Modélisation mathématique des néoplasmes myéloprolifératifs : de leur développement à leur traitement par Interféron alpha

Gurvan Hermange

université
PARIS-SACLAY

Modélisation mathématique des néoplasmes myéloprolifératifs : de leur développement à leur traitement par Interféron alpha

Mathematical modeling of myeloproliferative neoplasms: from their development to their treatment with Interferon alpha

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°573 INTERFACES
Spécialité de doctorat : Mathématiques Appliquées
Graduate School : Sciences de l'ingénierie et des systèmes. Référent : CentraleSupélec

Thèse préparée dans l'unité de recherche **Mathématiques et Informatique pour la Complexité et les Systèmes** (Université Paris-Saclay, CentraleSupélec), sous la direction de **Paul-Henry COURNÈDE**, Professeur, la co-direction d'**Isabelle PLO**, Docteur

Thèse soutenue à Paris-Saclay, le 08 février 2023, par

Gurvan HERMANGE

Composition du Jury

Membres du jury avec voix délibérative

Emmanuel VAZQUEZ Pr., Université Paris-Saclay	Président
Fabien CRAUSTE DR CNRS, HDR, Université Paris Cité	Rapporteur & Examinateur
Chantal GUIHENNEUC Pr., Université Paris Cité	Rapporteur & Examinatrice
Raphaël ITZYKSON Pr., Université Paris Cité	Examinateur

THESE DE DOCTORAT

NNT : 2023UPAST025

université PARIS-SACLAY | ÉCOLE DOCTORALE Interfaces: matériaux, systèmes, usages

Titre : Modélisation mathématique des néoplasmes myéloprolifératifs : de leur développement à leur traitement par Interféron alpha

Mots clés : Modèles mathématiques – Processus stochastiques – Cancérologie – Statistique Bayésienne – Hématopoïèse

Résumé : L'hématopoïèse est le processus de création de nos cellules sanguines à partir de cellules souches. C'est un processus complexe qui n'est pas encore totalement compris et qui fait intervenir de nombreux types cellulaires. L'apparition de certaines mutations au niveau des cellules souches hématopoïétiques peut entraîner le développement de certains cancers du sang tels que les néoplasmes myéloprolifératifs (NMP). Ces pathologies malignes induisent la surproduction d'un certain type de cellules, un risque de thromboses ou d'accidents cardio-vasculaires et peuvent dégénérer en leucémies.

Pour les traiter, voire induire une rémission, une thérapie à l'Interféron alpha (IFNα) a montré de bons résultats pour certains patients. L'action de ce traitement sur les NMP reste cependant mal compris, ce qui limite l'efficacité des thérapies actuelles.

Dans cette thèse, nous étudions sous l'angle de la modélisation mathématique l'hématopoïèse, le développement des NMP puis leur traitement à l'IFNα.

Title : Mathematical modeling of myeloproliferative neoplasms: from their development to their treatment with Interferon alpha

Keywords : mathematical models – Stochastic processes – Oncology – Bayesian inference – Hematopoiesis

Abstract : Hematopoiesis is the process of creating blood cells from stem cells. It is a complex process that is not yet fully understood and involves many cell types. The occurrence of specific mutations in hematopoietic stem cells can lead to the development of certain blood cancers, such as myeloproliferative neoplasms (MPN). These malignant pathologies induce the overproduction of blood cells, risk of thrombosis or cardiovascular events and can degenerate into leukemia. IFNα therapy has shown promising results for MPN patients, inducing for some of them a molecular remission.

However, the action of this treatment on MPNs remains poorly understood, limiting the efficacy of current therapies.

In this thesis, we study from a mathematical modeling perspective the hematopoiesis, the development of NMPs and then their treatment with IFNα.

Remerciements

Ce mémoire de thèse est plutôt long. Nul doute que je pourrais encore le rallonger avec cette section consacrée aux remerciements. Mais, une fois n'est pas coutume, je vais essayer de faire preuve d'un peu de concision.

Mes remerciements vont tout d'abord à Paul-Henry Cournède, mon directeur de thèse, qui dès le début de cette histoire a su me faire confiance. D'ailleurs, l'histoire commence avec lui. Après deux expériences en entreprise, alors que j'envisageais de reprendre mes études en mathématiques, je suis allé frapper à son bureau pour demander conseil. L'idée de faire une thèse était alors pour moi lointaine, si tant est qu'elle m'ait effleuré l'esprit. Paul-Henry a rendu cette idée concrète, possible. Il a dessiné avec moi les contours de mon nouveau projet professionnel, attentif à mes souhaits notamment en ce qui concerne mon futur domaine d'application, soucieux de me donner tous les moyens qui me permettraient de réussir. Je peux aujourd'hui affirmer que c'était un pari gagnant, bien au-delà de ce que j'aurais pu imaginer ; le travail de recherche m'a passionné et me passionne encore aujourd'hui.
Paul-Henry a été exactement le directeur de thèse qu'il me fallait. Il avait toujours des idées intéressantes de pistes à explorer - sans pour autant me les imposer - de telle façon qu'il me semblait que je n'aurais jamais pu me retrouver dans une impasse. Il me laissait l'autonomie dont j'avais besoin dans mon travail, tout en étant toujours soucieux de suivre régulièrement mes avancées. Et surtout, il faisait preuve d'un enthousiasme et d'une grande curiosité, deux ingrédients essentiels contribuant à ma motivation.
Enfin, Paul-Henry m'a également fait confiance pour encadrer des étudiants, et ce dès le début de ma thèse, m'offrant la possibilité de créer des projets en lien avec mes travaux de recherche, de transmettre ma passion et d'enseigner, ce qui sans nul doute m'a conforté dans mon choix de vouloir poursuivre dans l'enseignement et la recherche.

Vient ensuite ma rencontre avec Isabelle Plo, qui devint ma co-directrice de thèse. Jusqu'alors, j'avais un directeur de thèse, l'orientation mathématique de ma thèse - à savoir la modélisation et les statistiques Bayésiennes - ainsi que ma bourse doctorale. Restait à savoir précisément quel serait le domaine d'application de ma thèse. Par une belle matinée de septembre (permettez-moi - lecteur - d'édulcorer un peu lorsque j'ai oublié certains détails), nous nous étions rendus avec Paul-Henry à l'Institut Gustave Roussy pour rencontrer Isabelle, dans son bureau rempli d'orchidées. Elle nous présenta ses thématiques de recherche : l'hématopoïèse, les néoplasmes myéloprolifératifs, les mutations de *JAK2*, le traitement à l'Interféron alpha, etc. Bon, autant dire que je n'avais probablement pas compris grand-chose à l'issu de cette réunion. Mais Isabelle était passionnée par son sujet, passion qu'elle sut immédiatement nous transmettre. Elle avait l'intuition qu'on pourrait - grâce aux mathématiques - apporter des réponses aux questions qu'elle se posait ; elle sut nous motiver dès le début avec un sujet challengeant et de jolies données. Surtout, elle montra de l'intérêt pour les modèles mathématiques, cherchant à comprendre ce qu'on faisait et, réciproquement, elle fit preuve de temps et de patience pour m'expliquer petit à petit la biologie. Son implication et sa confiance furent deux éléments cruciaux sans lesquels ce travail de recherche n'aurait pas eu le même impact.

Bien, j'avais promis d'être concis, mais il reste tant de personnes à remercier !
De l'Institut Gustave Roussy, je souhaiterais remercier Amandine Tisserand, en thèse sous la direction d'Isabelle, qui réalisait les expériences sur lesquelles s'appuyaient mon travail d'inférence statistique. En plus de son don pour faire pousser les colonies de progéniteurs, Amandine a un talent pour réaliser de belles figures et faire passer les messages efficacement, ce qui est loin d'être anecdotique. Elle a contribué à la réalisation de nombreuses figures pour les articles scientifiques que nous avons écrits, dont plusieurs sont reprises dans ce manuscrit, et qui permettent de rendre compréhensible pour un public non-mathématicien certains de nos résultats. Je crois avoir également appris auprès d'elle et que mes figures sont un peu plus *biologistes-friendly* aujourd'hui qu'hier.

Toujours au sein de l'équipe d'Isabelle, je remercie Caroline Marty et Camélia Benlabiod, pour leur intérêt également pour les méthodes mathématiques, et le travail initié sur les données CyTOF. J'ai apprécié me faire des nœuds aux cerveaux lors de l'étude de leurs données, brainstormé avec elles sur la façon d'analyser et d'interpréter les résultats. J'ai l'impression d'avoir sans cesse changé ma méthode d'analyse, sans qu'elles ne perdent jamais patience. J'aurais aimé aboutir à plus de résultats - le travail reste cependant en cours - mais sans nul doute que j'ai beaucoup appris sur les techniques de cytométrie grâce à elles.
Plus généralement, je souhaiterais remercier l'ensemble des membres de l'unité de recherche à Gustave Roussy, avec qui j'ai eu plaisir à échanger, qui ont eu la patience de m'écouter parler à plusieurs reprises de modélisation mathématique, et pour qui il a bien fallu que je développe mes compétences en vulgarisation.

Dans cette histoire, une équipe de l'Institut Curie fait son apparition un peu plus tard, celle de Leïla Perié, que je souhaiterais remercier ainsi qu'Alessandro Donada. Ils ont ouvert de nouvelles perspectives à ce projet de recherche, perspectives aboutissant à la création du projet OptiMyN qui est l'objet de mon post-doc. Alessandro est intarissable quand il s'agit de parler de cellules souches et d'hématopoïèse, et c'est toujours un plaisir de discuter avec lui de ces sujets. Les conseils et feedbacks de Leïla, habituée à travailler à l'interface entre biologie et mathématiques, m'ont été d'une précieuse aide.

Il me faut maintenant recentrer le discours sur Gif-sur-Yvette, siège du laboratoire MICS. L'environnement y était parfait pour faire ma thèse. Même si un peu éloigné de Paris - il faut l'avouer - j'appréciais (et j'apprécie encore) venir au labo, aller au sport le midi avec les collègues (que ce soit pour du cross-fit, du badminton ou encore un peu de jogging), échanger sur les travaux des uns et des autres ou encore tester un nouveau thé. Parmi l'ensemble de mes collègues que je remercie, je voudrais mentionner en particulier Véronique Letort-Le Chevalier avec qui j'ai également beaucoup appris sur la modélisation du vivant, au travers cette fois-ci de l'enseignement et de ma participation en tant que chargé de TD sur certains de ses cours, Mahmoud Bentriou grâce à qui j'ai pu appliquer des méthodes ABC pour l'inférence des paramètres de certains de mes modèles, et enfin Antonin Della Noce et Brice Hannebicque, deux visages connus de mes années centraliennes à Châtenay-Malabry et que je retrouvais maintenant à Gif-sur-Yvette, comme si le labo m'était déjà familier avant même d'y mettre les pieds.

Pour finir, ou presque, petit détour par le sud de l'Europe, la Grèce, où travaille Samis Trevezas que je remercie pour son aide et son encouragement sur l'étude des temps de première division des cellules hématopoïétiques. Il me semble qu'il me resterait plein de choses à apprendre à ses côtés.

Pour terminer - en omettant néanmoins de nombreuses personnes, amis, collègues, que je ne cite pas nommément bien qu'elles le mériteraient - je remercie mes parents qui ont fourni des conditions plus que favorables à l'épanouissement de ce travail de thèse : pendant les deux confinements tout d'abord, avec au total près de neuf mois passés en Bretagne (les conditions y étant très favorables, peut-être ai-je fait un peu de zèle), puis au cours de l'été 2022, en partie passé dans l'appartement d'Erquy afin de rédiger ce manuscrit de thèse, face à la mer, entre une session de jogging sur la plage et une autre de paddle sur la mer.
Probablement, même, que les conditions nécessaires à l'épanouissement de ce travail de thèse avaient été créées bien avant le début de ma thèse.

Table des matières

Résumé

Dans cette thèse, nous étudions sous l'angle de la modélisation mathématique l'hématopoïèse, le développement des hémopathies malignes que sont les néoplasmes myéloprolifératifs (NMP) puis leur traitement à l'Interféron alpha (IFNα).

L'hématopoïèse est le processus de création de nos cellules sanguines à partir de cellules souches situées dans la moelle osseuse. C'est un processus complexe qui n'est pas encore totalement compris et qui fait intervenir de nombreux types cellulaires. L'apparition de certaines mutations au niveau des cellules souches hématopoïétiques peut entraîner le développement de certains cancers du sang tels que les NMP. Ces pathologies malignes induisent la surproduction d'un certain type de cellules, un risque de thromboses ou d'accidents cardio-vasculaires et peuvent dégénérer en leucémies. Pour les traiter, voire induire une rémission, une thérapie à l'IFNα a montré de bons résultats pour certains patients. L'action de ce traitement sur les NMP reste cependant mal compris, ce qui limite l'efficacité des thérapies actuelles.

Nous commençons cette thèse en étudiant l'hématopoïèse normale. Nous montrons tout d'abord, à partir de l'analyse de données obtenues par cytométrie de masse, que la prise en compte de nombreux marqueurs cellulaires, tant surfaciques qu'intracellulaires, montre une distribution des cellules hématopoïétiques suivant un continuum d'états. Néanmoins, les observations expérimentales utilisées pour la suite de ce travail de thèse reposeront sur l'utilisation de quelques marqueurs de surface et la mesure de l'expression de leur intensité par cytométrie de flux, justifiant notre répartition des cellules progénitrices suivant un ensemble discret de types cellulaires. Nous étudions alors la façon dont ces cellules, en fonction de leur type, prolifèrent et se différencient à court terme. Pour cela, nous modélisons les temps de division des cellules par différentes lois de probabilités que nous comparons entre elles, puis proposons un modèle plus complet de prolifération et différenciation que nous étudions en détail afin d'en estimer les paramètres.

Nous nous plaçons ensuite dans le cas pathologique, en étudiant le développement des NMP à partir de l'acquisition d'une mutation motrice (soit la mutation $JAK2^{V617F}$, soit la mutation au gène $CALR$). À partir d'un modèle stochastique formellement décrit par une chaîne de Markov à temps continu, puis une calibration de ce modèle en utilisant une méthode Approximate Bayesian Computation, nous mettons en évidence des différences dans l'âge d'acquisition des deux mutations motrices des NMP, la mutation $JAK2^{V617F}$ apparaissant plus tôt dans la vie (voire pendant la vie fœtale) comparée à la mutation $CALR^m$. Nous illustrons alors comment notre approche permet d'estimer un âge optimal pour détecter précocement la mutation $JAK2^{V617F}$.

Lorsque la maladie se déclenche, les patients sont pris en charge cliniquement. Le traitement à l'IFNα fait partie des thérapies prometteuses par sa capacité à induire une réponse moléculaire. Nous étudions alors une cohorte longitudinale de patients traités par IFNα pendant plusieurs années. À partir d'un modèle mathématique et d'une méthode d'inférence Bayésienne hiérarchique, nous obtenons une stratification des individus suivant différents critères : la mutation motrice du NMP, la zygosité ainsi que la dose moyenne reçue en début de traitement. Nous complexifions alors notre approche en étudiant les variations de posologie reçues sur 5 ans. Suite à une procédure de sélection de modèle, nous montrons alors comment nous pouvons déterminer pour chaque patient une dose personnalisée d'IFNα pour que l'individu commence à répondre au traitement. Enfin, nous montrons comment notre méthodologie peut s'intégrer dans un outil d'aide à la décision à destination des cliniciens. Cet outil, en cours de développement, sera accessible sous la forme d'une application web.

Summary

In this thesis, we study hematopoiesis, the development of hematological malignancies such as myeloproliferative neoplasms (MPN) and their treatment with Interferon alpha (IFNα) from the perspective of mathematical modeling.

Hematopoiesis is the process of creating blood cells from stem cells located in the bone marrow. It is a complex process that is not yet fully understood and involves many cell types. The occurrence of specific mutations in hematopoietic stem cells can lead to the development of certain blood cancers such as MPN. These malignant pathologies induce the overproduction of blood cells, risk of thrombosis or cardiovascular events and can degenerate into leukemia. IFNα therapy has shown promising results for MPN patients, inducing for some of them a molecular remission. However, the action of this treatment on MPN remains poorly understood, limiting the efficacy of current therapies.

We begin this thesis by studying normal hematopoiesis. Based on the analysis of mass cytometry data, we first show that taking into account numerous cellular markers, both surface and intracellular, shows a distribution of hematopoietic cells along a continuum of states. Nevertheless, the experimental observations used for the remainder of this thesis will rely on using a few surface markers and measuring their intensity expression by flow cytometry, justifying our distribution of progenitor cells along a discrete set of cell types. We then study how these cells, depending on their type, proliferate and differentiate in the short term. To do so, we model the cell division times by different probability distributions that we compare. Then, we propose a complete model of proliferation and differentiation that we study in detail to estimate its parameters.

We then place ourselves in the pathological case by studying the development of MPN from the acquisition of a driver mutation (either the $JAK2^{V617F}$ mutation or the $CALR^m$ mutation). From a stochastic model formally described by a continuous-time Markov chain, then calibration of this model using an Approximate Bayesian Computation method, we highlight differences in the age of acquisition of the two MPN driver mutations, $JAK2^{V617F}$ appearing earlier in life (or even during fetal life) compared to the $CALR^m$ mutation. We then illustrate how our approach allows estimating an optimal age for early detection of the $JAK2^{V617F}$ mutation.

When disease onset occurs, patients begin to be followed up clinically. IFNα treatment is among the promising therapies by its ability to induce a molecular response. We then study a longitudinal cohort of patients treated with IFNα for several years. Using a mathematical model and a hierarchical Bayesian inference method, we obtain a stratification of individuals according to different criteria : the MPN driver mutation, the zygosity, and the average dose received at the beginning of treatment. We then complexify our approach by studying the variations of dosage received over five years. Following a model selection procedure, we then show how we can determine for each patient a personalized dose of IFNα for the individual to begin responding to treatment. Finally, we show how our methodology can be integrated into a decision-support tool for clinicians. This tool, currently under development, will be accessible as a web application.

Avant-propos

Dans ce manuscrit, je présente l'intégralité de mes travaux de thèse réalisés entre octobre 2019 et septembre 2022. Mon doctorat a été réalisé au sein de l'École Doctorale Interfaces et, en effet, mes travaux se situent à l'interface entre mathématiques et biologie. Ainsi, j'ai bénéficié dans l'encadrement de ma thèse des connaissances et de l'expertise de Paul-Henry Cournède, du laboratoire MICS à l'École CentraleSupélec, spécialisé en modélisation mathématique, statistique et inférence Bayésienne, et d'Isabelle Plo, de l'équipe *"Des cellules souches hématopoïétiques aux mégacaryocytes"* à l'Institut Gustave Roussy, spécialisée dans l'étude des néoplasmes myéloprolifératifs. L'objectif de ma thèse, qui sera présenté plus en détail dans les chapitres suivants, était alors d'étudier sous l'angle de la modélisation mathématique et de l'inférence statistique l'hématopoïèse, l'apparition et le développement des néoplasmes myéloprolifératifs, puis leur traitement par Interféron alpha.
Pour rendre compte de la pluridisciplinarité de mes travaux, j'ai souhaité - dans l'écriture de mon manuscrit - qu'ils soient à la fois accessibles à des lecteurs mathématiciens s'intéressant à la modélisation de l'hématopoïèse et à des lecteurs biologistes s'intéressant à la mise en place de méthodes mathématiques pour l'analyse de leurs données expérimentales.
Ce manuscrit s'accompagne d'annexes qui sont disponibles en ligne au lien suivant :

`https://gitlab-research.centralesupelec.fr/2012hermangeg/supplementary-material-phd`

Les méthodes présentées dans ce manuscrit ont, pour la plupart, fait l'objet d'une implémentation en langage de programmation Julia. Le code est disponible sur GitLab au lien suivant.

`https://gitlab-research.centralesupelec.fr/2012hermangeg`

Enfin, les images de couverture qui illustrent chacun des chapitres font partie du projet Haematopoiesis par l'artiste et photographe Rubén Álvarez / *these images are part of the Haematopoiesis Project by Rubén Álvarez*. Elles sont reproduites dans ce manuscrit de thèse avec l'accord de l'auteur et, plus généralement, *"these images are royalty free for editorial and non-commercial uses in magazines, medical journals and everything else related with the research and treatment of cancer and rare diseases"*. Tel que présenté par l'artiste, *"before being photographed, these cells were created one by one with paint, liquid thickener, ferrofluid and magnets"*.
Personnellement, parmi les près de 300 images produites par l'artiste, j'ai choisi pour chacun de mes chapitres celle que je trouvais le plus en lien avec le sujet traité. Aurez-vous la même interprétation que moi, lecteur ?

Liste des principales abréviations utilisées dans ce manuscrit

Abréviations	Définition du terme
ABC	Approximate Bayesian Computation
AIC	Critère d'Information d'Akaike
BIC	Critère d'Information Bayésien
CF	Fraction Clonale
CMA-ES	Covariance Matrix Adaptation – Evolution Strategy
CTMC	Continuous Time Markov Chain
CyTOF	Cytometry by Time of Flight
DIC	Deviance Information Criterion
het	hétérozygote
hom	homozygote
HP	Hyper-paramètres
HPC	Hematopoietic Progenitor Cell
HSC	Cellule souche hématopoïétique
HSPC	Hematopoietic Stem and Progenitor cells
IFN	Interféron
k-NN	k-Nearest Neighbors
Lin-	Lineage negative
MCMC	Markov Chain Monte-Carlo
MFP	Myélofibrose primaire
MK	Mégacaryocyte
MPP	Progéniteur multipotent
NMP	Néoplasme myéloprolifératif
ODE	Équations différentielles ordinaires
PV	Poylglobulie de Vaquez
TE	Thrombocytémie essentielle
VAF	Variant Allele Frequency
WT	Wild-Type (type sauvage ou sain)

Chapitre 1
Introduction

Résumé

Dans ce chapitre, nous introduisons le contexte dans lequel s'inscrit le travail présenté tout au long de ce mémoire de thèse.
Nous commençons par décrire ce qu'est l'hématopoïèse - processus par lequel des cellules souches contribuent à la formation de toutes nos cellules sanguines - puis nous nous concentrons sur le dérèglement de la voie de signalisation JAK2/STAT suite à l'occurrence d'anomalies génétiques qui seront à l'origine d'hémopathies malignes appelées néoplasmes myéloprolifératifs. Un traitement prometteur pour ces cancers du sang est le traitement à l'interféron alpha.
Notre approche pour l'étude des néoplasmes myéloprolifératifs et de leur traitement à l'interféron alpha sera basée sur l'utilisation et l'analyse de modèles mathématiques. Nous présentons alors un état de l'art de modèles ayant permis de décrire l'hématopoïèse, que ce soit dans un cas non pathologique, dans le cas d'hémopathies ou encore pour l'étude de l'effet de certains traitements. L'analyse de tels modèles, une fois ces derniers construits, repose sur sur des méthodes mathématiques dont nous illustrerons les enjeux et difficultés, tels que les problématiques d'identifiabilité ou d'estimation des paramètres. Enfin, nous présenterons le plan de ce mémoire.

Une partie de ce chapitre a fait l'objet d'une revue (Hermange et al., Hématologie 2022).

Abstract

Hematopoiesis is the process by which all our blood cells (red blood cells, platelets, white blood cells) can be produced by some hematopoietic stem cells (HSC). In this chapter, we first describe hematopoiesis in terms of the cell populations involved (static view) and the differentiation and proliferation process (dynamical view). We also introduce the role of cytokines, especially in the JAK/STAT signalling pathways. When this pathway is altered because of mutation acquisition, it can lead to a deregulation of hematopoiesis and the development of hematological malignancies. Myeloproliferative Neoplasms (MPN) are clonal blood cancers in which one or more mature blood cell types are produced in excess. These diseases are due to genetic abnormalities affecting the JAK/STAT signalling pathway. In particular, the occurrence of MPN is mainly associated with somatic gain-of-function mutations in the genes encoding JAK2 ($JAK2^{V617F}$), CALR ($CALR^m$, type 1 del52 and type 2 ins5), or the thrombopoietin receptor (MPL^m) that are acquired in hematopoietic stem cells. Interferon alpha is a promising treatment that has been shown to induce some molecular remission in MPN patients, which we will study in this manuscript.

After having introduced the biological and clinical context, we present an overview of the state of the art concerning the mathematical models of hematopoiesis and its disorders, as well as methodological issues frequently met when dealing with experimental observations and models : data pretreatment, statistical inference, sensitivity analysis, model selection, uncertainty quantification.
Finally, we present the content of the manuscript.

Table des matières

1 Les promesses d'une médecine personnalisée

Le développement d'une médecine personnalisée, rendue possible par l'augmentation des capacités de calcul et l'accès facilité à des données biomédicales, fait partie des enjeux majeurs de la recherche en santé. Ainsi, début 2022, L'Institut Gustave Roussy dévoilait dans son plan 2030 ses ambitions portant sur une ultra-personnalisation en cancérologie. Il s'agit notamment de pouvoir, pour un patient atteint d'un cancer (ou susceptible d'en développer un), l'intégrer au sein d'un parcours personnalisé de soins basé sur ses données personnelles. Dans ce but, les initiatives pour récolter des données en grand nombre et hétérogènes se sont multipliées. Mais l'accès à des données ne constitue pas un accès à de l'information. Pour accéder à cette dernière, il est nécessaire de mettre en place des méthodes d'analyse mathématique et computationnelles. Les techniques basées sur de l'apprentissage statistique (*machine learning*) ont ouvert la voie à cette médecine personnalisée, avec des résultats prometteurs notamment en analyse d'images ou de données génomiques, par exemple pour détecter automatiquement des tumeurs ou encore l'identification de biomarqueurs [1, 2, 3]. Ces méthodes, très efficaces lorsque l'on a beaucoup de patients, le sont moins pour des maladies peu prévalentes. De plus, elles souffrent actuellement d'un manque d'explicabilité, même si on peut espérer des améliorations à venir au vu de la recherche active qui est faite en ce sens [4, 5]. Or, intégrer un patient dans un parcours de santé personnalisé, c'est également être en mesure de justifier le choix de tel soin plutôt qu'un autre.

Les modèles mathématiques mécanistes, construits à partir de règles physiques ou d'hypothèses biologiques, se trouvent alors être des outils adaptés lorsqu'il s'agit d'étudier des maladies rares ou d'étudier les mécanismes qui conduisent au développement ou à la rémission d'un cancer.
Ils pourraient également être perçus comme des alternatives aux modèles murins dans un contexte législatif européen peu favorable aux expérimentations animales. Ainsi, le parlement européen adoptait le 16 septembre 2021 une résolution "visant à accélérer le passage à une innovation sans recours aux animaux dans la recherche, les essais réglementaires et l'enseignement" [1]. Une des options possibles identifiée dans ce texte serait l'usage de "simulations intensives sophistiquées". Si les ressources informatiques actuelles ont en effet rendu possible les études *in silico*, celles-ci doivent néanmoins s'appuyer sur un travail de recherche considérable pour la construction de modèles adéquats, souvent sur mesure pour répondre à une problématique particulière, puis la calibration de ces modèles. Cette dernière étape, nécessaire pour ensuite pouvoir utiliser le modèle à des fins prédictives par exemple, pourra difficilement se passer de mesures expérimentales, souvent obtenues par l'intermédiaire d'expérimentations animales.

Dans ce mémoire de thèse, nous nous intéresserons aux néoplasmes myéloprolifératifs (NMP) de type BCR-ABL négatifs (on exclut donc l'étude des leucémies myéloïdes chroniques caractérisées par la présence du gène de fusion *BCR-ABL)* et à leur traitement à l'Interféron alpha (IFNα). Les NMP sont des hémopathies malignes faiblement prévalentes dans la population, justifiant le recours à des modèles mathématiques pour leur étude.
Nous présentons le contexte biomédical à la section 2. Notre étude de ces hémopathies malignes se fera à partir de modèles mathématiques. La littérature scientifique présentant de tels modèles, pour l'hématopoïèse saine ou pathologique, est abondante ; nous en donnerons un aperçu à la section 3. L'étude de tels modèles, de leur construction à l'estimation de leurs paramètres par exemple, passe par de nombreuses étapes ; nous en illustrerons les enjeux méthodologiques à la section 4. Les parties 3 et 4 font l'objet d'une revue "Modélisation mathématique de l'hématopoïèse et des hémopathies : développement, dynamique et traitement" publiée au journal Hématologie. Enfin, nous terminerons ce chapitre introductif en présentant le plan de ce mémoire de thèse.

1. Résolution 2021/2784(RSP)

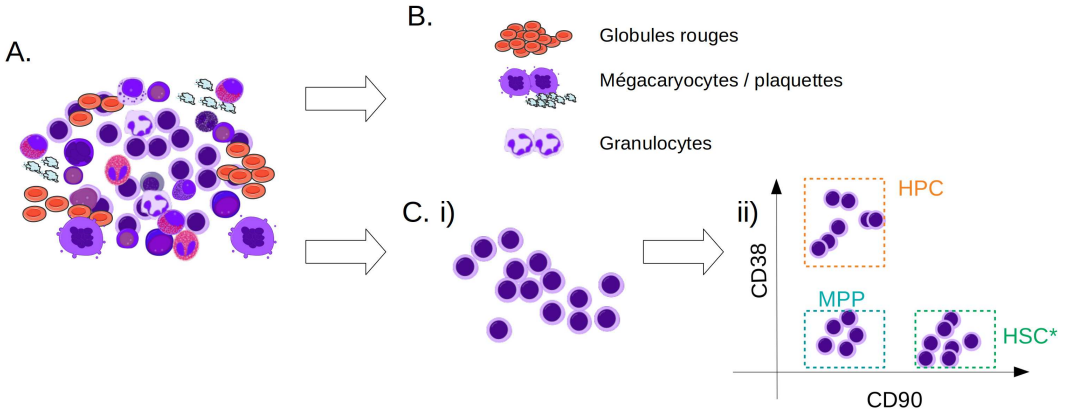


FIGURE 1 – Les cellules hématopoïétiques (A) consistent en un ensemble de cellules comprenant des cellules matures fonctionnelles (B) - incluant globules rouges, plaquettes et neutrophiles - et des cellules immatures : précurseurs, progéniteurs et cellules souches (C). Les cellules souches et progénitrices ne sont pas distinguables les unes des autres par leurs propriétés morphologiques (C.i). L'utilisation en cytométrie de certains anticorps - comme les anticorps anti-CD38 et anti-CD90 - couplés à des fluorochromes permet alors de distinguer différents types de cellules en fonction de certains marqueurs de surface (C.ii). En particulier, l'expression ou non des marqueurs CD90 et CD38 (pour un ensemble de cellules $CD34^+$) permet de distinguer les HPC (Hematopoietic Progenitor Cell) des MPP (Multipotent Progenitor) et des HSC* (Hematopoietic Stem Cell). Notons que, parmi les cellules $CD90^+CD38^-$, seule une faible proportion d'entre elles peuvent être considérées comme souches, dans le sens où elles seraient en mesure de permettre la création de toutes les cellules du sang sur le long terme. Pour faire la différence entre la définition conceptuelle de celle phénotypique, nous ajouterons le symbole * au terme HSC pour indiquer qu'il s'agit d'une caractérisation basée sur des marqueurs de surface.

2 Contexte biomédical

2.1 Cellules hématopoïétiques : caractérisation et structure de l'hématopoïèse

Cette histoire commence avec un grand nombre de cellules (de l'ordre de 10^{13} cellules matures chez l'homme [6]) : les cellules du sang appelées cellules hématopoïétiques (Fig. 1-A).
Pour un lecteur mathématicien non familier avec la biologie - ou plus particulièrement l'hématologie - nous rappellerons ici quelques fondamentaux. Dans l'ensemble des cellules hématopoïétiques, certaines assurent des fonctions vitales au sein de l'organisme : ce sont les cellules dites matures. Pour en citer quelques-unes (Fig. 1-B), il y a :

- Les globules rouges (ou érythrocytes), d'une durée de vie d'environ 120 jours, qui assurent le rôle de transport de l'oxygène,
- Les granulocytes neutrophiles, d'une durée de vie d'environ 24 heures, qui jouent un rôle dans le système immunitaire par la phagocytose,
- Les plaquettes - issues des mégacaryocytes, cellules polyploïdes de taille assez importante (50-100 μm) - qui participent à la coagulation du sang lorsque l'on se blesse.

Les cellules précédentes appartiennent à la lignée myéloïde, lignée qui nous intéressera dans ce manuscrit. La seconde lignée hématopoïétique est celle dite lymphoïde, qui regroupe les lymphocytes T et B assurant une fonction immunitaire.
Ces cellules matures ont très tôt pu être identifiées et caractérisées par observation directe au microscope. Elles sont produites par des cellules dites immatures - progéniteurs et précurseurs - elles-mêmes produites par les cellules souches hématopoïétiques (HSC - Hematopoietic Stem Cell), par un processus de prolifération et différenciation cellulaire (Fig. 2).
Les précurseurs hématopoïétiques sont des cellules déjà engagées vers une voie de différenciation,

mais qui ne sont pas encore matures. Les cellules souches et progénitrices (Fig. 1-C.i) sont des cellules plus immatures, avec un potentiel de différenciation plus important. On distingue les cellules dites multipotentes (HSC et MPP), capables de contribuer à toutes les lignées hématopoïétiques, de celles dites pluripotentes capables de produire toutes les cellules de l'organisme (tandis que les cellules souches embryonnaires sont dites totipotentes, c'est-à-dire qu'elles sont capables de produire un embryon entier). Le potentiel de différenciation des cellules est une des caractéristiques permettant de classer les cellules en différents types. La seconde caractéristique importante, notamment pour la définition de la cellule souche, est la capacité d'auto-renouvellement. L'auto-renouvellement est la capacité pour une cellule de donner naissance à au moins une cellule fille qui lui est fonctionnellement identique. Sans auto-renouvellement, les cellules filles issues des cellules souches seraient progressivement de plus en plus différenciées, et le stock de cellules s'épuiserait. Ces deux propriétés fondamentales - la multipotence et l'auto-renouvellement - permettent de définir la cellules souche hématopoïétique. Ces propriétés ont été très tôt exploitées en clinique pour le traitement de plusieurs hémopathies, notamment pour la transplantation après chimiothérapie ou irradiation [7, 8]. D'ailleurs, le concept de cellule souche hématopoïétique a justement été découvert par une expérience de transplantation sur souris, où il avait été montré que transplanter des cellules de moelle osseuse de souris adultes saines pouvait permettre à des souris létalement irradiées de survivre [9]. Parmi les cellules souches, on distingue parfois celles dites LT (Long-Term) de celles dites ST (Short-Term), les premières étant capables de maintenir une hématopoïèse sur le long terme (essentiellement, pendant toute une vie), quand les dernières ne seraient capable de le faire que sur une durée plus limitée.
La classification des cellules suivant leur potentiel de différenciation et leur capacité d'auto-renouvellement repose sur une définition conceptuelle des cellules, difficilement utilisable en pratique lors d'expérimentations. En effet, pouvoir dire qu'une cellule est une cellule souche nécessite d'observer après un temps suffisamment long si la cellule a abouti *in vivo* à des colonies de cellules matures de tous types.
Une caractérisation phénotypique des cellules hématopoïétiques a vu le jour avec le développement de techniques de cytométrie et la découverte de différents marqueurs de surface. Cette caractérisation repose sur l'existence, à la surface des cellules, de différentes protéines membranaires, permettant de discriminer les cellules entre elles. En cytométrie de flux, on utilise des anticorps spécifiques de certaines protéines de surface d'intérêt, qui vont alors se lier à ces dernières (voir par exemple [10]). En couplant les anticorps à des fluorochromes qui ont un certain spectre d'émission, on peut alors mesurer l'intensité de fluorescence pour chaque cellule, qui sera alors associée à la quantité de protéines à sa surface. En choisissant, lors d'une expérience, les bons marqueurs de surface permettant de classer les cellules en différents types, on peut alors caractériser les différentes cellules, voire les trier pour ensuite mener, par exemple, des expériences uniquement pour des cellules d'un type donné. Les principaux marqueurs qui nous intéresseront dans ce manuscrit sont les marqueurs CD34, CD38 et CD90. Le premier permet d'identifier les cellules immatures (qui seront dites $CD34^+$, c'est-à-dire pour lesquelles l'intensité de fluorescence associée au marqueur CD34 dépassera un certain seuil), les deux derniers permettent de distinguer les HPC des MPP et des HSC* (Fig. 1-C.ii). La notation avec le symbole * permettra de faire la distinction entre la définition immuno-phénotypique de la cellule souche et la définition conceptuelle. Ainsi, par HSC*, on définira les cellules qui sont $CD34^+CD38^-CD90^+$. Par HSC, on définira les cellules qui sont souches au sens conceptuel. On peut considérer qu'environ 10% des HSC* sont vraiment des cellules souches (i.e. capable de donner tous types de cellules hématopoïétiques à long-terme), et on définira parfois cette population cellulaire HSC* comme étant une population de cellules progénitrices enrichies en cellules souches.
L'utilisation des techniques de cytométrie a contribué à la structure standard de l'hématopoïèse, sous la forme d'arbre (Fig. 2-B). Avec la multiplication des marqueurs permettant de caractériser plus finement les cellules, cette structure s'est progressivement complexifiée, de nouveaux types cellulaires étant régulièrement identifiés. L'avènement des techniques single-cell (observations à l'échelon uni-cellulaire) et la possibilité de mesurer des centaines de *features*, en single-cell RNA-sequencing (scRNA-seq) par exemple, a récemment fait émerger de nouvelles structures,

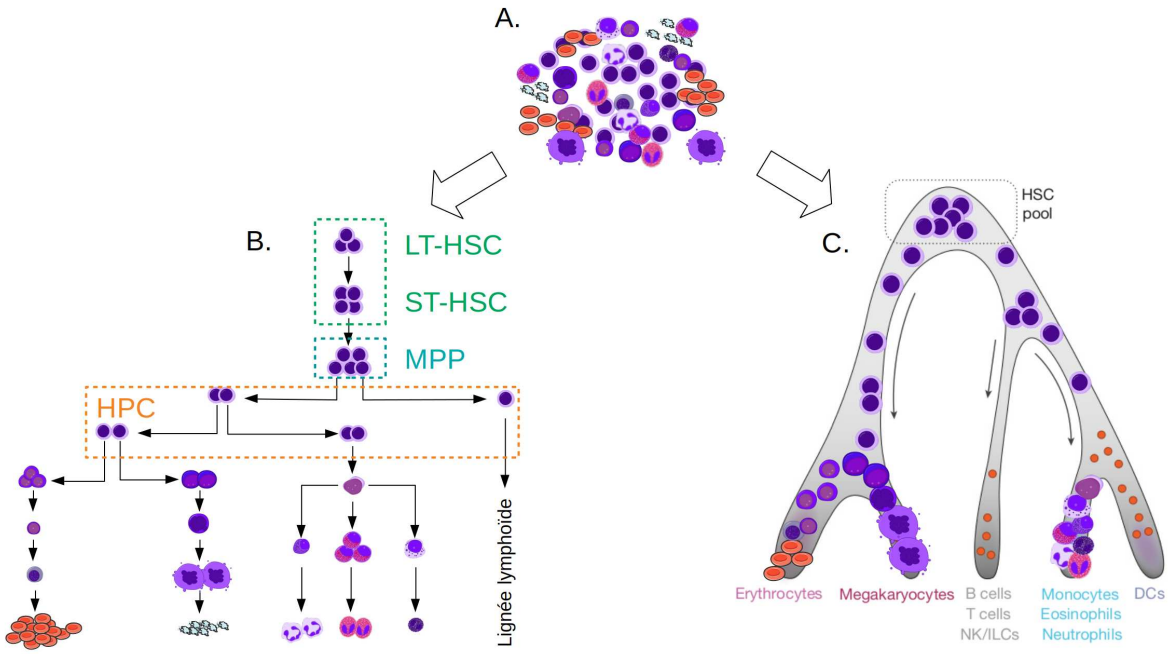


FIGURE 2 – L'hématopoïèse est un processus impliquant un grand nombre de cellules (A) qui peut se représenter par une structure hiérarchique. Une représentation classique de l'hématopoïèse est la structure en arbre (B). Dans celle-ci, les cellules sont regroupées en différents types cellulaires. Au sommet de la structure branchante (l'arbre), on retrouve les cellules souches hématopoïétiques (LT-HSC et ST-HSC - Long-Term et Short Term Hematopoietic Stem Cells), capables de s'auto-renouveler et de produire tous les types cellulaires. Les MPP (Multipotent Progenitors) sont des cellules progénitrices immatures issues des HSC, qui contribueront ensuite à la formation des HPC (Hematopoietic Progenitor Cells). Par une cascade de différenciation, on retrouve alors différents précurseurs hématopoïétiques qui aboutiront aux cellules matures, fonctionnelles. D'autres représentations possibles mettent l'accent sur le continuum d'états formé par les cellules hématopoïétiques (C - schéma adapté de Laurenti et al. [12]).

reposant sur l'hypothèse d'un continuum d'états cellulaires plutôt qu'un ensemble de cellules de types cellulaires distincts [11, 12, 13] (Fig. 2-C). Nous aborderons ce point plus en détail au chapitre 2.

2.2 Prolifération et différenciation cellulaire

Nous avons présenté au paragraphe précédent les principaux types de cellules qui composent l'ensemble des cellules hématopoïétiques. L'hématopoïèse n'est pas un système statique constitué de milliards de cellules : c'est un processus dynamique couplant prolifération et différenciation. Par prolifération, nous entendons la division cellulaire, c'est-à-dire la création de deux cellules filles par une cellule mère. C'est par ce processus qu'on peut obtenir des milliers de milliards de cellules [14] à partir de centaines de milliers de cellules souches [15]. Par différenciation, nous entendons le processus qui permet, à partir de cellules immatures multipotentes, d'aboutir progressivement à des cellules matures assurant un rôle clé dans l'organisme. La différenciation correspond ainsi aux changements cellulaires, transcriptomiques et métaboliques de la cellule, suite à une division ou au cours de sa vie, qui la font s'engager vers une lignée particulière, exprimer ou arrêter d'exprimer certains marqueurs de surface, synthétiser certaines molécules, etc.

Concernant la prolifération, nous nous intéresserons particulièrement, dans ce manuscrit, à celle des cellules souches. Nous l'avons mentionné plus haut, ces dernières sont notamment caractérisées par leur capacité d'auto-renouvellement. Il y a auto-renouvellement lorsqu'une cellule se

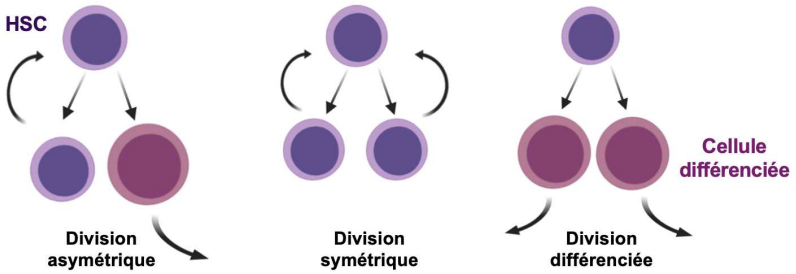


FIGURE 3 – Une cellule souche (HSC) peut se diviser suivant une division asymétrique, symétrique ou différenciée pour produire respectivement une HSC et un progéniteur (i.e. une cellule différenciée), deux HSC ou deux progéniteurs.

divise suivant une division que nous appellerons symétrique, i.e., lorsqu'une cellule donne naissance à deux cellules filles "identiques" à la cellule mère, ou suivant une division asymétrique, i.e., lorsqu'une cellule mère donne naissance à une cellule fille qui lui est identique et une autre qui s'engage vers la différenciation. Le troisième type de division, permettant d'assurer en moyenne un nombre constant de cellules souches dans l'organisme, est la division que nous appellerons différenciée par laquelle une cellule mère donne naissance à deux cellules filles engagées vers une voie de différenciation (voir Fig. 3). Sans l'existence de ce troisième type de division, le nombre de cellules souches dans l'organisme ne ferait qu'augmenter avec le temps (bien que des phénomènes d'apoptose pourraient également permettre la régulation de la taille du compartiment souche). L'existence de ce type de division pour des cellules souches est néanmoins sujet à controverse ; une cellule souche qui se divise en deux cellules progénitrices pourrait-elle toujours correspondre à la définition conceptuelle de la cellule souche, alors qu'elle ne permettrait par conséquent plus de maintenir une hématopoïèse de long-terme ? Nous considérerons dans ce travail que la cellule est souche par le potentiel qu'elle a de s'auto-renouveler indéfiniment. C'est-à-dire que nous supposerons qu'une cellule souche se divisant de façon différenciée ne le fait pas intrinsèquement (parce qu'elle serait déjà un peu plus différenciée qu'une autre cellule) mais que cette division résulte soit d'effets stochastiques [16, 17], soit d'une stimulation extérieure à la cellule (par l'action de cytokines par exemple, voir plus bas).
Lee-Six et al. ont mis en évidence que le stock de HSC connaîtrait une expansion dans les premières années de la vie, jusqu'à atteindre un plateau puis décroître vers des âges plus avancés [15]. L'hypothèse d'un nombre constant de cellules souches serait ainsi valable sur une certaine période dans la vie d'un individu sain (sans pathologie particulière). Ce nombre pourrait être maintenu constant (on parlera éventuellement d'homéostasie) grâce à différents mécanismes de régulation. De façon générale, l'hématopoïèse est un processus finement régulé qui doit être résilient à des perturbations, telles qu'une perte massive de globules rouges après hémorragie par exemple. Cette régulation se fait par l'intermédiaire de nombreuses protéines : les cytokines. Les cytokines constituent une famille de ligands extracellulaires, synthétisés par des cellules ou tissus, qui vont agir sur d'autres cellules, en se liant à des récepteurs de cytokines, activant alors certaines voies de signalisation [18]. Il s'agit donc de molécules centrales pour la signalisation cellulaire qui jouent également un rôle essentiel dans la régulation de l'hématopoïèse (Fig. 4). Parmi les différentes cytokines, on peut distinguer :

- L'érythropoïétine (EPO), synthétisée par les reins, sensible à la pression en oxygène du sang, qui favorise la production de globules rouges,
- La thrombopoïétine (TPO), synthétisée par le foie, a son récepteur associé - appelé MPL - présent sur les plaquettes et sur les mégacaryocytes,
- Les *colony-stimulating factors* (CSF) parmi lesquels on peut citer :
  — Le G-CSF qui favorise la différenciation vers les neutrophiles (granulocytes),
  — Le GM-CSF qui permet l'orientation vers la production de progéniteurs granulocytaires / monocytaires,
  — Le M-CSF qui permet la différenciation vers les monocytes.

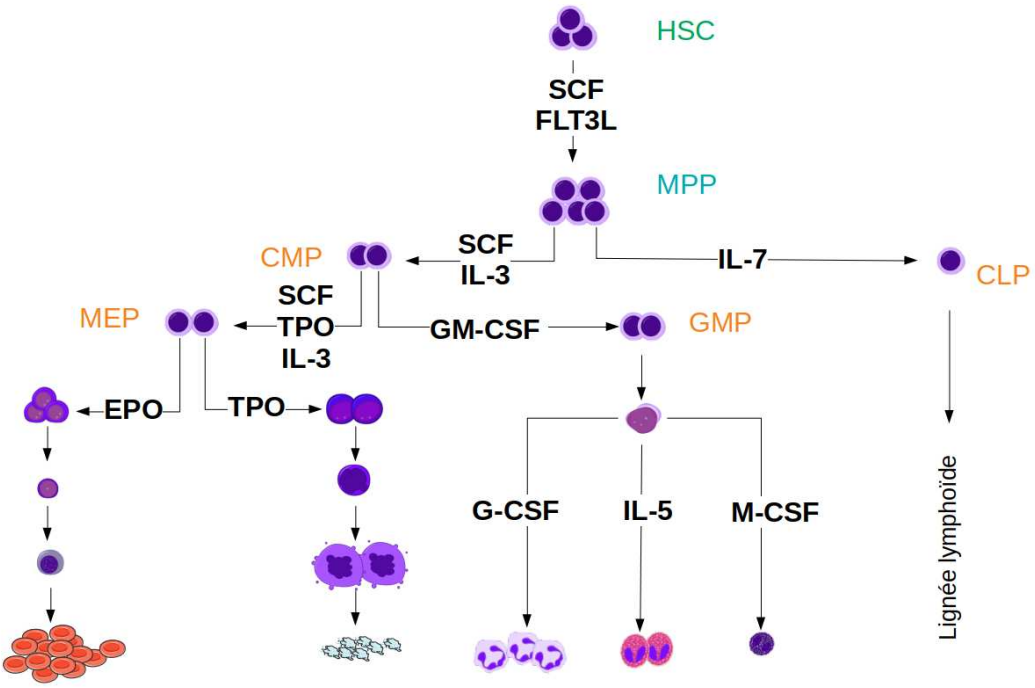


FIGURE 4 – Schéma du rôle des cytokines dans l'hématopoïèse (adapté de [18] et [19]). Les cytokines interviennent à différents niveaux, favorisant la prolifération / différenciation vers telle ou telle lignée cellulaire. Parmi les HPC (en orange), nous faisons la distinction entre les CLP (progéniteurs communs lymphoïdes), CMP (progéniteurs communs myéloïdes), MEP (progéniteurs communs mégacaryocytaires-érythroïdes) et GMP (progéniteurs granulocytaires-monocytaires). SCF (Stem cell factor) est le ligand qui se fixe au récepteur c-Kit. Il favoriserait notamment la prolifération et différenciation des progéniteurs hématopoïétiques et empêcherait l'apoptose des HSC [20], ce qui en fait une cytokine couramment utilisée pour la culture cellulaire *in vitro* [21].

- Les interférons (IFN),
- Les interleukines (IL).

Lorsque les cytokines se lient au récepteur qui leur est spécifique, cela induit une cascade de réactions conduisant à l'activation de facteurs de transcriptions, induisant alors l'expression de gènes pouvant être impliqués dans la survie cellulaire, la prolifération ou la différenciation.
Dans ce chapitre, nous nous intéresserons particulièrement aux récepteurs à la TPO (MPL), à l'EPO (EPOR) et au G-CSF (G-CSFR) qui sont ceux qui "gouvernent la différenciation" ([22]) de la lignée mégacaryocytaire, érythrocytaire et granuleuse, respectivement (voir Fig. 5).
Ces récepteurs (appartenant à la famille des récepteurs homodimériques de type I) sont dépourvus d'activité enzymatique (i.e. d'activité tyrosine kinase, activité qui consiste à transférer un groupement phosphate vers une protéine cible), c'est-à-dire qu'en eux-même, ils ne sont pas capables d'induire la cascade de réactions évoquée précédemment. Ils ne pourront le faire qu'associés à des Janus kinases (JAK), qui sont des protéines cytoplasmiques à activité tyrosine kinase non-récepteur [23]. JAK2 est la protéine kinase qui s'associe principalement aux trois récepteurs précédents. Une troisième famille de protéines intervient dans la composition de cette voie de signalisation : les protéines STAT (Signal Transducers and Activators of Transcription), qui permettront - après avoir été activées et avoir migré vers le noyau - la transcription de gènes cibles. Les récepteurs, les JAK et les STAT - suite à leur phosphorylation par les JAK - forment ainsi les trois constituants clés de l'axe de signalisation JAK/STAT [24], schématisé sur la figure 6. L'acquisition d'une mutation dans un gène codant pour une protéine impliquée dans cette voie de signalisation peut alors entraîner une production cellulaire incontrôlée et le développement de cancers, tels que les néoplasmes myéloprolifératifs qui seront présentés à la section suivante. Même si les STAT sont les principales protéines activées, la fixation de la cytokine sur son récepteur

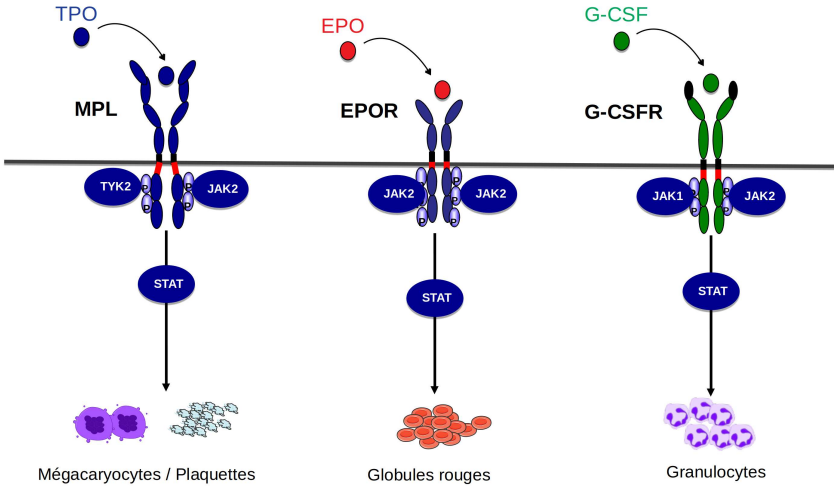


FIGURE 5 – Schéma des récepteurs à la TPO (MPL), à l'EPO (EPOR) et au G-CSF (G-CSFR) (Schéma par Isabelle Plo).

entraîne également l'activation des protéines de la voie MAPK et de la voie PI3K/AKT.

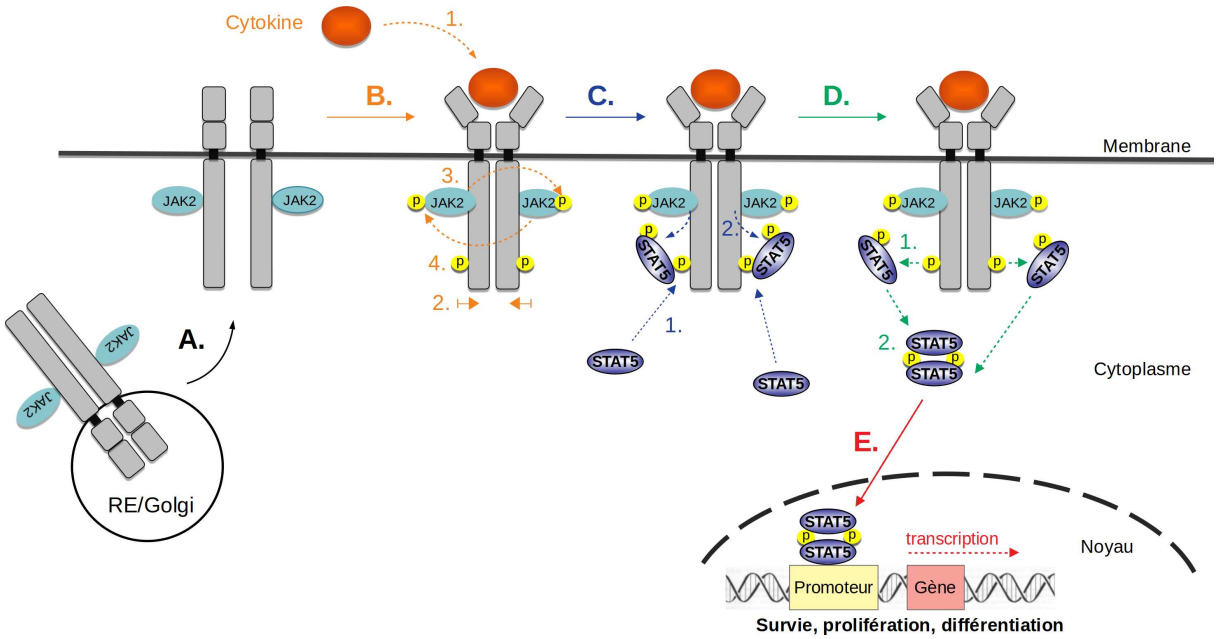


FIGURE 6 – Schéma de la voie de signalisation JAK2-STAT5 [22] (adapté d'un schéma de la thèse de Florence Pasquier [25]).
A. JAK2 se fixe à un récepteur (ici, par exemple, l'EPOR), jouant un rôle de chaperon au niveau du réticulum endoplasmique (RE), favorisant ensuite la montée du récepteur à la surface de la cellule.
B. Une cytokine spécifique du récepteur (ici, par exemple, l'EPO) va s'y fixer (1), entraînant un changement de conformation du récepteur, amenant les JAK2 associés au récepteur à proximité l'un de l'autre (2). les JAK2 vont se phosphoryler (transphosphorylation) l'un l'autre (3) augmentant leur activité tyrosine kinase [26]. Ils vont ensuite phosphoryler un résidu tyrosine au niveau de la partie intra-cytoplasmique du récepteur (4).
C. La phosphorylation du récepteur va servir de point d'ancrage pour d'autres molécules, par exemple ici STAT5 (1). Les molécules STAT5 seront à leur tour phosphorylées par JAK2 (2).
D. Une fois phosphorylées, les molécules STAT5 vont se dissocier du récepteur [27] (1) puis s'homodimériser (2).
E. L'homodimère ainsi formé migre vers le noyau, agissant en tant que facteur de transcription sur certains gènes cibles.

2.3 Néoplasmes Myéloprolifératifs

Les néoplasmes myéloprolifératifs BCR-ABL-négatifs sont des hémopathies malignes clonales dans lesquelles un ou plusieurs types de cellules sanguines matures sont produits en excès.
On classe les NMP en trois maladies suivant le type de cellules matures produites en excès [28] (Fig. 7) :

- La thrombocytémie essentielle (TE) caractérisée par une surproduction de plaquettes,
- La poylglobulie de Vaquez (PV) caractérisée par une surproduction de globules rouges,
- La myélofibrose primaire (MFP) - la forme la plus aiguë des NMP parfois caractérisée de stade pré-leucémique - caractérisée par une surproduction de mégacaryocytes et granuleux et associée à la présence de fibres de collagènes invalidantes dans la moelle osseuse.

Un patient peut évoluer d'une maladie à l'autre, sur un temps de l'ordre de la dizaine d'année ; plutôt que trois maladies distinctes, les NMP formeraient ainsi un continuum [29, 30] avec une certaine hétérogénéité d'évolution en grande partie due aux mutations associées.
Ces maladies sont dues à des anomalies génétiques affectant la voie de signalisation JAK/STAT décrite précédemment. En particulier, l'apparition des NMP est principalement associée à des mutations somatiques avec gain de fonction dans les gènes codant pour JAK2 ($JAK2^{V617F}$), la calréticuline ($CALR^{m}$; type 1 $CALR^{del52}$ et type 2 $CALR^{ins5}$), ou le récepteur de la thrombopoïétine (MPL^{m}) qui sont acquises dans les cellules souches hématopoïétiques [31, 32, 33, 34]. Plusieurs autres mutations somatiques impliquées dans la régulation épigénétique, l'épissage et les facteurs de transcription peuvent modifier le phénotype de la maladie [35].

Les NMP sont plutôt détectés tardivement, à des âges médians de l'ordre de 65 ans pour la TE et la PV et 70 ans pour la MFP [36]. Ce sont des hémopathies assez rares, avec une incidence qui augmente avec l'âge [37, 38]. Parmi les NMP, la TE serait la maladie la plus courante, avec une incidence annuelle de 1.03 pour 100,000, suivie de la PV (incidence annuelle de 0.84/100,000) et la MFP (incidence annuelle de 0.47/100,000), d'après les résultats issus d'une méta-analyse [38]. Une étude réalisée à l'échelle de l'Europe estime une prévalence (pour 100,000 individus) allant de 4.96 à 30.0 pour la PV, 4.0 à 24.0 pour la TE et 0.51 à 2.7 pour la MFP [39]. Néanmoins, les deux principales mutations motrices des NMP, à savoir $JAK2^{V617F}$ et $CALR^{m}$, seraient présentes dans la population générale à des fréquences bien plus élevées, avec des prévalences estimées (sur une cohorte danoise) à 3.1% et 0.16% respectivement [40], suggérant un temps de latence important entre l'acquisition de la mutation motrice du NMP et l'apparition de la maladie. Des études récentes ont ainsi montré que l'acquisition des mutations au niveau des HSC pourrait avoir lieu plusieurs décennies avant l'apparition des symptômes, voire pendant la vie fœtale [41, 42, 43, 44, 45]. Il en résulterait un développement clonal lent, sur plusieurs dizaines d'années, potentiellement accéléré par l'acquisition de mutations associées ou d'autres anomalies

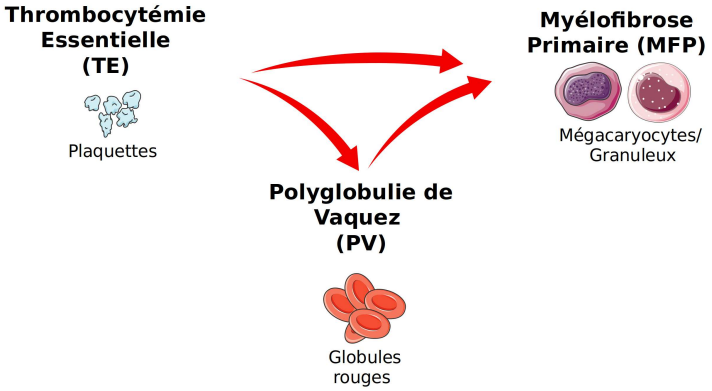


FIGURE 7 – Classification des NMP en trois maladies (schéma par Amandine Tisserand). La TE peut évoluer en PV ou MFP et la PV en MFP.

génétiques telles que la recombinaison mitotique par laquelle une cellule mutée hétérozygote devient homozygote [42, 43].
D'un point de vue biologique et étude de l'oncogenèse, les NMP peuvent être considérés comme des cancers du sang "modèles" de par leur développement clonal sur un temps très long.

La première mutation motrice des NMP qui a été découverte est la mutation $JAK2^{V617F}$ [31] (mentionnons également [46, 47]). Elle est retrouvée dans plus de 95% des cas de PV et environ 55% des TE et MFP [22]. La mutation $JAK2^{V617F}$ va favoriser l'activité kinase des protéines JAK2 (Fig. 6.B-3), et ce même en l'absence d'une cytokine qui se fixerait au récepteur, c'est-à-dire que la mutation entraîne une activation constitutive de JAK2 [48].

La seconde mutation motrice qui a été découverte affecte cette fois directement un récepteur de cytokines, à savoir le récepteur à la TPO : MPL. MPL^{W515L} a été la première mutation du gène *MPL* qui a été découverte [32], suivie de plusieurs autres [49]. Cette mutation affectant le récepteur à la TPO qui est impliqué dans la lignée des mégacaryocytes et plaquettes, on la retrouve dans certains cas de TE et MFP, mais néanmoins seulement pour une faible proportion d'entre eux (environ 3-5% [22]).

La dernière mutation motrice des NMP a été découverte plus récemment, il y a une dizaine d'année environ [33, 34]. Il s'agit d'une mutation du gène codant pour la protéine calréticuline (CALR) qui, *a priori*, n'était pas impliquée dans la voie de signalisation JAK/STAT. En fait, de nombreuses mutations du gène *CALR* ont été détectées, qu'on regroupe généralement en deux types : les mutations de type 1 ($CALR^{del52}$) et celles de type 2 ($CALR^{ins5}$). Les mutations consistent en des insertions ou délétions dans l'exon 9 du gène *CALR* qui vont entraîner un décalage du cadre de lecture ce qui entraîne une nouvelle séquence en C-terminal de la protéine et la délétion du signal de rétention dans le réticulum endoplasmique (KDEL) [33, 34]. La CALR est une protéine chaperon du réticulum endoplasmique ; lorsqu'elle est mutée, elle va se fixer sur le récepteur MPL et déclencher l'activation de la voie de signalisation JAK2/STAT [50, 51, 52]. On retrouve ainsi ces mutations dans les TE et MFP (dans 25-30% des cas [22]).

Pour résumer, les néoplasmes myéloprolifératifs (BCR-ABL négatifs) consistent en trois maladies, affectant chacune une lignée cellulaire en particulier, qui sont dues dans la très grande majorité des cas à l'acquisition de l'une des trois mutations dites motrices[2] (suffisantes pour l'apparition de la maladie) soit au gène *JAK2*, *MPL*, ou *CALR*. La surproduction de cellules hématopoïétiques matures, qui caractérise les NMP, fait peser des risques pour les patients, tels que des risques de thromboses ou d'événements cardiovasculaires. De plus, en l'absence de prise en charge clinique, la maladie peut évoluer vers des formes aiguës (dont font partie les MFP) puis vers des stades leucémiques, engageant alors fortement le pronostic vital. Ainsi, une fois la maladie (ou la mutation) détectée, il est crucial de suivre et prendre en charge les patients. Nous abordons à la section suivante les solutions thérapeutiques envisagées pour le traitement des patients atteints de NMP.

2.4 Traitement

Tout d'abord, la prise en charge médicale peut viser à limiter les risques associés aux NMP, tels que la thrombose ou les hémorragies. Parmi les médicaments prescrits, on retrouve ainsi l'aspirine (à faible dose) qui peut aider à limiter la formation de caillots, notamment pour les patients atteints de TE qui vont produire trop de plaquettes. On peut également prescrire des médicaments cytoréducteurs, qui vont faire baisser le nombre de plaquettes, tels que l'hydroxyurée (dont fait partie l'Hydrea) [53]. Lorsque le nombre de globules rouges en circulation est trop important et doit être réduit rapidement, on peut également pratiquer une phlébotomie , c'est-à-dire l'incision

2. Dans quelques rares cas de NMP, dits triple-négatifs, on ne retrouve pas la mutation à l'un de ces trois gènes.

d'une veine pour pratiquer une saignée.

Plusieurs essais cliniques ont également visé à montrer l'efficacité d'inhibiteurs de JAK2 [54], parmi lesquels on peut citer le ruxolitinib, dont la mise sur le marché a été approuvée en 2011 [55, 56] ou le fedratinib (Inrebic) autorisé plus récemment par la FDA (Food and Drug Administration) [57, 58]. Les inhibiteurs de JAK permettent généralement une réduction des symptômes, mais ne ciblent pas les cellules souches responsables de la maladie, et l'arrêt des traitements conduit généralement à une réapparition des symptômes.

Parmi les stratégies thérapeutiques développées ces dernières années pour le traitement des NMP, l'interféron alpha (IFNα) semblerait être le plus prometteur. Il a été montré que l'IFNα, en particulier l'IFNα pégylé (notamment Peg-IFNα2a), induit des réponses hématologiques dans les TE, PV et certaines MFP précoces [59, 60, 61, 62, 63], comme l'a récemment démontré un essai clinique de phase 2 réalisé chez des patients réfractaires/intolérants à l'hydroxyurée atteints de NMP $JAK2^{V617F}$ et $CALR^{m}$ [64]. Des essais cliniques randomisés de phase 3 ont démontré qu'un autre IFN pégylé, Ropeg-IFNα2b, augmentait le taux de réponse hématologique par rapport à l'hydroxyurée [65] ou à la phlébotomie [66] chez des patients PV (voir aussi [67]). Contrairement aux thérapies cytoréductrices ou aux inhibiteurs de JAK, l'IFNα serait capable de diminuer la charge allélique $JAK2^{V617F}$ dans les cellules sanguines chez environ 60% des patients, et d'induire des réponses moléculaires complètes dans 20% des cas même après arrêt du traitement [59, 64, 65], c'est-à-dire de cibler directement les cellules souches $JAK2^{V617F}$ responsables de la maladie.
Son effet sur les patients atteints d'un NMP avec comme mutation motrice $CALR^{m}$, serait moins important que dans le cas $JAK2^{V617F}$ [64, 68, 69, 70].
De plus, bien que très prometteur, l'IFNα n'est pas toléré par tous les patients, et peut induire certains effets secondaires, notamment des dépressions majeures [71, 72].

Le mécanisme d'action de l'IFNα n'est pas encore totalement élucidé, notamment la façon dont il cible les cellules souches.
Mieux comprendre comment il agit permettrait de mieux prendre en charge les patients, éventuellement en fonction de la mutation motrice du NMP et d'autres déterminants génétiques pouvant impacter la réponse au traitement. L'étude de ce traitement repose notamment sur l'utilisation de différents modèles murins, qui fournissent des pistes essentielles quant à l'action potentielle de l'IFNα sur les cellules souches (voir par exemple [73, 74]). Dans ce mémoire de thèse, nous adopterons une autre approche et étudierons des modèles mathématiques pour essayer de comprendre l'effet à long-terme de ce traitement, stratifier les patients suivant différents déterminants et enfin estimer quelles doses pourraient être les plus adaptées en vue d'obtenir une réponse moléculaire. Les modèles mathématiques constituent ainsi, non pas une alternative aux modèles murins, mais un outil complémentaire pour étudier l'effet de traitements. Dans la section suivante, nous présenterons un panorama de différents modèles et formalismes qui ont été utilisés pour l'étude de l'hématopoïèse et des hémopathies.

3 Modèles mathématiques de l'hématopoïèse et des hémopathies

3.1 Tour d'horizon des types de modèles

Les mathématiciens s'attellent à modéliser l'hématopoïèse depuis plus de cinquante ans. Dans sa revue, Pujo-Menjouet retrace une histoire de ces modèles, les différentes formes qu'ils ont pu prendre et les différentes tendances [75]. Les problématiques portant sur l'étude des oscillations périodiques survenant dans certaines maladies du sang ont rapidement intéressé les mathématiciens, dont les modèles permettaient de proposer des pistes quant aux causes potentielles des oscillations dans les quantités de cellules sanguines matures au cours du temps. En 1970, King-Smith et Morley suggéraient ainsi, à l'aide de simulations, que les oscillations survenant dans le cas des neutropénies cycliques pouvaient s'expliquer par la présence de mécanismes de régulation [76]. Plus tard, Mackey et al. formalisaient un modèle permettant de décrire les phénomènes d'oscillations dans certaines pathologies hématologiques, notamment les formes périodiques de la leucémie myéloïde chronique (LMC) [77]. Leur modèle repose sur une équation différentielle avec retard, c'est-à-dire une équation reliant les variations au cours du temps t d'une certaine variable – ici la quantité de globules blancs en circulation – à ses valeurs prises à un instant antérieur $t - \tau$ (et pas seulement à l'instant t, cas qui se modéliserait alors par une équation différentielle ordinaire). De nombreux auteurs ont depuis étudié l'hématopoïèse par des modèles d'équations différentielles avec retard [78, 79]. L'introduction d'un retard dans les équations peut par exemple permettre de prendre en compte le délai nécessaire à la production de cellules matures à partir de cellules souches ou progénitrices [80] ou encore le temps nécessaire à la cellule pour effectuer son cycle [81]. Notons que l'introduction de ce retard n'est pas toujours un choix de modélisation, mais que les équations différentielles à retard peuvent résulter de l'intégration, suivant des variables de structure, d'équations aux dérivées partielles. Les équations aux dérivées partielles, liant certaines variables à leurs variations dans le temps mais aussi dans l'espace, permettent de prendre en compte des effets spatiaux, par exemple pour modéliser l'organisation d'îlots érythroblastiques [82]. Les modèles compartimentaux, dans lesquels certaines populations de cellules sont assignées à des compartiments (considérées alors homogènes) et où le passage d'un compartiment à un autre permet par exemple de modéliser les phénomènes de maturation / différenciation, sont largement utilisés pour modéliser l'hématopoïèse. Leur usage est particulièrement adapté lorsque l'on souhaite analyser des données expérimentales, les observations pouvant directement être associées à certains des compartiments permettant alors la calibration du modèle. Ils sont généralement formalisés par des équations différentielles ordinaires (ODE), avec parfois des non-linéarités lorsque l'on modélise des mécanismes de régulation, comme dans le cas de Marciniak-Czochra et al. [83]. Enfin, les mathématiciens se sont intéressés à la modélisation de la dynamique de l'ensemble des lignées hématopoïétiques, étudiées séparément (par exemple pour la lignée mégacaryocytaire [84], pour la lignée érythrocytaire [85], pour la lignée lymphocytaire [86]) ou ensemble [87].

Les différents formalismes cités jusqu'à présent (équations différentielles ordinaires, équations aux dérivées partielles, équations différentielles avec retard) sont déterministes. Étant données des conditions initiales ($t = 0$) et des valeurs fixées (ou estimées) pour les différents paramètres, la sortie du modèle (par exemple la prédiction d'une certaine quantité de cellules à un instant t) sera déterminée de façon unique, sans ambiguïté (Fig. 8-C). À l'inverse, les modèles peuvent être stochastiques, c'est-à-dire introduire de l'aléatoire dans la description des dynamiques (Fig. 8-B). Alors que les modèles déterministes sont généralement valables lorsque l'on étudie la dynamique de populations cellulaires de grandes tailles, ils ne le sont plus lorsqu'il s'agit de s'intéresser à des comportements à l'échelle de la cellule, comme par exemple lorsque l'on souhaite modéliser le développement clonal d'un cancer à partir d'une unique cellule mutée (voir [88] pour une comparaison entre les approches déterministes et stochastiques dans le cas de modèles hématopoïétiques). Parmi les familles de processus stochastiques, mentionnons les processus Markoviens que l'on retrouve couramment dans la littérature. Il s'agit de processus dits « sans-mémoires », c'est-à-dire tels que leur état futur ne dépend que de leur état actuel (et non du passé). Ils ont été largement étudiés d'un point de vue théorique et sont relativement simples à simuler, ce qui

justifie leur usage dans de nombreux modèles. Citons par exemple Catlin et al. qui proposent la calibration d'un modèle d'hématopoïèse à deux compartiments, reposant sur un processus de Markov caché [89]. Le terme « caché » signifie ici qu'il n'est pas possible d'observer la dynamique dans le compartiment des cellules souches ; Catlin et al. vont alors inférer ce qui s'y passe à partir d'observations au niveau des progéniteurs. Notons que la définition des processus Markoviens (sans-mémoire) n'est pas si restrictive, et que l'espace d'état peut être suffisamment grand pour modéliser des processus complexes par des processus de Markov. Néanmoins, ces processus ne seraient plus adaptés lorsque l'on cherche explicitement à modéliser des phénomènes avec mémoire, par exemple une mémoire épigénétique [90]. Parmi les processus de Markov, mentionnons les processus de branchement (voir [91] pour les applications en biologie) qui permettent par exemple de décrire l'évolution au cours du temps d'une population de cellules – indépendantes les unes des autres – qui, lorsqu'elles se divisent, peuvent donner naissance à des cellules filles de différents types, suivant différentes probabilités (Fig. 8-A). Un tel formalisme est par exemple utilisé par Xu et al. pour un modèle de division et différenciation de cellules hématopoïétiques, qu'ils calibrent à partir de données de barcodes cellulaires [92]. Mentionnons également deux processus Markoviens qui ont été couramment utilisés en génétique des populations [93], et dont on retrouve l'usage pour inférer le développement clonal dans certains cancers du sang (comme nous l'illustrerons plus loin) : le processus de Moran [94] et le processus de Wright-Fisher [95, 96]. Il s'agit tous deux de processus à temps discret, c'est-à-dire que l'on s'intéresse à l'état du processus à différents instants $\{t, t+\Delta t, t+2\Delta t, \cdots\}$, par opposition aux modèles à temps continu. Contrairement aux processus de branchement pour lesquels le nombre de cellules total varie généralement au cours du temps, les processus de Moran ou de Wright-Fisher font l'hypothèse d'une population de taille constante.
Plus généralement, modéliser l'hématopoïèse, c'est définir un ensemble de règles et relations qui décrivent la dynamique et le comportement de différentes cellules hématopoïétiques, au cours du temps, éventuellement les unes par rapport aux autres, potentiellement dans l'espace. Les modèles n'aboutissent pas nécessairement à une formalisation sous la forme d'équations. Ainsi, les modèles multi-agents (un agent correspondant par exemple à une cellule), dont l'étude repose essentiellement sur des simulations numériques, se développent de plus en plus, favorisés par la diminution des coûts de calculs et l'usage de super-calculateurs. Bessonov et al. ont développé un logiciel basé sur un modèle multi-agents pour simuler et visualiser la dynamique de cellules hématopoïétiques [97]. Krinner et al. ont combiné un modèle déterministe avec un modèle multi-agents afin de décrire la dynamique de la granulopoïèse sous différents scénarios : une chimiothérapie ou une transplantation par exemple [98]. L'intérêt des modèles multi-agents repose notamment sur la possibilité de décrire des modèles aussi complexes qu'on le souhaite, mais ils s'accompagnent d'enjeux méthodologiques liés par exemple à leur calibration.

3.2 Inférer l'apparition et le développement des hémopathies malignes

Les hémopathies malignes sont encore trop souvent détectées tardivement, après apparition des symptômes et un développement clonal important. Comprendre leur dynamique d'apparition chez l'humain, à partir d'une cellule mutée, nécessite de pouvoir retracer le développement du cancer. Les modèles mathématiques sont particulièrement adaptés dans ce cas, avec pour objectif d'inférer ce qui n'a pas pu être observé (l'apparition et l'expansion clonale précoce), à partir de certaines règles décrivant le comportement des cellules.
Les modèles mathématiques peuvent par exemple être utilisés pour essayer de comprendre quelles sont les cellules à l'origine des cancers du sang. Modélisant la dynamique hématopoïétique par un processus de Moran, Traulsen et al. ont ainsi étudié le cas de l'hémoglobinurie paroxystique nocturne [99]. Avec une autre approche – basée sur des systèmes multi-agents et qui peut se généraliser à plusieurs cancers du sang – Haeno et al. ont étudié le cas des néoplasmes myéloprolifératifs positifs pour la mutation $JAK2^{V617F}$ [100]. Alors que l'hypothèse privilégiée consiste en un cancer de la cellule souche, Haeno et al. ont exploré des hypothèses alternatives, dont celle où la mutation $JAK2^{V617F}$ surviendrait au niveau d'un progéniteur qui aurait acquis une capacité d'auto-renouvellement. Leurs conclusions nécessiteraient d'être confirmées par des expériences

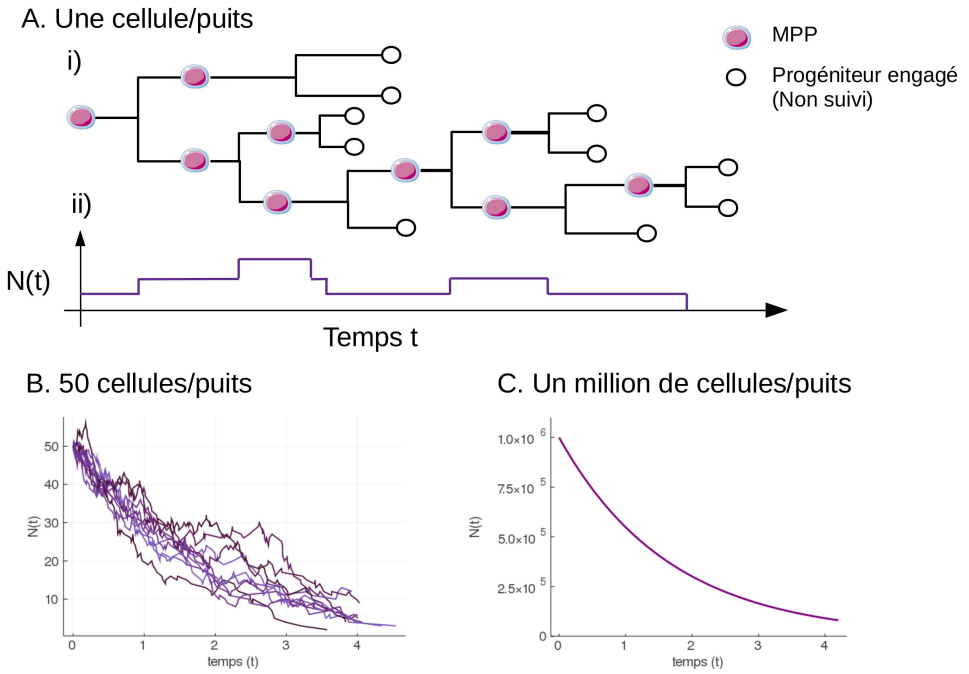


FIGURE 8 – Comparaison entre un modèle stochastique et son approximation déterministe. L'évolution du nombre de cellules hématopoïétiques d'une population donnée (considérons ici des progéniteurs multipotents – MPP) peut se modéliser par un processus de branchement à temps continu dans lequel un MPP se diviserait à un taux $\alpha = 0.5$ jour^{-1}, et lorsqu'il se diviserait pourrait produire 2 MPP avec une probabilité $p_2 = 0.2$, 1 MPP et un progéniteur engagé (pour lequel on ne fait pas le suivi dans le temps des divisions) avec une probabilité $p_1 = 0.3$ ou 0 MPP avec une probabilité $p_0 = 0.5$.

A. Représentation d'une trajectoire aléatoire pour le processus stochastique considéré, partant d'un MPP mis en culture dans un puits à $t = 0$. i) Schéma des branchements (divisions) et ii) évolution correspondante du nombre $N(t)$ de MPP dans le puits au cours du temps.

B. Représentation de 10 trajectoires aléatoires simulées suivant le modèle. Ces 10 trajectoires peuvent être interprétées comme l'évolution du nombre de MPP dans 10 puits dans lesquels on aurait mis 50 MPP initialement. Le modèle étant stochastique, pour des valeurs données des paramètres, on observe différentes dynamiques.

C. Lorsque le nombre de cellules devient très grand, la variabilité devient négligeable et on peut faire une approximation déterministe du modèle. Pour des valeurs données des paramètres, une seule dynamique déterministe est alors produite par le modèle, correspondant à la trajectoire de la population.

et des données, mais leur travail illustre la capacité des modèles mathématiques à tester différentes hypothèses biologiques et à en remettre en cause certaines. En l'occurrence, ils démontrent qu'il serait plus probable d'avoir l'acquisition de deux mutations – une mutation conférant une capacité d'auto-renouvellement puis la mutation $JAK2^{V617F}$ – au niveau d'un ensemble de cellules se divisant fréquemment (les progéniteurs) plutôt qu'une seule mutation au niveau d'un ensemble de cellules souches ayant déjà une capacité d'auto-renouvellement mais se divisant peu fréquemment. Leur modèle, comme tout modèle, repose sur des hypothèses, et permet également d'identifier certains paramètres clés pour lesquels le modèle est sensible et donc sur lesquels des investigations devraient être menées : le nombre de cellules souches, leur fréquence de divisions, le nombre de divisions subies par les progéniteurs. La question du nombre de cellules souches contribuant à l'hématopoïèse est une question centrale. Lyne et al., à partir de simulations d'un modèle de Moran, montrent ainsi que, contrairement à l'intuition, une évolution clonale linéaire (généralement attribuée à un mécanisme de sélection naturelle) pourrait également être obtenue sans que les mutations n'aient d'avantage sélectif, à condition que le nombre de cellules souches contribuant à l'hématopoïèse soit faible [101]. Or, les estimations de leur nombre varient fortement suivant les auteurs. Dingli et al., modélisant la taille du compartiment des cellules souches hématopoïétiques actives N_{HSC} entre les différents mammifères par une loi allométrique $N_{HSC} \sim M^{3/4}$ (avec M la masse du mammifère), estiment qu'environ 400 HSC pourraient contribuer activement à l'hématopoïèse chez l'humain [102]. Plus récemment, Lee-Six et al. ont estimé que leur nombre se situerait plutôt autour de 100,000 [15], estimation qui semble aujourd'hui privilégiée notamment parce qu'elle se base sur des données expérimentales obtenues chez l'homme. Leur méthode repose sur le séquençage complet de colonies de progéniteurs hématopoïétiques d'un individu (sans pathologie connue), la construction d'un arbre phylogénétique à partir de l'information sur les mutations somatiques accumulées par les cellules, puis la calibration d'un modèle de Moran à partir d'une méthode ABC (Approximate Bayesian Computation) [103]. Citons également Mitchell et al. [43] qui ont appliqué une méthodologie semblable avec les données de 10 individus, et ont trouvé des résultats comparables.

Les méthodes reposant sur la construction d'arbres phylogénétiques peuvent également être employées pour retracer l'histoire du développement clonal de certains cancers. Par l'identification d'ancêtres communs et l'hypothèse d'une accumulation des mutations somatiques linéaire au cours du temps [104] (qui en quelque sorte reproduirait une horloge moléculaire), on peut chercher à inférer l'âge d'apparition puis l'expansion d'un clone mutant responsable de la maladie ainsi que quantifier l'avantage prolifératif associé à la mutation. Ces méthodes ont notamment été utilisées dans le cas des NMP. Van Egeren et al., estimant les paramètres d'un modèle de Wright-Fisher à partir des données (structurées en arbres phylogénétiques) de deux patients atteints de NMP, ont ainsi montré que la mutation $JAK2^{V617F}$ pouvait apparaître tôt au cours de la vie, lors de l'enfance ou l'adolescence, et qu'elle entraînait un avantage prolifératif au niveau des cellules souches [41]. Par une approche similaire, l'étude de 12 patients et le choix comme modèle d'un processus de naissance et de mort (qui est un processus de Markov), Williams et al. ont quant à eux montré que le taux de croissance des clones mutés $JAK2^{V617F}$ était variable suivant les individus, qu'on pouvait le relier à la durée de latence entre l'acquisition de la mutation et le diagnostic de la maladie, et que des taux de croissance plus élevés étaient trouvés pour des clones abritant plusieurs mutations driver (motrices) [43]. Ils ont également trouvé que la mutation $JAK2^{V617F}$ pouvait apparaître durant la vie fœtale.

Les méthodes reposant sur la construction d'arbres phylogénétiques ne sont pas les seules permettant d'inférer le développement clonal. A partir de l'information sur la fréquence allélique (VAF – Variant Allele Frequency) mesurée dans le sang périphérique pour près de 500,000 individus, et l'utilisation d'un modèle de branchement pour modéliser la dynamique des HSC, Watson et al. ont étudié le cas de l'hématopoïèse clonale et notamment quantifié l'avantage sélectif de plusieurs mutations, telles que $JAK2^{V617F}$, *SFRSF2* ou *DNMT3A* [42]. Comme nous le présenterons au chapitre 5, nous avons pour notre part également travaillé sur un modèle de branchement, calibré à partir de données de fractions clonales au niveau de progéniteurs pour des patients mutés $CALR^m$ ou $JAK2^{V617F}$, pour estimer des différences dans la dynamique d'apparition des

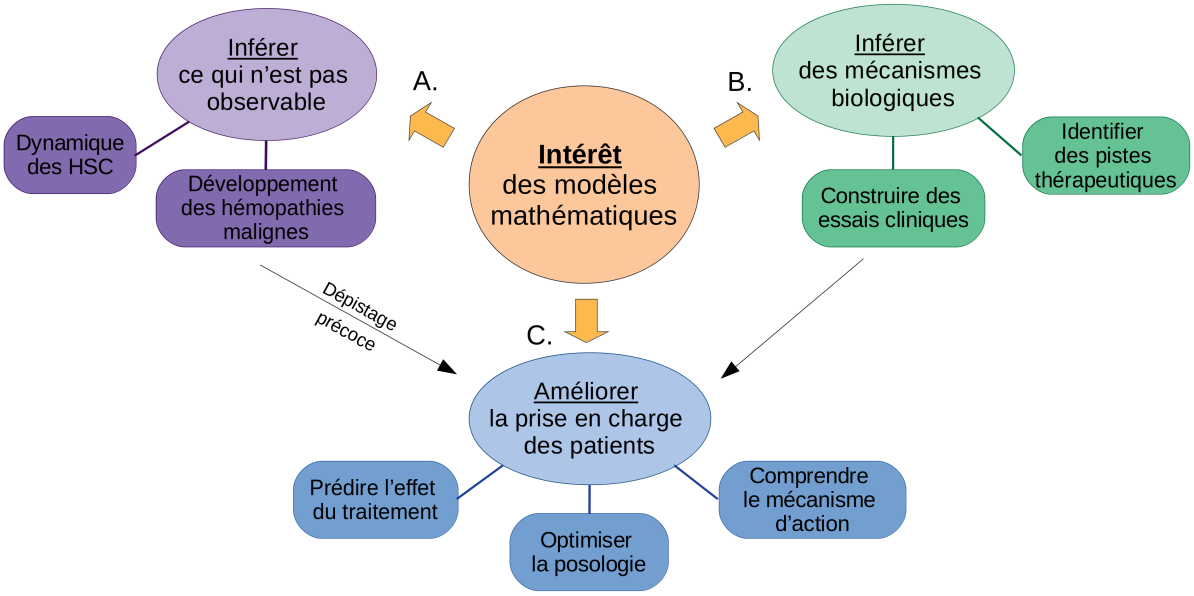


FIGURE 9 – Intérêt des modèles mathématiques. Différents intérêts (liste non exhaustive) des modèles mathématiques.
A. Les modèles mathématiques peuvent permettre d'inférer des processus latents non directement observables, tels que la dynamique au niveau de cellules souches (HSC) ou l'histoire de l'apparition et du développement des hémopathies malignes, ce qui peut permettre par la suite la mise en place de méthodes de dépistage précoce. Nous présenterons nos travaux portant sur ce sujet au chapitre 5.
B. Les modèles mathématiques peuvent inférer des mécanismes biologiques, en testant par exemple différentes hypothèses biologiques, ce qui peut servir à identifier des pistes thérapeutiques et à construire des essais cliniques qui permettront leur validation et leur usage en clinique (voir chapitres 6 et 7) .
C. Les modèles mathématiques sont des outils d'aide à la décision clinique, permettant de comprendre le mécanisme d'action des traitements, de prédire leurs effets et d'optimiser leur posologie (voir chapitre 8).

mutations ($CALR^m$ apparaissant plus tardivement) puis l'avantage prolifératif entre ces deux mutations motrices des NMP ($CALR^m$ conférant un avantage prolifératif plus important au niveau souche). Nos travaux illustrent également l'utilisation de modèles mathématiques pour la mise en place de méthodes de dépistage précoce [105].
Au total l'utilisation de ces différents modèles ont permis des avancées importantes pour la médecine prédictive notamment en ouvrant la voie à la détection des patients à risque de développer des NMP (ou d'autres hémopathies malignes) et encouragent l'utilisation de traitements visant à intercepter de façon précoce les clones malins (Fig. 9-A).

3.3 Modéliser l'hématopoïèse altérée

De nombreux auteurs ont modélisé l'hématopoïèse comme un système dynamique. Mathématiquement, une hématopoïèse physiologique impliquerait que le système soit stable. Les maladies, cancers ou stress hématopoïétiques peuvent alors être vus, dans ce cadre, comme une perturbation du système qui le déstabilise d'un état sain vers un état pathologique. C'est par exemple ce que font Stiehl et Marciniak-Czochra [106] où, partant d'un précédent modèle [83], étudient qualitativement les propriétés que devraient avoir les cellules leucémiques pour conduire à un développement de la maladie. Ils peuvent ainsi décliner leur modèle en plusieurs scénarios, suivant les valeurs prises par les paramètres, et par exemple décrire le cas des syndromes myélodyspla-

siques dans lesquels les cellules cancéreuses pourraient mettre plus de temps à se diviser mais avoir une augmentation de leur capacité d'auto-renouvellement, ou encore les lymphomes de Burkitt où la prolifération cellulaire serait plus importante dans les cellules mutantes comparées aux saines. La biologie des cellules cancéreuses est bien sûr plus complexe que celle décrite dans les modèles, mais ces derniers ont l'intérêt de se concentrer sur les paramètres les plus susceptibles d'induire la dynamique du cancer, et ainsi de mettre en évidence les caractéristiques des cellules mutées qu'on pourrait souhaiter infléchir, par l'usage de traitements, pour retrouver une hématopoïèse normale. On retrouve cette approche chez Moore et Li qui proposent un modèle de LMC (formalisé par un système d'équations différentielles non linéaires) dans lequel ils étudient l'influence des paramètres sur la valeur maximale atteinte par la concentration en cellules leucémiques [107]. Leur démarche correspond à une analyse de sensibilité ; au-delà de leurs conclusions biomédicales, ils illustrent une méthode qui permettrait de distinguer rigoureusement les paramètres d'un modèle qui devraient être estimés à partir d'observations de patients, de ceux qui pourraient être fixés constants à partir d'*a priori* biologiques (généralement issus de la littérature) [108, 109]. En effet, les modèles mathématiques sont généralement construits pour être aussi simples que possible, avec un minimum de paramètres à estimer, sans quoi il devient difficile de tirer des conclusions de leurs analyses. On retrouve par exemple cette démarche de construction d'un modèle parcimonieux chez Andersen et al. [110], qui étudient par modélisation mathématique le lien potentiel entre inflammation et développement des NMP, notamment la progression entre thrombocytémie essentielle, polyglobulie de Vaquez et myélofibrose primaire. Notons que, dans les exemples précédents, on ne retrouve pas d'étape de calibration de modèles à partir de données longitudinales de patients avant traitement du fait de leur prise en charge thérapeutique. Ainsi, l'estimation des paramètres à partir de données de patients est le plus souvent effectuée lorsque l'on étudie des modèles avec prise en compte d'un traitement, comme nous l'illustrerons dans la section suivante. Néanmoins, même si l'hématopoïèse murine diffère de celle de l'homme, il peut être intéressant de calibrer des modèles d'hématopoïèse pathologique à partir d'observations longitudinales chez des souris. Ce type de données est par exemple utilisé par Bonnet et al. pour étudier l'hématopoïèse de stress [111]. Bonnet et al. modélisent l'érythropoïèse avec un modèle à six compartiments, des HSC jusqu'aux érythrocytes, en passant par différents types de progéniteurs. À un système d'équations différentielles ordinaires permettant de décrire l'hématopoïèse saine, ils ajoutent de la régulation, leur permettant alors de modéliser également l'hématopoïèse de stress, en particulier une anémie hémolytique associée à de l'inflammation. Ils induisent ce phénomène chez des souris par l'administration de phénylhydrazine et utilisent les mesures expérimentales pour calibrer leur modèle. Pour trouver les paramètres qui minimisent l'écart entre les observations et la sortie du modèle (et donc résoudre un problème d'optimisation), ils utilisent l'algorithme CMA-ES (Covariance Matrix Adaptation - Evolution Strategy) [112].
Comme nous l'avons montré à l'aide de quelques exemples choisis parmi une littérature très riche sur le sujet, modéliser l'hématopoïèse altérée – dans le cas des cancers du sang notamment – peut permettre de mettre en évidence les types de cellules mises en cause dans les dérèglements de l'hématopoïèse pour éventuellement identifier des pistes thérapeutiques (Fig. 9-B). Nous présenterons dans le prochain paragraphe des exemples de modèles s'intéressant au traitement des hémopathies.

3.4 Comprendre, prévoir et optimiser l'effet d'un traitement

Pour reprendre les termes de Clapp et Levy [113], *« les modèles mathématiques sont un outil de recherche puissant qui peut être appliqué à la compréhension des leucémies et lymphomes. Ils peuvent identifier des mécanismes qui contrôlent la progression de la maladie, ou motiver et guider des expériences futures et des essais cliniques. En fin de compte, combiner modélisation mathématique, expériences et essais cliniques peut conduire à des améliorations significatives dans le traitement des leucémies et des lymphomes. »* Et de l'ensemble des cancers du sang, pourrions-nous ajouter. C'est lorsque les modèles sont confrontés à des données qu'ils nous semblent pleinement se révéler être de *« puissants outils de recherche »*. Michor et al. analysent

ainsi à partir d'un modèle mathématique les données de 169 patients atteints de LMC et traités à l'imatinib (mesure du niveau de transcrit de fusion BCR-ABL dans le sang en cours de traitement) [114]. À l'aide d'un modèle à 4 compartiments décrivant la dynamique des cellules sous traitement, ils suggèrent que l'imatinib agirait sur les progéniteurs leucémiques mais pas sur les cellules souches, et prédisent alors que l'arrêt du traitement conduirait à une rechute. À la suite de l'article de Michor et al., l'étude sous un angle mathématique de la LMC et son traitement à l'imatinib a alors connu un important développement. Roeder et al. ont construit un modèle multi-agents qu'ils ont simulé et comparé aux données de deux cohortes de patients atteints de LMC et traités à l'imatinib [115]. Contrairement aux résultats précédents de Michor et al. [114], Roeder et al. montrent que les données peuvent également être en accord avec un effet du traitement au niveau des cellules souches, et prédisent une potentielle rémission à long-terme (comme ils le soulignent néanmoins, le développement de clones résistants au traitement pourrait réduire les chances de succès de la thérapie), rémission qui pourrait être accélérée en stimulant la prolifération des HSC. Foo et al. ont alors proposé par la suite un modèle prenant en compte les HSC quiescentes pour étudier l'effet combiné du G-CSF (Granulocyte-Colony Stimulating Factor) et de l'imatinib [116]. Ils prédisent, à partir de leur modèle et de données patients, que l'ajout de G-CSF pourrait augmenter le risque de résistance à l'imatinib et déconseillent cette option. L'IFNα pourrait être ce candidat qui, combiné à l'imatinib, améliorerait la réponse au traitement. C'est ce que suggèrent Glauche et al. [117] où ils étendent le modèle multi-agents de Roeder et al. [115] pour y inclure un effet de l'IFNα. Plus récemment, Bunimovich-Mendrazisky et al. [118], à partir d'un modèle de LMC faisant intervenir des populations de cellules souches leucémiques, cellules matures leucémiques et lymphocytes T cytotoxiques, ont simulé l'effet de l'IFNα combiné avec l'imatinib. Dans leur article reposant sur la simulation de leur modèle sous différentes hypothèses, ils montrent que plusieurs scénarios seraient favorables à l'utilisation de l'IFNα qui, combiné à l'imatinib, pourrait permettre d'induire une réponse moléculaire complète et de l'accélérer. À la fois pour [117] et [118], la portée de leurs résultats est limitée par le fait que leurs modèles ne sont pas calibrés à partir de données de patients, mais ils proposent néanmoins des pistes intéressantes de thérapie qui pourraient être ensuite testées cliniquement. Comme Clapp et Levy [113] le soulignent dans leur revue, les modèles mathématiques peuvent en effet être utilisés pour tester différents scénarios *in silico*, et éventuellement orienter certaines pistes de recherche : chez [117] et [118], la modélisation mathématique peut être alors vue comme un outil permettant de sélectionner des pistes de recherche prometteuses ou pour construire des essais cliniques.

Nous avons mentionné plus haut l'IFNα comme traitement potentiel associé à l'imatinib dans le cas des LMC. Son utilisation dans un autre type de cancers du sang – les NMP non BCR-ABL – a également été démontrée puis étudiée mathématiquement par différents groupes. À partir des données d'une cohorte danoise (DALIAH trial, mesures de la VAF pour $JAK2^{V617F}$ au niveau des cellules matures), Ottesen, Pedersen et al. [119, 120] basant leur travail sur le modèle Cancitis [121, 110], ont modélisé l'effet de l'IFNα sur le taux de mortalité des cellules souches mutées. Les dynamiques observées pour les patients de leur cohorte sont en accord avec leur modèle. De plus, ils démontrent l'utilité de leur modèle comme outil d'aide à la décision pour prédire le résultat du traitement à l'échelle du patient. Comme nous le présenterons au chapitre 6, nous avons également étudié l'action de l'IFNα sur les progéniteurs hématopoïétiques en mesurant régulièrement l'architecture clonale des mutations chez des patients atteints de NMP, puis à partir de ces données, calibré un modèle mathématique permettant d'inférer l'action de l'IFNα au niveau des cellules souches initiatrices de la maladie, en distinguant l'effet suivant les clones hétérozygotes et homozygotes [122]. Étendant ce modèle pour prendre en compte les variations de posologies, nos travaux présentés au chapitre 7 suggèrent que l'usage de doses d'IFNα suffisamment élevées seraient nécessaires pour induire une rémission sur le long terme dans le cas de patients présentant la mutation (que ce soit hétérozygote ou homozygote) $JAK2^{V617F}$ [123].

Les quelques exemples précédents, malgré leur choix subjectif et ici axé sur les NMP, permettent d'illustrer l'usage de modèles mathématiques pour étudier l'effet de traitements, à partir de données patients ou *in silico*, pour ensuite faire de la prédiction ou de l'optimisation de trai-

tement (Fig. 9-C). Bien que nous n'ayons pas mis l'accent sur les questions techniques dans le paragraphe ci-dessus, les auteurs s'emploient également à mettre en place une méthodologie s'assurant de la robustesse des résultats, de leur reproductibilité et de leur validité. Nous évoquerons les enjeux méthodologiques à la section suivante. Soulignons néanmoins que les modèles mathématiques sont à chaque fois construits suivant certaines hypothèses. La validité des résultats est nécessairement conditionnée aux différentes hypothèses explicitées par leurs auteurs. Différentes hypothèses peuvent conduire à des interprétations qui diffèrent, comme sur l'exemple de Roeder et al. [115] qui obtenaient des conclusions différentes de celles de Michor et al. [114]. Les modèles mathématiques, lorsqu'ils sont utilisés pour élucider un mécanisme d'action d'un médicament par exemple, peuvent permettre de conclure en faveur ou en défaveur de certaines hypothèses ; ils ne remplacent pas l'étape de validation biologique mais la complètent.

4 Enjeux méthodologiques

La modélisation mathématique d'un processus biologique – ici l'hématopoïèse – est une démarche qui nécessite une étroite collaboration entre mathématiciens et biologistes. Aboutir à un modèle (ou un ensemble de modèles potentiels) prend du temps : il s'agit de comprendre le phénomène étudié et de faire des choix selon la question de recherche d'intérêt et l'objectif poursuivi. Nous ne rentrerons pas ici dans une description méthodologique des étapes nécessaires pour aboutir à la formalisation d'un modèle, mais nous nous intéresserons aux étapes qui viennent ensuite, lorsque l'on souhaite calibrer le modèle à partir de données réelles, afin par exemple d'en interpréter les valeurs des paramètres, de faire de la prédiction, d'inférer des mécanismes biologiques non observés ou encore d'optimiser l'administration d'un traitement. Nous considérons donc dans cette section avoir un modèle (ou un ensemble de plusieurs modèles potentiels) et que ce modèle est paramétrique, c'est-à-dire qu'il fait intervenir un nombre fini de paramètres inconnus qu'on souhaite estimer à partir d'observations expérimentales (notons que ce modèle peut également faire intervenir d'autres paramètres dont on connaît la valeur, qu'on appellera ici constantes). Ce modèle peut par exemple décrire la dynamique d'un système en simulant ses variables d'état $X(t)$ au cours du temps t : elles constitueront les sorties du modèle (qu'on nommera parfois « observations théoriques »). La figure 10 illustre, sur un exemple fictif, différents points abordés dans cette section.

4.1 Prétraitement des données

Pour calibrer un modèle, comme nous le verrons un peu plus loin, il faut pouvoir comparer la sortie du modèle à des observations qui lui correspondent. Les données brutes, issues d'expériences, sont rarement directement utilisables telles qu'elles. Un travail préliminaire sur les données (que nous appelons ici prétraitement) est souvent nécessaire. Chaque situation rencontrée est différente, mais évoquons cette problématique sur quelques exemples.

Tout d'abord, il est important d'évaluer l'incertitude que l'on a sur les observations. Les mesures ne sont jamais exactes : elles s'accompagnent d'un bruit de mesure (Fig. 10-A). Pour les gérer, on définit généralement un modèle statistique pour les observations, le plus courant étant de considérer que les observations expérimentales sont distribuées suivant une loi gaussienne, de moyenne la valeur prédite par le modèle (correspondant à l'hypothèse choisie), et avec une certaine dispersion (variance de la loi). Mais d'autres modèles peuvent être plus pertinents suivant la situation rencontrée, par exemple si les observations sont nécessairement positives ou appartenant à un certain intervalle. Gérer les incertitudes, c'est aussi être en mesure de prendre en considération les événements non observés. Par exemple si l'on observe des divisions de cellules seulement jusqu'à un certain temps, il ne faut pas interdire dans le modèle que des cellules puissent se diviser après ce temps, sinon cela introduirait un biais dû à l'observation incomplète du processus. Nous évoquerons ce cas au chapitre 3.

Les données peuvent également ne pas être analysables directement sous leur forme brute. Par exemple, les données de séquençage du génome (WGS – Whole Genome Sequencing) de plusieurs cellules peuvent être structurées sous la forme d'arbres phylogénétiques, forme qui servira ensuite à leur analyse et à la calibration de modèles.

Enfin, une problématique fréquemment soulevée est celle de l'hétérogénéité, par exemple au niveau des cellules hématopoïétiques. À partir de l'expression de marqueurs de surface et de la définition de seuils, on peut trier les cellules en populations, par exemple HSC* ou MPP suivant l'expression du marqueur CD90+ chez l'homme tout en négligeant l'hétérogénéité entre cellules au sein des populations. Les modèles compartimentaux sont adaptés lorsque les données sont structurées sous cette forme. Or, comme des analyses single-cell RNA-seq ont pu le mettre en évidence récemment (voir [12]), les cellules hématopoïétiques constituent plus un continuum d'états qu'un ensemble de populations distinctes (problématique que nous aborderdons au chapitre 2), ce qui pourrait se modéliser par des équations aux dérivées partielles. Ainsi, suivant la question de recherche, on peut choisir le niveau de détail à considérer dans les données et choisir le type de modèle en conséquence. La question de l'hétérogénéité se pose également lorsque l'on

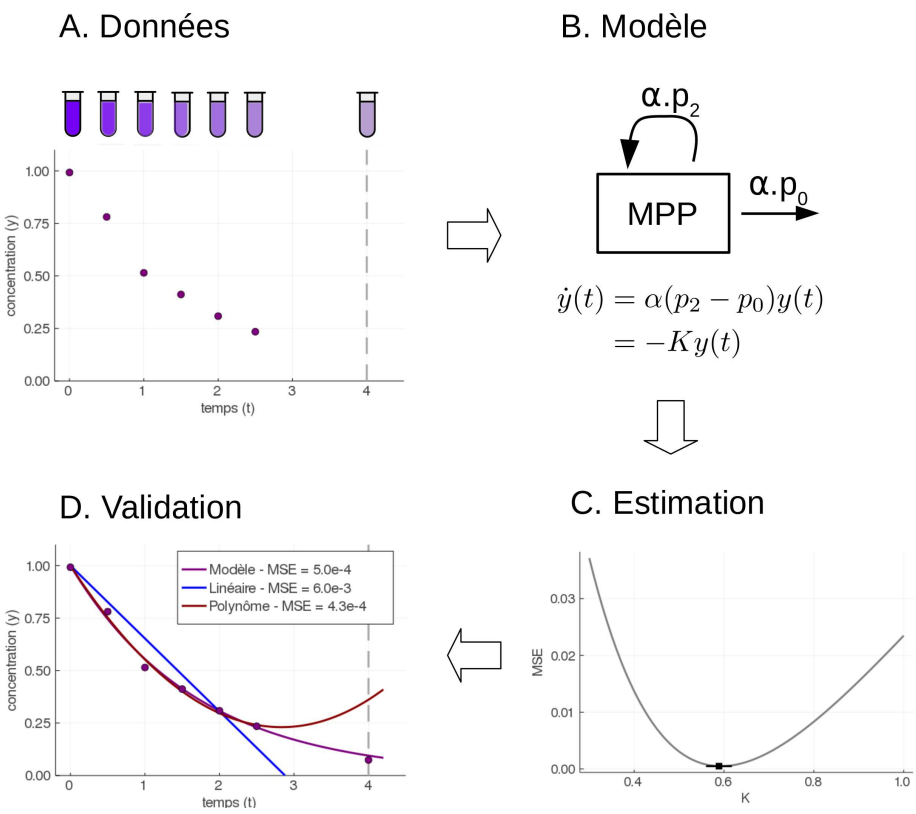


FIGURE 10 – Illustration de la méthodologie sur un exemple fictif.
A. Considérons une expérience fictive dans laquelle on mesure une concentration en MPP $y(t)$ à différents instants t (en jours). Les données sont bruitées. L'échantillon à 4 jours est l'échantillon contrôle. On souhaite prédire la dynamique de diminution du nombre de MPP au delà de 3 jours.
B. Pour répondre à cet objectif, on construit un modèle mathématique à un compartiment, modélisé par une équation différentielle ordinaire. La solution de cette équation est une loi exponentielle (modèle exponentiel). Il n'est pas possible d'estimer à la fois α, p_2 et p_0 (non-identifiabilité du modèle). En définissant $K = \alpha(p_0 - p_2)$, le modèle devient identifiable.
C. Pour estimer la valeur de $K_{\text{réel}}$, on minimise une distance (l'erreur quadratique moyenne – mse) entre les observations et les valeurs du modèle, en fonction de la valeur de K. On trouve comme valeur optimale $K_{\text{estimé}} = 0.59$ et comme intervalle de confiance à 95% : $K_{\text{estimé}} \in [0.56, 0.62]$.
D. On peut comparer les résultats du modèle exponentiel (mse : $e = 5.0e-4$) avec ceux qu'on obtiendrait par une régression linéaire ($e = 6.0e-3$) ou polynomiale ($e = 4.3e-4$). On obtient des résultats légèrement meilleurs avec un modèle polynomial plutôt qu'avec le modèle exponentiel, mais le polynôme fait aussi intervenir un paramètre en plus : il est moins parcimonieux. Pour valider le modèle exponentiel, on peut alors comparer l'erreur de prédiction sur l'observation à 4 jours et constater qu'il permet en effet de bien prédire la concentration en MPP, contrairement aux modèles polynomial et linéaire. Notons que sur cet exemple très simple, les données avaient été générées par le modèle exponentiel avec $K_{\text{réel}} = 0.6$ et un bruit additif Gaussien d'écart-type $\sigma = 0.05$.

analyse les données de plusieurs individus. Peut-on regrouper les données ensemble, ou faut-il considérer les individus indépendants les uns des autres ? Des tests d'hypothèses (par exemple Mann-Whitney [124] ou Kruskal-Wallis [125] lorsque l'on a plus que deux échantillons, ou des hypothèses sur la nullité des variances dans le cadre des modèles à effets mixtes [126]) peuvent être utiles pour indiquer si l'hypothèse d'observations distribuées suivant une distribution commune peut être rejetée, auquel cas il faudrait en toute rigueur traiter l'hétérogénéité inter-individuelle. Or, considérer les individus indépendants les uns des autres revient également à négliger un effet populationnel. Plutôt que d'un côté considérer les individus indépendants et d'un autre regrouper leurs données ensemble, on peut recommander l'usage de modèles à effets mixtes [127] ou Bayésiens hiérarchiques [128] qui prennent à la fois en compte la variabilité inter-individuelle et les similarités au sein de populations. Nous aborderons ce point au chapitre 5.

4.2 A priori biologiques

Les modèles mathématiques ne sont généralement pas créés *de novo*, mais reposent au contraire sur des *a priori* biologiques. C'est notamment un de leurs intérêts : ils incorporent au travers de nombreuses hypothèses des connaissances préalables sur le système biologique étudié. Les connaissances peuvent être précises, comme le choix de valeurs pour certaines constantes du modèle, ou plus vagues, comme par exemple les valeurs minimales ou maximales que peuvent prendre les paramètres du modèle qui seront à estimer. Le cadre Bayésien (dans lequel les paramètres du modèle sont considérés comme des variables aléatoires dont on souhaiterait estimer la distribution de probabilité à partir des observations à notre disposition) repose sur le choix de distributions *a priori* pour les paramètres. Avec une procédure d'estimation Bayésienne, les données viennent, en quelque sorte, actualiser notre connaissance sur le système étudié. Mentionnons Hoekstra et al. qui proposent (sur l'exemple de [129]) l'utilisation de l'approche Bayésienne pour plus de transparence dans la présentation des résultats lors de la rédaction d'articles scientifiques [130]. Les priors sur les paramètres (ou les modèles, lorsque l'on en compare plusieurs entre eux) pourraient ainsi être définis de trois façons différentes, correspondant aux stéréotypes suivants : le sceptique, l'agnostique et le convaincu. Partant de ces trois différents *a priori*, on peut alors comparer les distributions *a posteriori* obtenues (c'est-à-dire après prise en compte des données) et voir si elles sont proches, ce qui suggérerait alors que les observations expérimentales sont suffisantes pour convaincre même les plus réticents.

4.3 Identifiabilité et sensibilité

S'assurer de l'identifiabilité d'un modèle est essentiel quand on souhaite pouvoir interpréter les valeurs estimées pour les paramètres. Un modèle est dit identifiable si on peut estimer sans ambiguïté la valeur de ses paramètres à partir de données expérimentales. Sinon, il ne l'est pas. La non identifiabilité est un problème car cela signifie par exemple que deux valeurs différentes pour un même paramètre pourraient conduire aux mêmes observations théoriques (Fig. 10-B). Détecter les cas de non-identifiabilité n'est pas si simple, notamment lorsque le modèle est complexe et met en jeu de nombreux paramètres. Pour un exemple appliqué à l'hématopoïèse, citons Duchesne et al. [131] qui évaluent l'identifiabilité de leur modèle d'érythropoïèse à partir d'une méthode basée sur la vraisemblance profilée [132]. Pour lever le problème de non-identifiabilité, une fois celui-ci détecté, il faut soit avoir plus de données expérimentales, soit réduire le nombre de paramètres à estimer, en augmentant notre *a priori* sur les paramètres (par exemple en choisissant certains paramètres constants, fixés à des valeurs trouvées dans la littérature). Pour faire ce choix, une possibilité est de déterminer quels sont les paramètres qui ont le moins d'influence sur la sortie du modèle : c'est-à-dire trouver quels paramètres peuvent varier sans fortement impacter la sortie du modèle. On dit alors que le modèle est peu sensible à ces paramètres ; l'approche consiste ainsi à mener une analyse de sensibilité. Parmi les différentes techniques existantes, mentionnons le calcul d'indices de Sobol [133]. Nous illustrerons l'analyse de sensibilité au chapitre 4, et les problématiques d'identifiabilité aux chapitres 5 et 6.

4.4 Estimation des paramètres et de l'incertitude

Lorsque le modèle est formalisé, que les données ont été prétraitées et qu'on a défini les paramètres à estimer (après s'être assuré de l'identifiabilité du modèle), on peut procéder à la calibration du modèle. Cette étape revient à estimer les valeurs des paramètres telles que la sortie du modèle se rapproche le plus des observations. Cela nécessite de définir une distance entre les observations et la sortie du modèle, généralement déduite de la vraisemblance des paramètres étant donné les observations. La plus fréquemment rencontrée est l'erreur moyenne quadratique (MSE - mean squared error), associée à un modèle gaussien pour le modèle statistique des observations. En faisant varier les valeurs des paramètres, cette distance varie également, et on souhaite alors trouver le jeu de paramètres qui va la minimiser (Fig. 10-C). Le problème ainsi défini est un problème d'optimisation pour lequel il est assez rare de trouver une solution explicite. On résout alors ce problème numériquement. Il existe de nombreux algorithmes pour cela, parmi lesquels on peut citer les algorithmes de quasi-Newton très efficaces mais seulement adaptés aux problèmes convexes [134] et la famille des algorithmes évolutionnaires (mentionnons les algorithmes génétiques [135], les méthodes d'optimisation par essaims particulaires [136], ou encore les méthodes de recuit simulé [137]), dont fait partie l'algorithme CMA-ES (Covariance Matrix Adaptation - Evolution Strategy), développé assez récemment et qui montre de bonnes performances pour une large catégorie de problèmes, incluant ceux qui sont non-linéaires et en grande dimension [112]. Nous utiliserons l'algorithme CMA-ES tout au long de ce manuscrit (voir chapitres 4, 6 et 7).
L'approche présentée ci-dessus correspond à une approche dite fréquentiste dans laquelle on considère qu'il existe une vraie valeur pour les paramètres, qu'on souhaite estimer à partir des données. Dans la théorie Bayésienne, au contraire, les paramètres sont considérés comme des variables aléatoires dont on veut estimer la loi *a posteriori*, en combinant l'information *a priori* qu'on a sur elles (via la distribution *a priori*) et l'information provenant des données (via la vraisemblance, i.e. la probabilité d'avoir les données étant données les valeurs des paramètres). Avec une procédure d'estimation Bayésienne, on peut ainsi obtenir une estimation ponctuelle des paramètres (en choisissant par exemple la moyenne *a posteriori*), mais également un écart-type, ou mieux encore la probabilité que le paramètre soit compris entre telle et telle valeur. De nombreux algorithmes existent pour estimer la loi *a posteriori* suivant cette approche Bayésienne, dont le plus ancien est l'algorithme de Metropolis-Hasting [138]. Cet algorithme repose sur la construction d'une chaîne de Markov (MCMC – Markov Chain Monte-Carlo), et a connu de nombreuses améliorations permettant une convergence plus rapide et en plus grande dimension (par l'utilisation d'algorithmes adaptatifs par exemple [139]). Citons également les méthodes ABC (Approximate Bayesian Computation) [103] qui connaissent un fort essor ces dernières années et qui sont notamment utilisées pour la calibration de modèles complexes, par exemple des modèles multi-agents pour lesquels il est facile de simuler le modèle mais compliqué d'exprimer une vraisemblance. Nous présenterons une utilisation de la méthode ABC au chapitre 5. Une fois le modèle calibré, nous avons une estimation de la valeur de ses paramètres. Intuitivement, moins on a d'observations expérimentales et / ou plus elles sont bruitées, plus l'incertitude sur les paramètres (et, par propagation, sur la sortie) du modèle sera importante. Quantifier les incertitudes est primordial pour s'assurer de la confiance des prévisions du modèle. En théorie fréquentiste, plusieurs méthodes existent pour cela, dont par exemple les techniques de bootstrap (voir [140]) qui se basent sur du ré-échantillonnage de données. L'approche Bayésienne, quant à elle, permet naturellement de quantifier les incertitudes grâce à la distribution *a posteriori*.

4.5 Sélection et validation de modèles

Les modèles mathématiques permettent de décrire le comportement d'un système biologique à partir de différentes hypothèses. Différentes hypothèses conduisent ainsi à différents modèles. Sélectionner le meilleur modèle peut alors permettre de discriminer certaines hypothèses. Pour cela, plusieurs critères existent, par exemple le critère d'information d'Akaike (AIC) [141] ou le critère d'information Bayésien (BIC) [142] qu'on retrouve couramment. Sans rentrer dans les

détails, il s'agit de faire un compromis entre la qualité de l'ajustement du modèle aux données (donc minimiser la distance entre les observations et la sortie du modèle, ou maximiser la vraisemblance) et le nombre de paramètres à estimer. Le principe (correspondant au rasoir d'Occam) est – lorsque l'on a plusieurs modèles – de sélectionner celui qui s'adapte le mieux aux données tout en étant parcimonieux. Les modèles avec trop de paramètres (trop de degrés de liberté) étant plus susceptibles de s'adapter aux données (on parle parfois d'overfitting), ils doivent être pénalisés. Nous illustrerons l'emploi de critères de sélection de modèle aux chapitres 3 et 7.
Avec l'utilisation de ces critères, on peut alors sélectionner le meilleur modèle parmi ceux étudiés. Il reste alors la question de savoir si ce modèle est un bon modèle ; c'est-à-dire l'étape de validation des modèles. Mathématiquement, un bon modèle doit remplir certains pré-requis : être identifiable, parcimonieux, capable d'avoir généré les données, mais aussi de prédire des observations futures. Ainsi, une fois qu'un modèle est construit, calibré, ses capacités de prévisions doivent être évaluées à partir d'observations n'ayant pas été utilisées pour l'estimation des paramètres (par exemple des observations d'une cohorte de contrôle). Pour cela, on peut par exemple utiliser les observations d'un individu avant un temps T pour calibrer le modèle, puis utiliser les observations à $t > T$ pour mesurer l'erreur de prédiction (mean squared prediction error). Cette étape permet de s'assurer que le modèle est utilisable à des fins cliniques ; que ses prédictions ne seront pas erronées (Fig. 10-D). Nous illustrerons ce point au chapitre 8. Enfin, pour s'assurer qu'il décrit correctement le processus biologique modélisé, il restera à valider expérimentalement le modèle, en mettant au point par exemple de nouvelles expériences.

5 Au programme de ce mémoire

Les travaux que je vais présenter dans ce mémoire de thèse portent sur la modélisation de l'hématopoïèse de façon générale, avec l'accent mis sur la modélisation des néoplasmes myéloprolifératifs et l'effet de l'Interféron α sur ces hémopathies malignes. L'objectif de cette thèse est de mieux comprendre les dynamiques des cellules souches mutées dans les NMP lors de leur développement ou de leur traitement afin d'améliorer les stratégies curatives, de prédire l'effet de l'IFNα à long terme chez les patients pour enfin pouvoir optimiser les doses à administrer pour chacun d'entre eux. La réalisation de ces objectifs passe par la mise en place de différents modèles et différentes analyses que nous présenterons au fur et à mesure des chapitres.

L'hématopoïèse décrit un processus impliquant avant toutes choses un ensemble de cellules hématopoïétiques. C'est donc tout naturellement que nous commencerons ce mémoire par décrire ces dernières. En particulier, nous explorerons la question de la représentation de ces cellules comme un continuum d'états. Bien que dans la suite de ce travail, nous répartissons les cellules hématopoïétiques suivant un ensemble discret de types cellulaires et non selon un continuum, ce chapitre 2 permettra au lecteur de garder en mémoire qu'il s'agira bien là d'une première hypothèse simplificatrice. Pour explorer ce continuum d'états, nous nous baserons sur des mesures expérimentales effectuées par cytométrie de masse sur des cellules hématopoïétiques de souris. Nous nous focaliserons également ici sur la lignée mégacaryocytaire, et appliquerons notre méthode à l'étude comparée des mutations $CALR^m$ de type T1 *vs* T2, par rapport aux souris contrôle (wild-type).

Après avoir discuté la façon de représenter et caractériser les cellules hématopoïétiques, nous commencerons alors à en étudier la dynamique, dans le cas non pathologique, en mettant l'accent sur les cellules souches et progénitrices. Au chapitre 3, nous nous intéresserons à la modélisation du temps nécessaire à ces cellules pour se diviser. Nous étudierons plusieurs distributions de probabilité candidates, parmi lesquelles nous en sélectionnerons la plus adaptée.
Nous approfondirons ensuite la dynamique de prolifération et de différenciation de ces cellules progénitrices au chapitre 4. Nous proposerons un modèle stochastique à temps continu et à espace d'états discret, et présenterons les différentes étapes de modélisation, de la construction du modèle sur des hypothèses biologiques à l'estimation de ses paramètres par l'algorithme CMA-ES

en passant par une étape d'analyse de sensibilité.

Lorsque les cellules souches prolifèrent, des anomalies génétiques peuvent survenir. Celles qui nous intéressent dans ce travail sont celles qui conduisent au développement des NMP. A partir du chapitre 5, nous nous placerons ainsi dans le cas pathologique, en commençant par étudier la question de l'apparition et du développement des NMP. Dans ce chapitre, nous proposons un modèle décrivant l'envahissement du compartiment souche par des cellules mutées, soit au gène *CALR*, soit au gène *JAK2*. Ce modèle, décrit formellement par un CTMC, sera calibré par une méthode ABC-SMC. Nous détaillerons la façons dont nous avons démontré la robustesse de nos résultats conduisant à une proposition d'application au dépistage précoce de ces hémopathies. Ce travail présenté au chapitre 5 a fait l'objet d'une publication (PNAS 2022) [105].

En l'absence de dépistage précoce, les NMP se développent jusqu'à l'apparition de symptômes, puis la prise en charge clinique des patients. Parmi les traitements possibles, l'IFNα a démontré une certaine efficacité, à la fois dans la réponse hématologique et moléculaire. Son action au niveau des cellules souches reste cependant encore mal compris, et les préconisations quant à la dose à utiliser - forte *vs* faible - éventuellement en fonction de certains critères - $CALR^m$ *vs* $JAK2^{V617F}$ par exemple - sont rares. À partir d'observations expérimentales, d'une proposition de modèle mathématique et de l'estimation de ses paramètres par une méthode Bayésienne hiérarchique, nous avons proposé une stratification des patients suivant certains critères, des préconisations quant au niveau de dose à administrer ainsi que des pistes quant au mécanisme d'action de l'IFNα. Ce travail présenté au chapitre 6 a fait l'objet d'une publication (Blood 2021) [122].
Pour étudier plus en détail l'impact des variations de posologie au cours du traitement, nous avons alors étendu le travail précédent. À partir d'une procédure de sélection de modèles, nous proposons au chapitre 7 un nouveau modèle permettant de mieux décrire les dynamiques expérimentales, et à partir duquel nous pouvons estimer pour chaque patient une dose minimale de traitement en-dessous de laquelle il ne faudrait pas descendre. Ce travail a été soumis pour publication et déposé sur un serveur pre-print [123].

Enfin, nous présentons au chapitre 8 comment les développements précédents peuvent s'intégrer dans une application destinée au clinicien, lui permettant une prédiction de l'effet à long terme du traitement, et fournissant un outil d'aide à la décision quant à la posologie optimale à administrer.

Le chapitre 9 fournit une synthèse des résultats obtenus et des perspectives pour continuer ce travail. En particulier, nous présenterons comment nous envisageons à plus long terme de combiner les différents modèles et résultats obtenus dans les différents chapitres de ce mémoire pour une compréhension plus globale de l'effet de l'IFNα sur les cellules souches et, *de facto*, une amélioration de la prise en charge des patients.

Références

[1] Théo Estienne, Maria Vakalopoulou, Enzo Battistella, Alexandre Carré, Théophraste Henry, Marvin Lerousseau, Charlotte Robert, Nikos Paragios, and Eric Deutsch. Deep learning based registration using spatial gradients and noisy segmentation labels. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 87–93. Springer, 2020.

[2] Mahmoud Bentriou, Stefania Chounta, Rodrigue Allodji, Sarah Lemler, Duyen Thi Do, Florent de Vathaire, Ibrahima Diallo, Stergios Christodoulidis, Maria Vakalopoulou, and Veronique Letort Le Chevalier. Deep learning based representations for cardiac voxelised dosimetric data from childhood cancer therapy., 2022.

[3] Enzo Battistella, Maria Vakalopoulou, Théo Estienne, Marvin Lerousseau, Roger Sun, Charlotte Robert, Nikos Paragios, and Eric Deutsch. Gene expression high-dimensional clustering towards a novel, robust, clinically relevant and highly compact cancer signature. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 462–474. Springer, 2019.

[4] Martin Charachon, Céline Hudelot, Paul-Henry Cournède, Camille Ruppli, and Roberto Ardon. Combining similarity and adversarial learning to generate visual explanation : Application to medical image classification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7188–7195. IEEE, 2021.

[5] Martin Charachon, Paul-Henry Cournède, Céline Hudelot, and Roberto Ardon. Visual explanation by unifying adversarial generation and feature importance attributions. In *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data*, pages 44–55. Springer, 2021.

[6] Jason Cosgrove, Lucie SP Hustin, Rob J de Boer, and Leïla Perié. Hematopoiesis in numbers. *Trends in Immunology*, 42(12) :1100–1112, 2021.

[7] Edward A Copelan. Hematopoietic stem-cell transplantation. *New England Journal of Medicine*, 354(17) :1813–1826, 2006.

[8] Connie J Eaves. Hematopoietic stem cells : concepts, definitions, and the new reality. *Blood, The Journal of the American Society of Hematology*, 125(17) :2605–2613, 2015.

[9] LO Jacobson, EL Simmons, EK Marks, and JH Eldredge. Recovery from radiation injury. *Science*, 113(2940) :510–511, 1951.

[10] Julien Picot, Coralie L Guerin, Caroline Le Van Kim, and Chantal M Boulanger. Flow cytometry : retrospective, fundamentals and recent instrumentation. *Cytotechnology*, 64(2) :109–130, 2012.

[11] L Alexander Liggett and Vijay G Sankaran. Unraveling hematopoiesis through the lens of genomics. *Cell*, 182(6) :1384–1400, 2020.

[12] Elisa Laurenti and Berthold Göttgens. From haematopoietic stem cells to complex differentiation landscapes. *Nature*, 553(7689) :418–426, 2018.

[13] Serena Scala and Alessandro Aiuti. In vivo dynamics of human hematopoietic stem cells : novel concepts and future directions. *Blood advances*, 3(12) :1916–1924, 2019.

[14] Ron Sender and Ron Milo. The distribution of cellular turnover in the human body. *Nature medicine*, 27(1) :45–48, 2021.

[15] Henry Lee-Six, Nina Friesgaard Øbro, Mairi S Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J Osborne, Brian JP Huntly, Inigo Martincorena, Elizabeth Anderson, et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561(7724) :473–478, 2018.

[16] Arjun Raj and Alexander Van Oudenaarden. Nature, nurture, or chance : stochastic gene expression and its consequences. *Cell*, 135(2) :216–226, 2008.

[17] Mads Kaern, Timothy C Elston, William J Blake, and James J Collins. Stochasticity in gene expression : from theories to phenotypes. *Nature Reviews Genetics*, 6(6) :451–464, 2005.

[18] Lorraine Robb. Cytokine receptors and hematopoietic differentiation. *Oncogene*, 26(47) :6715–6723, 2007.

[19] Katja Fiedler and Cornelia Brunner. Mechanisms controlling hematopoiesis. In *Hematology-Science and Practice*, pages 3–28. IntechOpen London, 2012.

[20] HT Hassan and AR Zander. Stem cell factor as a survival and growth factor in human normal and malignant hematopoiesis. *Molecular Biology of Hematopoiesis 5*, pages 549–558, 1996.

[21] Cheng C Zhang and Harvey F Lodish. Cytokines regulating hematopoietic stem cell function. *Current opinion in hematology*, 15(4) :307, 2008.

[22] Matthieu Mosca, Gaëlle Vertenoeil, Katte Rao Toppaldoddi, Isabelle Plo, and William Vainchenker. Aspects biologiques de la voie jak/stat dans les néoplasmes myéloprolifératifs classiques négatifs pour bcr-abl. *Bulletin du Cancer*, 103(6) :S16–S28, 2016.

[23] Kunihiro Yamaoka, Pipsa Saharinen, Marko Pesu, Vance ET Holt, Olli Silvennoinen, and John J O'Shea. The janus kinases (jaks). *Genome biology*, 5(12) :1–6, 2004.

[24] David S Aaronson and Curt M Horvath. A road map for those who don't know jak-stat. *Science*, 296(5573) :1653–1655, 2002.

[25] Florence Pasquier. *Identification et fonction de nouvelles mutations des récepteurs à la thrombopoïétine et à l'érythropoïétine dans les néoplasmes myéloprolifératifs et les érythrocytoses.* PhD thesis, Université Paris Saclay (COmUE), 2015.

[26] Shashidhar S Jatiani, Stacey J Baker, Lewis R Silverman, and E Premkumar Reddy. Jak/stat pathways in cytokine signaling and myeloproliferative disorders : approaches for targeted therapies. *Genes & cancer*, 1(10) :979–993, 2010.

[27] Christian Schindler, David E Levy, and Thomas Decker. Jak-stat signaling : from interferons to cytokines. *Journal of Biological Chemistry*, 282(28) :20059–20063, 2007.

[28] A Tefferi and JW Vardiman. Classification and diagnosis of myeloproliferative neoplasms : the 2008 world health organization criteria and point-of-care diagnostic algorithms. *leukemia*, 22(1) :14–22, 2008.

[29] Jean-Jacques Kiladjian, Sylvie Chevret, Jean-Fran¸ois Abgrall, Yasmine Chait, and Jean Briere. Risk stratification for survival and clonal progression in essential thrombocythemia (et) : Result of a prospective study of 108 patients with very long term follow up. *Blood*, 112(11) :1747, 2008.

[30] Elisa Rumi, Daniela Pietra, Virginia Ferretti, Thorsten Klampfl, Ashot S Harutyunyan, Jelena D Milosevic, Nicole CC Them, Tiina Berg, Chiara Elena, Ilaria C Casetti, et al. Jak2 or calr mutation status defines subtypes of essential thrombocythemia with substantially different clinical course and outcomes. *Blood, The Journal of the American Society of Hematology*, 123(10) :1544–1551, 2014.

[31] Chloé James, Valérie Ugo, Jean-Pierre Le Couédic, Judith Staerk, François Delhommeau, Catherine Lacout, Loïc Garçon, Hana Raslova, Roland Berger, Annelise Bennaceur-Griscelli, et al. A unique clonal jak2 mutation leading to constitutive signalling causes polycythaemia vera. *nature*, 434(7037) :1144–1148, 2005.

[32] Yana Pikman, Benjamin H Lee, Thomas Mercher, Elizabeth McDowell, Benjamin L Ebert, Maricel Gozo, Adam Cuker, Gerlinde Wernig, Sandra Moore, Ilene Galinsky, et al. Mplw515l is a novel somatic activating mutation in myelofibrosis with myeloid metaplasia. *PLoS medicine*, 3(7) :e270, 2006.

[33] Jyoti Nangalia, Charles E Massie, E Joanna Baxter, Francesca L Nice, Gunes Gundem, David C Wedge, Edward Avezov, Juan Li, Karoline Kollmann, David G Kent, et al. Somatic calr mutations in myeloproliferative neoplasms with nonmutated jak2. *New England Journal of Medicine*, 369(25) :2391–2405, 2013.

[34] Thorsten Klampfl, Heinz Gisslinger, Ashot S Harutyunyan, Harini Nivarthi, Elisa Rumi, Jelena D Milosevic, Nicole CC Them, Tiina Berg, Bettina Gisslinger, Daniela Pietra, et al. Somatic mutations of calreticulin in myeloproliferative neoplasms. *New England Journal of Medicine*, 369(25) :2379–2390, 2013.

[35] William Vainchenker and Robert Kralovics. Genetic basis and molecular pathophysiology of classical myeloproliferative neoplasms. *Blood, The Journal of the American Society of Hematology*, 129(6) :667–679, 2017.

[36] M Rohrbacher, Ursula Berger, A Hochhaus, G Metzgeroth, K Adam, T Lahaye, Susanne Saußele, MC Müller, Joerg Hasford, Hermann Heimpel, et al. Clinical trials underestimate the age of chronic myeloid leukemia (cml) patients. incidence and median age of ph/bcr-abl-positive cml and other chronic myeloproliferative disorders in a representative area in germany. *Leukemia*, 23(3) :602–604, 2009.

[37] Jyotsna Mehta, Hongwei Wang, Sheikh Usman Iqbal, and Ruben Mesa. Epidemiology of myeloproliferative neoplasms in the united states. *Leukemia & lymphoma*, 55(3) :595–600, 2014.

[38] Glen J Titmarsh, Andrew S Duncombe, Mary Frances McMullin, Michael O'Rorke, Ruben Mesa, Frank De Vocht, Sarah Horan, Lin Fritschi, Mike Clarke, and Lesley A Anderson. How common are myeloproliferative neoplasms ? a systematic review and meta-analysis. *American journal of hematology*, 89(6) :581–587, 2014.

[39] O Moulard, J Mehta, J Fryzek, R Olivares, U Iqbal, and RA Mesa. Epidemiology of myelfibrosis (mf),essential thrombocythemia (et), and polycythemia vera (pv) in the european union (eu). *Eur J Haematol*, 2013.

[40] Sabrina Cordua, Lasse Kjaer, Vibe Skov, Niels Pallisgaard, Hans C Hasselbalch, and Christina Ellervik. Prevalence and phenotypes of jak2 v617f and calreticulin mutations in a danish general population. *Blood, The Journal of the American Society of Hematology*, 134(5) :469–479, 2019.

[41] Debra Van Egeren, Javier Escabi, Maximilian Nguyen, Shichen Liu, Christopher R Reilly, Sachin Patel, Baransel Kamaz, Maria Kalyva, Daniel J DeAngelo, Ilene Galinsky, et al. Reconstructing the lineage histories and differentiation trajectories of individual cancer cells in myeloproliferative neoplasms. *Cell Stem Cell*, 28(3) :514–523, 2021.

[42] Caroline J Watson, AL Papula, Gladys YP Poon, Wing H Wong, Andrew L Young, Todd E Druley, Daniel S Fisher, and Jamie R Blundell. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science*, 367(6485) :1449–1454, 2020.

[43] Nicholas Williams, Joe Lee, Emily Mitchell, Luiza Moore, E Joanna Baxter, James Hewinson, Kevin J Dawson, Andrew Menzies, Anna L Godfrey, Anthony R Green, et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature*, pages 1–7, 2022.

[44] P Hirsch, AC Mamez, R Belhocine, S Lapusan, R Tang, L Suner, D Bories, C Marzac, F Fava, O Legrand, et al. Clonal history of a cord blood donor cell leukemia with prenatal somatic jak2 v617f mutation. *Leukemia*, 30(8) :1756–1759, 2016.

[45] Nikolaos Sousos, Máire Ní Leathlobhair, Christina Simoglou Karali, Eleni Louka, Nicola Bienz, Daniel Royston, Sally-Ann Clark, Angela Hamblin, Kieran Howard, Vikram Mathews, et al. In utero origin of myelofibrosis presenting in adult monozygotic twins. *Nature Medicine*, pages 1–5, 2022.

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

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

[48] Rajintha M Bandaranayake, Daniela Ungureanu, Yibing Shan, David E Shaw, Olli Silvennoinen, and Stevan R Hubbard. Crystal structures of the jak2 pseudokinase domain and the pathogenic mutant v617f. *Nature structural & molecular biology*, 19(8) :754–759, 2012.

[49] Animesh D Pardanani, Ross L Levine, Terra Lasho, Yana Pikman, Ruben A Mesa, Martha Wadleigh, David P Steensma, Michelle A Elliott, Alexandra P Wolanskyj, William J Hogan, et al. Mpl 515 mutations in myeloproliferative and other myeloid disorders : a study of 1182 patients. *Blood*, 108(10) :3472–3476, 2006.

[50] Caroline Marty, Christian Pecquet, Harini Nivarthi, Mira El-Khoury, Ilyas Chachoua, Micheline Tulliez, Jean-Luc Villeval, Hana Raslova, Robert Kralovics, Stefan N Constantinescu, et al. Calreticulin mutants in mice induce an mpl-dependent thrombocytosis with frequent progression to myelofibrosis. *Blood, The Journal of the American Society of Hematology*, 127(10) :1317–1324, 2016.

[51] Ilyas Chachoua, Christian Pecquet, Mira El-Khoury, Harini Nivarthi, Roxana-Irina Albu, Caroline Marty, Vitalina Gryshkova, Jean-Philippe Defour, Gaëlle Vertenoeil, Anna Ngo, et al. Thrombopoietin receptor activation by myeloproliferative neoplasm associated calreticulin mutants. *Blood, The Journal of the American Society of Hematology*, 127(10) :1325–1335, 2016.

[52] Harini Nivarthi, Doris Chen, Ciara Cleary, Blanka Kubesova, Roland Jäger, Edith Bogner, Caroline Marty, Christian Pecquet, William Vainchenker, Stefan N Constantinescu, et al. Thrombopoietin receptor is required for the oncogenic function of calr mutants. *Leukemia*, 30(8) :1759–1763, 2016.

[53] Florence Pasquier, Xenia Cabagnols, Lise Secardin, Isabelle Plo, and William Vainchenker. Myeloproliferative neoplasms : Jak2 signaling pathway as a central target for therapy. *Clinical Lymphoma Myeloma and Leukemia*, 14 :S23–S35, 2014.

[54] Mamatha M Reddy, Anagha Deshpande, and Martin Sattler. Targeting jak2 in the therapy of myeloproliferative neoplasms. *Expert opinion on therapeutic targets*, 16(3) :313–324, 2012.

[55] Ruben A Mesa, Uma Yasothan, and Peter Kirkpatrick. Ruxolitinib. *Nature reviews Drug discovery*, 11(2) :103–105, 2012.

[56] Claire Harrison, Jean-Jacques Kiladjian, Haifa Kathrin Al-Ali, Heinz Gisslinger, Roger Waltzman, Viktoriya Stalbovskaya, Mari McQuitty, Deborah S Hunter, Richard Levy, Laurent Knoops, et al. Jak inhibition with ruxolitinib versus best available therapy for myelofibrosis. *New England Journal of Medicine*, 366(9) :787–798, 2012.

[57] Hannah A Blair. Fedratinib : first approval. *Drugs*, 79(15) :1719–1725, 2019.

[58] Ann Mullally, John Hood, Claire Harrison, and Ruben Mesa. Fedratinib in myelofibrosis. *Blood advances*, 4(8) :1792–1800, 2020.

[59] Jean-Jacques Kiladjian, Bruno Cassinat, Sylvie Chevret, Pascal Turlure, Nathalie Cambier, Murielle Roussel, Sylvia Bellucci, Bernard Grandchamp, Christine Chomienne, and Pierre Fenaux. Pegylated interferon-alfa-2a induces complete hematologic and molecular responses with low toxicity in polycythemia vera. *Blood, The Journal of the American Society of Hematology*, 112(8) :3065–3072, 2008.

[60] Alfonso Quintás-Cardama, Omar Abdel-Wahab, Taghi Manshouri, Outi Kilpivaara, Jorge Cortes, Anne-Laure Roupie, Su-Jiang Zhang, David Harris, Zeev Estrov, Hagop Kantarjian, et al. Molecular analysis of patients with polycythemia vera or essential thrombocythemia receiving pegylated interferon α-2a. *Blood, The Journal of the American Society of Hematology*, 122(6) :893–901, 2013.

[61] Thomas Stauffer Larsen, Katrine F Iversen, Esben Hansen, Anders Bruun Mathiasen, Claus Marcher, Mikael Frederiksen, Herdis Larsen, Inge Helleberg, Caroline Hasselbalch Riley, Ole W Bjerrum, et al. Long term molecular responses in a cohort of danish patients with essential thrombocythemia, polycythemia vera and myelofibrosis treated with recombinant interferon alpha. *Leukemia research*, 37(9) :1041–1045, 2013.

[62] Lucia Masarova, C Cameron Yin, Jorge E Cortes, Marina Konopleva, Gautam Borthakur, Kate J Newberry, Hagop M Kantarjian, Carlos E Bueso-Ramos, and Srdan Verstovsek. Histomorphological responses after therapy with pegylated interferon α-2a in patients with essential thrombocythemia (et) and polycythemia vera (pv). *Experimental hematology & oncology*, 6(1) :1–13, 2017.

[63] RichardT Silver. Recombinant interferon-alpha for treatment of polycythaemia vera. *The Lancet*, 332(8607) :403, 1988.

[64] Abdulraheem Yacoub, John Mascarenhas, Heidi Kosiorek, Josef T Prchal, Dmitry Berenzon, Maria R Baer, Ellen Ritchie, Richard T Silver, Craig Kessler, Elliott Winton, et al. Pegylated interferon alfa-2a for polycythemia vera or essential thrombocythemia resistant or intolerant to hydroxyurea. *Blood*, 134(18) :1498–1509, 2019.

[65] Heinz Gisslinger, Christoph Klade, Pencho Georgiev, Dorota Krochmalczyk, Liana Gercheva-Kyuchukova, Miklos Egyed, Viktor Rossiev, Petr Dulicek, Arpad Illes, Halyna Pylypenko, et al. Ropeginterferon alfa-2b versus standard therapy for polycythaemia vera (proud-pv and continuation-pv) : a randomised, non-inferiority, phase 3 trial and its extension study. *The Lancet Haematology*, 7(3) :e196–e208, 2020.

[66] T Barbui, AM Vannucchi, V De Stefano, A Masciulli, A Carobbio, A Ghirardi, F Ciceri, M Bonifacio, A Iurlo, F Palandri, et al. Phase ii randomized clinical trial comparing ropeginterferon versus phlebotomy in low-risk patients with polycythemia vera. results of the pre-planned interim analysis. *Hemasphere*, 4 :2602, 2020.

[67] John Mascarenhas, Heidi E Kosiorek, Josef T Prchal, Alessandro Rambaldi, Dmitriy Berenzon, Abdulraheem Yacoub, Claire N Harrison, Mary Frances McMullin, Alessandro M

Vannucchi, Joanne Ewing, et al. A randomized phase 3 trial of interferon-α vs hydroxyurea in polycythemia vera and essential thrombocythemia. *Blood, The Journal of the American Society of Hematology*, 139(19) :2931–2941, 2022.

[68] Emmanuelle Verger, Bruno Cassinat, Aurélie Chauveau, Christine Dosquet, Stephane Giraudier, Marie-Hélène Schlageter, Jean-Christophe Ianotto, Mohammed A Yassin, Nader Al-Dewik, Serge Carillo, et al. Clinical and molecular response to interferon-α therapy in essential thrombocythemia patients with calr mutations. *Blood, The Journal of the American Society of Hematology*, 126(24) :2585–2591, 2015.

[69] Lasse Kjær, Sabrina Cordua, Morten O Holmström, Mads Thomassen, Torben A Kruse, Niels Pallisgaard, Thomas S Larsen, Karin De Stricker, Vibe Skov, and Hans C Hasselbalch. Differential dynamics of calr mutant allele burden in myeloproliferative neoplasms during interferon alfa treatment. *PLoS One*, 11(10) :e0165336, 2016.

[70] Julia Czech, Sabrina Cordua, Barbora Weinbergerova, Julian Baumeister, Assja Crepcia, Lijuan Han, Tiago Maié, Ivan G Costa, Bernd Denecke, Angela Maurer, et al. Jak2v617f but not calr mutations confer increased molecular responses to interferon-α via jak1/stat1 activation. *Leukemia*, 33(4) :995–1010, 2019.

[71] Francis E Lotrich, Mordechai Rabinovitz, Patricia Gironda, and Bruce G Pollock. Depression following pegylated interferon-alpha : characteristics and vulnerability. *Journal of psychosomatic research*, 63(2) :131–135, 2007.

[72] Peter C Trask, Peg Esper, Michelle Riba, and Bruce Redman. Psychiatric side effects of interferon therapy : prevalence, proposed mechanisms, and future directions. *Journal of Clinical Oncology*, 18(11) :2316–2326, 2000.

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

[74] Ann Mullally, Claudia Bruedigam, Luke Poveromo, Florian H Heidel, Amy Purdon, Therese Vu, Rebecca Austin, Dirk Heckl, Lawrence J Breyfogle, Catherine Paine Kuhn, et al. Depletion of jak2v617f myeloproliferative neoplasm-propagating stem cells by interferon-α in a murine model of polycythemia vera. *Blood, The Journal of the American Society of Hematology*, 121(18) :3692–3702, 2013.

[75] Laurent Pujo-Menjouet. Blood cell dynamics : half of a century of modelling. *Mathematical Modelling of Natural Phenomena*, 11(1) :92–115, 2016.

[76] Eric A King-Smith and Alec Morley. Computer simulation of granulopoiesis : normal and impaired granulopoiesis. *Blood*, 36(2) :254–262, 1970.

[77] Michael C Mackey and Leon Glass. Oscillation and chaos in physiological control systems. *Science*, 197(4300) :287–289, 1977.

[78] Fabien Crauste. Delay model of hematopoietic stem cell dynamics : asymptotic stability and stability switch. *Mathematical Modelling of Natural Phenomena*, 4(2) :28–47, 2009.

[79] Mostafa Adimy, Fabien Crauste, My Lhassan Hbid, and Redouane Qesmi. Stability and hopf bifurcation for a cell population model with state-dependent delay. *SIAM Journal on Applied Mathematics*, 70(5) :1611–1633, 2010.

[80] Jacques Bélair, Michael C Mackey, and Joseph M Mahaffy. Age-structured and two-delay models for erythropoiesis. *Mathematical biosciences*, 128(1-2) :317–346, 1995.

[81] Mostafa Adimy, Abdennasser Chekroun, and Tarik-Mohamed Touaoula. Age-structured and delay differential-difference model of hematopoietic stem cell dynamics. *Discrete and Continuous Dynamical Systems-Series B*, 20(9) :27, 2015.

[82] Nathalie Eymard, Nikolai Bessonov, Olivier Gandrillon, MJ Koury, and Vitaly Volpert. The role of spatial organization of cells in erythropoiesis. *Journal of mathematical biology*, 70(1) :71–97, 2015.

[83] Anna Marciniak-Czochra, Thomas Stiehl, Anthony D Ho, Willi Jäger, and Wolfgang Wagner. Modeling of asymmetric cell division in hematopoietic stem cells—regulation of self-renewal is essential for efficient repopulation. *Stem cells and development*, 18(3) :377–386, 2009.

[84] Loïs Boullu, Laurent Pujo-Menjouet, and Jianhong Wu. A model for megakaryopoiesis with state-dependent delay. *SIAM Journal on Applied Mathematics*, 79(4) :1218–1243, 2019.

[85] Fabien Crauste, Laurent Pujo-Menjouet, Stéphane Génieys, Clément Molina, and Olivier Gandrillon. Adding self-renewal in committed erythroid progenitors improves the biological relevance of a mathematical model of erythropoiesis. *Journal of theoretical biology*, 250(2) :322–338, 2008.

[86] Salvador Chulián, Álvaro Martínez-Rubio, Anna Marciniak-Czochra, Thomas Stiehl, Cristina Blázquez Goñi, Juan Francisco Rodríguez Gutiérrez, Manuel Ramírez Orellana, Ana Castillo Robleda, Víctor M Pérez-García, and María Rosa. Dynamical properties of feedback signalling in b lymphopoiesis : A mathematical modelling approach. *Journal of Theoretical Biology*, 522 :110685, 2021.

[87] Caroline Colijn and Michael C Mackey. A mathematical model of hematopoiesis—i. periodic chronic myelogenous leukemia. *Journal of Theoretical Biology*, 237(2) :117–132, 2005.

[88] Marek Kimmel. Stochasticity and determinism in models of hematopoiesis. *A systems biology approach to blood*, pages 119–152, 2014.

[89] Sandra N Catlin, Janis L Abkowitz, and Peter Guttorp. Statistical inference in a two-compartment model for hematopoiesis. *Biometrics*, 57(2) :546–553, 2001.

[90] Patrick S Stumpf, Fumio Arai, and Ben D MacArthur. Modeling stem cell fates using non-markov processes. *Cell stem cell*, 28(2) :187–190, 2021.

[91] David Axelrod and Marek Kimmel. *Branching processes in biology*. Springer-Verlag, 2015.

[92] Jason Xu, Samson Koelle, Peter Guttorp, Chuanfeng Wu, Cynthia Dunbar, Janis L Abkowitz, and Vladimir N Minin. Statistical inference for partially observed branching processes with application to cell lineage tracking of in vivo hematopoiesis. *The Annals of Applied Statistics*, 13(4) :2091–2119, 2019.

[93] Alison Etheridge. *Some Mathematical Models from Population Genetics : École D'Été de Probabilités de Saint-Flour XXXIX-2009*, volume 2012. Springer Science & Business Media, 2011.

[94] Patrick Alfred Pierce Moran. Random processes in genetics. In *Mathematical proceedings of the cambridge philosophical society*, volume 54, pages 60–71. Cambridge University Press, 1958.

[95] Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2) :97, 1931.

[96] Ronald A Fisher. Xxi.—on the dominance ratio. *Proceedings of the royal society of Edinburgh*, 42 :321–341, 1923.

[97] Nikolai Bessonov, Laurent Pujo-Menjouet, and Vitaly Volpert. Cell modelling of hematopoiesis. *Mathematical Modelling of Natural Phenomena*, 1(2) :81–103, 2006.

[98] Axel Krinner, Ingo Roeder, Markus Loeffler, and Markus Scholz. Merging concepts-coupling an agent-based model of hematopoietic stem cells with an ode model of granulopoiesis. *BMC systems biology*, 7(1) :1–20, 2013.

[99] Arne Traulsen, Jorge M Pacheco, and David Dingli. On the origin of multiple mutant clones in paroxysmal nocturnal hemoglobinuria. *Stem Cells*, 25(12) :3081–3084, 2007.

[100] Hiroshi Haeno, Ross L Levine, D Gary Gilliland, and Franziska Michor. A progenitor cell origin of myeloid malignancies. *Proceedings of the National Academy of Sciences*, 106(39) :16616–16621, 2009.

[101] Anne-Marie Lyne, Lucie Laplane, and Leïla Perié. To portray clonal evolution in blood cancer, count your stem cells. *Blood, The Journal of the American Society of Hematology*, 137(14) :1862–1870, 2021.

[102] David Dingli and Franziska Michor. Successful therapy must eradicate cancer stem cells. *Stem cells*, 24(12) :2603–2610, 2006.

[103] SA Sisson and Y Fan. Abc samplers. *Handbook of Approximate Bayesian Computation*, pages 87–123, 2018.

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

[105] Gurvan Hermange, Alicia Rakotonirainy, Mahmoud Bentriou, Amandine Tisserand, Mira El-Khoury, François Girodon, Christophe Marzac, William Vainchenker, Isabelle Plo, and Paul-Henry Cournède. Inferring the initiation and development of myeloproliferative neoplasms. *Proceedings of the National Academy of Sciences of the United States of America*, 119(37) :e2120374119, 2022.

[106] Thomas Stiehl and Anna Marciniak-Czochra. Mathematical modeling of leukemogenesis and cancer stem cell dynamics. *Mathematical Modelling of Natural Phenomena*, 7(1) :166–202, 2012.

[107] Helen Moore and Natasha K Li. A mathematical model for chronic myelogenous leukemia (cml) and t cell interaction. *Journal of theoretical biology*, 227(4) :513–523, 2004.

[108] Andrea Saltelli, Marco Ratto, Stefano Tarantola, Francesca Campolongo, et al. Sensitivity analysis practices : Strategies for model-based inference. *Reliability Engineering & System Safety*, 91(10-11) :1109–1125, 2006.

[109] Qiong-Li Wu, Paul-Henry Cournède, and Amélie Mathieu. An efficient computational method for global sensitivity analysis and its application to tree growth modelling. *Reliability Engineering & System Safety*, 107 :35–43, 2012.

[110] Morten Andersen, Zamra Sajid, Rasmus K Pedersen, Johanne Gudmand-Hoeyer, Christina Ellervik, Vibe Skov, Lasse Kjær, Niels Pallisgaard, Torben A Kruse, Mads Thomassen, et al. Mathematical modelling as a proof of concept for mpns as a human inflammation model for cancer development. *PLoS One*, 12(8) :e0183620, 2017.

[111] Céline Bonnet, Panhong Gou, Simon Girel, Vincent Bansaye, Catherine Lacout, Karine Bailly, Marie-Hélène Schlagetter, Evelyne Lauret, Sylvie Méléard, and Stéphane Giraudier. Multistage hematopoietic stem cell regulation in the mouse : A combined biological and mathematical approach. *Iscience*, 24(12) :103399, 2021.

[112] Nikolaus Hansen. The cma evolution strategy : a comparing review. *Towards a new evolutionary computation*, pages 75–102, 2006.

[113] Geoffrey Clapp and Doron Levy. A review of mathematical models for leukemia and lymphoma. *Drug Discovery Today : Disease Models*, 16 :1–6, 2015.

[114] Franziska Michor, Timothy P Hughes, Yoh Iwasa, Susan Branford, Neil P Shah, Charles L Sawyers, and Martin A Nowak. Dynamics of chronic myeloid leukaemia. *Nature*, 435(7046) :1267–1270, 2005.

[115] Ingo Roeder, Matthias Horn, Ingmar Glauche, Andreas Hochhaus, Martin C Mueller, and Markus Loeffler. Dynamic modeling of imatinib-treated chronic myeloid leukemia : functional insights and clinical implications. *Nature medicine*, 12(10) :1181–1184, 2006.

[116] Jasmine Foo, Mark W Drummond, Bayard Clarkson, Tessa Holyoake, and Franziska Michor. Eradication of chronic myeloid leukemia stem cells : a novel mathematical model predicts no therapeutic benefit of adding g-csf to imatinib. *PLoS Computational Biology*, 5(9) :e1000503, 2009.

[117] I Glauche, K Horn, M Horn, L Thielecke, M AG Essers, Andreas Trumpp, and I Roeder. Therapy of chronic myeloid leukaemia can benefit from the activation of stem cells : simulation studies of different treatment combinations. *British journal of cancer*, 106(11) :1742–1752, 2012.

[118] Svetlana Bunimovich-Mendrazitsky, Natalie Kronik, and Vladimir Vainstein. Optimization of interferon–alpha and imatinib combination therapy for chronic myeloid leukemia : A modeling approach. *Advanced Theory and Simulations*, 2(1) :1800081, 2019.

[119] Rasmus K Pedersen, Morten Andersen, Trine A Knudsen, Vibe Skov, Lasse Kjær, Hans C Hasselbalch, and Johnny T Ottesen. Dose-dependent mathematical modeling of interferon-α-treatment for personalized treatment of myeloproliferative neoplasms. *Computational and Systems Oncology*, 1(4) :e1030, 2021.

[120] Johnny T Ottesen, Rasmus K Pedersen, Marc JB Dam, Trine A Knudsen, Vibe Skov, Lasse Kjær, and Morten Andersen. Mathematical modeling of mpns offers understanding and decision support for personalized treatment. *Cancers*, 12(8) :2119, 2020.

[121] Johnny T Ottesen, Rasmus K Pedersen, Zamra Sajid, Johanne Gudmand-Hoeyer, Katrine O Bangsgaard, Vibe Skov, Lasse Kjær, Trine A Knudsen, Niels Pallisgaard, Hans C Hasselbalch, et al. Bridging blood cancers and inflammation : The reduced cancitis model. *Journal of Theoretical Biology*, 465 :90–108, 2019.

[122] Matthieu Mosca, Gurvan Hermange, Amandine Tisserand, Robert Noble, Christophe Marzac, Caroline Marty, Cécile Le Sueur, Hugo Campario, Gaëlle Vertenoeil, Mira El-Khoury, et al. Inferring the dynamics of mutated hematopoietic stem and progenitor cells induced by ifnα in myeloproliferative neoplasms. *Blood, The Journal of the American Society of Hematology*, 138(22) :2231–2243, 2021.

[123] Gurvan Hermange, William Vainchenker, Isabelle Plo, and Paul-Henry Cournède. Mathematical modelling, selection and hierarchical inference to determine the minimal dose in ifn alpha therapy against myeloproliferative neoplasms. *arXiv preprint arXiv :2112.10688*, 2021.

[124] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[125] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260) :583–621, 1952.

[126] Charlotte Baey, Paul-Henry Cournède, and Estelle Kuhn. Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 135 :107–122, 2019.

[127] Marc Lavielle. *Mixed effects models for the population approach : models, tasks, methods and tools*. CRC press, 2014.

[128] A Gelman, JB Carlin, HS Stern, DB Dunson, A Vehtari, and DB Rubin. Bayesian data analysis, 3rd edn.(2013).

[129] Alexander Etz and Joachim Vandekerckhove. Introduction to bayesian inference for psychology. *Psychonomic bulletin & review*, 25(1) :5–34, 2018.

[130] Rink Hoekstra and Simine Vazire. Aspiring to greater intellectual humility in science. *Nature human behaviour*, 5(12) :1602–1607, 2021.

[131] Ronan Duchesne, Anissa Guillemin, Fabien Crauste, and Olivier Gandrillon. Calibration, selection and identifiability analysis of a mathematical model of the in vitro erythropoiesis in normal and perturbed contexts. *In silico biology*, 13(1-2) :55–69, 2019.

[132] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15) :1923–1929, 2009.

[133] Ilya M Sobol. Sensitivity analysis for non-linear mathematical models. *Mathematical modelling and computational experiment*, 1 :407–414, 1993.

[134] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[135] David E Goldberg. Genetic algorithms in search, optimization, and machine learning. addison. *Reading*, 1989.

[136] Maurice Clerc. *L'optimisation par essaims particulaires*. Hermes science publications, 2005.

[137] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598) :671–680, 1983.

[138] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications, 1970.

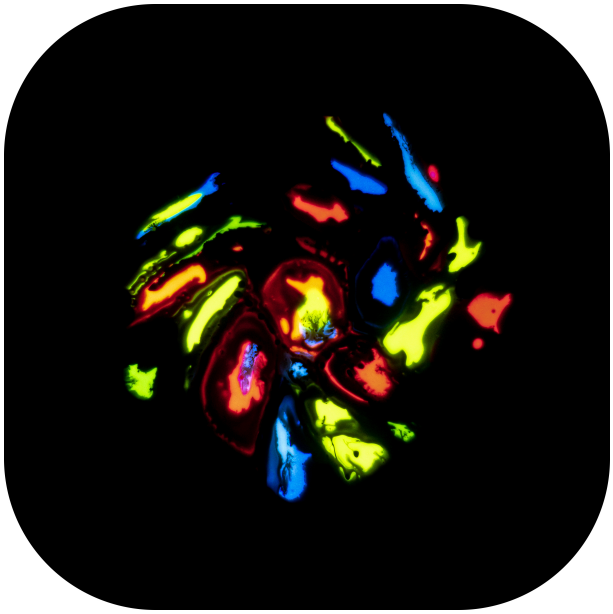
[139] Christophe Andrieu and Johannes Thoms. A tutorial on adaptive mcmc. *Statistics and computing*, 18(4) :343–373, 2008.

[140] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Cambridge university press, 1997.

[141] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6) :716–723, 1974.

[142] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

Chapitre 2
Cellules hématopoïétiques : un continuum d'états révélé par cytométrie de masse



Résumé

Partant de données en cytométrie de masse pour des cellules hématopoïétiques, nous étudions la façon dont celles-ci se distribuent de façon continue dans un espace de dimension 19, qui correspond au nombre de marqueurs de surface et intra-cellulaires considérés. En particulier, nous proposons une méthode empirique d'inférence de trajectoire de différenciation, en se concentrant sur la lignée mégacaryocytaire. Notre méthode se base sur la construction d'un graphe des plus proches voisins, puis son exploration à partir de marches aléatoires. Nous ajoutons des contraintes sur ces dernières, basées sur des *a priori* biologiques, pour parcourir le graphe en passant par des cellules dont le type est de plus en plus différencié. Une fois la trajectoire de différenciation inférée, nous pouvons étudier l'évolution de l'intensité de marqueurs - notamment des marqueurs de signalisation - au cours de celle-ci.
En appliquant alors cette méthode sur des souris mutées $CALR^m$ de type T1 ou T2, après correction d'effets batch, nous pouvons comparer les trajectoires de différenciation en fonction du génotype.

Abstract

It has recently been shown that hematopoiesis would be better described by a process involving a continuum of cells rather than a set of well distinct cell populations.
In this chapter, we explore how hematopoietic cells are continuously distributed. For that purpose, we analyse experimental observations of single-cell data obtained by mass cytometry (CyTOF). We get an expression intensity for each cell for 19 intracellular and surface markers. To study these data, we propose an empirical method of differentiation trajectory inference, focusing on the megakaryocytic lineage. Our method is based on constructing a nearest-neighbour graph and then its exploration using random walks. Based on prior biological knowledge, we add constraints on the latter to traverse the graph through cells of increasingly differentiated types. Once the differentiation trajectory has been inferred, we can study the evolution of the intensity of markers - in particular signalling markers - during this trajectory.
Then, we apply our method to understand how $CALR^m$ mutation (type 1 del52 and type 2 CALR ins5) might impact the differentiation trajectories.

Table des matières

1 Introduction

L'hématopoïèse, nous l'avons vu au chapitre précédent, a longtemps été décrite comme un processus très hiérarchisé, structuré en arbre, impliquant des types de cellules bien distincts. On y place au sommet les cellules souches hématopoïétiques, capables de s'auto-renouveler et de produire tous types de cellules sanguines, par l'intermédiaire de progéniteurs et précurseurs de plus en plus spécialisés (ou différenciés). Cette représentation reste largement utilisée, et justifiée lorsque les cellules souches et progénitrices ne peuvent être caractérisées que par un ensemble limité de marqueurs comme c'est souvent le cas en cytométrie de flux.
Néanmoins, les techniques pour caractériser les cellules - en particulier à l'échelle uni-cellulaire (*single-cell*) - se sont développées ces dernières années. D'une information limitée à quelques marqueurs par cellule, on est passé à des dizaines en cytométrie de masse, voire milliers en single-cell RNA-sequencing (scRNA-seq), ouvrant la voie à des méthodes d'analyses mathématiques plus sophistiquées. Ces méthodes reposent généralement sur l'analyse d'un large échantillon de cellules à un instant donné. Des algorithmes de clustering, tels que le clustering hiérarchique ou le clustering de Louvain [1] par exemple, ont pu être utilisés pour chercher à identifier des nouvelles populations de cellules hématopoïétiques [2, 3, 4], l'utilisation de techniques d'analyse différentielle sur les différents clusters permettant alors de les caractériser plus finement [5, 6]. Un autre domaine d'étude a connu de nombreux développements ces dernières années, celui de l'inférence de trajectoires. L'idée étant, à partir de l'échantillon d'un grand nombre de cellules, à un instant donné, de réussir à reconstruire le processus de différenciation qui a pu conduire aux observations. Ces méthodes se basent sur deux hypothèses principales : premièrement que les cellules du jeu de données soient représentatives du processus de différenciation étudié, et que ce dernier se fasse par des changements graduels dans l'expression des marqueurs. Pour satisfaire cette dernière hypothèse, un nombre important de marqueurs peut être nécessaire, ce qui explique que ces techniques ne sont généralement pas applicables en cytométrie de flux [1] et ne se sont développées que suite à l'augmentation du nombre de marqueurs possible en cytométrie de masse ou en scRNA-seq. Parmi les différents algorithmes d'inférence de trajectoires, l'algorithme Wanderlust développé par Bendall et al. [7] permet, à partir de l'expression d'une vingtaine de marqueurs en cytométrie de masse, de reconstruire la trajectoire de développement d'une lignée hématopoïétique. La méthode de Bendall et al. repose sur la construction de graphes des k plus proches voisins (k-NN graphe - k-Nearest Neighbors) puis le réalignement de toutes les cellules suivant un axe de développement, qui correspond dans leur application à la trajectoire de différenciation de la lignée lymphocytaire (lymphocytes B). L'algorithme Wanderlust nécessite de n'avoir que des cellules appartenant à la lignée considérée. En particulier, il ne peut pas s'appliquer pour décrire un processus de différenciation avec branchement. L'algorithme Whishbone [8] a été développé pour généraliser Wanderlust à plusieurs lignées, et a été appliqué au cas de la différenciation myéloïde chez la souris, à partir de données scRNA-seq. Parmi les autres algorithmes d'inférence de trajectoires, mentionnons STREAM [9], PAGA [10] ou encore SPADE [11].
L'utilisation combinée d'observations single-cell en grande dimension (c'est-à-dire, avec un grand nombre de marqueurs pour chaque cellule) et de méthodes d'analyse mathématiques a alors permis de faire émerger de nouvelles propositions de structures pour l'hématopoïèse, maintenant envisagée plutôt comme un processus continu impliquant un ensemble de cellules hétérogènes [12, 13, 14].

Dans ce contexte, une question nous intéresse particulièrement : l'étude de la lignée mégacaryocytaire et la façon dont évoluent les marqueurs de signalisation de la cellule souche (HSC) au mégacaryocyte. Ou, plus précisément, la trajectoire comparée de ces marqueurs au cours de la différenciation selon que les cellules étudiées sont saines (WT - Wild-Type), mutantes $CALR^m$ de type 1 (T1) ou de type 2 (T2).
Plusieurs méthodes d'inférence de trajectoires ont été développées au cours des dernières années,

1. Les techniques les plus récentes en cytométrie de flux permettent cependant l'utilisation de plus de marqueurs qu'auparavant.

principalement adaptées à des données scRNA-seq, et peu souvent appliquée à la mégacaryopoïèse. Mentionnons Prins et al. [15] qui ont étudié le cas de souris mutées $CALR^m$ à partir de données scRNA-seq, en appliquant l'algorithme PAGA [10], mettant en évidence une trajectoire de différenciation de la cellule souche aux mégacaryocytes et identifiant une nouvelle population cellulaire surreprésentée chez les souris mutées. Dans notre cas, nous baserons notre travail sur des données *single-cell* de souris, obtenues par cytométrie de masse, en se concentrant sur des marqueurs de signalisation et de surface et sur les différences entre les mutations au gène *CALR* de type T1 ou T2. Nous présentons dans la suite une méthode que nous avons développée pour analyser nos données de cytométrie de masse. En plus de la question de recherche précise à laquelle nous cherchons à répondre dans ce chapitre (à savoir : qu'est-ce qui différencie les mutations T1 et T2 en termes de trajectoire de différenciation des marqueurs intracellulaires, de la cellule souche au mégacaryocyte ?), ce dernier a aussi pour objectif - dans le cadre de ce travail de thèse - d'illustrer comment les cellules hématopoïétiques, en particulier celles immatures, se distribuent plutôt suivant un continuum d'états que selon un partitionnement discret. Ce chapitre, placé en début de ce mémoire de thèse, soulèvera ainsi la question de la caractérisation des progéniteurs hématopoïétiques.

2 Observations expérimentales et prétraitement des données

2.1 Expérience

Les expériences ayant permis d'obtenir les données expérimentales sur lesquelles se base la méthode décrite plus bas ont été réalisées par Camélia Benlabiod, alors en thèse sous la direction de Caroline Marty et d'Isabelle Plo. Les expériences et le contexte biologique y sont détaillés dans son manuscrit de thèse [16].
Pour cette première partie, nous ne considérerons que les données d'une seule souris WT (# 206). L'étude de l'impact du génotype muté ($CALR^m$ T1 ou T2) ne sera présentée qu'à la section 4, comme application de notre méthode.
Après sacrifice de la souris WT considérée, les cellules hématopoïétiques issues de la moelle osseuse ont été récoltées et purifiées. À cette étape, on dispose d'environ 3 millions de cellules, la plupart dites Lin^- (lineage negative), i.e. immatures. Plusieurs étapes sont nécessaires avant d'obtenir l'échantillon final, notamment deux étapes de marquage - une première pour les marqueurs de surface, la seconde pour ceux intra-cellulaires - ainsi que plusieurs étapes de lavage. Les étapes de marquages consistent à marquer les cellules avec des anticorps couplés à des métaux lourds. Les anticorps sont choisis pour être spécifiques de certaines protéines d'intérêt, soit des protéines de surface, soit des protéines intra-cellulaires. Enfin, après filtration, on récupère environ 2/3 des cellules de départ. Les métaux lourds fixés aux anticorps permettent leur détection par le cytomètre de masse (qu'on appellera également par la suite CyTOF, pour Cytometry by Time of Flight). Le principe de fonctionnement du CyTOF utilisé par la plateforme de Cytométrie de Gustave Roussy (Fig. 2) est illustré sur la figure 1.

2.2 Pré-traitement des données

Après passage au cytomètre de masse, nous disposons pour chaque cellule d'une mesure d'intensité associée à chaque marqueur du panel (ainsi que certaines informations additionnelles). Pour un marqueur donné, ces intensités donnent une information relative quant à la quantité d'anticorps fixés à la cellule. Les intensités ne sont pas comparables entre différents marqueurs. En fait, pour être plus précis, nous récupérons après passage au CyTOF les mesures associées à ce qu'on appelle, en cytométrie, "évènements". Les données récupérées étant structurées en tableaux, si l'information sur les marqueurs correspond aux colonnes, chaque ligne est alors un évènement. Tout évènement n'est pas nécessairement une cellule, et lorsque c'est le cas, ne correspond pas nécessairement à une cellule d'intérêt. Un évènement va en fait correspondre à une mesure au niveau d'une "*droplet*". Parfois, les droplets sont vides, ou contiennent une bille (permettant la calibration de l'appareil de mesure), voire contiennent plusieurs cellules.

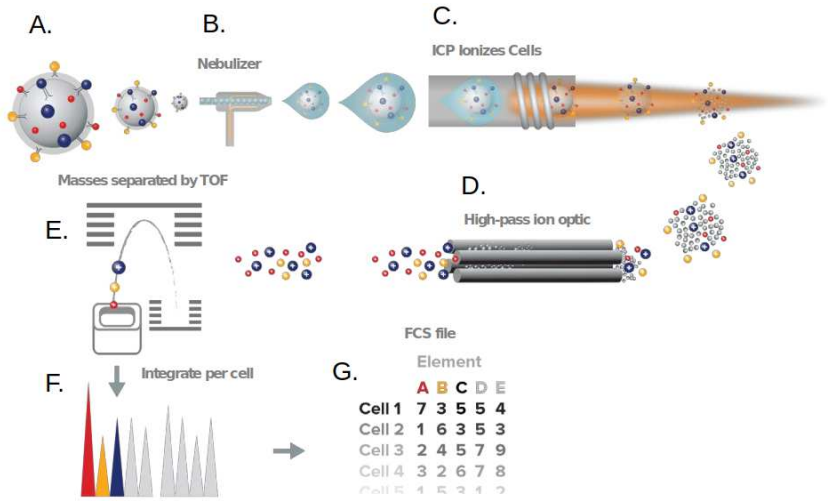


FIGURE 1 – Principe de fonctionnement du CyTOF (schéma provenant de Fluidigm). Les cellules sont marquées avec des anticorps auxquels sont fixés des métaux lourds (A). Elles passent ensuite, à une cellule par droplet, au nébuliseur (B), formant un nuage de particules. Chaque cellule est alors vaporisée (C) et les métaux liés aux anticorps ionisés au moyen d'une torche à plasma (ICP - Inductively coupled plasma). Seuls les ions associés aux métaux lourds sont sélectionnés (D), puis discriminés par leur temps de vols (TOF), qui dépend pour chaque ion de son rapport masse sur charge (E), pour aboutir ensuite à un spectrogramme de masse (F). Les données sont ensuite enregistrées sous forme matricielle (G) $\mathbf{Y} = [y_{i,j}]$ avec en ligne l'information pour chaque droplet et en colonne l'information sur chaque élément (i.e. marqueur).

Un prétraitement est ainsi nécessaire pour exclure tous les évènements indésirables et ne garder que des cellules d'intérêt. Dans ce prétraitement, réalisé par Camélia Benlabiod, on a les étapes suivantes :

- Exclusion Plomb / Barium
- Exclusion des billes (qui sont conçues pour avoir une forte intensité pour les isotopes 142Ce et 140Ce)
- Exclusion des cellules mortes : on ne conserve que des cellules exprimant fortement le marqueur associé à DNA1 et ayant une faible intensité pour le marqueur correspondant à l'isotope 198Pt.
- Exclusion des multiplets. Les singlets sont considérés avoir une "event length", d'après le protocole expérimental, inférieure à 50.
- Exclusion des globules rouges et lymphocytes B, en ne gardant que les cellules n'exprimant pas les marqueurs TER-119 et CD45R (B220)
- Exclusion des monocytes et lymphocytes T, en ne gardant pas les cellules exprimant CD3 et CD11b
- Exclusion des autres cellules Lin+, qui vont exprimer fortement Ly6G/GR1

Nous ajoutons également une étape - d'après le protocole d'analyse de Fluidigm - où nous excluons les évènements ayant une valeur inférieure à 0.01 pour le paramètre "center".
Par abus de langage, on dira indifféremment qu'une cellule (ou évènement) exprime la protéine (c'est-à-dire le marqueur), l'anticorps, ou l'isotope qui lui est associé, lorsque l'intensité mesurée est au-dessus d'un certain seuil (nous y reviendrons au § 4.2.1). En cytométrie, ces seuils sont propres à chaque expérience et généralement définis manuellement en visualisant leur distribution sur l'ensemble des évènements considérés (Figure 3).
La succession d'étapes d'exclusion / inclusion, en ne gardant à chaque fois que les évènements d'intérêt non exclus de l'étape précédente, correspond à ce qu'on appellera une stratégie de *gating*. Les stratégies de *gating* ne sont pas réservées à l'exclusion des évènements non désirés ;

FIGURE 2 – Photo du CyTOF (Helios) utilisé par la plateforme de cytométrie de Gustave Roussy.

elles s'appliquent également pour caractériser des cellules selon la valeur d'intensité de certains marqueurs. Formellement, il s'agit d'un partitionnement des évènements suivant un arbre de décision. Nous y reviendrons dans la section suivante.

Ainsi, après ces différentes étapes de filtration, on considère que nos données correspondent à un ensemble de cellules hématopoïétiques d'intérêt. Nous disposons alors de $N_c = 43,382$ cellules. Notre jeu de donnée est alors défini par la matrice $\mathbf{Y} = [y_{i,j}]$ avec $y_{i,j}$ l'intensité (brute) du marqueur j de la cellule i. Les marqueurs utilisés pour les étapes de filtration précédentes ne sont plus considérés. Le panel d'intérêt, sur lequel va se baser notre méthode, est alors celui du tableau 1.

2.3 Partitionnement

Notre objectif est d'explorer le continuum d'états, formé par les différentes cellules hématopoïétiques, par une méthode d'inférence de trajectoire. Cela ne se fera pas sans *a priori* biologiques ; nous commencerons ainsi par définir une première partition discrète de l'ensemble de nos cellules. Pour cela, nous utiliserons l'information de certains marqueurs de surface, qui sont classiquement utilisés pour affecter les cellules hématopoïétiques à un type donné. Ces marqueurs de surface sont : c-kit, CD41, CD42, Sca-1, CD150, CD48 et CD135. Pour chacun d'eux, nous avons défini des seuils au delà desquels on considère qu'une cellule exprimera le marqueur donné. Ces seuils sont définis ici en regardant la distribution de l'intensité du marqueur considéré, comme illustré dans le cas de CD150 sur la figure 3. Une fois ces seuils définis (voir tableau 1), on peut alors partitionner les cellules suivant différents types cellulaires, à partir de l'arbre de décision de la figure 4.
Cette idée de partitionner les cellules se retrouve dans l'algorithme PAGA (qui signifie d'ailleurs Partition-based graph abstraction) [10]. La méthode de partitionnement que Wolf et al. proposent dans leur article se base sur une méthode non supervisée, en l'occurrence l'utilisation de l'algorithme de Louvain, mais comme ils l'indiquent dans leur article, le partitionnement peut également se faire sur la base d'arguments biologiques, comme nous choisissons de le faire ici.

Les types cellulaires considérés sont ordonnés, du plus immature au plus mature, de la façon suivante :

1. SLAM (qui inclut les cellules souches) [2]
2. MPP (Progéniteurs multipotents)
3. LSK* (L pour Lin^-, S pour Sca-1^+ et K pour c-kit^+)
4. LK*
5. MkP (Progéniteurs Mégacaryocytaires)

2. SLAM signifie "Signaling lymphocyte activation molecule". Les récepteurs de la famille des SLAM incluent plusieurs marqueurs de surface, notamment CD150 et CD48 [17].

Protéine	Localisation	Rôle	Isotope associé à l'AC	Seuil
Ly6A/E (Sca-1)	Surface	Marqueur des HSC et progéniteurs	89 Y	20
CD117 (c-Kit)	Surface	Marqueur des HSC et progéniteurs	173 Yb	134.48
CD48	Surface	Marqueur des HSC et progéniteurs	154 Sm	48.49
CD150	Surface	Marqueur des HSC et progéniteurs	167 Er	10
CD135	Surface	Marqueur des HSC et progéniteurs	150 Nd	2.6
CD41	Surface	Marqueur des MK	143 Nd	54.16
MPL	Surface	Marqueur des MK	168 Er	
CD42d	Surface	Marqueur des MK	151 Eu	22.06
CALR	Surface		176 Yb	
pSTAT5	Intra-cellulaire	Marqueur de signalisation	147 Sm	
pSTAT3	Intra-cellulaire	Marqueur de signalisation	158 Gd	
STAT3	Intra-cellulaire	Marqueur de signalisation	162 Dy	
pSTAT1	Intra-cellulaire	Marqueur de signalisation	153 Eu	
pAKT	Intra-cellulaire	Marqueur de signalisation	152 Sm	
pERK1/2	Intra-cellulaire	Marqueur de signalisation	171 Yb	
JAK2	Intra-cellulaire	Marqueur de signalisation	161 Dy	
pEiF2d	Intra-cellulaire	Stress du réticulum endoplasmique	175 Lu	
CALR	Intra-cellulaire		142 Nd	
KI67	Intra-cellulaire	Marqueur de prolifération cellulaire	172Yb	

TABLE 1 – Panel des $m = 19$ marqueurs de l'analyse CyTOF utilisés pour la construction du k-NN graphe. AC signifie anticorps. Lorsque le marqueur est utilisé pour le partitionnement (voir figure 4), nous indiquons le seuil au-delà duquel le marqueur est considéré comme exprimé (valeur de l'intensité avant application de la transformation argsinh, et dans le cas de la souris WT # 206). Notons que, par rapport au panel initialement utilisé, les marqueurs de surface correspondant à CD45 et CD34 ont été exclus, le marquage n'ayant pas fonctionné dans leur cas.

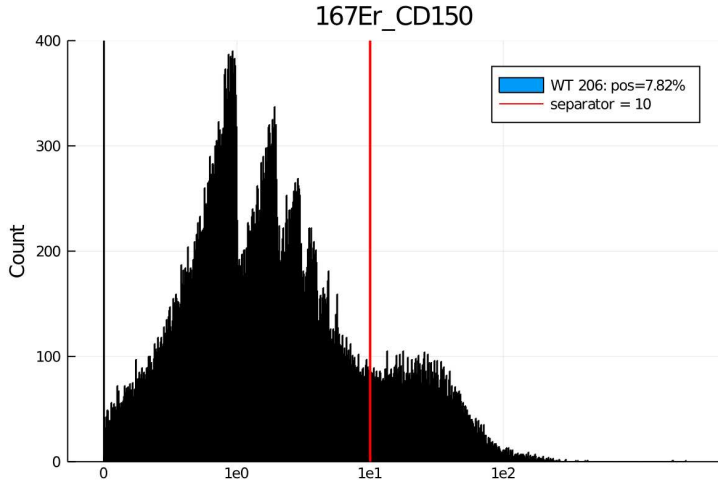


FIGURE 3 – Distribution des intensités (échelle logarithmique en abscisse) associées au marqueur CD150 pour les cellules de la souris WT # 206. On distingue trois zones d'intensité : 1) une intensité nulle qui correspond à une expression nulle du marqueur. Notons qu'ici l'axe des ordonnées est tronqué, mais qu'une forte proportion des cellules (51%) ont une intensité nulle pour ce marqueur ; 2) une intensité faible (à gauche de la ligne verticale rouge) qui correspondrait à une très faible expression ou à du bruit ; 3) une expression du marqueur au delà de la ligne verticale. Le choix du seuil (ici fixé à 10) est adapté en fonction de l'allure de la distribution. 7.82% des cellules sont ainsi considérées comme exprimant le marqueur CD150.

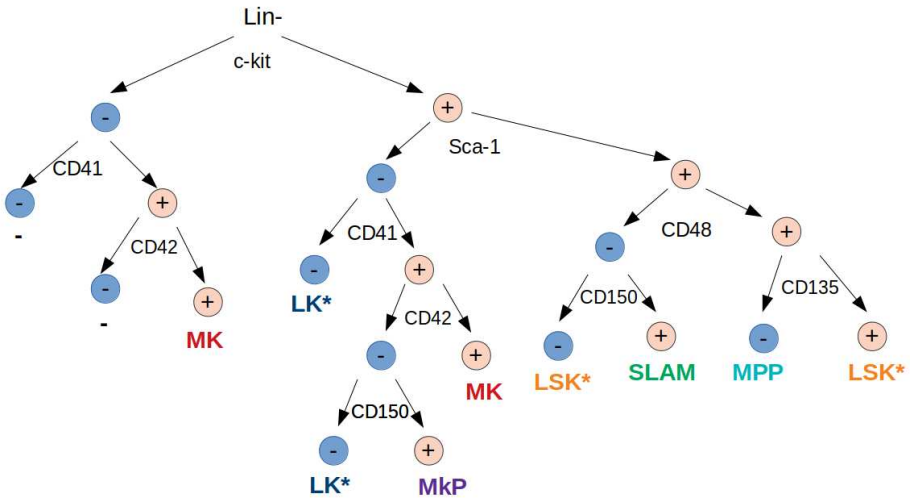


FIGURE 4 – Arbre de décision permettant d'attribuer à toute cellule Lin^- un et un seul type, suivant si l'expression de certains marqueurs est supérieure à un certain seuil (+) ou non (-). À noter que les cellules qui tombent dans la catégorie "-" sont exclues dans la suite.

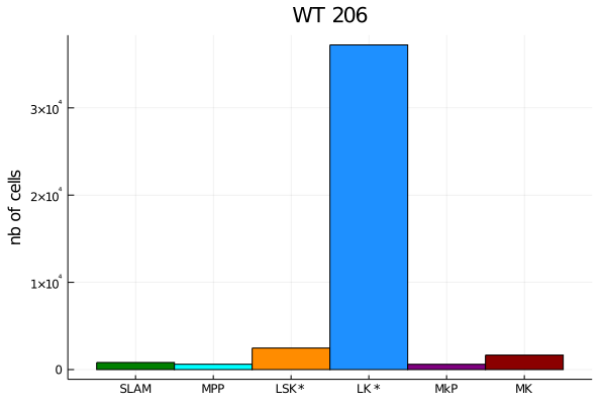


FIGURE 5 – Distribution des cellules de la souris WT (#206) suivant les différents types cellulaires considérés. Au total, le jeu de données est constitué de $N_c = 43,382$ cellules dont une très forte majorité sont des LK*.

6. MK (Mégacaryocytes)

Chaque cellule i est maintenant définie - en plus de l'intensité $\boldsymbol{Y}_i$ pour 9 marqueurs de surface et 10 marqueurs intracellulaires - par un type cellulaire parmi ceux de la liste précédente : $p_i \in \{1, \cdots, 6\}$. On applique la transformation argument sinus hyperbolique (choix standard en cytométrie de flux [18] mais également en cytométrie de masse [19]) aux intensités :

$$\text{argsinh} \ : y \longmapsto \log\left(y + \sqrt{1+y^2}\right) \tag{1}$$

Le jeu de données, après application de la transformation (1), est noté $\mathbf{X} = [x_{i,j}]_{1 \leq i \leq N_c, 1 \leq j \leq 19}$. La répartition des cellules suivant le type cellulaire est présenté sur l'histogramme de la figure 5. Sur la figure 6 nous présentons la distribution de ces cellules suivant une projection UMAP (Uniform Manifold Approximation and Projection) [20]. L'algorithme UMAP est une méthode de projection de données à un espace de plus petite dimension. Contrairement à l'ACP (Analyse en Composantes Principales) [21], la projection ici n'est pas linéaire. Cette technique de réduction de dimension est largement utilisée en bioinformatique pour la visualisation de données - lorsque l'on choisit de projeter sur un espace de dimension 2 - et maintenant privilégiée à t-SNE [22] pour ses performances.

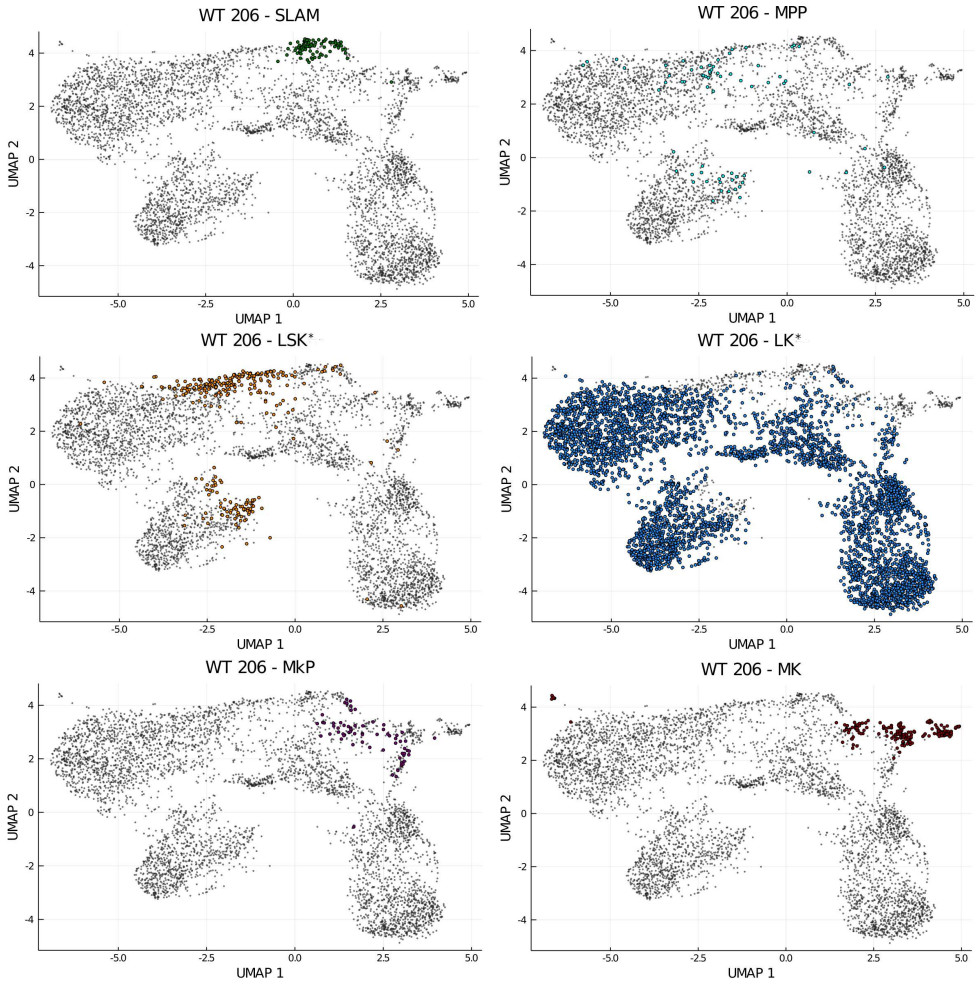


FIGURE 6 – Projection UMAP des cellules de la souris WT #206. La projection est calculée à partir de l'ensemble des cellules (Lin^-) de la souris (et 20 nearest neighbors). Les cellules grises sont un échantillon de 5,000 cellules. On représente sur chaque figure les populations SLAM (vert), MPP (cyan), LSK* (orange), LK* (bleu), MkP (violet) et MK (rouge), l'intégralité des cellules de ces populations étant projetées.

Comme nous pouvons le voir sur la figure 6, projetées en 2D, les cellules semblent bien former un continuum plutôt qu'un ensemble distinct d'états. En d'autres termes, il n'y a pas de limite claire entre différents types cellulaires. Nous allons dans la suite présenter la méthode utilisée pour explorer ce continuum d'états.

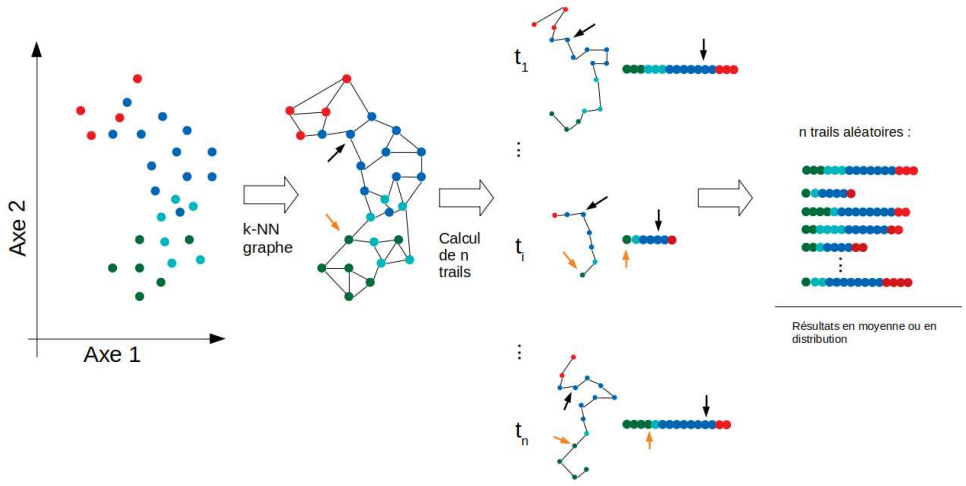


FIGURE 7 – Schéma de la méthode pour l'inférence de trails. À partir de cellules dont on connaît l'intensité d'expression de $m = 19$ marqueurs, ainsi que la répartition en types cellulaires (ici, pour simplifier, 4 types sont représentés : SLAM en vert, MPP en cyan, LK* en bleu et MK en rouge), on construit un graphe des k plus proches voisins (sur l'exemple, $k = 3$). On explore alors ce graphe de manière aléatoire. On calcule un ensemble de n trails, reliant une cellule immature (SLAM) et une cellule mature (MK) tirées au hasard (différents tirages pour chaque trail), en interdisant le retour à un type plus immature. On aboutit à n trails, constitués d'un nombre différent de cellules. Une cellule du jeu de données (par exemple celle pointée par la flèche orange ou noire) peut être présente dans plusieurs trails, à différentes localisations. L'information du dataset, initialement portée par un ensemble de nombreuses cellules, puis structurée en graphe, est ainsi transférée aux n trails qui ont exploré le graphe, suivant des règles issues d'un *a priori* biologique (le passage vers des types cellulaires de plus en plus matures). L'information est ensuite synthétisée en étudiant le comportement moyen des trails.

3 Méthode

3.1 Inférence de trails

Nous disposons, pour la souris WT considérée, d'un ensemble de cellules $\mathbf{X}$, où chaque cellule i est caractérisée par un vecteur $\boldsymbol{X}_i$ correspondant à l'intensité suivant m marqueurs (de surface et intracellulaires). Notre méthode d'inférence de trajectoire va se baser sur la construction d'un graphe à partir de ces cellules - comme le font par exemple les algorithmes Wanderlust ou Wishbone [7, 8] - puis l'exploration de ce graphe à partir de marches aléatoires, approche similaire à celle de Wolf et al. [10]. Nous ajoutons certaines contraintes basées sur un *a priori* biologique pour la construction de ces marches aléatoires, à savoir qu'elles ne peuvent que progresser vers des cellules d'un type cellulaire plus mature, qu'elles doivent commencer par une SLAM et se terminer par un MK. Ces marches aléatoires soumises à contraintes seront appelées "trails". Ces trails sont ainsi construits pour correspondre à la différenciation mégacaryocytaire.

L'algorithme, schématisé sur la figure 7, est le suivant : on part de l'ensemble des cellules $\mathbf{X}$. On commence par normaliser l'intensité en divisant, pour chaque marqueur, par la moyenne des intensités pour le marqueur considéré. On construit alors un graphes des k plus proches voisins (k-NN graph) [23] qu'on va parcourir de façon aléatoire pour inférer n trails. Un trail est calculé de la façon suivante : on commence par choisir au hasard une SLAM, cellule de départ, puis un MK, cellule d'arrivée. On effectue une marche aléatoire, où les voisins des cellules sont explorés de façon aléatoire, avec une probabilité plus forte lorsque la distance les séparant est plus faible. On interdit de repasser deux fois par la même cellule, et on force la progression en interdisant de revenir vers un type plus immature. Toutes les marches n'aboutiront pas. Celles qui aboutissent permettent de constituer un ensemble de n trails. Après exécution de l'algorithme, nous

avons donc n trails, qui consistent en des séquences de cellules, séquences qui n'ont pas la même longueur. Notons qu'une cellule du dataset peut apparaître dans plusieurs trails, et pas nécessairement à la même position.
Avec la règle de progression imposée, on fait alors l'hypothèse que les trails vont, en moyenne, permettre de décrire les étapes de la mégacaryopoïèse. C'est sous cette hypothèse que nous pourrons interpréter les résultats obtenus.

L'algorithme a été implémenté en Julia et le code est disponible sur GitLab au lien suivant :

`https://gitlab-research.centralesupelec.fr/2012hermangeg/cell-trajectory-inference`

Pour la construction des k-NN graphs, nous choisissons $k = 20$ et une distance euclidienne, choix dont nous n'avons cependant pas directement estimé l'influence. Nous exécutons le code jusqu'à l'obtention de $n = 100,000$ trails. Sur la figure 8, on représente quelques exemples de trails.

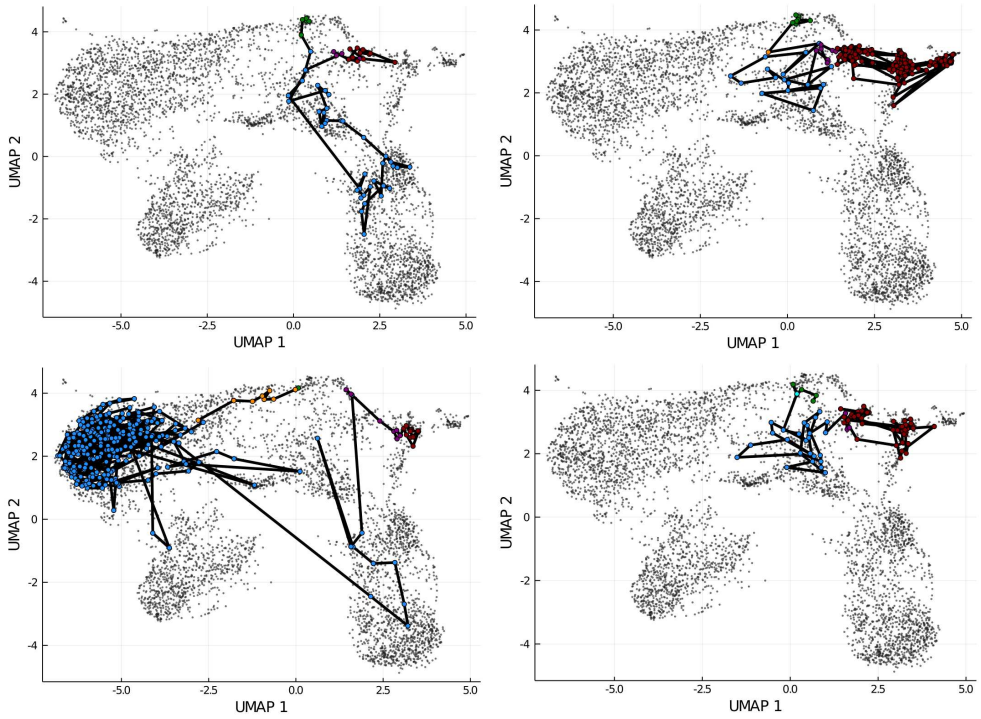


FIGURE 8 – Quatre exemples de trails. Les populations SLAM (vert), MPP (cyan), LSK* (orange), LK* (bleu), MKP (violet) et MK (rouge) ne peuvent être parcourues que dans cet ordre (mais sans qu'elles doivent nécessairement apparaître dans les trails, sauf pour les SLAM et MK qui en constituent respectivement les populations de départ et d'arrivée).

3.2 Normalisation

Pour une souris donnée, les n trails inferés n'ont pas tous la même longueur (que ce soit en nombre de cellules ou la somme des distances entre chacune des cellules de la séquence). Pour obtenir des résultats en moyenne ou en distribution, il est nécessaire de normaliser les trails. Cette étape de normalisation est représentée sur la figure 9.

Pour un trail t, on connaît l'enchaînement des cellules et la distance $D_{i\to i+1}$ entre deux cellules consécutives i et $i+1$. La distance considérée ici est une distance euclidienne. On normalise alors une première fois en rapportant chaque distance entre cellules à la distance totale du trail (D_t) : $d_{i\to i+1} = D_{i\to i+1}/D_t$. C'est ce que représente l'axe d sur la figure 9. Soit $l(t)$ le nombre total de cellules dans le trail t. On a $d_t = \sum_{i=1}^{l(t)-1} d_{i\to i+1} = 1$. Notons que, comme le trail n'est pas en ligne droite, D_t n'est pas égale à la distance entre la première et la dernière cellule du trail t.

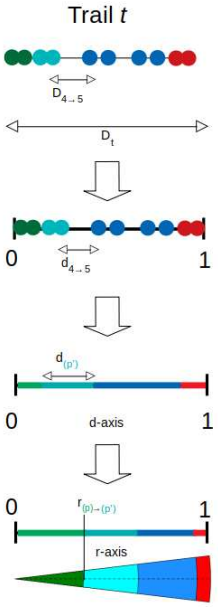


FIGURE 9 – Normalisation des trails. Un trail t quelconque, inféré par la méthode précédente, consiste en une séquence de cellules séparées entre elles d'une certaine distance (euclidienne) D. Une première étape de normalisation donne des distances d comprises entre 0 et 1. Les trails peuvent alors être interprétés de façon continue, en s'intéressant aux types p de cellules qui se succèdent, et aux proportions qu'ils occupent sur le trail $d_{(p)}$. On peut alors utiliser la transformation $r = \sqrt{d}$ qui permet de donner plus de visibilité aux types immatures (SLAM et MPP), et qui permet de représenter l'axe de progression des trails (r-axis) comme étant celui d'une portion de disque.

Si on ne s'intéresse pas aux cellules en particulier, mais plutôt aux types (sachant que nécessairement, pour un trail, les types ne peuvent s'enchaîner que dans un ordre précis) on peut alors s'intéresser à la distance $d_{(p)}$ parcourue le long du trail dans le type cellulaire $p \in \{1, \cdots, 6\}$. Au vu de la normalisation effectuée, $d_{(p)}$ correspondra au pourcentage d'occupation du trail considéré par le type cellulaire p. Les trails qui jusqu'à présent étaient des enchaînements de cellules, séparées chacune d'une certaine distance, sont maintenant décrits de façon continue comme des enchaînements de types, chaque type p représentant une portion $d_{(p)}$ de la distance $d_t = 1$ du trail.

On introduit également $r = \sqrt{d}$, distance qu'on préférera par la suite pour illustrer les résultats. Cette transformation permet de donner plus d'importance aux faibles valeurs de d, donc notamment aux cellules immatures type SLAM, MPP ou LSK*. Il s'agit avant tout d'un choix fait pour des questions de visibilité et clarté des résultats. Si on imagine représenter l'hématopoïèse (ou plutôt la mégacaryopoïèse) par une portion de disque, et qu'on choisit de faire correspondre à la quantité d les fractions de surface comme sur le schéma de la figure 9, alors la distance r fera référence à la progression le long de l'axe du disque.
D'autres choix auraient été possibles pour la transformation $r : d \longmapsto \sqrt{d}$, choix auxquels on pourrait faire correspondre une autre représentation visuelle.

3.3 Filtration

Tous les trails ne sont pas à garder, certains seront exclus car - empiriquement - ne correspondraient pas au processus que l'on souhaite décrire, à savoir la mégacaryopoïèse.
Pour cette partie, on s'intéressera en particulier au pourcentage d'occupation des LK* : $d_{(LK*)}$, sachant qu'il s'agit du type le plus représenté dans notre jeu de données.
Nous présentons la méthode dans le cas de la souris WT #206. Les étapes de filtration pour les

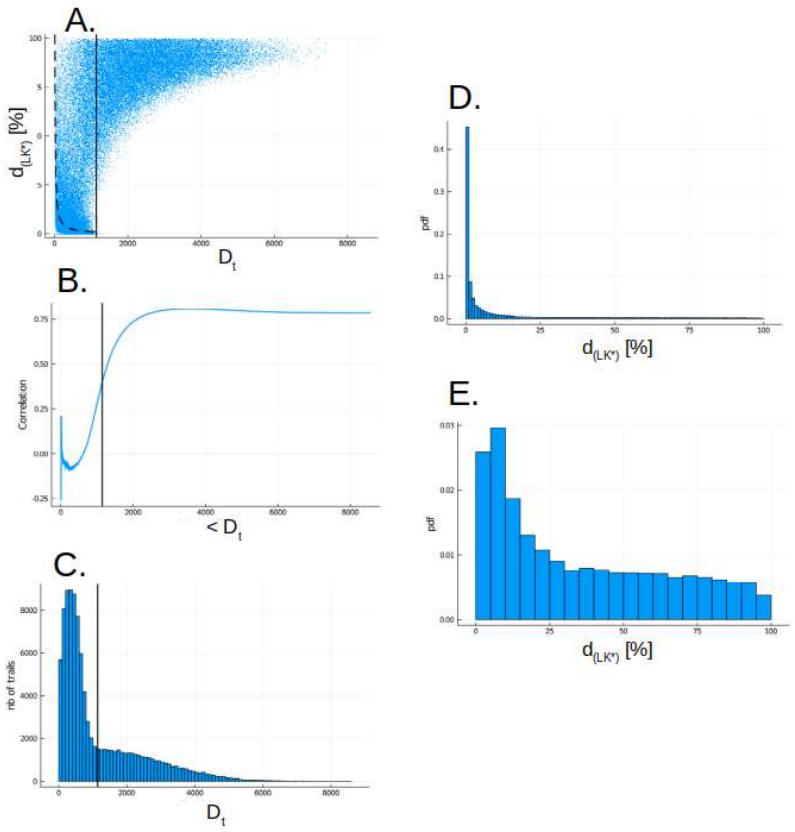


FIGURE 10 – Différents critères d'exclusion des trails. Les figures A, B et C sont réalisées à partir des $n = 100,000$ trails. La figure A représente la distribution de ces trails en fonction de leur distance totale D_t (en abscisse) et le pourcentage d'occupation des LK* $d_{(LK*)}$ (en ordonnée). La figure B représente le coefficient de corrélation (Pearson), calculé sur les trails d'une longueur D_t inférieure au seuil en abscisse, entre les quantités $d_{(LK*)}$ et D_t. On veut garder les trails tels que ce coefficient de corrélation ne soit pas trop élevé. On choisit comme seuil 0.4, c'est-à-dire qu'on ne conserve que les trails dont la longueur $D_t < 1,143$ (ligne verticale), qui représentent ici 65.4% des n trails. La figure C représente la distribution de D_t. Sur la figure D, on représente la distribution du pourcentage d'occupation des LK* $d_{(LK*)}$ sur les trails de longueur $D_t < 1,143$. Une forte proportion de ces derniers ne passent quasiment pas par les LK*. Après suppression de ces derniers (zone délimitée par la ligne pointillée sur la figure A), on se retrouve avec les 24,312 trails pour lesquels on montre la distribution de $d_{(LK*)}$ sur la figure E.

autres souris sont présentées dans l'annexe A à ce chapitre [3].

3.3.1 Exclusion des trails longs

Certains trails pour lesquels l'occupation des LK* représente un trop grand nombre de cellules (dans un sens qu'on définira par la suite), peuvent être vus, intuitivement, comme des trails qui se sont "égarés" dans les LK*, et qui ne sont par conséquent pas représentatifs de la lignée mégacaryocytaire. En effet, les LK* représentent un ensemble hétérogène de cellules, dont certaines seraient déjà plutôt engagées vers d'autres lignées que la lignée mégacaryocytaire.
Pour étudier cela rigoureusement, on va s'intéresser pour chaque trail t à sa taille totale D_t (somme des distances entre les cellules qui constituent le trail) et au pourcentage d'occupation des LK* : $d_{(LK*)}$. On représente sur la figure 10-A la répartition des trails en fonction de la valeur de ces deux quantités. Ce qu'on peut observer, c'est que dépassée une certaine longueur (absolue) de trail, on ne se trouve qu'en présence de trails qui occupent une part importante de LK*. Alors qu'en dessous de ce seuil, on peut avoir à la fois des trails avec un fort ou faible pourcentage

3. disponibles au lien : https ://gitlab-research.centralesupelec.fr/2012hermangeg/supplementary-material-phd

d'occupation de cette population. C'est ce qu'illustre également la figure 10-B qui représente la corrélation entre ces deux quantités, calculée sur l'ensemble des trails d'une longueur absolue inférieure à celle indiquée en abscisse. Afin d'avoir un ensemble de trails pour lesquels il n'y a pas une forte corrélation entre la longueur et le pourcentage des LK*, nous choisissons d'exclure les trails d'une longueur D_t supérieure à une certaine taille maximale. Afin d'avoir un critère le plus objectif possible, qui puisse être appliqué à chacune des souris étudiée à la section 4, on va définir cette taille maximale sur la base de la corrélation entre taille des trails et occupation des LK*. On choisit de ne pas dépasser une corrélation égale à 0.4, choix ajusté notamment pour bien séparer les distributions sur la figure10-C, et ce pour la plupart des souris étudiées à la section 4.

3.3.2 Exclusion des trails courts

Après exclusion des trails longs, on peut regarder la distribution des pourcentages d'occupation des LK* sur l'ensemble des trails restants. Ce qu'on observe (Fig. 10-D), c'est globalement des distributions de type "zero-inflated", c'est-à-dire avec un pic en zéro (ou aux faibles valeurs).
Ce qui indique que certains trails *bypassent* la population LK*. On considère que ces trails sont à exclure, car il y a un risque qu'ils correspondent à des artefacts, c'est-à-dire des cellules plus matures (MkP ou MK) connectées, dans le graphe, à des cellules immatures (SLAM, MPP ou LSK*), sans qu'il y ait de raison biologique pour les considérer proches.
Le critère d'exclusion, représenté par la ligne en pointillé sur la figure 10-A est le suivant : on exclut les trails dont la distance $D_{(LK*)}$ est inférieure à 1% de la taille maximale des trails (valant 1,143 pour la WT # 206).
Finalement, après exclusion des trails longs et courts, on se retrouve avec un nombre réduit de trails (Fig. 10-E) pour les analyses ultérieures. Sur $n = 100,000$ trails initialement, on en conserve $n' = 24,310$.
Il est à souligner que notre méthode de filtration repose sur des critères empiriques, dont il serait important d'évaluer plus en détail l'influence sur les résultats obtenus.

3.4 Trajectoire de différenciation

Les trails donnent une information sur la façon dont les types cellulaires se succèdent. Les résultats seront analysés en moyenne.
On ne va pas ici s'intéresser tout à fait aux types cellulaires, mais à ce que nous appelons "compartiments". Ces derniers sont simplement définis de la façon suivante :

- SLAM
- LSK : regroupe les types MPP et LSK*
- LK : regroupe les types LK* et MKP
- MK

Pour les analyses suivantes, il est nécessaire de se placer au niveau des compartiments plutôt que des types, ceci car certains types sont faiblement représentés dans le dataset, comme nous pouvions le constater sur la figure 5.

Pour chaque trail t, nous avons l'information sur la distance r parcourue avant de passer au prochain compartiment (voir Fig. 9). Ici, un trail sera donc caractérisé par les trois valeurs (Fig. 11-A) :

- $r_{SLAM \to LSK} = \sqrt{d_{(SLAM)}}$ qui est la distance normalisée à partir de laquelle on quitte le compartiment SLAM vers le compartiment moins immature LSK
- $r_{LSK \to LK} = \sqrt{d_{(SLAM)} + d_{(LSK)}}$ (qui est égale à $r_{SLAM \to LSK}$ dans le cas où il n'y aurait pas de LSK sur le trail considéré)
- $r_{LK \to MK} = \sqrt{d_{(SLAM)} + d_{(LSK)} + d_{(LK)}}$

Ainsi, à une proportion donnée $r \in [0, 1]$, on peut savoir pour chaque trail dans quel compartiment il est. En moyennant sur l'ensemble des trails, on peut alors estimer $\mathbb{P}[t \in c|r]$, probabilité qu'un

trail t quelconque soit dans le compartiment c après avoir parcouru la distance r. On peut tracer, pour chaque compartiment, l'évolution de cette quantité en fonction de $r \in [0,1]$ (Fig. 11-B.i). On définit enfin notre trajectoire de différenciation τ de la façon suivante (Fig. 11-B.ii) :

$$\begin{aligned} &[0,1] \longrightarrow [0,1] \\ &r \longmapsto \tau(r) = \frac{1}{3}\mathbb{P}[t \in LSK|r] + \frac{2}{3}\mathbb{P}[t \in LK|r] + \mathbb{P}[t \in MK|r] \end{aligned} \tag{2}$$

qui correspond à une espérance, celle d'appartenir à tel ou tel compartiment, en ayant associé les compartiments SLAM, LSK, LK et MK aux valeurs 0,1/3, 2/3 et 1 respectivement.

On peut également calculer la moyenne, sur l'ensemble des trails considérés, des quantités définies plus haut, à savoir $\mathbb{E}[r_{SLAM \to LSK}]$, $\mathbb{E}[r_{LSK \to LK}]$ et $\mathbb{E}[r_{LK \to MK}]$. Ce qui nous conduit à la représentation de la figure 11-C.

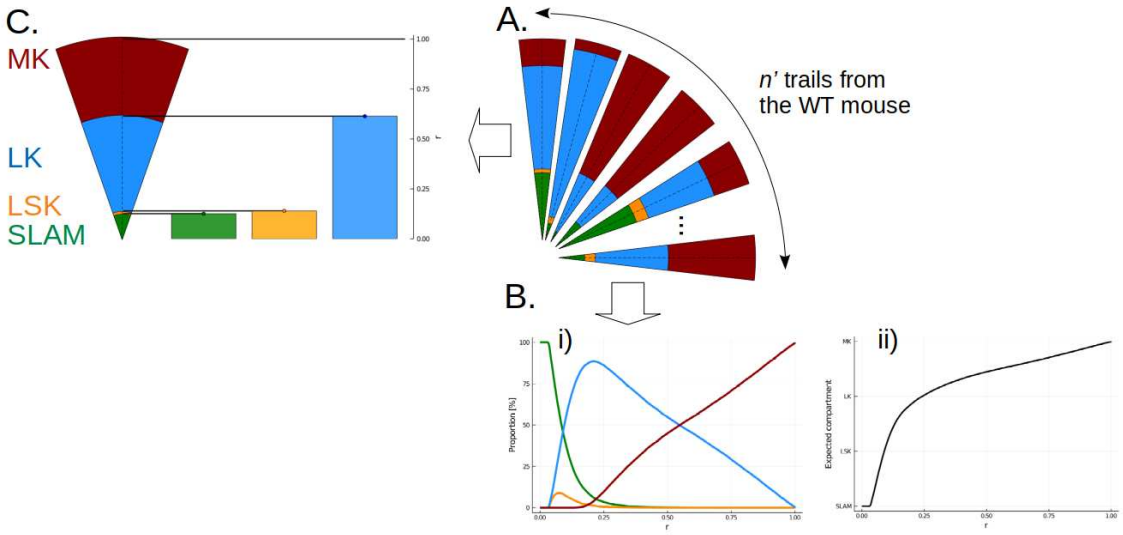


FIGURE 11 – Illustration de la construction de la trajectoire de différenciation (B.ii) à partir des n' trails inférés et sélectionnés (A). On peut calculer (B.i), pour toute valeur de r, le pourcentage de trails à être dans le compartiment des SLAM (vert), LSK (jaune), LK (bleu) ou MK (rouge). En combinant ces proportions suivant l'équation (2), on obtient la trajectoire de différenciation τ (B.ii). L'information sur les trails peut également être représentée par les valeurs moyennes de r auxquelles on passerait d'un compartiment à un autre (C).

3.5 Évolution des intensités au cours de la trajectoire de différenciation

L'objectif de ce travail est de pouvoir inférer l'évolution de l'intensité des différents marqueurs au cours de la différenciation. Pour cela, comme pour toute méthode d'inférence de trajectoires, il faut faire l'hypothèse que les cellules utilisées pour l'analyses sont suffisantes pour décrire le processus de développement de la lignée mégacaryocytaire, à savoir qu'elles sont en nombre suffisant, que les différents états intermédiaires sont représentés et que les changements d'expression sont graduels (hypothèse d'un continuum).
Pour étudier l'évolution de l'intensité des marqueurs le long de la trajectoire de différenciation, nous allons moyenner les intensités des différents des trails (ceux qui ont été sélectionnés à l'étape décrite au § 3.3).

Pour rappel, un trail est défini comme un enchaînement de cellules le long d'un axe de progression r. Ces cellules sont caractérisées par leur intensité pour m marqueurs. On représente sur la figure 12 (gauche) l'évolution de l'intensité d'expression d'un marqueur (en l'occurrence

CD41) pour un trail quelconque. On peut voir des fluctuations assez importantes de l'intensité d'expression de ce marqueur, qu'on choisit de lisser en faisant une moyenne flottante (sur une fenêtre $r \pm 0.03$). On moyenne alors l'ensemble des courbes (lissées) des n' trails pour obtenir l'évolution de l'intensité d'expression des marqueurs (Fig. 12 droite). Cette fonction, définie pour $r \in [0, 1]$, est notée $\mathcal{I}_j$ pour le marqueur j.

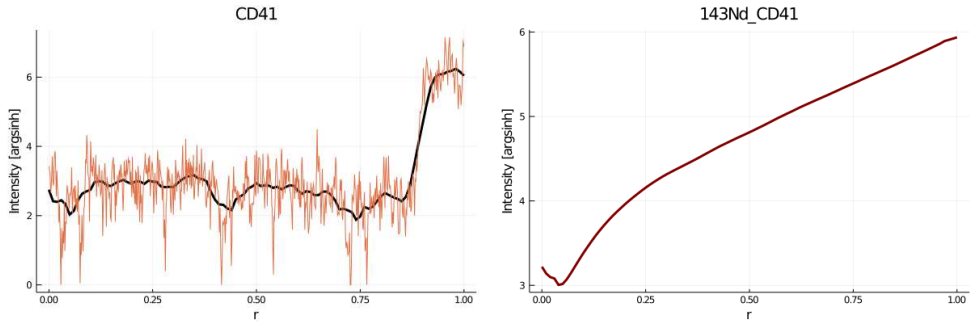


FIGURE 12 – Évolution de l'intensité d'expression du marqueur CD41 suivant l'axe r. À gauche, en orange l'évolution pour un trail quelconque. Les fluctuations sont lissées en calculant pour chaque trail une moyenne flottante (noir). Puis, en moyennant sur les n' trails, on obtient l'évolution moyenne $\mathcal{I}_{CD41}$ de l'intensité d'expression de CD41 le long de la trajectoire de différenciation.

On peut également regarder la trajectoire conjointe de certains marqueurs utilisés pour le partitionnement, comme présenté sur la figure 13. Par construction, on retrouve les différentes étapes de la mégacaryopoïèse liées aux différents types cellulaires qui avaient été considérés, et on peut les placer suivant l'axe r. On pourra alors regarder l'évolution de certains marqueurs intracellulaires pour comprendre comment certains marqueurs de signalisation évoluent au cours de la différenciation.

Soulignons que, contrairement à d'autres auteurs, nous évitons d'introduire un aspect temporel pour l'interprétation des résultats, et de parler de dynamique. L'hématopoïèse - et la mégacaryopoïèse - sont des processus dynamiques, que nous ne supposons pas pouvoir reconstruire par notre méthode d'inférence de trajectoires basée sur un échantillon obtenu à un seul instant. Notre trajectoire de différenciation doit ainsi être comprise comme une représentation de la façon dont les différents types cellulaires se succéderaient au cours de la différenciation, du plus immature au plus mature. En particulier, l'axe r ne doit pas être interprété comme un pseudo-temps.
C'est principalement dans la comparaison relative entre trajectoires obtenues pour souris de différents génotypes que nous pourrons obtenir des résultats pertinents d'un point de vue biologique. Ce sera l'objet de la section 4, mais avant cela, introduisons une méthode d'estimation de l'incertitude sur nos trajectoires d'évolution des intensités.

3.6 Estimation de l'incertitude

Pour estimer une incertitude sur nos trajectoires, nous appliquons la méthode précédente sur 10 sous-échantillons du jeu de données initial. Pour chacun, nous échantillonnons au hasard 80% de cellules du data set de départ (WT # 206). Nous reproduisons ensuite les mêmes étapes, à savoir inférence de 100,000 trails, normalisation, filtration des trails longs et courts, inférence de la trajectoire de différenciation et de l'évolution des intensités des marqueurs au cours de celle-ci. Pour chaque marqueur, nous pouvons représenter les 10 trajectoires, ou encore une trajectoire médiane, minimale et maximale (en calculant ces 3 quantiles pour toute valeur de r, sur les 10 trajectoires), comme représenté pour quatre exemples de marqueurs sur la figure 14. Nous observons une faible incertitude sur les résultats.

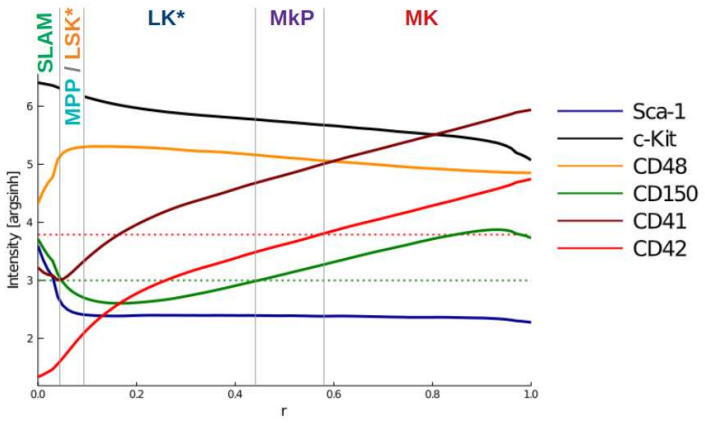


FIGURE 13 – Trajectoires d'évolution des intensités d'expression des marqueurs de surface Sca-1, c-Kit, CD48, CD150, CD41 et CD42 au cours de la trajectoire de différenciation. Les lignes horizontales en pointillées, correspondant aux seuils de positivité (voir table 1) des marqueurs CD150 (vert) et CD42 (rouge), permettent de donner une indications quant à la succession des types cellulaires (rappelons que les courbes représentent le comportement moyen des trails ; à une valeur de r donnée, nous avons en réalité une distribution de plusieurs types cellulaires sur nos n' trails. Les seuils de positivité des différents marqueurs ne sont en toute rigueur applicables qu'aux trails). On retrouve les SLAM lorsque CD150 est positif et l'expression de CD48 faible (relativement), puis les MPP/LSK* qui leur succèdent. Lorsque Sca-1 n'est plus exprimé, on se retrouve avec les LK*, suivis des MkP qui sont $CD150^+$, et enfin des MK qui sont $CD42^+$. Notons que, dans notre partitionnement schématisé sur la figure 4, on distingue les MK c-kit^+ de ceux c-kit^-, ces derniers étant considérés plus matures.

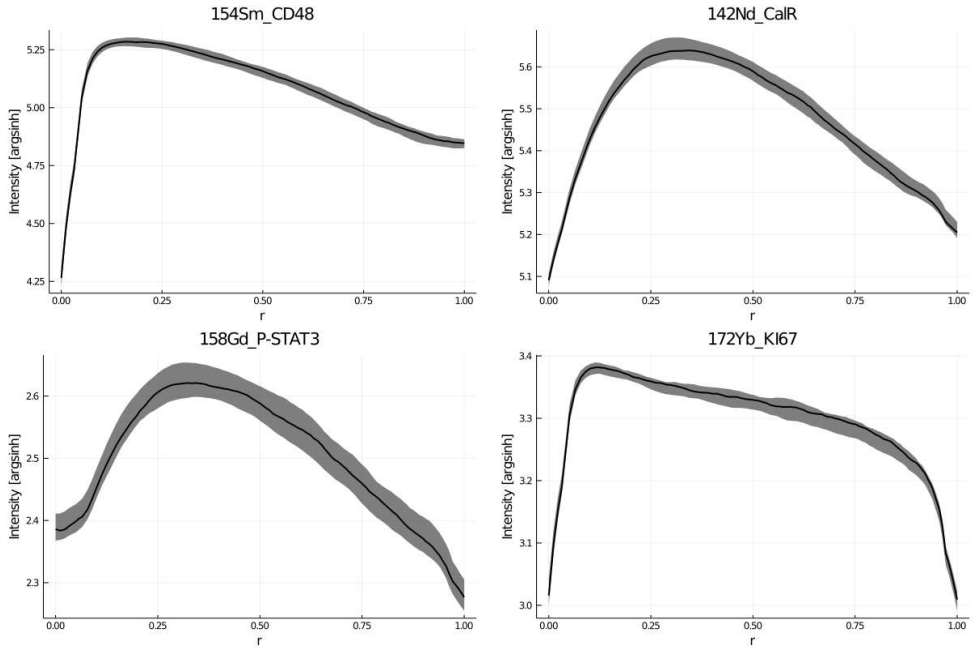


FIGURE 14 – Trajectoires des intensités de CD48, CALR (intra-cellulaire), P-STAT3 et KI67 avec estimation d'une incertitude. Pour chaque valeur de r et chaque marqueur, partant des 10 trajectoires estimées à partir d'un sous-échantillon du jeu de données initial, on calcule la valeur médiane (ligne noir) ainsi que les valeurs minimale et maximale (l'écart entre ces valeurs étant représenté par la marge grisée).

4 Application à l'effet des mutations $CALR^m$ T1 et T2

4.1 Contexte biologique et expériences

Les modèles murins, en reproduisant très bien les hémopathies malignes observées chez l'homme, ici les néoplasmes myéloprolifératifs (NMP), peuvent aider à mieux les comprendre [24] et à identifier des traitements prometteurs, comme par exemple une combinaison d'Interféron α avec de l'Arsenic [25].
Nous nous intéresserons ici à des modèles de souris knock-in *CALR del52* (T1) et *ins5* (T2). Chez l'homme, parmi les trois types de NMP, les thrombocytémies essentielles (TE) et les myélofibroses primaires (MF) sont des hémopathies impliquant la lignée mégacaryocytaire, la première conduisant à une surproduction de plaquettes, la seconde à une surproduction de mégacaryocytes et granulocytes. Dans environ 30% des cas, ces maladies sont dues à une mutation du gène *CALR* [26, 27]. La calréticuline (CALR) est une protéine chaperonne du réticulum endoplasmique qui joue notamment un rôle dans la régulation des ions calcium [28] ainsi que dans la conformation spatiale d'autres protéines [29].
La CALR mutante ($CALR^m$) va avoir une nouvelle fonction, se fixer sur MPL et déclencher l'activation de la voie de signalisation JAK2/STAT [30, 31, 32]. Or, MPL est le récepteur à la thrombopoïétine (TPO), facteur de croissance impliqué dans la mégacaryopoïèse. La présence de $CALR^m$ dans les cellules va alors induire une indépendance et hypersensibilité à la TPO, expliquant la surproduction de plaquettes et mégacaryocytes dans ce cas. Parmi les différentes mutations du gène *CALR* - toutes affectant le domaine C-terminal de la protéine impliquée dans la régulation du calcium et générant une nouvelle séquence chargée positivement ainsi qu'une délétion du signal KDEL de rétention dans le réticulum endoplasmique - la mutation la plus prévalente est la *CALRdel52* (Type 1) suivie de *CALRins5* (Type 2), avec une répartition estimée à 56 et 32% respectivement (sur une cohorte de 368 patients atteints de TE ou MF [33]). Pour étudier les différences entre ces deux types de mutations, Marty et al. [30] ont greffé, à des souris receveuses préalablement létalement irradiées, des cellules hématopoïétiques Lin- transduites avec des rétrovirus exprimant ces mutations. Ils ont observé que les mutants *CALRdel52* (T1) conduisaient à un phénotype plus sévère de la maladie, notamment par une augmentation de la production de plaquettes comparée aux mutants *CALRins5* (T2). Benlabiod et al. [34] ont montré que la génération de modèles murins plus physiopathologiques knock-in pour *CALRdel52* (T1) et *ins5* (T2) permettaient aussi de bien reproduire le phénotype de la maladie. La thrombocytose était plus importante avec la mutation T1 qu'avec la mutation T2. De plus, alors que la mutation T2 conduisait à une augmentation de la taille (ploïdie) des MK, la mutation T1 induisait à la fois une augmentation du nombre et de la taille des MK. De plus, la mutation T1 induisait une augmentation des cellules souches hématopoïétiques contrairement à la mutation T2. Benlabiod et al. ont également montré que le statut homozygote de ces mutations exacerbait ces différences par rapport au statut hétérozygote.
Dans ce contexte, la question qui nous intéresse est de savoir si ces différences phénotypiques, dans le cas de modèles murins *CALRdel52* et *ins5* homozygotes, pourraient s'expliquer par des différences au niveau des voies de signalisation activées au cours de la mégacaryopoïèse. On s'intéressera dans ce chapitre aux différences entre la CALR intracellulaire et extra-cellulaire, ainsi qu'à la voie de signalisation P-AKT impliquée dans la survie cellulaire (en inhibant l'apoptose), la voie JAK2/STAT5 (présentée au chapitre introductif) et la voie pEiF2 (impliquée dans la voie de réponse au stress ISR - integrated stress response).
Pour cela, la même expérience que celle présentée à la section 2 est réalisée, pour 11 souris (voir tableau 2). Les manipulations expérimentales ont été effectuées à chaque fois par groupe de 3 souris (une WT, une mutée T1 et une autre mutée T2), conduisant à différents batchs (qui correspondent à différents jours pour les expériences). Notons qu'une souris a été exclue du jeu de données, car les observations expérimentales pour cette dernière présentaient des écarts trop importants avec les autres.
Pour chaque souris, nous appliquons la même méthode d'inférence de trajectoire que celle présentée à la section 3 pour la souris WT#206 (à l'exception de l'estimation de l'incertitude, coûteuse

en temps de calcul). Avant de pouvoir comparer les souris entre elles, il est nécessaire de corriger des effets batchs. C'est l'objet du paragraphe suivant.

ID	Batch	Génotype	N_c
#202	1	WT	10,598
#201	1	T1	8,826
#69	1	T2	11,013
#206	2	WT	43,382
#199	2	T1	15,032
#71	2	T2	30,355
#70	3	WT	86,512
#198	3	T1	156,349
#239	4	WT	145,521
#238	4	T1	102,668
#132	4	T2	166,937

TABLE 2 – Liste des souris. Pour chacune, on indique le batch auquel elle appartient, son génotype (à savoir : si elle est WT ou mutée $CALR^m$ de type T1 ou T2), ainsi que la quantité de cellules Lin- N_c à disposition après les différentes étapes de prétraitement.

4.2 Proposition de méthode pour la correction des effets batch

4.2.1 Problématiques rencontrées

Avant de décrire la méthode proposée pour corriger - d'une certaine manière - les effets batch (ou plus précisément permettre la comparaison des trajectoires de différenciation et d'évolution des marqueurs entre souris), commençons par décrire certains problèmes que nous avons rencontrés lors de l'analyse du jeu de données, problèmes liés principalement à cette question de l'effet batch.
Idéalement, nous aurions souhaité avoir des jeux de données semblables entre souris de même type, et des différences claires en fonction du génotype. Puisqu'il s'agit d'une analyse multi-dimensionnelle (l'espace d'états est de taille $m = 19$), il y a plusieurs façons d'estimer dans quelle mesure deux jeux de données sont semblables. On peut par exemple s'intéresser au marqueurs un par un et visualiser la distribution de l'intensité de ces marqueurs pour les différentes souris (analyse uni-dimensionnelle). On peut aussi, comme proposé par Nowicka et al. [19], appliquer un Multidimensional Scaling (MDS)[35] - similaire à une ACP - à partir de l'expression médiane des m marqueurs (après application de la transformation (1)). À ce stade, nous observons - contrairement à nos attentes - que les différences entre jeux de données s'expliquent principalement par le batch, et non le génotype (Figure 15).
Ainsi, si on veut pouvoir comparer les génotypes afin de mettre en évidence des tendances, il est nécessaire de corriger ces effets batch.

Mentionnons également une caractéristique du jeu de données qui nous a posé problème, c'est la forte proportion de cellules ayant une intensité nulle. Les distributions uni-dimensionnelles pour les différents marqueurs sont ainsi de type *zero-inflated*. Après échanges avec Philippe Rameau, de la plateforme de cytométrie de Gustave Roussy, il apparaît qu'il n'y a pas de raison de considérer que les intensités peuvent être, pour certaines cellules, nulles par erreur, c'est-à-dire que leur intensité réelle n'aurait pas été mesurée (comme ça peut l'être dans le cas de données scRNA-seq). Ainsi, nous avons décidé de considérer qu'à une intensité nulle, pour un marqueur donné, correspondait bien en réalité une intensité nulle pour ce marqueur. Et donc qu'il n'y avait pas de raison de chercher à effectuer une correction pour ces valeurs. La réciproque n'est pas vraie. En analysant par exemple la distribution des intensités pour le marqueur CD42, nous avons remarqué une forte proportion de cellules ayant une intensité non nulle mais faible (≤ 10), proportion de cellules bien plus importante que celles devant réellement exprimer CD42. L'hypothèse que ces cellules pourraient avoir à leur surface un faible nombre de protéines CD42, justifiant les

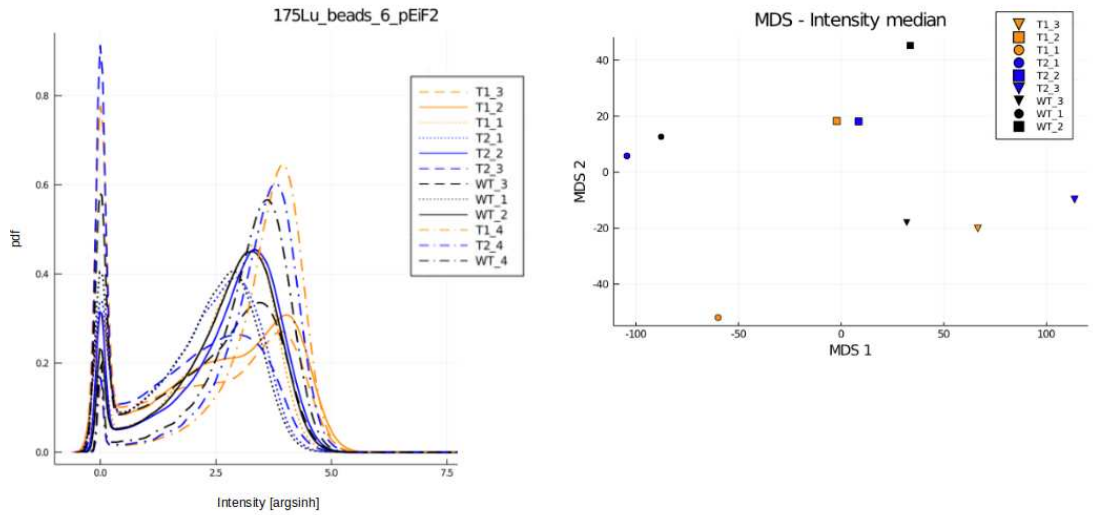


FIGURE 15 – Illustration de l'effet batch. À gauche, distribution de l'intensité pour le marqueur pEiF2 selon les différentes souris. À une couleur correspond un génotype. On observe d'importantes différences entre les courbes noires qui correspondent aux souris WT, i.e, aux souris de référence. Ces différences illustrent l'effet batch. De plus, on voit qu'au sein d'un même batch (même style de ligne), les différences dans l'allure des distributions ne sont pas marquées. Notons également que les distributions sont de type *zero-inflated*. À droite, graphe MDS, où chaque souris (ici, le calcul n'a été fait que pour les trois premiers batchs) est caractérisée par un vecteur correspondant à l'intensité médiane de chaque marqueur. Dans le cas idéal, il aurait fallu que les souris se répartissent par condition (i.e., par génotype) et non par batch comme c'est le cas ici.

faibles intensités mesurées, a été écartée. Ce qui laisse comme explication la présence d'un bruit de fond (qui peut s'expliquer par diverses sources de contaminations). Ainsi, on peut mesurer des faibles à très faibles intensités pour des marqueurs alors que ces derniers ne sont pas du tout exprimés, justifiant ainsi la définition de seuils d'expressions des marqueurs, comme présenté à la section 2.

Parmi les tentatives - n'ayant pas abouti - de correction des effets batch, nous avons tout d'abord tenté une approche unidimensionnelle. Il s'agissait de généraliser la transformation argsinh [19] en introduisant un paramètre d'échelle $a_{j,k}$ et un cofacteur $c_{j,k}$:

$$f_{j,k}(y_{i,j,k}) = a_{j,k}\text{asinh}(y_{i,j,k}/c_{j,k}) \tag{3}$$

où $y_{i,j,k}$ est l'intensité (brute) du marqueur j pour la cellule i de la souris WT du batch k.
L'objectif était de trouver, pour chaque $j \in \{1, \cdots, m\}$, les valeurs $a_{j,k}$ et $c_{j,k}$ qui minimisaient la divergence de Kullback-Leibler [36] entre $\boldsymbol{X}_{j,k} = f_{j,k}(\boldsymbol{Y}_{j,k})$ et $\boldsymbol{X}_{j,k'}$ pour $k \in \{1, 3, 4\}$ et $k' = 2$ (la souris du batch 2 ayant été choisie arbitrairement comme référence), puis d'appliquer ces coefficients pour toutes les souris du batch.
La seconde approche n'ayant pas non plus abouti reposait sur l'utilisation de la méthode de Johnson et al. [37]. Pour évaluer dans quelle mesure l'effet batch n'était pas suffisamment corrigé, nous regardions visuellement dans quelle mesure il y avait un bon mix entre les batchs par une projection UMAP des données.

Ainsi, ces premières approches ont échoué à corriger les effets batch sur les données. En perspective, nous pourrions utiliser des techniques reposant sur l'utilisation de réseaux de neurones (par exemple [38]) . Pour le moment, nous avons choisi de corriger les effets batch non pas au niveau des données d'expression des marqueurs, mais dans les trajectoires de différenciation.

4.2.2 Correction des trajectoires de différenciation (redimensionnement de l'axe r)

Lorsque nous inférons, pour chacune des souris, les trajectoires de différenciation définies par l'équation (2), nous obtenons les courbes de la figure 16 (en bas à gauche). En particulier, les courbes pour les souris WT des quatre batchs différents sont celles en haut à gauche. En faisant l'hypothèse que les différences entre souris WT sont uniquement dues à des effets batch, et non une hétérogénéité entre individus, leurs trajectoires de différenciation devraient se superposer une fois les effets batch corrigés. Nous notons $\tau_{k,WT} : [0,1] \longmapsto [0,1]$ les trajectoires de différenciation, définies équation (2), des souris WT du batch $k \in \{1,2,3,4\}$. Nous notons $\bar{\tau}_{WT}$ la trajectoire moyenne, sur les 4 batchs, correspondant à la courbe violette en haut à gauche de la figure 16 . Nous allons alors construire, pour chaque batch k, la bijection ϕ_k sur $[0,1]$ telle que :

$$\forall k \in \{1,2,3,4\}, \forall r \in [0,1] : \ \bar{\tau}_{WT}\left(\phi_k(r)\right) = \tau_{k,WT}(r) \tag{4}$$

Nous construisons ϕ_k numériquement. Pour cela, nous partitionnons $[0,1] = \bigcup_{i=1}^{1000} [r_i, r_{i+1}]$ avec $\forall i \in \{1, \cdots, 1000\}$, $r_{i+1} - r_i = 0.001$. Soit $i \in \{1, \cdots, 1001\}$ et $y_i = \tau_{k,WT}(r_i) \in [0,1]$. Alors $\exists! r'_i \in [0,1]$ tel que $\bar{\tau}_{WT}\left(r'_i\right) = y_i$. Nous définissons $\phi_{k,i} = \phi_k(r_i) := r'_i$.
L'effet de cette transformation de l'axe des r est visible en haut à droite de la figure 16 pour la souris WT#206. Cette isométrie est construite, pour chaque batch, à partir des trajectoires de différenciation des souris WT, puis est appliquée aux souris T1 et T2 (Fig. 16, en bas à droite). L'intérêt de cette transformation est de faire se superposer les trajectoires des WT pour ensuite pouvoir comparer facilement les trajectoires des T1 et T2 relativement aux WT.

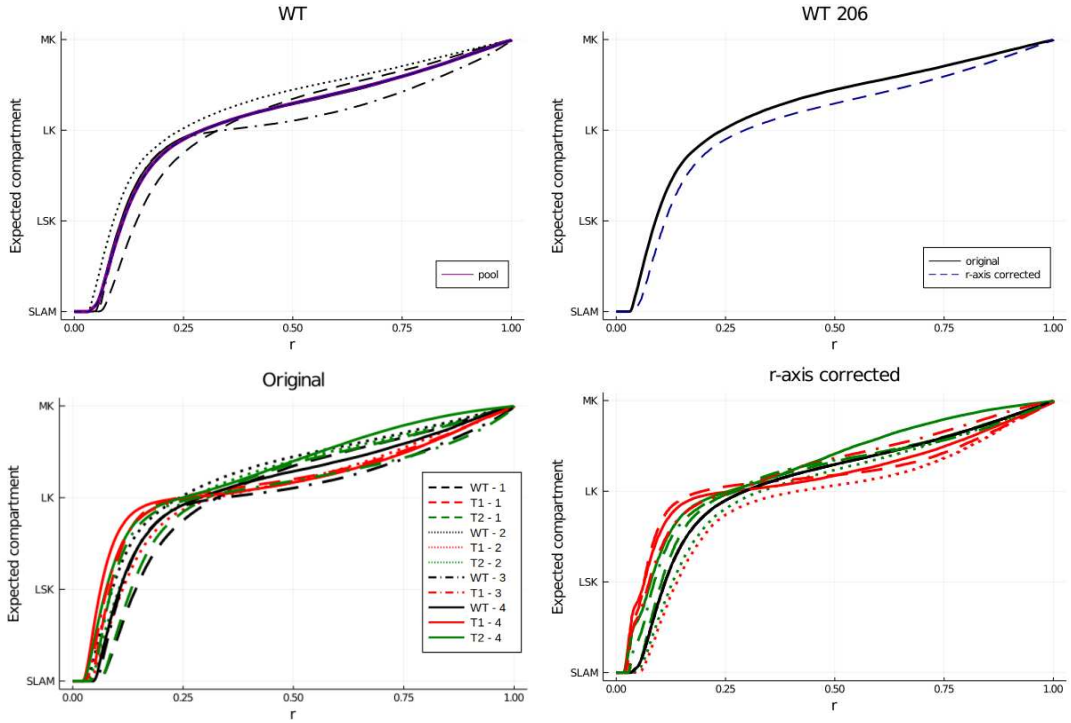


FIGURE 16 – Correction de l'effet batch sur les trajectoires de différenciation. En bas à gauche, on représente les trajectoire pour chacune des souris avant correction. En haut à gauche, on met en évidence les trajectoires $\tau_{k,WT}$ (k étant l'indice du batch) dans le cas WT ainsi que la trajectoire moyenne WT $\bar{\tau}_{WT}$ en violet. On définit une transformation ϕ_k de l'axe r permettant aux trajectoires des différentes WT de se superposer à la trajectoire moyenne, comme illustré en haut à droite dans le cas de la WT #206 (voir annexe B.1 pour les autres souris). cette transformation, construite sur les souris WT, est ensuite appliquée pour chaque batch aux souris T1 et T2 (en bas à droite). Dans ce cas, les trajectoires des souris WT se superposent.

4.2.3 Correction des intensités d'expression des marqueurs

Après application de la transformation définie en (4), on peut visualiser l'évolution de l'intensité des marqueurs le long de la trajectoire de différenciation (corrigée des effets batchs), par exemple celui de CALR (intracellulaire) à gauche de la figure 17. Notre précédente correction visait à faire se superposer les trajectoires de différenciation des WT. Il reste encore des différences quant au niveau d'expression des différents marqueurs entre batchs (i.e., entre souris WT). Pour corriger ces dernières, on adopte une approche similaire à celle précédente. Si on note $\mathcal{I}_{j,k,WT}(r)$ la trajectoire d'évolution du marqueur j pour la souris WT du batch k, en fonction de r (après correction), on définit la trajectoire WT $\bar{\mathcal{I}}_{j,WT}$ en moyennant sur les batchs. On construit ensuite, pour chaque batch k et chaque marqueur j la fonction λ_j :

$$\forall r \in [0,1],\ \lambda_{j,k}(r) = \frac{\bar{\mathcal{I}}_{j,WT}(r)}{\mathcal{I}_{j,k,WT}(r)} \tag{5}$$

Cette transformation permet de faire se superposer les trajectoires d'évolution des intensités pour les WT, et on peut ensuite l'appliquer aux souris T1 et T2 comme montré à droite de la figure 17. Cette transformation présente l'intérêt de conserver les relations d'ordre entre les trajectoires d'un même batch. Dans la présentation des résultats, au paragraphe suivant, nous utiliserons ces relations d'ordre pour évaluer dans quelle mesure certaines différences sont significatives au cours de la trajectoire.

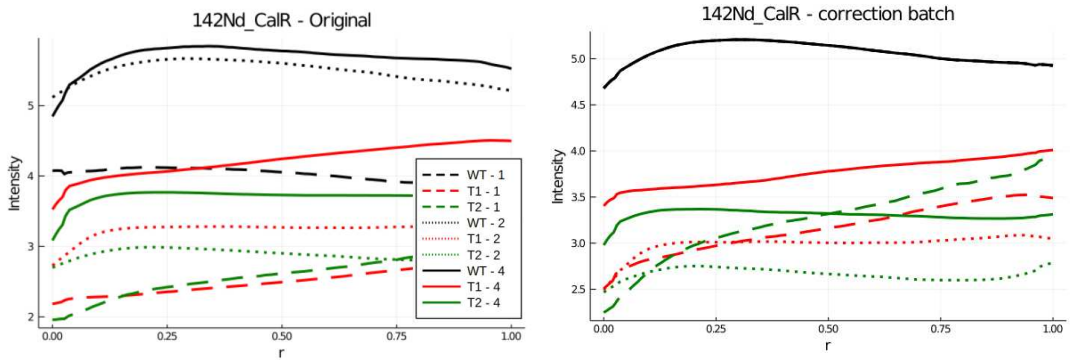


FIGURE 17 – À gauche, évolution de l'expression du marqueur CALR (intra-cellulaire) le long de l'axe r (après application de la transformation ϕ_k propre à chaque batch). On observe des différences entre les intensités, notamment entre les WT qui sont la référence. À droite, on multiplie chaque trajectoire d'origine - en fonction de son batch k - par $\lambda_{CALR,k}(r)$ (fonction de r) permettant ainsi de faire se superposer les trajectoires WT et de visualiser les relations d'ordre entre génotypes. Les résultats pour l'ensemble des marqueurs sont présentés en annexe B.2.

4.3 Résultats

Après avoir inféré les trajectoires de différenciation τ et d'évolution $\mathcal{I}$ de l'intensité d'expression des marqueurs pour chaque souris, puis appliqué notre méthode de correction des effets batchs présentée à la section précédente, on peut pour chaque marqueur visualiser l'évolution conjointe des intensités pour les différentes souris et estimer dans quelle mesure il y a des différences significatives entre génotypes (Figure 18).
Pour cela, on effectue différents tests statistiques se basant sur l'ordre entre les trajectoires, pour chaque valeur de r, et on en calcule la p-valeur. Ainsi, à une valeur de r donnée, pour un marqueur donné, on dispose de trois groupes : un formé des WT (4 souris, même valeur d'intensité après correction des effets batchs), un autre formé des T1 (4 souris) et un autre des T2 (3 souris). Soit on compare les groupes deux à deux en effectuant un test de Mann-Whitney [39], soit on compare les trois groupes ensemble en effectuant un test de Kruskall-Wallis [40]. Ces tests seront présentés plus en détail au chapitre suivant. À partir du résultat de ces tests, on peut alors

estimer sur quelles portions de la trajectoire de différenciation il y a des différences significatives ($p < 0.05$) entre génotypes (pour un marqueur donné). Lorsque ces différences sont significatives (ou ont tendance à l'être), on peut alors raisonnablement choisir de pooler ensemble (calcul d'une trajectoire moyenne) les souris par génotype, comme présenté sur la figure 19.

Nous étudions dans ce paragraphe cinq marqueurs, celui de P-AKT, celui de p-EiF2, celui de P-STAT5 et ceux de la calréticuline intra et extracellulaire (voir annexe C pour les résultats sur l'ensemble des marqueurs). Pour la CALR extracellulaire, nous obtenons de fortes différences dans l'allure des courbes, et aucune différence significative suivant le génotype, contrairement à la CALR intracellulaire pour laquelle les intensités des T1 et T2 sont toujours inférieures à celles des WT. En observant les trajectoires, nous observons une allure plutôt constante au cours de la différenciation. Nous en déduisons que les différences dans l'expression de la CALR intra qui pourraient être mesurées entre différents compartiments (ou types) cellulaires sont conservées lorsque l'on adopte une approche basée sur l'exploration du continuum d'états formés par les cellules. Notre approche n'apporte pas ici de nouveaux résultats.
L'approche basée sur l'analyse de trajectoires semble plus pertinente lorsque l'on s'intéresse à P-AKT. On observe en moyenne une augmentation de son expression dans les stades les plus immatures (SLAM puis LSK*/MPP), suivie d'une légère diminution au niveau des LK*, MkP et MK puis encore une légère diminution (dans le cas WT et T1) pour les MK les plus matures. L'allure des courbes, mises à part des différences dans l'intensité d'expression du marqueur, est globalement similaire entre WT, T1 et T2.
Concernant P-STAT5, nous n'identifions pas de différences entre les mutations T1 et T2, suggérant que la voie JAK2/STAT5 n'est pas activée différentiellement suivant le type de la mutation du gène *CALR*. Par rapport aux WT cependant, à partir des LK*, les niveaux d'intensité du marqueur P-STAT5 sont significativement plus faibles chez les souris mutantes que chez les WT. La voie P-EiF2 semblerait être activée différentiellement suivant le type de la mutation, avec une tendance à avoir des intensités plus faibles chez les souris T2 par rapport aux T1.

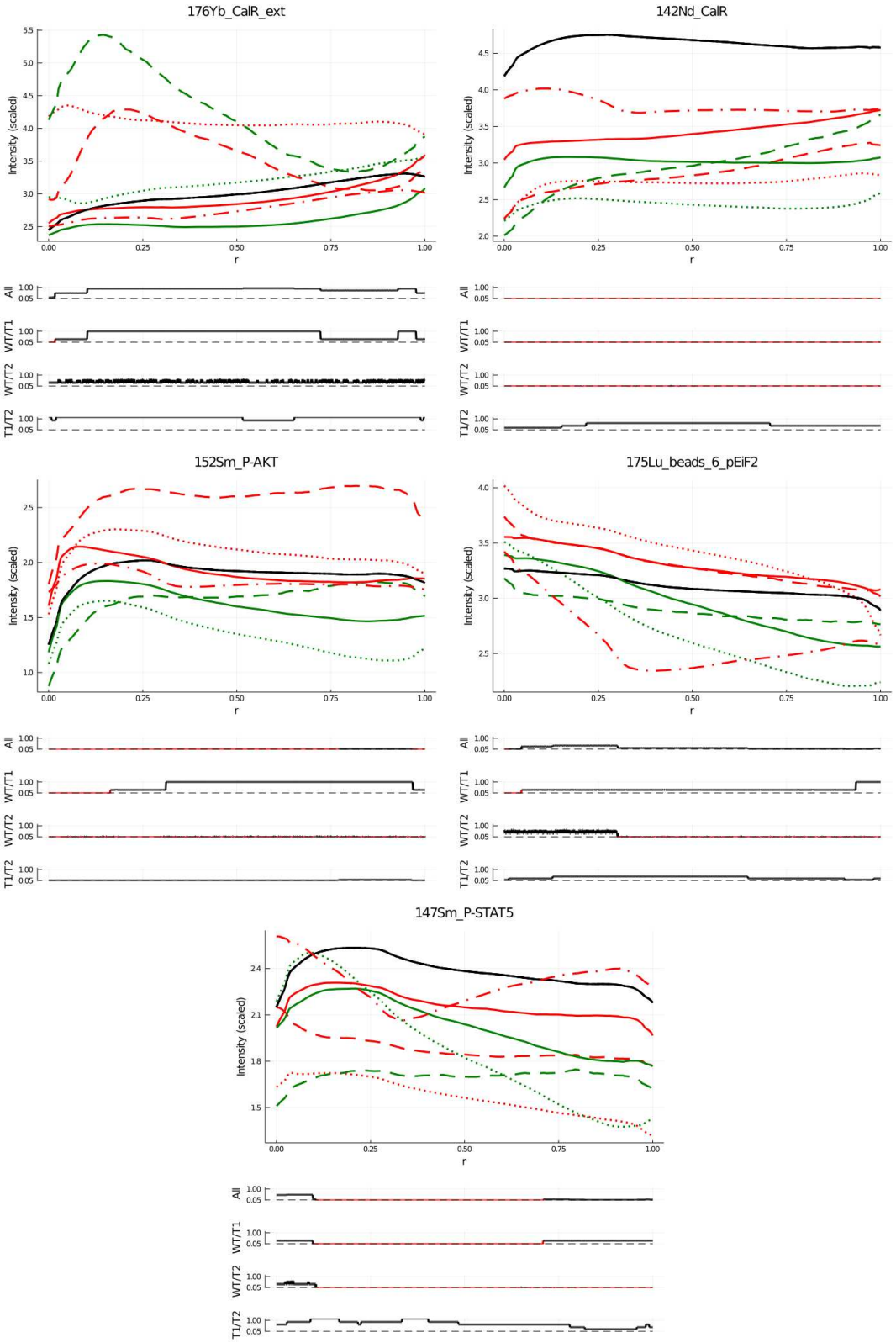


FIGURE 18 – Évolution de l'intensité des marqueurs CALR extracellulaire (en haut à gauche), CALR intracellulaire (en haut à droite), P-AKT (au milieu à gauche), pEiF2 (au milieu à droite) et P-STAT5 (en bas) pour les 11 souris du jeu de données (après correction des effets batch, donc avec superposition des souris WT). Les trajectoires des souris WT, T1 et T2 sont respectivement représentées en noir, rouge et vert. En bas, on présente l'évolution de la p-valeur, en fonction de r, basée sur un test de rang, à savoir le test de Mann-Whitney lors de la comparaison entre souris de deux génotypes, celui de Kruskall-Wallis lorsque l'on compare les souris des trois génotypes (All). Lorsque la p-valeur $p \leq 0.05$, la ligne est colorée en rouge.

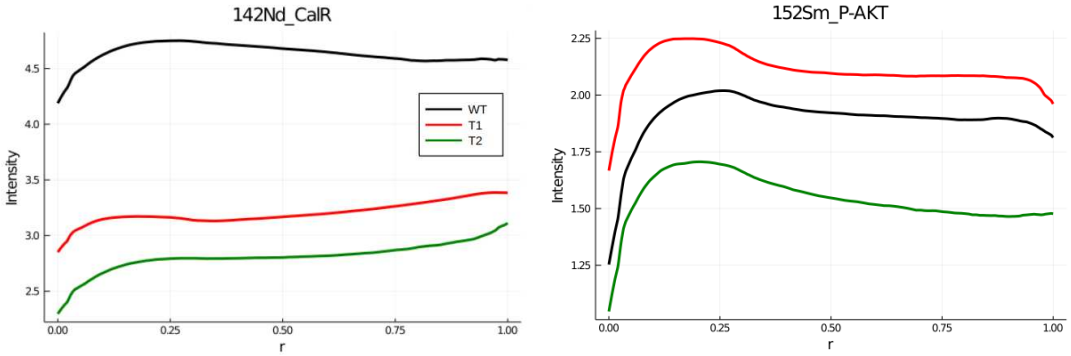


FIGURE 19 – Trajectoires d'évolution moyenne des souris WT (noir), T1 (rouge) et T2 (vert) pour la CALR intracellulaire (gauche) et P-AKT (droite).

5 Discussion

Pour étudier des différences dans l'expression de marqueurs intracellulaires entre souris mutées $CALR^m$ de type T1 et T2, au cours de la mégacaryopoïèse, nous sommes partis de mesures expérimentales obtenues par cytométrie de masse et avons proposé une méthode d'inférence de trajectoire. Celle-ci consiste à structurer les données sous la forme d'un k-NN graphe qu'on explore alors par des marches aléatoires soumises à certaines règles, marches aléatoires qu'on appellera trails. On applique ensuite certains critères de filtration pour exclure les trails longs et courts. On étudie ensuite le comportement moyen des trails, en particulier la succession des types cellulaires, pour inférer une trajectoire de différenciation ainsi que l'évolution moyenne des intensités d'expression des différents marqueurs. On applique alors la méthode aux différentes souris, en alignant les trajectoires des souris WT entre elles pour corriger - d'une certaine manière - les effets batchs.

Notre méthode, empirique, repose sur plusieurs hypothèses :

- Pour chaque souris, les cellules sont représentatives du processus étudié (la mégacaryopoïèse), c'est-à-dire qu'elles sont en nombre suffisant et que tout état intermédiaire, des SLAM aux MK, est représenté dans le jeu de données
- Le processus étudié est stationnaire, c'est-à-dire qu'on aurait obtenu une distribution similaire des cellules en réalisant l'expérience à un autre instant
- La différenciation se fait par des changements graduels dans l'expression des marqueurs de notre panel (hypothèse d'un continuum) de telle façon que le graphe se trouve être une structure adaptée pour représenter le jeu de données
- Il existe une trajectoire de différenciation, des SLAM jusqu'aux MK, qu'on peut inférer par l'exploration du graphe
- L'inférence de trails (c'est-à-dire l'exploration du graphe), lorsqu'on impose une règle pour la succession des types cellulaires et certains critères de filtre, permet, lorsqu'on les moyenne, d'inférer la trajectoire de différenciation
  - Une cellule d'un type donné (SLAM, MPP, LSK*, LK*, MkP, MK) ne pourra pas produire des cellules d'un type plus immature
  - Les trails courts sont des artefacts liés à la construction du graphe et à la possibilité de relier deux cellules qui ne seraient pas censées pouvoir avoir biologiquement un lien de parenté direct. C'est-à-dire qu'on ne considère pas pouvoir étudier ici une voie directe pour la mégacaryopoïèse (passage direct d'une SLAM à un MK ou MkP), et qu'on supprime les trails courts, en particulier ceux qui n'explorent aucun LK*
  - L'exclusion des trails longs, qui explorent dans une trop grande proportion les LK*, permet de filtrer les trails qui sont cohérents avec le processus de différenciation étudié (la mégacaryopoïèse), sachant que parmi les LK*, on retrouve des cellules qui sont engagées vers d'autres voies de différenciation

- Les différences entre les trajectoires des souris WT ne sont liées qu'à des effets de batchs, et non à une hétérogénéité entre souris

Les quatre premiers points sont des hypothèses classiques en inférence de trajectoires. Le cinquième point est notre hypothèse de construction des trails, permettant d'aboutir à une trajectoire de différenciation (Fig. 11-B.ii et 13) cohérente avec le processus étudié. La dernière hypothèse est nécessaire pour pouvoir, à partir des trajectoires inférées pour chaque souris, extrapoler des résultats en termes de génotypes. Néanmoins, en réalité, on peut s'attendre à ce qu'il y ait une certaine hétérogénéité entre souris WT.
Notre méthode repose également sur de nombreux choix dont nous n'avons pas évalué l'impact sur les résultats. Par exemple le choix d'une transformation argsinh pour les intensités brutes, d'une distance Euclidienne lors de la construction des k-NN graphes, de $k = 20$, ainsi que des différents critères de filtre pour les trails.
Parmi les limites de notre méthode, mentionnons également le fait qu'on ne prenne pas en compte l'expansion du nombre de cellules en cours de différenciation. Ainsi, nous avons une forte représentation des MK et LK dans nos trajectoires, et nous mettons insuffisamment l'accent sur les cellules plus immatures (LSK et SLAM). Nous n'avons pas non plus comparé notre méthode à celles de l'état de l'art (que ce soit sur notre jeux de données ou sur les jeux de données publiés). Pour prétendre que notre méthode explore correctement (i.e., en accord avec la réalité biologique) le continuum d'états formés par les cellules hématopoïétiques, il serait nécessaire d'effectuer cette comparaison et d'étudier l'influence de nos différents choix. Ainsi, la validité de notre méthode tient d'avantage dans la comparaison relative entre souris, puisque la méthode a été reproduite à l'identique pour chacune d'elles. Ce qui nous permet ainsi de les comparer les unes aux autres, en particulier quant à l'effet de la mutation $CALR^m$ de type T1 ou T2.

Enfin, ce chapitre aura introduit, sur un cas d'application particulier, la question de la caractérisation des cellules hématopoïétiques. Alors que l'utilisation d'un nombre restreint de marqueurs permet de partitionner les cellules en différents types, on peut observer en augmentant leur nombre que les cellules d'un même type ne forment pas un ensemble homogène, et que les frontières entre types cellulaires ne sont pas bien marquées. Les cellules hématopoïétiques se distribueraient plutôt suivant un continuum que suivant un partitionnement discret. Plusieurs représentations, et par extension classes de modèles (à espace d'états continu ou discret), sont possibles et valables. Elles dépendent de la question de recherche et des données à disposition.
Ainsi, dans les prochains chapitres, bien que nous ayons mis ici l'accent sur le continuum d'états formés par les cellules hématopoïétiques, nous adopterons une approche basée sur la caractérisation des cellules hématopoïétiques immatures en différents types sur la base de l'expression des marqueurs CD38, CD90 et CD34. Il s'agira là d'une première hypothèse dans la construction de nos modèles - la répartition des cellules en types distincts supposés homogènes - dont les résultats du présent chapitre montrent qu'il s'agit d'une hypothèse simplificatrice. Cette simplification sera néanmoins nécessaire du fait qu'on utilisera un nombre restreint de marqueurs, mais également - puisqu'on travaillera dans la suite sur l'homme et non plus la souris - un nombre relativement faible de cellules (en comparaison avec les milliers voire dizaines de milliers de cellules murines étudiées ici).

Références

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment*, 2008(10) :P10008, 2008.

[2] Stuart B Hay, Kyle Ferchen, Kashish Chetal, H Leighton Grimes, and Nathan Salomonis. The human cell atlas bone marrow single-cell interactive web portal. *Experimental hematology*, 68 :51–61, 2018.

[3] Meenakshi Venkatasubramanian, Kashish Chetal, Daniel J Schnell, Gowtham Atluri, and Nathan Salomonis. Resolving single-cell heterogeneity from hundreds of thousands of cells through sequential hybrid clustering and nmf. *Bioinformatics*, 36(12) :3773–3780, 2020.

[4] Iain C Macaulay, Valentine Svensson, Charlotte Labalette, Lauren Ferreira, Fiona Hamey, Thierry Voet, Sarah A Teichmann, and Ana Cvejic. Single-cell rna-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell reports*, 14(4) :966–977, 2016.

[5] Carolyn A De Graaf, Jarny Choi, Tracey M Baldwin, Jessica E Bolden, Kirsten A Fairfax, Aaron J Robinson, Christine Biben, Clare Morgan, Kerry Ramsay, Ashley P Ng, et al. Haemopedia : an expression atlas of murine hematopoietic cells. *Stem cell reports*, 7(3) :571–582, 2016.

[6] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7) :e47–e47, 2015.

[7] Sean C Bendall, Kara L Davis, El-ad David Amir, Michelle D Tadmor, Erin F Simonds, Tiffany J Chen, Daniel K Shenfeld, Garry P Nolan, and Dana Pe'er. Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3) :714–725, 2014.

[8] Manu Setty, Michelle D Tadmor, Shlomit Reich-Zeliger, Omer Angel, Tomer Meir Salame, Pooja Kathail, Kristy Choi, Sean Bendall, Nir Friedman, and Dana Pe'er. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature biotechnology*, 34(6) :637–645, 2016.

[9] Huidong Chen, Luca Albergante, Jonathan Y Hsu, Caleb A Lareau, Giosuè Lo Bosco, Jihong Guan, Shuigeng Zhou, Alexander N Gorban, Daniel E Bauer, Martin J Aryee, et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with stream. *Nature communications*, 10(1) :1–14, 2019.

[10] F Alexander Wolf, Fiona K Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. Paga : graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20(1) :1–9, 2019.

[11] Benedict Anchang, Tom DP Hart, Sean C Bendall, Peng Qiu, Zach Bjornson, Michael Linderman, Garry P Nolan, and Sylvia K Plevritis. Visualization and cellular hierarchy inference of single-cell data using spade. *Nature protocols*, 11(7) :1264–1279, 2016.

[12] Serena Scala and Alessandro Aiuti. In vivo dynamics of human hematopoietic stem cells : novel concepts and future directions. *Blood advances*, 3(12) :1916–1924, 2019.

[13] Elisa Laurenti and Berthold Göttgens. From haematopoietic stem cells to complex differentiation landscapes. *Nature*, 553(7689) :418–426, 2018.

[14] L Alexander Liggett and Vijay G Sankaran. Unraveling hematopoiesis through the lens of genomics. *Cell*, 182(6) :1384–1400, 2020.

[15] Daniel Prins, Hyun Jung Park, Sam Watcham, Juan Li, Michele Vacca, Hugo P Bastos, Alexander Gerbaulet, Antonio Vidal-Puig, Berthold Göttgens, and Anthony R Green. The stem/progenitor landscape is reshaped in a mouse model of essential thrombocythemia and causes excess megakaryocyte production. *Science advances*, 6(48) :eabd3139, 2020.

[16] Camélia Benlabiod. *Étude des mécanismes d'action permettant d'expliquer les différences phénotypiques entre les mutations de la calréticuline del52 et ins5 dans les néoplasmes myéloprolifératifs*. PhD thesis, Université Paris Cité, 2022.

[17] Mark J Kiel, Ömer H Yilmaz, Toshihide Iwashita, Osman H Yilmaz, Cox Terhorst, and Sean J Morrison. Slam family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *cell*, 121(7) :1109–1121, 2005.

[18] Greg Finak, Juan-Manuel Perez, Andrew Weng, and Raphael Gottardo. Optimizing transformations for automated, high throughput analysis of flow cytometry data. *BMC bioinformatics*, 11(1) :1–13, 2010.

[19] Malgorzata Nowicka, Carsten Krieg, Lukas M Weber, Felix J Hartmann, Silvia Guglietta, Burkhard Becher, Mitchell P Levesque, and Mark D Robinson. Cytof workflow : differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research*, 6, 2017.

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

[21] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6) :417, 1933.

[22] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[23] Wei Dong, Charikar Moses, and Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th international conference on World wide web*, pages 577–586, 2011.

[24] Camelia Benlabiod, Tracy Dagher, Caroline Marty, and Jean-Luc Villeval. Lessons from mouse models of mpn. *Cellular and Molecular Aspects of Myeloproliferative Neoplasms-Part B*, page 125, 2022.

[25] Tracy Dagher, Nabih Maslah, Valérie Edmond, Bruno Cassinat, William Vainchenker, Stéphane Giraudier, Florence Pasquier, Emmanuelle Verger, Michiko Niwa-Kawakita, Valérie Lallemand-Breitenbach, et al. Jak2v617f myeloproliferative neoplasm eradication by a novel interferon/arsenic therapy involves pml. *Journal of Experimental Medicine*, 218(2), 2021.

[26] Thorsten Klampfl, Heinz Gisslinger, Ashot S Harutyunyan, Harini Nivarthi, Elisa Rumi, Jelena D Milosevic, Nicole CC Them, Tiina Berg, Bettina Gisslinger, Daniela Pietra, et al. Somatic mutations of calreticulin in myeloproliferative neoplasms. *New England Journal of Medicine*, 369(25) :2379–2390, 2013.

[27] Jyoti Nangalia, Charles E Massie, E Joanna Baxter, Francesca L Nice, Gunes Gundem, David C Wedge, Edward Avezov, Juan Li, Karoline Kollmann, David G Kent, et al. Somatic calr mutations in myeloproliferative neoplasms with nonmutated jak2. *New England Journal of Medicine*, 369(25) :2391–2405, 2013.

[28] Marek Michalak, Jody Groenendyk, Eva Szabo, Leslie I Gold, and Michal Opas. Calreticulin, a multi-process calcium-buffering chaperone of the endoplasmic reticulum. *Biochemical Journal*, 417(3) :651–666, 2009.

[29] Maurizio Molinari, Klara Kristin Eriksson, Verena Calanca, Carmela Galli, Peter Cresswell, Marek Michalak, and Ari Helenius. Contrasting functions of calreticulin and calnexin in glycoprotein folding and er quality control. *Molecular cell*, 13(1) :125–135, 2004.

[30] Caroline Marty, Christian Pecquet, Harini Nivarthi, Mira El-Khoury, Ilyas Chachoua, Micheline Tulliez, Jean-Luc Villeval, Hana Raslova, Robert Kralovics, Stefan N Constantinescu, et al. Calreticulin mutants in mice induce an mpl-dependent thrombocytosis with frequent progression to myelofibrosis. *Blood, The Journal of the American Society of Hematology*, 127(10) :1317–1324, 2016.

[31] Ilyas Chachoua, Christian Pecquet, Mira El-Khoury, Harini Nivarthi, Roxana-Irina Albu, Caroline Marty, Vitalina Gryshkova, Jean-Philippe Defour, Gaëlle Vertenoeil, Anna Ngo, et al. Thrombopoietin receptor activation by myeloproliferative neoplasm associated calreticulin mutants. *Blood, The Journal of the American Society of Hematology*, 127(10) :1325–1335, 2016.

[32] Harini Nivarthi, Doris Chen, Ciara Cleary, Blanka Kubesova, Roland Jäger, Edith Bogner, Caroline Marty, Christian Pecquet, William Vainchenker, Stefan N Constantinescu, et al. Thrombopoietin receptor is required for the oncogenic function of calr mutants. *Leukemia*, 30(8) :1759–1763, 2016.

[33] Xenia Cabagnols, Jean-Philippe Defour, Valérie Ugo, Jean Christophe Ianotto, Pascal Mossuz, Julie Mondet, Francois Girodon, JH Alexandre, Olivier Mansier, Jean François Viallard, et al. Differential association of calreticulin type 1 and type 2 mutations with myelofibrosis and essential thrombocytemia : relevance for disease evolution. *Leukemia*, 29(1) :249–252, 2015.

[34] Camélia Benlabiod, Maira da Costa Cacemiro, Audrey Nédélec, Valérie Edmond, Delphine Muller, Philippe Rameau, Laure Touchard, Patrick Gonin, Stefan N Constantinescu, Hana Raslova, et al. Calreticulin del52 and ins5 knock-in mice recapitulate different myeloproliferative phenotypes observed in patients with mpn. *Nature communications*, 11(1) :1–15, 2020.

[35] Al Mead. Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society : Series D (The Statistician)*, 41(1) :27–39, 1992.

[36] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1) :79–86, 1951.

[37] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1) :118–127, 2007.

[38] Uri Shaham, Kelly P Stanton, Jun Zhao, Huamin Li, Khadir Raddassi, Ruth Montgomery, and Yuval Kluger. Removal of batch effects using distribution-matching residual networks. *Bioinformatics*, 33(16) :2539–2546, 2017.

[39] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[40] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260) :583–621, 1952.

Chapitre 3
Modélisation du temps de première division des cellules souches hématopoïétiques

Résumé

Nous étudions dans ce chapitre le temps nécessaire aux cellules souches hématopoïétiques pour effectuer leur première division en culture. Nous le modélisons par une variable aléatoire qui suit une loi choisie parmi plusieurs distributions paramétriques candidates. À partir d'observations censurées par intervalles, nous estimons les paramètres de chacun des modèles et sélectionnons le meilleur parmi ceux-ci. Le modèle gamma généralisé se trouve être celui qui donne les meilleurs résultats dans le cas des cellules souches. Un test de qualité d'ajustement confirme que ce modèle est approprié.
Dans le cas des cellules progénitrices, ce modèle ne donne pas de meilleurs résultats que ceux qu'ils généralisent - à savoir les modèles gamma et log-normal - ce qui suggère que le modèle gamma généralisé pourrait être adapté pour modéliser la sortie de quiescence des cellules souches.

Abstract

In vitro experiments, in which hematopoietic stem and progenitor cells divide in a growing medium, can provide key information about their proliferation dynamics.
In this chapter, we study the time required for hematopoietic stem cells (HSC) to complete their first division in culture. We model it by a random variable that follows a distribution chosen among several candidate parametric distributions. Several classes of distributions have already been proposed to model the division of cells in general and hematopoietic cells in particular, but few studies exist concerning HSC. An important feature of HSC is that they would spend more time in a quiescent (dormant) state than more differentiated cells. Such a property might not be appropriately modelled by standard models such as gamma or log-normal distributions and might require investigating other candidates. We will consider in this chapter the log-normal, gamma or Weibull distributions, with or without a shift, as well as the generalized gamma distribution. The latter class of distributions is large enough to include all the aforementioned parametric families of distributions as subclasses, thus making hypothesis testing easier. Using interval-censored observations, we estimate the parameters of each model and select the best among them. The generalized gamma model is found to give the best results in the case of stem cells. A goodness-of-fit test confirms that this model is appropriate.

Table des matières

1 Introduction

L'hématopoïèse est un processus dynamique ; chez l'homme, des centaines de milliards de cellules sanguines, globules rouges, plaquettes, neutrophiles, seraient produites chaque jour [1] à partir de quelques centaines de milliers de cellules souches [2]. Plusieurs cycles de division cellulaire sont ainsi nécessaires pour aboutir à la production des cellules matures, sans qu'on ait pu jusqu'à présent quantifier ce nombre de divisions chez l'homme. À titre d'exemple, Cosgrove et al. [3] estiment qu'il faudrait au moins 22 divisions chez la souris pour produire les érythrocytes. Beaucoup d'inconnu subsiste quant à la façon dont les cellules hématopoïétiques se divisent *in vivo* chez l'homme. La plupart des études ont ainsi été conduites *in vitro*. Dans ces dernières, il s'agit généralement de placer des cellules en culture dans un puits. En présence d'un cocktail de cytokines, c'est-à-dire de molécules (ou protéines) assurant une fonction de messagers chimiques, la cellule va se diviser. Plusieurs techniques existent pour suivre la division des cellules. La dilution de marqueurs de fluorescence au cours de la mitose, par exemple l'ester succinimidylique de diacétate de carboxyfluorescéine (CFSE) [4], qui se répartissent équitablement dans les cellules filles permet alors de déterminer le nombre de divisions subies par les cellules observées, ce qui peut ensuite se prêter à des analyses basées sur des modèles mathématiques, comme proposées par exemple par Zhang et al. [5] ou Bernard et al. [6]. Le suivi des cellules peut également se faire à partir de prises d'images, comme cela sera le cas pour les données présentées dans ce chapitre, ce qui s'accompagne de difficultés techniques quant à leur traitement (dont on peut trouver des exemples dans [7]). Enfin, les techniques de barcoding cellulaires [8] présentent un fort potentiel quant à la possibilité de suivre le processus de prolifération, mais également de différenciation, *in vivo* (voir Naik et al. [9] pour une revue sur le sujet).

Le temps au bout duquel la mitose survient n'est pas nécessairement le même entre deux cellules d'un même type, même pour des conditions de culture homogènes [10] : il est aléatoire. L'aléatoire, ici, pourrait être dû à un manque de connaissance sur l'état réel de la cellule [1] ou encore résulter d'une stochasticité dans l'expression des gènes [11, 12]. Nous ne chercherons pas ici à comprendre les mécanismes par lesquels les cellules se divisent, mais plutôt à modéliser le processus de prolifération. Dans ce chapitre, notre approche consistera à trouver une famille de distributions paramétriques appropriée pour décrire le temps nécessaire aux cellules souches et progénitrices pour effectuer leur première division.

Plusieurs classes de distributions ont déjà été proposées pour modéliser la division de cellules, notamment celles hématopoïétiques. Cheon et al. [13] ont par exemple étudié le cas des distributions exponentielle, log-normale, gamma ou Weibull, avec ou sans décalage (shift), pour modéliser le temps de divisions de lymphocytes. On retrouve également le choix du modèle log-normal chez Kuchen et al. [7] ou encore Duffy et al. [14]. À notre connaissance, peu d'études concernent les cellules souches hématopoïétiques (HSC). Une propriété importante des HSC est qu'elles passent plus de temps dans un état quiescent (dormant) que des cellules moins immatures. Cette caractéristique pourrait ne pas être décrite de façon adéquate par des modèles standards tels que les distributions gamma ou log-normale, et justifierait d'étudier d'autres candidats. Dans ce chapitre, nous optons pour l'utilisation de la distribution gamma généralisée pour modéliser la dynamique des HSC, en se concentrant sur l'étude du temps nécessaire à ces cellules pour effectuer leur première division en culture. Cette classe de distributions est assez vaste pour inclure, en tant que sous-classes, les familles de distributions paramétriques mentionnées plus haut. Nous évaluerons notre modèle sur des données expérimentales qui consistent en des observations des divisions cellulaires par imagerie en temps réel (Expérience Incucyte), et le comparerons aux modèles gamma, log-normal et Weibull, ainsi que leur version avec décalage. Nous appliquerons ensuite notre méthode à des types cellulaires moins immatures que les HSC, à savoir les MPP (Multipotent Progenitors) et HPC (Hematopoietic Progenitor Cells). Notons que, dans ce chapitre, la définition des HSC sera immuno-phénotypique, c'est-à-dire basée sur l'expression de marqueurs de surfaces ($CD34^+CD38^-CD90^+$). Nous utiliserons dans ce cas la notation HSC*. Parmi les

1. Comme nous l'avons vu au chapitre précédent, pour un type cellulaire donné, il peut en fait y avoir beaucoup d'hétérogénéité entre cellules.

HSC*, seule une partie d'entre elles sont vraiment des cellules souches (environ 10%), au sens conceptuel du terme introduit au chapitre 1, c'est-à-dire capable de maintenir une hématopoïèse sur le long terme.

2 Observations expérimentales

Les expériences ayant permis d'obtenir les données expérimentales sur lesquelles se base le travail présenté dans ce chapitre ont été réalisées par Alessandro Donada, de l'équipe de Leïla Perié à l'Institut Curie.

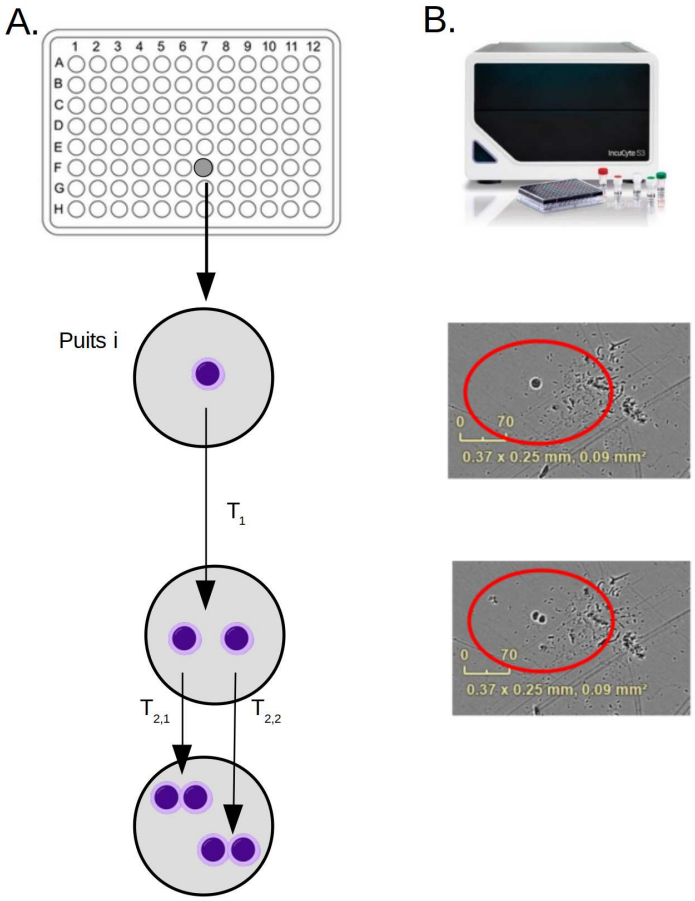


FIGURE 1 – Schéma de l'expérience Incucyte. Des cellules dont on connaît le type cellulaire sont mises en culture en présence d'un cocktail de cytokines, à une cellule par puits (A). La cellule va se diviser au bout d'un temps T_1, puis les divisions vont se poursuivre (associées à un processus de différenciation). Une prise d'image toutes les heures (B) permet alors de déterminer l'intervalle de temps pendant lequel les divisions ont eu lieu.

L'expérience Incucyte, schématisée sur la figure 1-A, consiste à déposer des cellules hématopoïétiques dans des puits, à une cellule par puits, et à les mettre en culture en présence d'un cocktail de cytokines. Les cellules ont été préalablement triées, sur la base de l'expression des marqueurs de surface CD34, CD38 et CD90, de telle sorte qu'on connaît le type cellulaire (indice p) de la cellule de départ, parmi les trois suivants :

- HSC* ($CD34^+CD38^-CD90^+$)
- MPP ($CD34^+CD38^-CD90^-$)
- HPC ($CD34^+CD38^+CD90^-$)

Les cellules proviennent de sang de cordon de différents individus. À chaque individu va alors correspondre une expérimentation différente (tableau 1). Deux cocktails de cytokines ont pu être

ID	Cytokines	HSC*	MPP	HPC
2	Diff	15	11	13
3	Diff	35	34	35
4	Diff	53	0	0
6	Diff	17	31	16
51	GT	44	47	31
57	GT	83	83	82

TABLE 1 – Liste des individus (ID) inclus dans l'étude. Dans la seconde colonne, on indique si les cellules ont été mises en culture dans le cocktail de cytokines Diff ou GT. Les trois dernières colonnes indiquent le nombre d'observations obtenues parmi les différents types cellulaires pour chaque individu.

utilisés : le cocktail Diff et celui GT (Gene Therapy).
Le premier est constitué de TPO (100 ng/mL), SCF (100 ng/mL), IL-6 (50 ng/mL), GM-CSF 25 (ng/mL), FLT3l (20 ng/mL), EPO (1 U/mL) et IL-3 (10 ng/mL). L'objectif de ce mélange de cytokines est de favoriser la différenciation des cellules vers différentes lignées, notamment la lignée érythrocytaire (favorisée par l'EPO) ou la lignée mégacaryocytaire (favorisée par la TPO).
Le second (GT - Gene Therapy) est constitué du mélange de cytokines suivant : TPO (300 ng/mL), SCF (300 ng/mL), FLT3l (300 ng/mL), IL-3 (60 ng/mL) [15]. Ce cocktail a été utilisé initialement dans les premiers essais cliniques de thérapie génique, avec pour objectif de conserver autant que possible les cellules souches et progénitrices, contrairement au précédent cocktail.

Une fois déposées dans les puits, les cellules sont mises en culture pendant $T_{max} = 96$ heures. Une photographie est prise toutes les heures ($\Delta t = 1$h), permettant alors de déterminer l'intervalle de temps pendant lequel a eu lieu la première division, ainsi que les intervalles de temps auxquels auront lieu les divisions ultérieures (Fig. 1-B). Au processus de prolifération s'ajoute un processus de différenciation : les cellules peuvent se différencier lors de la division. Dans l'expérience Incucyte, il n'est alors pas possible de connaître le type cellulaire des cellules filles. Pour cette raison, nous nous concentrerons ici sur le temps de première division, et nous étudierons la différenciation au prochain chapitre. Nous omettrons dans la suite de préciser qu'il s'agit de la première division et noterons par exemple T au lieu de T_1 la variable aléatoire correspondante. Les temps de première division, en fonction des expérimentations (i.e. des individus) et des types cellulaires de départ, sont représentés sur la figure 2.
Ces temps de première division (ainsi que ceux de deuxièmes voire troisièmes divisions) ont été reportés par Alessandro Donada. Ce travail a nécessité d'analyser et d'annoter manuellement les images une par une. Parmi les difficultés associées à ce travail, et qui rendent notamment l'utilisation de techniques d'analyse d'images automatiques compliquée, mentionnons l'existence d'un bruit de fond dans les puits (présence de débris qui peuvent être confondus avec les cellules comme on peut le voir sur la figure 1-B) ainsi que parfois des pertes de focale lors de la prise d'image.
Après 96h, les cellules sont laissées en culture jusqu'à 14 jours. La première division n'a pas nécessairement lieu avant 96h. Si, au bout de 14 jours, aucune cellule n'est observée, on suppose alors que la cellule ne s'est jamais divisée et on l'exclut du jeu de données. Finalement, le jeu de données utilisé est présenté sur la figure 2 et dans le tableau 1. Il correspond, pour un individu k et un type cellulaire de départ $p \in \{\text{HSC*, MPP, HPC}\}$, à des observations censurées par intervalles des temps de première division : $\mathcal{D}_k^p = \left\{ d_{k,1}^p, \cdots, d_{k,n_k^p}^p \right\}$, avec $d_{k,i}^p$ la borne inférieure de l'intervalle d'une heure (ou de l'intervalle $[T_{max}, +\infty]$ lors de la censure à droite due à la limite du temps d'observation) pendant laquelle la division a eu lieu pour la cellule i.
Dans la section 3, nous présenterons notre méthode d'analyse et de modélisation dans le cas des HSC*. Au paragraphe 4, nous l'appliquerons dans le cas des MPP et HPC.

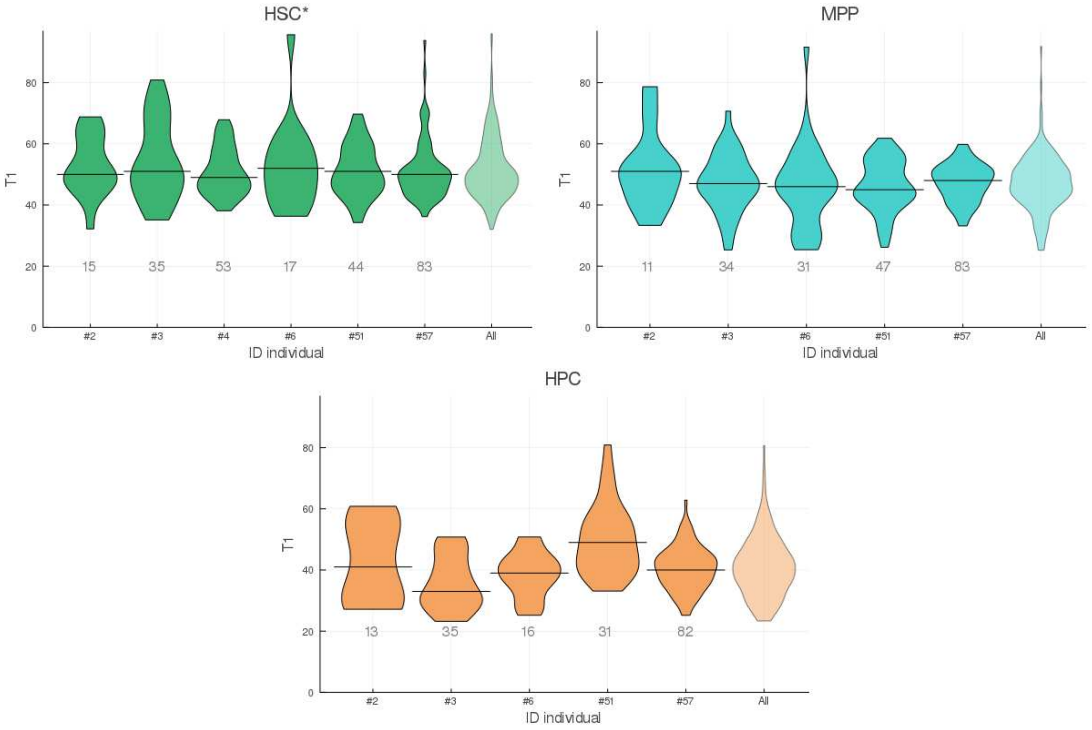


FIGURE 2 – Diagramme en violon des distributions des temps de première division des HSC* (en haut à gauche), MPP (en haut à droite) et HPC (en bas) pour les différents individus (en abscisse) inclus dans l'étude. Les lignes horizontales correspondent aux valeurs médianes. Les distributions plus claires à droite de chaque graphe correspondent aux distributions obtenues en regroupant les expérimentations. Les nombres en gris sous chaque distribution indiquent le nombre d'observations utilisées pour le tracé du diagramme.

3 Méthode

Dans cette section, nous nous concentrerons sur l'étude des temps de première division des HSC*. Nous omettrons ainsi dans la suite de mentionner l'indice p correspondant au type cellulaire de départ.

3.1 Mise en commun des données

Comme nous pouvons le voir sur la figure 2, visuellement, il semble y avoir une certaine homogénéité dans les distributions des temps de première division des HSC* entre les différents individus. Pour vérifier si on peut raisonnablement négliger l'hétérogénéité entre individus, et regrouper les données ensemble, nous allons nous baser sur les résultats du test U de Mann-Whitney [16] ainsi que celui de Kruskal-Wallis [17].

Pour un individu k, on note D_k la variable aléatoire correspondant à la borne inférieure de l'intervalle d'une heure pendant lequel la première division des HSC* a lieu. D_k est une variable aléatoire discrète. On note n_k le nombre d'observations pour l'individu k. Les observations $D_{k,1}, \cdots, D_{k,n_k}$ sont supposées indépendantes et identiquement distribuées suivant la loi de D_k. L'échantillon $\mathcal{D}_k = \{d_{k,1}, \cdots, d_{k,n_k}\}$ en est une réalisation aléatoire.
Nous allons appliquer le test de Mann-Whitney pour chaque couple d'individus. Pour un couple (k, k'), $k \neq k'$, nous considérons l'hypothèse nulle $H_{0;k,k'}$ suivante :

$$H_{0;k,k'} : \mathbb{P}[D_k > D_{k'}] = \mathbb{P}[D_k < D_{k'}]. \tag{1}$$

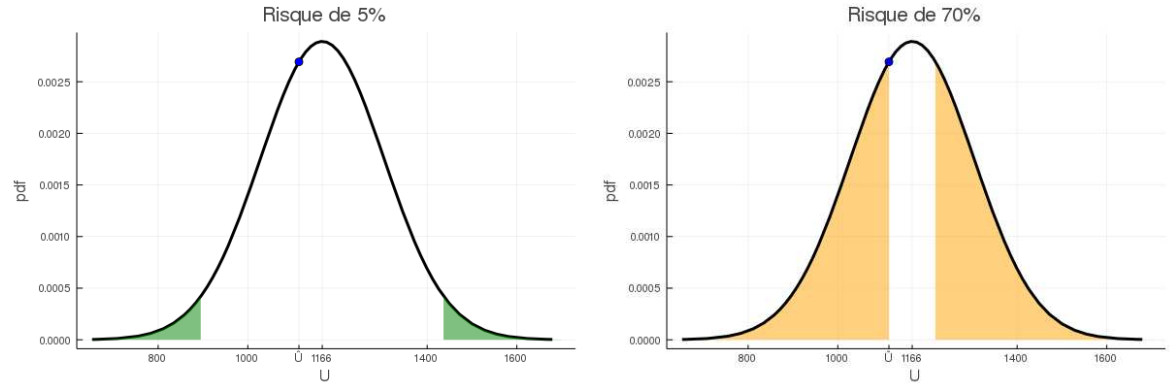


FIGURE 3 – Illustration de l'application du test de Mann-Whitney en choisissant comme couple d'individus $k = 4$ et $k' = 51$. Nous avons $n_4 = 53$ et $n_{51} = 44$. Dans ce cas, $U_{4,51}$ peut être approchée par une loi normale de moyenne 1,166 et variance 19,045. La courbe noire représente sa densité de probabilité (pdf). Nous calculons $\hat{U}_{4,51} = 1114$ (point bleu). Pour un risque maximal de 5%, on ne peut pas rejeter l'hypothèse $H_{0;4,51}$ (gauche). On calcule comme p-valeur $p = 0.7086$ (droite).

Nous définissons la variable aléatoire suivante :

$$U_{k,k'} = \sum_{i=1}^{n_k} \sum_{j=1}^{n_{k'}} \begin{cases} 1 & D_{k,i} > D_{k',j}, \\ \frac{1}{2} & D_{k,i} = D_{k',j}, \\ 0 & D_{k,i} < D_{k',j}. \end{cases} \tag{2}$$

Sous l'hypothèse nulle $H_{0;k,k'}$, nous pouvons calculer :

$$\mathbb{E}_{H_{0;k,k'}}[U_{k,k'}] = \frac{n_k \times n_{k'}}{2}, \tag{3}$$

$$\mathbb{V}_{H_{0;k,k'}}[U_{k,k'}] = \frac{n_k \times n_{k'} \times (n_k + n_{k'} + 1)}{12}. \tag{4}$$

On montre que, sous $H_{0;k,k'}$, la statistique $U_{k,k'}$ peut être approchée asymptotiquement (pour $n_k, n_{k'}$ suffisamment large) par une distribution Gaussienne dont la moyenne et la variance sont respectivement données par les relations (3) et (4).
La confrontation de la valeur $\hat{U}_{k,k'}$ calculée sur les observations aux quantiles de cette loi permet alors de rejeter ou non le test, et d'en calculer une p-valeur. Nous illustrons l'application de ce test sur la figure 3.

Les résultats du test de Mann-Whitney, présentés dans le tableau 2, montrent qu'on échoue à rejeter l'hypothèse selon laquelle les observations des temps de première division pour les HSC* pour chaque couple d'individu proviendraient de la même distribution.

Nous appliquons également le test de Kruskal-Wallis [17] qui généralise le test de Mann-Whitney à plusieurs échantillons. Nous obtenons alors des résultats similaires à ceux obtenus avec le test de Kruskal-Wallis ; le test de Kruskal-Wallis est rejeté avec une p-valeur $p = 0.8028$.

Nous négligerons donc dans la suite l'hétérogénéité entre individus (et omettrons donc de préciser l'indice k). Notre jeu de données correspond alors aux observations censurées par intervalles (l'observation correspond à la borne inférieure) des temps de première division de $n = 247$ HSC* : $\mathcal{D} = \{d_1, \cdots, d_n\}$, les données de tous les individus étant regroupées ensemble. Nous montrons sur la figure 4 la distribution empirique (histogramme) de ces observations.

$k \diagdown k'$	3	4	6	51	57
2	0.7667	0.2896	0.5966	0.4273	0.3373
3		0.2766	0.5581	0.4235	0.2723
4			0.9999	0.7086	0.9377
6				0.9999	0.8254
51					0.6385

TABLE 2 – p-valeurs obtenues par le test de Mann-Whitney, testant l'hypothèse nulle selon laquelle les observations d'un couple (ligne *vs* colonne) d'individus viendraient de la même distribution. Avec 95% de confiance, on échoue pour chaque couple à rejeter l'hypothèse nulle. D'après ces résultats, il paraît pertinent de regrouper les observations et de négliger l'hétérogénéité entre individus.

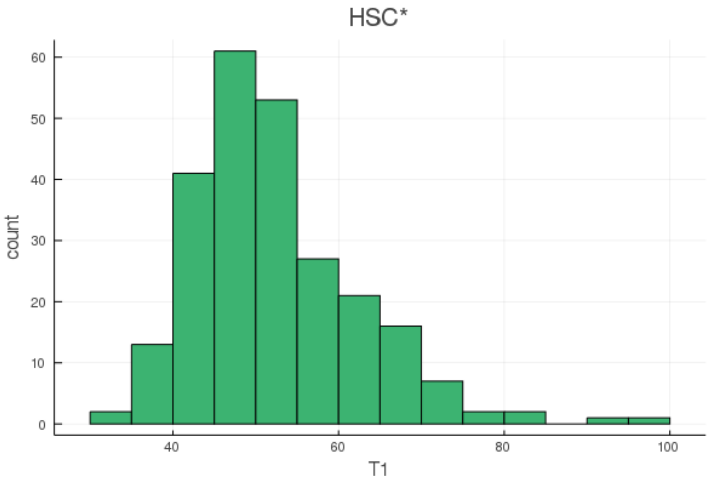


FIGURE 4 – Histogramme du temps nécessaire aux HSC* pour se diviser, toutes les observations ayant été regroupées ensemble.

3.2 Modèle statistique

Dans cette partie, nous modélisons la distribution du temps T nécessaire aux HSC* pour effectuer leur première division. Notre but consiste à choisir parmi différentes modèles paramétriques décrits à la section 3.3. Soit $\mathcal{M}$ l'un des modèles étudiés et $\boldsymbol{\theta}$ le vecteur de paramètres associé ; nous notons $T(\boldsymbol{\theta})$. Considérons n observations (censurées par intervalles) $\mathcal{D} = \{d_1, \cdots, d_n\}$. La vraisemblance est exprimée par :

$$\begin{aligned}\mathbb{P}[\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}] &= \prod_{1\leq i\leq n} \mathbb{P}[d_i|\boldsymbol{\theta}, \mathcal{M}] \\ &= \prod_{1\leq i\leq n} \mathbb{P}[d_i < T_i \leq d_i + \Delta t|\boldsymbol{\theta}, \mathcal{M}], \end{aligned} \tag{5}$$

avec $\Delta t = 1$ heure.
Étant donnés les choix de modèles considérés pour $\mathcal{M}$, la vraisemblance peut être évaluée pour toute valeur de $\boldsymbol{\theta}$ à partir de l'expression de la fonction de répartition (cdf - cumulated density function) du modèle considéré $\mathcal{M}$.

Notons que les divisions peuvent survenir après le temps d'observation $T_{max} = 96$h (cas standard de la censure à droite due à la durée limitée d'une observation). Nos observations sont censurées par intervalles, et le dernier intervalle correspond à $[T_{max}, \infty]$. Dans ce cas, nous remplaçons dans l'expression précédente $\mathbb{P}[d_i|\boldsymbol{\theta}, \mathcal{M}]$ par : $\mathbb{P}[T_{max} < T(\boldsymbol{\theta})|\boldsymbol{\theta}, \mathcal{M}]$.

Plutôt que la vraisemblance, nous préférerons étudier la log-vraisemblance :

$$\mathcal{L}(\boldsymbol{\theta}) := \log\left(\mathbb{P}[\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}]\right) \tag{6}$$

3.3 Distributions candidates

Pour modéliser le temps de première division, nous considérons plusieurs modèles, chacun d'eux correspondant à une distribution paramétrique parmi les suivantes :

3.3.1 Gamma

Nous étudions si le temps nécessaire aux HSC* pour se diviser suit une distribution gamma $T \sim \Gamma(\alpha, \beta)$ de paramètre de forme $\alpha > 0$ et de paramètre d'échelle $\beta > 0$, et dont la densité de probabilité (pdf - probability density function) est données par, pour $t > 0$:

$$f_{\alpha,\beta}(t) = t^{\alpha-1}\frac{\beta^\alpha \mathrm{e}^{-\beta t}}{\Gamma(\alpha)} \tag{7}$$

Nous considérons également le cas d'une loi gamma avec décalage (shift) T_c : $T - T_c \sim \Gamma(\alpha, \beta)$ pour $T \geq T_c$.

3.3.2 Log-normale

Nous étudions le cas de la loi log-normale $T \sim \mathcal{LN}(\mu, \sigma)$ dont la densité de probabilité, pour $t > 0$, est donnée par :

$$f_{\mu,\sigma}(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log(t) - \mu)^2}{2\sigma^2}\right) \tag{8}$$

Nous étudierons également la version avec décalage (shift T_c).

3.3.3 Weibull

Nous étudions également le cas de la distribution Weibull $T \sim W(\alpha, \lambda)$, de paramètres de forme $\alpha > 0$ et d'échelle $\lambda > 0$, dont la densité de probabilité, pour $t > 0$, est :

$$f_{\alpha,\lambda}(t) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \mathrm{e}^{-(t/\lambda)^{\alpha}} \tag{9}$$

Nous étudierons également la version avec décalage, introduisant le paramètre supplémentaire T_c.

3.3.4 Gamma généralisée

Dans la formulation originale de Stacy [18], la densité de probabilité de la loi gamma généralisée pour $t > 0$ est la suivante :

$$f_{a,d,p}(t) = \frac{\left(p/a^d\right) t^{d-1} e^{-(t/a)^p}}{\Gamma(d/p)} \tag{10}$$

avec $a, d, p > 0$, et Γ la function gamma.
Si $p = 1$, alors on retrouve la densité de probabilité de la distribution gamma (7) avec $d = \alpha$ et $a = 1/\beta$. Si $p = d$, nous retrouvons la densité de probabilité de la distribution Weibull (9) avec $p = d = \alpha$ et $a = \lambda$.

En notant $k = d/p > 0$, $\mu = \log(a) + \log(k)/p$, $\sigma = (pd)^{-1/2}$ et $Q = 1/\sqrt{k}$ signe(p) (dans la formuation originale de Stacy, $p > 0$ donc $Q > 0$), alors la densité de probabilité de la loi gamma généralisée est la suivante [19] :

$$f_{\mu,\sigma,Q}(t) = \frac{|Q|}{\sigma t \Gamma\left(Q^{-2}\right)} \left[Q^{-2}\left(\mathrm{e}^{-\mu} t\right)^{Q/\sigma}\right]^{Q^{-2}} \exp\left[-Q^{-2}\left(\mathrm{e}^{-\mu} t\right)^{Q/\sigma}\right] \tag{11}$$

Le cas limite quand $Q \to 0$ correspond à la loi log-normale. Prentice [20] propose une formule qui explicite le cas limite $Q = 0$ et autorise d'avoir $Q < 0$: si T est une variable aléatoire dont la densité de probabilité est donnée par la formule (11), alors $Y = \log(T)$ suit une loi de probabilité dont la densité est donnée par :

$$f_{\mu,\sigma,Q}(y) = \begin{cases} |Q|\left[\sigma\Gamma\left(Q^{-2}\right)\right]^{-1} \exp\left(wQ^{-2} - e^{w}\right) & (Q \neq 0) \\ \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y-\mu)^2/\sigma^2\right] & (Q = 0) \end{cases} \tag{12}$$

où $w = (y - \mu)\sigma^{-1}Q + \Psi(Q^{-2})$ et Ψ correspond à la fonction digamma.

On peut alors repasser à la paramétrisation de Stacy (avec cette fois la possibilité que $p, d < 0$ lorsque $Q < 0$) avec, pour $Q \neq 0$, $d = 1/(\sigma Q)$, $p = Q/\sigma$ et $a = |Q|^{2/p} \exp(\mu)$. Avec cette paramétrisation, la formule (10) donnant la densité de probabilité est toujours valable, à un signe près (pour rester positive). Dans la suite, lorsque T suit une loi gamma généralisée, nous notons $T \sim GG(\mu, \sigma, Q)$ et préférons la paramétrisation de Prentice [20] donnée par l'équation (12).
Dans notre implémentation de la distribution gamma généralisée en Julia (voir annexe A à ce chapitre [2]) nous utilisons la formulation (12) pour échantillonner suivant cette loi, mais utilisons la formulation de Stacy [18] pour évaluer la densité de probabilité et la fonction de répartition en une valeur $t > 0$ donnée.
Il nous a fallu implémenter en Julia la fonction de répartition de la loi gamma généralisée dont l'expression est donnée, avec la paramétrisation de Stacy, pour $t > 0$, par :

$$F_{a,d,p}(t) = \frac{\gamma\left(d/p, (t/a)^p\right)}{\Gamma(d/p)}$$

avec γ qui représente la fonction gamma incomplète. L'implémentation numérique de la fonction de répartition de la loi gamma généralisée repose sur l'observation selon laquelle : $F_{a,d,p}(t) = G((t/a)^p)$, avec G désignant ici la fonction de répartition d'une loi Gamma $\Gamma(\alpha = d/p, \beta = 1)$.

2. disponibles au lien : https ://gitlab-research.centralesupelec.fr/2012hermangeg/supplementary-material-phd

3.4 Vraisemblance profilée

Pour chaque modèle, nous calculons la fonction de vraisemblance pour différentes valeurs des paramètres, sur une grille à deux ou trois dimensions suffisamment fine. Pour étudier l'identifiabilité pratique des modèles [21], nous traçons alors, pour chaque modèle et chacun des paramètres θ_i, la log-vraisemblance profilée définie par :

$$\mathcal{L}_i(\theta_i) = \max_{\boldsymbol{\theta_{-i}}} \mathcal{L}(\boldsymbol{\theta}) \tag{13}$$

C'est-à-dire qu'on maximise la vraisemblance, pour chaque valeur de θ_i fixée, suivant toutes les autres composantes (notation $\boldsymbol{\theta_{-i}}$) du vecteur de paramètres $\boldsymbol{\theta}$. L'estimateur du maximum de vraisemblance pour θ_i se trouve au maximum de la vraisemblance profilée $\mathcal{L}_i(\theta_i)$. Si, au niveau de ce maximum, la vraisemblance profilée est plate, alors on est dans un cas de non-identifiabilité pratique, avec une forte incertitude sur l'estimateur du maximum de vraisemblance.

Les résultats pour les modèles gamma, shifted gamma, log-normal, shifted log-normal, Weibull, shifted Weibull et gamma généralisée sont présentés sur les figures 5, 6, 7, 8, 9, 10 et 11 respectivement. L'axe des ordonnées va jusqu'à -890 pour toutes les figures. Sur la base du maximum de vraisemblance estimé ici, on peut déjà écarter les modèles Weibull et shifted Weibull qui s'adaptent moins aux données que les autres modèles ayant respectivement deux et trois paramètres.
Parmi les deux modèles à deux paramètres restant, le modèle log-normal est meilleur que le modèle gamma, à la fois parce qu'il atteint une valeur plus élevée pour le maximum de vraisemblance, mais également parce que ce maximum se distingue plus nettement (donc une plus faible incertitude sur l'estimateur). Le modèle shifted log-normal ne possède, quant à lui, pas de maximum à se distinguer nettement. Le modèle gamma généralisé, à trois paramètres également, est légèrement meilleur que le modèle shifted log-normal en termes de maximum de vraisemblance et, contrairement à ce dernier, a un maximum qui se distingue plus nettement sur les courbes de log-vraisemblance profilée.
Nous poursuivons cette discussion au paragraphe suivant, basant la sélection de modèle sur des critères d'information Bayésienne et d'Akaike.

3.5 Sélection de modèle

Plusieurs critères existent pour comparer deux modèles sur la base du maximum de vraisemblance, les plus couramment utilisés étant le critère d'Akaike (AIC - Akaike Information Criterion) [22, 23] et le critère d'information Bayésienne (BIC - Bayesian Information Criterion) [24]. Ils sont respectivement définis par :

$$AIC = -2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2K \tag{14}$$

$$BIC = -2\mathcal{L}(\hat{\boldsymbol{\theta}}) + K\log(n) \tag{15}$$

avec K la dimension du vecteur de paramètre $\boldsymbol{\theta}$ et $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ l'estimateur du maximum de vraisemblance pour un modèle donné.
Ces deux critères cherchent à favoriser, entre deux modèles, celui qui maximise la vraisemblance (donc minimise la quantité $-2\mathcal{L}(\hat{\boldsymbol{\theta}})$) tout en étant le plus parcimonieux possible, c'est-à-dire avec le moins de paramètres à estimer. Les deux critères se distinguent l'un de l'autre sur la façon de pénaliser les modèles avec plus de paramètres, le critère BIC faisant intervenir le nombre n de données.
Entre plusieurs modèles, on choisira ainsi celui qui minimise le critère choisi. Les résultats de la procédure de sélection de modèle sont présentés dans le tableau 3. Nous avons considéré à la fois les critères AIC et BIC, qui nous amènent tous deux à sélectionner le même modèle, à savoir la distribution gamma généralisée, à une unité seulement devant le modèle shifted log-normal.

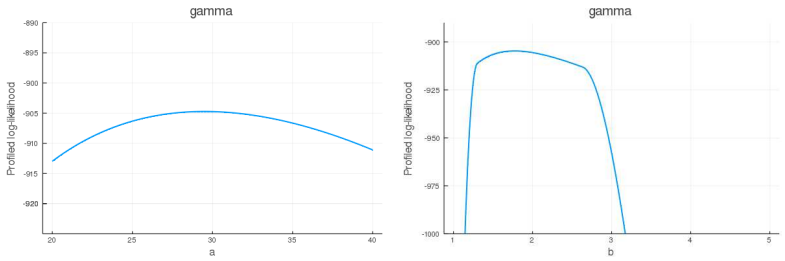


FIGURE 5 – Vraisemblance profilée, pour le modèle gamma, des paramètres α (gauche) et β (droite). Le maximum de vraisemblance, atteint sur l'ensemble des valeurs considérées pour les paramètres, vaut ici -904.72. Il est atteint pour $\hat{\alpha} = 29.6$, $\hat{\beta} = 1.765$.

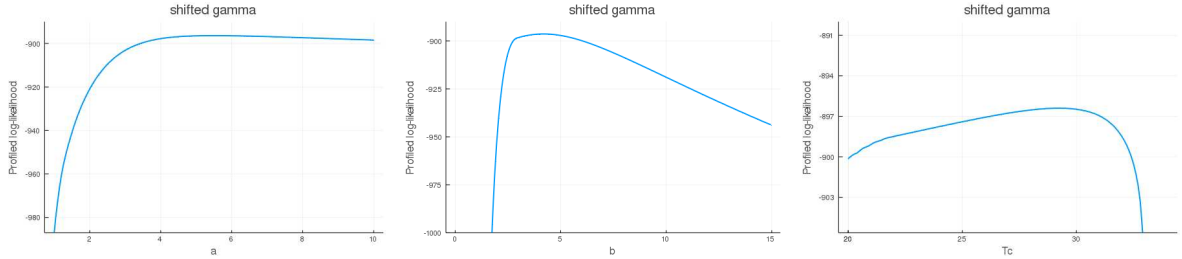


FIGURE 6 – Vraisemblance profilée, pour le modèle shifted gamma, des paramètres α (gauche), β (centre) et T_c (droite). Le maximum de vraisemblance, atteint sur l'ensemble des valeurs considérées pour les paramètres, vaut ici -896.40. Il est atteint pour $\hat{\alpha} = 5.45$, $\hat{\beta} = 4.21$ et $\hat{T}_c = 29.3$.

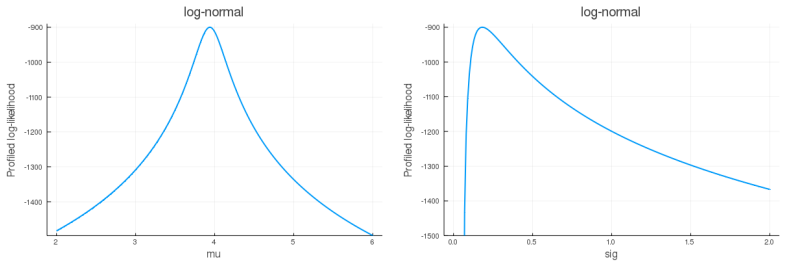


FIGURE 7 – Vraisemblance profilée, pour le modèle log-normal, des paramètres μ (gauche) et σ (droite). Le maximum de vraisemblance, atteint sur l'ensemble des valeurs considérées pour les paramètres, vaut ici -899.93. Il est atteint pour $\hat{\mu} = 3.94$, $\hat{\sigma} = 0.18$.

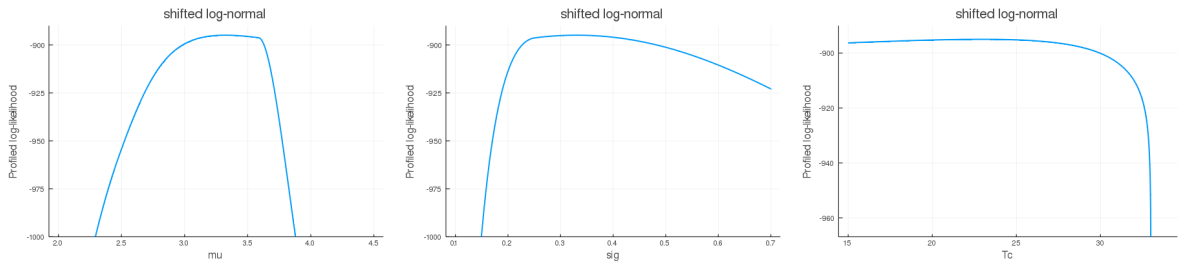


FIGURE 8 – Vraisemblance profilée, pour le modèle shifted log-normal, des paramètres μ (gauche), σ (centre) et T_c (droite). Le maximum de vraisemblance, atteint sur l'ensemble des valeurs considérées pour les paramètres, vaut ici -894.95. Il est atteint pour $\hat{\mu} = 3.325$, $\hat{\sigma} = 0.33$ et $\hat{T}_c = 22.9$.

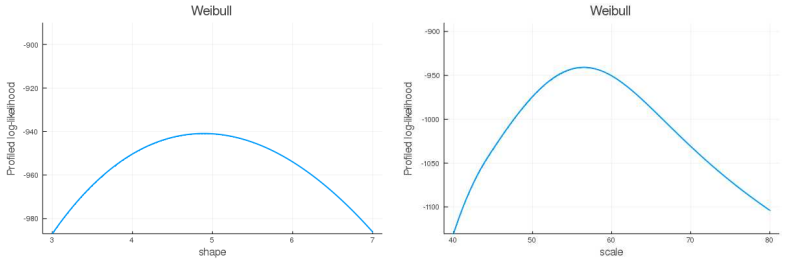


FIGURE 9 – Vraisemblance profilée, pour le modèle Weibull, des paramètres de forme (shape) α (gauche) et d'échelle (scale) λ (droite). Le maximum de vraisemblance, atteint sur l'ensemble des valeurs considérées pour les paramètres, vaut ici -940.92. Il est atteint pour $\hat{\alpha} = 4.9$ et $\hat{\lambda} = 56.5$.

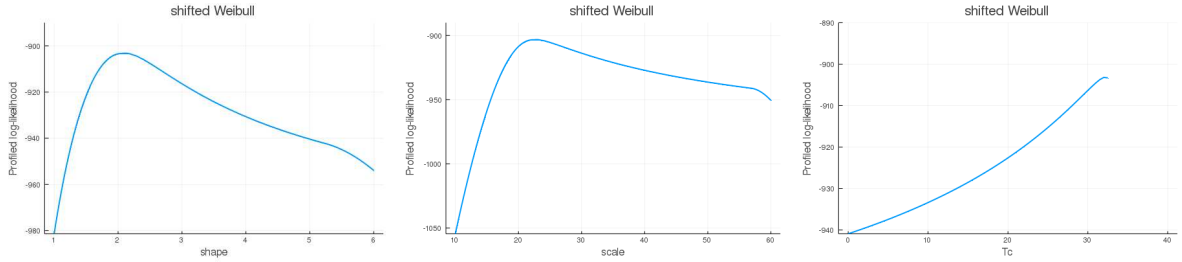


FIGURE 10 – Vraisemblance profilée, pour le modèle shifted Weibull, des paramètres de forme (shape) α (gauche), d'échelle (scale) λ (centre) et de shift T_c (droite). Le maximum de vraisemblance, atteint sur l'ensemble des valeurs considérées pour les paramètres, vaut ici -903.20. Il est atteint pour $\hat{\alpha} = 2.11$, $\hat{\lambda} = 22.9$ et $\hat{T_c} = 32.0$.

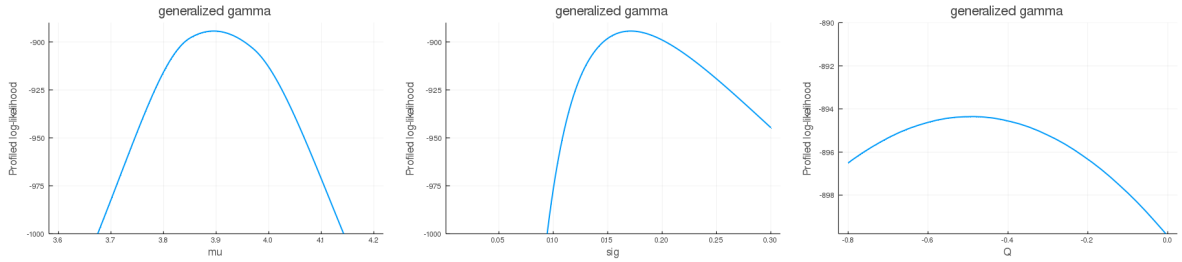


FIGURE 11 – Vraisemblance profilée, pour le modèle gamma généralisé, des paramètres μ (gauche), σ (centre) et Q (droite). Le maximum de vraisemblance, atteint sur l'ensemble des valeurs considérées pour les paramètres, vaut ici -894.36. Il est atteint pour $\hat{\mu} = 3.905$, $\hat{\sigma} = 0.170$ et $\hat{Q} = -0.493$.

Modèle $\mathcal{M}$	K	$\mathcal{L}(\hat{\boldsymbol{\theta}})$	AIC	BIC
Gamma généralisée	3	-894.36	1794.7	1805.2
Shifted log-normal	3	-894.95	1795.9	1806.4
Shifted gamma	3	-896.40	1798.8	1809.3
Log-normal	2	-899.93	1803.9	1810.9
Shifted Weibull	3	-903.20	1812.4	1822.9
Gamma	2	-904.72	1813.4	1820.5
Weibull	2	-940.92	1885.8	1892.9

TABLE 3 – Résultats de la procédure de sélection de modèle sur la base des critères AIC et BIC. Les modèles sont classés, du meilleur au moins bon, selon le critère AIC. $\mathcal{L}(\hat{\boldsymbol{\theta}})$ correspond au maximum de log-vraisemblance et K au nombre de paramètres à estimer pour le modèle $\mathcal{M}$. Le calcul du maximum de log-vraisemblance a été effectué numériquement à la section précédente. Les critères AIC et BIC aboutissent au même classement des modèles, sauf sur les modèles gamma et shifted Weibull, le critère BIC pénalisant plus fortement les modèles avec plus de paramètres.

Sur la figure 12, nous confrontons la distribution empirique des temps de première division des HSC* aux différents modèles (en choisissant comme vecteur de paramètre l'estimateur du maximum de vraisemblance). On voit que les modèles gamma généralisé, shifted log-normal et shifted gamma s'adaptent tous trois bien aux données, sans qu'il n'y ait de grandes différences entre eux. Le modèle gamma généralisé, sélectionné sur la base des critères AIC (et BIC), présente l'avantage d'être régulier, contrairement aux deux autres où la dépendance de support au paramètre T_c aboutit à un modèle non régulier [25, 26].

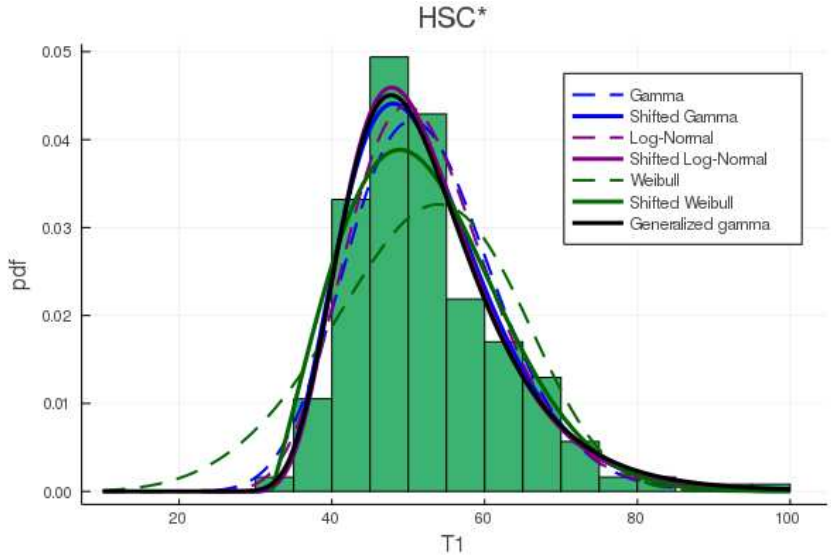


FIGURE 12 – Confrontation de la distribution empirique des temps de première division des HSC* aux sept modèles étudiés, en choisissant pour les paramètres ceux qui maximisent la vraisemblance. Les densité de probabilité (pdf) sont tracées, celles en pointillées correspondent à celles avec deux paramètres (sans shift).

3.6 Validation

Le modèle gamma généralisé a été sélectionné à l'issue de la procédure de sélection de modèle précédente. Visuellement (Fig. 12), ce modèle s'adapte assez bien aux données. La question qui se pose maintenant est de savoir si ce modèle, le meilleur parmi les sept étudiés, est un bon modèle. Pour y répondre, nous allons appliquer un test de qualité d'ajustement (goodness of fit) aux données. Pour cela, nous allons utiliser la méthode de Chen et Balakrishnan [27]. Cette méthode

n'est cependant pas adaptée à des données censurées par intervalles. Nous commençons alors par distribuer uniformément les observations d_i (pour $i \in \{1, \cdots, n\}$) sur les intervalles d'une heure associés. Par exemple, si nous avons $d_{i_1} = d_{i_2} = d_{i_3} = 39\text{h}$ et $d_i \neq 39\text{h}$ pour $i \notin \{i_1, i_2, i_3\}$, alors nous allons transformer les observations en $t_{i_1} = 39.25$, $t'_{i_2} = 39.5$ et $t_{i_3} = 39.75$. Ce nouveau jeu de données est noté $\mathcal{D}' = \{t_1, \cdots, t_n\}$. Toutes les valeurs sont différentes : $i \neq j \Rightarrow t_i \neq t_j$.
Notons que, pour les temps de première division des HSC*, nous n'avons qu'une seule cellule ι pour laquelle la division n'a pas été observée. Nous choisissons alors de fixer $t_\iota = 96\text{h}$.

Nous appliquons alors la méthode de Chen et Balakrishnan [27] pour tester l'hypothèse nulle H_0 selon laquelle $t_1, \cdots, t_n$ est un échantillon aléatoire issu d'une distribution gamma généralisée. Nous partons de l'estimateur $\hat{\boldsymbol{\theta}}$ calculé précédemment et calculons $v_i = F_{\hat{\boldsymbol{\theta}}}(t_i)$ avec F qui désigne ici la fonction de répartition de la loi gamma généralisée. Nous générons ensuite l'échantillon $\{u_1, \cdots, u_n\}$ par $u_i = \Phi^{-1}(v_i)$ où Φ^{-1} est l'inverse de la fonction de répartition de la loi normale centrée réduite, et effectuons un test de normalité (en l'occurrence, le test de Shapiro-Wilk [28]). Nous échouons à rejeter H_0 (avec une p-valeur $p = 0.2745$), c'est-à-dire que nous ne pouvons pas rejeter l'hypothèse selon laquelle les observations seraient bien distribuées suivant une loi gamma généralisée.

Une des limites de l'approche précédente est qu'elle repose sur une unique construction d'un jeu de données "continu" $\mathcal{D}'$ à partir de celui avec censure par intervalles $\mathcal{D}$. Plutôt que de répartir uniformément les temps de divisions sur l'intervalle qui leur correspond, on peut également choisir une répartition au hasard, et appliquer de nouveau la méthode de Chen et et Balakrishnan. On répète cette procédure 20 fois, le test étant rejeté à chaque fois (avec une p-valeur comprise entre 0.169 et 0.39 et une valeur médiane valant 0.269).

4 Application dans le cas des cellules progénitrices

Nous avons montré à la section 3 que la distribution des temps de première division des cellules HSC* pouvait être correctement modélisées par une loi gamma généralisée. Nous allons maintenant étudier si ce modèle est également adapté aux cellules progénitrices MPP et HPC, moins immatures que les HSC*.

Commençons par regarder si les observations issues de différents individus peuvent être regroupées ensemble. Le test de Kruskall-Wallis, appliqué au jeu de données correspondant aux MPP, échoue à rejeter l'hypothèse nulle selon laquelle les échantillons des différents individus seraient issus d'une même distribution ($p = 0.7416$). Il en va de même lorsque l'on applique le test de Mann-Whitney à chaque couple d'individus, justifiant ainsi de regrouper les observations ensemble dans le cas des MPP.
Dans le cas des HPC, au contraire, le test de Kruskall Wallis rejette l'hypothèse nulle ($p < 1e{-}4$) : au moins un échantillon domine stochastiquement les autres. En appliquant le test de Mann-Whitney à chaque couple d'individus, il ressort que les échantillons provenant des individus #51 et #57 ne devraient pas être regroupés entre eux ni avec les échantillons des autres individus. Il est intéressant de noter qu'il s'agit là des deux expérimentations conduites avec le cocktail GT, suggérant un effet significatif de ce cocktail de cytokines par rapport au cocktail Diff au niveau des temps de première division des HPC.
Néanmoins, pour simplifier, nous allons choisir de regrouper ensemble les temps de première division des HPC également. Alors que, dans le cas des HSC* et MPP, l'hypothèse de négliger l'hétérogénéité entre individus était justifiée mathématiquement, elle ne l'est plus ici, il s'agit d'une approximation et d'une limite de l'étude.

Nous calculons alors numériquement l'estimateur du maximum de vraisemblance dans le cas du modèle gamma généralisé, ainsi que pour les modèles gamma, log-normal et Weibull. Notons que nous n'étudions pas ici les versions shiftées de ces dernières qui, avec trois paramètres, ne

Modèle $\mathcal{M}$	MPP	HPC
Gamma généralisée	1498.1	1307.4
Gamma	1497.0	1306.8
Log-normale	1502.0	1305.4
Weibull	1526.6	1333.6

TABLE 4 – Calcul de l'AIC des modèles $\mathcal{M}$ dans le cas des observations issues soit de MPP (seconde colonne), soit de HPC (dernière colonne). Le modèle sélectionné est celui qui minimise ce critère, à savoir le modèle gamma dans le cas des MPP et le modèle log-normal dans le cas des HPC.

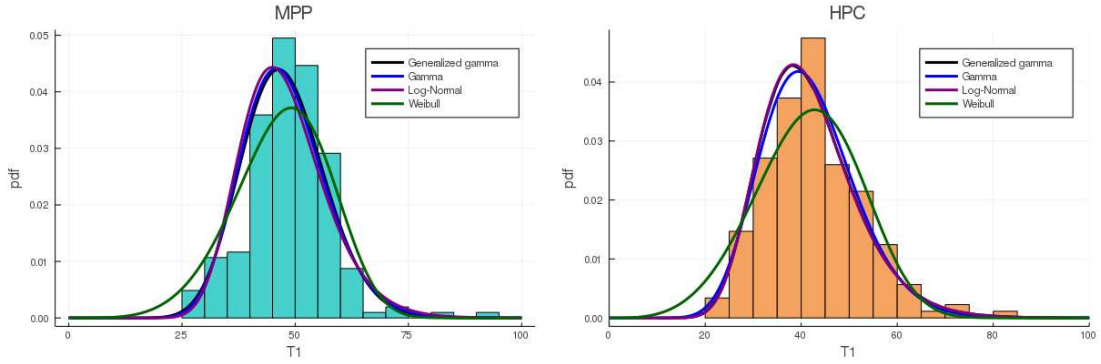


FIGURE 13 – Confrontation de la distribution empirique des temps de première division des MPP (gauche) et HPC (droite) aux modèles gamma généralisé, gamma, log-normal et Weibull (les paramètres des distributions sont ceux qui maximisent la vraisemblance).

donnaient pas de meilleurs résultats que le modèle gamma généralisé dans le cas des HSC*. Nous reportons les valeurs des AIC de chaque modèle dans le tableau 4 et confrontons les distributions empiriques (histogrammes) des temps de première division des MPP et HPC aux densités de probabilités des distributions candidates (avec comme choix de vecteur de paramètres l'estimateur du maximum de vraisemblance) sur la figure 13. Dans les deux cas, le modèle gamma généralisé n'est pas un bon candidat : il ne modélise pas mieux la distribution des temps de première division des MPP ou HPC que les modèles qu'il généralise, à savoir le modèle gamma et le modèle log-normal. Le premier est celui sélectionné dans le cas des MPP, le deuxième celui sélectionné dans le cas des HPC.

5 Discussion

Dans ce chapitre, nous avons étudié le choix du modèle gamma généralisé comme distribution paramétrique candidate pour modéliser la façon dont se distribuent les temps de première division des cellules hématopoïétiques exprimant le marqueur CD90 (cellules souches au sens immunophénotypique). Ce modèle généralise non seulement le modèle gamma, mais également celui log-normal et Weibull, par l'ajout d'un paramètre supplémentaire à estimer. Nous avons également considéré les versions avec décalage (shift) de ces derniers pour les comparer au modèle gamma généralisé, et ainsi étudier plusieurs distributions à trois paramètres.
Les données utilisées pour cette étude proviennent d'une expérience appelée Incucyte. Dans celle-ci, des cellules sont mises en culture, à une cellule par puits, et des prises d'images toutes les heures permettent de déterminer dans quel intervalle de temps chaque cellule s'est divisée. Les observations sont ainsi censurées par intervalle. De plus, on dispose de plusieurs échantillons provenant de dons de sang de cordon de différents individus. Nous avons regroupé ces échantillons ensemble, sur la base du résultat de deux tests statistiques : celui de Mann-Whitney et celui de Kruskall-Wallis. Nous avons ensuite exprimé la fonction de vraisemblance, avons calculé la vraisemblance profilée pour chaque modèle ainsi que l'estimateur du maximum de vraisemblance. Nous avons alors pu comparer le modèle gamma généralisé aux six autres, sur la base du critère d'information Bayésienne ainsi que le critère d'Akaike, nous conduisant à sélectionner le modèle gamma généralisé, légèrement meilleur que le modèle log-normale avec shift. En appliquant ensuite un test de qualité d'ajustement, basé sur la méthode de Chen et Balakrishnan [27], nous avons pu valider le choix de ce modèle. Néanmoins, les écarts dans les valeurs des AIC et BIC entre les modèles gamma généralisé, shifted log-normale et shifted gamma ne sont pas suffisants pour conclure que le modèle gamma-généralisé se distingue vraiment des deux autres, les trois modèles reproduisant les données de façon assez similaire.
Nous avons ainsi montré, avec notre méthode, que le modèle gamma généralisé était un bon candidat pour modéliser la distribution des temps de première division des HSC*. Nous n'avons cependant considéré qu'un nombre réduit de distributions paramétriques candidates, n'excluant pas la possibilité que d'autres modèles donnent de meilleurs résultats. Nous avons également supposé une homogénéité parmi les cellules de type HSC*, ce qui est une hypothèse simplificatrice, comme nous avons pu le voir au précédent chapitre, mais nécessaire au vu du nombre restreint de marqueurs utilisés ici. Nous avons trouvé que le modèle gamma généralisé n'était que légèrement meilleur que le modèle shifted log-normal. Par rapport à ce dernier, il présente cependant l'avantage d'être régulier, en l'occurrence de ne pas avoir une dépendance de son support à un paramètre. Cette propriété serait intéressante pour chercher à exprimer analytiquement l'estimateur du maximum de vraisemblance. Pour le moment, nous l'avons calculé numériquement ; en perspective nous pourrions envisager d'autres approches, par exemple basée sur l'utilisation de l'algorithme espérance-maximisation (EM - Expectation-Maximization) [29].
En appliquant la méthode dans le cas d'observations obtenues pour des cellules MPP et HPC, moins immatures que les HSC*, nous avons alors pu mettre en évidence des différences entre ces types cellulaires. Pour les échantillons correspondant aux HPC, nous avons dû cependant faire une hypothèse simplificatrice en regroupant les cellules ensemble, l'hypothèse d'une homogénéité inter-individus (ou entre expérimentations) étant rejetée, peut-être à cause de l'utilisation de différents cocktails de cytokines entre expérimentations, qui pourraient avoir un effet sur les temps de première division des HPC mais pas des MPP ou HSC* moins différenciées. Alors que, dans le cas des HSC*, le modèle gamma généralisé était privilégié (sur la base des critères AIC et BIC) par rapport aux modèles à deux paramètres qu'il généralise, ce n'est plus le cas pour les types MPP et HPC. On peut interpréter ce résultat par le fait que les cellules HSC* auraient une plus forte tendance à rester en quiescence que les cellules MPP et HPC. Nos résultats suggéreraient ainsi qu'un modèle gamma généralisé plutôt qu'un modèle gamma ou log-normal serait adapté pour décrire ce phénomène de temps de quiescence des cellules souches hématopoïétiques. Néanmoins, au vu des faibles différences obtenues entre les modèles lors du calcul de l'AIC, nos résultats ne sont pas significatifs. En perspective, il serait intéressant de voir ce que donneraient les résultats dans le cas de cellules mutées $JAK2^{V617F}$ ou $CALR^m$. On pourrait s'attendre *a*

priori à ce que les HSC* mutées soient moins quiescentes que celles Wild-Type, et ainsi à ce que le modèle gamma généralisé ne donne pas de meilleurs résultats comparé à un modèle avec moins de paramètres. Pour effectuer ce travail, il serait néanmoins nécessaire de travailler sur des échantillons de sang issus de dons de moelle osseuse ou de sang périphérique et non de sang de cordon. Or, les échantillons issus de dons de moelle présentent une forte hétérogénéité entre eux, qui pourraient s'expliquer par le fait qu'ils sont issus d'individus d'âge variable, et que l'âge a un impact sur le processus de division des cellules hématopoïétiques [30]. Ainsi, avant de pouvoir étudier le cas pathologique, il sera nécessaire de traiter l'hétérogénéité entre individus, ce qui pourra se faire par exemple par l'utilisation de modèles à effets mixtes [31] dans un cadre fréquentiste, ou l'utilisation d'un modèle hiérarchique en Bayésien [32]. Une première analyse de l'hétérogénéité entre les échantillons provenant de moelle osseuse, à partir d'un modèle simplifié et d'une estimation Bayésienne hiérarchique est proposée en annexe B de ce chapitre (voir annexes en ligne).

Dans ce chapitre, nous nous sommes concentrés sur l'étude du temps de première division des cellules souches et progénitrices lors de leur mise en culture. Ce temps inclut une durée de sortie de quiescence, notamment pour les HSC* comme le suggèrent nos résultats, mais également un temps d'adaptation au milieu, ce qui fait que ce temps est plus long que celui des divisions ultérieures, justifiant un traitement à part. Nous disposons également des observations des temps de seconde division. Néanmoins, lorsqu'une cellule d'un type donné se divise, ses cellules filles peuvent être d'un type différent de celui de leur mère (mais également entre elles). Ainsi, au delà de la première division, on ne peut pas étudier la prolifération sans étudier également le processus de différenciation. Ce sera l'objet du chapitre suivant.

Références

[1] Ron Sender and Ron Milo. The distribution of cellular turnover in the human body. *Nature medicine*, 27(1) :45–48, 2021.

[2] Henry Lee-Six, Nina Friesgaard Øbro, Mairi S Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J Osborne, Brian JP Huntly, Inigo Martincorena, Elizabeth Anderson, et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561(7724) :473–478, 2018.

[3] Jason Cosgrove, Lucie SP Hustin, Rob J de Boer, and Leïla Perié. Hematopoiesis in numbers. *Trends in Immunology*, 42(12) :1100–1112, 2021.

[4] A Bruce Lyons. Divided we stand : tracking cell proliferation with carboxyfluorescein diacetate succinimidyl ester. *Immunology and Cell Biology*, 77(6) :509–515, 1999.

[5] X-W Zhang, J Audet, JM Piret, and Y-X Li. Cell cycle distribution of primitive haematopoietic cells stimulated in vitro and in vivo. *Cell Proliferation*, 34(5) :321–330, 2001.

[6] Samuel Bernard, Laurent Pujo-Menjouet, and Michael C Mackey. Analysis of cell kinetics using a cell division marker : mathematical modeling of experimental data. *Biophysical journal*, 84(5) :3414–3424, 2003.

[7] Erika E Kuchen, Nils B Becker, Nina Claudino, and Thomas Höfer. Hidden long-range memories of growth and cycle speed correlate cell cycles in lineage trees. *Elife*, 9 :e51002, 2020.

[8] Koen Schepers, Erwin Swart, Jeroen WJ van Heijst, Carmen Gerlach, Maria Castrucci, Daoud Sie, Mike Heimerikx, Arno Velds, Ron M Kerkhoven, Ramon Arens, et al. Dissecting t cell lineage relationships by cellular barcoding. *The Journal of experimental medicine*, 205(10) :2309–2318, 2008.

[9] Shalin H Naik, Ton N Schumacher, and Leïla Perié. Cellular barcoding : a technical appraisal. *Experimental hematology*, 42(8) :598–608, 2014.

[10] JA Smith and L Martin. Do cells cycle ? *Proceedings of the National Academy of Sciences*, 70(4) :1263–1267, 1973.

[11] Arjun Raj and Alexander Van Oudenaarden. Nature, nurture, or chance : stochastic gene expression and its consequences. *Cell*, 135(2) :216–226, 2008.

[12] Mads Kaern, Timothy C Elston, William J Blake, and James J Collins. Stochasticity in gene expression : from theories to phenotypes. *Nature Reviews Genetics*, 6(6) :451–464, 2005.

[13] Hochan Cheon, Giulio Prevedello, Simone C Oostindie, Simon J Dovedi, Edwin D Hawkins, Julia Mary Marchingo, Susanne Heinzel, Ken R Duffy, and Philip D Hodgkin. Cyton2 : A model of immune cell population dynamics that includes familial instructional inheritance. *Frontiers in Bioinformatics*, page 50, 2021.

[14] Ken R Duffy, Cameron J Wellard, John F Markham, Jie HS Zhou, Ross Holmberg, Edwin D Hawkins, Jhagvaral Hasbold, Mark R Dowling, and Philip D Hodgkin. Activation-induced b cell fates are selected by intracellular stochastic competition. *Science*, 335(6066) :338–341, 2012.

[15] Alessandra Biffi, Eugenio Montini, Laura Lorioli, Martina Cesani, Francesca Fumagalli, Tiziana Plati, Cristina Baldoli, Sabata Martino, Andrea Calabria, Sabrina Canale, et al. Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. *Science*, 341(6148) :1233158, 2013.

[16] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[17] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260) :583–621, 1952.

[18] Edney W Stacy. A generalization of the gamma distribution. *The Annals of mathematical statistics*, pages 1187–1192, 1962.

[19] Christopher Cox, Haitao Chu, Michael F Schneider, and Alvaro Munoz. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in medicine*, 26(23) :4352–4374, 2007.

[20] Ross L Prentice. A log gamma model and its maximum likelihood estimation. *Biometrika*, 61(3) :539–544, 1974.

[21] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15) :1923–1929, 2009.

[22] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6) :716–723, 1974.

[23] Kenneth P Burnham and David R Anderson. A practical information-theoretic approach. *Model selection and multimodel inference*, 2, 2002.

[24] Schwarz Gideon et al. Estimating the dimension of a model. *The annals of statistics*, 6(2) :461–464, 1978.

[25] HaiYing Wang and Nancy Flournoy. On the consistency of the maximum likelihood estimator for the three parameter lognormal distribution. *Statistics & Probability Letters*, 105 :57–64, 2015.

[26] Hideo Hirose. Maximum likelihood parameter estimation in the three-parameter gamma distribution. *Computational statistics & data analysis*, 20(4) :343–354, 1995.

[27] Gemai Chen and N Balakrishnan. A general purpose approximate goodness-of-fit test. *Journal of Quality Technology*, 27(2) :154–161, 1995.

[28] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4) :591–611, 1965.

[29] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1) :1–22, 1977.

[30] M Carolina Florian, Markus Klose, Mehmet Sacma, Jelena Jablanovic, Luke Knudson, Kalpana J Nattamai, Gina Marka, Angelika Vollmer, Karin Soller, Vadim Sakk, et al. Aging alters the epigenetic asymmetry of hsc division. *PLoS biology*, 16(9) :e2003389, 2018.

[31] Marc Lavielle. *Mixed effects models for the population approach : models, tasks, methods and tools*. CRC press, 2014.

[32] A Gelman, JB Carlin, HS Stern, DB Dunson, A Vehtari, and DB Rubin. Bayesian data analysis, 3rd edn.(2013).

Chapitre 4
Modèle de prolifération et différenciation des cellules souches et progénitrices

Résumé

Afin de comprendre la dynamique de prolifération et différenciation des cellules hématopoïétiques souches et progénitrices, nous proposons un modèle mathématique dont nous estimons les paramètres à partir d'observations expérimentales provenant de deux expériences distinctes. Le modèle consiste en un processus stochastique à temps continu, où une cellule initiale souche ou progénitrice va se diviser pour donner naissance à deux cellules de types potentiellement différents de celui de leur mère, et ainsi de suite, pendant 96 heures. Pour étudier dans quelle mesure deux cellules sœurs auraient tendance à être d'un même type cellulaire et à se diviser de façon synchrone, nous avons modélisé une concordance et une synchronicité entre cellules sœurs.
Nous définissons un ensemble de statistiques descriptives pour effectuer une analyse de sensibilité puis définir une distance entre observations simulées et observations expérimentales. Nous recherchons alors le jeu de paramètres qui va minimiser cette distance à l'aide de l'algorithme d'optimisation CMA-ES.

Abstract

This chapter aims to understand the short-term proliferation and differentiation dynamics of hematopoietic stem and progenitor cells. For that purpose, we propose a mathematical model consisting of a continuous-time stochastic process, where an initial stem or progenitor cell will divide to give rise to two cells whose type might differ from their mother (and between them). The process of proliferation and differentiation continues for each daughter cell, and so on, until 96 hours. We also study to which extent sister cells might tend to be of the same type and/or divide synchronously. To study these hypotheses, we propose several declinations of our model to include a description of concordance and synchronicity. Model parameters are estimated from experimental observations from two separate experiments, namely the Incucyte and MultiGen assays. First, we define a set of summary statistics to perform a sensitivity analysis and then define a distance between simulated and experimental observations. We then calibrate the model using an optimisation algorithm, the CMA-ES algorithm.

Table des matières

1 Introduction

La production quotidienne de quelques milliards de monocytes, quelques dizaines de milliards de neutrophiles et quelques centaines de milliards de globules rouges chez l'homme [1] repose sur l'expansion importante d'un stock de cellules souches et progénitrices (HSPC, que nous définissons comme exprimant le marqueur CD34) qui représenteraient environ 2.5% des cellules mononucléées de la moelle osseuse [2, 3] (voir Cosgrove et al. pour une description quantitative de l'hématopoïèse [4]). Parmi ces cellules $CD34^+$, on distingue suivant l'expression de marqueurs CD90 et CD38 en cytométrie les cellules souches (HSC* qui sont définies ici comme étant $CD90^+CD38^-$), les progéniteurs multipotents (MPP - $CD90^-CD38^-$) et des progéniteurs au potentiel plus restreint (HPC - $CD90^-CD38^+$). Dans la conception standard de l'hématopoïèse, ces types cellulaires correspondent respectivement à des ensembles de cellules distincts et de plus en plus engagés vers une voie de différenciation.
Nous chercherons dans ce chapitre à modéliser la dynamique de prolifération et de différenciation de ces cellules, en particulier les effets de famille qui ont pu être observés par Tak et al. [5] chez la souris. On s'intéressera alors à comprendre dans quelle mesure deux cellules sœurs auraient tendance à être d'un même type cellulaire (concordance) et à se diviser en un temps très proche (synchronicité).
Pour cela, nous proposerons un modèle mathématique dont nous estimerons les paramètres à partir d'un riche jeu de données hétérogènes consistant en des observations provenant de deux expériences : l'Incucyte et le MultiGen.

2 Observations expérimentales

Le travail présenté dans ce chapitre se basera sur les observations issues de deux expériences : Incucyte et MultiGen. Les observations expérimentales ont été obtenues par Alessandro Donada, de l'équipe de Leïla Perié à l'Institut Curie.
Elles sont obtenues à partir d'échantillons de sang de cordon, échantillons purifiés pour conserver des cellules souches et progénitrices (i.e. des cellules $CD34^+$). Les échantillons proviennent de différents individus. Pour ce travail, nous regrouperons les données des expérimentations entre elles et faisons ainsi l'hypothèse qu'on peut négliger l'hétérogénéité entre individus, mais également entre conditions de culture (i.e. entre cocktails de cytokines utilisés : Diff ou GT). Cette hypothèse a été justifiée partiellement au chapitre précédent dans le cas des observations des temps de première division des HSC* et MPP. Elle est discutable s'agissant des expériences où la cellule initiale est de type HPC, avec dans ce cas une influence potentielle du cocktail de cytokines utilisé.

2.1 Expérience Incucyte

Dans cette expérience, décrite plus en détail au précédent chapitre, nous pouvons observer les temps de première, seconde ainsi que troisième divisions d'une cellule initiale d'un type $p \in \{\text{HSC*, MPP, HPC}\}$ donné, mise en culture dans un puits en présence d'un cocktail de cytokines pendant $T_{max} = 96$ heures. Nous appelons famille la colonie de cellules créées à partir de cette cellule initiale (voir discussion sur ce terme au paragraphe suivant). À partir de cette expérience, il est possible de connaître l'intervalle de temps (d'une durée de $\Delta t = 1\text{h}$) pendant lequel ont eu lieu les divisions d'intérêt. Il n'est cependant pas possible de connaître le type cellulaire des cellules survenant au cours de la dynamique (mise à part la cellule de départ).
Alors qu'au précédent chapitre, nous ne nous intéressions qu'à l'observation du temps de première division D_1 (borne inférieure de l'intervalle de temps pendant lequel a eu lieu la première division T_1), nous nous intéresserons ici également aux secondes divisions $D_{2,1}$ et $D_{2,2}$, bornes inférieures des intervalles de temps pendant lesquels ont eu lieu les divisions des deux cellules filles (temps compté depuis $t = 0$) avec $D_{2,1} \leq D_{2,2}$. Nous considérerons également l'information sur le nombre de cellules observées à 96h. Nous ne prendrons pas en compte ici les temps de

troisième division.
Les observations sont censurées par intervalles. Alors que la censure à droite due à la limite de temps d'observation était peu fréquente dans le cas des premières divisions (sur les échantillons issus de sang de cordon, qui sont ceux considérés ici), elle l'est naturellement plus dans le cas des secondes divisions. Nous excluons du jeu de données toute famille n'ayant pas au moins deux cellules à 96 heures.
Notre jeu de données consiste ainsi en l'observation de 246, 239 et 190 familles ayant respectivement comme cellule initiale une HSC*, MPP ou HPC. Par rapport au précédent chapitre, nous avons rajouté les observations de l'expérimentation #52 pour laquelle nous n'avons des observations que pour des MPP et HPC.
Nous noterons $\mathcal{I} = \bigcup_{p\in\{1,2,3\}} \mathcal{I}_p$ l'ensemble des familles observées par l'expérience Incucyte, en distinguant suivant la condition initiale, c'est-à-dire si le type cellulaire de départ est HSC* (indice 1), MPP (indice 2) ou HPC (indice 3).

2.2 Expérience MultiGen

L'expérience MultiGen, décrite initialement par Horton et al. [6] et utilisée par Tak et al. [5] pour l'étude des HSPC (cellules hématopoïétiques souches et progénitrices, incluant les HSC*, MPP et HPC) chez la souris, est schématisée sur la figure 1.
Dans cette expérience, des cellules sont mises en culture dans des puits, en présence d'un cocktail de cytokines (cocktail Diff introduit au chapitre 3), à quatre cellules par puits. On connaît le type de départ $p \in$ {HSC*, MPP, HPC} de chaque cellule, c'est-à-dire la condition initiale (Fig. 1-A). Pour un puits donné, les quatre cellules sont chacune marquées par des colorants fluorescents qui permettront de distinguer, au bout de $T_{max} = 96$ heures, les cellules issues de chacune d'entre elles (Fig. 1-D), mais également de connaître le nombre de divisions subies, comptées depuis la cellule initiale (Fig. 1-F). On emploiera parfois le terme de génération, la première génération étant celle de la cellule initiale. Une cellule en génération trois aura ainsi subi deux divisions.
Le principe du suivi des divisions repose sur la dilution du marqueur de fluorescence au cours de la mitose, avec une répartition équitable sur les deux cellules filles. Deux marqueurs de fluorescence sont utilisés ici, le CTV et le CFSE. Sur les quatre cellules de départ, une est marquée avec seulement du CFSE, l'autre avec seulement du CTV, les deux autres avec les deux marqueurs, mais dans des proportions différentes ce qui permettra de distinguer les familles de chacune d'elles (les proportions étant supposées rester les mêmes au cours des divisions).
L'ensemble des cellules qui dérivent d'un ancêtre commun est généralement appelé colonie. Néanmoins, ici, nous préfèrerons la définition de Tak et al. [5] et appellerons cet ensemble de cellules "famille". Le choix du terme n'est pas neutre, il met l'accent sur l'idée d'une transmission de certains éléments d'une génération à l'autre. En l'occurrence, Tak et al. ont montré chez la souris que les choix de différenciation (fate decisions en anglais) des cellules appartenant à une même famille pourraient être hérités de leur ancêtre.
Au bout de 96 heures, durée choisie en accord avec celle de l'expérience Incucyte, les cellules produites au cours de la dynamique de prolifération et différenciation sont récupérées. Elles subissent une étape de lavage puis sont marquées avec des marqueurs de surface, permettant alors de déduire leur type cellulaire (Fig. 1-E). Différents marqueurs de surface sont utilisés, à savoir le CD10, CD123, CD34, CD38, CD45 et CD90, ce qui permettrait en théorie de distinguer parmi les types cellulaires moins différenciés ($CD34^+CD38^-$) les HSC* des MPP et LMPP (progéniteurs communs lymphoïde-myéloïde), et parmi les cellules un peu plus différenciées ($CD34^+CD38^+$) les MEP (progéniteurs communs mégacaryocytaire-érythroïde), CMP (progéniteurs communs myéloïde) et GMP (progéniteurs granulocytaire-monocytaire). Pour simplifier - notamment parce que certains types cellulaires tels que les LMPP, GMP ou CMP étaient peu représentés - nous n'allons considérer ici que les marqueurs CD34, CD38 et CD90, et nous étudierons ainsi quatre types cellulaires : HSC*, MPP, HPC et $CD34^-$. Soulignons que le type cellulaire $CD34^-$, plus mature, est très hétérogène et inclut des cellules engagées vers des voies de différenciation différentes.

Pour définir le type cellulaire à partir de l'expression des marqueurs de surface, des seuils d'expression sont définis à partir de l'expression de cellules dans des puits contrôle (dans lesquels beaucoup plus de cellules, i.e. d'évènements en cytométrie, sont mesurés).
Lors de la récupération des cellules et des étapes expérimentales qui viennent ensuite, une partie des cellules du puits est perdue. On ne peut ainsi pas observer le contenu intégral du puits, mais seulement une fraction (Fig. 1-B et C). On appelle taux de récupération (recovery rate) la variable η égale au rapport du nombre de cellules récupérées par rapport au nombre de celles présentes dans le puits. Nous supposerons cette variable indépendante du nombre de cellules dans le puits.
Nous avons indiqué plus haut que quatre cellules distinctes (et différentiables) étaient déposées par puits. L'intérêt de ce choix est double. Il permet tout d'abord de multiplier par quatre le nombre de familles observées pour une seule plaque. Ensuite, cela permet de tester si la culture de quatre colonies dans un seul puits est équivalent à la culture de quatre colonies, chacune dans des puits distincts. Le cas contraire signifierait soit que le milieu de culture a un impact sur la dynamique (éventuellement à cause de conditions de culture non totalement homogènes), soit un effet des cellules les unes sur les autres, notamment entre familles, qui pourrait correspondre par exemple à des mécanismes de régulation. Nous n'étudions pas ces points ici, et faisons l'hypothèse que les colonies sont indépendantes entre elles. Sous cette hypothèse ainsi que celle selon laquelle le taux de récupération des cellules à 96 heures est indépendant du nombre de cellules réellement présentes, alors nous pouvons considérer qu'il est équivalent de considérer une expérience avec quatre fois plus de puits, dans lesquels une seule cellule serait mise en culture par puits. Nous nous placerons dans ce dernier cas, et proposons dans la suite un modèle à l'échelle de la famille.

Nous noterons $\mathcal{J} = \bigcup_{p\in\{1,2\}} \mathcal{J}_p$ l'ensemble des familles observées par l'expérience MultiGen, en distinguant suivant la condition initiale, c'est-à-dire si le type cellulaire de départ est HSC* (indice 1) ou MPP (indice 2). Notons que nous n'avons pas d'observations MultiGen ayant pour condition initiale une cellule de type HPC. Les données utilisées proviennent ici de deux expérimentations (la numéro 53 et la numéro 57) qui sont regroupées ensemble. Notre jeu de données MultiGen consiste alors en l'observation de 157 familles issues d'une HSC* et 163 familles issues d'un MPP.

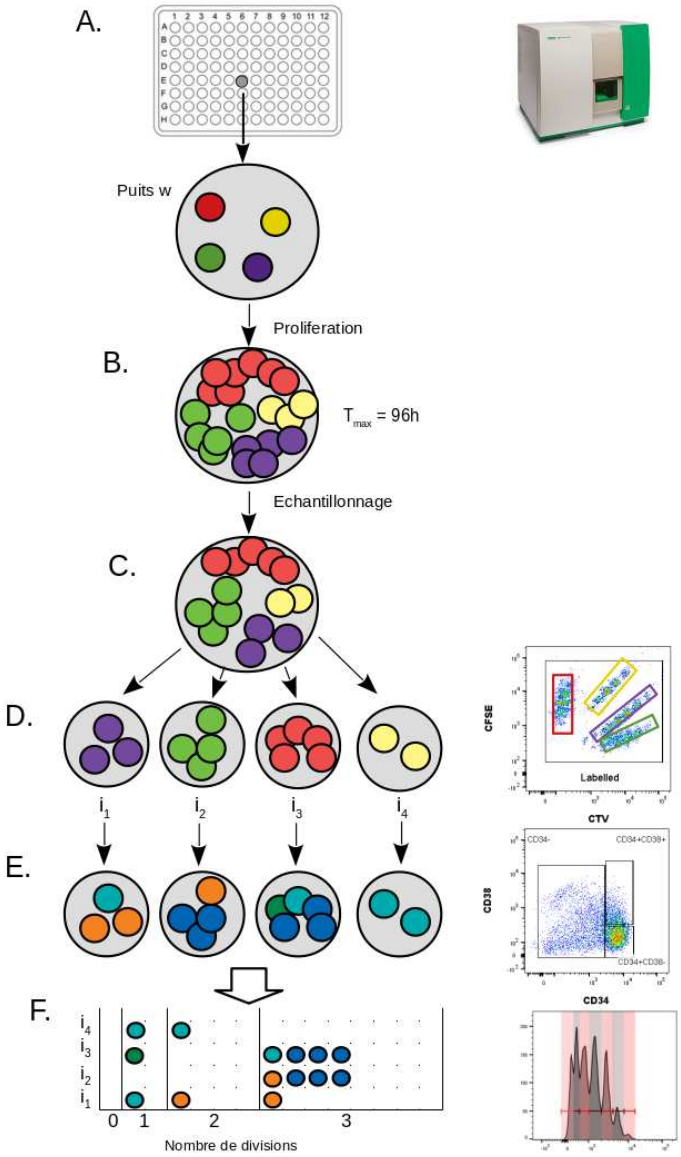


FIGURE 1 – Schéma de l'expérience MultiGen. Dans des plaques de culture de 96 puits, on dépose dans chaque puits (hormis les puits contrôle permettant de définir les seuils d'expression des marqueurs) quatre cellules d'un même type (A). À ces quatre cellules correspondront quatre familles distinctes (schématisées par quatre couleur différentes : rouge, jaune, vert pomme et violet). Les cellules seront mises en culture pendant 96 heures puis étudiées par cytométrie de flux. À 96 heures, les cellules initiales ont conduit à la création de colonies (familles) (B). Le contenu du puits est alors prélevé. Les cellules subissent un lavage puis un marquage ; un certain nombre de cellules seront perdues au cours de cette étape (échantillonnage) (C). On identifie pour les cellules échantillonnées la famille à laquelle elles appartiennent à partir de la concentration en deux marqueurs de fluorescence, CTV et CFSE (D). Les familles identifiées sont constituées de cellules de différents types parmi les HSC* (vert), MPP (cyan), HPC (orange) ou $CD34^-$ (bleu), qui pourront être identifiées par cytométrie de flux à partir de l'utilisation de marqueurs de surface (E). En étudiant également la dilution des marqueurs de fluorescence au cours des mitoses, on peut également déterminer le nombre de divisions subies pour chaque cellule (F).

3 Modèle

3.1 Formalisme

Nous allons présenter dans la suite une proposition de modèle permettant de décrire la dynamique de prolifération et différenciation des cellules souches et progénitrices. Ce modèle a été construit pour qu'on puisse en faire l'estimation de ses paramètres à partir des observations expérimentales décrites à la section 2 : il s'agit notamment d'un modèle stochastique de dynamique à court terme (96 heures), à temps continu, à l'échelle d'une famille, dans lequel chaque cellule ne peut être caractérisée que par un type discret $p \in \mathcal{P} = \{\text{HSC*, MPP, HPC, CD34}^-\}$.

Le système que nous modélisons sera entièrement caractérisé par les variable d'états N, $\boldsymbol{p} = (p_1, \cdots, p_N)$, $\boldsymbol{m} = (m_1, \cdots, m_N)$, et $\boldsymbol{T_A} = (T_{A,1}, \cdots, T_{A,N})$ (voir Fig. 2).
$N \geq 1$ correspond au nombre de cellules (d'une famille quelconque) ayant existé entre le début ($t = 0$) et la fin ($t = T_{max}$) de l'expérience. Les N cellules sont alors indexées par $j \in \{1, \cdots, N\}$ correspondant à leur ordre d'apparition (l'indice $j = 1$ correspondant à la cellule initiale).
Soit une cellule j donnée, elle est caractérisée par son type $p_j \in \mathcal{P} = \{\text{HSC*, MPP, HPC, CD34}^-\}$, l'indice de sa cellule mère $m_j \in \{0, \cdots, j-1\}$ (avec $m_1 = 0$) et le temps auquel elle est apparue $T_{A,j} \in [0, T_{max}]$ (avec $T_{A,1} = 0$).
Les indices correspondant à l'ordre d'apparition des cellules, nous avons :

$$\forall\ 1 \leq j \leq j' \leq N, T_{A,j} \leq T_{A,j'} \tag{1}$$

Une cellule ne pouvant se diviser qu'en deux et les deux cellules filles apparaissant au même instant, nous avons :

$$\forall\ 1 < j' < N, \text{ tel que } j' \text{ paire}, m_{j'} = m_{j'+1} \text{ et } T_{A,j'} = T_{A,j'+1} \tag{2}$$

Notons que nous faisons l'hypothèse qu'il n'y a pas de mort cellulaire ; N est nécessairement impair.

À toute valeur de N, $\boldsymbol{p}$, $\boldsymbol{m}$, et $\boldsymbol{T_A}$ respectant les contraintes des relations (1) et (2) correspond un état du système (Fig 2.A) dont on peut déduire le processus de prolifération et différenciation au cours du temps $t \in [0, T_{max}]$ (Fig. 2.C). En effet, à chaque temps t, on peut déduire l'ensemble $\mathcal{C}_t \subset \{1, \cdots, N\}$ de cellules présentes à cet instant (Fig. 2.B). Soit $j \in \{1, \cdots, N\}$, $j \in \mathcal{C}_t$ si et seulement si les deux conditions suivantes sont vérifiées :

1. $T_{A,j} \leq t$
2. $\forall\ j' \in \{j+1, \cdots, N\} : \big(m_{j'} = j\big) \Rightarrow \big(T_{A,j'} > t\big)$

Avec les définitions précédentes, il est ainsi équivalent de considérer le système ou le processus dynamique associé. Nous préférerons employer le terme de processus. Nous faisons dans la suite le choix d'un modèle stochastique ; N est alors une variable aléatoire et $\boldsymbol{p}$, $\boldsymbol{m}$, et $\boldsymbol{T_A}$ des vecteurs aléatoires (notons qu'ils sont de taille N). Modéliser ce processus reviendra à spécifier les lois de probabilité sous-jacentes.

Si par exemple nous avions choisi de modéliser les temps de divisions par des lois exponentielles, de décrire, conditionnellement au type cellulaire de la cellule mère, les types possibles des cellules filles et leur probabilités associées par une matrice de transition et enfin de supposer les cellules indépendantes entre elles, nous aurions alors été dans le cas d'un processus de branchement multi-types à temps continu, qui est un processus Markovien (voir par exemple [7]).

Dans les sections suivantes, nous décrirons plus en détails nos hypothèses de modélisation qui aboutissement à une spécification du modèle plus complexe que celle des processus de branchement, ce qui nécessitera alors d'étudier le modèle par des simulations de type Monte-Carlo, notamment pour en faire l'estimation des paramètres.

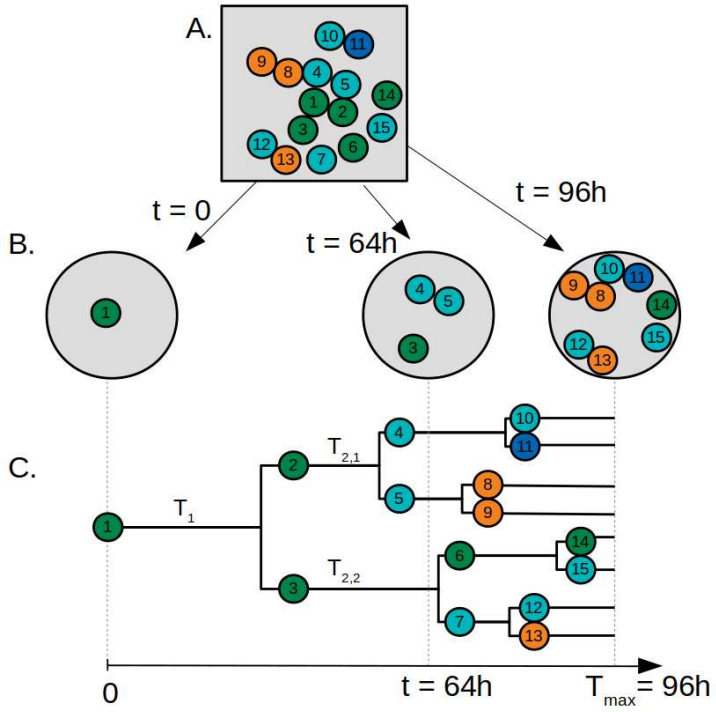


FIGURE 2 – Illustration du système étudié, avec comme valeurs pour les variables d'état : $N = 15$, $\boldsymbol{p} = (1,1,1,2,2,1,2,3,3,2,4,2,3,1,2)$ en associant les types HSC* (couleur verte), MPP (cyan), HPC (orange) et CD34$^-$ (bleu foncé) aux valeurs 1, 2, 3 et 4 respectivement, $\boldsymbol{m} = (0,1,1,2,2,3,3,5,5,4,4,7,7,6,6)$ et $\boldsymbol{T_A} = (0,34,34,60,60,65,65,68,68,72,72,76,76,90,90)$. À partir de l'état du système (A), on peut déduire l'état du système à un temps t quelconque (B) et reproduire le processus de prolifération et différenciation au cours du temps (C). T_1 va correspondre au temps de première division, et $T_{2,1}$ et $T_{2,2}$ aux temps de secondes divisions tels que $T_{2,1} \leq T_{2,2}$. Ici, $T_1 = T_{A,2} = T_{A,3}$, $T_{2,1} = T_{A,4} = T_{A,5}$ et $T_{2,2} = T_{A,6} = T_{A,7}$.

3.2 Différenciation

3.2.1 Transition entre types cellulaires

Dans ce chapitre, nous définissons la différenciation comme étant le passage d'un type cellulaire $p \in \mathcal{P}$ à un autre moins immature, l'ensemble $\mathcal{P}$ étant ainsi muni d'une relation d'ordre, avec, du type le plus immature au moins immature parmi ceux considérés :

1. HSC*
2. MPP
3. HPC
4. CD34$^-$

Nous confondrons ainsi dans la suite l'ensemble $\mathcal{P}$ avec l'ensemble $\{1,2,3,4\}$.
Nous négligeons l'hétérogénéité entre cellules d'un même type cellulaire, ce qui est une simplification (voir chapitre 2), particulièrement dans le cas des cellules CD34$^-$.
Nous faisons l'hypothèse qu'il ne peut y avoir différenciation que suite à une division, c'est-à-dire qu'une cellule ne changera pas de type cellulaire en cours de vie. Cette hypothèse est généralement justifiée [8] même si des observations ont pu faire état de différenciation cellulaire en l'absence de divisions [9, 10]. Enfin, nous supposons qu'il n'y a pas de phénomène de dé-différenciation, c'est-à-dire qu'une cellule ne pourra pas donner, en se divisant, une cellule d'un type plus immature.

Soit k l'une des deux cellules filles (quelconque) issues de la division de la cellule m_k. Son type p_k est une variable aléatoire discrète, prenant ses valeurs dans $\mathcal{P}$ avec la probabilité $\mathbb{P}[p_k = j | p_{m_k} =$

$i] = M_{i,j}$ avec $\mathbf{M} = (M_{i,j})_{1 \leq i,j \leq 4}$ la matrice de transition suivante :

$$\mathbf{M} = \begin{pmatrix} p_{1\to1} & p_{1\to2} & p_{1\to3} & 1 - p_{1\to1} - p_{1\to2} - p_{1\to3} \\ 0 & p_{2\to2} & p_{2\to3} & 1 - p_{2\to2} - p_{2\to3} \\ 0 & 0 & p_{3\to3} & 1 - p_{3\to3} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3}$$

En l'absence de concordance, c'est-à-dire si les deux cellules sœurs sont indépendantes quant au choix de leur type cellulaire, la probabilité d'avoir une division symétrique (i.e. que les deux cellules filles aient le même type que celui de leur mère) pour une cellule mère de type $a \in \mathcal{P}$ est donnée par $(p_{a\to a})^2$.
Nous noterons $I_c = 0$ l'hypothèse d'une absence de concordance.
Nous souhaitons confronter cette hypothèse à l'hypothèse alternative d'existence d'une concordance, que nous noterons $I_c = 1$ et présentons au paragraphe suivant.

3.2.2 Concordance

Tak et al. [5] ont montré, à partir d'expériences MultiGen réalisées sur des cellules souches et progénitrices de souris, une similarité dans les types cellulaires de cellules d'une même famille, c'est-à-dire descendant d'un même ancêtre, ce qui justifie d'étudier l'hypothèse d'une concordance dans notre modèle. Notons que notre définition de concordance diffère de celle de Tak et al. Nous choisissons de n'employer le terme de concordance qu'en lien avec le processus de différenciation, c'est-à-dire qu'en rapport avec le type cellulaire de cellules sœurs (ou plus généralement d'une même famille), alors que Tak et al. l'emploient également pour décrire une similarité dans le nombre de divisions (ou de générations) subies par des cellules d'une même famille. Or, nous préférons utiliser le terme de synchronicité lorsqu'il s'agira de décrire (et modéliser) des liens familiaux associés au processus de prolifération (voir § 3.3.3).

Pour modéliser cette concordance, nous allons introduire une dépendance entre les deux cellules sœurs quant au choix de leur type cellulaire.
Soient $k, k+1$ deux cellules sœurs issues de la division d'une cellule m_k.
Nous définissons alors la loi jointe pour leurs types cellulaires, conditionnement au type de leur cellule mère, de la façon suivante :

$$\mathbb{P}[p_k = i, p_{k+1} = j | p_{m_k} = a] = \rho_c p_{a\to i} \delta_{i,j} + (1 - \rho_c) p_{a\to i} p_{a\to j} \tag{4}$$

Avec $\rho_c \in [0,1]$ qui quantifie l'importance de la concordance et $\delta_{i,j}$ qui vaut 1 si $i = j$, 0 sinon.

On vérifie que loi marginale de p_k est la même que dans le cas sans concordance. En effet, on l'obtient en intégrant la relation (4) par rapport à p_{k+1} :

$$\begin{aligned} \mathbb{P}[p_k = i | p_{m_k} = a] &= \sum_{j=1}^{4} \mathbb{P}[p_k = i, p_{k+1} = j | p_{m_k} = a] \\ &= \rho_c p_{a\to i} + (1 - \rho_c) p_{a\to i} \sum_{j=1}^{4} p_{a\to j} \\ &= \rho_c p_{a\to i} + (1 - \rho_c) p_{a\to i} \\ &= p_{a\to i} \end{aligned}$$

Dans ce modèle où une hypothèse de concordance est faite ($I_c = 1$), nous avons maintenant une plus forte probabilité d'avoir une division symétrique que dans le cas $I_c = 0$. En effet, pour $p_{m_k} = a \in \mathcal{P}$, la probabilité de division symétrique est égale à :

$$\mathbb{P}[p_k = p_{k+1} = a | p_{m_k} = a] = \rho_c p_{a\to a} + (1 - \rho_c)(p_{a\to a})^2 \geq (p_{a\to a})^2 \tag{5}$$

On vérifie que le cas limite $\rho_c = 0$ correspond bien au cas sans concordance ($I_c = 0$). Dans le cas limite où $\rho_c = 1$, on voit dans la relation (4) que les cellules sœurs auront nécessairement le même type cellulaire.

Nous pouvons également exprimer la probabilité d'une division différenciée :

$$\mathbb{P}[p_k \neq a, p_{k+1} \neq a | p_{m_k} = a] = \sum_{i,j>a} \mathbb{P}[p_k = i, p_{k+1} = j | p_{m_k} = a] \tag{6}$$

La probabilité d'une division asymétrique se déduit alors des deux précédentes.

Dans notre modèle, les types cellulaires du couple de cellules sœurs ne dépendent que de celui de leur mère. Nous n'avons notamment pas introduit de dépendance avec le type d'un ancêtre plus lointain, ni avec le nombre de divisions subies, ni avec le temps mis par la cellule mère à se diviser. Ainsi, si l'on "oublie" la dynamique de prolifération, c'est-à-dire le temps mis entre chaque division, et que l'on ne considère que le processus discret extrait (qui correspond à notre processus de différenciation), alors on obtient un processus de Bienayme Galton-Watson multi-types que l'on pourrait étudier analytiquement.
Nous souhaitons cependant étudier également le processus de prolifération. Nous présentons dans la section suivante notre proposition de modélisation pour ce processus.

3.3 Division cellulaire

3.3.1 Première division

Nous introduisons la variable aléatoire T_1 correspondant au temps de première division, c'est à-dire au temps nécessaire à la cellule 1 pour se diviser. Dans le cas où $N > 1$, alors au moins une division aura lieu et T_1 correspondra au temps d'apparition de la cellule 2 : $T_{A,2}$ (ainsi que de la cellule 3), comme schématisé sur la figure 2.C. Dans le cas contraire, $T_{A,2}$ n'est pas défini. La façon dont se distribuent les temps de première division des cellules souches et progénitrices a été étudiée en détail dans le précédent chapitre. Nous reprendrons les résultats ici. Nous modéliserons ainsi le temps de première division des HSC* par une loi gamma généralisée :

$$T_{1,HSC*} \sim GG(\mu_{1,HSC*}, \sigma_{1,HSC*}, Q_{1,HSC*})$$

avec pour valeurs des paramètres (estimés au précédent chapitre) : $\mu_{1,HSC*} = 3.905, \sigma_{1,HSC*} = 0.170$ et $Q_{1,HSC*} = -0.493$.

Pour les temps de première division des MPP et HPC, nous choisissons de les modéliser tous deux par des lois log-normales :

$$T_{1,MPP} \sim \mathcal{LN}(\mu_{1,MPP}, \sigma_{1,MPP})$$

$$T_{1,HPC} \sim \mathcal{LN}(\mu_{1,HPC}, \sigma_{1,HPC})$$

avec $\mu_{1,MPP} = 3.843$, $\sigma_{1,MPP} = 0.196$, $\mu_{1,HPC} = 3.702$ et $\sigma_{1,HPC} = 0.236$. Ce choix est plus discutable dans le cas de la distribution des temps de première division des MPP qui, comme nous l'avions montré au précédent chapitre, était mieux décrite par un modèle gamma. Nous faisons néanmoins le choix de limiter les familles de distribution utilisées dans le modèle, et préférons ainsi l'usage la distribution log-normale à celle gamma. Nous discutons ce choix plus en détail dans le paragraphe suivant.

Concernant la loi log-normale, rappelons qu'une variable aléatoire $X \sim \mathcal{LN}(\mu, \sigma)$ si la variable $Y = \log(X)$ suit une loi normale d'espérance μ et de variance σ^2. L'espérance et la variance de X font alors intervenir les paramètres μ et σ dans leur expression :

$$\mathbb{E}[X] = \exp(\mu + \sigma^2/2)$$

$$\mathbb{V}[X] = \left(e^{\sigma^2} - 1\right) e^{2\mu+\sigma^2}$$

La médiane, elle, ne dépendra que de μ et vaudra $\exp(\mu)$.

Les observations des temps de première division issues de l'expérience Incucyte ayant été utilisées pour l'estimation des paramètres des lois précédentes, nous ne les utiliserons plus dans la suite.

3.3.2 Divisions ultérieures

À part pour la première division où le type de la cellule mère est connu, lorsque l'on étudie les divisions ultérieures, cette information est cachée. En effet, comme nous l'avons modélisé à la section 3.2, une cellule mère en se divisant donnera deux cellules filles dont chacune peut avoir un type parmi plusieurs possibles, selon différentes probabilités, qui dépendent du type de la cellule mère.
Nous l'avons vu au chapitre précédent, des cellules de types différents ont des distributions des temps de leur première division différentes. N'ayant pas de raison pour considérer qu'il n'en serait pas de même pour les divisions ultérieures, nous faisons alors l'hypothèse que la dynamique de prolifération, c'est-à-dire le temps nécessaire à une cellule pour se diviser, dépendra de son type. Soit $j > 1$ une cellule de type p_j, nous notons T_j la variable aléatoire correspondant à son temps de division (rappelons que nous faisons l'hypothèse qu'il n'y a pas de mort cellulaire). T_j dépendra ainsi de p_j.
Dans un modèle sans hypothèse de synchronicité, noté $I_s = 0$, nous ne faisons pas d'hypothèse quant à d'autres dépendances éventuelles. Notamment, nous ne considérons pas une relation de dépendance avec le temps mis par la cellule mère T_{m_j} pour se diviser, avec celui mis par sa cellule sœur $T_{m_{j'}}$ (cellule sœur j' telle que $j' \neq j$ et $m_{j'} = m_j$) pour se diviser ou avec le nombre de générations qui la séparent de la cellule 1 (cellule initiale).

Nous modélisons alors T_j par une loi log-normale dont les paramètres sont propres au type cellulaire $p_j \in \mathcal{P}$: $T_j|p_j \sim \mathcal{LN}(\mu_{p_j}, \sigma_{p_j})$. Nous faisons le choix de n'utiliser que la famille de lois log-normales, quel que soit le type de la cellule qui se divise. Nous aurions également pu faire le choix d'utiliser la loi gamma. Néanmoins, nous préférons la famille de lois log-normales, également utilisée par d'autres auteurs pour la modélisation de temps de division de cellules hématopoïétiques [11, 12, 13, 14], et qui présente une version multi-variée pratique d'emploi (pour modéliser la synchronicité, voir paragraphe suivant) et implémentée dans le langage Julia. Il aurait été intéressant de rester plus général et de considérer comme modèle unique la distribution gamma généralisée (que nous utilisons pour modéliser le temps de première division des HSC*), qui rappelons-le généralise à la fois le modèle gamma et le modèle log-normal. La difficulté dans ce cas aurait été celle de l'estimation des paramètres, puisque ce modèle - plus complexe - a un paramètre en plus à estimer.

Pour simplifier et restreindre le nombre de paramètres de notre modèle, nous faisons également l'hypothèse simplificatrice que le paramètre de la loi log-normale quantifiant la dispersion aléatoire autour de la médiane est le même quel que soit le type de la cellule qui se divise :

$$\sigma := \sigma_{\text{HSC*}} = \sigma_{\text{MPP}} = \sigma_{\text{HPC}} = \sigma_{\text{CD34}^-}$$

3.3.3 Synchronicité

Nous avons présenté ci-dessus une description du processus de prolifération sous l'hypothèse $I_s = 0$ d'absence de synchronicité. Or, de nombreux résultats tendent à montrer une forte corrélation dans les temps de division de cellules sœurs (mais pas avec celui de la cellule mère). Nous noterons $I_s = 1$ l'hypothèse selon laquelle il y aurait de la synchronicité entre cellules sœurs. Nous ne modéliserons une synchronicité qu'entre cellules sœurs. Notamment, nous ne considérerons pas ici de synchronicité entre cellules cousines.

Soient k, k' deux cellules sœurs, dont les temps de division T_k et $T_{k'}$ sont modélisés par des variables aléatoires. Sous l'hypothèse $I_s = 0$, les variables $T_k|p_k$ et $T_{k'}|p_{k'}$ étaient indépendantes, distribuées suivant une loi log-normale de paramètres (μ_{p_k}, σ) et $(\mu_{p_{k'}}, \sigma)$ respectivement.

Nous introduisons maintenant une relation de dépendance entre ces deux variables, tout en faisant en sorte que leurs lois marginales restent les mêmes que précédemment :

$$(T_k, T_{k'})|p_k, p_{k'} \sim MV\text{-}\mathcal{LN}\left((\mu_{p_k}, \mu_{p_{k'}}), \mathbf{\Sigma}\right) \tag{7}$$

où MV-$\mathcal{LN}$ correspond à une loi log-normale bivariée avec :

$$\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho_s \\ \rho_s & 1 \end{pmatrix}$$

et $\rho_s \in [0, 1]$ un paramètre qui quantifie l'importance de la synchronicité.
Nous faisons ainsi l'hypothèse qu'il y a synchronicité entre toutes cellules sœurs, même de types cellulaires différents, et que la synchronicité (c'est-à-dire, le paramètre ρ_s) ne dépend pas du type cellulaire. En particulier, avec ce modèle, deux cellules sœurs CD34$^-$ resteraient synchrones. Néanmoins, cette population de cellules est très hétérogène, comportant potentiellement des cellules déjà engagées vers certaines lignées cellulaires, ce qui fait que l'hypothèse d'une synchronicité entre cellules CD34$^-$ ne serait peut-être plus justifiée.

3.4 Modèle d'observations

L'état de notre système (ou processus) est entièrement caractérisé par N, $\boldsymbol{p}$, $\boldsymbol{m}$ et $\boldsymbol{T_A}$, variables (et vecteurs) aléatoires dont nous avons spécifié ci-dessus les relations de dépendance. Cependant, ces variables d'état ne sont pas directement observables : nous n'aurons accès à l'état du système qu'au travers des variables d'observation.

L'objectif, au travers les observations, est de comprendre le comportement du système, c'est-à-dire - puisque nous avons proposé un modèle paramétrique - d'estimer les valeurs des paramètres. Nous ferons l'estimation des paramètres du modèle en considérant que nous observons plusieurs réalisations aléatoires (considérées indépendantes et identiquement distribuées) de notre processus stochastique. Pour rappel, le système est décrit à l'échelle d'une famille issue d'une cellule unique dont nous connaissons le type à $t = 0$ (condition initiale). Nous observerons donc plusieurs familles. Une famille pourra soit être observée par l'expérience Incucyte, soit par l'expérience MultiGen. Nous indexons par $i \in \mathcal{I}$ et $j \in \mathcal{J}$ nos observations avec $\mathcal{I}$ et $\mathcal{J}$ qui correspondent à l'ensemble des $N_{\mathcal{I}}$ et $N_{\mathcal{J}}$ familles observées par les expériences Incucyte et MultiGen respectivement. La condition dite initiale est supposée connue sans incertitude, c'est-à-dire que nous avons l'information sur le type de la cellule 1 (p_1) et supposons $T_A = 0$ ce qui revient à négliger tout ce qui a pu influencer le devenir de la cellule avant le début de l'expérience.
Nous noterons alors $\mathcal{I} = \bigcup_{p \in \{1,2,3\}} \mathcal{I}_p$ (et de même pour $\mathcal{J}$) en différenciant les familles observées suivant si elles proviennent d'une HSC*, MPP ou HPC. Notons que nous n'avons pas d'observations commençant par une cellule CD34$^-$.

Si le système est observé via l'expérience Incucyte, nos variables d'observations seront D_1, $D_{2,1}$ et $D_{2,2}$ qui correspondent aux bornes inférieures des intervalles pendant lesquels ont eu lieu respectivement la première division (T_1) et les secondes divisions $T_{2,1}$ et $T_{2,2}$. Les observations étant également censurées à droite à cause de la limite du temps d'observation T_{max}, au cas où la division aurait lieu passé ce temps, ces variables seraient affectées à la valeur T_{max} (qui est la borne inférieure de l'intervalle $]T_{max}, +\infty]$). À part la censure, nous ne considérons pas d'incertitudes liées à la mesure de ces temps de division. Nous observons également $N(T_{max}) = \text{card}(\mathcal{C}_{T_{max}})$, c'est-à-dire le nombre de cellules présentes à T_{max}, que nous considérons estimé sans incertitude.

Si le système est observé via l'expérience MultiGen, il sera alors observé à $t = T_{max}$. Il y a un bruit d'échantillonnage : nous n'avons des informations que pour un ensemble de cellules $\hat{\mathcal{C}}_{T_{max}} \subset \mathcal{C}_{T_{max}}$. Nous faisons l'hypothèse que :

$$\mathbb{P}[i \in \hat{\mathcal{C}}_{T_{max}} | i \in \mathcal{C}_{T_{max}}] = \eta \tag{8}$$

avec $\eta \in [0, 1]$ appelé taux d'échantillonnage (recovery rate), qu'on suppose constant et indépendant de $\text{card}(\mathcal{C}_{T_{max}}) = N(T_{max})$.
Dans ce cas :

$$\mathbb{E}\left[\text{card}(\hat{\mathcal{C}}_{T_{max}})\right] = \eta\, \mathbb{E}\left[\text{card}(\mathcal{C}_{T_{max}})\right]$$

Nous pouvons alors estimer η par :

$$\eta = \mathbb{E}\left[\frac{\frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \text{card}(\hat{\mathcal{C}}_{T_{max},j})}{\frac{1}{N_{\mathcal{I}}} \sum_{i \in \mathcal{I}} \text{card}(\mathcal{C}_{T_{max},i})}\right] \tag{9}$$

et trouvons à partir des observations, en considérant $N_{\mathcal{I}}$ et $N_{\mathcal{J}}$ suffisamment grands, que $\eta = 0.71$.

Pour chaque cellule échantillonnée $k \in \hat{\mathcal{C}}_{T_{max}}$, nous observons les variables aléatoires correspondant à son type p_k ainsi qu'à son nombre de divisions n_k comptées depuis la cellule initiale, avec $n_1 = 0$ (voir Fig. 1.D). Avec la dilution des marqueurs de fluorescence, passé un certain nombre de divisions, nous considérons ne plus pouvoir les discerner. Toute cellule au-delà de la sixième génération (qui correspond à la génération maximale observée sur notre jeu de données) sera alors assignée à cette génération. Nous ne considérons pas d'autres incertitudes sur ces observations.

Finalement, les incertitudes liées aux observations sont associées soit à une censure par intervalles dans le cas des données Incucyte, soit à un échantillonnage dans le cas MultiGen.
À partir des variables d'observations précédentes et de leurs $N_{\mathcal{I}}$ et $N_{\mathcal{J}}$ réalisations, nous allons dans la suite chercher à estimer les paramètres de notre modèle. Le tableau 1 récapitule les paramètres impliqués dans notre modèle. Nous notons $\boldsymbol{\theta}$ le vecteur de paramètres à estimer :

$$\boldsymbol{\theta} = (p_{1\to1},\ p_{1\to2},\ p_{1\to3},\ p_{2\to2},\ p_{2\to3},\ p_{3\to3},\ I_c,\ \rho_c,\ \mu_{HSC*},\ \mu_{MPP},\ \mu_{HPC},\ \mu_{CD34^-},\ \sigma,\ I_s,\ \rho_s)$$

4 Statistiques descriptive et analyse de sensibilité

4.1 Statistiques descriptives

Notre modèle étant complexe, et l'état du système modélisé ne pouvant être que partiellement observé, il n'est pas possible d'exprimer une vraisemblance que nous pourrions alors maximiser pour obtenir un estimateur de nos paramètres.
Pour l'estimation de nos paramètres, nous nous baserons alors sur l'utilisation de statistiques descriptives (summary statistics en anglais) exprimées à partir des variables d'observation. Ces statistiques descriptives doivent alors pouvoir nous renseigner sur le processus étudié, c'est-à-dire sur le vecteur de paramètres $\boldsymbol{\theta}$ du modèle. Par exemple, Y_0, définie de la façon suivante :

$$Y_0 = \frac{\frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \text{card}(\hat{\mathcal{C}}_{T_{max},j})}{\frac{1}{N_{\mathcal{I}}} \sum_{i \in \mathcal{I}} \text{card}(\mathcal{C}_{T_{max},i})} \tag{10}$$

est une statistique descriptive, exprimée en fonction de certaines variables d'observation. On peut donc l'évaluer à partir des données expérimentales, ce qui donne ici : $\hat{y}_0 = 0.71$. $\mathbb{E}[Y_0]$ nous renseigne sur le processus en étant une fonction des paramètres. En l'occurrence, nous avons montré au paragraphe précédent que $\mathbb{E}[Y_0] = \eta$, ce qui nous permettait d'estimer $\eta \approx 0.71$.

Paramètre	Valeur	Prior	Référence
$p_{1\to1}$	à estimer	$\in [0,1]$	
$p_{1\to2}$	$p_{1\to2} = 1/3 p_{1\to1}$		Simplification eq. (13)
$p_{1\to3}$	$p_{1\to3} = p_{1\to2}$		Simplification eq. (13)
$p_{2\to2}$	à estimer	$\in [0,1]$	
$p_{2\to3}$	$p_{2\to3} = 1/2(1 - p_{2\to2})$		Simplification eq. (14)
$p_{3\to3}$	à estimer	$\in [0,1]$	
I_c	à estimer	$\in \{0,1\}$	
ρ_c	à estimer	$\in [0,1]$	
$\mu_{1,HSC*}$	3.905		Chapitre 3
$\sigma_{1,HSC*}$	0.17		Chapitre 3
$Q_{1,HSC*}$	-0.493		Chapitre 3
$\mu_{1,MPP}$	3.843		Chapitre 3
$\sigma_{1,MPP}$	0.196		Chapitre 3
$\mu_{1,HPC}$	3.702		Chapitre 3
$\sigma_{1,HPC}$	0.236		Chapitre 3
μ_{HSC*}	à estimer	$\in [2.2, 3.2]$	
μ_{MPP}	$\mu_{MPP} = \mu_{HSC*}$		Simplification eq. (12)
μ_{HPC}	à estimer	$\in [2.2, 3.2]$	
μ_{CD34^-}	à estimer	$\in [2.2, 3.2]$	
σ	à estimer	$\in [0.1, 0.4]$	
I_s	à estimer	$\in \{0,1\}$	
ρ_s	à estimer	$\in [0,1]$	
T_{max}	96h		Expérience
η	0.71		Calcul (9)

TABLE 1 – Liste des paramètres du modèle. La colonne "Valeur" indique si le paramètre sera à estimer, si il est fixé à une certaine valeur ou si il est lié à d'autres paramètres. Dans le cas où le paramètre est à estimer, nous indiquons dans la colonne "Prior" l'intervalle de valeurs autorisées. Dans le cas contraire, nous indiquons dans la colonne "Référence" la référence du choix de la valeur. Les paramètres pour lesquels il est indiqué "Simplification" sont ceux qui étaient considérés initialement comme degré de liberté et qui ont été fixés constant suite à l'analyse de sensibilité présentée au §4.3. Les intervalles de valeurs pour les paramètres μ_p ($p \in \mathcal{P}$) et σ sont choisis à partir d'une première analyse approximative basée sur le calcul de la médiane de l'échantillon $(D_{2,2,i} - D_{1,i})_{i\in\mathcal{I}}$.

De façon générale, on ne pourra pas exprimer directement les statistiques descriptives (leur espérance) en fonction des paramètres. Nous les calculerons numériquement, en simulant un grand nombre de fois le processus (voir § 4.2), pour différents jeux de paramètres, et les confronterons aux valeurs obtenues à partir des mesures expérimentales.
Nous noterons Y_i la variable aléatoire, $\hat{y}_i$ sa valeur évaluée sur les observations expérimentales, $y_i(\boldsymbol{\theta})$ la valeur évaluée sur un certain nombre de simulations du modèle, pour un jeu de paramètre $\boldsymbol{\theta}$ donné.
Une statistique descriptive $Y_i(\boldsymbol{\theta})$ pourra être comprise comme étant une sortie de notre modèle, fonction du vecteur de paramètres $\boldsymbol{\theta}$, que nous pourrons confronter à une valeur expérimentale $\hat{y}_i$, notamment pour trouver le jeu de paramètres $\hat{\boldsymbol{\theta}}$ qui minimise une certaine distance entre toutes les statistiques descriptives considérées, par exemple la norme L^2 :

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_i |\mathbb{E}[Y_i(\boldsymbol{\theta})] - \hat{y}_i|^2$$

Avant de définir une distance nous permettant de faire l'estimation des paramètres de notre modèle, il nous faut d'abord définir un ensemble de statistiques descriptives dont on souhaite qu'il soit suffisant pour caractériser notre processus. À partir de ces statistiques descriptives,

considérées comme les sorties de notre modèle, nous pourrons effectuer une analyse de sensibilité pour éventuellement réduire le nombre de paramètres à estimer dans notre modèle. Il nous faudra ensuite sélectionner un nombre restreint de statistiques descriptives sur lesquelles construire une distance à nos observations.
Nous définissons les statistiques descriptives soit à partir des variables d'observation associées à l'expérience Incucyte, soit à partir de celles associées à l'expérience MultiGen. De plus, nous faisons également la distinction entre familles suivant leur type cellulaire de départ. Nous avons abouti à la définition de 67 statistiques descriptives ($\times 3$ lorsque l'on considère les trois types cellulaires possible pour les cellules de départ) dont la liste totale est donnée en annexe A.1 disponible en ligne [1]. Nous en présentons quelques-unes dans la suite.

4.1.1 Incucyte

Pour un type $p \in \{1, 2, 3\}$ de départ, nous présentons ci-dessous, à titre d'exemple, quelques statistiques descriptives exprimées à partir des variables d'observation associées à l'expérience Incucyte. La liste complète est disponible en annexe A.1.1.

Y_{12}^p correspond à la moyenne du temps de seconde division :

$$Y_{12}^p = \frac{1}{N_{\mathcal{I}_p}} \sum_{i \in \mathcal{I}_p} \frac{1}{2} \left(D_{2,1,i} + D_{2,1,i} \right)$$

Y_{17}^p correspond à la moyenne de $(D_{2,1,i} - D_{1,i})_{i \in \mathcal{I}_p}$:

$$Y_{17}^p = \frac{1}{N_{\mathcal{I}_p}} \sum_{i \in \mathcal{I}_p} \left(D_{2,1,i} - D_{1,i} \right)$$

Y_{18}^p correspond à l'écart type de $(D_{2,1,i} - D_{1,i})_{i \in \mathcal{I}_p}$:

$$Y_{18}^p = \sqrt{\frac{1}{N_{\mathcal{I}_p} - 1} \sum_{i \in \mathcal{I}_p} \left(D_{2,1,i} - Y_{17}^p \right)^2}$$

4.1.2 MultiGen

Pour un type $p \in \{1, 2\}$ de départ, nous présentons ci-dessous, à titre d'exemple, quelques statistiques descriptives exprimées à partir des variables d'observation associées à l'expérience MultiGen et utilisées dans ce travail. La liste complète est disponible en annexe A.1.2.
Notons que nous n'utilisons pas ici de données correspondant à des expériences MultiGen faites avec des HPC comme condition initiale.

Y_{25}^p correspond au nombre moyen de cellules à avoir fait une division, sur les cellules échantillonnées correspondant à celles observées dans l'expérience MultiGen :

$$Y_{25}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \sum_{k \in \hat{\mathcal{C}}_{T_{max,j}}} 1_{n_k = 1}$$

On définit Y_{45}^p comme étant la proportion de familles dont les cellules observées sont réparties sur exactement trois générations :

$$Y_{45}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} 1_{\sum_n X_{j,n} = 3}$$

1. au lien suivant : https ://gitlab-research.centralesupelec.fr/2012hermangeg/supplementary-material-phd

où nous avons introduit $\boldsymbol{X_j}$, vecteur aléatoire de taille 6, tel que, pour une famille $j \in \mathcal{J}$ et un nombre de division $n \in \{0, \cdots, 5\}$:

$$X_{j,n} = \begin{cases} 1 & \text{si } \exists k \in \hat{\mathcal{C}}_{T_{max,j}}, n_k = n \\ 0 & \text{sinon} \end{cases}$$

Y_{61}^p est définie comme la proportion de familles à n'avoir que des cellules de type CD34$^-$ (parmi celles observées) :

$$Y_{61}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \prod_{k \in \hat{\mathcal{C}}_{T_{max,j}}} 1_{p_k=4}$$

4.2 Simulations

Nous avons défini un ensemble de statistiques descriptives exprimées à partir de nos variables d'observation. Ces statistiques descriptives dépendent des paramètres d'une façon complexe, qu'on ne peut en général pas obtenir analytiquement. Pour trouver, pour chaque statistique descriptive Y_i, à quels paramètres elle est sensible, c'est-à-dire trouver les paramètres qui influeront fortement sa valeur, nous devons commencer par échantillonner des jeux de paramètres dans l'espace des paramètres, simuler pour chacun un grand nombre de fois notre processus, puis calculer ensuite nos statistiques descriptives pour chaque type cellulaire de départ. Nous pourrons alors analyser les valeurs prises par nos statistiques descriptives et voir si les valeurs expérimentales peuvent en effet être atteintes par le modèle, pour au moins un jeu de paramètre donné. Nous pourrons ensuite mener une analyse de sensibilité pour estimer quels paramètres ont peu d'impact sur les sorties du modèles. Nous détaillons cette procédure dans la suite.

Nous souhaitons échantillonner un nombre N_{LHS} de vecteurs de paramètres $(\boldsymbol{\theta}_i)_{1 \leq i \leq N_{LHS}}$, avec N_{LHS} suffisamment grand pour que l'espace des paramètres soit bien couvert par notre échantillonnage. Nous avons choisi $N_{LHS} = 10,000$. Notre espace de paramètres étant de grande dimension (dimension 13), il n'est pas envisageable d'échantillonner à partir d'une grille régulière de cette espace. Nous choisissons alors d'échantillonner avec une méthode d'échantillonnage par hyper-cube latin, en l'occurrence celle de Urquhart et al. [15] basée sur les travaux de Bates et al. [16], qui permet de couvrir de manière adéquate l'espace des paramètres.

Pour chaque jeu de paramètres $\boldsymbol{\theta}_i$ nous simulons notre modèle un nombre $3 \times N_s$ de fois. À une simulation correspond une réalisation de notre processus (ou un état de notre système) obtenue pour un jeu de paramètres $\boldsymbol{\theta}_i$ et une cellule de départ p donnés.
Une simulation nous permet donc d'avoir l'information totale sur le système, c'est-à-dire que nous simulons des valeurs pour les variables d'état $N, \boldsymbol{p}, \boldsymbol{m}$ et $\boldsymbol{T_A}$ à partir desquelles nous pouvons calculer les variables d'observation qui seront comparables aux observations expérimentales.
Nous reproduisons les simulations un grand nombre de fois (N_s) pour calculer nos statistiques descriptives. Idéalement N_s doit être choisi suffisamment grand pour que la réalisation de Y_i (notée y_i) soit proche de la valeur théorique $\mathbb{E}[Y_i]$, sans être trop grand auquel cas les simulations deviendraient trop coûteuses en temps de calcul. Notons que nous n'avons pas nécessairement à choisir $N_s = N_{\mathcal{I}}$ ou $N_s = N_{\mathcal{J}}$ (les nombres de familles observées expérimentales par l'expérience Incucyte et MultiGen respectivement). Nous illustrons sur la figure 4 l'impact du choix de la valeur N_s. Nous choisissons $N_s = 1,000$, de l'ordre de grandeur de $N_{\mathcal{I}} = 675$, mais trois fois supérieur à $N_{\mathcal{J}} = 320$.
Une analyse plus fine devrait être conduite pour choisir N_s qui soit un bon compromis entre des statistiques descriptives y_i qui soient proches de $\mathbb{E}[Y_j]$, quelque soit la valeur des paramètres, et un temps de calcul pas trop élevé.
À titre d'illustration, il a fallu 5 heures de calculs (non parallélisés, sur un processeur Intel Xeon Gold 6230 20C @ 2.1GHz) pour réaliser toutes les simulations (pour l'ensemble des jeux de paramètres échantillonnés).

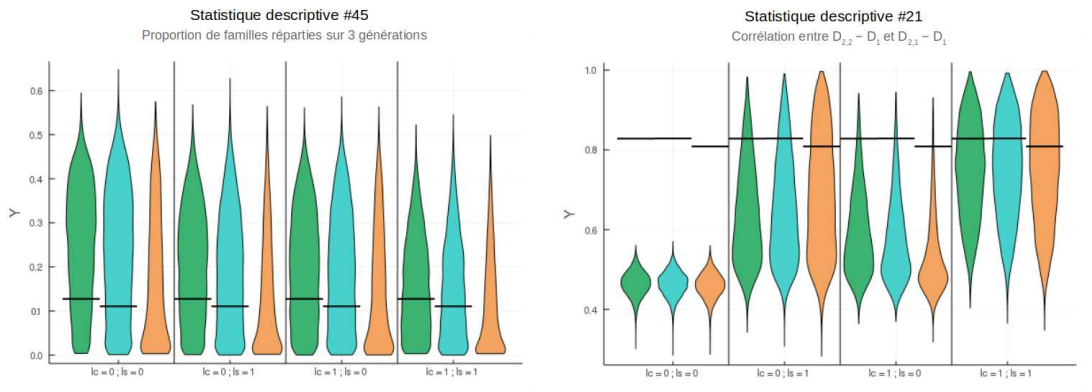


FIGURE 3 – Valeurs prises par les statistiques descriptives Y_{45} (gauche) et Y_{21} (droite) sur l'ensemble des paramètres échantillonnés par Latin Hypercube, suivant l'une des 4 valeurs prises par le couple (I_c, I_s) modélisant la présence (valeur égale à 1) ou absence (0) de concordance et synchronicité et l'une des 3 conditions initiales, à savoir si la cellule de départ est de type HSC* (vert), MPP (cyan) ou HPC (orange). Les lignes horizontales (représentées à l'identique pour chaque couple (I_c, I_s)) matérialisent la valeur de ces statistiques descriptives $\hat{y}^p_{45}$ et $\hat{y}^p_{21}$ calculées sur les données expérimentales en fonction du type p de la cellule de départ (notons que nous n'avons pas d'observations pour l'expérience MultiGen avec une HPC comme cellule de départ). La statistique descriptive Y_{45} correspond à la proportion de familles à avoir des cellules observées réparties sur exactement 3 générations. On observe par exemple qu'avec pour cellule de départ une HSC*, on peut trouver des paramètres tels que cette proportion aille jusqu'à 60% pour le modèle sans synchronicité ni concordance, 50% dans le cas avec synchronicité et concordance. La proportion observée vaut $\hat{y}^1_{45} = 0.127$; elle peut être obtenue pour au moins un jeu de paramètre quelles que soient les hypothèses de concordance et synchronicité. Lorsque l'on s'intéresse à la statistique descriptive Y^p_{21}, qui correspond à la corrélation entre entre les deux échantillons $(D_{2,2,i} - D_{1,i})_{i\in\mathcal{I}_p}$ et $(D_{2,1,i} - D_{1,i})_{i\in\mathcal{I}_p}$(selon le type p), on voit que les valeurs expérimentales $\hat{y}^p_{21}$ ne peuvent pas être obtenues dans le cas sans concordance ni synchronicité. On observe que les valeurs prises par Y^p_{21} se distribuent différemment suivant la valeur de I_c et I_s, c'est-à-dire que que cette statistique descriptive est sensible aux paramètres (ou aux hypothèses) I_c et I_s.

Nous pouvons ensuite calculer les statistiques descriptives pour chaque vecteur de paramètres (et chaque condition initiale). Nous pouvons avoir une idée des valeurs prises par ces statistiques descriptives, en regardant comment leurs valeurs se distribuent (voir Fig. 3). Nous pouvons alors pour chaque statistique descriptive Y^p_i estimer $\mathbb{V}_{\boldsymbol{\theta}}[Y^p_i(\boldsymbol{\theta})]$ ou encore les quantiles à 5 et 95% qui seront utilisés par la suite.
En parcourant chacune des statistiques descriptives (voir annexe A.2), nous pouvons vérifier que notre modèle (avec sa paramétrisation actuelle) n'échoue pas à reproduire les observations expérimentales. Nous constatons par exemple, à partir de la statistique descriptive Y_{21}, que le modèle avec l'hypothèse $I_c = I_s = 0$ n'est pas adapté, ce qui permettrait dès à présent d'exclure l'hypothèse d'une absence à la fois de synchronicité et de concordance.

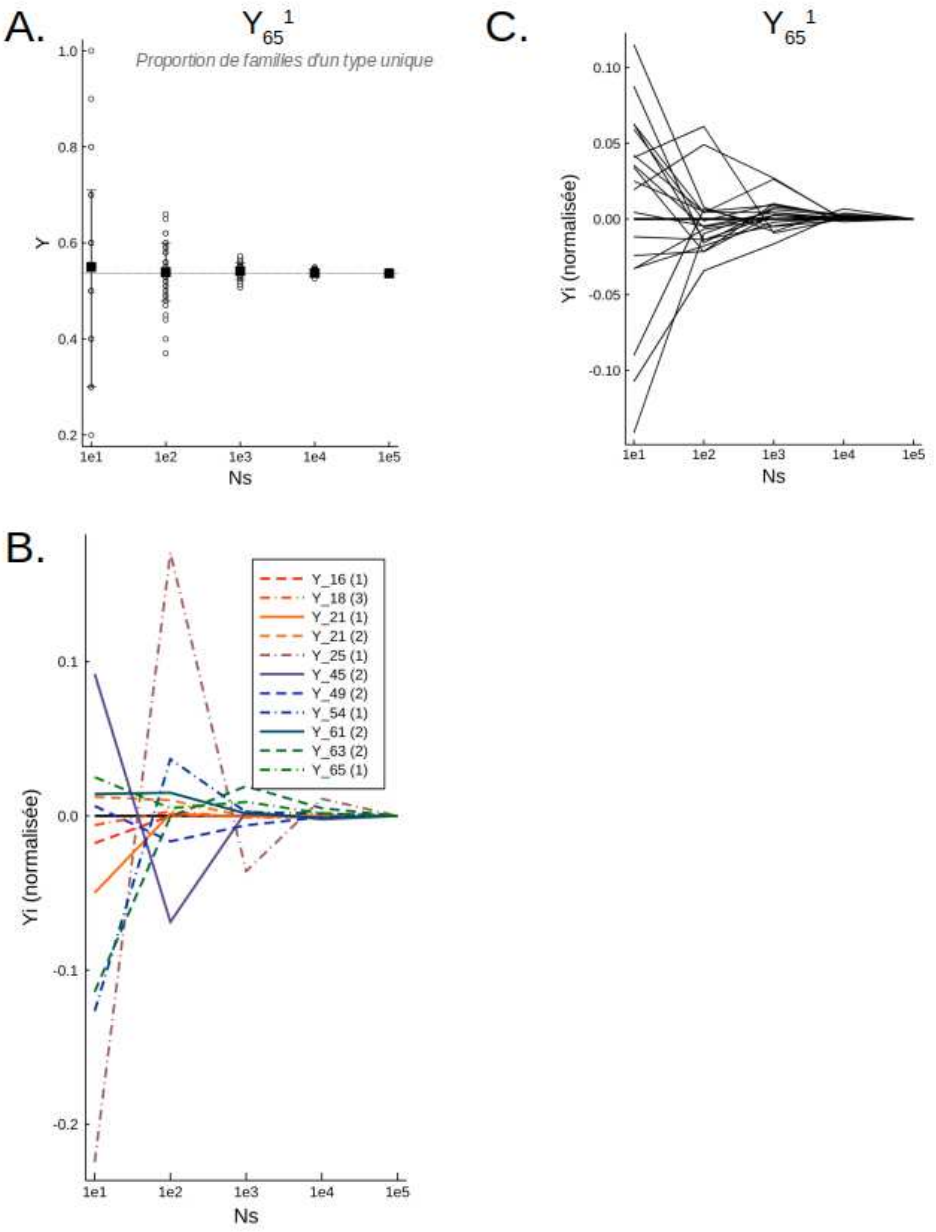


FIGURE 4 – Étude de l'influence du paramètre N_s sur le calcul des statistiques descriptives. Pour une valeur donnée $\boldsymbol{\theta}$ du vecteur de paramètres, nous pouvons simuler le modèle un nombre N_s de fois pour obtenir une réalisation y^1_{65} de la statistique descriptive Y^1_{65} (A). Pour chaque valeur de N_s, nous reproduisons cette procédure 50 fois, et obtenons 50 réalisations de la variable aléatoire (cercles blancs). Les carrés représentent la valeur moyenne de y^1_{65} sur les 50 réalisations, et les barres d'erreurs les quantiles à 10 et 90%. Plus N_s est grand, plus les réalisations y^1_{65} se rapprochent de l'espérance $\mathbb{E}[Y^1_{65}]$. Pour $N_s = 1,000$, les écarts sont faibles ; pour $N_s = 10^5$ ils sont négligeables. Sur la figure B, pour ce même jeu de paramètres (qui est en fait celui estimé à la section 5.3), nous regardons l'impact de N_s sur le calcul de plusieurs statistiques descriptives (celles qui correspondent au sous-ensemble $\mathcal{S}_A$ défini au § 4.4). Les résultats correspondent aux valeurs moyennées sur 50 calculs. On normalise ici les valeurs par la valeur obtenue pour $N_s = 10^5$ (supposée égale à l'espérance). Nous pouvons voir que certaines statistiques convergent plus lentement (comme par exemple Y^1_{25}) vers leurs espérances. Sur la figure C, nous regardons l'influence de N_s sur la valeur des réalisations de la statistique descriptive Y^1_{65}, cette fois-ci suivant le jeu de paramètres utilisé (en en choisissant 20 au hasard).

4.3 Analyse de sensibilité

Maintenant que nous avons calculé les statistiques descriptives $(y_{i,j})_{1\leq i\leq 67, 1\leq j\leq N_{LHS}}$ pour l'ensemble des $\boldsymbol{\theta}_j$ échantillonnés, nous pouvons pour chacune d'entre elles évaluer à quels paramètres elle est sensible. Pour cela, nous nous baserons sur l'indice de Sobol du premier ordre, défini, pour un paramètre θ_k et une statistique descriptive Y_i :

$$S_k = \frac{\mathbb{V}[Y_i(\boldsymbol{\theta})] - \mathbb{E}\left[\mathbb{V}\left[Y_i(\boldsymbol{\theta}) \mid \theta_k\right]\right]}{\mathbb{V}[Y_i(\boldsymbol{\theta})]} \tag{11}$$

Pour donner une intuition de ce que cet indice mesure : $\mathbb{V}[Y_i(\boldsymbol{\theta})]$ mesure la variation de Y_i sur l'ensemble des valeurs possibles pour les paramètres. Cela correspond à la variance des distributions représentées par exemple sur la figure 3. La quantité $\mathbb{E}\left[\mathbb{V}\left[Y_i \mid \theta_k\right]\right]$ quant à elle va mesurer la dispersion des valeurs prises par Y_i lorsque l'on fait varier tous les paramètres sauf θ_k. Si θ_k est entièrement responsable des variations prises par Y_i, alors en le fixant Y_i ne varierait plus, donc $\mathbb{E}\left[\mathbb{V}\left[Y_i \mid \theta_k\right]\right]$ vaudrait zéro et l'indice de Sobol S_k vaudrait 1. À l'inverse, si le paramètre θ_k ne permet pas d'expliquer la variabilité de Y_i (notamment dans le cas où, en changeant sa valeur, une autre combinaison de valeurs pour d'autres paramètres conduirait à la même valeur de Y_i), alors $\mathbb{E}\left[\mathbb{V}\left[Y_i \mid \theta_k\right]\right] = \mathbb{V}[Y_i]$ et $S_k = 0$.
Ainsi, $S_k \in [0, 1]$ mesure la sensibilité d'une sortie du modèle à un paramètre k. Une faible valeur indique que la sortie du modèle considérée n'est pas sensible aux variations du paramètre. À l'inverse, une forte valeur indique que le paramètre a une forte influence.
Si un paramètre est tel qu'aucune des statistiques descriptives ne lui est sensible, alors ce paramètre peut raisonnablement être fixé à une valeur constante : peut importe la valeur choisie, d'autres combinaisons de valeurs pour les autres paramètres permettraient d'obtenir les mêmes résultats.
L'analyse de sensibilité est ainsi une méthode qui nous permet de mieux comprendre l'influence de nos paramètres sur les sorties du modèle pour éliminer les moins influents et réduire ainsi le nombre de paramètres à estimer.

Une bonne estimation de la valeur des indices de Sobol est coûteuse en temps de calcul ; nous en ferons une approximation à partir de nos N_{LHS} échantillons de jeux de paramètres. Dans l'expression 11, les variances seront remplacées par les variances empiriques et le second terme qui intervient sera remplacé par :

$$\begin{aligned}
\mathbb{E}\left[\mathbb{V}\left[Y_i \mid \theta_k\right]\right] &= \int \mathbb{V}\left[Y_i \mid \theta_k = v\right] p[\theta_k = v] \mathrm{d}v \\
&\approx \sum_{1\leq j\leq 10} \mathbb{V}\left[Y_i \mid \theta_k \in [q_{j-1}, q_j]\right] \mathbb{P}[\theta_k \in [q_{j-1}, q_j]] \\
&= 1/10 \sum_{1\leq j\leq 10} \mathbb{V}\left[Y_i \mid \theta_k \in [q_{j-1}, q_j]\right]
\end{aligned}$$

en ayant défini les q_j pour $j \in \{0, 1, \cdots, 10\}$ comme étant les quantiles à $10\times j$ % de l'échantillon $(\theta_{k,n})_{1\leq n\leq N_{LHS}}$ de telle sorte que $\mathbb{P}[\theta_k \in [q_{j-1}, q_j]] = 1/10$. On a choisi ici de diviser l'intervalle de valeurs prises par θ_k en 10. Plus ce nombre est élevé, plus l'approximation précédente est bonne, mais moins la variance empirique - qui sera alors calculée sur des échantillons de taille plus réduite - sera proche de la variance théorique. Des méthodes plus sophistiquées devraient être utilisées pour avoir une bonne estimation de l'indice de Sobol du premier ordre mais également calculer les indices d'ordres supérieurs et l'indice d'ordre total.

Avec cette méthode de calcul, nous pouvons alors estimer les indices de Sobol du premier ordre pour l'ensemble des statistiques descriptives et l'ensemble des conditions initiales. Nous choisissons également de considérer nos paramètres I_c et I_s séparément, et présentons les résultats conditionnellement à la valeur du couple (I_s, I_c) L'intégralité des résultats est présenté en annexe B. Nous ne présentons ici que quelques exemples pour illustrer notre propos (voir Fig. 5).

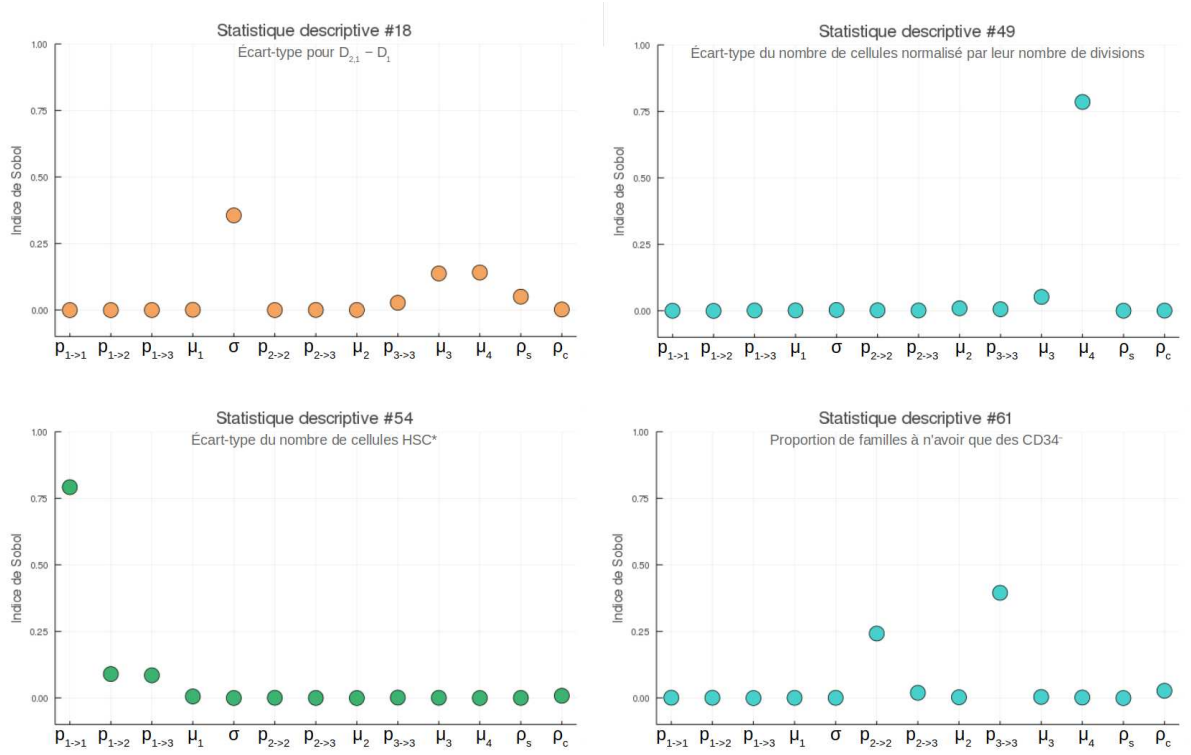


FIGURE 5 – Indice de Sobol du premier ordre S_k (en ordonnée) de tous les paramètres θ_k (en abscisse) pour $1 \leq k \leq 13$ pour les statistiques descriptives Y_{18}^3 (en haut à gauche), Y_{49}^2 (en haut à droite), Y_{54}^1 (en bas à gauche) et Y_{61}^2 (en bas à droite).
Y_{18}^3 correspond à l'écart-type pour $D_{2,1} - D_1$ (avec comme condition initiale une cellule HPC). Cette sortie du modèle est naturellement sensible aux paramètres σ, μ_{HPC} et μ_{CD34^-}. Y_{49}^2 correspond à l'écart-type du nombre de cellules échantillonnées, après normalisation en fonction du nombre de divisions subies. Cette statistique descriptive est fortement sensible au paramètre μ_{CD34^-}. Y_{54}^1 correspond à l'écart-type du nombre de cellules à être de type HSC* (partant d'une cellule HSC*). On observe ainsi que cette quantité est sensible aux paramètres liés à la dynamique des HSC*, principalement à la probabilité $p_{1\to1}$. Y_{61}^2 correspond à la proportion de familles à n'avoir que des cellules (après échantillonnage) de type CD34$^-$. On observe que cette statistique descriptive est sensible aux paramètres $p_{2\to2}$ et $p_{3\to3}$ ce qui n'était pas intuitif a priori. Les indices de Sobol sont calculés ici dans le cas avec synchronicité ($I_s = 1$) et concordance ($I_c = 1$). Plus l'indice de Sobol d'un paramètre est élevé, plus la statistique descriptive lui est sensible. Pour rappel, sur les noms des paramètres, les indices 1, 2, 3 et 4 correspondent respectivement aux types HSC*, MPP, HPC et CD34$^-$.

Si l'on regarde alors, pour chaque paramètre, l'indice maximal de Sobol, calculé sur l'ensemble des statistiques descriptives, nous obtenons les résultats de la figure 6.
Certains paramètres, tels que $p_{3\to3}$ (probabilité de division symétrique pour le type cellulaire HPC), μ_{CD34^-} (médiane du temps de division des cellules de type $CD34^-$) ou encore $p_{1\to1}$ (probabilité de division symétrique des HSC*) ont une forte influence, tandis que d'autres, tels que $p_{1\to2}$ et $p_{1\to3}$ (probabilité qu'une HSC* donne respectivement un MPP ou un HPC), $p_{2\to3}$ (probabilité qu'un MPP donne un HPC) ou μ_{HSC*} sont tels qu'aucune des statistiques descriptives ne leur soit vraiment sensible. On choisi alors, pour simplifier notre modèle, de ne pas estimer ces quatre derniers paramètres. On fixe alors :

$$\mu_{HSC*} = \mu_{MPP} \tag{12}$$

$$p_{1\to2} = p_{1\to3} = \frac{1 - p_{1\to1}}{3} \tag{13}$$

$$p_{2\to3} = \frac{1 - p_{2\to2}}{2} \tag{14}$$

En faisant ces choix, on perd alors la possibilité d'avoir une certaine information quant à certaines transitions possibles lors de la différenciation. On reste néanmoins en mesure d'estimer la probabilité qu'une cellule de type a se divise suivant une division symétrique, asymétrique ou différenciée, probabilités qui valent respectivement, dans le cas sans concordance : $(p_{a\to a})^2$, $2p_{a\to a}(1 - p_{a\to a})$ et $(1 - p_{a\to a})^2$.

Il nous reste alors 9 paramètres à estimer. Nous continuons à noter $\boldsymbol{\theta}$ le vecteur de ces 9 paramètres. Pour leur estimation, nous allons devoir définir un sous-ensemble de statistiques descriptives qui serviront ensuite pour définir une distance à nos observations.

4.4 Choix d'un sous-ensemble de statistiques descriptives

Après avoir enlevé certains de nos paramètres, nous aboutissons à un ensemble de statistiques descriptives suffisant pour caractériser le processus puisque, pour chaque paramètre restant à estimer, on aura au moins une statistique descriptive qui lui sera sensible.
Néanmoins, on a actuellement 67×3 statistiques descriptives. Le meilleur jeu de paramètres pour une statistique descriptive particulière ne le sera probablement pas pour une autre. Pour calibrer notre modèle, il nous faudra alors combiner ces statistiques descriptives. Or, leur nombre est actuellement beaucoup trop important. En méthode ABC (Approximate Bayesian Computation), méthode où l'on cherche à approcher la distribution a posteriori des paramètres du modèle en confrontant des observations expérimentales à des observations simulées au travers justement l'utilisation de statistiques descriptives, Prangle [17] fait référence à cette problématique sous le terme de "malédiction de la dimension" : un nombre limité de statistiques descriptives doivent être utilisées afin de réduire l'erreur d'approximation, et elles doivent alors être trouvées pour être informative des paramètres à estimer ou des hypothèses à tester [17].
Bien que nous n'utiliserons pas ici une méthode d'estimation Bayésienne, nous nous sommes placés dans un cadre proche de celui de l'utilisation de la méthode ABC : ayant une fonction de vraisemblance qui ne peut être évaluée, nous utilisons des statistiques descriptives dont nous confrontons les valeurs calculées à partir des observations expérimentales à celles évaluées sur des simulations. Nous nous limiterons cependant à un problème d'optimisation, cherchant non pas à estimer la distribution a posteriori des paramètres, mais le vecteur de paramètres qui minimise une certaine distance (qu'on aimerait proche de la log-vraisemblance). Cette étude pourra être considérée comme une étape préliminaire à l'utilisation d'un algorithme ABC, permettant par exemple son initialisation.
Ainsi, nous chercherons dans la suite à construire un ensemble de statistiques descriptives qui puissent être adaptées pour l'utilisation de la méthode ABC.

Partant d'un ensemble initial de statistiques descriptives, deux grandes catégories de méthodes existent pour aboutir à un ensemble de taille plus réduit [18].

La première revient à étudier des sous-ensembles de cet ensemble initial pour en choisir le meilleur. Le deuxième revient à créer des nouvelles statistiques descriptives, combinaisons linéaires (ou non) de celles de l'ensemble initial. Cette méthode, appelée technique de projection, est dans l'idée proche de l'analyse en composantes principales : il s'agit de projeter les variables sur un espace de taille réduite qui minimise la corrélation des nouvelles variables entre elles, et maximiser leur variabilité et leur corrélation avec le vecteur de paramètre $\boldsymbol{\theta}$, de telle façon que chacune des nouvelles variables soit informative (vis à vis de notre problématique d'estimation de paramètres). Cette méthode a l'avantage d'être relativement simple à mettre en place, et l'inconvénient d'aboutir à des nouvelles statistiques descriptives qui perdent leur signification biologique.
En particulier, mentionnons Fearnhead et Prangle [19] qui proposent la définition d'autant de statistiques descriptives qu'il y a de paramètres à estimer, chacune étant choisie pour sa capacité à prédire la valeur d'un paramètre dans un modèle de régression.

Notre approche sera celle de la définition d'un sous-ensemble de statistiques descriptives. Néanmoins, nous ne procéderons pas à une analyse systématique des sous-ensembles possibles mais chercherons à rationaliser sa construction. Pour cela, nous choisissons tout d'abord de fixer la taille du sous-ensemble au nombre de paramètres à estimer (comme le proposent Fearnhead et Prangle [19] avec leur méthode de projection), c'est-à-dire de choisir 11 statistiques descriptives. Notre construction du sous-ensemble, noté $\mathcal{S}_A$, repose alors sur l'intuition sur laquelle se base la méthode de projection, à savoir qu'une statistique descriptive devrait être choisie pour sa capacité à estimer un paramètre. Nous avons alors pour chaque paramètre listé les statistiques descriptives qui lui étaient sensibles, puis choisi l'une d'elles (de façon empirique). Nous aboutissons alors à :

$$\mathcal{S}_A = \{Y_{16}^1, Y_{18}^3, Y_{21}^1, Y_{21}^2, Y_{25}^1, Y_{45}^2, Y_{49}^2, Y_{54}^1, Y_{61}^2, Y_{63}^2, Y_{65}^1\}$$

Le tableau 2 décrit la composition de $\mathcal{S}_A$. Nous détaillons dans l'annexe D à ce chapitre les étapes ayant permis d'aboutir au choix précédent.

Nous ferons ainsi la calibration du modèle à partir d'une distance construite à partir de $\mathcal{S}_A$. Pour valider nos résultats, mais également pour étudier l'influence du choix du sous-ensemble sur les résultats de l'inférence, nous construisons un second sous-ensemble noté $\mathcal{S}_B$. Le choix a alors été laissé à Alessandro Donada, qui a construit ce sous-ensemble à partir à la fois de ses a priori biologiques et de la connaissance des résultats des expériences (voir Tab. 3) :

$$\mathcal{S}_B = \{Y_{17}^2, Y_{18}^2, Y_{19}^3, Y_{20}^1, Y_{50}^1, Y_{51}^1, Y_{52}^2, Y_{53}^2, Y_{63}^1, Y_{64}^1, Y_{65}^2\}$$

5 Estimation des paramètres

5.1 Distance

Soit $Y_i \in \mathcal{S}_A$ une statistique descriptive du sous-ensemble $\mathcal{S}_A$. Nous notons $\sigma_{Y_i} = \sqrt{\mathbb{V}[Y_i]}$ son écart-type calculé sur les valeurs prises par Y sur l'ensemble des paramètres échantillonnés $\boldsymbol{\theta}_{1 \leq j \leq N_{LHS}}$.
Soit $\boldsymbol{\theta}$ un vecteur de paramètres donné pour lequel nous simulons le modèle un grand nombre de fois (N_s par condition initiale). Nous calculons alors $y_i(\boldsymbol{\theta})$, réalisation de la variable aléatoire Y_i calculée à partir de l'échantillon simulé des N_s variables d'observation. Nous notons $\boldsymbol{y_A}(\boldsymbol{\theta}) = (y_i(\boldsymbol{\theta}))_{i \in \mathcal{S}_A}$ le vecteur de statistiques descriptives calculé à partir des simulations et $\hat{\boldsymbol{y}}_{\boldsymbol{A}}$ celui calculé à partir des observations expérimentales.

La distance aux observations, ou fonction de coût que nous chercherons à minimiser, est alors la suivante :

$$d_A(\boldsymbol{y_A}(\boldsymbol{\theta}), \hat{\boldsymbol{y}}_{\boldsymbol{A}}) = \sqrt{\sum_{Y_i \in \mathcal{S}_A} \left| \frac{y_i(\boldsymbol{\theta}) - \hat{y}_i}{\sigma_{Y_i}} \right|^2} \tag{15}$$

Y_i^p	Description	θ_j
Y_{16}^1	Corrélation entre les temps de seconde divisions $D_{2,1}$ et $D_{2,2}$	σ
Y_{18}^3	Écart-type pour $D_{2,1} - D_1$	μ_{HPC}
Y_{21}^1	Corrélation entre $D_{2,2} - D_1$ et $D_{2,1} - D_1$	ρ_s
Y_{21}^2	Corrélation entre $D_{2,2} - D_1$ et $D_{2,1} - D_1$	ρ_c
Y_{25}^1	Nombre moyen de cellules à avoir fait 1 division	μ_{HSC*}
Y_{45}^2	Proportion de familles réparties sur 3 générations	I_s
Y_{49}^2	Écart-type du nombre de cellules normalisé par leur nombre de divisions	μ_{CD34^-}
Y_{54}^1	Écart-type du nombre de cellules HSC*	$p_{1\to1}$
Y_{61}^2	Proportion de familles à n'avoir que des CD34$^-$	$p_{3\to3}$
Y_{63}^2	Proportion de familles réparties sur 3 types cellulaires	$p_{2\to2}$
Y_{65}^1	Proportion de familles d'un type unique	I_c

TABLE 2 – Liste des statistiques descriptives qui composent le sous-ensemble $\mathcal{S}_A$ qui servira pour l'estimation des paramètres du modèle. Dans la dernière colonne, nous listons les paramètres du modèle à estimer et les associons à une des statistiques descriptives choisies parce qu'elles leur étaient sensibles.

Y_i^p	Description
Y_{17}^2	Moyenne de l'écart $D_{2,1} - D_1$
Y_{18}^2	Écart-type pour $D_{2,1} - D_1$
Y_{19}^3	Moyenne de l'écart $D_{2,2} - D_1$
Y_{20}^1	Écart-type pour $D_{2,2} - D_1$
Y_{50}^1	Nombre moyen de cellules à être de type HSC*
Y_{51}^1	Nombre moyen de cellules à être de type MPP
Y_{52}^2	Nombre moyen de cellules à être de type HPC
Y_{53}^2	Nombre moyen de cellules à être de type CD34$^-$
Y_{63}^1	Proportion de familles réparties sur 3 types cellulaires
Y_{64}^1	Proportion de familles réparties sur 2 types cellulaires
Y_{65}^2	Proportion de familles d'un type unique

TABLE 3 – Liste des statistiques descriptives qui composent le sous-ensemble $\mathcal{S}_B$ qui servira pour l'étape de validation des résultats.

À noter que, plutôt que de normaliser par les variances $\sigma^2_{Y_i}$, il aurait été pertinent de calculer la matrice de covariance.
Nous chercherons ensuite à estimer le vecteur de paramètre $\hat{\boldsymbol{\theta}}$ qui minimise la distance précédente :

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ d_A(\boldsymbol{y_A}(\boldsymbol{\theta}), \hat{\boldsymbol{y}}_{\boldsymbol{A}}) \tag{16}$$

Pour un même jeu de paramètres, comme illustré sur la figure 4, nous pouvons avoir différentes valeurs de $\boldsymbol{y_A}(\boldsymbol{\theta})$. Cela revient à considérer que la fonction de coût est bruitée, ce qui posera certaines difficultés quand à la recherche numérique d'un optimum, et nécessitera le recours à des algorithmes d'optimisation adaptés.

On définit de même d_B, distance exprimée à partir des statistiques descriptives du sous-ensemble $\mathcal{S}_B$.

5.2 Algorithme d'optimisation CMA-ES

Pour faire l'estimation des paramètres de notre modèle, nous chercherons à résoudre un problème d'optimisation, à savoir trouver le jeu de paramètres qui minimise la distance (15). Nous faisons face à deux difficultés principales. La première est que nous avons un nombre élevé de paramètres

à estimer. La seconde difficulté est due au fait que le processus étudié est stochastique, et en particulier, que pour un jeu de paramètres donné $\boldsymbol{\theta}$, puisque nous simulons le modèle qu'un nombre fini N_s de fois, nous pouvons avoir plusieurs réalisations distinctes pour nos statistiques descriptives, ce qui revient à considérer que notre fonction de coût est bruitée.
Pour résoudre ce problème d'optimisation, nous allons procéder numériquement et utiliser un algorithme d'optimisation stochastique, à savoir l'algorithme CMA-ES (Covariance Matrix Adaptation - Evolution Strategy). Cet algorithme appartient à la famille des algorithmes évolutionnaires. C'est un algorithme d'optimisation stochastique qui donne de bons résultats pour une large catégories de problèmes, incluant ceux qui sont non-linéaires, en grande dimension ou encore impliquant des fonctions de coût bruitées ou non régulières [20]. Cet algorithme est ainsi particulièrement adapté à notre cas.
L'algorithme CMA-ES va rechercher le minimum d'une fonction sur plusieurs générations. À chaque génération, on génère un échantillon de λ points (c'est-à-dire des vecteurs de paramètres), selon une distribution normale multidimensionnelle dont la moyenne et la matrice de covariance sont calculées à partir des points sélectionnés de la génération précédente. Parmi ces λ descendants, nous en sélectionnons μ (ceux qui donnent les valeurs les plus faible pour la fonction de coût). Ce sont ceux que nous utilisons pour la génération suivante. La méthode itère jusqu'à ce que nous atteignions le nombre maximal de générations n_g ou lorsque la valeur estimée pour le maximum ne change plus suffisamment au fil des générations. L'équation de base pour l'échantillonnage des individus, pour le nombre de générations $g = 0, 1, 2, \cdots, n_g$ est la suivante [21] :

$$\theta_k^{(g+1)} \sim m^{(g)} + \sigma^{(g)} \mathcal{N}\left(\mathbf{0}, C^{(g)}\right) \quad \text{pour } k = 1, \ldots, \lambda \tag{17}$$

Avec $m^{(g)}$, $\sigma^{(g)}$ et $C^{(g)}$ respectivement la valeur moyenne, le pas et la matrice de covariance à la génération g.

Nous avons implémenté cet algorithme dans le langage de programmation Julia, dont le code est disponible sur GitLab [2].

Nous illustrons sur la figure 7 la recherche du jeu de paramètres qui minimise la distance d_A. Pour initialiser l'algorithme CMA-ES, nous commençons par choisir le meilleur jeu de paramètres $\hat{\boldsymbol{\theta}}_{LHS}$ à partir de l'ensemble des paramètres échantillonnés suivant l'échantillonnage par Latin Hypercube (Fig. 7-A).
Nous exécutons alors l'algorithme CMA-ES, en choisissant pour sa configuration $\lambda = 300$ et $n_g = 100$ (le script est disponible en annexe C à ce chapitre). Le nuage de particules va se déplacer au fur et à mesure des générations, jusqu'à atteindre un jeu de paramètres minimisant la distance (Fig. 7-B et C).
L'algorithme étant stochastique (mais également parce que notre fonction de coût est bruitée à cause de la stochasticité de notre processus), nous l'exécutons 10 fois, avec 10 graines aléatoires différentes à chaque fois, ce qui nous donne 10 candidats potentiels pour $\boldsymbol{\theta}$ (Fig. 7-D). Nous choisissons alors, parmi les dix estimations, celle qui conduit à une distance médiane (calculée sur les $\lambda = 300$ particules) la plus faible. Nous présentons les résultats de cette procédure d'estimation au paragraphe suivant.

5.3 Résultats

Après avoir exécuté notre méthode d'inférence présentée ci-dessus, nous obtenons pour les valeurs des paramètres qui minimisent la distance aux observations d_A celles présentées dans le tableau 4 (colonne correspondant à $\mathcal{S}_A$).
L'hypothèse d'une concordance ($\hat{I}_c = 1$) et d'une synchronicité ($\hat{I}_s = 1$) entre cellules sœurs est celle qui conduit aux meilleurs résultats. Néanmoins, on estime une valeur assez faible pour $\hat{\rho}_s = 0.381$, suggérant que la corrélation entre les temps de divisions de deux cellules sœurs

2. `https://gitlab-research.centralesupelec.fr/2012hermangeg/bayesian-inference`

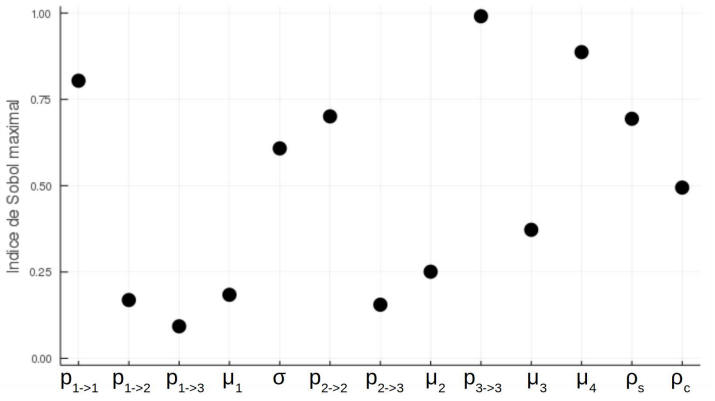


FIGURE 6 – Valeurs maximales des indices de Sobol, pour chacun des paramètres, calculées sur l'ensemble des statistiques descriptives considérées.

Paramètres	$\mathcal{S}_A$	$\mathcal{S}_B$
I_s	1	1
I_c	1	1
$p_{1\to1}$	0.427	0.408
$p_{2\to2}$	0.568	0.694
$p_{3\to3}$	0.234	0.081
μ_{HSC*}	3.127	2.918
μ_{HPC}	2.296	2.353
μ_{CD34^-}	2.859	2.794
σ	0.21	0.203
ρ_s	0.381	0.97
ρ_c	0.866	0.991

TABLE 4 – Résultats de la procédure d'estimation des paramètres. La colonne $\mathcal{S}_A$ présente les résultats principaux de notre estimation, basée sur la distance d_A. Nous notons $\hat{\boldsymbol{\theta}}$ ce vecteur de paramètres. La colonne $\mathcal{S}_B$ indique quels auraient été les résultats de l'estimation en choisissant de construire la distance à partir du sous-ensemble de statistiques descriptives $\mathcal{S}_B$.

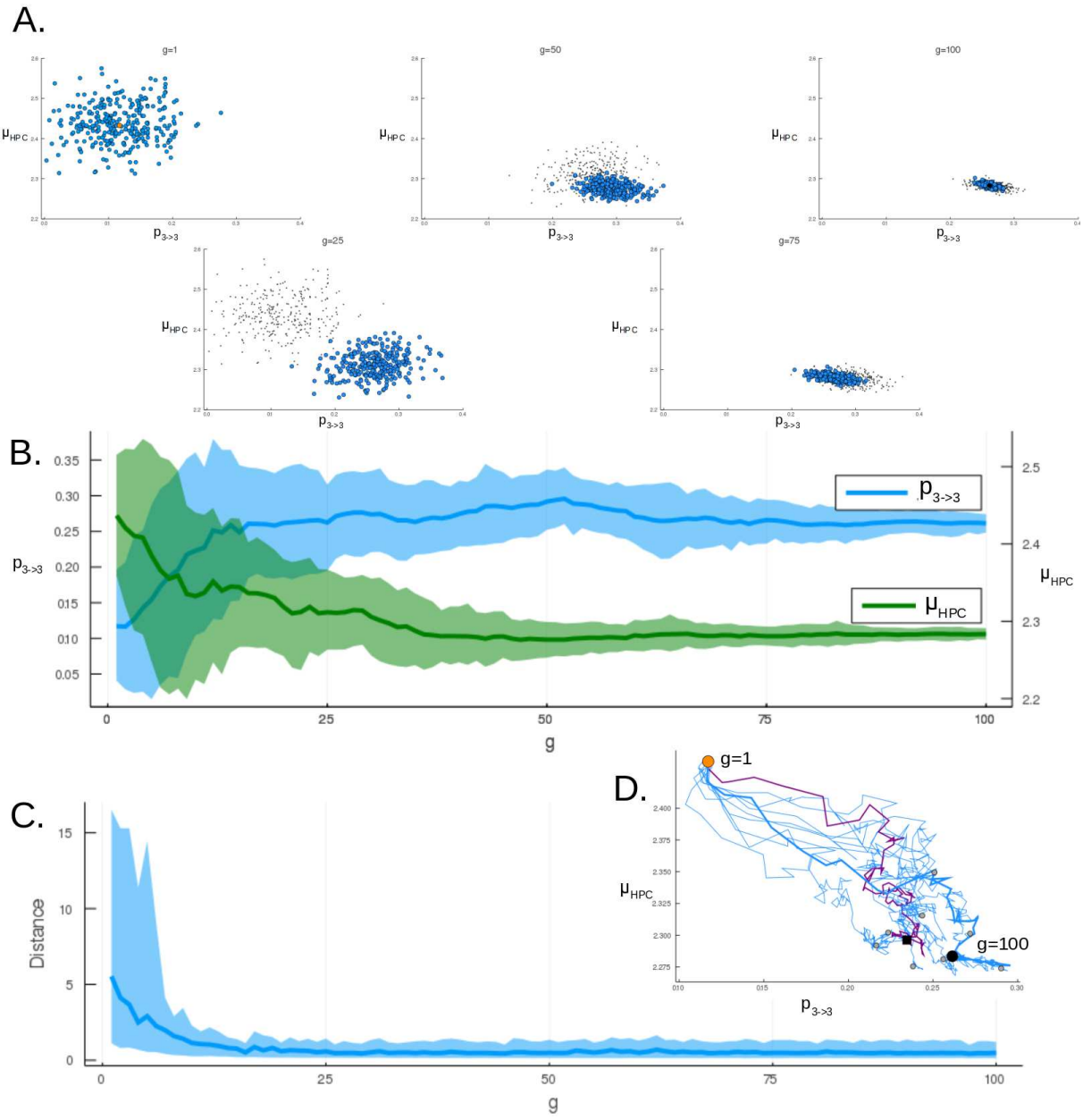


FIGURE 7 – Illustration de la procédure de recherche du meilleur jeu de paramètre $\hat{\boldsymbol{\theta}}$ à partir de l'algorithme CMA-ES, pour le choix de la distance d_A basée sur le sous-ensemble de statistiques descriptives $\mathcal{S}_A$. On montre (A) comment évoluent les valeurs prises par $p_{3\to3}$ (en abscisse) et μ_{HPC} (en ordonnée) sur les $\lambda = 300$ particules (en bleu) au cours des $n_g = 100$ générations (en représentant ici $g \in \{1, 25, 50, 75, 100\}$). L'algorithme est initialisé en distribuant les particules autour de $\hat{\boldsymbol{\theta}}_{LHS}$ (point orange). Le nuage de points va se déplacer puis se concentrer autour d'une valeur pour les paramètres qui minimisent la fonction de coût, à savoir ici la distance d_A. On représente l'évolution de la valeur $p_{3\to3}$ (bleu) et μ_{HPC} (vert) moyennée sur les 300 particules au cours des générations (B). La marge correspond aux quantiles à 5 et 95%. La distance d_A moyenne (ainsi que les quantiles à 5 et 95%) diminue au fur et et à mesure des générations (C). Nous répétons la procédure 10 fois, chaque fois avec une graine aléatoire différente, conduisant à des estimations légèrement différentes (D). La trajectoire violette correspond à la graine aléatoire ayant conduit au meilleur jeu de paramètres sur les 10 répétitions.

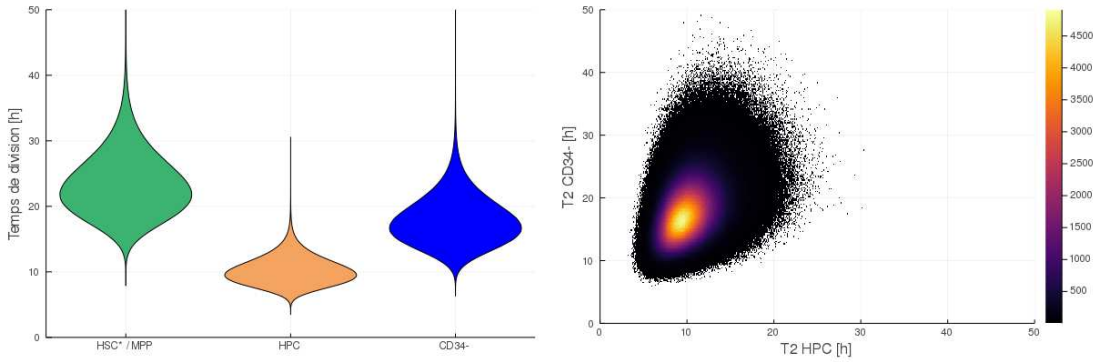


FIGURE 8 – Distribution des temps de division - suivant un modèle log-normal bivarié - avec $\hat{\mu}_{HSC*} = \hat{\mu}_{MPP} = 3.127$, $\hat{\mu}_{HPC} = 2.296$, $\hat{\mu}_{CD34^-} = 2.859$, $\hat{\sigma} = 0.21$ et $\hat{\rho}_s = 0.381$. À gauche, distribution marginale des temps de division suivant le type de la cellule. À droite, distribution conjointe des temps de division de deux cellules soeurs, l'une étant une HPC (en abscisse), l'autre une CD34$^-$ (en ordonnée). Les zones de couleur plus chaude correspondent aux zones de plus forte densité.

pourraient en fait s'expliquer par une forte concordance (le fait que les deux cellules sœurs soient souvent du même type) avec $\hat{\rho}_c = 0.866$.
Concernant le processus de prolifération, on estime $\hat{\mu}_{HPC} < \hat{\mu}_{CD34^-} < \hat{\mu}_{HSC*}$. L'ordre entre les médianes des temps de division pour les HSC* (/ MPP) et HPC était attendu. Il est plus surprenant d'avoir un temps médian de division des CD34$^-$ supérieur aux HPC : cela pourrait éventuellement s'expliquer par l'hétérogénéité entre cellules CD34$^-$. On représente sur la figure 8 la distribution des temps de division (à l'exception de la première division) en fonction du type de la cellule mère. Notons que nous obtenons pour $\hat{\mu}_{HPC}$ et $\hat{\mu}_{HSC*}$ des valeurs assez proches des bornes du prior, suggérant a posteriori de choisir un intervalle de valeurs admissibles qui soit plus large.

Concernant le processus de différenciation, on estime qu'une cellule fille aurait environ une chance sur deux d'être du même type que sa mère si cette dernière est une HSC* ou un MPP (probabilités égales à 0.427 et 0.568 respectivement), mais seulement une chance sur quatre dans le cas où la mère est une HPC ($\hat{p}_{3\to3} = 0.234$), suggérant que les HPC se différencient en moyenne en moins de divisions que les cellules plus immatures de type HSC* et MPP. La matrice de transition estimée vaut alors :

$$\hat{\mathbf{M}} = \begin{pmatrix} 0.427 & 0.14325 & 0.14325 & 0.14325 \\ 0 & 0.568 & 0.216 & 0.216 \\ 0 & 0 & 0.234 & 0.766 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

À partir des équations (5) et (6), en prenant en compte que $\hat{\rho}_c = 0.866$, on peut également estimer les probabilités qu'une cellule fasse une division symétrique, différenciée ou asymétrique. Les valeurs correspondantes sont indiquées dans le tableau 5.

Pour le vecteur de paramètres $\hat{\boldsymbol{\theta}}$ estimé, nous pouvons voir sur la figure 9 (gauche) que nous avons en effet une bonne adéquation (fit) entre les observations et les simulations, sur la base des statistiques descriptives de l'ensemble $\mathcal{S}_A$ utilisées pour la construction de la distance d_A. Nous pouvons alors évaluer la qualité de nos estimations sur les statistiques descriptives de $\mathcal{S}_B$ n'ayant pas servi pour l'inférence des paramètres (Fig. 9 - droite). Sur les 11 statistiques descriptives de ce sous-ensemble, nous obtenons une bonne adéquation pour sept d'entre elles, une surestimation - par rapport aux données expérimentales - de trois d'entre elles, et une forte différence pour Y_{51}^1 : cette paramétrisation du modèle conduit à sous-estimer fortement le nombre moyen de cellules MPP partant d'une HSC*.
Ces résultats suggèrent que notre choix de statistiques descriptives $\mathcal{S}_A$ n'est pas optimal pour

Type	Symétrique	Asymétrique	Différenciée
HSC*	0.394	0.209	0.397
MPP	0.535	0.066	0.399
HPC	0.210	0.048	0.742

TABLE 5 – Probabilités qu'une cellule d'un type a donné se divise suivant une division symétrique (deux cellules filles de type a), une division asymétrique (une et une seule cellule fille du type a) ou différenciée (deux cellules filles dont aucune n'est de type a). Les probabilités sont estimées à partir des relations (5) et (6) et des valeurs estimées pour les paramètres (Tab. 4), à partir de $\mathcal{S}_A$.

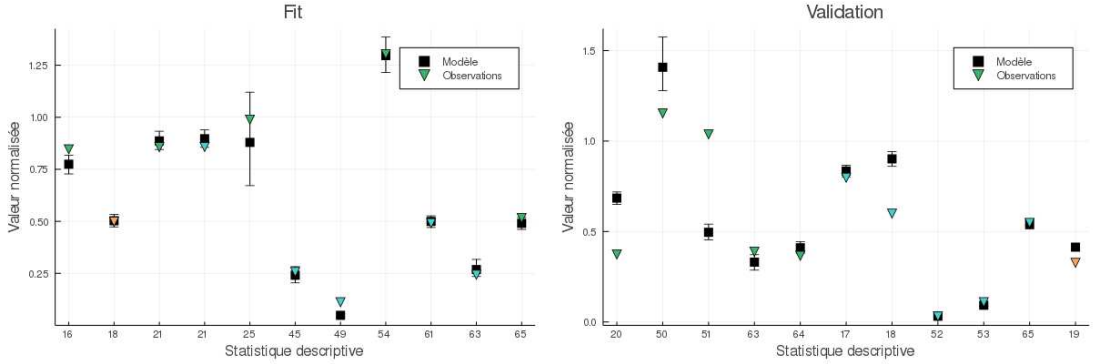


FIGURE 9 – Pour chaque statistique descriptive Y_i^p (en abscisse), nous confrontons sa valeur calculée à partir des données expérimentales $\hat{y}_i^p$ (triangle de couleur, dont la couleur correspond au type p de la cellule initiale : vert pour HSC*, cyan pour MPP et orange pour HPC) à celles calculées à partir du modèle $\left(y_{i,j}^p(\hat{\boldsymbol{\theta}})\right)_{1\leq j\leq 300}$ (en noir). Nous répétons les N_s simulations 300 fois, pour rendre compte de la variabilité de Y_i^p. Le carré noir représente la valeur médiane, les barres d'erreurs les quantiles à 5 et 95%. Nous normalisons les valeurs en calculant $\left(y_i^p - q_5(Y_i)\right) / \left(q_{95}(Y_i) - q_5(Y_i)\right)$ avec $q_5(Y_i)$ et $q_{95}(Y_i)$ qui correspondent respectivement aux quantiles à 5 et 95% de $Y_i(\boldsymbol{\theta})$ (estimés au § 4.2 à partir de la distribution des valeurs de Y_i suivant les différentes valeurs de $\boldsymbol{\theta}$ échantillonnées par Hypercube latin). À gauche, nous représentons les valeurs obtenues pour les statistiques descriptives utilisées pour l'estimation des paramètres ($\mathcal{S}_A$), à droite celles utilisées pour la validation ($\mathcal{S}_B$).

inférer au mieux le processus de prolifération et différenciation des HSPC. Nous pourrions augmenter la taille de ce sous-ensemble, en ajoutant de nouvelles statistiques pour lesquelles nous avons des mauvais résultats en validation, ce qui correspondrait dans l'idée à l'approche de Joyce et Marjoram [22]. Nous pourrions pénaliser également les sous-ensembles de taille plus importante, par l'utilisation des critères AIC ou BIC, comme proposé par Blum et al. [18]. Puisque dans leur définition, les critères d'AIC et BIC font intervenir la log-vraisemblance calculée au maximum de vraisemblance, et que nous n'avons pas d'expression de cette dernière, Blum et al. proposent d'adapter ces critères à partir d'un modèle de régression linéaire pour les paramètres, comme proposé par Beaumont et al. [23]. L'autre approche possible serait de s'orienter vers les techniques de projection pour construire de nouvelles statistiques descriptives de façon semi-automatique [19].
Enfin, si nous choisissons d'estimer les paramètres du modèle non plus à partir de la distance d_A basée sur les statistiques descriptives de l'ensemble $\mathcal{S}_A$, mais à partir de la distance d_B construite sur $\mathcal{S}_B$, nous obtenons les valeurs présentées dans la dernière colonne du tableau 4. Par rapport à nos précédents résultats, l'utilisation des statistiques $\mathcal{S}_B$ conduit à estimer une plus forte synchronicité que celle que nous avions inférée précédemment, et une bien plus faible valeur pour $p_{3\to 3}$. Les autres paramètres ont des valeurs estimées assez semblables à nos estimations précédentes. Cette comparaison permet d'illustrer l'influence du choix des statistiques descriptives dans l'estimation des paramètres, et l'importance de leur construction.

6 Discussion

Afin de comprendre la dynamique de prolifération et différenciation des cellules hématopoïétiques souches et progénitrices, et en particulier de comprendre dans quelle mesure deux cellules sœurs auraient tendance à être d'un même type cellulaire et / ou à se diviser de façon synchrone, nous avons proposé un modèle mathématique dont nous avons estimé les paramètres à partir d'un riche jeu de données hétérogènes consistant en des observations provenant de deux expériences : l'Incucyte et le MultiGen.

Notre modèle consiste en un processus stochastique à temps continu, où une cellule initiale d'un type connu à $t = 0$ va se diviser pour donner naissance à deux cellules de types potentiellement différents de celui de leur mère, et ainsi de suite, pendant 96 heures, durée limite de l'expérience. Nous avons ainsi un modèle qui décrit une famille de cellules sur un temps court, où l'aléatoire provient des choix de différenciation des cellules mais également du temps qu'elles mettent à se diviser.
Notre modèle, qui a pour objectif de décrire les premières étapes de l'hématopoïèse, est construit sur de nombreuses hypothèses :

- Une cellule ne peut être que de type HSC*, MPP, HPC ou $CD34^-$, et on néglige l'hétérogénéité entre cellules d'un même type
- Une cellule ne peut se différencier que suite à une division, et nécessairement en un type au moins aussi mature que celui de sa mère
- Il n'y a pas de mort cellulaire
- Les cellules se divisent au bout d'un temps aléatoire distribué selon une loi log-normale
  - Sauf dans le cas de la première division, étudiée en détail au chapitre 3
  - La médiane du temps de division est propre au type cellulaire de la cellule qui se divise
  - Le paramètre σ quantifiant la dispersion aléatoire autour de la médiane est supposé le même quel que soit le type cellulaire de la cellule se divisant
  - Lorsque l'on suppose que deux cellules sœurs sont synchrones, on modélise leurs temps de division par une loi log-normale bivariée. La corrélation dans les temps de division des deux cellules sœurs sera alors indépendante de leurs types.

Restreindre les premiers stades de l'hématopoïèse à quatre types cellulaires est une limitation de l'étude. Nous l'avons vu au chapitre 2, les cellules souches et progénitrices se distribueraient plutôt suivant un continuum d'états. Le nouveau paradigme pour la modélisation de l'hématopoïèse est maintenant celui d'un continuum de cellules qui se différencient progressivement de la cellule souche aux cellules matures [24, 25]. Un modèle plus réaliste devrait prendre en compte à la fois le fait qu'il y a une certaine hétérogénéité entre cellules d'un même type, mais également qu'il n'y aurait pas de délimitation stricte entre deux types cellulaires. Pour cela, des formalismes mathématiques autres que celui utilisé dans ce chapitre (qui est proche dans sa formulation des modèles de branchement à temps continu), comme par exemple l'utilisation de processus de Markov déterministes par morceaux (PDMP - Piecewise-Deterministic Markov Processes), pourraient être envisagés. Les PDMP constituent une classe très naturelle de processus stochastiques non-diffusifs [26] qui ont par exemple été utilisés dans de nombreux problèmes biologiques [27] mais pas encore à notre connaissance pour l'étude de l'hématopoïèse. Les PDMP combinent une dynamique déterministe avec sauts aléatoires et sont entièrement décrits par un espace d'état E, un flux Φ, une intensité de saut λ, et une loi de transition.
Le fait qu'on dispose pour chaque cellule de l'information, obtenue par cytométrie de flux, sur l'intensité d'expression de différents marqueurs de surface permettrait de s'orienter vers des modèles à espace d'états continu. Cela nécessiterait néanmoins un travail préliminaire de traitement des données pour, à partir d'une représentation des cellules dans un espace continu à plusieurs dimensions (autant qu'il y a de marqueurs et autres *features* telles que la morphologie des cellules par exemple), aboutir à une représentation uni-dimensionnelle qui pourrait correspondre à un degré de différenciation de la cellule. Pour cela, nous pourrions appliquer la méthode que nous avons développée au chapitre 2, proposée pour analyser des données obtenues par cytométrie de

masse mais qui pourrait se décliner facilement dans le cas d'observations en cytométrie de flux. Une difficulté éventuelle serait liée au nombre plus restreint de marqueurs à notre disposition dans ce cas. Nous pourrions également nous orienter vers une approche basée sur la théorie des *principal curves* de Hastie et Stuetzle [28].

Plusieurs hypothèses et simplifications ont été faites quant à la modélisation du temps nécessaire aux cellules pour se diviser. Mise à part la première division, nous avons choisi l'utilisation du modèle log-normal, utilisé également par d'autres auteurs. Nous aurions pu rester plus général avec le choix de la loi gamma généralisée. Surtout, nous avons simplifié en ne considérant pas de distinction du paramètre σ (écart-type de la loi normale associée à la loi log-normale) suivant le type cellulaire, hypothèse peut-être trop simplificatrice, notamment parce que les cellules de type $CD34^-$, qui constitue un type plus hétérogène que les trois autres, pourraient avoir des temps de division qui varient plus que ceux des cellules d'un autre type.
Pour modéliser la synchronicité, nous avons également opté pour un choix qui a l'avantage de n'introduire qu'un paramètre en plus à estimer, mais qui par conséquent est peut-être trop simplifié. En choisissant de modéliser les temps de division de deux cellules sœurs par une distribution log-normale bi-variée, nous nous sommes restreints à une structure de corrélation simple, quand de nombreuses autres seraient possibles, par exemple via l'utilisation de copules [29].

Enfin, nous avons proposé un modèle à l'échelle d'une famille de cellules, dans lequel nous n'avons considéré des liens directs qu'entre mère et filles, et entre cellules sœurs dans le cas où l'on ajoutait des hypothèses de synchronicité et concordance à notre modèle. Notre modélisation est dans ce cas assez abstraite, et ne permet pas de décrire la nature de ces liens. On pourrait chercher à modéliser ces hypothèses de façon plus directe, et également chercher à modéliser des liens familiaux sur plusieurs générations.
Néanmoins, l'objectif en construisant notre modèle était de rester aussi parcimonieux que possible, notamment pour pouvoir en faire sa calibration. D'ailleurs, suite à une analyse de sensibilité, nous avons simplifié un peu plus le modèle en supprimant quatre degrés de liberté. Nous nous sommes pour cela basé sur le calcul d'un indice de Sobol du premier ordre, mais il serait également pertinent de considérer l'indice d'ordre total.

Par la complexité du modèle, mais également parce que nos observations expérimentales sont, soit censurées par intervalles dans le cas de l'expérience Incucyte, soit associées à un bruit d'échantillonnage (observations incomplète) dans le cas de l'expérience MultiGen, l'estimation des paramètres constitue une tâche ardue. N'étant pas possible d'exprimer une vraisemblance que nous aurions pu chercher à maximiser, nous nous sommes alors placés dans un cadre couramment associé à l'estimation Bayésienne par méthode ABC, c'est-à-dire la construction de statistiques descriptives à partir desquelles nous définissons une distance entre les observations expérimentales et les observations simulées.
Partant d'un vaste ensemble de 3×67 statistiques descriptives, nous en avons sélectionné 13, autant que de paramètres à estimer. Notre sélection s'est basée sur une approche empirique, consistant à associer à chaque paramètre une statistique descriptive lui étant sensible. L'intuition repose sur la technique de projection de Fearnhead et Prangle [19] qui construisent un vecteur de statistiques descriptives, de la taille du vecteur de paramètres à estimer, en ajustant un modèle de régression pour que chaque nouvelle statistique soit en mesure de prédire la valeur d'un paramètre. De la méthode de Fearnhead et Prangle, nous avons gardé l'idée d'associer à chaque paramètre une statistique descriptive lui étant sensible. Néanmoins, nous souhaitions aboutir à un sous-ensemble de notre ensemble de 3×67 statistiques descriptives, c'est-à-dire s'orienter plus vers une méthode de sélection du meilleur sous-ensemble. Pour rester dans cette dernière approche, il conviendrait alors d'avoir une analyse plus systématique, étudiant différents sous-ensembles de statistiques descriptives pour en choisir le meilleur. Le sous-ensemble sélectionné à partir de notre approche pourrait alors en constituer le point de départ. Nous pourrions alors chercher à ajouter et enlever des statistiques descriptives, suivant par exemple la méthode

de Joyce et Marjoram [22] avec leur concept d'ε-*sufficiency*, ou éventuellement en appliquant une procédure de sélection sur la base d'un critère d'AIC adapté, comme proposé par Blum et al. [18].
Il serait également pertinent de mettre en place la technique de construction semi-automatique de Fearnhead et Prangle [19], ou d'étudier des techniques de projection, par exemple l'ACP. Même si les statistiques descriptives ainsi construites perdent leur sens biologique, on peut quand-même procéder à une évaluation des résultats à partir des statistiques descriptives initiales.
Nous l'avons montré dans ce chapitre, la façon de construire la distance aux observations, c'est-à-dire dans notre cas le choix des statistiques descriptives, a un impact sur les résultats de l'estimation et les conclusions. Ainsi, pour aller plus loin, il serait nécessaire d'explorer différentes méthodes de construction des statistiques descriptives, par exemple celles proposées dans la revue de Blum et al. [18], afin d'éviter de faire reposer notre méthode sur un choix insuffisamment justifié, surtout quand les résultats vont dépendre de ce choix.

Dans ce chapitre, nous avons fait une estimation ponctuelle de notre vecteur de paramètres en utilisant un algorithme d'optimisation, à savoir l'algorithme CMA-ES [21]. Néanmoins, le cadre dans lequel nous nous sommes placés, avec la définition de statistique descriptives, est en fait adapté à l'utilisation de méthodes d'estimation ABC, donc à une inférence Bayésienne des paramètres de notre modèle. L'utilisation d'une méthode Bayésienne permettrait non seulement une estimation ponctuelle des paramètres de notre modèle, mais également de leur distribution *a posteriori*. Ainsi, nous pourrions avoir une estimation de l'incertitude sur nos paramètres. L'estimation ponctuelle effectuée par l'algorithme CMA-ES resterait pertinente, mais plutôt comme façon d'initialiser les algorithmes d'inférence Bayésienne. Nous reprendrons par exemple cette idée au chapitre 6. Notamment, l'algorithme CMA-ES permet également d'estimer une matrice de covariance, ce qui est analogue à estimer l'inverse de la matrice Hessienne dans une méthode quasi-Newton [20]. On pourrait ainsi se baser sur ce résultat pour avoir une estimation de l'incertitude sur nos paramètres. Le recours à une méthode ABC pour l'estimation des paramètres de ce modèle est bien un des objectifs de ce travail. Cependant, il faut noter que l'utilisation de ces méthodes présente certaines difficultés en grande dimension, d'où l'intérêt de chercher à réduire le nombre de paramètres à estimer par une méthode d'analyse de sensibilité comme nous l'avons fait dans ce chapitre. Néanmoins, le processus étudié étant complexe, il sera délicat de réduire plus encore la dimension de l'espace des paramètres. Nous pourrions alors envisager de séparer le problème d'estimation de nos paramètres en plusieurs sous-problèmes. C'est par exemple ce que nous avons fait en se concentrant au chapitre 3 sur les temps de première division, ou dans ce chapitre en estimant directement le *recovery rate* η à partir d'une valeur d'une des statistiques descriptives (eq. (10)). En approfondissant l'étude théorique de notre modèle, notamment du processus de différenciation dont on peut - en faisant abstraction du temps mis par les cellules à se diviser - en extraire un processus de branchement multi-types pour lesquels il existe des résultats théoriques, on pourrait envisager exprimer certains de nos paramètres directement comme fonctions des variables d'observation.

Pour améliorer la qualité de nos estimations, il serait enfin nécessaire d'inclure plus d'observations à l'étude, notamment dans le cas de l'expérience MultiGen. Nous avons par exemple montré l'impact de la taille des échantillons de données sur le calcul des statistiques descriptives : plus nous avons d'observations, plus les statistiques descriptives vont s'approcher de leur espérance. Des expériences supplémentaires ont déjà été conduites, qu'il faudrait alors inclure à l'étude. Notamment, nous avons des expériences réalisées jusqu'à 72 heures qui pourraient alors nous renseigner un peu plus sur les premières divisions, ainsi que quelques expérimentations pour le MultiGen réalisées avec comme cellules de départ des HPC.

Nos premiers résultats, qui devraient être affinés en suivant les différentes pistes d'amélioration évoquées ci-dessus, montrent que les hypothèses de concordance et de synchronicité permettent en effet de mieux expliquer les données issues des expériences MultiGen et Incucyte que les hy-

pothèses alternatives, et que la probabilité pour une cellule de se diviser de façon différenciée augmenterait avec sa maturité. Nos résultats ne permettent cependant pas d'exclure que la forte synchronicité observée dans les temps de divisions des cellules sœurs puisse être principalement due au fait que deux cellules sœurs auraient une forte probabilité d'être du même type cellulaire (i.e. une valeur élevée pour notre paramètre ρ_c). L'hypothèse d'une synchronicité entre cellules sœurs, telle que nous l'avons modélisée avec l'introduction d'un paramètre ρ_s, pourrait ne pas être si importante. Auquel cas le modèle pourrait être simplifié, ce qui ouvrirait éventuellement la voie à une étude analytique du processus. Néanmoins, nous avons vu que le choix des statistiques descriptives utilisées pour faire l'estimation pouvait avoir une influence sur la valeur estimée de ρ_s, d'où l'importance déjà soulignée plus haut de construire des statistiques descriptives qui ne biaisent pas nos résultats dans un sens ou dans l'autre.
Au-delà des valeurs estimées pour nos paramètres, et de l'information que nous pouvons en tirer sur la dynamique de prolifération et différenciation des HSPC, l'objectif de ce travail est avant tout d'établir un modèle de référence dans le cas nominal, c'est-à-dire en l'absence de pathologie, et sans usage de traitement sur les cellules. Ce même modèle pourrait alors être étudié et calibré dans le cas de cellules mutées pour l'une des mutation motrice des néoplasmes myéloprolifératifs ($JAK2^{V617F}$, $CALR^m$ ou MPL^m). On pourrait alors estimer si le modèle reste valable et si oui, étudier dans quelle mesure les valeurs des paramètres diffèrent du cas pathologique. De même, on pourrait étudier le cas de cellules (saines ou mutées) exposées à des traitements, tels que l'interféron α, afin de mieux comprendre son impact sur la dynamique de prolifération et différenciation des cellules. Ces informations pourraient alors permettre une meilleure compréhension de l'impact des mutations sur les cellules souches et progénitrices, ouvrant la voie à l'utilisation de modèles plus complexes et plus représentatifs de la dynamique de prolifération des cellules souches mutées. Au prochain chapitre, nous nous placerons justement dans le cas de cellules mutées et proposerons un premier modèle de développement des néoplasmes myéloprolifératifs.

Références

[1] Ron Sender and Ron Milo. The distribution of cellular turnover in the human body. *Nature medicine*, 27(1) :45–48, 2021.

[2] Martha R Kirby and Robert E Donahue. Rare event sorting of cd34+ hematopoietic cells. *Annals of the New York Academy of Sciences*, 677(1) :413–416, 1993.

[3] Qian-Lin Hao, Ami J Shah, Flavia T Thiemann, Elzbieta M Smogorzewska, and Gay M Crooks. A functional comparison of cd34+ cd38-cells in cord blood and bone marrow. 1995.

[4] Jason Cosgrove, Lucie SP Hustin, Rob J de Boer, and Leïla Perié. Hematopoiesis in numbers. *Trends in Immunology*, 42(12) :1100–1112, 2021.

[5] Tamar Tak, Giulio Prevedello, Gaël Simon, Noémie Paillon, Camélia Benlabiod, Caroline Marty, Isabelle Plo, Ken R Duffy, and Leïla Perié. Hspcs display within-family homogeneity in differentiation and proliferation despite population heterogeneity. *Elife*, 10 :e60624, 2021.

[6] Miles B Horton, Giulio Prevedello, Julia M Marchingo, Jie HS Zhou, Ken R Duffy, Susanne Heinzel, and Philip D Hodgkin. Multiplexed division tracking dyes for proliferation-based clonal lineage tracing. *The Journal of Immunology*, 201(3) :1097–1103, 2018.

[7] David Axelrod and Marek Kimmel. *Branching processes in biology*. Springer-Verlag, 2015.

[8] Samik Upadhaya, Catherine M Sawai, Efthymia Papalexi, Ali Rashidfarrokhi, Geunhyo Jang, Pratip Chattopadhyay, Rahul Satija, and Boris Reizis. Kinetics of adult hematopoietic stem cell differentiation in vivo. *Journal of Experimental Medicine*, 215(11) :2815–2832, 2018.

[9] Aline Roch, Vincent Trachsel, and Matthias P Lutolf. Brief report : single-cell analysis reveals cell division-independent emergence of megakaryocytes from phenotypic hematopoietic stem cells. *Stem cells*, 33(10) :3152–3157, 2015.

[10] Tatyana Grinenko, Anne Eugster, Lars Thielecke, Beáta Ramasz, Anja Krüger, Sevina Dietz, Ingmar Glauche, Alexander Gerbaulet, Malte Von Bonin, Onur Basak, et al. Hematopoietic stem cells can differentiate into restricted myeloid progenitors before cell division in mice. *Nature communications*, 9(1) :1–10, 2018.

[11] Tatsiana Ryl, Erika E Kuchen, Emma Bell, Chunxuan Shao, Andrés F Flórez, Gregor Mönke, Sina Gogolin, Mona Friedrich, Florian Lamprecht, Frank Westermann, et al. Cell-cycle position of single myc-driven cancer cells dictates their susceptibility to a chemotherapeutic drug. *Cell systems*, 5(3) :237–250, 2017.

[12] Simon Mitchell, Koushik Roy, Thomas A Zangle, and Alexander Hoffmann. Nongenetic origins of cell-to-cell variability in b lymphocyte proliferation. *Proceedings of the National Academy of Sciences*, 115(12) :E2888–E2897, 2018.

[13] Erika E Kuchen, Nils B Becker, Nina Claudino, and Thomas Höfer. Hidden long-range memories of growth and cycle speed correlate cell cycles in lineage trees. *Elife*, 9 :e51002, 2020.

[14] Ken R Duffy, Cameron J Wellard, John F Markham, Jie HS Zhou, Ross Holmberg, Edwin D Hawkins, Jhagvaral Hasbold, Mark R Dowling, and Philip D Hodgkin. Activation-induced b cell fates are selected by intracellular stochastic competition. *Science*, 335(6066) :338–341, 2012.

[15] Magnus Urquhart, Emil Ljungskog, and Simone Sebben. Surrogate-based optimisation using adaptively scaled radial basis functions. *Applied Soft Computing*, 88 :106050, 2020.

[16] Stuart Bates, Johann Sienz, and Vassili Toropov. Formulation of the optimal latin hypercube design of experiments using a permutation genetic algorithm. In *45th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics & Materials Conference*, page 2011, 2004.

[17] Dennis Prangle. Summary statistics. In *Handbook of approximate Bayesian computation*, pages 125–152. Chapman and Hall/CRC, 2018.

[18] Michael GB Blum, Maria Antonieta Nunes, Dennis Prangle, and Scott A Sisson. A comparative review of dimension reduction methods in approximate bayesian computation. *Statistical Science*, 28(2) :189–208, 2013.

[19] Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate bayesian computation : semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 74(3) :419–474, 2012.

[20] Nikolaus Hansen. The cma evolution strategy : A tutorial. *arXiv preprint arXiv :1604.00772*, 2016.

[21] Nikolaus Hansen. The cma evolution strategy : a comparing review. *Towards a new evolutionary computation*, pages 75–102, 2006.

[22] Paul Joyce and Paul Marjoram. Approximately sufficient statistics and bayesian computation. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

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

[24] Elisa Laurenti and Berthold Göttgens. From haematopoietic stem cells to complex differentiation landscapes. *Nature*, 553(7689) :418–426, 2018.

[25] L Alexander Liggett and Vijay G Sankaran. Unraveling hematopoiesis through the lens of genomics. *Cell*, 182(6) :1384–1400, 2020.

[26] Mark HA Davis. Piecewise-deterministic markov processes : A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society : Series B (Methodological)*, 46(3) :353–376, 1984.

[27] Ryszard Rudnicki and Marta Tyran-Kamińska. Piecewise deterministic markov processes in biological models. In *Semigroups of operators-theory and applications*, pages 235–255. Springer, 2015.

[28] Trevor Hastie and Werner Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406) :502–516, 1989.

[29] Piotr Jaworski, Fabrizio Durante, Wolfgang Karl Hardle, and Tomasz Rychlik. *Copula theory and its applications*, volume 198. Springer, 2010.

Chapitre 5
Apparition et développement des Néoplasmes Myéloprolifératifs

Résumé

Les néoplasmes myéloprolifératifs sont souvent détectés à un âge avancé, après des complications telles que des thromboses et des événements cardiovasculaires. Comprendre la dynamique de ces hémopathies malignes (apparition et développement) pourrait aider à la mise en place de stratégies de dépistage précoce et, plus largement, à comprendre l'histoire du développement des cancers du sang. À cette fin, nous développons une approche mathématique combinant modélisation et inférence statistique et l'appliquons aux deux mutations les plus répandues dans les néoplasmes myéloprolifératifs, à savoir $JAK2^{V617F}$ et $CALR^m$.
Pour modéliser l'initiation du cancer, nous proposons un modèle mathématique qui consiste en une chaîne de Markov à temps continu. Dès qu'un certain nombre de cellules mutées est atteint, nous modélisons le développement de la maladie par une loi déterministe, approximation nécessaire pour calibrer notre modèle à partir d'observations de patients. L'estimation des paramètres de notre modèle se fait par une méthode ABC-SMC (Approximate Bayesian Computation - Sequential Monte Carlo).
Nos résultats mettent en évidence des différences dans l'expansion clonale de la mutation $CALR^m$ par rapport à $JAK2^{V617F}$, suggérant différents types de mécanismes pour l'acquisition de la mutation motrice.
Le contenu de ce chapitre a fait l'objet d'une publication (Hermange et al., PNAS 2022).

Abstract

Myeloproliferative Neoplasms (MPNs) blood cancers are characterized by abnormal proliferation of myeloid hematological cells. The developmental history of blood cancer begins with mutation acquisition and the resulting malignant clone expansion. The two most prevalent driver mutations found in MPNs - $JAK2^{V617F}$ and $CALR^{m}$ - occur in hematopoietic stem cells (HSCs), which are highly complex to observe *in vivo*. The precise role of these mutations in the development of clonal hematopoiesis and in the disease dynamics remains poorly understood. Of particular debate is the expansion kinetics of a malignant clone emanating from a single mutated cell. MPNs are generally detected at advanced ages when the malignant clone is present at significant Variant Allele Frequency in peripheral blood. Late detection, often after complications such as thrombosis and cardiovascular events, or more rarely at the leukemic stage, further increases the death rate. Methods to understand disease initiation and clonal expansion of mutated HSCs are thus crucial to develop adequate screening strategies and to avoid delayed medical care. A major difficulty for this task comes from the fact that human HSCs are difficult to access and *in vivo* expansion of mutated HSCs cannot be observed. We therefore propose a mathematical model to infer the dynamics of the disease, from the acquisition of the driver mutation to the appearance of the symptoms.

To understand how $JAK2^{V617F}$ and $CALR^{m}$ malignant clones might expand over time, we propose a model relying on a Continuous Time Markov Chain (CTMC) process. When a certain number of mutated HSCs is reached, we show numerically that a deterministic law can approximate the process. This approximation further facilitates the calibration of the model that otherwise would not be computationally feasible. We estimate the model parameters using an Approximate Bayesian Computation method based on Sequential Monte Carlo (ABC-SMC) with data from recently diagnosed $JAK2^{V617F}$ and $CALR^{m}$ patients. To achieve a faster convergence, we use an optimal affectation procedure based on the Hungarian algorithm. We apply this mathematical approach - combining modelling and statistical inference - to investigate potential differences in the disease dynamics of $CALR^{m}$ *vs* $JAK2^{V617F}$ MPN patients.

We evidence differences in the clonal expansion of the $CALR^{m}$ mutation compared to $JAK2^{V617F}$, suggesting different types of mechanisms for the driver mutation acquisition. Our findings suggest that $CALR^{m}$ mutations tend to occur later in life than $JAK2^{V617F}$. Our results confirm the higher proliferative advantage of the $CALR^{m}$ malignant clone compared to $JAK2^{V617F}$. Furthermore, we illustrate how mathematical modelling and Bayesian inference can be used for setting up early screening strategies.

Table des matières

1 Introduction

Les néoplasmes myéloprolifératifs (NMP) sont des hémopathies malignes caractérisées par une prolifération anormale des cellules myéloïdes. Ces pathologies résultent de mutations somatiques gain de fonction de gènes spécifiques. Trois mutations sont responsables de la maladie et affectent des gènes codant pour des protéines jouant un rôle crucial dans la signalisation cellulaire : MPL (MPL^m), JAK2 ($JAK2^{V617F}$) et la calréticuline (mutations de l'exon 9, $CALR^m$) [1, 2, 3, 4, 5]. Les deux mutations principales des NMP sont celles aux gènes *JAK2* et *CALR* ; ce seront celles que nous étudierons dans ce chapitre.
Jusqu'ici, nous nous sommes principalement placés dans le cas de l'hématopoïèse non pathologique, mis à part un bref détour au chapitre 2 pour étudier des différences de signalisation, au cours de la mégacaryopoïèse, chez des souris mutées $CALR^m$ par rapport à des souris Wild-Type (WT). À partir de ce chapitre, nous étudierions les NMP chez l'homme. Avant de considérer le traitement de patients atteints de cette maladie, nous commencerons par modéliser son apparition et développement sur plusieurs années. Le travail présenté dans ce chapitre est issu de notre article "*Inferring the initiation and development of Myeloproliferative Neoplasms*" publié dans le journal PNAS (2022).

Les mutations motrices des NMP surviennent dans les cellules souches hématopoïétiques (HSC). Le rôle précis de ces mutations dans le développement de l'hématopoïèse clonale et dans la dynamique de la maladie reste mal compris, en particulier concernant la cinétique d'expansion d'un clone malin émanant d'une unique cellule mutée. Alors qu'il a été rapporté que la mutation $JAK2^{V617F}$ était acquise des décennies avant l'apparition de la maladie, voire pendant la vie fœtale [6, 7, 8, 9], il n'y a pas de résultat similaire pour la mutation $CALR^m$. Les NMP sont généralement détectés à un âge avancé lorsque le clone malin est présent à une charge allélique (Variant Allele Frequency - VAF) significative dans le sang périphérique. La détection tardive, souvent après des complications telles que des thromboses et des événements cardiovasculaires, augmente le taux de mortalité. Les méthodes permettant de comprendre l'initiation de la maladie et l'expansion clonale des HSC mutées sont ainsi cruciales pour permettre le développement de stratégies de dépistage adéquates et éviter une prise en charge médicale tardive.
Une difficulté majeure pour cette tâche vient du fait que les HSC humaines sont difficiles d'accès et que l'expansion *in vivo* des HSC mutées n'est pas observable. Nous proposons dans ce chapitre un modèle mathématique permettant d'inférer la dynamique de la maladie, depuis l'acquisition de la mutation motrice jusqu'à l'apparition des symptômes. Les modèles stochastiques, tels que le modèle de Wright-Fisher utilisé en génétique des populations [10] ou les processus de branchements [11], sont des choix appropriés pour décrire l'expansion d'un clone malin à partir d'une cellule mutée unique. Cependant, leur calibration à partir d'observations réelles reste difficile et nécessite la mise en place d'une procédure d'optimisation adaptée afin d'obtenir une convergence des algorithmes en un temps réaliste.
Pour comprendre comment les clones mutés $JAK2^{V617F}$ et $CALR^m$ peuvent se développer au cours de la vie, nous proposons un modèle reposant sur une chaîne de Markov en temps continu (CTMC). Lorsqu'un certain nombre de HSC mutées est atteint, nous montrons numériquement qu'une loi déterministe peut approcher le processus. Cette approximation facilite la calibration du modèle qui, autrement, ne serait pas réalisable numériquement. Nous estimons la distribution *a posteriori* des paramètres du modèle à l'aide de la méthode Approximate Bayesian Computation - Sequential Monte Carlo (ABC-SMC) [12, 13, 14, 15, 16], avec des données provenant de patients atteints de NMP positifs pour la mutation $JAK2^{V617F}$ ou $CALR^m$. Pour obtenir une convergence plus rapide de nos algorithmes, nous utilisons une procédure d'affectation optimale basée sur l'algorithme hongrois [17, 18]. Nous appliquons notre méthode - combinant modélisation et inférence statistique - pour étudier les différences potentielles dans la dynamique de la maladie chez les patients atteints de NMP.

2 Observations expérimentales

2.1 Architecture clonale

Pour étudier le développement des néoplasmes myéloprolifératifs, nous nous basons sur les données de patients atteints de NMP, ayant soit comme mutation motrice $JAK2^{V617F}$, soit $CALR^m$. Pour chaque patient, nous disposons d'une observation, qui consiste en la proportion de cellules mutées parmi ses progéniteurs hématopoïétiques, et l'âge du patient à la date du prélèvement.
Les données sont obtenues en déterminant l'architecture clonale de progéniteurs $CD34^+$ purifiés, à partir d'échantillons de sang. Pour déterminer cette architecture clonale, à partir d'un échantillon de sang, de nombreuses étapes expérimentales sont nécessaires, telles que représentées sur la figure 1 (ainsi que sur les photos de la figure 2).
Dans ce travail, nous utilisons les données de 15 patients ayant la mutation $CALR^m$ (données publiées par El-Khoury et al. [19]) et de 11 patients ayant la mutation $JAK2^{V617F}$. Ont été exclus les patients présentant des HSC mutées homozygotes. En effet, les sous-clones mutés homozygotes apparaissent suite à une recombinaison homologue au niveau d'une cellule mutée hétérozygote, puis vont ensuite se développer avec un avantage plus important que pour les sous-clones hétérozygotes [7]. Intégrer les patients avec des clones homozygotes nécessiterait alors de modéliser le développement de deux sous-clones en parallèle, ainsi que l'apparition de l'homozygotie.
Les architectures clonales des progéniteurs hématopoïétiques et des cellules souches ont été mesurées avant que les patients ne commencent une thérapie qui ciblerait les HSC mutées (en l'occurrence, un traitement à l'IFNα, que nous étudierons au chapitre suivant). L'extraction et la purification des cellules souches hématopoïétiques (au sens conceptuel, c'est-à-dire capable de produire toute cellule hématopoïétique sur le long terme) est très difficile, et il est impossible d'obtenir des informations sur les véritables HSC. Ainsi, l'information que nous considérerons dans la plupart des cas - et lorsqu'elle est disponible - est la fraction clonale (CF, c'est-à-dire la fraction de cellules mutées) des cellules progénitrices enrichies en HSC ($CD90^+CD38^-CD34^+$, que nous notions aux chapitres précédents HSC*). Soit $(\hat{t}_i, \hat{\eta}_i)$ nos observations, où $\hat{\eta}_i$ est la CF des cellules progénitrices immatures du patient i mesurée à l'âge $\hat{t}_i$. La CF observée $\hat{\eta}_i$ parmi les cellules progénitrices immatures est donc une mesure indirecte du nombre de HSC mutées. Nous la considérerons comme une réalisation du processus stochastique $\eta(t)$ au temps $\hat{t}_i$ (voir eq. (2)).

2.2 Cohorte

Nous ne considérons que des patients ayant des clones hétérozygotes. Certains patients présentent des mutations associées à une hématopoïèse clonale ; les patients de notre cohorte peuvent donc être considérés comme représentatifs de la population atteinte de NMP.
Les tableaux 1 et 2 présentent la liste des patients dont les observations sont utilisées dans ce chapitre. La CF correspond à la Fraction Clonale en cellules hétérozygotes mesurée (en pourcentage) parmi le compartiment progéniteur correspondant (on choisit les cellules les plus immatures possibles). Pour les patients $CALR^m$, nous indiquons également le type de mutation (T1 ou T2), correspondant aux deux variants les plus fréquents chez les patients $CALR^m$. Nous donnons également les informations sur les autres mutations trouvées associées dans les mêmes HSC ou dans des clones séparés. Ces informations ont été obtenues à partir des mesures de VAF dans les cellules matures ; nous ne pouvons pas savoir si les autres mutations sont présentes ou non dans les sous-clones présentant la mutation motrice ($CALR^m$ ou $JAK2^{V617F}$). Nous ne prenons en compte que les observations des patients avant qu'ils ne commencent une thérapie sous IFNα (que nous étudierons plus en détail au chapitre suivant). De plus, nous ne considérons pas les patients présentant des sous-clones homozygotes, à l'exception des patients P4 et P28 qui présentaient respectivement une CF homozygote de 2% et 1.7%. Ces valeurs étant très faibles, nous avons supposé que les cellules mutées homozygotes avaient un comportement similaire à celui des cellules hétérozygotes et, par conséquent, la CF considérée dans le tableau est la somme des CF des cellules homozygotes et hétérozygotes.
La CF des cellules progénitrices a été mesurée sur la base des marqueurs de surface CD34, CD38 et CD90 et les cellules ont été triées selon les phénotypes suivants : $CD34^+CD38^-CD90^+$

(HSC* - progéniteurs enrichis en cellules souches), $CD34^{+}CD38^{-}CD90^{-}$ (MPP - progéniteurs multipotents) et $CD34^{+}CD38^{+}$ (HPC - progéniteurs au potentiel plus restreint).

ID	ID_0	Âge	CF [%]	Type cellulaire	Autres mutations	Maladie
1	P2	34	35	CD34+CD38-	Aucune	PV
2	P4	58	70.5	CD34+CD38-CD90+	TET2fS1906Rfs (25%)	PV
3	-	39	5	CD34+CD38-CD90+	TET2 Y1679Pfs* (45%) DNMT3A N727D (20%)	ET
4	P11	51	50	CD34+CD38-	SUZ12 Ala33Val (61%)	PV
5	P14	35	5	CD34+CD38+CD90-	Aucune	TE
6	-	57	14.5	CD34+CD38-CD90+	Aucune	PV
7	P20	45	2	CD34+CD38-	Aucune	PV
8	P22	53	19	CD34+CD38-CD90+	Aucune	PV
9	-	29	17	CD34+CD38-CD90+	Aucune	TE
10	-	61	32	CD34+CD38-CD90+	Aucune	TE
11	P28	51	13	CD34+CD38-	Aucune	PV

TABLE 1 – Observations pour les patients mutés $JAK2^{V617F}$. L'âge est exprimé en années. ID_0 fait référence au label utilisé au chapitre suivant. Les informations sur les autres mutations sont obtenues à partir de mesures dans des cellules matures. Le pourcentage fait référence à une mesure de VAF. La dernière colonne indique la maladie du patient. TE pour Thrombocytémie Essentielle ; PV pour Polyglobulie de Vaquez.

ID	ID_0	Type	Âge	CF [%]	Type cellulaire	Autres mutations	Maladie
1	P166	1	48	30.4	CD34+CD38-CD90+	Aucune	TE
2	P46	1	58	100	CD34+CD38-CD90+	Aucune	TE
3	P30	1	70	91.8	CD34+CD38-CD90+	Aucune	TE
4	P48	1	53	68.2	CD34+CD38-	TP53 V143M (7%)	TE
5	P90	1	58	7.7	CD34+CD38-CD90+	Aucune	TE
6	P169	1	46	100	CD34+CD38-CD90+	Aucune	TE
7	P38	1	66	100	CD34+CD38-CD90+	ASXL1 G658* (6%) ASXL1 E635Rfs (40%) EZH2 R690H (27%)	PMF
8	P157	1	44	50	CD34+CD38-CD90+	SF3B1 K666N (28%) JAK2V617F (2%)	PMF
9	P28	2	53	99.7	CD34+CD38-CD90+	Aucune	TE
10	P54	2	49	100	CD34+CD38-CD90+	Aucune	TE
11	P103	2	60	100	CD34+CD38-CD90+	Aucune	TE
12	P53	2	69	24	CD34+CD38-CD90+	DNMT3A L901R (22%)	TE
13	P187	2	86	79	CD34+CD38-CD90+	ASXL1 L775fs* (2%)	PMF
14	P170	2	70	94	CD34+CD38-CD90+	Aucune	TE
15	P52	2	44	5	CD34+CD38-CD90+	Aucune	TE

TABLE 2 – Observations pour les patients mutés $CALR^m$. L'âge est exprimé en années. ID_0 fait référence au label utilisé par El-Khoury et al. [19]. Nous indiquons si la mutation $CALR^m$ est de type 1 ou 2. L'information sur les autres mutations - associées dans les mêmes HSC ou dans des HSC différentes de celles portant la mutation motrice - est obtenue à partir de mesures dans des cellules matures. Le pourcentage fait référence à une mesure de VAF. La dernière colonne indique la maladie du patient. TE pour Thrombocytémie Essentielle ; PMF pour Myélofibrose Primaire.

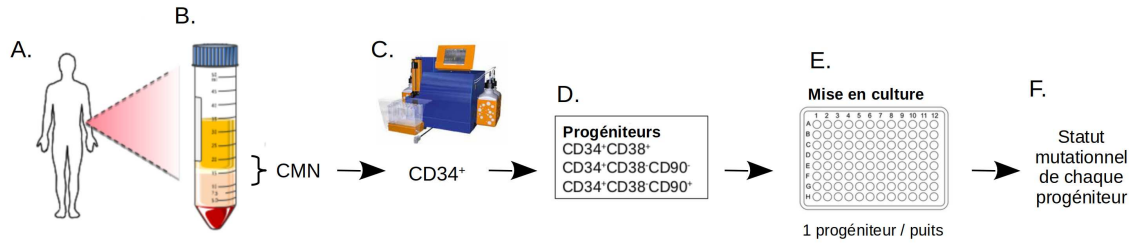


FIGURE 1 – Protocole expérimentale simplifié pour déterminer l'architecture clonale (adapté de celui d'Amandine Tisserand). Un échantillon de sang pour un patient i est prélevé au temps $\hat{t}_i$ (A). Le tube contenant le sang est tout d'abord placé dans une centrifugeuse pour séparer les plaquettes du reste du sang. L'ajout de Ficoll, suivi d'une étape de centrifugation, permet alors de séparer différentes phases (B) : de haut en bas, on distingue le plasma, un anneau blanc de cellules contenant les cellules mononucléées (CMN) et un culot contenant érythrocytes et granulocytes. Les progéniteurs $CD34^+$ seront trouvés dans l'anneau de CMN. Pour cela, on récupère cette couche de cellules à laquelle on ajoute des billes magnétiques couplées à des anticorps dirigés contre le marqueur CD34. Le tube ainsi constitué sera placé dans l'autoMACS (Automated cell isolation) pour séparer les cellules $CD34^+$ des $CD34^-$ (C). Les cellules sont ensuite marquées avec des anticorps couplés à des fluorochromes (anticorps anti-CD34 FITC, anti-CD38 PE et anti-CD90 APC) et triées par cytométrie de flux (D). Les cellules, dont on connaîtra alors le type cellulaire, sont mises en culture à une cellule par puits (E), en présence d'un cocktail de cytokines. 15 jours plus tard, les colonies issues des cellules initiales seront génotypées (*Taqman allelic discrimination* qPCR), permettant de déterminer rétrospectivement pour chacune le statut mutationel, à savoir si elle est mutée $JAK2^{V617F}$, ou $CALR^m$, et si la mutation est présente sur un allèle (cellule hétérozygote) ou deux allèles (cellule mutée homozygote). On en déduira la fraction clonale en cellules mutées hétérozygotes $\hat{\eta}_i$ du patient (F).

FIGURE 2 – Photos prises lors d'une manipulation expérimentale réalisée par Amandine Tisserand à l'Institut Gustave Roussy. L'expérience consistait en la purification et séparation de cellules de sang de patient. À gauche, photo de la paillasse, à droite photo de l'autoMACS.

3 Modèle

3.1 Développement d'une clone muté

Le modèle que nous proposons dans cette section a pour objectif de décrire l'expansion clonale à partir d'une première HSC mutée, parmi un ensemble de cellules souches saines (WT) dont on suppose le nombre N_{WT} maintenu constant par un mécanisme de régulation quelconque, correspondant à des conditions homéostatiques. T_0 est l'âge auquel la première mutation du gène considéré (soit $CALR$ ou $JAK2$) est acquise. Nous étudierons la dynamique du nombre de cellules mutées $N(t)$ au cours du temps $t \geq T_0$ (avec $N(T_0) = 1$).
Nous supposerons que toutes les HSC mutées sont indépendantes les unes des autres, et qu'elles forment un ensemble homogène. On considère que les HSC mutées se divisent à un taux α. Trois mécanismes de division, étudiés au précédent chapitre dans les cas des progéniteurs, sont admis pour les cellules souches (Fig. 3) :

- Une HSC peut effectuer une division différenciée (diff), avec une probabilité p_0, et générer deux cellules progénitrices (très immatures). Les cellules progénitrices continueront à proliférer et à se différencier, aboutissant à la production de nos cellules sanguines (voir § 3.2).
- Une HSC peut se diviser de manière asymétrique (asym), avec une probabilité p_1, donnant une cellule progénitrice et une cellule souche, cette dernière étant fonctionnellement identique à la cellule mère.
- Une HSC peut se diviser de manière symétrique (sym), avec une probabilité $p_2 = 1 - p_1 - p_0$, produisant deux HSC (voir eq. (1)).

Dans ce chapitre, nous nous concentrons sur les HSC, car ce sont elles qui sont à l'origine de la production de tous les types de cellules sanguines et qui seraient initialement porteuses de la mutation $JAK2^{V617F}$ ou $CALR^m$ dans le cas des NMP (même si d'autres auteurs ont étudié également la possibilité que la mutation soit initialement acquise par une cellule progénitrice [20]). Nous introduisons $\Delta = p_2 - p_0$. Pour les cellules WT, ce paramètre est égal à zéro afin de maintenir des conditions homéostatiques. Nous considérons que les cellules présentant des mutations ont un avantage prolifératif au niveau des cellules souches, c'est-à-dire que $\Delta > 0$ (sinon la probabilité d'extinction serait égale à 1).
Si le temps entre les divisions était constant, le processus décrit ci-dessus serait un processus de branchement multi-type standard, largement utilisé pour modéliser les processus de prolifération cellulaire [11]. Dans ce cas, nous pourrions calculer une probabilité d'extinction des cellules souches mutées égale à $q = p_0/p_2 < 1$ (puisque nous supposons que $\Delta > 0$). Dans notre modèle, le temps entre deux divisions n'est pas constant mais suit une distribution exponentielle avec un

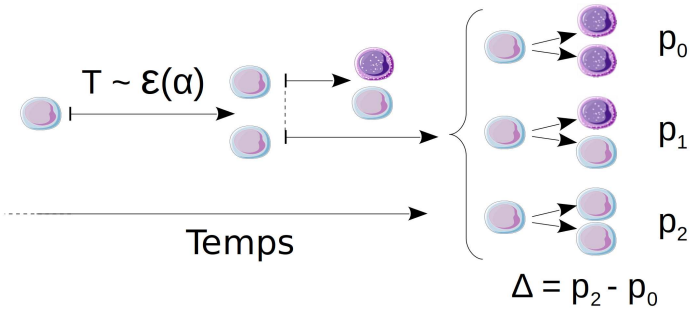


FIGURE 3 – Modélisation de la dynamique de prolifération des cellules souches : une HSC mutée donnée se divise après un temps aléatoire T qui suit une loi exponentielle $\mathcal{E}(\alpha)$ de taux α. Après s'être divisée, la HSC mutée donne naissance à 0, 1, ou 2 HSC mutées avec des probabilités p_0, p_1, ou p_2, respectivement. Les cellules filles, si elles ne sont pas des HSC, sont des cellules progénitrices (représentées par une couleur plus foncée). Sur cet exemple, la première division est symétrique, et la division d'une des cellules filles (en haut) est asymétrique.

taux égal à α. Ainsi, le nombre $N(t)$ de cellules mutées et son expansion potentielle au cours du temps t suit un processus stochastique qui peut être décrit comme une chaîne de Markov à temps continu (CTMC). Il peut également être écrit en utilisant le formalisme des réseaux de réactions chimiques :

$$\begin{aligned} &\text{diff :} && \text{HSC} \xrightarrow{\alpha p_0} \emptyset \\ &\text{asym :} && \text{HSC} \xrightarrow{\alpha p_1} \text{HSC} \\ &\text{sym :} && \text{HSC} \xrightarrow{\alpha p_2} 2\text{HSC} \end{aligned} \tag{1}$$

Avec ce processus, il existe une probabilité $q = p_0/p_2$ que le clone muté s'éteigne naturellement. S'il ne disparaît pas, le clone muté va s'étendre au fil du temps, jusqu'à ce que le nombre de HSC mutées devienne potentiellement très élevé. Ce type de modèle décrivant l'expansion d'un clone malin est classique, et n'est pas éloigné de celui utilisé par exemple par Watson et al. [8].
Dans tout ce chapitre, nous nous concentrons sur les trajectoires sans extinction.

3.2 Modèle pour les progéniteurs et cellules matures

Le processus stochastique introduit dans le paragraphe précédent décrit la dynamique des HSC mutées. Or, en pratique, il n'est pas possible d'obtenir des informations pour de véritables HSC, mais plutôt des cellules progénitrices (ou des cellules matures en routine clinique). Par progéniteur, nous entendons les cellules de type HPC, MPP ou HSC*. Ensuite, pour confronter notre modèle aux observations expérimentales (voir § 2), nous devons également décrire la dynamique des dernières étapes de l'hématopoïèse. Dans notre modèle, nous introduisons alors deux compartiments supplémentaires : un pour les cellules progénitrices et un pour les cellules matures (voir Fig. 4). Au niveau des cellules souches, nous considérons la dynamique $N(t)$ du nombre de HSC mutées au cours du temps $t > T_0$. Cependant, les observations associées aux cellules progénitrices ou aux cellules matures ne sont pas des mesures directes du nombre de cellules mutées mais plutôt des proportions.
En considérant qu'en moyenne, sur un temps court δt, $\alpha\delta t N_{WT}$ HSC WT génèrent la même quantité de cellules progénitrices (puisqu'il existe un équilibre entre les divisions différenciées et symétriques pour assurer les conditions homéostatiques de ces cellules saines) et $\alpha\delta t N$ HSC mutées donneront naissance à $\alpha(1-\Delta)\delta t N$ cellules progénitrices immatures, on obtient pour la fraction clonale des cellules progénitrices :

$$\eta(t) = \frac{(1-\Delta)N(t)}{(1-\Delta)N(t) + N_{WT}} \tag{2}$$

En fait, la relation (2) est une approximation qui ne serait pas totalement correcte dans les premiers stades d'expansion de la maladie, lorsque le nombre de HSC mutées est trop faible. Cependant, lorsque les patients sont diagnostiqués, le clone malin a déjà pris de l'ampleur, et notre relation devient valide.

Les cellules progénitrices donnent ensuite naissance à des cellules matures. En routine clinique, au lieu de mesurer une CF, c'est-à-dire une proportion de cellules mutées, nous mesurons une charge allélique (VAF), c'est-à-dire une proportion d'allèles mutés. Dans ce chapitre, comme nous ne considérons que les cellules mutées hétérozygotes, la CF est le double du pourcentage de la VAF. À noter que le terme CF est généralement associé aux cellules progénitrices alors que le terme VAF fera toujours référence - dans ce chapitre - à une mesure de la VAF dans le sang périphérique. La mesure de la VAF n'est pas un bon indicateur de la CF parmi les progéniteurs puisque les cellules progénitrices et précurseurs $JAK2^{V617F}$ mais aussi, dans une moindre mesure, $CALR^m$ prolifèrent davantage que les cellules WT au cours des derniers stades de l'hématopoïèse. En étendant l'équation (2), la VAF dans le sang périphérique au temps t est égale à :

$$VAF(t) = 0.5\frac{(1-\Delta)k_m N(t)}{(1-\Delta)k_m N(t) + N_{WT}} \tag{3}$$

où k_m modélise l'avantage prolifératif des cellules mutées, des progéniteurs aux cellules matures, par rapport aux cellules WT.

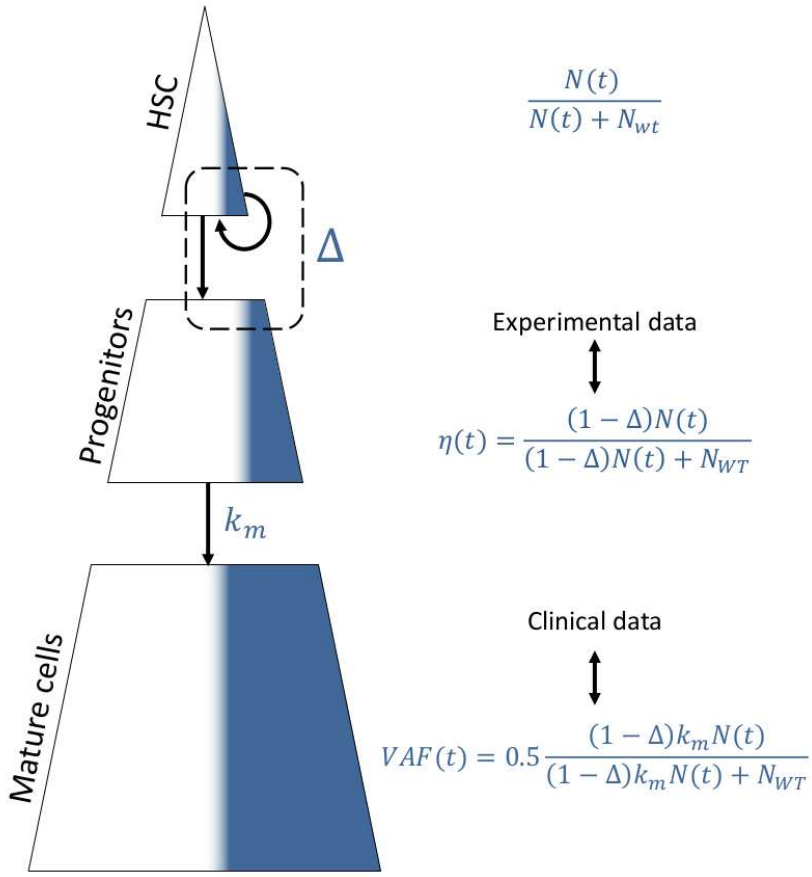


FIGURE 4 – Représentation schématique de notre modèle. Au sommet de l'arbre hématopoïétique, on trouve les HSC qui peuvent s'auto-renouveler. Si $N(t)$ représente le nombre de HSC mutées et N_{WT} le nombre de HSC WT, la CF parmi ce compartiment au temps t serait égale à $\frac{N(t)}{N(t)+N_{WT}}$. Or, les HSC ne peuvent pas être observées. Ce qui peut être mesuré est la CF parmi les cellules progénitrices $\eta(t)$. Les paramètres de notre modèle, notamment Δ, représentant l'équilibre entre les divisions symétriques et différenciées pour les HSC mutées, peuvent être inférrés à partir de mesures de CF parmi des cellules progénitrices. Ensuite, les cellules progénitrices se développent et se différencient pour devenir des cellules matures qui finissent par mourir. Les cellules progénitrices mutées auraient également un avantage prolifératif par rapport aux cellules WT, de sorte qu'une cellule progénitrice mutée produira plus de cellules matures qu'une cellule progénitrice WT. Cet avantage prolifératif au niveau des cellules engagées est décrit par le paramètre k_m. En routine clinique, on mesure la VAF parmi les cellules matures.

3.3 Acquisition de la mutation

Le processus stochastique décrit au § 3.1 commence avec une cellule mutée au temps $t = T_0$. le processus et la dynamique de l'acquisition des mutations motrices font encore l'objet de débats, différentes études aboutissant à des conclusions différentes dans le cas de la mutation $JAK2^{V617F}$ et presque aucune étude (mis à part l'article récemment paru de Sousos et al. [21]) pour la mutation du gène $CALR$. Williams et al. [7] ont rapporté des cas dans lesquels la mutation $JAK2^{V617F}$ a été acquise pendant la vie fœtale. Van Egeren et al. [6] ont montré, en reconstruisant la phylogénie de cellules hématopoïétiques, que la mutation pouvait être acquise après la naissance et des décennies avant l'apparition de la maladie. Watson et al. [8] ont obtenu des résultats similaires en utilisant une autre approche.
Nous faisons alors l'hypothèse d'un temps d'acquisition T_0 qui suit une distribution exponentielle. Sans connaissance *a priori*, la loi exponentielle est un choix approprié et pratique qui n'introduit qu'un seul paramètre supplémentaire à estimer, à savoir la moyenne de la distribution exponentielle désignée par $\lambda > 0$. $T_0 = 0$ correspond dans notre modèle à une acquisition durant la vie

fœtale ; il ne s'agit pas d'un événement ponctuel mais plutôt d'une période d'environ 0.75 an pendant laquelle le *pool* de cellules souches WT se développe rapidement.
Nous étudions donc deux hypothèses : soit la mutation se produit (pour tous les patients de l'une de nos deux populations d'intérêt, c'est-à-dire avec $CALR^m$ ou avec $JAK2^{V617F}$) pendant la vie fœtale, c'est-à-dire $T_0 = 0$, soit plus tard après la naissance, $T_0 > 0$ (voir figure 5). Nous considérons que cette dernière hypothèse correspond à $\lambda > 0$ alors que la première correspond au cas limite $\lambda = 0$, le problème de la sélection du modèle (sélection entre les deux hypothèses) se réduit donc à une question d'estimation des paramètres. Nous considérons *a priori* que $\lambda = 0$ avec une probabilité non nulle, de telle sorte que son *posterior* pourrait être une distribution *zero-inflated*. $\mathbb{P}[\lambda = 0|\mathcal{D}]$ correspondra à la probabilité d'acquisition de la mutation durant la vie fœtale (voir § 4.4).

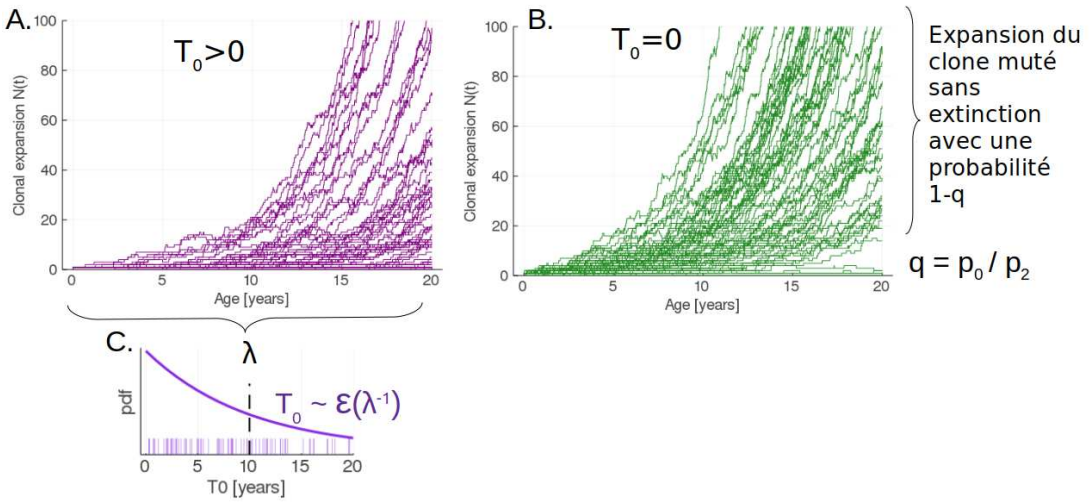


FIGURE 5 – Modélisation de l'expansion stochastique du clone muté à partir d'une seule cellule mutée, où $N(t)$ est le nombre de HSC mutées au cours du temps t. Deux hypothèses sont étudiées pour la modélisation de l'acquisition de la mutation : soit A) après la naissance ($T_0 > 0$), soit B) pendant la vie fœtale ($T_0 = 0$). Dans le premier cas, T_0 suit une loi exponentielle $\mathcal{E}(1/\lambda)$, de taux $1/\lambda$ (où $\lambda > 0$ correspond au temps moyen d'acquisition), dont la densité de probabilité (pdf) est représentée en (C). Pour chaque scénario, 100 simulations sont effectuées, avec $\alpha = 1/30$, $\Delta = 0.02$, $q = 1/3$, et $\lambda = 10$ (pour (A) uniquement).

3.4 Paramètres et priors

Le tableau 3 répertorie tous les paramètres du modèle avec leurs valeurs associées ou leurs distributions *a priori*. Les paramètres ne sont pas définis au niveau individuel (patient) mais au niveau d'une population de patients (soit ayant la mutation $JAK2^{V617F}$, soit $CALR^m$). On suppose donc que l'hétérogénéité entre les patients résulte de la stochasticité du modèle. À noter que les distributions *a posteriori* que nous allons estimer pour chaque population permettent cependant de prendre en compte en partie l'hétérogénéité intra-population.
Dans l'implémentation de notre modèle, nous considérons p_0 et p_2 qui peuvent varier entre 0 et 1, selon certaines contraintes. Cependant, nous préférons considérer Δ et q comme paramètres de notre modèle pour leur interprétation biologique.

Pour le paramètre α - qui modélise le taux auquel les HSC mutées sont recrutées pour se diviser - nous choisissons un *prior* tel que α prenne des valeurs autour de 1/30 [jours^{-1}]. Nous choisissons une distribution de probabilité définie sur $\mathbb{R}_+^*$, ici une loi gamma (mais d'autres choix auraient été possibles).
Il est difficile d'estimer précisément le temps moyen entre deux divisions d'une HSC WT, plus encore dans le cas d'une cellule mutée. Au chapitre 3, nous avions estimé des temps moyen

Paramètre	Valeur	Prior	Référence
N_{WT}	10^5	-	Lee-Six et al. [22]
α	À estimer	$\sim \begin{cases} \text{loi Gamma} \\ \text{valeur moyenne} = 1/30\ [\text{j}^{-1}] \\ \text{coefficient de variation} = 0.1 \end{cases}$	[22, 23, 24, 25]
(p_0, p_2)	À estimer	$\sim \begin{cases} \mathcal{U}([0,1]^2) \\ 0 \leq p_0 \leq p_2 \\ p_0 + p_2 \leq 1 \end{cases}$	-
p_1	$p_1 = 1 - p_2 - p_0$	$\in [0,1]$	-
Δ	$\Delta = p_2 - p_0$	$\Delta \geq 0$	-
q	$q = p_0/p_2$	$q \leq 1$	-
λ	À estimer	$\lambda \sim \begin{cases} 0 & \text{w.p } 0.5 \\ \mathcal{U}([0,100])\ [\text{years}] & \text{w.p } 0.5 \end{cases}$	[7, 6]
k_m	$\begin{cases} 7.5 & \text{pour } JAK2^{V617F} \\ 5.0 & \text{pour } CALR^m \end{cases}$	-	Chapitre 6

TABLE 3 – Liste des paramètres du modèle. Pour chacun d'entre eux, nous indiquons leur valeur, ou comment ils sont liés entre eux, ou "À estimer". Dans ce dernier cas, on indique la distribution *a priori* qu'ils suivent. $\mathcal{U}$ indique une distribution uniforme. "w.p" signifie "*with probability*".

de première division pour des HSC* (c'est-à-dire des progéniteurs enrichis en cellule souches $CD34^+CD38^-CD90^+$) de l'ordre de 2 jours. Néanmoins, il s'agissait de cellules progénitrices immatures étudiées *in vitro*, placées dans un cocktail de cytokines favorisant leur prolifération. D'après l'estimation de Catlin et al. [24] basée sur un modèle d'inactivation du chromosome X chez des femmes, les HSC s'auto-renouvelleraient (division symétrique) environ toutes les 40 semaines (entre 25 et 50 semaines). S'il en était de même pour les cellules mutées, ce temps correspondrait dans notre modèle à $1/(\alpha p_2)$. D'autres études ont porté sur l'étude de la diminution de la taille des télomères pour des cellules hématopoïétiques. Rufer et al. [25] ont estimé une diminution moyenne de 39 pb (paires de bases) par an pour des granulocytes (en *bulk*). Considérant qu'une cellule perdrait entre 50 et 100 pb par an [26, 27] (une autre estimation fait état d'une réduction de 30 à 100 pb par mitose pour des HSC [28]), Rufer et al. en déduisent qu'une cellule souche se diviserait en moyenne moins d'une fois par an. En construisant un arbre phylogénétique à partir de l'identification de mutations somatiques de cellules hématopoïétique d'un individu sain de 59 ans, Lee-Six et al. ont estimé que le temps entre deux divisions symétriques (auto-renouvellement) serait compris entre 2 et 20 mois [22]. Plus récemment, Mitchell et al. ont retrouvé des résultats comparables, estimant qu'il n'y aurait pas plus de une à deux divisions symétriques par an [23]. En considérant une division symétrique tous les 2 à 20 mois, et une probabilité d'avoir une division symétrique lors de la division d'une cellule souche valant 0.4 (qui est par exemple l'estimation faite au chapitre 3 dans le cas des HSC*), on estimerait qu'une HSC se diviserait en moyenne tous les 24 à 243 jours. Cet ordre de grandeur serait valable pour une cellule WT mais dans le cas muté, on pourrait s'attendre à ce que, lors de l'envahissement du compartiment des cellules souches, les HSCs mutées soient plus actives (mois quiescentes) que dans le cas WT, ce qui correspondrait ici à un taux de division plus élevé que dans le cas WT. À notre connaissance, il n'existe pas de valeurs quantifiant les différences dans le *timing* de divisions entre cellules mutées et WT. Nous choisissons ainsi de considérer *a priori* un temps moyen entre deux divisions dont l'ordre de grandeur est celui du mois.

Pour le paramètre λ, nous considérons *a priori* une distribution *zero-inflated*, où la probabilité

a priori que $\lambda = 0$ est fixée à 0.5, et sinon λ est uniformément distribué sur $[0, 100]$ [années]. On peut considérer que $\lambda = 0$ est une notation abusive. En fait, si $\lambda = 0$, alors nous n'échantillonnons pas - pour chaque patient de la population considérée - $T_0 \sim \mathcal{E}(\lambda^{-1})$, mais nous considérons plutôt que $T_0 = 0$. Autrement dit, $\lambda = 0$ correspond à l'hypothèse d'une acquisition de la mutation dans le fœtus alors que $\lambda > 0$ correspond à l'hypothèse d'une acquisition de mutation au cours de la vie. Cette problématique de l'estimation des paramètres avec une distribution *a priori zero-inflated* est équivalente à une problématique de sélection de modèle, où nos deux modèles seraient définis comme $\{T_0 = 0\}$ (acquisition dans la vie fœtale) et $\{T_0 > 0\}$ (acquisition au cours de la vie). Plus de détails sur l'implémentation au § 4.4.

4 Estimation des paramètres

4.1 Procédure ABC-SMC

Pour l'estimation des paramètres du modèle, nous nous plaçons dans un cadre Bayésien. Soit $\mathcal{D} = (\hat{t}_i, \hat{\eta}_i)_{i \in \{1,\cdots,N_p\}}$ nos N_p observations (CF $\hat{\eta}_i$ mesurées parmi les cellules progénitrices les plus immatures possible au temps $\hat{t}_i$ pour l'individu i de l'une des deux populations de patients considérées : soit $JAK2^{V617F}$ ou $CALR^m$) et $\boldsymbol{\theta}$ le vecteur de paramètres à estimer.
Notre objectif est d'estimer la distribution *a posteriori* de $\boldsymbol{\theta}$ étant donné les data $\mathcal{D}$:

$$p[\boldsymbol{\theta}|\mathcal{D}] \propto p[\mathcal{D}|\boldsymbol{\theta}]\ p[\boldsymbol{\theta}] \tag{4}$$

N'ayant pas d'expression analytique pour la vraisemblance, nous allons approcher notre *posterior* en utilisant une méthode ABC-SMC. Avec cette méthode, la distribution *a posteriori* est obtenue par *iterative rejection sampling*, en utilisant une séquence décroissante de tolérances [12, 16, 29]. La distribution *a posteriori* du vecteur de paramètres sera approchée à partir de M particules, c'est-à-dire M vecteurs de paramètres. Pour une particule donnée, un vecteur de paramètres est échantillonné (initialement à partir de la distribution *a priori*), le modèle est simulé, et génère autant de dynamiques que nous avons d'observations. L'erreur quadratique (distance d) entre les données et les simulations est calculée. Si d est inférieure à la tolérance considérée (à l'étape actuelle de l'algorithme), le vecteur de paramètres est attribué à la particule. Dans le cas contraire, les étapes précédentes sont répétées. Lorsque chaque particule est associée à une valeur de paramètres, la méthode est répétée avec une tolérance inférieure jusqu'à ce que l'algorithme ait convergé, c'est-à-dire jusqu'à ce que l'estimation de la distribution postérieure ne change plus lorsque l'on diminue la tolérance.

Notre modèle a été implémenté à l'aide du langage de programmation Julia (et disponible sur GitLab[1]). Le framework ABC-SMC utilisé pour calibrer le modèle a été implémenté en Julia par Mahmoud Bentriou dans le cadre de sa thèse de doctorat au laboratoire MICS sous la direction de Paul-Henry Cournède et Paolo Ballarini [14, 15]. Le code est disponible sur GitLab au lien suivant :
https ://gitlab-research.centralesupelec.fr/2017bentrioum/markovprocesses.jl.

Toutes nos estimations ont été calculées en utilisant 2,000 particules. Pour chaque particule, nous simulons autant de trajectoires (qui ne conduisent pas à extinction) que d'observations. Chaque trajectoire j nous donne le nombre de HSC mutées $N_j(t)$ au cours du temps. En utilisant l'équation (2), nous obtenons les trajectoires $\eta_j(t)$ des fractions clonales parmi les cellules progénitrices immatures. Nous assignons chaque trajectoire à une et une seule observation $(\hat{t}_i, \hat{\eta}_i)$ en utilisant une procédure d'affectation optimale basée sur l'algorithme hongrois [18, 17], minimisant une erreur quadratique entre les trajectoires et nos observations (voir § 4.3). La distance finale que nous calculons est la moyenne de la norme L^2 entre toutes nos trajectoires et les observations. Nous progressons d'une étape à l'autre en choisissant comme prochaine tolérance la médiane calculée sur l'étape précédente.

1. https ://gitlab-research.centralesupelec.fr/2012hermangeg/mpn-development

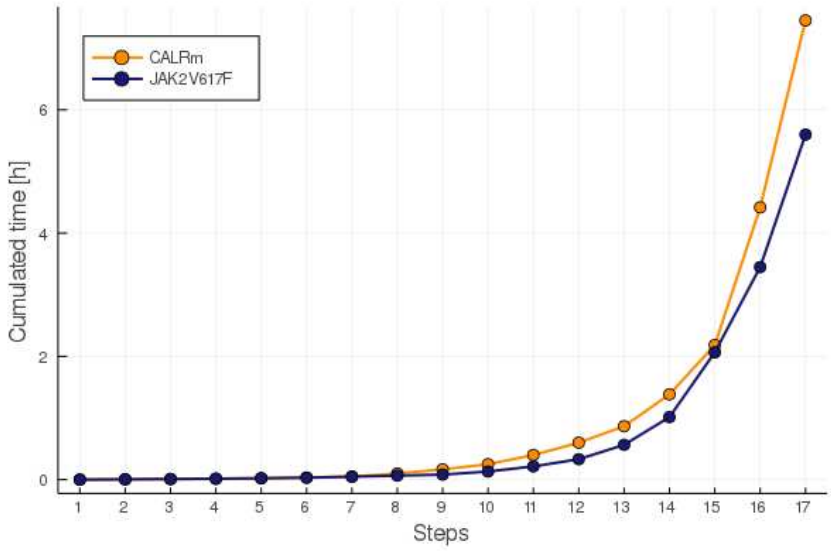


FIGURE 6 – Temps cumulé (en heures) nécessaire à la procédure ABC-SMC pour progresser au fil des étapes (avec 250 processeurs utilisés en parallèle) dans les cas $JAK2^{V617F}$ (bleu) et $CALR^m$ (orange).

À noter que ces calculs sont très coûteux en temps et ont été exécutés sur 250 processeurs en parallèle sur un cluster HPC.
Nous arrêtons l'algorithme lorsque la procédure ne passe pas à l'étape suivante en moins de 6 heures, avec 250 processeurs (voir Fig. 6). Nous regardons ensuite l'évolution de la distance L^2 (Fig. 7) et vérifions la convergence de l'algorithme en regardant l'évolution de la distribution *a posteriori* de nos paramètres sur les dernières étapes (Fig. 8). Enfin, nous exécutons également la procédure ABC-SMC une seconde fois avec une autre graine aléatoire pour nous assurer que nous obtenons les mêmes résultats dans les deux calculs. Pour chacune des populations de patients $JAK2^{V617F}$ et $CALR^m$, 17 étapes permettent une convergence correcte vers la distribution a *posteriori*, comme présenté sur la figure. 8. Ainsi, dans le cas $CALR^m$, nous effectuons nos calculs jusqu'à ce que nous atteignions une tolérance finale égale à 0.0021, et dans le cas $JAK2^{V617F}$ égale à 0.0038.
Comme on peut le voir sur les figures 6 et 7, le coût de calcul pour atteindre les étapes 16 et 17 est élevé alors que le gain sur l'erreur d'approximation est faible. Comme le montre la figure 8, la procédure pourrait raisonnablement être arrêtée à l'étape 16 (à la fois pour $CALR^m$ et $JAK2^{V617F}$).

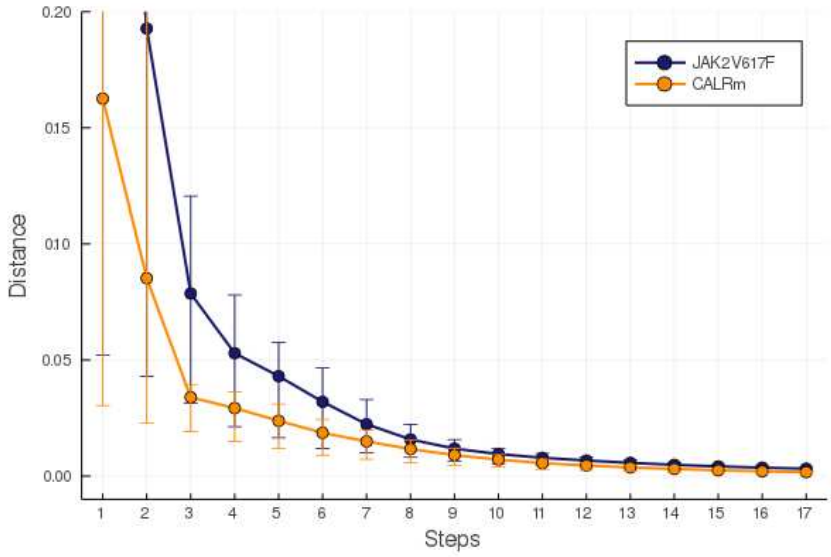


FIGURE 7 – Norme L^2 (distance) en fonction du nombre d'étapes utilisées pour la calibration du modèle dans les cas $JAK2^{V617F}$ (bleu) et $CALR^m$ (orange). Les points représentent les distances moyennes calculées sur les 2,000 particules. Les barres d'erreur représentent un intervalle de confiance de 90%. Pour plus de clarté, nous avons tronqué l'axe des ordonnées.

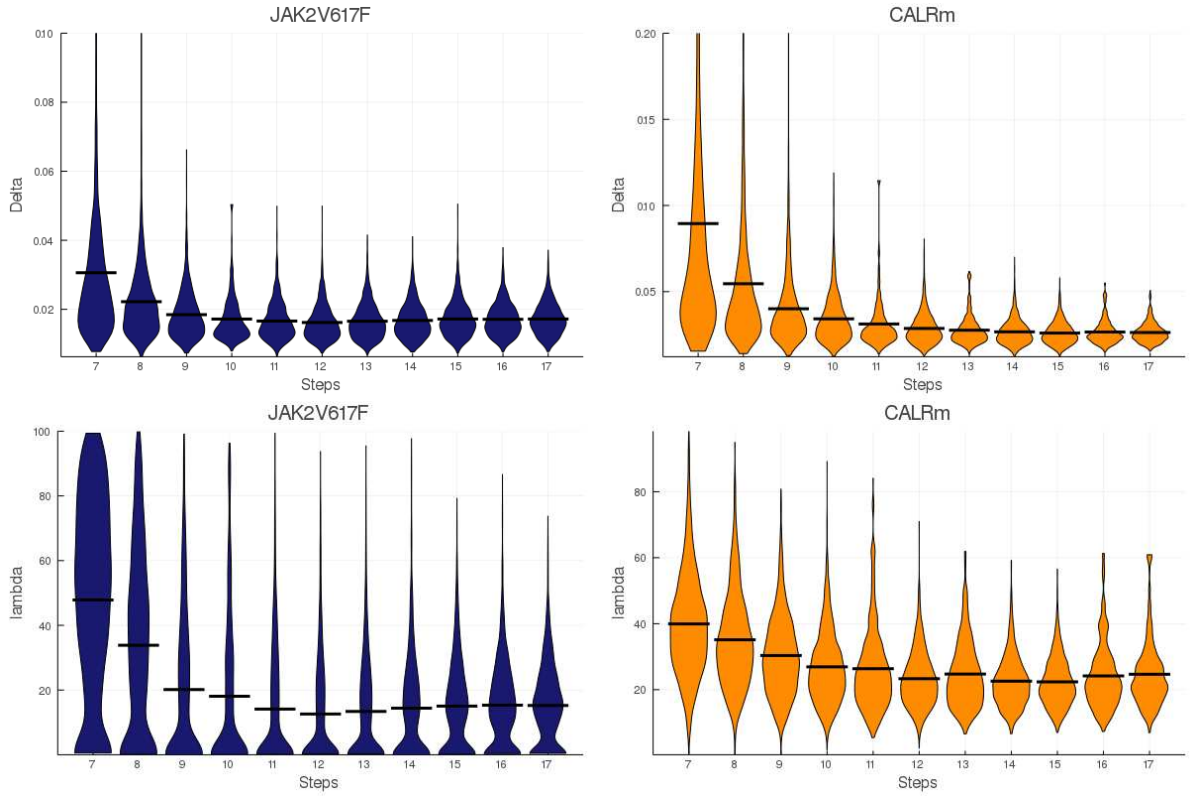


FIGURE 8 – Distributions *a posteriori* des paramètres Δ (en haut) et λ (en bas) pour les populations de patients $JAK2^{V617F}$ (à gauche) et $CALR^m$ (à droite) au fil des étapes. L'axe des abscisses commence à l'étape 7 pour plus de clarté. Les lignes horizontales indiquent les moyennes *a posteriori*.

4.2 Approximation déterministe

Notre modèle stochastique est bien adapté pour décrire l'expansion d'un clone malin à partir d'une seule cellule mutée. En effet, lorsque le nombre de cellules mutées est faible, les effets stochastiques jouent un rôle crucial (Fig. 9).

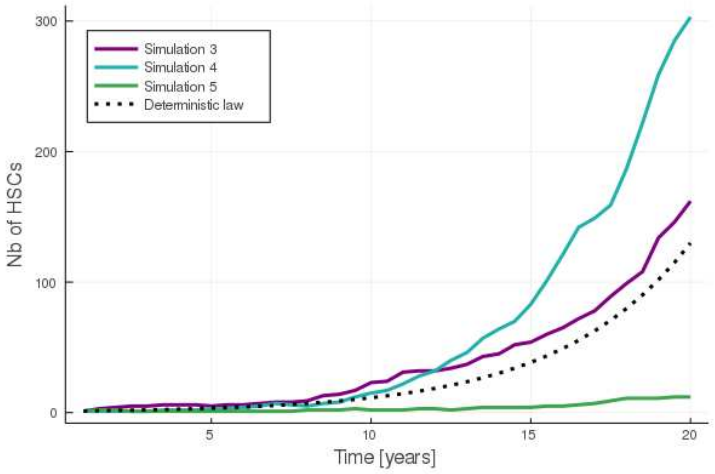


FIGURE 9 – Comparaison entre 3 simulations et l'approximation déterministe lorsque le processus commence avec une seule HSC mutée ($N_c = 1$). Les valeurs des paramètres utilisés pour les simulations sont $\alpha = 1/30$, $\Delta = 0.02$ et $p_2 = 0.03$.

Lorsque le nombre de cellules devient élevé, le formalisme stochastique reste valide, mais sa simulation devient plus difficile. Notre modèle est décrit par une CTMC. Dans le framework d'estimation ABC-SMC utilisé, une trajectoire est simulée en calculant successivement le temps auquel se produit la prochaine division d'une HSC mutée (ce que nous appelons une itération). Plus il y a de cellules mutées, plus le temps entre deux divisions consécutives est court. Ainsi, lorsque la taille du clone muté devient trop importante, il faut un très grand nombre d'itérations pour que la simulation progresse. Cette difficulté numérique est illustrée sur la figure 10. Nous avons simulé deux trajectoires, en fixant pour les valeurs des paramètres $\alpha = 1/30$, $\Delta = 0.02$ et $p_2 = 0.03$, jusqu'à ce qu'elles atteignent le nombre de $5 \cdot 10^5$ HSC mutées. Sachant que nous considérons l'expansion d'un clone malin dans un *pool* de cellules souches de 10^5 HSC de type sauvage (WT), $5 \cdot 10^5$ HSC mutées serait en accord avec les observations pour les patients $CALR^m$ où de nombreuses mesures des fractions clonales parmi les cellules progénitrices dépassent 80%. La dynamique de l'expansion des clones malins pour ces deux trajectoires est illustrée sur la gauche de la figure 10. Au cours des 30 premières années, l'expansion du clone muté est à peine perceptible, puis le nombre de cellules augmente rapidement en un court laps de temps.

Sur la droite de la figure 10, nous représentons le nombre d'itérations nécessaires pour simuler ces deux trajectoires. Un nombre raisonnable d'itérations est nécessaire pour simuler l'expansion dans les premières décennies, lorsque le nombre de HSC mutées n'est pas trop élevé, mais il devient ensuite presque impossible de simuler le processus jusqu'à 58 ans, qui est l'âge médian sur notre cohorte de patients $CALR^m$ (Tab. 2).
Il serait donc impossible de baser la calibration de notre modèle (qui nécessite la simulation de nombreuses trajectoires) uniquement sur ce processus stochastique.
Comme on peut l'observer sur la figure 10 (gauche), lorsqu'on considère un grand nombre de cellules, les effets stochastiques ne sont pas visibles et l'expansion clonale semble exponentielle. En effet, lorsqu'un nombre élevé de cellules N_c est atteint au temps t_c, alors la dynamique de l'expansion clonale au temps $t > t_c$ sera très proche de la moyenne conditionnelle du processus stochastique :

$$\mathbb{E}[N(t)|N(t_c) = N_c] = N_c \exp\left(\alpha\Delta(t - t_c)\right) \tag{5}$$

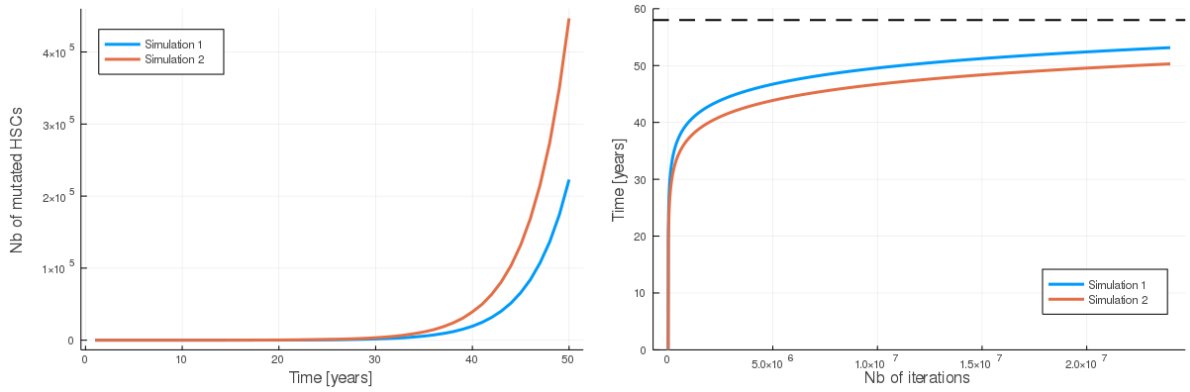


FIGURE 10 – Simulation de deux trajectoires avec $\alpha = 1/30$, $\Delta = 0.02$ et $p_2 = 0.03$. À gauche, augmentation du nombre de HSC mutées au cours du temps. À droite, nombre d'itérations nécessaires pour que les trajectoires progressent dans le temps. La ligne horizontale en tirets noirs représente l'âge médian des patients $CALR^m$.

Bien entendu, cette approximation déterministe ne peut pas remplacer notre modèle stochastique, comme l'illustre la figure 9 et n'est en fait pas valable lorsque N_c est trop faible.

Notre objectif est de déterminer une valeur de N_c telle que la loi déterministe précédente (5) puisse être utilisée sans faire une erreur d'approximation trop importante et qui ne dépende pas des valeurs des paramètres choisis. Un tel critère est difficile à obtenir théoriquement. Certains auteurs ont exploré la question [30], mais, en pratique, le choix d'une valeur pertinente pour N_c est un compromis entre erreur d'approximation et gain de temps de calcul. C'est pourquoi nous choisissons de la déterminer empiriquement.

Tout d'abord, nous considérons les mêmes valeurs de paramètres que celles utilisées précédemment. Nous simulons plusieurs trajectoires en commençant par N_c cellules mutées. Nous testons 20 valeurs différentes pour N_c :

$$N_c \in \{1, 2, 5, 10, 20, 30, \cdots, 3\ 000, 4\ 000, 5\ 000\}$$

et pour chacune d'entre elles, nous calculons 100 simulations jusqu'à ce que $N_{up} = 5 \cdot 10^5$ cellules soient atteintes (en recommençant les simulations en cas d'extinction). Chaque trajectoire représentera l'expansion clonale de N_c à N_{up} HSC mutées sur une période de plusieurs années. Le temps nécessaire pour atteindre cette limite supérieure N_{up} est une variable aléatoire. Il peut être comparé au temps qu'il faudrait pour atteindre ce niveau avec l'approche déterministe :

$$T_{det} = \frac{1}{\alpha\Delta} \log\left(\frac{N_{up}}{N_c}\right) \tag{6}$$

Si N_c est choisi suffisamment grand, alors l'erreur (par rapport à la valeur déterministe) sera faible. Sur la figure 11 (gauche), nous montrons que cette erreur est décroissante avec N_c. Pour $N_c = 2\ 000$, l'erreur relative moyenne passe en dessous de 0.005 et ne diminue ensuite que lentement avec des valeurs plus élevées de N_c. Par conséquent, $N_c = 2\ 000$ semble un critère valable : au-delà de ce nombre de cellules, la loi déterministe est une bonne approximation de notre processus pour les valeurs des paramètres que nous avons choisies.
Mais les valeurs des paramètres ne sont pas connues à l'avance puisque notre objectif est justement de les estimer. Nous devons donc vérifier si ce critère reste valable pour d'autres valeurs de paramètres. Pour cela, nous échantillonnons les valeurs des paramètres selon notre distribution *a priori* et simulons des trajectoires de $N_c = 2\ 000$ à $N_{up} = 5 \cdot 10^5$ cellules mutées et évaluons l'erreur relative comme précédemment. Comme présenté sur la droite de la figure 11, ce critère reste valable pour différentes valeurs de paramètres.

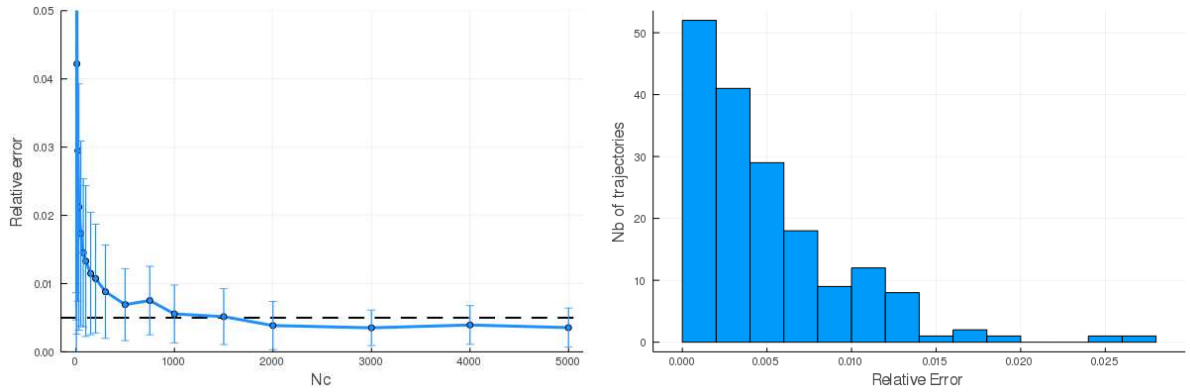


FIGURE 11 – Erreur relative entre l'approximation déterministe et les réalisations de notre processus stochastique. Les erreurs sont calculées sur la base du temps nécessaire pour atteindre $N_{up} = 5 \cdot 10^5$ HSC mutées. À gauche, l'erreur est représentée en fonction du nombre initial de HSC mutées N_c pour un jeu de paramètres donné. Les barres d'erreur représentent les écarts types. La ligne horizontale représente une erreur relative de 0.5%. À droite, N_c est fixé à 2,000 et nous représentons la distribution des erreurs relatives pour 175 trajectoires dont les paramètres sont échantillonnés aléatoirement à partir de notre distribution *a priori*.

Ce choix est conservatif et pourrait être sous-optimal, mais il est suffisant pour permettre une simulation suffisamment efficace des trajectoires pour que nous puissions exécuter notre procédure ABC-SMC en un délai raisonnable.

4.3 Algorithme hongrois

Dans la procédure ABC-SMC décrite au § 4.1, une simulation consiste à échantillonner un vecteur de paramètres et à calculer autant de trajectoires (sans extinction) que d'observations. Ensuite, nous assignons une trajectoire à une et une seule observation et calculons l'erreur quadratique. Notre distance est alors l'erreur quadratique moyenne. Si cette erreur est inférieure au seuil de tolérance, le vecteur de paramètres est rejeté.
Si N_p est notre nombre d'observations (et donc de trajectoires), il y aurait $N_p!$ de possibilités différentes pour affecter une trajectoire à une observation. Le choix aléatoire de l'affectation ne serait pas un choix pertinent. Bien que théoriquement convergent, cela déprécierait fortement la vitesse de convergence de la procédure puisqu'il en résulterait un taux élevé de rejets. Pour contourner cette difficulté, nous choisissons l'affectation optimale telle que l'erreur quadratique moyenne finale (calculée sur les N_p observations) soit la plus faible. Puisque le calcul de toutes les $N_p!$ possibilités ne serait pas faisable, nous utilisons l'algorithme hongrois d'affectation optimale [17, 18] dont la complexité temporelle est de $O\left(n^3\right)$.

4.4 distribution zero-inflated pour λ

L'âge auquel la mutation du gène *JAK2* apparaît est sujet à controverse, et il y a un manque d'étude concernant la mutation du gène *CALR*. Ce manque de connaissances justifie l'utilisation de modèles mathématiques pour tester les deux hypothèses : soit la mutation a été acquise au cours de la vie fœtale, soit elle a été acquise après la naissance. Dans notre modèle, les deux scénarios ne diffèrent que par la façon dont T_0 est modélisé, soit constant et égal à zéro pour tous les patients dans le premier cas, soit une variable aléatoire suivant une loi exponentielle avec le paramètre supplémentaire $\lambda > 0$ dans le second cas (voir § 3.3).
Le scénario d'acquisition lors de la vie fœtale peut donc être considéré comme le cas limite d'une acquisition suivant une loi exponentielle $\mathcal{E}(\lambda^{-1})$ lorsque $\lambda = 0$. Confronter les deux hypothèses équivaudrait à évaluer lorsque $\lambda = 0$ *vs* $\lambda > 0$. Cela implique de permettre à λ d'être égal à zéro avec une probabilité non nulle, c'est-à-dire de choisir une loi *zero-inflated* comme distribution *a priori*. En pratique, ceci est mis en œuvre en utilisant un paramètre binaire $\beta \in \{0, 1\}$, avec des

probabilités *a priori* $\mathbb{P}[\beta = 0] = \mathbb{P}[\beta = 1] = 0.5$, et un paramètre l étant *a priori* uniformément distribué sur $[0, 100]$ (années). Dans notre procédure ABC-SMC, lorsque le paramètre β est égal à zéro, le temps d'acquisition de la mutation T_0 est fixé à zéro pour tous les patients. En revanche, si $\beta = 1$, nous échantillonnons aléatoirement autant de temps d'acquisition qu'il y a d'individus dans la cohorte, en les échantillonnant selon une loi exponentielle $\mathcal{E}(l^{-1})$ de paramètre $l > 0$. Enfin, la moyenne *a posteriori* de λ est obtenue à partir de $\lambda = \beta\ l$.
Cette procédure équivaut à considérer deux modèles distincts : si $\beta = 0$, nous utilisons le modèle d'acquisition des mutations au cours de la vie fœtale, tandis que si $\beta = 1$, nous utilisons le modèle d'apparition au cours de la vie, procédure correspondant à un mélange de modèles en sélection de modèles Bayésienne.

4.5 Détermination d'un âge optimal de dépistage

Après avoir estimé les distributions *a posteriori* des paramètres, nous pouvons inférer la dynamique de l'expansion clonale et ainsi déduire les stratégies optimales en vue de réaliser un dépistage précoce (voir figure 12).
L'objectif est de détecter la mutation d'intérêt chez un individu le plus tôt possible, c'est-à-dire avant l'apparition des symptômes de la maladie. Le dépistage précoce nécessite de collecter des échantillons de sang des individus et de réaliser une analyse du gène d'intérêt pour mesurer la VAF dans le sang périphérique. Les techniques habituelles en pratique clinique permettent de détecter les mutations $CALR^m$ à une VAF supérieure à 2% avec la technique CALR *sizing* et à environ 0.1% pour $JAK2^{V617F}$ (0.01% à 1% par *allele-specific PCR*). Nous considérons que la détection est trop tardive si la CF dans les HSC est supérieure à 50% pour $CALR^m$ [19] et supérieure à 15% pour $JAK2^{V617F}$ [31], ce qui correspond aux seuils à partir desquels il existe un risque élevé de développement de NMP et de thrombose potentielle. Notre approche s'appuie sur la distribution *a posteriori* des paramètres du modèle pour calculer la probabilité de détecter la mutation à différents âges et peut donc aider à déterminer le meilleur *timing* pour le dépistage.

Nous sommes confrontés à un problème d'optimisation. Si le dépistage est trop précoce, il y aura un risque élevé de faux négatifs : une forte proportion de personnes développeraient la maladie sans être identifiées. D'un autre côté, si le dépistage est trop tardif, les personnes pourraient être détectées alors que leur *pool* de cellules souches est déjà envahi. Avec le modèle proposé, nous pouvons résoudre ce problème d'optimisation numériquement puisque nous déduisons la dynamique de l'initiation et de l'expansion du clone muté. Nous considérons que la détection est trop tardive lorsque la CF parmi les HSC est supérieure à un seuil CF_{late} :

$$\frac{N(t)}{N(t) + N_{WT}} > CF_{\text{late}} \tag{7}$$

avec $CF_{\text{late}} = 50\%$ pour $CALR^m$ [19] et égale à 15% pour $JAK2^{V617F}$ [31]. En routine clinique, seule la VAF dans le sang périphérique est mesurée. Cependant, la mesure de VAF n'est pas un bon indicateur de la CF parmi les progéniteurs, encore moins des cellules souches. Dans notre modèle, $VAF(t)$ au temps t est donnée par l'équation (3). Nous considérons qu'une mutation existante n'est pas détectée (faux négatif) lorsque $VAF(t) < VAF_{detection}$ où $VAF_{detection} = 0.1\%$ pour $JAK2^{V617F}$ et $VAF_{detection} = 2\%$ pour $CALR^m$. Enfin, le problème d'optimisation à résoudre est celui de trouver l'âge auquel la probabilité de dépistage précoce dans la population est la plus élevée possible :

$$T^* = \max_t \mathbb{P}\left[\text{VAF}(t) > \text{VAF}_{\text{detection}}\ ;\ \frac{N(t)}{N(t) + N_{WT}} < CF_{\text{late}}\ \mid \mathcal{D}\right] \tag{8}$$

Nous résolvons ce problème numériquement pour $JAK2^{V617F}$ et $CALR^m$ séparément, en évaluant la probabilité de détection précoce à l'âge t en échantillonnant les vecteurs des paramètres à partir de leur *posterior*. Nous simulons ainsi 20,000 trajectoires et estimons (Fig. 12.iii) la probabilité

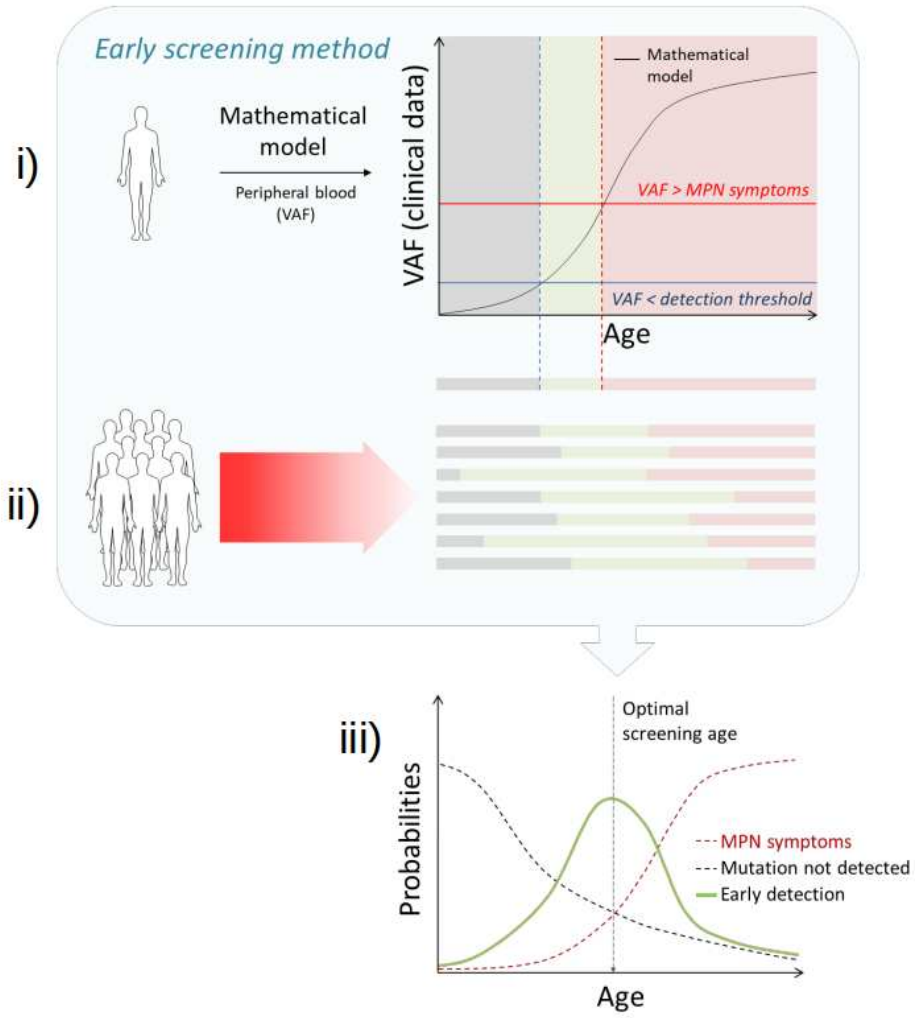


FIGURE 12 – Aperçu de la méthode permettant de déterminer l'âge optimal de dépistage.
i) Une trajectoire du modèle stochastique (après sa calibration à partir d'observations réelles) correspond à l'expansion de la maladie chez un patient quelconque (échantillonné au hasard). À partir du modèle, on obtient la progression de la VAF dans les cellules matures au cours des années. Trois périodes de la vie du patient peuvent être considérées. Tout d'abord, la VAF est inférieure à un seuil de détection (zone grise) ; le clone muté a déjà commencé à se développer mais est encore indétectable. Ensuite (zone verte), la mutation devient détectable encore avec une CF suffisamment faible pour qu'il y ait un faible risque de symptômes de NMP. Théoriquement, ce serait la période appropriée pour le dépistage précoce pour ce patient particulier. Finalement, le clone malin continue à s'étendre ; la VAF dépasse un seuil au-delà duquel il y a un risque de symptômes de NMP (zone rouge) : le dépistage serait trop tardif.
ii) En raison de l'hétérogénéité entre patients, les trois périodes définies précédemment (mutation non détectée - gris, détection précoce - vert, et symptômes des NMP - rouge) sont différentes selon les individus.
iii) En considérant un nombre élevé de patients, nous obtenons en fonction de l'âge les fréquences (ou probabilités) de dépistage précoce (ligne verte), de ne pas détecter la mutation (ligne grise en pointillé), ou d'avoir des symptômes (ligne rouge en pointillé). La probabilité de dépistage précoce atteint un maximum à l'âge de dépistage optimal, c'est-à-dire l'âge auquel il serait optimal de tester la population pour la mutation considérée. La valeur atteinte à l'âge de dépistage optimal correspond à la plus grande proportion de patients qui seraient détectés suffisamment tôt (selon notre modèle mathématique et sa calibration). Plus cette valeur est élevée, plus le dépistage est efficace.

de dépistage précoce (early detection) ainsi que le taux de faux négatifs (Mutation not detected) et la probabilité d'une détection trop tardive (MPN symptoms), ce qui nous permet de déduire l'âge optimal de dépistage.

5 Robustesse de la méthode

5.1 Inférence sur données simulées

Lorsque l'on conçoit des modèles mathématiques pour décrire des phénomènes biologiques et que nous voulons les calibrer à partir de données expérimentales, une étape importante consiste à valider d'abord l'identifiabilité pratique du modèle [32, 33]. Cela peut être fait en utilisant des ensembles de données synthétiques générés à partir du modèle mathématique, et donc pour lesquels la *ground truth* (la valeur réelle des paramètres) est connue. En particulier, une question importante est de savoir dans quelle mesure le nombre d'observations disponibles est suffisant pour obtenir des conclusions robustes.
Dans ce chapitre, nous étudions deux hypothèses (ou modèles) différents - $\lambda = 0$ *vs* $\lambda > 0$ - que nous voulons discriminer à partir des observations sur un nombre limité de patients. Nous cherchons également à interpréter les paramètres estimés car ils sont porteurs d'informations biologiques pertinentes.

N_i	Ground truth				Paramètres estimés		
	Δ	q	λ	$\bar{T}_0$	$\hat{\Delta}$	$\hat{q}$	$\hat{\lambda}$
15	0.030	0.850	20.0	19.2	0.027 [0.018, 0.037]	0.904 [0.781, 0.956]	17.6 [10.3, 27.3]
15	0.023	0.850	0.0	0.0	0.020 [0.016, 0.024]	0.924 [0.815, 0.963]	0.2 [0, 3.8]
14	0.139	0.554	0.0	0.0	0.146 [0.112, 0.184]	0.604 [0.285, 0.767]	0.0 [0, 0]
11	0.092	0.817	55.9	35.5	0.098 [0.056, 0.166]	0.658 [0.223, 0.861]	32.5 [17.4, 53]
15	0.119	0.336	38.2	19.4	0.127 [0.08, 0.193]	0.612 [0.244, 0.817]	21.3 [12.1, 33]
12	0.021	0.939	37.7	21.8	0.021 [0.014, 0.028]	0.912 [0.784, 0.966]	25.1 [13.4, 41.1]
14	0.197	0.546	96.6	40.1	0.058 [0.036, 0.085]	0.786 [0.488, 0.913]	41.3 [23.3, 65.1]
12	0.366	0.330	0.0	0.0	0.368 [0.284, 0.462]	0.285 [0.045, 0.481]	0.0 [0, 0]
15	0.134	0.664	0.0	0.0	0.142 [0.11, 0.179]	0.618 [0.328, 0.767]	0.0 [0, 0]
14	0.027	0.945	36.1	16.4	0.032 [0.022, 0.045]	0.869 [0.672, 0.947]	20.2 [11.5, 31.4]
12	0.251	0.427	0.0	0.0	0.247 [0.203, 0.294]	0.382 [0.068, 0.61]	0.0 [0, 0]

TABLE 4 – Résultats de la procédure d'estimation des paramètres en utilisant des données synthétiques. N_i observations sont générées à partir du modèle avec des valeurs de paramètres connues (*ground truth*). Les paramètres du modèle sont ensuite estimés à l'aide de la procédure ABC-SMC (paramètres estimés). Nous indiquons les moyennes *a posteriori* pour les paramètres Δ, q et λ et les intervalles de crédibilité à 95%. Par souci de clarté, nous omettons le paramètre α car il ne s'écarte pas de son *prior*. Pour les valeurs *ground truth*, nous indiquons également $\bar{T}_0$, qui est le temps d'acquisition moyen "observé" sur les N_i trajectoires simulées.

Nous générons différents jeux de données. Pour chacun d'eux (indice i), nous échantillonnons un vecteur de paramètres $\boldsymbol{\theta}_i$ (*ground truth*) et simulons N_i trajectoires à partir de notre modèle. N_i est choisi aléatoirement entre 11 (nombre de patients $JAK2^{V617F}$ que nous avons dans notre jeu de données réel) et 15 (nombre de patients $CALR^m$). Pour chaque trajectoire (indexée par $j \in \{1, \cdots, N_i\}$), nous échantillonnons une valeur y_j entre 0 et 1. Notre temps d'observation et la fraction clonale observée $(\hat{t}_j, \hat{\eta}_j)$ sont déterminés de telle sorte que $\eta(\hat{t}_j) = y_j = \hat{\eta}_j$ (avec η donnée par la relation (2)). Nous ne considérons que les trajectoires sans extinction et avec un temps d'observation $\hat{t}_j \in [0, 100]$ ans. Si l'une de ces deux conditions n'est pas respectée, nous

simulons une nouvelle trajectoire jusqu'à ce que les deux contraintes soient satisfaites. Nous définissons $\bar{T}_{0,i} = \frac{1}{N_i} \sum_{j=1}^{N_i} T_{0,i,j}$ le temps d'acquisition moyen "observé" sur les N_i trajectoires simulées. Ensuite, nous exécutons notre procédure d'estimation jusqu'à convergence et comparons nos paramètres estimés (moyennes *a posteriori* et intervalles de crédibilité à 95%) avec les valeurs réelles utilisées pour générer les données. Les résultats de cette étude sont présentés dans le tableau 4.
Les paramètres Δ et q sont globalement bien estimés. Nous avons également retrouvé correctement si les données étaient générées selon l'hypothèse d'une acquisition de la mutation au cours de la vie fœtale ($\lambda = 0$, c'est-à-dire $T_0 = 0$) ou celle d'une apparition au cours de la vie ($\lambda > 0$, c'est-à-dire $T_0 > 0$). Néanmoins, dans le cas d'une acquisition après la naissance, λ n'est pas très bien estimé. Mais si l'on considère, au lieu de λ, le temps d'acquisition moyen "observé" $\bar{T}_0$ (rappelons que $T_0 \sim \mathcal{E}(\lambda^{-1})$ tel que $\mathbb{E}[T_0] = \lambda$) calculé sur les N_i trajectoires simulées, on trouve que $\hat{\lambda}$ estime avec précision $\bar{T}_0$. Dans notre jeu de données simulées, la différence entre $\bar{T}_0$ et λ s'explique par le fait que nous simulons des temps d'observation censurés[2]. En effet, pour avoir un jeu de données synthétiques cohérent avec les observations réelles, nous imposons que le temps d'observation soit inférieur à 100 ans. Ainsi, pour des grandes valeurs de λ, les trajectoires avec des temps d'acquisition tardifs sont automatiquement exclues de notre jeu de données, donc $\bar{T}_0 < \lambda$. Par conséquent, notre estimation de λ peut être potentiellement biaisée. Nous explorons ce point plus en détail dans le paragraphe suivant.

5.2 Biais dans l'estimation de λ

Nos observations sont censurées : les individus qui acquerraient la mutation motrice à un âge trop avancé n'apparaîtraient pas dans notre cohorte puisqu'ils mourraient avant de présenter les symptômes de NMP et avant d'avoir une CF trop élevée. Dans notre modèle, λ est le temps moyen d'acquisition de la mutation, estimé à partir des observations censurées. Ainsi, au lieu d'estimer réellement λ, nous estimons le temps d'acquisition moyen "observé" $\bar{T}_0$, comme indiqué dans le paragraphe précédent. Intuitivement, si la valeur réelle de λ est faible, elle peut toujours être estimée avec précision puisqu'elle correspondra effectivement au temps d'acquisition moyen "observé". S'il est trop élevé, le temps d'acquisition moyen "observé" sur la population $\bar{T}_0$ sera tel que $\bar{T}_0 < \lambda$ et le paramètre λ sera sous-estimé. Nous quantifions cet effet de la censure sur l'estimation par une étude basée sur des simulations.
Pour $\lambda \in \{1, 2, \cdots, 100\}$, nous répétons pour $i \in \{1, \cdots, 1000\}$ la procédure suivante :

- Nous échantillonnons un vecteur de paramètres $\boldsymbol{\theta_i}$ à partir de notre *prior* (sauf λ qui est assigné à une valeur donnée)
- Nous simulons 15 trajectoires (sans extinction), $j \in \{1, 2, \cdots, 15\}$
- Nous gardons en mémoire les 15 temps d'acquisition $T_{0,i,j}$
- Pour chaque trajectoire j, nous échantillonnons une valeur uniformément comprise entre 0 et 1 pour la CF $\eta_{i,j}$, et calculons le temps d'observation associé $t_{i,j}$
- Nous censurons les individus (trajectoires) dont le temps d'observation $t_{i,j}$ dépasse 100 ans. On obtient $N_i \leq 15$ trajectoires non censurées
- Nous calculons le temps d'acquisition moyen "observé" $\bar{T}_{0,i}$ (calculé sur les N_i trajectoires non censurées)

Nous calculons ensuite le nombre moyen (sur les 1,000 itérations) d'individus (trajectoires) censurés sur les 15 trajectoires simulées pour chaque $\lambda \in \{1, 2, \cdots, 100\}$ (Fig. 13). Nous observons que, pour $\lambda \geq 25$, l'effet de censure commence. Plus le vrai λ est élevé, plus le nombre d'individus censurés est important.

Nous calculons également, pour chaque $\lambda \in \{1, 2, \cdots, 100\}$, la moyenne (sur les 1000 itérations) du temps d'acquisition "observé" $\bar{T}_0$ (calculé sur les individus non censurés) et les quantiles à 5 et 95% (Fig. 14). Pour $\lambda < 20$ ans, il n'y a pas beaucoup de différence entre λ et $\bar{T}_0$. Ensuite,

2. Sans censure, nous aurions $\lim_{N_i \longrightarrow \infty} \bar{T}_0 = \lambda$

nous observons progressivement l'effet de la censure des individus (trajectoires).
Avec notre procédure d'estimation des paramètres, nous estimons précisément $\bar{T}_0$ plutôt que le vrai λ (comme montré dans la section précédente). Dans le cas de $JAK2^{V617F}$, où nous estimons des valeurs assez faibles pour λ (avec une valeur moyenne d'environ 15 années, voir § 6), nous sommes dans la plage de valeurs où nous devrions l'estimer avec précision. Pour le cas $CALR^m$, pour lequel nous estimons une valeur moyenne de λ d'environ 25 (voir § 6), nous avons probablement sous-estimé le véritable temps d'acquisition. Avec notre modèle simple de censure, l'estimation d'un temps d'acquisition de 25 ans (qui correspondrait donc à $\bar{T}_0$) pourrait correspondre approximativement à un temps d'acquisition réel d'environ 28 ans (correspondant à la vraie valeur de λ).
Cette étude présente deux limites principales : premièrement, nous considérons un âge de censure de 100 ans, alors qu'une analyse plus précise tiendrait compte de valeurs démographiques plus réalistes. Deuxièmement, la CF est échantillonnée uniformément sur $[0, 1]$, alors qu'une étude plus approfondie rendrait compte d'une distribution de ces valeurs de CF plus conforme à celles trouvées dans la réalité. Néanmoins, même avec ce modèle simpliste de censure, nous pouvons trouver certains résultats qualitatifs, à savoir que nous sous-estimons probablement légèrement le temps d'acquisition réel de la mutation $CALR^m$ pour la population générale. Ainsi, notre estimation du paramètre λ doit être considérée comme le temps d'acquisition moyen estimé de la mutation pour les individus qui développeront une NMP, et ce sont en fait les individus sur lesquels nous nous concentrons dans ce chapitre.

Nous avons également étudié la robustesse dans l'estimation du paramètre Δ. Cette étude est présentée en annexe A.1 à ce chapitre, au lien suivant :

`https://gitlab-research.centralesupelec.fr/2012hermangeg/supplementary-material-phd`

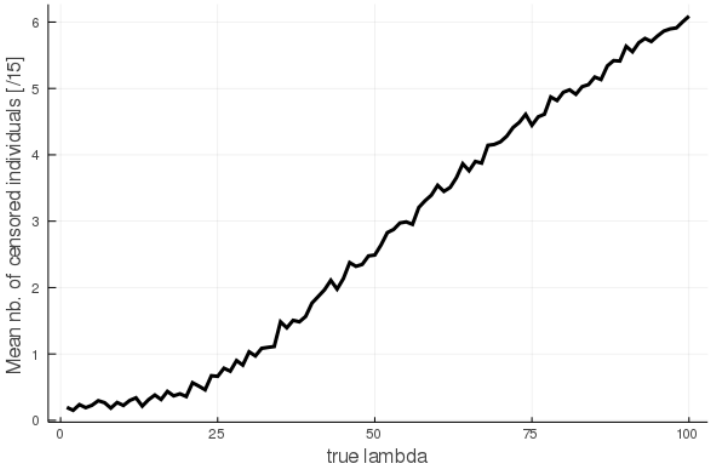


FIGURE 13 – Nombre d'individus censurés (temps d'observation supérieur à 100 ans) pour une cohorte simulée de 15 individus, pour différentes valeurs de λ (valeur moyenne calculée sur 1,000 simulations pour chaque valeur de λ).

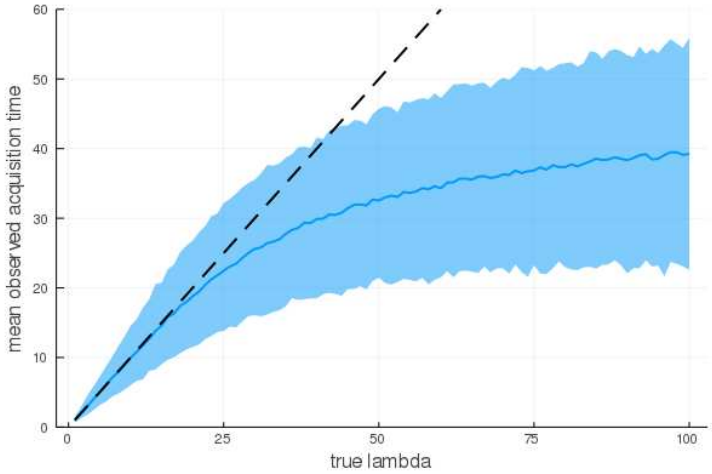


FIGURE 14 – Temps d'acquisition moyen "observé" $\bar{T}_0$ calculé pour les individus non censurés (axe des ordonnées) par rapport au temps d'acquisition moyen réel λ (axe des abscisses) de la mutation. La ligne bleue indique la valeur moyenne (pour chaque vrai λ, 1000×15 trajectoires sont simulées) et la zone ombrée représente un intervalle de confiance de 90%. La ligne noire en pointillés matérialise $\bar{T}_0 = \lambda$.

6 Résultats

6.1 Les mutations $CALR^m$ seraient acquises plus tard que celles $JAK2^{V617F}$

Les résultats sont obtenus en exécutant notre procédure d'estimation pour les populations de patients $JAK2^{V617F}$ et $CALR^m$, séparément.
Nous exécutons notre procédure ABC-SMC jusqu'à convergence (voir § 4.1) et obtenons les distributions *a posteriori* des paramètres pour la population de patients $JAK2^{V617F}$ et la population $CALR^m$ (voir annexe A.2).

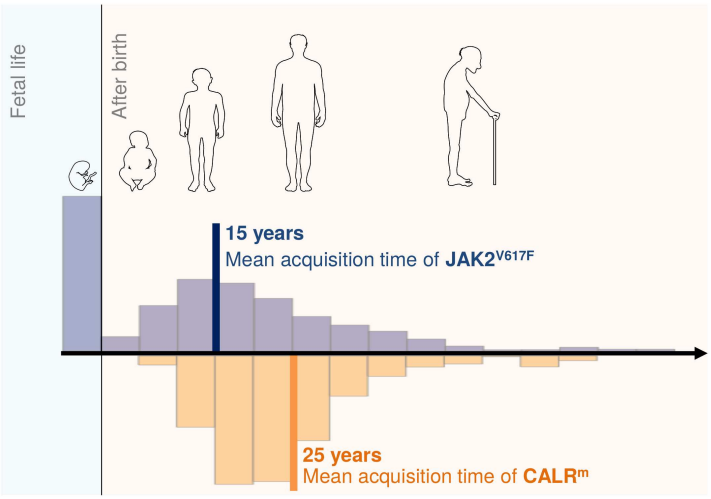


FIGURE 15 – La mutation $JAK2^{V617F}$ s'avère être acquise plus tôt en moyenne que la mutation $CALR^m$ (15 *vs* 25 ans), et potentiellement pendant la vie fœtale (avec une probabilité $\mathbb{P}[\lambda = 0|\mathcal{D}] = 0.24$ quand la même probabilité est estimée à zéro pour $CALR^m$). Les histogrammes représentent la distribution du paramètre λ.

Le temps moyen d'acquisition de la mutation $JAK2^{V617F}$ $\mathbb{E}[\lambda]$ est estimé à ~ 15 ans (Fig. 15). Par ailleurs, concernant l'hypothèse selon laquelle la mutation se produit pendant la vie fœtale pour la population de patients $JAK2^{V617F}$, nous estimons sa probabilité $\mathbb{P}[\lambda = 0|\mathcal{D}]$ égale à 0.24. Ces résultats sont cohérents avec les valeurs rapportées par Van Egeren et al. [6] et Williams et al. [7] (voir § 7.2.1).
Il est intéressant de noter que, pour la population de patients $CALR^m$, nous estimons que presque sûrement $\lambda > 0$, c'est-à-dire que l'hypothèse d'une acquisition de la mutation pendant la vie fœtale pour tous les patients $CALR^m$ est peu probable. Nous estimons également un temps d'acquisition moyen attendu plus élevé ($\mathbb{E}[\lambda] \sim 25$ ans) que pour $JAK2^{V617F}$.

6.2 Un avantage prolifératif plus important pour $CALR^m$

Dans le modèle proposé, l'avantage prolifératif des cellules mutées est décrit grâce au paramètre $\Delta = p_2 - p_0$, qui représente la balance entre les divisions symétriques et différenciées entraînant l'expansion clonale de la mutation lorsque $\Delta > 0$. Plus sa valeur est élevée, plus le clone muté se développe rapidement. Nous estimons Δ pour les clones $CALR^m$ et $JAK2^{V617F}$ séparément et en inférons que l'avantage prolifératif du clone malin $CALR^m$ au niveau des cellules souches est plus élevé que celui de $JAK2^{V617F}$ (Fig. 16.B), avec des valeurs moyennes de Δ respectivement égales à 0.026 et 0.017.
La propension des HSC $CALR^m$ à faire des divisions symétriques et donc à envahir le *pool* de cellules souches est plus élevée que celle des HSC $JAK2^{V617F}$ (Fig. 17). Cette observation explique que, même si la mutation $CALR^m$ est acquise plus tard dans la vie, le clone muté peut généralement atteindre une CF élevée plus rapidement. Notre estimation de Δ pour $JAK2^{V617F}$ est plus grande que celle trouvée par Watson et al. [8], puisque nous estimons Δ à partir d'ob-

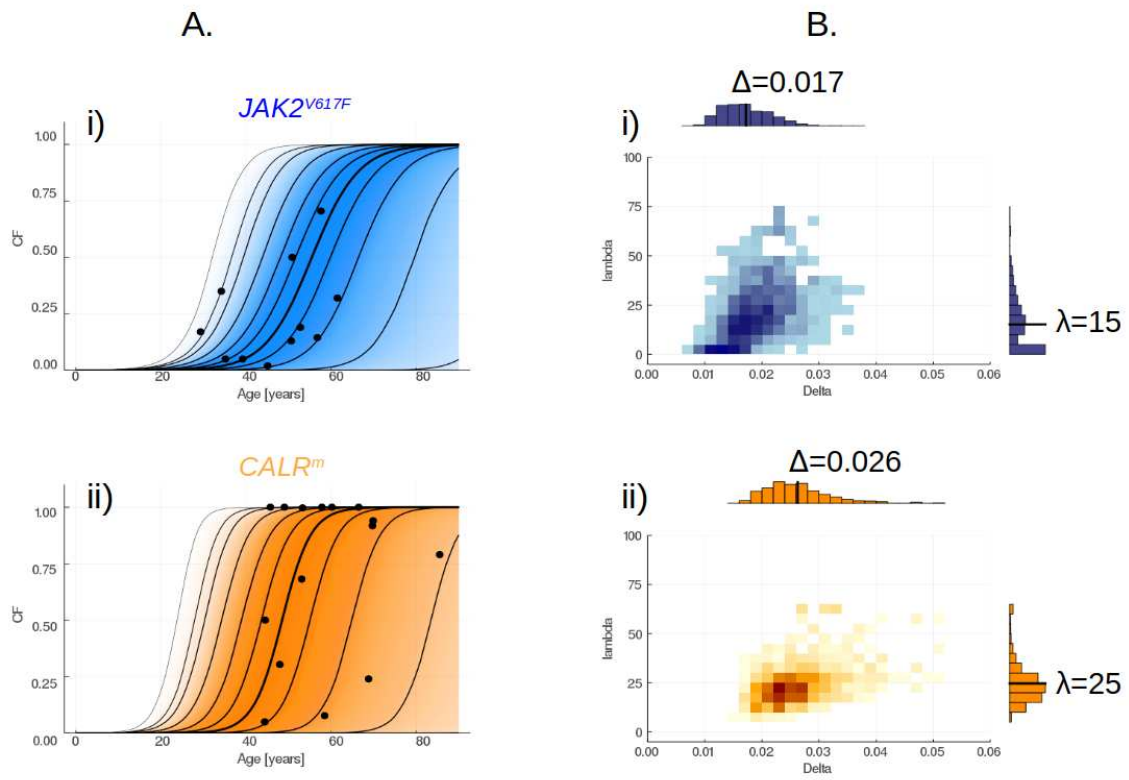


FIGURE 16 – A) Confrontation des observations des patients $JAK2^{V617F}$ (i) ou $CALR^m$ (ii) (points noirs) à l'évolution inférée de la CF parmi les cellules progénitrices. La dynamique de l'expansion clonale dans le temps est obtenue à partir de 10,000 simulations du modèle, en échantillonnant les paramètres à partir des *posteriors*. La ligne en gras représente la CF médiane, calculée sur 10,000 trajectoires simulées à chaque âge, tandis que les lignes de part et d'autre matérialisent des intervalles de crédibilité de 25, 50, 75, 90, 95 et 99%. Le gradient de couleur indique où les trajectoires peuvent être trouvées ; plus il est foncé, plus la probabilité est élevée.
B) Distributions *a posteriori* jointes estimées des paramètres Δ et λ, pour $JAK2^{V617F}$ (i) ou $CALR^m$ (ii). Les lignes pleines noires dans les histogrammes 1D correspondent aux valeurs moyennes. Les régions les plus sombres sur les histogrammes 2D correspondent à celles où la densité de probabilité est la plus élevée.

servations d'individus ayant un NMP alors que Watson et al. ont étudié l'hématopoïèse clonale de donneurs sains (voir § 7.2.2). En outre, nous estimons une forte probabilité d'extinction stochastique ($q = p_0/p_2$) pour les clones $JAK2^{V617F}$ et $CALR^m$, avec des valeurs moyennes de q respectivement égales à 0.94 et 0.87, ce qui signifie que l'acquisition de la mutation conduirait à une expansion clonale - et ensuite potentiellement à l'apparition de la maladie - dans seulement $\sim 10\%$ des cas (voir annexe A.2).

6.3 Le dépistage précoce - une option pour détecter les mutations $JAK2^{V617F}$

Après avoir estimé les valeurs moyennes des paramètres du modèle et leurs distributions de probabilité, nous pouvons déduire le développement des NMP et en déduire des stratégies de dépistage précoce.

L'âge optimal de dépistage se situe à 30 ans pour la mutation $JAK2^{V617F}$ (Fig. 18.A-i), et à 35 ans pour la mutation $CALR^m$ (Fig. 18.A-ii). À cet âge optimal de détection, il existe trois possibilités pour les individus porteurs de la mutation : la détection est trop tardive car ils souffraient déjà des symptômes de la maladie, nous parvenons à détecter la mutation précocement (vrai-positif),

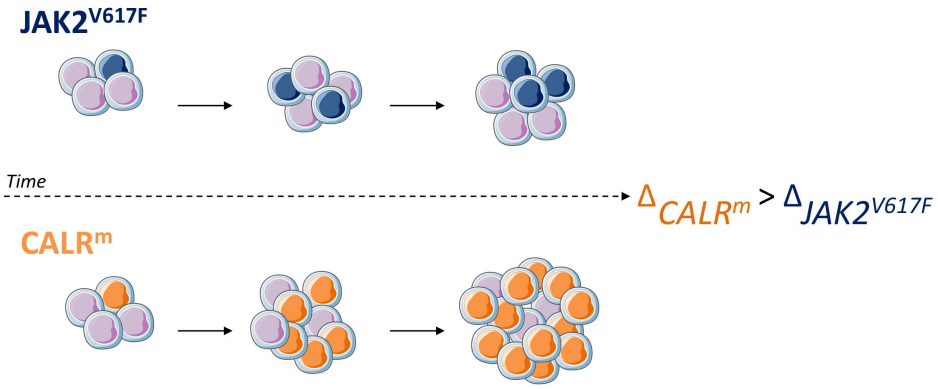


FIGURE 17 – La mutation $CALR^m$ s'avère avoir un avantage prolifératif plus important au niveau des cellules souches que $JAK2^{V617F}$. Le clone $CALR^m$ (en orange) se développe au fil du temps - dans un *pool* de HSC WT (cellules violettes) - à un rythme plus élevé que $JAK2^{V617F}$ (en bleu).

ou ne la détectons pas alors que l'individu développera un NMP (faux-négatif). Pour les patients $CALR^m$, la probabilité de détection précoce - calculée à l'âge optimal de dépistage - reste faible (42%), ce qui laisse 46% des individus qui développeraient plus tard la maladie non détectés, et 12% pour lesquels la détection serait trop tardive. En revanche, le dépistage précoce pourrait être une option clinique viable pour détecter les mutations du gène $JAK2^{V617F}$ - 79 % des individus étant détectés précocement à l'âge optimal de 30 ans.
De plus, nous étudions comment la sensibilité des techniques de dépistage influence les résultats précédents. Nous comparons différentes sensibilités avec des seuils de détection (correspondant à la VAF dans les cellules matures) allant de 0.01% à 2% (Fig. 18.B). Tant pour les clones malins $CALR^m$ que $JAK2^{V617F}$, des sensibilités plus élevées augmentent la probabilité de détecter la mutation à des âges plus faibles. Cependant, toujours à des niveaux ultra-sensibles (seuil de détection aussi bas que 0.01%), la probabilité de détection précoce à l'âge optimal de dépistage est inférieure à 65% pour $CALR^m$, ce qui est inférieur à la valeur obtenue pour $JAK2^{V617F}$ avec un seuil de VAF de 0.5% (Fig. 18.B-iii).

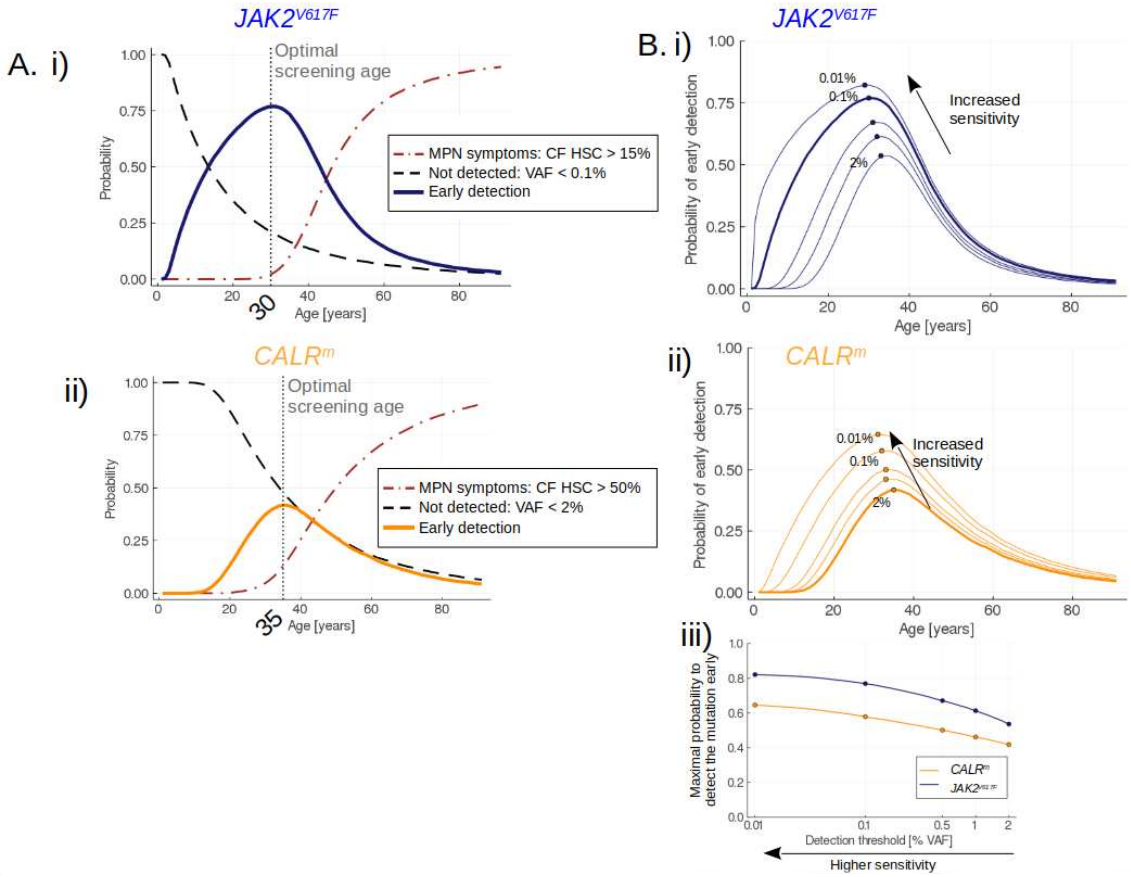


FIGURE 18 – A) Évolution de la probabilité d'une détection précoce (ligne continue, bleue pour $JAK2^{V617F}$ (i) et orange pour $CALR^m$ (ii)), du taux de faux-négatifs (ligne noire) et d'une détection trop tardive (ligne rouge) lorsqu'on teste la population pour la mutation $JAK2^{V617F}$ (i) ou $CALR^m$ (ii) à différents âges. On considère ici que la mutation $CALR^m$ est détectée dans les cellules matures dès lors que la VAF est supérieure à 2% (CALR *sizing*) et que $JAK2^{V617F}$ est détectée pour un VAF supérieure à 0.1% (*allele-specific* PCR).
B) Impact de la sensibilité des techniques pour la détection des mutations $JAK2^{V617F}$ (i) ou $CALR^m$ (ii). Comme prévu, des sensibilités plus élevées augmentent la probabilité de détecter les mutations, à des âges plus précoces. Sur la figure (iii), nous calculons la probabilité maximale de détecter la mutation (à l'âge de dépistage optimal correspondant) pour différents seuils de VAF allant de 2% à 0.01% (axe des abscisses en échelle logarithmique).

7 Validation

7.1 Analyse leave-one-out

Pour évaluer plus précisément la qualité des ajustements de notre modèle aux observations, nous procédons à une analyse *leave-one-out*. Pour chaque population de patients (soit $JAK2^{V617F}$ soit $CALR^m$) de taille N_p, nous effectuons N_p calibrations du modèle $i \in \{1, \cdots, N_p\}$ dans lesquelles l'individu i est retiré. Nous évaluons ensuite la capacité du modèle (calibré à partir des $N_p - 1$ observations) à généraliser ses résultats à un individu non inclus. En particulier, nous nous concentrons sur la capacité à prédire le moment où la fraction clonale d'un individu donné pourrait être atteinte. C'est l'information pertinente que nous devons estimer avec précision pour que notre méthodologie de dépistage précoce soit valide.

Plus précisément, pour une population de patients donnée, l'ensemble des données est noté $\mathcal{D}$. Pour un individu donné i, avec des observations $(\hat{t}_i, \hat{\eta}_i)$, l'ensemble de données sans ces observations est désigné par $\mathcal{D}_{-i}$. Le modèle est calibré sur la base de $\mathcal{D}_{-i}$. Nous estimons la distribution

a posteriori du vecteur de paramètres $\boldsymbol{\theta}$: $p[\boldsymbol{\theta}|\mathcal{D}_{-i}]$. En particulier, nous pouvons analyser comment $p[\Delta|\mathcal{D}_{-i}]$ s'écarte de $p[\Delta|\mathcal{D}]$ pour chaque patient i, et selon qu'il présente ou non d'autres mutations (Fig. 19). Cette analyse ne révèle pas que l'inclusion d'un patient présentant d'autres mutations (que celle qui nous intéresse) aurait un impact sur les estimations des paramètres du modèle.
Ensuite, pour la CF donnée $\hat{\eta}_i$, nous pouvons estimer $p[t_i|\mathcal{D}_{-i}, \hat{\eta}_i]$, où t_i est le moment auquel $\hat{\eta}_i$ devrait être atteint, c'est-à-dire $\eta(t_i) = \hat{\eta}_i$. Nous estimons numériquement la quantité précédente $p[t_i|\mathcal{D}_{-i}, \hat{\eta}_i]$ en simulant 2,000 trajectoires de notre modèle avec le vecteur de paramètres échantillonné à partir du *posterior* $p[\boldsymbol{\theta}|\mathcal{D}_{-i}]$. Nous confrontons $p[t_i|\mathcal{D}_{-i}, \hat{\eta}_i]$ au véritable temps d'observation $\hat{t}_i$ (Fig. 20).

Pour évaluer la précision de la prédiction, nous subdivisons la période $[0, 100]$ ans en vingt intervalles de 5 ans :

$$I_1 = [0, 5], I_2 = [5, 10], \cdots, I_{19} = [90, 95], I_{20} = [95, 100]$$

Nous désignons par $\hat{I}^i$ l'intervalle de 5 ans qui contient le temps d'observation : $\hat{t}_i \in \hat{I}^i$. Nous calculons :

$$\mathbb{P}[\hat{t}_i \in \hat{I}^i|\mathcal{D}_{-i}, \hat{\eta}_i] \tag{9}$$

Nous appelons la quantité précédente "score" et nous la comparons au score qui aurait été obtenu "par hasard" (c'est-à-dire avec un échantillonnage uniforme), dont la valeur est égale à $1/20 = 0.05$ (Fig. 21).
Plus le score est élevé, plus la prédiction est précise. Pour $JAK2^{V617F}$, seuls deux individus sur 11 (18%) ont un score inférieur à celui attendu "par hasard". Pour les autres, et notamment #2, #3, #4, #5 et #8, nous prédisons assez précisément la période à laquelle leur CF devrait être atteinte (avec des probabilités plus élevées).
Nous obtenons également de bons résultats pour les patients $CALR^m$, avec toutefois pour deux d'entre eux (13%) un score inférieur à celui attendu "par hasard", et pour trois d'entre eux (20%) un score approximativement égal à 0.05. Notons qu'en raison de la stochasticité de notre modèle, il ne faut pas s'attendre à n'avoir que d'excellents scores puisque les distributions *a posteriori* représentées sur la figure 20 sont destinées à représenter la variabilité sur une population de patients.

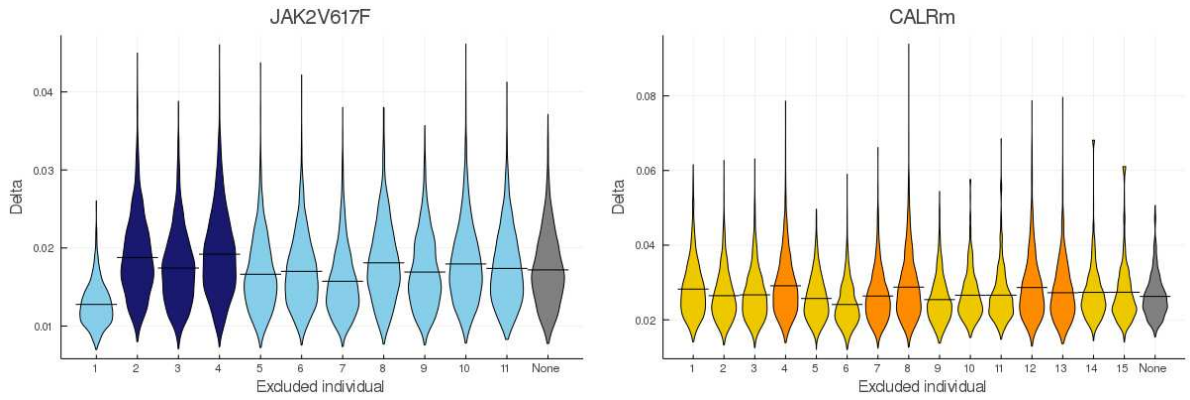


FIGURE 19 – Distribution *a posteriori* $p[\Delta|\mathcal{D}_{-i}]$ du paramètre Δ lors du retrait du patient i (axe des abscisses) des populations de patients $JAK2^{V617F}$ (gauche) ou $CALR^{m}$ (droite). La distribution de droite (grise) représente la distribution *a posteriori* calculée sans exclure aucun individu. Les lignes noires horizontales indiquent les moyennes *a posteriori*. Les individus présentant des mutations supplémentaires sont représentés en foncé.

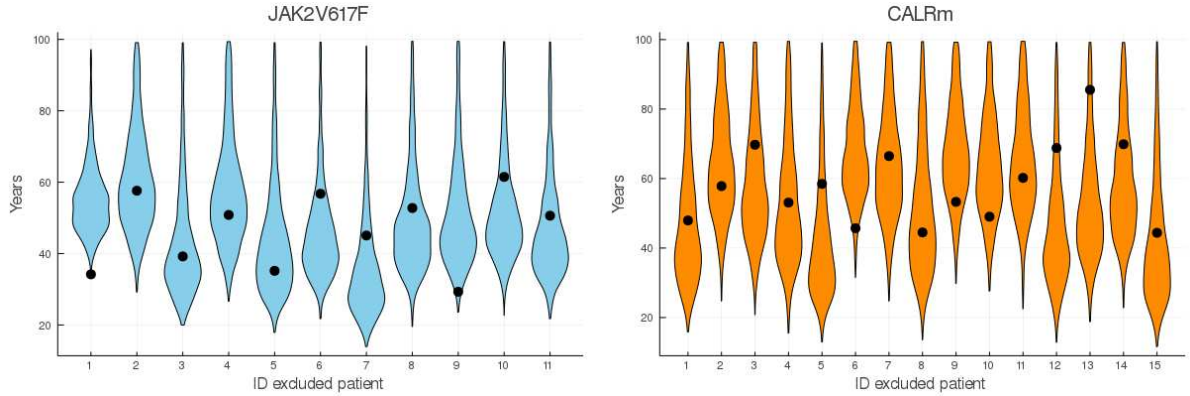


FIGURE 20 – Distributions *a posteriori* du temps t_i (axe des ordonnées) auquel la CF observée $\hat{\eta}_i$ du patient i (axe des abscisses) pourrait être atteinte : $p[t_i|\mathcal{D}_{-i}, \hat{\eta}_i]$ pour les populations de patients $JAK2^{V617F}$ (gauche) ou $CALR^{m}$ (droite). Les points noirs indiquent $p[t_i = \hat{t}_i|\mathcal{D}_{-i}, \hat{\eta}_i]$.

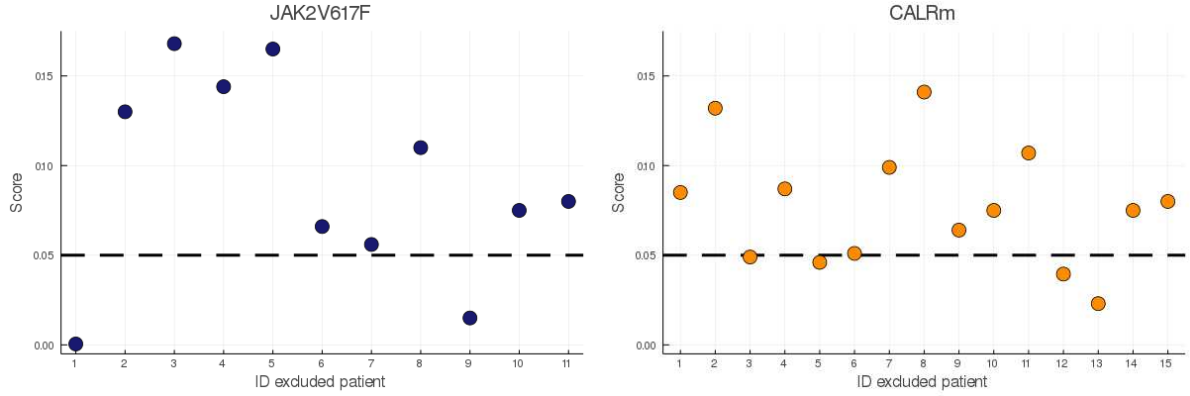


FIGURE 21 – Score - donné par l'équation (9) - pour les patients $JAK2^{V617F}$ (gauche) ou $CALR^{m}$ (droite). Nous les comparons au score que l'on pourrait attendre par hasard (égal à 0.05). Les scores élevés indiquent que le temps $\hat{t}_i$ auquel la CF $\hat{\eta}_i$ du patient i est mesurée est proche de ce qui serait attendu avec la plus grande probabilité par le modèle (calibré sans le patient i).

7.2 Comparaison avec d'autres études

Pour mieux évaluer la validité de notre modèle, nous vérifions que nos estimations dans le cas de la mutation $JAK2^{V617F}$ sont cohérentes avec celles rapportées par d'autres [6, 7, 8]. Nous nous concentrons d'abord sur nos estimations pour le temps d'acquisition moyen λ, puis pour l'avantage prolifératif Δ.

7.2.1 Temps d'acquisition

Van Egeren et al. [6] et Williams et al. [7] ont estimé, à partir de la construction d'arbres phylogénétiques, le temps d'acquisition individuel de la mutation $JAK2^{V617F}$ pour plusieurs patients tandis que, dans ce chapitre, nous inférons un temps d'acquisition moyen au niveau de la population. Nous confronterons nos résultats aux leurs. Plus précisément, Van Egeren et al. et Williams et al. estiment une période au cours de laquelle la mutation aurait dû être acquise.

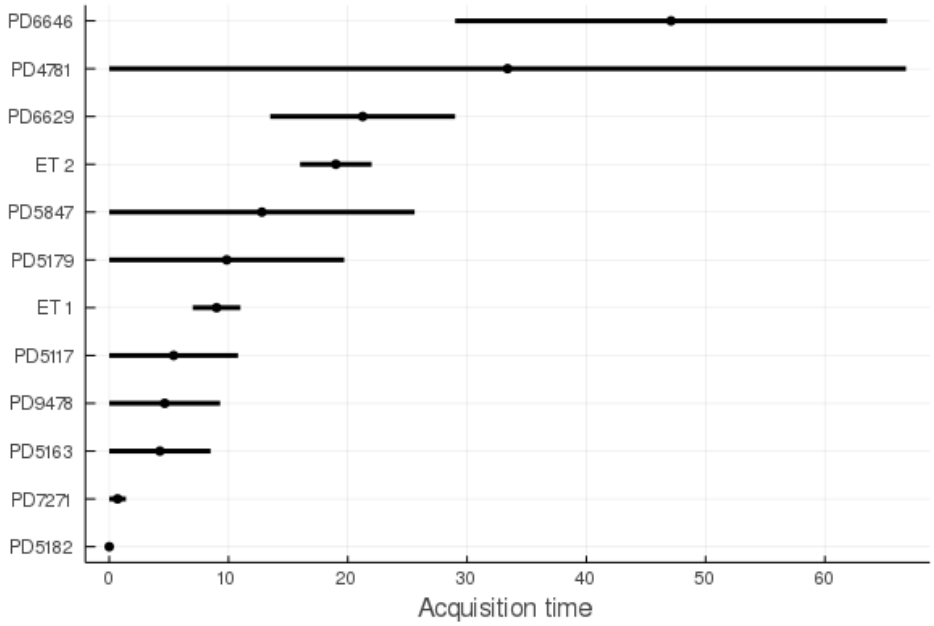


FIGURE 22 – Données de Williams et al. [7] (individus dont l'ID commence par "PD") et de Van Egeren et al. [6] (patients ET 1 et ET 2). Nous indiquons les estimations inférieures et supérieures du temps d'acquisition, telles que rapportées dans les articles correspondants. Les points noirs indiquent les estimations moyennes (la valeur moyenne calculée sur les 12 individus est égale à 14 ans). Les valeurs d'âge pendant la vie fœtale (entre 0 et 36 semaines après la conception) sont fixées à $T_0 = 0$.

Nous rapportons ci-dessous leurs conclusions (Fig. 22) :

- PD7271 : entre 6.9 semaines (post-conception) et 1.4 ans
- PD5163 : entre 3.1 semaines (post-conception) et 8.5 ans
- PD5117 : entre 2.9 semaines (post-conception) et 10.8 ans
- PD5182 : entre 6.4 semaines et 32.8 semaines (post-conception)
- PD5179 : entre 0 et 19.7 ans
- PD6629 : entre 13.5 et 29 ans
- PD5847 : entre 0 et 25.6 ans
- PD9478 : entre 0 et 9.3 ans
- PD6646 : entre 29 et 65.2 ans
- PD4781 : entre 0 et 66.8 ans
- ET1 : à l'âge de 9 ans $\pm$ 2
- ET2 : à l'âge de 19 ans $\pm$ 3

Nous désignons cet ensemble de données de validation par $\mathcal{D}'$. Pour chaque individu i, nous avons

une limite inférieure et supérieure pour le vrai temps d'acquisition $T_{0,i} \in [\hat{T}_{inf,i}, \hat{T}_{sup,i}]$. Notons que Williams et al. ont également rapporté des âges d'acquisition en semaines post-conception, c'est-à-dire qu'ils ont estimé un âge d'acquisition pendant la vie fœtale. Dans notre étude, toute la période de vie fœtale est modélisée par $T_0 = 0$. C'est aussi pourquoi nous devons autoriser que $\lambda = 0$ (équivalent à $T_0 = 0$) avec une probabilité non nulle.

Dans un premier temps, si l'on considère que le temps d'acquisition de chaque individu est de $\hat{T}_{0,i} = 0,5 \cdot (\hat{T}_{sup,i} - \hat{T}_{inf,i})$, une estimation grossière du temps d'acquisition moyen $\hat{\lambda}$ basée sur les données de validation $\mathcal{D}'$ serait la suivante :

$$\hat{\lambda} = \frac{1}{12} \sum_{i=1}^{12} \hat{T}_{0,i} = 14 \text{ années}$$

Cette estimation est comparable à la nôtre (une valeur moyenne pour λ estimée à 15.2 ans).

Ensuite, nous proposons une deuxième approche qui tient compte de l'incertitude sur le jeu de données de validation (c'est-à-dire des limites inférieures et supérieures). Notre méthode repose ici sur deux hypothèses. Premièrement, nous supposons que $T_{0,i}$ (pour l'individu i) est une variable aléatoire qui peut être égale à 0 avec une probabilité b ou sinon (avec une probabilité $1-b$) est distribuée sur $\mathbb{R}^+$ suivant une loi exponentielle de taux l^{-1}. Cette hypothèse est faite pour être en accord avec notre modèle (voir § 4.4). Dans ce cas, nous avons $\lambda = \mathbb{E}[T_{0,i}] = b \times l$.
Deuxièmement, nous supposons que $T_{0,i} - T_{inf,i} | T_{0,i} \sim \mathcal{E}(\mu)$ et $T_{sup,i} - T_{0,i} | T_{0,i} \sim \mathcal{E}(\mu)$ (avec μ identique dans les deux distributions). Autrement dit, nous considérons que les limites supérieures et inférieures doivent être trouvées "près" de la vraie valeur du temps d'acquisition.
Sous ces deux hypothèses, nous obtenons l'expression de la vraisemblance :

$$p[\mathcal{D}'|l,b] = \prod_{i=1}^{12} p[(\hat{T}_{inf,i}, \hat{T}_{sup,i})|l,b]$$

avec, pour $i \in \{1, \cdots, 12\}$:

$$\begin{aligned}
p[(\hat{T}_{inf,i}, \hat{T}_{sup,i})|l,b] &= \int_{\mathbb{R}} p[(\hat{T}_{inf,i}, \hat{T}_{sup,i})|T_{0,i} = t] \cdot p[T_{0,i} = t|l,b] \; dt \\
&= \int_{\hat{T}_{inf,i}}^{\hat{T}_{sup,i}} p[\hat{T}_{inf,i}|T_{0,i} = t] \cdot p[\hat{T}_{sup,i}|T_{0,i} = t] \cdot p[T_{0,i} = t|l,b] \; dt \\
&= \int_{\hat{T}_{inf,i}}^{\hat{T}_{sup,i}} \frac{1}{\mu} \exp(-\frac{1}{\mu}(t - \hat{T}_{inf,i})) \frac{1}{\mu} \exp(-\frac{1}{\mu}(\hat{T}_{sup,i} - t)) \cdot p[T_{0,i} = t|l,b] \; dt \\
&= \int_{\hat{T}_{inf,i}}^{\hat{T}_{sup,i}} \frac{1}{\mu^2} \exp(-\frac{1}{\mu}(\hat{T}_{sup,i} - \hat{T}_{inf,i})) \cdot p[T_{0,i} = t|l,b] \; dt \\
&= \frac{1}{\mu^2} \exp(-\frac{1}{\mu}(\hat{T}_{sup,i} - \hat{T}_{inf,i})) \int_{\hat{T}_{inf,i}}^{\hat{T}_{sup,i}} p[T_{0,i} = t|l,b] \; dt \\
&\propto \int_{\hat{T}_{inf,i}}^{\hat{T}_{sup,i}} p[T_{0,i} = t|l,b] \; dt \\
&= p[T_{0,i} \in [\hat{T}_{inf,i}, \hat{T}_{sup,i}]|l,b] \\
&= F_{l,b}(\hat{T}_{sup,i}) - F_{l,b}(\hat{T}_{inf,i})
\end{aligned}$$

avec $F_{l,b}$ la fonction de densité cumulée de $T_{0,i}$.

Nous obtenons la log-vraisemblance profilée des deux paramètres, affichée sur la figure 23. La log-vraisemblance profilée pour le paramètre b (et il en va de même pour l) est définie par :

$$\mathcal{L}_b = \max_l \log\left(p[\mathcal{D}'|l,b]\right)$$

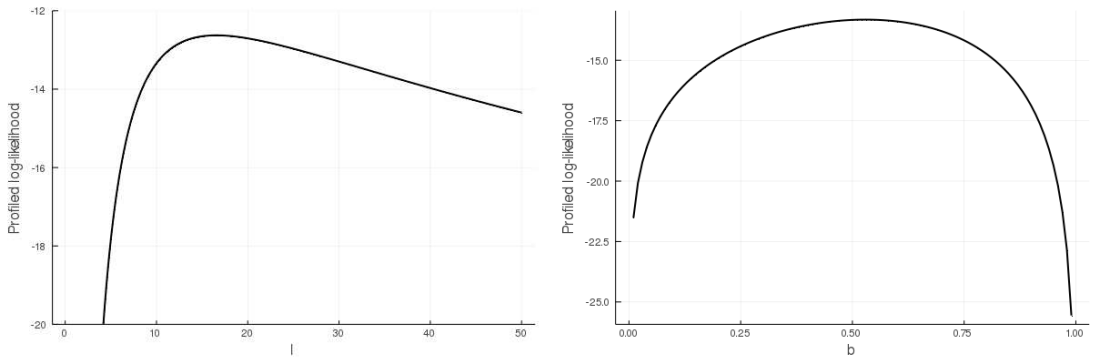


FIGURE 23 – Log-vraisemblance profilée des paramètres l (à gauche) et b (à droite). L'estimateur du maximum de vraisemblance (EMV) est $(\hat{l}, \hat{b}) = (17.71, 0.529)$.

Nous estimons les distributions *a posteriori* des deux paramètres en utilisant l'algorithme de Metropolis-Hastings (initialisé à l'EMV, 5 millions d'itérations, *burn-in length* de 2 000 000), avec des *priors* uniformes pour b sur $[0,1]$ et l sur $[0,100]$. Les distributions *a posteriori* de ces paramètres sont représentées sur la figure 24.

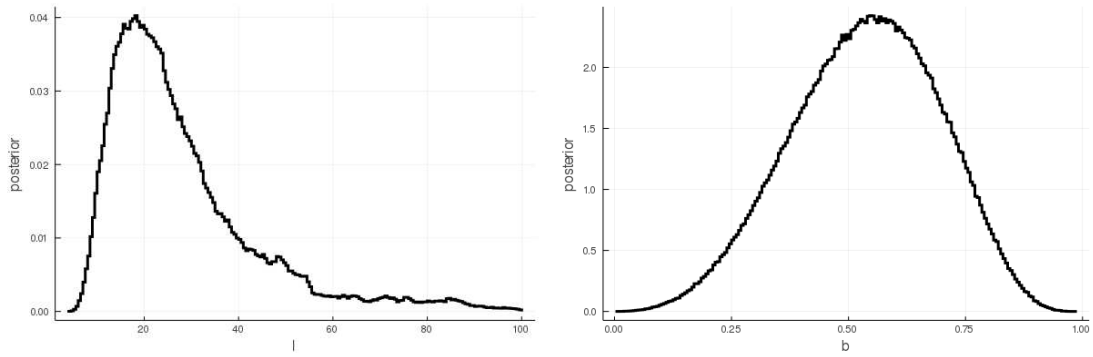


FIGURE 24 – Distribution *a posteriori* des paramètres l (gauche) et b (droite), étant donné le jeu de données de validation $\mathcal{D}'$.

Ensuite, nous pouvons estimer la distribution *a posteriori* du temps d'acquisition moyen, égal à $b \times l$ (voir Fig. 25). Le temps d'acquisition moyen (moyenne *a posteriori*) est estimé à 15.4 (valeur médiane égale à 12.5, et avec un intervalle de crédibilité de 90% : [4,0, 37,6]). Ainsi, avec cette seconde approche, nous retrouvons des résultats cohérents avec les nôtres.

7.2.2 Fitness

Notre modèle de la dynamique des HSC mutés n'est pas éloigné de celui utilisé par Watson et al. [8]. Pour décrire l'avantage prolifératif des clones mutés, ils introduisent un effet de *fitness* s. Nous pouvons déduire une relation entre leur paramètre s et notre paramètre Δ. En effet, d'après Watson et al. : "*In a time interval dt a single HSC can divide symmetrically producing two terminally differentiated cells [...] with probability D dt, divide symmetrically producing two stem cells with probability Bdt* $= (D + s)dt$, *[...]*" [8].

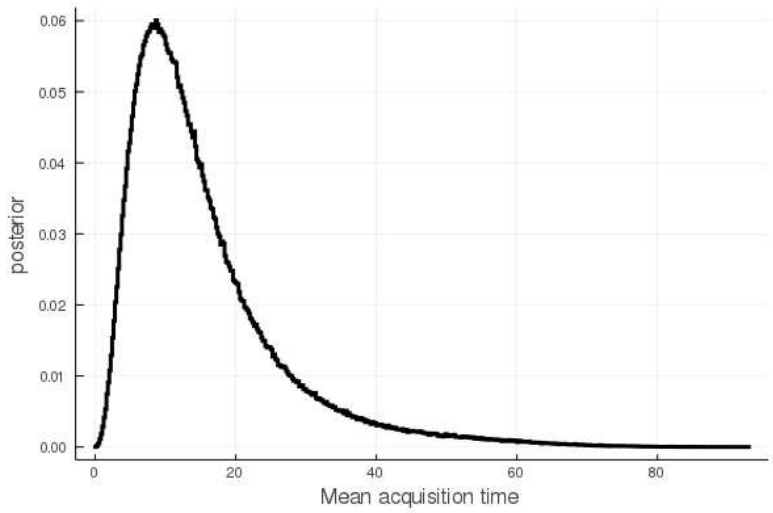


FIGURE 25 – Distribution *a posteriori* du temps d'acquisition moyen (égal à $b \times l$) étant donné le jeu de données de validation $\mathcal{D}'$.

Ainsi, nous déduisons les relations suivantes : $\alpha p_2 dt = (D + s)dt$ et $\alpha p_0 dt = D\ dt$, c'est-à-dire :

$$s = \alpha\Delta$$

Watson et al. ont estimé, pour l'effet de *fitness*, que $s = 14.6$ (% par an). Dans notre cas, nous estimons (moyenne *a posteriori*) $\alpha\Delta = 20.4$ (% par an) avec un intervalle de crédibilité de 95% : [14.2 , 28.6]. Par rapport à leurs résultats, nous estimons un effet de *fitness* plus élevé du clone muté $JAK2^{V617F}$. Cette différence peut s'expliquer par le fait que Watson et al. ont étudié l'hématopoïèse clonale, alors que nous inférons un avantage prolifératif sur la base des observations des patients atteints de NMP (cf. section 5.2).

8 Discussion

Nous avons développé dans ce chapitre une approche mathématique - combinant modélisation et inférence statistique - pour inférer la dynamique des NMP (initiation et développement) et mettre en place des stratégies de dépistage précoce, en particulier dans le cas où les NMP sont associés aux mutations motrices $JAK2^{V617F}$ et $CALR^m$. Ces deux mutations motrices sont acquises au niveau des cellules souches, ce qui rend difficile l'étude expérimentale de cette hémopathie maligne chez l'homme et justifie l'utilisation de modèles mathématiques. Ceux-ci sont nécessairement des simplifications de la réalité et reposent sur plusieurs hypothèses, sans lesquelles il ne serait pas possible de faire des inférences statistiques à partir des données. Nous avons supposé que les paramètres du modèle dépendaient uniquement du type de la mutation motrice principale ($JAK2^{V617F}$ *vs* $CALR^m$) et non des patients. Ainsi, nous avons considéré que l'hétérogénéité entre les patients ne résultait que d'effets stochastiques. Pour affiner nos résultats et explorer l'influence de facteurs génétiques additionnels, il faudrait augmenter le nombre de patients inclus dans notre cohorte pour les répartir en sous-groupes de taille suffisante. En particulier, il serait pertinent d'augmenter le nombre de patients $CALR^m$ pour tenir compte des différences entre les mutations $CALR^m$ T1 et T2, qui sont cliniquement distinctes [34] et dont nous avons vu au chapitre 2 (dans le cas de modèles murins) qu'il existait des différences de signalisation au niveau des cellules souches et progénitrices. Par ailleurs, une limite de notre travail est que nous n'incluons que des patients sans sous-clones homozygotes. Or, les cellules mutées homozygotes sont supposées avoir un avantage prolifératif accru qui pourrait accélérer le développement de la maladie [7]. Ainsi, pour étendre notre travail, nous pourrions complexifier le modèle en intégrant la dynamique des sous-clones homozygotes. Un tel modèle étendu pourrait également décrire l'expansion des sous-clones avec des mutations supplémentaires (*TET2*, *DNMT3A*,...). L'avantage prolifératif des clones mutés, modélisé par Δ, a également été considéré comme constant au cours de la vie. Cependant, des modèles plus précis devraient probablement prendre en compte des effets de seuil, éventuellement justifiés par des mécanismes de régulation, pour éviter la possibilité d'une forte expansion irréaliste. Les HSC WT sont également susceptibles d'acquérir des mutations qui augmenteraient leur avantage prolifératif, réduisant l'avantage sélectif relatif des clones $JAK2^{V617F}$ ou $CALR^m$, tandis que ces derniers pourraient également acquérir des mutations associées, augmentant leur avantage prolifératif. La valeur estimée pour le paramètre Δ doit donc être considéré comme une valeur moyenne qui englobe différents effets se produisant au cours de la vie. De plus, nous avons implicitement considéré que l'évolution des NMP est *drivée* par la sélection naturelle en supposant que $\Delta > 0$ et que la prolifération du clone muté se produit dans un *pool* important et constant de HSC WT. Comme discuté dans Lyne et al. [35], une estimation correcte du nombre de HSC WT serait essentielle pour distinguer la sélection naturelle de l'évolution neutre, mais aussi dans notre cas, pour quantifier plus précisément le paramètre Δ. De plus, le fait que les HSC âgées pourraient se diviser plutôt de façon symétrique, comme discuté dans Florian et al. impliquerait que le nombre de HSC WT augmente avec l'âge et ne reste pas constant [36]. Des modèles plus complexes pourraient être développés, mais ils nécessiteraient des observations longitudinales et des ensembles de données plus importants pour leur calibration. Notre modèle présente l'avantage de pouvoir être utilisé pour étudier d'autres mutations survenant dans divers cancers du sang et que sa calibration ne nécessite que des mesures de CF parmi les cellules progénitrices ($CD34^+$). Appliqué à la mutation $JAK2^{V617F}$, nous avons trouvé un temps d'acquisition moyen cohérent avec d'autres études [6, 7]. Notre estimation du paramètre Δ est plus grande que celle obtenue par Watson et al, qui ont étudié l'hématopoïèse clonale d'individus normaux [8] ; l'agressivité des NMP semble compatible avec un avantage prolifératif plus élevé des clones mutants par rapport à l'hématopoïèse clonale générale [37, 38, 39]. De plus, en appliquant la même approche aux patients $CALR^m$, nous avons pu mettre en évidence des différences entre l'expansion clonale des deux principales mutations motrices des NMP. Ces différences entre $JAK2^{V617F}$ et $CALR^m$ ont également été observées dans des modèles murins par le biais d'une greffe compétitive de cellules de moelle osseuse à une dilution limite de HSC dans laquelle les HSC $CALR^m$ supplantent les cellules WT plus rapidement que les HSC $JAK2^{V617F}$ pour induire la maladie [40, 41, 42]. En accord avec nos données, il a été montré dans les progé-

niteurs précoces de patients atteints de thrombocytémie essentielle que $JAK2^{V617F}$ donnait une dominance clonale plus faible que $CALR^m$ [6, 19, 31, 43]. De plus, nos résultats suggèrent que la mutation $CALR^m$ n'est probablement pas acquise pendant la vie fœtale, mais en moyenne 25 ans après la naissance, contrairement à la mutation $JAK2^{V617F}$. Cela est cohérent avec les observations selon lesquelles les jeunes patients atteints de NMP et présentant un syndrome de Budd-Chiari inaugural (âge moyen de 35 ans) auraient plutôt la mutation $JAK2^{V617F}$ que $CALR^m$ (90% contre 2%) [44], ou encore avec l'étude de Sobas et al. conduite sur une cohorte d'enfants et jeunes adultes parmi lesquels aucune mutation $CALR^m$ n'a été trouvée entre 0 et 9 ans [45]. Pourtant, des résultats récents ont également mis en évidence une possible acquisition de la mutation $CALR^m$ *in utero* [21].
Différents types de mécanismes semblent exister pour l'acquisition de la mutation motrice des NMP, qui pourraient s'expliquer par une différence dans le type de mutation : une mutation ponctuelle pour $JAK2^{V617F}$ et des délétions/insertions avec un décalage de cadre (*frameshift*) de $+1$ pour les mutations *CALR*, pourrait résulter de deux types distincts de mécanismes déficients de réparation de l'ADN (conduisant à une transversion G→T et à une délétion/insertion), et ces défauts pourraient apparaître à des âges différents de la vie. Une observation intéressante de Cordua et al. [46] est que même si la mutation $JAK2^{V617F}$ est plus fréquente que la mutation $CALR^m$ dans la population générale saine, la proportion de patients malades chez les individus $CALR^m$ est beaucoup plus élevée que chez les individus $JAK2^{V617F}$. Cela suggère que la latence de la maladie est plus courte avec la mutation $CALR^m$ et que le clone $CALR^m$ s'étend plus rapidement que $JAK2^{V617F}$, ce qui est conforme à nos résultats. De plus, même si la mutation $JAK2^{V617F}$ survient avant $CALR^m$, l'expansion plus rapide pour $CALR^m$ peut expliquer que la maladie $CALR^m$ débute 10 ans en moyenne avant la maladie $JAK2^{V617F}$ [47, 48].

Enfin, notre méthode d'inférence repose sur un cadre Bayésien permettant d'estimer les distributions de probabilité des paramètres, et donc une évaluation correcte de l'incertitude des paramètres. Une telle approche est coûteuse en termes de temps de calcul et repose sur des méthodes numériques efficaces pour permettre la convergence des algorithmes dans un temps réaliste. Nous avons mis en place une procédure d'optimisation pour rendre cette étude possible, en proposant une approximation déterministe de notre modèle stochastique, et en utilisant un algorithme d'affectation optimale. Ensuite, à partir des distributions *a posteriori* des paramètres, nous avons pu déduire plus précisément la dynamique de l'expansion clonale et explorer des stratégies de dépistage précoce. Nous avons constaté que près de la moitié des patients $CALR^m$ ne seraient pas détectés à l'âge optimal de dépistage. Par conséquent, le dépistage de la mutation *CALR* n'est pas recommandé à moins que les individus ne soient examinés plusieurs fois dans leur vie. En revanche, nous avons constaté que la détection précoce pourrait être une option clinico-biologique viable pour détecter la mutation $JAK2^{V617F}$, avec un âge optimal pour mesurer le VAF dans le sang périphérique d'environ 30 ans. Étant donné que cette mutation est présente dans un nombre important de cas de thromboses splanchniques et cérébrales inexpliquées chez des patients présentant un NMP non diagnostiqué avec une faible VAF pour $JAK2^{V617F}$, une détection précoce pourrait prévenir les événements thrombotiques en mettant les patients sous anti-coagulants ou aspirine.

En conclusion, notre méthode - appliquée aux deux mutations les plus répandues dans les NMP - illustre le potentiel de la modélisation mathématique pour aider à déduire le moment de l'apparition du cancer du sang, mieux comprendre son développement et concevoir des stratégies de détection précoce adaptées au type de mutation. Les individus qui n'auraient pas pu être dépistés précocement devraient alors être pris en charge cliniquement. Au chapitre suivant, nous étudierons le traitement de ces patients par Interféron α. À l'avenir, il sera peut-être possible de tester simultanément de nombreuses mutations malignes par NGS au lieu de considérer chacune d'entre elles séparément. Un tel effort de dépistage à grande échelle se traduirait par un problème d'optimisation plus complexe ; cependant, la flexibilité de notre cadre de modélisation rend son extension à un panel de gènes réalisable.

Références

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

[2] Chloé James, Valérie Ugo, Jean-Pierre Le Couédic, Judith Staerk, François Delhommeau, Catherine Lacout, Loïc Garçon, Hana Raslova, Roland Berger, Annelise Bennaceur-Griscelli, et al. A unique clonal jak2 mutation leading to constitutive signalling causes polycythaemia vera. *nature*, 434(7037) :1144–1148, 2005.

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

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

[5] Jyoti Nangalia, Charles E Massie, E Joanna Baxter, Francesca L Nice, Gunes Gundem, David C Wedge, Edward Avezov, Juan Li, Karoline Kollmann, David G Kent, et al. Somatic calr mutations in myeloproliferative neoplasms with nonmutated jak2. *New England Journal of Medicine*, 369(25) :2391–2405, 2013.

[6] Debra Van Egeren, Javier Escabi, Maximilian Nguyen, Shichen Liu, Christopher R Reilly, Sachin Patel, Baransel Kamaz, Maria Kalyva, Daniel J DeAngelo, Ilene Galinsky, et al. Reconstructing the lineage histories and differentiation trajectories of individual cancer cells in myeloproliferative neoplasms. *Cell stem cell*, 28(3) :514–523, 2021.

[7] Nicholas Williams, Joe Lee, Emily Mitchell, Luiza Moore, E Joanna Baxter, James Hewinson, Kevin J Dawson, Andrew Menzies, Anna L Godfrey, Anthony R Green, et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature*, 602(7895) :162–168, 2022.

[8] Caroline J Watson, AL Papula, Gladys YP Poon, Wing H Wong, Andrew L Young, Todd E Druley, Daniel S Fisher, and Jamie R Blundell. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science*, 367(6485) :1449–1454, 2020.

[9] P Hirsch, AC Mamez, R Belhocine, S Lapusan, R Tang, L Suner, D Bories, C Marzac, F Fava, O Legrand, et al. Clonal history of a cord blood donor cell leukemia with prenatal somatic jak2 v617f mutation. *Leukemia*, 30(8) :1756–1759, 2016.

[10] Warren John Ewens. *Mathematical population genetics : theoretical introduction*, volume 1. Springer, 2004.

[11] David Axelrod and Marek Kimmel. *Branching processes in biology*. Springer-Verlag, 2015.

[12] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6) :1760–1765, 2007.

[13] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31) :187–202, 2009.

[14] Mahmoud Bentriou, Paolo Ballarini, and Paul-Henry Cournède. Automaton-abc : A statistical method to estimate the probability of spatio-temporal properties for parametric markov population models. *Theoretical Computer Science*, 893 :191–219, 2021.

[15] Mahmoud Bentriou. *Statistical Inference and Verification of Chemical Reaction Networks*. PhD thesis, Université Paris-Saclay, 2021.

[16] Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert. Adaptive approximate bayesian computation. *Biometrika*, 96(4) :983–990, 2009.

[17] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2) :83–97, 1955.

[18] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1) :32–38, 1957.

[19] Mira El-Khoury, Xénia Cabagnols, Matthieu Mosca, Gaëlle Vertenoeil, Christophe Marzac, Fabrizia Favale, Olivier Bluteau, Florence Lorre, Amandine Tisserand, Graciela Rabadan Moraes, et al. Different impact of calreticulin mutations on human hematopoiesis in myeloproliferative neoplasms. *Oncogene*, 39(31) :5323–5337, 2020.

[20] Hiroshi Haeno, Ross L Levine, D Gary Gilliland, and Franziska Michor. A progenitor cell origin of myeloid malignancies. *Proceedings of the National Academy of Sciences*, 106(39) :16616–16621, 2009.

[21] Nikolaos Sousos, Máire Ní Leathlobhair, Christina Simoglou Karali, Eleni Louka, Nicola Bienz, Daniel Royston, Sally-Ann Clark, Angela Hamblin, Kieran Howard, Vikram Mathews, et al. In utero origin of myelofibrosis presenting in adult monozygotic twins. *Nature Medicine*, pages 1–5, 2022.

[22] Henry Lee-Six, Nina Friesgaard Øbro, Mairi S Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J Osborne, Brian JP Huntly, Inigo Martincorena, Elizabeth Anderson, et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561(7724) :473–478, 2018.

[23] Emily Mitchell, Michael Spencer Chapman, Nicholas Williams, Kevin J Dawson, Nicole Mende, Emily F Calderbank, Hyunchul Jung, Thomas Mitchell, Tim HH Coorens, David H Spencer, et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature*, pages 1–8, 2022.

[24] Sandra N Catlin, Lambert Busque, Rosemary E Gale, Peter Guttorp, and Janis L Abkowitz. The replication rate of human hematopoietic stem cells in vivo. *Blood, The Journal of the American Society of Hematology*, 117(17) :4460–4466, 2011.

[25] Nathalie Rufer, Tim H Brümmendorf, Steen Kolvraa, Claus Bischoff, Kaare Christensen, Louis Wadsworth, Michael Schulzer, and Peter M Lansdorp. Telomere fluorescence measurements in granulocytes and t lymphocyte subsets point to a high turnover of hematopoietic stem cells and memory t cells in early childhood. *The Journal of experimental medicine*, 190(2) :157–168, 1999.

[26] Richard C Allsopp, Homayoun Vaziri, Christopher Patterson, Samuel Goldstein, Edward V Younglai, A Bruce Futcher, Carol W Greider, and Calvin B Harley. Telomere length predicts replicative capacity of human fibroblasts. *Proceedings of the National Academy of Sciences*, 89(21) :10114–10118, 1992.

[27] Peter M Lansdorp. Telomere length and proliferation potential of hematopoietic stem cells. *Journal of Cell Science*, 108(1) :1–6, 1995.

[28] Homayoun Vaziri, Wieslawa Dragowska, Richard C Allsopp, Terry E Thomas, Calvin B Harley, and Peter M Lansdorp. Evidence for a mitotic clock in human hematopoietic stem cells : loss of telomeric dna with age. *Proceedings of the National Academy of Sciences*, 91(21) :9857–9860, 1994.

[29] Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate bayesian computational methods. *Statistics and Computing*, 22(6) :1167–1180, 2012.

[30] RWR Darling, James R Norris, et al. Differential equation approximations for markov chains. *Probability surveys*, 5 :37–79, 2008.

[31] Sabrina Dupont, Aline Massé, Chloé James, Irène Teyssandier, Yann Lécluse, Frédéric Larbret, Valérie Ugo, Patrick Saulnier, Serge Koscielny, Jean Pierre Le Couédic, et al. The jak2 617v> f mutation triggers erythropoietin hypersensitivity and terminal erythroid amplification in primary cells from patients with polycythemia vera. *Blood, The Journal of the American Society of Hematology*, 110(3) :1013–1021, 2007.

[32] Franz-Georg Wieland, Adrian L. Hauber, Marcus Rosenblatt, Christian Tönsing, and Jens Timmer. On structural and practical identifiability. *Current Opinion in Systems Biology*, 25 :60–69, 2021.

[33] Ronan Duchesne, Anissa Guillemin, Fabien Crauste, and Olivier Gandrillon. Calibration, selection and identifiability analysis of a mathematical model of the in vitro erythropoiesis in normal and perturbed contexts. *In silico biology*, 13(1-2) :55–69, 2019.

[34] Ayalew Tefferi, Emnet A Wassie, Paola Guglielmelli, Naseema Gangat, Alem A Belachew, Terra L Lasho, Christy Finke, Rhett P Ketterling, Curtis A Hanson, Animesh Pardanani, et al. Type 1 versus type 2 calreticulin mutations in essential thrombocythemia : a collaborative study of 1027 patients. *American journal of hematology*, 89(8) :E121–E124, 2014.

[35] Anne-Marie Lyne, Lucie Laplane, and Leïla Perié. To portray clonal evolution in blood cancer, count your stem cells. *Blood*, 137(14) :1862–1870, 2021.

[36] M Carolina Florian, Markus Klose, Mehmet Sacma, Jelena Jablanovic, Luke Knudson, Kalpana J Nattamai, Gina Marka, Angelika Vollmer, Karin Soller, Vadim Sakk, et al. Aging alters the epigenetic asymmetry of hsc division. *PLoS biology*, 16(9) :e2003389, 2018.

[37] Siddhartha Jaiswal, Pierre Fontanillas, Jason Flannick, Alisa Manning, Peter V Grauman, Brenton G Mar, R Coleman Lindsley, Craig H Mermel, Noel Burtt, Alejandro Chavez, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine*, 371(26) :2488–2498, 2014.

[38] Giulio Genovese, Anna K Kähler, Robert E Handsaker, Johan Lindberg, Samuel A Rose, Samuel F Bakhoum, Kimberly Chambert, Eran Mick, Benjamin M Neale, Menachem Fromer, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood dna sequence. *New England Journal of Medicine*, 371(26) :2477–2487, 2014.

[39] Thomas McKerrell, Naomi Park, Thaidy Moreno, Carolyn S Grove, Hannes Ponstingl, Jonathan Stephens, Charles Crawley, Jenny Craig, Mike A Scott, Clare Hodkinson, et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell reports*, 10(8) :1239–1245, 2015.

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

[41] Camélia Benlabiod, Maira da Costa Cacemiro, Audrey Nédélec, Valérie Edmond, Delphine Muller, Philippe Rameau, Laure Touchard, Patrick Gonin, Stefan N Constantinescu, Hana Raslova, et al. Calreticulin del52 and ins5 knock-in mice recapitulate different myeloproliferative phenotypes observed in patients with mpn. *Nature communications*, 11(1) :1–15, 2020.

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

[43] Shubha Anand, Frances Stedham, Philip Beer, Emma Gudgin, Christina A Ortmann, Anthony Bench, Wendy Erber, Anthony R Green, and Brian JP Huntly. Effects of the jak2 mutation on the hematopoietic stem and progenitor compartment in human myeloproliferative neoplasms. *Blood, The Journal of the American Society of Hematology*, 118(1) :177–181, 2011.

[44] Johanne Poisson, Aurélie Plessier, Jean-Jacques Kiladjian, Fanny Turon, Bruno Cassinat, Annalisa Andreoli, Emmanuelle De Raucourt, Odile Goria, Kamal Zekrini, Christophe Bureau, et al. Selective testing for calreticulin gene mutations in patients with splanchnic vein thrombosis : a prospective cohort study. *Journal of Hepatology*, 67(3) :501–507, 2017.

[45] Marta Sobas, Jean-Jacques Kiladjian, Yan Beauverd, Natalia Curto-Garcia, Parvis Sadjadian, Lee Yung Shih, Timothy Devos, Dorota Krochmalczyk, Serena Galli, Maria Bieniaszewska, et al. Real world study of children and young adults with myeloproliferative neoplasms identifying risks and unmet needs. *Blood Advances*, 2022.

[46] Sabrina Cordua, Lasse Kjaer, Vibe Skov, Niels Pallisgaard, Hans C Hasselbalch, and Christina Ellervik. Prevalence and phenotypes of jak2 v617f and calreticulin mutations in a danish general population. *Blood, The Journal of the American Society of Hematology*, 134(5) :469–479, 2019.

[47] A Tefferi, J Thiele, AM Vannucchi, and T Barbui. An overview on calr and csf3r mutations and a proposal for revision of who diagnostic criteria for myeloproliferative neoplasms. *Leukemia*, 28(7) :1407–1413, 2014.

[48] Elisa Rumi, Daniela Pietra, Virginia Ferretti, Thorsten Klampfl, Ashot S Harutyunyan, Jelena D Milosevic, Nicole CC Them, Tiina Berg, Chiara Elena, Ilaria C Casetti, et al. Jak2 or calr mutation status defines subtypes of essential thrombocythemia with substantially different clinical course and outcomes. *Blood, The Journal of the American Society of Hematology*, 123(10) :1544–1551, 2014.

Chapitre 6
Modélisation de l'effet du traitement à l'Interféron α

Résumé

Partant d'observations longitudinales pour des patients atteints de néoplasmes myéloprolifératifs et traités à l'Interféron α (IFNα), nous construisons un modèle mathématique permettant d'inférer la dynamique latente des cellules souches mutées, sous action du traitement, qui conduit aux dynamiques observées au niveau de progéniteurs hématopoïétiques et de granulocytes. Nous proposons un modèle compartimental déterministe, basé sur l'hypothèse que l'IFNα ciblerait les cellules souches mutées en favorisant leur sortie de quiescence et leur différenciation en progéniteurs. Après différentes simplifications nous permettant d'obtenir un modèle identifiable, nous mettons en place une méthode d'inférence Bayésienne hiérarchique pour estimer les paramètres individuels des patients.
Nos résultats montrent que l'IFNα ciblerait plus efficacement, pour les patients $JAK2^{V617F}$, les cellules souches et progénitrices homozygotes que celles hétérozygotes. Pour ces dernières, nous estimons qu'elles seraient ciblées plus rapidement à fort dosage. Enfin, nous estimons que l'IFNα aurait tendance à cibler les cellules souches mutées $CALR^m$ de type 2, mais peu celles de type 1.

Le contenu de ce chapitre a fait l'objet d'une publication (Mosca*, Hermange*, Tisserand*, Noble* et al., Blood 2021)

* denotes equal contribution

Abstract

Classical BCR-ABL-negative myeloproliferative neoplasms (MPN) are clonal disorders of hematopoietic stem cells (HSC) caused mainly by recurrent mutations in genes encoding JAK2 (*JAK2*), calreticulin (*CALR*), or the thrombopoietin receptor (*MPL*). Interferon alpha (IFNα) has demonstrated some efficacy in inducing molecular remission in MPN. In order to determine factors that influence the molecular response rate, we evaluated the long-term molecular efficacy of IFNα in MPN patients by monitoring the fate of cells carrying driver mutations in a prospective observational and longitudinal study of 48 patients over more than 5 years. We measured several times per year the clonal architecture of early and late hematopoietic progenitors (84,845 measurements) and the global variant allele frequency in mature cells (409 measurements). Using mathematical modeling and hierarchical Bayesian inference, we further inferred the dynamics of IFNα-targeted mutated HSC. Our data support the hypothesis that IFNα targets $JAK2^{V617F}$ HSC by inducing their exit from quiescence and differentiation into progenitors. Our observations indicate that treatment efficacy is higher in homozygous than heterozygous $JAK2^{V617F}$ HSC and increases with high IFNα dosage in heterozygous $JAK2^{V617F}$ HSC. Besides, we found that the molecular responses of $CALR^m$ HSC to IFNα were heterogeneous, varying between type 1 and type 2 $CALR^m$, and high dosage of IFNα correlates with worse outcomes. Together, our work indicates that the long-term molecular efficacy of IFNα implies an HSC exhaustion mechanism and depends on both the driver mutation type and IFNα dosage.

Table des matières

1 Introduction

Le développement des néoplasmes myéloprolifératifs, comme nous l'avons montré au chapitre précédent, se fait sur plusieurs décennies, avec une acquisition de la mutation motrice en moyenne vers l'âge de 25 ans dans le cas $CALR^m$, et vers l'âge de 15 ans dans le cas $JAK2^{V617F}$, avec dans ce dernier cas la possibilité d'une acquisition durant la vie fœtale. Ces maladies favorisent les événements thrombo-hémorragiques et peuvent se transformer en leucémie myéloïde aiguë secondaire. Il est ainsi crucial de prendre en charge les patients le plus tôt possible, pour normaliser leurs paramètres sanguins voire espérer une rémission.
L'interféron alpha (IFNα), en particulier l'IFNα pégylé (notamment Peg-IFNα2a), induit des réponses hématologiques dans les TE, PV et certaines MFP précoces [1, 2, 3, 4, 5], comme l'a récemment démontré un essai clinique de phase 2 réalisé chez des patients réfractaires/intolérants à l'hydroxyurée (HU) atteints de NMP $JAK2^{V617F}$ et $CALR^m$ [6]. Des essais cliniques randomisés de phase 3 ont démontré qu'un autre IFN pégylé, Ropeg-IFNα2b, augmentait le taux de réponse hématologique par rapport à l'HU [7, 8] ou à la phlébotomie [9] chez des patients PV. Il est important de noter que l'IFNα, contrairement aux thérapies cytoréductrices ou aux inhibiteurs de JAK, est capable de diminuer la charge allélique (Variant Allele Frequency - VAF) $JAK2^{V617F}$ dans les cellules sanguines chez environ 60% des patients, et surtout d'induire des réponses moléculaires complètes dans 20% des cas [1, 6, 7], alors que son impact pour des patients $CALR^m$ est sujet à débat [6, 10, 11, 12].
Malgré l'efficacité démontrée de ce traitement, les facteurs qui influencent le taux de réponse moléculaire à long terme induite par l'IFNα restent inconnus. Pour élucider ces déterminants, une analyse longitudinale prospective de la dynamique à long terme de $JAK2^{V617F}$, $CALR^m$ et MPL^m dans les cellules souches et progénitrices hématopoïétiques (HSPC) de patients NMP traités par IFNα a été réalisée à l'Institut Gustave Roussy, par l'équipe d'Isabelle Plo. Le suivi de 48 patients, pendant 5 ans, a permis la constitution d'un riche dataset consistant en des mesures de VAF au niveau de cellules matures ainsi que de l'architecture clonale de progéniteurs immatures. Dans ce chapitre, nous présenterons l'analyse que nous avons effectuée de ces données. Nous chercherons à comprendre comment l'IFNα cible différemment les HSPC en fonction du type de mutation, de la zygosité et du dosage. Pour cela, nous construirons un modèle mathématique dont nous estimerons les paramètres par une méthode d'inférence Bayésienne hiérarchique qui nous permettra d'inférer la dynamique des HSC mutées et leur réponse au traitement par IFNα.

Le travail présenté dans ce chapitre est issu de notre article "*Inferring the dynamics of mutated hematopoietic stem and progenitor cells induced by IFNα in myeloproliferative neoplasms*" publié dans le journal Blood (2021) [13].

2 Observations expérimentales

2.1 Information par patient

Les observations obtenues pour un patient i consistent en des mesures, à différents instants $t_k^{(i)}$ comptés depuis le début de la thérapie ($t = 0$), de l'architecture clonale parmi les cellules progénitrices ($\hat{n}_{k,wt}^{(i)}, \hat{n}_{k,het}^{(i)}, \hat{n}_{k,hom}^{(i)}$) et de la VAF $\hat{y}_k^{(i)}$ parmi les cellules matures (granulocytes), comme schématisé sur la figure 1.
Nous notons $\mathcal{I}^{(i)}$ l'ensemble des temps d'observations pour le patient i. Nous notons :

$$\mathcal{D}^{(i)} = \left(t_k^{(i)}, \hat{n}_{k,wt}^{(i)}, \hat{n}_{k,het}^{(i)}, \hat{n}_{k,hom}^{(i)}, \hat{y}_k^{(i)}\right)_{k \in \mathcal{I}^{(i)}}$$

$\hat{n}_{k,wt}^{(i)}, \hat{n}_{k,het}^{(i)}, \hat{n}_{k,hom}^{(i)}$ représentent respectivement les nombres de progéniteurs wild-type (wt), mutés hétérozygotes (het) et homozygotes (hom) pour la mutation motrice du NMP (soit $JAK2^{V617F}$, $CALR^m$ ou MPL^m), qui ont été échantillonnés et génotypés au temps $t_k^{(i)} \in \mathcal{I}^{(i)}$ pour le patient i. Nous notons :

$$\hat{N}_k^{(i)} := \hat{n}_{k,wt}^{(i)} + \hat{n}_{k,het}^{(i)} + \hat{n}_{k,hom}^{(i)}$$

Nous avons détaillé au chapitre précédent en quoi consistait la mesure de l'architecture clonale pour un échantillon de sang, à un instant donnée. La différence ici est qu'un patient sera suivi au cours de son traitement de telle façon que la manipulation expérimentale est répétée un très grand nombre de fois. Pour être précis l'architecture clonale a été réalisée (sur l'ensemble de la cohorte) pour 395 échantillons de sang - 84,845 colonies dérivées d'un progéniteurs ont ainsi été génotypées - ce qui représente un travail de grande ampleur, en partie réalisé par Matthieu Mosca puis Amandine Tisserand alors en thèse sous la direction d'Isabelle Plo.

Pour la mesure d'architecture clonale d'un seul échantillon de sang, environ 3-4 plaques de 96 puits sont utilisées. Toutes les cellules déposées initialement dans les puits n'aboutiront pas nécessairement à une colonie, et ce pour différentes raisons expérimentales. Aboutir à une colonie est néanmoins nécessaire pour permettre ensuite le génotypage, et donc déduire le génotype (en particulier si la cellule est wt, het ou hom) de la cellule initialement déposée dans le puits. Ainsi, la valeur de $\hat{N}_k^{(i)}$ - et donc l'incertitude sur la mesure - varie en fonction du patient i et de l'échantillon k considérés.
La clonogénicité, c'est-à-dire le rapport entre le nombre de colonies génotypées par rapport au nombre de progéniteurs mis en culture, est d'environ 70% avec une valeur médiane pour $\hat{N}_k^{(i)}$ (sur l'ensemble des 395 échantillons de sang) de 226 colonies génotypées par échantillon.

Les progéniteurs hématopoïétiques (cellules $CD34^+$) ont été marqués avec des anticorps anti-CD90, -CD34 et -CD38 (Becton Dickinson), permettant d'identifier leur type cellulaire parmi les suivants :

- Progéniteurs enrichis en HSC (HSC*) : $CD34^+CD38^-CD90^+$

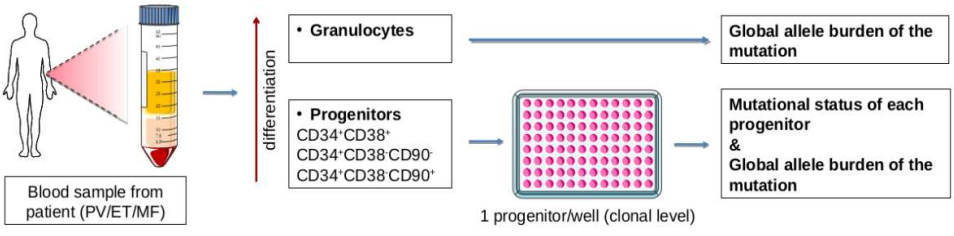


FIGURE 1 – Schéma de la manipulation expérimentale permettant la mesure de l'architecture clonale et de la VAF dans les granulocytes pour l'échantillon de sang d'un patient i à un instant $t_k^{(i)}$ (schéma réalisé par Amandine Tisserand).

- Progéniteurs multipotents (MPP) : $CD34^+CD38^-CD90^-$
- Progéniteurs au potentiel plus restreint (ou progéniteurs engagés, HPC) : $CD34^+CD38^+CD90^-$

Dans la suite, nous simplifierons et ne ferons pas la distinction entre les types de progéniteurs. Nous utiliserons le terme de cellules immatures pour désigner ces cellules progénitrices ($CD34^+$). Notons que, pour un patient, nous n'avons pas nécessairement de prélèvement avant traitement. Pour un prélèvement donné, nous n'avons pas non plus systématiquement à la fois la mesure de la VAF parmi les cellules matures et de l'architecture clonale parmi les cellules immatures.

Les patients sont considérés être sous traitement à partir du temps $t = 0$. Le traitement consiste en l'administration d'une certaine quantité d'Interféron α (entre 0 et 180μg) à une certaine fréquence (généralement de toutes les semaines à toutes les trois semaines). Il n'existe pas de consignes claires quant à l'utilisation de ce traitement, de telle sorte que la posologie est très variable entre les patients. Nous appelons dose d'interféron - notée $d^{(i)}(t)$ - le rapport au temps t de la quantité d'IFNα sur la période de temps entre deux administrations. En cours de traitement, cette dose peut varier. Généralement, le clinicien procède à une escalade de dose dans les premiers mois du traitement, jusqu'à obtenir une réponse hématologique (ou l'apparition d'effets secondaires), puis à une diminution de la dose. Là encore, il n'existe pas de stratégies bien définies sur la façon de faire varier les doses en cours de traitement. Nous étudierons les variations de dose au chapitre suivant. Nous appelons dosage - noté $D^{(i)}$ pour le patient i - la quantité moyenne d'IFNα reçue sur les $T = 450$ premiers jours de traitement :

$$D^{(i)} = \frac{1}{T} \int_{t=0}^{T} d^{(i)}(t)\, dt$$

Nous exprimerons le dosage en μg par semaine. Dans l'intégrale ci-dessus, on choisit de moyenner la dose sur $T = 450$ jours, temps à partir duquel on constate généralement une diminution de la dose par le clinicien.

2.2 Cohorte

Dans cette étude prospective, longitudinale et observationnelle ont été inclus $N = 48$ patients présentant un diagnostic de TE, de PV ou de MFP selon la classification de l'OMS (itération de 2016) [14] avec pour mutation motrice du NMP $JAK2^{V617F}$, $CALR^m$ ou MPL^m. Ces patients ont été inclus et suivis pendant au moins trois mois après le début du traitement par peg-IFNα. Les traitements antérieurs avec d'autres médicaments cytoréducteurs n'étaient pas exclus.
Nous avons ainsi un ensemble de N patients noté $\mathcal{P} = \{1, \cdots, N\}$. Le jeu de données total étudié dans ce chapitre consiste alors en $\mathcal{D} = \{D^{(i)}\}_{i \in \mathcal{P}}$.

Cette cohorte comprend 21 PV (44%), 22 TE (46%) et 5 MFP (10%). 32 patients NMP ont été détectés avec la mutation $JAK2^{V617F}$, 12 avec $CALR^m$ (7 type 1 et 5 type 2), 2 avec $MPL^{W515K/R}$, un ayant à la fois $JAK2^{V617F}$ et $CALR^{del46}$ et un ayant $JAK2^{V617F}$, $CALR^{del52}$ et MPL^{S505N} (Fig. 2 A et B). Les deux derniers cas ont été classés comme des NMP $CALR^m$ parce que les VAF $CALR^m > 0.4$ étaient clairement dominantes [15] et parce que les $JAK2^{V617F}$ et $CALR^m$ s'excluaient mutuellement dans les progéniteurs.
Le séquençage nouvelle génération d'un panel de 77 gènes axé sur les cellules myéloïdes, à partir d'ADN génomique isolé des granulocytes, a permis d'identifier des mutations additionnelles dans 31% de ces cas (Fig. 2 C).
L'âge médian au début du traitement était de 53 ans (intervalle [25,71]). Le dosage médian était de 71 μg/semaine (intervalle [11-157]).
Une réponse hématologique [1] a été observée dans 78% des cas $JAK2^{V617F}$, 72% des cas $CALR^m$ et dans les 2 cas MPL^m. Des effets secondaires conduisant à l'arrêt du traitement ont été observés chez 16% et 21% des patients atteints de NMP $JAK2^{V617F}$ et $CALR^m$, respectivement.

1. La réponse hématologique a été évaluée selon les critères de l'European LeukemiaNet (ELN) pour la TE et la PV.

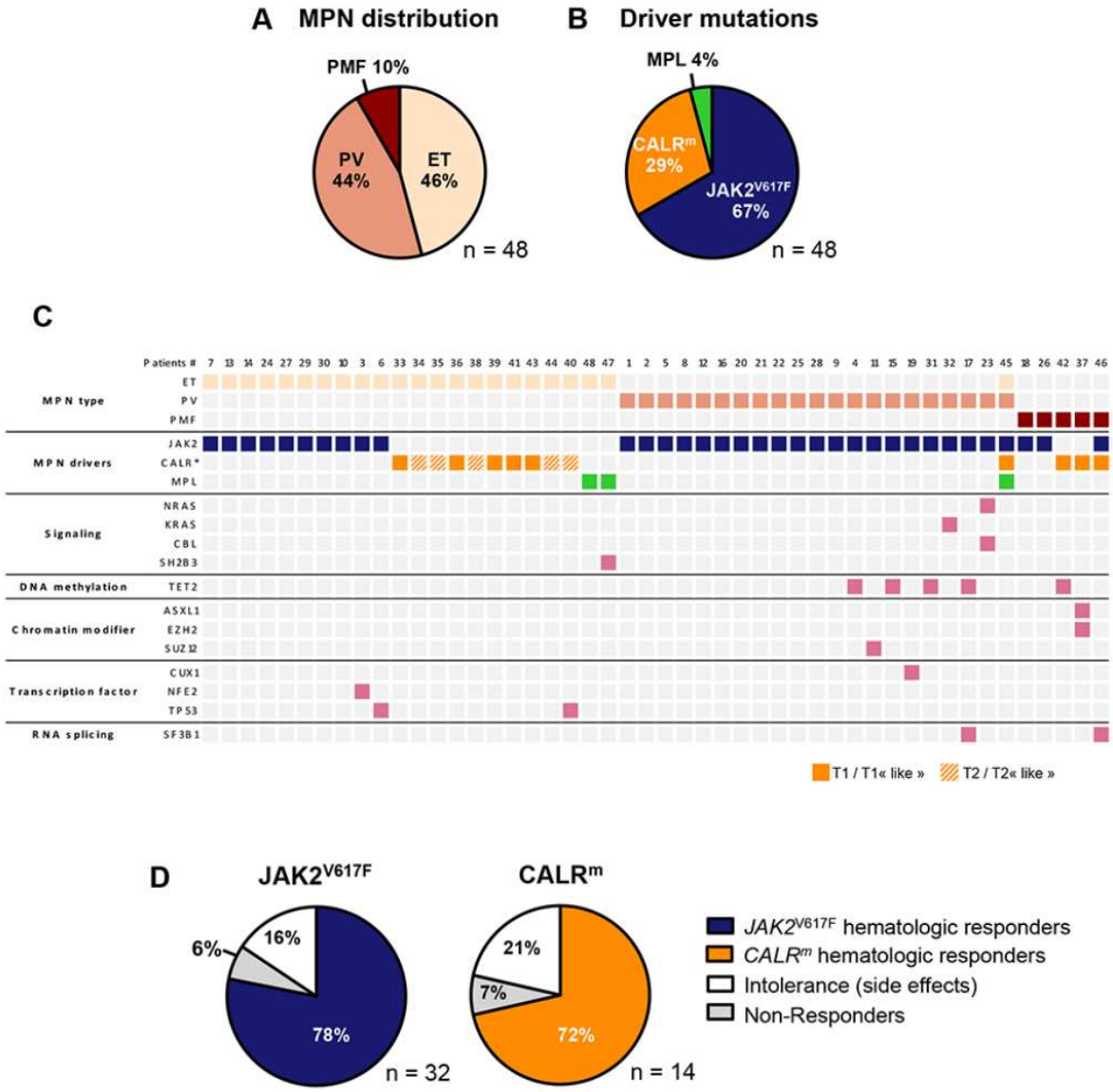


FIGURE 2 – Caractérisation de la cohorte de patients NMP traités par IFNα. (A) Distribution des NMP. (B) Distribution des mutations motrices du NMP. (C) Maladies et profil moléculaire déterminé à l'aide d'un panel myéloïde NGS de 77 gènes, sur le premier échantillon prélevé pour chacun des 48 patients de la cohorte. (D) Pourcentage de réponse hématologique, de non-réponse ou d'intolérance parmi les patients atteints de NMP $JAK2^{V617F}$ (gauche) ou $CALR^{m}$ (droite).

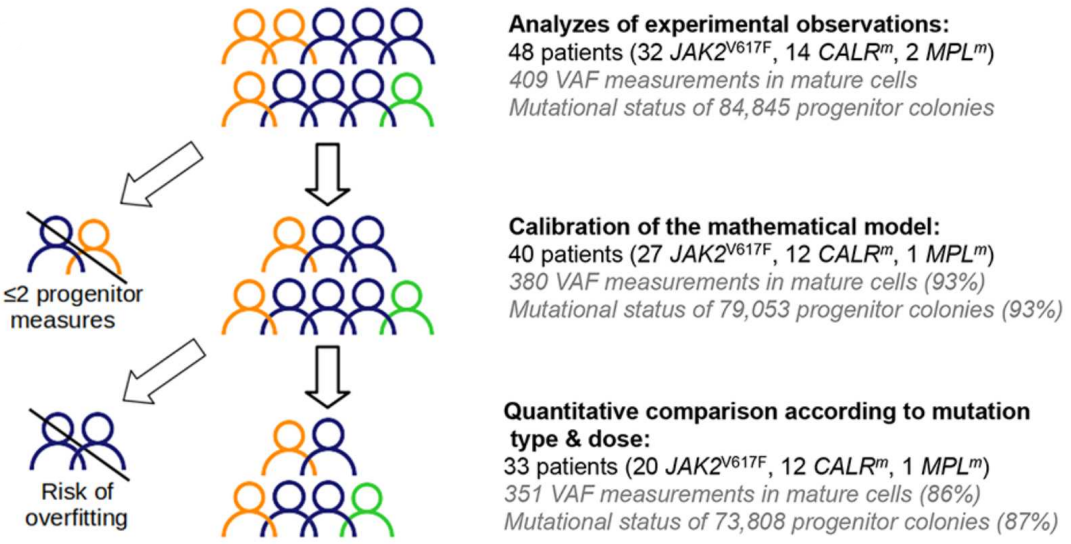


FIGURE 3 – Les observations expérimentales sont analysées à partir des cellules progénitrices et matures des 48 patients de la cohorte (§ 2.3). Nous excluons ensuite les patients qui présentent moins de deux observations (dans les progéniteurs et après le début du traitement) pour la calibration du modèle mathématique (§ 4). Ensuite, pour analyser statistiquement de manière rigoureuse la façon dont le dosage d'IFNα impacte différemment la réponse moléculaire en fonction du type de mutation et de la zygosité dans les HSC (§ 6), nous excluons les patients $JAK2^{V617F}$ qui n'ont pas plus de cinq observations (dans les progéniteurs et après le début du traitement). Aucun patient $CALR^m$ n'est exclu à ce stade, car le modèle utilisé pour ces patients est plus parcimonieux que dans le cas $JAK2^{V617F}$. Les nombres entre parenthèses correspondent aux pourcentages d'observations utilisées pour les analyses.

Nous avons trouvé 6-7% de non-répondeurs hématologiques dans les cas $JAK2^{V617F}$ et $CALR^m$ (Fig. 2 D).

Il s'agit d'une étude observationnelle : tous les patients n'ont pas été suivis à la même fréquence. Ainsi, pour certains patients, on disposera de plus de mesures que pour d'autres. L'intégralité des patients de la cohorte sera utilisée pour l'analyse des observations expérimentales (§ 2.3). Lorsqu'il s'agira de calibrer notre modèle mathématique (§ 4), partant du principe qu'avec le choix d'un cadre Bayésien hiérarchique même les patients avec peu d'observations contribuent à apporter de l'information au niveau de la population, nous exclurons seulement les patients n'ayant pas strictement plus de deux observations (au niveau des cellules immatures). Ensuite, pour les analyses statistiques ultérieures (§ 6) et les tests d'hypothèse entre sous-populations (en fonction du type de mutation, de la zygosité, du dosage) qui reposent sur l'estimation de paramètres individuels, nous devons exclure les patients pour lesquels il y a un risque d'*over-fitting*. Les patients $JAK2^{V617F}$ exclus seront ceux pour lesquels nous ne disposons pas de plus de cinq mesures de progéniteurs. Avec ce critère, nous avons principalement exclu les patients intolérants pour lesquels nous n'avions que trois ou quatre mesures, ainsi que le seul patient de notre cohorte qui avait cinq observations. Ce dernier patient était le patient #21, qui a été suivi tout au long du traitement mais pour lequel nous manquions d'informations au début de la thérapie.
Tous les patients $CALR^m$ utilisés pour la calibration du modèle seront utilisés pour les analyses statistiques. En effet, pour ces patients, étant donné que nous utiliserons un modèle plus parcimonieux que dans le cas $JAK2^{V617F}$ (voir § 3.1.1), le risque d'*over-fitting* est plus faible. Ces critères d'inclusion sont résumés sur la figure 3.

De plus, les patients peuvent avoir des HSC mutées homozygotes et hétérozygotes, mais pas dans les mêmes proportions. Lorsque nous analyserons, à partir des dynamiques inférées par le modèle, non seulement la VAF pour les HSC mais aussi la zygosité, par exemple lorsque nous

Study of zygosity for 20 *JAK2*^{V617F} patients		Patients carrying homozygous subclones	
		CF > 7% for at least 1 measure	CF < 7% for all measures
Patients carrying heterozygous subclones	CF > 7% for at least 1 measure	8	9
	CF < 7% for all measures	2	1

FIGURE 4 – Synthèse de la classification des patients $JAK2^{V617F}$ suivant si on considère qu'ils ont des sous-clones hétérozygotes et/ou homozygotes. Cette distinction n'est pas nécessaire dans le cas $CALR^m$ car seulement un faible nombre de patients (2 sur 12) présentaient des sous-clones homozygotes.

comparerons l'effet de l'IFNα sur les cellules homozygotes par rapport aux cellules hétérozygotes, il sera nécessaire d'exclure les patients dont les sous-clones présentent une CF (fraction clonale) trop faible. La plupart des patients $CALR^m$ dans notre cohorte n'ont que des cellules mutées hétérozygotes. Pour les patients $JAK2^{V617F}$, nous avons défini deux sous-groupes, l'un composé de patients présentant une proportion suffisante de sous-clones homozygotes, le second composé de patients présentant une proportion suffisante de sous-clones hétérozygotes. L'intersection entre les deux sous-groupes n'étant pas vide. Un patient $JAK2^{V617F}$ sera considéré comme ayant des sous-clones homozygotes (respectivement hétérozygotes) lorsque $>7\%$ de progéniteurs homozygotes (respectivement hétérozygotes) seront identifiés dans au moins un des échantillons collectés. En utilisant ce critère, 10 patients $JAK2^{V617F}$ sur 20 sont considérés comme ayant des sous-clones homozygotes et 17 sur 20 des sous-clones hétérozygotes. Pour être plus précis, 8 patients sur 20 appartiennent au sous-groupe $JAK2^{V617F}$ des patients porteurs de sous-clones homozygotes et hétérozygotes, 9 sur 20 au groupe de patients porteurs (uniquement) de sous-clones hétérozygotes et 2 sur 20 au groupe de patients porteurs (uniquement) de sous-clones homozygotes. Un seul patient n'appartient à aucun de ces deux groupes puisque sa CF parmi les cellules immatures est inférieure à 7% tant pour les cellules homozygotes qu'hétérozygotes. Ce patient ne sera considéré que dans les analyses portant sur la VAF, mais pas celles portant sur la CF. Ces informations sont résumées sur la figure 4.

2.3 Effet de l'IFNα sur les progéniteurs

Avant d'utiliser un modèle mathématique pour inférer la dynamique au niveau des HSC qui - rappelons-le - ne sont pas observables, commençons par analyser les données expérimentales brutes.

Nous pouvons étudier la réponse au traitement pour chaque patient en regardant l'évolution de la VAF dans les compartiments immatures (progéniteurs CD34$^+$) et matures (granulocytes), comme présenté sur la figure 5. Alors que le VAF des patients $CALR^m$ reste globalement stable pendant le traitement, la VAF des patients $JAK2^{V617F}$ a tendance à diminuer au cours du traitement à la fois dans les cellules hématopoïétiques matures et les progéniteurs. En particulier, on observe une différence significative entre les patients $CALR^m$ et $JAK2^{V617F}$ après 600 jours de traitement suggérant un effet différencié de l'IFNα selon le type de mutation motrice considéré.
Nous avons alors cherché à savoir si le dosage d'IFNα pouvait impacter la réponse dans le cas des progéniteurs de patients $CALR^m$ (Fig. 6.A). Pour cela, nous avons séparé les patients $CALR^m$ en deux sous-groupes, suivant si le dosage du patient était en-dessous (Low-Dose - LD) ou au-dessus (High Dose - HD) de la valeur médiane calculée sur le groupe. Nous avons fait de même dans le cas des patients $JAK2^{V617F}$ (Fig. 6.B). Alors que nous n'avons pas détecté d'effet significatif du dosage chez les patients $CALR^m$, la VAF mesurée dans les progéniteurs de patients $JAK2^{V617F}$

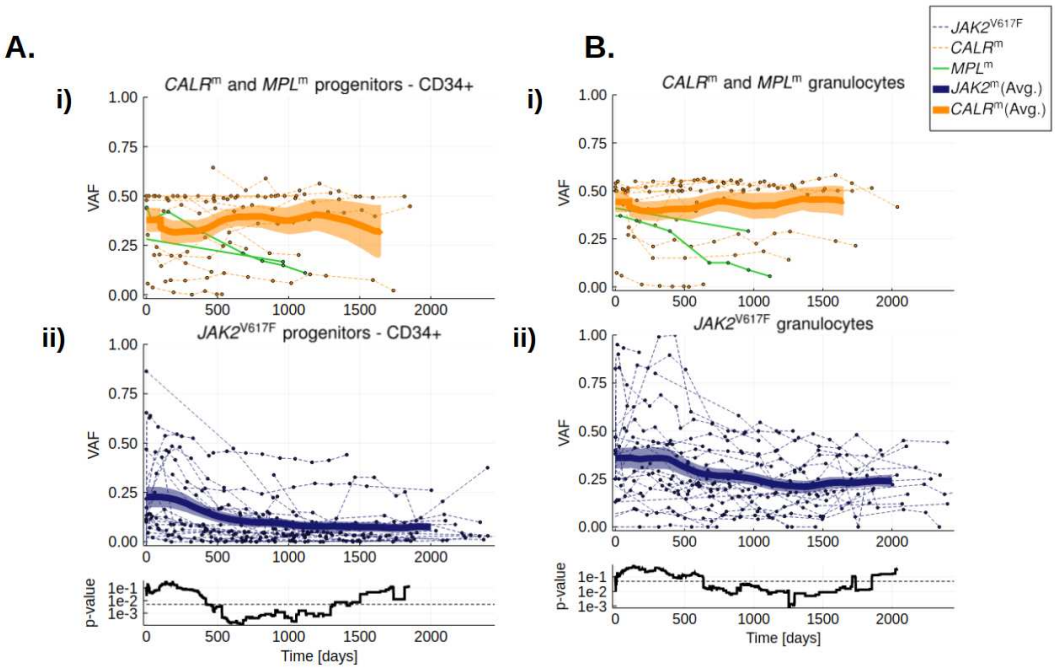


FIGURE 5 – Effet de l'IFNα dans différents compartiments hématopoïétiques au cours de l'étude clinique réalisée sur les 48 patients. On représente l'évolution de la VAF dans les progéniteurs CD34$^+$ (A) et dans les granulocytes (B) pour les patients $CALR^m$ et MPL^m (i) et les patients $JAK2^{V617F}$ (ii). Sur chaque figure, une ligne correspond à un patient. $t = 0$ correspond au début du traitement. Les lignes en trait fort correspondent aux moyennes flottantes (sur $\pm$100 jours), avec de part et d'autres l'erreur standard de la moyenne. À chaque instant, on effectue un test U de Mann-Whitney (ce test statistique basé sur le rang a été décrit plus en détail au chapitre 3) pour tester si on peut rejeter l'hypothèse selon laquelle la distribution des mesures de VAF pour les patients $JAK2^{V617F}$ pourrait être proche de celle pour les patients $CALR^m$. On trace la p-valeur associée à ce test selon que l'on considère la VAF dans les progéniteurs (A) ou dans les granulocytes (B). Au cours des 300 premiers jours, il n'y a pas de différence significative entre la VAF des patients $CALR^m$ et $JAK2^{V617F}$. Il est ensuite intéressant de constater qu'on obtient des différences significatives à partir de 600 jours de traitement dans le compartiment des progéniteurs ($p<0.0005$), suggérant un effet différencié de l'IFNα selon le type de mutation motrice considéré. Une différence moindre entre les VAF $CALR^m$ et $JAK2^{V617F}$ est observée dans les cellules matures ($p<0.025$ après 650 jours).

diminue plus fortement pour les patients traités à fort dosage, avec une différence significative entre LD et HD partir de $t = 1,000$ jours.
Nous remarquons également que, chez certains individus atteints de NMP $JAK2^{V617F}$ traités par de fortes doses d'IFNα, il y a une augmentation de la VAF au début du traitement suivie d'une diminution importante. Nous désignerons cet effet par le terme de "courbe en cloche" et chercherons notamment, dans la construction du modèle, à être en mesure de reproduire cette caractéristique. À noter que l'augmentation de la VAF (qui, pour rappel, est une proportion) dans les progéniteurs est également associée à une augmentation de la quantité de progéniteurs et granulocytes en circulation.
Pour les patients $JAK2^{V617F}$ traités à faible dose, on observe plutôt une diminution progressive et continue de la VAF. Cela suggérerait un effet différencié de l'IFNα, dans le cas $JAK2^{V617F}$, selon le dosage.
Nous avons enfin comparé, dans le cas $JAK2^{V617F}$, l'évolution de la CF des progéniteurs mutés hétérozygotes et homozygotes (Fig. 6.C) et trouvons des différences significatives à partir de 600 jours de traitement, suggérant que l'IFNα ciblerait plus rapidement les cellules mutées homozygotes que celles mutées hétérozygotes.

L'analyse faite ci-dessus, directement sur les données expérimentales, présente quelques limites :

- L'incertitude sur les observations, notamment celle pour les progéniteurs qui est variable suivant les échantillons, n'est pas prise en compte.
- La dynamique de la réponse au traitement est extrapolée linéairement entre chaque instant où un échantillon de sang de patient a été analysé.
- Pour certains patients, nous n'avons pas d'observations en début de traitement, rendant impossible l'extrapolation précédente.
- Nous comparons des VAF entre patients, sans prendre en compte l'hétérogénéité de la VAF entre patients en début de traitement.
- Nous n'observons la dynamique de traitement qu'au niveau de progéniteurs immatures, et pas au niveau des cellules souches, alors que ces dernières seraient vraisemblablement la cible de l'IFNα, sans quoi il n'y aurait pas de réponse moléculaire possible.
- Il est difficile d'extrapoler une réponse à long-terme en l'absence de modèle.

Pour pallier à ces limites, et également chercher à comprendre l'effet possible de l'IFNα au niveau des cellules souches, nous allons dans la suite analyser ces données par l'intermédiaire d'un modèle dont les paramètres seront estimés à partir des observations expérimentales.
Notre modèle sera basé sur l'hypothèse de travail selon laquelle l'IFNα induirait une dynamique latente et non observée au niveau des HSC, qui induirait à son tour la dynamique observable au niveau des des cellules progénitrices puis des granulocytes [16]. Le formalisme mathématique qui nous paraît alors naturel pour l'étude de l'effet de l'IFNα est alors celui des modèles déterministes compartimentaux, parmi lesquels nous pouvons citer celui de Michor et al. [17] utilisé pour décrire l'effet du traitement à l'imatinib chez des patients atteints de leucémie myéloïde chronique.

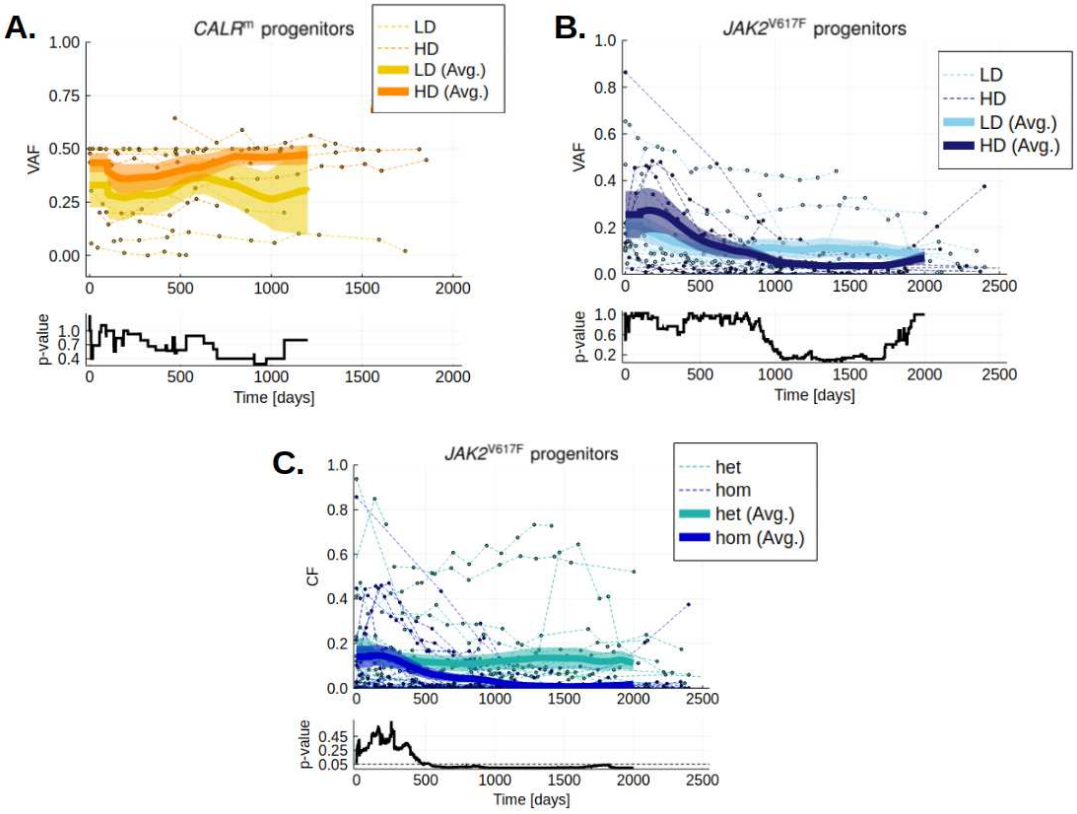


FIGURE 6 – Effet de l'IFNα sur les progéniteurs hématopoïétiques.
(A) Évolution de la VAF dans les progéniteurs de patients $CALR^m$, selon s'ils ont été traités à un fort dosage (High dosage - HD, $D > 78$ μg/semaine) ou à un faible dosage (Low Dosage - LD, $D < 78$ μg/semaine).
(B) Évolution de la VAF dans les progéniteurs de patients $JAK2^{V617F}$, selon s'ils ont été traités à un fort dosage (High dosage - HD, $D > 96.5$ μg/semaine) ou à un faible dosage (Low Dosage - LD, $D < 96.5$ μg/semaine).
(C) Évolution de la CF en progéniteurs hétérozygotes (het - en vert) et homozygotes (hom - en bleu) dans le cas des patients $JAK2^{V617F}$.
Sur chaque figure, une ligne correspond à un patient. $t = 0$ correspond au début du traitement. Les lignes en trait fort correspondent aux moyennes flottantes (sur $\pm$100 jours), avec de part et d'autres l'erreur standard de la moyenne. La p-valeur est celle d'un test U de Mann-Whitney, testant à chaque instant si on peut rejeter l'hypothèse selon laquelle les VAF (ou CF) se distribueraient de la même façon suivant les deux sous-groupes considérés. Pour la comparaison suivant la zygosité, dans le cas $JAK2^{V617F}$ (C), il est intéressant d'observer qu'en début de traitement, il n'y a pas de différence significative, puis que $p < 0.03$ après 600 jours de traitement et qu'on obtient une différence très significative après 1,000 jours de traitement (test U de Mann et Whitney, $p<0.003$), suggérant un effet différencié de l'IFNα selon la zygosité.

3 Modèle

3.1 Modèle compartimental

3.1.1 Un seul compartiment pour les HSC

Au chapitre précédent, nous étudions l'expansion d'un clone muté hétérozygote ($JAK2^{V617F}$ ou $CALR^m$) à partir de la première cellule souche ayant acquis la mutation motrice du NMP. Dans ce modèle, présenté sur la figure 7.A, des cellules souches mutées hétérozygotes pouvaient se diviser à un taux α, en faisant une division symétrique, asymétrique ou différenciée avec des probabilités respectives valant p_2, p_1 et p_0, et produisant respectivement 0, 1 ou 2 cellules filles progénitrices. On introduisait alors le paramètre $\Delta = p_2 - p_0$ qui modélisait la capacité du clone mutant à envahir le *pool* de cellules souches lorsque $\Delta > 0$. La dynamique d'expansion des HSC mutées était alors modélisée par un processus stochastique, à savoir une CTMC (Continuous Time Markov Chain). Le choix d'un modèle stochastique était justifié par le fait qu'on étudiait, lors du développement initial du NMP, un nombre faible de cellules.
Dans ce chapitre, nous souhaitons étudier l'effet de l'IFNα sur la dynamique des cellules mutées, en particulier les HSC. Notre modèle reposera sur l'hypothèse que l'IFNα induira une dynamique latente et non observée des HSC qui se répercute sur les cellules progénitrices et les granulocytes [16].
Juste avant traitement, les patients, puisqu'ils ont développé les symptômes de la maladie, présentent un nombre important de cellules mutées, ce qui justifie le choix d'un modèle déterministe. On peut tout d'abord considérer l'équivalent déterministe du modèle étudié au chapitre 5. Considérons pour le moment la dynamique de populations de cellules d'un seul génotype, par exemple des cellules mutées $CALR^m$ hétérozygotes. On introduit alors trois compartiments, c'est-à-dire trois types de cellules pour les cellules mutées : les HSC, les progéniteurs (ou cellules immatures) et les cellules matures (en l'occurrence, les granulocytes). Les cellules matures sont entièrement différenciées ; elles ne subissent plus de divisions cellulaires et meurent à un taux δ_m. Ce compartiment mature correspondra aux mesures expérimentales de la VAF parmi les granulocytes (après normalisation, voir § 3.2).
On fait l'hypothèse qu'on peut négliger l'hétérogénéité entre progéniteurs $CD34^+$, ce qui, comme nous l'avons montré au chapitre 2, est une hypothèse simplificatrice. Le compartiment des progéniteurs dans notre modèle regroupe plusieurs progéniteurs distincts qui correspondent expérimentalement aux cellules triées selon les marqueurs de surface : $CD90^-CD34^+CD38^+$, $CD90^-CD34^+CD38^-$ et $CD90^+CD34^+CD38^-$. Ces progéniteurs, considérés être issus de HSC, sont supposés avoir subi en moyenne plusieurs divisions (modélisées par le paramètre κ_i) et quitter leur compartiment immature au taux de différenciation δ_i. Pour tenir compte des cellules qui subissent plusieurs cycles de prolifération et de différenciation avant de devenir complètement différenciées, nous avons introduit le paramètre κ_m.
Ainsi, si $N_s(t)$, $N_i(t)$ et $N_m(t)$ sont respectivement les quantités (les nombres absolus) de cellules souches, immatures et matures mutées hétérozygotes, avec le modèle décrit précédemment, leur dynamique est décrite par le système d'équations différentielles ordinaires (ODE) suivant :

$$\begin{cases} \frac{dN_s(t)}{dt} &= \alpha\Delta N_s(t) \\ \frac{dN_i(t)}{dt} &= \alpha(1-\Delta)\kappa_i N_s(t) - \delta_i N_i(t) \\ \frac{dN_m(t)}{dt} &= \delta_i\kappa_m N_i(t) - \delta_m N_m(t) \end{cases} \quad (1)$$

Ce système est linéaire ; on peut en calculer une solution explicite au temps t, en fonction des conditions initiales. Ce modèle est proche de celui de Michor et al. [17]. On suppose que les équations restent valables avant et après traitement mais, suivant l'idée de Michor et al., on considérera que le traitement agira en modifiant la valeur de certains paramètres du modèle à partir du temps $t = 0$, faisant basculer le système d'un "équilibre" vers un autre. Nous estimions par exemple $\Delta_{het} = 0.026$ dans le cas $CALR^m$; ce qui donnerait avant traitement une croissance exponentielle de la population de cellules souches mutées hétérozygotes. En faisant l'hypothèse que l'IFNα agit sur le paramètre Δ, dans le cas d'un patient répondant positivement au traitement, on obtiendrait $\Delta^*_{het} < 0$ (le symbole * indique que le paramètre est modifié par l'action de

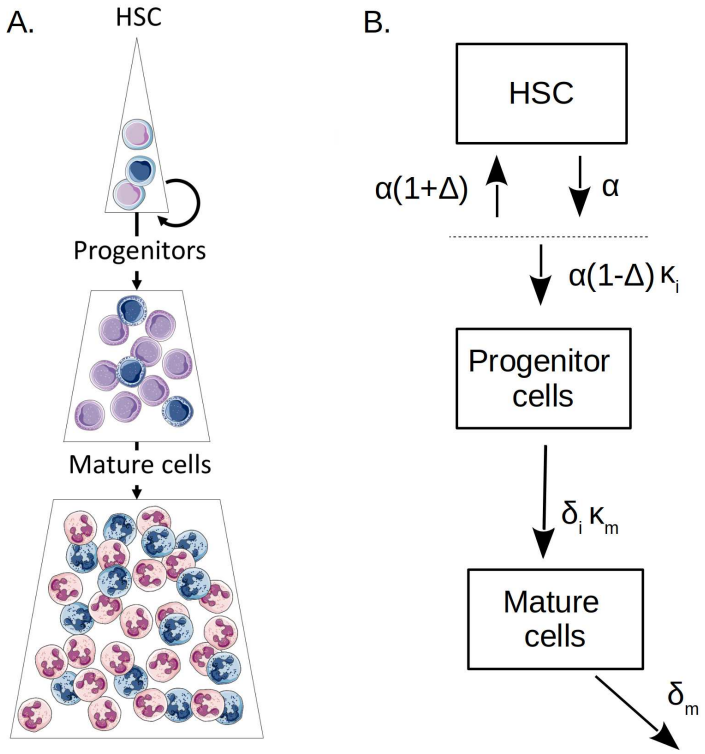


FIGURE 7 – Construction d'un modèle déterministe à trois compartiments (B) à partir du modèle stochastique présenté au chapitre 5 (A).

l'IFNα), c'est-à-dire qu'on basculerait d'une croissance à une décroissance exponentielle jusqu'à tendre vers une disparition des cellules mutées hétérozygotes.
À noter que, lorsque la quantité N_s devient faible (voire < 1 ce qui n'aurait pas de sens biologique), le formalise déterministe n'est plus adapté et il faudrait rebasculer vers un modèle stochastique, par exemple celui étudié au chapitre précédent.

Ci-dessus, nous avons décrit la dynamique de populations de cellules mutées hétérozygotes. Dans les faits, nous considérerons jusqu'à trois populations de cellules : les wild-type, les mutées hétérozygotes et les mutées homozygotes. Nous supposerons ces trois populations indépendantes les unes des autres, et considérerons que la dynamique de chacune obéit à un même système d'équations, mais avec des valeurs de paramètres éventuellement différentes. Nous y reviendrons au paragraphe 3.2.

3.1.2 Distinction entre les HSC actives et quiescentes

Le modèle présenté au paragraphe précédent, avec un seul compartiment pour décrire la dynamique des HSC, est adapté dans le cas de cellules mutées $CALR^m$, mais pas dans le cas $JAK2^{V617F}$.
En effet, dans le cas $CALR^m$, les HSC mutées sont supposées avoir un avantage prolifératif plus important au niveau souche que les $JAK2^{V617F}$ (comme nous l'avions montré au chapitre précédent) ce qui suggère qu'elles pourraient rester dans un état actif [18, 6]. Le modèle en l'état, avec trois compartiments, est alors en accord avec l'état de nos connaissances sur la biologie des cellules mutées $CALR^m$, et n'a pas de raison d'être complexifié. En particulier, ce modèle permettra de reproduire correctement les dynamiques observées sous traitement.
Par contre, dans le cas $JAK2^{V617F}$, il échouera à reproduire l'effet courbe en cloche évoqué au § 2.3. Les cellules $JAK2^{V617F}$ sont supposées avoir un moindre avantage que les $CALR^m$ au niveau souche, et avoir un comportement - en l'absence de traitement - qui se rapprocherait des WT. Pour modéliser l'hématopoïèse et en particulier la dynamique de cellules souches, il

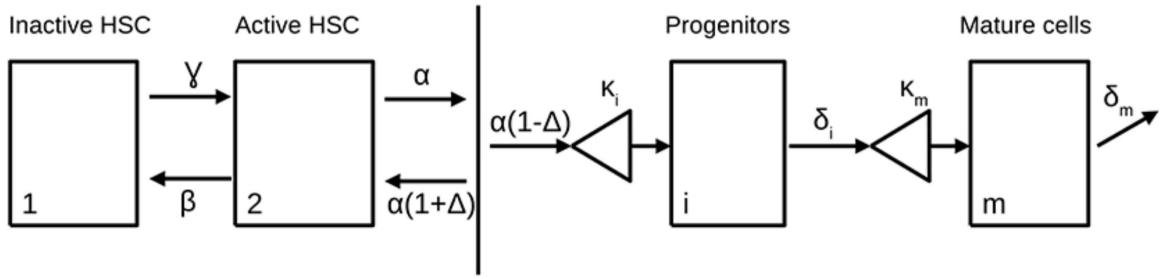


FIGURE 8 – Conception du modèle mathématique à quatre compartiments. Les cellules matures et entièrement différenciées ne se divisent plus et meurent à un taux δ_m. Nous avons modélisé les cellules progénitrices comme provenant d'une HSC active qui se divise et subit plusieurs divisions (modélisées par le paramètre κ_i). Les progéniteurs sortent de leur compartiment au taux de différenciation δ_i et prolifèrent (modélisé par le paramètre κ_m) avant d'entrer dans le compartiment mature. Nous avons également introduit deux compartiments de cellules souches selon que la HSC est considérée comme active ou inactive (quiescente), les paramètres γ et β modélisent les échanges entre ces deux compartiments. Nous avons supposé que les HSC actives pouvaient être recrutées pour se différencier à un taux α afin de contribuer à l'hématopoïèse. Le paramètre Δ modélise la propension du *pool* de cellules souches à s'épuiser (si $\Delta < 0$) ou à s'étendre (si $\Delta > 0$).

est courant d'introduire la notion de cellules quiescentes (ou dormantes) et actives. Citons par exemple Catlin et al. qui étudient un modèle à deux compartiments, consistant en une réserve de cellules souches et un compartiment de cellules contribuant à l'hématopoïèse [19], ou encore Adimy et al. qui étudient un modèle de dynamique des HSC incluant une phase proliférative et une phase quiescente [20].
Nous allons alors séparer le compartiment des HSC précédent en deux compartiments fonctionnels : les HSC sont soit quiescentes (ou inactives), soit actives et peuvent donc contribuer directement à l'hématopoïèse. Une cellule souche peut passer de quiescente à active et vice versa. Les paramètres γ et β modélisent ces échanges entre compartiments. Cette distinction entre HSC actives et inactives sera valable pour la population de cellules WT et et de cellules mutées $JAK2^{V617F}$ ou MPL^m. Comme nous l'avons dit précédemment, les HSC mutées $CALR^m$ seront supposées rester actives. Dans ce cas, avec $N_1(t)$, $N_2(t)$, $N_i(t)$ et $N_m(t)$ qui sont respectivement les quantités de cellules souches quiescente, souches actives, immatures et matures, par exemple mutées $JAK2^{V617F}$ hétérozygotes, leur dynamique est décrite par le système d'ODE suivant :

$$\begin{cases} \frac{dN_1(t)}{dt} &= -\gamma N_1(t) + \beta N_2(t) \\ \frac{dN_2(t)}{dt} &= \gamma N_1(t) + (\alpha\Delta - \beta) N_2(t) \\ \frac{dN_i(t)}{dt} &= \alpha(1-\Delta)\kappa_i N_2(t) - \delta_i N_i(t) \\ \frac{dN_m(t)}{dt} &= \delta_i \kappa_m N_i(t) - \delta_m N_m(t) \end{cases} \tag{2}$$

Ce système - schématisé sur la figure 8 - est linéaire ; nous présentons sa solution analytique en annexe A à ce chapitre [2].

3.2 Normalisation, conditions initiales et effet de l'IFNα

Les données à notre disposition pour l'estimation des paramètres du modèle ne fournissent pas d'informations sur les valeurs absolues des quantités de cellules wt, het ou hom séparément (qui sont les solutions des systèmes d'ODE (2)), mais plutôt sur leurs proportions relatives (pour plus de détails, voir § 4.1). Pour en tenir compte, nous considérons comme sorties de notre modèle non plus le nombre de cellules mais les proportions de cellules hétérozygotes immatures (et de

2. disponible au lien : https ://gitlab-research.centralesupelec.fr/2012hermangeg/supplementary-material-phd

même pour les cellules homozygotes) :

$$z_{het}(t) = \frac{N_{i,het}(t)}{N_{i,wt}(t) + N_{i,het}(t) + N_{i,hom}(t)}$$

ainsi que la VAF parmi les granulocytes :

$$y(t) = \frac{0.5N_{m,het}(t) + N_{m,hom}(t)}{N_{m,wt}(t) + N_{m,het}(t) + N_{m,hom}(t)}$$

Jusqu'à présent, par souci de clarté, nous avions omis de mentionner en indice la zigosité de la cellule. Nous la rajoutons maintenant.

Suivant l'idée de Michor et al. [17], on considère que l'IFNα agit en modifiant les valeurs de certains paramètres du modèle. Le temps $t = 0$ correspond au début du traitement. Avant cet instant, les équations (2) sont toujours valides. Néanmoins, on simplifie en considérant le système dans un état quasi stationnaire, c'est-à-dire que $\Delta \approx 0$ (quelque soit la zigosité : wt, het ou hom). Dans le cas de cellules WT, supposer $\Delta = 0$ revient à supposer que des conditions homéostatiques sont satisfaites permettant d'assurer un nombre constant de cellules WT dans l'organisme. Ainsi, sous l'hypothèse qu'on peut négliger les variations des quantités de cellules WT sur un temps court, on trouve, en annulant les dérivées dans le système (2), les conditions initiales suivantes :

$$\begin{aligned}
N_{1,wt}(0) &= \frac{\beta}{\beta+\gamma} N_{HSC} \\
N_{2,wt}(0) &= \frac{\gamma}{\beta+\gamma} N_{HSC} \\
N_{i,wt}(0) &= \frac{\kappa_i \alpha}{\delta_i} N_{2,wt}(0) \\
N_{m,wt}(0) &= \frac{\kappa_m \delta_i}{\delta_m} N_{i,wt}(0)
\end{aligned}$$

avec N_{HSC}, le nombre total de HSC WT, supposé constant.
Dans le cas des cellules mutées, l'hypothèse de quasi-stationnarité est simplificatrice puisque les cellules mutées ont tendance à envahir le *pool* de HSC. Néanmoins, si on considère l'envahissement lent, nous pouvons faire l'approximation que $\Delta_{het} \approx \Delta_{hom} \approx 0^+$. Cette hypothèse se justifie notamment au vu des valeurs estimées au chapitre précédent pour les HSC mutées hétérozygotes, à savoir 0.017 et 0.026 pour les HSC $JAK2^{V617F}$ et $CALR^m$ respectivement.
Nous introduisons :

$$\eta_{het} = \frac{N_{1,het}(0) + N_{2,het}(0)}{N_{HSC}}$$

et :

$$\chi_{het} = \frac{N_{2,het}(0)}{N_{1,het}(0) + N_{2,het}(0)}$$

permettant d'exprimer les conditions initiales dans les compartiments hétérozygotes :

$$\begin{aligned}
N_{1,het}(0) &= \eta_{het}\left(1 - \chi_{het}\right) N_{HSC} \\
N_{2,het}(0) &= \chi_{het}\eta_{het} N_{HSC} \\
N_{i,het}(0) &= \frac{\kappa_{i,het}\alpha_{het}}{\delta_{i,het}} N_{2,het}(0) \\
N_{m,het}(0) &= \frac{\kappa_{m,het}\delta_{i,\ het}}{\delta_{m,het}} N_{i,het}(0)
\end{aligned}$$

Et de même pour les cellules homozygotes.

Étant donné les solutions du système d'ODE présentées en annexe A, et la façon dont nous les normalisons pour obtenir des VAF ou CF, nous aboutissons à quelques simplifications. Le

paramètre N_{HSC} n'est plus pertinent : quelle que soit sa valeur, il ne changera pas la sortie finale du modèle (en considérant les ratios z_{het}, z_{hom}, et y). En introduisant $k_{i,het}$ et $k_{i,hom}$ tels que $k_{i,het} = \kappa_{i,het}/\kappa_{i,wt}$ et $k_{i,hom} = \kappa_{i,hom}/\kappa_{i,wt}$, les paramètres $\kappa_{i,het}$ et $\kappa_{i,wt}$ n'ont plus à être estimés. Il en va de même pour $\kappa_{m,wt}$ et $\kappa_{m,het}$.
À partir de $t = 0$, les patients sont sous traitement. On suppose que l'IFNα modifie les valeurs de certains paramètres, potentiellement de différentes manières selon la zygosité des cellules. En termes de notation, l'exposant * est ajouté aux paramètres impactés par le traitement. Nous définissons $k^*_{i,wt}$ tel que $\kappa^*_{i,wt} = k^*_{i,wt}\kappa_{i,wt}$ et de même pour $k^*_{i,het}$, $k^*_{i,hom}$,$k^*_{m,wt}$, $k^*_{m,het}$ et $k^*_{m,hom}$. Puisque l'on considérera des ratios de cellules et non les nombres absolus, $k^*_{i,wt}$ et $k^*_{m,wt}$ se simplifient et n'ont pas à être estimés.
L'intégralité des paramètres impliqués dans notre modèle est présenté dans le tableau 9. Au paragraphe suivant, nous présentons un ensemble de simplifications permettant de réduire le nombre de paramètres à estimer afin d'aboutir à un modèle identifiable.

3.3 Paramètres et simplifications

Le modèle présenté plus haut comporte potentiellement de nombreux paramètres spécifiques à un patient donné. Un nombre excessif de paramètres risque d'entraîner un risque d'*over-fitting* et de non-identifiabilité du modèle. Pour résoudre ce problème, nous avons fait des hypothèses supplémentaires concernant les paramètres et leurs valeurs.
Comme il est particulièrement difficile de déduire les valeurs des paramètres associés aux compartiments des HSC alors qu'aucune observation pour ces cellules n'est disponible, nous avons choisi de supposer que le paramètre α était indépendant du génotype et de la zygosité (wt, het ou hom) et non affecté par l'IFNα. Nous avons choisi $\alpha_{WT} = 1/30$ [jours^{-1}], $\gamma_{WT} = 1/300$ [jours^{-1}] (c'est-à-dire un ordre de grandeur inférieur à α), et $\chi_{WT} = 0.1$ (ce qui signifie que les HSC actives WT représenteraient 10% du *pool* de HSC WT). À notre connaissance, pour l'hématopoïèse humaine *in vivo*, il n'existe pas de valeurs exactes pour ces paramètres que l'on puisse trouver dans la littérature. Les valeurs de ces trois paramètres, décrivant la dynamique des HSC WT, ont été fixées à des ordres de grandeur qui nous semblaient raisonnables.
Il aurait pu être judicieux, plutôt que d'imposer aux paramètres précédents d'être constants, de choisir des *priors* centrés sur ces valeurs, comme nous avions fait par exemple pour α_{het} au chapitre précédent. Néanmoins, dans l'étape d'estimation de paramètres, cela aurait rallongé les temps de calculs qui sont déjà importants avec la paramétrisation choisie. De plus, cela aurait remis en cause l'hypothèse selon laquelle la dynamique des cellules WT est la même pour tous les patients (voir la remarque plus bas).
Un point important à souligner est que la construction du modèle (et le choix des hypothèses simplificatrices) s'est effectuée par étapes successives. Alors qu'il serait élégant rétrospectivement de reconstruire une démarche rigoureuse ayant permis d'aboutir à la version finale du modèle présenté dans ce chapitre, en pratique, le travail de modélisation a consisté en une succession de propositions de modèles et d'hypothèses (non nécessairement consignées), constamment retravaillées suite aux échanges avec les équipes d'Isabelle Plo à Gustave Roussy, conduisant à de nombreuses remises en question des hypothèses et choix de modélisation.
Nous présentons ainsi en annexe B à ce chapitre une version intermédiaire de notre modèle, qui avait été présentée au congrès ISMCO 2020 (*International Symposium on Mathematical and Computational Oncology*). Dans cette version, la structure du modèle (c'est-à-dire le système d'ODE (2)) est la même que le modèle présenté dans ce chapitre (modèle à deux compartiments souches). Les hypothèses simplificatrices différaient néanmoins de celles choisies actuellement, et surtout, nous avions un nombre plus important de paramètres à estimer (10 au lieu de 7 actuellement). En particulier, nous cherchions à estimer plus de paramètres décrivant le comportement des cellules souches, ce qui - à cause du manque d'observations au niveau des HSC - conduisait finalement à un modèle non identifiable : les dynamiques inférées pour le patient étudié - en l'occurrence le patient #32 - étaient très proches de cellules obtenues avec nos hypothèses simplificatrices actuelles.
Pour lever ce problème de non identifiabilité, nous avons alors imposé aux paramètres α_{WT}, γ_{WT}

et χ_{WT} d'être constants. Les valeurs ont été choisies dans des ordres de grandeurs proches de ceux estimés[3] dans notre première version (uniquement sur un patient, le patient #32, qui présentait le phénomène de courbe en cloche), et donc pas *a priori*. Les choix des valeurs pour α_{WT}, γ_{WT} et χ_{WT} peuvent alors ne pas correspondre aux valeurs biologiques réelles (pour lesquelles il n'existe à notre connaissance pas de valeurs connues). Derrière le choix d'imposer à α_{WT}, γ_{WT} et χ_{WT} d'être constants, il y a l'hypothèse implicite selon laquelle la dynamique des cellules souches WT serait la même pour chaque individu. Ce choix nous permet alors de quantifier l'effet du traitement sur les patients, relativement les uns aux autres.

Pour continuer sur nos hypothèses, étant donné que l'exposition chronique, contrairement à l'exposition aiguë à l'IFNα, induit une prolifération transitoire des HSC de souris suivie d'un retour rapide à la quiescence [21, 22, 23], nous avons supposé que l'IFNα influence à peine les paramètres des HSC humaines WT, c'est-à-dire $\gamma^*_{wt} = \gamma_{wt}$ et $\beta^*_{wt} = \beta_{wt}$. Nous avons considéré qu'avant le début du traitement, les transitions entre les compartiments quiescent et actif pour les cellules mutées hétérozygotes et homozygotes étaient similaires au cas WT, c'est-à-dire $\gamma_{het} = \gamma_{hom} = \gamma_{wt}$ et $\beta_{hom} = \beta_{het} = \beta_{wt}$.

Dans ce chapitre, une de nos hypothèses de travail est celle d'un effet de l'IFNα sur la sortie de quiescence des cellules mutées hétérozygotes et homozygotes. Concernant les cellules hétérozygotes (et de même pour les homozygotes), on pourrait considérer que le traitement affecte γ_{het}, β_{het}, ou les deux paramètres. Ce dernier choix entraînerait un problème de non-identifiabilité. Nous avons choisi de considérer que le traitement affecte γ_{het} ; nous estimons alors γ^*_{het} et fixons $\beta^*_{het} = \beta_{wt}$. Pour les patients $CALR^m$, les hypothèses précédentes concernant γ, β et χ ne sont pas nécessaires car nous ne considérons pas de HSC mutées $CALR^m$ inactives.

Pour les progéniteurs hématopoïétiques et les cellules matures, nous avons plus de connaissances biologiques pouvant nous aider à fixer certaines valeurs. Tout d'abord, nous avons considéré que δ_i et δ_m ne dépendaient pas de la zygosité ou du traitement. Nous avons fixé $\delta_m = 1$ [jours^{-1}] (sachant que les granulocytes WT ont une durée de vie de l'ordre de grandeur du jour) et $\delta_i = 1/6$ [jours^{-1}] (supposant que l'ordre de grandeur de la durée de vie des progéniteurs devrait être de quelques jours).
Nous avons supposé que les cellules progénitrices prolifèrent de la même manière quel que soit le type de cellule, c'est-à-dire $\kappa^*_{i,het} = \kappa^*_{i,hom} = \kappa^*_{i,wt}$ et $\kappa_{i,het} = \kappa_{i,hom} = \kappa_{i,wt}$. Cette hypothèse permet de déduire les conditions initiales η_{het} et η_{hom} à partir des données des cellules progénitrices.
Enfin, nous avons supposé que les cellules mutées peuvent proliférer plus rapidement que les cellules WT aux derniers stades de l'hématopoïèse, mais n'avons considéré aucune différence entre les cellules hétérozygotes ou homozygotes (puisque les données dont nous disposions pour les granulocytes ne faisaient pas la distinction entre la zygosité) : ainsi, nous prenons $k_{m,het} = k_{m,hom}$. Finalement, nous avons supposé que le traitement affecte également toutes les cellules matures : $k^*_{m,hom} = k^*_{m,wt} = k^*_{m,het}$.

Pour chaque patient, nous aboutissons alors à cinq paramètres (trois pour les patients $CALR^m$) et deux conditions initiales au niveau des cellules souches (pour les clones homozygotes et hétérozygotes) à estimer à partir des données. Ces paramètres sont Δ^*_{het} et Δ^*_{hom} liés à la différenciation des HSC mutées hétérozygotes et homozygotes sous IFNα, et γ^*_{het} et γ^*_{hom} pour modéliser l'effet du traitement sur la sortie de quiescence des cellules mutées. Ces deux derniers paramètres n'apparaissaient pas dans le modèle simplifié pour $CALR^m$. Nous estimons également $k_{m,het} = k_{m,hom}$ qui modélise l'avantage prolifératif des cellules mutées aux derniers stades de l'hématopoïèse (paramètre utilisé au chapitre précédent). Enfin, nous devons estimer les conditions initiales inconnues η_{het} et η_{hom} qui correspondent aux proportions de HSC mutées hétérozygotes et ho-

3. Dans l'article présent à l'ISMCO'2020, nous avions estimé pour le patient #32 $\gamma_{WT} = 0.0051$, $\chi_{WT} = 0.050$ et $\alpha^*_{wt} = 0.027$ (voir Annexe B).

mozygotes (par rapport à l'ensemble des HSC WT).
Le tableau de la figure 9 résume les paramètres du modèle et nos hypothèses.

3.4 Réponse moléculaire

Pour chaque patient, une fois que nous aurons estimé la valeur des paramètres du modèle, ce dernier étant déterministe, nous aurons accès à toute l'information sur la dynamique du traitement, c'est-à-dire l'évolution des fractions clonales au sein des compartiments souches, immatures et matures, à tout instant. A partir de ces quantités, nous pouvons en calculer de nouvelles qui nous permettrons de caractériser la réponse moléculaire (au niveau des HSC) pour chaque patient, et d'effectuer différentes comparaisons entre individus.

La première quantité que nous introduisons est le facteur de réponse moléculaire R (R-factor, ou facteur de rémission). Il est défini comme le rapport de la fraction clonale (ou de la VAF) inférée pour les HSC à un moment donné par rapport à sa valeur initiale. Pour les HSC mutées $JAK2^{V617F}$ hétérozygotes, le facteur de réponse est :

$$R_{het}(t) = \frac{N_{1,het}(t) + N_{2,het}(t)}{N_{1,het}(0) + N_{2,het}(0)} \tag{3}$$

De même pour les cellules homozygotes.
La réponse globale sera calculée sur la base de la VAF :

$$R(t) = \frac{0.5\,[N_{1,het}(t) + N_{2,het}(t)] + [N_{1,hom}(t) + N_{2,hom}(t)]}{0.5\,[N_{1,het}(0) + N_{2,het}(0)] + [N_{1,hom}(0) + N_{2,hom}(0)]} \tag{4}$$

En particulier, on étudiera la réponse à long terme $t = 3,000$ jours. Nous noterons $R := R(t = 3000)$. Ces R-factors seront calculés par propagation des incertitudes, des paramètres du modèle aux variables décrivant la dynamique des populations de cellules au cours du temps. Ils permettront de quantifier l'intensité de la réponse. $R > 1$ indique qu'il n'y a pas de réponse ou une réponse défavorable, car la VAF parmi les HSC est plus élevée à $t = 3,000$ jours qu'initialement. Si $R < 1$, il y a une réponse au traitement. Plus cette valeur est faible, meilleure est la réponse. Si R atteint approximativement zéro, cela signifie qu'il y a une rémission complète (CMR - Complete Molecular Response).

Ce facteur de réponse pourra être calculé non seulement au temps $t = 3,000$ jours, mais aussi en fonction du temps. Il vaut 1 à $t = 0$ puis évolue dans le temps. Son évolution pourra être analysée. Si le R-factor passe en dessous de 0.5, on parle de rémission moléculaire partielle (PMR). L'instant où cette valeur est atteinte sera noté T_{PMR} :

$$R(T_{PMR}) = 0.5$$

La comparaison des temps de PMR entre répondeurs permettra d'avoir un aperçu de la rapidité à laquelle les patients répondent au traitement.

Ces deux quantités, le R-factor et le temps T_{PMR}, pourront être utilisés pour faire des comparaisons entre patients et, plus intéressant encore, entre groupes de patients. Ainsi, nous pourrons comparer des groupes en fonction de leurs mutations motrice des NMP, ou de leur dosage d'IFNα, ou même comparer la réponse hétérozygote à la réponse homozygote.

	Before treatment and initial conditions		Under IFNα (t≥0)	
	Parameter	Value	Parameter	Value
WT	α	1/30	α^*	$= \alpha$
	Δ	0	Δ^*	$= \Delta$
	γ	1/300	γ^*	$= \gamma$
	β	$= \gamma(1-\chi)/\chi$	β^*	$= \beta$
	κ_i	NR	$k_i^* = \kappa_i^* / \kappa_i$	NR
	δ_i	1/6	δ_i^*	$= \delta_i$
	κ_m	NR	$k_m^* = \kappa_m^* / \kappa_m$	NR
	δ_m	1	δ_m^*	$= \delta_m$
	χ	0.1		
	N_{HSC}	NR		
Het	α_{het}	$= \alpha$	α^*_{het}	$= \alpha$
	Δ_{het}	0	$\boldsymbol{\Delta^*_{het}}$	**To estimate**
	γ_{het}	NR	$\boldsymbol{\gamma^*_{het}}$	**To estimate**
	β_{het}	NR	β^*_{het}	$= \beta$
	$k_{i,het} = \kappa_{i,het} / \kappa_i$	1	$k_{i,het}^* = \kappa_{i,het}^* / \kappa_{i,het}$	$= k_i^*$
	$\delta_{i,het}$	$= \delta_i$	$\delta_{i,het}^*$	$= \delta_i$
	$\boldsymbol{k_{m,het} = \kappa_{m,het} / \kappa_m}$	**To estimate**	$k_{m,het}^* = \kappa_{m,het}^* / \kappa_{m,het}$	$= k_m^*$
	$\delta_{m,het}$	$= \delta_m$	$\delta_{m,het}^*$	$= \delta_m$
	χ_{het}	$= \chi$		
	$\boldsymbol{\eta_{het}}$	**To estimate**		
Hom	α_{hom}	$= \alpha$	α^*_{hom}	$= \alpha$
	Δ_{hom}	0	$\boldsymbol{\Delta^*_{hom}}$	**To estimate**
	γ_{hom}	NR	$\boldsymbol{\gamma^*_{hom}}$	**To estimate**
	β_{hom}	NR	β^*_{hom}	$= \beta$
	$k_{i,hom} = \kappa_{i,hom} / \kappa_i$	1	$k_{i,hom}^* = \kappa_{i,hom}^* / \kappa_{i,hom}$	$= k_i^*$
	$\delta_{i,hom}$	$= \delta_i$	$\delta_{i,hom}^*$	$= \delta_i$
	$k_{m,hom} = \kappa_{m,hom} / \kappa_m$	$= k_{m,het}$	$k_{m,hom}^* = \kappa_{m,hom}^* / \kappa_{m,hom}$	$= k_m^*$
	$\delta_{m,hom}$	$= \delta_m$	$\delta_{m,hom}^*$	$= \delta_m$
	χ_{hom}	$= \chi$		
	$\boldsymbol{\eta_{hom}}$	**To estimate**		

FIGURE 9 – Tableau récapitulatif des paramètres utilisés dans notre modélisation. Le tableau montre tous les paramètres du modèle à 4 compartiments (pour les patients $JAK2^{V617F}$ et MPL^m). Nous faisons la distinction suivant la zigosité (wt, het ou hom) et suivant si on considère le temps avant ou pendant le traitement. On omet de préciser l'indice wt dans le cas des populations de cellules WT. Nous avons indiqué dans ce tableau les paramètres fixés à des valeurs constantes, les relations entre les paramètres résultant de nos hypothèses simplificatices, les paramètres qui ne sont pas pertinents (NR), c'est-à-dire qui se simplifient après normalisation, et enfin les 7 paramètres à estimer (To estimate).

4 Estimation des paramètres

4.1 Modèle statistique

Pour prendre en compte de manière rigoureuse les erreurs de mesure et les autres sources d'incertitude dans les données, nous avons défini des modèles statistiques pour l'inférence Bayésienne.

4.1.1 Cellules matures

Pour un patient i donné (nous omettrons de mentionner l'indice i dans ce paragraphe par soucis de clarté), au temps t_k, la VAF mesurée $\hat{y}_k$ parmi les cellules matures correspond à une fraction réelle inconnue y_k. On a : $\hat{y}_k, y_k \in [0,1]$. Les modèles d'incertitude classiques sont des bruits gaussiens additifs ou multiplicatifs qui ne sont pas tout à fait appropriés dans notre cas. En effet, dans le cas où $y_k = 1$ par exemple, les deux modèles de bruit autoriseraient $\hat{y}_k \neq 1$. Mais si la vraie VAF parmi les cellules matures est égale à 1, expérimentalement, nous ne nous attendrions pas à détecter des allèles non mutés dans l'échantillon de sang. Ainsi, nous choisissons de généraliser les deux modèles de bruit comme suit :

$$\hat{y}_k \mid y_k \sim \mathcal{N}\left(y_k, \sigma^2(y_k)\right)$$

avec $\sigma^2 : [0,1] \longrightarrow \mathbb{R}$. En choisissant pour la fonction σ^2 une constante, on se retrouve avec un bruit additif et en choisissant $\sigma^2(y_k) = \sigma_m^2 y_k^2$ on se retrouve avec un bruit multiplicatif. Afin d'avoir un bruit symétrique autour de $y_k = 1/2$ et qui s'annule pour $y_k = 0 = 1$, nous choisissons :

$$\sigma^2(y_k) = y_k\,(1 - y_k)\,\sigma_m^2$$

avec σ_m^2 à estimer.

4.1.2 Cellules immatures

Pour modéliser le bruit d'échantillonnage lié aux fractions clonales (CF) mutées des progéniteurs, nous supposons que nous avons tiré au hasard avec remplacement des cellules immatures du corps du patient. Cette approche est utilisée, par exemple, par Catlin et al. [19] et nous permet de modéliser l'incertitude par une distribution multinomiale. Pour un grand nombre de cellules immatures dans le corps, l'approche est presque identique à celle de Xu et al. [24] qui considèrent une loi hypergéométrique multivariée utilisée pour modéliser un échantillonnage sans remplacement.
Considérons qu'à l'instant $t_k^{(i)}$ (pour un patient i donné, là encore, nous omettons l'indice i pour plus de clarté), les proportions réelles de cellules immatures hétérozygotes et homozygotes sont respectivement $z_{k,het}$ et $z_{k,hom}$ (et pour les WT, $z_{k,wt} = 1 - z_{k,het} - z_{k,hom}$). De l'ensemble des cellules immatures, dont le nombre est inconnu mais très important (de l'ordre de $\sim 3 \cdot 10^{10}$ pour un individu sain d'après les estimations de Cosgrove et al. [25] basées sur [26, 27, 28]), on tire un nombre $\hat{N}_k := \hat{n}_{k,wt} + \hat{n}_{k,het} + \hat{n}_{k,hom}$ de cellules. Parmi ces cellules, nous avons exactement $\hat{n}_{k,het}$ et $\hat{n}_{k,hom}$ cellules mutées hétérozygotes et homozygotes respectivement. On suppose que ces variables aléatoires suivent une loi multinomiale :

$$\mathbb{P}\left[\hat{n}_{k,wt} = n_1, \hat{n}_{k,het} = n_2, \hat{n}_{k,hom} = n_3 \mid z_{k,het}, z_{k,hom}\right] = \frac{(n_1 + n_2 + n_3)!}{n_1! n_2! n_3!} z_{k,wt}^{n_1} z_{k,het}^{n_2} z_{k,hom}^{n_3}$$

4.2 Estimation Bayésienne hiérarchique

Nous considérons N patients d'une population donnée (soit ayant la mutation $JAK2^{V617F}$, $CALR^m$ ou MPL^m). $\boldsymbol{\theta} = \left\{\boldsymbol{\theta}^{(i)}\right\}_{i \in \{1,\cdots,N\}}$ désigne l'ensemble des paramètres :

$$\boldsymbol{\theta}^{(1)} = \left(\theta_1^{(1)}, \cdots, \theta_P^{(1)}\right)$$
$$\vdots$$
$$\boldsymbol{\theta}^{(N)} = \left(\theta_1^{(N)}, \cdots, \theta_P^{(N)}\right)$$

avec $P = 7$ le nombre de paramètres à estimer pour chaque patient $JAK2^{V617F}$ et MPL^m (et $P = 5$ dans le cas $CALR^m$), et N est le nombre de patients dans la population considérée.
Dans un cadre Bayésien non hiérarchique, chaque vecteur de paramètres est estimé individuellement, en ne considérant que les données du patient i. Cette approche peut entraîner un risque d'*overfitting*, notamment pour les patients pour lesquels nous avons peu d'observations. Pour améliorer la robustesse de nos estimations, nous nous sommes placés dans un cadre Bayésien hiérarchique, qui tend à réduire la variance entre les patients et à produire des inférences plus précises [29, 30].
Ce cadre revient à considérer qu'*a priori*, les valeurs des paramètres pour chaque patient sont issues d'une distribution de population. Le cadre hiérarchique tient ainsi compte de la variabilité inter-individuelle (avec la distribution de probabilité) et de la similarité dans la population (le fait que la distribution soit commune à tous les individus de la population). Les paramètres de cette distribution de population sont appelés hyper-paramètres (HP).

Nous illustrons sur la figure 10 l'effet de l'estimation Bayésienne hiérarchique par rapport à une estimation standard, où chaque patient est considéré indépendant. Sur cette figure, nous montrons en haut le *posterior* de l'un des paramètres du modèle, estimé par une approche Bayésienne non hiérarchique pour différents patients. Nous pouvons voir que, pour certains patients, la distribution *a posteriori* du paramètre présente une variance élevée, car nous disposons de peu d'observations pour le patient considéré et/ou parce que sa dynamique est peu informative. En dessous, nous montrons les résultats issus de la méthode d'inférence hiérarchique associée. Les distributions *a posteriori* des paramètres des différents patients sont plus proches de la moyenne de la population, cette dernière ayant été inférée via le cadre hiérarchique. De plus, on observe une réduction de la variabilité et de l'incertitude sur les résultats.

Techniquement, cette approche populationnelle a été mise en œuvre en considérant un modèle hiérarchique à un niveau et en introduisant certains hyper-paramètres
Avec la méthode d'inférence hiérarchique, au lieu d'estimer chaque $\boldsymbol{\theta}^{(i)}$ indépendamment, nous introduisons des hyper-paramètres : $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_P)$ et $\boldsymbol{\sigma^2} = \left(\sigma_1^2, \cdots, \sigma_P^2\right)$ de sorte que, *a priori* :

$$\forall i \in \{1, \cdots, N\}, \forall k \in \{1, \cdots, P\}, \theta_k^{(i)} \mid \tau_k, \sigma_k^2 \sim \mathcal{N}_{c,k}\left(\tau_k, \sigma_k^2\right) \tag{5}$$

où $\mathcal{N}_{c,k}$ est définie comme une distribution gaussienne tronquée (sur un intervalle qui dépend du paramètre considéré k, comme décrit au § 4.5). À noter que, pour les paramètres η_{het} et η_{hom}, nous ne considérons pas d'hyper-paramètres, considérant qu'il n'y a pas de raison d'avoir des conditions initiales échantillonnées à partir de distributions de population.
Nous pouvons alors estimer les distributions *a posteriori* jointes de $\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}$ et des hyper-paramètres $\boldsymbol{\tau}$ et $\boldsymbol{\sigma^2}$:

$$\begin{aligned}
p\left[\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}, \boldsymbol{\tau}, \boldsymbol{\sigma^2} \mid \mathcal{D}\right] &\propto p\left[\mathcal{D} \mid \boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}, \boldsymbol{\tau}, \boldsymbol{\sigma^2}\right] p\left[\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}, \boldsymbol{\tau}, \boldsymbol{\sigma^2}\right] \\
&\propto p\left[\mathcal{D} \mid \boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}\right] p\left[\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}, \boldsymbol{\tau}, \boldsymbol{\sigma^2}\right] \\
&\propto \prod_{i \in \{1, \cdots, N\}} \left(p\left[\mathcal{D}_i \mid \boldsymbol{\theta}^{(i)}\right]\right) p\left[\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)} | \boldsymbol{\tau}, \boldsymbol{\sigma^2}\right] p\left[\boldsymbol{\tau}, \boldsymbol{\sigma^2}\right] \\
&\propto \prod_{i \in \{1, \cdots, N\}} \left(p\left[\mathcal{D}_i \mid \boldsymbol{\theta}^{(i)}\right] p\left[\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\tau}, \boldsymbol{\sigma^2}\right]\right) p[\boldsymbol{\tau}] p\left[\boldsymbol{\sigma^2}\right]
\end{aligned} \tag{6}$$

La relation précédente est obtenue en considérant l'indépendance entre les patients, conditionnellement aux valeurs des hyper-paramètres. De plus, nous simplifions encore la relation ci-dessus en supposant que les composantes de nos vecteurs de paramètres et hyper-paramètres sont indépendantes. Nous obtenons pour le patient i :

$$p\left[\boldsymbol{\theta}^{(i)} \mid \tau, \sigma^2\right] = \prod_{k \in \{1, \cdots, P\}} p\left[\theta_k^{(i)} \mid \tau_k, \sigma_k^2\right]$$

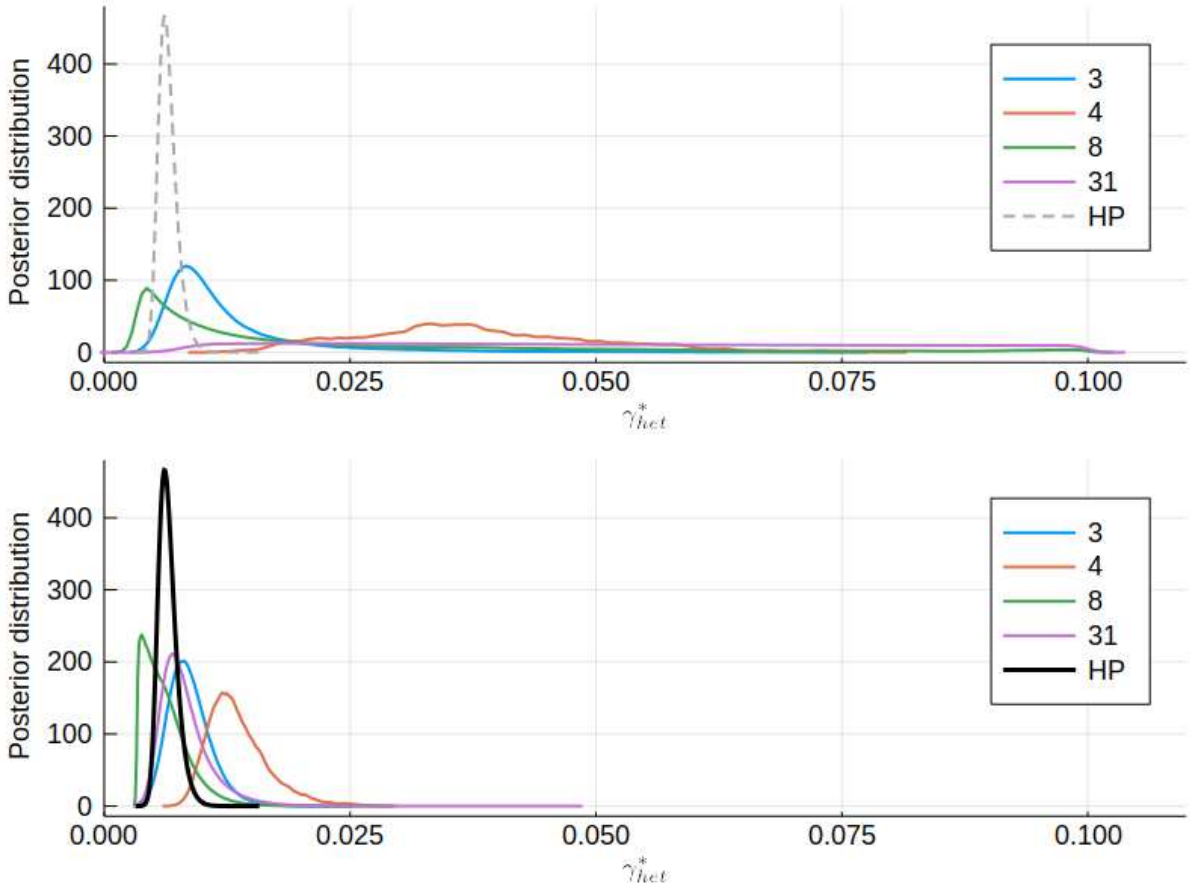


FIGURE 10 – Comparaison entre une approche populationnelle et une estimation Bayésienne standard. Étude de l'effet de l'estimation Bayésienne hiérarchique (en bas) et d'une estimation standard (en haut) avec l'exemple du paramètre γ^*_{het}. La figure du haut montre une estimation des paramètres individuels de quatre patients. Pour le patient #4, nous n'avions des observations qu'en 3 instants, donc pour lui γ^*_{het} a été estimé avec une variance élevée. Il y avait beaucoup d'observations pour le patient #31 mais une dynamique très "plate" au niveau des CF des progéniteurs hétérozygotes mutés donc, étant données les observations, son paramètre γ^*_{het} pouvait potentiellement avoir des valeurs très différentes avec des probabilités égales sans impact sur la qualité des ajustements. Les patients #3 et #8 avaient des dynamiques plus caractéristiques, leurs paramètres γ^*_{het} ont donc été estimés avec une grande précision : des valeurs éloignées du maximum *a posteriori* entraîneraient de mauvais ajustements.
Le graphique du bas montre les résultats de l'estimation hiérarchique basée sur l'approche populationnelle ; nous n'avons pas estimé les paramètres de chaque patient indépendamment mais avons utilisé un cadre hiérarchique. Nous avons considéré *a priori* que le vecteur des paramètres de chaque patient est un échantillon d'une même distribution, décrivant à la fois la variabilité inter-individuelle (avec la distribution de probabilité) et la similarité dans la population (le fait que la distribution est commune à tous les individus de la population). Techniquement, nous avons introduit un paramètre de population appelé hyper-paramètre (HP) qui a été estimé. Cette figure montre clairement l'effet de l'HP sur tous les patients, leurs distributions *a posteriori* (c'est-à-dire l'estimation du paramètre γ^*_{het} avec les incertitudes) est plus concentrée. Cette méthode permet également de réduire la variance des paramètres.

et :

$$p[\boldsymbol{\tau}] = \prod_{k\in\{1,\cdots,P\}} p\left[\tau_k\right]$$

$$p\left[\boldsymbol{\sigma^2}\right] = \prod_{k\in\{1,\cdots,P\}} p\left[\sigma_k^2\right]$$

La vraisemblance $p\left[\mathcal{D}_i \mid \boldsymbol{\theta}^{(i)}\right]$ est exprimée selon le modèle d'observation décrit au paragraphe 4.1.

4.3 Lois conditionnelles pour les hyper-paramètres

La distribution *a posteriori* exprimée précédemment sera estimée numériquement à l'aide de l'algorithme de Metropolis-Hasting *within* Gibbs [31, 32]. Avant d'utiliser cette méthode de calcul, il est utile d'exprimer les distributions *a posteriori* conditionnelles des HP, qui seront utilisée dans le *proposal* de l'algorithme de Metropolis-Hasting.
Concernant l'hyper-paramètre $\boldsymbol{\tau}$, pour $k \in \{1, \cdots, P\}$, en ne gardant dans la formule (6) que les termes impliquant τ_k (puisque nous ne nous intéressons ici qu'à sa distribution à un facteur de multiplication près), on obtient :

$$p\left[\tau_k|\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}, \boldsymbol{\sigma^2}, \mathcal{D}\right] \propto p\left[\theta_k^{(1)}|\tau_k, \sigma_k^2\right] \cdots p\left[\theta_k^{(N)}|\tau_k, \sigma_k^2\right] p\left[\tau_k\right]$$

Les distributions *a priori* pour τ_k sont choisies uniformes sur les intervalles $[a_k, b_k]$ avec les mêmes limites supérieure et inférieure que celles utilisées pour tronquer la loi gaussienne pour $\theta_k|\tau_k, \sigma_k^2$. Ainsi, nous déduisons la densité de probabilité (*a posteriori* et conditionnelle) f_{τ_k} pour $t \in \mathbb{R}$:

$$\begin{aligned} f_{\tau_k}(t) &\propto \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(\theta_k^{(1)} - t)^2\right) \cdots \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(\theta_k^{(N)} - t)^2\right) 1_{[a_k,b_k]}(t) \\ &\propto \exp\left(-\frac{1}{2\sigma_k^2}\left((\theta_k^{(1)} - t)^2 + \cdots + (\theta_k^{(N)} - t)^2\right)\right) 1_{[a_k,b_k]}(t) \end{aligned}$$

Puisque nous avons :

$$\begin{aligned} (\theta_k^{(1)} - t)^2 + \cdots + (\theta_k^{(N)} - t)^2 &= \sum_{i\in\{1,\cdots,N\}} (\theta_k^{(i)})^2 + Nt^2 - 2t \sum_{i\in\{1,\cdots,N\}} \theta_k^{(i)} \\ &= N(t - \frac{1}{N} \sum_{i\in\{1,\cdots,N\}} \theta_k^{(i)})^2 + \cdots \end{aligned}$$

avec le symbole $\cdots$ qui indique des termes n'impliquant pas t, nous obtenons finalement que :

$$f_{\tau_k}(t) \propto \exp\left(-\frac{N}{2\sigma_k^2}\left(t - \frac{1}{N} \sum_{i\in\{1,\cdots,N\}} \theta_k^{(i)}\right)^2\right) 1_{[a_k,b_k]}(t) \tag{7}$$

Ainsi, nous déduisons que τ_k (plus précisément, sa distribution *a posteriori* conditionnelle) suit une loi Gaussienne également tronquée sur l'intervalle $[a_k, b_k]$:

$$\tau_k \sim \mathcal{N}_c(\frac{1}{N} \sum_{i\in\{1,\cdots,N\}} \theta_k^{(i)}, \frac{\sigma_k^2}{N}) \tag{8}$$

Concernant l'HP σ_k^2 (pour $k \in \{1, \cdots, P\}$), nous faisons de même que pour τ_k pour exprimer $\sigma_k^2|\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}, \boldsymbol{\tau}, \mathcal{D}$ et trouvons pour sa densité de probabilité, pour $t \in \mathbb{R}$:

$$f_{\sigma_k^2}(t) \propto \left(\frac{1}{t}\right)^{N/2+1} \exp\left(-\frac{1}{2t} \sum_{i\in\{1,\cdots,N\}} \left(\theta_k^{(i)} - \tau_k\right)^2\right) 1_{[a'_k,b'_k]}(t)$$

en ayant considéré que σ_k^2 suivait un *prior* impropre, à savoir une loi inverse gamma (0,0) - tronquée sur $[a'_k, b'_k]$ (Nous choisissons $a'_k = 0$ and $b'_k = 2$). Nous reconnaissons l'expression d'une loi gamma inverse (tronquée). Pour rappel, une variable aléatoire X qui suit une loi gamma inverse de paramètres α, β a comme densité, pour $x \in \mathbb{R}$:

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}(1/x)^{\alpha+1}\exp(-\beta/x)$$

Ainsi, $\sigma_k^2|\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}, \boldsymbol{\tau}, \mathcal{D}$ suit une loi gamma inverse (tronquée) de paramètres $\alpha_k = \alpha = N/2$ et $\beta_k = 1/2 \sum_{i\in\{1,\cdots,N\}} \left(\theta_k^{(i)} - \tau_k\right)^2$.

4.4 Metropolis-Hasting *within* Gibbs

La distribution *a posteriori* jointe (6) - pour laquelle nous n'avons pas d'expression analytique - est estimée en générant une chaîne de Markov (MCMC - Monte Carlo Markov Chain). Nous utilisons l'algorithme de Metropolis-Hasting (MH) *within* Gibbs [31, 32], qui permet d'échantillonner itérativement les valeurs des paramètres $\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}$ et des hyper-paramètres $\boldsymbol{\tau}$ et $\boldsymbol{\sigma^2}$. Après avoir initialisé la chaîne MCMC, nous échantillonnons successivement les valeurs des HP puis les valeurs des paramètres pour chaque patient (indépendamment, conditionnellement aux valeurs des HP).

Pour initialiser les valeurs des paramètres pour chaque patient, nous exécutons d'abord un algorithme d'optimisation - à savoir l'algorithme CMA-ES [33] introduit au chapitre 4 - pour trouver le vecteur de paramètres qui maximise la vraisemblance. L'algorithme CMA-ES fournit également une matrice de covariance que nous utiliserons pour le *proposal* de l'algorithme Metropolis-Hasting. Ce couplage entre les algorithmes CMA-ES et Metropolis-Hasting within Gibbs est inspiré par les méthodes de type Bayes empirique. Les HP sont initialisés de façon aléatoire.

Nous utilisons la méthode de Gibbs pour échantillonner les valeurs des hyperparamètres (composante par composante). Le *proposal* de l'algorithme de Metropolis est dans ce cas la loi *a posteriori* conditionnelle (que nous avons explicitée dans la section précédente). Avec ce choix, le nouvel échantillon est toujours accepté.

Ensuite, conditionnellement aux valeurs des HP précédemment échantillonnées, nous échantillonnons les valeurs des paramètres pour chaque patient, un par un. Ceci est réalisé en utilisant un schéma standard de Metropolis-Hasting, avec, pour le choix du *proposal*, une distribution normale multivariée de moyenne nulle et de matrice de covariance $\boldsymbol{\Sigma}$.
Lorsque la dimension de l'espace des paramètres est grande, l'algorithme de Metropolis-Hastings s'avère souvent inefficace, car il devient compliqué de définir une matrice de covariance permettant une bonne convergence de l'algorithme. Des algorithmes adaptatifs ont été proposés pour contourner ce problème [34]. Nous proposons ici une méthode alternative et plus simple qui s'est avérée très efficace en pratique : nous choisissons de fixer la matrice de covariance du *proposal* à celle estimée par l'algorithme CMA-ES (ajustée par un facteur multiplicatif près réglé pour obtenir des taux d'acceptation convenables). Ce choix se justifie du fait que l'apprentissage de la matrice de covariance par l'algorithme CMA-ES est analogue à l'apprentissage de la matrice Hessienne inverse dans une méthode quasi-Newton [35].

Nous avons implémenté les méthodes en utilisant le langage de programmation Julia. Le framework utilisé pour l'estimation des paramètres (et qui peut être utilisé pour un large éventail de problèmes) est disponible à l'adresse :

`https://gitlab-research.centralesupelec.fr/2012hermangeg/bayesian-inference`

Les calculs sont exécutés sur 25 millions d'itérations, un nombre suffisant pour atteindre la convergence de l'algorithme (lorsque les moyennes ergodiques ne varient plus). Nous choisissons une *burn-in length* de 15 millions d'itérations. Pour chaque estimation (une pour chaque population), nous exécutons l'algorithme deux fois avec des graines aléatoires différentes pour vérifier que nous obtenons des résultats similaires à chaque fois.

4.5 Priors

Cinq paramètres (trois pour un individu $CALR^m$) et deux conditions initiales sont estimés par patient.
Pour η_{het} et η_{hom}, nous considérons des distributions *a priori* uniformes sur un intervalle [0,3]. Pour les autres paramètres, nous les considérons *a priori* échantillonnés suivant des distributions de population (voir équation (5)). Ces dernières sont des distributions gaussiennes tronquées, avec une moyenne τ_k et une variance σ_k^2 pour le paramètre k. Pour la variance du paramètre k, nous avons choisi une distribution *a priori* impropre : une loi inverse-gamma (0,0). Les hyperparamètres correspondant aux moyennes des distributions gaussiennes suivent *a priori* une distribution uniforme, sur les mêmes intervalles utilisés pour tronquer la distribution gaussienne. Ces intervalles sont [1,20], $[1/300, 0.1]$ et [-1,1] pour les paramètres $k_{m,het}$, γ^*_{het} (et γ^*_{hom}) et Δ^*_{het} (et Δ^*_{hom}), respectivement.

5 Inférence sur données simulées

5.1 Simulation de données

Après les différentes hypothèses simplificatrices proposées plus haut, nous aboutissons (pour le modèle à quatre compartiments) à 7 paramètres devant être estimés. Nous souhaitons alors vérifier l'identifiabilité pratique du modèle. Comme l'ont souligné Duchesne et al. [36], le test d'identifiabilité ne peut être évité si l'on souhaite utiliser le modèle à des fins de prédiction. Notre approche ici sera basée sur l'identification du modèle à partir de jeux de données synthétiques.

Pour vérifier l'identifiabilité du modèle (celui à quatre compartiments, utilisé pour la population de patients $JAK2^{V617F}$), nous générons 30 patients virtuels. Nous échantillonnons 30 vecteurs de paramètres (à partir d'une distribution de population que nous choisissons), conduisant à 30 dynamiques différentes. Pour chacune d'entre elles, nous échantillonnons les temps d'observation et ajoutons un bruit aux observations "théoriques" pour reproduire un jeu de donnés simulé en accord avec celui de notre cohorte de patients.

Cinq valeurs de paramètres et deux conditions initiales sont nécessaires pour générer la dynamique virtuelle d'un patient (voir Fig. 9). Les distributions de population utilisées pour générer les paramètres sont les suivantes :

- $\eta_{het} \sim \mathcal{U}([0,1])$ (uniforme)
- $\eta_{hom} \sim \mathcal{U}([0,2.5])$ (uniforme)
- $\Delta^*_{het} \sim \mathcal{N}(-0.1, 0.1^2)$ (tronquée sur $[-1,1]$)
- $\Delta^*_{hom} \sim \mathcal{N}(-0.3, 0.1^2)$ (tronquée sur $[-1,1]$)
- $\gamma^*_{het} \sim \mathcal{N}(1/40, 0.01^2)$ (tronquée sur $[1/300, 1/10]$)
- $\gamma^*_{hom} \sim \mathcal{N}(1/120, 0.01^2)$ (tronquée sur $[1/300, 1/10]$)
- $k_{m,het} \sim \mathcal{N}(10,1)$ (tronquée sur $[1,20]$)

Sur la figure 11, nous montrons tous les paramètres qui ont été échantillonnés et les distributions de population correspondantes.
Les paramètres ayant été fixés, nous disposons de la dynamique réelle pour les 30 patients virtuels (puisque notre modèle est déterministe). Nous voulons générer des données bruitées pour chacun d'entre eux. Pour être en accord avec nos données réelles, tous les patients virtuels n'auront pas

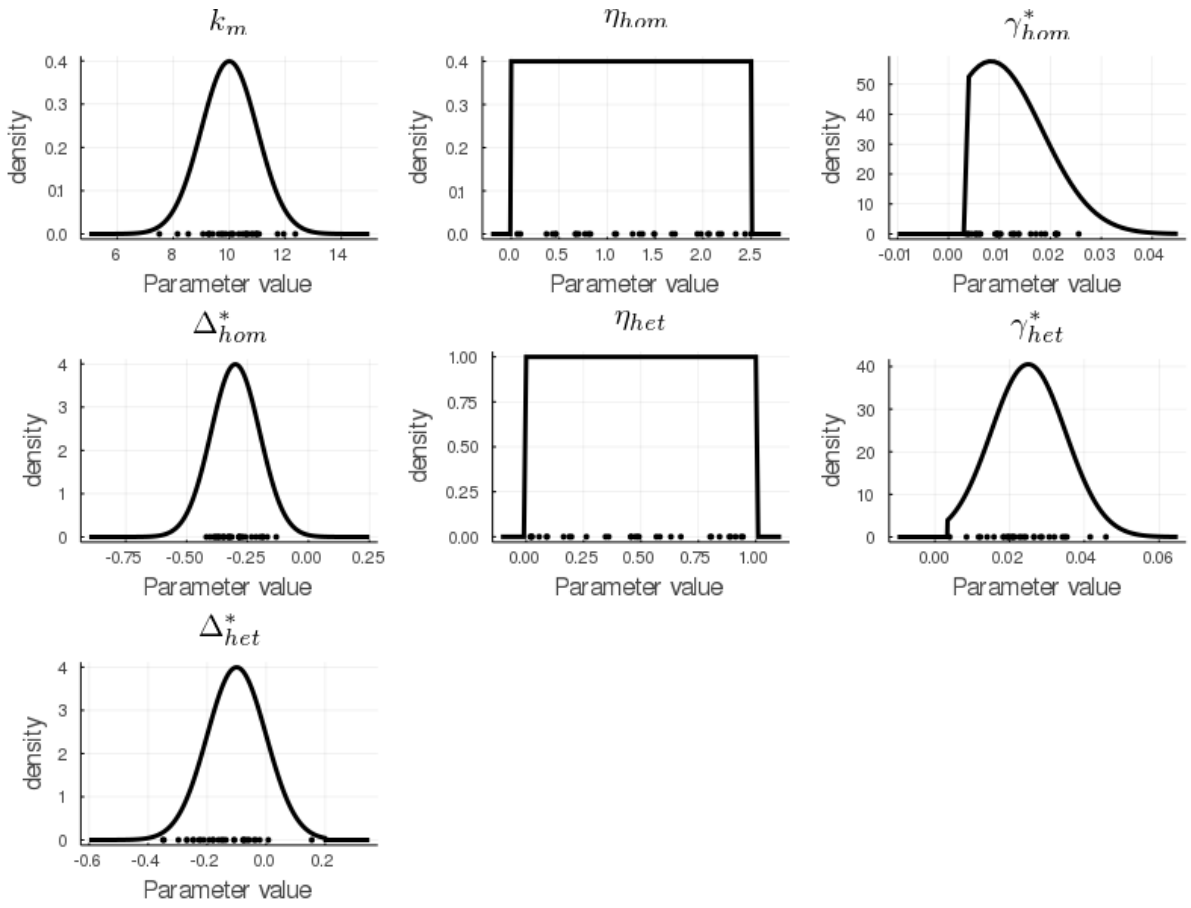


FIGURE 11 – Paramètres générés pour 30 patients virtuels. Les points sont des paramètres échantillonnés. La ligne représente la densité réelle de la population, à partir de laquelle nous avons échantillonné des valeurs.

le même nombre d'observations n_i (indice i pour les patients). La figure 12 montre la distribution du nombre d'échantillons sur la cohorte virtuelle. Le nombre moyen d'observations est égal à 10, certains patients virtuels n'auront des observations qu'à trois instants, tandis que d'autres en ont 15.
Nous pouvons alors générer un nombre n_i d'observations bruitées à partir de la dynamique réelle i de chaque patient virtuel. Tout d'abord, nous générons les temps d'observations (en jours). Nous considérons toujours que la première mesure est effectuée au moment initial. Ensuite, nous générons les autres de manière aléatoire selon la méthode suivante, qui revient à considérer une observation environ tous les trimestres :

```
function sample_time(n)
    t = zeros(n)
    t[1] = 0.0
    for i in 2:n
        t[i] = t[i-1] + rand(truncated(Normal(100, 10),30, 300))
    end
    return t
end
```

Ensuite, à partir des ratios (VAF ou CF) réels de cellules immatures et matures à ces moments-là, nous ajoutons du bruit. Nous utilisons le même modèle de bruit que celui décrit au § 4.1.
À ce stade, nous disposons de 30 ensembles de données comparables à ceux de notre cohorte de patients $JAK2^{V617F}$. L'objectif est de retrouver la dynamique réelle pour chaque patient en utilisant notre méthode d'inférence.

5.2 Résultats sur données simulées

Nous avons exécuté la méthode d'estimation précédente sur 13 millions d'itérations, avec une *burn-in length* de 2 millions.
Le code permettant de l'exécuter (implémentation du modèle, données simulées et *settings* pour la méthode d'inférence) est disponible au lien suivant :

`gitlab-research.centralesupelec.fr/2012hermangeg/identifiability-base-model`

Toutes les dynamiques inférées sont présentées sur la figure 14 pour nos 30 patients. Nous comparons également la dynamique prédite (lignes pleines) avec la dynamique réelle (lignes pointillées). Il y a un bon accord entre les deux.
La figure 13 compare les fractions clonales (CF) inférées *vs* réelles pour les HSC mutées hétérozygotes et homozygotes. Même si nous ne disposons pas de données pour les HSC, notre méthode d'inférence permet de retrouver les vraies valeurs avec précision.
Dans ce chapitre, la principale grandeur d'intérêt est notre facteur de réponse R. Il sera utilisé pour comparer des sous-groupes de patients en fonction du dosage qu'ils ont reçu, ou pour comparer différentes réponses en fonction de la zigosité. Il est donc important de l'estimer avec précision. Nous pouvons voir sur la figure 15 que c'est bien le cas.

Au niveau individuel, les résultats de l'estimation des paramètres sont présentés sur la figure 16. η_{het}, η_{hom}, γ^*_{het}, γ^*_{hom} et Δ^*_{het} sont correctement estimés, avec des intervalles de crédibilité à 95% qui contiennent toujours la vraie valeur. k_m est mal estimé : toutes les valeurs se concentrent autour de 10.5 de sorte que l'effet de la population semble prédominer sur la variabilité inter-individuelle, ce qui suggère que la sortie du modèle pourrait ne pas être si fortement sensible à la valeur k_m. Contrairement à Δ^*_{het}, Δ^*_{hom} n'est pas prédit avec précision, et nous observons un effet populationnel important, ce qui n'a pas été observé pour les HSC hétérozygotes. Cela pourrait être dû au fait que les valeurs réelles (virtuelles) de Δ^*_{hom} sont plus faibles que celles de Δ^*_{het}. Des valeurs faibles pourraient induire une augmentation plus forte de la CF dans les premiers jours de la thérapie, de sorte qu'il soit plus difficile d'estimer la valeur du paramètre

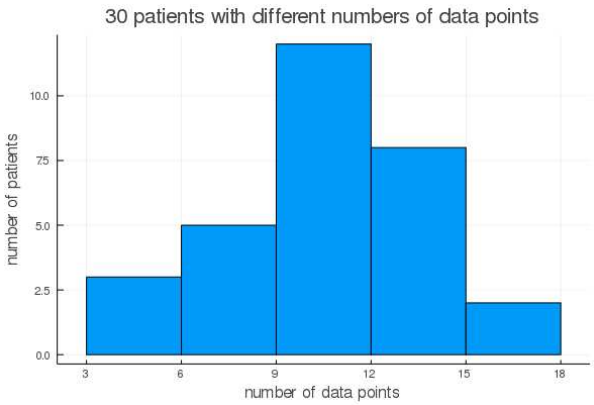


FIGURE 12 – Distribution du nombre d'instants n pour lesquels on dispose d'observations parmi nos 30 patients virtuels.

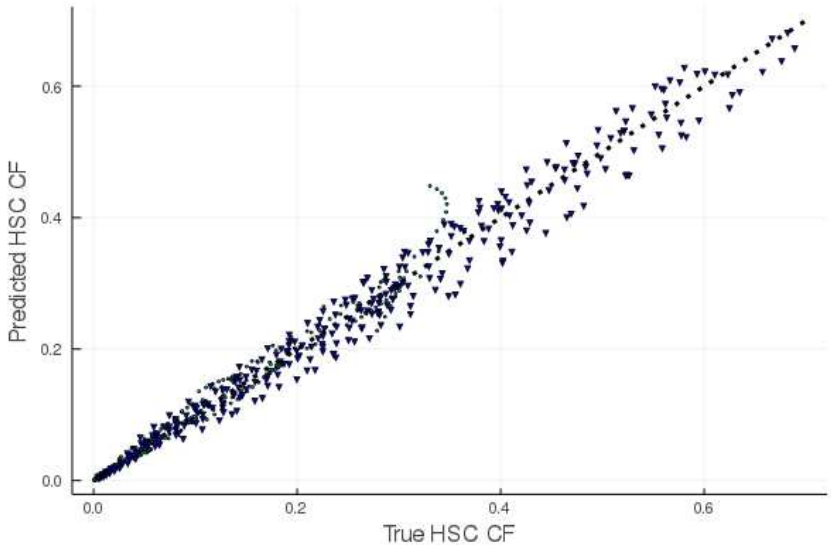


FIGURE 13 – Comparaison entre les CF réelles et les CF inférées, au niveau des HSC. Les valeurs sont prises au moment des observations à partir de notre ensemble de données virtuelles. Les triangles bleus font référence aux HSC homozygotes, les cercles verts aux HSC hétérozygotes.

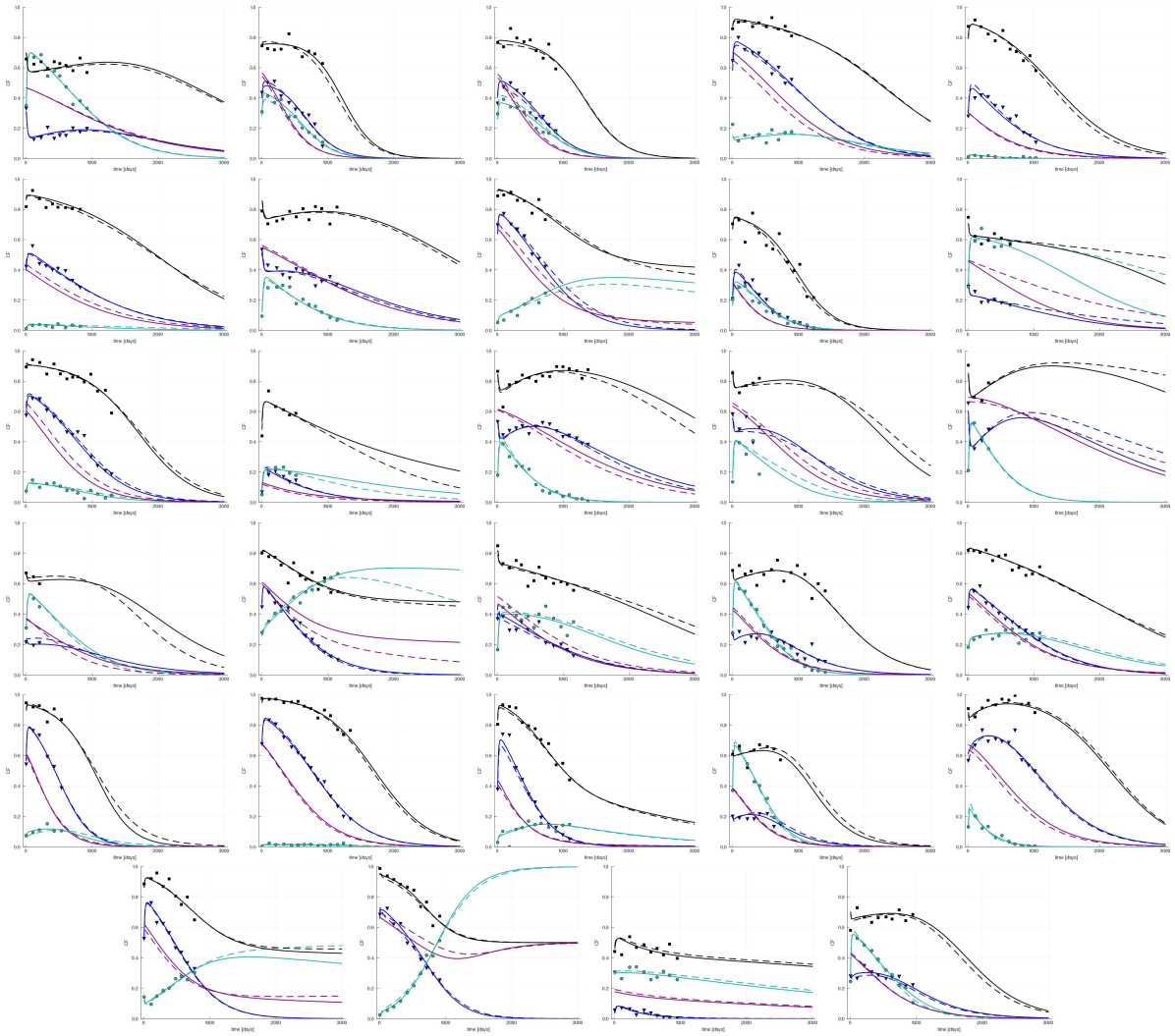


FIGURE 14 – Comparaison des dynamiques réelles et estimées pour nos 30 patients virtuels. Les lignes en tirets représentent la dynamique réelle. Les lignes pleines représentent la dynamique inférée (basée sur la médiane du *posterior* du vecteur de paramètres). Les carrés noirs, les cercles verts et les triangles bleus sont nos données bruitées pour la VAF mature et les CF hétérozygote et homozygote des cellules immatures respectivement. La ligne violette fait référence à la VAF mutée au niveau des HSC.

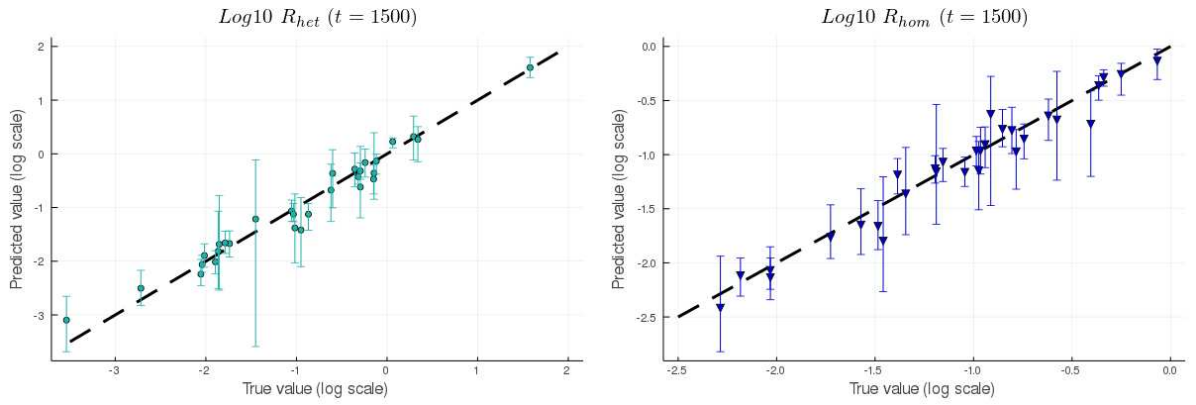


FIGURE 15 – Comparaison entre le R-factor réel (en abscisse) et inféré (en ordonnée). Le facteur de réponse est ici calculé à $t = 1,500$. Les axes sont en échelle logarithmique. À gauche, nous avons le R-factor hétérozygote, et à droite celui calculé sur les HSC homozygotes. La barre d'erreur correspond à un intervalle de crédibilité à 95%.

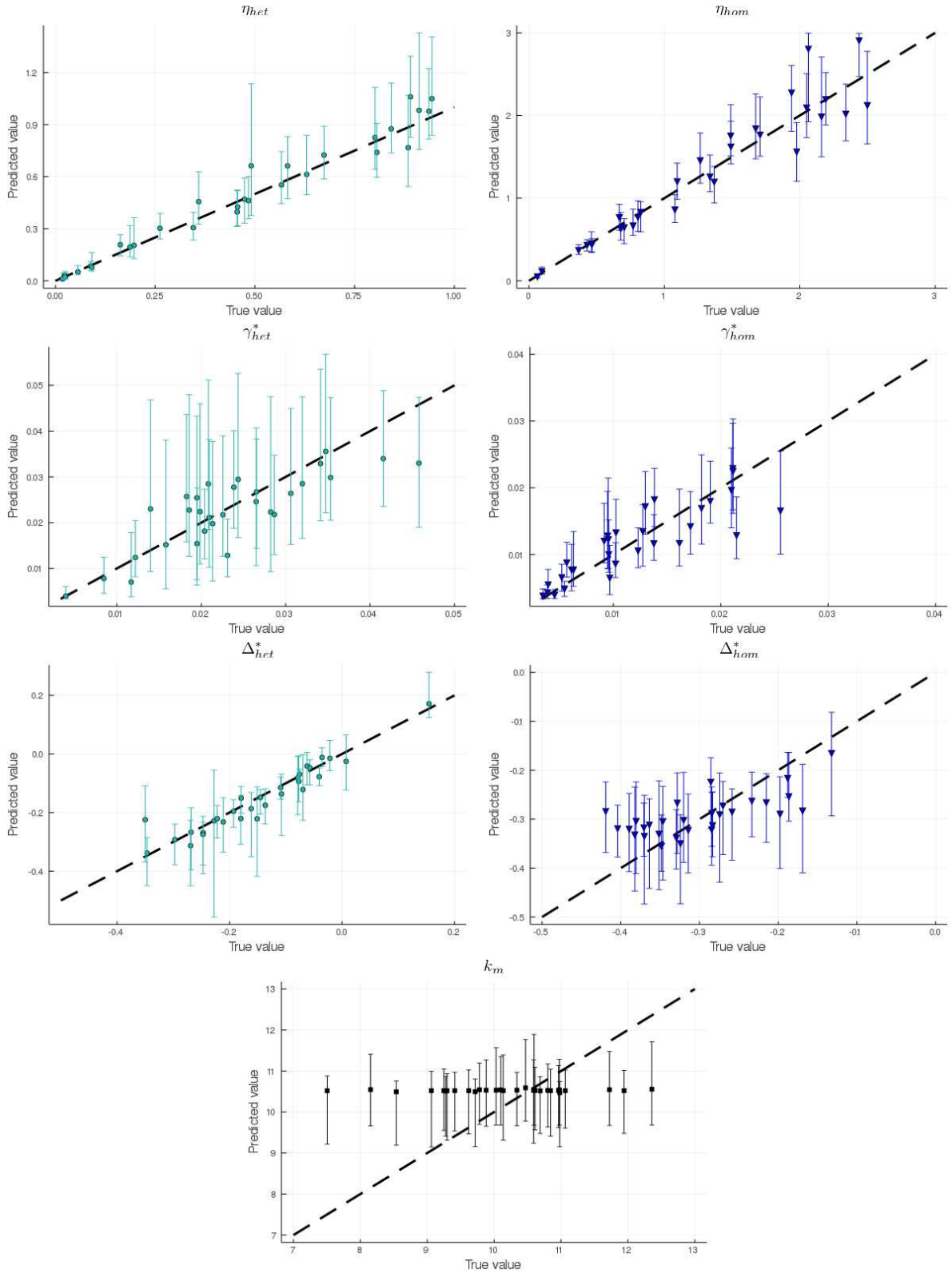


FIGURE 16 – Comparaison entre les valeurs réelles (en abscisse) et inférées (en ordonnée) des paramètres pour nos 30 patients virtuels. Les valeurs inférées sont présentées avec leurs intervalles de crédibilité à 95%.

avec précision puisque nous ne disposons pas de données pour les premiers jours de la thérapie (sauf au moment initial).
Nous analysons maintenant les résultats précédents au niveau de la population : nous comparons la densité de la population estimée à la densité réelle. Les résultats sont présentés sur la figure 17. En orange, nous avons les résultats de l'estimation. Nous voyons sur cette figure comment nous pouvons inférer non seulement les valeurs individuelles mais aussi la densité de la population. Cela fonctionne très bien pour γ^*_{het}. Pour Δ^*_{hom} nous avons également de bons résultats, même si nous prédisons moins de variance qu'en réalité, mais la moyenne de la densité de la population est très bien estimée. Ceci est intéressant car les estimations au niveau individuel n'étaient pas aussi bonnes. Pour k_m, la moyenne de la densité de la population est assez bien estimée, mais la variance est plus faible que dans la réalité.

Pour conclure, cette étude d'identifiabilité pratique démontre la capacité de notre méthode d'estimation à inférer avec précision la dynamique et les réponses à long-terme des patients à partir des observations qu'on pourra avoir pour notre cohorte réelle. Le facteur de réponse est la principale quantité d'intérêt utilisée pour présenter les résultats dans ce chapitre, et est notamment utilisé pour évaluer l'effet du dosage et stratifier les patients en différents groupes (§ 6.2). Nous avons montré qu'il peut être estimé avec précision.
De plus, nous présenterons au § 6.3 les valeurs estimées pour les paramètres de notre modèle, afin d'inférer le mécanisme d'action de l'IFNα. En particulier, nos paramètres sont utilisés pour comparer le mécanisme entre les cellules hétérozygotes et homozygotes, de sorte qu'il est important d'estimer avec précision l'effet populationnel. Nous constatons que la plupart des paramètres sont bien estimés au niveau individuel et que le cadre hiérarchique permet de retrouver les distributions de population qui ont généré les paramètres individuels.

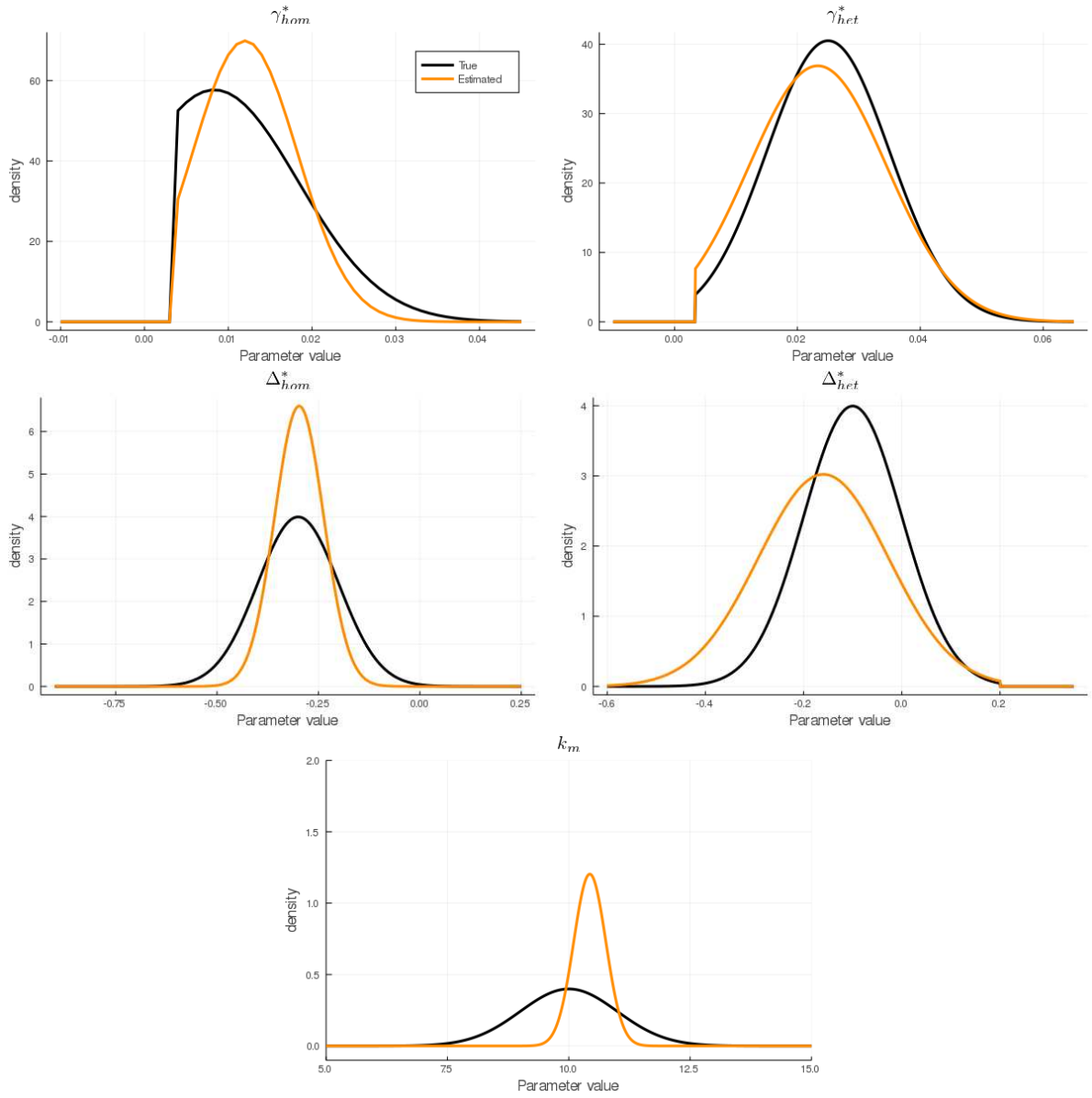


FIGURE 17 – Comparaison de la densité réelle de la population (en noir) utilisée pour générer les valeurs des paramètres et celle estimée (en orange) par notre cadre Bayésien hiérarchique. La densité en orange est la densité estimée de la population, c'est-à-dire une loi gaussienne tronquée avec une moyenne et une variance qui sont les hyper-paramètres estimés (valeur moyenne de leurs distributions *a posteriori*).

6 Résultats

6.1 Inférence de la dynamique des cellules mutées sous IFNα

Pour caractériser précisément la dynamique induite par l'IFNα sur les HSC mutées, les progéniteurs et les cellules matures, et pour déduire les effets du traitement à long terme sur les HSC mutées, nous avons conçu un modèle mathématique compartimental, reposant sur l'hypothèse que l'IFNα induit une dynamique latente et non observée des HSC qui impacte ensuite celle des cellules progénitrices et des granulocytes [16].
Nous avons calibré le modèle sur la base des données de 27 patients $JAK2^{V617F}$, 12 $CALR^m$ et 1 MPL^m, en utilisant une méthode d'estimation Bayésienne hiérarchique pour augmenter la robustesse des résultats. Partant du principe que les patients présentant la même mutation motrice du NMP auraient des valeurs de paramètres comparables, nous avons considéré trois sous-populations indépendantes, c'est-à-dire trois distributions de paramètres indépendantes dans le cadre hiérarchique. Avec relativement peu de degrés de liberté, notre modèle simple peut s'adapter aux données moléculaires longitudinales (Fig. 18) et décrire la dynamique des progéniteurs et des cellules matures $JAK2^{V617F}$, $CALR^m$ et MPL^m chez les patients jusqu'à par exemple 3,000 jours de traitement (Fig. 19, 20, 22, 23 et 21).
Cette approche basée sur un modèle nous a permis d'inférer les effets de l'IFNα sur les HSC mutées en fonction du type de mutation et de la zygosité. Nous avons estimé qu'il y avait une déplétion rapide du compartiment des HSC homozygotes $JAK2^{V617F}$, concomitante dans certains cas à une augmentation initiale rapide suivie d'une diminution drastique à la fois de la population de progéniteurs homozygotes et des granulocytes $JAK2^{V617F}$ (courbe en cloche, voir par exemple le patient #32). Parmi les cellules hétérozygotes $JAK2^{V617F}$, nous inférons une déplétion plus lente du compartiment des HSC, des progéniteurs et des granulocytes.
La dynamique des cellules $CALR^m$ (10/12 patients n'avaient aucun sous-clone homozygote) était plus hétérogène, avec une déplétion lente chez les patients présentant une réponse moléculaire. Enfin, nous avons inféré une déplétion rapide des HSC MPL^m hétérozygotes.
En résumé, la dynamique latente inférée au niveau des HSC est en accord avec l'analyse effectuée à partir des observations dans les progéniteurs et les cellules matures (§ 2.3) ; les données expérimentales sont en accord avec notre hypothèse de travail selon laquelle l'IFNα agirait sur les HSC mutées, favorisant leur sortie de quiescence et leur différenciation.

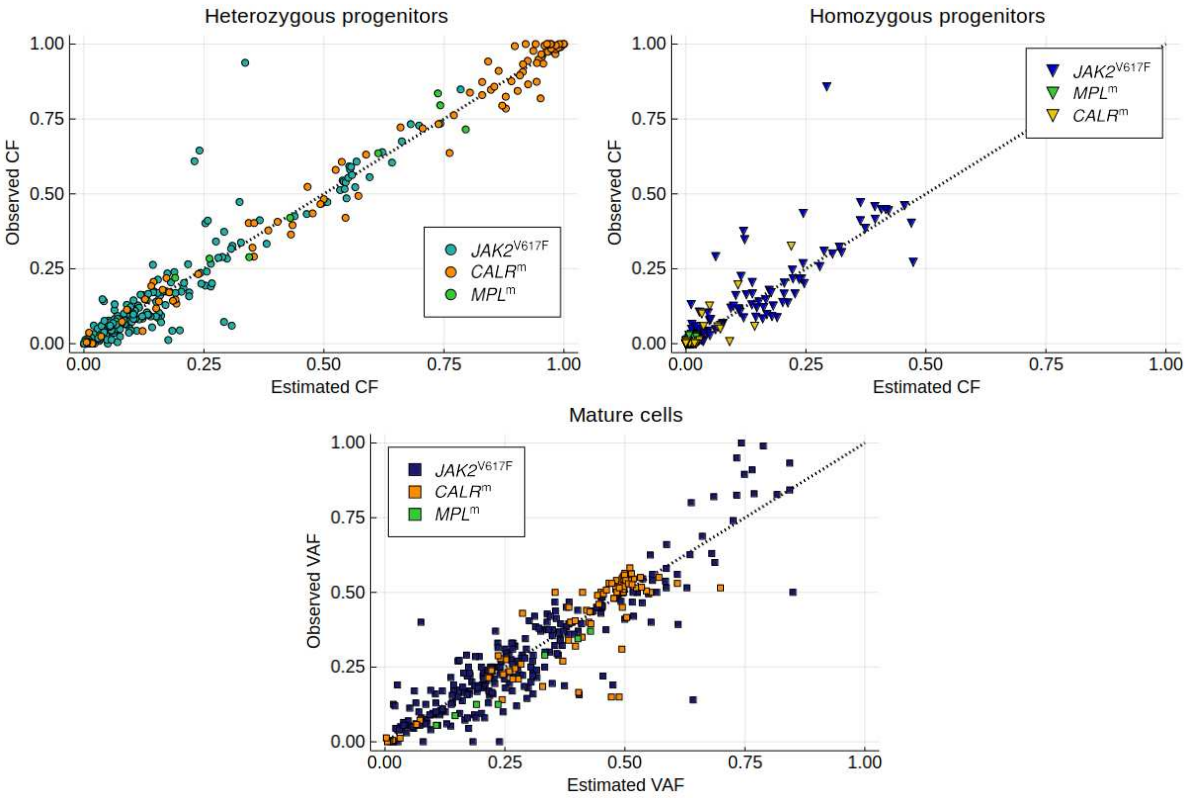


FIGURE 18 – Comparaisons entre les valeurs observées (axe des ordonnées) et inférées (axe des abscisses, valeur médiane) de la CF ou VAF parmi les progéniteurs et les cellules matures. La ligne en pointillés représente les ajustements exacts, les points éloignés de cette ligne ne sont pas en accord avec le modèle. Cette figure donne un aperçu de l'ensemble des observations utilisées pour la calibration du modèle, et la qualité des ajustements au modèle.
Pour les cellules matures, nous avons 98, 264 et 7 observations pour les cas $CALR^m$, $JAK2^{V617F}$ et MPL^m, respectivement. La plupart des carrés sont localisés près de la première bissectrice, ce qui indique que le modèle décrit correctement la dynamique de la VAF mature. Néanmoins, pour une VAF expérimentale élevée dans les cellules matures, le modèle sous-estime la VAF théorique. Pour les progéniteurs, nous avons 96, 258 et 6 observations pour les cas $CALR^m$, $JAK2^{V617F}$ et MPL^m, respectivement. Il y a un très bon ajustement entre la CF hétérozygote observée et celle estimée pour la plupart des observations, avec des valeurs allant de 0 à 1. Il est plus difficile d'évaluer la qualité des estimations pour les cellules homozygotes car la plupart des points expérimentaux ont une valeur négligeable.

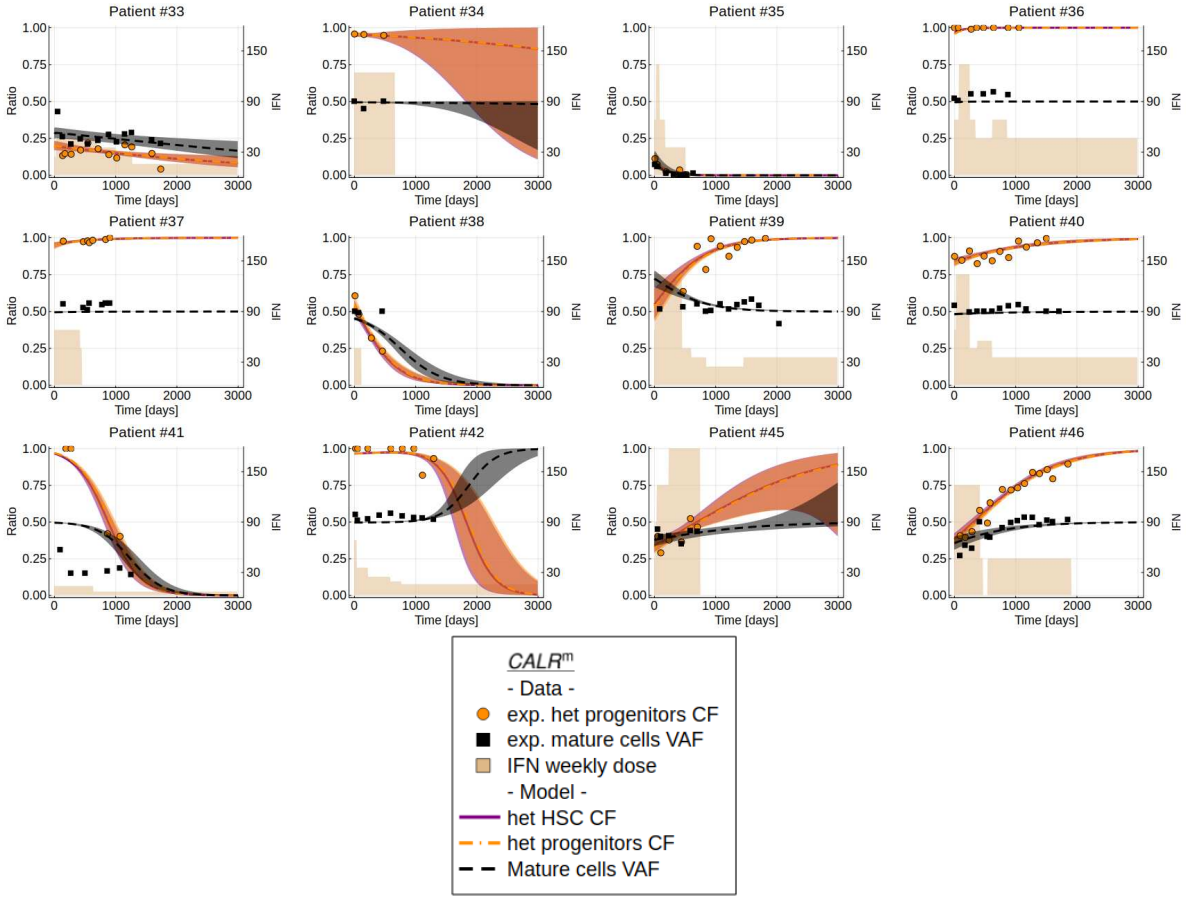


FIGURE 19 – Dynamique inférée (pour les populations de cellules mutées hétérozygotes) pour les patients $CALR^m$. Les points, triangles et carrés représentent les observations expérimentales. Les courbes ont été déterminées avec le modèle mathématique (valeurs médianes). La ligne violette représente la dynamique inférée des HSC mutés. Les zones ombrées représentent les intervalles de crédibilité à 95%. Les zones grisées en beige correspondent à la dose d'IFNα reçue au cours du traitement. Dans l'ensemble, les ajustements sont très bons. Le patient #41, qui présentait une faible VAF mature ($\sim$0.2-0.25) au cours des premiers mois de traitement, ce qui n'était pas compatible - selon notre modèle - avec une CF ~ 1 parmi les progéniteurs, est le seul cas $CALR^m$ qui présentait une divergence entre les données et le modèle.

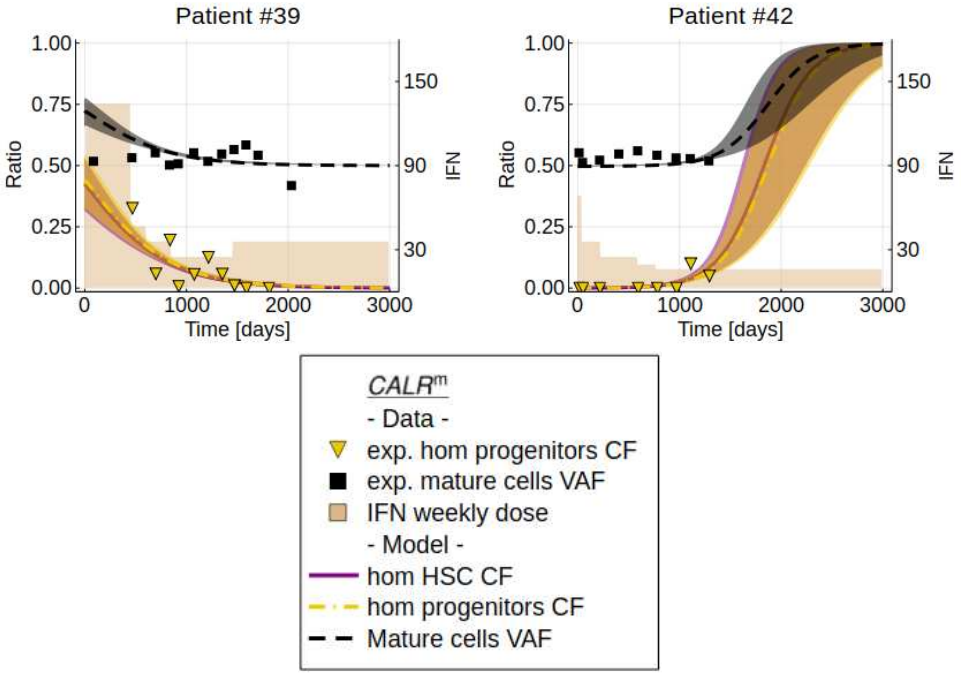


FIGURE 20 – Dynamique inférée, pour les populations de cellules mutées homozygotes, pour les patients $CALR^m$ qui avaient des sous-clones homozygotes.

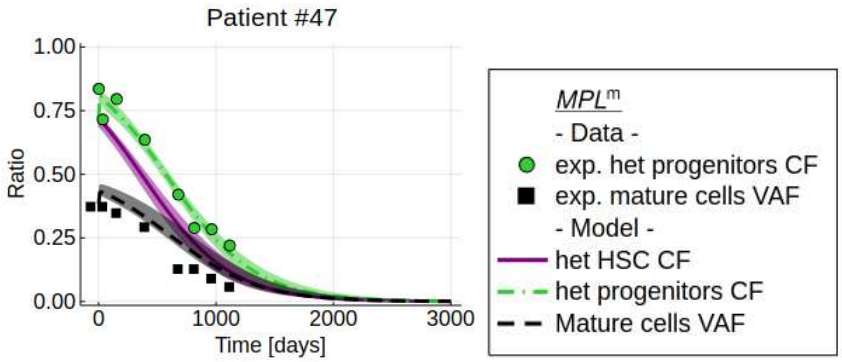


FIGURE 21 – Dynamique inférée dans le cas du patient MPL^m.

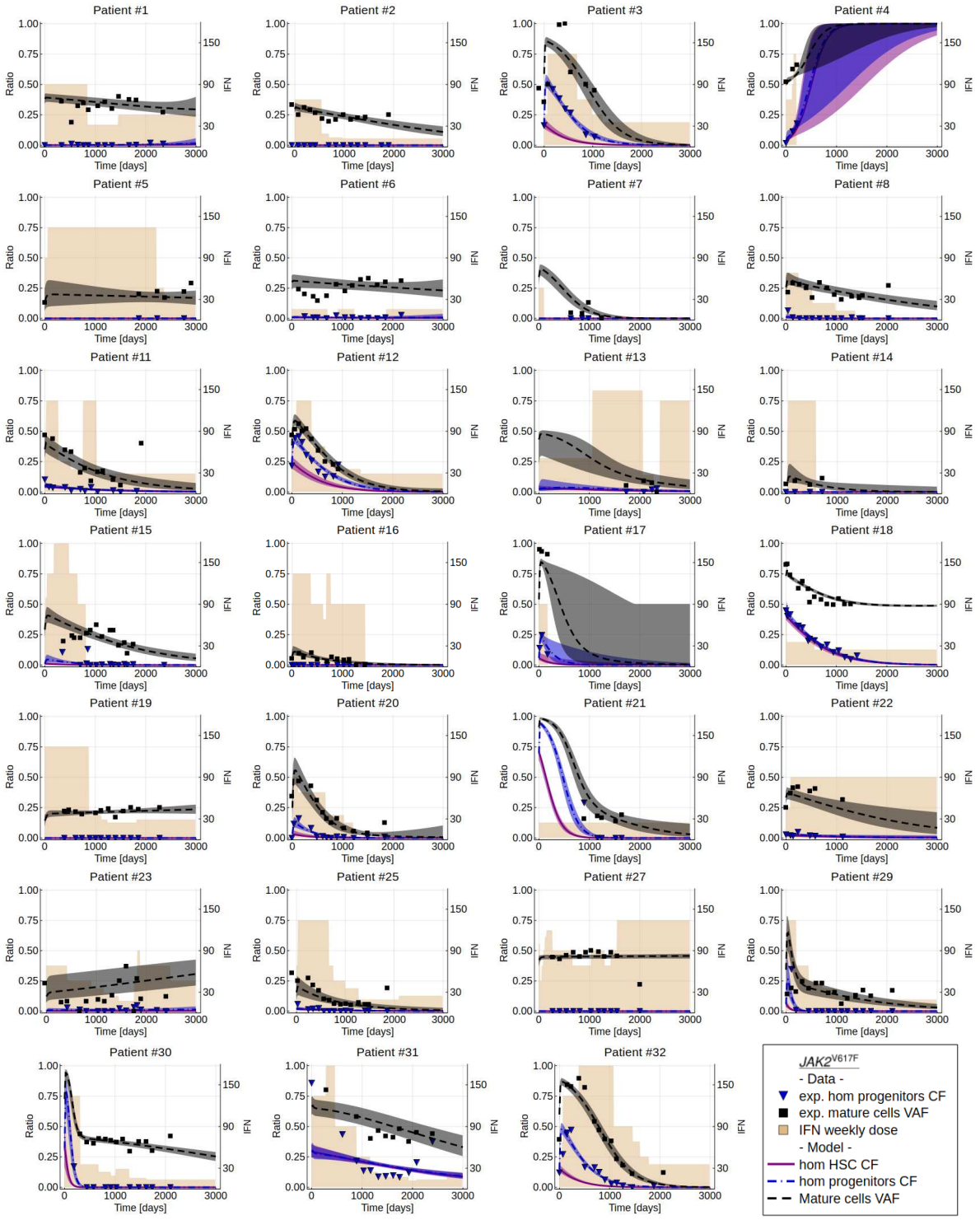


FIGURE 22 – Dynamique inférée, pour les populations de cellules mutées homozygotes, pour les patients $JAK2^{V617F}$. Dans la plupart des cas, il y a un bon ajustement entre les données expérimentales et les dynamiques inférées. En particulier, notre modèle a reproduit et expliqué l'augmentation suivie de la diminution des CF progéniteurs homozygotes observée au cours des premiers mois de traitement. Pour le patient #31 uniquement, le modèle n'a pas réussi à décrire la dynamique expérimentale de la CF homozygote, avec d'abord une diminution suivie d'un plateau et enfin d'une augmentation.

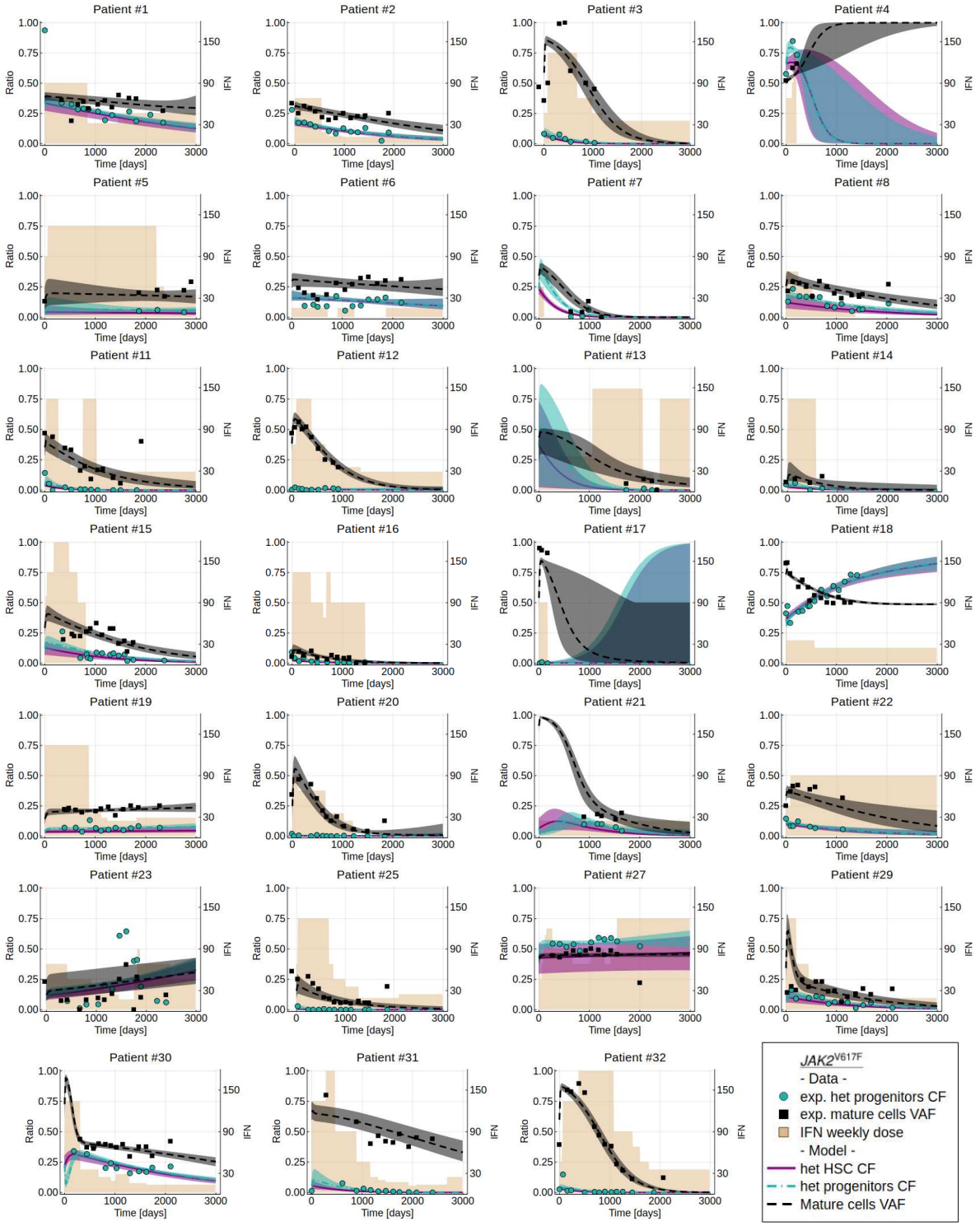


FIGURE 23 – Dynamique inférée, pour les populations de cellules mutées hétérozygotes, pour les patients $JAK2^{V617F}$. Dans la plupart des cas, il y a une bonne adéquation entre notre dynamique estimée et les données expérimentales, tant pour le VAF mature que pour la CF hétérozygote. Seuls quelques patients présentent une dynamique expérimentale qui ne peut pas être décrite par le modèle. C'est le cas du patient #23, qui a présenté une augmentation très tardive puis une diminution à la fois de sa CF hétérozygote progénitrice et de sa VAF mature. Le patient #15 présentait une dynamique similaire. Il est intéressant de noter que les patients #2 ou #6 avaient une dynamique oscillatoire. Ces phénomènes ne peuvent pas être capturés par notre modèle dans son état actuel car nous soupçonnons qu'il repose sur des variations de dose au cours du traitement. À noter que l'incertitude n'est pas la même pour toutes les mesures d'architecture clonale, selon le nombre effectif de clones génotypés par échantillon de sang (voir § 4.1), ce qui peut justifier que les courbes théoriques sont plus susceptibles de passer par certains points que d'autres.

6.2 Stratification des patients

La plupart des patients $CALR^m$ de notre cohorte ont des cellules mutées hétérozygotes. Nous avons alors quantifié la réponse moléculaire hétérozygote à long terme dans les HSC après 3,000 jours en estimant le facteur de réponse (R-factor), défini comme la proportion de HSC mutées (ou fraction clonale CF) par rapport à sa valeur initiale. Nous avons observé une réponse moléculaire hétérogène dans les HSC $CALR^m$, avec des dosages d'IFNα élevés corrélées à une moins bonne réponse au niveau des HSC (Fig. 24-i et 25-i). Même si peu de patients ont été analysés, les HSC $CALR^m$ de type 2 semblent être ciblées plus efficacement que les HSC $CALR^m$ de type 1 (Fig. 24-ii).

En revanche, nous avons observé une réponse moléculaire pour la plupart des patients $JAK2^{V617F}$, avec un R-factor significativement meilleur dans le cas de HSC mutées homozygotes que hétérozygotes (Fig. 24-iii). Pour les clones hétérozygotes $JAK2^{V617F}$, nous avons estimé que la réponse était meilleure lorsque le patient était traité à un plus fort dosage d'IFNα (p=0.0745 avec le test U de Mann-Whitney ; p=0.0498 en testant la nullité du coefficient de régression linéaire, voir figure 25) (Fig. 24-iv). Nous n'avons constaté aucun effet du dosage dans le cas des clones mutés homozygotes (Fig. 24-v). Lors de l'analyse de la VAF globale pour le compartiment souche, les patients $JAK2^{V617F}$ traités par des dosages élevés d'IFNα étaient ceux ayant le mieux répondu (Fig. 24-vi et 25-ii, iv).

Nous avons ensuite estimé le temps médian T_{PMR} pour atteindre une rémission moléculaire partielle (chez les patients répondants). D'après nos estimations, les HSC homozygotes $JAK2^{V617F}$ étaient ciblées plus rapidement que celles hétérozygotes chez les patients répondeurs (350 jours contre 920 jours) (Fig. 26-i). Un fort dosage d'IFNα réduirait de manière significative T_{PMR} pour les HSC hétérozygotes (Fig. 26-ii), mais pas dans le cas homozygote (Fig.26-iii). En comparaison, alors qu'ils recevaient un faible dosage d'IFNα, nous avons estimé que les patients MPL^m et $CALR^m$ de type 2 atteignaient le temps de PMR en moyenne à 600 jours (Fig. 26-iv).

Chez les patients $JAK2^{V617F}$, nous n'avons trouvé aucune évidence que le type de maladie (PV ou TE), l'âge du patient au début du traitement, le sexe ou la présence de mutations associées pouvaient influencer la dynamique des cellules souches mutées.
En conclusion, nous avons estimé que la réponse moléculaire au niveau des HSC induite par l'IFNα dépendait du type de mutation ($JAK2^{V617F}$,$CALR^m$ type 1/2), de la zygosité de $JAK2^{V617F}$ et du dosage de l'IFNα.

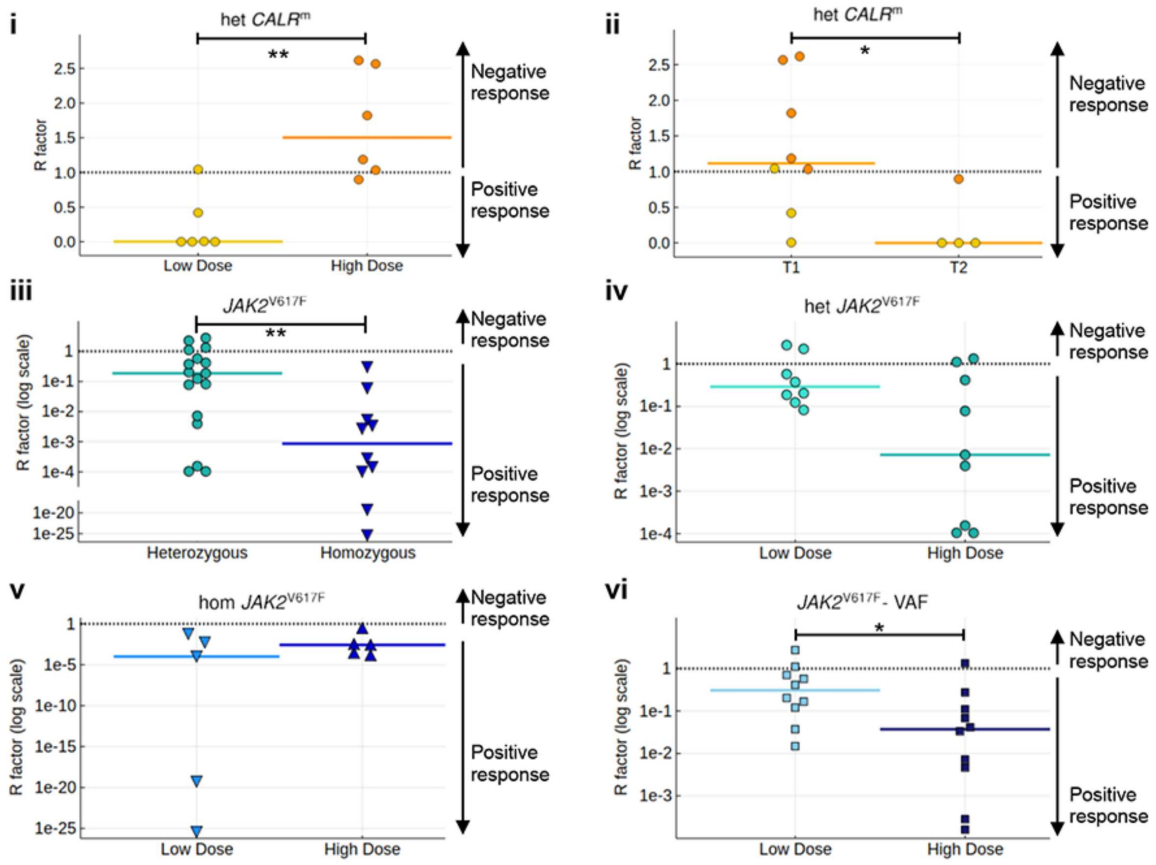


FIGURE 24 – Le facteur de réponse des cellules souches moléculaires (R-factor) a été inféré à $t = 3,000$ jours pour (i) les HSC $CALR^m$ hétérozygotes chez les patients traités avec des dosages élevés *vs* faibles d'IFNα ; (ii) les HSC $CALR^m$ hétérozygotes de type 1 par rapport aux HSC $CALR^m$ hétérozygotes de type 2 ; (iii) les HSC $JAK2^{V617F}$ hétérozygotes par rapport aux HSC $JAK2^{V617F}$ homozygotes ; (iv) les HSC $JAK2^{V617F}$ hétérozygotes chez les patients traités avec des dosages élevés par rapport aux dosages faibles d'IFNα ; (v) les HSC $JAK2^{V617F}$ homozygotes chez des patients traités par des dosages élevés d'IFNα par rapport à des dosages faibles et (vi) les HSC $JAK2^{V617F}$ hétérozygotes et homozygotes (VAF) chez des patients traités par des dosages élevés d'IFNα par rapport à des dosages faibles. Selon le contexte, le R factor est défini comme un rapport (ici, valeur médiane) de CF hétérozygote, CF homozygote ou VAF. Les lignes pointillées indiquent $R = 1$ pour une absence de réponse, $R > 1$ pour une réponse négative et $R < 1$ pour une réponse positive. $R < 0.5$ correspond à une réponse moléculaire partielle (PMR), et $R \approx 0$ correspond à une réponse moléculaire complète (CMR). Les lignes de couleur correspondent à la médiane calculée par sous-population. R diffère significativement entre les HSC $CALR^m$ hétérozygotes selon le dosage (test U de Mann-Whitney, p=0.0087) et selon si les $CALR^m$ sont de type 1 ou de type 2 (p=0.0162). R diffère significativement entre les HSC hétérozygotes et homozygotes $JAK2^{V617F}$ (p=0.0047). Il y a une tendance à avoir des valeurs de R différentes pour la réponse des HSC hétérozygotes de patients $JAK2^{V617F}$ selon si le dosage était faible *vs* fort (p=0.0745). La VAF globale de $JAK2^{V617F}$ dans les HSC diffère significativement entre les patients traités par des dosages élevés *vs* faibles d'IFNα (p=0.0288). Les dosages élevés par rapport aux dosages faibles sont définis en fonction du dosage médian des groupes de patients considérés. Le seuil est calculé automatiquement pour comparer deux sous-groupes de patients de même taille. Les seuils de dosage de l'IFNα sont de 78 μg/semaine pour les $CALR^m$ hétérozygotes, 96.5 μg/semaine pour les HSC $JAK2^{V617F}$, 96 μg/semaine pour les HSC $JAK2^{V617F}$ hétérozygotes et 108 μg/semaine pour les HSC $JAK2^{V617F}$ homozygotes. Les différences statistiques ont été calculées à l'aide du test U de Mann Whitney ; *p<0.05, **p<0.01.

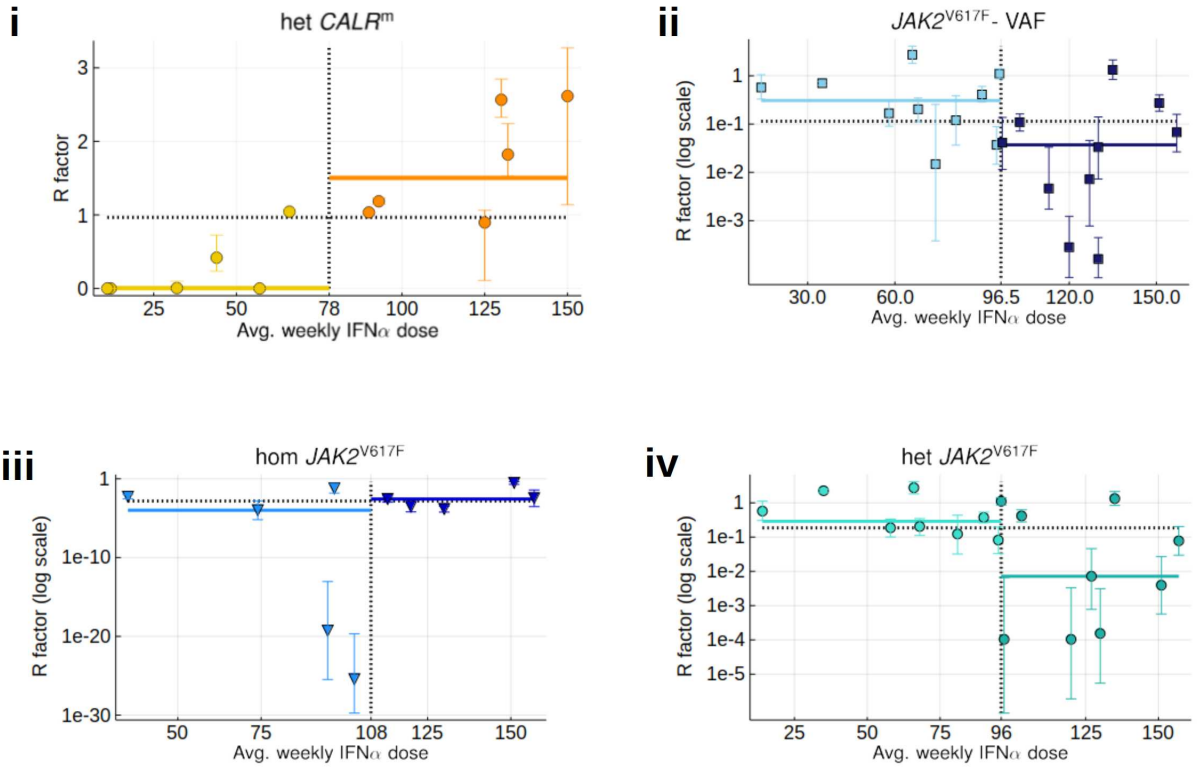


FIGURE 25 – Le facteur de réponse moléculaire (R) a été calculé à $t = 3,000$ jours pour chaque patient et tracé en fonction du dosage reçu par le patient pour (i) les HSC hétérozygotes $CALR^m$; (ii) les HSC $JAK2^{V617F}$ (hétérozygotes et homozygotes, VAF) ; (iii) les HSC homozygotes $JAK2^{V617F}$ et (iv) les HSC hétérozygotes $JAK2^{V617F}$. Les points représentent, pour chaque patient, la valeur médiane du R-factor (après propagation des incertitudes). Nous avons observé une tendance significative à une plus mauvaise réponse (R>1) pour les HSC hétérozygotes $CALR^m$ corrélée à une augmentation du dosage (p<1e-4 en testant la nullité du coefficient de régression linéaire). Au contraire, on estime une tendance à une meilleure réponse ($R \to 0$) pour les HSC $JAK2^{V617F}$ (en termes de VAF) lorsque le dosage est plus élevé (p=0.0747 en testant la nullité du coefficient de régression linéaire). Il n'y a pas d'effet du dosage mais dans tous les cas une très bonne réponse au niveau des HSC homozygotes $JAK2^{V617F}$, alors qu'il y a une tendance à avoir une meilleure réponse au niveau des HSC hétérozygotes $JAK2^{V617F}$ en augmentant le dosage (p=0.0498 en testant la nullité du coefficient de régression linéaire). Les barres d'erreur représentent les intervalle de crédibilité à 95%. Les lignes pointillées verticales délimitent les groupes de patients en deux sous-groupes de même taille, selon qu'ils ont reçu un dosage inférieur ou supérieur au dosage médian. Les lignes pointillées noires horizontales représentent la valeur médiane du R-factor pour tous les patients du groupe et les lignes pleines colorées horizontales représentent les valeurs médianes du R-factor calculées parmi les patients de chaque sous-groupe.

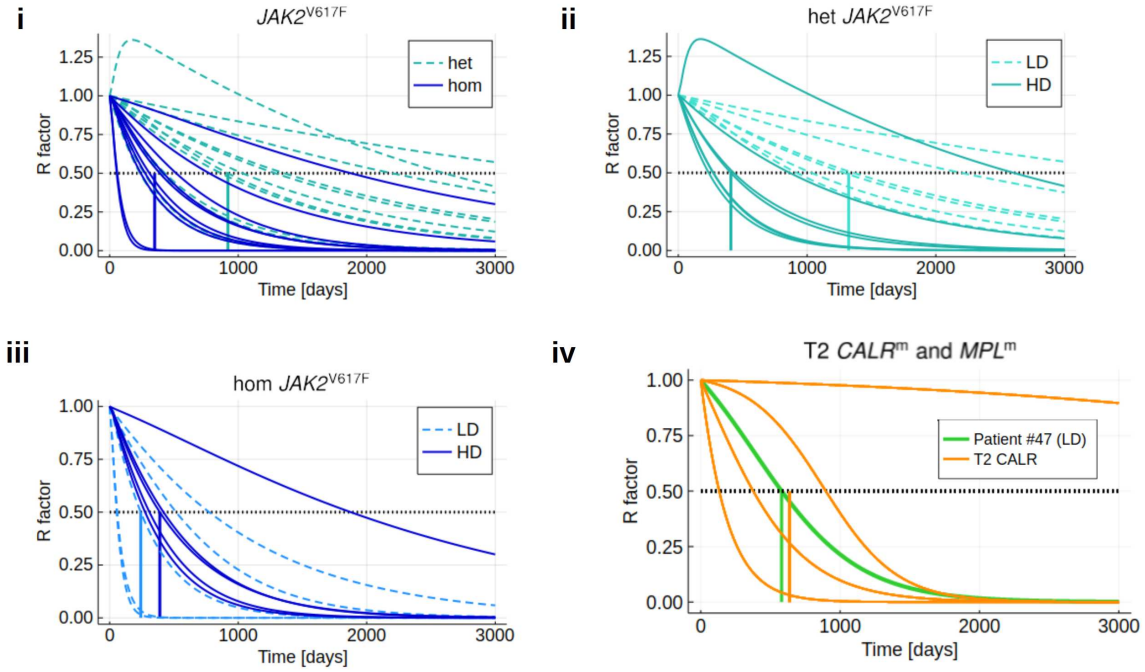


FIGURE 26 – Dynamique de réponse au traitement (évolution du R-factor) inférée au cours du temps pour les populations de (i) HSC $JAK2^{V617F}$ hétérozygotes *vs* homozygotes ; (ii) HSC $JAK2^{V617F}$ hétérozygotes chez les patients traités avec des dosages élevés *vs* faibles d'IFNα ; (iii) HSC $JAK2^{V617F}$ homozygotes chez les patients traités avec des dosages élevés *vs* faibles d'IFNα et (iv) HSC MPL^m hétérozygotes et HSC $CALR^m$ de type 2. Seuls les patients répondeurs sont considérés ici. Le temps médian pour atteindre une diminution de 50% du facteur de réponse R (lignes horizontales en tirets noirs) a été calculé pour chaque patient (et clone). Les HSC homozygotes $JAK2^{V617F}$, avec un temps médian pour atteindre une diminution de 50% du R-factor (PMR) de 350 jours, ont été ciblées plus rapidement que les hétérozygotes ($T_{PMR} = 920$ jours). Le dosage a eu un effet significatif (test U de Mann-Whitney, p=0.0221) sur le T_{PMR} des patients hétérozygotes (408 jours pour les hauts dosages contre 1323 jours pour les faibles dosages). Aucune différence significative n'a été trouvée pour les répondeurs homozygotes avec un T_{PMR} médian égal à 242 et 390 jours pour les LD (Low Dosage) et les HD (High Dosage), respectivement.

6.3 Effet de l'IFNα

Le modèle mathématique construit pour décrire la dynamique des HSPC vise également à fournir des indications sur le mécanisme d'action de l'IFNα. Les deux paramètres critiques introduits dans le modèle pour décrire les effets de l'IFNα sont Δ^* et γ^* : Δ^*_{het} et γ^*_{het} pour les cellules hétérozygotes et Δ^*_{hom} et γ^*_{hom} pour les cellules homozygotes. Le premier paramètre Δ^* modélise la balance entre divisions différenciées et symétriques des HSC mutées sous IFNα et devient négatif si les divisions génèrent plus de progéniteurs que de cellules souches, conduisant finalement à l'épuisement du *pool* de HSC mutées. Le deuxième paramètre γ^* décrit la vitesse à laquelle les HSC mutées quiescentes deviennent actives (le taux de sortie de quiescence). Une valeur élevée de γ^* (par rapport à la valeur choisie $\gamma = 1/300$ [j^{-1}]) signifie que davantage de HSC mutées vont être actives dans les premières semaines du traitement et disponibles pour contribuer à l'hématopoïèse.
Chez les patients $JAK2^{V617F}$, nous avons estimé que l'IFNα induit des valeurs négatives pour Δ^*_{het} et Δ^*_{hom} (-0.20 et -0.29, respectivement, en moyenne sur les patients) (Fig. 27-i et 28-i), indiquant - d'après notre modèle - que l'IFNα favoriserait les divisions différenciées des HSC mutées conduisant à l'épuisement de ce compartiment.

Les valeurs Δ^* se sont avérées différer significativement entre les patients $JAK2^{V617F}$ et ceux $CALR^m$, suggérant un mécanisme d'action de l'IFNα reposant sur la déplétion des HSC $JAK2^{V617F}$ mais pas celles mutées $CALR^m$ (Fig. 27-i et 28i, iii).
Il faut néanmoins rappeler que nous avons utilisé un modèle différent, à un seul compartiment souche, dans le cas des patients $CALR^m$, pouvant en partie expliquer ce résultat.
Nous estimons cependant une différence significative entre les valeurs Δ^*_{het} estimées dans le cas des patients $CALR^m$ de type 2 *vs* type 1 (Fig. 27-ii).
Enfin, l'IFNα favoriserait davantage la sortie de quiescence des HSC homozygotes $JAK2^{V617F}$ par rapport aux HSC hétérozygotes $JAK2^{V617F}$, comme l'indiquent les valeurs estimées de γ^* (Fig. 27-iii et 28-ii).
Dans l'ensemble, nous avons estimé que l'IFNα entraîne une déplétion des HSC $JAK2^{V617F}$ et $CALR^m$ de type 2 en diminuant leur tendance à l'auto-renouvellement. De plus, l'IFNα favoriserait la sortie de quiescence des HSC homozygotes $JAK2^{V617F}$ mais pas celles hétérozygotes.

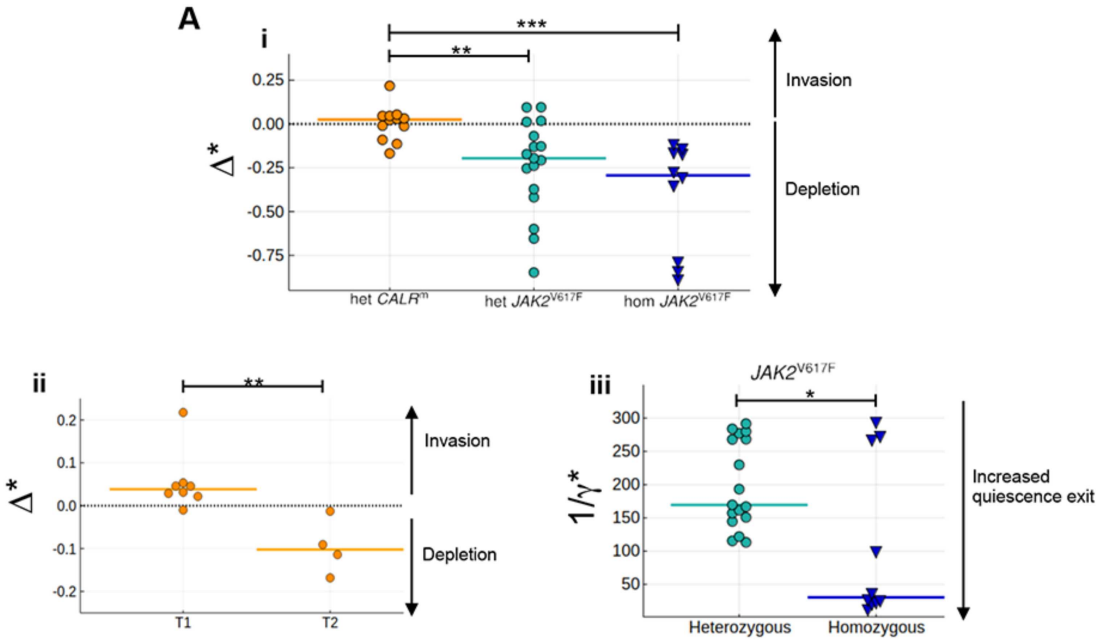


FIGURE 27 – Valeurs estimées des paramètres du modèle (moyennes *a posteriori*). (i) Les paramètres Δ^* ont été calculés pour les populations de HSC hétérozygotes et homozygotes $JAK2^{V617F}$ et pour les populations de HSC hétérozygotes $CALR^m$. Δ^*_{het} diffère significativement chez les patients présentant un NMP $CALR^m$ par rapport à ceux présentant un NMP $JAK2^{V617F}$ (test U de Mann-Whitney, p=0.0031). Δ^*_{het} chez les patients atteints du NMP $CALR^m$ est significativement différent de Δ^*_{hom} chez ceux atteints d'un NMP $JAK2^{V617F}$ (p<0.0001). (ii) Les paramètres Δ^*_{het} ont été calculés pour les populations de HSC $CALR^m$ de type 1 et de type 2. Δ^*_{het} diffère significativement chez les patients atteints ayant une mutation $CALR^m$ de type 1 par rapport à ceux de type 2 (p=0.004). (iii) Les rapports inverses du paramètre γ^* ont été calculés dans les populations de HSC hétérozygotes *vs* homozygotes $JAK2^{V617F}$. La valeur $1/\gamma^*$ peut être considérée comme un temps relatif passé par les cellules dans le compartiment inactif de notre modèle. Le rapport $1/\gamma^*$ diffère significativement entre les populations hétérozygotes *vs* homozygotes $JAK2^{V617F}$ (p=0.027).

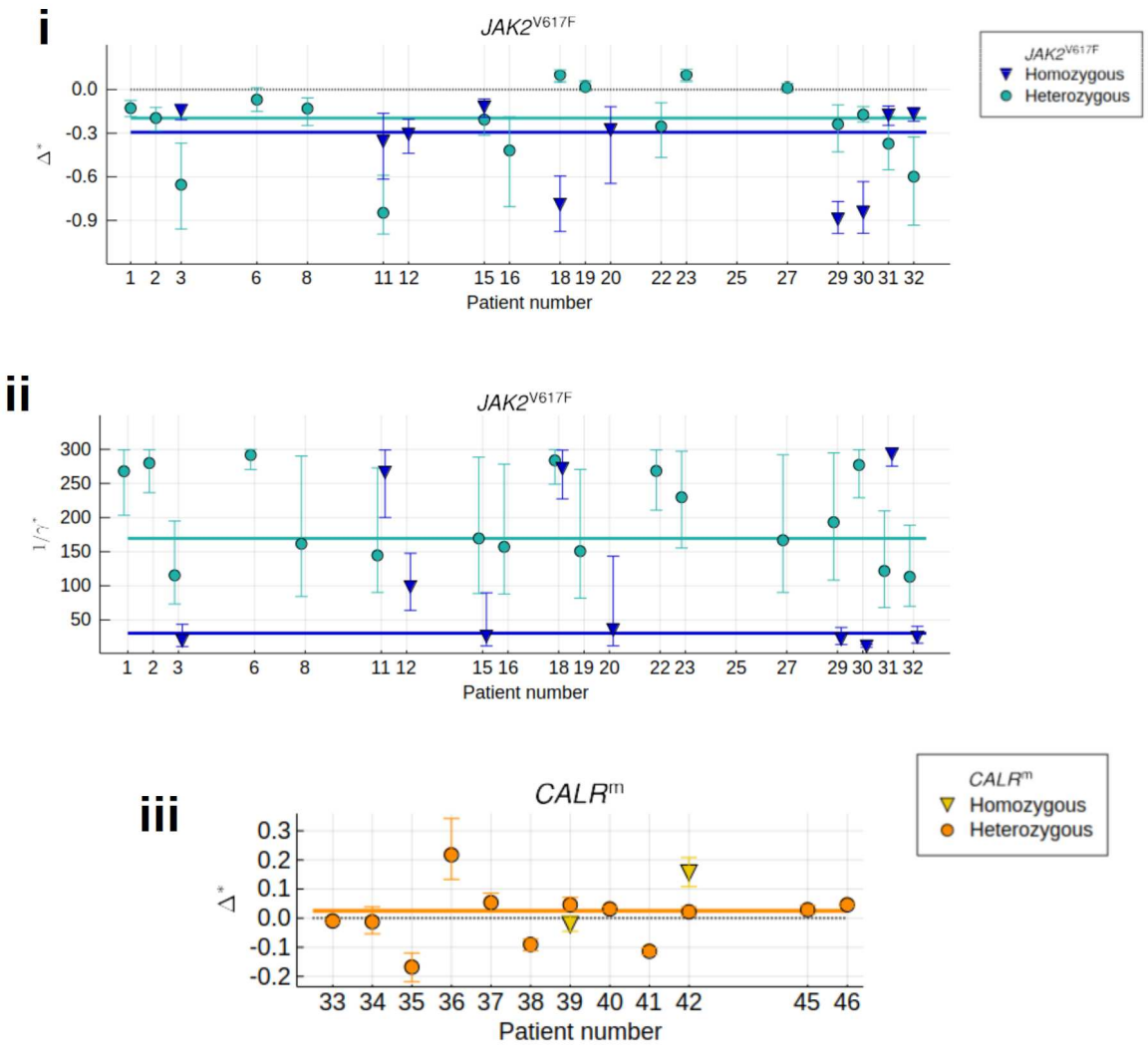


FIGURE 28 – Paramètres estimés (valeurs moyennes et intervalles de crédibilité à 95%) pour chaque patient. (i) Δ^*_{hom} et Δ^*_{het} pour les HSC homozygotes et hétérozygotes de $JAK2^{V617F}$ estimés pour chaque patient. Une tendance pour les cellules homozygotes à subir plus de divisions différenciées que les cellules hétérozygotes sous IFNα a été estimée. Les cercles verts sont les valeurs moyennes pour les HSC hétérozygotes et les triangles bleus sont les valeurs moyennes pour les HSC homozygotes, accompagnées d'un intervalle de crédibilité de 95% ; (ii) les rapports inverses des paramètres γ^*_{het} et γ^*_{hom} sont estimés pour chaque patient $JAK2^{V617F}$. Cet inverse peut être vu comme le temps passé par les cellules mutées dans le compartiment inactif de notre modèle (ces temps doivent être considérés relativement à la valeur fixe $\gamma = 1/300$). Les cercles verts sont les valeurs moyennes pour les HSC hétérozygotes et les triangles bleus sont les valeurs moyennes pour les HSC homozygotes, avec un intervalle de crédibilité de 95% ; (iii) Δ^*_{hom} et Δ^*_{het} pour les HSC $CALR^m$. Les cercles orange sont les valeurs moyennes pour les HSC hétérozygotes et les triangles jaunes sont les valeurs moyennes pour les HSC homozygotes, avec un intervalle de crédibilité de 95%.

7 Discussion

L'analyse prospective et longitudinale de patients NMP traités par IFNα sur une période de 5 ans, sur laquelle se base le travail présenté dans ce chapitre, a permis de générer un dataset qui, combiné à de la modélisation mathématique et de l'inférence bayésienne hiérarchique, nous a permis d'identifier des réponses moléculaires différentielles entre NMP $JAK2^{V617F}$, MPL^m et $CALR^m$.

Nous avons commencé par analyser les données expérimentales sans l'intermédiaire d'un modèle. La réponse globale dans les granulocytes montre une diminution de la VAF dans le cas $JAK2^{V617F}$ en accord avec la littérature [1, 2, 6, 7]. De plus, nos observations ont confirmé l'hétérogénéité dans la réponses des patients $CALR^m$. Globalement, les NMP $CALR^m$ sont moins susceptibles que les NMP $JAK2^{V617F}$ de présenter une diminution de la VAF dans les cellules matures [6, 10, 11, 12, 37, 38]. Ces observations au niveau des cellules matures se confirment lorsque l'on étudie la réponse au niveau des cellules immatures où nous observons généralement de meilleures réponses dans les progéniteurs $JAK2^{V617F}$ que dans les $CALR^m$. Ces résultats contrastent avec la réponse hématologique améliorée à l'IFNα et la survie plus longue des patients $CALR^m$ que des patients $JAK2^{V617F}$ dans plusieurs études [6, 37, 39], qui pourrait s'expliquer parce que l'IFNα a un effet délétère général pendant la mégacaryopoïèse, ce qui limite la thrombocytose dans une maladie limitée à la lignée mégacaryocytaire [40].

Notre hypothèse, pour expliquer la dynamique de réponse au traitement, est que l'IFNα ciblerait les cellules souches mutées - cellules initiatrices du développement des NMP sur plusieurs décennies comme étudié au chapitre précédent - induisant une dynamique latente non observée parmi le compartiment souche, qui se répercuterait ensuite sur les compartiments immature et mature. Cette hypothèse repose notamment sur des résultats obtenus dans des modèles murins $JAK2^{V617F}$, où il a été démontré que l'IFNα favorise la prolifération et la sortie de quiescence des HSC mutées, entraînant une augmentation du nombre total de progéniteurs [16, 41, 42].
Pour vérifier si les observations expérimentales pouvaient être en accord avec cette hypothèse, nous avons proposé un modèle compartimental décrivant la dynamique de populations de cellules souches, immatures et matures pouvant être wild-type, mutées hétérozygotes ou homozygotes. Suivant l'idée de Michor et al. [17], nous avons supposé que le traitement induisait une nouvelle dynamique en changeant la valeur de certains des paramètres du modèle.
Nous avons alors estimé pour chaque patient de la cohorte la valeur de ces paramètres impactés par l'IFNα, pour ensuite comparer les estimations entre patients et groupes de patients. Pour plus de robustesse, nous avons utilisé une méthode d'inférence Bayésienne hiérarchique et considéré différentes populations de patients suivant la mutation motrice du NMP.
Les dynamiques inférées par notre modèle pour les compartiments des progéniteurs et des granulocytes s'ajustaient bien aux données expérimentales. Nous pouvions alors étudier la dynamique latente inférée au niveau des cellules souches mutées hétérozygotes ou homozygotes, et en particulier estimer la réponse à long-terme. En comparant les réponses des clones hétérozygotes et homozygotes des patients, suivant qu'ils étaient mutés $JAK2^{V617F}$, MPL^m ou $CALR^m$ type 1 et 2, et suivant le dosage (la dose moyenne sur les 450 premiers jours de traitement) d'IFNα qui leur avait été administré, nous avons pu mettre en évidence des différences quant à la réponse moléculaire au niveau des HSC.
Concernant la réponse en fonction du génotype, nous avons estimé que l'IFNα ciblait efficacement la plupart des HSC $JAK2^{V617F}$, et préférentiellement celles homozygotes $JAK2^{V617F}$, en accord avec ce qui avait pu être montré au niveau des cellules matures [43]. Bien que les patients soient peu nombreux, nos résultats suggèrent que l'IFNα ciblerait préférentiellement les HSC $CALR^m$ de type 2 que celles de type 1.
Concernant la réponse en fonction du dosage administré, nous avons estimé que des dosages élevés d'IFNα seraient plus efficaces que des dosages faibles pour cibler les HSC hétérozygotes $JAK2^{V617F}$. Par contre, pour les HSC $CALR^m$ hétérozygotes, des dosages d'IFNα plus élevés étaient corrélés à une moins bonne réponse. Même si corrélation ne signifie pas nécessairement

causalité, nos données indiquent qu'il n'y a pas de preuve pour soutenir un traitement à haute dose pour obtenir une réponse moléculaire à long terme chez les patients $CALR^m$.
Nous n'avons pas mis en évidence, pour les patients $JAK2^{V617F}$ (qui sont les plus représentés dans notre cohorte), que le type de maladie, l'âge, ou le nombre de mutations associées (1, 2 ou 3) au début du traitement pourraient avoir un impact sur la dynamique des HSC. Néanmoins, notre étude a une pertinence limitée s'agissant de comprendre les effets potentiels des mutations de *TET2*, *DNMT3A* ou *TP53* [2] car très peu de patients de notre cohorte présentaient de telles mutations. Notre capacité à déduire les effets des polymorphismes génétiques impliqués dans la réponse à l'IFNα, tels que *IFNL4*, est limitée de par leurs effets modérés [44].
En perspective, il serait intéressant d'étudier la dynamique clonale pour les patients qui présentent une CF suffisamment élevée pour certaines mutations associées aux NMP (par exemple de *TET2*), pour comprendre si les mutations associées ont un impact sur la réponse moléculaire au traitement ou si l'IFNα peut induire la sélection de certains clones mutés qui pourraient induire une évolution péjorative de la maladie. Notre modèle, de même qu'il permet de faire la distinction entre les cellules suivant leur zigosité, pourrait facilement être étendu pour faire la distinction entre les cellules ayant ou non certaines mutations associées.

Après avoir quantifié, par l'intermédiaire de l'analyse de la dynamique latente inférée au niveau des HSC, la réponse à l'IFNα en fonction de certains déterminants, nous avons cherché à l'interpréter par l'analyse des paramètres estimés. L'hypothèse sur laquelle se base la construction de notre modèle mathématique est celle d'un effet de l'IFNα au niveau de la différenciation des cellules mutées (effet sur le paramètre noté Δ^*) ainsi que, dans le cas $JAK2^{V617F}$ et MPL^m où nous introduisons une distinction entre HSC actives et quiescentes, un effet sur la sortie de quiescence des HSC mutées (paramètre noté γ^*). Sous ces hypothèses de travail, nous avons alors pu étudier dans quelle mesure les estimations des paramètres de notre modèle pouvaient différer suivant qu'ils concernaient des cellules mutées hétérozygotes *vs* homozygotes. Nous avons estimé que l'IFNα épuisait lentement les HSC homozygotes et hétérozygotes $JAK2^{V617F}$ avec des demi-vies (T_{PMR}) d'environ 12 et 31 mois, respectivement. Le traitement favoriserait préférentiellement la sortie de quiescence des HSC homozygotes $JAK2^{V617F}$, par rapport aux hétérozygotes. Nos résultats sont cohérents avec ceux observés par Pedersen et al. [45] sur une cohorte indépendante de taille similaire à la nôtre. Pedersen et al. ont récemment rapporté que, chez les répondeurs au traitement par IFNα, la VAF $JAK2^{V617F}$ diminuait dans les granulocytes avec une demi-vie typique comprise entre un et deux ans. Ils ont également observé qu'une augmentation initiale des granulocytes mutés précédait les taux de diminution les plus rapides comme nous avons également observé chez plusieurs patients une augmentation temporaire de la fréquence des progéniteurs homozygotes $JAK2^{V617F}$ suivie d'une diminution (effet courbe en cloche). Ainsi, une augmentation de la VAF en début du traitement ne serait pas nécessairement associée à une mauvaise réponse moléculaire à long terme. Ces résultats s'accordent également avec une augmentation précédemment observée des progéniteurs dans la moelle osseuse de patients $JAK2^{V617F}$ peu après le traitement par IFNα [46]. Dans le cas de patients $CALR^m$, l'IFNα favoriserait la différenciation des HSC en progéniteurs seulement pour un faible nombre de patients, principalement ceux de type 2.

Le travail présenté dans ce chapitre a plusieurs limites. Tout d'abord, notre modèle repose sur l'hypothèse d'un effet de l'IFNα sur la différenciation et sortie de quiescence des cellules souches mutées, hypothèse dont nous avons montré qu'elle pouvait être en accord avec les données expérimentales à notre disposition. Néanmoins, il n'est pas exclu que d'autres modèles, construits sur des hypothèses alternatives, permettent également de retrouver les dynamiques observées au niveau des progéniteurs et cellules matures.
Contrairement à nous, Tong et al. [47] ont par exemple observé - à partir d'une analyse Target-Seq après traitement par IFNα - une augmentation de la quiescence dans les HSPC homozygotes et de l'apoptose dans les HSPC hétérozygotes $JAK2^{V617F}$. Une limite potentielle de l'étude de Tong et al. est qu'elle ne repose pas sur des observations longitudinales. Il est possible que de

telles voies se retrouvent dans les HSPC $JAK2^{V617F}$ non ciblées restantes à un moment ultérieur ou que plusieurs mécanismes tels que la sénescence ou l'apoptose coopèrent pour renforcer l'épuisement des HSC [46, 41, 47, 48, 49].
De plus, pour permettre de reproduire le phénomène de courbe en cloche observé chez certains patients $JAK2^{V617F}$, nous avons proposé, par rapport au modèle utilisé pour le cas $CALR^m$, de séparer le compartiment des HSC en deux, afin notamment d'étudier l'hypothèse selon laquelle l'IFNα favoriserait la sortie de quiescence des cellules mutées [16]. Une hypothèse alternative serait que cet effet courbe en cloche soit dû à l'arrêt d'un traitement à l'Hydrea utilisé avant traitement à l'IFNα. En perspective, nous prévoyons de tester cette hypothèse alternative par la construction d'un modèle qui reposerait - comme pour les $CALR^m$ - sur un unique compartiment pour les cellules souches.
Ainsi, nous avons proposé dans ce chapitre une interprétation possible du mécanisme d'action de l'IFNα au niveau des HSC mutées, sans être en mesure d'exclure certaines hypothèses alternatives. Pour continuer ce travail, il serait intéressant de proposer différents modèles, construits sur des hypothèses biologiques alternatives ou faisant intervenir des mécanismes biologiques plus complexes, et d'en sélectionner celui qui semble le meilleur sur la base de critères tels que les critères AIC ou BIC étudiés par exemple au chapitre 3. Néanmoins, l'utilisation seule de modèles et critères mathématiques pour inférer la dynamique latente non observée des HSC mutées sous IFNα ne saurait être suffisante. Pour élucider le mécanisme d'action de l'IFNα, d'autres expériences seront ainsi nécessaires. Il serait ainsi essentiel d'aborder le mécanisme de signalisation exact de l'épuisement des HSC sous l'effet de l'IFNα. Une augmentation des ROS, une suractivation de P53 et STAT1 ainsi qu'un amorçage spécifique de la signalisation IFNα ont été rapportés avec $JAK2^{V617F}$ [41] mais pas avec $CALR^m$ [12]. En revanche, $CALR^m$, mais pas $JAK2^{V617F}$, est sécrété et pourrait déréguler la réponse immunitaire induite par l'IFNα [50]. Notre étude suggère également que les répondeurs moléculaires les plus faibles sont les patients $CALR^m$ de type 1, potentiellement parce que ce type de mutation dérégule des voies de signalisation spécifiques qui conduisent à une amplification plus forte au niveau des HSC par rapport au type 2 [18, 51] (voir également l'étude menée au chapitre 2).

Notre modèle actuel repose également sur de nombreuses hypothèses simplificatrices. Nous avons ainsi négligé l'hétérogénéité de l'ensemble du compartiment des progéniteurs ($CD34^+$), ce qui est une simplification, comme nous avions pu le montrer aux chapitres 2, 3 et 4. Le modèle pourrait par exemple être complexifié en considérant un continuum d'états, des cellules souches aux progéniteurs, ce qui conduirait à un système d'équations aux dérivées partielles, avec l'introduction d'une variable de structure, en l'occurrence le degré de différenciation, comme par exemple étudié de façon théorique par Adimy et al. [20]. On pourrait également envisager l'introduction de mécanismes de régulation, comme le font par exemple Marciniak-Czochra [52] (voir également Annexe C), résultant en des systèmes d'ODE non linéaires. Plus simplement, on pourrait séparer le compartiment des progéniteurs en trois, en faisant la distinction entre les HSC*, MPP et HPC, comme nous l'avions fait au chapitre 4. Le modèle de la dynamique de prolifération et différenciation étudié dans ce chapitre-là a justement pour objectif de nous permettre de raffiner notre modèle, et de rationaliser le choix de certaines hypothèses. En effet, pour avoir un modèle identifiable, nous avons fait jusqu'à présent plusieurs choix sur les paramètres qui nous semblaient devoir différer en fonction du statut mutationnel (wt, het ou hom) et en fonction de si on est avant ou après traitement.
En étudiant plus précisément la dynamique des HSPC, même *in vitro*, comme effectué au chapitre 4, avec cette fois-ci non seulement des cellules WT mais également des cellules mutées avec et sans IFNα, nous devrions être en mesure de raffiner nos hypothèses.
Il restera néanmoins toujours difficile de proposer des valeurs numériques pour les paramètres associés à la dynamique des cellules souches HSC, pour lesquelles les résultats actuels reposent principalement sur des études de modèles murins, même si l'essor des techniques de barcoding [53] ou encore des analyses phylogénétiques [54, 55] ouvre la voie à une meilleure compréhension du comportement des cellules souches hématopoïétiques.

Dans l'étude présentée dans ce chapitre, nous avons également négligé les variations de dose d'IFNα en cours de traitement, et fait l'hypothèse d'un effet unique et durable de l'IFNα à partir du début du traitement. Nous n'avons pas non plus introduit directement le dosage, défini comme la dose moyenne reçue sur les 450 jours, dans l'estimation des paramètres (ni d'autres variables caractérisant les patients tels que l'âge, le sexe, la maladie) comme cela peut être fait dans des modèles à effet mixte [56]. Les paramètres propres aux patients, incluant le dosage reçu, n'étaient utilisés qu'à posteriori pour étudier la réponse au traitement en séparant les patients en différents groupes. Le choix de moyenner les doses sur 450 jours pourrait également être remis en question. Les sous-groupes de patients constitués en fonction de s'ils ont reçu un faible ou fort dosage pourraient différer suivant ce choix. Nous avons néanmoins vérifié que nos conclusions n'étaient pas sensibles à ce choix. Cependant, négliger les variations de dose est une hypothèse forte, sachant que les cliniciens procèdent généralement à une augmentation des doses suivie d'une désescalade, et que certains patients subissent des interruptions de traitement. En particulier, il est difficilement envisageable de prédire l'effet à long-terme du traitement sans prendre en considération les changements de dose qui peuvent survenir. Au chapitre suivant, nous présenterons une extension de notre modèle, dans le cas des patients $JAK2^{V617F}$, qui prend en compte ces variations de dose.

Finalement, le travail présenté dans ce chapitre permet déjà de mettre en évidence des différences quant à l'effet de l'IFNα en fonction du type de mutation motrice du NMP, de la zigosité, et du dosage. Si les interprétations du mécanisme d'action de l'IFNα au niveau des HSC doivent être soumises à caution, les résultats quand à la stratification des patients sont robustes. Notre modèle dans ce cas est en effet utilisé comme moyen d'extrapoler les données expérimentales à long-terme et de prendre en compte l'incertitude sur les mesures. Les résultats obtenus concernant la réponse moléculaire au niveau des HSC (estimés par l'intermédiaire du facteur de réponse R) sont ainsi faiblement sensibles aux choix de modélisation. Nous avons notamment des résultats qui sont cohérents avec l'analyse directe des données, mais également en accord avec un modèle alternatif étudié par Robert Noble (non publié).
À ce jour, l'administration actuelle de l'IFNα en pratique clinique n'est ni standardisée ni adaptée à la mutation motrice de NMP et les stratégies de traitement sont principalement guidées par la réponse hématologique et la tolérabilité. Notre étude propose pour la première fois des recommandations cliniques basées sur la réponse moléculaire au niveau des HSC. Nos résultats suggèrent que le titrage de la dose jusqu'à la dose maximale tolérée serait une stratégie plus susceptible de permettre une réduction de la quantité de HSC mutées $JAK2^{V617F}$ et donc d'atteindre une réponse hématologique et moléculaire. Par conséquent, l'intensité de la dose doit être idéalement maintenue pendant le traitement. Dans l'ensemble, cette étude ouvre de nouvelles voies de recherche visant à comprendre les effets différentiels précis de $JAK2^{V617F}$ et de $CALR^{m}$ sur les HSC à l'origine des NMP.

Références

[1] Jean-Jacques Kiladjian, Bruno Cassinat, Sylvie Chevret, Pascal Turlure, Nathalie Cambier, Murielle Roussel, Sylvia Bellucci, Bernard Grandchamp, Christine Chomienne, and Pierre Fenaux. Pegylated interferon-alfa-2a induces complete hematologic and molecular responses with low toxicity in polycythemia vera. *Blood, The Journal of the American Society of Hematology*, 112(8) :3065–3072, 2008.

[2] Alfonso Quintás-Cardama, Omar Abdel-Wahab, Taghi Manshouri, Outi Kilpivaara, Jorge Cortes, Anne-Laure Roupie, Su-Jiang Zhang, David Harris, Zeev Estrov, Hagop Kantarjian, et al. Molecular analysis of patients with polycythemia vera or essential thrombocythemia receiving pegylated interferon α-2a. *Blood, The Journal of the American Society of Hematology*, 122(6) :893–901, 2013.

[3] Thomas Stauffer Larsen, Katrine F Iversen, Esben Hansen, Anders Bruun Mathiasen, Claus Marcher, Mikael Frederiksen, Herdis Larsen, Inge Helleberg, Caroline Hasselbalch Riley, Ole W Bjerrum, et al. Long term molecular responses in a cohort of danish patients with essential thrombocythemia, polycythemia vera and myelofibrosis treated with recombinant interferon alpha. *Leukemia research*, 37(9) :1041–1045, 2013.

[4] Lucia Masarova, C Cameron Yin, Jorge E Cortes, Marina Konopleva, Gautam Borthakur, Kate J Newberry, Hagop M Kantarjian, Carlos E Bueso-Ramos, and Srdan Verstovsek. Histomorphological responses after therapy with pegylated interferon α-2a in patients with essential thrombocythemia (et) and polycythemia vera (pv). *Experimental hematology & oncology*, 6(1) :1–13, 2017.

[5] RichardT Silver. Recombinant interferon-alpha for treatment of polycythaemia vera. *The Lancet*, 332(8607) :403, 1988.

[6] Abdulraheem Yacoub, John Mascarenhas, Heidi Kosiorek, Josef T Prchal, Dmitry Berenzon, Maria R Baer, Ellen Ritchie, Richard T Silver, Craig Kessler, Elliott Winton, et al. Pegylated interferon alfa-2a for polycythemia vera or essential thrombocythemia resistant or intolerant to hydroxyurea. *Blood*, 134(18) :1498–1509, 2019.

[7] Heinz Gisslinger, Christoph Klade, Pencho Georgiev, Dorota Krochmalczyk, Liana Gercheva-Kyuchukova, Miklos Egyed, Viktor Rossiev, Petr Dulicek, Arpad Illes, Halyna Pylypenko, et al. Ropeginterferon alfa-2b versus standard therapy for polycythaemia vera (proud-pv and continuation-pv) : a randomised, non-inferiority, phase 3 trial and its extension study. *The Lancet Haematology*, 7(3) :e196–e208, 2020.

[8] John Mascarenhas, Heidi E Kosiorek, Josef T Prchal, Alessandro Rambaldi, Dmitriy Berenzon, Abdulraheem Yacoub, Claire N Harrison, Mary Frances McMullin, Alessandro M Vannucchi, Joanne Ewing, et al. A randomized phase 3 trial of interferon-α vs hydroxyurea in polycythemia vera and essential thrombocythemia. *Blood, The Journal of the American Society of Hematology*, 139(19) :2931–2941, 2022.

[9] T Barbui, AM Vannucchi, V De Stefano, A Masciulli, A Carobbio, A Ghirardi, F Ciceri, M Bonifacio, A Iurlo, F Palandri, et al. Phase ii randomized clinical trial comparing ropeginterferon versus phlebotomy in low-risk patients with polycythemia vera. results of the pre-planned interim analysis. *Hemasphere*, 4 :2602, 2020.

[10] Emmanuelle Verger, Bruno Cassinat, Aurélie Chauveau, Christine Dosquet, Stephane Giraudier, Marie-Hélène Schlageter, Jean-Christophe Ianotto, Mohammed A Yassin, Nader Al-Dewik, Serge Carillo, et al. Clinical and molecular response to interferon-α therapy in essential thrombocythemia patients with calr mutations. *Blood, The Journal of the American Society of Hematology*, 126(24) :2585–2591, 2015.

[11] Lasse Kjær, Sabrina Cordua, Morten O Holmström, Mads Thomassen, Torben A Kruse, Niels Pallisgaard, Thomas S Larsen, Karin De Stricker, Vibe Skov, and Hans C Hasselbalch. Differential dynamics of calr mutant allele burden in myeloproliferative neoplasms during interferon alfa treatment. *PLoS One*, 11(10) :e0165336, 2016.

[12] Julia Czech, Sabrina Cordua, Barbora Weinbergerova, Julian Baumeister, Assja Crepcia, Lijuan Han, Tiago Maié, Ivan G Costa, Bernd Denecke, Angela Maurer, et al. Jak2v617f but not calr mutations confer increased molecular responses to interferon-α via jak1/stat1 activation. *Leukemia*, 33(4) :995–1010, 2019.

[13] Matthieu Mosca, Gurvan Hermange, Amandine Tisserand, Robert Noble, Christophe Marzac, Caroline Marty, Cécile Le Sueur, Hugo Campario, Gaëlle Vertenoeil, Mira El-Khoury, et al. Inferring the dynamics of mutated hematopoietic stem and progenitor cells induced by ifnα in myeloproliferative neoplasms. *Blood, The Journal of the American Society of Hematology*, 138(22) :2231–2243, 2021.

[14] Daniel A Arber, Attilio Orazi, Robert Hasserjian, Jürgen Thiele, Michael J Borowitz, Michelle M Le Beau, Clara D Bloomfield, Mario Cazzola, and James W Vardiman. The 2016 revision to the world health organization classification of myeloid neoplasms and acute leukemia. *Blood, The Journal of the American Society of Hematology*, 127(20) :2391–2405, 2016.

[15] Hugo Campario, Matthieu Mosca, Bernard Aral, Valentin Bourgeois, Pauline Martin, Antoine Brustel, Mathilde Filser, Christophe Marzac, Isabelle Plo, and François Girodon. Impact of interferon on a triple positive polycythemia vera. *Leukemia*, 34(4) :1210–1212, 2020.

[16] Ann Mullally, Claudia Bruedigam, Luke Poveromo, Florian H Heidel, Amy Purdon, Therese Vu, Rebecca Austin, Dirk Heckl, Lawrence J Breyfogle, Catherine Paine Kuhn, et al. Depletion of jak2v617f myeloproliferative neoplasm-propagating stem cells by interferon-α in a murine model of polycythemia vera. *Blood, The Journal of the American Society of Hematology*, 121(18) :3692–3702, 2013.

[17] Franziska Michor, Timothy P Hughes, Yoh Iwasa, Susan Branford, Neil P Shah, Charles L Sawyers, and Martin A Nowak. Dynamics of chronic myeloid leukaemia. *Nature*, 435(7046) :1267–1270, 2005.

[18] Mira El-Khoury, Xénia Cabagnols, Matthieu Mosca, Gaëlle Vertenoeil, Christophe Marzac, Fabrizia Favale, Olivier Bluteau, Florence Lorre, Amandine Tisserand, Graciela Rabadan Moraes, et al. Different impact of calreticulin mutations on human hematopoiesis in myeloproliferative neoplasms. *Oncogene*, 39(31) :5323–5337, 2020.

[19] Sandra N Catlin, Janis L Abkowitz, and Peter Guttorp. Statistical inference in a two-compartment model for hematopoiesis. *Biometrics*, 57(2) :546–553, 2001.

[20] Mostafa Adimy, Abdennasser Chekroun, and Tarik-Mohamed Touaoula. Age-structured and delay differential-difference model of hematopoietic stem cell dynamics. *Discrete and Continuous Dynamical Systems-Series B*, 20(9) :27, 2015.

[21] Eric M Pietras, Ranjani Lakshminarasimhan, Jose-Marc Techner, Sarah Fong, Johanna Flach, Mikhail Binnewies, and Emmanuelle Passegué. Re-entry into quiescence protects hematopoietic stem cells from the killing effect of chronic exposure to type i interferons. *Journal of Experimental Medicine*, 211(2) :245–262, 2014.

[22] Marieke AG Essers, Sandra Offner, William E Blanco-Bose, Zoe Waibler, Ulrich Kalinke, Michel A Duchosal, and Andreas Trumpp. Ifnα activates dormant haematopoietic stem cells in vivo. *Nature*, 458(7240) :904–908, 2009.

[23] Dagmar Walter, Amelie Lier, Anja Geiselhart, Frederic B Thalheimer, Sina Huntscha, Mirko C Sobotta, Bettina Moehrle, David Brocks, Irem Bayindir, Paul Kaschutnig, et al. Exit from dormancy provokes dna-damage-induced attrition in haematopoietic stem cells. *Nature*, 520(7548) :549–552, 2015.

[24] Jason Xu, Samson Koelle, Peter Guttorp, Chuanfeng Wu, Cynthia Dunbar, Janis L Abkowitz, and Vladimir N Minin. Statistical inference for partially observed branching processes with application to cell lineage tracking of in vivo hematopoiesis. *The Annals of Applied Statistics*, 13(4) :2091–2119, 2019.

[25] Jason Cosgrove, Lucie SP Hustin, Rob J de Boer, and Leïla Perié. Hematopoiesis in numbers. *Trends in Immunology*, 42(12) :1100–1112, 2021.

[26] Ron Sender and Ron Milo. The distribution of cellular turnover in the human body. *Nature medicine*, 27(1) :45–48, 2021.

[27] Martha R Kirby and Robert E Donahue. Rare event sorting of cd34+ hematopoietic cells. *Annals of the New York Academy of Sciences*, 677(1) :413–416, 1993.

[28] Qian-Lin Hao, Ami J Shah, Flavia T Thiemann, Elzbieta M Smogorzewska, and Gay M Crooks. A functional comparison of cd34+ cd38-cells in cord blood and bone marrow. 1995.

[29] Artémis Llamosi, Andres M Gonzalez-Vargas, Cristian Versari, Eugenio Cinquemani, Giancarlo Ferrari-Trecate, Pascal Hersen, and Gregory Batt. What population reveals about individual cell identity : single-cell parameter estimation of models of gene expression in yeast. *PLoS computational biology*, 12(2) :e1004706, 2016.

[30] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[31] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[32] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6) :721–741, 1984.

[33] Nikolaus Hansen. The cma evolution strategy : a comparing review. *Towards a new evolutionary computation*, pages 75–102, 2006.

[34] Christophe Andrieu and Johannes Thoms. A tutorial on adaptive mcmc. *Statistics and computing*, 18(4) :343–373, 2008.

[35] Nikolaus Hansen. The cma evolution strategy : A tutorial. *arXiv preprint arXiv :1604.00772*, 2016.

[36] Ronan Duchesne, Anissa Guillemin, Fabien Crauste, and Olivier Gandrillon. Calibration, selection and identifiability analysis of a mathematical model of the in vitro erythropoiesis in normal and perturbed contexts. *In silico biology*, 13(1-2) :55–69, 2019.

[37] Jean-Christophe Ianotto, Aurélie Chauveau, Françoise Boyer-Perrard, Emmanuel Gyan, Kamel Laribi, Pascale Cony-Makhoul, Jean-Loup Demory, Benoit de Renzis, Christine Dosquet, Jerome Rey, et al. Benefits and pitfalls of pegylated interferon-α2a therapy in patients with myeloproliferative neoplasm-associated myelofibrosis : a french intergroup of myeloproliferative neoplasms (fim) study. *Haematologica*, 103(3) :438, 2018.

[38] Bruno Cassinat, Emmanuelle Verger, and Jean-Jacques Kiladjian. Interferon alfa therapy in calr-mutated essential thrombocythemia. *New England Journal of Medicine*, 371(2) :188–189, 2014.

[39] Richard T Silver, Ariella C Barel, Elena Lascu, Ellen K Ritchie, Gail J Roboz, Paul J Christos, Attilio Orazi, Duane C Hassane, Wayne Tam, and Nicholas CP Cross. The effect of initial molecular profile on response to recombinant interferon-α (rifnα) treatment in early myelofibrosis. *Cancer*, 123(14) :2680–2687, 2017.

[40] Akiko Yamane, Takanori Nakamura, Hidenori Suzuki, Mamoru Ito, Yasuyuki Ohnishi, Yasuo Ikeda, and Yoshitaka Miyakawa. Interferon-α2b–induced thrombocytopenia is caused by inhibition of platelet production but not proliferation and endomitosis in human megakaryocytes. *Blood, The Journal of the American Society of Hematology*, 112(3) :542–550, 2008.

[41] Rebecca J Austin, Jasmin Straube, Claudia Bruedigam, Gabor Pali, Sebastien Jacquelin, Therese Vu, Joanne Green, Julius Gräsel, Lianne Lansink, Leanne Cooper, et al. Distinct effects of ruxolitinib and interferon-alpha on murine jak2v617f myeloproliferative neoplasm hematopoietic stem cell populations. *Leukemia*, 34(4) :1075–1089, 2020.

[42] Tata Nageswara Rao, Nils Hansen, Jan Stetka, Damien Luque Paz, Milena Kalmer, Julian Hilfiker, Max Endele, Nouraiz Ahmed, Lucia Kubovcakova, Margareta Rybarikova, et al. Jak2-v617f and interferon-α induce megakaryocyte-biased stem cells characterized by decreased long-term functionality. *Blood*, 137(16) :2139–2151, 2021.

[43] S Hasan, B Cassinat, N Droin, JP Le Couedic, F Favale, B Monte-Mor, C Lacout, M Fontenay, C Dosquet, C Chomienne, et al. Use of the 46/1 haplotype to model jak2v617f clonal architecture in pv patients : clonal evolution and impact of ifnα treatment. *Leukemia*, 28(2) :460–463, 2014.

[44] Roland Jäger, Heinz Gisslinger, Elisabeth Fuchs, Edith Bogner, Jelena D Milosevic Feenstra, Jakob Weinzierl, Fiorella Schischlik, Bettina Gisslinger, Martin Schalling, Michael Zörer, et al. Germline genetic factors influence the outcome of interferon-α therapy in polycythemia vera. *Blood*, 137(3) :387–391, 2021.

[45] Rasmus K Pedersen, Morten Andersen, Trine A Knudsen, Zamra Sajid, Johanne Gudmand-Hoeyer, Marc JB Dam, Vibe Skov, Lasse Kjær, Christina Ellervik, Thomas S Larsen, et al. Data-driven analysis of jak2v617f kinetics during interferon-alpha2 treatment of patients with polycythemia vera and related neoplasms. *Cancer medicine*, 9(6) :2039–2051, 2020.

[46] Katherine Y King, Katie A Matatall, Ching-Chieh Shen, Margaret A Goodell, Sabina I Swierczek, and Josef T Prchal. Comparative long-term effects of interferon α and hydroxyurea on human hematopoietic progenitor cells. *Experimental hematology*, 43(10) :912–918, 2015.

[47] Jingyuan Tong, Ting Sun, Shihui Ma, Yanhong Zhao, Mankai Ju, Yuchen Gao, Ping Zhu, Puwen Tan, Rongfeng Fu, Anqi Zhang, et al. Hematopoietic stem cell heterogeneity is linked to the initiation and therapeutic response of myeloproliferative neoplasms. *Cell stem cell*, 28(3) :502–513, 2021.

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

[49] Tracy Dagher, Nabih Maslah, Valérie Edmond, Bruno Cassinat, William Vainchenker, Stéphane Giraudier, Florence Pasquier, Emmanuelle Verger, Michiko Niwa-Kawakita, Valérie Lallemand-Breitenbach, et al. Jak2v617f myeloproliferative neoplasm eradication by a novel interferon/arsenic therapy involves pml. *Journal of Experimental Medicine*, 218(2), 2021.

[50] Peng Liu, Liwei Zhao, Friedemann Loos, Caroline Marty, Wei Xie, Isabelle Martins, Sylvie Lachkar, BO Qu, Emmanuelle Waeckel-Énée, Isabelle Plo, et al. Immunosuppression by mutated calreticulin released from malignant cells. *Molecular cell*, 77(4) :748–760, 2020.

[51] Camélia Benlabiod, Maira da Costa Cacemiro, Audrey Nédélec, Valérie Edmond, Delphine Muller, Philippe Rameau, Laure Touchard, Patrick Gonin, Stefan N Constantinescu, Hana Raslova, et al. Calreticulin del52 and ins5 knock-in mice recapitulate different myeloproliferative phenotypes observed in patients with mpn. *Nature communications*, 11(1) :1–15, 2020.

[52] Anna Marciniak-Czochra, Thomas Stiehl, Anthony D Ho, Willi Jäger, and Wolfgang Wagner. Modeling of asymmetric cell division in hematopoietic stem cells—regulation of self-renewal is essential for efficient repopulation. *Stem cells and development*, 18(3) :377–386, 2009.

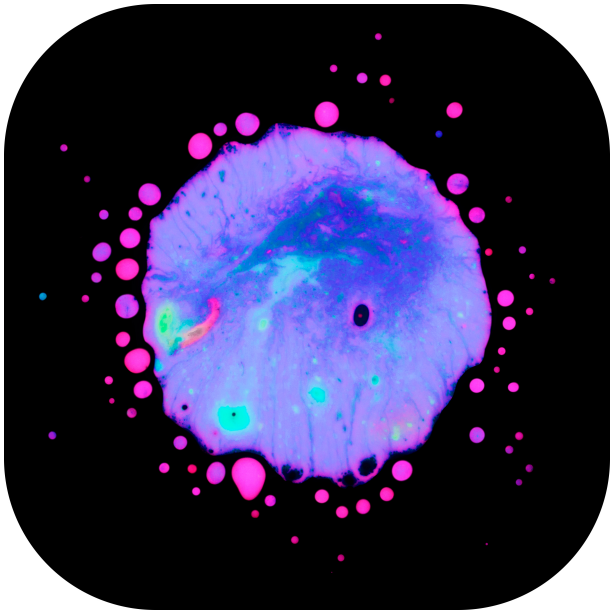
[53] Shalin H Naik, Ton N Schumacher, and Leïla Perié. Cellular barcoding : a technical appraisal. *Experimental hematology*, 42(8) :598–608, 2014.

[54] Nicholas Williams, Joe Lee, Emily Mitchell, Luiza Moore, E Joanna Baxter, James Hewinson, Kevin J Dawson, Andrew Menzies, Anna L Godfrey, Anthony R Green, et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature*, 602(7895) :162–168, 2022.

[55] Emily Mitchell, Michael Spencer Chapman, Nicholas Williams, Kevin J Dawson, Nicole Mende, Emily F Calderbank, Hyunchul Jung, Thomas Mitchell, Tim HH Coorens, David H Spencer, et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature*, pages 1–8, 2022.

[56] Marc Lavielle. *Mixed effects models for the population approach : models, tasks, methods and tools*. CRC press, 2014.

Chapitre 7
Déterminer une dose minimale d'Interféron α pour patients ayant la mutation $JAK2^{V617F}$



Résumé

Nous étendons le modèle du chapitre 6 pour prendre en compte les variations de doses d'Interféron α en cours de traitement, dans le cas de patients atteints de néoplasmes myéloprolifératifs positifs pour la mutation $JAK2^{V617F}$. En particulier, nous cherchons à modéliser des relations dose-réponse appropriées. Nous proposons une procédure de sélection en deux étapes pour déterminer le meilleur modèle, en partant d'un ensemble de 225 modèles potentiels. Dans un premier temps, nous considérons chaque modèle et chaque patient de façon indépendante et basons la sélection de modèle sur le critère d'information d'Akaike. Dans un deuxième temps, nous effectuons une estimation Bayésienne hiérarchique sur une sous-sélection de modèles pour trouver le meilleur selon le critère d'information de déviance. Enfin, nous analysons le modèle sélectionné afin de mieux caractériser l'impact de la dose sur la réponse au traitement par IFNα des patients et de déterminer pour chacun d'entre eux quelle serait la dose minimale sous laquelle une rémission pourrait ne pas être atteinte.

Le contenu de ce chapitre a fait l'objet d'une soumission pour publication et a été déposé sur un serveur *pre-print*.

Abstract

To determine the minimal dose in IFNα therapy against Myeloproliferative Neoplasms (MPN), we proposed a method combining mathematical modelling, model selection, and hierarchical Bayesian inference. We extended the model presented in chapter five to take into account the variations of posology along treatment, for $JAK2^{V617F}$ patients.

Here, we aim to accurately model the changes of posology in IFNα therapy and derive appropriate dose-response relationships. We propose a two-step selection procedure to determine the best model, starting with a very large set of potential models. Considering the high number of models that we choose to compare (225), it is computationally too expensive to estimate the parameters of these models in a hierarchical framework. Therefore, our two-step procedure is designed to first eliminate most models with a coarser but computationally feasible strategy and then conduct a more refined statistical analysis for the few selected models. First, the estimation considers all individuals independently, and models are compared based on the Akaike Information Criterion (AIC). In a second step, we perform a hierarchical Bayesian estimation on a subselection of models to find the best one according to the Deviance Information Criterion (DIC). Finally, we analyze the selected model to characterize better how the dose impacts the response to the IFNα therapy for MPN patients and to determine for each of them which would be the minimal dose under which a remission might not be reached. In addition, taking advantage of the hierarchical framework that allows us to estimate the population effect, we can determine the most suitable starting IFNα dose for a new MPN patient. Altogether, this work aims to improve clinicians' decision-making regarding the doses of IFNα to be prescribed, be it for the initial dose or the lower limit when de-escalating the dose.

Our results suggest that IFNα increases the quiescence exit of $JAK2^{V617F}$ mutated HSCs, especially the homozygous ones, but we found no evidence suggesting that this response may be sensitive to variations of dose along therapy. The identified major mechanism of action by which the stem cell pool is depleted is HSCs differentiation. IFNα may increase the propensity of $JAK2^{V617F}$ mutated HSCs to differentiate into progenitor cells as the dose increases.

We estimated that an initial dose of 45 μg/week, classically used in clinical trials, should only induce a long-term remission for 86% of the $JAK2^{V617F}$ patients, and we advocate instead to start at about 70 μg/week. A dose escalation remains relevant.

Table des matières

1 Introduction

Au chapitre précédent, nous avons montré que, pour des patients $JAK2^{V617F}$ traités par IFNα, les HSC mutées homozygotes étaient ciblées plus efficacement que les hétérozygotes et que ces dernières répondaient mieux à des doses plus fortes. Comprendre et quantifier l'impact du traitement à l'IFNα sur les cellules hématopoïétiques (mutées) est essentiel pour développer une médecine personnalisée. Grâce à leur pouvoir potentiel de prédiction de la dynamique des populations cellulaires, les modèles mathématiques sont des outils prometteurs pour guider les décisions cliniques [1]. Plusieurs modèles mathématiques ont été proposés pour modéliser l'effet d'un traitement donné contre les hémopathies malignes [2, 3, 4, 5], mais peu d'entre eux étudient l'impact des variations de dose lors de thérapies à long terme, alors qu'en routine clinique, les médecins prescrivent rarement une posologie constante sur plusieurs années. Au contraire, ils augmentent souvent la dose de manière continue jusqu'à atteindre la dose maximale tolérée (ou une réponse hématologique suffisante) et procèdent ensuite à une désescalade de la dose. De telles stratégies ont été observées sur la cohorte de patients étudiée au précédent chapitre.
Pour étudier l'effet de l'IFNα sur les cellules souches hématopoïétiques, nous avons proposé au précédent chapitre un modèle hiérarchique calibré à partir des données d'une cohorte de patients atteints de NMP et suivis pendant plusieurs années. De meilleures réponses moléculaires ont été obtenues chez les patients $JAK2^{V617F}$traités en moyenne avec des doses d'IFNα plus élevées. Cependant, le modèle n'a pas pris en compte les variations de dose dans le temps et, par conséquent, la dose minimale d'IFNα à administrer aux patients pour améliorer leurs chances d'obtenir une rémission moléculaire à long terme n'a pas pu être déterminée.
Dans ce chapitre, nous cherchons à modéliser avec précision les changements de posologie dans le traitement par IFNα et à déduire des relations dose-réponse appropriées. Nous proposons une procédure de sélection en deux étapes pour déterminer le meilleur modèle, en partant d'un ensemble très large de modèles potentiels. Compte tenu du nombre élevé de modèles que nous choisissons de comparer (225), il est trop coûteux en termes de temps de calcul d'estimer les paramètres de ces modèles dans un cadre hiérarchique comme cela était fait au chapitre précédent. Par conséquent, notre procédure en deux étapes est conçue pour éliminer d'abord la plupart des modèles à l'aide d'une stratégie plus grossière mais faisable sur le plan computationnel, puis ensuite pour effectuer une analyse statistique plus fine pour les quelques modèles sélectionnés. Dans un premier temps, l'estimation considère tous les individus indépendamment, et les modèles sont comparés sur la base du critère d'information d'Akaike (AIC). Dans un deuxième temps, nous effectuons une estimation bayésienne hiérarchique sur une sous-sélection de modèles pour trouver le meilleur selon le critère d'information de déviance (DIC), tel que défini par Spiegelhalter et al. [6]. Enfin, nous analysons le modèle sélectionné afin de mieux caractériser l'impact de la dose sur la réponse au traitement par IFNα des patients $JAK2^{V617F}$et de déterminer pour chacun d'entre eux quelle serait la dose inférieure sous laquelle une rémission pourrait ne pas être atteinte.
De plus, tirant parti du cadre hiérarchique qui nous permet d'estimer l'effet populationnel, nous pouvons déterminer la dose initiale d'IFNα la plus appropriée pour un nouveau patient atteint de NMP. Dans l'ensemble, ce travail vise à améliorer la prise de décision des cliniciens concernant les doses d'IFNα à prescrire, que ce soit pour la dose initiale ou la limite inférieure lors de la désescalade de la dose.

Le travail présenté dans ce chapitre est issu de de l'article "*Mathematical modelling, selection, and hierarchical inference to determine the minimal dose in IFNα therapy against Myeloproliferative Neoplasms*" déposé en pre-print sur arXiv [7].

2 Observations expérimentales

Les observations expérimentales sur lesquelles se base le travail présenté dans ce chapitre sont les mesures de la charge allélique (VAF) parmi les cellules matures (granulocytes) et les mesures d'architecture clonale au niveau des progéniteurs hématopoïétiques, permettant de déduire les fractions clonales (CF) en cellules mutées hétérozygotes et homozygotes. Ces observations expérimentales ont été présentées en détail aux chapitre 5 et 6.
Dans ce chapitre, nous étudions 19 patients $JAK2^{V617F}$. Ces derniers sont ceux pour lesquels nous avions suffisamment d'observations, et qui étaient considérés avoir au moins un sous-clone hétérozygote ou homozygote (i.e. au moins une CF mesurée $> 7\%$ pour l'un des sous-clones).
Par rapport au chapitre précédent, nous allons maintenant considérer comme donnée d'entrée la dose d'IFNα reçue au cours du traitement. Les patients sont considérés être sous traitement à partir du temps $t = 0$. Le traitement consiste en l'administration d'une certaine quantité d'IFNα (entre 0 et 180μg) à une certaine fréquence (généralement de toutes les semaines à toutes les trois semaines). Nous appelons dose d'interféron le ratio entre la quantité d'IFNα administrée et la durée écoulée entre deux administrations. Nous introduisons la variable $d(t) \in [0, 1]$ qui décrit la dose hebdomadaire d'IFNα administrée au moment t ($t = 0$ correspond au début du traitement) normalisée par la dose maximale observée sur la cohorte (égale à 180 μg/semaine).
Les doses administrées sont très hétérogènes entre patients, comme nous pouvons le voir sur la figure 8 qui inclut les données expérimentales utilisées pour calibrer le modèle et la visualisation des variations de dose.

3 Méthodes

3.1 Rappels

Pour comprendre comment les variations des doses d'IFNα pégylé (Pegasys) impactent précisément les HSC mutées de patients atteints de NMP, nous étendons le modèle proposé au chapitre précédent qui décrit la dynamique des cellules hématopoïétiques mutées.
Pour rappel, ce modèle (considéré dans ce chapitre comme le modèle de base que nous étendrons) considère des HSC quiescentes (compartiment 1) qui peuvent devenir actives à un taux γ, des HSC actives (compartiment 2) qui peuvent retourner à la quiescence à un taux β ou être recrutées pour se diviser à un taux α. Dans ce dernier cas, selon le type de division, la cellule donne naissance à 0, 1 ou 2 cellules progénitrices (compartiment i). Le paramètre Δ modélise l'équilibre entre la division différenciée et la division symétrique ; autrement dit, Δ est égal à la probabilité qu'une HSC génère deux HSC moins la probabilité qu'elle génère deux cellules progénitrices. Nous avons en général $\Delta \in [-1, 1]$, et dans des conditions homéostatiques, $\Delta = 0$. Les cellules progénitrices sortent de leur compartiment au taux δ_i pour devenir, après expansion au taux κ_m, des cellules matures (compartiment m). Les cellules matures sont des cellules entièrement différenciées qui meurent à un taux δ_m.
Dans ce travail, seuls les granulocytes sont considérés, c'est-à-dire que nous étudions la granulopoïèse. Cependant, le modèle lui-même reste général et valable pour d'autres types de cellules hématopoïétiques matures (en fonction principalement de la valeur que nous fixons pour δ_m). D'autres cellules différenciées sont également produites au cours de l'hématopoïèse, comme par exemple les plaquettes, les érythrocytes ou les lymphocytes. Certains modèles mathématiques se sont attachés à décrire la production d'un type cellulaire mature donné (mégacaryopoïèse [8], érythropoïèse [9], lymphopoïèse [10]) quand d'autres ont modélisé de multiples lignées cellulaires [11, 12].
Dans notre cas, la dynamique hématopoïétique est décrite par le système suivant d'équations

différentielles ordinaires (linéaires) (ODEs) :

$$\begin{cases} \frac{dN_1(t)}{dt} & = -\gamma N_1(t) + \beta N_2(t) \\ \frac{dN_2(t)}{dt} & = \gamma N_1(t) + (\alpha\Delta - \beta) N_2(t) \\ \frac{dN_i(t)}{dt} & = \alpha(1-\Delta)\kappa_i N_2(t) - \delta_i N_i(t) \\ \frac{dN_m(t)}{dt} & = \delta_i \kappa_m N_i(t) - \delta_m N_m(t) \end{cases} \tag{1}$$

où $N_1(t)$, $N_2(t)$, $N_i(t)$ et $N_m(t)$ décrivent respectivement le nombre de HSC inactives, de HSC actives, de progéniteurs et de cellules matures. Pour éviter une résolution numérique de ces équations, qui entraînerait un coût de calcul supplémentaire lors de la procédure d'estimation des paramètres, nous avions calculé précédemment une solution analytique à ce système d'ODE.

En fait, plusieurs populations de cellules sont considérées selon que les cellules sont de type sauvage (WT, indice wt) ou présentent la mutation $JAK2^{V617F}$sur un (hétérozygote, indice het) ou deux (homozygote, indice hom) allèles. Pour chacune de ces trois populations de cellules, un système d'ODE est proposé comme présenté précédemment. En termes de notation, les indices correspondants font référence aux quantités wt, het ou hom, et l'exposant * indique les paramètres influencés par l'IFNα.

3.2 Modèle dose-réponse à l'IFNα

Nous avons montré au précédent chapitre que l'IFNα pouvait favoriser la sortie de quiescence des HSC homozygotes et hétérozygotes, et accroître leur propension à se différencier en progéniteurs. En d'autres termes, un effet dose potentiel a été identifié sur les paramètres Δ^*_{het}, Δ^*_{hom}, γ^*_{het} et γ^*_{hom} de notre modèle.
Précédemment, suivant l'idée de Michor et al. [3], nous avions seulement considéré que le traitement agissait en modifiant les valeurs des paramètres dès le début de la thérapie, sans considérer d'autres variations de la posologie, ce qui correspond à une relation dose-réponse constante.
Or, les patients sous traitement subissent généralement de nombreuses variations de posologie, et parfois même des interruptions temporaires de traitement. Pour décrire plus précisément l'effet de l'IFNα sur les patients atteints de NMP, nous devons intégrer ces variations de posologie comme entrées du modèle et déduire des relations dose-réponse appropriées. À cette fin, nous introduisons la variable $d(t) \in [0,1]$ qui décrit la dose hebdomadaire d'IFNα administrée au moment t ($t = 0$ correspond au début du traitement) normalisée par la dose maximale observée (égale à 180 μg/semaine).
$\bar{\Delta}^*_{het}$, $\bar{\Delta}^*_{hom}$, $\bar{\gamma}^*_{het}$ et $\bar{\gamma}^*_{hom}$ sont maintenant des fonctions de la variable d - une entrée du modèle - et non plus des paramètres. Nous faisons la distinction entre fonctions et paramètres en utilisant le symbole $\bar{\circ}$.
À noter que $d : t \longmapsto d(t)$ est une fonction constante par morceaux. Cela implique que nous pouvons toujours obtenir une solution analytique du système d'ODE (1). Il est à noter que, pour obtenir une dynamique inférée plus lisse lors des changements de dosage, nous aurions pu utiliser une équation pharmacocinétique comme l'ont fait Ottesen et al. [5] lors de la modélisation de l'absorption de l'IFNα.

Nous adoptons une approche basée sur un modèle pour étudier la dose-réponse à l'IFNα et considérons plusieurs relations potentielles. Pour $\bar{\Delta}^*_{het}$ (et de manière équivalente pour les cellules homozygotes) qui modélise la propension des HSC mutées hétérozygotes à se différencier en progéniteurs, outre la relation constante $\bar{\Delta}^*_{het} : d \longmapsto \Delta^*_{het}$ utilisée dans le modèle de base, nous considérons :

- une relation linéaire :

$$\bar{\Delta}^*_{het} : d \longmapsto \Delta^*_{het} \cdot d \tag{2}$$

- une relation affine :

$$\bar{\Delta}^*_{het} : d \longmapsto \Delta^*_{het} \cdot d + \Delta_{het} \tag{3}$$

- une relation sigmoïde :

$$\bar{\Delta}^*_{het} : d \longmapsto \frac{-2}{1 + e^{-\Delta^*_{het} \cdot d}} + 1 \tag{4}$$

- et une relation affine sigmoïde :

$$\bar{\Delta}^*_{het} : d \longmapsto -2 \left(\frac{1}{1 + e^{-\Delta^*_{het} \cdot d}} - 0.5 \right) \cdot (1 + \Delta_{het}) + \Delta_{het} \tag{5}$$

Dans l'équation (3) et (5), nous obtenons $\bar{\Delta}^*_{het}(d = 0) = \Delta_{het}$; c'est-à-dire que nous étudions la possibilité que les HSC mutées $JAK2^{V617F}$envahissent naturellement le *pool* de cellules souches (en supposant que $\Delta_{het} > 0$) en l'absence de traitement, comme l'ont montré Van Egeren et al. [13] ou comme nous l'avons montré au chapitre 5. Dans les équations (2) et (4), à l'opposé, nous forçons explicitement Δ_{het} à être égal à zéro, comme nous l'avions fait au chapitre précédent, avec donc un degré de liberté en moins, en considérant que l'envahissement lent du clone muté nous permette de faire l'approximation $\Delta_{het} \approx 0$.
Nous devons avoir $\forall d \in [0, 1], \bar{\Delta}^*_{het}(d) \in [-1, 1]$. Cette condition est assurée par un choix approprié des distributions *a priori* (plus précisément, par le choix des limites inférieure et supérieure du support des distributions *a priori*) dans le cas des relations constantes, linéaires et affines et est automatiquement vérifiée dans les deux relations sigmoïdes.

Pour $\bar{\gamma}^*_{het}$ (et de manière équivalente pour $\bar{\gamma}^*_{hom}$) qui modélise la sortie de quiescence des HSC mutées, en plus de la relation constante $\bar{\gamma}^*_{het} : d \longmapsto \gamma^*_{het}$ utilisée dans le modèle de base, nous considérons une relation affine :

$$\bar{\gamma}^*_{het} : d \longmapsto \gamma^*_{het} \cdot d + \gamma_{het}, \tag{6}$$

et une relation inverse :

$$\bar{\gamma}^*_{het} : d \longmapsto \frac{1}{\tau^*_{het} \cdot d + 1/\gamma_{het}}, \tag{7}$$

ce qui correspond à une relation affine pour l'inverse de $\bar{\gamma}^*_{het}$. En fait, il n'y a pas de raison particulière de privilégier le paramètre γ, qui correspond à un taux de sortie de quiescence, à son inverse $\tau = 1/\gamma$, qui correspondrait à un temps moyen de sortie de quiescence.
En suivant les choix faits au précédent chapitre, nous considérons que $\bar{\gamma}^*_{het}(d = 0) = \gamma_{het} = 1/300$ [jours^{-1}], et donc, nous n'étudions pas les relations linéaires comme nous le faisons pour $\bar{\Delta}^*_{het}$. Là encore, des distributions *a priori* appropriées sont choisies pour garantir que $d \longmapsto \bar{\gamma}^*_{het}(d)$ soit une fonction croissante et positive.

De nombreuses autres relations dose-réponse auraient pu être étudiées, comme cela a été fait par exemple dans [14], mais comme nous combinons les relations dose-réponse pour quatre paramètres différents dans notre modèle dynamique, il en résulterait un très grand ensemble de modèles à calibrer. Nous avons donc choisi de nous limiter à quelques relations standard.

3.3 Procédure de sélection de modèle en deux étapes

3.3.1 Sélection basée sur l'AIC

Étant donné les relations dose-réponse du § 3.2, nous nous retrouvons avec un large ensemble de modèles mathématiques différents à comparer. En effet, nous étudions 5 relations pour $\bar{\Delta}^*_{het}$ et $\bar{\Delta}^*_{hom}$, et 3 pour $\bar{\gamma}^*_{het}$ et $\bar{\gamma}^*_{hom}$. Comme il serait possible d'avoir des relations dose-réponse différentes selon la zygosité (het ou hom), on se retrouve avec $5^2 \times 3^2 = 225$ modèles à comparer. Pour $j \in \{1, \cdots, 225\}$, le modèle j ($\mathcal{M}_j$) correspond au modèle dynamique (décrit par trois systèmes d'ODE indépendants (1), un pour la population de cellules wt, un pour les cellules het et le troisième pour les hom) avec une combinaison particulière de relations dose-réponse pour les quatre paramètres précédents.
Dans une première approche, étant donné le grand nombre de modèles que nous voulons comparer, nous utilisons d'abord une méthode grossière mais rapide basée sur le critère d'information

d'Akaike (AIC) [15, 16], pour comparer les différents modèles et sélectionner le plus adéquat.
À cette fin, nous utilisons les données (désignées par $\mathcal{D} = \{\mathcal{D}_i\}_{i\in\{1,\cdots,N\}}$) de $N = 19$ patients $JAK2^{V617F}$étudiés au chapitre précédent.
Les données consistent en des proportions (ou Fraction Clonale, CF) de cellules progénitrices mutées (qui correspondent au compartiment i dans le modèle dynamique), et des VAF parmi les cellules matures (correspondant au compartiment m) mesurées à différents moments, du début de la thérapie jusqu'à 5 ans de traitement.
Pour chaque patient i et modèle j, nous calculons un AIC (Fig. 1) :

$$AIC_{i,j} = -2\log(\mathcal{L}_{i,j}) + 2k_j \tag{8}$$

avec $\mathcal{L}_{i,j}$ le maximum de vraisemblance et k_j le nombre de paramètres à estimer avec le modèle j.
À noter que nous aurions également pu utiliser le critère d'information bayésien (BIC) [17], ce qui aurait conduit, dans notre cas, aux mêmes conclusions.
Le modèle statistique utilisé pour exprimer la vraisemblance est le même qu'au chapitre précédent. Le maximum de la vraisemblance est calculé à l'aide de l'algorithme CMA-ES (Covariance Matrix Adaptation - Evolution Strategy) [18] que nous avions présenté plus en détail au chapitre 4. Enfin, pour comparer les performances des différents modèles sur l'ensemble de la cohorte, nous calculons un AIC global (Fig. 2) qui additionne la contribution de tous les patients :

$$AIC_j = \sum_{1\leq i\leq N} AIC_{i,j} \tag{9}$$

Il correspond à la vraisemblance d'un modèle considérant tous les patients ensemble mais observés indépendamment les uns des autres, et avec des paramètres individuels indépendants. Sur la base de ce critère global, nous pouvons trier les différents modèles et sélectionner ceux qui donnent les meilleurs résultats (c'est-à-dire les plus petites valeurs pour l'AIC global).

3.3.2 Sélection basée sur une estimation bayésienne hiérarchique

Un des inconvénients du modèle statistique sur lequel le critère (9) est construit est que tous les patients sont considérés indépendamment. Aucun effet de population n'est pris en compte, avec un risque d'*overfitting* [19].
Par conséquent, dans une deuxième étape, nous appliquons une méthode d'estimation Bayésienne hiérarchique pour estimer les distributions des paramètres pour chaque patient ainsi que les paramètres de population (appelés hyper-paramètres, HP).
Cette méthode a été utilisée et détaillée au chapitre précédent. En bref, si l'on considère une population $\mathcal{P} = \{1, \cdots, N\}$ de N patients, dont la dynamique hématopoïétique est décrite selon le modèle $\mathcal{M}_j$, $\boldsymbol{\theta} = \left\{\boldsymbol{\theta}^{(i)}\right\}_{i\in\mathcal{P}}$ désigne l'ensemble de tous les paramètres des patient avec :

$$\begin{aligned}\boldsymbol{\theta}^{(1)} &= \left(\theta_1^{(1)}, \cdots, \theta_P^{(1)}\right)\\ &\vdots\\ \boldsymbol{\theta}^{(N)} &= \left(\theta_1^{(N)}, \cdots, \theta_P^{(N)}\right)\end{aligned}$$

où P est le nombre de paramètres à estimer pour le modèle $\mathcal{M}_j$. Avec la méthode d'inférence hiérarchique, au lieu d'estimer chaque $\boldsymbol{\theta}^{(i)}$ indépendamment, nous supposons que tous les vecteurs de paramètres individuels sont des réalisations de la même variable aléatoire de distribution inconnue dans un modèle statistique. Ainsi, le modèle hiérarchique (également appelé modèle à effets aléatoires) peut rendre compte de la variabilité inter-individuelle mais aussi de la similarité entre les patients. En pratique, on considère ici :

$$\forall i \in \mathcal{P}, \forall k \in \{1, \cdots, P\}, \theta_k^{(i)} \mid \tau_k, \sigma_k^2 \sim \mathcal{N}_{c,k}\left(\tau_k, \sigma_k^2\right) \tag{10}$$

où la distribution de population pour chacun des paramètres est une distribution gaussienne tronquée $\mathcal{N}_{c,k}$ (sur un intervalle qui dépend du paramètre considéré k), et $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_P)$ et $\boldsymbol{\sigma^2} = (\sigma_1^2, \cdots, \sigma_P^2)$ sont les hyperparamètres.

Nous pouvons alors estimer les distributions *a posteriori* jointes de $\boldsymbol{\theta}$ et des hyperparamètres $\boldsymbol{\tau}$ et $\boldsymbol{\sigma^2}$ (comme effectué et présenté en détails au précédent chapitre) :

$$\begin{aligned} p\left[\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\sigma^2} \mid \mathcal{D}\right] &\propto p\left[\mathcal{D} \mid \boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}, \boldsymbol{\tau}, \boldsymbol{\sigma^2}\right] p\left[\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(N)}, \boldsymbol{\tau}, \boldsymbol{\sigma^2}\right] \\ &\propto \prod_{i \in \mathcal{P}} \left(p\left[\mathcal{D}_i \mid \boldsymbol{\theta}^{(i)}\right] p\left[\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\tau}, \boldsymbol{\sigma^2}\right]\right) p[\boldsymbol{\tau}] p\left[\boldsymbol{\sigma^2}\right] \end{aligned} \tag{11}$$

Ensuite, nous échantillonnons à partir de la distribution *a posteriori* en utilisant une méthode MCMC, à savoir l'algorithme de Metropolis-Hastings *within* Gibbs [20, 21]. Conditionnellement aux hyperparamètres, les patients sont indépendants et leurs paramètres peuvent être échantillonnés en utilisant un schéma standard de Metropolis-Hastings.

En raison du coût de calcul élevé de la méthode d'inférence hiérarchique, nous ne l'utilisons que pour la comparaison d'un nombre limité de modèles : le modèle de base et ceux que nous avons d'abord sélectionnés sur la base de l'AIC.

Après avoir exécuté la procédure d'estimation des paramètres jusqu'à convergence, nous calculons le critère d'information de déviance (DIC) de chaque modèle afin de sélectionner le meilleur [6] (Tab. 1). Pour le modèle $\mathcal{M}_j$, DIC_j est défini par :

$$DIC_j = D\left(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}_j]\right) + 2p_{D_j} \tag{12}$$

Avec la déviance définie par $D(\boldsymbol{\theta}) = -2\log\left(p[\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}_j]\right)$ et p_{D_j} le nombre effectif de paramètres définis, suivant Gelman et al. [22], par $p_{D_j} = 0.5\mathbb{V}[D(\boldsymbol{\theta})]$. Enfin, nous sélectionnons le modèle présentant la valeur DIC la plus faible.

3.4 Estimation de l'impact de l'IFNα et d'une dose minimale

Une fois que nous avons exécuté notre procédure de sélection de modèles et sélectionné le meilleur modèle (c'est-à-dire les relations dose-réponse les plus appropriées pour $\bar{\Delta}^*_{het}$, $\bar{\Delta}^*_{hom}$, $\bar{\gamma}^*_{het}$ et $\bar{\gamma}^*_{hom}$), nous pouvons analyser plus en détail les résultats de cette procédure (section 4.1). Nous pouvons étudier si les clones mutés $JAK2^{V617F}$ hétérozygotes et homozygotes répondent différemment aux variations des doses d'IFNα. En effet, notre méthode permet de comparer des modèles dont les relations dose-réponse pourraient être différentes selon le génotype, à l'instar de Tong et al. [23] qui ont suggéré que les HSC hétérozygotes pourraient répondre à l'IFNα différemment des cellules homozygotes. En outre, pour le meilleur modèle (que nous désignons maintenant par $\mathcal{M}$), nous pouvons comparer les distributions *a posteriori* de nos hyper-paramètres $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_P)$ et étudier comment elles diffèrent entre les cellules hétérozygotes ou homozygotes (Fig. 4).

De plus, à partir des résultats de la calibration du modèle $\mathcal{M}$, nous pouvons étudier comment les patients répondent individuellement au traitement (section 4.2). Nous pouvons d'abord comparer les distributions *a posteriori* de leurs paramètres individuels (Fig. 6).
Ensuite, en échantillonnant dans ces distributions à l'aide d'une méthode de Monte-Carlo, nous pouvons propager l'incertitude des paramètres à la sortie des modèles (c'est-à-dire la dynamique de la CF dans chaque compartiment hématopoïétique) et afficher la dynamique inférée accompagnée d'un intervalle de crédibilité à 95% pour chaque patient (Fig. 8).

Plus intéressant encore, nous pouvons étudier pour chaque patient comment $\bar{\Delta}^*_{het}$ et $\bar{\Delta}^*_{hom}$ dépendent de la dose d'IFNα d (section 4.3). $\bar{\Delta}^*_{het}$ et $\bar{\Delta}^*_{hom}$ sont en fait des quantités clés de notre modèle. En effet, une rémission moléculaire ne peut être obtenue que si des valeurs

négatives sont atteintes pour ces deux quantités, ce qui signifie, selon notre modèle, que les HSC $JAK2^{V617F}$ subiront plus de divisions différenciées que de divisions symétriques (d'auto-renouvellement), conduisant à une déplétion du *pool* de cellules souches mutantes.
Puisque l'IFNα peut potentiellement induire certains effets secondaires, notamment des dépressions majeures [24, 25], ce serait crucial de connaître la dose minimale qui est nécessaire et suffisante pour éliminer les clones mutés, c'est-à-dire telle que $\bar{\Delta}^*_{het}$ et $\bar{\Delta}^*_{hom}$ soient négatifs. Une dose trop élevée pourrait augmenter la toxicité de la thérapie alors qu'une dose trop faible pourrait ne pas être capable d'induire une réponse moléculaire (majeure) pour le patient. Ainsi, nous introduisons $d^{(i)}_{min}$ comme étant la valeur minimale de la dose d, pour le patient i, telle que $P_{rem}(d) > 95\%$ avec :

$$P_{rem}(d) = \mathbb{P}[\bar{\Delta}^*_{het}(d) < 0, \bar{\Delta}^*_{hom}(d) < 0 \mid \mathcal{M}, \mathcal{D}] \tag{13}$$

En calculant cette quantité, nous pouvons estimer pour chaque patient de notre cohorte la limite inférieure de la dose d'IFNα à administrer (Fig. 11). Comme ces patients sont toujours sous traitement, cette information pourrait être utile aux cliniciens.

Un avantage de la méthode d'inférence Bayésienne hiérarchique est que nous n'estimons pas seulement les paramètres individuels, mais nous déduisons également un effet populationnel. En supposant que les patients de notre cohorte sont représentatifs de la population des patients atteints de NMP, notamment en ce qui concerne la gamme des doses d'IFNα administrées, nous pouvons considérer que l'effet de population que nous avons inféré (par l'estimation de la distribution *a posteriori* des hyper-paramètres) peut être généralisé pour d'autres patients atteints de NMP en dehors de notre cohorte.
Plus précisément, nous avons supposé que les paramètres individuels du modèle $\theta^{(i)}_k$ (paramètre k, patient $i \in \{1, \cdots, N\}$) suivaient une distribution gaussienne (tronquée) (10) de moyenne et de variance τ_k et σ^2_k respectivement : $\theta^{(i)}_k \sim \mathcal{N}_{c,k}\left(\tau_k, \sigma^2_k\right)$, où τ_k et σ^2_k étaient des hyper-paramètres à estimer et pour lesquels nous n'avions que de vagues *priors*. Après l'estimation sur les données $\mathcal{D}$, nous obtenons la distribution *a posteriori* de τ_k et σ^2_k, de sorte que nous pouvons mettre à jour nos connaissances au niveau de la population. Pour un nouveau patient $N+1$, nous pouvons maintenant considérer comme nouveau *prior* :

$$\theta^{(N+1)}_k \sim \mathcal{N}_{c,k}\left(\mathbb{E}[\tau_k|\mathcal{D}], \mathbb{E}[\sigma^2_k|\mathcal{D}]\right)$$

En échantillonnant dans la distribution précédente, nous pouvons déduire les relations dose-réponse générales (c'est-à-dire au niveau de la population) pour $\bar{\Delta}^*_{het}$ et $\bar{\Delta}^*_{hom}$ (section 4.4, Fig. 12) et estimer la dose minimale d_{min} qui devrait être administrée à un nouveau patient ayant soit des HSC mutées hétérozygotes, soit des HSC homozygotes, soit les deux (et avant toute autre information médicale ou observation clinique pertinente) pour maximiser ses chances d'obtenir une rémission moléculaire à long terme (Fig. 13).
Nous pouvons également évaluer, pour des posologies croissantes, la proportion de patients qui pourraient être finalement guéris, et confronter nos résultats à des stratégies standard d'escalade de dose allant de 45 μg/semaine à 135 μg/semaine [26, 27, 28, 29].

4 Résultats

4.1 Différences suivant la zygosité : d'une procédure de sélection de modèle à l'analyse des distributions *a posteriori* des hyper-paramètres

Pour mieux comprendre l'impact de l'IFNα sur la dynamique des HSC mutées chez les patients atteints de NMP $JAK2^{V617F}$, et en particulier comprendre son impact en fonction de la zygosité (het *vs* hom), nous appliquons notre procédure de sélection de modèles en deux étapes.

La première étape de notre procédure de sélection de modèles consiste à calibrer l'ensemble des 225 modèles pour chacun des 19 patients. Il en résulte une grande quantité de modèles à calibrer,

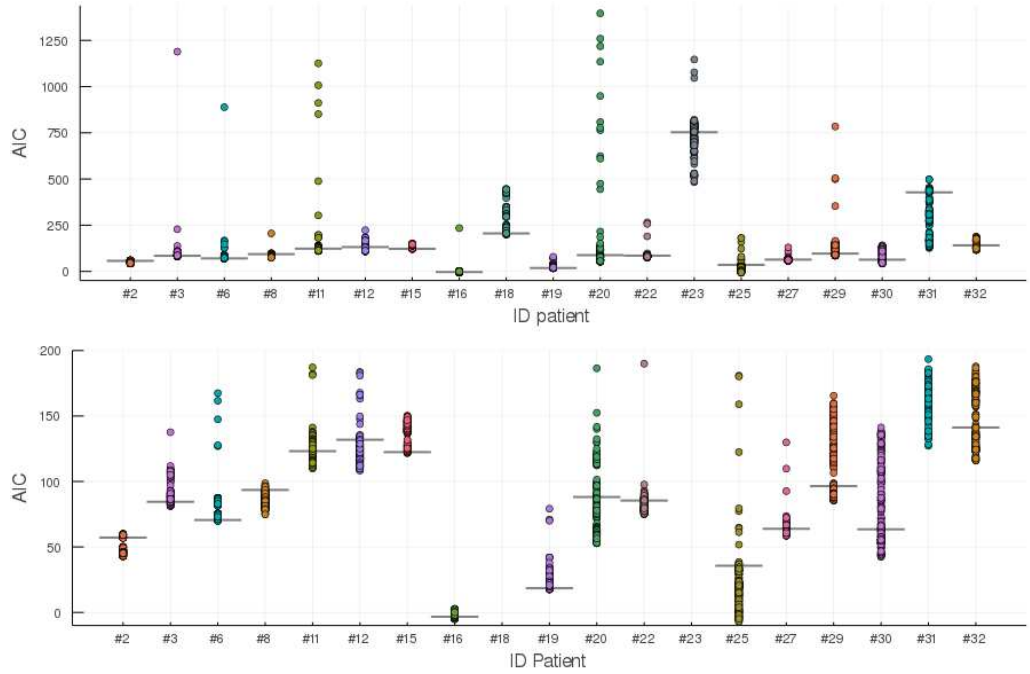


FIGURE 1 – Valeurs AIC pour chaque modèle j (représenté par un point) et patient i. Pour chaque patient, la ligne horizontale affiche la valeur AIC du modèle de base. Sur la figure du bas, l'axe des ordonnées est tronqué pour plus de clarté et permet de faire un focus sur les meilleurs modèles.

et il est donc nécessaire d'utiliser une méthode d'inférence efficace. À cette fin, nous utilisons l'algorithme CMA-ES pour calculer le maximum de vraisemblance $\mathcal{L}_{i,j}$ pour chaque modèle $\mathcal{M}_j$, étant donné les observations du patient i. Pour tous les modèles, la vraisemblance est exprimée en utilisant le même modèle d'observation que celui décrit au précédent chapitre.
Ensuite, nous calculons l'AIC (voir eq. (8)) pour chaque combinaison : patient × modèle. Les résultats sont affichés sur la figure 1. On observe que le modèle de base (celui du chapitre précédent) était déjà un bon modèle, avec de très bons ajustements pour les patients #6, 15, 16, 19. Cependant, le modèle de base présentait l'inconvénient de donner lieu à de très mauvais ajustements pour d'autres patients, tels que les patients #20, 23 et 31. Il était donc nécessaire de trouver un modèle plus performant pour tous les patients de la cohorte. Pour presque tous les patients, le modèle de base est amélioré en considérant des relations dose-réponse plus complexes que les relations constantes.
Néanmoins, le meilleur modèle pour un patient donné n'est pas nécessairement le meilleur pour un autre. Pour comparer les modèles, non pas au niveau individuel mais au niveau de la population (de nos 19 patients $JAK2^{V617F}$), nous calculons un AIC global (voir eq. (9)) qui additionne la contribution de tous les AIC individuels. Sommer les AIC individuels se justifie en considérant un modèle complet pour les N patients, mais avec des observations indépendantes. La vraisemblance du modèle complet s'obtient alors comme le produit des vraisemblances et la log-vraisemblance comme la somme des log-vraisemblances. Le nombre de paramètres du modèle complet s'obtient de même en sommant les nombres de paramètres pour chacun des modèles individuels.
Sur la figure 2, nous affichons l'AIC global de tous les modèles et les comparons à l'AIC global du modèle de base (ligne horizontale orange). De nombreux modèles améliorent le modèle de base sur l'ensemble de la cohorte. En particulier, trois modèles se distinguent, avec un AIC global d'environ 2,000 alors que l'AIC global du modèle de base est d'environ 2,700. Ils sont présentés dans le tableau 1, et nous affichons leurs performances pour chaque patient sur la figure 3. Ces modèles sont les suivants :

- Modèle "orange" : une relation dose-réponse constante pour $\bar{\gamma}^*_{het}$ et $\bar{\gamma}^*_{hom}$, une relation sigmoïde affine pour $\bar{\Delta}^*_{hom}$, et une relation affine pour $\bar{\Delta}^*_{het}$.

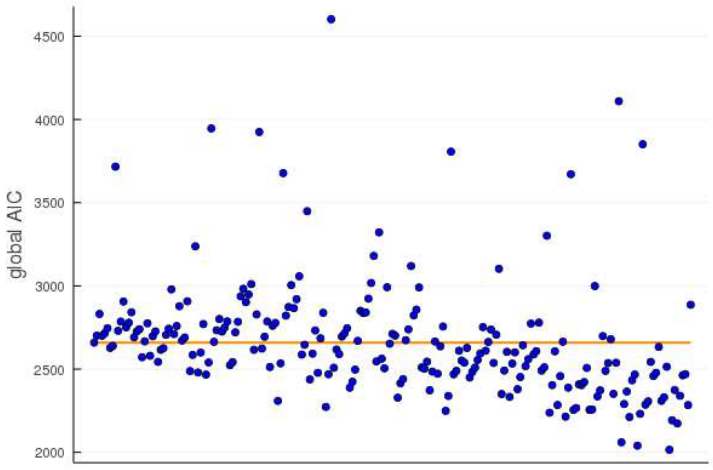


FIGURE 2 – AIC global AIC_j pour chaque modèle j. La ligne horizontale orange représente l'AIC global du modèle de base (celui du chapitre 6).

- Modèle "rouge" : relation dose-réponse constante pour $\bar{\gamma}^*_{het}$, affine pour $\bar{\gamma}^*_{hom}$, affine sigmoïde pour $\bar{\Delta}^*_{hom}$ et affine pour $\bar{\Delta}^*_{het}$.
- Modèle "bleu" : relations dose-réponse constantes pour $\bar{\gamma}^*_{het}$ et $\bar{\gamma}^*_{hom}$, et une relation affine sigmoïde pour $\bar{\Delta}^*_{hom}$ et $\bar{\Delta}^*_{het}$.

Les modèles "orange" et "rouge" sont meilleurs que le modèle de base pour 11 patients (sur 19), et le modèle "bleu" améliore l'ajustement de 12 patients.

À noter qu'au lieu de l'AIC, nous aurions pu utiliser le critère d'information bayésien (BIC), défini par :

$$BIC_{i,j} = -2\log\left(\mathcal{L}_{i,j}\right) + k_j \cdot \log\left(N_i\right)$$

avec N_i le nombre d'observations (à la fois pour les cellules matures et progénitrices) pour le patient i. Les mêmes trois meilleurs modèles sont sélectionnés lors de l'application de ce critère.

Dans un second temps, nous ne considérons que les trois modèles qui se démarquent et les comparons à l'aide d'une méthode d'inférence Bayésienne hiérarchique plus rigoureuse. Les résultats sont présentés dans le tableau 1.
Le modèle sélectionné est le modèle "bleu", avec des relations dose-réponse constantes pour $\bar{\gamma}^*_{het}$ et $\bar{\gamma}^*_{hom}$ et une relation affine sigmoïde pour $\bar{\Delta}^*_{hom}$ et $\bar{\Delta}^*_{het}$. Le modèle sélectionné est celui qui a le meilleur AIC, DIC, et qui améliore également les résultats pour 12 patients au lieu de 11 pour les deux autres.
Il est intéressant de noter que les résultats de cette procédure de sélection de modèle suggèrent que l'IFNα a un mécanisme d'action similaire envers les sous-clones mutés hétérozygotes et homozygotes, car nous aboutissons au même type de relation dose-réponse dans les deux cas (mais avec des paramètres différents suivant la zygosité).

En comparant sur la figure 4 les distributions *a posteriori* des hyperparamètres $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_P)$,

Label	$\bar{\Delta}^*_{hom}$	$\bar{\Delta}^*_{het}$	$\bar{\gamma}^*_{hom}$	$\bar{\gamma}^*_{het}$	global AIC	DIC
"bleu"	sigmoïde affine	sigmoïde affine	constante	constante	2,016	2,292
"rouge"	sigmoïde affine	affine	affine	constante	2,040	2,423
"orange"	sigmoïde affine	affine	constante	constante	2,060	2,309
Base	constante	constante	constante	constante	2,659	2,878

TABLE 1 – Les meilleurs modèles basés sur les critères AIC et DIC, et comparaison avec le modèle de base.

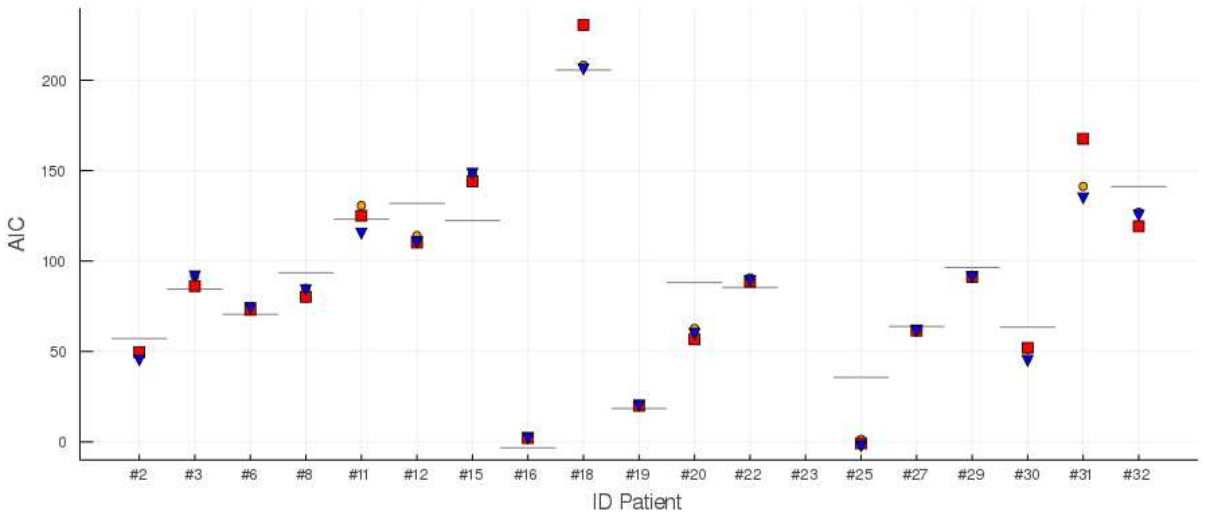


FIGURE 3 – Focus sur les 3 meilleurs modèles en fonction de leur AIC global. Le cercle orange correspond au modèle avec une relation dose-réponse constante pour $\bar{\gamma}^*_{het}$ et $\bar{\gamma}^*_{hom}$, une relation sigmoïde affine pour $\bar{\Delta}^*_{hom}$, et une relation affine pour $\bar{\Delta}^*_{het}$. Le carré rouge correspond au modèle avec une relation dose-réponse constante pour $\bar{\gamma}^*_{het}$, affine pour $\bar{\gamma}^*_{hom}$, affine sigmoïde pour $\bar{\Delta}^*_{hom}$ et affine pour $\bar{\Delta}^*_{het}$. Le triangle bleu correspond au modèle avec une relation dose-réponse constante pour $\bar{\gamma}^*_{het}$ et $\bar{\gamma}^*_{hom}$, et une relation affine sigmoïde pour $\bar{\Delta}^*_{hom}$ et $\bar{\Delta}^*_{het}$. La ligne noire horizontale correspond au modèle de base. L'axe des ordonnées est tronqué pour plus de clarté.

qui correspondent aux moyennes des distributions de population (10) associées à :

- $1/\gamma^*_{hom}$ et $1/\gamma^*_{het}$ (liées à la sortie de quiescence),
- $\Delta_{hom}(0)$ et $\Delta_{het}(0)$ (liées à la propension initiale - ou en l'absence de traitement - des HSC mutées à envahir le *pool* de cellules souches),
- et Δ^*_{hom} et Δ^*_{het} (liés à l'intensité à laquelle décroît la capacité d'auto-renouvellement des HSC mutées avec l'augmentation de la dose d'IFNα),

nous constatons que l'ampleur de la réponse à l'IFNα diffère entre les cellules mutées hétérozygotes et homozygotes.
Comme nous l'avions déjà mis en évidence au précédent chapitre, globalement, la sortie de quiescence sous IFNα est plus importante dans les sous-clones homozygotes que dans les hétérozygotes (Fig. 4 gauche). Au niveau de la population, l'intensité à laquelle décroît la capacité d'auto-renouvellement des HSC mutées avec la dose ne diffère pas significativement selon la zygosité (Fig. 4 droite). Enfin, nous constatons une plus grande propension à l'auto-renouvellement pour les HSC mutées homozygotes, comparées aux hétérozygotes, en l'absence de traitement (Fig. 4 milieu). Ce résultat est biologiquement cohérent puisque les HSC homozygotes présentent la mutation sur deux allèles de sorte que l'avantage sélectif conféré par la mutation $JAK2^{V617F}$ devrait être accru par rapport aux HSC hétérozygotes [30].

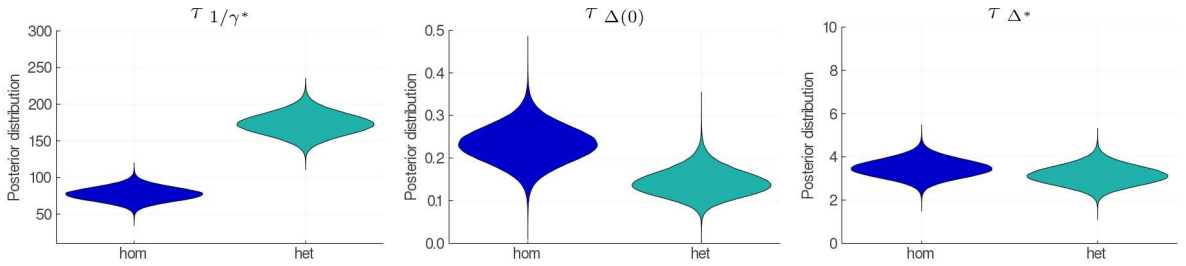


FIGURE 4 – Comparaison des distributions *a posteriori* des hyper-paramètres (HP) de la population $\boldsymbol{\tau}$ pour les clones hétérozygotes (vert) et homozygotes (bleu). À gauche : HP lié à la sortie de quiescence ($1/\tau^*$). Des valeurs plus faibles indiquent un taux plus élevé de sortie de la quiescence. Au milieu : HP lié à la propension initiale (c'est-à-dire sans traitement) des HSC mutées à envahir le *pool* de cellules souches ($\Delta(0)$). Des valeurs plus élevées indiquent une plus grande capacité à envahir le *pool* de cellules souches, c'est-à-dire plus de divisions symétriques (auto-renouvellement) que de divisions différenciées. À droite : HP lié à l'effet de la dose dans la relation dose-réponse (5) (Δ^*). Des valeurs plus élevées indiquent une augmentation plus importante de la différenciation des HSC mutées en progéniteurs en fonction de la dose.

4.2 Réponses individuelles au traitement

Dans le modèle choisi $\mathcal{M}$, les variations de l'IFNα ont un impact sur $\bar{\Delta}^*_{het}$ et $\bar{\Delta}^*_{hom}$ à travers la relation sigmoïde affine (5) qui fait intervenir deux paramètres : la valeur initiale Δ (sans IFNα, que nous écrivons également pour plus de clarté : $\Delta(0)$) et Δ^* qui peut être interprétée comme une pente pour la relation dose-réponse (ou l'intensité à laquelle décroît la capacité d'auto-renouvellement des HSC mutées avec l'augmentation de la dose d'IFNα).
Pour visualiser dans quelle mesure ce modèle s'adapte bien aux données de la cohorte, nous comparons sur la figure 5 les valeurs observées et inférées, à la fois pour les cellules matures (VAF) et pour les progéniteurs het et hom (CF). Nous observons globalement un bon accord entre les observations et les valeurs inférées.

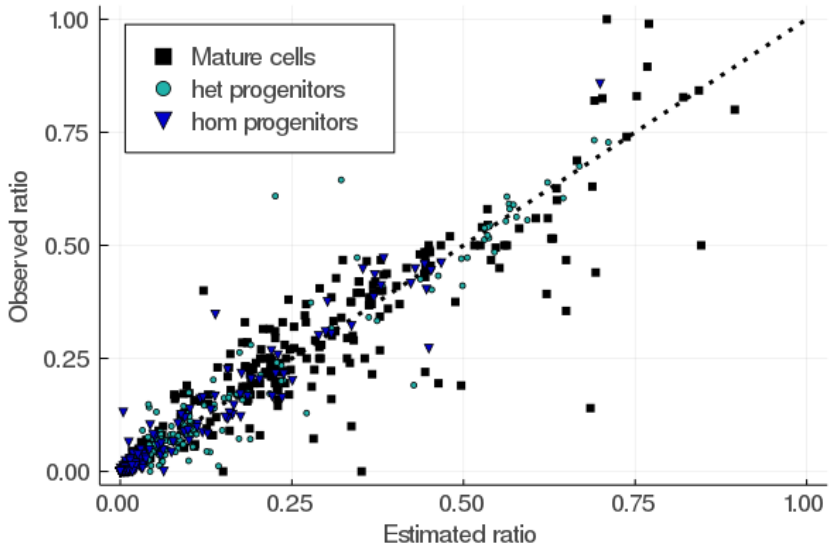


FIGURE 5 – Comparaison entre les ratios observés (CF pour les cellules progénitrices, VAF pour les cellules matures) et ceux estimés (basés sur la moyenne *a posteriori* du vecteur de paramètres). Sur nos 19 patients, nous avons 232 observations pour les cellules matures et 255 pour les progéniteurs hétérozygotes et homozygotes. La ligne en pointillés matérialise la première bissectrice qui correspond à un ajustement qui serait exact.

Les distributions *a posteriori* de ces paramètres, à la fois pour les sous-clones mutés hétérozy-

gotes et homozygotes, ont été estimées pour chaque patient, ainsi qu'un effet populationnel, à partir d'une méthode d'inférence Bayésienne hiérarchique. Ces distributions sont affichées sur la figure 6.
À noter que tous les patients de notre cohorte ne présentent pas de HSC homozygotes. Nous observons une certaine hétérogénéité inter-individuelle ; les patients répondent différemment au traitement, comme nous l'avions déjà montré au chapitre précédent. L'effet populationnel inféré, décrit par une loi normale (tronquée) (voir éq. (10)) avec une moyenne (*a posteriori*) donnée par $\mathbb{E}[\boldsymbol{\tau}|\mathcal{D}]$ et une variance donnée par $\mathbb{E}[\boldsymbol{\sigma^2}|\mathcal{D}]$, est également représenté sur la figure 6. Les estimations des autres paramètres impliqués dans le modèle $\mathcal{M}$ sont présentées sur la figure 7.

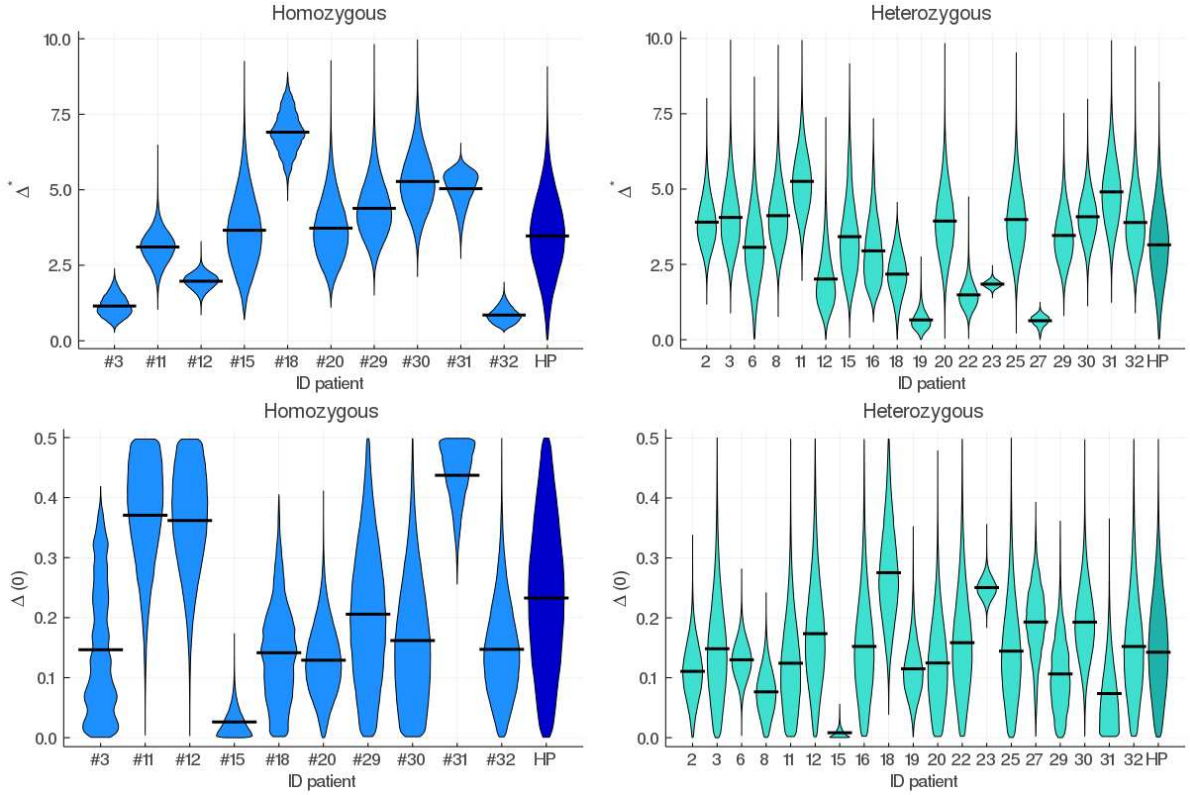


FIGURE 6 – Distributions *a posteriori* des paramètres impliqués dans les relations dose-réponse de $\bar{\Delta}^*_{hom}$ (gauche) et $\bar{\Delta}^*_{het}$ (droite), estimées pour chaque patient. HP indique la distribution de la population, décrite par une distribution gaussienne (tronquée) (10) avec une moyenne $\mathbb{E}[\boldsymbol{\tau}|\mathcal{D}]$ et une variance $\mathbb{E}[\boldsymbol{\sigma^2}|\mathcal{D}]$.
En haut, nous affichons les distributions pour le paramètre Δ^* qui peut être vu comme la pente dans les relations sigmoïdes affines (5), et en bas la propension - en l'absence de traitement - des cellules mutées à envahir le *pool* de cellules souches ($\Delta(0)$). Pour les paramètres liés aux cellules homozygotes, seuls les patients qui présentent des clones homozygotes sont présentés. Les lignes horizontales indiquent les valeurs moyennes. L'axe des ordonnées correspond à l'intervalle de valeurs autorisées (intervalles sur lesquels les distributions de population sont tronquées).

Ensuite, pour chaque patient, en échantillonnant à partir de sa distribution *a posteriori* à l'aide d'une méthode de Monte-Carlo, nous propageons les incertitudes des paramètres à la sortie du modèle et affichons la dynamique hématopoïétique au cours du traitement, avec un intervalle de crédibilité à 95%, pour la VAF dans les cellules matures et la CF dans les progéniteurs. La figure 8 présente toutes les dynamiques inférées.
Il y a un bon accord entre les valeurs observées et estimées pour la plupart des patients. Nous montrons également sur cette figure comment les doses d'IFNα varient au cours du traitement pour chaque patient. Il est intéressant d'observer comment la proportion de cellules mutées augmente lorsque la dose diminue de manière trop importante. Ce modèle est plus performant que

le modèle de base, en particulier pour les patients #23, 31, 20 et 25.
Par rapport aux résultats du chapitre précédent, il est intéressant d'observer que certaines observations tardives (à environ 2,000 jours pour les patients #20, 25 et 32), qui étaient auparavant considérées comme aberrantes, sont maintenant, avec notre modèle amélioré, le reflet d'une diminution de la dose (ou même d'une interruption du traitement) et le signe d'une rechute. Ces résultats indiquent que la dose ne devrait pas être diminuée de manière trop importante.

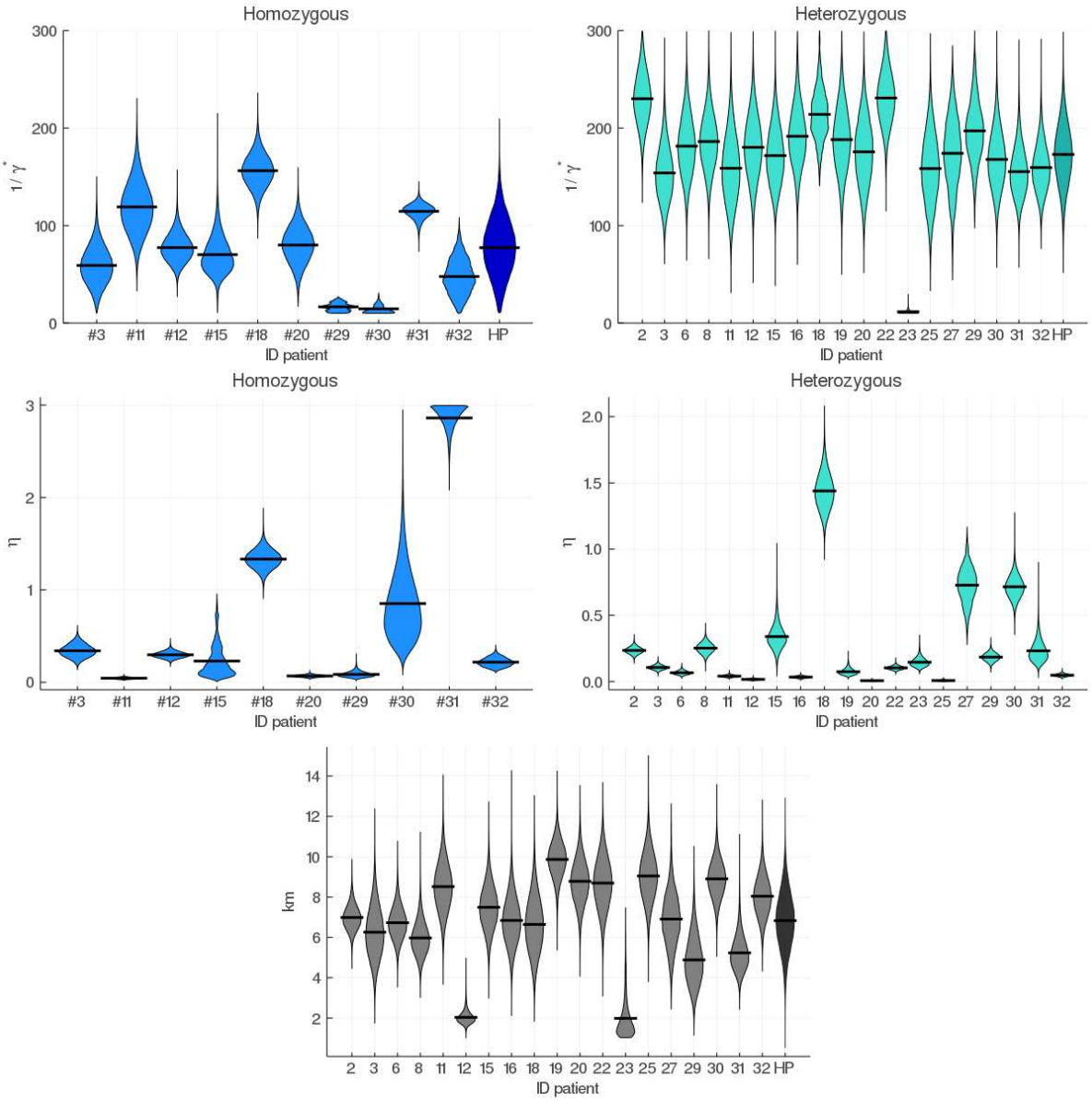


FIGURE 7 – Distributions *a posteriori* des paramètres. HP correspond à la distribution populationnelle, décrite par une distribution gaussienne (tronquée) (formule (10)) avec une moyenne $\mathbb{E}[\boldsymbol{\tau}|\mathcal{D}]$ et une variance $\mathbb{E}[\boldsymbol{\sigma^2}|\mathcal{D}]$. Pour les paramètres relatifs aux cellules homozygotes (bleu), seuls les patients qui présentent des clones homozygotes sont présentés. Les lignes horizontales indiquent les valeurs moyennes. L'axe des ordonnées correspond à l'intervalle de valeurs autorisées (sauf pour η_{het} qui est tronqué pour des raisons de clarté, et dont l'intervalle de valeurs *a priori* est le même que pour η_{hom}).

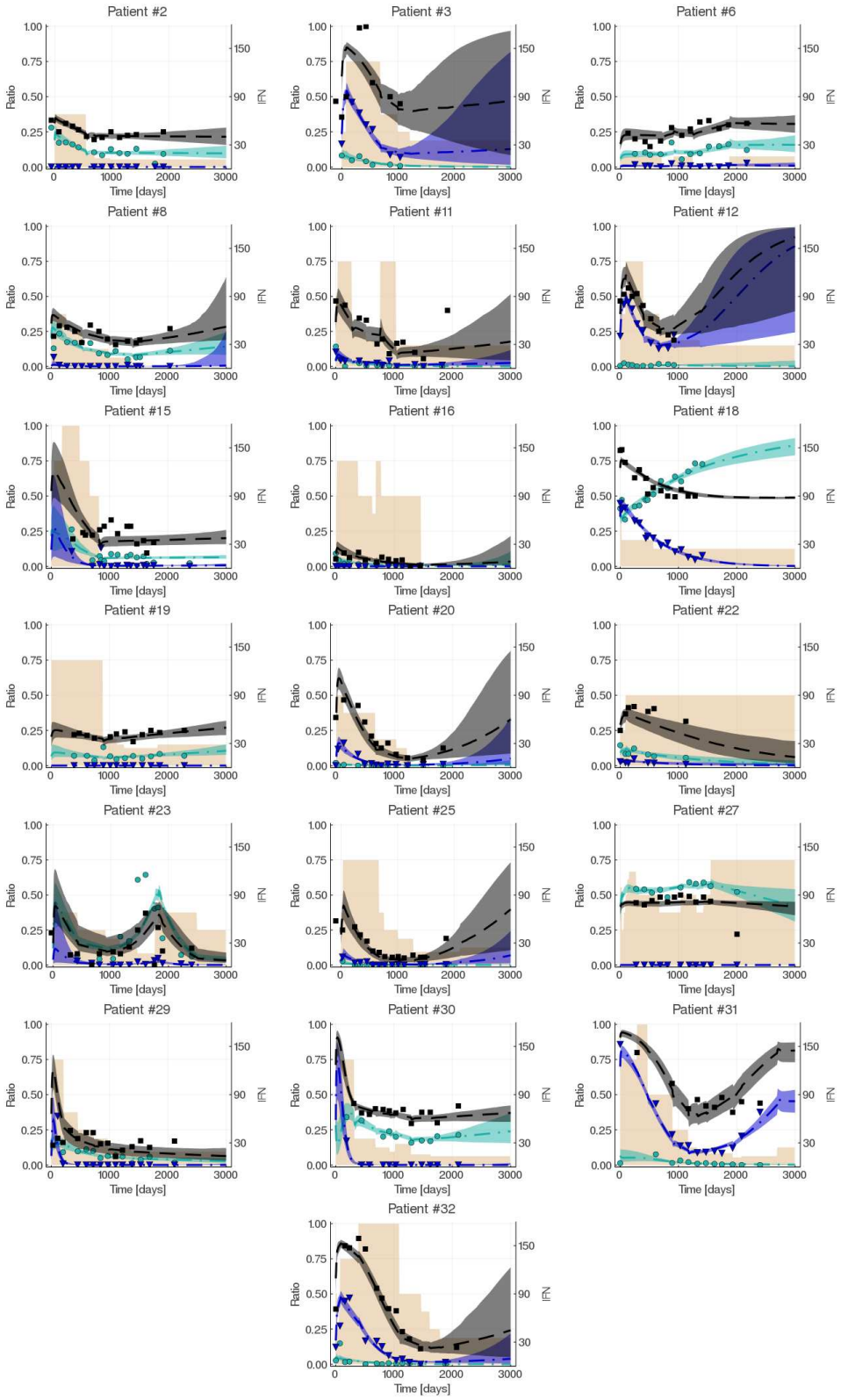


FIGURE 8 – Les dynamiques des progéniteurs mutés homozygotes et hétérozygotes inférées (CF) et des cellules matures (VAF) sont présentées pour 19 patients $JAK2^{V617F}$ pour lesquels nous avons effectué l'estimation des paramètres en utilisant notre méthode d'estimation Bayésienne hiérarchique (pour le modèle sélectionné). Les triangles, les points et les carrés représentent les observations expérimentales. Les courbes ont été déterminées avec le modèle (valeurs médianes). Les zones ombrées représentent les intervalles de crédibilité à 95%. Les zones grisées beiges correspondent à la dose d'IFNα reçue au cours du traitement.

4.3 Estimation d'une dose individuelle minimale

D'après les résultats de notre estimation, nous constatons que les patients $JAK2^{V617F}$ réagissent différemment au traitement et qu'il existe un risque de rechute si la dose d'IFNα est diminuée. Dans le modèle sélectionné $\mathcal{M}$, la rechute s'explique par le fait que la diminution de la dose d augmente également la propension des HSC mutées à envahir le *pool* de cellules souches, c'est-à-dire que $\bar{\Delta}^*$ est une fonction décroissante de d.
Sur la figure 9, nous affichons pour tous les patients les relations dose-réponse estimées $\bar{\Delta}^*_{het}$ et $\bar{\Delta}^*_{hom}$ (pour cette dernière, uniquement si le patient possède des sous-clones homozygotes) en prenant, pour les paramètres impliqués dans les deux relations, les moyennes *a posteriori* des distributions de $\Delta_{het}(0), \Delta_{hom}(0), \Delta^*_{het}$ et Δ^*_{hom}.

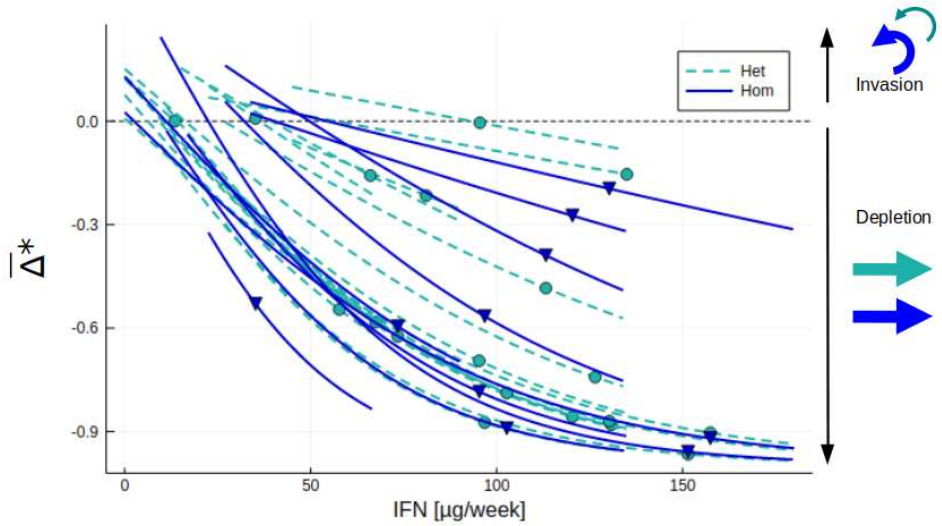


FIGURE 9 – Relations dose-réponse pour $\bar{\Delta}^*_{het}$ (vert) et $\bar{\Delta}^*_{hom}$ (bleu) en fonction de la dose hebdomadaire d d'IFNα. Chaque courbe représente un patient. Les extrémités des courbes correspondent aux doses minimale et maximale qu'ils ont reçues au cours du traitement. Les points et les triangles correspondent respectivement à la valeur de $\bar{\Delta}^*_{het}(D)$ et de $\bar{\Delta}^*_{hom}(D)$ où D est le dosage (défini au chapitre précédent comme la dose moyenne reçue sur les 450 premiers jours de traitement) administré au patient.

Puisque $\Delta^*_{het}(0) > 0$ et que $\bar{\Delta}^*$ est une fonction décroissante de d, il existe, pour chaque patient hétérozygote (sans HSC homozygote), une dose d' telle que $\bar{\Delta}^*_{het}(d') = 0$ et $\bar{\Delta}^*_{het}(d) > 0$ pour $d < d'$ (et nous étendons naturellement la définition de d' dans le cas de patients ayant simultanément des HSC homozygotes et hétérozygotes). Puisque nous obtenons les distributions (*a posteriori*) des paramètres, nous pouvons donner les intervalles de crédibilité des relations dose-réponse et quantifier leur incertitude, comme illustré sur la figure 10 (gauche) pour le patient #32.
En particulier, d' ne peut être connu avec certitude, mais nous pouvons estimer pour toutes les doses possibles d la probabilité de rémission $P_{rem}(d)$ (voir eq. (13)), qui est également égale à $P_{rem}(d) = \mathbb{P}[d > d']$. Cette probabilité est représentée pour le patient #32 sur la figure 10 (à droite). Nous définissons la dose minimale d'IFNα pour un patient i par $d^{(i)}_{min}$ telle que $P_{rem}(d^{(i)}_{min}) = 0.95$. Pour le patient #32, cette dose minimale est estimée à 72 μg/semaine.

La dose minimale peut alors être calculée pour chaque patient de notre cohorte (Fig. 11), de manière similaire à ce qui a été présenté pour le patient #32. Nos résultats peuvent alors être utilisés comme des recommandations à destination des cliniciens, notamment pour décourager une désescalade brutale de la dose ou une diminution de la dose d'IFNα pour un patient donné en dessous de la limite inférieure estimée, car cela augmenterait son risque de rechute.

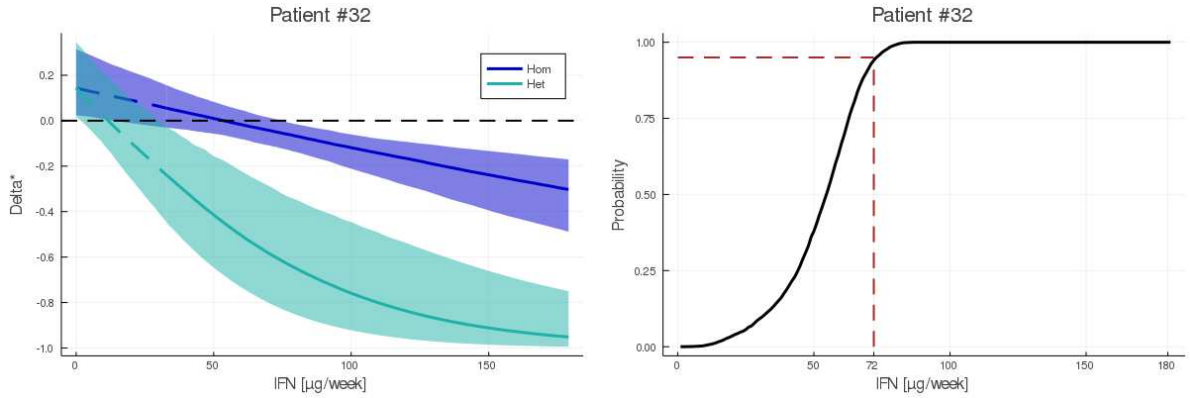


FIGURE 10 – À gauche, relations dose-réponse de $\bar{\Delta}^*_{het}$ et $\bar{\Delta}^*_{hom}$ pour le patient #32 ainsi que les intervalles de crédibilité à 95%. À droite, probabilité de rémission P_{rem} en fonction de la dose. Les lignes pointillées rouges indiquent la dose au-dessus de laquelle il y a 95% de chances d'obtenir une rémission, soit $d^{(i)}_{min} = 72$ μg/semaine pour le patient $i := 32$.

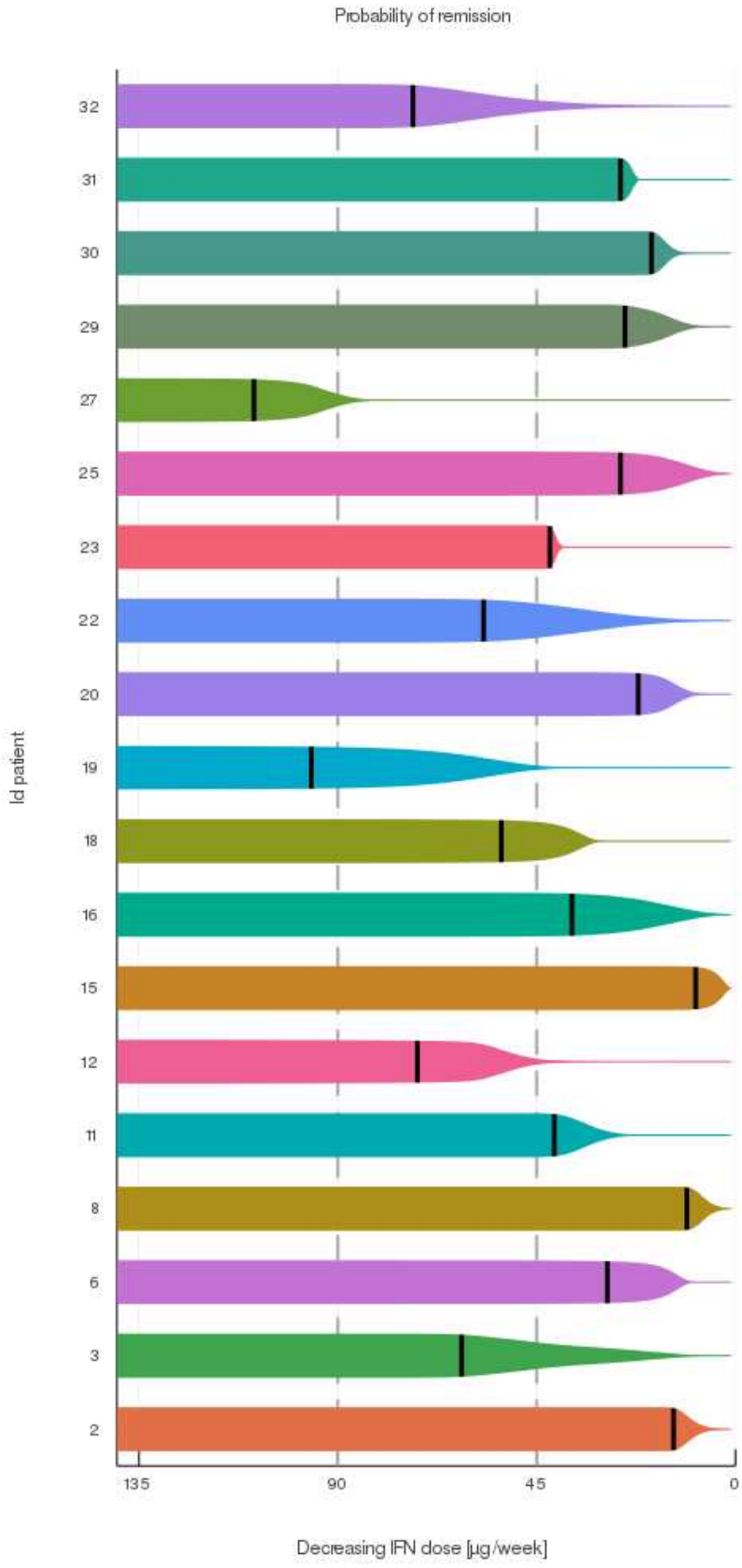


FIGURE 11 – Estimation de la dose minimale qui doit être administrée à chaque patient. L'axe des abscisses indique une désescalade de la dose, de 135 μg/semaine (à gauche) à 0 μg/semaine (à droite). La diminution de la dose réduit également la probabilité de rémission. Les lignes noires verticales indiquent la dose minimale estimée d'IFNα $d_{min}^{(i)}$ qui devrait être administrée au patient i pour qu'il ait 95% de chances d'obtenir une rémission moléculaire à long terme.

4.4 Estimation d'une dose initiale adaptée

Dans la section précédente, nous avons calculé pour tous les patients de notre cohorte une dose minimale personnalisée d'IFNα, à la fois nécessaire et suffisante pour obtenir une rémission à long terme. De plus, grâce au cadre Bayésien hiérarchique, l'inférence ne se fait pas seulement au niveau individuel, mais aussi au niveau de la population.
Pour un nouveau patient, nous pouvons maintenant considérer, comme distributions *a priori* pour ses paramètres, des distributions gaussiennes (tronquées) de moyennes et de variances les moyennes *a posteriori* estimées des hyper-paramètres correspondants.
Ce nouveau *prior* peut être utilisé *a priori*, avant d'avoir des observations pour ce patient, et peut aider à déterminer une dose de départ appropriée. En échantillonnant à partir de ces nouvelles distributions *a priori*, nous pouvons déterminer le comportement *a priori* de $\bar{\Delta}^*_{het}$ et de $\bar{\Delta}^*_{hom}$ en fonction de la dose (figure 12).

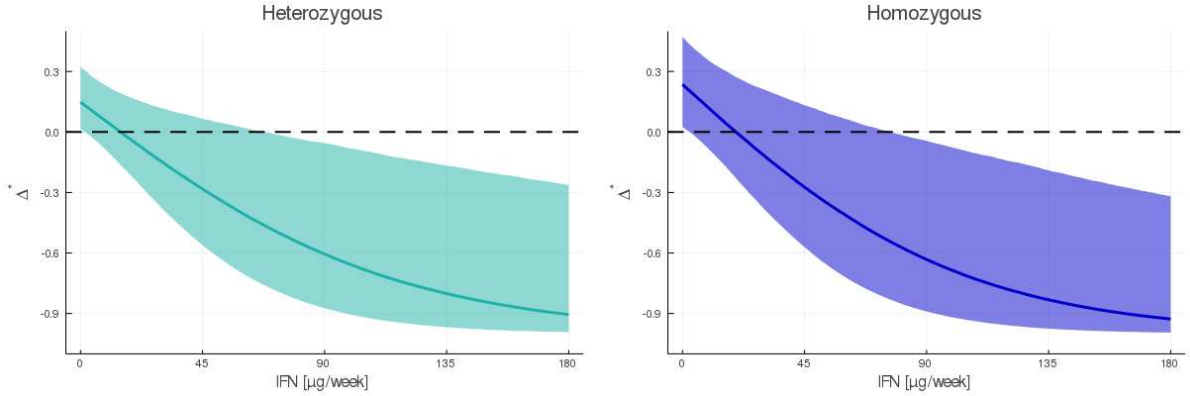


FIGURE 12 – Relations dose-réponse *a priori* de $\bar{\Delta}^*_{het}$ (gauche) et de $\bar{\Delta}^*_{hom}$ (droite) pour des nouveaux patients (d'après l'estimation faite à partir de notre cohorte). Les lignes indiquent les valeurs médianes, les zones ombragées indiquent les intervalles de crédibilité à 95%. La ligne horizontale en pointillés noirs indique $\Delta^* = 0$. En dessous de cette limite, le traitement induit une rémission à long terme, selon notre modèle.

Ensuite, nous pouvons calculer la probabilité de rémission à long terme, en fonction de la dose, en considérant un patient ayant uniquement des sous-clones hétérozygotes, uniquement des sous-clones homozygotes, ou les deux sous-clones mutés (Fig. 13). Pour une population de patients ayant des HSC homozygotes et hétérozygotes $JAK2^{V617F}$, nous estimons qu'une dose initiale de 45 μg/semaine pourrait conduire à une rémission à long terme dans 86% des cas. Nos résultats suggèrent que la dose initiale minimale telle que 95% des patients atteignent une rémission, devrait être égale à 71 μg/semaine.

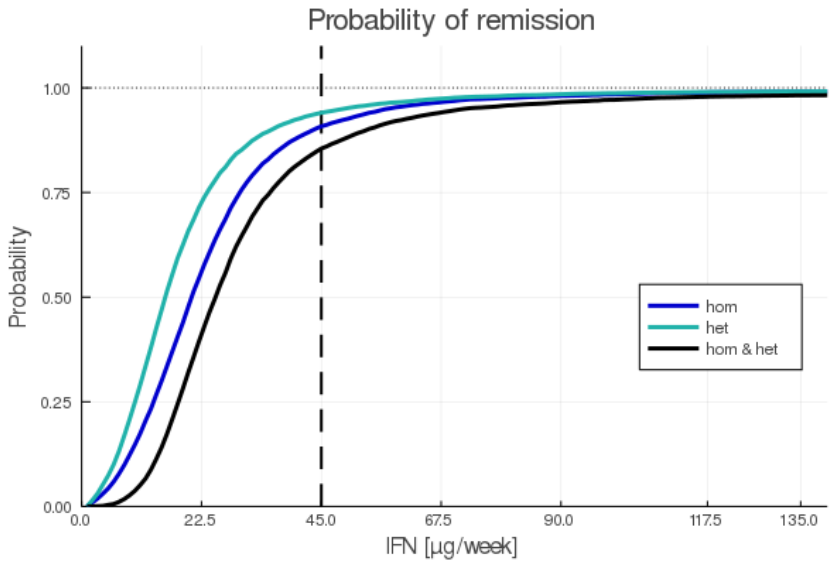


FIGURE 13 – Pour un nouveau patient $JAK2^{V617F}$, probabilité d'avoir une rémission à long terme en fonction de la posologie initiale. Nous distinguons les cas où les patients ne présentent que des HSC homozygotes (bleu), que des HSC hétérozygotes (vert) ou les deux (noir). La ligne verticale en pointillés indique la dose initiale utilisée dans les essais cliniques [27, 26].

5 Discussion

Pour déterminer la dose minimale d'IFNα dans le traitement contre les NMP $JAK2^{V617F}$, nous avons proposé une méthode combinant modélisation mathématique, sélection de modèles et inférence Bayésienne hiérarchique. Nous avons étendu le modèle présenté au chapitre 6 pour prendre en compte les variations de la posologie au cours du traitement.
Nous avons proposé plusieurs modèles alternatifs et utilisé une procédure de sélection de modèle en deux étapes : d'abord, nous avons écarté la plupart des modèles sur la base de l'AIC, puis nous avons appliqué une méthode d'inférence Bayésienne hiérarchique pour calibrer les modèles restants. Nous avons sélectionné le modèle le plus performant en fonction du critère DIC. Enfin, nous avons analysé en détail les résultats obtenus pour le modèle sélectionné. Les données utilisées pour la calibration provenaient de 19 patients $JAK2^{V617F}$.
En accord avec nos résultats du chapitre 6, nos estimations suggèrent que l'IFNα favorise la sortie de quiescence des HSC mutées, en particulier celles homozygotes, mais nous n'avons trouvé aucune preuve suggérant que cette augmentation de la sortie de quiescence puisse être sensible aux variations de dose au cours du traitement. Une explication possible serait que les HSC homozygotes sortent de quiescence au début du traitement et ne reviennent plus à cet état. Le principal mécanisme d'action identifié par lequel le *pool* de cellules souches mutées est appauvri est la différenciation des HSC en progéniteurs, contrairement à ce qui se passerait dans le cas WT, tel que rapporté par Pietras et al. [31]. L'IFNα peut augmenter la propension des HSC mutées à se différencier en cellules progénitrices, d'autant plus que la dose augmente. Le fait que notre procédure de sélection de modèles nous ait conduit à sélectionner les mêmes types de relations dose-réponse dans le cas hétérozygote et homozygote pourrait suggérer que le mécanisme d'action ne diffèrerait pas suivant la zygosité, en accord nos précédents résultats, mais contrairerment à ceux de Tong et al. [23], qui ont montré, par analyse single-cell RNA-seq, que les cellules homozygotes seraient quiescentes tandis que les cellules hétérozygotes subiraient une apoptose, chez des patients traités par IFNα.
Lorsqu'une dose minimale est atteinte, les HSC mutées subiraient plus de divisions différenciées qu'auto-renouvelées ; une rémission moléculaire à long terme pourrait être obtenue. Nous avons estimé cette dose minimale nécessaire et suffisante pour épuiser à long terme les clones mutés, pour chaque individu de notre cohorte. D'après nos résultats, il serait déconseillé de procéder à une trop forte désescalade de la dose, pendant le traitement, et de passer en dessous de cette

limite inférieure. Avec le modèle sélectionné, tel qu'il est construit, nous ne pouvons pas recommander une interruption du traitement car elle conduirait à une rechute. En effet, dans ce modèle, nous considérons que les HSC mutées ont une propension initiale (et en l'absence de traitement) à envahir le *pool* de cellules souches, condition nécessaire à l'expansion du clone muté et à l'apparition des symptômes du NMP, comme décrit dans [13] et étudié au chapitre 5. Ainsi, en l'absence d'IFNα, même une infime fraction de cellules mutées (qui subsistent toujours dans notre modèle déterministe) continuerait à se développer. En réalité, lorsque le nombre de HSC mutées devient très faible, notre modèle déterministe n'est plus adapté puisque les effets stochastiques prédominent. On pourrait en tenir compte en étendant de façon appropriée notre modèle et ainsi étudier le moment opportun à partir duquel on peut interrompre le traitement.
Selon nos estimations, l'avantage sélectif (initial) des cellules homozygotes est plus élevé que celui des cellules hétérozygotes, ce qui signifie que les HSC homozygotes subiraient en moyenne plus de divisions symétriques (auto-renouvellement) que celles hétérozygotes. Ce résultat est biologiquement cohérent avec le fait que les HSC homozygotes présentent la mutation $JAK2^{V617F}$ sur les deux allèles, ce qui augmenterait leur *fitness* [30]. La capacité initiale à envahir le compartiment souche (paramètre $\Delta(0)$) est cependant plus importante que celle estimée au chapitre 5. Cela pourrait peut-être s'expliquer par le fait que le modèle étudié au chapitre 5 ne comporte qu'un compartiment pour les cellules souches, et se concentre sur le développement initial du NMP, alors que dans ce chapitre, nous considérons à la fois un compartiment actif et quiescent pour les cellules souches, et le paramètre $\Delta(0)$ correspondrait dans notre modèle à la fois à une capacité d'envahissement initiale (avant traitement) mais également pendant une interruption du traitement.
Avec notre méthode d'estimation Bayésienne hiérarchique, nous avons inféré un effet populationnel, estimé la dose minimale pour tous les individus de notre cohorte et déterminé la dose initiale la plus appropriée pour un nouveau patient. Nous avons estimé qu'une dose initiale de 45 μg/semaine, classiquement utilisée dans les essais cliniques [27, 26, 28], ne devrait induire une rémission à long terme que dans 86% des cas, et nous préconisons plutôt de commencer à environ 70 μg/semaine. Une escalade de dose reste pertinente. Même si nous prévoyons une rémission avec une faible dose, elle pourrait n'être atteinte qu'au bout d'un temps très long, et l'augmentation de la dose accélérerait la réponse au traitement. D'ailleurs, les cliniciens traitent rarement à de faibles doses (par exemple 45 μg/semaine) pendant une longue période, mais augmentent plutôt rapidement la posologie jusqu'à 90 ou 135 μg/semaine, jusqu'à obtenir une réponse hématologique. Bien sûr, en routine clinique, les médecins doivent également prendre en compte des contraintes supplémentaires, telles que l'apparition d'effets secondaires, principalement la dépression [24, 25]. L'apparition d'effets secondaires peut obliger le clinicien à diminuer la posologie ou même à interrompre temporairement le traitement. En perspective, il serait intéressant de relier les paramètres du modèles à des caractéristiques cliniques, telles que l'âge, le sexe ou la maladie du patient (TE, PV ou MFP).

En conclusion, l'approche mathématique proposée, utilisée pour déterminer une dose minimale à prescrire, reste générale et peut être appliquée dans un large éventail de problèmes. Notre méthode illustre comment une procédure de sélection de modèles peut aider à comprendre le mécanisme d'action d'un traitement et montre le potentiel des méthodes d'inférence Bayésiennes hiérarchiques, tant pour augmenter la robustesse des estimations faites sur une cohorte d'individus observés que pour généraliser les résultats à de nouveaux patients.
Au chapitre suivant, nous proposerons des pistes pour prolonger ce travail, étudiant des méthodes de contrôle optimal et l'application de ce modèle à des fins prédictives.

Références

[1] Katja Hoffmann, Katja Cazemier, Christoph Baldow, Silvio Schuster, Yuri Kheifetz, Sibylle Schirm, Matthias Horn, Thomas Ernst, Constanze Volgmann, Christian Thiede, et al. Integration of mathematical model predictions into routine workflows to support clinical decision making in haematology. *BMC medical informatics and decision making*, 20(1) :1–12, 2020.

[2] Ingo Roeder, Matthias Horn, Ingmar Glauche, Andreas Hochhaus, Martin C Mueller, and Markus Loeffler. Dynamic modeling of imatinib-treated chronic myeloid leukemia : functional insights and clinical implications. *Nature medicine*, 12(10) :1181–1184, 2006.

[3] Franziska Michor, Timothy P Hughes, Yoh Iwasa, Susan Branford, Neil P Shah, Charles L Sawyers, and Martin A Nowak. Dynamics of chronic myeloid leukaemia. *Nature*, 435(7046) :1267–1270, 2005.

[4] Svetlana Bunimovich-Mendrazitsky, Natalie Kronik, and Vladimir Vainstein. Optimization of interferon–alpha and imatinib combination therapy for chronic myeloid leukemia : A modeling approach. *Advanced Theory and Simulations*, 2(1) :1800081, 2019.

[5] Johnny T Ottesen, Rasmus K Pedersen, Marc JB Dam, Trine A Knudsen, Vibe Skov, Lasse Kjær, and Morten Andersen. Mathematical modeling of mpns offers understanding and decision support for personalized treatment. *Cancers*, 12(8) :2119, 2020.

[6] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society : Series b (statistical methodology)*, 64(4) :583–639, 2002.

[7] Gurvan Hermange, William Vainchenker, Isabelle Plo, and Paul-Henry Cournède. Mathematical modelling, selection and hierarchical inference to determine the minimal dose in ifn alpha therapy against myeloproliferative neoplasms. *arXiv preprint arXiv :2112.10688*, 2021.

[8] Loïs Boullu, Laurent Pujo-Menjouet, and Jianhong Wu. A model for megakaryopoiesis with state-dependent delay. *SIAM Journal on Applied Mathematics*, 79(4) :1218–1243, 2019.

[9] Fabien Crauste, Laurent Pujo-Menjouet, Stéphane Génieys, Clément Molina, and Olivier Gandrillon. Adding self-renewal in committed erythroid progenitors improves the biological relevance of a mathematical model of erythropoiesis. *Journal of theoretical biology*, 250(2) :322–338, 2008.

[10] Salvador Chulián, Álvaro Martínez-Rubio, Anna Marciniak-Czochra, Thomas Stiehl, Cristina Blázquez Goñi, Juan Francisco Rodríguez Gutiérrez, Manuel Ramírez Orellana, Ana Castillo Robleda, Víctor M Pérez-García, and María Rosa. Dynamical properties of feedback signalling in b lymphopoiesis : A mathematical modelling approach. *Journal of Theoretical Biology*, 522 :110685, 2021.

[11] Caroline Colijn and Michael C Mackey. A mathematical model of hematopoiesis—i. periodic chronic myelogenous leukemia. *Journal of Theoretical Biology*, 237(2) :117–132, 2005.

[12] Jason Xu, Samson Koelle, Peter Guttorp, Chuanfeng Wu, Cynthia Dunbar, Janis L Abkowitz, and Vladimir N Minin. Statistical inference for partially observed branching processes with application to cell lineage tracking of in vivo hematopoiesis. *The Annals of Applied Statistics*, 13(4) :2091–2119, 2019.

[13] Debra Van Egeren, Javier Escabi, Maximilian Nguyen, Shichen Liu, Christopher R Reilly, Sachin Patel, Baransel Kamaz, Maria Kalyva, Daniel J DeAngelo, Ilene Galinsky, et al. Reconstructing the lineage histories and differentiation trajectories of individual cancer cells in myeloproliferative neoplasms. *Cell stem cell*, 28(3) :514–523, 2021.

[14] Frank Bretz, José C Pinheiro, and Michael Branson. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, 61(3) :738–748, 2005.

[15] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6) :716–723, 1974.

[16] Kenneth P Burnham and Davfd R Anderson. A practical information-theoretic approach. *Model selection and multimodel inference*, 2, 2002.

[17] Schwarz Gideon et al. Estimating the dimension of a model. *The annals of statistics*, 6(2) :461–464, 1978.

[18] Nikolaus Hansen. The cma evolution strategy : A tutorial. *arXiv preprint arXiv :1604.00772*, 2016.

[19] Artémis Llamosi, Andres M Gonzalez-Vargas, Cristian Versari, Eugenio Cinquemani, Giancarlo Ferrari-Trecate, Pascal Hersen, and Gregory Batt. What population reveals about individual cell identity : single-cell parameter estimation of models of gene expression in yeast. *PLoS computational biology*, 12(2) :e1004706, 2016.

[20] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[21] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6) :721–741, 1984.

[22] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. Bayesian data analysis chapman & hall. *CRC Texts in Statistical Science*, 2004.

[23] Jingyuan Tong, Ting Sun, Shihui Ma, Yanhong Zhao, Mankai Ju, Yuchen Gao, Ping Zhu, Puwen Tan, Rongfeng Fu, Anqi Zhang, et al. Hematopoietic stem cell heterogeneity is linked to the initiation and therapeutic response of myeloproliferative neoplasms. *Cell stem cell*, 28(3) :502–513, 2021.

[24] Francis E Lotrich, Mordechai Rabinovitz, Patricia Gironda, and Bruce G Pollock. Depression following pegylated interferon-alpha : characteristics and vulnerability. *Journal of psychosomatic research*, 63(2) :131–135, 2007.

[25] Peter C Trask, Peg Esper, Michelle Riba, and Bruce Redman. Psychiatric side effects of interferon therapy : prevalence, proposed mechanisms, and future directions. *Journal of Clinical Oncology*, 18(11) :2316–2326, 2000.

[26] Trine Alma Knudsen, Vibe Skov, Kristen Stevenson, Lillian Werner, William Duke, Charles Laurore, Christopher J Gibson, Anwesha Nag, Aaron R Thorner, Bruce Wollison, et al. Genomic profiling of a randomizedtrial of interferon-α versus hydroxyureain mpnreveals mutation-specific responses. *Blood advances*, 2021.

[27] Abdulraheem Yacoub, John Mascarenhas, Heidi Kosiorek, Josef T Prchal, Dmitry Berenzon, Maria R Baer, Ellen Ritchie, Richard T Silver, Craig Kessler, Elliott Winton, et al. Pegylated interferon alfa-2a for polycythemia vera or essential thrombocythemia resistant or intolerant to hydroxyurea. *Blood*, 134(18) :1498–1509, 2019.

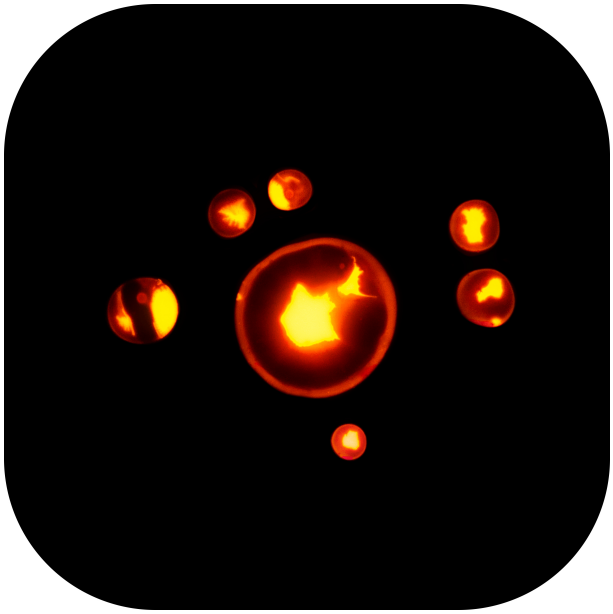
[28] Heinz Gisslinger, Christoph Klade, Pencho Georgiev, Dorota Krochmalczyk, Liana Gercheva-Kyuchukova, Miklos Egyed, Viktor Rossiev, Petr Dulicek, Arpad Illes, Halyna Pylypenko, et al. Ropeginterferon alfa-2b versus standard therapy for polycythaemia vera (proud-pv and continuation-pv) : a randomised, non-inferiority, phase 3 trial and its extension study. *The Lancet Haematology*, 7(3) :e196–e208, 2020.

[29] Tiziano Barbui, Alessandro Maria Vannucchi, Valerio De Stefano, Arianna Masciulli, Alessandra Carobbio, Alberto Ferrari, Arianna Ghirardi, Elena Rossi, Fabio Ciceri, Massimiliano Bonifacio, et al. Ropeginterferon alfa-2b versus phlebotomy in low-risk patients with polycythaemia vera (low-pv study) : a multicentre, randomised phase 2 trial. *The Lancet Haematology*, 8(3) :e175–e184, 2021.

[30] Nicholas Williams, Joe Lee, Emily Mitchell, Luiza Moore, E Joanna Baxter, James Hewinson, Kevin J Dawson, Andrew Menzies, Anna L Godfrey, Anthony R Green, et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature*, 602(7895) :162–168, 2022.

[31] Eric M Pietras, Ranjani Lakshminarasimhan, Jose-Marc Techner, Sarah Fong, Johanna Flach, Mikhail Binnewies, and Emmanuelle Passegué. Re-entry into quiescence protects hematopoietic stem cells from the killing effect of chronic exposure to type i interferons. *Journal of Experimental Medicine*, 211(2) :245–262, 2014.

Chapitre 8
Prédire l'effet du traitement à l'IFNα : vers un outil d'aide à la décision clinique



Résumé

Dans ce chapitre, nous utilisons notre modèle à des fins de prédiction. Nous considérons le cas de patients mutés $JAK2^{V617F}$ et le modèle prenant en compte les variations de dose. Nous montrons que notre modèle est en capacité de prédire la réponse à long-terme du traitement à l'Interféron α, même avec peu d'observations, en utilisant un *a priori* appris suite à une estimation Bayésienne hiérarchique. En particulier, nous nous intéressons au temps nécessaire pour obtenir une rémission.

Nous utilisons ensuite notre modèle pour étudier des scénarios alternatifs de dose, et déterminer quelle aurait pu être la dose optimale à administrer au patient.

Enfin, nous montrons comment nous envisageons de rendre accessibles ces méthodes aux cliniciens, par l'intermédiaire d'une application web - en cours de développement au moment de l'écriture de ce manuscrit - qui servira d'outil d'aide à la décision.

Abstract

In this chapter, we use our model - the one that accounts for dose variation - for predicting purposes and consider the case of patients having the mutation $JAK2^{V617F}$. We show that our model can predict the long-term response to Interferon alpha treatment, even with few observations, using prior knowledge previously learned from a hierarchical Bayesian estimation.

After having illustrated our method's predictive capacity, we are interested in using our model to optimize IFNα therapy. In the example of a given patient, we study alternative dose scenarios than the actual ones and aim to determine which would have been the most suitable dose to administrate to our patient. We consider the optimal dose as being the one allowing to reach a fast remission while still limiting the risk of drug toxicity.

Finally, we show how we plan to make our methods accessible to clinicians through a web application - currently under development - that will serve as a decision support tool.

Table des matières

1 Introduction

Notre modèle, sélectionné au chapitre précédent, décrit la dynamique sous traitement à l'interféron alpha (IFNα) de la fréquence allélique (VAF) parmi des cellules matures (granulocytes) et de la fraction clonale (CF) en cellules mutées hétérozygotes ou homozygotes parmi des cellules immatures, en prenant en compte des changements de dose. Il permet ainsi d'estimer, pour chacun des patients ayant un néoplasme myéloprolifératif (NMP) positif $JAK2^{V617F}$, la dose minimale nécessaire pour permettre la rémission à long-terme.
Si une rémission peut être obtenue, la question qui se pose est celle du temps nécessaire pour l'atteindre. On s'intéressera ainsi dans ce chapitre à prédire l'effet à long terme du traitement. Pour cela, nous enlèverons à tour de rôle certains patients de la cohorte, calibrerons le modèle sans eux pour apprendre un effet populationnel qui sera ensuite utilisé comme *a priori* pour estimer les paramètres individuels du patient exclu. Nous pourrons alors étudier comment l'ajout progressif d'observations permet d'ajuster les prédictions à long-terme. Cette étude nous permettra également de valider le modèle, en évaluant la capacité qu'il a de prédire correctement les observations expérimentales n'ayant pas été utilisées pour sa calibration.

Si on est en mesure de prédire correctement la dynamique de la VAF sous traitement, pour une posologie donnée, on peut alors se poser la question de ce qui se serait passé en optant pour des stratégies différentes. Nous étudierons ainsi la question de l'optimisation du traitement, et illustrerons sur un exemple simple le potentiel de notre modèle comme un outil d'aide à la décision clinique.

Enfin, nous terminerons ce chapitre en présentant une première version, en cours de développement au moment de l'écriture de ce manuscrit, d'une application en ligne destinée aux cliniciens qui aura pour objectif d'intégrer les méthodes de prédiction et d'optimisation présentées dans ce chapitre.

2 Méthodes

2.1 Leave-one-out

Pour étudier les capacités prédictives de notre modèle, puis ensuite étudier différentes stratégies de dosage, nous considérons la cohorte de 19 patients mutés $JAK2^{V617F}$ étudiée au chapitre précédent, à laquelle nous allons enlever un individu (leave-one-out) pour lequel nous ferons de l'assimilation de données (voir § 2.2).
Nous notons $\mathcal{D}$ le jeu de données pour les 19 patients, et $\mathcal{D}_{-i}$ celui pour tous les patients à l'exception de l'individu i. Pour reprendre les notations introduites au chapitre 6, nous notons $\mathcal{I}^{(i)}$ l'ensemble des temps d'observations pour le patient i, et

$$\mathcal{D}^{(i)} = \left(t_k^{(i)}, \hat{n}_{k,wt}^{(i)}, \hat{n}_{k,het}^{(i)}, \hat{n}_{k,hom}^{(i)}, \hat{y}_k^{(i)}\right)_{k\in\mathcal{I}^{(i)}}$$

avec $\hat{n}_{k,wt}^{(i)}, \hat{n}_{k,het}^{(i)}, \hat{n}_{k,hom}^{(i)}$ qui représentent respectivement les nombres de progéniteurs wild-type (wt), mutés hétérozygotes (het) et homozygotes (hom) qui ont été échantillonnés et génotypés au temps $t_k^{(i)} \in \mathcal{I}^{(i)}$ pour le patient i, et $\hat{y}_k^{(i)}$ la mesure de sa VAF au niveau des cellules matures.

Nous étudierons plus particulièrement trois individus que nous exclurons à tour de rôle de la cohorte : les patients #12, 18 et 32. Ces patients sont choisis parce qu'on dispose pour eux de plusieurs observations avant $t = 300$ jours de traitement.
Nous considérons alors le modèle qui a été sélectionné à l'étape de la procédure de sélection présentée au chapitre précédent. Nous lançons trois procédures d'estimation hiérarchique (détails aux chapitres 6 et 7) correspondant respectivement aux jeux de données $\mathcal{D}_{-12}$, $\mathcal{D}_{-18}$ et $\mathcal{D}_{-32}$. De ces trois calibrations, nous ne retiendrons que l'estimation de la distribution de population.

Dans l'estimation Bayésienne hiérarchique, nous supposons que les paramètres individuels du modèle $\theta_k^{(j)}$ (paramètre k, patient $j \neq i$ avec i l'individu exclu) suivent *a priori* une distribution gaussienne (tronquée) de moyenne et de variance $\tau_k^{(i)}$ et $\sigma_k^{2\ (i)}$ respectivement :

$$\theta_k^{(j)} | \tau_k^{(i)}, \sigma_k^{2\ (i)} \sim \mathcal{N}_{c,k}\left(\tau_k^{(i)}, \sigma_k^{2\ (i)}\right)$$

où $\tau_k^{(i)}$ et $\sigma_k^{2\ (i)}$ sont des hyper-paramètres (ou paramètres de populations) qui sont également estimés. Après exécution de la procédure d'estimation, nous avons pour $\tau_k^{(i)}$ et $\sigma_k^{2\ (i)}$ l'estimation de leur distribution *a posteriori* à partir de laquelle nous estimons la moyenne *a posteriori*, étant données les observations $\mathcal{D}_{-i}$:

$$\bar{\tau}_k^{(i)} = \mathbb{E}[\tau_k^{(i)} | \mathcal{D}_{-i}]$$
$$\bar{\sigma}_k^{2\ (i)} = \mathbb{E}[\sigma^{2\ (i)} | \mathcal{D}_{-i}]$$

2.2 Assimilation de données

Soit i l'individu exclu de la cohorte. Les paramètres de ce patient (à l'exception des paramètres correspondant aux conditions initiales) sont supposés suivre *a priori* la distribution de population modélisée comme une loi gaussienne (tronquée) dont les paramètres de population (i.e. les hyper-paramètres) ont été estimés à l'étape décrite au paragraphe précédent.
Ainsi, le *prior* pour le paramètre k du patient i est le suivant :

$$\theta_k^{(i)} \sim \mathcal{N}_{c,k}\left(\bar{\tau}_k^{(i)}, \bar{\sigma}_k^{2\ (i)}\right) \tag{1}$$

La prise en compte progressive d'observations pour ce patient, par l'intermédiaire de la vraisemblance, permettra la mise à jour de cette distribution. C'est ce que nous appelons "assimilation de données".
Soit $T \in \{300, 600, 1000\}$ [jours] le temps (d'assimilation) jusqu'auquel nous considérons les observations de ce patient. Nous notons $\mathcal{I}_T^{(i)}$ l'ensemble des temps d'observation avant le temps d'assimilation. Le jeu de données du patient i utilisé pour la calibration du modèle sera alors :

$$\mathcal{D}_T^{(i)} = \left\{t_k^{(i)}, \hat{y}_k^{(i)}\right\}_{k \in \mathcal{I}_T^{(i)}} \bigcup \left\{\hat{n}_{k=1,wt}^{(i)}, \hat{n}_{k=1,het}^{(i)}, \hat{n}_{k=1,hom}^{(i)}\right\} \bigcup \left\{\hat{n}_{k',wt}^{(i)}, \hat{n}_{k',het}^{(i)}, \hat{n}_{k',hom}^{(i)}\right\}$$

avec k' correspondant à l'indice de l'observation effectuée à un temps proche de 300 jours après traitement.
Ainsi, nous ne faisons pas le choix d'utiliser toutes les observations disponibles pour le patient avant le temps d'assimilation, ceci afin de se rapprocher de conditions réalistes en vue d'une utilisation de notre modèle par des cliniciens. En effet, en routine clinique, on peut mesurer aisément une VAF dans les cellules matures, mais l'observation répétée de l'architecture clonale, elle, n'est pas envisageable. Nous choisissons ainsi de nous limiter à la donnée de deux observations parmi les cellules progénitrices : une en début de traitement aidant à estimer les conditions initiales, notamment concernant le statut mutationnel (la zigosité), et une autre à environ 300 jours après traitement, afin de voir la réponse au traitement au niveau des progéniteurs environ un an après traitement, choix que nous discuterons plus loin.
Le jeu de données contrôle, permettant d'évaluer la qualité des prédictions, sera noté :

$$\mathcal{D}_c^{(i)} = \mathcal{D}^{(i)} \setminus \mathcal{D}_T^{(i)}$$

Pour le patient i, pour le temps d'assimilation $T \in \{300, 600, 1000\}$, nous estimerons alors la distribution *a posteriori* de chacun de ses paramètres à partir des observations $\mathcal{D}_T^{(i)}$ et du prior (1) issu de l'estimation hiérarchique effectuée sur tous les autres patients.
Nous propagerons ensuite les incertitudes des paramètres à la sortie du modèle pour obtenir la dynamique en cours de traitement de la VAF dans les cellules matures ainsi que de la CF hétérozygote et homozygote dans les progéniteurs. Nous considérerons notamment, à chaque instant,

la valeur médiane et les intervalles de crédibilité à 95% que nous pourrons tracer et confronter aux données, que ce soit celles utilisées pour la calibration du modèle ou celles jouant le rôle de contrôle. La confrontation aux observations contrôles nous permettra d'évaluer la qualité des prédictions.
À noter que notre modèle prend en entrée les variations de dose en cours de traitement. Ici, nous les supposerons connues à tout instant (donc avant et après le temps choisi pour l'assimilation). Pour des temps supérieurs à la dernière posologie connue, nous considérerons alors cette dernière maintenue constante.

Nous pourrons alors estimer la réponse au traitement, au niveau des cellules matures (valeur médiane de VAF et intervalle de crédibilité à 95%), sur le long-terme, à 600, 1000 et 3000 jours de traitement, et lorsque applicable, estimer le temps pour atteindre la rémission, défini ici comme le temps au bout duquel la VAF dans les cellules matures passe en dessous de 5%.
Notons que cette approche a été étudiée notamment par Pedersen, Ottesen et al. [1, 2] dans le cas d'un modèle mathématique permettant de décrire la dynamique de réponse à l'IFNα pour des patients atteints de NMP. Une des différences majeures avec notre travail est qu'ils ne prennent pas en compte la zigosité, n'ayant accès qu'à l'information sur la VAF de cellules matures.

2.3 Optimisation du traitement

Si on est en mesure de bien prédire la dynamique future d'un patient sous traitement, ce qui revient à avoir un bon modèle et une estimation correcte de ses paramètres individuels, on peut alors chercher à aller plus loin dans l'analyse et s'interroger sur la façon dont on peut faire varier les doses pour obtenir une meilleure réponse, c'est-à-dire une rémission plus rapidement.
Nous considérerons pour cette étude le patient #12, qui a été soumis à une désescalade de dose à partir d'environ 300 jours de traitement, et pour lequel les prédictions concernant la VAF (qui sont celles qui nous intéressent pour estimer le temps de rémission) étaient correctes avec un temps d'assimilation $T = 600$ jours (comme nous le verrons en section 3.1 et sur la figure 1). Parmi les trois patients étudiés, il s'agit du seul pour lequel nous pouvions estimer un temps de rémission avec la posologie qui lui avait été administrée.
Nous considérons donc pour ce patient avoir estimé correctement la distribution *a posteriori* de ses paramètres :

$$p[\boldsymbol{\theta}^{(i)}|\mathcal{D}^{(i)}_{T=600}, \bar{\boldsymbol{\tau}}^{(i)}, \bar{\boldsymbol{\sigma}}^{\mathbf{2}(i)}]$$

Nous étudions alors différents scénarios pour la dose à administrer au-delà de 600 jours.
Dans ce chapitre, la seule stratégie étudiée sera celle du maintien d'une dose constante. Nous pourrons alors étudier ce qui se serait passé si, à partir de 600 jours de traitement, au lieu d'une désescalade de dose :

- La posologie avait été maintenue constante à 135 μg toutes les deux semaines ($d = 0.375$),
- La dose avait été diminuée à 135 μg toutes les 3 semaines ($d = 0.25$),
- La dose avait été augmentée à 135 μg / semaine ($d = 0.75$).

Rappelons que la quantité $d \in [0, 1]$ est définie comme la dose hebdomadaire rapportée à la dose maximale de 180 μg/semaine.

Plus généralement, nous pouvons estimer $\bar{T}_{rem}(d)$, définie comme le temps de rémission médian, c'est-à-dire la médiane de la distribution de T_{rem} obtenue après propagation des incertitudes du *posterior* de $\boldsymbol{\theta}^{(i)}$, pour toute valeur $d \in [0, 1]$.
Par la façon dont le modèle est construit, une augmentation de la dose d va conduire à une diminution de $\bar{T}_{rem}$, puisque $\bar{\Delta}^*_{het}$ et $\bar{\Delta}^*_{hom}$ sont des fonctions décroissantes de la dose d (relations sigmoïdes affines).
Ainsi, si on souhaitait seulement optimiser le temps de rémission pour qu'il soit le plus petit possible, il faudrait choisir la dose maximale de 180 μg/semaine ($d = 1$).
Néanmoins, le traitement à l'IFNα peut s'accompagner d'effets secondaires [3, 4], et présente également une certaine toxicité qui peut augmenter avec la dose. Ainsi, dans le problème d'op-

timisation qui consiste à déterminer une dose constante optimale, il faudrait certes prendre en compte le temps de rémission, mais également pénaliser les doses élevées. Intuitivement, si une forte augmentation de la dose ne permet de diminuer le temps de rémission que de quelques mois, au vu des échelles de temps considérées, cela ne serait pas intéressant. Les données concernant la toxicité de l'IFNα, qui auraient pu nous permettre de rationaliser la façon dont on pénaliserait l'augmentation de la dose, n'existent cependant pas dans la littérature scientifique. Nous allons ainsi choisir de minimiser la quantité totale d'IFNα administrée entre $T = 600$ jours et le temps nécessaire pour obtenir une rémission $\bar{T}_{rem}(d)$, c'est-à-dire que la dose optimale à administrer sera définie ici par :

$$d^* = \min_d \left(\bar{T}_{rem}(d) - 600\right) \times d \tag{2}$$

On pourrait également considérer que, en diminuant cette quantité, on optimise le coût financier lié au traitement. Mentionnons par exemple Pedersen et al. [5] qui prennent en compte des aspects économiques pour rationaliser le traitement à l'IFNα. Ils ne prennent néanmoins pas en compte le coût associé à la dose (considérant un coût approximatif de 500 USD pour un mois de traitement), mais plutôt le coût associé au traitement en fonction de la VAF en début de traitement. Dans leur travail, ils considèrent une rémission atteinte lorsque la VAF passe en dessous de 1%, et modélisent la dynamique de la VAF en cours de traitement par, soit une décroissance exponentielle, soit bi-exponentielle. Ainsi, avec leur modèle, on obtient naturellement des rémissions plus rapides lorsque le traitement débute à des VAF plus faibles.

3 Résultats

3.1 Prédictions

Les résultats de l'assimilation de données sont présentées sur les figures 1, 2 et 3 pour les patients #12, 18 et 32 respectivement. Sur ces figures on représente, en fonction du temps utilisé pour l'assimilation $T \in \{300, 600, 1000\}$ [jours], les dynamiques prédites jusqu'à 3,000 jours. On évalue également la qualité des prédictions et on prédit la VAF à long-terme, jusqu'à 3,000 jours.

Pour le patient #12, même avec un temps d'assimilation faible ($T = 300$ jours), on obtient des bons résultats en prédiction pour la VAF dans les cellules matures. Néanmoins, on sous-estime systématiquement la CF homozygote parmi les progéniteurs. L'ajout progressive d'observations expérimentales pour l'estimation des paramètres du patient #12 permet d'obtenir des dynamiques prédites pour la VAF pour lesquelles les intervalles de crédibilité sont plus restreints, et également d'affiner la prédiction sur le temps de rémission.
Quand on considère les données avant 300 jours, on prédit une rémission à environ 16 ans (valeur médiane), temps probablement surestimé. En effet, l'ajout des mesures de VAF jusqu'à 600 jours ajuste la prédiction à la baisse, à environ 14 ans, puis à 8 ans lorsque l'on inclut les données de VAF jusqu'à 1,000 jours. La borne supérieure de l'intervalle de crédibilité à 95% pour ce temps estimé de rémission est dans tous les cas égale à $+\infty$, ce qui signifie que, d'après notre modèle et l'estimation de ses paramètres, il y aurait toujours un risque que la baisse de la dose d'IFNα jusqu'à environ 30 μg/semaine conduise à une rechute.

Pour le patient #18, le seul des trois à avoir des clones hétérozygotes (en plus de clones homozygotes), les prédictions, lorsqu'on les compare aux valeurs réellement observées (mais non utilisées), sont plutôt satisfaisantes pour un temps d'assimilation $T = 300$ puis deviennent très bonnes pour $T = 600$ et $T = 1,000$, et ce à la fois pour les mesures de VAF, mais également les mesures de CF hétérozygote et homozygote. Pour ce patient, on prédit une VAF qui va se stabiliser à 50%. Le traitement de ce patient par des faibles doses permettrait d'éradiquer le clone homozygote qui, comme nous l'avions montré au chapitre 6, est ciblé efficacement par l'IFNα, même à faible dose. Néanmoins, la dose administrée serait trop faible pour cibler le clone hétérozygote, dont on prédirait une expansion au cours du temps. Ainsi, on ne prédit pas de rémission pour ce patient. D'après nos estimations du chapitre précédent, une dose minimale d'environ 50 μg/semaine serait requise pour permettre d'obtenir une rémission chez ce patient.

Pour le patient #32, les prédictions étaient plutôt correctes lorsqu'on considérait les observations avant $T = 300$ jours, même si les intervalles de crédibilité étaient très larges. En ajoutant alors les mesures de VAF jusqu'à $T = 600$ jours, les prédictions étaient fortement détériorées, le modèle s'ajustant trop fortement aux données utilisées pour sa calibration. Cela pourrait suggérer que le *prior* est trop vague, ou alors que la dynamique de ce patient est trop éloignée de celles des autres individus de la cohorte, de telle sorte que le *prior* soit largement dominé par la vraisemblance. Il se pose également la question du choix des observations à utiliser, en l'occurrence du choix de l'instant auquel on effectuerait la mesure d'architecture clonale (en plus de l'observation initiale). Nous abordons ce point au paragraphe suivant.
Pour $T = 1,000$, les prédictions sont correctes (sur les données de contrôle), avec néanmoins des VAF sur-estimées, conduisant à prédire une rechute plus rapide qu'elle ne le serait en réalité. Néanmoins, on estime que la diminution de la posologie à des doses légèrement inférieures à 45 μg/semaine va conduire à une rechute. D'après nos estimations du chapitre précédent, la dose à administrer à ce patient devrait être supérieure à 72 μg/semaine pour atteindre une rémission avec 95% de probabilité.

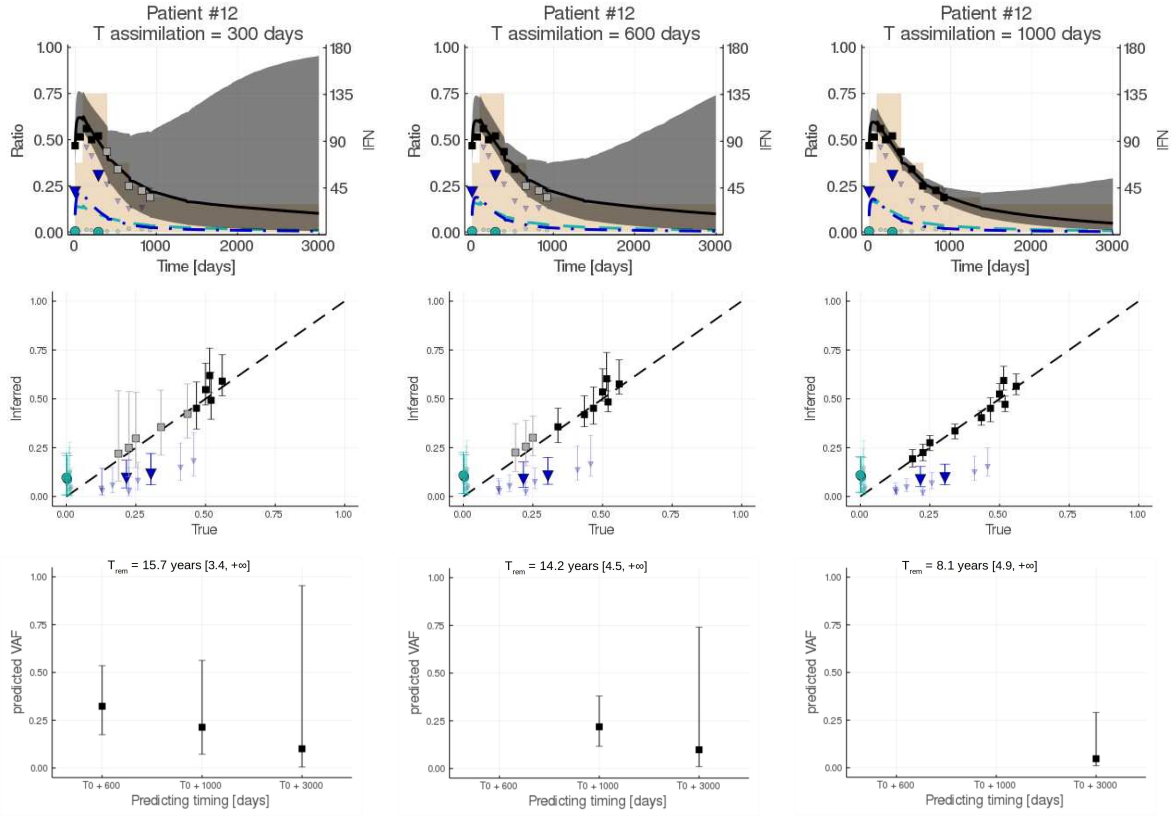


FIGURE 1 – Résultats de l'assimilation de données pour le patient #12. De gauche à droite, les temps d'assimilation sont $T = 300$, 600 et 1,000 jours.
En haut, on présente les dynamiques prédites. En noir la VAF (valeur médiane et intervalle de crédibilité à 95%), en vert et en bleu les CF (médianes) des progéniteurs mutés hétérozygotes et homozygotes respectivement. Les carrés, cercles et triangles de couleur correspondent aux données utilisées pour l'estimation des paramètres, ceux en grisés correspondent au jeu de données contrôle. En beige, on représente les variations de dose en cours de traitement.
Au milieu, on confronte les valeurs inférées / prédites (valeurs médianes, en ordonnées) par rapport à celles observées (en abscisses), soit pour le jeu de données contrôle (en grisé), soit celui utilisé pour l'estimation des paramètres (en couleur). Les barres d'erreur correspondent aux intervalles de crédibilité à 95%.
En bas, on prédit la VAF parmi les cellules matures (valeur médiane et intervalles de crédibilité à 95%) à 600 (uniquement dans le cas où le temps d'assimilation vaut $T = 300$), 1,000 (dans le cas $T = 300$ ou $T = 600$ jours) ou 3,000 jours après traitement. On indique sur ces figures l'estimation du temps de rémission (valeur médiane et intervalle de crédibilité à 95%), temps requis pour avoir une diminution de la VAF en dessous de 5%.

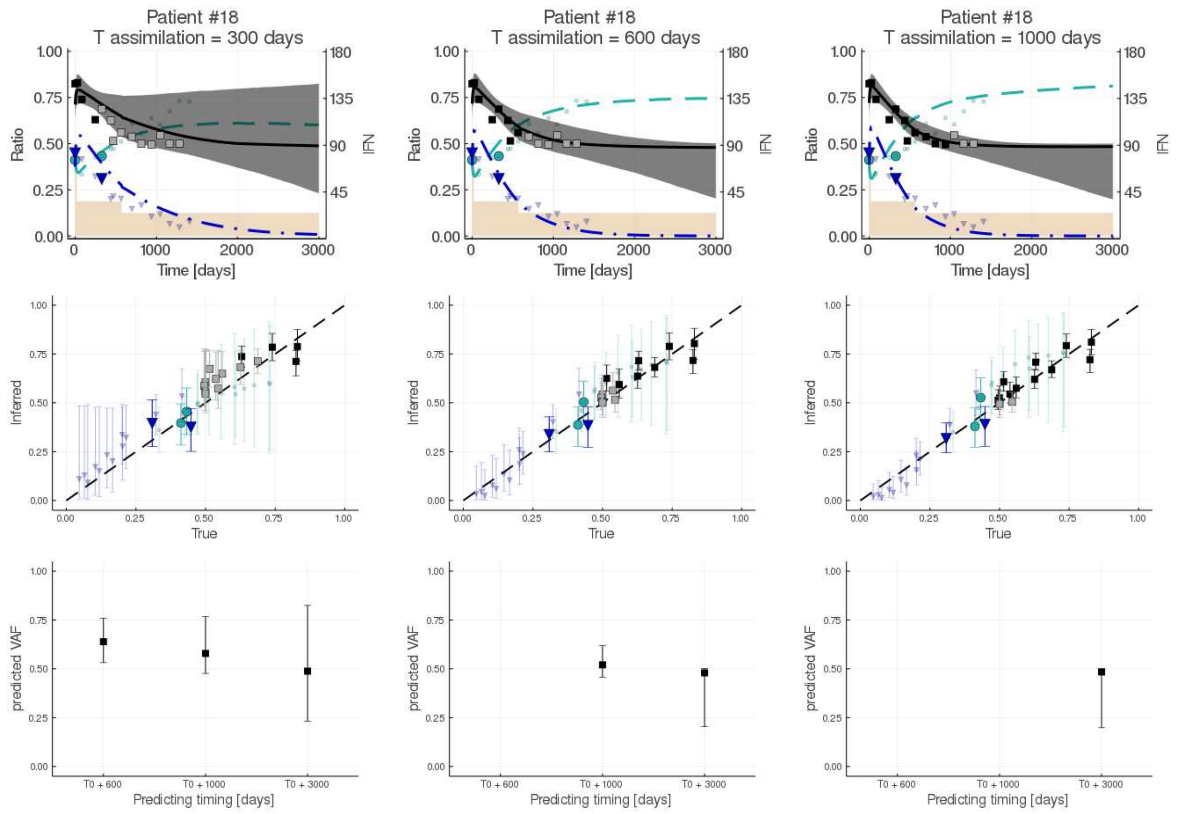


FIGURE 2 – Résultats de l'assimilation de données pour le patient #18.

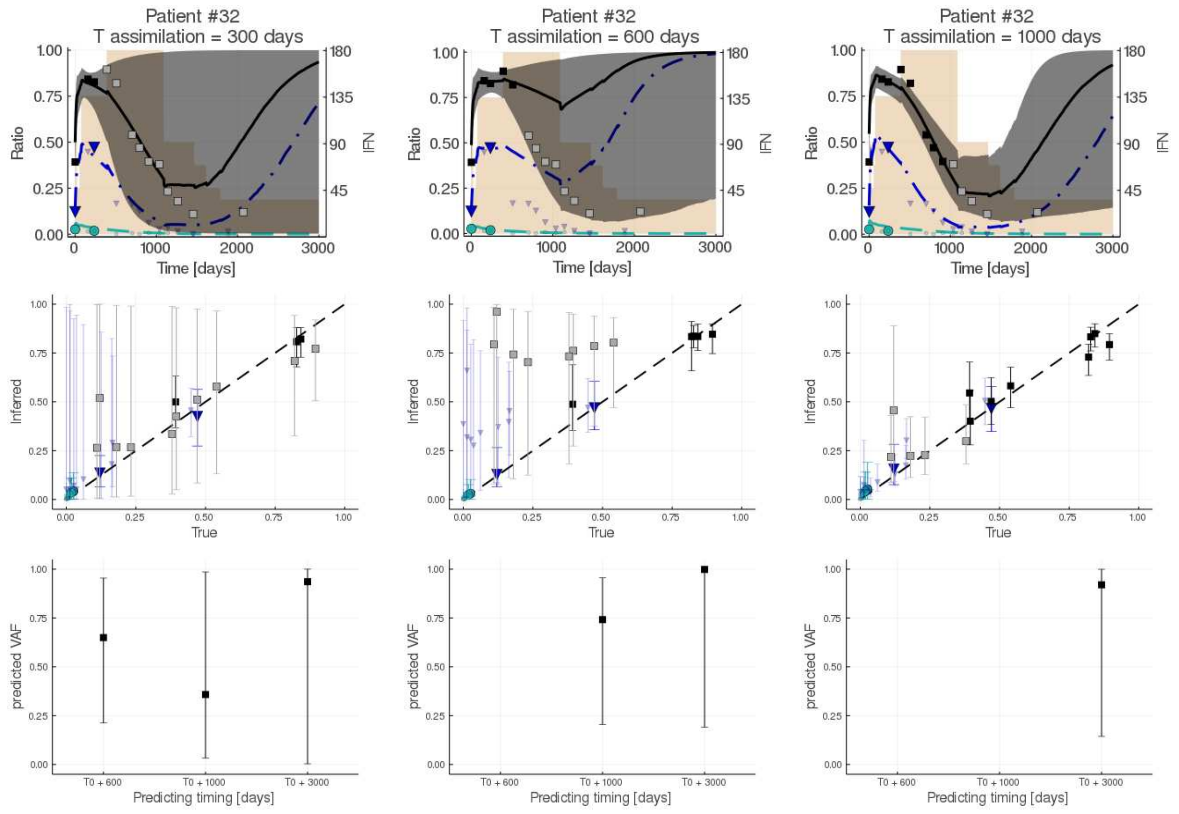


FIGURE 3 – Résultats de l'assimilation de données pour le patient #32.

3.2 Problématique de design expérimental

Pour le jeu de données servant à l'estimation des paramètres, nous avons fait le choix de considérer deux temps d'observation pour les cellules immatures, l'un en tout début de traitement, l'autre à environ 300 jours. Ce dernier choix est discutable. Nous illustrons sur la figure 4 son impact sur les inférences du patient #32, en montrant quels auraient été les résultats de l'estimation si on avait plutôt choisi comme seconde observation celle à $\sim$ 600 jours de traitement. Soulignons que la seule différence entre les deux prédictions présentées sur cette figure tient uniquement à une observation au niveau des cellules immatures.

Dans le cas où l'observation avait été choisie à $\sim$ 300 jours, les résultats sur les données de contrôle sont bien moins bons que dans le cas où l'observation est choisie à $\sim$ 600 jours.
Ces résultats indiquent que la qualité des estimations est sensible au choix de seconde observation pour les cellules immatures. Une bonne estimation étant cruciale, non seulement pour obtenir des bonnes prédictions, mais également pour le choix d'une dose optimale de traitement, il serait ainsi important de pouvoir déterminer à quel moment effectuer cette seconde mesure pour maximiser les chances d'obtenir de bonnes prédictions.
Il serait également pertinent, plus généralement, d'étudier la question du design expérimental optimal, c'est-à-dire la façon dont on devrait choisir le timing des observations expérimentales, au fur et à mesure du suivi d'un patient, afin de maximiser l'information obtenue.

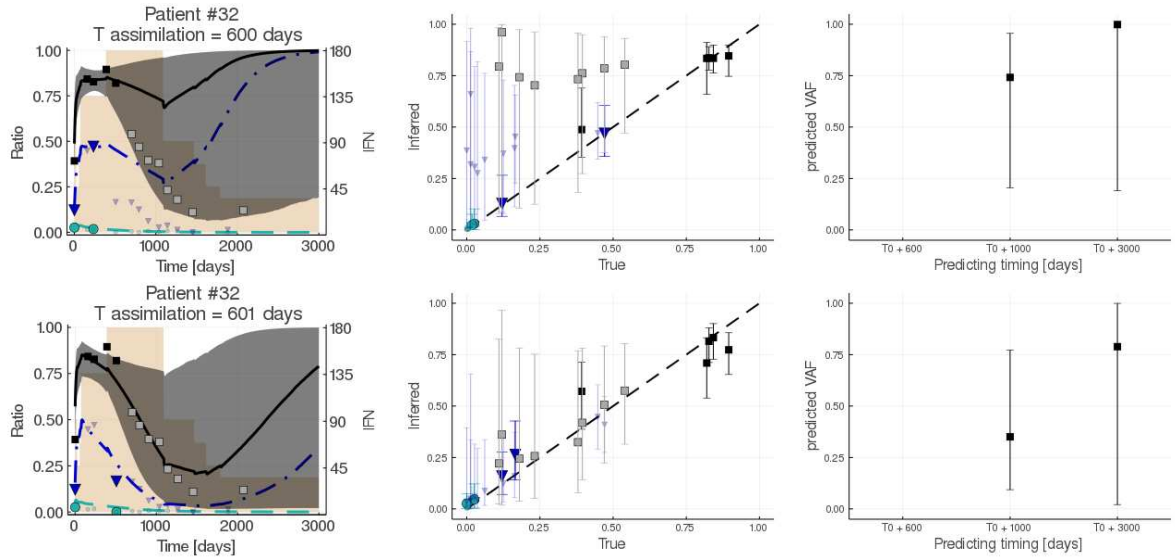


FIGURE 4 – Illustration de l'influence du choix du timing de la seconde observation sur la qualité de la prédiction, sur l'exemple du patient #32. Les résultats en haut correspondent à ceux de la figure 3 (pour une assimilation à $T = 600$ jours). En bas, on présente les résultats qu'on obtient lorsque, au lieu de choisir une mesure à $\sim$ 300 jours après traitement, on la choisit à $\sim$ 600 jours. Les résultats en prédiction, sur les données de contrôle, sont bien meilleurs dans ce dernier cas. Ici, cela peut s'expliquer par le fait qu'à environ 600 jours de traitement, la CF des progéniteurs homozygotes a commencé à diminuer, ce qui n'est pas encore le cas dans les cellules matures. Les prévisions à 1,000 et 3,000 jours après traitement vont également être très différentes entre les estimations étudiées ici.

3.3 Optimisation du traitement

Nous illustrons ici, sur l'exemple du patient #12, la façon dont on peut chercher à optimiser le traitement grâce à notre modèle et une estimation correcte de ses paramètres.
Les résultats sont présentés sur la figure 5. On se place dans le cas où le patient aurait été suivi jusqu'à 600 jours. En haut de la figure, nous rappelons les prédictions réalisées en considérant comme posologie, après 600 jours, celle réellement administrée. Ce patient a subi une désescalade de dose, passant de 135 μg d'IFNα toutes les 2 semaines (soit une dose équivalente à 67.5 μg/semaine) à $\sim$ 40 μg/semaine au-delà de 1,000 jours. Cette diminution de la dose au cours du temps, d'après nos estimations, ralentirait l'efficacité du traitement et conduirait à une rémission (i.e. une diminution de la VAF en-dessous de 5%) au bout de 14 ans.
On peut alors se demander ce qui aurait pu se passer pour ce patient si, passé 600 jours, il avait reçu une autre posologie que celle réellement administrée. Nous considérons, à titre d'exemple, 3 configurations :

1. Après 600 jours, on maintient constant la posologie (c'est-à-dire 135 μg d'IFNα toutes les 2 semaines, soit une dose de 67.5 μg/semaine)
2. Après 600 jours, on diminue la dose, passant de 135 μg d'IFNα toutes les 2 semaines à 135 μg d'IFNα toutes les 3 semaines, et on maintient cette dernière posologie constante
3. Après 600 jours, on double la dose, passant de 135 μg d'IFNα toutes les 2 semaines à 135 μg d'IFNα toutes les semaines, et on maintient cette dernière posologie constante

Dans le premier cas, le maintien de la posologie administrée à 600 jours aurait conduit à une rémission estimée (en moyenne) à $\sim$ 5 ans, soit une réponse obtenue bien plus rapidement que dans le cas de la désescalade de dose réellement subie par le patient. Dans le second cas, en diminuant la dose, passant à une administration toutes les 3 semaines (donc une dose de 45 μg/semaine), nous estimons que la réponse aurait été ralentie de deux ans, avec une rémission obtenue à environ 7 ans après début du traitement. Surtout, la borne supérieure de l'intervalle de crédibilité pour le temps estimé de rémission vaut $+\infty$, ce qui signifie qu'avec cette baisse de la dose, il y aurait un risque de rechute. Ces résultats suggéreraient qu'il ne serait pas judicieux de baisser la dose à 45 μg/semaine. Dans le troisième cas, en doublant la dose, passant à 135 μg par semaine au-delà de 600 jours, nous estimons que le temps moyen de rémission serait d'environ 4 ans, soit un an plus tôt que dans le cas un. Si on considère uniquement le gain dans le temps de réponse au traitement (temps pour atteindre une rémission), cette posologie pourrait sembler plus favorable. Néanmoins, pour gagner une année (sur un traitement chronique, donc de long terme), il nous faudrait ici doubler la dose, s'exposant à des risques liés à la toxicité du traitement.

Ainsi, sur les trois exemples précédents, on obtient que l'augmentation de la dose diminue le temps nécessaire avant d'obtenir une rémission, mais cette relation n'est pas linéaire. On peut ainsi généraliser et étudier la relation entre ce temps de rémission $\bar{T}_{rem}$ et la dose d administrée. Dans le cas du patient #12 "observé" jusqu'à 600 jours, cette relation est tracée en violet sur la figure 6.
Pour rationaliser le choix de la dose optimale, c'est-à-dire faire l'arbitrage entre gain de temps (avant d'atteindre la rémission) et la perte associée à l'augmentation de la dose (que ce soit en termes de coût économique ou de risques liés à la toxicité qui augmente avec la dose), nous choisissons de minimiser la quantité totale d'IFNα administrée jusqu'à atteindre cette rémission. Cette quantité, en fonction de la dose, est tracée en bleu sur la figure 6. Elle atteint un minimum pour $d^* = 54$ μg/semaine. À cette dose, nous prédisons une rémission à environ 5 ans et demi.

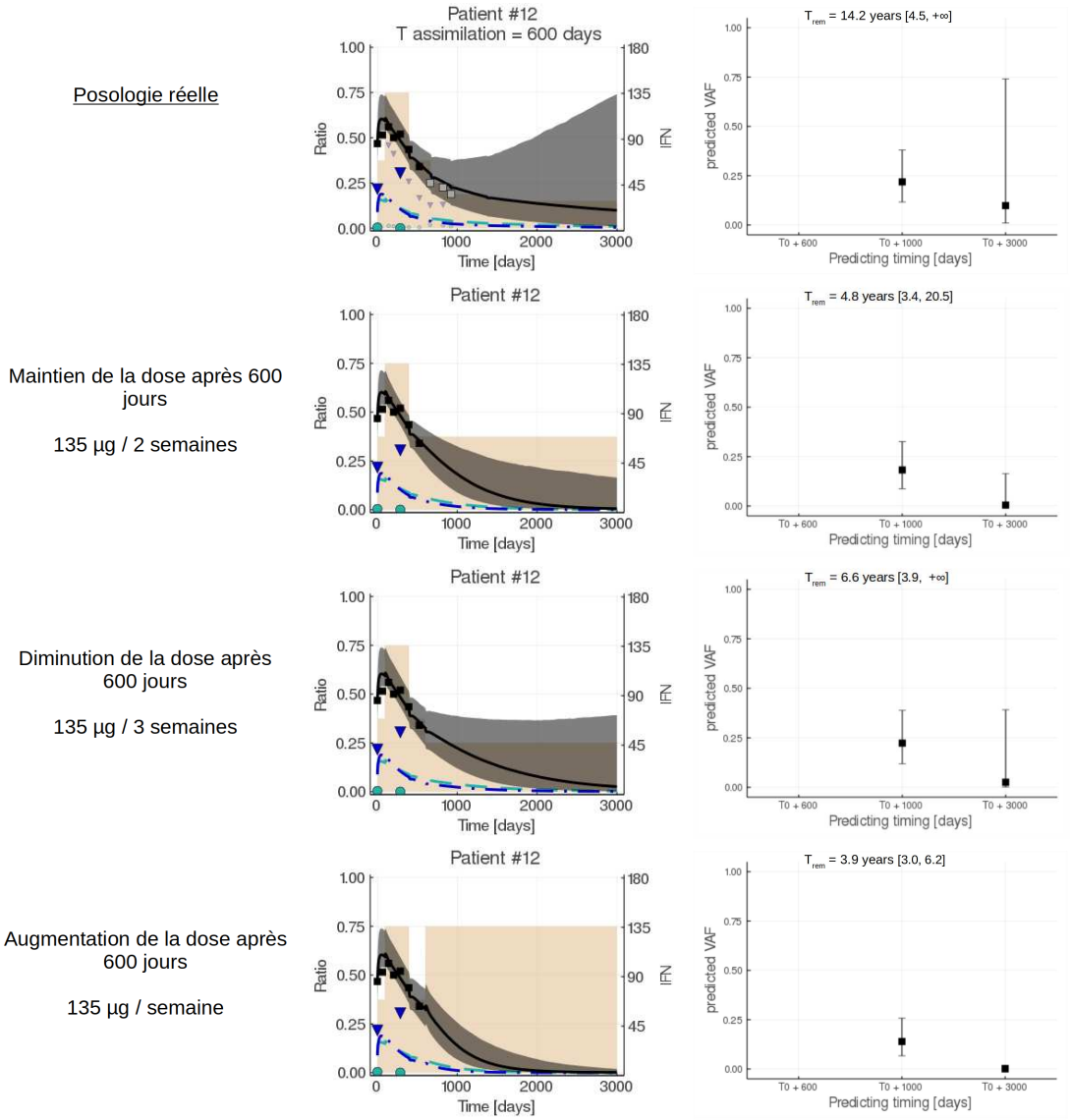


FIGURE 5 – Comparaison de différents scénarios de traitement, à partir de 600 jours, sur l'exemple du patient #12. Tout en haut, les résultats correspondent à ceux de la figure 1, où la posologie passé 600 jours est celle qui a réellement été administrée. On étudie ensuite, de haut en bas, le cas où on maintiendrait, passé 600 jours, une administration de 135 μg d'IFNα toutes les 2 semaines, au cas où on la réduirait à toutes les 3 semaines, et le cas où on l'augmenterait à toutes les semaines. Sur les figures de droite, on indique également l'estimation du temps de rémission (valeur médiane et intervalle de crédibilité à 95%), temps requis pour avoir une diminution de la VAF en dessous de 5%. L'augmentation de la dose conduit à une diminution du temps nécessaire pour atteindre la rémission qui n'est pas linéaire.

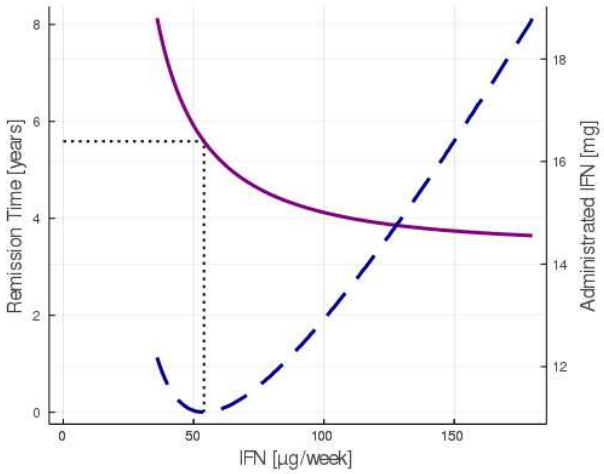


FIGURE 6 – Détermination d'une dose optimale de traitement à administrer (de façon constante) après 600 jours pour le patient #12. En violet, nous représentons le temps pour atteindre une rémission $\bar{T}_{rem}$ en fonction de la dose. La ligne bleue hachurée représente la quantité totale d'IFNα qui serait administrée entre 600 jours et le temps auquel on obtiendrait la rémission, en fonction de la dose. Cette quantité est minimale pour $d^* = 54$ μg/semaine correspondant à $\bar{T}_{rem}(d^*) = 5.6$ années.

4 Vers un outil d'aide à la décision

4.1 Description

Les méthodes que nous avons décrites plus haut, à savoir la prédiction de la réponse au traitement, du temps de rémission et de la dose optimale à administrer, ont vocation à être accessibles facilement aux cliniciens.
Nos efforts de recherche portent en effet sur le développement d'une médecine personnalisée, dans le cas du traitement des NMP à l'IFNα. Ainsi, nos méthodes doivent pouvoir être appliquées pour tout nouveau patient, et pas seulement ceux inclus dans la cohorte ayant été étudiée tout au long de ce manuscrit.

Notre choix, pour rendre accessibles les modèles et méthodes étudiés dans le cadre de cette thèse, est celui d'une application en ligne (un site web), appelée OptiMyN (du nom du projet "Optimize IFNα Therapy in Myeloproliferative Neoplasms").
Un clinicien pourrait s'inscrire sur le site et ajouter ses patients. Les fonctionnalités cibles et le parcours utilisateur sont présentés sur le schéma de la figure 7.
Dans le cas d'un patient non encore sous IFNα, notre application pourrait aider le clinicien à décider de la dose initiale (voir chapitre 7), comparer le profil du patient avec d'autres enregistrés (non étudié pour le moment), et recommander un design expérimental, c'est-à-dire préconiser les prochaines dates auxquelles il serait idéal d'avoir des mesures expérimentales pour bien prédire l'effet à long-terme du traitement (non étudié).
Dans le cas d'un patient déjà sous traitement, le clinicien peut lui ajouter des informations, telles que les variations de posologie et les mesures expérimentales effectuées pendant le suivi clinique. Trois fonctionnalités clé devraient alors lui être accessibles :

- Prédire l'effet à long-terme du traitement (voir § 3.1)
- Recommander une dose optimale (voir § 3.3)
- Recommander le timing auquel effectuer la (ou les) prochaine mesure expérimentale (voir § 3.2)

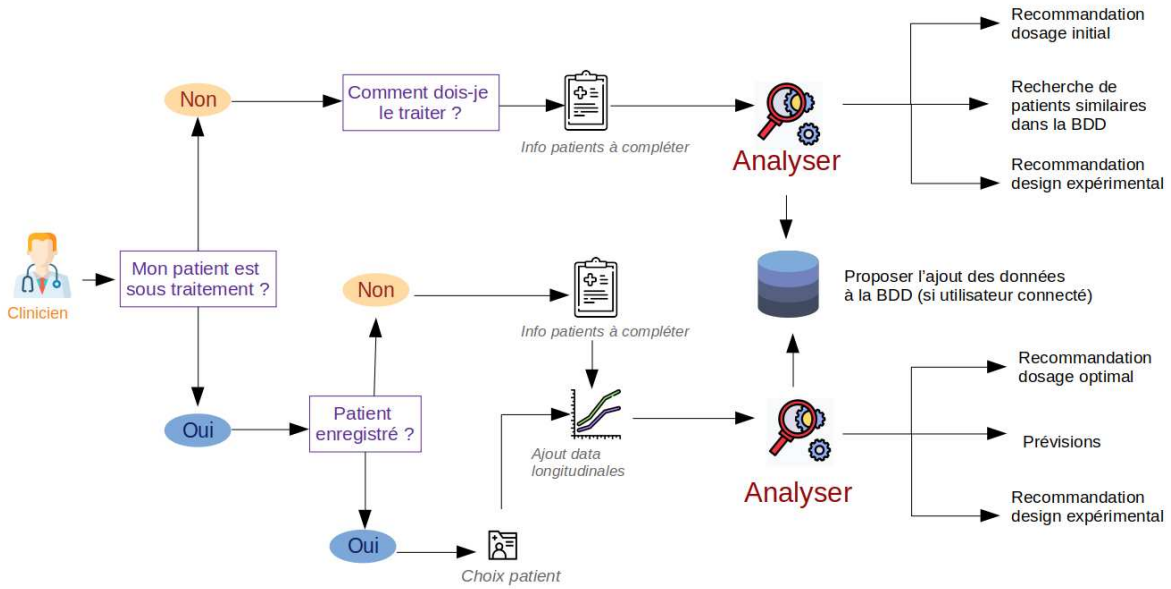


FIGURE 7 – Schéma du parcours utilisateur (en l'occurrence, l'utilisateur principal sera un clinicien) et des différentes fonctionnalités envisagées pour l'application OptiMyN.

4.2 Implémentation

Le développement de l'application OptiMyN a été initié par Agathe Le Galiot - alors ingénieure de recherche au laboratoire MICS - avec laquelle j'ai collaboré pour le design et le lien avec mon code de simulation. Sans rentrer dans les détails techniques, l'application peut se décomposer en trois composantes principales (voir figure 7) :

- Le front-end : c'est la partie interface avec l'utilisateur, lui permettant de remplir les formulaires pour la saisie des données, de naviguer entre les pages, d'exécuter les fonctionnalités de l'application.
- La base de données (BDD) qui permet le stockage des données enregistrées, permettant leur utilisation pour l'exécution des différentes fonctionnalités.
- Le back-end : c'est la partie codée en Julia, qui intègre le code implémenté dans le cadre de ce travail de thèse, pour effectuer les calculs concernant la prédiction et l'optimisation.

L'architecture est similaire à celle utilisée pour une application développée au laboratoire sur un sujet différent [6].

La version actuelle de l'application est la version alpha-test, que nous présentons plus en détail au paragraphe suivant.
Nous poursuivrons son développement afin de répondre aux besoins identifiés, que ce soit en termes d'interface (front-end) qu'en termes de méthodes implémentées (back-end).

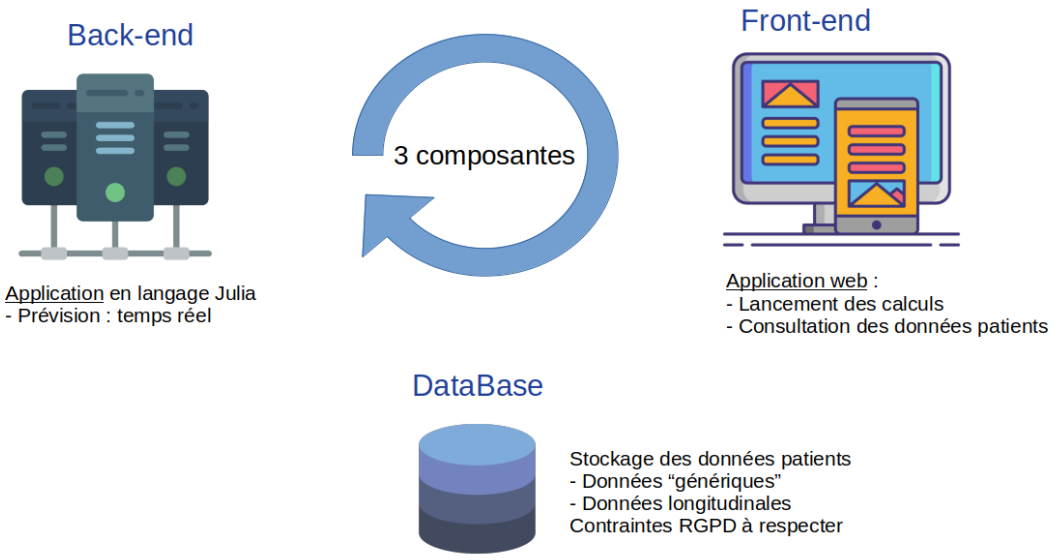


FIGURE 8 – Schéma simplifié de l'architecture de l'application OptiMyN. Une base de données (BDD) permettra le stockage des observation des patients (observations anonymisées pour respecter le Règlement général sur la protection des données - RGPD). Le back-end correspond au code écrit en langage Julia, développé dans le cadre de ce travail de thèse, permettant par exemple d'obtenir les résultats présentés à la section 3. Le front-end, écrit principalement en ReactJS, permet de naviguer sur l'application pour exécuter les différents calculs.

4.3 Version alpha-test

La version actuelle de l'application est la version alpha-test : c'est-à-dire que toutes les fonctionnalités cibles ne sont pas encore implémentées. Le site n'est ainsi accessible qu'en local.
La page d'accueil est présentée sur la figure 9. L'utilisateur, en particulier le clinicien, peut alors se connecter avec son identifiant (adresse e-mail) et mot de passe.
Il accédera alors à sa page (Fig. 10) où il peut suivre ses patients, et en ajouter de nouveaux.
En sélectionnant le patient de son choix, il arrive alors sur une page (Fig. 11) sur laquelle il peut ajouter les données longitudinales, à savoir les informations sur les changements de posologie (Fig. 11-A) ou sur les observations expérimentales (Fig. 11-B). Lorsque les données sont enregistrées, le clinicien pourra alors lancer les analyses (Fig. 12).
Pour le moment, nous n'avons intégré qu'une version simplifiée de la méthode de prédiction, se basant sur l'estimateur du maximum *a posteriori* des paramètres du modèle, à partir de l'algorithme CMA-ES.
Les méthodes présentées plus haut dans ce chapitre sont encore à ajouter et à adapter légèrement pour pleinement s'intégrer dans l'application.

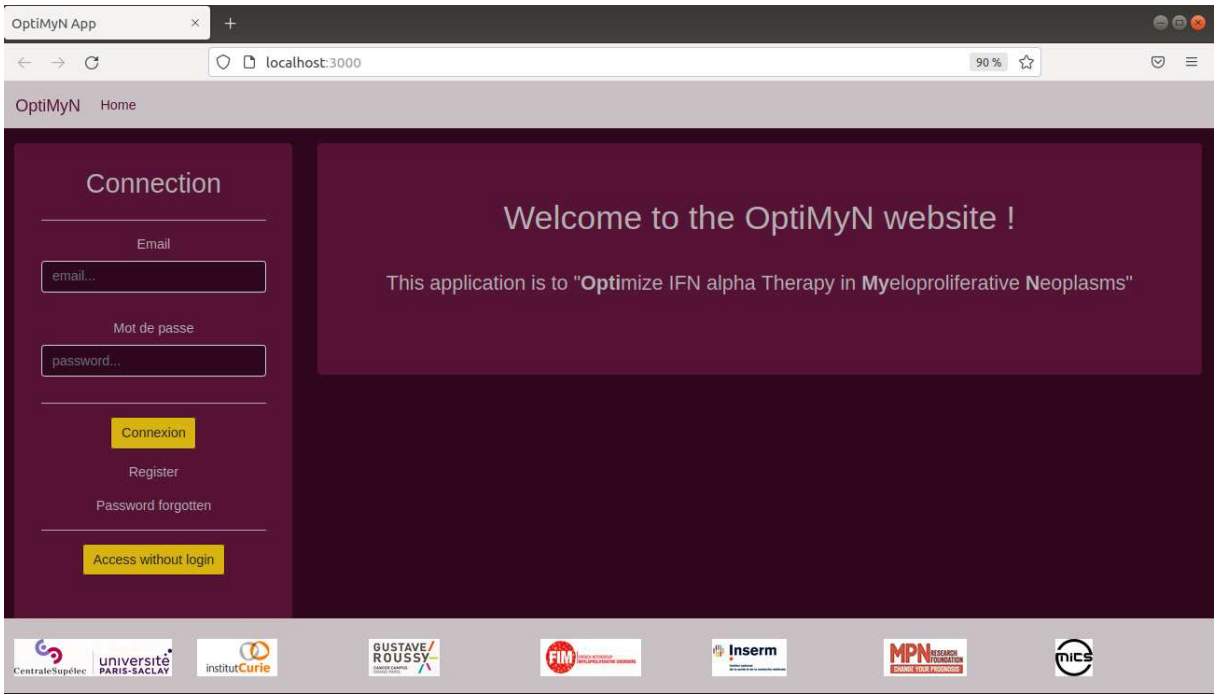


FIGURE 9 – Page d'accueil de l'application OptiMyN.

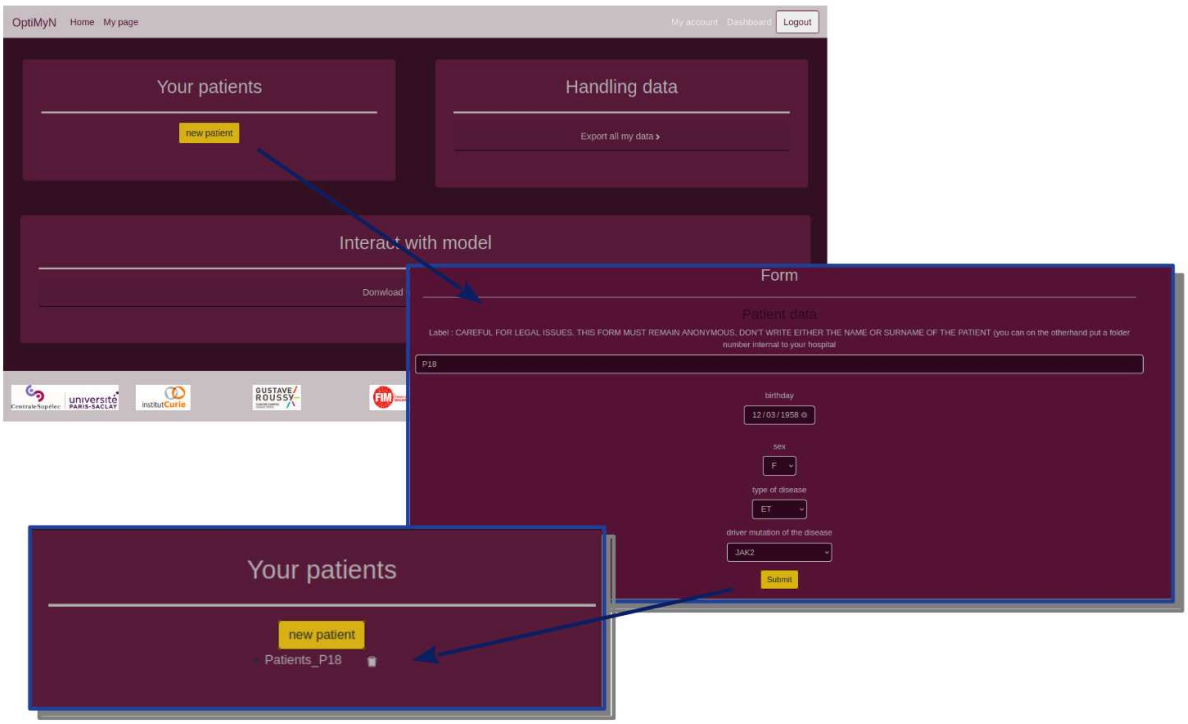


FIGURE 10 – Une fois connecté, le clinicien peut gérer ses patients, notamment en ajouter de nouveaux. Il lui sera alors demandé un identifiant pour le patient (qui doit garantir son anonymat), sa date de naissance, son sexe, sa maladie (TE, PV ou MFP) et la mutation motrice de cette dernière.

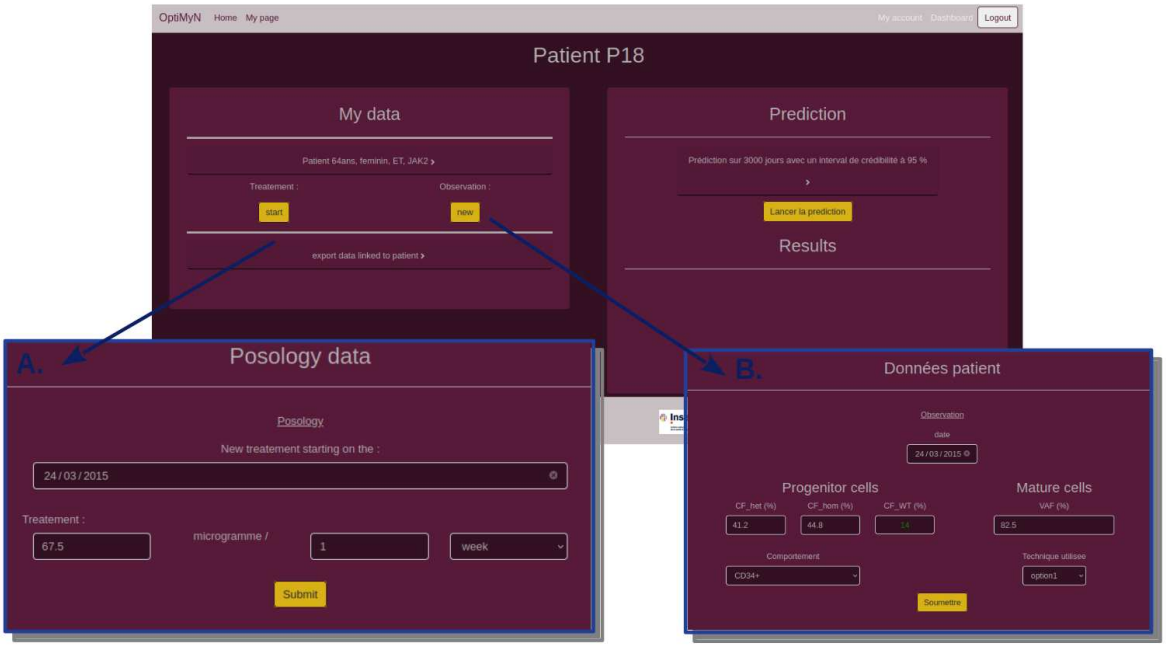


FIGURE 11 – En sélectionnant un patient, le clinicien peut alors ajouter des informations concernant les changements de posologie (A) ou les mesures expérimentales (B).

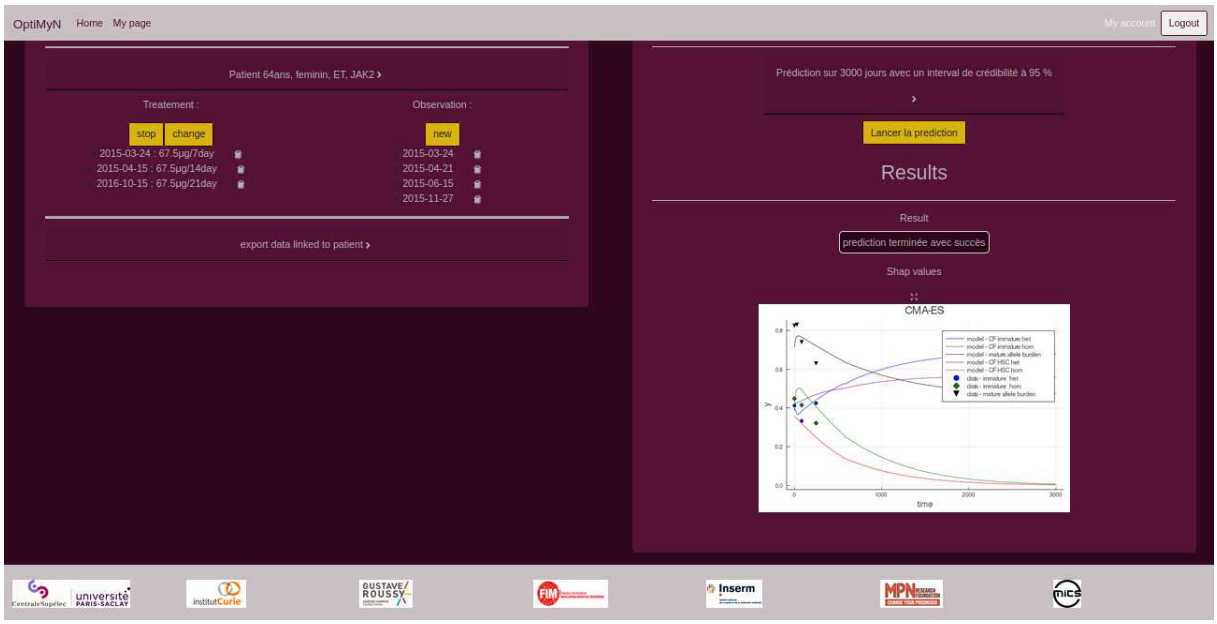


FIGURE 12 – Lorsque suffisamment de données sont disponibles, le clinicien peut lancer une prédiction pour le patient considéré.

5 Discussion

Dans ce chapitre, nous avons illustré comment le modèle étudié au chapitre précédent pouvait être utilisé à des fins de prédiction et d'optimisation du traitement. Nous avons notamment montré que notre modèle était en mesure de prédire correctement, pour un individu, ses observations futures à partir de quelques observations et d'un *prior* basé sur les estimations d'une cohorte (à laquelle l'individu considéré ne faisait pas partie). Ainsi, nous utilisons une distribution de population apprise sur d'autres patients pour améliorer l'estimation individuelle des paramètres du patient considéré, et éviter un risque d'*over-fitting* qui nuirait à la qualité des prédictions. Néanmoins, une limite du travail mené jusqu'à présent est que nous n'avons pas étudié en détail la question du choix des *priors*. Dans l'estimation Bayésienne hiérarchique, nous avons ainsi considéré que les paramètres étaient *a priori* indépendants les uns des autres, et que chacun suivait une loi gaussienne (tronquée sur un intervalle propre au paramètre), avec pour moyennes et variances des hyper-paramètres (ou paramètres de population) dont l'estimation des moyennes *a posteriori* est utilisée pour l'assimilation de données (pour un nouveau patient). Nous aurions pu faire d'autres choix que des lois normales, et également étudier le cas où les paramètres pourraient *a priori* être dépendants les uns des autres. De plus, l'estimation hiérarchique repose également sur le choix de *priors* pour les hyper-paramètres, en particulier les hyper-paramètres correspondant à la variance des lois gaussiennes. Pour ces hyper-paramètres, on impose *a priori* qu'ils soient les plus faibles possibles, afin de favoriser un rapprochement des paramètres individuels autour d'une valeur moyenne. Pour cela, nous avons fait le choix classique d'un *prior* distribué suivant une loi inverse-gamma (0,0) (voir chapitres 6 et 7). Néanmoins, d'autres choix sont possibles, permettant de moduler notre *a priori* sur l'importance de l'effet population (nous explorions ce sujet en annexe B du chapitre 3). Il serait ainsi intéressant d'étudier plus en détail le choix des *priors*. En favorisant encore plus l'effet population, on pourrait espérer avoir des prédictions meilleures avec moins de données, à condition que l'hétérogénéité inter-individuelle dans la réponse au traitement ne soit pas trop forte. Il serait également envisageable d'étudier un modèle hiérarchique à plusieurs niveaux, stratifiant par exemple les patients $JAK2^{V617F}$ suivant si leur maladie est la TE, MFP ou PV par exemple.
Nous ne considérons pas non plus d'*a priori* pour les conditions initiales. Néanmoins, l'âge au moment du traitement et le type de maladie pourraient éventuellement nous renseigner sur les conditions initiales, notamment la zygosité. Ainsi, il serait *a priori* moins probable de trouver une forte CF homozygote pour des patients à un stade peu avancé du NMP (par exemple la TE) ou pour des patients jeunes. Le modèle de développement des NMP, présenté au chapitre 5, permet d'inférer, à tout âge, la distribution de la CF hétérozygote qui pourrait ainsi servir de distribution *a priori*.
Nous avons présenté une méthode d'assimilation de données pour montrer comment l'ajout progressif d'observations pour un nouveau patient permettait de mettre à jour le *posterior* sur ses paramètres, et par conséquent les prédictions. Le terme assimilation de données utilisé dans ce chapitre peut paraître légèrement abusif. Il est souvent employé dans des cas de modèles complexes, coûteux à simuler, par exemple en météorologie. Dans ces cas-là, lorsque l'on rajoute des observations à un ensemble de données déjà existantes, on ne relance pas tout les calculs depuis le début, mais on adapte les résultats obtenus avec le précédent jeu de données, pour prendre en compte les nouvelles. Dans notre cas néanmoins, les observations ne sont pas si nombreuses (elles sont obtenues à une fréquence d'environ un trimestre) et il n'est pas coûteux, lorsque l'on rajoute des données, de relancer entièrement les estimations (dans le cas de l'estimation pour un patient, et non pas l'estimation Bayésienne hiérarchique qui elle est coûteuse en temps de calcul).
Pour faire nos prédictions et l'assimilation de données pour un nouveau patient, nous avons considéré avoir accès à ses mesures de VAF, à l'architecture clonale initiale ainsi qu'à un autre moment après traitement. Nous avons illustré l'influence du choix du *timing* de cette dernière observation sur la qualité de nos prédictions. Il serait ainsi intéressant d'étudier la question du design expérimental, afin d'estimer à quel moment effectuer la mesure expérimentale afin de maximiser nos chances d'avoir de bonnes prédictions. Nous nous sommes limités dans ce chapitre

à deux mesures d'architecture clonale, quand nous aurions pu en utiliser bien plus, ce qui aurait certainement amélioré la qualité de nos estimations. Néanmoins, notre objectif est le déploiement de nos méthodes en routine clinique. Or, dans ce cas là, la mesure répétée de l'architecture clonale n'est pas envisageable. D'ailleurs, considérer avoir cette information à deux instants est probablement trop optimiste, et il faudrait tendre vers des méthodes n'ayant pas du tout recours à l'information sur la CF dans les progéniteurs. Notre modèle, en l'état, a probablement trop de paramètres pour se passer de ces informations. D'où l'intérêt évoqué plus haut de travailler sur les *priors* pour favoriser encore plus l'effet populationnel, ainsi qu'avoir un *prior* sur la zigosité à l'instant initial. Notons que Pedersen, Ottesen et al. [1, 2] ont étudié des modèles basés uniquement sur la mesure de la VAF, et qu'il pourrait être intéressant de tester leurs modèles avec nos données. La limite de leur travail est qu'ils ne considéraient alors pas l'impact de la zigosité, dont nous avons néanmoins démontré l'importance dans le traitement à l'IFNα. Un moyen détourné existe pour avoir accès à une information sur la zigosité dans les cellules matures, à partir d'une mesure de VAF : celui d'utiliser l'haplotype 46/1 [7], dans le cas uniquement de patients hétérozygotes pour ce polymorphisme (mais il serait possible de déterminer d'autre polymorphismes aussi informatifs). En effet, il a été montré que généralement, dans le cas d'une recombinaison mitotique par laquelle une cellule mutée $JAK2^{V617F}$ hétérozygote donne une cellule mutée homozygote, la cellule deviendra également homozygote pour l'haplotype 46/1. Ainsi, la mesure de la VAF pour cet haplotype nous renseignera sur la proportion de cellules mutées $JAK2^{V617F}$ homozygotes. Notre modèle pourrait être adapté pour prendre en compte ces données.

Une fois qu'on est capable de prédire correctement la réponse à long-terme pour une posologie donnée, on peut alors étudier des stratégies de traitement alternatives afin d'estimer celle qui pourrait être la plus adéquate. Dans ce chapitre, nous n'avons que très partiellement étudié la question de l'optimisation, en se limitant au cas de l'étude d'une dose constante. Nous aurions pu nous intéresser à des stratégies du type "stop and go", néanmoins les cliniciens se sont montrés défavorables à ce type de stratégie, car il faut un certain temps pour s'adapter au traitement, et son interruption temporaire obligerait le patient à de nouveau s'y adapter lorsqu'il le reprendrait. Dans ce travail, nous n'avons parlé que de la dose, c'est-à-dire la quantité hebdomadaire moyenne d'IFNα administrée. Nous avons alors implicitement négligé l'influence potentielle de la posologie, supposant qu'il était équivalent d'avoir par exemple 90 μg d'IFNα par semaine *vs* 180 μg toutes les deux semaines. Nous pourrions étudier cette question par l'intermédiaire de modèles de types PK/PD (Pharmacokinetic-Pharmacodynamic) couramment utilisés dans ce type de problèmes. La difficulté néanmoins, pour prendre en compte finement la posologie, est qu'il est difficile de savoir, sur une thérapie de très long terme, les instants précis auxquels les patients ont pris leur dose d'IFNα.
Dans notre étude portant sur la détermination d'une dose optimale de traitement, nous avons considéré qu'il fallait minimiser la quantité totale de médicament administrée jusqu'à un temps de rémission. Cela repose sur l'argument que la toxicité du traitement augmenterait avec la dose, et ne prend pas en compte directement la posologie. Pour raffiner notre problème d'optimisation, il faudrait être en mesure de mieux définir le coût (i.e. la pénalisation) associé à l'augmentation de la dose. Les données sur la toxicité du traitement sont néanmoins manquantes dans la littérature scientifique. On pourrait également considérer, dans une approche dont l'idée est proche de celle de Pedersen et al. [5], le coût économique associé au traitement. Nous avons défini la rémission comme étant le moment au delà duquel la VAF dans les cellules matures passe en dessous de 5%. Ce critère est néanmoins arbitraire, et il serait intéressant d'étudier à quel niveau de VAF on peut effectivement décider d'arrêter le traitement. Cela pourrait se faire en considérant le temps qui serait nécessaire pour que l'envahissement clonal conduise de nouveau au déclenchement d'un NMP. L'arbitrage se ferait en considérant l'âge du patient, et le temps nécessaire à l'envahissement. Il serait également pertinent de considérer le fait que l'IFNα puisse ne plus cibler efficacement les cellules mutées lorsque celles-ci sont présentes en trop faible quantité. Ainsi, pour étudier la rémission et la question de l'arrêt du traitement, il faudrait adapter notre modèle, pour passer d'un formalisme déterministe à un formalisme stochastique.

Comme nous l'avons mentionné précédemment, notre objectif est de permettre le déploiement de nos méthodes en contexte clinique. Pour cela, nous avons initié le développement d'une application web à destination des cliniciens, qui devrait leur permettre d'ajouter leurs données patients pour directement utiliser nos méthodes de calculs. On pourrait également, sous condition d'avoir l'accord du clinicien, profiter de cette application pour étudier plus de patients et raffiner ainsi nos méthodes de prédiction et optimisation.
Pour être vraiment utilisable, l'application devrait néanmoins pouvoir se baser sur des mesures de VAF. Il sera également essentiel d'inclure les cliniciens dans les futurs développements de cette application, pour s'assurer qu'elle correspond à leur besoin, et faire naître leur intérêt pour l'utilisation de cet outil. De plus, il serait intéressant de pouvoir combiner différents modèles, suivant différentes pondérations, pour améliorer la robustesse de nos résultats.

Références

[1] Rasmus K Pedersen, Morten Andersen, Trine A Knudsen, Vibe Skov, Lasse Kjær, Hans C Hasselbalch, and Johnny T Ottesen. Dose-dependent mathematical modeling of interferon-α-treatment for personalized treatment of myeloproliferative neoplasms. *Computational and Systems Oncology*, 1(4) :e1030, 2021.

[2] Johnny T Ottesen, Rasmus K Pedersen, Marc JB Dam, Trine A Knudsen, Vibe Skov, Lasse Kjær, and Morten Andersen. Mathematical modeling of mpns offers understanding and decision support for personalized treatment. *Cancers*, 12(8) :2119, 2020.

[3] Peter C Trask, Peg Esper, Michelle Riba, and Bruce Redman. Psychiatric side effects of interferon therapy : prevalence, proposed mechanisms, and future directions. *Journal of Clinical Oncology*, 18(11) :2316–2326, 2000.

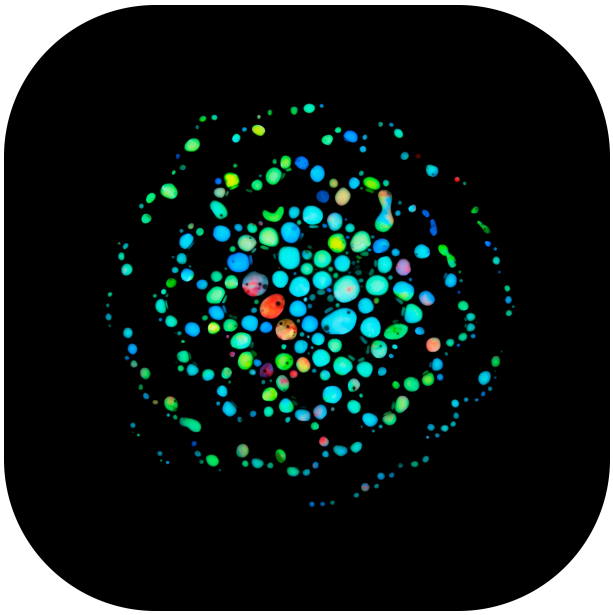
[4] Francis E Lotrich, Mordechai Rabinovitz, Patricia Gironda, and Bruce G Pollock. Depression following pegylated interferon-alpha : characteristics and vulnerability. *Journal of psychosomatic research*, 63(2) :131–135, 2007.

[5] Rasmus K Pedersen, Morten Andersen, Trine A Knudsen, Zamra Sajid, Johanne Gudmand-Hoeyer, Marc JB Dam, Vibe Skov, Lasse Kjær, Christina Ellervik, Thomas S Larsen, et al. Data-driven analysis of jak2v617f kinetics during interferon-alpha2 treatment of patients with polycythemia vera and related neoplasms. *Cancer medicine*, 9(6) :2039–2051, 2020.

[6] Benjamin Besse, Alison Dormieux, Laura Mezquita, Renaud Monnet, Melodie Tazdait, Ludovic Lacroix, Etienne Rouleau, Julien Adam, Jordi Remon Masip, María Bluthgen, et al. Prediction of the molecular status in non-small cell lung cancer based on metastatic pattern : A free webtool powered by artificial intelligence., 2020.

[7] S Hasan, B Cassinat, N Droin, JP Le Couedic, F Favale, B Monte-Mor, C Lacout, M Fontenay, C Dosquet, C Chomienne, et al. Use of the 46/1 haplotype to model jak2v617f clonal architecture in pv patients : clonal evolution and impact of ifnα treatment. *Leukemia*, 28(2) :460–463, 2014.

Chapitre 9
Synthèse et perspectives



Résumé

Dans ce dernier chapitre, nous reviendrons sur les principaux résultats biologiques et cliniques présentés tout au long de ce manuscrit, et récapitulerons les principales méthodes mathématiques employées pour les obtenir.
Plutôt que de conclure, nous présenterons les différentes perspectives en vue d'améliorer nos méthodes et nos modèles et indiquerons la façon dont les différents travaux présentés dans ce manuscrit seront liés les uns aux autres et poursuivis dans le cadre du projet **OptiMyN**.

Table des matières

1 Synthèse

1.1 Principaux résultats biomédicaux

Nous présentons dans cette section quelques-uns des principaux résultats biologiques et cliniques obtenus pour chacun des chapitres de ce manuscrit (à l'exception du chapitre introductif).

Chapitre 2 - Cellules hématopoïétiques : un continuum d'états révélé par cytométrie de masse

Au chapitre 2, nous avons analysé des données *single-cell* de souris, obtenues par cytométrie de masse. À partir d'une proposition de méthode d'inférence de trajectoire, nous avons cherché à reconstruire la voie de différenciation mégacaryocytaire, ceci afin de mettre en évidence des différences entre souris WT et souris mutées $CALR^m$ de type T1 ou T2 quant aux voies de signalisation différentiellement activées. Alors que les trajectoires d'expression de l'intensité de la CALR extracellulaire ne présentaient pas de différences significatives entre souris WT, T1 et T2, au cours de la différenciation, ce n'était pas le cas de la CALR intracellulaire pour laquelle les intensités d'expression dans les cas T1 et T2 étaient systématiquement en dessous de ceux WT. Quand on s'intéressait à comprendre ce qui pouvait différer entre les cellules mutées T1 *vs* T2, nous avons trouvé une tendance à ce que la voie P-AKT soit activée différentiellement, avec des intensités d'expression pour le marqueur P-AKT qui restaient plus faibles dans le cas des souris T2 par rapport aux T1 tout au long de la trajectoire de différenciation.

Chapitre 3 - Modélisation du temps de division des cellules souches hématopoïétiques

Au chapitre 3, nous avons trouvé que la distribution des temps de première division de cellules souches HSC*, mises en culture en présence d'un cocktail de cytokines, était correctement décrite par un modèle gamma généralisé, alors que les temps de division des cellules moins immatures, à savoir les MPP et HPC, se distribuaient plutôt suivant des modèles plus parcimonieux (gamma ou log-normal). Les HSC* pourraient avoir besoin de plus de temps pour sortir de quiescence, phénomène qui pourrait être pris en compte par des modèles tels que le modèle gamma généralisé.

Chapitre 4 - Modèle de prolifération et différenciation des cellules souches et progénitrices

Nous avons trouvé au chapitre 4 que la dynamique de prolifération et de différenciation des cellules souches et progénitrices était plus en adéquation avec des modèles qui faisaient l'hypothèse d'une synchronicité (dans les temps de division des cellules sœurs) et d'une concordance (quant au choix des types cellulaires des cellules sœurs).

Chapitre 5 - Apparition et développement des Néoplasmes Myéloprolifératifs

Nous avons estimé que les mutations $CALR^m$ pourraient être acquises plus tard au cours de la vie que les mutations $JAK2^{V617F}$ (en moyenne 25 *vs* 15 ans), et conférer un avantage prolifératif aux cellules souches plus important dans le cas $CALR^m$ par rapport à $JAK2^{V617F}$. Par une méthode différente de celles de Williams et al. [1] ou Van Egeren et al. [2], nous avons retrouvé les estimations concernant le *timing* d'acquisition de la mutation $JAK2^{V617F}$. La méthode que nous avons proposée, puisqu'elle se place non pas à l'échelle d'un individu mais d'une population de patients, ouvre la voie à la mise en place de méthodes de dépistage précoce. En particulier, nous avons montré qu'il pourrait être pertinent de chercher à dépister la mutation $JAK2^{V617F}$, et que l'âge optimal pour effectuer un prélèvement de sang serait aux alentours de 30 ans [3].

Chapitre 6 - Modélisation de l'effet du traitement à l'Interféron α

Nous avons proposé un modèle de l'effet de l'IFNα sur des patients atteints de NMP. En utilisant une méthode d'estimation Bayésienne hiérarchique, à partir d'observations longitudinales obtenues pour une cohorte de 48 patients, nous avons mis en évidence que l'IFNα ciblait plus efficacement les HSC mutées $JAK2^{V617F}$ homozygotes que celles hétérozygotes et que, pour ces dernières, le dosage pouvait avoir un impact, avec des doses plus élevées d'IFNα qui conduisent à des meilleures réponses moléculaires.
Pour les patients $CALR^m$, au contraire, nous avons plutôt mis en évidence que les fortes doses étaient associées à des moins bonnes réponses moléculaires. Nous préconisions alors l'utilisation de l'IFNα à des doses élevées pour permettre d'atteindre une rémission moléculaire chez les patients $JAK2^{V617F}$ [4].

Chapitre 7 - Déterminer une dose minimale d'Interféron α pour patients ayant la mutation $JAK2^{V617F}$

Au chapitre 7, Nous avons étendu le modèle du chapitre 6 pour prendre en compte les variations de dose au cours du traitement, chez les patients $JAK2^{V617F}$. Nous avons montré qu'en dessous d'une certaine dose, il y avait un risque que le traitement ne permette plus d'obtenir une rémission sur le long-terme. Nous avons illustré la possibilité de déterminer une dose limite (inférieure) d'IFNα qui soit propre à chaque patient. De plus, pour un patient non encore traité à l'IFNα, nos résultats suggèrent de choisir comme dose de départ 70 μg/semaine [5].

Chapitre 8 - Prédire l'effet du traitement à l'IFN α : vers un outil d'aide à la décision clinique

Dans ce chapitre, nous avons illustré comment le modèle du chapitre 7 pouvait être utilisé à des fins prédictives et d'optimisation du traitement. Nous avons alors implémenté ces méthodes dans une application destinée à être accessible en ligne, pour les cliniciens.

1.2 Méthodes mathématiques utilisées

Pour réaliser le travail présenté dans ce manuscrit, nous nous sommes attachés à mettre en place différentes méthodes mathématiques en vue de répondre à différentes questions de recherche, toutes liées à l'étude de la dynamique de populations de cellules hématopoïétiques, saines ou mutantes, avec ou sans traitement par Interféron alpha.

À chaque question de recherche était généralement associée un jeu de données expérimentales, qui pouvait nécessiter un prétraitement préliminaire, comme dans le cas des données obtenues par cytométrie de masse (voir chapitre 2), ainsi qu'une modélisation appropriée de l'incertitude : censure par intervalle (chapitre 3), perte d'information par échantillonnage (chapitre 4), incertitudes sur la mesure (chapitre 6).

Nous cherchions ensuite à modéliser les phénomènes biologiques susceptibles d'avoir produit de telles observations. Nous avons été amenés à construire des modèles stochastiques (chapitres 3 et 4) lorsque les dynamiques étaient étudiées à l'échelle de quelques cellules, des modèles déterministes (chapitres 6 et 7) lorsque nous nous intéressions à des populations cellulaires de grande taille, ou encore hybride (chapitre 5). Les modèles déterministes employés consistaient en des systèmes d'équations différentielles ordinaires dont nous pouvions facilement obtenir une solution analytique. Nos modèles stochastiques s'apparentaient à des modèles de branchement ; nous les étudions à partir de simulations numériques.
À chaque fois, l'objectif était de construire un modèle approprié pour répondre à la question de recherche, qui soit le plus simple possible (notamment en termes de parcimonie, c'est-à-dire du nombre de degrés de liberté) tout en essayant de capturer les facteurs biologiques les plus déterminants. La construction des modèles représente une part importante de ce travail, qui a

nécessité de nombreux échanges avec les biologistes.
L'autre part importante de ce travail de thèse était celle de l'inférence statistique. Une fois les modèles construits, nous voulions en estimer les paramètres à partir des observations expérimentales disponibles. Notre approche a alors principalement consisté à utiliser des méthodes d'inférence Bayésienne - méthodes ABC-SMC (chapitre 5), algorithme de Métropolis-Hasting (chapitre 8), méthodes Bayésiennes hiérarchiques (chapitre 6 et 7) - ou encore des méthodes d'optimisation numérique - à savoir l'algorithme CMA-ES - pour trouver des paramètres minimisant une certaine distance (chapitre 4). Nous avons d'ailleurs souvent utilisé l'algorithme CMA-ES pour initialiser nos algorithmes d'inférence Bayésienne (chapitres 6, 7 et 8).

Entre la partie modélisation et la partie inférence statistique, il y a en fait un va-et-vient. Pour s'assurer de l'identifiabilité des modèles étudiés, nous générions des jeux de données synthétiques puis nous essayions de retrouver les paramètres des modèles les ayant générés, en utilisant nos méthodes d'estimation (chapitres 5 et 6). Lorsque les modèles n'étaient pas identifiables, nous reprenions alors le travail sur la modélisation, en cherchant à faire des simplifications supplémentaires pour fixer certains paramètres constants, à partir d'*a priori* biologiques. L'utilisation de méthodes d'analyse de sensibilité, à partir du calcul d'indices de Sobol (chapitre 4), a également été un moyen de simplifier certains modèles.

L'utilisation de modèles mathématiques avait parfois pour objectif de tester différentes hypothèses. Nous nous placions alors dans un contexte de sélection de modèles, en essayant non seulement de trouver les valeurs des paramètres des modèles, mais également de trouver quel modèle était le plus susceptible de conduire aux observations expérimentables étudiées. Nous avons par exemple utilisé des méthodes de sélection de modèles basées sur les critères AIC ou BIC et l'estimation du maximum de vraisemblance (chapitre 3) ou sur le critère DIC après estimation Bayésienne hiérarchique (chapitre 7). Au chapitre 5, nous avons utilisé une procédure correspondant à un mélange de modèles en sélection de modèles Bayésienne.

Enfin, nous évaluions la robustesse de nos résultats et conclusions, notamment en testant la capacité prédictive de nos modèles (chapitre 5 et 8).

2 Limites et perspectives

2.1 Problématiques d'estimation de paramètres

La question de l'estimation des paramètres s'accompagne d'enjeux méthodologiques. Au cours de ce travail, nous avons fait face à deux problématiques principales. La première se rencontre lorsque le modèle est trop complexe et qu'on ne peut pas en exprimer de vraisemblance. C'était le cas avec les modèles étudiés au chapitre 4 et 5. La méthode ABC, utilisée pour approcher la distribution *a posteriori* des paramètres, repose alors sur la construction de statistiques descriptives. Nous avons pu constater l'influence du choix des statistiques descriptives sur les résultats de l'estimation, et le besoin de recourir à des méthodes de construction de statistiques descriptives qui ne biaisent pas les résultats. Pour poursuivre le travail initié au chapitre 4, nous prévoyons d'explorer plus en détail les problématiques associées à la construction des statistiques descriptives, que ce soit pour l'estimation des paramètres ou la sélection de modèles.

Une autre problématique fréquemment rencontrée en statistique Bayésienne est celle du choix des *priors*. Nous avons rencontré cette problématique principalement dans le cas de l'estimation hiérarchique, où l'utilisation d'hyper-paramètres permet d'introduire un effet populationnel. Ces hyper-paramètres ont eux-mêmes des *priors* qui leur sont associés. En particulier, on choisit généralement comme *a priori* sur la variance de la distribution de population qu'elle soit la plus faible possible. Derrière cette formulation un peu vague, on peut en fait effectuer différents choix, qui vont conduire les paramètres individuels à être plus ou moins proches. En perspective, nous souhaiterions étudier l'effet de ces *priors* sur les hyper-paramètres, notamment concernant leur

importance en vue d'utiliser les résultats de l'estimation pour de l'assimilation de données et à des fins prédictives.

2.2 Analyse de données de cytométrie de masse

Au chapitre 2, nous avons exploré le continuum d'états formé par les cellules hématopoïétiques. Nous nous sommes intéressés à la distribution de ces cellules, et avons cherché à reconstruire la dynamique de différenciation et de prolifération par la construction de chemins aléatoires parcourant des graphes. Notre méthode devrait être étudiée plus en détail, notamment d'un point de vue théorique, et confrontée à d'autres algorithmes de l'état de l'art (par exemple celle de Wolf et al. [6]). Il serait également intéressant d'étudier la pertinence des méthodes de Graph Attention Network [7] dans ce contexte et de voir comment les effets batch peuvent être corrigés. Une des problématiques rencontrée sur ce sujet porte sur la prise en compte de la dynamique de prolifération : certaines cellules sont plus représentées que d'autres car elles sont plus différenciées et par conséquent ont subi plus de divisions. Ceci représente un enjeu en termes d'analyse, car les cellules les plus intéressantes à étudier sont également celles les moins différenciées (donc les plus souches) qui sont par conséquent les plus rares. Il pourrait alors être intéressant d'étudier ces données sous l'angle de la modélisation. Nous pourrions construire un modèle dynamique de mégacaryopoïèse, en étudier son état stationnaire et en estimer les paramètres à partir des observations par cytométrie de masse. L'utilisation d'un modèle pourrait également permettre de générer des données synthétiques afin d'évaluer comment les méthodes de l'état de l'art permettent d'inférer des trajectoires de différenciation dans ce cas.

Enfin, les méthodes étudiées pourraient également être utilisées dans le cas d'observations par cytométrie en flux. Même si, dans ce cas, moins de marqueurs sont accessibles à l'observation, on observe quand-même que les cellules immatures se distribuent suivant un continuum, avec une frontière flou entre les HSC*, MPP et HPC. L'utilisation des méthodes développées précédemment pourrait alors permettre de caractériser les cellules progénitrices, étudiées aux chapitres 3 et 4, non plus par un type discret mais par un degré de différenciation.

2.3 Développement et initiation des NMP

Nous avons proposé au chapitre 5 un modèle de développement des NMP. Ce modèle est pour le moment assez simple et pourrait notamment être complexifié de deux manières. Tout d'abord, nous pourrions prendre en compte également les patients avec des sous-clones homozygotes, dont nous pourrions modéliser l'apparition à partir d'un sous-clone hétérozygote puis le développement du clone hétérozygote et homozygote en parallèle. Ensuite, afin d'éviter une croissance irréaliste du nombre de cellules mutantes au cours du temps, nous pourrions envisager d'introduire des mécanismes de régulation dans notre modèle.
Dans les deux cas, ces améliorations conduiraient à des modèles moins parcimonieux, soulevant la problématique de la bonne estimation des paramètres. Dans le cas d'une modification du modèle pour prendre en compte les cellules homozygotes, on pourrait alors utiliser les données de patients possédant des clones homozygotes, ce qui ferait que le modèle pourrait rester identifiable. Une des pistes à l'étude pour complexifier le modèle tout en rendant possible l'estimation de ses paramètres serait d'utiliser des données expérimentales supplémentaires, en l'occurrence celles de Williams et al. [1] et Van Egeren et al. [2]. Ils observent par séquençage WGS plusieurs cellules progénitrices puis construisent des arbres phylogénétiques. Pour tirer profit de ces données, il faudrait alors être en mesure de générer une phylogénie à partir de notre modèle de développement des NMP. On pourrait alors avoir une approche à mi-chemin entre la nôtre et celle de Williams et al. [1]. Williams et al. faisaient l'estimation des paramètres d'un modèle pour chacun des patients, quand nous considérions un modèle dont les paramètres étaient les mêmes pour toute une population de patients. En perspective, on pourrait ainsi envisager une approche de type hiérarchique.

De plus, notre modèle de développement des NMP pourrait également être calibré à partir

d'observations expérimentales obtenues pour d'autres populations d'individus. En particulier, nous en avons identifié deux : des patients avec des syndromes de Budd Chiari (patients NMP révélés à l'âge de 35 ans en moyenne par une thrombose hépatique) et des individus issus de familles avec des prédispositions génétiques au développement des NMP. Ces deux types de populations sont caractérisées par le fait que les patients déclenchent très tôt des NMP. Il serait alors intéressant de regarder si on estime pour eux qu'il est plus probable que la mutation $JAK2^{V617F}$ soit apparue à la naissance ou dans l'enfance.

2.4 Projet OptiMyN

Né de la collaboration entre les équipes d'Isabelle Plo (Institut Gustave Roussy), de Leïla Perié (Institut Curie), et de Paul-Henry Cournède (CentraleSupélec), le projet OptiMyN a vu le jour pendant ce travail de thèse. Le projet bénéficie d'un financement de l'Institut Thématique Multi-Organismes (ITMO) sur des fonds administrés par l'INSERM. La poursuite de ce travail se fera majoritairement dans le cadre de ce projet. C'est notamment à travers ce projet que le lien entre les différents chapitres de ce manuscrit sera fait.

OptiMyN signifie : Optimize IFN alpha Therapy in Myeloproliferative Neoplasms. Dans ce projet, il s'agit alors de réussir à optimiser le traitement à l'IFNα et de poursuivre ainsi les travaux introduits au chapitre 8.
Pour cela, comme schématisé sur la figure 1, nous allons chercher à mettre en place des modèles et des méthodes mathématiques permettant l'intégration d'observations hétérogènes, en provenance de différentes expériences, pour répondre à trois objectifs principaux qui sont :

- Approfondir notre compréhension du mécanisme d'action de l'IFNα ;
- Prédire avec le moins d'incertitude possible, l'effet de l'IFNα sur le long terme ;
- Optimiser le traitement, si possible en fonction de différentes caractéristiques cliniques.

Pour cela, nous allons chercher à améliorer notre modèle de l'effet de l'IFNα (introduit au chapitre 6, complexifié au chapitre 7), à la fois sa structure mais également la façon d'en estimer les paramètres. Nous illustrons sur la figure 2 la façon dont nous envisageons de complexifier le modèle. Il s'agira notamment de continuer le travail sur le modèle d'hématopoïèse à court terme (chapitres 3 et 4), en intégrant des données de patients avant et après mise sous traitement à l'IFNα. Les résultats concernant l'effet de l'IFNα sur l'hématopoïèse à court terme pourront alors permettre de complexifier le modèle long-terme, mais également de justifier ou de remettre en question certaines simplifications que nous avons pu être amenés à faire au chapitre 6.

Dans le cadre de ce projet, nous voulons également développer des modèles qui se baseront sur l'hypothèse d'un continuum et non plus d'un ensemble de compartiments de populations, ceci afin d'être plus en accord avec les découvertes récentes concernant la structure de l'hématopoïèse. Les travaux présentés au chapitre 2 ont été réalisés dans ce sens. Enfin, nous chercherons à prendre en compte plus finement l'impact des mutations associées sur la réponse au traitement.

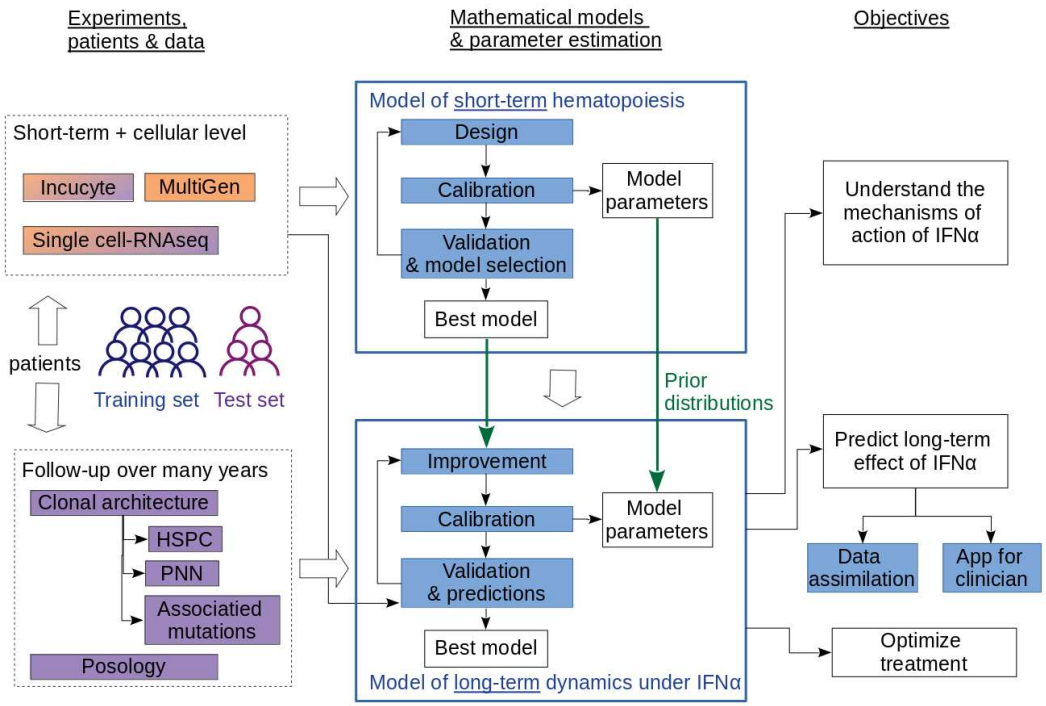


FIGURE 1 – Vue d'ensemble de la méthodologie envisagée pour mener à bien le projet OptiMyN. Les tâches en bleu correspondent aux tâches à réaliser par l'équipe de Paul-Henry Cournède, celles en violet à celles devant être réalisées par l'équipe d'Isabelle Plo et celles en orange par l'équipe de Leïla Perié.

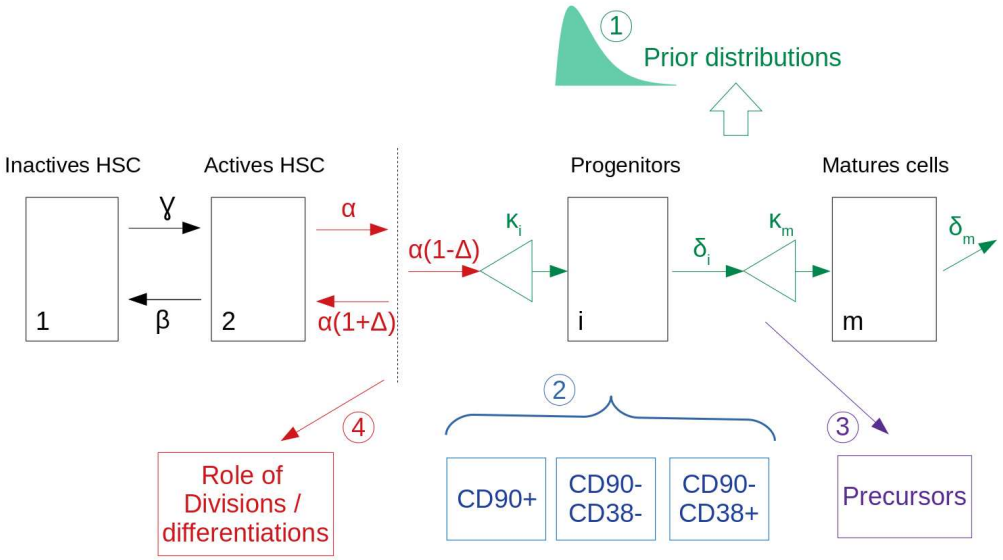


FIGURE 2 – Schéma du modèle compartimental du chapitre 6 et des différentes étapes envisagées pour l'améliorer. 1) Utilisation des connaissances apprises lors du travail sur la modélisation de l'hématopoïèse à court terme (chapitre 4), et qui pourraient être incluses au modèle compartimental par l'utilisation de *priors* ; 2) Séparation des populations de progéniteurs, entre les HSC* ($CD90^+$), MPP ($CD90^-CD38^-$) et HPC ($CD90^-CD38^+$) (chapitres 3 et 4) et prise en compte du fait que l'hématopoïèse est un processus continu plutôt que discret (chapitre 2) ; 3) Ajout d'informations pour les cellules précurseurs ; et 4) Affiner le modèle concernant la dynamique des HSC (chapitre 5).

Références

[1] N. Williams, J. Lee, E. Mitchell, L. Moore, E. J. Baxter, J. Hewinson, K. J. Dawson, A. Menzies, A. L. Godfrey, A. R. Green, *et al.*, "Life histories of myeloproliferative neoplasms inferred from phylogenies," *Nature*, vol. 602, no. 7895, pp. 162–168, 2022.

[2] D. Van Egeren, J. Escabi, M. Nguyen, S. Liu, C. R. Reilly, S. Patel, B. Kamaz, M. Kalyva, D. J. DeAngelo, I. Galinsky, *et al.*, "Reconstructing the lineage histories and differentiation trajectories of individual cancer cells in myeloproliferative neoplasms," *Cell stem cell*, vol. 28, no. 3, pp. 514–523, 2021.

[3] G. Hermange, A. Rakotonirainy, M. Bentriou, A. Tisserand, M. El-Khoury, F. Girodon, C. Marzac, W. Vainchenker, I. Plo, and P.-H. Cournède, "Inferring the initiation and development of myeloproliferative neoplasms," *Proceedings of the National Academy of Sciences*, vol. 119, no. 37, p. e2120374119, 2022.

[4] M. Mosca, G. Hermange, A. Tisserand, R. Noble, C. Marzac, C. Marty, C. Le Sueur, H. Campario, G. Vertenoeil, M. El-Khoury, *et al.*, "Inferring the dynamics of mutated hematopoietic stem and progenitor cells induced by ifnα in myeloproliferative neoplasms," *Blood, The Journal of the American Society of Hematology*, vol. 138, no. 22, pp. 2231–2243, 2021.

[5] G. Hermange, W. Vainchenker, I. Plo, and P.-H. Cournède, "Mathematical modelling, selection and hierarchical inference to determine the minimal dose in ifn alpha therapy against myeloproliferative neoplasms," *arXiv preprint arXiv :2112.10688*, 2021.

[6] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis, "Paga : graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells," *Genome biology*, vol. 20, no. 1, pp. 1–9, 2019.

[7] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv :1710.10903*, 2017.