically generated via cell lysis, reverse transcription into first-strand cDNA using uniquely barcoded beads, second-strand synthesis, and cDNA amplification (Goldman et al., 2019).Thus, thanks to barcoded beads, scRNA-seq has the ability to resolve noise in bulk-NGS through the additional ability to trace generated reads back to their cell of origin.

Single Cell RNA sequencing applications One of the potential and most used applications of scRNA-seq technology is certainly the new cell type identification. Indeed, the possibility to look to the differences in gene expression between individual cells could bring to the identification of rare populations that cannot be detected with bulk methods. For example, good advancements have been done in hematopoiesis allowing to deconvolute many diverse immune cell populations in healthy and diseased states through scRNA-seq (Ranzoni, Strzelecka, and Cvejic, 2019)(Yamamoto et al., 2013; Naik et al., 2013; Sanjuan-Pla et al., 2013). All together these studies allowed the discovery of new immune cell types and revealed that haematopoiesis is a continuous rather than a stepwise process, thus challenging the classical haematopoietic lineage tree model. Another successful application of scRNA-seq is the cell hierarchy reconstruction which enables the capture of different instantaneous time points along an entire trajectory (Goldman et al., 2019; Shin et al., 2015). Hierarchy tracing is a long-standing fundamental question in biology aimed at understanding in particular how a single-cell embryo gives rise to various cells types that are organised into complex tissue and organs (Goldman et al., 2019; Shin et al., 2015). In this context, another interesting potential application of scRNA-seq includes the identification of genes involved in stem cell regulatory networks, whose information is essential for understanding the basic biological processes underlying human health and diseases. In cancer research field, scRNA-seq has revealed itself

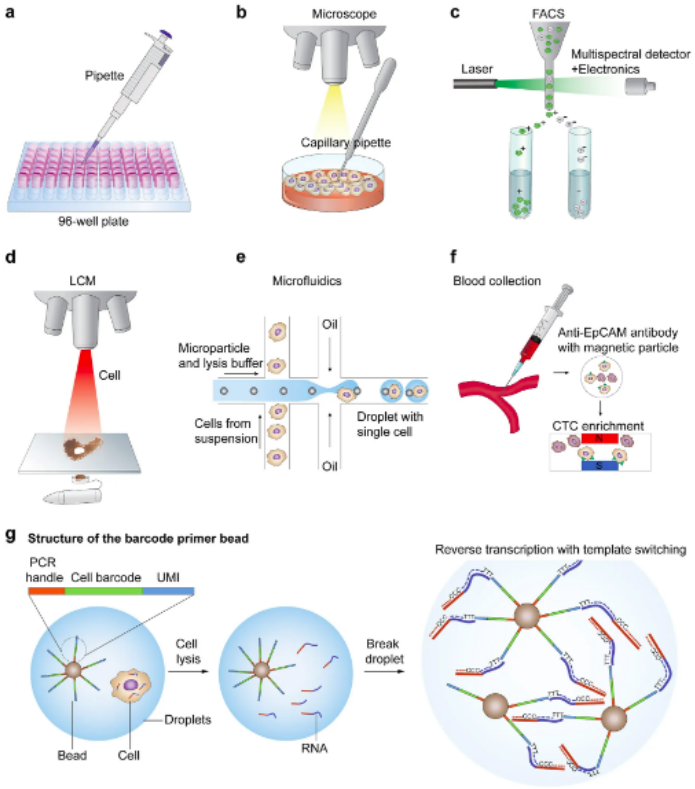


FIGURE 1.2: **Single-cell isolation and library preparation.** (a) The limiting dilution method isolates individual cells, leveraging the statistical distribution of diluted cells. (b) Micromanipulation involves collecting single cells using microscope-guided capillary pipettes. (c) FACS isolates highly purified single cells by tagging cells with fluorescent marker proteins. (d) Laser capture microdissection (LCM) utilises a laser system aided by a computer system to isolate cells from solid samples. (e) Microfluidic technology for single-cell isolation requires nanoliter-sized volumes. An example of in-house microdroplet-based microfluidics (e.g., Drop-Seq). (f) The CellSearch system enumerates CTCs from patient blood samples by using a magnet conjugated with CTC binding antibodies. (g) A schematic example of droplet-based library generation. Libraries for scRNA-seq are typically generated via cell lysis, reverse transcription into first-strand cDNA using uniquely barcoded beads, second-strand synthesis, and cDNA amplification. (Figure and caption from Hwang, Lee, and Bang, 2018)

also a good method to study intratumour heterogeneity (Goldman et al., 2019). Tumour heterogeneity is a common event that can occur both within and between tumours, and the hope is that scRNA-seq can be applied to reveal unknown tumour features that cannot be discerned from conventional bulk transcriptomic studies. For example, single-cell RNA sequencing has revealed significant tumor heterogeneity in primary glioblastomas (Patel et al., 2014) that inversely correlated with survival, indicating that intratumour heterogeneity should be considered as an essential clinical factor (Henssen et al., 2017). Metastatic melanoma is also transcriptionally heterogeneous, and this heterogeneity is associated with several biological processes including cell cycle stage, location, and also chemotherapeutic resistance (Tirosh et al., 2016). Thus, the ability to find and characterise outlier cells within a population has potential implications for further understanding about drug resistance and relapse in cancer treatment (Hwang, Lee, and Bang, 2018). Additionally, single-cell RNA sequencing could

also be applied to reconstruct clonal and phylogenetic relationships between tumour cells by modelling transcriptional kinetics (Goldman et al., 2019). Thus, further integration of single-cell transcriptomics with other cancer omics data will expand the knowledge in cancer research field.

Chapter 2

Characterisation of DNA signatures: identification of clonal IGH rearrangements to quantify Minimal Residual Disease in B-cell lymphoma from deep sequencing data

2.1 Background

Cancers represent abnormal outgrowth of tumour cells that acquire cell growth, survival, and metabolism selective advantages, allowing territorial expansion, proliferative self-renewal, migration and invasion (Greaves and Maley, 2012). By now, cancers are considered complex pathological processes due to the multiple and various characteristics evident in the same or different tumour types. Thus, each cancer can be considered individually unique due to both the variable time frames of evolution (~1–50 years) and to the physical structure, genotype and phenotype shifting over time among patients (Greaves and Maley, 2012). However, a large debate is still present in the research community regarding tumours growth. Actually, researchers debate is focused among three theories (Adams and Strasser, 2008)(Figure 2.1 from Adams and Strasser, 2008):

- Cancer stem cell model (Bonnet and Dick, 1997)
- Clonal evolution model (Nowell, 1976)
- Clonal succession creating a dominant clone (Adams and Strasser, 2008)

The first theory of tumour growth is the *Cancer Stem Cell (CSC) model*. It is based on the idea that tumours are hierarchically organised with CSCs lying at the apex (Bonnet and Dick, 1997; Kreso and Dick, 2014) (Figure 2.1 A, from Adams and Strasser, 2008). CSC are tumourigenic cells biologically distinct from other subpopulations with two defining features: their long-term ability to self-renew and their capacity to differentiate into progeny that is non-tumourigenic but still contributes to the growth of the tumour (Kreso and Dick, 2014). From literature it is known that both CSCs and normal tissue stem cells have a self-renewal capacity but deregulation of this property is typical of CSCs. For many cancers, CSCs is a distinct population that can be physically isolated from the other tumour cells and can be shown to have clonal long-term repopulation (Kreso and Dick, 2014; Nguyen et al., 2012). However, some cancer types seem to be homogeneous or possess a very shallow hierarchy, and it has not been possible to distinguish CSCs from non-CSCs. Moreover, last researches are considering the idea that certain cancer cells exhibit plasticity converting between a stem and non-stem-cell state (Kreso and Dick, 2014). In general, CSCs model has been studied

and identified in several solid tumours, including Brain (Singh et al., 2003), Breast (Al-Hajj et al., 2003) and Ovarian (Alvero et al., 2009).

The second theory is the *clonal evolution model* that was first proposed in 1976 by Peter Nowell (Nowell, 1976). In this model, tumours arise from a single mutated cell, accumulating additional mutations as it progresses. With successive cellular division and collecting mutations, additional subpopulations rise, and each of these subpopulations has the ability to divide and mutate further. This heterogeneity may give rise to subclones that possess an evolutionary advantage within the tumour environment and override normal cells for space, energy and nutrient requirements, becoming the dominant population over time (Nowell, 1976) (Figure 2.1 B, from Adams and Strasser, 2008). This feature appears to be in line with an evolutionary process as those reported by Darwin about species evolution (Greaves and Maley, 2012). As predicted by Nowell and thanks to the advent of NGS technologies, the massive sequencing of genomes revealed that clones withing tumours contain cells that harbor hundreds to thousands of genetic mutations, chromosomal alterations and epigenetic aberrations (Sabaawy, 2013). The acquisition of mutations is random but the selection and propagation of dominant neoplastic clones that ultimately lead to a malignant transformation may be independently sustained by multiple different combinations of driver mutations during certain stages of tumour progression; while the remaining mutations are considered neutral (Gerlinger et al., 2012). Finally, also the tumour microenvironment may contribute to this branched tumour expansion due to the selective pressures that the subclones may be exposed to (Greaves and Maley, 2012). Examples of tumours that follow clonal evolution model are Leukemia and Lymphoma (Adams and Strasser, 2008; Ma et al., 2019; Jan et al., 2012), on which this thesis is focused.

The last theory is the *mixed model* based on both theories assumptions. Indeed, many tumours that initially follow the cancer stem cell paradigm could progress acquiring additional mutations and finally follow the clonal evolution model (Adams and Strasser, 2008)(Figure 2.1 C, from Adams and Strasser, 2008). Such tumours might exhibit features of both models, such as a relatively high frequency of tumour-propagating cells, high mutation frequency as well as a substantial proportion of cells non-tumourigenic that sustains tumour growth (Adams and Strasser, 2008).

Regardless the type of mechanisms a tumour undergoes, the evolution from the founder tumour cell to the final disease bring along also differences in the morphological and phenotypic profiles of the evolving tumour cells resulting in heterogeneity of cellular morphology, gene expression, metabolism, motility, proliferation, and metastatic potential (Marusyk and Polyak, 2010). Tumours originating from different tissues and cell types vary in genomic landscapes, prognosis and their response to therapies, probably due to genetic events of transformation interacting with cell-intrinsic biological properties (Burrell et al., 2013). However, considerable variations in terms of genomic aberrations, aggressiveness and drug sensitivity are also observed among tumours that originate from the same tissue and cell type. For these reasons, *intertumour heterogeneity* is defined as genetic and phenotypic variation among tumours while *intratumour heterogeneity* is referred to variation within individual tumours (Figure 2.2 from Felipe De Sousa et al., 2013) (Burrell et al., 2013).
Intertumour heterogeneity is based on several aspects of the tumours including morphological differences of cells among different types and also in their development as well as genetic heterogeneity (Marusyk and Polyak, 2010). A common source of intertumour heterogeneity is genomic instability, a term used to indicate the increased mutation rate in tumour cells that often arises when key regulatory pathways are disrupted in the cells (Burrell et al., 2013). Examples include impaired DNA repair mechanisms, which can lead to increased replication errors or defects in the mitosis, that allow gain or loss of entire chromosomes. Moreover, due

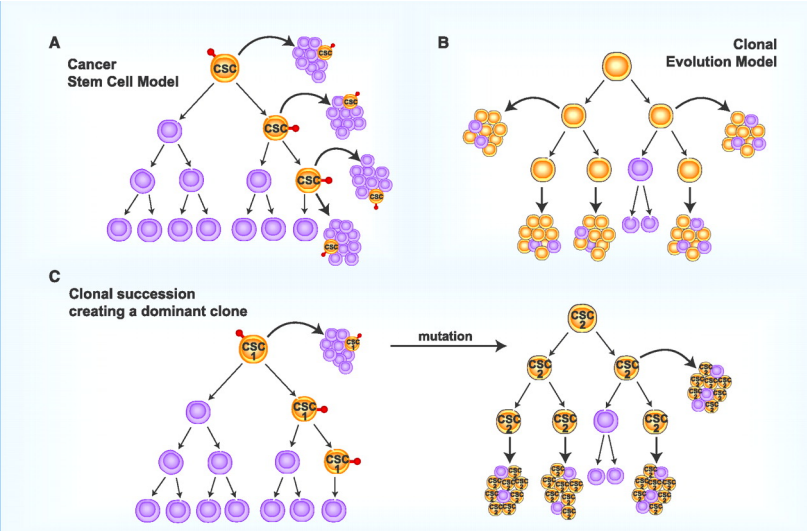


FIGURE 2.1: **Models for the nature of sustained tumour growth.** (A) The cancer stem cell (CSC) model focuses on the unique property of self-renewal activity of the CSC (gold) and, hence, represents the only relevant target for therapy. CSC can be isolated prospectively by surface markers (red). (B) In the clonal evolution model, a substantial proportion of the tumour cells (gold) can sustain its growth, and hence, therapy must attempt to eliminate all the cells. (C) In the mixed model, whereas a tumour is originally driven primarily by rare cells of one phenotype (CSC1), a mutation enhancing self-renewal in a differentiated derivative creates a dominant subclone driven by cells of a different phenotype (CSC2). In some human tumours (e.g., AML), CSC1 but not CSC2 might be able to engraft mice. (Figure from Adams and Strasser, 2008)

to epigenetic changes, tumour cells can also show heterogeneity between their expression profiles with different mutational frequencies of oncogenes and tumour suppressors between tumours of different tissues (Burrell et al., 2013).

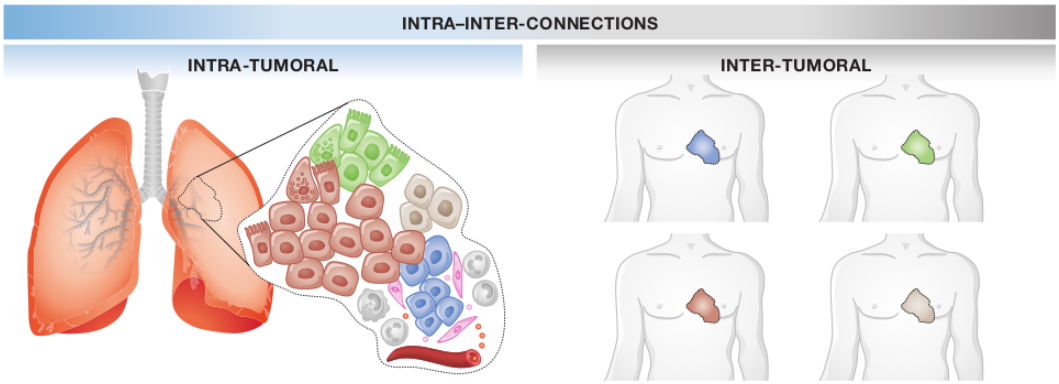


FIGURE 2.2: **Intra- and intertumour heterogeneity.** Intratumour heterogeneity revealed by multiple subclones existing in the same tumour. Each of these is represented by a different subtype of lung cancer resulting in intertumour heterogeneity. (Figure from Felipe De Sousa et al., 2013)

Another factor supporting intertumour heterogeneity is the mutational burden (i.e. total number of somatic/acquired mutations per coding area of a tumour genome) that can be highly variable across tumour types (Lawrence et al., 2013). For example, Leukemias tend to have the lowest number of mutations per tumour compared to adult solid tumours. Moreover a further prove of intertumour heterogeneity regards the fact that the same driver mutations can occur in different tumour types, suggesting that the same pathways can be active in different tumours (Alexandrov et al., 2013).
Tumour evolution is also associated with significant intratumour heterogeneity due to differences at genetic and epigenetic levels within tumour itself. All tumour cells contain shared somatic mutations that reflect a clonal origin, but additional mutations are present in subpopulations of cells that define tumour subclones (Klco et al., 2014). An extended analysis of genetic and intratumour heterogeneity across and within tumour types was proposed by Andor et al., 2016 (Figure 2.3 from Andor et al., 2016). The authors used bioinformatics tools to quantify intratumour heterogeneity from exome-sequencing data in The Cancer Genome Atlas. The authors measured the number and size of genetically diverse clones (present at a $\geq$ 10% frequency) in 1,165 primary tumour samples across 12 cancer types. On average, four clones were estimated to coexist in a tumour at the time of biopsy or surgical resection. Even for thyroid carcinoma, the least heterogeneous tumour type, two or more clones were predicted to coexist in >50% of the samples. (Figure 2.3 from Andor et al., 2016). The most important result of this analysis is that 86% of analyzed tumours had at least two clones, so the genetic intratumour heterogeneity occurs in the vast majority of cancers represented among the 12 types included in the study. Thus, subclonal genetic heterogeneity is a common feature of many types of cancers.

The mechanism leading to heterogeneity in subclonal populations is a consequence of inter-cellular genetic variation, followed by selective outgrowth of clones that have a phenotypic advantage within a given tumour micro-environmental resulting in extensive subclonal diversity (Burrell et al., 2013). Moreover, genetically distinct subclonal populations of cells that have distinctive mutational and phenotypic profiles, can also be territorially segregated. It is not clear precisely how spatial separation of genetically distinct clones arises in primary tumours but it is possible that separation reflects the presence of distinct micro-enviromental niches in the primary tumour, such that the subclones occupying each niche evolve relatively independently of one another(Burrell et al., 2013; Marusyk and Polyak, 2010).

2.2 Clonal evolution and heterogeneity in Lymphoma

Lymphoma is a cancer of the lymphatic system, which is part of immunity system. Lymphoma occurs when lymphocytes, a type of white blood cell (i.e. immune system), grow and multiply uncontrollably with characteristic formation of solid tumours at lymph nodes level (Marcus, Sweetenham, and Williams, 2013). Clonal evolution and intratumour heterogeneity have been documented in a range of Lymphoma types and the overall studies have provided a novel perspective to understand the mechanisms behind Lymphoma evolution and its implication in terms of treatment failure and relapse or recurrence (Burrell et al., 2013; Adams and Strasser, 2008; Ma et al., 2019). Several types of intratumour heterogeneity have been reported in Lymphoma patients. One of the most common examples is the evolution of a low-grade lymphoma into a high-grade lymphoma in the same patient during the disease course (i.e. lymphoma transformation). Secondly but less frequently, some patients can be simultaneously diagnosed with a low-grade and a high-grade lymphoma at different anatomical sites. Thirdly and least frequently, two different, usually clonally unrelated lymphoma entities coincide within the same organ (i.e. composite lymphoma) (Schürch et al., 2018).

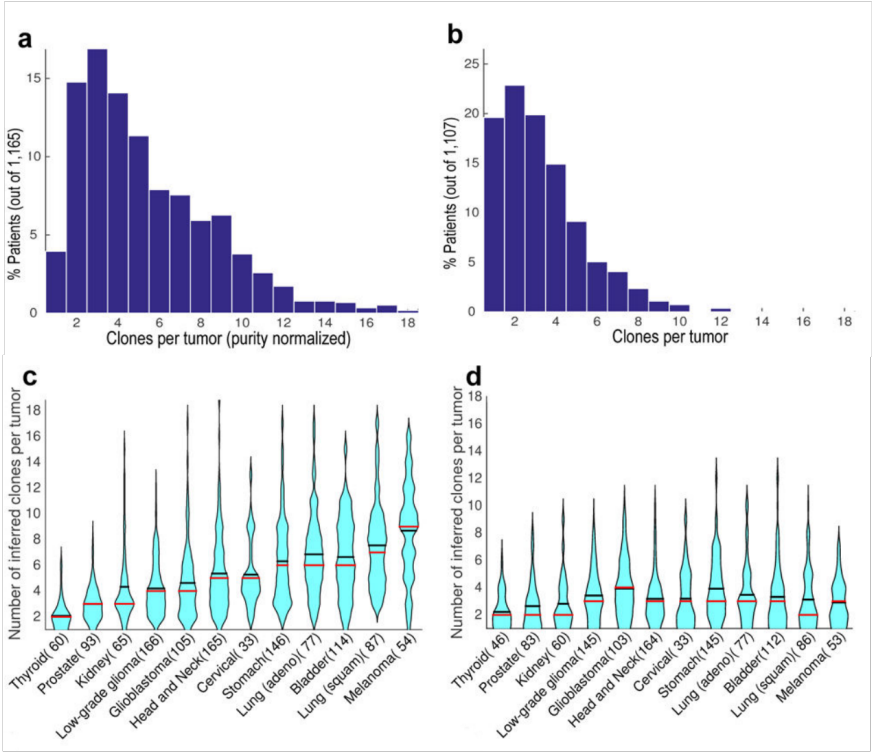


FIGURE 2.3: **Intratumour genetic heterogeneity in 12 tumour types** (a,b) Clone number distribution as predicted by two bioinformatics tools: EXPANDS(a) (Andor et al., 2013) and PyClone(b) (Roth et al., 2014) across tumour types. (c,d) Violin plots of clone number distributions as predicted by EXPANDS (c) and PyClone (d) within different tumour types. The exact number of tumours of each type is indicated in brackets. (Figure from Andor et al., 2016)

One of the most important questions in understanding clonal progression in lymphoma is to determine the nature and number of different subclones within an individual cancer. Moreover, the co-existence of genetically distinct clones within a tumour can result in a network of biological interactions among the distinct clones. Thus, the behaviour of Lymphoma composed of distinct clones might be different from that of a monoclonal tumour (Marusyk and Polyak, 2010).

Pathogenesis of Lymphoma To understand pathogenesis and the molecular mechanism behind Lymphoma, it is first convenient to describe the physiological process of lymphocytes maturation and how the alteration of this process leads to malignant lymphocytes characteristic of Lymphoma.

A lymphocyte is one of the subtypes of white blood cell in a vertebrate's immune system. Lymphocytes include Natural Killer cells (which function in cell-mediated, cytotoxic innate immunity), T cells (for cell-mediated, cytotoxic adaptive immunity) and B cells (for humoral, antibody-driven adaptive immunity). The function of T cells and B cells is to recognise specific non-self antigens, during a process known as antigen presentation. Once they have identified an invader, the cells generate specific responses that are tailored to maximally eliminate specific pathogens or pathogen-infected cells. During the early stages of T and B

cell maturation, a recombination of genes coding for antibodies/immunoglobulins (Igs), B cell receptors (BCR) and T cell receptors (TCR) occurs, resulting in the numerous immune repertoire able to recognise invading pathogens (Figure 2.4). The recombination occurs in the primary lymphoid organs (bone marrow for B cells and thymus for T cells) and it consists of a random fashion rearranges of the variable (V), joining (J), and in some cases, diversity (D) segments that constitute Igs and TCRs genes. The process ultimately results in novel amino acid sequences in the antigen-binding regions of Igs and TCRs that allow for the recognition of antigens from nearly all pathogens including bacteria, viruses, parasites (Figure 2.4).

Igs molecules are composed of heavy and light chains with both constant (C) and variable (V) regions that are encoded by genes on three loci (Figure 2.4):

- Immunoglobulin heavy locus (IgH) on chromosome 14, containing gene segments for the immunoglobulin heavy chain
- Immunoglobulin kappa locus (IgK) on chromosome 2, containing gene segments for the immunoglobulin light chain
- Immunoglobulin lambda locus (IgL) on chromosome 22, containing gene segments for the immunoglobulin light chain

Each heavy chain and light chain gene contains multiple copies of three different types of gene segments for the variable regions of the antibody proteins.

In contrast, in 95% of T cells, TCRs are composed of alpha chain and beta chain. In 5% of T cells the TCR consists of gamma and delta chains (Figure 2.4). The T cell receptor genes are similar to immunoglobulins genes because they both contain multiple V, D and J gene segments in their beta chains (and V and J gene segments in their alpha chains).

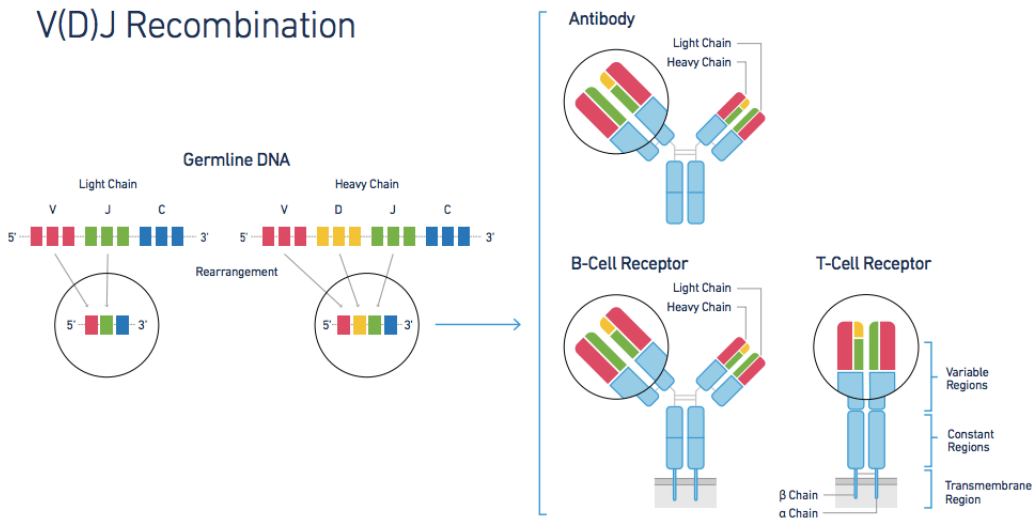


FIGURE 2.4: **TCR and BCR composition.** Antigen specificity and immune system diversity are generated by V(D)J recombination and somatic hypermutation in a clonal manner. Antigen specificity is determined by two co-expressed genes, the heavy and light chains of the B-cell receptor and the alpha and beta chains of the T-cell receptor. Single-cell sequencing reveals the true clonality and diversity of the immune repertoire. (Figure from 10xGenomics website)

To generate a receptor repertoire with sufficient diversity to recognise the high number of potential antigens, the extracellular TCR and Igs contain a variety of complementary determining region 3 (CDR3) that is the main responsible for recognising processed antigen (Figure 2.5 from (Ralph and Matsen IV, 2016)). Like the most variable parts of the molecules, CDRs are crucial to the diversity of antigen specificities generated by lymphocytes. The CDR3 region is created by splicing together specific copies of three gene segments V, D and J segments. At the junctions between V-D and D-J segments, a varying number of nucleotides (N) are deleted and randomly inserted, creating a unique receptor. It is these unique insertions and deletions that are responsible for the vast diversity found in the adaptive immune repertoire. The total repertoire of Ig and TCR molecules is estimated to include nearly 1012 molecules, resulting from combinatory of V(D)J recombinations, somatic mutations, deletions at junction sites and the addition of N-diversity regions between the rearranged genes.

The study of Igs and TCRs rearrangements are powerful markers able to identify the variation patterns of the clonal subpopulations in Lymphoma. Indeed, deletion as well as random insertion of nucleotides between the joined gene segments (N-regions) creates an enormous junctional diversity (Figure 2.5 from Ralph and Matsen IV, 2016). Therefore, the junction regions of rearranged genes are unique 'fingerprint-like' sequences that are different in each healthy B-lymphoid cell (polyclonal), but constant in tumour population (monoclonal), that retains the IGH rearrangement of the B cell giving rise to the tumour clone.

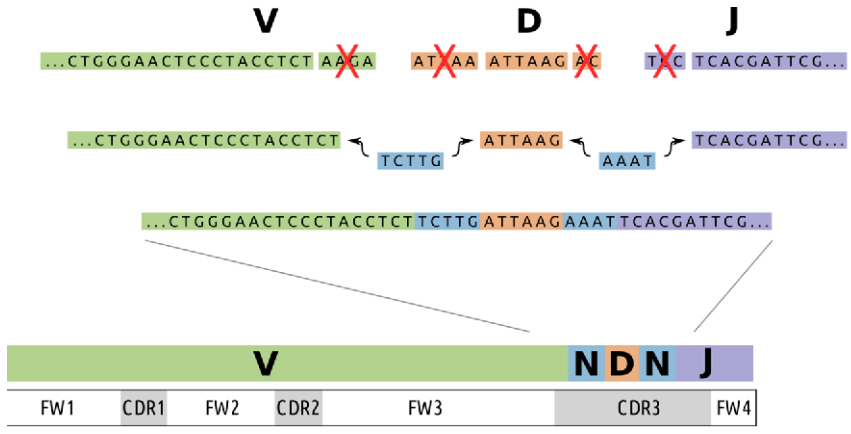


FIGURE 2.5: **VDJ rearrangement.** A V(D)J recombination in a lymphocyte derives from two (or three) germlines V, (D), and J genes that may have been truncated or mutated. The N-diversity regions correspond to random nucleotides inserted between the rearranged genes. (Figure from Ralph and Matsen IV, 2016)

In principle, all tumour cells of a lymphoid malignancy have a common clonal origin with identically rearranged Ig and/or TCR genes. Consequently, the junction regions of these Ig/TCR gene rearrangements can be considered as 'DNA-fingerprints' of the malignant cells, which can be used as *specific signature* for their detection (Dongen et al., 2015). Indeed, the rearrangements in junction regions of IGH and TCR genes are patient-specific and considered **unique molecular signatures that differ in length and composition in each clone and consequently in each patients** (Velden et al., 2003). Specifically, due to the clonal origin of the rearrangement, the analysis of Ig and TCR genes configuration can be used to evaluate *the persistence of malignant clones during the disease course*. Monitoring requires the analysis of both lymphoblasts and normal lymphocytes, and these cells are counted according to their V(D)J recombinations (Beccuti et al., 2017a; Dongen et al., 2015).

Mantle Cell Lymphoma Mantle Cell Lymphoma is a rare subtype of non-Hodgkin Lymphomas (NHL) in which the tumour cells originally come from the mantle zone of the lymph node. It is estimated that MCL patients represent only about 6% (about 4,200 cases) of all new cases of Lymphomas in the United States (Fu et al., 2017). MCL occurs more frequently in older adults where the average age at diagnosis is the mid-60s. It is more often diagnosed in males than in females and white men and women are at a higher risk than black men and women for MCL diagnosis (Fu et al., 2017). Up to now in over 90% of the cases MCL, tumour cells over-express cyclin D1 (BCL1 gene) due to a t(11:14) chromosomal translocation in the DNA, leading to a clonal expansion of malignant B lymphocytes (Zhang et al., 2011). The analysis of the chromosomal breakpoints suggests that this lesion takes place very early during the B-cell ontogenesis as an error during the recombination of the V(variable)-D(diversity)-J(joining) segments of the immunoglobulins gene. Indeed, MCL cells were considered to be derived from naïve B-cell (e.g. pre-germinal B cells not yet exposed to an antigen) which are characterised by a lack of somatic mutations within genes coding for immunoglobulin V-segment heavy chain. These somatic mutations can be detected in 15%-40% of MCLs (Zhang et al., 2011).

2.2.1 Minimal Residual Disease

Clinical markers for NHL patients prognosis are called prognostic factors and include age, initial white blood cell count, cancer subtype, chromosome abnormalities and Minimal Residual Disease (MRD) (Dongen et al., 1998). The MRD is related to monitor the after treatment tumour cells that can not be found in the bone marrow using standard technique (cytomorphologic and cytochemistry exams) but they can be still detected with more sensitive tests (Brüggemann, Raff, and Kneba, 2012). During the last decade, a large number of studies (Schultze and Gribben, 1996; Szczepariski et al., 2001) have shown that detection of very low numbers of malignant cells significantly correlates with clinical outcome in many haematologic malignancies. MRD information is important for clinical decision-making especially during the initial phase of therapy allowing significantly better stratification of patients into risk groups as compared with classical risk groups based on other relevant clinical and biological characteristics (Velden et al., 2003; Dongen et al., 2015). This stratification is useful not only to precisely assess early treatment response and detect relapse but also to identify "low risk" patient and "high risk" patient that could benefit of a reduction or intensification of therapy. Although qualitative MRD information can be highly significant, it only gives limited information and does not allow precise analysis of tumour load kinetics. Therefore (semi) quantitative methods were developed that enable accurate assessment of the number of tumour cells at consecutive follow-up time points. Indeed, quantitative MRD data appeared to be crucial for appropriate evaluation of treatment in NHL patients. (Velden et al., 2003).

For performing MRD analysis, it is necessary to identify ab initio a molecular marker that will be followed during or after treatment: in general, for NHL, rearrangements of IGH or of immunoglobulin light chains (kappa or lambda) can be used. Several studies show that clonal IGH rearrangements detection and MRD monitoring based on these markers are powerful early predictors of therapy response and outcome in B-mature lymphoid tumours (Dongen et al., 2015; Kotrova et al., 2017). However, IGH rearrangement detection is currently part of the routine clinical management of patients affected by Leukemias and currently under validation in other mature lymphoid tumours, as MCL (Pott et al., 2010). In the detail of NHLs, for T-cell lymphoma the molecular marker is the rearrangement of T-cell receptor; for FL, the BCL2-IGH rearrangement; for MCL, the BCL1-IGH rearrangement; for hairy

cell leukaemia, the B-RAF V600E mutation, and for Waldenstrom's macroglobulinaemia the MYD88 L265P mutation (Galimberti et al., 2019).

2.3 Methodologies to detect MRD

The different methods that are used to monitoring the MRD in NHL differ for their sensitivity in discriminating the malignant nature and quantity of remaining blast cells in bone marrow (Campana, 2010). As shown in Table 2.1, three different techniques can be actually used to monitor MRD: Qualitative PCR, Real Time Quantitative Polimerase Chain Reaction (ASOq-PCR), NGS.

TABLE 2.1: **Sensitivity in current technique for MRD detection.** (Data from Galimberti et al., 2019)

Technique	Sensitivity	Target
Qualitative PCR	10^{-5}	IGH,TCR,BCL1-IGH, BCL2-IGH
ASOq-PCR	10^{-4}-10^{-5}	IGH,TCR,BCL1-IGH,BCL2-IGH
NGS	10^{-4}-10^{-5}	IGH,TCR,Mutations

In the next paragraphs, all the methodologies will be explained with the application on Mantle Cell Lymphoma.

Qualitative PCR Qualitative PCR was the pioneer molecular approach implemented for NHLs in 1990. PCR approach is a method widely used in molecular biology to e exponentially amplify specific DNA regions to generate thousands to millions of more copies of that particular DNA segment. Thus, qualitative PCR can be performed using fluorescent primers in order to run the PCR product on a DNA sequencer and reveal the molecular marker under analysis. Qualitative PCR is able to detect up to one clonal cell among 100,000 analyzed ($1x10^{-5}$) (2.1, Galimberti et al., 2019). Molecular targets detectable by Qualitative PCR in MCL are: IGH clonal rearrangements (detectable in more than 60% of patients), translocation of BCL1 or BCL2 genes and TCR marker.

Real Time Quantitative Polimerase Chain Reaction - ASOq-PCR The introduction in molecular diagnosis of quantitative PCR technique, has amplified the knowledge about MRD significance, due to the possibility to identify and quantify specific sequence of DNA and complementary DNA (cDNA) , through the use of fluorescence probes. The fragmentation of these probes during elongation phase of amplification reaction permits fluorescence emission that is captured e quantified. This leads to an elevate reproducibility and specificity in results (Kotrova et al., 2015).

For ASOq-PCR assays, IGH rearrangements are screened using single or multiplex PCR approaches followed by Sanger sequencing allowing to obtain the complete IGH sequence revealing the *patient-specific and clone-specific insertions* (N) among the V, D, J recombined regions (Figure 2.5). Indeed, the rearrangement in junction region of IGH and TCR genes are patient-specific and considered a **unique molecular signatures that differ in length and composition in each clone and consequently in each patient** (Velden et al., 2003). The N insertions are in particular required to set the Allele Specific Oligonucleotide primer for the ASOq-PCR technique. (Galimberti et al., 2019). ASO primer is a typically oligonucleotide of 15–21 nucleotide bases in length designed (and used) in a way that makes it specific for only one version or allele of the DNA being tested. ASO can be used in different approaches:

- ASO probe approach: the probe is positioned in the tumour-specific sequence, in particular the junction region of Ig/TCR gene rearrangements. The probe is used in combination with a forward and reverse primer, which are positioned in germline sequences opposite of the tumour-specific sequence.
- ASO forward primer approach: forward primer positioned in the junction region in combination with a germline reverse primer and a germline probe.
- ASO reverse primer approach: it is comparable to the ASO forward primer approach, but with opposite location of the germline primer and probe relative to the tumour-specific sequence (Velden et al., 2003).

If IGH/TCR gene rearrangements are chosen as MRD-PCR target, the ASO forward primer approach (using a probe generally located in the J gene segments) has several advantages over the ASO reverse primer approach (using a probe generally located in the V gene segments). Firstly, the number of J gene segments is lower than the number of V gene segments, and consequently a lower number of probes need to be made. Secondly, the occurrence of somatic hypermutations in the V gene segment may result in less optimal primer and probe annealing (Velden et al., 2003).

ASOq-PCR is actually defined as the gold standard approaches for MRD evaluation because of the reached sensitivity levels, detecting up to one clonal cell among 100,000 analyzed ($1x10^{-5}$) (Table 2.1). Nevertheless, ASOq-PCR still has some potential technical and biological pitfalls that affect its use in NHLs: for example, applicability of methods is limited to patients with the presence of genetic rearrangements appropriate to reach a sufficient sensibility. Moreover, ASOq-PCR requires a time- and labor-intensive process of patient-specific ASO primer development. As a result, outside of centres where ASOq-PCR has been established and validated, access to this method is limited. Moreover, for the IGH-positive cases, the set of the standard curve necessary for MRD quantification could be affected by the entity of tissue tumour infiltration at diagnosis, so making the quantification not always accurate.

Next Generation Sequencing techniques Next Generation Sequencing, also known as high-throughput sequencing, is the term used to describe a number of different modern sequencing technologies that permit to sequence DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing (the principal method of DNA sequencing since its invention in the 1970s). NGS applied to Lymphoma samples means that it is possible to universally amplify IGH genes belonging to the tumour cells and identifies all clonal gene rearrangements at diagnosis, allowing monitoring of disease progression and clonal evolution during therapy (Faham et al., 2012). However, NGS does require the availability of a sufficient amount and quality of tumour tissue to establish the tumour clonotype. The general protocol implies a first fragmentation of the genomic strand under analysis (e.g. VDJ gene segments), and the identification of the bases in each fragment by emitted signals when the fragments are ligated against a template strand. The strong point of NGS is the possibility to perform millions of these reactions in parallel, resulting in very high speed and throughput at a reduced cost (Xuan et al., 2013). Next generation methods of DNA sequencing have three general steps:

1. Library preparation: libraries are created using random fragmentation of DNA gene segments, followed by ligation with custom linkers
2. Cluster amplification: the library is loaded in a flow cell and it is amplified using clonal amplification methods and PCR

3. Sequencing: during a sequencing reaction, sequenced base pairs or "reads" are generated with a length around 100-250 bp reads. First of all, longer fragments are ligated to generic adaptors and annealed to a slide using the adaptors. PCR is carried out to amplify each read, creating a spot with many copies of the same read. They are then separated into single strands to be sequenced. The slide is flooded with nucleotides and DNA polymerase. These nucleotides are fluorescently labelled, with the colour corresponding to the base. They also have a terminator, so that only one base is added at a time. In each read location, there will be a fluorescent signal indicating the base that has been added. The process is repeated, adding one nucleotide at a time. The flow cell is imaged and the emission from each cluster is recorded. This cycle is repeated n times to create read longer n bases. All of the sequence reads will be of the same length since the read length depends on the number of cycles carried out

4. Data analysis: The result obtained from NGS experiments is a collection of reads of the same length representing the VDJ rearrangements of the tumour clones and also of the normal immune repertoire of normal lymphocytes (i.e. VDJ rearrangements is a physiological event), divided on the basis of the samples analyzed (e.g. timeline of samples from diagnosis to several time points). For MRD monitoring purpose, data analysis of the reads need to respect some rules. Recognition of clones has to be based: (i) on mismatches or indels in sequences rearrangements (especially in the junction region that is reach of N nucleotides); (ii) on clustering similar clones taking into consideration the sequences with very small differences that can be derived from different clones, especially if these differences occur in N-diversity regions (Giraud et al., 2014); (iii) on quantification of each clone. Moreover, one of the most frequent problem after the sequencing is the presence of incomplete short reads that did not contain the whole rearrangement. Several short reads had to be assembled to obtain longer reads covering the whole rearrangement, requiring that the reads were sufficiently redundant. All these features can not be managed with standard reads mapping tools because principally they cannot deal with reads containing recombinations, somatic mutations or large insertions and therefore a large amount of data is lost.

ASOq-PCR vs NGS: advantages and disadvantages ASOq-PCR and NGS are both powerful methods for MRD monitoring and IGH rearrangements detection. However, recently, several studies have suggested that the IGH monitoring through deep sequencing techniques can produce not only comparable results to ASOq-PCR methods, but also might overcome the classical technique in terms of feasibility and sensitivity (Kotrova et al., 2015; Kotrova et al., 2017; Beccuti et al., 2017a; Ladetto et al., 2014).

1. A property of ASOq-PCR for the amplification of clonal rearrangement of IGH and TCR genes is an accurate quantification of MRD due to the fact that these rearrangements are a unique molecular signature which can be detected with a sensitivity of 0.01% to 0.001%. However, applicability of methods is limited to patients with IGH/TCR genetic rearrangements with junction regions appropriate to reach a sufficient sensibility (Kotrova et al., 2015). NGS can overcome this limitation since it has the potentiality to reach a major sensibility (until $1x10^{-6}$) and it allows the detection of clones at lower concentrations than is possible with conventional techniques in the study of V(D)J repertoires.

2. ASOq-PCR amplification requires considerable expertise and a labor intensive target identification that can be further complicated by the loss of same target because of clonal evolution. Moreover, careful consideration must be given to the presence of minor clones, which might be undetected at diagnosis and can that become predominant

during the course of the disease, with the potential of false-negative results. Targeting two or more different rearrangements has been recommended to avert this problem, but multiple sensitive probes are not identifiable in approximately 30% of cases (Campana, 2010). Instead, NGS allows multiclone follow-up and the detection of emerging subclones at diagnostic concentrations far below that of the main clone identified at diagnosis, as well as a full repertoire analysis (Faham et al., 2012).

3. ASOq-PCR requires the development of reagents and assay conditions for each individual patient, which is laborious and time-consuming (Campana, 2010). NGS can overcome this disadvantage of PCR-based methods like need of specific probes because the sequencing assay uses a set of universal primers and it can assess multiple clonal rearrangements in every patient without the need of individualised procedures and further increase sensitivity, specificity, accuracy and reproducibility.

Several studies have been performed to prove the strength of NGS approach with respect to ASOq-PCR (Ladetto et al., 2014; Kotrova et al., 2015). Ladetto et al., 2014 analyzed 378 samples from 55 patients with ALL, MCL or multiple myeloma (MM) with ASOq-PCR and NGS for clonotype identification, clonotype identity and comparability of MRD results between the two methods (Figure 2.6 and 2.7 from Ladetto et al., 2014). Figure 2.6 shows a comparison of the feasibility of the two approaches (ASOq-PCR versus NGS). All 15-ALL cases were evaluable by PCR and NGS. Among the 30 MCL cases, 22 patients were evaluable by PCR and 26 cases could be effectively monitored by NGS, including four in which ASOq-PCR failed. Instead for MM both PCR and NGS evaluated the same number of samples; however two of MM cases that were not evaluable by ASOq-PCR due to sequencing failure, then could be evaluated by NGS. Overall 70% of samples could be evaluated using both methods and were employed for the following concordance analysis. The concordance analysis demonstrated a significant level of concordance (R=0,791) between the ASOq-PCR and NGS methods for MRD quantification indeed 79.6% of cases were classified as fully concordance and 20.4% were scored as discordant (Ladetto et al., 2014)(Figure 2.7).

Table 1. Rates of success of RQ-PCR and NGS among ALL, MCL and MM by patient

Disease	*Patients*	*Patients evaluable PCR*	*Patients evaluable NGS*	*Patients evaluable with both tools*	*Patients evaluable with at least one tool*	*Patients not evaluable*
ALL	15	15	15	15	15	0
MCL	30	22	26	22	26	4
MM	10	8	8	6	10	0
TOT	55	45	49	43	51	4

Abbreviations: ALL, acute lymphoblastic leukemia; MCL, mantle cell lymphoma; MM, multiple myeloma; NGS, next-generation sequencing; RQ-PCR, real-time quantitative polymerase chain reaction.

FIGURE 2.6: **Rates of success of ASOq-PCR and NGS among ALL, MCL and MM by patient.** All 15-ALL patients were evaluable by PCR and NGS (5th column). Among the 30 MCL cases, 22 patients were evaluable by PCR (3th column) and 26 cases could be effectively monitored by NGS (4th column), including four in which ASOq-PCR failed. Instead for MM both PCR and NGS evaluated the same number of samples (3th and 4th column). (Figure from Ladetto et al., 2014)

Also Kotrova et al., 2015 prove NGS strength in MRD detection sequencing 210 samples from 76 ALL-patients treated by Berlin-Frankfurt-Munster-based protocol and investigated the changes in the treatment stratification (Kotrova et al., 2015). The results showed that NGS provided a precise prediction of relapse and NGS risk group assignment was optimal. Moreover, it has the potential to achieve a higher level of sensitivity (0.0001%) respect to ASOq-PCR (0.001%) (Kotrova et al., 2015). In the context of Lymphomas, studies from

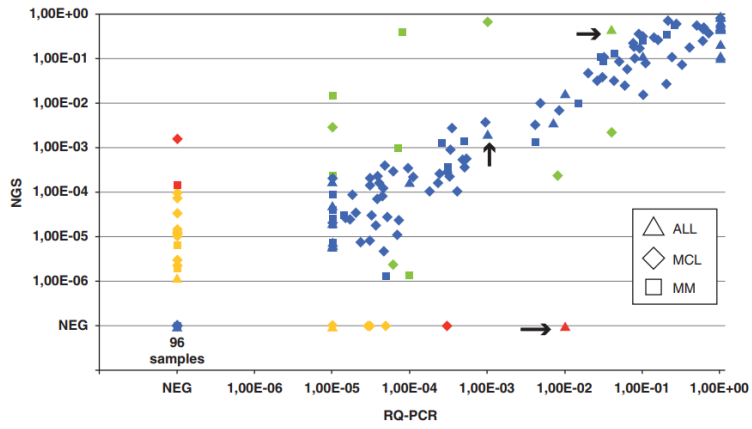


FIGURE 2.7: **Correlation analysis of ASOq-PCR and NGS results.** The regression curve showed a significant concordance (P<0.001, R=0,791). Concordant cases are shown with a blue icon; arrows indicate cases in which clonal evolution was documented. The majority of the discordance could be attributed to minor quantitation differences related to low input cells (Figure from Ladetto et al., 2014).

(Ladetto et al., 2014; Herrera et al., 2016) have demonstrated that NGS can identify a suitable immunoglobulin clone for subsequent MRD assessment in 83% to 94% of patients with Diffuse Large B-cell Lymphoma (DLBCL) and 85% to 100% of patients with MCL (Ladetto et al., 2014; Herrera et al., 2016).

2.4 Aim

In Mantle Cell Lymphoma the analysis of B cell clonality is used to monitor the Minimal Residual Disease (MRD) response to long-lasting remission and improve patients outcome. ASOq-PCR based method is a useful predictive technique for MRD detection in Lymphoma, stratifying patients in prognostic groups based on relapse risk. This is crucial for an adapted therapy, offering intensive treatment only to patients with suboptimal response. NGS methods are powerful tools for MRD monitoring and IGH rearrangements detection and can exceed some limitations of ASOq-PCR method thanks to NGS high sensitivity, specificity, accuracy and reproducibility (Beccuti et al., 2017a). More importantly, NGS approach allows multi-clone follow-up and the detection of emerging subclones at diagnostic concentrations far below that of the main clones identified at diagnosis, as well as full repertoire analysis. However, the huge data obtained from NGS must be analyzed with appropriate bioinformatics algorithms able to manage complex data. Indeed, considering the total repertoire of samples collected for each patient (e.g. diagnosis and follow-up samples) and analysed through NGS, an average of 300.000 reads per sample is present and each obtained read could possibly contain a putative clone. Therefore, a combinatorial research of all possible clones contained in each read is necessary, considering all the recombinations of known VDJ gene segments. Moreover, to identify the most significant tumour clones during the evolution of MCL, the combinatorial research has to be applied for diagnosis sample and for all patient samples collected during time. Thus, it is a long-time work and it is necessary to develop efficient algorithms in term of accuracy, execution time and space complexity. **The role of this part of the thesis is to present an innovative bioinformatics pipeline for NGS-MRD analysis that can be applied to MRD detection and tumour clones identification without**

the supporting ASOq-PCR technique. This pipeline includes an easy-to-use and reliable bioinformatics tool, **HashClone** that provides a clonality assessment and the MRD detection over time (Beccuti et al., 2017a; Romano et al., 2018). *HashClone is based on a parallel implementation of four C++ applications for the data processing based on the analysis of all set of sample reads simultaneously. HashClone returns the corresponding set of tumour clones aligned with respect to V-GENE, J-GENE, and D- GENE sequences and associated with their frequency in all input samples. Moreover, the fifth application of the tool is dedicated to data visualisation with a graphic user interface that allows the inspection of the tumour clones and their associated features in all the samples under study.* HashClone is based on an ad-hoc C++ hash table class implementation to optimise the trade-off between the memory utilisation and the execution time. Finally, HashClone is integrated in a Docker container, a platform that allows to easily install and run the application packaged with all its dependencies and libraries.

In this thesis we demonstrated that HashClone has best performance to investigate clonotypes and clone abundance in Lymphoma samples respect to the standard technique (ASOq-PCR). We applied HashClone in the context of a particular rare type of Lymphoma, Mantle Cell Lymphoma, with the aims to detect the major B-cell tumour clone and monitor it over time during clinical follow ups. We performed an entire clonality analysis with NGS technology in the experimental procedure and HashClone algorithm for data analysis.

We performed three types of experiments:

- Study 1: Simulated datasets to test HashClone performances
- Study 2: Clonality analysis of cohort 1 (i.e. 5 MCL patients)
- Study 3: Clonality analysis of cohort 2 (i.e. 23 MCL patients)

In Study 2 and 3, patients were investigated for IGH detection and MRD monitoring using a new designed amplicon-based NGS approach and HashClone analysis. Both studies involved patients enrolled in a phase III multicenter, randomized trial with Lenalidomide drug maintenance after induction regimen containing Rituximab followed by high dose chemotherapy in adult patients with advanced MCL (Fondazione Italiana Linfomi).
Through Study 2, we tested the ability of HashClone to identify clones with the best performance in sensibility, sensitivity and flexibility to manage different kinds of samples. Moreover, we showed that HashClone can overpass limitation of the current technique, ASOq-PCR and the state-of-art tool, Vidjil (Beccuti et al., 2017a; Romano et al., 2018). Moreover, through Study 3, we used HashClone to perform a clonality analysis of 23 patients without ASOq-PCR counterpart and showed differences among patients in terms of characteristic MRD clones and their trend during the entire therapy protocol.
HashClone pipeline has a big application potentiality not only with Lymphoma but also with a multitude of cancer types since it can be putative apply to all kinds of samples in order to establish the clonal evolution. This part of the thesis was conducted at Department of Computer Science, University of Torino, in collaboration with the Division of Hematology, Department of Molecular Biotechnologies and Health Sciences, University of Torino, and Euroclonality division of ESLHO (European Scientific foundation for Laboratory HematoOncology).

2.5 State-of-art of the computational methodologies for MRD monitoring

NGS MRD approach might provide a full repertoire analysis through multi-clones detection at diagnosis, and it gives the opportunity to monitor all the neoplastic clones at several follow

ups. However, this issue requires suitable computational algorithm. Actually, the large volume of data, collected thanks to the advent of deep sequencing technologies, raises multiple challenges in data storage and data analysis, to efficiently finding the best match between a rearranged sequence and the V, D and J germlines. In literature, there are several tools as JoinSolver (Souto-Carneiro et al., 2004), HighV-QUEST (Giudicelli, Chaume, and Lefranc, 2004), iHMMune-align (Gaëta et al., 2007), SoDA2 (Munshaw and Kepler, 2010) , VDJSeq-Solver (Paciello et al., 2015), ARRest/Interrogate (Bystry et al., 2016) and ViDJil (Giraud et al., 2014) currently implemented for marker screening and detection of IGH rearrangements on a set of reads obtained from deep sequencing experiment of a single sample. All of them try to assign a specific V, D and J alleles to a unique sequence, extracted in laboratory via high-throughput sequencing experiments (Giraud et al., 2014).

- **JOINSOLVER** (Souto-Carneiro et al., 2004) is a web-based software program developed for human immunoglobulin V(D)J recombination analysis created by the National Institutes of Health, National Institute of Arthritis and Musculosketelal and Skin Diseases and the Center for Information Technology. JOINSOLVER was developed to analyze the complementarity-determining region 3 (CDR3) of the immunoglobulin genes in human B cells which includes the IGHD gene and its junction with the IGHV and IGHJ genes. The length of the D segment in CDR3 is extremely short (3-5 nucleotides) making the identification of this segment very difficult. JOINSOLVER uses a consecutive matching approach to assign the D segments and to limit the portion of the sequences that must be analyzed to identify V - J regions. JOINSOLVER interrogates the input sequence to find the beginning of the CDR3 region that is characterised by conserved motif in most of the human V germline genes. When, the V segment is defined, JOINSOLVER screens the sequence in order to identify the limits of the J segment. Once the CDR3 region is identified, V, J, and D assignment is done comparing the sequences stored in IMGT reference database (Giudicelli, Chaume, and Lefranc, 2005).

- **IMGT HighV-QUEST** (Giudicelli, Chaume, and Lefranc, 2004) is a web portal supported by the international ImMunoGeneTics information system (IMGT) consortium for the analysis of rearranged nucleotide sequences of the antigen receptors (immunoglobulins or antibodies and T cell receptors) obtained from deep sequencing data. IMGT HighV-QUEST is the high-throughput version of IMGT V-QUEST, it is able to analyze 500.000 nucleotide sequences per run. IMGT HighV-QUEST allows (i) the identification of the closest V, D and J genes and alleles (ii) the IMGT/JunctionAnalysis (iii) the description of mutations, and (iv) the characterisation of IMGT clonotypes. The IMGT HighV-QUEST standard output includes a summary file which contains the V(D)J annotation, the score of the alignment, the V(D)J identity percentage and the identity nuclotides for each sequence. Moreover, the aminoacid sequences, the junction frame and potential insertions/deletions in the V(D)J region are reported. In addition to the summary file, IMGT HighV-QUEST reports more detailed files with the complete analysis of nucleotide sequences, aminoacids sequences, junction regions and possible nucleotide mutations.

- **iHMMune-align** (Gaëta et al., 2007) is an alignment program that uses a Hidden Markov Model (HMM) to model the processes involved in human IGH gene rearrangement and maturation. iHMMune-align was developed by the School of Biotechnology and Biomolecular Sciences, School of Computer Science and Engineering, The University of New South Wales (Sydney, Australia). First of all, iHMMune-align

identify the IGHV gene with a local alignment step of the sequence with the human IGHV germline repertoire stored in IMGT. The pre-alignment of IGHV gene allows iHMMune-align to calibrate the emission probabilities of the HMM through the estimation of somatic mutation amount over the sequence. Initially, the HMM is developed on rearranged sequences without mutations. The emission probabilities in the match states are re-computed to model the process of somatic mutation. This probability is adjusted to take into account the position of the putative mutations, the local sequence context and the effect of antigen selection. The HMM is finally aligned with the rearranged, mutated sequence, using the Viterbi algorithm. The program outputs the alignment corresponding to the optimal path along the HMM and reports the germline genes.

- **SoDA2 (Somatic Diversification Analysis 2)** (Munshaw and Kepler, 2010) is a tool based on a HMM that compute the posterior probabilities of candidate VDJ rearrangements of Ig genes and find those with the highest values among them. SoDA2 was developed by the Center for Computational Immunology, Computational Biology and Bioinformatics Program and the Department of Biostatistics and Bioinformatics (Durham,USA). SoDA2 aligns the target sequence with a consensus-like sequence of the V families to determine if the input sequence is a heavy, kappa or lambda chain. Then a pre-alignment step of V and J gene is performed in order to submit to the HMM all V and J segments with highest likelihood alignments. The HMM is used to compute the emission probability in 10 states (see Munshaw and Kepler, 2010 for more details about the states). Then, the total probability of a proposed rearrangement is calculated using the forward and backward algorithms. The gene segments leading to the highest posterior probabilities are selected to identify the path with the highest posterior probability for each possible rearrangement through posterior Viterbi algorithm approach. The final output is composed of the top rearrangement candidate (highest posterior probability) and also rearrangements associated with high posterior probabilities in order to have a complete picture of the input sequences.

- **VDJSeq-Solver** (Paciello et al., 2015) is a tool for the identification of clonal lymphocyte populations from paired-end RNA Sequencing reads. The tool was developed by the Department of Control and Computer Engineering, Politecnico di Torino (Torino, Italy). The tool detects the main clone characterising the tissue of interest by recognizing the most abundant V(D)J rearrangement in the sample. VDJSeq-Solver implement a unique pipeline composed of several existing tools: TopHat, Bowtie, BEDTools, BLAST alignment and Shrimp. VDJSeq-Solver performs a first alignment of the reads to the reference genome using both Bowtie and Tophat. The output of Bowtie contains the reads not mapped on the genome (VDJ unmapped); the output from Tophat contains all reads mapped on the reference genome that become the input of BEDTools to extract all reads mapped on the V,D,J gene segments (VDJ mapped). Then, VDJ mapped and VDJ unmapped reads are aligned against V and J gene segments using Blast in order to identify a VJ recombination (VDJ encompassing reads). Finally D gene segments are identified using Shrimp tool. The output of VDJSeq-Solver represents the list of all identified clones.

- **ARRest/Interrogate** (Bystry et al., 2016) is a web-based interactive application for Ig and TCR (T cell receptor) immunoprofiling. ARRest/Interrogate was developed by CEITEC – Central European Institute of Technology, Masaryk University (Brno,

Czech Republic). The application includes four functions: input processing, data selection and filtering, comparative calculation and data visualisation. The first function deals with V,D and J gene annotation through IMGT. Genes and alleles are combined with the amino acid sequence of the junction regions to construct IMGT-like clonotypes. Data selection and filtering allow the user to select groups of samples depending on several features. The third function (i.e. comparative calculation) can calculate and visualise differences among samples comparing the features (single feature or entire feature type) on the basis of their abundance across the samples. The visualisation function allows to visualise the results using several type of graph as bubble charts, heatmaps, bar graph, PCA scatterplots or statistical plots.

- **Vidjil** (Giraud et al., 2014) is an open-source platform for the analysis of high-throughput sequencing data from lymphocytes, developed by the Bonsai bioinformatics team and its collaborators. Vidjil processes NGS sequencing data in order to extract V(D)J junctions and gather them into clones for quantification. To quantify the clonotype abundances starting from a set of reads, the method proceeds through a first ultrafast prediction of short string-sequence, overlapping the third complementarity-determining region (CDR3) in order to contain the junction region and part of the V region and J region of each rearrangement. Each clonotype abundance is then estimated using the number or reads containing the same string-sequence. Finally it is selected one representative sequence per clone. Vidjil is implemented in C++ open-source program (Giraud et al., 2014).

2.6 HashClone pipeline

HashClone pipeline was published in 2017 on BMC Bioinformatics (Beccuti et al., 2017a).

The HashClone strategy is organised on three steps (Figure 2.8). The *significant k-mer identification* (Step 1) and the *Generation of read signatures* (Step 2) implement an *alignment-free* prediction method that identifies a set of putative tumour clones from patient's samples; while in *characterisation and evaluation of the cancer clones* (Step 3) the IGHV, IGHJ and IGHD identification is obtained via the alignment of rearrangements with respect to the IMGT reference database (Giudicelli, Chaume, and Lefranc, 2005). A detailed description of these three steps is now reported.

HashClone - Description of the strategy

Significant k-mer identification (Step 1) In this step the entire set of reads for each of the n patient's samples is scanned and a set of sub-strings of length k, namely k-mers, is generated using a *sliding window* approach. For instance given the read *ATCCCGTC* the following k-mers with $k = 3$ are generated: ATC, TCC, CCC, CCG, CGT and GTC.

Formally, given an alphabet $\mathcal{L} = \{A, C, T, G\}$ where the letters correspond with DNA-bases we define ρ, namely *read*, as a string over $\mathcal{L}$ of arbitrary length m, and A_k^* as the set of strings of length k constructed from $\mathcal{L}$. Then, $A_k^\rho = \{\alpha_1^k, \alpha_2^{k+1}, \ldots, \alpha_{m-k+1}^m\}$ is the set of strings of length k generated from ρ using *sliding window* approach s.t. α_p^{k+p-1} is the sub-string of ρ starting at position p, spanning k characters and ending at $k + p - 1$. We define the function:

$$\mathcal{C} : A_k^* \rightarrow \mathbb{N}^n \tag{2.1}$$

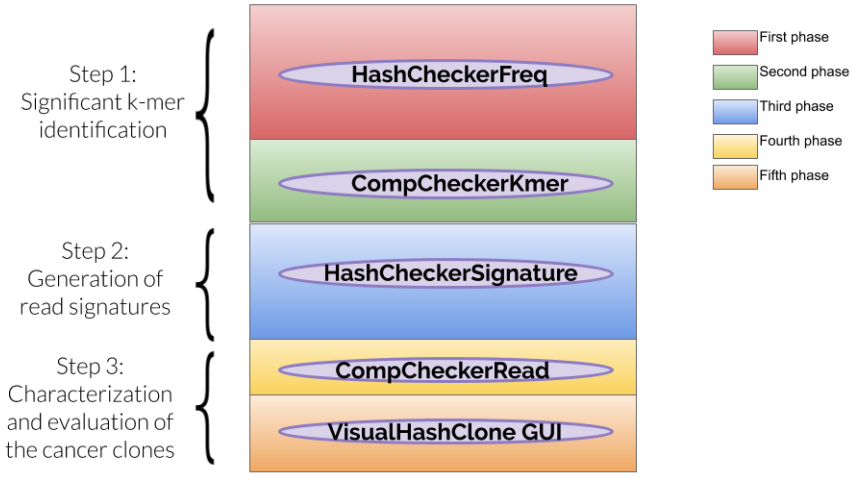


FIGURE 2.8: **HashClone pipeline.** The first step regards the *significant k-mer identification* considering all samples to be analyzed and generating the set of *k*-mers; the second step is focused on the *generation of read signatures* leading to the identification of the set of putative clones from patient's samples; the third step is dedicated to the *characterisation and evaluation of the cancer clones*.

s.t. for each k-mer returns a vector listing the total number of times this k-mer appears in any patient's sample (i.e. k-mer frequencies for patient's samples). Thus, $\mathcal{C}(\alpha)[i] = h$ with $1 \leq i \leq n$, iff k-mer α is present in h reads of the sample i.

Then, a k-mer α is defined as *significant* iff $\exists 1 \leq i, j \leq n$ such that:

$$\begin{cases} |log_{10}(\mathcal{C}(\alpha)[i]) - log_{10}(\mathcal{C}(\alpha)[j])| \geq \tau, & if\ \mathcal{C}(\alpha)[i] \neq 0 \wedge \mathcal{C}(\alpha)[j] \neq 0 \\ log_{10}(\mathcal{C}(\alpha)[j]) \geq \tau, & if\ \mathcal{C}(\alpha)[i] = 0 \wedge \mathcal{C}(\alpha)[j] \neq 0 \\ log_{10}(\mathcal{C}(\alpha)[i]) \geq \tau, & if\ \mathcal{C}(\alpha)[i] \neq 0 \wedge \mathcal{C}(\alpha)[j] = 0 \end{cases} \tag{2.2}$$

where τ is a user-defined parameter. The choice of an appropriated τ value can impact on the capability of HashClone to identify clones. A detailed analysis about this aspect and the set of τ value used in the *Pilot1* and *Pilot2* experiments are reported in the Supplementary Material (6.1).

Moreover, we introduce the following function:

$$\mathcal{CH} : A_k^* \rightarrow \{\textbf{TRUE,FALSE}\} \tag{2.3}$$

that takes as input a k-mer α and returns ***TRUE*** iff α is a *significant k-mer* otherwise ***FALSE***. For instance, assuming $n = 3$, $\tau = 1$, and $\mathcal{C}(ATC) = \langle 1000, 2000, 25000 \rangle$ then $\mathcal{CH}(ATC)$ returns ***TRUE*** because $|log_{10}(\mathcal{C}(ATC)[1]) - log_{10}(\mathcal{C}(ATC)[3])| \geq 1$.

Thus, $\mathcal{CH}$ function is used to derive the set of *significant k-mers* $\Psi = \{\psi_1, \ldots, \psi_t\}$.

Generation of read signatures (Step 2) This step takes as input the set Ψ of all the *significant k-mers*, and it generates the read signatures. Given a patient's sample i, for each read ρ all its k-mers are analyzed to derive the corresponding read signature. A k-mer $\alpha \in A_k^\rho$

is selected iff $\alpha \in \Psi$, then all the selected k-mers are combined to generate a read signature according to their positions in ρ.

For instance, considering the read *ATCCCGTC* and assuming CCC, CCG, CGT the only *significant k-mers* in the read the corresponding signature is CCGT. Defined $\Gamma_i = \{\gamma_1, \ldots, \gamma_e\}$ the set of read signatures obtained for the sample i, the function:

$$\mathcal{CS} : \Gamma_i \rightarrow \mathbb{N} \tag{2.4}$$

returns the total number of reads of sample i in which the signature γ appears (i.e. signature frequency in patient's sample i).

When the entire set of reads of sample i is scanned, the set of generated signatures Γ_i is processed to identify those similar (with respect to a fixed number of mismatches, insertions and deletions) using a Smith-Waterman algorithm. Practically in this correction step two signatures $\gamma, \gamma' \in \Gamma_i$ are considered similar if their alignment score computed by Smith-Waterman algorithm is greater than a specified threshold T. Hence, the signature γ with lower frequency is removed from the set of signatures and its frequency is added to the frequency of the other signature γ', i.e. $\mathcal{CS}(\gamma') = \mathcal{CS}(\gamma) + \mathcal{CS}(\gamma')$

Characterisation and evaluation of the cancer clones (Step 3) This step takes as input the sets of signatures $\Gamma_1, \ldots, \Gamma_n$ generated from each patient's sample in the Step 2. We define the set of putative cancer clones Δ (initially empty), and the function:

$$\mathcal{CC} : \Delta \rightarrow \mathbb{N}^n \tag{2.5}$$

that for each clone δ returns a vector listing the total number of times this clone appears in any patient's sample.

Δ is incrementally updated processing the signatures into each set Γ_i (starting from Γ_1 to Γ_n). For each signature $\gamma \in \Gamma_i$ a similar putative cancer clone is searched in Δ. The similarity between a clone and a signature is evaluated using the same strategy proposed for the correction step. If a similar clone is not found then a new one identified by the signature sequence γ is inserted in Δ and its associated frequencies are defined as follows: let γ be a signature in Γ_i and δ the corresponding new clone then $\forall 1 \leq j \leq n \wedge j \neq i \Rightarrow \mathcal{CC}(\delta)[j] = 0$, while for $j = i \Rightarrow \mathcal{CC}(\delta)[j] = \mathcal{CS}(\gamma)$. Instead, if a similar clone is found then its frequencies are updated as follows: let γ be a signature in Γ_i and the δ the corresponding similar clone then $\mathcal{CC}(\delta)[i] = \mathcal{CC}(\delta)[i] + \mathcal{CS}(\gamma)$.

Finally, the putative cancer clones in Δ are verified exploiting biological knowledge. Indeed, all the identified putative clones are analyzed and evaluated using IMGT reference database (`http://www.imgt.org/download/GENE-DB/`, Giudicelli, Chaume, and Lefranc, 2005). For each clone, its best alignments with respect to V-GENE, J-GENE, and D-GENE are reported and ranked according to a similarity measure (i.e. matched bases divided matched and unmatched bases).

2.6.1 HashClone - Implementation details

HashClone strategy described above, has been implemented thanks to tool suite specifically developed for this purpose. This tool suite, called HashClone, is composed of four C++ applications for data processing and one HTML5+Javascript application for the data visualisation. Moreover, VisualHashClone Java-GUI has been also developed to simplify the data visualisation phase (Figure 2.8).

Data processing applications are:

- Phase 1: *HashCheckerFreq* takes as input reads of a patient's sample and returns the corresponding set of k-mers associated with their frequency in the input reads. The k-mers and their frequency are stored in RAM as an associative array achieved through a C++ hash table class specifically implemented to optimize the trade-off between the memory utilization and the execution time. Observe that this class implements a *separate chaining* as collision resolution policy to deal with the case of different k-mers having a similar hash value.

- Phase 2: *CompCheckerKmer* takes as input all the k-mers derived by all the patient's samples and their frequencies, and it analyses the k-mer frequencies in each patient's sample to derive the set Ψ of *significant k-mers* (as defined in Eq.2.2). This is achieved by exploiting an associative array, implemented through a *red-black tree* data structure. Hence, in this associative array the array keys are the k-mer sequences and the array values the k-mer frequencies. In this application, a *red-black tree* data structure was used (instead of hash table) because we are going to investigate the possibility of implementing an efficient correction step (up to m mismatches) based on the characteristic of this data structure.

- Phase 3: *HashCheckerSignature* takes as input the *significant k-mers* and the set of reads of i^{th} sample and returns the set of read signatures for this sample (i.e. Γ_i) with their frequencies. The k-mers are stored using the implemented hash table class, while the generated signatures are stored using red-black tree. A correction step identifying similar signatures (with respect to mismatches, insertions and deletions) is performed exploiting the implementation of the Smith-Waterman algorithm provided by SIMD Smith-Waterman C++ library (Zhao et al., 2013). In our implementation the T threshold previously introduced (in the Step 2, Generation of the read signature) to discriminate between similar reads is automatically derived as follows:

IF $max(size_{\gamma_1}, size_{\gamma_2}) * 0.7 > min(size_{\gamma_1}, size_{\gamma_2})$
THEN RETURN $(max(size_{\gamma_1}, size_{\gamma_2}) * M)$
ELSE RETURN $((M * 4/5 - MM * 2/50 - IN * 2/10) * max(size_{\gamma_1}, size_{\gamma_2}))$

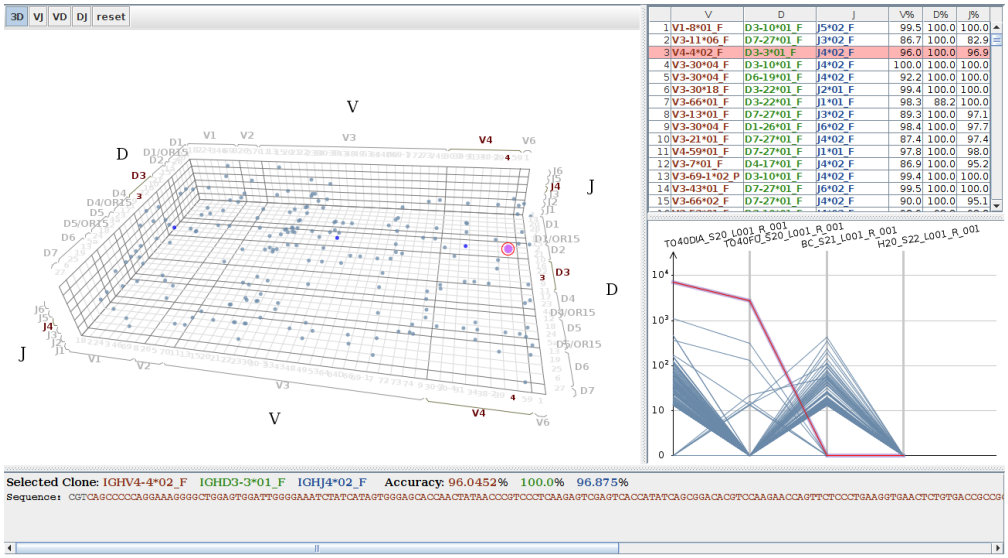
where $size_{\gamma_1}$ and $size_{\gamma_2}$ are the lengths of the two input signatures γ_1, γ_2, and M, MM and IN are the match, mismatch and insertion/deletion scores defined in the Smith-Waterman algorithm. Moreover, in our experiment we set M and MM score values equal to 2, and IN score value equals to 3. Observe that if the length of the smaller read is less than 70% of the length of the other then the reads γ_1, γ_2 are always considered different.

- Phase 4: *CompCheckerRead* takes as input the sets of signatures for each patient's sample (i.e. $\Gamma_1, \ldots, \Gamma_n$), and it derives the set of putative cancer clone Δ. Similar signatures among the samples are identified using the Smith-Waterman algorithm provided by SIMD Smith-Waterman C++ library. Then each identified putative tumour clone is analyzed to identify its best alignment with respect to V-GENE, J-GENE, and D-GENE. This task is performed thanks to a specifically developed aligner which uses a modified version of Smith-Waterman algorithm to find the best alignment of such clones with respect to the IMGT reference database. Since the experimental design (2, Experimental materials and Methods) can include the amplification of IGH, IGK,

IGL or TCR (alpha,beta,gamma,delta), CompCheckerRead uses for the alignment the appropriate reference database on the basis of the users choice.

- Phase 5: *VisualHashClone GUI*

  VisualHashClone graphic user interface allows the inspection of the tumour clones and their associated features in all the samples under study. The GUI is organised in four panels (Figure 2.9): the first panel contains 3D grid-plot orientated on three axes where each dot represents a clone. This panel allows an interactive analysis of all features of the clone: selecting a dot on the plot, the GUI highlights all the features (gene names, homology values, frequency and sequence) of the specific clone in all panels. The second panel contains a data-grid where the first column reports the clone identity in terms of V, J and D gene names with their associated homology values; The third panel contains a trend-plot that describes the clones frequencies and their trend in the samples; the third panel allows to visualise the DNA sequence for each clone. The sequence is characterised by three different colours to easily identify the V,J and D nucleotides. The fourth panel allows to visualise the DNA sequence for each clone. The sequence is characterised by three different colours to easily identify the V,J and D nucleotides. The user can easily manipulate and query the data presented in the data-grid. For instance all the clones can be ordered with respect to each column or set of columns, and they can be filtered according to their frequencies or the occurrence of a specific sub-sequence. The obtained graph can be also exported as a png file. Since physicians are interested for medical purpose to particularly follow the clone with the highest frequency at diagnosis (i.e. Major clone), another feature of VisualHashClone allows to highlight with a different colour (violet) the Major clone both in the first and in all the other panels. Thus, physicians are helped in the visualisation and tracking of the clone.



	V	D	J	V%	D%	J%
1	V1-8*01_F	D3-10*01_F	J5*02_F	99.5	100.0	100.0
2	V3-11*06_F	D7-27*01_F	J3*02_F	86.7	100.0	82.9
3	V4-4*02_F	D3-3*01_F	J4*02_F	96.0	100.0	96.9
4	V3-30*04_F	D3-10*01_F	J4*02_F	100.0	100.0	100.0
5	V3-30*04_F	D6-19*01_F	J4*02_F	92.2	100.0	100.0
6	V3-30*18_F	D3-22*01_F	J2*01_F	99.4	100.0	100.0
7	V3-66*01_F	D3-22*01_F	J1*01_F	98.3	88.2	100.0
8	V3-13*01_F	D7-27*01_F	J3*02_F	89.3	100.0	97.1
9	V3-30*04_F	D1-26*01_F	J6*02_F	98.4	100.0	97.7
10	V3-21*01_F	D7-27*01_F	J4*02_F	87.4	100.0	97.4
11	V4-59*01_F	D7-27*01_F	J1*01_F	97.8	100.0	98.0
12	V3-7*01_F	D4-17*01_F	J4*02_F	86.9	100.0	95.2
13	V3-69-1*02_P	D3-10*01_F	J4*02_F	99.4	100.0	100.0
14	V3-43*01_F	D7-27*01_F	J6*02_F	99.5	100.0	100.0
15	V3-66*02_F	D7-27*01_F	J4*02_F	90.0	100.0	95.1

FIGURE 2.9: **VisualHashClone graphic user interface.** The first panel on the top-left contains 3D grid-plot orientated on three axes where each dot represents a clone. The second panel on the top-right contain a data-grid where the first column reports the clone identity in terms of V, J and D gene names with their associated homology values. The third panel on bottom-right allows to visualise the trend-plot that describes the clones frequencies and their trend in the samples. The fourth panel on bottom allows to visualise the DNA sequence for each clone.

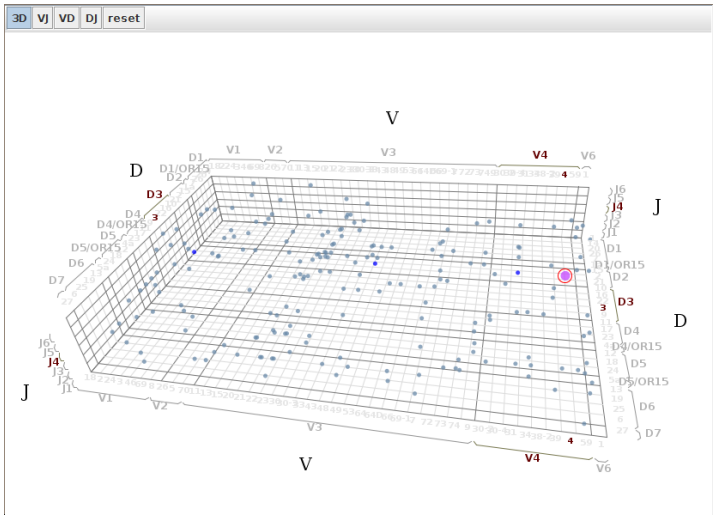


FIGURE 2.10: **HashClone graphic user interface.** First panel, 3D visualisation for each V, D and J segments. In violet the major clone is highlighted.

Figure 2.8 shows how the above described C++ applications are combined in a workflow to implement HashClone strategy for B-cells clonality assessment and MRD monitoring from collected samples of a single patient. Practically, *HashCheckerFreq* is executed on each patient's sample at a time to derive the k-mers and their associated frequencies. The collected set of k-mers generated by all the patient's samples are the input of *CompCheckerKmer*, which computes the set of *significant k-mers*. Then, *HashCheckerSignature* is run on each patient's sample to obtain the set of read signatures from the set of *significant k-mers*. Then, *CompCheckerRead* is executed to derive the putative clones from the read signatures obtained by all patient's samples. Finally, the VisualHashClone GUI is used for the visualisation of the results.

Parallel implementation of HashClone HashClone is composed by four C++ applications combined to implement B-cells clonality assessment in patient's samples. In Romano et al., 2018, we provided a parallel implementation of two out of five applications of HashClone suite. The parallelization of the first, *HashCheckerFreq*, and the third, *HashCheckerSignature* application allows to analyze more efficiently the samples from the same patient in parallel.

In the first phase of the pipeline, the application *HashCheckerFreq* is called on each patient's sample at time to derive the k-mers and their associated frequencies. Since each run is called separately on each patient's sample then this task is independent and can be performed in parallel (Figure 2.11). The parallelization of this phase allows us to derive the k-mers and their frequencies simultaneously from all the samples under analysis. Since the extrapolation of the k-mers proceed in parallel among all the samples, the next phase (implemented by *CompCheckerKmer*) can start only when all the samples are processed.

Then, the collected set of k-mers generated by all the patient's samples are the input of *CompCheckerKmer*, which computes the set of significant k-mers in the second phase.

Afterwards, in the third phase, *HashCheckerSignature* takes as input the set of significant k-mers and the set of the sample reads to obtain the set of read signatures from the set of significant k-mers. During the third phase, this application runs separately on each patient's sample and for this reason, as in the first phase, it was parallelized to speed up the analysis of the set of read signatures from the set of significant k-mers (Figure 2.11). Also in this case, the next phase (implemented by *CompCheckerRead*) can start only when all the samples are processed.

Finally, in the fourth phase *CompCheckerRead* is executed on the sets of signatures obtained for each patient's sample to derive the putative tumour clones and evaluate them using the Smith-Waterman algorithm provided by SIMD Smith-Waterman C++ library (Zhao et al., 2013).

Spike-in research module All analysis that use NGS technique platform for comparing changes between two or more conditions often require the use of methods that allow the researchers to consider the differences in DNA or RNA quantity in the samples under study. For this reason, lab procedures involve the use of spike-in control sequences which are known in sequences and quantity and used to calibrate measurements. A spike-in control has to be added in an amount proportional to the number of cells for subsequent normalization of the data, in order to allow accurate interpretations of whether there are increases or decreases in signal at each region of the genome between samples. Since IG and MRD monitoring lab procedures imply the use of spike-in sequences, *CompCheckerRead* application of HashClone

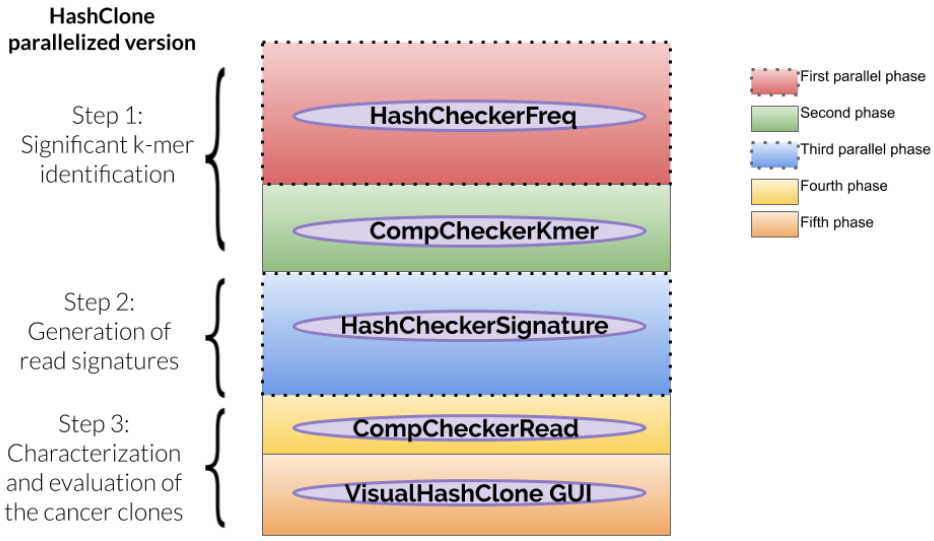


FIGURE 2.11: **HashClone parallelized version.** Phases with dots rectangles are parallelized (i.e. First phase: HashCheckerFreq; Third phase: HashCheckerSignature). Phases with lines rectangles are in series.

algorithm was improved with a spike-in research function. If the user employed spike-in sequences during the experimental protocol, he can easily upload as additional HashClone input, a file in FASTA format containing the list of the spike-in sequences that are supposed to be found in the samples (Figure 2.12).

Thus, *CompCheckerRead* application of HashClone algorithm exploits the Smith-Waterman algorithm (Zhao et al., 2013) to check the presence of the spike-in among the clones sequences. In detail, spike-in sequences similar among the samples are identified through a modified version of the Smith-Waterman algorithm to find the best alignment of such spike-in sequences with respect to the clone sequences. If a clone is identified as a spike-in sequence (i.e. the minimum alignemnt score is equal to 40 bases), *CompCheckerRead* memorises this information and reporting the name, the sequence and the frequency in each sample of the spike-in in the output file.

FIGURE 2.12: **Example of file in FASTA format.** Example of file in FASTA format containing three sequences. For each sequence, the header is identified by ">" symbol, the second line represents the sequence.

Strategy for clonality analysis, major B-cell clone detection, Minimal Residual Disease monitoring. HashClone output displays the entire list of the identified B-cell clones associated with the frequency value, the IGH rearrangement (in terms of VDJ genes and alleles), and homology identity values. Among all the reported B-cell clones, it is necessary to define the predominant clones that should be followed for MRD purpose (i.e. major clone). For this reason, we designed a *filtered strategy* composed of three phases.

In the *Filter-A* we selected a set of predominant clones based on the frequency values observed in the diagnostic samples. As reported by Faham and colleagues in Faham et al., 2012 any clonotype associated with low frequency ($< 5\%$) value was prudentially not considered representative of the disease.
In the *Filter-B* we considered the identity values associated with each B-cell clones: only the clones associated with more than 80% of homology in each IGHV, IGHD, and IGHJ genes are considered.
Finally, for the major clone identification (i.e. *Major clone selection* phase) we refer to criteria reported by Faham et al., 2012 and Ladetto et al., 2014. In detail, a clone can be declared as major in NGS sequencing analysis if at least one segment of V,D and J genes is above 80% of homology and if its composing number of reads represents at least $\pm 5\%$ of the total number of reads composing the entire patients clonotypes. Thus, we consider the clones that passed firstly *Filter-A* then *Filter-B* and we check the presence of a clone respecting the above criteria. If none of these clones is suitable, thus we check the clonotypes of passed only *Filter-A*. If no major clone is founded inspecting both the filters, thus we declare the patient as polyclonal (i.e. few clones with low frequency percentage and no major clone identified).
To make the MRD quantifications comparable between the two techniques, we set up a proportion between the total reads number of the major MCL clone at diagnosis (HashClone) and the ASOq-PCR value as shown in equation 2.6 where S_{MRD} is the MRD value of the sample (e.g. FD1/FA1) to calculate, S_{reads} is the number of reads derived from HashClone of that sample, DIA_{qPCR} is the qPCR value at Diagnosis of the patient and DIA_{reads} is the number of reads from HashClone of Diagnosis sample.

$$S_{MRD} = (S_{reads} * DIA_{ASO-qPCR}) / DIA_{reads} \tag{2.6}$$

Docker Container We included HashClone in the Docker container that is an open platform for developing, shipping, and running applications. Docker provides the ability to package and run an application in a loosely isolated environment called a container. A Docker container is a lightweight, stand-alone, executable package of a piece of software that includes everything needed to run a specific software: code, runtime, system tools, system libraries, settings. The container runs completely isolated from the host environment. The advantage of using Docker images is that the whole environment is fixed, the images are available in the Docker repository, and the identity of the images is the only element needed to reproduce the results.

The execution of the Docker images, implementing HashClone, is done by Docker4seq (Beccuti et al., 2017b), a R package which contains a collection of functions to execute NGS computing demanding applications and provides the connection between users and docker containers. The Docker is available at `https://www.docker.com/`. Then, HashClone can be run using the R enviroment, by installing previously *devtools* package and *Docker4seq*. HashClone command line run with Docker platform requires *group, data.folder, k-mer, hash_size, collision_list_size, threshold, type* and *input.files* parameters.

- *group*: setted as "docker";
- *data.folder*: is the folder where the output will be saved;
- *k-mer*: is the size of k-mers encoded in the hash table. This must be a value between 1 and 32;

- *hash_size*: is a (prime) number indicating the size of the hash table. Increasing this value reduces the execution time but increases the memory utilization. Ideally, this value should be close to the number of different k-mers stored in the hash table;
- *collision_list_size*: is the maximum number of different k-mers that the tool might need to store in the hash table. This parameter is required to optimize the memory utilization;
- *threshold* (tau): this value is the threshold used to select significant *k-mers*. We suggest to set tau equal to 1;
- *types* : a character string indicating which type of immunoglobulin (IGH, IGK, IGL) or T cell receptor chain (alpha,beta,gamma,delta) were amplified during the experimental procedure.
- *spike-in* : a character string indicating the path of the spike-in file in FASTA format otherwise set this parameter as 'null'
- *input.files*: the list of input patient files (FASTQ format).

2.7 Application of HashClone to MCL patients to determine Minimal Residual Disease and major clone detection

We applied HashClone in the context of Mantle Cell Lymphoma, with the aims to detect the major clone and monitor it over time during clinical follow ups. We performed an entire clonality analysis with NGS technology in the experimental procedure and HashClone algorithm for data analysis.

We performed three types of studies:

1. Study 1: Simulated datasets to test HashClone performances
2. Study 2: Clonality analysis of cohort 1 (i.e. 5 MCL patients)
3. Study 3: Clonality analysis of cohort 2 (i.e. 23 MCL patients)

In the first study, we used three simulated dataset to test the efficiency and the computational performance of HashClone. In the second and third studies, a total of 28 MCL patients were investigated for IGH detection and MRD monitoring using a new designed amplicon-based NGS approach and HashClone analysis. Both studies involved patients enrolled in Fondazione Italiana Linfomi in a phase III multicenter, randomized trial with maintenance using Lenalidomide drug, after induction regimen containing Rituximab, followed by high dose chemotherapy.

2.7.1 Study 1: Simulated dasasets to test HashClone performance

In the first study, we used three simulated datasets (Jackson et al., 2010) to test HashClone performance in execution time. The first, StandfordS_22 dataset, was taken from Jackson et al., 2010 paper and was used to test the performance of HashClone in terms of the IGHV, IGHD and IGHJ assignments. While, the second and third were synthetic in-house dataset created from the random concatenation of real patients samples and were used to compare the performance of the parallelized version of HashClone (Figure 2.11) versus the original one (Figure 2.8).

Performance on IGH alignment using StandfordS_22 dataset We tested the performance of the IGH alignment implemented in HashClone using the Stanford_S22 dataset. In the paper of Jackson et al., 2010 the authors evaluated the performance of seven algorithms handling the thousands of VDJ rearrangements in Stanford_S22 dataset to identify the IGHV, IGHD and IGHJ assignments and compare these back to the known genes from the inferred genotype for the subject. The thousands of VDJ rearrangements in this dataset allow the individual genotype at the IGH variable gene loci to be inferred.

We modified the Stanford_S22 dataset in order to test the performance of HashClone in terms of the IGHV, IGHD and IGHJ assignments. HashClone identifies 111 clones out of 500 having the percentage of alignment major than 90% for IGHV, IGHJ and IGHD. In the following are reported the data grouped by gene:

IGHV gene and allele correct: **99**
IGHV gene correct but different allelic variant: **12**
IGHV different gene not in the genotype: **0**

IGHJ gene and allele correct: **107**
IGHJ gene correct but a different allelic variant: **4**
IGHJ different gene not in the genotype: **0**

IGHD gene and allele correct: **105**
IGHD gene correct but a different allelic variant: **4**
IGHD different gene not in the genotype: **2**

Figure 2.13 shown the update version of Table 1 of the Jackson et al paper that reports the percentage of alignments from the Stanford_S22 dataset that were made, by various utilities, to IGHV, IGHD and IGHJ genes and allelic variants absent from the S22 genotype

Utility	IGHV (%)[a]	IGHD (%)[a]	IGHJ (%)[a]	Total (%)[b]
iHMMune-align	3.21 (0.21)	2.21 (1.27)	1.95 (0.0)	7.11
IMGT/VQ+JA	4.90 (0.22)	5.09 (2.81)	1.55 (0.0)	10.87
IgBLAST	3.84 (0.75)	3.96 (2.16)	0.85 (0.0)	8.39
Ab-origin	4.06 (0.22)	7.94 (5.53)	2.53 (0.0)	13.74
JOINSOLVER	6.17 (0.86)	6.93 (4.92)	1.24 (0.0)	7.89
SoDA	2.68 (0.29)	6.82 (6.63)	1.50 (0.0)	10.37
VDJSolver	6.87 (0.48)	1.96 (0.79)	0.71 (0.0)	9.09
HashClone	0 (10.8)	1.8 (5.4)	0 (3.6)	0

FIGURE 2.13: **Performance of the algorithms in IGH detection.** [a]Errors involving an incorrect gene, rather than an incorrect allelic variant, are shown in brackets. [b] Percentage of sequences that include an incorrect gene or allele for either the V, D or J.

.

The overall error for HashClone is equal to 1.8% that is the lowest value compared to the overall error percentages reported by Jackson, ranging between 7.1% (using iHMMune-align algorithm) and 13.7% (using Ab-origin algorithm).

Performance on in-house simulated datasets We created two datasets, Dataset 1 and 2, starting from the concatenation of diagnosis and follow up samples of four MCL patients[1]. Each MCL patient was composed of one diagnosis and one follow up sample. Diagnosis of Dataset1 was generated from the concatenation of the four MCL diagnostic samples; the follow up time point (FU1) derived from the concatenation of the four FUs samples. Dataset2 contained one diagnosis (DIA) and three follow ups time points (FU1,FU2,FU3). To improve the complexity of the samples and the total number of reads to analyze through the two pipelines, we created the Dataset2 concatenating the samples of Dataset1. The diagnosis of Dataset2 was composed through the concatenation of the diagnostic time point of Dataset1 two times (i.e. 21.746.096 reads $*2 = 43.492.192$). The follow ups (FU1,FU2,FU3) derived from the two- three- and four-time concatenation of follow up sample of Dataset1, respectively (e.g. 22.456.848 reads of Dataset1-FU1 $*4 = 89.827.392$). Table 2.2 reports the number of reads contained in each sample of Dataset1 and Dataset2 (Romano et al., 2018).

Execution time comparison: HashClone parallelized version vs HashClone original version

We performed the analysis with *HashClone parallelized* (Figure 2.11) and *HashClone* original version (Figure 2.8) of Dataset1 and Dataset2 and we retained the execution time in both versions. Since this analysis is performed only to analyze *HashClone parallelized*

[1]These patients were not part of the Study 2 and 3 but belong to previous experiment

TABLE 2.2: **Number of reads for Dataset1 and Dataset2**

	Number of reads	
	Dataset1	***Dataset2***
Diagnosis	21,746,096	43,492,192
FU1	22,456,848	44,913,696
FU2	/	67,370,544
FU3	/	89,827,392
Average	22,101,472	61,400,956

execution time and it is based on a random concatenation of the sequences of the MCL patients we did not consider and report clonality assessment.

For Dataset1, *HashClone parallelized* and HashClone performed a 2-samples analysis (one diagnosis and one follow up), instead for Dataset2 a 4-samples analysis (one diagnosis and three follow ups samples). *HashClone parallelized* and HashClone were executed on a server with 2 12-core AMD Opteron 6176 SE (48 Hyper-Threads) 2.3GHz, with 12 MiB L3 cache and 512 GiB of main memory, running Linux x86_64. For Dataset1 and Dataset2 two and four computational cores were used, respectively.

We investigated the time necessary to the first phase (Figure 2.8 and 2.11, red boxes), implemented by *HashCheckerFreq* to create the *hash table* for each sample, and the third phase (Figure 2.8 and 2.11, blue boxes), implemented by *HashCheckerSignature* to find the overlapping between the *significant k-mers* lists and the sample reads, in *HashClone parallelized* and HashClone. We reported the maximum time for the version parallelized and the sum of the samples times for the original version. Indeed, since *HashClone parallelized* first phase proceeds in parallel, its total time corresponds to the needed time to create the hash table among all the samples. Instead, for the first version it is given by the sum of the times of each sample (e.g. four samples = four times), because in this version the analysis is done one by one.

In Dataset1, the first phase was performed by HashClone parallelized in 10 minutes (37,5% time less) with respect to 16 minutes of HashClone. Moreover, the third phase requested 28 minutes for *HashClone parallelized* (46% time less) respect to 52 minutes of HashClone (Table 2.3).
In Dataset2,*HashClone parallelized* performed the first phase in 34 minutes (60% time less) and the third phase in 186 minutes (57% time less). Instead, HashClone required 86 minutes and 435 minutes for the first and third phase,respectively (Table 2.3).

TABLE 2.3: **Execution time of the first and third phase by *HashClone parallelized* and HashClone original**

	Execution Times			
	HashClone parallelized		**HashClone original**	
	Dataset1	***Dataset2***	***Dataset1***	***Dataset2***
First phase	**10 min**	**34 min**	**16 min**	**86 min**
Second phase	70min	107min	51min	105min
Third phase	**28 min**	**186 min**	**52 min**	**435 min**
Fourth phase	62min	727min	62min	732min

Table 2.4 reports the total execution times for both datasets of the complete *HashClone parallelized* and HashClone analysis. In Dataset1, *HashClone parallelized* was 5% more efficient than HashClone. This is probably influenced by the low numbers and the small differences between the input reads of the two samples (average per sample: 22,101,472 reads;*HashClone parallelized*: 2 hours 50 minutes ; HashClone: 3 hours). In Dataset2, we performed a 4-samples run and we increased the number of reads (average per sample: 61,400,956 reads). In this case *HashClone parallelized* was more effiecient in terms of execution time of 23% (*HashClone parallelized*: 17 hours; HashClone: 22 hours) (Romano et al., 2018).

TABLE 2.4: **Total execution time comparison: *HashClone parallelized* vs HashClone original**

	Total execution time	
	HashClone parallelized	***HashClone original***
Dataset1	2h 50min	3h
Dataset2	17h	22h

The maximum theoretical speedup (S) can be evaluated with Amdahl's law (Amdahl, 1967) that means:

$$S(s) = \frac{1}{(1-p) + \frac{p}{s}} \tag{2.7}$$

where s is the theoretical maximum earnings using 4 processors; p is the proportion of execution time of the part that can be parallelized.
Considering Dataset2, p is equal to 0.38 and s is equal to 4, then S results in 1.40. This is due to the second and fourth phases that are still sequential and their execution times prevail on the parallelized phases. Despite the maximum earnings is small in terms of scalability, it is still good for the absolute execution times of *HashClone parallelized* (Romano et al., 2018).

In the first dataset (Dataset1),*HashClone parallelized* performed slightly better (5% time less) than HashClone original version (2 hours 50 minutes *vs* 3 hours), probably due to the

small differences between the two samples. Indeed, in Dataset2 we increased the complexity of the data and *HashClone parallelized* was more efficient of 23% in terms of execution time since it performed the analysis in 17 hours respect to HashClone, which finished in 22 hours (Romano et al., 2018).

2.7.2 Study 2: Clonality analysis of cohort 1

Patients data collection Five MCL patients (PatA-E) were investigated for IGH detection and MRD monitoring using a new designed amplicon-based NGS approach. Two Pilot studies, namely *Pilot1* and *Pilot2* were performed, details about the sample are summarized in Figure 2.14. Patients were enrolled in MCL0208 protocol of FIL that consists of a phase III multicenter, open-label, randomized, controlled study to determine the efficacy and safety of Lenalidomide drug as maintenance therapy (i.e. Maintenance 1 and 2) in patients with MCL in complete or partial remission after first line intensified and high-dose chemotherapy additioned with Rituximab (i.e. Induction phase) and followed by ASCT (i.e. Consolidation therapy)(Figure 2.14). In *Pilot1* the five diagnostic samples and two (for PatD) and three (for PatA, B, C, and E) artificial dilution samples were analyzed. These samples were prepared diluting the diagnostic material in a pooled DNA derived from healthy subjects ("buffycoat"); the same buffycoat was included in the experiment, as polyclonal control. The 19 libraries were prepared using 500 ng of gDNA and sequenced using Illumina V2 kit chemistry 500 cycles PE on MiSeq platform as described in Supplementary Material (6.1). The average number of reads in each sample is equal to 481,289 (range: from 328,950 to 1,042,206 reads). The buffycoat sample contains 301,772 reads and the negative control (water) contains 466,348 reads. The quality check of the runs was performed using FastQC software (Patel and Jain, 2012) among the features considered the base quality (average value equals to 36) and the N content passed the check.

Pilot2 was composed of three diagnostic samples and three (PatA) or four (PatB and E) real FU samples. The first follow ups (FU1) were collected after three months, during the induction therapy (i.e. first treatment given for MCL); the second follow up (FU2) after four months from induction therapy when patients underwent to consolidation therapy (e.g. radiation therapy or stem cell transplant)(i.e. Pat A and B) or post Autologous Stem Cell Transplantation (i.e. ASCT, Pat E). The third followup (FU3) were collected Post ASCT for patient A and at Maintenance 1 (i.e. Pat B and E). Finally, last follow up (FU4) was collected after 6 months from maintenance 1 only for patients B and E. To test the efficiency of our wet lab procedures, 14 libraries were prepared reducing the gDNA input to 100 ng each. The average number of reads is equal to 316,789 (range: from 6,554 to 1,509,538 reads), while the buffycoat sample contains 478 reads and the negative sample (HELA cell line not carrying IGH rearrangements) contains 788 reads. As performed in *Pilot1*, we checked the quality of the data by FastQC software, but both base sequence quality (average value equals to 20) and N content features failed the check. For the whole experimental methodology refers to the Additional Figure 6.6, Supplementary Materials (6).
Five and three runs of HashClone were executed, one for each patient of *Pilot1* and *Pilot2*, respectively. Each run simultaneously analyzed the diagnostic sample and all artificial or clinical follow ups. In the next paragraphs, clonotypes identification, major B-cell clone selection and MRD monitoring will be reported for *Pilot1* and *Pilot2*.

Clonality The set of B-cell clones obtained by HashClone on both the *Pilot1* and *Pilot2* are processed following the *Filtering strategy* presented in section 2.6.1. In the diagnostic samples of the five patients of *Pilot1*, HashClone identified an average value of 1547 clonotypes (min 870, PatD; max 2149, PatC). The application of the *Filter-A* selected on average

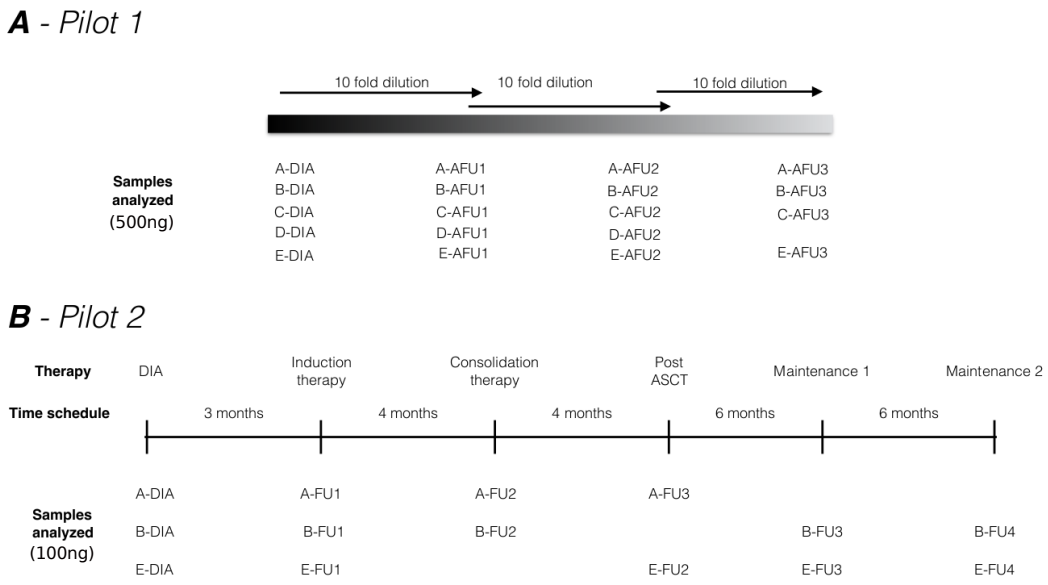


FIGURE 2.14: **Timeline of Study 2.A** Dilution gradient for samples of *Pilot1*. The five diagnostic samples (DIA) and two (for PatD) and three (for PatA, B, C, and E) artificial dilution samples (FU1-FU3) were analyzed. **B** The time line for patients of *Pilot2* where three diagnostic samples (DIA) and three (PatA) or four (PatB and E) real follow-up (FU1-FU4) samples were analyzed

.

38 clones. Indeed, the threshold of 5% reported by Faham et al., 2012 and collegues corresponds in our experiments to 100 reads. Thus only the clones associated with a frequency value major than 100 in the diagnostic sample were considered. From the *Filter-A*, on average 22 B-cell clones were retained in the analysis after the *Filter-B*. The average number of reads supporting these selected clonotypes is 100,929. The *Major clone selection phase* is not related to the two previous filters as explained in section 2.6.1. Indeed, for all the 5 patients a major clone is still selected and overcame the criteria of at least one gene segment homology >80% and the composing percentage of reads $> \pm 5\%$ of the total (Table 2.5).
In *Pilot2* HashClone identifies an average value of 96 clonotypes (min 77, PatE; max 278, PatB). The *Filter-A* filters out around 18% of the clonotypes: on average 18 clones were passed to the *Filter-B*. On average 6 clones passed the *Filter-B*, the average number of reads supporting the selected clonotype is 141,570 (Table 2.5). *Major clone selection phase* selected one clone for each patient (following the rule presented in section 2.6.1). Indeed, PatA presented one major clone with 97% of homology for V segment and frequency of 99.414 reads (92% out of 107.825 total reads), while PatB showed V and D homology of 93% and 80%, respectively and 169.175 reads (73% out of 231.281 total reads) . Finally, PatE showed 99% of V homology and 85.027 reads (99% out of 85.603 total reads) in the major clone (Additional Figure 6.2).

Major B-cell clone detection In *Pilot1* each of the five diagnostic samples clearly displayed one major clone with an average frequency of 93% (min 82%, PatB; max 98% PatA); while the other identified B-cell rearrangements showed an average frequency value equals to 7% (min 2% PatA; max 18% PatB), see Figure 2.15 and Additional Figure 6.2 in Supplementary Materials (6). In *Pilot2* the predominant clone is easily identified since its average frequency is 88% (min 73%, PatB; max 99% PatE) (Figure 2.16) while the other B-cell

TABLE 2.5: **Clonotypes identified with HashClone analysis and IMGT validation.** For each patient, the total number of identified clonotypes (third column) is reported and the average value across all patients. For each of these set, clonotypes with a frequency greater than 100 were selected and passed the Filter-A (fourth column). Then from the Filter-A, clonotypes with a VDJ homology greater than 80% were selected and passed the Filter-B (fifth column). Finally, the last column reports the number of major clone identified for each patient following the rules described in 2.6.1.

			Filter-A	**Filter-B**	**Major clone**
Study	**Patient (only diagnosis samples)**	**Clonotype identified**	**Clonotype with frequency $>$ 100**	**Clonotype with VDJ homology $>$ 80%**	**selection**
Pilot 1	**A**	1616	12	7	✓
	B	1703	59	33	✓
	C	2149	72	44	✓
	D	870	10	5	✓
	E	1398	35	21	✓
	Average value	**1547**	**38**	**22**	**1**
Pilot 2	**A**	96	18	6	✓
	B	278	72	11	✓
	E	77	5	0	✓
	Average value	**150**	**32**	**6**	**1**

clones showed an average frequency value of 12%. See Additional Figure 6.2 in Supplementary Materials (6) for more details. For each patient the predominant clone identified by HashClone was compared with the IGH monoclonal rearrangement identified by Sanger sequencing, in terms of IGHV, IGHD and IGHJ nucleotide homology, using BLASTn algorithm http://blast.ncbi.nlm.nih.gov. Four out of five diagnostic samples of *Pilot1* (PatA, C, D and E) showed exactly the same IGH rearrangement, in terms of IGH gene annotation and 100% nucleotide homology with respect to the Sanger sequence. Also Patient B showed the same rearrangement excepted for three nucleotide mismatches. On the other hand, a lower nucleotide homology (ranging from 44% to 66%) was noticed in *Pilot2*, due to the high number of unknown base calls (N) introduced by sequencing in the variable regions. Nevertheless, HashClone was still be able to assign the correct IGHV and IGHJ annotations, perfectly comparable with the Sanger results. These results are reported in Table 2.6.

TABLE 2.6: **HashClone and Sanger Sequence comparison.** This table reports the comparison in terms of IGHV, IGHD, and IGHJ nucleotide homology between the predominant clone identified by HashClone and the IGH monoclonal rearrangement identified by Sanger sequencing for each patient. Last column reports the homology between the two sequences as difference in nucleotide content and percentage. Red nucleotides in the sequences are those who differ between two sequences. N: unknown base calls.

Study	Patient	CDR3 Sanger Sequence	CDR3 HashClone Sequence	Homology
Pilot 1	A	GCGAGAGA**TCCA**GGGTATAGCAGTGGCTGGAA C**CTGGGA**TACTACTACTACGGTATGGACGTC TGTGCGAGAAGCAATTTTGGAGTGG**TCTAAAT** **TACAT**GGACGTCT	GCGAGAGA**TCCA**GGGTATAGCAGTGGCTGGAA C**CTGGGA**TACTACTACTACGGTATGGACGTC TGTG**T**G**C**GAA**T**CAATTTTGGAGTGG**TCTAAAT** **TACAT**GGACGTCT	100% (63/63 nt) 93% (42/45 nt)
	B	CGAGAGA**TTACACAGCCCC**GGGTATAGCAGAA CCAGGC**CCCT** TGCGAGAGG**CGCGA**ATAACTGGAAC**CCCA**TTG ACTA	CGAGAGA**TTACACAGCCCC**GGGTATAGCAGAA CCAGGC**CCCT** TGCGAGAGG**CGCGA**ATAACTGGAAC**CCCA**TTG ACTA	100% (42/42 nt) 100% (36/36 nt)
	C	GCGA**CCCAGCGAA**ATTACGATATTTTGACCGG **G**TTTGACTACT	GCGA**CCCAGCGAA**ATTACGATATTTTGACCGG **G**TTTGACTACT	100% (43/43 nt)
Pilot 2	A	GCGAGAGA**TCCA**GGGTATAGCAGTGGCTGGAA C**CTGGGA**TACTACTACTACGG TGTGCGAGAAGCAATTTTGGAGTGG**TCTAAAT** **TACAT**GGACGTCT	GCGAGANN**NNCA**NNNTATANCANNNGCTGGAA C**NNGGA**TACTACTACTACGG TGTGCGNNAA**TG**ANTTNNNNNGNNG**TCTAAAT** **TAAAT**NNNCNTCT	66% (39/59 nt) 64% (28/45 nt)
	B	GCGA**CCCAGCGAA**ATTACGATATTTTGACCGG **G**TTTGACTACT	GCGA**CNNTGNNNNN**TTNNNNNNTTTNGANCNN **N**NNTNA**A**NACT	44% (19/43 nt)
	E			

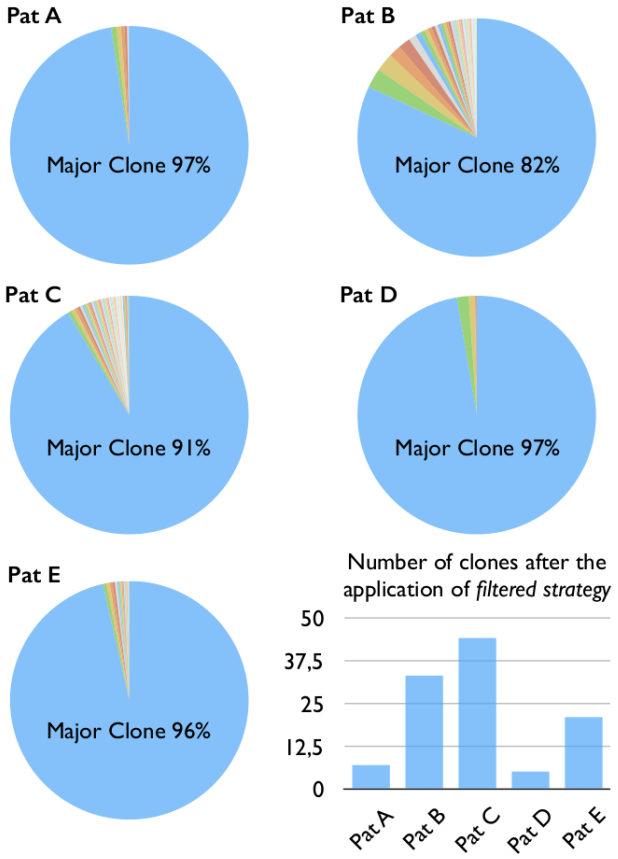


FIGURE 2.15: **Clonality analysis in MCL patients of *Pilot1*.** Pie plots showing the distribution of the frequency percentage associated with the B-cell clones passed the *filter strategy* in the five diagnostic samples of *Pilot1*. Into each pie plots it is reported the frequency percentages associated with the major clone. The histogram reports the number of B-cell clones passed the *filter strategy* in each patient.

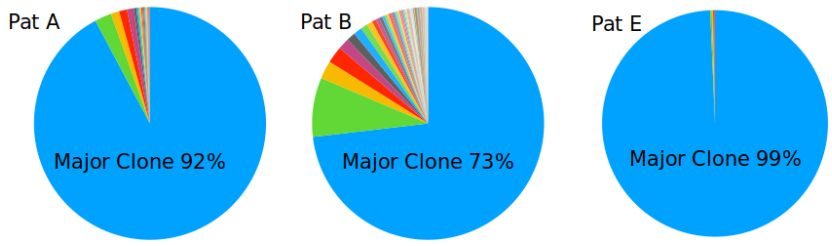


FIGURE 2.16: **Clonality analysis in MCL patients of *Pilot2*.** Pie plots show the distribution of the frequency percentage associated with the B-cell clones passed the *filter strategy* in the three diagnostic samples of *Pilot2*.

Minimal residual disease monitoring To monitor the MRD, HashClone tracks the clonotypes evolutions analyzing simultaneously the data from the diagnostic and the serial dilutions (*Pilot1*) or FU samples (*Pilot2*). Therefore, we compared the HashClone performance with the standardized results of the classical ASO q-PCR.

To make the MRD quantifications comparable between the two approaches, we set up

a proportion between the total reads number of the major MCL clone at diagnosis (HashClone) and the ASO q-PCR value. In details, patients A, C, D, and E had a high tumour infiltration (ASO q-PCR value of 1E+00 according to EuroMRD guidelines) (Velden et al., 2007); while patient B started from an ASO q-PCR value of $1E-01$, according to a lower tumour infiltration. These data are confirmed by a 2.5% $CD5^{+}/CD19^{+}$ MCL cells rate by flow cytometry.

HashClone was able to perfectly extract the MRD trend kinetics in the dilution/FU samples of the five MCL patients in both Pilot studies. Figure 2.17 reports the trends of PatB and Pat E (*Pilot1*) and PatA and PatE (*Pilot2*). Overall, the correlation analysis showed a high concordance between ASO q-PCR and the NGS technology (R^2=0.86), see Figure 4 Panel A. Indeed 30 out of 33 points are concordant: in *Pilot1* HashClone overestimates the frequency value in one case point; in *Pilot2* ASO q-PCR overestimates the frequency value in two cases.

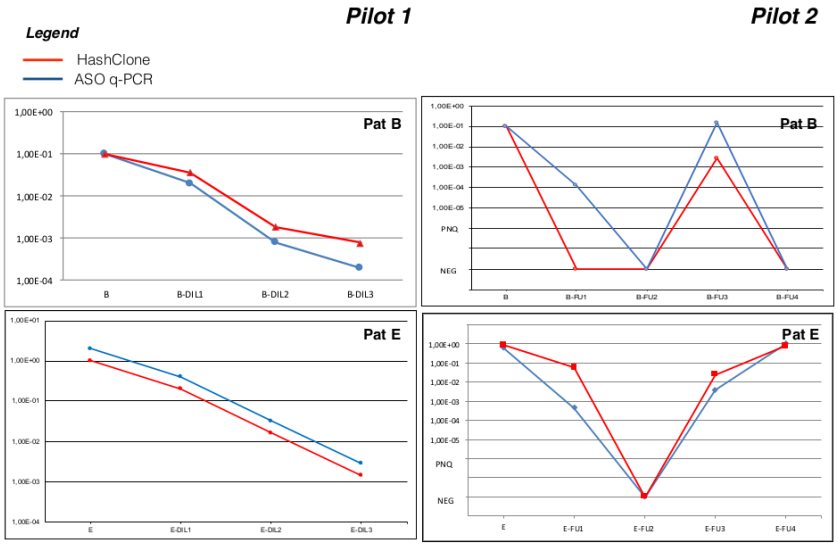


FIGURE 2.17: **MRD trend comparison.** MRD trend obtained from ASO q-PCR (blue line) and HashClone (red line) of Patient B and E of *Pilot1* and patient A and E of *Pilot2*.

Comparison of Hashclone accuracy with respect to the state-of-art algorithm: ViDJil We compared the accuracy of HashClone with respect to ViDJil algorithm. At the best of our knowledge, ViDJil is the only tool currently able to analyze the high-throughput sequencing data from lymphocytes, to extract IGHV, IGHD, and IGHJ junctions and to gather them into **clones** for quantification. ViDJil quantifies the clonotype abundances through a first ultra-fast prediction of putative rearrangements by a seed-based heuristic analysis and it outputs a window overlapping the CDR3 with the IMGT reference database. The putative clone sequence identified is further processed to obtain its full IGHV, IGHD, and IGHJ segmentation. Moreover, ViDJil can carry out the MRD analysis thanks to a web multi-sample application able to track selected clones in the diagnostic samples through different runs on different FU samples.

The strategy used to analyze the ViDJil results is composed of two phases: the *Phase-A* is the same implemented for HashClone, in the *Phase-B* since ViDJil associates the clones with the VDJ genes and alleles without reporting the homology values, we consider only the clones associated with one IGH rearrangement.

The set of B-cell clones obtained by ViDJil on both the *Pilot1* and *Pilot2* and those filtered by *Phase-A* and *Phase-B* are reported inAdditional Figure 6.3 in Supplementary Materials (6). More details about the number of reads associated with each clone are reported in Additional Figure 6.4 in Supplementary Materials (6). In *Pilot1* ViDJil is able to detect the major B-cell clone in all patients, the CDR3 regions detected in patients A, C, D and E have 100% homology with respect to the Sanger sequence, while patient B has an homology value equal to 93%, as reveled by HashClone. In *Pilot2* the elevated number of N base calls masking the CDR3 regions did not allow ViDJil to correctly annotate the IGHV, IGHD, and IGHJ in any patient, so that the nucleotide homology value dropped to 0 with respect to the Sanger sequence, see Additional Figure 6.5 in Supplementary Materials (6). In contrast, as described above, the HashClone performance was not hampered by the number of N base calls in the *Pilot2*.

We also compared the MRD quantification of all samples of both *Pilot1* and *Pilot2* between ViDJil and the ASO q-PCR data. Figure 2.18 reports the correlation analysis of all samples between HashClone and the ASO q-PCR data (Panel A) and between ViDJil and the ASO q-PCR data (Panel B). It is worthwhile to note that the concordance between HashClone and ASO q-PCR is higher than the concordance between ViDJil and ASO q-PCR, 86% versus 80% respectively.

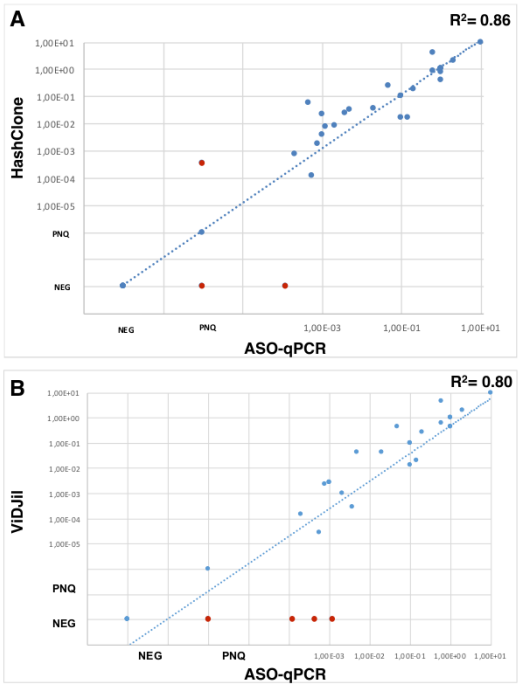


FIGURE 2.18: **Correlation analysis.** Scatter plot of the correlation analysis between HashClone and the ASO q-PCR data (Panel A) and between ViDJil and the ASO q-PCR data (Panel B). In Panel A, three discordances (red dots) are detected, one of them is quantifiable only by HashClone. While in Panel B there are four samples quantifiable only by ASO q-PCR. NEG, Negative; PNQ, Positive Not Quantifiable.

Discussion HashClone is an easy-to-use and reliable bioinformatics suite that provides B-cells clonality assessment and IGH-based MRD monitoring over time. To test its performances we analyzed two NGS experiments targeting the IGH rearrangements in samples obtained from 5 patients affected by MCL. The comparison was done on two MCL pilot studies generated using either 500 ng (Pilot1) or 100 ng (Pilot2) of gDNA as input in library preparation.

The two experimental protocols considered reflect different clinical/biological situations. *Pilot1* reproduces in the NGS setting the optimal requirements of a classical IGH screening experiment and a dilution curve. On the other hand *Pilot2* investigates the effects of a decrease in DNA quantity, mimicking a real-life situation that typically occurs in the routine of haematological laboratories. The restricted DNA availability can be due to the low cellularity of the biological samples (i.e low disease infiltration or material lack) or to specific sample conditions (i.e DNA extracted from formalin fixed paraffin embedded-FFPE- samples, or cell-free DNA from serum, plasma, or urine).

Our NGS experiments showed that, even though the mean number of reads obtained from the two studies was similar (481,298 *Pilot1* and 316,789 *Pilot2*), the base sequence quality was poorer in the *Pilot2*. This is reported by the base N content (FastQC check failed for the *Pilot2*) and the base sequence quality (mean value of 36 in *Pilot1* compared to a mean value of 20 in *Pilot2*). The limited quality of the *Pilot2* data is reflected on a lower number of detected clones that passed the filtering phases (i.e. Phase A, B and C, Table 2.5) and in a very low homology level of the CDR3 regions with respect to the Sanger sequence (average value of 99% in *Pilot1* with respect to an average value of 58% in *Pilot2*, p-value=0.02, computed by Student's t-test) (Table 2.6).

To assess the accuracy of HashClone to identify the major B-cell clone and to monitor the MRD we compared its performance with respect to the results obtained by state-of-art tool, ViDJil (Giraud et al., 2014). HashClone and ViDJil correctly identified the major clones in *Pilot1*. However, in *Pilot2* the elevate number of N base calls masked the IGHD region and reduced the nucleotide homology, leading to a decrease in the efficiency of ViDJil. In contrast, HashClone was able to identify the major clone in all the diagnostic samples. Moreover, in MRD monitoring the performance of HashClone outperformed the ViDJil results (concordance percentage: 86% HashClone, 80% ViDJil) in the comparison with respect to the gold-standard technique, ASO q-PCR data.

2.7.3 Study 3: Clonality analysis of cohort 2

Patients data collection In this third study, 23 MCL patients (i.e. cohort 2) were investigated for IGH detection and MRD monitoring using the amplicon-based NGS approach outlined in the Additional Figure 6.6 in Supplementary Materials (6). Patients of cohort 2 were enrolled in the same protocol of cohort 1 (i.e. MCL0208 protocol of FIL) but with a longer period of maintenance therapy. FD1/FA1 samples were taken during the induction therapy, after three months from diagnosis (DIA), FD2/FA2 samples after four months from induction therapy while FD3/FA3 samples post ASCT (i.e. autologous stem cell transplantation). Maintenance samples (i.e. from MD1/MA1 to MD6/MA6) were taken every 6 months in the observation period (Figure 2.19). For cohort 2, a total of 68 samples were available divided among diagnosis, follow up and maintenance samples (Table 2.7). The diagnosis accounts for 23 samples, FD1/FA1 accounts for 12 samples, FD2/FA2 for three samples, FD3/FA3 for 8 samples, MD1/MA1 and MD2/MA2 for four samples each, two samples for MD3/MA3, MD4/MA4 for four samples and finally one MA5 and MA6 samples. Patient 49 is the one with almost the timeline complete while 23,74,267, 205 and 264 patients displayed only the diagnosis samples.On average three samples for patient were available (Table 2.7).

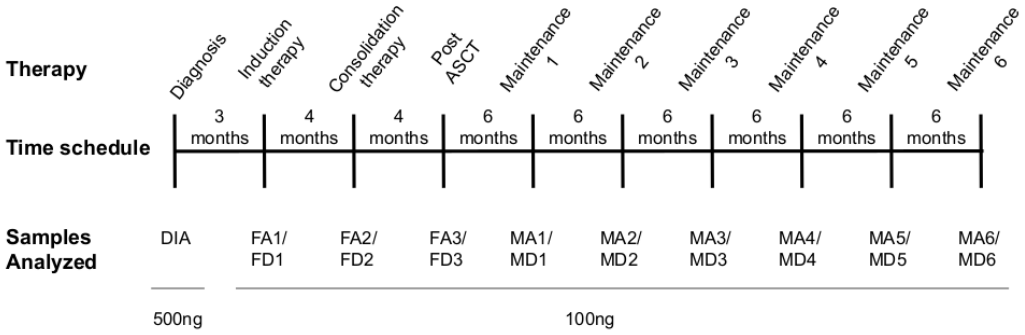


FIGURE 2.19: **Time line of cohort 2 composed of 23 MCL patients.**

Patients were divided on the basis of the type of markers showed at diagnosis in three categories (see Table 2.1). The first is composed of 17 patients with a positive marker for IGH and negative for BCL1, the second of four patients with IGH negative marker and positive for BCL1 while the third of two patients with both IGH and BCL1 markers negative.

- *Patients IGH+/BCL1-* (n=17 patients): 17 patients showed at diagnosis IGH positive marker and negative for BCL1. 13 patients out of 17 were analyzed with both NGS and ASOq-PCR followed by Sanger sequencing for clonality assessment and MRD monitoring (pat 1,10,19,22,74,90,140,155,175,182,237,251,267,284). For the remaining four patients, NGS, qualitative PCR and Sanger technique were performed, due to lack of materials or failure of primers during the ASOq-PCR experimental procedure (pat 23,255,262 and 267).

- *Patients IGH-/BCL1+* (n=4 patients): Patients were negative for IGH rearrangement and positive for BCL1 markers (IGH-/BCL1+). NGS sequencing and qualitative PCR data were reported since no ASOq-PCR and Sanger sequencing primers for BCL1 were available (i.e. pat 49,61,111,179).

- *Patients IGH-/BCL1-* (n=2 patients): Patients with both IGH and BCL1 markers negative (IGH-/BCL1-). Since any of the two markers can be detected, only NGS sequencing and qualitative PCR data were available (pat 205,264).

All the libraries were prepared using 500ng of gDNA and 100ng for FA/FD and MA/MD samples, from peripheral blood (D) or bone marrow (A) and sequenced as described in Supplementary 6.1 using primer for FR1 region from EuroClonality NGS (Brüggemann et al., 2019). The average number of reads at diagnosis is equal to 1.129.804 (range: from 414.350 to 2.610.814 reads), while follow up samples have an average number of reads equal to 809.872 (range: from 414.350 to 2.610.814 reads). The quality check of the runs was performed using FastQC software (Patel and Jain, 2012) among the features considered the base quality (average value equals to 36) and the N content passed the check. As for Study 1, buffycoat sample and negative control (water) were used (Buffycoat: 834.768 reads, water: 180.308).

TABLE 2.7: **Patients samples details of the Study 3.**

Patient ID	Tumour infiltration	Marker	Number of samples	Samples available
1	48.5	IGH+/BCL1-	3	DIA (D) FD1 MA3
10	79.3	IGH+/BCL1-	3	DIA (D) FD1 MA4
19	44.1	IGH+/BCL1-	3	DIA (D) FD1 MA1
22	32	IGH+/BCL1-	3	DIA (A) MA1 MA3
23	na	IGH+/BCL1-	2	DIA (A)
74	0.6	IGH+/BCL1-	1	DIA (A)
90	45	IGH+/BCL1-	2	DIA (A) MA2
140	na	IGH+/BCL1-	3	DIA (A) FA1 FA3
155	6.4	IGH+/BCL1-	3	DIA (D) FD1 FD3
175	82	IGH+/BCL1-	3	DIA (A) FA1 MD2
182	22	IGH+/BCL1-	3	DIA (A) FD2
237	80	IGH+/BCL1-	5	DIA (D) FA1 FA2 FD3 MA1
251	3.9	IGH+/BCL1-	3	DIA (D) FA1
255	30	IGH+/BCL1-	4	DIA (A) FA1 FA2 FA3
262	20	IGH+/BCL1-	3	DIA (A) FA1 FA3
267	4	IGH+/BCL1-	2	DIA (A)
284	9	IGH+/BCL1-	2	DIA (D) FA1
49	4	IGH-/BCL1+	7	DIA (A) FA3 MA1 MA2 MA4 MA5 MA6
61	1	IGH-/BCL1+	2	DIA (D) FD3
111	1.7	IGH-/BCL1+	3	DIA (A) MA4
179	1.8	IGH-/BCL1+	5	DIA (D) FD1 FA3 MA2 MD4
205	10.9	IGH-/BCL1-	1	DIA (D)
264	0.1	IGH-/BCL1-	2	DIA (A)

Clonality analysis We performed a clonality analysis of the three patients categories with HashClone performing 23 total runs, including diagnosis sample, available follow up (i.e. FA/FD and MA/MD) samples and buffycoat/negative control for each of the patient. Then, the filtering strategy explained in section 2.6.1 was applied using two filters (*Filter A* and *Filter B*) for the selection of clonotypes and the *Major Clone selection* for major B-cell clones identification. For the first category (n=17 patients IGH+/BCL1-), HashClone identified in the diagnostic samples an average value of 157 clonotypes (min 17, Pat 74; max 454, Pat 155). The application of the *Filter-A* selected on average 60 clones of which on average 10 B-cell clones were retained in the analysis after the *Filter-B*. One major clone was identified for 12 out of 17 patients (pat 1,10,19,22,90,140,155,175, 237,255,262,284) while the clones of the remaining five patients did not pass the limits imposed by the *Major Clone selection* phase of section 2.6.1. The second category was composed of four patients positive for BCL1 marker and not for IGH. HashClone identified an average value of 296 clonotypes (min 91, Pat 61; max 446, Pat 49). The *Filter-A* filtered around 34 clonotypes and on average 6 clones passed the *Phase-B*. In the *Major clone selection* phase, three patients displayed a major clone that passed the limits (pat 49,61,179), while clones of patient 111 did not. In the last category composed of patients negative for both markers (IGH-/BCL1-), HashClone identified an average value of 136 clonotypes of which 77 clonotypes passed the *Filter-A* and four clones from *Filter-B*. Only patient 205 displayed a major clone while patient 264 did not. Details about the results in both the Pilot studies are reported in Table 2.8.

Major B-cell clone detection The major B-cell clones of all patients were identified using the filtering strategy presented in 2.6.1 and reported their presence in Table 2.5. For 12 out of 17 patients belonging to IGH+/BCL1- category HashClone identified one major clone in the diagnostic sample. Patients 1 and 237 displayed one major clone with a frequency of 100% while patients 10,19,90,140 near 98% (Figure 2.22). The homology percentage between HashClone and Sanger sequencing are reported in Table 2.9. In the remaining patients 22,155,175,255,262,284 the frequency of the major clone is included between a minimum value of 77.2% and max 96.1% (Figure 2.22 and Table 2.9). However, for patient 22 the major clone identified with NGS (2.22) was not the same of that identified with Sanger sequencing. The clone identified by Sanger technique belongs to the repertoire identified by HashClone (Figure 2.22 and Table 2.9) and represents the second biggest clone. The supporting number of reads for major clone identified by HashClone is 13.174 (81.8%) while for sanger clone is 1.353 (8.4%). The presence of this minor clone identified by Sanger can be explained as an artificial clone created from the merge of the forward and reverse primers whose pairing dephase the sequencing lecture of the real major clone sequence or subject to worker-behaviour. However, the bigger sensibility and the use of universal primer of NGS technique (10^{-4}-10^{-5}) allowed to detect both the major and artificial clone. Patient 262 presents one major clone with 96.1% of frequency and several other minor clones (Figure 2.22). This patient represents one case of discordance among NGS and PCR technique since the major clone of NGS is not the same clone identified by Sanger sequencing (Table 2.9). However, this major clone identified by NGS technique display a frequency of 30% on the total reads (i.e. 1.159.703 reads), thus it overcame the limit imposed by Faham et al., 2012 of 5% out of the total number of reads. Moreover, we didn't find the clone derived from Sanger technique in HashClone results and this is probably due to an artefact of Sanger technique that was amplified during PCR in fewer copies. For five patients out of 17 (pat 23,74,182,251,267), HashClone did not find a major clone since from the repertoire of Table 2.5 no clones passed the $\pm$5% limit imposed by Faham et al., 2012 in the *Major Clone selection* phase, identifying these patients as polyclonal (i.e. few clones with low frequency percentage and no major clone identified). In detail, for patient 23 standard technique was

TABLE 2.8: **Clonotypes identified with HashClone analysis and IMGT validation.** For each patient, the total number of identified clonotypes (third column) is reported and the average value across all patients. For each of these set, clonotypes with a frequency greater than 100 were selected and passed the Filter-A (fourth column). Then from the Filter-A, clonotypes with a VDJ homology greater than 80% were selected and passed the Filter-B (fifth column). Last column represents the number of major clones found in the Major Clone selection phase.

			Filter-A	**Filter-B**	**Major Clone**
Markers	**Patient (only diagnosis samples)**	**Clonotype identified**	**Clonotype with frequency $>$ 100**	**Clonotype with VDJ homology $>$ 80%**	**selection**
IGH+/BCL1-	**1**	264	112	1	✓
	10	153	39	4	✓
	19	101	53	4	✓
	22	113	54	4	✓
	23	161	91	2	x
	74	17	14	1	x
	90	141	61	4	✓
	140	215	46	3	✓
	155	454	204	95	✓
	175	110	21	4	✓
	182	49	7	1	x
	237	164	40	1	✓
	251	111	52	3	x
	255	191	21	3	✓
	262	193	102	28	✓
	267	115	58	3	x
	284	109	46	4	✓
	Average value	**157**	**60**	**10**	**1**
IGH-/BCL1+	**49**	446	23	4	✓
	61	91	20	2	✓
	111	271	50	16	x
	179	377	43	1	✓
	Average value	**296**	**34**	**6**	**1**
IGH-/BCL1-	**205**	168	87	3	✓
	264	103	66	4	x
	Average value	**136**	**77**	**4**	**1**

concordant with NGS since no result about a putative major clone was retrieved. However, for patient 74,182,251 and 267, standard technique identified four putative major clones (i.e. one for each patient). Thanks to the high sensitivity of NGS technique, the clones identified by the standard technique were actually revealed as artificial derived probably from forward and reverse primers self-annealing problems. The second category was composed of four patients BLC1 marker positive and IGH negative (IGH-/BCL1+) (i.e. pat 49, 61, 111, 179). Major clones of patients 49,61 and 179 were identified by HashClone (Figure 2.20). The three major clones overcame the *Major clone selection phase* and displayed a frequency of 98.3%, 99.7% and 100% respectively. Only qualitative PCR was performed for these patients due to the lack of IGH markers. However, the highest NGS sensibility allowed to find major clones whose presence improve cases with single markers, changing the initial classification (IGH-/BCL1+) to double markers (IGH+/BCL1+) patients (Figure 2.20). Patient 111, instead, showed 16 minor clones and no major one with highest frequency. HashClone result was consistent with standard technique which also gave a negative outcome, concluding that patient 111 is polyclonal (Figure 2.20). Last category was composed of two patients both IGH-/BCL1- (Figure 2.21). Patient 205 displayed at diagnosis one major clone with 97.1% of frequency and two other clones with frequency less than 2% (Figure 2.21). However,

standard technique and Sanger sequencing did not amplified any major clone and also qualitative PCR was negative. Thus, this is another case of discordance of the two techniques, where highest NGS sensibility allowed to find a putative major clone that helped to find a real marker associated with these patients, where the qualitative PCR did not succeed, changing the initial classification of the patient from no marker to IGH positive. Finally, patient 264 is composed of four clones with frequencies of 55.5%, 19.7%, 16.8% and 8.1%, respectively. HashClone did not find a predominant major clone and also standard technique gave negative results. Thus, we can conclude that patient 264 had a polyclonal background at the diagnosis sample and probably the disease is carried out by another marker rather then IGH or BCL1. (Figure 2.21).

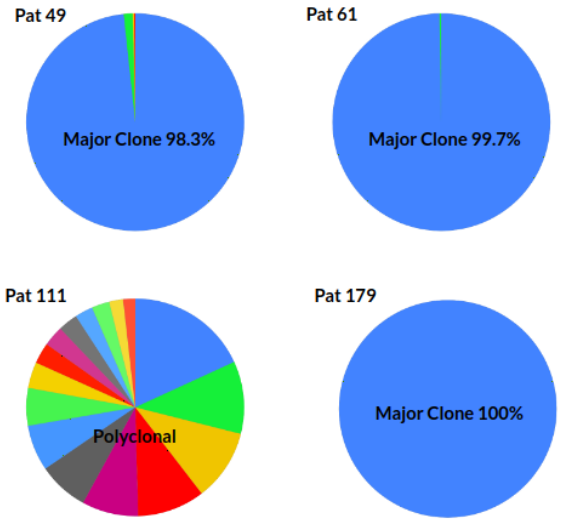


FIGURE 2.20: **Major B-cell clone detection in IGH-/BCL1+ patients.** Pie plots show the distribution of the frequency percentage associated with the B-cell clones passed the *filter strategy* in the four diagnostic samples of IGH-/BCL1+ patients. Note: patient 111 is polyclonal.

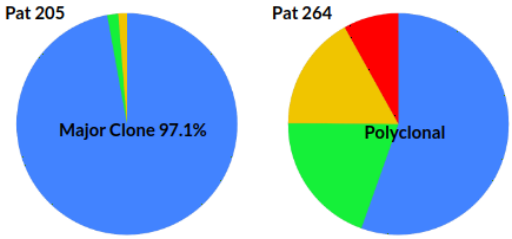


FIGURE 2.21: **Major B-cell clone detection in IGH-/BCL1- patients.**Pie plots show the distribution of the frequency percentage associated with the B-cell clones passed the *filter strategy* in the two diagnostic samples of IGH-/BCL1- patients. Note: patient 264 is polyclonal.

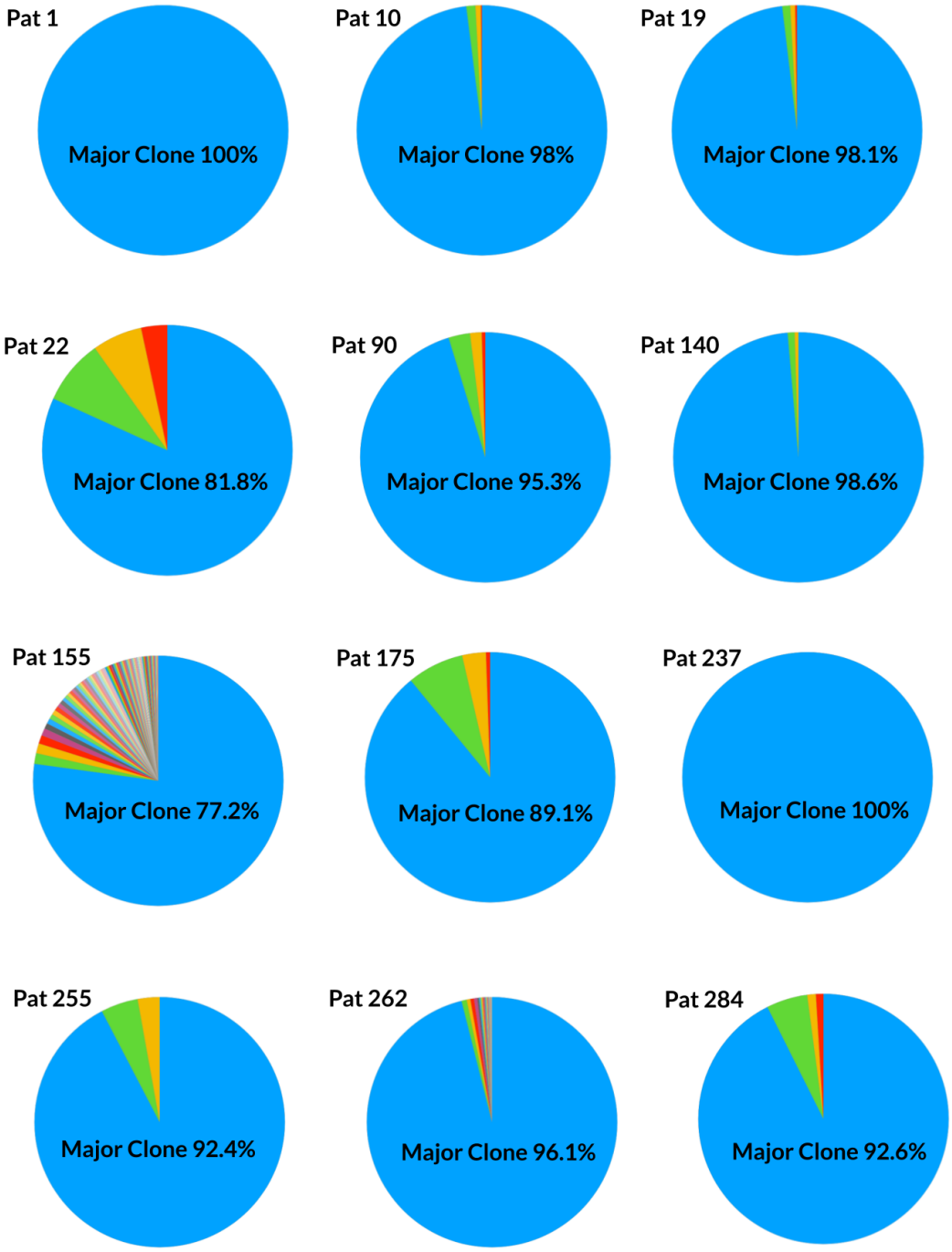


FIGURE 2.22: **Major B-cell clone detection in IGH+/BCL1- patients** Pie plots show the distribution of the frequency percentage associated with the B-cell clones passed the *filter strategy* in the 12 diagnostic samples of IGH+/BCL1- patients.

TABLE 2.9: **HashClone and Sanger Sequence comparison for cohort 2.** This table reports the comparison in terms of IGHV, IGHD, and IGHJ nucleotide homology between the predominant clone identified by HashClone and the IGH monoclonal rearrangement identified by Sanger sequencing for each patient.

Marker	Patient	CDR3 Sanger Sequence	CDR3 HashClone Sequence	Homology
IGH+/BCL1-	1	TGTGCGAGCTCGCCCAACTA CTACTACGGTATGGACGTCTGG	TGTGCGAGCTCGCCCAACTA CTACTACGGTATGGACGTCTGG	100% (42/42)
	10	TGTGCTAGAATCTTTGCTTCG GGGAGCGCTTTTTGACTTCTGG	TGTGCTAGAATCTTTGCTTCG GGGAGCGCTTTTTGACTTCTGG	100% (42/42)
	19	TGTGCGAGAGATGTCGATGAGTACAATATTTTGACT GGTTATTACATGTACTACTTTGACTACTGG	TGTGCGAGAGATGTCGATGAGTACAATATTTTGACT GGTTATTACATGTACTACTTTGACTACTGG	100% (66/66)
	22	TGTGCGAGTGCGGGGATTTTTGGAGTGGTT ATCACCGACAGCTACTACTACGGTATGGACGTCTGG	TGTACCACAGGGATGGTTCGGGGAGTTATT ATACTACTACTACGGTATGGACGTCTGG	0% (0/59 nt)
	90	TGTGCGAGACCGGGAGTAGGTGATGGCTAC AATTTTTGGGACCGTGGGCCTACCGAATGG GCGTACTACGGTATGGACGTCTGG	TGTGCGAGACCGGGAGTAGGTGATGGCTAC AATTTTTGGGACCGTGGGCCTACCGAATGG GCGTACTACGGTATGGACGTCTGG	100% (84/84 nt)
	140	TGTGCGTATGCCGGAATAAGTAG TGGTAGAACAAATGACTACTGG	TGTGCGTATGCCGGAATAAGTAG TGGTAGAACAAATGACTACTGG	100% (45/45 nt)
	155	TGTGCGAGACAGATGGTTCGG CGGGGGAGCTACGACTACTGG	TGTGCGAGACAGATGGTTCGG CGGGGGAGCTACGACTACTGG	100% (42/42 nt)
	175	TGTGCGAGAGATATCTGGAGCGGTGGCTACG ATTACTACTACTACGGTATGGACGTCTGG	TGTGCGAGAGATATCTGGAGCGGTGGCTACG ATTACTACTACTACGGTATGGACGTCTGG	100% (63/63 nt)
	237	GTGCGAGAGCTCTCAACTCAG GGCTGCCCTACTTTGACTACTGG	GTGCGAGAGCTCTCAACTCAG GGCTGCCCTACTTTGACTACTGG	100% (44/44 nt)
	255	TGTGCGAGAGGCCTTTTTGGAGTGGTTA TATCCTACTACTACTTCGGTATGGACGTCTGG	TGTGCGAGAGGCCTTTTTGGAGTGGTTA TATCCTACTACTACTTCGGTATGGACGTCTGG	100% (60/60)
	284	TGTGCGAGACAGACAGCAGT GGCTGGAGCTGACTACTGG	TGTGCGAGACAGACAGCAGT GGCTGGAGCTGACTACTGG	100% (39/39)
	262	TGTGCGCGAGCACAGGAATACGCTGATACT GGTGGAGGATGGTTCGACCCCTGG	TGTGCGAGGGGGCTCCGTAGGAGATTGTAGTGG TGGCAGCTGCTACTCAGGCCTTGTCTGGTTCGACCCCTGG	0% (0/73 nt)

MRD monitoring HashClone performances were compared to the standardized results of the classical ASOq-PCR in terms of MRD monitoring. As reported in the introduction of the cohort 2, only for 13 patients (belonging to the first category IGH+/BCL1-) out of 23, ASOq-PCR was perfomed and we were able to retrieve MRD values at Diagnosis. Unfortunately, for the last 13 patients, one MRD value at one follow up of three patients was not available, reducing the final set to 10 patients. While for IGH-/BCL1+ and IGH-/BCL1- negative patients ASOq-PCR values were not available. To make the MRD quantifications comparable between the two techniques, we set up a proportion between the total reads number of the major MCL clone at diagnosis (HashClone) and the ASOq-PCR value as shown in section 2.6.1.

The 10 patients (pat 1,10,19,22,90,140,155,175,237,284) had various tumour infiltration ranging from a minimum of 6.4 (pat 155) to 82 (pat 175). The ASO q-PCR value of patients 1,10,155,175,237 are 1.00E+00 according to EuroMRD guidelines (Velden et al., 2007), while patients 19,22,90,140,284 started from lower ASO q-PCR value with a minimum of 2.22E-03 (pat 19) to 0.6 (pat 90). Figure 2.23 reports the trends of the patients in the samples timeline available. For each patient of this category, 7 samples composed of one diagnosis, one follow up and one maintenance sample (i.e. pat 1,10,19,175), two follow up samples (i.e. pat 140,155) or two maintenance sample (i.e. pat 22). Then, the samples timeline of two patients was composed of one diagnosis and one follow up sample (i.e. pat 90 and 284). Finally, only patient 237 was composed of a timeline with 5 samples with complete follow up series and one maintenance. In Figure 2.23 are reported the MRD trend of these patients detected by HashClone and standard technique (HashClone: red; ASOq-PCR: blue). HashClone (Figure 2.23 ; red) was able to perfectly extract the MRD trend kinetics in the follow up and maintenance samples of this category as well as ASOq-PCR (Figure 2.23 ; blue). In Figure 2.24 is present correlation analysis between the two techniques using Pearson correlation, showing that there is a good concordance of all the 36 points (NGS and ASOq-PCR; Pearson correlation = 1).

Discussion The purpose of Study 3 was to analyze a cohort of 23 patients affected by MCL, enrolled in Fondazione Italiana Linfomi and following a protocol therapy of four years from diagnosis. Patients were divided in three categories on the basis of the type of markers showed at diagnosis: IGH+/BCL1- (first category), IGH-/BCL1+ (second category), IGH-/BCL1- (third category). We performed clonality analysis to investigate the number of valid clonotypes (i.e. according to the filtering strategy of section 2.6.1), identification of the major clone characterising each patient (i.e. clone with the highest frequency at diagnosis and followed by phisicians for clinical purpose) and MRD monitoring to track trend kinetics of the disease during the samples timeline (i.e. on average three samples for patient were available)(Table 2.10). The first category contained 17 patients with 165 different clonotypes. 12 out of 17 patients displayed one major clone (i.e. one for patient) and of these, 10 showed 100% of homology in terms of VDJ gene segment with Sanger sequencing clone. While, two patients displayed a major clone different from Sanger sequencing due to experimental technical reasons. Finally, MRD monitoring was performed showing that the disease trend kinetics was concordant in these 10 patients between standard technique and HashClone tool (Table 2.10). For the second category (IGH-/BCL1), four patients were analyzed retrieving 23 total different clonotypes. Three out of four displayed one major clone passing the criteria of *Major clone identification* phase. However, Sanger sequencing and ASO-qPCR values were not available for these patients due to different marker positivity (BCL1+), thus the comparison can not be reported (Table 2.10). Last category (IGH-/BCL1-) was composed of two patients affected by MCL but with both negative markers. We identified 7 different clonotypes among these patients and one out of two displayed a valid major clone while the other

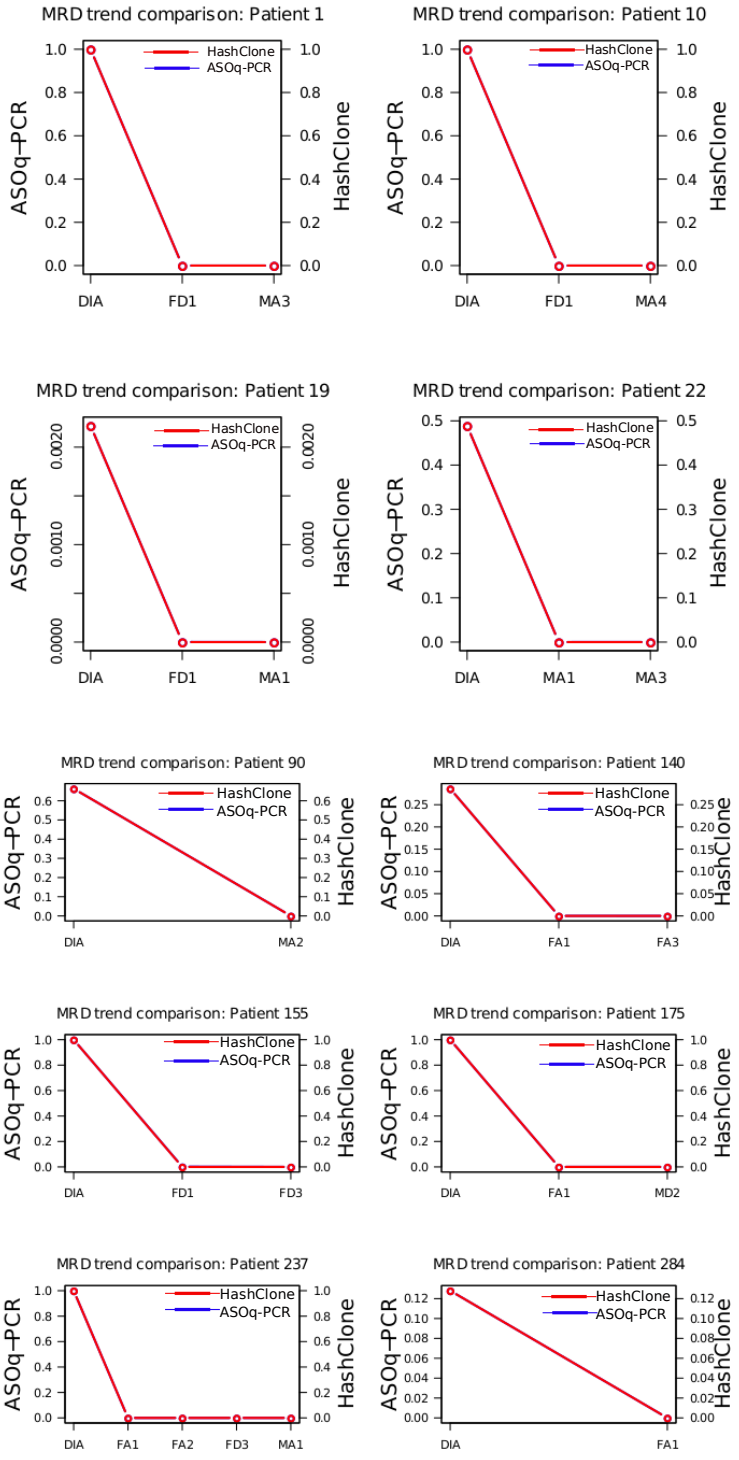


FIGURE 2.23: **MRD monitoring for IGH+/BCL1- patients.** MRD trend obtained from ASO q-PCR (blue line) and HashClone (red line) of IGH+/BCL1- Patient.

one was confirmed as polyclonal. Also in this case, Sanger sequencing and ASO-qPCR values were not available and the comparison was not reported (Table 2.10). Overall, HashClone was able to perform the entire clonality analysis and MRD monitoring in all patients. Moreover for 6 patients, the results of HashClone overcame those of gold-standard technique due

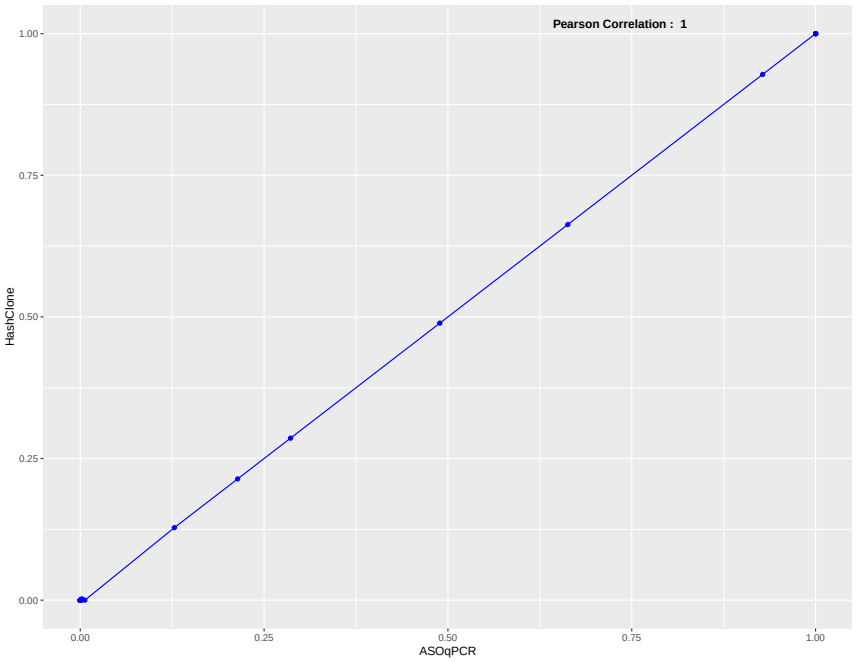


FIGURE 2.24: **Correlation analysis of HashClone and ASOq-PCR technique.** Scatter plot of the correlation analysis between HashClone and the ASO q-PCR data. No discordances (red dots) are detected between the two methods.

to an improved technical sensibility allowing to distinguish real major clones (i.e. patients 22 and 205) or polyclonal background in the tree categories (i.e. patients 23,74,182,251,267 IGH+/BCL1-, patient 111 IGH-/BCL1+ and patient 264 IGH-/BCL1-) with respect to putative technical artefacts.

TABLE 2.10: **Summary of results obtained from cohort 2.** (*according to the filtering strategy of section 2.6.1)

Category	Number of patients	Number of valid* clonotypes	Number of Major Clones	Comparison with Sanger sequencing	MRD monitoring
IGH+/BCL1+	17	165	12	10/12	10
IGH-/BCL1+	4	23	3	na	na
IGH-/BCL1-	2	7	1	na	na
Total	23	195	16	10/12	10

Chapter 3

Development of a new computational approach to identify cluster-specific RNA signatures from single cell RNA sequencing data

3.1 Background

Cells are the basic building blocks of organisms and each cell is unique. Thus, the idea of heterogenous cellular systems has been well-documented, both theoretically (Elsasser, 1984) and experimentally (Altschuler and Wu, 2010). Heterogeneity arises at many different levels and for a number of different reasons, in order to improve survival and functionality.

The use of scRNA-sequencing to define the cellular heterogeneity is key, as transcriptomics captures fine details that other methods may have missed. scRNA-sequencing technologies are now increasingly focused on the characterization of individual cells to uncover new and potentially unexpected biological discoveries with respect to traditional profiling bulk populations methods. Indeed, scRNA-seq methods can reveal complex and rare cell populations, uncover regulatory relationships between genes, and track the trajectories of distinct cell lineages in development (Hwang, Lee, and Bang, 2018).

3.2 Computational workflow and challenges to analyze scRNA-seq data

Single-cell RNA sequencing is essential for investigating cellular heterogeneity and highlighting cell subpopulation-specific signatures. Experimental methods for scRNA-seq are now increasingly accessible to many laboratories and computational pipelines for handling raw data files are increasing day by day.

scRNA-seq data analysis requires several steps to go from count generation to the cell subpopulation identification and they account for (Figure 3.1):

1. Phase 1: pre-processing (i.e. quality control of the reads, normalization, identification of confounding factors)
2. Phase 2: clustering methods (e.g. k-means, hierarchical clustering, density-based clustering)
3. Phase 3: bio-identification of the clusters (e.g. enrichment analysis)

In the Phase 1 of pre-processing, once reads are obtained from well-designed scRNA-seq experiments, quality control (QC) is performed for inspecting quality distributions across entire reads. Low-quality bases (usually at the 3' end) and adapter sequences can be removed

at this pre-processing step. Read alignment is the next step of scRNA-seq analysis, and the tools available for this procedure include the Burrows-Wheeler Aligner BWA (Li and Durbin, 2009) and STAR (Dobin et al., 2013) ,as those used in the bulk RNA-seq analysis pipeline (Hwang, Lee, and Bang, 2018). After alignment, reads are allocated to exonic, intronic, or intergenic features using transcript annotation and only reads that map to exonic loci with high mapping quality are considered for generation of the gene expression matrix. The matrix is often composed of columns representing cells and rows representing genes. In this context, a distinctive feature of scRNA-seq data is the presence of zero counts due to reasons such as dropout (i.e. a transcript is expressed in a cell but is entirely undetected in its mRNA profile) or transient gene expression. To account for this feature, data normalization is usually suggested to remove cell-specific bias, which can affect the determination of differential gene expression (Hwang, Lee, and Bang, 2018). After normalization, the next step is to estimate confounding factors that can be due to technical or biological variations. Technical variations include batch effect (i.e. systematic differences that are unrelated to any biological variation and result from sample preparation conditions), cell-specific capture efficiency, transcripts amplification bias and dropout. Biological variation instead include stochastic gene expression and cell cycle effect. The correction of confounding factors as amplification can be performed using for example spike-in sequences to calibrate the experiment or batch effects can be avoid repeating the analysis of multiple cells from a condition. More in general, statistical models that incorporate random noise can be used to manage known and unknown variables (Hwang, Lee, and Bang, 2018). Behind the confounding factors problem, another challenge regards the optimal statistical approach that permit to simultaneously compare the expression of thousands of genes of the dataset. In general, the total number of genes measured in a dataset is referred to as the dimensionality (Andrews and Hemberg, 2018). When comparing cells in a high dimensional gene expression space, distances between cells become more homogenous, making it difficult to distinguish differences between populations from variability within a population. There are two main approaches, firstly, data can be projected into a lower dimensional space (i.e. dimensionality reduction) in order to optimally preserve some characteristics of the original data. Some approaches comprise Principal Component Analysis (PCA) and T-distributed stochastic neighbor embedding (tSNE). Secondly, it is also possible to remove uninformative genes (i.e. feature selection) to reduce the number of dimensions used in the analysis to facilitates visualization and reduce noise and speed up calculations (Andrews and Hemberg, 2018).

The Phase 2, characterization of the numerous cells in the dataset, is a really delicate task. The problem has been widely studied in the machine learning literature, and there are several well-established strategies that have been adapted for scRNASeq data.

- *K-means* is a commonly used clustering algorithm for single-cell analysis and it consists in the iteratively assignation of cells to the nearest cluster centre (or centroid), and then recomputes the cluster centroids. However, k- means requires the number of clusters to be predetermined and uses stochastic starting locations for each cluster, thus requiring it to be run multiple times to check robustness to these parameters.

- *Hierarchical clustering* is another popular general-purpose clustering method commonly used to identify cell-populations. The most common assumption assume round equally-sized clusters like k-means and it has the advantage of being able to determine relationships between clusters of different granularities since the result can be visualized as a dendrogram. This dendrogram is then cut at different heights to generate different numbers of clusters.

- *Density-based clustering* identifies clusters as contiguous regions with a high density of cells. Unlike hierarchical clustering or k-means, density-based clustering assumes that all clusters are equally dense (i.e. cell populations are equally homogenous).

Thus, a key decision for clustering methods is how many groups to identify. Coarse clusterings identify a small number of very distinct clusters which are more likely to correspond to cell-types; whereas fine clustering identifies a large number of less distinct clusters which may correspond to different cell-states. Most clustering algorithms require either the number of clusters (k) or parameters relating with the density of the cluster to be defined a priori by the user (Andrews and Hemberg, 2018). However, choosing an appropriate k is difficult since there is no generally accepted method to do so and the identity of any of the cells assayed a priori is quite unrealistic. Finally, basing the choose of k on reliable marker genes can be misleading since they are only available for a few well-characterized cell-types.

Last phase (i.e. Phase 3) of scRNA-seq data analysis regards the interpretation of groups identified by the clustering algorithm that is not trivial. The biological interpretation of the final results (i.e. association to the cell types) often requests other times and external tools (e.g. functional enrichment analysis using external annotations). Although novel cell populations can be readily identified using external methods, these findings must be validated using external data or experiments to ensure they are not technical artefacts. Indeed, due to the heuristic nature of clustering algorithms, they will always find some partitioning and even when clusters are a result of biological effects, rather than noise, those effects may not represent differences in cell-type. Today, there are no accepted standards for the criteria required to label a cell population as a novel cell type. Defining cell-types based on transcriptional differences is difficult since transient differences in cell-state (e.g. cell-cycle stage) can have a larger effect on the global transcriptome than cell-type (Andrews and Hemberg, 2018).

Finally, even though several computational pipelines are now present that include all the cited phases, only few of them provide analysis flexibility while also achieving functional (i.e. information about the data and the tools used are saved as metadata) and computational reproducibility (i.e. a real image of the computational environment used to generate the data is stored) through a user-friendly environment.

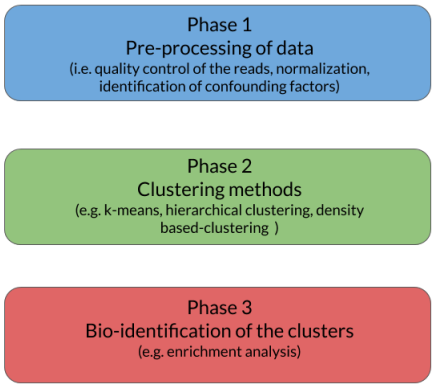


FIGURE 3.1: **Typical workflow of scRNA-seq data analysis.** The three phases represent the typical steps to follow for a scRNA-seq data analysis (i.e. Pre-processing, Clustering, Biological identification of the clusters obtained

.

3.3 State-of-art pipelines for scRNA-seq data analysis

At the best of our knowledge, four are the workflows available for a complete single-cell analysis: (i) simpleSingleCell, Bioconductor workflow package (Lun, McCarthy, and Marioni, 2016); (ii) Granatum, web-based single-cell RNA-seq analysis suite (Zhu et al., 2017); (iii) SCell, graphical workflow for single-cell analysis (Diaz et al., 2016); and (iv) R toolkit Seurat (Butler et al., 2018).

- simpleSingleCell (Lun, McCarthy, and Marioni, 2016) is a computational workflow for basic analysis of scRNA-seq data, using software packages from the open-source Bioconductor project (release 3.4) (Huber et al., 2015). Starting from a count matrix, this workflow contains the steps required for quality control to remove problematic cells; normalization of cell-specific biases, with and without spike-ins; cell cycle phase classification from gene expression data; data exploration to identify putative subpopulations; and finally, HVG and marker gene identification to prioritize interesting genes.
- Granatum (Zhu et al., 2017) is a web-based scRNA-Seq analysis pipeline. Granatum has a comprehensive list of modules, including plate merging and batch-effect removal, outlier-sample removal, gene-expression normalization, imputation, gene filtering, cell clustering, differential gene expression analysis, pathway/ontology enrichment analysis, protein network interaction visualization, and pseudo-time cell series construction.
- SCell (Diaz et al., 2016) is an integrated software tool for quality filtering, normalization, feature selection, iterative dimensionality reduction, clustering and the estimation of gene-expression gradients from large ensembles of single-cell RNA-seq datasets. Moreover, SCell can regress/interpolate gene expression on PCA space, visualize expression gradients, and estimate expression kinetics along minimum spanning trees and minimum weight paths.
- Seurat (Butler et al., 2018) is an R package designed for quality control, analysis, and exploration of single-cell RNA-seq data. Seurat aims to enable users to identify and interpret sources of heterogeneity from single-cell transcriptomic measurements, and to integrate diverse types of single-cell data.

3.4 Aim

The single-cell sequencing results (Butler et al., 2018; Alessandri et al., 2019) suggest the presence of a complex organization of heterogeneous cell states that produce a system-level functionalities. A mandatory element of scRNA-seq is **the availability of dedicated bioinformatics workflows that allow exhaustive analysis of data. The analysis has to retrieve the corresponding cellular organization and differentiation in cellular types.** For this purpose, many related algorithms and tools have been developed, but few computational workflows provide complete analysis from raw data to biological cluster identification. Moreover, few of these pipelines provide flexibility while also achieving functional (i.e. information about the data and the tools used are saved as metadata) and computational reproducibility (i.e. a real image of the computational environment used to generate the data is stored) through a user-friendly environment. Beyond the computational challenges, there is also a biological interpretation challenge related to the post-processing phases of the analysis. Indeed, the majority of the pipelines rely on external tools to improve the biological interpretation of the clusters obtained, using enrichment, pathway analysis of the genes or functional association network. In this context we propose rCASC, a pipeline for reproducible classification analysis of single-cell sequencing data (Alessandri et al., 2019). rCASC is a modular

workflow providing an integrated analysis environment (from count generation to cell subpopulation identification) exploiting Docker containerization to achieve both functional and computational reproducibility in data analysis. The main advantages of rCASC pipeline are the many pre-processing tools to remove low-quality cells and/or specific bias (e.g. cell cycle), the different clustering techniques based on different distance metrics and a new metric (Cell Stability Score) for the evaluation of clusters quality. In the context of biological interpretation challenge, we also provide rMLSC (reproducible Machine Learning analysis for single-cell sequencing data), a new module based on machine learning approach to identify cluster-specific gene signatures from rCASC results. At the best of our knowledge, this type of approach has never been applied to single cell data with the purpose of improving cluster interpretation. rMLSC strategy exploits the classification algorithm based on Random Forest decision tree to analyze the clusters obtained from rCASC.
rMLSC identifies for each cluster a subset of genes whose disposition in the decision trees is a reflection of the differences among the clusters. This approach is organized in five phases and tasks that converge to the derivation of a set of gene signatures for all the clusters. The entire strategy of single cell data analysis is reported in Figure 3.2.

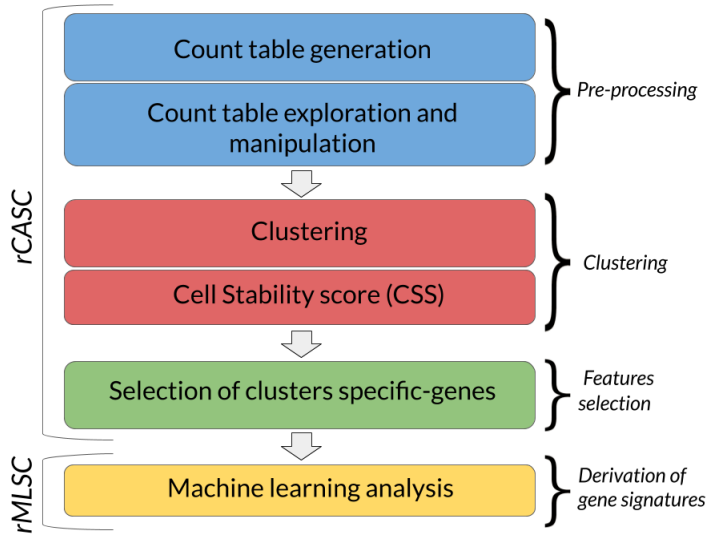


FIGURE 3.2: **Computational workflow for single cell data analysis.** The workflow is composed of rCASC pipeline where blue boxes indicate pre-processing tools, red boxes define clustering tools and green boxes indicates gene signature tools, and rMLSC where yellow box indicates the machine learning implementation.

3.5 Computational strategy to analyze scRNA-seq data

In this section, rCASC computational workflow (paragraph 3.5.1) and rMLSC (paragraph 3.5.3) are explained. rCASC pipeline was published in 2019 (Alessandri et al., 2019) while rMLSC is part of a new implementation and the manuscript is in preparation.

3.5.1 rCASC computational pipeline

The key elements of the rCASC workflow are shown in Figure 3.3 and the main functionalities are summarized in the following sections.

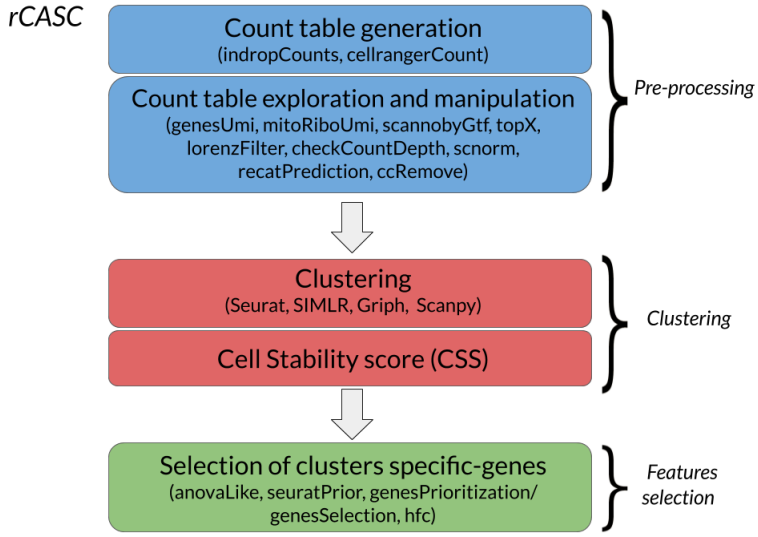


FIGURE 3.3: **rCASC workflow.** Blue boxes indicate pre-processing tools. Red boxes define clustering tools. Green boxes indicates gene signature tools.

Count table generation In rCASC, the generation of the count table start from fastq files generated with inDrop or 10X Genomics platforms , or with a count matrices as directed input (Figure 3.3). The inDrop and 10X Genomics methods involve co-encapsulation of cells along with a reverse transcription (RT) mix and hydrogel particles carrying primers that can be released upon UV excitation (Zhang et al., 2011). The barcoding strategy for both methods uses initial universal sequence for PCR/sequencing, cell barcode and a unique molecular identifier (UMI) sequence with a poly-dT sequence to capture mRNAs (Zhang et al., 2011). However, the primer on the inDrop beads also has a photo-cleavable moiety and a T7 promoter. The cell capture efficiency can reach markedly higher levels in both approaches with ~80% bead occupancy for and a cell capture rate of ~50% (Zhang et al., 2011). After encapsulation, the entire beads from 10X are dissolved to release all of the primers into the solution phase to boost the efficiency of mRNA capture. inDrop instead mobilizes the primers by UV-irradiation-induced cleavage. Then, the RT reaction is carried out inside the droplets (Zhang et al., 2011) with inDrop employing CEL-seq method (Hashimshony et al., 2012), whereas 10X following a template-switching protocol (Macosko et al., 2015), similar to the popular Smart-seq chemistry (Ramsköld et al., 2012). The transcription step in inDrop extends the library preparation time beyond 24 hour while 10X Genomics can be completed within a day. In rCASC, data derived from inDrop platform are reformatted by *indropIndex* function and allows the generation of the transcript index required to convert fastq in counts. Finally, the *indropCounts* function converts reads in unique molecular identifier (UMI) counts. Data derive from 10X Genomics platforms, the function *cellrangerCount* converts fastq to UMI matrix using any of the genome indexes with the *cellrangerIndexing* function.

Count table exploration and manipulation rCASC provides various data inspection and preprocessing tools (Figure 3.3, Dark grey boxes with white character). First of all, the *genesUmi* function generates a plot where the number of detected genes is plotted for each cell with respect to the number of UMI. Moreover, *mitoRiboUmi* function calculates the percentage of mitochondrial/ribosomal genes with respect to the total number of detected genes in each cell and plots the percentage of mitochondrial genes with respect to percentage of ribosomal genes. *mitoRiboUmi* allows researchers to identify cells with low information

content (i.e. those cells with few detectable genes)(e.g. <100 genes/cell), little ribosomal content, and high content of mitochondrial genes, which indicate cell stress.

Then, the function *scannobyGtf* uses ENSEMBL gtf and the R package refGenome to associate gene symbol with the ENSEMBL gene ID and remove from the dataset mitochondrial/ribosomal genes and stressed cells, detected by *mitoRiboUmi function*.

The function *lorenzFilter* embeds the Lorenz statistics developed by Diaz et al., 2016, a cell quality statistic correlated with cell live-dead staining. Specifically, the outlier filtering for single-cell RNA-seq experiments designed by Diaz et al., 2016 estimates which genes are expressed at background levels in each sample; then samples with significantly high background levels are discarded (Diaz et al., 2016).

We implemented the functions *checkCountDepth/scnorm* to detect the presence of sample-specific count–depth relationship (Bacher et al., 2017) (i.e. the relationship existing between transcript-specific expression and sequencing depth) and to adjust the count table for it. Specifically, *checkCountDepth* initially executes a quantile regression, thus estimating the dependence of transcript expression on sequencing depth for every gene. Then, genes with similar dependence are aggregated. *Scnorm* function, after executing *checkCountDepth*, performs a new quantile regression to estimate scale factors within each group of genes. Then, sequencing depth adjustment is done within each group using the estimated scale factors. Furthermore, we added two other functions *recatPrediction* and *ccRemove* which are based on Liu et al., 2017 and Barron and Li, 2016 respectively. The function *recatPrediction* organizes the single-cell data to reconstruct cell cycle pseudo time-series and is used to understand whether a cell cycle effect is present. The *ccRemove* function embeds their scLVM (single-cell latent variable model) algorithm, which uses a sophisticated Bayesian latent variable model to reconstruct hidden factors in the expression profile of the cell cycle genes. This algorithm is able to remove cell cycle effect from real single-cell RNA-seq datasets.

Clustering For the identification of cell subpopulations we implemented four approaches: Seurat (Butler et al., 2018), SIMLR (Wang et al., 2017), Griph (Serra et al., 2019), and Scanpy (Wolf, Angerer, and Theis, 2018) (Figure 3.3, Dark grey boxes with black character).

- Seurat (Butler et al., 2018) is a toolbox for single-cell RNA-seq data analysis. We implemented in rCASC one of the clustering procedures present in the Seurat toolbox. The function *seuratPCAEval* has to be run before executing the clustering program to identify the metafeatures (i.e. the subset of PCA components describing the relevant source of cell heterogeneity) to be used for clustering. The *seuratBootstrap* function implements data reduction and clustering. Specifically, cells undergo global scaling normalization (i.e. LogNormalize method) and scaling factor 10,000. Subsequently, a linear dimensional reduction is done using the range of principal components defined with *seuratPCAEval*. Then, clustering is performed using the cell PCA scores. The Seurat clustering procedure, embedded in seuratBootstrap, is based on the Louvain modularity optimization algorithm.

- SIMLR (Wang et al., 2017) implements a k-mean clustering, where the number of clusters (i.e. k) is taken as input. SIMLR requires as input raw counts that are log10 transformed. SIMLR is capable of learning an appropriate cell-to-cell similarity metric from the input single-cell data and can exploit it for the clustering task. In the learning phase SIMLR identifies a distance metric that better fits the structure of the data by combining multiple Gaussian kernels (Wang et al., 2017). Thus, the tool can deal with the large noise and dropout effects (i.e. a transcript is expressed in a cell but is entirely undetected in its mRNA profile) of single-cell data, which could not easily fit with specific statistical assumptions made by standard dimension reduction algorithms

(Wang et al., 2017). The function *simlrBootstrap* controls the clustering procedure and the function *nClusterEvaluationSIMLR* is exploited to estimate the (sub)optimal number "k" of clusters.

- Griph clustering (Serra et al., 2019) is closer to agglomerative clustering methods because every node is initially assigned to its own community and communities are subsequently built by iterative merging.
- Scanpy (Wolf, Angerer, and Theis, 2018) uses for clustering also Louvain algorithm that detects communities as groups of cells that have more links between them than expected from the number of links the cells have in total. The optimized modularity function includes a resolution parameter, which allows the user to determine the scale of the cluster partition.

Cell Stability Score We developed, for Seurat, SIMLR, griph, and scanpy, a procedure to measure the cluster quality on the basis of data structure. The rationale of our approach is that cells belonging to a specific cluster should be little affected by changes in the size of the dataset (e.g. removal of 10% of the total number of cells used for clustering). Thus, we developed a metric called **Cell Stability Score (CSS)**, which describes the persistence of a cell in a specific cluster upon jackknife resampling and therefore offers a peculiar way of describing cluster stability (Figure 3.4).

In detail, a set of cells to be organized in clusters (Figure 3.4 A) is analyzed with one of the clustering methods, applying a user defined k number of clusters (Figure 3.4 B). A user defined percentage of cells is removed from the original data set and these cells are clustered again (Figure 3.4 C). The clusters obtained in each bootstrap step are compared with the clusters generated on the full dataset using Jaccard index (Figure 3.4 D,E). If the Jaccard index is greater of a user defined threshold, e.g. 0.8, the cluster is called confirmed in the bootstrap step (Figure 3.4 F). Then to each cell, belonging to the confirmed cluster, cell stability score value is increased of 1 unit (Figure 3.4 G). At the end of the bootstrap procedure, cells are labeled with different symbols describing their cell stability score in a specific cluster. See Supplementary 6.3.1 for further details.

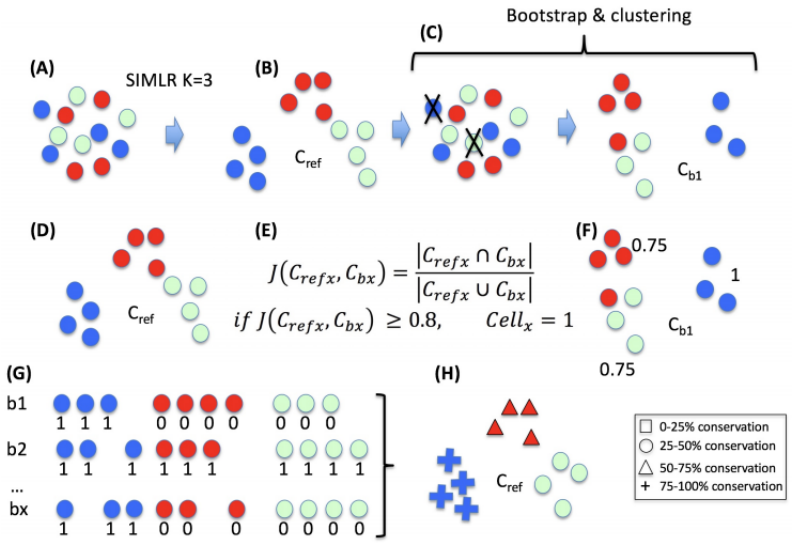


FIGURE 3.4: **Cell Stability Score (CSS).** Figure from Alessandri et al., 2019

.

Selection of clusters specific-genes To select the most important features of each cluster we implemented the *anovaLike* function with the edgeR ANOVA-like method for single cells (Robinson, McCarthy, and Smyth, 2010) and in the functions *seuratPrior* and *genesPrioritization/ genesSelection*, respectively, the Seurat and SIMLR gene prioritization methods. The *hfc* function allows visualization of the genes prioritized with the above methods as a heat map and provides plots of prioritized genes in each single cell (Figure 3.3) (Alessandri et al., 2019).

3.5.2 rCASC scalability, reproducibility and comparison with state-of-art pipelines

rCASC computing time for one analysis depends on multiple parameters: (i) the number of permutations performed in parallel, (ii) the number of cells under analysis, (iii) the clustering tool in use, and (iv) the hardware used for the analysis. Concerning the amount of RAM required for each permutation run in parallel, for up to 5,000 cells the maximum amount of RAM required is about 4 GB; from 10,000 to 100,000 cells, the maximum RAM required is about 20 GB. Independently by the clustering approach and the size of the dataset, it is suggested running $\geq$ 100 permutations to correctly estimate CSS (Alessandri et al., 2019). All the computational tools in rCASC are embedded in Docker images stored in a public repository on the Docker hub. Parameters are delivered to Docker containers via a set of R functions, part of the rCASC R github package. To simplify the use of the rCASC package for users without scripting experience, R functions can be controlled by a dedicated GUI, integrated in the 4SeqGUI tool (Beccuti et al., 2017b), which is also available as a github package (Figure 3.5).
The overall characteristics of rCASC were compared with the four other workflows for single-cell analysis (Figure 3.6): simpleSingleCell (Lun, McCarthy, and Marioni, 2016),Granatum (Zhu et al., 2017), SCell (Diaz et al., 2016) and R toolkit Seurat (Butler et al., 2018). The comparison was based on the following elements:

- supported single-cell platforms
- types of tools provided by the workflow
- type of reproducibility granted by the workflow
- tool flexibility

rCASC is the only workflow providing support at the fastq level because all the other packages require as input the processed count table. Cell quality control and outlier identification is available in all the workflows but Granatum. Association of ENSEMBL gene IDs to gene symbols is only provided by rCASC. All the workflows provide gene-filtering tools but simpleSingleCell. All packages provide normalization procedures to be applied to raw count data. However, rCASC is the only tool providing both Seurat specific normalization (Butler et al., 2018) and count-depth specific normalization (Bacher et al., 2017). The workflows implement different data reduction and clustering methods. rCASC integrates four clustering tools (i.e. Seurat (Butler et al., 2018), SIMLR (Wang et al., 2017), griph (Serra et al., 2019), and scanpy (Wolf, Angerer, and Theis, 2018), which differ in the metrics driving the clustering analysis. In rCASC, we have implemented a cell stability score, which uses the Jaccard index to estimate the stability of each cell in each cluster. The CSS provides an enhanced description of each cluster because it allows the identification of a subset of cells, in any cluster, that are particularly sensitive to perturbation of the overall dataset structure (i.e. cell bootstrapping). Specifically, we have implemented the *clusterboot* function from the fpc R package (Hennig, 2013), which allows the evaluation of cluster stability using a personalized clustering function. To the best of our knowledge, rCASC is the only workflow

performing clustering in the presence of data perturbation (i.e. removal of a subset of cells), and measuring cluster quality using the CSS and silhouette score (SS), a cluster quality metric measuring the consistency within clusters of data. Gene feature selection approaches are implemented in a different way in the five workflows. Granatum is the only one providing biological inference. Granatum and Seurat implement various statistical methods to detect cluster-specific gene signatures. rCASC embeds an ANOVA-like statistics derived from the EdgeR Bioconductor package (Robinson, McCarthy, and Smyth, 2010) and Seurat/SIMLR gene prioritization procedures. Visualization of gene signatures by heat map, with cells colored on the basis of gene expression, is only provided by rCASC. Considering reproducibility, only rCASC provides both computational and functional reproducibility. Finally, rCASC is the only one providing both a command line interface and GUI.

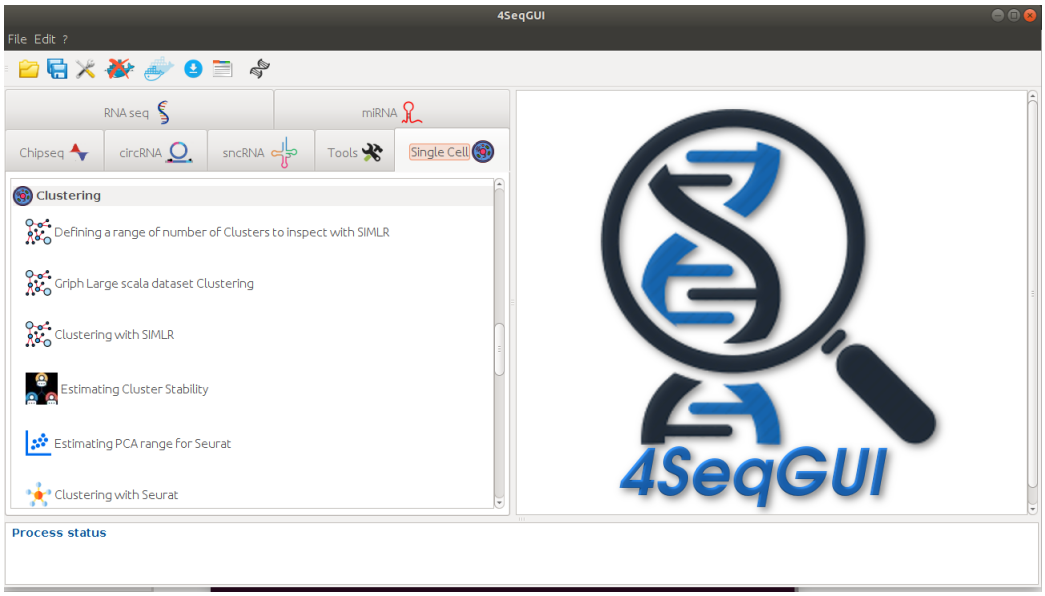


FIGURE 3.5: **rCASC graphical interface within 4seqGUI.**

.

		rCASC (stand alone)	simpleSingleCell (stand alone)	Granatum (web)	Scell (stand alone)	Seurat (stand alone)
Platforms		10Xgenomics, inDrop, counts table	counts table	counts table	counts table	counts table
Tools	*Fastq conversion in counts table*	Y	-	-	-	-
	Quality Control / Outlier Filtering	Y	-	Y	Y	Y
	Annotation	ENSEMBL ID -> Gene Symbol	-	-	-	-
	Genes filter	Y	-	Y	Y	Y
	Data normalization	Y	Y	Y	Y	Y
	Cell cycle bias removal	Y	-	-	Y	Y
	Data dimensionality reduction	Y	Y	Y	Y	Y
	Supported clustering methods	tSne, SIMLR, Seurat (PCA), griph, scanpy (UMAP)	Walktrap	Non-negative matrix factorization, K-mean (Euclidean), K-mean (tSne)	PCA, k-means, Gaussian mixture, Minkowski weighted k-means, DBSCAN	PCA, tSne, ica, dmap
	Cluster quality score	Silhuette, Cell Stability Score	-	-	-	-
	Features selection and visualization	Y	Y	Y	-	-
	Supported methods	ANOVA-like (edgeR), SIMLR and Seurat genes prioritization	filtering on expression	NODES, SCDE, EdgeR, Limma	-	wilcox, bimod, roc, t-test, tobit, negbinom, MAST, DESeq2
	Biological inference	-	-	Y	-	-
Reproducibility	*Functional reproducibility*	Y	Y	-	-	Y
	Computational reproducibility	Y	-	Y	Y	-
Flexibility	*line command execution*	Y	Y	-	-	Y
	graphical interface	Y	-	Y	Y	-

FIGURE 3.6: **Comparison of analysis features available in rCASC vs other single-cell analysis workflows (simpleSingleCell, Granatum, Scell, Seurat).** Y = yes; - = not present.

.

3.5.3 rMLSC

The purpose of rMLSC workflow is to exploit machine learning approaches to mainly improve the biological interpretation of the clusters obtained by rCASC analysis, deriving a set of gene signatures for each cluster. This task is performed by a classification algorithms (e.g. Random Forest or Decision Tree) through the comparison of all the genes in the dataset and their expression values, generating decision trees for each cluster. Then, exploration and filtering of the trees through several methods allows the identification of a subset of genes whose disposition in the decision trees is a reflection of a difference among the clusters. The biological characterization of these signatures, both individually and compared among the clusters, help the user for a deeper understanding of dataset composition as cellular types, putative hierarchical lineage of the cells or differences in cells behaviour upon different conditions.

The analysis of scRNA data based on the application of classification learning methods is organized in five phases to identify s set of gene signatures for all the clusters (Figure 3.7). In detail:

1. Phase 1: Pre-processing of input data
2. Phase 2: Machine learning classification analysis
3. Phase 3: Analysis of classification results
4. Phase 4: Exploration of the set of decision trees for each cluster
5. Phase 5: Derivation of gene signatures set

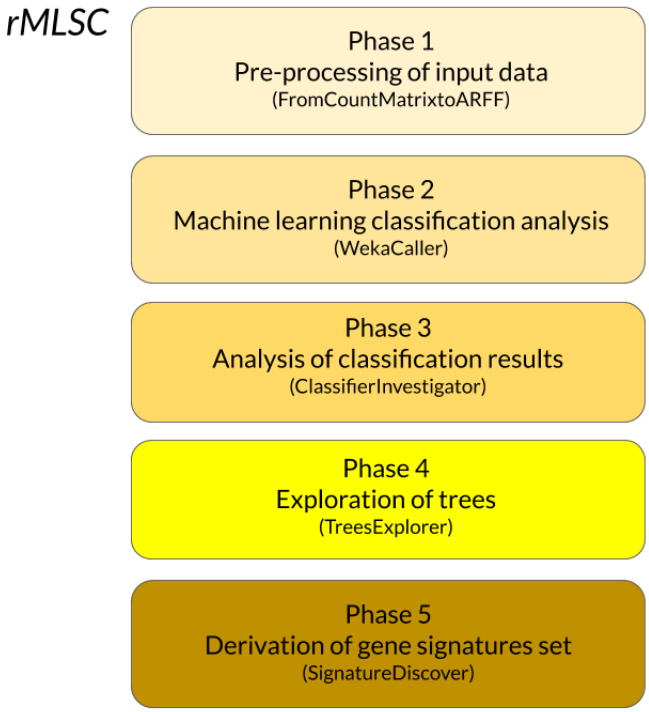


FIGURE 3.7: **Machine learning workflow**

.

Phase 1: Pre-processing of input data

The machine learning analysis of scRNA data starts with a pre-processing R script able to convert the output of rCASC pipeline in a proper input for classification algorithms. Gene expression matrix, derived from a scRNA-seq experiment and sequenced with platforms as 10X Genomics or inDrop protocol, are often composed as spreadsheet format with columns representing cells and rows representing genes. Each cell of the spreadsheet contains the value associated with the expression of one gene in one cell.
We developed the R function, *FromCountMatrixtoARFF* (1), that takes as input two outputs of rCASC pipeline: the scRNA gene expression count matrix (a) and the association cell-cluster file (b) (i.e. a tab-separated file with two columns indicating the cell barcode identification and the associated cluster retrieved from rCASC analysis), the number of clusters obtained (c), the percentage of instances in training (d) and test (e) and the number of runs to perform in the classification analysis (h).

Algorithm 1 Creation of Training and Test files

procedure FROMCOUNTMATRIXTOARFF(a, b, c, d, e, h)
    INPUT:
    a = scRNA count matrix
    b = Association cell-cluster file
    c = Number of clusters
    d = Percentage of instances in training
    e = Percentage of instances in test
    h = Number of runs
    OUTPUT:
    $Training$ = $\{\text{Training}_i\}_{i=1,h}$
    $Test$ = $\{\text{Test}_i\}_{i=1,h}$
end procedure

FromCountMatrixtoARFF is able to convert the matrix in h training and test datasets (in ARFF format) for each cluster (i.e. c), usable directly for the classification analysis.
FromCountMatrixtoARFF function randomly splits the matrix in two datasets that contain respectively the d% (i.e. *training set*; default 70%) and e% (i.e. *test set*; default 30%) of the cells (i.e. instances). The final ARFF files for *training and test set* are composed of an upper list of all attributes genes and a bottom list where each cell is represented by a row indicating the relative gene expression values separated by commas with last column indicating the belonging cluster (i.e. for training set) or a question mark (i.e. for test set). Since the goal of the analysis is to find a signature able to differentiate one cluster against the others, the last column of training set files will contain the label of the cluster to investigate in that analysis (e.g. $\text{Cluster}_{i=1}$) versus the others indicated as generic Cluster_x, where x meanings one cluster at time or the merge of other interested clusters.

For example, supposing to have a count matrix with 1000 cells clustered from rCASC in three clusters distributed as following: $\text{Cluster}_{i=1}$ 200 cells, $\text{Cluster}_{i=2}$ 100 cells, $\text{Cluster}_{i=3}$ 700 cells, and want to perform the classification analysis 600 times for each cluster. Thus starting from $\text{Cluster}_{i=1}$, *FromCountMatrixtoARFF* function will create:

- Training_h = 600 training inputs containing 650 cells (i.e. 70% of cells of cluster_x where x is equal to the merge of $\text{cluster}_{i=2}$ and $\text{cluster}_{i=3}$ cells) and 140 cells of $\text{cluster}_{i=1}$ randomly choose every time. All the training input can display for each cell two labels: $\text{cluster}_{i=1}$ or cluster_x.

- Test_h = 600 test inputs containing 240 cells (i.e. 30% of the total cluster_x) and 60 cells of $\text{cluster}_{i=1}$, randomly choose. All the test input display a question mark in place of cluster labels.

Phase 2: Machine learning classification analysis

In the second phase, *WekaCaller* bash script was developed to exploit Weka classification algorithms (i.e. choose by the user) (Witten et al., 2016) on the generated set of ARFF training and test input (2). *WekaCaller* calls in iterative way a classification algorithm using standard Weka command line on each Training_h and Training_h files h times. The Weka parameters setting are the main explained in the previous section "Classification analysis through Weka tool". *WekaCaller* creates for each h run a folder that will contain: (i) summary output of the classification with the relative decision tree of the specific h run as Figure 6.8, (ii) a prediction file for test set with detailed results of classification for each instance (i.e. correct/incorrect classification for each cell on the test set).

Algorithm 2 Classification analysis

1: **procedure** WEKACALLER($c, h, Training, Test$)
2: **INPUT:**
3: c = Number of clusters
4: h = Number of runs
5: $Training$ = { Training_i }$_{i=1,h}$
6: $Test$ = { Test_i }$_{i=1,h}$
7: **OUTPUT:**
8: { $summary$ }$_{i=1,h}$
9: { $prediction$ }$_{i=1,h}$
10: **end procedure**

Resuming the previous example, *WekaCaller* will perform the analysis for h runs firstly for $\text{Cluster}_{i=1}$, then for $\text{Cluster}_{i=2}$ and finally for $\text{Cluster}_{i=3}$. As previously explained, starting from $\text{Cluster}_{i=1}$ classification, the model will be firstly trained with the training input (i.e. known label cell-cluster) and then used to classify the cells in test input (i.e. unknown label cell-cluster). Thus, each prediction file of $\text{Cluster}_{i=1}$ will contain the classification of each cell of test dataset that can return $\text{Cluster}_{i=1}$ or Cluster_x. Moreover, if h runs are performed for each cluster, a total of $h * c$ trees and classification will be returned.

Phase 3: Analysis of classification results

To retrieve and summarize the results of the classification we developed *ClassifierInvestigator* bash script.

For the task 1, *ClassifierInvestigator* takes as input the set of { $prediction$ }$_{i=1,h}$ files with the complete results of the classification for each cluster_i and (i) count the percentage of correct classifications; (ii) calculate the accuracy of the classifier as the total percentage of cells of one cluster correctly classified among all the h runs (3).

Resuming the previous example, supposing to have 1000 total cells divided in three clusters (e.g. 200 cells of $\text{Cluster}_{i=1}$, 100 cells of $\text{Cluster}_{i=2}$, 700 cells of Cluster $_{i=1}$. The test set input contains 300 cells (i.e. 60 cells for Cluster and 240 for Cluster 2 plus Cluster 3) and for each cluster $prediction = h$ files were generated. Starting with $\text{Cluster}_{i=1}$, *ClassifierInvestigator* takes as input all the $prediction$ files and calculate singularly, for each of these,

Algorithm 3 Analysis of the classification results

procedure CLASSIFIERINVESTIGATOR($c, prediction$)
    INPUT:
    c = Number of clusters
    $\{\ prediction\ \}_{i=1,h}$
    OUTPUT:
    $count$= $\{\ \text{count}_i\ \}_{i=1,h}$
    $estimate$ = $\{\ \text{estimate}_k\ \}_{k=1,c}$
end procedure

the correct classification percentage.

Phase 4: Exploration of the set of decision trees for each cluster

For this phase purpose, we developed *TreesExplorer* tool (i.e. Phyton tool) able to parse and explore all the trees generated from the classification algorithms.

Algorithm 4 Exploration of decision trees

procedure TREESEXPLORER($c, summary$)
    INPUT:
    c = Number of clusters
    $\{\ summary\ \}_{i=1,h}$
    OUTPUT:
    $\{Level - gene\ \}_{k=1,c}$
    $\{Entropy\ \}_{k=1,c}$
end procedure

In detail, *Trees Explorer* takes as input, for each cluster the folders containing all the summary outputs generated from 2 (i.e. Figure 6.8), extracts and parses all the decision trees returning $\{Level - gene\ \}_{k=1,c}$ file, for each cluster, with the list of genes and levels of trees where they are present. Moreover, $\{Entropy\ \}_{k=1,c}$ files for each cluster are also reported and contain the entropy values for each level of the trees. Entropy as it relates to information theory, is a measure of the amount of the information being processed. A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances belonging to the same class. Decision tree algorithms use an impurity measure, such as entropy or Gini index to calculate the homogeneity of a dataset. For example, if the dataset is completely homogeneous the entropy is zero otherwise if it is equally divided it has entropy near to one. *Trees Explorer* is able to calculate the entropy measure for each tree generated in the analysis.

Considering one tree, the entropy is calculated firstly for each node n of the tree (i.e. starting from the root) with 3.1 equation where the entropy of the node is multiplied for the number of positive example that fall into that node.

$$H_n = -p_n * \log_2(p_n) - (1 - p_n) * \log_2(1 - p_n) \tag{3.1}$$

Then, when the entropies of all the nodes are calculated, the 3.2 is exploited to derive the entropy of each level (l) as a *weighted average of the entropy of nodes (n) at level l*.

$$H_l = \frac{\sum_{n \in N_l} H_n * |n|}{\sum_{n \in N_l} |n|} \tag{3.2}$$

Phase 5: Derivation of gene signatures set

In the last phase, the set of gene signatures is derived.

Algorithm 5 Derivation of gene signatures

```
1: procedure SIGNATUREDISCOVER(Level − gene, Entropy, c)
2:     INPUT:
3:     {Level − gene }k=1,c
4:     {Entropy }k=1,c
5:     c = number of clusters
6:     OPERATIONS:
7:     Filter1 = threshold on entropy measure
8:     Filter2 = threshold on the number of cells
9:     OUTPUT:
10:    {Level − gene − filtered }k=1,c
11: end procedure
```

SignatureDiscover processes the $\{Level - gene\}_{k=1,c}$ files of the clusters, derived from *TreeExplorer* (i.e. CSV file containing for each row, the gene and the level in which the gene is found (i.e. from the root to the leaf nodes)) appling two different filters. The first filter ($Filter1$) takes into account the $\{Entropy\}_{k=1,c}$ file where median of weighted entropy on each level of the trees are reported. *SignatureDiscover* filters out all the levels whose entropy value is under a median entropy threshold (i.e. *Entropy threshold*).

Then, *SignatureDiscover* calculates the percentage of cells expressing the genes in the cluster under analysis, presented in $Level - gene$ files. For this estimate, we considered a gene as expressed in a cell if it has at least 3 UMIs assigned (Alessandri et al., 2019). Then, the tool applies the $Filter2$ on the $Level - gene$ files that is to retain only genes expressed above a percentage threshold of cells. After filtering, the set of gene signatures derived are compared among the clusters. In particular, in this task, the number of unique and putative overlapping genes are researched and reported in a table format. Finally, we visualize the resulted gene signatures on the basis of gene expression data exploiting the heatmap visualization method in ggplots R package.

3.6 Results

3.6.1 Analysis of single cell data using rCASC and rMLSC

rCASC/rMLSC were used to analyze the single-cell dataset from Pace et al., 2018. The authors explored the role of histone methyltransferase Suv39h1 in murine CD8+ T cells activated after Listeria monocytogenes infection. CD8+ T cells are T lymphocytes (a type of white blood cell) that play a role in immune system when normal cells are infected (particularly with viruses), damaged or when the presence of cancer cells is revealed. In geneal, after the antigen recognition, naïve $CD8^+$ T lymphocytes establish specific heritable transcription programs that define progression to long-lasting memory cells or to short-lived effector cells. Since it remains unclear how chromatin dynamics contributes to the control of gene expression programs, the authors explored the role of gene silencing by the histone methyltransferase Suv39h1. In murine CD8+ T cells activated after an infection, Suv39h1-dependent trimethylation of histone H3 lysine 9 controls the expression of a set of stem cell-related memory genes. When Suv39h1 is silenced in murine CD8+ T effector cells, a defect in silencing of stem/memory genes is revealed. As a result, Suv39h1-defective CD8+ T cells show sustained survival and increased long-term memory reprogramming capacity. Thus, Suv39h1 plays a critical role in marking chromatin to silence stem/memory genes during CD8+ T effector terminal differentiation. The single-cell sequencing dataset presented in this paper is composed as described in Table 3.1. 11.918 cells were derived from WT and KO-Suv39h1 mice distributed as following: for naive cells, they sequenced 2426 and 2372 cells from wild-type and Suv39h1-KO mice, respectively. Then, they processed two technical replicates for both wild-type and Suv39h1-KO infected mice (approximately 3161 and 2293 cells, respectively) to retrieve WT CD8+ T cells and KO Suv39h1-defective CD8+ T cells, respectively. Finally, an additional biological replicate (from different mice, 1005 wild-type and 661 Suv39h1-defective CD8+ T cells) was used. The authors performed the experiment firstly isolating immune cells by FACS 7 days after a Listeria monocytogenes infection and then using Chromium single-cell RNA-seq kit for sequencing. Briefly, the initial step taking place into the Chromium System consisted in performing an emulsion where individual cells were isolated into droplets together with gel beads coated with unique primers bearing 10X cell barcodes, UMI (unique molecular identifiers) and poly(dT) sequences. For a second time, reverse transcription reactions were engaged to generate barcoded full-length cDNA followed by the disruption of emulsions using the recovery agent and cDNA clean up. Global Amplification of bulk cDNA was achieved using Applied Biosystems kit. From each cDNA preparation, indexed libraries were constructed following these five steps: (1) fragmentation, end repair and A-tailing; (2) size selection with SPRI select beads; (3) adaptor ligation; (4) post-ligation cleanup with SPRI select beads; (5) sample index PCR and final cleanup with SPRI select beads. Library quantification and quality assessment were achieved by Qubit fluorometric assay (Invitrogen) and Bioanalyzer Agilent 2100 System using a High Sensitivity DNA chip. At the end, indexed libraries were tested for quality, equimolarly pooled and sequenced on an Illumina HiSeq2500 using paired-end 26x98 bp as sequencing mode, using a full Rapid flow cell, with a coverage around 75M reads per sample. Approximately 500-1,000 cells were obtained corresponding to 86,000 reads/cell. Each pool of cells was tested for library quality and concentration. CellRanger pipeline (10X Genomics company) was used for demultiplexing of sequencing output and single cell count table generation. For data analysis, they used Seurat pipeline (Butler et al., 2018) and retrieved four clusters (Figure 3.8). The clusters were biologically identified using differential gene expression analysis provided by Seurat in Naive cells, Memory precursors, Cycling cells and Effector cells both for WT and KO mice. The scRNA-seq analysis revealed the alternative expression of the memory precursors with stem cell–like properties and effector signatures in the two

cell types, respectively, and the concomitant low or high expression of these two signatures in the cycling cells. These results suggest that cycling cells may represent bipotent differentiation intermediates expressing both effector and stem/memory potential. Furthermore, the commitment to effector differentiation paths appears to be acquired by the silencing of stem/memory genes (Figure 3.9 from Pace et al., 2018). Thus, the stem cell/memory gene expression program is under the control of Suv39h1 by imposing the H3K9me3 modification on chromatin at the corresponding loci. However, the identification of the four cell types through single cell analysis performed by the authors is actually driven by a set of 24 immunological genes that are known in literature as involved in the commitment process of the immune cells. Thus, any other gene, even not strictly related to immune functions, is investigated as putative contributor to the classification in the four cell types. Thus, to perform an unbiased and more exhaustive identification of the dataset, we performed the entire single cell analysis in two steps. The first step (Step 1) is performed with rCASC and is focused on the identification of cell clusters and DEGs with a particular attention to the assessment of cells and clusters stability. Then in the second step (Step 2) we investigate a new classification approach, rMLSC, based on machine learning techniques to identify clusters-gene signatures, retrieved from the total set of genes available from the experiment of Pace et al., 2018. The final goal is to improve the biological interpretation of the clusters picking new putative hide information from the whole dataset.

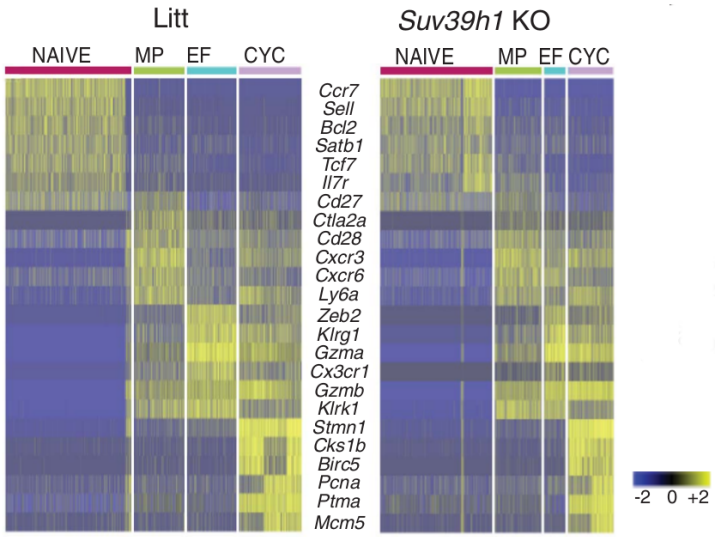


FIGURE 3.8: **Heatmaps of 24 genes identified by the authors as representative of the four clusters.** For each subset category (Naive, Memory Precursor, Effector and Cycling cells) a signature of 6 genes is identified as representative. Columns represent cells; rows represent genes; Litt. (i.e. wildtype mice); Suv39h1 KO (i.e. knockout mice). (Figure from Pace et al., 2018)

3.6.2 Step 1: Clusters derivation of Pace et al., 2018 dataset by rCASC

The entire dataset of Pace et al., 2018 was analyzed with rCASC starting from the function *cellrangerCount* using Mus musculus genome assembly GRCm38 (mm10) genome as reference to generate the count table. With the pre-processing tools of rCASC, we have performed the following steps: inspecting the dataset with *genesUmi* and *mitoRibouMI* functions to plot

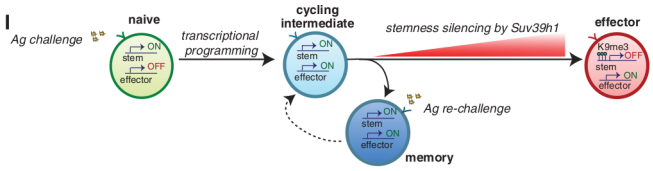


FIGURE 3.9: **Working model depicting the pivotal role of Suv39h1 during CD8 + T cell lineage differentiation and commitment.** After priming, cycling CD8+ T lymphocytes reprogram both self-renewing and effector gene expression profiles. Cycling cells may represent bipotent intermediates, which would then repress either the effector or stem cell/memory programs while they differentiate to memory precursors or effectors, respectively. The silencing of the stem cell/memory gene expression program is under the control of Suv39h1 by imposing the H3K9me3 modification on chromatin at the corresponding loci. (Figure from Pace et al., 2018)

TABLE 3.1: **Distribution of cells in each subset category from scRNA-seq experiment of Pace et al., 2018.**

Type of cell	WT mice	KO mice	Technical Replicate	Biological Replicate
WT Naive (N)	2426	-	-	-
KO Naive Suv39h1-defective (Nd)	-	2372	-	-
WT CD8+ T cells (NA)	-	-	3161	1005
KO CD8+ T cells Suv39h1-defective (NdA)	-	-	2293	661

the number of detected genes with respect to the number of UMI and to calculates the percentage of mitochondrial and ribosomal genes (Figure 3.10 and 3.11). WT naive (**N**) and KO naive (**Nd**) showed similar number of called genes (around 150-200 genes) with respect to the total number of UMIs/reads mapped on each cell in log10 format, while WT CD8+ T cells (**NA**) and KO Suv39h1-defective CD8+ T cells (**NdA**) showed higher number of called genes Figure 3.10. In Figure 3.11 the percentage of mitochondrial genes with respect to the percentage of ribosomal genes is plotted. WT naive (**N**) and KO naive (**Nd**) black and green cells were probably low informative cells composed of few genes and mainly ribosomal/mitochondrial. While WT CD8+ T cells (**NA**) and KO Suv39h1-defective CD8+ T cells (**NdA**) are also composed of cells with genes number included among 250 and 1000 (i.e. yellow and red cells). Moreover, we annotated the counts table and removing the ribosomal and mitochondrial genes with *scannobyGtf* function, selecting from the total genes (i.e. 27.785) the most variant 10K genes for the analysis. Thus, to summarize we removed all the cells with:

- more than 20% of mitochondrial genes
- less than 20% of ribosomal genes
- less than 100 called genes

From the 11.918 cells dataset, after filtering the following number of cells were present:

- WT naive (**N**): 2010
- KO naive (**Nd**): 1890
- WT CD8+ T cells (**NA**): 422
- KO Suv39h1-defective CD8+ T cells (**NdA**): 1636

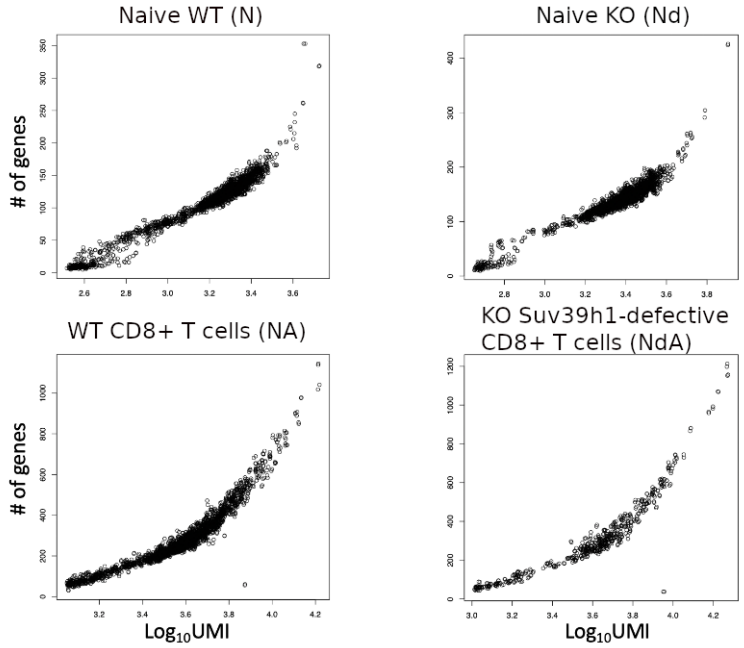


FIGURE 3.10: **Number of detected genes plotted for each cell with respect to the total number of UMI/reads in that cell.**

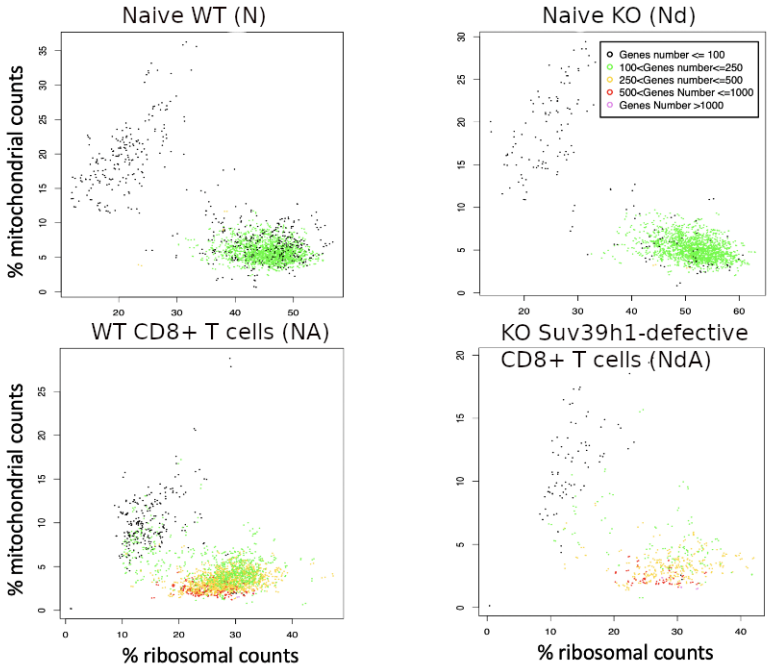


FIGURE 3.11: **Percentage of mitochondrial protein genes plotted with respect to percentage of ribosomal protein genes.** Cells are colored on the basis of total number detected genes.

From the 11.918 cells dataset, 400 cells for the four conditions were randomly selected to obtain a final dataset of 1600 cells. The choice of 400 number was dictate to uniform the entire dataset to the group with the minimum number of cells (i.e. WT CD8+ T cells (**NA**)). Moreover, for these cells we select the 10K most variant genes for the analysis and retained

the top 5000 expressed out of top variant genes.

For the clustering phase, we used *nClusterEvaluationSIMLR* function to detect a range of k number of clusters to be used for data partitioning (5-7 clusters). k=5 was the most represented data organization (Additional Figure 6.9). The range of k clusters detected using *nClusterEvaluationSIMLR* was then investigated with SIMLR. SIMLR was run for each k of the k-range defined with griph tool and we also evaluated the cell stability in each cluster with simlrBootstrap. 160 bootstraps were performed in which 10% of the cells was randomly removed from the initial data set. To understand how subgroups were organized in the various clusters and if there was any cluster showing asimmetry between WT CD8+ T cells (NA) and KO Suv39h1-defective CD8+ T cells (NdA), we investigated the cells composition of the different 5-7 partitions (Table 3.2, 3.3 and 3.4). Both in five clusters composition and in 6 clusters composition, there was a cluster where NA (WT) and NdA (KO) were in the same proportion (cluster 2 in 3.2, cluster 1 in 3.3). However, there was also one cluster in which NdA (KO) were more than NA (WT) (cluster 4 in 3.2, cluster 2 in 3.3) and one in which there was the opposite of the latter (cluster 5 in 3.2, cluster 4 in 3.3). The composition in 7 clusters showed totally different situation with naive cells (N and Nd) spread in three clusters with similar proportion

TABLE 3.2: **Distribution of cells in the 5 clusters**

Cluster	N cells (WT)	Nd cells (KO)	NA (WT)	NdA (KO)	Total
C1	249	218	2	0	469
C2	0	0	182	141	323
C3	150	182	7	0	339
C4	0	0	96	173	269
C5	1	0	113	86	200

TABLE 3.3: **Distribution of cells in the 6 clusters**

Cluster	N cells (WT)	Nd cells (KO)	NA (WT)	NdA (KO)	Total
C1	0	0	188	143	331
C2	0	0	96	172	268
C3	20	12	9	0	41
C4	1	0	107	84	192
C5	246	240	0	1	486
C6	133	148	0	0	281

TABLE 3.4: **Distribution of cells in the 7 clusters**

Cluster	N cells (WT)	Nd cells (KO)	NA (WT)	NdA (KO)	Total
C1	0	0	83	132	215
C2	0	0	120	120	240
C3	1	0	93	65	159
C4	20	6	17	13	56
C5	133	149	0	0	282
C6	246	245	0	0	491
C7	0	0	87	70	157

CSS violin plot (Figure 3.12) shows that the mean stability for k=5 (CSS ~0.83) is slightly higher than the others ks with CSS ~0.81 and ~0.82, respectively. Thus, Clusters k=6 and k=7 did not represent the most stable organizations in terms of CSS. Moreover, even if their CSS were near k=5, in any case they were not the most frequent organizations observed in *nClusterEvaluationSIMLR* analysis (Additional Figure 6.9 in Supplementary Material 6).

Since the best CSS is observed in k=5, we explored these clusters. Referring to the Table 3.2, all the clusters are quite homogeneous with cluster 1 composed of Naive cells equally

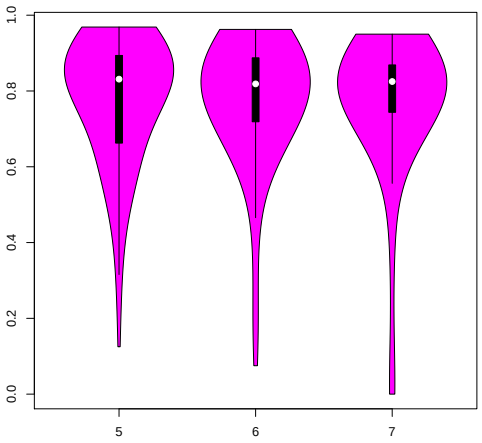


FIGURE 3.12: **Cell stability score detected by simlrBootstrap.** The mean stability for k=5 (CSS 0.83) is higher than k=6 and k=7 with CSS 0.81 and 0.82, respectively.

distributed among WT and KO (53% and 46% respectively); clusters 2,4 and 5 are composed of WT and KO CD8+ T cells (56% and 44% for cluster 2, 36% and 64% for cluster 4, 56% and 43% for cluster 5 respectively). However, cluster 3 is mixed containing 44% Naive WT and 54% KO but also 2% CD8+ T cells WT (NA).
In Figure 3.13 clusters 1, 2, 4 and 5 show a quite good stability, since cells stay in these clusters between 75 to 100% of the bootstraps. While, cluster 3 is characterized by a lower CSS (50-75%) and it is possible that this cluster contains cells localized at the boundaries of clusters 1 (Figure 3.13). Figure 3.14 reports in a different graphical way the stability of the single cells, coloured on the basis of their Stability Score and on cellular labels type. Finally, the plot of Figure 3.15, shows the genes detectable in each cell in function of the total number of reads/cell. In this plot cells are colored with the same color of their belonging cluster.

This plot is useful to observe if the clustering is biased by the number of genes called in each cluster. In this specific example, only the violet cluster is characterized by a number of detected genes larger of those detectable in the other clusters. To detect the genes playing the major role in clusters formation, we used *anovaLike* function (*adjusted p-value* < 0.05; *logFC threshold: 0.5*) and we identified a total of 583 differentially expressed genes. As reported in Table 3.2, cluster 1 and 3 were mainly composed of cells already labeled from sequencing as Naive WT (N) and KO (Nd) and a total of 225 DEGs for the two clusters were identified. Then, the analysis of DEGs was focused on the remaining clusters (cluster 2,4 and 5) classified overall as WT CD8+ T cells (NA) and KO Suv39h1-defective CD8+ T cells (NdA). In detail, we compared cluster 2 and cluster 4, cluster 2 and cluster 5, cluster 4 and cluster 5 and select differentially expressed genes (DEGs) for each comparison.

As reported in Figure 3.16, cluster 2 vs cluster 4 showed two DEGs (Gpx4, Cks1b), where Cks1b is unique and has function in cell cycle (Figure 3.16, blue circle). Gpx4 is in common among the three comparison and is also known as glutathione peroxidase 4, which catalyze the reduction of hydrogen peroxide, organic hydroperoxides and lipid hydroperoxides, and thereby protect cells against oxidative damage. Cluster 5 vs cluster 2 showed in total 255 DEGs with 31 genes unique (Figure 3.16, red circle). Enrichment analysis of the 255 DEGS revelead that these genes are involved in DNA metabolic process, DNA replication, G1/S transition of mitotic cell cycle, DNA-dependent DNA replication and nucleotide-excision

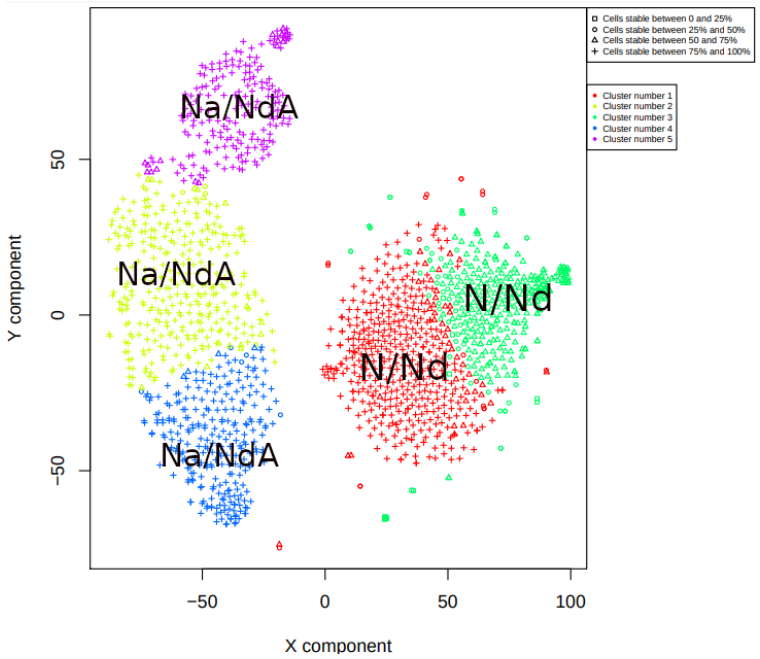


FIGURE 3.13: **CSS of the 5 clusters obtained with SIMLR. N**: WT naive; **Nd**: KO naive; **NA**: WT CD8+ T cells; **NdA**: KO Suv39h1-defective CD8+ T cells. Clusters 1, 2, 4 and 5 show a quite good stability (75 to 100% of the bootstraps). While, cluster 3 is characterized by a lower CSS (50-75%).

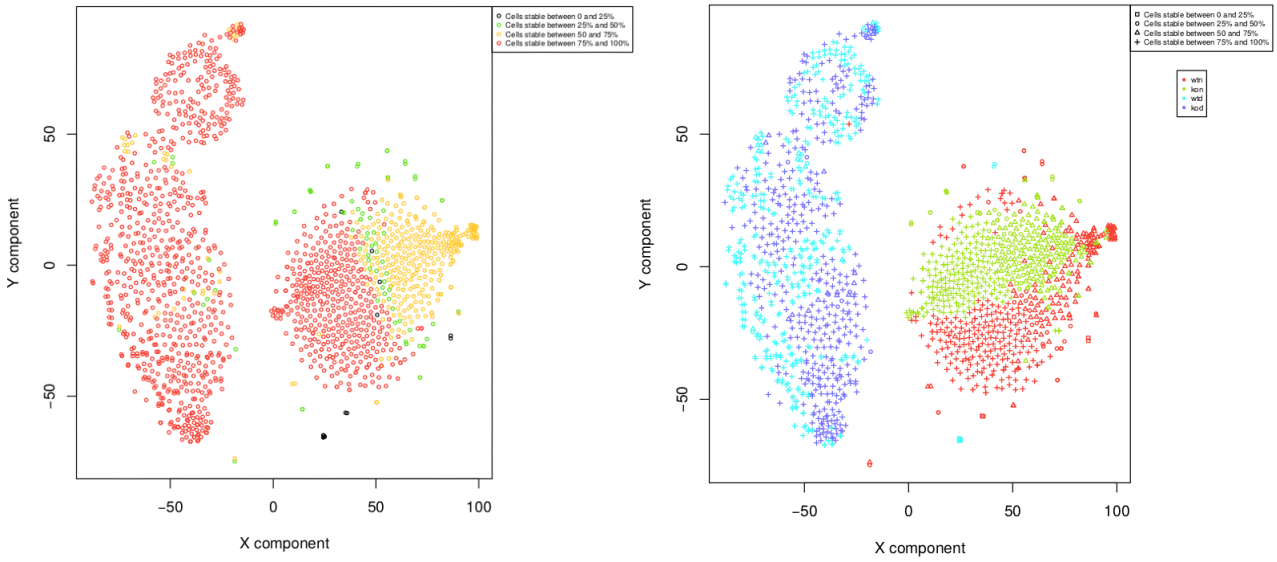


FIGURE 3.14: **Cells are colored on the basis of their Stability Score (left) and cellular types (right).**

repair (Figure 3.17).

Finally, the comparison among cluster 4 and cluster 5 identified 326 DEGs of which one, Gpx4, in common with all and 223 in common with cluster 5 vs 4 and 102 unique genes (Figure 3.16, green circle and Figure 3.18). Also for this comparison, enrichment analysis revealed genes implicated in DNA processing and cell cycle.

Finally, we compared the differential expression for cluster 5 vs 2 and cluster 5 vs 4.

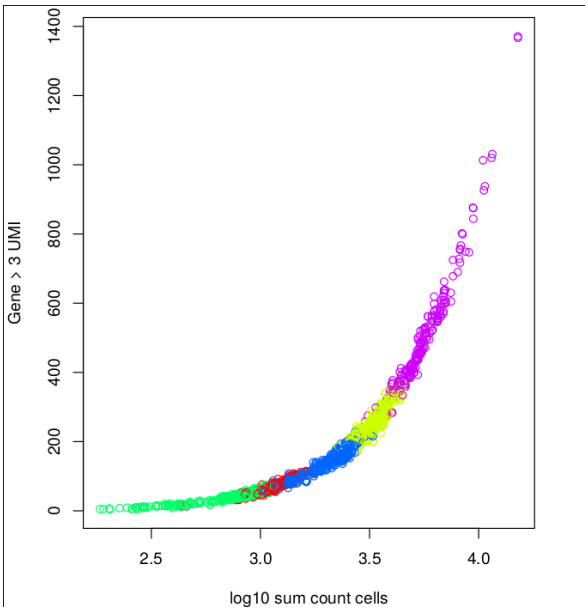


FIGURE 3.15: **Genes versus UMI counts.** The plot shows the genes detectable in each cell in function of the total number of reads/cell. The cells are colored on the basis of their belonging cluster.

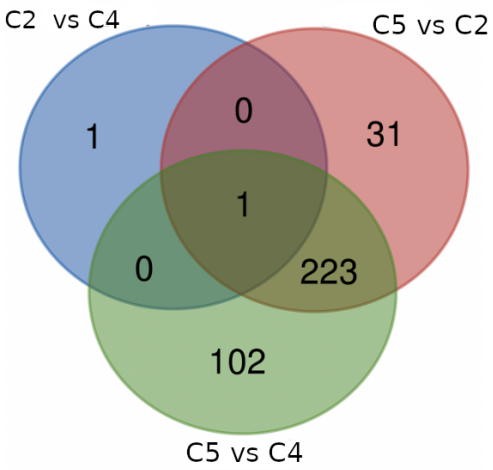


FIGURE 3.16: **Comparison of DEGs among cluster 2,4 and 5.** Blue circle: comparison of cluster 2 vs 4. Red circle: comparison of cluster 5 vs 2. Green circle: comparison cluster 5 vs 4. Intersections among circles indicate the DEGs shared among the comparison.

As reported in Figure 3.19, cluster 2 and 4 are very similar since the two clusters linearly correlate with their differential expression with respect to cluster 5.

3.6.3 Step 2: Gene signatures identification of Pace et al. dataset by rMLSC

The results of rCASC pipeline (i.e. association cells-clusters) were used as input for the machine learning analysis based on classification algorithms (Witten et al., 2016).

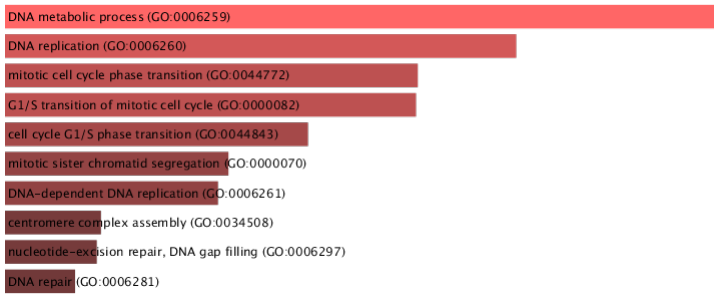


FIGURE 3.17: **Enrichment analysis of the 255 DEGs in comparison cluster 5 vs cluster 2**

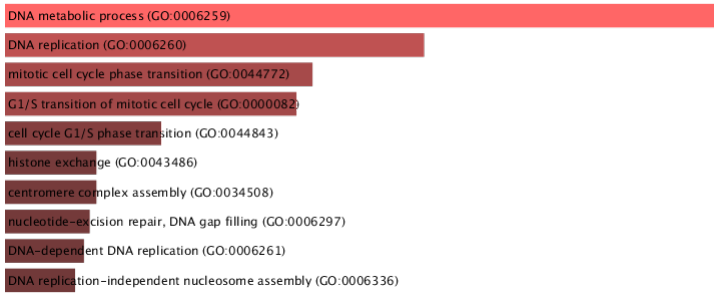


FIGURE 3.18: **Enrichment analysis of the 326 DEGs in comparison cluster 5 vs cluster 4**

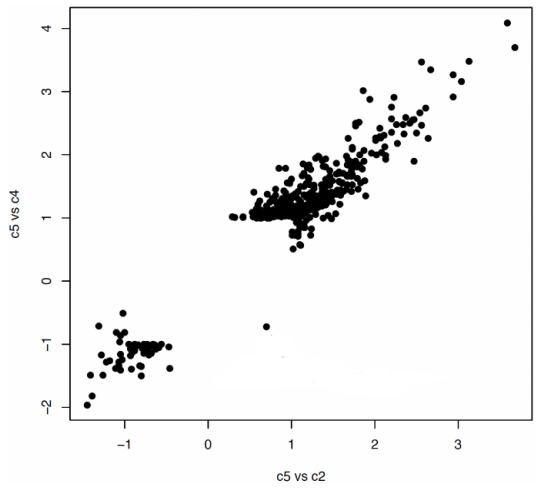


FIGURE 3.19: **Comparison of the differential expression for cluster 5 vs cluster 2 and cluster 5 vs cluster 4.** Cluster 2 and 4 are very similar since the two clusters linearly correlate with their differential expression with respect to cluster 5.

For the Phase 1 (i.e. input data pre-processing phase) we used *FromCountMatrixtoARFF* function to generate the ARFF files divided in Training and Test, with 70% and 30% of the cells randomly chosen, respectively. In Table 3.5 are reported the composition in terms of number of cells of the two input files (Training and Test). For cluster 1, Training was composed of 328 cells of cluster 1 and 792 cells of cluster x (i.e. merge of all the other clusters), representing 70% of instances, respectively. While the Test was composed of 141 and 339 cells belonging to cluster 1 and cluster x, respectively. The final number of cells was 1120 and 480 cells for Training file and Test, respectively. Respecting the same criteria, we

generated 600 Training and 600 Test files for the other clusters (i.e. cluster 2,3,4,5). The final number of cells was 1050 and 450 cells for Training file and Test of cluster 2, 1120 and 480 cells for Training file and Test of cluster 3, 4 and 5, respectively.

TABLE 3.5: **Composition of Training and Test files**

Cluster	Composition	Training (70%)	Test (30%)
C1	C1: 469	328	141
	CX=C2+C3+C4+C5: 1131	792	339
	Total	1120	480
C2	C2: 323	226	97
	CX=C1+C3+C4+C5: 1177	824	353
	Total	1050	450
C3	C3: 339	237	102
	CX=C1+C2+C4+C5: 1261	883	378
	Total	1120	480
C4	C4: 269	188	81
	CX=C1+C2+C3+C5: 1331	932	399
	Total	1120	480
C5	C5: 200	140	60
	CX=C1+C2+C3+C4: 1400	980	420
	Total	1120	480

For the Phase 2 (i.e. machine learning classification analysis) we used *WekaCaller* to perform the classification analysis using Random Forest algorithm performing 600 iterations for each cluster and using default parameters.
After the classification analysis we explored the results in the Phase 3 through *ClassifierInvestigator* tool that is able to summarize the results of the classification for all the clusters.

TABLE 3.6: **Classification performance of Random Forest on dataset of Pace et al., 2018**

Cluster	Number of cells in the cluster	Number of cells in Test set	Average Accuracy
C1	469	141	71%
C2	323	97	74%
C3	339	102	64%
C4	269	81	75%
C5	200	60	89%

The overall performance of the classifier was good in all the clusters (Table 3.6). The maximum percentage of correct classification (89%) was obtained for cluster 5 with 32.040 out of 36.000 cells correctly classified in all the 600 runs. In cluster 4 36.450 out of 48.600 cells (75%), in cluster 2 43.068 out of 58.200 cells (74%) and in cluster 1 60.066 out of 84.600 (71%) were correctly classified. The lower performance of classification regards the cluster 3 with 39.168 out of 61.200 (64%) correctly classified. Remarkable, cluster 3 was the only cluster with a CSS ranging from 50% and 75% (Figure 3.13).

For the Phase 4, *TreesExplorer* tool was used to extract and parse the trees generated from phase 2. In Figure 3.20 is reported the plots of entropy calculated by *TreesExplorer* in the 600 trees of each of the 5 clusters. In all the clusters the entropy decreases as the number of levels increases due to the continuous splitting of the dataset in partitions such that each partition contains the highest number of cells belonging to the same cluster. Considering all the plots in Figure 3.20, the first levels contain the genes whose expression values allow us to

partition the highest number of cells in the same cluster. Then, the coloured traces reach an elbow point where entropy is near the minimum value meaning that in these levels the further splitting of partitions is due to genes whose information to cell-cluster classification is lower.

In Table 3.7 is reported, for each cluster, the total number of levels available and the total number of unique genes retrieved from the performed 600 runs. Despite the number of levels among cluster 1 and 3 is quite near (i.e. 40 and 31 levels, see also Figure 3.20), cluster 2, 4 and 5 show instead a decrease in the number of levels, 25, 29 and 14 respectively (Table 3.7 and Figure 3.20. Concerning the number of genes, cluster 1 to 4 display on average 3187 unique genes. Cluster 5 in particular shows the smallest number of levels (i.e. 14) and unique genes (i.e. 1881).

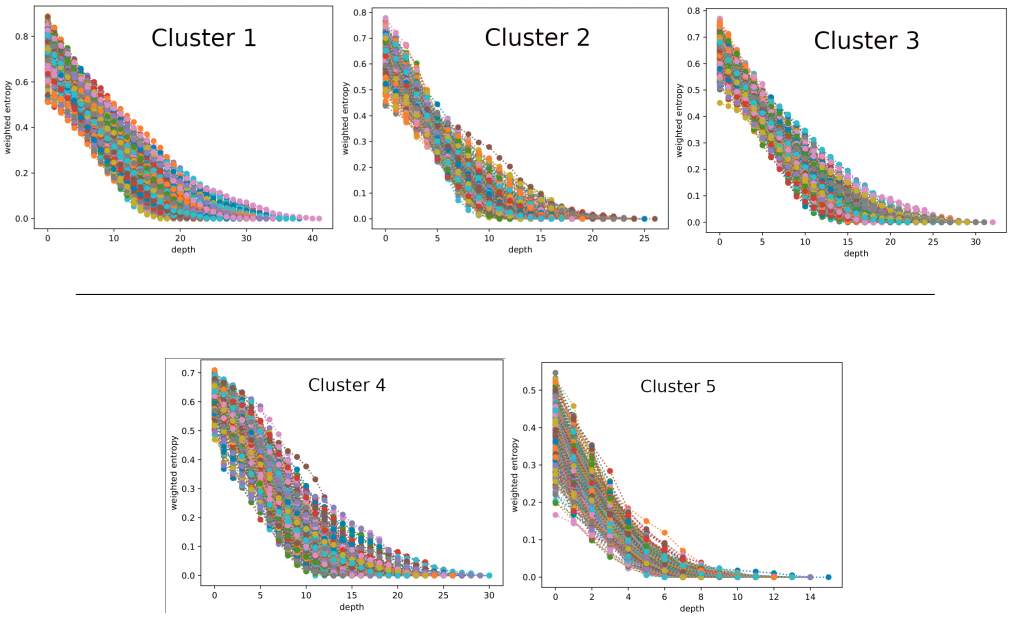


FIGURE 3.20: **Entropy plots for all the clusters.** Trees of the cluster are represented by 600 traces coloured with different colours. x axis represents the depth (i.e. levels) reached by each tree and y axis represent the weighted entropy.

TABLE 3.7: **Number of levels and genes retrieved from the 600 trees in each cluster**

Cluster	Number of levels	Number of genes
C1	40	3484
C2	25	3049
C3	31	3092
C4	29	3126
C5	14	1881

Last phase of the rMLSC workflow is dedicated to find the final set of gene signatures. To find signatures most informative as possible, we used *SignatureDiscover* function to select the *Entropy threshold* that allow us to choose how many levels of the trees (and the consequent number of genes) retain in the further analysis. Thus, we plotted the median of the entropy values for each level in each cluster (Figure 3.21). To adapt the chose of the *Entropy threshold* to all the clusters, we set *Entropy threshold* equal to 0.1 weighted entropy (Filter 1), since this is the maximum entropy value beyond which considering more gene levels results in any further useful information to derive the gene signatures. The number of genes and levels retained after *Entropy threshold* is reported in Table 3.8, Filter 1. The maximum decrease regarding levels affected cluster 1 (i.e. from 40 to 16 levels), while regarding genes, 456 genes were filtered out from cluster 5. The minimum number of genes filtered out is 242 genes of cluster 4, while only 11 levels of cluster 5 were removed (Table 3.8, Filter 1). Overall for all clusters, the number of levels is decreased less than half, while the number of genes is decreased on average of 300 genes (Table 3.8, Filter 1).

SignatureDiscover also calculates, for each gene present in the levels of the trees of clusters, the percentage of cells expressing that gene in the cluster under analysis. This allow the user to understand the proportion of genes retrieved from the levels of the tree that are

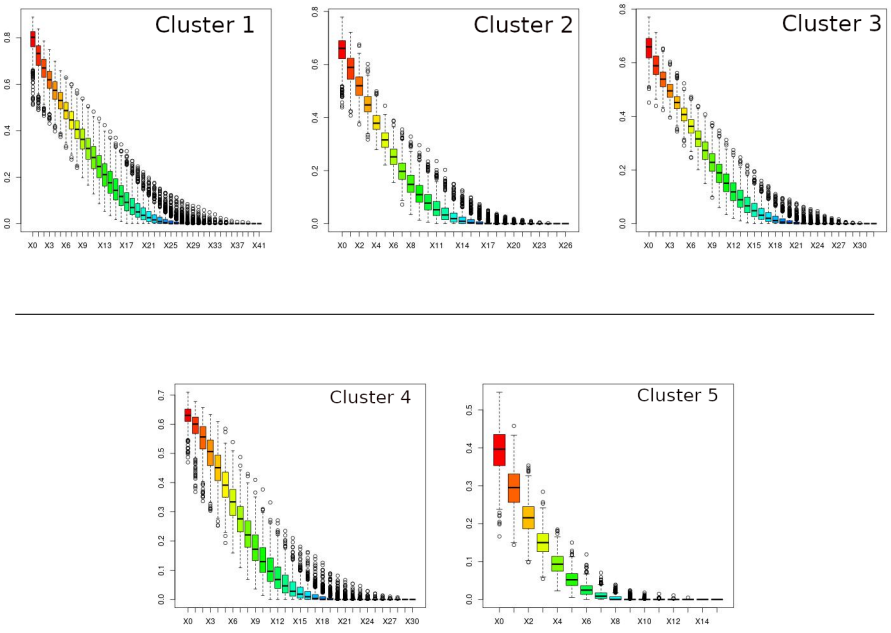


FIGURE 3.21: **Boxplots for clusters 1,2,3,4 and 5.** x axis represents the depth (i.e. levels) reached by each tree and y axis represent the weighted entropy. Each boxplot refers to the entropy of the single level.

actually expressed in the cells of the clusters ranging from 0 to 100%. In this case, we considered only the genes expressed in at least 70% of cells of the belonging cluster in order to obtain informative signatures (Filter 2). Applying this filter allow a further skimming of the genes in the signatures with a consistent reduction. As it is shown in Table 3.8, with the second filter we reduced further the number of genes to the final signatures composed of 15 genes for cluster 1, 71 for cluster 2, 9 for cluster 3, 41 for cluster 4 and 151 for cluster 5. Table 3.9 shows the gene names of the signatures associated with each cluster.

TABLE 3.8: **Number of levels and genes after the *EntropyCutting* (Filter 1) and gene presence in the clusters (Filter 2)**

	Original number		Filter 1		Filter 2
Cluster	Levels	Genes	Levels	Genes	Genes
C1	40	3484	16	3187	15
C2	25	3049	9	2767	71
C3	31	3092	12	2755	9
C4	29	3126	11	2884	41
C5	14	1881	3	1425	151

TABLE 3.9: **Gene signatures for all the clusters**

Cluster	Number of genes	Gene signature									
C1	15	Actb	B2m	Cd52	Cd8b1	Eef1a1	Eef2	Fau	H2-D1	mt-Cytb	mt-Nd1
		mt-Nd4	Rack1	Tmsb10	Tmsb4x	Tpt1					
C2	71	Abracl	Actb	Actr3	Arhgdib	Arpc1b	Arpc2	Atp5h	B2m	Btf3	Calm1
		Ccl5	Cd3d	Cd3g	Cd52	Cd8b1	Cdc42	Cfl1	Clic1	Coro1a	Cox4i1
		Cox8a	Crip1	Ctsd	Cyba	Eef1a1	Eef2	Eif1	Fau	Fth1	Ftl1
		Fxyd5	Gzmb	H2-D1	H2-K1	H3f3b	Hspa8	Itm2b	Laptm5	Lck	Lgals1
		Lsp1	Ms4a4b	Ms4a6b	mt-Cytb	mt-Nd1	mt-Nd4	Myl12a	Myl6	Naca	Nkg7
		Oaz1	Pfn1	Ppia	Ppp1ca	Ptprcap	Pycard	Rac2	Rack1	S100a10	S100a4
		S100a6	Selplg	Sh3bgrl3	Shisa5	Slc25a5	Sub1	Tmsb4x	Tpt1	Uba52	Ubb
		Uqcrh									
C3	9	Actb	B2m	Eef1a1	Fau	mt-Cytb	mt-Nd1	Tmsb10	Tmsb4x	Tpt1	
C4	41	Actb	Arhgdib	Arpc1b	Arpc2	Atp5h	B2m	Calm1	Ccl5	Cd3d	Cd52
		Cd8b1	Cfl1	Coro1a	Cox4i1	Crip1	Eef1a1	Eef2	Eif1	Fau	H2-D1
		H2-K1	Hspa8	Itm2b	Lgals1	Lsp1	Ms4a4b	mt-Cytb	mt-Nd1	Myl6	Naca
		Nkg7	Rac2	Rack1	S100a10	S100a6	Sh3bgrl3	Tmsb10	Tmsb4x	Tpt1	Uba52
		Ubb									
C5	151	Abracl	Actb	Actr3	Anxa2	Anxa6	Arhgdib	Arpc1b	Arpc2	Arpc3	Arpc5
		Atp5b	Atp5e	Atp5f1	Atp5h	Atp5j	Atp5j2	Atp5mpl	B2m	Btf3	Calm1
		Capzb	Ccl5	Cd3d	Cd3e	Cd3g	Cd47	Cd48	Cd52	Cd8b1	Cdc42
		Cfl1	Chchd2	Clic1	Coro1a	Cotl1	Cox4i1	Cox5a	Cox5b	Cox6a1	Cox6b1
		Cox6c	Cox7a2	Cox7c	Cox8a	Crip1	Ctsd	Cyba	Ddx5	Dnajc15	Eef1a1
		Eef1b2	Eef1g	Eef2	Eif1	Eif3h	Eif5a	Elob	Emp3	Eno1	Epsti1
		Fau	Fth1	Ftl1	Fxyd5	Gapdh	Gimap4	Gmfg	Gzma	Gzmb	H2-D1
		H2-K1	H2-Q7	H2afz	H3f3b	Hint1	Hmgb2	Hnrnpa2b1	Hnrnpf	Hnrnpk	Hsp90ab1
		Hspa8	Itm2b	Krtcap	Laptm5	Lat	Lck	Ldha	Lgals	Lgals3	Lsp1
		Ms4a4b	mt-Cytb	mt-Nd1	mt-Nd2	mt-Nd4	Myl12a	Myl6	Naca	Ndufa4	Ndufb7
		Nkg7	Npm1	Oaz1	Ostf1	Pabpc1	Pclaf	Pfn1	Ppia	Ppib	Ppp1ca
		Prdx1	Psmb1	Psmb8	Ptma	Ptprc	Ptprcap	Pycard	Rac2	Rack1	Ran
		S100a10	S100a11	S100a13	S100a4	S100a6	Sec61b	Selplg	Sem1	Sept1	Serbp1
		Serf2	Sh3bgrl3	Slc25a5	Snrpe	Snrpg	Stmn1	Sub1	Tagln2	Thy1	Tmsb10
		Tmsb4x	Tpm3	Tpt1	Tspo	Tubb5	Txn1	Uba52	Ubb	Ubl5	Uqcrh
		Vim									

Exploration of gene signatures The machine learning approach (i.e. rMLSC) allowed us to derive for each cluster a set of genes that together should better characterize the cells grouped in one cluster. We user Enrichr web tool (Kuleshov et al., 2016) to perform an enrichment analysis of the gene signatures characterizing each set for molecular function (Figure 3.22).As showed in Table 3.9, cluster 1 and 3 are both composed of cells previously labeled as Naive cells. Thus, as expected, they show the same genes in the signatures even if with different numerousness (cluster 1: 15 genes, cluster 3: 9 genes). Genes of the signature are enriched for metabolism activity that mainly involve NADH dehydrogenase activity (mt-Nd4, mt-Nd1) and for cytoskeleton activity with actin involvement (Tmsb4x, Tmsb10, Actb). However, the majority of genes are enriched for RNA and protein kinase binding (Eef1a1, Tmsb4x, Rack1, Fau, Eef2,Tpt1, Actb) (Figure 3.22). The genes of cluster 2 signatures (71 genes) are enriched for immune function related to T cell receptor binding and CD4 receptor binding (Lck, Cd3g). However, also genes involved in cytoskeleton activity are present as actin filament and monomer binding (Actr3, Arpc2, Arpc1b, Tmsb4x, Cfl1, Abracl, Lsp1, Fxyd5, Coro1a) and tropomyosin binding (Pycard, S100a6) and genes of apoptotic process (Pycard, Rack1). Signature of cluster 4 contains 41 genes enriched for actin monomer binding (Arpc2, Arpc1b, Tmsb4x, Cfl1, Lsp1, Coro1a, Tmsb10), calcium signaling and (insitol trisphosphate kinase activity, Calm1) and phospholipid metabolism (phospholipase activator, lipase activator, GDP activity). Finally, genes of the cluster 5 signature (151 genes) are enriched, as cluster 2, for function in T cell receptor binding (Lck, Cd3g, Cd3e), functions in chain elongation during polypeptide synthesis at the ribosome (Eif5a,Eef1a1, Eef1b2,Eef1g). As for cluster 1 and 3, genes of cluster 5 signature are also involved in NADH deidrogenase activity (mt-Nd4, Ndufb7, Ndufa4, mt-Nd2, mt-Nd1). Overall,the enrichment analysis revealed that signatures are enriched mainly for basic cells activity spanning from cytoskeleton to metabolism. However, cluster 2 and 5 genes also shared function related to immune system, in particular to T cell receptor activity.

FIGURE 3.22: **Enrichment analysis of gene signatures for molecular function.**

To better explore the gene signatures and discover how many genes were unique or shared among the clusters, we generate an upset plot shown in Figure 3.23. Each column corresponds to a set, and each row corresponds to one cluster. Cells are either empty (light-gray), indicating that this set is not part of that intersection, or filled, showing that the set is participating in the intersection.

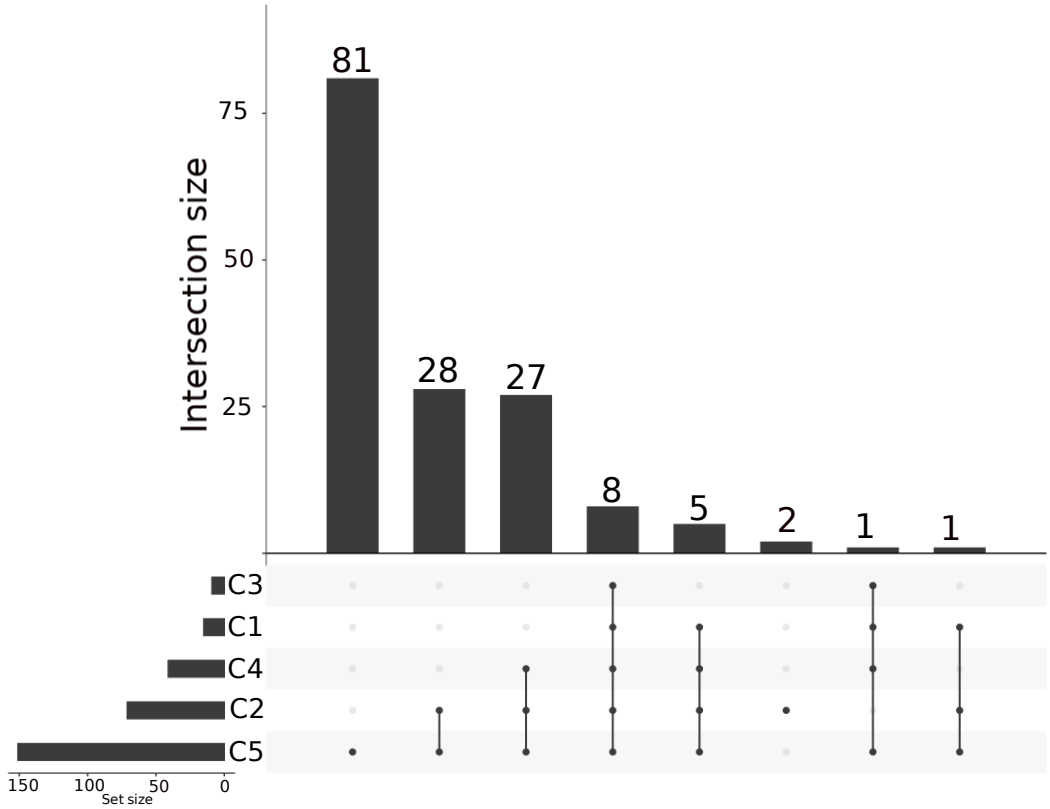


FIGURE 3.23: **Upset plot of the 5 gene signatures.** At the top of each bar is reported the number of genes unique/shared of the clusters. In the Bottom left corner it is reported the size of signature. Black dots indicate the cluster among which genes are shared.

The first row in the figure corresponds to cluster 3 that shared 8 genes (Actb,B2m, Eef1a1, Fau, mt-Cytb, mt-Nd1, Tmsb4x, Tpt1) with all the clusters (Figure 3.23 and Table 3.10). Gene enrichment analysis and literature research of this set of genes revealed association with cellular metabolism and regulation of cells cytoskeleton,functions that are conserved in all the cell types. Cluster 3 shared also one gene, Tmsb10, with cluster 1,4 and 5 that plays an important role in the organization of the cytoskeleton (Figure 3.23 and Table 3.10). Cluster 1 shares 5 genes (Cd52,Cd8b1,Eef2,H2-D1,Rack1) with cluster 2, 4 and 5 (Figure 3.23 and Table 3.10). Cd8b1 and H2-D1 are genes involved in the mechanism of T cell receptor interaction with antigen-bearing MHC Class I and antigen processing and presentation mechanism (i.e. typical functions of Naive T cells that are considered mature but , unlike activated or memory T cells, have not encountered its cognate antigen within the periphery). While Cd52, Eef2 and Rack1 are genes coding for proteins related to synthesis and post-translational modification of proteins. Cluster 1 shared also one mytochondrial gene, mt-Nd4, with cluster 2 and 5. Cluster 2 had two unique genes, Shisa5 and Ms4a6b (Figure 3.23 and Table 3.10). The first plays a role in p53/TP53-dependent apoptosis, while the second may be involved in signal transduction as a component of a multimeric receptor complex. Moreover, cluster 2 shares 27 genes with clusters 4 and 5 and other 28 genes only with cluster 5 (Figure 3.23 and Table 3.10). The set of 27 genes is enriched for Prostaglandin Synthesis and Regulation (S100A6 S100A10 genes), Macrophage markers (RAC2 gene), Microglia Pathogen Phagocytosis Pathway (ARPC1B, RAC2 genes), EBV signaling (Ccl5 gene), Fc

gamma R-mediated phagocytosis (ARPC2, ARPC1B, CFL1, RAC2 genes), Chemokine signaling pathway (CCL5 and RAC2 genes) and G13 Signaling Pathway (i.e. G13 regulates actin cytoskeletal remodeling)(ARHGDIB,CFL1,CALM1 genes). The first mechanisms are functions related to active immune cells whose function is immuno-surveillance and defense against pathogens while the latter is more related to cell migration functions. Finally, cluster 2 shared 28 genes only with cluster 5 (Figure 3.23 and Table 3.10). Looking more inside this set revealed gene functions related to IL-3 Signaling Pathway (CDC42 and LCK genes), Oxidative Stress (CYBA gene), Inflammatory Response Pathway (LCK gene), p38 MAPK Signaling Pathway (CDC42 gene), Microglia Pathogen Phagocytosis Pathway (CYBA gene) but also Necroptosis (PYCARD, FTL1, FTH1, SLC25A5, PPIA genes), Salmonella infection (CDC42,PYCARD,PFN1 genes), Ferroptosis (FTL1, FTH1 genes), T cell receptor signaling pathway (CDC42, LCK, CD3G genes) and Leukocyte transendothelial migration (CDC42, CYBA, MYL12A genes). Finally cluster 5 shows 81 unique genes in the signature that are related to T-cell Apoptosis (PTPRC,CD3E,LAT genes), Granzyme A mediated Apoptosis (GZMA,HMGB2 genes), Initiation of TCR Activation (PTPRC,CD3E genes) and Antigen Processing and Presentation (PSMB8 gene)(Figure 3.23 and Table 3.10).

We then plotted the expression values of gene signatures on the basis of the results obtained by the upset plot. Figure 3.24 shows the heatmap in which columns represent cells grouped on the belonging cluster (cluster 1: red, cluster 2: yellow, cluster 3: green, cluster 4: blue, cluster 5 violet) while rows contain genes grouped in 8 blocks (i.e. see Figure 3.23 and Table 3.10) with different color-label.

The first block in green water is composed of the 81 genes unique of cluster 5. Then, the second pink block is composed of the 28 genes shared between cluster 2 and 5. The third in blue contains 27 genes shared among cluster 4,2 and 5. Then, the 8 genes shared among all clusters have orange label. Yellow block is composed of 5 genes shared among cluster 1,4,2 and 5. Then the two genes unique of cluster two have light Blue label. Finally, green label belongs to one gene shared among cluster 1,3,4 and 5 and red label to one gene shared among cluster 1,2 and 5.

Starting from the first gene, mt-Nd4 (i.e. bottom of the Heatmap, in red, contained in the signature of cluster 1,2 and 5) shows high expression level in all the cells of clusters 5 and lower expression in cells of cluster 1 and 2. Tmsb10 gene (green block, shared among all cluster except for cluster 2) expression is opposite inside the cells of cluster 1,3,4 and 5. Indeed, cluster 1 show about 60% of cells with lowest and about 40% with the highest expression. The same proportion is present also in the other clusters. Further analysis showed that the cells with high expression of Tmsb10 are KO cells for Suv39h1 reflecting the possibility that the function of this gene (i.e. organization of the cytoskeleton) is request when Suv39h1 is silenced. Ms4a6b and Shisa5 (ligh blue block) are genes present uniquely in cluster 2 signature and they show in it the highest expression with respect to other clusters. Yellow block is composed of 5 genes (Cd52,Cd8b1,Eef2,H2-D1,Rack1) with heterogeneous expression among clusters with highest levels in cells of cluster 4 and 5 cells with respect to 1 and 3. Orange block is composed of 8 genes shared among the 5 clusters (Actb,B2m, Eef1a1, Fau, mt-Cytb, mt-Nd1, Tmsb4x, Tpt1). mt-Cytb, mt-Nd1, Tpt1, and Eef1a1 are overexpressed in cluster 1,2 and 5 while cluster 3 and 4 show lowest expression. Particularly for cluster 3, KO cells show higher expression with respect to WT cells. This is probably due to a major involvement of these genes in absence of Suv39h1 gene. For the remaining four genes (Actb,B2m, Fau and Tmsb4x) the situation is opposite. Indeed, cluster 1 and 3 show lower expression while cluster 2, 4 and 5 highest. However, also in this case, KO cells of cluster 1 and 3 show high expression of Fau gene with respect to WT cells. Blue and violet blocks contain the 27 and 28 genes shared among cluster 2,4,5 and cluster 2 and 5, respectively and showing similar behaviour. Indeed, genes of the two blocks are overexpressed in cluster 2, 4

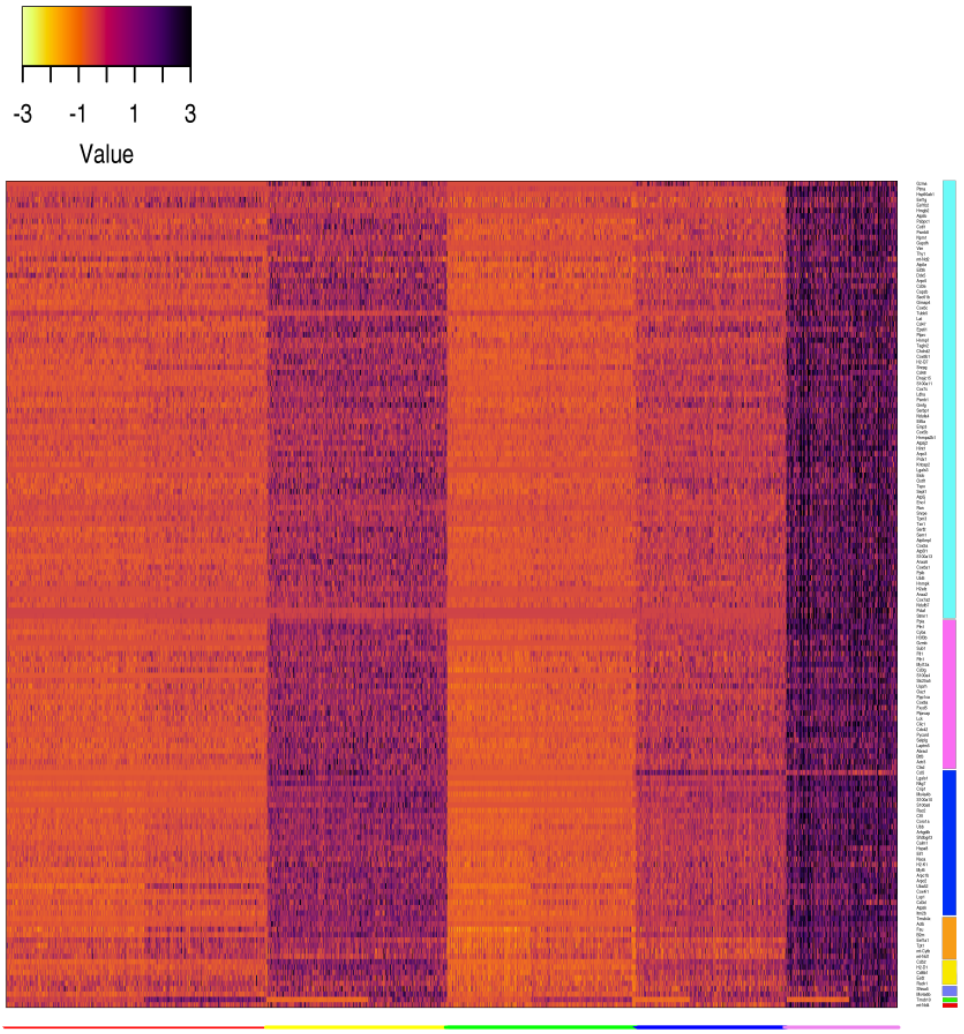


FIGURE 3.24: **Heatmap of gene signatures based on Upset plot.** Rows representing genes organized as Upset plot results. Green water: 81 genes unique of cluster 5. Pink: 28 genes shared between cluster 2 and 5. Blue: 27 genes shared among cluster 4,2 and 5. Orange: 8 genes shared among all clusters. Yellow: 5 genes shared among cluster 1,4,2 and 5. Light Blue: Two genes unique of cluster two. Green: one gene shared among cluster 3,1,4 and 5. Red: one gene shared among cluster 1,2 and 5. Columns represent clusters: Cluster 1-red; Cluster 2-yellow; Cluster 3-green; Cluster 4-blue; Cluster 5: violet.

and 5 while lower expression is detected in cluster 1 and 3. However, Uba52 gene belonging to 27 genes (blue block) also shows a higher expression in KO cells of in cluster 1 and 3 with respect to WT. In the pink block composed of 28 genes common in cluster 2 and 5, Ccl5 gene has a particular role in differentiating the cluster for its expression. This gene is is one of several chemokine genes clustered on the q-arm of chromosome 17. Chemokines form a superfamily of secreted proteins involved in immunoregulatory and inflammatory processes. Cluster 1 and 3 cells as well as cluster 5 cells (both WT and KO) show lowest expression of Ccl5 gene while cluster 2 and 4 show overexpression. Finally, 81 genes with green water label color are unique of cluster 5 and this is particularly reflected in their highest expression in these cells. On the opposite part, cluster 1 and 3 show overall lower expression while cluster

2 and 4 are more similar to cluster 5. Despite some differences in the expression of single genes among WT and KO cells of the same cluster, the overall expression behaviour of the 8 gene blocks (81, 28,27,8,5,2,1,1) is concordant with the previous analysis where we showed that cells of cluster 2,4 and 5 are mature immune cell types while cluster 1 and 3 gene signatures displayed functions more related to naive immune cells. Indeed, cluster 1 (red) and 3 (green) show overall a lower expression of all the sets of genes, followed by cluster 4 (blue), cluster 2 (yellow) and finally cluster 5 (violet).

To summarize and understand how the gene signatures are associated with the final cellular types, we also performed an enrichment analysis using Enrichr of the set of gene signatures (see Table 3.9). The results are shown in Figure 3.25. Cluster 1 and 3 contains the same genes in the signature however the first has 6 genes more. Cluster 2, 4 and 5 have the most robust set of gene signatures. Enrichr analysis showed that genes of the cluster 1 and 3 (i.e. Eef1A1, Cd52, mt-ND4, Tmsb4X, Rack1, Fau, B2m, Tmsb10, Actb, Tpt1 and mt-Nd1 genes) are particularly enriched for blood dendritic cells (i.e. antigen-presenting cells that initiate and direct adaptive immune responses), CD34+ cell (i.e. Hematopoietic stem cells with CD34 marker on the surface) and T lymphocytes (See also Supplementary Materials, Figure 6.10 and 6.12). Cluster 2 gene signatures is enriched for CD4+ thymocyte (i.e. Arpc2, Lck, Cd3g, Itm2b genes), CD8+ T lymphocytes (i.e. Ms4a6b, Ms4a6b, Ms4a4b genes) and mast cells (i.e. H3f3b, Rac2, Laptm5, Gzmb, Fxyd5 genes) (See also Supplementary Materials, Figure 6.11). Cluster 4 genes are enriched for CD8+ T lymphocytes (i.e. CD8B1, MS4A4B genes) that are cytotoxic T cells recognising peptides presented by MHC Class I molecules and kill infected or malignant cells (See also Supplementary Materials, Figure 6.13). Finally, cluster 5 genes are enriched as cluster 2 for CD4+ thymocyte (i.e. ARPC2, Lck, Gimap4, CD3G, CD3E, Itm2b genes), B-cells positive for GL7 activation antigen (i.e. Btf3, Cox8A, Ddx5, H3f3b, Ostf1, Cox5B, Cox7C) and NK cells (i.e. Ndufb7, Atp5J, Atp5H, Atp5F1, Cox5B, Cox6C, Cox5A, Atp5J2, Cox6B1) cells of the innate immune system that respond quickly to a wide variety of pathological challenges (See also Supplementary Materials, Figure 6.14). From the last analysis we can conclude that each signature contains genes strictly related to immune phenotype. Indeed, cluster 1 and 3 contain cells mainly developing from stem cells originating in the bulk tissue (i.e. bone marrow) and differentiating in central lymphoid tissues in an antigen-independent manner. Thus, cluster 1 and 3 cells are the most probable precursors of cells contained in cluster 2,4 and 5. Indeed for the successive antigen-dependent differentiation, cells undergo do positive/negative selection and migrate into peripheral lymphoid tissues (i.e. lymph nodes, spleen, and mucosa-associated lymphoid tissues) at sites where these cells can react with antigen and differentiate in lymphocytes T CD4+ or CD8+ (cluster 2,4,5).

FIGURE 3.25: **Enrichment analysis of gene signatures for cellular type.**

3.7 Discussion

Chapter 3 of this thesis was dedicated to develop a computational workflow able to improve the biological characterisation and the analysis of data derived from single cell RNA sequencing. The workflow proposed is divided in two main steps. The first step is performed by rCASC. The second by rMLSC. rCASC is a reproducible pipeline for scRNA-seq data analysis, composed of several phases. rCASC allows to select clusters of cells exploiting different clustering algorithm. rMLSC is a post-processing workflow that analyze rCASC results in five phases, exploiting random forest approach, to derive sets of gene signatures for each cluster who contribute is to improve the biological explanation of the clusters. rCASC and rMLSC were tested on a single cell dataset presented in Pace et al., 2018 paper, composed of immune cells derived from WT and KO for Suv29h1 mice. The analysis of rCASC derived five clusters with stability score (CSS) ranging from 75% and 100% for more than a half clusters. Then, rMLSC workflow was applied on rCASC results and we derived five sets of gene signatures using random forest algorithm to build a cell-cluster classification model, obtaining good average accuracy. We performed several enrichment analyses to explore the genes from the molecular function and cell type point of view. We also investigated the number of genes unique and shared among the clusters highlighting which of these had the major contribution in inter-cluster division as well as intra-cluster separation among WT and KO cells. The final analysis of the five gene signatures supports the idea of an immune cell timeline with cells belonging to cluster 1 and 3 as earlier precursors of more differentiated cells belonging to cluster 2,4 and 5. Clusters 1 and 3 are more probably partially differentiated cells of the bone marrow due to the presence of genes in the signature typical of precursors of blood dendritic and CD34+ cell. Clusters 2 and 5 show genes typical of CD4+ immune cell branch and NK cells while cluster 4 contains genes more CD8+ oriented. Thus, these findings support the idea that cluster 2 is composed by cells belonging to CD4 Helper Lymphocytes with the role of releasing cytokines and help in suppress or regulate immune responses. The cells of cluster 4 are more related to CD8+ (cytotoxic) T cells functions (effector) that are crucial for immune defence against intra-cellular pathogens, including viruses and bacteria, and for tumour surveillance, while cluster 5 contains a mixture of CD4+ thymocyte and NK cells. Pace et al., 2018 identified four specific immune cell types (i.e. Naive, Memory, Effector and Cycling cells) through a the state-of-art single cell pipeline, Seurat (Butler et al., 2018). However, the identification of the four clusters by the authors was actually strongly driven by the use of a set of 24 immunological genes that are known in literature as involved in the commitment process of the immune cells (Figure 3.9). Thus, the construction and the biological interpretation of the clusters, in this case, did not consider all other genes of the cells. Indeed, the majority of the pipelines for single cell data analysis have as ultimate step the derivation of Differentially Expressed Genes (DEGs), that are often few explanatory of the biological meanings and underestimate the real complexity of scRNA-seq data. Our workflow has allowed a greater biological comprehension of the cells contained in the dataset going further the ultimate derivation of DEGs of classical analysis pipelines. Finally, the combination of the two approaches, rCASC and rMLSC, allows to correlate the obtained results with the current biological evidences.

TABLE 3.10: **Comparison of gene signatures: Gene shared and unique of the clusters**

Clusters	Number of genes	Gene names									
C5	81	Ptma	Hsp90ab1	Gapdh	Atp5e	Atp5b	Hmgb2	Arpc5	Psmb8	Sec61b	Cox6b1
		Hnrnpf	Cox6c	Thy1	Pabpc1	Ldha	Cd48	Atp5j	Ndufa4	Hint1	Vim
		Cotl1	Capzb	Snrpg	Cox5b	Lat	Prdx1	S100a11	Cd47	Tagln2	Ran
		Npm1	Eef1g	Cox7c	Txn1	Eef1b2	Chchd2	Gzma	H2afz	Eif3h	Tpm3
		Ppib	Gimap4	Cox5a	Emp3	Anxa2	Krtcap2	Dnajc15	Atp5f1	Sem1	Psmb1
		Snrpe	mt-Nd2	Hnrnpa2b	Elob	Eif5a	Atp5j2	Cd3e	S100a13	Stmn1	Ptprc
		Ubl5	Tubb5	Serbp1	Pclaf	Cox7a2	Anxa6	H2-Q7	Gmfg	Eno1	Cox6a1
		Ddx5	Arpc3	Ostf1	Lgals3	Hnrnpk	Sept1	Tspo	Atp5mpl	Ndufb7	Serf2
		Epsti1									
C2-C5	28	Abracl	Actr3	Btf3	Cd3g	Cdc42	Clic1	Cox8a	Ctsd	Cyba	Fth1
		Ftl1	Fxyd5	Gzmb	H3f3b	Laptm5	Lck	Myl12a	Oaz1	Pfn1	Ppia
		Ppp1ca	Ptprcap	Pycard	S100a4	Selplg	Slc25a5	Sub1	Uqcrh		
C2-C4-C5	27	Arhgdib	Arpc1b	Arpc2	Atp5h	Calm1	Ccl5	Cd3d	Cfl1	Coro1a	Cox4i1
		Crip1	Eif1	H2-K1	Hspa8	Itm2b	Lgals1	Lsp1	Ms4a4b	Myl6	Naca
		Nkg7	Rac2	S100a10	S100a6	Sh3bgrl3	Uba52	Ubb			
C1-C2-C3-C4-C5	8	Actb	B2m	Eef1a1	Fau	mt-Cytb	mt-Nd1	Tmsb4x	Tpt1		
C1-C2-C3-C4	5	Cd52	Cd8b1	Eef2	H2-D1	Rack1					
C2	2	Ms4a6b	Shisa5								
C1-C3-C4-C5	1	Tmsb10									
C1-C2-C5	1	mt-Nd4									

Chapter 4

Characterization of a trend signature for Multiple Sclerosis

4.1 Background

Multiple Sclerosis (MS) is a chronic and potentially highly disabling disease with considerable social impacts and economic consequences. In Europe it is the leading cause of non-traumatic disabilities in young adults, since more than 700,000 EU people suffer from MS (Dutta and Trapp, 2011).

Multiple sclerosis is an inflammatory autoimmune disease in which the patient's immune system reacts against itself by damaging CNS nerve cells - i.e. compromising the ability of the neurons to send electrical signals - resulting in a progression of physical handicap until complete paralysis within 25 years in more than 30% of patients (Trapp and Nave, 2008).

In literature four courses of MS are identified: Relapsing-Remitting MS (RRMS), Secondary Progressive MS (SPMS), Primary Progressive MS (PPMS), and Progressive Relapsing MS (PRMS). Among them the RRMS is the most common course since it is diagnosed in about 85% of MS cases. It is characterized by episodes of neurological dysfunction (i.e. relapses) followed by a complete or partial recovery (i.e. remissions). Unfortunately, within 25 years RRMS usually changes to SPMS (in about 90% of cases) increasing the severity of the disease. (Dutta and Trapp, 2011)

Despite the etiology of MS is unknown, researchers agree that also environmental factors can act as triggers of MS, leading to the inflammatory process in the Central Nervous System (CNS). In particular, viruses may play a role in MS pathogenesis acting as such environmental triggers. Some studies linked MS with Epstein Barr Virus (EBV) infection due to the presence of higher titers of EBV antibodies in MS patients compared to age-matched controls (Virtanen, 2012).

Besides environmental factors, physiological factors also impact on the outcome of the MS disease. In particular, pregnancy represents a period of immune tolerance for patients that has important consequences on the relapse rate (Yamout and Alroughani, 2018). Indeed, pregnancy condition seems to have beneficial effects on women patients which have been associated with fewer relapses in RRMS. This process has been related with an increase in a particular type of immune cells, the Regulatory T lymphocytes cells (Treg), which confers fetal tolerance and thus shows a protective effect of pregnancy to patients (McCombe, 2018).

In the last two decades the advances in the understanding of the immune pathogenesis of MS and the advent of Monoclonal Antibodies (mAb) allowed researchers to define novel treatments against this disease. In particular mAb are powerful new tools to modify the course of MS based on a molecular targeted approach. Indeed, they are potentially able to break the immune cascade of events that brings to the autoimmune reaction causing the myelin loss. Therefore, these treatments that include several mAbs such as Natalizumab, Rituximab, and Alemtuzumab, constitute nowadays the most effective first and second line treatments in the

therapy of MS (Clerico et al., 2017; Salzer et al., 2016; Guarnera, Bramanti, and Mazzon, 2017).

Moreover, when the first and second line treatments provide an inadequate response in patients, daclizumab (DAC) treatment (Wynn et al., 2010) represented the only third line treatment to be used as a valid alternative. Differently from the other mAbs, DAC is a humanized monoclonal IgG1 antibody tailored against InterLeukin-2 Receptor (IL2R), thus able to break the autoimmune reaction by suppressing the immune cells expansion.

The basic mechanism of MS is, however, not fully understood yet and, despite its promising results, in 2018 DAC was withdrawn from the EU marketing authorization process due to the observation of twelve cases of patients who developed, after the beginning of the treatment, serious immune-mediated adverse reactions at the level of the CNS, including encephalitis and meningoencephalitis. More studies are needed to understand these effects as well as to explain why women affected by MS seem to improve when they become pregnant and during the pregnancy period. To improve our understanding of these process, we extend the RRMS models presented in Pernice et al., 2018; Pennisi et al., 2013; Beccuti et al., 2018 proposing a new computational methodology to analyse the RRMS behaviour. Hence, we firstly describe how the Extended Stochastic Symmetric Net (ESSN) formalism can be efficiently used to derive a graphical and parametric description of the system under study. Then, we show how the system of Ordinary Differential Equations (ODEs), that can be automatically derived from an ESSN model, reproduces the disease dynamics and how uncertainty and sensitivity analysis can be used to make more robust the results provided by the model. Finally we tested the proposed methodology constructing a model which allows to represent two different scenarios where the effect of the daclizumab administration is investigated and the RRMS in pregnant women is considered.

4.2 Methods

In this section we provide an intuitive introduction to the formalism used to model and analyze our case study. The Petri Nets formalism is firstly introduced, then a specific type of high-level extension, called Stochastic Symmetric Net (SSN) (Chiola et al., 1993), is described. After, the technique used to derive the qualitative properties of systems modelled with this formalism is discussed showing how these results can be computed efficiently using a fluid approximation. Then a new extension of the SSN formalism, called ESSN, is introduced to deal easily with more complex biological laws different by Mass Action one. Finally, in the last part of this section we describe how the model sensitivity analysis can be carried out using a sampling-based method.

Petri Nets and Stochastic Symmetric Net

Petri Nets (PNs) Marsan et al., 1995 and their extensions are well-known computational and mathematical formalisms which provide a graphical intuitive and formal description of the important features of the system under study. They allow the use of different analysis techniques to derive the qualitative and quantitative properties of a system.

PNs are bipartite directed graphs with two types of nodes, namely places and transitions. Places, correspond to state variables of the system and are graphically represented as circles. For instance, in Figure 4.1 an example PN model is presented, where it is described i) the Effector T cells (Teff) attack to the myelin sheaths due to the structure similarity of the viral protein with myelin proteins, and ii) the Oligodentrocytes cells (ODC) recovery of the lost myelin when the damage is not irreversible. Indeed, these events play a central role in RRMS progression, and more details will be given in Sec. *Model description*. Here the *Teff*

and *ODC* nodes are model places representing the Effector T cells and the Oligodentrocytes cells, respectively.

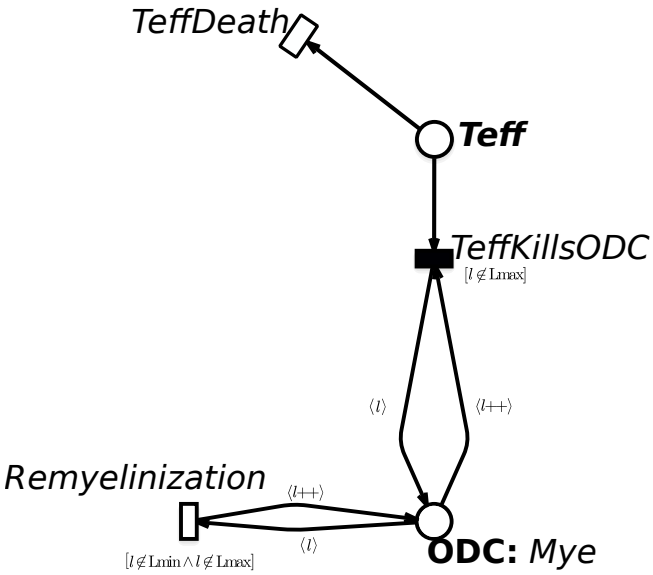


FIGURE 4.1: **Example of SSN.** Example of SSN representing the Effector T cells (place on the top named as Teff) which damage the Oligodentrocytes cells (place on the bottom named as ODC), and their partially recovery of the lost myelin when the damage is not excessive. This is a sub net of the SSN represented in Figure 4.2.

Differently, transitions correspond to the events that can induce a state change and are graphically represented as boxes. Referring again to Figure 4.1, transitions are *TeffDeath*, *Remyelinization* and *TeffKillsODC* which simulate the Teff death, the ODC recovery, and the damages of the Teff over the ODC cells, respectively.

The arcs connecting places to transitions (and vice-versa) express the relation between states and event occurrences. Places can contain tokens, drawn as black dots. The state of a PN, called *marking*, is defined by the number of tokens in each place of the model.

The system evolution derives from the firing of enabled transitions, where a transition is enabled if and only if each input place contains a number of tokens greater or equal than a given threshold defined by the cardinality (multiplicity) of the corresponding input arc. Thus, the firing of an enabled transition removes a fixed number of tokens from its input places and adds a fixed number of tokens into its output places, according to the cardinality of its input/output arcs.

Among the PN generalisations proposed in literature, SSNs Chiola et al., 1993 extend PNs providing a more compact and readable representation of the system, thanks to the possibility of using tokens belonging to different classes and thus graphically represented in the models as dots of different colors.

In SSNs each place p has an associated color domain (a data type) denoted $cd(p)$ and each token in a given place has an associated value defined by $cd(p)$. Color domains are defined by the Cartesian product of elementary types called *color classes* $\mathcal{C} = \{C_1, \ldots, C_n\}$, so that $cd(p) = C_1^{e_1} \times C_2^{e_2} \times \ldots \times C_n^{e_n}$ where e_i is the number of times C_i appears in $cd(p)$. Color classes are finite and disjoint sets. They can be ordered (in this case a successor function is defined on the class, inducing a circular order among the elements in the class), and can be partitioned into (static) subclasses (e.g $C_{i,j}$ is the i^{th} static subclass of the j^{th} color class).

In the example model represented in Figure 4.1 the ODC color domain is defined by one color class, the myelination levels of ODC cells, named *Mye*. This is divided into 5 static

subclasses (i.e. L_{min}, $L1$, $L2$, $L3$ and L_{max}) so that myelination level ranges from an irreversible damage (L_{min}, no myelination) to no damages (L_{max}, full myelination). Similarly, a color domain is associated with transitions and is defined as a set of typed variables, where the variables are those appearing in the functions labeling the transition arcs and their types are the color classes. For instance, the color domain of transition *Remyelinization*, representing the recovery of a ODC cell, is Mye and the variable characterizing its input arc is $l \in Mye$.

An instance of a given transition t is an assignment of the transition variables to a specific color of proper type. Hence, we use the notation $\langle t, c \rangle$ to denote an instance, where c is the assignment, also called binding. Moreover, a guard can be used to define restrictions on the allowed instances of a transition. A guard is a logical expression defined on the color domain of the transition, and its terms, called basic predicates, allow (*i*) to compare colors assigned to variables of the same type $(x = y,\ x \neq y)$; (*ii*) to test whether a color element belongs to a given static subclass $(x \in C_{i,j})$; (*iii*) to compare the static sub-classes of the colors assigned to two variables $(d(x) = d(y), d(x) \neq d(y))$.

The *marking* of an SSN is defined by the number of colored tokens in each place. For instance, a possible marking of the place *DAC* in Figure 4.1, is $500\langle Lmax \rangle$ corresponding to 500 ODC cells with a full myelination.

Each arc connecting a place p to a transition t, namely an input arc of t, is labeled with an expression defined by the function $I[p,t] : cd(t) \rightarrow Bag[cd(p)]$, where $Bag[A]$ is the set of multisets built on set A, and if $b \in Bag[A] \wedge a \in A$, $b[a]$ denotes the multiplicity of a in the multiset b. Similarly, each arc connecting a transition t to a place p, namely an output arc of t, is denoted by the function $O[p,t] : cd(t) \rightarrow Bag[cd(p)]$. Thus, the evaluation of $I[p,t]$ (resp. $O[p,t]$), given a legal binding of t, provides the multiset of colored tokens that will be withdrawn from (input arc) or will be added to (output arc) the place connected to that arc by the firing of such transition instance. Moreover, we denote with ${}^\bullet\mathbf{t}$ the set of input places of the transition t and with $\mathbf{t}^\bullet$ the set of output places of t, i.e. ${}^\bullet\mathbf{t} := \{p \in P |\ \exists\ c \in cd(p)\ s.t.\ I[p,\mathbf{t}](c')[c] > 0\}$ and $\mathbf{t}^\bullet := \{p \in P |\ \exists\ c \in cd(p)\ s.t.\ O[p,\mathbf{t}](c')[c] > 0\}$. In details, a transition instance $\langle t, c \rangle$ is enabled and can fire in an marking m, iff: (1) its guard evaluated on c is true; (2) for each place p we have that $I[p,t](c) \leq m(p)$, where $\leq$ is the comparison operator among multisets. We use the notation $E(t,m)$ to denote the set of all instances of t enabled in marking m. The firing of the enabled transition instance $\langle t, c \rangle$ in m produces a new marking m' such that, for each place p, we have $m'(p) = m(p) + O[p,t](c) - I[p,t](c)$.

In the SSNs, the firing time of an enabled transition instance $\langle t, c \rangle$ is sampled from a negative exponential distribution whose rate is given by the function ω, i.e.

$$\omega(t,c) = \begin{cases} r_i & cond_i(c)\ i = 1, \ldots, n, \\ r_{n+1} & otherwise, \end{cases}$$

where $cond_i$ are boolean and mutually exclusive expressions comprising standard predicates on the transition color instance. In this manner, the firing rate r_i of a transition instance can depend only on the static sub-classes of the objects assigned to the transition parameters and on the comparison of variables of the same type. Thus, these stochastic firing delays, sampled from a negative exponential distribution, permit to automatically derived the stochastic process, i.e. a Continuous Time Markov Chain (CTMC), that describes the dynamics of the SSN model. Specifically, the CTMC states are isomorphic to SSN markings and the state changes correspond to the marking changes in the model.

Hereafter we recall the formal definition of SSN.

Definition 1 (Stochastic Symmetric Net)**.** *An SSN is a nine-tuple:*

$$\mathcal{N}_{SSN} = \langle P, T, \mathcal{C}, I, O, cd, \Theta, \omega, \mathbf{m}_0 \rangle$$

where

- *P and T are two disjoint finite non empty sets (representing places and transitions respectively).*
- *$\mathcal{C} = \{C_1, \ldots, C_n\}$ is the finite set of basic color classes.*
- *$cd : \bigotimes_{i=1}^{n} \bigotimes_{j}^{e_i} C_i^j$ is a function defining the color domain of each place and transition (where $e_i \in \mathbb{N}$ is the number of occurrences of the class C_i); for places it is expressed as Cartesian product of basic color classes, for transitions it is expressed as a list of variables with their types. Observe that a place may contain undistinguished tokens only or a transition may have no parameters, in this case their domain is* neutral*.*
- *$I, O[p,t] : cd(t) \to Bag[cd(p)]$ are the pre- and post- matrices, whose elements are in the form of the arc functions defined above.*
- *Θ is the vector of guards and maps each element of T into a standard predicate ($\Theta(t)$ may be the constant* true*, which is also a standard predicate).*
- *$\omega(t,c)$ is the function returning the rate of transition t assuming the firing of the instance $\langle t, \boldsymbol{c} \rangle$.*
- *$\mathbf{m}_0 : P \to Bag[cd(p)]$ is the initial marking, mapping each place p on a multiset on $cd(p)$.*

Assuming that all the transitions of the SSN are characterized by a Mass Action law, the intensity (also calles the transition speed) of $\langle t, c \rangle$ in marking m is defined as follows:

$$\varphi(m,t,c) = \omega(t,c) \prod_{\langle p,c' \rangle |\ p \in {}^\bullet \mathbf{t} \ \wedge\ c' \in cd(p)} m[p][c']^{I[p,t](c)[c']} \tag{4.1}$$

where $m[p][c']$ denotes the marking of place p for color c'.

In the literature, different techniques are proposed to solve (or analyse) the CTMC underlying an SSN; in particular, in case of very complex models, the so-called deterministic approach Kurtz, 1978 can be efficiently exploited. According to this, in Beccuti et al., 2015 we described how to derive a deterministic process, represented through a system of ODEs, which well approximates the stochastic behavior of an SSN model. In particular for each place p and possible color tuple $c \in cd(p)$ we have the following ODE:

$$\frac{dx_{p,c}(\nu)}{d\nu} = \sum_{\langle t',c' \rangle \in E(t',x(\nu))\ \wedge\ t' \in T} \varphi(x(\nu), t', c')(L[p,t'](c')[c]) \tag{4.2}$$

where $x_{p,c}(\nu)$ is the average number of tokens of color c in the place p at time ν, $L[p,t'](c')[c] = O[p,t'](c')[c] - I[p,t'](c')[c]$, T is the set of transitions of the SSN, and $E(t, x(\nu))$ the set of the enabled instances of t in $x(\nu)$, i.e. the vector of the average number of tokens at time ν for each place and possible color tuple. In this case eq. 4.1 becomes

$$\varphi(x(\nu), t, c) = \omega(t,c) \prod_{\langle p_j,c' \rangle |\ p \in {}^\bullet \mathbf{t} \ \wedge\ c' \in cd(p_j)} x_{p_j,c'}(\nu)^{I[p_j,t](c')[c]}. \tag{4.3}$$

Extended Stochastic Symmetric Net

It is important to highlight that the reactions velocity can be defined by more complex laws than Mass Action (MA), for instance Michaelis Menten and Hill kinetics. In Pernice et al., 2019a the Extended Stochastic Petri Nets (ESPNs) formalism was presented to extend SPN with general functions which make easier to model reactions with more complex biological laws. Similarly here we propose a new formalism, called Extended Stochastic Symmetric Net (ESSN), which extends the SSN exploiting the same ideas discussed in the proposal of the ESPN formalism Pernice et al., 2019a.

In details, the set of transitions T is split in two subsets T_{ma} and T_g. The former subset contains all transitions which fire with a rate following a MA law. The latter includes instead all the transitions whose random firing times have rates that are defined as general real functions. Hence, we will refer to the transitions belonging to T_{ma} as standard transitions and as general transitions those in T_g. For instance, considering the Figure 4.1 again, the general transition is graphically represented as black box and is that simulating the myelin damage, i.e. *TeffKillODC*. In details, the function of the general transition is given in the Additional File 1.

Definition 2 (Extended Stochastic Symmetric Net)**.** *An ESSN is a ten-tuple:*

$$\mathcal{N}_{ESSN} = \langle P, T, \mathcal{C}, I, O, cd, \Theta, \omega, \Omega, m_0 \rangle$$

where

- $P, \mathcal{C}, I, O, cd, \Theta, m_0$ *are defined as in SSN (see definition 1).*
- T *is the set of transitions and is defined as* $T = T_{ma} \cup T_g$*, with* $T_{ma} \cap T_g = \emptyset$*. Where* $T_{ma} = \{t_i^*\}_{1 \leq i \leq n_{T_{ma}}}$ *is the set of the* $n_{T_{ma}}$ *transitions whose speeds follow the MA law, and* $T_g = \{\bar{t}_i\}_{1 \leq i \leq n_{T_g}}$ *is the set of the* n_{T_g} *transitions whose speeds are defined as general functions.*
- $\omega(t, c)$ *is the rate of transition* $t \in T_{ma}$ *assuming the firing of the instance* $\langle t, c \rangle$*.*
- $\Omega = \{f_{\langle t,c \rangle}\}_{t \in T \wedge c \in cd(t)}$ *is set grouping all the transition speeds* $\forall t \in T$*. In detail, with* $t \in T_{ma}$ *then* $f_{\langle t,c \rangle} = \varphi(\cdot, t, c)$*, where* φ *is defined in Eq. 4.1.*

Similarly to what discussed in Sec. *Petri Nets and Stochastic Symmetric Nets*, let $x_{p,c}(\nu) \in \mathbb{R}^+$ be the continuous approximation of the number of tokens in place p and color c so that the vector $x(\nu) \in \mathbb{R}^{n_P}$ is the marking of the ESPN at time ν.
Let define $\hat{x}(\nu) = x(\nu)_{|\bullet t}$ as the subset of the marking $x(\nu)$ concerning only the input places to the transition t. Given $\langle t, c \rangle$ at the time ν, with transition $t \in T = T_{ma} \cup T_g$, the firing of $\langle t, c \rangle$ will move tokens in state $x_{\langle p,c \rangle}(\nu)$ with speed $F(\hat{x}(\nu), t, c, \nu)$ defined as follows:

$$F(\hat{x}(\nu), t, c, \nu) := \begin{cases} \varphi(\hat{x}(\nu), t, c), & t \in T_{ma}, \\ f_{\langle t,c \rangle}(\hat{x}(\nu), \nu), & t \in T_g, \\ & f_{\langle t,c \rangle} \in \Omega(t, c) \end{cases} \tag{4.4}$$

where $\varphi(\hat{x}(\nu), t, c)$ is defined as in Eq. 4.3. Observe that $\varphi(\hat{x}(\nu), t, c)$ and $f_{\langle t,c \rangle}(\hat{x}(\nu), \nu)$ can depend only on the marking of the input places of transition t at time ν.

Finally the ODE characterizing the p and color tuple $c \in cd(p)$ is defined as:

$$\begin{aligned}\frac{dx_{p,c}(\nu)}{d\nu} &= \sum_{\langle \mathbf{t}',\mathbf{c}'\rangle \in E(\mathbf{t}',x(\nu))} F(\hat{x}(\nu),\mathbf{t}',\mathbf{c}',\nu)(L[p,\mathbf{t}'](\mathbf{c}')[c]) \\ &= \sum_{\substack{\langle \mathbf{t}',\mathbf{c}'\rangle \in E(\mathbf{t}',x(\nu)) \\ \wedge \mathbf{t}' \in T_{ma}}} \varphi(\hat{x}(\nu),\mathbf{t}',\mathbf{c}')(L[p,\mathbf{t}'](\mathbf{c}')[c]) \\ &+ \sum_{\substack{\langle \mathbf{t}',\mathbf{c}'\rangle \in E(\mathbf{t}',x(\nu)) \\ \wedge \mathbf{t}' \in T_{g}}} f_{\langle \mathbf{t}',\mathbf{c}'\rangle}(\hat{x}(\nu),\nu)(L[p,\mathbf{t}'](\mathbf{c}')[c]) \end{aligned} \tag{4.5}$$

where $\hat{x}(\nu) = x(\nu)_{|\bullet \mathbf{t}'}$.

Sensitivity analysis

Sensitivity analysis is broadly used in computational modelling to study which parameters affect mostly the variability of the outcomes produced by the model. Several approaches are proposed in the literature to achieve this task, such as Pearson correlation coefficient (CC) method (for linear relationships), Partial Rank Correlation Coefficient (PRCC) method (for non-linear and monotonic relationships) or Fourier Amplitude Sensitivity Test (FAST) method (for any non-linear relationships) Marino et al., 2008; Saltelli et al., 2005. In this work we focus on a sampling-based method which combines Latin Hypercube Sampling (LHS) McKay, Beckman, and Conover, 1979 with PRCC index. Practically LHS, a well-known stratified sampling method, is adopted to generate samples of the model input variables. Then the model is run N times in a chosen interval: one for each generated input variable sample combination. Finally PRCC between the generated input variables and the obtained model outputs are evaluated on the same chosen interval. In this way PRCC analysis and corresponding significance tests (i.e significant p-value) are used to identify key model parameters and to select time points which need an additional in-depth investigation. Specifically, PRCC values close to 1 (-1) identify positive (negative) monotone relationships between inputs and outputs; while the significance tests permit to discover those correlations that are important, despite having relatively small PRCC values.

4.3 Model description

In this section we first describe how the healthy immune system achieves the immunosurveillance. Then we focus on the pathogenesis correlated to MS, highlighting the roles of the immune system cells, CNS cells, and EBV virus. Finally, we describe the structure of the model dividing it into seven modules.

4.3.1 Healthy case

The immune system represents the entire compartment of cells leading the defense of the human body against potentially damaging foreign molecules and pathogens with a highly specific response to the encountered infectious agents. This specific response is conducted by a stringent selection and maturation of the naive T lymphocytes in the lymphoid organs (e.g. thymus), prior to exit in the periphery of the body as mature T lymphocytes cells. Consequently to the pathogens entrance, the naive T lymphocytes become activated T lymphocytes by Antigens Presenting Cells (APCs) via the T Cell Receptor (TCR). The activated T lymphocytes produce Interleukin-2 (IL2) an immunomodulating cytokine released for self-stimulating in order to duplicate and propagate their actions, via the binding of IL2 to the

receptor IL2R, located on the surface of the cells. Thanks to this bond, activated T lymphocytes undergo the clonal expansion process (i.e. lymphocytes multiplication). During clonal expansion, a portion of the activated T lymphocytes are destined to become Effector T lymphocytes cells (Teff) and others as Memory T lymphocytes cells (Tmem). Thanks to the presence of Tmem cells, the responses to subsequent attacks from the same pathogen are faster and greater than the first one. Since the response of Teff is reversed when no longer needed, the Treg - response T lymphocytes - contributes to suppress the Teff cells activity. (see Supplementary Figure 6.15). In addiction, IL2 brings to the activation of another family of immune cells, named Natural Killer (NK) cells that act as host-rejection of infected cells and, in same cases, against auto-reactive Teff populations, T lymphocytes recognizing self-molecules as foreign (Handel, Irani, and Holländer, 2018).

4.3.2 Multiple Sclerosis

MS is an autoimmune disease in which the immune system removes the myelin sheath from neuronal axons of the CNS preventing the efficient transmission of the nervous signals. RRMS is the predominant type of MS in which the disease alternates phases of active neural inflammation and disease worsening (relapses) to remission phases in which there is a complete or partial lack of the symptoms. The occurrences of the relapses periods range from mild to severe, depending on the course and history of the disease. The occurrence and severity of relapses are the only measures to estimate the efficacy of a drug treatment.

Several studies agree with the idea that the damage at CNS is probably due to auto-reactive T lymphocytes which recognize the myelin as foreign. Indeed, during the relapse phase, the continuous attack of T lymphocytes leads to a progressive decrease of the quantity of myelin (Kock and Yong, 2013). It is worth to note that in some cases the Oligodentrocytes cells (ODC) are able to partially recover the lost myelin if the damage is not excessive.

As already pointed out in the previous section, several T lymphocytes with different roles take part in the general immunosurveillance (e.g. Teff, Treg, Tmem). An imbalance in the T lymphocytes differentiation can lead to a strong or long-term response that goes beyond its original purpose. In fact, in literature it has been hypothesized that homeostasis of Treg and Teff cells may be crucial to prevent autoimmunity. Moreover, a breakdown of the peripheral tolerance mechanisms of Treg (represented, for example, by a lower duplication rate of Treg compared to Teff) can bring to the selection of self-reactive Teff cells leading to damage in the CNS (Dendrou, Fugger, and Friese, 2015). The etiology of MS remains elusive. Some studies linked MS with signs of Epstein Barr Virus (EBV) infection (Virtanen, 2012; Belbasis et al., 2015). A hypothesis is that the EBV first infection as well as the reactivation of the latent infection can cause the activation of auto-reactive Teff lymphocytes through a process called molecular mimicry. Molecular mimicry would cause the activation of Teff lymphocytes that recognize an EBV virus protein. However, due to the structure similarity between the viral protein and myelin proteins, these Teff could also be able to attack myelin sheaths leading to a neural damage (Virtanen, 2012).

A very interesting case of study is the development and progression of MS in pregnant women. The RRMS pregnant women have a lower relapse rate until the delivery (Vukusic et al., 2004). Indeed, during the pregnancy period the number of Treg cells increase to establish tolerance against fetal antigens. As a consequence, pregnancy represents a moment of immune tolerance and well-being for the patients.

4.3.3 Relapsing-Remitting Multiple Sclerosis model

The cell and molecular interactions involved in the Relapsing-Remitting Multiple Sclerosis(RRMS) are described by the model showed in Figure 4.2. Our model consists of 10 places

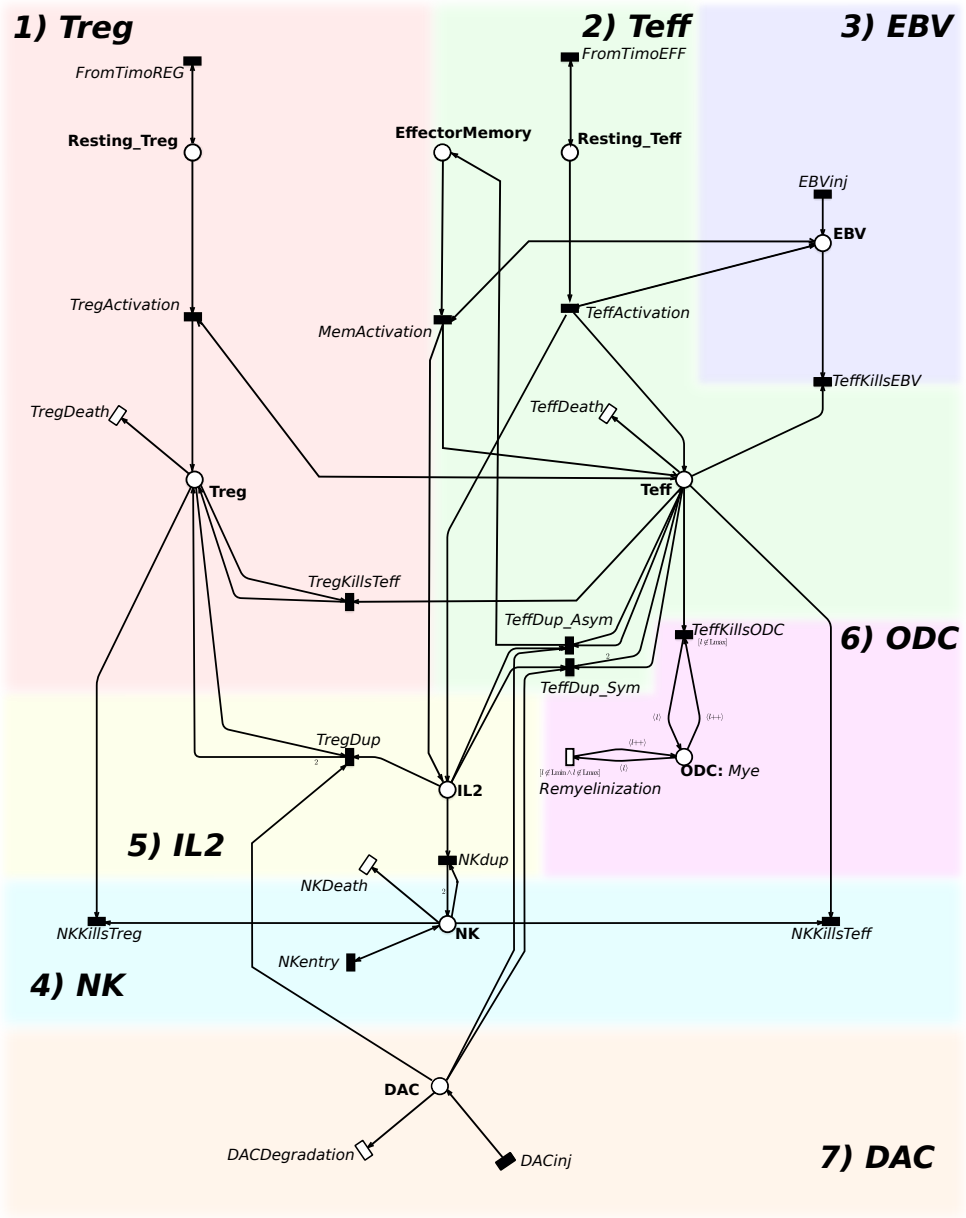


FIGURE 4.2: **The RRMS model.** The RRMS PN model is composed by places (graphically represented by circles) corresponding to cells or molecules, and by transitions (graphically represented by rectangles) corresponding to the interactions among the entities, injections or death of molecules. The RRMS model is composed by seven modules: Treg, Teff, EBV, NK, IL2, ODC and DAC.

and 22 transitions. For sake of clarity, the white transitions model mass-action functions, while the black transitions model different kinetics described in details in the Additional File 1. All the constants and numerical values associated with the transitions are reported in the Supplementary Tables 6.1 and 6.2. This model is organized in seven modules corresponding to the biological entities characterizing RRMS: Treg, Teff, EBV, NK, IL2, ODC, and DAC.

1) Treg module. The Treg cells are characterized by two places: the *Resting_Treg* and *Treg*. The transition *FromTimoREG* represents the arrival of new resting Treg cells from thymus. Its rate is defined in order to keep constant the number of resting cells. The transition

TregActivation represents the activation of the resting Treg depending on the Teff cell number and EBV concentration, while *TregDeath* represents the death of Treg. The transition *TregKillsTeff* models the homeostatic regulation operated by Treg cells against self-reactive Teff cells, and *TregDup* models the Treg duplication.

2) Teff module. The second module is characterized by three places: $Resting\_Teff$, $Teff$, and *EffectorMemory*. The transitions *FromTimoEFF*, *TeffActivation*, and *TeffDeath* behave similarly to those described in module 1, but they are referred to the Teff population.

The Teff proliferation takes place in two different manners called $Symmetrical$ and $Asymmetrical$ processes. These two possibilities are captured in the model by assuming that one happens with probability p_{eff}^{dup} and the other with probability $p_{eff}^{mem} = 1 - p_{eff}^{dup}$. Given a replication speed named r_{dup}^{eff}, the transition *TeffDup_Sym* generates two Teff cells with the rate equals to $r_{dup}^{eff} * p_{eff}^{dup}$, for more details see Additional Material. Otherwise, the transition *TeffDup_Asym* takes place with a speed resulting from the product $r_{dup}^{eff} * p_{eff}^{mem}$ replicating one Teff cell into one T Memory effector cell and one Teff cell. The transitions *TeffKillsEBV* and *TeffKillsODC* encode the killing effect of Teff cells against EBV and ODC, respectively. Finally, *MemActivation* models the rapid activation of the Effector Memory depending on both the EBV and the Tmem concentrations.

The transitions *TeffKillsEBV* and *TeffKillsODC* encode the killing effect of Teff cells against EBV and ODC, respectively. Finally, *MemActivation* models the rapid activation of the Effector Memory depending on both the EBV and the Tmem concentrations.

3) EBV module. The third module describes the EBV behaviour. Transition *EBVinj* models the infection. The *TeffKillsEBV* transition summarizes all steps from antigen processing and presentation by EBV infected cells to Teff cells, to the activation of Teff cells.

4) NK module. In this modules the role of the NK cells is described. The transition *NKentry* models the arrival of new NK cells. The death of NK is then modeled by transition *NKDeath*. Transitions *NKKillsTeff* and *NKKillsTreg* encode the killing of self-reactive Teff and Treg cells respectively due to NK cells. Finally *NKdup* models the proliferation of the NK cells led by the presence of IL2.

5) IL2 module. The fifth module is focused on the $IL2$ role. $IL2$ is involved in the $Treg$, $Teff$ and NK proliferation. All these types of cells consume IL2 which is produced by the transition *TeffActivation*.

6) ODC module. The sixth module encodes the ODC behaviour. The transition *TeffKillsODC* models the damage caused by Teff cells on ODC cells. When the myelination level reaches the lowest value, an irreversible damage occurs and the remyelinization is no more possible (i.e. the transition *Remyelinization* is permanently disabled by its guard).

To model this effect, we used the color class *Mye* encoding the myelination levels of ODC. *Mye* is divided into five static subclasses ranging from *Lmin* (no myelination) to *Lmax* (full myelination).

7) DAC module. In the last module the daclizumab behaviour is modeled through the place *DAC*. The drug administration is modeled by transition *DACinj*, while the pharmacokinetic inhibiting the expansion of Treg and Teff decreases the velocity of transitions *TregDup*,*TeffDup_Sym* and *TeffDup_Asym*. Finally, the its degradation is modeled by the

transition *DACDegradation*.

4.4 Results

In this section we present the prototype computational framework developed to study the RRMS. Then, we describe how our model was calibrated to mimic the real behaviours in healthy and MS subjects. Finally we discuss also the results coming from two case studies in which we investigated the effect of the daclizumab administration on MS patients and the modification of the MS evolution due to the occurring of a pregnant condition.

Framework Architecture

The prototype framework specifically developed for this project consists of a few modules which interact with GreatSPN (Babar et al., 2010), a software suite for modelling and analyzing complex systems using the PN formalism and its extensions. The architecture of this framework is depicted in Figure 4.3. The GreatSPN GUI is used to construct the ESSN model. First the coloured portion of a model "NetName" (similar to that depicted in Figure 4.2) is drawn on the GUI interface, then the "Unfolding" module is called to transform the coloured model in a basic Petri Nets representation where all the tokens are assumed to be indistinguishable and the files "NetName.net" and "NetName.def" are produced. Moreover the generic firing rate functions characteristic of the ESSN formalism must be specified in a separate file called "name.rate". Then the files "NetName.net" and "NetName.def" and "NetName.rate" are used by the module "PN2ODE" to automatically derive the corresponding system of ODEs producing a file "NetName.r" with a format suitable for being processed by the "R" software. Finally the "R" environment is used for the solution of the set of ODEs, for carrying out the sensitivity analysis with respect to model parameters, and for computing interesting results in terms of numerical values and diagrams. As the whole framework is not yet completely integrated with GreatSPN, calling the "Unfolding" and "PN2ODE" modules, as well as the interaction with the "R" software, must be done using command-line directives. GreatSPN extended with this prototype framework can be downloaded at `https://github.com/greatspn`. Instead, all the R files and the GreatSPN files of the net are freely available at `https://github.com/qBioTurin/ESSNandRRMS/DeterministicModel`.

4.4.1 Model calibration for healthy and MS individuals

The model calibration was addressed to select the input values (transition parameters, and the concentrations of EBV and DAC) leading outputs of the model towards the values obtained from observing the behaviour of the specific quantities both in healthy and MS affected subjects. The calibration is performed using the LHS with PRCC index to identify which parameters have more impact on the model outcomes. Then, the identified parameters were thoroughly investigated in the healthy and MS individuals.

From our model (without the *DAC* module, Figure 4.2) a system of 13 ODEs with nine input parameters is derived. The values of these nine parameters were sampled by means of LHS method. Hence, 5000 parameter combinations were generated using a *uniform* distribution whose ranges are showed in the third column of the Supplementary Table 6.1.

For all the simulations, we assumed as initial marking the following parameters consistent with a space of $1mm^3$ of blood and $4mm^3$ of neural tissue: 500 ODC with level L_{max} of neuronal myelinization, 1687 resting Teff cells, 63 resting Treg cells, 375 NK cells and 1000 IL2 molecules, and zero cells in the other places (see Supplementary Table 6.3).

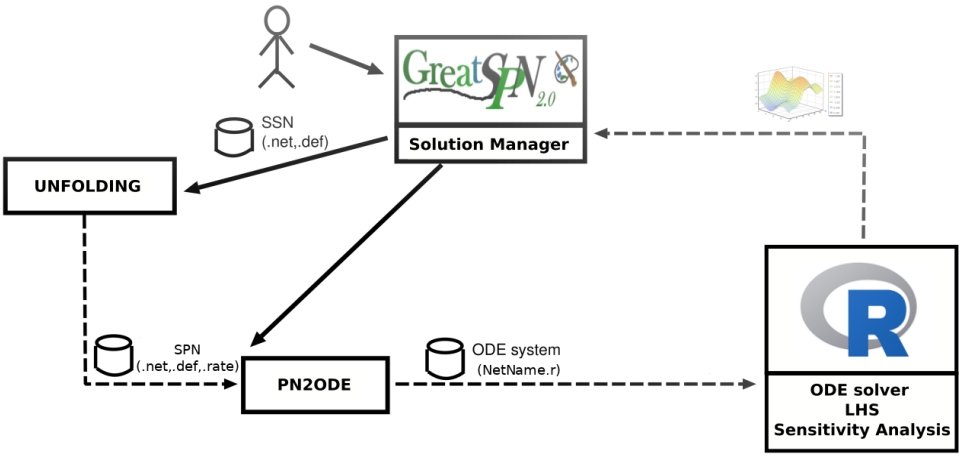


FIGURE 4.3: **Framework structure.** Outline of the prototype framework combining GreatSPN suite with R. The components are shown by rectangles, component invocations by solid arrows, models/data exchanges by dashed arrows.

Moreover, we defined the disease occurrence when the L_{min} level of neuronal myelinization is reached for each ODC cell, representing an irreversible damage. Then, five virus injections are simulated at regular times (every two months), introducing into the system 1000 EBV copies per injection. Finally, model solutions were calculated for each parameter combination over one year interval, $[0, 365]$ days.

Analyzing the 5000 trajectories generated, three scenarios have been identified: (i) the occurrence of the MS, represented by a huge number of dead ODC cells; (ii) the complete remission of the MS disease, characterized by a low number of dead ODC cells and with a complete elimination of the EBV virus; (iii) the partial remission of the MS disease specified by a partial elimination of the EBV virus. The Supplementary Figure 6.16 reports the EBV and ODC dynamics generated considering different set of parameters.

On these trajectories the PRCC analysis was applied to identify key model parameters affecting the system behaviour. The PRCCs values are calculated for each parameter over the entire time period. Figure 4.4 shows the PRCC values for all the model parameters over the time interval considered.

The rates associated with transitions *TeffKillODC*, *TregKillTeff*, *TeffKillEBV* and *Recovery* result to be the crucial parameters affecting the *ODC* behaviour. Figure 4.5 reports a scatter plot in which each point corresponds to a generated trajectory, its color represents the percentage of irreversible ODC damaged at the final time point (i.e. a grey color corresponds to a lowest percentage of damaged ODC and a red color to highest one). The simulations are performed changing the rates of *TeffKillEBV* (in the x-axis), the *TregfKillTeff* (in the y-axis) and *TeffKillODC* (in the z-axis). A few number of irreversible damaged ODC are obtained increasing the *TregfKillTeff* rate and the decreasing the others two rates.

The key parameters identified were deeply studied exploiting the LHS method computing 500 new combinations varying only *TeffKillEBV*, *TregfKillTeff* and *TeffKillODC*. We defined two sets of parameters (see Supplementary Table 6.5), one for the MS patients and one for the healthy subjects, see Figure 4.6.

The MS patients are modelled by a set of parameters which maximizes the ODC damage maintaining the Treg and Teff cell numbers consistent with those measured in the reality, see red trajectories in Figure 4.6 a and b panels.

For the healthy cases, we selected the trajectory providing a Teff-Treg regulatory balance

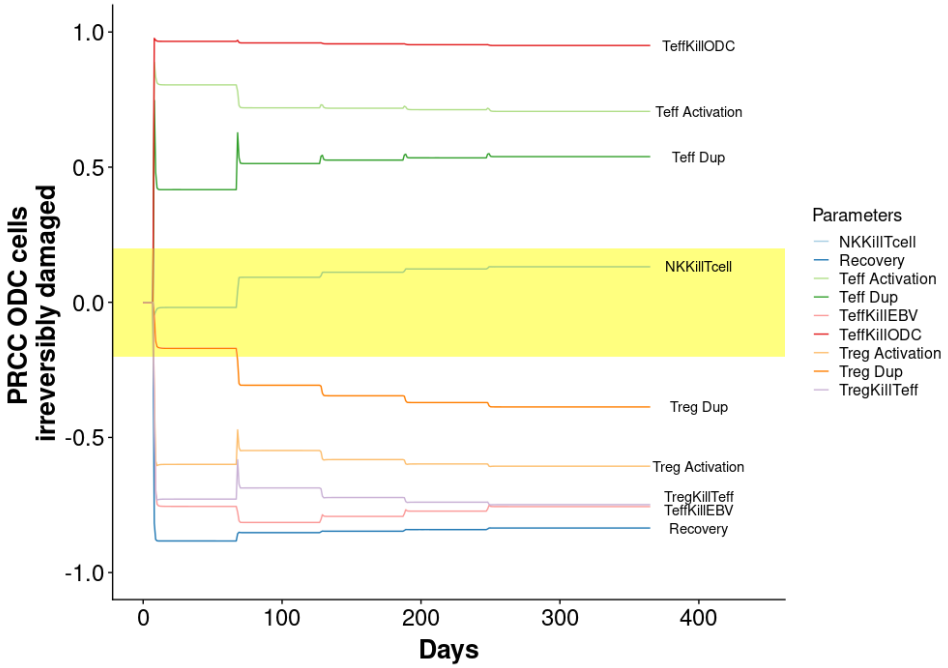


FIGURE 4.4: **Sensitivity analysis.** PRCCs over the whole time interval for each model parameter is reported. Yellow area represents the zone of non-significant PRCC values.

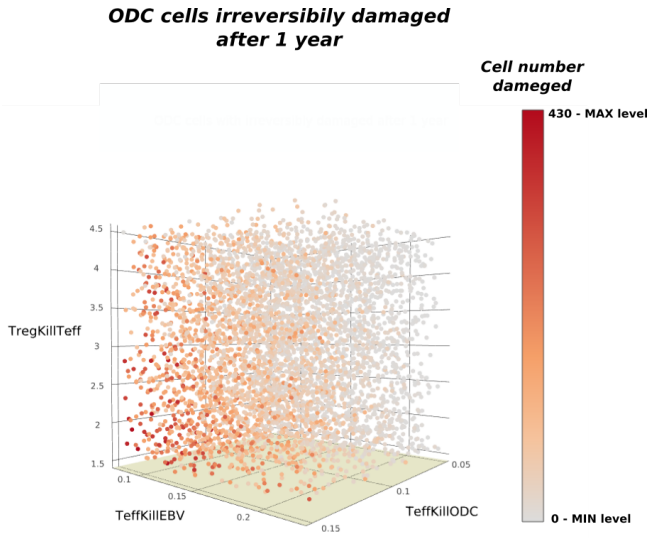


FIGURE 4.5: **Parameters scatter plot.** 3D scatter plot of the ODC irreversible damaged at the fixed time 365 versus the*TeffKillODC*, *TregKillTeff*, *TeffKillEBV* parameters variation.

able to control the spread of the EBV virus and to minimize the irreversible damage to ODC cells, even if the amount of EBV in each injection is substantially increased, see blue trajectories in Figure 4.6, panels a and b. In particular, we performed 500 simulations varying the amount of EBV injected in a range of $[1000 - 5000\ particles/mm^3]$. From Figure 4.7 it is possible to observe that, even for large quantities of EBV injected, the percentage of irreversibly damaged ODCs reaches 17% (Figure 4.7(a)). This value is very small if compared with respect to the 77% of irreversibly damaged ODCs in the case of the disease occurrence.

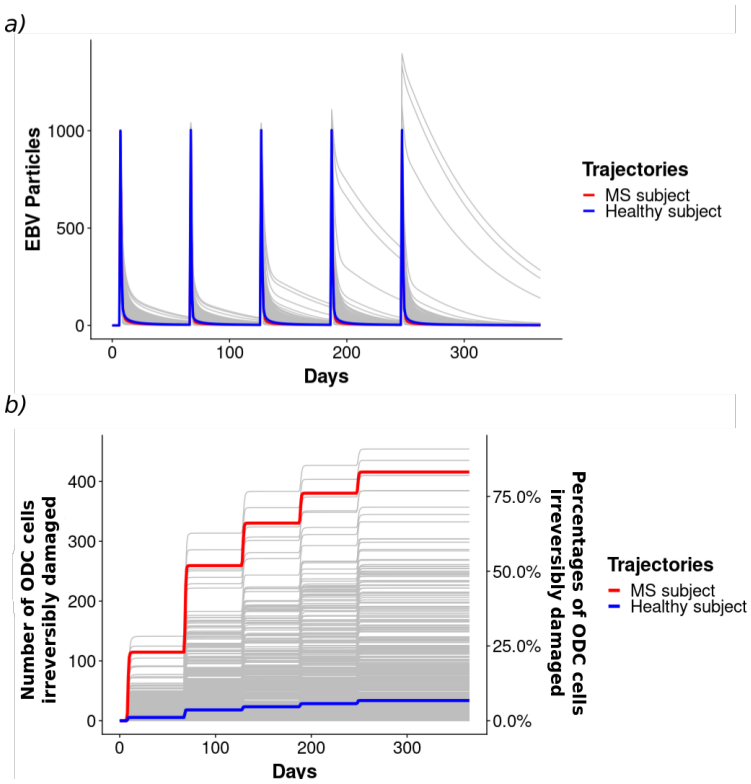


FIGURE 4.6: **Parameters choice.** A set of the 500 trajectories generated by LHS of the EBV virus (a) and the ODC cells with an irreversible damage (b) over the whole time interval varying the *TeffKillODC*, *TregKillTeff*, and *TeffKillEBV* transition parameters.

Moreover, independently of the quantity of EBV injected, Teff are able to eliminate EBV completely (Figure 4.7(d)), and the abundance of EBV does not drastically affect the number of effectors or regulators in the system (Figure 4.7(b,c)).

DAC therapy

To investigate the effect of the DAC therapy in our RRMS model calibrated for MS patients, we simulated the DAC administration at the 53rd day after the first EBV injection. Our results are reported considering two important aspects in the modulation of a therapy: the drug dose and the drug degradation time. The quantity of *DAC* administrated per injection and the *DAC* cells deterioration were studied by means of LHS method. The values of these two parameters are then sampled according to two uniform distributions whose ranges are reported into the Supplementary Table 6.1.

From the LHS analysis we clearly observed that the drug degradation time has a greater impact on the elimination of EBV virus than the amount of DAC administered, see the number of ODC irreversibly damaged in Supplementary Figure 6.17. Therefore, we decided to focus our attention on the *DACDegradation* parameter variation. Knowing that the half-life of DAC was detected around 22 days, we considered that a complete degradation of DAC ranges between 30 and 90 days (Kim and Baker, 2016). The results of the simulations are reported in Figure 4.8(a) in which it is possible to appreciate that a greater DAC permanence has the effect of reducing the number of irreversibly damaged ODC cells with respect to the case in which no therapy was considered (red line). Moreover, it is interesting to note that the RRMS model with DAC injections highlights a decrease of the long term ability of the

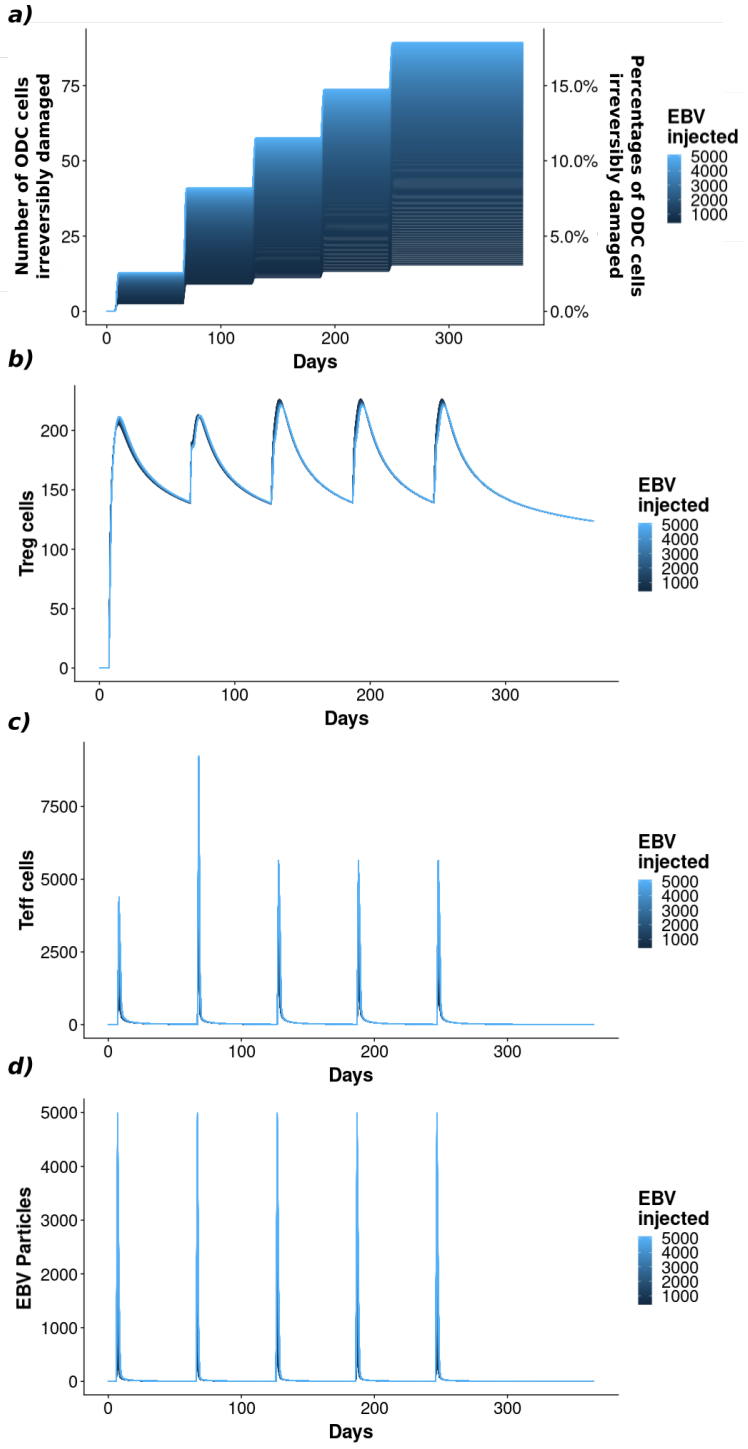


FIGURE 4.7: **Effect of EBV quantities.** Different injections of EBV (d) are considered to check if the Teff-Treg (c-b) regulatory loop is able to control the virus spreading minimizing the irreversible damages to the ODC cells (a).

immune system to eliminate EBV Figure 4.8(b). Finally, in Supplementary Figure 6.18 is reported the trend of the NK cells that increase with respect to the DAC degradation rate.

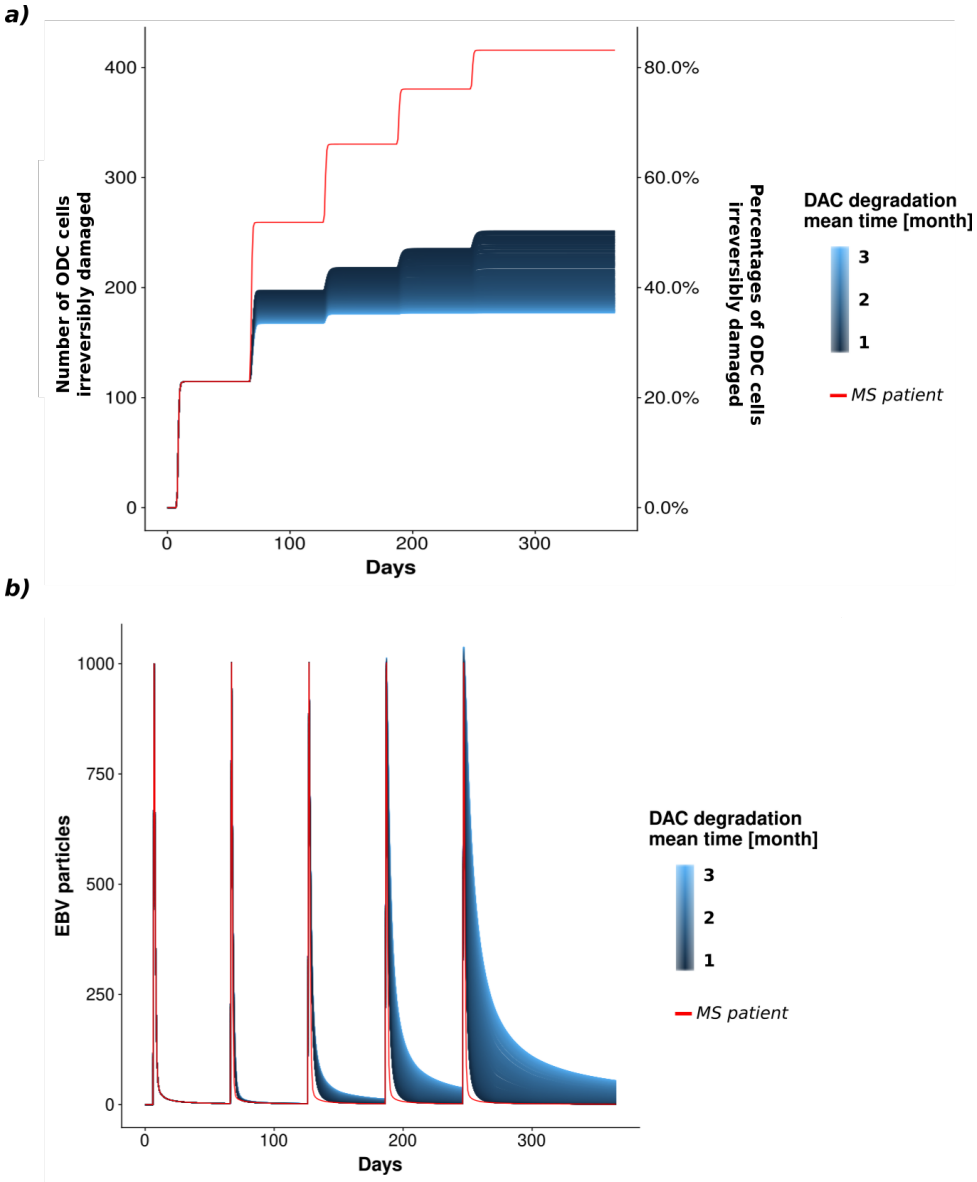


FIGURE 4.8: **Varying the DAC degradation rate.** *ODC* and *EBV* trajectories colored depending on *DAC* degradation rate (expressed in months). The red line represents the starting sample without drug administration.

Pregnancy

In this subsection we investigate the RRMS in pregnant women. As already pointed out before, pregnancy was associated with fewer relapses in RRMS and reduced activity of disease in autoimmune encephalomyelitis (EAE). Beneficial effects of pregnancy are thought to be related to pregnancy-associated changes in the maternal immune system. One of the observations is that Treg cells increase in number establishing the fetal tolerance and conferring a temporary protection to women with RRMS (Sánchez-Ramón et al., 2005; Somerset et al., 2004).

According to the literature, we modelled the pregnancy condition changing the proportion between the activate Treg cells and the activate Teff cells decreasing the Teff activation rate and increasing the Treg activation rate proportionally to the pregnancy phase (Somerset et al., 2004). Three pregnancy phases, corresponding to the three trimesters, have been simulated. When a new trimester begins, we increased the ratio of *TregActivation* rate to *TeffActivation*

rate; while at delivery time both rates return to their initial values.

Thus we simulated 100 different scenarios with a increasing variation of parameters, obtaining different levels of protection from ODC damage. As expected, the model behaviour shows a substantial reduction of the ODCs damage (see Figure 4.9).

Regarding the immune system cells, we observed that Treg cells increase during pregnancy and then suffer a sharp decline at the time of delivery. The same effect, but in the opposite direction is showed on Teff cells. It is interesting to note that a rebound of Teff is reported in the week following pregnancy, see Supplementary Figures 6.19.

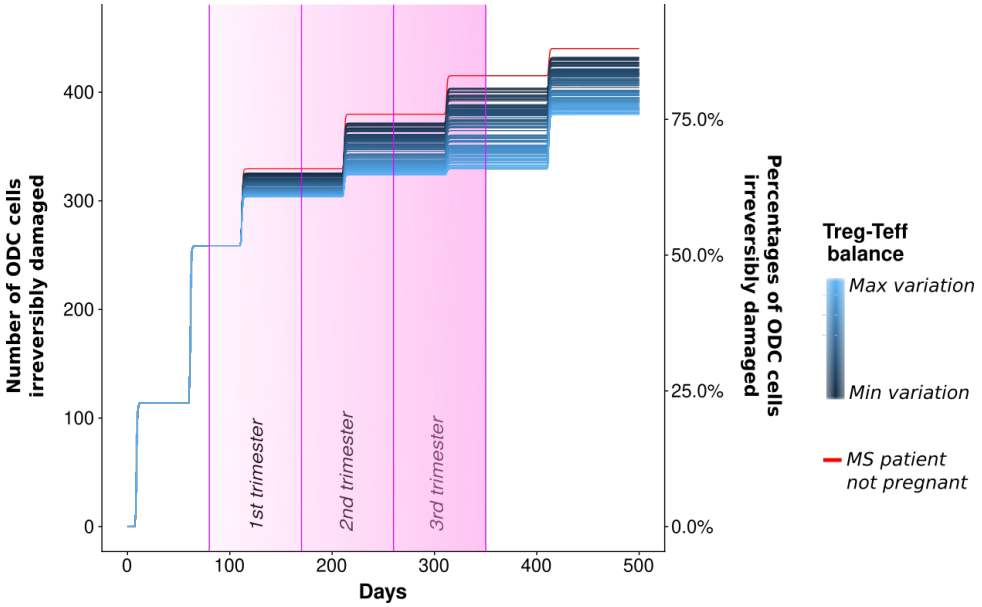


FIGURE 4.9: **Pregnant woman case: ODC.** The ODCs irreversibly damaged considering the pregnant woman case of study. 100 trajectories colored depending on different variations of the *TregActivation* and *TeffActivation* parameters. The red line represents the starting sample without pregnancy. Furthermore each trimester another variation is applied to these parameters in order to represent the increasing of the maternal immune system.

4.5 Discussion

Because of nowadays the computational modelling is widely recognized to succeed in helping scientists in the study of the complex mechanisms of different diseases, the aim of this part of the thesis is to present a new computational methodology and an associated model to better elucidate the dynamics behind the RRMS. Indeed, RRMS represents a very challenging case study, due to the complexity of the disease which involves many different biological agents, ranging from molecular to environmental factors.

We exploited the descriptive power of Extended Stochastic Symmetric Nets to provide a graphical representation of the complex biological system in a compact and parametric way. Moreover, we used LHS method with PRCC index to calibrate the model parameters. Hence, we showed the ability of the model to reproduce the typical oscillatory behavior relating to the onset of RRMS by supposing a breakdown of the cross-balance regulation mechanisms at the peripheral level. Moreover, the simulation of DAC injections in the RRMS model can help scientists to define the mechanisms of actions of this drug and to theorize the possible causes of its observed side-effect on the patients. Instead, the experiments simulating RRMS

in pregnant women can contribute to define the mechanisms at the basis of the variation of the Treg and Teff cells.

A challenging issue in the definition of the RRMS model is the calibration of the transition parameters and EBV and DAC concentrations. The LHS with PRCC index identified *TeffKillODC*, *TregKillTeff*, *TeffKillEBV* as the most critical parameters to the model outcomes. This result agrees with our expectation since these parameters play a central role in the disease progression. In the analysis of the EBV behaviour in a healthy subject, it is interesting to note the effect of immune memory which increases the number of activated Teff cells from the time of second injection(see Supplementary Figure 6.20). In particular, thanks to the faster activation of the Tmem cells with respect to the Teff cells, from the second EBV injection it is possible to observe a more rapid virus annihilation. Indeed, Tmem cells have a faster activation (i.e. since they already have the memory of a previous contact with the EBV) than Teff cells, leading to a more rapid virus annihilation during the relapses.

In the first set of our experiments we inspected the effect of the DAC therapy in our model calibrated for reproducing the behaviors of RRMS patients. A greater amount of latent EBV is present in the system with respect to the case in which no therapy was considered.

In the DAC therapy module there are two parameters form whom depend the elimination of EBV virus: the DAC degradation time and the DAC concentration. Using the parameter sensitivity analysis, we identified the drug degradation time as the crucial parameter, while the DAC concentration has ho effect on the EBV treatment. Indeed, this slight effect of the DAC concentration in the treatment of MS patients was recently described by Gold and coworkers (Gold et al., 2013). In this paper the authors described a clinical trial involving 76 centers in which the MS patients have been treated by subcutaneous injections of DAC HYP 150 mg or 300 mg, or placebo, every 4 weeks for 52 weeks. The annualised relapse rate was lower for patients given DAC HYP (150 mg or 300 mg) than for those that received the placebo, However, no significance difference in terms of relapse between the different DAC doses was reported. Finally, the DAC analysis revealed the secondary effects of the DAC immunosuppressive therapy that can actually increase susceptibility to secondary infections.

The second set of experiments were devoted to study the effect of RRMS in pregnant patients. The mechanisms at the basis of a partial MS remission during the pregnancy are not fully understood yet, leading this case particularly interesting. During pregnancy, the maternal immunotolerance to the fetus induced Treg proliferation and reduced the relapse rate, therefore our model predicts a reduction of ODC damage. Our results are in line with what comes to us from biological knowledge and clinical observations since the resetting of the immune system is what significantly influences the course of the disease during pregnancy and also has been related with clinical manifestation of increased relapses associated with the post-partum period. From our results it is possible to appreciate the difference between the ODC cells irreversibly damages in the case of MS no pregnant patients and MS pregnant patients (specially in the case of max variation of the Treg-Teff balance). This difference increases from the first trimester to the time of delivery then returns to become not significant.

Chapter 5

Conclusions

All the cells composing an organism are surrounded by several levels of organisation (i.e. genome, epigenome, transcriptome, proteome) whose intersection gives rise to a variety of cellular phenotypes. The research of a signature in each level, whose elements can help in classifying groups of cells both in physiological and pathological contexts, is one of the main issues in biological and medical research. Deciphering the set of genes composing a signature could be a powerful method to identify and characterise cell type behaviour in physiological and pathological condition as cancer. The aim supporting this thesis project is to contribute to the research of gene signatures in several contexts that involve immune system and the great amount of cell types engaged. This objective was reached exploiting different deep sequencing techniques and developing bioinformatics algorithms and tools able to decipher the data obtained. In the first part, we focused our attention on DNA signature developing HashClone tool able to analyze the complex tumour clonal pattern of Lymphoma patients starting from the derivation of VDJ gene signatures from tumour lymphocytes. HashClone is an easy-to-use and reliable bioinformatics suite that provides B-cells clonality assessment and IGH-based MRD monitoring over time. The HashClone strategy to identify a set of putative clones is based on an alignment-free prediction method that identifies the set of putative clones belonging to the repertoire of the patient under study. The advantage of using an alignment-free prediction is twofold: (i) it may provide new rearrangements because no reference is used to select the putative clones (ii) it may be more robust to detect genome-scale events as rearrangements, recombination, and duplications. Moreover, the alignment-free prediction method provides an elevate accuracy, because the putative clones are identified through an integrated analysis of all the patient's samples collected over time. The candidate clones will be analysed to identify the germline origins of IGH rearrangements based on alignment of the putative B-cell clones with respect to the IMGT reference database (Giudicelli, Chaume, and Lefranc, 2004). Notice that the current tool implementation allows the users to exploit different datasets, leading to a broadly applications of HashClone to biological projects dedicated to the clonality detection from NGS data. To test its performances, we performed three studies:

1. Study 1: Simulated datasets to test HashClone performances
2. Study 2: Clonality analysis of cohort 1 (5 MCL patients)
3. Study 3: Clonality analysis of cohort 2 (23 MCL patients)

Our results showed that HashClone had good performance on simulated datasets (Study 1) in terms of execution times and compared with other tools results (Section 2.7.1). In Study 2 and 3 we confirmed HashClone performances on real patients cohorts (Section 2.7.2 and 2.7.3) since it was able to detect the major B-cell clone in both studies that were confirmed through the classical Sanger sequencing approach. Moreover, HashClone efficiently analyzed NGS data to monitor the MRD, providing highly comparable data with respect to the standardized ASO q-PCR. We also compared HashClone performance in the Study 2 with the

state-of-art tool, VIDJil (Giraud et al., 2014). Actually, Hashclone has two main distinct features with respect to VIDJil, the first is the reference free strategy, that allows Hashclone not to use biological knowledge until the last step in which it is necessary to assign to the putative clonotype the IGHV, IGHD and IGHJ composition. Secondly, Hashclone is specifically designed for the MRD detection working simultaneously on all sets of samples belonging to a patient. Instead, VIDJil is not specifically designed to simultaneously work on all samples. VIDJil provides a set of additional tools able to merge the set of rearrangements obtained from each sample generating the set of clones associated with the temporal trend.

In the second part of the thesis, we focused on RNA signature and we proposed a new workflow composed of rCASC and rMLSC. The workflow was thought to be a method able to retrieve clusters-specific gene-signature and improve biological data interpretation. The first phase, rCASC is a modular workflow for reproducible single cell data analysis with new features that could help researchers to define cell subpopulations and detect subpopulation specific markers. rCASC is composed of many pre-processing tools to remove low-quality cells and/or specific bias (e.g. cell cycle), different clustering techniques based on different distance metrics and a new metric (Cell Stability Score) for the evaluation of clusters quality. Moreover, rCASC uses Docker for ease of installation and to achieve a computation-reproducible analysis and a Java GUI is also provided to welcome users without computational skills in R. The second phase, rMLSC, is a new approach based on machine learning algorithms able to identify cluster-specific gene signatures starting from rCASC result. This task is performed exploiting classification algorithms and generating decision trees for each cluster. Then, the exploration and filtering of the trees through several *in-house* methods allows the identification of subsets of genes for each cluster (the gene signatures) whose disposition in the decision trees is a reflection of a difference among the clusters. The biological characterisation of these signatures, both individually and compared among the clusters, help the user for a deeper understanding of dataset composition as cellular types, putative hierarchical lineage of the cells or differences in cells behaviour upon different conditions. rCASC and rMLSC were tested on single cell dataset from Pace et al., 2018 characterized by immune cells at different state of differentiation. Overall, our combined workflow was able to identify the same results of the authors identifying four immune cellular types but adding a further layer of information regarding the putative hierarchical link among the cells. Moreover, the analysis performed by Pace et al., 2018 was actually strongly driven due to the use of a set of 24 immunological genes, while our workflow has allowed a greater biological comprehension of the cells contained in the dataset using all the set of top expressed genes as input of the classification algorithm and without any *a priori* biological knowledge. Thus, rCASC and rMLSC actually go further the classical pipeline of analysis whose ultimate goal is the derivation of DEGs that are often few explanatory of the biological meanings and underestimating the real complexity of scRNA-seq data.

Finally, in the last part of the thesis, a new methodology based on mathematical models was used to model complex biological scenario as Multiple Sclerosis (MS) disease. In detail, we provide a promising application of a computational framework based on Coloured Petri Nets and sensitivity analysis to perform *in silico* experiments helping to improve the understanding of the Relapse Remitting Multiple Sclerosis disease (RRMS), possibly giving some indications that may ameliorate the clinical management. We tested our approach in two different conditions: (i) patients affected by MS and (ii) pregnant women affected by MS and we inspected the effect of the daclizumab therapy in our model calibrated for reproducing the behaviours of RRMS patients in both conditions. In the first part, we showed how our approach was able to identify MS disease trend in patients under the effect of the DAC therapy and we also revealed the secondary effects of the DAC immunosuppressive therapy that can actually increase susceptibility to secondary infections, as reported in literature. In the second part, we showed how the simulation in the pregnancy in MS patients mimicked the

resetting of the immune system that significantly influences the course of the disease during pregnancy and also has been related with clinical manifestation of increased relapses associated with the post-partum period.

In conclusion, as future works we plan, for the first part, to improve HashClone pipeline with a further parallelization of the remaining phases, to make HashClone usable also for detecting signature in other types of cancer as solid tumours adapting the several code phases. For the second part, we will implement other classification algorithms in rMLSC or more in general to test different machine learning methods to improve the accuracy in the prediction of the RNA signature.
Finally, for the last part, we will expand the mathematical model of MS including different cell populations of the immune system for a better integration between adaptive and innate immunity. Moreover, we plan to insert new colour classes in the model to encode the spatial coordinates of all entities in a cubic tissue portion. In the context of precision medicine, we may also exploit the model to predict the patient-specific outcome when the DAC or other therapies are administered with an initial model calibration with patient-specific clinical data.

5.1 Contributions

Chapter 1
HashClone suite:

- Implementation of parallelization algorithm
- Implementation of the spike-in function
- All the computational analysis

Chapter 2
rCASC and rMLSC:

- Implementation of *cellrangerCount* function in rCASC
- Implementation of rMLSC pipeline
- All the computational analysis

Chapter 3
RRMS model:

- Supervisor of the biological aspects implemented in the model
- Interpretation of the results

Chapter 6

Supplementary Materials

6.1 Supplementary Materials Chapter 2

Sample processing and genomic DNA extraction For Diagnostic and follow up samples 5 ml of bone marrow (BM) and/or 10 ml of peripheral blood (PB) were collected in Sodium citrate or Lithium-heparine vacutainers. Cells recovery was performed using Ficoll density stratification and red blood cells lysis, obtaining mononuclear cells (MNCs) and total white blood cells (Total WBCs) for PB and BM respectively. Firstly, for MNCs isolations PB samples were diluted with NaCl 0,9% solution (ratio 1:2) and then stratified on Ficoll layer (Sigma-Aldricht; Germany). After density separation, performed as recommended by manufacturer's instructions, MNCs were isolated and stored at -80°C in dried pellets until genomic DNA (gDNA) extraction. WBCs were recovered from bone marrow (BM) mixed to lysis buffer 1X [NH4Cl, KHCO3 and EDTA (pH7,3)] (Qiagen, Germany) (ratio 1:2). After incubation at room temperature for 10 minutes, the samples were centrifuged twice at 1500 rpm for 15 and 10 minutes, respectively. Total WBCs were stored as dry pellets at -80°C until gDNA extraction.
gDNA extraction was extracted from MNCs and total WBCs using DNAzol reagent (Thermo Scientific). Briefly, MNCs and WBCs were lysed in 1 ml of DNAzol and then centrifuged. gDNA precipitation and washes were performed with 100-70% ethanol solutions. After the mixing of the sample by inversion, DNA should quickly become visible as a cloudy precipitate, it was removed with a pipette and it was put into a new Eppendorf. Finally, gDNA was eluted using NaOH 8 mM.
gDNA quantity (ng) and purity (OD ratio A260/A280 and A260/A230) were evaluated by Nanodrop2000 Spectrophotometer (Thermo Scientific). Housekeeping gene control amplification (p53 exon 8) was performed on MNCs and total WBC samples in order to assess gDNA integrity.

IGH rearrangements screening and MRD monitoring IGH rearrangements screening and MRD study were performed using both an NGS approach and the gold standard techniques, i.e. Sanger sequencing and ASO q-PCR.

Next generation sequencing approach The DNA libraries were prepared using 500 ng and 100 ng of gDNA by two-steps PCR approach: in the first round, the IGH regions were amplified using FR1 BIOMED II primers for Study 2 (Cohort 1 of 5 MCL patients)(Van Dongen et al., 2003) and primer for FR1 region from EuroClonality NGS for Study 3 (Cohort 2 of 23 MCL patients), modified with an universal Illumina adapter linker sequence (Brüggemann et al., 2019). While in the second PCR round, Illumina specific indexes (Illumina; Sigma-Aldrich) were incorporated to the first round PCR IGH amplicons [21]. After a Bioanalyzer QC control (Agilent), the purified PCR products were serially dilute and pooled to a final concentration of 9pM adding 10% PhiX. The sequencing run was carried out by Illumina

V2 kit chemistry 500 cycles PE on MiSeq platform. A polyclonal sample, called buffycoat DNA, and negative control (water or HELA cell line) were added to each run.

Sanger sequencing and ASO q-PCR approach Diagnostic gDNA was screened for IGH rearrangements using consensus primers (Leader and Framework Regions (FR) 1 and 2). Purified post PCR products were directly sequenced and analyzed using the IGH reference database published in IMGT/V-QUEST tool (http://imgt.org) [23]. MRD monitoring was conducted by ASO q-PCR on 500 ng of gDNA, using patient specific primers and consensus probes designed on Complementarity-Determining Region 2 (CDR2) sequences, on CDR3 and FR3 IGH regions, respectively [24]. MRD results were interpreted according to the ESLHO-Euro MRD guidelines [4].

How the τ choice affects HashClone performance As already highlighted in the Material and Methods section, the choice of an appropriated τ can impact on the capability of HashClone to identify clones; therefore here we report some considerations and suggestions to help the user in this task. First of all, the information associated with the biological experiment (e.g. the dilution factor in case of artificial experiments or an estimation of the residual tumour cells) should be exploited by the user to find an appropriate initial τ value. Then, τ value can be decreased to refine the HashClone solution. Indeed smaller τ values leads to a higher number of significant k-mers resulting in more specific clone signatures. For instance, in Pilot1 the initial τ value, as already stated, can be selected equal to 1 since for each MCL patient, a 10-fold standard curve was prepared diluting diagnostic sample in gDNA pool from five healthy donors. Then, smaller τ values can be chosen without severally influencing the major clone characterization. Instead, values of τ greater than 1 do not make sense due to the design of the experiments. An example is showed in Figure 6 in which the major clone MRD trends obtained decreasing τ from 1 to 0.25 for Patient A are reported and compared with ASO q-PCR result. As expected all these trends are similar and close to the ASO q-PCR trend.

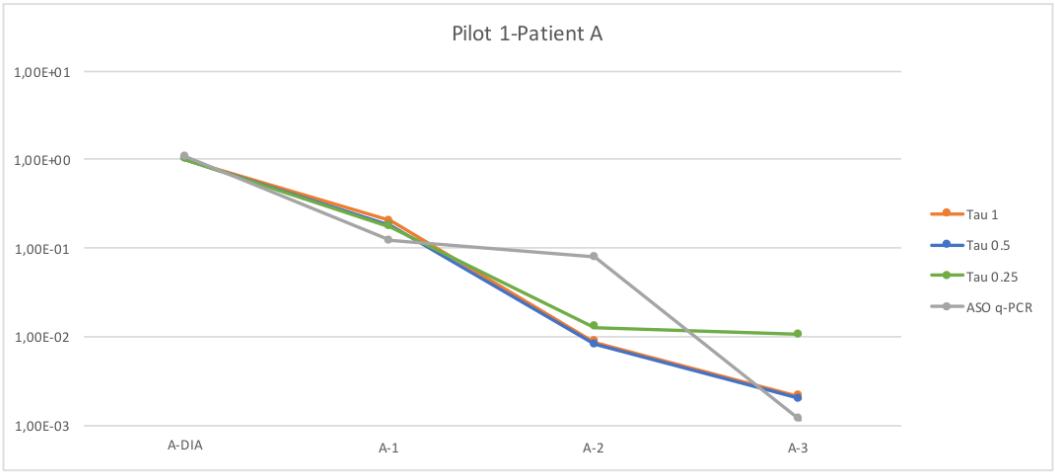


FIGURE 6.1: **Assesment of τ value.** The major clone trends obtained with different τ values on the data of the Patient A of Pilot1.

6.2 Additional figures and tables Chapter 2

Study	Patient	Number of clonotypes	Number of read associated with clonotype
Pilot 1	**A**	7	Major clone: 72698/74542 (98%) Other clones: 1844/74542 (2%)
	B	32	Major clone:46694/56964(82%) Others:10270/56964(18%)
	C	44	Major clone:88674/97018 (91%) Others:8344/97018 (9%)
	D	5	Major clone:104908/107808 (97%) Others:2900/107808 (3%)
	E	21	Major clone:185456/192947 (96%) Others: 7491/192947 (4%)
	Average value	**22**	**Major clone: 93%** **Others: 7%**
Pilot 2	**A**	18	Major clone: 99414/107825 (92%) Others: 8411/107825 (8%)
	B	72	Major clone: 169175/231281 (73%) Others: 62106/231281 (27%)
	E	5	Major clone: 85027/85603 (99%) Others: 576/85603 (0,1%)
	Average value	**32**	**Major clone: 88%** **Others: 12%**

Figure S1 - Clonotypes quantification by Hashclone

FIGURE 6.2: **Clonotype quantification by HashClone for Study 1.** Hashclone identifies an average number of clones equals to 21 in *Pilot1* and 32 in *Pilot2*. In the last column of the table is reported for each major clone the number of reads associates to it with respect to the total number of reads. The same data are also reported for the other clones identified.

			Phase A	Phase B
Study	**Patient (only diagnosis samples)**	**Clonotype identified**	**Clonotype with frequency > 100**	**Clonotype associated with VDJ**
Pilot 1	**A**	159	49	30
	B	171	51	31
	C	163	51	43
	D	129	50	42
	E	165	52	38
	Average value	**157**	**51**	**37**
Pilot 2	**A**	138	43	34
	B	196	72	50
	E	108	45	39
	Average value	**114**	**53**	**41**

Figure S2 - Clonotypes quantification by ViDJil

FIGURE 6.3: **Clonotype identification by ViDJil for Study 1.** The clonotypes identified by ViDJil in *Pilot1* and *Pilot2* are reported in the third column. In the fourth column are reported the clones passed the *Phase A* while in the fifth column there are the number of clones passed the *Phase B*.

Study	Patient	Number of clonotype	Number of read associated with Clonotype
Pilot 1	**A**	30	Major clone: 83974/86619 (97%) Other clones: 2645/86619 (3%)
	B	31	Major clone: 58620/65302 (90%) Others: 6682/65302 (10%)
	C	43	Major clone: 108386/121593 (98%) Others: 13207/121593 (2%)
	D	42	Major clone: 180075/183559 (98%) Others: 3484/183559 (2%)
	E	38	Major clone: 217524/217917(99%) Others: 393/217917 (1%)
	Average value	**37**	**Major clone: 96%** **Others: 4%**
Pilot 2	**A**	0	/
	B	0	/
	E	0	/

Figure S3 - Clonotypes quantification by ViDJil

FIGURE 6.4: **Clonotype quantification by ViDJil for Study 1.** ViDJil identifies an average number of clones equals to 37 in *Pilot1* while in *Pilot2* it does not identified any clonotypes. In the last column of the table is reported for each major clone the number of reads associates to it with respect to the total number of reads. The same data are also reported for the other clones identified.

Study	Patient	CDR3 Sanger Sequence	CDR3 Vidjil sequence	Homology
Pilot 1	A	GCGAGAGA**TCCA**GGGTATAGCAGTGGCTGGAA C**CTGGGA**TACTACTACTACGGTATGGACGTC	GCGAGAGA**TCCA**GGGTATAGCAGTGGCTGGAA C**CTGGGA**TACTACTACTACGGTATGGACGTC	100%(63/63nt)
	B	TGTGCGAGAAGCAATTTTGGAGTGG**TCTAAAT TACAT**GGACGTCT	TGTGC**NN**GAA**T**CAATTTTGGAGTGG**TCTAAAT TACAT**GGACGTCT	93%(42/45nt)
	C	CGAGAGA**TTACACAGCCCC**GGGTATAGCAGAA CCAGGC**CCC**T	CGAGAGA**TTACACAGCCCC**GGGTATAGCAGAA CCAGGC**CCC**T	100%(42/42)
	D	TGCGAGAGG**CGCGA**ATAACTGGAAC**CCCA**TTG ACTA	TGCGAGAGG**CGCGA**ATAACTGGAAC**CCCA**TTG ACTA	100%(36/36nt)
	E	GCGA**CCCAGCGAA**ATTACGATATTTTGACCGG **G**TTTGACTACT	GCGA**CCCAGCGAA**ATTACGATATTTTGACCGG **G**TTTGACTACT	100%(43/43nt)
Pilot 2	A	GCGAGAGA**TCCA**GGGTATAGCAGTGGCTGGAA C**CTGGGA**TACTACTACTACGG	CACGGNTGTGTATTNNTGTGCNNGNANNNNNG NGTNTANNNGTGNNNGNNANCNNGNANNCTNN	0%(0/59nt)
	B	TGTGCGAGAAGCAATTTTGGAGTGG**TCTAAAT TACAT**GGACGTCT	GNAAAACGACGGCCAGTTGGATCCGTCAGCCC CCAGGGAAGGTCACCGTCTCCTCAGGTAAGCC CTATAGTGAGTCGTATTA	0%(0/45nt)
	E	GCGA**CCCAGCGAA**ATTACGATATTTTGACCGG **G**TTTGACTACT	GNAAAACGACGGCCAGTTTGGGTGCGACAGGC CCCTGGACAAGGGCTTGAGTGGNTGGGATGGA	0%(0/43nt)

FIGURE 6.5: **ViDJil and Sanger Sequence comparison for Study 1.** Nucleotide alignments between the complementary region 3 sequences (CDR3, indicated in bold and underline) Sanger sequence and the sequence identified by ViDJil.

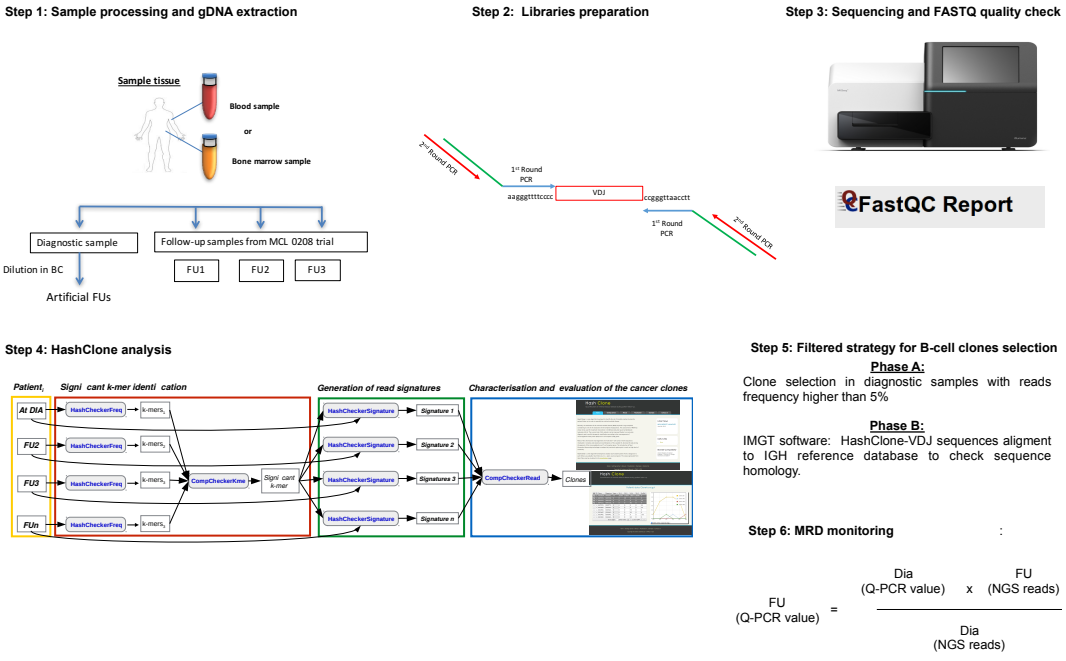


FIGURE 6.6: **The whole experimental and computational methodology**

6.3 Supplementary materials for Chapter 3.

6.3.1 Cell stability score implementation details

Stability score Algorithm Let be C the count matrix with $N \times M$ dimension where N is the gene number and M is cell number.

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots \\ \vdots & \ddots & \\ c_{N1} & & c_{NM} \end{bmatrix}$$

Then, we define C^p the matrix generated by p^{th} permutation removing q random columns from the original matrix C. Moreover considering the p^{th} permutation we denote L^p the set of all the removed cells in p^{th} permutation with $|L^p| = q$ and $\mathbf{cl}^p$ the vector with length $M - q$ encoding the relation between cells and clusters in p. Hence, $\mathbf{cl}^p_i$ identified the cluster in which the ith cell is inserted in permutation p. Finally, we use the notation $\mathbf{cl}^C$ for indicating the output of the clustering algorithm obtained by all the cells (i.e. matrix C).

The relation symmetric matrix R^p with dimension $M \times M$ is defined as follows:

$$R^p = \begin{bmatrix} r^p_{1,1} & r^p_{1,2} & \dots \\ \vdots & \ddots & \\ r^p_{M,1} & & r^p_{M,M} \end{bmatrix}$$

where $r^p_{i,j}$ is:

$$r^p_{i,j} = \begin{cases} 1 \; if \; \mathbf{cl}^p_i = \mathbf{cl}^p_j \\ 0 \; otherwise \end{cases} \tag{6.1}$$

Similarly we defined R^C the relation symmetric matrix obtained considering all cells and R^{C-L^P} as the relation symmetric matrix R^C in which the columns and the rows associated with cells in L^p are removed. Observe that the sum $R^{C-L^P} + R^p$ will always gives as results a matrix with 3 possible values: 0,1 or 2.

$$R^{C-L^p}[j,i] + R^p[j,i] = \begin{cases} 1 & if \; cell \; i \; and \; cell \; j \; clustered \; together \; in \; R^{C-L^p} \\ & or \; in \; R^P \; only; \\ 2 & if \; cell \; i \; and \; cell \; j \; are \; always \; in \; the \; same \; cluster; \\ 0 & if \; cell \; i \; and \; cell \; j \; are \; never \; in \; the \; same \; cluster. \end{cases}$$

Let $\delta(i,k)$ be the kronecker delta, a function of two variables that return 1 if the variables are equal, and 0 otherwise:

$$\delta(i,k) = \begin{cases} 1 \; if \; i = k \\ 0 \; otherwise \end{cases} \tag{6.2}$$

then we define the function *length* that counts the occurrence of a value k fixing the row j in the matrix $R^{C-L^P} + R^p$.

$$length(j,p,k) = \sum_{i=1}^{M} \delta(R^{C-L^P}[j,i] + R^p[j,i], k) \tag{6.3}$$

We use the function *length* to count the occurrence of 1 or 2 in in the matrix $R^{C-L^P} + R^p$.

Finally, we define the permutation score $pscore_{j,p}$ as:

$$pscore_{j,p} = \frac{length(j,p,2)}{length(j,p,2) + length(j,p,1)} \tag{6.4}$$

where p is a permutation and j a cell. Then, this returns the percentage of cells initially clustered with cell j that remain clustered with cell j in the permutation p.

Then, we define $tscore_{j,s}$ as follow:

$$tscore_{j,s} = \frac{1}{P} \sum_{p \in P} 1_{pscore_{j,p} >= s} \tag{6.5}$$

where P is the total number of permutations and s is user-defined threshold. This metric compute the probability that a cell j is always clustered with the same set of cells given that $pscore_{j,p} >= s$.

6.3.2 Example

Be $\mathbf{cl} = \{1 \quad 2 \quad 2 \quad 1 \quad 2 \quad 1\}$
Be $\mathbf{L} = \{6 \quad 2 \quad 2 \quad 4\}$
Be $\mathbf{cl}^1 = \{1 \quad 2 \quad 1 \quad 1 \quad 2\}$
Be $\mathbf{cl}^2 = \{1 \quad 2 \quad 1 \quad 2 \quad 2\}$
Be $\mathbf{cl}^3 = \{1 \quad 2 \quad 1 \quad 1 \quad 2\}$
Be $\mathbf{cl}^4 = \{1 \quad 2 \quad 2 \quad 1 \quad 2\}$
$\forall\ p \in \{1,2,3,4\},\ R_p$ is calculated
for instance hereafter I reported R and R^1

$$R = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$R^1 = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

$\forall\ p\ R^{p'} + R^p$ *is calculated*
for instance hereafter I reported $R^{p'} + R^1$

$$R'^1 + R^1 = \begin{bmatrix} 2 & 0 & 1 & 2 & 0 \\ 0 & 2 & 1 & 0 & 2 \\ 1 & 1 & 2 & 1 & 1 \\ 2 & 0 & 1 & 2 & 0 \\ 0 & 2 & 1 & 0 & 2 \end{bmatrix}$$

$\forall\ p,\ pscore$ *is evaluated with* $S = 0.6$ for instance hereafter I reported $pscore_1$

$$pscore_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

This means that in permutation 1 the third cell is unstable "jumping" from cluster number 1 to cluster number 2 in P. $\text{tscore}_{m,s}$ is evaluated then for each cell $tscore_s = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0.25 \\ 0.25 \\ 0 \end{bmatrix}$

In this Example number of permutation is 4, for statistical relevance, an higher number of permutation is required.

6.3.3 Decision tree

A decision tree is a flowchart-like structure where an internal node represents *feature (or attribute)*, the branch represents a *decision rule*, and each leaf node represents the outcome of the classification of the *sample instances* (Hunt, Marin, and Stone, 1966). The topmost node in a decision tree is known as the root node. Decision tree learns to partition on the basis of the attribute value in recursively manner (i.e. recursive partitioning). The paths from root to leaf represent classification rules. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery (Hunt, Marin, and Stone, 1966).

Random Forest Random forests (RF) algorithm was developed by (Breiman, 2001) and it is a supervised classification learning algorithm. In general, supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs (Russell and Norvig, 2016). RF are built by combining the predictions of several decision trees, each of which is trained in isolation. While in standard trees, each node is split using the best split among all variables, in a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. Indeed, during the construction of the individual trees in the RF, randomization is applied when selecting the best node to split on. Typically, this is equal to $\sqrt{M}$, where M is the number of features (or attributes) in the data set (Fawagreh, Gaber, and Elyan, 2014). Each tree in the ensemble acts as a base classifier to determine the class label of an unlabeled instance and finally the predictions of the trees are combined through averaging. This is done via majority voting where each classifier casts one vote for its predicted class label, then the class label with the most votes is used to classify the instance (Fawagreh, Gaber, and Elyan, 2014).

Breiman, 2001 introduced additional randomness during the construction of decision trees using the *classification and regression trees (CART) technique*. In this case, the subset of features selected in each interior node is evaluated with the **Gini index heuristics**. The feature with the highest Gini index is chosen as the split feature in that node Breiman et al., 1984. The index is a function that is used to measure the impurity of data, that is, how uncertain there is if an event will occur. In classification, this event would be the determination of the class label. In its general form, it can be calculated as:

$$Gini(t) = 1 - \sum P(C_i|t)^2 \tag{6.6}$$

In order to perform a classification with RF algorithm, it is often suggested to divide the initial dataset containing the instances in several subsamples called *training and test dataset*. The *training dataset* is the subset of instances used to fit the optimal parameters of the model. Statistically, it is suggested that training set is likely to have about 64% of instances (Fawagreh, Gaber, and Elyan, 2014). Successively, the fitted model is used to predict the responses for the observations in a second dataset called the *test dataset* that usually contains the remaining 36% of the instances. This approach of training/test set allows to compare different models in an unbiased way, by basing the comparisons in data that were not use in any part of your training selection process.

6.3.4 Weka tool

Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization (Witten et al., 2016). Weka is open source software developed in JAVA and issued under the GNU General Public License. Weka provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning. As well as a wide variety of learning algorithms, it includes a wide range of preprocessing tools. This diverse and comprehensive toolkit is accessed through a common graphic interface so that its users can compare different methods and identify those that are most appropriate for the problem at hand (Witten et al., 2016). Weka's main user interface is the Explorer, but essentially the same functionality can be accessed from the command line.

The Explorer interface features several panels providing access to the main components of the workbench (Figure 6.7 from Witten et al., 2016):

- the Preprocess panel has facilities for importing data from a comma-separated values (CSV) file or ARFF file and for preprocessing this data using a filtering algorithm. These filters can be used to transform the data (e.g. turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria.
- the Classify panel enables applying classification and regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, receiver operating characteristic (ROC) curves or the model itself. Weka implements several classification algorithm as ADTree, DecisionStump, J48, RandomForest, RandomTree, REPTree, SimpleCart and others.
- the Associate panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data.
- the Cluster panel gives access to the clustering techniques in Weka (e.g. the simple k-means algorithm).
- the Select attributes panel provides algorithms for identifying the most predictive attributes in a dataset.
- the Visualize panel shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

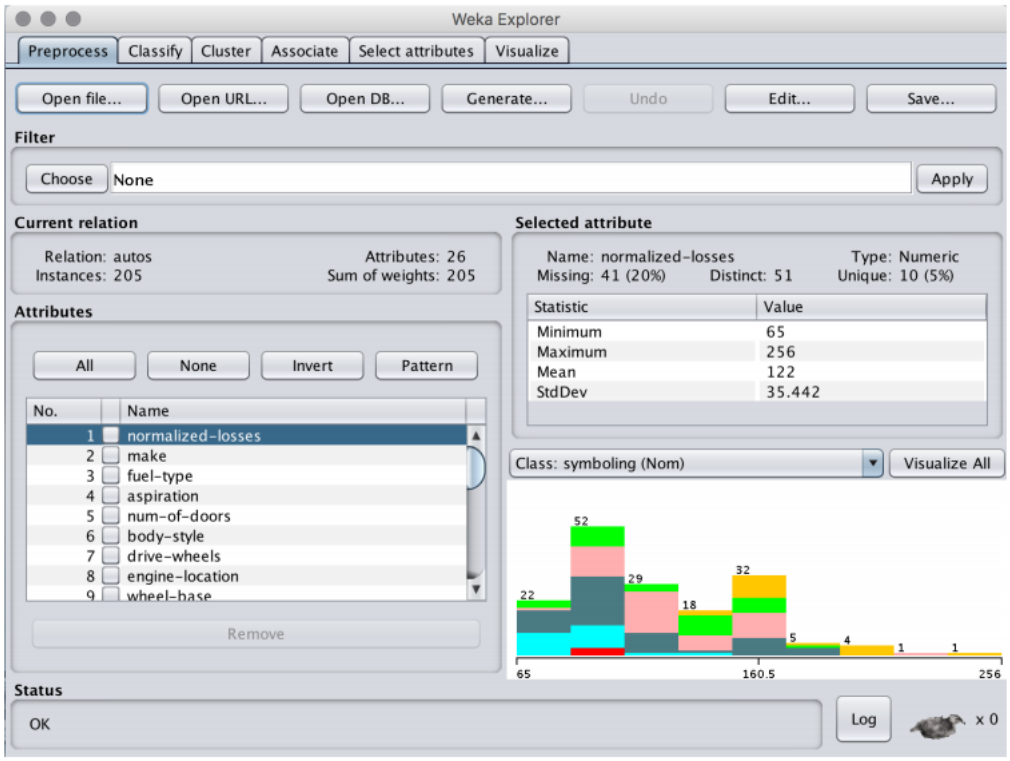


FIGURE 6.7: **WEKA's main graphical user interface.** Figure from Witten et al., 2016.

Classification analysis through Weka tool The input data that can be analyzed with WEKA can be presented as spreadsheet or ARFF format. When using training and test set, training ARFF file contains a first list of all the attributes in the upper part of the file. While, in the bottom of the file the attribute values separated by commas with last column indicating the correct classification label are reported for the instances considered in the training set (around 64% of the total dataset). The test ARFF file contains again the list of all attributes in the upper part and then the attribute values of the instances considered in the test set (around 36% of the total dataset) with last column replaced with question mark.

The classifiers in WEKA are divided into Bayesian classifiers, trees, rules, functions, lazy classifiers and a final miscellaneous category. Several trees methods are available in WEKA tool: Decision tree J48 (e.g. implementation of algorithm ID3 (Iterative Dichotomiser 3), developed by the WEKA project team), Random Forest, Random Tree, DecisionStump, REPTree and more. Once choose the optimal trees classifier, several parameters can be selected:

- percentage of the training set size (default 100)
- number of iterations (default 100)
- number of execution slots (default 1 - i.e. no parallelism)
- number of attributes to randomly investigate (default 0)
- set minimum number of instances per leaf (default 1)
- set minimum numeric class variance proportion of train variance for split (default 1e-3)
- seed for random number generator (default 1)

- the maximum depth of the tree, 0 for unlimited
- number of folds for backfitting (default 0)
- allow unclassified instances.
- break ties randomly when several attributes look equally good.
- the desired batch size for batch prediction (default 100).

Figure 6.8 shows the full output of a classification analysis using J48 decision tree. The first part of the output is a summary of the dataset with a pruned decision tree in textual form. In the tree structure, a colon introduces the class label that has been assigned to a particular leaf, followed by the number of instances that reach that leaf, expressed as a decimal number because of the way the algorithm uses fractional instances to handle missing values. If there were incorrectly classified instances their number would appear means that two instances reached that leaf, of which one is classified incorrectly. Beneath the tree structure the number of leaves is printed; then the total number of nodes (Size of the tree). The next part of the output gives estimates of the tree's predictive performance reporting the number and the percentage of correctly and incorrectly classified instances. As well as the classification error, the evaluation module also outputs the Kappa statistic (i.e. metric that compares an Observed Accuracy with an Expected Accuracy (random chance)), the mean absolute error, and the root mean-squared error of the class probability estimates assigned by the tree. The root mean-squared error is the square root of the average squared loss. The mean absolute error is calculated in a similar way using the absolute instead of the squared difference. It also outputs relative errors, which are based on the prior probabilities. Finally it is also reported the per-class average of each statistic, weighted by the number of instances from each class. The confusion matrix at the bottom is also reported for summarizing the performance of the classification algorithm.

6.4 Additional figures and tables Chapter 3

Visualization of gene signatures for each cluster In all the heatmaps rows represent genes of one signature and columns represent of the cells dataset (i.e. n=1600). The clustering of the rows on the heatmaps is also shown as a dendrogram to highlight the hierarchical relationship between genes. Cells in the columns are clustered together on the basis of their belonging cluster (i.e. cluster 1 in red, cluster 2 in yellow, cluster 3 in green, cluster 4 in blue and cluster 5 in violet). Color key is a z-score (i.e. score calculated subtracting the mean and then divide by the standard deviation) with value from -3 to +3. Figure 6.10 shows the expression value of the gene signature of cluster 1, that is composed of 15 genes in the dataset. Red cells (i.e. cluster 1) show similar downregulation of genes in the upper part of dendrogram while bottom genes determine intra-cluster differentiation among wt (i.e. cells on the left) and ko (i.e. cells on the right)of cluster 1. On the opposite, cluster 2 (i.e. yellow) shows this gene signature mostly upregulated while cluster 3 (i.e. green) displays mostly downregulation. Gene signature expression in cluster 4 (i.e. blue) shows distinct behaviour since the major part of the gene signature is not expressed except for some cells turning towards a light downregulation. Finally cluster 5 display the most evident expression of the signature since the majority of cells shows gene signature upregulation. It is worth to note that there is only one gene, Tmsb10, that allow a precise intra-cluster differentiation among wt and ko cells in all the clusters. Figure 6.11 shows the gene signature (i.e. 71 genes) of cluster 2 (yellow) in the cell dataset. It is quite evident how cells of cluster 2 show a low expression of the gene signature and no intra-cluster division is present. However, cluster

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

outlook = sunny
|   humidity <= 75: yes (2.0)
|   humidity > 75: no (3.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)

Number of Leaves  : 5

Size of the tree : 8

Time taken to build model: 0.27 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           9               64.2857 %
Incorrectly Classified Instances         5               35.7143 %
Kappa statistic                          0.186
Mean absolute error                      0.2857
Root mean squared error                  0.4818
Relative absolute error                 60      %
Root relative squared error             97.6586 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.778    0.6        0.7       0.778     0.737      0.789    yes
                 0.4      0.222      0.5       0.4       0.444      0.789    no
Weighted Avg.    0.643    0.465      0.629     0.643     0.632      0.789

=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 3 2 | b = no
```

FIGURE 6.8: **Example of an output from the J4.8 decision tree learner**

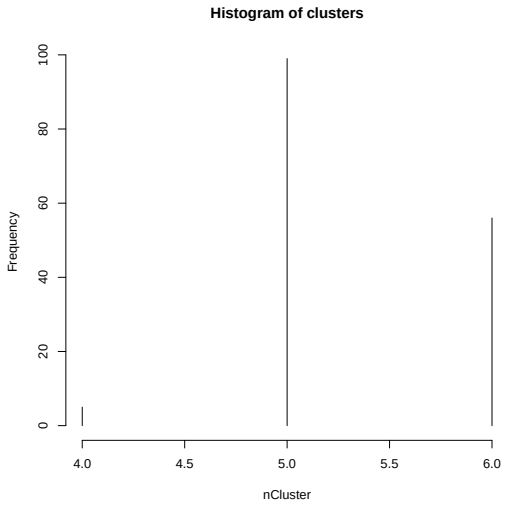


FIGURE 6.9: **Clusters detected by nClusterEvaluationSIMLR**

5 shows again the most evident expression of all the genes while cluster 1 and 3 have the same behaviour and turn towards the downregulation. Cluster 4 instead shows no expression of this gene signature. From the figure it is also visible that some blocks of genes show light/strong difference in the expression inside each cluster. For example, cluster 1 in red shows cells with a detectable expression for Cd3d, Cd3g, Uqcrh, Uba52 and Fau genes as well as mitochondrial genes block in the bottom part of the figure. In the same way, cluster

4 in blue shows the same mitochondrial genes block and few other genes that shift toward a downregulation. Finally, Ccl5 gene is the only gene with no expression in cluster 5 (violet) but detectable in cluster 4 (violet).

Figure 6.12 displays the gene signature (i.e. 9 genes) of cluster 3 (green). All these genes are contained in the cluster 1 gene signature however it is possible to better appreciate the differences of expression due to the lowest number of genes. As for Figure 6.10 cluster 3 shows a not detectable expression of the own gene signature. Again, Tmsb10 is the gene that allows to obtain a strong intra-cluster expression difference among wt and ko in all the clusters as well as Fau and Tpt1 (for cluster 1).

Figure 6.13 represents the gene signature of cluster 4 (violet) composed of 41 genes. This gene signature shows a similar pattern as figure 6.11 since cluster 1 (red) and 3 (yellow) have, for the upper genes, no detectable expression, cluster 4 (blue) shows a slight presence of the signature followed by cluster 2 (yellow) and cluster 5 (violet). Again, in the signature are present genes with strong expression differences among the clusters as Ccl5 and Tmsb10 with the latter giving also intra-cluster differences. However, in the bottom part of the signature, 11 genes belonging to the mitochondrial blocks but also others as Eef1a1, Rack1, Eef2, Hspa8 and Naca give intra-cluster differences in particular in cluster 1 and 3.

Finally, figure 6.14 displays the signature of cluster 5 (violet) composed of 151 genes, 81 of these uniques of this signature. The expression pattern is repeated among the clusters, however some differences are well visible. Genes confering intra-cluster division are Pabpc1, Uba52, Fau, mytochondrial genes, Rack1, Eef1a1, Tpt1, Npm1, Eef1b2 and Ddx5 (in cluster 1), Tmsb10 (in all the clusters). A distinctive trait in this signature is the presence of 7 genes whose expression is not detectable in any cluster except for cluster 5 (violet) and it is composed of Stmn1, Pclaf, Tubb5, H2afz, Ran and Ptma. Moreover, Ccl5 gene expression is again heterogeneous with no expression in cluster 5, downregulation in cluster 1 and 3 and strong upregulation in cluster 2 and 4.

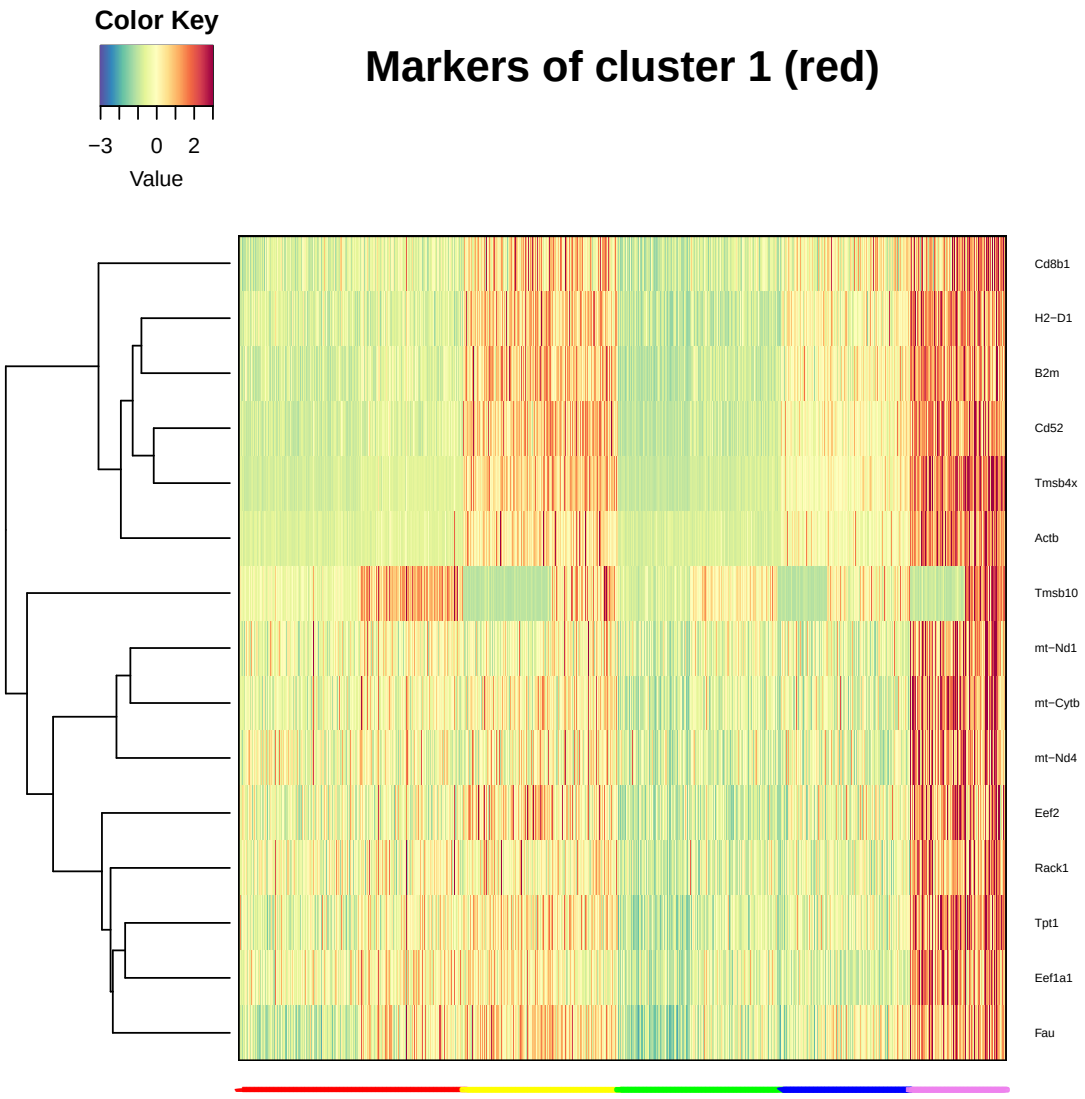


FIGURE 6.10: **Heatmap of the 15 genes of cluster 1 signature.** Columns represent cells and rows genes. Cells are coloured on the basis of their belonging cluster (C1:red, C2:yellow ,C3:green, C4:blue, C5:violet)

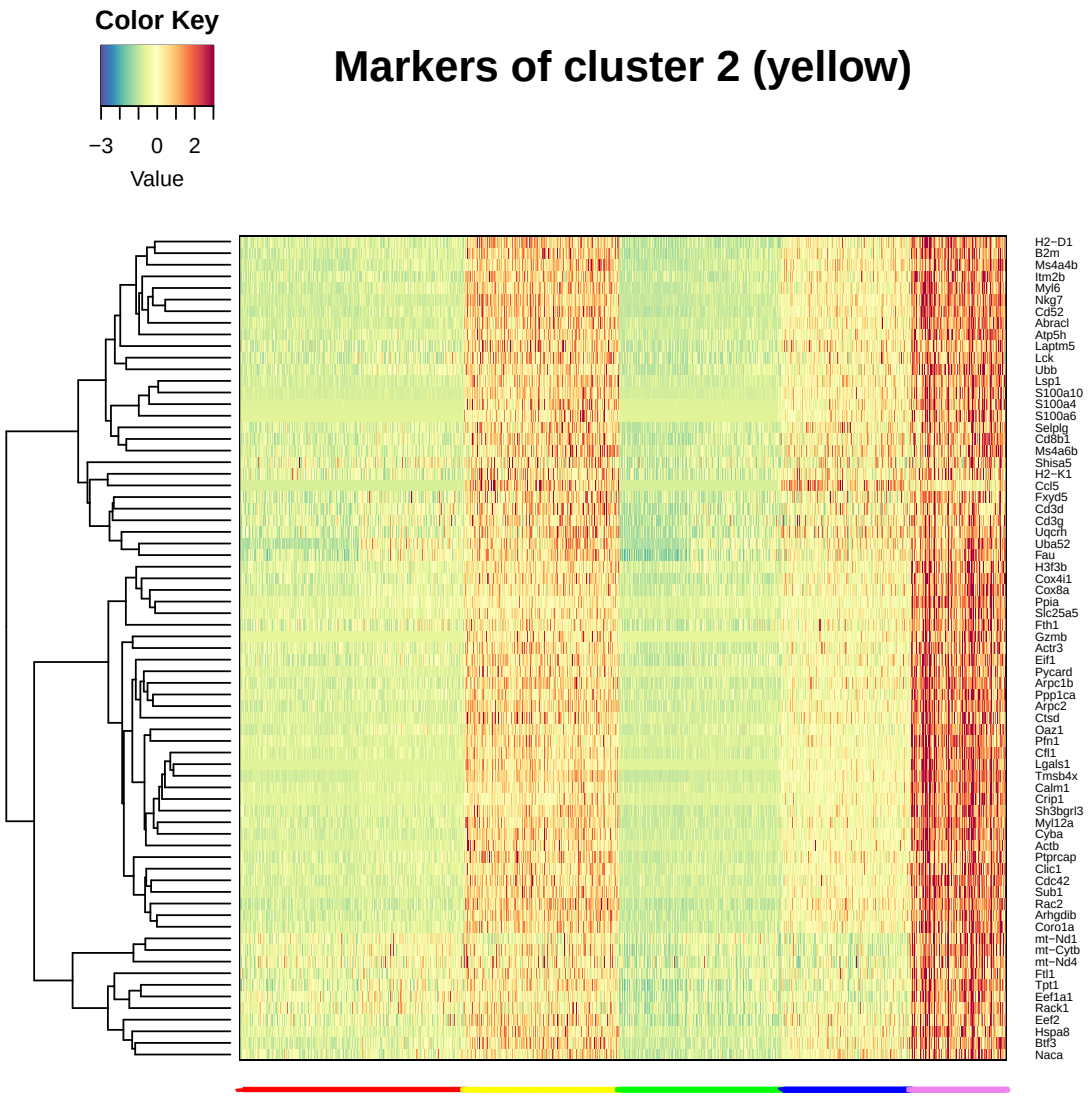


FIGURE 6.11: **Heatmap of the 71 genes of cluster 2 signature.** Columns represent cells and rows genes. Cells are coloured on the basis of their belonging cluster (C1:red, C2:yellow ,C3:green, C4:blue, C5:violet)

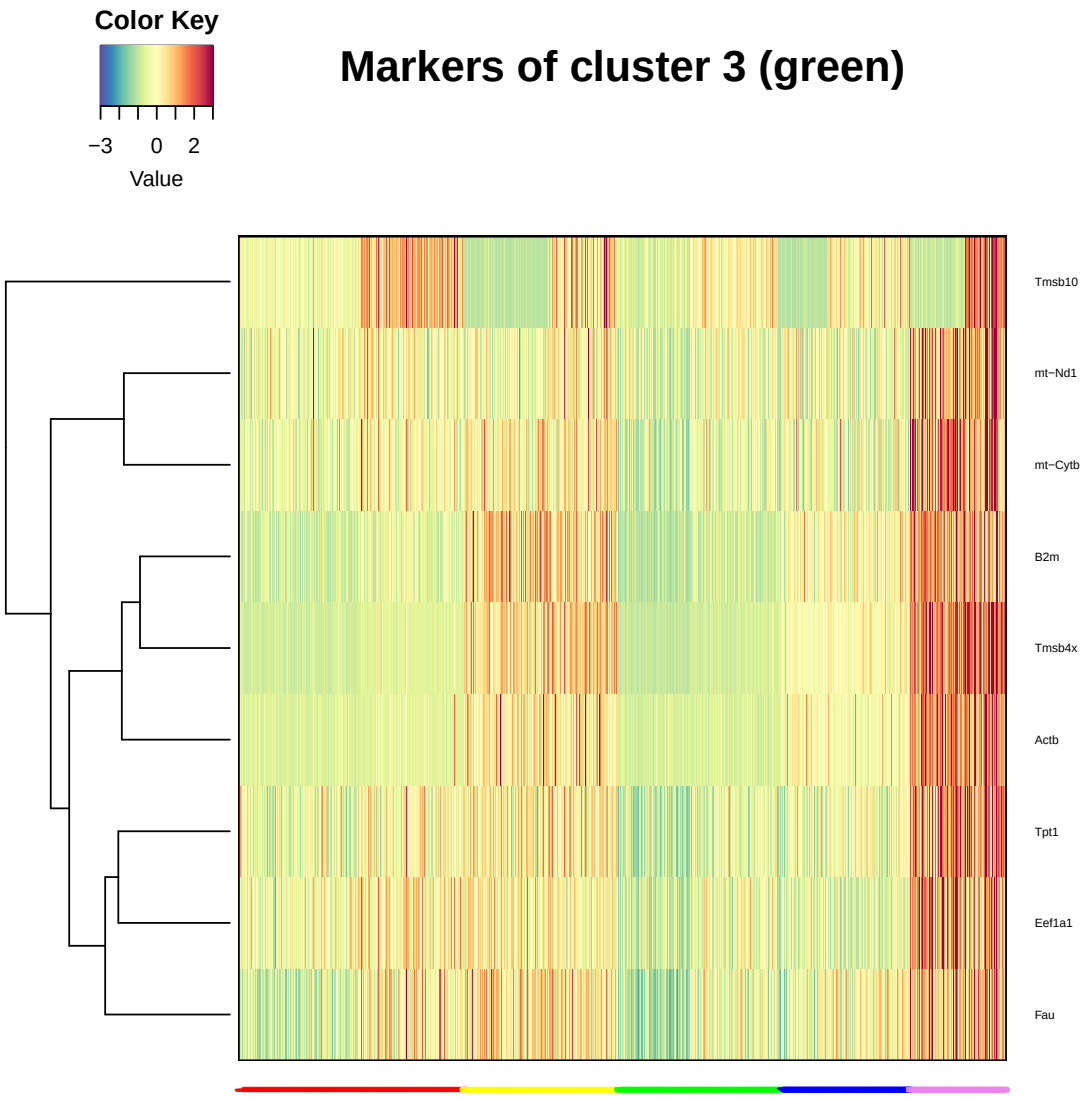


FIGURE 6.12: **Heatmap of the 9 genes of cluster 3 signature.** Columns represent cells and rows genes. Cells are coloured on the basis of their belonging cluster (C1:red, C2:yellow ,C3:green, C4:blue, C5:violet)

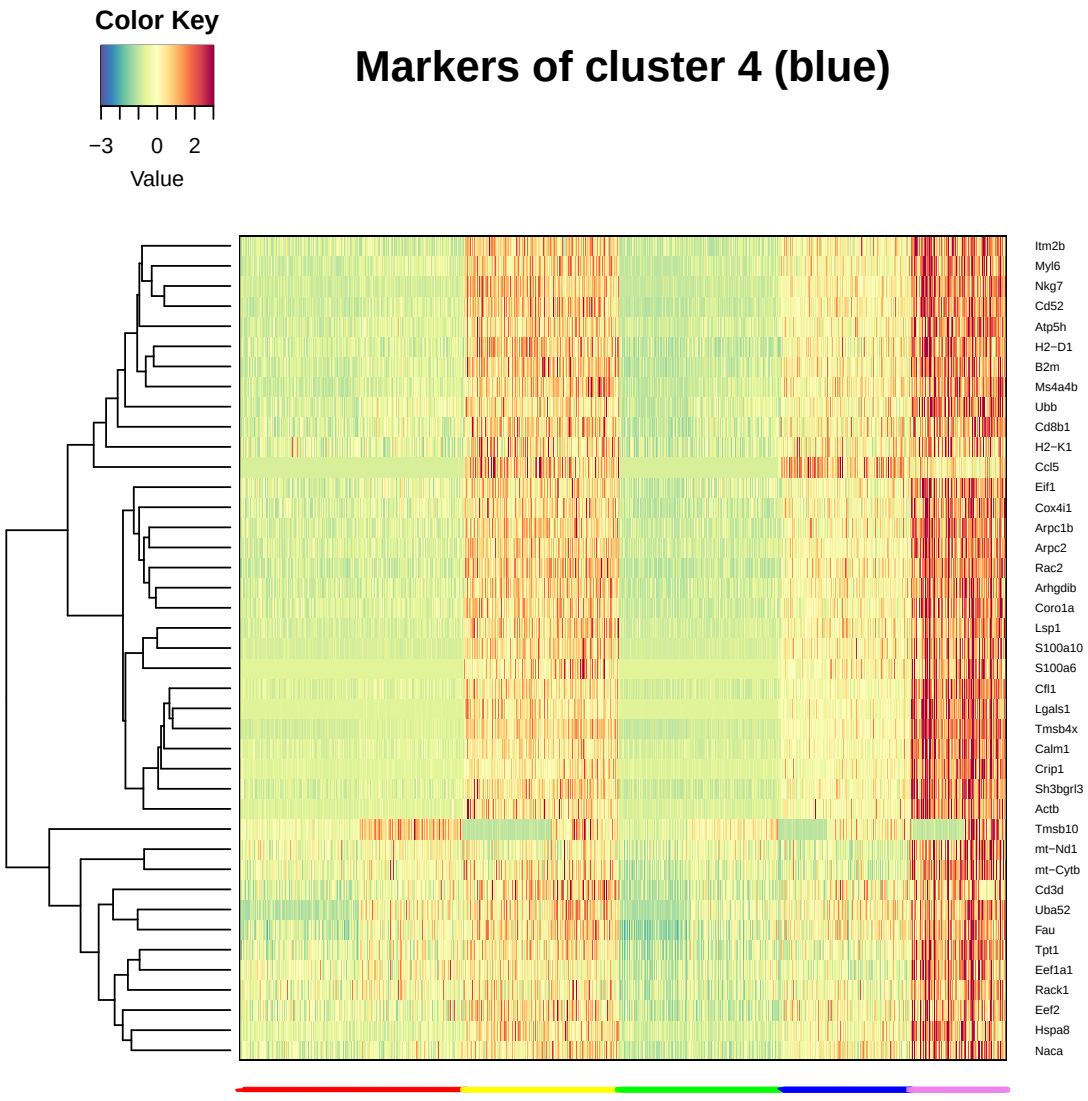


FIGURE 6.13: **Heatmap of the 41 genes of cluster 4 signature.** Columns represent cells and rows genes. Cells are coloured on the basis of their belonging cluster (C1:red, C2:yellow ,C3:green, C4:blue, C5:violet)

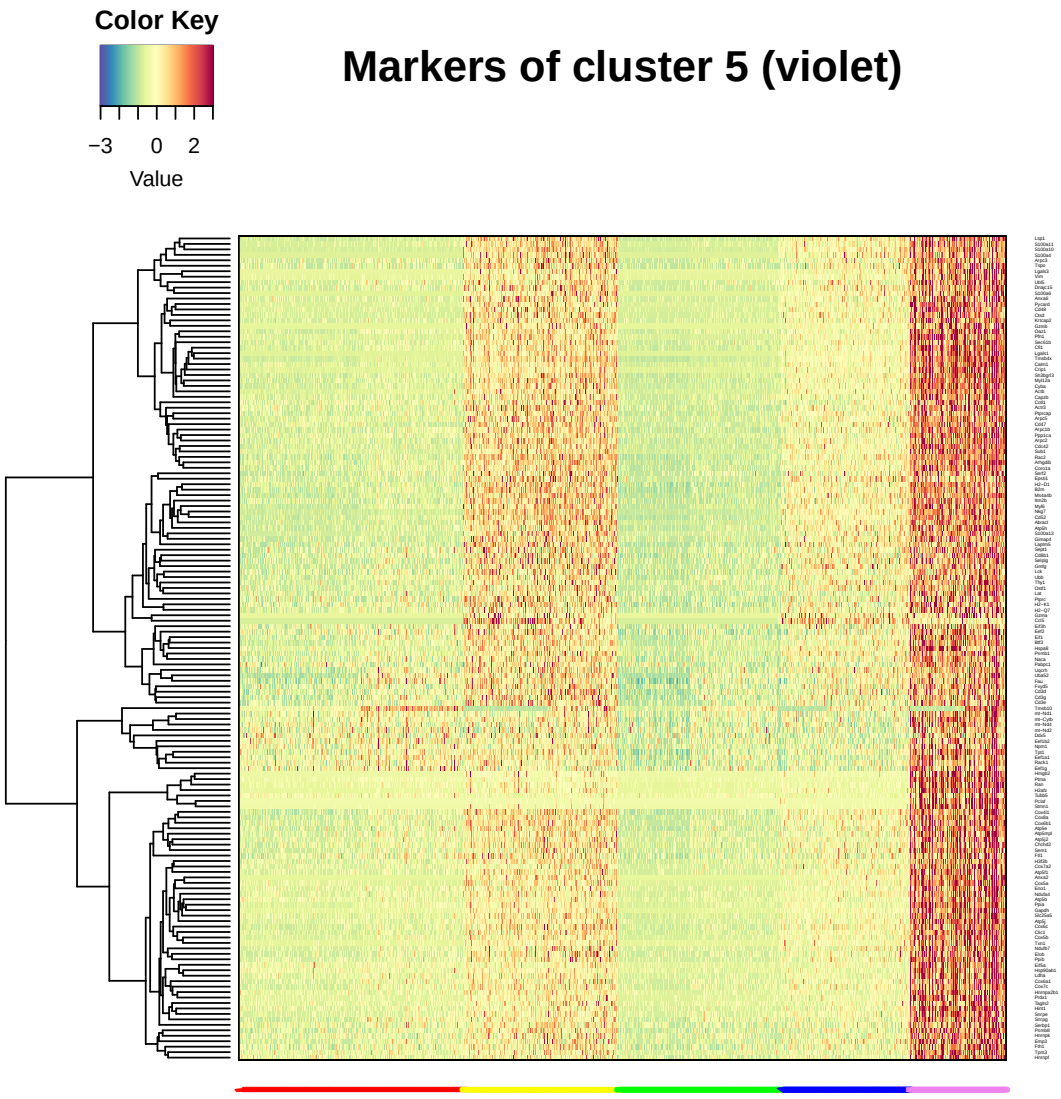


FIGURE 6.14: **Heatmap of the 151 genes of cluster 5 signature.** Columns represent cells and rows genes. Cells are coloured on the basis of their belonging cluster (C1:red, C2:yellow ,C3:green, C4:blue, C5:violet)

6.5 Supplementary Figure and Tables Chapter 4

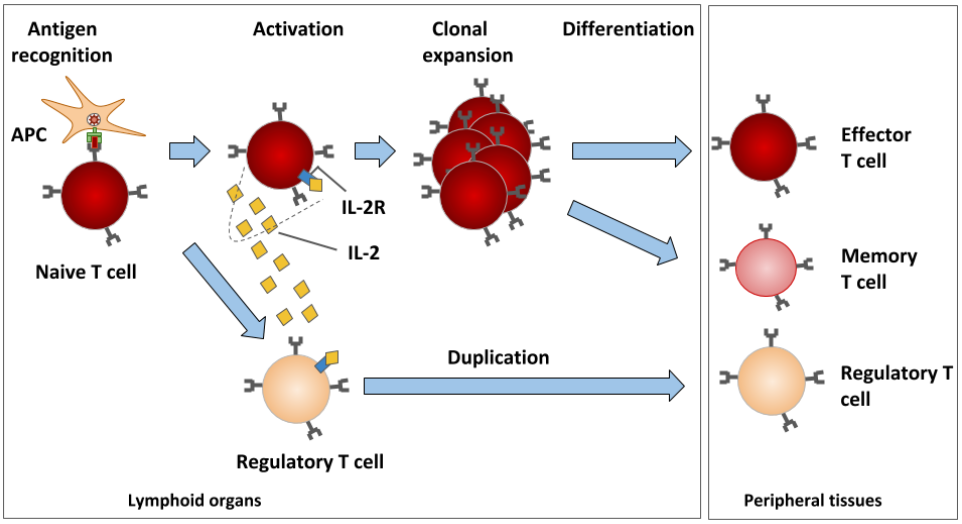


FIGURE 6.15: **T Lymphocytes activation.** The naive T Lymphocytes are activated through the bound of the TCR receptor with an Antigen Presenting Cell (APC). Then, T Lymphocytes start to produce and release IL2 in the environment. IL2 is simultaneously internalized by T Lymphocytes in order to self-stimulate the duplication and differentiation in Effector T cells (Teff), Memory T cells (Tmem) and Regulatory T cells (Treg).

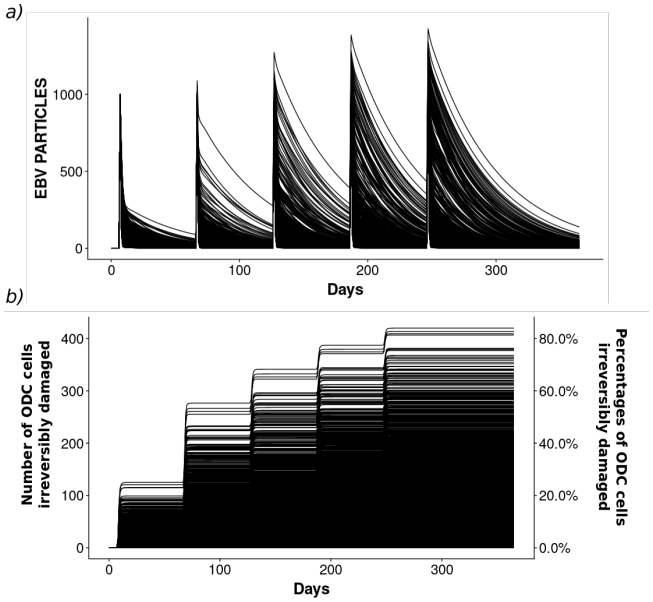


FIGURE 6.16: **EBV and ODC dynamics.** A subset of the 5000 trajectories generated by LHS of the EBV cells (a) and the ODC cells with an irreversible damage (b) over the whole time interval

[1]DACinjected represents the quantity of DAC injected per time and with this formula we estimate automatically the constant in order to have $exp(-DACinjected/C_{DAC}) = .1$, i.e. the T-cells duplication rate is reduced of the 90% when all the DAC particles are present.

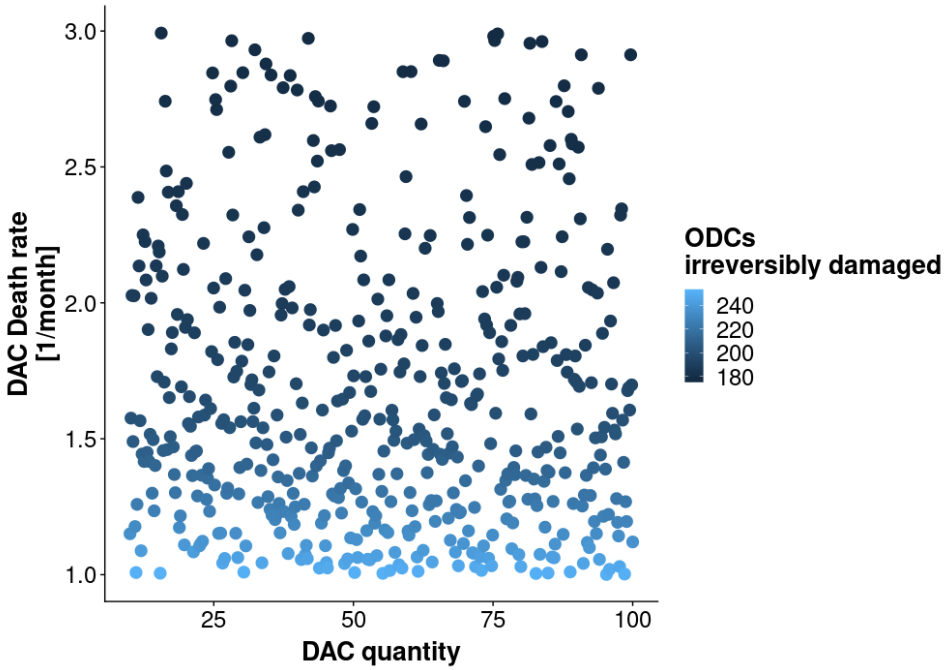


FIGURE 6.17: **Varying the quantity of DAC injected and its degradation.** Scatter plot of the ODC irreversibly damaged variable at the fixed time 365 depending on the DAC injected (x-axis) and the DAC death rates (y-axis). The colour depends on the number of ODCs irreversibly damaged. The number of ODC is strongly dependent by the DAC degradation: the decrease of the number of damaged ODC is more influenced by an increase of the permanence time of DAC drug in the body.

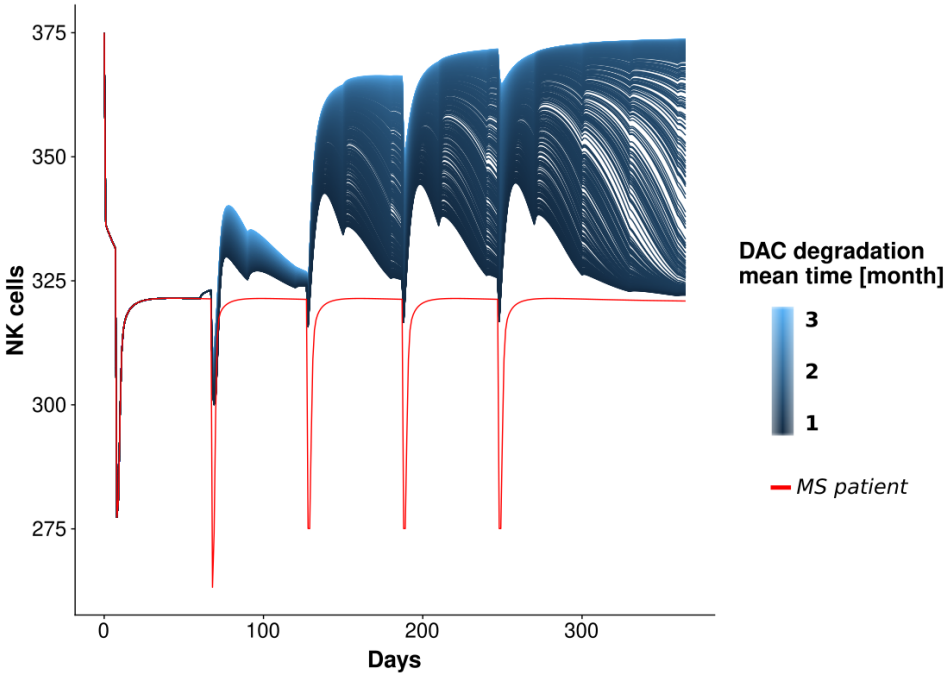


FIGURE 6.18: **Result of the variation of DAC degradation rate in NK cells.***NK* trajectory colored depending on *DAC* degradation rate (expressed in months). The red line represents the starting sample without drug administration.

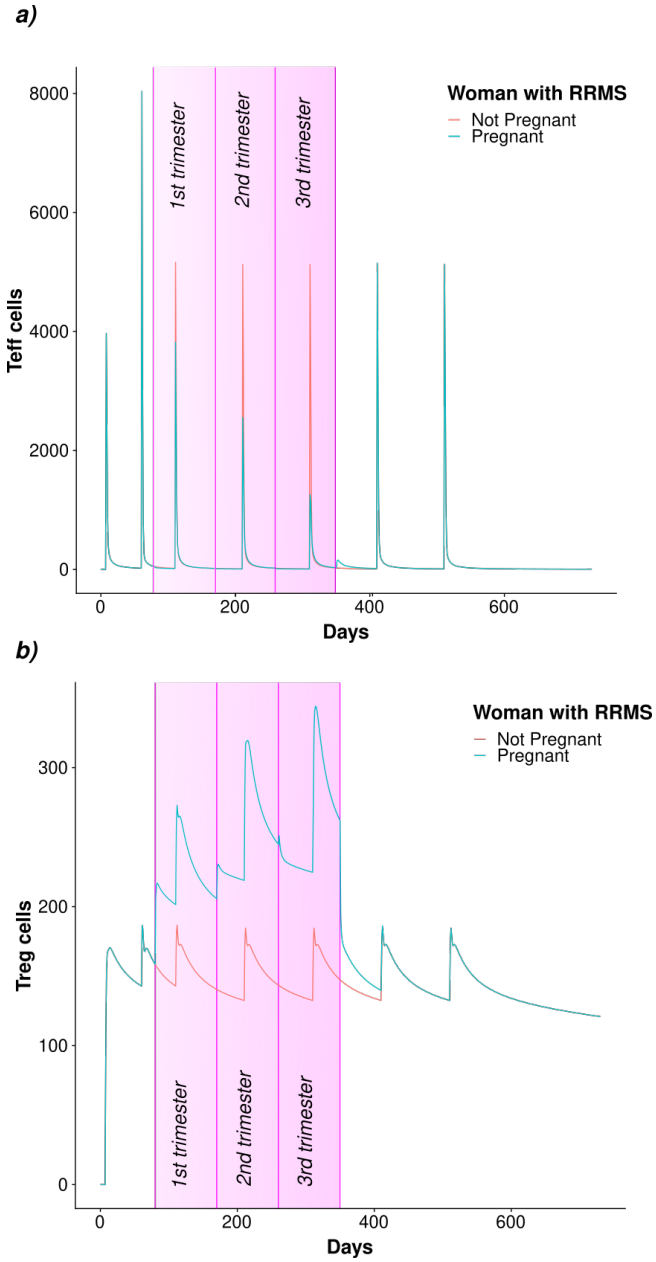


FIGURE 6.19: **Pregnant woman case: Teff and Treg.** The Teff a) and Treg b) dynamics before, during and after the pregnancy. The red line represents the starting sample without pregnancy.

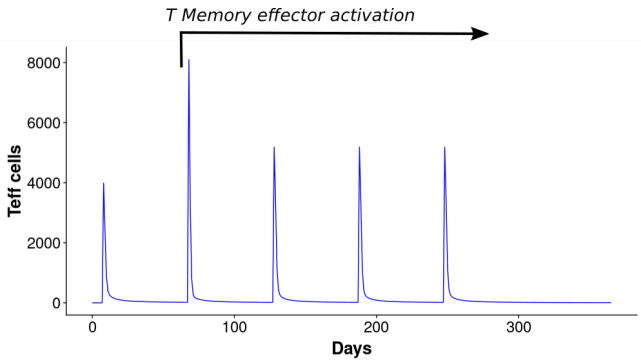


FIGURE 6.20: **Teffs considering an MS subject.** The Teff dynamics considering an MS subject.

TABLE 6.1: List of the model fixed and unknown parameters with their corresponding values or (in the latter case) ranges on whose the Uniform distribution is defined.

Transitions/events	Parameters	Range
Treg Death	r_{TregD}	$1/24\ h^{-1}$
Teff Death	r_{TeffD}	$1/24\ h^{-1}$
NK Death	r_{NKD}	$1/24\ h^{-1}$
NK Dup	r_{NKDup}	$1/24\ h^{-1}$
Teff Activation	r_{TeffA}	$[0.2, 0.6]\ h^{-1}$
Treg Activation	r_{TregA}	$[0.1, 0.3]\ h^{-1}$
Treg Dup	$r_{TregDup}$	$[0.045, 0.135]\ h^{-1}$
Teff Dup	$r_{TeffDup}$	$[0.25, 0.75]\ h^{-1}$
TeffKillODC	r_{TeKodc}	$[0.05, 0.15]\ h^{-1}$
TregKillTeff	r_{TrKTe}	$[1.5, 4.5]\ h^{-1}$
TeffKillEBV	r_{TeKebv}	$[0.075, 0.225]\ h^{-1}$
Recovery	r_{rec}	$[0.075, 0.225]\ h^{-1}$
NKKillTcell	r_{NKkTc}	$[0.05, 0.15]\ h^{-1}$
DACDeath	r_{DacD}	$[0.0004, 0.001]\ h^{-1}$
DACinjection	r_{DacInj}	$[5, 100]\ h^{-1}$

TABLE 6.2: List of the model constants.

Constant	Value
$q_{RestTreg}$	20
$q_{RestTeff}$	500
q_{NK}	100
C_{EBV}	1000
C_{Mem}	200
C_{DAC}	$(DACinjected)/log(.1)$ [1]
C_{IL2}	200
C_{Tcell}	200
C_{Teff}	200
p_{eff}^{dup}	2/3
p_{eff}^{mem}	1/3

TABLE 6.3: List of the cell numbers used in the model.

Cell	Value	Reference
$TLymphocytes$	$[3 * 10^3 cells/mm^3]$	Al-Mawali et al., 2018; Warny et al., 2018
$RestingTeff$	$[1687 cells/mm^3]$	Santagostino et al., 1999; Bisset et al., 2004; Choi et al., 2014
$RestingTreg$	$[63 cells/mm^3]$	Somerset et al., 2004
NK	$[375 cells/mm^3]$	Santagostino et al., 1999; Bisset et al., 2004; Choi et al., 2014
ODC	$[125 cells/mm^3]$	Segal, Schmitz, and Hof, 2009
$EBVinfection$	$[50 - 70 days]$	Balfour Jr et al., 2005

TABLE 6.4: Parameters used for simulating the Healthy version (first row) and Sick version (second row) of the disease.

Transition	Teff Activation	Treg Activation	Treg Dup	Teff Dup	TeffKillODC	TregKillTeff	TeffKillEBV	Recovery	NKKillTcell
Healthy	0.4	0.2	0.09	0.5	0.1	3	0.15	0.1	0.1
Sick	0.4	0.2	0.09	0.5	0.15	1	0.1	0.1	0.1

Bibliography

Adams, Jerry M and Andreas Strasser (2008). "Is tumor growth sustained by rare cancer stem cells or dominant clones?" In: *Cancer research* 68.11, pp. 4018–4021.

Al-Hajj, Muhammad, Max S Wicha, Adalberto Benito-Hernandez, Sean J Morrison, and Michael F Clarke (2003). "Prospective identification of tumorigenic breast cancer cells". In: *Proceedings of the National Academy of Sciences* 100.7, pp. 3983–3988.

Al-Mawali, Adhra, Avinash Daniel Pinto, Raiya Al-Busaidi, Rabab H Al-Lawati, and Magdi Morsi (2018). "Comprehensive haematological indices reference intervals for a healthy Omani population: First comprehensive study in Gulf Cooperation Council (GCC) and Middle Eastern countries based on age, gender and ABO blood group comparison". In: *PloS one* 13.4, e0194497.

Alessandri, Luca, Francesca Cordero, Marco Beccuti, Maddalena Arigoni, Martina Olivero, Greta Romano, Sergio Rabellino, Nicola Licheri, Gennaro De Libero, Luigia Pace, et al. (2019). "rCASC: reproducible classification analysis of single-cell sequencing data". In: *GigaScience* 8.9, giz105.

Alexandrov, Ludmil B, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. (2013). "Signatures of mutational processes in human cancer". In: *Nature* 500.7463, p. 415.

Altschuler, Steven J and Lani F Wu (2010). "Cellular heterogeneity: do differences make a difference?" In: *Cell* 141.4, pp. 559–563.

Alvero, Ayesha B, Rui Chen, Han-Hsuan Fu, Michele Montagna, Peter E Schwartz, Thomas Rutherford, Dan-Arin Silasi, Karina D Steffensen, Marianne Waldstrom, Irene Visintin, et al. (2009). "Molecular phenotyping of human ovarian cancer stem cells unravels the mechanisms for repair and chemoresistance". In: *Cell cycle* 8.1, pp. 158–166.

Amdahl, Gene M (1967). "Validity of the single processor approach to achieving large scale computing capabilities". In: *Proceedings of the April 18-20, 1967, spring joint computer conference*. ACM, pp. 483–485.

Andor, Noemi, Julie V Harness, Sabine Mueller, Hans W Mewes, and Claudia Petritsch (2013). "EXPANDS: expanding ploidy and allele frequency on nested subpopulations". In: *Bioinformatics* 30.1, pp. 50–60.

Andor, Noemi, Trevor A Graham, Marnix Jansen, Li C Xia, C Athena Aktipis, Claudia Petritsch, Hanlee P Ji, and Carlo C Maley (2016). "Pan-cancer analysis of the extent and consequences of intratumor heterogeneity". In: *Nature medicine* 22.1, p. 105.

Andrews, Tallulah S and Martin Hemberg (2018). "Identifying cell populations with scRNASeq". In: *Molecular aspects of medicine* 59, pp. 114–122.

Babar, Junaid, Marco Beccuti, Susanna Donatelli, and Andrew S. Miner (2010). "GreatSPN Enhanced with Decision Diagram Data Structures". In: *Application and Theory of Petri Nets.PETRI NETS 2010*. Vol. 6128. LNCS, pp. 308–317.

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

Balfour Jr, Henry H, Carol J Holman, Kristin M Hokanson, Meghan M Lelonek, Jill E Giesbrecht, Dana R White, David O Schmeling, Chiu-Ho Webb, Winston Cavert, David H

Wang, et al. (2005). “A prospective clinical study of Epstein-Barr virus and host interactions during acute infectious mononucleosis”. In: *Journal of Infectious Diseases* 192.9.

Barron, Martin and Jun Li (2016). “Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data”. In: *Scientific reports* 6, p. 33892.

Beccuti, M., C. Fornari, G. Franceschinis, S.M. Halawani, O. Ba-Rukab, A.R. Ahmad, and G. Balbo (2015). “From symmetric nets to differential equations exploiting model symmetries”. In: *Computer Journal* 58.1, pp. 23–39.

Beccuti, M., P. Cazzaniga, M. Pennisi, D. Besozzi, M. S. Nobile, S. Pernice, G. Russo, A. Tangherloni, and Pappalardo F. (2018). “GPU accelerated analysis of Treg-Teff cross regulation in relapsing-remitting multiple sclerosis.” In: *4th International European Conference on Parallel and Distributed Computing (Euro-Par 2018)*.

Beccuti, Marco, Elisa Genuardi, Greta Romano, Luigia Monitillo, Daniela Barbero, Mario Boccadoro, Marco Ladetto, Raffaele Adolfo Calogero, Simone Ferrero, and Francesca Cordero (2017a). “HashClone: a new tool to quantify the minimal residual disease in B-cell lymphoma from deep sequencing data”. In: *BMC bioinformatics* 18.1, p. 516.

Beccuti, Marco, Francesca Cordero, Maddalena Arigoni, Riccardo Panero, Elvio G Amparore, Susanna Donatelli, and Raffaele A Calogero (2017b). “SeqBox: RNAseq/ChIPseq reproducible analysis on a consumer game computer”. In: *Bioinformatics* 34.5, pp. 871–872.

Belbasis, Lazaros, Vanesa Bellou, Evangelos Evangelou, John PA Ioannidis, and Ioanna Tzoulaki (2015). “Environmental risk factors and multiple sclerosis: an umbrella review of systematic reviews and meta-analyses”. In: *The Lancet Neurology* 14.3.

Bisset, Leslie R, Thomas L Lung, Matthias Kaelin, Elisabeth Ludwig, and Rolf W Dubs (2004). “Reference values for peripheral blood lymphocyte phenotypes applicable to the healthy adult population in Switzerland”. In: *European journal of haematology* 72.3, pp. 203–212.

Bonnet, Dominique and John E Dick (1997). “Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell”. In: *Nature medicine* 3.7, p. 730.

Brady, Gerard, Mary Barbara, and Norman N Iscove (1990). “Representative in vitro cDNA amplification from individual hemopoietic cells and colonies”. In: *Methods Mol Cell Biol* 2.1, pp. 17–25.

Breiman, Leo (2001). “Random forests”. In: *Machine learning* 45.1, pp. 5–32.

Breiman, Leo, Jerome Friedman, Richard Olshen, and Charles Stone (1984). “Classification and regression trees. Wadsworth Int”. In: *Group* 37.15, pp. 237–251.

Brown, Terence A (2002). “Transcriptomes and proteomes”. In: *Genomes. 2nd edition*. Wiley-Liss.

Brüggemann, Monika, Thorsten Raff, and Michael Kneba (2012). “Has MRD monitoring superseded other prognostic factors in adult ALL?” In: *Blood* 120.23, pp. 4470–4481.

Brüggemann, Monika, Michaela Kotrová, Henrik Knecht, Jack Bartram, Myriam Boudjogrha, Vojtech Bystry, Grazia Fazio, Eva Froňková, Mathieu Giraud, Andrea Grioni, et al. (2019). “Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study”. In: *Leukemia*, p. 1.

Buettner, Florian, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle (2015). “Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells”. In: *Nature biotechnology* 33.2, p. 155.

Burrell, Rebecca A, Nicholas McGranahan, Jiri Bartek, and Charles Swanton (2013). “The causes and consequences of genetic heterogeneity in cancer evolution”. In: *Nature* 501.7467, p. 338.

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija (2018). "Integrating single-cell transcriptomic data across different conditions, technologies, and species". In: *Nature biotechnology* 36.5, p. 411.

Bystry, Vojtech, Tomas Reigl, Adam Krejci, Martin Demko, Barbora Hanakova, Andrea Grioni, Henrik Knecht, Max Schlitt, Peter Dreger, Leopold Sellner, et al. (2016). "ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data". In: *Bioinformatics* 33.3, pp. 435–437.

Campana, Dario (2010). "Minimal residual disease in acute lymphoblastic leukemia". In: *ASH Education Program Book* 2010.1, pp. 7–12.

Cantini, Laura, Laurence Calzone, Loredana Martignetti, Mattias Rydenfelt, Nils Blüthgen, Emmanuel Barillot, and Andrei Zinovyev (2017). "Classification of gene signatures for their information value and functional redundancy". In: *NPJ systems biology and applications* 4.1, p. 2.

Chen, Shu-Hwa, Wen-Yu Kuo, Sheng-Yao Su, Wei-Chun Chung, Jen-Ming Ho, Henry Horng-Shing Lu, and Chung-Yen Lin (2018). "A gene profiling deconvolution approach to estimating immune cell composition from complex tissues". In: *BMC bioinformatics* 19.4, p. 154.

Chiola, G., C. Dutheillet, G. Franceschinis, and S. Haddad (1993). "Stochastic well-formed coloured nets for symmetric modelling applications". In: *IEEE Transactions on Computers* 42.11, pp. 1343–1360.

Choi, Joungbum, Su Jin Lee, Yun A Lee, Hyung Gun Maeng, Jong Kyun Lee, and Yong Won Kang (2014). "Reference values for peripheral blood lymphocyte subsets in a healthy korean population". In: *Immune network* 14.6, pp. 289–295.

Clerico, Marinella, Carlo Alberto Artusi, Alessandra Di Liberto, Simona Rolla, Valentina Bardina, Pierangelo Barbero, Stefania Federica De Mercanti, and Luca Durelli (2017). "Natalizumab in Multiple Sclerosis: Long-Term Management". In: *Int. J. Mol. Sci.* 18.5. Ed. by Christoph Kleinschnitz and Sven Meuth, p. 940. ISSN: 1422-0067.

Dendrou, Calliope A, Lars Fugger, and Manuel A Friese (2015). "Immunopathology of multiple sclerosis". In: *Nature Reviews Immunology* 15.9, p. 545.

Di Tizio, Daniela, Alessandra Di Serafino, Prabin Upadhyaya, Luca Sorino, Liborio Stuppia, and Ivana Antonucci (2018). "The Impact of Epigenetic Signatures on Amniotic Fluid Stem Cell Fate". In: *Stem cells international* 2018.

Diaz, Aaron, Siyuan J Liu, Carmen Sandoval, Alex Pollen, Tom J Nowakowski, Daniel A Lim, and Arnold Kriegstein (2016). "SCell: integrated analysis of single-cell RNA-seq data". In: *Bioinformatics* 32.14, pp. 2219–2220.

Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1, pp. 15–21.

Dongen, Jacques JM van, Taku Seriu, E Renate Panzer-Grümayer, Andrea Biondi, Marja J Pongers-Willemse, Lilly Corral, Frank Stolz, Martin Schrappe, Giuseppe Masera, Willem A Kamps, et al. (1998). "Prognostic value of minimal residual disease in acute lymphoblastic leukaemia in childhood". In: *The Lancet* 352.9142, pp. 1731–1738.

Dongen, Jacques JM van, Vincent HJ van der Velden, Monika Brüggemann, and Alberto Orfao (2015). "Minimal residual disease diagnostics in acute lymphoblastic leukemia: need for sensitive, fast, and standardized technologies". In: *Blood* 125.26, pp. 3996–4009.

Dupont, Cathérine, D Randall Armant, and Carol A Brenner (2009). "Epigenetics: definition, mechanisms and clinical perspective". In: *Seminars in reproductive medicine*. Vol. 27. 05. © Thieme Medical Publishers, pp. 351–357.

Dutta, Ranjan and Bruce D Trapp (2011). "Mechanisms of Neuronal Dysfunction and Degeneration in Multiple Sclerosis". In: *Prog. Neurobiol.* 93.1, pp. 1–12. ISSN: 0301-0082.

Eberwine, James, Hermes Yeh, Kevin Miyashiro, Yanxiang Cao, Suresh Nair, Richard Finnell, Martha Zettel, and Paul Coleman (1992). "Analysis of gene expression in single live neurons". In: *Proceedings of the National Academy of Sciences* 89.7, pp. 3010–3014.

Elsasser, Walter M (1984). "Outline of a theory of cellular heterogeneity". In: *Proceedings of the National Academy of Sciences* 81.16, pp. 5126–5129.

Faham, Malek, Jianbiao Zheng, Martin Moorhead, Victoria EH Carlton, Patricia Stow, Elaine Coustan-Smith, Ching-Hon Pui, and Dario Campana (2012). "Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia". In: *Blood* 120.26, pp. 5173–5180.

Fawagreh, Khaled, Mohamed Medhat Gaber, and Eyad Elyan (2014). "Random forests: from early developments to recent advancements". In: *Systems Science & Control Engineering: An Open Access Journal* 2.1, pp. 602–609.

Felipe De Sousa, E Melo, Louis Vermeulen, Evelyn Fessler, and Jan Paul Medema (2013). "Cancer heterogeneity—a multifaceted view". In: *EMBO reports* 14.8, pp. 686–695.

Ferreira, Rita, Daniel Moreira-Gonçalves, Ana Lúcia Azevedo, José Alberto Duarte, Francisco Amado, and Rui Vitorino (2015). "Unraveling the exercise-related proteome signature in heart". In: *Basic research in cardiology* 110.1, p. 454.

Ferrero, Simone, Daniela Drandi, Barbara Mantoan, Paola Ghione, Paola Omedè, and Marco Ladetto (2011). "Minimal residual disease detection in lymphoma and multiple myeloma: impact on therapeutic paradigms". In: *Hematological oncology* 29.4, pp. 167–176.

Fu, Shuangshuang, Michael Wang, David R Lairson, Ruosha Li, Bo Zhao, and Xianglin L Du (2017). "Trends and variations in mantle cell lymphoma incidence from 1995 to 2013: A comparative study between Texas and National SEER areas". In: *Oncotarget* 8.68, p. 112516.

Gaëta, Bruno A, Harald R Malming, Katherine JL Jackson, Michael E Bain, Patrick Wilson, and Andrew M Collins (2007). "iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences". In: *Bioinformatics* 23.13, pp. 1580–1587.

Galimberti, Sara, Elisa Genuardi, Francesco Mazziotta, Lorenzo Iovino, Fortunato Morabito, Susanna Grassi, Elena Ciabatti, Francesca Guerrini, and Mario Petrini (2019). "The Minimal Residual Disease in Non-Hodgkin's Lymphomas: From the Laboratory to the Clinical Practice". In: *Frontiers in Oncology* 9.

Garrels, JI (2001). "Proteome". In: *Encyclopedia of Genetics*.

Gerlinger, Marco, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, et al. (2012). "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing". In: *New England journal of medicine* 366.10, pp. 883–892.

Giraud, Mathieu, Mikaël Salson, Marc Duez, Céline Villenet, Sabine Quief, Aurélie Caillault, Nathalie Grardel, Christophe Roumier, Claude Preudhomme, and Martin Figeac (2014). "Fast multiclonal clusterization of V (D) J recombinations from high-throughput sequencing". In: *BMC genomics* 15.1, p. 409.

Giudicelli, V, D Chaume, and MP Lefranc (2005). "IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes". In: *Nucleic Acids Research* 33, pp. 256–261.

Giudicelli, Veronique, Denys Chaume, and Marie-Paule Lefranc (2004). "IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V–J and V–D–J rearrangement analysis". In: *Nucleic acids research* 32.suppl_2, W435–W440.

Gold, R, G Giovannoni, K Selmaj, E Havrdova, X Montalban, EW Radue, D Stefoski, R Robinson, K Riester, J Rana, J Elkins, G O'Neill, and SELECT study investigators.

(2013). "Daclizumab high-yield process in relapsing-remitting multiple sclerosis (SELECT): a randomised, double-blind, placebo-controlled trial." In: *Lancet* 381.9884, pp. 2167–2175.

Goldman, Samantha L, Matthew JD MacKay, Ebrahim Afshinnekoo, Ari Melnick, Shixiu Wu, and Christopher E Mason (2019). "The impact of heterogeneity on single-cell sequencing". In: *Frontiers in Genetics* 10, p. 8.

Greaves, Mel and Carlo C Maley (2012). "Clonal evolution in cancer". In: *Nature* 481.7381, p. 306.

Grob, Jean-Jacques, Laurent Mortier, Lionel D'Hondt, Florent Grange, Jean Francois Baurain, Brigitte Dréno, Céleste Lebbe, Caroline Robert, Anne Dompmartin, Bart Neyns, et al. (2017). "Safety and immunogenicity of MAGE-A3 cancer immunotherapeutic with dacarbazine in patients with MAGE-A3-positive metastatic cutaneous melanoma: an open phase I/II study with a first assessment of a predictive gene signature". In: *ESMO open* 2.5, e000203.

Guarnera, Cristina, Placido Bramanti, and Emanuela Mazzon (2017). "Alemtuzumab: a review of efficacy and risks in the treatment of relapsing remitting multiple sclerosis". In: *Ther. Clin. Risk Manag.* 13, pp. 871–879. ISSN: 1176-6336.

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

Handel, Adam E, Sarosh R Irani, and Georg A Holländer (2018). "The role of thymic tolerance in CNS autoimmune disease". In: *Nature Reviews Neurology*, p. 1.

Hartman, John L, Barbara Garvik, and Lee Hartwell (2001). "Principles for the buffering of genetic variation". In: *Science* 291.5506, pp. 1001–1004.

Hashimshony, Tamar, Florian Wagner, Noa Sher, and Itai Yanai (2012). "CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification". In: *Cell reports* 2.3, pp. 666–673.

Hennig, Christian (2013). *fpc: Flexible procedures for clustering. R package version 2.1-5*.

Henssen, Anton, Richard Koche, Jiali Zhuang, Eileen Jiang, Casie Reed, Amy Eisenberg, Eric Still, Ian MacArthur, Elias Rodriguez-Fos, Santiago Gonzalez, et al. (2017). "Human PGBD5 DNA transposase promotes site-specific oncogenic mutations in rhabdoid tumors". In: *bioRxiv*, p. 111138.

Herrera, Alex F, Haesook T Kim, Katherine A Kong, Malek Faham, Heather Sun, Aliyah R Sohani, Edwin P Alyea, Victoria E Carlton, Yi-Bin Chen, Corey S Cutler, et al. (2016). "Next-generation sequencing-based detection of circulating tumour DNA After allogeneic stem cell transplantation for lymphoma". In: *British journal of haematology* 175.5, pp. 841–850.

Hoshida, Yujin, Augusto Villanueva, Angelo Sangiovanni, Manel Sole, Chin Hur, Karin L Andersson, Raymond T Chung, Joshua Gould, Kensuke Kojima, Supriya Gupta, et al. (2013). "Prognostic gene expression signature for patients with hepatitis C–related early-stage cirrhosis". In: *Gastroenterology* 144.5, pp. 1024–1030.

Huang, Sui (2009). "Non-genetic heterogeneity of cells in development: more than just noise". In: *Development* 136.23, pp. 3853–3862.

Huber, Wolfgang, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, et al. (2015). "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nature methods* 12.2, p. 115.

Hunt, Earl B, Janet Marin, and Philip J Stone (1966). "Experiments in induction." In:

Hwang, Byungjin, Ji Hyun Lee, and Duhee Bang (2018). "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental & molecular medicine* 50.8, pp. 1–14.

Itadani, Hiraku, Shinji Mizuarai, and Hidehito Kotani (2008). "Can systems biology understand pathway activation? Gene expression signatures as surrogate markers for understanding the complexity of pathway activation". In: *Current genomics* 9.5, pp. 349–360.

Jackson, Katherine JL, Scott Boyd, Bruno A Gaëta, and Andrew M Collins (2010). "Benchmarking the performance of human antibody gene alignment utilities using a 454 sequence dataset". In: *Bioinformatics* 26.24, pp. 3129–3130.

Jan, Max, Thomas M Snyder, M Ryan Corces-Zimmerman, Paresh Vyas, Irving L Weissman, Stephen R Quake, and Ravindra Majeti (2012). "Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia". In: *Science translational medicine* 4.149, 149ra118–149ra118.

Karp, G (2009). *Biologia cellulare e molecolare-3 edizione*.

Kim, Anne P and Danial E Baker (2016). "Daclizumab". In: *Hospital pharmacy* 51.11, pp. 928–939.

Klco, Jeffery M, David H Spencer, Christopher A Miller, Malachi Griffith, Tamara L Lamprecht, Michelle O'Laughlin, Catrina Fronick, Vincent Magrini, Ryan T Demeter, Robert S Fulton, et al. (2014). "Functional heterogeneity of genetically defined subclones in acute myeloid leukemia". In: *Cancer cell* 25.3, pp. 379–392.

Kock Metz, Agrawal and Yong (2013). "Environmental factors and their regulation of immunity in multiple sclerosis". In: *J Neurol Sci* 324.1-2, pp. 10–6.

Kotrova, Michaela, Katerina Muzikova, Ester Mejstrikova, Michaela Novakova, Violeta Bakardjieva-Mihaylova, Karel Fiser, Jan Stuchly, Mathieu Giraud, Mikaël Salson, Christiane Pott, et al. (2015). "The predictive strength of next-generation sequencing MRD detection for relapse compared with current methods in childhood ALL". In: *Blood* 126.8, p. 1045.

Kotrova, Michaela, Jan Trka, Michael Kneba, and Monika Brüggemann (2017). "Is next-generation sequencing the way to go for residual disease monitoring in acute lymphoblastic leukemia?" In: *Molecular diagnosis & therapy* 21.5, pp. 481–492.

Kreso, Antonija and John E Dick (2014). "Evolution of the cancer stem cell model". In: *Cell stem cell* 14.3, pp. 275–291.

Kukurba, Kimberly R and Stephen B Montgomery (2015). "RNA sequencing and analysis". In: *Cold Spring Harbor Protocols* 2015.11, pdb–top084970.

Kuleshov, Maxim V, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. (2016). "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update". In: *Nucleic acids research* 44.W1, W90–W97.

Kurtz, Thomas G (1978). "Strong approximation theorems for density dependent Markov chains". In: *Stoc. Proc. Appl.* 6.3, pp. 223–240.

Kwan, Tanya T, Aditya Bardia, Laura M Spring, Anita Giobbie-Hurder, Mark Kalinich, Taronish Dubash, Tilak Sundaresan, Xin Hong, Joseph A LiCausi, Uyen Ho, et al. (2018). "A digital RNA signature of circulating tumor cells predicting early therapeutic response in localized and metastatic breast cancer". In: *Cancer discovery* 8.10, pp. 1286–1299.

Ladetto, Marco, M Brüggemann, Luigia Monitillo, Simone Ferrero, F Pepin, Daniela Drandi, Daniela Barbero, Antonio Palumbo, R Passera, Mario Boccadoro, et al. (2014). "Next-generation sequencing and real-time quantitative PCR for minimal residual disease detection in B-cell disorders". In: *Leukemia* 28.6, p. 1299.

Lawrence, Michael S, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. (2013). "Mutational heterogeneity in cancer and the search for new cancer-associated genes". In: *Nature* 499.7457, p. 214.

Li, Bryan T, Jin X Lim, and Maurice HT Ling (2019). "Analyzing Transcriptome-Phenotype Correlations". In: *Encyclopedia of Bioinformatics and Computational Biology* 3, pp. 819–824.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *bioinformatics* 25.14, pp. 1754–1760.

Liu, Jiangang, Andrew Campen, Shuguang Huang, Sheng-Bin Peng, Xiang Ye, Mathew Palakal, A Keith Dunker, Yuni Xia, and Shuyu Li (2008). "Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data". In: *BMC medical genomics* 1.1, p. 39.

Liu, Rui, Xinhao Wang, Grace Y Chen, Piero Dalerba, Austin Gurney, Timothy Hoey, Gavin Sherlock, John Lewicki, Kerby Shedden, and Michael F Clarke (2007). "The prognostic role of a gene signature from tumorigenic breast-cancer cells". In: *New England Journal of Medicine* 356.3, pp. 217–226.

Liu, Zehua, Huazhe Lou, Kaikun Xie, Hao Wang, Ning Chen, Oscar M Aparicio, Michael Q Zhang, Rui Jiang, and Ting Chen (2017). "Reconstructing cell cycle pseudo time-series via single-cell transcriptome data". In: *Nature communications* 8.1, p. 22.

Lodish, Harvey, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell (2000). "Molecular cell biology 4th edition". In: *National Center for Biotechnology Information, Bookshelf*.

Lun, Aaron TL, Davis J McCarthy, and John C Marioni (2016). "A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor". In: *F1000Research* 5.

Ma, Jiao, David Redmond, Ayako Miyaguchi, Anna S Nam, Kui Nie, Susan Mathew, Olivier Elemento, and Wayne Tam (2019). "Exploring tumor clonal evolution in bone marrow of patients with diffuse large B-cell lymphoma by deep IGH sequencing and its potential relevance in relapse". In: *Blood cancer journal* 9.9, pp. 1–6.

Macosko, Evan Z, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. (2015). "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets". In: *Cell* 161.5, pp. 1202–1214.

Marcus, Robert, John W Sweetenham, and Michael E Williams (2013). *Lymphoma: pathology, diagnosis, and treatment*. Cambridge University Press.

Mardis, Elaine R (2013). "Next-generation sequencing platforms". In: *Annual review of analytical chemistry* 6, pp. 287–303.

Marino, Simeone, Ian B. Hogue, Christian J. Ray, and Denise E. Kirschner (2008). "A methodology for performing global uncertainty and sensitivity analysis in systems biology". In: *Journal of Theoretical Biology* 254.1, pp. 178 –196. ISSN: 0022-5193. URL: http://www.sciencedirect.com/science/article/pii/S0022519308001896.

Marsan, M. Ajmone, G. Balbo, G. Conte, S. Donatelli, and G. Franceschinis (1995). *Modelling with Generalized Stochastic Petri Nets*. New York, NY, USA: J. Wiley.

Marusyk, Andriy and Kornelia Polyak (2010). "Tumor heterogeneity: causes and consequences". In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1805.1, pp. 105–117.

McCombe, Pamela (2018). "The short and long-term effects of pregnancy on multiple sclerosis and experimental autoimmune encephalomyelitis". In: *Journal of clinical medicine* 7.12, p. 494.

McKay, M. D., R. J. Beckman, and W. J. Conover (1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". In: *Technometrics* 21.2, pp. 239–245. ISSN: 00401706. URL: http://www.jstor.org/stable/1268522.

Merienne, Nicolas, Cécile Meunier, Anne Schneider, Jonathan Seguin, Satish S Nair, Anne B Rocher, Stephanie Le Gras, Céline Keime, Richard Faull, Luc Pellerin, et al. (2019). "Cell-Type-Specific Gene Expression Profiling in Adult Mouse Brain Reveals Normal and Disease-State Signatures". In: *Cell reports* 26.9, pp. 2477–2493.

Moreno-Romero, Jordi, Gerardo Del Toro-De León, Vikash Kumar Yadav, Juan Santos-González, and Claudia Köhler (2019). "Epigenetic signatures associated with imprinted paternally expressed genes in the Arabidopsis endosperm". In: *Genome biology* 20.1, p. 41.

Munshaw, Supriya and Thomas B Kepler (2010). "SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements". In: *Bioinformatics* 26.7, pp. 867–872.

Naik, Shalin H, Leïla Perié, Erwin Swart, Carmen Gerlach, Nienke van Rooij, Rob J de Boer, and Ton N Schumacher (2013). "Diverse and heritable lineage imprinting of early haematopoietic progenitors". In: *Nature* 496.7444, p. 229.

Ng, Charlotte KY, Britta Weigelt, Roger A'Hern, Francois-Clement Bidard, Christophe Lemetre, Charles Swanton, Ronglai Shen, and Jorge S Reis-Filho (2014). "Predictive performance of microarray gene signatures: impact of tumor heterogeneity and multiple mechanisms of drug resistance". In: *Cancer research* 74.11, pp. 2946–2961.

Nguyen, Hao G, Christopher J Welty, and Matthew R Cooperberg (2015). "Diagnostic associations of gene expression signatures in prostate cancer tissue". In: *Current opinion in urology* 25.1, pp. 65–70.

Nguyen, Long V, Robert Vanner, Peter Dirks, and Connie J Eaves (2012). "Cancer stem cells: an evolving concept". In: *Nature Reviews Cancer* 12.2, p. 133.

Nowell, Peter C (1976). "The clonal evolution of tumor cell populations". In: *Science* 194.4260, pp. 23–28.

Oldenhuis, CNAM, SF Oosting, JA Gietema, and EGE De Vries (2008). "Prognostic versus predictive value of biomarkers in oncology". In: *European journal of cancer* 44.7, pp. 946–953.

Ozsolak, Fatih and Patrice M Milos (2011). "RNA sequencing: advances, challenges and opportunities". In: *Nature reviews genetics* 12.2, p. 87.

Pace, Luigia, Christel Goudot, Elina Zueva, Paul Gueguen, Nina Burgdorf, Joshua J Waterfall, Jean-Pierre Quivy, Geneviève Almouzni, and Sebastian Amigorena (2018). "The epigenetic control of stemness in CD8+ T cell fate commitment". In: *Science* 359.6372, pp. 177–186.

Paciello, Giulia, Andrea Acquaviva, Chiara Pighi, Alberto Ferrarini, Enrico Macii, Alberto Zamo', and Elisa Ficarra (2015). "VDJSeq-Solver: in silico V (D) J recombination detection tool". In: *PLoS One* 10.3, e0118192.

Patel, Anoop P, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. (2014). "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma". In: *Science* 344.6190, pp. 1396–1401.

Patel, Ravi K and Mukesh Jain (2012). "NGS QC Toolkit: a toolkit for quality control of next generation sequencing data". In: *PloS one* 7.2, e30619.

Pennisi, Marzio, Abdul Mateen Rajput, Luca Toldo, and Francesco Pappalardo (2013). "Agent based modeling of Treg-Teff cross regulation in relapsing-remitting multiple sclerosis". In: *BMC Bioinformatics*.

Pernice, S., M. Beccuti, P. Do', M. Pennisi, and F. Pappalardo (2018). "Estimating Daclizumab effects in Multiple Sclerosis using Stochastic Symmetric Nets". In: *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018, Madrid, Spain, December 3-6, 2018*, pp. 1393–1400.

Pernice, S., L. Follia, G. Balbo, G. Sartini, N. Totis, P. Lió, I. Merelli, F. Cordero, and Beccuti M. (2019a). "Integrating Petri nets and Flux Balance methods in computational biology models: a methodological and computational practice." In: *Fundamenta Informaticae*.

Pernice, Simone, Marzio Pennisi, Greta Romano, Alessandro Maglione, Santina Cutrupi, Francesco Pappalardo, Gianfranco Balbo, Marco Beccuti, Francesca Cordero, and Raffaele Adolfo Calogero (2019b). "A computational approach based on the Colored Petri Net formalism for studying Multiple Sclerosis". In: *BMC bioinformatics*.

Pimienta, Genaro, Douglas M Heithoff, Alexandre Rosa-Campos, Minerva Tran, Jeffrey D Esko, Michael J Mahan, Jamey D Marth, and Jeffrey W Smith (2019). "Plasma Proteome Signature of Sepsis: a Functionally Connected Protein Network". In: *Proteomics* 19.5, p. 1800389.

Pott, Christiane, Eva Hoster, Marie-Hélène Delfau-Larue, Kheira Beldjord, Sebastian Böttcher, Vahid Asnafi, Anne Plonquet, Reiner Siebert, Evelyne Callet-Bauchu, Niels Andersen, et al. (2010). "Molecular remission is an independent predictor of clinical outcome in patients with mantle cell lymphoma after combined immunochemotherapy: a European MCL intergroup study". In: *Blood* 115.16, pp. 3215–3223.

Pujol, Jean-Louis, Johan F Vansteenkiste, Tommaso Martino De Pas, Djordje Atanackovic, Martin Reck, Michiel Thomeer, Jean-Yves Douillard, Gianpiero Fasola, Vanessa Potter, Paul Taylor, et al. (2015). "Safety and immunogenicity of MAGE-A3 cancer immunotherapeutic with or without adjuvant chemotherapy in patients with resected stage IB to III MAGE-A3-positive non-small-cell lung cancer". In: *Journal of Thoracic Oncology* 10.10, pp. 1458–1467.

Ralph, Duncan K and Frederick A Matsen IV (2016). "Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation". In: *PLoS computational biology* 12.1, e1004409.

Ramsköld, Daniel, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, et al. (2012). "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells". In: *Nature biotechnology* 30.8, p. 777.

Ranzoni, Anna M, Paulina M Strzelecka, and Ana Cvejic (2019). "Application of single-cell RNA sequencing methodologies in understanding haematopoiesis and immunology". In: *Essays in biochemistry* 63.2, pp. 217–225.

Ricci, Pierbruno, Pedro Magalhães, Magdalena Krochmal, Martin Pejchinovski, Erica Daina, Maria Rosa Caruso, Laura Goea, Iwona Belczacka, Giuseppe Remuzzi, Muriel Umbhauer, et al. (2019). "Urinary proteome signature of Renal Cysts and Diabetes syndrome in children". In: *Scientific reports* 9.1, p. 2225.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1, pp. 139–140.

Romano, Greta, Elisa Genuardi, Raffaele Calogero, and Simone Ferrero (2018). "ParallelHashClone: a parallel implementation of HashClone suite for clonality assessment from NGS data". In: *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*. IEEE, pp. 415–422.

Roth, Andrew, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah (2014). "PyClone: statistical inference of clonal population structure in cancer". In: *Nature methods* 11.4, p. 396.

Russell, Stuart J and Peter Norvig (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,

Sabaawy, Hatem E (2013). "Genetic heterogeneity and clonal evolution of tumor cells and their impact on precision cancer medicine". In: *Journal of leukemia (Los Angeles, Calif.)* 1.4, p. 1000124.

Saltelli, Andrea, Marco Ratto, Stefano Tarantola, and Francesca Campolongo (2005). "Sensitivity Analysis for Chemical Models". In: *Chemical Reviews* 105.7, pp. 2811–2828.

Salzer, Jonatan, Rasmus Svenningsson, Peter Alping, Lenka Novakova, Anna Björck, Katharina Fink, Protik Islam-Jakobsson, Clas Malmeström, Markus Axelsson, Mattias Vågberg, Peter Sundström, Jan Lycke, Fredrik Piehl, and Anders Svenningsson (2016). "Rituximab in multiple sclerosis: A retrospective observational study on safety and efficacy". In: *Neurology* 87.20, pp. 2074–2081. ISSN: 0028-3878.

Sánchez-Ramón, Silvia, Joaquín Navarroa, Carol Aristimuño, Margarita Rodríguez-Mahou, José Ma Bellón, Eduardo Fernández-Cruz, and Clara de Andrés (2005). "Pregnancy-induced expansion of regulatory T-lymphocytes may mediate protection to multiple sclerosis activity". In: *Immunology letters* 96.2.

Sanger, Frederick, Steven Nicklen, and Alan R Coulson (1977). "DNA sequencing with chain-terminating inhibitors". In: *Proceedings of the national academy of sciences* 74.12, pp. 5463–5467.

Sanjuan-Pla, Alejandra, Iain C Macaulay, Christina T Jensen, Petter S Woll, Tiago C Luis, Adam Mead, Susan Moore, Cintia Carella, Sahoko Matsuoka, Tiphaine Bouriez Jones, et al. (2013). "Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy". In: *Nature* 502.7470, p. 232.

Santagostino, Alberto, Germano Garbaccio, Angela Pistorio, Vittorio Bolis, Giovanni Camisasca, Pasqualepaolo Pagliaro, and Mauro Girotto (1999). "An Italian national multicenter study for the definition of reference ranges for normal values of peripheral blood lymphocyte subsets in healthy adults". In: *Haematologica* 84.6.

Schultze, JL and JG Gribben (1996). "Minimal residual disease in non-Hodgkin's lymphoma". In: *Biomedicine & pharmacotherapy* 50.9, pp. 451–458.

Schürch, Christian M, Birgit Federmann, Leticia Quintanilla-Martinez, and Falko Fend (2018). "Tumor heterogeneity in lymphomas: a different breed". In: *Pathobiology* 85.1-2, pp. 130–145.

Segal, Devorah, Christoph Schmitz, and Patrick R Hof (2009). "Spatial distribution and density of oligodendrocytes in the cingulum bundle are unaltered in schizophrenia". In: *Acta neuropathologica* 117.4, p. 385.

Serra, Denise, Urs Mayr, Andrea Boni, Ilya Lukonin, Markus Rempfler, Ludivine Challet Meylan, Michael B Stadler, Petr Strnad, Panagiotis Papasaikas, Dario Vischi, et al. (2019). "Self-organization and symmetry breaking in intestinal organoid development". In: *Nature* 569.7754, p. 66.

Shawky, Rabah M (2014). "Reduced penetrance in human inherited disease". In: *Egyptian Journal of Medical Human Genetics* 15.2, pp. 103–111.

Shin, Jaehoon, Daniel A Berg, Yunhua Zhu, Joseph Y Shin, Juan Song, Michael A Bonaguidi, Grigori Enikolopov, David W Nauen, Kimberly M Christian, Guo-li Ming, et al. (2015). "Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis". In: *Cell stem cell* 17.3, pp. 360–372.

Singh, Sheila K, Ian D Clarke, Mizuhiko Terasaki, Victoria E Bonn, Cynthia Hawkins, Jeremy Squire, and Peter B Dirks (2003). "Identification of a cancer stem cell in human brain tumors". In: *Cancer research* 63.18, pp. 5821–5828.

Somerset, David A, Yong Zheng, Mark D Kilby, David M Sansom, and Mark T Drayson (2004). "Normal human pregnancy is associated with an elevation in the immune suppressive CD25+ CD4+ regulatory T-cell subset". In: *Immunology* 112.1.

Souto-Carneiro, M Margarida, Nancy S Longo, Daniel E Russ, Hong-wei Sun, and Peter E Lipsky (2004). "Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER". In: *The Journal of Immunology* 172.11, pp. 6790–6802.

Szczepariski, Tamasz, Alberto Orfão, Vincent HJ van der Valden, Jésus F San Miguel, and Jacques JM van Dongen (2001). "Minimal residual disease in leukaemia patients". In: *The lancet oncology* 2.7, pp. 409–417.

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

Tirosh, Itay, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. (2016). "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq". In: *Science* 352.6282, pp. 189–196.

Trapp, Bruce D. and Klaus-Armin Nave (2008). "Multiple Sclerosis: An Immune or Neurodegenerative Disorder?" In: *Annu. Rev. Neurosci.* ISSN: 0147-006X. arXiv: NIHMS150003.

Urdinguio, Rocio G, Jose V Sanchez-Mut, and Manel Esteller (2009). "Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies". In: *The Lancet Neurology* 8.11, pp. 1056–1072.

Van Dongen, JJM, AW Langerak, Monika Brüggemann, PAS Evans, Michael Hummel, FL Lavender, E Delabesse, Frédéric Davi, Ed Schuuring, Ramon García-Sanz, et al. (2003). "Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936". In: *Leukemia* 17.12, p. 2257.

Velden, VHJ Van der, Andreas Hochhaus, Gianni Cazzaniga, Tomasz Szczepanski, J Gabert, and JJM Van Dongen (2003). "Detection of minimal residual disease in hematologic malignancies by real-time quantitative PCR: principles, approaches, and laboratory aspects". In: *Leukemia* 17.6, p. 1013.

Velden, VHJ Van der, G Cazzaniga, A Schrauder, J Hancock, P Bader, ER Panzer-Grumayer, T Flohr, R Sutton, H Cave, HO Madsen, et al. (2007). "Analysis of minimal residual disease by Ig/TCR gene rearrangements: guidelines for interpretation of real-time quantitative PCR data". In: *Leukemia* 21.4, p. 604.

Verhaak, Roel GW, Chantal S Goudswaard, Wim van Putten, Maarten A Bijl, Mathijs A Sanders, Wendy Hugens, André G Uitterlinden, Claudia AJ Erpelinck, Ruud Delwel, Bob Löwenberg, et al. (2005). "Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance". In: *Blood* 106.12, pp. 3747–3754.

Virtanen, Jacobson (2012). "Viruses and Multiple Sclerosis". In: *CNS Neurol Disord Drug Targets* 5, pp. 528–544. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4758194/.

Vukusic, Sandra, Michael Hutchinson, Martine Hours, Thibault Moreau, Patricia Cortinovis-Tourniaire, Patrice Adeleine, and Christian Confavreux (2004). "Pregnancy and multiple sclerosis (the PRIMS study): clinical predictors of post-partum relapse". In: *Brain* 127.6.

Wang, Bo, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou (2017). "Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning". In: *Nature methods* 14.4, p. 414.

Warny, Marie, Jens Helby, Børge Grønne Nordestgaard, Henrik Birgens, and Stig Egil Bojesen (2018). "Lymphopenia and risk of infection and infection-related death in 98,344 individuals from a prospective Danish population-based study". In: *PLoS medicine* 15.11, e1002685.

Witten, Ian H, Eibe Frank, Mark A Hall, and Christopher J Pal (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wolf, F Alexander, Philipp Angerer, and Fabian J Theis (2018). "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome biology* 19.1, p. 15.

Wouters, Bas J, Bob Löwenberg, Claudia AJ Erpelinck-Verschueren, Wim LJ van Putten, Peter JM Valk, and Ruud Delwel (2009). "Double CEBPA mutations, but not single CEBPA

mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome". In: *Blood* 113.13, pp. 3088–3091.

Wynn, Daniel, Michael Kaufman, Xavier Montalban, Timothy Vollmer, Jack Simon, Jacob Elkins, Gilmore O'Neill, Lauri Neyer, James Sheridan, Chungchi Wang, Alice Fong, and John W. Rose (2010). "Daclizumab in active relapsing multiple sclerosis (CHOICE study): a phase 2, randomised, double-blind, placebo-controlled, add-on trial with interferon beta". In: *Lancet Neurol.* 9.4, pp. 381–390. ISSN: 14744422.

Xuan, Jiekun, Ying Yu, Tao Qing, Lei Guo, and Leming Shi (2013). "Next-generation sequencing in the clinic: promises and challenges". In: *Cancer letters* 340.2, pp. 284–295.

Yamamoto, Ryo, Yohei Morita, Jun Ooehara, Sanae Hamanaka, Masafumi Onodera, Karl Lenhard Rudolph, Hideo Ema, and Hiromitsu Nakauchi (2013). "Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells". In: *Cell* 154.5, pp. 1112–1126.

Yamout, B. I. and R. Alroughani (2018). "Multiple Sclerosis". In: *Semin Neurol* 38.2, pp. 212–225.

Zhang, Hui-lai, Hua-qing Wang, Xi-shan Hao, Daniela Capello, Sergio B Cogliatti, Francesco Bertoni, and Franco Cavalli (2011). "Biased immunoglobulin genes rearrangement in mantle cell lymphoma: Hints to identify the normal B-cell counterpart". In: *Clinical Oncology and Cancer Research* 8.2, p. 65.

Zhao, Mengyao, Wan-Ping Lee, Erik P Garrison, and Gabor T Marth (2013). "SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications". In: *PloS one* 8.12, e82138.

Zhu, Kaiyi, Tai-Hsien Ou Yang, Vincent Dorie, Tian Zheng, and Dimitris Anastassiou (2019). "Meta-analysis of expression and methylation signatures indicates a stress-related epigenetic mechanism in multiple neuropsychiatric disorders". In: *Translational psychiatry* 9.1, p. 32.

Zhu, Xun, Thomas K Wolfgruber, Austin Tasato, Cédric Arisdakessian, David G Garmire, and Lana X Garmire (2017). "Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists". In: *Genome medicine* 9.1, p. 108.