

# THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université  
le 17 septembre 2021 par

Léonard HÉRAULT

## Étude du vieillissement des cellules souches hématopoïétiques par la biologie des systèmes :

Séquençage d'ARN et réseau booléen de régulation génétique à l'échelle de la cellule

**Discipline**

Informatique

**École doctorale**

ED 184

**Laboratoires**

Institut de Mathématiques  
de Marseille (I2M)  
CNRS UMR7373

Centre de Recherche en  
Cancérologie de Marseille  
(CRCM)  
CNRS UMR7258  
Inserm U1068

**Composition du jury**

Elisabeth REMY DR,  
Univ. Aix-Marseille (CNRS UMR7373)

Directrice de thèse

Estelle Duprez DR,  
Univ. Aix-Marseille  
(CNRS UMR7258, Inserm U1068)

Co-Directrice de thèse

Loïc PAULEVÉ CR,  
Univ. Bordeaux (CNRS UMR5800)

Rapporteur

Thierry JAFFREDO DR,  
Univ. Sorbonne (CNRS UMR7622)

Rapporteur

Laurence CALZONE IR,  
Mines ParisTech (Inserm U900)

Examinateuse

Denis THIEFFRY PR,  
ENS, Univ. PSL  
(CNRS UMR8197, Inserm U1024)

Examinateur

Carl Hermann MCF,  
Univ. Heidelberg (Health Data Science Unit)

Invité





Je soussigné, Léonard Hérault, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique d'Élisabeth Remy et Estelle Duprez, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 15/07/2021



Cette œuvre est mise à disposition selon les termes de la [Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International](#).

# Résumé

L'avancée récente des techniques de séquençage à l'échelle de la cellule unique permet aujourd'hui l'analyse des systèmes biologiques à une échelle beaucoup plus fine qu'auparavant via le développement de nouvelles approches bioinformatiques qui s'appuient sur l'intelligence artificielle et la masse considérable de données et connaissances biologiques disponibles.

Dans cette thèse, nous avons mis en œuvre ces nouvelles technologies pour étudier l'effet du vieillissement sur l'hématopoïèse, le processus de formation et de renouvellement des cellules sanguines. L'hématopoïèse repose sur les propriétés uniques d'un réservoir de cellules, les cellules souches hématopoïétiques (CSH) capables soit de s'auto-renouveler pour maintenir leur caractère souche, soit de se différencier en cellules fonctionnelles, à durée de vie limitée. L'homéostasie de ce réservoir qui repose sur des groupes de CSH différentes se détériore avec l'âge. En effet, il a été observé une accumulation de CSH peu fonctionnelles dans la moelle osseuse de souris âgées ce qui promeut l'immunosénescence ainsi que le développement d'hémopathies.

Afin de comprendre les mécanismes qui régulent cette perte d'équilibre associée au vieillissement physiologique, nous avons séquencé le transcriptome à l'échelle de la cellule unique de plus de 15000 CSH isolées à partir de souris jeunes et âgées. Ces données de séquençage ont été analysées à l'aide de différentes approches d'apprentissage supervisées et non supervisées pour la classification et la représentation des cellules dans des espaces de dimensions réduits, notamment avec la construction d'une pseudo-trajectoire de différenciation. Ces analyses transcriptomiques nous ont permis de caractériser avec précision la population de CSH et ses changements liés à l'âge. Nous avons notamment mis en évidence chez les souris âgées une diminution de l'amorçage des CSH vers les différents lignages hématopoïétiques à l'exception du lignage mégacaryocytaire. Nous avons pu montrer que cette diminution se traduisait dans une accumulation de cellules âgées quiescentes, concomitante à un blocage précoce de leur différenciation.

Dans un second temps nous avons utilisé ces résultats ainsi que la connaissance actuelle sur la biologie de la CSH pour bâtir un réseau booléen de régulation génétique afin de modéliser le processus de différenciation de ces cellules. Pour ce faire, nous avons mis en œuvre et

adapté une méthode récemment proposée pour inférer des réseaux de régulations génétiques sous la forme d'un problème de satisfiabilité booléenne à l'aide de la programmation de contraintes dynamiques. Nous avons également travaillé à la sélection d'un modèle parmi l'ensemble des solutions, puis analysé ce modèle et l'impact des perturbations en simulant les altérations observées lors du vieillissement au niveau transcriptomique. Nous avons ainsi pu mettre en évidence le rôle majeur des facteurs de transcription clés, Egr1, Junb et Gata2, dans les changements de la dynamique de différenciation des CSH au cours du vieillissement.

Mots clés : Cellule souche hématopoïétique, séquençage d'ARN cellule unique, inférence de réseaux de régulation génétique, modélisation logique, vieillissement

# Abstract

Recent advances in single cell sequencing technologies now enable biological system analysis at a much finer scale than before, thanks to the development of new computational biology approaches relying on artificial intelligence and the considerable amount of biological data and knowledge now available.

In this thesis, we implemented these new technologies to study the ageing alterations of haematopoiesis, the process of formation and renewal of all blood cells. The constant production of these short-life cells is based on an undifferentiated cell, the haematopoietic stem cell (HSC). This crucial process relies on the unique properties of HSC population to either self-renew to maintain the HSC pool or to differentiate into functional cells. The functionality of this pool relies on different groups of HSCs and deteriorates with age. Thus, an accumulation of poorly functioning HSCs is observed in the bone marrow of aged mice animals, promoting immunosenescence and development of myeloid leukemias and anaemias.

In order to understand the mechanisms regulating this loss of balance, we sequenced the single-cell transcriptome of more than 15,000 HSCs isolated from young and old wild-type mice. These sequencing data were analysed using different supervised and unsupervised learning approaches for the classification and representation of cells in reduced dimensional spaces, including the construction of a differentiation trajectory. This allowed us to accurately characterise the HSC population and its age-related changes at the transcriptome level. In particular, we have shown a decrease in the priming of HSCs to the different haematopoietic lineages in aged animals, with the exception of the megakaryotic lineage. We were able to show that this decrease was reflected in an accumulation of quiescent aged cells concomitant with an early blockage of their differentiation.

In a second step we used these results together with the current knowledge on HSC biology to build a Boolean gene regulatory network to model the differentiation of these cells. To do so, we implemented and adapted a recently proposed method for inferring such networks in the form of a Boolean satisfiability problem by defining dynamical constraints with Answer Set Programming. We also worked on the selection of a model from the solution set. Then, we analysed this model and the impact of perturbations simulating the alterations observed during ageing at the transcriptomic level. We were finally able to highlight the major role

of the key transcription factors, Egr1, Junb and Gata2, in the changes of the differentiation dynamics of HSCs during ageing.

Key words: Haematopoietic stem cell, single cell RNA seq, gene regulatory network inference, logical modelling, ageing

# Remerciements

Tout d'abord, j'aimerais remercier mes superviseures Élisabeth Remy et Estelle Duprez pour m'avoir offert l'opportunité de réaliser ce travail et pour m'avoir guidé et encouragé tout au long de ces 4 années et même avant lors de la préparation du concours de l'école doctorale. Je leur suis très reconnaissant d'avoir toujours été disponibles et à l'écoute dans les moments difficiles, notamment avec les perturbations liées au contexte sanitaire.

Je remercie également les membres du jury pour avoir accepté de consacrer du temps à évaluer mon travail et venir à ma soutenance. Je remercie tout particulièrement Loïc Paulevé et Thierry Jaffredo pour leur rapport sur mon manuscrit et leurs précieux commentaires.

Je souhaite remercier tous les gens avec qui j'ai eu l'occasion de travailler et d'échanger durant mon doctorat à l'I2M et au CRCM. Merci à Anaïs Baudot, Laurent Tichit et Brigitte Mossé, pour leurs conseils et encouragements. Je souhaiterais dire un grand merci à Nadine Platet et Mathilde Poplineau pour leur soutien et le travail que nous avons accompli ensemble, tout particulièrement pour avoir réalisé la partie expérimentale à la base de mes recherches. Je remercie tous l'open-space bioinfo du CRCM, en particulier mes anciens coloc Quentin Da Costa et Benoît Gouthorbe, ainsi qu'Adrien Mazuel et Julien Vernerey avec qui nous sommes devenus experts en débogage de pipelines. Merci à Alberto Valdeolivas, Firas Hammami, Elva María Novoa Del Toro, Sylvain Garciaz et Bochra Zidi, les anciens doctorants que j'ai côtoyés qui m'ont montré la voie, et, Saran Pankaew, Clara Tellez-Quijorna, Laureen Haboub Taha et Nadine Ben Boina à qui je souhaite toute la réussite pour leur thèse. Merci également à tous les autres membres passés et actuels des 2 équipes, Sara Karaki, Marie Miniville, Lia Dasi, Sara Nael, Stephen Chapman, Francesca Zaccagnino, Maxime Lucas, Ozan Özışık et tous les autres pour les moments qu'on a passés ensemble. Travailler dans deux laboratoires différents a vraiment été une expérience enrichissante et source de très nombreuses rencontres. Merci aussi aux matheux Bastien Pacifico, Alejandro J. Giangreco Maidana et Elena Berardini et les autres pour les parties de foot endiablées à Luminy qui m'ont coûté un genou (à priori maintenant réparé).

Un grand merci également aux membres de mon comité de thèse Stéphane Mancini, Lionel Spinelli et Claudine Chaouiya qui m'ont apporté leurs points de vue complémentaires très utiles sur mon travail. Je remercie aussi vivement Manuela Carenzi pour m'avoir donné

l'opportunité d'enseigner et poursuivre mes recherches une 4<sup>ème</sup> année.

Merci beaucoup à Denis Puthier pour m'avoir donné goût à la bioinformatique quand j'étais à Polytech et pour tous les nombreux conseils qu'il a pu m'apporter par la suite en stage et quand je travaillais au TAGC. Un grand merci aussi à Béatrice Loriod pour m'avoir fait confiance à la fin de Polytech en me recrutant au TGML et pour m'avoir ainsi permis de faire mes premières armes en bioinformatique.

Enfin, j'aimerais remercier mes parents et mon frère, toute ma famille (la vraie et celle de Polytech) et mes amis pour m'avoir soutenu pendant ces 4 années. Merci à Albane, Cindy, Adrien et Luca pour ces moments de retrouvailles et d'escapades loin des lignes de codes. Merci également à Macaron pour son soutien dans les péripéties administratives de dernière minute ainsi que sa bonne humeur et son rire contagieux.

Ces remerciements pourraient encore se poursuivre sur des pages. Pour finir, j'aurais une dernière pensée pour Rodrigue, Robin et mon grand-père qui m'ont vu commencer ma thèse et qui malheureusement nous ont quitté avant que je ne la termine, je suis sûr qu'ils auraient été très heureux de me voir conclure cette aventure.

# Table des matières

<b>Résumé</b>	<b>1</b>
<b>Abstract</b>	<b>3</b>
<b>Remerciements</b>	<b>5</b>
<b>Table des matières</b>	<b>7</b>
<b>Table des figures</b>	<b>9</b>
<b>Liste des tableaux</b>	<b>10</b>
<b>Liste des acronymes</b>	<b>11</b>
<b>Glossaire</b>	<b>14</b>
<b>1 Introduction</b>	<b>15</b>
1.1 Propriétés et vieillissement de la cellule souche hématopoïétique (CSH) . . . . .	17
1.1.1 La CSH : une cellule quiescente et multipotente . . . . .	17
1.1.2 Régulation du cycle cellulaire de la CSH . . . . .	22
1.1.3 Le vieillissement de la CSH . . . . .	25
1.2 Évolution des technologies cellules uniques pour appréhender l'hétérogénéité cellulaire et fonctionnelle des CSH . . . . .	32
1.2.1 Technologies de séquençage d'ARN cellules uniques (scRNA-seq) . . . . .	32
1.2.2 Méthodes d'analyse du scRNA-seq . . . . .	33
1.2.3 Les biais d'analyses en scRNA-seq et leurs résolutions possibles . . . . .	38
1.2.4 Une hétérogénéité et une différenciation complexe des CSH mis en évidence par le scRNA-seq . . . . .	41
1.2.5 Apport des techniques d'analyses cellules uniques complémentaires au scRNA-seq . . . . .	46
1.2.6 Applications des technologies cellules uniques à l'étude de l'hémato-poïèse en conditions perturbées . . . . .	49
1.3 Biologie des systèmes et réseaux de régulation moléculaire . . . . .	50
1.3.1 Modélisation des réseaux de régulation moléculaire . . . . .	51

1.3.2	Inférence de réseaux de régulation moléculaire à partir de données cellules uniques . . . . .	64
1.3.3	Inférence de modèles logiques à partir de données cellules uniques . . . . .	69
<b>2</b>	<b>Objectifs de la thèse</b>	<b>73</b>
2.1	Caractérisation des sous populations des HSPC et de leur vieillissement . . . . .	73
2.2	Identification des acteurs moléculaires des changements phénotypiques des CSH âgées . . . . .	74
2.3	Construction du graphe d'influence des acteurs de l'hématopoïèse précoce altérés par le vieillissement . . . . .	74
2.4	Construction d'un modèle logique de l'hématopoïèse précoce permettant d'étudier le vieillissement . . . . .	75
2.5	Développement de pipelines d'analyse . . . . .	75
<b>3</b>	<b>Résultat : Analyse scRNA-seq haut débit de l'hématopoïèse précoce</b>	<b>76</b>
3.1	Avant propos . . . . .	76
3.2	Pipeline d'analyse . . . . .	77
3.3	Article . . . . .	81
3.4	Complément biais et corrections . . . . .	113
3.5	Complément sur les pseudo-traj ectoires de différenciation . . . . .	114
3.6	Discussion . . . . .	117
<b>4</b>	<b>Résultat : Inférence d'un modèle logique de l'hématopoïèse précoce altérée par le vieillissement</b>	<b>119</b>
4.1	Avant propos . . . . .	119
4.1.1	Introduction . . . . .	119
4.1.2	Inférence de BN à partir de contraintes dynamiques avec Bonesis . . . . .	120
4.2	Pipeline d'analyse . . . . .	122
4.3	Résultats . . . . .	125
4.4	Discussion . . . . .	157
<b>5</b>	<b>Conclusion et perspectives</b>	<b>160</b>
5.1	Cartographie du compartiment HSPC en condition jeune et âgée . . . . .	160
5.2	Modélisation logique en sémantique MP de l'hématopoïèse précoce . . . . .	161
5.3	Aspects méthodologiques de la construction de BN à partir de données cellules uniques . . . . .	163
	<b>Bibliographie</b>	<b>165</b>

# Table des figures

1.1 L'arbre hématopoïétique . . . . .	19
1.2 Régulation de la transcription chez les mammifères . . . . .	21
1.3 Acteurs majeurs de la régulation de la sortie de quiescence de la CSH. . . . .	24
1.4 Le vieillissement de la CSH . . . . .	27
1.5 Etude scRNA-seq de l'hématopoïèse . . . . .	37
1.6 Evolution de la vision de l'hématopoïèse avec le scRNA-seq . . . . .	46
1.7 Un BN possible d'un graphe d'influence et ses trajectoires synchrones, asynchrones et généralisées . . . . .	54
1.8 Sémantique MP des BN . . . . .	59
1.9 Perturbations des BN . . . . .	61
1.10 Inférence de BN à partir de données scRNA-seq . . . . .	70
3.1 Pipeline d'analyse de données scRNA-seq . . . . .	78
3.2 Corrections des effets de lots et du bruit du cycle cellulaire . . . . .	114
3.3 Pseudo-trajectoire obtenue avec STREAM . . . . .	116
4.1 Pipeline d'inférence d'un BN à partir de données scRNA-seq . . . . .	123

# Liste des tableaux

1.1	Classification des HSPC . . . . .	18
1.2	Sélections d'études scRNA-seq sur les CSH et leur devenir . . . . .	43
1.3	Sélections d'études scRNA-seq sur les CSH et leur devenir (2) . . . . .	44
4.1	Nombre de fonctions booléennes monotones possibles pour un composant selon son nombre de régulateurs . . . . .	122

# Liste des acronymes

## **ACP**

Analyse en Composantes Principales. [35, 38](#)

## **AML**

Leucémies aiguës myéloïdes –ou *Acute Myeloïd Leukemias*–. [49](#)

## **ARNm**

ARN messagers. [20](#)

## **ASP**

*Answer Set Programming*. [69, 122](#)

## **ATAC-seq**

analyse de la chromatine accessible à la transposase par séquençage –ou *Assay for Transposase-Accessible Chromatin with sequencing*–. [48](#)

## **BN**

réseau booléan –ou *Boolean Network*–. [54, 69, 75, 157](#)

## **CDK**

*Cyclin-Dependant kinases*. [22, 119](#)

## **CITE-seq**

indexage cellulaire des transcriptomes et des épitopes par séquençage –ou *Cellular Indexing of Transcriptomes and Epitopes by Sequencing*–. [40, 47, 48](#)

## **CKI**

*Cycline-Kinase Inhibitors*. [22–24, 30, 31, 119](#)

## **CLP**

progéniteurs communs lymphocytaires –ou *Common Lymphocytic Progenitors*–. [18, 19, 27, 63](#)

## **CMP**

progéniteurs communs myéloïdes –ou *Common Myeloid Progenitors*–. [18, 19, 42, 63](#)

## **CSH**

Cellule Souche Hématopoïétique. [17, 19, 63](#)

**CSM**

Cellule Stromale Mésenchimateuse. [63](#)

**DEG**

gènes différentiellement exprimés –ou *Differentially Expressed Genes*–. [36](#), [37](#)

**FACS**

tri cellulaire induit par fluorescence –ou *Fluorescence Activated Cell Sorting*–. [17](#), [46](#), [161](#)

**GMP**

progéniteurs granulo-monocytaires –ou *Granulo-Monocytic Progenitors*–. [18](#), [19](#), [27](#), [42](#), [63](#), [120](#)

**HSPC**

cellules souches et progéniteurs hématopoïétiques –ou *Hematopoietic Stem and Progenitor Cells*–. [18](#), [32](#), [39](#), [40](#), [42](#), [49](#), [63](#), [73](#), [76](#), [117](#), [160](#)

**LMPP**

progéniteurs amorcés lymphoïdes multipotents –ou *Lymphoid-primed MultiPotent Progenitor*–. [18](#), [19](#), [64](#)

**LTHSC**

cellules souches hématopoïétiques à long terme –ou *Long Term Haematopoietic Stem Cells*–. [18](#), [24](#), [26](#), [63](#), [115](#), [117](#)

**MAPK**

protéines kinases activées par les mitogènes –ou *Mitogen-activated protein kinases*–. [22](#), [24](#), [25](#), [29](#), [30](#), [60](#), [61](#)

**MEP**

progéniteurs érythroïdes et mégacaryocytaires –ou *Megakaryocytic and Erythroid Progenitors*–. [18](#), [19](#), [22](#), [63](#), [64](#)

**MEX**

échange de marché –ou *Market Exchange*–. [34](#), [37](#)

**MLLE**

*Modified Locally Linear Embedding*. [115](#)

**MO**

Moelle Osseuse. [17–19](#), [46](#), [73](#), [119](#), [123](#), [161](#)

**MP**

sémantique la plus permissive des BN –ou *Most Permissive (MP) semantic of BN*–. [58](#), [71](#), [120](#), [161](#)

**MPP**

progéniteurs multipotents –ou *multipotent progenitors*–. 18, 19, 21, 22, 45, 46, 63

**mTOR**

cible de la rapamycine chez les mammifères –ou *mammalian target of rapamycin*–. 29

**NGS**

séquençage nouvelle génération –ou *Next Generation Sequencing*–. 32, 50, 51, 62

**ODE**

équations différentielles ordinaires –ou *Ordinary Differential Equations*–. 65

**PCR**

réactions de polymérisation en chaîne –ou *Polymérase Chain Reactions*–. 32, 34

**PI3K**

phosphoinositide 3-kinase. 23, 24

**ROS**

espèces réactives de l'oxygène –ou *Reactive Oxygen Species*–. 28–30, 64

**scATAC-seq**

ATAC-seq en cellules uniques –ou *single-cell ATAC-seq*–. 48

**sc-multi-omics**

multi-omiques en cellules uniques –ou *single-cell multi-omics*–. 48

**scRNA-seq**

séquençage d'ARN à l'échelle de la cellule –ou *single cell RNA sequencing*–. 32, 33, 35, 36

**STHSC**

cellules souche hématopoïétiques à court terme –ou *Short Term Haematopoietic Stem Cells*–. 18, 63

**TF**

facteurs de transcription –ou *Transcription Factors*–. 20, 21, 23, 68, 158

**tSNE**

*t-distributed stochastic neighbor embedding*. 35

**TSS**

site d'initiation de la transcription –ou *transcriptionnal Start Site*–. 20, 21, 68

**UMAP**

*Uniform Manifold Approximation and Projection*. 35, 38

**v.a.**

variable aléatoire. 65, 66

# Glossaire

## **ChIP-seq**

Technique d'analyse des interactions ADN-protéines combinant l'immunoprecipitation de la chromatine (*Chromatine Immuno-Precipitation*) et le séquençage haut débit. [48](#)

## **CRISPR**

Séquence d'ADN présentant de courtes répétitions en palindrome, regroupées et régulièrement espacées (*Clustered Regularly Interspaced Short Palindromic Repeats*). Ces séquences sont utilisées par l'enzyme Cas9 qui peut couper l'ADN du génome lié à l'ARN complémentaire de ces séquences. Ce "ciseau moléculaire" est utilisé en génie génétique pour introduire des mutations à des endroits précis du génome. [49](#)

## **drop-outs**

Transcrit non détecté, bien que présent dans la cellule, à cause de la faible couverture de séquençage en scRNA-seq.. [34](#), [39](#), [71](#)

## **inflammaging**

Inflammation chronique, stérile et de faible intensité résultat de la stimulation physiologique chronique du système immunitaire inné sur le long terme, qui peut devenir nuisible au cours du vieillissement. [26](#)

## **priming**

Amorçage de la cellule souche au niveau transcriptionnel. La cellule commence à exprimer certains gènes propres à une lignée sans qu'il n'y ait encore de conséquence au niveau du phénotype cellulaire. [21](#), [32](#), [68](#)

# 1 Introduction

## Sommaire

1.1	Propriétés et vieillissement de la cellule souche hématopoïétique (CSH) . . . . .	17
1.1.1	La CSH : une cellule quiescente et multipotente . . . . .	17
1.1.1.1	Découverte et définition de la CSH . . . . .	17
1.1.1.2	Une diversité de devenirs possibles pour la CSH . . . . .	17
1.1.1.3	Phénotypes de la CSH . . . . .	19
1.1.1.4	Mécanismes moléculaires gouvernant le devenir de la CSH . . . . .	20
1.1.2	Régulation du cycle cellulaire de la CSH . . . . .	22
1.1.2.1	Activation de la CSH quiescente . . . . .	22
1.1.2.2	Cycle cellulaire, autorenouvellement et départ en différenciation	24
1.1.3	Le vieillissement de la CSH . . . . .	25
1.1.3.1	La CSH âgée à l'origine de l'immunosénescence . . . . .	25
1.1.3.2	Facteurs intrinsèques du vieillissement de la CSH . . . . .	28
1.1.3.3	Facteurs extrinsèques du vieillissement de la CSH . . . . .	30
1.1.3.4	Signature épigénétique et transcriptomique de la CSH agée . . . . .	30
1.2	Évolution des technologies cellules uniques pour appréhender l'hétérogénéité cellulaire et fonctionnelle des CSH . . . . .	32
1.2.1	Technologies de séquençage d'ARN cellules uniques (scRNA-seq) . . . . .	32
1.2.2	Méthodes d'analyse du scRNA-seq . . . . .	33
1.2.2.1	Traitements primaire et contrôle qualité . . . . .	33
1.2.2.2	S'affranchir du bruit et de la dimensionnalité . . . . .	34
1.2.2.3	Classification des cellules . . . . .	35
1.2.2.4	Recherche des marqueurs de populations . . . . .	36
1.2.2.5	Construction de pseudo-traj ectoires de différenciation . . . . .	36
1.2.3	Les biais d'analyses en scRNA-seq et leurs résolutions possibles . . . . .	38
1.2.3.1	Biais des représentations dans des espaces réduits . . . . .	38
1.2.3.2	Biais des mesures d'expression à cause de la faible couverture de séquençage . . . . .	39
1.2.3.3	Biais dus à des effets biologiques . . . . .	39
1.2.3.4	Biais dus aux effets de lots . . . . .	40
1.2.3.5	La difficulté de l'inférence de trajectoire . . . . .	41

1.2.4	Une hétérogénéité et une différenciation complexe des CSH mis en évidence par le scRNA-seq . . . . .	41
1.2.5	Apport des techniques d'analyses cellules uniques complémentaires au scRNA-seq . . . . .	46
1.2.5.1	Traçage de lignages . . . . .	46
1.2.5.2	Analyses multi-omiques cellules uniques . . . . .	47
1.2.5.3	Criblage génétique par CRISPR-seq . . . . .	49
1.2.6	Applications des technologies cellules uniques à l'étude de l'hématopoïèse en conditions perturbées . . . . .	49
1.2.6.1	Étude du vieillissement de la CSH par scRNA-seq . . . . .	49
1.2.6.2	Applications des technologies cellules uniques à l'étude des myélopathies . . . . .	49
1.3	Biologie des systèmes et réseaux de régulation moléculaire . . . . .	50
1.3.0.1	Représentation des régulations moléculaires par un graphe d'influence . . . . .	50
1.3.0.2	Ressources pour la construction des graphes d'influence . . . . .	51
1.3.1	Modélisation des réseaux de régulation moléculaire . . . . .	51
1.3.1.1	Intérêt de la modélisation . . . . .	51
1.3.1.2	La modélisation booléenne . . . . .	52
1.3.1.3	Sémantiques des réseaux booléens . . . . .	55
1.3.1.4	Lien entre le graphe d'influence et la dynamique du BN . . . . .	57
1.3.1.5	Réseaux Booléens les plus permissifs . . . . .	58
1.3.1.6	Analyses <i>in silico</i> et perturbations des BN de systèmes biologiques	60
1.3.1.7	Applications au système hématopoïétique . . . . .	62
1.3.2	Inférence de réseaux de régulation moléculaire à partir de données cellules uniques . . . . .	64
1.3.2.1	Diversité des méthodes mathématiques . . . . .	64
1.3.2.2	Inférence de graphe d'influence à partir de matrices d'expressions de données cellules uniques . . . . .	65
1.3.2.3	Apport des données génomiques et épigénétiques . . . . .	68
1.3.3	Inférence de modèles logiques à partir de données cellules uniques . . . . .	69
1.3.3.1	Présentation du problème . . . . .	69
1.3.3.2	Revues des méthodes existantes . . . . .	70

## 1.1 Propriétés et vieillissement de la cellule souche hématopoïétique (CSH)

### 1.1.1 La CSH : une cellule quiescente et multipotente

#### 1.1.1.1 Découverte et définition de la CSH

Le terme cellule souche a été employé la première fois à la fin du 19<sup>ème</sup> siècle par Ernst Heackel pour parler de l'être vivant unicellulaire duquel descendrait tous les organismes multicellulaires selon la théorie Darwinienne. Dans ses écrits, Ernst Heackel compare cet être originel à l'œuf fertilisé à la base du développement embryonnaire et de la différenciation cellulaire chez ces organismes (HAECKEL, 1877). Ce concept de cellule souche siégeant à la base d'un arbre généalogique ramifié a ensuite été repris en hématologie la première fois en 1896 par Pappenheim pour décrire la cellule multipotente au départ de l'hématopoïèse, processus de différenciation cellulaire qui permet la production des cellules matures du sang (PAPPENHEIM, 1896). Le débat sur l'existence de cette [Cellule Souche Hématopoïétique \(CSH\)](#) dura jusqu'aux années 1960, lorsqu'elle fut pleinement démontrée par le suivi d'anomalies cytogénétiques lors de greffes de [Moelle Osseuse \(MO\)](#) permettant la reconstitution du système hématopoïétique de souris irradiées (BECKER et al., 1963). La cellule souche est définie sur deux propriétés essentielles : l'autorenouvellement et le départ en différenciation ; deux propriétés qui permettent de maintenir la population de cellules souches tout en assurant la formation des cellules matures fonctionnelles, à la durée de vie limitée (RAMALHO-SANTOS et WILLENBRING, 2007). Dans le cas de l'hématopoïèse, l'homéostasie du tissu sanguin dépend donc de la CSH multipotente capable après différenciation de maintenir les différentes lignées du sang : lymphoïde, myeloïde, érythroïde et mégacaryocytaire. Au fur et à mesure de sa différenciation, la CSH perd en capacité d'autorenouvellement en passant par des progéniteurs intermédiaires au potentiel de différenciation de plus en plus restreint vers un lignage spécifique.

#### 1.1.1.2 Une diversité de devenirs possibles pour la CSH

Grâce au développement du [tri cellulaire induit par fluorescence –ou Fluorescence Activated Cell Sorting– \(FACS\)](#), la purification des CSH et des progéniteurs a été grandement facilitée, permettant leur greffe et donc l'étude de leur potentiel de différenciation (SPANGRUDE et al., 1988). Une vision de l'hématopoïèse en arbre hiérarchique a ainsi pu émerger dans les années 2000, suite à la caractérisation de différents stades de progéniteurs multi puis oligo et finalement unipotents sur la base de leur capacité à reconstituer tout ou une partie du système hématopoïétique (Figure 1.1 ; ADOLFSSON et al., 2005 ; AKASHI et al., 2000 ; DOULATOV et al., 2010).

Cet arbre a ensuite été régulièrement affiné notamment en ce qui concerne la classification

## 1 Introduction – 1.1 Propriétés et vieillissement de la cellule souche hématopoïétique (CSH)

de la population des CSH. Ainsi, ont été définis : les **cellules souches hématopoïétiques à long terme** –ou *Long Term Haematopoietic Stem Cells*– (LTHSC) comme les CSH capables de reconstituer l'hématopoïèse à long terme suite à leur greffe dans la **MO** de souris irradiées ; les **cellules souche hématopoïétiques à court terme** –ou *Short Term Haematopoietic Stem Cells*– (STHSC), CSH capables de se différencier dans n'importe quelle lignées mais ne reconstituant pas la **MO** sur un temps long. Les CSH se différencient en différents **progéniteurs multipotents** –ou *multipotent progenitors*– (MPP) dont le potentiel de différenciation est biaisé vers certains lignages (PIETRAS et al., 2015). C'est notamment le cas des **progéniteurs amorcés lymphoïdes multipotents** –ou *Lymphoid-primed MultiPotent Progenitor*– (LMPP). La population cellulaire de ce premier étage de l'hématopoïèse, ou hématopoïèse précoce, est couramment englobée sous le terme des **cellules souches et progéniteurs hématopoïétiques** –ou *Hematopoietic Stem and Progenitor Cells*– (HSPC) (Tableau 1.1).

Fraction	Phenotypage (surface cellulaire)	fonctionnalité
LTHSC	LSK; FLT3 <sup>-</sup> ; CD150 <sup>+</sup> ; CD48 <sup>-</sup>	repeuplement à long terme après greffe
STHSC/MPP1	LSK; c-Kit <sup>+</sup> ; FLT3 <sup>-</sup> ; CD150 <sup>-</sup> ; CD48 <sup>-</sup>	repeuplement à court terme après greffe
MPP2	LSK; c-Kit <sup>+</sup> ; FLT3 <sup>-</sup> ; CD150 <sup>+</sup> ; CD48 <sup>+</sup>	amorcés MkE
MPP3	LSK; c-Kit <sup>+</sup> ; FLT3 <sup>-</sup> ; CD150 <sup>-</sup> ; CD48 <sup>+</sup>	amorcés Mye
MPP4	LSK; c-Kit <sup>+</sup> ; FLT3 <sup>+</sup> ; CD150 <sup>-</sup>	amorcés Lymph
LMPP	LSK; c-Kit <sup>+</sup> ; FLT3 <sup>hi</sup> ; CD150 <sup>-</sup> ; CD34 <sup>+</sup>	amorcés Lymph

Tableau 1.1 – Classification des HSPC

Phénotypage et fonctionnalité des différents types cellulaires de HSPC. + (resp -) : cellules positives pour le marqueurs de surface (resp. negatives), hi : cellules fortement positives pour le marqueur. LSK : Lin<sup>-</sup>; Sca-1<sup>+</sup>; c-Kit<sup>+</sup>. MkE : Mégacaryocytaire et Érythroïde, Mye : Myéloïde, Lymph : Lymphoïde (d'après ADOLFSSON et al., 2005; PIETRAS et al., 2015; A. WILSON et al., 2008) .

Dans l'étage inférieur de l'arbre, différents types de progéniteurs oligopotents ont pu être isolés. Les **progéniteurs communs myéloïdes** –ou *Common Myeloid Progenitors*– (CMP) donnent les **progéniteurs érythroïdes et mégacaryocytaires** –ou *Megakaryocytic and Erythroid Progenitors*– (MEP) et les **progéniteurs granulo-monocytaires** –ou *Granulo-Monocytic Progenitors*– (GMP). Les **MEP** se différencient ensuite en mégacaryocytes et en érythrocytes, les **GMP** donnent les mastocytes, les monocytes, les granulocytes et les cellules dendritiques. Quant aux **progéniteurs communs lymphocytaires** –ou *Common Lymphocytic Progenitors*– (CLP) ils peuvent se différencier en lymphocytes N, T, en cellules NK ainsi qu'en cellules dendritiques (LAURENTI et GÖTTGENS, 2018; Figure 1.1).

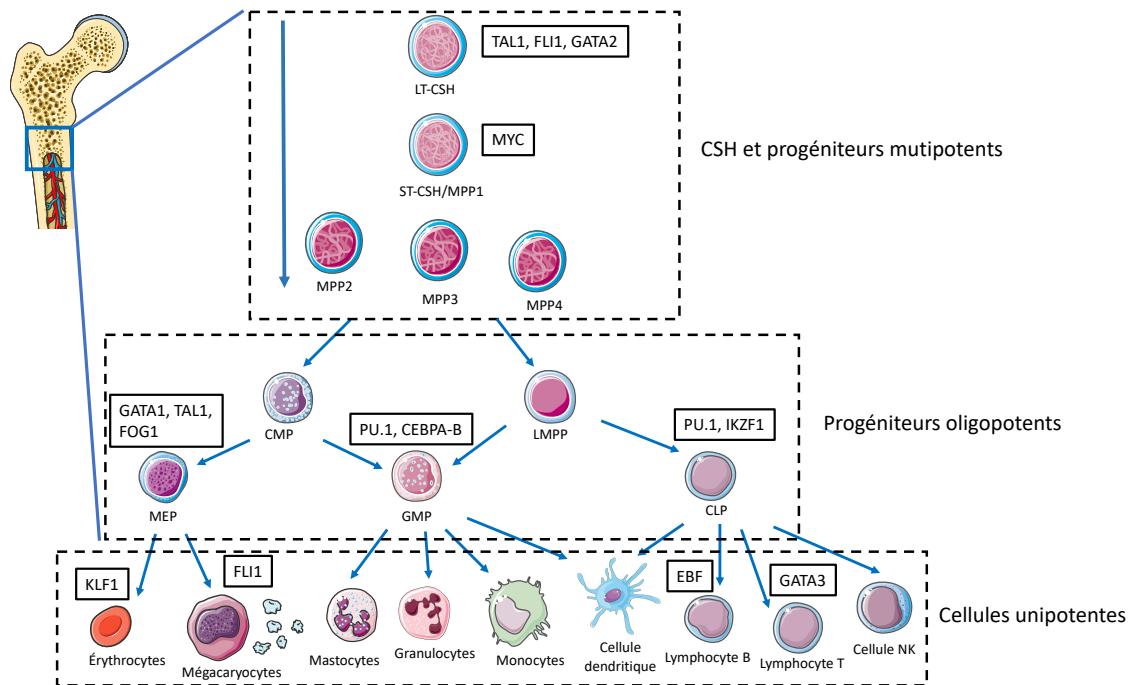


FIGURE 1.1 – L’arbre hématopoïétique

Représentation schématique d’une des vues classiques de l’hématopoïèse comme un arbre de différenciation en plusieurs étapes hiérarchisées. À la racine de l’arbre les **CSH** et les **MPP**, qui se différencient ensuite en progéniteurs **CMP**, **LMPP** puis **MEP**, **GMP** et **CLP**. Finalement, aux niveaux des feuilles de l’arbre, se trouvent les cellules différencierées et fonctionnelles. Chacun des choix de la cellule aux embranchements est associé à la mise en place d’un programme transcriptionnel précis dont certains facteurs de transcription majeurs sont indiqués en encadré.

### 1.1.1.3 Phénotypes de la CSH

La CSH est majoritairement en phase de dormance, une phase caractérisée par une faible activité métabolique ainsi qu’un état quiescent qui se définit par une absence totale de division cellulaire (phase G0 du cycle cellulaire). Occasionnellement, la CSH peut s’autorenouveler ou partir en différenciation. Ces trois états ont des caractéristiques phénotypiques particulières qui se manifestent notamment au niveau du métabolisme et des marqueurs de surface et sont influencés par le microenvironnement de la **MO**.

Les études génomiques de **CSH** murines couplées à des études génétiques (inactivation de gènes notamment) ont en premier lieu révélé des caractéristiques cellulaires en lien avec le maintien de leur intégrité. En effet, contrairement aux **MPP** directement en aval dans la différenciation, la CSH se maintient dans un état quiescent et assure ses besoins énergétiques par la glycolyse. Elle n’a ainsi qu’une faible activité mitochondriale (SIGNER et al., 2014). La CSH se démarque aussi par une forte activation de l’autophagie qui lui permet de dégrader efficacement ses organites endommagés par le stress métabolique résultant de sa privation de facteurs de croissance (WARR et al., 2013).

Occasionnellement, pour assurer le maintien de sa population, la CSH se divise pour s'autorenouveler. Chez la souris, il a été estimé qu'une CSH se divise de cette manière jusqu'à 4 fois dans sa vie avant de perdre sa capacité d'autorenouvellement (BERNITZ et al., 2016). La CSH peut aussi s'activer puis proliférer transitoirement en réponse au stress lors d'une demande urgente en cellules sanguines matures, suite à une blessure par exemple, puis retourner à un état quiescent (A. WILSON et al., 2008). Ces caractéristiques intrinsèques de la CSH sont soutenues par des signaux extérieurs via un dialogue complexe avec les cellules de la niche hématopoïétique, le microenvironnement hautement spécialisé dans lequel réside la population de CSH (BARYAWNO et al., 2017). On distingue la niche endostéale physiquement proche de l'os où résiderait la majorité des CSH à l'état quiescent, de la niche perivasculaire proche des vaisseaux sanguins (MENDELSON et FRENETTE, 2014).

#### 1.1.1.4 Mécanismes moléculaires gouvernant le devenir de la CSH

Le devenir de la CSH est contrôlé par des facteurs intrinsèques qui interviennent en premier lieu au niveau de la transcription des gènes au sein du noyau. Ce phénomène régulé par des complexes transcriptionnels comprenant entre autres l'ARN polymérase, des facteurs chromatiniens et des [facteurs de transcription –ou \*Transcription Factors\*– \(TF\)](#), permet la synthèse des [ARN messagers \(ARNm\)](#). Ces ARNm sont ensuite exportés dans le cytoplasme où ils sont traduits par les ribosomes en protéines qui assurent les fonctions biologiques propres à la cellule.

La spécificité lors de l'initiation puis le maintien de la transcription par la machinerie transcriptionnelle est médiée par l'action des [TF](#). Il s'agit de protéines qui ont la capacité de se lier à des séquences spécifiques d'ADN, d'une dizaine de paires de bases, situées sur des régions du génome appelées éléments régulateurs cis (du fait de leur présence sur la même molécule d'ADN que le gène régulé). Une fois liés à ces régions, les [TF](#) recrutent des cofacteurs qui activent ou inhibent l'activité de la transcription. Chez les mammifères, on distingue les promoteurs qui sont les éléments régulateurs situés jusqu'à 500 paires de bases du [site d'initiation de la transcription –ou \*transcriptional Start Site\*– \(TSS\)](#), des enhancers et silenciers qui peuvent être localisés jusqu'à plusieurs dizaines de kilobases du [TSS](#) et qui interagissent avec les promoteurs par la formation de boucles d'ADN (Figure 1.2). L'accessibilité des [TF](#) aux éléments régulateurs est, quant à elle, régulée par des facteurs épigénétiques qui modulent le niveau de compaction de l'ADN. Ces facteurs épigénétiques ajoutent, suppriment ou lisent et interprètent les modifications post-traductionnelles (methylation/acetylation) des histones, protéines sur lesquelles s'enroulent l'ADN compactée. Le niveau de méthylation de l'ADN régule lui aussi la transcription de l'ADN avec la méthylation des cytosines au niveau des éléments régulateurs qui a un effet répressif sur la transcription par les ADN-methyltransferases (Figure 1.2).

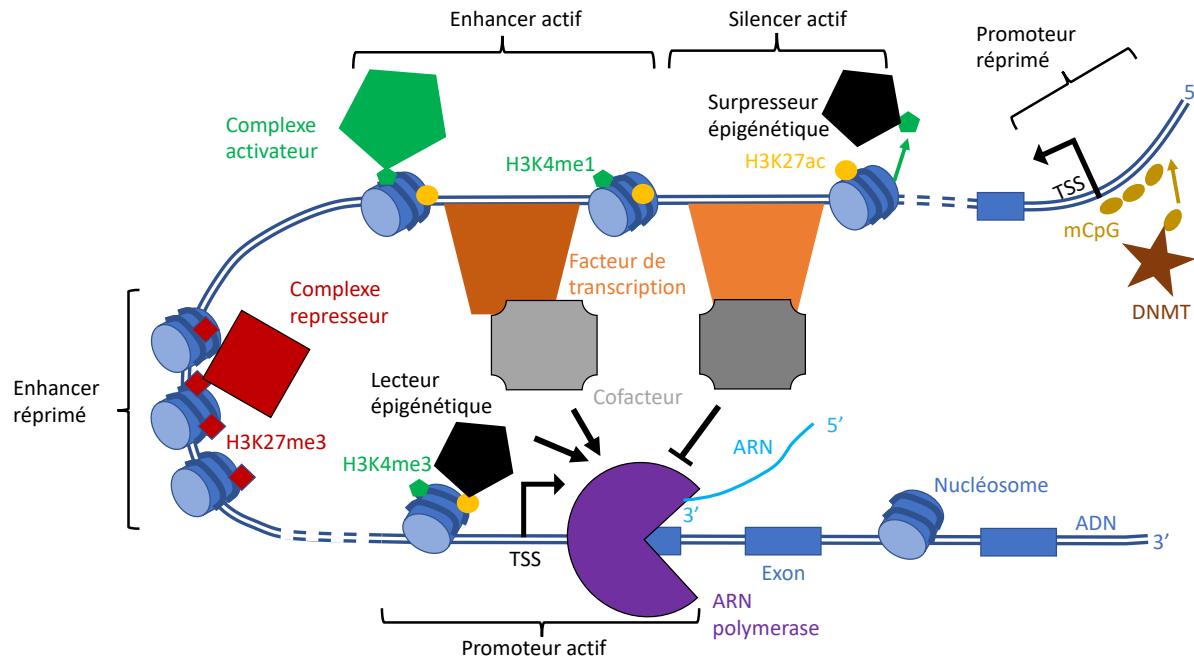


FIGURE 1.2 – Régulation de la transcription chez les mammifères

La transcription de l'ADN en ARN se fait dans le sens 5' vers 3' par l'ARN polymerase à partir du TSS. L'activité de la polymerase est accrue (resp. diminuée) par le recrutement de cofacteurs par des TF qui se lient aux éléments régulateurs cis enhancer (resp. silencer). L'accès à ces éléments ainsi qu'au promoteur est modulé par des facteurs épigénétiques qui contrôlent la compaction de l'ADN via la lecture, l'ajout et la suppression de marques épigénétiques activatrices (H3K4me1, H3K27ac pour les enhancers et silencers, H3K4me3 pour les promoteurs) ou répressives (H3K27me3) au niveau des histones qui forment les nucléosomes. Des ADN (dé)methyltransferases peuvent également moduler la méthylation de l'ADN aux éléments régulateurs pour (dé)réprimer l'expression.

Les différents facteurs (transcriptionnels et épigénétiques) agissent au sein de réseaux complexes d'interactions pour réguler l'expression de différents programmes de différenciation ou au contraire maintenir l'état souche comme le souligne une étude multi-omiques des CSH et de leurs descendants directes les MPP (CABEZAS-WALLSCHEID et al., 2014). Les TF dictent donc l'activation ou la répression d'un programme transcriptionnel qui va permettre de réguler le devenir de la CSH. Au sein de la CSH, les gènes codant pour les protéines de la différenciation vers les différentes lignées hématopoïétiques sont réprimés alors que ceux nécessaires aux fonctions souches sont actifs. Lors du processus de différenciation de la CSH la transcription des gènes codant pour les protéines d'un programme de différenciation particulier est activée, on parle alors d'amorçage ou *priming* de la CSH (LAURENTI et GÖTTGENS, 2018).

Dans l'hématopoïèse, plusieurs facteurs clés dont l'activité instruit le choix de la CSH vers différents lignages ont été identifiés. C'est le cas de GATA1 et TAL1 pour la lignée érythroïde et

mégacaryocytaire, PU.1 (codé par le gène *Spi1*) pour la lignée myéloïde et lymphoïde, CEBPA pour la lignée myéloïde (LAIOSA et al., 2006). De la même façon, d'autres facteurs spécifient la différenciation des progéniteurs déjà engagés vers un type cellulaire précis. KLF1 par exemple au niveau des MEP marque le choix vers les érythrocytes tandis que que FLI1 détermine la différenciation en mégacaryocytes (STARCK et al., 2003). Ces facteurs une fois transcrits et actifs se stabilisent souvent par des boucles d'auto-activation, ils interviennent aussi dans la répression de la transcription des programmes vers les autres lignages. Certains sont déjà actifs au niveau de la CSH comme TAL1, FLI1 et GATA2 et sont considérés comme des facteurs qui maintiennent le potentiel souche, tandis que d'autres comme MYC marquent l'amorçage de la différenciation de la CSH vers les MPP (A. WILSON et al., 2004). Ceci souligne toute la complexité du réseau d'interactions régissant la dynamique de la CSH (BONZANNI et al., 2013; PIMANDA et al., 2007). Un ensemble des TF déterminant dans les choix possibles de la CSH est présenté Figure 1.1.

## 1.1.2 Régulation du cycle cellulaire de la CSH

### 1.1.2.1 Activation de la CSH quiescente

Seulement 5% des CSH adultes présentent une activité de division. L'état très majoritaire de dormance n'est atteint à l'âge adulte qu'après une période de forte activité du cycle cellulaire et de prolifération pendant la vie du fœtus qui a pour but d'acquérir rapidement la quantité de cellules hématopoïétiques nécessaire à la fois pour le transport de l'oxygène et le système immunitaire (BOWIE et al., 2006). Chez l'adulte, des travaux ont montré que l'activation de la division de la CSH a lieu environ toutes les 20 semaines chez la souris et toutes les 40 semaines chez l'homme (CATLIN et al., 2011).

La régulation de la division de la CSH se fait principalement à deux points de contrôle du cycle cellulaire : (i) au niveau du passage de la phase G1, période de croissance cellulaire, à la phase S de réPLICATION de l'ADN, et (ii) au niveau du passage de la phase G0 de quiescence, à la réentrée dans le cycle en phase G1 (PIETRAS et al., 2011). Les acteurs clés de ces points de contrôle dans la cellule sont les complexes de protéines kinases dépendantes de cyclines ou *Cyclin-Dependant kinases* (CDK) et leurs complexes inhibiteurs ou *Cycline-Kinase Inhibitors* (CKI). Ils sont régulés au niveau transcriptionnel, notamment via la synthèse des cyclines et au niveau post-traductionnel par (dé)phosphorylation au sein d'un réseau complexes d'acteurs intra et extra cellulaires (Figure 1.3).

La progression dans la phase G1 est régulée par Le complexe formé par les cyclines D1 à D3 et les CDK 4 et 6 (CDK4/6). La transcription des Cyclines D est activée par différentes voies de signalisation comme la signalisation des protéines kinases activées par les mitogènes –ou *Mitogen-activated protein kinases*– (MAPK) et la voie morphogène de Wnt (REYA et al.,

## 1 Introduction – 1.1 Propriétés et vieillissement de la cellule souche hématopoïétique (CSH)

2003). Lorsqu'un niveau seuil, appelé point R, de protéines cyclines D est atteint le complexe Cyclines-D-CDK4/6 est activé notamment à l'aide de la voie de signalisation de l'enzyme phosphoinositide 3-kinase (PI3K). Il peut alors inhiber les répresseurs transcriptionnels rétinoblastomes (RB), ce qui autorise la transcription des gènes nécessaires à la progression du cycle dans la phase G1 par les facteurs de transcription E2F. Les gènes des cyclines de type E, cibles des E2F, sont transcrits permettant la formation du complexe Cyclines-E-CDK2 qui maintient l'inhibition des RB par phosphorylation. Cette boucle d'auto activation des E2F permet finalement l'entrée en phase S de la cellule (CHOI et ANDERS, 2014).

La sortie de quiescence par l'entrée en G1 peut être régulée par différents facteurs extérieurs à la cellule. Par exemple, Junb est un TF médiateur de la quiescence des CSH via la signalisation TGF-beta (PASSEGUE et al., 2004 ; SANTAGUIDA et al., 2009) tout comme Egr1 dans le contexte de leucémies myéloïdes aiguës suite à la production de TGF-beta par les mégacaryocytes (GONG et al., 2018). L'activation de ces facteurs de transcription conduit à une hausse de la transcription de *Cdkn1a* codant pour P21, membre du complexe CKI CIP/KIP, qui maintient la CSH quiescente. Cette relation entre les mégacaryocytes, la signalisation TGF-beta et la quiescence des CSH a également été caractérisée après une blessure (ZHAO et al., 2014) ce qui illustre la régulation de la population de la CSH à l'échelle de l'organisme.

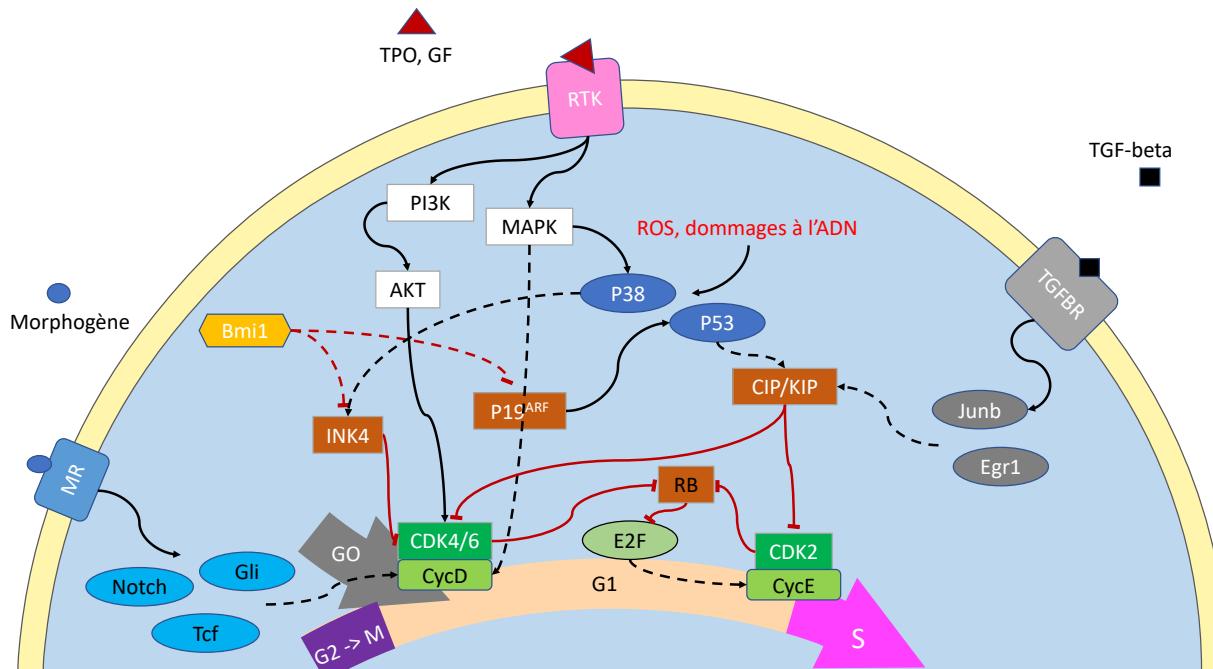


FIGURE 1.3 – Acteurs majeurs de la régulation de la sortie de quiescence de la CSH.

La régulation de la sortie de la quiescence de la CSH est réalisée par les complexes Cyclines-D-CDK4/6 pour l'entrée dans le cycle cellulaire en phase G1 depuis la phase de quiescence G0 et par le complexe Cyclines-E-CDK2 pour le passage de la phase G1 à S. Ces deux complexes sont inhibés par les CKI CIP/KIP et INK4. Ces complexes et leurs inhibiteurs sont régulés par un ensemble de facteurs internes et de voies de signalisations (MAPK, PI3K, P53, TGF-beta) qui permettent à la CSH d'adapter sa réponse au stress interne (ROS, dommages à l'ADN) et aux signaux extérieurs (GF : facteurs de croissance, TPO : thrombopoïétine, et TGF-beta). A l'état LTHSC, le répresseur épigénétique Bmi1 et les voies de signalisation développementales Wnt, Notch et Hedgehog via des signaux morphogènes venant de la niche régulent l'activation du cycle cellulaire liée à l'autorenouvellement. TGFBR : récepteur au TGF-beta, RTK : récepteur Tyrosine Kinase, MR : récepteur de la signalisation morphogène. Les flèches pleines noires (resp. rouge) indiquent les activations (resp. répressions) directes tandis que les flèches en pointillées indiquent les régulations (activations et répressions) transcriptionnelles (adaptée de PIETRAS et al., 2011).

### 1.1.2.2 Cycle cellulaire, autorenouvellement et départ en différenciation

A l'âge adulte, la différenciation de la CSH s'accompagne d'une hausse de la prolifération ce qui a alimenté plusieurs théories sur l'importance du cycle cellulaire pour le départ en différenciation de la CSH (KOWALCZYK et al., 2015 ; ORFORD et SCADDEN, 2008 ; PHAM et al., 2014). Plusieurs études postulent que les divisions symétriques qui donnent deux cellules filles identiques seraient majoritaires chez les LTHSC s'auto-renouvelant, alors que les divisions asymétriques seraient majoritaires pour la différenciation en progéniteurs générant deux cellules filles au matériel biologique différent (PHAM et al., 2014). L'étude de la division symétrique versus asymétrique de la CSH est particulièrement délicate du fait que celle-ci est un événement extrêmement rare. L'héritage asymétrique de plusieurs déterminants cellulaires tels qu'AP2A2 (TING et al., 2012), CDC42 (FLORIAN et al., 2012) a permis d'avancer

dans cette caractérisation mais il est particulièrement difficile d'écartier la possibilité que cette asymétrie observée ne soit pas la cause d'une régulation différente de deux cellules filles identiques au départ. Plus récemment, des travaux utilisant l'imagerie et le suivi quantitatif à long terme en direct de cellules uniques suggèrent que l'héritage asymétrique des lysosomes est impliqué dans l'engagement de la CSH dans la différenciation via la régulation de la clairance mitochondriale et de l'autophagie (LOEFFLER et al., 2019). Une étude récente propose quant à elle que l'héritage asymétrique de mitochondries dysfonctionnelles est utilisé par la CSH comme mémoire du nombre de divisions pour limiter son nombre de divisions d'autorenouvellement (HINGE et al., 2020) . Cette étude suggère donc un rôle primordial des divisions asymétriques dans la régulation de l'autorenouvellement de la CSH mais reste en contradiction avec des études montrant que les divisions de la CSH sont en grande majorité symétriques (BARILE et al., 2020; BERNITZ et al., 2016; TAK et al., 2019).

Le choix de départ en différenciation de la CSH pourrait également être régulé par la longueur du cycle cellulaire et plus particulièrement par le temps passé en phase G1. En effet, la longueur de cette phase est allongée par les facteurs maintenant l'état cellule souche neurale et embryonnaire tandis que les facteurs de différenciation la rallongent (SALOMONI et CALEGARI, 2010; SINGH et DALTON, 2009). De ce constat est née l'hypothèse de *la longueur du cycle cellulaire* selon laquelle la phase G1, du fait d'une plus grande disponibilité des facteurs et ressources que durant les autres phases, serait la période critique durant laquelle sont prises les décisions du destin de la cellule (LANGE et CALEGARI, 2010). Chez la CSH plus spécifiquement, il a été suggéré que l'allongement de la durée de la phase G1 se faisait par l'élévation du seuil R avec la stimulation de la synthèse des cyclines D par les voies MAPK. Ceci favoriserait l'engagement en différenciation de la CSH. Par opposition, l'activation du complexe Cyclines-D-CDK4/6 par les voies de signalisation morphogène Wnt ou Hedgehog abaisserait le seuil R et permettrait alors un passage rapide à travers la phase G1 (ORFORD et SCADDEN, 2008). Ainsi une phase G1 réduite empêcherait que de telles décisions soient prises et favoriserait l'autorenouvellement des cellules souches (Figure 1.3). Partant de cette hypothèse, des travaux ont cherché un lien entre la longueur de la phase G1 et le départ en différenciation de la CSH (KOWALCZYK et al., 2015; LAURENTI et al., 2015). Cependant la causalité de ce lien est difficile à prouver, notamment car l'allongement de la phase G1 pourrait tout aussi bien être une conséquence qu'une cause du choix de la CSH d'amorcer sa différenciation.

### 1.1.3 Le vieillissement de la CSH

#### 1.1.3.1 La CSH âgée à l'origine de l'immunosénescence

Chez les mammifères, le vieillissement se caractérise par une usure des différents tissus à laquelle n'échappe pas le système hématopoïétique. Avec le vieillissement global de la

## 1 Introduction – 1.1 Propriétés et vieillissement de la cellule souche hématopoïétique (CSH)

population que connaît notre époque ceci se traduit par une augmentation de l'incidence des cancers et maladies du sang comme les leucémies myéloïdes, les myélodysplasies ainsi que les anémies (ROSSI et al., 2008). Les altérations de l'immunité avec l'âge regroupées sous le terme d'immunosénescence en sont la cause et se manifestent en premier lieu par une diminution de l'immunité adaptative. En effet, la diversité du répertoire des lymphocytes B mémoires est réduite et le nombre de lymphocytes T naïfs décroît chez les personnes âgées. Leur réponse aux vaccins est ainsi diminuée et leur susceptibilité aux maladies infectieuses accrue (HENRY et al., 2011). La fonctionnalité des cellules de l'immunité innée est elle aussi altérée avec notamment une baisse de l'aptitude à la phagocytose des macrophages et granulocytes neutrophiles (C. Q. WANG et al., 1995). Par ailleurs, l'observation d'un état inflammatoire prolongé appelé *inflammaging* avec des niveaux élevés de certaines cytokines médiatrices de l'immunité innée (interleukines IL6 et IL-1beta, facteur de nécrose tumoral TNF) chez les personnes âgées conduit au développement de maladies dégénératives liées au vieillissement (FRANCESCHI et al., 2000).

De par leur capacité de différenciation et d'autorenouvellement, la population de CSH est souvent vue comme une « fontaine de jouvence » dont l'altération serait responsable du vieillissement de notre système immunitaire. Ainsi, la CSH étant à la source de la production des lignées myéloïde et lymphoïde elle apparaît naturellement comme une origine probable des mécanismes de l'immunosénescence (GEIGER et al., 2013). Au cours du vieillissement, on observe une augmentation des LTHSC (MORRISON et al., 1996) avec cependant des fonctionnalités réduites, comme leur capacité d'autorenouvellement ou leur différenciation biaisée vers le lignage myéloïde (Figure 1.4.A&B ; DYKSTRA et al., 2011).

La plupart des caractéristiques de la CSH et de son vieillissement sont observées dans plusieurs espèces de mammifères, malgré quelques différences. En effet, contrairement à ce qui est observé chez la souris, la baisse des CLP ne s'accompagne pas d'une hausse des GMP chez l'humain (PANG et al., 2011). Pour des raisons expérimentales la souris s'est rapidement imposée comme le modèle d'étude de choix de l'immunosénescence. Il est admis qu'une souris C57BL6 de 2 à 6 mois correspond à un jeune adulte et qu'au-delà de 20 mois les animaux sont considérés vieux (ZHANG et al., 2020). Des expériences de greffes de CSH de souris âgées dans des souris jeunes ont montré une conservation du phénotype de la CSH âgée, ainsi les mécanismes du vieillissement de la CSH semblent avoir une composante intrinsèque importante. Ces études ont également montré qu'au sein de la niche de la jeune souris receveuse, la CSH âgée semble se localiser plus loin de l'endosteum (tissu conjonctif des cavités) que la CSH jeune et préférer la niche vasculaire (où sont présents les vaisseaux sanguins FLORIAN et al., 2012 ; KÖHLER et al., 2009). Elle est ainsi moins attachée à la niche et plus facilement mobilisable dans le sang que la CSH jeune (Figure 1.4.A&B).

## 1 Introduction – 1.1 Propriétés et vieillissement de la cellule souche hématopoïétique (CSH)

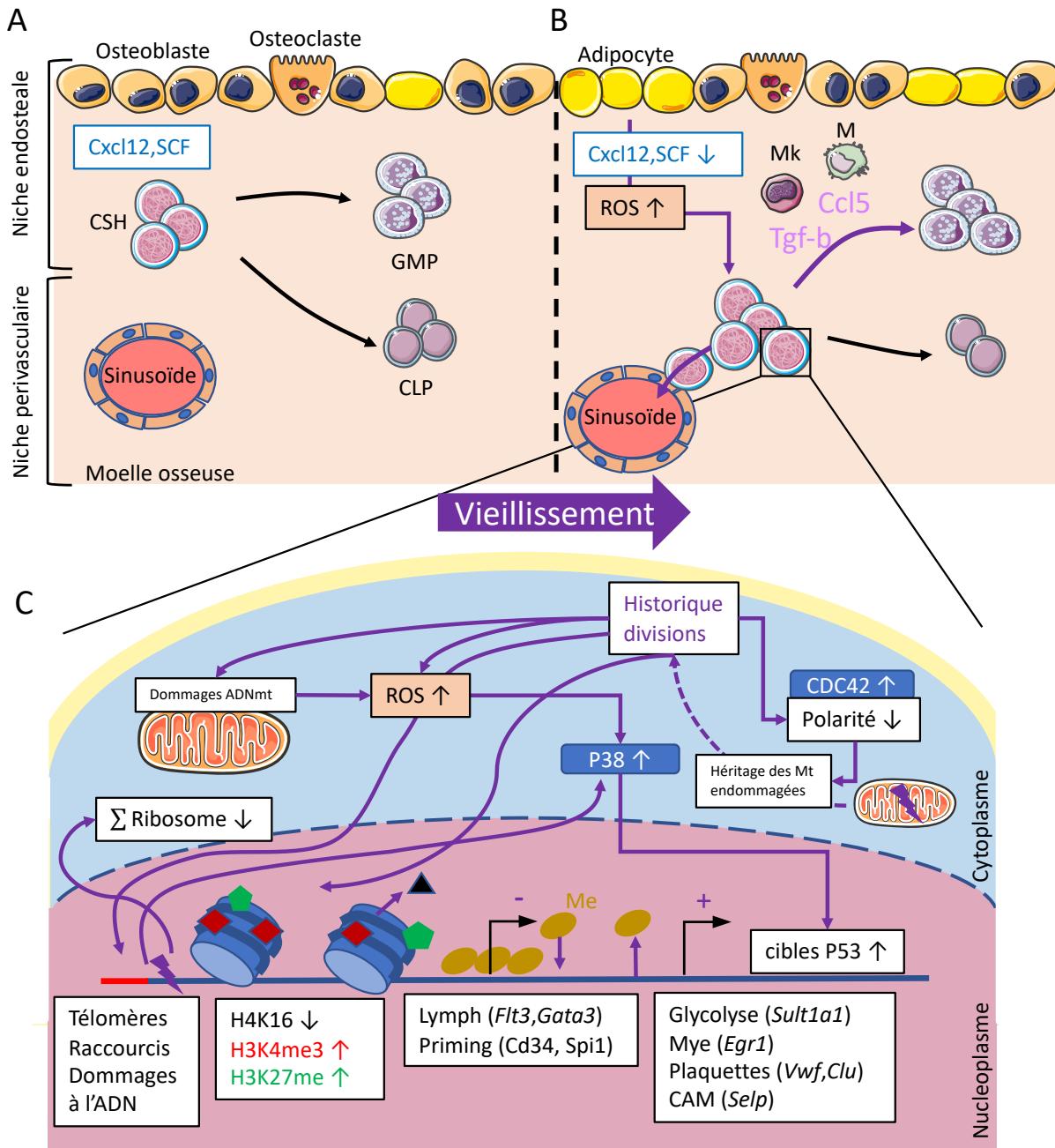


FIGURE 1.4 – Le vieillissement de la CSH

A : Population de CSH jeunes à l'équilibre entre autorenouvellement et départ en différenciation vers les **GMP** et **CLP**. Les CSH sont localisées près de l'endostéum. Les facteurs extrinsèques (Cxcl12, SCF) produits par la niche assurent l'équilibre de la population de CSH. B : Population de CSH âgées qui croît avec un départ en différenciation biaisé vers les **GMP** au détriment des **CLP**. Les CSH sont localisées près des vaisseaux sanguins et sont plus facilement mobilisables que chez la souris jeune. La population de la niche est altérée avec plus d'adipocytes et de cellules myéloïdes comme les macrophages (M) et les mégacaryocytes (Mk). En découle une altération de la signalisation avec une augmentation des signaux inflammatoires (Tgf-beta, CCL5) produits par les cellules myéloïdes au détriment des facteurs d'équilibres (SCF, Cxcl12). C : Facteurs intrinsèques du vieillissement de la CSH avec pour origine commune l'historique de division de la CSH. Lymph : Lymphotoïde, Mye : Myéloïde, CAM : Molécules d'adhésion cellulaire, Me : groupement méthyles (répression de la transcription), Mt : mitochondrie,  $\Sigma$  : biosynthèse (adaptée de GEIGER et al., 2013).

### 1.1.3.2 Facteurs intrinsèques du vieillissement de la CSH

Au regard des altérations que présente la CSH âgée, plusieurs mécanismes internes à la cellule pouvant être à l'origine de son vieillissement ont été proposés (Figure 1.4.C).

#### Raccourcissement des télomères, dommages à l'ADN et stress répliquatif

Une augmentation de foci H2AX, foyers d'histone 2A phosphorylée, considérés comme des marqueurs de cassure de l'ADN double brin, a été observée chez les CSH âgées murines et humaines (ROSSI et al., 2007; RÜBE et al., 2011). Le raccourcissement des télomères, séquences répétées d'ADN à l'extrémité des chromosomes pourrait être impliqué dans cette instabilité du génome. Située de part et d'autre de l'ADN linéaire, ces séquences sont spécialisées pour ne pas être reconnues ni réparées comme dommage à l'ADN. Elles évitent ainsi la fusion linéaire des chromosomes tout en protégeant l'intégrité des séquences internes. Les télomères sont maintenus par des telomérasées sans qui leur longueur est réduite à chaque division cellulaire ou du fait d'une augmentation du stress oxydatif. En sortant de quiescence, à plusieurs reprises au cours de sa vie, la CSH est soumise à ce stress et se révèle donc, tout comme d'autres types de cellules souches, particulièrement dépendante de ces enzymes (ALLSOPP et al., 2003). Suite à une activité ralentie ou absente des telomérasées l'érosion des télomères laisse sans protection les extrémités des chromosomes qui sont alors reconnues comme des dommages à l'ADN. L'accumulation de ces dommages pourrait interférer avec les fonctions de la CSH et provoquer sa sénescence : un arrêt irréversible du cycle cellulaire qui conduit à la mort de la cellule (GEIGER et al., 2013; ROSSI et al., 2007).

Cependant, une autre étude a montré que les foci H2AX dans les CSH âgées ne seraient pas dus à des dommages à l'ADN, puisqu'ils ne sont ni associés aux protéines clés de ce phénomène (53BP1 et pCHK1) ni à une fragmentation de l'ADN (FLACH et al., 2014). Ils seraient plutôt le résultat d'un blocage de la fourche de réPLICATION lors de la division de la CSH âgée. Leur persistance lors du retour à la quiescence de la CSH réprimerait l'expression des gènes des ARN ribosomaux (ARNr). Ce stress répliquatif et le stress ribosomal qui en découle participeraient au déclenchement du déclin de la fonctionnalité des CSH âgées (Figure 1.4.C).

#### Stress oxydatif et vieillissement mitochondrial

Lorsqu'elle sort de quiescence pour se diviser, la CSH passe à un métabolisme oxydatif. L'activation de la respiration cellulaire produit alors des *espèces réactives de l'oxygène –ou Reactive Oxygen Species– (ROS)*, incluant l'anion superoxyde O<sub>2</sub><sup>-</sup>. Le stress oxydatif important qui en résulte pourrait avoir un rôle majeur dans la perte de fonctionnalité de la CSH avec l'âge en étant responsable du raccourcissement des télomères et en affectant la structure des lipides et le repliement des protéines. Les *ROS* provoquent aussi des dommages et mutations sur l'ADN des mitochondries altérant leur fonction et augmentant en retour le niveau de ROS dans une boucle d'auto-activation (Figure 1.4.C). Dans le cas de l'hématopoïèse, l'inhibition des

ROS par la cystéine N-acetyl (NAC) dans la CSH restaure sa capacité d'autorenouvellement après greffe en séries (ITO et al., 2006). Les voies de signalisation de la kinase [cible de la rapamycine chez les mammifères](#) –ou *mammalian target of rapamycin*– (mTOR) et P38 MAPK dont les activités sont accrues dans la CSH suite à une élévation des ROS, ainsi que le facteur de transcription régulant le métabolisme FOXO3A ont été identifiés comme des acteurs principaux de ces mécanismes (GEIGER et al., 2013). Le lien entre mutations de l'ADN mitochondriale et perte de fonctionnalité de la CSH a également été suggéré avec l'étude de souris KO pour la polymérase mitochondriale POLG qui présentent une lymphopénie et une anémie (NORDDAHL et al., 2011). Cependant, le vieillissement prématûr de ce modèle ne se traduit pas par les altérations épigénétiques et transcriptomiques observées chez la CSH âgée. Ceci montre bien la multiplicité des facteurs impliqués dans le vieillissement et remet en question la vision d'une altération des mitochondries comme un de ses éléments déclencheurs.

### La division de la CSH comme source de son vieillissement

Chez la CSH âgée, les stress réplicatif et oxydatif ont pour origine commune la sortie de quiescence et la division cellulaire. Il n'est donc pas surprenant que des travaux montrent que les capacités d'autorenouvellement et de différenciation de la CSH sont anti-correlées à son historique de division (J. QIU et al., 2014). Une altération liée à l'âge provenant de la nature même du type de division a également été mise en cause par des études montrant une perte de la polarité de la CSH avec l'âge. Cette altération serait due à un changement vers un mode de division symétrique d'autorenouvellement, au détriment de la division asymétrique de différenciation. Ce changement aurait pour conséquence l'augmentation de la population de CSH non fonctionnelles observées (FLORIAN et al., 2012; FLORIAN et al., 2018). Ces travaux reposent sur l'observation d'une hausse de l'activité de la GTPase Rho CDC42 chez la CSH âgée qui aboutit à une perte de la polarité de la cellule avec une distribution aléatoire de la tubuline, de la marque épigénétique H4K16 et de CDC42 elle-même. La polarité de la CSH âgée peut être retrouvée en inhibant CDC42 ce qui conduit à un rajeunissement de la cellule qui retrouve une partie de sa fonctionnalité avec une augmentation de la lymphopoïèse B et une diminution de la myélopoïèse (FLORIAN et al., 2012). Les auteurs ont ensuite suggéré un lien entre le statut polaire ou apolaire de la cellule avant la division et la nature symétrique ou asymétrique de celle-ci. Les CSH jeunes polaires privilégieraient des divisions asymétriques et les CSH âgées plus apolaires des divisions symétriques (FLORIAN et al., 2018). Ces divisions asymétriques seraient gardées en mémoire grâce l'héritage des mitochondries endommagées de la CSH agée (voir section 1.1.2.2, page 25 et Figure 1.4.C). Cependant, ce modèle reste discutable de par la difficulté des études sur la symétrie des divisions dans la CSH et les résultats contradictoires déjà évoqués (section 1.1.2.2, page 25).

### 1.1.3.3 Facteurs extrinsèques du vieillissement de la CSH

Avec l'âge, l'environnement de la CSH se détériore également. La composition de la niche hématopoïétique évolue avec une altération des composants de la matrice extracellulaire, la diminution de la formation des os et une augmentation de la formation des tissus adipeux. Ceci s'accompagne d'une diminution des niveaux plamatiques de facteurs CXCL12, SCF et IL7 essentiels à la régulation de l'équilibre de la CSH (GEIGER et al., 2013). En parallèle, le nombre de mégacaryocytes, de macrophages, de cellules B associées à l'âge et de cellule myéloïdes suppressives augmentent dans la niche. Ces différentes cellules produisent des cytokines pro-inflammatoires comme CCL5, TNF-alpha, l'interferon-gamma et le TGF-beta1 qui participent ainsi au biais myéloïde de la différenciation de la CSH âgée (ZHANG et al., 2020).

Au niveau métabolique, l'altération de la niche endostéale se caractérise par une perte de l'environnement hypoxique nécessaire au maintien de la CSH quiescente et une hausse de la production de ROS. Ceci pourrait avoir comme origine une altération des jonctions gap entre CSH et cellules stromales. Ces jonctions permettent l'évacuation des ROS de la CSH prévenant ainsi l'activation des voies de signalisation de réponse aux dommages à l'ADN comme P38-MAPK (TANIGUCHI ISHIKAWA et al., 2012). L'implication potentielle dans le vieillissement de la CSH des CKI P16 (JANZEN et al., 2006) et P21 (CHOUDHURY et al., 2007) en aval de cette signalisation qui agissent directement sur le cycle cellulaire de la CSH renforce cette hypothèse (Figure 1.3 et 1.4.A&B).

### 1.1.3.4 Signature épigénétique et transcriptomique de la CSH agée

Les CSH âgées présentent 3 fois plus d'ARN total que les CSH jeunes en conséquence d'une activité de transcription accrue corrélée à une plus grande ouverture de l'ADN (SVENDSEN et al., 2021). Les mécanismes du vieillissement de la CSH qu'ils soient le résultat de facteurs intrinsèques, extrinsèques ou une combinaison des deux ont donc une signature qui peut être capturée par l'analyse globale de l'épigénome et du transcriptome.

#### Epigénétique

Les facteurs épigénétiques ont un rôle clé dans la régulation des états quiescent et multipotent de la CSH comme mentionné au début de l'introduction (section 1.1.1.4, page 20). De ce fait, plusieurs d'entre eux ont été identifiés comme liés au vieillissement du système hématopoïétique. On peut citer en premier lieu certains membres des complexes répressifs polycombs comme EZH2 et BMI1 (DE HAAN et GERRITS, 2007; NITTA et al., 2020). C'est également le cas de PLZF, un facteur recrutant des agents modificateurs de la chromatine, impliqué dans la différenciation des progéniteurs, dont la perte chez la souris provoque un phénotype de vieillissement amplifié (VINCENT-FABERT et al., 2016). De plus, on retrouve fréquemment

chez les personnes âgées souffrant de syndromes myélodysplasiques, des mutations dans les méthyltransferases et déméthylases d'ADN tel que *Tet2* et *Dnmt3a-b* (KRAMER et CHALLEN, 2017). Le profilage épigénomique de CSH jeunes et âgées a révélé une augmentation avec l'âge de la méthylation au niveau des gènes associés à la lignée lymphoïde (comme *Flt3* et *Gata3*) alors qu'une diminution est observée au niveau des gènes de la lignée myéloïde en accord avec le biais myéloïde de la CSH âgée (D. SUN et al., 2014). Au niveau des marques histones, les CSH âgées présentent des marques H3K4me et H3K27me plus larges et moins de marques H4K16ac. L'apparition de ces altérations épigénétiques serait aussi liée à l'historique de division de la CSH qui perdrait progressivement sa mémoire épigénétique au fur et à mesure de ses sorties de quiescence et de ses activations (BEERMAN et al., 2013; Figure 1.4.C). La conséquence de ces changements au niveau de l'expression des gènes n'est pas toujours bien visible au niveau de la CSH elle-même (voir ci-dessous) mais impacte plus la fonctionnalité des progéniteurs qui en découlent et héritent de ses altérations épigénétiques.

## Transcriptomique

Les facteurs du vieillissement ont largement été étudiés de façon globale à travers la comparaison de transcriptomes de CSH jeunes et âgées via de nombreuses plateformes, incluant les biopuces (CHAMBERS et al., 2007; FLACH et al., 2014; NORDDAHL et al., 2011), le séquençage de l'ARN de population (MARYANOVICH et al., 2018; RENDERS et al., 2021; D. SUN et al., 2014) et plus récemment le séquençage à l'échelle de la cellule (GROVER et al., 2016; HÉRAULT et al., 2021; KOWALCZYK et al., 2015 et voir Tableau 1.2). Une méta-analyse récente des données disponibles chez la souris a mis en évidence une grande variabilité des marqueurs transcriptomiques du vieillissement, entre les différentes études. Ceci s'explique en partie par la technologie de séquençage utilisée mais surtout par l'hétérogénéité du vieillissement lorsqu'il est étudié sur des pools de quelques souris. (SVENDSEN et al., 2021). Un consensus se dégage toutefois vers une altération de la transcription des gènes codant pour des protéines associées à la membrane avec une augmentation des gènes se rapportant à l'adhésion et aux jonctions cellulaires comme *Selp* codant pour la selectine P, tandis que les gènes des récepteurs cytokiniques de la différenciation comme *Flt3* sont réprimés. Certains régulateurs de la réPLICATION (*Mcm5-7*), de l'inflammation et de la prolifération (*Fap*, *Prtn3*, *Prtp*) ressortent également surexprimés avec l'âge dans cette méta-analyse mais dans des proportions beaucoup plus faibles qu'attendues. De façon tout aussi surprenante, les gènes *Cdkn1a* et *Cdkn2a* codant pour les CKI P21<sup>cip/kip</sup> et P16<sup>ink4</sup>, plusieurs fois identifiés comme altérés avec le vieillissement, ne sont pas retrouvés dans cette signature (CHOUDHURY et al., 2007; JANZEN et al., 2006). *Nuprl1* qui code pour une protéine senseur du stress et *Sult1a1* lié au métabolisme glycolytique de la CSH quiescente ont été identifié parmi les marqueurs transcriptomiques les plus robustes du vieillissement de la CSH. Un biais plaquettaire de la CSH âgée qui sur-exprime *Vwf*, *Clu*, *Mef2c* par rapport à la CSH jeune est également retrouvé

avec consistance (Figure 1.4.C).

## 1.2 Évolution des technologies cellules uniques pour appréhender l'hétérogénéité cellulaire et fonctionnelle des CSH

En plus de l'hétérogénéité du vieillissement, la variabilité des études transcriptomiques de CSH âgées trouve en partie ses origines dans l'hétérogénéité de la population de CSH. En effet, il ressort finalement de la première partie de cette introduction une diversité non seulement de devenirs mais aussi d'états et de comportements (quiescence, autorenouvellement, activation, biais de *priming*) qui s'avèrent être altérés par le vieillissement. Il apparaît plus juste d'étudier les HSPC comme une agrégation de différents groupes fonctionnels aux proportions changeantes en fonction des conditions de stress ou d'âge. Ce changement d'échelle dans l'étude de l'hématopoïèse précoce a été rendu possible ces 5 dernières années par les avancées dans les technologies de traçage de lignages et de séquençage à l'échelle de la cellule (technologies cellules uniques ou *single-cell technologies*).

### 1.2.1 Technologies de séquençage d'ARN cellules uniques (scRNA-seq)

C'est au début des années 1990 qu'a été réalisée pour la première fois la quantification de l'expression d'une dizaine de gènes cibles à l'échelle d'une cellule unique (neurone) (EBERWINE et al., 1992). Une quinzaine d'années plus tard, suite au développement des méthodes de réactions de polymérisation en chaîne –ou *Polymérase Chain Reactions*– (PCR) et à l'avènement du séquençage nouvelle génération –ou *Next Generation Sequencing*– (NGS) qui parallélise massivement les lectures, une première étude globale du transcriptome d'un blastomère unique de souris fut réalisée (TANG et al., 2009). Dans un premier temps, les analyses de séquençage d'ARN à l'échelle de la cellule –ou *single cell RNA sequencing*– (scRNA-seq) ont eu pour objectif la définition en profondeur de quelques types cellulaires bien définis. Par la suite, un changement s'est opéré lorsque le nombre de cellules analysées est monté à plusieurs centaines rendant possible l'identification de nouveaux types cellulaires (GUO et al., 2010). S'en est suivi une croissance exponentielle du nombre de cellules analysées, permettant la découverte de types et états cellulaires de plus en plus rares. Dans le cas de systèmes avec un turnover rapide comme l'hématopoïèse cette montée en charge a également permis d'obtenir des aperçus de l'état du transcriptome à différents temps (WATCHAM et al., 2019). De quelques centaines de cellules en 2010, il est désormais possible aujourd'hui d'en séquencer plus d'une centaine de milliers en une seule expérimentation. Ce saut d'échelle repose sur différentes avancées technologiques qui ont également diminué les volumes et coûts des réactifs. Le multiplexage des échantillons a permis dans un premier temps d'atteindre

le millier de cellules séquencées, puis des milliers de cellules ont pu être caractérisés avec l'introduction de la microfluidique. Finalement, la robotique et le barcoding *in situ* ont rendu possible l'analyse de plusieurs dizaines de milliers de cellules (SVENSSON et al., 2018).

En plus du nombre de cellules séquencées, l'autre paramètre clé d'une étude **scRNA-seq** est la profondeur de séquençage de chaque cellule qui détermine le nombre de gènes détectés dans chacune d'elle. Comme dans un run de séquençage le nombre de fragment d'ARN lus (reads) est limité par la taille de la flowcell, plus le nombre de cellules séquencées est important plus le nombre de reads par cellule diminue. Pour dépasser cette limitation, des plateformes de séquençage **scRNA-seq** ont fait le choix de ne séquencer que l'extrémité (une cinquantaine de paires de bases) 5' ou 3' des transcrits. Ces technologies à haut débit comme par exemple celle développée par *10X genomics* font donc l'impasse sur la détection et la quantification des isoformes pour privilégier la quantité de cellules analysées. Ce choix caractérise ainsi les méthodes dites haut-débit pouvant détecter de 2000 à 3000 gènes pour plus de 10 000 cellules séquencées. Il s'agit principalement de méthodes basées sur des gouttelettes dans lesquelles sont capturées les cellules ainsi que les réactifs nécessaires à la préparation de l'ADN complémentaire qui sera séquencé (Figure 1.5.A). D'autres méthodes dites bas débit (Smart-seq2, CelSeq2) permettent quant à elles un séquençage plus en profondeur des cellules sur toute la longueur des transcrits et la détection de plus de 5000 gènes pour environ un millier de cellules. Elles sont toujours utilisées aujourd'hui bien que les méthodes haut débit plus récentes sont de plus en plus privilégiées dans le but de caractériser de nouveaux types cellulaires rares (WATCHAM et al., 2019).

## 1.2.2 Méthodes d'analyse du scRNA-seq

De très nombreuses méthodes de clustering, de réduction de dimension, et d'analyse de l'expression différentielles fréquemment intégrées au sein de pipelines sont disponibles sous R (KISELEV et al., 2017; LUN et al., 2016; STUART et al., 2019) ou python pour des jeux de données plus importants (WOLF et al., 2018). Ces méthodes ont également été évaluées (TIAN et al., 2019). Dans cette partie nous décrirons plus particulièrement l'approche de Seurat qui est l'un des outils les plus utilisés et soutenus à l'heure actuelle (BUTLER et al., 2018; SATIJA et al., 2015; STUART et al., 2019) et dont les grandes étapes du pipeline d'analyse (sélections d'élément, réduction de dimension, clustering, recherche de marqueurs) sont reprises par de nombreux autres outils.

### 1.2.2.1 Traitement primaire et contrôle qualité

Le traitement primaire comprend le démultiplexage des fichiers brutes d'appel de base BCL en fichier texte de séquence qualité FASTQ. Les séquences de ces fichiers sont ensuite alignées sur le génome (obtention de fichier d'alignements BAM). Dans le cas de la technologie *10X ge-*

*nomics* chaque fragment aligné comprend 2 reads. Le premier contient la séquence biologique de la molécule d'ARN lue. Le deuxième inclut le barcode d'index qui renseigne sur la cellule d'origine, et le barcode appelé UMI marquant la molécule d'ARN qui permet de repérer (et de soustraire lors de l'analyse) les duplicités PCR (Figure 1.5.A). À partir des fichiers BAM, les UMI par cellules sont comptés pour générer la matrice gènes/barcodes qui est ensuite utilisée pour la suite de l'analyse. Cette matrice donne pour chaque cellule, pour chaque gène détecté dans celle-ci, un comptage d'UMI qui correspond au niveau d'expression du gène mesuré. Ce format [échange de marché –ou Market Exchange– \(MEX\)](#) permet d'éviter de stocker les comptages nuls très nombreux en scRNA-seq (Figure 1.5.B). Plusieurs critères de contrôle qualité peuvent ensuite être utilisés pour éliminer les cellules de faible qualité (faible nombre de gènes/d'UMI détectés, fort pourcentage d'ARN mitochondrial, caractéristiques de cellules mourantes) ou les doublets (cellules capturées et séquencées ensemble, caractérisées par un fort nombre de gènes/d'UMI détectés) (Figure 1.5.C).

D'une cellule à une autre, le nombre d'UMI, ou profondeur de comptage, peut varier du fait de la variabilité des étapes de capture, transcription inverse et séquençage d'une molécule d'ARN. Il est donc nécessaire de normaliser les données pour obtenir des abondances relatives d'expression géniques correctes entre les cellules. L'approche par défaut de Seurat consiste simplement à diviser chacun des comptages d'UMI pour chaque gène détecté dans une cellule par le nombre total d'UMI détectés dans la cellule. La distribution des données ainsi obtenue tendance à être fortement déséquilibrée vers les faibles comptages. Afin de travailler avec des distributions se rapprochant davantage de la normalité requise pour la plupart des analyses en aval, les comptages relatifs  $x$  sont multipliés par un facteur de 10000 puis transformés au logarithme ( $\log(10^4 x + 1)$ ). De nombreuses méthodes de normalisation plus complexes faisant appel à une modélisation paramétrique des distributions existent et peuvent s'avérer plus précises notamment dans le cas où des différences importantes de profondeur de comptages existent entre les différents groupes de cellules analysées (LUECKEN et THEIS, 2019).

### 1.2.2.2 S'affranchir du bruit et de la dimensionnalité

Le séquençage d'ARN en cellule unique génère des matrices de grandes tailles difficilement analysables telle quelles. Celle-ci présentent en effet un bruit de fond important, notamment pour les technologies haut débit utilisant des gouttelettes, avec de nombreux événements dits de [drop-outs](#) pour lesquels le transcrit bien que présent dans la cellule n'est pas détecté à cause de la faible couverture de séquençage. Après le traitement brut des données de séquençage (alignement sur le génome, comptage, contrôle qualité des cellules) une première étape d'analyse consiste donc à s'affranchir de ce bruit important. L'approche couramment utilisée consiste d'abord à identifier les gènes qui varient le plus (généralement les 2000 premiers) au regard de la relation entre variance et moyenne d'expression après normalisa-

tion du jeu de données puis à réaliser une première réduction de dimension pour résumer l'information. Des méthodes linéaires comme l'[Analyse en Composantes Principales \(ACP\)](#) sont couramment utilisées dont on retient les composantes informatives, généralement les 15 à 30 premières pour les analyses en aval. Cette sélection est souvent réalisée en identifiant la chute de la proportion de variance expliquée par les principales composantes (Figure 1.5.D). La linéarité de l'ACP permet une interprétation régulière des distances de l'espace réduit dans toutes les régions de celui-ci. Il est ainsi possible d'évaluer la corrélation des différentes composantes avec certaines covariables de nuisances techniques (nombre d'UMI par cellules, proportion d'ARN mitochondriaux) pour étudier les performances des étapes de contrôle qualité et de normalisation notamment (BUTLER et al., 2018; LUECKEN et THEIS, 2019).

### 1.2.2.3 Classification des cellules

Dans le but de distinguer différents types cellulaires, la plupart des pipelines d'analyse proposent, dans ce premier espace réduit, une classification non supervisée ou *clustering* des cellules. En parallèle, peut être menée une nouvelle réduction de dimension pour aboutir à une représentation des données dans un espace à 2 ou 3 dimensions visualisable par l'être humain qui vient confirmer ou non le résultat du clustering. L'approche de Seurat consiste à construire un graphe de cellules en prenant pour chacune d'elles les 20 à 30 plus proches voisines dans l'espace des composantes principales retenues. Seurat utilise ensuite sur ce graphe des algorithmes de clustering comme Louvain ou Leiden (TRAAG et al., 2019). En parallèle, l'espace des composantes principales retenues est réduit à deux ou trois dimensions pour la visualisation à l'aide d'un [Uniform Manifold Approximation and Projection \(UMAP\)](#) qui a l'avantage de mieux préserver la structure globale des données que les [t-distributed stochastic neighbor embedding \(tSNE\)](#) auparavant très utilisés (Figure 1.5.E; BECHT et al., 2019).

En plus d'une classification non supervisée, il peut également être intéressant d'utiliser la connaissance disponible d'autres jeux de données [scRNA-seq](#) en lien avec le processus étudié pour caractériser les clusters identifiés. Il est par exemple possible de classer les cellules en s'appuyant sur des étiquettes de types cellulaires (LIEBERMAN et al., 2018) ou de phases du cycle cellulaire (SCIALDONE et al., 2015) préalablement déterminées pour chaque cellule à partir de jeux de données [scRNA-seq](#) étiquetés. Ces techniques d'apprentissages supervisées à partir de jeux de données d'entraînement s'appuient sur des méthodes variées telles que : (i) l'intégration des deux jeux de données via l'identification d'ancres (paires de cellules identiques) entre les données étiquetées et les données à classifier (STUART et al., 2019) ou (ii) l'entraînement de classificateurs à l'aide de méthodes d'apprentissage comme les arbres de décisions boostés (LIEBERMAN et al., 2018) ou les réseaux de neurones (MA et PELLEGRINI,

2020). Dans le cas de l'hématopoïèse précoce on peut citer plus particulièrement le hscScore un classificateur entraîné sur des cellules souches hautement purifiées qui donne un score d'état souche aux cellules analysées (HAMEY et GÖTTGENS, 2019).

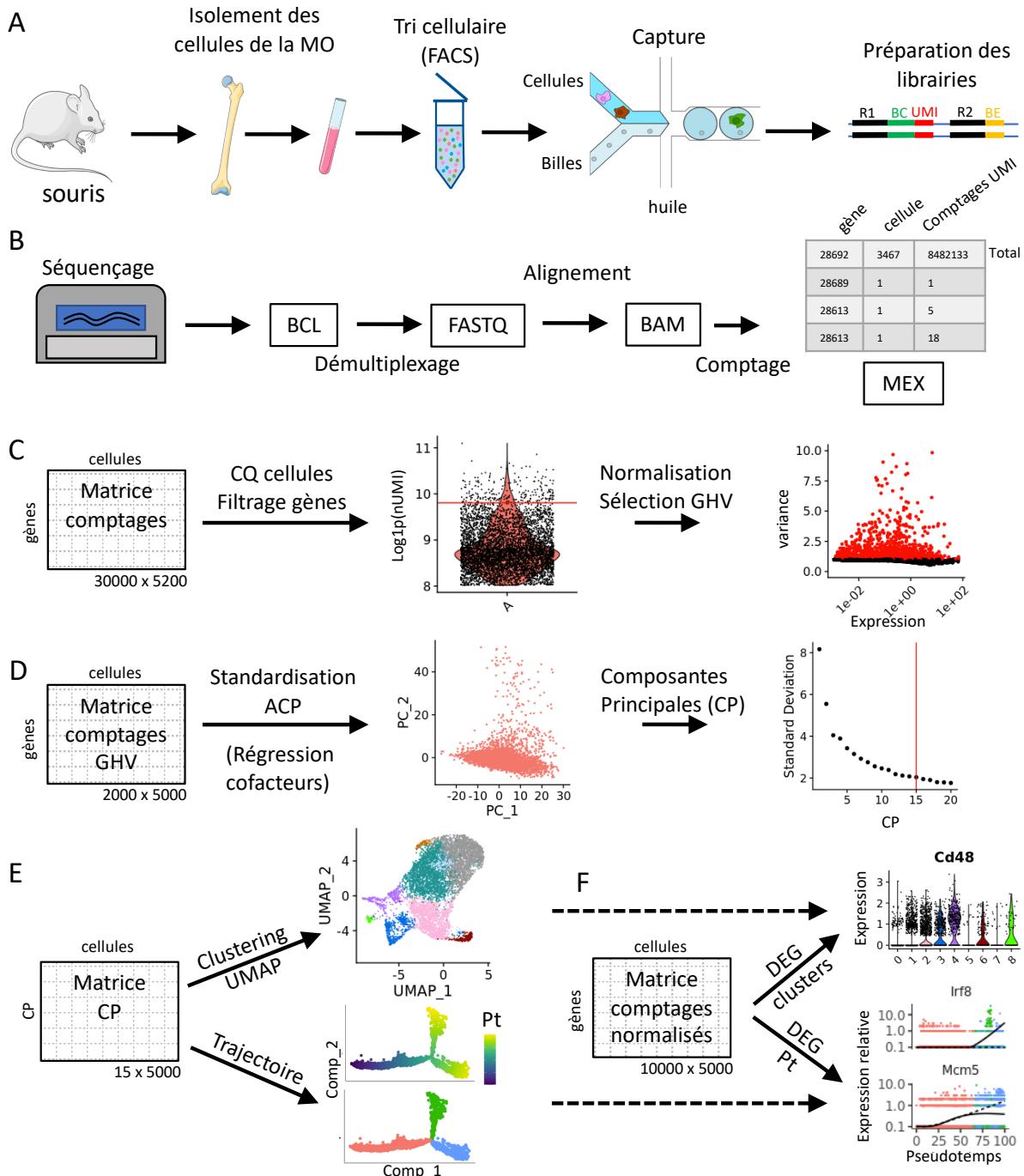
#### 1.2.2.4 Recherche des marqueurs de populations

L'identification des groupes (ou clusters) de cellules ainsi obtenus se fait ensuite par la recherche de leurs marqueurs par l'analyse des *gènes différemment exprimés* –ou *Differentially Expressed Genes*– (DEG) entre les différents groupes. Des seuils de significativité et/ou de différence d'expression sont utilisés pour sélectionner les marqueurs les plus pertinents pour chaque groupe (Figure 1.5.F). Il ressort des études d'évaluation des méthodes statistiques disponibles utilisables pour le scRNA-seq que les tests classiques de Student ou non paramétriques de Wilcoxon réussissent aussi bien voire mieux que les méthodes statistiques plus complexes élaborées pour le RNA-seq en "bulk". Ceci s'explique par la grande taille des échantillons puisqu'en scRNA-seq chaque cellule d'un groupe est un réplicat (SONESON et ROBINSON, 2018).

Finalement, ces marqueurs permettent de définir biologiquement les différents clusters aussi bien en termes de types cellulaires attendus que d'états biologiques (prolifération, quiescence par exemple). Une expertise biologique du système biologique étudié est souvent la clé pour une définition pertinente. Elle peut être assistée par des tests d'enrichissements des marqueurs annotées pour des termes d'ontologies ou de voies de signalisations dans les différentes bases de données biologiques (Gene Ontologies, Kegg pathways, Reactome).

#### 1.2.2.5 Construction de pseudo-traj ectoires de différenciation

Dans le cas de l'étude de processus biologiques continus comme la différenciation ou le développement, des méthodes de construction de trajectoire ont également été développées. Elles partent du principe que chaque cellule séquencée peut être considérée comme une photo à un instant T du processus étudié. Il faut donc les réordonner en fonction des changements continus de l'expression des gènes à partir d'un état initial vers un ou plusieurs états terminaux ou différenciés (Figure 1.5.E&F). Ces approches s'appuient en général encore une fois sur des méthodes de réduction de dimension qui visent plus particulièrement dans ce cas à préserver la structure globale (petite et grandes distances entre cellules, embranchements) de l'espace initial de haute dimension plutôt qu'à séparer les différents groupes de cellules (CHEN et al., 2019 ; X. QIU et al., 2017). La construction d'un chemin de différenciation avec ses bifurcations dans l'espace ainsi réduit permet ensuite l'obtention d'un pseudotemps. Celui-ci se définit pour chaque cellule comme la distance de cette cellule à la cellule initiale, qui a été en général déterminée manuellement en s'appuyant sur l'expression de marqueurs connus.



## FIGURE 1.5 – Etude scRNA-seq de l'hématopoïèse

A : Isolement, tri, capture des cellules d'intérêt et préparation des librairies. Exemple de la technologie 10X, R1 : Read 1 biologique, BC : barcode cellule, BE : barcode échantillon. B : Traitement primaire des données après séquençage. BCL : Fichier d'appel de base, FASTQ, fichier séquence qualité, BAM : fichier d'alignement, MEX : fichier matrice de comptage. C : Contrôle qualité (CQ), dans cet exemple 10 000 gènes sont conservés, normalisation et sélection des gènes les plus variables, ici 2000 gènes hautement variables (GHV). D : Réduction de la dimensionnalité par ACP. E : Visualisation en 2 dimensions par UMAP, clustering des cellules et inférence de trajectoires. F : Analyse des DEG entre clusters et au cours du pseudotemps (Pt).

Depuis leur apparition dans le milieu des années 2010, le développement des méthodes d'inférence de trajectoire a suscité un très fort intérêt de la communauté scientifique. Plus de 70 méthodes ont été proposées et l'évaluation d'une partie de celles-ci a mis en lumière leur grande hétérogénéité en termes de résultats (SAELENS et al., 2019). Ces méthodes peuvent être classées selon la topologie (linéaire, trajectoires divergentes, cycles) des trajectoires qu'elles peuvent inférer. Le choix de la méthode doit donc être pris en fonction des connaissances au préalable du système étudié mais également en fonction des dimensions du jeu de données. En effet, les différents algorithmes proposés peuvent présenter des temps de calcul allant de quelques secondes à plusieurs heures pour un jeu de données standard de 10 000 cellules avec 10 000 gènes analysés.

### 1.2.3 Les biais d'analyses en scRNA-seq et leurs résolutions possibles

Le très grand nombre d'études de scRNA-seq menées au cours de ces dix dernières années a permis d'approfondir grandement nos connaissances des différents processus biologiques notamment dans le cadre de l'hématopoïèse (voir section 1.2.4, page 42). Cependant, des travaux sur des même types cellulaires ont parfois abouti à des résultats divergents. C'est le cas par exemple des études des modifications du cycle cellulaire de la CSH au cours du vieillissement (voir section 3.6, page 117). Ces différences proviennent en partie du choix de la technologie et du rapport entre profondeur de séquençage et nombre de cellules séquencées qui influent directement sur la précision des signatures transcriptionnelles (favorisée par une profondeur de séquençage importante) mais aussi du clustering des cellules (favorisé par un nombre élevé de cellules étudiées). Néanmoins, ces différences soulèvent également la question des biais d'analyses des données scRNA-seq relatifs aux méthodes utilisées.

#### 1.2.3.1 Biais des représentations dans des espaces réduits

Un premier biais majeur des méthodes actuelles et difficilement contournable vient de la réduction de dimension nécessaire à la visualisation des données. Aujourd'hui, le pipeline qui comprend (1) une réduction linéaire (ACP) pour résumer l'information puis (2) une réduction non linéaire pour la visualiser dans un espace de 2 à 3 dimensions (UMAP) fait de plus en plus consensus (voir section 1.2.2.3, page 35; Figure 1.5.D&E; LUECKEN et THEIS, 2019). Ce n'était pas vraiment le cas il y a 5 ans, ce qui a abouti à des stratégies de réduction de dimension souvent différentes d'une étude à l'autre. Le choix de la dimensionnalité des données pour résumer l'information (1ère réduction de dimension) est source de biais puisqu'il se fait bien souvent de manière heuristique par l'observation du coude dans le graphe de la part de variance expliquée de chaque composante (Figure 1.5.D). L'étude des gènes majeurs de chaque composante avec une expertise du processus biologique étudié est recommandée pour arrêter ce choix. D'autre part, bien que des progrès importants ont été réalisés, il existe

toujours une perte d'information lors de la réduction de dimension, dommageable pour les interprétations biologiques faites dans les espaces réduits et finaux de seulement 2 ou 3 dimensions (WATCHAM et al., 2019).

### 1.2.3.2 Biais des mesures d'expression à cause de la faible couverture de séquençage

La faible couverture conduit à la non quantification de transcrits pourtant bien exprimés par la cellule. Ces événements de *drop-outs* sont un autre biais majeur inhérent aux données scRNA-seq. Si ce biais est maintenant assez bien pris en compte pour la représentation des données dans un espace réduit, le clustering des cellules et l'inférence de trajectoires, il impacte encore les conclusions biologiques qui en découlent. Des méthodes d'imputation de comptages manquants (dus aux *drop-outs*) peuvent être utilisées pour une analyse individuelle plus précise de l'expression des gènes, mais celles-ci sont déconseillées pour les étapes de réduction de dimension et la classification des cellules (HOU et al., 2020). Une autre correction efficace de ce biais et peu coûteuse en temps de calcul est la recherche d'empreintes d'activités de facteurs de transcription ou de voies de signalisation dans chaque cellule (HOLLAND et al., 2020). Cette approche consiste à tirer profit des nombreuses signatures transcriptomiques de facteurs de transcription ou de voies de signalisation établies par des études de RNA-seq en bulk pour calculer des scores dans chaque cellule. Ces signatures étant généralement constituées de plusieurs dizaines de gènes, elles sont beaucoup moins sensibles aux *drop-outs* que l'analyse individuelle d'un marqueur. Le calcul des scores de signatures par cellule peut se faire de manière assez simple en considérant tous les gènes de celles-ci au même niveau et en évaluant leur enrichissement dans le quantile supérieur des gènes les plus exprimés par chaque cellule (AIBAR et al., 2017). Il peut aussi se faire par des méthodes plus complexes qui combinent par exemple un modèle linéaire à l'ensemble des gènes de la signature (DUGOURD et SAEZ-RODRIGUEZ, 2019).

### 1.2.3.3 Biais dus à des effets biologiques

L'hétérogénéité transcriptomique d'une population peut être masquée en scRNA-seq par des processus biologiques comme le métabolisme ou le cycle cellulaire qui brouillent les différences plus discrètes entre types cellulaires. Il n'est donc pas surprenant que les premières études en scRNA-seq du compartiment HSPC ont dans leur ensemble pointé majoritairement des différences de cycle (KOWALCZYK et al., 2015; N. K. WILSON et al., 2015; YANG et al., 2017) et de métabolisme (CABEZAS-WALLSCHEID et al., 2017) sans que ne soit proposé de pseudo-trajectoire de différenciation. Ces études ne corrigeaient pas ou que partiellement les effets du cycle en écartant simplement les gènes du cycle cellulaire de l'analyse. Comme il est devenu évident que tout le transcriptome était affecté par la position de la cellule dans le cycle cellulaire (SCIALDONE et al., 2015), une correction de l'ensemble du transcriptome

par rapport au cycle cellulaire s'est imposée par la suite, permettant la détection de nouvelles sous populations (BUETTNER et al., 2015) et la construction de trajectoires, au sein du compartiment HSPC (HÉRAULT et al., 2021). Dans leur version la plus simple, ces corrections consistent à régresser linéairement l'effet indésirable de l'expression de l'ensemble des gènes avant l'étape de réduction de dimension à l'aide d'un modèle linéaire calculé à partir de scores de signatures (voir ci-dessus) du cycle cellulaire (BUTLER et al., 2018; WOLF et al., 2018).

D'autres covariables, biologique (stress mesuré par le taux d'ARN mitochondriaux) ou technique (nombre de molécule d'ARN par cellules) peuvent être corrigées de la même façon. Il est important de souligner que très souvent ces covariables indésirables sont liées et, que si la décision d'en corriger plusieurs est prise, il est alors préférable de régresser leurs effets à l'aide d'un modèle unique qui les considère toutes ensembles (LUECKEN et THEIS, 2019). Ces corrections doivent être menées avec précaution car elles peuvent amener à masquer des processus biologiques d'intérêts liés aux variables corrigées. Il est ainsi recommandé de ne pas les utiliser que pour les analyses de réduction de dimension et d'inférence de trajectoire une fois le biais constaté sans correction (Figure 1.5.D; LUECKEN et THEIS, 2019).

#### 1.2.3.4 Biais dus aux effets de lots

Les études de scRNA-seq sont aussi particulièrement sensibles aux effets de lots d'expériences ou *batch effect* à partir du moment où l'ensemble des cellules étudiées est manipulé en plusieurs lots et ce même si les protocoles expérimentaux sont strictement identiques entre chaque lot. Au niveau des profils transcriptomiques obtenus ce biais se manifeste notamment par des profondeurs de séquençage différentes entre chaque lot mais également par des activations plus ou moins fortes de groupes de gènes liés à des processus cellulaires de stress comme les gènes des protéines ribosomales ou mitochondrielles. Ces biais peuvent être évités ou du moins fortement atténués en s'appuyant sur une conception expérimentale habile et en utilisant des techniques récentes de marquage cellulaire. Une technique désormais répandue, l'[indexage cellulaire des transcriptomes et des épitopes par séquençage – ou Cellular Indexing of Transcriptomes and Epitopes by Sequencing – \(CITE-seq\)](#) utilise par exemple des anticorps oligo-tagués dirigés contre des protéines de surface ubiquitaires, ce qui permet de marquer de manière unique des cellules provenant d'échantillons distincts. Le séquençage de ces étiquettes couplé au transcriptome cellulaire permet d'associer chaque cellule à son échantillon d'origine (STOECKIUS et al., 2018). Si cela s'avère insuffisant, il existe des méthodes efficaces pour corriger l'expression des gènes en fonction des lots (BÜTTNER et al., 2019). Certaines ont été développées bien avant l'arrivée du scRNA-seq et s'aident de modèles linéaires prenant en compte la variable de lot comme ComBat (JOHNSON et al., 2007).

En l'absence de marquage cellulaire, le biais sera d'autant plus problématique lorsque

les lots correspondent à des conditions expérimentales différentes (test d'une perturbation, d'un traitement, récolte des cellules à différents temps d'un processus). La composition des populations cellulaires n'est alors pas forcément la même entre les différents lots et les méthodes qui réduisent l'effet de lots en prenant en compte toutes les cellules comme ComBat ne seront pas capables de différencier l'effet de la condition testée de l'effet de lots. Dans ce cas, une correction efficace est néanmoins possible en intégrant les différents jeux de données ensemble dans un espace commun. Plusieurs méthodes ont été proposées dans ce but et s'appuient sur l'identification de sous-ensembles d'états cellulaires partagés entre les jeux de données (BUTLER et al., 2018; HAGHVERDI et al., 2018; KORSUNSKY et al., 2019). Par exemple, Seurat utilise l'analyse canonique des corrélations pour trouver des paires de cellules communes, ou ancrés, entre les jeux de données et projeter ceux-ci dans un nouvel espace dont les composantes informatives sont ensuite utilisées pour les analyses en aval (réduction de dimension, clustering de cellules et inférence de trajectoires; BUTLER et al., 2018). L'intégration de données ne se limite pas à la correction de l'effet de lots, elle est également essentielle pour l'analyse conjointe de différentes technologies omiques à l'échelle de la cellule unique (voir section 1.2.5.2, page 47).

### 1.2.3.5 La difficulté de l'inférence de trajectoire

Une grande hétérogénéité des résultats concerne les analyses de pseudo-trajectoires de différenciation. En effet, les très nombreuses méthodes disponibles et la variabilité de leur résultat amène à la plus grande prudence quant aux embranchements et aux valeurs de pseudotemps calculées. Tandis que les analyses de trajectoires et/ou de réduction de dimension donnent une position de chaque cellule dans l'espace du processus biologiques étudié, les analyses de vitesse de l'ARN permettent d'estimer la dérivée temporelle des cellules (vecteur vitesse) en distinguant les ARNm épissés et non épissés du jeu de données (LA MANNO et al., 2018). Il peut donc être intéressant de confronter les résultats d'inférence de trajectoires aux analyses de la vitesse de l'ARN en scRNA-seq pour obtenir une meilleure vision de la dynamique du processus. Cependant, relativement peu d'études sur l'hématopoïèse en scRNA-seq ont utilisé cette démarche sans doute du fait du rôle de l'épissage dans la régulation du devenir cellulaire, notamment au niveau de la CSH (BOWMAN et al., 2006).

### 1.2.4 Une hétérogénéité et une différenciation complexe des CSH mis en évidence par le scRNA-seq

Comme ce fut le cas pour la microscopie, la radiologie ou encore le tri cellulaire, la communauté scientifique en hématologie s'est rapidement emparée de la technologie scRNA-seq, avec une étude pionnière présentant les transcriptomes de plus de 2700 progéniteurs murins myélo et érythroïdes (PAUL et al., 2015). S'en est suivi, le profilage de tout le compartiment cellulaire de la moelle osseuse chez l'humain (plus de 90 000 cellules, X. HAN et al., 2018),

des études couplant le scRNA-seq au traçage de lignages issus de CSH murines (RODRIGUEZ-FRATICELLI et al., 2020; RODRIGUEZ-FRATICELLI et al., 2018; WEINREB et al., 2020) et plus récemment, une étude pré-publiée de single cell multi-omiques de CSH humaines (SOMMARIN et al., 2021). Des dizaines d'études en hématologie s'appuyant sur le scRNA-seq ont ainsi été publiées ces 7 dernières années et une sélection des principales études scRNA-seq menées sur le compartiment cellules souches et progéniteurs avec leur caractéristiques et résumés est donnée Table 1.2.

Ces travaux dont la chronologie montre bien la montée en puissance du nombre de cellules séquencées ont tout d'abord remis en question la vision traditionnelle de l'hématopoïèse comme un processus en étapes hiérarchisées, représenté comme un arbre. En effet, les analyses de pseudo-traj ectoires de différenciation ont révélé un continuum de différenciation à partir de la population de CSH murines (Figure 1.6.A&B; HERMAN et al., 2018; NESTOROWA et al., 2016). Bien que globalement les trois trajectoires vers les lignées érythroïde/mégacaryocytaire, myéloïde et lymphoïde soient retrouvées, les localisations des différents embranchements sont loin de faire consensus et la notion même de progéniteurs intermédiaires est remise en question par ces études (WATCHAM et al., 2019). C'est notamment le cas des GMP et CMP dont des études scRNA-seq ont mis en évidence leur grande hétérogénéité avec des paysages transcriptomiques complexes (GILADI et al., 2018; OLSSON et al., 2016).

Ces études ont également identifié un point d'amorçage précoce vers les différents lignages dans les HSPC, caractérisé par un continuum de différenciation et non de groupes de cellules distincts (Figure 1.6.C; HÉRAULT et al., 2021; RODRIGUEZ-FRATICELLI et al., 2018). Ainsi, la vision révisée de l'hématopoïèse se rapproche beaucoup plus des paysages de la différenciation cellulaire de Waddington dans lesquels la différenciation d'une cellule est représentée par une bille descendant d'une colline vers des vallées divergentes, chacune d'elles menant finalement à un type cellulaire différent (Figure 1.6.D; WADDINGTON et KACSER, 1957). Le relief de ces paysages représente le choix entre une vallée et une autre pour la cellule en différenciation qui reposera sur les facteurs de transcription, sorte de haubans, capables de recruter des complexes modificateurs de la chromatine, et de promouvoir des phénotypes cellulaires (Figure 1.6.E; BUENROSTRO et al., 2018).

Le scRNA-seq a permis de distinguer des CSH quiescentes et actives (CABEZAS-WALLSCHEID et al., 2017; N. K. WILSON et al., 2015; YANG et al., 2017) avec la mise en évidence du rôle de la signalisation de l'acide rétinoïque pour le maintien de l'état de dormance hypoxique (CABEZAS-WALLSCHEID et al., 2017; LAURIDSEN et al., 2018) ce qui influence probablement la différenciation des CSH.

## 1 Introduction – 1.2 Évolution des technologies cellules uniques pour appréhender l'hétérogénéité cellulaire et fonctionnelle des CSH

Tableau 1.2 – Sélections d'études scRNA-seq sur les CSH et leur devenir

TC : Techniques complémentaires, NC : Nombre de cellules, PI : Phénotypage Immunologique, V : Vieillissement, KO : Knock Out. Populations : sauf si le contraire est indiqué, les cellules sont triées de la moelle osseuse de souris (\*) ou d'humain (\*\*)

Référence	Populations	NC	Plateforme	TC	Perturbations	Résumé
N. K. Wilson, 2015	LTHSC*	>90	SMART-Seq2	PI	Induction d'anémie	cluster CSH en autorenouvellement
Yang, 2017	HSC*, EML (Ery., Mye., Lymph.)*	>220	C1	PI	differences de cycle	
Cabezas-Wallscheid , 2017	LTHSC*	>310	C1 + SMARTer		differences de cycle; Ac. Rétinoïque CSH quiescente	
Lauridsen, 2018	LTHSC*	>1200	MARS-seq	rapporteur CFP	differences de cycle; Ac. Rétinoïque CSH quiescentes	
Rodriguez-Fraticelli, 2018	LTHSC, STHSC, MPP2-3-4*	>4900	inDrops	tracage de lignages, PI	amorçage précoce, différenciation CSH en Mk directe	
Kowalczyk, 2015	LTHSC , STHSC, MPP3-4*	>1100	SMART-seq	V	difference de cycle (G1 plus court CSH agée)	
Young, 2016	LMPP/MPP4*	>90	C1 Autoprep	V(1/14)	prolifération +, potentiellement lymphoïde et MPP - chez souris agées	
Grover, 2016	LTHSC*	>130	C1 + SMARTer	V	biais plaquette des CSH agées	
Kirschner, 2017	LTHSC*	>420	SMART-Seq2	V, JAKSTAT	cluster CSH Quiesc/mye étendu par JAKSTAT	
Florian, 2018	CSH (LSK,CD34-Flik2-)*	>290	SMART-Seq4	V, CASIN, Wnt5a	assymétrie, polarité CSH jeunes/agées	
Mann, 2018	LTHSC, STHSC, MPP3-4*	>940	SMART-seq2	PI	V,LPS	cluster LTHSC mye CD61-high amorcé en réponse au LPS, amplifié avec l'âge
Hérault, 2021	LTHSC, STHSC, MPP2-3*	>14900	10X Chromium	V	trajectoire d'amorçage des CSH retard des CSH agées quiesc.	
Paul, 2015	Lin-, Kit+, Sca1- *	>2700	MARS-Seq	PI	Cebpe-a KO	biais de lignages des CMP
Nestorowa, 2016	LTHSC, MPP,LMPP, MEP, CMP, GMP*	>1600	SMART-Seq2	PI	Trajectoires de différenciation vers les lymph. B, CD, neutro/monocytes et erythrocytes	
Olsson, 2016	LSK, CMPS, GMPS*	>380	C1 + SMARTer		Biais de lignages des progéniteurs oligopotents	

1 Introduction – 1.2 Évolution des technologies cellules uniques pour appréhender l'hétérogénéité cellulaire et fonctionnelle des CSH

Tableau 1.3 – Sélections d'études scRNA-seq sur les CSH et leur devenir (2)

suite légende : Les types murins LTHSC, STHSC, MPP1-4 et LMPP font référence au phénotypage Tableau 1.1. SC : sang du cordon ombilical. LK : Lin<sup>-</sup>; cKit<sup>+</sup>, LSK : Lin<sup>-</sup>; Sca1<sup>+</sup>; c-Kit<sup>+</sup> (Adapté de WATCHAM et al., 2019).

Weinreb, 2020	LSK, LK (et descendants)*	>300 000	InDrops	greffe, traçage de lignages		Carte de l'hématopoïèse, Devenir d'avantage prédict par la filiation que par le transcriptome
Rodriguez-Fraticelli, 2020	LTHSC, STHSC, MPP2-4, MkP (LK CD150+ CD41+)*	>100 000	InDrops	traçage de lignages, CRISPR-seq	greffe, CRISPR	LTHSC fonctionnelles TCF15 <sup>+</sup>
Giladi, 2018	MO totale, Lin <sup>-</sup> , LK cells*	>80 000	MARS-seq	CRISPR-seq, PI	CRISPR	trajectoires de différenciation au cours de l'hématopoïèse normale et perturbée
Dahlin, 2018	LSK, LK *	>58600	Smart-seq2; 10X Chromium		c-Kit KO	trajectoires de différenciation, pas de différenciation Mast avec la perte de cKit
han, 2018	51 tissues dont MO totale*	>400 000 (MO: >93000)	Microwell-seq			Vision gloable des cellules de la MO et des des progéniteurs
Dong, 2020	LT/STHSC, MPP cellules matures*	>2300	SMART-seq2	PI	greffe	Cinétique et trajectoire de différenciation des CSH transplantées
Velten, 2017	HSPC (CD34+, Lin-)**	>1400	Quartz-Seq	PI		NUAGE de CSH chez l'homme, suivi par des progéniteurs unipotents
Sommarin, 2021	CSH (Lin- CD34+ CD38-)** SC (CD49f+CSH)***	>46 000	10X Chromium	Cite-seq ATAC-seq	✓	Hétérogénéité du nuage-HSC au cours de la vie; CSH+, MPP lymph - avec l'âge
Buenrostro, 2018	CSH, CMP, GMP**	>7800	10X Chromium	PI ATAC-seq		La variabilité de la chromatine associée aux motifs de TF dans les CSH suit les voies érythroïde/lymphoïde.

## *1 Introduction – 1.2 Évolution des technologies cellules uniques pour appréhender l'hétérogénéité cellulaire et fonctionnelle des CSH*

Ces études ont aussi souligné des différences entre l'humain et la souris. L'analyse du compartiment CSH et **MPP** humain (cellules CD34<sup>+</sup>) suggère en effet un nuage de CSH aux transcriptomes très proches donnant directement des progéniteurs unipotent avec encore moins de hiérarchie de différenciation entre les cellules que chez la souris (VELTEN et al., 2017). Des pseudo-trajectoires de différenciation de la CSH humaine vers les lignées lymphocytes B, DC, neutrophiles/monocytes et érythrocytes ont par ailleurs pu être construites (HERMAN et al., 2018).

Ces études proposent également bien souvent des trajectoires de différenciation de la CSH légèrement différentes pour un même organisme, notamment en ce qui concerne la localisation des embranchements. En plus des biais d'analyses discutés précédemment (voir section 1.2.3, page 38), une autre explication de ces différents résultats pour un même organisme est que le transcriptome ne serait pas le seul guide du destin cellulaire et que des décisions majeures seraient prises à d'autre niveaux, notamment épigénétique et post-transcriptionnel (LAURENTI et GÖTTGENS, 2018; WATCHAM et al., 2019). Cette vision multi-niveau de la régulation du devenir de la CSH a conduit au développement récent de technologies complémentaires au scRNA-seq toujours à l'échelle de la cellule unique.

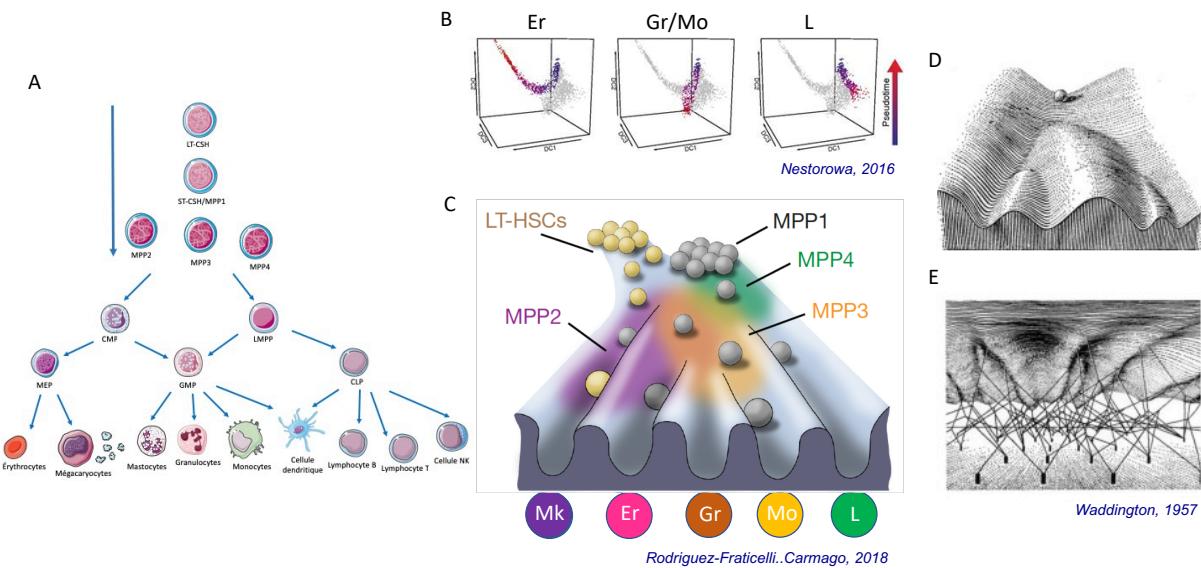


FIGURE 1.6 – Evolution de la vision de l'hématopoïèse avec le scRNA-seq

A : Vision classique de l'hématopoïèse comme un arbre de différenciation avec plusieurs étapes hiérarchisées. B : Trajectoires de différenciation vers les lignées Érythroïde (Er), Granulo (Gr)/ Monocytaire (Mo) et Lymphoïde (L) inférées à partir de données scRNA-seq (NESTOROWA et al., 2016). C : Proposition de modèle d'amorçage transcriptionnel continu de la CSH établie par inférence de trajectoire à partir de scRNA-seq et traçage de lignages (RODRIGUEZ-FRATICELLI et al., 2018). D : Paysage de différenciation transcriptionnelle d'une cellule souche au sommet d'une colline avec plusieurs choix possibles de vallées dont le relief en E est modelé par les facteurs de transcription (BUENROSTRO et al., 2018; WADDINGTON et KACSER, 1957).

## 1.2.5 Apport des techniques d'analyses cellules uniques complémentaires au scRNA-seq

### 1.2.5.1 Traçage de lignages

Les études de scRNA-seq ont montré un amorçage précoce des CSH vers les différentes lignées avec un nuage de MPP qui tend à se diviser en plusieurs branches. Cependant, ces études peinent à caractériser de façon précise les états cellulaires transitoires cruciaux lors desquels sont prises les décisions d'amorçage vers telle ou telle lignée. La compréhension des mécanismes moléculaires internes, gouvernant les choix de la CSH dans son amorçage est encore partielle. Une population de CSH potentiellement biaisées pour une lignée identifiées par scRNA-seq peut éventuellement être triée par FACS. Puis son devenir peut être suivi après greffe dans la MO d'une souris receveuse irradiée chez qui finalement est mesurée la proportion des populations différencierées après reconstitution du système hématopoïétique (MANN et al., 2018). Cette stratégie présente de nombreux biais et n'est parfois pas réalisable (tri impossible).

Pour lever cette limite, plusieurs groupes ont développé des méthodes de traçage de lignages qui couplées au scRNA-seq permettent de suivre la destinée d'une CSH donnée et de lier son transcriptome avec sa destinée (RODRIGUEZ-FRATICELLI et CAMARGO, 2021). Ces méthodes s'appuient sur une bibliothèque de barcodes d'ADN exprimés au sein d'un transgène qui sont intégrés de façon stable dans les génomes des CSH étudiées. La descendance après plusieurs divisions d'une CSH particulière hérite ainsi du barcode qui lui est propre. L'expression de ce barcode est ensuite induite et l'ARN correspondant peut alors être identifié par scRNA-seq. Après analyse, il est possible de retracer la descendance d'une CSH dans une pseudo-trajectoire de différenciation inférée. Cette approche a permis une analyse de l'hématopoïèse en condition naturelle dans une souris (WEINREB et al., 2020) et couplée au scRNA-seq offre une vision non biaisée de la différenciation. En premier lieu, l'amorçage précoce de CSH vers différentes lignées a été conforté (RODRIGUEZ-FRATICELLI et al., 2018). Par ailleurs, une différenciation directe des CSH en mégacaryocytes a pu être observée. Une des premières études de traçage de lignage avait montré de façon surprenante qu'assez peu de cellules matures présentaient un barcode intégré au niveau des CSH (mis à part pour la lignée mégacaryocytaire), suggérant que les MPP plutôt que les CSH étaient les contributeurs principaux de l'hématopoïèse non perturbée (J. SUN et al., 2014). Ces résultats ont été confirmés par la suite et une étude plus récente couplant traçage de lignages, scRNA-seq et CRISPR-seq (voir plus loin section 1.2.5.3, page 49) a également montré que la sous population de CSH avec la plus faible contribution à l'hématopoïèse est également celle qui a le plus fort potentiel d'autorenouvellement (RODRIGUEZ-FRATICELLI et al., 2020).

En se basant sur l'hypothèse que les CSH se divisent majoritairement de façon symétrique, des expériences de traçage de lignages ont d'autre part montré que deux cellules soeurs issues d'une même division ont tendance à avoir une destinée beaucoup plus proche que deux cellules aux transcriptomes similaires (WEINREB et al., 2020). Ceci montre l'importance de combiner des approches expérimentales différentes pour étudier les mécanismes décisionnels de la destinée de la CSH.

### 1.2.5.2 Analyses multi-omiques cellules uniques

Comme nous l'avons vu la transcriptomique a été le premier champ de la génomique à être étudié à l'échelle de la cellule et a redéfini en profondeur notre vision de la CSH et de son devenir. Néanmoins, suite aux nombreuses zones d'ombres relevées notamment avec le traçage de lignages d'autres techniques d'analyses omiques se sont également développées à cette échelle avec l'espoir d'accéder aux "variables encore cachées" (protéines, épigénétique, localisation cellulaire) de la régulation de la CSH (WATCHAM et al., 2019). Le CITE-seq décrit précédemment (section 1.2.3.4, page 40) est encore en développement notamment pour le

ciblage des protéines intra-cellulaires. Il offre cependant un débit beaucoup plus important que la cytométrie en flux pour la quantification des protéines de surfaces et ouvre la voie aux analyses protéomiques en cellules uniques. Il est pour le moment plutôt utilisé pour le marquage des échantillons et types cellulaires mais est amené, via l'augmentation des débits, à être de plus en plus couplé au scRNA-seq pour l'analyse de panels de protéines d'intérêt.

Au niveau épigénétique, le ChIP-seq, qui permet de séquencer l'ADN lié à une protéine donnée (souvent des histones ou des facteurs de transcription) est assez peu développé à l'échelle de la cellule à cause de la quantité importante d'ADN nécessaire. Cette limite impacte moins l'[analyse de la chromatine accessible à la transposase par séquençage –ou Assay for Transposase-Accessible Chromatin with sequencing– \(ATAC-seq\)](#), qui mesure l'ouverture de la chromatine et qui permet notamment de connaître l'accessibilité des éléments régulateurs cis cibles de facteurs de transcription spécifiques. L'[ATAC-seq en cellules uniques –ou single-cell ATAC-seq– \(scATAC-seq\)](#) est ainsi à un stade avancé de développement. Plusieurs études intégrant le scRNA-seq et le scATAC-seq (BUENROSTRO et al., 2018; FLORIAN et al., 2018) voire en plus le CITE-seq (SOMMARIN et al., 2021) dans le cadre de l'hématopoïèse ont été réalisées. Elles ont permis en particulier de montrer que la variabilité de la chromatine aux motifs des facteurs de transcription suit les trajectoires de différenciation de la CSH. Elles apportent donc un niveau d'information supplémentaire essentiel à la compréhension des mécanismes de régulation des choix de la CSH. Couplé au scRNA-seq et à l'inférence de trajectoire le scATAC-seq pourrait permettre en théorie de relever l'activation ou la répression d'un gène par un facteur de transcription à un instant donné du processus de différenciation (section 1.3.2.3, page 68) et amène ainsi à construire un modèle de l'hématopoïèse comme un paysage de différenciation de Waddington (section 1.2.4, page 42 Figure 1.6).

Si pour le moment, scRNA-seq et CITE-seq sont réalisés dans la même cellule les études combinant scATAC-seq et scRNA-seq sont pour la plupart réalisées sur des cellules différentes. Ceci nécessite une intégration des différentes types de données (discutée section 1.2.3.4, page 41). Pour ce faire la recherche de sous ensembles de cellules communs entre les deux types de données se fait alors par une étape préalable d'estimation de l'activité des gènes selon le niveau de signal ATAC mesuré dans leur région sur l'ADN ce qui mène inévitablement à des imprécisions, notamment car les matrices de scATAC-seq sont encore plus clairsemées que celles de scRNA-seq. Cependant de véritables plateformes [multi-omiques en cellules uniques –ou single-cell multi-omics– \(sc-multi-omics\)](#) sont en développement et devraient permettre dans les prochaines années de réaliser scRNA-seq, scATAC-seq et CITE-seq dans une même cellule (CHEN et al., 2019; LIU et al., 2019).

### 1.2.5.3 Criblage génétique par CRISPR-seq

Dans le but d'étudier les réseaux de régulation génétiques soutenant le paysage de différenciation, le scRNA-seq a été couplé à l'introduction de perturbations génétiques ciblées par CRISPR/Cas9. Ce criblage génétique consiste à réprimer des dizaines de gènes au cours de la différenciation étudiée pour identifier le ou les quelques facteurs clés régissant les choix possibles de différenciation (JAITIN et al., 2016). Des cellules multipotentes receveuses sont ainsi infectées par des vecteurs viraux apportant ces modifications puis greffées dans de nouvelles souris. Le devenir des cellules modifiées et greffées est analysé par scRNA-seq et les biais de différenciation (perte de lignées) sont réattribués à la fonction d'un gène inactivé. Une étude de référence a ainsi pu mettre en évidence l'activité essentiel de *Cebpa* pour la myélopoïèse précoce et le rôle de *Spi1* et *Irf8* dans la spécification ultérieure vers les destinées granulo-monocytaires (GILADI et al., 2018). Le rôle de *Tcf-15* dans l'autorenouvellement des CSH a également pu être mis en évidence par cette approche (RODRIGUEZ-FRATICELLI et al., 2020).

## 1.2.6 Applications des technologies cellules uniques à l'étude de l'hématopoïèse en conditions perturbées

### 1.2.6.1 Étude du vieillissement de la CSH par scRNA-seq

La continuité des trajectoires de différenciation et l'absence de délimitation claire entre les différents sous-types cellulaires des HSPC ont dans leur ensemble été retrouvées et précisées dans les études cellules uniques comparant l'hématopoïèse adulte en conditions non perturbée et perturbée, plus particulièrement lors du vieillissement (Tableau 1.2). Dans ce processus de vieillissement, les études ont tout particulièrement cherché à mettre en évidence l'apparition d'une sous-population de CSH âgées au potentiel biaisé vers une différenciation myéloïde (GROVER et al., 2016; KIRSCHNER et al., 2017; MANN et al., 2018) ainsi que la perte du potentiel lymphoïde des CSH (SOMMARIN et al., 2021; YOUNG et al., 2016) qui expliqueraient le phénotype de vieillissement observé à l'échelle de l'individu. Elles ont également cherché à caractériser l'altération avec le vieillissement du cycle cellulaire de la CSH (HÉRAULT et al., 2021; KOWALCZYK et al., 2015). Bien qu'apportant de nouvelles connaissances précieuses sur le vieillissement de la CSH ces travaux ne se recoupent que partiellement (voir discussion section 3.6, page 117).

### 1.2.6.2 Applications des technologies cellules uniques à l'étude des myélopathies

Les technologies cellules uniques trouvent aussi leur application dans l'étude des différentes myélopathies notamment celles liées à l'âge comme les *Leucémies aiguës myéloïdes* –ou *Acute Myeloid Leukemias*– (AML) et les myélodysplasies (CAMPILLO-MARCOS et al., 2021).

Plusieurs sous types cellulaires leucémiques à différents stades de l'hématopoïèse (CSH, progéniteurs, cellules myéloïdes matures) et leurs implications dans la suppression de l'activité des lymphocytes T au sein du micro-environnement tumoral ont par exemple pu être mis en évidence dans le cas des AML (van GALEN et al., 2019). Le profilage mutationnel à l'échelle de la cellule (single-cell DNA-seq) a quant à lui permis d'analyser l'évolution de la clonalité des AML de cohortes de patients au cours de la progression de la maladie. De ces résultats a pu être tiré l'enchaînement probable de l'apparition des mutations affectant les facteurs épigénétiques et gènes des voies de signalisation à l'origine du développement tumoral (MILES et al., 2020; MORITA et al., 2020). L'histoire mutationnelle de quelques patients souffrant du syndrome myélodysplasique avec délétion isolée du chromosome 5q a de la même manière pu être étudiée (ACHA et al., 2021). Ces études ont souligné l'hétérogénéité tumorale inter patients de ces cancers, appelant à une caractérisation moléculaire fine de chaque malade pour le choix d'un traitement dans le cadre d'une approche de médecine personnalisée. Dans cette optique le scRNA-seq se révèle être également un outil intéressant pour étudier la toxicité d'un traitement personnalisé sur les cellules d'un patient donné (BJÖRN et al., 2020).

## 1.3 Biologie des systèmes et réseaux de régulation moléculaire

Traditionnellement, les biologistes étudient un système biologique en le décomposant en plusieurs blocs de base indépendants les uns des autres. Aujourd'hui, les technologies NGS apportent une description exhaustive des différents niveaux de régulations (génomique, épigénome, transcriptome, protéome) permettant l'étude d'un système biologique dans son ensemble. Cependant, ces données ne peuvent être considérées comme de la connaissance biologique telles quelles et un travail de modélisation doit être entrepris afin d'extraire la connaissance utile de ces différentes couches d'informations. L'approche de la biologie des systèmes peut ainsi être définie en trois étapes : Elle permet dans un premier temps d'organiser et de formaliser les observations multi-niveaux pour dans un second temps expliquer le fonctionnement du système à un haut niveau, c'est à dire en terme de comportements et/ou phénotypes cellulaires voire de l'organisme. Finalement, à partir de ces résultats la biologie des systèmes vise à prédire de nouveaux comportements du système, par exemple suite à des perturbations génétiques ou environnementales afin de proposer de nouvelles hypothèses à tester (BARILLOT et al., 2013).

### 1.3.0.1 Représentation des régulations moléculaires par un graphe d'influence

Les signaux biologiques de différenciation ou d'activation cellulaire se propagent entre les différents niveaux analysés, de la fixation du ligand sur un récepteur à la transmission du

signal par les kinases jusqu'à l'activation de la transcription de certains gènes par les facteurs de transcription. Les protéines une fois traduites auront la capacité d'agir à différents niveaux, par exemple en remodelant l'accès à la chromatine ou bien en opérant un changement de métabolisme. Ainsi les différents niveaux sont organisés en réseaux de régulation moléculaire complexes de plusieurs centaines de noeuds et d'interactions qui régissent le comportement et le devenir cellulaire. Les noeuds de ces réseaux peuvent être différentes entités moléculaires (gènes, protéines, TF, etc) qui s'influencent les uns les autres via différentes natures d'interactions (activation / repression transcriptionnelle, liaison physique, modification chimique, etc). Une approche de la biologie des systèmes consiste ainsi à placer ces différents composants et leurs interactions au même niveau sous la forme d'un **graphe d'influence** orienté et signé. Dans cette représentation, dite en flux d'activité ou *activity flows*, les étapes de la signalisation sont décrites comme des chaînes causales de relations binaires (un composant  $G_1$  inhibe/active un composant  $G_2$  figure 1.7.A; LE NOVÈRE, 2015).

### 1.3.0.2 Ressources pour la construction des graphes d'influence

Si la biologie des systèmes vise à analyser un système dans son ensemble les études "réductionnistes" menées encore jusqu'à aujourd'hui ont permis d'identifier une à une de nombreuses interactions au sein des réseaux de régulation moléculaire. Cette connaissance, enrichie depuis le début du 21<sup>ème</sup> siècle par les données de NGS, est désormais accessible en grande partie dans différentes bases de données qui sont ainsi des outils indispensables à la biologie des systèmes (BARILLOT et al., 2013). SIGNOR (LICATA et al., 2020) et KEGG PATHWAY (KANEHISA et GOTO, 2000) sont deux bases de données qui ont adopté la représentation en flux d'activité particulièrement utilisées du fait de leur curation manuelle. On peut également citer STRING un outil combinant l'analyse de données de NGS, le contexte génomique et l'exploration de bases de données et de la littérature dans le but de prédire des interactions fonctionnelles entre protéines (SZKLARCZYK et al., 2017). Concernant spécifiquement les réseaux transcriptionnels, des bases de données comme JASPAR (FORNES et al., 2020) et TRANSFAC (MATYS et al., 2003) donnent accès aux cibles potentielles des TF via l'analyse de leurs motifs. Plus récemment, on peut citer la base de données Trrust qui s'appuie sur une approche d'apprentissage automatique (*machine learning*) d'exploration du texte (*text mining*) des articles de PUBMED dans le but d'identifier des régulations transcriptionnelles ensuite vérifiées manuellement (H. HAN et al., 2018).

## 1.3.1 Modélisation des réseaux de régulation moléculaire

### 1.3.1.1 Intérêt de la modélisation

Les graphes d'influences présentent des structures complexes comme des circuits de retroaction (positif ou négatif) et des boucles d'anticipation (positive ou négative; Figure 1.7.A&B)

qui produisent une dynamique complexe nécessitant le développement de modèles mathématiques pour être analysée. Ces modèles peuvent se construire à partir du réseau de régulation en définissant pour chacun des composants une fonction mathématique décrivant son évolution au cours du temps du processus étudié, en fonction de l'influence de ces régulateurs.

Brièvement, on peut distinguer les modélisations déterministes ou non déterministes qui sont quantitatives ou qualitatives (WHICHARD et al., 2010). Tandis qu'un modèle déterministe vise à décrire la dynamique d'un réseau en l'absence de phénomène aléatoire et attribuera toujours le même résultat en sortie à des données fixées en entrée, les modèles stochastiques permettent quant à eux l'analyse de plusieurs comportements possibles du système découlant de phénomène aléatoires ou de rendre compte de l'incertitude des connaissances du système étudié. De plus, la modélisation quantitative permet l'estimation précise de la quantité de chaque composant du réseau étudié à l'aide d'équations différentielles alors qu'une modélisation qualitative simplifiera l'observation du réseau en discréétisant ces quantités sur deux (actif/inactif) ou plusieurs niveaux.

Ce type d'approche se révèle ainsi particulièrement intéressant pour la modélisation de réseaux de régulation moléculaire de taille importante puisqu'il permet de s'affranchir de l'estimation délicate des paramètres d'équations différentielles tout en donnant une vision globale de la dynamique du système étudié dans un temps de calcul raisonnable (KAUFFMAN, 1969). Ce formalisme permet, de plus, en adoptant un mode de mise à jour non déterministe (voir ci dessous) d'étudier les différentes évolutions possibles d'un état initial donné du système (R. THOMAS, 1973), comme par exemple une cellule souche se différenciant vers plusieurs types cellulaires matures. Enfin, les modèles logiques peuvent être étendus en modèles stochastiques, par exemple dans le but de modéliser la dynamique d'une population de cellules à l'aide de simulations de Monte-Carlo (STOLL et al., 2012).

Dans cette partie nous donnerons les principales notions mathématiques de la modélisation logique booléenne des réseaux de régulation moléculaire en adoptant la démarche de la biologie des systèmes. Nous présenterons d'abord le graphe d'influence des composants qui résume les régulations possibles entre ceux-ci. Ce graphe peut ensuite être paramétré pour étudier la dynamique du processus biologique étudié. Finalement, ce modèle peut être utilisé pour faire des prédictions biologiques.

### 1.3.1.2 La modélisation booléenne

**Notations.** Soient  $n \in \mathbf{N}$ ,  $[n] = \{1, \dots, n\}$ ,  $\mathbf{B} = \{0, 1\}$  et une configuration  $x \in \mathbf{B}^n$ . L'ensemble des valeurs différentes entre deux configurations  $x, y \in \mathbf{B}^n$  est notée  $\Delta(x, y) := \{i \in [n] \mid x_i \neq y_i\}$ . les connecteurs logiques seront notés  $\wedge$  pour *ET*,  $\vee$  pour *OU* et  $\neg$  pour *NON*

Le réseau de régulation génétique étudié est formé de n composants biologiques du système (gènes, protéines, complexes, etc) qui s'influencent les uns avec les autres. Ces interactions forment un graphe  $G$  de dimensions  $n$  qui est défini par un ensemble de n noeuds  $V = \{1, \dots, n\}$  liés entre eux par un ensemble d'arêtes  $E = \{(i, j); i, j \in V\}$ . Un composant peut avoir une influence positive (activation) et/ou négative (inhibition) sur un composant cible. Dans l'exemple Figure 1.7.A le composant  $G_1$  active le composant  $G_2$ . L'ensemble de ces régulations génétiques peut ainsi être représenté mathématiquement par un graphe orienté et signé appelé graphe d'influence.

**Définition 1 (Graphe d'influence)** *Un graphe d'influence est un graphe orienté signé sur  $[n]$   $G = \{E, V\}$  tel que  $V = \{1, \dots, n\}$  et  $E = \{(i, j)^s; i, j \in V; S = \{+; -\}\}$ . Un arc de  $G$  de  $i$  à  $j$  de signe  $s$  est noté  $(i, j)^s$ ,*

$G$  est simple si pour tous  $i, j \in [n]$  il existe au plus un arc de  $j$  à  $i$ . Un circuit positif (resp. négatif) est un circuit orienté contenant un nombre pair (resp. impair) d'arcs négatifs. La longueur d'un circuit est le nombre d'arcs dont il est constitué.

Dans le formalisme booléen, l'activité de chaque composant du système est représentée par une variable booléenne (0 composant inactif, 1 composant actif) et reflète sa capacité à réguler ses cibles. Une configuration  $x = \{x_1, \dots, x_n\} \in \mathbf{B}^n$  du système donne l'ensemble des niveaux d'activité des composants dans un état donné du système.

Pour étudier la dynamique du système, un modèle booléen se construit à partir du graphe d'influence et de la connaissance biologique préalable en définissant pour chaque composant  $G_i$  du système une fonction logique  $f_i : \mathbf{B}^n \rightarrow \mathbf{B}$  qui donne le niveau d'activité de  $G_i$  en fonction de celui de ces régulateurs à l'aide des connecteurs *ET* ( $\wedge$ ), *OU* ( $\vee$ ) et *NON* (! ou  $\neg$ ).

Dans l'exemple Figure 1.7.C,  $G_2$  est régulé par les activateurs  $G_1$  et  $G_3$  et l'inhibiteur  $G_2$  (auto-inhibition). Ainsi, la fonction logique  $f_2$  décrit l'évolution de la variable booléenne du composant  $G_2$  en fonction des trois variables  $x_1$ ,  $x_2$  et  $x_3$ . Par exemple on peut choisir la règle "G<sub>2</sub> est actif si ses deux activateurs sont présents et son inhibiteur absent" :

$$f_2(x) = x_1 \wedge \overline{x_2} \wedge x_3$$

Ces fonctions logiques peuvent être écrites sous la forme normale conjonctive (conjonctions de disjonctions de variables booléennes FNC) et la forme normale disjonctive (disjonctions de conjonctions de variables booléenne FND). Par exemple pour la fonction logique du composant  $G_3$  on a :

FNC :

$$f_3(x) = (x_1 \vee x_2) \wedge (x_1 \vee x_3)$$

FND :

$$f_3(x) = x_1 \vee (x_2 \wedge x_3)$$

La FND de  $f_3(x)$  présente ici deux clauses  $(x_1)$  et  $(x_2 \wedge x_3)$ .

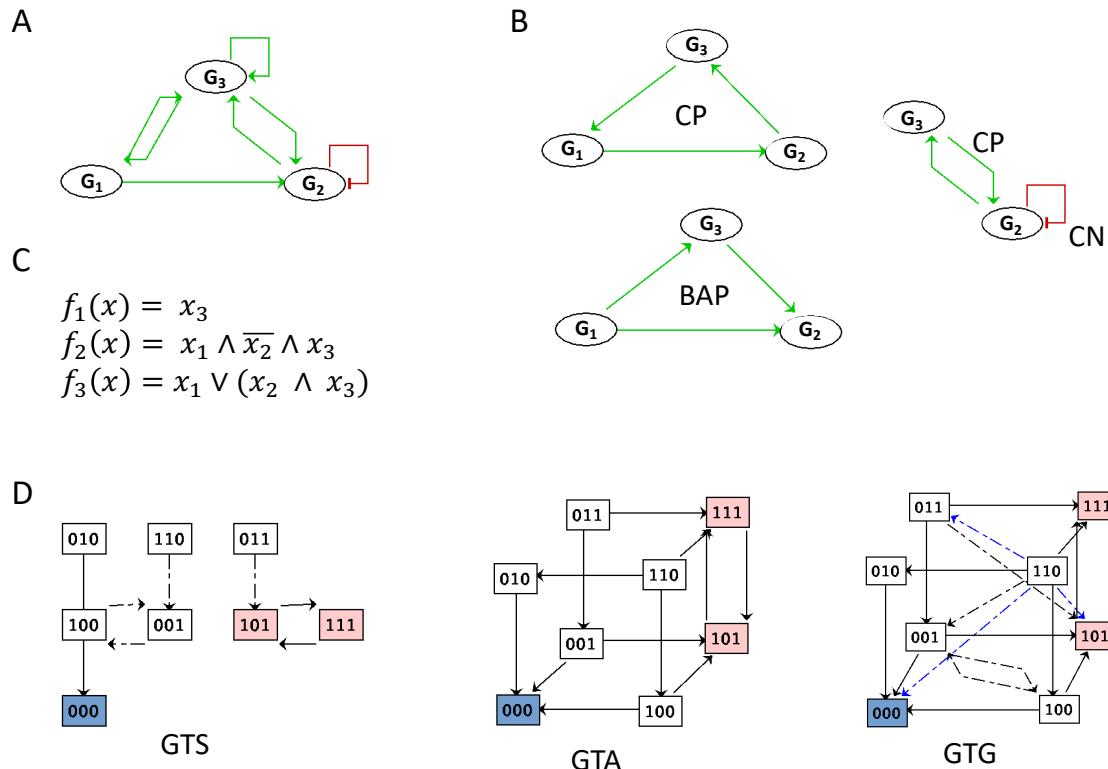


FIGURE 1.7 – Un BN possible d'un graphe d'influence et ses trajectoires synchrones, asynchrones et généralisées

A : Graphe d'influence entre 3 composants biologiques notés  $G_1, G_2, G_3$ . B : Sélection de motifs du graphe d'influence : Circuit Positif (CP), Circuit Negatif (CN), Boucle d'Anticipation Positive (BAP). C : Un BN  $f$  possible pour le graphe d'influence en A. D : Graphe de Transition Synchrone (GTS), Asynchrone (GTA) et Généralisée (GTG) du BN  $f$  défini en C. Les noeuds sont les configurations du BN  $\{x_1, x_2, x_3\}$  dans un état actif (1) ou inactif (0). Les transitions synchrones sont représentées par des arcs tiret-points, les transitions asynchrones par des arcs pleins et les transitions uniquement généralisées par des arcs tiret-point bleus. La configuration  $\{0, 0, 0\}$  colorée en bleu est un point fixe de  $f$ . Les configurations  $\{1, 0, 1\}$  et  $\{1, 0, 1\}$  font partie d'un attracteur cyclique de  $f$  pour les 3 sémantiques synchrone et asynchrone et généralisée.

Étant donné  $n$  variables booléennes  $\{x_1, \dots, x_n\}$ , l'ensemble des  $f_i$  pour  $i \in [n]$  définit un réseau booléen –ou *Boolean Network*– (BN)  $f$  :

**Définition 2 (Réseau Booléen (Boolean Network BN))** *Un BN de dimension  $n$  est une fonction  $f = \{f_1, \dots, f_n\} : \mathbf{B}^n \rightarrow \mathbf{B}^n$ . Pour tout  $i \in [n]$ ,  $f_i : \mathbf{B}^n \rightarrow \mathbf{B}$  est la fonction locale du  $i^{\text{ème}}$  composant. Le vecteur  $x \in \mathbf{B}^n$  est appelé une configuration du BN. Un BN est dit localement monotone*

*lorsque pour chacun de ses composants chaque régulateur est soit activateur soit inhibiteur (mais ne peuvent pas avoir un effet dual).*

### 1.3.1.3 Sémantiques des réseaux booléens

Étant donné une configuration du BN, plusieurs composants peuvent être appelés pour mettre à jour leur niveau par les fonctions logiques (définies dans l'ensemble  $\Delta(x, f(x))$ ). Différentes sémantiques de mise à jour  $\sigma$  sont utilisées pour définir les transitions possibles entre deux configurations  $x$  et  $y$  d'un BN  $f$  notées  $x \xrightarrow[\sigma]{f} y$  et ainsi décrire les évolutions des composants d'un système dynamique discret. Les plus connues sont les sémantiques synchrone, asynchrone et généralisée :

**Définition 3 (sémantiques synchrone, asynchrone et généralisée)** Pour tout  $x, y \in \mathbf{B}^n$  deux configurations d'un BN  $f$  :

- $x \xrightarrow[s]{f} y$  est une transition synchrone de  $x$  vers  $y$  si et seulement si  $x \neq y$  et  $y = f(x)$ .
- $x \xrightarrow[a]{f} y$  est une transition asynchrone de  $x$  vers  $y$  si et seulement si  $\exists i \in [n]$  tel que  $\Delta(x, y) = \{i\}$  et  $y_i = f_i(x)$ .
- $x \xrightarrow[g]{f} y$  est une transition généralisée de  $x$  vers  $y$  si et seulement si  $x \neq y$  et  $\forall i \in \Delta(x, y), y_i = f_i(x)$ .

On peut alors définir  $\rho_\sigma^f(x)$  l'ensemble des configurations accessibles depuis une configuration  $x$  d'un BN  $f$  avec la sémantique  $\sigma$ . Ceci permet par exemple de définir quels sont les états cellulaires oligopotents qui peuvent être atteints par un progéniteur multipotent.

**Définition 4 (Accessibilité)** Soient deux configurations  $x, y \in \mathbf{B}^n$  d'un BN  $f$ ,  $y$  est accessible depuis  $x$  dans la sémantique  $\sigma$  si  $x = y$  ou s'il existe une séquence de transitions de  $x$  à  $y$ . L'ensemble des configurations de  $f$  accessibles depuis  $x$  est  $\rho_\sigma^f(x) := \{y \in \mathbf{B}^n | x = y \vee x \xrightarrow[\sigma]{f} \dots \xrightarrow[\sigma]{f} y\}$

Les trajectoires des sémantiques synchrone, asynchrone et généralisée peuvent être décrites par un graphe de transition  $GT$  dont les noeuds sont l'ensemble des configurations  $\mathbf{B}^n$  et les arcs  $x \rightarrow y$  sont les transitions possibles  $x \xrightarrow[\sigma]{f} y$  entre les configurations du BN représentées par les noeuds du graphe. Les configurations accessibles depuis une configuration  $x$  sont alors identifiables par les noeuds sur les chemins partant de  $x$  dans le graphe de transition  $GT$  (Figure 1.7.D).

**Définition 5 (Graphes de transitions)** Soit  $f : \mathbf{B}^n \rightarrow \mathbf{B}^n$  un BN,  $x, y \in \mathbf{B}^n$  deux configurations de  $f$ .

Le graphe de transition de  $f$  dans la sémantique  $\sigma$ ,  $GT(f)$  est un graphe orienté dont l'ensemble des sommets est  $\mathbf{B}^n$  et l'ensemble des arcs  $E = \{(x, y); x, y \in \mathbf{B}^n\}$  tel que  $(x, y) \in E \Leftrightarrow x \xrightarrow[\sigma]{f} y$ .

- Le graphe de transition synchrone (GTS) de  $f$ ,  $GTS(f)$  est le graphe orienté sur  $\mathbf{B}^n$  qui contient un arc  $x \rightarrow y$  si et seulement si  $x \xrightarrow[s]{f} y$ .
- Le graphe de transition asynchrone (GTA) de  $f$ ,  $GTA(f)$  est le graphe orienté sur  $\mathbf{B}^n$  qui contient un arc  $x \rightarrow y$  si et seulement si  $x \xrightarrow[a]{f} y$ .
- Le graphe de transition généralisé (GTG) de  $f$ ,  $GTG(f)$  est le graphe orienté sur  $\mathbf{B}^n$  qui contient un arc  $x \rightarrow y$  si et seulement si  $x \xrightarrow[g]{f} y$ .

La sémantique synchrone déterministe attribue au plus une transition pour chaque configuration. Biologiquement, il est considéré comme peu probable que tous les composants du système changent de niveaux en même temps et il apparaît plus raisonnable d'étudier la dynamique du système avec la sémantique asynchrone qui décrit plusieurs transitions à partir d'une configuration (R. THOMAS, 1973). Ce non déterminisme convient bien pour la modélisation d'une différenciation cellulaire à partir d'une configuration initiale souche du système qui peut prendre plusieurs voies de différenciation. La sémantique généralisée quant à elle englobe les transitions synchrones et asynchrones en plus d'autres transitions en permettant la mise à jour d'un nombre indéterminé de composants par itérations.

L'exemple Figure 1.5.D illustre ces trois types de sémantiques et leur représentation graphique pour l'exemple de BN présenté. On peut remarquer que les GTA et GTG sont connexes (quelques soient les sommets  $u, v$  du GT il existe au moins un chemin non orienté reliant  $u$  à  $v$ ) contrairement au GTS.

L'étude de la dynamique des BN porte en premier lieu sur le comportement asymptotique du système qui au gré des transitions à partir d'une configuration finit par atteindre des sous ensembles de configurations desquels il ne peut sortir appelés attracteurs :

**Définition 6 (Attracteurs)** *Un sous ensemble  $A \subseteq \mathbf{B}^n$  est un attracteur du BN  $f$  avec la sémantique  $\sigma$  si pour tout  $x \in A$ ,  $\rho_\sigma^f(x) = A$ . Si  $A = \{x\}$  pour  $x \in \mathbf{B}^n$ ,  $x$  est appelé point fixe de la dynamique. Les attracteurs de dimension supérieure ou égale à 2 sont appelés attracteurs cycliques.*

Dans un graphe de transition  $GT(f)$ , les attracteurs sont les composantes fortement connexes terminales (plus petits sous ensembles non nuls de noeuds sans arc sortant) et un point fixe est un noeud sans aucun arc sortant. Dans l'exemple Figure 1.7.D, les trois GT présentent un point fixe commun  $\{000\}$  et un attracteur cyclique  $\{111,101\}$ .

On peut noter que pour la plupart des sémantiques, dont celles considérées dans ce manuscrit, les points fixes de la dynamique correspondent exactement aux points fixes de  $f$  ( $\{x\}$  est un attracteur si et seulement si  $x = f(x)$ ).

Biologiquement les attracteurs sont souvent interprétés comme des comportements ou états cellulaires stables, par exemple des types cellulaires différenciés atteints à partir d'une configuration initiale de cellule indifférenciée.

#### 1.3.1.4 Lien entre le graphe d'influence et la dynamique du BN

La fonction  $f$  définissant un BN contient toute l'information nécessaire pour l'étude de sa dynamique. Cependant comme la taille de l'espace des configurations croît de façon exponentielle avec le nombre de composants, l'étude de la dynamique directement avec  $f$  peut s'avérer difficile voir impossible pour des BN de tailles importantes. Dans cette situation, certaines propriétés de la dynamique du système peuvent néanmoins être déduites de l'analyse statique du BN à partir de son graphe d'influence. On peut noter qu'à un BN  $f$  correspond un unique graphe d'influence mais que plusieurs BN peuvent être définis à partir d'un graphe d'influence. Ces analyses ne peuvent donc apporter qu'une information partielle sur la dynamique.

Les circuits des graphes d'influence sont assez tôt apparus comme des responsables potentiels des comportements asymptotiques du système. En 1981, René Thomas formalisa sous forme de règles (conjectures) le rôle des circuits positifs et négatifs au sein dans la dynamique des systèmes biologiques : Les premiers étant nécessaires à la multistabilité d'un système, les seconds permettant des oscillations entretenues du système entre différentes configurations (R. THOMAS, 1981). Ces théorèmes ont par la suite été démontrés dans la sémantique asynchrone dans différents formalismes, dont le formalisme logique (REMY et al., 2008) :

Soit  $f : \mathbf{B}^n \rightarrow \mathbf{B}^n$  un BN,  $T$  son graphe de transition et  $G$  son graphe d'influence.

**Théorème 1** *Si  $G$  n'a pas de circuit négatif alors  $f$  n'a pas d'attracteur cyclique et donc  $f$  a au moins un point fixe.*

**Théorème 2** *Si  $G$  n'a pas de circuit positif alors  $f$  a au plus un point fixe.*

Dans le cas d'une mise à jour asynchrone ce théorème liant multistabilité et circuit positif est précisé pour montrer la nécessité d'un circuit positif pour la présence de plusieurs attracteurs (RICHARD et COMET, 2007) :

**Théorème 3** *Si  $G$  n'a pas de circuit positif alors  $f$  à au plus un attracteur dans la sémantique asynchrone.*

D'autres propriétés de la dynamique des BN peuvent être établies à partir de l'analyse de leur graphe d'influence (PAULEVÉ et RICHARD, 2012). Ces propriétés trouvent leur application dans les méthodes de réduction des BN par suppression de certains de leurs composants qui permettent l'analyse de BN de tailles importantes.

### 1.3.1.5 Réseaux Booléens les plus permissifs

D'après R. Thomas la sémantique asynchrone représente une bonne abstraction pour modéliser des systèmes biologiques (R. THOMAS, 1973) et elle est aujourd'hui communément utilisée dans les études de modélisation de systèmes biologiques. Cependant elle peut introduire des comportements trompeurs et échouer à en expliquer d'autres au vue des observations biologiques (PAULEVÉ et al., 2020). Récemment une nouvelle sémantique des BN à ainsi été développée qui garantit de n'exclure aucun comportement du système réalisable dans tout raffinement quantitatif du modèle. Cette [sémantique la plus permissive des BN – ou Most Permissive \(MP\) semantic of BN – \(MP\)](#) fournit une sur-approximation formelle de la dynamique et tout comportement qu'elle prédit est réalisable par un raffinement quantitatif du BN en utilisant la sémantique asynchrone. La sémantique MP permet également de réduire significativement la complexité des analyses de recherches attracteurs et d'accessibilité en évitant le problème d'explosion combinatoire de l'espace des configurations avec la taille du BN. Dans cette partie, nous allons décrire brièvement cette sémantique sur laquelle repose la méthode d'inférence de BN présentée section 4.1.2 qui a été utilisée dans le résultat 4.3.

Le niveau d'activité nécessaire pour l'influence d'un composant peut varier selon ses cibles ce qui n'est pas pris en compte par la sémantique asynchrone en formalisme booléen. Lorsque ces différents niveaux sont connus un modèle logique multi-valué peut être construit. Cependant les seuils véritables d'influence sont souvent ignorés pour la plupart des composants. Ainsi, pour tenir compte de cette incertitude la sémantique MP ajoute à chaque composant, deux états dynamiques  $\nearrow$  et  $\searrow$ , aux états booléens actifs et inactifs. lorsqu'il est dans un état dynamique, un composant peut être au dessus du seuil d'influence pour une de ses cibles mais en dessous pour une autre et peut ainsi aussi bien être considéré actif qu'inactif pour la mise à jour MP.

Formellement, la sémantique MP attribue à chaque composant du BN un parmi quatre états notés  $\mathbf{P} \in \{0, \nearrow, \searrow, 1\}$ . Les interprétations binaires possibles d'une configuration  $x \in \mathbf{P}^n$  sont notés  $\gamma(x) := \{z \in \mathbf{B}^n | \forall i \in [n], x_i \in \mathbf{B} \Rightarrow z_i = x_i\}$ .

Un composant inactif à 0 (resp actif à 1) peut passer dans l'état dynamique  $\nearrow$  (resp  $\searrow$ ) chaque fois qu'il peut interpréter la valeur de ses régulateurs d'une manière qui rend sa fonction logique  $f_i$  vraie (resp. fausse) :

**Définition 7 (Sémantique MP des BN)** Soient  $x, y \in \mathbf{P}^n$ ,  $x \xrightarrow[\text{mp}]{f} y$  est une transition MP d'un

BN f de x vers y si et seulement si  $\exists i \in [n] : \Delta(x, y) = \{i\}$

$$\text{et } y_i = \begin{cases} \nearrow \text{ si } x_i \neq 1 \text{ et } \exists z \in \gamma(x) \text{ telle que } f_i(z) = 1 \\ 1 \text{ si } x_i = \nearrow \\ \searrow \text{ si } x_i \neq 0 \text{ et } \exists z \in \gamma(x) \text{ telle que } f_i(z) = 0 \\ 0 \text{ si } x_i = \searrow \end{cases}$$

la Figure 1.8 illustre cette sémantique et donne un exemple d'exécution pour l'exemple de BN précédemment étudié. A partir de la configuration  $\{0, 1, 1\}$ ,  $G_2$  s'auto-inhibe jusqu'à un seuil qui n'est plus capable d'activer  $G_3$ .  $G_3$  décroît vers 0 en étant encore capable d'activer  $G_1$ . Ce dernier croît alors vers 1 ce qui réactive  $G_2$  qui atteint 1.  $G_3$  atteint 0 avant que  $G_1$  n'atteigne 1 ce qui pousse  $G_1$  à décroître vers 0. Le système est ainsi passé de la configuration  $\{0, 1, 1\}$  à  $\{0, 1, 0\}$ , une transition impossible avec la sémantique généralisée (et donc aussi avec les sémantiques synchrone et asynchrone).

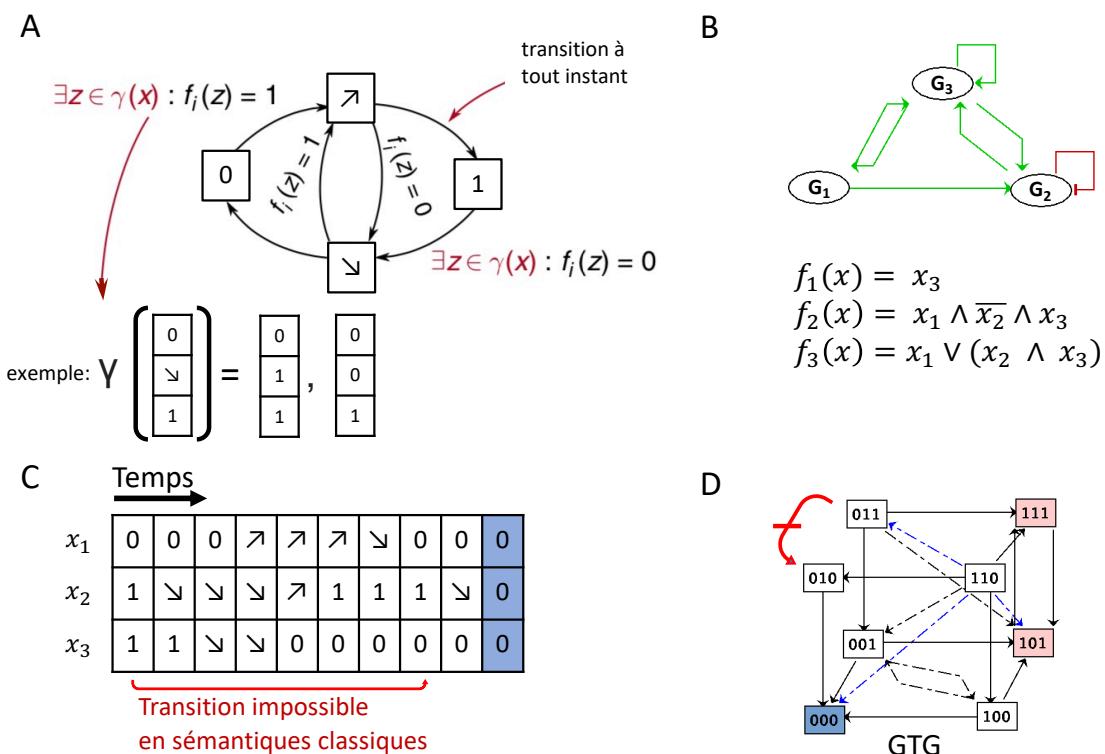


FIGURE 1.8 – Sémantique MP des BN

A : Illustration de la sémantique MP des BN. B : Précédent exemple de BN f et son graphe d'influence. C : Exemple d'exécution possible du BN f avec la sémantique MP. D : Graphe de Transition Généralisée (GTG) du BN f. Les noeuds sont les configurations du BN  $\{x_1, x_2, x_3\}$  dans un état actif (1) ou inactif (0). Les transitions synchrones sont représentées par des arcs tiret-points, les transitions asynchrones par des arcs pleins et les transitions uniquement généralisées par des arcs tiret-points bleus. La configuration  $\{0, 0, 0\}$  colorée en bleu est un point fixe de f. Les configurations  $\{1, 0, 1\}$  et  $\{1, 0, 1\}$  font partie de l'attracteur cyclique de f. En rouge la transition impossible de  $\{0, 1, 1\}$  vers  $\{0, 1, 0\}$

On peut relever les liens suivants entre sémantique MP et asynchrone (CHEVALIER et al., 2020) :

**Propriété 1 (Liens entre sémantique MP et asynchrone)**

Soient  $x, y \in \mathbf{B}^n$  deux configurations d'un BN  $f$ .

- $x \in \mathbf{B}^n$  est un point fixe de  $f$  en sémantique MP si et seulement si il est point fixe en sémantique asynchrone ( $x = f(x)$ ).
- $y \in \mathbf{B}^n$  est accessible depuis  $x \in \mathbf{B}^n$  avec la sémantique asynchrone seulement si il est accessible avec la sémantique MP ( $\rho_a^f(x) \subseteq \rho_{mp}^f(x)$ ).
- Le nombre d'attracteurs en sémantique MP est égal ou plus petit que le nombre d'attracteurs avec la sémantique asynchrone.

Les questions d'accessibilité entre deux configurations d'un BN et d'existence d'attracteurs d'un BN sont les deux principales propriétés dynamiques utilisées dans l'analyse de modèles logiques de système biologiques. Ces deux problèmes sont de complexité PSPACE complet avec les sémantiques classiques des BN. La sémantique MP simplifie grandement la complexité de ces deux analyses l'accessibilité devenant un problème P pour des BN localement monotone (ou  $P^{NP}$  sinon) et l'appartenance d'une configuration à un attracteur devenant coNP pour des BN localement monotones (ou  $coNP^{coNP}$  sinon; PAULEVÉ et al., 2020).

### 1.3.1.6 Analyses *in silico* et perturbations des BN de systèmes biologiques

En adressant les questions d'attracteurs et d'accessibilités les études *in silico* de modélisation logique d'un réseau de régulation moléculaire visent en premier lieu à tester l'aptitude du BN construit à décrire les observations expérimentales dans un contexte donné. Des processus biologiques primordiaux comme le cycle cellulaire, la réparation de l'ADN et la sénescence chez les mammifères notamment ont été modélisés avec succès par un formalisme logique (FAURÉ et al., 2006; MOMBACH et al., 2014; VERLINGUE et al., 2016). Des études se sont également penchées sur des contextes plus spécifiques comme la signalisation MAPK et son rôle dans le choix de la cellule entre prolifération et apoptose dans le contexte du cancer de la vessie (GRIECO et al., 2013).

Si les résultats du modèle sont cohérents avec les observations disponibles dans un contexte donné, l'approche de la biologie des systèmes consiste ensuite à prédire *in silico* le comportement du système dans de nouveaux contextes. Ceci est réalisé classiquement par l'étude de trajectoires de la dynamique à partir de nouvelles configurations en entrée (changement de la combinaison d'input de signaux extra-cellulaires par exemple) ou bien en perturbant le modèle par l'altération des règles logiques (fonction  $f_i$  de chaque composant). Ces altérations des règles permettent ainsi de modéliser des mutations d'expression ectopiques (Knock In KI) ou au contraire d'inactivation totale d'un gène (Knock Out KO) ou bien des mutations

marginales qui n'affectent qu'en partie la régulation d'un gène (par exemple la perte d'influence d'un des régulateurs). L'exemple ci dessous illustre ces trois types de mutations pour le composant  $G_2$  du BN donné en exemple Figure 1.7.

- $f_2(x) = 1$  est la mutation KI de  $G_2$ .
  - $f_2(x) = 0$  est la mutation KO de  $G_2$ .
  - $f_2(x) = x_3 \wedge x_1$  est une mutation marginale de  $G_2$  (perte de l'auto-inhibition).

Les conséquences de ces 3 mutations sur la dynamique asynchrone de  $f$  sont données Figure 1.9. Dans les trois cas, la suppression de l'auto-inhibition supprime le circuit négatif à l'origine de l'attracteur cyclique en condition normale. On peut également remarquer que chaque mutation aboutit à des ensembles de points fixes différents.

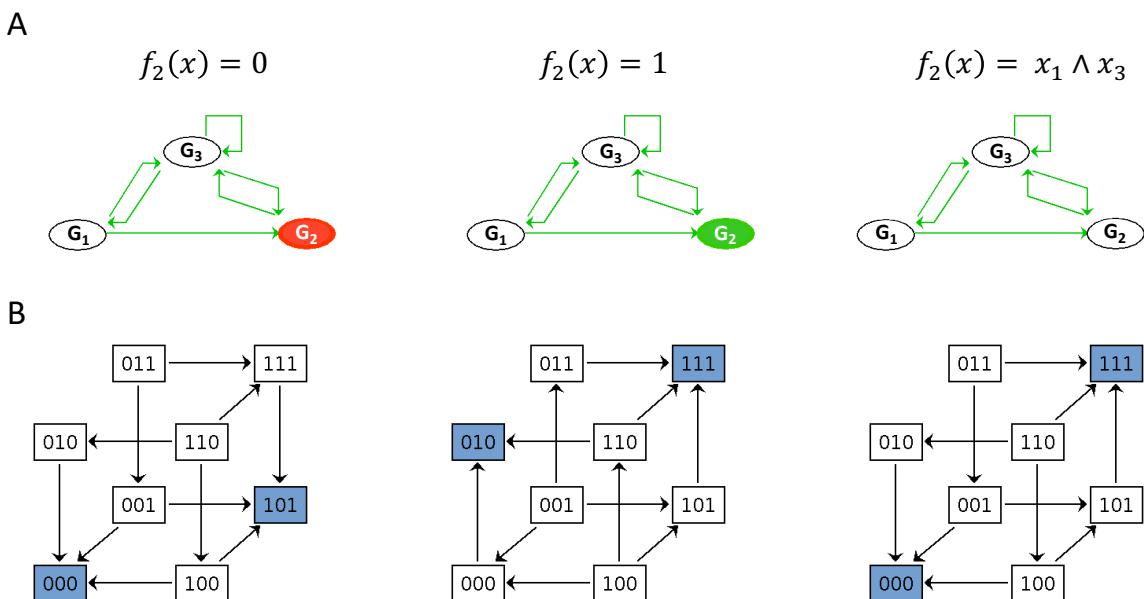


FIGURE 1.9 – Perturbations des BN

A : Trois perturbations du composant  $G_2$  ( $f_2 = x_2 \wedge \overline{x_2} \wedge x_3$ ) dans le BN  $f$  présenté Figure 1.7 : KI ( $f_2(x) = 1$ ), KO ( $f_2(x) = 0$ ) et perte de l'auto-inhibition ( $f_2(x) = x_3 \wedge x_1$ ). B : Graphes de Transition Asynchrones (GTA) du BN  $f$  découlant des perturbations. Les noeuds sont les configurations du BN  $\{x_1, x_2, x_3\}$  dans un état actif (1) ou inactif (0), les transition asynchrones sont les arcs. Les configurations colorées en bleu sont des points fixes du BN  $f$  perturbé.

Lorsque ces mutations ont été testées expérimentalement auparavant, leurs analyses participent à l'étape de validation du modèle comme par exemple avec l'étude *in silico* du KI EGFR conjoitement à un KO de P53 dans le modèle MAPK conduisant à un unique attracteur caractérisant un état cellulaire de prolifération en accord avec ce qui est observé dans les cancers de la vessie les plus agressifs (GRIECO et al., 2013). Ces analyses de mutations peuvent

aussi être utilisées pour prédire le comportement de systèmes dans des conditions qui n'ont pas encore été testées expérimentalement. Des synergies de drogues antiprolifératives sur des lignées cellulaires de cancers gastriques ont par exemple été mises en évidence par cette approche (FLOBAK et al., 2015).

Souvent le premier modèle construit ne décrit pas parfaitement le système étudié, certaines trajectoires ou attracteurs peuvent s'avérer manquant ou en contradictions avec les données expérimentales. Un travail itératif peut alors être entrepris qui consiste à partir des résultats incohérents du modèle et faire des hypothèses sur des régulations et/ou des noeuds manquants qui sont testés expérimentalement avant d'être ajoutés au modèle qui peut ensuite à nouveau être enrichi de cette façon jusqu'à ce que son comportement soit au plus en accord avec les observations expérimentales. De cette façon a été suggérée la régulation négative de *Cebpa* par Foxo1 durant la spécification lymphoïde (COLLOMBET et al., 2017).

De nombreux outils bioinformatiques ont été développés pour réaliser ces analyses *in silico*. On peut citer en premier lieu GINSim qui permet la construction, la simulation, la perturbation et l'analyse de BN (NALDI et al., 2009). Cet outil propose en plus d'un package python une interface graphique accessible par les biologistes, et ses développeurs entretiennent une base de données de BN de systèmes biologiques régulièrement mise à jour. Plusieurs packages en R comme BoolNet (MÜSSEL et al., 2010) et en python comme biolqm (NALDI, 2018) ou PINT (PAULEVÉ, 2017) sont également disponibles. Par ailleurs un effort important a été réalisé ces dernières années pour uniformiser le panel d'outils disponibles pour l'étude de BN sous la forme d'un carnet de notes interactif en python (NALDI et al., 2018). Ce cadre permet d'éditer, d'exécuter, de partager et de reproduire des analyses de modèles qualitatifs de réseaux de régulation moléculaire. Il permet notamment l'accès aux outils MaBoSS pour réaliser des simulations stochastiques de BN (STOLL et al., 2017) et mpbn pour les étudier dans la sémantique MP (PAULEVÉ et al., 2020).

### 1.3.1.7 Applications au système hématopoïétique

Du fait de la facilité à récolter des échantillons de sang ou de moelle osseuse, le système hématopoïétique a été très tôt sujet à diverses études de modélisations mathématiques (WHICHARD et al., 2010). Dans un premier temps ce sont d'avantage des études au niveau de la production des différentes populations cellulaires sanguines et leurs interactions qui ont été modélisées à l'aide d'équations différentielles (ARINO et KIMMEL, 1986). L'arrivée du NGS a ensuite facilité grandement l'analyse des réseaux de régulation moléculaire et de nombreux modèles logiques en lien avec l'hématopoïèse ont été établis, notamment sur la différenciation des lymphocytes T (CACACE et al., 2020; MENDOZA, 2006; NALDI et al.,

2010) et leur activation (RODRÍGUEZ-JORGE et al., 2019) dans le but de mieux comprendre les cancers résistants aux thérapies immunes (anti-PD1, anti-CTL4; KONDRAHOVA et al., 2020). Plus proche du sujet de cette thèse un certain nombre de modèles ont été développés pour décrire la différenciation et le comportement de HSPC.

Krumsiek et ses collègues ont proposé un modèle logique de différenciation des CMP vers les progéniteurs érythrocytaires, mégacaryocytaires, granulocytaires et monocytaire représentés par 4 points fixes du BN (KRUMSIEK et al., 2011). La dynamique de ce modèle récapitule également la hiérarchie de différenciation connue (voir Figure 1.1) avec la présence d'un premier embranchement entre GMP et MEP dans le graphe de transition asynchrone. Plus en amont dans l'hématopoïèse encore, un modèle logique de la régulation de la maintenance de l'état HSPC a été établi (BONZANNI et al., 2013). Ce modèle présente un attracteur cyclique interprété comme le reflet de l'hétérogénéité de l'état HSPC observable dans les données d'expressions cellules uniques (12 cellules à l'époque). Les auteurs suggèrent que ce réseau très connecté de TF en configuration HSPC doit être perturbé par des signaux extérieurs pour atteindre des états différenciés. Leurs simulations montrent par exemple que l'activation extérieure transiente de *Gata1*, quand le réseau est dans l'attracteur cyclique, permet un échappement vers un point fixe correspondant à un érythrocyte. Le modèle de Collombet et al, quant à lui décrit une différenciation partant de configurations MPP qui peuvent atteindre 2 points fixes correspondants aux CLP, GMP pouvant ensuite sous l'effet de stimulations cytokiniques atteindre deux nouveaux points fixes correspondants respectivement aux cellules pré-B, et aux Macrophages (COLLOMBET et al., 2017). À partir de ce résultat les auteurs décrivent également la reprogrammation de cellules pré-B en macrophages par activation transiente de *Cebpa*.

D'autres études se sont concentrées pour leur part à construire des modèles logiques décrivant le comportement prolifératif ou quiescent de la CSH en réponse aux signaux émis par la niche. Une modélisation du dialogue entre la CSH et la **Cellule Stromale Mésenchimale (CSM)** a par exemple été proposée (ENCISO et al., 2016). Les points fixes du BN construit correspondent à l'état attaché et détaché de la CSH à la CSM. Cette étude suggère que l'expression aberrante de NF-kB induite par des facteurs intrinsèques ou extrinsèques à la CSH peut contribuer à créer un microenvironnement tumoral. Les modèles évoqués jusqu'alors ont tous été étudiés avec une sémantique asynchrone (celui de Collombet également à l'aide de simulations stochastiques) mais on peut également relever une modélisation récente d'un réseau de régulation moléculaire régissant la quiescence et le réveil de la CSH étudiée avec une sémantique synchrone (IKONOMI et al., 2020). Dans ce modèle, le réseau peut transiter entre les configurations stables LTHSC, STHSC et CSH proliférantes selon la combinaison d'input de signaux de la niche promouvant la quiescence ou l'activation du cycle cellulaire. Les auteurs expliquent ces transitions par un nouveau mécanisme de régulation de P53 impli-

quant des régulateurs des ROS et des TF activés par RAS.

Si certains des modèles évoqués ont été utilisés pour étudier des mécanismes tumoraux (ENCISO et al., 2016; KONDRATOVA et al., 2020), aucun des modèles évoqués n'a été utilisé pour l'étude du vieillissement du système hématopoïétique avec notamment la perte de capacité de la CSH à se différencier vers les lignées lymphoïdes. Par ailleurs des études de modélisations du système hématopoïétique ont été mené en partant de l'analyses de données d'expression single-cell pour la construction du graphe d'influence et l'inférence de règles logiques définissant des BN. Des ensembles de BN pour le développement embryonnaire du système hématopoïétique (MOIGNARD et al., 2015) et pour la différenciation de la CSH vers les LMPP et les MEP (HAMEY et al., 2017) ont pu être construits de cette façon. Dans la partie suivante, nous allons donner un aperçu de ces différentes méthodes d'inférence de graphes d'influences et de modèles logiques à partir des données d'expression à l'échelle de la cellule.

## 1.3.2 Inférence de réseaux de régulation moléculaire à partir de données cellules uniques

### 1.3.2.1 Diversité des méthodes mathématiques

Le scRNA-seq a rendu possible l'étude de nouveaux types cellulaires et de trajectoires de différenciation. L'étape suivante consiste alors à élucider le réseau de régulation moléculaire qui contrôle la transition d'une cellule d'un type cellulaire vers un autre notamment au cours de la différenciation. Cette question avait déjà abordée à partir des données d'expression en vrac (*bulk*) (MARBACH et al., 2012), mais aujourd'hui la quantité d'information apportée par les données cellules uniques améliore grandement la qualité de l'inférence des interactions entre les composants biologiques de ces réseaux, tout particulièrement pour les régulations entre les TF et leurs gènes cibles en s'appuyant sur les dépendances d'expression observées dans les données (HU et al., 2020). De plus, les constructions de pseudo-trajectoires de différenciation ordonnent dans le temps ces données, rendant possible la construction de modèles mathématiques pour étudier la dynamique de ces réseaux, notamment à l'aide de BN comme nous venons de le voir.

De nombreuses méthodes mathématiques ont été proposées dans ce but (HU et al., 2020) et évaluées (PRATAPA et al., 2020) s'appuyant sur des équations différentielles (AUBIN-FRANKOWSKI et VERT, 2020; MATSUMOTO et al., 2017), des approches de régression (HUYNH-THU et al., 2010; PAPILI GAO et al., 2018), les corrélations d'expression (SPECHT et LI, 2017), la théorie de l'information (CHAN et al., 2017; VERNY et al., 2017) et les réseaux booléens (MOIGNARD et al., 2015). Cette diversité des méthodes se répercute sur leur pré-requis : nécessité de données temporellement ordonnées ou non, sur une trajectoire linéaire ou pouvant bifurquer; leurs hypothèses : relation linéaire ou non entre les expressions du régulateur et de sa cible, faible

nombre de régulateurs pour un composant ou non; leur résultat : modèle dynamique ou réseau de régulation statique orienté signé (graphe d'influence) ou non; et leur capacité en matière de taille du réseau inféré et temps de calcul.

Les méthodes qui ont besoin d'un pseudotemps incluent notamment celle s'appuyant sur des modèles d'[équations différentielles ordinaires –ou Ordinary Differential Equations– \(ODE\)](#) généralement de la forme :

$$\frac{dy}{dt} = f(x),$$

où  $x$  représente l'expression de TF ( $x_1, x_2, \dots, x_n$ ) et  $y$  l'expression du gène cible qu'ils régulent,  $x$  et  $y$  étant liés aux séries temporelles de temps (ici pseudotemps)  $t$ . Bien que ces méthodes semblent à première vue idéales en aboutissant directement à l'obtention d'un graphe d'influence et de sa dynamique, une évaluation récente sur des données scRNA-seq simulées à partir de réseaux de la littérature ou synthétiques a mis en évidence une plus faible précision des méthodes nécessitant des cellules pseudo-ordonnées dont celles utilisant des [ODE](#) (PRATAPA et al., 2020).

Par ailleurs, étant donné l'intérêt de la modélisation logique pour les systèmes biologiques discuté précédemment (voir section 1.3.1.1 page 51) et mise en oeuvre dans cette thèse (voir section 4.3 page 125), nous ne détaillerons pas les approches d'inférence de modèles d'[ODE](#) dans cette introduction et nous discuterons dans la partie suivante plus précisément des méthodes permettant l'inférence de BN à partir de données cellules uniques en deux étapes. Tout d'abord, nous aborderons les méthodes qui, à partir des données cellules uniques permettent d'inférer un graphe d'influence, particulièrement celles utilisant des arbres de décisions (AIBAR et al., 2017; HUYNH-THU et al., 2010). Nous discuterons ensuite des méthodes permettant d'inférer des règles logiques pour paramétriser ce graphe, selon la dynamique observée dans les données cellules uniques, et obtenir finalement un BN (CHEVALIER et al., 2019; LIM et al., 2016; MOIGNARD et al., 2015).

### 1.3.2.2 Inférence de graphe d'influence à partir de matrices d'expressions de données cellules uniques

**Corrélations.** La corrélation de Pearson  $\rho_{x1,x2}$  est la statistique la plus simple pour caractériser l'association des expressions de deux gènes  $X1$  et  $X2$  :

$$\rho_{X1,X2} = \frac{cov(X1, X2)}{\sigma_{X1}\sigma_{X2}},$$

où  $\sigma_X$  est la variance de la [variable aléatoire \(v.a.\)](#)  $X$  (ici l'expression du gène  $X1$  ou  $X2$ ) et  $cov(X1, X2)$  la covariance entre les [v.a.](#)  $X1$  et  $X2$ . Cette statistique facilement calculable est cependant trop naïve pour rendre compte de régulations complexes au sein du réseau car elle mesure une dépendance linéaire et ne permet pas de distinguer les dépendances directes

des dépendances indirectes.  $X_1$  et  $X_2$  peuvent en effet être fortement corrélées aussi bien dans le cas où  $X_1$  interagit avec  $X_2$  que dans le cas où  $X_1$  interagit avec un gène  $X_3$  qui lui agit directement sur  $X_2$ . Pour y remédier, il est possible de calculer les corrélations partielles qui suppriment l'effet des autres gènes en assumant des relations linéaires entre ceux ci (LAWRANCE, 1976).

**Information mutuelle.** Cette hypothèse de linéarité est loin d'être vérifiée dans les régulations biologiques. Des méthodes ont ainsi été développées en s'appuyant sur l'information mutuelle qui quantifie la dépendance entre deux v.a. à partir de leur distribution de probabilité et permet ainsi de quantifier des relations non linéaires (COVER et THOMAS, 1991). Via une discréétisation des données et une estimation des distributions de probabilité de l'expression des gènes, l'information mutuelle permet de mesurer la réduction de l'incertitude d'une v.a.  $X$  quand une autre v.a.  $Y$  est connue. Différentes mesures multivariées de l'information peuvent également être définies pour quantifier la dépendance entre trois variables ou plus et sont utilisées par des outils d'inférence de réseaux d'interactions comme PIDC (CHAN et al., 2017) et MIIC (VERNY et al., 2017). Si les méthodes utilisant l'information mutuelle se révèlent assez performantes lorsqu'elles sont testées pour retrouver des interactions signées directes de réseaux simulés ou bien décrit par la littérature (PRATAPA et al., 2020), elles ne donnent pas directement des orientations aux interactions inférées. Néanmoins des méthodes complémentaires peuvent être utilisées pour orienter partiellement le graphe obtenu. Ces approches ont été mises en oeuvre pour construire un graphe d'influence partiellement orienté des acteurs moléculaires du soutien des CSH par les cellules stromales de la niche hématopoïétique (DESTERKE et al., 2020).

**Régressions.** Les méthodes de régression quant à elle permettent l'obtention directe de régulations orientées et signées en partant de l'hypothèse que l'expression d'un gène cible peut être prédit par l'expression de ses régulateurs :

$$y = f(x) + \epsilon$$

où  $y$  est l'expression du gène cible,  $x = (x_1, x_2, \dots, x_n)$  l'expression de l'ensemble de ses régulateurs et  $\epsilon$  l'erreur du modèle due au bruit dans les données. Lorsque la forme de la fonction  $f$  (qui peut être non linéaire) est connue la méthode des moindres carrées est communément employée pour trouver les coefficients impliqués dans  $f$ . Cette approche consiste à minimiser la somme des erreurs au carré entre la prédition  $f(x)$  et la valeur observée  $y$  :

$$\min ||f(x) - y||^2$$

La régression linéaire où  $f(x) = a_1x_1 + a_2x_2 + \dots + a_nx_n$  (avec  $a_1, \dots, a_n \in \mathbf{R}$ ) est le type de

régression la plus utilisée et a été employée avec différents types de régularisation pour l’inférence de graphes d’influence à partir de plusieurs jeux de données d’expression en vrac (OMRANIAN et al., 2016). A partir de données de scRNA-seq, on peut citer SINCERITIES qui combine régression linéaire régularisée et analyse des corrélations partielles pour reconstruire un graphe d’influence sur la base des changements dans les distributions d’expression des gènes le long du pseudotemps (PAPILI GAO et al., 2018).

**Arbre de régression.** Les approches s’appuyant sur des arbres de régression (HUYNH-THU et al., 2010; MOERMAN et al., 2019) ne font quant à elle aucune hypothèse sur la nature de  $f$  et n’utilise pas de données temporelles. Elles apparaissent ainsi particulièrement adaptée à l’inférence de graphe d’influence à partir de données de scRNA-seq. Ces méthodes décomposent la prédiction d’un réseau de régulation entre  $q$  TF  $x_1, \dots, x_q$  et  $p$  gènes  $y_j$  en  $p$  différents problèmes de régression définis sur les échantillons d’apprentissage  $EA_j$  sur les données d’expression de  $N$  cellules :

$$EA_j = \{(x_{1,k}, \dots, x_{q,k}, y_j), k = 1, \dots, N\}$$

Pour chacun de ces  $EA_j$  l’objectif est de trouver le modèle le plus approprié  $f_j$  pour le gène  $y_j$  tel que :

$$y_j = f_j(x_1, \dots, x_q) + \epsilon$$

qui minimise :

$$\sum_{k=1}^N (y_{j,k} - f_j(x_{1,k}, \dots, x_{q,k}))^2$$

Où  $\epsilon$  est l’erreur du modèle due au bruit dans les données.

Les  $f_j$  sont modélisés par des arbres de décision qui sont construits par divisions récursives de l’échantillon d’apprentissage EA avec des tests binaires basés chacun sur l’expression d’un régulateur ( $x_1, x_2, \dots, x_q$ ) en entrée, en essayant de réduire autant que possible la variance de la variable de sortie  $y_j$  dans les sous-ensembles d’EA résultants. Les divisions d’EA se font en comparant les valeurs des régulateurs en entrée à un seuil déterminé pendant la croissance de l’arbre.

Une mesure d’importance (MI) d’un régulateur en entrée  $x_k$  dans la prédiction du modèle d’expression du gène cible  $y$  peut être calculée et donne une estimation de la pertinence de la régulation de  $y$  par  $x_k$ . Les régulations putatives sont ensuite agrégées sur tous les gènes et classées selon leur MI normalisée. Le réseau de régulation global peut finalement être reconstruit en choisissant un seuil de MI pour ne garder que les interactions pertinentes. Le signe (activation ou inhibition) des régulations peut être déterminé en analysant la corrélation

du gène cible et de son régulateur. Plutôt qu'un unique arbre les approches citées utilisent des modèles d'ensembles d'arbres comme les forêt aléatoires (HUYNH-THU et al., 2010) et les arbres boostés (MOERMAN et al., 2019). Dans ces approches les prédictions de plusieurs arbres se complémentent ce qui améliore grandement les prédictions réalisées.

### 1.3.2.3 Apport des données génomiques et épigénétiques

L'inférence de graphe d'influence à partir de données scRNA-seq peut être améliorée par ajout de niveaux d'informations épigénétiques et génomiques (HU et al., 2020). Ces dernières années, les données épigénétiques de ChIP-seq de TF et de marques d'histones ainsi que d'ATAC-seq ont permis l'identification de régions régulatrices et de motifs de liaison à l'ADN pour des centaines de TF. Des collections de motifs de TF sont ainsi disponibles et peuvent être utilisées pour identifier des ensemble de gènes corégulés par un même TF (HERRMANN et al., 2012). Cette stratégie est employée par SCENIC qui recherche des enrichissements de motifs dans les régions régulatrices (autour du TSS) des gènes ciblés par un même TF dans le graphe d'influence inféré sur des données scRNA-seq par des arbres de régression (AIBAR et al., 2017). Ces modules composés d'un TF et de ses cibles, appelés régulons, peuvent alors être raffinés en écartant les gènes cibles ne présentant pas d'enrichissement de motif pour le TF régulateur. Cette méthode a été utilisée pour bâtir un graphe d'influence des TF régulateurs du **priming** de la CSH dans le résultat section 4.3, page 125.

Cette méthode permet de réduire considérablement le nombre de fausses connections transcriptionnelles dans le graphe d'influence inféré à partir des données d'expression. Cependant, elle ne tient pas compte du contexte épigénétique et génomique particulier des cellules étudiées. En effet, un gène cible peut effectivement présenter un motif pour un TF régulateur aux alentours de son TSS mais la séquence génomique concernée peut s'avérer être fermée par la compaction de la chromatine. L'intégration de données scATAC-seq aux données scRNA-seq permettrait de raffiner encore plus les régulons en ne recherchant les motifs qu'au niveau des régions ouvertes de la chromatine. Par ailleurs l'intégration avec des données de génomique cellules uniques permettrait également d'identifier des régulations atypiques dues à des mutations sur les régions régulatrices des gènes cibles dans certaines sous population particulières étudiées qui ne peuvent être retrouvées par SCENIC, notamment dans le cas des cancers.

Cependant les technologies multi-omiques véritablement cellules uniques ne sont encore pas très répandues et les études multi-omiques à l'échelle de la cellule sont aujourd'hui principalement menées avec des échantillons de cellules différents pour chaque niveau d'information qui nécessitent d'être intégrés avant de pouvoir être analysés (voir section 1.2.5.2, page 48). La difficulté et les biais de cette étape d'intégration sont à l'heure actuelle un obs-

tacle au développement de méthodes d'inférence de réseaux de régulation moléculaire tirant pleinement usage des données multi-omiques cellules uniques (HU et al., 2020).

### 1.3.3 Inférence de modèles logiques à partir de données cellules uniques

#### 1.3.3.1 Présentation du problème

Les données scRNA-seq représentent des observations de milliers voire de dizaines de milliers d'états cellulaires. En ordonnant les cellules le long d'une pseudo-trajectoire l'espace des données peut être interprété après binarisation de l'activité des composants d'intérêt (généralement l'expression des gènes) comme une observation d'un graphe de transition d'un BN. Ainsi à partir des transitions entre états observées dans les données, une approche de rétro-ingénierie peut être entreprise pour trouver les fonctions logiques possibles pour chaque composant au vue de la dynamique observée (Figure 1.10). Plusieurs approches ont été proposées dans ce but, certaines inférant directement le graphe d'influence et sa paramétrisation logique (MOIGNARD et al., 2015), d'autres contraignant la recherche de BN issu d'un graphe d'influence construit au préalable par les méthodes que nous venons de décrire (CHEVALIER et al., 2019; HAMEY et al., 2017).

Les méthodes dont nous allons discuter s'appuient sur la définition de problèmes de satisfiabilité booléenne à l'aide de langages de programmations déclaratifs comme l'*Answer Set Programming (ASP)* (CHEVALIER et al., 2019; OSTROWSKI et al., 2016) ou impératifs comme F# (HAMEY et al., 2017; MOIGNARD et al., 2015). Ces langages permettent d'énoncer des contraintes logiques que doivent satisfaire les BN recherchés. Un solveur, clingo pour l'ASP (GEBSER et al., 2012), Z3 pour F# (de MOURA et BJØRNER, 2008), est ensuite utilisé pour énoncer toutes les solutions possibles. Des optimisations peuvent également être ajoutées pour réduire le nombre de solutions par exemple en cherchant les solutions optimisant un score définissant l'adéquation aux données (LIM et al., 2016).

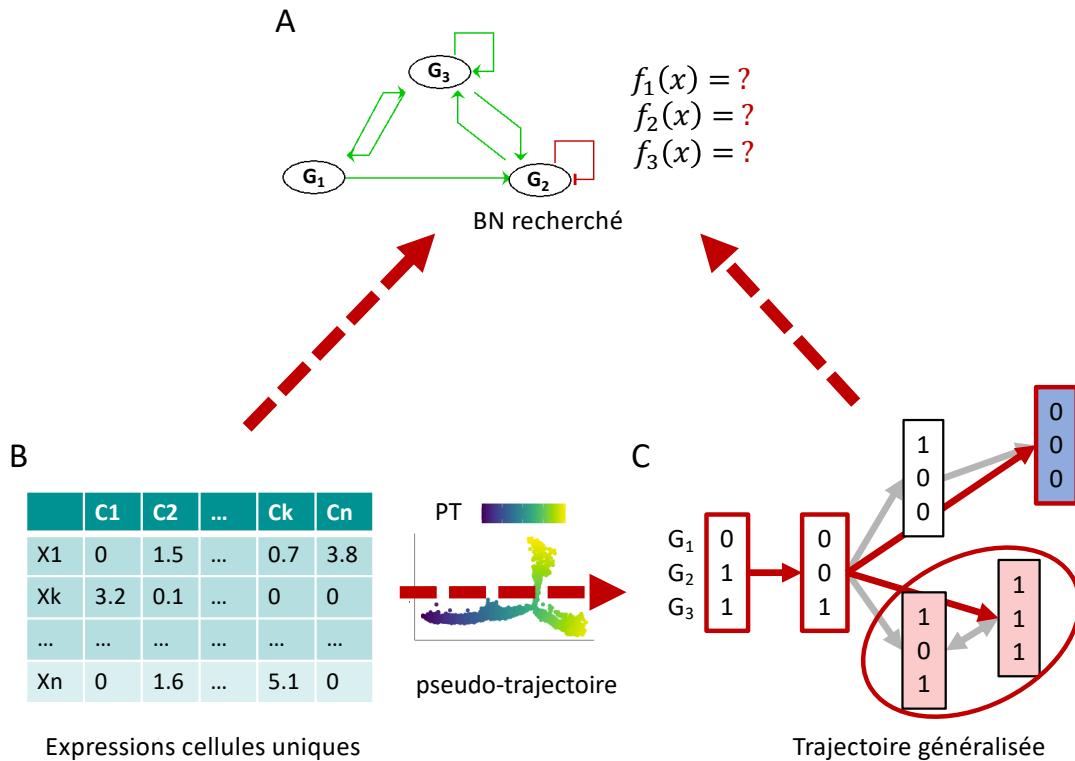


FIGURE 1.10 – Inférence de BN à partir de données scRNA-seq

A : BN entre trois composants  $\{G_1, G_2, G_3\}$  recherché (le graphe d'influence de l'exemple précédent est repris) B : Matrice de données scRNA-seq continues qui peut être utilisée pour inférer un graphe d'influence avec les méthodes présentées section 1.3.2.2, page 65. PT : Pseudotemps . C : Trajectoire généralisée possible du BN recherché obtenue après construction d'une pseudo-trajectoire et discréétisation des données d'expression des trois composants étudiés. Souvent, seules certaines transitions entre configurations (en rouge) sont clairement identifiables dans la pseudo-trajectoire construite. Ces données pseudo-ordonnées peuvent être utilisées pour inférer un BN qui reproduit la dynamique observée.

### 1.3.3.2 Revues des méthodes existantes

Le problème de la reconstruction de BN biologiques à partir de séries temporelles de données d'expression et d'un graphe d'influence a été abordé par la recherche en informatique bien avant l'arrivée du scRNA-seq. A la fin des années 90 et au début des années 2000 des méthodes ont été proposées pour inférer des BN dont les mises à jour synchrones reproduisent au mieux les transitions d'états d'expression observées sur une trajectoire (LÄHDESMÄKI et al., 2003; LIANG et al., 1998). Ces méthodes pionnières ne considèrent que des transitions synchrones et ne tiennent pas compte du signe des interactions dans le graphe d'influence. Tester l'accessibilité d'une configuration depuis une autre en sémantique asynchrone ou généralisée est un problème PSPSACE complet. Ainsi dans le but de pouvoir inférer des BN dans une sémantique autorisant plusieurs successeurs à une configuration comme on peut l'observer en biologie notamment dans les pseudo-trajectoires de différenciation, une approche basée sur une sur-approximation des transitions généralisées par la sémantique MP a été proposée

(OSTROWSKI et al., 2016). Cette méthode tient compte du signe des interactions dans le graphe d'influence en entrée et permet après une étape de vérification des modèles (*model-checking*) d'obtenir des BN dont les trajectoires simulées par une sémantique généralisée présentent des embranchements en accord avec les données en entrée.

Des méthodes ont également été développées spécifiquement dans le cas de données d'expression cellules uniques. Parmi elles SCNS toolkit apprend en même temps le graphe d'influence et sa paramétrisation logique étant donné des données d'expression cellules uniques pour un ensemble donné de composants (MOIGNARD et al., 2015). Cette méthode binarise les données puis connecte entre elles les cellules qui ne diffèrent que pour le niveau d'un gène pour avoir un aperçu du graphe de transition asynchrone que le BN recherché doit être en mesure de produire. En se focalisant sur les plus courts chemins dans ce graphe entre les états initiaux et finaux déterminés biologiquement au préalable, l'ensemble des BN possibles est finalement retrouvé. Une autre méthode considérant également des transitions asynchrones repose elle sur l'identification de paires de cellules sur un pas de pseudotemps donné (HAMEY et al., 2017). À partir d'un graphe d'influence construit à l'aide des corrélations partielles d'expression, les BN reproduisant au mieux les transitions entre états binaires donnés par les paires de cellules sont recherchés. Ces deux méthodes ont été développées pour des données d'expression cellules uniques de qPCR et leurs résultats sont très sensibles à la qualité de la binarisation des données. Leur utilisation sur des données scRNA-seq semblent ainsi compromise par le bruit beaucoup plus important dans ces données notamment à cause des *drop-outs* qui compliquent grandement la binarisation.

Pour faire face à ce biais de la discrétisation l'outil BTR propose lui d'entraîner un BN existant des composants d'intérêts à partir de données continues d'expressions cellules uniques (LIM et al., 2016). Cette méthode améliore un BN initial en cherchant à minimiser une distance entre les données biologiques continues d'expression et les données générées par les trajectoires asynchrones du BN partant d'un état initial donné. Cette méthode fait le choix de ne pas utiliser l'information du pseudotemps et ne garanti donc pas l'obtention de BN présentant des attracteurs qui correspondent aux états finaux qui sont pourtant souvent bien identifiables dans les analyses de pseudo-trajectoires. En outre bien que le BN initial puisse être construit de façon aléatoire cette option diminue fortement les performances de la méthode et n'est pas conseillée.

Finalement une méthode récente, Bonesis, s'appuyant sur la sémantique MP opte pour une approche dans lesquelles des contraintes dynamiques en ASP entre des configurations sont données en entrée selon les observations biologiques (CHEVALIER et al., 2019). Ces contraintes peuvent être l'existence de points fixes, d'attracteurs, de (non) accessibilité entre deux métaconfigurations. Cette méthode présentent deux avantages majeurs dans le cas de données

scRNA-seq. D'une part elle ne cherche pas à contraindre toutes les transitions observées dans les données. Ceci amoindrit les biais dus à l'incertitude des valeurs de pseudotemps discutés précédemment, sous réserve de définir des contraintes entre des configurations qui semblent robustes dans la pseudo-trajectoire (accessibilité d'attracteurs à partir d'une configuration initiale souche multipotent ou progéniteur oligopotent par exemple). D'autre part en considérant des méta-configurations pour la définition de ces contraintes elle permet de tenir compte de l'incertitude des niveaux d'expression mesurés en scRNA-seq qui rend pour certains gènes une binarisation hasardeuse. Bonesis a été utilisée dans le résultat section 4.3, page 125 pour chercher un BN en sémantique MP du de la CSH. Une description de Bonesis est donnée en avant propos de cette section (section 4.3, page 125).

# 2 Objectifs de la thèse

## Sommaire

2.1 Caractérisation des sous populations des HSPC et de leur vieillissement . . . . .	73
2.2 Identification des acteurs moléculaires des changements phénotypiques des CSH âgées . . . . .	74
2.3 Construction du graphe d'influence des acteurs de l'hématopoïèse précoce altérés par le vieillissement . . . . .	74
2.4 Construction d'un modèle logique de l'hématopoïèse précoce permettant d'étudier le vieillissement . . . . .	75
2.5 Développement de pipelines d'analyse . . . . .	75

Le but de ma thèse a été d'étudier le vieillissement de la CSH en adoptant la démarche de la biologie des systèmes. Nous avons commencé par une approche de scRNA-seq haut débit pour obtenir une description approfondie des populations cellulaires concernées. Puis nous avons construit un graphe d'influence des composants biologiques (TF, complexes régulant le cycle cellulaire) qui nous sont apparus les plus pertinents pour notre étude d'après nos résultats et la connaissance préalable du système étudié. Nous avons ensuite construit un modèle logique à partir de ce graphe d'influence dans le but de reproduire la dynamique de la CSH dans l'hématopoïèse précoce en condition normale. En perturbant le modèle selon nos observations du vieillissement nous avons finalement cherché à mettre en évidence des mécanismes moléculaires précis du vieillissement à tester expérimentalement.

## 2.1 Caractérisation des sous populations des HSPC et de leur vieillissement

En s'appuyant sur les avancées récentes du scRNA-seq, aussi bien en terme de débit que de méthodes d'analyse, le premier objectif de ma thèse a été de proposer une carte précise du transcriptome des populations de cellules souches et progéniteurs de la MO de souris. Je me suis plus particulièrement intéressé à la différenciation précoce des CSH au sommet de l'arbre hématopoïétique et me suis focalisé sur la population de HSPC FLT3<sup>-</sup> (Tableau 1.1). Les études précédentes de scRNA-seq bas débit sur le vieillissement de la CSH ont suggéré

l’existence de sous populations de CSH âgées jusqu’alors non-nobservée à la fois en matière d’amorçage de lignage (GROVER et al., 2016; MANN et al., 2018; SOMMARIN et al., 2021; YOUNG et al., 2016)) et en matière d’état du cycle cellulaire (KIRSCHNER et al., 2017; KOWALCZYK et al., 2015). Cependant, les résultats issus de ces différentes études sont assez divergents, du fait notamment, des différentes techniques utilisées et du faible nombre de cellules analysées. Aucune trajectoire de différenciation pour cette population de cellules très immature n’a par ailleurs été établie. Partant de ce constat, une étude scRNA-seq haut débit (plusieurs milliers de cellules) des HSPC FLT3<sup>-</sup> nous est apparue essentielle pour dresser une cartographie complète des différentes sous populations de HSPC, de leur différenciation, et de leurs altérations avec le vieillissement. Ce travail a été entrepris en collaboration avec les biologistes de l’équipe *Facteurs épigénétiques dans l’hématopoïèse normale et pathologique* du CRCM et a abouti au résultat présenté section 3.3.

## 2.2 Identification des acteurs moléculaires des changements phénotypiques des CSH âgées

Une fois les sous populations de HSPC précisément identifiées ainsi que leur altérations (transcriptomique, proportion et cycle cellulaire) au cours du vieillissement, mon travail s’est porté sur la recherche des principaux acteurs de ces changements. Cette recherche a été réalisée à partir des résultats de l’analyse d’expression différentielle issus du scRNA-seq couplée à des analyses d’enrichissements pour des processus biologiques ou signatures en lien avec la CSH précédemment établies. De façon plus originale, ce travail a été mené également via l’identification de régulons et la mesure de leur activité au sein des cellules étudiées. Les résultats de cette recherche sont présentés sections 3.3 et 4.3.

## 2.3 Construction du graphe d’influence des acteurs de l’hématopoïèse précoce altérés par le vieillissement

Mon travail a ensuite eu pour objectif de sélectionner un sous ensemble pertinent de ces acteurs pour une étude approfondie de leur activité et leurs interactions dans la régulation de l’hématopoïèse précoce qui se retrouve altérée avec le vieillissement. En partant de nos données scRNA-seq, de notre interprétation de celles-ci, et des données et connaissances préalables de l’hématopoïèse nous avons construit un graphe d’influence composé de 13 facteurs de transcription et 2 complexes régulateurs du cycle cellulaire qui est présenté dans le résultat de la section 4.3.

## 2.4 Construction d'un modèle logique de l'hématopoïèse précoce permettant d'étudier le vieillissement

Par la suite, à partir de ce graphe d'influence, mon travail s'est focalisé sur la construction d'un modèle logique décrivant la dynamique de la CSH observée dans les données scRNA-seq en condition normale. Ce travail a été réalisé grâce au développement d'une stratégie originale de synthèse de BN assistée par scRNA-seq, utilisant une méthode d'inférence de BN à partir de contraintes dynamiques (voir section 4.1.2, page 120 et CHEVALIER et al., 2019). Les perturbations *in silico* du modèle obtenu selon nos observations du vieillissement ont permis de cibler des mécanismes moléculaires pouvant être à l'origine du vieillissement qui seront prochainement testés expérimentalement au CRCM. Ces travaux sont détaillés dans le résultat de la section 4.3.

## 2.5 Développement de pipelines d'analyse

D'un point de vue pratique, ma thèse a également eu pour objectif le développement de pipelines d'analyse robuste à l'aide du gestionnaire de workflow Snakemake (KÖSTER et RAHMANN, 2018; MÖLDER et al., 2021) et d'environnements Conda (<https://docs.conda.io>). Ce développement a pour but d'une part d'assurer la reproductibilité des analyses, et d'autre part de fournir à l'équipe des outils facilement utilisables sur de nouveaux jeux de données cellules uniques dans des contextes biologiques différents. Ces pipelines sont présentés brièvement sections 3.1 et 4.1.

# 3 Résultat : Analyse scRNA-seq haut débit de l'hématopoïèse précoce

## Sommaire

3.1 Avant propos . . . . .	76
3.2 Pipeline d'analyse . . . . .	77
3.3 Article . . . . .	81
3.4 Complément biais et corrections . . . . .	113
3.5 Complément sur les pseudo-traj ectoires de différenciation . . . . .	114
3.6 Discussion . . . . .	117

### 3.1 Avant propos

Les études précédentes de scRNA-seq pour étudier les HSPC jeunes et âgés ont été conduites sur des pools de sous-types de HSPC isolés, ce qui limitait l'étude de leur proportions (GROVER et al., 2016; KIRSCHNER et al., 2017; KOWALCZYK et al., 2015; MANN et al., 2018; YOUNG et al., 2016). Ainsi, dans le but d'établir une cartographie de l'hématopoïèse précoce en condition jeune et âgée nous avons choisi d'analyser les HSPC FLT3<sup>-</sup> (Tableau 1.1), de façon globale pour être en mesure d'étudier les proportions de chaque sous populations et leurs variations avec le vieillissement. De plus, en choisissant une approche de scRNA-seq haut-débit, notre étude a aussi eu pour but d'apporter une résolution très fine de ce compartiment dans les deux conditions, afin de distinguer clairement de potentielles sous populations exclusives au CSH âgées ou bien des sous populations jeunes complètement perdues avec le vieillissement.

Les sous-type de HSPC classiquement étudiés sont définis par un ensemble de marqueurs de surface (Tableau 1.1). En scRNA-seq haut débit, du fait de la faible couverture de séquençage, l'expression protéique d'un marqueur de surface peut ne pas être corrélée au niveau de son ARNm mesuré (notamment à cause des *drop-outs*). Par conséquent, notre protocole expérimental (tri global des HSPC FLT3<sup>-</sup> non étiquetés) ne nous permet pas de connaître directement à quel sous type de HSPC, chacune de nos cellules séquencées appartient. Nous avons ainsi utilisé une approche de classification supervisée pour attribuer un sous type à chacune de nos cellules à partir d'un jeu de données étiqueté (RODRIGUEZ-FRATICELLI et al., 2018). Nous avons également utilisé ce type d'approche pour attribuer une phase du cycle

cellulaire à chacune de nos cellules comme cela a pu être fait auparavant (KOWALCZYK et al., 2015).

Un des points cruciaux de notre étude a été la construction d'une pseudo-trajectoire de différenciation de la CSH. Jusqu'alors, aucune trajectoire n'avait été proposée pour les HSPC, probablement du fait de l'important bruit biologique du cycle cellulaire pour lequel les méthodes de corrections n'étaient pas encore bien établies. Nous avons tout particulièrement cherché à caractériser si le vieillissement altérait la structure même de cette trajectoire par la perte ou l'apparition d'embranchements, ou bien modifiait le "courant" de cellules le long de celle-ci.

Finalement, cette étude a eu pour objectif d'identifier les acteurs moléculaires de l'amorçage de la CSH et leurs altérations éventuelles avec le vieillissement. Ces acteurs ont été recherchés tout d'abord à l'aide de l'analyse de l'expression différentielle entre sous-populations et avec le vieillissement. Cependant, la faible couverture due au haut débit peut brouiller ces analyses. Nous avons donc également caractérisé les différences transcriptomiques par la recherche et la mesure d'empreintes (DUGOURD et SAEZ-RODRIGUEZ, 2019) ; tout d'abord à partir des nombreuses signatures d'états de CSH établie dans la littérature, et de façon plus originale, pour les TF à l'aide de SCENIC (AIBAR et al., 2017).

## 3.2 Pipeline d'analyse

Cette étude a été l'occasion de développer un pipeline d'analyse de données scRNA-seq à l'aide du gestionnaire de workflow Snakemake (MÖLDER et al., 2021). Dans un pipeline Snakemake, chaque étape d'analyse est définie sous la forme d'une règle. Chacune d'elle décrit une étape de l'analyse définissant comment obtenir des fichiers de sortie à partir de fichiers d'entrée. Les dépendances entre les règles sont déterminées automatiquement et chaque règle peut faire appel à des environnements Conda différents ce qui permet de gérer certains conflits de versions entre les différents outils utilisés. Avec l'écriture de fichiers de configurations, un pipeline d'analyse peut être facilement mobilisable sur de nouveaux jeux de données. Ce pipeline a ainsi été réutilisé avec des ajustements mineurs pour l'analyse de jeux de données scRNA-seq pour d'autres projets de l'équipe du CRCM (HSPC de souris KO pour PLZF, Promyélocytes, HSPC et progéniteurs de souris KO pour P16 et PLZF)

### 3 Résultat : Analyse scRNA-seq haut débit de l'hématopoïèse précoce – 3.2 Pipeline d'analyse

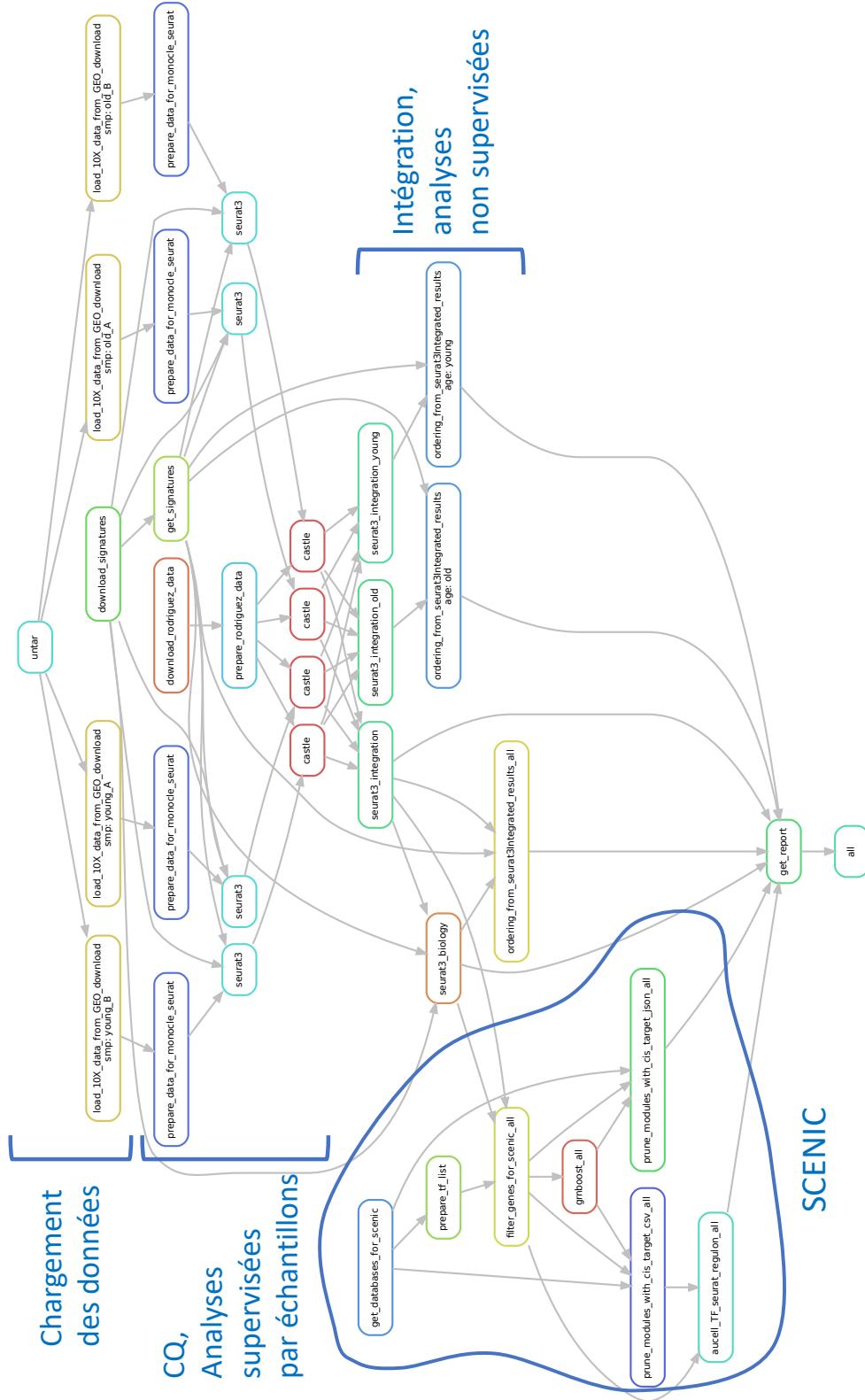


FIGURE 3.1 – Pipeline d'analyse de données scRNA-seq

Pipeline d'analyse présenté sous la forme d'un graphe dirigée acyclique partant de la règle (étape) initiale untar (décompression des données de cellranger) jusqu'à la règle finale all. Les règles sont les noeuds et leur dépendances en terme de fichier d'entrées et de sorties sont les arêtes du graphe.

Le pipeline développé comprend 4 grandes parties (Figure 3.1) : D'abord le chargement des données issues du traitement primaire (alignement, comptage) réalisé avec cellranger (le pipeline développé par 10X Genomic) ; ensuite le contrôle qualité et les classifications en sous types de HSPC et phases du cycle ; puis l'intégration et les analyses non supervisées de classification (avec seurat) et de pseudo-trajectoire (avec Monocle2) ; enfin une analyse des régulons avec SCENIC.



### 3.3 Article

HÉRAULT, L., POPLINEAU, M., MAZUEL, A., PLATET, N., REMY, É. & DUPREZ, E. (2021). Single-cell RNA-seq reveals a concomitant delay in differentiation and cell cycle of aged hematopoietic stem cells. *BMC Biology*, 19. <https://doi.org/10.1186/s12915-021-00955-z>

## RESEARCH ARTICLE

## Open Access

# Single-cell RNA-seq reveals a concomitant delay in differentiation and cell cycle of aged hematopoietic stem cells

Leonard Hérault<sup>1,2</sup>, Mathilde Poplineau<sup>1</sup>, Adrien Mazuel<sup>1</sup>, Nadine Platel<sup>1</sup>, Elisabeth Remy<sup>2†</sup> and Estelle Duprez<sup>1\*†</sup>**Abstract**

**Background:** Hematopoietic stem cells (HSCs) are the guarantor of the proper functioning of hematopoiesis due to their incredible diversity of potential. During aging, heterogeneity of HSCs changes, contributing to the deterioration of the immune system. In this study, we revisited mouse HSC compartment and its transcriptional plasticity during aging at unicellular scale.

**Results:** Through the analysis of 15,000 young and aged transcriptomes, we identified 15 groups of HSCs revealing rare and new specific HSC abilities that change with age. The implantation of new trajectories complemented with the analysis of transcription factor activities pointed consecutive states of HSC differentiation that were delayed by aging and explained the bias in differentiation of older HSCs. Moreover, reassigning cell cycle phases for each HSC clearly highlighted an imbalance of the cell cycle regulators of very immature aged HSCs that may contribute to their accumulation in an undifferentiated state.

**Conclusions:** Our results establish a new reference map of HSC differentiation in young and aged mice and reveal a potential mechanism that delays the differentiation of aged HSCs and could promote the emergence of age-related hematologic diseases.

**Keywords:** Single-cell RNA-seq, Hematopoietic stem cell, Aging, Trajectories, Differentiation, Cell cycle

**Background**

The hematopoietic stem cell (HSC) is an adult tissue stem cell residing in the bone marrow (BM), with multipotent differentiation, regenerative and self-renewal abilities, the proper functioning of which is a guarantee of a healthy immune system. HSC properties have been extensively studied thanks to the use of specific surface markers and multicolored fluorescence-assisted cell sorting (FACS) analyses that have made it possible to isolate them and test their properties during serial grafts [1, 2].

This cell-surface marker-based HSC characterization has shaped the classical but largely revisited hematopoietic model, in which the long-term HSC (LTHSC), at the top of the hierarchy, undergoes a lineage commitment through a series of discrete intermediate progenitor stages in a stepwise manner. This approach has helped to categorize short-term HSC (STHSC) and multipotent progenitor populations (MPP2, MPP3, and MPP4) [3–5].

HSCs are not a homogeneous cell population and each HSC does not contribute equivalently to all blood lineages. HSC heterogeneity was first suggested with single cell transplantation experiments showing that phenotypically identical HSC differs in self-renewal abilities and lineage differentiation potential [6–8]. Next, single cell transcriptomic approaches combined with lineage tracing suggested an initiation of transcriptional lineage

\* Correspondence: estelle.duprez@inserm.fr

†Elisabeth Remy and Estelle Duprez contributed equally to this work.

<sup>1</sup>Epigenetic Factors in Normal and Malignant Hematopoiesis Team, Aix Marseille Université, CNRS, INSERM, Institut Paoli-Calmettes, CRCM, Marseille, France

Full list of author information is available at the end of the article



© The Author(s). 2021, corrected publication 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

programs in HSCs, which bias their differentiation potential [9, 10] supporting an early HSC lineage segregation and a continuous differentiation model [11]. Thus, it is now admitted that each individual HSC, although sharing the same marker combination, differs in terms of functional outputs and molecular signature [12–14].

This HSC heterogeneity has physiological consequences upon aging. Hematopoietic aging is associated with a reduced production of red blood cells and lymphocytes concomitant to an increase of myeloid and megakaryocytic cells that promote immunosenescence and myeloid malignancies [15, 16]. Evidence indicates that these alterations of the hematopoietic system come from an age-related modification of the HSC compartment. Intrinsic changes such as accumulation of DNA damage, changes in the activity of epigenetic modulators, and imbalance between repressive and activating histone marks in HSCs have emerged as contributing factors of hematopoiesis aging [17, 18]. HSCs that are heterogeneous with respect to their self-renewal and differentiation capacities at birth pass through clonal selection over time due to environmental cues [19]. This results in an increase in myeloid- and megakaryocytic-biased HSCs within the phenotypic LTHSC compartment [20, 21]. Thus, aging is not only reflecting an intrinsic uniform change in lineage output of the HSCs but is rather due to a shift in the relative proportion of HSCs with different characteristics [22].

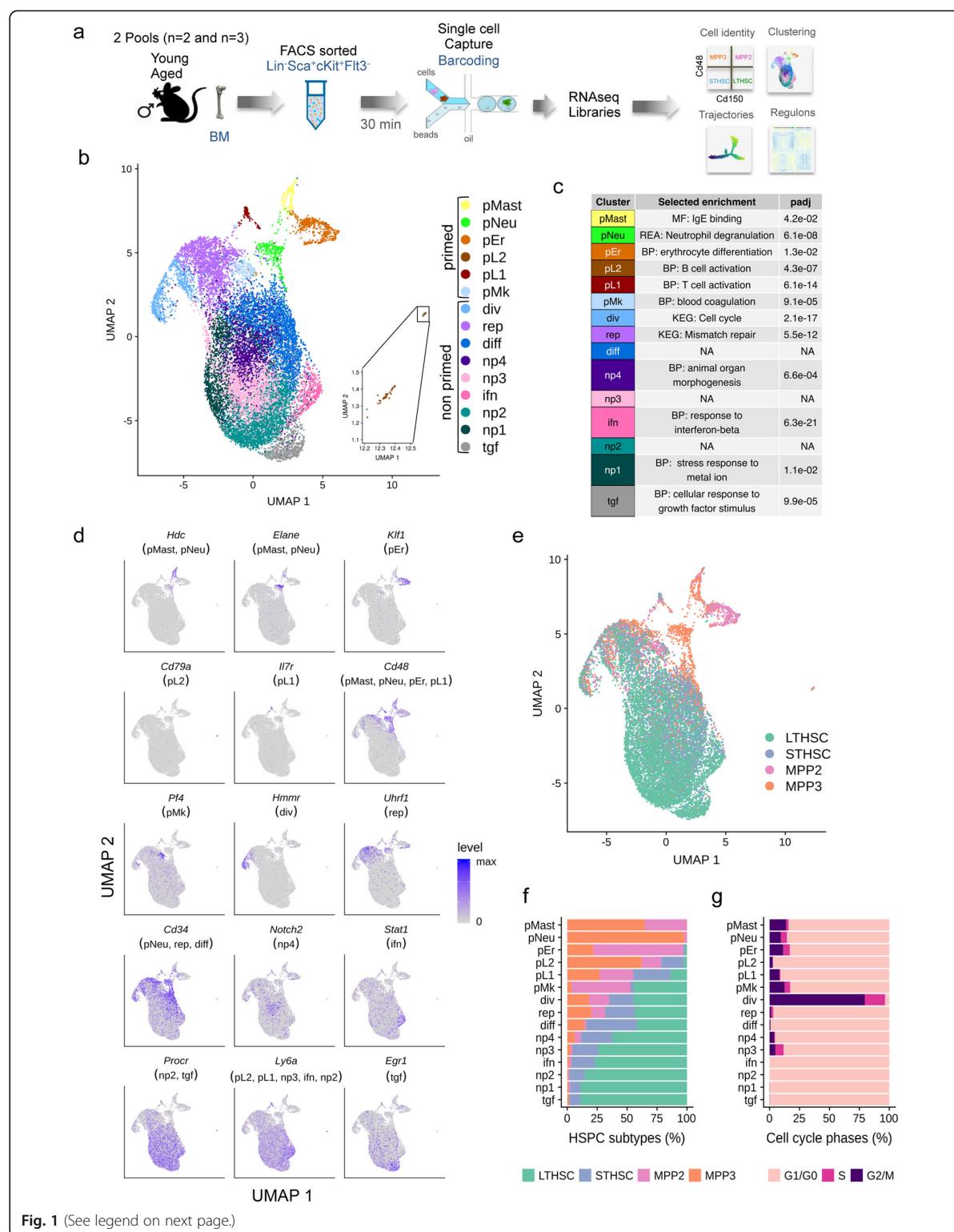
Previous studies on age-related transcriptomic changes of HSCs at the single cell resolution have revealed an expansion of platelet-primed HSCs [23] and a gain of a self-renewal expression program [24] with aging. However, the resolution of the analyses in particular regarding the proportion of the different HSC populations and their variation upon aging were limited due to the small number of analyzed cells and sorting strategies. Here, we took advantage of the 10x Genomics approaches and the development of new bioinformatic methods and tools to increase the resolution and revisit the transcriptional heterogeneity and change upon aging of the HSC compartment. By analyzing 15,000 single murine hematopoietic stem and progenitor cells (HSPC) transcriptomes, we detected new rare HSC subpopulations that accumulate upon aging. We also highlighted transcriptional program changes linked to cell cycle activity during aging that participate to the HSC age-related alterations.

## Results

### Stratification of HSPCs using single-cell transcriptome analysis highlighted 15 different clusters

To characterize HSC populations by single cell RNAseq (scRNA-seq), we purified HSPCs, including LTHSCs, STHSCs, MPP2, and MPP3 by FACS from BM pools of young ( $n = 5$ ; 2–3 months) and aged ( $n = 5$ ; 17–18

months) mice applying the widely used Lin<sup>-</sup>, Sca1<sup>+</sup>, cKit<sup>+</sup> (LSK) marker strategy with the addition of the Flt3 marker to exclude the Flt3<sup>+</sup> LSKs also referenced as MPP4 (Fig. 1a and Additional file 1: Fig. S1A). Four pools (2 pools of young and 2 of aged) of thousands HSPCs were subjected to 10x Genomics Chromium capture platform and a total of 15,000 single HSPC transcriptomes were sequenced (young pools, with 5189 and 2244 cells and aged pools with 3328 and 4154 cells after quality control; Additional file 2: Table S1). As we made the assumption that aging would not dramatically modify HSC identity, we first analyzed young and aged HSPCs together using Seurat workflow [25] for the integration of the different samples to correct batch effect. Reduction of dimension and unsupervised clustering were performed on cell-cycle-corrected data using Uniform Manifold Approximation and Projection (UMAP) [26]. A total of 15 clusters were identified (Fig. 1b), which were characterized further by identifying their markers using differential expressed gene (DEG) analysis on the log-normalized data without any correction (Additional file 3: Table S2) and by deducing their characteristics (Fig. 1c) from gene set enrichment analysis (Gene Ontology, KEG, and Reactome pathways; Additional file 4: Table S3) and gene signatures related to hematopoiesis (Additional file 5: Table S4a). Six clusters were classified as lineage-primed clusters as they were clearly enriched for HSPCs with megakaryocyte (pMk), erythroid (pEr), neutrophil (pNeu), mastocyte (pMast), and lymphocyte (pL1 and pL2) commitment gene markers (Fig. 1b–d; Additional file 3: Table S2, Additional file 4: Table S3 and Additional file 5: Table S4a). Nine clusters were considered as non-primed due to their lack in expression of lineage restricted-genes. They accounted for a large majority of the analyzed cells (90%) (Fig. 1b–d; Additional file 3: Tables S2 and Additional file 5: S4b). The 4 phenotypically distinct HSPCs, LTHSCs, STHSCs, MPP2, and MPP3 were assigned by supervised classification using previously published scRNA-seq data of FACS-labeled HSPCs [11] (Additional file 1: Fig. S1B) and were superimposed on the UMAP (Fig. 1e). This showed that globally MPP2 and MPP3 were composed of lineage-primed clusters, suggesting their “more differentiated” state in comparison to the remaining clusters while LTHSCs were enriched with non-primed clusters (Fig. 1e). Distribution of the different populations among the clusters showed that the neutrophil-biased cluster (pNeu) was almost exclusively enriched with MPP3 (98%), while pMast and pEr were enriched with both MPP2 and MPP3 (Fig. 1f and Additional file 5: Table S4c). The pMK cluster was composed of almost 50% of LTHSCs, supporting previous work suggesting that platelet-biased stem cells reside at the apex of the HSC hierarchy [27].



**Fig. 1** (See legend on next page.)

(See figure on previous page)

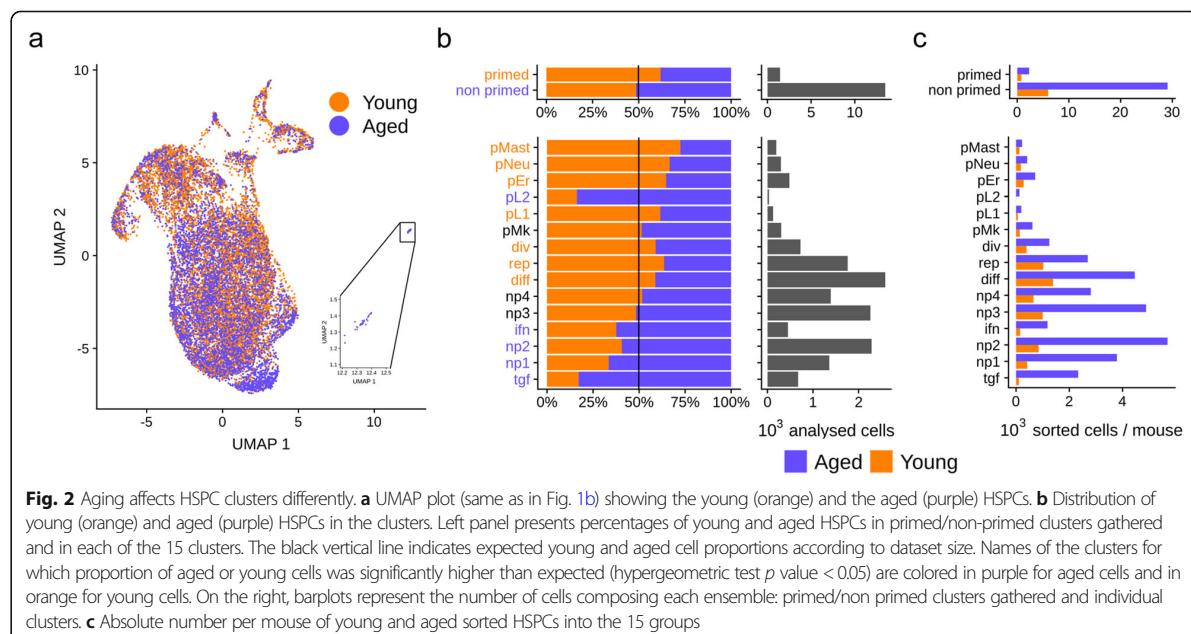
**Fig. 1** Unsupervised clustering of young and aged HSPCs revealed 15 clusters gathering mainly immature and to a lesser extend lineage-primed HSPCs. **a** Overview of the scRNA-seq sample preparation and analysis. Cells were isolated from bone marrow (BM) of young and aged mice and pooled to obtain 2 pools for each age. Pools of 2 and 3 BMs for both young and aged mice were analyzed. BM cells were FACS sorted to purify Lin<sup>-</sup>, Sca-1<sup>+</sup>, c-Kit<sup>+</sup> (LSK) Flt3<sup>-</sup> cells that defined the HSPCs. The four pools of HSPCs were processed using droplet-based single cell sequencing (10X Genomics) and multiple analyses were performed using bioinformatics tools to characterize aging effects. **b** UMAP plot of young and aged HSPCs (15,000 cells) analyzed using Seurat. Colors marked the 15 distinct clusters identified by unsupervised clustering and characterized with differential gene expression and gene set enrichment analyses. Each dot represents a cell. **c** Selected enrichment of our analysis (Gene Ontology, KEGG and Reactome pathways Supplemental Table S3) for markers of each cluster and corresponding *p* values adjusted for multiple testing (*p*adj). NA indicates non-relevant enrichment. **d** UMAP plots colored by expression of selected cluster markers. Cluster names are indicated in parenthesis. **e** Localization in the UMAP of LTHSCs, STHSCs, MPP2 and MPP3, identified by supervised classification. **f** Percentage of LTHSCs, STHSCs, MPP2 and MPP3 within the HSPC population, in each of the 15 clusters. **g** Percentage of computationally assigned cell cycle (G1/G0, S and G2/M) phases in each of the 15 clusters

Four non-primed clusters, np1, np2, np3, and np4, were overlapping and positioned at the center of the UMAP (Fig. 1b) with few specific gene markers (Fig. 1c, d; Additional file 3: Table S2 and Additional file 1: Fig. S2). They were characterized by a high percentage of LTHSCs (Fig. 1e, f; Additional file 3: Table S2). By contrast, 2 clusters, also composed mainly of LTHSCs, harbored a very distinguishable signature for growth factor signaling (tgf) and interferon response (ifn) respectively (Fig. 1b–d and Additional file 4: Table S3), witnessing the existence of cells with signaling features at the top of the differentiation hierarchy. The remaining 3 clusters (diff, div and rep) were composed of less than 50% of LTHSCs (Fig. 1e and Additional file 5: Table S4c) suggesting their intermediate state in term of differentiation. The cluster named diff had very few distinguishable markers but was enriched with *Cd34* expressing cells (Fig. 1d and Additional file 1: Fig. S2). Interestingly, this cluster was the most enriched with STHSCs (Fig. 1e and Supplemental Table S4c), which have been characterized by the appearance of the *Cd34* at their surface [2]. The div cluster, characterized by enrichment for the cell cycle KEGG pathway (Fig. 1c and Additional file 4: Table S3) and genes involved in division such as *Hmmr2* (Fig. 1d and Additional file 1: Fig. S2), was particularly different from the other clusters by its enrichment in G2/M cells (Fig. 1g and Additional file 5: Table S4d). The rep cluster was characterized by genes involved in DNA repair and replication and presented a specific high expression of *Uhhfr* (Fig. 1c, d and Additional file 1: Fig. S2 and Additional file 4: Table S3).

As a whole, these results highlight the interest of gene expression signature to identify heterogeneity in the HSC population. They support the presence of differentiation-biased cells in the immature hematopoietic compartment and demonstrate that transcriptional programs can subdivide HSPCs in different clusters besides their classical differentiation state defined by cell surface markers.

#### Aging affects HSPC clusters differently

To assess the aging effect at the level of HSC populations, we first confirmed by FACS analyses and by transcriptomic-based cell population predictions the well-described accumulation of LTHSCs that occurs at the expense of the STHSCs and the MPP3 upon aging (Additional file 1: Fig. S1C and D). Analysis of young versus aged cells in the UMAP plot showed that aged cells were significantly more distributed in the non-primed clusters while lineage-primed clusters were enriched with young HSPCs (Fig. 2a, b). Indeed, the primed lymphoid (pL1) and the myeloid primed (pMast, pNeu, and pEr) clusters were predominantly composed of young cells (Fig. 2b and Additional file 5: Table S4e). An exception was observed for the pL2 cluster; although representing very few cells, this cluster, characterized with B cell markers, was comprised mainly of aged ones (Fig. 2b and Additional file 5: Table S4b and e). Interestingly, these potentially aged B-biased cells were characterized, in addition to the expression of early B cell markers such as *Ly6d* and *Cd79a* (Fig. 1d, Additional file 1: Fig. S3 and Additional file 3: Table S2), by *Trp53inp1* expression, for which we recently showed its involvement in the blockage of early B cell developmental step [28]. Analysis of absolute numbers of sorted HSPCs per mouse confirmed the higher increase of the non-primed clusters in comparison to the primed ones upon aging (Fig. 2c). This increase in HSC subpopulations was specially marked for the non-primed np1, np2, ifn, and tgf clusters that were largely amplified in aged condition (Fig. 2c and Additional file 5: Table S4e). This result highlights an amplification of LTHSCs being able to respond to different stimuli such as ifn and tgf signaling that may overlap with previously reported HSC sub-populations, which promotes differential responses to inflammatory challenge in aged HSCs [29]. Noticeably, we observed that the age-induced decrease of pL1 cluster and increase of tgf cluster were mainly driven by one batch, specific for



each of them (Additional file 5: Table S4e) witnessing heterogeneity of aging inter mouse groups.

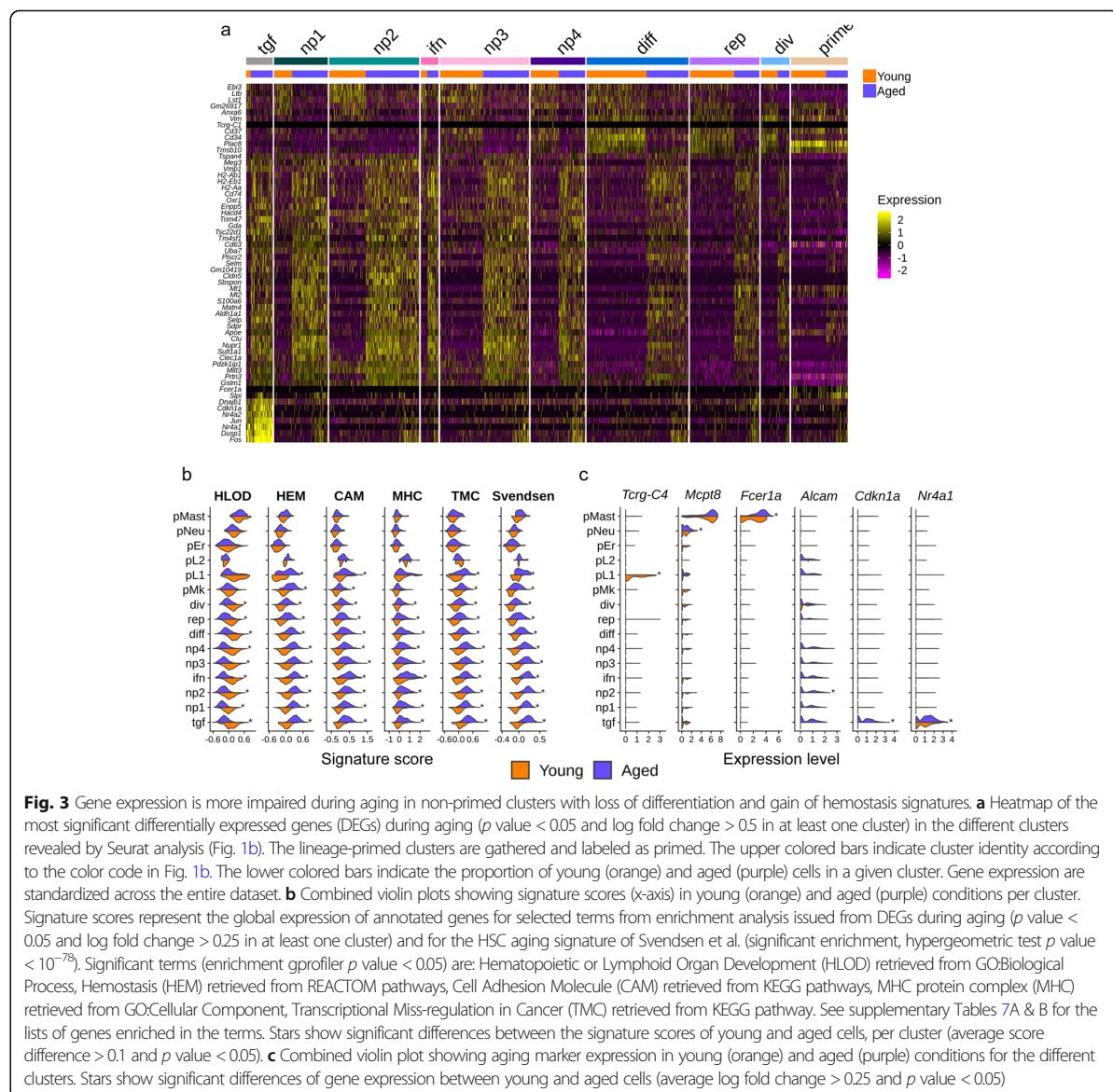
From these results, we conclude that neither HSPCs nor individuals are affected equally by aging. Globally, aged hematopoiesis is stemming from HSPCs that are not lineage primed and but characterized by specific signaling signatures.

#### Gene expression is more altered upon aging in non-primed clusters, with a loss of differentiation and a gain of hemostasis signatures

To reveal age-dependent changes in gene expression, we first compared the transcriptomes of young and aged HSPCs. Differentially expressed gene (DEG) analysis highlighted a global HSC aging signature that was characterized by an upregulation of the stress gene *Nupr1*, the platelet-lineage markers *Vwf* and *Clu*, and markers of undifferentiated HSPCs such as *Procr* and *Slamf1*, as well as by a downregulation of genes that mark HSC differentiation, such as *Cd34* and *Cd48* (Additional file 1: Fig. S4 and Additional file 6: Table S5). These results are in line with the altered differentiation potential and platelet bias of aged HSPCs [23] and a recently published comprehensive HSC aging signature [30].

In order to assess the heterogeneity of transcriptome changes upon aging according to HSC clusters, we analyzed changes in gene expression of each cluster separately (Additional file 7: Table S6). Heatmap of the most differentially expressed genes (DEGs) (log fold change  $> 0.5$ ) upon aging analyzed per clusters showed that the non-primed clusters exhibited more differences in their

transcriptome than the primed ones and that these differences were towards an increase of gene expression rather than a decrease, suggesting an increased cell-to-cell transcriptional variability upon aging (Fig. 3a and Additional file 1: Fig. S6). For these non-primed clusters, except for the *tgf* cluster, the differential gene expression analysis per cluster followed the aging signature that was observed when analyzing the totality of the cells ( $R^2 > 0.8$ ; Additional file 6: Fig. S5). Enrichment analysis of DEGs upon aging revealed a negative regulation of hematopoietic or lymphoid organ development (HLOD) marked by the downregulation of *Cd34*, *Plac8*, and *Foxo3* (Additional file 8: Table S7A), together with a positive regulation of hemostasis with *Clu* and *Selp* increased expression, Cell Adhesions Molecule (CAM) genes such as *Alcam*, *Jam2*, Major Histocompatibility Complex (MHC) H-2 genes and genes involved in transcriptional miss-regulation in cancer (TMC) (Additional file 8: Table S7B). TMC enrichment, in addition to TFs such as *Fli1* and *Pbx1*, relies on cell cycle kinase inhibitors *Cdkn1a* and *Cdkn2c* and the stress response gene *Nupr1* suggesting a deregulation of the cell cycle phases upon aging. Globally, we found that aging feature-score differences were more pronounced in the non-primed clusters than in the lineage-primed ones (Fig. 3b; Additional file 8: Table S7C). However, we highlighted that two lineage-primed clusters, the *pL1* and *pMK* clusters, were transcriptionally affected by aging, with an increase in HEM, TMC, and CAM signatures (Fig. 3b and Additional file 8: Table S7C). These observations were confirmed by analyzing a computed aging signature



score retrieved from a comprehensive aging signature [30] across the clusters (Fig. 3b and Additional file 8: Table S7C). Looking at some genes individually, we were able to highlight some age-related changes affecting particular clusters. We observed a downregulation of the T cell gene *Tcrg-C4* in the aged pL1 cluster and an upregulation of the protease mast cell gene *Mcpt8* and myeloid integrin gene *Fcer1a* in aged pMast and pNeu cluster respectively (Fig. 3c and Additional file 7: Table S6). We observed an upregulation of *Alcam* required for HSC maintenance in aged np2 cluster (Fig. 3c and Additional file 7: Table S6). Finally, we also observed a very specific transcriptome in aged tgf cluster characterized by an

increase of genes involved in HSC quiescence such as *Cdkn1a*, *Nr4a1* (Fig. 3c), which were clustered together in the heatmap of DEGs upon aging (Fig. 3a).

Altogether, our results revealed particular age-related changes mostly affecting the transcriptome of HSPCs from non-primed clusters and characterized by a loss of differentiation genes that could account for the functional changes of the aged hematopoietic compartment.

**Differentiation trajectory shows a HSPC progression towards T, Mast/Neu and Mk/Er fates that is altered with age**  
It has been recently suggested that HSCs undergo a continuous differentiation process rather than a stepwise

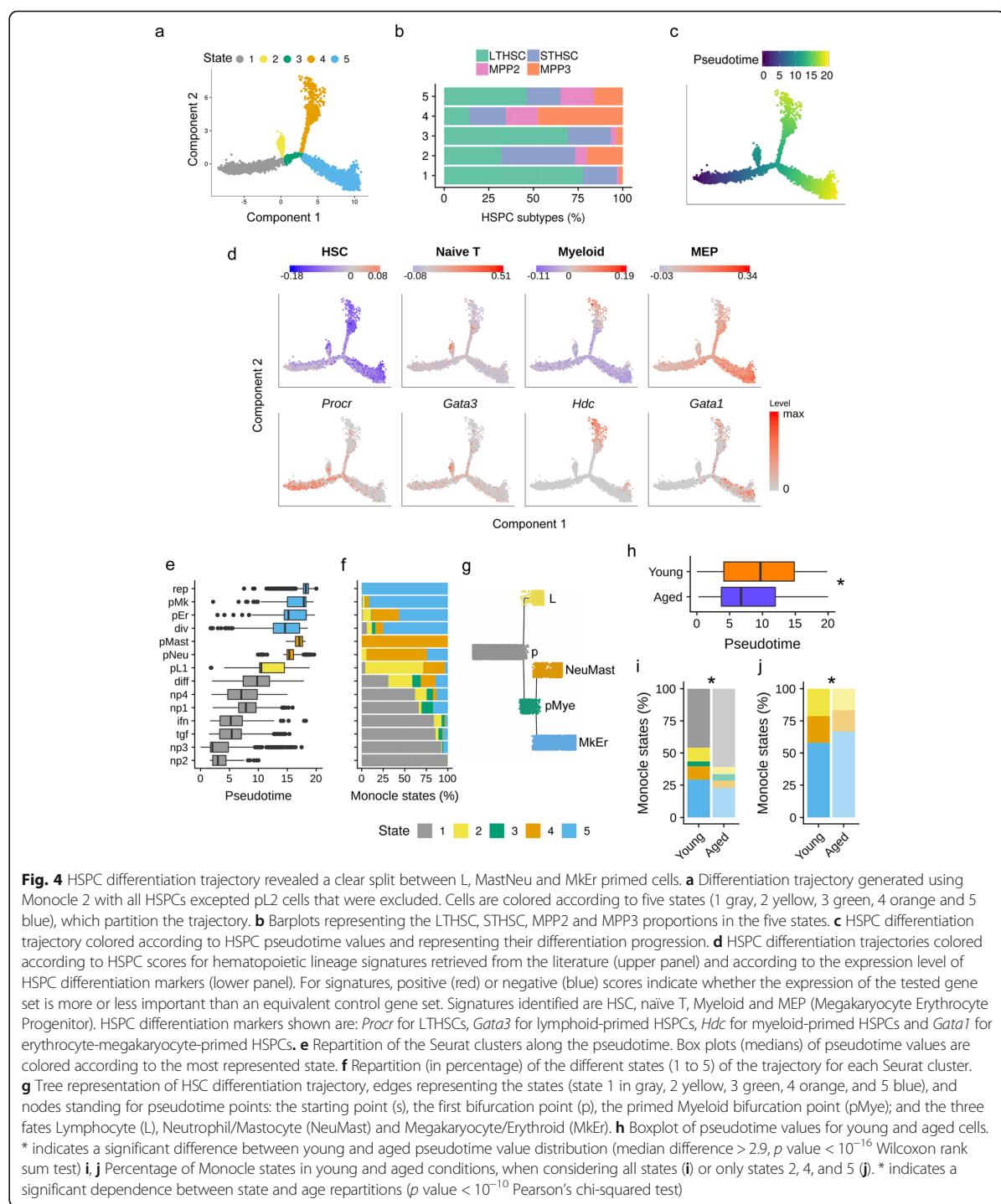
process [12]. In order to better capture and understand the progression of this differentiation process during aging, we constructed pseudotime trajectories by ordering HSPCs based on the similarities between their expression profiles with Monocle [31]. We first generated the trajectories of young and aged HSPCs separately, which showed a very similar shape, with the exception of a group of cells standing apart from the aged trajectory, and made exclusively of pL2 cells (Additional file 1: Fig. S7A). Because these cells were detected only in aged HSPCs and were clearly distant from the rest of the cells in the UMAP (Fig. 1b), we excluded them for cell pooling and ordering for both ages. Thus, we analyzed the differentiation trajectory inferred from young and aged cells pooled together, without pL2 cells. The resulting trajectory was partitioned into 5 segments, called Monocle states 1, 2, 3, 4, and 5 (Fig. 4a). The departure of the trajectory was identified at the extremity of the state 1, as this state possessed the highest percentage of LTHSCs (Fig. 4b, c). States 2, 4, and 5 were enriched with MPPs suggesting their progression towards differentiated states (Fig. 4b, c). We characterized the 5 states of the trajectory with previously published signatures related to HSPCs and hematopoiesis (referenced in Additional file 9: Table S8A) and with our state marker analysis (Additional file 9: Table S8B & S8C). We revealed that HSPCs in state 1 expressed a HSC signature with especially cells expressing the dormant HSC marker (*Procr*); state 2 cells (after the first bifurcation) were characterized with naive T cell signature and were expressing *Gata3*, suggesting a primed-lymphoid differentiation state (Fig. 4d); state 4 cells were characterized by a myeloid signature [32] and high expression of *Hdc*, previously reported as a marker of myeloid biased HSPCs [33], while cells in state 5 presented a Megakaryocyte Erythrocyte Progenitors (MEP) signature [34] and expressed *Gata1* a MEP marker (Fig. 4d).

Analysis of Seurat cluster position and spreading on the trajectory strengthened the pseudotime differentiation relevance with lineage-primed clusters located at the two extremities of the trajectory and suggested a differentiation specificity of the states (Fig. 4e; Additional file 1: Fig. S8). In addition, we assessed the similarity of our pseudotime cell values with a published HSC score to validate their degree of stemness [35]. By doing so, we identified the two non-primed clusters np2 and np3, located at the beginning of the trajectory, as the most immature ones (Additional file 1: Fig. S9). Analysis of the five state proportions across the clusters revealed a first bifurcation separating pL1 cells (state 2), from cells primed for myeloid lineages (state 3), and then a clear branching between Neu/Mast-primed (NeuMast) HSPCs (state 4) and Mk/Er-primed (MkEr) HSPCs (state 5) (Fig. 4f). The specificity of state 5 for megakaryocyte differentiation was supported

by the high representation of the rep cluster (Fig. 4f), characterized by a reparation gene signature (Additional file 4: Table S3), which was previously associated with megakaryocyte fate [36]. Separate pseudotime ordering of young and aged HSPCs provided very similar segregation between the lineage-primed HSPCs, with one bifurcation from LTHSC (state 6) towards Neu/Mast-primed (NeuMast) HSPCs (state 7) and Mk/Er-primed (MkEr) HSPCs (state 8) (Additional file 1: Fig. S7A-E). However, the bifurcation towards the lymphocyte fate was not retrieved probably because of the reduction of the pL1 cell number due to the sample splitting (Additional file 1: Fig. S7A). Hence, to synthetize our analyses, we proposed a tree-representation of the HSC differentiation trajectory (Fig. 4g) where nodes stand for pseudotime points, and edges for Monocle states. It contains 6 nodes: a root, the starting point (s); two internal nodes, the first bifurcation point (p) and primed Myeloid bifurcation point (pMye); and three leave nodes, the three fates Lymphoid (L), Neutrophils/Mastocytes (NeuMast), and Megakaryocyte/Erythroid (MkEr).

Next, we compared the differentiation progression of young and aged HSPCs. Aged HSPCs appear to be significantly delayed in the pseudotime (Fig. 4h) while Seurat cluster spreading along the trajectory showed no clear differences of any cluster pseudotime position according to age (no median difference higher than 0.8 unit of pseudotime; Additional file 1: Fig. S10A). Looking at the proportion of the different Monocle states of the trajectory according to age revealed an increase in aged HSPCs in states 1 and 3 in comparison to young ones (Fig. 4i), belonging to the non-primed clusters np3, tgf, ifn, np4, diff, and div (Additional file 1: Fig. S10B). When focusing on cells from states 2, 4, and 5, which reflect the 3 lineage-primed HSPC states, we observed that the proportion of state 5 (MkEr fate) was larger in aged than young condition (Fig. 4j), although age was not affecting the percentage of the Monocle states from lineage-primed cluster cells (Additional file 1: Fig. S10B). This suggests that while less aged HSCs were detected in the three differentiation paths, cells with MkEr fate are more maintained upon aging than the ones towards NeuMast and L fates.

In conclusion, our trajectory analysis revealed a priming of HSPCs for lymphoid lineage that occurs early in the differentiation process and evidenced a clear split between the NeuMast and the MkEr HSC fate identifying an early lineage specification of HSCs (Fig. 4g). While the global shape of the trajectory and the lineage specification of the HSPCs are conserved upon aging, repartition of the aged HSPCs along the differentiation trajectory is altered with a decrease in terminal states 2 and 4 conducting respectively to L and NeuMast fates, in favor to cells of the initial states 1 and 3.



**HSPC differentiation trajectory is associated with transcriptional programs that are altered upon aging**  
Cell fate decision and proper function of HSCs rely on tightly controlled transcriptional programs orchestrated

by transcription factor (TF) activity [37]. Since level of the expression of TFs is not sufficient to assess their activity, we measured changes in TF activity during differentiation and aging of HSPCs. For that, we took

advantage of Single-Cell Regulatory Network Inference and Clustering (SCENIC) approach [38] that calculates the activity of a given TF (regulon score) based on target expression and cis-regulatory elements. We considered 154 TFs, selected from the literature or from our single cell expression data analysis (Seurat cluster markers), out of which, 58 were identified as active regulons in our HSPCs (Additional file 10: Table S9A). By looking at regulon activities of young HSPCs along the trajectory, we revealed a specific regulon signature for each state (Additional file 10: Table S9B). State 1 was characterized with activity of the stress sensors Atf3, the interferon signaling factors, Irf1, Irf7, Irf9, and the downstream targets of the Tgfbeta signaling, Stat1, Klf4, Egr1, Klf6, Junb, depicting a stemness state (regulon clusters C1a and C1b Fig. 5a and C1a Fig. 5b, young panel). Comparison of TF activities between state 2 and state 3 at the first bifurcation (p) emphasized the L fate of state 2 with the detection of high activity of the T cell transcription factors Ikzf1, Sox4 (regulon cluster C2 Fig. 5a, young panel) while state 3 cells enter a more general differentiation program with a slight increase of regulon activities such as Myc (regulon cluster C3 Fig. 5a, young panel). As expected, aging reduced the activity of the two regulons in state 2 witnessing the reduced lymphoid activity during aging. By contrast, Klf6, Junb, Jun, and Stat1 activities of aged HSCs were spread and increased in aged states 1 and 3, (Fig. 5a, b and Additional file 10: Table S9C), which was consistent with the stem cell activity of aged states 1 and 3 containing mainly LTHSCs (Fig. 4b).

By looking at the second bifurcation (pMye) between state 4 and state 5, we confirmed that state 4 was neutrophil- and mast-biased as it was indorsed with a high activity of C/ebpa-e, Runx1, and Irf8, involved in myeloid differentiation (regulon cluster C4 Fig. 5b, young panel). Noticeably, aging decreased the activity of regulons involved in myeloid fate such as Cebpa and -e in state 4 (Fig. 5b and Additional file 10: Table S9C). This result was consistent with the decrease of neutrophils and mastocyte primed-cell number with aging (observed in Fig. 2b) and strengthened our hypothesis that myeloid bias of aged hematopoiesis would not come from this path of the trajectory. Cluster C5 of the heatmap shows that State 5 was characterized with a strong activity of Klf1, E2f8, Ybx1, Gfi1b, and Ezh2, all of which are implicated in the erythroid/megakaryocyte development (regulon cluster C5 Fig. 5b, young panel). Interestingly, the activity of E2f8 was significantly reduced with aging in state 5 whereas Gfi1b activity was considerably increased in this aged state. It should be noted that Gfi1b is the regulon that experienced the greatest increase in activity with aging, not only in state 5, but also at the beginning of the trajectory in state 1. As Gfi1b is a master regulator of thrombopoiesis (reviewed in [39]) and as we found

some of its targets such as *Clu*, *Esam*, and *Serpina1a*, annotated for hemostasis (Additional file 10: Table S9A) upregulated with aging (Additional file 8: Table S7B), we suggested that Gfi1b sustains the platelet bias of aged HSPCs.

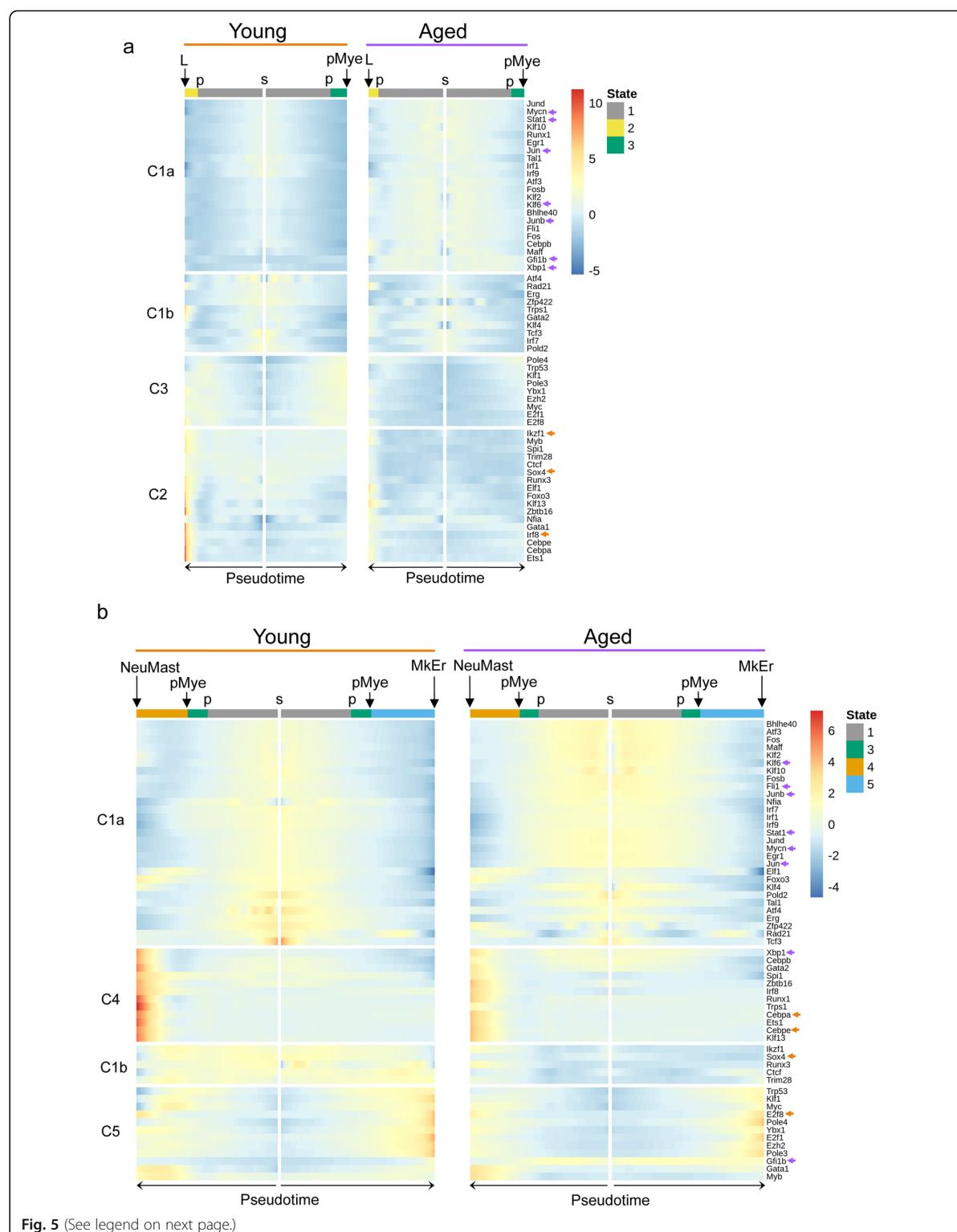
Thus, TF activity analyses over the pseudotime corroborated the trajectory features and clearly identified a separation in TF activity that explains the L priming (Fig. 5a) and the two distinct myeloid fates, NeuMast and MkEr (Fig. 5b). It also indicated that aging is associated with marked changes in TF expression and activity with a gain of TFs involved in stemness and platelet activity and a loss of lineage-specific factors that drive lineage commitment and terminal differentiation.

#### Cell cycle analysis along pseudotime highlights a delay in differentiation associated with cell cycle arrest in aged condition

As one of the hallmarks of HSC aging is a reduction of cycling HSCs [40], we analyzed the cell cycle phases according to BM age. We showed an increase of non-cycling HSPCs (G1/G0) at the expense of the S and G2/M phases in aged BM in comparison to young one (Fig. 6a). Analysis of LTHSC, STHSC, MPP2, and MPP3 population separately showed that age did not typically affect the proportion of cycle phases within each subtype, with the exception of a slight but significant change in LTHSCs and MPP2 (Fig. 6b). This suggests that the increase of the G1/G0 phase proportion observed upon aging is mainly due to the accumulation of quiescent LTHSCs that are known to be arrested in G1/G0 phase [41], and to a lesser extent to LTHSC and MPP2 intrinsic cell cycle changes induced by aging.

Positioning quiescent versus proliferative cells along the trajectories showed that quiescent cells were at the departure of the trajectory while proliferating cells were towards the differentiated states (Fig. 6c, left panel). Comparison of the quiescence and proliferation signatures between young and aged HSPCs showed a quiescence gain in the aged condition in the first part of the trajectory (states 1, 2 and 3) while the proliferation signature remained unchanged (Fig. 6c, right panel and Additional file 11: Table S10A).

Next, we addressed the question of cell cycle and its influence on HSPC aging. We first looked at the distribution of young and aged HSPCs along the trajectory, analyzing T, NeuMast, and MkEr fates separately (Fig. 6d). Doing so, we confirmed the accumulation of aged HSPCs in state 1 before the first bifurcation point p and the decrease of aged cells in the differentiated states 2, 4, and 5 (Fig. 6d). To associate cell-cycle status and cell accumulation, we performed high-resolution analysis of HSC cell cycle along the trajectory by plotting the ratio of dividing cells on pseudotime bins for young and aged

**Fig. 5** (See legend on next page.)

(See figure on previous page.)

**Fig. 5** HSPC differentiation trajectory associates with transcriptional programs that are altered upon aging. **a, b** Heatmaps showing standardized regulon activity scores, recovered with the AUCell procedure of Scenic, for young (left panel) and aged (right panel) HSPCs across Monocle states. Cells (columns) were ordered according to their pseudotime, and color bars at the top of the heatmaps indicate the state at which cell belongs (1 gray, 2 yellow, 3 green, 4 orange, and 5 blue). Regulons (rows) were hierarchically clustered, based on their activity score in young HSPCs. In **a**, 4 clusters of regulons are highlighted when analyzing regulon activity along pseudotime trajectories from s to L fate and from s to pMye bifurcation point (i.e., across Monocle states 1, 2 and 1, 3). In **b**, regulon activity along pseudotime trajectories from s to NeuMast and from s to MkEr fates (i.e., across states 1, 3, 4 and 1, 3, 5) is analyzed and 4 other clusters of regulons are highlighted. Arrows mark regulons for which a significant difference of activity with aging (average AUCell score difference between young and aged cells  $> 0.002$  and  $p$  value  $< 0.05$ ) were found in at least one of the considered states (i.e., states 1, 2, and 3 in **a** and 1, 3, 4, and 5 in **b**). The color indicates if regulon activity is increased (purple) or decreased (orange) in aged condition

cells in Lymphoid, NeuMast, and MkEr fates separately (Fig. 6e). This highlighted a dramatic loss of dividing cells in aged condition in state 1 with the exception of cells located at the very beginning of the trajectory (Fig. 6e). We hypothesized that these dividing cells (that are LTHSCs and belong to np3 cluster) represent cell-cycle activity of self-renewing LTHSCs. Interestingly, we found no difference in cell cycle phase proportion between these young and aged LTHSCs ( $p$  value  $> 0.3$  Pearson's chi-squared test; Additional file 1: Fig. S12), suggesting a conservation of self-renewal potential in aged HSCs. By opposition, the absence of cell cycle activity of aged HSPCs later in state 1, which may represent cell cycle activity linked to differentiation, underlines a default in cell division of aged HSPCs associated to differentiation (Fig. 6e). Division rate of aged HSPCs became positive after the first bifurcation and was similar to what we observed in young HSPCs (Fig. 6e), with the exception of a decrease in aged cycling cells in state 4 (towards NeuMast fate) suggesting a default of cell cycle in aged Neu-primed HSPCs. Visualization of the distribution of the different HSPC subsets confirmed the accumulation of aged LTHSCs at the expense of the STHSCs and revealed a dramatic loss of NeuMast-primed cells upon aging (Fig. 6f).

We extracted from our DEG analyses with aging (Additional file 11: Table S10B) DEGs involved in proliferation, differentiation, and cell cycle and analyzed their expression profile in young and aged cells along the trajectory. We observed a pronounced increase in expression of the two proliferation-division genes, *Ccnb1* and *Mki67*, in young HSPCs that was occurring in state 1 concomitant to the increase of the marker of differentiation *Cd48* (Fig. 6g). In aged cells, increase in the expression of these three genes was also detected but was delayed until the branching point pMye suggesting a delay in the commitment of aged HSPCs. To grasp molecular mechanism(s) that could be involved in this delay, we compared cell cycle inhibitor expression across young and aged HSPC trajectories. *Cdkn1a* and *Cdkn2c* were upregulated along the aged trajectory (except in state 4 for *Cdkn2c*) especially in the first part of the trajectory

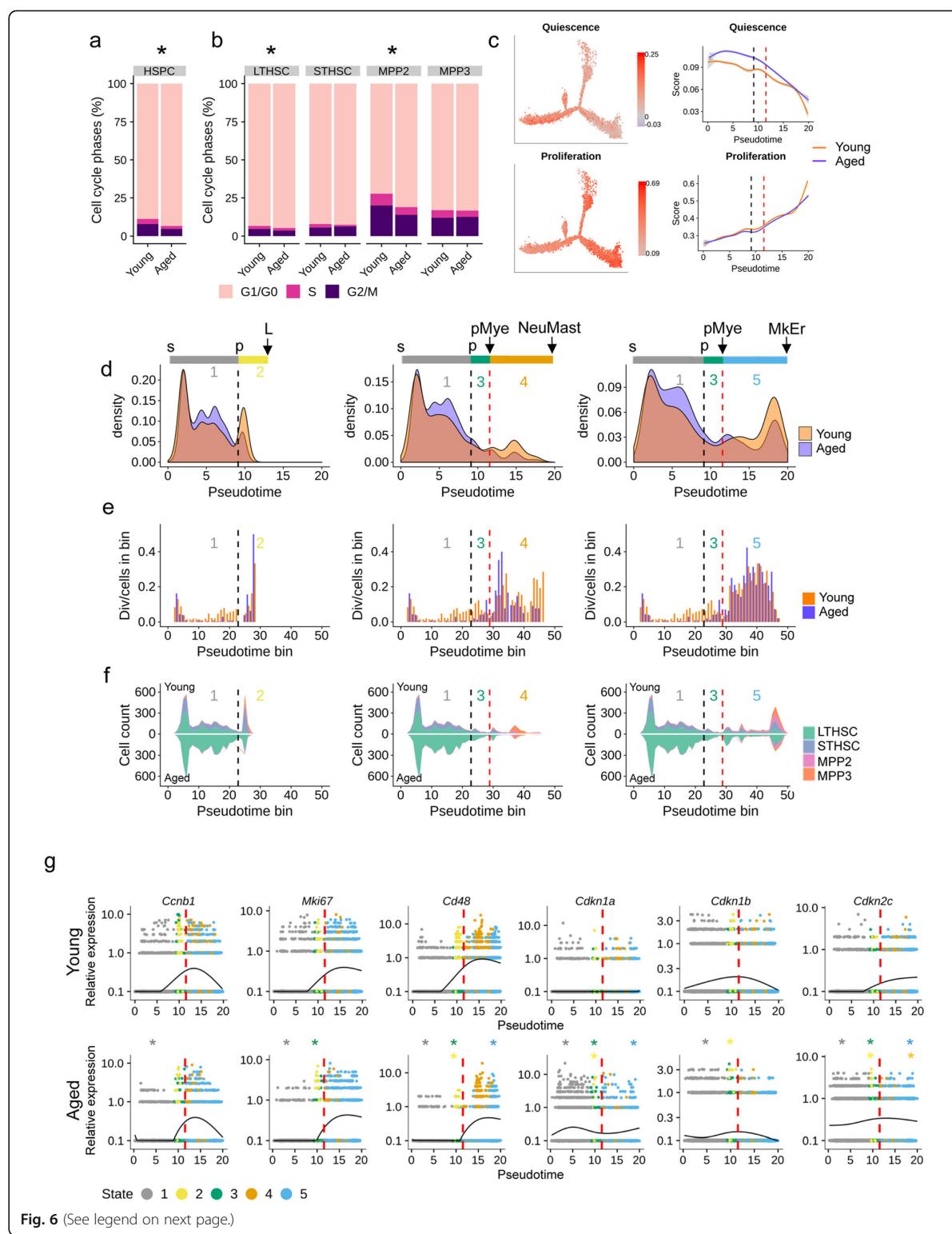
(states 1, 2, and 3) in comparison to young one. By contrast, *Cdkn1b* was downregulated in states 1 and 2 of the aged trajectory (Fig. 6g and Supplemental Table S10B). The change in expression with aging of the three cell cycle inhibitors known to control HSC fate indicates deregulation of cell cycle phases in aged HSCs. It is interesting to note that *Cdkn1a* was found to be a target of Stat1, Jun, and Junb which are themselves targets of the Klf6 regulon (Additional file 10: Table S9A), four regulons whose activities increased with aging in the same range of pseudotime as changes in the level of *Cdkn1a* expression (Figs. 6g and 5b).

Together, these results suggest that aged HSCs have a default in cell cycle, concomitant to a delay in their differentiation program, which occurs before the lineage priming of the HSPCs.

## Discussion

In this study, we questioned the effect of aging on the heterogeneity of HSCs and their properties using scRNA-seq, which provides a powerful method for defining cell subtypes as well as a detailed description of the functional properties specific to these subtypes [42].

At first, the large number of cells analyzed provided us new insights of HSPC heterogeneity, through the identification of 15 distinct HSPC clusters that we divided in two categories, the non-primed clusters by opposition to the lineage-primed clusters composed of low-abundant HSPCs with restricted lineage potential. We identified distinct lineage-primed HSPCs such as HSPCs with mastocytes, neutrophils, erythrocytes, and lymphoid-restricted lineage signatures in addition to the previously reported Mk-restricted HSPCs [8, 11, 27]. The lineage potentials of HSPCs detected in this study, which favors an early HSPC uni-lineage segregation [11, 43], may be the result of our cell cycle correction that diminish cell cycle gene expression noise, a dominant source of transcriptional heterogeneity in the HSC compartment [44]. Our pseudotemporal reconstruction of differentiation trajectories together with our clustering and transcriptional activity analyses highlighted a clear separation in the fates of specific lineage-primed HSPCs and clearly



(See figure on previous page)

**Fig. 6** Cell cycle analysis along pseudotime highlights a delay in differentiation associated with cell cycle arrest in aged condition. **a** Repartition (in percentage) of the cell cycle phases (estimated with cyclone) in young and aged HSPCs. **b** Repartition (in percentage) of the cell cycle phases (estimated with cyclone) in LTHSCs, STHSCs, MPP2 and MPP3 in young and aged conditions. For **a** and **b**, \* indicates a significant dependence between cell cycle phase and age repartitions ( $p$  value < 0.05 Pearson's chi-squared test). **c** Left panel, differentiation trajectory of HSPCs colored in accordance to their score for previously published quiescence and proliferation signatures. Right panel, comparison of the scores for the quiescence and proliferation signatures between young and aged HSPCs in pseudotime. **d** Density plot of young (orange) and aged (purple) cells along pseudotime for the T (left), NeuMast (middle), and MkEr (right) fates. Black and red dashed lines mark respectively p and pMye bifurcation points. **e** Division rate along pseudotime for young (orange) and aged (purple) HSPCs for the T (left), NeuMast (middle), and MkEr (right) fates. On x-axis, pseudotime was cut into 50 bins and a division rate is calculated for each bin, by dividing the number of young (resp. aged) cells assigned to G2M phase by the total number of young (resp. aged) cells of the bin. Black and red stretched lines mark p and pMye pseudotime bifurcation point respectively. **f** Stacked plot of predicted cell types along pseudotime cut into 50 bins for young (upper part of the plots) and aged (lower part of the plots), for the L (left), NeuMast (middle) and MkEr (right) fates. Black and red stretched lines mark p and pMye bifurcation point pseudotime respectively. **g** Smoothed gene expression along pseudotime of selected markers for young (upper panel) and aged (lower panel) HSPCs. Points represent cells, which are colored according to their belonging to the 5 different states (1 gray, 2 yellow, 3 green, 4 orange and 5 blue). The y-axis is in log scale. \* indicates significant differences in gene expression between young and aged cells ( $p$  value < 0.05) and star color indicates the state where the difference is found

characterized two bifurcation points, revealing three distinct HSPC fates towards Lymphocyte, Neu/Mast, or Mk/Er lineages. In addition, our analysis showed that lineage priming of HSPCs is not delineate by a specific HSPC subset such as LTHSC, STHSC, MPP2, and MPP3 in any instance. Although we showed that Neu priming is clearly stemming from MPP3 and Er priming from MPP2, lineage priming could also arise from a combination of HSPC subsets. This is the case for Mk priming which stems from LTHSC and MPP2 subsets, in line with previous studies [11, 23]. It is also the case for B and T lymphoid potential found in the four subsets of HSPCs, suggesting a lymphoid-priming occurring earlier in the BM and not restricted to the more engaged Flt3-positive MPP4 as previously reported [45]. The fact that we detected lineage primed cells in the very early subset of HSPCs goes in line with a previous studies showing the existence of four distinct and closely related stages of self-renewing LTHSCs in adult BM that stably adopt lineage-restricted fates (platelet, B and T lymphoid, erythroid, and myeloid lineages) despite remaining multipotent [46]. In addition, we were able to distinguish subtle differences in the LT-HSC compartment. Our study delineates a discrete hierarchy of differentiation within the LTHSCs, thanks to the pseudotime and cluster characterization, which posits the np2 cluster as being the most immature one.

If the accumulation of very immature HSCs in the BM of aged individuals is now an accepted criterion of hematopoietic aging, we still do not fully understand what are the characteristics of these aged HSCs and what causes them to accumulate. By looking at the transcriptomic changes at the single cell scale, we confirmed the global increase of the LTHSC fraction within the HSPCs. However, by analyzing our aged HSPCs by clusters or individually, we could demonstrate that HSPCs are not affected uniformly by aging and grasp some interesting aging feature. At first, we showed that the proportion of aged

HSPCs in pMast, pNeu pEr, and pL1-primed clusters was decreased while increased in ifn, tgf, np1, and np2 clusters. In addition, young and aged cells were found in the expected ratio in the pMk cluster. This clearly indicates that the platelet and myeloid bias observed upon aging [23] is not due to an amplification of the pool of lineage-primed HSPCs but stem from other HSPC subsets. Secondly, we highlighted some specific amplification of LTHSCs such as LTHSCs with miss-regulated interferon signaling (ifn cluster). As the increase in interferon response with aging in a number of different tissues has been observed [47] and is consistent with the concept of inflammaging [48], this amplification could afford for the myeloid bias observed in aging. Another interesting HSC group that we detected amplified during aging was the cluster of LTHSCs presenting a Tgf signature that may correspond to the accumulation of the HSC subtypes with differential responses to Tgf that was previously identified [49]. These two types of aged HSCs need to be further analyzed but considering their characteristics, it is tempting to hypothesize that their proportion was increased under stress selection pressures to compensate for the loss of mature cell production that occurs upon aging. Moreover, cluster amplification during aging has not been observed in the same way in our different animals, this is particularly true for the amplification of HSCs marked for Tgf, driven by a batch of aged mice. This heterogeneity of aging might witness the emergence of competitive clones that amplify during aging and fit quite well with the clonal hematopoiesis model. In another perspective, the apparition of the pL2-primed cluster that we observed quasi exclusively in the aged BM might also represent clonal evolution. Since this cluster was characterized with the expression of *Trp53imp1*, a gene limiting B-lymphoid differentiation upon aging, it could correspond to an accumulation of aged HSPCs altered in their lymphoid differentiation [28] but resulting for a pressure of immune deficiency.

Pseudotime trajectory analysis led us to address the question concerning the differentiation state of aged LTHSCs, which were thought to accumulate in a more undifferentiated state compared to young LTHSCs [24]. First, the outcome of our analyses is in favor of no difference in term of differentiation state between young and aged LTHSCs as when plotted together along the trajectory the most immature aged cells were not positioned at an anterior pseudotime compared to the young ones. Second, our results support that aged LTHSCs are delayed in their differentiation journey in comparison to young ones and that this delay occurs pretty early in the pseudotime, before the first bifurcation point that splits lymphoid fate from myeloid fate. This was clearly emphasized by our regulon activity analysis of transcription factors such as Myc, Trp53, or Spi1 that were previously described involved in multipotency and commitment of HSCs [50] and for which we could observe a delay in their activity along the differentiation trajectory.

Thus, the aged HSCs are not more undifferentiated than the young ones but seem to have intrinsic defaults that would delay their commitment. This finding is interesting when putting in perspective what causes the accumulation of LTHSCs. Increase of LTHSCs with aging could originate from an increase in the self-renewal rate of HSCs or/and from a blockade or at least a slowdown of the LTHSCs along their differentiation journey. It was also hypothesized that label-retaining HSCs (LR-HSCs), which divide minimally over time, accumulate in aged BM after completing four traceable symmetric self-renewal divisions to expand its size before entering a state of dormancy [51]. Although we could not directly address the question of self-renewal, we can argue based on our regulon and cell cycle analyses that aged LTHSCs have kept their capacity to self-renew and have not reached a state of complete dormancy but have reduced their proliferation linked to differentiation. Interestingly, we could associate this reduced and age-related proliferation/differentiation potential to a high level of Mycn activity, known to contribute to the stemness and self-renewal of different stem cells [52] and a high level of Gfi1b activity known to promote self-renewal of HSCs [53].

One interesting outcome of our analysis is the link between the delay in differentiation and cell cycle activity changes of aged HSPCs. We deduced from our computational cell cycle classification that lineage-primed HSPCs were less in G1/G0 than the non-primed LTHSCs. This observation is fully consistent with current knowledge that the most undifferentiated HSCs reside in the G0 phase and cycle infrequently and that cell cycle overall becomes more frequent as HSCs are gradually committed [41, 54]. In addition, we detected an increase in HSCs in G1/G0 phases in older BMs and an increase

in older LTHSCs in G1/G0 phases compared to younger LTHSCs, which reflects the decrease in cell cycle activity of older HSCs when considered as a whole [55]. Finally, when calculating a division rate per cells and studying division gene expression along the trajectory, we could detect a loss of aged dividing HSPCs located before the first bifurcation of the differentiation trajectory. These cells partially overlap in our trajectory with the div cluster, marked by genes related to asymmetric division such as *gpm2*, *Ragcap*, and *Ccnb1* [56, 57], suggesting that the delay in differentiation could be linked to an altered capacity of aged HSPCs to divide asymmetrically. In addition, gene expression of cell cycle inhibitors clearly shows that HSPCs in the first part of the trajectory have increased expression in *Cdkn1a* and *Cdkn2c*, promoters of quiescence but a reduction in *Cdkn1b* activation, which promotes commitment [58]. Interestingly, our analysis pointed out *Cdkn1a* as a direct target of Junb, itself target of Klf6. As the activation of *Cdkn1a* by Junb has been previously described to limit hematopoietic stem cell proliferation [59] and as Klf6 is a key factor in Tgfbeta signaling pathway [60, 61], our work unveils an interesting pathway controlled by the cytokine Tgfbeta involving Klf6 as a key regulon and *Cdkn1a* as a cell cycle regulator that is enhanced upon aging, endorses quiescence, and limits HSC differentiation.

## Conclusions

Our single-cell transcriptome-based identification of cell identity and its modifications associated with aging provides new information on cellular heterogeneity and intrinsic changes that will be useful for future investigation of the role of other regulators on the aged HSC phenotype.

## Methods

### Mouse model and cell sorting

C57BL/6-CD45.2 mice were purchased from Charles River Laboratories and were aged at the CRCM animal facility under specific pathogen-free conditions and handled in accordance with the French Guidelines for animal handling (Agreement #02294.01). Only males were analyzed, at 2–3 months (young) and 17–18 months (aged) of age. HSPCs were collected from the BM of 5 young and 5 aged mice over 2 independent batches with cells from 2 pooled young (Young\_A sample) and 3 pooled aged (Old\_A sample) mice for one batch, and cells from 3 pooled young (Young\_B sample) and 2 pooled aged (Old\_B sample) mice for the other one (Supplemental Table S1). For each sample, the BM was lineage depleted by using the Lineage Cell Depletion Kit (Miltenyi Biotec) and labeled with the following antibody cocktail: anti CD45.2, anti Sca-1, anti-cKit, anti CD150, anti Cd48, anti Cd34, and anti Flt3 antibodies

(Additional file 12: Table S11) to purify Lin-Sca1+cKit+ Flt3 cells (HSPCs) by multi-parameter fluorescence-activated cell sorting (FACS) on a FACSAriaII (Special Order Research Products; BD Biosciences). Flow cytometry analyses were performed using a BD-LSRII cytometer and analyzed using BD-DIVA Version 6.1.2 software (Special Order Research Products; BD Biosciences).

#### Single cell RNA-seq and data processing

We used the 10x genomics platform from two facilities: HalioDX for samples Young\_A and Old\_A (Marseille, France) and TGML for samples Young\_B and Old\_B (Marseille, France). In both facilities, FACS purified HSPCs were loaded 30 min after the sorting onto a Chromium Single Cell Chip and processed with the Chromium Controller (10x Genomics) according to the manufacturer's instructions for single cell barcoding at a target capture rate of 4000 individual cells per sample. Libraries were prepared using Chromium Single-Cell 3' Reagent Kits v2 (10x Genomics) and were sequenced using an Illumina NextSeq500 sequencer to an average depth of about 45,000 reads per cell for Young\_A and Old\_B samples and about 130,000 reads per cell for Old\_A and Young\_B samples. Cell ranger software v2.2 was used to align reads to the (GRCm38) mm10 mouse reference genome. Cell counts and transcript detection rates are summarized in Supplemental Table S1.

#### Quality control and data normalization

Cells outside 2 standard deviations (SDs) from the mean UMI log-counts were filtered out for each sample to discard poor quality cells and doublets. In total, 7433 young and 7482 aged cells were kept. For each dataset (our four samples and the Rodriguez-Fraticelli dataset), genes with no UMI count in more than 0.5% of the cells were discarded. All gathering, 17,513 genes were kept. Then, UMI counts were normalized with the NormalizeData Seurat function. For each cell, we considered the log transformation of the ratio of UMI counts per gene by the total UMI counts of the cell, multiply by a scaling factor of 10,000 ( $\log(10,000(\text{UMI}_{\text{gene}}/\text{UMI}_{\text{cell}}) + 1)$ ).

#### Cell cycle phase classification

Prediction of cell cycle phase for each cell was made with the cyclone [62], which relies on a pre-defined classifier for cell division constructed from a training dataset of synchronized mouse embryonic stem cells [63]. For each cell, a score based on raw count data before gene filtering was computed for each phase (G2/M, S and G1) and used to assign a phase to the cells. As quiescent HSCs are closer transcriptionally to G1 than S or G2/M cells of the cyclone training dataset, we classified them with the cyclone G1 cells and named this category G1/G0.

#### HSPC subtype assignment

In order to assign known FACS cell identity in our HSPC pool, we used CaSTLe (Classification of single cells by transfer learning), a supervised classification method consisting in labeling cells in a scRNA-seq experiment, using knowledge learnt from other experiments on similar subtypes [64]. We chose as source dataset a published scRNA-seq dataset obtained from FACS isolated HSPCs [11]. Cells from this data set (approximately 2000/per type) were divided into 4 subsets: the LTHSC (Lin- Sca1<sup>+</sup> Kit<sup>+</sup> Flt3<sup>-</sup> Cd150<sup>+</sup> Cd48<sup>-</sup>), the STHSC (Lin- Sca1<sup>+</sup> Kit<sup>+</sup> Flt3<sup>-</sup> Cd150<sup>-</sup> Cd48<sup>-</sup>), the MPP2s (Lin- Sca1<sup>+</sup> Kit<sup>+</sup> Flt3<sup>-</sup> Cd150<sup>+</sup> Cd48<sup>+</sup>), and the MPP3 (Lin- Sca1<sup>+</sup> Kit<sup>+</sup> Flt3<sup>-</sup> Cd150<sup>-</sup> Cd48<sup>+</sup>). HscScores were computed as previously described [35].

#### Dataset integration, data scaling, and cell cycle regression

To minimize batch effect between datasets, we integrated our 4 sample datasets (Young\_A, Young\_B, Old\_A, Old\_B) following the procedure of Seurat 3 [25]. Integration was done also for young and aged conditions separately. Briefly, the highly variable genes for each dataset were selected with the FindVariableFeatures function (selection.method = "vst") and ranked according to the number of datasets in which they were independently identified as highly variable. The top highly variable 2000 genes were thus integrated by iteratively merging pairs of datasets according to a given distance. Integration anchors, representing two cells that are predicted to originate from a common biological state in both datasets using a Canonical Correlation Analysis (CCA), were done using the FindIntegrationAnchors function (dims=1:15). Then, the expression of the target dataset was corrected using the difference in expression between the two expression vectors for each pair of anchor cells. This step was realized using IntegrateData function (dims=1:15). This process resulted in an expression matrix that contains the batch-effect-corrected expression for the 2000 selected genes of the cells from the 4 samples.

Standardized (i.e., centered and reduced) expression values with cell to cell variations due to cell cycle effect regressed were obtained with the ScaleData function of Seurat using the G2/M, S and G1/G0 scores previously computed for each cell by cyclone for the var.to.regress argument (cf Cell cycle phase classification).

#### Dimension reduction and clustering

A PCA was performed on the scaled data using RunPCA Seurat function (npc = 40). The 15 first principal components of the PCA were kept for nonlinear dimension reduction and cell clustering. Uniform Manifold Approximation and Projection (UMAP), [26], a nonlinear dimension reduction

method, was run using RunUMAP Seurat function package in order to embed cells in a 2-dimensional space. A K-nearest neighbor graph (KNN) based on the Euclidean distance in PCA space was constructed ( $k = 20$ ) to cluster the cells with the Louvain algorithm (resolution = 0.5) using the FindNeighbors and FindClusters Seurat functions respectively.

#### Pseudotime ordering

Unsupervised ordering of the HSPCs was done with the Seurat 3 integrated results as input to build a tree like differentiation trajectory using the DDRTree algorithm of the R package Monocle v2 [31]. Integrated data from (i) all samples (young and aged) excluding the primed pL2 cluster cells, (ii) young cells only, or (iii) aged cells only were processed with Monocle. For the three pseudotime ordering analyses (all cells, young only, and aged only), the 2000 gene expression matrix, scaled and regressed for cell cycle effect (see *Data scaling and cell cycle regression*) issued from the Seurat 3 integrated samples was loaded into Monocle using the newCellDataSet function (lowerDetectionLimit = 0.1, expressionFamily = uninormal()). The 2000 genes were set as ordering genes and trajectory building was made by calling the reduceDimension Monocle function (max\_components = 2, reduction\_method = 'DDRTree', norm\_method = "none", pseudo\_expr = 0). For each of the three trajectories, the root state was identified by selecting the Monocle state with the highest proportion of LTHSC predicted subtype (Fig. 4b; Additional file 1: Fig. S6B) in order to compute pseudotime values for the cells using the orderCells Monocle function. Expression of some genes as a function of pseudotime (Fig. 6g) was plotted with the plot\_gene\_expression Monocle function (using the Monocle normalization method with the estimateSizeFactor Monocle function).

#### Differential gene expression analyses

Specific markers for each cluster (Additional file 3: Table S2) and for each Monocle state (Additional file 9: Table S8B & C) were identified using FindAllMarkers Seurat function, with default parameters on log-normalized data without any cell cycle correction. Genes significantly overexpressed in one cluster/state versus all the others (positive markers) were tested with Wilcoxon rank sum tests on the log-normalized data of the given cluster against all the others. To further characterize state 2, which shared 72% of its markers with state 4, we identified DEGs between the two states using FindMarkers Seurat function (Additional file 9: Table S8C). Only genes expressed in at least 10% of the cells in either of the two groups (min.pct = 0.1) and with a log fold change threshold of 0.25 (logfc.threshold = 0.25) were tested. A  $p$ -adjusted value (Bonferroni correction)

threshold of 0.05 was applied to filter out non-significant markers.

Aging markers for the global population were obtained with the FindConservedMarkers Seurat function (min.pct = 0.1, logfc.threshold = 0) using the sequencing platform as grouping variable to minimize batch effect (Young\_A, Old\_A were processed on HalioDx platform and Young\_B, Old\_B on TGML platform). The Wilcoxon rank-sum test was performed on the log-normalized data between all young versus all aged cells (Additional file 6: Table S5) from each batch separately and the two  $p$  values for each gene were combined using the Tippett's method. Genes presenting an opposite variation between the 2 batches were filtered out.

Aging markers for each cluster (Additional file 7: Table S6) and for each Monocle state (Additional file 11: Table S10B) were obtained with the same method by looking at the difference cluster per cluster and state per state (min.pct = 0.1, logfc.threshold = 0.25 for each cluster and min.pct = 0, logfc.threshold = 0 for each state). No tests were performed in the pL2 cluster cells because it contained less than 3 cells in one young pool. From these results, for each cluster and each state only significant aging markers (combined  $p$  value  $< 0.05$  and same direction of variation in the 2 batches) were kept.

Among these markers the highly variable ones (average log fold change  $> 0.5$  with aging in at least one cluster in both batches) were selected to generate heatmap for all clusters with primed clusters gathered (Fig. 3a) and for primed clusters only (Additional file 1: Fig. S4) by adapting the DoHeatmap Seurat function. Genes (raw) were ordered using hclust R function on standardized aging gene expression of the subset. Euclidian distance and unweighted pair group method with arithmetic mean (UPGMA) were used. Up- and down-regulated genes with aging were ordered separately.

Volcano plots for the global aging markers were drawn (Additional file 1: Fig. S3) with EnhancedVolcano function from the R package of the same name [65].

#### Gene set enrichment analysis

To characterize the identified clusters with Seurat, we performed gene set enrichment analysis on cluster markers with gProfiler v0.6.7 [66] with default arguments except for background set to all genes expressed in the whole dataset (i.e., genes that passed filtering during quality control). We tested enrichments in GO terms (GO:BP, GO:MF, GO:CC) as well as in terms from KEGG, REAC, TF, MI, CORUM, HP, HPA, and OMIM databases (Fig. 1c and Additional file 4: Table S3). Cluster markers were also tested for enrichment in previously published gene set signatures related to HSPCs (Additional file 13: Table S12). Signatures tested were Bcell\_Chambers, Diff\_Chambers, Gran\_Chambers, HSC\_Chambers, Lymph\_Chambers,

Mono\_Chambers, Mye\_Chambers, NK\_Chambers, NaiveT\_Chambers, and Ner\_Chambers [32], lineage priming of HSC signatures C1, C2, C3, Mk, Er, Ba, Neu, Mo, Mo2, preDC, preB and preT [11], and HSCs and aging signatures Mm\_HSC\_Runx1\_Wu, Mm\_HSC\_Tcf7\_Wu [67], Mm\_LT\_HSC\_Venezia, Mm\_Proliferation\_Venezia, Mm\_Quiescence\_Venezia [68], Polarity\_factors\_Ting, Novel\_HSC\_regul\_polar\_Ting [57], HSC aging Svendsen [30] and MGA-MEP [34]. Cluster marker enrichment for the different signatures in comparison to all dataset genes was tested using a hypergeometric test (phyper R function). To perform enrichment analysis of aging markers with a consistent gene number, we gathered the overexpressed (resp. underexpressed) markers from at least one cluster and used gprofiler as described above (Additional file 8: Table S7A & B). Expression scores of the signatures or of selected aging features from the enrichment analysis were calculated for each individual cell using the AddModule-Score Seurat function (on log-normalized data) with default parameters, using as input the genes of the signatures or the aging markers annotated for the selected features. The Svendsen signature score was computed in the same way taking the aging markers common to our study and those of Svendsen's re-analysis [30].

#### Differential signature score analysis

Signature markers of Monocle state were tested in the same way as gene state markers (see above) using FindAllMarkers (min.pct = 0, logfc.threshold = 0) with Student's *t* tests. Only signatures with an average score differences above 0.015 between one state and all were kept. A *p*-adjusted value (Bonferroni correction) threshold of 0.05 was applied to filter out non-significant differences.

Signature score differences with aging in each state were tested in the same way as the aging markers per clusters (see above) using the FindConservedMarkers Seurat function (sequencing platform as grouping variable, min.pct and logfc.threshold set to 0) with Student's *t* tests. For each Monocle state, only average score differences of same sign and above 0.015 in the two batches presenting a combined *p* value  $< 0.05$  were kept (Additional file 9: Table S8A).

The selected aging features expression score differences with aging in each cluster were tested in the same way as the aging markers per clusters (see above) using the FindConservedMarkers Seurat function (sequencing platform as grouping variable, min.pct and logfc.threshold set to 0) with Student's *t* tests (Additional file 8: Table S7C). For each cluster, only average score differences of same sign and above 0.1 in the two batches presenting a combined *p* value  $< 0.05$  are considered as significant (Fig. 3b). No tests were performed in the

primed B cells clusters because it contained less than 3 cells in one young pool.

#### Regulon analysis

pySCENIC (1.10.0) was used with its command line implementation [38]. The raw expression matrix for the cells of all samples was filtered, by keeping genes with a total expression greater than  $2 * 0.01 * (\text{number of cell})$ . 10,698 genes passed the filtering. pyscenic grn command was used with grnboos2 method and default options and a fixed seed to derive co-expression modules between transcription factors and potential targets. We used as input all the markers of the Seurat clusters for which a transcription factor binding motif was available in the motifs-v9-nr.mgi-m.0.001-o.0 database provided by Scenic, plus several TFs involved in early hematopoiesis, Spi1, Tal1, Zfp1, Cbfa2t3, Erg, Fli1, Gata1, Gata2, Hhex, Runx1, Smad6 [69], Gfi1b [70], and Zbtb16 [71]. The obtained modules were refined by pruning targets that did not have an enrichment for a corresponding motif of the TF with pyscenic ctx command with –maskdropouts option using the motif database motifs-v9-nr.mgi-m.0.001-o.0 and the cis-target database mm9-tss-centered-10 kb-7species.mc9nr. Only positive regulons (i.e., those with a positive correlation between the TF and its targets) were kept for downstream analysis (Additional file 10: Table S9A). AUCell scores (regulon activities) in each cell were computed with pycenic aucell command (default options). To be noted that number of target genes was highly variable from a regulon to another (Additional file 1: Fig. S9).

For young and aged HSPCs, two heatmaps of regulon activity scores, along pseudotime, were made in order to analyze transcriptional activity at the two bifurcation points for both ages. See Additional file 10; Supplementary methods for detailed regulons heatmaps construction.

#### Analysis of HSPC subtypes and cell cycle phases in the differentiation trajectory depending on age

Cell density (Fig. 6d), division rate (Fig. 6e), and stacked plot of HSPC subtypes (Fig. 6f) were computed and plotted along pseudotime at each age for the 3 HSPC fates: the lymphoid (Monocle states 1 and 2), the Mastocytes/Neutrophils (Monocle states 1, 3, and 4), and the Megakaryocytes/Erythrocytes (Monocle states 1, 3, and 5). For division rate and stacked plot of HSPC subtypes, pseudotime was cut into 50 bins. For each age, in each pseudotime bin, division rate was computed as the ratio of the number of cells with a G2/M phase assigned to the total number of cells in the bin.

#### Statistics

Statistics were computed with R software v3.5.1. The statistical tests for gene expression and signature or

regulon activity scores were performed with Seurat and are detailed above. In each cluster and in non-primed/primed clusters gathered, the enrichment of age was tested using a hypergeometric test (phyper R function Fig. 2b). Chi<sup>2</sup> tests (chisq.test R function) were performed to test independence between cell cycle phase and age, in all cells (Fig. 6a) and in each HSPC subtype separately (Fig. 6b), and in the cells at the departure of the trajectory (Pseudotime < 2, Additional file 11: Fig. S10) and to test independence between Monocle state and age in all Monocle states (Fig. 4i) and in the states 2, 3, and 5 only (Fig. 4j). Fisher's exact test (fisher.test R function) was performed to test independence between Monocle state and age in each Seurat cluster (Additional file 9: Supplemental Fig. S8B). Wilcoxon rank-sum test was used to test for median difference between pseudotime value distributions of young and aged cells (Fig. 4h). In each cluster, a linear regression was computed between the average log fold change (in the cluster) and the global (in all cells) average log fold change of the aging markers recovered in the cluster (lm R function Additional file 6: Supplemental Fig. S5). Smooth curves of module score expression in pseudotime through the different fates for young and aged cells were drawn for quiescence and proliferation signature (Fig. 6c) using the geom\_smooth function ggplot2 R package [72] with the gam function of mgcv R package [73].

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12915-021-00955-z>.

**Additional file 1. Supplementary Methods.** **Figure S1.** LTHSCs accumulate upon aging. (A) FACS profiles of young and aged HSPCs. (B-D) Cell type classification: Proportions of LTHSC, STHSC, MPP2 and MPP3 determined by FACS and by supervised classification with CaSTLe when considering (B) all HSPCs, (C) young and aged HSPCs separately and (D) the 4 samples separately. **Figure S2.** Representative gene markers used to identify HSPC clusters. Violin plots showing gene markers expressed by the 15 clusters revealed in the UMAP shown in Fig. 1b. The complete list of significantly up-and down-regulated genes for the 15 clusters is shown in Supplemental Table S2. **Figure S3.** Violin plots showing Ly6d and Trp53inp1 expression significantly up regulated in the pL2 cells cluster in comparison to the other cells ( $p\text{-value} < 0.05$  & log fold change > 0.25). **Figure S4.** Volcano plot of differential expression upon aging tested on all cells. Black dots indicate significant DEGs ( $p\text{-value} < 0.05$  and log fold change > 0.25). A total of 3362 genes were tested. **Figure S5.** Heatmap of the most significant differentially expressed genes upon aging ( $p\text{-value} < 0.05$  and log fold change > 0.5 in at least one cluster) in the 6 lineage-primed clusters revealed by the Seurat analysis (Fig. 1b). Gene expression is standardised across the entire dataset. **Figure S6.** Comparison of cluster gene expression changes with global gene expression changes upon aging. For each significant aging marker in a given cluster ( $p\text{-value} < 0.05$  and log fold change > 0.25), its global log fold change (logFC; x-axis) is plotted with its cluster log fold change (logFC; y-axis). For each cluster, a regression line is drawn in blue, formula is indicated at the left top corner with its square regression coefficient R<sup>2</sup>. **Figure S7.** (A) Monocle trajectories for young and aged HSPCs ordered separately. Cells are coloured according to their belonging to the 3 states (6 grey, 7 yellow, 8 blue) or to the pL2 cluster (brown). Both trajectories present a similar segregation between the lineage-primed HSPCs, with one bifurcation from LTHSC (state 6) towards

Neu/Mast-primed (NeuMast) HSPCs (state 7) and MkErprimed (MkEr) HSPCs (state 8). The bifurcation to the lymphocyte fate was not retrieved probably due to the reduction in pL1 cell number due to sample splitting. (B) Barplots representing the LTHSC, STHSC, MPP2 and MPP3 proportions in the three states. (C) Monocle trajectories of young and aged HSPCs coloured in accordance to their pseudotime values and representing their differentiation progression. (D) Repartition of the Seurat clusters along the pseudotime of young and aged HSPC trajectories. Box plots of pseudotime values are coloured according to the most represented state. (E) Repartition (in percentage) of the different states (6 to 8) of the trajectory for each Seurat cluster for young and aged HSPCs. **Figure S8.** Localization of the different Seurat clusters in Monocle trajectory. Cells belonging to a given cluster are coloured in orange for young and in purple for aged HSPCs.

**Figure S9.** Analysis of the hscScore according to Seurat clusters. Violin plots of hscScore distribution is presented in the 15 clusters.

**Figure S10.** Repartition of young and old HSPCs in Monocle pseudotime and in states per Seurat cluster. (A) Boxplots of Monocle pseudotime values of the young (dark) and aged (pale) cells from the different clusters obtained with Seurat (except pL2 cluster). Box plots showing medians are coloured according to the most represented state. (B) Comparison of Monocle state percentage in the different clusters between young (Y, dark colours) and aged (A, pale colours). Stars indicate a significant dependence between state repartition of the cells and age ( $p\text{-value} < 0.05$  Fisher's Exact Test). **Figure S11.** Number of targets recovered for each regulon identified with scenic. Y axis is in log scale. **Figure S12.** Young and aged HSPCs located at the very beginning of the trajectory cycled the same. (A) Highlight in the trajectory of the starting cells (coloured in black, pseudotime < 2). (B) Cell cycle phase prediction of young and aged starting cells highlighted in A. NS: no significant dependence between age and phase repartition ( $p\text{-value} > 0.3$  Pearson's Chi-squared test).

**Additional file 2: Table S1.**

**Additional file 3: Table S2.**

**Additional file 4: Table S3.**

**Additional file 5: Table S4.**

**Additional file 6: Table S5.**

**Additional file 7: Table S6.**

**Additional file 8: Table S7.**

**Additional file 9: Table S8.**

**Additional file 10: Table S9.**

**Additional file 11: Table S10.**

**Additional file 12: Table S11.**

**Additional file 13: Table S12.**

## Acknowledgements

We thank Dr. Lionel Spinelli for critically reading the manuscript and the code. We are grateful to the core flow cytometry and the animal facilities of the CRCM for providing supportive help and to the HaliodX and the TGML sequencing facilities for the single cell capture and sequencing. Computing resources for this study was provided both by the computing facilities DISC (Datacenter IT and Scientific Computing) of the Centre de Recherche en Cancérologie de Marseille and by the facilities of the Institut de Mathématiques de Marseille thanks to Dr. Olivier Chabrol.

## Authors' contributions

Conceptualization: E.R., E.D. Methodology: L.H., M.P., E.R., E.D. Acquisition of data and materials: L.H., M.P., N.P. Formal analysis: L.H., A.M. Writting: L.H., E.R., E.D. Review and editing, M.P. Supervision: E.R., E.D. Funding acquisition: E.R., E.D. All authors approved the final version.

## Funding

This work was partly supported by the Excellence Initiative of Aix-Marseille University – A\*MIDEX, a French "Investissements d'Avenir" program to ED and ER, Institut Thématisque Multi-Organisme-cancer, by l'Institut National du Cancer (grant number 20141PLBIO06 to ED) and the ARC foundation (PJA#20161204989 to ED). LH was the recipient of an interdisciplinary PhD

grant from Aix Marseille University; MP's postdoctoral fellowship was supported by the Fondation de France (#2017-00076284).

#### Availability of data and materials

The single-cell RNA-seq data generated here are available in the Gene Expression Omnibus database under accession code GSE147729 [74]. All R and python codes used for data analysis are integrated in a global snakemake workflow available at: [https://gitcrcm.marseille.inserm.fr/herault/schSC\\_herault](https://gitcrcm.marseille.inserm.fr/herault/schSC_herault) [75].

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare no competing interests.

#### Author details

<sup>1</sup>Epigenetic Factors in Normal and Malignant Hematopoiesis Team, Aix Marseille Université, CNRS, INSERM, Institut Paoli-Calmettes, CRCM, Marseille, France. <sup>2</sup>Aix Marseille Université, CNRS, Centrale Marseille, I2M, Marseille, France.

Received: 23 September 2020 Accepted: 8 January 2021

Published online: 01 February 2021

#### References

- Spangrude GJ, Heimfeld S, Weissman IL. Purification and characterization of mouse hematopoietic stem cells. *Science*. 1988;241(4861):58–62.
- Osawa M, Hanada K, Hamada H, Nakauchi H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science*. 1996;273(5272):242–5.
- Akashi K, Traver D, Miyamoto T, Weissman IL. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*. 2000; 404(6774):193–7.
- Adolfsson J, Mansson R, Buza-Vidas N, Hultquist A, Liuba K, Jensen CT, Bryder D, Yang L, Borge OJ, Thoren LA, et al. Identification of Flt3+ lymphomyeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell*. 2005;121(2):295–306.
- Oguro H, Ding L, Morrison SJ. SLAM family markers resolve functionally distinct subpopulations of hematopoietic stem cells and multipotent progenitors. *Cell Stem Cell*. 2013;13(1):102–16.
- Dykstra B, Olthof S, Schreuder J, Ritsema M, de Haan G. Clonal analysis reveals multiple functional defects of aged murine hematopoietic stem cells. *J Exp Med*. 2011;208(13):2691–703.
- Morita Y, Ema H, Nakauchi H. Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment. *J Exp Med*. 2010;207(6): 1173–82.
- Yamamoto R, Morita Y, Ooehara J, Hamanaka S, Onodera M, Rudolph KL, Ema H, Nakauchi H. Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. *Cell*. 2013; 154(5):1112–26.
- Naik SH, Perie L, Swart E, Gerlach C, van Rooij N, de Boer RJ, Schumacher TN. Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature*. 2013;496(7444):229–32.
- Paul F, Arkin Y, Giladi RA, Jaitin DA, Kenigsberg E, Keren-Shaul H, Winter D, Lara-Astiaso D, Gury M, Weinier A, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*. 2015;163(7):1663–77.
- Rodriguez-Fraticelli AE, Wolock SL, Weinreb CS, Panero R, Patel SH, Jankovic M, Sun J, Calogero RA, Klein AM, Camargo FD. Clonal analysis of lineage fate in native haematopoiesis. *Nature*. 2018;553(7687):212–6.
- Haas S, Trumpp A, Milson MD. Causes and consequences of hematopoietic stem cell heterogeneity. *Cell Stem Cell*. 2018;22(5):627–38.
- Zhang Y, Gao S, Xia J, Liu F. Hematopoietic Hierarchy - an updated roadmap. *Trends Cell Biol*. 2018;28(12):976–86.
- Laurenti E, Gottgens B. From haematopoietic stem cells to complex differentiation landscapes. *Nature*. 2018;553(7689):418–26.
- Geiger H, de Haan G, Florian MC. The ageing haematopoietic stem cell compartment. *Nat Rev Immunol*. 2013;13(5):376–89.
- Chung SS, Park CY. Aging, hematopoiesis, and the myelodysplastic syndromes. *Blood Adv*. 2017;1(26):2572–8.
- Sun D, Luo M, Jeong M, Rodriguez B, Xia Z, Hannah R, Wang H, Le T, Faull KF, Chen R, et al. Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal. *Cell Stem Cell*. 2014;14(5):673–88.
- Li X, Zeng X, Xu Y, Wang B, Zhao Y, Lai X, Qian P, Huang H. Mechanisms and rejuvenation strategies for aged hematopoietic stem cells. *J Hematol Oncol*. 2020;13(1):31.
- Cooper JN, Young NS. Clonality in context: hematopoietic clones in the marrow environment. *Blood*. 2017;130(22):2363–72.
- Beerman I, Bhattacharya D, Zandi S, Sigvardsson M, Weissman IL, Bryder D, Rossi DJ. Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *Proc Natl Acad Sci U S A*. 2010;107(12):5465–70.
- Yamamoto R, Wilkinson AC, Ooehara J, Lan X, Lai CY, Nakauchi Y, Pritchard JK, Nakauchi H. Large-scale clonal analysis resolves aging of the mouse hematopoietic stem cell compartment. *Cell Stem Cell*. 2018;22(4):600–7 e604.
- de Haan G, Lazare SS. Aging of hematopoietic stem cells. *Blood*. 2018; 131(5):479–87.
- Grover A, Sanjuan-Pla A, Thongjuea S, Carrelha J, Giustacchini A, Gambardella A, Macaulay I, Mancini E, Luis TC, Mead A, et al. Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. *Nat Commun*. 2016;7:11075.
- Kowalczyk MS, Tirosh I, Heckl D, Rao TN, Dixit A, Haas BJ, Schneider RK, Wagers AJ, Ebert BL, Regev A. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res*. 2015;25(12):1860–72.
- Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. *Cell*. 2019;177(7):1888–902 e1821.
- McInnes L, Healy J, Melville J: UMAP: uniform manifold approximation and projection for dimension reduction. In: Preprint at <https://arxiv.org/abs/1802.03426>, 2018.
- Sanjuan-Pla A, Macaulay IC, Jensen CT, Woll PS, Luis TC, Mead A, Moore S, Carella C, Matsuoka S, Bouriez Jones T, et al. Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature*. 2013; 502(7470):232–6.
- Zidi B, Vincent-Fabert C, Pouyet L, Seillier M, Vandevelde A, N'Guessan P, Poplineau M, Guittard G, Mancini SJC, Duprez E, et al. TP53INP1 deficiency maintains murine B lymphopoiesis in aged bone marrow through redox-controlled IL-7R/STAT5 signaling. *Proc Natl Acad Sci U S A*. 2019;116(1):211–6.
- Mann M, Mehta A, de Boer CG, Kowalczyk MS, Lee K, Haldeman P, Rogel N, Knecht AR, Farouq D, Regev A, et al. Heterogeneous responses of hematopoietic stem cells to inflammatory stimuli are altered with age. *Cell Rep*. 2018;25(11):2992–3005 e2995.
- Svendsen A, Yang D, Lazare S, Zwart E, Ausema A, de Haan G, Bystrykh L: A comprehensive transcriptome signature of murine hematopoietic stem cell aging. In: bioRxiv. preprint: <https://doi.org/10.1101/2020.08.10.244434>; 2020.
- Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods*. 2017;14(10):979–82.
- Chambers SM, Boles NC, Lin KY, Tierney MP, Bowman TV, Bradfute SB, Chen AJ, Merchant AA, Sirin O, Weksberg DC, et al. Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell*. 2007;1(5):578–91.
- Chen X, Deng H, Churchill MJ, Luchsinger LL, Du X, Chu TH, Friedman RA, Middelhoff M, Ding H, Tailor YH, et al. Bone marrow myeloid cells regulate myeloid-biased hematopoietic stem cells via a histamine-dependent feedback loop. *Cell Stem Cell*. 2017;21(6):747–60 e747.
- Su A, Wiltshire T, Batalov S, Lapp H, Ching K, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*. 2004;101(16):6062–7.
- Hamey F, Gottgens B. Machine learning predicts putative hematopoietic stem cells within large single-cell transcriptomics data sets. *Exp Hematol*. 2019;78:11–20.
- Lu YC, Sanada C, Xavier-Ferrucio J, Wang L, Zhang PX, Grimes HL, Venkatasubramanian M, Chetal K, Aronow B, Salomonis N, et al. The molecular signature of megakaryocyte-erythroid progenitors reveals a role for the cell cycle in fate specification. *Cell Rep*. 2018;25(8):2083–93 e2084.
- Muench DE, Grimes HL. Transcriptional control of stem and progenitor potential. *Curr Stem Cell Rep*. 2015;1(3):139–50.

38. Aibar S, Gonzalez-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geurts P, Aerts J, et al. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods.* 2017;14(11):1083–6.
39. Moroy T, Vassen L, Wilkes B, Khandanpour C. From cytopenia to leukemia: the role of Gfi1 and Gfi1b in blood formation. *Blood.* 2015;126(24):2561–9.
40. Pietras EM, Warr MR, Passegue E. Cell cycle regulation in hematopoietic stem cells. *J Cell Biol.* 2011;195(5):709–20.
41. Seita J, Weissman IL. Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip Rev Syst Biol Med.* 2010;2(6):640–53.
42. Yuan GC, Cai L, Elowitz M, Enver T, Fan G, Guo G, Irazarry R, Kharchenko P, Kim J, Orkin S, et al. Challenges and emerging directions in single-cell analysis. *Genome Biol.* 2017;18(1):84.
43. Notta F, Zandi S, Takayama N, Dobson S, Gan Ol, Wilson G, Kaufmann KB, McLeod J, Laurenti E, Dunant CF, et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* 2016, 351(6269):aab2116.
44. Tsang JC, Yu Y, Burke S, Buettner F, Wang C, Kolodziejczyk AA, Teichmann SA, Lu L, Liu P. Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells. *Genome Biol.* 2015;16:178.
45. Pietras EM, Reynaud D, Kang YA, Carlin D, Calero-Nieto FJ, Leavitt AD, Stuart JM, Gottgens B, Passegue E. Functionally distinct subsets of lineage-biased multipotent progenitors control blood production in normal and regenerative conditions. *Cell Stem Cell.* 2015;17(1):35–46.
46. Carrelha J, Meng Y, Ketty LM, Luis TC, Norfo R, Alcolea V, Boukarabila H, Grasso F, Gambardella A, Grover A, et al. Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. *Nature.* 2018; 554(7690):106–11.
47. Benayoun BA, Pollina EA, Singh PP, Mahmoudi S, Harel I, Casey KM, Dulken BW, Kundaje A, Brunet A. Remodeling of epigenome and transcriptome landscapes with aging in mice reveals widespread induction of inflammatory responses. *Genome Res.* 2019;29(4):697–709.
48. Xia S, Zhang X, Zheng S, Khanabadi R, Kalionis B, Wu J, Wan W, Tai X. An update on Inflamm-aging: mechanisms, prevention, and treatment. *J Immunol Res.* 2016;2016:8426874.
49. Challen GA, Boles NC, Chambers SM, Goodell MA. Distinct hematopoietic stem cell subtypes are differentially regulated by TGF-beta1. *Cell Stem Cell.* 2010;6(3):265–78.
50. Klimmeck D, Cabezas-Wallscheid N, Reyes A, von Paleske L, Renders S, Hansson J, Krijgsfeld J, Huber W, Trumpp A. Transcriptome-wide profiling and posttranscriptional analysis of hematopoietic stem/progenitor cell differentiation toward myeloid commitment. *Stem Cell Reports.* 2014;3(5): 858–75.
51. Bernitz JM, Kim HS, MacArthur B, Sieburg H, Moore K. Hematopoietic stem cells count and remember self-renewal divisions. *Cell.* 2016;167(5):1296–309 e1210.
52. Park CS, Lewis A, Chen T, Lacorazza D. Concise review: regulation of self-renewal in normal and malignant hematopoietic stem cells by Kruppel-like factor 4. *Stem Cells Transl Med.* 2019;8(6):568–74.
53. Zeng H, Yucel R, Kosan C, Klein-Hitpass L, Moroy T. Transcription factor Gfi1 regulates self-renewal and engraftment of hematopoietic stem cells. *EMBO J.* 2004;23(20):4116–25.
54. Cheung TH, Rando TA. Molecular regulation of stem cell quiescence. *Nat Rev Mol Cell Biol.* 2013;14(6):329–40.
55. Flach J, Bakker ST, Mohrin M, Conroy PC, Pietras EM, Reynaud D, Alvarez S, Diolaiti ME, Ugarte F, Forsberg EC, et al. Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells. *Nature.* 2014; 512(7513):198–202.
56. Zhu J, Wen W, Zheng Z, Shang Y, Wei Z, Xiao Z, Pan Z, Du Q, Wang W, Zhang M. LGN/mlnsc and LGN/NuMA complex structures suggest distinct functions in asymmetric cell division for the Par3/mlnsc/LGN and Galphai/LGN/NuMA pathways. *Mol Cell.* 2011;43(3):418–31.
57. Ting SB, Deneault E, Hope K, Cellot S, Chagraoui J, Mayotte N, Dorn JF, Laverdure JP, Harvey M, Hawkins ED, et al. Asymmetric segregation and self-renewal of hematopoietic stem and progenitor cells with endocytic Ap2a2. *Blood.* 2012;119(11):2510–22.
58. Hao S, Chen C, Cheng T. Cell cycle regulation of hematopoietic stem or progenitor cells. *Int J Hematol.* 2016;103(5):487–97.
59. Santaguida M, Schepers K, King B, Sabnis AJ, Forsberg EC, Attema JL, Braun BS, Passegue E. JunB protects against myeloid malignancies by limiting hematopoietic stem cell proliferation and differentiation without affecting self-renewal. *Cancer Cell.* 2009;15(4):341–52.
60. Botella LM, Sanz-Rodriguez F, Komi Y, Fernandez LA, Varela E, Garrido-Martin EM, Narla G, Friedman SL, Kojima S. TGF-beta regulates the expression of transcription factor KLF6 and its splice variants and promotes co-operative transactivation of common target genes through a Smad3-Sp1-KLF6 interaction. *Biochem J.* 2009;419(2):485–95.
61. Dhaouadi N, Li JY, Feugier P, Gustin MP, Dab H, Kacem K, Bricca G, Cerutti C. Computational identification of potential transcriptional regulators of TGF-ss1 in human atherosclerotic arteries. *Genomics.* 2014;103(5–6):357–70.
62. Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O, Marioni JC, Buettner F. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods.* 2015;85:54–61.
63. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol.* 2015;33(2):155–60.
64. Lieberman Y, Rokach L, Shay T. CaSTLE - classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS One.* 2018;13(10):e0205499.
65. Blighe K, Rana S, Lewis M: EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. In: <https://bioconductor.org/packages/devel/bioc/vignettes/EnhancedVolcano/inst/doc/EnhancedVolcano.html>; 2018.
66. Reimand J, Kull M, Peterson H, Hansen J, Vilo J: gProfiler—a web-based toolkit for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* 2007, 35(Web Server issue):W193–200.
67. Wu JQ, Seay M, Schulz VP, Hariharan M, Tuck D, Lian J, Du J, Shi M, Ye Z, Gerstein M, et al. Tcf7 is an important regulator of the switch of self-renewal and differentiation in a multipotential hematopoietic cell line. *PLoS Genet.* 2012;8(3):e1002565.
68. Venezia TA, Merchant AA, Ramos CA, Whitehouse NL, Young AS, Shaw CA, Goodell MA. Molecular signatures of proliferation and quiescence in hematopoietic stem cells. *PLoS Biol.* 2004;2(10):e301.
69. Bonzani N, Garg A, Feenstra KA, Schutte J, Kinston S, Miranda-Saavedra D, Heringa J, Xenarios I, Gottgens B. Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. *Bioinformatics.* 2013;29(13):i80–8.
70. Schutte J, Wang H, Antoniou S, Jarratt A, Wilson NK, Riepsaame J, Calero-Nieto FJ, Moignard V, Basilico S, Kinston SJ, et al. An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability. *Elife.* 2016;5:e11469.
71. Poplineau M, Vernerey J, Platet N, N'Guyen L, Herault L, Esposito M, Saurin AJ, Guilouf C, Iwama A, Duprez E. PLZF limits enhancer activity during hematopoietic progenitor aging. *Nucleic Acids Res.* 2019;47(9):4509–20.
72. Wickham H: ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York 2016.
73. Wood SN, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. *J Am Stat Assoc.* 2016;111(516):1548–63.
74. Herault L, Poplineau M, Mazuel A, Platet N, Remy E, Duprez E: Single-cell RNA-seq revealed a concomitant delay in differentiation and cell cycle of aged hematopoietic stem cells. In: *GEO*; 2020: GSE147729.
75. Herault L: Single-cell RNA-seq reveals a concomitant delay in differentiation and cell cycle of aged hematopoietic stem cells. In: [https://gitrcm.marseille.inserm.fr/herault/schSC\\_herault](https://gitrcm.marseille.inserm.fr/herault/schSC_herault); GitLab; 2020.

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## ADDITIONAL FILE 1

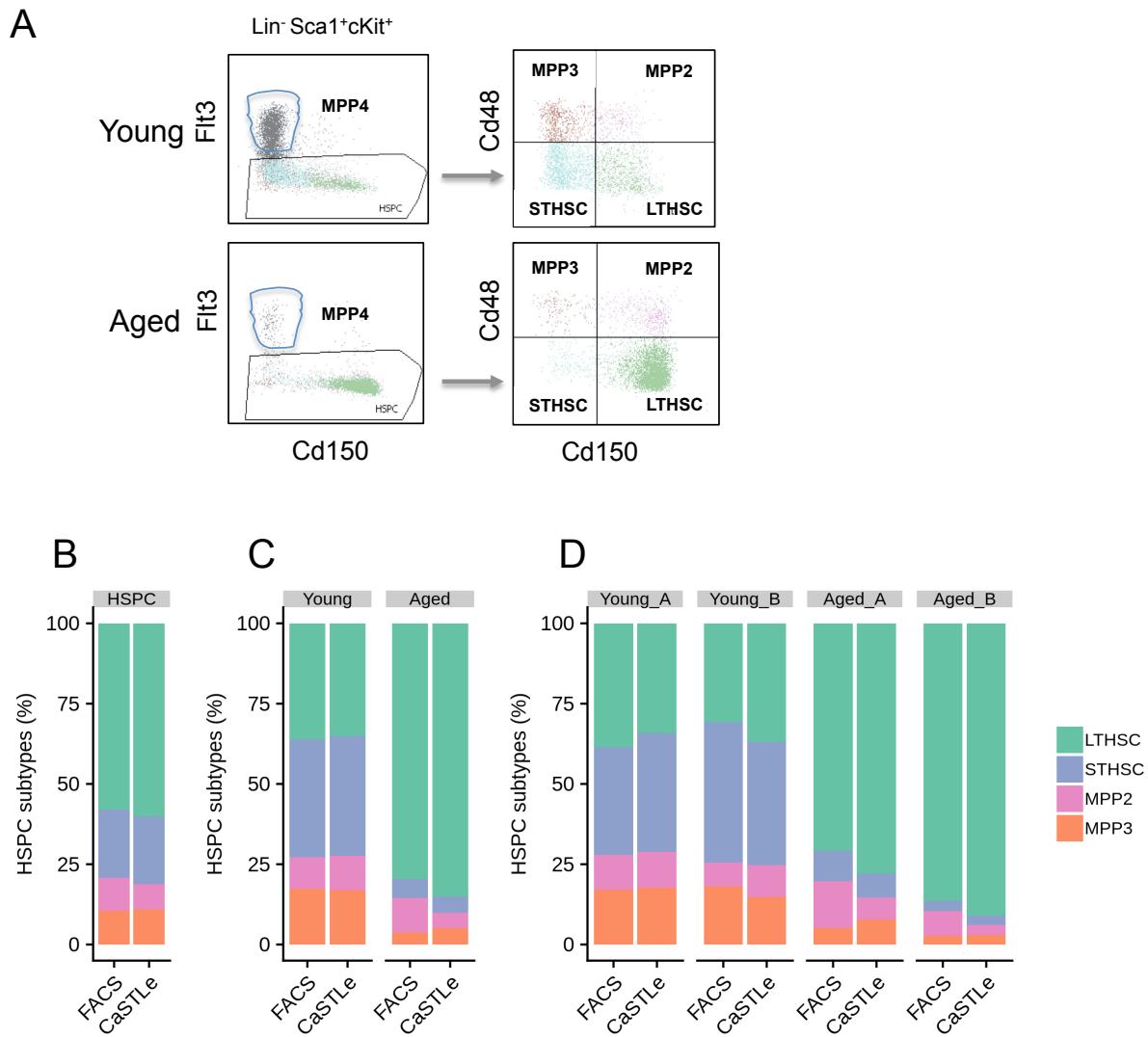
### SUPPLEMENTARY METHODS

#### Regulon heatmaps

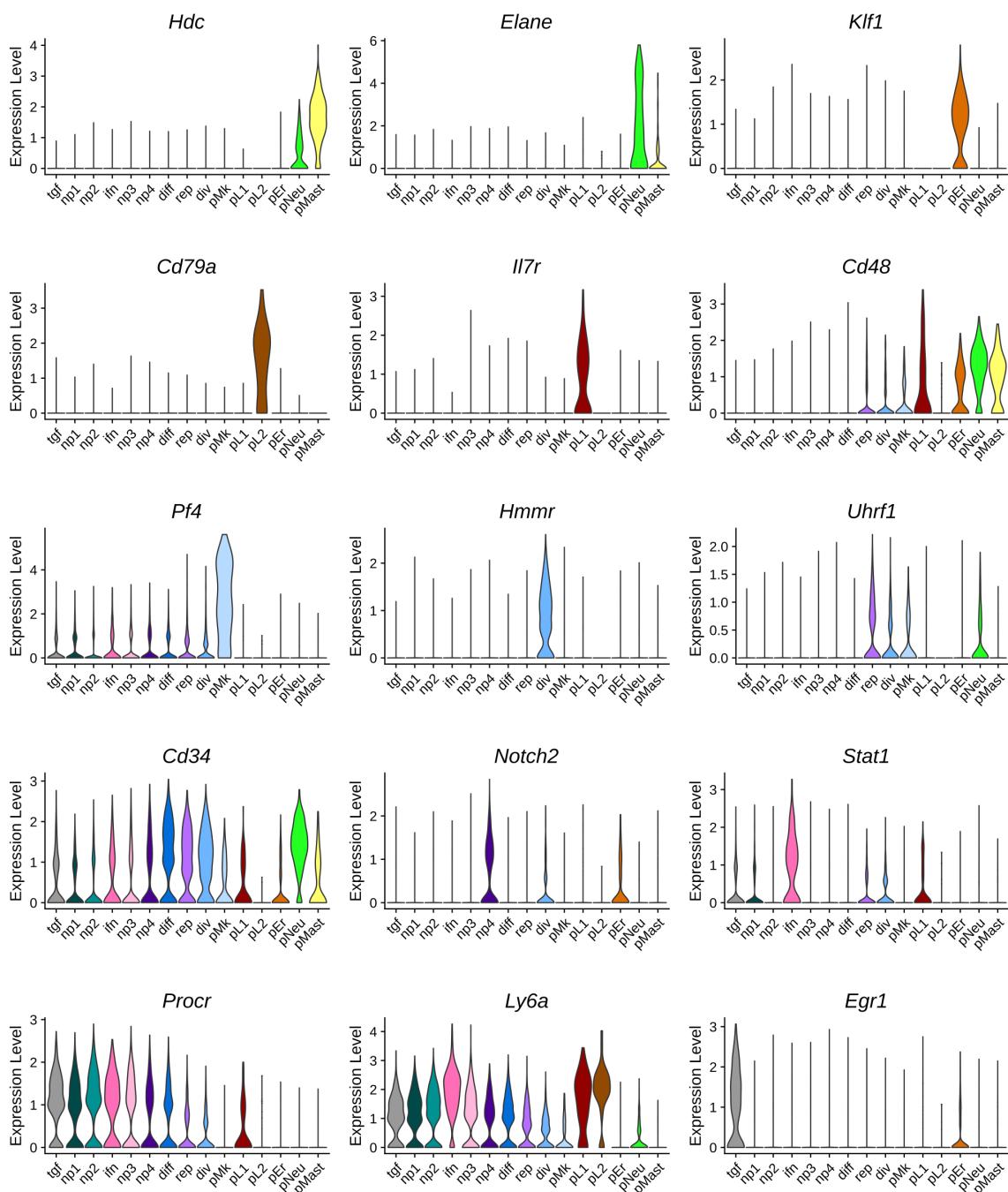
The first heatmaps measured the regulon activity from the departure of the trajectory (state 1) towards the ends of Monocle state 2 and Monocle state 3 (Figure 5a). The second ones measured the regulon activity from the departure of the trajectory (excluding state 2) towards Monocle state 4 and Monocle state 5 extremities (Figure 5b). Thus, each heatmap displays one bifurcation point and two paths. First, for each path at each age, a generalized linear model is fitted to the activity scores of each regulon as a function of pseudotime using the gam function of mgcv R package (1). Pseudotime is cut into 100 bins of equal length. For each bin and for each regulon the mean of the regulon activity score on all cells belonging to the bin is computed. The resulted matrix with data from the two paths for the two ages was scaled and regulons were hierarchically clustered on the young data subset thanks to the hclust R function using Euclidian distance and ward.D2 clustering method (4 clusters for the first and the second heatmaps). The regulon order obtained was then used to build the final heatmap on all the data with the pheatmap of pheatmap R package (2). Regulon markers of monocle states were tested in the same way as gene state markers (see above) with their AUCell scores using FindAllMarkers Seurat function (min.pct= 0.1, logfc.threshold=0) with Wilcoxon rank sum tests. Only regulons with an average AUCell score differences above 0.002 between one state versus all the others were kept. A *p*-adjusted value (Bonferroni correction) threshold of 0.05 was applied to filter out non-significant differences.

Regulon activity differences with aging in each state were tested in the same way as the aging markers per clusters using the FindConservedMarkers Seurat function (sequencing platform as grouping variable, min.pct = 0.1 and logfc.threshold = 0) with Wilcoxon rank sum tests. For each state, only average AUCell score differences of same sign and above 0.002 in the two batches presenting a combined *p* value < 0.05 were kept (Supplementary Table 9B).

- (1) Wood, S.N., Pya, N., and Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association* *111*, 1548-1563.
- (2) Kolde (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12. (<https://CRAN.R-project.org/package=pheatmap>).

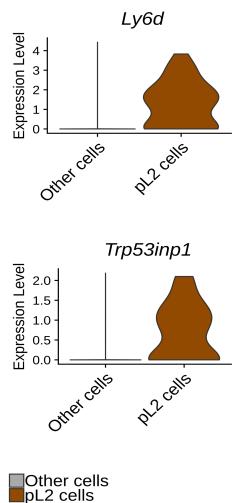


**Supplemental Figure S1: LTHSCs accumulate upon aging.** (A) FACS profiles of young and aged HSPCs. (B-D) Cell type classification: Proportions of LTHSC, STHSC, MPP2 and MPP3 determined by FACS and by supervised classification with CaSTLe when considering (B) all HSPCs, (C) young and aged HSPCs separately and (D) the 4 samples separately.

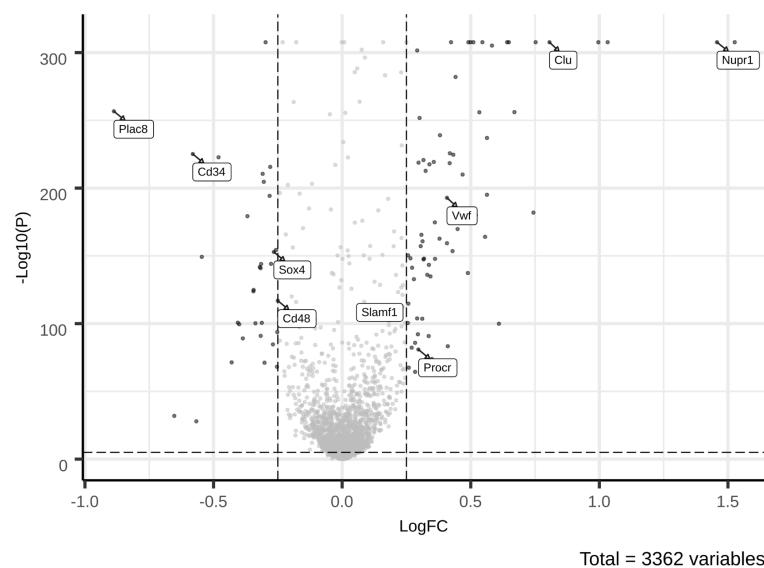


**Supplemental Figure S2. Representative marker genes used to identify HSPC clusters.**

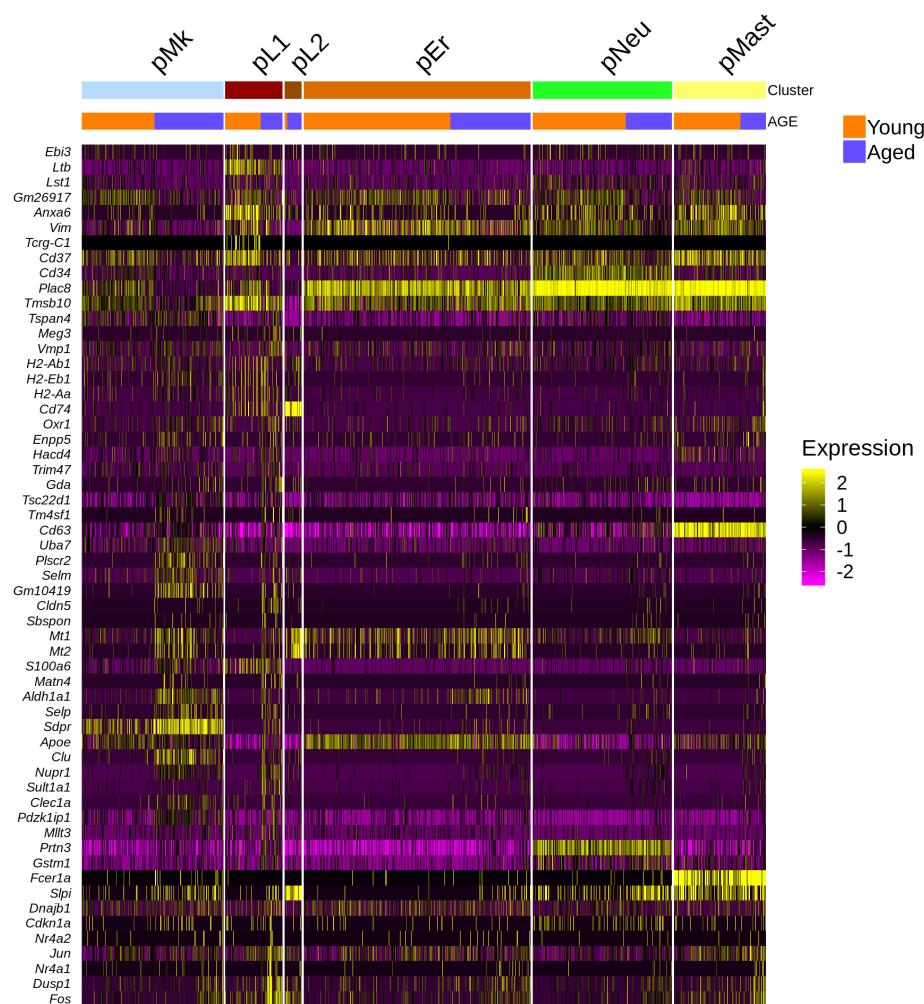
Violin plots showing gene markers expressed by the 15 clusters revealed in the UMAP shown in Fig. 1b. The complete list of significantly up-and down-regulated genes for the 15 clusters is shown in Supplemental Table S2.



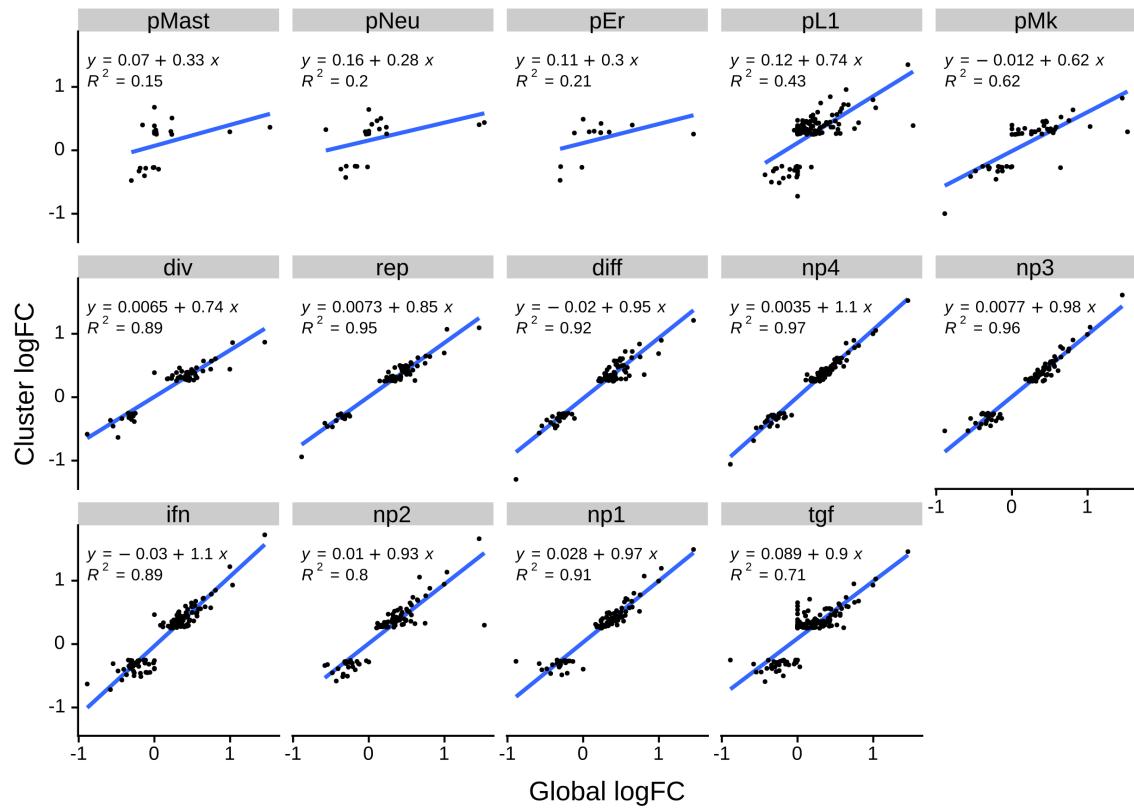
**Supplemental Figure S3.** Violin plots showing Ly6d and Trp53inp1 expression significantly up regulated in the pL2 cells cluster in comparison to the other cells ( $p\text{-value} < 0.05$  & log fold change  $> 0.25$ ).



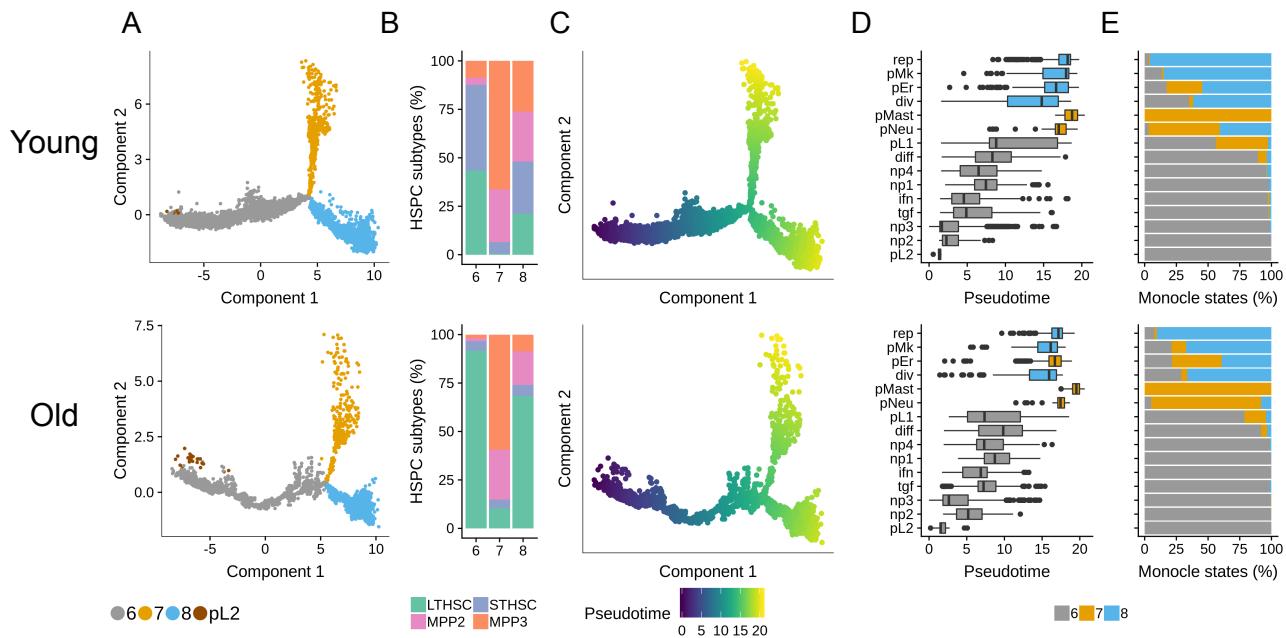
**Supplemental Figure S4.** Volcano plot of differential expression upon aging tested on all cells. Black dots indicate significant differentially expressed genes (DEGs;  $p\text{-value} < 0.05$  and log fold change  $> 0.25$ ). A total of 3362 genes were tested.



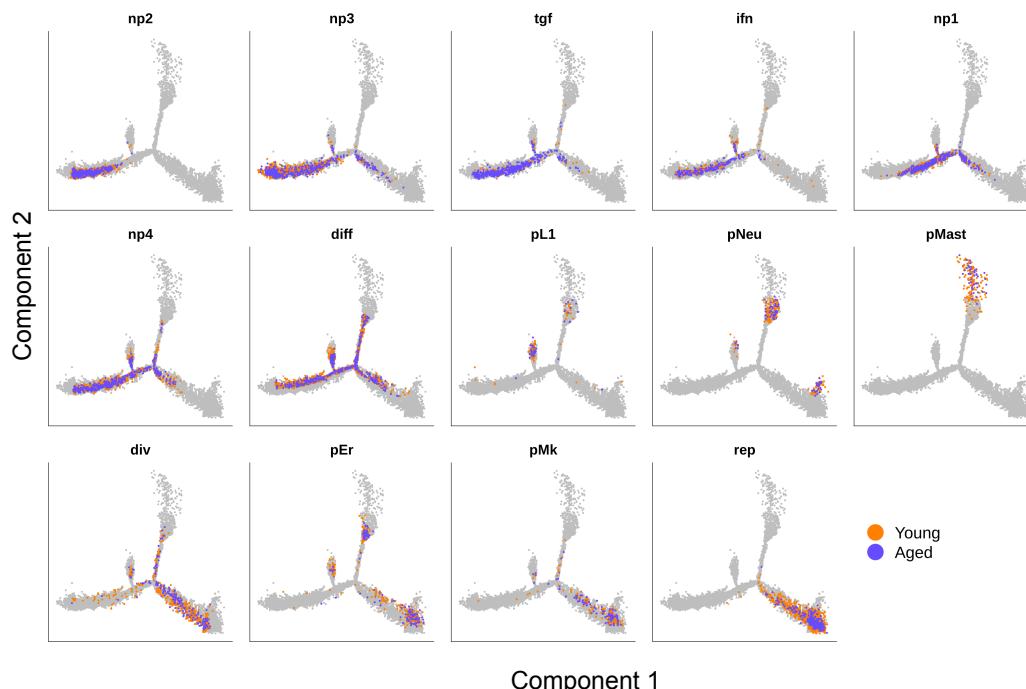
**Supplemental Figure S5:** Heatmap of the most significant DEGs upon aging ( $p\text{-value} < 0.05$  and log fold change  $> 0.5$  in at least one cluster) in the 6 lineage-primed clusters revealed by the Seurat analysis (Fig. 1b). Gene expression is standardised across the entire dataset.



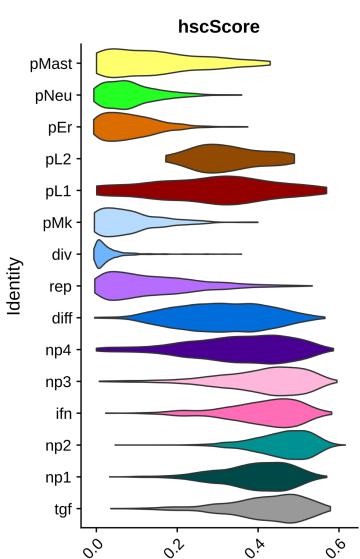
**Supplemental Figure S6. Comparison of cluster gene expression changes with global gene expression changes upon aging.** For each significant aging marker in a given cluster (p-value < 0.05 and log fold change > 0.25), its global log fold change (logFC; x-axis) is plotted with its cluster log fold change (logFC; y-axis). For each cluster, a regression line is drawn in blue, formula is indicated at the left top corner with its square regression coefficient  $R^2$ .



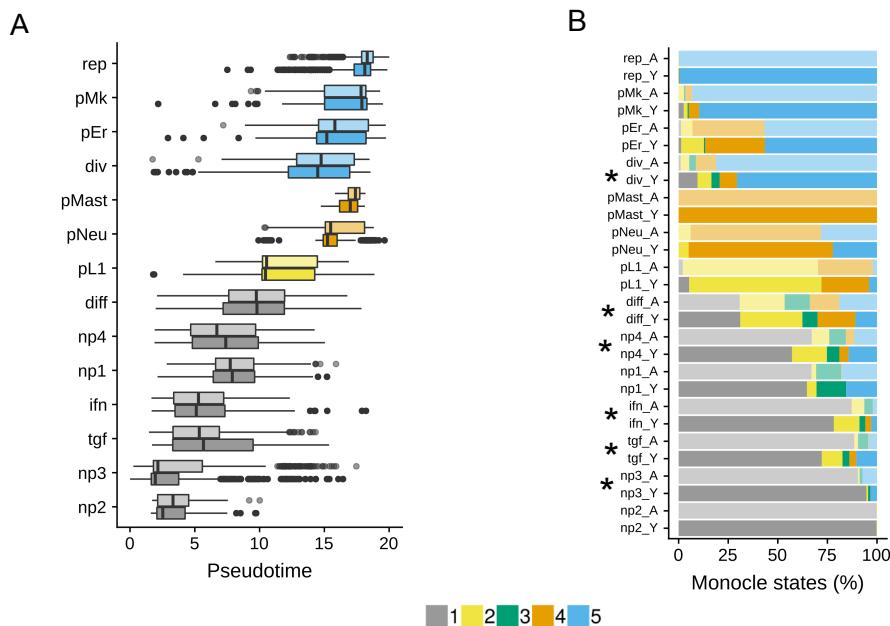
**Supplemental Figure S7.** (A) Monocle trajectories for young and aged HSPCs ordered separately. Cells are coloured according to their belonging to the three states (6 grey, 7 yellow, 8 blue) or to the pL2 cluster (brown). Both trajectories present a similar segregation between the lineage-primed HSPCs, with one bifurcation from LTHSC (state 6) towards Neu/Mast-primed (NeuMast) HSPCs (state 7) and Mk/Er-primed (MkEr) HSPCs (state 8). The bifurcation to lymphocyte fate was not retrieved, probably due to the reduction in pL1 cell number due to sample splitting. (B) Barplots representing the LTHSC, STHSC, MPP2 and MPP3 proportions in the three states. (C) Monocle trajectories of young and aged HSPCs coloured in accordance to their pseudotime values and representing their differentiation progression. (D) Repartition of the Seurat clusters along the pseudotime of young and aged HSPC trajectories. Box plots of pseudotime values are coloured according to the most represented state. (E) Repartition (in percentage) of the different states (6 to 8) of the trajectory for each Seurat cluster for young and aged HSPCs.



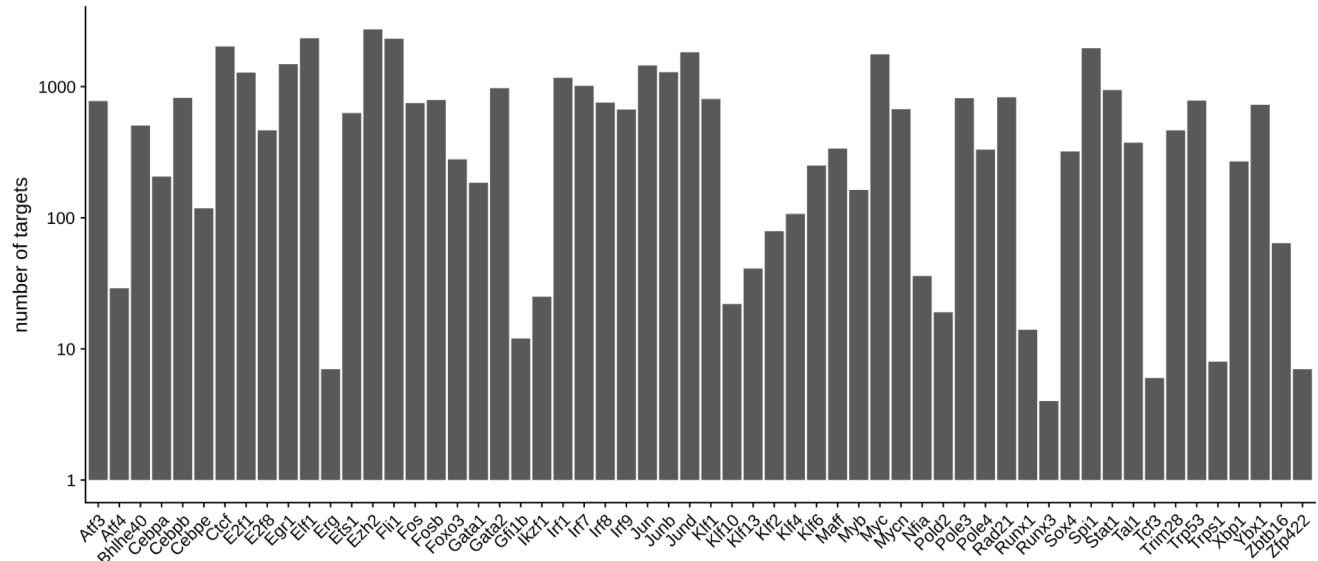
**Supplemental Figure S8. Localization of the different Seurat clusters in Monocle trajectory.**  
Cells belonging to a given cluster are coloured in orange for young and in purple for aged HSPCs.



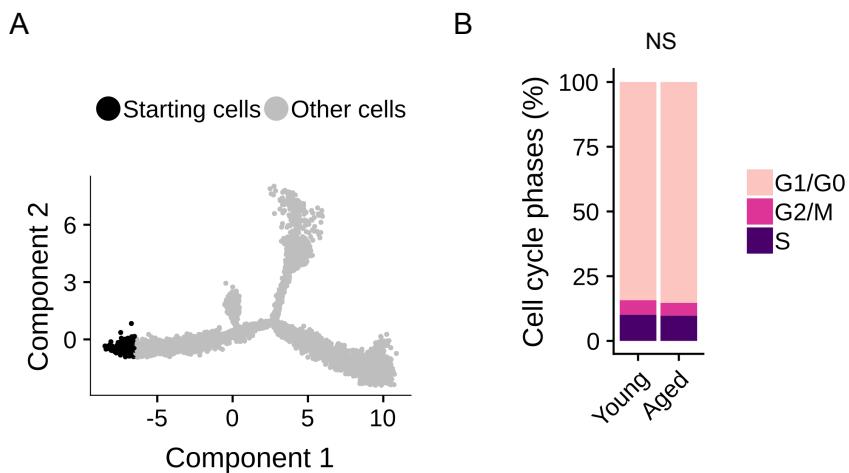
**Supplemental Figure S9. Analysis of the hscScore according to Seurat clusters.** Violin plots of hscScore distribution is presented in the 15 clusters.



**Supplemental Figure S10. Repartition of young and old HSPCs in Monocle pseudotime and in states per Seurat cluster.** (A) Boxplots of Monocle pseudotime values of the young (dark) and aged (pale) cells from the different clusters obtained with Seurat (except pL2 cluster). Box plots showing medians are coloured according to the most represented state. (B) Comparison of Monocle state percentage in the different clusters between young (Y, dark colours) and aged (A, pale colours). Stars indicate a significant dependence between state repartition of the cells and age ( $p$ -value  $< 0.05$  Fisher's Exact Test).



**Supplemental Figure S11.** Number of targets recovered for each regulon identified with scenic. Y axis is in log scale.



**Supplemental Figure S12. Young and aged HSPCs located at the very beginning of the trajectory cycled the same.** (A) Highlight in the trajectory of the starting cells (coloured in black, pseudotime < 2). (B) Cell cycle phase prediction of young and aged starting cells, highlighted in A. NS: no significant dependence between age and phase repartition (p-value > 0.3 Pearson's Chi-squared test).

## 3.4 Complément biais et corrections

Dans cette étude nous avons été confronté à deux biais importants que nous avons pu corriger en grande partie grâce aux développements récents des méthodes d'analyses scRNA-seq. Tout d'abord, nos expérimentations ont commencé il y a 4 ans et la technologie CITE-seq n'était pas encore disponible. Ainsi pour les 4 échantillons (deux lots d'un échantillon jeune et d'un échantillon âgé), les étapes de tri et de capture ont été menées en parallèle. De plus pour des raisons administratives les cellules des deux lots d'échantillons ont été capturées et séquencées sur des sites différents (A et B). En conséquence, l'utilisation du pipeline classique de seurat sans correction a montré un fort effet de lots avec des échantillons clairement distincts dans le UMAP qui se regroupent par sites de traitement (Figure 3.2.A). En utilisant la procédure d'intégration de Seurat il a été possible de corriger efficacement ce biais pour les analyses de réduction de dimension (UMAP présenté Figure 3.2.B, clustering, et pseudo-trajectoire).

Un autre biais important relevé dans les précédentes études scRNA-seq de HSPC est le fort bruit du cycle cellulaire, notamment pour les analyses de pseudo-trajectoires. En effet, nous avons pu observer que sans correction, les cellules assignées en phase G2/M et S par cyclone (SCIALDONE et al., 2015) avaient tendance à se regrouper majoritairement en fin de trajectoire (Figure 3.2.C) suggérant un ordre des cellules sur la trajectoire dicté par leur avancement dans le cycle cellulaire plus que par leur avancement dans la différenciation. Nous avons alors choisi de régresser les scores des phases du cycles sur les données d'expression utilisées pour les analyses de réduction de dimension (matrice d'expression des 2000 gènes les plus variable standardisée) selon les approches conseillées pour corriger ce biais (LUECKEN et THEIS, 2019). Nous avons obtenu une correction assez nette du bruit du cycle sur les pseudo-trajectoires construites par Monocle2 avec une répartition plus homogène des cellules en G2/M et S le long de la trajectoire (Figure 3.2.D). Cette correction a également abouti à l'observation de cellules très indifférenciées en début de trajectoire en division, que nous avons interprétées comme des CSH en autorenouvellement.

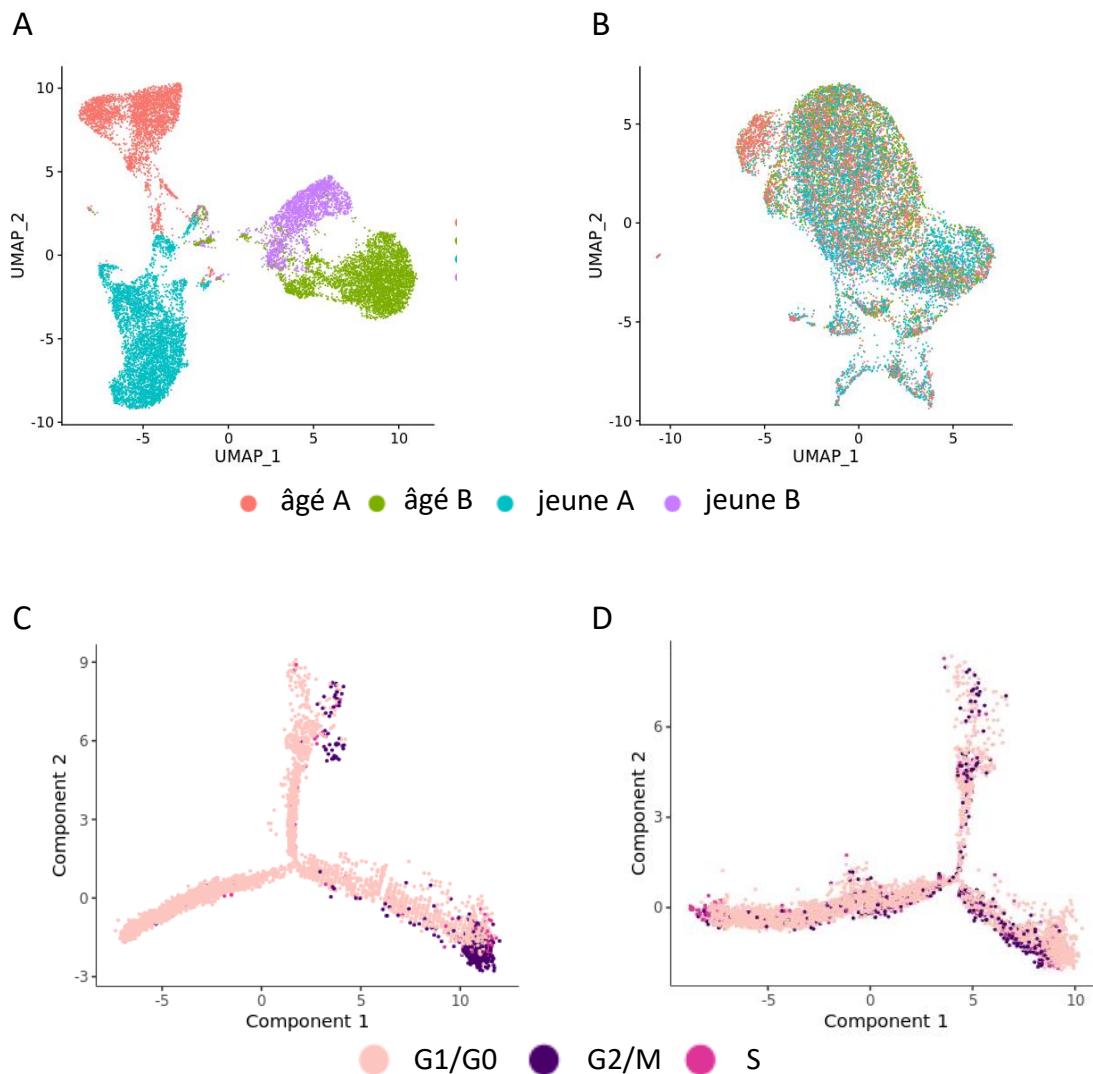


FIGURE 3.2 – Corrections des effets de lots et du bruit du cycle cellulaire

A : UMAP obtenu avec Seurat 3 sans la procédure d'intégration des échantillons. B : UMAP obtenu à partir des données des échantillons intégrés ensemble par Seurat3. C : Pseudo-trajectoire des HSPC jeune obtenue avec Monocle2 sans la régression du cycle cellulaire. D : Pseudo-trajectoire des HSPC jeune obtenue avec Monocle2 avec la régression du cycle cellulaire

## 3.5 Complément sur les pseudo-trajectoires de différenciation

Du fait de la disparité des résultats de pseudo-trajectoire obtenus d'une méthode à l'autre (SAELENS et al., 2019), il nous est apparu indispensable de confronter les résultats de Monocle2 avec au moins un autre algorithme d'inférence de trajectoire. Étant donné nos résultats intéressants concernant les différences de densité de cellules selon leur âge le long de la trajectoire, nous avons décidé d'utiliser STREAM une méthode qui vise tout particulièrement ce

genre d'analyse (CHEN et al., 2019). Les traj ectoires de STREAM sont établies en construisant des ensembles de graphes des cellules dans un espace réduit à l'aide de calculs d'énergie élastique. Nous avons utilisé STREAM\_v1 sur les mêmes données que Monocle2, c'est à dire les 15 premières composantes principales de la matrice d'expression des données intégrées des 4 échantillons (expression standardisée des 2000 gènes les plus variables, avec l'effet du cycle cellulaire corrigé). Nous avons utilisé les options par défaut de STREAM avec la réduction de dimension *Modified Locally Linear Embedding* (MLLE). Dans cette espace réduit de 3 dimensions, un premier graphe de 10 nœuds a été construit avec la fonction `seed_elastic_principal_graph` puis la traj ectoire a été obtenue avec la fonction `elastic_principal_graph`. Celle-ci a ensuite été ajustée avec la fonction `optimize_branching` et `extend_elastic_principal_graph`. Les fonctions `plot_flat_tree`, `plot_stream` et `plot_stream_sc` ont été utilisées pour la visualisation (figure 3.3.A).

La forme de la traj ectoire de STREAM obtenue est similaire à celle de Monocle2 en termes de nombre de branches et de points terminaux, et les pseudotemps des deux méthodes sont corrélés (figure 3.3.B&C). Les distributions des âges et des phases du cycle cellulaire le long de la traj ectoire STREAM sont proches de celles observées avec Monocle2 et on retrouve avant les embranchements d'amorçage une accumulation de CSH âgées quiescentes (figure 3.3.D&F). Cependant, les embranchements dans la traj ectoire STREAM sont un peu différents. Ils se produisent plus tôt dans le pseudotemps et sont moins clairs que dans la traj ectoire Monocle2 (Figure 3.3.A&B). Contrairement à Monocle2, STREAM fait partir les amorcages vers les lignages érythroïde et mégacaryocytaire en premier, dans le pseudotemps, tandis que les amorcages neutrophile/mastocytaire et lymphoïde se séparent après (figure 3.3.B&D).

Le hscScore que nous avons calculé précédemment montre que les HSPC amorcés lymphoïdes sont plus proches au niveau transcriptionnel des LTHSC que les autres HSPC amorcés (Figure S9 section 3.3; HÉRAULT et al., 2021). Ce résultat conforte davantage les résultats de Monocle2 que ceux de STREAM avec un amorçage lymphoïde survenant avant les amorcages érythroïde et mégacaryocytaire. Ainsi, nous avons décidé de nous appuyer sur les résultats de Monocle2 pour la suite de nos travaux.

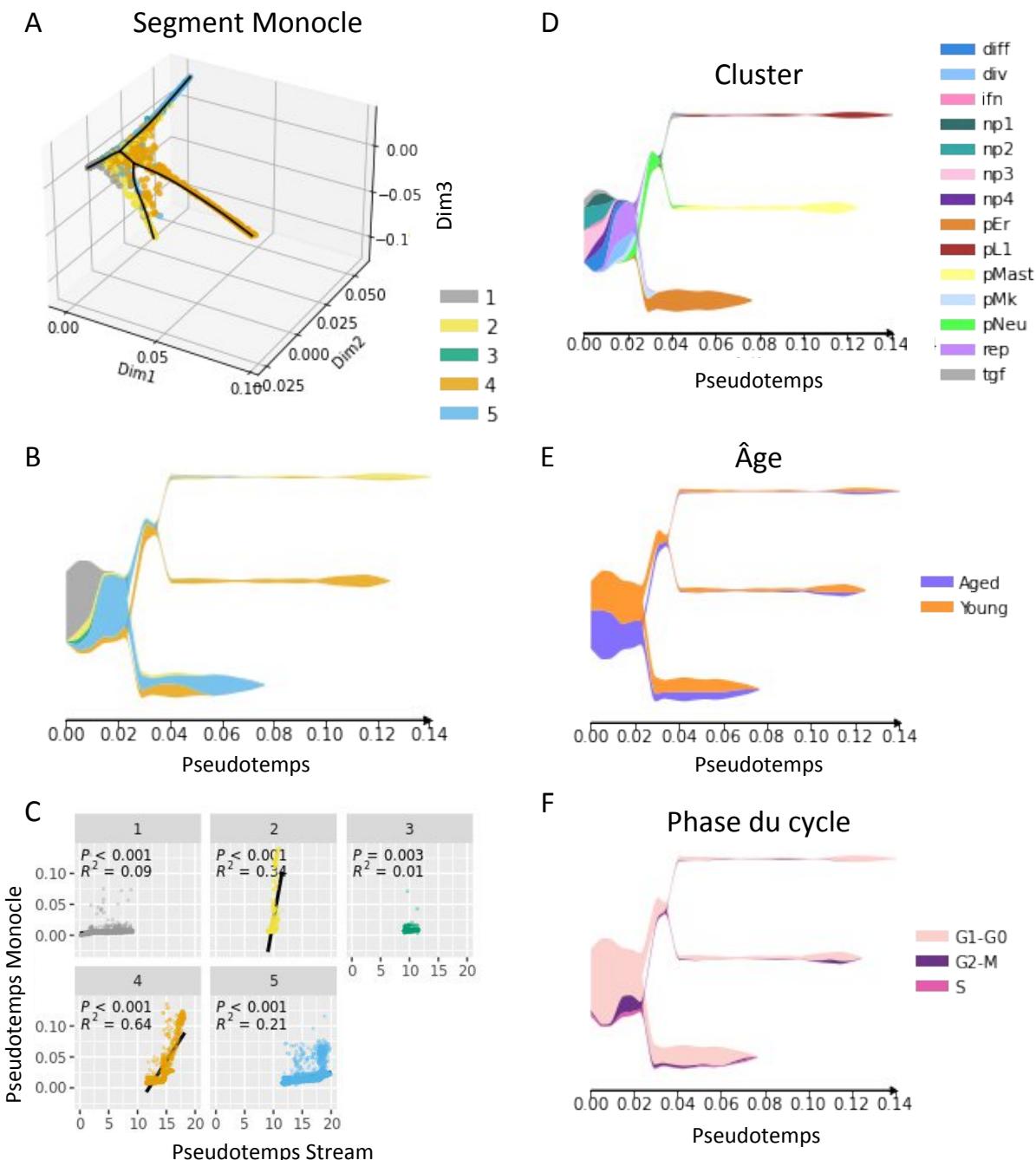


FIGURE 3.3 – Pseudo-trajectoire obtenue avec STREAM

A : Pseudo-trajectoire construite dans le MLLE en 3 dimensions avec STREAM. Les points sont les cellules colorées selon le segment de trajectoire Monocle2. B : Courant de cellule le long de la trajectoire obtenue avec STREAM. L'épaisseur de chaque couche représente la proportion de cellules des segments de trajectoire Monocle le long du pseudotemps STREAM. C : Corrélation linéaire entre pseudotemps de STREAM et Monocle2 dans chaque segment de trajectoire Monocle. P : p-valeur du test de régression linéaire,  $R^2$  : coefficient de détermination de la corrélation linéaire. DEF : courants de cellules le long de la trajectoire de stream colorés selon le cluster déterminé avec seurat (D), l'âge de la cellule (E) et la phase du cycle cellulaire (F). En B, D, E et F Une échelle logarithmique est utilisée en ordonnées pour la visualisation des branches d'amorçage qui contiennent peu de cellules par rapport au premier segment de la trajectoire.

## 3.6 Discussion

Le vieillissement de la CSH, du fait de ses conséquences sur l'hématopoïèse a fait l'objet de nombreuses études. Dans cette partie, nous voudrions reprendre les trois points importants de nos résultats et discuter leur recouplement avec les études précédentes.

### ***Le biais plaquettaire à l'origine du biais myéloïde ?***

Le fait que nous ayons effectué notre étude sur un pool de HSPC FLT3<sup>-</sup> par une approche scRNA-seq haut débit nous a permis de confirmer les biais préalablement décrits tout en précisant leur amplitude et leur mécanisme. Les études précédentes, menées avec beaucoup moins de cellules et sur des fractions triées de HSPC avaient mis en évidence des biais plaquettaire (GROVER et al., 2016) et myéloïde (KIRSCHNER et al., 2017) des HSPC, liés à l'âge. L'amorçage lymphoïde quant à lui a été retrouvé diminué au sein des MPP4 et LMPP (YOUNG et al., 2016). Une expansion avec l'âge d'une sous population de CSH amorcées pour répondre à un stimulus inflammatoire a également été mise en évidence (MANN et al., 2018).

Notre analyse de plus de 15000 HSPC isolées dans leur ensemble à des âges différents a permis de retrouver le biais plaquettaire sur l'ensemble des CSH âgées. Elle a permis de mettre en évidence une population de CSH âgée quiescente, caractérisée par une activité de facteurs de transcription, du biais myéloïde, Cebpb et Egr1, qui possède des similitudes avec celle observée par Kirschner (KIRSCHNER et al., 2017).

Nous avons en outre construit une pseudo-trajectoire de différenciation au sein des HSPC révélant une perte de la proportion de CSH s'amorçant pour les lignées lymphoïde, neutrophile/mastocytaire et érythroïde alors que l'amorçage megacaryocytaire est bien conservé. Nos résultats convergent donc vers un biais plaquettaire qui domine le vieillissement de la CSH. Une publication récente vient renforcer ce résultat, et rapporte une diminution des MPP amorcés pour les lignées lymphoïde et myéloïde chez l'humain (SOMMARIN et al., 2021).

### ***Origine de l'accumulation des CSH***

L'accumulation de CSH non fonctionnelles au cours du vieillissement a été aussi très étudiée et on pense qu'elle est en lien avec l'activité de cycle des CSH et un déséquilibre entre autorenouvellement et départ en différenciation (voir section 1.1.3.2 page 29 en introduction). Concernant la régulation de l'équilibre entre maintien de la population et départ en différenciation, une étude basée sur du scRNAseq a suggéré une hausse de l'autorenouvellement qui serait liée à un raccourcissement de la phase G1 du cycle cellulaire (KOWALCZYK et al., 2015).

Notre travail ne supporte pas cette hypothèse. En effet, nous ne retrouvons aucune différence avec l'âge de la proportion de LTHSC qui cyclent et que nous interprétons comme des CSH en autorenouvellement, au départ de la pseudo-trajectoire de différenciation. D'après

nos résultats, l'accumulation de CSH âgées non fonctionnelles semble plutôt avoir pour origine un blocage de la différenciation des CSH âgées quiescentes et biaisées pour le lignage myéloïde. En fouillant dans nos données, nous avons mis en évidence une signalisation TGF accrue dans ces cellules, ce qui nous amène à suggérer le rôle de la signalisation TGF-beta pour expliquer cette accumulation de CSH âgées en quiescence. De façon intéressante, le rôle de la voie JAK-STAT et de P53 a été mis en avant pour expliquer l'accumulation de cette population CSH âgée (KIRSCHNER et al., 2017). Notre étude et celle de Kirschner s'accordent sur les marqueurs de cette population, en effet les mêmes gènes clés ont été retrouvés (*Hes1*, facteurs KLF, AP-1, *Junb*, *Nr4a1*, *Cdkn1a*). Ces gènes sont compatibles avec les observations sur l'état de quiescence des CSH discutées en introduction (voir section 1.1.2.1 page 23 et PASSEGUÉ et al., 2004; SANTAGUIDA et al., 2009).

### ***Une hétérogénéité du vieillissement***

Finalement, notre étude a apporté une cartographie précise de l'hématopoïèse précoce au sein des HSPC jeunes et âgés. Celle-ci englobe en grande partie les sous populations qui avaient été identifiées indépendamment dans les études scRNA-seq de HSPC jeunes et âgées.

Cependant, certaines altérations du vieillissement que nous observons proviennent majoritairement d'un des deux échantillons âgés (diminution de l'amorçage lymphoïde T et augmentation des CSH quiescentes biaisées myéloïdes). Bien que des biais expérimentaux soient envisageables, cette différence pourrait très bien être la conséquence d'un vieillissement hétérogène des animaux des deux échantillons selon la théorie de l'hématopoïèse clonale (JAISWAL et EBERT, 2019). Il est fort possible que des clones compétitifs apparaissent et s'amplifient au détriment des autres au cours de la vie d'un individu.

# 4 Résultat : Inférence d'un modèle logique de l'hématopoïèse précoce altérée par le vieillissement

## Sommaire

4.1 Avant propos . . . . .	119
4.1.1 Introduction . . . . .	119
4.1.2 Inférence de BN à partir de contraintes dynamiques avec Bonesis . . . . .	120
4.2 Pipeline d'analyse . . . . .	122
4.3 Résultats . . . . .	125
4.4 Discussion . . . . .	157

## 4.1 Avant propos

### 4.1.1 Introduction

Notre analyse de la pseudo-trajetoire à la lumière des connaissances préalables sur l'héméostasie de la population de la CSH établit qu'avant d'atteindre des états amorcés vers les lignées lymphoïde, neutrophile, mastocytaire, érythroïde et mégakaryocytaire, la CSH peut passer par des états transitoires de quiescence et d'autorenouvellement. De plus, l'analyse de la distribution des cellules âgées a mis en évidence des altérations de l'aptitude des CSH agées à atteindre ces différents états avec une augmentation globale des CSH quiescentes avec l'âge au détriment de tous les états amorcés, mis à part l'état pMk. Ces résultats nous ont également permis d'identifier des acteurs potentiels (TF, [CDK](#), [CKI](#)) de ce processus biologique et de sa perturbation avec le vieillissement. Dans cette seconde étude, nous avons commencé par préciser les états HSPC clés du processus et le réseau transcriptionnel des acteurs sous-jacents via une nouvelle analyse exhaustive des régulons (en prenant tous les TF avec un motif disponible) que nous avons confrontée aux données publiées de ChIP-seq de 224 TF dans la [MO](#) de souris.

La modélisation logique est une approche pertinente pour l'étude de la dynamique de ce réseau de régulation génétique (voir introduction section [1.3.1.1](#) page [51](#)). Plusieurs modèles logiques de la différenciation de la CSH ont été proposés ([BONZANNI et al., 2013](#); [COLLOMBET](#)

et al., 2017; HAMEY et GÖTTGENS, 2019), mais aucun n'est en mesure de décrire nos observations des données scRNA-seq. À l'inverse, La dynamique d'un précédent modèle de la différenciation précoce des progéniteurs myéloïdes (KRUMSIEK et al., 2011) a retenu notre attention en se révélant proche de nos observations. Si ce modèle se situe à première vue en aval du système que nous étudions, ses états stables Érythrocyte et Mégakaryocyte présentent un profil d'activité correspondant assez bien aux expressions et activités de TF dans nos états pER et pMk. C'est également le cas pour la configuration pivot GMP de ce modèle qui se révèle proche de nos états pNeu et pMast combinés. Ce rapprochement rend compte de l'identification de plus en plus tôt dans l'hématopoïèse de l'amorçage de la CSH que nous et d'autres groupes avons rapportée (RODRIGUEZ-FRATICELLI et al., 2018).

Ce modèle constitue ainsi une base intéressante pour notre travail. Nous avons donc décider de nous appuyer sur le graphe d'influence de 8 TF de celui-ci auxquels nous avons ajouté 5 TF et deux complexes régulant la sortie (Cyclines D-CDK4/6) et le maintient (CIP/KIP) de la quiescence de la CSH. Le graphe d'influence a ensuite été enrichi d'interactions provenant d'études plus récentes et des régulations transcriptionnelles entre ces 15 composants inférées avec SCENIC (AIBAR et al., 2017).

Nous avons ensuite utilisé Bonesis (CHEVALIER et al., 2019), pour inférer un BN sur ce graphe d'influence vérifiant des contraintes dynamiques entre les états du processus, définies selon la pseudo-trajectoire et les connaissances préalables de l'homéostasie des CSH. Finalement, nous avons perturbé le modèle en fonction des altérations d'activité de régulons observées lors du vieillissement. Ceci nous a conduit à proposer certains facteurs et mécanismes à l'origine du biais de différenciation des CSH que nous observons dans nos données scRNA-seq, en particulier le rôle des facteurs Junb et Egr1 dans le déclin de l'amorçage vers les différents lignages à l'exception de l'amorçage mégacaryocytaire.

#### 4.1.2 Inférence de BN à partir de contraintes dynamiques avec Bonesis

L'objectif de Bonesis est d'inférer un BN localement monotone (cf. définition 2 en introduction page 54) reposant sur un graphe d'influence donné qui reproduit en sémantique MP (cf. définition 7 en introduction page 58) les observations de la dynamique du processus étudié (CHEVALIER et al., 2019; CHEVALIER et al., 2020). Ce problème est énoncé en définissant des contraintes dynamiques entre configurations partiellement décrites d'après les observations expérimentales que doit vérifier le BN recherché. Ces contraintes peuvent être l'accessibilité (A, cf. définition 4 en introduction page 55) ou la non accessibilité (NA) d'une configuration du BN depuis une autre. Des contraintes peuvent également être formulées sur les comporte-

ments asymptotiques du BN recherché imposant l'existence de points fixes (PF cf. définition 6 en introduction page 56), d'un ensemble de points fixes possibles (points fixes universels PFU), ou encore d'un ensemble de points fixes possibles accessibles depuis une configuration donnée (points fixes universels accessibles PFUA). Une observation partielle  $o$  à partir des données d'une configuration d'intérêt est un ensemble de couples associant un composant ( $i$ ) à une variable booléenne (0 ou 1) :  $o \subseteq [n] \times \mathbf{B}$ , en supposant qu'il n'existe pas de  $i \in [n]$  tel que  $\{(i, 0); (i, 1)\} \subseteq o$ .

Formellement, le problème de satisfaisabilité booléenne qui exprime la synthèse d'un BN entre  $n$  composants vérifiant ces contraintes est le suivant (CHEVALIER et al., 2020) :

Étant donné :

- un graphe d'influence  $G$  sur  $[n]$  (cf définition 1 en introduction page 53)
- $p$  observations partielles  $o^1, \dots, o^p$
- des ensembles de PF et de PFU d'indices d'observations ( $PF, PFU \subseteq [p]$ )
- des ensembles d'A, NA et PFUA de couples d'indices d'observations ( $A, NA, PFUA \subseteq [p]^2$ )

Trouver un BN  $f$  localement monotone de dimension  $n$  tel que :

- $G(f) \subseteq G$
- Il existe  $p$  configurations  $x^1, \dots, x^p$  telles que :
  - (lien avec les observations)  $\forall m \in [p], \forall (i, v) \in o^m, x_i^m = v$
  - (accessibilité)  $\forall (m, m') \in A, x^{m'} \in \rho_{mp}^f(x^m)$
  - (non accessibilité)  $\forall (m, m') \in NA, x^{m'} \notin \rho_{mp}^f(x^m)$
  - (points fixes)  $\forall m \in PF, f(x^m) = x^m$
  - (points fixes universels)  $\forall z \in \mathbf{B}^n, f(z) = z \Rightarrow \exists m \in PFU : \forall (i, v) \in o^m, x_i = v$
  - (points fixes universels accessibles)  $\forall (q, m) \in PFUA, \forall z \in \rho_{mp}^f(x^q), f(z) = z \Rightarrow \exists (q, m) \in PFUA : \forall (i, v) \in o^m, z_i = v$

En pratique, les contraintes sont donc définies sur des méta-configurations à partir des observations partielles (le niveau d'activité de certains composants peut être inconnu). Dans chaque méta-configuration impliquée dans une contrainte d'A, de NA et de PF, l'existence d'au moins une configuration satisfaisant la contrainte est demandée. Ceci permet de tenir compte de l'imprécision de la binarisation de l'activité de certains composants à partir des données expérimentales. Bonesis permet également de définir des *trap spaces* (partie de l'espace des phases dont on ne peut plus sortir) et de contraindre l'utilisation de l'ensemble du graphe d'influence en entrée. Par ailleurs, des contraintes peuvent également être formulées sur le comportement de mutants pour des composants du BN, dans ce cas là les propriétés

demandées doivent être vérifiées dans le BN f muté (voir section 1.3.1.6 en introduction page 61). Bonesis offre une interface python pour définir l'ensemble des contraintes en un unique programme logique exprimé en **ASP** pour lequel chaque solution correspond à un réseau booléen vérifiant les propriétés indiquées (PAULEVÉ et al., 2020). Ce langage déclaratif est utilisé pour résoudre des problèmes de satisfaisabilité combinatoire (GEBSER et al., 2012). Bonesis utilise le solveur clingo pour énumérer les fonctions Booléennes solutions sous la forme normale disjonctive (voir section 1.3.1.2 en introduction page 53).

Le problème posé peut n'avoir aucune solution ou au contraire un nombre trop important pour en faire une étude exhaustive notamment lorsque des composants ont un nombre de régulateurs importants et sont peu contraints. En effet, sous la forme normale disjonctive, le nombre de solutions possibles sans contrainte pour un composant avec  $d$  régulateur correspond au nombre de Dedekind connu jusqu'à  $d = 8$  (WIEDEMANN, 1991). Le tableau 4.1 donne le nombre de fonctions possibles (nombre de Dedekind) pour  $d$  allant de 0 à 8.

# régulateurs	# fonctions booléennes monotones
0	2
2	6
4	168
6	7828354
8	$> 5.6 \times 10^{22}$

Tableau 4.1 – Nombre de fonctions booléennes monotones possibles pour un composant selon son nombre de régulateurs

Lorsque le nombre de solutions est très grand, Bonesis offre la possibilité d'échantillonner des ensembles de BN solutions divers permettant une analyse des différents scénarios de modélisation (CHEVALIER et al., 2020). Par ailleurs, le nombre de solutions peut être limité par l'utilisateur en fixant un nombre maximum de clauses par composants. Typiquement, les BN de systèmes biologiques présentés dans la littérature présentent rarement plus de 3 clauses dans leurs fonctions logiques sous la forme FND. Pour restreindre le nombre de solutions possibles, l'ASP permet également de rechercher des solutions minimisant ou maximisant certains paramètres comme le nombre de noeuds ou d'arêtes du graphe d'influence utilisés ou bien le nombre total de clauses du BN.

## 4.2 Pipeline d'analyse

Un second pipeline d'analyse en Snakemake a été développé pour cette étude. Il prend en entrée les résultats de l'étude précédente (matrice d'expression filtrée après le contrôle qualité, clustering de Seurat, pseudo-trajectoire de monocle, classification dans les phases

du cycle cellulaire et les sous-types HSPC). La partie SCENIC (GRNboost et cisTarget) est ici lancée 50 fois par ensemble de cellules considérés (toutes, que les jeunes ou que les âgées) dans le but de stabiliser les résultats d'inférence de régulations. Les jeux de données ChIP-seq de TF sur des échantillons de MO de souris sont analysés en parallèle avec BETA (S. WANG et al., 2013). Un graphe d'influence est ensuite construit à partir de ces résultats ainsi que des données de régulations de la littérature sur un ensemble de 13 TF et 2 complexes du cycle cellulaire (Cyclines-D-CDK4/6 et CIP/KIP). Bonesis est finalement utilisée pour inférer un BN en utilisant la stratégie présentée Figure 2 section 4.3 suivante.

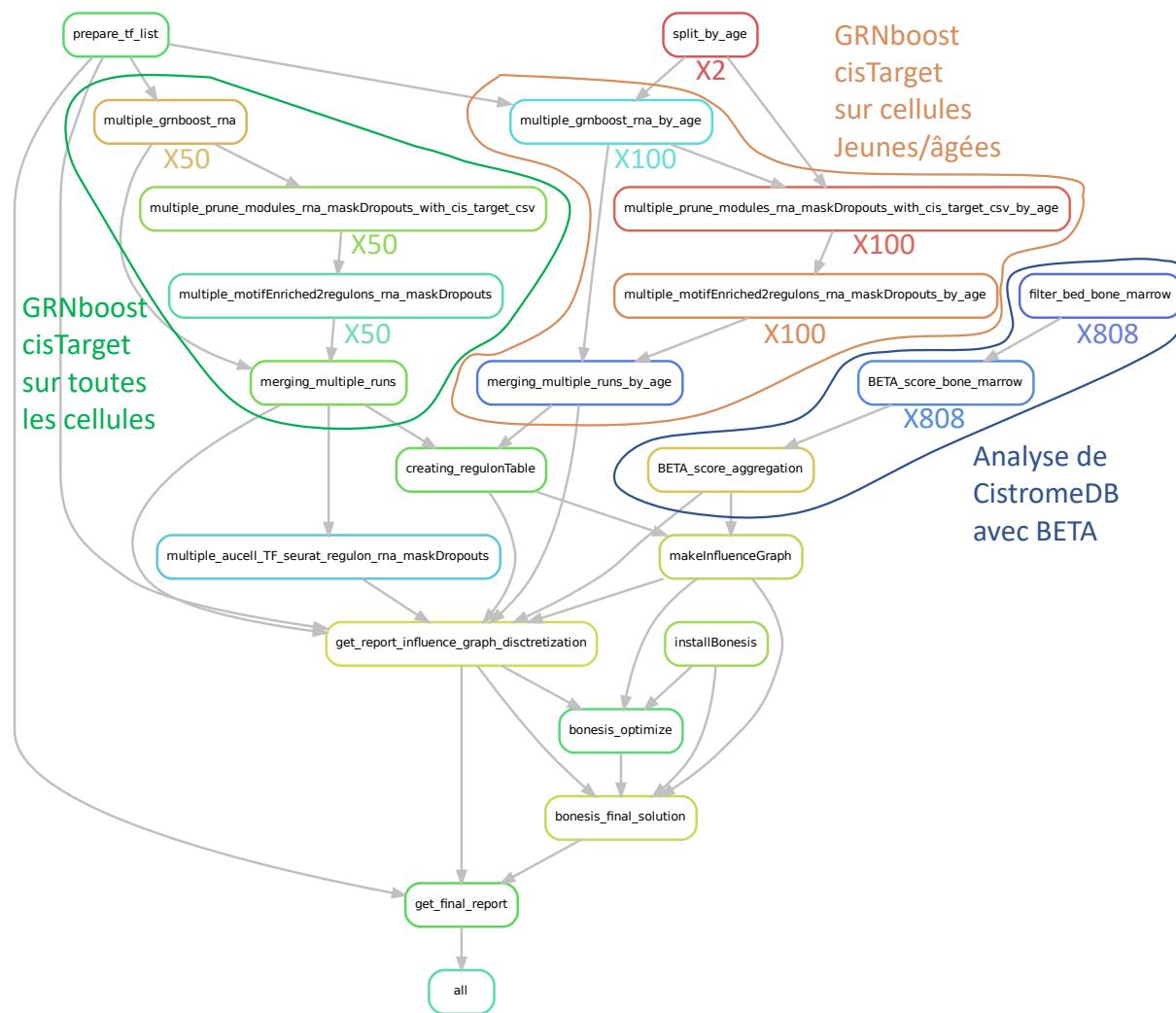


FIGURE 4.1 – Pipeline d'inférence d'un BN à partir de données scRNA-seq

Pipeline d'analyse présenté sous la forme d'un graphe dirigé acyclique partant des règles (étape) initiales `split_by_age` (séparation des matrices d'expression jeunes et âgées) et `prepare_tf_list` (obtention des TF avec un motif référencé dans la base de données utilisée par cisTarget) jusqu'à la règle finale `all`. Les règles sont les noeuds et leur dépendances en terme de fichiers d'entrée et de sortie sont les arêtes du graphe. 150 runs des étapes GRNboost et cisTarget (50 sur toutes les cellules, 50 sur les cellules jeunes, 50 sur les cellules âgées) ont été réalisés. 808 jeux de données ChIP-seq pour 224 TF sur des échantillons de MO de souris ont été analysés.



## **4.3 Résultats**

Un manuscrit présentant ce travail est en cours de rédaction. Cette section présente les parties Résultats et Méthodes du manuscrit (rédigés en anglais).

# SC-RNA-seq assisted synthesis of a Boolean gene network to model early haematopoiesis and its alteration with aging

Léonard Hérault<sup>1,2</sup>, Mathilde Poplineau<sup>2</sup>, ..., Estelle Duprez<sup>2\*#</sup> and Élisabeth Remy<sup>1\*#</sup>

1. Aix Marseille Université, CNRS, Centrale Marseille, I2M, Marseille, France
2. Epigenetic Factors in Normal and Malignant Hematopoiesis Team, Aix Marseille Université, CNRS, INSERM, Institut Paoli-Calmettes, CCRM, Marseille, France

\* These authors contributed equally: Elisabeth Remy, Estelle Duprez  
#Corresponding authors.

## Abstract

We previously analysed 15 000 transcriptomes of mouse hematopoietic stem and progenitor cells (HSPCs) from young and old mice and characterized the early differentiation of the hematopoietic stem cells (HSCs) according to age, thanks to cell clustering and pseudotime analysis (Herault et al, 2021). In this study, we propose an original strategy to build a Boolean gene network explaining HSC priming and homeostasis based on our previous single cell data analysis and the actual knowledge of these biological processes (**graphical abstract**).

After a global regulon analysis on selected HSPC states in the differentiation trajectory of HSCs, we chose to focus on 15 components, 13 selected TFs (Tal1, Fli1, Gata2, Gata1, Zfpml1, Egr1, Junb, Ikzf1, Myc, Cebpa, Bclaf1, Klf1, Spi1) and two complexes regulating the ability of HSC to cycle (CDK4/6 - Cyclines D and CIP/KIP). We then defined the relations in the differentiation dynamics we want to model ((non) reachability, attractors) between the HSPC states that are partial observations of binarized activity configurations of the 15 components. Besides, we defined an influence graph of possibly involved TF interactions in the dynamic using regulon analysis on our single cell data and interactions from the literature. Next, using Answer Set Programming (ASP) and considering these inputs, we obtained a Boolean model as a final solution of a Boolean satisfiability problem. Finally, we perturbed the model according to aging differences underlined from our regulon analysis. This led us to propose new regulatory mechanisms at the origin of the differentiation bias of aged HSCs, explaining the decrease in HSC priming toward all mature cell types except megakaryocytes.

## Results

### Regulon activities and their transcriptional network sustain early haematopoiesis

In order to define the reference to establish dynamical constraints for the inference method, we relied on our previous scRNASeq analysis (clustering, pseudotime trajectory and cell cycle phase assignment, Herault et al., 2021) to define HSPC states and characterized their activity.

The clusters identified in Herault et al, 2021 provide a meaningful functional partition of the cells, and the shape of the trajectory, that is robust with respect to different analyses, reflects well the priming of HSC toward different lineages. Hence, we defined four lineage-primed HSPC states corresponding to the cells of lineage-primed clusters (Herault et al, 2021) and that localized at the terminal branches of the trajectory: pLymph (primed lymphoid clusters pL1 on branch 2), pNeuMast (primed neutrophils and primed mastocytes clusters gathered together, on branch 4), pEr (primed erythrocytes, on branch 5) and pMk (primed megakaryocytes, on branch 5) (**Figure 1A**). We defined a state preDiff that contains all cells of the diff cluster, gathering most of the short-term hematopoietic stem cells (STHSC) and spreading on the branches of the trajectory. Upstream in the trajectory, four HSPC states of cells were considered (**Figure 1A**). The ifnHSC state that gathers the cells of the ifn cluster (interferon reponse signature) and the qHSC state that gathers the cells of tgf cluster which present the highest quiescence signature related to TGF-beta in the dataset and are, for almost all of them, in G1/G0 cell cycle phase. We also defined two HSPC states at the beginning of the pseudotime trajectory (pseudotime <2), the non-cycling cells that form the initiating HSC state (iHSC) and the ones being in the G2/M cell cycle phase that form the self-renewing HSC state (srHSC).

The nine defined states together represent 60% of the cells previously analysed (Hérault et al., 2021) and are linked to initial (iHSC, srHSC), terminal (pEr, pMk, pNeuMast, pLymph) and branching (preDiff) points of the pseudo-trajectory, they also recapitulate key HSC states usually described: LTHSC (iHSC), SRHSC (srHSC), committed STHSC (preDiff), quiescent HSC (qHSC), primed HSC (pEr, pMk, pNeuMast, pLymph) plus the two non-primed HSC states showing pathway signature we previously described (tgfHSC, ifnHSC). Thus, they provide a detailed view of the key states an HSC can reach during its life (**Figure 1A**).

Then, we studied regulons to functionally characterize these HSPC states. Regulons are defined as modules constituted of a transcription factor and its potential targets. We used the SCENIC workflow (Van de Sande et al., 2020) to identify them with a regression per target approach, followed by cis-regulatory motif discovery in order to discard indirect targets. The regression step being stochastic, we ran the SCENIC workflow 50 times on all the cells of the dataset. We selected regulons found in at least 80% of the runs, and such that their targets are also found in at least 80% of the runs. Doing so, we characterised 197 activating regulons and 132 inhibiting ones (**supplementary Table 1**).

We next quantified the activities of the regulons with AUCell enrichment score (Aibar et al., 2017) and performed a hierarchical clustering (**Figure 1B**). This revealed specific regulon activities for each of the HSPC states we defined. The lineage-primed HSPC states were associated with expected regulon activities: Klf1 was active in pEr, Gata1 in pEr and pMk, Spi1 and Cebpa were active in pNeuMast and Ikkzf1 and Zbtb16 in pLymph. Concerning the non-primed states, we observed the activity of Stat and Irf regulons for ifnHSC, Gata2, Junb, Egr1 and Klf regulons for qHSC and Bclaf1 and Srf regulons for respectively iHSC and srHSC states. Fli1 regulons was active in both iHSC and pMk. In the preDiff state some regulons of priming showed an activity such as Spi1 and Myc contrary to iHSC. Thus, performing differential activity test we identified 140 activating regulons markers of our 9 HSPC states (**supplementary Table 2**). It can be noted that Zbtb7a is the

unique regulon that significantly marked srHSC, probably because of the low cell number and the stemness of this state (71). These findings highlight the transcriptional differences between HSCP states that are each of them marked by a particular set of regulons.

To go further we built a transcriptional network whose nodes are TFs at the head of regulons significantly marking at least one of the HSPC states, and directed edges represent the transcriptional regulations between them. We only considered the regulations recovered in at least 90% of SCENIC runs. After removing auto-regulations, we obtained a directed graph of 133 nodes and 670 edges (**Figure 1C**). In addition, to further support these interactions, we analysed ChIP-seq data of mouse TFs with a Bone Marrow tissue annotation from Cistrome database (Liu et al., 2011), which covered 33 % of the source TFs of the transcriptional network. Around 60% (302) of the network interactions whose source node is a TF available in the Cistrome database were supported by a TF peak (**supplementary Table 1**). Thus, the edges of the regulon network obtained using SCENIC are consistent with the public ChIP-seq data, generated from cell populations close to ours.

We then performed a clustering analysis by weighting the network using a Normalized Interaction Score (NIS) computed from SCENIC outputs and with applying Louvain clustering. We underlined 10 regulon communities and three isolated regulons (Zbtb7b, Brf2, Sp4). By associating each TF in the network to the HSPC state that its regulon most characterizes, we observed that half of the communities consists of a set of TFs whose activity characterizes a particular HSPC state (**Figure 1C and supplementary Table 2**). Indeed, most of TFs from the C1 community (Klf factors, Jun and fos AP-1 factors, Egr1) are known to be related to quiescence and their regulons were markers of the qHSC state, whereas the C2 community contains mainly TFs leading regulon markers of ifnHSC state (eg Irf1-7-9, Stat1-2). In the same way, C3 community was associated with pEr state, C4 community with pNeuMast and C5 with iHSC (Kdm5b, Foxp1) and preDiff (Sox4, Hoxa9) states. It was more difficult to analyse the smaller communities C6 to C10 as they present a more heterogeneous composition of TFs.

Altogether, our regulon analysis revealed a functional relevance of the HSPC states that we defined by integrating 3 layers of information (cell cluster, trajectory branch and cell cycle phase), with their specific transcriptional activity characterization. It also pointed out that the TFs, regulators of these activities, interact with each other in a structured network that governs the differentiation journey of HSCs from iHSC to one of the possible primed HSPC states, passing or not through the other transient HSPC states (sr-qHSC, preDiff).

## Inference of a gene Boolean network to model HSC priming

One of the key questions is the underlying dynamics that govern the choice of HSCs between maintaining stemness or priming it to different hematopoietic lineages. To decipher the key molecular mechanisms governing HSC fate, we conducted a gene Boolean network construction. To do so, we implemented a strategy to build this network by leveraging both current knowledge of the biological processes and our analysis of single cell data, which we combined with the use of Bonesis, a recently proposed approach for Boolean network inference (Chevalier et al., 2019). From an influence graph, and dynamical constraints (expressed for instance in terms of stability or reachability of configurations of the network), Bonesis solves a Boolean satisfiability problem using Answer Set Programming (ASP), and enumerates all the possible Boolean models that satisfy the constraints in the Most Permissive (MP) semantic (Paulevé et al., 2020).

**Influence graph synthesis.** We built our gene network based on a previous published Boolean model of early myeloid differentiation (Krumsiek et al., 2011) that encompasses 11 TFs. Eight of them were regulon markers of some of the HSPC states in our analysis: Gata1 a marker of pEr and

pMk; Fli1 of pMk; Klf1 of pEr; Spi1 and Cebpa of pNeuMast; Tal1, Fli1 and Gata2 of qHSC (**supplementary Table 2**). Zfpm1 the cofactor of Gata1 was not detected as a regulon but was expressed in pEr and pMk HSPC states (**Supplementary Figure 2**). These expression and activity patterns were in a very close agreement with the fixed point Er and Mk and with the branching state before the granulocytes and monocytes fixed points generated by the Krumsiek model. Thus, we included these nine TFs and their interactions in the influence graph of the Krumsiek model. To represent lymphoid priming, we added the TF Ikzf1 whose regulon marked pLymph state according to our analysis (**supplementary Table 2**), in agreement with prior knowledge of early lymphoid specification in HSPC (Ng et al., 2009).

Next, we added two components regulating the ability of HSC to cycle: The CDK4/6-Cyclines D (CDK4/6CycD) complex (*Ccnd1-3* and *CDK4/6 genes*) required for the HSC quiescence exit, and its inhibitory complex CIP/KIP (*Cdkn1a-b-c genes*) marking the quiescence of the HSCs (Pietras et al., 2011). To connect CIP/KIP complex to the network we added Junb, Egr1 both identified in our regulon analysis as markers of qHSC state and activators of CIP/KIP genes (**supplementary Tables 1&2**). Both were also previously found involved in HSC quiescence (Min et al., 2008; Santaguida et al., 2009). Finally, to connect CDK4/6CycD to the network we added Myc and Bclaf1, two factors involved in HSC cell cycle ((Dell'Aversana et al., 2017; Wilson et al., 2004)). Both were active regulons in the preDiff state whereas only Bclfa1 regulon was active in srHSC state with CDK4/6CycD complex genes in its targets (**Figure 1B and supplementary Table 1&2**). In addition to the interactions originating from Krumsiek model, we added interactions between the 15 components identified with SCENIC analysis (interactions found in at least 90% of the runs, discarding self-inhibitions because of their uncertainty (Van de Sande et al., 2020)), and from the literature. Finally, we obtained an influence graph encompassing 15 components and 60 interactions (**Figure 2Ai**). More than 75% of them were confirmed by at least two sources (Scenic, literature references or Cistrome database analysis, cf **supplementary Figure 1A & supplementary Table 3**). Note that one can compute with the Dedekind numbers that more than  $10^{22}$  locally monotonic Boolean networks are compatible with the influence graph (Wiedemann, 1991).

We associated to each HSPC state, a meta-configuration, *i.e.* a vector representing the discretized activity level of each of the 15 components. If the node of the network represents a TF heading a regulon with more than 10 targets, we considered AUCell scores of all cells of the HSPC state and used a Kmeans clustering (with K=2) to decide whether the node was active (1) or inactive (0). Otherwise, because the AUCell was not always available and less reliable, we discretized the activity on 3 levels, active (1), inactive (0) or free/unknown (\*) using Kmeans clustering (K=3) on averaged RNA levels per HSPC states. We did the same for each gene of the cell cycle complexes. The activity levels of the CDK4/6CycD and CIP/KIP complexes were set according to the sum of the computed discretization level of each of their genes (**Figure 2Aii and supplementary Figure 2**).

**Dynamical constraints.** In order to get possible Boolean parametrisations of the influence graph, we enounced dynamical constraints between the HSPC states to define the Boolean satisfiability problem solved by Bonesis. The constraints are resumed in a HSC differentiation journey (**Figure 2Aiii**) deduced from the pseudotime trajectory and prior knowledge of HSC biology. We required that the model presents at least a fixed point, reachable from iHSC, in each of the 4 primed meta-configurations pLymph, pNeuMast, pER and pMk. We also imposed that all the fixed points of the model reachable from the iHSC correspond to at least one of these 4 primed meta-configurations. A cell can go back and forth from iHSC to srHSC state (enter in self-renewal) or qHSC state (quiescence) (Bernitz et al., 2016; Wilson et al., 2008). From iHSC, cells pass through the pre-differentiating HSPC state preDiff according to the pseudo-trajectory. Because this transient state is composed of the majority of STHSCs that are known to be definitely committed in the differentiation we considered it as a “non-return state”: from preDiff, state iHSC is not accessible, but any of the 4 primed fixpoints are reachable. We decided not to use the ifnHSC state in the

definition of dynamical constraints as we did not have enough knowledge on how it dynamically communicates with the others HSPC states.

This first set of constraints proved to be too stringent. Thus, we relaxed some constraints on the component activities, by replacing their discretized value 0 or 1 to a free (\*) status (Egr1 in srHSC, Bclaf1 in pLymph, Myc in pNeuMast, pEr and pMk, Gata2 in pNeuMast, CDK4/6CycD in pMk and CIP/KIP in pLymph, see **Figure 2Aiii & supplementary note 1**). We added a zeros configuration in which all the components are inactive in the list of reachable fixpoints. The reachability of pLymph from preDiff condition prevented from finding a solution, so we discarded it. This is not in contradiction with the pseudo-trajectory as the branching towards pLymph was the earliest, very close to non-primed HSC regarding its hscScore (herault et al). Finally, we obtained more than 100 000 solutions with this first inference using Bonesis.

**Adding mutant constraints.** In order to refine our search toward relevant solutions, we limited the number of clauses per logical rules to 3 and used altered behaviours previously described to add constraints. To do that, we inferred a subset of 1000 “diverse” solutions (as previously done in (Chevalier et al., 2020), and for each of them we simulated Knock-Out/Knock-In (KO/KI) perturbations on all the nodes one by one. We compared attractors of the altered models with mutant phenotypes reported in the literature (see **supplementary Table 4 & supplementary note 2**) and aging phenotypes observed in our data (**supplementary note 2**). When an altered behaviour agreed with the data, we constrained it in a subsequent inference step with Bonesis (**Figure 2Bi** and **supplementary Figure 3A&B**).

**Influence graph pruning.** To this updated constraint set, we added two optimization criteria: maximization of the confident interaction number (strong literature evidence in our studied cell types **supplementary Table 3**) and then minimization of the other interaction number in order to prune the influence graph and reduce the number of solutions. More than 80% of the 36 remaining interactions were supported by at least two sources (SCENIC, Cistrome, literature evidences **supplementary Figure 1B**). We then kept interactions retrieved in at least one solution, and required solutions containing all these 36 interactions (**Figure 2Bii**). Among the 616 possible solutions, logical rules were different for the components CDK4/6CycD, Fli1, Gata1 (2 inferred rules each) and Gata2 (77 inferred rules). Taking into account literature and complexity (clause number) of these rules (**supplementary note 3**), we selected the final Boolean network presented **Table 1**.

Thus, the coupling of an implementation of Bonesis method with multiple biological knowledge sources allowed us to overcome the problem of dealing with the huge number of possible solutions provided by Bonesis. We successfully synthetized a Boolean network of early hematopoiesis based on a regulatory network constituted of 15 components and 36 interactions.

## The inferred Boolean model of early hematopoiesis shows a hierarchy in HSC priming

Although the topology of the gene regulatory network consisted in a unique strongly connected component and 4 output nodes (CDK4/6CycD, CIP/KIP, Tal1 and Klf1), we distinguished two modules in the gene regulatory network, one constituted of the nodes Bclaf1, Myc, Junb regulating cell cycle complexes CDK4/6CycD and CIP/KIP and a second one with the other highly connected TF nodes governing HSC fate (**Figure 3A**). Cebpa, Fli1 and especially Egr1 are making the connections between these two parts.

Simulations of the model done within the MP semantic provided a complete description of the attractors (**Figure 3B**). We highlighted that the Tal1 was active in the fixed point pEr, as it is also in

the erythroid fixed point in the Krumsiek model (Krumsiek et al., 2011) while its value has been left free in the constraints. According to our requests, all fixed points were reachable from iHSC, regardless the initial value of Zfpm1 component (**Figure 3B**).

A fine analysis of the most permissive trajectories highlighted events at the origin of the dynamic properties along the trajectory. Note that this trajectory was captured thanks to the MP semantic in Boolean formalism. Our analysis indicated that Gata2 is active at the initial states in iHSC and inactive when the cell reaches preDiff and can no longer be reactivated. This event may explain a first branching of the trajectory between preDiff and pLymph that is marked by the activity of Iκzf1 whose only regulator is the activator Gata2 (**Figure 3B&C**). Moreover, we observed that cells can reach the configuration pME (defined in **Figure 3B**) from the preDiff state, with an activation of *Junb* and inactivation of *Spi1* while this configuration was no longer reachable from pNeuMast (**Figure 3C**).

Interestingly, the choice between pMk and pEr fixed points relied on two different levels of activity of Fli1. Indeed, starting from the branching point pME, Fli1 activity can increase allowing *Gata1* to be also activated. Then, activation of *Klf1* by *Gata1* can occur as long as Fli1 has not reached its activity level allowing to inhibit *Klf1*. This scenario leads to pEr fixed point. It means that a necessary condition to reach pEr from pME is that the inhibition threshold of *Klf1* is greater than the activation threshold of *Gata1*. In addition, our model is suggesting that a proliferation configuration is accessible from the iHSC when CDK4/6CycD and Myc are active, and CIP/KIP inactive, and that all fixed points can be reached from this configuration. This points out that preDiff state is the major waypoint of HSC differentiation toward the stable states as suggested by our previous study (Herault et al, 2021)

We previously showed that a large part of HSC proliferates when the priming toward the different lineages occur (Herault et al, 2021). Our model agrees with this observation as a proliferation configuration (CDK4/6CycD active, Myc active and CIP/KIP inactive) is reachable from iHSC and any fixed points can be reached from it.

To assess the consistency of the model with literature we also conducted KO mutation simulations for all the sources (TF) nodes one by one and compared the resulting dynamics with wet lab experiments in the literature. Regarding the reachability of HSPC meta-configurations from iHSC, the large majority of the in-silico KO simulations matched the corresponding *in-vivo/in-silico* perturbations reported in the literature (**supplementary Table 4**). For example, our *in silico* *Fli1* KO led to the loss of pMK fixed point from iHSC in agreement with the *Fli1* KO BM that harbours a megakaryopoiesis defect (Moussa et al., 2010).

To summarise, the dynamical analysis of our MPBN of early hematopoiesis gives new insights about early priming hierarchy of HSC. It highlights a decisive role of *Gata2* inactivation to reach preDiff at the expense of pLymph. From pDiff, the inactivation of *Spi1* with the activation of *JunB* leads to pME a branching point that depends on fine tuning of Fli1 to commit the priming toward pMK or pER.

## Perturbations of early hematopoiesis model explains some HSC aging features

Our previous single cell RNA-seq analysis revealed an alteration of HSC priming with an accumulation of quiescent HSCs in aged animals at the expense of the priming toward pLymph, pEr and pNeuMast (Hérault et al., 2021). To decipher molecular mechanisms and factors at the origin of this alteration, we simulated perturbations in the inferred Boolean network according to alterations in the transcriptome of aged HSPCs to recover aging phenotypes. First, we identified alterations of

regulon activity due to aging by comparing for each states the regulon activities of young and aged HSPCs. Regulon transcriptional activity differences were found mainly (80%) in the non-primed iHSC, ifnHSC, srHSC states with similar amount of decrease and increase in activity (**supplementary Figure 4**), and very little in pNeuMast and pEr. Almost all regulon activity alterations were found in more than one state. Aging features consisted mainly in a decrease of the activity in regulons related to HSC activation (Runx3, Sox4, Myc and Spi1) and NF-kappaB pathway (Rel and Nfkb factors), and an increase in regulon from the AP-1 complex (Atf, Jun and Fos factors) and involved in quiescence of HSCs (Egr1, Klf factors, Gata2) (**supplementary Figure 5 & supplementary Table 5**). To be noted that we observed a specific increase in Cebpe-b regulon activity in qHSC state marking the myeloid bias of these quiescent aged cells. Eight of the 13 TF components of our models were altered upon aging in their regulon activities (Myc, Spi1, Junb, Egr1, Fli1, Klf1, Gata2 and Gata1, **supplementary Table 5**). More precisely we found Junb, Egr1 and Fli1 (resp. Spi1) activities significantly increased (resp. decreased) in more than a half of the 8 HSPC states considered for the model inference (**Figure 4A**).

As aging of HSCs is characterized by alterations of the chromatin structure that influence the ability of TFs to regulate their targets (Sun et al., 2014), we also considered altered TF regulations. To identify such possible altered regulations, we compared for each regulation the normalized interaction score (NIS) of young and aged cells analysed separately using SCENIC workflow and computed a score difference between young and aged conditions (**supplementary Table 1 and supplementary Figure 6**). The distribution of these score differences showed that most of the regulations were not strongly altered (14% of the interactions have a score difference above 0.4; **supplementary Figure 7**). When focusing on the interactions of the model supported by Scenic, we noticed an alteration of *Cebpa* activation by Gata2 (decrease of the NIS of 0.4 upon aging), which was compensated by an increase of activation by Spi1 (NIS increase by 0.2) upon aging (**Figure 4B**).

In order to stimulate aging alteration, we either performed node mutations with KI mutations on *Junb*, *Egr1* and *Fli1* and KO on *Spi1* or an edgetic mutation (loss of *Cebpa* activation by Gata2) was used in the model (**Figure 4C**).

The *Spi1* KO mutant led to the loss of pLymph and pNeuMast fixed points. *Egr1*, *Junb* and *Fli1* KI Mutants presented a unique pMk fixed point that is quiescent (pMk with CIP/KIP active) for *Egr1* and *Junb* KI. To be noted that we imposed pMk to be the unique reachable fixed point from iHSC for *Egr1* and *Junb* KI in the inference step. The loss of activation of *Cebpa* by Gata2 (edgetic mutation) led to the loss of reachability of pLymph and pNeuMast fixed points from any of i-sr-qHSC configurations (**Table 2**). Thus, simulations of the 5 perturbations simulating aging led to the loss of reachability of pLymph and pNeuMast fixed points from i-sr-qHSC configurations, and the 3 overexpressed mutations (KI) to the loss of reachability of pEr. These results agree with our single cell analysis at the population scale, as the 4 fixed points correspond to the primed HSPC states whose cell proportion significantly decreases with aging, while pMk remains reachable in any of our model perturbations (**Table 2**). We also observed that preDiff was no longer reachable from i-sr-qHSC configurations with the mutation *Junb* KI and the edgetic mutation *Cebpa*-Gata2, suggesting that HSC priming toward pMk in aged mice take an alternative differentiation path. This path could be directly from qHSC state whose proportion increases with aging (**Figure 4D**) and that spread at the end of the first pseudo-trajectory branch in pseudotime near the appearance of the early pMk cells (branch 3 and beginning of branch 5 of pseudotime trajectory **Figure1A**).

The model (**Table1**) emphasizes a clear dependence between the 5 perturbations related to aging. Indeed, *Egr1* KI implies a definitive activation of *Junb* which in turn activates definitively *Fli1*. Besides, the KO of *Cebpa* activation by Gata2 prevents Spi1 to become active from any of the i-sr-qHSC configurations. Thus, we highlighted the major roles of *Egr1* overexpression and loss of *Cebpa* activation by Gata2 in early hematopoiesis aging. On another note, regarding the global TF network from scenic (**supplementary Table 1**) *Junb* and *Egr1* upregulation with aging could be mediated by Klf factors such as Klf2-4-6 also upregulated with aging (**supplementary Table 5**).

These Klf factors are known to be downstream of TGF-beta pathway (Botella et al., 2009; He et al., 2015, p. 4; Yan et al., 2018), that is the main signature of qHSC HSPC we retrieved.

Therefore, our model perturbation analysis of aging of early haematopoiesis highlights *Egr1/Junb* upregulation through TGF-beta pathway and loss of *Cebpa* activation by Gata2 alterations as two major molecular mechanisms that lead to HSC aging resulting in the decrease in all lineage priming except the megakaryocyte one.

## METHOD DETAILS

### sc-RNA-seq dataset

We used the scRNA-seq dataset presented in our previous study available in the Gene Expression Omnibus database under accession code GSE147729 (Héault et al., 2021). This dataset is composed of two pools of young (2/3 months) mice HSPCs and two pool of aged (18 months) mice HSPCs. Our previous results (cell cycle phase, cell clustering, pseudotime ordering) were considered in this study.

### Definition of HSPC states

HSPC states were defined as follow according to our previous study: For the primed state, pEr (resp pMk) gathers cells belonging to pEr (resp pMk) cluster and trajectory branch 5, pNeuMast gathers cells belonging to pNeu and pMast clusters and trajectory branch 4, pLymph state gathers cells belonging to cluster pL1 and branch 2. The non-primed states preDiff, qHSC and ifnHSC gather respectively cells of clusters tgf, ifn and diff. iHSC (resp srHSC) gathers cells with a pseudotime value lower than 2 and assigned to G1/G0 (resp G2/M) cell cycle phases.

### Regulon analysis with pyScenic

We ran scenic workflow using pySCENIC v1.10.0 with its command line implementation (Van de Sande et al., 2020) as in our previous study regarding gene filtering, TF motifs (motifs-v9-nr.mgi-m.001-o0.0), cis-target (+/- 10 kb from TSS mm9-tss-centered-10 kb-7species.mc9nr) databases and command line options (Héault et al., 2021). In this study we used the whole 1721 TFs with an available motif in the motif database as input. We processed with Scenic workflow all cells together as well as only young or only old cells. For each cell set, regression per target step with grnboost2 followed by cis-target motif discovery and target pruning were run 50 times using a different seed for the pySCENIC grn command. The regulons and their targets recovered in at least 80% (e.g. 40) runs were kept.

For each remaining interaction representing a transcriptional regulation of a gene  $g$  by  $r_t$  one of its n regulators ( $r_1, \dots, r_n$ ) we computed a normalized interaction score NIS as follow:

$$NIS(r_t, g) = \frac{IS(r_t, g)}{\sum_{j=1}^n IS(r_j, g)}$$

Where IS the interaction score defined as the product of the number of scenic runs where the interaction was found by the mean importance score given by grnboost2 for the interaction across these scenic runs. The results for interactions found in all cell set pySCENIC analysis are available in **supplementary Table 1**.

## Regulon marker analysis

For scenic results obtained on all cells, AUCell scores of activating regulons (i.e., those with a positive correlation between the TF and its targets) resulting from the multiple run filtering were computed with pycenic aucell command (default option with a fixed seed).

Averaged AUCell scores by HSPC states were computed. These scores were standardized in order to hierarchically cluster the regulons using ward.D2 method of the R function hclust with Euclidean distance. The DoHeatmap function from the Seurat package (Stuart et al., 2019), was used to display the results (**Figure 1B**). Averaged AUCell enrichment scores by HSPC states plus cell age were also computed, standardized and the results were displayed with the previous clustering results (**supplementary Figure 5**).

Activating regulon markers of HSPC states were identified based on their AUCell scores using FindAllMarkers Seurat function (min.pct= 0.1, logfc.threshold=0) with Wilcoxon rank sum tests. Only regulon with an average AUCell score difference above 0.001 between one state versus all the others were kept. A p-adjusted value (Bonferroni correction) threshold of 0.001 was applied to filter out non-significant markers (**supplementary Table 2**)

Activating regulon activity differences with aging in each state were identified using the FindConservedMarkers Seurat function (sequencing platform as grouping variable, min.pct = 0.1 and logfc.threshold = 0) with Wilcoxon rank sum tests. For each HSPC state, only average AUCell score differences of same sign and above 0.001 in the two batches presenting a combined p value < 0.001 were kept (**supplementary Table 5**)

## Cistrome database analysis

Available mouse TF ChIP-seq experiments annotated for bone marrow tissue in the Cistrome database were analysed using Cistrome database workflow (Liu et al., 2011). More precisely, for each bed file of selected experiment from the databases, the top 10,000 peaks with more than 5-fold signal to background ratio were conserved for downstream analysis. Then, target transcripts were identified with BETA in each TF experiment (Wang et al., 2013). All TF peaks in an experiment  $i$  inside a +/- 10 kb window from a Transcriptional Start Site (TSS) were considered to compute a regulatory score for each  $TSS_i^g$  of potential target genes of the TF. Then we defined a global cistrome regulatory score (*CRS*) for a TF  $t$  on a potential target gene  $g$  as follow:

$$CRS(t, g) = \frac{m}{N} \times \sum_{j=1}^m \sum_{i=1}^{n_m} s_j(t, TSS_i^g)$$

Where, the  $TSS_i^g$  are the  $n_m$  TSSs of  $g$  for which a regulatory score  $s_j$  by  $t$  is obtained in experiment  $j$  among the  $m$  experiments where the regulation is found. This score is weighted by the  $m/N$  ratio where  $N$  is the number of experiments for the given TF available in the considered cistrome datasets. Only *CRS* for Scenic interactions (**supplementary Tables 1&3**) or referenced regulations (**supplementary Table 3**) were retained.

## Regulon network analysis

A network based on interactions between TFs recovered in 90% (eg. 45) Scenic runs on all cells was built. The cluster\_louvain function, from igraph R package (Csardi et al., 2006), was used to find TF communities in the undirected transformation of this network with edges weighted by the NIS scores and without its self-loops. The Cytoscape software (Shannon et al., 2003), was used to visualize the results from graph clustering (**Figure 1C**).

## Activity discretization method

For the inference of a Boolean model of early haematopoiesis we selected 13 TFs Egr1, Junb, Bclaf1, Myc, Fli1, Gata2, Spi1, Cebpa, Gata1, Klfl1, Tal1, Ikzf1, Zfpm1 and 2 cells cycle complexes CDK4/6 CyclineD and CIP/KIP and we defined a set of configurations that are Boolean vectors of activities of these 15 elements. That is to say, in a configuration, elements can be active (1), inactive (0) or free (\*).

The chosen configurations are linked to the HSPC states defined on the single cell data: We considered an initial and a self-renewal, HSC configurations (iHSC, srHSC), a pre-differentiating configuration (preDiff) and the primed configuration pLymph, pNeuMast and pEr all linked to the corresponding HSPC states. We also defined a quiescent HSC configuration (qHSC) linked to the qHSC HSPC state.

To set Boolean values for each component in these configurations we discretized the regulon activities in the corresponding HSPC states of single cell data (**Figure 2Aii**). For TF with a identified activating regulons with more than 10 targets (all TF except Ikzf1, Tal1 and Zfpm1), AUCell score activity is more precise than TF RNA levels. Thus, Kmeans clustering on these scores was used to separate the cells where the regulon is active and the ones where is not. Discretized value in the configuration was then chosen according the majority cells group in the HSPC state, active if the majority of the cells present an active regulon, inactive otherwise. For the cell cycle complex genes (*Cdkn1a-b-c* for CIP/KIP and *Ccnd1-2-3, Cdk4-6* for Cyclines D-CDK4/6), Tal1 and Ikzf1 Zfpm1, because of drop out events and the resulting uncertainty of RNA level measurement in single cell RNA seq data, we averaged the RNA expression per states and discretized the results with Kmean in three groups (inactive, \* free/unknown, active **supplementary Figure 2**). For the two cell cycle complexes we took the discretization of the RNA levels genes coding for their component and attributed them a value from -1 to 1 (-1 inactive, 0 free/unknown, 1 active). Then we considered the sum of gene values for each complex: >1 active complex, <-1 inactive complex, between -1 and 1 free/unknown complex activity, **supplementary Figure 2 and 3**).

## Influence graph building

Two interactions sources were considered to build the influence graph giving the possible activations or inhibitions between the 15 elements retained (13 TF and 2 cell cycle complexes): Interaction identified thanks to Scenic workflow in at least 90% (eg 45) runs on all cells, Interactions previously found in hematopoietic cell lines or tissues (Barbeau et al., 1999; Bockamp et al., 1995; Chickarmane et al., 2009; Chou et al., 2009; Cooper et al., 2015; Crossley et al., 1994; Friedman, 2007; Grass et al., 2003; Iwasaki et al., 2003; Laiosa et al., 2006; Le Clech et al., 2006; Leddin et al., 2011; Malinge et al., 2013; Ohneda & Yamamoto, 2002; Rao et al., 2013; Starck et al., 2003; Tsai et al., 1991; Tsang et al., 1998; Walsh et al., 2002; Yeamans et al., 2007; Zhang et al., 2000). For these interactions we also considered their CRS score computed before when it was available. Some of the interactions retrieved in the literature are protein-protein physical interactions, therefor they could not be identified in the Cistrome or Scenic analysis. To be noted that we did not retained self-inhibitions recovered by scenic without any evidence in the literature because firstly inhibition from Scenic are much trustable less than activations (ref scenic) and secondly because self-regulations from Scenic rely only on the cis target step.

Finally, we defined the high confident interaction of the influence graph the ones with a strong literature evidence, that is to say described in a cell line or tissue matching with the HSCP sorting plus supported by the Cistrome analysis (for transcriptional regulations when the TF experiment is available) and/or identified with Scenic (**supplementary Table 3**).

## Dynamical analysis of Boolean networks

Dynamical analysis (eg attractors reachability from iHSC state, (un)reachabilities between states) of the inferred Boolean models was done in the Most Permissive (MP) semantics thanks to the MPBN python package (ref). In perturbed conditions, node were locked to 1 (resp 0) for Knock In alteration (resp Knock out) both in the logical rules and in the definition of the configurations. For instance, for the Knock In alteration of *Junb* related to aging, *Junb* rule becomes  $\text{Junb} = 1$  and *Junb* is leveled at one in iHSC, srHSC, qHSC, preDiff and in the primed configurations for reachability analysis.

The edgetic alteration of *Cebpa* regulation was conducted by removing the clause of the rule in which the activator is lost with aging. *Cebpa* rule becomes  $\text{Cebpa} = \text{Spi1} \wedge \overline{\text{Ikzf1}}$ .

*In-silico* KO were compared to corresponding *in-vivo/in-vitro* previous studies in mouse BM (Dell’Aversana et al., 2017; Guo et al., 2018; Gutiérrez et al., 2008; Lim et al., 1997; Mancini et al., 2012; Menendez-Gonzalez et al., 2019; Mikkola et al., 2003; Min et al., 2008; Moussa et al., 2010; Ng et al., 2009; Passegue et al., 2004; Scott et al., 1994; White et al., 2018; Wilson et al., 2004; D. E. Zhang et al., 1997).

## Boolean network inference with Bonesis

Influence graph, configurations, and dynamical constraints between them (eg. (un)reachability and attractors) were encode in Answer Set Programming (ASP) language thanks to Bonesis tool as previously described (Chevalier et al., 2019). In this way the search for a Boolean model satisfying the constraints in the MP semantics was defined as a Boolean Satisfiability Problem. Number of clauses per logical rules were limited to 3. The solver Clingo (Gebser et al., 2017) was used in the successive inference steps: firstly, a first set of dynamical constraints established from pseudotime trajectory analysis was used to make an exploration of the solution space (**Figure 2Ai**). In order to do this the heuristic of the solver was tweaked as previously described by randomly selecting a subset of variable assignments at each solution and avoiding their use by the solver in the next iterations (Chevalier et al., 2020; Gebser et al., 2017). By this way a sample of 1000 diverse solutions were obtained. For each of them in silico KO perturbation on each source node one by one were performed in the aim of recovering some mutant phenotype previously described experimentally. In the same way, in silico KI perturbation for node of TF activity upregulated upon aging were conducted. This leaded to the addition of new mutant constraints for the next inferences step (**Figure 2Bi supplementary Figure 3**).

Next inference step consisted in pruning the influence graph by adding two optimizations to reduce the number of possible solutions: in priority a maximization of the confident interaction number and then a minimization in the other interactions number in the inferred models (**Figure 2Bii**).

Finally, we considered all the interactions retrieved in inferred models maximizing the use of confident interaction and minimizing the use of ether ones. With this set of interactions, we made a final inference constraining the model to use all of them. Selection of a final model among the final set of solution was done manually, regarding previous study and favouring simpler rules (**supplementary note 3**).

## Code availability

All R, python, and ASP codes used in this study are integrated in a global snakemake workflow available at: [https://gitcrem.marseille.inserm.fr/herault/scRNA\\_infer](https://gitcrem.marseille.inserm.fr/herault/scRNA_infer).

## Statistics

Statistics were computed with R software v4.0.2. The statistical tests for regulon activity scores were performed with Seurat and are detailed above. In each primed HSPC state and in non-primed clusters gathered, the enrichment of age was tested using a hypergeometric test (phyper R function **Figure 4C**).

## Supplementary notes

### Supplementary note 1: constraint release on meta-configuration.

As no solution were obtained in a first inference try we decided to release empirically some constraints on meta-configuration by letting some component to be in a free (\*) state. First, we consider that we can allow the cycling state of some primed meta-configuration to be free as they are linked to HSPC state made of cell in different cell cycle phases. Thus, we let free CDK4/6CycD in pMk and CIPKIP in pLymph. Following the same idea, we let free the CDK4/6CycD activators Bclaf1 in pLymph, and Myc in pNeuMast, pEr and pMk. We also found that the pNeuMast states present a bimodal activity for Gata2, with this gene marking pMast and not pNeu cells (see supplementary table 1 in Héault et al., 2021). Thus, we let free Gata2 in pNeuMast HSCP states. Finally we also let free Egr1 in srHSC in agreement with a previous study suggesting its role in HSC maintenance in the niche (Min et al., 2008, p. 1).

### Supplementary note 2: Dynamical constraints from KO/KI phenotypes.

Following KO behaviours, we found for some models obtained in the solution space exploration (supplementary Figure 4): *Ikzf1* KO conducing to an absence pLymph fixpoint (Ng et al., 2009), *Spi1* KO to absences of both pNeuMast and pLymph fixpoints (Scott et al., 1994), *Klf1* KO to an absence of pEr fixpoints (Lim et al., 1997) and *Junb* KO to an apparition of an additional proliferative (active CDK4/6CycD complex) pNeuMast fixpoint (Passegué et al., 2004).

We also tested Knock In (KI) mutations on *Egr1* and *Junb* as we and others previously found upregulated in HSC upon aging (Kirschner et al., 2017) and obtained an interesting behaviour with the loss of reachability of all fixpoints expect a quiescent (CIP/KIP active) pMk. This aging phenotype matches with our previous results as we found that the pMk was the only primed cell cluster not decreased in proportion upon aging (see below and Héault et al., 2021). These 6 altered behaviours added in the constraints set of Bonesis.

### Supplementary note 3: final rule selection.

For the CDK4/6CycD we choose the rules with two clauses making the activation possible through Myc in the preDiff cells or through Bclaf1 in the self-renewal cells. For Fli1 we choose the simplest 2 clauses rule. For Gata1 we choose the rule where the auto activation is possible only when the two repressors *Ikzf1* and *Spi1* are inactive. Finally, for Gata2 on the 77 possibilities we consider the 7 simplest rules having only two clauses and chose the one where the inhibition by Gata1 and its co-factor Zfpm1 is present in both clauses.

## Supplementary table legends

**Supplementary Table 1: Transcriptional network inferred with SCENIC.** The table gives all the transcriptional interaction recovered in at least 80% (40) runs of SCENIC on all cell dataset from a TF head of a regulon toward a target gene with a mor (mode of regulation) of 1 for activation and -1 for inhibition. The recoveredTimes columns give the number of SCENIC runs in which the regulation is recovered. NIS: Normalized Interaction Score computed from importance score of SCENIC.

NIS\_diff : Normalized Interaction Score differences between NIS obtained from aged cell analysis versus NIS obtained from young cell analysis. NIS difference is 1 when the interaction is recovered in young (old) cell analysis and not in old (young) cell analysis. NIS is 0 when the interaction is recovered neither in young or old cell analysis and only in all dataset analysis. Cistrome\_BM column indicate if some ChIP-seq experiments in Cistrome database for the TF head of regulon were available and analysed (TRUE) or not (FALSE) enabling the computation of the CRS: Cistrome Regulatory Score for the interaction.

**Supplementary Table 2: Regulon activity markers of HSPC states.** Thresholds: average AUCell score difference of (avg\_diff one state vs all others) > 0.001 and p adjusted value (p\_val\_adj Wilcoxon rank sum test, Bonferroni correction) < 0.05. Thanks to Louvain clustering on the transcriptional network from SCENIC TF head of regulon were assigned to a community from C1 to C10.

**Supplementary Table 3: Influence graph interactions.** **A** Interactions between the 15 components considered for the Influence graph to infer a Boolean network. All source node are transcription factors (tf) activating, mode of regulation (mor) of 1 or inhibiting mor of -1 a target. When available reference (ref) studies characterizing experimentally the interaction are given. In that case the interaction proof level can be a transcriptional regulation, a physical protein-protein interaction (physical interaction), or a functional interaction: Knock Down (KD), KO (Know Out). This level of proof was retrieved by the reference studies in the specified cell line (cell\_line) and or cell type/tissue (cell\_type\_tissue). The interactions was confidently identified by SCENIC (present in more than 90% of the runs) analysis or not and when it was possible a Cistrome Regulatory Score (CRS) was computed. For cell cycle complexes (CIP/KIP, CDK4/6-CycD) the CRS is the sum of the CRS of each regulation of a considered TF toward one of the genes of the complex. A confidence level of A (high) or B (low) was given depending of references information and CRS and NIS (see B) score. After the pruning of low confidence level interactions 36 interactions remained in the solution (solution = TRUE). **B** SCENIC interactions considered for the Influence graph to infer a Boolean network. The table gives all the transcriptional interaction recovered in at least 90% (45) runs of SCENIC on all cell dataset from a TF head of a regulon toward a target gene with a mor (mode of regulation) of 1 for activation and -1 for inhibition. The recoveredTimes(\_young/\_aged) columns give the number of SCENIC runs on all dataset (young dataset/aged dataset) in which the regulation is recovered. NIS(\_young/\_aged): Normalized Interaction Score computed from importance score of SCENIC on all dataset (young dataset/aged dataset). NIS\_diff = NIS\_aged – NIS\_young. NIS difference is 1 when the interaction is recovered in young (old) cell analysis and not in old (young) cell analysis. NIS is 0 when the interaction is recovered neither in young or old cell analysis and only in all dataset analysis. Cistrome\_BM column indicate if some ChIP-seq experiments in Cistrome database for the TF head of regulon were available and analysed (TRUE) or not (FALSE) enabling the computation of the CRS: Cistrome Regulatory Score for the interaction. After the pruning of low confidence level interactions 36 interactions remained in the solution (solution = TRUE)

**Supplementary Table 4: Comparison of *in silico* TF KO in the the final BN selected with previous *in vivo/in vitro* studies.**

The reachability of fixed points from iHSC in the perturbed BN was assessed. Some additional fixed points compare to wild type condition were found. Some of these mutant behaviours were observed in the 1000 diverse BN solutions and constrained for the inference of the final solution.

**Supplementary Table 5: Regulons markers of aging in the different HSPC states.**

In each HSPC state Wilcoxon Rank sum tests were performed on the AUCell activity scores between young versus aged cells in batch A and B separately. Only regulons with an activity in at least 10% of either young or aged cells of the state in both batches were tested. The two p-values for each regulon were combined using the Tipett's method (minimum\_pval column). In each cluster only

regulon differences presenting the same variation and with an average score difference > 0.001 in the two batches were kept.

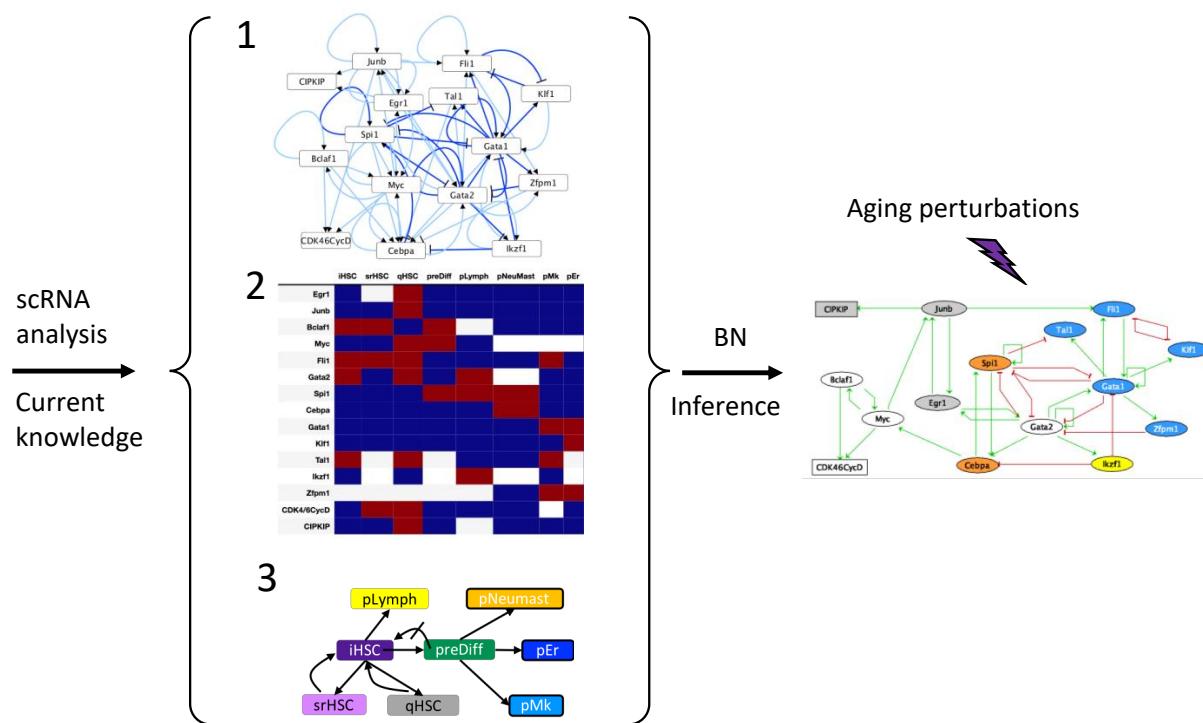
## References

- Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., & Aerts, S. (2017). SCENIC : Single-cell regulatory network inference and clustering. *Nature Methods*, 14(11), 1083-1086. <https://doi.org/10.1038/nmeth.4463>
- Barbeau, B., Barat, C., Bergeron, D., & Rassart, E. (1999). The GATA-1 and Spi-1 transcriptional factors bind to a GATA/EBS dual element in the Fli-1 exon 1. *Oncogene*, 18(40), 5535-5545. <https://doi.org/10.1038/sj.onc.1202913>
- Bernitz, J. M., Kim, H. S., MacArthur, B., Sieburg, H., & Moore, K. (2016). Hematopoietic Stem Cells Count and Remember Self-Renewal Divisions. *Cell*, 167(5), 1296-1309.e10. <https://doi.org/10.1016/j.cell.2016.10.022>
- Bockamp, E. O., McLaughlin, F., Murrell, A. M., Göttgens, B., Robb, L., Begley, C. G., & Green, A. R. (1995). Lineage-restricted regulation of the murine SCL/TAL-1 promoter. *Blood*, 86(4), 1502-1514.
- Botella, L. M., Sanz-Rodriguez, F., Komi, Y., Fernandez-L, A., Varela, E., Garrido-Martin, E. M., Narla, G., Friedman, S. L., & Kojima, S. (2009). TGF-beta regulates the expression of transcription factor KLF6 and its splice variants and promotes co-operative transactivation of common target genes through a Smad3-Sp1-KLF6 interaction. *The Biochemical Journal*, 419(2), 485-495. <https://doi.org/10.1042/BJ20081434>
- Chevalier, S., Froidevaux, C., Paulevé, L., & Zinovyev, A. (2019). Synthesis of Boolean Networks from Biological Dynamical Constraints using Answer-Set Programming. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 34-41. <https://doi.org/10.1109/ICTAI.2019.00014>
- Chevalier, S., Noël, V., Calzone, L., Zinovyev, A., & Paulevé, L. (2020). Synthesis and Simulation of Ensembles of Boolean Networks for Cell Fate Decision. *18th International Conference on Computational Methods in Systems Biology (CMSB)*, 12314, 193-209. [https://doi.org/10.1007/978-3-030-60327-4\\_11](https://doi.org/10.1007/978-3-030-60327-4_11)
- Chickarmane, V., Enver, T., & Peterson, C. (2009). Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLoS Computational Biology*, 5(1), e1000268. <https://doi.org/10.1371/journal.pcbi.1000268>
- Chou, S. T., Khandros, E., Bailey, L. C., Nichols, K. E., Vakoc, C. R., Yao, Y., Huang, Z., Crispino, J. D., Hardison, R. C., Blobel, G. A., & Weiss, M. J. (2009). Graded repression of PU.1/Sfp1 gene transcription by GATA factors regulates hematopoietic cell fate. *Blood*, 114(5), 983-994. <https://doi.org/10.1182/blood-2009-03-207944>
- Cooper, S., Guo, H., & Friedman, A. D. (2015). The +37 kb Cebpa Enhancer Is Critical for Cebpa Myeloid Gene Expression and Contains Functional Sites that Bind SCL, GATA2, C/EBP $\alpha$ , PU.1, and Additional Ets Factors. *PloS One*, 10(5), e0126385. <https://doi.org/10.1371/journal.pone.0126385>
- Crossley, M., Tsang, A. P., Bieker, J. J., & Orkin, S. H. (1994). Regulation of the erythroid Kruppel-like factor (EKLF) gene promoter by the erythroid transcription factor GATA-1. *The Journal of Biological Chemistry*, 269(22), 15440-15444.
- Csardi, G., Nepusz, T., & others. (2006). The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5), 1-9.
- Dell'Aversana, C., Giorgio, C., D'Amato, L., Lania, G., Matarese, F., Saeed, S., Di Costanzo, A., Belsito Petrizzi, V., Ingenito, C., Martens, J. H. A., Pallavicini, I., Minucci, S., Carissimo, A., Stunnenberg, H. G., & Altucci, L. (2017). MiR-194-5p/BCLAF1 deregulation in AML tumorigenesis. *Leukemia*, 31(11), 2315-2325. <https://doi.org/10.1038/leu.2017.64>
- Friedman, A. D. (2007). C/EBP $\alpha$  induces PU.1 and interacts with AP-1 and NF-kappaB to

- regulate myeloid development. *Blood Cells, Molecules & Diseases*, 39(3), 340-343. <https://doi.org/10.1016/j.bcmd.2007.06.010>
- Gebser, M., Kaminski, R., Kaufmann, B., & Schaub, T. (2017). Multi-shot ASP solving with clingo. *CoRR, abs/1705.09811*.
- Grass, J. A., Boyer, M. E., Pal, S., Wu, J., Weiss, M. J., & Bresnick, E. H. (2003). GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proceedings of the National Academy of Sciences of the United States of America*, 100(15), 8811-8816. <https://doi.org/10.1073/pnas.1432147100>
- Guo, H., Barberi, T., Suresh, R., & Friedman, A. D. (2018). Progression from the Common Lymphoid Progenitor to B/Myeloid PreproB and ProB Precursors during B Lymphopoiesis Requires C/EBP $\alpha$ . *Journal of Immunology (Baltimore, Md.: 1950)*, 201(6), 1692-1704. <https://doi.org/10.4049/jimmunol.1800244>
- Gutiérrez, L., Tsukamoto, S., Suzuki, M., Yamamoto-Mukai, H., Yamamoto, M., Philipsen, S., & Ohneda, K. (2008). Ablation of Gata1 in adult mice results in aplastic crisis, revealing its essential role in steady-state and stress erythropoiesis. *Blood*, 111(8), 4375-4385. <https://doi.org/10.1182/blood-2007-09-115121>
- He, M., Zheng, B., Zhang, Y., Zhang, X.-H., Wang, C., Yang, Z., Sun, Y., Wu, X.-L., & Wen, J.-K. (2015). KLF4 mediates the link between TGF- $\beta$ 1-induced gene transcription and H3 acetylation in vascular smooth muscle cells. *The FASEB Journal*, 29(9), 4059-4070. <https://doi.org/10.1096/fj.15-272658>
- Héault, L., Poplineau, M., Mazuel, A., Platet, N., Remy, É., & Duprez, E. (2021). Single-cell RNA-seq reveals a concomitant delay in differentiation and cell cycle of aged hematopoietic stem cells. *BMC Biology*, 19. <https://doi.org/10.1186/s12915-021-00955-z>
- Iwasaki, H., Mizuno, S., Wells, R. A., Cantor, A. B., Watanabe, S., & Akashi, K. (2003). GATA-1 converts lymphoid and myelomonocytic progenitors into the megakaryocyte/erythrocyte lineages. *Immunity*, 19(3), 451-462. [https://doi.org/10.1016/s1074-7613\(03\)00242-5](https://doi.org/10.1016/s1074-7613(03)00242-5)
- Kirschner, K., Chandra, T., Kiselev, V., Flores-Santa Cruz, D., Macaulay, I. C., Park, H. J., Li, J., Kent, D. G., Kumar, R., Pask, D. C., Hamilton, T. L., Hemberg, M., Reik, W., & Green, A. R. (2017). Proliferation Drives Aging-Related Functional Decline in a Subpopulation of the Hematopoietic Stem Cell Compartment. *Cell Reports*, 19(8), 1503-1511. <https://doi.org/10.1016/j.celrep.2017.04.074>
- Krumsiek, J., Marr, C., Schroeder, T., & Theis, F. J. (2011). Hierarchical Differentiation of Myeloid Progenitors Is Encoded in the Transcription Factor Network. *PLoS ONE*, 6(8). <https://doi.org/10.1371/journal.pone.0022649>
- Laiosa, C. V., Stadtfeld, M., & Graf, T. (2006). Determinants of lymphoid-myeloid lineage diversification. *Annual Review of Immunology*, 24, 705-738. <https://doi.org/10.1146/annurev.immunol.24.021605.090742>
- Le Clech, M., Chalhoub, E., Dohet, C., Roure, V., Fichelson, S., Moreau-Gachelin, F., & Mathieu, D. (2006). PU.1/Spi-1 binds to the human TAL-1 silencer to mediate its activity. *Journal of Molecular Biology*, 355(1), 9-19. <https://doi.org/10.1016/j.jmb.2005.10.055>
- Leedlin, M., Perrod, C., Hoogenkamp, M., Ghani, S., Assi, S., Heinz, S., Wilson, N. K., Follows, G., Schönheit, J., Vockentanz, L., Mosammam, A. M., Chen, W., Tenen, D. G., Westhead, D. R., Göttgens, B., Bonifer, C., & Rosenbauer, F. (2011). Two distinct auto-regulatory loops operate at the PU.1 locus in B cells and myeloid cells. *Blood*, 117(10), 2827-2838. <https://doi.org/10.1182/blood-2010-08-302976>
- Lim, S. K., Bieker, J. J., Lin, C. S., & Costantini, F. (1997). A shortened life span of EKLF-/- adult erythrocytes, due to a deficiency of beta-globin chains, is ameliorated by human gamma-globin chains. *Blood*, 90(3), 1291-1299.
- Liu, T., Ortiz, J. A., Taing, L., Meyer, C. A., Lee, B., Zhang, Y., Shin, H., Wong, S. S., Ma, J., Lei, Y., Pape, U. J., Poidinger, M., Chen, Y., Yeung, K., Brown, M., Turpaz, Y., & Liu, X. S. (2011). Cistrome : An integrative platform for transcriptional regulation studies. *Genome Biology*, 12(8), R83. <https://doi.org/10.1186/gb-2011-12-8-r83>
- Malinge, S., Thiollier, C., Chlon, T. M., Doré, L. C., Diebold, L., Bluteau, O., Mabialah, V.,

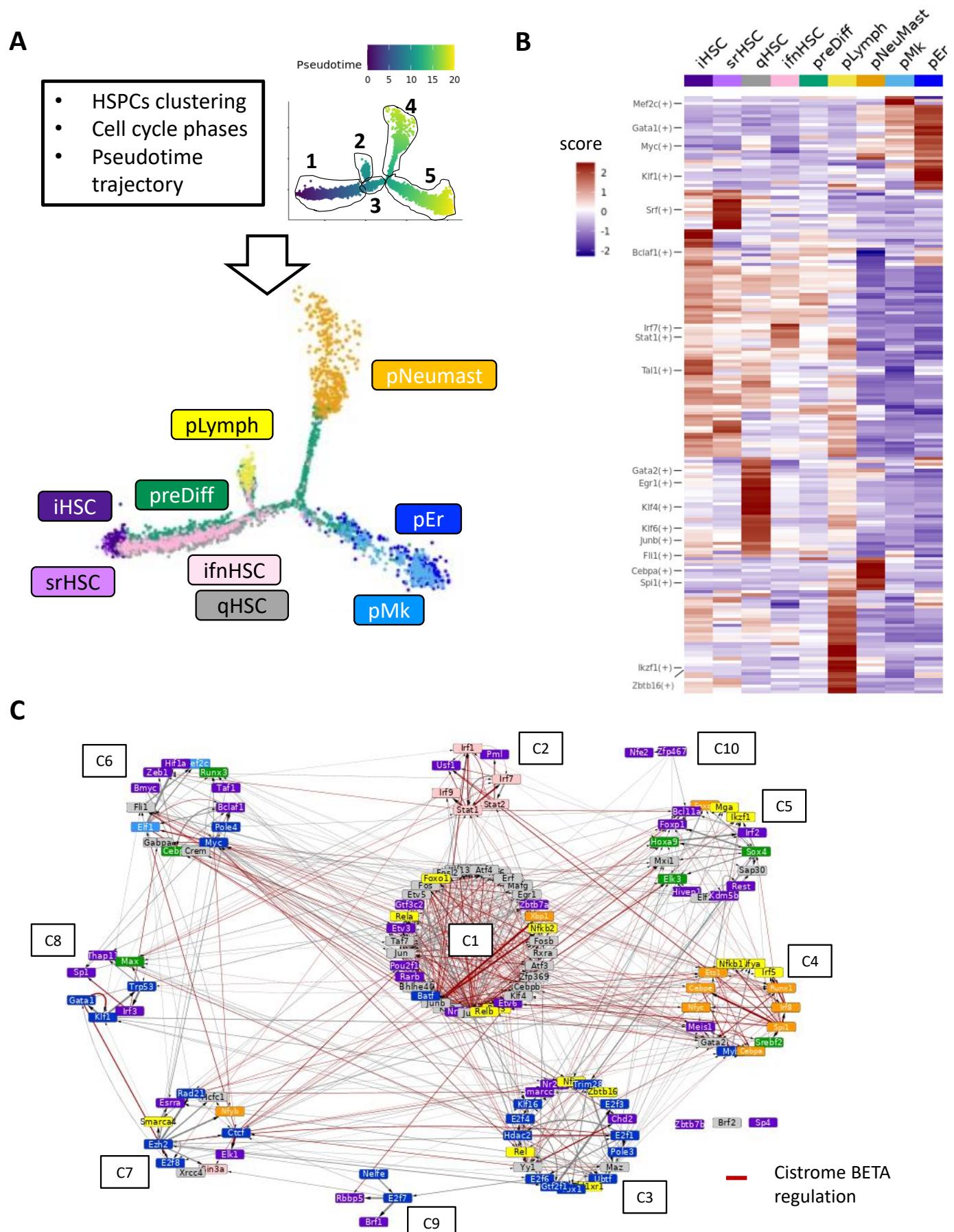
- Vainchenker, W., Dessen, P., Winandy, S., Mercher, T., & Crispino, J. D. (2013). Ikaros inhibits megakaryopoiesis through functional interaction with GATA-1 and NOTCH signaling. *Blood*, 121(13), 2440-2451. <https://doi.org/10.1182/blood-2012-08-450627>
- Mancini, E., Sanjuan-Pla, A., Luciani, L., Moore, S., Grover, A., Zay, A., Rasmussen, K. D., Luc, S., Bilbao, D., O'Carroll, D., Jacobsen, S. E., & Nerlov, C. (2012). FOG-1 and GATA-1 act sequentially to specify definitive megakaryocytic and erythroid progenitors. *The EMBO Journal*, 31(2), 351-365. <https://doi.org/10.1038/emboj.2011.390>
- Menendez-Gonzalez, J. B., Vukovic, M., Abdelfattah, A., Saleh, L., Almotiri, A., Thomas, L., Agirre-Lizaso, A., Azevedo, A., Menezes, A. C., Tornillo, G., Edkins, S., Kong, K., Giles, P., Anjos-Afonso, F., Tonks, A., Boyd, A. S., Kranc, K. R., & Rodrigues, N. P. (2019). Gata2 as a Crucial Regulator of Stem Cells in Adult Hematopoiesis and Acute Myeloid Leukemia. *Stem Cell Reports*, 13(2), 291-306. <https://doi.org/10.1016/j.stemcr.2019.07.005>
- Mikkola, H. K. A., Klintman, J., Yang, H., Hock, H., Schlaeger, T. M., Fujiwara, Y., & Orkin, S. H. (2003). Haematopoietic stem cells retain long-term repopulating activity and multipotency in the absence of stem-cell leukaemia SCL/ tal-1 gene. *Nature*, 421(6922), 547-551. <https://doi.org/10.1038/nature01345>
- Min, I. M., Pietramaggiori, G., Kim, F. S., Passegue, E., Stevenson, K. E., & Wagers, A. J. (2008). The transcription factor EGR1 controls both the proliferation and localization of hematopoietic stem cells. *Cell Stem Cell*, 2(4), 380-391. <https://doi.org/10.1016/j.stem.2008.01.015>
- Moussa, O., LaRue, A. C., Abangan, R. S., Williams, C. R., Zhang, X. K., Masuya, M., Gong, Y. Z., Spyropoulos, D. D., Ogawa, M., Gilkeson, G., & Watson, D. K. (2010). Thrombocytopenia in mice lacking the carboxy-terminal regulatory domain of the Ets transcription factor Fli1. *Molecular and Cellular Biology*, 30(21), 5194-5206. <https://doi.org/10.1128/MCB.01112-09>
- Ng, S. Y.-M., Yoshida, T., Zhang, J., & Georgopoulos, K. (2009). Genome-wide lineage-specific transcriptional networks underscore Ikaros-dependent lymphoid priming in hematopoietic stem cells. *Immunity*, 30(4), 493-507. <https://doi.org/10.1016/j.immuni.2009.01.014>
- Ohneda, K., & Yamamoto, M. (2002). Roles of hematopoietic transcription factors GATA-1 and GATA-2 in the development of red blood cell lineage. *Acta Haematologica*, 108(4), 237-245. <https://doi.org/10.1159/000065660>
- Passegue, E., Wagner, E. F., & Weissman, I. L. (2004). JunB deficiency leads to a myeloproliferative disorder arising from hematopoietic stem cells. *Cell*, 119(3), 431-443. <https://doi.org/10.1016/j.cell.2004.10.010>
- Paulevé, L., Kolčák, J., Chatain, T., & Haar, S. (2020). Reconciling qualitative, abstract, and scalable modeling of biological networks. *Nature Communications*, 11. <https://doi.org/10.1038/s41467-020-18112-5>
- Pietras, E. M., Warr, M. R., & Passegue, E. (2011). Cell cycle regulation in hematopoietic stem cells. *The Journal of Cell Biology*, 195(5), 709-720. <https://doi.org/10.1083/jcb.201102131>
- Rao, K. N., Smuda, C., Gregory, G. D., Min, B., & Brown, M. A. (2013). Ikaros limits basophil development by suppressing C/EBP- $\alpha$  expression. *Blood*, 122(15), 2572-2581. <https://doi.org/10.1182/blood-2013-04-494625>
- Santaguida, M., Schepers, K., King, B., Sabnis, A. J., Forsberg, E. C., Attema, J. L., Braun, B. S., & Passegue, E. (2009). JunB protects against myeloid malignancies by limiting hematopoietic stem cell proliferation and differentiation without affecting self-renewal. *Cancer Cell*, 15(4), 341-352. <https://doi.org/10.1016/j.ccr.2009.02.016>
- Scott, E. W., Simon, M. C., Anastasi, J., & Singh, H. (1994). Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science (New York, N.Y.)*, 265(5178), 1573-1577. <https://doi.org/10.1126/science.8079170>
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape : A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498-2504. <https://doi.org/10.1101/gr.1239303>
- Starck, J., Cohet, N., Gonnet, C., Sarrazin, S., Doubeikovskaya, Z., Doubeikovski, A., Verger, A., Duterque-Coquillaud, M., & Morle, F. (2003). Functional cross-antagonism between transcription

- factors FLI-1 and EKLF. *Molecular and Cellular Biology*, 23(4), 1390-1402. <https://doi.org/10.1128/mcb.23.4.1390-1402.2003>
- Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, 177(7), 1888-1902.e21. <https://doi.org/10.1016/j.cell.2019.05.031>
- Sun, D., Luo, M., Jeong, M., Rodriguez, B., Xia, Z., Hannah, R., Wang, H., Le, T., Faull, K. F., Chen, R., Gu, H., Bock, C., Meissner, A., Göttgens, B., Darlington, G. J., Li, W., & Goodell, M. A. (2014). Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal. *Cell stem cell*, 14(5), 673-688. <https://doi.org/10.1016/j.stem.2014.03.002>
- Tsai, S. F., Strauss, E., & Orkin, S. H. (1991). Functional analysis and in vivo footprinting implicate the erythroid transcription factor GATA-1 as a positive regulator of its own promoter. *Genes & Development*, 5(6), 919-931. <https://doi.org/10.1101/gad.5.6.919>
- Tsang, A. P., Fujiwara, Y., Hom, D. B., & Orkin, S. H. (1998). Failure of megakaryopoiesis and arrested erythropoiesis in mice lacking the GATA-1 transcriptional cofactor FOG. *Genes & Development*, 12(8), 1176-1188. <https://doi.org/10.1101/gad.12.8.1176>
- Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Aibar, S., Seurinck, R., Saelens, W., Cannoodt, R., Rouchon, Q., Verbeiren, T., De Maeyer, D., Reumers, J., Saeys, Y., & Aerts, S. (2020). A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nature Protocols*, 15(7), 2247-2276. <https://doi.org/10.1038/s41596-020-0336-2>
- Walsh, J. C., DeKoter, R. P., Lee, H. J., Smith, E. D., Lancki, D. W., Gurish, M. F., Friend, D. S., Stevens, R. L., Anastasi, J., & Singh, H. (2002). Cooperative and antagonistic interplay between PU.1 and GATA-2 in the specification of myeloid cell fates. *Immunity*, 17(5), 665-676. [https://doi.org/10.1016/s1074-7613\(02\)00452-1](https://doi.org/10.1016/s1074-7613(02)00452-1)
- Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., Tang, Q., Meyer, C. A., Zhang, Y., & Liu, X. S. (2013). Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nature Protocols*, 8(12), 2502-2515. <https://doi.org/10.1038/nprot.2013.150>
- White, L. S., Soodgupta, D., Johnston, R. L., Magee, J. A., & Bednarski, J. J. (2018). Bclaf1 Promotes Maintenance and Self-Renewal of Fetal Hematopoietic Stem Cells. *Blood*, 132(Supplement 1), 1269-1269. <https://doi.org/10.1182/blood-2018-99-114144>
- Wiedemann, D. (1991). A computation of the eighth Dedekind number. *Order*, 8(1), 5-6. <https://doi.org/10.1007/BF00385808>
- Wilson, A., Laurenti, E., Oser, G., van der Wath, R. C., Blanco-Bose, W., Jaworski, M., Offner, S., Dunant, C. F., Eshkind, L., Bockamp, E., Lió, P., MacDonald, H. R., & Trumpp, A. (2008). Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair. *Cell*, 135(6), 1118-1129. <https://doi.org/10.1016/j.cell.2008.10.048>
- Wilson, A., Murphy, M. J., Oskarsson, T., Kaloulis, K., Bettess, M. D., Oser, G. M., Pasche, A.-C., Knabenhans, C., MacDonald, H. R., & Trumpp, A. (2004). C-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation. *Genes & Development*, 18(22), 2747-2763. <https://doi.org/10.1101/gad.313104>
- Yan, X., Xiong, X., & Chen, Y.-G. (2018). Feedback regulation of TGF-β signaling. *Acta Biochimica et Biophysica Sinica*, 50(1), 37-50. <https://doi.org/10.1093/abbs/gmx129>
- Yeamans, C., Wang, D., Paz-Priel, I., Torbett, B. E., Tenen, D. G., & Friedman, A. D. (2007). C/EBPalpha binds and activates the PU.1 distal enhancer to induce monocyte lineage commitment. *Blood*, 110(9), 3136-3142. <https://doi.org/10.1182/blood-2007-03-080291>
- Zhang, D. E., Zhang, P., Wang, N. D., Hetherington, C. J., Darlington, G. J., & Tenen, D. G. (1997). Absence of granulocyte colony-stimulating factor signaling and neutrophil development in CCAAT enhancer binding protein alpha-deficient mice. *Proceedings of the National Academy of Sciences of the United States of America*, 94(2), 569-574. <https://doi.org/10.1073/pnas.94.2.569>
- Zhang, P., Zhang, X., Iwama, A., Yu, C., Smith, K. A., Mueller, B. U., Narravula, S., Torbett, B. E., Orkin, S. H., & Tenen, D. G. (2000). PU.1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. *Blood*, 96(8), 2641-2648.



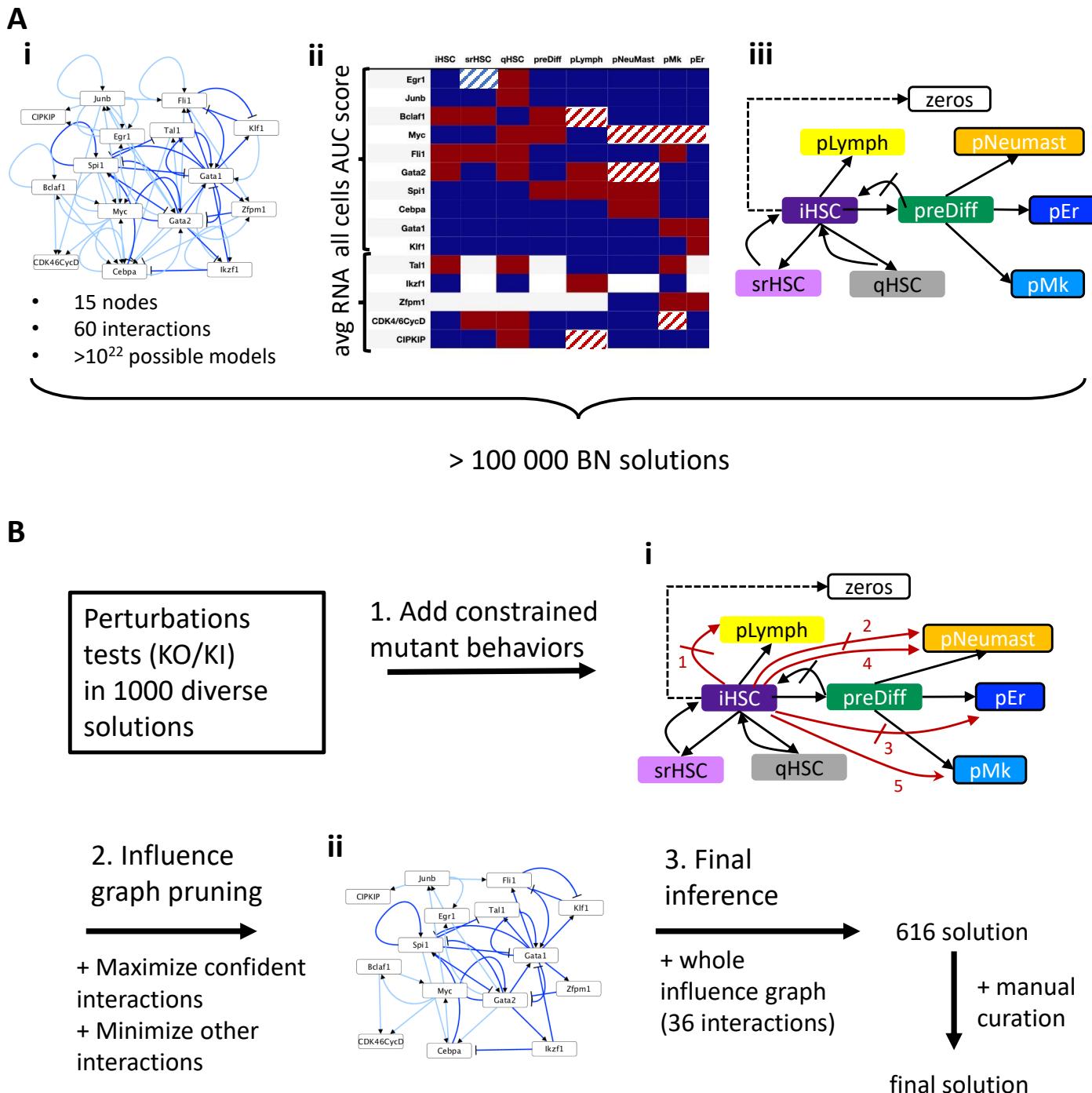
**Graphical abstract: Overview of scRNA seq assisted gene Boolean network synthesis strategy.** From single cell-RNA seq data and current knowledge in early hematopoiesis (literature and biological database investigation) 3 inputs are obtained to define the network synthesis as a Boolean Satisfiability Problem depending on observations of states in the differentiation process : **1** influence graph of the possible component interactions, **2** discretized component activity levels in the considered states (blue: 0, inactive, white: \*, unknown/free, red: 1, active). **3** dynamic relations ((non) reachability, attractors) between the considered states. Then, these inputs are encoded as constraints in Answer Set Programming (ASP) thanks to Bonesis tool and after the solving, a final solution of a Boolean model of early hematopoiesis was obtained. This model was altered regarding aging features in single cell RNA seq data to identify key aging molecular actors and mechanisms.

## 4 Résultat : Inférence d'un modèle logique de l'hématopoïèse précoce altérée par le vieillissement – 4.3 Résultats



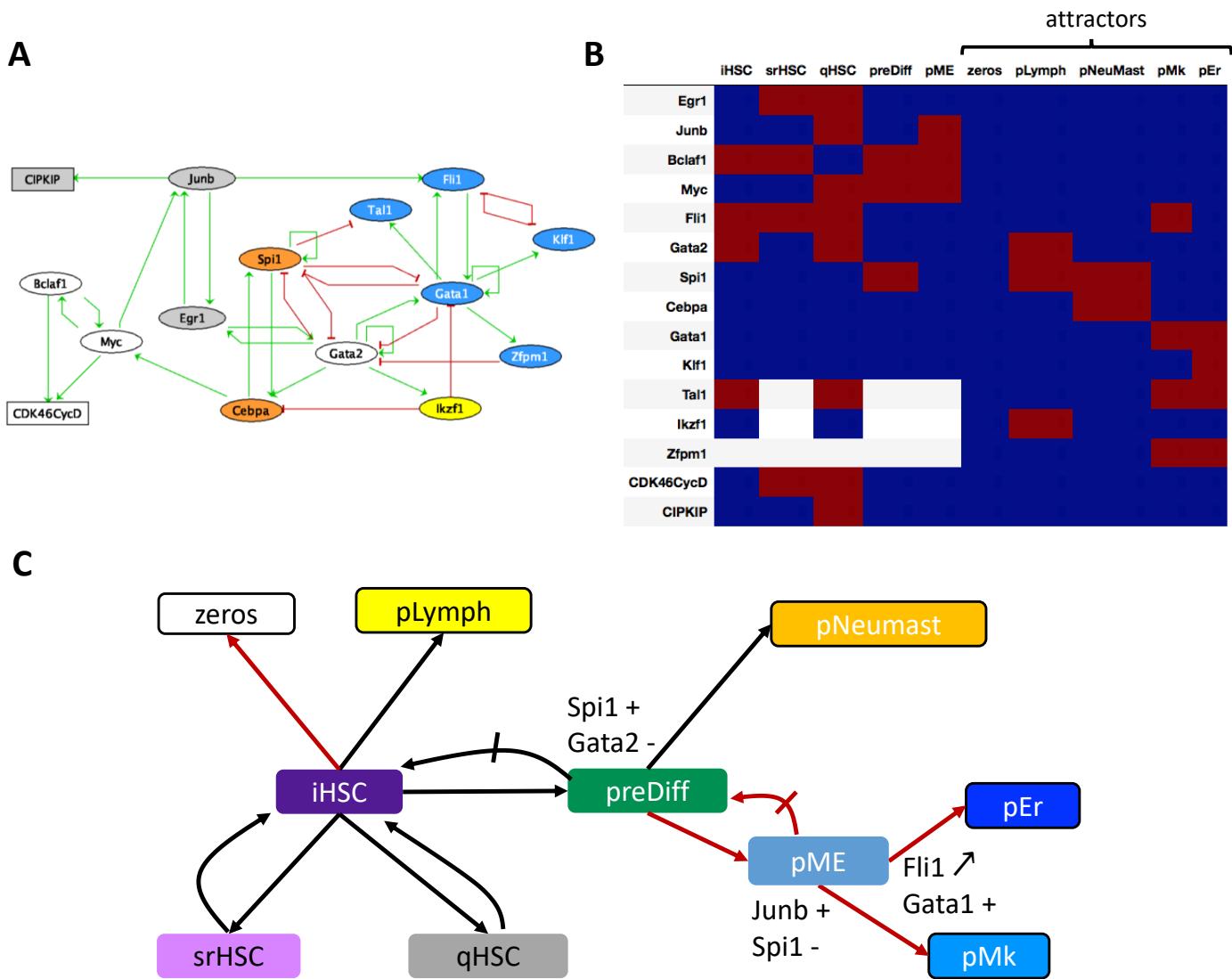
**Figure 1: Regulon activities and network sustain HSC homeostasis and priming.**  
Legend next page

**Figure 1: Regulon activities and network sustain HSC homeostasis and priming.** A HSPC states defined according to results of cell clustering, cell cycle phase assignment and pseudotime trajectory analysis of scRNA-seq data (Héroult et al., 2021). On the right of the upper panel, trajectory is presented; cells are coloured according to their pseudotime values and the 5 branches of the trajectory are circled. The main panel presents cells from the defined HSPC states (labels) in the pseudotime trajectory and coloured accordingly. B Heatmap of AUCell scores for regulons averaged by groups of cells from the HSPC states. Scores were standardized and used to hierarchically cluster the regulons. C Transcriptional regulation network of the regulon markers of the defined HSPC states. Regulons were clustered in 10 communities (from C1 to C10 plus 3 isolated nodes) with Louvain graph clustering. Node label color highlights the states where the regulon is the most active (same color code as in B). Red (grey) edges indicate that the transcriptional regulation is (is not) supported by peak analysis in Cistrome database with BETA tool. Edge thickness represents the normalized interaction score (NIS) obtained from SCENIC.



**Figure 2: Inference of Boolean gene network fitting scRNA-seq observations with Answer Set Programming.** A. First inference with wild-type constraints. i Retained influence graph of possible component interactions from literature investigation and SCENIC results. Interactions with a high (low) confidence level are in dark (pale) blue. Legend continues on the next page.

**ii** Discretization of component activities in the configurations. TF head of a regulons with more than 10 targets were discretized with a 2 clustering Kmeans on all cell regulons activities scores (major cell proportion group in each corresponding state in the data). Blue: inactive; red: active. Other TFs (Tal1, Ikzf1 and Zfpm1) and gene belonging to the two complexes CDK4/6CycD and CIPKIP were discretized with a 3 clustering Kmeans on averaged RNA levels in the corresponding states. Blue: inactive; white: unknown/free; red: active. For the complexes the final discretization was deduced from the sum of the discretization of each of its genes (see **supplementary Figure 2**). Red (resp. blue) hatched cases mark node activities freed from 1 (resp 0) to \* in the final configuration settings compare to the discretized data in order to get some solutions. **iii** Graph representation of the dynamical constraints (edges) defined between the configurations (nodes). Arrow (crossed arrow) indicate reachability (resp. unreachability) between source and target configurations. Framed configurations are constrained as fixpoints. **B** Strategy to refine the solution search and obtain a final solution. **i** Updated constraints (in red) after perturbations analysis of 1000 diverse solutions inferred with the wild type constraints. 1: loss of pLymph reachability with *Ikzf1/Spi1* KO, 2: loss of pNeuMast reachability with *Spi1* KO, 3 loss of pEr reachability with *Klf1* KO, 4: additional pNeuMast cycling fixpoint with *Junb* KO, 5: a unique pMk quiescent fixpoint with *Junb/Egr1* KI. **ii** Pruned influence graph through maximization (minimization) of high-confident (others) interactions. A last inference on this whole pruned influence graph resulted in 616 possible solutions. A final model was selected with manual curation of the possible rules.

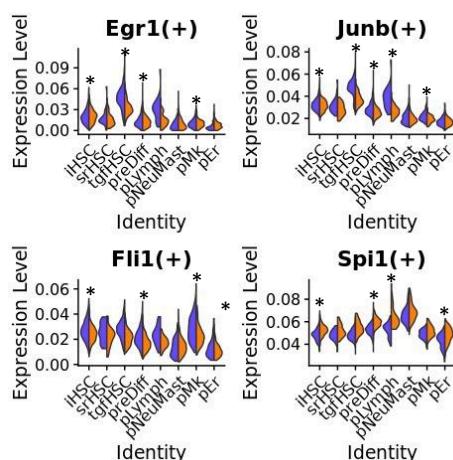


**Figure 3: The inferred Boolean model of early hematopoiesis gives some insights about branching dynamic during HSC priming.** A Gene regulatory network of the inferred Boolean model. Rectangular nodes are cell cycle complexes and ellipse node TFs. Nodes are colored according to the HSPC states in which they are highly active according to our single cell analysis: Grey for qHSC, yellow for pL, orange for pNeuMast , blue for pMk and pEr, white for the nodes highly active in several HSPC states. B Table relating HSPC and pME configurations identified by the model analysis (column: HSPC states, lines: components of the model). Colors represent the activation levels of the nodes (blue : inactive; red : active, white: free). C Graph representation of the dynamics between the configurations (nodes) from iHSC toward the different fixed points (framed nodes). Arrows (resp. crossed arrows) indicate reachability (resp. unreachability) between source and target configuration. Black edges are constrained dynamic properties whereas the red ones result from the dynamic study of the model. Zfpm1\* highlights the two possible values of this node in iHSC . Spi1+, Gata2 – indicates the irreversible inactivation of Gata2 by Spi1 observed in the preDiff non-return configuration. Junb+, Spi1- indicates the necessary change in preDiff to reach the branching configuration pME reveled by the dynamic analysis.

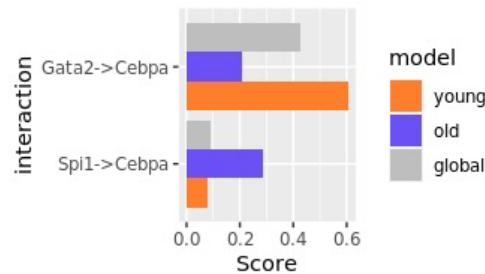
Component	Logical rules
<i>Egr1</i>	<i>Gata2</i> $\wedge$ <i>Junb</i>
<i>Junb</i>	<i>Egr1</i> $\vee$ <i>Myc</i>
<i>Bclaf1</i>	<i>Myc</i>
<i>Myc</i>	<i>Cebpa</i> $\wedge$ <i>Bclaf1</i>
<i>Fli1</i>	<i>Junb</i> $\vee$ ( <i>Gata1</i> $\wedge$ $\overline{Klf1}$ )
<i>Gata2</i>	( <i>Gata2</i> $\wedge$ $\overline{Gata1}$ $\wedge$ $\overline{Zfpm1}$ ) $\vee$ ( <i>Egr1</i> $\wedge$ $\overline{Gata1}$ $\wedge$ $\overline{Zfpm1}$ $\wedge$ $\overline{Spi1}$ )
<i>Spi1</i>	( <i>Spi1</i> $\wedge$ $\overline{Gata1}$ ) $\vee$ ( <i>Cebpa</i> $\wedge$ $\overline{Gata1}$ $\wedge$ $\overline{Gata2}$ )
<i>Cebpa</i>	( <i>Gata2</i> $\wedge$ $\overline{Ikzf1}$ ) $\vee$ ( <i>Spi1</i> $\wedge$ $\overline{Ikzf1}$ )
<i>Gata1</i>	<i>Fli1</i> $\vee$ ( <i>Gata2</i> $\wedge$ $\overline{Spi1}$ ) $\vee$ ( <i>Gata1</i> $\wedge$ $\overline{Ikzf1}$ $\wedge$ $\overline{Spi1}$ )
<i>Klf1</i>	<i>Gata1</i> $\wedge$ <i>Fli1</i>
<i>Tal1</i>	<i>Gata1</i> $\wedge$ <i>Spi1</i>
<i>Ikzf1</i>	<i>Gata2</i>
<i>Zfpm1</i>	<i>Gata1</i>
<i>CDK46CycD</i>	<i>Bclaf1</i> $\vee$ <i>Myc</i>
<i>CIPKIP</i>	<i>Junb</i>

Table 1 : Logical rules of HSC priming Boolean model.

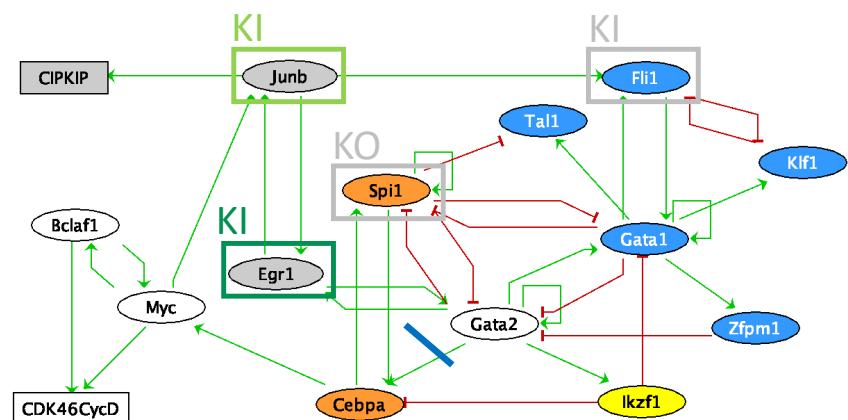
A



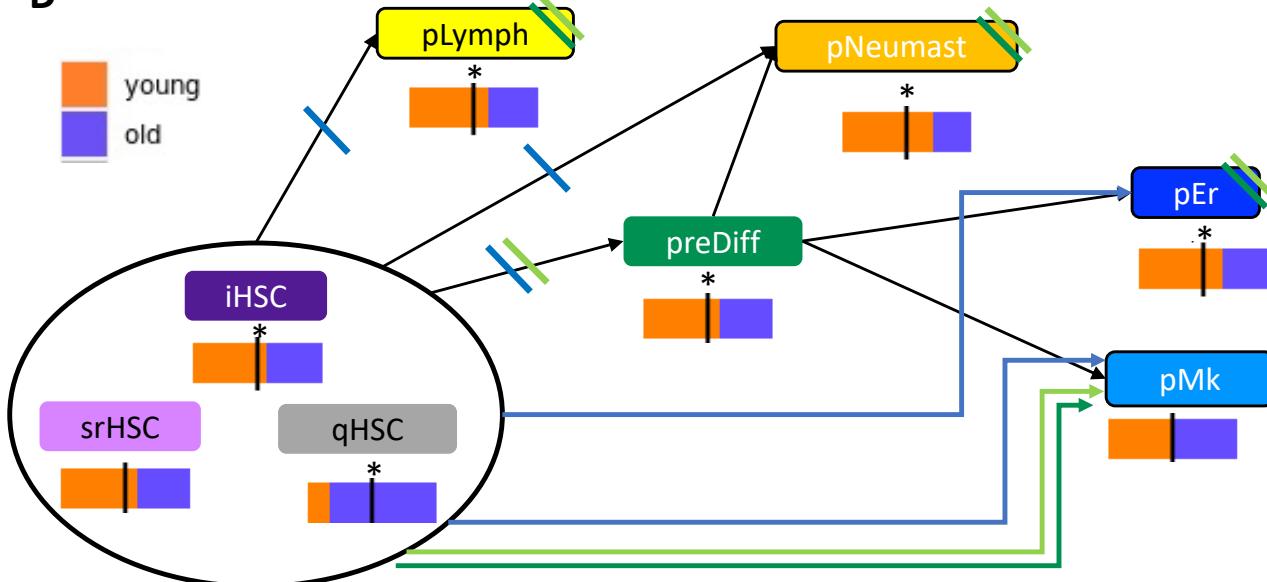
B



C



D

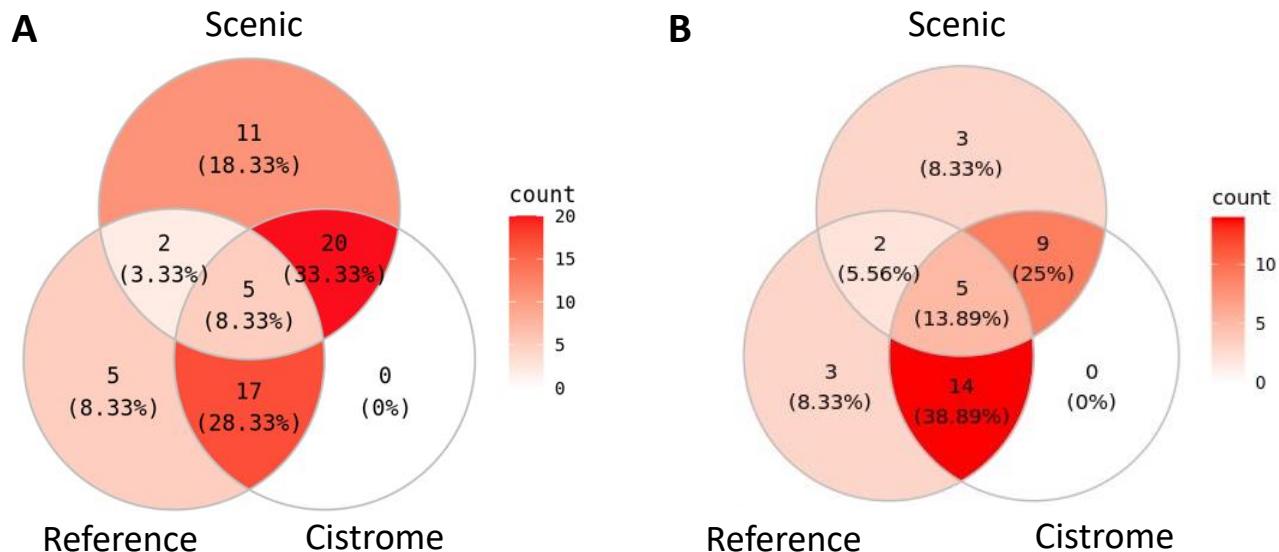


**Figure 4: Perturbations of early hematopoiesis model explains some HSC aging features.** A Combined violin plot of most altered TF (of the model) activities upon aging in young (orange) and aged (purple) cells from the different HSPC state. Stars show significant differences of activity score between young and aged cells (average difference  $> 0.001$  and  $p$  value  $< 10^{-3}$ ). B Aging perturbations of the Boolean gene network of early hematopoiesis. Rectangular nodes are cell cycle complexes and ellipse nodes TFs. Nodes are coloured according to the HSPC states in which they are highly active according to our single cell analysis: Grey for qHSC, yellow for pL, orange for pNeuMast , blue for pMk and pEr, white for the nodes highly active in several HSPC states. Framed nodes highlight KO/KI perturbations. Crossed activation of Cebpa by Spi1 illustrates its edgetic mutation. Legend continues on the next page.

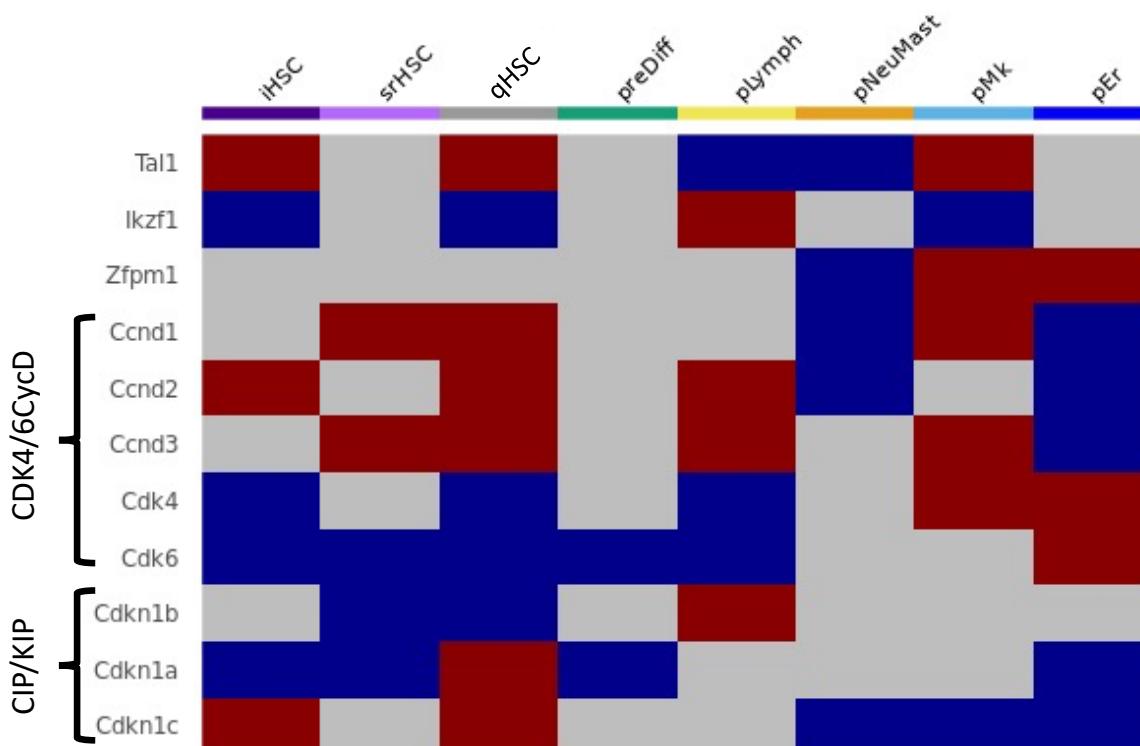
**C** Normalized interaction scores of *Cebpa* activation by *Spi1* and *Gata2* from scenic multiple runs on all cells (grey), only young cells (orange) and only aged cells (purple).  
**D** Altered dynamic of the model following aging perturbation *Egr1* (dark green) *Junb* (pale green) and *Cebpa* edgetic mutation (blue). In the considered perturbation, crossed fixpoints are lost and crossed arrow highlights a loss of reachability of a HSCP state from any configurations in i-sr-qHSCs. Young (orange) and aged (purple) cell proportion is given below each HSPC state node. A star highlights a significant shift from expected proportion (hypergeometric test p value < 0.05).

Perturbation	scRNA seq data evidences	Reachabilities from i-sr-qHSC
WT		pEr, pMk, preDiff, pLymph, pNeuMast
<i>Junb</i> KI	Junb ↑ in iHSC, qHSC, preDiff, pLymph, pMk;	pEr, pMk, preDiff, pLymph, pNeuMast
<i>Egr1</i> KI	Egr1 ↑ in iHSC, qHSC, preDiff, pMk	pEr, pMk, preDiff, pLymph, pNeuMast,
<i>Spi1</i> KO	Spi1 ↓ in iHSC, preDiff, pLymph, pEr	pEr, pMk, preDiff, pLymph, pNeuMast
<i>Fli1</i> KI	Fli1 ↑ in iHSC, preDiff, pEr, pMk;	pEr, pMk, preDiff, pEr, pLymph, pNeuMast,
<i>Cebpa</i> = ( <i>Spi1</i> ∧ $\overline{Ikzf1}$ )	Gata2 → Cebpa ↓ Spi1 → Cebpa ↑	pEr, pMk, preDiff, pNeuMast, pLymph,

**Table 2: Aging perturbation of early hematopoiesis model inferred.** Aging perturbations were modelized by logical rule modification supported by our scRNA seq data analysis and lead to alteration of the dynamics regarding reachability from i-sr-qHSC configurations. Up (resp.down) arrows indicate an increase (resp a decrease) upon aging in component activity or interaction score revealed by our scRNA-seq analysis. States becoming unreachable from i-sr-qHSC with the perturbation are striped. For instance the update Egr1 rule to *Egr1* = 1 is supported by an increase in Egr1 activity in the specified states and lead to the loss of pLymph and pNeuMast fixpoints reachability from i-sr-qHSC.

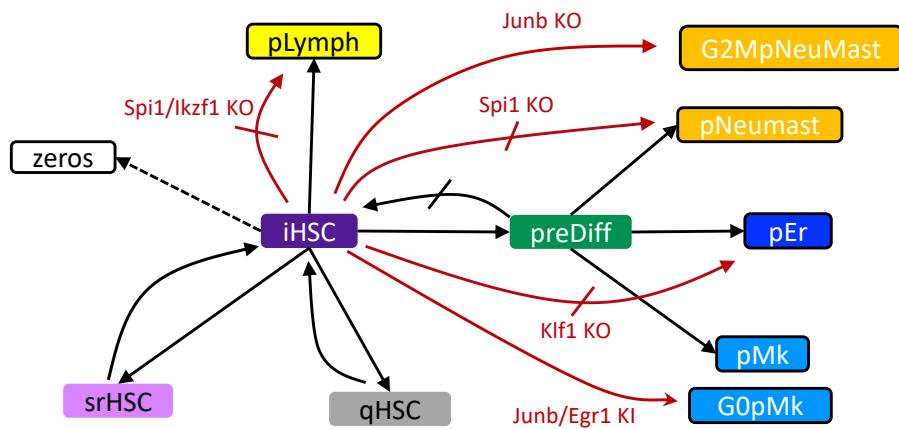


**Supplementary Figure 1: Venn diagram of influence graph interactions sources.** **A** Initial influence graph interactions retrieved from Scenic results and or literature investigation and supported or not by the Cistrome database analysis (see supplementary table 3). **B** Same for interactions remaining after the influence graph pruning.

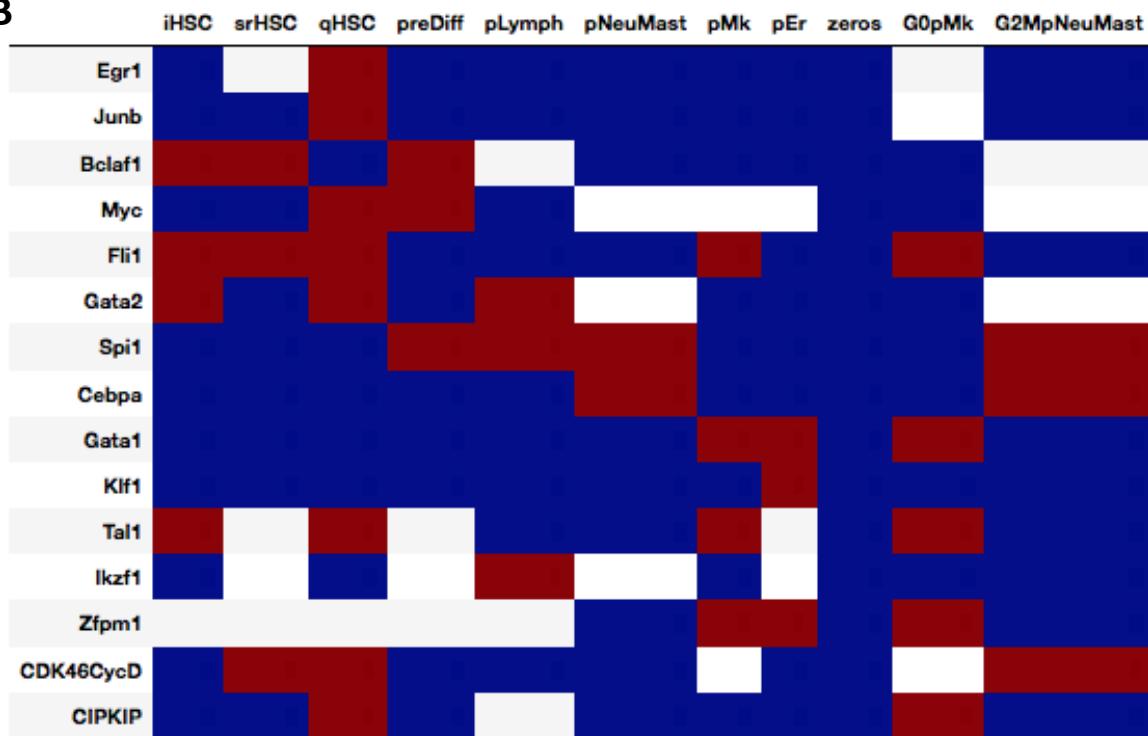


**Supplementary Figure 2: Discretization of gene expressions for genes and regulons with less than 10 targets.** Results of k-means clustering on averaged RNA levels of the selected HSPC states. Blue: inactivated; grey: unknown/free; red: activated.

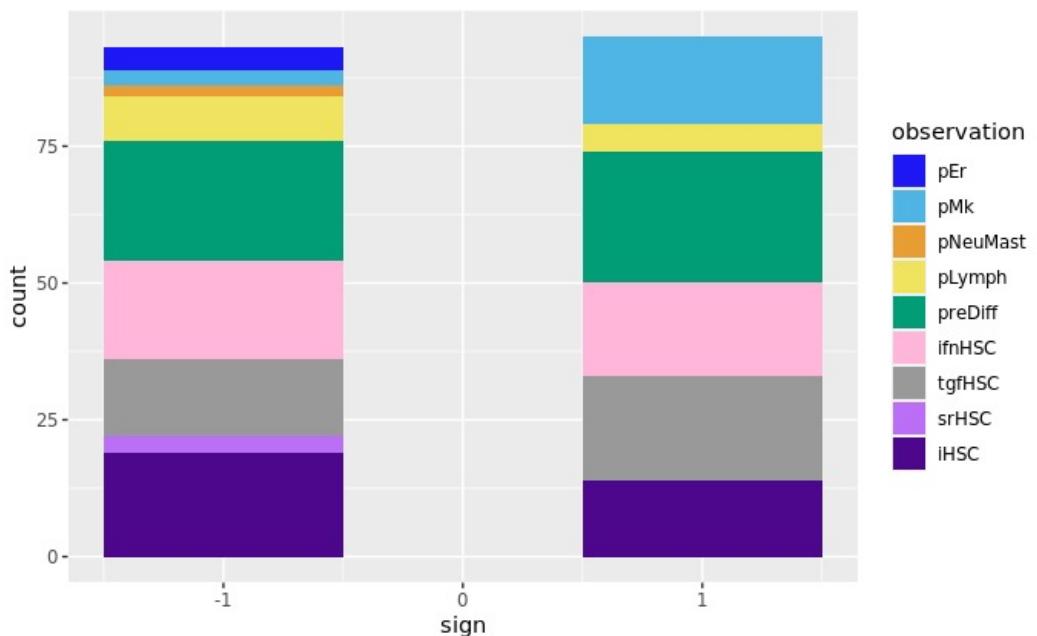
A



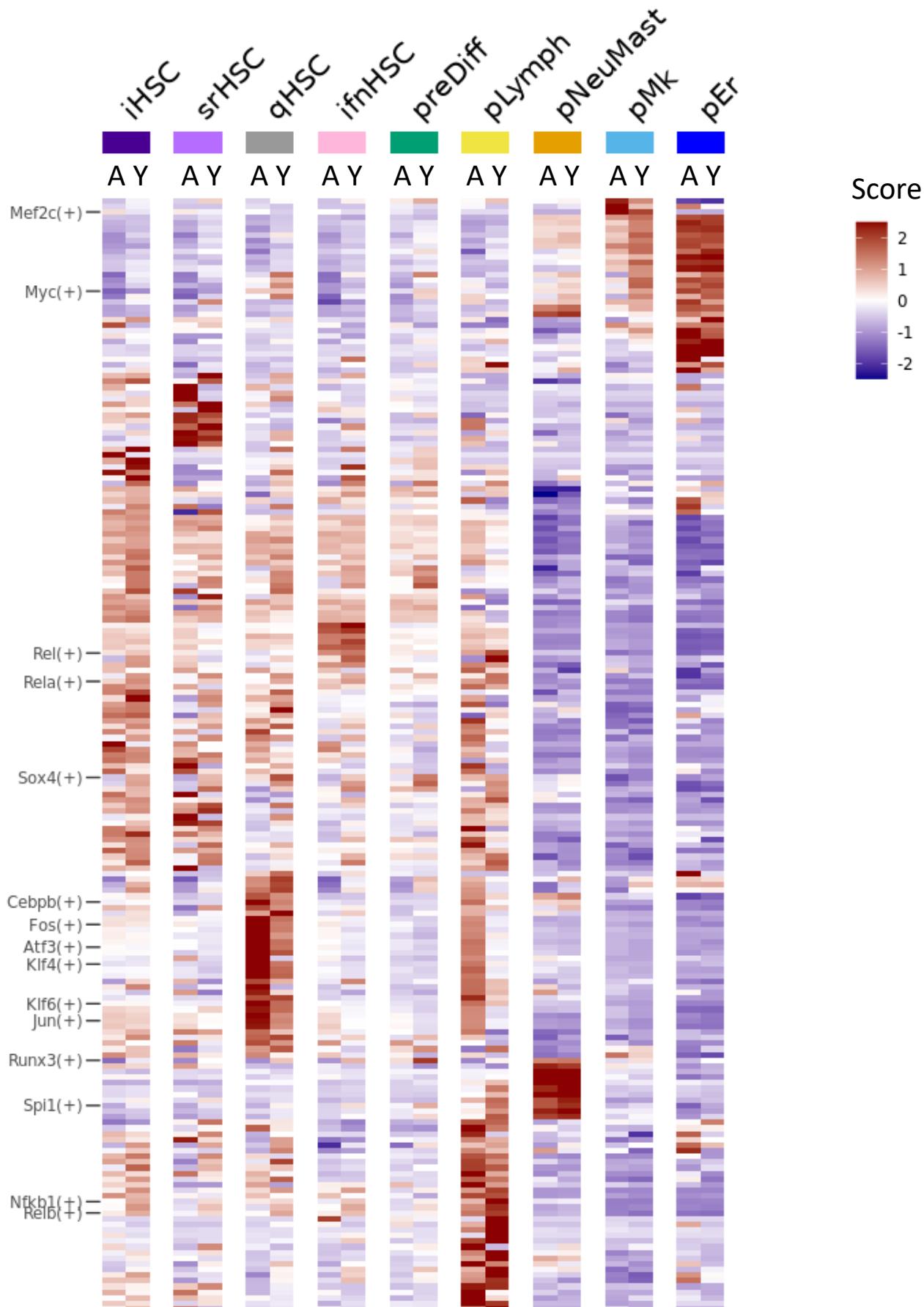
B



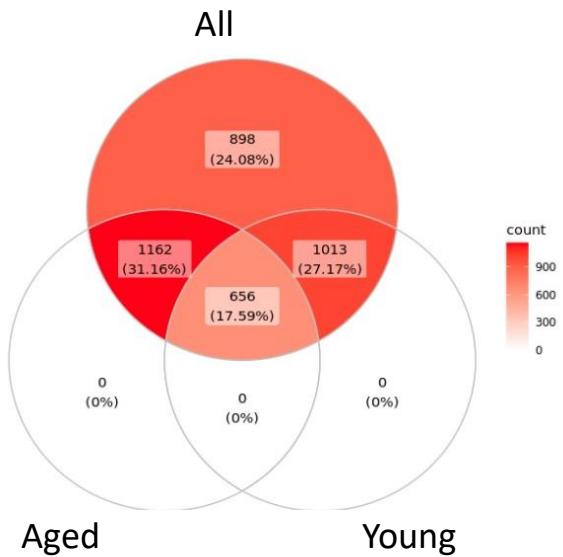
**Supplementary Figure 3: Constraints and discretization used for the influence graph pruning and the final rule inference.** A Dynamical constraints used for the pruning of the influence graph and the final rule inference. Arrows (resp. crossed arrows) indicate reachability (resp. unreachability) between source and target configuration. Framed configurations are constrained as fixpoints. Dashed line highlights the allowed reachability of a fixpoint with all node activities at 0 from iHSC. Red (crossed) arrow highlights the additional (non) reachabilities constraints of mutant behaviors. G0pMk is the only reachable fixpoint from iHSC in pEr/pMk KI (large blue arrow). B Discretization of component activities in the configurations used for the pruning of the influence graph and the final rule inference. Blue: 0, inactivated; white: \*, unknown/free; red: 1, activated. G0pMk and G2MpNeuMast configuration were defined according to the first solution space exploration.



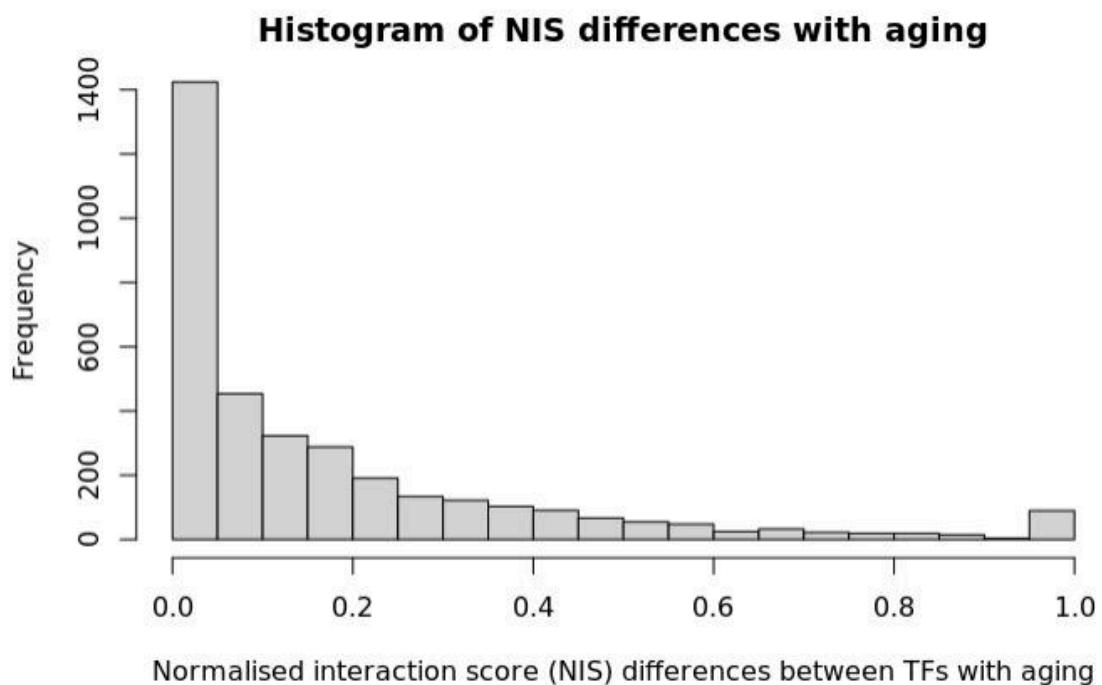
**Supplementary Figure 4: HSPC state repartition of the regulon activity alterations upon aging.** 1 (resp -1) marks significant increase (decrease) in activity upon aging (average differences > 0.001, p-value < 10-3). An alteration can be recovered in several HSPC state.



**Supplementary Figure 5: Heatmap of AUCell enrichment scores of regulons activity averaged by group of cells from the selected HSPC states in young and aged cells.** Scores were standardized on aged (A) and young (Y) cells of the different states. Row are ordered as in Figure 1.



**Supplementary Figure 6:** retrieved scenic interaction from all cell analysis in aged or young only cell analysis. We did not consider interactions retrieved only in young or only in aged cells.



**Supplementary Figure 7: Histogram of normalized interaction score differences with aging.** Interactions non retrieved in SCENIC analysis of aged or young cells have a null difference. A difference of one points out that a regulation was retrieved only in young or only in aged cells analysis

## 4.4 Discussion

Plusieurs BN ont auparavant été établis pour décrire le choix de différenciation de la CSH vers les progéniteurs lymphoïdes ou myéloïdes (COLLOMBET et al., 2017; HAMEY et al., 2017). Cependant, les trajectoires de ces modèles se révèlent incomplètes vis à vis des 5 états amorcées situés aux 3 extrémités de la pseudo-trajectoire que nous avons établie à partir de nos données scRNA-seq (HÉRAULT et al., 2021). Un autre BN décrit quant à lui le maintien d'un état non amorcé de HSPC qui se différencie ensuite vers les différents types cellulaires matures (BONZANNI et al., 2013), ce qui ne correspond pas à l'amorçage anticipé que nous observons dans le compartiment HSPC. Afin de proposer un nouveau modèle de l'amorçage de la CSH en accord avec nos observations, nous nous sommes appuyés dans cette étude sur un modèle de la différenciation précoce des progéniteurs myéloïdes qui, à l'inverse des modèles précédemment cités, présente des configurations d'états stables et d'embranchements très proches des profils d'activités de TF que nous avons mesurés (KRUMSIEK et al., 2011). Nous avons enrichi cette base avec la connaissance actuelle de la différenciation de la CSH par l'ajout de composants et d'interactions nouvellement caractérisées, notamment en ce qui concerne le rôle d'Ikzf1 pour l'amorçage lymphoïde (MALINGE et al., 2013; NG et al., 2009; RAO et al., 2013).

De précédentes approches d'inférence de BN à partir de données scRNA-seq ont été proposées (HAMEY et al., 2017; MOIGNARD et al., 2015). Ces approches s'appuient uniquement sur les données, ainsi leurs résultats peuvent être biaisés par l'imprécision des valeurs de pseudotemps et la discréétisation des données qu'elles requièrent. Dans son ensemble, notre approche originale de synthèse de BN permet quant à elle de tirer pleinement partie des données scRNA-seq tout en tenant compte de leurs incertitudes. Elle incorpore également la connaissance riche disponible dans la littérature et les bases de données pour compenser les "zones d'ombres" du scRNA-seq (WATCHAM et al., 2019).

La synthèse de notre BN a été ainsi assistée par nos données scRNA-seq à plusieurs niveaux, d'une part en inférant de nouvelles régulations transcriptionnelles possibles entre les composants sélectionnés à l'aide d'arbres de régression (ajoutées aux régulations de la littérature), et d'autre part en définissant les contraintes dynamiques ((non) accessibilité, états stables) entre les meta-configurations définies sur nos observations d'états clés du processus dans les données. Nous avons ensuite mis en place une stratégie d'inférence de BN à l'aide de Bonesis pour aboutir à la sélection d'un BN de l'hématopoïèse précoce le plus en accord possible avec la connaissance préalable de comportements de mutants et de régulations caractérisées expérimentalement.

L'étude de la dynamique de notre BN solution dans la sémantique MP nous a permis de retracer une succession d'événements conduisant à l'amorçage hiérarchisé vers les différents

lignages. L'activation d'*Ikzf1* par Gata2 stabilise précocement l'amorçage lymphoïde de la CSH. D'un état engagé avec *Spi1* et *Myc* actif et Gata2 inactif la CSH peut s'amorcer pour le lignage neutrophile/mastocytaire ou bien l'activation de *Fli1* régule le choix d'amorçage vers le lignage érythroïde ou mégacaryocytaire.

Notre modèle autorise également des trajectoires d'amorçage alternatives pour les différents lignages qui n'empruntent pas l'état engagé preDiff. Ceci est possible par exemple avec une activation précoce de *Gata1* par Gata2 conduisant directement à l'état amorcé mégacaryocyte ou érythrocyte, ce qui s'avère être en accord avec de précédentes études de traçage de lignages soulignant la coexistence de hiérarchies hématopoïétiques multiples (RODRIGUEZ-FRATICELLI et CAMARGO, 2021; WEINREB et al., 2020). Cependant, nous suggérons qu'en condition normale la CSH s'amorce majoritairement en suivant la trajectoire observée dans nos données scRNA-seq sur laquelle ont été définies les contraintes dynamiques satisfaites par notre modèle.

Notre modèle reproduit correctement le comportement observé *in vivo/in vitro* de mutants pour une majorité des TF du réseau. Ce n'est pas le cas pour le KO de *Myc* pour lequel nous n'avons observé aucune différence en matière d'accessibilité des états amorcés *in silico* alors qu'une augmentation des CSH et une diminution de sa différenciation ont été rapportées expérimentalement avec cette mutation (A. WILSON et al., 2004). Selon cette étude, ceci est dû à des interactions intercellulaires non prises en compte dans notre modèle. Nous n'avons pas non plus observé de changements dans la dynamique pour le mutant *Egr1* KO bien que, de la même manière, une étude précédente ait montré une diminution de l'amorçage des CSH en même temps qu'une augmentation des CSH (MIN et al., 2008). Ces observations pourraient correspondre à une accumulation transitoire avant un amorçage retardé que notre formalisme de modélisation ne peut capturer.

En condition perturbée par le vieillissement notre étude suggère une altérations de l'accessibilité des états amorcés. Nos perturbations du modèle selon les altérations d'activité de TF (sur-activation de *Egr1* et *Junb*) et de régulations (perte de l'activation de *Cebpa* par Gata2) observé avec le vieillissement dans nos données scRNA-seq, aboutissent dans leur ensemble à la perte de l'amorçage vers tous les lignages à l'exception du lignage mégacaryocytaire. Ces observations sont en accord avec nos données scRNA-seq qui montrent une diminution de la proportion des cellules âgées dans tous les clusters d'états amorcés mis à part dans le cluster de cellule amorcé pour la mégacaryopoïèse (section résultat 3.3). Notre modèle est donc en mesure de décrire en partie les mécanismes de l'immunosénescence ayant lieu lors de l'hématopoïèse précoce avec une perte du potentiel lymphoïde et un biais myéloïde (DYKSTRA et al., 2011).

Nos résultats mettent en avant le rôle des facteurs Egr1 et Junb suractivés dans les CSH quiescentes biaisées myéloïdes qui s'accumulent avec le vieillissement (section 3.3 et KIRSCHNER et al., 2017) et qui ont auparavant été identifiés comme des facteurs de quiescence de la CSH (MIN et al., 2008; SANTAGUIDA et al., 2009). Notre modèle montre que ces altérations impactent un circuit positif entre Egr1 et Junb qui pourrait être nécessaire à la multiplicité des amorcages vers les différents points fixes du modèle. De façon intéressante, le réseau transcriptionnel global inféré avec SCENIC montre que ces deux facteurs sont activés par les facteurs Klf2-4-6 connus pour être en aval de la signalisation TGF-beta dans d'autres contextes biologiques (BOTELLA et al., 2009; HE et al., 2015; YAN et al., 2018). Nous avons vu dans le résultat précédent (section 3.3) que les cellules de l'état qHSC présentent une forte signature TGF-beta à laquelle se rajoute avec le vieillissement un biais myéloïde (sur-activation de *Cebpe-b* notamment). Par ailleurs, il est connu que les mégacaryocytes peuvent promouvoir la quiescence des CSH en produisant du TGF-beta (GONG et al., 2018; ZHAO et al., 2014). Nos résultats nous permettent ainsi de proposer une boucle d'auto-activation du vieillissement de la CSH avec une différenciation vers les mégacaryocytes qui promouvrait un état quiescent biaisé myéloïde de la CSH à partir duquel un amorçage unique vers le lignage mégacaryocytaire serait occasionnellement possible. En accord avec cette hypothèse, une différenciation directe de la CSH en mégacaryocytes a été montrée dans une étude de traçage de lignages (RODRIGUEZ-FRATICELLI et al., 2018). Cette trajectoire serait ainsi la mieux conservée par la CSH agée quiescente biaisée myéloïde.

En parallèle, la perte de l'activation de *Cebpa* par Gata2 participe aussi à la perte des amorçage lymphoïde et neutrophile/mastocyte. Cette altération pourrait avoir une origine épigénétique étant donné l'impact très fort du vieillissement sur les marques d'histones régulant l'ouverture de la chromatine aux éléments régulateurs des gènes de différenciation de la CSH (D. SUN et al., 2014). Elle pourrait également avoir pour origine l'émergence avec l'âge d'un clone avec une mutation au niveau d'un élément régulateur de *Cebpa* reconnu par Gata2.

Si notre modèle décrit la différenciation d'une CSH, il pourrait dans de nouveaux travaux être utilisé pour modéliser une population de CSH à l'aide de simulations stochastiques (STOLL et al., 2017). Ceci pourrait permettre de décrire encore plus finement les observations du scRNA-seq en proposant en sortie des proportions d'états stables probables qui correspondent aux proportions de cellules que l'on retrouve dans les clusters amorcés des données scRNA-seq.

# 5 Conclusion et perspectives

## Sommaire

5.1	Cartographie du compartiment HSPC en condition jeune et âgée . . . . .	160
5.2	Modélisation logique en sémantique MP de l'hématopoïèse précoce . . . . .	161
5.3	Aspects méthodologiques de la construction de BN à partir de données cellules uniques . . . . .	163

## 5.1 Cartographie du compartiment HSPC en condition jeune et âgée

Nos travaux proposent une nouvelle cartographie du compartiment HSPC murin au niveau transcriptomique en conditions jeune et âgée (section 3.3). 15 sous populations de HSPC ont pu être identifiées en combinant des approches de classifications supervisées et non supervisées. Ces sous populations se distinguent les unes des autres par leurs caractéristiques d'amorçage vers les différents lignages lymphocytaire, neutrophile, mastocytaire, érythroïde et mégacaryocytaire. Ces sous populations diffèrent également selon leur état d'avancement dans le cycle cellulaire. Notre construction d'une pseudo-trajectoire de différenciation montre effectivement que les cellules les plus immatures non amorcées sont majoritairement quiescentes tandis que l'engagement dans la différenciation s'accompagne d'une hausse de la prolifération.

En condition âgée, la répartitions des cellules dans les sous populations de HSPC et dans la pseudo-trajectoire de différenciation est altérée. Nous observons une diminution de l'amorçage vers les différents lignages à l'exception de l'amorçage mégacaryocytaire tandis que la proportion de cellules non amorcées immatures augmente. Nous avons pu mettre plus précisément en évidence une accumulation de CSH quiescentes biaisées myéloïdes avec une forte signature TGF-beta, localisée dans la pseudo-trajectoire juste avant l'amorçage des cellules vers les différents lignages. Ces résultats montrent clairement un lien entre l'amorçage perturbé et la régulation du cycle cellulaire de la CSH âgée. Précédemment, une étude scRNA-seq de HSPC murins suggère une diminution de la longueur de la phase G1 responsable d'une augmentation de l'autorenouvellement pour expliquer l'accumulation de CSH dans la MO (KOWALCZYK et al., 2015). Notre étude des variations avec l'âge des proportions

des différentes sous populations du compartiment HSPC pointe plutôt vers une capacité d'autorenouvellement inchangé mais un blocage du départ en différenciation des CSH qui s'accumuleraient dans la MO dans un état quiescent avec des biais myéloïde et plaquettaire dans la [MO](#) des animaux.

Dans son ensemble notre étude scRNA-seq haut débit retrouve et précise les précédentes sous populations particulières, identifiées dans des études scRNA-seq bas débits du vieillissement de CSH et MPP triées séparément (GROVER et al., 2016; KIRSCHNER et al., 2017; KOWALCZYK et al., 2015; MANN et al., 2018; YOUNG et al., 2016). Nous avons également identifié une nouvelle sous population de cellules non amorcée présentant une signature de réponse à l'interféron qui augmente avec le vieillissement. Dans le but de caractériser fonctionnellement ces sous populations, un nouveau panel d'anticorps est développé au sein de l'équipe du CRCM pour pouvoir isoler et identifier ces différentes cellules par [FACS](#).

Notre étude a mis en avant des disparités entre les 2 lots d'échantillons. Nous avons notamment observé que la diminution de l'amorçage lymphoïde et l'augmentation des CSH quiescentes avec une signature TGF-beta étaient principalement retrouvées pour un des lots. De notre point de vue, ceci illustre l'hétérogénéité du vieillissement entre les groupes de souris qui est sans doute également à l'origine des différences entre les études scRNA-seq du vieillissement des HSPC déjà relevées entre les études RNA-seq en "bulk" de ce processus (SVENDSEN et al., 2021). Comme pour les études de RNA-seq en "bulk" une analyse globale de ces études serait pertinente dans le but de mieux appréhender cette hétérogénéité en distinguant les différences résultant de biais techniques ou expérimentaux de celles pouvant s'expliquer par l'émergence de clones différents au cours de la vie d'un individu à l'autre. Cette étude nécessiterait une intégration de l'ensemble de ces jeux de données et pourrait s'accompagner de l'ajout de données aux niveaux épigénétique et génomique grâce au développement récents des technologies multi-omiques cellules uniques (voir section 1.2.5.2 en introduction). Les jeux de données scRNA-seq publiés présentent une condition âgée à des temps différents (de 18 à 22 mois). Avec l'ajout d'un temps intermédiaire (souris de 12 mois par exemple) ce travail pourrait aboutir à retracer l'historique d'une hématopoïèse clonale aboutissant à l'accumulation des CSH âgées non fonctionnelles.

## 5.2 Modélisation logique en sémantique MP de l'hématopoïèse précoce

Notre seconde étude propose un modèle logique de l'amorçage de la CSH vers les lignages lymphoïde, neutrophile/mastocytaire, érythroïde et mégacaryocytaire qui décrit dans la sémantique [MP](#) notre pseudo-trajectoire de différenciation et son altération avec le vieillissement (section 4.3). Pour construire ce réseau booléen de 13 TF et deux complexes régulant la

quiescence de la CSH nous avons tenu compte des connaissances actuelles de l'hématopoïèse précoce plutôt que d'essayer d'inférer un BN uniquement à partir des données scRNA-seq comme cela a pu être fait auparavant (HAMEY et al., 2017). Bien que l'analyse scRNA-seq apporte une information riche du processus étudié, les données présentent encore selon nous des incertitudes trop importantes pour une stratégie d'inférence directe. En divisant la construction du modèle avec d'abord la construction d'un graphe d'inférence puis sa paramétrisation nous avons pu combiner efficacement la connaissance actuelle et l'apport des données scRNA-seq avec l'utilisation de SCENIC et Bonesis.

La solution retenue est valide en sémantique MP qui autorise beaucoup plus de transitions que les sémantiques asynchrone et généralisée classiquement utilisées. Une comparaison des trajectoires asynchrones et MP de notre modèle serait intéressante dans le but d'identifier l'ensemble des raffinements multivalués nécessaires pour reproduire en sémantique asynchrone les trajectoires MP. Nous avons réalisé ce travail pour expliquer le choix de différenciation vers les points fixes pEr et pMk depuis l'état preDiff qui repose sur l'existence de deux seuils d'influence de Fli1 sur ces cibles *Klf1* et *Gata1*.

Actuellement il n'existe pas à ma connaissance d'outils pour l'analyse globale des trajectoires d'un BN en sémantique MP. Le graphe de transition de notre BN est d'une taille trop importante pour être analysé tel quel ( $n=2^{15}$ ) et les relations entre configurations de (non) accessibilité entre deux configurations ont été analysées une par une dans notre étude. La méthode de construction de graphes de transitions hiérarchiques développée pour la sémantique asynchrone pourrait, je pense, être adaptée pour la sémantique MP bien que du fait du nombre de transitions plus important son temps de calcul sera plus long (BÉRENGUIER et al., 2013). Ceci permettrait de caractériser les trajectoires alternatives de la différenciation des CSH qui n'empruntent pas les chemins que nous avons contraints pour l'inférence.

L'altération du modèle selon nos observations du vieillissement nous a permis de reproduire à l'échelle d'une cellule la perte de la capacité d'amorçage vers l'ensemble des lignages à l'exception de l'amorçage mégacaryocytaire. Grâce au modèle nous sommes en mesure de proposer deux mécanismes moléculaires spécifiques qui pourraient être à l'origine de cette altération. Les facteurs Junb et Egr1 sont d'après nos observations les médiateurs de signaux extrinsèques qui favorisent un état quiescent de la CSH âgée duquel l'amorçage mégacaryocytaire est la seule différenciation possible. Nous suggérons que ces signaux extrinsèques proviennent de la signalisation TGF-beta ce qui constitue une boucle d'auto activation du vieillissement de la CSH dans laquelle les mégacaryocytes produisent du TGF-beta qui favorise la quiescence des CSH et empêche leur différenciation vers les autres lignées. Cette hypothèse pourrait être testée expérimentalement en réduisant l'expression (Knock-down) de *Junb* et/ou *Egr1* dans les CSH âgée.

La perte de l'activation de *Cebpa* par Gata2 pourrait également être impliquée dans l'altération de l'amorçage de la CSH avec le vieillissement selon nos analyses. Nous envisageons avec l'équipe du CRCM de tester cette hypothèse par une analyse ChIP-seq du facteur Gata2 chez les CSH jeunes et âgées pour essayer d'identifier une diminution de la fixation de Gata2 aux niveaux des éléments régulateurs de *Cebpa* avec le vieillissement.

## 5.3 Aspects méthodologiques de la construction de BN à partir de données cellules uniques

Notre travail a permis de montrer la faisabilité de la construction de BN à partir de données cellules uniques en combinant des arbres de régression pour la construction du graphe d'influence et l'inférence des règles logiques avec la programmation de contraintes dynamiques. Cette approche pourrait être améliorée en plusieurs points par l'ajout de données multi-omiques cellules uniques et l'amélioration des outils utilisés.

Pour la construction du graphe d'influence, des données scATAC-seq permettraient de définir les régions ouvertes de la chromatine dans les cellules. Une première étape consisterait alors à identifier les motifs de TF aux éléments régulateurs sur les régions ouvertes de l'ADN des cellules étudiées. Pour chaque gène considéré, l'ensemble des régulateurs possibles serait alors défini avant l'étape de régression ce qui aboutirait à l'inférence d'un réseau transcriptionnel plus précis (et contextualisé) et réduirait le temps de calcul par rapport à SCENIC qui effectue un filtrage une fois le réseau inféré à partir d'une banque de motifs.

Dans notre étude, nous avons adapté en nous justifiant la discréétisation de certains composants pour obtenir des solutions avec Bonesis. Plutôt que de faire ce travail de façon empirique je pense qu'il serait possible de chercher en ASP la matrice des méta-configurations (en entrée de Bonesis) la plus proche de la matrice des données discréétisées qui permet d'obtenir des solutions.

Les approches de construction de BN à partir de données scRNA-seq permettent d'obtenir un BN qui décrit le comportement d'une cellule. Comme évoqué précédemment des simulations stochastiques peuvent être utilisées pour modéliser le comportement de la population de cellules analysées avec les technologies cellules uniques (STOLL et al., 2012). Ces simulations se font par la paramétrisation de taux d'activation/d'inhibition de chacun des composants du BN. Il conviendrait donc de développer des méthodes pour adapter cette paramétrisation vis à vis des observations sur les données cellules uniques. Cependant, une telle modélisation serait encore imparfaite étant donné qu'au cours de processus de différenciation

ciation biologiques la taille de la population varie en fonction des états de prolifération ou d'apoptose à différents temps du processus. De nouvelles approches de modélisation logique à l'échelle de la population tenant compte de ces variations pourraient ainsi être développées pour améliorer notre compréhension de l'hématopoïèse et des systèmes biologiques en général.

# Bibliographie

- ACHA, P., PALOMO, L., FUSTER-TORMO, F., XICOY, B., MALLO, M., MANZANARES, A., GRAU, J., MARCÉ, S., GRANADA, I., RODRÍGUEZ-LUACES, M., DIEZ-CAMPELO, M., ZAMORA, L. & SOLÉ, F. (2021). Analysis of Intratumoral Heterogeneity in Myelodysplastic Syndromes with Isolated del(5q) Using a Single Cell Approach [Number : 4 Publisher : Multidisciplinary Digital Publishing Institute]. *Cancers*, 13(4), 841. <https://doi.org/10.3390/cancers13040841> (cf. p. 50)
- ADOLFSSON, J., MÄNSSON, R., BUZA-VIDAS, N., HULTQUIST, A., LIUBA, K., JENSEN, C. T., BRYDER, D., YANG, L., BORGE, O.-J., THOREN, L. A. M., ANDERSON, K., SITNICKA, E., SASAKI, Y., SIGVARDSSON, M. & JACOBSEN, S. E. W. (2005). Identification of Flt3+ Lympho-Myeloid Stem Cells Lacking Erythro-Megakaryocytic Potential : A Revised Road Map for Adult Blood Lineage Commitment [Number : 2]. *Cell*, 121(2), 295-306. <https://doi.org/10.1016/j.cell.2005.02.013> (cf. p. 17, 18)
- AIBAR, S., GONZÁLEZ-BLAS, C. B., MOERMAN, T., HUYNH-THU, V. A., IMRICOVA, H., HULSELMANS, G., RAMBOW, F., MARINE, J.-C., GEURTS, P., AERTS, J., van den OORD, J., ATAK, Z. K., WOUTERS, J. & AERTS, S. (2017). SCENIC : single-cell regulatory network inference and clustering [Number : 11]. *Nature Methods*, 14(11), 1083-1086. <https://doi.org/10.1038/nmeth.4463> (cf. p. 39, 65, 68, 77, 120)
- AKASHI, K., TRAVER, D., MIYAMOTO, T. & WEISSMAN, I. L. (2000). A clonogenic common myeloid progenitor that gives rise to all myeloid lineages [Number : 6774 Publisher : Nature Publishing Group]. *Nature*, 404(6774), 193-197. <https://doi.org/10.1038/35004599> (cf. p. 17)
- ALLSOPP, R. C., MORIN, G. B., DEPINHO, R., HARLEY, C. B. & WEISSMAN, I. L. (2003). Telomerase is required to slow telomere shortening and extend replicative lifespan of HSCs during serial transplantation [Number : 2]. *Blood*, 102(2), 517-520. <https://doi.org/10.1182/blood-2002-07-2334> (cf. p. 28)
- ARINO, O. & KIMMEL, M. (1986). Stability analysis of models of cell production systems. *Mathematical Modelling*, 7(9-12), 1269-1300. [https://doi.org/10.1016/0270-0255\(86\)90081-3](https://doi.org/10.1016/0270-0255(86)90081-3) (cf. p. 62)
- AUBIN-FRANKOWSKI, P.-C. & VERT, J.-P. (2020). Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. *Bioinformatics*, 36(18), 4774-4780. <https://doi.org/10.1093/bioinformatics/btaa576> (cf. p. 64)
- BARILE, M., BUSCH, K., FANTI, A.-K., GRECO, A., WANG, X., OGURO, H., ZHANG, Q., MORRISON, S. J., RODEWALD, H.-R. & HÖFER, T. (2020). Hematopoietic stem cells self-renew symmetrically or gradually proceed to differentiation. *bioRxiv*, 2020.08.06.239186. <https://doi.org/10.1101/2020.08.06.239186> (cf. p. 25)
- BARILLOT, E., CALZONE, L., HUPE, P., VERT, J.-P. & ZINOVYEV, A. (2013). *Computational systems biology of cancer*. CRC Press Boca Raton, FL. (Cf. p. 50, 51).
- BARYAWNO, N., SEVERE, N. & SCADDEN, D. T. (2017). Hematopoiesis : Reconciling Historic Controversies about the Niche [Number : 5]. *Cell Stem Cell*, 20(5), 590-592. <https://doi.org/10.1016/j.stem.2017.03.025> (cf. p. 20)

- BECHT, E., MCINNES, L., HEALY, J., DUTERTRE, C.-A., KWOK, I. W. H., NG, L. G., GINHOUX, F. & NEWELL, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP [Number : 1 Publisher : Nature Publishing Group]. *Nature Biotechnology*, 37(1), 38-44. <https://doi.org/10.1038/nbt.4314> (cf. p. 35)
- BECKER, A. J., MCCULLOCH, E. A. & TILL, J. E. (1963). Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature*, 197, 452-454. <https://doi.org/10.1038/197452a0> (cf. p. 17)
- BEERMAN, I., BOCK, C., GARRISON, B. S., SMITH, Z. D., GU, H., MEISSNER, A. & ROSSI, D. J. (2013). Proliferation-dependent alterations of the DNA methylation landscape underlie hematopoietic stem cell aging [Number : 4]. *Cell Stem Cell*, 12(4), 413-425. <https://doi.org/10.1016/j.stem.2013.01.017> (cf. p. 31)
- BÉRENGUIER, D., CHAOUIYA, C., MONTEIRO, P. T., NALDI, A., REMY, E., THIEFFRY, D. & TICHIT, L. (2013). Dynamical modeling and analysis of large cellular regulatory networks [Publisher : American Institute of Physics]. *Chaos : An Interdisciplinary Journal of Nonlinear Science*, 23(2), 025114. <https://doi.org/10.1063/1.4809783> (cf. p. 162)
- BERNITZ, J. M., KIM, H. S., MACARTHUR, B., SIEBURG, H. & MOORE, K. (2016). Hematopoietic Stem Cells Count and Remember Self-Renewal Divisions [Number : 5]. *Cell*, 167(5), 1296-1309.e10. <https://doi.org/10.1016/j.cell.2016.10.022> (cf. p. 20, 25)
- BJÖRN, N., JAKOBSEN, I., LOTFI, K. & GRÉEN, H. (2020). Single-Cell RNA Sequencing of Hematopoietic Stem and Progenitor Cells Treated with Gemcitabine and Carboplatin [Number : 5]. *Genes*, 11(5). <https://doi.org/10.3390/genes11050549> (cf. p. 50)
- BONZANNI, N., GARG, A., FEENSTRA, K. A., SCHÜTTE, J., KINSTON, S., MIRANDA-SAAVEDRA, D., HERINGA, J., XENARIOS, I. & GÖTTGENS, B. (2013). Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model [Number : 13]. *Bioinformatics*, 29(13), i80-i88. <https://doi.org/10.1093/bioinformatics/btt243> (cf. p. 22, 63, 119, 157)
- BOTELLA, L. M., SANZ-RODRIGUEZ, F., KOMI, Y., FERNANDEZ-L, A., VARELA, E., GARRIDO-MARTIN, E. M., NARLA, G., FRIEDMAN, S. L. & KOJIMA, S. (2009). TGF-beta regulates the expression of transcription factor KLF6 and its splice variants and promotes co-operative transactivation of common target genes through a Smad3-Sp1-KLF6 interaction [Number : 2]. *The Biochemical Journal*, 419(2), 485-495. <https://doi.org/10.1042/BJ20081434> (cf. p. 159)
- BOWIE, M. B., MCKNIGHT, K. D., KENT, D. G., MCCAFFREY, L., HOODLESS, P. A. & EAVES, C. J. (2006). Hematopoietic stem cells proliferate until after birth and show a reversible phase-specific engraftment defect [Number : 10]. *The Journal of Clinical Investigation*, 116(10), 2808-2816. <https://doi.org/10.1172/JCI28310> (cf. p. 22)
- BOWMAN, T. V., MCCOOEY, A. J., MERCHANT, A. A., RAMOS, C. A., FONSECA, P., POINDEXTER, A., BRADFUTE, S. B., OLIVEIRA, D. M., GREEN, R., ZHENG, Y., JACKSON, K. A., CHAMBERS, S. M., MCKINNEY-FREEMAN, S. L., NORWOOD, K. G., DARLINGTON, G., GUNARATNE, P. H., STEFFEN, D. & GOODELL, M. A. (2006). Differential mRNA Processing in Hematopoietic Stem Cells. *STEM CELLS*, 24(3), 662-670. <https://doi.org/https://doi.org/10.1634/stemcells.2005-0552> (cf. p. 41)
- BUENROSTRO, J. D., CORCES, M. R., LAREAU, C. A., WU, B., SCHEP, A. N., ARYEE, M. J., MAJETI, R., CHANG, H. Y. & GREENLEAF, W. J. (2018). Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation [Number : 6]. *Cell*, 173(6), 1535-1548.e16. <https://doi.org/10.1016/j.cell.2018.03.074> (cf. p. 42, 46, 48)

- BUETTNER, F., NATARAJAN, K. N., CASALE, F. P., PROSERPIO, V., SCIALDONE, A., THEIS, F. J., TEICHMANN, S. A., MARIONI, J. C. & STEGLE, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells [Number : 2]. *Nature Biotechnology*, 33(2), 155-160. <https://doi.org/10.1038/nbt.3102> (cf. p. 40)
- BUTLER, A., HOFFMAN, P., SMIBERT, P., PAPALEXI, E. & SATIJA, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species [Number : 5]. *Nature Biotechnology*, 36(5), 411-420. <https://doi.org/10.1038/nbt.4096> (cf. p. 33, 35, 40, 41)
- BÜTTNER, M., MIAO, Z., WOLF, F. A., TEICHMANN, S. A. & THEIS, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction [Number : 1 Publisher : Nature Publishing Group]. *Nature Methods*, 16(1), 43-49. <https://doi.org/10.1038/s41592-018-0254-1> (cf. p. 40)
- CABEZAS-WALLSCHEID, N., BUETTNER, F., SOMMERKAMP, P., KLIMMECK, D., LADEL, L., THALHEIMER, F. B., PASTOR-FLORES, D., ROMA, L. P., RENDERS, S., ZEISBERGER, P., PRZYBYLLA, A., SCHÖNBERGER, K., SCOGNAMIGLIO, R., ALTAMURA, S., FLORIAN, C. M., FAWAZ, M., VONFICHT, D., TESIO, M., COLLIER, P., ... TRUMPP, A. (2017). Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy [Number : 5]. *Cell*, 169(5), 807-823.e19. <https://doi.org/10.1016/j.cell.2017.04.018> (cf. p. 39, 42)
- CABEZAS-WALLSCHEID, N., KLIMMECK, D., HANSSON, J., LIPKA, D. B., REYES, A., WANG, Q., WEICHENHAN, D., LIER, A., von PALESKE, L., RENDERS, S., WÜNSCHE, P., ZEISBERGER, P., BROCKS, D., GU, L., HERRMANN, C., HAAS, S., ESSERS, M. A. G., BRORS, B., EILS, R., ... TRUMPP, A. (2014). Identification of Regulatory Networks in HSCs and Their Immediate Progeny via Integrated Proteome, Transcriptome, and DNA Methylome Analysis [Number : 4]. *Cell Stem Cell*, 15(4), 507-522. <https://doi.org/10.1016/j.stem.2014.07.005> (cf. p. 21)
- CACACE, E., COLLOMBET, S. & THIEFFRY, D. (2020). Logical modeling of cell fate specification-Application to T cell commitment. *Current Topics in Developmental Biology*, 139, 205-238. <https://doi.org/10.1016/bs.ctdb.2020.02.008> (cf. p. 62)
- CAMPILLO-MARCOS, I., ALVAREZ-ERRICO, D., ALANDES, R. A., MEREU, E. & ESTELLER, M. (2021). Single-cell technologies and analyses in hematopoiesis and hematological malignancies. *Experimental Hematology*, 98, 1-13. <https://doi.org/10.1016/j.exphem.2021.05.001> (cf. p. 49)
- CATLIN, S. N., BUSQUE, L., GALE, R. E., GUTTORP, P. & ABKOWITZ, J. L. (2011). The replication rate of human hematopoietic stem cells in vivo [Number : 17]. *Blood*, 117(17), 4460-4466. <https://doi.org/10.1182/blood-2010-08-303537> (cf. p. 22)
- CHAMBERS, S. M., SHAW, C. A., GATZA, C., FISK, C. J., DONEHOWER, L. A. & GOODELL, M. A. (2007). Aging hematopoietic stem cells decline in function and exhibit epigenetic dysregulation [Number : 8]. *PLoS biology*, 5(8), e201. <https://doi.org/10.1371/journal.pbio.0050201> (cf. p. 31)
- CHAN, T. E., STUMPF, M. P. H. & BABTIE, A. C. (2017). Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Systems*, 5(3), 251-267.e3. <https://doi.org/10.1016/j.cels.2017.08.014> (cf. p. 64, 66)
- CHEN, H., ALBERGANTE, L., HSU, J. Y., LAREAU, C. A., LO BOSCO, G., GUAN, J., ZHOU, S., GORBAN, A. N., BAUER, D. E., ARYEE, M. J., LANGENAU, D. M., ZINOVYEV, A., BUENROSTRO, J. D., YUAN, G.-C. & PINELLO, L. (2019). Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM [Number : 1 Publisher : Nature Publishing

- Group]. *Nature Communications*, 10(1), 1903. <https://doi.org/10.1038/s41467-019-09670-4> (cf. p. 36, 48, 115)
- CHEVALIER, S., FROIDEVAUX, C., PAULEVÉ, L. & ZINOVYEV, A. (2019). Synthesis of Boolean Networks from Biological Dynamical Constraints using Answer-Set Programming [ISSN : 2375-0197]. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 34-41. <https://doi.org/10.1109/ICTAI.2019.00014> (cf. p. 65, 69, 71, 75, 120)
- CHEVALIER, S., NOËL, V., CALZONE, L., ZINOVYEV, A. & PAULEVÉ, L. (2020). Synthesis and Simulation of Ensembles of Boolean Networks for Cell Fate Decision. *18th International Conference on Computational Methods in Systems Biology (CMSB)*, 12314, 193-209. [https://doi.org/10.1007/978-3-030-60327-4\\_11](https://doi.org/10.1007/978-3-030-60327-4_11) (cf. p. 60, 120-122)
- CHOI, Y. J. & ANDERS, L. (2014). Signaling through cyclin D-dependent kinases [Number : 15]. *Oncogene*, 33(15), 1890-1903. <https://doi.org/10.1038/onc.2013.137> (cf. p. 23)
- CHOUDHURY, A. R., JU, Z., DJOJOSUBROTO, M. W., SCHIENKE, A., LECHEL, A., SCHAETZLEIN, S., JIANG, H., STEP CZYNSKA, A., WANG, C., BUER, J., LEE, H.-W., von ZGLINICKI, T., GANSER, A., SCHIRMACHER, P., NAKAUCHI, H. & RUDOLPH, K. L. (2007). Cdkn1a deletion improves stem cell function and lifespan of mice with dysfunctional telomeres without accelerating cancer formation [Number : 1]. *Nature Genetics*, 39(1), 99-105. <https://doi.org/10.1038/ng1937> (cf. p. 30, 31)
- COLLOMBET, S., OVELEN, C. v., ORTEGA, J. L. S., ABOU-JAOUDÉ, W., STEFANO, B. D., THOMAS-CHOLIER, M., GRAF, T. & THIEFFRY, D. (2017). Logical modeling of lymphoid and myeloid cell specification and transdifferentiation [Publisher : National Academy of Sciences Section : Colloquium Paper]. *Proceedings of the National Academy of Sciences*, 114(23), 5792-5799. <https://doi.org/10.1073/pnas.1610622114> (cf. p. 62, 63, 119, 157)
- COVER, T. M. & THOMAS, J. A. (1991). *Elements of information theory*. Wiley. (Cf. p. 66).
- DE HAAN, G. & GERRITS, A. (2007). Epigenetic control of hematopoietic stem cell aging the case of Ezh2. *Annals of the New York Academy of Sciences*, 1106, 233-239. <https://doi.org/10.1196/annals.1392.008> (cf. p. 30)
- de MOURA, L. & BJØRNER, N. (2008). Z3 : An Efficient SMT Solver. In C. R. RAMAKRISHNAN & J. REHOF (Éd.), *Tools and Algorithms for the Construction and Analysis of Systems* (p. 337-340). Springer. [https://doi.org/10.1007/978-3-540-78800-3\\_24](https://doi.org/10.1007/978-3-540-78800-3_24). (Cf. p. 69)
- DESTERKE, C., PETIT, L., SELLA, N., CHEVALLIER, N., CABELI, V., COQUELIN, L., DURAND, C., OOSTENDORP, R. A. J., ISAMBERT, H., JAFFREDO, T. & CHARBORD, P. (2020). Inferring Gene Networks in Bone Marrow Hematopoietic Stem Cell-Supporting Stromal Niche Populations [Publisher : Elsevier]. *iScience*, 23(6). <https://doi.org/10.1016/j.isci.2020.101222> (cf. p. 66)
- DOULATOV, S., NOTTA, F., EPPERT, K., NGUYEN, L. T., OHASHI, P. S. & DICK, J. E. (2010). Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development [Number : 7]. *Nature Immunology*, 11(7), 585-593. <https://doi.org/10.1038/ni.1889> (cf. p. 17)
- DUGOURD, A. & SAEZ-RODRIGUEZ, J. (2019). Footprint-based functional analysis of multiomic data. *Current Opinion in Systems Biology*, 15, 82-90. <https://doi.org/10.1016/j.coisb.2019.04.002> (cf. p. 39, 77)
- DYKSTRA, B., OLTHOF, S., SCHREUDER, J., RITSEMA, M. & de HAAN, G. (2011). Clonal analysis reveals multiple functional defects of aged murine hematopoietic stem cells [Number : 13]. *The Journal of Experimental Medicine*, 208(13), 2691-2703. <https://doi.org/10.1084/jem.20111490> (cf. p. 26, 158)

- EBERWINE, J., YEH, H., MIYASHIRO, K., CAO, Y., NAIR, S., FINNELL, R., ZETTEL, M. & COLEMAN, P. (1992). Analysis of gene expression in single live neurons [Number : 7]. *Proceedings of the National Academy of Sciences of the United States of America*, 89(7), 3010-3014. <https://doi.org/10.1073/pnas.89.7.3010> (cf. p. 32)
- ENCISO, J., MAYANI, H., MENDOZA, L. & PELAYO, R. (2016). Modeling the Pro-inflammatory Tumor Microenvironment in Acute Lymphoblastic Leukemia Predicts a Breakdown of Hematopoietic-Mesenchymal Communication Networks [Publisher : Frontiers]. *Frontiers in Physiology*, 7. <https://doi.org/10.3389/fphys.2016.00349> (cf. p. 63, 64)
- FAURÉ, A., NALDI, A., CHAOUIYA, C. & THIEFFRY, D. (2006). Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 22(14), e124-e131. <https://doi.org/10.1093/bioinformatics/btl210> (cf. p. 60)
- FLACH, J., BAKKER, S. T., MOHRIN, M., CONROY, P. C., PIETRAS, E. M., REYNAUD, D., ALVAREZ, S., DIOLAITI, M. E., UGARTE, F., FORSBERG, E. C., LE BEAU, M. M., STOHR, B. A., MÉNDEZ, J., MORRISON, C. G. & PASSEGUÉ, E. (2014). Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells [Number : 7513]. *Nature*, 512(7513), 198-202. <https://doi.org/10.1038/nature13619> (cf. p. 28, 31)
- FLOBAK, Å., BAUDOT, A., REMY, E., THOMMESEN, L., THIEFFRY, D., KUIPER, M. & LÆGREID, A. (2015). Discovery of Drug Synergies in Gastric Cancer Cells Predicted by Logical Modeling [Number : 8]. *PLOS Computational Biology*, 11(8), e1004426. <https://doi.org/10.1371/journal.pcbi.1004426> (cf. p. 62)
- FLORIAN, M. C., DÖRR, K., NIEBEL, A., DARIA, D., SCHREZENMEIER, H., ROJEWSKI, M., FILIPPI, M.-D., HASENBERG, A., GUNZER, M., SCHARFFETTER-KOCHANEK, K., ZHENG, Y. & GEIGER, H. (2012). Cdc42 Activity Regulates Hematopoietic Stem Cell Aging and Rejuvenation [Number : 5]. *Cell Stem Cell*, 10(5), 520-530. <https://doi.org/10.1016/j.stem.2012.04.007> (cf. p. 24, 26, 29)
- FLORIAN, M. C., KLOSE, M., SACMA, M., JABLANOVIĆ, J., KNUDSON, L., NATTAMAI, K. J., MARKA, G., VOLLMER, A., SOLLER, K., SAKK, V., CABEZAS-WALLSCHEID, N., ZHENG, Y., MULAW, M. A., GLAUCHE, I. & GEIGER, H. (2018). Aging alters the epigenetic asymmetry of HSC division. *PLoS biology*, 16(9), e2003389. <https://doi.org/10.1371/journal.pbio.2003389> (cf. p. 29, 48)
- FORNES, O., CASTRO-MONDRAGON, J. A., KHAN, A., van der LEE, R., ZHANG, X., RICHMOND, P. A., MODI, B. P., CORREARD, S., GHEORGHE, M., BARANAŠIĆ, D., SANTANA-GARCIA, W., TAN, G., CHÈNEBY, J., BALLESTER, B., PARCY, F., SANDELIN, A., LENHARD, B., WASSERMAN, W. W. & MATHELIER, A. (2020). JASPAR 2020 : update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 48(D1), D87-D92. <https://doi.org/10.1093/nar/gkz1001> (cf. p. 51)
- FRANCESCHI, C., BONAFÈ, M., VALENSIN, S., OLIVIERI, F., DE LUCA, M., OTTAVIANI, E. & DE BENEDICTIS, G. (2000). Inflamm-aging. An evolutionary perspective on immunosenescence. *Annals of the New York Academy of Sciences*, 908, 244-254. <https://doi.org/10.1111/j.1749-6632.2000.tb06651.x> (cf. p. 26)
- GEBSER, M., KAMINSKI, R., KAUFMANN, B. & SCHAUB, T. (2012). Answer Set Solving in Practice [Publisher : Morgan & Claypool Publishers]. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(3), 1-238. <https://doi.org/10.2200/S00457ED1V01Y201211AIM019> (cf. p. 69, 122)
- GEIGER, H., HAAN, G. d. & FLORIAN, M. C. (2013). The ageing haematopoietic stem cell compartment [Number : 5]. *Nature Reviews Immunology*, 13(5), 376-389. <https://doi.org/10.1038/nri3433> (cf. p. 26-30)

- GILADI, A., PAUL, F., HERZOG, Y., LUBLING, Y., WEINER, A., YOFE, I., JAITIN, D., CABEZAS-WALLSCHEID, N., DRESS, R., GINHOUX, F., TRUMPP, A., TANAY, A. & AMIT, I. (2018). Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis [Number : 7 Publisher : Nature Publishing Group]. *Nature Cell Biology*, 20(7), 836-846. <https://doi.org/10.1038/s41556-018-0121-4> (cf. p. 42, 49)
- GONG, Y., ZHAO, M., YANG, W., GAO, A., YIN, X., HU, L., WANG, X., XU, J., HAO, S., CHENG, T. & CHENG, H. (2018). Megakaryocyte-derived excessive transforming growth factor 1 inhibits proliferation of normal hematopoietic stem cells in acute myeloid leukemia. *Experimental Hematology*, 60, 40-46.e2. <https://doi.org/10.1016/j.exphem.2017.12.010> (cf. p. 23, 159)
- GRIECO, L., CALZONE, L., BERNARD-PIERROT, I., RADVANYI, F., KAHN-PERLÈS, B. & THIEFFRY, D. (2013). Integrative Modelling of the Influence of MAPK Network on Cancer Cell Fate Decision [Publisher : Public Library of Science]. *PLOS Computational Biology*, 9(10), e1003286. <https://doi.org/10.1371/journal.pcbi.1003286> (cf. p. 60, 61)
- GROVER, A., SANJUAN-PLA, A., THONGJUEA, S., CARRELHA, J., GIUSTACCHINI, A., GAMBARDELLA, A., MACAULAY, I., MANCINI, E., LUIS, T. C., MEAD, A., JACOBSEN, S. E. W. & NERLOV, C. (2016). Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. *Nature Communications*, 7, 11075. <https://doi.org/10.1038/ncomms11075> (cf. p. 31, 49, 74, 76, 117, 161)
- GUO, G., HUSS, M., TONG, G. Q., WANG, C., LI SUN, L., CLARKE, N. D. & ROBSON, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst [Number : 4]. *Developmental Cell*, 18(4), 675-685. <https://doi.org/10.1016/j.devcel.2010.02.012> (cf. p. 32)
- HAECKEL, E. (1877). *Anthropogénie; ou, Histoire de l'évolution humaine, leçons familiaires sur les principes de l'embryologie et de la phylogénie humaines*. Reinwald. (Cf. p. 17).
- HAGHVERDI, L., LUN, A. T. L., MORGAN, M. D. & MARIONI, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors [Number : 5]. *Nature Biotechnology*, 36(5), 421-427. <https://doi.org/10.1038/nbt.4091> (cf. p. 41)
- HAMEY, F. K. & GÖTTGENS, B. (2019). Machine learning predicts putative hematopoietic stem cells within large single-cell transcriptomics data sets. *Experimental Hematology*, 78, 11-20. <https://doi.org/10.1016/j.exphem.2019.08.009> (cf. p. 36, 120)
- HAMEY, F. K., NESTOROWA, S., KINSTON, S. J., KENT, D. G., WILSON, N. K. & GÖTTGENS, B. (2017). Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proceedings of the National Academy of Sciences*, 114(23), 5822-5829 (cf. p. 64, 69, 71, 157, 162).
- HAN, H., CHO, J.-W., LEE, S., YUN, A., KIM, H., BAE, D., YANG, S., KIM, C. Y., LEE, M., KIM, E., LEE, S., KANG, B., JEONG, D., KIM, Y., JEON, H.-N., JUNG, H., NAM, S., CHUNG, M., KIM, J.-H. & LEE, I. (2018). TRRUST v2 : an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46(D1), D380-D386. <https://doi.org/10.1093/nar/gkx1013> (cf. p. 51)
- HAN, X., WANG, R., ZHOU, Y., FEI, L., SUN, H., LAI, S., SAADATPOUR, A., ZHOU, Z., CHEN, H., YE, F., HUANG, D., XU, Y., HUANG, W., JIANG, M., JIANG, X., MAO, J., CHEN, Y., LU, C., XIE, J., ... GUO, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq [Number : 5]. *Cell*, 172(5), 1091-1107.e17. <https://doi.org/10.1016/j.cell.2018.02.001> (cf. p. 41)
- HE, M., ZHENG, B., ZHANG, Y., ZHANG, X.-H., WANG, C., YANG, Z., SUN, Y., WU, X.-L. & WEN, J.-K. (2015). KLF4 mediates the link between TGF-1-induced gene transcription and

- H3 acetylation in vascular smooth muscle cells. *The FASEB Journal*, 29(9), 4059-4070. <https://doi.org/10.1096/fj.15-272658> (cf. p. 159)
- HENRY, C. J., MARUSYK, A. & DEGREGORI, J. (2011). Aging-associated changes in hematopoiesis and leukemogenesis : what's the connection ? [Number : 6]. *Aging*, 3(6), 643-656. <https://doi.org/10.18632/aging.100351> (cf. p. 26)
- HÉRAULT, L., POPLINEAU, M., MAZUEL, A., PLATET, N., REMY, É. & DUPREZ, E. (2021). Single-cell RNA-seq reveals a concomitant delay in differentiation and cell cycle of aged hematopoietic stem cells. *BMC Biology*, 19. <https://doi.org/10.1186/s12915-021-00955-z> (cf. p. 31, 40, 42, 49, 81, 115, 157)
- HERMAN, J. S., SAGAR, n. & GRÜN, D. (2018). FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data [Number : 5]. *Nature Methods*, 15(5), 379-386. <https://doi.org/10.1038/nmeth.4662> (cf. p. 42, 45)
- HERRMANN, C., VAN DE SANDE, B., POTIER, D. & AERTS, S. (2012). i-cisTarget : an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Research*, 40(15), e114-e114. <https://doi.org/10.1093/nar/gks543> (cf. p. 68)
- HINGE, A., HE, J., BARTRAM, J., JAVIER, J., XU, J., FJELLMAN, E., SESAKI, H., LI, T., YU, J., WUNDERLICH, M., MULLOY, J., KOFRON, M., SALOMONIS, N., GRIMES, H. L. & FILIPPI, M.-D. (2020). Asymmetrically Segregated Mitochondria Provide Cellular Memory of Hematopoietic Stem Cell Replicative History and Drive HSC Attrition [Number : 3]. *Cell Stem Cell*, 26(3), 420-430.e6. <https://doi.org/10.1016/j.stem.2020.01.016> (cf. p. 25)
- HOLLAND, C. H., TANEVSKI, J., PERALES-PATÓN, J., GLEIXNER, J., KUMAR, M. P., MEREU, E., JOUGHIN, B. A., STEGLE, O., LAUFFENBURGER, D. A., HEYN, H., SZALAI, B. & SAEZ-RODRIGUEZ, J. (2020). Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data [Number : 1]. *Genome Biology*, 21(1), 36. <https://doi.org/10.1186/s13059-020-1949-z> (cf. p. 39)
- HOU, W., JI, Z., JI, H. & HICKS, S. C. (2020). A systematic evaluation of single-cell RNA-sequencing imputation methods [Number : 1]. *Genome Biology*, 21(1), 218. <https://doi.org/10.1186/s13059-020-02132-x> (cf. p. 39)
- HU, X., HU, Y., WU, F., LEUNG, R. W. T. & QIN, J. (2020). Integration of single-cell multi-omics for gene regulatory network inference. *Computational and Structural Biotechnology Journal*, 18, 1925-1938. <https://doi.org/10.1016/j.csbj.2020.06.033> (cf. p. 64, 68, 69)
- HUYNH-THU, V. A., IRRTHUM, A., WEHENKEL, L. & GEURTS, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods [Publisher : Public Library of Science]. *PLOS ONE*, 5(9), e12776. <https://doi.org/10.1371/journal.pone.0012776> (cf. p. 64, 65, 67, 68)
- IKONOMI, N., KÜHLWEIN, S. D., SCHWAB, J. D. & KESTLER, H. A. (2020). Awakening the HSC : Dynamic Modeling of HSC Maintenance Unravels Regulation of the TP53 Pathway and Quiescence. *Frontiers in Physiology*, 11, 848. <https://doi.org/10.3389/fphys.2020.00848> (cf. p. 63)
- ITO, K., HIRAO, A., ARAI, F., TAKUBO, K., MATSUOKA, S., MIYAMOTO, K., OHMURA, M., NAKA, K., HOSOKAWA, K., IKEDA, Y. & SUDA, T. (2006). Reactive oxygen species act through p38 MAPK to limit the lifespan of hematopoietic stem cells [Number : 4]. *Nature Medicine*, 12(4), 446-451. <https://doi.org/10.1038/nm1388> (cf. p. 29)
- JAISWAL, S. & EBERT, B. L. (2019). Clonal hematopoiesis in human aging and disease [Publisher : American Association for the Advancement of Science Section : Review]. *Science*, 366(6465). <https://doi.org/10.1126/science.aan4673> (cf. p. 118)

- JAITIN, D. A., WEINER, A., YOFE, I., LARA-ASTIASO, D., KEREN-SHAUL, H., DAVID, E., SALAME, T. M., TANAY, A., van OUDENAARDEN, A. & AMIT, I. (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq [Number : 7]. *Cell*, 167(7), 1883-1896.e15. <https://doi.org/10.1016/j.cell.2016.11.039> (cf. p. 49)
- JANZEN, V., FORKERT, R., FLEMING, H. E., SAITO, Y., WARING, M. T., DOMBKOWSKI, D. M., CHENG, T., DEPINHO, R. A., SHARPLESS, N. E. & SCADDEN, D. T. (2006). Stem-cell ageing modified by the cyclin-dependent kinase inhibitor p16INK4a [Number : 7110]. *Nature*, 443(7110), 421-426. <https://doi.org/10.1038/nature05159> (cf. p. 30, 31)
- JOHNSON, W. E., LI, C. & RABINOVIC, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods [Number : 1]. *Biostatistics*, 8(1), 118-127. <https://doi.org/10.1093/biostatistics/kxj037> (cf. p. 40)
- KANEHISA, M. & GOTO, S. (2000). KEGG : Kyoto Encyclopedia of Genes and Genomes [Publisher : Oxford Academic]. *Nucleic Acids Research*, 28(1), 27-30. <https://doi.org/10.1093/nar/28.1.27> (cf. p. 51)
- KAUFFMAN, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3), 437-467. [https://doi.org/10.1016/0022-5193\(69\)90015-0](https://doi.org/10.1016/0022-5193(69)90015-0) (cf. p. 52)
- KIRSCHNER, K., CHANDRA, T., KISELEV, V., FLORES-SANTA CRUZ, D., MACAULAY, I. C., PARK, H. J., LI, J., KENT, D. G., KUMAR, R., PASK, D. C., HAMILTON, T. L., HEMBERG, M., REIK, W. & GREEN, A. R. (2017). Proliferation Drives Aging-Related Functional Decline in a Subpopulation of the Hematopoietic Stem Cell Compartment [Number : 8]. *Cell Reports*, 19(8), 1503-1511. <https://doi.org/10.1016/j.celrep.2017.04.074> (cf. p. 49, 74, 76, 117, 118, 159, 161)
- KISELEV, V. Y., KIRSCHNER, K., SCHaub, M. T., ANDREWS, T., YIU, A., CHANDRA, T., NATARAJAN, K. N., REIK, W., BARAHONA, M., GREEN, A. R. & HEMBERG, M. (2017). SC3 : consensus clustering of single-cell RNA-seq data [Number : 5 Publisher : Nature Publishing Group]. *Nature Methods*, 14(5), 483-486. <https://doi.org/10.1038/nmeth.4236> (cf. p. 33)
- KÖHLER, A., SCHMITHORST, V., FILIPPI, M.-D., RYAN, M. A., DARIA, D., GUNZER, M. & GEIGER, H. (2009). Altered cellular dynamics and endosteal location of aged early hematopoietic progenitor cells revealed by time-lapse intravital imaging in long bones [Number : 2]. *Blood*, 114(2), 290-298. <https://doi.org/10.1182/blood-2008-12-195644> (cf. p. 26)
- KONDRATOVA, M., BARILLOT, E., ZINOVYEV, A. & CALZONE, L. (2020). Modelling of Immune Checkpoint Network Explains Synergistic Effects of Combined Immune Checkpoint Inhibitor Therapy and the Impact of Cytokines in Patient Response [Number : 12 Publisher : Multidisciplinary Digital Publishing Institute]. *Cancers*, 12(12), 3600. <https://doi.org/10.3390/cancers12123600> (cf. p. 63, 64)
- KORSUNSKY, I., MILLARD, N., FAN, J., SLOWIKOWSKI, K., ZHANG, F., WEI, K., BAGLAENKO, Y., BRENNER, M., LOH, P.-R. & RAYCHAUDHURI, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony [Number : 12]. *Nature Methods*, 16(12), 1289-1296. <https://doi.org/10.1038/s41592-019-0619-0> (cf. p. 41)
- KÖSTER, J. & RAHMANN, S. (2018). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 34(20), 3600-3600. <https://doi.org/10.1093/bioinformatics/bty350> (cf. p. 75)
- KOWALCZYK, M. S., TIROSH, I., HECKL, D., RAO, T. N., DIXIT, A., HAAS, B. J., SCHNEIDER, R. K., WAGERS, A. J., EBERT, B. L. & REGEV, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells

- [Number : 12]. *Genome Research*, 25(12), 1860-1872. <https://doi.org/10.1101/gr.192237.115> (cf. p. 24, 25, 31, 39, 49, 74, 76, 77, 117, 160, 161)
- KRAMER, A. & CHALLEN, G. A. (2017). The epigenetic basis of hematopoietic stem cell aging [Number : 1]. *Seminars in Hematology*, 54(1), 19-24. <https://doi.org/10.1053/j.seminhematol.2016.10.006> (cf. p. 31)
- KRUMSIEK, J., MARR, C., SCHROEDER, T. & THEIS, F. J. (2011). Hierarchical Differentiation of Myeloid Progenitors Is Encoded in the Transcription Factor Network [Number : 8]. *PLoS ONE*, 6(8). <https://doi.org/10.1371/journal.pone.0022649> (cf. p. 63, 120, 157)
- LA MANNO, G., SOLDATOV, R., ZEISEL, A., BRAUN, E., HOCHGERNER, H., PETUKHOV, V., LIDSCHREIBER, K., KASTRITI, M. E., LÖNNERBERG, P., FURLAN, A., FAN, J., BORM, L. E., LIU, Z., van BRUGGEN, D., GUO, J., HE, X., BARKER, R., SUNDSTRÖM, E., CASTELO-BRANCO, G., ... KHARCHENKO, P. V. (2018). RNA velocity of single cells [Number : 7719 Publisher : Nature Publishing Group]. *Nature*, 560(7719), 494-498. <https://doi.org/10.1038/s41586-018-0414-6> (cf. p. 41)
- LÄHDESMÄKI, H., SHMULEVICH, I. & YLI-HARJA, O. (2003). On Learning Gene Regulatory Networks Under the Boolean Network Model. *Machine Learning*, 52(1), 147-167. <https://doi.org/10.1023/A:1023905711304> (cf. p. 70)
- LAIOSA, C. V., STADTFELD, M. & GRAF, T. (2006). Determinants of lymphoid-myeloid lineage diversification. *Annual Review of Immunology*, 24, 705-738. <https://doi.org/10.1146/annurev.immunol.24.021605.090742> (cf. p. 22)
- LANGE, C. & CALEGARI, F. (2010). Cdks and cyclins link G1 length and differentiation of embryonic, neural and hematopoietic stem cells [Number : 10]. *Cell Cycle*, 9(10), 1893-1900. <https://doi.org/10.4161/cc.9.10.11598> (cf. p. 25)
- LAURENTI, E., FRELIN, C., XIE, S., FERRARI, R., DUNANT, C. F., ZANDI, S., NEUMANN, A., PLUMB, I., DOULATOV, S., CHEN, J., APRIL, C., FAN, J.-B., ISCOVE, N. & DICK, J. E. (2015). CDK6 Levels Regulate Quiescence Exit in Human Hematopoietic Stem Cells [Number : 3]. *Cell Stem Cell*, 16(3), 302-313. <https://doi.org/10.1016/j.stem.2015.01.017> (cf. p. 25)
- LAURENTI, E. & GÖTTGENS, B. (2018). From haematopoietic stem cells to complex differentiation landscapes [Number : 7689]. *Nature*, 553(7689), 418-426. <https://doi.org/10.1038/nature25022> (cf. p. 18, 21, 45)
- LAURIDSEN, F. K. B., JENSEN, T. L., RAPIN, N., ASLAN, D., WILHELMSON, A. S., PUNDHIR, S., REHN, M., PAUL, F., GILADI, A., HASEMANN, M. S., SERUP, P., AMIT, I. & PORSE, B. T. (2018). Differences in Cell Cycle Status Underlie Transcriptional Heterogeneity in the HSC Compartment [Number : 3]. *Cell Reports*, 24(3), 766-780. <https://doi.org/10.1016/j.celrep.2018.06.057> (cf. p. 42)
- LAWRANCE, A. J. (1976). On Conditional and Partial Correlation [Publisher : Taylor & Francis \_eprint : <https://www.tandfonline.com/doi/pdf/10.1080/00031305.1976.10479163>]. *The American Statistician*, 30(3), 146-149. <https://doi.org/10.1080/00031305.1976.10479163> (cf. p. 66)
- LE NOVÈRE, N. (2015). Quantitative and logic modelling of molecular and gene networks [Number : 3 Publisher : Nature Publishing Group]. *Nature Reviews Genetics*, 16(3), 146-158. <https://doi.org/10.1038/nrg3885> (cf. p. 51)
- LIANG, S., FUHRMAN, S. & SOMOGYI, R. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing*. *Pacific Symposium on Biocomputing*, 18-29 (cf. p. 70).
- LICATA, L., LO SURDO, P., IANNUCCELLI, M., PALMA, A., MICARELLI, E., PERFETTO, L., PELUSO, D., CALDERONE, A., CASTAGNOLI, L. & CESARENI, G. (2020). SIGNOR 2.0, the SIGnaling

- Network Open Resource 2.0 : 2019 update. *Nucleic Acids Research*, 48(D1), D504-D510. <https://doi.org/10.1093/nar/gkz949> (cf. p. 51)
- LIEBERMAN, Y., ROKACH, L. & SHAY, T. (2018). CaSTLe – Classification of single cells by transfer learning : Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments [Number : 10]. *PLOS ONE*, 13(10), e0205499. <https://doi.org/10.1371/journal.pone.0205499> (cf. p. 35)
- LIM, C. Y., WANG, H., WOODHOUSE, S., PITERMAN, N., WERNISCH, L., FISHER, J. & GÖTTGENS, B. (2016). BTR : training asynchronous Boolean models using single-cell expression data [Number : 1]. *BMC Bioinformatics*, 17(1), 355. <https://doi.org/10.1186/s12859-016-1235-y> (cf. p. 65, 69, 71)
- LIU, L., LIU, C., QUINTERO, A., WU, L., YUAN, Y., WANG, M., CHENG, M., LENG, L., XU, L., DONG, G., LI, R., LIU, Y., WEI, X., XU, J., CHEN, X., LU, H., CHEN, D., WANG, Q., ZHOU, Q., ... XU, X. (2019). Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity [Number : 1]. *Nature Communications*, 10(1), 470. <https://doi.org/10.1038/s41467-018-08205-7> (cf. p. 48)
- LOEFFLER, D., WEHLING, A., SCHNEITER, F., ZHANG, Y., MÜLLER-BÖTTICHER, N., HOPPE, P. S., HILSENBECK, O., KOKKALIARIS, K. D., ENDELE, M. & SCHROEDER, T. (2019). Asymmetric lysosome inheritance predicts activation of haematopoietic stem cells [Number : 7774]. *Nature*, 573(7774), 426-429. <https://doi.org/10.1038/s41586-019-1531-6> (cf. p. 25)
- LUECKEN, M. D. & THEIS, F. J. (2019). Current best practices in single-cell RNA-seq analysis : a tutorial [Number : 6 Publisher : John Wiley & Sons, Ltd]. *Molecular Systems Biology*, 15(6), e8746. <https://doi.org/10.15252/msb.20188746> (cf. p. 34, 35, 38, 40, 113)
- LUN, A. T., MCCARTHY, D. J. & MARIONI, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, 5, 2122. <https://doi.org/10.12688/f1000research.9501.2> (cf. p. 33)
- MA, F. & PELLEGRINI, M. (2020). ACTINN : automated identification of cell types in single cell RNA sequencing [Number : 2]. *Bioinformatics*, 36(2), 533-538. <https://doi.org/10.1093/bioinformatics/btz592> (cf. p. 35)
- MALINGE, S., THIOLIER, C., CHLON, T. M., DORÉ, L. C., DIEBOLD, L., BLUTEAU, O., MABIALAH, V., VAINCHENKER, W., DESSEN, P., WINANDY, S., MERCHER, T. & CRISPINO, J. D. (2013). Ikaros inhibits megakaryopoiesis through functional interaction with GATA-1 and NOTCH signaling [Number : 13]. *Blood*, 121(13), 2440-2451. <https://doi.org/10.1182/blood-2012-08-450627> (cf. p. 157)
- MANN, M., MEHTA, A., de BOER, C. G., KOWALCZYK, M. S., LEE, K., HALDEMAN, P., ROGEL, N., KNECHT, A. R., FAROUQ, D., REGEV, A. & BALTIMORE, D. (2018). Heterogeneous Responses of Hematopoietic Stem Cells to Inflammatory Stimuli Are Altered with Age [Number : 11]. *Cell Reports*, 25(11), 2992-3005.e5. <https://doi.org/10.1016/j.celrep.2018.11.056> (cf. p. 46, 49, 74, 76, 117, 161)
- MARBACH, D., COSTELLO, J. C., KÜFFNER, R., VEGA, N., PRILL, R. J., CAMACHO, D. M., ALLISON, K. R., KELLIS, M., COLLINS, J. J. & STOLOVITZKY, G. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8), 796-804. <https://doi.org/10.1038/nmeth.2016> (cf. p. 64)
- MARYANOVICH, M., ZAHALKA, A. H., PIERCE, H., PINHO, S., NAKAHARA, F., ASADA, N., WEI, Q., WANG, X., CIERO, P., XU, J., LEFTIN, A. & FRENETTE, P. S. (2018). Adrenergic nerve degeneration in bone marrow drives aging of the hematopoietic stem cell niche [Number : 6]. *Nature Medicine*, 24(6), 782-791. <https://doi.org/10.1038/s41591-018-0030-x> (cf. p. 31)

- MATSUMOTO, H., KIRYU, H., FURUSAWA, C., KO, M. S. H., KO, S. B. H., GOUDA, N., HAYASHI, T. & NIKAIKO, I. (2017). SCODE : an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*, 33(15), 2314-2321. <https://doi.org/10.1093/bioinformatics/btx194> (cf. p. 64)
- MATYS, V., FRICKE, E., GEFFERS, R., GÖSSLING, E., HAUBROCK, M., HEHL, R., HORNISCHER, K., KARAS, D., KEL, A. E., KEL-MARGOULIS, O. V., KLOOS, D.-U., LAND, S., LEWICKI-POTAPOV, B., MICHAEL, H., MÜNCH, R., REUTER, I., ROTERT, S., SAXEL, H., SCHEER, M., ... WINGENDER, E. (2003). TRANSFAC ® : transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1), 374-378. <https://doi.org/10.1093/nar/gkg108> (cf. p. 51)
- MENDELSON, A. & FRENETTE, P. S. (2014). Hematopoietic stem cell niche maintenance during homeostasis and regeneration [Number : 8]. *Nature Medicine*, 20(8), 833-846. <https://doi.org/10.1038/nm.3647> (cf. p. 20)
- MENDOZA, L. (2006). A network model for the control of the differentiation process in Th cells. *Biosystems*, 84(2), 101-114. <https://doi.org/10.1016/j.biosystems.2005.10.004> (cf. p. 62)
- MILES, L. A., BOWMAN, R. L., MERLINSKY, T. R., CSETE, I. S., OOI, A. T., DURRUTHY-DURRUTHY, R., BOWMAN, M., FAMULARE, C., PATEL, M. A., MENDEZ, P., AINALI, C., DEMAREE, B., DELLEY, C. L., ABATE, A. R., MANIVANNAN, M., SAHU, S., GOLDBERG, A. D., BOLTON, K. L., ZEHIR, A., ... LEVINE, R. L. (2020). Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature*, 587(7834), 477-482. <https://doi.org/10.1038/s41586-020-2864-x> (cf. p. 50)
- MIN, I. M., PIETRAMAGGIORI, G., KIM, F. S., PASSEGÜÉ, E., STEVENSON, K. E. & WAGERS, A. J. (2008). The Transcription Factor EGR1 Controls Both the Proliferation and Localization of Hematopoietic Stem Cells. *Cell Stem Cell*, 2(4), 380-391. <https://doi.org/10.1016/j.stem.2008.01.015> (cf. p. 158, 159)
- MOERMAN, T., AIBAR SANTOS, S., BRAVO GONZÁLEZ-BLAS, C., SIMM, J., MOREAU, Y., AERTS, J. & AERTS, S. (2019). GRNBoost2 and Arboreto : efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12), 2159-2161. <https://doi.org/10.1093/bioinformatics/bty916> (cf. p. 67, 68)
- MOIGNARD, V., WOODHOUSE, S., HAGHVERDI, L., LILLY, A. J., TANAKA, Y., WILKINSON, A. C., BUETTNER, F., MACAULAY, I. C., JAWAID, W., DIAMANTI, E., NISHIKAWA, S.-I., PITERNAN, N., KOUSKOFF, V., THEIS, F. J., FISHER, J. & GÖTTGENS, B. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements [Number : 3 Publisher : Nature Publishing Group]. *Nature Biotechnology*, 33(3), 269-276. <https://doi.org/10.1038/nbt.3154> (cf. p. 64, 65, 69, 71, 157)
- MÖLDER, F., JABLONSKI, K. P., LETCHER, B., HALL, M. B., TOMKINS-TINCH, C. H., SOCHAT, V., FORSTER, J., LEE, S., TWARDZIOK, S. O., KANITZ, A., WILM, A., HOLTGREWE, M., RAHMANN, S., NAHNSEN, S. & KÖSTER, J. (2021). Sustainable data analysis with Snakemake. *F1000Research*, 10, 33. <https://doi.org/10.12688/f1000research.29032.1> (cf. p. 75, 77)
- MOMBACH, J. C., BUGS, C. A. & CHAOUIYA, C. (2014). Modelling the onset of senescence at the G1/S cell cycle checkpoint. *BMC Genomics*, 15(7), S7. <https://doi.org/10.1186/1471-2164-15-S7-S7> (cf. p. 60)
- MORITA, K., WANG, F., JAHN, K., HU, T., TANAKA, T., SASAKI, Y., KUIPERS, J., LOGHAVI, S., WANG, S. A., YAN, Y., FURUDATE, K., MATTHEWS, J., LITTLE, L., GUMBS, C., ZHANG, J., SONG, X., THOMPSON, E., PATEL, K. P., BUESO-RAMOS, C. E., ... TAKAHASHI, K. (2020). Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics.

- Nature Communications*, 11(1), 5327. <https://doi.org/10.1038/s41467-020-19119-8> (cf. p. 50)
- MORRISON, S. J., WANDYCZ, A. M., AKASHI, K., GLOBERSON, A. & WEISSMAN, I. L. (1996). The aging of hematopoietic stem cells [Number : 9]. *Nature Medicine*, 2(9), 1011-1016. <https://doi.org/10.1038/nm0996-1011> (cf. p. 26)
- MÜSSEL, C., HOPFENSITZ, M. & KESTLER, H. A. (2010). BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics*, 26(10), 1378-1380. <https://doi.org/10.1093/bioinformatics/btq124> (cf. p. 62)
- NALDI, A., BERENGUIER, D., FAURÉ, A., LOPEZ, F., THIEFFRY, D. & CHAOUIYA, C. (2009). Logical modelling of regulatory networks with GINsim 2.3. *Biosystems*, 97(2), 134-139. <https://doi.org/10.1016/j.biosystems.2009.04.008> (cf. p. 62)
- NALDI, A. (2018). BioLQM : A Java Toolkit for the Manipulation and Conversion of Logical Qualitative Models of Biological Networks [Publisher : Frontiers]. *Frontiers in Physiology*, 9. <https://doi.org/10.3389/fphys.2018.01605> (cf. p. 62)
- NALDI, A., CARNEIRO, J., CHAOUIYA, C. & THIEFFRY, D. (2010). Diversity and Plasticity of Th Cell Types Predicted from Regulatory Network Modelling [Publisher : Public Library of Science]. *PLOS Computational Biology*, 6(9), e1000912. <https://doi.org/10.1371/journal.pcbi.1000912> (cf. p. 62)
- NALDI, A., HERNANDEZ, C., LEVY, N., STOLL, G., MONTEIRO, P. T., CHAOUIYA, C., HELIKAR, T., ZINOVYEV, A., CALZONE, L., COHEN-BOULAKIA, S., THIEFFRY, D. & PAULEVÉ, L. (2018). The CoLoMoTo Interactive Notebook : Accessible and Reproducible Computational Analyses for Qualitative Biological Networks [Publisher : Frontiers]. *Frontiers in Physiology*, 9. <https://doi.org/10.3389/fphys.2018.00680> (cf. p. 62)
- NESTOROWA, S., HAMEY, F. K., PIJUAN SALA, B., DIAMANTI, E., SHEPHERD, M., LAURENTI, E., WILSON, N. K., KENT, D. G. & GÖTTGENS, B. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation [Number : 8]. *Blood*, 128(8), e20-31. <https://doi.org/10.1182/blood-2016-05-716480> (cf. p. 42, 46)
- NG, S. Y.-M., YOSHIDA, T., ZHANG, J. & GEORGOPoulos, K. (2009). Genome-wide lineage-specific transcriptional networks underscore Ikaros-dependent lymphoid priming in hematopoietic stem cells [Number : 4]. *Immunity*, 30(4), 493-507. <https://doi.org/10.1016/j.jimmuni.2009.01.014> (cf. p. 157)
- NITTA, E., ITOKAWA, N., YABATA, S., KOIDE, S., HOU, L.-B., OSHIMA, M., AOYAMA, K., SARAYA, A. & IWAMA, A. (2020). Bmi1 counteracts hematopoietic stem cell aging by repressing target genes and enforcing the stem cell gene signature [Number : 3]. *Biochemical and Biophysical Research Communications*, 521(3), 612-619. <https://doi.org/10.1016/j.bbrc.2019.10.153> (cf. p. 30)
- NORDDAHL, G. L., PRONK, C. J., WAHLESTEDT, M., STEN, G., NYGREN, J. M., UGALE, A., SIGVARDSSON, M. & BRYDER, D. (2011). Accumulating mitochondrial DNA mutations drive premature hematopoietic aging phenotypes distinct from physiological stem cell aging [Number : 5]. *Cell Stem Cell*, 8(5), 499-510. <https://doi.org/10.1016/j.stem.2011.03.009> (cf. p. 29, 31)
- OLSSON, A., VENKATASUBRAMANIAN, M., CHAUDHRI, V. K., ARONOW, B. J., SALOMONIS, N., SINGH, H. & GRIMES, H. L. (2016). Single-cell analysis of mixed-lineage states leading to a binary cell fate choice [Number : 7622]. *Nature*, 537(7622), 698-702. <https://doi.org/10.1038/nature19348> (cf. p. 42)
- OMRANIAN, N., ELOUNDOU-MBEBI, J. M. O., MUELLER-ROEBER, B. & NIKOLOSKI, Z. (2016). Gene regulatory network inference using fused LASSO on multiple data sets. *Scientific Reports*, 6(1), 20533. <https://doi.org/10.1038/srep20533> (cf. p. 67)

- ORFORD, K. W. & SCADDEN, D. T. (2008). Deconstructing stem cell self-renewal : genetic insights into cell-cycle regulation [Number : 2]. *Nature Reviews Genetics*, 9(2), 115-128. <https://doi.org/10.1038/nrg2269> (cf. p. 24, 25)
- OSTROWSKI, M., PAULEVÉ, L., SCHAUB, T., SIEGEL, A. & GUZILOWSKI, C. (2016). Boolean Network Identification from Perturbation Time Series Data combining Dynamics Abstraction and Logic Programming [Publisher : Elsevier]. *BioSystems*. <https://doi.org/10.1016/j.biosystems.2016.07.009> (cf. p. 69, 71)
- PANG, W. W., PRICE, E. A., SAHOO, D., BEERMAN, I., MALONEY, W. J., ROSSI, D. J., SCHRIER, S. L. & WEISSMAN, I. L. (2011). Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age [Number : 50]. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), 20012-20017. <https://doi.org/10.1073/pnas.1116110108> (cf. p. 26)
- PAPILI GAO, N., UD-DEAN, S. M. M., GANDRILLON, O. & GUNAWAN, R. (2018). SINCERITIES : inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34(2), 258-266. <https://doi.org/10.1093/bioinformatics/btx575> (cf. p. 64, 67)
- PAPPENHEIM, A. (1896). Ueber Entwicklung und Ausbildung der Erythroblasten. *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin*, 145(3), 587-643. <https://doi.org/10.1007/BF01969901> (cf. p. 17)
- PASSEGUÉ, E., WAGNER, E. F. & WEISSMAN, I. L. (2004). JunB deficiency leads to a myeloproliferative disorder arising from hematopoietic stem cells [Number : 3]. *Cell*, 119(3), 431-443. <https://doi.org/10.1016/j.cell.2004.10.010> (cf. p. 23, 118)
- PAUL, F., ARKIN, Y., GILADI, A., JAITIN, D. A., KENIGSBERG, E., KEREN-SHAUL, H., WINTER, D., LARA-ASTIASO, D., GURY, M., WEINER, A., DAVID, E., COHEN, N., LAURIDSEN, F. K. B., HAAS, S., SCHLITZER, A., MILDNER, A., GINHOUX, F., JUNG, S., TRUMPP, A., ... AMIT, I. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors [Number : 7]. *Cell*, 163(7), 1663-1677. <https://doi.org/10.1016/j.cell.2015.11.013> (cf. p. 41)
- PAULEVÉ, L. (2017). Pint : A Static Analyzer for Transient Dynamics of Qualitative Networks with IPython Interface [Series Title : Lecture Notes in Computer Science]. In J. FERET & H. KOEPLI (Éd.), *Computational Methods in Systems Biology* (p. 309-316). Springer International Publishing. [https://doi.org/10.1007/978-3-319-67471-1\\_20](https://doi.org/10.1007/978-3-319-67471-1_20). (Cf. p. 62)
- PAULEVÉ, L., KOLČÁK, J., CHATAIN, T. & HAAR, S. (2020). Reconciling qualitative, abstract, and scalable modeling of biological networks. *Nature Communications*, 11. <https://doi.org/10.1038/s41467-020-18112-5> (cf. p. 58, 60, 62, 122)
- PAULEVÉ, L. & RICHARD, A. (2012). Static Analysis of Boolean Networks Based on Interaction Graphs : A Survey [Publisher : Elsevier]. *Electronic Notes in Theoretical Computer Science*, 284, 93-104. <https://doi.org/10.1016/j.entcs.2012.05.017> (cf. p. 57)
- PHAM, K., SACIRBEGOVIC, F. & RUSSELL, S. M. (2014). Polarized Cells, Polarized Views : Asymmetric Cell Division in Hematopoietic Cells. *Frontiers in Immunology*, 5. <https://doi.org/10.3389/fimmu.2014.00026> (cf. p. 24)
- PIETRAS, E. M., REYNAUD, D., KANG, Y.-A., CARLIN, D., CALERO-NIETO, F. J., LEAVITT, A. D., STUART, J. M., GÖTTGENS, B. & PASSEGUÉ, E. (2015). Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions [Number : 1]. *Cell Stem Cell*, 17(1), 35-46. <https://doi.org/10.1016/j.stem.2015.05.003> (cf. p. 18)

- PIETRAS, E. M., WARR, M. R. & PASSEGUÉ, E. (2011). Cell cycle regulation in hematopoietic stem cells [Number : 5]. *Journal of Cell Biology*, 195(5), 709-720. <https://doi.org/10.1083/jcb.201102131> (cf. p. 22, 24)
- PIMANDA, J. E., OTTERSBACK, K., KNEZEVIC, K., KINSTON, S., CHAN, W. Y. I., WILSON, N. K., LANDRY, J.-R., WOOD, A. D., KOLB-KOKOCINSKI, A., GREEN, A. R., TANNAHILL, D., LACAUD, G., KOUSKOFF, V. & GÖTTGENS, B. (2007). Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development [Number : 45]. *Proceedings of the National Academy of Sciences*, 104(45), 17692-17697. <https://doi.org/10.1073/pnas.0707045104> (cf. p. 22)
- PRATAPA, A., JALIHAL, A. P., LAW, J. N., BHARADWAJ, A. & MURALI, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2), 147-154. <https://doi.org/10.1038/s41592-019-0690-6> (cf. p. 64-66)
- QIU, J., PAPATSENKO, D., NIU, X., SCHANEL, C. & MOORE, K. (2014). Divisional History and Hematopoietic Stem Cell Function during Homeostasis [Number : 4]. *Stem Cell Reports*, 2(4), 473-490. <https://doi.org/10.1016/j.stemcr.2014.01.016> (cf. p. 29)
- QIU, X., MAO, Q., TANG, Y., WANG, L., CHAWLA, R., PLINER, H. A. & TRAPNELL, C. (2017). Reversed graph embedding resolves complex single-cell trajectories [Number : 10]. *Nature Methods*, 14(10), 979-982. <https://doi.org/10.1038/nmeth.4402> (cf. p. 36)
- RAMALHO-SANTOS, M. & WILLENBRING, H. (2007). On the Origin of the Term “Stem Cell” [Number : 1]. *Cell Stem Cell*, 1(1), 35-38. <https://doi.org/10.1016/j.stem.2007.05.013> (cf. p. 17)
- RAO, K. N., SMUDA, C., GREGORY, G. D., MIN, B. & BROWN, M. A. (2013). Ikaros limits basophil development by suppressing C/EBP- expression [Number : 15]. *Blood*, 122(15), 2572-2581. <https://doi.org/10.1182/blood-2013-04-494625> (cf. p. 157)
- REMY, E., RUET, P. & THIEFFRY, D. (2008). Graphic requirements for multistability and attractive cycles in a Boolean dynamical framework [Publisher : Academic Press]. *Advances in Applied Mathematics*, 41(3), 335-350. <https://doi.org/10.1016/j.aam.2007.11.003> (cf. p. 57)
- RENDERS, S., SVENDSEN, A. F., PANTEN, J., RAMA, N., MARYANOVICH, M., SOMMERKAMP, P., LADEL, L., REDAVID, A. R., GIBERT, B., LAZARE, S., DUCAROUGE, B., SCHÖNBERGER, K., NARR, A., TOURBEZ, M., DETHMERS-AUSEMA, B., ZWART, E., HOTZ-WAGENBLATT, A., ZHANG, D., KORN, C., ... TRUMPP, A. (2021). Niche derived netrin-1 regulates hematopoietic stem cell dormancy via its receptor neogenin-1 [Number : 1]. *Nature Communications*, 12(1), 608. <https://doi.org/10.1038/s41467-020-20801-0> (cf. p. 31)
- REYA, T., DUNCAN, A. W., AILLES, L., DOMEN, J., SCHERER, D. C., WILLERT, K., HINTZ, L., NUSSE, R. & WEISSMAN, I. L. (2003). A role for Wnt signalling in self-renewal of haematopoietic stem cells [Number : 6938]. *Nature*, 423(6938), 409-414. <https://doi.org/10.1038/nature01593> (cf. p. 22)
- RICHARD, A. & COMET, J.-P. (2007). Necessary conditions for multistationarity in discrete dynamical systems. *Discrete Applied Mathematics*, 155(18), 2403-2413. <https://doi.org/10.1016/j.dam.2007.04.019> (cf. p. 57)
- RODRIGUEZ-FRATICELLI, A. E. & CAMARGO, F. (2021). Systems analysis of hematopoiesis using single-cell lineage tracing [Number : 1]. *Current Opinion in Hematology*, 28(1), 18-27. <https://doi.org/10.1097/MOH.0000000000000624> (cf. p. 47, 158)
- RODRIGUEZ-FRATICELLI, A. E., WEINREB, C., WANG, S.-W., MIGUELES, R. P., JANKOVIC, M., USART, M., KLEIN, A. M., LOWELL, S. & CAMARGO, F. D. (2020). Single-cell lineage

- tracing unveils a role for TCF15 in haematopoiesis [Number : 7817]. *Nature*, 583(7817), 585-589. <https://doi.org/10.1038/s41586-020-2503-6> (cf. p. 42, 47, 49)
- RODRIGUEZ-FRATICELLI, A. E., WOLOCK, S. L., WEINREB, C. S., PANERO, R., PATEL, S. H., JANKOVIC, M., SUN, J., CALOGERO, R. A., KLEIN, A. M. & CAMARGO, F. D. (2018). Clonal analysis of lineage fate in native haematopoiesis [Number : 7687]. *Nature*, 553(7687), 212-216. <https://doi.org/10.1038/nature25168> (cf. p. 42, 46, 47, 76, 120, 159)
- RODRÍGUEZ-JORGE, O., KEMPIS-CALANIS, L. A., ABOU-JAOUDÉ, W., GUTIÉRREZ-REYNA, D. Y., HERNANDEZ, C., RAMIREZ-PLIEGO, O., THOMAS-CHOLIER, M., SPICUGLIA, S., SANTANA, M. A. & THIEFFRY, D. (2019). Cooperation between T cell receptor and Toll-like receptor 5 signaling for CD4+ T cell activation [Publisher : American Association for the Advancement of Science Section : Research Article]. *Science Signaling*, 12(577). <https://doi.org/10.1126/scisignal.aar3641> (cf. p. 63)
- ROSSI, D. J., JAMIESON, C. H. M. & WEISSMAN, I. L. (2008). Stems cells and the pathways to aging and cancer [Number : 4]. *Cell*, 132(4), 681-696. <https://doi.org/10.1016/j.cell.2008.01.036> (cf. p. 26)
- ROSSI, D. J., SEITA, J., CZECHOWICZ, A., BHATTACHARYA, D., BRYDER, D. & WEISSMAN, I. L. (2007). Hematopoietic Stem Cell Quiescence Attenuates DNA Damage Response and Permits DNA Damage Accumulation During Aging [Number : 19]. *Cell Cycle*, 6(19), 2371-2376. <https://doi.org/10.4161/cc.6.19.4759> (cf. p. 28)
- RÜBE, C. E., FRICKE, A., WIDMANN, T. A., FÜRST, T., MADRY, H., PFREUNDSCHUH, M. & RÜBE, C. (2011). Accumulation of DNA damage in hematopoietic stem and progenitor cells during human aging [Number : 3]. *PLoS One*, 6(3), e17487. <https://doi.org/10.1371/journal.pone.0017487> (cf. p. 28)
- SAELENS, W., CANNOODT, R., TODOROV, H. & SAEYS, Y. (2019). A comparison of single-cell trajectory inference methods [Number : 5 Publisher : Nature Publishing Group]. *Nature Biotechnology*, 37(5), 547-554. <https://doi.org/10.1038/s41587-019-0071-9> (cf. p. 38, 114)
- SALOMONI, P. & CALEGARI, F. (2010). Cell cycle control of mammalian neural stem cells : putting a speed limit on G1. *Trends in Cell Biology*, 20(5), 233-243. <https://doi.org/10.1016/j.tcb.2010.01.006> (cf. p. 25)
- SANTAGUIDA, M., SCHEPERS, K., KING, B., SABNIS, A. J., FORSBERG, E. C., ATTEMA, J. L., BRAUN, B. S. & PASSEGUÉ, E. (2009). JunB protects against myeloid malignancies by limiting hematopoietic stem cell proliferation and differentiation without affecting self-renewal [Number : 4]. *Cancer Cell*, 15(4), 341-352. <https://doi.org/10.1016/j.ccr.2009.02.016> (cf. p. 23, 118, 159)
- SATIJA, R., FARRELL, J. A., GENNERT, D., SCHIER, A. F. & REGEV, A. (2015). Spatial reconstruction of single-cell gene expression data [Number : 5]. *Nature Biotechnology*, 33(5), 495-502. <https://doi.org/10.1038/nbt.3192> (cf. p. 33)
- SCIALDONE, A., NATARAJAN, K. N., SARAIVA, L. R., PROSERPIO, V., TEICHMANN, S. A., STEGLE, O., MARIONI, J. C. & BUETTNER, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85, 54-61. <https://doi.org/10.1016/j.ymeth.2015.06.021> (cf. p. 35, 39, 113)
- SIGNER, R. A. J., MAGEE, J. A., SALIC, A. & MORRISON, S. J. (2014). Haematopoietic stem cells require a highly regulated protein synthesis rate [Number : 7498]. *Nature*, 509(7498), 49-54. <https://doi.org/10.1038/nature13035> (cf. p. 19)
- SINGH, A. M. & DALTON, S. (2009). The Cell Cycle and Myc Intersect with Mechanisms that Regulate Pluripotency and Reprogramming. *Cell Stem Cell*, 5(2), 141-149. <https://doi.org/10.1016/j.stem.2009.07.003> (cf. p. 25)

- SOMMARIN, M. N. E., DHAPOLA, P., SAFI, F., WARFVINGE, R., ULFSSON, L. G., ERLANDSSON, E., KONTUREK-CIESLA, A., THAKUR, R. K., BÖIERS, C., BRYDER, D. & KARLSSON, G. (2021). Single-Cell Multiomics Reveals Distinct Cell States at the Top of the Human Hematopoietic Hierarchy [Publisher : Cold Spring Harbor Laboratory Section : New Results]. *bioRxiv*, 2021.04.01.437998. <https://doi.org/10.1101/2021.04.01.437998> (cf. p. 42, 48, 49, 74, 117)
- SONESON, C. & ROBINSON, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis [Number : 4 Publisher : Nature Publishing Group]. *Nature Methods*, 15(4), 255-261. <https://doi.org/10.1038/nmeth.4612> (cf. p. 36)
- SPANGRUD, G. J., HEIMFELD, S. & WEISSMAN, I. L. (1988). Purification and characterization of mouse hematopoietic stem cells [Number : 4861]. *Science (New York, N.Y.)*, 241(4861), 58-62. <https://doi.org/10.1126/science.2898810> (cf. p. 17)
- SPECHT, A. T. & LI, J. (2017). LEAP : constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics*, 33(5), 764-766. <https://doi.org/10.1093/bioinformatics/btw729> (cf. p. 64)
- STARCK, J., COHET, N., GONNET, C., SARAZIN, S., DOUBEIKOVSKAIA, Z., DOUBEIKOVSKI, A., VERGER, A., DUTERQUE-COQUILLAUD, M. & MORLE, F. (2003). Functional cross-antagonism between transcription factors FLI-1 and EKLF [Number : 4]. *Molecular and Cellular Biology*, 23(4), 1390-1402. <https://doi.org/10.1128/mcb.23.4.1390-1402.2003> (cf. p. 22)
- STOECKIUS, M., ZHENG, S., HOUCK-LOOMIS, B., HAO, S., YEUNG, B. Z., MAUCK, W. M., SMIBERT, P. & SATIJA, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics [Number : 1]. *Genome Biology*, 19(1), 224. <https://doi.org/10.1186/s13059-018-1603-1> (cf. p. 40)
- STOLL, G., CARON, B., VIARA, E., DUGOURD, A., ZINOVYEV, A., NALDI, A., KROEMER, G., BARILLOT, E. & CALZONE, L. (2017). MaBoSS 2.0 : an environment for stochastic Boolean modeling. *Bioinformatics*, 33(14), 2226-2228. <https://doi.org/10.1093/bioinformatics/btx123> (cf. p. 62, 159)
- STOLL, G., VIARA, E., BARILLOT, E. & CALZONE, L. (2012). Continuous time boolean modeling for biological signaling : application of Gillespie algorithm. *BMC Systems Biology*, 6(1), 116. <https://doi.org/10.1186/1752-0509-6-116> (cf. p. 52, 163)
- STUART, T., BUTLER, A., HOFFMAN, P., HAFEMEISTER, C., PAPALEXI, E., MAUCK, W. M., HAO, Y., STOECKIUS, M., SMIBERT, P. & SATIJA, R. (2019). Comprehensive Integration of Single-Cell Data [Number : 7]. *Cell*, 177(7), 1888-1902.e21. <https://doi.org/10.1016/j.cell.2019.05.031> (cf. p. 33, 35)
- SUN, D., LUO, M., JEONG, M., RODRIGUEZ, B., XIA, Z., HANNAH, R., WANG, H., LE, T., FAULL, K. F., CHEN, R., GU, H., BOCK, C., MEISSNER, A., GÖTTGENS, B., DARLINGTON, G. J., LI, W. & GOODELL, M. A. (2014). Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal [Number : 5]. *Cell stem cell*, 14(5), 673-688. <https://doi.org/10.1016/j.stem.2014.03.002> (cf. p. 31, 159)
- SUN, J., RAMOS, A., CHAPMAN, B., JOHNNIDIS, J. B., LE, L., HO, Y.-J., KLEIN, A., HOFMANN, O. & CAMARGO, F. D. (2014). Clonal dynamics of native haematopoiesis [Number : 7522 Publisher : Nature Publishing Group]. *Nature*, 514(7522), 322-327. <https://doi.org/10.1038/nature13824> (cf. p. 47)
- SVENDSEN, A. F., YANG, D., KIM, K. M., LAZARE, S. S., SKINDER, N., ZWART, E., MURA-MESZAROS, A., AUSEMA, A., EYSS, B. v., de HAAN, G. & BYSTRYKH, L. V. (2021). A comprehensive transcriptome signature of murine hematopoietic stem cell aging. *Blood*. <https://doi.org/10.1182/blood.2020009729> (cf. p. 30, 31, 161)

- SVENSSON, V., VENTO-TORMO, R. & TEICHMANN, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade [Number : 4]. *Nature Protocols*, 13(4), 599-604. <https://doi.org/10.1038/nprot.2017.149> (cf. p. 33)
- SZKLARCZYK, D., MORRIS, J. H., COOK, H., KUHN, M., WYDER, S., SIMONOVIC, M., SANTOS, A., DONCHEVA, N. T., ROTH, A., BORK, P., JENSEN, L. J. & von MERING, C. (2017). The STRING database in 2017 : quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1), D362-D368. <https://doi.org/10.1093/nar/gkw937> (cf. p. 51)
- TAK, T., PREVEDELLO, G., SIMON, G., PAILLON, N., DUFFY, K. R. & PERIÉ, L. (2019). *Simultaneous tracking of division and differentiation from individual hematopoietic stem and progenitor cells reveals within-family homogeneity despite population heterogeneity* (preprint). *Immunology*. <https://doi.org/10.1101/586354>. (Cf. p. 25)
- TANG, F., BARBACIORU, C., WANG, Y., NORDMAN, E., LEE, C., XU, N., WANG, X., BODEAU, J., TUCH, B. B., SIDDIQUI, A., LAO, K. & SURANI, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell [Number : 5]. *Nature Methods*, 6(5), 377-382. <https://doi.org/10.1038/nmeth.1315> (cf. p. 32)
- TANIGUCHI ISHIKAWA, E., GONZALEZ-NIETO, D., GHIAUR, G., DUNN, S. K., FICKER, A. M., MURALI, B., MADHU, M., GUTSTEIN, D. E., FISHMAN, G. I., BARRIO, L. C. & CANCELAS, J. A. (2012). Connexin-43 prevents hematopoietic stem cell senescence through transfer of reactive oxygen species to bone marrow stromal cells [Number : 23]. *Proceedings of the National Academy of Sciences of the United States of America*, 109(23), 9071-9076. <https://doi.org/10.1073/pnas.1120358109> (cf. p. 30)
- THOMAS, R. (1981). On the Relation Between the Logical Structure of Systems and Their Ability to Generate Multiple Steady States or Sustained Oscillations. In J. DELLA DORA, J. DEMONGEOT & B. LACOLLE (Ed.), *Numerical Methods in the Study of Critical Phenomena* (p. 180-193). Springer. [https://doi.org/10.1007/978-3-642-81703-8\\_24](https://doi.org/10.1007/978-3-642-81703-8_24). (Cf. p. 57)
- THOMAS, R. (1973). Boolean formalization of genetic control circuits [Number : 3]. *Journal of Theoretical Biology*, 42(3), 563-585. [https://doi.org/10.1016/0022-5193\(73\)90247-6](https://doi.org/10.1016/0022-5193(73)90247-6) (cf. p. 52, 56, 58)
- TIAN, L., DONG, X., FREYTAG, S., LÊ CAO, K.-A., SU, S., JALALABADI, A., AMANN-ZALCENSTEIN, D., WEBER, T. S., SEIDI, A., JABBARI, J. S., NAIK, S. H. & RITCHIE, M. E. (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments [Number : 6]. *Nature Methods*, 16(6), 479-487. <https://doi.org/10.1038/s41592-019-0425-8> (cf. p. 33)
- TING, S. B., DENEAULT, E., HOPE, K., CELLOT, S., CHAGRAOUI, J., MAYOTTE, N., DORN, J. F., LAVERDURE, J.-P., HARVEY, M., HAWKINS, E. D., RUSSELL, S. M., MADDOX, P. S., ISCOVE, N. N. & SAUVAGEAU, G. (2012). Asymmetric segregation and self-renewal of hematopoietic stem and progenitor cells with endocytic Ap2a2 [Number : 11]. *Blood*, 119(11), 2510-2522. <https://doi.org/10.1182/blood-2011-11-393272> (cf. p. 24)
- TRAAG, V. A., WALTMAN, L. & van ECK, N. J. (2019). From Louvain to Leiden : guaranteeing well-connected communities [Number : 1 Publisher : Nature Publishing Group]. *Scientific Reports*, 9(1), 5233. <https://doi.org/10.1038/s41598-019-41695-z> (cf. p. 35)
- van GALEN, P., HOVESTADT, V., WADSWORTH II, M. H., HUGHES, T. K., GRIFFIN, G. K., BATTAGLIA, S., VERGA, J. A., STEPHANSKY, J., PASTIKA, T. J., LOMBARDI STORY, J., PINKUS, G. S., POZDNYAKOVA, O., GALINSKY, I., STONE, R. M., GRAUBERT, T. A., SHALEK, A. K., ASTER, J. C., LANE, A. A. & BERNSTEIN, B. E. (2019). Single-Cell RNA-Seq Reveals AML Hie-

- rarchies Relevant to Disease Progression and Immunity. *Cell*, 176(6), 1265-1281.e24. <https://doi.org/10.1016/j.cell.2019.01.031> (cf. p. 50)
- VELTEN, L., HAAS, S. F., RAFFEL, S., BLASZKIEWICZ, S., ISLAM, S., HENNIG, B. P., HIRCHE, C., LUTZ, C., BUSS, E. C., NOWAK, D., BOCH, T., HOFMANN, W.-K., HO, A. D., HUBER, W., TRUMPP, A., ESSERS, M. A. G. & STEINMETZ, L. M. (2017). Human haematopoietic stem cell lineage commitment is a continuous process [Number : 4]. *Nature Cell Biology*, 19(4), 271-281. <https://doi.org/10.1038/ncb3493> (cf. p. 45)
- VERLINGUE, L., DUGOURD, A., STOLL, G., BARILLOT, E., CALZONE, L. & LONDOÑO-VALLEJO, A. (2016). A comprehensive approach to the molecular determinants of lifespan using a Boolean model of geroconversion. *Aging Cell*, 15(6), 1018-1026. <https://doi.org/10.1111/acel.12504> (cf. p. 60)
- VERNY, L., SELLA, N., AFFELDT, S., SINGH, P. P. & ISAMBERT, H. (2017). Learning causal networks with latent variables from multivariate information in genomic data. *PLoS computational biology*, 13(10), e1005662. <https://doi.org/10.1371/journal.pcbi.1005662> (cf. p. 64, 66)
- VINCENT-FABERT, C., PLATET, N., VANDEVELDE, A., POPLINEAU, M., KOUBI, M., FINETTI, P., TIBERI, G., IMBERT, A.-M., BERTUCCI, F. & DUPREZ, E. (2016). PLZF mutation alters mouse hematopoietic stem cell function and cell cycle progression [Number : 15]. *Blood*, 127(15), 1881-1885. <https://doi.org/10.1182/blood-2015-09-666974> (cf. p. 30)
- WADDINGTON, C. H. & KACSER, H. (1957). *The Strategy of the Genes : A Discussion of Some Aspects of Theoretical Biology*. Allen & Unwin. (Cf. p. 42, 46).
- WANG, C. Q., UDUPA, K. B., XIAO, H. & LIPSCHITZ, D. A. (1995). Effect of age on marrow macrophage number and function. *Aging (Milan, Italy)*, 7(5), 379-384. <https://doi.org/10.1007/BF03324349> (cf. p. 26)
- WANG, S., SUN, H., MA, J., ZANG, C., WANG, C., WANG, J., TANG, Q., MEYER, C. A., ZHANG, Y. & LIU, X. S. (2013). Target analysis by integration of transcriptome and ChIP-seq data with BETA [Number : 12]. *Nature Protocols*, 8(12), 2502-2515. <https://doi.org/10.1038/nprot.2013.150> (cf. p. 123)
- WARR, M. R., BINNEWIES, M., FLACH, J., REYNAUD, D., GARG, T., MALHOTRA, R., DEBNATH, J. & PASSEGUÉ, E. (2013). FOXO3A directs a protective autophagy program in haematopoietic stem cells [Number : 7437]. *Nature*, 494(7437), 323-327. <https://doi.org/10.1038/nature11895> (cf. p. 19)
- WATCHAM, S., KUCINSKI, I. & GOTTGENS, B. (2019). New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing [Number : 13]. *Blood*, 133(13), 1415-1426. <https://doi.org/10.1182/blood-2018-08-835355> (cf. p. 32, 33, 39, 42, 44, 45, 47, 157)
- WEINREB, C., RODRIGUEZ-FRATICELLI, A., CAMARGO, F. D. & KLEIN, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation [Number : 6479]. *Science (New York, N.Y.)*, 367(6479). <https://doi.org/10.1126/science.aaw3381> (cf. p. 42, 47, 158)
- WHICHARD, Z. L., SARKAR, C. A., KIMMEL, M. & COREY, S. J. (2010). Hematopoiesis and its disorders : a systems biology approach. *Blood*, 115(12), 2339-2347. <https://doi.org/10.1182/blood-2009-08-215798> (cf. p. 52, 62)
- WIEDEMANN, D. (1991). A computation of the eighth Dedekind number. *Order*, 8(1), 5-6. <https://doi.org/10.1007/BF00385808> (cf. p. 122)
- WILSON, A., LAURENTI, E., OSER, G., van der WATH, R. C., BLANCO-BOSE, W., JAWORSKI, M., OFFNER, S., DUNANT, C. F., ESHKIND, L., BOCKAMP, E., LIÓ, P., MACDONALD, H. R. & TRUMPP, A. (2008). Hematopoietic Stem Cells Reversibly Switch from Dormancy to

- Self-Renewal during Homeostasis and Repair [Number : 6]. *Cell*, 135(6), 1118-1129. <https://doi.org/10.1016/j.cell.2008.10.048> (cf. p. 18, 20)
- WILSON, A., MURPHY, M. J., OSKARSSON, T., KALOULIS, K., BETTESS, M. D., OSER, G. M., PASCHE, A.-C., KNABENHANS, C., MACDONALD, H. R. & TRUMPP, A. (2004). c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation [Number : 22]. *Genes & Development*, 18(22), 2747-2763. <https://doi.org/10.1101/gad.313104> (cf. p. 22, 158)
- WILSON, N. K., KENT, D. G., BUETTNER, F., SHEHATA, M., MACAULAY, I. C., CALERO-NIETO, F. J., SÁNCHEZ CASTILLO, M., OEDEKOVEN, C. A., DIAMANTI, E., SCHULTE, R., PONTING, C. P., VOET, T., CALDAS, C., STINGL, J., GREEN, A. R., THEIS, F. J. & GÖTTGENS, B. (2015). Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations [Number : 6]. *Cell Stem Cell*, 16(6), 712-724. <https://doi.org/10.1016/j.stem.2015.04.004> (cf. p. 39, 42)
- WOLF, F. A., ANGERER, P. & THEIS, F. J. (2018). SCANPY : large-scale single-cell gene expression data analysis [Number : 1]. *Genome Biology*, 19(1), 15. <https://doi.org/10.1186/s13059-017-1382-0> (cf. p. 33, 40)
- YAN, X., XIONG, X. & CHEN, Y.-G. (2018). Feedback regulation of TGF- signaling. *Acta Biologica et Biophysica Sinica*, 50(1), 37-50. <https://doi.org/10.1093/abbs/gmx129> (cf. p. 159)
- YANG, J., TANAKA, Y., SEAY, M., LI, Z., JIN, J., GARMIRE, L. X., ZHU, X., TAYLOR, A., LI, W., EUSKIRCHEN, G., HALENE, S., KLUGER, Y., SNYDER, M. P., PARK, I.-H., PAN, X. & WEISSMAN, S. M. (2017). Single cell transcriptomics reveals unanticipated features of early hematopoietic precursors [Number : 3]. *Nucleic Acids Research*, 45(3), 1281-1296. <https://doi.org/10.1093/nar/gkw1214> (cf. p. 39, 42)
- YOUNG, K., BORIKAR, S., BELL, R., KUFFLER, L., PHILIP, V. & TROWBRIDGE, J. J. (2016). Progressive alterations in multipotent hematopoietic progenitors underlie lymphoid cell loss in aging [Number : 11]. *Journal of Experimental Medicine*, 213(11), 2259-2267. <https://doi.org/10.1084/jem.20160168> (cf. p. 49, 74, 76, 117, 161)
- ZHANG, L., MACK, R., BRESLIN, P. & ZHANG, J. (2020). Molecular and cellular mechanisms of aging in hematopoietic stem cells and their niches [Number : 1]. *Journal of Hematology & Oncology*, 13(1), 1-22. <https://doi.org/10.1186/s13045-020-00994-z> (cf. p. 26, 30)
- ZHAO, M., PERRY, J. M., MARSHALL, H., VENKATRAMAN, A., QIAN, P., HE, X. C., AHAMED, J. & LI, L. (2014). Megakaryocytes maintain homeostatic quiescence and promote post-injury regeneration of hematopoietic stem cells [Number : 11]. *Nature Medicine*, 20(11), 1321-1326. <https://doi.org/10.1038/nm.3706> (cf. p. 23, 159)