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

<table>
<thead>
<tr>
<th>1</th>
<th>Introduction</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Cancer Evolution</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1</td>
<td>The major biological events in cancer progression</td>
<td>3</td>
</tr>
<tr>
<td>1.1.2</td>
<td>Heterogeneity in cancer</td>
<td>5</td>
</tr>
<tr>
<td>1.1.3</td>
<td>Driver event interactions</td>
<td>7</td>
</tr>
<tr>
<td>1.1.4</td>
<td>Modelling cancer growth and resistance</td>
<td>9</td>
</tr>
<tr>
<td>1.2</td>
<td>Healthy and abnormal blood development in adults</td>
<td>11</td>
</tr>
<tr>
<td>1.2.1</td>
<td>Hematopoiesis</td>
<td>11</td>
</tr>
<tr>
<td>1.2.2</td>
<td>Blood cancers</td>
<td>15</td>
</tr>
<tr>
<td>1.2.3</td>
<td>JAK2 and TET2 in Myeloproliferative Neoplasms</td>
<td>18</td>
</tr>
<tr>
<td>1.3</td>
<td>Computational modelling of cancer progression and evolution</td>
<td>23</td>
</tr>
<tr>
<td>1.3.1</td>
<td>Space scales in cancer modelling</td>
<td>24</td>
</tr>
<tr>
<td>1.3.2</td>
<td>Model abstraction: discrete or continuous variables? Stochastic or deterministic simulations?</td>
<td>26</td>
</tr>
<tr>
<td>1.3.3</td>
<td>Parameter estimation as a barrier to cancer modelling</td>
<td>28</td>
</tr>
<tr>
<td>1.4</td>
<td>Thesis goals and plan</td>
<td>30</td>
</tr>
</tbody>
</table>

## Methods

<table>
<thead>
<tr>
<th>2</th>
<th>Methods</th>
<th>33</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>33</td>
</tr>
<tr>
<td>2.2</td>
<td>Dataset description</td>
<td>34</td>
</tr>
<tr>
<td>2.2.1</td>
<td>In-vivo lymphoma tumour growth in mouse models</td>
<td>34</td>
</tr>
<tr>
<td>2.2.2</td>
<td>RNAseq of AML patients</td>
<td>35</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Microarrays of MPN mouse cohorts</td>
<td>36</td>
</tr>
<tr>
<td>2.2.4</td>
<td>RNAseq of pediatric ALAL patients</td>
<td>37</td>
</tr>
<tr>
<td>2.2.5</td>
<td>RNAseq of adult MPAL patients</td>
<td>38</td>
</tr>
<tr>
<td>2.3</td>
<td>Computational tools to simulate and evaluate lymphoma rule-based models</td>
<td>38</td>
</tr>
</tbody>
</table>
2.3.1 Rule-based models ........................................... 38  
2.3.2 BioPEPA for the time-series and distribution analyses .... 39  
2.3.3 ProPPA for the parameter inference .......................... 39  
2.3.4 Stochastic simulation algorithms ............................ 40  
2.3.5 Approximate Bayesian Computation (ABC) method and pseudo-code 41  
2.3.6 Fluid approximation algorithm ............................. 43  
2.3.7 Scoring system for ProPPA inference results ............... 44  
2.4 Construction and analyses of the MPN network ............... 45  
2.4.1 Qualitative Networks ...................................... 45  
2.4.2 BioModelAnalyzer ........................................ 47  
2.4.3 Linear Temporal Logic (LTL) in BMA ........................ 49  
2.4.4 Stabilisation Analysis in BMA ............................. 50  
2.5 XGBoost to rank and identify JAK2 correlated pathways in AML patients . 52  
2.5.1 Tree Ensembles ........................................... 52  
2.5.2 XGBoost .................................................. 54  
2.5.3 Ranking pathway association for gene expression level using XGBoost 55  
2.5.4 SHAP score .............................................. 56  
2.6 Model checking using satisfiability modulo theories (SMT) .... 57  

3 Modelling therapeutic resistance dynamics of lymphoma mice cohorts. 61  
3.1 Introduction .................................................. 62  
3.2 Results ....................................................... 62  
3.2.1 A simple rule-based model can describe lymphoma growth in mice. 62  
3.2.2 Quantitative analysis of in vitro experiments of [1] is used to evaluate some parameters. ........................................... 64  
3.2.3 Logistic LD model reproduces lymphoma growth under control and continuous treatment regimes. ........................ 66  
3.2.4 Competitive regrowth is selected in alternative treatment regimes with parameter values matching in vitro experiments .......... 70  
3.2.5 Continuous regime can be improved with adaptive treatment regimes. 74  
3.3 Discussion .................................................... 77  

4 HOXA9 acts as an epigenetic switch in blood malignancies. 81  
4.1 Introduction .................................................. 82  
4.2 Results ....................................................... 84  
4.2.1 HOXA9 expression separates distinct cohorts of AML patients with distinct clinical features. ........................................... 84
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.2</td>
<td>The JAK2/TET2/HOXA9 motif explains divergent clinical outcomes in MPN</td>
<td>88</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Analyses of public MPN datasets validate NOTCH role in MPN as well as HOXA9 bimodality and prognosis role in blood diseases</td>
<td>98</td>
</tr>
<tr>
<td>4.3</td>
<td>Discussion</td>
<td>101</td>
</tr>
<tr>
<td>5</td>
<td>APP determines cell fate in blood diseases</td>
<td>103</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>104</td>
</tr>
<tr>
<td>5.2</td>
<td>Results</td>
<td>105</td>
</tr>
<tr>
<td>5.2.1</td>
<td>HOXA9 and APP form patient clusters with distinct clinical features</td>
<td>105</td>
</tr>
<tr>
<td>5.2.2</td>
<td>APP stratifies AML patients and is an important prognosis marker.</td>
<td>108</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Differential gene expression analyses suggest a role for APP in the lymphoid/myeloid differentiation process</td>
<td>112</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Investigation of APP as a biomarker for MPAL</td>
<td>115</td>
</tr>
<tr>
<td>5.2.5</td>
<td>Exploring the degree of similarity between the low-low APP AML cohort and MPAL</td>
<td>125</td>
</tr>
<tr>
<td>5.3</td>
<td>Discussion</td>
<td>129</td>
</tr>
<tr>
<td>6</td>
<td>Boolean motif inference applied to HOXA9 and APP in AML</td>
<td>133</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>133</td>
</tr>
<tr>
<td>6.2</td>
<td>Results</td>
<td>137</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Implementation of a motif inference algorithm</td>
<td>138</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Identification of motifs for the HOXA9/APP AML cohorts</td>
<td>143</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Investigation of possible motif gene candidates</td>
<td>145</td>
</tr>
<tr>
<td>6.2.4</td>
<td>Candidate gene validation using gene expression and clinical data</td>
<td>146</td>
</tr>
<tr>
<td>6.3</td>
<td>Discussion</td>
<td>150</td>
</tr>
<tr>
<td>7</td>
<td>Discussion</td>
<td>153</td>
</tr>
<tr>
<td>7.1</td>
<td>Liquid versus solid tumours</td>
<td>155</td>
</tr>
<tr>
<td>7.2</td>
<td>Hematopoietic dysregulation: the importance of gene expression dynamics in cell fate</td>
<td>156</td>
</tr>
<tr>
<td>7.3</td>
<td>Novel experimental technologies for tumour progression analyses</td>
<td>158</td>
</tr>
<tr>
<td>7.4</td>
<td>Conclusion</td>
<td>159</td>
</tr>
</tbody>
</table>

References 161
# List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>Approximate Bayesian Computation</td>
</tr>
<tr>
<td>AD</td>
<td>Alzheimer Disease</td>
</tr>
<tr>
<td>ALL</td>
<td>Acute Lymphoblastic Leukemia</td>
</tr>
<tr>
<td>AML</td>
<td>Acute Myeloid Leukemia</td>
</tr>
<tr>
<td>APL</td>
<td>Acute Promyelocytic Leukemia</td>
</tr>
<tr>
<td>APP</td>
<td>Amyloid Precursor Protein</td>
</tr>
<tr>
<td>BMA</td>
<td>BioModelAnalyzer</td>
</tr>
<tr>
<td>BP-MPN</td>
<td>Blast-Phase Myeloproliferative Neoplasm</td>
</tr>
<tr>
<td>CD</td>
<td>Cluster Differentiation</td>
</tr>
<tr>
<td>CLP</td>
<td>Common Lymphoid Progenitor</td>
</tr>
<tr>
<td>CMP</td>
<td>Common Myeloid Progenitor</td>
</tr>
<tr>
<td>DC</td>
<td>Dendritic Cells</td>
</tr>
<tr>
<td>DEGs</td>
<td>Differentially Expressed Genes</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>ECM</td>
<td>Extracellular Matrix</td>
</tr>
<tr>
<td>ET</td>
<td>Essential Thrombocytemia</td>
</tr>
<tr>
<td>FAB</td>
<td>French-American-British</td>
</tr>
<tr>
<td>GMP</td>
<td>Granulocyte Monocyte Progenitor</td>
</tr>
</tbody>
</table>
List of abbreviations

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  Myelodysplastic 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)
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>Root Mean Squared Error</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>SAT</td>
<td>Boolean satisfiability problem</td>
</tr>
<tr>
<td>SHAP</td>
<td>SHapley Additive exPlanations</td>
</tr>
<tr>
<td>SMT</td>
<td>Satisfiability Modulo Theories</td>
</tr>
<tr>
<td>ST-HSC</td>
<td>Short Term Hematopoietic Stem Cell</td>
</tr>
<tr>
<td>TCGA</td>
<td>The Cancer Genome Atlas</td>
</tr>
<tr>
<td>TCR</td>
<td>T Cell Receptor</td>
</tr>
<tr>
<td>TET2</td>
<td>Tet methylcytosine dioxygenase 2</td>
</tr>
<tr>
<td>VAF</td>
<td>Variant Allele Frequency</td>
</tr>
<tr>
<td>WBC</td>
<td>White Blood cell Count</td>
</tr>
<tr>
<td>XGBoost</td>
<td>eXtreme Gradient Boosting</td>
</tr>
<tr>
<td>t-SNE</td>
<td>t-distributed stochastic neighbor embedding</td>
</tr>
</tbody>
</table>
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.
1.1 Cancer Evolution

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 tumorigenesis, for example in a mouse model of soft-tissue sarcoma, the timing at which TP53 gene is lost after \( K\text{-ras}^{G12D} \) mutation influences
1.1 Cancer Evolution

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 tumorigenesis: 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.

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

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, granulocyte-monocyte progenitors; MEP, megakaryocyte-erythroid progenitors. Cell images from smart.servier.com, licensed under CC BY 3.0, edited from original.
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 dendritic 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 nucleated 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 thrombocytemia (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.
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 (bilineal) 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 Lymphocytic 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 megakaryocyctic/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, granulocyte-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 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 \textit{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 \textit{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. \textit{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 \textit{TET2} loss. \textit{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 \textit{TET2} loss mutation, papers indicate that \textit{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 \textit{TET2} loss [166, 168].

To conclude on \textit{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 \textit{TET2} in the erythroid lineage remains unclear. One could however hypothesise that \textit{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 challenges 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 disadvantageous 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 \( \text{JAK2} \) mutants in HSCs. Despite their initial observation reported above, Mullally et al [175] in a more recent study notice a gradual clonal advantage for \( \text{JAK2} \) 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 \( \text{JAK2} \) 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 \( \text{JAK2} \) mutants with an increased number of LSK cells in mice.

Several studies however support a deleterious effect of \( \text{JAK2} \) mutation for HSC growth. A significant number of papers [177, 172, 148, 178] demonstrate \( \text{JAK2} \) 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 \textit{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 \( \text{JAK2} \) 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 \( \text{JAK2} \) 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 \( \text{JAK2} \) 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 erythroid lineage.
In conclusion, \textit{JAK2} is essential for MEP and GMP expansion as well as for skewing differentiation towards myeloid and erythroid lineages. \textit{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 \textit{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 \textit{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 lymphoma.

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?

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

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 \textit{HOXA9} and \textit{APP} expression. While \textit{HOXA9} is a well-studied gene in AML, the poor characterisation of \textit{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 \textit{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 \textit{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.
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 \textit{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 (periodic 3) 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.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Survival (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>22, 22, 26, 26</td>
</tr>
<tr>
<td>Continuous 14 days</td>
<td>32, 35, 35, alive at 35</td>
</tr>
<tr>
<td>Every 3 days for 13 days</td>
<td>32, 32, 32, 32</td>
</tr>
<tr>
<td>Every 5 days for 16 days</td>
<td>25, 26, 30, 31</td>
</tr>
</tbody>
</table>

Table 2.1 \textbf{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 patient. 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 ArrayExpress 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 ≤ 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 us-
ing 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 $\alpha$ and die at rate $\beta$ can be written as follows:

$$
C \xrightarrow{\alpha} C + C \\
C \xrightarrow{\beta} \emptyset
$$

(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
Methods

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 $t < t_{\text{max}}$ 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 $i = 0, 1, \ldots, N$, $x_i$ the $N$ variables of the model and $t_{\text{max}}$ 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 $\tau$, 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, \( \theta \) 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 $i_{\text{max}}$ 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 $d_{x^*} < \text{eps}$ 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
  
  - $x(i+1) = x(i)$;
  
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’. $\text{eps}$ is the accepted tolerance threshold and $i_{\text{max}}$ the maximum number of samples. $i_{\text{max}}, \text{eps}$ 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 $d_{x^*} < \text{eps}$, 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 $i_{\text{max}}$ 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: $r = \frac{p(x^*) L(x^*)}{p(x(i)) L(x(i))}$;

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, $i_{\text{max}}$, 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 \textit{JAK2} and \textit{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 \( f_i : \{0, 1\}^k \rightarrow \{0, 1\} \) 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 \( \alpha_{ij} \) 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:

\[
f_i(s) = \begin{cases} 
0 & \text{if } \sum_{c_j \in C} \alpha_{ji}c_j < 0 \\
1 & \text{if } \sum_{c_j \in C} \alpha_{ji}c_j > 0 \\
c_i & \text{if } \sum_{c_j \in C} \alpha_{ji}c_j = 0
\end{cases}
\] (2.2)

Here the term \( \sum_{c_j \in C} \alpha_{ji}c_j \) 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 \( \{0, 1, \ldots, N\} \). \( N \) 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:

\[
c_i(t+1) = \begin{cases} 
c_i(t) + 1 & \text{if } \text{target}_i(s(t)) > c_i(t) \\
c_i(t) - 1 & \text{if } \text{target}_i(s(t)) < c_i(t) \\
c_i(t) & \text{if } \text{target}_i(s(t)) = c_i(t)
\end{cases}
\] (2.3)

with \( s(t) \) the current state of the network and \( \text{target}_i \in T \) the target function of \( c_i \). \( \text{target}_i \) returns a value in \( \{0, 1, \ldots, N\} \). 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:

\[
act_i = \frac{\sum_{\alpha_{ji} > 0} \alpha_{ji} c_j}{\sum_{\alpha_{ji} > 0} \alpha_{ji}}
\]

\[
inh_i = \frac{\sum_{\alpha_{ji} < 0} \alpha_{ji} c_j}{\sum_{\alpha_{ji} < 0} \alpha_{ji}}
\]

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

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

(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.](image)

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:

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

where $p_i$s 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].
Methods

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 (eXtreme Gradient Boosting) 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
2.5 XGBoost to rank and identify JAK2 correlated pathways in AML patients

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 (SHapley Additive exPlanations) 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 \textit{JAK2}. Genes with a high SHAP scores are most likely correlated to \textit{JAK2}.

\section{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 \textit{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 \(\text{max}\) with \(\text{max}\) 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 \(\land\), \(\lor\) 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 \((x \lor y) \land \neg x \lor \neg y) \land x\), with \((\neg x \lor \neg y)\) 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–Loveland (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
\[(x - x^2 \leq 2) \land (\exp(x) \neq y) \land (\sin(x) \geq 0 \land x + y = 1)\]
can be written as \(r \land s \land (p \land q)\) 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
3.2 Results

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](image)

**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:

$$
\begin{align*}
A & \xrightarrow{A_A} A + A \\
A & \xrightarrow{A_M} A + M \\
M & \xrightarrow{M_M} M + M \\
A & \xrightarrow{d} \varnothing \quad \text{when treatment is applied}
\end{align*}
$$

(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](image)

*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](image)

**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 $A_0 e^{A t}$ and $M_0 e^{M t}$, a linear regression is performed on the logarithmic data using the "stats" R package [309] to find $A$ and $M$ 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 $A$ and $M$ 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.
Modelling therapeutic resistance dynamics of lymphoma mice cohorts.

<table>
<thead>
<tr>
<th>Tumour 6</th>
<th>Tumour 6 Tam</th>
<th>Tumour 29 Tam</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_A = 0.93 \pm 0.12 )</td>
<td>( M_M = 1.12 \pm 0.02 ) (ctrl)</td>
<td>( M_M = 1.14 \pm 0.02 ) (ctrl)</td>
</tr>
<tr>
<td>( M_M = 0.91 \pm 0.03 ) (Tam)</td>
<td>( M_M = 0.92 \pm 0.06 ) (Tam)</td>
<td></td>
</tr>
</tbody>
</table>

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 addition, 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
3.2 Results

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.
Modelling therapeutic resistance dynamics of lymphoma mice cohorts.

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.
### 3.2 Results

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.

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

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.

<table>
<thead>
<tr>
<th>( A(t = 0) )</th>
<th>Luria-Delbrück ( M_M )</th>
<th>( d )</th>
<th>( A_M )</th>
<th>score</th>
<th>Logistic LD ( M_M )</th>
<th>( d )</th>
<th>( A_M )</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^3</td>
<td>0.68 [0.6802, 0.6814]</td>
<td>1.3</td>
<td>5.3 ( \times 10^{-5} ) [5.21, 5.46]</td>
<td>5.1</td>
<td>1.1 [1.118, 1.123]</td>
<td>0.56</td>
<td>5.0 ( \times 10^{-5} ) [5.624, 5.661]</td>
<td>4.3</td>
</tr>
<tr>
<td>10^4</td>
<td>0.61 [0.6123, 0.6139]</td>
<td>1.2</td>
<td>3.5 ( \times 10^{-5} ) [3.12, 3.54]</td>
<td>5.0</td>
<td>0.98 [0.9835, 0.9861]</td>
<td>0.50</td>
<td>5.0 ( \times 10^{-5} ) [5.023, 5.042]</td>
<td>4.3</td>
</tr>
<tr>
<td>10^5</td>
<td>0.56 [0.5604, 0.5632]</td>
<td>0.90</td>
<td>8.2 ( \times 10^{-5} ) [7.92, 8.49]</td>
<td>4.8</td>
<td>0.86 [0.8627, 0.5641]</td>
<td>0.46</td>
<td>5.0 ( \times 10^{-5} ) [0.4621, 0.4640]</td>
<td>4.4</td>
</tr>
<tr>
<td>10^6</td>
<td>0.47 [0.4677, 0.4685]</td>
<td>0.55</td>
<td>5.0 ( \times 10^{-5} ) [0.7117, 0.7130]</td>
<td>4.7</td>
<td>0.71 [0.7117, 0.7130]</td>
<td>0.20</td>
<td>5.0 ( \times 10^{-5} ) [0.2041, 0.2054]</td>
<td>4.6</td>
</tr>
</tbody>
</table>
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
Modelling therapeutic resistance dynamics of lymphoma mice cohorts.

$A$ 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:

$$
A \xrightarrow{A_A} A + A \\
A \xrightarrow{A_M} A + M \\
M \xrightarrow{M_M} M + M \\
M + R \xrightarrow{rM_M} M + M \\
A + R \xrightarrow{rA_A} M + M \\
A \xrightarrow{d} R \\
$$

when treatment is applied

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.

<table>
<thead>
<tr>
<th>Name</th>
<th>Value (days$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitive proliferation rate $A_A$</td>
<td>0.82</td>
</tr>
<tr>
<td>Resistant proliferation rate $M_M$</td>
<td>0.75</td>
</tr>
<tr>
<td>Mutation rate $A_M$</td>
<td>$5 \times 10^{-5}$</td>
</tr>
<tr>
<td>Carrying capacity $K$</td>
<td>$2.6 \times 10^{10}$</td>
</tr>
<tr>
<td>Sensitive death rate due to treatment $d$</td>
<td>2.0</td>
</tr>
<tr>
<td>Replacement rate $r$</td>
<td>0.7</td>
</tr>
</tbody>
</table>

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.
Modelling therapeutic resistance dynamics of lymphoma mice cohorts.

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
3.2 Results

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.
Modelling therapeutic resistance dynamics of lymphoma mice cohorts.

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 periodic 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 $2.5 \times 10^5$ 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
4.1 Introduction

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 func-
tion 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 branch-
ing 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 patients (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.
4.2 Results

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-RAR$\alpha$ and CBF translocations (CBFB-MYH11 and RUNX1-RUNX1) 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 \textit{HOXA9} impacts AML subtypes, I look at the distribution of French-American-British (FAB, named M0-M7), and molecular classifications among the two \textit{HOXA9} cohorts. Distributions show that different \textit{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-\textit{HOXA9} cohort while all 10 patients with M3 AML are observed in the low-\textit{HOXA9} cohort. This specific distribution among cohorts suggests the important role of \textit{HOXA9} expression in the clinical characteristics of blood disease patients.

In light of these findings, I looked to characterise the common features of \textit{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 \textit{HOXA9} stratifies patients with different molecular classification (Fig. 4.1D). The 10 patients with PML-RAR\(\alpha\) translocation and the 9 with core binding factor (CBF) translocations, RUNX1-RUNXT1 and CBFB-MYH11, are all found in the low-\textit{HOXA9} cohort. Within the high cohort, the specific cytogenetic subtypes are complex cytogenetics and the MLL/NUP98 translocation. MLL-induced leukaemia has been linked to high \textit{HOXA9} [332], while M3 AML subtype is characterised by PML-RAR\(\alpha\) translocation and low \textit{HOXA9} in the literature [337]. Low \textit{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 \textit{HOXA9} expression and the M0 and M5 FAB subtypes as well as complex cytogenetics. Finally, \textit{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-\textit{HOXA9} cohort are significantly older and have a higher WBC. They also tend to display higher percentage of blasts in the bone marrow.

PML-RAR\(\alpha\), RUNX1-RUNXT1 (AML1-ETO) or CBFB-MYH11 chromosomal abnormalities confer good prognosis in AML patients [138, 339]. All these aberrations are linked to low \textit{HOXA9} expression which also exhibits good survival prognosis among patients compared to high expression. To confirm that high \textit{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 \textit{HOXA9} expression. Survival curves and log rank tests within both subtypes \((p = 9 \times 10^{-4} \text{ for M2 patients and } p = 0.051 \text{ for M4 patients})\) confirm the poor prognosis marker function of high \textit{HOXA9} (Fig. 4.2). These results as well as the diverging clinical features among patient cohorts support the switch role of \textit{HOXA9} within AML diseases.
4.2 Results

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.
HOXA9 acts as an epigenetic switch in blood malignancies.

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

Based on this \textit{JAK2/TET2/HOXA9} motif, the MPN model is extended to reproduce the observed biological differences between patients with different combinations of \textit{JAK2} and \textit{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).
4.2 Results

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 Temporal 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...
HOXA9 acts as an epigenetic switch in blood malignancies.

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 CEBPα [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 phospholylated 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, RUNXI links those hematopoietic genes to the JAK2/TET2/HOXA9 motif thanks to RUNXI activation by HOXA9. RUNXI 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. RUNXI is found in the earliest stages of haematopoiesis and is essential for a fully functional hematopoiesis [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.
### 4.2 Results

<table>
<thead>
<tr>
<th>Upstream</th>
<th>Interaction</th>
<th>Downstream</th>
<th>Reference</th>
<th>Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>TET2</td>
<td>Activates</td>
<td>HOXA9</td>
<td>[351]</td>
<td>in vitro/in vivo murine</td>
</tr>
<tr>
<td>JAK2</td>
<td>Activates</td>
<td>HOXA9</td>
<td>[348]</td>
<td>in vitro human</td>
</tr>
<tr>
<td>HOXA9</td>
<td>Activates</td>
<td>HOXA9</td>
<td>[334]</td>
<td>in vitro murine and human</td>
</tr>
<tr>
<td>HOXA9</td>
<td>Activates</td>
<td>RUNX1</td>
<td>[375, 376]</td>
<td>ex vivo /in vivo murine</td>
</tr>
<tr>
<td>RUNX1</td>
<td>Activates</td>
<td>SPI1</td>
<td>[380]</td>
<td>in vitro human</td>
</tr>
<tr>
<td>RUNX1</td>
<td>Activates</td>
<td>GATA1</td>
<td>[378, 379]</td>
<td>in vitro human/ in vivo</td>
</tr>
<tr>
<td>SPI1</td>
<td>Inhibits</td>
<td>GATA1</td>
<td>[373]</td>
<td>in vitro/in vivo human</td>
</tr>
<tr>
<td>GATA1</td>
<td>Inhibits</td>
<td>SPI1</td>
<td>[373]</td>
<td>in vitro/in vivo human</td>
</tr>
<tr>
<td>SPI1</td>
<td>Activates</td>
<td>CEBPα</td>
<td>[356]</td>
<td>in vitro murine and human</td>
</tr>
<tr>
<td>CEBPα</td>
<td>Inhibits</td>
<td>E2F1</td>
<td>[363]</td>
<td>in vivo in vitro murine</td>
</tr>
<tr>
<td>CEBPα</td>
<td>Activates</td>
<td>GMP differentiation</td>
<td>[358]</td>
<td>in vivo murine</td>
</tr>
<tr>
<td>E2F1</td>
<td>Activates</td>
<td>GMP expansion</td>
<td>[361]</td>
<td>in vivo murine</td>
</tr>
<tr>
<td>RUNX1</td>
<td>Inhibits</td>
<td>MYB</td>
<td>Prediction</td>
<td></td>
</tr>
<tr>
<td>MYB</td>
<td>Activates</td>
<td>CMP expansion</td>
<td>[360]</td>
<td>in vivo murine</td>
</tr>
<tr>
<td>TET2</td>
<td>Activates</td>
<td>SC self renewal</td>
<td>[168]</td>
<td>in vitro/vivo mice/human PDXs</td>
</tr>
<tr>
<td>GATA1</td>
<td>Activates</td>
<td>KLF1</td>
<td>[367]</td>
<td>in vitro/in vivo murine</td>
</tr>
<tr>
<td>GATA1</td>
<td>Inhibits</td>
<td>MYC</td>
<td>[381]</td>
<td>in vitro murine</td>
</tr>
<tr>
<td>MYC</td>
<td>Activates</td>
<td>MEP expansion</td>
<td>[381]</td>
<td>in vitro murine</td>
</tr>
<tr>
<td>KLF1</td>
<td>Activates</td>
<td>Erythroid differentiation</td>
<td>[367]</td>
<td>in vitro murine</td>
</tr>
<tr>
<td>TET2</td>
<td>Activates</td>
<td>TET2p</td>
<td>[368]</td>
<td>in vitro human</td>
</tr>
<tr>
<td>JAK2</td>
<td>Activates</td>
<td>TET2p</td>
<td>[368]</td>
<td>in vitro human</td>
</tr>
<tr>
<td>TET2p</td>
<td>Activates</td>
<td>KLF1</td>
<td>[368]</td>
<td>in vitro human</td>
</tr>
<tr>
<td>JAK2</td>
<td>Activates</td>
<td>STAT3</td>
<td>[382, 383]</td>
<td>in vitro/in vivo murine and human</td>
</tr>
<tr>
<td>JAK2</td>
<td>Activates</td>
<td>STAT5</td>
<td>[382, 383]</td>
<td>in vitro/in vivo murine and human</td>
</tr>
<tr>
<td>STAT3</td>
<td>Activates</td>
<td>SPI1</td>
<td>[372]</td>
<td>in vitro murine</td>
</tr>
<tr>
<td>STAT5</td>
<td>Activates</td>
<td>AKT</td>
<td>[369]</td>
<td>Human PV transplantation (mice)</td>
</tr>
<tr>
<td>AKT</td>
<td>Activates</td>
<td>GATA1</td>
<td>[369]</td>
<td>Human PV transplantation (mice)</td>
</tr>
<tr>
<td>STAT5</td>
<td>Activates</td>
<td>MAPK</td>
<td>[370]</td>
<td>in vivo murine</td>
</tr>
<tr>
<td>MAPK</td>
<td>Activates</td>
<td>BCL2L1</td>
<td>[370]</td>
<td>in vivo murine</td>
</tr>
<tr>
<td>BCL2L1</td>
<td>Activates</td>
<td>MEP expansion</td>
<td>[371]</td>
<td>in vitro human</td>
</tr>
<tr>
<td>TET2</td>
<td>Activates</td>
<td>MEP expansion</td>
<td>[158]</td>
<td>in vivo murine</td>
</tr>
<tr>
<td>MAPK</td>
<td>Activates</td>
<td>JNK1</td>
<td>[384]</td>
<td>in vitro human</td>
</tr>
<tr>
<td>JNK1</td>
<td>Activates</td>
<td>ITCH</td>
<td>[385]</td>
<td>in vitro human</td>
</tr>
<tr>
<td>ITCH</td>
<td>Inhibits</td>
<td>NOTCH</td>
<td>[386]</td>
<td>in vitro human</td>
</tr>
<tr>
<td>NOTCH</td>
<td>Inhibits</td>
<td>GMP expansion</td>
<td>[387]</td>
<td>in vivo murine</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Stem Cell Self Renewal</th>
<th>WT</th>
<th>TET2</th>
<th>JAK2</th>
<th>TET2 first</th>
<th>JAK2 first</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMP expansion</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GMP expansion</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GMP differentiation</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Erythroid differentiation</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MEP expansion</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

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.
### Target Functions of the BMA Model Variables for the MPN Network

<table>
<thead>
<tr>
<th>Node</th>
<th>Target Function</th>
<th>References</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOXA9</td>
<td>( JAK2^* TET2 + 2 \times \max(0, (HOXA9-1)) \times \max(0, JAK2-1) )</td>
<td>[348, 351, 334]</td>
<td>Activation by ( JAK2, TET2 ) and itself. Memory property prediction</td>
</tr>
<tr>
<td>SPI1</td>
<td>( STAT3 + \min( (RUNX1-1), (1-GATA1) ) )</td>
<td>[380, 373, 369]</td>
<td>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 ).</td>
</tr>
<tr>
<td>GATA1</td>
<td>( AKT+RUNX1-\max(SPI1,1) )</td>
<td>[378, 373, 372]</td>
<td>Activation by ( AKT ) and ( RUNX1 ) and inhibition loop with ( SPI1 ). Maximum function necessary as loss of ( SPI1 ) does not increase erythroid differentiation.</td>
</tr>
<tr>
<td>MEP</td>
<td>min( BCL2L1+TET2 , max( MYC, BCL2L1))</td>
<td>[381, 371, 158, 388]</td>
<td>( TET2 ) loss reduces erythroid progenitors by skewing toward myeloid lineage. ( JAK2 ) positive regulation of the erythroid lineage is stronger than ( TET2 ) through ( MYC ) activation.</td>
</tr>
</tbody>
</table>

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
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 CEBPa 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 (βα) 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 βα segment [402]. Once released in the extracellular space, βα 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 \textit{HOXA9}/\textit{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 \textit{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 \textit{HOXA9} and \textit{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 separate quite early in hematopoiesis and both have their own characteristics. Thus, the antagonistic \textit{HOXA9}/\textit{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 \textit{HOXA9}/\textit{APP} patient cohorts (chi-squared test, $p = 1.4 \times 10^{-21}$). As expected, the PML-RAR\textalpha 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 \textit{HOXA9} and \textit{APP} markers which could help AML diagnoses and clinical outcome prediction. This also supports the well-known role of \textit{HOXA9} in blood diseases, as well as the important \textit{APP} role in other diseases than brain disorders.
5.2 Results

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-RUNX1 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](image)

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 \textit{APP} peak. To do so, the 32 patients within the low \textit{APP} peak are split into two new subsets of patients with low and high expression for \textit{APP}. Survival analysis shows that patients within the low peak with a low \textit{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 \textit{APP} expression (Fig. 5.5). In the rest of the chapter, I refer to the 32 patients the low peak of \textit{APP} as the "low \textit{APP}" patients as opposed to the "high \textit{APP}" group (Fig. 5.2). I define the 16 patients within the low peak with low expression for \textit{APP} as the "low-low \textit{APP}" cohort and the patients in the low peak with high expression for \textit{APP} as the "low-high \textit{APP}" patients.

![Survival of AML1-ETO patients within the low and high APP peaks.](image)

**Fig. 5.4 Survival of AML1-ETO patients within the low and high \textit{APP} peaks.** The patients with AML1-ETO present in the low \textit{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 \textit{APP} patients supports the critical effect of reduced \textit{APP} expression in AML. I compare the available clinical characteristics of low-low and low-high \textit{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 \textit{APP} biological functions in blood, I compare low-low and low-high \textit{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], \( KNYU \) [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
5.2 Results

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.
5.2 Results

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.

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. 5.10 **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.
5.2 Results

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, BAAALC, a poor prognosis marker in leukemia already identified in AML analyses, is also upregulated in the low-low APP cohort. KLF4, one of BAAALC known repressed gene target [453], is downregulated in the same cohort. To conclude on these pediatric
5.2 Results

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 $TRBC2$.](image)

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 ($p = 0.12$).](image)

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.
5.2 Results

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 \textit{BAALC} downstream target genes is not significantly different between the \textit{APP} cohorts in any datasets (Fig. 5.17). Thus, it can be assumed that low \textit{APP} poor survival is not linked to \textit{BAALC}.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig5.17.png}
\caption{\textbf{APP is not correlated to \textit{BAALC} downstream target expression.} Expression of \textit{BAALC} upregulated (A) and downregulated (B) downstream genetic targets in the low-low versus low-high \textit{APP} AML patients do not show any significant trend that could validate a potential connection between \textit{BAALC} and \textit{APP} to explain the survival diagnosis of these patients. Similar results are found in the pediatric and adult MPAL datasets.}
\end{figure}

\subsection{5.2.5 Exploring the degree of similarity between the low-low \textit{APP} AML cohort and MPAL.}

To further explore the possibility of low \textit{APP} expression to be a marker of MPAL, I investigate the degree of similarity between the low-low \textit{APP} AML cohort and MPAL. I assume that if the low-low \textit{APP} AML patients share MPAL characteristics, the number of genes that are differentially expressed between MPAL patients and low-low \textit{APP} AML patients should be lower than against AML patients with higher \textit{APP} expression. Due to format issues, the adult MPAL data \cite{250} cannot be compared to the AML patients. Further analyses therefore use pediatric MPAL sequencing data \cite{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 \cite{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 \cite{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 \textit{APP} patients. In standard DESeq2 analyses, DEGs are selected using the log2 fold change (LFC) and the adjusted p-value (\textit{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 \( p_{adj} \) and LFC that allow a decent number of DEGs between MPAL and AML groups (Fig. 5.18A,B,C). I first set \( p_{adj} < e^{-80} \) 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 \( p_{adj} < e^{-20} \). 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.
5.2 Results

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 \( \text{pad} \) 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, \( \text{pad} < e^{-80} \) 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 \texttt{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 \texttt{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 \texttt{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 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 \( pad j < e^{-20} \) 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 βα 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 βα 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 βα 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 separated 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
Boolean motif inference applied to HOXA9 and APP in AML.

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).
6.2 Results

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:

<table>
<thead>
<tr>
<th>Input</th>
<th>A</th>
<th>B</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

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
Boolean motif inference applied to HOXA9 and APP in AML.

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:

<table>
<thead>
<tr>
<th>Input</th>
<th>A</th>
<th>B</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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: $\text{TF-Input-Input/False-A/False-B/False-Output/False} = \text{True}$. The first part of the variable $\text{TF-Input}$ gives the name of the node, here Input, and the second part $\text{Input/False-A/False-B/False-Output/False}$ 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.

![Fig. 6.4 Example of a one-step simulation using target functions.](image)

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.
6.2 Results

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.
Boolean motif inference applied to HOXA9 and APP in AML.

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 \textit{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 BioModel-Analyzer (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.

\subsection*{6.2.2 Identification of motifs for the \textit{HOXA9/APP} AML cohorts}

As shown in the previous chapters 4 and 5, the analysis of the TCGA AML data highlighted the importance of \textit{HOXA9} as a marker for AML stratification in a first instance, and then the unforeseen \textit{APP} role in leukemia. Further work on the data showed the presence of three clusters of patients with distinct \textit{HOXA9} and \textit{APP} expression (Fig. 5.1B). These three clusters are unexpected: no link between \textit{HOXA9} and \textit{APP} has been previously identified in the literature and the role of \textit{APP} in AML is not well understood [517, 414, 518]. To untangle why and how \textit{HOXA9} and \textit{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 \textit{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, \textit{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 \textit{APP} and \textit{HOXA9}, called the high/high cohort, low
expression for $\textit{HOXA9}$ and high expression for $\textit{APP}$, the low/high cohort and finally high expression for $\textit{HOXA9}$ and low expression for $\textit{APP}$, 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 $\textit{APP}$ and $\textit{HOXA9}$ 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 $\textit{HOXA9}$ and $\textit{APP}$ nodes. Network dynamics are described in the following paragraphs.

![Fig. 6.6 Motif search for $\textit{HOXA9}$/$\textit{APP}$ clusters identifies three potential networks](image)

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

If the marker activates both $\textit{APP}$ and $\textit{HOXA9}$ (Fig. 6.6A and Table 6.1), the low/high cohort should have a low expression for the marker (Trace 1). Here, $\textit{APP}$ 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 $\textit{HOXA9}$ and $\textit{APP}$ expression and thus the high/high cohort is expected to have a high expression for the marker gene (Trace 2). Finally, if $\textit{HOXA9}$ is turned on before $\textit{APP}$ while the marker stays off, $\textit{APP}$ expression remains low due to $\textit{HOXA9}$ inhibition and lack of activation (Trace 3). Then, the high/low cohort should have a low expression for the marker gene.

<table>
<thead>
<tr>
<th>Trace 1</th>
<th>Trace 2</th>
<th>Trace 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker</td>
<td>$\textit{HOXA9}$</td>
<td>$\textit{APP}$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.1 *Traces of the three $\textit{HOXA9}$/$\textit{APP}$ clusters when the marker activates both genes.*
If the marker inhibits both \textit{APP} and \textit{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: \textit{APP} turns itself on if and only if \textit{HOXA9} is not expressed before which leads to the low/high case (Trace 2). However, if \textit{HOXA9} is on and stays on thanks to its positive self loop, \textit{APP} expression is inhibited by \textit{HOXA9} and the marker, which results in the high/low situation (Trace 3).

<table>
<thead>
<tr>
<th>Trace 1</th>
<th>Trace 2</th>
<th>Trace 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker</td>
<td>\textit{HOXA9}</td>
<td>\textit{APP}</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.2 \textbf{Traces of the three \textit{HOXA9}/\textit{APP} clusters when the marker inhibits both genes.}

The last scenario illustrates the case in which the marker inhibits \textit{HOXA9} gene and activates \textit{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, \textit{HOXA9} is turned on before the marker and stays in this state thanks to its positive feedback loop. However in the low/high cohort, \textit{HOXA9} has low expression when the marker is turned on which results in its inhibition while \textit{APP} is activated by the marker (Trace 2). Finally, the absence of marker activates \textit{HOXA9} while \textit{APP} expression stays low, which leads to the high/low case where patients should have low expression for the marker (Trace 1).

<table>
<thead>
<tr>
<th>Trace 1</th>
<th>Trace 2</th>
<th>Trace 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker</td>
<td>\textit{HOXA9}</td>
<td>\textit{APP}</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.3 \textbf{Traces of the three \textit{HOXA9}/\textit{APP} clusters when the marker inhibits \textit{HOXA9} gene and activates \textit{APP}.}

\subsection*{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 \textit{HOXA9} inhibition by \textit{APP} or \textit{APP} inhibition by \textit{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 \textit{HOXA9} and \textit{APP} connections. No direct interaction between \textit{HOXA9} and \textit{APP} can be found. I next look for genes that are downstream of \textit{HOXA9} and upstream of \textit{APP} to explore intermediate genes. One gene fits the inhibition of \textit{APP} by \textit{HOXA9}: TGF\(\beta\)1. Similar work is performed for \textit{HOXA9} inhibition by \textit{APP}, but no gene matches the request. In Omnipath, TGF\(\beta\)1 is marked as repressed by \textit{HOXA9} and as a positive regulator of \textit{APP}. While many references for the activation of \textit{APP} by TGF\(\beta\)1 are available [519–521], Omnipath only provides one questionable source for TGF\(\beta\)1 inhibition by \textit{HOXA9}. After literature curation, one paper showing \textit{HOXA9} repression of TGF\(\beta\)1 [522] is found. \textit{HOXA9} inhibition of \textit{APP} is confirmed thanks to another study demonstrating the role of \textit{HOXA9} as a co-repressor partner of SMAD6 [523]. The latter has been shown to repress \textit{SMAD4} activity which positively regulates \textit{APP} [521] by competing for the binding with SMAD1 [524]. Collectively, an inhibition of \textit{APP} by \textit{HOXA9} via the TGF\(\beta\) 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 \textit{HOXA9} by \textit{APP} and focus on networks in which \textit{HOXA9} inhibits \textit{APP} and activates itself [334].

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

11 markers, I expect a gene to be a good marker if its p-value is lower than $4 \times 10^{-3}$ (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 $8 \times 10^{-3}$, 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.
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 summarise 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 $8 \times 10^{-3}$ for the Bonferroni test (threshold is $4 \times 10^{-3}$).

### 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 \textit{TET2} is mutated before \textit{JAK2}, its function loss decreases the gene expression of some \textit{JAK2} downstream targets. Consequently, \textit{JAK2} inhibitor drugs become inefficient for those patients despite the \textit{JAK2} mutation as a result of the evolution switch property of \textit{JAK2} and \textit{TET2} common downstream target \textit{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 \textit{APP} expression as well as high \textit{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.

![Diagram of cancer progression](image)

**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 \textit{JAK2} and \textit{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 \( CEBP\alpha \) and \( CEBP\beta \) 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 \( CEBP\alpha \) expression. If upregulation of \( GATA2 \) in GMPs is followed by the downregulation of \( CEBP\alpha \), then GMPs differentiate into eosophil progenitors. On the other hand, maintained \( CEBP\alpha \) 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 \textit{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 \textit{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


References


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


[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.


References


