Modelling timing in blood cancers

Laure Talarmain

MRC Cancer Unit, Hutchison/MRC Research Centre
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Corpus Christi College

March 2021

I would like to dedicate this thesis to my mum and my brother.

Declaration

This dissertation describes work done between October 2016 and December 2020 at the Hutchison/MRC Research Centre, Cambridge, UK under the supervision of Dr. Benjamin A. Hall.

I hereby declare that this thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.

This thesis does not exceed the prescribed word limit (60,000) for the Clinical Medicine and Clinical Veterinary Medicine Degree Committee. These limits exclude figures, photographs, tables, appendices and bibliography.

Laure Talarmain
March 2021

Acknowledgements

Foremost, I would like to express my deepest gratitude to my supervisor Dr. Benjamin Hall. His support and scientific enthusiasm have made these four years as a PhD student a great and unforgettable adventure despite the numerous challenges. I wish every student to be as lucky as I have been to find a perfect mentor like Ben.

I am deeply grateful for my colleagues and our great lab environment without which these four years would not have been as enjoyable. In particular, I would like to thank Dr. David Shorthouse for his prompt scientific advice and general help whenever I needed it. You have been a scientific model for me and I know you will be an excellent PI. Special thank you also goes to Vicky Kostiou and Cassandra Kosmidou for their friendly support and our fun and very appreciated girl talks. Finally, I would like to thank Michael Hall for his help with all my mathematical questions.

I would like to extend my appreciation to Jasmine Fisher and Matthew Clarke for their collaboration, but also for their thesis and career advice.

Last but not least, I am forever grateful for my amazing family. You raised me to be an open-minded adult and taught me important human values. I owe you so much, and would not have been to complete this thesis without your endless support.

Abstract

Dysregulation of biological processes in normal cells can lead to the abnormal growth of tumours. Oncogenesis requires the acquisition of advantageous mutations to expand in a fluctuating environment. Cancer cells gain these genetic and epigenetic alterations at different timing in their development, resulting in the formation of heterogeneous cell populations which interact and compete with each others inside tumours. At later stages, by escaping the immune system and acquiring malignant properties, some cancer cells manage to evade the primary tumour and spread in different organs to form metastases. Hence, tumour development in healthy tissues endure several biological changes whilst progressing and the order between these molecular and cellular events may modify prognosis.

This thesis addresses the influence of biological event timing on blood cancer progression and clinical outcomes. It first investigates the therapeutic efficacy of p53 restoration in a lymphoma mouse model. While several therapy schedules are tested, all fail due to resistance emergence. Computational modelling establishes the cell dynamics in these tumours and how to use it to propose alternative treatment strategies. Data availability leads this work to explore the impact of molecular evolution in myeloid malignancies. Notably, one study has found that Myeloproliferative Neoplasms patients with both *JAK2* and *TET2* mutations have different disease characteristics with distinct mutation order. My analyses identify *HOXA9* as a potential prognosis marker and biological switch responsible for patient stratification in these patients and in Acute Myeloid Leukemia. Additionally, a molecular network identifies the hematopoietic regulators involved in the branching evolution of Myeloproliferative Neoplasms. Further investigations of the Acute Myeloid Leukemia data show the possible involvement of *APP*, a gene associated to Alzheimer disease, in early cell fate commitment in hematopoiesis and in poor survival prognosis in undifferentiated leukemia when lowly expressed. Finally, this thesis examines the regulatory dynamics behind three clusters of Acute Myeloid Leukemia patients with distinct levels of *HOXA9* and *APP* expression. By building a program inferring molecular motifs from biological observations, genes which may interact with *HOXA9* and *APP* are identified.

Table of contents

List of abbreviations

ABC	Approximate Bayesian Computation
AD	Alzheimer Disease
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
APL	Acute Promyelocytic Leukemia
APP	Amyloid Precursor Protein
BMA	BioModelAnalyzer
BP-MPN	Blast-Phase Myeloproliferative Neoplasm
CD	Cluster Differentiation
CLP	Common Lymphoid Progenitor
CMP	Common Myeloid Progenitor
DC	Dendritic Cells
DEGs	Differentially Expressed Genes
DNA	Deoxyribonucleic Acid
ECM	Extracellular Matrix
ET	Essential Thrombocythemia
FAB	French-American-British
GMP	Granulocyte Monocyte Progenitor

GSEA	Gene Set Enrichment Analysis
HOXA9	Homeobox 9
HSC	Hematopoietic Stem Cell
JAK2	Janus Kinase 2
LD	Luria-Delbrück
LFC	Log2 Fold Change
LSK	Lineage$^-$, Sca-1$^+$, cKit$^+$
LT-HSC	Long Term Hematopoietic Stem Cell
LTL	LinearTemporal Logic
MCC	Matthew Correlation Coefficient
MDS	Myelodyplastic Syndrome
MEP	Megakayocyte Erythrocyte Progenitor
MHC	Major Histocompatibility Complex
MPAL	Mixed Phenotypic Acute Leukemia
MPN	Myeloproliferative Neoplasm
MPO	Myeloperoxidase
MPP	Multipotent Progenitor
NK	Natural Killer
ODE	Ordinary Differential Equation
PCA	Principal Component Analysis
PDE	Partial Differential Equation
PMF	Primary Myelofibrosis
PV	Polycythemia Vera
ProPPA	Probabilistic Programming Process Algebra)

RMSE	Root Mean Squared Error
RNA	Ribonucleic acid
SAT	Boolean satisfiability problem
SHAP	SHapley Additive exPlanations
SMT	Satisfiability Modulo Theories
ST-HSC	Short Term Hematopoietic Stem Cell
TCGA	The Cancer Genome Atlas
TCR	T Cell Receptor
TET2	Tet methylcytosine dioxygenase 2
VAF	Variant Allele Frequency
WBC	White Blood cell Count
XGBoost	eXtreme Gradient Boosting
t-SNE	t-distributed stochastic neighbor embedding

Chapter 1

Introduction

Summary

Dysregulation of biological processes in normal cells can lead to the abnormal growth of cancerous tumours. Such aberrations alter the basic functions of cells such as apoptosis, DNA repair or mobility and include a wide range of genetic and epigenetic transformations. When the first mutant cells appear, they start to expand, can form a lesion and eventually a tumour. This mass of cancerous cells grows over time by acquiring new abilities enabling progress and survival in the developing environment. Amongst the challenges faced by cancer cells, they must find resources to keep proliferating, but also avoid attacks from the immune system.

Inside the tumour, DNA alterations emerge within individual cells at distinct stages of disease evolution. These mutational processes result in the formation of various cell populations called clones which possess their own characteristics, interact and compete with each other to survive. These genetic and cellular dynamics contribute to intratumour heterogeneity which is a major barrier for cancer therapies. Simultaneous existence of heterogeneous clones in tumours often prevents clinicians targeting a mutation present in all tumour cells and eases the emergence of therapeutic resistant phenotype. It follows that treating tumours early enables better prognosis by reducing the risk of resistance and intratumour clinical diversity. Similarly, the sequence of aberrations in cells determine the morphology, structure and biological properties of tumours. It therefore further influences the diagnosis of the disease and the optimum treatment protocol. Moreover, different orderings between alterations require personalised care to temper distinct tumour aggressiveness. Hence, better understanding of timing in cancer development offers new opportunities to improve the prognosis of this complex disease. This thesis addresses the influence of biological event timing on blood cancer progression and clinical outcomes. I approach this by applying methods that address different aspects of this broad problem. I first explore cell dynamics and resistance emergence in a lymphoma mouse model. Then as a result of data availability, following studies explore the impact of molecular evolution in myeloid malignancies: notably, mutation order in Myeloproliferative Neoplasms and the effect of gene interactions in Acute Myeloid Leukemia patient stratification.

In this chapter, I first introduce cancer evolution by explaining what are the main biological events in cancer and what are their clinical impact and type of interactions inside tumours. I finish the section by characterising two contradictory models which illustrate therapeutic emergence in cancer. In the next section, I give a biological overview of hematopoiesis and blood cancers. I also provide a literature review of the effect of *JAK2* and *TET2* gene mutations in a myeloid malignancy called Myeloproliferative Neoplasm. Finally, I illustrate how cancer modelling is currently achieved in the mathematical and computer science community, which techniques are available and how I use them in my thesis to answer important biological questions.

1.1 Cancer Evolution

Tumour cells face severe environmental changes during their development (Fig 1.1). Studying how cancer progresses and evolves to adapt to these events can be seen as an ecological problem. Cancer cells can be represented as evolving individuals interacting and competing with each other for resources and survival while adjusting to their environment. One of the dominant models currently used for tumour evolution was introduced by Nowell in 1976 [2]. In this model, tumour evolution is illustrated as a Darwinian process with random mutations accumulating in cancer. Advantageous mutations result in clonal expansions at the expense of unfit clones. However, this model is incomplete as it ignores the non-genetic variability of cancer cells. Genetic, epigenetic and environmental variations as well as their timing in tumour development make modelling cancer progression quite challenging, but better insights into the impact of these changes is essential for designing optimal therapies.

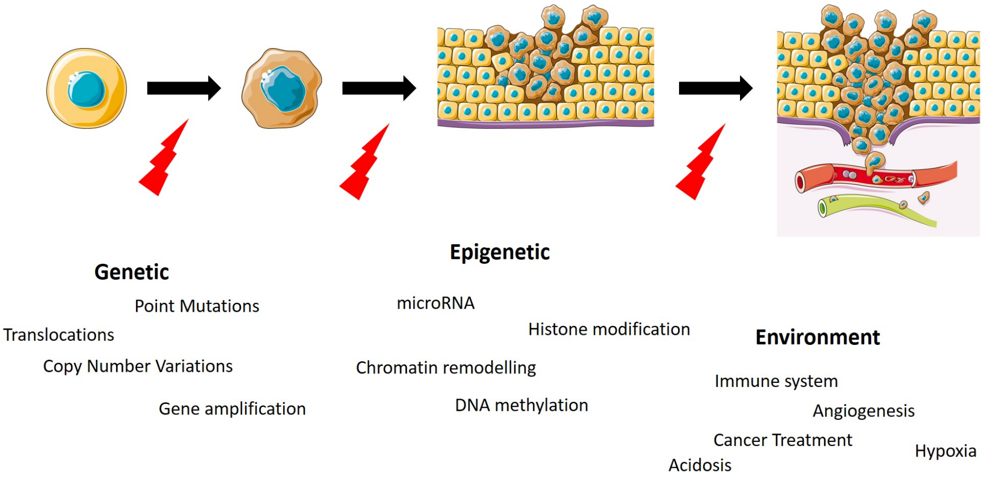


Fig. 1.1 **Cancer Progression**. Cancer evolves through several stages leading to uncontrolled tumour growth and immune invasion until it escapes the primary site to invade other tissues. This step-wise progression towards malignancy can be a long process and involve several alterations in tumour cells and their surrounding. These modifications can be genetic or epigenetic and often respond to new environmental challenges. Cell images from smart.servier.com, licensed under CC BY 3.0, edited from original.

1.1.1 The major biological events in cancer progression

One of the major cancerous alterations are genetic mutations. Gene mutations determine the different routes cancer cell populations can take: they can be deleterious, neutral or advantageous for the cell fitness [3]. While deleterious mutations are capable of driving a population to its own extinction, advantageous genetic modifications can lead to clonal dominance [4]. They can affect a wide variety of cellular characteristics such as cell proliferation [5], cell death [6] or cell division [7] with clear implications for tumour growth. Cancer mathematical models often have a relatively low number of mutations in their equations to simplify analyses and therefore only include those affecting important cell phenotypes for the disease progression. These advantageous mutations also called driver mutations alter genes involved in essential cellular mechanisms and suffice to model and globally interpret how tumour cells grow and interact [8, 9]. The number of advantageous mutations needed to drive healthy cells into tumour formation is tissue-specific: some cancers such as acute myeloid leukemia (AML) requires two or fewer driver mutations while uterine corpus endometrial carcinoma (UCEC) can accumulate six or more [10]. However, passenger mutations are also fundamental to capture the heterogeneous complexity of tumours [11]. These genetic alterations are thought to alter the cell genome with no or little consequences on cell fitness. However, a recent study [12] demonstrates that mildly-deleterious mutations can impact the clinical outcomes of patients due to their higher number in tumours and negative selection effect. Therefore, such mutations could be used as gene targets for future therapies.

Gene mutations are not the only biological alterations affecting tumour cells. This thesis defines epigenetics as the genome modifications which do not implicate alterations of the nucleotide sequence. Epigenetic plays an important role in cancer evolution. For example the biological process which consists of adding of a methyl group to the DNA (DNA methylation) has been reported in several cancers. DNA hypomethylation can increase chromosomal instability [13] while site-specific methylation represses tumour suppressor genes such as the Retinoblastoma gene [14]. Histone modification is another well-studied epigenetic change. Disrupted histone acetylation or methylation alter access to chromatin and can result in major alterations in favour of tumorigenesis [15].

Another challenge cancer cells face in their existence is the developing diversity of their surrounding microenvironment. The environment can induce challenges and opportunities for the development of tumours. In crowded tumours, the formation of new blood vessels, or angiogenesis, is critical to avoid cell starvation and provide vital resources for tumours to continue to expand [16]. The immune system can also be a mortal enemy as well as an essential ally in cancer progression [17]. Indeed, immunosuppressive medication and immunosuppression mechanisms induced by viruses have been shown to increase cancer incidence [18] while tumour-associated macrophages have demonstrated pro-tumoral phenotype by stimulating angiogenesis, suppressing NK and T cell response and promoting metastasis [19]. Treatments also induce new adaptations in cancer cells such as senescence [20] or phenotypic switch [21].

The timings of these diverse biological events highly affect the disease progression. Driver mutations are often found in the majority of cancer cells, as they mostly arise in the early stage of cancers [22]. Detecting early aberrations leading to abnormal cell growth is crucial to identify markers of disease initiation. Exposure of these markers help clinicians to recognise patients with higher risk of developing tumours but also to find the appropriate treatment to target a maximum number of cells. However, early lesions can be challenging to expose due to their smaller size. Barrett's esophagus (BE) is a premalignant state of esophageal adenocarcinoma (EAC). While most of BE cases do not progress to EAC, a study has carried out genomic analyses on BE and EAC samples from the same patients to highlight the oncogenic precursor events [23]. They find two pathways for BE oncogenic transformation, one with accumulation of specific tumour suppressor gene mutations and a second with *TP53* loss followed by whole genome doubling. Sometimes unique sequence of oncogenic events are decisive for tumorgenesis, for example in a mouse model of soft-tissue sarcoma, the timing at which *TP53* gene is lost after *K*-ras$^{\mathrm{G12D}}$ mutation influences tumour formation [24]. Despite the dependence on particular ordering of these events in the precursor lesions, specific genetic and epigenetic alterations of tumours also modify the disease progression in later stages [25]. Unlike in most cancers, amplifications of genes involved in oncogenic cell pathways intervene in the last stages of EAC [23]. Similarly, in the later stages of cancers, some cells exit primary tumours, travel through blood and penetrate new tissues and organs to form metastases. Comparing primary tumours with metastases can reveal metastasis-specific genes and help clinicians to limit tumour spread by targeting them. Turajlic et al [26] demonstrate the substantial role of chromosomal aberrations in clear-cell renal cell carcinoma (ccRCC) metastasis in patients samples. Their work also shows two distinct paths to metastasis among tumours with high or low genetic heterogeneity, underlining the impact of early evolution on later stages.

1.1.2 Heterogeneity in cancer

The large variety of genetic, epigenetic and environmental variations described in the previous subsection and observed within tumours and amongst patients is a major barrier to our full comprehension of tumour progression [27]. Intratumour heterogeneity is caused by the emergence of various clones with distinct aberrations, sometimes in response to their environment. While unfit clones are doomed to disappear and become extinct, competition between subpopulations of cancer cells does not always lead to extinction and several subclones can coexist [28]. This evolution process is called branched evolution and is responsible for the diversity observed in tumours [29]. Intertumour heterogeneity refers to diversity between tumours and is related to intratumour heterogeneity. The major source of intertumour heterogeneity is the tissue of origin which shape the disease fate as a consequence of distinct tissue properties such as cell types and tissue structure. For instance, cell death can be managed by apoptotic or non-apoptotic mechanisms and the chosen mechanism vary among tissues, which was shown to affect DNA-damaging agent potency in tumours from distinct tissue [30]. Furthermore, heterogeneity is also observed in tumours with the same cell type of origin within a tissue. Variations in molecular and cellular processes among tumours contribute to diverse morphologic, phenotypic and metastatic characteristics. Breast cancer is a good example of intertumour heterogeneity. Distinct grades are observed in breast carcinoma and are determined by morphologic parameters such as the mitotic rate [31]. The grade as well as the expression of fundamental hormonal biomarkers, such as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) establish the patient profile and guide physicians for relevant therapeutic protocols [32]. Finally, intertumour heterogeneity also refers the clonal divergence between a primary tumour and its metastases. Cells need to gain certain properties to invade adjacent tissues [33]. Evaluation of 32 clear-cell renal-cell carcinoma metastases has allowed authors to identify important chromosomal aberrations in metastases which frequency correlates with the dissemination tissue suggesting the spread of tumour cells in various tissues correlates with particular molecular alterations [34].

Heterogeneity in tumours manifests alarming clinical consequences in disease progression. First, as a result of mutation accumulation in cancer cells, subclones of different size coexist in tumours. We already know that early driver events are detected in most cells of the tumour, and often are the targets of drug treatments to eliminate a maximum number of cancer cells and reduce tumour burden [35]. However, some minor subclones initially undetectable can be resistant to therapies, expand and cause treatment relapses [36]. Genomic follow-up of a high risk patient with multiple melanoma shows the competitive dynamics of subclones when therapies are applied, but also reveals that the tumour cell subpopulation responsible for the patient death was hardly detectable at diagnosis [37]. Secondly, spatial tumour heterogeneity induces sampling bias and complicate the identification of good prognostic factors. Gerlinger et al [38] found gene-expression signatures from good and bad prognostic markers in a same tumour but in different areas. With 63 to 69% of somatic mutations not detectable across all tumour regions, this study raises the concern about single tumour samples which might not reflect the full complexity of certain cancers. Finally, despite the importance of genomic in heterogeneity, the tumour microenvironment and epigenetic fluctuations also influence tumour organisation. Different environmental conditions between core and peripheric tumour cells induce phenotypic changes, core cells promoting nutrient flow and peripheric cells showing invasion and high proliferation traits [39]. Similarly, epigenetic heterogeneity has been studied in different cancer types, involving distinct DNA methylation, chromosomal and histone aberrations in tumour subclones [40]. For instance, Pastore et al [41] analyse methylation and histone modifications across patients with chronic lymphocytic leukemia (CLL) and identify a large diversity in epigenetic markers resulting in permissive chromatin states across cells. This acquired property might encourage cells to stochastically alternate between different gene expression programs to facilitate the emergence of new cell phenotypes. Overall, it is clear that a better knowledge of tumour heterogeneity will encourage treatment successes.

Assessing heterogeneity in tumours supports untangling tumour progression and predicting patient clinical outcome. One method to evaluate genetic diversity within and between tumours as well as among patients is DNA sequencing. Whole-genome, whole-exome and next-generation sequencing determine the genetic code of individuals by retrieving the order of nucleotides in the DNA. All these methods also recognise mutated DNA that have acquired genetic alterations in their sequence such as substitutions, deletions, insertions, duplications or copy number changes [42], and are therefore important tools for cancer diagnosis and treatment design. Sequencing at different time points and/or in diverse tumour areas has a broad potential of applications from clonal and subclonal evolution history, timing of aberrations or identification of biomarkers [43–45]. In order to study heterogeneity in large samples while conserving spatial information, *in situ* techniques provide a range of methods, including immunohistochemistry (IHC), immunofluorescence (IF) or hybridization. After extraction of tumour sections, these techniques identify heterogeneous properties such as cell morphology, molecular expression or DNA alterations which all can be measured and quantified using computational methods [46]. Non-invasive imaging methods such as x-ray computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET) can also be used to evaluate heterogeneity [47]. Those techniques can currently identify morphologic, vascular and necrosis variations in different regions. New studies aim at developing imaging methods to analyse additional tumour traits such as hypoxia [48]. Finally, statistical methods can also quantify heterogeneity in samples. The true variant allele frequency (VAF) distribution [49] measures clone frequency. It can be used to classify tumours based on their evolution modes and how intratumoral heterogeneity accumulates.

1.1.3 Driver event interactions

The previous subsections identify the numerous biological and environmental players of tumour progression and how the resulting heterogeneity has clinical consequences in cancer. Malignancy certainly necessitates several alterations to drive cancer cells to progress and invade surrounding tissues. Interaction and timing among these events produce specific cell traits impacting their fitness and induce competition dynamics among clones and subclones. Genetic and clonal interactions are keys in tumour development from initiation to primary tumour escape. Better understanding of how they interact will contribute to our ability to fight tumour progression.

"Epistasis" refers to the mechanism by which the outcome of a given gene on a biological trait is enhanced or masked by one or several other genes [50]. This phenomenon obscures our understanding of relationship between gene expression and cell phenotype, encouraging further studies to categorise genes and their cooperation effect on cellular functions. Three types of gene interactions are described in tumorgenesis: synthetic lethality, synthetic viability and synthetic sickness [51]. In cancer, two genes are said synthetically lethal if a mutation in either one of these gene is harmless to cell survival but both gene mutations are deadly to the cell. This concept has major consequences in cancer therapies as targeting a synthetic lethal gene in tumours possessing the other compatible mutation could kill the cancerous cells while being inoffensive for healthy cells [52]. A good example of synthetic lethality is the case of ARID1A mutation in ovarian clear cell carcinomas which mutation was shown to be lethal in combination with EZH2 inhibition due to their antagonistic role on PIK3IP1 expression, a negative regulator of cell proliferation and promoter of apoptosis [53]. The opposed gene interaction to synthetic lethality is synthetic viability, by which the combination of two gene mutations rescue the lethal effect of each single mutation. Drug resistance is often caused by these interactions. Melanomas with a BRAF oncogene mutation are frequently treated with BRAF inhibitors, however, the increased expression of COT reactivates the MAPK pathway in a RAF-independent manner and therefore save cancer cells from death [54]. Finally, synthetic sickness defines the combination of gene alterations that result in a greater reduction of fitness expected by each single event on the cell phenotype. Important in therapeutic marker research, this gene interaction can increase the drug effect on specific tumours. For instance, senescence induction by ECT2 was shown to be particularly enhanced on tumours with a *KRAS* mutation [55]. Overall, better exploitation of this complex genetic interaction landscape will help to identify important cross talks between genetic pathways which will be essential to our comprehension of cell dynamics in clinical settings.

Epigenetic modifiers have also been reported to interact and play an important role in cancer epistasis. Both DNA methylation and histone modifications have been shown to work together in diverse cancer contexts by enhancing gene silencing [56]. In colorectal cancer, the expression of three tumour suppressor genes (*P16*, *MLH1*, and *MGMT*) are silenced through H3K9 hypermethylation and deletion of DNA methylation can reactivate these genes [56]. Similarly, several studies report epigenetic interactions between miRNA, which are important post-transcriptional regulators of gene expression [57], and DNA methylation [58, 59]. Finally, epistasis can be observed among genes altering identical epigenetic mechanisms. *DNMT3A* and *TET2* null mutations were shown to cooperate to accelerate T cell lymphoma despite their antagonistic role in DNA methylation, indicating complex interactions with a combination of independent and interdependent roles for *DNMT3A* and *TET2* [60]. These studies demonstrate the importance of including epigenetics in cancer evolution models to fully capture the complexity of tumour development.

Cancer treatments are also important driver events in tumour progression, they may lead to disease curation as well as therapeutic failure. The clinical outcomes of treatments not solely rely on the drugging protocols and the drug effect on tumour cells. The cell dynamics with their surrounding contribute greatly to the success or failing of therapies. Some treatments themselves have been shown to alter the cell functions in unpredictable ways which can obstruct the primary killing goal of the drug. Jackson et al [61] demonstrate that chemotherapy induced senescence in cells with a wild-type *p53* gene which contribute to reduced tumour regression and relapses. Even more alarming, some treatments have been shown to promote faster regrowth with acceleration of cancerous cell expansion and loss of any treatment benefit [62, 63]. Cancer therapies can therefore modify tumour progression in unexpected ways due to the complex dynamics inside and outside heterogeneous tumours. A good understanding of these mechanisms can prevent therapeutic relapses and failure.

1.1.4 Modelling cancer growth and resistance

I have demonstrated in the previous subsections how interactions between the important biological events of tumour evolution impact the phenotype of cells and the clinical outcomes of therapies. A major consequence of these events and tumour heterogeneity is resistance emergence. A wide range of treatments and therapies is currently available for fighting tumour progression. While some cancer types have excellent prognosis thanks to efficient therapeutic protocols, others remain lethal after sequential administration of drugs and/or radiotherapy due to resistance emergence and relapses. Resistance emergence has distinct origins and in order to better treat patients, understanding how resistance arises and impacts tumour growth will facilitate clinical protocol designs. I have identified in the literature two contradictory scenarios for cancer growth and resistance emergence in tumours, with both presenting two different therapeutic challenges. Both concepts are used in Chapter 3 of this thesis, which focus on the dynamics of sensitive and resistant cells in a lymphoma mouse model.

The first scenario is illustrated by the Luria-Delbruck model. The first biological resistance studies appeared in the 1940s with penicillin and antibiotic resistant bacteria [64]. Despite the slow appearance of antibiotic resistance at that time, some scientists started studying the underlying mechanisms causing this resistance [65, 66]. From these bacteria studies, one model is now well used in cancer resistance appearance: the Luria-Delbrück model [67]. Luria and Delbrück showed in theoretical and experimental work that mutations conferring antibiotic resistance in bacteria are random and not induced by a selective environment. In cancer development, a Luria-Delbrück-like (LD-like) model is typically treated as well-mixed subpopulations of cells, with exponential growth and no competition for resources or space. Using a LD-like model, Diaz et al [68] study acquired resistance through *KRAS* mutation in colorectal cancers treated with EGFR blockade. Results from the model indicate that the resistance, already present before treatment, repopulates the lesion due to the drug killing the sensitive clones allowing a fast expansion of resistant cells with a fully regrown lesion occurring between five to six months after starting the treatment. Later studies have also expanded LD-like models with various growth dynamics, such as the logistic [69] or Gompertz growth [70], to capture realistic population dynamics which might be constrained by space or resource limitation. For tumours with Luria-Delbrück-like dynamics, resistance regrowth is a *fait accompli*, which will, in most cases, end by therapeutic failure due to resistance emergence. In spite of curing these patients, the best treatment strategy may be to improve their quality-of-life by increasing their survival while minimising drug toxicity.

The second concept is more recent and is inspired by game theory. In some cancers, resistance has a fitness cost, as a resistance phenotype might require a higher amount of resources or might slow cell proliferation due to costly phenotypic changes. This fitness discrepancy between sensitive and resistant subpopulations triggers competitive behaviours among populations for space and resource. In the absence of therapy, the fitter sensitive cells will expand and inhibit the proliferation of resistant cells, while during therapy, resistant clones will take over and sensitive cells will die due to treatment. Gatenby et al [71] use these dynamics and propose spatial cancer cell models in which space and resource competition is included. Based on this model, Gatenby also proposed a new treatment strategy that could be used called "adaptive therapy". Adaptive therapy differs from the usual maximal killing therapies, that fails due to resistance, in its intention of not "curing" cancer. Instead, its objective is to stabilise the tumour over time by permitting a significant proportion of chemosensitive cells to survive to avoid the rapid proliferation of resistant cells. To do so, monitoring tumour response to treatment is crucial, dosing/schedule of the next drug application is defined by the previous responses, which help researchers designing optimal therapy strategies. Although this is not a curative therapy, this technique demonstrated its efficiency with an important increased survival for the patients treated with adaptive therapy [72, 73]. Enriquez et al [72] compare different treatment strategies and conclude that high drug concentrations injected at the beginning of treatment followed by regular application of drugs with decreased concentration when the tumour is successfully reacting to treatment cause the exponential tumour growth to plateau and lead to a no-apparent tumour state with no treatment necessary in 60-80% of mice. Increasing patient survival is also beneficial for researchers and clinicians that gain time to search for additional treatment strategies or to start therapies with longer curative responses such as immunotherapy [74]. LD-like and Gatenby models have different therapeutic consequences, and determining which dynamics the tumour belongs to is a first step to optimise treatment strategies.

1.2 Healthy and abnormal blood development in adults

1.2.1 Hematopoiesis

This thesis use distinct modelling techniques to depict the complexity of blood cancer evolution. However, first understanding how healthy hematopoiesis works is crucial to untangle the aberrant dynamics of hematopoietic malignancies. This subsection reviews the main hematopoiesis biological terms and players. Hematopoiesis is the continuous and tightly controlled biological process responsible for the production of our blood cells. Blood has various roles in the human body, from nutrients and hormones transportation, body temperature regulation, blood clot formation following injuries and pathogen protection. While definitive hematopoiesis is located in the bone marrow, primitive hematopoiesis which occurs during early embryogenesis starts in the yolk sac. For the rest of the thesis as our main focus is adult blood diseases, the term hematopoiesis refers to definitive hematopoiesis.

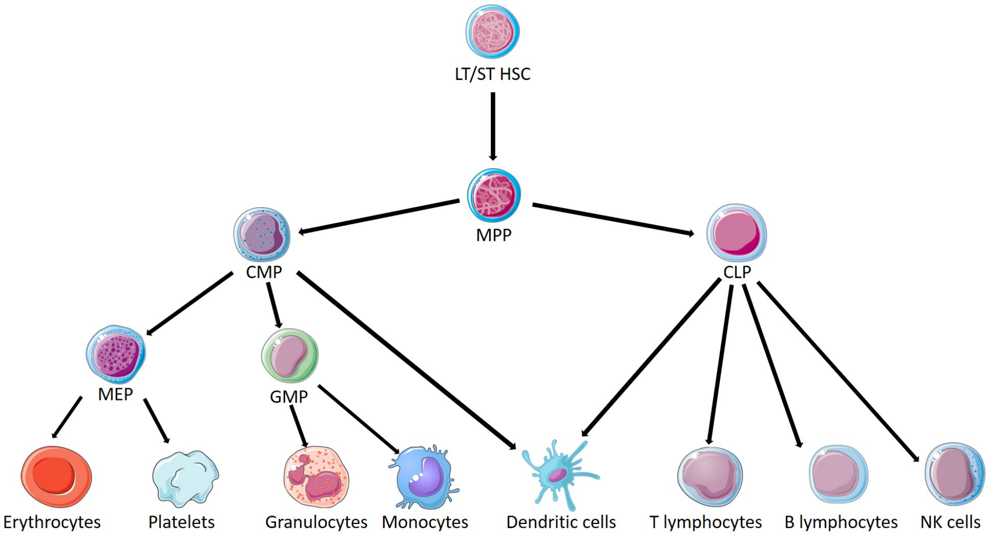


Fig. 1.2 **Hematopoiesis schematic**. LT/ST HSC, long-term/short-term hematopoietic stem cells ; MPP, multipotent progenitors ; CMP, common myeloid progenitors ; CLP, common lymphoid progenitors ; GMP, granuloctye-monocyte progenitors ; MEP, megakaryocyte-erythroid progenitors. Cell images from smart.servier.com, licensed under CC BY 3.0, edited from original.

Hematopoiesis maintains our different blood lineages at a stable level through hematopoietic stem cells (HSCs) [75]. Their ability to self-regenerate, called self-renewal, is responsible for the preservation of normal blood cell production. HSCs are defined as pluripotent as a result of their ability to differentiate into any type of blood cells [76]. Lifespan of HSC is also extremely long. Initially thought to be immortal cells, a recent study shows that HSC lifespan varies from 10 to 60 months [77]. This lengthy lifespan is another crucial attribute of HSC opposed to differentiated cells whose lifespan is much shorter and varies among lineages [78–80].

Hematopoietic precursors can be grouped into different subpopulations as a result of their markers and their ability to differentiate: there are long-term HSC (LT-HSC), short-term HSC (ST-HSC), multipotent progenitor 2–4 (MPP2, MPP3, and MPP4), common lymphoid progenitor (CLP), common myeloid progenitor (CMP), megakaryocyte-erythroid progenitor (MEP), and granulocyte-macrophage progenitor (GMP). The Lineage$^-$, Sca-1$^+$, cKit$^+$ (LSK) cells refers to LT/ST-HSC and MPP, while the LK subset refers to CMP, GMP and MEP subpopulations [81].

We distinguish LT-HSC from ST-HSC by their different aptitude in regenerating all lineages after radiation: LT-HSCs are considered as the true stem cells and can sustain hematopoiesis almost indefinitely while ST-HSCs only produce for several weeks [82]. LT-HSCs committing to differentiation give rise to ST-HSCs which can differentiate into MPP. The self-renewal ability of these cells decreases linearly, with MPP self-regeneration being less than two weeks [83]. The multipotent progenitors commit to lymphoid or myeloid lineage by dividing into common myeloid progenitor (CMP) or common lymphoid progenitor (CLP) [76]. CLPs can differentiate and produce progenitors for the T, B, and Natural Killer (NK) cells [84], while granulocyte monocyte committed progenitors (GMP) or megakaryocyte erythrocyte progenitors (MEP) derive from the CMPs [85]. Finally, MEP cells are responsible for megakaryocyte and erythrocyte production and GMPs produce macrophages and granulocytes including neutrophils, eosinophils, and basophils [86].

Functions of our immune cells can be split between the innate versus the adaptive immune system [87]. The innate immune system is the first defense mechanism against infections and cells from both myeloid and lymphoid lineages are involved in this rapid immune response. This natural immunity is nonspecific, which means that it targets any non-self or foreign organisms, and includes the granulocytes, the NK cells and finally the monocytes which refer to macrophages and some types of dentritic cells.

Macrophages are large monocytic cells found in all tissues of the human body and their main function is phagocytosis which consists of ingesting harmful mechanisms but also dead cells. They recruit and trigger other hematopoietic cells' immune response by presenting foreign antigens to T cells or releasing cytokines [88, 89]. Macrophages can be split into two types, the M1 and M2 macrophages: M1 macrophages are activated by IFNγ and/or tumour necrosis factor (TNF) and have antimicrobial properties while M2 macrophages are activated by IL-4 and/or IL-13 and are associated with tissue repair properties and apoptotic cell clearance [90, 91].

Granulocytes are also members of our innate immune system and are named after the granules present in their cytoplasm. There are three main types of granulocytes: neutrophils, eosinophils, and basophils. Neutrophils compose the vast majority of our innate immune system cells as our body produce about 100 billions of them per day [92]. They are therefore the first cells to arrive on the infection site and as macrophages, phagocytosis is one of their main action against pathogens. Another microbial killing strategy of neutrophils is carried out through degranulation. Degranulation process consists of the release of granular antimicrobial molecules such as myeloperoxidase (MPO) [93] or neutrophil elastase (NE) [94]. Neutrophils are also capable of neutrophil extracellular traps (NETs) generation, composed of deoxyribonucleic acid (DNA), histones, and antimicrobial granule proteins. Those NETs aims to trap and kill invasive bacteria [95]. Finally, neutrophils recruits antigen-presenting cells (monocytes and dendritic cells) by secreting cytokines. Despite releasing a lower amount of cytokines than other immune recruiters, the large number of neutrophils results in an efficient immune cell recruitment during infections [96]. Eosinophils and basophils are the two other types of granulocytes, and unlike neutrophils they are non-phagocytic but similarly to neutrophils, eosinophils can trigger invader death by degranulation [97]. Eosinophils are mostly activated during parasite infections and are associated to allergic diseases [98]. Finally, basophils form the least abundant immune cell category and are largely associated to allergic inflammation through secretion of histamine, a known compound triggering allergic reactions [99].

Natural killer (NK) cells are cytotoxic lymphocytes which can induce the death of cancerous and infected cells through different killing processes such as lysis [100]. Lysis results in membrane destruction and release of the cell cytoplasmic compounds including the virions if the cell was infected by a virus [101]. NK cells contain of granules which can release proteins such as perforin that form pores in cell membranes. This process allows associated proteins such as granzymes to enter the cell and induce apoptosis [102].

Finally, dendritic cells (DCs) cannot be clearly classified as lymphoid or myeloid lineage. Those cells are important players in the coordination between innate and adaptive immune system and can be subdivided into three different types: the conventional DCs (cDCs), the plasmacytoid DCs (pDCs), and the monocyte-derived DCs (moDCs), with the latter being quite a controversial category due to their ambiguous presence *in vivo* [103]. The major role of cDCs is to process and present antigens to T cells [104]. The plasmacytoid DCs are specialised dendritic cells that produces type I interferons (IFNs) during viral infection and therefore promote antiviral immune responses [105]. pDCs are generated by a common DC progenitor (CDP) that also generates conventional DCs and can be derived from the CMP and CLP cells [86].

Our second defense barrier is composed of lymphocytes, the T and B cells, which are part of our adaptive immune system. T and B cells derive their name from the site they both mature: thymus for the T cells and the bone marrow for the B cells [106]. The adaptive immune system is a slow but gives a specific immune response against foreign organisms that our immune cells can recognise and target through response to pathogen antigens. When an unknown antigen is recognised, the adaptive immune system is in charge of developing an immunological memory, so that a quicker response can be induced in later infections from the same pathogen [107].

Lymphocyte B cells secrete antibodies, also known as immunoglobulins, which targets specific pathogens by antigen recognition [108]. After their development in the bone marrow, B cells are released to be matured in the spleen and the lymph nodes. Mature B cells are also called "naive" B cells until they first encounter an antigen that fits its antibodies. After this exposure, naive B cells become memory B cells or plasma cells [109]. Memory B cells express the specific antibodies that have initially recognised the pathogens. They are dormant cells in charge of immunological memory, which will trigger a faster and stronger response if a second infection from the same pathogen happens. Plasma cells possess the same antibodies, but are only able to secrete them. They do not proliferate but have a memory role for later infections [110].

Lymphocytes T, unlike lymphocytes B, do not produce and secrete antibodies, but express important receptors, such as T cell receptors (TCRs), CD8 and CD4 [111]. These receptors are essential for the three fundamental functions of T cells: activation of the immune system after infection, autoimmune disease prevention as well as pathogens and infected cell removal. All T cells have TCRs, but not all express CD4 and CD8. Cytotoxic T cells only carry the CD8 receptor, while helper T cells only the CD4 receptor [112]. T cells can detect self and non-self cells thanks to the Major Histocompatibility Complex (MHC), which is a group of genes that code for receptor molecules on the surface of individual cells and play a role in antigen presentation [113]. All healthy human individuals express MHC but two individuals rarely share identical MHC molecules. Cytotoxic T cells recognise self MHC class I (MHCI) molecules which are found on our nulceated cells. If a cell presents MHC molecules not recognised by the T cell, the latter is destroyed and an immune response is triggered. Cytotoxic T lymphocytes are also activated when their TCR and CD8 co-receptors detect MHCI presenting mutated or viral proteins [114, 115]. Once activated, these lymphocytes kill the infected or cancerous cells by releasing perforin and granzyme molecules. On the other hand, helper T cells activates our immune cells by releasing cytokines during an infection: they trigger the maturation of B cells and activate cytotoxic T cells [116]. Helper T lymphocyte stimulation is induced by the recognition with their TCR/CD8 receptors of

MHC class II (MHCII) presenting cells which are expressed on the membrane of antigen-presenting immune cells (macrophages, dendritic cells, B cells). These antigens are loaded on the MHCII of specific immune cells and include proteins, peptides or polysaccharides that originate from the individual (self-antigen) or from the external environment (non-self antigen). Finally, regulatory T cells are characterised by both CD4 and CD25 co-receptors which regulate our other immune cells to reduce autoimmune diseases [117]. They can suppress the activation, proliferation and cytokine production of helper T cells and cytotoxic T cells, as well as suppress B cells and dendritic cells.

1.2.2 Blood cancers

Four types of blood malignancies are studied in this thesis: lymphoma in Chapter 3, acute myeloid leukemia in Chapter 4, 5 and 6, myeloproliferative neoplasm diseases in Chapter 4 and mixed phenotype acute leukemia in Chapter 6. Blood cancers are diseases arising from any type of our blood cells and dramatically affect hematopoiesis. The maintenance of blood cell production lies on hematopoietic stem cells. Stem cells after division can differentiate into two main lineages: myeloid and lymphoid. Perturbation in either the myeloid or lymphoid lineages can result into three types of hematologic malignancies: leukemia, lymphoma and myeloma [118]. Leukemia affects the white blood cells of the bone marrow, lymphoma the lymphocytes in the lymph nodes and finally myeloma the plasma cells [119–121]. In this classification of blood cancers, leukemia also includes two additional myeloid malignancies: the myelodysplastic syndromes (MDS) and the myeloproliferative neoplasms (MPNs). MDS is a group of diseases characterised by an ineffective hematopoiesis producing a low number of blood cells, which causes anaemia or cytopenia in patients [122]. Myeloproliferative Neoplasms (MPNs) are chronic diseases of the myeloid lineage characterised by an excessive production of fully functional terminally differentiated blood cells [123].

Classical MPNs have been classified into 3 entities: polycythemia vera (PV), essential thrombocythemia (ET), and primary myelofibrosis (PMF) (Fig. 1.3). PV is the most common MPN and is characterised by *JAK2* mutation and erythrocytosis, an abnormal increase of the number of red blood cells. ET is identified by thrombocytosis, an overproduction of platelets. Finally, PMF is the least common and most aggressive and is defined by its bone marrow fibrosis, an excessive fibrous tissue formation. PMF can be primary or be a later stage of PV and ET.

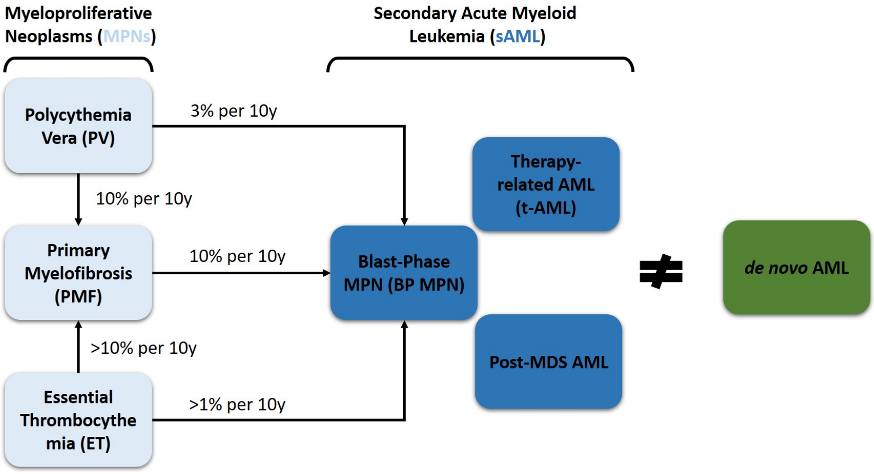

Myeloproliferative Neoplasms (MPNs)
Secondary Acute Myeloid Leukemia (sAML)
Polycythemia Vera (PV)
3% per 10y
10% per 10y
Primary Myelofibrosis (PMF)
10% per 10y
>10% per 10y
Essential Thrombocythe mia (ET)
>1% per 10y
Blast-Phase MPN (BP MPN)
Therapy-related AML (t-AML)
Post-MDS AML
≠
de novo AML

Fig. 1.3 **Myeloid disorders**. MPN diseases can progress to AML with different frequencies. While secondary AML arise from a previous myeloid malignancy such as MDS or MPN, *de novo* AML patients have no clinical history of blood diseases.

Despite the relatively good prognosis of these diseases, MPN patients are at high risk of thrombosis and can develop blast phase MPN (MPN-BP) [124]; a subtype of the blood cancer Acute Myeloid leukemia (AML) with poor survival outcomes [125]. The frequency of MPN transformation to blast phase MPN is highly related to the initial MPN disease type: PMF has the highest incidence with a risk of 10-20% to develop MPN-BP in the first ten years [126, 127] against 3% for PV patients [128] and less than 1% for ET patients [129].

AML itself is an aggressive blood and bone marrow malignancy defined by the uncontrolled growth of the myeloid progenitor cells along with a myeloid-lineage differentiation arrest [130]. AML has one of the lowest number of mutations per case among cancer types [131]. In a study with 200 *de novo* AML patients, authors found on average 13 coding mutations per patient with only 5 recurrent mutations in all genomes [132]. Two types of AML have been described in the literature: *de novo* AML and secondary AML (sAML) (Fig. 1.3). The general term secondary AML refers to AML transformation after MDS, MPN (MPN-BP) or after therapy and represents a high proportion of AML patients [133]. Clinicians and researchers initially believed that secondary AML patients had worst prognosis than *de novo* patients who do not possess any known medical history of blood diseases [134]. However, several recent studies have shown that age and cytogenetic risk could be the actual markers for good or bad prognosis in AML [7, 133, 135]. Indeed, secondary AML patients are on average older than *de novo* patients [7, 135]. Plus, the spectrum of cytogenetic abnormalities is similar between de novo and secondary AML, only higher frequency for complex karyotypes are observed in secondary AML [136]. This shows that AML should be classified by their genetic characteristics instead of preleukemic history.

AML patients present a broad range of morphologic, cytogenic and immunologic features which all are associated with diverse clinical effects. Two AML classification systems are widely used nowadays: the French–American–British (FAB) classification emerged first using morphology, cytochemistry and blast percentage, then the World Health Organization (WHO) added cytogenetics, dysplastic features and AML history (primary or secondary) [137]. In this thesis, FAB classification and cytogenetics are used for AML patient stratification in various analyses. FAB clusters patients into 8 groups from M0 to M7. M0 is the most undifferentiated subtype, while M3 and M5 display higher number of early monocytic/granulocytic blasts. M6 and M7 are rare subtypes and are respectively associated with the erythroid and megakaryocyte lineages. Cytogenic abnormalities such as gene translocations are common genetic dysregulation in AML [138] and are not included in FAB classification. Identification of these alterations, as well as FAB subtypes, are essential for disease diagnosis and the choice of treatment protocol.

In some rare cases, assigning a single lineage of origin in leukemia is difficult. Sometimes, both lymphoid and myeloid blasts (bilienal) develop in leukemia or some tumours are found with blasts (biphenotypic) expressing both lymphoid and myeloid markers. These patients are diagnosed with Mixed Phenotype Acute Leukemia (MPAL) [139]. As reviewed in [140], research in MPAL is quite sparse for several reasons. The rarity of the disease, estimated to vary from 1 to 5 % of leukemia cases, but also its subjective diagnostic definitions and large phenotypic and genotypic diversity among patients make MPAL a complex disease to diagnose and treat. In a recent study comparing MPAL survival to other leukemia types, authors found that MPAL patients have the worst prognosis regardless of patient age [141]. It should be noted that often associated to children, MPAL also exists in adults and the genomic landscape and prognosis between pediatric and adult MPAL differ. The overall survival already quite poor of MPAL patients is even poorer in adults [142]. Regarding classification, two systems have tempted to classify MPAL: the European Group for the Immunological Characterization of Leukemias (EGIL) [143] and the World Health Organization (WHO) in 2008 and updated in 2017 [144]. Both use immunophenotype characterisation, but diverge on specific genomic alteration characteristics such as KMT2A rearranged (KMT2Ar) or BCR-ABL fusions. Finally, regarding MPAL treatment, current methodology is to apply Acute Lymhphocytic Leukemia (a blood cancer type with an overproduction of immature lymphocytes) directed therapy which shows better results than AML directed therapy. However, due to patient heterogeneous characteristics, other treatments may improve clinical outcomes of diseases with specific aberrations [140].

1.2.3 *JAK2* and *TET2* in Myeloproliferative Neoplasms

In Chapter 4, I investigate the impact of mutation order of two genes in MPN diseases: *JAK2* and *TET2*. JAK2V617F is the most common oncogenic event in MPN and is therefore well described in the literature [145–147]. It has been shown that a single *JAK2* mutation in an unique hematopoietic stem cell can initiate MPN in a mouse model [148]. *JAK2* mutation increases red blood cell production, also called erythropoiesis, by skewing myeloid differentiation toward the erythroid lineage and by expanding disproportionately megakaryocytic/erythroid progenitors (MEP) over other myeloid progenitors [149]. The role of JAK2V617F in stem cell expansion however remains unclear [150]. In contrast, the interest for *TET2* in blood studies is quite recent. First discovered in MPN in 2008 by Delhommeau et al [151], *TET2* loss has been associated with myelodysplastic syndromes (MDS), chronic myelomonocytic leukemia (CMML), acute myeloid leukemias (AML) and secondary AML (sAML) [152–155]. In MPN diseases, mutational frequencies of *TET2* loss approximate 16% in PV, 5% in ET, 17% in PMF, 14% in post-PV MF and 14% in post-ET MF [156]. Several studies have shown the importance of *TET2* in hematopoiesis as loss of *TET2* leads to increased self-renewal of hematopoietic stem cells (HSC), expansion of the hematopoietic stem/progenitor cells as well as a skewed differentiation into the monocyte/macrophage lineage [157–159]. Finally, it has been shown that *JAK2*/*TET2* double mutant cells develop severe MPN diseases [160] and insights into how these two genes interact will help to untangle how gene interactions impact on disease prognosis.

Despite some unquestionable roles for *JAK2* and *TET2* mutations in hematopoiesis, the functions of both proteins in diverse lineages is highly controversial and contradictory findings can be found in the literature. These dissimilar results can be explained by distinct experiment protocols or cells at different differentiated states. In the next subsections, I review the literature for a better understanding of *JAK2* and *TET2* role in different blood cell populations. Conclusions of this literature analysis will be used for the molecular networks in Chapter 4. Summary of this review is shown in Figure 1.4.

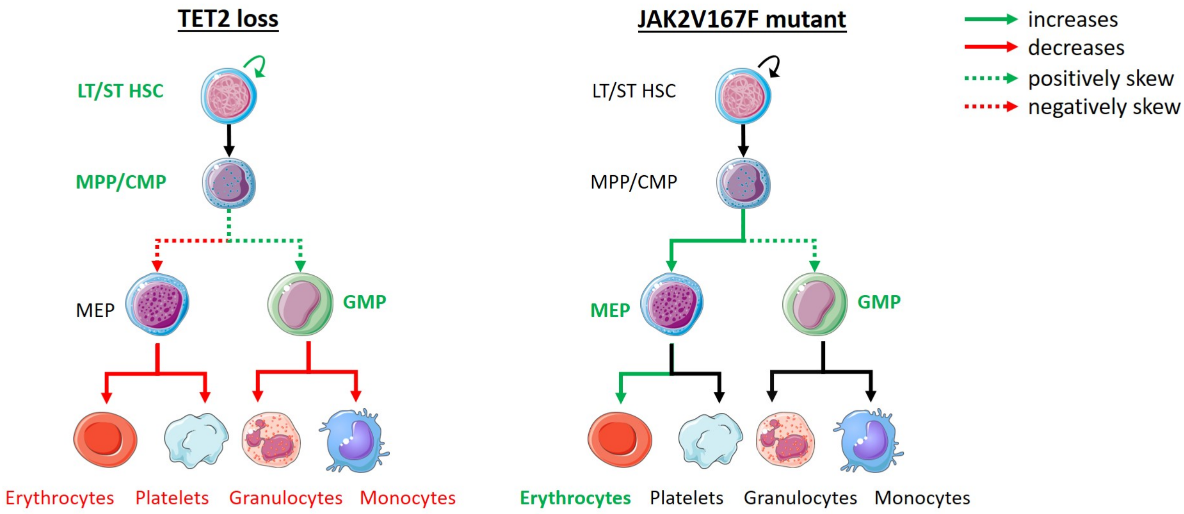


Fig. 1.4 **The effect of JAK2V617F and *TET2* loss mutations in the different hematopoietic cell populations.** This figure shows in green immune cells that are expanded while red indicates a decreased number of cells. Arrows are also coloured in green and red to indicate respectively a positive or negative differentiation. Dotted arrows represent weak skew towards certain lineages. LT/ST HSC, long-term/short-term hematopoietic stem cells ; MPP, multipotent progenitors ; CMP, common myeloid progenitors ; GMP, granuloctye-monocyte progenitors ; MEP, megakaryocyte-erythroid progenitors.

TET2 mutation phenotype

The critical role of *TET2* in genome stability has been recently highlighted [161]. Moreover, healthy hematopoietic stem cells have high levels of *TET2* expression [162]. It is therefore not surprising that *TET2* loss in hematopoietic stem and progenitor cells is associated with myeloid malignancies [163]. *TET2* role in cancer is uncontroversial, as well as is its involvement in hematopoiesis and monocyte differentiation [164]. However, its precise role in the different lineages of hematopoiesis is unclear. Researchers have designed *TET2* knockdown mouse models and study the consequences of this mutation in hematopoiesis. I review in the next paragraphs their different findings.

TET2 has been reported to have different effects on hematopoietic cell populations. Most papers agree on the increased number and enhanced self-renewal capacity of stem cells harbouring a *TET2* mutation [165, 150, 158, 157, 162, 166–168, 60, 160, 169]. Several papers mention an upregulation of the c-kit marker [150, 168, 167]. Similarly, Chen et al [150] observe an elevated self-renewal gene signatures in *TET2* mutant mice. However, few papers found dissimilar results. Kameda et al [160] do not find increased LK and LT HSC, while Quivoron et al [166] only observe a modest increased of LT HSC compared to ST HSC. Overall, most paper agree to a larger number of HSC as well as increased self-renewal capacity in *TET2* mutant models and therefore this is the hypothesis retained for the computational model.

Regards to the progenitor cells, literature results diverge. Several papers note the increased number of CMP cells in *TET2* deficient mice [158, 157, 166], while other notice the expansion of immature myeloid precursors such as MPPs [167, 169]. However, not all papers agree with these findings. Li et al [162] observe no change in blasts and immature myeloblasts and similarly, Pronier et al [170] found constant levels of immature myeloid precursors. *TET2* loss effect on GMP is debatable as well. Several papers indicate an increased number of GMP [150, 158, 162], while some studies do not see any change [157, 166, 169]. Finally, several experiments [150, 157, 169] report no increase in megakaryocyte-erythroid progenitors (MEP), while Quivoron et al [166] notice a significant increase in proerythroblasts and a decrease in the number of late erythroblasts. Also, erythroid infiltration and increased CFU-E/BFU-E (cells preceding proerythroblasts) are observed in mutant mice [162].

With respect to differentiated cell populations, the granulocyte-monocyte lineage is increased by *TET2* loss. *TET2* mutation enhances both granulocyte and monocyte lineages [150, 158, 162, 60, 170, 161]. However, Pronier et al [170] observe a higher number of monocyte compared to granulocyte. For the erythroid lineage, most papers notice a decrease in the red blood cells [162, 170, 165]. Nonetheless, a couple of papers find an increased number of erythoblast and erythroid precursors [161, 162].

Finally, for the differentiation status of hematopoietic cells with a *TET2* loss mutation, papers indicate that *TET2* loss induces decreased differentiation with less differentiated cells [158, 157, 166]. Most studies highlight a skew toward the granulocyte-monocyte lineage [158, 157, 166, 168] and some indicate a preference for monocytes [170, 162]. Impaired erythroid differentiation with insufficient erythropoiesis and accumulation of erythroblasts is mentioned in Li et al [162]. However, some experiments observe a predominance of the erythroid lineage associated with the myeloid expansion generated by *TET2* loss [166, 168].

To conclude on *TET2* mutation impact on blood cells, myeloproliferation, enhanced self-renewal of hematopoietic stem cells and immature progenitors, as well as reduced differentiation with a skew towards the granulocyte-monocyte lineage are confirmed phenotypes in mutant experiments. The role of *TET2* in the erythroid lineage remains unclear. One could however hypothesise that *TET2* loss increases myeloid immature precursors and obstructs erythroid differentiation which leads to an increase of early erythroid precursors with a reduced number of erythrocytes as a consequence of the granulocyte-monocyte predominance.

JAK2V617F phenotype

JAK2V617F mutation is the most frequent aberration in Myeloproliferative Neoplasm diseases [171]. There are several lines of evidence supporting its role in erythropoiesis and PV/ET diseases, but its exact function in other cell types and lineages is more ambiguous. In this short review, I describe the outcomes of JAK2V617F knock-in mouse models in the hematopoietic stem cells and progenitors as well as its potential role in differentiation.

One of the main challenge in JAK2V617F function description is to untangle its role in hematopoietic stem cells. A broad and contradictory set of phenotypes have been reported. Distinct studies identify neutral, advantageous or disavantageous function of JAK2V617F on stem cell survival and proliferation.

Several knock-in mouse models demonstrate neutral change in the number of HSC. Among them, two papers [149, 172] do not observe any increase in the number of LSK cells. Similarly, Akada et al [173] observe only a modest increase of HSC number. Finally, a last study [174] report JAK2V617F as a non-driver mutation for clonal expansion.

On the other hand, some papers suggest a positive role for *JAK2* mutants in HSCs. Despite their initial observation reported above, Mullally et al [175] in a more recent study notice a gradual clonal advantage for *JAK2* mutant cells after one year. Authors interpret the neutral change in HSC number in their first paper by analyses done over a short period of 16 weeks. Another paper find a cell cycling advantage for *JAK2* mutant LSK cells which results in increased number of LSK cells in spleen and LT HSC and MPP cells in the bone marrow [150]. Finally, Kubovcakova et al [176] also underline the advantageous competition of *JAK2* mutants with an increased number of LSK cells in mice.

Several studies however support a deleterious effect of *JAK2* mutation for HSC growth. A significant number of papers [177, 172, 148, 178] demonstrate *JAK2* role in DNA damage. As a consequence, HSC have reduced self-renewal and increased senescence properties [177, 172]. HSC self-renewal is also reduced in *in vitro* experiments in Kent et al [179]. Authors also demonstrate an increased symmetric HSC cell division with a trend toward differentiation, which leads to a reduced number of stem cells. Finally, Kameda et al [160] find that *JAK2* mutant mice show over time a reduced number of LSK/LK cells with decreased competitive advantage and self-renewal. They do not find elevated DNA damage, but increased cell cycle which can explain the exhaustion of LSK cells.

Finally, some experiments find conflicting results. For example, Lundberg et al [148] observe increased quiescence gene signatures with increased DNA damage in their *JAK2* mutants, however the total number of LSK is two-fold increased compared to wild type animals due to cell division boost. This finding could be explained by increased DNA repair mechanisms in which *JAK2* could be involved [178]. Another knock-in mouse model [180] show that *JAK2* mutation gives a competitive advantage to LT HSC with reduced apoptosis and increased proliferation. Authors note however that the discrepancies in LSK, GMP, CMP cell numbers between the knock-in and wild type mice models increase with mouse age and therefore is mostly apparent in older mice, whereas the increased number of MEP in knock-in mice seemed prematurely established. These results seem to agree with Mullally et al findings [175] as only older mice showed increased number of HSC. The overall conclusion of their study is that JAK2V617F might only give a subtle advantage to HSCs which can be detected after several months in mice and several years in humans [180].

The effect of *JAK2* mutation on the granulocyte and macrophage cell lines is unclear. One evident point is the involvement of *JAK2* in increased MEPs [177, 172, 149, 150, 173]. I therefore focus on other lineages in the rest of the subsection.

Some papers find decreased or neutral change in myeloid/granulocyte/macrophage progenitor numbers in their *JAK2* knock-in mouse models. For example, Li et al [177] show a decreased number of myeloid progenitors, while one of their recent paper [172] show no change in GMP frequency. Finally, Mullally et al [149] found an overall increased myeloid progenitors number due to MEP increase while GMP number was unchanged.

However, some studies contest these findings. For example, a couple of papers [173, 179] observe an increased number of GMPs and Chen et al [150] notice a signature increase from genes involved in myeloid precursors. Increased metamyelocytes (early cells in granulopoiesis) is described in Marty et al [181]. Kameda et al [160] observe highly proliferative GMPs, while more papers describe an increased number of myeloid progenitors [180, 148]. Overall, most papers agree on the increased myeloid progenitors, indicating an increased number of GMPs in *JAK2* mutant models.

The case of the common myeloid progenitor (CMP) number is also controversial. CMP numbers can be unchanged [149], decreased [150], or slightly increased [173]. Globally, the CMP population does not seem highly impacted by *JAK2* mutation.

Finally, papers studying differentiation in *JAK2* mutants show clashing results. Li et al [177] do not observe any sign of abnormal erythroid or myeloid differentiation. Nevertheless, Kent et al [179] report an increase for differentiation markers in stem cells. Mullaly et al [149] find that *JAK2* mutation directs hematopoietic differentiation within the LSK compartment into CMP/GMP/MEP, which confirm the results found another paper [148] in which authors demonstrate an increased expression of genes involved in myeloid/erythroid differentiation in *JAK2* mutant LSK cells. Finally, Kameda et al [160] notice an increased of pre-erythroid colonies signature genes. Collectively, one can assume that *JAK2* mutation favours differentiation toward myeloid cells, with a main increase for the erythoid lineage.

In conclusion, *JAK2* is essential for MEP and GMP expansion as well as for skewing differentiation towards myeloid and erythroid lineages. *JAK2* role in HSC and early stage of hematopoiesis most likely is only subtle and noticeable after a long period of time, I therefore assume in the rest of this thesis that *JAK2* mutation has no impact on stem cells or early myeloid progenitors.

1.3 Computational modelling of cancer progression and evolution

This thesis interrogates the importance of timing in blood cancers by simulating and analysing biological computational models. A computational model is a formalisation of fundamental mechanisms, which can be analysed with computers and compared against data. These models aim to reproduce the observed characteristics of a biological system by including the known properties of the different elements [182]. However, biological systems are remarkably complex. The incorporation of every detail would be computationally highly costly and prone to error due to uncertainty. Abstraction by capturing only necessary processes is therefore an ideal option for most systems [183, 184]. While these models cannot replace experiments, they represent a faster and economical way of reproducing important biological processes by reducing the number of experiments and avoiding failed experiments which is primordial when using animal models. Experiments suggest new hypotheses that the model can assess while making new predictions [185]. Predictions can be verified in experiments which will determine the next level of model refinement.

Diverse computational models have been used to study hematopoiesis and blood cancers. In 2008, a paper shows using ordinary differential equations (ODEs) with stochastic simulations and experiments the presence of a dormant population of HSCs which is important for homeostasis and which can be activated following injuries or biological stimulation [186]. Another study builds a genetic network of Chronic Myeloid Leukemia progression and proposes thanks to *in silico* deterministic simulations new combinatorial therapeutic targets [187]. Both models find interesting results using different biological scales, simulation features and modelling techniques. This emphasises the great ability of computational modelling to clarify complex biological dynamics in blood.

However, many questions remain unanswered in blood cancer evolution. The appropriate models to answer those depend on the biological context and the available data. Different data origins (primary tumour, mouse model or cell line) induce different dynamics due to the species characteristics and the spatial set-up of the experiments which also impact on how we model biological systems. In order to investigate tumour progression in diverse hematopoietic malignancies, I must determine the spatial scale, the variable characteristics, the simulation type and finally how to efficiently parametrise the models.

1.3.1 Space scales in cancer modelling

Four main space scales are typically considered in cancer modelling: the atomic, the molecular, the multi-cellular and the macroscopic scales (Fig. 1.5). While most studies initially focused on a single level [188], multi-scale models are developed to fully capture tumour development complexity [189]. This thesis mainly focuses on the molecular and multi-cellular scales, but a brief overview of the four scales is given here.

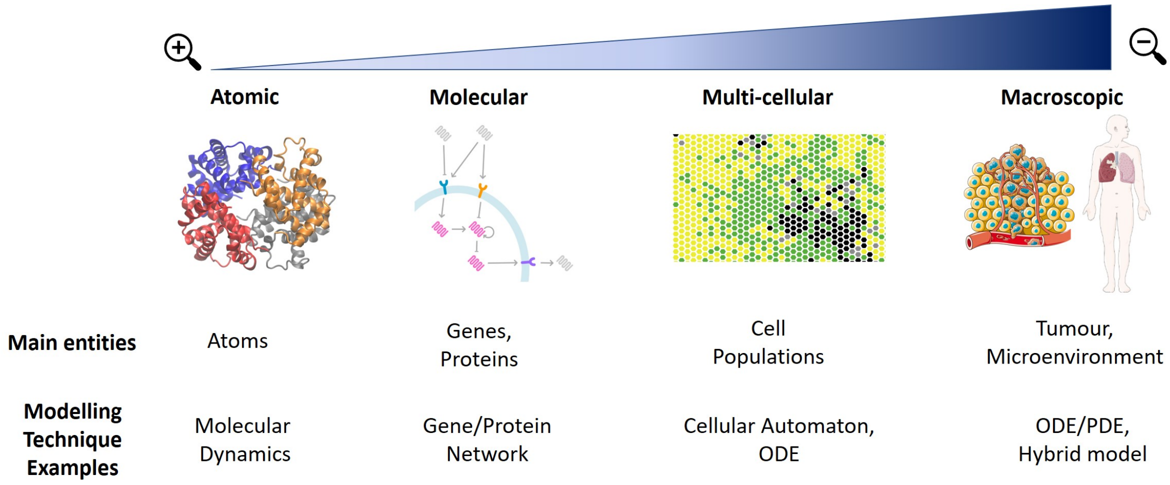


Fig. 1.5 **The four scales in cancer computational modelling**. Atomic, molecular, multi-cellular and macroscopic scales focus on distinct biological entities. The length and time ranges broadly increase from atomic to the macroscopic scale. Some modelling techniques such as molecular dynamics (MD) are specific to a scale, while ODEs and PDEs can be applied to most scales. Atomic scale image from VMD [190], molecular network image from BioModelAnalyzer [191] and cell automaton image from [192]. Cell images from smart.servier.com, licensed under CC BY 3.0, edited from original.

The smallest scale, called the atomic scale, studies the structure and dynamics of biological molecules as well as their interaction with their environment. Atoms interact with each other through a wide variety of interactions such as hydrogen bondings or electrostatic interactions [193]. The modelling of these systems is mainly carried out with Molecular Dynamics (MD) simulations. Molecular dynamics simulations capture the position and motion of every atom at every time point and have helped to untangle many biomolecular challenges such as protein folding and conformational changes [194]. A major limitation of

MD is its limited ability to model phenomena occurring over long timescale as a result of long simulation times [195].

Similarly to MD, the molecular scale focuses on the molecular components of the cells. However, unlike MD which is limited to nanosecond reactions, molecular modelling illustrates behaviours and interactions of molecules in their environment occurring over longer timescale. The predominant methods are gene regulatory networks and protein interaction networks [196, 197]. They explore important cell signalling mechanisms from signal transduction to ligand/receptor activity leading to intra-cellular molecule interactions. These complex cellular pathways converge toward the modification of a cell phenotype and function. These models are particularly important when searching and designing new treatment to ensure for example the virtual screening of potential drug compounds prior to testing [198]. However, the absence of cell interactions in these models limits insights about how cell phenotype alterations impact tumours with a large variety of cell populations. I use the molecular scale to build a molecular network in Chapter 4 and investigate important gene interactions in MPN progression. In Chapter 6, an algorithm inferring small molecular motifs is applied to explain patient stratification in the TCGA AML dataset.

The next scale integrates cell interactions and is called the multi-cellular scale. As the name suggests, the main entities of these models consist of cell populations in which each cell has its own characteristics and can interact with its neighbours. Models include the features of population dynamics studies such as cell-cell interactions and competition, cell-matrix interactions, resource distribution and cell phenotypic alterations [199]. Several techniques are available for these models, such as cellular automaton, rule-based models or partial/ordinary differential equations (PDE/ODE) [200–202]. Heterogeneity and cell competition in tumours are the important domains of investigation at this scale [203]. Work in Chapter 3 explore with rule-based models (described in the Methods chapter) the multi-cellular scale by looking at sensitive and resistance cell dynamics in lymhpoma.

Lastly, the largest scale is the macroscopic scale where the tumour itself is the main model entity. Models examine the dynamics of the tumour including its shape and morphology as well as how it vascularises and spreads to form metastases [204]. The elevated number of cells constrains modellers to specific modelling techniques such as PDEs which include continuous spatial characteristics and are optimal to study tumour dynamics and whole organism models [205].

Multi-scale models integrate distinct biological levels to simulate natural processes which involve various range of space and/or time [206]. Cancer progression solicits various genetic signalling and cell interactions which may take minutes or hours, but also activates complex phenomena such as angiogenesis and metastasis which may occur over months or years.

Hence, multi-scale models are crucial to simulate these spatio-temporal ranges observed in tumour development [207]. They are particularly important and well-used in cancer therapeutic research for the discovery of new molecular drug targets to associate molecular dynamics to cell phenotype alterations [208] or to measure treatment effectiveness on distinct cell traits and microenvironment [209]. Despite the undeniable appeal of multi-scale models, they are also confronted to several technical challenges, a major one being their mathematical complexity. The construction of the biological levels must be well-organised as all levels influence each other and might not use the same abstraction (discrete or continuous, stochastic or deterministic), therefore keeping the model consistent might prove to be a difficult task [210]. The diversity of techniques for the distinct stages can also require different domain of expertise and encourage collaborations [211]. Finally, the large number of biological entities composing these models can also lengthen simulation time and therefore force models to be simplified or to reduce the number of cells for example [212].

1.3.2 Model abstraction: discrete or continuous variables? Stochastic or deterministic simulations?

	Discrete	Continuous	Deterministic	Stochastic
Advantages	• Model interpretability • Biologically heterogeneous individuals • Biological tracking	• Large systems • Numerical and Analytic Analyses • Computationally Inexpensive	• Large systems • Fast simulations • Simpler analyses	• Include biological randomness • Explore stochastic outcomes
Limitations	• Computationally expensive if many entities • Small number of individuals	• Biologically homogeneous populations • Parameter and model structure dependence	• Ignore biological randomness	• Slow simulations • Complex analyses • Small systems

Table 1.1 **Overview of the advantages and limitations of discrete/continuous and deterministic/stochastic model resolution**.

The scale of the model can impact on the variable characteristics and the level of abstraction: how should the biological elements of a model be represented? Should they have discrete or continuous values? How is time treated at different scales? Broadly, studies focusing on molecular interactions tend to choose discrete models such as Boolean networks while atomic, multi-cellular and macroscopic models use continuous or hybrid models [213]. Hybrid models combine discrete and continuous variables. Discrete modelling techniques such as agent-based models, cellular automata or discrete networks are attractive methods to track each biological entity. Update of variable states are determined with simple transition rules, which are interpretable to any scientists with or without mathematical/computational background and makes those models easy to visualise. Plus, the model execution is independent of algorithms. However, the major drawback of these methods is that they can be computationally lengthy if the number of variables is high due to the storage of all biological entity properties [214]. Model reduction of discrete models can help to simplify computations by reducing the number of reactions for example [215], which is unachievable in non-discrete models. On the other hand, continuous models such as ODEs and PDEs perform efficiently for larger biological systems, as biological entities are viewed as a population with similar characteristics. Differential equations can be solved analytically or numerically and their analyses are often computationally inexpensive. The trade-offs compared to discrete models are their lack of interpretability and biological diversity within cell populations [216]. Consequently, continuous models are often inappropriate to describe the full heterogeneous capacity of tumours. They also often require complex mathematical functions to represent biological systems which can complicate communication with non-mathematical expert. A study also demonstrates the limitations of continuous approximations which sometimes fail to reproduce the dynamics of a system due to its parameter value and model structure dependability [217]. To avoid drawbacks from both methods, several hybrid models have emerged. For example, a cellular automaton for cell interaction and migration combined with PDEs describing chemical and matrix dynamics was built to demonstrate the role of cell adhesion in solid tumour invasion [218]. Finally, mathematical models such as ODEs and PDEs describe quantitative variables and are opposed to computational methods such as networks or petri nets which can include qualitative relationships, that is values represent the state of an entity and not a quantity [219].

Qualitative modelling is particularly attractive when some biological details are missing or when the number of variables grows and requires qualitative relationships like thresholds. For this reason, discrete modelling is favoured in this thesis. Some biological details are unknown in the studied systems and therefore Boolean networks and rule-based models are appropriate modelling techniques to answer cancer progression questions. Moreover, by choosing these approaches, I show that computational modelling is accessible to most scientists and simple models can generate reliable predictions.

Once the scale and the variable characteristics have been defined, one should determine how the output simulations are generated. If the modeller desires to reproduce identical output simulations for a given initial condition and a set of parameter values then, the model should be built as deterministic [213]. Deterministic models in biological systems have been developed to study the dynamics of populations in which individual characteristics can be averaged over time, for example, tumour growth when behaviours of cells inside tumours can be averaged [220]. Their main advantages are their ability to represent large systems and the ease for scientists to analyse them. However, this simulation method ignores the environmental noise and the uncontrollable factors that affect cell proliferation and death. The alternative is to construct a stochastic model. Stochastic models include uncertainty and probabilistic events and as a result, the output vary for identical initial states and parameter values [221]. These models can mimic biological fluctuations in tumour progression, for example, a recent study has developed a stochastic model with immune and cancer cells with random birth and death to demonstrate the important role of random events in immunotherapy outcomes [222]. Stochasticity is particularly important in early stage cancer where the number of cells is low and therefore the impact of random events is strong [223]. A striking example is when a stochastic model is capable of showing clonal extinction in some simulations for a system for which a deterministic model could not [224]. However, including stochastic events in larger systems can be computationally expensive [225], as one gram of tumour contains in on average about 10^9 cells [226]. Often large tumours can be averaged using deterministic methods, especially as stochastic models are also often more difficult to analyse as a result of the randomness [227]. However, new stochastic simulation algorithms emerge and improve simulation times [228]. Information on parameter values also greatly impacts the choice between deterministic and stochastic simulations. As deterministic model outcomes are greatly affected by its parameter values, wrong values could false the results and therefore stochastic models should be preferred.

To simplify simulations of a large system, deterministic simulations are applied to the lymphoma model in Chapter 3. However, as mentioned in the previous paragraph, stochastic events have a greater impact during tumour initiation. Thus, in addition to the deterministic algorithms, stochastic simulations are performed to investigate the degree of variability of resistance emergence before the treatment is applied.

1.3.3 Parameter estimation as a barrier to cancer modelling

One of the central work of Chapter 3 is to find the appropriate parameter values for the lymphoma rule-based model. Parameter estimation of biological systems faces diverse challenges nowadays. Several methods can be applied to find parameter values such as literature searching or statistical inference. Literature can be a fast and convenient technique to obtain parameter values from already published experiments. For example, prostate cancer growth kinetics including cell loss and growth rate have been determined in a recent paper to evaluate the influence of tumour stages on these parameters [229]. Various computational models have demonstrated the reliability of such methods with good biological predictions [230–232]. However, despite the evident appeal for literature-based parameters, many parameters have not yet been evaluated experimentally, or even cannot be determined experimentally. Moreover, most parameter values are tissue-dependent. For example, cell division rates or the number of stem cells in a tissue are radically divergent between cancer types with different origins [233]. Therefore, papers used to find parameter values should be chosen cautiously.

Alternatively, Bayesian parameter inference is a well-used method as it reduces the risk of overfitting by giving the full parameter distribution [234]. The main feature but highly controversial matter of this technique is that parameters are not defined as constants but as probability distributions. However, by considering parameter as random variables, Bayesian techniques avoid fixing parameter values which can considerably affect the outcomes of a model [235]. The choice of priors has also been described as subjective as different individuals might specify different prior distributions for the same model which will impact on the posterior distributions [236]. Furthermore, Bayesian methods such as Approximate Bayesian Computation (ABC) simulate data with the proposed parameters instead of calculating the likelihood which is much faster but does require to know the likelihood function, however, ABC can be computationally intensive if many parameters are unknowns [237]. Despite some drawbacks, Bayesian methods still offer traceable computations and interpretable answers.

There exists another type of statistical inference opposed to Bayesian inference and which does not require a prior distribution named frequentist inference. Frequentist inference searches for constant values for unknown parameters using relative frequency of occurrence and confidence intervals. A common hypothesis test method is the use of the null hypothesis and its p-value, which is the probability to obtain the observed results or more extreme outcomes when the null hypothesis is true. The advantage of frequentist methods is the lack of prior knowledge, plus they tend to be less computationally intensive. However, their high reliance on how data are sampled due to the absence of priors and the misinterpretation of confidence intervals are major drawbacks [238]. The 95% confidence intervals contain the true mean of the estimated parameter in 95% of cases, however, it is incorrect to assume that confidence intervals represent a probability of 95% to cover the true mean. Generally, diverse options currently exist for parameter estimation, most if not all have limitations, however, it should be noted that imprecise parameter values still can address challenging biological problem and give reliable qualitative conclusions [239].

Chapter 3 combines literature and Bayesian inference to estimate parameter values of our lymphoma model. Bayesian inference is preferred over frequentist approaches as it requires no prior knowledge and can be easily implemented with the program used in this work (ProPPA, described in Methods). The mutation rate heavily studied in cancer research is estimated thanks to literature [240].

1.4 Thesis goals and plan

The main purpose of this work is to determine how the timing of different biological events shape blood cancer evolution and alter their clinical outcomes. Computational modelling of biological systems facilitate our comprehension of complex mechanisms by formalising and recreating the dynamics of entities composing these systems with fast and easily reproducible simulations. In this thesis, I use various computational models to study the abnormal development of hematopoiesis leading to blood malignancy. Blood cancers can be sampled non-invasively which allow a relatively easy access to tumour information, encouraging experimental and computational modelling of these diseases. The longitudinal sampling informs about clonal diversity and therefore makes blood disorders an ideal system to study cancer heterogeneity and evolution. To explore and answer these questions about blood cancer evolution, I first describe in chapter 2 the tools and methods used in my work. Chapter 3 first focuses on timing and cancer evolution at the cellular scale. For data availability reasons, the following chapters examine the molecular scales to explain the impact of timing at larger scales. These chapters are organised as follows:

In chapter 3, I demonstrate how multi-cellular models with simple cancer population growth dynamics can be applied to study complex dynamics in lymphoma. Using rule-based models and parameter inference, I focus on resistance emergence in a lymphoma mouse model treated with p53 restoration therapy to highlight the effect of treatment schedules on resistance. Further *in silico* treatment simulations help to improve existing treatment schedules and to propose new ones to increase survival.

In chapter 4, I explore the underlying mechanisms below branching evolution in two blood diseases, AML and MPN. I suggest *HOXA9* acts as a biological switch leading to AML patient stratification and MPN clinical variations when *JAK2* and *TET2* mutations have different orders. I implement a molecular network describing MPN progression with both mutations. The model recapitulates the disease symptoms, brings to light the genes responsible for the branching evolution and helps to clarify observed clinical characteristics with molecular explanation.

In chapter 5, I further study AML patient stratification by the identification of patient clusters with distinct *HOXA9* and *APP* expression. While *HOXA9* is a well-studied gene in AML, the poor characterisation of *APP* in hematopoiesis leads further analyses to investigate its function in leukemia clinical prognosis. I find that this gene, well-known for its involvement in Alzheimer disease, could play an important role in leukemia as a prognosis marker, but also for cell fate commitment.

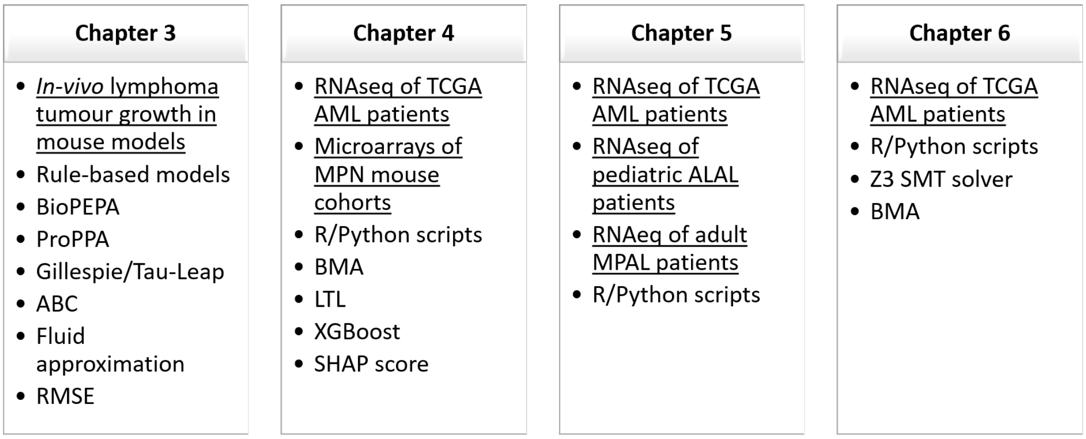
Finally, in chapter 6, I focus on the molecular dynamics underlying the *HOXA9*/*APP* patient cohorts found in Chapter 5. To determine how these patients can present distinct gene expression levels and which genes are responsible for this clustering, I build a motif inference algorithm which can generate from biological observations small molecular networks. This work allows users to find for complex molecular and cellular phenomena the responsible gene/protein motifs. This tool used on the AML data is able to select gene candidates associated with the identified relevant molecular motifs reproducing *HOXA9*/*APP* clusters. These findings demonstrate the algorithm ability to identify important disease markers that might have a role in AML patient classification and stratification. This program can be applied to various biological data and explain diverse cellular evolution mechanisms.

Chapter 2

Methods

2.1 Introduction

Methods used in this thesis are described in this chapter (summarised in Figure 2.1). I start by giving a brief description of the different datasets used in the chapters. Then, I introduce the tools applied to the rule-based models to describe lymphoma growth in mice. Specifically, I present two software BioPEPA and ProPPA, as well as their underlying algorithms to simulate and analyse models. I finish the section by comparing evaluation metrics which are employed to compare different lymphoma models. In the third section, I define qualitative networks and illustrate the software used in this thesis to analyse them, BioModelAnalyzer which permits their construction and stability analysis. In the following section, I introduce XGBoost a machine learning program used in this work to rank and identify *JAK2* correlated pathways in leukemia. Finally, I give a brief overview of what are satisfiability modulo theories (SMT) and how they work in model checking.



Chapter 3

- *In-vivo* lymphoma tumour growth in mouse models
- Rule-based models
- BioPEPA
- ProPPA
- Gillespie/Tau-Leap
- ABC
- Fluid approximation
- RMSE

Chapter 4

- RNAseq of TCGA AML patients
- Microarrays of MPN mouse cohorts
- R/Python scripts
- BMA
- LTL
- XGBoost
- SHAP score

Chapter 5

- RNAseq of TCGA AML patients
- RNAseq of pediatric ALAL patients
- RNAeq of adult MPAL patients
- R/Python scripts

Chapter 6

- RNAseq of TCGA AML patients
- R/Python scripts
- Z3 SMT solver
- BMA

Fig. 2.1 **Overview of the datasets and methods used in each chapter**. Except for Chapter 3, all analyses in this thesis are carried out using my own R and python scripts. Datasets are underlined. ABC: Approximate Bayesian Computation, RMSE: Root Mean Squared Error, AML: Acute Myeloid Leukemia, BMA: BioModelAnalzer, LTL: Linear Temporal logic, SHAP: SHapley Additive exPlanations, ALAL: Acute Leukemia of Ambiguous Lineage, MPAL: Mixed phenotype acute leukaemia, SMT: Satisfiability Modulo Theories.

2.2 Dataset description

This thesis uses a range of datasets to explore the theme of cancer progression in blood diseases. A brief description of each set is given in the following subsections.

2.2.1 *In-vivo* lymphoma tumour growth in mouse models

Almost 50% of human cancers have a mutant p53 protein which has lost its tumour protective effect as a result of its transcriptional activity depletion through multiple mechanisms [241]. Restoration of p53 functions in tumours has legitimately been proposed as an effective cancer therapy, however resistance development currently obstruct treatment efficiency and its process remains unclear. Understanding how this resistance arises would have considerable implications for existing p53 based treatments, such as Nutlin based MDM2 inhibition [242] and p53 gene therapy [243]. Martins et al [1] studied p53 restoration therapy with a reversibly switchable p53 knockin in $E\mu - myc$ lymphoma mouse model. These mice develop clonal B cell lineage lymphomas after being injected with a transgenic cell line which overexpresses Myc oncoprotein. Administration of a drug called tamoxifen makes the p53 protein functional again for about 30h. Whilst restoration of p53 is initially effective, rapid tumour regrowth with tamoxifen resistance is observed.

An interesting question arising from this study is: can we improve treatment strategies to avoid or delay this resistance emergence while increasing patient survival? Martins et al attempt to answer this question by developing additional experiments with *in vivo* quantification of tumour growth through fluorescent markers. To do so, $E\mu - myc$ lymphoma tumour cells were harvested from mice with the reversibly switchable p53 knockin and modified to include the fluorescent marker MSCV-Luciferase-Puro. Those cells were grown in cultures and re-injected in four new cohorts of mice (day 0). Each cohort consisted of four mice and followed a different treatment schedule: no tamoxifen injection (control), daily injections for 14 days (continuous), injection every 3 days (periodic3) and every 5 days (periodic 5). The first injection was done at day 10 and experiments concluded at day 35. Using imaging and luciferin techniques, the relative size of the tumours was assessed as photon counts which are translated into cell numbers in this thesis. Mice survival varies among groups, with lowest survival for the control group and highest for the continuous and periodic treatment schedules (Table 2.1). I study this lymphoma growth dataset in Chapter 3 to determine how to use cancer cell dynamics to delay resistance growth and improve treatment strategies to increase survival.

Treatment	Survival (days)
Control	22, 22, 26, 26
Continuous 14 days	32, 35, 35, alive at 35
Every 3 days for 13 days	32, 32, 32, 32
Every 5 days for 16 days	25, 26, 30, 31

Table 2.1 **Mouse survival for each p53 restoration treatment procedure in lymphoma mice cohorts.** Except for one mouse from the continuous treatment regime which survived, all mice succumbed to the disease regardless of the treatment procedure. As expected, survival is the lowest in the control group that did not receive any treatment. Mice in the continuous regime display the longest survival. However, the periodic schedule with tamoxifen injection every 3 days also leads to good survival outcomes.

2.2.2 RNAseq of AML patients

To study timing in human blood malignancies, I use in the next chapters (Chapter 4, 5 and 6) an AML public dataset. Data is available on The Cancer Genome Atlas (TCGA) project website [132]. Initiated in 2005, the TCGA project aims to regroup in a single platform patient samples and genomic studies of different cancer types. Data includes patient characteristics, their diagnosis and prognosis as well as their genetic mutational profile and gene expression profiling.

The data used in this work consists of 200 patients with *de novo* AML, among which 173 have RNA sequencing (RNAseq) information. RNA sequencing is a next-generation sequencing (NGS) technique which provides an overview of the transcriptome (mRNA, tRNA, sRNA) state of a sample at a given time point. This information is important to highlight which genes are turned on or off and what is their level of expression. The main advantage of RNAseq is that it does not require any prior sequence information (unlike microarrays) and it can detect structural variations such as gene fusions and alternative splicing events. Its drawbacks are its cost and more complex bioinformatic analyses compared to methods such as microarrays. However, RNA sequencing cost has been decreasing over the years with improving modern technologies and a large amount of packages and tutorials are now available to study RNAseq data [244].

AML RNAseq data obtained from TCGA is a gene expression matrix with the patients as the matrix columns and the genes as the rows. Each cell gives the raw counts of each gene of each patient. Data contains 20531 genes. As raw counts are not comparable between samples and genes due to experimental and gene length variations [245], it is necessary to normalise the gene raw counts to compare samples. I therefore choose to normalise raw counts into Transcripts Per Million (TPM). TPM is a well-used technique which was shown to be the best normalisation method to analyse RNAseq data thanks to its preservation of the biological signal [246]. All analyses are performed with the normalised data except for the DESeq2 differential analysis in Chapter 6. Using python programming, I filter unknown genes and genes with low expression (more than 50 patients have the gene expressed at less than 1 TPM). Outcome file contains 11832 genes for 173 patients.

Fundamental for the clinical analyses, TCGA provides a file containing the clinical features of each patients. Data contains: sex, race, age, FAB subtype (M0-M7), the percentage of blasts in the bone marrow and in the peripheral blood, the white blood cell count (WBC), the cytogenetics, the gene fusions and rearrangements, the cytogenetic and molecular classification, the cytogenetic and molecular risk, the single variants, the event free and overall survival as well as each patient mutation profile and some clinical details.

2.2.3 Microarrays of MPN mouse cohorts

To validate the predictions of the MPN computational model in Chapter 4, I work with a publicly available dataset from a recent study on MPN [150]. In this paper, authors compare the effect of *JAK2V617F* and *TET2* loss mutations in different cohorts of genetically engineered mice. Experiments start with a *JAK2* knockin mouse model in which *JAK2V617F* is expressed from its endogenous promoter and which can reproduce human MPN characteristics [149]. Using similar experiment settings, mice with four distinct genotypes are generated:

wild-type (no mutation), *JAK2* single mutant, *TET2* single mutant and a double mutant (both genes are mutated). From these mice, LSK cells are isolated to perform gene expression profiling. Detailed experimental protocols can be found in [150, 149].

To compare the different genotypes, authors use microarray data set, an high-throughput technology. Data has been deposited in the ArrayExpress repository at European Molecular Biology Laboratory–European Bioinformatics Institute and is accessible through the Array-Express accession number E-MTAB-2986 (http://www.ebi.ac.uk/arrayexpress/). Microarrays are relatively cheap and reliable tools to detect gene expression (as a signal) in a sample using fluorescence [247]. However, one of the main drawbacks of microarrays compared to RNAseq is the need of probes (fragments of DNA or RNA) which are necessary to detect gene expression. When using microarrays, a limited number of gene expression levels can be assessed. It is therefore crucial before using microarrays to select the appropriate probes for the genes of interest. Microarray experiments from [150] contain all the genes included in the MPN molecular network. Comparison of their expression in the different mouse cohorts can validate our model findings in Chapter 4 (section 4.2.3).

Data contains 4 mice with a wild-type (WT) genotype, 3 with a single *JAK2V617F* mutation, 2 with a *TET2* loss mutation, and 4 with a double mutant (*JAK2V617F*/*TET2*) genotype. Using a R script, genes that have a low detection signal (p-value $\leq$ 0.05) are removed as they represent poorly detected genes and might introduce errors. However, all gene probes are kept (one gene can have several probes for distinct isoforms). Finally, to exclude potential technical variations in the analyses, quantile normalisation, a popular and robust technique when appropriately performed is applied [248].

2.2.4 RNAseq of pediatric ALAL patients

In the prospect of establishing a link between a gene of interest *APP* and MPAL, two datasets from two different studies are analysed in Chapter 5. Both are publicly available RNA sequencing data from published papers on mixed phenotype/undifferentiated leukemias. The first study focuses on pediatric acute leukemia of ambiguous lineage (ALAL) patients [249] while the second described in the following subsection 2.2.5 examines adult MPAL [250].

In their paper, Alexander et al [249] gather the gene expression information of 115 pediatric ALAL patients: 35 are diagnosed with B-Myeloid MPAL, 49 with T-Myeloid MPAL, 16 with KTM2Ar MPAL and the 15 remaining have other subtypes of ALAL. ALAL consists of mixed phenotype acute leukemia (MPAL) and acute undifferentiated leukemia (AUL) [251]. The normalised gene expression matrix as well as the raw counts of each patient can be found on the National Cancer Institute website (https://target-data.nci.nih.gov/Public/ALL/mRNA-seq/Phase3/L3/expression/StJude/). The gene expression matrix has been normalised using the rlog transformation by DESeq2 [252]. To be able to compare those data with AML, I combine using python programming all patient raw counts into one matrix without gene filtering which has already been done by authors. Gene annotation is performed using the BiomaRt R package [253]. As identically implemented in AML data, raw counts are normalised using TPM. Outcome file contains 50635 genes for 89 patients. Clinical information of patients can be downloaded on the same website (https://target-data.nci.nih.gov/Public/ALL/clinical/Phase3/harmonized/). In the file, gender, survival, WBC, diagnosis and disease classification characteristics are available.

2.2.5 RNAseq of adult MPAL patients

The second dataset I use in Chapter 5 consists of adult MPAL patients [250]. The paper gives a summary of the clinical characteristics of 31 patients. Among those, RNA sequencing information of 24 patients diagnosed with B-Myeloid or T-Myeloid MPAL are publicly available in the Gene Expression Omnibus repository with the following accession numbers: GSE113601. Raw counts and specific patient clinical details are unavailable. Analyses in Chapter 5 are carried out with the accessible gene expression matrix which has been processed and normalised by DESeq2 [252]. This file contains 57773 genes. Absence of the original count information prevents the required transformation of these data into TPM for AML/MPAL comparison.

2.3 Computational tools to simulate and evaluate lymphoma rule-based models

2.3.1 Rule-based models

Rule-based models are used to describe the *in vivo* tumour growth of lymphoma mouse models in Chapter 3. This modelling technique simulate the behaviour and interactions between "species". The interest for such models is still emerging, but their application to diverse biological systems is currently growing thanks to their ease of use [254–256]. Species can be any biological components such as proteins and molecular complexes, as well as cells or living organisms. In rule-based modelling, rules refer to conditions for the reaction/interaction to happen and include the reaction rate, which means how often the reaction happens, and what the outcome of the reaction is. The main advantage of this method is that not all the different model states need to be specified, meaning that a single rule can be applied to several different reactants but all have to possess the required conditions for the reaction to happen. This characteristic is particularly attractive for large and complex systems which cannot be described by traditional methods such as ordinary differential equations due to combinatorial explosion [255].

An example of rule-based model for a population of cells C that divide at rate α and die at rate β can be written as follows:

$$\begin{aligned} C &\overset{\alpha}{\rightarrow} C+C \\ C &\overset{\beta}{\rightarrow} \emptyset \end{aligned} \tag{2.1}$$

2.3.2 BioPEPA for the time-series and distribution analyses

BioPEPA is a language for describing, modelling and analysing biochemical networks. BioPEPA is an extension of PEPA which is a process algebra defined for the performance analysis of computer systems. A comprehensive description of BioPEPA language can be found here [257]. One of the main feature of Bio-PEPA is the possibility to represent explicitly the characteristics of biochemical models, such as stoichiometry and the role of biological species in diverse reactions. The software enables the user to apply different mathematical and statistical analyses on the same biochemical model. However, I mostly use BioPEPA time series simulations for the lymphoma rule-based models. Simulations in this thesis are performed using the Tau-Leaping algorithm when stochastic simulations are desired, and the Implicit-Explicit Runge Kutta ODE solver for deterministic simulations. This Bio-PEPA functionality enables to study the evolution and proportion of the sensitive and resistant populations in the lymphoma model during treatment administration.

2.3.3 ProPPA for the parameter inference

ProPPA (Probabilistic Programming Process Algebra) is a process algebra that allows uncertainty in the model description [258, 259]. By using Bayesian inference, a machine learning method, and "observation datasets" compatible with a defined model, ProPPA can be used to infer the parameters of this model. This tool is particularly useful for biological models in which parameter values could not be found, are uncertain or when current techniques are not accurate enough to have a precise idea of some parameter values. This software is used to infer most parameters in the lymphoma computational model as additional biological experiments were not available to obtain parameter values experimentally.

To infer parameter values, the choice of the algorithm is crucial and depends on the model and the data (details on how to choose the algorithm can be found in Chapter 5 of Georgoulas [259]). Each algorithm uses either stochastic or deterministic approaches, but all start with a proposed sample of parameter values. The algorithm will accept or reject this sample according to input criteria. The next proposed sample is chosen through a function depending on the last accepted sample. ProPPA output is a text file with the set of accepted parameters.

Parameter inference are first carried out using the Approximate Bayesian Computation (ABC). This algorithm is chosen as it explicitly takes into account stochastic effects and does not use likelihood calculations which can be computationally expensive. A more detailed description of the method and pseudo code of the ProPPA ABC algorithm can be found in the subsection 2.3.5. However, as a result of the Gillespie algorithm and the large number of cells, simulation times of the lymphoma rule-based models with ABC are long (see the stochastic simulation algorithm subsection 2.3.4 for Gillespie algorithm). Unfortunately, switching Gillespie to Tau-Leap algorithm did not improve much simulation timing, and as the ABC method is highly dependent on configuration values, I opt for a deterministic approach called the fluid approximation algorithm which is faster to run (details and pseudo code in the subsection 2.3.6). As the models contain a large number of cells, continuous simulations can be considered as a good average of stochastic simulations. However, with this approximation, the variability property of tumour growth and random resistance emergence are ignored.

2.3.4 Stochastic simulation algorithms

For the ABC parameter inference, ProPPA uses the Gillespie algorithm to simulate the model with the proposed parameters. Gillespie generates exact stochastic trajectories for a system and has originally been designed to stochastically simulate chemical reactions [260]. However, Gillespie is now applied to many other biological fields such as epidemiology [261] or ecology and evolutionary dynamics [262]. Pseudo-code of the algorithm is found below. The first step of this algorithm consists in evaluating all the possible reactions of the system among which each reaction has a corresponding rate, for instance division, differentiation or death rates in the case of cellular models. The probability for this reaction to happen as a next event is proportional to its rate and the number of input variables affected by this reaction. Then, two numbers are randomly generated: one will determine the next reaction and the other the time interval at which it will happen. Next, the new time and number of variables are updated accordingly to the picked reaction and interval. A new iteration can then start and randomly pick new numbers until the system reaches a time threshold initially defined by the user. The outcome is a trajectory describing the evolution of all the variables of the system. Generation of a large number of trajectories gives all possible solutions for the final states of the model. Gillespie is therefore a useful and accurate tool to simulate biological models, however, reactions with a lot of input variables or happening at a fast rate considerably slow the simulations by generating very small time intervals and the algorithm can therefore become ineffective.

```
Result: Gives the model trajectories of each variable $x_i$
Initialisation of the variables $x_i(0)$ and time $t = 0$;
while $t < t_{max}$ do
    Generate two uniform random numbers $u_1$ and $u_2$;
    Determine the next time step $dt$ using $u_1$;
    Compute each reaction probability;
    Determine the next reaction using reaction probabilities and $u_2$;
    Calculate the new system state $x_i(t + dt)$;
    Set $t = t + dt$;
end
```

Algorithm 1: Gillespie pseudo code
with $i = 0, 1, .., N$, x_i the N variables of the model and t_{max} the time threshold.

To overcome slow simulations in larger systems, Gillespie proposed an approximate algorithm called Tau-Leaping in which the time interval is constant and therefore the system is less often updated [263]. For a defined time step called τ, Tau-Leaping generates for each reaction a random number from a Poisson distribution giving the number of times each event happens. Similarly to Gillespie, the number of reactions per time step is proportional to its probability to happen. By doing so, Tau-Leaping is less accurate than Gillespie, however, this approximation applied to systems that does not change much in a time step can considerably improve simulation time with a relatively low impact on the accuracy of the trajectories [264]. The time step should therefore be small enough that the system does not substantially change during this interval.

2.3.5 Approximate Bayesian Computation (ABC) method and pseudo-code

Bayesian analysis gathers methods to find unknown parameters and requires three main inputs: some data, a model and priors [265]. Priors are the information one already has on the model and the parameters and constitute one of the main assets of using such methods. Inputs combined with the application of the Bayes theorem generate an output solution for the unknown parameters called the posterior probability distribution. The Bayes theorem is defined as follows:

$$P(\theta|D) = \frac{P(\theta) \times P(D|\theta)}{P(D)}$$

with D the data, θ the unknown parameters, $P(\theta|D)$ the posterior distribution, $P(\theta)$ the prior distribution for the unknown parameters, $P(D|\theta)$ the likelihood of observing the data with a sample of parameters and finally $P(D)$ the model evidence which can be interpreted as the probability that randomly selected parameters from the prior would generate D.

The Approximate Bayesian Computation (ABC) method uses model simulations to approximate the likelihood of the proposed parameter values and a rejection algorithm for the parameter selection [266]. Parameter values are accepted or rejected by comparing the model simulations to the data. If the distance between both is too large, values are rejected. The set of accepted samples constitute the desired posterior distribution. This Bayesian technique can however be computationally costly [267].

ProPPA uses a Markov Chain Monte Carlo algorithm for sampling parameter values. Unlike the standard ABC which picks up randomly parameter values from the prior distributions, Markov Chain Monte Carlo algorithms create samples of parameters by first picking up values from the prior distributions while the next samples are generated using the previous simulation [268]. The appearance of combinations of parameters is proportional to the probability of these parameter values in the prior distribution. The pseudo code of the

ABC algorithm is the following:

```
Result: Gives the accepted samples for the inferred parameters
Initialisation of the inferred parameters $x(0)$;
for $i$ = 0 to $i_{max}$ do
    Sample the parameters $x^*$ from $q(x^*|x(i))$;
    Run model with $x^*$ using Gillespie algorithm ;
    Compute Euclidean distance $d_{x^*}$ between the simulated trace and observation data ;
    if $d_{x^*} < eps$ then
        Sample $k$ using an uniform distribution of interval $(0,1)$ ;
        if $k < min(1, \frac{p(x^*)}{p(x(i))})$ then
            $x(i+1) = x^*$;
        else
            $x(i+1) = x(i)$;
        end
    else
        $x(i+1) = x(i)$;
    end
end
```

Algorithm 2: ProPPA ABC algorithm for inference parameters

with p the prior distribution of the inferred parameters and q is the distribution that suggests the next candidate for the parameter sample with a mean $x(i)$ (the previous accepted sample) and a variance named '*proposal*'. *eps* is the accepted tolerance threshold and i_{max} the maximum number of samples. i_{max}, *eps* and *proposal* are configuration constants that the user can define in an input text file, which allows a large flexibility for the outcomes of the algorithm. It should be noted that if the prior p is a uniform distribution, then $\frac{p(x^*)}{p(x(i))}$ always equals one and thus if $d_{x^*} < eps$, the sample x^* will always be accepted.

2.3.6 Fluid approximation algorithm

As the number of cells is large enough to consider continuous approximation for the simulation of the lymphoma models, the fluid approximation sampler is chosen over ABC for parameter inference analyses. Also known as mean field approximation, the fluid approximation algorithm estimates the stochastic dynamics of a system by deterministic ones by generating for each variable an ordinary differential equation (ODE) [269]. ODEs give the average number of variables at each time point. Once generated, the ODE solution is compared to the data. The likelihood L of the proposed parameter values is calculated using the distance between the generated simulation and the data assuming a Gaussian noise. A Gaussian noise is a random variable with a normal distribution and is a useful addition to deterministic approaches that do not take into account noise in the data experiments. Adding a noise to the likelihood is therefore an accurate way to include noisy measurements into the algorithm. The proposed parameter sample is accepted or rejected by comparing a randomly picked number between 0 and 1 and a ratio r which is a function of the prior distribution p and the likelihood L between the previous and the proposed sample. The pseudo code of the fluid approximation defined in ProPPA is as follows:

```
Result: Gives the accepted samples for the inferred parameters
Initialisation of parameters to infer $x(0)$;
for $i$ = 0 to $i_{max}$ do
    Sample the parameters $x^*$ from $q(x^*|x(i))$;
    Solve the model with parameters $x^*$ using an ODE solver ;
    Compute the difference $d_{x^*}$ with observation data ;
    Compute the likelihood $L$ of the parameter values $x^*$ which is a function of $d_{x^*}$ and
      some noise ;
    Compute the ratio: $r = \frac{p(x^*)}{p(x(i))} \frac{L(x^*)}{L(x(i))}$ ;
    Sample $k$ a random number between 0 and 1;
    if $k < min(1, r)$ then
        $x(i+1) = x^*$;
    else
        $x(i+1) = x(i)$;
    end
end
```

Algorithm 3: Fluid approximation algorithm in ProPPA

with p the prior distribution of the inferred parameters and q is the distribution that suggests the next candidate for the parameter sample with a mean $x(i)$ (the previous accepted sample) and a variance named '*proposal*'. Similarly to ABC, the Gaussian noise, i_{max}, and *proposal* are constant configurations that the user can define in a text file.

2.3.7 Scoring system for ProPPA inference results

Once parameter values have been evaluated, work in Chapter 3 aims to find the simplest model which can best fit and explain the lymphoma mouse models. A scoring system is designed to compare the *in vivo* experiments against the BioPEPA simulations with the

ProPPA inferred parameters. Scores are computed with the root mean square error (RMSE) which is defined as follows:

$$RMSE = \sqrt{\frac{\sum_i (Y_i - Y_i')^2}{N}}$$

with for this work, Y the $\log_{10}$ of mouse experiment values, Y' the $\log_{10}$ of BioPEPA simulation values and N the number of observations in one experiment. The lowest the score the closest the simulation is from the experimental data.

Many evaluation metrics are available to compare observations against simulations, among which the most popular are the mean absolute error (MAE), the mean squared error (MSE), the root mean square error (RMSE) and the root mean square log error (RMSLE) [270]. However, MAE and RMSE are the most commonly used metrics and therefore are subject to many debates and comparison [271]. While RMSE computes the square root of the average of squared differences between simulations and actual data, MAE is the average of the mean difference between both, that is:

$$MAE = \frac{1}{N}\sum_i |Y_i - Y_i'|$$

Both RMSE and MAE lower values predict a greater similarity between observations and simulations and both neglect the direction of the errors ($Y_i - Y_i'$). However, by squaring the errors instead of taking their absolute value, RMSE penalises samples with large errors and outliers. Despite its sensibility to outliers which can be removed before calculating the RMSE and some criticisms about its ambiguity, the RMSE by giving high weights to large errors is better for comparing model performance [272]. I therefore decide to use this scoring metric to compare different rule-based models against the *in vivo* mouse experiments.

2.4 Construction and analyses of the MPN network

2.4.1 Qualitative Networks

To understand the regulatory dynamics of MPN patients with *JAK2* and *TET2* mutations, I build a qualitative network including important hematopoietic regulators. Qualitative networks are an extension of Boolean networks with two major modifications [273]. First, discrete instead of Boolean variables are assigned to the nodes representing molecular expression. This means higher resolution in expression levels. The second change is the addition of more complex interactions between nodes along with the usual activation and inhibition.

First, the formal definition of a Boolean network can be defined as follows: a Boolean network $B(C,F)$ can be seen as a variant of a graph $G(V,E)$ with V the set of nodes and E the set of edges. In Boolean networks, C is similar to V and includes all the components of the Boolean networks, while F is a list of Boolean functions f_i assigned to each component $c_i \in C$. A Boolean function f_i computes the next state of a component at time $t+1$ using the current state s of the network at time t:

$$c_i(t+1) = f_i(s(t))$$

As all the components have Boolean values, then $f_i : \{0,1\}^k \to \{0,1\}$ with k the number of components in the network. Boolean network interactions includes activation and inhibition. The connection between two components c_i and c_j in these networks is attached to a weight α_{ij} which can be positive or negative for an activation or inhibition respectively. Therefore, the Boolean function f_i of a component c_i can be defined as follows:

$$f_i(s) = \begin{cases} 0 & \text{if } \sum\limits_{c_j \in C} \alpha_{ji} c_j < 0 \\ 1 & \text{if } \sum\limits_{c_j \in C} \alpha_{ji} c_j > 0 \\ c_i & \text{if } \sum\limits_{c_j \in C} \alpha_{ji} c_j = 0 \end{cases} \tag{2.2}$$

Here the term $\sum\limits_{c_j \in C} \alpha_{ji} c_j$ gauges the overall effect of all components of the network on c_i. If this effect is positive, the next value for c_i is 1, and if it is negative, the next value is 0. Finally, if the effect is neutral, the component keeps its current value.

Components of a qualitative network are discrete variables, and therefore can express a higher range of expression. A qualitative network $Q(C,T,N)$ consists of components in C which can take values in $\{0,1,..,N\}$. N is a constant integer with any possible values above one, and is called the node granularity. If N is one, the model is a Boolean network. T is the set of target functions, which are the functions that determine the level toward which each component moves at the following time step. At each time step, each component moves by a maximum of one level. The update of each component $c_i \in C$ can therefore be mathematically defined as:

$$c_i(t+1) = \begin{cases} c_i(t) + 1 & \text{if } target_i(s(t)) > c_i(t) \\ c_i(t) - 1 & \text{if } target_i(s(t)) < c_i(t) \\ c_i(t) & \text{if } target_i(s(t)) = c_i(t) \end{cases} \tag{2.3}$$

with $s(t)$ the current state of the network and $target_i \in T$ the target function of c_i. $target_i$ returns a value in $\{0,1,..,N\}$. The default target function is the difference between the amount of activation act_i and the amount of inhibition inh_i a component c_i is subject to. Both are averaged by the total number of activating/inhibiting components and are defined as follows:

$$\begin{aligned} act_i &= \frac{\sum\limits_{\alpha_{ji}>0} \alpha_{ji} c_j}{\sum\limits_{\alpha_{ji}>0} \alpha_{ji}} \\ inh_i &= \frac{\sum\limits_{\alpha_{ji}<0} \alpha_{ji} c_j}{\sum\limits_{\alpha_{ji}<0} \alpha_{ji}} \end{aligned}$$

The definition of the default target function for $c_i \in C$ is:

$$target_i = \begin{cases} max(0, act_i(s) - inh_i(s)) & \text{if } max(\alpha_{ji}) > 0 \\ N - inh_i(s) & \text{if } max(\alpha_{ji}) \leq 0 \end{cases} \tag{2.4}$$

where $max(\alpha_{ji}) \leq 0$ represents the specific scenario when only inhibition interactions are applied to c_i. In this case, the component returns to its maximum value N in absence of any activity in the inhibitors. However, target functions can also be customised by the users to represent more complicated biological connections. They include in their formula mathematical operators such as the addition and multiplication, but also additional functions such as the floor and ceiling functions which respectively take the value of the closest largest or smallest integer of the input variable.

2.4.2 BioModelAnalyzer

The BioModelAnalyzer platform (BMA) is a graphical tool for the construction and analysis of biological systems [191]. Any users can generate simple to complex biological models which are internally translated into qualitative networks that can be automatically analysed. The appeal of such method is the absence of required computer science knowledge background, while advanced users can still make the most of qualitative network features such as Linear Temporal Logic (LTL) model checking (see LTL section 2.4.3 for more details). I use BMA to construct and analyse attractors of the MPN network in Chapter 4, but also to verify that the motifs found by the inference algorithm in Chapter 6 can reproduce the input observations.

BMA is a visual tool where models are built by dragging and dropping the different elements into a gridded canvas (Fig. 2.2). The graphical items of BMA are the cells, the proteins and the edges. Cells make the model visually comprehensible and allow users to copy and paste motifs, but they have no role in the underlying network and analysis. Proteins are respectively illustrated by three icons for the receptors, extra and intra-cellular proteins. Interactions between proteins have an arrow edge shape for the activations or a bar-arrow edge shape for the inhibitions. The granularity of the components as well as their target function are also directly accessible on the interface. The default granularity is one which generates a Boolean network. Default target functions are as in equation (2.4), that is the weighted average of activations minus the weighed average of inhibitions that act on a component. When no activation is present, the variable value stays null. However, these variable features can be customised by the user by right-clicking on the desired element. A small window opens with the possibility to add a name, description and target function to the component as well as change its granularity. Customised target functions includes constants, other components and simple mathematical operators such as addition (+), substraction (-), multiplication (*) and division (/). The user can make more complex functions using the minimum, maximum and average operations. Finally, BMA also includes the ceiling and floor functions. The target functions simply return, respectively, the closest greatest or lowest integer when their input is a real number.

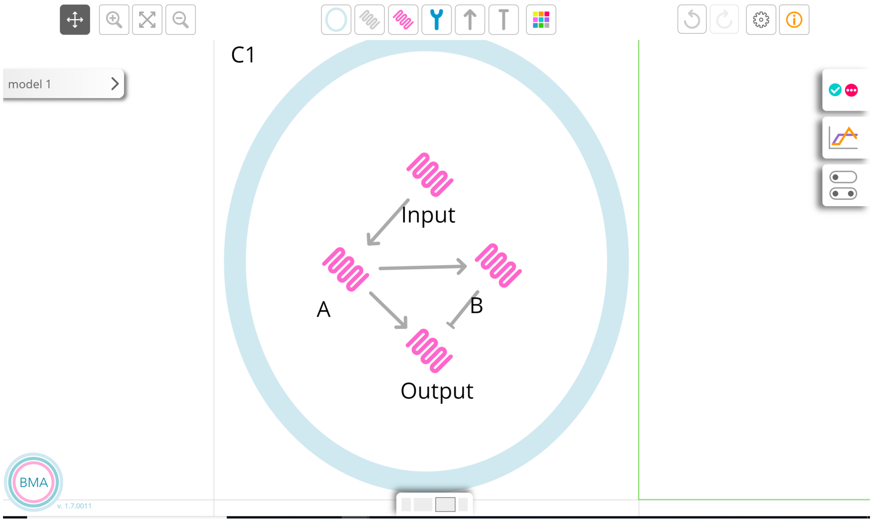


Fig. 2.2 **BMA interface**. All model elements are grouped at the middle top of the screen and are, in the right order: the cell, the extra-cellular protein, the intra-cellular protein, the receptor, the activation edge, the inhibition edge and finally the last icon allows the user to change colours of elements. On the side at the top right, the three icons are for the model analyses. They are in order from top to bottom: the stability analysis (described in section 2.4.4), the model simulation and the LTL model checking tool (described in section 2.4.3).

Once the model is built and translated into a qualitative network, the next step is the stabilisation analysis (see [274] and Stabilisation analysis section 2.4.4 for details). BMA models are discrete and deterministic, therefore the number of states a model can reach is finite. As part of the possible network analyses, BMA tries to prove the stability of the model for different initial conditions. We say a model is stable if, for any initial condition, all simulations reach a single fixed point attractor. An attractor is a numerical valued solution the system progresses to. A fixed point is reached if and only if $s = T(s)$ with s a state of the model, here the fixed point, and T the set of target functions. If the model does not stabilise, another possible attractor is a cycle of states. In this case, we say the states are recurring and are defined by $s = T_n(s)$. Here, n defines the size of the cycle, in other words, the number of recurring states. Finally, the stabilisation analysis can also find bifurcations. BMA performs the stabilisation analysis for all initial conditions, and different attractors can be reached for different initial states.

2.4.3 Linear Temporal Logic (LTL) in BMA

Temporal logics are concise languages in the form of mathematical logics, more specifically modal logics, where one appends temporal modalities to a proposition to add information about its evolution in time. This means a proposition can be false at a certain moment and become true later on. Temporal logics include the classic logical operators as well as temporal operators. We refer logical operators as "OR", "AND", "NOT" and "IMPLIES" which are all available in BMA Linear Temporal Logic (LTL) analyses. I use the LTL tool in BMA to analyse the stability of the network and verify that the specifications for all *JAK2* and *TET2* mutants match the literature.

LTL is one of the most popular temporal logic language and it was introduced by Pnueli [275]. Its syntax includes two temporal operators written **X** for next and **U** for until, respectively called "NEXT" and "UPTO" in BMA. **X** implies that a proposition should hold in the next step while **U** verifies that a proposition is true up to another becomes true. Derived from the "UPTO" operator, **F** eventually and **G** always respectively specify that a proposition should be eventually true and that a proposition should forever hold ("EVENTUALLY" and "ALWAYS" in BMA). Using LTL in biological modelling studies is essential to untangle patterns in the biological events [276, 277]. LTL can be applied to search for different temporal properties in models such reachability, liveness, invariance, stability and oscillation [278].

As LTL queries are model dependent, it has to be manually performed. However, LTL syntax can be challenging for users with no computer science and formal verification background. To facilitate LTL queries, BMA includes in its functionalities a graphical interface for LTL analyses using visual icons representing different LTL operators and formulas. Users can drag and drop different elements and form formulas for LTL analyses.

LTL queries can be performed following two simple steps: first by defining the network states of interest and second by adding the temporal and logic operators to these states.

State definition can be carried out by adding value constraints on the different components of the network. This is achieved through a drop down menu as illustrated in Figure 2.3A. The second step consists in including the newly created states into the logic and temporal operators to create new queries. Each operator includes the right number of sockets for the addition of operands that can be states or formulas made of states and operators. Some default states have also been included to simplify query production. Self-loop and oscillations can be added to the formulas. These added features enable users to quickly check the stability of its model, for example, by simply creating a query with the icons `Self-loop` (state), `Always` (operator) and `Eventually` (operator) as shown in Figure 2.3B.

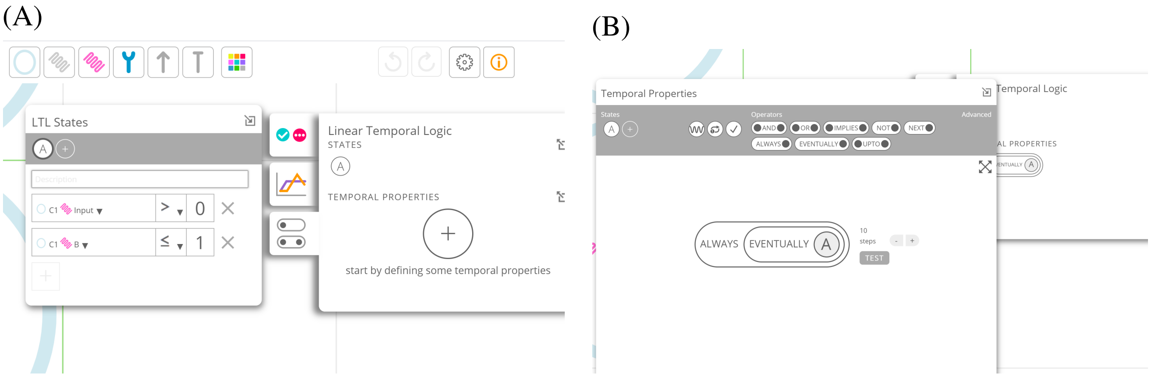


Fig. 2.3 **LTL interface in BMA.** The two steps to create LTL queries is shown in this figure with (A) the generation of the LTL state A and (B) the addition of the logic and temporal operators "ALWAYS" and "EVENTUALLY" to manually perform a stability analysis.

After a query is submitted, three possible outcomes can be obtained: the query is true for all traces, for some or for none. Respectively, the colour of the query indicates the outcome and will be either blue, stripes of blue and pink or pink. BMA includes in its LTL analysis examples and/or counterexamples of simulations demonstrating the outcome of the query.

2.4.4 Stabilisation Analysis in BMA

Stabilisation analysis in BMA is based on the algorithms described in [274]. As mentioned in the previous section 2.4.2, BMA models have three possible outcomes. The system is said stable if all simulations reach a unique attractor that is a fixed point. A bifurcation is observed if the simulations attain a fixed point but there exist at least two different fixed points for different initial conditions. If none of the above occurs, we can deduce that the model reaches a cycle of length greater than one.

To prove stabilisation in BMA, the algorithm generates small lemmas on the components of the network that are solvable through their target function. For example, if a component c_i

is induced by another component c_j with a target function $target_i = 1 + c_j$, one can deduce that c_i is always greater than 1 and lower than N, with N the node granularity. This can be defined using LTL language such that $FG(0 \leq c_j \leq N) \Rightarrow FG(1 \leq c_i \leq N)$. In LTL language, "$F$" means *eventually* and "G" *always* [275]. In other words, "FG" determines how components behave in the long term in simulations. Iterative generation of lemmas on the all components can be therefore described in the general form:

$$FG(p_1) \wedge FG(p_2) \wedge ...FG(p_m) \Rightarrow FG(q)$$

where p_is are the lemmas of the m input variables associated to a component c_i and q is the resulting lemma defining c_i. All lemmas are of the form $m \leq c \leq M$ with c a component of the model, and m and M two constants in $\{0,..,N\}$. All lemmas p_i are initially generated using target functions defined in the BMA models to find m and M. Additional lemmas can be found using lemmas found with the target functions. The algorithm computes lemmas until no new ones can be found. Stabilisation is proved if for all the component c_i its lemma entails $FG(c_i = k_i)$ with k_i a constant.

If stabilisation cannot be proved using lemma generation, the algorithm searches for counterexamples until exhaustion. If no counterexamples can be found, stabilisation is then proved. To find multiple fixed points (bifurcation) and prove unstability, the algorithm encodes a problem of existence using a formula satisfiability problem. Let $(x_1,...,x_N)$ and $(y_1,...,y_N)$ be two different states of the network such that $\exists i \in \{0,..,N\} \Rightarrow x_i \neq y_i$ and $\forall i \in \{0,..,N\}$, $(x_i(t+1) = x_i, y_i(t+1) = y_i)$ where $x_i(t+1)$ and $y_i(t+1)$ are determined via Eq. (2.3). If the problem is unsatisfiable, the algorithm searches for cycles using model checking [279] and the encoding of liveness to safety [280]. Briefly, the algorithm first determines the length of the longest path of the network to limit the cycle search. Then, for all paths of the network, it iteratively verifies if the i-th state of the state sequence defining a path is different from state 0 (Fig. 2.4). State 0 is formerly certified as not in a self loop or equal to state 1, in other words, the algorithm check that state 0 is not a fixed point. If the i-th state of the sequence is identical of state 0, then this means a cycle of length i is found. Procedures for finding counterexamples is a fast process thanks to the previous lemma generation that bounded the components and therefore limits the search. Further details on the algorithm can be found in [274].

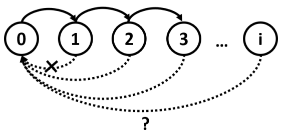


Fig. 2.4 **Cycle search in BMA.** Cycle search steps are: 1) The longest path of the network is defined as the algorithm simulation threshold. 2) Check and confirm that state 0 is not a fixed point. 3) Are state i ($i \geq 2$) and 0 equal? If so, cycle has a length i, else, we look at state $i+1$.

2.5 XGBoost to rank and identify *JAK2* correlated pathways in AML patients

XGBoost (e**X**treme **G**radient **Boost**ing) is a machine learning algorithm using gradient tree boosting to learn from large structured data [281]. In this section, I first define supervised machine learning and tree ensembles which are two important concepts to understand Gradient Boosting algorithms. I then give an overview of the essential steps of the XGBoost algorithm and describe how I customise it for the pathway correlation analyses of Chapter 4. I finish the section by explaining how SHAP scores can help to identify genes correlated with *JAK2*.

2.5.1 Tree Ensembles

Supervised learning is a set of Machine Learning techniques that uses information from some training data to create a learned function which can be applied to new data in order to make new predictions. For example, supervised machine learning algorithms can be used to classify pictures of lions and zebras. A first step consists in feeding the algorithm with different pictures of lions and zebras (the training set) paired with their correct output value "lion" or "zebra". From these pictures, the algorithm learns how to split the two types of pictures in a learned function, for example, by using the main colour of the picture as a classifier. If the picture mainly contains the colour yellow, the algorithm can predict that it is most likely a picture of lion, but if it is mainly black and white, it can assume it should be a zebra. The final stage is to feed the algorithm with new pictures without annotations (the testing set) and score the new predictions the algorithm made. The more lions and zebras correctly detected in the new pictures and the higher the score will be, scaling the performance of the algorithm and its learned function. Supervised learning is opposed to unsupervised learning methods which look for patterns in data, and do not require human supervision. A well-known example of unsupervised learning is clustering analysis which aims at grouping data into subgroups based on the presence or absence of common characteristics.

Also called classification and regression trees (CART), decision trees are supervised machine learning algorithms mainly used for classification and regression analyses. Classification tree analysis is used when the prediction is a discrete variable, as for the lion/zebra example. When the predictions are continuous values, decision trees are called regression trees. The goal of both algorithms is to predict the value of a target variable based on several observations (or input variables). Decision trees can be seen as tree-shaped flowchart, where each internal node represents a test, each branch is the outcome of the test and each leaf (terminal node) is the prediction (Fig. 2.5). In the previous example, a test could be "is the main colour in this picture yellow?" and the two output branches would be "yes" and "no". From the "no" branch, a new node/test could be added "are the main colours white and black?". Leaves for this same example would be the values "lion" and "zebra" linked to the correct branches.

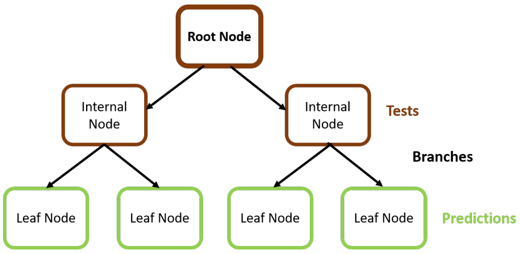


Fig. 2.5 Basic structure of a decision tree.

Accuracy in these models is dependent on the approach used by trees to decide how to split nodes. The decision criteria will be different for regression and classification trees. The tree first split the nodes using all available input variables and then selects the split that generates the most homogeneous subnodes (data are from the same class inside nodes). To obtain homogeneity, different algorithms are available. Popular algorithms for splitting nodes include the Gini Impurity and the Information Gain (IG) for quantitative variables [282] and the variance reduction for continuous variables [283].

CART have many advantages, starting by the simplicity to visualise, understand and interpret them. Plus as mentioned before, it can handle both continuous and quantitative variables and does not require any normalisation while many other techniques usually do. The main disadvantages, however, are issues with overfitting and robustness. Decision trees can create complex trees that will not represent well general data and depend on the training data. Consequently, a small change in the training data could impact the tree and its predictions. A recognised way to cope with these issues is the use of boosting algorithms that sequentially build a tree ensemble and where each new model learns from the errors of the previous one.

2.5.2 XGBoost

XGBoost success amongst the research community can be mostly explained by its execution speed. The scalability of this algorithm lies on its parallel and distributed computing which offers efficient memory usage. In addition to its interesting speed feature, XGBoost model performance has been demonstrated by many developers in Kaggle competitions, a well-known worldwide competition for the data science community. Chen et al [281] reported that among the 29 challenge winning solutions in 2015, 17 used XGBoost.

XGBoost is a gradient boosting algorithm. Boosting algorithm technique resides in combining weak learners (here a decision tree) while weighting the outcomes of each classifier. The algorithm sequentially builds trees in order to build a new stronger model by untangling the misclassification error of the previous tree and try to reduce it. In my work, the target variable is *JAK2* and is the binary variable I want to predict values from. It takes only two values: "low" and "high" (translated to 0 and 1 in the code) which represent both groups of patients with low and high expression level for *JAK2*.

The first step of XGBoost algorithm consists in calculating the initial prediction for every observation. The default initial prediction for regression or classification is 0.5. For this work, this means that it initially assumes that each new individual has 50% chance of being in the high or low cohort for the gene *JAK2*. In Gradient Boosting, trees are built to predict the residuals of the data, which are found by calculating the difference between the last predicted values minus the observed values. XGBoost starts by computing the first residuals with the initial prediction and builds a first tree. The first leaf of the tree always contains all the residuals. The next step consists in computing a quality score, also called similarity score. Similarity scores allow XGBoost to compare different nodes and their homogeneity. If residuals with similar values are in the same node, the residual sum will be high while residuals with very different values (positive and negative) will cancel each other out and the score will be low.

The next step consists in determining the best way to split the residuals into two groups for a certain input variable. To do so, several threshold values from the desired input variable are defined by calculating the means between two consecutive measurements. For example, if the input variable is $[0,2,5,7]$, the threshold values are 1, 3.5 and 6. These thresholds separate the residuals into two nodes. Similarity scores of both nodes are estimated and for each threshold the algorithm computes another score called Gain. XGBoost compares the gains between the thresholds and chooses the split with the highest gain (as this is the split that has best separated the residuals homogeneously).

Then, XGBoost repeats the same workflow for each node when there are more than one residual. The default number of levels per tree is 6, but this parameter can be customised as most parameters. Once splitting is achieved, XGBoost avoids model overfitting by deciding if nodes must be pruned (removed) based on its gain and a parameter γ. The final step for building an XGBoost tree consists in calculating the output value that the algorithm use to make the predictions. Once these new predictions are found, new residual values can be computed and a new tree is built starting by the root node with all the updated residual values. The algorithm stops building new trees if the number of trees has reached its maximum or residuals are small enough.

2.5.3 Ranking pathway association for gene expression level using XGBoost

Work in Chapter 4 applies XGBoost algorithm to rank gene pathways well-described in cancer studies to identify which pathways and genes have the highest correlation with *JAK2* in AML patients. Input of the algorithm is the TCGA AML data which is a gene expression matrix, where columns represent patients and rows the genes. Patients are split into two groups: one with the highest expression for *JAK2* and another with the lowest expression. The goal is to determine which pathways are the best to classify patients between these two cohorts. This approach is called binary classification as the target variable *JAK2* is quantitative and takes only two values "low" and "high". The input variables for the classification are the rows of the matrix, the genes.

Analyses are programmed in python using the XGBoost package. The Matthew Correlation Coefficient (MCC) is used to score the competence of each pathway to classify the patients into the right cohort. The MCC is currently considered as the most reliable measure to rate binary classification, as it takes into account the four classes of the confusion matrix (True Positive, True Negative, False Positive and False Negative) while accuracy and F1 scores only uses True Positive and True Negative:

$$MCC = \frac{TP \times TN - FP \times FN}{TP + TN + FP + FN}$$

True Positive (TP) and True Negative (TN) are counts representing how many times a pathway have correctly classified a patient into the high or low cohort respectively. False Positive (FP) represents the number of times a pathway have classified a patient with low expression into the high cohort, and vice versa for the False Negative (FN). MCC score varies between -1 and 1, but is translated into percentage in the figures.

Evaluation metrics are scores to quantify the quality of a statistical or Machine learning model. The default evaluation metric in XGBoost is the error rate for binary classification, which is calculated as the ratio of the number of wrongly classified observations over the total number of observations. This default function is changed for the logarithmic loss (logloss) function. The Logloss is thought to be the best evaluation metric for binary classification as it heavily penalises classifiers (trees) that are confident about a wrong classification [284].

In this work, the `colsample_bytree` parameter is set to 0.3 instead of 1 (default value). A pathway can be identified by XGBoost as a good classifier because one of its gene is highly correlated to *JAK2* although the other genes in this pathway are not as good for clustering patients. The column sampling parameter is a solution to reduce the accuracy of such models. By setting its value to 0.3, XGBoost builds trees with randomly picked 30% genes per pathway. Therefore, it excludes in some iterations the genes that could be highly correlated to the gene of interest, *JAK2*, and therefore reduces the classifier score if the other genes of the pathway are not good.

2.5.4 SHAP score

SHAP is used to explain the output of the XGBoost models for *JAK2*. SHAP (**SH**apley **A**dditive ex**P**lanations) aims to ease the interpretability of complex models by representing the importance of model features with shapley values [285]. Complete documentation on SHAP and how to use it can be found on github (http://github.com/slundberg/shap).

The Shapley value is a solution concept of cooperative game theory. Cooperative game theory uses games in which players are forming coalitions (group of players) due to the possibility of external enforcement of cooperative behaviour. Analysis of such games includes predicting how players will form coalitions. Each coalition obtains an overall gain from the cooperation, but each player does not contribute equally to this gain. Shapley values allows one to determine the importance of each player in the overall cooperation and its expected pay off.

In machine learning, the coalition can be interpreted as a subset of input variables of the model, also called features, and the gain of the coalition as the predicted value of the model for those input variables. Some feature values have a large impact on the prediction, while others have a small impact. The Shapley value is the average marginal contribution of a feature value across all possible coalitions. In other words, in a first step, the algorithm takes different coalitions, or subsets of input variables, and computes the predicted value for each subset. Then, the algorithm takes these same subsets and adds the feature we wish to evaluate. Then, it computes the difference for each subset between the two predicted values to determine the contribution of the feature to this subset. The average of all contributions gives the Shapley value for this feature. SHAP scores are part of XGBoost package and enable in my work to determine in the tested pathways which genes contribute the most to the good clustering of patients with distinct expression levels of *JAK2*. Genes with a high SHAP scores are most likely correlated to *JAK2*.

2.6 Model checking using satisfiability modulo theories (SMT)

To infer molecular network motifs from biological observations in Chapter 6, I use Z3 theorem prover, a SMT solver by Microsoft [286]. This section aims to present SMT, which are model checking methods. In the following paragraphs, I first define what is model checking, then I introduce Boolean satisfiability problems (SAT) to finally explain what are SMT.

A fundamental step in the network modelling process involves the addition of specifications. A specification is the formal definition of an observation. For example, a specification can be the loss of DNA repair mechanisms when the gene *TP53* is not functional [287]. The validation that a network satisfies all specifications can be challenging as a result of the numerous possible outcomes some model execution can provide.

Model checking is an automatic verification of the model correctness given a specification in a finite state system [288]. Model checking is a formal verification technique initially designed to verify hardware circuits [289], now applied to many different fields such as executable cell biology [219] and communication [290]. Several model checking tools have now emerged. Among them, SPIN (Simple Promela INterpreter) was released to the public in 1991 but originally designed in the 80s and first applied by Holzmann et al [290] to diverse problems such as controlling telephone exchanges or leader election. SPIN includes a graphical interface called Xpsin which allows users to run simulations and do model checking. Another tool widely used in model checking is NuSMV [291]. NuSMV is an extension of Symbolic Model Verifier (SMV), another tool for model checking, which uses binary decision diagrams (BDDs) to compactly represent sets of states in symbolic model checking [292]. Boolean functions can be represented by truth tables in which every state and outcome of the system is included in a table. However, such representation is only adequate for small systems as the number of states grow exponentially with the number of variables [293]. BDDs are compressed representations of truth tables which can be reduced by removing unnecessary states. One of the main applications of BDDs is model checking, and can be used in many fields such as biochemical networks [294]. Extension of BDDs to non-binary discrete systems is achieved with multivalued decision diagrams (MDDs) where variables take values between 0 and *max* with *max* as an integer [295].

Another model checking method consists in solving Boolean satisfiability problems (SAT). SAT solvers reduce a problem to a Boolean formula and ask whether this formula is solvable via mathematical proofs. A formula in SAT is built with Boolean variables taking TRUE or FALSE values and 3 operators: AND (conjunction), OR (disjunction) and NOT (negation) which mathematical representations are $\wedge$, $\vee$ and $\neg$ respectively. SAT formulas are generally written in the conjunctive normal form (CNF - also called AND of ORs), that is a conjunction (a sum) of one or several clauses written with disjunctions of literals. Literals are Boolean variables and their negation and clauses are finite expression formed with literals. An example of formula in CNF is $(x \vee y) \wedge (\neg x \vee \neg y) \wedge x$, with $(\neg x \vee \neg y)$ an example of clause and x, $\neg x$, y and $\neg y$ the literals. CNF is opposed to disjunction normal form (DNF - OR OF ANDS). SAT solvers ask if a formula is satisfiable or unsatisfiable. After assigning TRUE or FALSE values to all variables, the solver determines if the formula returns TRUE (satisfiable) or FALSE (unsatisfiable). The main advantage of SAT is that they are NP-problems, that is they are in general relatively quick to solve [296].

The main application areas of SAT solvers are computer science and machine learning [297–299]. However, recently, few papers studying cancer and regulatory gene networks have used SAT solvers on biological models. Lin et al [300] use SAT solvers to optimise drug therapies with driver gene mutations defined as 'faults'. Optimisation using these solvers are extremely fast (one second) and predict optimum combination of drugs for different mutational profiles. More recently, a new study succeeded using SAT to design an improved method for drug therapy optimisation with a reduced number of constraints and increased number of results [301]. If no drug is available for a set of mutations, the algorithm predicts a potential target for new drug design. Both studies demonstrate the fast solving property of SAT solvers for complex biological problems.

SAT can be extended for the more powerful satisfiability modulo theories (SMT) problem which consists of a SAT solver with complex non-binary variables. Similarly to SAT, a SMT problem consists of variables and the three operators AND, OR and NEG. However, in SMT, Boolean variables are predicates which contain a large range of possible operators and functions. Predicates are Boolean-valued function, that is they are functions whose binary outcome, TRUE or FALSE, depends on the values of non-binary predicate variables. "$P(x,y) : x > y$" is a predicate with x and y the predicate variables. SMT solvers ask whether a problem is satisfiable or unsatisfiable by reducing it to a SAT formula. It therefore first transforms predicates of a problem into Boolean variables and then uses a SAT solver to solve the problem.

In Chapter 6, I use the SMT solver Z3 to infer molecular motifs from biological observations. Z3 uses DPLL(T) to solve problems [302]. This framework uses the Davis–Putnam–Logemann–Lovela (DPLL) algorithm to solve the Boolean formula generated from the problem and a theory solver to verify the consistency of the values attributed to the predicates by DPLL algorithm. For example, the SMT problem $(x - x^2 \leq 2) \wedge (\exp(x) \neq y) \wedge (sin(x) \geq 0 \wedge x + y = 1)$ can be written as $r \wedge s \wedge (p \wedge q)$ with r, s, q and q the predicates. DPLL will define if the later is satisfiable while the theory solver will check if the assigned values TRUE and FALSE to r, s, q and q are consistent among all the predicates.

The DPLL algorithm starts by assigning a value to an unassigned variable. If there are none, it returns SAT. Otherwise, it simplifies the formula by removing all clauses that are TRUE with this assigned value. From the remaining clauses, DPLL assigns values to the unassigned literals so that all clauses become TRUE. If the current assignment cannot satisfy the formula, then it takes the negation of the first assigned value and tries the simplification again. If it cannot simplify it, the algorithm chooses another variable for the assignment. The algorithm ends when it has assigned values to all literals so that all clauses are TRUE (SAT) or when there is no variable left to assign a value to (UNSAT). In the case of SAT, the theory solver then evaluates the "feasibility" of the assigned value of each predicate, and determines if the problem is satisfiable or not. Detailed mathematical background about DPLL(T) can be found in [302].

Chapter 3

Modelling therapeutic resistance dynamics of lymphoma mice cohorts.

Abstract

The stochastic evolutionary process of mutations in cancer leads to the development of drug resistance and often to inefficient therapies. Insights about these processes can lead to better predictions of tumour growth and with that improved treatment strategies. To this end, I have developed computational models describing the evolution of sensitive and resistant subpopulations of a tumour with Luria-Delbrück-like growth, under different treatment regimes. Using rule-based modelling and parameter inference, this work recapitulates biological experiments depicting resistance emergence with a p53 restoration model in Eμ-myc lymphoma under daily drug administration or in the absence of treatment. However, the calculated growth parameters appear to contradict *in vitro* experiments and alternative periodic treatment regimes do not fit our model. The addition of a "regrowth" process is necessary to fix this issue and identify correct parameters, indicating increased competition between cells in some treatment strategies. This competitive regrowth process suggests a counterintuitive response to the removal of drugs, where the substantially larger sensitive cell population is able to regrow faster than the resistant population despite their apparently similar relative fitnesses. Further *in silico* simulations of alternative drugging strategies demonstrate that maximal survival can be obtained with shorter daily drug administration.

3.1 Introduction

Tumour heterogeneity plays a key role in tumour development and often complicates treatment strategies due to resistance emergence [303]. Despite this, less than 1% of published cancer clinical trials cited evolutionary principles [304], which potentially alter interpretations of therapeutic complications. Knowledge about tumour dynamics and resistance evolution then represents a major advantage when designing treatment strategies to eradicate cancer, increase patient's survival and/or decrease treatment toxicities. These dynamics can also explain the stratification of patient treatment outcomes and help to improve drug administration.

Here, I present how rule-based models (Methods sections 2.3.1 for details) and parameter inference can help to describe sensitive and resistant population dynamics in a lymphoma mouse model receiving p53 restoration therapy [1] (data description in Methods subsection 2.2.1). Parameter inference is achieved on computational tumour growth models with the help of BioPEPA and ProPPA, two computational languages that perform stochastic and deterministic analysis using a single model syntax (Methods sections 2.3.2 and 2.3.3 for details). In particular, analyses show that lymphoma tumours in these experiments have a Luria-Delbrück-like (LD-like) growth. Despite the weak competition between clones, the addition of a "regrowth" process is necessary to fully capture the dynamics of cancerous cells. Thanks to these insights, additional simulations help to optimise a treatment strategy *in silico*. Notably, reduced drug dosage and regular injections could significantly improve survival.

3.2 Results

3.2.1 A simple rule-based model can describe lymphoma growth in mice.

Rule-based models are very effective and succinct way to study complex biological systems using a simple language. I therefore choose rule-based modelling to describe and understand clonal evolution and resistance emergence in lymphoma tumours. This first section aims to depict the framework for the lymphoma model construction. The presented model is inspired by traditional cancer growth models from the literature [240, 305, 306] and from the Luria-Delbrück paper [67] for the mutation emergence. The model assumes that resistance arises from sensitive cell division where a sensitive cell can give birth to one sensitive cell and one resistant cell. Reverse mutations are ignored as expected to be rare events [307]. Cancer cells divide and die which are often associated to distinct death and birth rates in computational models. However, the inference of the death and birth rates in the first model describing lymphoma growth in absence of treatment gives a straight line suggesting that both parameters are correlated (Fig. 3.1). Hence, to reduce the number of unknown parameters and simplify inference, the birth and death rates are combined in one single parameter called proliferation rate or viability rate in the rest of this chapter.

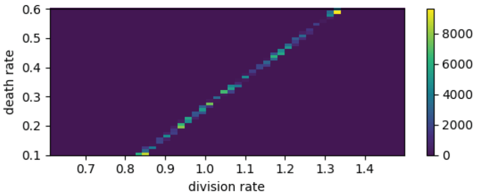


Fig. 3.1 **Inference of the birth and death rates in the control lymphoma model.** Accepted parameters for the birth and death rates of the lymphoma model form a straight line, indicating a correlation between both values.

In this model, A represents the sensitive cells and M the resistant cells. Resistant and sensitive cells have different proliferation rates, as this study investigates a potential fitness cost for the resistant phenotype which can be interpreted as a decreased proliferation rate compared to sensitive cells. When the treatment is applied, an additional death rate is included to the proliferation rate of the sensitive cells. The treatment used on the lymphoma tumours activates p53 thanks to the administration of tamoxifen [1]. The activation is fully competent 1 or 2h after administration and reverts back to null state at 30h. To simplify the model, I assume that the treatment is fully efficient for 24h after administration. Considering all previously described assumptions, the rule-based model for the lymphoma system can be written as follows:

$$\begin{array}{cccl}
A & \xrightarrow{A_A} & A+A & \\
A & \xrightarrow{A_M} & A+M & \\
M & \xrightarrow{M_M} & M+M & \\
A & \xrightarrow{d} & \oslash & \text{when treatment is applied}
\end{array} \tag{3.1}$$

with A_A the sensitive proliferation rate, M_M the resistant proliferation rate, d the treatment killing rate for sensitive cells, and A_M the mutation rate. The unit for all parameters is day^{-1}. All parameters are inferred by ProPPA, except for A_M which gives poor inference results in some analyses. When the inference of A_M value is not possible, I set the mutation rate at $A_M = 5 \times 10^{-5}$, which is the rate used by Iwasa et al [240] for their cancer growth model that shares many similar characteristics with this model.

3.2.2 Quantitative analysis of *in vitro* experiments of [1] is used to evaluate some parameters.

To validate subsequent inference results and minimise the risk of fitting, I approximate the values of some parameters using *in vitro* experiment results described in Martins et al paper [1]. I first attempt to find the treatment killing ratio d using the Annexin V/PI staining data plot from the paper (Figure 3.2). Authors observed that six hours after treatment only 6.5% of the harvested cells are viable against 52% in the control assays indicating an elevated killing rate. Using those values and the exponential decay equation $N(t) = N_0 e^{-Kt}$, I find a 0.35 per hour decay rate difference between cells that had tamoxifen and the ones that did not. Therefore, if the treatment efficiency is the same for 24h, the killing rate d should equal 8.3 cells per day. However, this value seems excessively high. If d was this high, I should observe in the mice experiments a faster tumour reduction right after the first injection (assuming resistance is underdeveloped at that time). A first explanation for this excessive value is that the apoptosis induced by the treatment is not constant during 24h, and 0.35 would be the average decay rate for the first 6h, while it would decrease for the remaining 18h. Another explanation is that harvesting and extracting cells from their environment could be lethal to cells and induce a higher apoptotic rate, which either way falsify the computation for d.

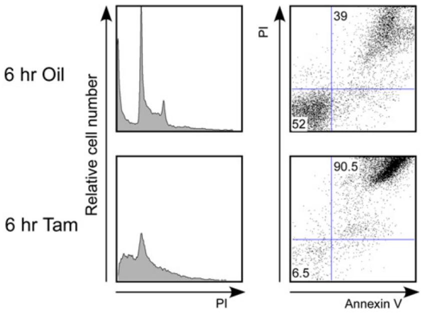


Fig. 3.2 ***in vivo* impact of p53 restoration on lymphoma tumours.** Lymphoma cell analyses by flow cytometry for DNA content (left) or viability using Annexin V/PI staining (right) from control mice (Oil) and mice treated with tamoxifen (Tam). Annexin V/PI staining is a technique to determine cell death rate. Percentage of viable cells in each sample is given in the right upper corner while the left lower corner gives the percentage of dying cells. Mice are sacrificed 6 hours post injections. More details about specific experimental settings can be found in [1].

To confirm the unrealistic high value found for d, I use the mice experiments describing *in vivo* lymphoma growth to compute how many cancer cells survived 24h after the first tamoxifen injection. On average, 29.9% of the cells present at day 10 survived at day 11.

This average however does not include two mice in which treatment had no effect. In these mice, *in vivo* experiments reveal an increased number of tumour cells between day 10 and 11. If I include all the mice, cell survival after the first injection is 61.8%. Therefore, a sensitivity ratio close to 8.3 seems improbable, except if the overall growth rate of lymphoma cells is about 9. This would imply that these cells divide every three hours which is biologically unrealistic. I conclude that d has to be inferred by ProPPA.

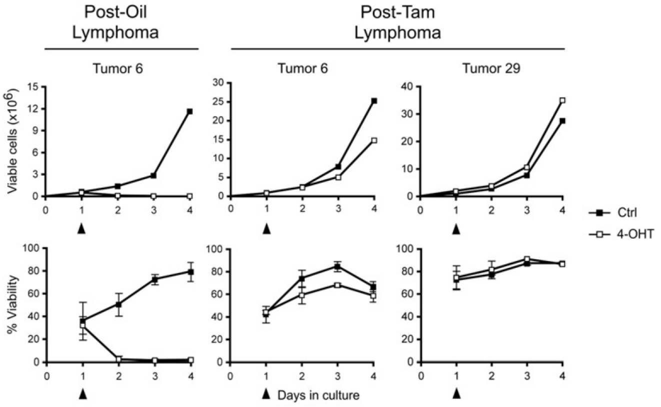


Fig. 3.3 **Resistance emerges in lymphoma tumours treated with tamoxifen.** Cells harvested from mice previously treated with oil (Oil) or tamoxifen (Tam) are cultured in absence (Ctrl) or presence (4-OHT) of tamoxifen. The proliferative rate (upper panels) and viability (lower panels) of tumor cells was determined by Trypan Blue exclusion. This figure demonstrates that cells from mice that had received tamoxifen injections are not responsive to new injections, while control cells show an important decrease in cell viability. Tamoxifen or oil were added to the cultures at day 1, indicated by arrowhead. More details about specific experimental settings can be found in [1].

An approximation of the sensitive and resistant cell proliferation rates can be deduced thanks to Figure 3.3 from [1]. The software Engauge Digitizer [308] is used to extract data from the picture. Considering that cells have an exponential growth of shape $A_0e^{A_At}$ and $M_0e^{M_Mt}$, a linear regression is performed on the logarithmic data using the "stats" R package [309] to find A_A and M_M in the different tumours (see Table 3.1). The non-treated tumour 6 helps to recover the sensitive proliferation rate, as I assume that resistant cells are present in significant lower proportion when no treatment is applied. Mutant proliferation rate is found thanks to the treated tumour (Tam). However, Tam curves most likely also include the death of the remaining sensitive cells still present in the tumours and therefore can slightly decrease the resistant proliferation rate. As *in vitro* experiments are used to find A_A and M_M which are *in vivo* proliferation rate and as the extraction of figure points and the regression might bring some imprecision, these values only serve for comparison purpose with the ProPPA inferred parameters.

Tumour 6	Tumour 6 Tam	Tumour 29 Tam
$A_A = 0.93 \pm 0.12$	$M_M = 1.12 \pm 0.02$ (ctrl) $M_M = 0.91 \pm 0.03$ (Tam)	$M_M = 1.14 \pm 0.02$ (ctrl) $M_M = 0.92 \pm 0.06$ (Tam)

Table 3.1 **Values of cell proliferation rates calculated from the *in vitro* experiments of Martins et al paper [1].** Sensitive A_A and resistant M_M proliferation rates are found using regression on data extracted by Engauge Digitizer from the result figures in Martins et al [1].

3.2.3 Logistic LD model reproduces lymphoma growth under control and continuous treatment regimes.

Once the lymphoma model and the parameters to infer have been defined, I start tumour dynamics analyses by examining the most common treatment strategies, typified by Martins et al [1]. The first approach is a control regime with no drug applied, while a second strategy consists of a daily drug administration regime, dubbed "continuous treatment" due to the treatment 30h efficacy. In their paper, Martins et al present *in vitro* experimental data showing tumour growth and treatment efficacy for the control and continuous regimes, alongside *in vivo* measurement of survival. In additional unpublished experiments given by Martins personal communication, *in vivo* tumour growth dynamics in mice are quantified by fluorescence for various treatment strategies. The mean survival in the control group equals 24 days post tumour infection and as expected, is the lowest of all treatment regimes. When tamoxifen is daily administrated for 7 or 14 days at day 10 post tumour injection, survival is extended to 34-35 days.

I first want to validate or refute the hypothesis that lymphoma in these quantified *in vivo* tumour growth experiments possesses LD-like growth characteristics. To do so, I build rule-based models with LD-like growth characteristics for each treatment schedule and infer unknown parameters with ProPPA. To simplify parameter inferences, the proliferation of sensitive cells A_A inferred with the model without tamoxifen is not inferred again in the other models. I assume low resistance in this model as a result of the high tamoxifen killing efficacy in these control tumours [1]. The treatment killing rate d, the mutation rate A_M and the proliferation rate of the resistant cells M_M are therefore only inferred for the continuous schedule. I first apply approximate Bayesian computation to infer parameters from the data (Methods section 2.3.5 for details). However, due to slow stochastic inferences and the high number of cells, I decide to change for a deterministic approach, called the fluid approximation algorithm (Methods section 2.3.6 for details). ProPPA identifies parameter values that satisfy the input criteria defined by the user so that model simulations with sampled parameter values are the closest from the observations. As such, the output from each inference is a set of accepted parameter values for the computational model. Accepted parameter values which appear at the highest frequency in the set are considered the ones describing better the experiments.

Using BioPEPA, I simulate the control and continuous models with the inferred parameter values found by ProPPA for different initial number of sensitive cells. I compare the inference results using the root mean squared errors (RMSE) which help to identify the best fitting parameters and initial condition (Methods section 2.3.7). These simulations highlight that the rule-based model with LD growth is incomplete. Figure 3.4A and 3.4B show that while simulated cells grow exponentially all along, cancer cells in experiments have a reduced growth the days immediately preceding the mouse death. This phenomenon is observed for both control and continuous regimes. I decide to modify the exponential growth described in standard LD models and add a carrying capacity K to the tumour proliferation rates A_A and M_M. K represents the maximum number of cells a mouse can support before it dies due to the tumour invasion. This modified growth function is referred as logistic and is commonly used in cancer growth models to depict space and resource limitation [310]. For a range of initial numbers of cells in the tumours ($A(t=0)$), I infer in the control regime A_A the sensitive population proliferation rate and K the carrying capacity for the logistic growth, before inferring the rest of the parameters in the continuous model. Tables of parameter values and scores for both models can be found in Table 3.2 and Table 3.3. As shown in Figure 3.4C and 3.4D, the growth update improves data fitting with the simulations of the new model.

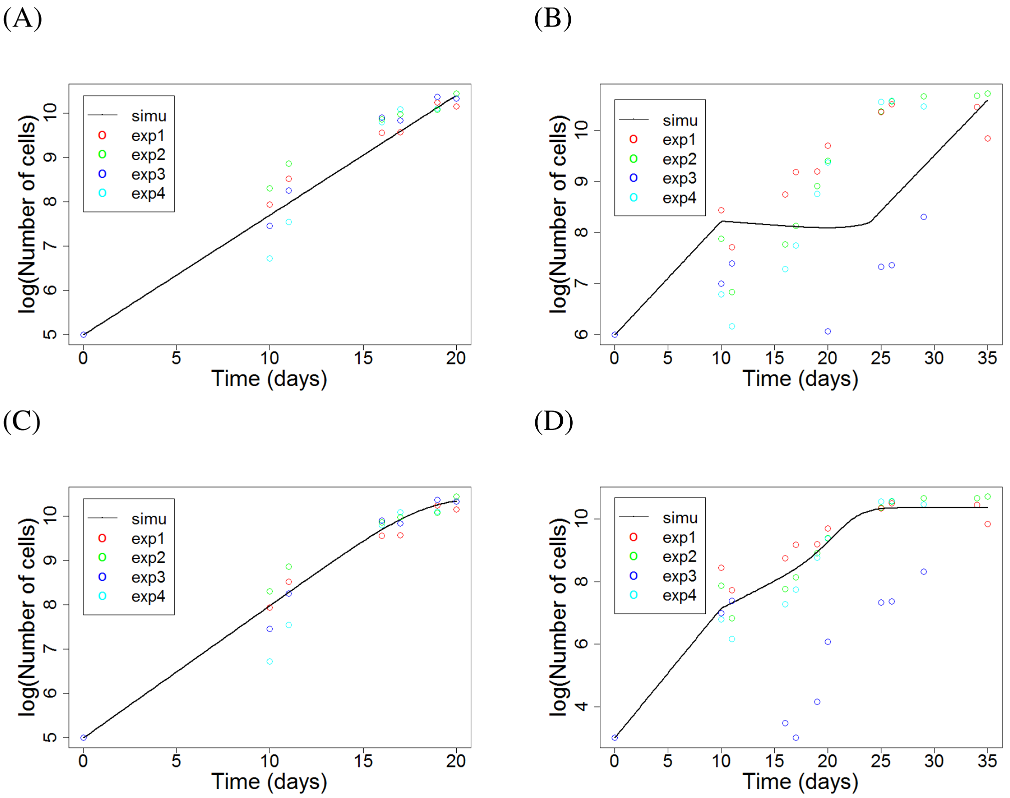


Fig. 3.4 **Best fit of the standard (A,B) and logistic (C,D) Luria-Delbrück models for the control (A,C) and continuous (B,D) simulations and experiments.** I use the root mean squared error, RMSE, to compare the inferences. The RMSE scores and visualisation of the plots confirm that logistic LD growth better captures lymphoma tumour growth in the mice. Therefore, high tumour burden in later stages reduces the growth rate *in vivo*. Black lines show the simulation of the model and initial conditions with the highest RMSE. Dots represent the quantified *in vivo* lymphoma growth in the four mice.

The RMSE scores for the control and continuous regimens indicate that the LD-like model with a carrying capacity better describes the experiments. The observed logistic growth suggests that carrying capacity and tumour burden for mortality are related. However, whilst simulations of this model correctly describe the tumour growth in both treatment regimes, the parameter values for the logistic LD model appear incompatible with *in vitro* experiments from Martins et al [1]. In particular, Figure 2C from Martins et al [1] emphasises that the drug is very efficient at killing sensitive cells. This experiment indicates that I should find in the inference a lower value for the sensitive proliferation rate A_A compared to the killing ratio d, which is not what I observe (Tables 3.2 and 3.3). Similarly, as shown by Table 3.1, resistant and sensitive cells have similar proliferation rates in experiments which is not what I find in the inferences. Therefore, this data in isolation is not sufficient to resolve the tumour dynamics between different experimental measures of lymphoma growth.

	Luria-Delbrück		Logistic LD		
$A(t=0)$	A_A	score	A_A	K	score
10^2	0.97 [0.9668, 0.9672]	4.1	1.1 [1.0950, 1.0958]	$2.2\ [2.224, 2.237] \times 10^{10}$	3.6
10^3	0.86 [0.8625, 0.8627]	3.8	0.96 [0.9613, 0.9616]	$2.3\ [2.325, 2.362] \times 10^{10}$	3.4
10^4	0.74 [0.7377, 0.7380]	3.5	0.82 [0.8207, 0.8211]	$2.6\ [2.629, 2.640] \times 10^{10}$	3.3
10^5	0.62 [0.6222, 0.6226]	3.4	0.69 [0.685, 0.686]	$2.9\ [2.940, 2.945] \times 10^{10}$	3.3
10^6	0.51 [0.5071, 0.5078]	3.4	0.54 [0.5434, 0.5439]	$3.8\ [3.795, 3.806] \times 10^{10}$	3.5

Table 3.2 **ProPPA inference results for the vehicule experiments.** ProPPA inference results between the standard and logistic Luria-Delbrück models for the vehicule experiments, that is when no treatment is applied.

	Luria-Delbrück				Logistic LD			
$A(t=0)$	M_M	d	A_M	score	M_M	d	A_M	score
10^3	0.68 [0.6802,0.6814]	1.3 [1.322,1.334]	5.3×10^{-5} [5.21,5.46]	5.1	1.1 [1.118,1.123]	0.56 [0.5624,0.5661]	5.0×10^{-5}	4.3
10^4	0.61 [0.6123,0.6139]	1.2 [1.165,1.186]	3.5×10^{-5} [3.12,3.54]	5.0	0.98 [0.9835,0.9861]	0.50 [0.5023,0.5042]	5.0×10^{-5}	4.3
10^5	0.56 [0.5604,0.5632]	0.90 [0.8948,0.8994]	8.2×10^{-5} [7.92,8.49]	4.8	0.86 [0.8627,0.5641]	0.46 [0.4621,0.4640]	5.0×10^{-5}	4.4
10^6	0.47 [0.4677,0.4685]	0.55 [0.5437,0.5513]	5.0×10^{-5}	4.7	0.71 [0.7117,0.7130]	0.20 [0.2041,0.2054]	5.0×10^{-5}	4.6

Table 3.3 **ProPPA inference results for the continuous treatment experiments.** ProPPA inference results between the standard and logistic Luria-Delbrück models for the continuous regime experiments, that is when treatment is applied every day.

3.2.4 Competitive regrowth is selected in alternative treatment regimes with parameter values matching *in vitro* experiments

Toxicity in chemotherapy is a major concern for researchers and clinicians [311, 312]. Cancer patients often endure heavy treatments with many side effects. Clinicians first aim to prolong survival but also increase their patient quality-of-life by reducing treatment toxicity. Periodic treatments have been designed for this purpose, to avoid daily treatment and increase intervals of drugs intake. In addition to control and continuous regimes, Martins et al also examine two periodic regimes: drug injection every 3 and 5 days (referred to as periodic3 and periodic5). Survival in the periodic5 regimes vary from 25 to 31 days, with a mean at 28 days post infection, whilst periodic3 has a mean survival of 32 days. Using the same inference technique as in the previous section 3.2.3, I examine if the same logistic LD model can explain alternative treatment regimes such as periodic regimes. By doing so, robustness of the parameter inference method is tested while these new analyses allow further investigation on the contradicting parameter values found in the previous section.

As shown in Figures 3.5A and 3.5B, BioPEPA simulations of the rule-based model with the logistic LD growth underestimate tumour growth during periodic3 treatment, suggesting that the model lacks some biological mechanisms to fully explain the lymphoma growth in these mice. Alongside the low treatment death rate d found for the continuous regime by ProPPA, this result suggests a link between the growth and the treatment death processes and that a growth process is missing in the model.

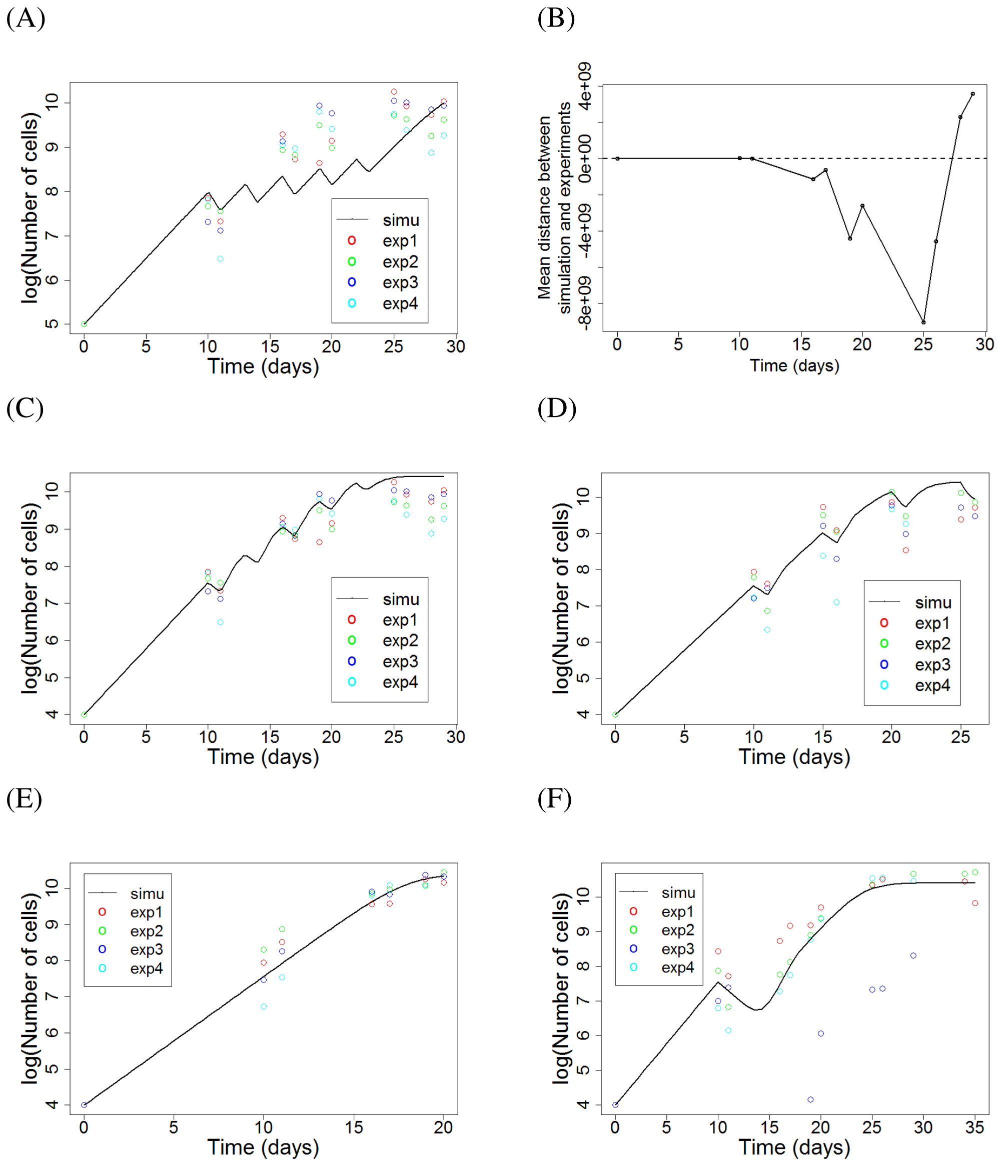


Fig. 3.5 **Addition of a replacement process is necessary for the model to fit the data.** (A,B) The underestimated growth in periodic3 with logistic LD model is improved by the addition of a replacement process (C-F). The simulations of the new model fit periodic3 (C) and periodic5 (D) treatment data as well as the control (E) and continuous (F) regimens.

To address this underestimated growth, an additional mechanism needs to be included to the model. Dying cancer cells release nutrients and modify their microenvironment which enable surviving cells to boost their proliferation and repopulate tumours after cytotoxic therapies [313]. We therefore decide to include a new "replacement" process in our model: when treatment kills sensitive cells, the proliferation of neighbouring resistant *M* and sensitive

A cells is induced to fill the empty spaces *R* or use the nutrients released after cell death. The updated rule-based model is as follows:

$$\begin{array}{rcll}
A & \xrightarrow{A_A} & A+A & \\
A & \xrightarrow{A_M} & A+M & \\
M & \xrightarrow{M_M} & M+M & \\
M+R & \xrightarrow{rM_M} & M+M & \\
A+R & \xrightarrow{rA_A} & A+A & \\
A & \xrightarrow{d} & R & \text{when treatment is applied}
\end{array} \tag{3.2}$$

with *R* the empty space left by a dying sensitive cell *A* and *r* the replacement rate. This additional competitive regrowth improves the fit of simulations for the periodic experiments as shown in Figures 3.5C and 3.5D, but also for the control and continuous regimens (Figure 3.5E and 3.5F). Another beneficial consequence of the addition of this regrowth process is that new inferred parameter values for this replacement model match the *in vitro* experiments of Martins et al [1]. As shown by the parameter values of the replacement model (Table 3.4), treatment efficacy *d* is now significantly superior to the sensitive proliferation rate A_A. Also, the proliferation rate of the resistant cells is very close to the proliferation rate of the sensitive cells suggesting that the resistant phenotype does not have a fitness cost to the cells, which can explain the absence of strong competition between populations during tumour growth.

Name	Value (days^{-1})
Sensitive proliferation rate A_A	0.82
Resistant proliferation rate M_M	0.75
Mutation rate A_M	5×10^{-5}
Carrying capacity K	2.6×10^{10}
Sensitive death rate due to treatment d	2.0
Replacement rate r	0.7

Table 3.4 **Inferred parameter values for the replacement model with a logistic LD growth.** All parameter values are inferred, except for the mutation rate A_M and the replacement *r* which could not be inferred by ProPPA. A_M is taken from Iwasa et al study [240] and *r* is found by comparing and scoring BioPEPA simulations to experiments with RMSE for a range of *r* values.

An interesting finding with this competitive regrowth model is the addition of the Heaviside function for the replacement rate in the model. The Heaviside function is a discontinuous function such that it equals zero when the argument is negative and one when it is positive. In the rule-based model, each death of a sensitive cell due to treatment creates an empty space or available resource called R. The Heaviside function acts on R such that the proliferation of neighbouring cells is only activated when R is positive, that is when additional space/resource is available. Different replacement processes with and without the Heaviside function were tested and visually compared, but only the use of the Heaviside on R produces a close match with the experiments. This finding suggests that the tumours are not well-mixed, and an increase of available space or resource will not increase the proliferation rates (except if the resource R goes from 0 to any positive number which activates the regrowth process). However, it should be noted that the discrimination between the tested functions was also dependent on the limited number of data of this study.

Simulations of the replacement model with BioPEPA shown in Figure 3.6 give supplementary information about resistance dynamics in the different regimes. Increased drug administration induces a higher proportion of resistant cells when the mouse dies. Consequently, the proportion of resistant cells in periodic5 regimes is almost nonexistent while continuous treatment kills most sensitive cells. Also, even if periodic treatments have decreased drug intake compared to the daily treatment, the frequency of the intake still plays a role in resistant emergence and evolution. Periodic3 selects resistant population more than periodic5 and mice in periodic3 die with a higher proportion of resistant cells in their body (Fig. 3.6C). These insights are crucial for the design of potential new treatment regimes.

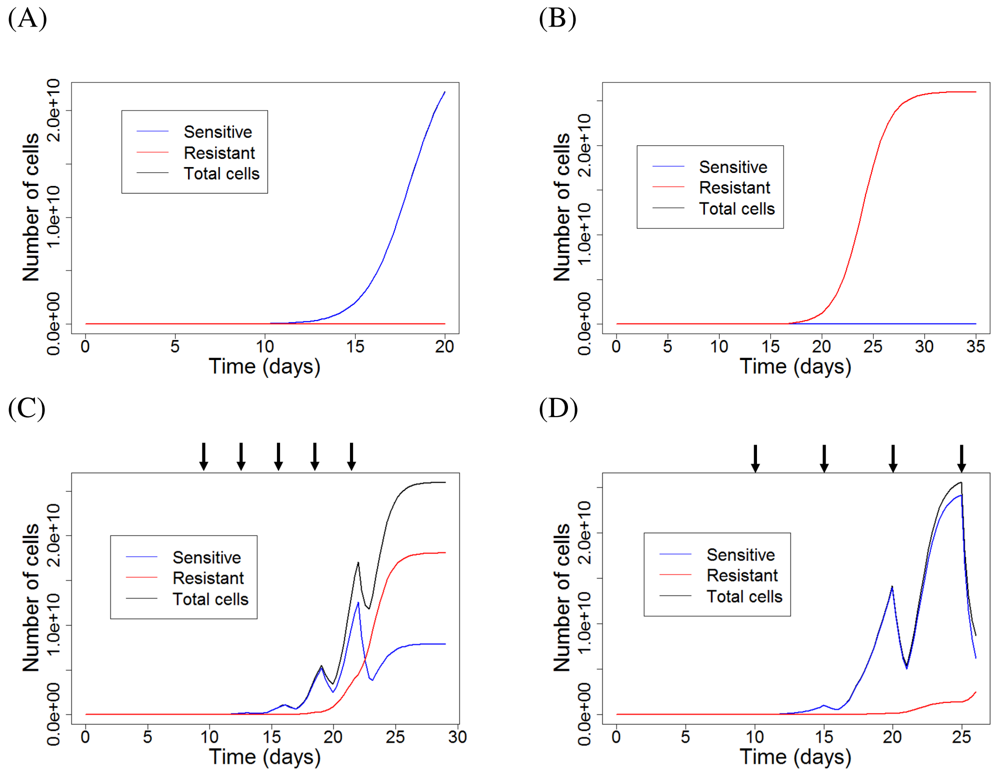


Fig. 3.6 **Time series analyses with BioPEPA of all the treatment regimes with the replacement model.** (A) In the control model, only sensitive cells are growing as the absence of treatment does not select for resistant clones. (B) The continuous regime kills almost all sensitive cells leaving the space and resources to the resistant population which quickly takes over the tumour. (C) Periodic3 does not select for a particular cell population which creates a competitive regrowth process between resistant and sensitive clones. In this regime, mice die with high proportion of both populations. (D) Despite the low drug intake, periodic5 still allows the growth of the resistant cells, however when the mice die, tumours are mostly composed of sensitive cells. Black arrows give the days of the tamoxifen injections in the periodic regimens.

Lastly, I want to use this model to estimate the initial number of resistant cells in the experiments. I find that mice have fewer than 1 in 1000 resistant cells at the start of the experiment. Simulations of tumours with higher initial resistance (1 in 100 resistant cells) have an earlier resistant regrowth which do not fit the *in vivo* lymphoma data points.

3.2.5 Continuous regime can be improved with adaptive treatment regimes.

Next, I hypothesise whether I can use the replacement model to find a better treatment strategy alternative to increase mice survival. A major advantage of *in silico* cancer models is the fast and cheap execution of simulations of potential new treatment regimes. Although biological experiments are necessary to validate computational models, this can considerably speed up treatment designing processes. Using the replacement model, I therefore investigate simulations of different treatment regimes to see if a new treatment schedule can improve the survival of the continuous regime.

As illustrated in the previous section 3.2.4, continuous regime kills all the sensitive cells before the end of the treatment (Fig. 3.6B), suggesting that past a certain time, treatment is inefficient. Daily drug administration can be harmful for patients, thus reducing the drug intake can improve patient's quality-of-life. I design several *in silico* continuous regimes with different treatment period length, and compare them to the 14-day continuous treatment used in experiments. *In silico* simulations demonstrate that once the resistant population has overtaken the sensitive one, any further drug administration becomes ineffective. When the drug is given daily, resistance overtakes sensitive cells at about 5 day post first treatment injection. For this reason, a 5-day continuous treatment regime has a very similar curve and survival than the 14-day continuous regime (Fig. 3.7A). Surprisingly, 6-day continuous regime overlaps the original 14-day continuous curve, indicating that the last 8 days of treatment have no impact on survival while increasing toxicity for the mice.

As described in section 1.4.2, adaptive therapy aims to stabilise tumours and increase patient survival by promoting competition between sensitive and resistant populations [71]. Despite the lack of a resistance cost in our lymphoma model, the existence of a competitive regrowth process suggests an interaction between both population which could be used in treatments. Inspired by the concept of adaptive therapy, a rule-based model in which sensitive population is kept constant is designed as a thought experiment. I find that keeping sensitive clones to a stable state considerably increase the survival which could potentially be extended from 35 days, mean survival of mice in continuous regime in experiments, to 40 days (Fig. 3.7B).

The competitive regrowth process potentially offers an opportunity to extend survival times. Additional *in silico* simulations suggest that daily injections with reduced drug efficacy could in principle extend life (Fig. 3.7C). The lower efficacy treatment leads to a transient but potentially long lasting equilibrium between cell populations which is reached below carrying capacity. This is enabled by the effective first-mover advantage [314] of the sensitive population. Due to its relative higher number, the larger sensitive population is able to quickly reuse available resources and thereby effectively suppress the growth of the resistant population.

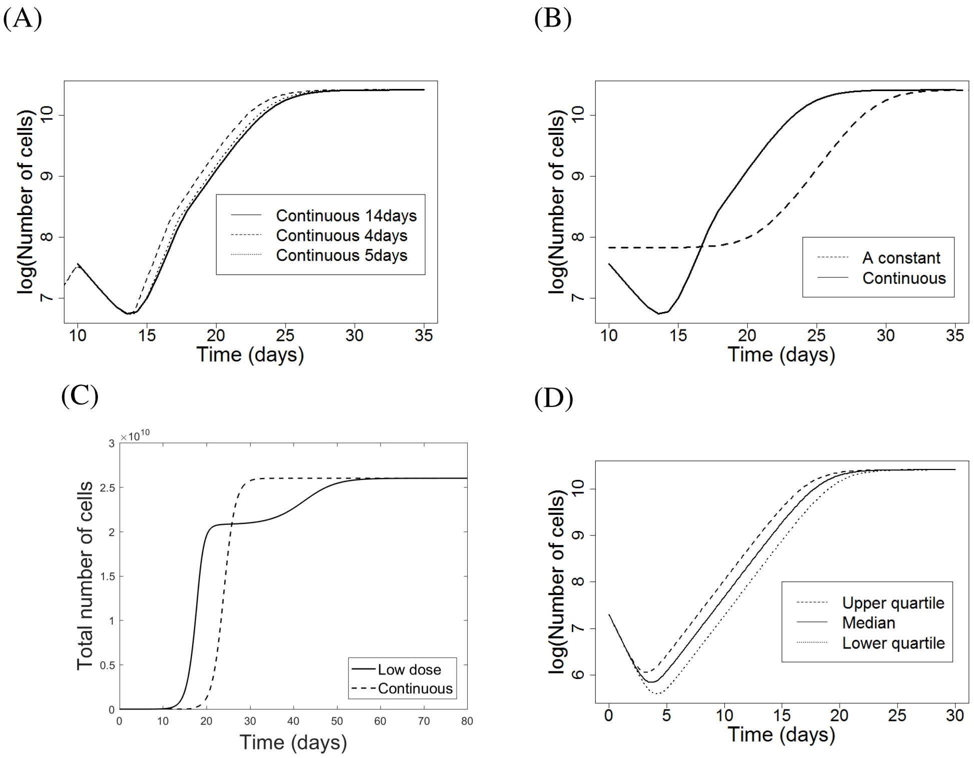


Fig. 3.7 **Additional *in silico* simulations help to identify strategies which reduce treatment toxicity and improve survival.** (A) Reducing treatment length in continuous regime gives similar survival with diminished drug intake. (B) Keeping the sensitive cells A to a constant level with a theoretical treatment regime increases the survival by five days. (C) Additional simulations with new drugging strategies suggest that reducing the efficacy of treatment may allow a temporary restriction of tumour growth due to competition between sensitive and resistant cell populations. (D) Survival can vary by two days between a patient with high resistance profile and another with initial low resistance.

Finally, survival times are prone to variations due to the stochastic emergence of the resistant population in tumours. This is particularly true for the periodic5 mouse cohort which has the highest survival variation of all regimes. Due to the high number of cells in tumours, models have been simulated deterministically. Consequently, these deterministic models ignore the variability in patient resistance profiles. To determine this variability and how it impacts survival, I compute using stochastic simulations of the lymphoma rule-based model the resistance distribution at day 10, that is just before the first injection of tamoxifen. From this distribution, the lower quartile, the median and the upper quartile are defined as the new initial numbers of resistant cells before the treatment is applied. Then, using these new initial conditions, deterministic simulations of the continuous regime model are performed. Simulations show that tumours with higher or lower resistance reach the same maximal tumour burden with a difference of two days (Fig. 3.7D). In periodic regimes, no survival difference is observed between different levels of initial resistance.

3.3 Discussion

In this work, I develop rule-based models and apply Bayesian inference to find parameter values to describe the growth dynamics in different treatment regimes of an *in vivo* lymphoma mouse model from Martins et al [1]. To my knowledge, this is the first study using such approach to propose alternative treatment schedules. Alongside a logistic LD-like model that explains tumour growth in control and continuous treatment regimes, a replacement process, defined as a competitive regrowth induced by sensitive death release of nutrients and space, is necessary to fully describe all treatment regimes including periodic treatments. Addition of a regrowth process also enables to find compatible inferred parameters with *in vitro* experiments. Model simulations confirm that applying daily tamoxifen injections quickly kills all sensitive cells, which leaves space and resources for a rapid proliferation of the resistant clone. In contrast, periodic treatments with longer intervals between injections do not promote such a strong resistance invasion. Finally, additional *in silico* treatment simulations suggest three potential improvements for therapy schedule. First, reducing the 14-day continuous treatment by 8 days should give similar survival with less drug administration. Secondly, aiming at keeping a constant proportion of sensitive cells should increase survival. Lastly, daily injection with a reduced drug efficacy can temporarily stabilise tumours before reaching the maximal tumour burden, and potentially increase mice survival. However, the broader utility of this approach depends on both the treatment type and the relationship between the tumour size and animal death, but it suggests that competitive processes even in liquid cancers may be generally exploitable.

Models in this study reveal interesting biological mechanisms. First, despite the similar proliferation rates of sensitive and resistant cells, this competitive regrowth shows that sensitive cells use their substantially higher population size after drug removal as a fitness advantage to repopulate the tumour faster than the resistant clone. Therefore, a short gap between injections allows the sensitive cells to still outnumber resistant cells and take advantage of the remaining space and resources. Secondly, the logistic behaviour of the tumours suggests that growth is slowed down when approaching the mouse death, probably due to the organ dysfunctions involved in the killing process. Thus, with further studies, estimating the tumour growth rate could predict patient survival as growth rate gradually decreases with tumour burden. Finally, the periodic3 treatment regime leads to a higher competition between cancer cells and increases survival to a similar level as observed with the continuous treatment. This finding agrees with adaptive therapy models described by Gatenby et al [71] in which triggering high competition between populations in tumours capable of such competitive behaviours increases patients' survival without requiring daily drug injections. However, as demonstrated by the theoretical treatment in which sensitive cells are kept to a constant state, triggering competition in these lymphoma tumours is inefficient for a complete recovery as resistance always takes over sensitive cells. This suggests that due to the LD-like behaviours of this particular cancer cell line, competition cannot contribute to extreme survival increase like the one observed in adaptive therapies.

Another interesting finding is the use of the Heaviside function for the replacement rate in the rule-based model. This function implies poorly-mixed tumour populations. Experiments with their few number of mice might have also biased this particular function choice. However, this idea of poorly-mixed populations despite tumour heterogeneity has been described in several papers, among which Lloyd et al [39] describe two distinct populations in the core and edge of the tumours due to microenvironment pressure and Raz et al [315] which show that hypoxia induces resistant phenotypes in the core of tumours. A plausible explanation for the good match of the Heaviside function is that treatment kills sensitive cells in some specific and easy-to-access areas creating a large space of resources R with only the outlying resources/spaces accessible to neighbouring cells while further central resources R become accessible once the outlying ones have been consumed.

I run parameter inference in ProPPA using a deterministic approach. However, I first started with a stochastic method called Approximate Bayesian Computation (ABC). Due to the high number of cells (reaching up to 10^{10} cells), simulations with this approach were extremely time consuming. Changing the Gillespie algorithm used by ABC to Tau-Leap, an algorithm which is faster, did not improve simulation times. Moreover, ABC outcomes can vary substantially with the input configuration values. For these reasons, I decided to switch to a deterministic algorithm and ignore stochasticity in the models for faster simulations. However, as illustrated with simulations of patients with distinct resistant profiles (Fig. 3.7D), the survival varies only by one day if the resistance is low or high compared the median and deterministic behaviour. Therefore, I believe that ignoring the stochastic resistance emergence does not considerably alter the results of this study.

Some comments about the models should be pointed out. First, further biological experiments to validate findings should be performed. In particular, testing the continuous low dose regimen on a new cohort of mice could confirm the increased survival we observe *in silico*. Secondly, quantified tumour growth by photon counts was translated into number of cells using Figure 3.8 from Sweeney et al paper [316]. This necessary conversion could have introduced errors. To explore this possibility, further inference work on the photon count data was performed. Inference and model fit results were unchanged which therefore invalidate potential errors caused by the conversion.

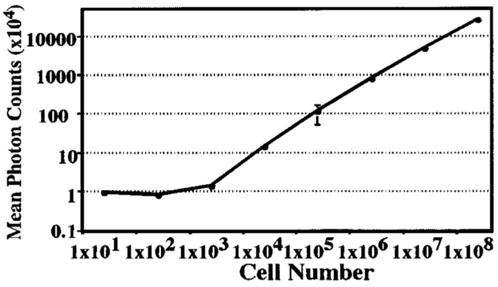


Fig. 3.8 **Mean photon counts per cell number.** Mean photon counts quantified using the signals of the entire abdominal region of distinct mouse cohorts. More details on experimental settings can be found in [316]. All variations are $< 0.01\%$ except for 2.5×10^5 which variation is $< 5\%$. Copyright (1999) National Academy of Sciences, U.S.A.

Finally, this work suggests that tools and approaches used to describe lymphoma growth could be exploited for other types of cancer with different growth dynamics. Rule-based models and ProPPA parameter inference have helped to better understand $E\mu - myc$ lymphoma dynamics, and propose treatment strategies which could improve currently existing continuous regimes. The same protocol could be applied on more complex biological cancer models, for example which use multiple drugs, or have drug processes that require several steps to kill the cells.

Chapter 4

HOXA9 acts as an epigenetic switch in blood malignancies.

Abstract

Blood malignancies arise from the dysregulation of haematopoiesis. The type of blood cell and the specific order of oncogenic events initiating abnormal growth determine the ultimate cancer subtype and subsequent patient clinical outcome. *HOXA9* plays an important role in leukaemia prognosis by promoting blood cell expansion and altering differentiation. However, the function of *HOXA9* in other blood diseases is still unclear. Here, I demonstrate the importance of this gene in Myeloproliferative Neoplasms and highlight the biological switch and prognosis marker properties of *HOXA9* in AML and MPN. This switch function can explain the branching evolution of these two blood disorders. First, I establish the ability of *HOXA9* to stratify AML patients with distinct cellular and clinical outcomes. Then, modelling MPN as a qualitative network, I show that the self-activation of *HOXA9* and its relationship to *JAK2* and *TET2* can justify the branching progression of *JAK2*/*TET2* mutant MPN patients towards divergent clinical characteristics. Finally, I predict a connection between *RUNX1* and *MYB* genes and a suppressive role for the NOTCH pathway in MPN diseases.

4.1 Introduction

AML and MPN diseases are hematologic malignancies affecting the myeloid lineage and resulting in blood cell overproduction. Despite these similarities, genetic alterations, symptoms and prognosis differ between both diseases. For example, the *JAK2* mutation is the main driver event of MPN diseases yet is rarely found in AML [317]. However, myeloid lineage dysregulation by both MPN and AML as well as the ability of MPN to evolve to AML indicate that both diseases may share dysregulated biological mechanisms. The identification of these processes will help identify aberrant genes and pathways in blood malignancies that could be ideal targets for new drugs. It is known that the frequency of MPN transformation to secondary AML is highly related to the initial MPN disease type [126–129]. Therefore, a better understanding of the molecular events driving the different subtypes of MPNs is essential to help clinicians diagnose patients with higher risk of thrombosis and AML progression.

Better understanding of the patterns of genetic alterations in cancer cells can be used for the classification of analogous blood diseases and evaluation of the risk of developing severe later stage diseases [318]. Mutations may occur multiple times in a clone over a lifetime, and therefore modelling how one mutation affects another is important to understand how the cancer progresses for a better classification. How different combinations and orders of mutations lead to different subtypes of cancer remains a major open question [319]. In adrenocortical tumours for example, the order between *RAS* and *TP53* mutations leads to either highly malignant or benign tumours [320]. Yet whilst it is now possible using experimental and genomic data to infer the order of mutation in neoplasm evolution[321–323], the interaction between different mutations as the tumour evolves, how this determines the future of the tumours, and the influence of order is relatively poorly characterised [324]. The importance of mutation order has been demonstrated in MPN by Ortmann et al [325] who show that two subpopulations of patients with MPN can be determined by the order of mutation acquisition between *TET2* and *JAK2* genes. Further analyses of these cohorts show that patients with *JAK2* mutated before *TET2* are younger at presentation of the disease in clinics, are more likely to present PV, have a higher risk of thrombosis and respond better to Ruxolitinib, a JAK2 inhibitor drug. However, the molecular interplay between both mutations within cancer cells and how their order rather their combination triggers dissimilar clinical characteristics have not been investigated.

Overexpression of a single homeobox gene, *HOXA9*, has been reported as sufficient to quickly induce myeloproliferation, gradually followed by AML progression after a period of time [326]. Homeobox genes or HOX genes were first identified in the fruit fly Drosophila melanogaster as essential regulators of early embryogenesis [327], and are thought to have a critical role in cancer development [328]. Among these homeoproteins, the HOXA cluster is essential to normal human haematopoiesis [329] and is often involved in leukemogenesis [330]. In the HOXA family, *HOXA9* is the most described gene in literature as its expression was shown to be the single most highly correlating factor, out of 6817 genes tested, for poor prognosis in AML [331]. The importance of *HOXA9* in AML has been widely explored. However, this has mainly focused on specific AML subtypes such as MLL-rearranged leukaemia [332] and NUP98-HOXA9 induced leukaemia [333], while its role in other blood malignancies such as MPN or other AML subtypes is poorly characterised. Recently, the oncogenic property of *HOXA9* has been associated with its self-positive feedback loop in myeloid precursor cells as a result of its ability to bind its own promoter [334]. This study investigates the hypothesis that this specific biological skill can help stratify patients with blood cancers affecting the myeloid lineage.

A biological switch is defined in this thesis as a molecule producing distinct cellular phenotype when its expression is activated or inhibited. Using public datasets from AML patients and MPN studies, the work presented here demonstrates the biological switch function of *HOXA9* by splitting data into two distinct cohorts of patients/mice with antagonistic expression for this gene. The bimodal expression of this homeobox protein induces a branching evolution in these blood diseases by separating individuals into two clinically divergent populations. Clinical heterogeneity includes distinct prognosis outcomes, but also specific disease type classification. First, *HOXA9* bimodal expression affects the clinical features such as age and white blood cell counts as well as the stratification of AML patients into specific FAB or molecular subtypes. This stratification is supported by the genetic profile of each cohort. Next, a molecular network model is designed to describe MPN progression in patients with *JAK2* and *TET2* mutations. *HOXA9* in this model forms with *JAK2* and *TET2* a memory motif, causing a phenotypic switch in double mutant cells with different mutation order and producing distinct types of diseases. Finally, the network also predicts a suppressive role for the NOTCH pathway in MPN and a new interaction between *RUNX1* and *MYB*. Overall, this work shows the significant influence of *HOXA9* in two distinct myeloid blood disorders, MPN and AML, all subtypes included, and establish new essential properties for *HOXA9* in blood cells.

4.2 Results

4.2.1 *HOXA9* expression separates distinct cohorts of AML patients with distinct clinical features.

Ectopic expression of *HOXA9* in AML has been widely demonstrated, but few studies have investigated the biological attributes of this transcription factor contributing to leukemogenesis. Zhong et al [334] have shown that *HOXA9* in murine myeloid cell lines can potentially induce its own expression thanks to a positive feedback loop, which promotes a continuous differentiation block and self-renewal increase of hematopoietic stem cells leading to leukaemia development. To validate *HOXA9* self-activation and its oncogenic role in leukaemia in general, I study its expression in AML patient RNAseq data from The Cancer Genome Atlas (TCGA) [132] (data description in subsection 2.2.1). Data show that *HOXA9* has bimodal expression in these patiens (Fig. 4.1A). This bimodality separates patients into two cohorts: 31 patients in the low expression peak, and 80 patients in the high expression peak. Survival analyses are performed for both groups using Kaplan-Meier survival curves and the log-rank test (R package survival [335]) and confirm that *HOXA9* can be used as a marker of poor prognosis in AML (Fig. 4.1B). Kaplan-Meier curves estimate the percentage of survival at each time point taking into account censored data such as patients withdrawn before the end of the study. The log rank test is a statistical test checking if survival curves differ significantly. The survival probability for AML patients with high *HOXA9* expression is 0.19 after 3 years versus 0.60 for patients with low expression. This patient stratification based upon *HOXA9* expression supports the reported positive feedback loop characteristic of this gene and suggests that once activated or inhibited, the gene would have the ability to keep its expression as is and create a branching evolution in the disease progression.

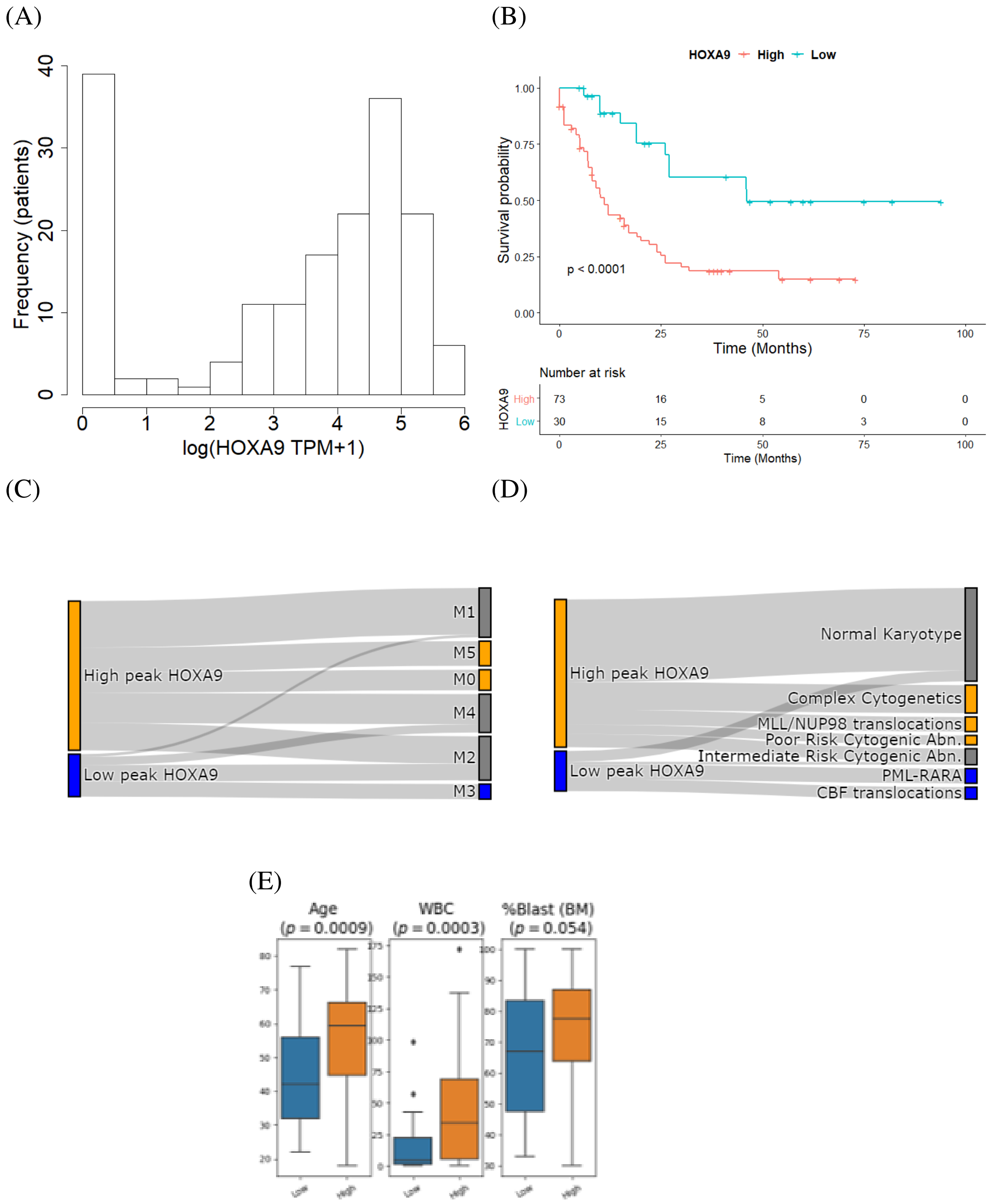


Fig. 4.1 ***HOXA9* low and high expression stratify patients in AML.** (A) *HOXA9* expression in AML patients is significantly bimodal (ACR unimodality test rejected with $p < 2.2 \times 10^{-16}$, using the R package multimode [336]), suggesting a role as a genetic switch. (B) *HOXA9* high and low cohorts have divergent AML prognosis, consistent with known *HOXA9* biology in AML. *HOXA9* expression also partially explains (C) FAB and (D) molecular classifications of AML patients. Using Sankey diagrams, I find that the M3 FAB subtype as well as the PML-RARα and CBF translocations (CBFB-MYH11 and RUNX1-RUNXT1) are solely linked to low expression of *HOXA9*. Similarly, the high cohort possesses specific AML subclasses: FAB M0, FAB M5, MLL/NUP98 translocations and complex cytogenetics. (E) *HOXA9* cohorts show distinct clinical characteristics: high cohort patients are older, display a higher white blood cell counts (WBC) and tend to have higher percentage of blasts in the bone marrow.

To investigate how the switch role property of *HOXA9* impacts AML subtypes, I look at the distribution of French-American-British (FAB, named M0-M7), and molecular classifications among the two *HOXA9* cohorts. Distributions show that different *HOXA9* expression cohorts exclude specific FAB subtypes (Fig. 4.1C). In this dataset, none of the 10 M0 and 15 M5 AML patients are found in the low-*HOXA9* cohort while all 10 patients with M3 AML are observed in the low-*HOXA9* cohort. This specific distribution among cohorts suggests the important role of *HOXA9* expression in the clinical characteristics of blood disease patients.

In light of these findings, I looked to characterise the common features of *HOXA9* expression cohorts. Cytogenic aberrations and gene rearrangements are frequent in AML and are known to alter the disease morphology as well as the clinical features and prognosis [136, 132]. I find that *HOXA9* stratifies patients with different molecular classification (Fig. 4.1D). The 10 patients with PML-RARα translocation and the 9 with core binding factor (CBF) translocations, RUNX1-RUNXT1 and CBFB-MYH11, are all found in the low-*HOXA9* cohort. Within the high cohort, the specific cytogenic subtypes are complex cytogenetics and the MLL/NUP98 translocation. MLL-induced leukaemia has been linked to high *HOXA9* [332], while M3 AML subtype is characterised by PML-RARα translocation and low *HOXA9* in the literature [337]. Low *HOXA9* expression in AML with RUNX1-RUNXT1 and CBFB-MYH11 abnormalities, which constitute the core binding factor (CBF) AML, was also established but unexplained in literature [338]. The work presented here further establishes the correlation between high *HOXA9* expression and the M0 and M5 FAB subtypes as well as complex cytogenetics. Finally, *HOXA9* stratification effect is expanded to other clinical features such as age, white blood cell count (WBC) and blast percentage in the bone marrow (Fig. 4.1E). Patients within the high-*HOXA9* cohort are significantly older and have a higher WBC. They also tend to display higher percentage of blasts in the bone marrow.

PML-RARα, RUNX1-RUNXT1 (AML1-ETO) or CBFB-MYH11 chromosomal abnormalities confer good prognosis in AML patients [138, 339]. All these aberrations are linked to low *HOXA9* expression which also exhibits good survival prognosis among patients compared to high expression. To confirm that high *HOXA9* is a poor prognosis marker independently of its associated molecular aberrations or FAB subtypes, I look at survival outcomes within FAB classes. M0, M3 and M5 being all specific to one cohort, I examine the survival of patients within the M2 and M4 subtypes for high and low *HOXA9* expression. Survival curves and log rank tests within both subtypes ($p = 9 \times 10^{-4}$ for M2 patients and $p = 0.051$ for M4 patients) confirm the poor prognosis marker function of high *HOXA9* (Fig. 4.2). These results as well as the diverging clinical features among patient cohorts support the switch role of *HOXA9* within AML diseases.

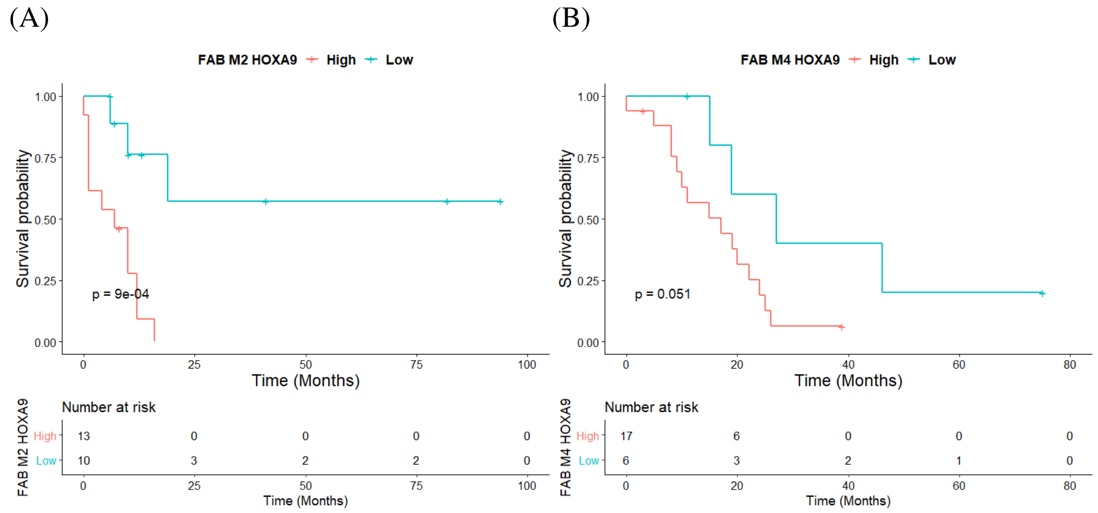


Fig. 4.2 ***HOXA9* high and low cohorts of M2 and M4 AML patients show distinct survival probabilities.** (A) None of the 13 M2 patients with high expression for *HOXA9* survive past 20 months while three among the 10 patients with low expression reach 25 months ($p = 9 \times 10^{-4}$). (B) The Kaplan Meier curves for the survival of M4 patients indicates a trend towards lower survival probability for patients with high *HOXA9* expression compared to the low-*HOXA9* cohort ($p = 0.051$).

The relationship between *HOXA9* expression and clinical characteristics among *HOXA9* cohorts suggests that this HOXA gene would be expected to lead to wide-ranging gene expression changes within blood cancer cells. To explore this, I study differentially expressed genes between low and high *HOXA9* AML cohorts ranked by the absolute value of the mean expression difference. From the 30 most differentially expressed genes, the majority can be classified into 5 functional groups (Fig. 4.3). When a gene has several known biological functions including haematopoiesis, the gene is associated with haematopoiesis. Genes from the HOX family obtain the highest differential expression values and their expression follows *HOXA9* expression. The second top group is composed of markers for the stem/progenitor cells and are all up-regulated with high *HOXA9* expression. Other identified genes are involved in hematopoietic differentiation or are markers for the innate immune system and T cells. Those genes are all down-regulated when *HOXA9* is highly expressed. These observations are consistent with known *HOXA9* behaviours from the literature. This is consistent with the previously identified roles of *HOXA9* in proliferation and repopulating ability of hematopoietic stem cells [340] and with suppression of differentiation [341]. Overall, these results suggest that *HOXA9* bimodality in blood cells acts as a switch in AML, controlling key blood development phenotypes involved in haematopoiesis and the determination of specific lineages, and ultimately the clinical features of the disease.

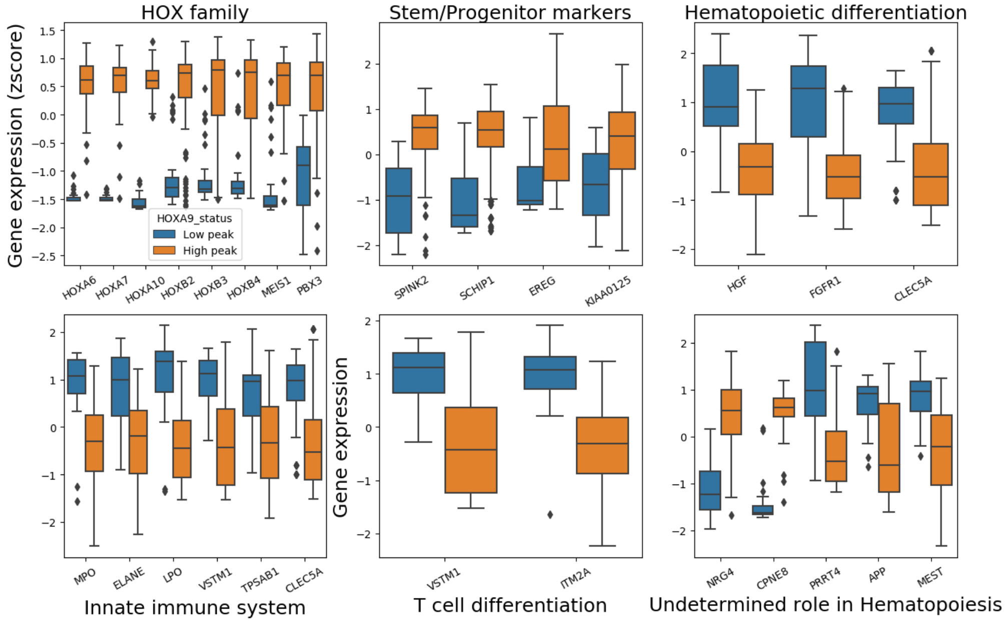


Fig. 4.3 **Top 30 most differentially expressed genes for *HOXA9* cohorts in AML separated by gene family or biological function.** The HOX family is the most represented group, followed by upregulation of the genes involved in stem and progenitor cells and downregulation of genes involved in hematopoietic differentiation and more mature cells. All 30 genes have a significantly different mean between the two cohorts.

4.2.2 The *JAK2/TET2/HOXA9* motif explains divergent clinical outcomes in MPN.

JAK2 is the main driver mutation found in all MPN patients, but different subtypes of the disease with distinct clinical traits are observed [146]. I showed in the previous section 4.2.1 that *HOXA9* can induce clinical stratification in AML thanks to its positive feedback loop. This raises the question whether *HOXA9* expression can also explain the divergent clinical symptoms among MPN patients with both *JAK2* and *TET2* mutations in different orders.

To address this question, I construct a qualitative network in a multistep process. First, I identify the fundamental network features required to achieve branching [342, 343]. The mutation order described in Ortmann et al [325] shows that the first mutation between *JAK2* and *TET2* in the cell sets up the biological background for the second mutation. Therefore, the history of a cell mutational profile impacts its current state. This “memory” property can be induced by a positive feedback loop in cells [344]. A first simple gene motif is designed with *JAK2* and *TET2* genes and a third hypothetical gene target that can provide this memory property (Fig. 4.4). The gene target must possesses a self-activation loop and is necessarily downstream of *JAK2* and *TET2* in order to respond to mutations. Furthermore, as *JAK2* and *TET2* mutations have constitutive activation and loss of function effects respectively, each must have the same signed interaction with the gene target. This motif recapitulates the fundamental properties of mutation ordering; unmutated and single-mutant models are stable (i.e. they have a unique fixpoint attractor and no cycles - methods section 2.4.4 for details on stabilisation analysis), whilst the double mutant has no cycles and two fixpoints that arise from different mutation orders and subsequent activity of the gene target. The cellular phenotypes are determined by their interaction and relationship with the three genes and their mutation combination. The resulting bifurcation provides a potential explanation for the dependence upon order of mutation of the phenotypes observed by Ortmann et al [325].

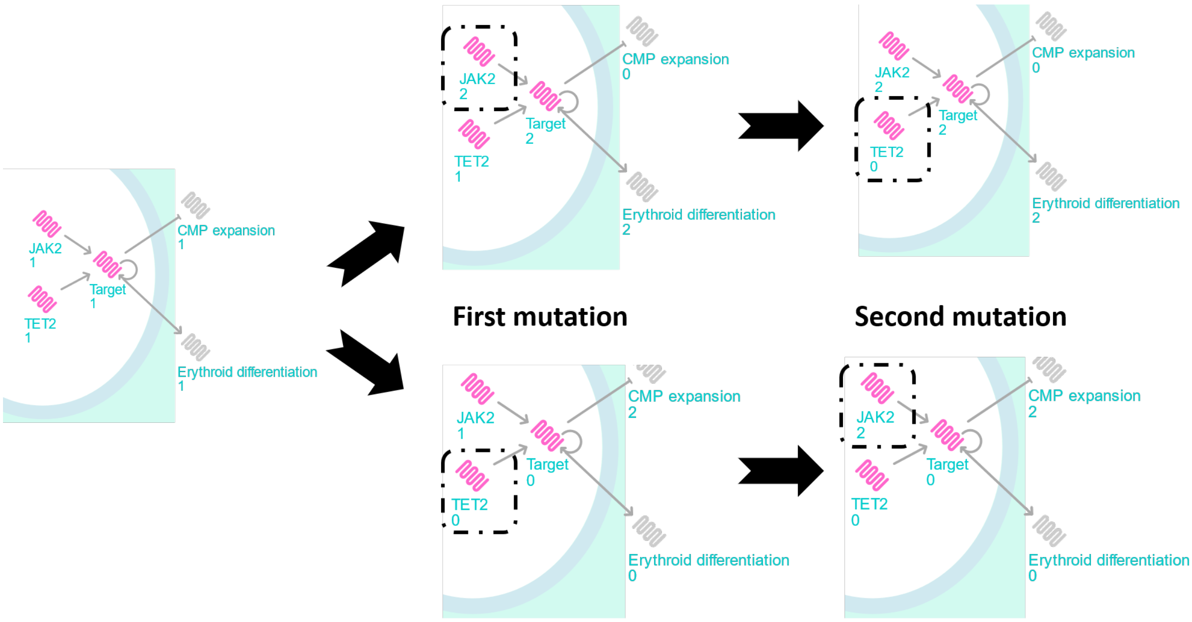


Fig. 4.4 ***JAK2/TET2* double mutant bifurcation illustrated by a simple gene motif.** The different clinical characteristics of MPN patients with *JAK2/TET2* can be explained by a simple gene motif including a memory property. Model starts from a healthy state on the left (wild type) and sequentially acquire mutations in *JAK2* and *TET2* genes. The first mutation affects the gene target expression (middle networks) which remains stable when the second mutation appears (networks on the right). The order in which mutations occur impacts on the gene target expression but also the phenotypes, CMP expansion and erythroid differentiation. Despite the identical final mutational state, the cell shows different hematopoietic behaviours. The value 1 represents the healthy state, 0 the lowered/inactive state and 2 the overactive state.

Once the core motif was defined, I then sought to confirm *HOXA9* as a candidate for the gene target. *TET2* and *JAK2* have been indirectly and directly linked to *HOXA9* activity (Fig. 4.5A). *STAT5* is a well-known downstream target of *JAK2* [345], and it is also established that *STAT5* and *HOXA9* act as binding partners in hematopoietic cells [346, 347]. Furthermore, it was recently shown that tyrosine phosphorylation of *HOXA9* is *JAK2*-dependent [348]. This *JAK2*/*HOXA9* observation mirrors the finding, a decade earlier, of HOXA10 tyrosine phosphorylation by *JAK2* [349]. Moreover, *HOXA9* and HOXA10 share many similarities [350]. In the case of *HOXA9*, this tyrosine phosphorylation seems to increase *HOXA9* effect on its downstream targets [348]. Regarding the interaction of *TET2* with *HOXA9*, Bocker et al found a significant reduced expression of HOXA genes when *TET2* expression is lost [351]. In particular *HOXA9* expression in kidney is significantly decreased by *TET2* lost. *TET1* was also found to positively target *HOXA9* in Mixed Lineage Leukaemia [352], and despite the rare involvement of *TET1* in leukaemia compared to *TET2* [353], both *TET1* and *TET2* expression disruptions share many clinical similarities such as increased hematopoietic stem cells and alterations of both the lymphoid and myeloid lineages [354]. There is evidence that *HOXA9* is activated by both *JAK2* and *TET2* and possesses a self-positive feedback loop property [334]. Therefore, the *JAK2*/*TET2*/*HOXA9* motif shares all the required properties for observing a branching evolution in blood diseases. I therefore propose that *HOXA9* acts as the "memory" of *JAK2* and *TET2* mutation order in MPN. It should be noted that in the complete MPN network of this work, *HOXA9* requires both *JAK2* and *TET2* expression to be active (Table 4.3). Upregulation of either *JAK2* or *HOXA9* results in the increased expression of *HOXA9* while *TET2* loss decreases its expression, and once in either mutant states, *HOXA9* expression retains its activity level as a result of its feedback loop.

Based on this *JAK2*/*TET2*/*HOXA9* motif, the MPN model is extended to reproduce the observed biological differences between patients with different combinations of *JAK2* and *TET2* mutations. To do so, six phenotypes relevant to cancer development are included in the model: stem cell self-renewal, common myeloid progenitor (CMP) expansion, granulocyte-monocyte progenitor (GMP) expansion, GMP differentiation, erythroid differentiation and megakaryocyte-erythroid progenitor (MEP) expansion (Fig. 4.5B).

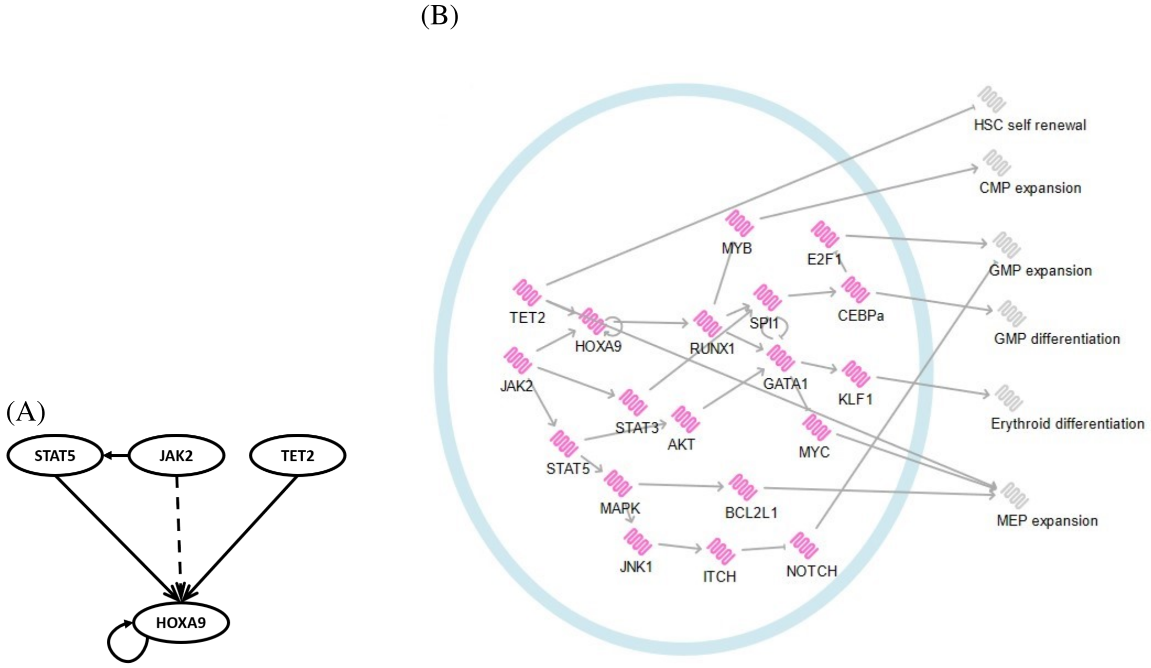


Fig. 4.5 ***JAK2/TET2/HOXA9* molecular network for MPN progression.** (A) *HOXA9* is downstream of both *TET2* and *JAK2* and acts as a developmental switch through its self-positive feedback loop. Direct interaction drawn as continuous arrows and post translation modification as dotted arrows. (B) The *JAK2/TET2/HOXA9* molecular network is built with the BioModelAnalyzer (BMA) tool [191] and integrates six phenotypes considered as the model outputs. Using the Linear Temportal Logic (LTL) model checking tool available in BMA (Methods section 2.4.3), a bifurcation is observed after simulation of the double mutant state. The bifurcation analysis identifies two stable states with different phenotype values that fit the mutation order characteristics observed in MPN patients with *JAK2* and *TET2* mutations.

To capture the wider biology of MPN progression, important hematopoietic markers are incorporated to the molecular network (Table 4.1 and Fig. 4.5B). As shown in Table 4.1, all interactions are found in studies with distinct experimental settings but all focusing on hematopoietic cells. For the stem cell self-renewal, I assume *TET2* is the only gene that can modify stem cell properties as in all the genes in the network, it is the only potential genetic regulator having a clear and undoubted role in early stem cells [158, 162]. I then search for transcription factors involved in the myeloid lineage. *SPI1* and *CEBPα* are identified as important markers of the monocyte and macrophage lineages [355]. *SPI1* activates *CEBPα* in early progenitors [356, 357] and *CEBPα* helps the transition from CMP to GMP [358]. Both genes are essential for myeloid differentiation and are downregulators of progenitor proliferation. The process of differentiation requires a reduced cell proliferation [359], therefore, genes involved in downregulating cell proliferation are essential for haematopoiesis daily cell production. *MYB* and *E2F1* are found as upstream effectors for CMP and GMP expansion [360, 361] and both are inactivated by *SPI1* and *CEBP*α [362–364]. For the erythroid lineage, *GATA1* is added to the network and the *JAK2* pathway is extended. Both have been shown to be important players in MEP progenitor production and erythropoiesis [365, 366]. *GATA1* is required for *KLF1* activation, which is a marker of erythroid differentiation [367]. *KLF1* is also a downstream target of phospholyrated *TET2* in erythroid cell lines and *TET2* phosphorylation is induced by *JAK2* [368]. *JAK2* is an important upstream regulator of *GATA1* via *AKT* [369], but also plays an important role in MEP expansion with *STAT5* and *MAPK* activation of *BCL2L1* anti-apoptotic gene [370, 371]. Despite *JAK2* role in driving erythroid differentiation, it can also give to the myeloid lineage an advantage via *STAT3* activation of *SPI1* [372]. Another essential feature of the model is the *SPI1/GATA1* inhibition loop with *GATA1* repressing *SPI1* and *SPI1* repressing *GATA1* expression [373]. This cross talk between two important hematopoietic transcription factors is crucial in the erythroid/myeloid lineage commitment [374]. Finally, *RUNX1* links those hematopoietic genes to the *JAK2/TET2/HOXA9* motif thanks to *RUNX1* activation by *HOXA9*. *RUNX1* upregulation has been associated to *HOXA9* upregulation in early stem and progenitor cells [375, 376] and therefore is an ideal candidate to link the motif with the hematopoietic genetic regulators. *RUNX1* is found in the earliest stages of haematopoiesis and is essential for a fully functional haematopoiesis [377]. I connect this gene to the rest of the network using its downstream targets *SPI1* for the myeloid lineage [346] and *GATA1* for the erythroid lineage [378, 379].

Finally, four cancer genotypes are defined: the wild type (no mutation), the *TET2* single mutant, the *JAK2* single mutant and the double mutant (Table 4.2). The wild type model illustrates haematopoiesis in its healthy state. When the system contains no mutation, all variables and phenotypes are equal to 1. It should be noted that in the network all variables and phenotypes can take values between zero and two, with one being the normal state, zero a reduced expression state and two an increased expression state. The second specification is the *TET2* single mutant. *TET2* mutation results in loss of its function with increased stem cell self-renewal [168], elevated CMP expansion [169] and diminished overall differentiation [157] with a skew towards the granulocyte-monocyte lineage [158]. I include this skew in *TET2* mutant state by increasing the GMP expansion and leaving the MEP expansion at its wild type state. The third specification is the *JAK2* single mutant, where *JAK2* mutation results in *JAK2* constitutive expression. As a consequence, erythroid differentiation and MEP expansion are both increased [388, 150]. Myeloid lineage being advantaged when *JAK2* is overexpressed, GMP expansion is also increased, but not GMP differentiation as erythroid lineage is preferred [173]. A detailed review of single mutant phenotype is given

Upstream	**Interaction**	**Downstream**	**Reference**	**Experiments**
TET2	Activates	*HOXA9*	[351]	*in vitro*/*in vivo* murine
JAK2	Activates	*HOXA9*	[348]	*in vitro* human
HOXA9	Activates	*HOXA9*	[334]	*in vitro* murine and human
HOXA9	Activates	*RUNX1*	[375, 376]	*ex vivo* /*in vivo* murine
RUNX1	Activates	*SPI1*	[380]	*in vitro* human
RUNX1	Activates	*GATA1*	[378, 379]	*in vitro* human/ *in vivo* murine
SPI1	Inhibits	*GATA1*	[373]	*in vitro*/*in vivo* human
GATA1	Inhibits	*SPI1*	[373]	*in vitro*/*in vivo* human
SPI1	Activates	*CEBPα*	[356]	*in vitro* murine and human
CEBPα	Inhibits	*E2F1*	[363]	*in vivo*/ *in vitro* murine
CEBPα	Activates	GMP differentiation	[358]	*in vivo* murine
E2F1	Activates	GMP expansion	[361]	*in vivo* murine
RUNX1	Inhibits	*MYB*	Prediction	
MYB	Activates	CMP expansion	[360]	*in vivo* murine
TET2	Activates	SC self renewal	[168]	*in vitro/vivo* mice/human PDXs
GATA1	Activates	*KLF1*	[367]	*in vitro*/*in vivo* murine
GATA1	Inhibits	*MYC*	[381]	*in vitro* murine
MYC	Activates	MEP expansion	[381]	*in vitro* murine
KLF1	Activates	Erythroid differentiation	[367]	*in vitro/vivo* murine
TET2	Activates	*TET2*p	[368]	*in vitro* human
JAK2	Activates	*TET2*p	[368]	*in vitro* human
*TET2*p	Activates	*KLF1*	[368]	*in vitro* human
JAK2	Activates	*STAT3*	[382, 383]	*in vitro/vivo* murine and human
JAK2	Activates	*STAT5*	[382, 383]	*in vitro/vivo* murine and human
STAT3	Activates	*SPI1*	[372]	*in vitro* murine
STAT5	Activates	*AKT*	[369]	Human PV transplantation (mice)
AKT	Activates	*GATA1*	[369]	Human PV transplantation (mice)
STAT5	Activates	*MAPK*	[370]	*in vivo* murine
MAPK	Activates	*BCL2L1*	[370]	*in vivo* murine
BCL2L1	Activates	MEP expansion	[371]	*in vitro* human
TET2	Activates	MEP expansion	[158]	*in vivo* murine
MAPK	Activates	*JNK1*	[384]	*in vitro* human
JNK1	Activates	*ITCH*	[385]	*in vitro* human
ITCH	Inhibits	NOTCH	[386]	*in vitro* human
NOTCH	Inhibits	GMP expansion	[387]	*in vivo* murine

Table 4.1 **Gene interaction table for *JAK2*/*TET2* BMA model.** PDX: patient derived xenograft.

in section 1.3.3. The final genotype is the double mutant which can lead to one of two fixpoint attractors. Each fixpoint represents either *TET2* first or *JAK2* first double mutants and are defined from results presented in Ortmann et al's paper [325]. Both double mutants share four identical phenotype values and only CMP expansion and erythroid differentiation differ between the two fixed points. The model as shown in Figure 4.5B with the defined target functions (Table 4.3) reproduces the specifications described in Table 4.2 and therefore the branching evolution observed in [325]. The increased differentiation in the *JAK2* first double mutants can partly explain the divergent clinical behaviours between the two groups of patients, including the increased risk of thrombosis and the faster diagnosis as a result of the abnormally high number of differentiated cells in these patients.

	WT	*TET2*	*JAK2*	*TET2* first	*JAK2* first
Stem Cell Self Renewal	1	2	1	2	2
CMP expansion	1	2	1	2	1
GMP expansion	1	2	2	2	2
GMP differentiation	1	0	1	1	1
Erythroid differentiation	1	0	2	1	2
MEP expansion	1	1	2	2	2

Table 4.2 ***JAK2*/*TET2* mutant specification table for BMA model.** In order from the left to the right columns, the specifications are: the wild type state, the *TET2* single mutant, the *JAK2* single mutant and finally the double mutants, which consists of a bifurcation with two state attractors that represent the case where *TET2* is mutated before *JAK2* (*TET2* first) and the alternative case where *JAK2* is mutated first (*JAK2* first). Phenotype values are determined using literature for the single mutants (reviewed in the Introduction chapter) and the Ortmann et al paper [325] for the double mutants. The value 1 represents the healthy state, 0 the lowered/inactive state and 2 the overactive state.

The model identifies new interactions as part of the MPN disease progression. Whilst building single mutant specifications, I discovered a path between *JAK2* and GMP expansion is required to match the increased number of myeloid progenitors observed in *JAK2* first patients. To explore possible downstream pathways of *JAK2* that could contribute to this observation, a machine learning approach (XGBoost - see Methods section 2.5) is applied to AML TCGA data, AML being a relevant and closely related blood disease. AML patients are split for high and low levels of *JAK2* expression. XGBoost then ranks a set of 14 classical cancer pathways on their ability to cluster patients into the right *JAK2* expression level group. Analyses show that *JAK2* is highly correlated with the NOTCH pathway (Fig. 4.6A), which has been found to act as a tumour suppressor in leukaemia due to the great expansion of

GMP cells after loss of NOTCH signalling [387]. From the SHAP scores (see Methods section 2.5.5) of NOTCH genes plotted in Figure 4.6B, *ITCH* is identified as among the genes with the highest score in the NOTCH pathway for *JAK2* expression. *ITCH* controls the degradation of NOTCH [386] and is found to be induced by *JNK1* [385] from the MAPK pathway which is a well-known downstream pathway of *JAK2*/*STAT5* [389, 390, 383]. I therefore hypothesise that *JAK2* path to GMP expansion upregulation could be MAPK and NOTCH dependent.

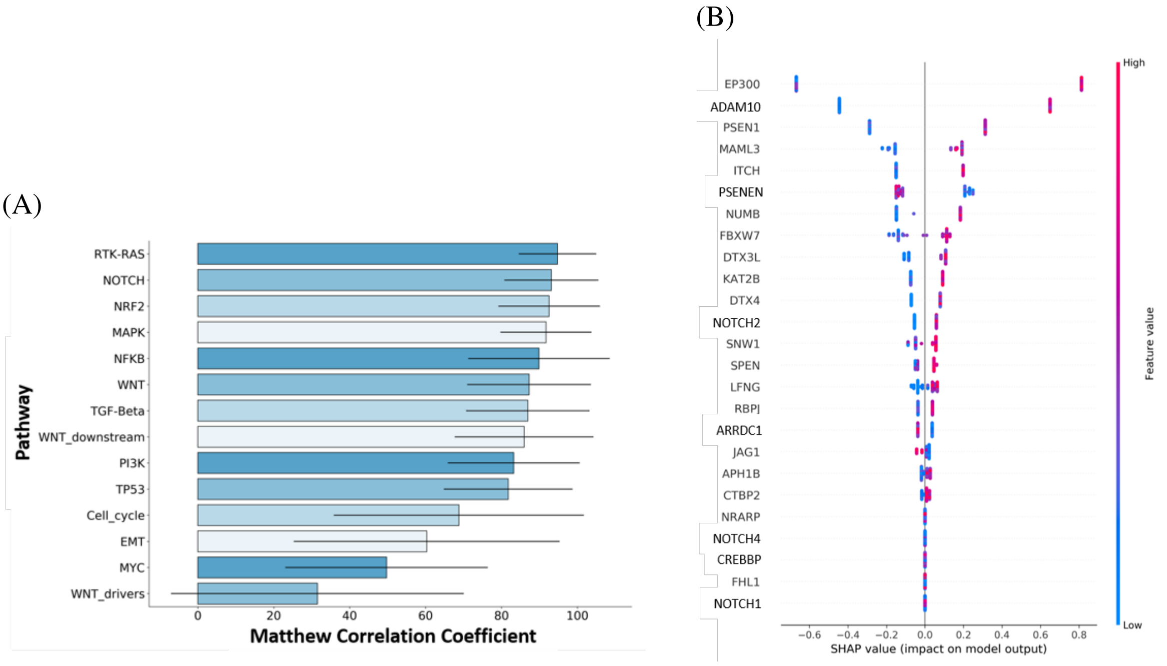


Fig. 4.6 **The MPN network predicts NOTCH could be a downstream regulator of *JAK2* for GMP expansion.** (A) Ranking of major cancer pathways in *JAK2* analyses determine the RTK-RAS pathway as the most correlated to *JAK2*. This pathway contains *JAK2* and so this result is as expected. The second pathway is NOTCH. The overall accuracy of the pathway is computed with the Matthews Correlation Coefficient. (B) SHAP scores are computed for the NOTCH pathway to determine which genes of the pathway have been important classifiers, that is genes with an important expression correlation with *JAK2*. Among them, *PSEN1* and *ADAM10* are involved in the cleavage of NOTCH membrane receptors. *NUMB* activates *ITCH* which degrades NOTCH.

Another new interaction predicted from the network is the inhibition of *RUNX1* on *MYB*. Common myeloid progenitors (CMP) are found to be differentially expanded between *JAK2* and *TET2* first patients in Ortmann et al [325]. The initial model integrates an inhibition interaction between *SPI1* and *MYB*, the model CMP expansion marker, a connection which has been found experimentally [362]. This inhibition and the stable *SPI1* expression in the double mutant states prevented the known bifurcation in CMP expansion in double mutants. Further investigations lead to the hypothesis that the bifurcation could be obtained by replacing *SPI1* by *RUNX1* for the *MYB* inhibition which is supported by different studies. *RUNX1* activates *SPI1* and *GATA1*, and both are found to be inhibitors of *MYB* [362, 391]. Additionally, conditional knockout of *RUNX1* in mice results in enhanced CMP frequencies [392, 393]. All together, these findings suggest that *RUNX1* can be linked to CMP expansion via *MYB* inhibition.

Node	Target Function	References	Comments
HOXA9	JAK2*TET2 + 2*max(0,(HOXA9-1))*max(0,JAK2-1)	[348, 351, 334]	Activation by *JAK2*, *TET2* and itself. Memory property prediction
SPI1	STAT3 + min((RUNX1-1) , (1-GATA1))	[380, 373, 369]	Activation by *STAT3* and *RUNX1* and inhibition loop with *GATA1*. Minimum function to have normal *SPI1* expression when *GATA1* is underexpressed but decreased expression when *GATA1* is overexpressed independently of *RUNX1*.
GATA1	AKT+RUNX1-max(SPI1,1)	[378, 373, 372]	Activation by *AKT* and *RUNX1* and inhibition loop with *SPI1*. Maximum function necessary as loss of *SPI1* does not increase erythroid differentiation.
MEP expansion	min(BCL2L1+TET2 , max(MYC, BCL2L1))	[381, 371, 158, 388]	*TET2* loss reduces erythroid progenitors by skewing toward myeloid lineage. *JAK2* positive regulation of the erythroid lineage is stronger than *TET2* through *MYC* activation.

Table 4.3 **Target functions of the BMA model variables for the MPN network.** Each node of the network has a target function which is a function that determines the level toward which each molecular component moves at the following time step considering the current state of the model (Methods section 2.4.1). Except for the nodes indicated by this table, all nodes have the default target function of the BioModelAnalyzer (BMA) platform which is the difference between the average state of all the variables activating the current node and the average of all the variables inhibiting it. If there is no activating variable, the target function equals the difference between a constant representing the basal activity of the node and the inhibiting variables. The constant is calculated so that in the healthy state (no mutation), all variables equal 1.

4.2.3 Analyses of public MPN datasets validate NOTCH role in MPN as well as *HOXA9* bimodality and prognosis role in blood diseases

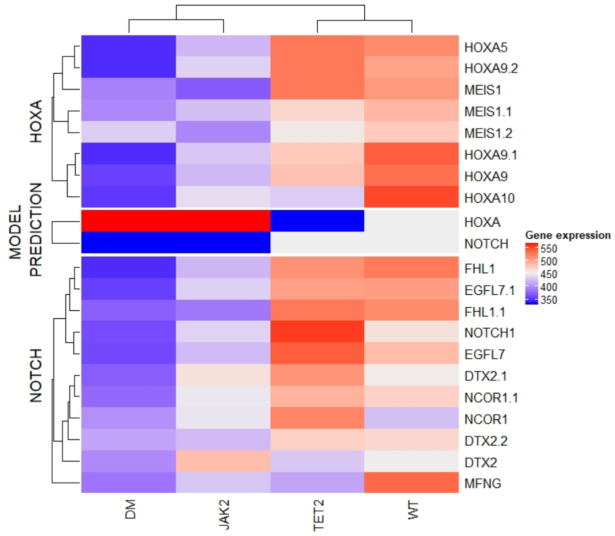


Fig. 4.7 **Model validation using public MPN transcriptomic data.** The heatmap of the NOTCH pathway and HOXA family generated using MPN microarray datasets from [150] validate NOTCH expression in the MPN model highlighting the importance of this pathway in MPN disease progression. HOXA heatmap confirms HOXA bimodality but show different levels of expression to what is found in the model. Heatmap is generated using the ComplexHeatmap R package [394]. For the transcriptomic data, a red/blue colour scale (red: high; blue: low) illustrates gene expression levels normalised with quantile normalisation on raw gene expression expressed in RFU (Relative Fluorescence Units). For the model prediction, blue depicts the value 0, grey the value 1 and red the value 2, which represent decreased, unaltered and increased expression respectively. "JAK2" and "TET2" refer to the single mutant mouse models, and "DM" is the double mutant with *JAK2* mutated first. "WT" designates the wild type (no mutation) genotype.

To validate the predictions arising from the MPN model, the network findings are compared to public MPN data not used for model construction. Chen et al [150] compare MPN with different *JAK2* and *TET2* mutational profiles using transcriptomic mouse data. The authors use microarrays to carry out gene expression profiling of the different mouse genotypes: WT, *JAK2* single mutant, *TET2* single mutant and *JAK2*/*TET2* double mutant (*JAK2* being mutated first). The gene expression of pathways/gene subsets are compared to those included in the network to determine if the MPN model fits their data. Heatmap reveals that the NOTCH pathway behaves as predicted (Fig. 4.7): its expression is lowered in *JAK2* and the double mutant mouse models while WT and *TET2* mutant mice show a higher expression. Further analyses show that the trend in the expression of *RUNX1*, *MYC* and *MYB* support the model

(Fig. 4.8A,B and C). Confusingly *HOXA9* expression displays a “switching” behaviour in this mouse model, but heatmap associates a low expression of *HOXA9* with *JAK2* mutations and high expression with *TET2* mutations.

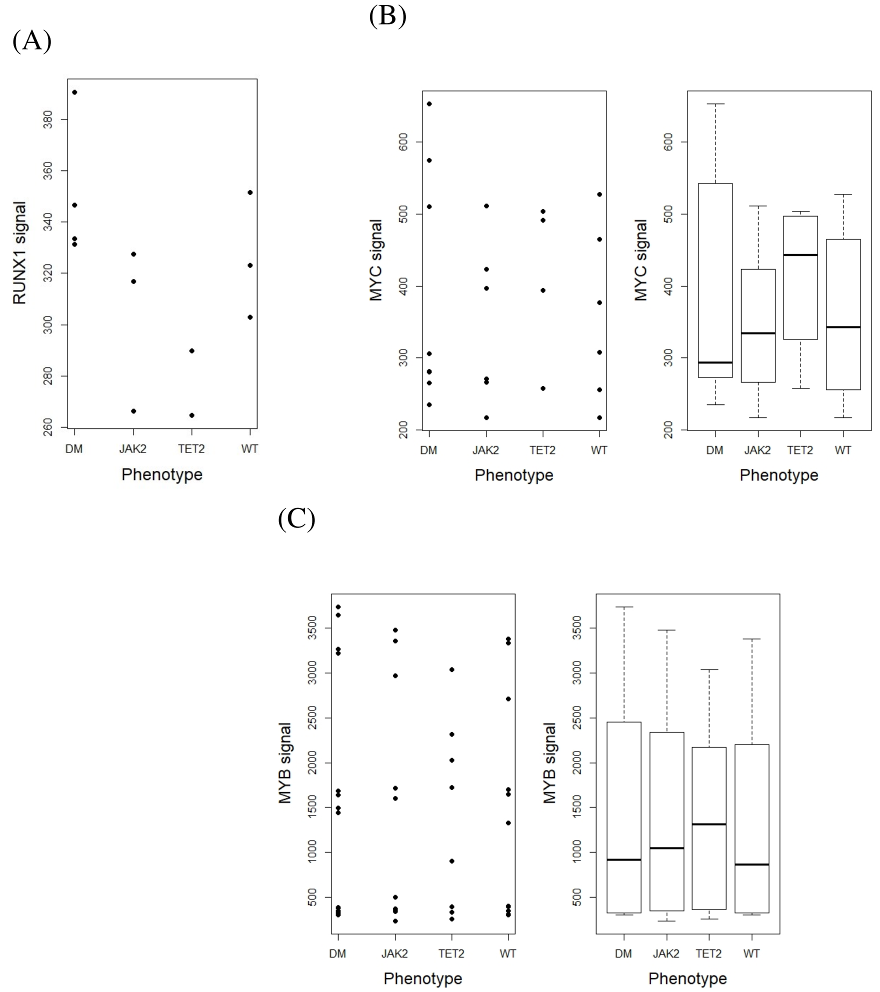


Fig. 4.8 ***RUNX1*****,** ***MYC*** **and** ***MYB*** **expression in public MPN data share the same trend than the MPN model.** Despite the lack of significance, some genes in the microarray experiments support what is observed in the model. (A) The MPN network predicts *RUNX1* to be higher in the *JAK2* single and *JAK2* first double mutants and lower in *TET2* single mutant compared to the wild type. Despite the low number of data points, the trends for *RUNX1* expression in the different phenotypes exactly fit the model findings. (B) The network predicts *MYB* expression to be higher in *TET2* single mutant while *JAK2* single and *JAK2* first double mutants have a similar *MYB* expression compared to the wild type state. Boxplot figures highlight that the trends for *MYB* expression in the different genotypes fit the model predictions. (C) Finally, the model expects *MYC* expression to show similar levels of expression than *MYB* in the different genotypes. Trends for *MYC* expression in the different mutants fit the model results for at least the *TET2* single mutant and the *JAK2* first double mutant.

Jeong et al [368] have shown direct activation of *TET2* by *JAK2* in a combination of in vitro human/murine hematopoietic cell lines with erythroid characteristics, and that in a murine cell line *JAK2* mutation leads to *HOXA9* upregulation. These findings are consistent with the *JAK2*/*TET2*/*HOXA9* motif but are opposed to Chen's microarray experiments where *HOXA9* expression is lowered in *JAK2* single and double mutants (Fig. 4.7). Given the downstream genes follow the expected expression, this raises the question of whether the interactions in the original motif should be replaced by a pair of inhibitions rather than activations if the Chen data is correct. Whilst there exist possible routes to connect *TET2* and *HOXA9* through an inhibition, I am however unable to find evidence of inhibition of *HOXA9* by *JAK2*. It should be further noted that as the Jeong data is human derived, it may be a more representative experimental model system. Moreover, data on individual genes presented in [368] is consistent with the model: *RUNX1*, *ITCH*, *GATA1*, *KLF1* and *BCL2L1* are overexpressed in MPN patients with a *JAK2* mutation in our network as well as in [368]. Both datasets however support the role of *HOXA9* as a switch in MPN.

4.3 Discussion

Out of 6817 genes tested *HOXA9* is the single most highly correlated factor for poor prognosis due to treatment failure in AML [331]. This finding has made *HOXA9* the most studied gene in the HOXA family. The work presented here demonstrates that this gene may act in AML as a discrete switch rather than a spectrum, which impacts AML clinical characteristics such as classification and survival. This study further suggests that the prognosis marker role of the *HOXA9* gene could be extended to another blood disorder, MPN. In MPN diseases with *JAK2*/*TET2* mutations, *HOXA9* high expression is found in the *JAK2* first patients while *TET2* first patients display lower *HOXA9* expression. *JAK2* first patients have a higher risk of developing thrombosis compared to *TET2* first patients. As thrombotic events are the main causes of death in MPN patients [395], this suggests again a deleterious influence of *HOXA9* high expression on patient clinical outcomes in another myeloid disease and emphasises the role of *HOXA9* as a poor prognosis marker in blood malignancies.

This work presents the first molecular network showing and proposing an explanation for the impact of mutation order in a blood disease. In addition to insights into the epigenetic control of cancer cell fate through *HOXA9*, the MPN model recapitulates the disease symptoms using well-known hematopoietic transcription factors such as *GATA1* and *CEBP*α but also the NOTCH pathway. Further investigations of these genes could benefit clinicians by designing new drugs or applying already existing treatments to reduce symptoms and the risk of developing blast phase MPN. In addition to the specific claims of the model, several other clinical implications arise. One key feature of *TET2*-first MPN patients is their reduced sensitivity to Ruxolitinib, a *JAK2* inhibitor drug [325]. It is intriguing to note based on the model that after *TET2* loss, most common *JAK2* targets are unchanged by *JAK2* activation mutation due to the “memory” property exerted by *HOXA9* self-loop. It follows that *JAK2* inhibition is therefore inefficient for those genes. Also, whilst *JAK2* is the main driver mutation found in all MPN patients, different diseases with distinct clinical traits can be observed [146]. Until now, the source of this clinical diversity following *JAK2* mutation was unclear. Here, the network predicts that patients who first had a *TET2* mutation have a reduced number of erythroid cells as a result of *TET2* indirect downregulation of *GATA1* and *KLF1* which explains the reduced number of PV diseases in *TET2* first patients despite the presence of *JAK2* mutation [325]. While *JAK2* dysregulation may be the principal driver of MPNs, other mutations shape the disease clinical type by altering the normal development of distinct hematopoietic subpopulations. Finally, our work assumes the involvement of the NOTCH pathway in MPN diseases. NOTCH shows both oncogenic and tumour suppressor roles in different tissues and in the hematopoietic system: NOTCH favours cancer growth in T acute lymphoblastic leukaemia (T-ALL) through its *MYC* activation but is also found to augment the host immune response against cancer by activation of M1 macrophages [396]. The role of NOTCH in hematopoietic stem and progenitor cells is still an on-going debate, however, it seems that a certain level of NOTCH signalling is required to protect individuals from hematological malignancies [397]. The MPN model suggests that *JAK2* increases GMP expansion through its inhibitory effect on NOTCH via the MAPK pathway and *ITCH* and so predicts a tumour suppressor role for NOTCH in the GMP cell population. This molecular network offers a novel mechanism for understanding how cancer fate can be determined through epigenetic switches, and highlights several new areas for further study.

Chapter 5

APP determines cell fate in blood diseases.

Abstract

Personalised medicine combines patient genome information and clinical characteristics to make precise diagnoses and predict disease outcomes. Patient stratification is therefore essential for choosing the appropriate treatment protocol. I find that AML patients can be clustered into three clinically distinct groups for different *HOXA9* and *APP* gene expression levels. Further investigation of *APP* expression in these patients highlights the important role of this gene in AML patient stratification and survival prognosis. Despite the substantial number of medical studies for *APP* involvement in brain and Alzheimer disease, its role in blood has been overlooked. This work investigates the unexpected poor survival feature of leukemia patients with low *APP* expression using genetic information and computational tools. Data from various leukemia malignancies show reduced myeloid marker expression and increased lymphoid characteristics in low *APP* patients. Similarities between AML patients expressing low *APP* and Mixed Phenotype Acute Leukemia (MPAL) support the potential implication of *APP* in the lymphoid versus myeloid differentiation. I therefore extend with this work the importance of *APP* in other medical disorders and point out its role in cell fate commitment and patient stratification in blood malignancies.

5.1 Introduction

The amyloid precursor protein (APP) is a transmembrane protein expressed in many tissues, but particularly well-characterised in brain studies due to its critical implication in Alzheimer disease (AD). APP undergoes complex processing which generates among other molecular fragments the β-amyloid peptide which is produced in excess in AD patients and might cause neurodegeneration [398]. Two pathways have been described in the literature for APP cleavage by secretases: the amyloidogenic and non-amyloidogenic pathways [399]. First, the secreted extracellular domain of APP can be cleaved by the α- or β-secretase producing the soluble APP-α for the non-amyloidogenic pathway and the soluble APP-β for the amyloidogenic pathway [400]. Further APP cleavage by the α-secretase leads to the generation of the p3 segment in the non-amyloidogenic pathway, the β amyloid ($\beta\alpha$) fragment in the amyloidogenic pathway and the APP intracellular domain (AICD) which is released in the cytosol in both pathways and translocated to the nucleus to regulate the gene expression of many important biological pathways [401].

The precise biology behind AD pathology remains currently uncertain, however, one hypothesis called the amyloid hypothesis implicate the $\beta\alpha$ segment [402]. Once released in the extracellular space, $\beta\alpha$ can aggregate to form oligomers and generate β plaques [403]. Formation of those plaques would cause synaptic and neuritic injury and lead to major neuronal dysfunctions, cell death and transmission deficit [404].

APP expression is not restricted to the brain [405]. *APP* is ubiquitously expressed in human tissues, and its function in skin, intestine and muscle among many other biological systems have been well studied [406]. For example, *APP* is highly expressed in adipose tissues and was shown to be upregulated in obese patients with the development of insulin resistance and adipose tissue inflammation [407]. *APP* plays a central role in various diseases among which many different cancer types [408]. In melanoma, downregulation of *APP* induces proliferation reduction, melanocyte pigmentation/differentiation marker increase, and higher sensitivity to chemotherapy drugs [409]. Studying *APP* biological function in non neuronal tissues is therefore a compelling approach against oncogenesis.

Despite the extensive research on APP, its function in the immune system and blood cancers is relatively poorly characterised although few studies have highlighted a role for *APP* in several hematopoietic populations [410, 411, 406]. For example, Sondag et al [411] illustrate the activation of peripheral monocytic cells by *APP* as well as its adhesion involvement in monocytes to type I collagen. The importance of *APP* in blood malignancies originally arises with two studies investigating acute lymphoblastic leukemia (ALL) and lymphoma [412, 413]. In B-precursor ALL, *APP* is found underexpressed among patients bearing an MLL translocation compared to those who do not [412], while *APP* overexpression characterises Epstein-Barr virus-negative Burkitt's lymphoma [413]. In their work, Baldus et al [414] highlight the role of *APP* in AML. Authors find *APP* as the most overexpressed gene in a subset of AML patients with complex karyotypes compared to patients with normal cytogenetics. These studies collectively establish the critical role of *APP* in hematopoiesis and blood cancers, but the exact biological function of *APP* in blood development and lineage differentiation remains unclear. Therefore, insights into *APP* biological role in blood could establish its precise implication in liquid tumours and aid patient classification.

This study focuses on *APP* expression in different blood malignancies. *APP* is first identified in chapter 4 as differentially expressed in *HOXA9* low and high cohorts. Deeper investigation of the TCGA AML data (data description in subsection 2.2.1) highlights the bimodality of *APP* expression and the presence of patient clusters with clinically divergent characteristics and distinct levels of expression for *HOXA9* and *APP*. Analyses also show that AML patients with opposed *APP* expression exhibit different clinical and genetic characteristics. Notably, I identify a cohort of AML patients with extremely low *APP* expression and poor survival probability. Data from various leukemia malignancies show reduced myeloid marker expression and increased lymphoid characteristics in patients with very low *APP* expression. The potential role of *APP* in myeloid versus lymphocyte differentiation as well as the upregulation of poor prognosis markers in these patients indicate a substantial effect of *APP* reduced expression on blood malignancies. This study therefore highlights the imperative need to investigate *APP* function in blood in order to treat patients with aberrant *APP* expression. Finally, *APP* involvement in Mixed Phenotype Acute Leukemia (MPAL) is also proposed. MPAL being a rare and difficult to diagnose disease, the identification of marker involved in this cancer could fasten diagnosis.

5.2 Results

5.2.1 *HOXA9* and *APP* form patient clusters with distinct clinical features.

APP is first identified as a gene of interest in the work on *HOXA9* in blood malignancies. As shown by Figure 4.3 in chapter 4, *APP* is among the most differentially expressed genes between the high and low *HOXA9* cohorts. Investigation of *APP* expression in AML patients shows that similarly to *HOXA9*, *APP* expression is bimodal (Fig. 5.1A). Further exploration of the TCGA data uncovers the presence of clusters in AML data for *HOXA9* and *APP* expression (Fig. 5.1B). Patients can be classified into three groups with distinct *HOXA9*/*APP* levels of expression: High/High, High/Low and Low/High. The Low/Low cases are ignored due to their high sparsity in the data.

To explore how clinically similar these patients are, I look for the classification distribution of the *HOXA9*/*APP* cohorts. First, histograms show that the French-American-British (FAB) classification of AML is unevenly distributed among patients (Fig. 5.1C). To confirm this observation, a chi-squared test of independence is performed on the data and is found to be significant ($p = 9.5 \times 10^{-13}$). This statistical test examines if it exists a relationship between two categorical variables. As the p-value is significant, the null hypothesis stating no relationship between the variables is rejected and therefore FAB and *HOXA9* cohorts are not independent variables. While the M1, M2 and M4 subtypes do not show any particular pattern, M3 characterises the Low/High group, M0 the High/High group and M5 the High/Low cohort. AML is defined by the uncontrolled growth of the myeloid progenitor cells along with a myeloid-lineage differentiation arrest [130]. However, malignancy can originate from different types of blood cells and stop the maturation at separated stages. The High/High cohort with high expression of both *HOXA9* and *APP* is characterised by M0 which has minimal differentiation compared to other AML subtypes. Strikingly, the Low/High and High/Low patients are discriminated by two opposite subtypes: M3 the Acute Promyelocytic Leukemia have mostly progranulocytic cells and M5 the Acute Monocytic Leukemia possess essentially monocytic blasts. The granulocyte and monocyte lineages seperate quite early in hematopoiesis and both have their own characteristics. Thus, the antagonistic *HOXA9*/*APP* expression levels leading to opposite differentiated blood precursors seem to indicate a role for these genes in myeloid maturation mechanisms.

AML classification includes molecular grouping of patients with distinct fusion or translocation aberrations [415]. Similarly to FAB subtypes, the fusion and translocation distributions show significant unequal repartition in the distinct *HOXA9*/*APP* patient cohorts (chi-squared test, $p = 1.4 \times 10^{-21}$). As expected, the PML-RARα specific to the M3 FAB subtype is solely found in the Low/High patients. The core binding factor (RUNX1-RUNXT1 and CBFB-MYH11) translocations [416] are also particular features of the Low/High patients. Regarding the other cohorts, all patients with complex cytogenetics belong to the High/High cluster. Lastly, the High/Low cohort possesses all NUP98 and MLL translocations. These findings indicate a strong categorisation effect of combined *HOXA9* and *APP* markers which could help AML diagnoses and clinical outcome prediction. This also supports the well-known role of *HOXA9* in blood diseases, as well as the important *APP* role in other diseases than brain disorders.

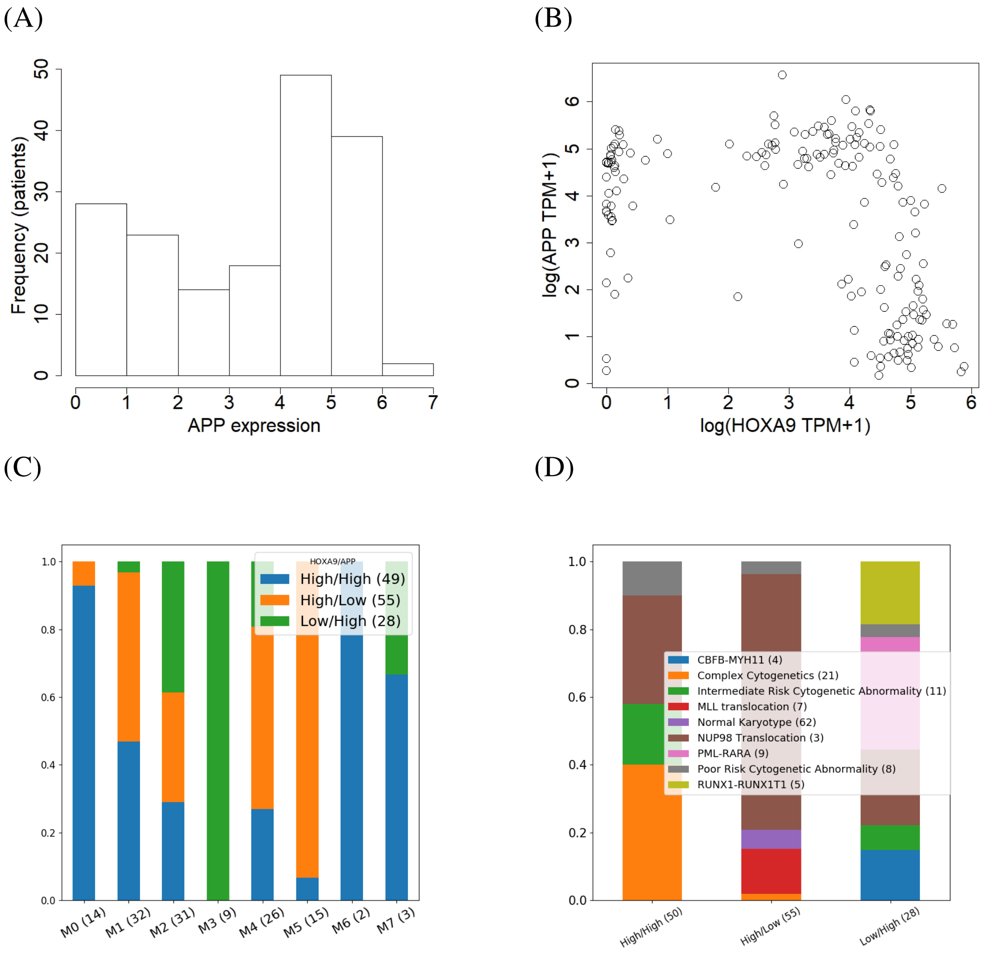


Fig. 5.1 ***HOXA9* and *APP* cluster characteristics in AML patients.** *APP* and *HOXA9* expression stratify patients and their clinical characteristics. (A) Similarly to *HOXA9*, *APP* has a bimodal expression in AML (ACR unimodality test rejected with $p = 0.004$ [336]). (B) Bimodality of both genes generates three different *HOXA9*/*APP* clusters of patients, referred as the High/High, the Low/High and the High/Low cohorts. Due the high sparsity in the patients with low expression for both genes, I ignore these patients in the rest of the analyses. Each cohort show distinct (C) FAB and (D) molecular classifications. (C) The M3 AML subtype also called Acute Promyelocytic Leukemia (APL) is uniquely characterised by low *HOXA9* and high *APP* expressions. Patient diagnosed with M0 (undifferentiated) AML have high expression of *HOXA9* and *APP* while M3 patients have high expression of *HOXA9* but low expression of APP (chi-squared test, $p = 9.5 \times 10^{-13}$). (D) Similarly, molecular classification is stratified among cohorts: CBFB-MYH11 and RUNX1-RUNXT1 translocations, part of the core binding factor (CBF) complex, characterise the Low/High cohort, while complex cytogenetics are observed in the High/High group and NUP98/MLL translocations in the High/Low patients (chi-squared test, $p = 1.4 \times 10^{-21}$).

5.2.2 *APP* stratifies AML patients and is an important prognosis marker.

The observed clinical differences between patients with different levels of *HOXA9* and *APP* suggest that *APP* could be involved in hematopoiesis and liquid malignancy development. To explore the biological functions of *APP* in blood, I extract patients in the two peaks of *APP* expression (Fig. 5.1A). 66 patients constitute the cohort for high *APP* expression, called the "high *APP*" cohort, and 32 for the low peak, referred as the "low *APP*" cohort (Fig. 5.2). Next, I compare survival (Fig. 5.3A), age, WBC, bone marrow blast percentage (Fig. 5.3B), molecular aberrations (Fig. 5.3C) and FAB subtypes (Fig. 5.3D) between both cohorts and find that they exhibit distinct clinical characteristics. Patients in the low peak have worse survival probability ($p = 0.043$) and higher number of white blood cells ($p = 1.9 \times 10^{-6}$). Chi-squared tests find a significant relationship between FAB and *APP* cohorts ($p = 2.8 \times 10^{-3}$) as well as between molecular distribution and *APP* cohorts ($p = 3.3 \times 10^{-6}$). The M5 FAB subtype, NUP98/NLL translocations and normal karyotype are mainly observed in the low cohort, while the M0 FAB subtype, complex and intermediate cytogenetics are specific to the high cohort. These results support the role of *APP* in patient stratification even in absence of *HOXA9*.

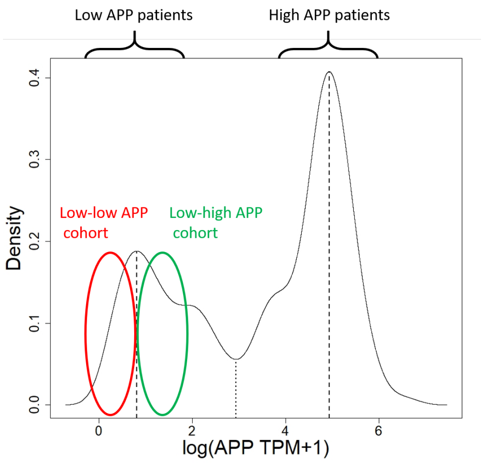


Fig. 5.2 ***APP* cohorts in AML.** The 173 AML patients are separated into groups with different *APP* level of expression. As *APP* is bimodal, the first split distinguishes the low and high peaks and are respectively composed of 32 and 66 patients. The second split divides equally patients in the low peak into two groups of 16 patients.

The poor survival in the low peak patients is unexpected. A recent paper studying the role of *APP* in AML1-ETO-positive AML shows after dividing patients into two cohorts with high and low expression of *APP* that the low *APP* cohort has a better overall survival probability compared to the high cohort [417]. Their finding contradicts the *APP* survival findings in the TCGA AML dataset, however, TCGA data includes all subtypes of AML.

To confirm the poor survival of low *APP* patients, individuals with a reported AML1-ETO translocation are extracted from the TCGA data and their survival between *APP* peaks is compared (Fig. 5.4). Seven patients possess this translocation in the data among which only two have low *APP* expression. However, these two patients seem to have a poorer survival probability compared to the other patients ($p = 0.046$). This is consistent with previous findings.

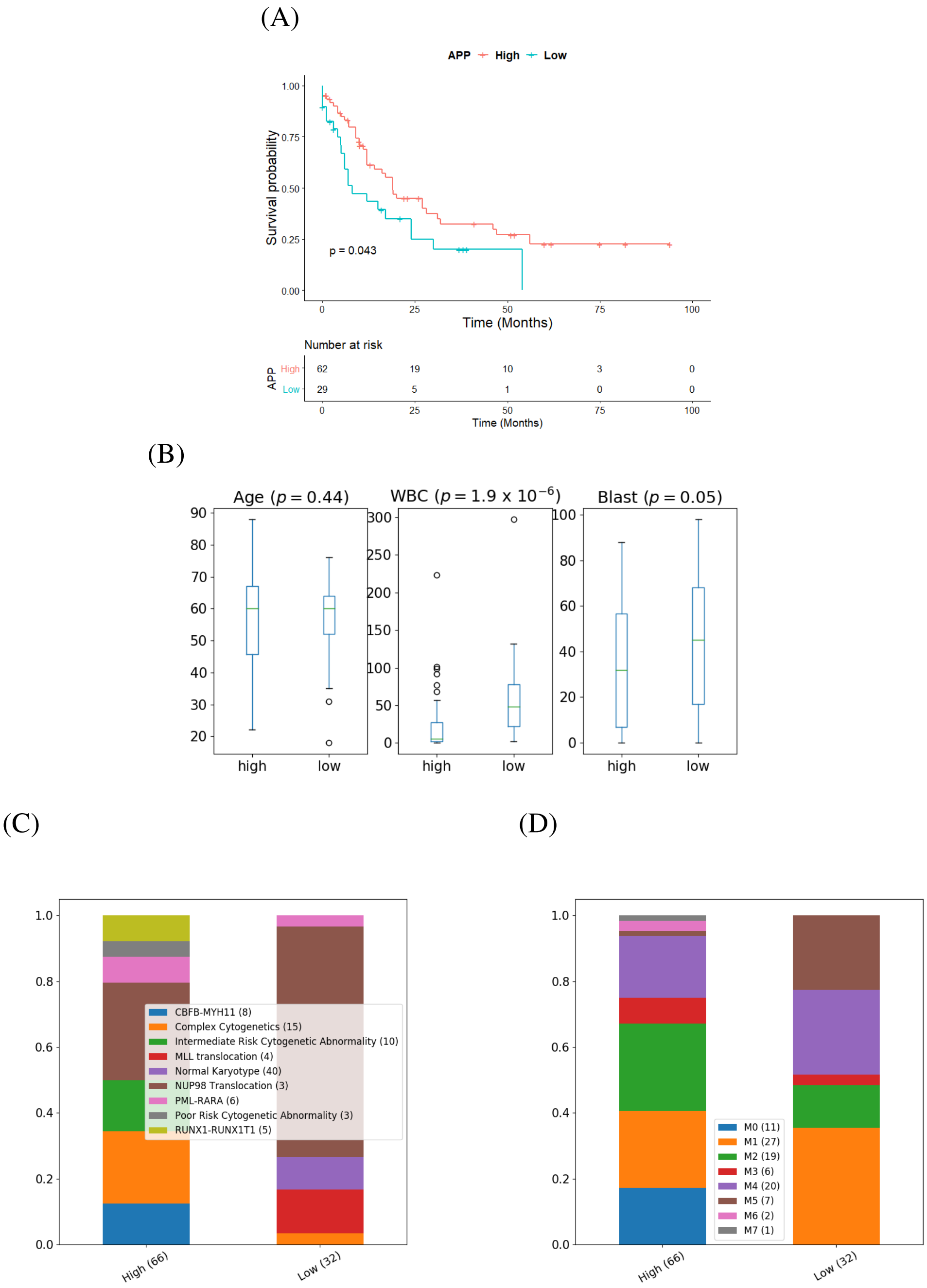


Fig. 5.3 ***APP* peaks stratify patients and their clinical characteristics.** (A) Low *APP* is a poor prognosis marker in AML (log rank test, $p = 0.043$). (B) Low peak *APP* cohort displays a higher number of WBC (Mann–Whitney U test, $p = 1.9 \times 10^{-6}$), but no difference in age or blast number is observed with the high peak. (C) MLL and NUP98 translocations are specific to the low peak patients while the core binding factor (CBF) translocations including RUNX1-RUNXT1 and CBFB-MYH11 are specific to the high peak. Complex and intermediate cytogenetics tend to be mostly present in the high cohort while the normal karyotype repartition is skewed towards the low peak patients (chi-squared test, $p = 3.3 \times 10^{-6}$). (D) M0 FAB subtype characterises high peak patients while M5 is unique to the low peak patients (chi-squared test, $p = 1.9 \times 10^{-6}$).

Next, I explore the clinical features which could explain the poor prognosis of patients within the low *APP* peak. To do so, the 32 patients within the low *APP* peak are split into two new subsets of patients with low and high expression for *APP*. Survival analysis shows that patients within the low peak with a low *APP* expression have poor survival probability ($p = 0.0018$) with a median overall survival time of 6 months against 24 months for the patients within the same peak but with higher *APP* expression (Fig. 5.5). In the rest of the chapter, I refer to the 32 patients the low peak of APP as the "low *APP*" patients as opposed to the "high *APP*" group (Fig. 5.2). I define the 16 patients within the low peak with low expression for *APP* as the "low-low *APP*" cohort and the patients in the low peak with high expression for APP as the "low-high *APP*" patients.

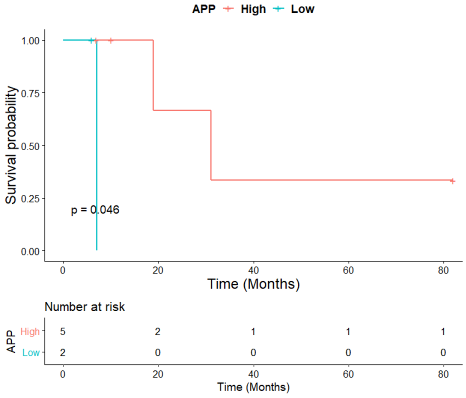


Fig. 5.4 **Survival of AML1-ETO patients within the low and high *APP* peaks.** The patients with AML1-ETO present in the low *APP* peak seem to present poorer survival than the patients in the high peak (log rank test, $p = 0.046$).

The poor prognosis of low-low *APP* patients supports the critical effect of reduced *APP* expression in AML. I compare the available clinical characteristics of low-low and low-high *APP* cohorts to find an explanation. I find that age, sex, WBC, blast number, molecular and FAB classifications are not significantly different between the two groups. In following analyses studying *APP* biological functions in blood, I compare low-low and low-high *APP* cohorts which have similar clinical classifications and ignore the 66 patients in the high peak due to their divergent disease characteristics. By removing this cohort, results are not biased by disease subtypes.

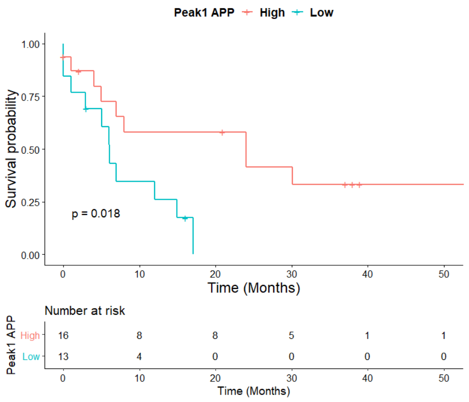


Fig. 5.5 **Survival comparison between low-low and low-high *APP* cohorts.** The two groups of patients in the low *APP* peak have distinct survival probabilities (log rank test, $p = 0.0018$). With an average survival time of 6 months, patients with extreme low *APP* expression succumb to the disease much faster than the high *APP* group.

5.2.3 Differential gene expression analyses suggest a role for *APP* in the lymphoid/myeloid differentiation process.

As none of the clinical features seems to justify the observed poor prognosis of patients with very low *APP* expression, I hypothesise that gene expression could point out dissimilarities between *APP* cohorts. Comparing gene expression levels can indicate which hematopoietic markers are poorly or highly expressed in patient groups with different *APP* expression. These markers can be lineage or prognosis markers. Finding such genes can highlight how *APP* expression modify clinical characteristics. I look at the 30 most differentially expressed genes (DEGs) between low-low and low-high *APP* patients by computing the absolute difference and the fold change between gene expression. DEGs based on the absolute difference includes important markers of the mature B cell population such as *FCGR2C* [418], *LY86/MD-1* [419] and *CD200* [420] which are all upregulated in low-low *APP* (Fig. 5.6). Absolute difference analysis also shows the downregulation of many myeloid genes such as *CPA3* [421], *HBG1/2* [422], *PRNT3* [423] and *EREG* [424]. Finally, high *HGF* [425], STAB1 [426], *KYNU* [427] and *DDIT4* [428] have been found to be poor prognosis markers in AML and are all upregulated in the low-low *APP* group.

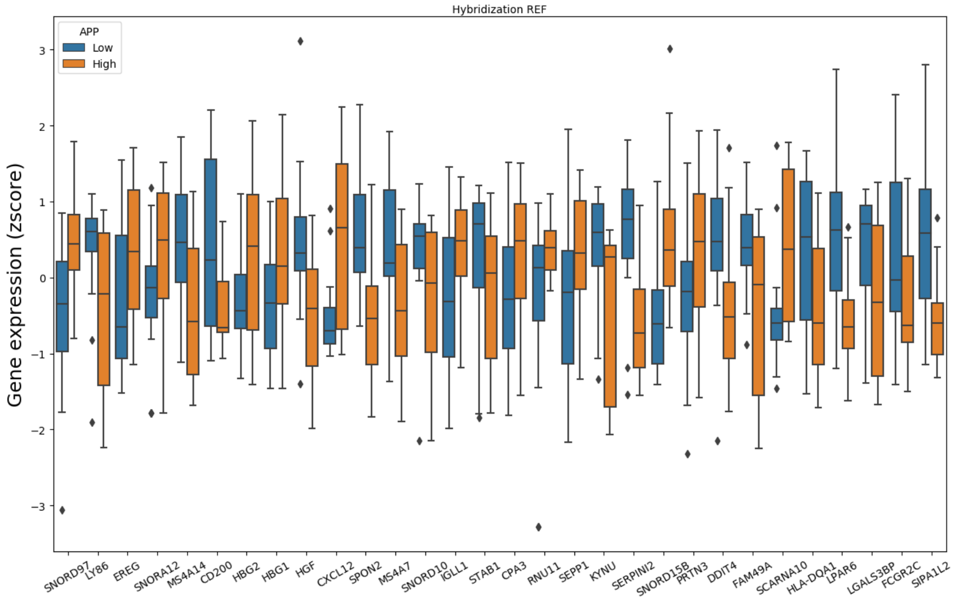


Fig. 5.6 **30 first DEGs between low-low and low-high *APP* patients based on the absolute difference.** DEGs include important regulators of hematopoiesis such as *LY86* and *CD200* for the lymphoid lineage, but also *CPA3* and *HBG1/2* for the myeloid cells. Poor prognosis markers are upregulated in the low-low *APP* cohort: *HGF*, *STAB1*, *KYNU* and *DDIT4*.

DEGs in the fold change analysis also display the upregulation of important B lymphocyte genes such as *CD200* [429], FZD6 [430] and *PAX5* [431] (Fig. 5.7). *FGFR1* found to be involved in Mixed Phenotypic Acute Leukemia (MPAL), a blood malignancy characterised by mixed lymphoid and myeloid traits, is upregulated in low-low *APP* [432]. For the myeloid lineage, downregulation of *PF4* expressed in platelet granules [433] is observed. Finally, increased expression of *BAALC* [434] and *ZNF667* [435], which are both poor prognosis markers in blood malignancies, characterise the low-low *APP* cohort. *BAALC* expression has also been found to be increased in biphenotypic acute leukemia (BAL), a subtype of MPAL, compared to acute T-lymphoblastic leukemia (T-ALL) and AML [436].

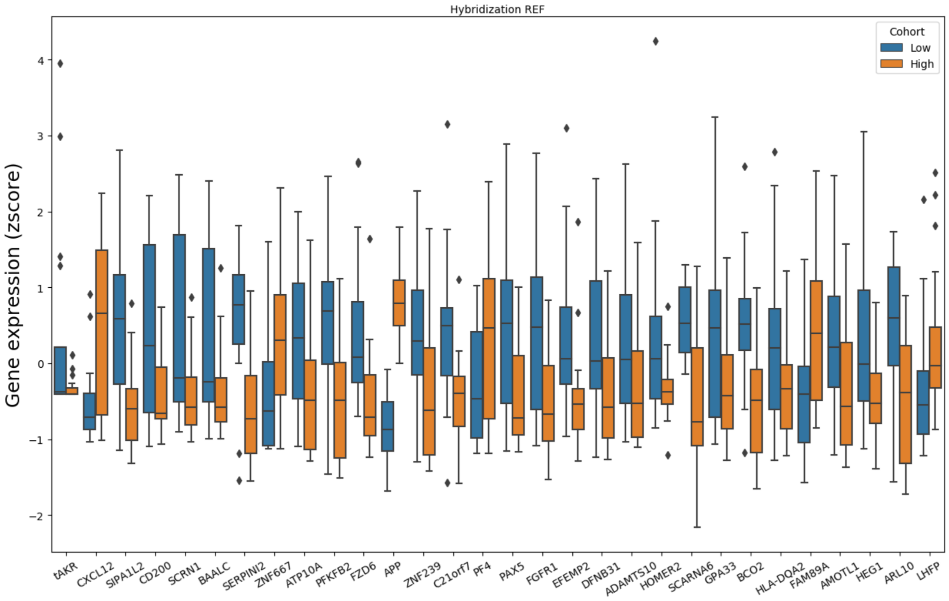


Fig. 5.7 **30 first DEGs between low-low and low-high *APP* patients based on the fold change.** Similarly to the absolute difference analysis, DEGs based on fold change show a pattern with increased B-lymphoid marker (*FZD6*, *PAX5*) and myeloid gene downregulation (*PF4*). Poor prognosis marker (*BAALC*, *ZNF667*) and genes associated with Mixed Phenotypic Acute Leukemia (MPAL) such as *BAALC* and *FGFR1* are upregulated in the low cohort.

Both DEGs analyses highlight the upregulation of genes involved in B lymphocyte development and genes with poor prognosis features, but also the downregulation of myeloid lineage markers. Interestingly, DEGs also include genes associated with MPAL. To further investigate the biological functions in hematopoiesis of identified DEGs, I perform a Gene Set Enrichment Analysis (GSEA) on curated gene sets associated with hematopoietic studies. GSEA is a computational method which identifies gene sets that show statistically different expression among two samples [437]. A large number of gene sets are available, but customised groups of genes can also be added to compare samples. Genes of each set share common features, such as their biological functions for example but also their connection to a particular signalling process and many other possibilities. Users include in the analyses as many sets they desire according to the question they want to answer. Here I include gene sets associated to blood development and lineage markers to identify the hematopoietic features that distinguish both *APP* patient groups. I find that the two first upregulated gene sets for each cohort derive from the same study [438] (Fig. 5.8). In this paper, authors compare genes that are either upregulated or downregulated between a common lymphoid progenitor (CLP) cell and a multipotent progenitor. GSEA results suggest that the low-low *APP* group is associated with genes upregulated in CLP while genes downregulated in CLP are specific to the low-high *APP* cohort. Collectively, differential gene expression studies in AML data support a role for *APP* in the myeloid/lymphoid cell fate commitment, with low *APP* expression skewing cells towards the B lymphocyte lineage.

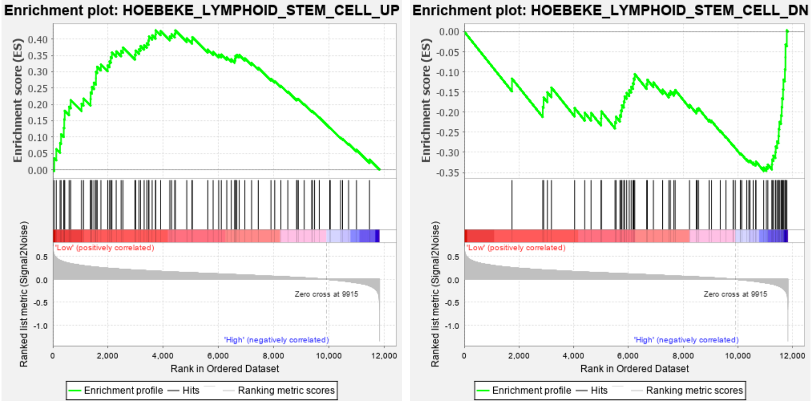


Fig. 5.8 **First GSEA gene sets associated with the *APP* cohorts.** GSEA for the DEGs identified by fold change between the two low-low and low-high *APP* cohorts finds that the two most upregulated gene set for each group is from the same study which study common lymphoid progenitor cells [438]. Low-low *APP* is characterised by the upregulation of genes involved in the lymphoid progenitors while low-high *APP* is defined by genes that are downregulated in those progenitors.

5.2.4 Investigation of *APP* as a biomarker for MPAL.

AML patients with low *APP* expression show common characteristics with MPAL disease.

The lack of studies on *APP* function in the immune system encourages further exploration of the role of *APP* in other blood cancers. In particular, the poor survival as well as the lymphoid skewing of low-low *APP* patients lead to the investigation of papers studying MPAL which is a blood malignancy with poor survival prognosis and the cytochemical and/or immunophenotypic characteristics of both myeloid and lymphoid lineages [439]. Following previous gene expression analyses, I hypothesise that *APP* could be a marker of B-Myeloid MPAL whose patients display high expression of markers of the myeloid and B lymphoid lineages and generally a M1/M5 cell morphology when first diagnosed as AML [439]. More than half of low *APP* peak patients in the TCGA AML data are diagnosed with M1 or M5 AML (Fig. 5.3). This is consistent with our hypothesis.

To assess the potential of the low-low *APP* cohort to be diagnosed as MPAL, I use a combination of lineage markers defined by the European Group for the Immunological Characterization of Leukemias (EGIL) classification [440] as well as gene markers defined in the literature [441]. In the 2008 WHO classification scheme (included in the EGIL classification), three sets of markers are defined for each lineage: myeloid, B and T cells. A patient is diagnosed as biphenotypic acute leukemia if its leukemic blasts express MPO or at least two other monocytic markers for the myeloid lineage, cytoplasmic or surface *CD3* for T cells and finally strong *CD19* as well as strong expression of another B cell marker or weak *CD19* and two other markers for the B lineage. I compare the expression of these markers between the low-low and low-high *APP* cohorts, but cannot find two myeloid markers significantly overexpressed in the low-low *APP* cohort. However, as all cohort patients are initially diagnosed as AML, all should express myeloid markers (Fig. 5.9). For the lymphoid lineage, none of the T lineage markers is significantly upregulated but several B lineage genes such as *CD19* and *CD40* are significantly overexpressed in the low-low *APP* cohort, supporting the initial assumption of B-Myeloid MPAL diagnosis for the low-low *APP* patients. It should be noted that *CD40* is not included in the EGIL criteria, however, its expression in B cells is well known [442]. In addition to the significant upregulation of *CD19* and *CD40*, the higher expression of *CD24* (another important B cell marker) [443] seems sufficient to assume a B lineage preference in low-low *APP* blood cells.

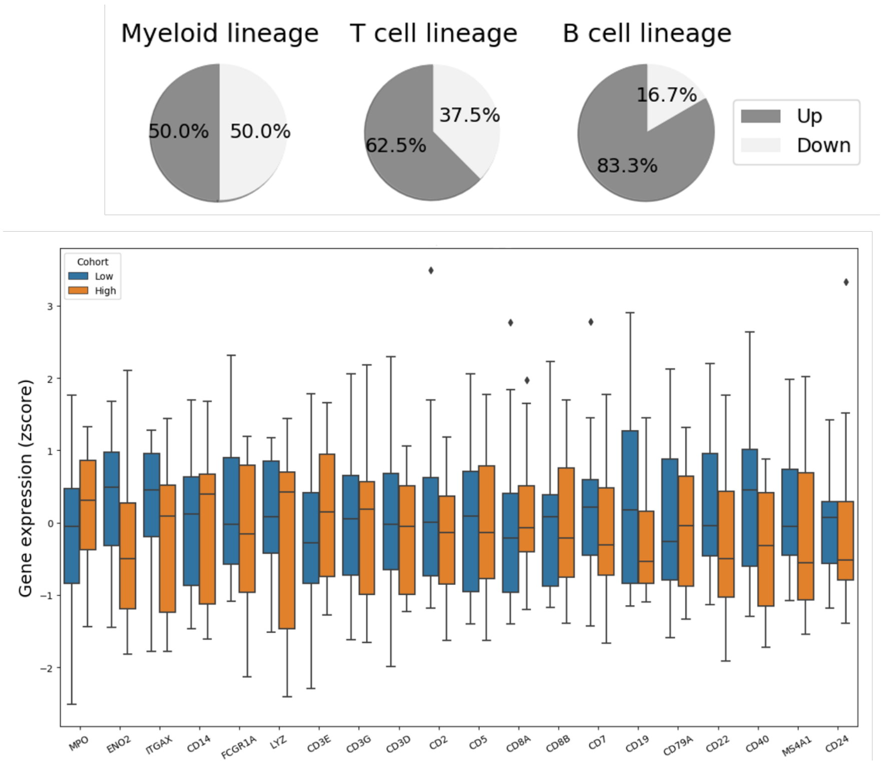


Fig. 5.9 **Analysis highlights the upregulation of B lineage markers in the low-low *APP* cohort.** Comparison of the expression of the myeloid, T and B lineage markers between the low-low and low-high *APP* cohorts shows the significant upregulation of *CD19* and *CD40* for the B lineage in the low-low *APP* patients. Conversely, no marker of the T lineage is significantly increased. Pie charts indicate the proportion of significantly and insignificantly upregulated and downregulated markers for each lineage in the low-low *APP* cohort. Myeloid markers: *MPO*, *ENO2*, *ITGAX*, *CD14*, *FCGR1A* and *LYZ*. T cell markers: *CD3E*, *CD3G*, *CD3D*, *CD2*, *CD5*, *CD8A*, *CD8B* and *CD7*. B cell markers: *CD19*, *CD79A*, *CD22*, *CD40*, *MS4A1* and *CD24*.

Pediatric MPAL data support the poor prognosis role of *APP* in leukemia and its involvement in lymphoid versus myeloid differentiation process.

To test the hypothesis that low *APP* expression could be a B-Myeloid MPAL specificity, I examine *APP* expression in MPAL public datasets. First, I use the RNAseq data from Alexander et al [249] (see section 2.2.4 for data description). To compare *APP* expression between ALAL subtypes, patients are separated into two groups, B-Myeloid versus other subtypes (Fig. 5.10A). *APP* expression is significantly higher in patients with B-Myeloid MPAL compared to other patients. This result clashes with the findings in AML data in which patients with low *APP* expression have increased expression of important B cell markers such as *PAX5* and *CD40*. However, a possible explanation for this discrepancy is that pediatric blood cancers might display different genetic and immunologic profiles compared to adults, which is discussed at the end of this chapter.

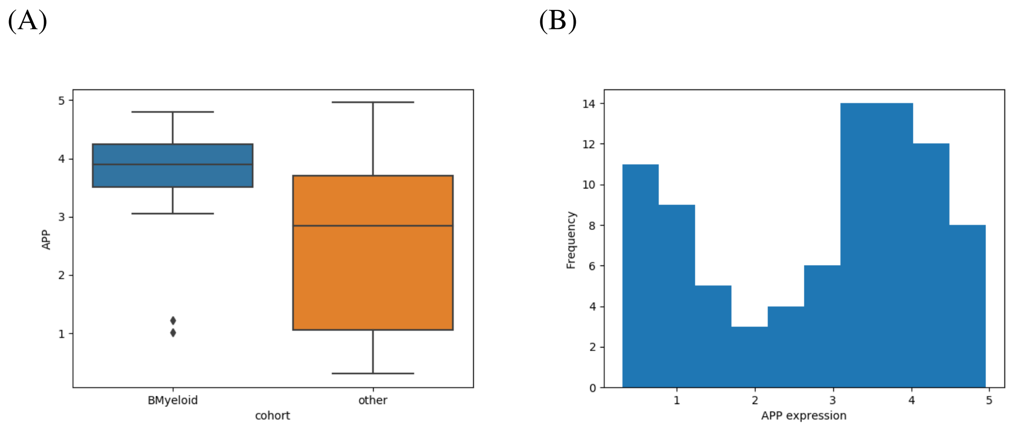


Fig. 5.10 ***APP* expression in pediatric ALAL.** (A) *APP* expression is higher in B-Myeloid MPAL versus other ALAL patients. (B) As observed in AML, *APP* expression in the pediatric ALAL data is bimodal.

Despite the unexpected result of increased *APP* expression in B-Myeloid MPAL patients, I further investigate *APP* expression in pediatric ALAL. I find that *APP* has a bimodal expression as observed in AML patients (Fig. 5.10B). Clinical characteristics between patients in these two peaks are explored. Unlike AML patients, no difference in survival is observed between pediatric ALAL patients in the high and low *APP* peaks (Fig. 5.11A). However, in line with AML, within the low peak, patients with the lowest *APP* expression tend to have worse survival probability despite the insignificant p-value ($p = 0.11$, Fig. 5.11B). Age and WBC are not significantly different between patients in the two peaks. Finally, I explore ALAL subtype distribution between peaks (Fig. 5.11C). As found in AML, low *APP* peak patients are characterised by MLL translocation while high peak patients have the highest proportion of B-Myeloid MPAL which is consistent with Figure 5.10A. Within the low peak patients, B-Myeloid MPAL is exclusively found in the low-high *APP* cohort (Fig. 5.11D).

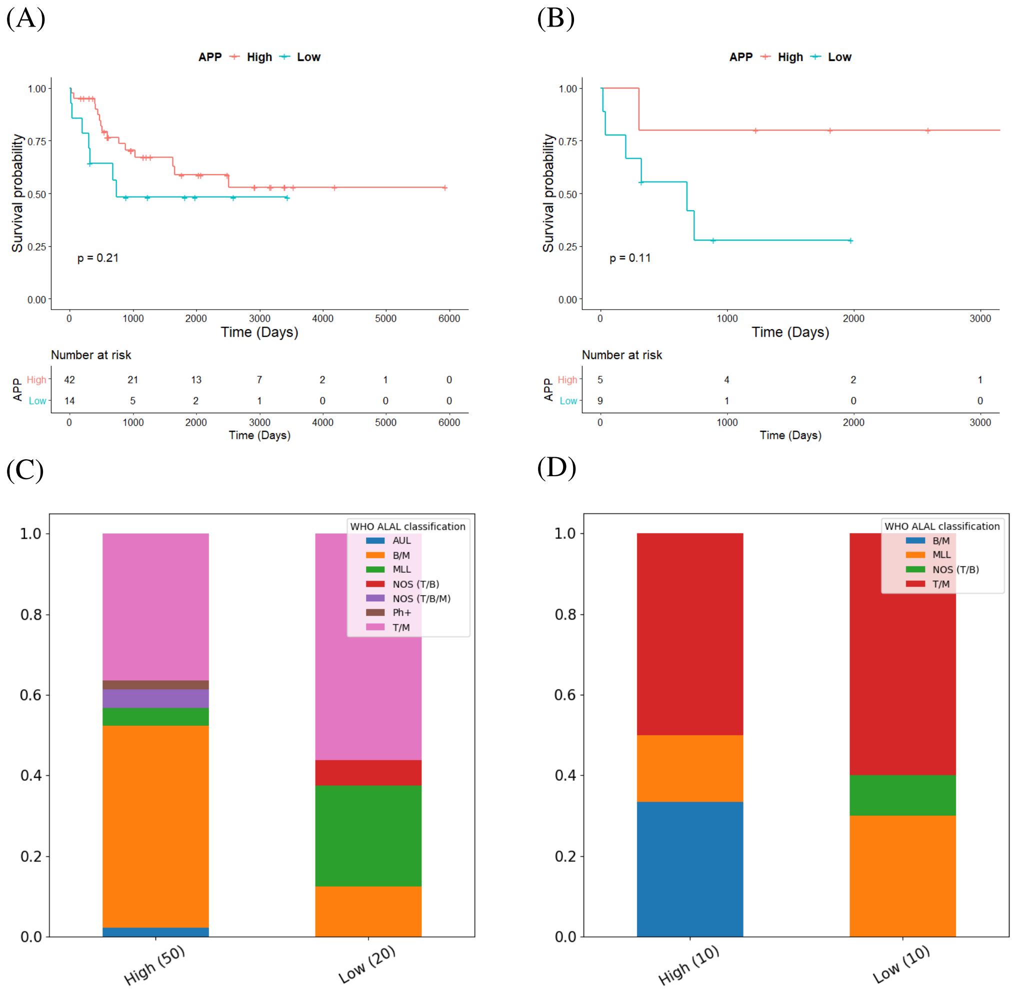


Fig. 5.11 **Pediatric ALAL clinical characteristics for different level of *APP* expression.** Survival between *APP* peaks (A) and within the low peak (B) is not significant ($p = 0.21$ and $p = 0.11$ respectively). Despite insignificance, as found in AML, I observe poorer survival probability for patients with the lowest *APP* expression. ALAL subtype distribution between peaks (C) and within the low peak (D) confirm that high *APP* expression correlates with B-Myeloid subtype in pediatric MPAL. Molecular distribution also shows that MLL translocation is specific to the low peak cohort. AUL: acute undifferentiated leukemia, B/M: B-Myeloid MPAL, MLL: patients with MLL translocation, NOS: not other specified, Ph+: Philadelphia Chromosome positive ALAL, T/B: T-B MPAL, T/B/M: T-B-Myeloid MPAL, T/M: T-Myeloid MPAL.

To further explore the possible roles of *APP* in blood malignancies, gene expression is compared between the high and low *APP* pediatric ALAL patients within the low peak. First, I look at gene markers defined by EGIL for MPAL diagnosis (Fig. 5.12) and find an upregulation of T lineage markers in the low-low *APP* cohort. This result suggests that T lineage is preferred in pediatric MPAL patients with very low *APP* expression and therefore explains the absence of B-Myeloid MPAL in this cohort.

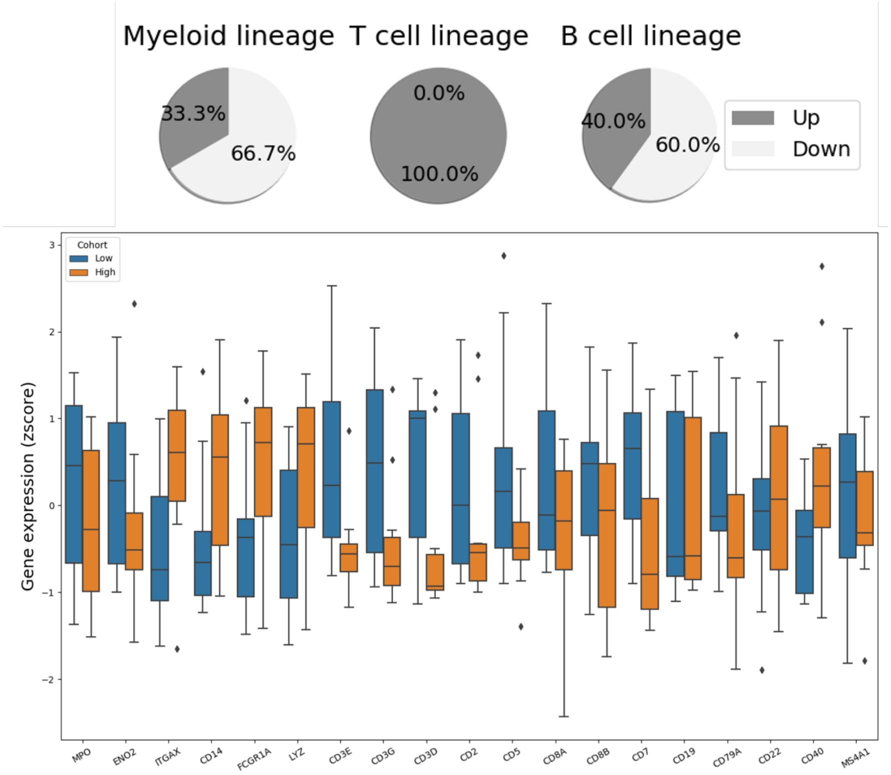


Fig. 5.12 **T cell markers are upregulated in the low-low *APP* cohort in pediatric ALAL.** EGIL genes for the T lineage are all significantly upregulated at the exception of *CD8A* in low-low *APP* patients compared to the low-high *APP* group. The expression patterns for the myeloid and B lineage are more ambiguous. Pie charts indicate the proportion of significantly and insignificantly upregulated and downregulated markers for each lineage in the low-low *APP* cohort. Myeloid markers: *MPO*, *ENO2*, *ITGAX*, *CD14*, *FCGR1A* and *LYZ*. T cell markers: *CD3E*, *CD3G*, *CD3D*, *CD2*, *CD5*, *CD8A*, *CD8B* and *CD7*. B cell markers: *CD19*, *CD79A*, *CD22*, *CD40* and *MS4A1*.

Finally, I search for DEGs between the same two *APP* ALAL cohorts and plot the distribution of the 30 most differentially expressed genes (Fig. 5.13). Genes associated with T cells such as *CD3D* [444], *CD7* [445], *SH2D1A* [446] and *TRBC2* [447] are all upregulated in the low-low *APP* cohort. Upregulation of *PROM1* a stem cell marker [448] is identified in low-low *APP* while genes affiliated with the myeloid lineage such as *CD300E* [449], *IL1B* [450], *MPEG1* [451], and *MAFB* [452] are all downregulated in low-low *APP*. Finally, *BAALC*, a poor prognosis marker in leukemia already identified in AML analyses, is also upregulated in the low-low *APP* cohort. *KLF4*, one of *BAALC* known repressed gene target [453], is downregulated in the same cohort. To conclude on these pediatric

ALAL data analyses, some findings such as upregulation of T cell markers in low-low *APP* patients contradict the observed B lineage skewing in AML patients with low *APP* expression. However, these data support the poor prognosis property of low *APP* expression and its skew towards lymphoid at the expense of the myeloid lineage.

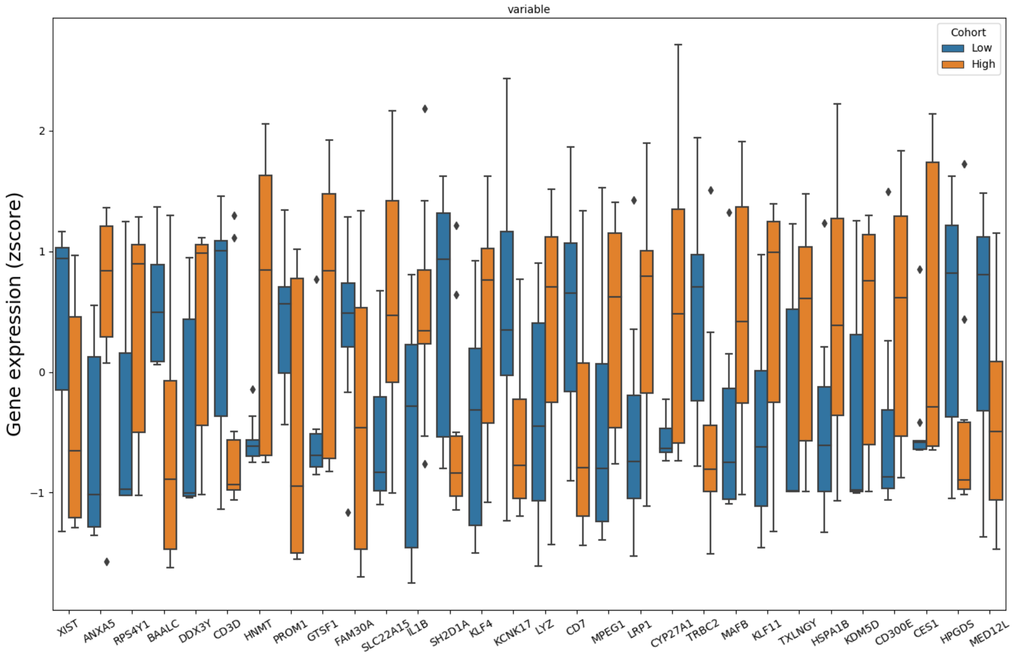


Fig. 5.13 **30 first DEGs between low-low and low-high *APP* patients in pediatric ALAL.** DEGs analysis in the pediatric ALAL patients within the low *APP* peak confirms the repressed myeloid marker expression in patients with poor *APP* expression, but also shows increased T cell markers such as *CD3D*, *CD7* and *TRBC2*.

Adult MPAL data support *APP* involvement in lymphoid versus myeloid differentiation process and suggest that in adults, low *APP* expressing cells are biased towards the B lineage.

To compare *APP* role in adult MPAL against pediatric MPAL and AML, I investigate *APP* expression in a new dataset which contains RNA sequencing information of 24 adult patients diagnosed with B-Myeloid or T-Myeloid MPAL [250] (see section 2.2.5 for data description). In this study, B-Myeloid MPAL patients express the B lineage marker *CD19*, while none of the T-Myeloid cases do. As patient clinical diagnosis and characteristics are missing in the data, the seven patients expressing high *CD19* are defined as B-Myeloid patients. High *PAX5* expression is also found in this group, which supports the B-Myeloid diagnosis. By splitting patients into two cohorts of high and low *CD19* expression, I can compare *APP* expression in B-Myeloid versus other MPAL patients (Fig. 5.14). Though this is not significant, analysis shows that *APP* expression tends towards lower values in the B-Myeloid cohort (high *CD19*) compared to other MPAL ($p = 0.12$). This result differs from the findings in pediatric MPAL where B-Myeloid patients have higher *APP* expression level, however it supports the AML analyses in which B cell markers are increased in the low-low *APP* cohort.

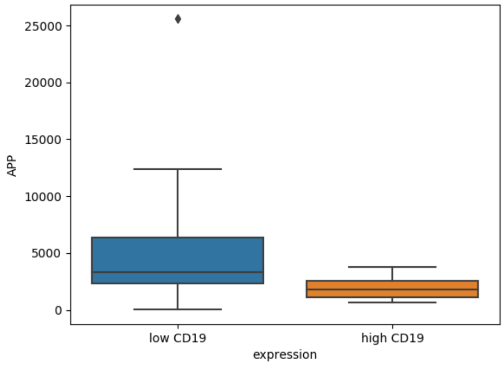


Fig. 5.14 ***APP* expression in adult MPAL with high or low *CD19*.** Reduced *APP* expression is found in high *CD19* cohort which represents the B-Myeloid MPAL patients ($p = 0.12$).

Next, *APP* expression is tested for bimodality in adult MPAL. Unlike to what is observed in AML and pediatric ALAL, *APP* expression is not bimodal in this dataset. To further examine the effect of low *APP* in blood malignancies, patients with the lowest and highest *APP* expression are gathered into two groups of six patients. Comparison of EGIL marker expression levels between these patients show that most myeloid markers (*ITGAX*, *CD14*, *FCGR1A* and *LYZ*) are significantly downregulated in the low *APP* patients, while half of T lineage (*CD3E*, *CD3G*, *CD3D*, *CD2*) and two B lineage (*CD19* and *PAX5*) genes are significantly upregulated (Fig. 5.15). In agreement with AML and pediatric ALAL data, these results suggest that blood cells expressing low *APP* are biased in favour of a lymphoid phenotype at the cost of the myeloid lineage.

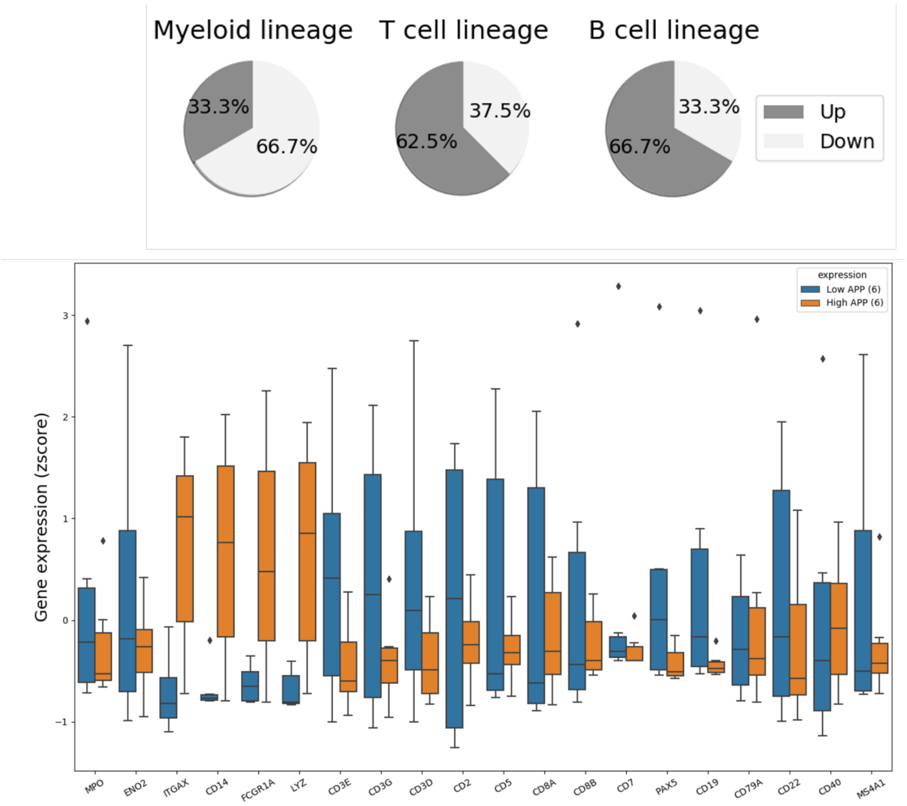


Fig. 5.15 **T and B lineage markers are upregulated in the low *APP* cohort in adult MPAL.** Most myeloid markers (*ITGAX*, *CD14*, *FCGR1A* and *LYZ*) have reduced expression in the low *APP* cohort, while many lymphoid markers are significantly increased from both T and B lymphocyte lineages (*CD3*, *CD2*, *CD19* and *PAX5*). Pie charts indicate the proportion of significantly and insignificantly upregulated and downregulated markers for each lineage in the low *APP* cohort. Myeloid markers: *MPO*, *ENO2*, *ITGAX*, *CD14*, *FCGR1A* and *LYZ*. T cell markers: *CD3E*, *CD3G*, *CD3D*, *CD2*, *CD5*, *CD8A*, *CD8B* and *CD7*. B cell markers: *PAX5*, *CD19*, *CD79A*, *CD22*, *CD40* and *MS4A1*.

Finally, I examine the DEGs between patients with lowest and highest *APP* expression. As observed in previous DEG analyses of other blood malignancies, many genes involved in myeloid development are downregulated in the low *APP* cohort such as *CTSS* [454], *MPEG1* [451], *LYZ* [455], and *CYBB* [456], but I also notice the downregulation of genes associated with T cell lineage such as *AHNAK* [457], *LCP1/L-plastin* [458], *PTPRC/CD45* [459] or *TXNIP/VDUP1*, an essential gene for natural killer cell development [460]. Surprisingly, a subset of genes involved in motility and cell adhesion shows a decreased expression in the low *APP* cohort: *VCAN* [461], *ITGB2/CD18* [462], *RHoA* [463] and *IQGAP1* [464]. Lastly, *MALAT1*, *CDK6* and *NPM1* are also increased in patients with low APP. *MALAT1* has been described as a poor prognosis marker in acute monocytic leukemia [465] and in many other cancer types [466–468]. *CDK6* is a negative regulator of myeloid differentiation [469], while

NPM1 mutations are frequent in AML (between 25-41%) and confer good prognosis if no *FLT3* mutation is present [470]. To summarise, results from adult MPAL validate the role of *APP* in the lymphoid/myeloid differentiation ramification. Furthermore, AML and adult MPAL datasets support that in adults, low *APP* expression skews hematopoiesis development towards the B lineage. This work also suggests that the poor survival of blood malignancies with low *APP* expression could be explained by the upregulation of several poor prognosis genes, but also the mixed and most likely undifferentiated lineage characteristic of tumour cells expressing low APP.

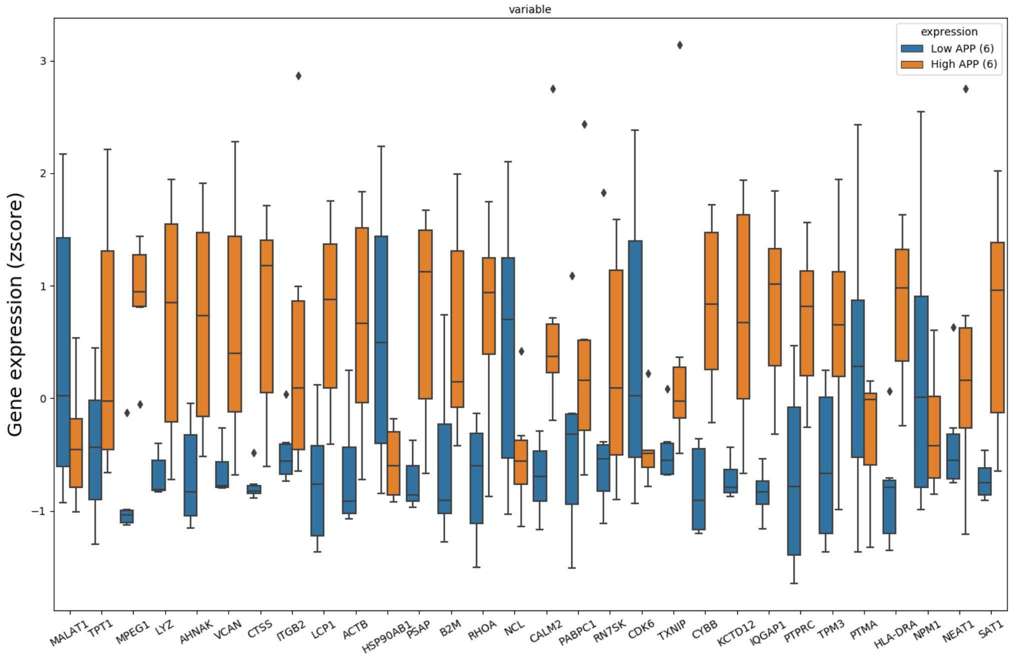


Fig. 5.16 **30 first DEGs between low and high *APP* patients in adult MPAL.** DEGs analysis in the adult MPAL patients with lowest and highest *APP* expression identifies the downregulation of myeloid genes such as *CTSS*, *MPEG1*, *LYZ* and *CYBB*, but also several T lineage genes like *AHNAK*, *LCP1* and *PTPRC*. A subset of genes involved in cell motility and adhesion is also found to have reduced expression in the low *APP* cohort: *VCAN*, *ITGB2*, *RHoA* and *IQGAP1*.

BAALC, a poor prognosis marker in leukemia, is found upregulated in the low *APP* patients of two distinct datasets (Fig. 5.7 and 5.13). I therefore wonder if low *APP* poor survival could be explained by a potential direct or indirect interaction between *APP* and *BAALC*. To investigate this possibility, I search for *BAALC* downstream targets in literature [471, 472] and compare their expression in the *APP* cohorts of all three previously studied leukemia data. The expression of *BAALC* downstream target genes is not significantly different between the *APP* cohorts in any datasets (Fig. 5.17). Thus, it can be assumed that low *APP* poor survival is not linked to *BAALC*.

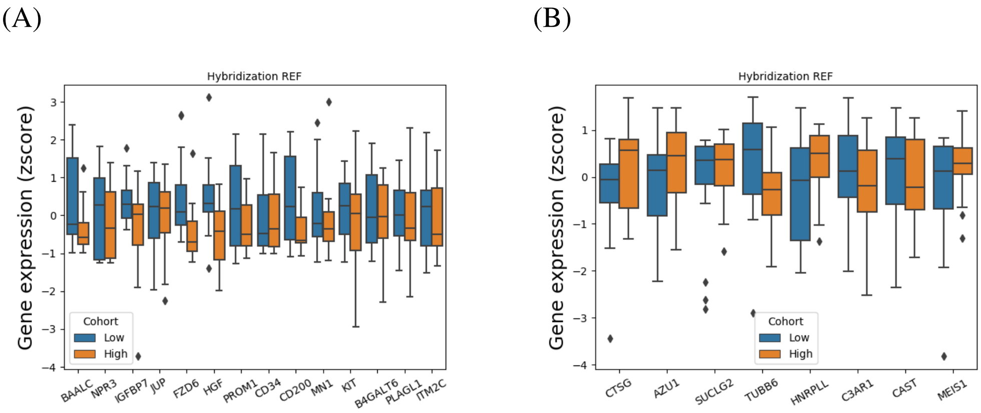


Fig. 5.17 ***APP* is not correlated to *BAALC* downstream target expression.** Expression of *BAALC* upregulated (A) and downregulated (B) downstream genetic targets in the low-low versus low-high *APP* AML patients do not show any significant trend that could validate a potential connection between *BAALC* and *APP* to explain the survival diagnosis of these patients. Similar results are found in the pediatric and adult MPAL datasets.

5.2.5 Exploring the degree of similarity between the low-low *APP* AML cohort and MPAL.

To further explore the possibility of low *APP* expression to be a marker of MPAL, I investigate the degree of similarity between the low-low *APP* AML cohort and MPAL. I assume that if the low-low *APP* AML patients share MPAL characteristics, the number of genes that are differentially expressed between MPAL patients and low-low *APP* AML patients should be lower than against AML patients with higher *APP* expression. Due to format issues, the adult MPAL data [250] cannot be compared to the AML patients. Further analyses therefore use pediatric MPAL sequencing data [249] despite the few dissimilarities already observed.

Gene expression matrices from both diseases are compared by performing a differential gene expression analysis with the R package DESeq2 [252]. This package uses a negative binomial model and gene count data of patient samples with different characteristics to do pairwise differential expression tests. In the following analyses, the MPAL dataset [249] is set as the reference group to compare the number of differentially expressed genes against the three AML subgroups: low-low, low-high and high *APP* patients. In standard DESeq2 analyses, DEGs are selected using the log2 fold change (LFC) and the adjusted p-value ($padj$), that is a transformed p-value accounting for multiple comparison testing. The log2 fold change can be improved by using a LFC shrinkage method that reduce the LFC of genes with low information such as low counts or high dispersion values. For these DEG analyses using DESeq2, I choose the Approximate Posterior Estimation for generalized linear model, apeglm, as the shrinkage estimator [473]. This method uses an adaptive Cauchy prior instead of the default normal distribution, which results in reduced variances for LFC estimates, but also protects true large LFCs.

Moreover, there exists an alternative to the default Wald statistical test used by DESeq2 to evaluate differential expression. While Wald tests if the estimated standard error of a log2 fold change between two conditions is equal to zero, the likelihood ratio test (LRT) identifies genes that show different gene expression across multiple factors. LRT compare two models based on the gene expression count matrix, a full and a reduced model. The full model contains all the parameters to explain the counts while some have been removed in the reduced model. LRT tests if the increased likelihood of the full model with the extra terms is significant and therefore if the removed terms are necessary to explain the data. LRT in combination with the R package DEGreport [474] and its `degPattern` function can determine clusters of genes with similar patterns. I therefore choose this test to find and analyse DEGs between our datasets (Figure 5.18 and 5.19).

To select the appropriate parameter values to filter differentially expressed genes (DEGs), volcano plots are used to establish the $padj$ and LFC that allow a decent number of DEGs between MPAL and AML groups (Fig. 5.18A,B,C). I first set $padj < e^{-80}$ and $LFC > 0.58$ and plot the Venn diagram (Fig. 5.18D). However, the number of specific DEGs between the low peak cohorts and MPAL patients is almost null. To further distinguish low peak subgroups with MPAL, DEG filtering is reduced by defining $padj < e^{-20}$. After this modification, analyses show that the low-low *APP* cohort have the lowest (50) number of DEGs against MPAL, while high peak patients obtain the highest (1207) number (Fig. 5.19A). These findings indicate a greater similarity between the low-low AML cohort and MPAL patients than between the other AML and MPAL patients.

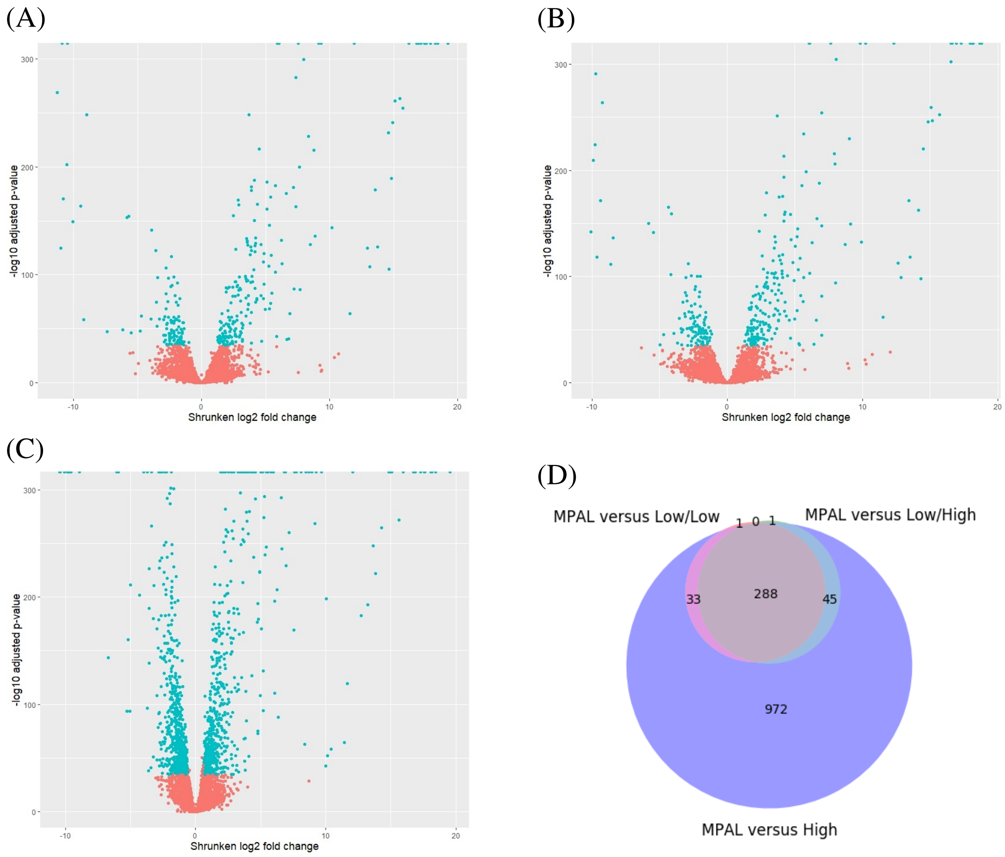


Fig. 5.18 **Volcano plots (A,B,C) and Venn Diagram (D) of DEGs between AML cohorts and MPAL.** Representation of DEGs with volcano plots helps to choose $padj$ and LFC values to filter DEGs between MPAL and the three AML cohorts: (A) Low-Low, (B) Low-High and (C) High APP cohorts. Here, $padj < e^{-80}$ and $LFC > 0.58$. The resulting Venn Diagram (D) identifies a high number of DEGs for the high AML cohort against MPAL.

An elevated number of genes are differentially expressed between MPAL and all AML patients. This finding could be explained by different experimental conditions between both studies as well as adult and children distinct biology. To test this hypothesis and link these genes to biological functions, the `degPattern` function is carried out to identify groups of genes with similar gene change patterns across cohorts. To do so, samples are first gathered into their condition groups and the mean of each gene is calculated. Then `degPattern` creates a distance matrix from all pair-wise gene expression correlation. DIANA (DIvisive ANAlysis) [475], a hierarchical clustering algorithm, builds a hierarchical tree which is cut by `degPattern` to generate groups of genes with similar expression profiles. For the 1384 DEGs shared between MPAL and AML, the analysis detects two clusters of genes: the first cluster with 680 genes consists of genes upregulated in AML and the second cluster with 704 DEGs upregulated in MPAL (Fig. 5.19B). Using the clusterProfiler R package [476], a functional analysis of these two clusters is performed (Fig. 5.19C, D). Except for the neutrophil pathways found in genes upregulated in AML, ClusterProfiler [476] identifies in both gene cohorts biological pathways correlated to general cell functions such as transcription and protein modifications but unrelated to blood malignancies. This finding supports that genes differentially expressed between the AML groups and MPAL can be explained by different experiment protocols or by distinct cellular processes adult and children possess.

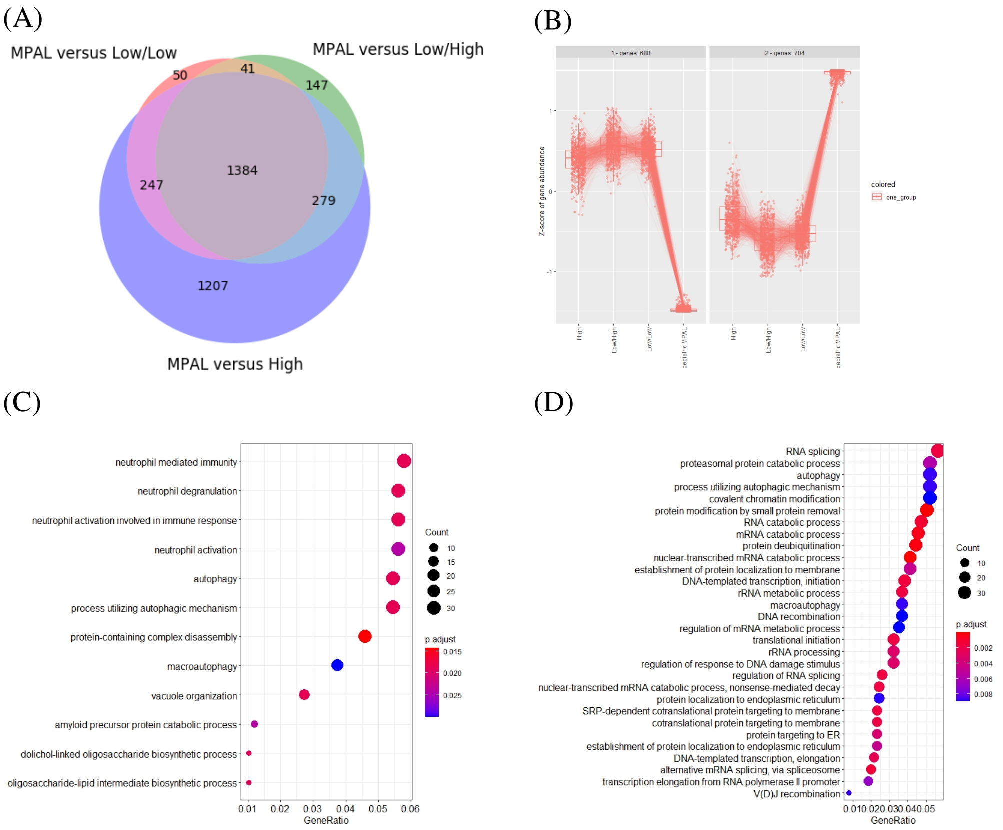


Fig. 5.19 **Low-low *APP* AML patients have more genetic similarity with MPAL patients compared to the two other AML groups.** (A) The Venn diagram of DEGs among MPAL and AML cohorts for $padj < e^{-20}$ and $LFC > 0.58$ identifies more DEGs in the low-high (147) and high (1207) AML groups than in the low-low (50) group. (B) The degPattern function finds two DEG expression patterns for the 1384 DEGs that are shared between AML and MPAL patients. Gene set 1 defines all genes upregulated in AML versus MPAL while group 2 the downregulated genes. A gene ontology (GO) analysis of both groups is performed. (C) The GO specific to the list of genes upregulated in AML and downregulated in pediatric MPAL contain neutrophil characteristics, but also autophagy and glycan biosynthesis. (D) For the genes upregulated in MPAL and downregulated in AML, the GO analysis identifies many pathways correlated to transcription and translation.

5.3 Discussion

In this chapter, three clusters of AML patients with distinct expression levels of *HOXA9* and *APP* are identified. Each group shows distinct clinical characteristics supporting a potential role for *APP* in blood development and malignancies. Further exploration of the AML data leads to the unexpected poor survival feature of patients with very low *APP* expression. These patients also present upregulated expression of B cell lineage markers such as *PAX5* and *CD40*. Upregulation of lymphoid markers and downregulation of myeloid markers in patients with low *APP* expression are found in public data of two other leukemia malignancies. Collectively, these findings in distinct liquid cancers suggest a function for *APP* in early blood cell fate commitment. As a result of the poor survival of AML patients with low *APP* expression and its potential role in lymphoid versus myeloid differentiation, I wonder if *APP* could be a marker of MPAL, a blood disease with lymphoid and myeloid characteristics and poor survival prognosis. To explore this possibility, I compare genetic expression between MPAL and three AML cohorts with different levels of *APP*. The lowest number of differentially expressed genes is found between the AML patients with very low *APP* expression and MPAL. Hence, this study is the first to propose that low *APP* could be a characteristic in MPAL patients explained by its involvement in early blood cell differentiation.

In the scientific literature, the role of *APP* in the lymphoid lineage is poorly characterised. Despite the lack of recent research on this subject, several papers between 1990 and 1996 report the important function of *APP* in white blood cells and more particularly in T cells [477–480]. Monning et al [477, 478] suggest that *APP* is secreted by T cells and plays a role in the immune system activation. In another paper, authors show that granulocytes do not express any level of *APP* unlike monocytic, T and B cells and speculate *APP* as a cell surface receptors in immune cells [480]. To my knowledge, no further study looked at *APP* role in the lymphocyte lineage until a paper about the alteration of T cell development in autism was published in 2012 [481]. As autism patients express elevated secreted soluble *APP* α (sAPPα) and have aberrant T cell development, Bailey et al [481] conceive a transgenic mouse overexpressing sAPPα to observe changes in the immune system. They find increased levels of cytokines involved in T cell activation, but also an elevated number of CD8+ T cells at the expense of B and CD4+ T cells compared to control mice. This study further supports the AML analyses presented here in which low *APP* expression correlates with higher expression of B lineage markers and lower expression of T cell gene markers. Generally, the role of *APP* in blood cell fate commitment suggests that this gene could be a marker for leukemia classification and prognosis.

In this chapter, divergent immune phenotypes typify low *APP* expressing blood cells from distinct liquid cancers. While adult AML and MPAL show an upregulation of B cell markers, pediatric ALAL is characterised by T cell marker increase. Patient age difference could explain these findings. Abnormal processing of APP and accumulation of $\beta\alpha$ plaques are associated to dementia in the elderly [482]. Additionally, APP processing has been shown to be downregulated during aging of normal human fibroblasts as a result of the modified expression of secretases involved in APP cleavage [483]. Another study demonstrates that at older ages, the soluble APPα neuroprotective role is blocked by $\beta\alpha$ oligomers [484]. Moreover, even without any change in *APP* gene expression, age-related changes in APP protein processing in neurons can affect the cell phenotype as observed by Burrinha et al [485]. Authors show that despite the unchanged cellular levels of APP, accumulation of $\beta\alpha$ in aged brains is partly explained by increased APP endocytosis. Endosomes are the principal site of interaction between APP and its secretases which perform APP processing. Therefore, even if no changes are observed in *APP* expression between adult and pediatric blood cancers, *APP* might be diversely processed and have modified functions in different age groups.

Many biological mechanisms are differentially processed between adults and children [486]. In the immune system, cytokine production is altered in healthy children compared to adults, which explains their higher susceptibility to infections [487]. A recent study have explored the immune system and more particularly the different lymphocyte subsets of five groups of individuals clustered by age from infants to elderly [488]. In this paper, authors examine the dynamics of the T, B and NK cell populations in the different cohorts and show that except for B cells which have a negative correlation, all other lymphocyte numbers are positively correlated with age. B cell overall decline with age is consistent with another study [489]. Authors in this paper demonstrate a reduced number of naive B population in growing populations with a stable count of memory B lymphocytes in adulthood. Therefore, age is undoubtedly an important parameter to consider when comparing *APP* expression in pediatric against adult blood diseases. Fluctuations in hematopoietic populations as well as distinct APP processing could differentially impact results. In particular, B and T cells seem to demonstrate opposite trend with ageing, supporting the T versus B cell skew in low-low *APP* cohort in pediatric versus adult MPAL datasets. APP being differentially processed with age further investigations on its potential involvement in lymphocyte differentiation could explain the B cell decrease with ageing.

DEGs analysis in adult MPAL data from [250] highlights the reduced expression of several genes involved in cell motility and cell adhesion in the low *APP* cohort. *APP* increases cell migration in several cancers. For example, *APP* knockout in keratinocytes leads to reduced migration velocity as a result of damaged cell substrate adhesion [490]. This study is therefore consistent with the DEGs analysis. However, reduced migration should be a good prognosis property, which is not what is observed in the low *APP* cohorts of AML and pediatric ALAL patients. A possible explanation is that in liquid cancers, increased motility does not give a fitness advantage to cells. Leukemic cells already circulate in blood and are therefore less affected by changes in adhesion or motility cell functions.

Chapter 6

Boolean motif inference applied to HOXA9 and APP in AML.

Abstract

Understanding molecular evolution in tumours confers an asset for predicting cancer progression and clinical outcomes. Knowledge about regulatory dynamics includes insights into DNA sequence expression changes and how it affects other genes. As shown in chapter 4, these gene interactions are decisive for future disease characteristics and different dynamics can considerably alter prognosis. In chapter 5, I identified in AML patients three clusters with distinct levels of *HOXA9* and *APP* expression. These patients show distinct clinical characteristics, however, the regulatory dynamics underlying these features are unclear. To uncover the genes involved in AML patient stratification, I build a program inferring a list of Boolean motifs reproducing biological system dynamics from input observations. The program returns several motifs, thanks to which 12 gene candidates are identified as potentially involved in the AML clusters. Each gene is computationally tested against AML data to evaluate its probability as a good marker. Collectively, this motif inference program highlights the importance of molecular evolution in blood disease diagnosis and therefore could be applied to other systems to explain diverse complex biological mechanisms.

6.1 Introduction

Signalling pathways consist of a combination of molecular interactions that results in the modification of a cell phenotype or function, such as its growth or DNA repair ability. Signalling often starts with an input signal from the cell environment, such as an hormone or growth factor. This ligand then forms a complex with a receptor, which once activated, will send a signal to another protein inside the cell. The number of interactions and proteins involved varies from one pathway to another. This leads to a cellular response to the external stimulus. Cell responses are broad and vary with gene expression changes such as post-translational modifications and metabolism alterations. A small disruption in this series of events can result in aberrant cell functions and cell death [491]. To better understand how these disruptions affect the cells, signalling pathways can be modelled with networks [492].

A network motif is a recurring subgraph which repeats itself in complex networks and which is defined by a particular pattern of interconnections between nodes [493]. Motifs can be seperated into two broad classes which explain cell signalling with different network properties. The first class is the sequential interactions of responsive elements, while the second consists of biological switches with a feedback process (Fig. 6.1). The first type of signalling involves transcription factors and response elements and can be illustrated as a sequential linear series of molecules interacting one after another (Fig. 6.1A). In this type of motif, the outcome phenotype is entirely dependent on the input signal. Varying input values directly modify the cell phenotype. Such linearity also contributes to modified cell function when either component of the pathway is mutated [494, 495]. The Notch pathway involved in many cellular functions and cancer types has been described as a linear signalling pathway [496].

In contrast to linear signalling, the second type of motifs including switches can retain the history of the previous states. This memory property emerges from the presence of a feedback loop as seen with *HOXA9* in chapter 4. As illustrated in Figure 6.1B, the molecular motif once activated by its environment stays in this state even after removal of the upstream activation due to the feedback loop. This type of signalling is important for cell fate commitment. A well-known example of switch signalling is the maturation of *xenopous* oocytes [344]. Immature oocytes commit to maturation after hormone stimulation and stay in the mature state several days after hormone removal. This irreversible commitment is induced by protein kinase activation that forms a positive feedback loop and therefore all kinases in this loop stay highly expressed even after hormone removal. Perturbation of any genes involved in this feedback loop abrogates the irreversible property of the maturation. Feedback loops are certainly at the origin of many important cellular signalling regulation. Negative feedback loops for example exert a hold on temporal expression of certain pathways once a certain threshold has been reached while positive feedback loops can amplify and prolong a signal. Finally, a combination of both could explain the complex process of pattern formation [497]. Inclusion of these feedback loops in executable biological models is essential to reproduce experiments and untangle complex biological mechanisms [342]. Hence, feedback loops are important in cell signalling as they provide an intermediate control of the pathway regulation in addition to the input signal that would not be sufficient on its own for cell fate commitment and complex biological mechanisms.

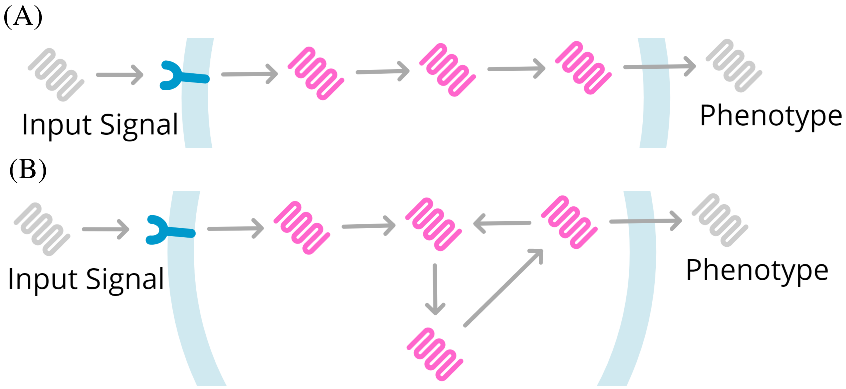


Fig. 6.1 **Examples of linear (A) and switch (B) signalling pathways.** Both pathways are initiated by an input signal from outside the cell and result in the modification of the cell phenotype. However, in the linear pathway (A) the outcome phenotype is directly dependent on the cell environment while the positive feedback loop in (B) stores events.

Linking gene expression and phenotype is challenging as a result of the considerable number of possible protein interactions and feedback loops. Which proteins and pathways modify a cell trait remains unclear. Proteins can have several functions which might be cell-type or tissue specific, and even vary among different organisms. Similarly, protein interactions depend on the biological context. These biological regulatory systems can be represented as networks of regulatory interactions between molecules and many modelling approaches are available to study, analyse and interpret them [498]. A common approach consists in using continuous models based on ordinary differential equations (ODEs) which are defined by continuous and quantitative variables over a continuous timescale. ODEs for biological systems use biochemical kinetic reaction equations describing how the concentration of molecular elements evolve over time. They are powerful models thanks to the rich diversity of biological details they can represent [499]. However, a major drawback of ODE models is their limited application to small networks to avoid high computational cost [500]. Alternatively, Boolean networks offer a promising approach to model large complex systems. First introduced in 1969 by Kauffman [501], Boolean networks describe the evolution of discrete variables with binary states. Variables can be seen as nodes which represent any biological molecules or phenotypes. Interactions between molecules and phenotypes are illustrated by edges and Boolean functions. These Boolean functions are rules written in formal logic connecting different nodes and representing their relationship, such as activation or repression of a gene by one or several other genes. The appeal of this modelling technique is the abstraction of the system. Boolean networks can be applied to large signalling networks and are relatively easy to build and interpret due to the structure of the model [500, 499].

Boolean networks have proved their value in many cancer studies. Among them, Fumia et al [502] demonstrate thanks to a network of 96 nodes and 249 edges representing the main cancer pathways that their model reproduces coherent healthy cell response to different environments such as hypoxia or DNA damage, but also identifies well-known mutational sequences leading to tumour malignancy. Importantly, authors show the emergence of resistant clones after therapy targeting only one cancer cell phenotype and therefore highlight the value of combinatorial series of drugs applied concurrently to block multiple cancer pathways. Boolean networks also highlights the role of biological switches in tumour progression and cancer stage evolution [503]. In this paper, authors use Boolean networks built from gene expression and protein interaction experiments to establish sets of genes which flip expression between different disease stages and act as cellular phenotypic switches to allow disease progression. Both studies confirm the strong potential of Boolean networks in predicting protein and cellular pathways involved in tumour progression leading to malignancy in different cancers, emphasising the robustness of abstract models in biological systems.

Boolean networks can incorporate deterministic or stochastic updates. The dynamics of networks are often updated on discrete time steps, where every state at a time $t+1$ is evaluated through its Boolean function and the values of the other variables at time t. Nodes can be updated simultaneously, in a synchronous manner, or asynchronously, that is nodes are selected for update randomly at each time step [504]. As genes do not update simultaneously in cells, asynchronous simulations are more realistic. The choice between update techniques must be done cautiously as different methods can give different results [505]. In synchronous update, a transition state has a unique successor which is not always the case in asynchronous. Different attractors can be reached with different probability when starting from the same initial condition in asynchronous update. Asynchronous updates tend to give additional outcomes, which are often complex and difficult to analyse [506]. Finally, asynchronous network are suitable for reduction techniques, which is untrue for synchronous networks as reduction can lead to loss of reachable states [507]. Reduction techniques aim to simplify a Boolean network while preserving its main properties.

Public databases provide easy-to-access information for network construction. The increasing amount of new omic data have developed the urgent need of biological databases listing metabolic and signalling pathways. These databases regroup biological data to facilitate researchers work by accelerating information searches often scattered in several papers and experiments. If the modeller already has a list of molecules in mind, these tools can build models by finding the genetic or physical interactions between components of the list and deliver a visualisation of the network. Amongst popular tools, KEGG [508], Reactome [509] and WikiPathway [510] all generate their databases through literature curation and experimental publications for different cell types in different organisms. However, these databases have different focuses and level of information for different molecular compounds. They are often biased by the shared community interest for common drug targeted molecules which limits for example the number of databases for protein-RNA interaction [511]. To overcome this issue, Omnipath collects 61 resources, with a focus on public resources containing literature-curated signalling interactions, which provide a current total of 99,255 interactions and has recently created an application linked to Cytoscape [512] for network analysis and visualisation [513]. The development of tools like Omnipath ensures the reliability of prior information used in network construction and help in the analysis of gene regulatory networks.

Studies in previous chapters have identified two genes, *HOXA9* and *APP*, with important patient stratification properties in blood malignancies. Further investigations have shown that both genes are not independent and actually form three groups of patients with different clinical characteristics in TCGA AML data (data description in subsection 2.2.1). Relationship between both genes is unknown, therefore, a better understanding of the regulatory dynamics behind these clusters could aid patient stratification and classification. To do so, I use Z3 theorem prover, a SMT solver by Microsoft [286], to identify network motifs explaining the three AML patient groups (see Methods section 2.6 for more details on SMT). Using the Z3Python package, I build a program which infers a list of gene motifs from biological observations. I apply this tool to the *HOXA9*/*APP* clusters. This work helps to identify a set of genes potentially involved in AML patient stratification and are therefore important for AML subtype characterisation and personalised treatment design. This approach can be applied generally to other systems to decipher gene interactions and understand complex biological mechanisms.

6.2 Results

Key findings of this work are divided into four subsections: in the first subsection, I describe the conceptualisation and construction of the motif search algorithm. The next subsection illustrates how the algorithm can be applied to biological systems. Here I use this algorithm to untangle the regulatory dynamics in the three AML cohorts of patients with distinct *HOXA9* and *APP* gene expressions. In the third subsection, I search for candidate genes for the newly found motifs. Finally, the last subsection shows how poor or good candidates are discriminated for the gene motifs.

6.2.1 Implementation of a motif inference algorithm

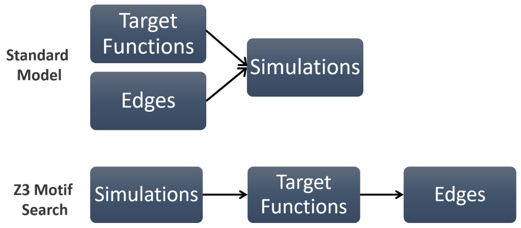


Fig. 6.2 **Standard network studies and the motif search program differ in their main purpose and input information.** Standard network studies aim to understand how the model inputs alter the network outcomes by finding attractors. In this work, the simulation outcomes and how the motif behaves to different initial conditions are known information. From this prior input, the algorithm aims to determine the networks which are able to reproduce these observations. To do so, it needs to find the corresponding edges and variable update functions, here the target functions, between input variables. To identify correct edges, target functions are defined from the input simulations which helps to identify the relationships between nodes.

To determine the regulatory dynamics behind the three *HOXA9/APP* clusters, I use the Z3Python package to build a python program which returns a list of Boolean networks with the ability to reproduce particular biological observations. This program takes as inputs one or several traces representing the biological observations and a list of genes that the motifs should contain. From all the possible motifs existing for a certain number of genes, the code selects and returns all the networks that match the input simulations as well as the smallest network and the consensus interactions. The smallest network is defined as the network with the lowest number of interactions and the consensus interactions as the required edges for the motifs to match the observations. This work diverges from standard studies which aim to find attractors for a defined network. Here, the biological outcomes are known and the aim is to design a network to represent those (Fig. 6.2).

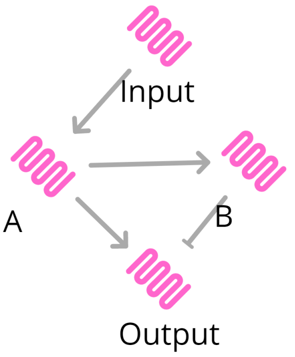


Fig. 6.3 **Network example.**

Z3 solves systems of constraints between defined variables. Constraints are important rules, properties or variable definitions a model must possess for Z3 to define the system as satisfiable. In this program, constraints are established by the traces which determine how the variable update functions should be defined for each gene. For example for the network in Figure 6.3, one possible trace is:

Input	A	B	Output
0	0	0	0
1	0	0	0
1	1	0	0
1	1	1	1
1	1	1	0
1	1	1	0

Rows of a trace show network states at each time step, while columns represent the network variables. In traces, the initial conditions are defined in the first row. Here, all genes are initially set to zero (False), meaning no or low activity. As shown by the second row of the trace, Input is a basal gene, which means it can activate itself when all genes are inactive at the previous step. In this example, the network simulations are synchronous which means all genes are updated at the same time. Therefore, when A is activated, B and Output are both active in the next time step. It should be noted that the last two steps are similar which means the network has reached a steady state. Due to the synchronous update, the network is deterministic. Consequently, for identical initial conditions, simulations always converge to the same attractor which is a single state where Input, A and B are active and Output inactive. Synchronous Boolean networks can also converge to a repeating sequence of states, called cycles [514]. A network can reach several attractors with different initial conditions. Finally, several traces can define the same network. Different definitions can be established for the activation and inhibition edges. In the example, Output is inactive when A and B are active. However, one could consider that Output activation by A is stronger than its inhibition by B, and thus another possible trace for the same network could be:

Input	A	B	Output
0	0	0	0
1	0	0	0
1	1	0	0
1	1	1	1
1	1	1	1

Variable update functions are essential features of Boolean networks. To determine the next state of a network, one must know the values of each node at the previous state as well as their corresponding variable update functions (Fig. 6.4). Here, a target function is defined as a variable update function that establishes the set of rules linking a node to the other nodes with logic operators (AND, NOT, OR). A target function in the motif inference algorithm is a Z3 Boolean variable giving the next state of a specific node given the current network state. In the example, the target function of Input when all nodes equal zero (False) is one (True) due to its basal activity. This means that if all genes are inactive at a certain time t, Input will be active at time $t+1$. Its resulting target function is defined as follows: `TF-Input-Input/False-A/False-B/False-Output/False = True`. The first part of the variable `TF-Input` gives the name of the node, here Input, and the second part `Input/False-A/False-B/False-Output/False` displays the current state of the network. The value of this target function gives the state of the node Input at the next time step, here True.

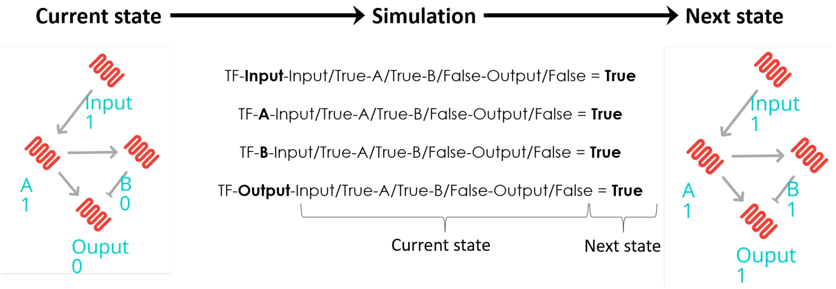
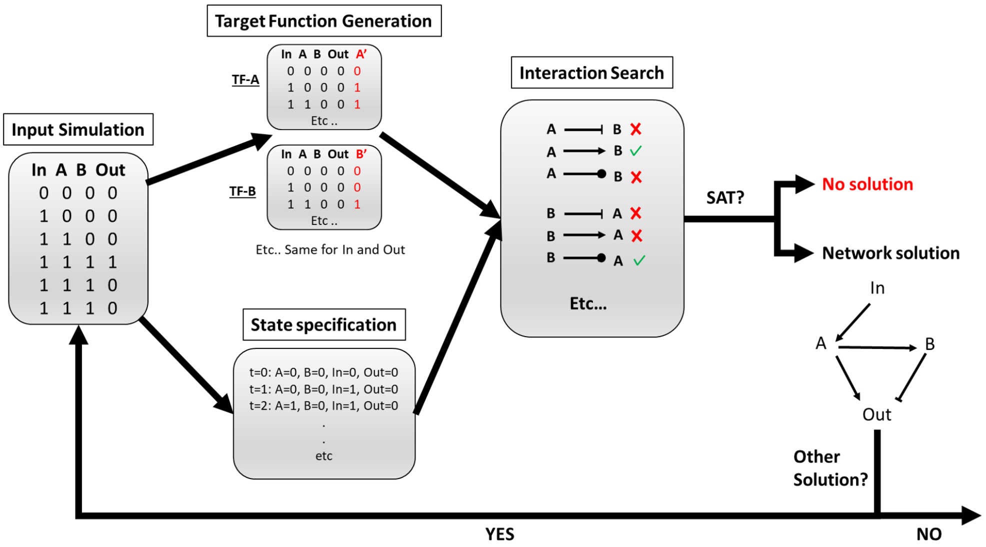

Current state
Simulation
Next state
Input
1
A
1
B
0
Ouput
0
TF-**Input**-Input/True-A/True-B/False-Output/False = **True**
TF-**A**-Input/True-A/True-B/False-Output/False = **True**
TF-**B**-Input/True-A/True-B/False-Output/False = **True**
TF-**Output**-Input/True-A/True-B/False-Output/False = **True**
Current state
Next state
Input
1
A
1
B
1
Ouput
1

Fig. 6.4 **Example of a one-step simulation using target functions.** Figures on the left and right represent the same network at different time steps. At each time step of a simulation, the program computes the next network state by looking at the target functions for all nodes given the current state. A target function is a variable update function represented by a Z3 Boolean variable for a specific node and network state.

The algorithm first creates all target functions for all nodes and all possible current states. Then, a "step" function assigns values to the target functions using the input traces (Fig. 6.2). Target functions not defined by the traces have their value assigned by the Z3 solver to obtain a satisfiable model.

Once all target functions are generated and assigned to a Boolean value, edges between the nodes are defined by establishing for each pair of genes the nature of their interaction (Fig. 6.2). The algorithm assumes 3 types of interactions: activation, inhibition and none. For simplification, mixed interactions are ignored, meaning genes cannot have complex interactions combining activation and inhibition. Indeed, molecular interactions can be context dependent. For example, Wise, a secreted protein involved in posterior neural marker induction, acts as both activator and inhibitor of the Wnt signalling pathway [515]. However, these interactions are rare [516], and therefore ignoring mixed interactions should not alter results. To test the interaction between a gene source and a gene target, all the target functions of the gene target are gathered. Then, the algorithm checks if the gene target expression is constant, increased or decreased when the gene source is activated but all other genes have kept the same activity. For example, in order to find the effect of a gene A over a gene B, the program compares all the paired target functions of B such that `TF-B-A/False-{rest of state}` and `TF-B-A/True-{rest of state}` have the same rest of state. If all paired target functions have the same value for B, lack of interaction between A and B is assumed. If B is activated in at least one paired target function, that is its target function value goes from False to True when A is activated, A activation of B is expected. Reciprocally, if activation of A inactivates B in at least one paired target function, the algorithm speculates that A inhibits B. If both activation and inhibition are found for B, the constraint is set to False which reciprocates to ignore mixed interactions and rejects motifs including these interactions.

A system is said satisfiable if and only if Z3 finds a solution to the problem with all constraints returning a True value. In this code, Z3 solver returns satisfiable if it finds at least one network containing target functions and interactions between all input genes which satisfy the constraints established by the observations/traces. The schematic of the code workflow can be seen in Figure 6.5.



In	A	B	Out
0	0	0	0
1	0	0	0
1	1	0	0
1	1	1	1
1	1	1	0
1	1	1	0

In	A	B	Out	A’
0	0	0	0	0
1	0	0	0	1
1	1	0	0	1
Etc ..				

In	A	B	Out	B’
0	0	0	0	0
1	0	0	0	0
1	1	0	0	1
Etc ..				

t=0: A=0, B=0, In=0, Out=0
t=1: A=0, B=0, In=1, Out=0
t=2: A=1, B=0, In=1, Out=0
.
.
etc

Fig. 6.5 **Code workflow for the motif inference program.** From one or several input traces representing biological observations, the code generates target functions for each input variable. In this example, input variables are In, A, B and Out. In parallel, it associates values to all variables at each time step using the trace and its corresponding time starting at $t = 0$ for the first row. Combining both state specification and target function information, Z3 can now explore different sets of interactions between all genes and test if one can reproduce the observations. If a set is found, the problem is said satisfiable and the code returns the motif with the corresponding interactions. The final step consists in excluding this motif from the solutions to look for other potential networks. The algorithm searches for a new solution until none can be found.

When the Z3 solver determines the problem as satisfiable for the input traces and genes, a network has been found. As mentioned formerly, one network can fit several traces in regard to the activation/inhibition rules applied to gene interactions. The converse is also true. Therefore, it is possible that several networks can reproduce the input observations. In order to find all possible networks, the first network is saved in a list and a new rule is added to constrain the solver to exclude this network from the possible solutions. If a new network is found, the result is saved again in the list and this process is repeated until no network can satisfy the problem. From this catalogue of motifs, the algorithm searches for interactions that are present in all the networks. These edges are defined as consensus interactions and represent necessary node interactions for the motifs to reproduce the observations. Finally, the algorithm searches for the smallest network, that is the network with the highest number of none interactions. Finding the smallest network and the consensus interactions facilitates network explorations by starting with smaller models. It also helps to quickly reject variables that do not possess the essential interactions.

Lastly, I include two functions to facilitate the motif search for the *HOXA9*/*APP* clusters which can be used for other biological problems. The first function converts the target functions of a network found by Z3 into target functions which can be read by the BioModelAnalyzer (BMA) tool [191] (Methods section 2.4.2). This functionality allows the validation of the networks found by Z3. The second function allows the addition of "switches" into the network. These switches represent biological events such as mutations and are not real biological molecules. For example, a switch can induce a loss of function in a gene, and is represented by an inhibition between the switch and its receptor. Switches only interact and modify the expression of one gene, called the switch receptor, while none gene can interact or modify the switch. Finally, the switch has a constant value and stays in its on or off state at all time steps. To include a switch in the observations, a new input needs to be added in the code specifying the name of the switch and its receptor.

6.2.2 Identification of motifs for the *HOXA9* /*APP* AML cohorts

As shown in the previous chapters 4 and 5, the analysis of the TCGA AML data highlighted the importance of *HOXA9* as a marker for AML stratification in a first instance, and then the unforeseen *APP* role in leukemia. Further work on the data showed the presence of three clusters of patients with distinct *HOXA9* and *APP* expression (Fig. 5.1B). These three clusters are unexpected: no link between *HOXA9* and *APP* has been previously identified in the literature and the role of *APP* in AML is not well understood [517, 414, 518]. To untangle why and how *HOXA9* and *APP* clusters appear in AML patients, I search for networks that can generate these clusters using the motif inference algorithm. Identification of molecular motifs can help to understand how these genes are linked and predict patient classification.

The aim of this work is to find the smallest motifs reproducing the three *HOXA9*/*APP* clusters. I therefore increase sequentially the number of variables until a fitting motif is found. The motif inference is initiated with two genes, but with these inputs, traces reproducing the three clusters cannot be generated. I increase the search to two genes with a switch event. With these variables and some hypothetical traces reproducing the clusters, Z3 returns unsatisfiable as a result of one of the switch constraints stating that a switch has one-way interaction with a single gene. Removing the switch constraints on the third variable is a sufficient condition for the Z3 solver to find Boolean networks for the clusters. Three genes are then required and sufficient to explain the formation of the three clusters.

All possible traces reproducing the clusters found by hand are tested in the program for the three genes, *APP*, *HOXA9* and a third unknown gene, called marker here. A set of three traces leads to three steady states representing the clusters and depending on the initial conditions: high expression for *APP* and *HOXA9*, called the high/high cohort, low expression for *HOXA9* and high expression for APP, the low/high cohort and finally high expression for *HOXA9* and low expression for APP, the high/low cohort. Only traces able to reproduce these clusters and resulting in a list of inferred networks are kept. In the possible motifs, three types of relationship between the marker and *APP* and *HOXA9* are found: the marker activates both genes, the marker inhibits them, and finally the marker inhibits one and activates the other. The smallest motifs for each interaction type are illustrated in Figure 6.6. It should be noted that motifs are symmetrical, that is each motif is one of the two solutions producing the same traces but with swapped *HOXA9* and *APP* nodes. Network dynamics are described in the following paragraphs.

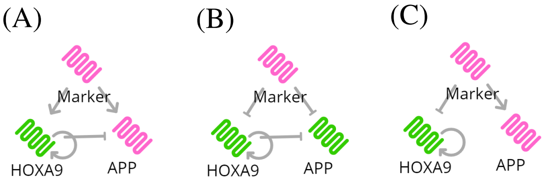


Fig. 6.6 **Motif search for *HOXA9*/*APP* clusters identifies three potential networks** (six with the symmetric networks). Proteins in green have a basal activity. (A) The marker activates both *HOXA9* and *APP*. *HOXA9* activates itself and inhibits *APP*. (B) The marker inihibits both *HOXA9* and *APP*. *HOXA9* activates itself and inhibits *APP*. (C) The marker activates *HOXA9* and inhibits *APP* and *HOXA9* activates itself.

If the marker activates both *APP* and *HOXA9* (Fig. 6.6A and Table 6.1), the low/high cohort should have a low expression for the marker (Trace 1). Here, *APP* have a basal activity, which means when all genes are turned off, it can activate itself. Activation of the marker is sufficient to induce both *HOXA9* and *APP* expression and thus the high/high cohort is expected to have a high expression for the marker gene (Trace 2). Finally, if *HOXA9* is turned on before *APP* while the marker stays off, *APP* expression remains low due to *HOXA9* inhibition and lack of activation (Trace 3). Then, the high/low cohort should have a low expression for the marker gene.

Trace 1			Trace 2			Trace 3		
Marker	*HOXA9*	*APP*	Marker	*HOXA9*	*APP*	Marker	*HOXA9*	*APP*
0	0	0	1	0	0	0	1	1
0	0	1	1	1	1	0	1	0
0	0	1	1	1	1	0	1	0

Table 6.1 **Traces of the three *HOXA9*/*APP* clusters when the marker activates both genes.**

If the marker inhibits both *APP* and *HOXA9* (Fig. 6.6B and Table 6.2), the high/high cohort should have a low expression for the marker gene (Trace 1). If the marker gene is initially turned on, two cases are possible: *APP* turns itself on if and only if *HOXA9* is not expressed before which leads to the low/high case (Trace 2). However, if *HOXA9* is on and stays on thanks to its positive self loop, *APP* expression is inhibited by *HOXA9* and the marker, which results in the high/low situation (Trace 3).

Trace 1			Trace 2			Trace 3		
Marker	*HOXA9*	*APP*	Marker	*HOXA9*	*APP*	Marker	*HOXA9*	*APP*
0	0	0	1	0	0	1	1	1
0	1	1	1	0	1	1	1	0
0	1	1	1	0	1	1	1	0

Table 6.2 **Traces of the three *HOXA9*/*APP* clusters when the marker inhibits both genes.**

The last scenario illustrates the case in which the marker inhibits *HOXA9* gene and activates *APP* (Fig. 6.6C and Table 6.3). The motif search predicts that for the high/high cohort, the marker should also have a high expression (Trace 3). In this case, *HOXA9* is turned on before the marker and stays in this state thanks to its positive feedback loop. However in the low/high cohort, *HOXA9* has low expression when the marker is turned on which results in its inhibition while *APP* is activated by the marker (Trace 2). Finally, the absence of marker activates *HOXA9* while *APP* expression stays low, which leads to the high/low case where patients should have low expression for the marker (Trace 1).

Trace 1			Trace 2			Trace 3		
Marker	*HOXA9*	*APP*	Marker	*HOXA9*	*APP*	Marker	*HOXA9*	*APP*
0	0	0	1	0	0	1	1	1
0	1	0	1	0	1	1	1	1
0	1	0	1	0	1	1	1	1

Table 6.3 **Traces of the three *HOXA9*/*APP* clusters when the marker inhibits *HOXA9* gene and activates *APP*.**

6.2.3 Investigation of possible motif gene candidates

To discriminate between the symmetrical possible networks in the cases of double activation or inhibition, I must determine which of *HOXA9* inhibition by *APP* or *APP* inhibition by *HOXA9* is the most likely. The R package Omnipath gathers genetic interaction information from a large number of public databases [513]. Using Omnipath, I search for *HOXA9* and *APP* connections. No direct interaction between *HOXA9* and *APP* can be found. I next look for genes that are downstream of *HOXA9* and upstream of *APP* to explore intermediate genes. One gene fits the inhibition of *APP* by *HOXA9*: TGFβ1. Similar work is performed for *HOXA9* inhibition by APP, but no gene matches the request. In Omnipath, TGFβ1 is marked as repressed by *HOXA9* and as a positive regulator of *APP*. While many references for the activation of *APP* by TGFβ1 are available [519–521], Omnipath only provides one questionable source for TGFβ1 inhibition by *HOXA9*. After literature curation, one paper showing *HOXA9* repression of TGFβ1 [522] is found. *HOXA9* inhibition of *APP* is confirmed thanks to another study demonstrating the role of *HOXA9* as a co-repressor partner of SMAD6 [523]. The latter has been shown to repress *SMAD4* activity which positively regulates *APP* [521] by competing for the binding with SMAD1 [524]. Collectively, an inhibition of *APP* by *HOXA9* via the TGFβ pathway is therefore likely, however, further biological experiments are required for confirmation of this connection in blood. In the rest of this work, I ignore networks including an inhibition of *HOXA9* by *APP* and focus on networks in which *HOXA9* inhibits *APP* and activates itself [334].

To determine the most biologically realistic motifs, identification of markers with a known relationship to both *HOXA9* and *APP* is essential. Omnipath does not find genes directly upstream of both *HOXA9* and *APP*. To fasten the search of markers in literature, I save the two lists of upstream genes for *HOXA9* and *APP* found by Omnipath and manually look in literature for an interaction with the second gene *HOXA9* or *APP* that Omnipath did not report. This work identifies nine genes interacting with both *HOXA9* and *APP*: *PRMT5*, *DNMT3A*, *TFAP2A*, *CTCF*, *STAT1*, *SP1*, *CDK1*, *GATA1* and *GATA2*. To this list, I add *JAK2* and *TET2* which were found to be upstream of *HOXA9* in the MPN work and literature [368]. Lastly, I also include *SMAD4* as *SMAD4* positively regulates *APP* and inhibits *HOXA9* activity [521, 525]. The markers can be classified into three categories. *JAK2*, *TET2*, *PRMT5*, *TFAP2A*, *STAT1*, *SP1*, *CDK1*, *GATA1* and *GATA2* form the largest category and all positively regulate *HOXA9* and *APP*. *SMAD4* and *CTCF* represents the second category and are positive regulators of *APP* and repressors of *HOXA9* expression. Finally, *DNMT3A* inhibits the activity of *APP* and *HOXA9* via its methylation function.

6.2.4 Candidate gene validation using gene expression and clinical data

Public databases and literature curation have helped to list 12 genetic candidates for the motifs. To validate the markers for the *HOXA9*/*APP* clusters found in AML, two tests are performed using the TCGA AML data to discriminate good and poor markers. If the selected genes are good markers, their expression in the three cohorts should follow a specific and predictable pattern (Tables 6.1, 6.2 and 6.3), but also stratify AML subtypes in the cohorts. Summary of the validation tests can be found in Table 6.4.

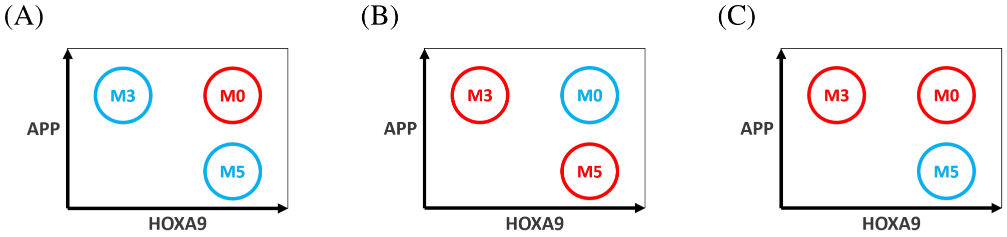


Fig. 6.7 **Expected marker expression level and FAB classification in the AML cohorts.** As shown in Tables 6.1, 6.2 and 6.3, different expression levels are expected for the marker in the *HOXA9*/*APP* clusters. The marker expression level in the cohorts depends on its interaction with *HOXA9* and *APP*: (A) the marker activates both *HOXA9* and APP, (B) the marker inhibits both genes and (C) the marker inhibits *HOXA9* and activates *APP*. Red indicates an expected high expression and blue a low expression. Moreover, each area is associated to a AML subtype from the French-American-British (FAB) classification: M0 the undifferentiated acute myeloblastic, M3 the acute promyelocytic leukemia (APL) and M5 the acute monocytic leukemia.

Each category of markers should show different expression level in the patient cohorts as a result of its relationship with *HOXA9* and *APP* (Fig. 6.7). For example as illustrated in Figure 6.7A, genes activating *HOXA9* and *APP* should have high expression in the cohorts with high *APP* and high *HOXA9* expression (Trace 2 in Table 6.1). To assess how well our candidate genes reproduce these patterns, AML patients are split into two groups with the lowest and highest expression for each marker. I look at their distribution in the *HOXA9*/*APP* cohorts and score each marker by calculating the percentage of patients found in the right cohorts. For example, if *GATA1* is highly expressed in a patient, this patient is expected to be found in the cohort with high expression for *HOXA9* and *APP* as *GATA1* activates both genes. This scoring method is repeated with diverse cut-offs for the marker expression level, that is more or less patients are included in the groups with high and low expression for the marker. This cut-off range analysis examines the possibility of extreme gene expressions giving better scores.

To test score significance, permutation tests are carried out by randomly shuffling patients in the low and high expression groups. Then a new score for the shuffled data is computed. This method is reiterated ten thousand times to obtain a score distribution for the shuffled data. How the original score falls into this distribution allows to compute the p-value for the hypothesis that this score is obtained by chance. P-value for this permutation test is the percentage of scores above the original score (Fig. 6.8). After Bonferroni correction for the

11 markers, I expect a gene to be a good marker if its p-value is lower than 4×10^{-3} (original threshold is 0.05). Through this permutation test, three classes of markers are found. The first class composed of *STAT1* and *CTCF* has a significant p-value after Bonferroni correction. I also include in this category *SMAD4* which has a p-value of 8×10^{-3}, considered close enough from significance in light of the conservative property of Bonferroni adjustment. The second group with *SP1*, *PRMT5* and *JAK2* has an insignificant p-value after Bonferroni correction, but a pattern can be identified in p-value significance for the different cut-offs before correction. Patients with extreme expression levels for those genes are more likely to be found in the expected *HOXA9*/*APP* cohorts. Lastly, the markers of the third class are found as bad markers in all the permutation tests and consist of *TET2*, *CDK1*, *DNMT3A* and *GATA2*. *GATA1* due to noisy results cannot be classified: this gene possesses a significant p-value after Bonferroni correction, but for some smaller cut-offs the tests without Bonferroni correction resume as insignificant. I exclude *TFAP2A* from testing as absent in the AML dataset.

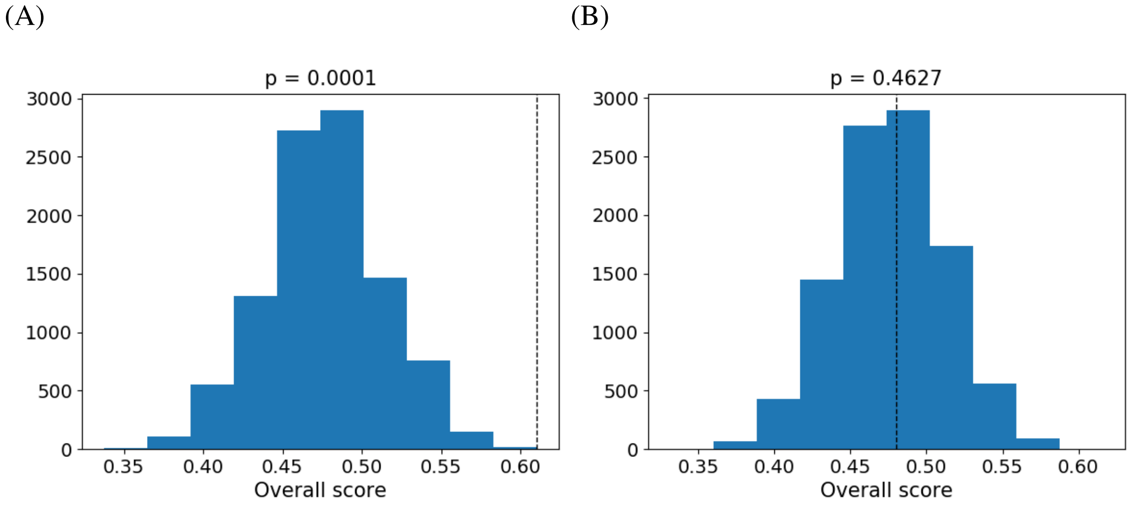


Fig. 6.8 ***STAT1* and *TET2* score distributions for the permutation test.** Score distributions for the shuffled data of (A) *STAT1*, found as a good marker by the permutation test ($p = 0.0001$), and (B) *TET2*, found as a poor marker ($p = 0.4627$). The dotted line represents the original score.

To further investigate the genes found significant in the permutation tests, I examine the AML French-American-British (FAB) classification distribution in the high and low expression groups of each marker (Fig. 6.9). For the genes activating *HOXA9* and APP, that is for *STAT1*, *GATA1* and *SP1*, M0 AML subtype patients should be mostly present in the group with higher expression level for the marker while M3 and M5 should be in the low expression group (Fig. 6.7A). I include *SP1* and *GATA1* in this new test as a result of their close significance for several permutation tests.

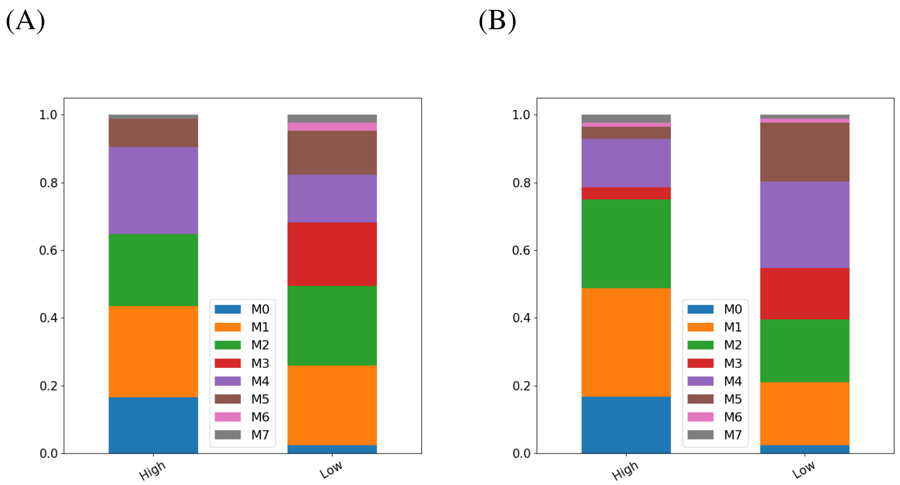


Fig. 6.9 **AML French-American-British (FAB) classification distribution in the high and low expression groups of (A) *SP1* and (B) *SMAD4*.** *SP1* activates both *APP* and *HOXA9* while *SMAD4* represses *HOXA9* and activates *APP*. Low *SP1* expression should be characterised by M3 and M5 subtypes and M0 should be observed in the high group. *SP1* expression fails to stratify M5 patients, but seems as an excellent marker for M3 patients. Per contrast, *SMAD4* wrongly classifies M3 patients which should be characterised by high *SMAD4* expression but is excellent for M0 and M5 patient stratification.

Results from this test show that no one of the genes seems very good at classifying all subtypes (Table 6.4). *STAT1* and *SP1* are good classifiers for M3 and M0 subtypes. *SP1* is an excellent M3 marker as the high expression cohort does not have any patients with M3 subtype. M5 patients seem randomly distributed between the two *SP1* groups. Distribution of M5 patients in the *STAT1* cohorts shows a good trend towards good classification. On the other hand, *GATA1* is good at classifying M5 patients but not M3 and M0. For the genes activating *APP* and inhibiting *HOXA9*, *CTCF* and *SMAD4*, patients with M0 and M3 subtypes should be found in the marker high expression group and M5 should be in the low expression one. However, both genes wrongly classify M3 patients which are mostly found in the low expression cohort. However, other subtypes are very well classified.

Marker	Bonferroni significant	Significant (< 0.05)	Good M0 Marker	Good M3 Marker	Good M5 Marker
Markers activating *HOXA9* and *APP*					
STAT1	Yes	Yes	Good	Good	Average
GATA1	Yes	Noisy	Poor	Poor	Good
SP1	No	Yes	Good	Excellent	Poor
PRMT5	No	Yes			
JAK2	No	Yes			
TET2	No	No			
CDK1	No	No			
GATA2	No	No			
Marker inhibiting *HOXA9* and *APP*					
DNMT3	No	No			
Markers inhibiting *HOXA9* and activating *APP*					
CTCF	Yes	Yes	Excellent	Poor	Excellent
SMAD4	Close	Yes	Excellent	Poor	Excellent

Table 6.4 **Marker validation test summary.** Column 2 and 3 of the table indicates if the p-value is significant for the permutation tests with or without Bonferroni correction. The last three columns summarises if the markers stratify well M0, M3, M5 AML patients as expected by the *HOXA9*/*APP* cohorts (visually using the histograms). *GATA1* is referred as noisy due to few insignificant permutation tests for random cut-offs despite its significant p-value of the Bonferroni test. *SMAD4* is said to be close to significance as its p-value is 8×10^{-3} for the Bonferroni test (threshold is 4×10^{-3}).

6.3 Discussion

In this chapter, I investigate the regulatory dynamics behind the observed *HOXA9*/*APP* patient cohorts in AML by developing a program inferring Boolean motifs from biological observations. Inputs of the algorithm consist of a list of genes and one or several traces representing the known biological dynamics between the genes. The program returns a set of Boolean networks generating the traces. It also identifies the smallest motif as well as the gene interactions required in all the accepted motifs. The input observations are transformed into constraints in the form of Boolean variables. The set of constraint variables must have True values for the theorem prover Z3 to find the appropriate edges between the genes and solve the motif inference problem. To determine gene motifs reproducing the AML clusters, various traces representing the different levels of *HOXA9* and *APP* expression in

AML patients are generated. Using the motif inference algorithm, I find that three genes, *HOXA9*, *APP* and a third unknown gene, are required and sufficient to generate the clusters. Using Omnipath and literature curation, 12 candidates are identified for the unknown gene. Two tests are performed to classify genes into good or poor markers by splitting patients into two groups with the lowest and highest expression for each marker. I then compare how patients in both cohorts are located in the *HOXA9*/*APP* clusters and what are their AML classification distribution. No gene perfectly passes both tests, but some genes such as *STAT1*, *CTCF* and *SMAD4* do better than others. Further biological experiments will be necessary to confirm these findings. The knockout of these markers with CRISPR in human blood cells to detect their effect on *HOXA9* and *APP* expression could for example confirm the motifs. Additionally, monitoring expression levels of these genes in patients with M0, M3 and M5 subtypes would also highlight their stratification property in leukemia. It should be noted that multiple gene candidates may work together which could also be tested in experiments. Identification of genetic markers are important for personalised diagnosis and treatment. I expect this program to have further biological and clinical applications by establishing links between gene expression and phenotypic changes or by finding new gene interactions.

This study concentrates the motif search on smallest models to explain the patient clusters observed in AML. As three genes are required and sufficient to reproduce the three levels of expression of *HOXA9* and *APP*, the list of unknown genes is narrowed down to one. However, one could argue the value of searching for larger motifs with higher number of components to improve the test results. For several reasons, I believe this would not lead to better conclusions. First, increasing the number of unknowns in a model increases prediction uncertainty [526]. By reducing the number of genes in the motifs, the number of possible gene interactions is reduced which therefore avoids addition of assumptions. Consequently, reducing unknowns strengthens the predictions based on fewer hypotheses. Second, larger networks can result in non-interpretable complex dynamics [527]. Increasing the number of genes in the motifs could lead to unexpected dynamics which can be difficult to biologically interpret. With smaller networks, the misinterpretation of the results is prevented. Moreover, small networks have been widely used in diverse biological systems and give satisfiable and relevant answers [528, 529]. Lastly, the validation of the motifs by biological experiments would be prone to more errors if many gene expression and interactions have to be tested. A large source of errors may be introduced in experiments and can be classified as "human" errors, systematic errors or random errors [530]. Human errors refer to avoidable mistakes such as misreading a quantity or using the wrong reactant. Conversely, systematic and random errors are inevitable regardless the number of similar experiments. Systematic errors often involve flaws in experimental instruments leading to inaccurate quantities, and random errors can be due to environmental variations such as temperature. By reducing the number of experiments, protocols are simplified and the error rate is reduced.

This work is subject to possibilities of improvement. First, I ignore complex interactions between genes and only consider activation or inhibition. Despite their known existence [516, 531], complex molecular interactions are often ignored in interaction inference tools [532, 533]. By neglecting mixed interactions, the algorithm possibly ignores gene connections explaining some biological mechanisms. However, I believe neglecting these relationships in the motifs could be apprehended by adding new variables representing the same gene with a distinct biological state. Another potential limitation of this study is that the program infers Boolean networks, which means variables can only take two values, True or False. Quantitative networks with a wider range of discrete values could integrate intermediate level of gene expression which might be biologically relevant. However, increasing granularity would most likely result in longer simulation time. Moreover, it has been shown that network structure is more relevant than kinetic details [534]. Boolean networks can also be used as groundwork for quantitative modelling when additional information or experiments become available. These findings further emphasise the benefit of Boolean networks for the motif inference algorithm.

Lastly, this work highlights the valuable use of theorem prover in biological dilemmas. Wolkenhauer et al define theorem proving in system biology as a tool to identify conditions entities must possess for the system properties to hold [535]. Although some life science studies have proved its performance for solving clinical and biological pathway problems [536–538], I believe theorem provers are underused in the medical field and could provide quick and useful answers to complex disease studies. Using theorem provers, de Maria et al [537] model the p53/Mdm2 DNA-damage repair mechanism and prove the importance of several properties for the system to behave properly. With a Boolean model and rules written in HyLL, a linear logic language, they show that in absence of DNA damage, the system stays at its initial state while DNA damage generates oscillations between two states. These observations establish the regulatory dynamics in normal cells and constitute valuable knowledge for deciphering the progression of dysregulated processes in diseases.

Chapter 7

Discussion

This thesis presents a wide range of computational, statistical and mathematical techniques to study tumour progression in different blood malignancies. I show that important driver events considerably impact on the cancer dynamics and the tumour clinical characteristics. The order in which oncogenic aberrations have appeared alters the underlying cell fate and functions, and therefore leads to distinct disease subtypes. Similarly, the timing between resistance emergence and beginning of treatment affects the diagnosis outcome. Better insights into the evolution of critical oncogenic aberrations can help to find the right treatment strategy. In the following paragraphs, I discuss how the work presented in this thesis could affect clinical monitoring and treatment.

This thesis demonstrates that resistance take-over and mutation order influence patient treatment protocols. Specifically, for blood cancers with weak competition between cancerous populations, a daily treatment at maximal dose becomes redundant once the number of resistant cells exceeds the sensitive one. Hence, timing of cancer cell dynamics determines when treatment should be stopped to avoid unnecessary toxicity. Similarly, timing between two mutations alter the phenotype of cells as shown in the work in Myeloproliferative Neoplasms. When *TET2* is mutated before *JAK2*, its function loss decreases the gene expression of some *JAK2* downstream targets. Consequently, JAK2 inhibitor drugs become inefficient for those patients despite the *JAK2* mutation as a result of the evolution switch property of *JAK2* and *TET2* common downstream target *HOXA9*.

This work also highlights the value of frequent tumour monitoring to predict clinical outcomes (Fig. 7.1). As shown with the model of lymphoma growth, measuring proliferation rate of cancer cells could help predict patient survival, as with increasing tumour burden, organ dysfunction results in lower proliferation rates when approaching maximum burden and death. Loss of *APP* expression as well as high *HOXA9* expression in leukemia are valid indicators for poor survival in patients. Control and stratification of patients for these genetic markers can help clinicians initiate a faster care of those aggressive tumours.

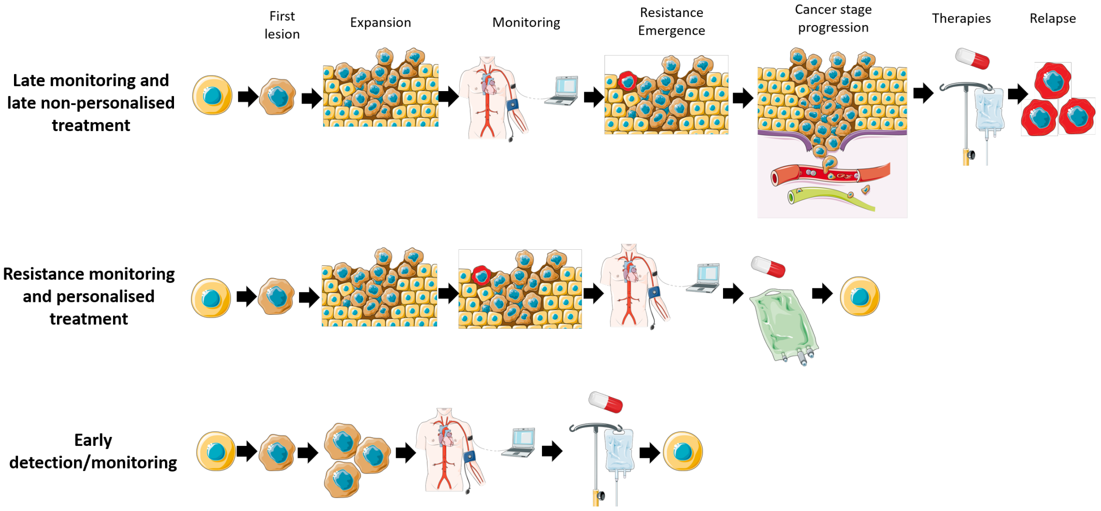

First lesion
Expansion
Monitoring
Resistance Emergence
Cancer stage progression
Therapies
Relapse
Late monitoring and late non-personalised treatment
Resistance monitoring and personalised treatment
Early detection/monitoring

Fig. 7.1 **Monitoring and oncogenic event timing in cancer progression.** Three scenarios of cancer progression demonstrate the importance of monitoring in cancer evolution. A good monitoring is an early monitoring in which clinicians have identified the disease markers and potential resistant clones. Efficient monitoring also examines specific gene expressions which correlate with prognosis and disease classification. Identification of important markers and clones for cancer evolution improves disease monitoring and therefore helps clinicians to develop personalised cancer therapies.

Finally, this thesis emphasises the benefit of biological networks to link regulatory dynamics with tumour evolution and clinical features. Gene and protein networks possess many functionalities which allow users to simulate major molecular events such as genetic mutations but also drug application which has not been investigated in this work. Changes on the cell phenotype induced by these events justify the observed disease characteristics. Two distinct analyses in which molecular networks are central to the methodology are carried out in this thesis. First, a network is constructed to understand which molecules are responsible for the evolution switch property observed in Myeloproliferative Neoplasms patients bearing both *JAK2* and *TET2* mutations. This model helps to assess important genes involved in the disease progression. Then in Chapter 6, a program is built to infer gene motifs reproducing patient clusters observed in AML. Both studies help to decipher the complexity of gene dynamics in blood diseases and identify potential drug targets to fight tumour progression.

In the following sections, I discuss new insights related to specific clinical aspects this thesis highlights. Specifically, I review the distinct clinical characteristics of solid and liquid cancers and how treatment strategies are directly impacted by those. I also argue about how the specific order of biological processes for lineage commitment can be viewed as the premises for the influence of oncogenic event timing in blood diseases. In a third section, I examine the novel promising technologies to study cancer evolution and how these could be applied to my work. Finally, I conclude this chapter with a list of the detailed results of this thesis.

7.1 Liquid versus solid tumours

Blood cancers arise from hematopoietic cells. Unlike solid tumours, they do not form a mass, but circulate in our blood system and are therefore easier to access. Consequently, despite their lower prevalence in adults, publications and researches on hematological malignancies are more frequent than solid cancers [539]. As for treatments, liquid cancers can be treated with chemotherapy and targeted therapies, but also with stem cell transplant which is primarily used in blood diseases [540, 541]. Stem cell transplant allows clinicians to transfer to the patient bone marrow new hematopoietic stem cells after damaging the old ones with toxic therapies. However, haematological malignancies cannot be removed by surgical interventions [542]. Thanks to their lack of physical barriers, the infiltration of immune cells and drug targeting against cancer cells are facilitated in liquid cancers [543]. Liquid and solid tumours are also genetically different, for example, chromosome aberrations are mostly found in blood cancers [544]. Finally, well-known oncogenic processes such as angiogenesis and metastasis are also radically different in blood cancers [545, 546]. Hence, liquid and solid cancers display many dissimilarities, resulting in distinct cell dynamics, disease progression and clinical treatment strategies.

Among possible cancer treatment strategies, adaptive therapy aims at stabilising therapeutic sensitive cells by giving tumours resting periods between drugging applications [71]. Rest periods are necessary to allow the sensitive cells, the primary competitors of resistant clones, to be saved from complete extinction. Adaptive therapy by keeping constant competition interactions between clones avoids resistance expansion and therefore increases patient survival. However, adaptive therapy has only been used and investigated in solid cancers [547, 73, 548]. I believe adaptive therapy due to the differences in spatial properties and cell dynamics of liquid cancers might not succeed to prolong survival as long as observed in prostate cancer [548].

One of the major results of Chapter 3 is that the therapeutic sensitive and therapeutic resistant cell populations have similar proliferation rates in lymphoma. This result can be interpreted as both populations having similar fitness in absence of treatment. This lack of resistance cost reduces competition between clones and therefore shortens possible therapy strategies such as adaptive therapy. Despite the improved survival after stabilisation of the sensitive population, the best survival outcome is obtained by daily injections of reduced efficacy drugging. This finding correlates with a recently published paper which highlights the importance of using spatial models to describe solid tumour dynamics treated with adaptive therapies [385]. In their model, trapping resistance cells inside tumours due to spatial constraints can improve treatments. This cannot be achieved in liquid tumours as lymphoma cells circulate freely in blood and cannot be trapped. I hypothesise that in absence of resistance cost, adaptive therapy is not the optimal strategy in liquid cancers due to limited spatial pressure. However, treating tumours close from carrying capacity seems to improve adaptive therapy results [549]. Competition between cancerous clones is indeed enhanced when tumours are closed from the carrying capacity. This conclusion is supported by the reduced-efficacy daily drugging treatment in lymphoma which greatly improves survival by creating a stable equilibrium close from maximal tumour burden between sensitive and resistance cells.

7.2 Hematopoietic dysregulation: the importance of gene expression dynamics in cell fate

In multicellular organisms, each cell has a defined type and function to support life and reproduction of the individual [550]. Here, I define cell fate as the last differentiated state a cell can be in in a stable environment. Cell fate commitment is determined by cell-type specific transcriptional programs which consist of genetic, epigenetic and environmental processes leading to fully functional and mature cells [551]. Progressively, cells transition from one state to another losing their immaturity and self-renewal ability while gaining important features for their final role in a particular tissue or organ [552]. Through acquisition of their identity, cells also gain specific structure and morphology which are important to their function. Cell fate differentiation is therefore a central mechanism in embryogenesis [553], but also in tissue homeostasis [554] and hematopoiesis [555].

Hematopoietic stem cells (HSC) possess the ability to regenerate the entire blood system and represent therefore the first state of hematopoietic cell fate commitment [75]. Production of all mature cells by HSC requires the involvement of many biological processes. Important signalling pathways such Notch, Shh, Wnt and Smad have been shown to control both self-renewal and differentiation of blood cells [556]. Cytokines, small proteins produced by immune cells, play also an important role in lineage differentiation [557–559].

Timing of events in hematopoiesis is crucial for the healthy development of all lineages and differentiated cells which reflects on blood disease evolution. As a large variety of mechanisms are involved in blood cell production, the ordering between all events leading to the specification of the cell identity must be thoroughly performed. Consequently, activation of markers in the wrong lineage can alter cell fate as shown by the artificial induction of the *CEBPα* and *CEBPβ* myeloid markers in lymphoid progenitors reprogrammed into macrophages [560]. It is therefore not surprising that different orders between important hematopoietic transcription factors can also modify cell fate. In their paper, Iwasaki et al [561] show how the lineage commitment of Granulocyte-Monocyte Progenitors (GMPs) is regulated by different orders of *GATA2* and *CEBPα* expression. If upregulation of *GATA2* in GMPs is followed by the downregulation of *CEBPα*, then GMPs differentiate into eosophil progenitors. On the other hand, maintained *CEBPα* expression results in GMPs differentiation into basophil/mast cell progenitors. This study confirms the importance of studying oncogenic event order to clarify disease clinical characteristics. Systems such as hematopoiesis require clear defined steps towards healthy cell development. Hence, even when dysregulated by genetic or epigenetic aberrations, distinct blood diseases necessitate precise order of events to display their own characteristics. For example, as shown in this thesis, Polycythemia vera (PV) which overproduces erythroid cells needs the upregulation of several *JAK2* downstream targets. Apparition of a mutation such as *TET2* loss impacting on *JAK2* targets before *JAK2* is mutated disables erythroid overproduction and therefore PV diagnosis.

Finally, the investigation of gene expression dynamics in a strictly organised system such as hematopoiesis reveals the complexity of genotype-phenotype interactions in cells. As mentioned above, slight disruption in important lineage marker expression can greatly impact differentiation and cell fate. This work illustrates that the earliest the hematopoietic stage the greater the impact of a marker aberration. Dissociation between the lymphoid and the myeloid lineages is one of the first step towards hematopoietic stem cell differentiation [76], in which the Amyloid Precursor Protein (APP) seems to participate. Its dysregulation, specifically its low expression, induces poor survival probability in leukemia patients. Similarly, *HOXA9* expression has been shown to be high in early progenitors and downregulated with differentiation [341]. This thesis further demonstrates the strong influence of its loss or overactivation on the development of haematological malignancies. By directly or indirectly acting upstream of important hematopoietic markers, *HOXA9* is a master regulator of hematopoiesis and an important prognosis and clinical marker in blood cancers. This also raises the question of why *HOXA9* is often impacted by upstream mutations or translocations in blood diseases, but rarely mutated. Two explanations could justify this observation. First, *HOXA9* belongs to a large family of genes which interact with each other and consequently, upstream mutations such as *MEIS1* often lead to the aberrant expression of several HOX genes [562]. Recent studies have shown that loss of one HOX gene is not compensated by other HOX genes [563] and that mice with multiple null HOX genes demonstrate worse clinical characteristics than single HOX null mice [564]. Therefore, cancer cells would benefit better from mutations upstream of *HOXA9* rather than its single mutation. A second explanation is that *HOXA9* mutation could result in its extreme expression which might be lethal for cells as a result of vital *HOXA9* downstream targets. Similar finding has been shown for overexpression of wild-type *RAS* which can promote tumour expansion while "too much" *RAS* would lead to cell senescence [565]. A mutation in an upstream regulator of *HOXA9* could therefore promote its expression while preserving cell vital needs. To conclude, this thesis shows that studying the dynamics between frequently mutated genes and their upstream and downstream targets sometimes better explain disease progression and resulting clinical outcomes than studies focusing on single mutation. Further work on non-mutated oncogenic genes could enlightens unknown important cancerous mechanisms.

7.3 Novel experimental technologies for tumour progression analyses

This thesis highlights the importance of studying tumour progression and evolution to tackle the complex disease that is cancer. As shown by the diverse methods used in this work, a large range of computational and modelling tools are currently publicly available to explore cancer evolution. However, the recent development of "omics" technologies have considerably improved our understanding of cancer dynamics [566], some of which could be applied to the work carried out in this thesis.

Among these emergent promising technologies, RNA velocity has already shown its great value to determine single cell future state on a short timescale in various cancer studies [567, 568]. Established in 2018 by La Manno et al [569], the spliced mRNA abundance, or RNA velocity, can be determined by solving a simple model which includes the mRNA degradation and the production of spliced mRNA from unspliced mRNA. Indeed, the increased transcription of a gene induces the upregulation of nascent/unspliced mRNA, followed by increase of mature/spliced mRNA (and opposite when the gene is repressed). By measuring the current (spliced mRNA) and future (unspliced mRNA) states of the cell, this technique predicts gene expression dynamics using scRNAseq data. Velocity can after be displayed as vector fields onto existing reduced dimension plots such as t-SNE to visualise directionality in differentiation trajectories [570]. This method should have a great impact in the cancer evolution field. RNA velocity could for example confirm the role of APP in leukemia as well as in the lymphoid versus myeloid differentiation process by comparing differentiation trajectories of hematopoietic stem cells with and without APP knockout.

Predicting cancer evolution remains an attractive but complex solution to control cancer [571]. Based on this idea, Hosseini et al [572] estimate mutational pathway probabilities using driver mutation data and conjunctive Bayesian networks (CBN). CBN are networks describing combination of events, here mutations, with their order. The technique described by authors possesses the great advantage of not requiring to measure the effect of mutations on cell fitness and therefore avoids simulation of fitness landscapes which can be experimentally costly [573]. In their paper, authors find that cancer progression is highly predictable in most cancer types, and even higher in tumour samples from metastatic sites. Their method could be used to further investigate the impact of *JAK2* and *TET2* mutation order. For example, it could help to stratify MPN diseases (PV, ET, PMF) according to *JAK2* and *TET2* mutation order. Another application of this method could be to examine the proportion of secondary AML patients with an initial *JAK2* or *TET2* mutation to help clinicians identify the most dangerous mutational routes from MPN to leukemogenesis.

7.4 Conclusion

Studying the influence of biological event timing in blood cancer progression has permit to highlight the importance of studying gene, cell and tumour dynamics to understand patient stratification. Combined together these events determine the path toward which tumours develop, and therefore how they should be treated. Using computational models of blood cancer progression, this thesis proposes alternative treatment strategies and identifies potential new drug targets. I recapitulate here the main findings of this work:

- Coupled with the lack of strong space constraints in blood tumours, absence of therapeutic resistance fitness cost reduces cell competition and possibilities of treatment strategies such as adaptive therapy when resistance is present before patient treatment starts.

- In blood tumours with a high proportion of sensitive compared to resistant cells, treatment holidays can improve patient survival by increasing sensitive cell proliferation at the expense of resistant cells as a result of the released nutrients and space of dying cells.

- Mutation order greatly impact the regulatory dynamics of important hematopoietic genes in blood cancer patients and the identification of biological switch such as *HOXA9* in one malignancy can explain tumour development and patient stratification in closely related diseases.

- Identification of markers involved in the early stages of hematopoiesis differentiation such as *APP* is important to highlight genes able to induce an undifferentiated state in blood cells, resulting in deadly blood malignancies.

- As shown by our motif inference program, the use of computational models for deciphering the regulatory dynamics of a blood disease at the molecular scale can help to decipher disease dynamics at larger scales. Notably, gene networks are valuable tools to study patient stratification and identify new potential drug targets.

References

[1] Carla P Martins, Lamorna Brown-Swigart, and Gerard I Evan. Modeling the therapeutic efficacy of p53 restoration in tumors. *Cell*, 127(7):1323–1334, 2006.

[2] Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[3] Laurence Loewe and William G Hill. The population genetics of mutations: good, bad and indifferent, 2010.

[4] Ruchira S. Datta, Alice Gutteridge, Charles Swanton, Carlo C Maley, and Trevor A Graham. Modelling the evolution of genetic instability during tumour progression. *Evolutionary applications*, 6(1):20–33, 2013.

[5] Armelle Calipel, Gaelle Lefevre, Celio Pouponnot, Frédéric Mouriaux, Alain Eychène, and Frédéric Mascarelli. Mutation of b-raf in human choroidal melanoma cells mediates cell proliferation and transformation through the mek/erk pathway. *Journal of Biological Chemistry*, 278(43):42409–42418, 2003.

[6] Weijia Zhu, Alison Cowie, Gihane W Wasfy, Linda Z Penn, Brian Leber, and David W Andrews. Bcl-2 mutants with restricted subcellular location reveal spatially distinct pathways for apoptosis in different cell types. *The EMBO journal*, 15(16):4130–4141, 1996.

[7] Richard A Larson. Is secondary leukemia an independent poor prognostic factor in acute myeloid leukemia? *Best practice & research Clinical haematology*, 20(1):29–37, 2007.

[8] Michael Korenjak and Jiri Zavadil. Experimental identification of cancer driver alterations in the era of pan-cancer genomics. *Cancer science*, 110(12):3622, 2019.

[9] Akira Yokoyama, Nobuyuki Kakiuchi, Tetsuichi Yoshizato, Yasuhito Nannya, Hiromichi Suzuki, Yasuhide Takeuchi, Yusuke Shiozawa, Yusuke Sato, Kosuke Aoki, Soo Ki Kim, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature*, 565(7739):312–317, 2019.

[10] Cyriac Kandoth, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew A Wyczalkowski, et al. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, 2013.

[11] Christopher D McFarland, Kirill S Korolev, Gregory V Kryukov, Shamil R Sunyaev, and Leonid A Mirny. Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences*, 110(8):2910–2915, 2013.

[12] Christopher D McFarland, Leonid A Mirny, and Kirill S Korolev. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proceedings of the National Academy of Sciences*, 111(42):15138–15143, 2014.

[13] Amir Eden, François Gaudet, Alpana Waghmare, and Rudolf Jaenisch. Chromosomal instability and tumors promoted by dna hypomethylation. *Science*, 300(5618):455–455, 2003.

[14] Valerie Greger, Eberhard Passarge, Wolfgang Höpping, Elmar Messmer, and Bernhard Horsthemke. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Human genetics*, 83(2):155–158, 1989.

[15] Ping Chi, C David Allis, and Gang Greg Wang. Covalent histone modifications—miswritten, misinterpreted and mis-erased in human cancers. *Nature reviews cancer*, 10(7):457–469, 2010.

[16] Naoyo Nishida, Hirohisa Yano, Takashi Nishida, Toshiharu Kamura, and Masamichi Kojiro. Angiogenesis in cancer. *Vascular health and risk management*, 2(3):213, 2006.

[17] Karin E De Visser, Alexandra Eichten, and Lisa M Coussens. Paradoxical roles of the immune system during cancer development. *Nature reviews cancer*, 6(1):24–37, 2006.

[18] Alexandre Corthay. Does the immune system naturally protect against cancer? *Frontiers in immunology*, 5:197, 2014.

[19] Roy Noy and Jeffrey W Pollard. Tumor-associated macrophages: from mechanisms to therapy. *Immunity*, 41(1):49–61, 2014.

[20] Igor B Roninson. Tumor cell senescence in cancer treatment. *Cancer research*, 63(11):2705–2715, 2003.

[21] Xiangjun Kong, Thomas Kuilman, Aida Shahrabi, Julia Boshuizen, Kristel Kemper, Ji-Ying Song, Hans WM Niessen, Elisa A Rozeman, Marnix H Geukes Foppen, Christian U Blank, et al. Cancer drug addiction is relayed by an erk2-dependent phenotype switch. *Nature*, 550(7675):270–274, 2017.

[22] Nicholas McGranahan, Francesco Favero, Elza C de Bruin, Nicolai Juul Birkbak, Zoltan Szallasi, and Charles Swanton. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science translational medicine*, 7(283):283ra54–283ra54, 2015.

[23] Matthew D Stachler, Amaro Taylor-Weiner, Shouyong Peng, Aaron McKenna, Agoston T Agoston, Robert D Odze, Jon M Davison, Katie S Nason, Massimo Loda, Ignaty Leshchiner, et al. Paired exome analysis of barrett's esophagus and adenocarcinoma. *Nature genetics*, 47(9):1047, 2015.

[24] Nathan P Young, Denise Crowley, and Tyler Jacks. Uncoupling cancer mutations reveals critical timing of p53 loss in sarcomagenesis. *Cancer research*, 71(11):4040–4047, 2011.

[25] Karen Gomez, Sayaka Miura, Louise A Huuki, Brianna S Spell, Jeffrey P Townsend, and Sudhir Kumar. Somatic evolutionary timings of driver mutations. *BMC cancer*, 18(1):85, 2018.

[26] Samra Turajlic, Hang Xu, Kevin Litchfield, Andrew Rowan, Tim Chambers, Jose I Lopez, David Nicol, Tim O'Brien, James Larkin, Stuart Horswell, et al. Tracking cancer evolution reveals constrained routes to metastases: Tracerx renal. *Cell*, 173(3):581–594, 2018.

[27] Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, 2013.

[28] John S Welch, Timothy J Ley, Daniel C Link, Christopher A Miller, David E Larson, Daniel C Koboldt, Lukas D Wartman, Tamara L Lamprecht, Fulu Liu, Jun Xia, et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell*, 150(2):264–278, 2012.

[29] Mel Greaves and Carlo C Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012.

[30] J Martin Brown and Laura D Attardi. The role of apoptosis in cancer development and treatment response. *Nature reviews cancer*, 5(3):231–237, 2005.

[31] Christopher W Elston and Ian O Ellis. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991.

[32] Gulisa Turashvili and Edi Brogi. Tumor heterogeneity in breast cancer. *Frontiers in medicine*, 4:227, 2017.

[33] Gaorav P Gupta and Joan Massagué. Cancer metastasis: building a framework. *Cell*, 127(4):679–695, 2006.

[34] Heidi Bissig, Jan Richter, Richard Desper, Verena Meier, Peter Schraml, Alejandro A Schäffer, Guido Sauter, Michael J Mihatsch, and Holger Moch. Evaluation of the clonal relationship between primary and metastatic renal cell carcinoma by comparative genomic hybridization. *The American journal of pathology*, 155(1):267–274, 1999.

[35] Eric R Fearon and Bert Vogelstein. A genetic model for colorectal tumorigenesis. *cell*, 61(5):759–767, 1990.

[36] R Fisher, L Pusztai, and C Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479–485, 2013.

[37] Jonathan J Keats, Marta Chesi, Jan B Egan, Victoria M Garbitt, Stephen E Palmer, Esteban Braggio, Scott Van Wier, Patrick R Blackburn, Angela S Baker, Angela Dispenzieri, et al. Clonal competition with alternating dominance in multiple myeloma. *Blood, The Journal of the American Society of Hematology*, 120(5):1067–1076, 2012.

[38] Marco Gerlinger, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl j Med*, 2012(366):883–892, 2012.

[39] Mark C Lloyd, Jessica J Cunningham, Marilyn M Bui, Robert J Gillies, Joel S Brown, and Robert A Gatenby. Darwinian dynamics of intratumoral heterogeneity: not solely random mutations but also variable environmental selection forces. *Cancer research*, 76(11):3136–3144, 2016.

[40] Mingzhou Guo, Yaojun Peng, Aiai Gao, Chen Du, and James G Herman. Epigenetic heterogeneity in cancer. *Biomarker research*, 7(1):23, 2019.

[41] Alessandro Pastore, Federico Gaiti, Sydney X Lu, Ryan M Brand, Scott Kulm, Ronan Chaligne, Hongcang Gu, Kevin Y Huang, Elena K Stamenova, Wendy Béguelin, et al. Corrupted coordination of epigenetic modifications leads to diverging chromatin states and transcriptional heterogeneity in cll. *Nature communications*, 10(1):1–11, 2019.

[42] Ali J Marian. Challenges in medical applications of whole exome/genome sequencing discoveries. *Trends in cardiovascular medicine*, 22(8):219–223, 2012.

[43] Hege G Russnes, Nicholas Navin, James Hicks, and Anne-Lise Borresen-Dale. Insight into the heterogeneity of breast cancer through next-generation sequencing. *The Journal of clinical investigation*, 121(10):3810–3818, 2011.

[44] Alexander W Wyatt, Fan Mo, Yuzhuo Wang, and Colin C Collins. The diverse heterogeneity of molecular alterations in prostate cancer identified through next-generation sequencing. *Asian journal of andrology*, 15(3):301, 2013.

[45] Kornelius Schulze, Jean-Charles Nault, and Augusto Villanueva. Genetic profiling of hepatocellular carcinoma using next-generation sequencing. *Journal of hepatology*, 65(5):1031–1042, 2016.

[46] Vanessa Almendro, Andriy Marusyk, and Kornelia Polyak. Cellular heterogeneity and molecular evolution in cancer. *Annual Review of Pathology: Mechanisms of Disease*, 8:277–302, 2013.

[47] James PB O'Connor. Cancer heterogeneity and imaging. In *Seminars in cell & developmental biology*, volume 64, pages 48–57. Elsevier, 2017.

[48] Daniel Zips, Klaus Zöphel, Nasreddin Abolmaali, Rosalind Perrin, Andrij Abramyuk, Robert Haase, Steffen Appold, Jörg Steinbach, Jörg Kotzerke, and Michael Baumann. Exploratory prospective trial of hypoxia-specific pet imaging during radiochemotherapy in patients with locally advanced head-and-neck cancer. *Radiotherapy and Oncology*, 105(1):21–28, 2012.

[49] Ruping Sun, Zheng Hu, Andrea Sottoriva, Trevor A Graham, Arbel Harpak, Zhicheng Ma, Jared M Fischer, Darryl Shibata, and Christina Curtis. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nature genetics*, 49(7):1015–1024, 2017.

[50] Jason H Moore. A global view of epistasis. *Nature genetics*, 37(1):13–14, 2005.

[51] Alan Ashworth, Christopher J Lord, and Jorge S Reis-Filho. Genetic interactions in cancer progression and treatment. *Cell*, 145(1):30–38, 2011.

[52] Nigel J O'Neil, Melanie L Bailey, and Philip Hieter. Synthetic lethality and cancer. *Nature Reviews Genetics*, 18(10):613–623, 2017.

[53] Benjamin G Bitler, Katherine M Aird, Azat Garipov, Hua Li, Michael Amatangelo, Andrew V Kossenkov, David C Schultz, Qin Liu, Ie-Ming Shih, Jose R Conejo-Garcia, et al. Synthetic lethality by targeting ezh2 methyltransferase activity in arid1a-mutated cancers. *Nature medicine*, 21(3):231, 2015.

[54] Cory M Johannessen, Jesse S Boehm, So Young Kim, Sapana R Thomas, Leslie Wardwell, Laura A Johnson, Caroline M Emery, Nicolas Stransky, Alexandria P Cogdill, Jordi Barretina, et al. Cot drives resistance to raf inhibition through map kinase pathway reactivation. *Nature*, 468(7326):968–972, 2010.

[55] Claire J Cairney, Lauren S Godwin, Alan E Bilsland, Sharon Burns, Katrina H Stevenson, Lynn McGarry, John Revie, Jon D Moore, Ceri M Wiggins, Rebecca S Collinson, et al. A 'synthetic-sickness' screen for senescence re-engagement targets in mutant cancer backgrounds. *PLoS genetics*, 13(8):e1006942, 2017.

[56] Motonari Kondo, Amy J Wagers, Markus G Manz, Susan S Prohaska, David C Scherer, Georg F Beilhack, Judith A Shizuru, and Irving L Weissman. Biology of hematopoietic stem cells and progenitors: implications for clinical application. *Annual review of immunology*, 21(1):759–806, 2003.

[57] David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.

[58] Manu Shivakumar, Younghee Lee, Lisa Bang, Tullika Garg, Kyung-Ah Sohn, and Dokyoon Kim. Identification of epigenetic interactions between mirna and dna methylation associated with gene expression as potential prognostic markers in bladder cancer. *BMC medical genomics*, 10(1):30, 2017.

[59] Prashant Kumar Singh and Moray J Campbell. The interactions of microrna and epigenetic modifications in prostate cancer. *Cancers*, 5(3):998–1019, 2013.

[60] Xiaotian Zhang, Jianzhong Su, Mira Jeong, Myunggon Ko, Yun Huang, Hyun Jung Park, Anna Guzman, Yong Lei, Yung-Hsin Huang, Anjana Rao, et al. Dnmt3a and tet2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells. *Nature genetics*, 48(9):1014, 2016.

[61] James G Jackson, Vinod Pant, Qin Li, Leslie L Chang, Alfonso Quintás-Cardama, Daniel Garza, Omid Tavana, Peirong Yang, Taghi Manshouri, Yi Li, et al. p53-mediated senescence impairs the apoptotic response to chemotherapy and clinical outcome in breast cancer. *Cancer cell*, 21(6):793–806, 2012.

[62] Sherif Y El Sharouni, HB Kal, and JJ Battermann. Accelerated regrowth of non-small-cell lung tumours after induction chemotherapy. *British journal of cancer*, 89(12):2184–2189, 2003.

[63] Samer Tohme, Richard L Simmons, and Allan Tsung. Surgery for cancer: a trigger for metastases. *Cancer research*, 77(7):1548–1552, 2017.

[64] Randall S Singer, Roger Finch, Henrik C Wegener, Robin Bywater, John Walters, and Marc Lipsitch. Antibiotic resistance—the interplay between antibiotic use in animals and human beings. *The Lancet infectious diseases*, 3(1):47–51, 2003.

[65] Milisav Demerec. Origin of bacterial resistance to antibiotics. *Journal of bacteriology*, 56(1):63, 1948.

[66] William MM Kirby. Extraction of a highly potent penicillin inactivator from penicillin resistant staphylococci. *Science*, 99(2579):452–453, 1944.

[67] Salvador E Luria and Max Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491, 1943.

[68] Luis A Diaz Jr, Richard T Williams, Jian Wu, Isaac Kinde, J Randolph Hecht, Jordan Berlin, Benjamin Allen, Ivana Bozic, Johannes G Reiter, Martin A Nowak, et al. The molecular evolution of acquired resistance to targeted egfr blockade in colorectal cancers. *Nature*, 486(7404):537–540, 2012.

[69] Ron Sorace and Natalia L Komarova. Accumulation of neutral mutations in growing cell colonies with competition. *Journal of theoretical biology*, 314:84–94, 2012.

[70] Anup Dewanji, EG Luebeck, and Suresh H Moolgavkar. A generalized luria–delbrück model. *Mathematical biosciences*, 197(2):140–152, 2005.

[71] Robert A Gatenby, Ariosto S Silva, Robert J Gillies, and B Roy Frieden. Adaptive therapy. *Cancer research*, 69(11):4894–4903, 2009.

[72] Pedro M Enriquez-Navas, Yoonseok Kam, Tuhin Das, Sabrina Hassan, Ariosto Silva, Parastou Foroutan, Epifanio Ruiz, Gary Martinez, Susan Minton, Robert J Gillies, et al. Exploiting evolutionary principles to prolong tumor control in preclinical models of breast cancer. *Science translational medicine*, 8(327):327ra24–327ra24, 2016.

[73] Ariosto S Silva, Yoonseok Kam, Zayar P Khin, Susan E Minton, Robert J Gillies, and Robert A Gatenby. Evolutionary approaches to prolong progression-free survival in breast cancer. *Cancer research*, 72(24):6362–6370, 2012.

[74] Steven A Rosenberg, James C Yang, and Nicholas P Restifo. Cancer immunotherapy: moving beyond current vaccines. *Nature medicine*, 10(9):909, 2004.

[75] Gerald J Spangrude, Shelly Heimfeld, and Irving L Weissman. Purification and characterization of mouse hematopoietic stem cells. *Science*, 241(4861):58–62, 1988.

[76] Sean J Morrison, Nobuko Uchida, and Irving L Weissman. The biology of hematopoietic stem cells. *Annual review of cell and developmental biology*, 11(1):35–71, 1995.

[77] Hans B Sieburg, Betsy D Rezner, and Christa E Muller-Sieburg. Predicting clonal self-renewal and extinction of hematopoietic stem cells. *Proceedings of the National Academy of Sciences*, 108(11):4370–4375, 2011.

[78] DA Fulcher and A Basten. B cell life span: a review. *Immunology and cell biology*, 75(5):446–455, 1997.

[79] C Ricketts, A Jacobs, and I Cavill. Ferrokinetics and erythropoiesis in man: the measurement of effective erythropoiesis, ineffective erythropoiesis and red cell lifespan using 59fe. *British Journal of Haematology*, 31(1):65–75, 1975.

[80] John G Kelton, Cedric J Carter, Claudia Rodger, George Bebenek, Jack Gauldie, David Sheridan, Yasmin B Kassam, WF Kean, WW Buchanan, and PJ Rooney. The relationship among platelet-associated igg, platelet lifespan, and reticuloendothelial cell function. 1984.

[81] Bas Pilzecker, Olimpia Alessandra Buoninfante, Paul van den Berk, Cesare Lancini, Ji-Ying Song, Elisabetta Citterio, and Heinz Jacobs. Dna damage tolerance in hematopoietic stem and progenitor cells in mice. *Proceedings of the National Academy of Sciences*, 114(33):E6875–E6883, 2017.

[82] RK Zhong, Clinton M Astle, and David E Harrison. Distinct developmental patterns of short-term and long-term functioning lymphoid and myeloid precursors defined by competitive limiting dilution analysis in vivo. *The Journal of Immunology*, 157(1):138–145, 1996.

[83] Judith A Shizuru, Robert S Negrin, and Irving L Weissman. Hematopoietic stem and progenitor cells: clinical and preclinical regeneration of the hematolymphoid system. *Annu. Rev. Med.*, 56:509–538, 2005.

[84] Anne Galy, Marilyn Travis, Dazhi Cen, and Benjamin Chen. Human t, b, natural killer, and dendritic cells arise from a common bone marrow progenitor cell subset. *Immunity*, 3(4):459–473, 1995.

[85] Koichi Akashi, David Traver, Toshihiro Miyamoto, and Irving L Weissman. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, 404(6774):193–197, 2000.

[86] Markus G Manz, Toshihiro Miyamoto, Koichi Akashi, and Irving L Weissman. Prospective isolation of human clonogenic common myeloid progenitors. *Proceedings of the National Academy of Sciences*, 99(18):11872–11877, 2002.

[87] Charles A Janeway Jr, Paul Travers, Mark Walport, and Mark J Shlomchik. Principles of innate and adaptive immunity. In *Immunobiology: The Immune System in Health and Disease. 5th edition*. Garland Science, 2001.

[88] Emil R Unanue. Antigen-presenting function of the macrophage. *Annual review of immunology*, 2(1):395–428, 1984.

[89] Alan Aderem and David M Underhill. Mechanisms of phagocytosis in macrophages. *Annual review of immunology*, 17(1):593–623, 1999.

[90] Tamás Rőszer. Understanding the mysterious m2 macrophage through activation markers and effector mechanisms. *Mediators of inflammation*, 2015, 2015.

[91] David G Russell, Lu Huang, and Brian C VanderVen. Immunometabolism at the interface between macrophages and pathogens. *Nature Reviews Immunology*, 19(5):291–304, 2019.

[92] Jennifer W Leiding. Neutrophil evolution and their diseases in humans. *Frontiers in immunology*, 8:1009, 2017.

[93] Mark J Quade and James A Roth. A rapid, direct assay to measure degranulation of bovine neutrophil primary granules. *Veterinary immunology and immunopathology*, 58(3-4):239–248, 1997.

[94] S Sinha, W Watorek, S Karr, J Giles, W Bode, and James Travis. Primary structure of human neutrophil elastase. *Proceedings of the National Academy of Sciences*, 84(8):2228–2232, 1987.

[95] Volker Brinkmann, Ulrike Reichard, Christian Goosmann, Beatrix Fauler, Yvonne Uhlemann, David S Weiss, Yvette Weinrauch, and Arturo Zychlinsky. Neutrophil extracellular traps kill bacteria. *science*, 303(5663):1532–1535, 2004.

[96] Carl Nathan. Neutrophils and immunity: challenges and opportunities. *Nature reviews immunology*, 6(3):173–182, 2006.

[97] Barbara Geering, Christina Stoeckle, Sébastien Conus, and Hans-Uwe Simon. Living and dying for inflammation: neutrophils, eosinophils, basophils. *Trends in immunology*, 34(8):398–409, 2013.

[98] Elizabeth A Jacobsen, Anna G Taranova, Nancy A Lee, and James J Lee. Eosinophils: singularly destructive effector cells or purveyors of immunoregulation? *Journal of Allergy and Clinical Immunology*, 119(6):1313–1320, 2007.

[99] Mark C Siracusa, Brian S Kim, Jonathan M Spergel, and David Artis. Basophils and allergic inflammation. *Journal of Allergy and Clinical Immunology*, 132(4):789–801, 2013.

[100] Eric Vivier, David H Raulet, Alessandro Moretta, Michael A Caligiuri, Laurence Zitvogel, Lewis L Lanier, Wayne M Yokoyama, and Sophie Ugolini. Innate or adaptive immunity? the example of natural killer cells. *Science*, 331(6013):44–49, 2011.

[101] Sara W Bird and Karla Kirkegaard. Escape of non-enveloped virus from intact cells. *Virology*, 479:444–449, 2015.

[102] Jerome Thiery, Dennis Keefe, Steeve Boulant, Emmanuel Boucrot, Michael Walch, Denis Martinvalet, Swie Goping, R Chris Bleackley, Tomas Kirchhausen, and Judy Lieberman. Perforin pores in the endosomal membrane trigger the release of endocytosed granzyme b into the cytosol of target cells. *Nature immunology*, 12(8):770, 2011.

[103] Dorine Sichien, BN Lambrecht, Martin Guilliams, and CL Scott. Development of conventional dendritic cells: from common bone marrow progenitors to multiple subsets in peripheral tissues. *Mucosal immunology*, 10(4):831–844, 2017.

[104] Dirk M Anderson, Eugene Maraskovsky, William L Billingsley, William C Dougall, Mark E Tometsko, Eileen R Roux, Mark C Teepe, Robert F DuBose, David Cosman, and Laurent Galibert. A homologue of the tnf receptor and its ligand enhance t-cell growth and dendritic-cell function. *Nature*, 390(6656):175–179, 1997.

[105] Melissa Swiecki and Marco Colonna. The multifaceted biology of plasmacytoid dendritic cells. *Nature Reviews Immunology*, 15(8):471–485, 2015.

[106] Max D Cooper and Matthew N Alder. The evolution of adaptive immune systems. *Cell*, 124(4):815–822, 2006.

[107] Rafi Ahmed and David Gray. Immunological memory and protective immunity: understanding their relation. *Science*, 272(5258):54–60, 1996.

[108] M Papamichail, JC Brown, and EJ Holborow. Immunoglobulins on the surface of human lymphocytes. *The Lancet*, 298(7729):850–852, 1971.

[109] Christophe Arpin, Julie Dechanet, Cees Van Kooten, Pierre Merville, Geraldine Grouard, Francine Briere, Jacques Banchereau, and Yong-Jun Liu. Generation of memory b cells and plasma cells in vitro. *Science*, 268(5211):720–722, 1995.

[110] Thomas Dörner and Andreas Radbruch. Selecting b cells and plasma cells to memory. *Journal of Experimental Medicine*, 201(4):497–499, 2005.

[111] Jennifer E Smith-Garvin, Gary A Koretzky, and Martha S Jordan. T cell activation. *Annual review of immunology*, 27:591–619, 2009.

[112] Wai-Ping Fung-Leung, Marco W Schilham, Amin Rahemtulla, Thomas M Kündig, Maja Vollenweider, Julia Potter, Willem van Ewijk, and Tak W Mak. Cd8 is needed for development of cytotoxic t but not helper t cells. *Cell*, 65(3):443–449, 1991.

[113] MHC The et al. Complete sequence and gene map of a human major histocompatibility complex. *Nature*, 401(6756):921, 1999.

[114] Olaf Rötzschke, Kirsten Falk, Karl Deres, Hansjörg Schild, Maria Norda, Jörg Metzger, Günther Jung, and Hans-Georg Rammensee. Isolation and analysis of naturally processed viral peptides as recognized by cytotoxic t cells. *Nature*, 348(6298):252–254, 1990.

[115] Adrian F Ochsenbein, Sophie Sierro, Bernhard Odermatt, Marcus Pericin, Urs Karrer, Jan Hermans, Silvio Hemmi, Hans Hengartner, and Rolf M Zinkernagel. Roles of tumour localization, second signals and cross priming in cytotoxic t-cell induction. *Nature*, 411(6841):1058–1064, 2001.

[116] Tim R Mosmann and Robert L Coffman. Heterogeneity of cytokine secretion patterns and functions of helper t cells. In *Advances in immunology*, volume 46, pages 111–147. Elsevier, 1989.

[117] Ethan M Shevach. Regulatory t cells in autoimmmunity. *Annual review of immunology*, 18(1):423–449, 2000.

[118] Deqing Hu and Ali Shilatifard. Epigenetics of hematopoiesis and hematological malignancies. *Genes & development*, 30(18):2021–2041, 2016.

[119] John W Berg. The incidence of multiple primary cancers. i. development of further cancers in patients with lymphomas, leukemias, and myeloma. *Journal of the National Cancer Institute*, 38(5):741–752, 1967.

[120] Marshall A Lichtman. Obesity and the risk for a hematological malignancy: leukemia, lymphoma, or myeloma. *The oncologist*, 15(10):1083, 2010.

[121] Gregory A Abel and Heidi D Klepin. Frailty and the management of hematologic malignancies. *Blood, The Journal of the American Society of Hematology*, 131(5):515–524, 2018.

[122] Ayalew Tefferi and James W Vardiman. Myelodysplastic syndromes. *New England Journal of Medicine*, 361(19):1872–1885, 2009.

[123] Jerry L Spivak. Myeloproliferative neoplasms. *New England Journal of Medicine*, 376(22):2168–2181, 2017.

[124] Ayalew Tefferi, Mythri Mudireddy, Francesco Mannelli, Kebede H Begna, Mrinal M Patnaik, Curtis A Hanson, Rhett P Ketterling, Naseema Gangat, Meera Yogarajah, Valerio De Stefano, et al. Blast phase myeloproliferative neoplasm: Mayo-agimm study of 410 patients from two separate cohorts. *Leukemia*, 32(5):1200–1210, 2018.

[125] Meera Yogarajah and Ayalew Tefferi. Leukemic transformation in myeloproliferative neoplasms: a literature review on risk, characteristics, and outcome. In *Mayo Clinic Proceedings*, volume 92, pages 1118–1128. Elsevier, 2017.

[126] Francisco Cervantes, Dolores Tassies, Camino Salgado, Montserrat Rovira, Arturo Pereira, and Ciril Rozman. Acute transformation in nonleukemic chronic myeloproliferative disorders: actuarial probability and main characteristics in a series of 218 patients. *Acta haematologica*, 85(3):124–127, 1991.

[127] Ayalew Tefferi, Paola Guglielmelli, Dirk R Larson, Christy Finke, Emnet A Wassie, Lisa Pieri, Naseema Gangat, Rajmonda Fjerza, Alem A Belachew, Terra L Lasho, et al. Long-term survival and blast transformation in molecularly annotated essential thrombocythemia, polycythemia vera, and myelofibrosis. *Blood, The Journal of the American Society of Hematology*, 124(16):2507–2513, 2014.

[128] Ayalew Tefferi, E Rumi, G Finazzi, H Gisslinger, AM Vannucchi, F Rodeghiero, ML Randi, R Vaidya, M Cazzola, A Rambaldi, et al. Survival and prognosis among 1545 patients with contemporary polycythemia vera: an international study. *Leukemia*, 27(9):1874–1881, 2013.

[129] Tiziano Barbui, Juergen Thiele, Francesco Passamonti, Elisa Rumi, Emanuela Boveri, Marco Ruggeri, Francesco Rodeghiero, Emanuele SG d'Amore, Maria Luigia Randi, Irene Bertozzi, et al. Survival and disease progression in essential thrombocythemia are significantly influenced by accurate morphologic diagnosis: an international study. *Journal of clinical oncology*, 29(23):3179–3184, 2011.

[130] Carolyn S Grove and George S Vassiliou. Acute myeloid leukaemia: a paradigm for the clonal evolution of cancer? *Disease models & mechanisms*, 7(8):941–951, 2014.

[131] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

[132] Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine*, 368(22):2059–2074, 2013.

[133] Lene Sofie Granfeldt Østgård, Eigil Kjeldsen, Mette Skov Holm, Peter De Nully Brown, Bjarne Bach Pedersen, Knud Bendix, Preben Johansen, Jørgen Schøler Kristensen, and Jan Maxwell Nørgaard. Reasons for treating secondary aml as de novo aml. *European journal of haematology*, 85(3):217–226, 2010.

[134] James L Gajewski, Winston G Ho, Stephen D Nimer, Karim F Hirji, Linda Gekelman, Andrew D Jacobs, and Richard E Champlin. Efficacy of intensive chemotherapy for acute myelogenous leukemia associated with a preleukemic syndrome. *Journal of Clinical Oncology*, 7(11):1637–1645, 1989.

[135] Erik Hulegårdh, Christer Nilsson, Vladimir Lazarevic, Hege Garelius, Petar Antunovic, Åsa Rangert Derolf, Lars Möllgård, Bertil Uggla, Lovisa Wennström, Anders Wahlin, et al. Characterization and prognostic features of secondary acute myeloid leukemia in a population-based setting: A report from the s wedish a cute l eukemia r egistry. *American journal of hematology*, 90(3):208–214, 2015.

[136] David Grimwade, Helen Walker, Georgina Harrison, Fiona Oliver, Stephen Chatters, Christine J Harrison, Keith Wheatley, Alan K Burnett, and Anthony H Goldstone. The predictive value of hierarchical cytogenetic classification in older adults with acute myeloid leukemia (aml): analysis of 1065 patients entered into the united kingdom medical research council aml11 trial. *Blood, The Journal of the American Society of Hematology*, 98(5):1312–1320, 2001.

[137] Ulrike Bacher, Wolfgang Kern, Susanne Schnittger, Wolfgang Hiddemann, Claudia Schoch, and Torsten Haferlach. Further correlations of morphology according to fab and who classification to cytogenetics in de novo acute myeloid leukemia: a study on 2,235 patients. *Annals of hematology*, 84(12):785–791, 2005.

[138] John C Byrd, Krzysztof Mrózek, Richard K Dodge, Andrew J Carroll, Colin G Edwards, Diane C Arthur, Mark J Pettenati, Shivanand R Patil, Kathleen W Rao, Michael S Watson, et al. Pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse, and overall survival in adult

patients with de novo acute myeloid leukemia: results from cancer and leukemia group b (calgb 8461) presented in part at the 43rd annual meeting of the american society of hematology, orlando, fl, december 10, 2001, and published in abstract form. 59. *Blood, The Journal of the American Society of Hematology*, 100(13):4325–4336, 2002.

[139] Nathan J Charles and Daniel F Boyer. Mixed-phenotype acute leukemia: diagnostic criteria and pitfalls. *Archives of pathology & laboratory medicine*, 141(11):1462–1468, 2017.

[140] Thomas B Alexander and Etan Orgel. Mixed phenotype acute leukemia: Current approaches to diagnosis and treatment. *Current Oncology Reports*, 23(2):1–10, 2021.

[141] Runhua Shi and Reinhold Munker. Survival of patients with mixed phenotype acute leukemias: A large population-based study. *Leukemia research*, 39(6):606–616, 2015.

[142] Maria Maruffi, Richard Sposto, Matthew J Oberley, Lynn Kysh, and Etan Orgel. Therapy for children and adults with mixed phenotype acute leukemia: a systematic review and meta-analysis. *Leukemia*, 32(7):1515–1528, 2018.

[143] MC Bene, G Castoldi, W Knapp, Wolf-Dieter Ludwig, Estela Matutes, Alberto Orfao, and MB Van't Veer. Proposals for the immunological classification of acute leukemias. european group for the immunological characterization of leukemias (egil). *Leukemia*, 9(10):1783–1786, 1995.

[144] SH Swerdlow, Elias Campo, N Lee Harris, ES Jaffe, SA Pileri, HWHO Stein, J Thiele, D Arber, R Hasserjian, M Le Beau, et al. Who classification of tumours of haematopoietic and lymphoid tissues (revised 4th edition). *IARC: Lyon*, 421, 2017.

[145] E Joanna Baxter, Linda M Scott, Peter J Campbell, Clare East, Nasios Fourouclas, Soheila Swanton, George S Vassiliou, Anthony J Bench, Elaine M Boyd, Natasha Curtin, et al. Acquired mutation of the tyrosine kinase jak2 in human myeloproliferative disorders. *The Lancet*, 365(9464):1054–1061, 2005.

[146] Ross L Levine, Martha Wadleigh, Jan Cools, Benjamin L Ebert, Gerlinde Wernig, Brian JP Huntly, Titus J Boggon, Iwona Wlodarska, Jennifer J Clark, Sandra Moore, et al. Activating mutation in the tyrosine kinase jak2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer cell*, 7(4):387–397, 2005.

[147] Robert Kralovics, Francesco Passamonti, Andreas S Buser, Soon-Siong Teo, Ralph Tiedt, Jakob R Passweg, Andre Tichelli, Mario Cazzola, and Radek C Skoda. A gain-of-function mutation of jak2 in myeloproliferative disorders. *New England Journal of Medicine*, 352(17):1779–1790, 2005.

[148] Pontus Lundberg, Hitoshi Takizawa, Lucia Kubovcakova, Guoji Guo, Hui Hao-Shen, Stephan Dirnhofer, Stuart H Orkin, Markus G Manz, and Radek C Skoda. Myeloproliferative neoplasms can be initiated from a single hematopoietic stem cell expressing jak2-v617f. *Journal of Experimental Medicine*, 211(11):2213–2230, 2014.

[149] Ann Mullally, Steven W Lane, Brian Ball, Christine Megerdichian, Rachel Okabe, Fatima Al-Shahrour, Mahnaz Paktinat, J Erika Haydu, Elizabeth Housman, Allegra M Lord, et al. Physiological jak2v617f expression causes a lethal myeloproliferative neoplasm with differential effects on hematopoietic stem and progenitor cells. *Cancer cell*, 17(6):584–596, 2010.

[150] Edwin Chen, Rebekka K Schneider, Lawrence J Breyfogle, Emily A Rosen, Luke Poveromo, Shannon Elf, Amy Ko, Kristina Brumme, Ross Levine, Benjamin L Ebert, et al. Distinct effects of concomitant jak2v617f expression and tet2 loss in mice promote disease progression in myeloproliferative neoplasms. *Blood*, 125(2):327–335, 2015.

[151] François Delhommeau, Sabrina Dupont, Véronique Della Valle, Chloe James, Severine Trannoy, Aline Masse, Olivier Kosmider, Jean-Pierre Le Couedic, Fabienne Robert, Antonio Alberdi, et al. Mutation in tet2 in myeloid cancers. *New England Journal of Medicine*, 360(22):2289–2301, 2009.

[152] Saskia MC Langemeijer, Roland P Kuiper, Marieke Berends, Ruth Knops, Mariam G Aslanyan, Marion Massop, Ellen Stevens-Linders, Patricia van Hoogen, Ad Geurts van Kessel, Reinier AP Raymakers, et al. Acquired mutations in tet2 are common in myelodysplastic syndromes. *Nature genetics*, 41(7):838, 2009.

[153] Ayalew Tefferi, KH Lim, O Abdel-Wahab, TL Lasho, J Patel, Mrinal M Patnaik, CA Hanson, A Pardanani, DG Gilliland, and RL Levine. Detection of mutant tet2 in myeloid malignancies other than myeloproliferative neoplasms: Cmml, mds, mds/mpn and aml. *Leukemia*, 23(7):1343–1345, 2009.

[154] L Couronne, E Lippert, J Andrieux, O Kosmider, I Radford-Weiss, D Penther, N Dastugue, F Mugneret, M Lafage, Nathalie Gachard, et al. Analyses of tet2 mutations in post-myeloproliferative neoplasm acute myeloid leukemias. *Leukemia*, 24(1):201–203, 2010.

[155] Wen-Chien Chou, Sheng-Chieh Chou, Chieh-Yu Liu, Chien-Yuan Chen, Hsin-An Hou, Yuan-Yeh Kuo, Ming-Cheng Lee, Bor-Sheng Ko, Jih-Luh Tang, Ming Yao, et al. Tet2 mutation is an unfavorable prognostic factor in acute myeloid leukemia patients with intermediate-risk cytogenetics. *Blood, The Journal of the American Society of Hematology*, 118(14):3803–3810, 2011.

[156] Ayalew Tefferi, A Pardanani, KH Lim, O Abdel-Wahab, TL Lasho, J Patel, N Gangat, CM Finke, S Schwager, A Mullally, et al. Tet2 mutations and their clinical correlates in polycythemia vera, essential thrombocythemia and myelofibrosis. *Leukemia*, 23(5):905–911, 2009.

[157] Myunggon Ko, Hozefa S Bandukwala, Jungeun An, Edward D Lamperti, Elizabeth C Thompson, Ryan Hastie, Angeliki Tsangaratou, Klaus Rajewsky, Sergei B Koralov, and Anjana Rao. Ten-eleven-translocation 2 (tet2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. *Proceedings of the National Academy of Sciences*, 108(35):14566–14571, 2011.

[158] Kelly Moran-Crusio, Linsey Reavie, Alan Shih, Omar Abdel-Wahab, Delphine Ndiaye-Lobry, Camille Lobry, Maria E Figueroa, Aparna Vasanthakumar, Jay Patel, Xinyang

Zhao, et al. Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer cell*, 20(1):11–24, 2011.

[159] K Shide, T Kameda, H Shimoda, T Yamaji, H Abe, A Kamiunten, M Sekine, T Hidaka, K Katayose, Y Kubuki, et al. Tet2 is essential for survival and hematopoietic stem cell homeostasis. *Leukemia*, 26(10):2216, 2012.

[160] Takuro Kameda, Kotaro Shide, Takumi Yamaji, Ayako Kamiunten, Masaaki Sekine, Yasuhiro Taniguchi, Tomonori Hidaka, Yoko Kubuki, Haruko Shimoda, Kousuke Marutsuka, et al. Loss of tet2 has dual roles in murine myeloproliferative neoplasms: disease sustainer and disease accelerator. *Blood*, 125(2):304–315, 2015.

[161] Feng Pan, Thomas S Wingo, Zhigang Zhao, Rui Gao, Hideki Makishima, Guangbo Qu, Li Lin, Miao Yu, Janice R Ortega, Jiapeng Wang, et al. Tet2 loss leads to hypermutagenicity in haematopoietic stem/progenitor cells. *Nature communications*, 8:15102, 2017.

[162] Zhe Li, Xiaoqiang Cai, Chen-Leng Cai, Jiapeng Wang, Wenyong Zhang, Bruce E Petersen, Feng-Chun Yang, and Mingjiang Xu. Deletion of tet2 in mice leads to dysregulated hematopoietic stem cells and subsequent development of myeloid malignancies. *Blood*, 118(17):4509–4518, 2011.

[163] Zhigang Zhao, Shi Chen, Xingguo Zhu, Feng Pan, Rong Li, Yuan Zhou, Weiping Yuan, Hongyu Ni, Feng-Chun Yang, and Mingjiang Xu. The catalytic activity of tet2 is essential for its myeloid malignancy-suppressive function in hematopoietic stem/progenitor cells. *Leukemia*, 30(8):1784, 2016.

[164] Maja Klug, Sandra Schmidhofer, Claudia Gebhard, Reinhard Andreesen, and Michael Rehli. 5-hydroxymethylcytosine is an essential intermediate of active dna demethylation processes in primary human monocytes. *Genome biology*, 14(5):R46, 2013.

[165] Jungeun An, Edahí González-Avalos, Ashu Chawla, Mira Jeong, Isaac F López-Moyado, Wei Li, Margaret A Goodell, Lukas Chavez, Myunggon Ko, and Anjana Rao. Acute loss of tet function results in aggressive myeloid cancer in mice. *Nature communications*, 6:10071, 2015.

[166] Cyril Quivoron, Lucile Couronné, Véronique Della Valle, Cécile K Lopez, Isabelle Plo, Orianne Wagner-Ballon, Marcio Do Cruzeiro, Francois Delhommeau, Bertrand Arnulf, Marc-Henri Stern, et al. Tet2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. *Cancer cell*, 20(1):25–38, 2011.

[167] Maria E Figueroa, Omar Abdel-Wahab, Chao Lu, Patrick S Ward, Jay Patel, Alan Shih, Yushan Li, Neha Bhagwat, Aparna Vasanthakumar, Hugo F Fernandez, et al. Leukemic idh1 and idh2 mutations result in a hypermethylation phenotype, disrupt tet2 function, and impair hematopoietic differentiation. *Cancer cell*, 18(6):553–567, 2010.

[168] Luisa Cimmino, Igor Dolgalev, Yubao Wang, Akihide Yoshimi, Gaëlle H Martin, Jingjing Wang, Victor Ng, Bo Xia, Matthew T Witkowski, Marisa Mitchell-Flack, et al. Restoration of tet2 function blocks aberrant self-renewal and leukemia progression. *Cell*, 170(6):1079–1095, 2017.

[169] Hiroyoshi Kunimoto, Yumi Fukuchi, Masatoshi Sakurai, Ken Sadahira, Yasuo Ikeda, Shinichiro Okamoto, and Hideaki Nakajima. Tet2 disruption leads to enhanced self-renewal and altered differentiation of fetal liver hematopoietic stem cells. *Scientific reports*, 2:273, 2012.

[170] Elodie Pronier, Carole Almire, Hayat Mokrani, Aparna Vasanthakumar, Audrey Simon, Barbara da Costa Reis Monte Mor, Aline Massé, Jean-Pierre Le Couédic, Frédéric Pendino, Bruno Carbonne, et al. Inhibition of tet2-mediated conversion of 5-methylcytosine to 5-hydroxymethylcytosine disturbs erythroid and granulomonocytic differentiation of human hematopoietic progenitors. *Blood*, 118(9):2551–2555, 2011.

[171] Isabelle Plo and William Vainchenker. Molecular and genetic bases of myeloproliferative disorders: questions and perspectives. *Clinical Lymphoma and Myeloma*, 9:S329–S339, 2009.

[172] Juan Li, David G Kent, Anna L Godfrey, Harriet Manning, Jyoti Nangalia, Athar Aziz, Edwin Chen, Kourosh Saeb-Parsy, Juergen Fink, Rachel Sneade, et al. Jak2v617f homozygosity drives a phenotypic switch in myeloproliferative neoplasms, but is insufficient to sustain disease. *Blood*, 123(20):3139–3151, 2014.

[173] Hajime Akada, Dongqing Yan, Haiying Zou, Steven Fiering, Robert E Hutchison, and M Golam Mohi. Conditional expression of heterozygous or homozygous jak2v617f from its endogenous promoter induces a polycythemia vera–like disease. *Blood*, 115(17):3589–3597, 2010.

[174] Edwin Chen and Ann Mullally. How does jak2v617f contribute to the pathogenesis of myeloproliferative neoplasms? *ASH Education Program Book*, 2014(1):268–276, 2014.

[175] Ann Mullally, Luke Poveromo, Rebekka K Schneider, Fatima Al-Shahrour, Steven W Lane, and Benjamin L Ebert. Distinct roles for long-term hematopoietic stem cells and erythroid precursor cells in a murine model of jak2v617f-mediated polycythemia vera. *Blood*, 120(1):166–172, 2012.

[176] Lucia Kubovcakova, Pontus Lundberg, Jean Grisouard, Hui Hao-Shen, Vincent Romanet, Rita Andraos, Masato Murakami, Stephan Dirnhofer, Kay-Uwe Wagner, Thomas Radimerski, et al. Differential effects of hydroxyurea and inc424 on mutant allele burden and myeloproliferative phenotype in a jak2-v617f polycythemia vera mouse model. *Blood*, 121(7):1188–1199, 2013.

[177] Juan Li, Dominik Spensberger, Jong Sook Ahn, Shubha Anand, Philip A Beer, Cedric Ghevaert, Edwin Chen, Ariel Forrai, Linda M Scott, Rita Ferreira, et al. Jak2 v617f impairs hematopoietic stem cell function in a conditional knock-in mouse model of jak2 v617f–positive essential thrombocythemia. *Blood*, 116(9):1528–1538, 2010.

[178] Theodoros Karantanos and Alison R Moliterno. The roles of jak2 in dna damage and repair in the myeloproliferative neoplasms: Opportunities for targeted therapy. *Blood reviews*, 32(5):426–432, 2018.

[179] David G Kent, Juan Li, Hinal Tanna, Juergen Fink, Kristina Kirschner, Dean C Pask, Yvonne Silber, Tina L Hamilton, Rachel Sneade, Benjamin D Simons, et al. Self-renewal of single mouse hematopoietic stem cells is reduced by jak2v617f without compromising progenitor cell expansion. *PLoS biology*, 11(6):e1001576, 2013.

[180] Salma Hasan, Catherine Lacout, Caroline Marty, Marie Cuingnet, Eric Solary, William Vainchenker, and Jean-Luc Villeval. Jak2v617f expression in mice amplifies early hematopoietic cells and gives them a competitive advantage that is hampered by ifnα. *Blood*, 122(8):1464–1477, 2013.

[181] Caroline Marty, Catherine Lacout, Antoine Martin, Salma Hasan, Sylvie Jacquot, Marie-Christine Birling, William Vainchenker, and Jean-Luc Villeval. Myeloproliferative neoplasm induced by constitutive expression of jak2v617f in knock-in mice. *Blood*, 116(5):783–787, 2010.

[182] Zhiwei Ji, Ke Yan, Wenyang Li, Haigen Hu, and Xiaoliang Zhu. Mathematical and computational modeling in complex biological systems. *BioMed research international*, 2017, 2017.

[183] Richard Banks and L Jason Steggles. An abstraction theory for qualitative models of biological systems. *Theoretical Computer Science*, 431:207–218, 2012.

[184] Benedikt von Bronk, Alexandra Götz, and Madeleine Opitz. Complex microbial systems across different levels of description. *Physical biology*, 15(5):051002, 2018.

[185] Kathryn Atwell, Sara-Jane Dunn, James M Osborne, Hillel Kugler, and E Jane Albert Hubbard. How computational models contribute to our understanding of the germ line. *Molecular reproduction and development*, 83(11):944–957, 2016.

[186] Anne Wilson, Elisa Laurenti, Gabriela Oser, Richard C van der Wath, William Blanco-Bose, Maike Jaworski, Sandra Offner, Cyrille F Dunant, Leonid Eshkind, Ernesto Bockamp, et al. Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell*, 135(6):1118–1129, 2008.

[187] Ryan Chuang, Benjamin A Hall, David Benque, Byron Cook, Samin Ishtiaq, Nir Piterman, Alex Taylor, Moshe Vardi, Steffen Koschmieder, Berthold Gottgens, et al. Drug target optimization in chronic myeloid leukemia using innovative computational platform. *Scientific reports*, 5(1):1–9, 2015.

[188] Santiago Schnell, Ramon Grima, and Philip K Maini. Multiscale modeling in biology: New insights into cancer illustrate how mathematical tools are enhancing the understanding of life from the smallest scale to the grandest. *American Scientist*, 95(2):134–142, 2007.

[189] Zhihui Wang, Joseph D Butner, Romica Kerketta, Vittorio Cristini, and Thomas S Deisboeck. Simulating cancer growth with multiscale agent-based modeling. In *Seminars in cancer biology*, volume 30, pages 70–78. Elsevier, 2015.

[190] William Humphrey, Andrew Dalke, Klaus Schulten, et al. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[191] David Benque, Sam Bourton, Caitlin Cockerton, Byron Cook, Jasmin Fisher, Samin Ishtiaq, Nir Piterman, Alex Taylor, and Moshe Y Vardi. Bma: Visual tool for modeling and analyzing biological networks. In *International Conference on Computer Aided Verification*, pages 686–692. Springer, 2012.

[192] Vasiliki Kostiou, Michael WJ Hall, Philip Jones, and Benjamin A Hall. Different responses to cell crowding determine the clonal fitness of p53 and notch inhibiting mutations in squamous epithelia. *BioRxiv*, 2020.

[193] Pnina Dauber-Osguthorpe and Arnold T Hagler. Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there? *Journal of computer-aided molecular design*, 33(2):133–203, 2019.

[194] Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.

[195] Ron O Dror, Morten Ø Jensen, David W Borhani, and David E Shaw. Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. *Journal of General Physiology*, 135(6):555–562, 2010.

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

[197] Gozde Kar, Attila Gursoy, and Ozlem Keskin. Human cancer protein-protein interaction network: a structural perspective. *PLoS computational biology*, 5(12), 2009.

[198] Roland Somogyi and Larry D Greller. The dynamics of molecular networks: applications to therapeutic discovery. *Drug Discovery Today*, 6(24):1267–1277, 2001.

[199] Yuri Mansury, Mark Kimura, Jose Lobo, and Thomas S Deisboeck. Emerging patterns in tumor systems: simulating the dynamics of multicellular clusters with an agent-based spatial agglomeration model. *Journal of Theoretical Biology*, 219(3):343–370, 2002.

[200] An-Shen Qi, Xiang Zheng, Chan-Ying Du, and Bao-Sheng An. A cellular automaton model of cancerous growth. *Journal of theoretical biology*, 161(1):1–12, 1993.

[201] Minki Hwang, Marc Garbey, Scott A Berceli, and Roger Tran-Son-Tay. Rule-based simulation of multi-cellular biological systems—a review of modeling techniques. *Cellular and molecular bioengineering*, 2(3):285–294, 2009.

[202] Anyue Yin, Dirk Jan AR Moes, Johan GC van Hasselt, Jesse J Swen, and Henk-Jan Guchelaar. A review of mathematical models for tumor dynamics and treatment resistance evolution of solid tumors. *CPT: pharmacometrics & systems pharmacology*, 8(10):720–737, 2019.

[203] AR Kansal, Salvatore Torquato, EA Chiocca, and TS Deisboeck. Emergence of a subpopulation in a computational model of tumor growth. *Journal of Theoretical Biology*, 207(3):431–441, 2000.

[204] Elaine L Bearer, John S Lowengrub, Hermann B Frieboes, Yao-Li Chuang, Fang Jin, Steven M Wise, Mauro Ferrari, David B Agus, and Vittorio Cristini. Multiparameter computational modeling of tumor invasion. *Cancer research*, 69(10):4493–4501, 2009.

[205] Christophe Deroulers, Marine Aubert, Mathilde Badoual, and Basil Grammaticos. Modeling tumor cell migration: from microscopic to macroscopic models. *Physical Review E*, 79(3):031917, 2009.

[206] Bastien Chopard, Joris Borgdorff, and Alfons G Hoekstra. A framework for multiscale modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2021):20130378, 2014.

[207] Thomas E Yankeelov, Gary An, Oliver Saut, E Georg Luebeck, Aleksander S Popel, Benjamin Ribba, Paolo Vicini, Xiaobo Zhou, Jared A Weis, Kaiming Ye, et al. Multiscale modeling in clinical oncology: opportunities and barriers to success. *Annals of biomedical engineering*, 44(9):2626–2641, 2016.

[208] Zhihui Wang, Veronika Bordas, and Thomas S Deisboeck. Discovering molecular targets in cancer with multiscale modeling. *Drug development research*, 72(1):45–52, 2011.

[209] Yangjin Kim, Gibin Powathil, Hyunji Kang, Dumitru Trucu, Hyeongi Kim, Sean Lawler, and Mark Chaplain. Strategies of eradicating glioma cells: a multi-scale mathematical model with mir-451-ampk-mtor control. *PloS one*, 10(1), 2015.

[210] John S Lowengrub, Hermann B Frieboes, Fang Jin, Yao-Li Chuang, Xiangrong Li, Paul Macklin, Steven M Wise, and Vittorio Cristini. Nonlinear modelling of cancer: bridging the gap between cells and tumours. *Nonlinearity*, 23(1):R1, 2009.

[211] Jitin Singla, Kyle M McClary, Kate L White, Frank Alber, Andrej Sali, and Raymond C Stevens. Opportunities and challenges in building a spatiotemporal multi-scale model of the human pancreatic β cell. *Cell*, 173(1):11–19, 2018.

[212] Katarzyna A Rejniak. An immersed boundary framework for modelling the growth of individual cells: an application to the early tumour development. *Journal of theoretical biology*, 247(1):186–204, 2007.

[213] S Yu Jessica and Neda Bagheri. Multi-class and multi-scale models of complex biological phenomena. *Current opinion in biotechnology*, 39:167–173, 2016.

[214] Zhihui Wang and Thomas S Deisboeck. Computational modeling of brain tumors: discrete, continuum or hybrid? In *Scientific modeling and simulations*, pages 381–393. Springer, 2008.

[215] Shodhan Rao, Arjan Van der Schaft, Karen Van Eunen, Barbara M Bakker, and Bayu Jayawardhana. A model reduction method for biochemical reaction networks. *BMC systems biology*, 8(1):52, 2014.

[216] Hasitha N Weerasinghe, Pamela M Burrage, Kevin Burrage, and Dan V Nicolau. Mathematical models of cancer cell plasticity. *Journal of Oncology*, 2019, 2019.

[217] Alireza Pourranjbar, Jane Hillston, and Luca Bortolussi. Don't just go with the flow: Cautionary tales of fluid flow approximation. In *Computer Performance Engineering*, pages 156–171. Springer, 2012.

[218] Kristina Anderson, Christoph Lutz, Frederik W Van Delft, Caroline M Bateman, Yanping Guo, Susan M Colman, Helena Kempski, Anthony V Moorman, Ian Titley, John Swansbury, et al. Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature*, 469(7330):356–361, 2011.

[219] Jasmin Fisher and Thomas A Henzinger. Executable cell biology. *Nature biotechnology*, 25(11):1239–1249, 2007.

[220] Jeffrey West and Paul K Newton. Cellular interactions constrain tumor growth. *Proceedings of the National Academy of Sciences*, 116(6):1918–1923, 2019.

[221] Steven S Andrews, Tuan Dinh, Adam P Arkin, and Robert Meyers. Stochastic models of biological processes. *Encyclopedia of Complexity and Systems Science*, 2009:8730–8749, 2009.

[222] Martina Baar, Loren Coquille, Hannah Mayer, Michael Hölzel, Meri Rogava, Thomas Tüting, and Anton Bovier. A stochastic model for immunotherapy of cancer. *Scientific reports*, 6:24169, 2016.

[223] Stephen J Merrill. Stochastic models of tumor growth and the probability of elimination by cytotoxic cells. *Journal of mathematical biology*, 20(3):305–320, 1984.

[224] Grant Lythe and Carmen Molina-París. Some deterministic and stochastic mathematical models of naïve t-cell homeostasis. *Immunological reviews*, 285(1):206–217, 2018.

[225] Sayuri K Hahl and Andreas Kremling. A comparison of deterministic and stochastic modeling approaches for biochemical reaction systems: on fixed points, means, and modes. *Frontiers in genetics*, 7:157, 2016.

[226] Ugo Del Monte. Does the cell number 109 still really fit one gram of tumor tissue? *Cell Cycle*, 8(3):505–506, 2009.

[227] Niko Beerenwinkel, Roland F Schwarz, Moritz Gerstung, and Florian Markowetz. Cancer evolution: mathematical models and computational inference. *Systematic biology*, 64(1):e1–e25, 2015.

[228] Anne Auger, Philippe Chatelain, and Petros Koumoutsakos. R-leaping: Accelerating the stochastic simulation algorithm by reaction leaps. *The Journal of chemical physics*, 125(8):084103, 2006.

[229] Gennady M Zharinov, Oleg A Bogomolov, Natalia Yu Neklasova, Grigory A Raskin, Irina V Chepurnaya, Sergey N Bugrov, and Vladimir N Anisimov. Prognostic value of tumor growth kinetic parameters in prostate cancer patients. *Oncotarget*, 10(49):5020, 2019.

[230] Xiaoxiao Zhang, Holger Fröhlich, Dima Grigoriev, Sergey Vakulenko, Jörg Zimmermann, and Andreas Günter Weber. A simple 3-parameter model for cancer incidences. *Scientific reports*, 8(1):1–12, 2018.

[231] Mohammad Mamunur Rahman, Yusheng Feng, Thomas E Yankeelov, and J Tinsley Oden. A fully coupled space–time multiscale modeling framework for predicting tumor growth. *Computer methods in applied mechanics and engineering*, 320:261–286, 2017.

[232] Michael Welter, Thierry Fredrich, Herbert Rinneberg, and Heiko Rieger. Computational model for tumor oxygenation applied to clinical data on breast tumor hemoglobin concentrations suggests vascular dilatation and compression. *PloS one*, 11(8), 2016.

[233] Kamel Lahouel, Laurent Younes, Ludmila Danilova, Francis M Giardiello, Ralph H Hruban, John Groopman, Kenneth W Kinzler, Bert Vogelstein, Donald Geman, and Cristian Tomasetti. Revisiting the tumorigenesis timeline with a data-driven generative model. *Proceedings of the National Academy of Sciences*, 117(2):857–864, 2020.

[234] Anis Ben Abdessalem, Nikolaos Dervilis, David Wagg, and Keith Worden. Model selection and parameter estimation in structural dynamics using approximate bayesian computation. *Mechanical Systems and Signal Processing*, 99:306–325, 2018.

[235] Yunn-Kuang Chu and Jau-Chuan Ke. Computation approaches for parameter estimation of weibull distribution. *Mathematical and Computational Applications*, 17(1):39–47, 2012.

[236] Richard David Wilkinson. Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, 12(2):129–141, 2013.

[237] Philip B Holden, Neil R Edwards, James Hensman, and Richard D Wilkinson. Abc for climate: dealing with expensive simulators. *Handbook of approximate Bayesian computation*, pages 569–95, 2018.

[238] Eric-Jan Wagenmakers, Michael Lee, Tom Lodewyckx, and Geoffrey J Iverson. Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses*, pages 181–207. Springer, 2008.

[239] Maksat Ashyraliyev, Johannes Jaeger, and Joke G Blom. Parameter estimation and determinability analysis applied to drosophila gap gene circuits. *BMC Systems Biology*, 2(1):83, 2008.

[240] Yoh Iwasa, Martin A Nowak, and Franziska Michor. Evolution of resistance during clonal expansion. *Genetics*, 172(4):2557–2566, 2006.

[241] Christian Frezza and Carla P Martins. From tumor prevention to therapy: empowering p53 to fight back. *Drug Resistance Updates*, 15(5):258–267, 2012.

[242] Lyubomir T Vassilev, Binh T Vu, Bradford Graves, Daisy Carvajal, Frank Podlaski, Zoran Filipovic, Norman Kong, Ursula Kammlott, Christine Lukacs, Christian Klein, et al. In vivo activation of the p53 pathway by small-molecule antagonists of mdm2. *Science*, 303(5659):844–848, 2004.

[243] Loretta L Nielsen and Daniel C Maneval. P53 tumor suppressor gene therapy for cancer. *Cancer gene therapy*, 5(1):52–63, 1998.

[244] Sayed Mohammad Ebrahim Sahraeian, Marghoob Mohiyuddin, Robert Sebra, Hagen Tilgner, Pegah T Afshar, Kin Fai Au, Narges Bani Asadi, Mark B Gerstein, Wing Hung Wong, Michael P Snyder, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum rna-seq analysis. *Nature communications*, 8(1):1–15, 2017.

[245] Shanrong Zhao, Zhan Ye, and Robert Stanton. Misuse of rpkm or tpm normalization when comparing across samples and sequencing protocols. *Rna*, 26(8):903–909, 2020.

[246] Zachary B Abrams, Travis S Johnson, Kun Huang, Philip RO Payne, and Kevin Coombes. A protocol to evaluate rna sequencing normalization methods. *BMC bioinformatics*, 20(24):1–7, 2019.

[247] John Quackenbush. Computational analysis of microarray data. *Nature reviews genetics*, 2(6):418–427, 2001.

[248] Yaxing Zhao, Limsoon Wong, and Wilson Wen Bin Goh. How to do quantile normalization correctly for gene expression data analyses. *Scientific reports*, 10(1):1–11, 2020.

[249] Thomas B Alexander, Zhaohui Gu, Ilaria Iacobucci, Kirsten Dickerson, John K Choi, Beisi Xu, Debbie Payne-Turner, Hiroki Yoshihara, Mignon L Loh, John Horan, et al. The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature*, 562(7727):373–379, 2018.

[250] Koichi Takahashi, Feng Wang, Kiyomi Morita, Yuanqing Yan, Peter Hu, Pei Zhao, Abdallah Abou Zhar, Chang Jiun Wu, Curtis Gumbs, Latasha Little, et al. Integrative genomic analysis of adult mixed phenotype acute leukemia delineates lineage associated molecular subtypes. *Nature communications*, 9(1):1–12, 2018.

[251] Shunsuke Nakagawa. Acute leukemia of ambiguous lineage (alal). In *Pediatric Acute Lymphoblastic Leukemia*, pages 141–149. Springer, 2020.

[252] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[253] Steffen Durinck, Paul T Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4(8):1184, 2009.

[254] Lily A Chylek, Leonard A Harris, Chang-Shung Tung, James R Faeder, Carlos F Lopez, and William S Hlavacek. Rule-based modeling: a computational approach for studying biomolecular site dynamics in cell signaling systems. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 6(1):13–36, 2014.

[255] Melanie I Stefan, Thomas M Bartol, Terrence J Sejnowski, and Mary B Kennedy. Multi-state modeling of biomolecules. *PLoS Comput Biol*, 10(9):e1003844, 2014.

[256] John AP Sekar and James R Faeder. Rule-based modeling of signal transduction: a primer. *Computational Modeling of Signaling Networks*, pages 139–218, 2012.

[257] Federica Ciocchetta and Jane Hillston. Bio-pepa: A framework for the modelling and analysis of biological systems. *Theoretical Computer Science*, 410(33):3065–3084, 2009.

[258] Anastasios-Andreas Georgoulas. Formal language for statistical inference of uncertain stochastic systems. 2016.

[259] Anastasis Georgoulas, Jane Hillston, Dimitrios Milios, and Guido Sanguinetti. Probabilistic programming process algebra. In *International Conference on Quantitative Evaluation of Systems*, pages 249–264. Springer, 2014.

[260] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[261] Xiaomin Deng and Xiaomeng Wang. The application of gillespie algorithm in spreading. In *3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019)*. Atlantis Press, 2019.

[262] John P DeLong and Jean P Gibert. Gillespie eco-evolutionary models (gem s) reveal the role of heritable trait variation in eco-evolutionary dynamics. *Ecology and evolution*, 6(4):935–945, 2016.

[263] Daniel T Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of chemical physics*, 115(4):1716–1733, 2001.

[264] Yang Cao, Daniel T Gillespie, and Linda R Petzold. Avoiding negative populations in explicit poisson tau-leaping. *The Journal of chemical physics*, 123(5):054104, 2005.

[265] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[266] Mikael Sunnåker, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate bayesian computation. *PLoS Comput Biol*, 9(1):e1002803, 2013.

[267] Gloria Isabel Valderrama Bahamondez and Holger Fröhlich. Mcmc techniques for parameters estimation of ode based models in systems biology. *Frontiers in Applied Mathematics and Statistics*, 5:55, 2019.

[268] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

[269] Luca Bortolussi and Jane Hillston. Model checking single agent behaviours by fluid approximation. *Information and Computation*, 242:183–226, 2015.

[270] A Singh. Evaluation metrics for regression models-mae vs mse vs rmse vs rmsle, 2019.

[271] Gary Brassington. Mean absolute error and root mean square error: which is the better metric for assessing model performance? *EGUGA*, page 3574, 2017.

[272] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250, 2014.

[273] Marc A Schaub, Thomas A Henzinger, and Jasmin Fisher. Qualitative networks: a symbolic approach to analyze biological signaling networks. *BMC systems biology*, 1(1):4, 2007.

[274] Byron Cook, Jasmin Fisher, Elzbieta Krepska, and Nir Piterman. Proving stabilization of biological systems. In *International Workshop on Verification, Model Checking, and Abstract Interpretation*, pages 134–149. Springer, 2011.

[275] Amir Pnueli. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pages 46–57. IEEE, 1977.

[276] Emmanuelle Gallet, Matthieu Manceny, Pascale Le Gall, and Paolo Ballarini. Adapting ltl model checking for inferring biological parameters. *Proceedings of the Approches Formelles dans l'Assistance au Développement de Logiciels (AFADL)*, pages 46–60, 2014.

[277] Steven Eker, Merrill Knapp, Keith Laderoute, Patrick Lincoln, and Carolyn Talcott. Pathway logic: Executable models of biological networks. *Electronic Notes in Theoretical Computer Science*, 71(0):144–161, 2004.

[278] Ezio Bartocci and Pietro Lió. Computational modeling, formal analysis, and tools for systems biology. *PLoS computational biology*, 12(1):e1004591, 2016.

[279] Edmund Clarke, Armin Biere, Richard Raimi, and Yunshan Zhu. Bounded model checking using satisfiability solving. *Formal methods in system design*, 19(1):7–34, 2001.

[280] Armin Biere, Cyrille Artho, and Viktor Schuppan. Liveness checking as safety checking. *Electronic Notes in Theoretical Computer Science*, 66(2):160–177, 2002.

[281] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[282] Laura Elena Raileanu and Kilian Stoffel. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1):77–93, 2004.

[283] L Breiman, JH Friedman, RA Olshen, and CJ Stone. Cart. *Classification and Regression Trees, Wadsworth and Brooks/Cole, Monterey, CA*, 1984.

[284] Amichai Painsky and Gregory W Wornell. Bregman divergence bounds and universality properties of the logarithmic loss. *IEEE Transactions on Information Theory*, 66(3):1658–1673, 2019.

[285] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[286] Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient smt solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 337–340. Springer, 2008.

[287] Ashley B Williams and Björn Schumacher. p53 in the dna-damage-repair process. *Cold Spring Harbor perspectives in medicine*, 6(5):a026070, 2016.

[288] Edmund M Clarke Jr, Orna Grumberg, Daniel Kroening, Doron Peled, and Helmut Veith. *Model checking*. MIT press, 2018.

[289] Jerry R Burch, Edmund M Clarke, Kenneth L McMillan, David L Dill, and Lain-Jinn Hwang. Symbolic model checking: 1020 states and beyond. *Information and computation*, 98(2):142–170, 1992.

[290] Gerard J. Holzmann. The model checker spin. *IEEE Transactions on software engineering*, 23(5):279–295, 1997.

[291] Alessandro Cimatti, Edmund Clarke, Enrico Giunchiglia, Fausto Giunchiglia, Marco Pistore, Marco Roveri, Roberto Sebastiani, and Armando Tacchella. Nusmv 2: An opensource tool for symbolic model checking. In *International Conference on Computer Aided Verification*, pages 359–364. Springer, 2002.

[292] Kenneth L McMillan. The smv language. *Cadence Berkeley Labs*, pages 1–49, 1999.

[293] Sheldon B. Akers. Binary decision diagrams. *IEEE Transactions on computers*, (6):509–516, 1978.

[294] Miguel Carrillo, Pedro A Góngora, and David A Rosenblueth. An overview of existing modeling tools making use of model checking in the analysis of biochemical networks. *Frontiers in plant science*, 3:155, 2012.

[295] Aurélien Naldi, Denis Thieffry, and Claudine Chaouiya. Decision diagrams for the representation and analysis of logical models of genetic networks. In *International Conference on Computational Methods in Systems Biology*, pages 233–247. Springer, 2007.

[296] Craig A Tovey. A simplified np-complete satisfiability problem. *Discrete applied mathematics*, 8(1):85–89, 1984.

[297] Peixin Zhong, Margaret Martonosi, Pranav Ashar, and Sharad Malik. Solving boolean satisfiability with dynamic hardware configurations. In *International Workshop on Field Programmable Logic and Applications*, pages 326–335. Springer, 1998.

[298] Mathias Soeken, Robert Wille, Mirco Kuhlmann, Martin Gogolla, and Rolf Drechsler. Verifying uml/ocl models using boolean satisfiability. In *2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010)*, pages 1341–1344. IEEE, 2010.

[299] Jia Hui Liang, Vijay Ganesh, Pascal Poupart, and Krzysztof Czarnecki. Learning rate based branching heuristic for sat solvers. In *International Conference on Theory and Applications of Satisfiability Testing*, pages 123–140. Springer, 2016.

[300] Pey-Chang Kent Lin and Sunil P Khatri. Application of max-sat-based atpg to optimal cancer therapy design. *BMC genomics*, 13(S6):S5, 2012.

[301] Anuj Deshpande and Ritwik Kumar Layek. Fault detection and therapeutic intervention in gene regulatory networks using sat solvers. *BioSystems*, 179:55–62, 2019.

[302] Robert Nieuwenhuis, Albert Oliveras, and Cesare Tinelli. Solving sat and sat modulo theories: From an abstract davis–putnam–logemann–loveland procedure to dpll (t). *Journal of the ACM (JACM)*, 53(6):937–977, 2006.

[303] Isaac Garcia-Murillas, Gaia Schiavon, Britta Weigelt, Charlotte Ng, Sarah Hrebien, Rosalind J Cutts, Maggie Cheang, Peter Osin, Ashutosh Nerurkar, Iwanka Kozarewa, et al. Mutation tracking in circulating tumor dna predicts relapse in early breast cancer. *Science translational medicine*, 7(302):302ra133–302ra133, 2015.

[304] C Athena Aktipis, Virginia SY Kwan, Kathryn A Johnson, Steven L Neuberg, and Carlo C Maley. Overlooking evolution: a systematic analysis of cancer relapse and therapeutic resistance research. *PloS one*, 6(11):e26100, 2011.

[305] Ivana Bozic, Johannes G Reiter, Benjamin Allen, Tibor Antal, Krishnendu Chatterjee, Preya Shah, Yo Sup Moon, Amin Yaqubie, Nicole Kelly, Dung T Le, et al. Evolutionary dynamics of cancer in response to targeted combination therapy. *Elife*, 2:e00747, 2013.

[306] Bartlomiej Waclaw, Ivana Bozic, Meredith E Pittman, Ralph H Hruban, Bert Vogelstein, and Martin A Nowak. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature*, 525(7568):261–264, 2015.

[307] Dan I Andersson and Diarmaid Hughes. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nature Reviews Microbiology*, 8(4):260–271, 2010.

[308] Muftakhidinov Mitchell and Winchen et al. Engauge digitizer software. http://markummitchell.github.io/engauge-digitizer.

[309] R Core Team et al. R: A language and environment for statistical computing, 2013.

[310] John A Spratt, D Von Fournier, John S Spratt, and Ernst E Weber. Decelerating growth and human breast cancer. *Cancer*, 71(6):2013–2019, 1993.

[311] Takuya Iwamoto. Clinical application of drug delivery systems in cancer chemotherapy: review of the efficacy and side effects of approved drugs. *Biological and Pharmaceutical Bulletin*, 36(5):715–718, 2013.

[312] Anne E Kayl and Christina A Meyers. Side-effects of chemotherapy and quality of life in ovarian and breast cancer patients. *Current Opinion in Obstetrics and Gynecology*, 18(1):24–28, 2006.

[313] Ming-jie Jiang, Dian-na Gu, Juan-juan Dai, Qian Huang, and Ling Tian. Dark side of cytotoxic therapy: Chemoradiation-induced cell death and tumor repopulation. *Trends in Cancer*, 2020.

[314] Roger A Kerin, P Rajan Varadarajan, and Robert A Peterson. First-mover advantage: A synthesis, conceptual framework, and research propositions. *Journal of marketing*, 56(4):33–52, 1992.

[315] S Raz, D Sheban, N Gonen, M Stark, B Berman, and YG Assaraf. Severe hypoxia induces complete antifolate resistance in carcinoma cells due to cell cycle arrest. *Cell death & disease*, 5(2):e1067, 2014.

[316] Thomas J Sweeney, Volker Mailänder, Amanda A Tucker, Adesuwa B Olomu, Weisheng Zhang, Yu-an Cao, Robert S Negrin, and Christopher H Contag. Visualizing the kinetics of tumor-cell clearance in living animals. *Proceedings of the National Academy of Sciences*, 96(21):12044–12049, 1999.

[317] Jason Aynardi, Rashmi Manur, Paul R Hess, Seble Chekol, Jennifer JD Morrissette, Daria Babushok, Elizabeth Hexner, Heesun J Rogers, Eric D Hsi, Elizabeth Margolskee, et al. Jak2 v617f-positive acute myeloid leukaemia (aml): a comparison between de novo aml and secondary aml transformed from an underlying myeloproliferative neoplasm. a study from the bone marrow pathology group. *British journal of haematology*, 182(1):78–85, 2018.

[318] R Coleman Lindsley. Uncoding the genetic heterogeneity of myelodysplastic syndrome. *Hematology 2014, the American Society of Hematology Education Program Book*, 2017(1):447–452, 2017.

[319] Matthew A Clarke, Steven Woodhouse, Nir Piterman, Benjamin A Hall, and Jasmin Fisher. Using state space exploration to determine how gene regulatory networks constrain mutation order in cancer evolution. In *Automated reasoning for systems biology and medicine*, pages 133–153. Springer, 2019.

[320] Maryline Herbet, Aude Salomon, Jean-Jacques Feige, and Michael Thomas. Acquisition order of ras and p53 gene alterations defines distinct adrenocortical tumor phenotypes. *PLoS genetics*, 8(5), 2012.

[321] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90, 2011.

[322] Kathleen Sprouffske, John W Pepper, and Carlo C Maley. Accurate reconstruction of the temporal order of mutations in neoplastic progression. *Cancer prevention research*, 4(7):1135–1144, 2011.

[323] Yong Wang, Jill Waters, Marco L Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160, 2014.

[324] Charles Swanton. Cancer evolution constrained by mutation order. *New England Journal of Medicine*, 372(7):661–663, 2015.

[325] Christina A Ortmann, David G Kent, Jyoti Nangalia, Yvonne Silber, David C Wedge, Jacob Grinfeld, E Joanna Baxter, Charles E Massie, Elli Papaemmanuil, Suraj Menon, et al. Effect of mutation order on myeloproliferative neoplasms. *New England Journal of Medicine*, 372(7):601–612, 2015.

[326] Christian Bach, Sebastian Buhl, Dorothée Mueller, María-Paz García-Cuéllar, Emanuel Maethner, and Robert K Slany. Leukemogenic transformation by hoxa cluster genes. *Blood, The Journal of the American Society of Hematology*, 115(14):2910–2918, 2010.

[327] Edward B Lewis. A gene complex controlling segmentation in drosophila. In *Genes, development and cancer*, pages 205–217. Springer, 1978.

[328] Seema Bhatlekar, Jeremy Z Fields, and Bruce M Boman. Hox genes and their role in the development of human cancers. *Journal of molecular medicine*, 92(8):811–823, 2014.

[329] Raed A Alharbi, Ruth Pettengell, Hardev S Pandha, and Richard Morgan. The role of hox genes in normal hematopoiesis and acute leukemia. *Leukemia*, 27(5):1000–1008, 2013.

[330] Elizabeth Eklund. The role of hox proteins in leukemogenesis: insights into key regulatory events in hematopoiesis. *Critical Reviews™ in Oncogenesis*, 16(1-2), 2011.

[331] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

[332] Joerg Faber, Andrei V Krivtsov, Matthew C Stubbs, Renee Wright, Tina N Davis, Marry van den Heuvel-Eibrink, Christian M Zwaan, Andrew L Kung, and Scott A Armstrong. Hoxa9 is required for survival in human mll-rearranged acute leukemias. *Blood, The Journal of the American Society of Hematology*, 113(11):2375–2385, 2009.

[333] Y Shima, M Yumoto, T Katsumoto, and I Kitabayashi. Mll is essential for nup98-hoxa9-induced leukemia. *Leukemia*, 31(10):2200–2210, 2017.

[334] Xiaoxia Zhong, Andreas Prinz, Julia Steger, Maria-Paz Garcia-Cuellar, Markus Radsak, Abderrazzak Bentaher, and Robert K Slany. Hoxa9 transforms murine myeloid cells by a feedback loop driving expression of key oncogenes and cell cycle control genes. *Blood advances*, 2(22):3137–3148, 2018.

[335] Terry M Therneau and Thomas Lumley. Package ‘survival’. *R Top Doc*, 128(10):28–33, 2015.

[336] Jose Ameijeiras-Alonso, Rosa M Crujeiras, and Alberto Rodríguez-Casal. multimode: An r package for mode assessment. *arXiv preprint arXiv:1803.00472*, 2018.

[337] Katerina Rejlova, Alena Musilova, Karolina Skvarova Kramarzova, Marketa Zaliova, Karel Fiser, Meritxell Alberich-Jorda, Jan Trka, and Julia Starkova. Low hox gene expression in pml-rarα-positive leukemia results from suppressed histone demethylation. *Epigenetics*, 13(1):73–84, 2018.

[338] A Lasa, MJ Carnicer, A Aventin, C Estivill, S Brunet, J Sierra, and JF Nomdedeu. Meis 1 expression is downregulated through promoter hypermethylation in aml1-eto acute myeloid leukemias. *Leukemia*, 18(7):1231–1237, 2004.

[339] Zhen-Yi Wang and Zhu Chen. Acute promyelocytic leukemia: from highly fatal to highly curable. *Blood*, 111(5):2505–2515, 2008.

[340] H Jeffrey Lawrence, Julie Christensen, Stephen Fong, Yu-Long Hu, Irving Weissman, Guy Sauvageau, R Keith Humphries, and Corey Largman. Loss of expression of the hoxa-9 homeobox gene impairs the proliferation and repopulating ability of hematopoietic stem cells. *Blood*, 106(12):3988–3994, 2005.

[341] Katherine R Calvo, Paul S Knoepfler, David B Sykes, Martina P Pasillas, and Mark P Kamps. Meis1a suppresses differentiation by g-csf and promotes proliferation by scf: potential mechanisms of cooperativity with hoxa9 in myeloid leukemia. *Proceedings of the National Academy of Sciences*, 98(23):13120–13125, 2001.

[342] Yasmin Z Paterson, David Shorthouse, Markus W Pleijzier, Nir Piterman, Claus Bendtsen, Benjamin A Hall, and Jasmin Fisher. A toolbox for discrete modelling of cell signalling dynamics. *Integrative Biology*, 10(6):370–382, 2018.

[343] John J Tyson, Katherine C Chen, and Bela Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current opinion in cell biology*, 15(2):221–231, 2003.

[344] Wen Xiong and James E Ferrell. A positive-feedback-based bistable ‘memory module’that governs a cell fate decision. *Nature*, 426(6965):460–465, 2003.

[345] Shengming Zhao, Karen Zoller, Masayoshi Masuko, Ponlapat Rojnuckarin, Xuexian O Yang, Evan Parganas, Kenneth Kaushansky, James N Ihle, Thalia Papayannopoulou, Dennis M Willerford, et al. Jak2, complemented by a second signal from c-kit or flt-3, triggers extensive self-renewal of primary multipotential hemopoietic cells. *The EMBO journal*, 21(9):2159–2167, 2002.

[346] Yongsheng Huang, Kajal Sitwala, Joel Bronstein, Daniel Sanders, Monisha Dandekar, Cailin Collins, Gordon Robertson, James MacDonald, Timothee Cezard, Misha Bilenky, et al. Identification and characterization of hoxa9 binding sites in hematopoietic cells. *Blood, The Journal of the American Society of Hematology*, 119(2):388–398, 2012.

[347] Charles E de Bock, Sofie Demeyer, Sandrine Degryse, Delphine Verbeke, Bram Sweron, Olga Gielen, Roel Vandepoel, Carmen Vicente, Marlies Vanden Bempt, Antonis Dagklis, et al. Hoxa9 cooperates with activated jak/stat signaling to drive leukemia development. *Cancer discovery*, 8(5):616–631, 2018.

[348] L Bei, C Shah, Hao Wang, Weiqi Huang, Leonidas C Platanias, and Elizabeth A Eklund. Regulation of cdx4 gene transcription by hoxa9, hoxa10, the mll-ell oncogene and shp2 during leukemogenesis. *Oncogenesis*, 3(12):e135–e135, 2014.

[349] Renu Kakar, Bryan Kautz, and Elizabeth A Eklund. Jak2 is necessary and sufficient for interferon-γ-induced transcription of the gene encoding gp91phox. *Journal of leukocyte biology*, 77(1):120–127, 2005.

[350] Christina M Ferrell, Sheri T Dorsam, Hideaki Ohta, R Keith Humphries, Mika Kakefuda Derynck, Chris Haqq, Corey Largman, and H Jeffrey Lawrence. Activation of stem-cell specific genes by hoxa9 and hoxa10 homeodomain proteins in cd34+ human cord blood cells. *Stem Cells*, 23(5):644–655, 2005.

[351] Michael T Bocker, Francesca Tuorto, Günter Raddatz, Tanja Musch, Feng-Chun Yang, Mingjiang Xu, Frank Lyko, and Achim Breiling. Hydroxylation of 5-methylcytosine by tet2 maintains the active state of the mammalian hoxa cluster. *Nature communications*, 3(1):1–12, 2012.

[352] Hao Huang, Xi Jiang, Zejuan Li, Yuanyuan Li, Chun-Xiao Song, Chunjiang He, Miao Sun, Ping Chen, Sandeep Gurbuxani, Jiapeng Wang, et al. Tet1 plays an essential oncogenic role in mll-rearranged leukemia. *Proceedings of the National Academy of Sciences*, 110(29):11994–11999, 2013.

[353] Omar Abdel-Wahab, Ann Mullally, Cyrus Hedvat, Guillermo Garcia-Manero, Jay Patel, Martha Wadleigh, Sebastien Malinge, JinJuan Yao, Outi Kilpivaara, Rukhmi Bhat, et al. Genetic characterization of tet1, tet2, and tet3 alterations in myeloid malignancies. *Blood*, 114(1):144–147, 2009.

[354] Kasper Dindler Rasmussen and Kristian Helin. Role of tet enzymes in dna methylation, development, and cancer. *Genes & development*, 30(7):733–750, 2016.

[355] Frank Rosenbauer and Daniel G Tenen. Transcription factors in myeloid development: balancing differentiation with transformation. *Nature Reviews Immunology*, 7(2):105–117, 2007.

[356] Pavel Burda, Nikola Curik, Juraj Kokavec, Petra Basova, Dana Mikulenkova, Arthur I Skoultchi, Jiri Zavadil, and Tomas Stopka. Pu. 1 activation relieves gata-1–mediated repression of cebpa and cbfb during leukemia differentiation. *Molecular Cancer Research*, 7(10):1693–1703, 2009.

[357] Sachin Pundhir, Felicia Kathrine Bratt Lauridsen, Mikkel Bruhn Schuster, Janus Schou Jakobsen, Ying Ge, Erwin Marten Schoof, Nicolas Rapin, Johannes Waage, Marie Sigurd Hasemann, and Bo Torben Porse. Enhancer and transcription factor dynamics during myeloid differentiation reveal an early differentiation block in cebpa null progenitors. *Cell reports*, 23(9):2744–2757, 2018.

[358] Pu Zhang, Junko Iwasaki-Arai, Hiromi Iwasaki, Maris L Fenyus, Tajhal Dayaram, Bronwyn M Owens, Hirokazu Shigematsu, Elena Levantini, Claudia S Huettner, Julie A Lekstrom-Himes, et al. Enhancement of hematopoietic stem cell repopulating capacity and self-renewal in the absence of the transcription factor c/ebpα. *Immunity*, 21(6):853–863, 2004.

[359] Alan D Friedman. Runx1, c-myb, and c/ebpα couple differentiation to proliferation or growth arrest during hematopoiesis. *Journal of cellular biochemistry*, 86(4):624–629, 2002.

[360] Yen K Lieu and E Premkumar Reddy. Impaired adult myeloid progenitor cmp and gmp cell function in conditional c-myb-knockout mice. *Cell Cycle*, 11(18):3504–3512, 2012.

[361] Bo T Porse, David Bryder, Kim Theilgaard-Monch, Marie S Hasemann, Kristina Anderson, Inge Damgaard, Sten Eirik W Jacobsen, and Claus Nerlov. Loss of c/ebpα cell cycle control increases myeloid progenitor proliferation and transforms the neutrophil granulocyte lineage. *The Journal of experimental medicine*, 202(1):85–96, 2005.

[362] Teresa Bellon, Danilo Perrotti, and Bruno Calabretta. Granulocytic differentiation of normal hematopoietic precursor cells induced by transcription factor pu. 1 correlates with negative regulation of the c-myb promoter. *Blood, The Journal of the American Society of Hematology*, 90(5):1828–1839, 1997.

[363] Bo T Porse, Thomas Å Pedersen, Xiufeng Xu, Bo Lindberg, Ulla M Wewer, Lennart Friis-Hansen, and Claus Nerlov. E2f repression by c/ebpα is required for adipogenesis and granulopoiesis in vivo. *Cell*, 107(2):247–258, 2001.

[364] Lisa M Johansen, Atsushi Iwama, Tracey A Lodie, Koichi Sasaki, Dean W Felsher, Todd R Golub, and Daniel G Tenen. c-myc is a critical target for c/ebpα in granulopoiesis. *Molecular and cellular biology*, 21(11):3789–3806, 2001.

[365] Xu Han, Jieying Zhang, Yuanliang Peng, Minyuan Peng, Xiao Chen, Huiyong Chen, Jianhui Song, Xiao Hu, Mao Ye, Jianglin Li, et al. Unexpected role for p19ink4d in posttranscriptional regulation of gata1 and modulation of human terminal erythropoiesis. *Blood, The Journal of the American Society of Hematology*, 129(2):226–237, 2017.

[366] Hans Neubauer, Ana Cumano, Mathias Müller, Hong Wu, Ulrike Huffstadt, and Klaus Pfeffer. Jak2 deficiency defines an essentialdevelopmental checkpoint in definitivehematopoiesis. *Cell*, 93(3):397–409, 1998.

[367] Felix Lohmann and James J Bieker. Activation of eklf expression during hematopoiesis by gata2 and smad5 prior to erythroid commitment. *Development*, 135(12):2071–2082, 2008.

[368] Jong Jin Jeong, Xiaorong Gu, Ji Nie, Sriram Sundaravel, Hui Liu, Wen-Liang Kuo, Tushar D Bhagat, Kith Pradhan, John Cao, Sangeeta Nischal, et al. Cytokine-regulated phosphorylation and activation of tet2 by jak2 in hematopoiesis. *Cancer discovery*, 9(6):778–795, 2019.

[369] Ifat Geron, Annelie E Abrahamsson, Charlene F Barroga, Edward Kavalerchik, Jason Gotlib, John D Hood, Jeffrey Durocher, Chi Ching Mak, Glenn Noronha, Richard M Soll, et al. Selective inhibition of jak2-driven erythroid differentiation of polycythemia vera progenitors. *Cancer cell*, 13(4):321–330, 2008.

[370] Merav Socolovsky, Amy EJ Fallon, Carlo Brugnara, and Harvey F Lodish. Fetal anemia and apoptosis of red cell progenitors in stat5a-/- 5b-/- mice: a direct role for stat5 in bcl-xl induction. *Cell*, 98(2):181–191, 1999.

[371] Masaki Mori, Mie Uchida, Tomoko Watanabe, Keita Kirito, Kiyohiko Hatake, Keiya Ozawa, and Norio Komatsu. Activation of extracellular signal-regulated kinases erk1 and erk2 induces bcl-xl up-regulation via inhibition of caspase activities in erythropoietin signaling. *Journal of cellular physiology*, 195(2):290–297, 2003.

[372] Athanasia D Panopoulos, David Bartos, Ling Zhang, and Stephanie S Watowich. Control of myeloid-specific integrin αmβ2 (cd11b/cd18) expression by cytokines is regulated by stat3-dependent activation of pu. 1. *Journal of Biological Chemistry*, 277(21):19001–19007, 2002.

[373] Pu Zhang, Gerhard Behre, Jing Pan, Atsushi Iwama, Nawarat Wara-Aswapati, Hanna S Radomska, Philip E Auron, Daniel G Tenen, and Zijie Sun. Negative cross-talk between hematopoietic regulators: Gata proteins repress pu. 1. *Proceedings of the National Academy of Sciences*, 96(15):8705–8710, 1999.

[374] P Burda, P Laslo, and T Stopka. The role of pu. 1 and gata-1 transcription factors during normal and leukemogenic hematopoiesis. *Leukemia*, 24(7):1249–1257, 2010.

[375] Shinobu Tsuzuki and Masao Seto. Expansion of functionally defined mouse hematopoietic stem and progenitor cells by a short isoform of runx1/aml1. *Blood, The Journal of the American Society of Hematology*, 119(3):727–735, 2012.

[376] Valeria Azcoitia, Miguel Aracil, Carlos Martínez-A, and Miguel Torres. The homeodomain protein meis1 is essential for definitive hematopoiesis and vascular patterning in the mouse embryo. *Developmental biology*, 280(2):307–320, 2005.

[377] Joseph D Growney, Hirokazu Shigematsu, Zhe Li, Benjamin H Lee, Jennifer Adelsperger, Rebecca Rowan, David P Curley, Jeffery L Kutok, Koichi Akashi, Ifor R Williams, et al. Loss of runx1 perturbs adult hematopoiesis and is associated with a myeloproliferative phenotype. *Blood*, 106(2):494–504, 2005.

[378] Kamaleldin E Elagib, Frederick K Racke, Michael Mogass, Rina Khetawat, Lorrie L Delehanty, and Adam N Goldfarb. Runx1 and gata-1 coexpression and cooperation in megakaryocytic differentiation. *blood*, 101(11):4333–4341, 2003.

[379] Tomomasa Yokomizo, Kazuteru Hasegawa, Hiroyuki Ishitobi, Motomi Osato, Masatsugu Ema, Yoshiaki Ito, Masayuki Yamamoto, and Satoru Takahashi. Runx1 is involved in primitive erythropoiesis in the mouse. *Blood*, 111(8):4075–4080, 2008.

[380] Martha S Petrovick, Scott W Hiebert, Alan D Friedman, Christopher J Hetherington, Daniel G Tenen, and Dong-Er Zhang. Multiple functional domains of aml1: Pu. 1 and c/ebpα synergize with different regions of aml1. *Molecular and cellular biology*, 18(7):3915–3925, 1998.

[381] Veerendra Munugalavadla, Louis C Dore, Bai Lin Tan, Li Hong, Melanie Vishnu, Mitchell J Weiss, and Reuben Kapur. Repression of c-kit and its downstream substrates by gata-1 inhibits cell proliferation during erythroid maturation. *Molecular and cellular biology*, 25(15):6747–6759, 2005.

[382] Amin Al-Shami, Wahib Mahanna, and Paul H Naccache. Granulocyte-macrophage colony-stimulating factor-activated signaling pathways in human neutrophils: selective activation of jak2, stat3, and stat5b. *Journal of Biological Chemistry*, 273(2):1058–1063, 1998.

[383] Warren J Leonard and John J O'Shea. Jaks and stats: biological implications. *Annual review of immunology*, 16(1):293–322, 1998.

[384] Hsin-Lien Huang, Ming-Ju Hsieh, Ming-Hsien Chien, Hui-Yu Chen, Shun-Fa Yang, and Pei-Ching Hsiao. Glabridin mediate caspases activation and induces apoptosis through jnk1/2 and p38 mapk pathway in human promyelocytic leukemia cells. *PLoS One*, 9(6):e98943, 2014.

[385] Ewen Gallagher, Min Gao, Yun-Cai Liu, and Michael Karin. Activation of the e3 ubiquitin ligase itch through a phosphorylation-induced conformational change. *Proceedings of the National Academy of Sciences*, 103(6):1717–1722, 2006.

[386] Patricia Chastagner, Alain Israel, and Christel Brou. Aip4/itch regulates notch receptor degradation in the absence of ligand. *PloS one*, 3(7), 2008.

[387] Apostolos Klinakis, Camille Lobry, Omar Abdel-Wahab, Philmo Oh, Hiroshi Haeno, Silvia Buonamici, Inge van De Walle, Severine Cathelin, Thomas Trimarchi, Elisa Araldi, et al. A novel tumour-suppressor function for the notch pathway in myeloid leukaemia. *Nature*, 473(7346):230–233, 2011.

[388] Catriona HM Jamieson, Jason Gotlib, Jeffrey A Durocher, Mark P Chao, M Rajan Mariappan, Marla Lay, Carol Jones, James L Zehnder, Stan L Lilleberg, and Irving L Weissman. The jak2 v617f mutation occurs in hematopoietic stem cells in polycythemia vera and predisposes toward erythroid differentiation. *Proceedings of the National Academy of Sciences*, 103(16):6224–6229, 2006.

[389] Renata Mendes de Freitas and Carlos Magno da Costa Maranduba. Myeloproliferative neoplasms and the jak/stat signaling pathway: an overview. *Revista brasileira de hematologia e hemoterapia*, 37(5):348–353, 2015.

[390] Jason S Rawlings, Kristin M Rosler, and Douglas A Harrison. The jak/stat signaling pathway. *Journal of cell science*, 117(8):1281–1283, 2004.

[391] Wei Zhao, Claire Kitidis, Mark D Fleming, Harvey F Lodish, and Saghi Ghaffari. Erythropoietin stimulates phosphorylation and activation of gata-1 via the pi3-kinase/akt signaling pathway. *Blood*, 107(3):907–915, 2006.

[392] Motoshi Ichikawa, Takashi Asai, Toshiki Saito, Go Yamamoto, Sachiko Seo, Ieharu Yamazaki, Tetsuya Yamagata, Kinuko Mitani, Shigeru Chiba, Hisamaru Hirai, et al. Aml-1 is required for megakaryocytic maturation and lymphocytic differentiation, but not for maintenance of hematopoietic stem cells in adult hematopoiesis. *Nature medicine*, 10(3):299–304, 2004.

[393] Motoshi Ichikawa, Susumu Goyama, Takashi Asai, Masahito Kawazu, Masahiro Nakagawa, Masataka Takeshita, Shigeru Chiba, Seishi Ogawa, and Mineo Kurokawa. Aml1/runx1 negatively regulates quiescent hematopoietic stem cells in adult hematopoiesis. *The Journal of Immunology*, 180(7):4402–4408, 2008.

[394] Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, 2016.

[395] F Cervantes, F Passamonti, and G Barosi. Life expectancy and prognostic factors in the classic bcr/abl-negative myeloproliferative disorders. *Leukemia*, 22(5):905–914, 2008.

[396] Jon C Aster, Warren S Pear, and Stephen C Blacklow. The varied roles of notch in cancer. *Annual Review of Pathology: Mechanisms of Disease*, 12:245–275, 2017.

[397] Fabio Pereira Lampreia, Joana Gonçalves Carmelo, and Fernando Anjos-Afonso. Notch signaling in the regulation of hematopoietic stem cell. *Current stem cell reports*, 3(3):202–209, 2017.

[398] David R Howlett, Kevin H Jennings, David C Lee, Michael SG Clark, Frank Brown, Ronald Wetzel, Stephen J Wood, Patrick Camilleri, and Gareth W Roberts. Aggregation state and neurotoxic properties of alzheimer beta-amyloid peptide. *Neurodegeneration*, 4(1):23–32, 1995.

[399] Vivian W Chow, Mark P Mattson, Philip C Wong, and Marc Gleichmann. An overview of app processing enzymes and products. *Neuromolecular medicine*, 12(1):1–12, 2010.

[400] Christian Haass, Christoph Kaether, Gopal Thinakaran, and Sangram Sisodia. Trafficking and proteolytic processing of app. *Cold Spring Harbor perspectives in medicine*, 2(5):a006270, 2012.

[401] Hassan Bukhari, Annika Glotzbach, Katharina Kolbe, Gregor Leonhardt, Christina Loosse, and Thorsten Müller. Small things matter: Implications of app intracellular domain aicd nuclear signaling in the progression and pathogenesis of alzheimer's disease. *Progress in neurobiology*, 156:189–213, 2017.

[402] Dennis J Selkoe and John Hardy. The amyloid hypothesis of alzheimer's disease at 25 years. *EMBO molecular medicine*, 8(6):595–608, 2016.

[403] Steven J Roeters, Aditya Iyer, Galja Pletikapić, Vladimir Kogan, Vinod Subramaniam, and Sander Woutersen. Evidence for intramolecular antiparallel beta-sheet structure in alpha-synuclein fibrils from a combination of two-dimensional infrared spectroscopy and atomic force microscopy. *Scientific reports*, 7(1):1–11, 2017.

[404] John Hardy and Dennis J Selkoe. The amyloid hypothesis of alzheimer's disease: progress and problems on the road to therapeutics. *science*, 297(5580):353–356, 2002.

[405] Dennis J Selkoe, Marcia Berman Podlisny, Catharine L Joachim, Elizabeth A Vickers, Gloria Lee, Lawrence C Fritz, and Tilman Oltersdorf. Beta-amyloid precursor protein of alzheimer disease occurs as 110-to 135-kilodalton membrane-associated proteins in neural and nonneural tissues. *Proceedings of the National Academy of Sciences*, 85(19):7341–7345, 1988.

[406] Kendra L Puig and Colin K Combs. Expression and function of app and its metabolites outside the central nervous system. *Experimental gerontology*, 48(7):608–611, 2013.

[407] Yong-Ho Lee, William G Tharp, Rhonda L Maple, Saraswathy Nair, Paska A Permana, and Richard E Pratley. Amyloid precursor protein expression is upregulated in adipocytes in obesity. *Obesity*, 16(7):1493–1500, 2008.

[408] Poomy Pandey, Bailee Sliker, Haley L Peters, Amit Tuli, Jonathan Herskovitz, Kaitlin Smits, Abhilasha Purohit, Rakesh K Singh, Jixin Dong, Surinder K Batra, et al. Amyloid precursor protein and amyloid precursor-like protein 2 in cancer. *Oncotarget*, 7(15):19430, 2016.

[409] Michelle G Botelho, Xiaolei Wang, Donna J Arndt-Jovin, Dorothea Becker, and Thomas M Jovin. Induction of terminal differentiation in melanoma cells on down-regulation of β-amyloid precursor protein. *Journal of investigative dermatology*, 130(5):1400–1410, 2010.

[410] M Anna Kowalska and Karen Badellino. β-amyloid protein induces platelet aggregation and supports platelet adhesion. *Biochemical and biophysical research communications*, 205(3):1829–1835, 1994.

[411] Cindy M Sondag and Colin K Combs. Adhesion of monocytes to type i collagen stimulates an app-dependent proinflammatory signaling response and release of aβ1-40. *Journal of neuroinflammation*, 7(1):22, 2010.

[412] Scott A Armstrong, Jane E Staunton, Lewis B Silverman, Rob Pieters, Monique L den Boer, Mark D Minden, Stephen E Sallan, Eric S Lander, Todd R Golub, and Stanley J Korsmeyer. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature genetics*, 30(1):41–47, 2002.

[413] Yoshitomo Maesako, Takashi Uchiyama, and Hitoshi Ohno. Comparison of gene expression profiles of lymphoma cell lines from transformed follicular lymphoma, burkitt's lymphoma and de novo diffuse large b-cell lymphoma. *Cancer science*, 94(9):774–781, 2003.

[414] Claudia D Baldus, Sandya Liyanarachchi, Krzysztof Mrózek, Herbert Auer, Stephan M Tanner, Martin Guimond, Amy S Ruppert, Nehad Mohamed, Ramana V Davuluri, Michael A Caligiuri, et al. Acute myeloid leukemia with complex karyotypes and abnormal chromosome 21: Amplification discloses overexpression of app, ets2, and erg genes. *Proceedings of the National Academy of Sciences*, 101(11):3915–3920, 2004.

[415] James W Vardiman, Jüergen Thiele, Daniel A Arber, Richard D Brunning, Michael J Borowitz, Anna Porwit, Nancy Lee Harris, Michelle M Le Beau, Eva Hellström-Lindberg, Ayalew Tefferi, et al. The 2008 revision of the world health organization (who) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood*, 114(5):937–951, 2009.

[416] Sabrina Opatz, Stefanos A Bamopoulos, Klaus H Metzeler, Tobias Herold, Bianka Ksienzyk, Kathrin Bräundl, Sebastian Tschuri, Sebastian Vosberg, Nikola P Konstandin, Christine Wang, et al. The clinical mutatome of core binding factor leukemia. *Leukemia*, pages 1–10, 2020.

[417] Guopan Yu, Changxin Yin, Ling Jiang, Dan Xu, Zhongxin Zheng, Zhixiang Wang, Chunli Wang, Hongsheng Zhou, Xuejie Jiang, Qifa Liu, et al. Amyloid precursor protein has clinical and prognostic significance in aml1-eto-positive acute myeloid leukemia. *Oncology letters*, 15(1):917–925, 2018.

[418] Xinrui Li, Jianming Wu, Travis Ptacek, David T Redden, Elizabeth E Brown, Graciela S Alarcón, Rosalind Ramsey-Goldman, Michelle A Petri, John D Reveille, Richard A Kaslow, et al. Allelic-dependent expression of an activating fc receptor on b cells enhances humoral immune responses. *Science translational medicine*, 5(216):216ra175–216ra175, 2013.

[419] Yoshinori Nagai, Rintaro Shimazu, Hirotaka Ogata, Sachiko Akashi, Katsuko Sudo, Hidetoshi Yamasaki, Shin-Ichi Hayashi, Yoichiro Iwakura, Masao Kimoto, and Kensuke Miyake. Requirement for md-1 in cell surface expression of rp105/cd180 and b-cell responsiveness to lipopolysaccharide. *Blood, The Journal of the American Society of Hematology*, 99(5):1699–1705, 2002.

[420] David M Dorfman and Aliakbar Shahsafaei. Cd200 (ox-2 membrane glycoprotein) expression in b cell–derived neoplasms. *American journal of clinical pathology*, 134(5):726–733, 2010.

[421] WILLIAM E Serafin, ELAHE T Dayton, PETER M Gravallese, K FRANK Austen, and RICHARD L Stevens. Carboxypeptidase a in mouse mast cells. identification, characterization, and use as a differentiation marker. *The Journal of Immunology*, 139(11):3771–3776, 1987.

[422] Satish K Nandakumar, Jacob C Ulirsch, and Vijay G Sankaran. Advances in understanding erythropoiesis: evolving perspectives. *British journal of haematology*, 173(2):206–218, 2016.

[423] Fabien Loison, Haiyan Zhu, Kutay Karatepe, Anongnard Kasorn, Peng Liu, Keqiang Ye, Jiaxi Zhou, Shannan Cao, Haiyan Gong, Dieter E Jenne, et al. Proteinase 3–dependent caspase-3 cleavage modulates neutrophil death and inflammation. *The Journal of clinical investigation*, 124(10):4445–4458, 2014.

[424] Hitoshi TOYODA, Toshi KOMURASAKI, Daisuke UCHIDA, and Sigeo MORIMOTO. Distribution of mrna for human epiregulin, a differentially expressed member of the epidermal growth factor family. *Biochemical Journal*, 326(1):69–75, 1997.

[425] Jong Gwang Kim, Sang Kyun Sohn, Dong Hwan Kim, JIn Ho Baek, Nan Young Lee, Jang Soo Suh, Shung-Chull Chae, Kun Soo Lee, and Kyu Bo Lee. Clinical implications of angiogenic factors in patients with acute or chronic leukemia: hepatocyte growth factor levels have prognostic impact, especially in patients with acute myeloid leukemia. *Leukemia & lymphoma*, 46(6):885–891, 2005.

[426] Sheng-Yan Lin, Fei-Fei Hu, Ya-Ru Miao, Hui Hu, Qian Lei, Qiong Zhang, Qiubai Li, Hongxiang Wang, Zhichao Chen, and An-Yuan Guo. Identification of stab1 in multiple datasets as a prognostic factor for cytogenetically normal aml: Mechanism and drug indications. *Molecular Therapy-Nucleic Acids*, 18:476–484, 2019.

[427] Hai-Yan Gao, Xin-Guo Luo, Xi Chen, and Jing-Hua Wang. Identification of key genes affecting disease free survival time of pediatric acute lymphoblastic leukemia based on bioinformatic analysis. *Blood Cells, Molecules, and Diseases*, 54(1):38–43, 2015.

[428] Zhiheng Cheng, Yifeng Dai, Yifan Pang, Yang Jiao, Yan Liu, Longzhen Cui, Liang Quan, Tingting Qian, Tiansheng Zeng, Chaozeng Si, et al. Up-regulation of ddit4 predicts poor prognosis in acute myeloid leukaemia. *Journal of Cellular and Molecular Medicine*, 24(1):1067–1075, 2020.

[429] GJ Wright, M Jones, MJ Puklavec, MH Brown, and AN Barclay. The unusual distribution of the neuronal/lymphoid cell surface cd200 (ox2) glycoprotein is conserved in humans. *Immunology*, 102(2):173–179, 2001.

[430] Qing-Li Wu, Claudia Zierold, and Erik A Ranheim. Dysregulation of frizzled 6 is a critical component of b-cell leukemogenesis in a mouse model of chronic lymphocytic leukemia. *Blood*, 113(13):3031–3039, 2009.

[431] César Cobaleda, Alexandra Schebesta, Alessio Delogu, and Meinrad Busslinger. Pax5: the guardian of b cell identity and function. *Nature immunology*, 8(5):463–470, 2007.

[432] MS Khodadoust, B Luo, BC Medeiros, RC Johnson, MD Ewalt, AS Schalkwyk, CD Bangs, AM Cherry, S Arai, DA Arber, et al. Clinical activity of ponatinib in a patient with fgfr1-rearranged mixed-phenotype acute leukemia. *Leukemia*, 30(4):947–950, 2016.

[433] M Anna Kowalska, Lubica Rauova, and Mortimer Poncz. Role of the platelet chemokine platelet factor 4 (pf4) in hemostasis and thrombosis. *Thrombosis research*, 125(4):292–296, 2010.

[434] Daniela Damiani, Mario Tiribelli, Alessandra Franzoni, Angela Michelutti, Dora Fabbro, Margherita Cavallin, Eleonora Toffoletti, Erica Simeone, Renato Fanin, and Giuseppe Damante. Baalc overexpression retains its negative prognostic role across all cytogenetic risk groups in acute myeloid leukemia patients. *American Journal of Hematology*, 88(10):848–852, 2013.

[435] Nashwa El-Khazragy, Marwa A Esmaiel, Magdy M Mohamed, and Nahla S Hassan. Upregulation of long noncoding rna lnc-irf2-3 and lnc-znf667-as1 is associated with poor survival in b-chronic lymphocytic leukemia. *International Journal of Laboratory Hematology*, 42(3):284–291, 2020.

[436] Sandra Heesch, Martin Neumann, Stefan Schwartz, Isabelle Bartram, Cornelia Schlee, Thomas Burmeister, Matthias Hänel, Arnold Ganser, Michael Heuser, Clemens-Martin Wendtner, et al. Acute leukemias of ambiguous lineage in adults: molecular and clinical characterization. *Annals of hematology*, 92(6):747–758, 2013.

[437] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[438] Inge Hoebeke, Magda De Smedt, Frank Stolz, K Pike-Overzet, FJT Staal, Jean Plum, and Georges Leclercq. T-, b-and nk-lymphoid, but not myeloid cells arise from human cd34+ cd38- cd7+ common lymphoid progenitors expressing lymphoid-specific genes. *Leukemia*, 21(2):311–319, 2007.

[439] Ofir Wolach and Richard M Stone. How i treat mixed-phenotype acute leukemia. *Blood, The Journal of the American Society of Hematology*, 125(16):2477–2485, 2015.

[440] OK Weinberg and DA Arber. Mixed-phenotype acute leukemia: historical overview and a new definition. *Leukemia*, 24(11):1844–1851, 2010.

[441] Fatih M Uckun, Kazimiera Gajl-Peczalska, DE Myers, W Jaszcz, S Haissig, and JA Ledbetter. Temporal association of cd40 antigen expression with discrete stages of human b-cell ontogeny and the efficacy of anti-cd40 immunotoxins against clonogenic b-lineage acute lymphoblastic leukemia as well as b-lineage non-hodgkin's lymphoma cells. 1990.

[442] Pierre Garrone, Eve-Marie Neidhardt, Eric Garcia, Laurent Galibert, C Van Kooten, and Jacques Banchereau. Fas ligation induces apoptosis of cd40-activated human b lymphocytes. *The Journal of experimental medicine*, 182(5):1265–1273, 1995.

[443] Marcela Vlková, Eva Froňková, Veronika Kanderová, Aleš Janda, Šárka Ržičková, Jiří Litzman, Anna Šedivá, and Tomáš Kalina. Characterization of lymphocyte subsets in patients with common variable immunodeficiency reveals subsets of naive human b cells marked by cd24 expression. *The Journal of immunology*, 185(11):6431–6438, 2010.

[444] Uma Malhotra and Patrick Concannon. Human t-cell receptor cd3-delta (cd3d)/mspi dna polymorphism. *Nucleic acids research*, 17(6):2373, 1989.

[445] Alejandro Aruffo and Brian Seed. Molecular cloning of two cd7 (t-cell leukemia antigen) cdnas by a cos cell expression system. *The EMBO journal*, 6(11):3313–3316, 1987.

[446] Noémi Nagy, Cristina Cerboni, Karin Mattsson, Akihiko Maeda, Péter GOGOLak, János Sümegi, Arpád Lanyi, László SZekely, Ennio Carbone, George Klein, et al. Sh2d1a and slam protein expression in human lymphocytes and derived cell lines. *International journal of cancer*, 88(3):439–447, 2000.

[447] Paul M Maciocia, Patrycja A Wawrzyniecka, Brian Philip, Ida Ricciardelli, Ayse U Akarca, Shimobi C Onuoha, Mateusz Legut, David K Cole, Andrew K Sewell, Giuseppe Gritti, et al. Targeting the t cell receptor β-chain constant region for immunotherapy of t cell malignancies. *Nature medicine*, 23(12):1416, 2017.

[448] Hugo J Snippert, Johan H van Es, Maaike van den Born, Harry Begthel, Daniel E Stange, Nick Barker, and Hans Clevers. Prominin-1/cd133 marks stem cells and early progenitors in mouse small intestine. *Gastroenterology*, 136(7):2187–2194, 2009.

[449] Francisco Borrego. The cd300 molecules: an emerging family of regulators of the immune system. *Blood*, 121(11):1951–1960, 2013.

[450] S Kotake, M Higaki, K Sato, S Himeno, H Morita, KANG JUNG Kim, N Nara, N Miyasaka, K Nishioka, and S Kashiwazaki. Detection of myeloid precursors (granulocyte/macrophage colony forming units) in the bone marrow adjacent to rheumatoid arthritis joints. *The Journal of Rheumatology*, 19(10):1511–1516, 1992.

[451] Felix Ellett, Luke Pase, John W Hayman, Alex Andrianopoulos, and Graham J Lieschke. mpeg1 promoter transgenes direct macrophage-lineage expression in zebrafish. *Blood, The Journal of the American Society of Hematology*, 117(4):e49–e56, 2011.

[452] Louise M Kelly, Ursula Englmeier, Isabelle Lafon, Michael H Sieweke, and Thomas Graf. Mafb is an inducer of monocytic differentiation. *The EMBO journal*, 19(9):1987–1997, 2000.

[453] K Morita, Y Masamoto, K Kataoka, J Koya, Y Kagoya, H Yashiroda, T Sato, S Murata, and M Kurokawa. Baalc potentiates oncogenic erk pathway through interactions with mekk1 and klf4. *Leukemia*, 29(11):2248–2256, 2015.

[454] Caroline S Hughes, Liza M Colhoun, Baljinder K Bains, Joanne D Kilgour, Roberta E Burden, James F Burrows, Ed C Lavelle, Brendan F Gilmore, and Christopher J Scott. Extracellular cathepsin s and intracellular caspase 1 activation are surrogate biomarkers of particulate-induced lysosomal disruption in macrophages. *Particle and fibre toxicology*, 13(1):1–13, 2015.

[455] Alphonse Krystosek and Leo Sachs. Control of lysozyme induction in the differentiation of myeloid leukemic cells. *Cell*, 9(4):675–684, 1976.

[456] Jacinta Bustamante, Andres A Arias, Guillaume Vogt, Capucine Picard, Lizbeth Blancas Galicia, Carolina Prando, Audrey V Grant, Christophe C Marchal, Marjorie Hubeau, Ariane Chapgier, et al. Germline cybb mutations that selectively affect macrophages in kindreds with x-linked predisposition to tuberculous mycobacterial disease. *Nature immunology*, 12(3):213–221, 2011.

[457] Didi Matza, Abdallah Badou, Mithilesh K Jha, Tim Willinger, Andrey Antov, Shomyseh Sanjabi, Koichi S Kobayashi, Vincent T Marchesi, and Richard A Flavell. Requirement for ahnak1-mediated calcium signaling during t lymphocyte cytolysis. *Proceedings of the National Academy of Sciences*, 106(24):9785–9790, 2009.

[458] Chen Wang, Sharon Celeste Morley, David Donermeyer, Ivan Peng, Wyne P Lee, Jason Devoss, Dimitry M Danilenko, Zhonghua Lin, Juan Zhang, Jie Zhou, et al. Actin-bundling protein l-plastin regulates t cell activation. *The Journal of immunology*, 185(12):7487–7497, 2010.

[459] Charles A Janeway Jr. The t cell receptor as a multicomponent signalling machine: Cd4/cd8 coreceptors and cd45 in t cell activation. *Annual review of immunology*, 10(1):645–674, 1992.

[460] Kee Nyung Lee, Hyung-Sik Kang, Jun-Ho Jeon, Eun-Mi Kim, Suk-Ran Yoon, Hyunkeun Song, Chil-Youl Lyu, Zheng-Hao Piao, Sun-Uk Kim, Ying-Hao Han, et al. Vdup1 is required for the development of natural killer cells. *Immunity*, 22(2):195–208, 2005.

[461] Thomas N Wight. Versican: a versatile extracellular matrix proteoglycan in cell biology. *Current opinion in cell biology*, 14(5):617–623, 2002.

[462] JMH Kijas, TR Bauer Jr, S Gäfvert, S Marklund, G Trowald-Wigh, A Johannisson, Å Hedhammar, M Binns, RK Juneja, DD Hickstein, et al. A missense mutation in the β-2 integrin gene (itgb2) causes canine leukocyte adhesion deficiency. *Genomics*, 61(1):101–107, 1999.

[463] Ann P Wheeler and Anne J Ridley. Why three rho proteins? rhoa, rhob, rhoc, and cell motility. *Experimental cell research*, 301(1):43–49, 2004.

[464] Jennifer M Mataraza, Michael W Briggs, Zhigang Li, Alan Entwistle, Anne J Ridley, and David B Sacks. Iqgap1 promotes cell motility and invasion. *Journal of Biological Chemistry*, 278(42):41237–41245, 2003.

[465] Jin-Long Huang, Wei Liu, Li-Hong Tian, Ting-Ting Chai, Yang Liu, Feng Zhang, Hai-Ying Fu, Hua-Rong Zhou, and Jian-Zhen Shen. Upregulation of long non-coding rna malat-1 confers poor prognosis and influences cell proliferation and apoptosis in acute monocytic leukemia. *Oncology reports*, 38(3):1353–1362, 2017.

[466] Kang-xiao Ma, Hong-jie Wang, Xiao-rong Li, Tao Li, Gang Su, Pan Yang, and Jian-wen Wu. Long noncoding rna malat1 associates with the malignant status and poor prognosis in glioma. *Tumor Biology*, 36(5):3355–3359, 2015.

[467] Hong-Tu Zheng, De-Bing Shi, Yu-Wei Wang, Xin-Xiang Li, Ye Xu, Pratik Tripathi, Wei-Lie Gu, Guo-Xiang Cai, and San-Jun Cai. High expression of lncrna malat1 suggests a biomarker of poor prognosis in colorectal cancer. *International journal of clinical and experimental pathology*, 7(6):3174, 2014.

[468] Lars Henning Schmidt, Tilmann Spieker, Steffen Koschmieder, Julia Humberg, Dominik Jungen, Etmar Bulk, Antje Hascher, Danielle Wittmer, Alessandro Marra, Ludger Hillejan, et al. The long noncoding malat-1 rna indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth. *Journal of thoracic oncology*, 6(12):1984–1992, 2011.

[469] T Fujimoto, Kristina Anderson, Sten Eirik W Jacobsen, S-i Nishikawa, and C Nerlov. Cdk6 blocks myeloid differentiation by interfering with runx1 dna binding and runx1-c/ebpα interaction. *The EMBO journal*, 26(9):2361–2370, 2007.

[470] EM Heath, SM Chan, MD Minden, T Murphy, LI Shlush, and AD Schimmer. Biological and clinical consequences of npm1 mutations in aml. *Leukemia*, 31(4):798–807, 2017.

[471] Sebastian Schwind, Guido Marcucci, Kati Maharry, Michael D Radmacher, Krzysztof Mrózek, Kelsi B Holland, Dean Margeson, Heiko Becker, Susan P Whitman, Yue-Zhong Wu, et al. Baalc and erg expression levels are associated with outcome and distinct gene and microrna expression profiles in older patients with de novo cytogenetically normal acute myeloid leukemia: a cancer and leukemia group b study. *Blood, The Journal of the American Society of Hematology*, 116(25):5660–5669, 2010.

[472] Christian Langer, Michael D Radmacher, Amy S Ruppert, Susan P Whitman, Peter Paschka, Krzysztof Mrózek, Claudia D Baldus, Tamara Vukosavljevic, Chang-Gong Liu, Mary E Ross, et al. High baalc expression associates with other molecular prognostic markers, poor outcome, and a distinct gene-expression signature in cytogenetically normal patients younger than 60 years with acute myeloid leukemia: a cancer and leukemia group b (calgb) study. *Blood*, 111(11):5371–5379, 2008.

[473] Anqi Zhu, Joseph G Ibrahim, and Michael I Love. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, 35(12):2084–2092, 2019.

[474] Pantano L. Degreport: Report of deg analysis. r package version 1.13.8. http://lpantano.github.io/DEGreport/, 2020.

[475] L Kaufman and PJ Rousseeuw. Divisive analysis (program diana), volume 1 of 1, chapter 6. *New York: Wiley Inter-Science*, 9:14–22, 1990.

[476] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287, 2012.

[477] Ursula Mönning, Gerhard König, Reinhard Prior, Hans Mechler, Ursula Schreiter-Gasser, Colin L Masters, and Konrad Beyreuther. Synthesis and secretion of alzheimer amyloid βa4 precursor protein by stimulated human peripheral blood leucocytes. *FEBS letters*, 277(1-2):261–266, 1990.

[478] U Mönning, G König, RB Banati, H Mechler, Christian Czech, J Gehrmann, U Schreiter-Gasser, CL Masters, and K Beyreuther. Alzheimer beta a4-amyloid protein precursor in immunocompetent cells. *Journal of Biological Chemistry*, 267(33):23950–23956, 1992.

[479] Ryuichi Fukuyama, Yohko Murakawa, and Stanley I Rapoport. Induction of gene expression of amyloid precursor protein (app) in activated human lymphoblastoid cells and lymphocytes. *Molecular and chemical neuropathology*, 23(2-3):93–101, 1994.

[480] Maria J Bullido, Maria A Muñoz-Fernadez, Maria Recuero, Manuel Fresno, and Fernando Valdivieso. Alzheimer's amyloid precursor protein is expressed on the surface of hematopoietic cells upon activation. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1313(1):54–62, 1996.

[481] Antoinette R Bailey, Huayan Hou, Demian F Obregon, Jun Tian, Yuyan Zhu, Qiang Zou, William V Nikolic, Michael Bengtson, Takashi Mori, Tanya Murphy, et al. Aberrant t-lymphocyte development and function in mice overexpressing human soluble amyloid precursor protein-α: implications for autism. *The FASEB Journal*, 26(3):1040–1051, 2012.

[482] Richard J O'Brien and Philip C Wong. Amyloid precursor protein processing and alzheimer's disease. *Annual review of neuroscience*, 34:185–204, 2011.

[483] Andreas Kern, Birgit Roempp, Kai Prager, Jochen Walter, and Christian Behl. Down-regulation of endogenous amyloid precursor protein processing due to cellular aging. *Journal of Biological Chemistry*, 281(5):2405–2413, 2006.

[484] Sebastian Jimenez, Manuel Torres, Marisa Vizuete, Raquel Sanchez-Varo, Elisabeth Sanchez-Mejias, Laura Trujillo-Estrada, Irene Carmona-Cuenca, Cristina Caballero, Diego Ruano, Antonia Gutierrez, et al. Age-dependent accumulation of soluble amyloid β (aβ) oligomers reverses the neuroprotective effect of soluble amyloid precursor protein-α (sappα) by modulating phosphatidylinositol 3-kinase (pi3k)/akt-gsk-3β pathway in alzheimer mouse model. *Journal of Biological Chemistry*, 286(21):18414–18425, 2011.

[485] Tatiana Burrinha, Ricardo Gomes, Ana Paula Terrasso, and Cláudia Guimas Almeida. Neuronal aging potentiates beta-amyloid generation via amyloid precursor protein endocytosis. *bioRxiv*, page 616540, 2019.

[486] Kathryn L McCance and Sue E Huether. *Pathophysiology: The biologic basis for disease in adults and children*. Elsevier Health Sciences, 2014.

[487] Desa Lilic, Andrew J Cant, Mario Abinun, Jane E Calvert, and Gavin P Spickett. Cytokine production differs in children and adults. *Pediatric research*, 42(2):237–240, 1997.

[488] R Valiathan, M Ashman, and D Asthana. Effects of ageing on the immune system: infants to elderly. *Scandinavian journal of immunology*, 83(4):255–266, 2016.

[489] H Morbach, EM Eichhorn, JG Liese, and HJ Girschick. Reference values for b cell subpopulations from infancy to adulthood. *Clinical & Experimental Immunology*, 162(2):271–279, 2010.

[490] Christina Siemes, Thomas Quast, Christiane Kummer, Sven Wehner, Gregor Kirfel, Ulrike Müller, and Volker Herzog. Keratinocytes from app/aplp2-deficient mice are impaired in proliferation, adhesion and migration in vitro. *Experimental cell research*, 312(11):1939–1949, 2006.

[491] Francisco Sanchez-Vega, Marco Mina, Joshua Armenia, Walid K Chatila, Augustin Luna, Konnor C La, Sofia Dimitriadoy, David L Liu, Havish S Kantheti, Sadegh Saghafinia, et al. Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173(2):321–337, 2018.

[492] Evren U Azeloglu and Ravi Iyengar. Signaling networks: information flow, computation, and decision making. *Cold Spring Harbor perspectives in biology*, 7(4):a005934, 2015.

[493] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[494] Adam Friedman and Norbert Perrimon. Genetic screening for signal transduction in the era of network biology. *Cell*, 128(2):225–231, 2007.

[495] Norbert Perrimon, Chrysoula Pitsouli, and Ben-Zion Shilo. Signaling mechanisms controlling cell fate and embryonic patterning. *Cold Spring Harbor perspectives in biology*, 4(8):a005975, 2012.

[496] Spyros Artavanis-Tsakonas, Matthew D Rand, and Robert J Lake. Notch signaling: cell fate control and signal integration in development. *Science*, 284(5415):770–776, 1999.

[497] Matthew Freeman. Feedback control of intercellular signalling in development. *Nature*, 408(6810):313–319, 2000.

[498] Hidde De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103, 2002.

[499] John J Tyson, Teeraphan Laomettachit, and Pavel Kraikivski. Modeling the dynamic behavior of biochemical regulatory networks. *Journal of theoretical biology*, 462:514–527, 2019.

[500] Lian En Chai, Swee Kuan Loh, Swee Thing Low, Mohd Saberi Mohamad, Safaai Deris, and Zalmiyah Zakaria. A review on the computational approaches for gene regulatory network construction. *Computers in biology and medicine*, 48:55–65, 2014.

[501] Stuart A Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, 22(3):437–467, 1969.

[502] Herman F Fumia and Marcelo L Martins. Boolean network model for cancer pathways: predicting carcinogenesis and targeted therapy outcomes. *PloS one*, 8(7):e69008, 2013.

[503] Sriganesh Srihari, Venkatesh Raman, Hon Wai Leong, and Mark A Ragan. Evolution and controllability of cancer networks: a boolean perspective. *IEEE/ACM transactions on computational biology and bioinformatics*, 11(1):83–94, 2013.

[504] Desheng Zheng, Guowu Yang, Xiaoyu Li, Zhicai Wang, Feng Liu, and Lei He. An efficient algorithm for computing attractors of synchronous and asynchronous boolean networks. *PloS one*, 8(4):e60593, 2013.

[505] Eric Goles and Lilian Salinas. Comparison between parallel and serial dynamics of boolean networks. *Theoretical Computer Science*, 396(1-3):247–253, 2008.

[506] Adrien Fauré, Aurélien Naldi, Claudine Chaouiya, and Denis Thieffry. Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 22(14):e124–e131, 2006.

[507] Aurélien Naldi, Elisabeth Remy, Denis Thieffry, and Claudine Chaouiya. Dynamically consistent reduction of logical regulatory graphs. *Theoretical Computer Science*, 412(21):2207–2218, 2011.

[508] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29–34, 1999.

[509] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl_1):D428–D432, 2005.

[510] Alexander R Pico, Thomas Kelder, Martijn P Van Iersel, Kristina Hanspers, Bruce R Conklin, and Chris Evelo. Wikipathways: pathway editing for the people. *PLoS Biol*, 6(7):e184, 2008.

[511] Gary D Bader, Michael P Cary, and Chris Sander. Pathguide: a pathway resource list. *Nucleic acids research*, 34(suppl_1):D504–D506, 2006.

[512] Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, 2011.

[513] Francesco Ceccarelli, Denes Turei, Attila Gabor, and Julio Saez-Rodriguez. Bringing data from curated pathway resources to cytoscape with omnipath. *Bioinformatics*, 36(8):2632–2633, 2020.

[514] István Albert, Juilee Thakar, Song Li, Ranran Zhang, and Reka Albert. Boolean network simulations for life scientists. *Source code for biology and medicine*, 3(1):1–8, 2008.

[515] Nobue Itasaki, C Michael Jones, Sara Mercurio, Alison Rowe, Pedro M Domingos, James C Smith, and Robb Krumlauf. Wise, a context-dependent activator and inhibitor of wnt signalling. *Development*, 130(18):4295–4305, 2003.

[516] Soorin Yim, Hasun Yu, Dongjin Jang, and Doheon Lee. Annotating activation/inhibition relationships to protein-protein interactions using gene ontology relations. *BMC systems biology*, 12(1):9, 2018.

[517] Qiao-Xin Li, Stephanie J Fuller, Konrad Beyreuther, and Colin L Masters. The amyloid precursor protein of alzheimer disease in human brain and blood. *Journal of leukocyte biology*, 66(4):567–574, 1999.

[518] Mary Y Chang, Christina K Chan, Kathleen R Braun, Pattie S Green, Kevin D O'Brien, Alan Chait, Anthony J Day, and Thomas N Wight. Monocyte-to-macrophage differentiation synthesis and secretion of a complex extracellular matrix. *Journal of Biological Chemistry*, 287(17):14122–14135, 2012.

[519] Stefan Bodmer, Marcia Berman Podlisny, Dennis J Selkoe, Irma Heid, and Adriano Fontana. Transforming growth factor-beta bound to soluble derivatives of the beta amyloid precursor protein of alzheimer's disease. *Biochemical and biophysical research communications*, 171(2):890–897, 1990.

[520] Darrell D Mousseau, Sarah Chapelsky, Gregory De Crescenzo, Marina D Kirkitadze, Joanne Magoon, Sadayuki Inoue, David B Teplow, and Maureen D O'Connor-McCourt. A direct interaction between transforming growth factor (tgf)-βs and amyloid-β protein affects fibrillogenesis in a tgf-βreceptor-independent manner. *Journal of Biological Chemistry*, 278(40):38715–38722, 2003.

[521] Sylvain Lesné, Fabian Docagne, Cecılia Gabriel, Géraldine Liot, Debomoy K Lahiri, Luc Buée, Laurent Plawinski, André Delacourte, Eric T MacKenzie, Alain Buisson, et al. Transforming growth factor-β1 potentiates amyloid-β generation in astrocytes and in transgenic mice. *Journal of Biological Chemistry*, 278(20):18408–18418, 2003.

[522] Thomas M Williams, Melissa E Williams, Joanne H Heaton, Thomas D Gelehrter, and Jeffrey W Innis. Group 13 hox proteins interact with the mh2 domain of r-smads and modulate smad transcriptional activation functions independent of hox dna-binding capability. *Nucleic acids research*, 33(14):4475–4484, 2005.

[523] Shuting Bai, Xingming Shi, Xiangli Yang, and Xu Cao. Smad6 as a transcriptional corepressor. *Journal of Biological Chemistry*, 275(12):8267–8270, 2000.

[524] Akiko Hata, Giorgio Lagna, Joan Massagué, and Ali Hemmati-Brivanlou. Smad6 inhibits bmp/smad1 signaling by specifically competing with the smad4 tumor suppressor. *Genes & development*, 12(2):186–197, 1998.

[525] Ronan Quéré, Göran Karlsson, Falk Hertwig, Marianne Rissler, Beata Lindqvist, Thoas Fioretos, Peter Vandenberghe, Marilyn L Slovak, Jörg Cammenga, and Stefan Karlsson. Smad4 binds hoxa9 in the cytoplasm and protects primitive hematopoietic cells against nuclear activation by hoxa9 and leukemia transformation. *Blood, The Journal of the American Society of Hematology*, 117(22):5918–5930, 2011.

[526] Leon Arriola and James M Hyman. Sensitivity analysis for uncertainty quantification in mathematical models. In *Mathematical and statistical estimation approaches in epidemiology*, pages 195–247. Springer, 2009.

[527] Yanika Borg, Ekkehard Ullner, Afnan Alagha, Ahmed Alsaedi, Darren Nesbeth, and Alexey Zaikin. Complex and unexpected dynamics in simple genetic regulatory networks. *International Journal of Modern Physics B*, 28(14):1430006, 2014.

[528] Timothy S Gardner, Charles R Cantor, and James J Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, 2000.

[529] Michael B Elowitz and Stanislas Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

[530] Felipe Barraza, Marcelo Arancibia, Eva Madrid, and Cristian Papuzinski. General concepts in biostatistics and clinical epidemiology: Random error and systematic error. *Medwave*, 19(07), 2019.

[531] Elena Kuzmin, Benjamin VanderSluis, Wen Wang, Guihong Tan, Raamesh Deshpande, Yiqun Chen, Matej Usaj, Attila Balint, Mojca Mattiazzi Usaj, Jolanda Van Leeuwen, et al. Systematic analysis of complex genetic interactions. *Science*, 360(6386), 2018.

[532] Apichat Suratanee, Martin H Schaefer, Matthew J Betts, Zita Soons, Heiko Mannsperger, Nathalie Harder, Marcus Oswald, Markus Gipp, Ellen Ramminger, Guillermo Marcus, et al. Characterizing protein interactions employing a genome-wide sirna cellular phenotyping screen. *PLoS Comput Biol*, 10(9):e1003814, 2014.

[533] Tengjiao Wang, Yanghe Feng, and Qi Wang. Pairs: Prediction of activation/inhibition regulation signaling pathway. *Computational intelligence and neuroscience*, 2017, 2017.

[534] Sarah M Assmann and Réka Albert. Discrete dynamic modeling with asynchronous update, or how to model complex systems in the absence of quantitative information. In *Plant Systems Biology*, pages 207–225. Springer, 2009.

[535] Olaf Wolkenhauer, Darryl Shibata, and Mihajlo D Mesarović. The role of theorem proving in systems biology. *Journal of Theoretical Biology*, 300:57–61, 2012.

[536] Pavel Rusnok and K Adlassnig. Detection of inaccuracy in a medical knowledge base using a classical theorem prover. *Health Informatics Meet Ehealth, Vienna, Austria*, 2010.

[537] Elisabetta De Maria, Joëlle Despeyroux, and Amy P Felty. A logical framework for systems biology. In *International Conference on Formal Methods in Macro-Biology*, pages 136–155. Springer, 2014.

[538] Adnan Rashid, Osman Hasan, Umair Siddique, and Sofiene Tahar. Formal reasoning about systems biology using theorem proving. *PloS one*, 12(7):e0180179, 2017.

[539] A Chizuka, M Suda, T Shibata, E Kusumi, A Hori, T Hamaki, Y Kodama, K Horigome, Y Kishi, K Kobayashi, et al. Difference between hematological malignancy and solid tumor research articles published in four major medical journals. *Leukemia*, 20(10):1655–1657, 2006.

[540] Edward A Copelan. Hematopoietic stem-cell transplantation. *New England Journal of Medicine*, 354(17):1813–1826, 2006.

[541] Taner Demirer, Lisbeth Barkholt, Didier Blaise, Paolo Pedrazzoli, Massimo Aglietta, Angelo Michele Carella, Jacques-Olivier Bay, Fikret Arpaci, Giovanni Rosti, Gunhan Gurman, et al. Transplantation of allogeneic hematopoietic stem cells: an emerging treatment modality for solid tumors. *Nature clinical practice Oncology*, 5(5):256–267, 2008.

[542] Peter Hohenberger and Dieter Buchheidt. Surgical interventions in patients with hematologic malignancies. *Critical reviews in oncology/hematology*, 55(2):83–91, 2005.

[543] Kristen B Long, Regina M Young, Alina C Boesteanu, Megan M Davis, J Joseph Melenhorst, Simon F Lacey, David A DeGaramo, Bruce L Levine, and Joseph A Fraietta. Car t cell therapy of non-hematopoietic malignancies: detours on the road to clinical success. *Frontiers in immunology*, 9:2740, 2018.

[544] Felix Mitelman. Recurrent chromosome aberrations in cancer. *Mutation Research/Reviews in mutation research*, 462(2-3):247–253, 2000.

[545] D Ribatti. Is angiogenesis essential for the progression of hematological malignancies or is it an epiphenomenon? *Leukemia*, 23(3):433–434, 2009.

[546] Matthew Trendowski. The inherent metastasis of leukaemia and its exploitation by sonodynamic therapy. *Critical reviews in oncology/hematology*, 94(2):149–163, 2015.

[547] Ryan Cordner, Keith L Black, and Christopher J Wheeler. Exploitation of adaptive evolution in glioma treatment. *CNS oncology*, 2(2):171–179, 2013.

[548] Jingsong Zhang, Jessica J Cunningham, Joel S Brown, and Robert A Gatenby. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. *Nature communications*, 8(1):1–9, 2017.

[549] Maximilian Andreas Roland Strobl, Jeffrey West, Yannick Viossat, Mehdi Damaghi, Mark Robertson-Tessi, Joel Brown, Robert Gatenby, Philip Maini, and Alexander Anderson. Turnover modulates the need for a cost of resistance in adaptive therapy. *BioRxiv*, 2020.

[550] William C Ratcliff, R Ford Denison, Mark Borrello, and Michael Travisano. Experimental evolution of multicellularity. *Proceedings of the National Academy of Sciences*, 109(5):1595–1600, 2012.

[551] Berenika Plusa and Anna-Katerina Hadjantonakis. Current topics in developmental biology cell fate in mammalian development introduction, 2018.

[552] Kyoko Ito and Keisuke Ito. Metabolism and the control of cell fate decisions and stem cell renewal. *Annual review of cell and developmental biology*, 32:399–409, 2016.

[553] Makiko Iwafuchi-Doi and Kenneth S Zaret. Cell fate control by pioneer transcription factors. *Development*, 143(11):1833–1837, 2016.

[554] Daniel Jun-Kit Hu and Heinrich Jasper. Control of intestinal cell fate by dynamic mitotic spindle repositioning influences epithelial homeostasis and longevity. *Cell reports*, 28(11):2807–2823, 2019.

[555] Jos Domen and Irving L Weissman. Self-renewal, differentiation or death: regulation and manipulation of hematopoietic stem cell fate. *Molecular medicine today*, 5(5):201–208, 1999.

[556] Ulrika Blank, Göran Karlsson, and Stefan Karlsson. Signaling pathways governing stem-cell fate. *Blood*, 111(2):492–503, 2008.

[557] Michael A Rieger, Philipp S Hoppe, Benjamin M Smejkal, Andrea C Eitelhuber, and Timm Schroeder. Hematopoietic cytokines can instruct lineage choice. *Science*, 325(5937):217–218, 2009.

[558] Noushine Mossadegh-Keller, Sandrine Sarrazin, Prashanth K Kandalla, Leon Espinosa, E Richard Stanley, Stephen L Nutt, Jordan Moore, and Michael H Sieweke. M-csf instructs myeloid lineage fate in single haematopoietic stem cells. *Nature*, 497(7448):239–243, 2013.

[559] Ciaran James Mooney, Alan Cunningham, Panagiotis Tsapogas, Kai-Michael Toellner, and Geoffrey Brown. Selective expression of flt3 within the mouse hematopoietic stem cell compartment. *International journal of molecular sciences*, 18(5):1037, 2017.

[560] Huafeng Xie, Min Ye, Ru Feng, and Thomas Graf. Stepwise reprogramming of b cells into macrophages. *Cell*, 117(5):663–676, 2004.

[561] Hiromi Iwasaki, Shin-ichi Mizuno, Yojiro Arinobu, Hidetoshi Ozawa, Yasuo Mori, Hirokazu Shigematsu, Kiyoshi Takatsu, Daniel G Tenen, and Koichi Akashi. The order of expression of transcription factors directs hierarchical specification of hematopoietic lineages. *Genes & development*, 20(21):3010–3021, 2006.

[562] Aline Mamo, Jana Krosl, Evert Kroon, Janet Bijl, Alexander Thompson, Nadine Mayotte, Simon Girard, Richard Bisaillon, Nathalie Beslu, Mark Featherstone, et al. Molecular dissection of meis1 reveals 2 domains required for leukemia induction and a key role for hoxa gene activation. *Blood*, 108(2):622–629, 2006.

[563] Kira Orlovsky, Alexander Kalinkovich, Tanya Rozovskaia, Elias Shezen, Tomer Itkin, Hansjuerg Alder, Hatice Gulcin Ozer, Letizia Carramusa, Abraham Avigdor, Stefano Volinia, et al. Down-regulation of homeobox genes meis1 and hoxa in mll-rearranged acute leukemia impairs engraftment and reduces proliferation. *Proceedings of the National Academy of Sciences*, 108(19):7956–7961, 2011.

[564] Mattias Magnusson, Ann CM Brun, H Jeffrey Lawrence, and Stefan Karlsson. Hoxa9/hoxb3/hoxb4 compound null mice display severe hematopoietic defects. *Experimental hematology*, 35(9):1421–e1, 2007.

[565] Bingying Zhou, Channing J Der, and Adrienne D Cox. The role of wild type ras isoforms in cancer. In *Seminars in cell & developmental biology*, volume 58, pages 60–69. Elsevier, 2016.

[566] Miaolong Lu and Xianquan Zhan. The crucial role of multiomic approach in cancer research and clinically relevant outcomes. *EPMA Journal*, 9(1):77–102, 2018.

[567] Meromit Singer and Ana C Anderson. Revolutionizing cancer immunology: the power of next-generation sequencing technologies. *Cancer immunology research*, 7(2):168–173, 2019.

[568] Lei Zhang, Ziyi Li, Katarzyna M Skrzypczynska, Qiao Fang, Wei Zhang, Sarah A O'Brien, Yao He, Lynn Wang, Qiming Zhang, Aeryon Kim, et al. Single-cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer. *Cell*, 181(2):442–459, 2020.

[569] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018.

[570] Manu Setty, Vaidotas Kiseliovas, Jacob Levine, Adam Gayoso, Linas Mazutis, and Dana Pe'er. Characterization of cell fate probabilities in single-cell data with palantir. *Nature biotechnology*, 37(4):451–460, 2019.

[571] Michael Lässig, Ville Mustonen, and Aleksandra M Walczak. Predicting evolution. *Nature ecology & evolution*, 1(3):1–9, 2017.

[572] Sayed-Rzgar Hosseini, Ramon Diaz-Uriarte, Florian Markowetz, and Niko Beerenwinkel. Estimating the predictability of cancer evolution. *Bioinformatics*, 35(14):i389–i397, 2019.

[573] J Arjan Gm De Visser and Joachim Krug. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15(7):480–490, 2014.