Étude à l'échelle unicellulaire du compartiment des cellules souches et progénitrices des syndromes myélodysplasiques

Charles Dussiau

Université de Paris

École doctorale Hématologie, Oncologie et Biothérapies (HOB, ED 561)

Équipe « Hématopoïèse normale et pathologique »
Département Développement, Reproduction et Cancer
Institut Cochin, INSERM U1016, Paris

Étude à l'échelle unicellulaire du compartiment des cellules souches et progénitrices des syndromes myélodysplasiques

Par **Charles Dussiau**

Thèse de doctorat en Hématologie

Dirigée par le Pr Olivier Kosmider

Présentée et soutenue publiquement le 15 octobre 2020

Devant un jury composé de :

Dr Françoise Pflumio : Présidente, Université Paris Sud

Dr Camille Lobry : Rapporteur, Université Paris-Saclay

Dr Pierre Milpied : Rapporteur, Université Aix Marseille

Pr Valérie Bardet : Examinatrice, Université Versailles Saint-Quentin

Dr Diana Passaro : Examinatrice, Université de Paris

Dr Nathalie Droin : Examinatrice, Université Paris-Saclay

Dr Olivier Gandrillon : Examinateur, Université Lyon1

Remerciements

En premier lieu, je tiens à remercier le Pr Olivier Kosmider de m'avoir encadré tout au long de cette thèse. Je te remercie d'avoir eu l'idée de démarrer cette aventure scRNA-Seq. Merci également pour ta confiance et pour m'avoir permis de me lancer dans la bioinformatique.

Merci au Pr Michaela Fontenay de m'avoir accueilli au laboratoire d'hématologie de l'hôpital Cochin à la sortie de mon internat. Merci de m'avoir permis de réaliser ce travail de thèse au sein de votre équipe.

Je tiens également à remercier les membres du jury :
Merci au Dr Camille Lobry et au Dr Pierre Milpied d'avoir accepté d'être rapporteur et d'avoir pris le temps d'évaluer ce travail.
Merci au Dr Françoise Pflumio d'avoir accepté de présider ce jury. Je te remercie également pour ta bienveillance et pour toutes les discussions scientifiques fructueuses que nous avons pu avoir au cours de ma thèse. Je suis très heureux d'avoir eu l'opportunité de travailler avec toi.
Merci au Pr Valérie Bardet, d'avoir accepté de juger ce travail. Je te remercie pour ta pédagogie et les connaissances que tu m'as transmises en hématologie.
Merci au Dr Diana Passaro et au Dr Nathalie Droin d'avoir accepté d'être examinateur de cette thèse.
Merci au Dr Olivier Gandrillon d'avoir accepté de faire partie de ce Jury. Je te remercie pour m'avoir ouvert les yeux sur le domaine de recherche passionnant qu'est la biologie des systèmes. Merci pour tout ce que tu as su me transmettre dans ce domaine et merci pour ton aide précieuse dans ce travail de thèse.

Merci au Dr Pierre Sujobert, je te remercie pour ton aide dans ce travail et pour les nombreuses discussions scientifiques passionnantes que nous avons eu. C'est un plaisir de collaborer avec toi.

Merci au Dr Nicolas Chapuis, pour ton aide au cours de cette thèse. Je te remercie également pour ta disponibilité et tout ce que tu m'as transmis en cytométrie.

Merci au Dr Marie Laure Arcangeli, pour ton aide en cytométrie, pour ta participation à mes comités de suivi de thèse, et pour m'avoir encouragé tout au long de ce travail. Je suis très heureux d'avoir pu travailler avec toi.

Merci au Dr Raphaël Itzykson, pour avoir participé à mes comités de suivi de thèse. Merci pour ton regard critique qui m'a permis d'améliorer ce travail.

Merci au Dr Philippe Asquier, merci tonton d'avoir eu la patience de m'envoyer ces échantillons de têtes fémorales, sans lesquels ce travail n'aurait pas été possible.

Merci à Laïla Zaroili, pour ton aide précieuse dans la mise au point des protocoles expérimentaux, et ton aide lors des manips critiques tout au long de ma thèse. Merci pour ton expertise, ta rigueur scientifique, et ta bonne humeur.

Merci à Clotilde Bravetti, pour ton aide précieuse dans ce travail. Je te remercie pour tes compétences, ton calme et ta concentration lors des manips scRNA-Seq.

Merci à Agathe Boussaroque, pour avoir eu le courage de te lancer avec moi dans R, Python, linux et autres sujets qui nous paraissaient tellement obscure au début de ton master 2. Je te remercie pour les nombreux packages installés et les nombreuses lignes de codes et scripts écrits.

Merci au Dr Camille Knosp, pour ton aide et tes compétences en cytométrie. Je te remercie pour nos discussions cytométriques, ainsi que pour ta bonne humeur contagieuse.

Merci à Amandine Houvert, pour ton aide dans la prise en charge des échantillons de routine.

Je souhaite également remercier les membres statutaires de l'équipe hématopoïèse normale et pathologique : Merci au Pr Didier Bouscary, au Dr Patrick Mayeux, au Dr Evelyne Lauret, au Dr Isabelle Dusanter pour leur avis critique tout au long de ma thèse.

Je remercie également tous les autres membres de l'équipe et notamment ceux avec qui j'ai partagé l'open space et autre (basket, shoot contest, NBA playground…) : Rudy, Ismael, Zubaidan, Justine, Salomé, Sabrina, Fetta, Maria-Lilia, Eric, Tony, Natacha.

Je remercie également l'équipe de cytométrie du laboratoire d'hématologie : Laurence, Catherine, Loé et Bruno, je vous remercie pour vos compétences et votre bonne humeur de tous les instants, ça a été un grand plaisir de travailler avec vous toutes ces années. Merci également de m'avoir prêté le Navios à des horaires raisonnables.

Je remercie aussi l'équipe de biologie moléculaire du laboratoire d'hématologie : Christine, Marlène, Angélique, Graziella, Christophe. Merci pour vos compétences, votre disponibilité, votre aide et votre gentillesse.

Merci également à l'équipe de biologiste du laboratoire d'hématologie cellulaire de Cochin : Alexa, Anna, Fatou, Carole, Anne-Sophie, Sylvain, Chloé, Loria. Ce fut un plaisir de travailler à vos côtés.

Merci à ma famille, à mes parents et à mon frère pour leur soutien sans faille depuis le début de mes études.

Merci à mes amis pour m'avoir permis de décrocher de la thèse lors de la période difficile de l'écriture. Merci notamment à Nabih pour s'être embarqué dans la même galère.

Enfin merci Stéphanie d'être à mes côtés et de m'avoir soutenu pendant ces 3 ans.

Table des matières

Liste des abréviations

ADN : Acide Désoxyribonucléique

ARN : Acide Ribonucléique

ARNm : Acide Ribonucléique messager

CAR : CXCL12 Abundant Reticular cells

CD : Cluster of Differenciation

CMN : Cellules Mononuclées

CMP : Common Myeloid Progenitor

CSH : Cellule Souche Hématopoïétique

CSL : Cellule Souche Leucémique

CSM : Cellule Stromale Mésenchymateuse

EPO : Erythropoïétine

ETP : Early-T-Precursor

FA : ForceAtlas2

G-CSF : Granulocyte Colony Stimulating Factor

GEM : Gel Bead in Emulsion

GM-CSF : Granulocyte Macrophage Colony Stimulating Factor

GMP : Granulocyte Macrophage Progenitor

HSPC : Hematopoïetic Stem and Progenitor Cells

IPSS-R : Revised-International Pronostic Scoring System

LAM : Leucémie Aigue Myéloïde

Lin : Lineage

LMC : Leucémie Myéloïde Chronique

LMMC : Leucémie Myélomonocytaire Chronique

LMPP : Lymphomyeloid Primed Progenitor

MAST : Model-based Analysis of Single Cell Transcriptomics

MEP : Megacaryocyte Erythroid Progenitor

miARN : micro Acide Ribonucléique

MLP : MultiLymphoid Progenitor

MPP : Multipotent Progenitor

NGS : Next Generation Sequencing

NSG : Non Obsese Diabetic Severe Combined Immunodeficient IL-2Rγ-null

pb : paire de base

PCR : Polymerase Chain Reaction

scRNA-Seq : Single-Cell RNA-Seq

SMD : Syndrome Myélodysplasique

SMP : Syndrome Myéloprolifératif

STAR : Spliced Transcripts Alignment to a Reference

TPO : Thrombopoïétine

UMAP : Uniform Manifold Approximation and Projection

UMI : Unique Molecular Identifier

Table des Figures

Introduction

1. Single-Cell RNA-Seq (Transcriptome à l'échelle unicellulaire)

1.1. Historique

La première expérience de transcriptome à l'échelle unicellulaire (scRNA-Seq) a été publiée en 2009[1]. A cette époque les techniques de séquençage de nouvelle génération (NGS) nécessitaient une quantité d'ARN de l'ordre du microgramme ce qui équivaut à environ 100000 cellules. Cette quantité de cellules représentait un facteur limitant pour l'étude de processus tels que l'embryogénèse où les étapes précoces ne sont constituées que de quelques cellules.

Les technologies à haut débit permettent désormais l'analyse de centaines de milliers de cellules en parallèles offrant ainsi une vision précise de l'hétérogénéité des cellules individuelles au sein d'une population[2]. Cette technologie a permis d'entrevoir la réponse à certaines questions jusqu'alors inaccessibles[3] :

- Quelle sont les sous populations cellulaires présentes au sein des tissus ?
- Existe-t-il de nouvelles populations jusqu'alors inconnues et inexplorées ?
- Comment se font les processus de développement et de transitions entre deux états cellulaires ?
- Qu'en est-il de la stochasticité et de la spécificité allélique de l'expression des gènes ?
- Comment les réseaux de régulation de gènes sont-ils construits ?

Suite à cette première expérience, de nombreuses technologies ont été mises au point afin d'améliorer la qualité et la quantité de cellules analysables.

1.2. Technologie

Pour séquencer les ARN messagers (ARNm) d'une seule cellule il faut surmonter deux défis. Premièrement la capture de la cellule unique. Deuxièmement l'amplification d'une quantité minime d'ARNm contenue dans chaque cellule. Toutes les technologies de scRNA-Seq suivent une stratégie de base similaire. Tout d'abord une seule cellule est capturée, lysée,

puis une transcription inverse est effectuée pour obtenir l'ADN complémentaire (ADNc) à l'ARNm. Par la suite les quantités infimes d'ADNc obtenues sont amplifiées par PCR. Enfin l'ADNc amplifié est utilisé pour la préparation des librairies destinées au séquençage[3]. De très nombreuses techniques ont été développées permettant le séquençage du transcriptome à l'échelle unicellulaire et il serait impossible de toutes les décrire. Tout au long de ce chapitre, je vais donc présenter et détailler la technologie Chromium développée par 10x Genomics, technologie qui a été utilisée au cours des expériences de scRNA-Seq présentées dans ce travail de thèse.

1.2.1. Capture d'une cellule unique.

La technologie Chromium développée par 10x Genomics utilise un circuit microfluidique qui permet de séparer individuellement les cellules. Cette séparation se fait en mélangeant ensemble les trois principaux composants qui sont déposés dans une puce microfluidique. Ces composants sont les perles de gel barcodées (Gel Beads), les cellules d'intérêt en solution avec les réactifs, et l'huile de séparation. Chaque composant s'écoule dans un canal de la puce microfluidique, pour finir par se mélanger et former une gouttelette lipidique contenant idéalement une cellule, une perle de gel barcodée et les réactifs en solution. Chaque gouttelette est ainsi nommée GEM pour « Gel Bead in Emulsion » **(Figure 1).**

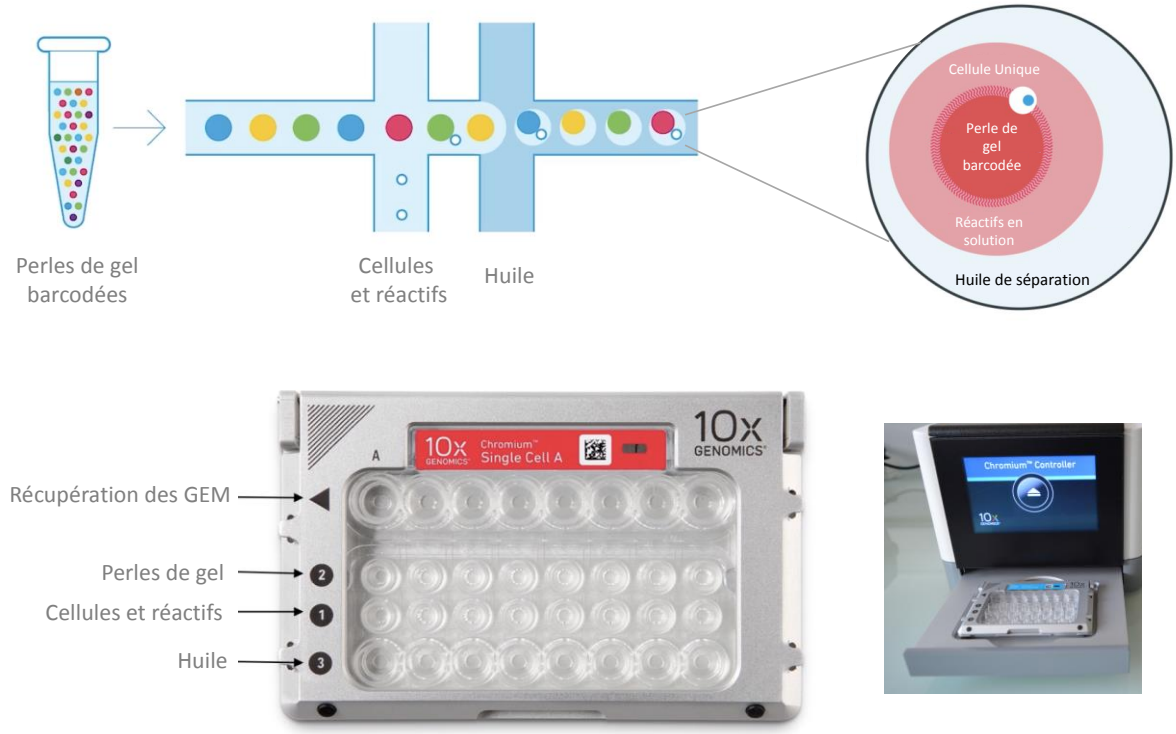


Figure 1 : Formation des GEM (Gel Bead in Emulsion).

Principe de la technologie permettant l'encapsulation des cellules dans les gouttelettes lipidiques pour former des GEM[4].

La perle de gel est barcodée, signifiant que des oligonucléotides identifiables après séquençage sont accrochés à sa surface. Ces oligonucléotides sont constitués des éléments suivants **(Figure 2)** :

- Le R1 ou read 1, cette séquence est identique sur tous les oligonucléotides de chaque perle de gel. Elle va servir à fixer l'amorce destinée à l'amplification de l'ADN complémentaire obtenus dans les étapes ultérieures. La séquence va également permettre lors du séquençage la fixation d'une amorce (cette partie sera détaillée dans la suite du manuscrit).
- Le 10x Barcode : cette séquence est caractéristique d'une perle de gel, et permet de marquer les ARN messagers de la cellule encapsulée dans la même gouttelette. Ainsi, tous les ARN messagers provenant d'une même cellule seront marqués par le même barcode.

- L'UMI : pour « unique molecular identifier », cette séquence est spécifique de chaque molécule d'ARN messager, elle permet de limiter les biais lors de l'amplification de l'ADNc. En effet cela permet de détecter et de normaliser si un ADNc provenant d'une molécule d'ARN messager est amplifié anormalement lors de la PCR comparativement aux autres.
- Le Poly(dT)VN : séquence composée de « T » répétés permettant l'hybridation sur la queue PolyA des ARN messagers.

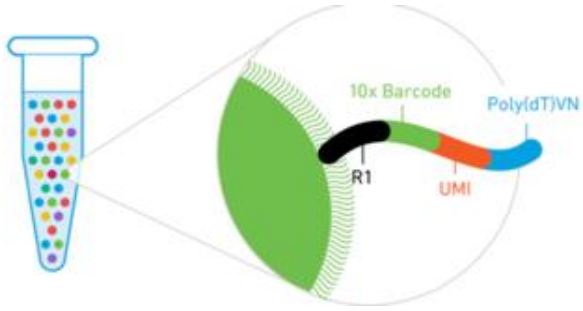


Figure 2 : Composition de la perle de gel barcodée.
Sur chaque perle de gel sont accrochés les oligonucléotides contenant les séquences nécessaires au bon déroulement du processus[4].

1.2.2. Lyse, transcription inverse et amplification

Au sein de chaque GEM, dans l'environnement aqueux, la cellule est lysée, la perle de gel se dissout et les oligonucléotides fixés ainsi libérés vont pouvoir s'hybrider par leur poly(dT)VN à la queue PolyA des ARN messagers de la cellule. Les GEM sont ensuite incubés, la réaction de transcription inverse peut se dérouler afin de générer l'ADNc à partir des transcrits d'ARNm. Une enzyme, la transcriptase inverse (RT enzyme) va compléter la synthèse du brin complémentaire à l'ARNm à partir du polyDT de l'oligonucléotide barcodé et rajouter des cytosines à l'extrémité 5' du transcrit. Ces cytosines vont permettre ensuite l'amorçage du « Switch 0ligo », et la partie complémentaire au « Switch Oligo » va être synthétisée par la transcriptase inverse **(Figure 3)**.

A l'intérieur de chaque GEM

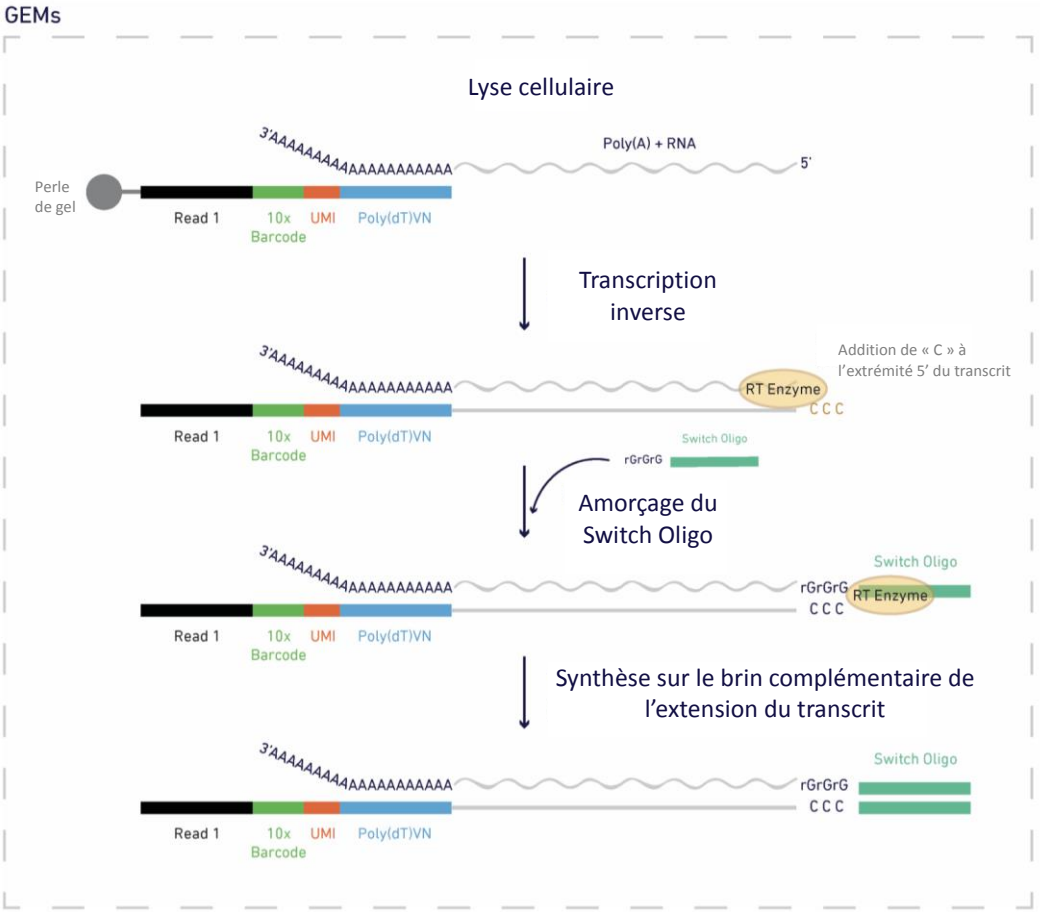


Figure 3 : Transcription inverse à l'intérieur des GEM.
Après lyse cellulaire et dissolution de la perle de gel, la transcription inverse peut se dérouler[4].

L'émulsion contenant les GEM est ensuite « cassée », regroupant ainsi les molécules d'ADNc de chaque cellule. Les ADNc sont ensuite amplifiés par PCR à l'aide d'amorces se fixant sur les extrémités correspondant aux séquences read 1 et « Switch Oligo ». Une étape de fragmentation permet d'optimiser la taille de l'insert d'ADNc avant la construction de la librairie. La séquence Read 2 est ajoutée par ligation à la suite de l'insert. Ensuite, les adaptateurs P5, P7 ainsi que l'index spécifique à chaque échantillon sont ajoutés par PCR.
La librairie finale est constituée de fragments contenant les séquences P5, P7, Read1 et Read 2 nécessaire au séquençage par la technique Illumina (détaillée dans la suite du manuscrit). Chaque fragment contient également le code barre 10x spécifique de chaque cellule, l'UMI

spécifique de chaque molécule d'ARNm, et la séquence d'ADNc d'interêt qui y est insérée[4] **(Figure 4)**.

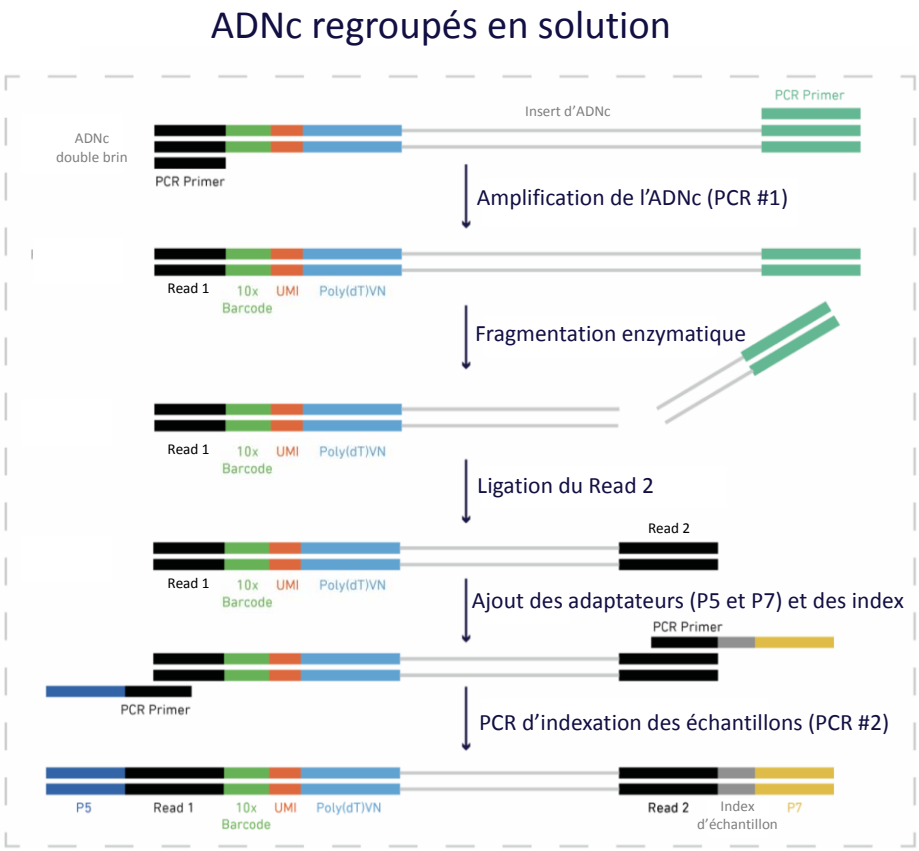


Figure 4 : Amplification de l'ADNc et génération des librairies.
Après amplification de l'ADNc par PCR, la fragmentation enzymatique permet la sélection de la taille des inserts. Une deuxième PCR permet l'ajout des adaptateurs P5 et P7, ainsi que des index spécifiques aux échantillons[4].

1.2.3. Séquençage des librairies

Dans la stratégie prise en exemple, les librairies obtenues sont séquencées par technologie de séquençage nouvelle génération Illumina. Les adaptateurs P5 et P7 permettent la liaison à la Flowcell, la séquence read 1 sur laquelle se fixe l'amorce complémentaire permet le séquençage du barcode 10x (16pb) et de l'UMI (12pb) soit 28 cycles, la séquence read 2 sur laquelle se fixe l'amorce complémentaire (sens 5' vers 3') permet le séquençage de l'index I7 (8pb) soit 8 cycles, et la fixation de l'amorce complémentaire (sens 3' vers 5') permet le séquençage de l'insert d'ADNc (98pb) avec 98 cycles **(Figure 5).**

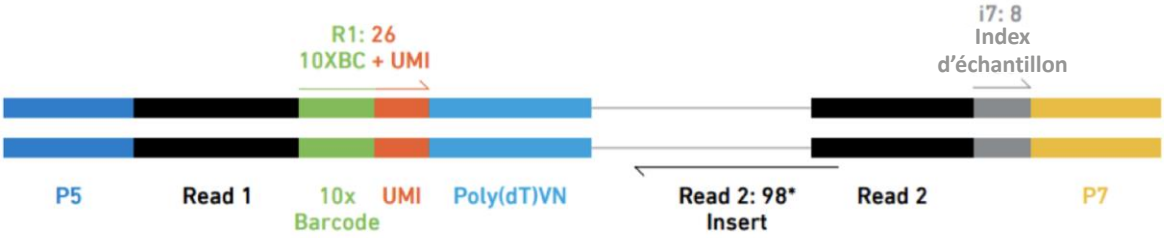


Figure 5 : Schéma d'un fragment de la librairie finale Chromium Single Cell.
Les fragments qui composent la libraire finale contiennent les adaptateurs P5 et P7 nécessaire à la fixation sur la Flowcell, et les séquences read1 et read2 qui vont permettre le séquençage[4].

La première partie du séquençage est appelée étape de « clustering » et permet de générer les clusters. La Flowcell est une fine lame de verre composé de plusieurs lignes. Sur chaque ligne sont accrochés des fragments d'oligonucléotides complémentaire aux séquences adaptatrices P5 et P7. Dans un premier temps, Les fragments de la librairie finale Chromium vont s'hybrider sur la flowcell grâce à l'adaptateur P5. Une polymérase va synthétiser le brin complémentaire au fragment hybridé. La molécule double-brin est ensuite dissociée et le fragment original est lavé. Les brins restants sont ensuite amplifiés clonalement grâce à la technique de « bridge amplification ». Dans ce processus, les brins se plient et l'adaptateur P7 se lie à l'oligonucléotide correspondant présent sur la Flowcell. La polymérase génère ensuite le brin complémentaire formant ainsi un pont double-brin. Le pont est alors dénaturé, il en résulte deux copies simple brin de la molécule qui sont attachées à la Flowcell. Le processus est alors répété de nombreuses fois et se produit simultanément pour des millions de clusters, entraînant une amplification clonale de tous les fragments. Après cette « bridge amplification », les brins Antisens sont déshybridés et lavés. Il ne reste alors sur la Flowcell que le brin sens. L'extrémité 3' est bloquée pour éviter les amorçage non voulu **(Figure 6).**

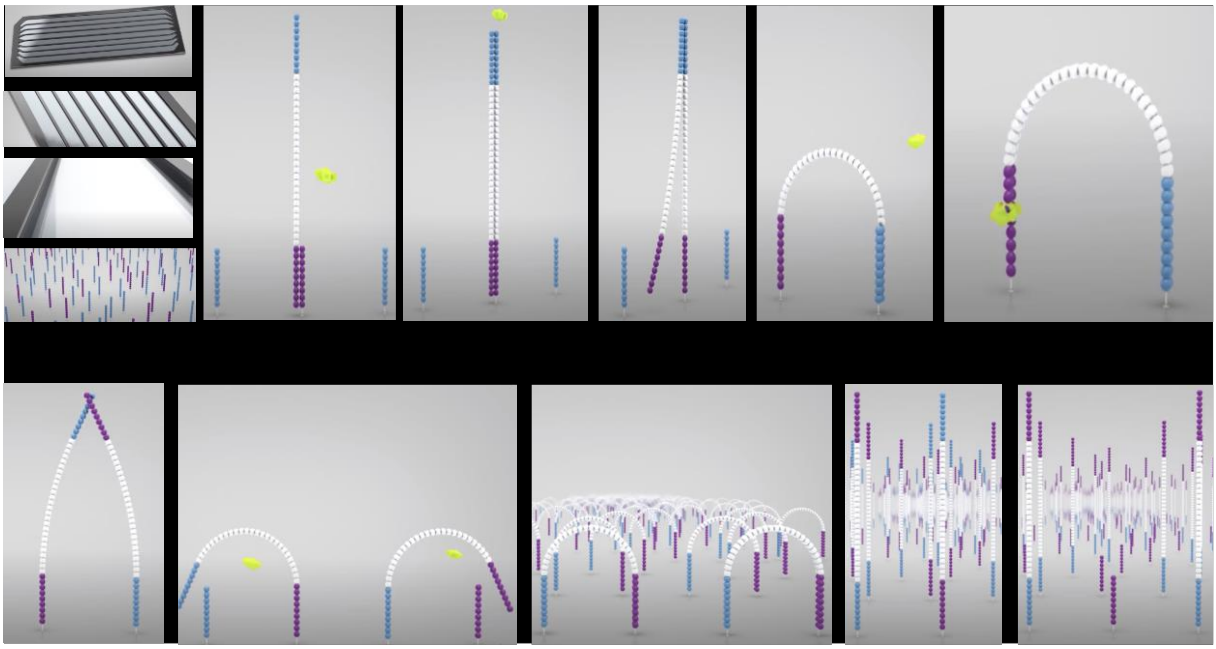


Figure 6 : Technologie Illumina : génération des clusters.
Les fragments de la librairie finale vont se fixer sur la Flowcell grâce aux séquences complémentaires aux adaptateurs P5 et P7 puis, seront amplifiés de manière clonale par la technique de « bridge amplification »[5].

Le séquençage commence par la fixation de l'amorce complémentaire à la région read 1, puis, à chaque cycle, des nucléotides marqués avec des fluorochromes vont entrer en compétition pour allonger le fragment. Seul un nucléotide est incorporé à la fois en fonction de sa complémentarité avec le nucléotide sur le fragment à séquencer. Après l'addition de chaque nucléotide, les clusters sont excités par une source lumineuse, un signal fluorescent caractéristique du nucléotide incorporé est émis. Ce processus est appelé « séquençage par synthèse ». Le nombre de cycles correspond à la longueur du read. La longueur d'onde d'émission ainsi que l'intensité du signal déterminent la nature de la base incorporée. Pour un cluster donné, tous les brins identiques sont lus simultanément. Des centaines de millions de clusters sont ainsi séquencés en parallèle. Après la fin du read 1, le produit du read est lavé. Lors de l'étape suivante, l'amorce se fixant à la région read 2 va s'hybrider et permettre le séquençage de l'index I7. Le read de séquençage de l'index I7 est généré de la même manière que le premier read (read1). Ensuite, le produit de ce read est lavé et l'extrémité 3' du fragment est déprotégé. Le fragment peut ainsi se plier et se lier au deuxième oligonucléotide de la flowcell, puis la polymérase forme le pont double brin. Les deux brins sont déshybridés et deviennent alors linéaires, les extrémités 3' sont bloquées. Le brin sens original est

déshybridé et lavé, il ne reste accroché à la Flowcell que le brin antisens. L'amorce Read 2 va se fixer sur la séquence correspondante et ainsi permettre le séquençage de l'insert (98pb). A la fin du séquençage, le produit du read est déshybridé et lavé. Ce processus génère des millions de reads qui seront ensuite alignés sur le génome de référence[5] **(Figure 7).**

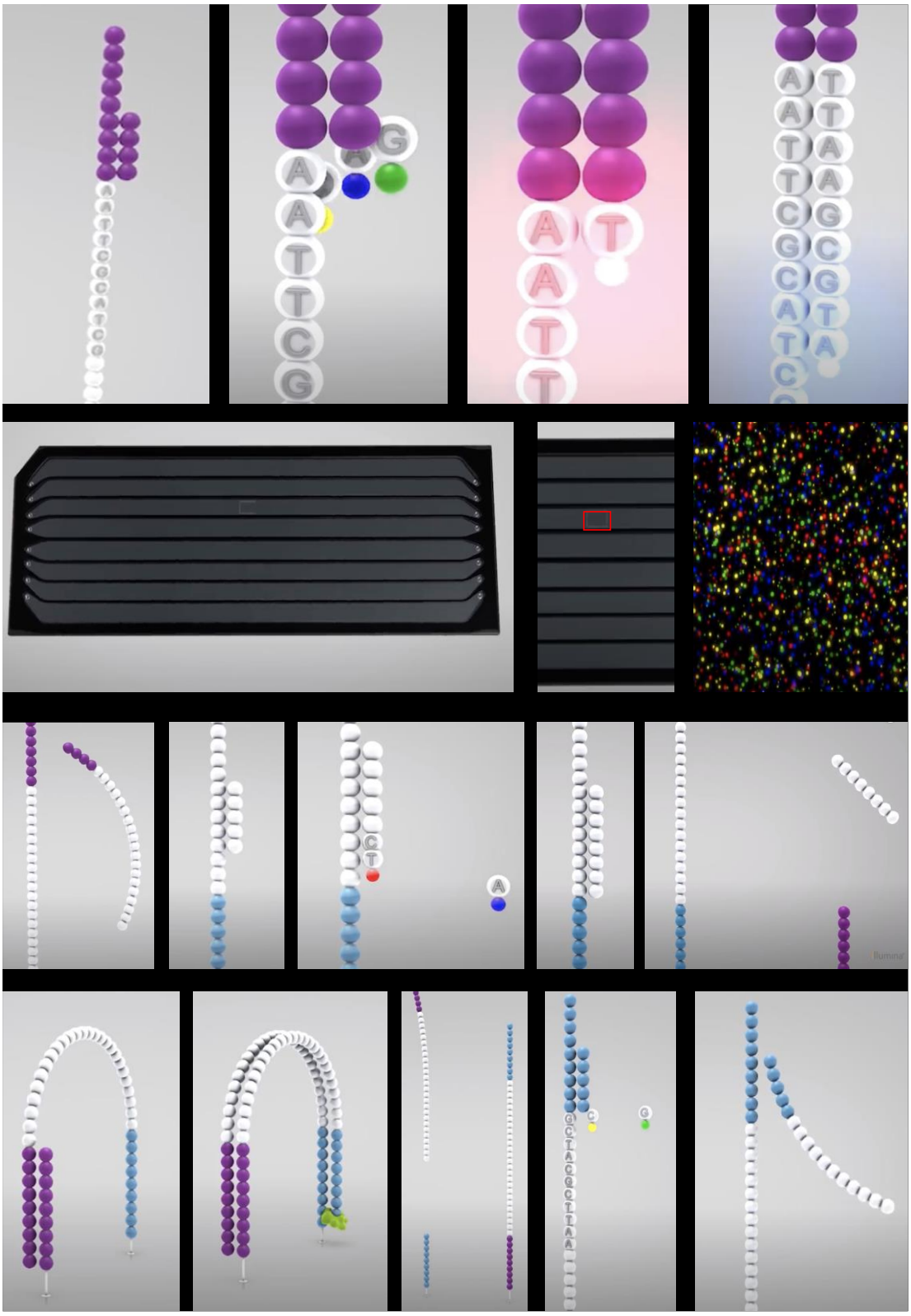


Figure 7 : Technologie Illumina : séquençage.

Le séquençage se fait par la technique du « séquençage par synthèse ». En effet, à chaque cycle, le nucléotide qui s'incorpore émet une fluorescence caractéristique qui sera détectée par une caméra. Le séquençage se fait de manière concomitante à la synthèse du brin complémentaire[5].

1.3. Analyse bioinformatique

L'analyse bioinformatique des données de scRNA-Seq est un domaine en plein essor, il existe actuellement plus de 600 outils bioinformatiques référencés, et de nouveaux outils sont publiés presque chaque semaine[6] **(Figure 8)**. Ce chapitre est volontairement restreint aux méthodes sélectionnées et utilisées au cours de mon travail de thèse.

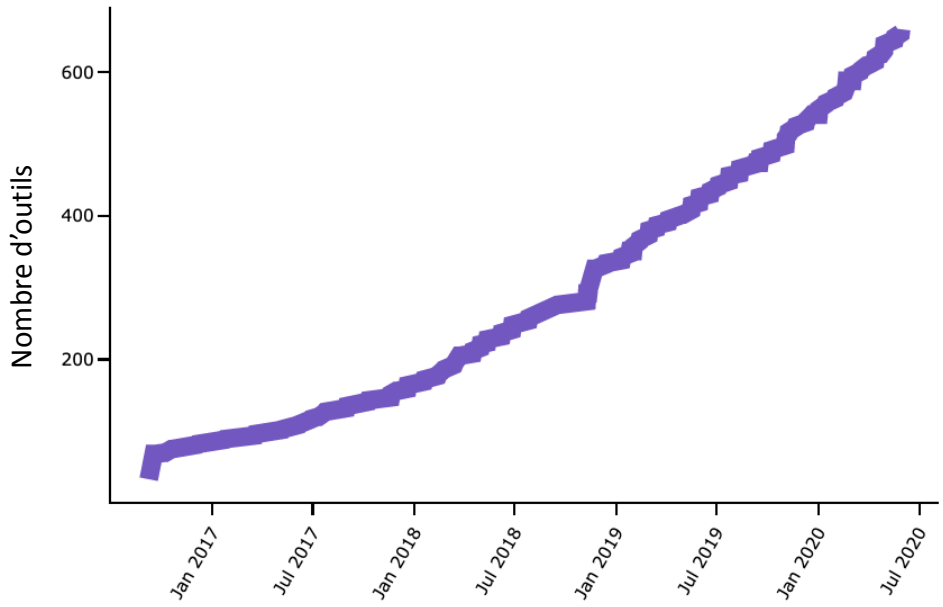


Figure 8 : Outils bioinformatiques disponibles pour l'analyse de données de scRNA-Seq [6].
Entre Janvier 2017 et Juillet 2020, l'évolution du nombre d'outils disponibles suit une courbe d'allure exponentielle.

1.3.1. Construction de la matrice gènes-cellules.

La première étape de toute analyse des données de séquençage consiste à convertir les fichiers renvoyés par le séquenceur dans un format adapté à l'analyse. Pour les données de scRNA-Seq produites à l'aide de la plate-forme Chromium de chez 10x Genomics, l'outil standard pour accomplir cette tâche est le logiciel Cell Ranger développé par 10x Genomics. Dans un premier temps, Cell Ranger permet la conversion des fichiers BCL issus du séquençage Illumina en fichiers de séquence FASTQ classiquement utilisés dans l'analyse de données RNA-Seq et scRNA-Seq.

A partir des fichiers FASTQ (qui contiennent donc les séquences Read1, Read2 et Index I7), Cell Ranger va ensuite construire la matrice gènes cellules en plusieurs étapes. Dans un premier temps, Cell Ranger va extraire les Barcode 10x et les UMI. Il va ensuite aligner les reads sur le génome de référence à l'aide de STAR (Spliced Transcripts Alignement to a Reference)[7], ce qui permet d'attribuer le read au gène d'où provient le transcrit capturé. Les reads sont ensuite regroupés par cellules grâce aux barcodes. Puis Cell Ranger va compter les UMI pour chaque gène dans chaque cellule et va éliminer les duplicats **(Figure 9).** On obtient alors une matrice gènes-barcodes non filtrée qui correspond à un préalable indispensable pour la création de la matrice gènes-cellules.

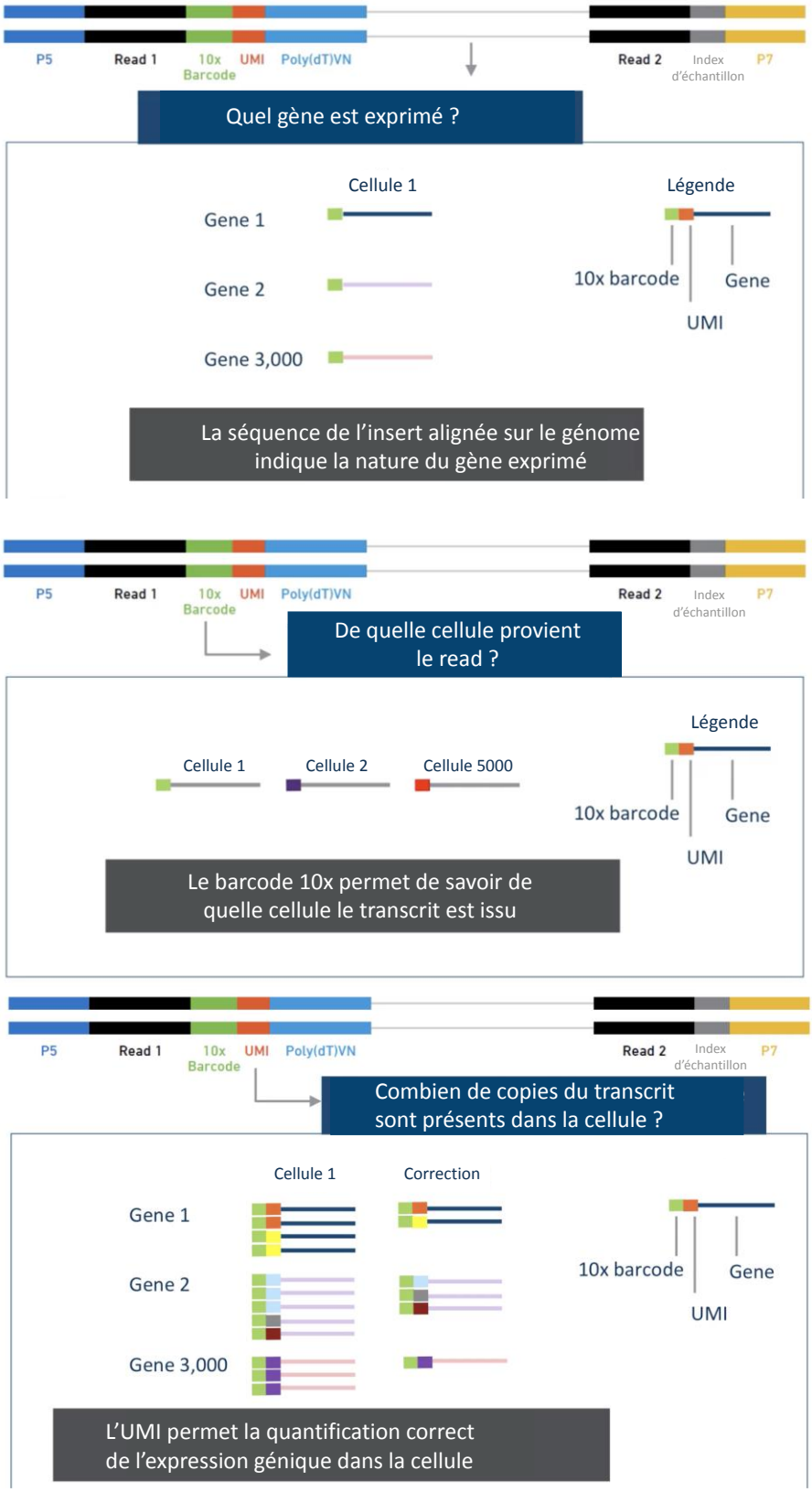


Figure 9 : Construction par Cell Ranger de la matrice gènes barcodes.

L'alignement du read sur le génome de référence permet de déterminer la nature du gène correspondant au transcrit capturé, le barcode 10x permet de savoir de quelle cellule provient le transcrit, et l'UMI permet la quantification correcte de l'expression génique[4].

De nombreuses gouttelettes sont produites au cours du processus de capture cellulaire, mais la plupart d'entre elles ne contiennent pas de cellules et tous les reads qui leur sont associées seront issus de débris cellulaires, ou de l'ARN libre présent en solution. Il faut donc uniquement sélectionner les barcodes qui correspondent à l'encapsulation d'une cellule. Pour sélectionner les barcodes correspondant à des cellules encapsulées, on utilise un graphique à échelle logarithmique qui représente les barcodes détectés pendant le séquençage, rangés dans l'ordre décroissant du nombre d'UMI associés. Le nombre d'UMI associé à chaque barcode est utilisé par Cell Ranger pour déterminer quels barcodes correspondent à une cellule encapsulée. Les barcodes correspondant aux cellules encapsulées sont associés à un nombre d'UMI plus élevé que les barcodes correspondant aux gouttelettes vides ou contenant seulement de l'ARN issu de débris cellulaires ou libre en solution **(Figure 10).** Cette sélection permet d'obtenir une matrices gènes-barcodes filtrée, qui correspond à la matrice gènes-cellules utilisée pour les analyses ultérieures).

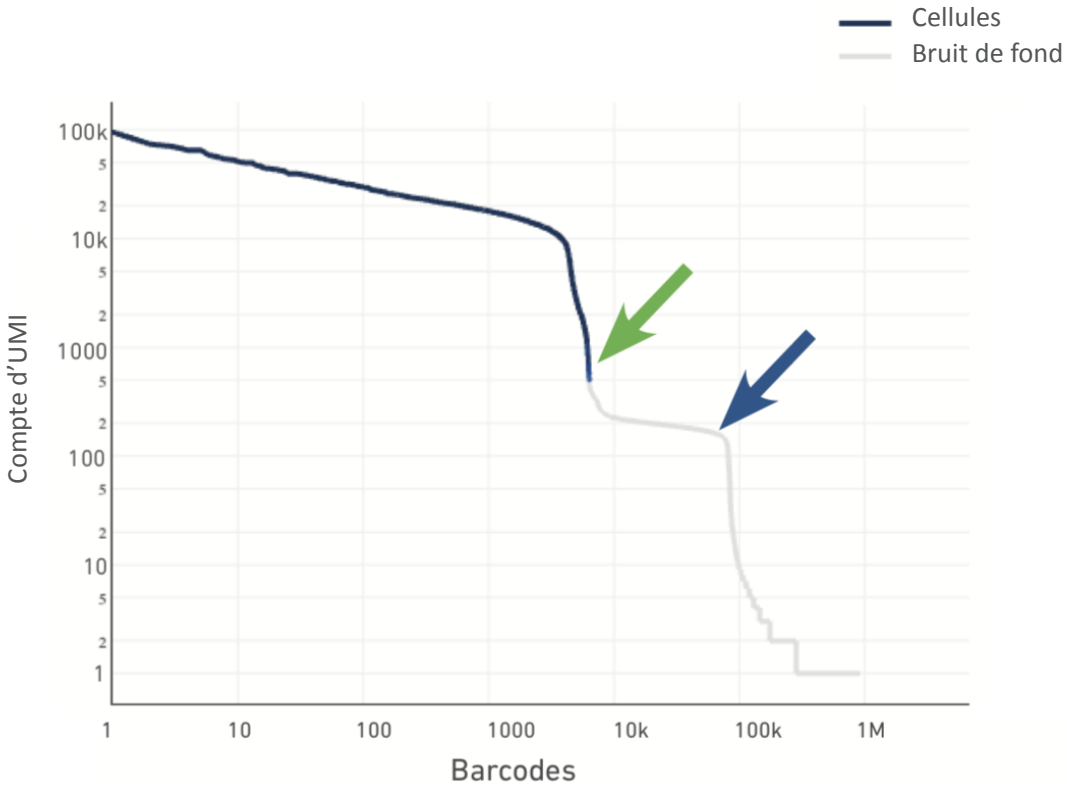


Figure 10 : Graphique des barcodes rangés selon le compte d'UMI.
Une chute abrupte indique une bonne séparation entre les barcodes associés aux cellules et les barcodes associés aux gouttelettes vides. Un tracé idéal présente une forme distinctive appelée « falaise et genou ». La transition bleu-gris (flèche verte) est appelée falaise, la transition marquée par la flèche bleue est appelée genou[4].

1.3.2. Contrôle qualité

L'étape précédente a permis de sélectionner les barcodes spécifiques des gouttelettes qui contenaient des cellules. L'étape de contrôle qualité permet de s'assurer que les barcodes sélectionnés correspondent à des cellules qui étaient vivantes lors de l'encapsulation. Pour rappel, un barcode correspond à une cellule.

Le contrôle qualité est effectué à partir de la mesure de trois variables : le nombre d'UMI par barcode, le nombre de gènes par barcode, ainsi que le pourcentage de gènes mitochondriaux par rapport aux gènes détéctés[8]. On examine la distribution de ces trois variables pour éliminer les barcodes qui présentent des valeurs aberrantes. En effet, les barcodes présentant peu d'UMI, peu de gènes détectés et un important pourcentage de gènes mitochondriaux indique que l'ARNm cytoplasmique s'est échappé au travers de la cellule encapsulée, que la membrane cellulaire est endommagée et que seul l'ARN mitochondrial a pu être conservé lors de l'encapsulation. La cellule correspondant à ce barcode était donc en train de mourir. Par ailleurs, les barcodes présentant un nombre très élevé de gènes et d'UMI détectés par rapport aux autres barcodes ont une grande chance de correspondre à des doublets, c'est-à-dire des gouttelettes ayant encapsulé plus d'une cellule. Dans certains cas, le fait pour un barcode de présenter des valeurs aberrantes pour l'un des trois paramètres évalués peut avoir une signification biologique. Par exemple des barcodes présentant une importante fraction de gènes mitochondriaux peuvent correspondre à des cellules parfaitement viables impliquées dans un processus cellulaire spécifique. Les barcodes présentant peu d'UMI et de gènes détectés peuvent également correspondre à des cellules quiescentes. A l'inverse, les barcodes qui ont beaucoup d'UMI et de gènes détectés peuvent correspondre à des cellules de grande taille. Il est donc important de ne pas analyser ces variables individuellement mais de les prendre en compte de manière simultanée. La filtration des barcodes doit se faire avec les seuils les plus permissifs possibles afin de ne pas éliminer de l'analyse des cellules importantes pour le tissu ou le processus biologique étudié.

En plus d'éliminer dans la matrice gènes-cellules les cellules mourantes et de mauvaise qualité, il peut également être nécessaire d'éliminer certains gènes. En effet, les matrices de comptage brut incluent souvent plus de 20 000 gènes. Ce nombre peut être considérablement

réduit en ne gardant que les gènes exprimés dans un nombre minimum de cellules afin d'éliminer ceux qui ne sont pas informatifs de l'hétérogénéité cellulaire. L'une des stratégies utilisées pour définir le nombre minimum de cellules dans lequel un gène s'exprime pour le garder dans la matrice, consiste à utiliser le nombre minimum de cellules susceptibles de former un cluster d'intérêt. Par exemple, éliminer les gènes exprimés dans moins de 20 cellules peut rendre difficile la détection de clusters composés de moins de 20 cellules. Le choix du seuil doit donc prendre en compte le nombre de cellules total du jeu de données, ainsi que les analyses prévues en aval.

Le contrôle qualité de la matrice gènes cellules est effectué pour garantir que la qualité des données est suffisante pour les analyses réalisées en aval. La qualité des données ne peut être déterminée *a priori*, elle est évaluée sur la performance des analyses réalisées en aval (par exemple sur l'annotation des clusters). Ainsi, il peut être nécessaire de modifier plusieurs fois les seuils des contrôles qualités après analyses des données. Il est donc préférable de commencer avec des seuils permissifs, et d'étudier les effets de ces seuils sur les analyses en aval avant de revenir à des seuils plus stricts. Cette approche est particulièrement pertinente dans l'analyse des jeux de données où les populations cellulaires analysées sont très hétérogènes[9].

1.3.3. Normalisation

Dans les données de scRNA-Seq, il existe des variations entre les cellules qui ne sont pas due à l'hétérogénéité biologique, mais liées à des contraintes techniques[10]. En effet, chaque compte dans la matrice gènes cellules représente pour une cellule la capture, la transcription inverse, l'amplification par PCR et le séquençage réussi d'une molécule d'ARN messager. Même si l'utilisation d'UMI supprime les variations techniques associées à la PCR, la profondeur de séquençage entre des cellules biologiquement identiques peut différer en raison de la variabilité inhérente à chacune de ces étapes. On ne peut donc pas comparer l'expression génique entre les cellules sur la base des données de comptage bruts. La normalisation résout ce problème en remettant à la même échelle les données de comptage pour obtenir des niveaux d'expression génique relatifs comparables entre les cellules[10]. Ce

problème déjà observé dans les données de RNA-Seq est exacerbé dans les données de scRNA-Seq par les nombreux zéros présents dans la matrice de comptage[11].

Parmis les nombreuses méthodes de normalisation développées pour le scRNA-Seq, nous avons fait le choix d'utiliser la méthode « SCtransform » qui est implémenté dans le package R Seurat[12]. Les auteurs de la méthode proposent que des données correctement normalisées doivent présenter les caractéristiques suivantes :

- le niveau d'expression normalisé d'un gène ne doit pas être corrélé avec la profondeur de séquençage d'une cellule. Les analyses en aval (réduction dimensionnelle, expression différentielle…) ne devraient pas non plus être influencées par la variation de la profondeur de séquençage.

- la variance d'un gène normalisé (entre les cellules) devrait refléter principalement l'hétérogénéité biologique, indépendamment de l'abondance du gène ou de la profondeur de séquençage. Par exemple, les gènes avec une variance élevée après normalisation devraient être exprimés de manière différentielle entre les types de cellules, tandis que les gènes de ménage devraient présenter une faible variance. De plus, la variance d'un gène doit être similaire lorsque l'on compare des cellules dont la profondeur de séquençage est différente.

Les auteurs ont ainsi développé une nouvelle approche statistique pour la modélisation, la normalisation et la stabilisation de la variance des données de comptage. Ils ont construit un modèle linéaire généralisé (GLM) pour chaque gène avec le compte d'UMI comme variable dépendante et la profondeur de séquençage comme variable explicative. En regroupant les informations entre les gènes ayant des abondances similaires, ils ont pu régulariser les estimations des paramètres et obtenir des modèles d'erreur reproductibles. Les résidus de leur « régression binomiale négative régularisée » représentent des valeurs de données effectivement normalisées qui ne sont plus influencées par les caractéristiques techniques, mais préservent l'hétérogénéité induite par des états biologiques distincts. Les résidus positifs pour un gène donné dans une cellule donnée indiquent que le nombre d'UMI observé est plus important que prévu étant donné l'expression moyenne du gène dans la population et la profondeur de séquençage cellulaire, tandis que les résidus négatifs indiquent l'inverse. Ces valeurs normalisées permettent les analyses couramment réalisées en aval,

telles que la réduction dimensionnelle et les tests d'expression différentielle, sans que les résultats obtenus ne soient influencés par la profondeur de séquençage cellulaire.

1.3.4. Réduction dimensionnelle et visualisation.

Dans un jeu de données de scRNA-Seq, même après l'étape de contrôle qualité, la matrice gènes-cellules contient encore beaucoup de gènes qui ne sont pas informatifs ce qui peut représenter jusqu'à 20000 dimensions. Pour alléger la charge de calcul des outils d'analyse en aval, réduire le bruit de fond et visualiser les données, il existe plusieurs étapes pour réduire la dimensionnalité d'un jeu de données.

La première étape de la réduction de la dimensionnalité est la sélection des gènes les plus « informatifs » de la variabilité (HGV pour Highly Variable Genes). Dans Seurat[13] par exemple, les gènes sont séparés en groupes selon leur moyenne d'expression, et les gènes ayant le rapport variance-moyenne le plus élevé dans chaque groupe sont sélectionnés comme gènes les plus informatifs.

Après la sélection des gènes les plus informatifs, la dimensionnalité de la matrice gènes cellules peut encore être réduite par des algorithmes spécialement dédiés à la réduction dimensionnelle **(Figure 11).** Ces algorithmes permettent d'intégrer la matrice d'expression gènes cellules dans un espace de faible dimension, conçu pour capturer la structure sous-jacente des données dans le moins de dimensions possible. Cette approche fonctionne car les données de scRNA-Seq sont intrinsèquement de faible dimension[14]. En d'autres termes, la variété biologique expliquée par les profils d'expression génique cellulaires peut être décrite par un nombre de dimensions bien moins grande que le nombre de gènes. La réduction de la dimensionnalité vise à définir ces dimensions.

L'analyse en composante principale (ACP)[15] est une méthode de réduction dimensionnelle linéaire qui permet d'extraire et de visualiser les informations importantes contenues dans la matrice gènes cellules. L'ACP synthétise cette information en seulement quelques nouvelles variables appelées composantes principales. Ces nouvelles variables correspondent à une combinaison linéaire des variables originelles. Le nombre de composantes principales est inférieur ou égal au nombre de variables d'origine. L'information contenue dans un jeu de données correspond à la variance ou l'inertie totale qu'il contient.

L'objectif de l'ACP est d'identifier les directions (axes principaux ou composantes principales) le long desquelles la variation des données est maximale. En pratique, l'ACP réduit les dimensions de la matrice gènes cellules à deux ou trois composantes principales, qui peuvent être visualisées graphiquement, en perdant le moins possible d'information. En général, l'ACP ne permet pas de capturer aussi bien la structure globale des données comme peut le faire les méthodes de réduction dimensionnelle non linéaires. L'ACP est cependant couramment utilisée comme une étape de prétraitement des données avant d'utiliser ces méthodes[9].

Les méthodes de réduction dimensionnelle non linéaires comme UMAP pour (Uniform Approximation and Projection method)[16] et FA (ForceAtlas2)[17] utilisent des algorithmes mathématiques complexes mais faciles à mettre en œuvre, rapides, et permettent les représentations en deux dimension les plus fidèles à la réalité biologique. Ces deux méthodes sont les plus utilisées pour représenter graphiquement des données de scRNA-Seq en deux dimensions.

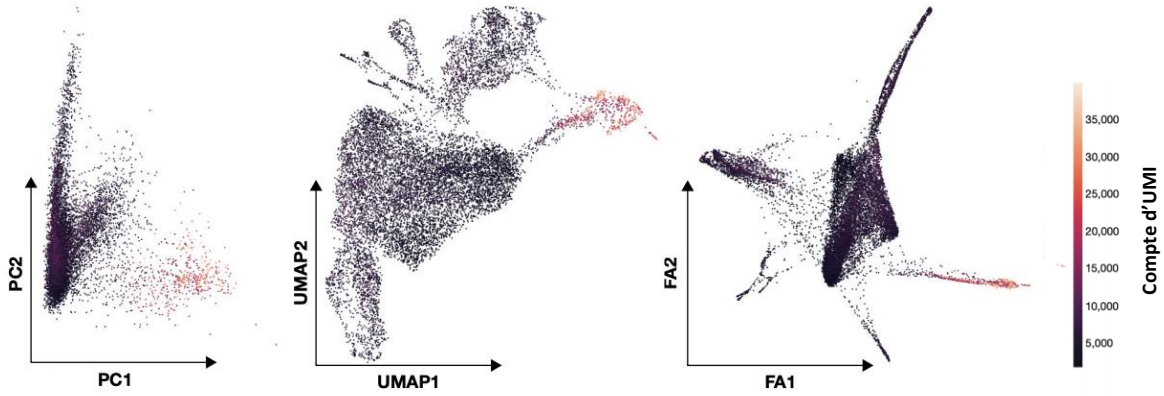


Figure 11 : Méthodes de réduction dimensionelle utilisées pour la représentation graphique de données de scRNA-Seq [9].

Les cellules issues de données d'épithelium intestinal de souris[18] sont colorées en fonction du nombre d'UMI détectés et ordonnées dans un espace à deux dimensions selon trois méthodes différentes : l'analyse en composante principale (PC), la UMAP, et le ForceAtlas2 (FA).

1.3.5. Clustering (formation de groupes de cellules)

Le regroupement de cellules en « clusters » (groupes de cellules) est généralement le premier résultat d'une analyse de données de scRNA-Seq. Les méthodes les plus communément utilisées sont appelées méthodes de détection des communautés. Ce sont des algorithmes de séparation basés sur des graphiques. En effet, ces méthodes s'appuient sur une représentation graphique des données. Cette représentation graphique est obtenue par la méthode des k plus proches voisines ou KNN (pour k-nearest neighbors). Les cellules y sont représentées comme des nœuds dans le graphique. Chaque cellule est connectée à ses k cellules les plus similaires en termes d'expression génique. La similarité entre les cellules est basée sur les distances euclidiennes dans l'espace dimensionnel réduit par l'analyse en composante principale. Le choix de la valeur de k dépend de la taille du jeu de donnée et il est généralement défini entre 5 et 100. Ainsi, les régions qui contiennent des cellules très proches dans l'espace dimensionnel d'expression génique sont représentées comme des régions denses en cellules dans l'espace graphique déterminé par la méthode KNN. Ces régions denses en cellules sont ensuite détectées par les méthodes de détection des communautés[9]. L'algorithme de Louvain[19] est la méthode de détection des communautés implémentée dans Seurat. Il permet de détecter les communautés comme des groupes de cellules qui ont plus de liens entre elles que le nombre de liens théorique moyens entre les cellules.

Suite à cette étape de clustering, les cellules sont regroupées entre cellules proches au niveau de l'expression génique, chaque cluster devrait correspondre à un type cellulaire particulier.

1.3.6. Annotation des clusters

Pour identifier à quel type cellulaire un cluster correspond, il est nécessaire de s'intéresser à l'expression génique des cellules qui compose ce cluster. Idéalement, il faut identifier des gènes marqueurs qui sont exprimés à un niveau relativement élevé et dont la fonction est suffisamment documentée pour être interprétée en tant que tel. La façon la plus courante d'identifier les gènes marqueurs est de mettre en évidence les gènes différentiellement exprimés entre les cellules d'un cluster et toutes les autres cellules de

l'ensemble des données, puis d'inspecter les gènes fortement régulés à la hausse et de les comparer aux données de la littérature[9].

Néanmoins, les cellules qui composent un cluster ne sont pas forcément représentatives d'un type cellulaire. L'identité cellulaire est un concept qui n'est pas clairement défini[20]. Par exemple, identifier un cluster comme étant composé de lymphocytes T peut être satisfaisant dans un certain contexte, alors qu'il serait nécessaire de différencier les lymphocytes T CD4+ et CD8+ dans un autre contexte. De plus, les cellules d'un même type mais dans un état différent (en cycle, en apoptose…) peuvent être regroupées dans des clusters différents. Il est donc nécessaire lors des étapes de clustering et d'annotation des clusters de définir le niveau de détail nécessaire en rapport avec la question expérimentale posée. Il faut donc avoir une bonne connaissance de la biologie des cellules étudiées pour annoter correctement les clusters et garder un œil critique sur l'étape de clustering.

Récemment, de nombreuses bases de données de scRNA-Seq ont été rendues disponibles, notamment le « Human Cell Atlas » [21] qui facilite grandement l'annotation des clusters. Lorsque des données de référence existent, certains algorithmes permettent l'annotation automatique des clusters en comparant les profils d'expression génique des clusters avec des données de référence et ainsi transférer les annotations entre la référence et le jeu de donnée[22].

1.3.7. Annotation des cellules

S'il est possible d'annoter automatiquement les clusters, on peut également annoter indépendamment les cellules une par une, ce qui permet de s'affranchir de l'étape de clustering et des biais qui lui sont associés. L'algorithme implémenté dans le package R SingleR[23] entre autres permet cette annotation à l'échelle unicellulaire. Tout d'abord, SingleR calcul un coefficient de corrélation de Spearman entre le profil d'expression génique de chaque cellule et le profil d'expression génique de chaque type cellulaire des données de référence. La corrélation est calculée sur les gènes les plus spécifiques de chaque sous population des données de référence. Ensuite, pour chaque cellule, les coefficients de corrélation avec les types cellulaires des données de référence sont comparés. Le type cellulaire qui a le coefficient de corrélation le plus faible est éliminé, puis SingleR refait le calcul

des coefficients de corrélation sur les types cellulaires restants. La méthodologie est appliquée jusqu'à ce qu'il ne reste qu'un seul type cellulaire issu des données de référence qui est ainsi attribué à la cellule.

1.3.8. Analyse des gènes différentiellement exprimés

Une des questions à laquelle on peut répondre avec des données de scRNA-Seq, est quels sont les gènes différentiellement exprimés entre deux conditions (ex : malade et sujets sains). Contrairement aux analyses de gènes différentiellement exprimés en RNA-Seq, les données de scRNA-Seq permettent de prendre en compte l'hétérogénéité cellulaire, de comparer des populations cellulaires précises et non pas une « soupe » contenant des cellules de types parfois très différents.

S'il apparait que les outils statistiques développés pour mettre en évidence les gènes différentiellement exprimés en RNA-Seq sont utilisable de la même manière pour les données de scRNA-Seq, de nombreux algorithmes ont été développés prenant en compte la spécificité des données qui contiennent plus de bruit de fond et de nombreux comptes de gènes à zéro (appelé aussi dropout). Récemment, une étude comparant des méthodes d'analyse des gènes différentiellement exprimés suggère que les méthodes utilisées pour le RNA-Seq sont aussi performantes que celles développées spécialement pour le scRNA-Seq[24]. De plus, lorsque les techniques utilisées pour le RNA-Seq ont été adaptées en prenant en compte le poids des gènes dans les tests, elles sont clairement supérieures aux techniques développées spécialement pour le scRNA-Seq.

Le problème de ces algorithmes est qu'ils nécessitent une puissance de calcul informatique et temps de calcul important. Ce défaut va devenir de plus en plus problématique au vu de la tendance à l'augmentation du nombre de cellules étudiées dans les expériences de scRNA-Seq. L'algorithme MAST représente alors une alternative intéressante[25]. Celui-ci utilise un modèle statistique pour prendre en compte le dropout tout en modélisant les variations d'expression génique qui sont liés aux variables techniques. De plus, l'algorithme MAST est beaucoup moins exigeant en termes de puissance et en temps de calcul.

1.3.9. Analyses des gènes regroupés par fonction

Les listes de gènes différentiellement exprimés contiennent souvent un grand nombre de gènes qu'il apparait difficile d'interpréter. On peut tenter de faciliter l'interprétation en regroupant les gènes en fonction de leurs caractéristiques communes et tester si ces caractéristiques communes sont surreprésentées dans la liste de gènes différentiellement exprimés.

Les regroupements de gènes (ou genesets) impliqués dans des processus biologiques communs sont décrits dans les bases de données comme « MSigDB »[26], et « Gene Ontology » [27,28]. Ceux impliqués dans les voies de signalisation sont décrits dans « KEGG »[29] et « Reactome » [30]. L'enrichissement de la liste de gènes différentiellement exprimés en gènes appartenant à des genesets définis par les bases de données précédemment cités ont été comparés par Tarca et al[31].

1.3.10. Analyse des trajectoires

La diversité cellulaire ne peut pas être objectivée seulement par l'annotation de clusters ou l'annotation des cellules. Dans les processus de développement (par exemple l'hématopoïèse), les mécanismes biologiques impliqués et responsables de l'hétérogénéité des populations les plus matures sont des phénomènes continus[32]. Pour décrire les transitions entre les différents types cellulaires, et les branchements entre les différentes étapes de développement, différentes méthodes ont été développées. Ces méthodes sont appelées inférences de trajectoire. Elles permettent de voir les données scRNA-Seq comme une photo d'un processus continu. Ce processus est reconstruit en trouvant des chemins à travers l'espace cellulaire en minimisant les changements transcriptionnels entre les cellules voisines. L'ordre des cellules tout au long de ces chemins est décrit par une variable appelée pseudotemps (pseudotime). Cette variable correspond pour chaque cellule à sa distance avec la cellule située au départ du chemin. De nombreuses méthodes d'inférence de trajectoire ont été développées et ont été comparées de manière quasi exhaustive par Saelens et al[33]. Les auteurs concluent que l'algorithme du package R Slingshot[34] est celui qui est le meilleur dans la description de trajectoires linéaires et celles présentant une ou plusieurs bifurcations. Pour

les trajectoires plus complexes, c'est l'algorithme PAGA implémenté dans le package python Scanpy[35] qui est le plus approprié.

1.3.11. Dynamique d'expression des gènes au cours des trajectoires

Les gènes qui varient de manière régulière au cours du pseudotemps caractérisent la trajectoire et peuvent être utiles pour identifier le processus biologique sous-jacent. De plus, les listes de gènes associées à la trajectoire ont de bonne chance de contenir des gènes qui régulent le processus modélisé. Les premières approches permettant de trouver des gènes associés aux trajectoires testaient les gènes différentiellement exprimés entre les clusters le long du pseudotemps[36]. On peut désormais identifier les gènes qui varient au cours du pseudotemps grâce à des techniques de régression de l'expression génique par rapport au pseudotemps. Les auteurs de Slingshot proposent d'utiliser ce type de méthode dans le package R TradeSeq[37].

1.4. Conclusion

Le scRNA-Seq est une technologie récente et toujours en pleine évolution. Des améliorations du point de vue de la microfluidique, de la biochimie, et de l'analyse bio-informatique sont publiées chaque semaine. C'est une technologie dont le coût (environ 1 euro par cellule analysée) est encore important à l'heure actuelle, mais qui tend à diminuer. Le scRNA-Seq parait incontournable pour étudier l'hétérogénéité des systèmes biologiques et sera probablement dans le futur un outil de base dans les laboratoires de biologie.

2. L'hématopoïèse humaine

2.1. Généralités

L'hématopoïèse correspond à l'ensemble des mécanismes impliqués dans la production des cellules sanguines matures à partir d'un pool restreint de cellules souches hématopoïétiques (CSH). C'est un processus complexe, encore imparfaitement compris. Chez l'adulte, l'hématopoïèse se déroule principalement dans la moelle osseuse des os courts et plats tels que le sternum, les côtes, le bassin, le crâne ou encore les vertèbres. La production quotidienne est estimée à environ 10^{12} cellules sanguines par jour, toutes lignées confondues. L'hématopoïèse est un processus finement régulé par de multiples facteurs intrinsèques et extrinsèques qui permettent à la CSH de se différencier en un lignage cellulaire défini en fonction des besoins de l'organisme. Les progéniteurs hématopoïétiques en aval des CSH sont des cellules engagées dans le processus de différenciation mais pouvant encore donner naissance à plusieurs lignages. Ces progéniteurs vont ensuite donner naissance à des précurseurs qui sont morphologiquement identifiables en microscopie optique, spécifiques d'un lignage précis et qui se différencient en cellule mature circulante.

2.2. Cellule souche hématopoïétique (CSH)

Ernst Haeckel fut le premier à expliciter la notion de cellule souche en 1868. Il l'utilisa au sens darwiniste du terme pour désigner l'organisme unicellulaire d'où toute vie multicellulaire est issue. Les histopathologistes ont ensuite appliqué ce concept de cellule souche à l'hématopoïèse normale, mettant en avant le concept d'un progéniteur commun aux globules rouges et blancs. Dès le début, le concept de cellule souche est vu comme un modèle arborescent dans lequel les cellules souches multipotentes donnent naissance à leur descendance à travers une série ordonnée d'étapes de ramification.

Le premier essai « fonctionnel » *in vivo* de caractérisation des cellules souches fut le sauvetage par greffe de moelle osseuse d'une souris après irradiation létale. L'identification et le compte de cellules capables de s'auto-renouveler et de se différencier dans la rate de souris préalablement irradiées, a permis d'estimer le nombre de CSH à 1 cellule pour 10000 dans la moelle osseuse de souris [38].

Les scientifiques se sont ensuite intéressés à développer des méthodes pour purifier les CSH à partir de la moelle osseuse, afin de les étudier. L'isolement des CSH a été rendu possible grâce à la technique de tri cellulaire par cytométrie en flux. En 1988, l'équipe de Weissman a pu isoler chez la souris à l'aide d'une combinaison de marqueurs de surface une population de cellules enrichies en CSH, c'est-à-dire des cellules capables de proliférer, de se différencier et de reconstituer l'hématopoïèse après transplantation chez la souris irradiée[39]. Ils ont plus tard décrit la CSH humaine caractérisée par les marqueurs CD34+ et CD90+ et la capacité à générer *in vitro* et après transplantation chez la souris immunodéprimé les lignées lymphoïdes et myéloïdes. Au fur et à mesure de l'identification de nouveaux marqueurs de surface la caractérisation des CSH a été de plus en plus précise.

Les CSH ont été historiquement définies par deux propriétés : l'auto-renouvellement et la multipotence. A l'inverse, les progéniteurs sont définis par l'absence d'auto-renouvellement prolongé et une capacité de différenciation de lignée restreinte à un ou deux lignages.

2.3. Modèle classique de l'hématopoïèse

2.3.1. Progéniteurs et hiérarchie arborescente de l'hématopoïèse

Le modèle de hiérarchie arborescente basé sur l'immunophénotypage et les cultures de progéniteurs *in vitro* a été en grande partie établi par le groupe de Weissman **(Figure 12)**[40]. Dans ce modèle classique, les CSH se divisent en progéniteurs multipotents (MPP) qui ont été décrits comme ayant des capacités de multipotence mais des capacités d'auto-renouvellement incomplètes[41]. La première bifurcation se produit entre les progéniteurs myéloïdes communs (CMP) et les progéniteurs lymphoïdes communs (CLP) qui dérivent tous deux des MPP. Le deuxième point de branchement sépare les progéniteurs bipotents granulo-monocytaires (GMP) et érythro-mégacaryocytaires (MEP)[42]. Les CLP donnent naissance aux lignées T, B, NK et dendritiques, les GMP se différencient en granulocytes et monocytes, alors que les MEP génèrent les mégacaryocytes et les érythrocytes.

Toutes ces populations forment un modèle hiérarchique arborescent et équilibré au sein duquel les facteurs de transcriptions et les cytokines assurent le rôle de facteurs intrinsèques et extrinsèques permettant l'engagement vers un lignage et la différenciation progressive des CSH en cellules sanguines matures [40]

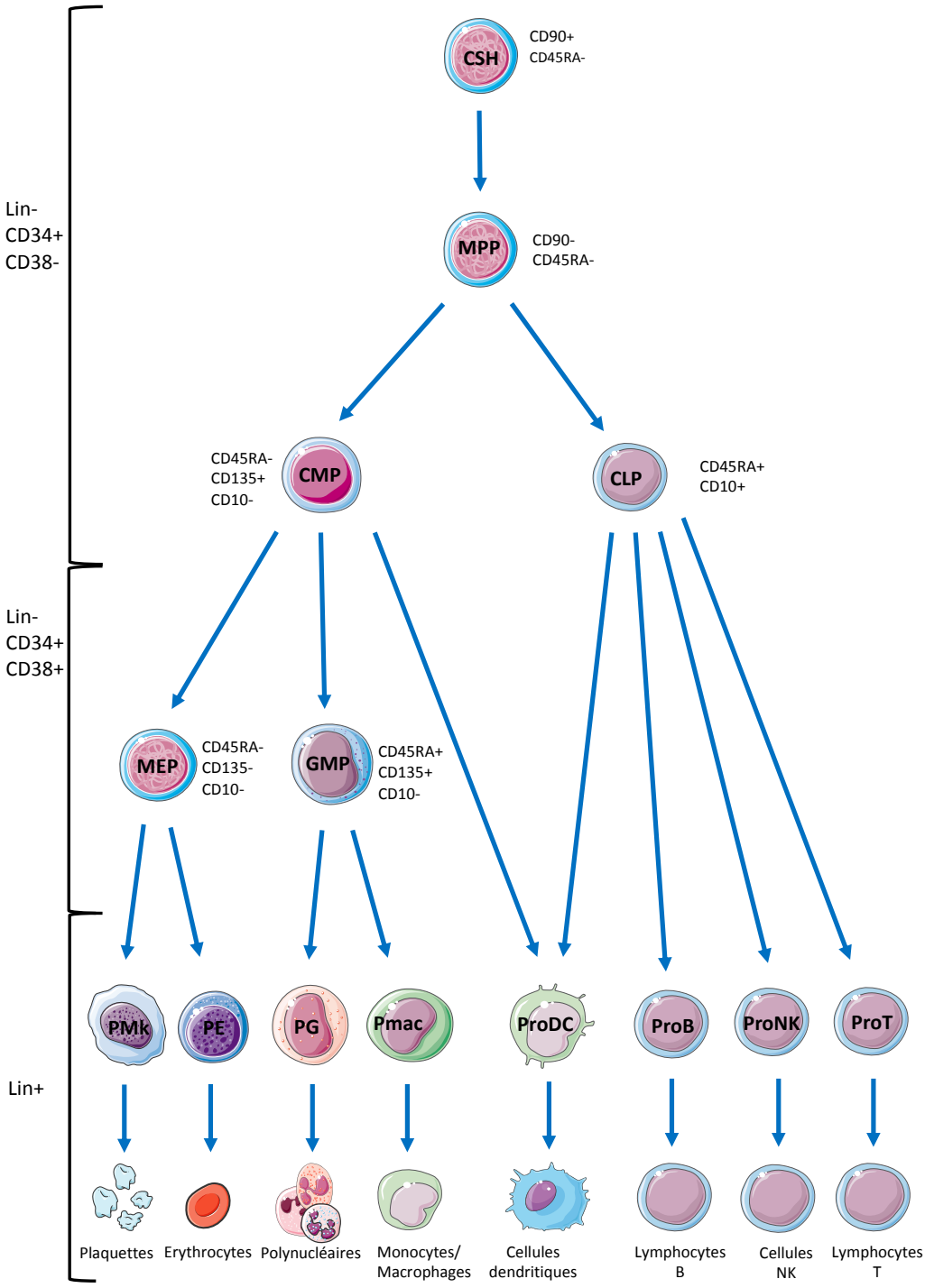


Figure 12 : Modèle de la hiérarchie hématopoïétique selon Weissman[40].

Les principales classes de cellules souches et progénitrices sont définies par leurs caractéristiques immunophénotypiques répertoriés à coté de chaque population et dans les accolades à gauche du schéma. Lin : cocktail contenant les marqueurs de surface cellulaire de toutes les populations différenciées en phase terminale. **CSH** : cellule souche hématopoïétique ; **MPP** : progéniteur multipotent ; **CLP** : progéniteur lymphoïde commun ; **MEP** : progéniteur érythro-mégacaryocytaire ; **CMP** : progéniteur myéloïde commun ; **GMP** : progéniteur granulo-monocytaire ; **PMk** : progéniteur mégacaryocytaire ; **PE** : progéniteur érythrocytaire ; **PG** : progéniteur granuleux ; **MacP** : progéniteur mono/macrophagique ; **ProDC** : progéniteur dendritique ; **ProB :** progéniteur lymphoïde B ; **ProNK** : progéniteur lymphoide NK ; **ProT** :progéniteur lymphoide T

Ce modèle classique a été amélioré par la suite notamment grâce à plusieurs études concernant les progéniteurs lymphoïdes. L'équipe de Dick a décrit une population appelée MLP (pour progéniteur multilymphoïde) qui englobe tous les progéniteurs qui ont un potentiel lymphoïde B, NK et T et qui peuvent avoir ou non d'autres potentiels (par exemple myéloïde) [43]. L'équipe de Vyas a également mis en évidence chez l'homme un progéniteur lymphomyéloïde qui a un potentiel d'engagement de lignée principalement lymphoide mais qui garde également un potentiel myéloïde et qu'ils ont choisi de nommer LMPP (lymphomyeloid primed progenitor)[44]. On peut tenter de résumer toutes ces études sur la figure 13. Dans ce modèle de hiérarchie arborescente , les principales différences par rapport au modèle précédent sont les interconnections entre les lignages lymphoïde et myéloïde ainsi qu'une définition plus précise des progéniteurs lymphoïdes avec la description des LMPP et MLP (Multilymphoid progenitor) [45]. En effet, par rapport au modèle précédent, les LMPP donnent naissance aux MLP mais pourraient aussi engendrer les GMP, et même directement ou via d'autres intermédiaires non décrits les cellules dendritiques. Les MLP (qui pourraient aussi provenir des MPP) donnent naissance aux progéniteurs B/NK et ETP (early T precursor), mais pourrait aussi engendrer directement ou *via* d'autres intermédiaires non décrits les cellules dendritiques **(Figure 13)**.

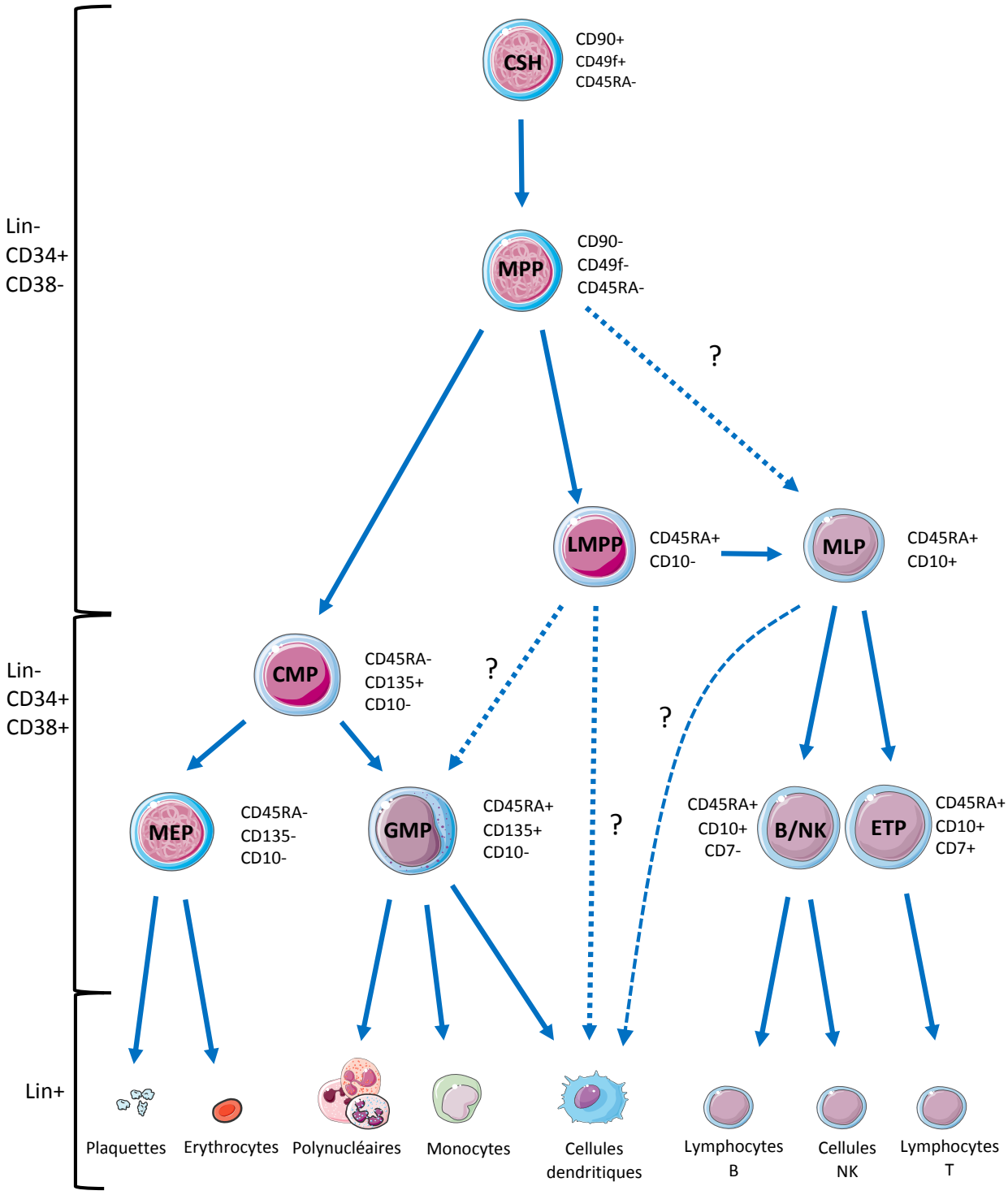


Figure 13 : Modèle classique amélioré adapté de Doulatov et al [45].

Les principales classes de cellules souches et progénitrices sont définies par leurs marqueurs immunophénotypiques répertoriés à coté de chaque population et dans les accolades à gauche du schéma. Lin : cocktail contenant les marqueurs de surface cellulaire de toutes les populations différenciées en phase terminale. Les flèches continues représentent des présomptions fortes tandis que les flèches en pointillées sont incertaines. **LMPP** : lymhomyeloid primed progenitor ; **MLP** : Multilymphoid progenitor.

2.3.2. Destin cellulaire et régulation de l'hématopoïèse

Le destin d'une cellule souche implique un choix entre des programmes d'expression génique alternatifs exécutés par des régulateurs transcriptionnels et épigénétiques en réponse à des signaux extracellulaires. Les facteurs de transcription se lient à des motifs de séquence ADN spécifiques au niveau des promoteurs, enhancers et silencers des gènes cibles ; ou à distance de ceux-ci. La spécificité d'un type cellulaire est obtenue grâce à des interactions de facteurs de transcription, qui forment les éléments constitutifs clés de réseau de régulation plus larges. Les facteurs de transcriptions peuvent former entre eux des complexes qui recrutent des régulateurs épigénétiques pour moduler le statut d'activation d'un locus génique. Ce statut peut être transmis aux générations cellulaires suivantes comme « mémoire épigénétique ». Une cellule progénitrice peut ainsi amorcer son destin cellulaire en permettant l'accès à des éléments régulateurs associés à des gènes impliqués dans la différenciation dans une lignée spécifique mature [46]. Les différents mécanismes de régulation de l'hématopoïèse sont exposés dans les paragraphes suivants.

2.3.2.1. Signaux extracellulaires

Les cytokines sont des facteurs solubles qui circulent dans l'organisme et peuvent être sécrétés par un ou plusieurs types cellulaires. Au cours de l'hématopoïèse, les cytokines en se fixant sur des récepteurs membranaires spécifiques vont activer des voies de signalisation et induire la production de facteurs de transcriptions spécifiques. Par exemple, l'érythropoïétine (EPO), et la thrombopoïétine (TPO) sont respectivement impliqués dans la régulation de la différenciation érythroïde et mégacaryocytaire. Les différenciations granuleuses et monocytaires sont régulées par l'association de plusieurs cytokines, notamment l'interleukine 3, l'interleukine 6, le G-CSF, et le GM-CSF. L'expression des récepteurs aux cytokines à la surface des progéniteurs est variable. C'est la régulation de leur expression qui permet l'engagement progressif vers un lignage particulier[47].

2.3.2.2. Facteurs de transcriptions

L'activation des récepteurs des cytokines par leur ligand induit l'expression de facteurs de transcription par l'intermédiaire de cascades de signalisation. En se fixant à des séquences

d'ADN appelés promoteurs, ils peuvent induire l'expression de gènes cibles. L'expression de ces gènes cibles constitue un programme transcriptionnel spécifique permettant l'engagement vers un lignage particulier. De plus, les facteurs de transcription peuvent réguler négativement d'autres facteurs de transcription antagonistes[48]. Ainsi, l'expression du facteur de transcription GATA1 qui permet la différenciation érythro-mégacaryocytaire régule négativement le facteur de transcription antagoniste PU.1 impliqué dans la différenciation granulo-monocytaire[49]. L'antagonisme GFI-1/PU.1 régule la différenciation des GMP en granuleux et monocytes[50], et les facteurs de transcriptions FLI-1 et EKLF régule la différenciation des MEP en érythrocytes et mégacaryocytes[51,52].

2.3.2.3. Régulateurs épigénétiques

Les modifications épigénétiques induisent des modifications de l'expression génique sans modifier la séquence de l'ADN correspondant. Ces modifications peuvent être transmises à la descendance lors des divisions cellulaires mais contrairement aux mutations qui affectent la séquence d'ADN, les modifications épigénétiques sont réversibles.

La méthylation de l'ADN dans les régions promotrices d'un gène, va entrainer une moindre accessibilité de ces promoteurs à la machinerie transcriptionnelle et ainsi entrainer une diminution de l'expression génique. Cette méthylation est médiée par les protéines de la famille DNMT (pour DNA Methyl Transférases) qui vont transférer des groupements méthyles sur les cytosines et guanines de l'ADN[53 54].

Les histones jouent un rôle dans la compaction de l'ADN et donc son accessibilité à la machinerie transcriptionnelle. Les modifications post traductionnelles des histones peuvent moduler cette accessibilité. L'acétylation des résidus lysine des histones réalisé par les HAC (histones acetyl transférase) permet le relâchement de la chromatine et favorisent la transcription. Inversement la désacétylation des histones par les HDAC (histones desacetylases) régule négativement la transcription puisqu'il en résulte une structure de chromatine fermée et donc une accessibilité réduite pour la machinerie transcriptionnelle[55].

2.3.2.4. Micro ARN (miARN)

Les miARN sont de petites molécules d'ARN simple brin d'une vingtaine de nucléotides qui peuvent moduler l'expression génique. Ils se fixent principalement dans la région 3'UTR des ARNm ce qui entraine leur prise en charge par le complexe RISC (RNA-induced silencing complex) qui va inhiber la traduction spécifique et dans une moindre mesure entrainer la dégradation de l'ARNm cible[56].

2.3.3. La niche hématopoïétique

La niche hématopoïétique a été décrite pour la première fois en 1978 comme étant un site anatomique défini où les CSH pourraient être maintenues et se reproduire, où leur différenciation serait inhibée, et un espace qui limite le nombre de cellules souches[57]. La niche est aujourd'hui considérée comme un microenvironnement constitué de plusieurs sous-types cellulaires responsables de la sécrétion, de l'acheminement, et de la régulation des facteurs extrinsèques nécessaire au contrôle de la quiescence et de l'auto-renouvellement des CSH[58] **(Figure 14)**.

Il existe au moins deux niches hématopoïétiques différentes dans la moelle osseuse : La niche centrale qui est située au centre de la moelle et la niche endostéale à proximité immédiate de la surface osseuse. La niche centrale qui contient la majorité des sinusoïdes et des artérioles, représente 90% du volume de la moelle osseuse et abrite 85% des CSH.[59] Les fonctions spécifiques des niches artériolaires et sinusoïdales restent un sujet de controverse, différents laboratoires obtenant des résultats opposés selon les systèmes expérimentaux utilisés.[60–62] Etant une niche beaucoup plus petite (moins de 10% du volume de la moelle),la niche endostéale est relativement enrichie en CSH (15 % de toutes les CSH)[63], et contient tous les vaisseaux de la zone de transition[64]. Il est probable que les deux niches soient fonctionnellement différentes, en effet, les CSH activent peuvent migrer à travers la niche sinusoïdale[65] qui est responsable de la production sanguine quotidienne et qui est sensible au stress génotoxique induit par l'irradiation ou la myéloablation[66]. La migration des CSH à travers les sinusoïdes est régulée par les fibres nerveuses sympathiques [67,68]. En revanche, la niche endostéale n'est pas impactée, et semble être dédiée au maintien d'une réserve de CSH nécessaire à la régénération hématopoïétique à long terme[69]. A la fois dans les parties

centrales et endostéales de la moelle osseuse, les CSH résident dans des niches péri vasculaires [70], où les cellules endothéliales[71] et les cellules souches mésenchymateuses (CSM) péri vasculaires associées maintiennent et régulent les CSH[72]. Parmi ces CSM formant la niche, les cellules NG2+ sont associées aux artérioles dans toute la moelle osseuse [60] et aux vaisseaux de la zone de transition situés près de la surface osseuse[73], tandis que les CAR CSM[74 75] sont associées aux sinusoïdes de la partie centrale de la moelle osseuse. Chez l'homme, les CSM sont définis par l'immunophénotypage comme étant Lin- CD45- CD271+ CD140a-/dim [76]. Les CSM périsinusoidales sont CD146+ tandis que les CSM endostéales sont CD146-/dim[77]. Dans la niche endostéale, plusieurs signaux ont été décrits comme favorisant la mise au repos des CSH [69,70], ce qui est essentiel pour préserver la capacité d'auto-renouvellement des CSH. Cependant des CSH au repos peuvent également être retrouvées près des sinusoïdes [59], où d'autres cellules de la niche peuvent limiter la prolifération des CSH. Par exemple, les mégacaryocytes sont pour la plupart adjacents aux sinusoïdes et peuvent favoriser la quiescence des CSH par la sécrétion de TGFβ ou de TPO (thrombopoïétine)[78,79]. De plus, les cellules de Schwann non myélinisantes associées aux fibres nerveuses peuvent favoriser la quiescence des CSH en activant le TGFβ latent dans les niches centrales et endostéales de la moelle osseuse [80]. Par conséquent, il est possible que la quiescence des CSH soit régulée différemment selon leur localisation pour répondre aux exigences de l'hématopoïèse à l'état d'équilibre ou dans des conditions de stress.[81]

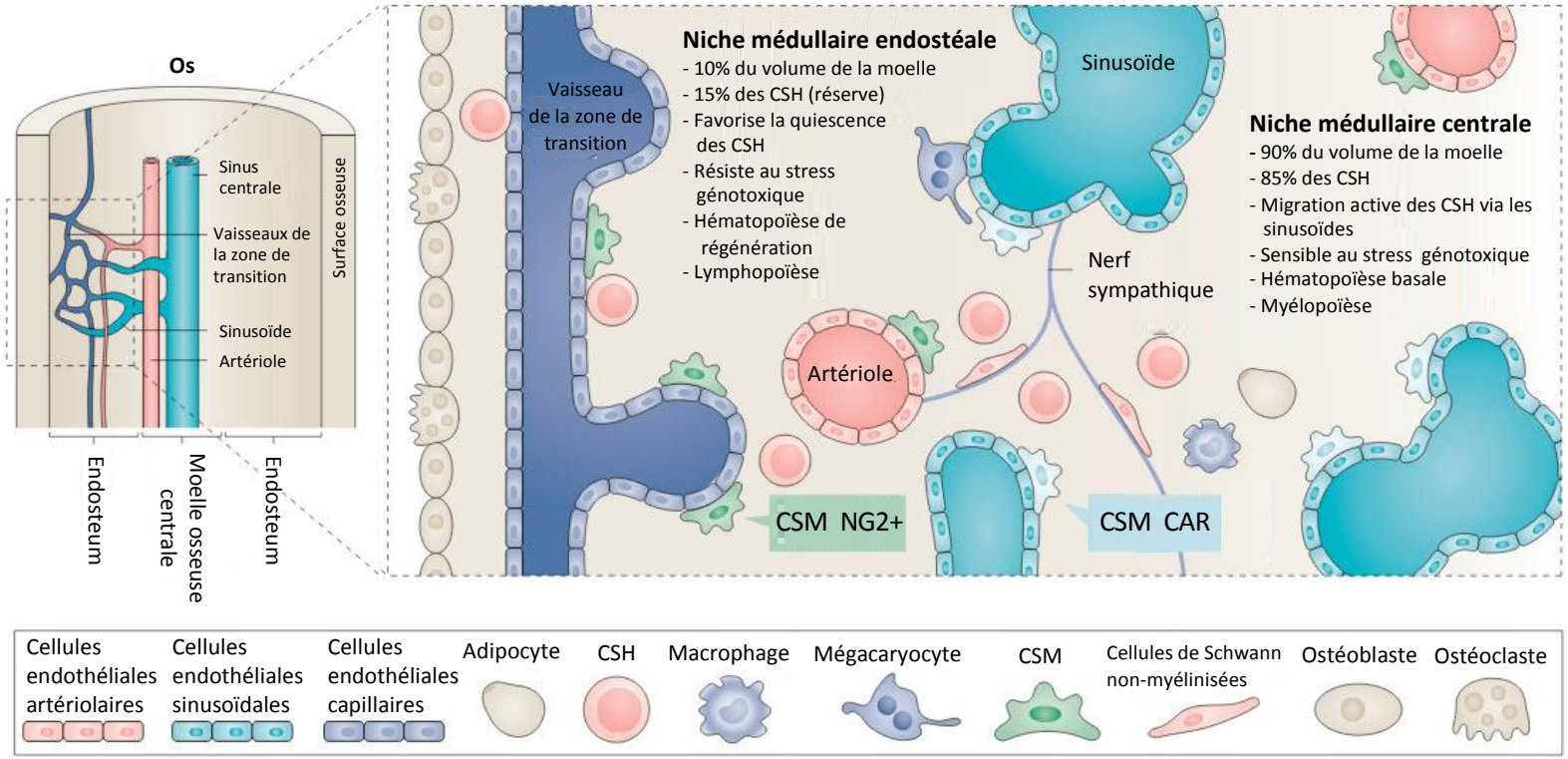


Figure 14 Représentation simplifiée de la niche hématopoïétique[81].

Schéma résumant les principaux types de cellules et les caractéristiques fonctionnelle de la niche hématopoïétique centrale et endostéale.
Les cellules souches mésenchymateuses (CSM) donnent naissance aux ostéoblastes et adipocytes, tandis ue les ostéoclastes partagent une origine monocytaire avec les macrophages. Une partie des CSM (NG2+ pour Neural/glial antigen 2) sont associées aux vaisseaux et artérioles de la zone de transition endostéale, tandis que les CSM CAR (pour CXCL12 abundant reticular) sont associées aux sinusoïdes dans la zone centrale de la moelle osseuse. Les fibres nerveuses du système sympathique régulent la migration des CSH à travers les sinusoïdes. Les différentes populations de CSM, les cellules endothéliales, les cellules de Schwann non myélinisantes et les mégacaryocytes pourraient contribuer à réguler l'équilibre entre la quiescence et la prolifération des CSH nécessaire à l'hématopoïèse basale ou régénérative. Des modifications dans les niches spécialisées pourrait affecter directement la balance entre la production myéloïde et lymphoïde. De même, la production déséquilibrée de cellules hématopoïétiques matures dans les niches spécifiques pourrait à son tour remodeler le microenvironnement. Attention, les marqueurs de différenciation des CSM utilisés sur ce schéma sont ceux décrits chez la souris. (D'après Méndez-Ferrer et al [81])

2.3.4. Conclusion

Bien que ces modèles classiques aient été très utiles pour comprendre le processus de différenciation des CSH, ceux-ci présentent des lacunes. En effet, le modèle arborescent simplifie la complexité des cellules souches et progénitrices (HSPC) en étant basé uniquement sur les marqueurs de surfaces et des expériences de transplantations de cellules en « bulk ». L'analyse des cellules en « bulk » suppose en effet que chaque cellule qui possède le même phénotype possède une fonction identique. Le destin de chaque cellule et les processus de différenciation se produisant à l'échelle individuelle, il parait donc impératif pour les comprendre que les expériences réalisées se fassent à l'échelle unicellulaire.

2.4. Modèle révisé de l'hématopoïèse

Dans le but de raffiner notre vision de l'hématopoïèse, le groupe de John Dick a trié les progéniteurs MPP, CMP et MEP du foie fœtal et de la moelle osseuse adulte, pour étudier leur potentiel de lignée à différents stades de développement[82]. Ils ont montré grâce à des techniques d'analyse à l'échelle unicellulaire (clonogénicité, expression génique sur 12 gènes), que les MPP, CMP et MEP définies précédemment sont hétérogènes, et que les MEP tels que définis de manière classique sont principalement des précurseurs érythroïdes. Le foie fœtal contient quant à lui plusieurs types de progéniteurs oligopotents distincts ce qui n'est pas le cas dans la moelle osseuse adulte. Dans celle-ci, deux classes de progéniteurs prédominent, multipotents et unipotents. Ainsi, cette étude fournit un modèle révisé de l'hématopoïèse normale et démontre que celui-ci est flexible au cours du développement **(Figure 15).**

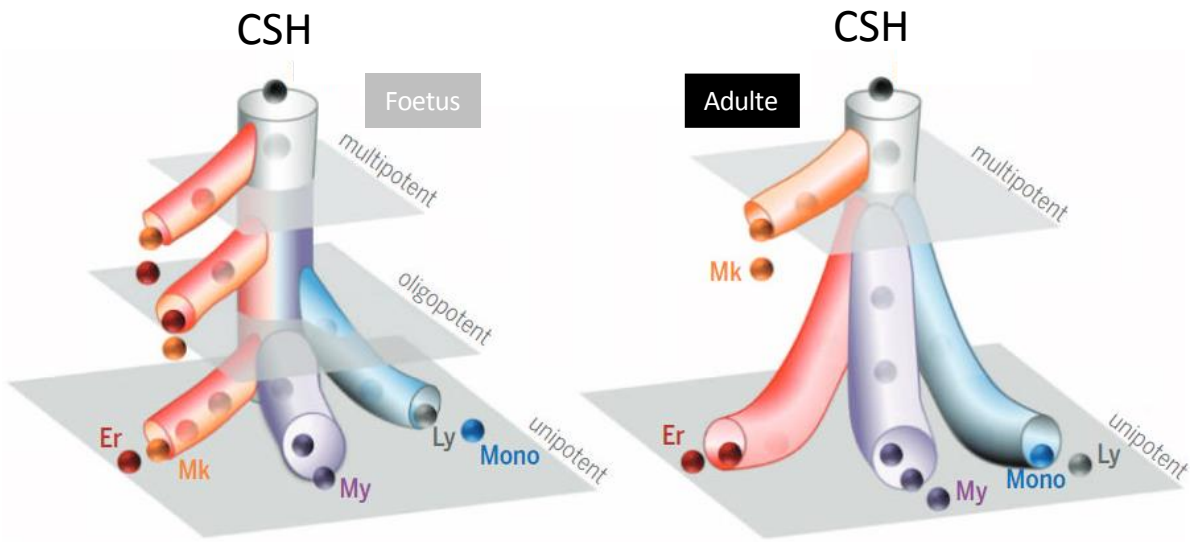


Figure 15 : Modèle révisé l'hématopoïèse selon Notta[82].
Le modèle révisé propose une évolution dans l'architecture des cellules progénitrices au cours du développement. Dans l'hématopoïèse fœtale, de nombreux types de cellules souches et progénitrices sont multipotentes. A l'âge adulte, le compartiment des cellules souches est multipotent, mais les progéniteurs sont unipotents. Les plans grisés représentent des niveaux théoriques de différenciation

2.5. Hématopoïèse et technologie single-cell haut débit

2.5.1. Généralités

Si l'équipe de John Dick a été la première à utiliser des techniques de mesure d'expression génique à l'échelle unicellulaire pour étudier l'hématopoïèse, ces techniques ne permettaient de mesurer l'expression que d'un nombre limité de gènes (une centaine). L'avènement des techniques de scRNA-Seq a permis de mesurer simultanément l'expression de plusieurs milliers de gènes dans chaque cellule offrant ainsi un niveau de détail sans précédent. Cette technique a permis ces dernières années de grandement préciser notre connaissance de l'hématopoïèse. En effet, pour un processus aussi rapide que l'hématopoïèse, le scRNA-Seq permet d'avoir un « instantané » des cellules et de leurs états d'expression génique à un moment donné. Bien que l'information temporelle soit manquante, le profil d'expression génique de chaque cellule permet de les placer dans un espace multidimensionnel sur des chemins de maturation, des progéniteurs les plus immatures jusqu'aux cellules matures des diff érentes lignées hématopoïétiques [83 84].

2.5.2. L'hématopoïèse: un processus continu?

Grâce au scRNA-Seq et à l'aide de modélisations informatiques étayées par des expériences *in vitro* à l'échelle unicellulaire, les chercheurs ont étudié le compartiment souche et progéniteur CD34+ de la moelle osseuse humaine[85]. Les auteurs montrent que les populations progénitrices unipotentes précédemment décrites forment effectivement des sous populations discrètes. Par contre, les CSH/MPP semblent former une structure plutôt continue. Au vu de l'absence d'amorçage évident dans ces populations, les auteurs concluent que celles-ci évoluent dans une sorte de « nuage » qui donne directement lieu à des progéniteurs engagés sans transition majeure à travers des étapes multi ou bipotentes. De plus, mêmes les populations les moins engagées peuvent donner lieu à un lignage unique *in vitro*, indiquant que les vraies cellules multipotentes ne constituent qu'une petite fraction de la population des CSH conventionnelles. Ainsi, les auteurs suggèrent une restriction de lignée précoce, ce qui les conduit à proposer un modèle dans lequel l'acquisition du destin de lignée est un processus continu. En utilisant des cellules progénitrices lympho-myéloïdes de sang de cordon humain tels que des LMPP, des GMP et des MLP, une étude suggère une modèle dans

lequel un continuum de progéniteurs est en charge de la différenciation lymphoïde et myéloïde[86]. Au niveau de la différenciation érythro-mégacaryocytaire, une étude montre que la population des MEP serait en fait hétérogène et composé par plusieurs sous types de progéniteurs unipotents[87] ; une autre étude suggère que la principale voie de différenciation des mégacaryocytes proviendrait directement du compartiment des CSH[88]. Certains auteurs ne sont pas entièrement d'accord avec cette idée de processus continu à partir d'un « nuage » de cellules souches et suggèrent un processus hiérarchique astructuré avec notamment une division qui séparent les progéniteurs érythro-mégacaryocytaires des progéniteurs lympho-myéloïdes puis un processus de différenciation continu en aval[89].

Quoiqu'il en soit, ces études changent notre point de vue sur la différenciation hématopoïétique qui correspondrait davantage à un processus continu plutôt qu'à un processus hiérarchique structuré et figé. Les avis divergents sur la théorie du « nuage » de CSH/MPP pourrait provenir de la manière dont les différents algorithmes mathématiques utilisées pour interpréter les données de scRNA-Seq, placent les cellules dans l'espace multidimensionnel.

2.5.3. Destin cellulaire et scRNA-Seq

Les donnéees actuelles laissent à penser que les cellules multipotentes présentent un amorçage multiligné, ce qui implique l'activation simultanée à bas niveau de plusieurs programmes d'expression spécifiques de différentes lignées. Le choix de différenciation d'une cellule vers un lignage résulterait de l'activation d'un programme gagnant alors que les programmes alternatifs sont éteints. L'antagonisme croisé entre des paires de facteurs de transcription déterminant le lignage final représente un modèle mécanistique attrayant, basé initialement sur le facteur de transcription érythroïde GATA1 et le facteur de transcription myéloïde PU.1. Cependant, une expérience d'imagerie en « time lapse » à l'échelle unicellulaire a montré pour PU1 et GATA1 que leur dynamique d'expression observée est incompatible avec l'hypothèse que la commutation stochastique entre PU.1 et GATA1 précède et initie la prise de décision entre la différenciation vers la lignée érythro-mégacaryocytaire par rapport à la lignée granulo-monocytaire. Cette étude suggère que les facteurs de transcriptions ne déterminent pas le choix vers un lignage mais le renforce une

fois que celui-ci est fait [90]. De plus, avec des méthodes de modélisation mathématique, les auteurs ont montré que les niveaux des facteurs de transcription PU.1 ne change pas pendant, mais longtemps après la décision d'orientation vers le lignage prévu. Ce qui indique que l'antagonisme PU.1/GATA1 ne peut être à l'origine du choix d'orientation vers une lignée[91].

2.5.4. Nouvelle représentation de l'hématopoïèse

Une nouvelle représentation de l'hématopoïèse à la lumière des études de scRNA-Seq a été proposé par Laurenti et Gottgens[92]. Selon les auteurs, il semble clair qu'un arbre dans lequel chaque cercle représente un état successif de restriction de potentiel et dans lequel chaque cercle est connecté à quelques autres par des flèches est une simplification excessive. Premièrement, le cercle est considéré de manière intuitive comme un ensemble de cellules avec des caractéristiques spécifiques ce qui est incompatible avec le grand degré d'hétérogénéité observé expérimentalement. Deuxièmement, ces arbres représentent un ensemble restreint de transitions possibles entre les cercles, ce qui sous-estime probablement les voies de différenciations possibles in vivo. Etant donné que l'isolement via des marqueurs de surface permet de décrire des sous populations très différentes du point de vue de leurs capacités d'auto-renouvellement, de différenciation, ou de capacités de prolifération, un modèle dans lequel toutes les lignées se ramifient directement à partir du compartiment CSH (modèle du « nuage ») semble également irréaliste. Ainsi les auteurs présentent une visualisation alternative **(Figure 16a),** dans laquelle les trajectoires de différenciation sont cartographiées sur des zones qui ont longtemps été représentées en cercle, mettant en évidence à la fois la diversité des itinéraires possibles et la prévalence du choix précoce de la lignée. De plus, il est important de se rappeler qu'une CSH produira un très grand nombre de descendants, qui augmente exponentiellement avec chaque division, élément jusqu'ici ignoré dans les représentations graphiques de l'hématopoïèse. C'est pourquoi les auteurs incluent ce paramètre dans leur deuxième modèle complémentaire du premier **(Figure 16b).**

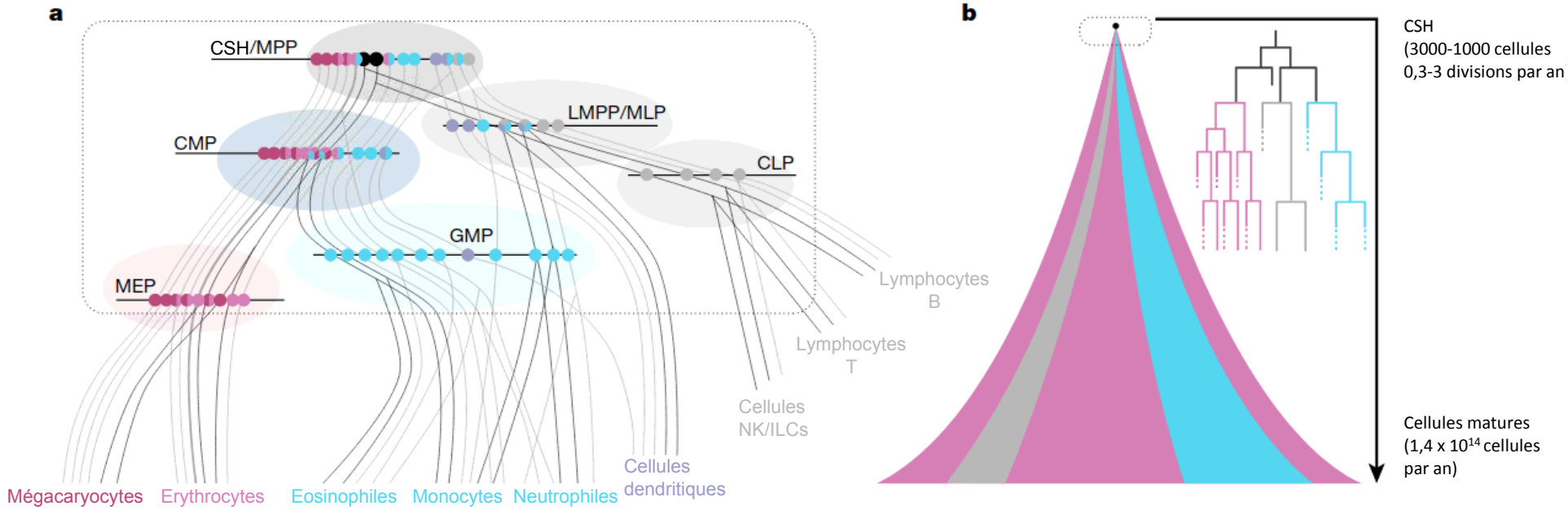


Figure 16 : Visualisation de la hiérarchie hématopoïétique basée sur la trajectoire[92].

a) Illustration bidimensionnelle de l'hématopoïèse précoce. Les lignes continues indiquent des trajectoires de différenciation pour différents types de cellules présentes dans le compartiment phénotypique HSPC (HSC et MPP). Le long de ces trajectoires, les cellules et leurs progénies, traversent des compartiments de progéniteurs généralement définis par des combinaisons spécifiques de marqueurs de surface cellulaire. Les lignes horizontales représentent des instantanés du potentiel de lignée des cellules présentent dans chaque compartiment phénotypique (les cercles monochromes désignent les cellules unipotentes, les cercles bicolores désignent les cellules bipotentes, les cercles tricolores désignent les cellules tripotentes, et les cercles noirs désignent les cellules multipotentes). Ainsi la figure illustre les trajectoires de différenciation rapportées jusqu'à présent dans la littérature, mais leurs proportions peuvent ne pas refléter la situation in vivo
b) Illustration de la descendance d'une seule CSH. Le rose, le bleu et le gris représentent respectivement les lignées érythroïdes, myéloïdes et lymphoïdes. L'amplification à partir de quelques milliers de cellules souches hématopoïétiques est gigantesque. Pour cela, cette amplification doit inclure un compartiment amplificateur transitoire. De plus, il y a beaucoup plus de cellules érythroïdes différenciées en phase terminale que de cellules myéloïdes, et encore moins de cellules lymphoïdes, toutes avec des taux de renouvellement différents, le flux dans chaque compartiment doit donc être hautement régulé. (D'après Laurenti et al[92])

Un consortium appelé *Human Cell Atlas* (https://www.humancellatlas.org/) a pour objectif de créer des cartes de référence à l'échelle de la cellule pour plus de cinquante tissus humains. Cela permettra la création de profils cellulaires de référence, de gènes marqueurs et de réseaux de régulation génique pour une meilleure compréhension des cellules saines et pathologiques. Les données hématologiques ont récemment été publiées, à partir de moelles osseuses de huit donneurs sains, âgés de 26 à 52 ans les auteurs ont pu analyser par scRNA-Seq plus de 200 000 cellules.

Il apparait chez l'homme comme un besoin primitif de tout faire rentrer dans des cases. Ainsi, malgré les fortes preuves en faveur d'un modèle continu de l'hématopoïèse, un total de 35 groupes de cellules ayant une cohérence transcriptomique ont été définis. Ces données mettent en évidence de nouvelles populations de progéniteurs de lignées mixtes et des trajectoires supposées vers les différenciations granulocytaires, monocytaires, lymphoïdes, érythroïdes, mégacaryocytaires et éosinophiles. Les auteurs décrivent notamment avec une précision inégalée les différentes sous populations du compartiment des cellules souches et progénitrices CD34+ **(Figure 17).**

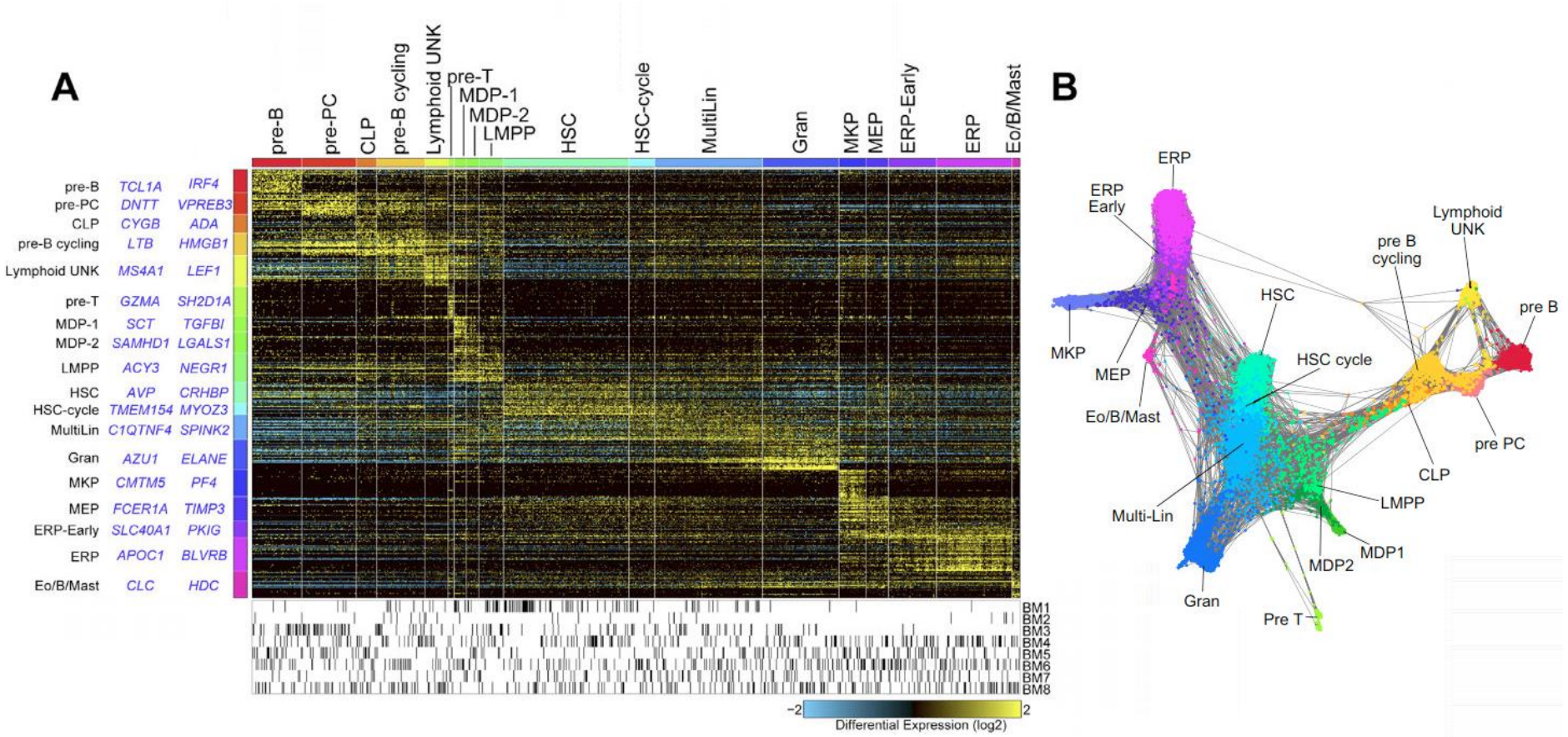


Figure 17 : Sous-populations hématopoïétiques du compartiment CD34+ normal selon Hay et al [93].

A) Heatmap représentant les marqueurs spécifiques des différentes sous populations du compartiment CD34+ déterminées après clustering itératifs des données de scRNA-Seq.

B) Représentation en 2 dimensions de ces sous populations à l'aide de l'algorithme SPRING.

2.5.5. Limites du scRNA-Seq

L'analyse des données scRNA-Seq produit un paysage des cellules souches et progénitrices dans lequel il apparait difficile de mettre en évidence des états cellulaires clairement distincts. Certains aspects de la structure des populations peuvent être masqués par le bruit de fond provoqué par des processus tel que le cycle cellulaire ou le métabolisme. Les données actuelles de scRNA-Seq représentent une vue incomplète des états cellulaires en raison de leur inaptitude à détecter les gènes faiblement exprimés et de l'absence d'informations sur l'expression protéique, le statut épigénétique et la localisation des cellules. De plus, il faut noter que la plupart des études ont été analysées chacune à l'aide de méthodes bioinformatiques différentes et une analyse croisée et minutieuse sera requise avant de tirer des conclusions solides[94].

2.6. Hématopoïèse : les perspectives futures

A la lumière des données de scRNA-Seq accumulées, le concept de CSH et progéniteurs clairement démarqués est remis en question. Il parait difficile de continuer à représenter la hiérarchie hématopoïétique par une succession de progéniteurs clairement définis à part pour des raisons de simplicité ou esthétiques. En effet, mêmes les marqueurs de surface utilisés pour l'isolement des progéniteurs présentent généralement des niveaux d'expression continues plutôt que discret.

La notion de différents types de progéniteurs a été historiquement dictée par des limitations techniques : capacité à observer des cellules en utilisant seulement un petit nombre de marqueurs et un nombre limité de tests fonctionnels. Par contre dans le paysage transcriptomique, chaque cellule est positionnée en utilisant l'information provenant de plusieurs milliers de gènes. Il convient de noter que le regroupement des cellules dans le paysage transcriptomique en une forme continue n'implique pas un manque d'informations. Au contraire il est évident que les positions dans cet espace sont fonctionnellement pertinentes et associées à des qualités fonctionnelles clés, telle que la persistance de l'autorenouvellement[95] et la différenciation [96] [97]. Cependant en raison de cette nature continue, les approches par cytométrie en flux séparent arbitrairement le paysage des progéniteurs isolant ainsi un mélange de cellules qui ont des propriétés fonctionnelles et de

différenciation différentes. Bien qu'utile, l'isolement de populations spécifiques par ce type d'approche n'offre qu'un aperçu restreint de l'organisation et de la dynamique des cellules souches et progénitrices.[94]

Les données transcriptomiques offrent une représentation plus complexe et probablement plus fidèle sans le besoin de distinguer des sous populations de manière subjective. Bien que les données de scRNA-Seq soient statiques, les informations encodées, le sont telles qu'elles existent in vivo. Le scRNA-Seq a le potentiel de capturer des états moléculaires représentatifs des transitions cellulaires. Cela signifie que pour chaque emplacement dans le paysage transcriptionnel, il peut devenir possible de déduire les directions de transition et les probabilités qui leur sont associées dans les conditions natives. La description quantitative de ce flux cellulaire à travers l'espace multidimensionnel constituera une avancée majeure et nécessitera plusieurs avancées techniques, analytiques, et expérimentales :

2.6.1. Une caractérisation plus fine des cellules

Le transcriptome à l'échelle unicellulaire constitue un cadre essentiel pour caractériser les cellules. Toutefois, celui-ci est limité aux informations de l'ARNm et manque d'informations potentiellement importantes tel que : l'expression protéique, l'épigénétique ou la position de la cellule au sein du tissu. La poursuite du développement de la technologie scRNA-Seq permettra l'analyse d'un plus grand nombre de cellules [98] (plusieurs millions) et une couverture plus importante du transcriptome[99]. Combiné à de nouvelles techniques, elle permettra la détection simultanée de l'expression protéique (CITE-seq[100] et REAP-Seq[101]), de l'ouverture de la chromatine[102], et de la méthylation de l'ADN[103]. La transcriptomique associée à l'imagerie a également été développée, permettant de compléter les données avec les informations spatiales[104]. Ensemble, ces informations permettront de localiser précisément les cellules dans un espace multidimensionnel et de relier ensemble les informations moléculaires, cellulaires et tissulaires.

2.6.2. Des outils analytiques avancés

Les nouvelles technologies décrites précédemment entrainent l'accumulation de données de plus en plus complexes dont l'analyse devient de plus en plus difficile. Cela comprend l'inférence de trajectoire, l'identification de points de ramification et l'extraction des informations concernant la régulation des gènes.

La description quantitative des flux cellulaires à travers les progéniteurs en est encore à ses balbutiements. Mais des méthodes bio-informatiques nouvellement développées tentent déjà d'approximer les transitions cellulaires à partir de données d'instantanés [105,35]. Néanmoins, les informations expérimentales parallèles restent essentielles pour fournir des informations en temps réel sur les processus de différenciation cellulaire.

2.6.3. De nouveaux outils expérimentaux

Plutôt que de caractériser les cellules souches et progénitrices en populations discrètes par immunophénotypage, une approche plus prometteuse consisterait à relier les positions des cellules dans le paysage transcriptionnel avec les informations expérimentales. Ainsi, chaque position encoderait des informations sur le flux cellulaire permettant la quantification de la différenciation et de l'auto-renouvellement. De nouvelles technologies permettant le barcoding *in vivo/in vitro* et le transcriptome à l'échelle unicellulaire de manière simultanée (un code barre peut être assigné aux cellules au sein des données scRNA-Seq) font leur apparition. Ces techniques utilisent soit le marquage par transposon[106] soit le marquage CRISPR[107]. Cela devrait permettre d'intégrer les informations d'ascendance cellulaire en temps réel avec les données transcriptomiques et potentiellement l'identification de signaux transcriptionnels générés uniquement transitoirement pendant la différenciation.

Une question pertinente est de savoir dans quelle mesure les souris de laboratoire maintenues dans des conditions stériles et exemptes d'agents pathogènes sont un modèle approprié pour l'hématopoïèse humaine. En effet, l'homme est constamment exposé à des agents infectieux et a une durée de vie beaucoup plus longue que la souris. Le suivi à long terme des patients ayant subi une greffe de moelle autologue a déjà révélé des aspects fonctionnels de l'hématopoïèse humaine inconnus auparavant, tels que le nombre, la stabilité, et la dynamique des CSH individuelles pendant de nombreuses années [108] Les

mutations somatiques silencieuses représentent des codes-barres uniques qui peuvent être exploités pour reconstruire l'architecture clonale des lignages et ainsi étudier la dynamiques des cellules souches et progénitrices humaines[109]. Les mutations de l'ADN mitochondrial peuvent également être un outil puissant dans ce domaine[110].

Pour des échelles de temps plus précises et un aperçu des effets liés au cycle cellulaire, les expériences de « pulse chase » devraient se révéler importantes.

Les réseaux de régulation des gènes et l'état d'accessibilité de la chromatine peuvent désormais être interrogés efficacement avec des techniques récemment développées combinant le criblage CRISPR avec le scRNA-Seq[111]. Cibler plusieurs gènes et observer les effets globaux permettra de comprendre comment l'ARN, les protéines, l'épigénétique et les signaux extrinsèques sont impliqués dans la différenciation et les flux cellulaires

2.6.4. Conclusion

Ces nouveaux outils présentent tous des limites intrinsèques, et aucune technologie à ce jour n'est susceptible de résoudre complètement la complexité de l'hématopoïèse. Au contraire c'est l'intégration de ces techniques moléculaires, génétiques et fonctionnelles à l'échelle unicellulaire, associée au développement d'outils bio-informatiques performants ; qui permettra de comprendre le rôle et la fonction des populations souches et progénitrices au cours de l'hématopoïèse physiologique, de stress et pathologique[112].

3. Désordre, hasard et différenciation cellulaire

3.1. Stochasticité de l'expression des gènes

Les populations clonales de cellules présentent une variabilité phénotypique intrinsèque. Cette hétérogénéité est essentielle à de nombreux processus biologiques et est supposée provenir de la stochasticité dans l'expression des gènes (ou caractère aléatoire de l'expression des gènes). La stochasticité de l'expression des gènes a été démontrée expérimentalement en 2002 par Elowitz et al[113]. Deux gènes rapporteurs codants pour des protéines fluorescentes ont été introduits dans des bactéries issues d'une même colonie, sous le contrôle du même promoteur. L'observation de la fluorescence des bactéries a alors révélé que les protéines fluorescentes n'étaient pas exprimées en quantité identique dans toutes les cellules, possédant pourtant le même fond génétique. Les cellules isogéniques soumises au même environnement présentent toujours des fluctuations aléatoires de l'expression des gènes. Cette variabilité aléatoire se produit entre les cellules d'une même population, un gène donné étant exprimé dans une cellule à un moment donné mais non exprimé dans une autre cellule. Pour cette raison, les gènes ne peuvent plus être considérés comme étant simplement actifs ou inactifs dans un type cellulaire donné ou dans un état cellulaire donné. Au lieu de cela, il faut leur attribuer une probabilité d'être actif (même si cette probabilité peut dans certains cas être proche de 1 ou 0) correspondant à ce qu'on appelle respectivement les états actif et réprimé dans la vision déterministe classique de l'expression des gènes[114].

Pour décrire ce concept, le terme "hasard" est souvent utilisé pour remplacer le mot "stochasticité", laissant penser que l'expression des gènes est soumise à un phénomène totalement aléatoire. Cependant, il est difficile d'envisager dans un organisme multicellulaire parfaitement structuré, qu'un tissu ou organe puisse exercer une fonction très précise, si le comportement des cellules qui le composent dépend d'une expression totalement aléatoire et chaotique des gènes. Il convient alors de préciser que le terme "hasard" fait référence à une expression probabiliste, et non pas anarchique, des gènes. On peut ainsi parler de hasard contraint ou de variabilité contrôlée. L'état de l'ouverture de la chromatine pourrait par exemple représenter une forme de contrainte dynamique de cette stochasticité. De façon imagée et simplifiée, on peut comparer la stochasticité de l'expression des gènes à un jeu de

dés. Les contraintes du dé sont liées à son architecture, composée de six faces planes parfaitement égales, offrant à chaque face exactement le même taux de probabilité de 1/6 d'être la face qui sera au-dessus du dé lors d'un lancer. Par analogie, la probabilité pour un gène d'être exprimé à un temps t dans une cellule serait de 1/x, où x peut prendre plusieurs valeurs en fonction des contraintes appliquées au système biologique à laquelle la cellule appartient. Une vision dite déterministe de l'expression des gènes représenterait alors un exemple extrême de cette représentation probabiliste et s'apparenterait à un système binaire où chacun des gènes possèderait une probabilité uniquement égale à 1 ou 0 de s'exprimer[115].

3.2. Entropie de Shannon

Il existe différents types et différentes définitions de l'entropie. Initialement, l'entropie a été définie en thermodynamique comme mesurant la quantité de « désordre » dans un système donné[116]. Claude Shannon a utilisé ce concept d'entropie dans la théorie de l'information[117]. L'entropie de Shannon mesure la quantité d'information délivrée par le système. Quand le message fourni par le système (par exemple une suite de chiffres) est uniforme, l'entropie est minimale. A l'inverse, plus le message fourni est hétérogène (désordonné et aléatoire), plus l'entropie est grande. Ainsi une répétition du chiffre 1 pioché aléatoirement parmi les chiffres de 1 à 9 aura une entropie de Shannon nulle. Au contraire, si le système fourni une information hétérogène comme une suite aléatoire de chiffres de 1 à 9, l'entropie de Shannon est maximale[115] **(Figure 18).**

L'entropie de Shannon est un des moyen utilisé pour mesurer la stochasticité de l'expression des gènes dans une population cellulaire, et ainsi permettre d'étudier son rôle dans les processus biologiques notamment la différenciation[118,119] **(Figure 19).** Pour calculer l'entropie de Shannon d'un gène dans une population cellulaire donnée (10 cellules dans cet exemple), on utilise la formule suivante : Soit H(X) l'entropie de Shannon du gène X mesuré sur les cellules numérotées de 1 à 10, soit n le nombre de valeurs que peut prendre l'expression du gène 1, et pi la probabilité pour que l'expression du gène 1 soit égale à i :

$$H(X) = -\sum_{i=1}^{n} pi \times \log(pi)$$

Ainsi, l'entropie est nulle si l'expression du gène X est la même dans toutes les cellules (pi = 1 donc log (pi) = 0), et celle-ci est élevée si le gène est exprimé différemment par les cellules **(Figure 19).**

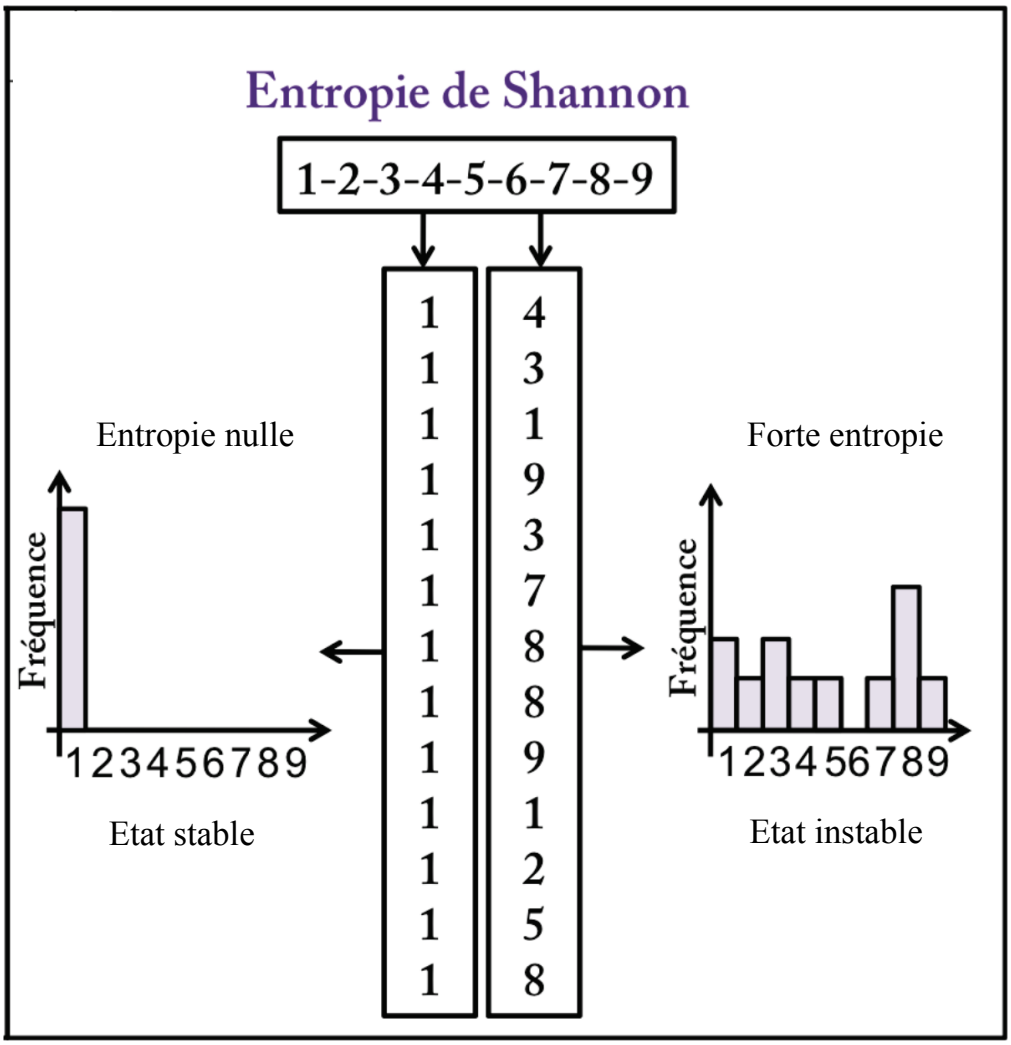


Figure 18 : Illustration de l'Entropie de Shannon[115].

Lorsque l'information donnée par un système est uniforme, l'entropie est nulle et le système est dans un état stable. Inversement lorsque l'information donnée par le système est aléatoire et hétérogène, l'entropie est forte et le système est instable.

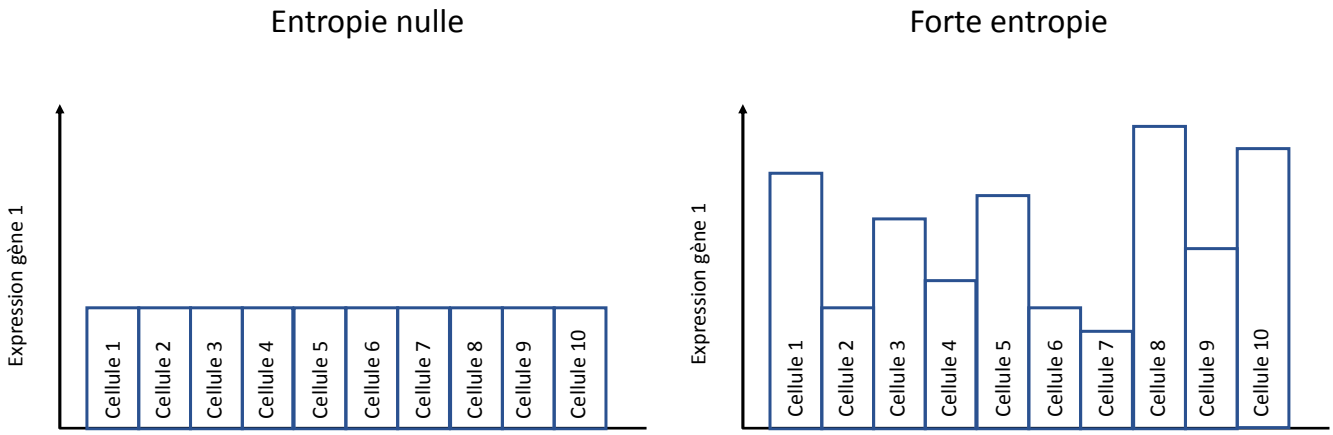


Figure 19 : Entropie de Shannon appliquée à l'expression d'un gène sur 10 cellules.

Lorsque toutes les cellules expriment le gène 1 de la même manière, l'entropie est nulle, au contraire, si le gène 1 prend des valeurs différentes d'expression pour chaque cellule, l'entropie est élevée.

3.3. Entropie et différenciation cellulaire

La stochasticité de l'expression des gènes mesurée par l'entropie de Shannon est impliquée dans des processus biologiques tels que la reprogrammation cellulaire[120], et le développement embryonnaire[121], suggérant ainsi un rôle dans la différenciation cellulaire. D'autres études sur les cellules souches hématopoïétiques ont mis en évidence l'importance de l'hétérogénéité moléculaire dans la différenciation[122,123], et également lors d'un processus de différenciation *ex vivo*[124].

Plusieurs modèles de différenciation cellulaire basés sur la stochasticité de l'expression des gènes ont été proposés, au cours desquels un pic de variabilité de l'expression génique devrait se produire. Dans le premier modèle la stochasticité de l'expression des gènes est la force motrice de la différenciation cellulaire qui permet de générer la diversité cellulaire, sur laquelle s'exerce alors une contrainte sélective[125]. Dans le second modèle, le bruit de fond dans l'expression des gènes provoque des bifurcations dans la dynamique des réseaux de régulation des gènes[126]. Dans le troisième modèle, la différenciation cellulaire est considérée comme un processus dynamique où les cellules représentées par des points se déplacent dans un espace multidimensionnel dans lequel chaque dimension correspond à un gène, le profil d'expression génique déterminant ainsi la position des cellules[127,128]. Les cellules humaines se déplacent par exemple dans un espace comprenant environ 30000 dimensions.

Par conséquent, si l'expression d'un gène change dans une cellule, celle-ci se déplacera dans l'espace multidimensionnel. La représentation d'un ensemble de cellules souches ou progénitrices en état d'auto-renouvèlement dans cet espace, consiste donc en un nuage de points, représentatif d'un profil d'expression génique donné. À cet instant, les cellules se situent dans un état dit "attracteur", dans lequel leur expression génique est, certes stochastique mais, stable. Le processus de transition vers un autre état attracteur (cellules différenciées par exemple) nécessite donc de sortir de l'état attracteur d'origine, ce qui peut être rendue possible par l'augmentation de la stochasticité de l'expression des gènes[129].

Quelles que soient les différences entre ces modèles, ils supposent tous que le processus de différenciation est représenté par des trajectoires cellulaires menant d'un état attracteur stable à un autre à travers une phase instable de modification aléatoire de l'expression des gènes. Les variations de l'expression des gènes étant stochastiques, chaque cellule emprunte un chemin différent, générant ainsi une augmentation significative de la variabilité intercellulaire. Cette phase est suivie d'une stabilisation (convergence) vers un modèle particulier d'expression génique correspondant à un état d'attracteur stable (état final différencié), dans lequel les fluctuations de l'expression génique sont minimisées par l'effet stabilisant de l'attracteur.

Par conséquent, les changements observés dans la variabilité cellule-cellule pourraient être une nouvelle métrique permettant de caractériser le processus de différenciation cellulaire[118].

3.4. Entropie et érythropoïèse

L'équipe d'Olivier Gandrillon a étudié la stochasticité de l'expression des gènes au cours de l'érythropoïèse[118]. Pour ce faire, ils ont utilisé un modèle physiologiquement pertinent de cellules primaires érythrocytaires aviaires pour analyser le niveau d'expression de 92 gènes dans des cellules individuelles recueillies à différents temps du processus de différenciation. Afin de tester l'hypothèse selon laquelle l'engagement en différenciation s'accompagne d'une forte augmentation de la variabilité de l'expression génique, ils ont mesuré l'entropie. Pour chaque temps de différenciation, ils ont calculé une valeur d'entropie par gène et comparé la distribution de ces valeurs au cours de la différenciation **(Figure 20).**

Ainsi, l'entropie augmente significativement à 8h, reste stable jusqu'à 24h, et diminue progressivement jusqu'à 72h, suggérant ainsi que la différenciation érythrocytaire aviaire s'accompagne d'un pic de la variabilité d'expression génique intercellulaire à 8h-24h du processus. Cette augmentation de la variabilité s'accompagne d'une forte chute dans le niveau de corrélation entre les gènes et précède l'engagement irréversible dans la différenciation.

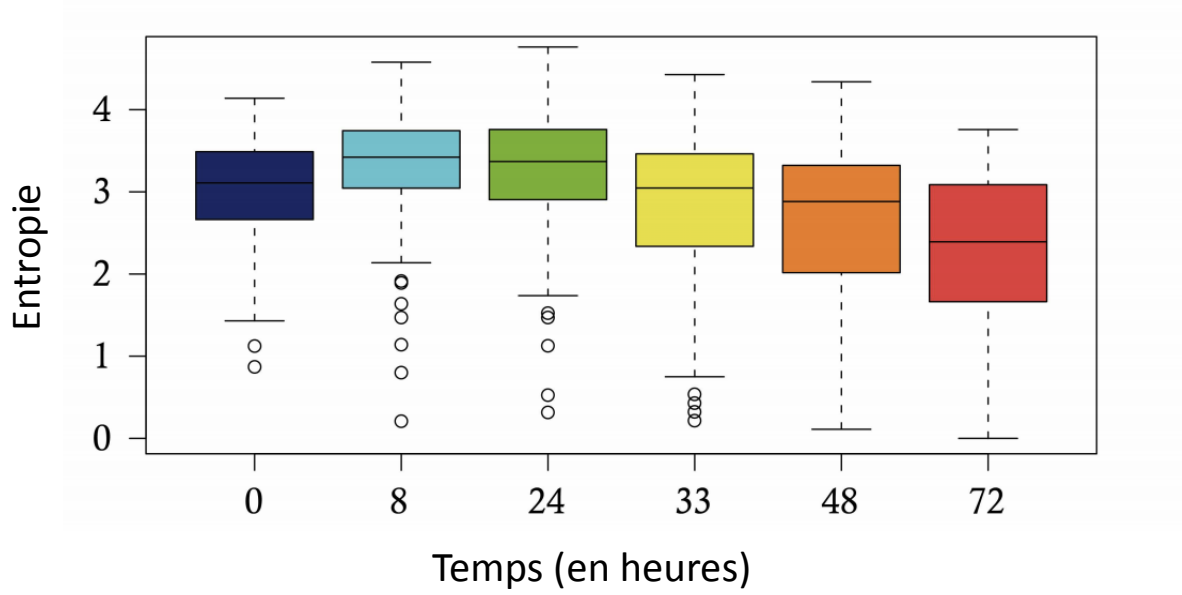


Figure 20 : Variation de l'entropie dans un modèle in vitro d'érythropoïèse aviaire[118]

Dans ce modèle, la variabilité de l'expression génique mesurée par l'entropie augmente au cours de la différenciation pour atteindre un pic puis redescend dans les cellules les plus matures

Les auteurs montrent pour la première fois dans cette étude l'implication de la stochasticité de l'expression des gènes pendant la différenciation. Ce concept prouvé expérimentalement est confirmé par d'autres études dans d'autres processus de différenciation[119,130,131]. Ces résultats soutiennent ainsi l'idée que la différenciation n'est pas une "simple" série d'événements moléculaires bien ordonnés et exécutés par toutes les cellules à l'identique, cette hypothèse ne pouvant expliquer ni l'augmentation constatée de variabilité, ni la baisse soudaine des corrélations. L'idée est donc que la différenciation résulte du comportement d'un réseau de gènes dynamique sous-jacent, qui reste à déterminer.

3.5. Entropie : cause ou conséquence de la différenciation ?

Les données expérimentales montrent clairement que la stochasticité de l'expression des gènes est un phénomène important dans la différenciation cellulaire. Pour préciser cette observation, l'équipe d'Olivier Gandrillon a étudié le rôle causal de cette stochasticité dans la différenciation[132]. Pour cela, ils ont manipulé expérimentalement les niveaux de stochasticité de l'expression génique pour analyser l'impact résultant de ces variations sur la différenciation cellulaire. Ils ont ainsi pu mettre en évidence un lien direct de cause à effet entre la variabilité de l'expression des gènes et le processus de différenciation érythrocytaire. En mesurant l'effet de drogues sur le niveau de stochasticité de l'expression génique pendant l'érythropoïèse aviaire, ils en ont sélectionné 3 : deux qui réduisent l'entropie de l'expression des gènes entre cellules (Artemisinine et Indomethacine) et une qui l'augmente (MB-3). Ainsi, l'artémisinine et l'indométhacine ont diminué la quantité de cellules différenciées au contraire du MB-3 qui a augmenté la quantité de cellules différenciées.

Cette étude fourni les preuves expérimentales que, dans un système cellulaire physiologiquement pertinent, la stochasticité de l'expression génique à un rôle causal dans la différenciation.

4. Les syndromes myélodysplasiques (SMD)

4.1. Diagnostic et classification

Les syndromes myélodysplasiques (SMD) sont des hémopathies clonales acquises de la cellule souche hématopoïétique médullaire, avec prolifération excessive de progéniteurs myéloïdes qui se différencient de manière anormale, on parle alors de dysmyélopoïèse. L'apoptose excessive intramédullaire des précurseurs anormaux aboutit à une hématopoïèse inefficace, c'est à dire un défaut de production des cellules sanguines matures et à des cytopénies périphériques.

La médiane d'âge au diagnostic est de 70 ans[133]. Le diagnostic se fait le plus souvent de manière fortuite après découverte de cytopénies périphériques lors d'un hémogramme. Dans la classification révisée OMS 2016 les cytopénies sont définies par un taux d'hémoglobine inférieur à 10 g/dL (anémie), et /ou un taux de polynucléaires neutrophiles inférieur à 1800/µL (neutropénie), et/ou un chiffre de plaquette inférieur à 100 000/µL (thrombopénie) [134]. Le diagnostic de SMD est basé sur des preuves morphologiques de dysplasie et un compte du pourcentage de blastes lors de l'analyse cytologique au microscope d'un frottis de moelle osseuse. Sur le frottis médullaire réalisé chez un patient SMD, on observe des dysplasies pouvant toucher la lignée mégacaryocytaire, la lignée granuleuse, la lignée érythroïde, deux de ces lignées ou toutes ces lignées. Un certain nombre de tests supplémentaires peuvent aider à affiner le diagnostic[135].

Le plus important est l'analyse cytogénétique de la moelle osseuse. Il est bien établi que les anomalies cytogénétiques sont très hétérogènes dans les SMD[136]. Elles sont retrouvées dans moins de 50 % des cas *de novo*, et sont plus fréquentes dans les SMD secondaires (jusqu'à 80 %). Les anomalies les plus fréquentes sont les délétions partielles et monosomies des chromosomes 5 (-5/5q-) et 7 (-7/7q-)[137]. L'analyse cytogénétique permet de calculer le pronostic des patients et dans certains cas de choisir la thérapie la plus efficace[138].

Les études par cytométrie en flux n'entrent pas encore dans les critères obligatoires pour le diagnostic des SMD, mais sont fortement recommandées pour affiner le diagnostic des SMD. Une combinaison d'un certain nombre de marqueurs des lignées érythroïdes, granuleuses et lymphoïdes sont recherchés, afin de mettre en évidence des expressions

aberrantes, des intensités de fluorescences différentes et / ou des asynchronismes de maturation pouvant aider ou conforter le diagnostic d'un SMD[139]. Il existe de nombreux scores publiés tels que l'Ogata et le Red Score. Le score d'Ogata a une bonne spécificité mais une mauvaise sensibilité. Ce score repose sur 4 critères : la dégranulation cellulaire, le pourcentage d'hématogones, le pourcentage de blastes, et l'intensité d'expression du CD45 sur les cellules CD34+. Si ce score est ≥ 2, celui-ci est en faveur d'un SMD [139], [140]. Le Red Score se base, quant à lui sur trois paramètres : le coefficient de variation de l'expression du CD71 (récepteur à la transferrine), du CD36 (récepteur de nombreux ligands tel que la thrombospondine, la fibronectine, le collagène, les acides gras…) par les érythroblastes, et le taux d'hémoglobine. Un score ≥ 3 est en faveur d'un SMD. En combinant le score d'Ogata et le Red Score, la sensibilité augmente[141]. Les examens de cytométrie en flux sont donc recommandés afin d'aider au diagnostic quand il existe des doutes sur la cytologie médullaire, et que le caryotype est normal.

La recherche de mutations somatiques dans les SMD peut également être utile pour compléter les autres outils de diagnostic. En effet, plus de 80% des patients atteints de SMD, seront porteurs d'une ou plusieurs mutations somatiques dans les gènes impliqués dans la méthylation de l'ADN, la conformation de la chromatine, l'épissage de l'ARN, la régulation de la transcription, la réparation de l'ADN , les cohésines, ou la transduction du signal [142,143]. Bien qu'aucune de ces mutations ne fasse partie des critères de diagnostic des SMD, à l'exception de la mutation *SF3B1* (qui définit les SMD avec sidéroblastes en couronne), certaines mutations ainsi que leur association sont étroitement liés à des sous-types de MDS spécifiques et pourraient être impliqués à l'avenir dans le diagnostic, le pronostic et la prise en charge thérapeutique[144].

Ainsi, la classification OMS 2016 distingue plusieurs types de SMD en fonction du nombre de cytopénies, des signes de dysplasie observés sur le myélogramme, de la présence de sidéroblastes en couronne, des anomalies cytogénétiques et du taux de blastes[134]. Les SMD avec excès de blastes de type 1 (MDS-EB1) présentent 5 à 9% de blastes dans la moelle ou 2 à 4 % dans le sang périphérique, et ceux de type 2 (MDS-EB2) 10 à 19% de blastes dans la moelle ou 5 à 19% dans le sang périphérique. Les SMD avec sidéroblastes en couronne (SMD-RS) doivent présenter à la coloration de Perls au moins 15% de sidéroblastes en couronne de type

III sur le frottis médullaire. S'il y a entre 5% et 15% de sidéroblastes en couronne, une mutation du gène *SF3B1* doit être mise en évidence. Les SMD avec une del(5q) isolée représentent une catégorie à eux seuls **(Figure 21).**

Maladie	Lignée(s) dysplasique(s)	Cytopénie	Aspect sanguin	Aspect médullaire
SMD avec dysplasie unilignée	1	1 ou 2	1 ou 2 cytopénie(s) blastes <1%	Dysplasie d'une lignée Blastes <5% Sidéroblastes en couronne (SC) <15%
SMD avec dysplasie multilignée	2 ou 3	1-3	Cytopénie(s) blastes <1%	Dysplasie de 2 ou 3 lignées Blastes <5% Absence de corps d'Auer SC<15%
SMD avec sidéroblastes en couronne			Cytopénie(s) blastes <1%	SC ≥ 15% ou ≥5% si SF3B1 muté Blastes < 5% Absence de corps d'Auer
SMD-SC avec dysplasie unilignée	1	1 ou 2		
SMD-SC avec dysplasie multilignée	2 ou 3	1- 3		
SMD-EB-1	0-3	1-3	Cytopénie(s) Blastes 2 à 4% Absence de corps d'Auer	Dysplasie uni- ou multilignée Blastes : 5-9% Absence de corps d'Auer
SMD-EB-1	0-3	1-3	Cytopénie(s) Blastes 5 à 19 % ou corps d'Auer	Dysplasie uni- ou multilignée Blastes : 10-19% ou corps d'Auer
SMD associé à une del(5q) isolée	1-3	1-2	blastes <1%	Mégacaryocytes à noyau hypolobé Blastes < 5% del(5q) +/- 1 autre anomalie sauf -7 Absence de corps d'Auer
SMD inclassable				SC <15% Blastes < 5% Absence de corps d'Auer
Avec 1% de blastes sanguin	1-3	1-3	1%	
Pancytopénie et 1 dysplasie	1	3	<1%	
Anomalie cytogénétique seulement	0	1-3	<1%	

Figure 21 : Classification OMS 2016 des syndromes myélodysplasiques[134].

La classification OMS 2016 est basée sur le nombre de cytopénies, les signes de dysplasie observés sur le myélogramme, la présence de sidéroblastes en couronne, les anomalies cytogénétiques et le taux de blastes médullaire et sanguin.

4.2. Scores pronostic

Les SMD sont des syndromes pré-leucémiques et se transforment en leucémie aigüe myéloïde de pronostic sombre dans 40% des cas. Le pronostic des patients peut être prédit par différents scores.

Le score IPSS (International Prognostic Scoring System) est basé sur le caryotype et le taux de blastes médullaires, le nombre et l'étendue des cytopénies sanguines, et classe les patients en quatre catégories. Au cours des deux dernières décennies, les patients ont été regroupés en fonction de ce score en SMD de bas risque (low et intermediate-1) et SMD de haut risque (intermediate-2 et high) [145].

La version révisée de l'IPSS (IPSS-R) utilisée aujourd'hui, stratifie davantage les patients, sur la base des mêmes facteurs, en cinq groupes de risque avec des résultats différents en termes d'évolution et de survie[146] **(Figure 22)**. En utilisant l'IPSS-R, un quart des SMD de bas risque selon l'IPSS classique ont été reclassés comme présentant un risque plus élevé, tandis qu'environ un cinquième des patients atteints de SMD de haut risque selon l'IPSS classique étaient reclassés comme présentant un risque inférieur. Par conséquent, bien que les médicaments actuellement disponibles soient souvent homologués sur la base de l'IPSS classique, la définition du risque basée sur l'IPSS-R est préférable : le terme SMD « de bas risque » s'applique donc généralement aux cas ayant un IPSS-R jusqu'à 3,5 ; incluant les catégories IPSS-R bas et très bas, ainsi qu'une partie des patients de la catégorie in-intermédiaire. Les SMD «de haut risque» inclurait alors les patients avec un IPSS-R supérieur ou égale à 4, c'est-à-dire un score de risque IPSS-R haut et très haut, ainsi que les autres patients IPSS-R intermédiaires[147] .

Catégorie cytogénétique	Anomalie cytogénétique
Très bon	-Y , del(11q)
Bon	Caryotype normale, del(5q), del(12p), del(20q) ou 2 anomalies dont del(5q)
Intermédiaire	del(7q), +8, +19, i(17q), toute autre anomalie simple ou double
Mauvais	-7, inv(3)/t(3q)/del(3q), deux anomalies dont -7/del(7q) caryotype complexe avec 3 anomalies
Très mauvais	caryotype complexe avec > 3 anomalies

Calcul du score de risque							
POINTS	**0**	**0,5**	**1**	**1,5**	**2**	**3**	**4**
Catégorie cytogénétique	Très bonne	-	bonne	-	interm	mauvaise	Très mauvaise
% de blastes médullaires	< ou = 2%	-	Entre 2 et 5%	-	5-10%	> 10%	-
Taux Hb	> ou = 10 g/dl	-	Entre 8 et 10 g/dl	< 8 g/dl	-	-	-
Chiffre plaquettes	> Ou = 100	Entre 50 et 100	< 50	-	-	-	-
Nb de PNN	> ou = 0,8	< 0,8	-	-	-	-	-

Catégorie de risque	Score de risque	Survie globale (années)
Très bas	≤ 1.5	9
Bas	> 1.5-3	5,5
Intermédiaire	> 3-4.5	2,9
Haut	> 4.5-6	1,7
Très haut	> 6	0,7

Figure 22 : Score IPSS-R méthode de calcul et groupes pronostic [146].

A partir des résultats du caryotype, de l'hémogramme et du myélogramme, le score IPSS révisé est calculé et permet d'estimer la survie des patients.

4.3. Génétique et physiopathologie

La cellule à l'origine de l'apparition de la maladie appartient au compartiment des cellules souches et progénitrices de la moelle osseuse (cellules CD34+). L'hypothèse actuellement défendue est qu'à la suite d' interactions complexes entre l'acquisition d'anomalies génétiques et épigénétiques, le microenvironnement médullaire, et le système immunitaire, un clone malin va se développer au détriment de l'hématopoïèse normale et ce sur plusieurs années[148].

4.3.1. Paysage mutationnel des SMD

Plus de 45 mutations somatiques pathogènes récurrentes ont été identifiées dans les SMD au diagnostic. Ces mutations touchent des gènes impliqués dans diverses processus cellulaires tel que la modification des histones, la réparation de l'ADN, la régulation épigénétique, l'épissage des ARN, la transcription, le remodelage de la chromatine et la signalisation[149] **(Figure 23).** Au moment du diagnostic, la plupart des cas de SMD présentent un paysage mutationnel complexe, avec de nombreux clones contenant plusieurs mutations coopérantes qui peuvent contribuer à la progression de la maladie et / ou à la rechute, bien que l'hématopoïèse soit généralement dominée par un clone spécifique[150]. Grâce à l' adoption généralisée de panels de séquençage ciblés, il est maintenant reconnu que plus de 80% des patients atteints de SMD sont porteurs d' au moins une mutation récurrente[142]. Si certaines mutations peuvent cohabiter et coopérer chez un même patient, certaines mutations sont mutuellement exclusives les unes des autres (par exemple, les mutations impliquées dans l'épissage)[151] . Les études de séquençage des SMD à l'échelle unicellulaire suggère des architectures clonales complexes, comme cela avait été prédit par les études de séquençage en « bulk » [152]. Ces informations sont importantes et auront dans l'avenir des implications pour l'identification des clones au diagnostic et leur suivi au cours de l'évolution de la maladie[153,154]. En effet, la dynamique de l'architecture clonale des SMD évolue au cours de la maladie et en réponse au traitement. Ainsi, les traitements peuvent entrainer une pression de sélection qui peut modifier les proportions relatives des clones existants et conduire à l'émergence de nouveaux clones résistants au traitement[155,156].

Fonction	Gènes
Méthylation de l'ADN	*DNMT3A, TET2, IDH1,* IDH2,** and *WT1*
Modification de la chromatine	*EZH2, SUZ12, EED, JARID2, ASXL1, KMT2, KDM6A, ARID2, PHF6,* and *ATRX*
Epissage de l'ARN	*SF3B1, SRSF2, U2AF1, U2AF2, ZRSR2, SF1, PRPF8, LUC7L2*
Complexe des cohésines	*STAG2, RAD21, SMC3,* and *SMC1A* (*PDS5B, CTCF, NIPBL,* and *ESCO2*)
Transcription	*RUNX1,† ETV6,† GATA2,† IRF1, CEBPA, BCOR, BCORL1, NCOR2,* and *CUX1*
Récepteur de cytokine/tyrosine kinase	*FLT3, KIT, JAK2,* and *MPL, CALR,* and *CSF3R*
Signalisation RAS	*PTPN11, NF1, NRAS, KRAS,* and *CBL* (*RIT1* and *BRAF*)
Signalisation autre	*GNAS, GNB1, FBWX7,* and *PTEN*
Cycle cellulaire/Checkpoint	*TP53* and *CDKN2A*
Réparation de l'ADN	*ATM, BRCC3,* and *FANCL*
Autres	*NPM1, SETBP1,* and *DDX41†*

*Seul les mutants IDH1 et IDH2 affectent la méthylation de l'ADN, pas les formes sauvages.
†Les mutations germinales sont impliquées dans la prédisposition aux SMD/LAM.

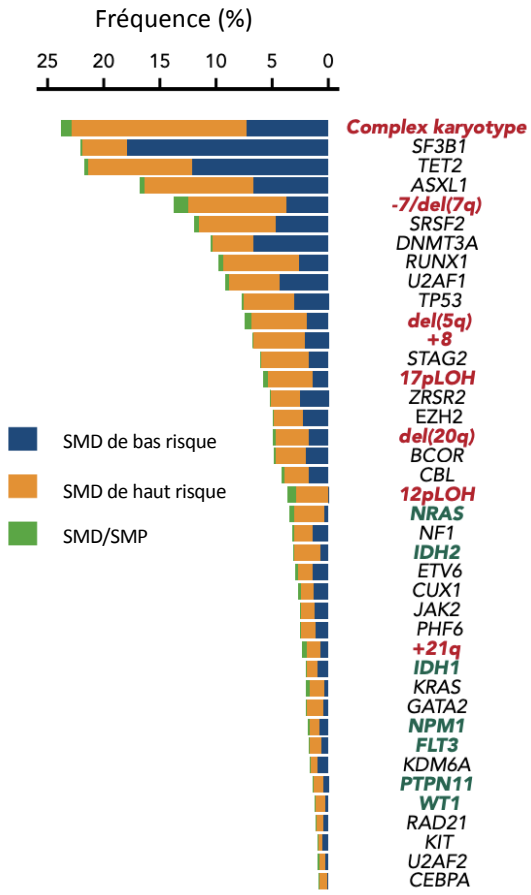


Figure 23 : Mutations somatiques et anomalies cytogénétiques majeurs retrouvées dans les SMD[149].

Les mutations somatiques retrouvées dans les SMD et formes frontières touchent des gènes impliqués dans diverses fonctions cellulaires. **SMP** : syndrome myéloprolifératif.

Deux gènes majeurs vont être associés aux études menées au cours de ma thèse :

a) Les mutations dans le gène de la sous-unité 1 du facteur d'épissage 3b (*SF3B1*), qui code pour une sous-unité clé du complexe multiprotéique d'épissage ou spliceosome, se révèlent être les mutations somatiques les plus fréquente dans le les SMD[157,158]. Les mutations *SF3B1* sont retrouvées plus fréquemment dans le SMD avec sidéroblastes en couronne (environ 80% des cas)[159]. Dans les SMD avec sidéroblastes en couronne, les mutations du gène *SF3B1* apparaissent dans la cellule souche hématopoïétique, puis se propagent dans le compartiment des progéniteurs myéloïdes plus différenciés mais aussi parfois dans le compartiment des progéniteurs lymphoides B[160]. Un mécanisme d'épissage perturbé conduit

par la suite à une synthèse protéique anormale au niveau des cellules hématopoïétiques ce qui conduit à l'hématopoïèse anormale et inefficace observée dans les SMD. Cependant, il reste difficile de savoir par quel mécanisme les CSH mutées sont sélectionnées pour l'expansion clonale [161]. Il a été précédemment démontré que les mutations *SF3B1* sont associées à la régulation à la baisse de voies de signalisation clés de la mitochondrie qui deviennent chargées de fer, et se disposent autour du noyau pour former les sidéroblastes en couronne. Récemment au laboratoire, une étude a montré que la mutation *SF3B1* entraine un épissage alternatif du gène *FAM132B* qui code pour l'érythroferrone. Cet épissage alternatif entraine la production d'une protéine anormale qui pourrait avoir un rôle dans la surcharge en fer observé dans les SMD mutés *SF3B1* [162]. Au niveau du phénotype clinique, les mutations de *SF3B1* sont associées à des cytopénies modérées, à une meilleure survie et à un risque de transformation en LAM diminué par rapports aux autres SMD [157].

b) Les mutations du gène *TET2* sont présentes chez près de 20 % des patients atteints de SMD et sont également observées dans les syndromes myéloprolifératifs (10 %), la leucémie myélomonocytaire chronique (30 à 50 %), et dans les LAM secondaire (25%) [163,164]. Elles ne sont pas étroitement associées à d'autres mutations ou anomalies cytogénétiques. La présence de mutations *TET2* dans les SMP indique qu'elles ne provoquent pas de dysplasie puisque la différenciation n'est pas altérée dans ces maladies. Elles ne sont pas non plus responsables des processus prolifératifs retrouvées dans les SMP et LAM, puisqu'on retrouve ces mutations dans les SMD de bas risque. Au contraire, les mutations *TET2* pourraient avoir une rôle pathogène commune à tous les cancers myéloïdes dans lesquels elles se trouvent, telle que l'établissement ou le renforcement de la dominance clonale de la cellule initiatrice de la maladie [165]. Une étude a noté que 26% des cas de SMD avec mutation TET2 possèdent une deuxième mutation de TET2, ce qui soutient la notion que la perte fonctionnelle biallélique de TET2 contribue probablement à la pathogénèse du SMD[166]. Aucun impact pronostique des mutations TET2 n'a été mis en évidence dans les SMD, mais la présence d'une mutation TET2 (en particulier en l'absence d'une mutation ASXL1) a été associée à une meilleure réponse aux agents déméthylants[167].

4.3.2. Modèle physiopathologique.

Pour être associé à des manifestations cliniques, un clone de SMD doit *a priori* passer par plusieurs étapes[168].

Si une seule mutation somatique peut contribuer à plusieurs étapes du processus, celui-ci nécessite l'accumulation de plusieurs anomalies génétiques. De plus, chaque étape du processus peut être associée à la présence de mutations touchant différents gènes. Il existe ainsi de multiples voies moléculaires permettant l'apparition de la maladie.

Les étapes associées à la pathogenèse du SMD comprennent (1) l'augmentation de l'auto-renouvellement d'une cellule souche hématopoïétique ou l'acquisition de la capacité d'auto-renouvellement dans une cellule progénitrice, (2) l'augmentation de la capacité de prolifération du clone SMD porteur ou de sa descendance plus différenciée, (3) un blocage ou une altération de la différenciation, (4) une instabilité génétique et épigénétique, (5) des mécanismes anti-apoptotiques,(6) une évasion vis-à-vis du système immunitaire et (7) la suppression de l'hématopoïèse normale.

La capacité d'auto-renouvellement doit être présente dans la cellule initiatrice de la maladie[169]. Cette cellule peut être une CSH (qui par définition possède cette capacité d'auto-renouvellement), ou alors un progéniteur myéloïde plus différencié ayant acquis une capacité d'auto-renouvellement. Ensuite, l'expansion du clone myélodysplasique peut être facilitée par l'acquisition de capacités de prolifération accrues, la résistance à l'apoptose, ainsi que par un microenvironnement médullaire anormal. Un SMD apparait lorsqu'au moins l'une des lésions moléculaires présentes dans le clone dominant ou dans son microenvironnement provoque également une différenciation dysplasique d'une ou plusieurs des lignées myéloïdes ce qui entraine une hématopoïèse inefficace. Le degré d'atteinte de chaque étape décrite précédemment a un impact sur les manifestations cliniques de la maladie, et sur sa rapidité d'évolution[165] (**Figure 24).**

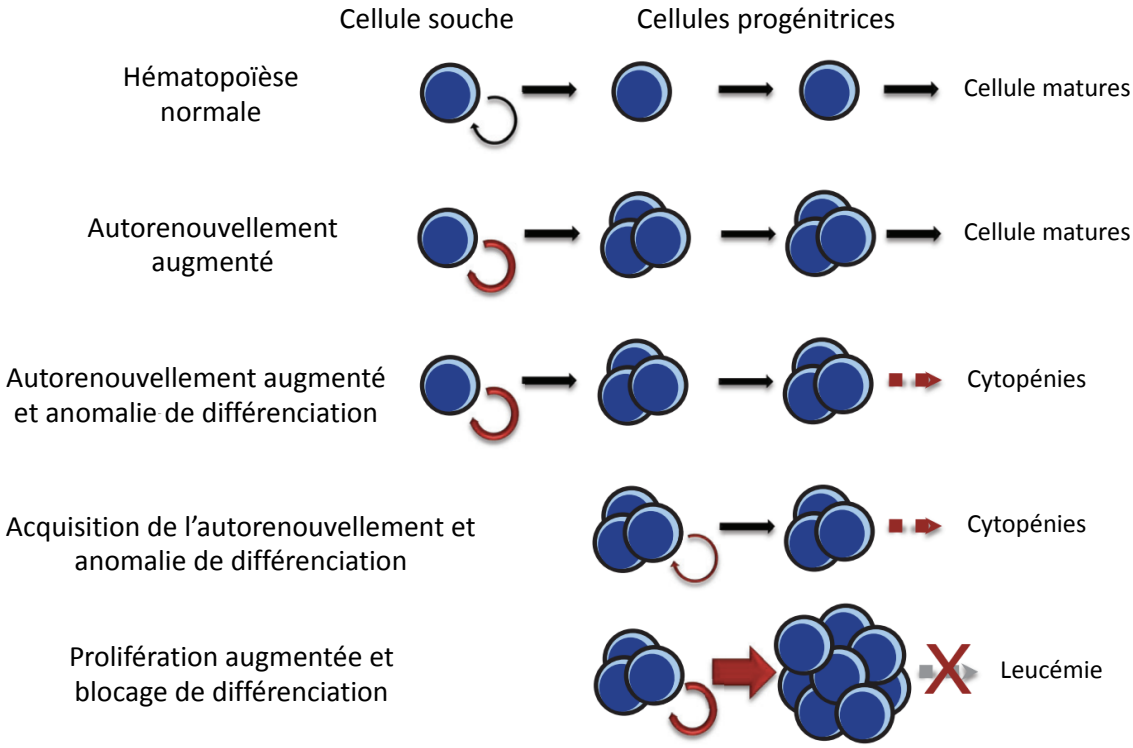


Figure 24 : Modèle de la physiopathologie des SMD[165].
Différentes voies vers la transformation. La caractéristique qui définit les syndromes myélodysplasiques est une hématopoïèse clonale et inefficace qui provoque des cytopénies. Un ensemble hétérogène d'anomalies génétiques et épigénétiques sont responsables de processus cellulaires à l'origine des phénotypes cliniques. Des lésions individuelles pourraient être responsable d'une seule étape de transformation (auto-renouvellement amélioré, anomalie de différenciation…) et être cliniquement silencieux. Ou elles pourraient entraîner plusieurs anomalies (par exemple, auto-renouvellement acquis et altération de la différenciation). La coopération entre deux ou plusieurs lésions est probablement nécessaire pour l'expression complète du phénotype de la maladie.

4.4. Traitements

Différents traitements sont utilisés dans la prise en charge des patients SMD, cependant l'efficacité reste modérée. Pour les patients de bas risque, le traitement vise à corriger les cytopénies, tandis que pour les patients SMD de haut risque, il vise à retarder la progression de la maladie en LAM et à prolonger la survie. Toutefois, la présence de certaines mutations peut également avoir un impact sur le choix et l'efficacité du traitement. En effet, la présence de clones mutés TP53 à une fréquence significative (> 10%) chez les patients SMD de bas risque avec del 5q (5q-) indique un risque élevé de progression de la maladie et une faible survie. Ainsi, cela suggère un renforcement du traitement ou l'allogreffe lorsque cela est possible. Au contraire, la présence d'une mutation *SF3B1* isolée dans les SMD avec

sidéroblastes en couronne prédit un taux faible de progression, ce qui conduit à éviter les chimiothérapies intensives ou autres traitements toxiques chez ces patients[147].

4.4.1. SMD de bas risque

Dans la plupart des cas, lorsque les cytopénies sont modérées et que le patient est asymptomatique, aucun traitement n'est requis. Ainsi, le traitement vise principalement à améliorer la ou les cytopénie(s) symptomatiques. L'anémie est généralement asymptomatique lorsque les taux d'hémoglobine restent supérieurs à 10 g/dL, bien que le seuil en dessous duquel se manifestent les symptômes de mauvaise tolérance varie en fonction de l'âge et des comorbidités du patient. La thrombopénie ne provoque généralement pas de saignement lorsque la numération plaquettaire est supérieure à 50 G/L et des infections sont rarement observées chez les patients atteints de SMD avec un nombre de polynucléaires neutrophiles supérieurs à 0,5 G/L. Par ailleurs, certains patients présentent des anomalies fonctionnelles des neutrophiles ou des plaquettes et peuvent présenter des infections ou des saignements au-dessus de ces seuils habituels.

L'anémie peut être corrigée transitoirement par des transfusions répétées de culots globulaires. Cependant, une telle attitude ne conduit qu'à une correction transitoire de l'anémie, une qualité de vie plus médiocre, nécessite des ressources hospitalières importantes (lits d'hôpitaux, etc.), entraine pour les patients une « dépendance » à l'égard du système hospitalier et conduit à une surcharge en fer. Il est donc préférable, dans la mesure du possible d' utiliser des médicaments pour augmenter le taux d'hémoglobine et d'éviter les transfusions de culots globulaire[170]. Les agents stimulant l'érythropoïèse ou ASE (érythropoïétines recombinante) sont donc les médicaments de première intention utilisés dans le traitement de l'anémie des patients SMD de bas risque[171]. En cas d'échec ou de résistance à l'action des ASE, des traitements de deuxième ligne peuvent être utilisés tel que le Lénalidomide (agent antinéoplasique, immunomodulateur)[172], certains immunosuppresseurs (sérum anti-lymphocytaire et la ciclosporine)[173] , ainsi que les agents déméthylants (5-azacytidine VIDAZA). Ceux-ci n'ont toutefois pas montré d'amélioration de la survie par rapport aux soins de support dans les SMD de bas risque[174]. Les traitements de deuxième intention ayant généralement des taux de réponse modérés, la plupart des patients atteints de SMD de bas

risque sont ensuite transfusés en culots globulaires. Le Déférasirox est alors le médicament chélateur du fer le plus utilisé pour les patients ayant une surcharge martiale (liée à la transfusion) avec une ferritine supérieure à 1000 ng/mL[175]. Enfin, un autre médicament, le Luspatercept, chef de file des agents de maturation de l'érythropoïèse (AME), a récemment montré une activité prometteuse dans un essai clinique de phase 3[176].

Dans les SMD de bas risque, la neutropénie et la thrombopénie sont moins profondes et moins fréquentes que l'anémie. En cas de neutropénie fébrile, la mise en place d'une antibiothérapie à large spectre est recommandée. L'utilisation de facteurs de croissance tel que le G-CSF est possible sur de courtes périodes, même si le risque de stimuler la progression vers un SMD de haut risque ou une LAM n'est pas complètement exclu. La thrombopénie si elle est symptomatique peut être traitée par des agonistes du récepteur de la thrombopoïétine (TPO)[177].

4.4.2. SMD de haut risque

Pour les SMD de haut risque l'allogreffe de cellules souches est le seul traitement curatif. Ainsi l'allogreffe permet de prolonger la survie prolongée chez 40% à 50 % des patients[178]. L'allogreffe peut généralement être proposée aux patients jusqu'à 70 ans. La décision d'un traitement par allogreffe est basée sur le score de risque de la maladie (score IPSS / IPSS-R) et l'évaluation des comorbidités. Néanmoins, seule une minorité de patients atteints de SMD de haut risque se voient proposer une allogreffe de moelle osseuse.

La chimiothérapie intensive (principalement basée sur des combinaisons anthracycline-cytarabine) dans les SMD de haut risque a été en partie éclipsée par l'avènement des agents hypométhylants[179]. Elle permet toutefois d'obtenir des rémissions plus complètes que les agents hypométhylants, et semble donc être une option intéressante chez les patients jeunes notamment en pré allogreffe de moelle osseuse.

Les agents hypométhylants tels que l'azacitidine (et le decitabine non approuvé en Europe) sont souvent les traitements de première intention dans la plupart des cas de SMD de haut risque[179]. La réponse à l'azacitidine est souvent retardée avec un délai médian de réponse de 3 mois. Il faut donc attendre au moins 6 cures (soit 6mois) pour juger l'efficacité

de ce traitement. La réponse peut être prédite par les résultats du caryotype, les besoins transfusionnels, la présence de blastes circulants ou de certaines mutations somatiques[180,181]. Après échec de l'azacitidine, si l'allogreffe n'est pas envisageable, les pronostics sont sombres avec une survie médiane de 5 à 6 mois[147].

Des mutations des gènes IDH1 et IDH2 sont présentes dans 5 à 10% des SMD, avec une prédominance dans les SMD de haut risque, et sont associées à un gain de fonction conduisant à l'accumulation du 2-hydroxyglutarate (onco-métabolite) dans les cellules hématopoïétiques[182]. Les inhibiteurs d'IDH1 et d'IDH2 ont fait la preuve de leur efficacité dans les LAM et sont actuellement testés dans les SMD.

La mutation TP53 est présente dans environ 10% des SMD de haut risque et dans 50% de ceux ayant un caryotype complexe. Elle est associée à des taux de réponse plus faibles et des durées de réponse plus courte aux agents déméthylants[183]. Des études cliniques évaluent actuellement l'innocuité et l'efficacité de l'APR-246 (un composé permettant de reconformer la protéine TP53 mutante) en combinaison avec de l'azacitidine dans les SMD mutés TP53. Les résultats préliminaires sont prometteurs mais nécessitent une confirmation[184].

Objectifs du travail de thèse

A mon arrivée au laboratoire, un important travail sur l'architecture clonale et l'hétérogénéité des syndromes myélodysplasiques avait été effectué par la précédente étudiante en thèse[152,155]. J'ai pu prendre connaissance de ce travail à travers l'écriture d'une revue sur l'hétérogénéité des SMD[185]. Ainsi, l'objectif de mon travail de thèse était de poursuivre ces travaux permettant d'avancer dans la compréhension de la physiopathologie des SMD. Pour cela, nous avons choisi de nous focaliser sur le compartiment HSPC des SMD, et d'utiliser des approches à l'échelle unicellulaire telles que la cytométrie en flux et le scRNA-Seq. L'analyse des données de scRNA-Seq nécessitant des compétences en bioinformatique, j'ai eu l'opportunité de me former dans ce domaine afin d'analyser les résultats expérimentaux obtenus. Nous avons débuté au cours de ma thèse, une collaboration avec Olivier Gandrillon et Pierre Sujobert de l'université de Lyon dans le but d'étudier et de comparer la variabilité de l'expression génique dans un contexte d'hématopoïèse normale et pathologique jusque-là inexploré. Les résultats de ce travail font actuellement l'objet de la préparation d'un manuscrit.

Les parties résultats et discussion de mon manuscrit seront orientées sur cet axe principal de mon travail.

Parallèlement à ce travail j'ai eu la chance de participer à l'obtention et à l'exploitation de données moléculaires issues de xénogreffe de SMD dans un modèle de souris NSG avec l'équipe de Sophie Park[186]. J'ai également été impliqué dans la mise au point d'un modèle de xénogreffe de SMD par utilisation de niches humanisées en collaboration avec l'équipe de Françoise Pflumio au CEA. J'ai aussi pu prendre part aux expériences de scRNA-Seq et à l'analyse bioinformatique d'une étude menée par l'équipe de Nathalie Drouin à l'IGR, sur le compartiment des cellules stromales de patients porteurs de LMMC. Enfin, j'ai été impliqué dans les expériences de scRNA-Seq d'une étude récemment publiée permettant de distinguer les sujets atteints de formes modérées et sévères du COVID 19[187].

Résultats partie 1 : Etude du compartiment souche CD34+ des SMD par cytométrie en flux

1. Introduction

Les syndromes myélodysplasiques (SMD) sont des pathologies hétérogènes clonales de la cellule souche hématopoïétique (CSH). Les CSH qui sont au sommet de la hiérarchie hématopoïétique normale, sont porteuses de mutations somatiques dans les SMD. Ces mutations, associées à d'autres facteurs environnementaux, entrainent une dysplasie cytologique des lignées myéloïdes, conduisant à des cytopénies. Dans 30% des cas, les SMD évoluent en leucémie aiguë myéloïde de pronostic sombre.

Dans la moelle osseuse, les CSH génèrent des progéniteurs multipotents (MPP) possédant une activité d'auto-renouvellement réduite. Les MPP vont perdre leur capacité de multipotence en générant d'abord une série de progéniteurs oligopotents puis unipotents et enfin tous les types de cellules matures hématopoïétiques. Dans leur étude[82] Notta et ses collaborateurs affirment qu'il existe dans la moelle normale trois catégories différentes de MPP (MPP1,2,3) qui donnent naissance à des progéniteurs myéloïdes unipotents (CMP1,2,3; MEP1,2,3 et GMP). Cette première partie de mon travail a eu pour but d'établir la présence de ces populations dans le compartiment HSPC de moelles issues de sujets âgés et de patients SMD, et si, le cas échéant, l'étude de leur répartition peut aider au diagnostic, et à prédire l'évolution de la maladie.

2. Matériel et Méthodes

2.1. Recueil et conservation des cellules

Les échantillons médullaires des patients atteints de SMD ont été obtenus au décours d'une ponction de moelle osseuse (dans le sternum ou l'os iliaque) réalisée dans le cadre du diagnostic ou du suivi de la maladie dans le service d'hématologie biologique de l'hôpital Cochin. Suite au prélèvement réalisé sur un tube Hanks hépariné, les cellules mononucléées (CMN) ont été isolées sur un gradient de Ficoll. Les échantillons de moelle osseuse issus de patients sains âgés ont été obtenus par extraction des cellules médullaires issues de l'os de la tête fémorale. Les têtes fémorales sont obtenues après consentement éclairé, au décours d'une chirurgie pour pose de prothèse de hanche. Celles-ci sont coupées en deux et recueillies dans un milieu de conservation (Hanks balanced salt solution with NaHCO3, Eurobio™), supplémenté en héparine puis acheminées au laboratoire à température ambiante. Celles-ci sont grattées avec une spatule, broyées dans un mortier, et lavées avec une solution de PBS additionnée de DNAse à 100 ug/mL. Les CMN sont alors, comme pour les cellules issues des ponctions médullaires des patients SMD, isolées sur un gradient de Ficoll, et lavées dans une solution de PBS.

Une fois isolées, les CMN de moelle osseuse sont comptées et resuspendues dans une solution adéquate (IMDM, SVF 40%, DMSO 15%), puis congelées et conservées dans l'azote liquide à une concentration de 20 à 30 millions de cellules par ampoule de 1 mL.

2.2. Cytométrie en flux

Les CMN sont décongelées dans du milieu de culture IMDM à 37°C supplémenté en DNAse (1ug/mL) pour éviter la présence d'agrégats et limiter ainsi la perte de cellules. Les CMN sont ensuite lavées avec du PBS puis triées sur colonnes magnétiques (technologie MACS MicroBead de Miltenyi Biotec™) afin de récupérer les cellules CD34 positives. Les cellules CD34 positives sont marquées à l'abri de la lumière, et incubées pendant 20 min dans un tampon Brillant Stain Buffer, BD Horizon™ (100µL par tube) avec le panel d'anticorps présenté dans le **tableau 1**. Les cellules sont ensuite lavées avec du PBS (tampon Dulbecco's PBS™) et un deuxième marquage dans du PBS est réalisé à l'abri de la lumière pendant 20 min avec du

Zoombie acqua Amcyan (BioLegend™) pour distinguer les cellules mortes et vivantes. Le phénotypage est effectué sur l'automate LSR Fortessa™ (BD Biosciences).

Tableau 1 : caractéristiques des anticorps utilisés pour la caractérisation du compartiment des HSPC normales et pathologiques par cytométrie en flux.

Cible	Fluorochrome	Isotype	Réactivité	Clone	Dilution	Fournisseur
CD34	Pc7	Souris IgG1	Humain	581	1/100	Beckman Coulter
CD38	PECF564	Souris IgG1, κ	Humain	HIT2	1/100	BD Horizon
CD90	Bv421	Souris IgG1, κ	Humain	5E10	1/50	BD Horizon
CD45RA	Buv737	Souris IgG2b,κ	Humain	HI100	1/50	BD Horizon
CD71	Bv786	Souris IgG2a, κ	Humain	M-A712	1/50	BD Horizon
CD110	PE	Souris IgG1, κ	Humain	BAH-1	1/50	BD Pharmingen
CD7	APC Vio770	Souris IgG2a, κ	Humain	6B7	1/50	MACS Miltenyi Biotec
CD10	PerCP CY5.5	IgG1 Recombinant	Humain	REA877	1/50	MACS Miltenyi Biotec
CD123 (Flt3)	Bv711	Souris IgG1, κ	Humain	9F5	1/50	BD Horizon
CD49f	AlexaFluor 647	Rat IgG2a, κ	Humain	GoH3	1/25	BD Pharmingen
CD2	FITC	Souris IgG2a	Humain	S5.2	1/100	BD Biosciences
CD3	FITC	Souris IgG1, κ	Humain	SK7	1/100	BD Biosciences
CD4	FITC	Souris IgG1, κ	Humain	SK3	1/100	BD Biosciences
CD8	FITC	Souris IgG1, κ	Humain	SK1	1/100	BD Biosciences
CD11b/MAC1	FITC	Souris IgG1, κ	Humain	ICRF44	1/100	BD Pharmingen
CD14	FITC	Souris IgG2b, κ	Humain	MφP9	1/100	BD Biosciences
CD16	FITC	Souris IgG1, κ	Humain	NKP15	1/100	BD Biosciences
CD19	FITC	Souris IgG1, κ	Humain	4G7	1/100	BD Biosciences
CD235a	FITC	Souris IgG2b, κ	Humain	GA-R2	1/100	BD Pharmingen

2.3. Analyse des données de cytométrie en flux.

Les fichiers .FCS ont été analysés avec le logiciel Kaluza (Kaluza Analysis Software, Beckman Coulter Life Sciences) puis grâce à la plateforme en ligne d'outils Cytobank (https://www.cytobank.org). La méthode de réduction dimensionnelle par t-SNE (t-distributed stochastic neighbor embedding) et de clustering FlowSOM[188] ont été appliquées aux données regroupées de tous les échantillons. La classification hiérarchique des échantillons en fonction de la répartition des clusters FlowSom a été réalisé dans R avec la fonction hclust du package heatmap.2.

2.4. Calcul de l'entropie de la répartition cellulaire.

Pour évaluer l'entropie de Shannon de la répartition des cellules au sein des clusters, nous avons utilisé la formule suivante dans le langage statistique R. Soit H(X) l'entropie de Shannon de la répartition des cellules d'un patient au sein d'un nombre de clusters déterminés, soit k le nombre de valeurs que peut prendre le pourcentage de cellules au sein d'un cluster, et pi la probabilité pour que le pourcentage de cellules au sein du cluster soit égale à i, avec k = nombre de clusters/2 :

$$H(X) = -\sum_{i=1}^{k} pi \times \log(pi)$$

3. Résultats

Nous avons appliqué le panel publié par Notta[82] aux cellules CD34+ de 36 échantillons **(Tableau 2)** répartis en 9 échantillons de sujets âgés (CTRL), 17 échantillons de SMD de bas risques (LRMDS) et 10 échantillons de SMD de haut risque (HRMDS) comme définis précédemment[147] **(Figure 22).** Pour la première fois, nous avons pu ainsi mettre en évidence la présence des différentes sous populations décrites dans des échantillons de SMD. Les cellules CD34+ identifiées sont séparées en deux compartiments en fonction de l'expression du CD38. Le compartiment CD34+CD38- se divise en quatre populations : les CSH CD90+ CD45RA- CD10-, les MPP (progéniteurs multipotents) CD90- CD45RA- CD10-, les LMPP (progéniteurs multipotents lympho-myéloïde) CD90- CD45RA+ CD10-, et les MLP (progéniteurs multilymhpoïde) CD45RA+ CD10+. Les MPP sont subdivisés en MPP1 (CD71- CD10-), MPP2 (CD71+ CD110-), et MPP3 (CD71+ CD110+). Dans le compartiment CD34+CD38+, les cellules CD10- comprennent les GMP (progéniteur granulocytaire/macrophagique) CD45RA+, CMP (progéniteur myéloïde commun) CD45RA- CD123- et MEP (progéniteur mégacaryocytaire/érythrocytaire) CD45RA- CD123-, tandis que les cellules CD10+ sont les progéniteurs B/NK. Les sous populations CMP1, 2, 3 et MEP1, 2, 3 se définissent en fonction de l'expression des marqueurs CD71 et CD110 (respectivement double négatif, simple positif CD71, double positive) **(Figure 25).**

Tableau 2 : caractéristiques clinico-biologiques des 36 patients étudiés en cytométrie en flux, selon le panel de Notta et al[82].

Echantillon	Age	Sexe	WHO 2016	Hb	VGM	Plaquettes	PNN	Blastes médullaires	Caryotype/FISH	Score R-IPSS	Risque R-IPSS	Classification
CTRL1	58	F	CTRL	NA	NA	NA	NA	NA	NA	NA	NA	CTRL
CTRL2	67	H	CTRL	NA	NA	NA	NA	NA	NA	NA	NA	CTRL
CTRL3	68	F	CTRL	NA	NA	NA	NA	NA	NA	NA	NA	CTRL
CTRL4	72	F	CTRL	NA	NA	NA	NA	NA	NA	NA	NA	CTRL
CTRL5	45	F	CTRL	NA	NA	NA	NA	NA	NA	NA	NA	CTRL
CTRL6	83	F	CTRL	NA	NA	NA	NA	NA	NA	NA	NA	CTRL
CTRL7	78	F	CTRL	NA	NA	NA	NA	NA	NA	NA	NA	CTRL
CTRL8	56	F	CTRL	NA	NA	NA	NA	NA	NA	NA	NA	CTRL
CTRL9	74	H	CTRL	NA	NA	NA	NA	NA	NA	NA	NA	CTRL
LRMDS1	94	F	SMD-EB1	11	91	65	0,56	8	46, XX[20]	3,5	intermediate	LRMDS
LRMDS2	90	H	SMD dysplasie multilignée	10,2	91	156	3,12	2	45, X,-Y [4]/46,XY [16]	0	very low	LRMDS
LRMDS3	62	F	SMD dysplasie multilignée	11	87	78	1,7	3	NA	NA	NA	LRMDS
LRMDS4	77	H	SMD dysplasie multilignée	10,4	111	61	NA	4	46, XY[20]	2,5	low	LRMDS
LRMDS5	84	H	SMD-EB1	8,3	79	717	35,93	6	46, XY[20]	4	intermediate	LRMDS
LRMDS6	84	H	SMD dysplasie multilignée	10	103	60	1,04	3	46, XY[20]	2,5	low	LRMDS
LRMDS7	64	H	SMD dysplasie unilignée	14,6	96,7	181	1,59	2	NA	NA	NA	LRMDS
LRMDS8	82	H	SMD avec sidéroblastes en couronne	10,6	89,3	287	3,46	1	46, XY[20]	1	very low	LRMDS
LRMDS9	86	H	SMD-EB1	13	106	156	1,6	7	46, XY[20]	3	low	LRMDS
LRMDS10	79	F	SMD dysplasie unilignée	11,6	95	301	4,4	3	46, XX[20]	1	very low	LRMDS
LRMDS11	62	F	SMD avec del5q	NA	NA	NA	NA	3	46,XX[20].ish 5q31(D5S1518E-D5S1976,EGR1,RPS14)x2[5].nuc ish(D5S1518E-D5S1976,EGR1,RPS14)x2[251/264]	NA	NA	LRMDS
LRMDS12	87	F	SMD avec sidéroblastes en couronne	10	94	342	1,4	3	46,XX,del(11)(q14][3]/46,sl,?del(20)(q12)[2]/ 46,X,del(X)(q21),add(5)(q?31),del(11)(q14), add(17)(q2?2)[12]/46,XX[2]	4	Intermediate	LRMDS
LRMDS13	58	F	SMD dysplasie unilignée	11,3	67,3	142	2,02	2	46, XX[20]	2	low	LRMDS
LRMDS14	70	F	SMD dysplasie multilignée	8,8	85	253	2	3	46, XX[20]	2	low	LRMDS
LRMDS15	74	H	SMD dysplasie multilignée	10,2	103	123	3,01	3	NA	NA	NA	LRMDS
LRMDS16	67	H	SMD dysplasie multilignée	12,9	96	136	11,1	2	46, XY[20]	3	low	LRMDS
LRMDS17	91	F	SMD dysplasie unilignée	11,7	93,6	136	1,27	1	46,XX,ider(20)(q10)del(20)(q11q13)[20]	1	very low	LRMDS
HRMDS1	57	F	SMD-EB1	10,8	79	NA	1,5	2	45,XX,-7[15]/46,XX[4]	5	high	HRMDS
HRMDS2 (évolution de LRMDS1)	94	F	SMD EB2	7,8	97	53	7,2	13	46, XX[20]	6	high	HRMDS
HRMDS3	74	F	SMD-EB1	NA	NA	NA	NA	9	47, XX,+8,9ph[3]/46, XX,9ph[17]	5	high	HRMDS
HRMDS4	67	H	SMD-EB2	11	106	65	3,27	13	46,XY,t(2;12)(q24;q23),del(6)(q15q23),?t(6;20)(q12;q11), del(11)(q22),add(21)(q21)[20].ish del(11)(KMT2A-)[9] .nuc ish(KMT2Ax1)[90/134]	7,5	very high	HRMDS
HRMDS5	81	F	SMD-EB2	13,6	103	157	0,39	13	46,XX,del(5)(q14q34)[5]/46,XX[16] .unc ish(D5S1518E-D5S176x2,EGR1x1,RPS14x1[16/100] (D7Z1,D8Z1)x2[100]	4,5	intermediate	HRMDS
HRMDS6	80	H	SMD EB2	8,4	97	10	7,63 (12% blastes circulants)	11	45,XY,del(3)(p12p26),del(9)(q12q31) ,-12-16,-20,+22[6]/46,idem,+mar[13]/46,XY[1]	9	very high	HRMDS
HRMDS7	69	H	SMD-EB2	10,1	89	21	34,2 (myélémie : 5% blastes, 4% myélo, 3% méta)	13	NA	NA	NA	HRMDS
HRMDS8	63	F	SMD-EB2	8,9	94	12	3,35	17	46,XX,der(14)t(14:21)(pll;q22),der(15)t(15:21)(pll:q22), der(21)t(21:21) (Pl 1 ;q22)dup(2 1) (q22q?22)[I 5]/?46,XX[5]	7	very high	HRMDS
HRMDS9	77	H	SMD-EB1	9,1	84,2	243	0,39	7	46, XY[20]	4,5	intermediate	HRMDS
HRMDS 10	67	F	SMD-EB1	7,6	104	215	3,01	8	46,XX,del(5)(q13q33)[5]/47,idem,+21[10]/46, XX[4]	4,5	intermediate	HRMDS

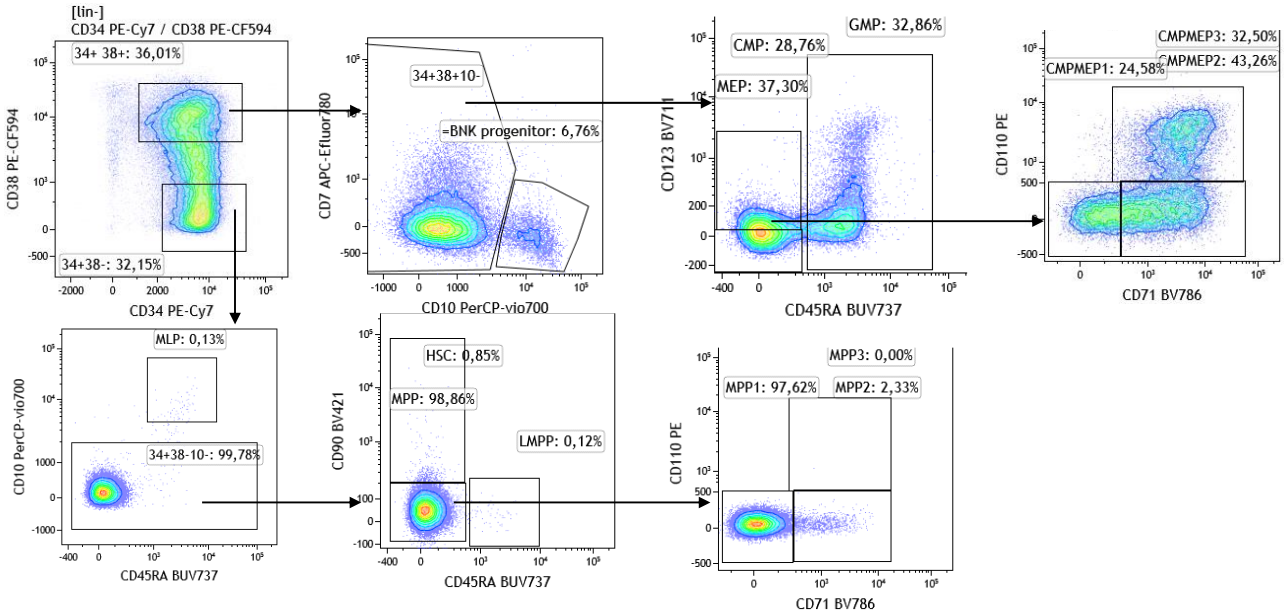


Figure 25 : Exemple d'application du panel de cytométrie en flux d'étude du compartiment CD34 + décrit par Notta et al [82].

Exemple de répartition des sous populations du compartiment CD34+ de la moelle osseuse d'un sujet sain âgé.

La séparation manuelle des sous populations à partir de l'expression des marqueurs est parfois difficile lorsque les nuages de points sont continus. Par exemple, il est difficile de placer la fenêtre séparant les cellules CD38+ des cellules CD38- de manière reproductible. Ce manque de reproductibilité ne permet pas de comparer correctement la répartition des différentes sous populations entre les échantillons. Nous avons donc choisi d'utiliser des techniques de « machine learning » (apprentissage automatisé) pour interpréter nos données. Dans un premier temps, nous avons utilisé le t-SNE sur la totalité des cellules (995616 évènements) de tous nos échantillons pour visualiser nos données en deux dimensions. Ensuite FlowSOM nous a permis de clusteriser les cellules afin de les séparer en sous populations homogènes **(Figure 26, Figure 27).**

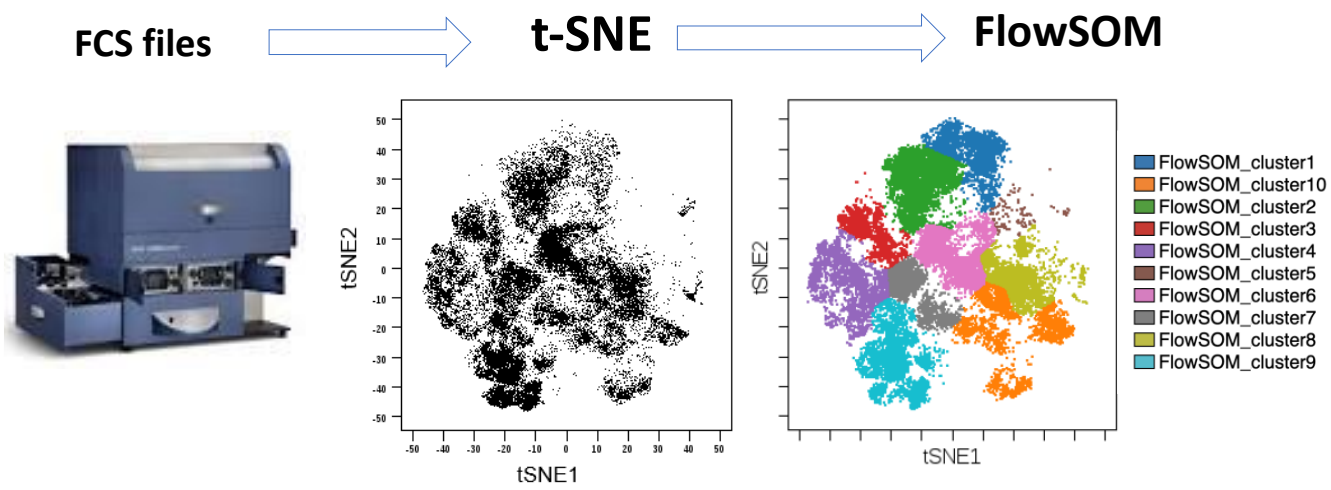
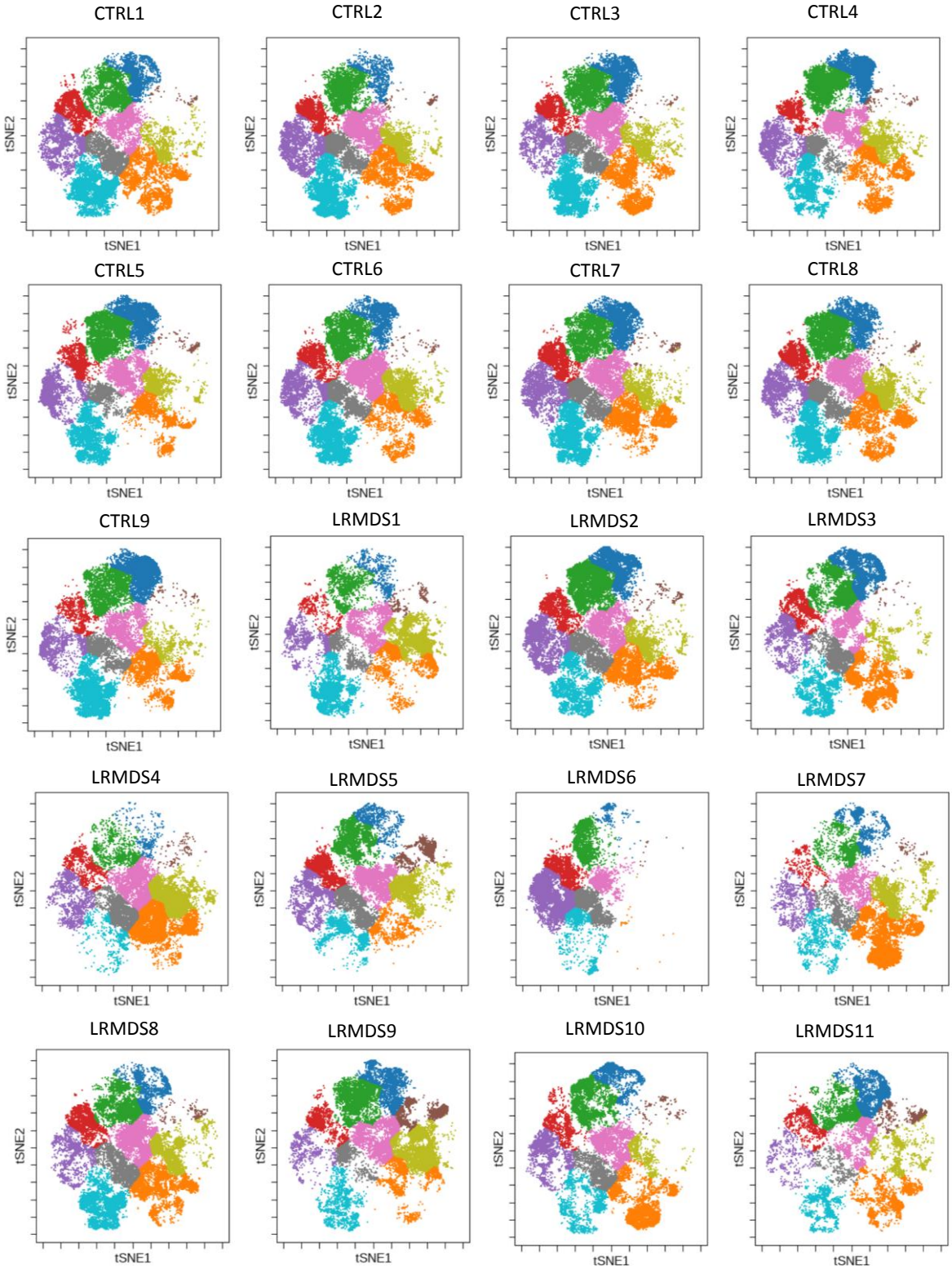


Figure 26 : Description du la stratégie d'analyse des données de cytométrie en flux.

Les fichiers FCS de tous les échantillons ont été regroupés, un t-SNE a été réalisé pour réduire la dimensionnalité, puis l'algorithme de clustering FlowSom a été appliqué sur les données réduites.


CTRL1
CTRL2
CTRL3
CTRL4
CTRL5
CTRL6
CTRL7
CTRL8
CTRL9
LRMDS1
LRMDS2
LRMDS3
LRMDS4
LRMDS5
LRMDS6
LRMDS7
LRMDS8
LRMDS9
LRMDS10
LRMDS11
tSNE1
tSNE2

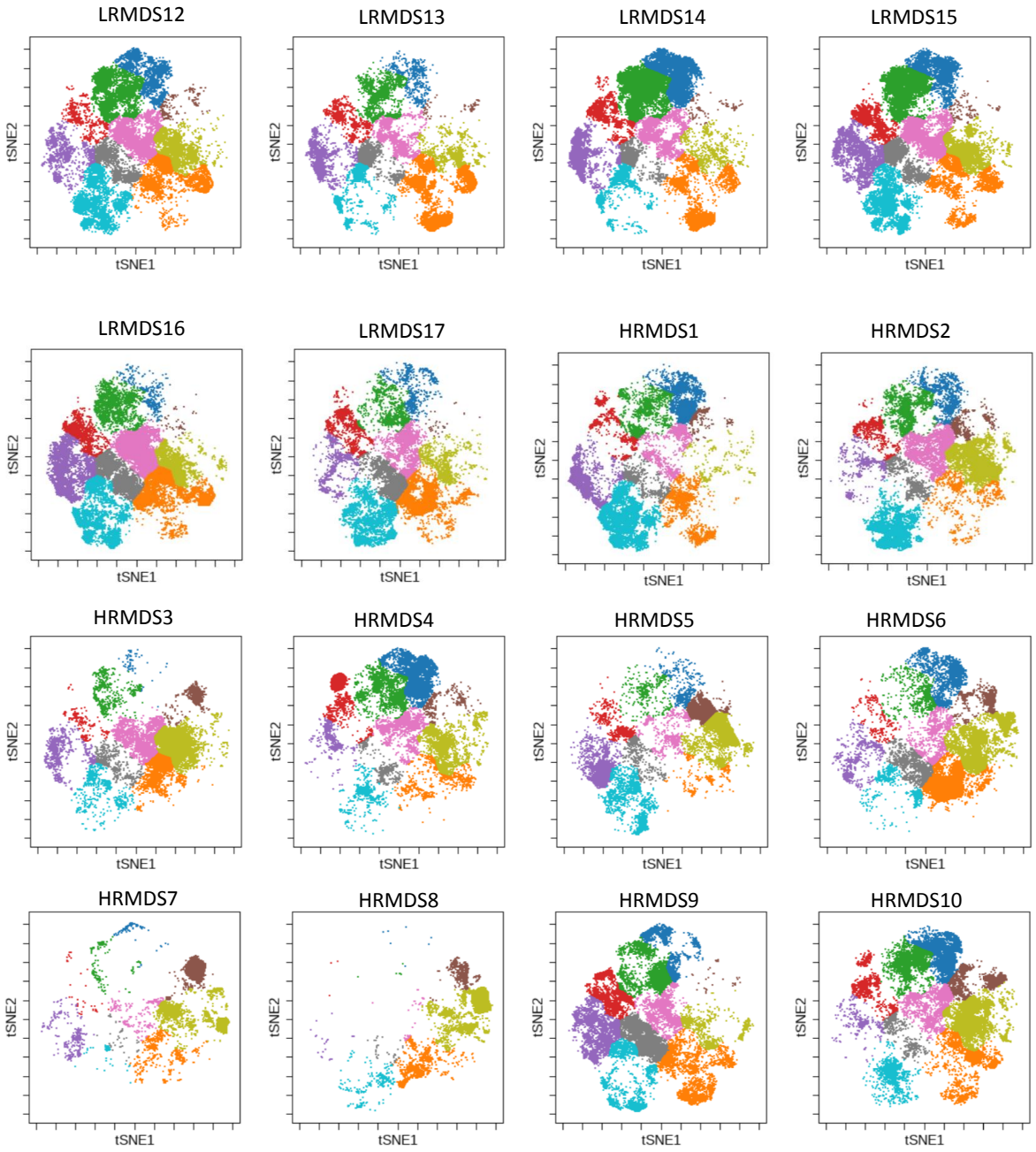


Figure 27 : Représentation individuelle par t-SNE et clustering FlowSOM des cellules CD34+ de 36 échantillons incluant 9 sujets sains âgés, 17 SMD de bas risque, et 10 SMD de haut risque.

La réduction dimensionnelle par t-SNE et le clustering FlowSOM a été réalisée sur le regroupement des cellules des 36 échantillons étudiés. Chaque échantillon est ensuite représenté individuellement

FlowSOM a séparé les cellules de nos 36 échantillons en 10 clusters. Nous avons ensuite pu comparer la répartition de ces clusters chez les sujets âgés contrôles (CTRL), les SMD de bas risque (LRMDS) et les SMD de haut risque (HRMDS) en réalisant une classification

hiérarchique des échantillons. Trois sous-groupes sont clairement mis en évidence en fonction de la répartition des cellules dans les différents clusters FlowSOM. Un premier groupe « cluster 8 » est composé principalement de SMD de haut risque (n=9). Le groupe « cluster 10 » est constitué de SMD de bas risque (n=9). Le groupe « cluster 9 » est constitué de sujets contrôles et d'un SMD de haut risque (n=7). Les autres échantillons sont principalement des contrôles et des SMD de bas risque et n'appartiennent pas à un groupe clairement distinct **(Figure 28).** A noter que l'échantillon LRMDS1 et l'échantillon HRMDS2 sont issus du même patient, l'échantillon LRMDS1 correspond au diagnostic, tandis que l'échantillon HRMDS2 correspond à l'évolution de la maladie. De manière intéressante, l'échantillon LRMDS1 est classé dans le groupe cluster 8 avec l'échantillon HRMDS2 et d'autres SMD de haut risque.

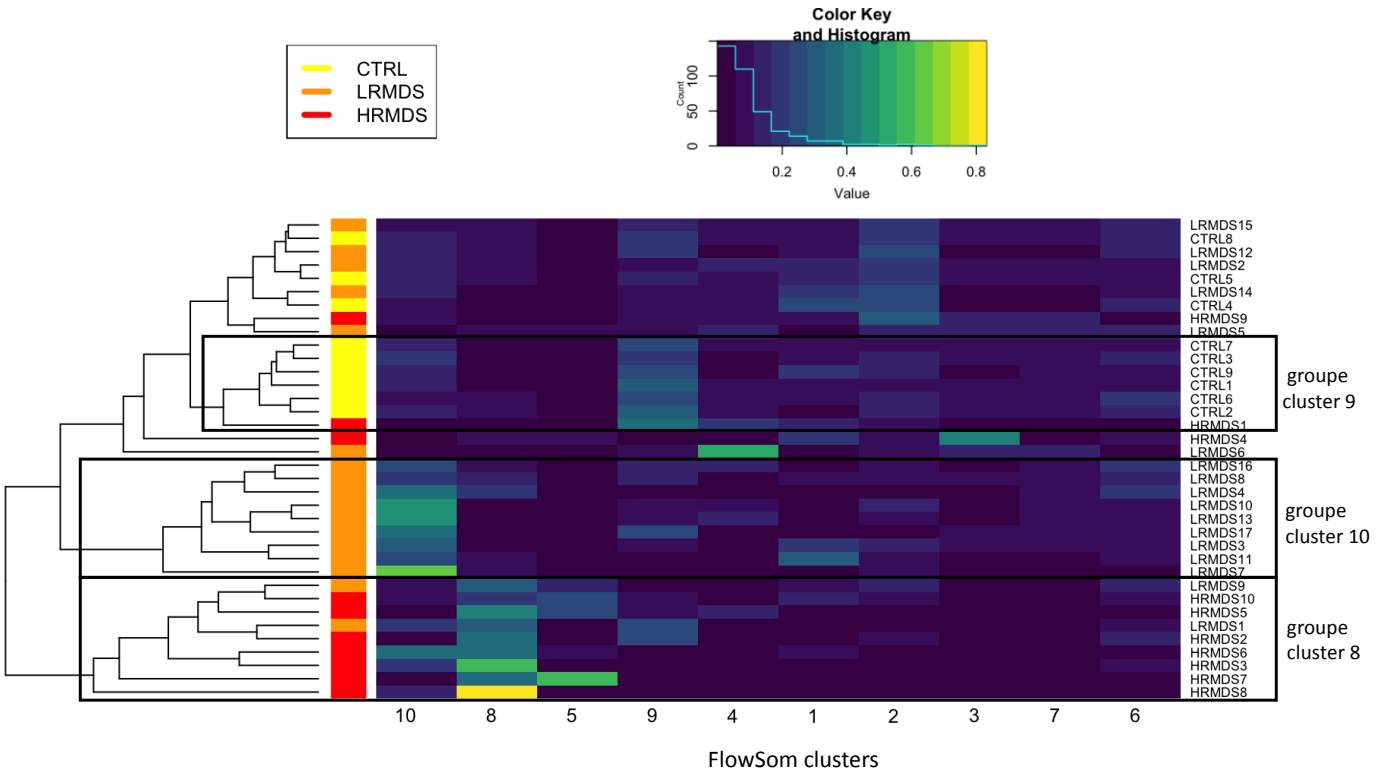


Figure 28 : Heatmap de la répartition des cellules des échantillons sains (CTRL), SMD de bas risques (LRMDS) et SMD de haut risque (HRMDS) au sein des clusters FlowSom.

Le clustering hiérarchique des échantillons en fonction de la répartition des cellules au sein des clusters FlowSom permet de distinguer 3 principaux groupes : « cluster 8 » contient principalement des SMD de haut risque, « cluster 10 » contient des SMD de bas risques, et cluster 9 contient des sujets contrôles et un SMD de haut risque. La couleur de chaque rectangle de la heatmap représente le rapport entre le nombre de cellules appartenant au cluster correspondant et le nombre de cellules totales de l'échantillon correpondant. Par conséquent, le total de chaque ligne est égal à 1.

Nous avons voulu ensuite caractériser les cellules appartenant aux clusters 8, 9 et 10. Pour cela nous avons étudié les cellules appartenant à ces clusters avec les méthodes classiques d'analyse de données de cytométrie en flux **(Figure 29)**. Nous avons pu établir que le cluster 8 est constitué de cellules correspondant aux LMPP et aux GMP. Le cluster 9 est quant à lui composé de cellules correspondant aux populations CMP3 et MEP3, et le cluster 10 correspond aux progéniteurs B/NK.

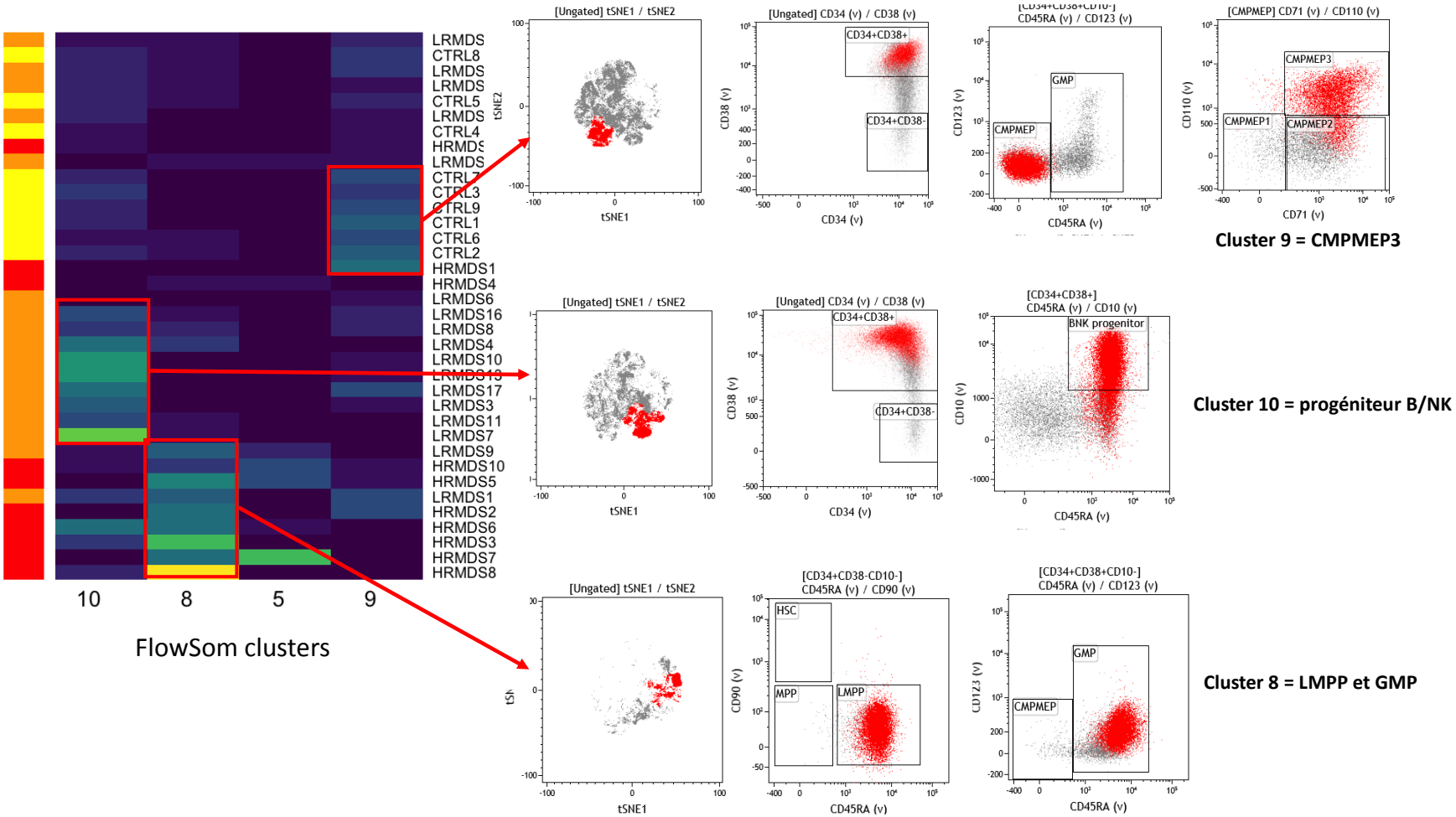


Figure 29 : Caractérisation des cellules des échantillons sains (CTRL), SMD de bas risques (LRMDS) et SMD de haut risque (HRMDS) appartenant aux clusters FlowSOM 8,9 et 10.

Les cellules appartenant aux clusters 8,9 et 10 qui permettent de distinguer les sujets âgés des SMD de bas risque et de haut risque sont caractérisées par l'expression de leurs marqueurs analysées de manière classique sur des graphiques biparamétriques.

Nous avons ensuite démontré que les SMD de haut risque présente une quantité plus importante de LMPP et GMP par rapport aux SMD de bas risques et aux contrôles. Les SMD de bas risque et de haut risque présentent un nombre diminué de cellules de type CMP3 et MEP3 par rapport aux sujets sains d'âge comparable. Enfin, il existe un groupe de patients SMD de bas risque (Groupe cluster 10) qui ont un nombre de progéniteurs B/NK plus important comparé aux autres échantillons analysés **(Figure 30)**.

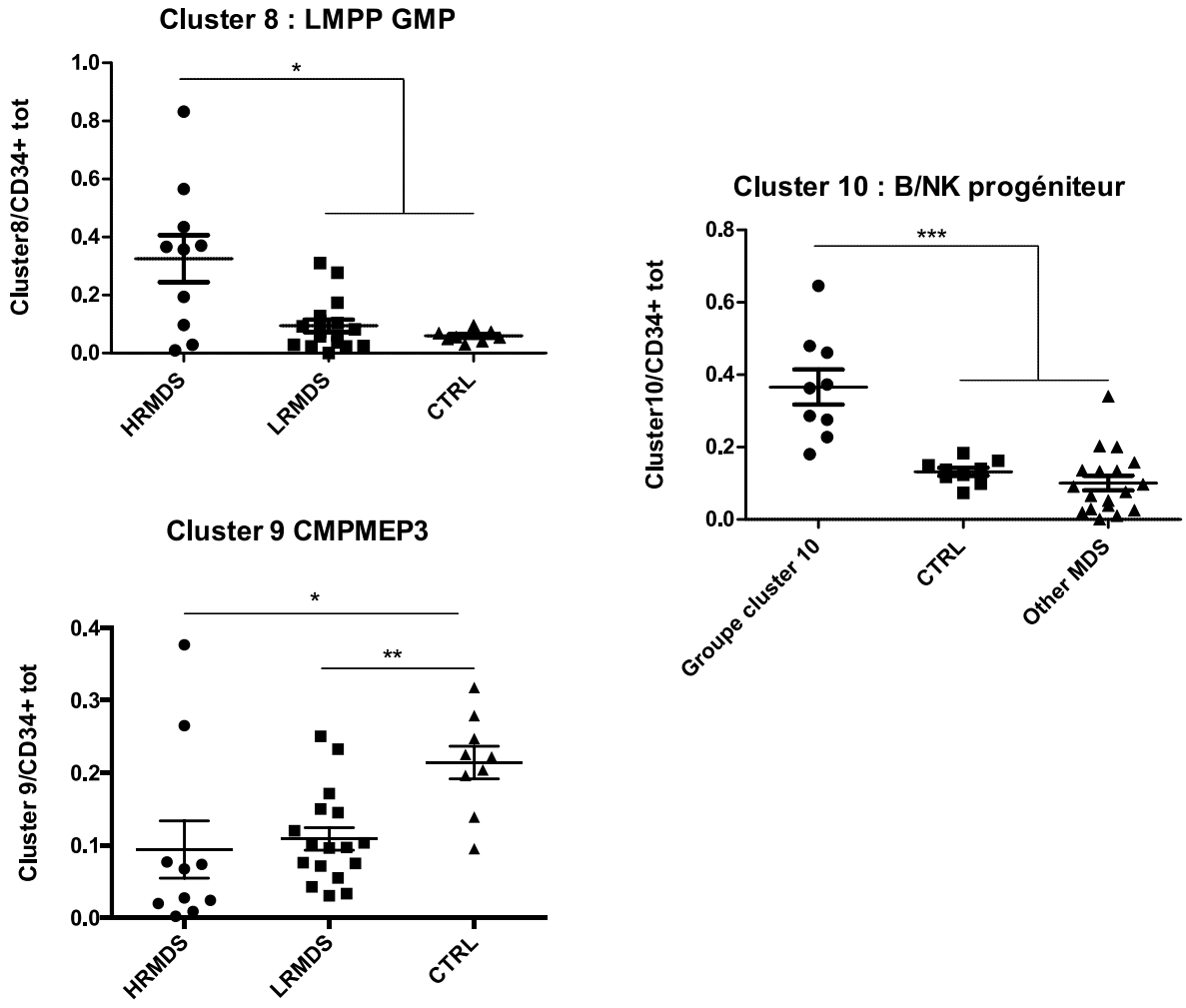


Figure 30 : Comparaison de la répartition des cellules CD34+ au sein des clusters FlowSOM 8,9 et 10 entre les sujets sains âgés, les SMD de bas risque et de haut risque.

Le nombre de cellules appartenant aux clusters 8,9 et 10 par rapport au nombre de cellule total de chaque échantillon est représenté en ordonnée. Les différents groupes de patients (CTRL, LRMDS, HRMDS, Groupe cluster 10 et Other MDS) ont été comparés avec un test non paramétrique de Mann-Whitney. (* : p<0,05 ; ** : p<0,01 ; *** : p<0,001)

En observant l'allure des t-SNE de chaque échantillon, nous avons remarqué que les cellules des SMD de haut risque étaient regroupées dans un nombre restreint de clusters par rapport aux témoins sains et aux SMD de bas risques **(Figure 27).** Pour estimer par une valeur cette impression visuelle, nous avons calculé l'entropie de Shannon de la répartition des cellules au sein des clusters **(Figure 31)**. En pratique, une valeur d'entropie élevée reflète une répartition des cellules au sein des 10 clusters, alors que, lorsque les cellules se répartissent dans un nombre limité de clusters, la valeur d'entropie est basse. Nous observons que l'entropie est significativement plus basse dans les SMD de haut risque par rapport aux SMD de bas risque et aux sujets sains âgés. Ceci montre que les cellules CD34+ de SMD haut risque ne sont pas aussi diversifiées que chez les SMD de bas risque et les sujets sains âgés.

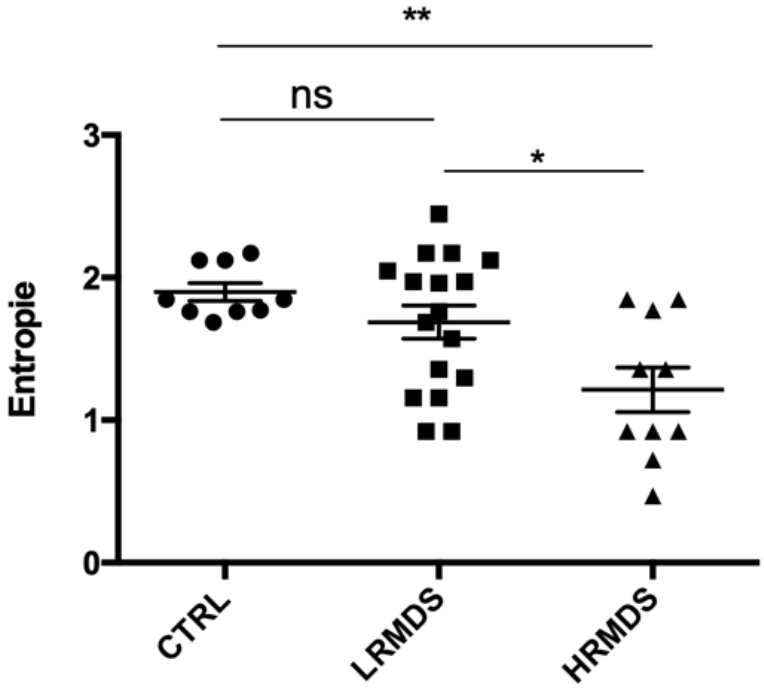


Figure 31 : Entropie de la répartition des cellules au sein des clusters FlowSOM.
Pour chaque patient, le calcul de l'entropie de Shannon est effectué sur les pourcentages de cellules appartenant aux clusters déterminés par FlowSOM. L'entropie reflète ainsi l'hétérogénéité de la répartition des cellules au sein des clusters. Les différents groupes de patients (CTRL, LRMDS, HRMDS) ont été comparés avec un test non paramétrique de Mann-Whitney. (* : p<0,05 ; ** : p<0,01)

4. Conclusion

Nous démontrons dans cette étude que la caractérisation des cellules souches et progénitrices CD34+ des SMD peut permettre de les différencier des sujets sains d'âge comparable. En effet, une partie des SMD semble avoir une diminution des populations CMPMEP3 par rapport aux témoins. De plus un sous-groupe de SMD semble présenter un nombre augmenté de progéniteurs B/NK par rapport aux autres SMD et aux témoins. Nous montrons également que les populations CMP et LMPP sont surreprésentés dans une partie des SMD de haut risque.

L'entropie de la répartition des cellules au sein des clusters reflète en quelque sorte la diversité cellulaire du compartiment CD34+, et celle-ci est significativement diminué dans les SMD de haut risque en accord avec l'expansion d'un clone pathologique au détriment de la diversité de l'hématopoïèse normale.

Ces résultats seront par la suite confrontés aux données clinico-biologiques et aux données de génotypage.

Résultats partie 2 : Rôle de la variabilité de l'expression génique dans l'hématopoïèse normale et pathologique

1. Introduction

L'hématopoïèse est un processus complexe dont la compréhension a été grandement améliorée par les expériences de scRNA-Seq. Les syndromes myélodysplasiques ont pour caractéristique commune une hématopoïèse anormale (dysplasique) qui peut toucher une ou plusieurs lignées, liée en partie à la présence d'anomalies cytogénétiques et de mutations somatiques. Ces anomalies apparaissent dans le compartiment des cellules souches et progénitrices hématopoïétiques médullaires. L'azacytidine est l'un des médicaments le plus utilisé pour tenter de ralentir l'évolution de ces maladies. Nous avons voulu à travers différentes expériences de scRNA-Seq comprendre de quelle manière le transcriptome est altéré dans les cellules du compartiment souche et progéniteurs des SMD, et quelles sont les effets de l'azacytidine sur l'expression génique en lien avec la réponse au traitement. Pour cela, nous avons débuté au cours de ma thèse une collaboration avec Olivier Gandrillon et Pierre Sujobert de l'université de Lyon. L'idée était de mettre en commun nos connaissances sur l'hématopoïèse normale, les SMD, et la biologie intégrative pour répondre aux questions posées.

2. Matériels et Méthodes

2.1. Patients et échantillons

Les caractéristiques des 8 échantillons issus des 6 individus étudiés dans cette partie de mon travail de thèse sont résumées dans le **Tableau 3.** Les patients Ctrl1 et Ctrl3 sont des sujets âgés sains, présentant un hémogramme normal et ne prenant pas de médicaments susceptibles d'affecter l'hématopoïèse (chimiothérapie ou immunosuppresseurs). Les patients MDS2 et MDS4 sont des malades atteints de SMD avec sidéroblastes en couronne mutés pour le gène *SF3B1*. Le patient NR est atteint de SMD avec excès de blastes de type 1, traité par azacytidine et non répondeur au traitement. Le patient R est atteint de LMMC 2, traité par azacytidine et répondeur au traitement. Les échantillons NR_BF et R_BF ont été collectés respectivement le jour de début de traitement juste avant la première injection. Les échantillons NR_AF et R_AF ont été collectés respectivement après 13 et 12 cures d'azacytidine. Le prélèvement a eu lieu 46 (NR_AF) et 48 jours (R_AF) après la dernière injection.

Les échantillons médullaires des patients atteints de SMD (MDS2, MDS4, NR_BF, NR_AF, R_BF et R_AF) ont été obtenus au décours d'une ponction de moelle osseuse réalisée dans le cadre du diagnostic ou du suivi de la maladie. Suite au prélèvement, les cellules mononuclées ont été isolées sur un gradient de Ficoll. Les échantillons de moelle osseuse issus de patients sains âgés (Ctrl1 et Ctrl3) ont été obtenu par extraction des cellules médullaires issues de l'os de la tête fémorale.

Les échantillons ont ensuite été pris en charge de la même manière que ceux utilisés dans les expériences de cytométrie en flux (page 79).

Tableau 3 : Caractéristiques clinico-biologiques des 8 échantillons médullaires étudiés en scRNA-Seq.

Patient	Age	Sex	Diagnosis	R-IPSS	Hb, g/dL	Platelets, G/L	Neutrophils, G/L	Monocytes, G/L	Lymphocytes, G/L	Bone marrow cellularity	Erythroid precursors, %	Myeloid precursors, %	Megakaryocytes	% bone marrow blasts	Ring sideroblasts, %	Karyotype	Somatic mutations (VAF%)
Ctrl1	68	F	Healthy ctrl	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Not detected
Ctrl3	74	M	Healthy ctrl	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Not detected
MDS2	87	F	MDS-RS	low	10	342	1,4	0,1	0,5	High	60	30	Presents	2	25	46,XX,del(11)(q14)[3] /46,sl,?del(20)(q12)[2] /46,X,del(X)(q21),add(5)(q?31), del(11)(q14),add(17)(q2?2)[12] /46,XX[2]	TET2, p.Q769X (44%) SF3B1, p.K700E (43%)
MDS4	82	M	MDS-RS	very low	10,6	287	3,46	0,96	1,74	Very high	52	43	Numerous	1	50	46, XY [20]	SF3B1, p.K666RK (35%)
NR_BF	61	F	MDS-EB1	high	11,1	13	6,45	0,62	1,37	Normal	18	64	Presents	7	NA	46,XX,add(21)(q2?2)[20].ish add(21)(q22)(AML1-)[10] .	SRSF2 p.P95RP (44%)
NR_AF	63	F			8,6	24 (after CPA)	2,29	1,4	3,69	Very High	38	36	rares	14	NA	NA	SRSF2,p.P95RP (42%) KRAS,p.G12GA (45%)
R_BF	76	M	CMML-2	Intermediate	13,1	63	2,5	1,5	2,7	Very High	12	62	Numerous	15	NA	46, XY [24]	TET2,p.K693Nfs (50%) TET2,p.M1656Ifs (39%) CBL,p.R420QR (65%) CBL,Splice (8%) ASXL1,p.D979Afs (22%) NRAS,p.G13GD (2%) CSF3R,p.R698CR (54%) KRAS,p.G60R (8%)
R_AF	77	M			14	91	6,72	0,9	1,39	Very High	11	66	Numerous	7	NA	NA	TET2,p.K693Nfs (48%) TET2,p.M1656Ifs (39%) CBL,Splice (19%) CBL,p.R420QR (54%) ASXL1,p.D979Afs (13%) CSF3R,p.R698CR (52%) KRAS,p.G60R (8%)

2.2. Études génomiques

La recherche de mutations sur les CMN de moelle osseuse de tous les échantillons analysés a été réalisée après extraction de l'ADN (Maxwell 16 DNA Purification system de chez Promega) par séquençage nouvelle génération (NGS). Le séquençage NGS de 45 gènes (*ASXL1, BCOR, BCORL1, BRAF, CALR, CBL, CSF3R, CSNK1A1, CUX1, DDX41, DNMT3A, ETNK1, ETV6, EZH2, FLT3, GATA2, HRAS, IDH1, IDH2, JAK2, KDM6A, KIT, KRAS, MPL, MYD88, NRAS, PHF6, PPM1D, PTEN, PTPN11, RAD21, RHOA, RIT1, RUNX1, SETBP1, SF3B1, SH2B3, SRSF2, STAG2, STAT3, TET2, TP53, U2AF1, WT1, ZRSR2*) a été effectué à l'aide du kit Ion AmpliSeq library kit2 384 (Life Technologies). L'amplification par PCR multiplex (606 amplicons) a été réalisée à partir de 2x10 ng d'ADN génomique. Après amplification, les codes-barres et les adaptateurs ont été ajoutés aux amplicons par ligation. Les produits de PCR ont été purifiés sur billes AMPure (Life Technologies). La PCR en émulsion a été réalisée à l'aide de l'instrument Ion Chef (Life Technologies). Le séquençage a été effectué sur Ion S5XL (Life Technologies) sur la puce 530 (32 échantillons par puce). De plus, pour chaque échantillon, la mutations c.1934dupG *d'ASXL1* a été recherchée par analyse de fragments. Pour l'analyse, la détection des variants à partir des fichiers .bam issus du séquenceur a été réalisée à l'aide du logiciel NextGENe (Softgenetics). En parallèle, le pipeline « Polydiag » développé par l'institut IMAGINE a été utilisé.

2.3. Préparation des cellules et scRNA-Seq

Après décongélation, les cellules mortes ont été éliminées par tri négatif immunomagnétique (technologie MACS MicroBead de Miltenyi Biotec™). Après élimination des cellules mortes, les cellules CD34 positives (cellules souches et progénitrices) ont été récupérées après un deuxième tri immunomagnétique et lavées avec du PBS contenant 0,04% de BSA. La concentration et la viabilité des cellules ont été déterminées au microscope sur cellule de Malassez après coloration au bleu trypan.

Les expériences de scRNA-Seq ont été réalisées avec le système Chromium de 10x genomics, avec les kits Chromium Single Cell 3 'V2 et Chromium Single Cell 3 'V3 selon le protocole du fabricant (www.10xgenomics.com)[189]. Deux expériences ont été réalisées, la

première sur une même puce chromium contenait les cellules des échantillons Ctrl1, Ctrl3, MDS2 et MDS4 (avec le kit Chromium Single Cell 3 ’V2), la seconde sur une autre puce contenait les cellules des échantillons R1, R2, NR1, NR2 (avec le kit Chromium Single Cell 3 ’V3). Le nombre de cellules visées par échantillon était de 5000 cellules.

Les librairies ont été séquencées par la société Integragen sur un séquenceur HiSeq4000 avec une profondeur visée de 50000 reads par cellules.

2.4. Analyse bioinformatique

Après séquençage, le logiciel Cell Ranger (https://support.10xgenomics.com/single-cellgene-expression/software/pipelines/latest/what-is-cell-ranger) a été utilisé pour traiter les données brutes, afin d’aligner les reads sur le génome et générer les matrices d'expression gènes-cellules. Les reads ont été alignés sur le génome de référence GRCh38 par STAR avec l’annotation issue d’ENSEMBL.

Des données provenant de CMN de moelle totale de donneur sain publiées par l’équipe de Greenleaf ont été réanalysées. Les analyses ont été réalisées sur la matrice gènes cellules filtrée par les auteurs[190]. Les auteurs ont choisi de retirer les gènes ribosomiques et mitochondriaux de la matrice gène cellule. Pour le calcul d’entropie, les fichier .bam d’origines ont donc été démultiplexés à l’aide du logiciel DropEst[191] et l’expression des gènes mitochondriaux et ribosomaux a été réincorporée à la matrice gènes cellules déjà filtrée.

Après l’obtention de la matrice gène cellules, l’analyse a été réalisée à l’aide de Seurat (v3), outil développé par l’équipe de Satija[13], conçu pour le contrôle de la qualité et l'exploration de données de scRNA-Seq.

Au cours de ce contrôle de qualité et dans le but de ne conserver que des cellules donnant lieu à des données informatives, nous avons d’abord observé la répartition des cellules en fonction du nombre de gènes exprimés et en fonction du pourcentage de gènes mitochondriaux exprimés **(Figure 32).** Sur la base de ces graphiques nous avons fait le choix de ne conserver pour la suite de l’analyse que les cellules qui exprimaient entre 500 et 5500 gènes avec un pourcentage de gènes mitochondriaux inférieur à 10 % en ce qui concerne les échantillons de la première expérience (Ctrl1, Ctrl3, MDS2 et MDS4). Pour les échantillons de la deuxième expérience (R1, R2, NR1 et NR2), nous avons choisi de conserver les cellules qui

exprimaient entre 500 et 7000 gènes avec un pourcentage de gènes mitochondriaux inférieur à 15%. Nous avons également choisi de ne conserver que les gènes exprimés dans au moins 3 cellules de chaque échantillon.

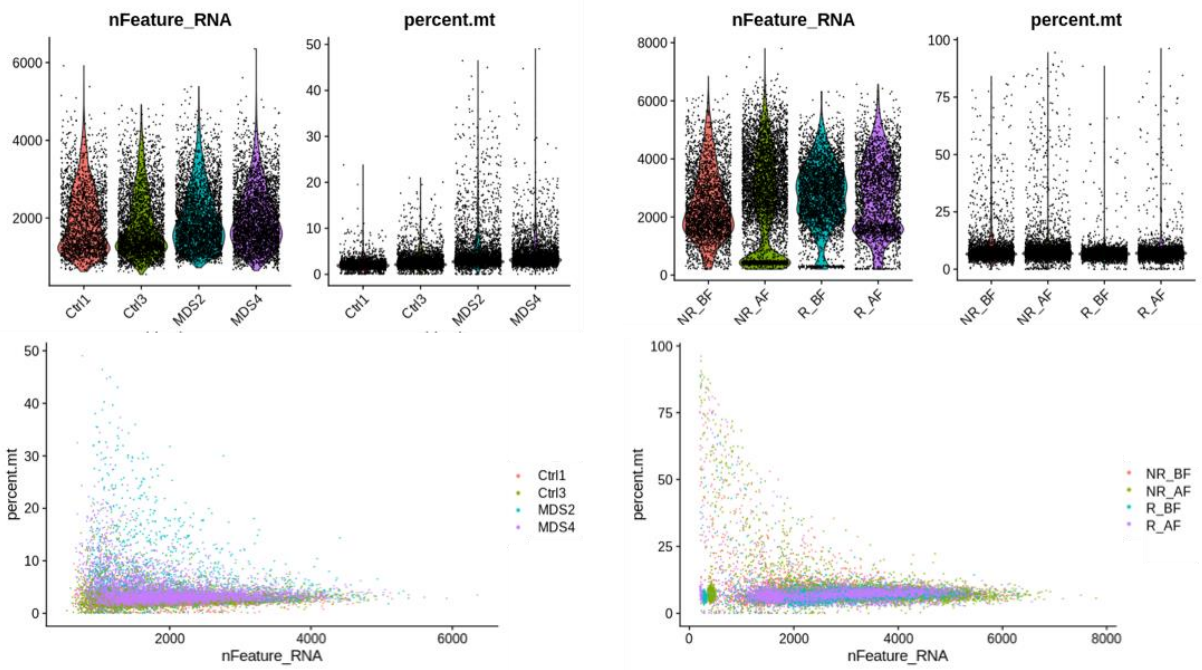


Figure 32 : Contrôle qualité des données de scRNA-Seq.
Le nombre de gènes par cellules (nFeature_RNA), ainsi que le pourcentage de gènes mitochondriaux (percent.mt) sont représentés pour toutes les cellules (qui correspondent à un point sur les graphiques).

Les matrices gènes-cellules de chaque échantillon obtenues ont ensuite été normalisées avec la technique SCtransform[12]. Après normalisation, les matrices gènes-cellules ont été soumises à des techniques de réduction de dimensionnalité telles que l'analyse en composante principale (ACP)[15] et la UMAP[16] à l'aide de Seurat. Nous avons également utilisé le package python Scanpy[192] pour calculer le ForceAtlas2 (FA)[17]. Le clustering a été effectué par Seurat avec l'algorithme de Louvain[19].

Les cellules ont été annotées individuellement à l'aide du logiciel SingleR[23] en comparant leur profil d'expression génique avec celui des 34 populations médullaires normales publiées par Hay et al [93]. Les clusters ont également été annotés par SingleR à partir de la même référence.

L'attribution d'une phase du cycle cellulaire à chaque cellule a été réalisé avec Seurat, basé sur la stratégie décrite par Tirosh et al[193]

Pour ordonner les cellules à partir des CSH, jusqu'aux cellules les plus matures, les sous populations cellulaires appartenant à chaque voie de différenciation sont sélectionnées et représentées graphiquement en deux dimensions selon les coordonnées de la UMAP globale. Si des cellules sont positionnées de manière aberrante dans cet espace, elles sont éliminées manuellement avec la fonction Cellselector de Seurat. Pour chaque voie de différenciation, un pseudotemps représentatif de la différenciation (pseudotime) a été calculé par Slingshot[34]. Les options suivantes ont été utilisées : la population cellulaire de départ est toujours spécifiée comme étant les CD34+ HSC ; l'espace multidimensionnel spécifiée était soit la UMAP, soit la FA, soit la PCA ; la population cellulaire finale du pseudotemps était parfois spécifiée comme étant la sous population correspondant aux cellules les plus matures de la voie de différenciation étudiée. Ainsi en fonction des options choisies, plusieurs pseudotemps ont été calculés pour chaque voie de différenciation. Le pseudotemps choisi pour les analyses en aval étant celui qui permettait d'ordonner les cellules pour chaque voie de différenciation de la façon la plus cohérente avec nos connaissances de l'hématopoïèse.

L'intégration des échantillons entre eux permet de les représenter dans un espace dimensionnel commun pour pouvoir par la suite calculer à partir de ce nouvel espace dimensionnel un pseudotemps commun. Ainsi, afin de comparer les échantillons entre eux, les matrices gènes cellules des patients Ctrl1, Ctrl3,MDS2, et MDS4 ont été intégrés à l'aide de Seurat[194] après normalisation avec la technique SCtransform[12]. Les matrices des patients R_BF et R_AF d'un côté et des patients NR_BF et NR_AF de l'autre ont été intégrées de la même manière.

Concernant l'intégration des échantillons Ctrl1, Ctrl3, MDS2 et MDS4, après annotations des cellules par SingleR, nous avons fait le choix pour la suite des analyses de retirer des types cellulaires sans lien évident avec le compartiment CD34 et dont la représentation n'était pas constante/suffisante dans tous les échantillons (Naive T-cell, NK cells, Plasma Cell, Stromal, Neutrophil, Immature-Neutrophil, soit moins de 20 cellules par échantillon).

2.5. Calcul de l’entropie

L’entropie de Shannon est calculée avec R pour chaque gène sur les données d’expressions bruts pour un nombre précis de cellules. Soit H(X) l’entropie de Shannon du gène X mesuré sur un groupe de 50 cellules déterminé, soit k le nombre de valeurs que peut prendre l’expression du gène X, et pi la probabilité pour que l’expression du gène X soit égale à i :

$$H(X) = -\sum_{i=1}^{k} pi \times \log(pi)$$

Pour ce calcul, il est important de définir le nombre k, c’est-à-dire le nombre de valeurs que peut prendre l’expression du gène X. Nous avons choisi de définir k comme étant égale au rapport N / m ; N étant le nombre de cellules utilisées pour calculer l'entropie, et m étant le plus grand nombre de molécules d'ARNm issues du gène X trouvées dans une cellule.

.

3. Résultats

3.1. Un pic de variabilité de l'expression génique est observé au cours de l'hématopoïèse normale.

Pour étudier la variabilité de l'expression génique au cours de l'hématopoïèse normale, nous avons utilisé les données de scRNA-Seq de cellules mononuclées provenant de la moelle osseuse d'une seul donneuse saine publiées par Granja et al [190]. Les données d'expressions ont été réanalysées avec Seurat[13], puis les cellules ont été annotées individuellement avec SingleR[23] en comparant l'expression génique de chaque cellule avec le profil d'expression génique des populations hématopoïétiques décrites par Hay et al à partir des données du Human Cell Atlas[93] **(Figure 33).** Les marqueurs spécifiques des sous populations distinguées par SingleR sont en accord avec les études précédemment réalisées sur la moelle osseuse **(Figure 34)**[89,93,195,196]. On distingue sur la représentation UMAP les différentes populations hématopoïétiques physiologiquement présentes dans la moelle osseuse (en accord avec les données du Human Cell Atlas), ce qui nous permet de valider ces données comme représentatives de la moelle osseuse saine **(Figure 33)**. Quatre voies de différenciation qui partent des cellules souches hématopoïétiques (CD34+ HSC) se distinguent clairement sur la UMAP. L'érythropoïèse, la granulopoïèse, la maturation dendritique, et la lymphopoïèse. Nous avons ainsi à notre disposition les données d'expression génique presque exhaustives (15962 gènes, 12602 cellules) de la quasi-totalité des sous populations hématopoïétiques.

A partir de ces données, nous avons voulu savoir si le pic de variabilité de l'expression génique observé par Richard et al[118] sur un modèle *in vitro* d'érythropoïèse aviaire était également présent dans l'érythropoïèse humaine, mais aussi dans les autres voies de différenciation hématopoïétique.

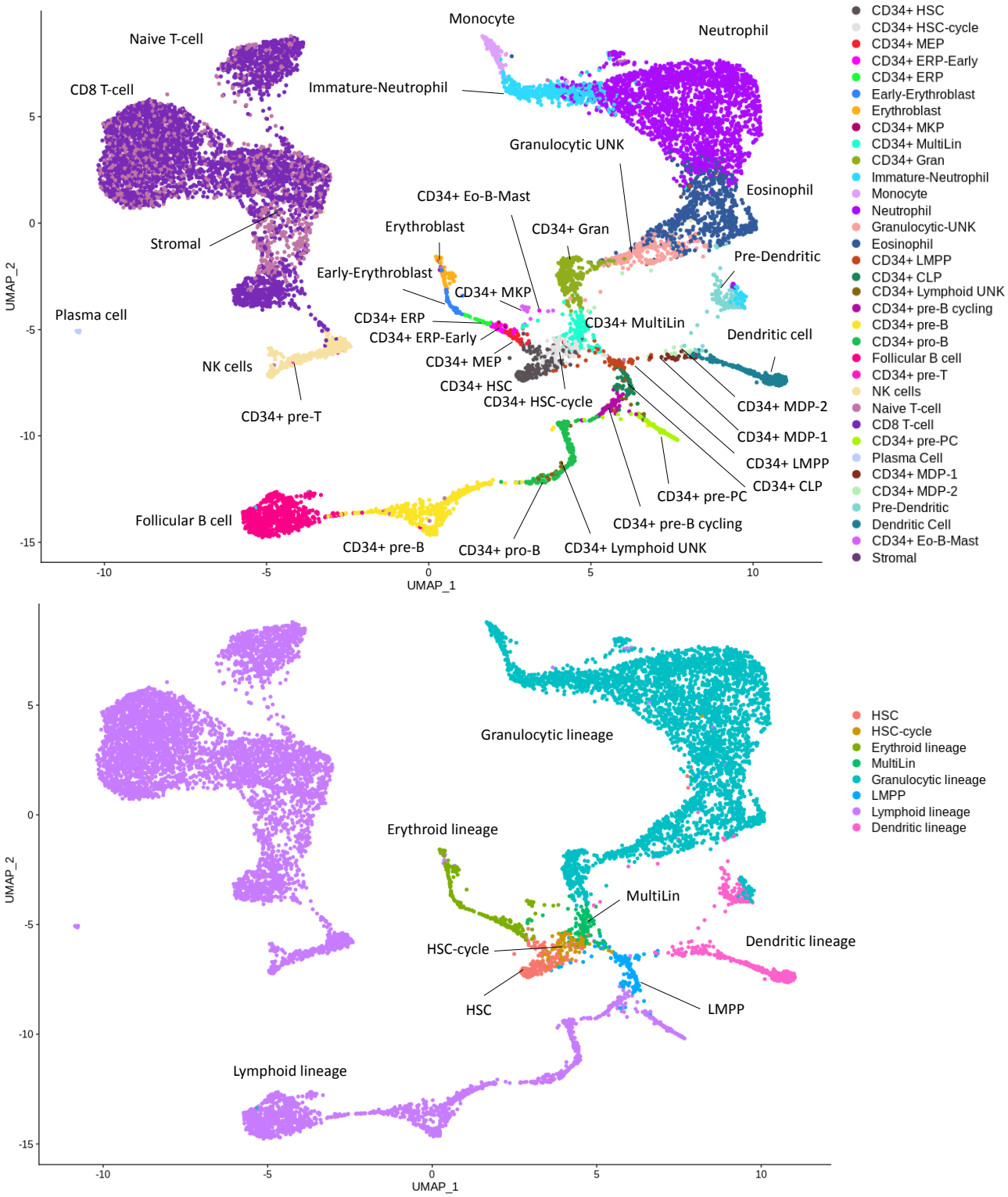


Figure 33 : Représentation par UMAP du paysage transcriptionnel des 12602 cellules mononucléées issues d'une moelle de donneur sain[190].

34 des 35 populations décrites par Hay et al sont représentées sur la UMAP de cet échantillon de moelle normale (seules les plaquettes sont manquantes). On distingue facilement les principales voies de la différenciation hématopoïétique que sont l'érythropoïèse, la granlulopoïèse, la maturation dendritique, et la lymphopoïèse.

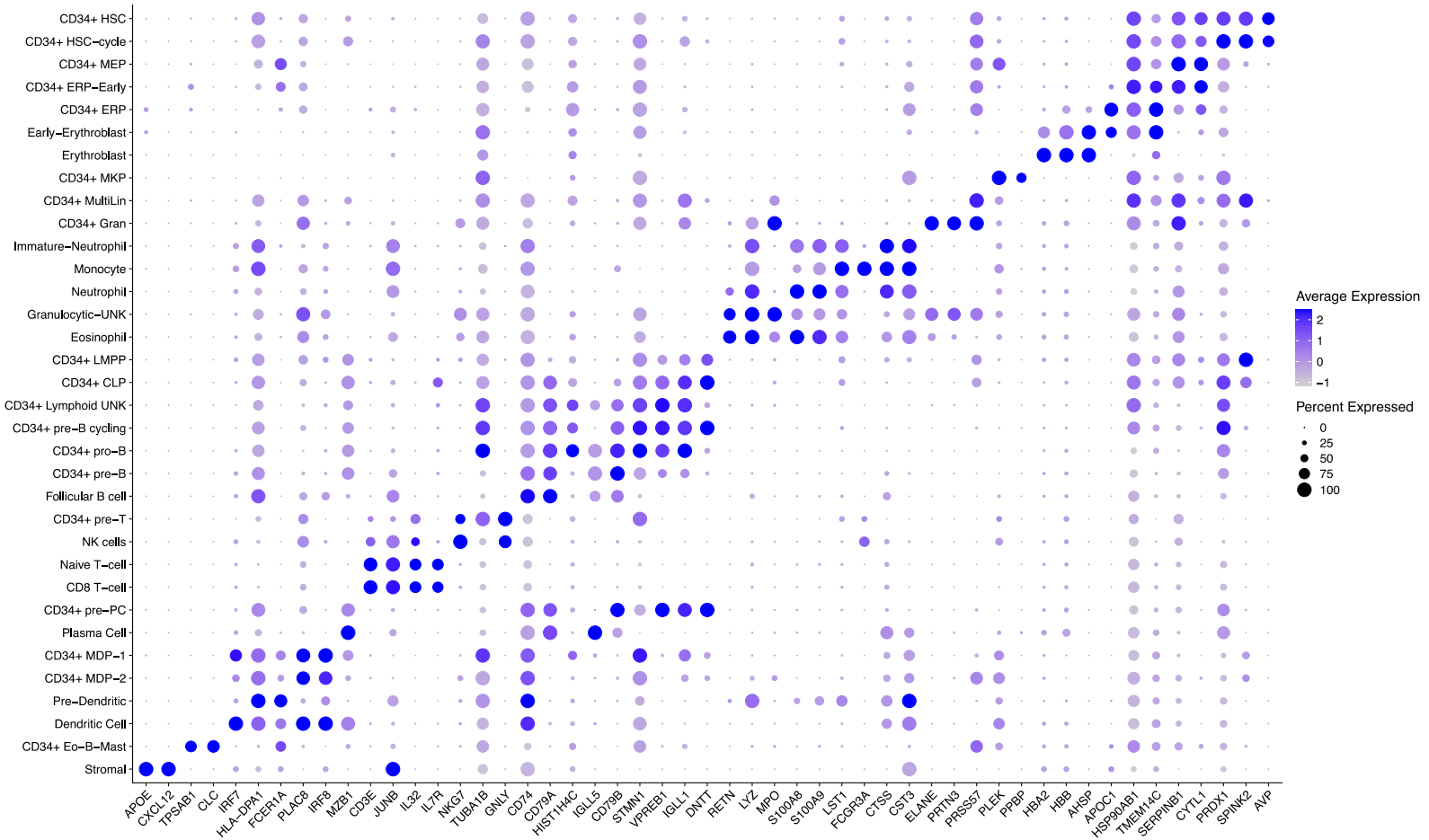


Figure 34 : Expression transcriptionnelle des marqueurs les plus spécifiques des sous populations de la moelle osseuse normale annotées par SingleR.

Les gènes les plus différentiellement exprimés de chaque sous population annotées par SingleR en comparaison avec toutes les autres ont été déterminés par la fonction FindMarkers de Seurat. Les deux gènes les plus spécifiques de chaque sous population sont représentés sur ce graphique. La couleur du cercle indique les valeurs d'expression génique et la taille du cercle représente la proportion de cellules exprimant le marqueur dans la sous population.

Les données de scRNA-Seq que nous avons analysé correspondent à un « instantané » de l'hématopoïèse, il est donc nécessaire de classer les cellules dans l'ordre qui correspond le mieux possible à l'hématopoïèse telle qu'elle est dans la réalité actuellement admise de ce processus dynamique. Pour cela, nous avons dans un premier temps sélectionné les cellules par voies de différenciation afin de les ordonner de la cellule la plus immature (CD34+ HSC), jusqu'aux cellules les plus matures. A l'aide du package R Slingshot[34], les cellules ont été ordonnées le long d'un « pseudotemps » (pseudotime). Ainsi, pour chaque voie de différenciation, les cellules sont caractérisées chacune par une valeur qui les positionne le long d'un pseudotemps qui reflète de manière plus ou moins précise l'hématopoïèse physiologique. L'entropie de Shannon reflète la variabilité de l'expression d'un gène entre les cellules issues d'une même population. Pour avoir une représentation précise de l'évolution de l'entropie moyenne au cours des différentes voies de différenciation hématopoïétique, nous avons choisi de calculer l'entropie de Shannon de chaque gène, puis de faire la moyenne, sur une fenêtre glissante de 50 cellules qui se déplace tout au long du pseudotemps avec un pas de 10 cellules. Chaque point sur le graphique correspond donc à une fenêtre de 50 cellules **(Figure 35) :**

a) pour l'érythropoïèse, les 444 cellules sont classées dans un ordre cohérent avec ce qui est communément admis. On note qu'après les cellules souches (CD34+ HSC), la population cellulaire qui vient juste après correspond aux MEP (CD34+ MEP). Les cellules souches en cycles (CD34+ HSC-cycle) ainsi que les progéniteurs multilignés (CD34+ MultiLin) ne semble pas se trouver sur le chemin de la différenciation érythropoïétique **(Figure 33)**. Après les MEP, les cellules ordonnées le long du pseudotemps correspondent dans l'ordre aux progéniteurs érythroïdes précoces (CD34+ ERP-early), tardifs (CD34+ ERP), aux érythroblastes immatures (Early-Erythroblast) et enfin aux érythroblastes. L'analyse de ces données montre qu'au cours de la différenciation, l'entropie augmente pour atteindre un pic au niveau de la jonction entre les MEP et les progéniteurs érythroïdes précoces puis redescend en dessous du niveau de base dans la population des érythroblastes matures. Nous montrons ainsi pour la première fois qu'il existe chez l'homme un pic de variabilité de l'expression génique au cours de l'érythropoïèse, ce qui confirme les observations de Richard et al[118] effectuées dans un modèle *in vitro* d'érythropoïèse aviaire.

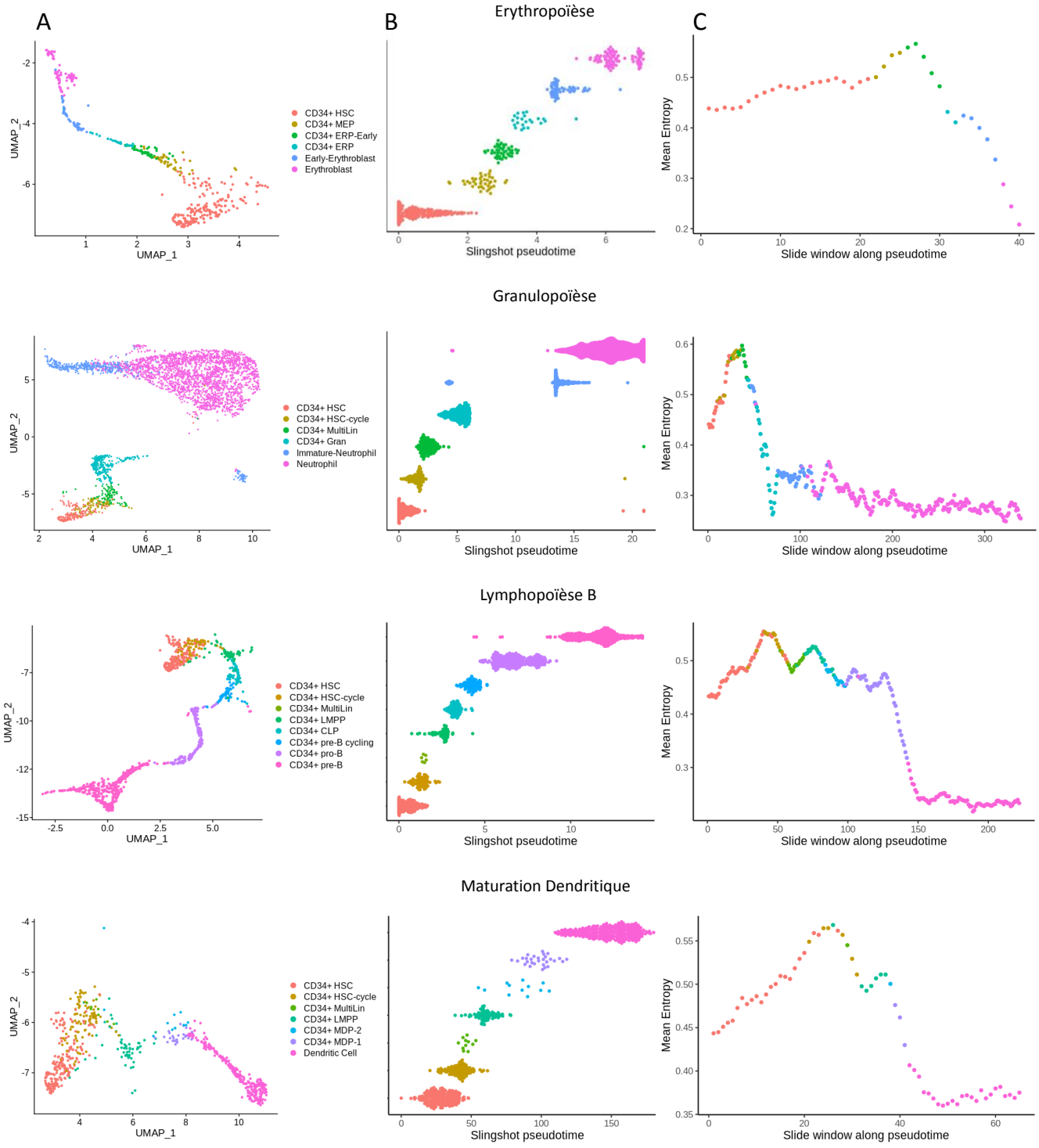


Figure 35 : Evolution de la moyenne d'entropie au cours des principales voies de la différenciation hématopoïétique normale.

A) Les populations cellulaires spécifiques de chaque voie de différenciation sont sélectionnées et représentées en deux dimensions selon les coordonnées de la UMAP globale (Figure 33). **B)** Les cellules sont ordonnées selon le pseudotemps calculé par Slingshot. **C)** Chaque point correspond à la moyenne d'entropie de tous les gènes calculée sur une fenêtre de 50 cellules. La fenêtre avance le long du pseudotemps avec un pas de 10 cellules. La couleur de chaque point sur le graphique correspond à la nature de la première cellule de la fenêtre correspondante.

b) Pour la granulopoïèse les 3440 cellules sélectionnées sont classées par slingshot dans l'ordre communément admis. En effet, on observe après les cellules souches (CD34+ HSC), les cellules souches en cycles (CD34+ HSC-cycle), les progéniteurs multilignées (CD34+ MultiLin), les progéniteurs granuleux (CD34+ Gran), les neutrophiles immatures (Immature Neutrophil) et enfin les neutrophiles matures (Neutrophil). Pendant la granulopoïèse, l'entropie augmente pour atteindre un pic au niveau des progéniteurs multilignées (CD34+ MultiLin) puis redescend pour atteindre un minimum au niveau des neutrophiles (Neutrophil). Nous montrons ainsi qu'il existe chez l'homme un pic de variabilité de l'expression génique au cours de la granulopoïèse.

c) En ce qui concerne la lymphopoïèse B nous avons choisi de sélectionner uniquement les sous populations dont la maturation se fait dans la moelle, en excluant les cellules qui sont connues comme subissant une maturation au niveau des ganglions lymphatiques (cellules folliculaires et plasmocytaires). Ainsi les 1161 cellules sont correctement ordonnées sur le pseudotemps des cellules souches hématopoïétiques jusqu'aux cellules pré-B (CD34+ pre-B). Durant la maturation lymphocytaire B, l'entropie augmente pour atteindre son pic au niveau des cellules souches en cycle (CD34+ HSC-cycle), puis un second pic est observé au niveau de la jonction entre les progéniteurs multilignées (CD34+ MultiLin), et les progéniteurs lympho-myéloïdes (CD34+ LMPP). L'entropie diminue ensuite pour atteindre un minimum au niveau des cellules pré-B. Ceci montre qu'il existe au cours de la lymphopoïèse B une augmentation de la variabilité de l'expression génique qui diminue ensuite dans les cellules plus matures.

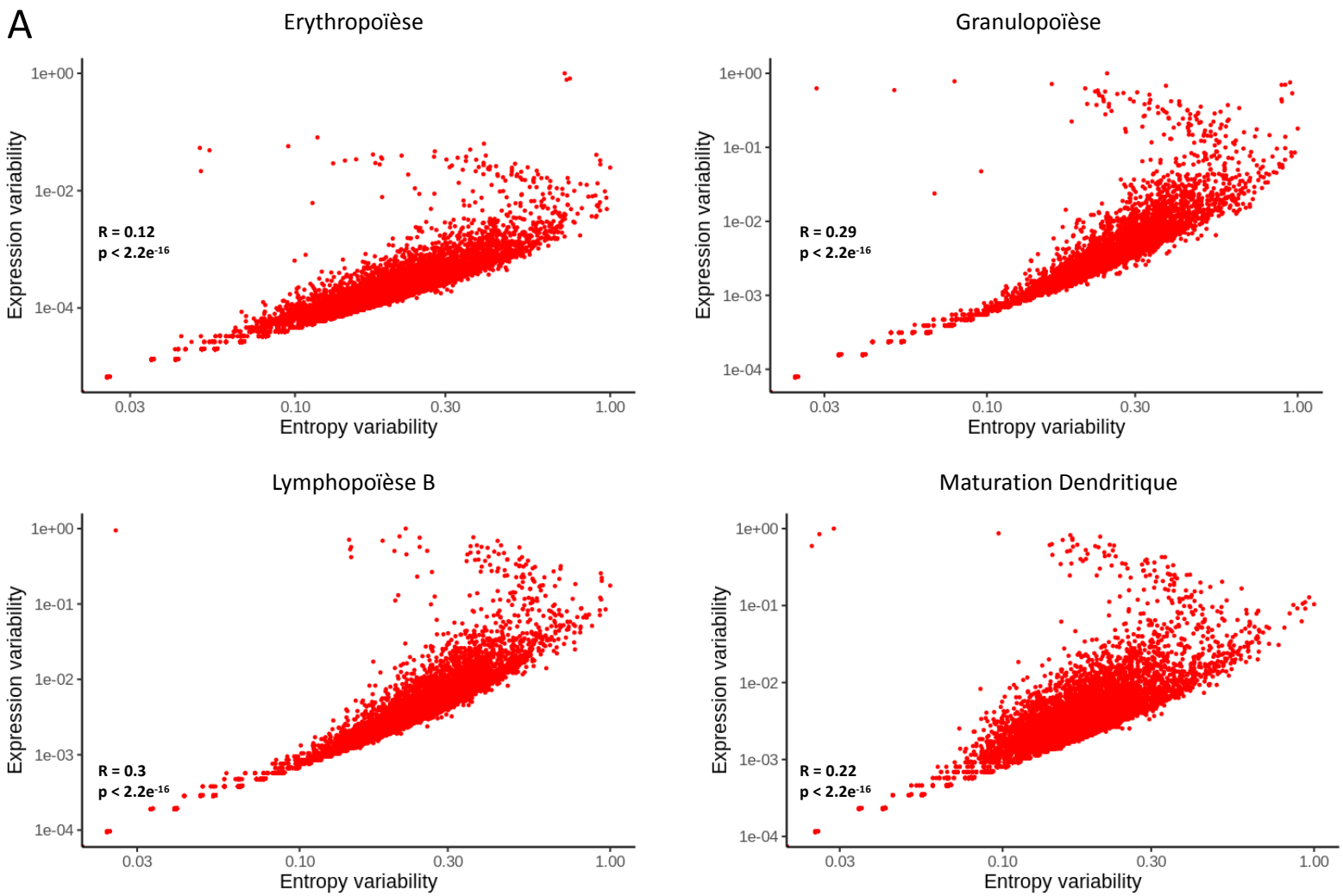
d) A propos de la maturation dendritique, les 699 cellules sélectionnées sont ordonnées par slingshot en commençant par les cellules souches hématopoïétiques, puis les cellules souches en cycles (CD34+ HSC-cycle), les progéniteurs multilignés (CD34+ MultiLin), les progéniteurs lympho-myéloïdes (CD34+ LMPP), les progéniteurs mono-dendritiques de type 2 (CD34+ MDP-2), les progéniteurs mono-dendritiques de type 1 (CD34+ MDP-1), pour finir par les cellules dendritiques matures (Dendritic cell). Au cours de la maturation des cellules dendritiques, l'entropie augmente pour atteindre un maximum au niveau de la jonction entre les populations cellules souches en cycle, progéniteurs multilignés, et progéniteurs lymph-myéloides. Puis l'entropie redescend pour atteindre un minimum dans les cellules dendritiques matures. On observe qu'il existe au cours de la maturation dendritique une

augmentation de la variabilité de l'expression génique qui diminue ensuite dans les cellules les plus matures.

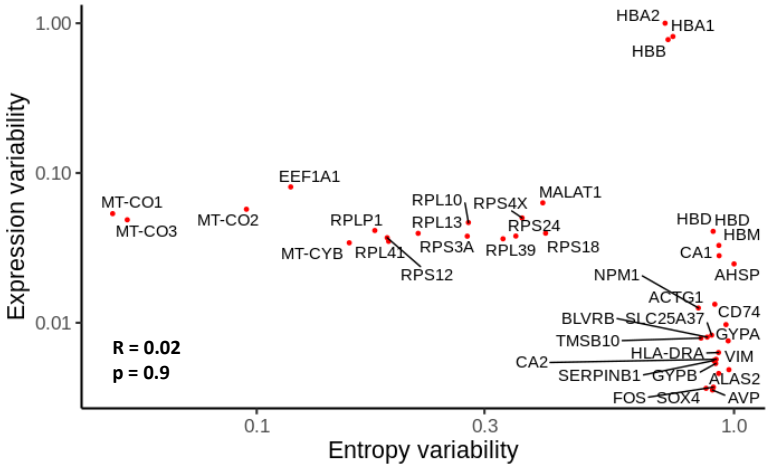
Ainsi, nos données indiquent pour la première fois à notre connaissance, qu'il existe au cours de l'hématopoïèse humaine normale, un phénomène d'augmentation de la variabilité de l'expression des gènes qui est une caractéristique commune aux différentes voies de différenciation. A ce stade, nous avons cherché à identifier les gènes impliqués dans ce pic d'entropie et spécifiques à chaque processus de différenciation.

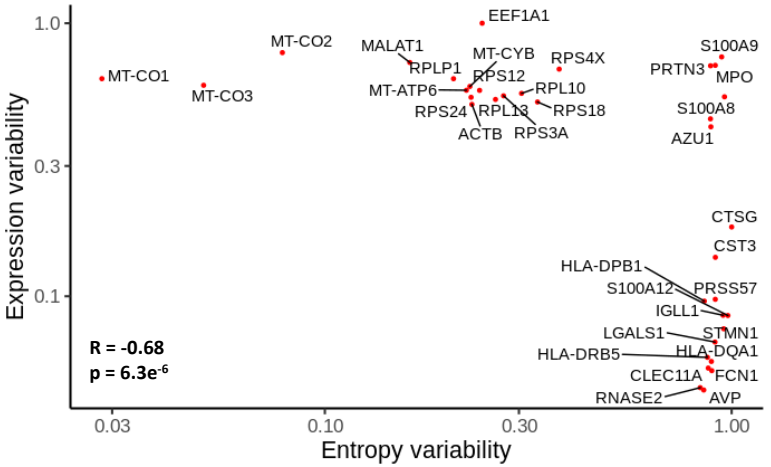
3.2. Identification des gènes les plus variablement entropiques au sein de chaque voie de différenciation.

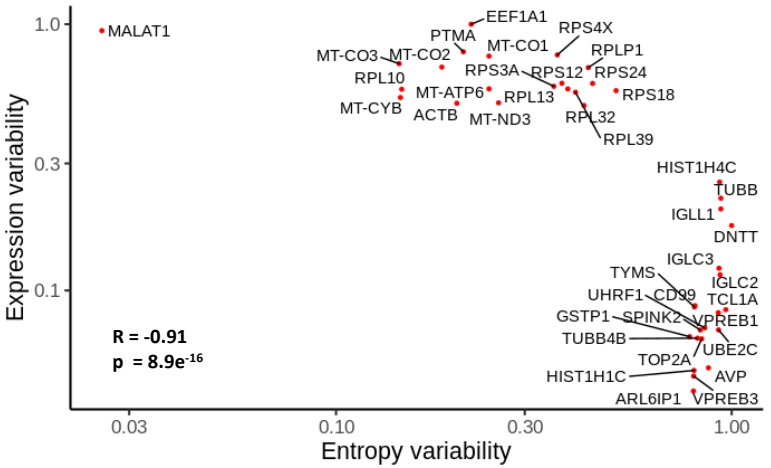
Dans un premier temps, nous avons voulu comparer la variation d'entropie et la variation d'expression des gènes au cours de la différenciation. Pour avoir des valeurs comparables d'entropie et d'expression, nous avons calculé l'expression moyenne de chaque gène sur une fenêtre glissante se déplaçant le long du pseudotemps, de la même manière que celle utilisée pour le calcul de l'entropie. En prenant en compte la totalité des gènes étudiés (15962), il semble que pour une grande majorité d'entre eux, plus l'entropie varie au cours de la différenciation, plus l'expression varie. Il apparait tout de même que certains gènes ont une variation d'expression importante qui n'est pas corrélée à la variation d'entropie. Nous avons alors comparé les 20 gènes dont la variation d'expression est la plus grande et les 20 gènes dont la variation d'entropie est la plus grande. Pour la plupart de ces gènes, les variations d'expression et d'entropie ne sont pas corrélées (érythropoïèse) voire même corrélées négativement (granulopoïèse, lymphopoïèse B et maturation dendritique).Pour l'érythropoïèse et la granulopoïèse, certains gènes ont une forte variation d'expression et ont également une forte variation d'entropie (*HBA1*, *HBA2*, *HBB, HBD, HBM, CA1, AHSP* dans l'érythropoïèse ; *S100A8*, *S100A9*, *MPO*, *PRTN3*, *AZU1* dans la granulopoïèse) **(Figure 36)**. Il est intéressant de noter que ces gènes sont connus pour être spécifiques respectivement de l'érythropoïèse et de la granulopoïèse.


A
Erythropoïèse
Granulopoïèse
Lymphopoïèse B
Maturation Dendritique
Expression variability
Entropy variability
1e+00
1e-01
1e-02
1e-03
1e-04
0.03
0.10
0.30
1.00
R = 0.12
p < 2.2e-16
R = 0.29
p < 2.2e-16
R = 0.3
p < 2.2e-16
R = 0.22
p < 2.2e-16

B


Erythropoïèse
Expression variability
Entropy variability
1.00
0.10
0.01
0.1
0.3
1.0
R = 0.02
p = 0.9
HBA2
HBA1
HBB
EEF1A1
MT-CO1
MT-CO3
MT-CO2
RPLP1
MT-CYB
RPL41
RPS12
RPL10
RPS4X
RPL13
RPS3A
MALAT1
RPS24
RPL39
RPS18
HBD
HBM
CA1
AHSP
NPM1
ACTG1
CD74
BLVRB
SLC25A37
GYPA
TMSB10
HLA-DRA
VIM
CA2
SERPINB1
GYPB
ALAS2
FOS
SOX4
AVP


Granulopoïèse
Expression variability
Entropy variability
1.0
0.3
0.1
0.03
0.10
0.30
1.00
R = -0.68
p = 6.3e-6
EEF1A1
MT-CO2
MALAT1
MT-CYB
RPS4X
S100A9
MT-CO1
RPLP1
RPS12
PRTN3
MPO
MT-CO3
MT-ATP6
RPL10
S100A8
RPS24
RPL13
RPS18
AZU1
ACTB
RPS3A
CTSG
CST3
HLA-DPB1
S100A12
PRSS57
IGLL1
LGALS1
STMN1
HLA-DRB5
HLA-DQA1
CLEC11A
FCN1
RNASE2
AVP


Lymphopoïèse B
Expression variability
Entropy variability
1.0
0.3
0.1
0.03
0.10
0.30
1.00
R = -0.91
p = 8.9e-16
MALAT1
EEF1A1
PTMA
RPS4X
MT-CO1
MT-CO3
MT-CO2
RPLP1
RPS3A
RPS12
RPS24
RPL10
MT-ATP6
RPL13
RPS18
MT-CYB
ACTB
MT-ND3
RPL32
RPL39
HIST1H4C
TUBB
IGLL1
DNTT
IGLC3
TYMS
IGLC2
UHRF1
CD99
TCL1A
GSTP1
SPINK2
VPREB1
TUBB4B
UBE2C
TOP2A
HIST1H1C
AVP
ARL6IP1
VPREB3

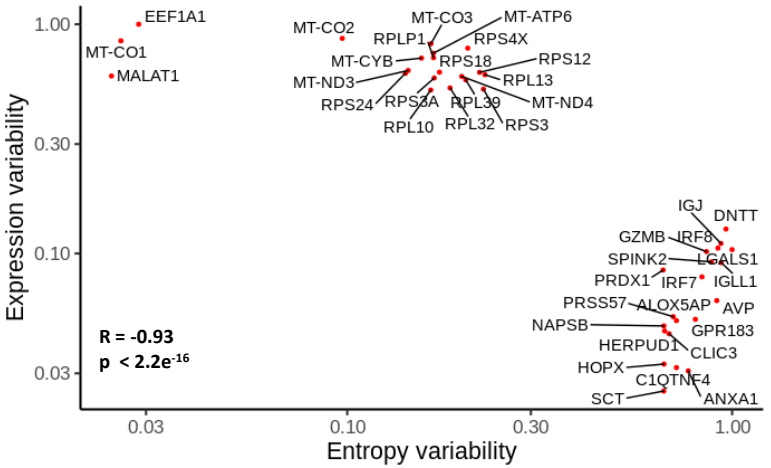

Maturation Dendritique
Expression variability
Entropy variability
1.00
0.30
0.10
0.03
0.10
0.30
1.00
R = -0.93
p < 2.2e-16
EEF1A1
MT-CO1
MALAT1
MT-CO2
MT-CO3
MT-ATP6
RPLP1
RPS4X
MT-CYB
RPS18
RPS12
MT-ND3
RPL13
RPS24
RPS3A
RPL39
MT-ND4
RPL10
RPL32
RPS3
IGJ
DNTT
GZMB
IRF8
SPINK2
LGALS1
PRDX1
IRF7
IGLL1
PRSS57
ALOX5AP
AVP
NAPSB
GPR183
HERPUD1
CLIC3
HOPX
C1QTNF4
SCT
ANXA1

Figure 36 : Corrélation entre la variation d'entropie et la variation d'expression génique au cours de l'hématopoïèse.

Pour chaque gène (points rouges sur les graphiques), la variation d'entropie est représentée en fonction de la variation d'expression (échelle logarithmique), dans les différentes voies de différenciation de l'hématopoïèse. La variation d'entropie (Entropy variability), correspond pour chaque gène à la différence entre l'entropie maximale et l'entropie minimale au cours de chacune des voies de différenciation. La variation d'expression de chaque gène, correspond à la différence entre le maximum et le minimum d'expression au cours des différenciations. **A)** tous les gènes (15962) ; **B)** 20 gènes dont les variations d'entropie et / ou d'expression génique sont les plus grande). **R :** coéfficient de corrélation. **p** : p-valeur.

Pour chacune des voies de différenciations étudiées, parmi les 20 gènes dont la variation d'entropie est la plus forte au cours de la différenciation, on observe des gènes qui jouent un rôle important dans le processus de différenciation ainsi que dans la fonction des cellules les plus matures **(Figure 37).** En effet pour l'érythropoïèse, cette liste comprend des gènes de globines (*HBD, HBM*), des gènes impliqués dans la synthèse de l'hème (*BLVRB, ALAS2, SLC25A37),* dans l'assemblage de l'hémoglobine (*AHSP*), la respiration (*CA1, CA2*), ainsi que des gènes qui codent pour des glycoprotéines transmembranaires des érythrocytes *(GYPA, GYPB).* Pour la granulopoïèse, on retrouve parmi les gènes les plus différentiellement entropiques, des gènes d'alarmines (*S100A8, S100A9, S100A12*), des gènes codants pour des protéines antibactériennes et antivirales (*AZU1, MPO*, *PRTN3, ELANE, CTSG, CST3, RNASE2*), des molécules présentatrice d'antigène (*HLA-DPB1, HLA-DQA1, HLA-DRB5*), et des lectines (*FCN1, CLEC11A*). Concernant la lymphopoïèse B, on retrouve parmi les 20 gènes les plus variablement entropiques, des gènes impliqués dans la formation du pré-BCR (*VPREB1, VPREB3, IGLL1*), les chaines légères d'immunoglobulines (*IGLC2, IGLC3*), et l'ajout de nucléotides lors des réarrangements des chaines d'immunoglobulines (*DNTT*). Pour la maturation dendritique, la liste contient des gènes régulant la réponse à l'interféron et qui sont des marqueurs des populations dendritiques matures et progénitrices (*IRF7, IRF8*), des gènes jouant un rôle dans la réponse inflammatoire (*ALOX5AP, ANXA1*), l'immunité innée (*GZMB, C1QTNF4*), dans la différenciation dendritique (*LGALS1*), et également un autre marqueur des populations dendritiques (*IGJ*). On peut noter que le gène commun aux quatre voies de différenciation : *AVP*, est un gène signature des populations cellules souches hématopoïétiques et cellules souches en cycle**.**

Nous démontrons ainsi que les gènes les plus variablement entropiques sont spécifiques et impliqués dans des processus clés des voies de différenciation hématopoïétique étudiées.

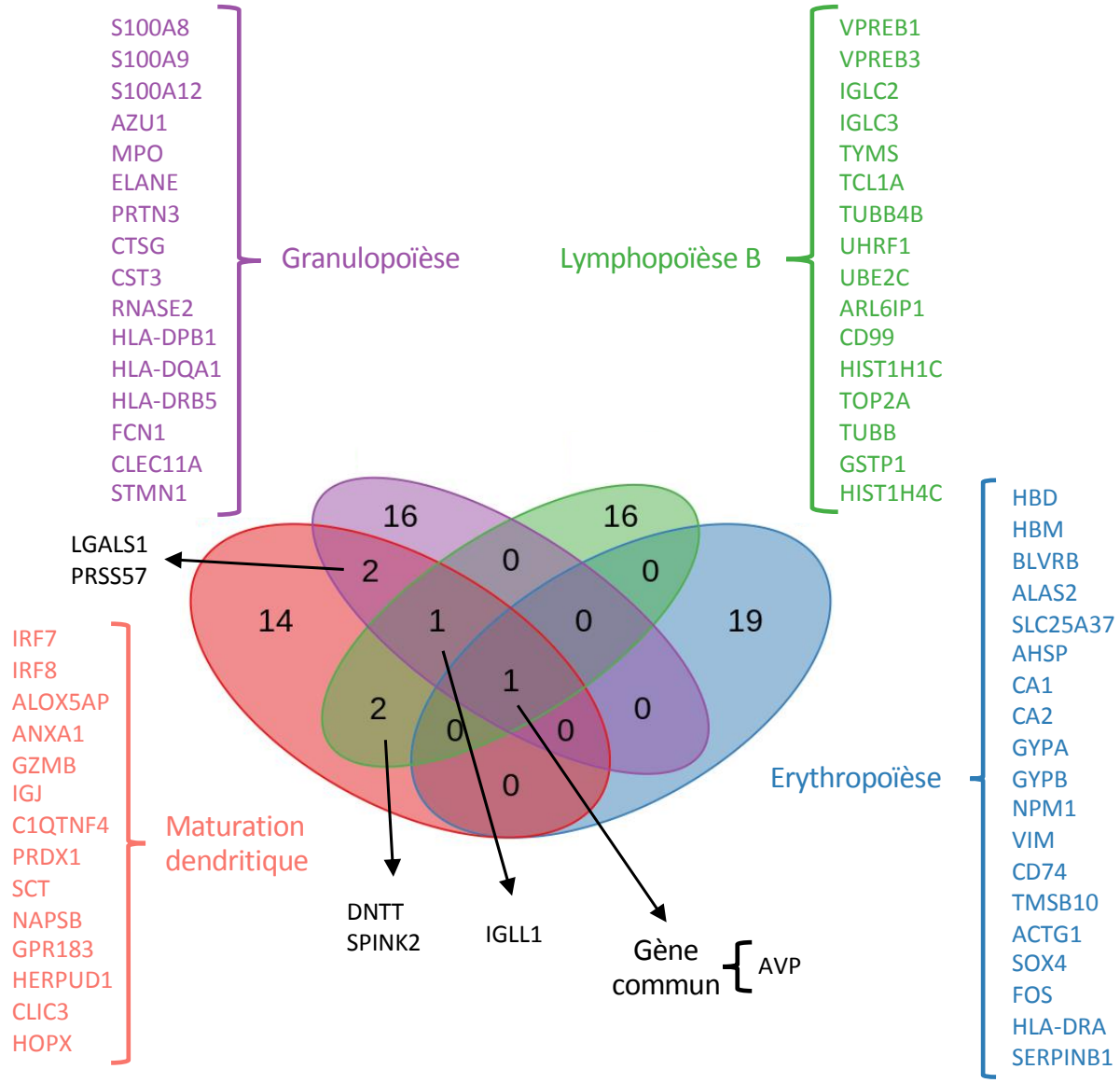


Figure 37 : Diagramme de Venn des 20 gènes dont la variation d'entropie est la plus forte au cours des différentes voies de différenciation hématopoïétiques.

Les 20 gènes les plus variablement entropiques des 4 voies de différenciation étudiées sont comparés pour voir lesquels leur sont communs ou spécifiques.

Nous avons ensuite voulu comparer les 20 gènes dont les variations d'expression sont les plus fortes au cours des 4 voies de différenciation étudiées. Même si certains de ces gènes sont spécifiques d'une voie de différenciation, la grande majorité des gènes identifiés (n=14) sont communs aux 4 voies de différenciation **(Figure 38).** Ces gènes codent pour des protéines ribosomiques (*RPS4X, RPL13, RPLP1, RPL10, RPS24, RPS12, RPS18, RPS3A*), sont impliqués

dans la traduction (*EEF1A*), les processus de respiration cellulaire (*MT-CO1, MT-CO2, MT-CO3, MT-CYB*), ou codent pour un des long ARN non codant les plus exprimés dans les cellules (*MALAT1*) qui a de nombreuses fonctions qui ne sont pas encore très bien définies[197].

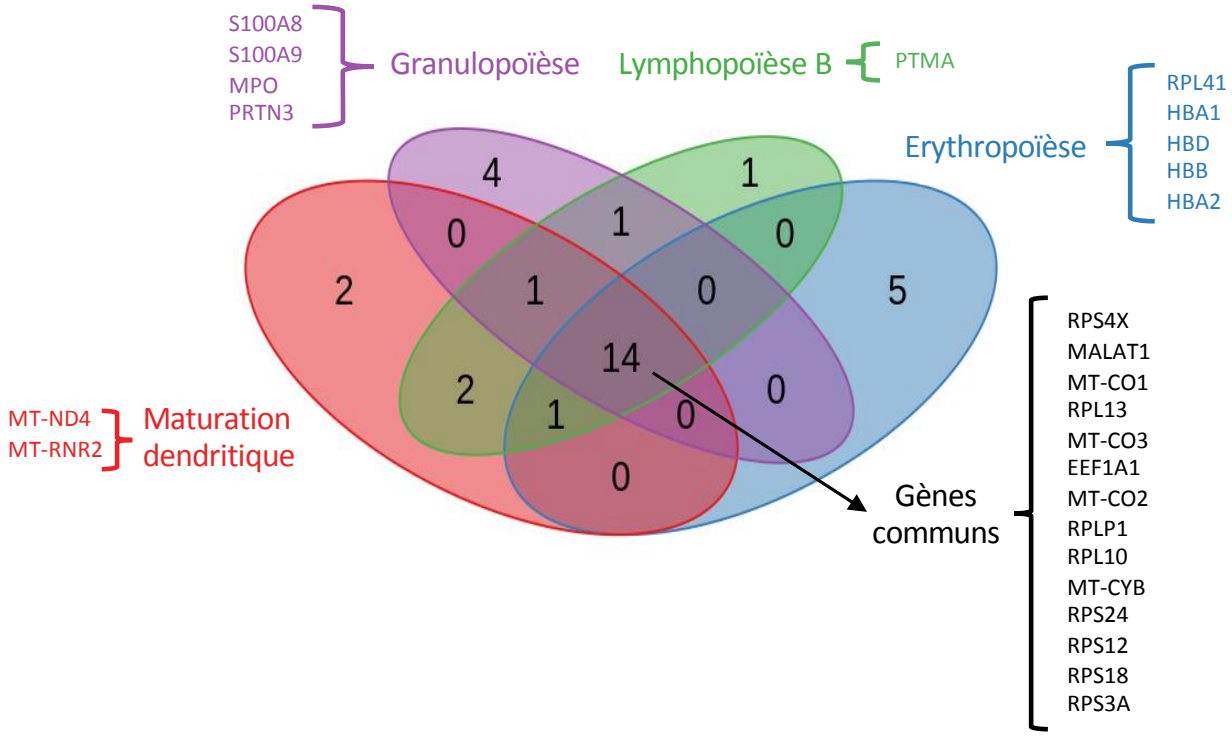


Figure 38 : Diagramme de Venn des 20 gènes dont la variation d'expression est la plus forte au cours des différentes voies de différenciation hématopoïétiques.
Les 20 gènes dont la variation d'expression est la plus importante des 4 voies de différenciation étudiées sont comparés pour voir lesquels leur sont communs ou spécifiques.

Ces résultats montrent que les gènes dont la variation d'expression est la plus importante et qui sont communs aux 4 voies de différenciations étudiées sont impliqués dans des mécanismes généraux tel que la synthèse protéique et le métabolisme énergétique. Ces gènes présentent également une variation d'entropie beaucoup plus faible que les autres gènes.

3.3. Paysage transcriptionnel du compartiment souche et progéniteurs chez les sujets sains âgés et les SMD

Nous avons mis en place les outils bioinformatiques nécessaires à l'analyse des données de scRNA-Seq à partir de données publiques et établi qu'il existait des variations d'entropie au cours des principaux processus de différenciation hématopoïétique. Nous avons ensuite voulu transposer cette approche à nos propres échantillons, normaux et pathologiques de patients porteurs de SMD, pathologie hématologique ou des anomalies des processus de différenciation sont observés à différents niveaux.

Nous avons décidé de caractériser le paysage transcriptionnel de cellules de moelle de patients porteurs d'un SMD avec mutation du gène *SF3B1* et de rechercher des anomalies qui pourraient nous aider à mieux comprendre la physiopathologie de ces maladies. Nous avons choisi de comparer ces échantillons pathologiques à des sujets sains appariés en âge. Les dons de moelle osseuse en France sont limités aux personnes ayant moins de 60 ans. Les patients dont on prélève la moelle dans un but diagnostic présentent des anomalies clinico-biologiques susceptibles d'avoir une influence sur l'hématopoïèse. C'est pourquoi nous avons choisi d'extraire les cellules médullaires issues de tête fémorale de sujets âgés dont l'hémogramme est normal. Au moment du prélèvement, ces patients ne prenaient pas de médicaments susceptibles d'interférer avec l'hématopoïèse (immunosupresseurs, chimiothérapie). Nous avons donc utilisé dans cette étude les cellules souches et progénitrices (HSPC) issues de deux témoins sains et de deux patients atteint de SMD avec sidéroblastes en couronnes au diagnostic **(Tableau 3)**. La technologie chromium 10x genomics appliquée à ces échantillons nous a permis de disposer de données transcriptomiques à l'échelle unicellulaire d'un total de 12689 HSPC réparties sur nos 4 échantillons (2252 provenant de Ctrl1, 3347 provenant de Ctrl3, 3302 provenant de MDS2, et 3788 provenant de MDS4) **(Figure 39A).**

Pour générer une carte de référence de toutes les cellules analysées, nous avons combiné toutes les cellules provenant des 4 échantillons avec la méthode d'intégration implémenté dans Seurat[194]. Les cellules ont été annotées par SingleR à partir des profils d'expression génique des populations hématopoïétiques décrites par Hay et al[93]. La UMAP qui en résulte distingue 21 sous types cellulaires différents qui sont organisés selon les voies de

différenciations hématopoïétiques principales (érythropoïèse, granulopoïèse, lymphopoïèse et maturation dendritique) **(Figure 39B).** Nous avons comparé l'annotation par SingleR des clusters déterminés par Seurat avec l'annotation cellule par cellule. Les deux méthodes sont cohérentes, mais l'annotation cellule par cellule étant plus précise nous avons choisi de l'utiliser pour les analyses présentées dans la suite du manuscrit **(Figure 40)**.

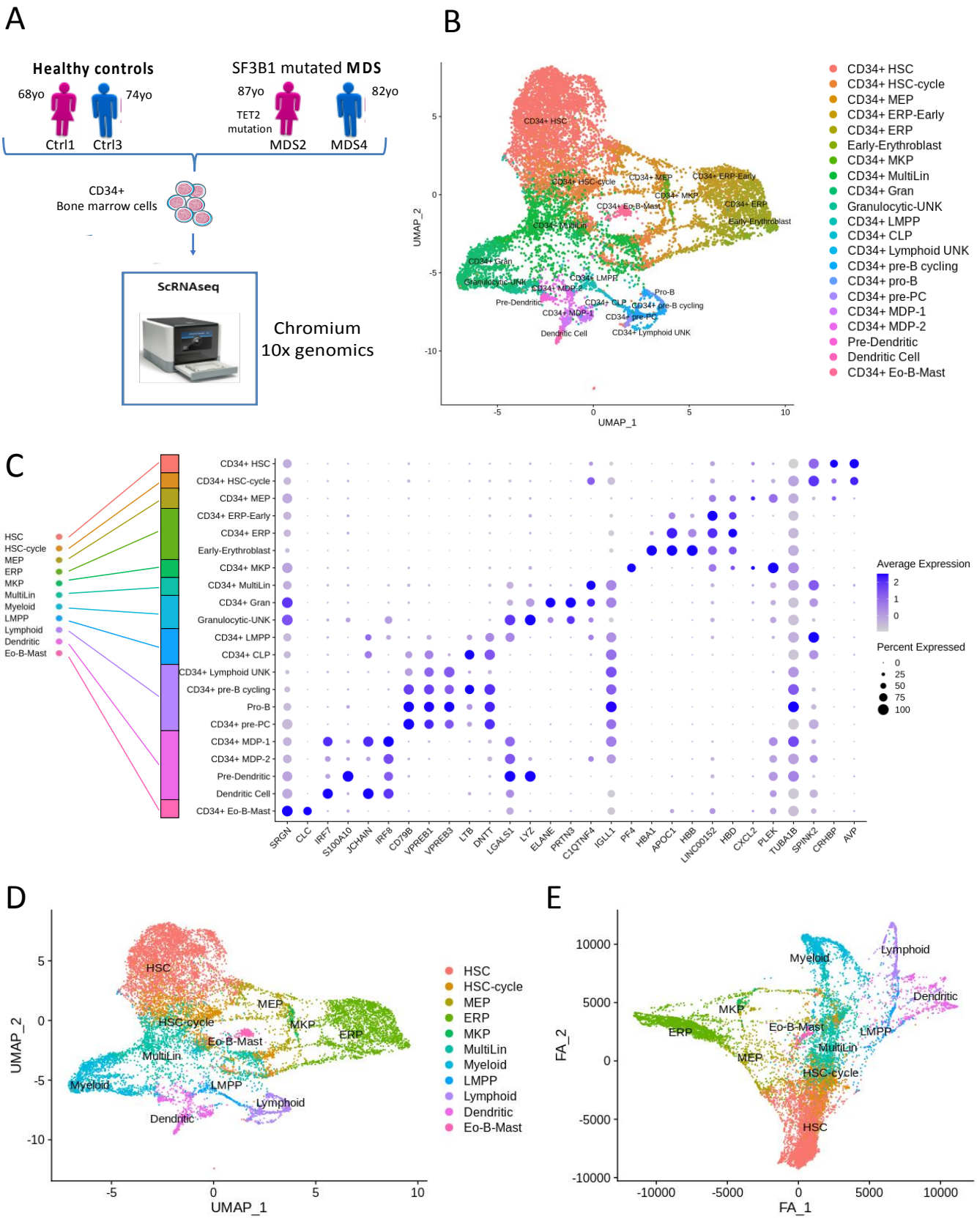
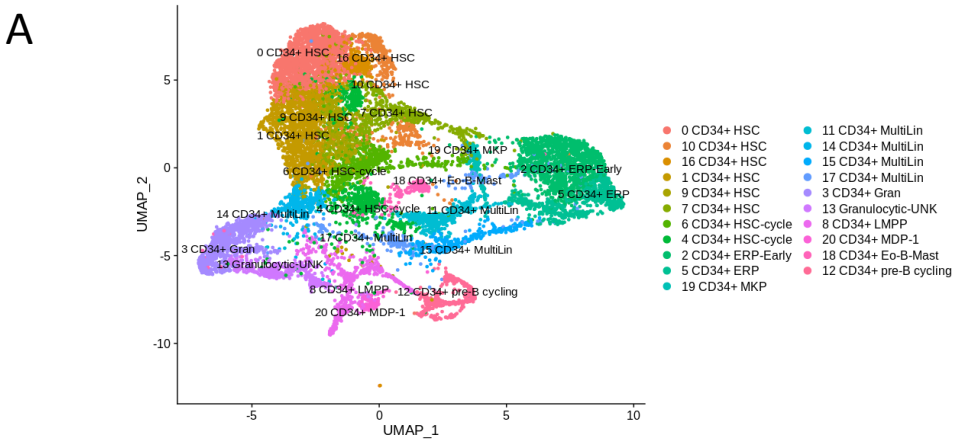
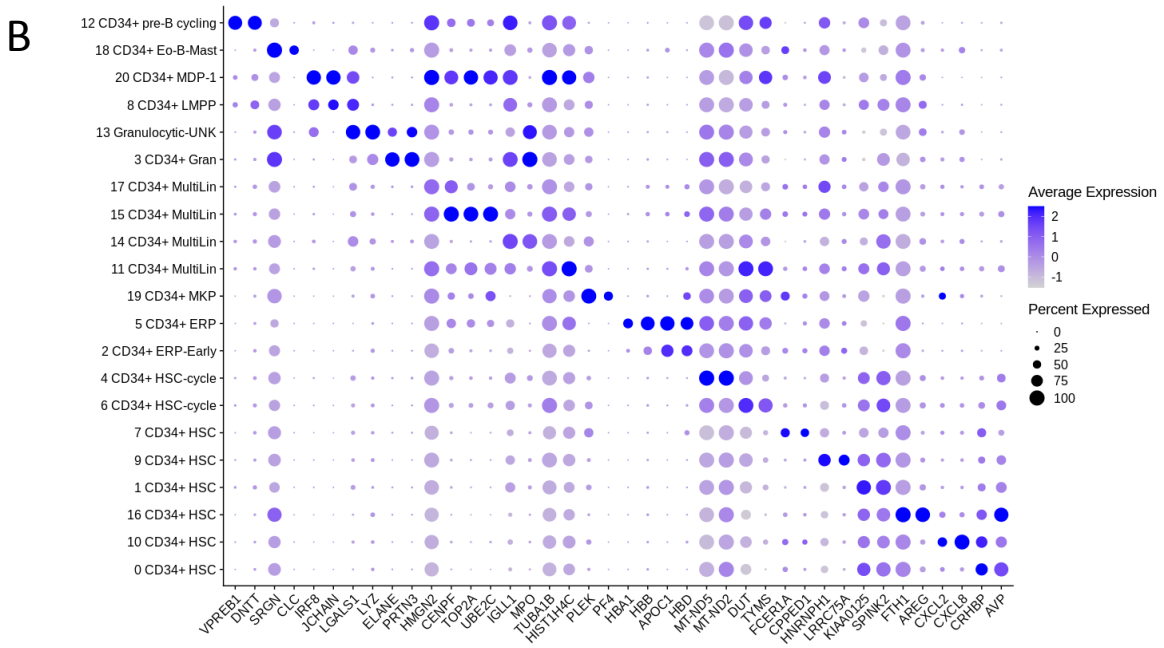

A
Healthy controls
68yo
74yo
Ctrl1
Ctrl3
SF3B1 mutated MDS
87yo
82yo
TET2 mutation
MDS2
MDS4
CD34+ Bone marrow cells
ScRNAseq
Chromium 10x genomics
B
CD34+ HSC
CD34+ HSC-cycle
CD34+ MEP
CD34+ ERP-Early
CD34+ ERP
Early-Erythroblast
CD34+ MKP
CD34+ MultiLin
CD34+ Gran
Granulocytic-UNK
CD34+ LMPP
CD34+ CLP
CD34+ Lymphoid UNK
CD34+ pre-B cycling
CD34+ pro-B
CD34+ pre-PC
CD34+ MDP-1
CD34+ MDP-2
Pre-Dendritic
Dendritic Cell
CD34+ Eo-B-Mast
UMAP_1
UMAP_2
C
HSC
HSC-cycle
MEP
ERP
MKP
MultiLin
Myeloid
LMPP
Lymphoid
Dendritic
Eo-B-Mast
SRGN
CLC
IRF7
S100A10
JCHAIN
IRF8
CD79B
VPREB1
VPREB3
LTB
DNTT
LGALS1
LYZ
ELANE
PRTN3
C1QTNF4
IGLL1
PF4
HBA1
APOC1
HBB
LINC00152
HBD
CXCL2
PLEK
TUBA1B
SPINK2
CRHBP
AVP
Average Expression
Percent Expressed
D
UMAP_1
UMAP_2
E
FA_1
FA_2

Figure 39 : Paysage trancriptionnel du compartiment HSPC des SMD SF3B1 mutés et sujets âgés

A) Les cellules souches et progénitrices CD34+ des témoins (Ctrl1, Ctrl3) et des SMD au diagnostic (MDS2, MDS4) sont isolées de la moelle osseuse par tri immunomagnétique. L'analyse par scRNA-Seq est effectuée selon la technologie chromium 10x genomics. **B)** Représentation UMAP en deux dimensions de la matrice gènes cellules (12689 cellules, 19420 gènes) intégrées des 4 échantillons. Les cellules annotées par SingleR sont classées en 21 sous types différents représenté chacun par une couleur différente. **C)** Les gènes les plus différentiellement exprimés de chaque sous population annotées par SingleR en comparaison avec toutes les autres ont été déterminés par la fonction FindMarkers de Seurat. Les deux gènes les plus spécifiques de chaque sous population sont représentés sur ce graphique. La couleur du cercle indique les valeurs d'expression génique et la taille du cercle représente la proportion de cellules exprimant le marqueur dans la sous population. **D-E)** représentation par UMAP et FA des sous populations HSPC regroupées par voie de différenciation.


A
0 CD34+ HSC
16 CD34+ HSC
10 CD34+ HSC
9 CD34+ HSC
7 CD34+ HSC
1 CD34+ HSC
19 CD34+ MKP
6 CD34+ HSC-cycle
18 CD34+ Eo-B-Mast
2 CD34+ ERP-Early
5 CD34+ ERP
4 CD34+ HSC-cycle
11 CD34+ MultiLin
14 CD34+ MultiLin
17 CD34+ MultiLin
3 CD34+ Gran
15 CD34+ MultiLin
13 Granulocytic-UNK
8 CD34+ LMPP
12 CD34+ pre-B cycling
20 CD34+ MDP-1
UMAP_2
UMAP_1
5
0
-5
-10
-5
0
5
10


B
12 CD34+ pre-B cycling
18 CD34+ Eo-B-Mast
20 CD34+ MDP-1
8 CD34+ LMPP
13 Granulocytic-UNK
3 CD34+ Gran
17 CD34+ MultiLin
15 CD34+ MultiLin
14 CD34+ MultiLin
11 CD34+ MultiLin
19 CD34+ MKP
5 CD34+ ERP
2 CD34+ ERP-Early
4 CD34+ HSC-cycle
6 CD34+ HSC-cycle
7 CD34+ HSC
9 CD34+ HSC
1 CD34+ HSC
16 CD34+ HSC
10 CD34+ HSC
0 CD34+ HSC
VPREB1 DNTT SRGN CLC IRF8 JCHAIN LGALS1 LYZ ELANE PRTN3 HMGN2 CENPF TOP2A UBE2C IGLL1 MPO TUBA1B HIST1H4C PLEK PF4 HBA1 HBB APOC1 HBD MT-ND5 MT-ND2 DUT TYMS FCER1A CPPED1 HNRNPHL LRRC75A KIAA0125 SPINK2 FTH1 AREG CXCL2 CXCL8 CRHBP AVP
Average Expression
2
1
0
-1
Percent Expressed
0
25
50
75
100

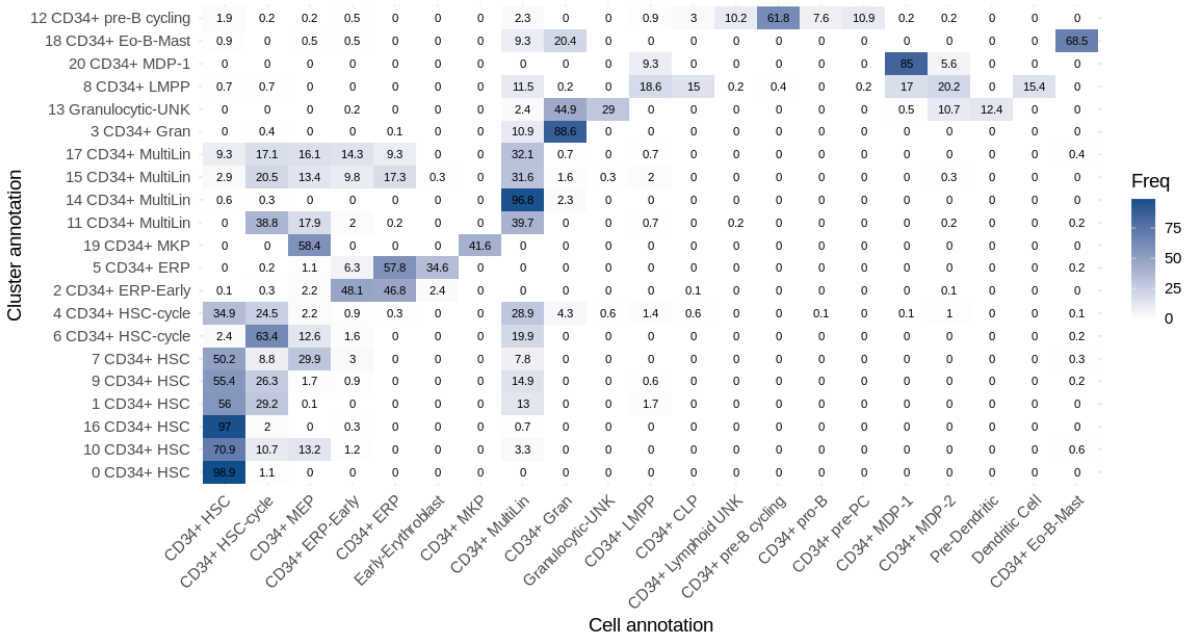

C
Cluster annotation
12 CD34+ pre-B cycling 1.9 0.2 0.2 0.5 0 0 0 2.3 0 0 0.9 3 10.2 61.8 7.6 10.9 0.2 0.2 0 0 0
18 CD34+ Eo-B-Mast 0.9 0 0.5 0.5 0 0 0 9.3 20.4 0 0 0 0 0 0 0 0 0 0 0 68.5
20 CD34+ MDP-1 0 0 0 0 0 0 0 0 0 0 9.3 0 0 0 0 0 85 5.6 0 0 0
8 CD34+ LMPP 0.7 0.7 0 0 0 0 0 11.5 0.2 0 18.6 15 0.2 0.4 0 0.2 17 20.2 0 15.4 0
13 Granulocytic-UNK 0 0 0 0.2 0 0 0 2.4 44.9 29 0 0 0 0 0 0 0.5 10.7 12.4 0 0
3 CD34+ Gran 0 0.4 0 0 0.1 0 0 10.9 88.6 0 0 0 0 0 0 0 0 0 0 0 0
17 CD34+ MultiLin 9.3 17.1 16.1 14.3 9.3 0 0 32.1 0.7 0 0.7 0 0 0 0 0 0 0 0 0 0.4
15 CD34+ MultiLin 2.9 20.5 13.4 9.8 17.3 0.3 0 31.6 1.6 0.3 2 0 0 0 0 0 0 0.3 0 0 0
14 CD34+ MultiLin 0.6 0.3 0 0 0 0 0 96.8 2.3 0 0 0 0 0 0 0 0 0 0 0 0
11 CD34+ MultiLin 0 38.8 17.9 2 0.2 0 0 39.7 0 0 0.7 0 0.2 0 0 0 0 0.2 0 0 0.2
19 CD34+ MKP 0 0 58.4 0 0 0 41.6 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5 CD34+ ERP 0 0.2 1.1 6.3 57.8 34.6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.2
2 CD34+ ERP-Early 0.1 0.3 2.2 48.1 46.8 2.4 0 0 0 0 0 0.1 0 0 0 0 0 0.1 0 0 0
4 CD34+ HSC-cycle 34.9 24.5 2.2 0.9 0.3 0 0 28.9 4.3 0.6 1.4 0.6 0 0 0.1 0 0.1 1 0 0 0.1
6 CD34+ HSC-cycle 2.4 63.4 12.6 1.6 0 0 0 19.9 0 0 0 0 0 0 0 0 0 0 0 0 0.2
7 CD34+ HSC 50.2 8.8 29.9 3 0 0 0 7.8 0 0 0 0 0 0 0 0 0 0 0 0 0.3
9 CD34+ HSC 55.4 26.3 1.7 0.9 0 0 0 14.9 0 0 0.6 0 0 0 0 0 0 0 0 0 0.2
1 CD34+ HSC 56 29.2 0.1 0 0 0 0 13 0 0 1.7 0 0 0 0 0 0 0 0 0 0
16 CD34+ HSC 97 2 0 0.3 0 0 0 0.7 0 0 0 0 0 0 0 0 0 0 0 0 0
10 CD34+ HSC 70.9 10.7 13.2 1.2 0 0 0 3.3 0 0 0 0 0 0 0 0 0 0 0 0 0.6
0 CD34+ HSC 98.9 1.1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
CD34+ HSC
CD34+ HSC-cycle
CD34+ MEP
CD34+ ERP-Early
CD34+ ERP
Early-Erythroblast
CD34+ MKP
CD34+ MultiLin
CD34+ Gran
Granulocytic-UNK
CD34+ LMPP
CD34+ CLP
CD34+ Lymphoid UNK
CD34+ pre-B cycling
CD34+ pro-B
CD34+ pre-PC
CD34+ MDP-1
CD34+ MDP-2
Pre-Dendritic
Dendritic Cell
CD34+ Eo-B-Mast
Cell annotation
Freq
75
50
25
0

Figure 40 : Annotation par clusters du compartiment HSPC des patients SMD et des témoins sains âgés.

A) Représentation par UMAP du compartiment HSPC intégré des 4 échantillons. Les 21 clusters distingués par Seurat et annotées par SingleR sont représentés chacun par une couleur différente**. B)** Les gènes les plus différentiellements exprimés de chaque cluster annoté par SingleR en comparaison avec toutes les autres ont été déterminés par la fonction FindMarkers de Seurat. Les deux gènes les plus spécifiques de chaque sous population sont représentés sur ce graphique. La couleur du cercle indique les valeurs d'expression génique et la taille du cercle représente la proportion de cellules exprimant le marqueur dans la sous population. **C)** Comparaison de l'annotation par cluster avec l'annotation par cellule, chaque case représente le pourcentage de cellule du cluster correspondant à l'annotation cellule par cellule.

Les marqueurs spécifiques des sous populations distinguées par SingleR sont en accord avec les études précédemment réalisées sur le compartiment HSPC médullaire[89,93,195,198] **(Figure 39C).** Pour faciliter la compréhension et les analyses en aval, les populations cellulaires proches ont été regroupées en 11 populations. La représentation de ces populations par UMAP **(Figure 39D)** et FA **(Figure 39E),** permet de distinguer plus facilement les différentes voies de différenciation.

Il existe des différences dans la répartition des cellules entre témoins et SMD qu'il est difficile d'interpréter au vu du nombre d'échantillons comparés (**Figure 41A, Figure 42).**

Nous avons ensuite calculé le pourcentage de cellules en cycle (phase S, G2 et M) dans les sous-populations simplifiées avec l'algorithme d'assignation de phase du cycle cellulaire mis au point par Tirosh et al et implémenté dans Seurat[193]. En comparant les sujets âgés sains regroupés, avec les deux SMD, on observe que le patient MDS2 présente un pourcentage de cellules en cycle significativement supérieur dans la plupart des sous populations. On ne retrouve pas ces différences pour le patient MDS4 **(Figure 41B).** Cette différence de proportions de cellules en cycle pourrait être en lien avec la présence chez le patient MDS2 de la mutation du gène *TET2* dans ses cellules médullaires, en effet il a récemment été démontré que l'inactivation de *TET2* pouvait augmenter la prolifération cellulaire[199].

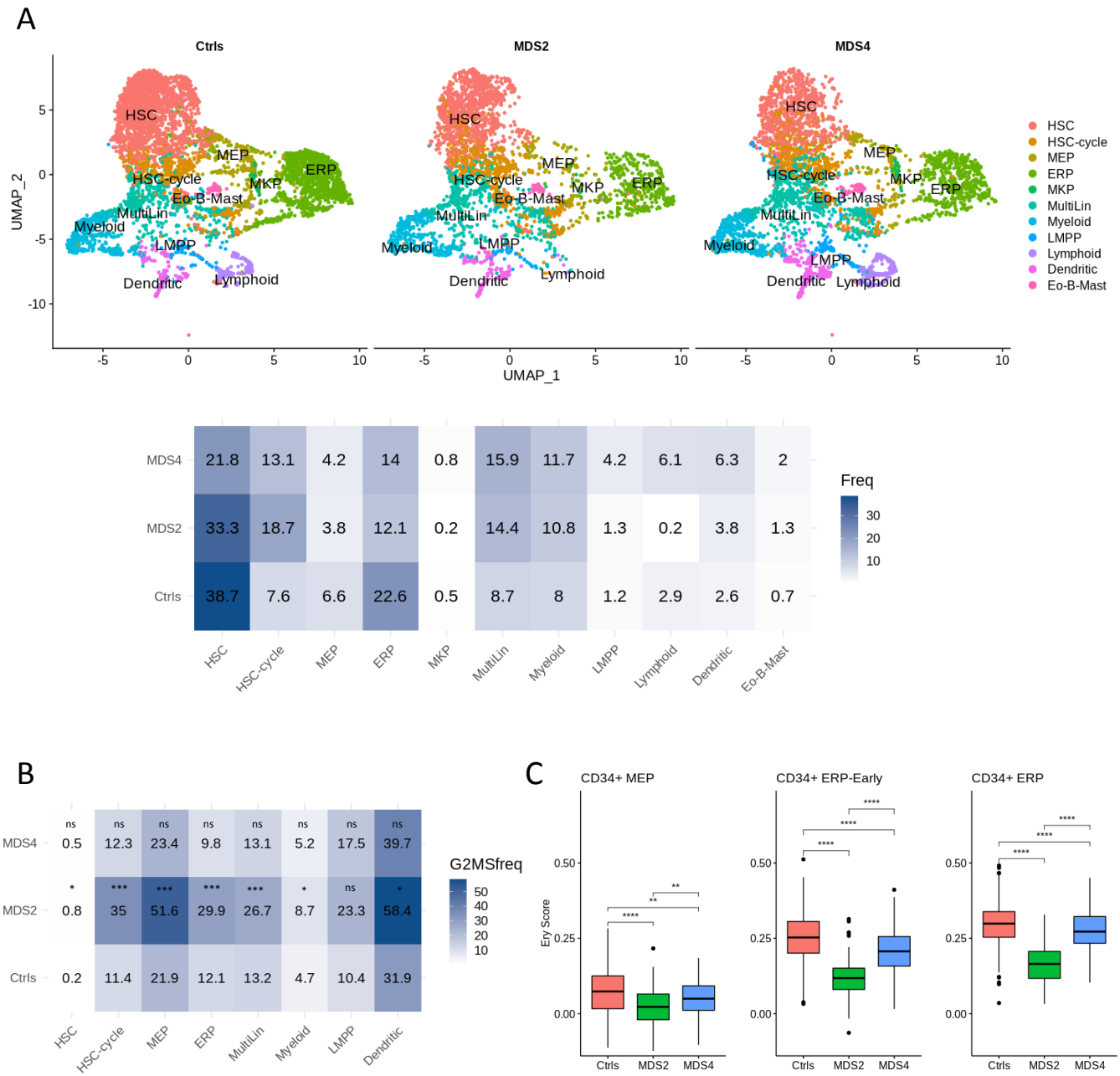


Figure 41 : Comparaison des échantillons de SMD avec les témoins matchés en âge.

A) Proportion relative par échantillon, de chaque sous population cellulaire simplifiée au sein du paysage transcritpitonnel des HSPC. Les Ctrl1 et Ctrl3 sont combinés pour cette analyse. **B)** Proportion relative de cellules en cycle (phase S, G2 et M) au sein de chaque sous population cellulaire simplifiée. Les proportions ont été comparés avec la fonction prop.test du package R stats. **C)** Un score d'amorçage de lignage érythroide (Ery Score) a été calculé en utilisant l'expression moyenne de gènes connus pour être régulés à la hausse au cours de l'érythropoïèse. L'Eryscore a été comparé entre les échantillons par un test de wilcoxon avec la fonction compare_means du package R ggpubr. (* : p<0.05 ; ** : p<0,01 ; *** : p<0,001 ; **** : p<0.0001)

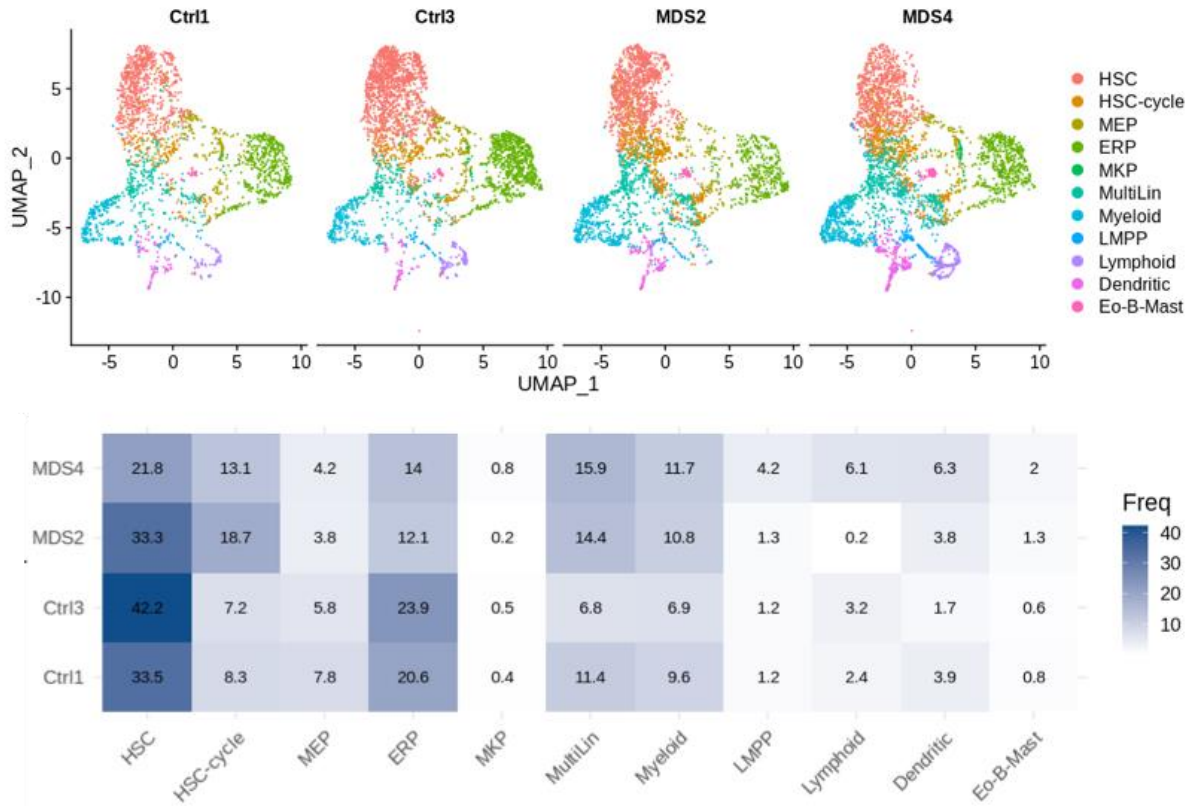


Figure 42 : Répartition des 11 sous populations cellulaires simplifiées des HSPC chez les SMD et témoins sains âgés.

Proportion relative par échantillon, de chaque sous population cellulaire simplifiée au sein du paysage transcriptionnel des HSPC

L'hématopoïèse dysplasique des SMD mutés pour *SF3B1* affecte principalement la lignée érythroïde. Nous avons donc voulu étudier plus précisément les progéniteurs de cette lignée. Pour explorer l'amorçage érythroïde des progéniteurs, nous avons calculé un score (Eryscore) à partir de 100 gènes régulés à la hausse pendant l'érythropoïèse humaine [87,200]. Cette analyse montre que l'Eryscore augmente au cours de l'érythropoïèse (des CD34+ MEP jusqu'aux CD34+ ERP). De plus ce score est diminué significativement dans les SMD par rapport aux témoins sains âgés. Il est également significativement plus bas chez MDS2 par rapport à MDS4 **(Figure 41C)**. Ces données suggèrent que dans les SMD mutés *SF3B1*, un amorçage érythroïde altéré pourrait être relié à la dysérythropoïèse et à l'anémie observée chez ces patients. La mutation de *TET2* connue pour entrainer un biais de différenciation granuleux, pourrait également jouer un rôle dans l'altération de l'amorçage érythroïde des progéniteurs.

Ainsi, l'ensemble de ces données suggèrent qu'au niveau de la répartition des sous populations, la hiérarchie hématopoïétique du compartiment HSPC de nos deux patients SMD mutés *SF3B1* est globalement conservée par rapport aux sujets sains âgés. Par contre, il existe une anomalie qualitative de l'érythropoïèse chez nos patients SMD avec une diminution de l'amorçage des progéniteurs vers la lignée érythroblastique, ce qui est en accord avec les connaissances actuelles sur la physiopathologie de ces entités.

3.4. Un pic de variabilité de l'expression génique est observé au cours de l'hématopoïèse chez les sujets sains âgés et les SMD de bas risque.

Nous avons démontré à partir de données de scRNA-Seq qu'il existe un pic de variabilité de l'expression génique au cours de l'hématopoïèse chez un sujet sain. Nous avons voulu confronter et confirmer cette observation sur des échantillons issus de sujets âgés, et de patients atteints de SMD. Nous avons dans un premier temps analysé les patients individuellement **(Figure 43).**

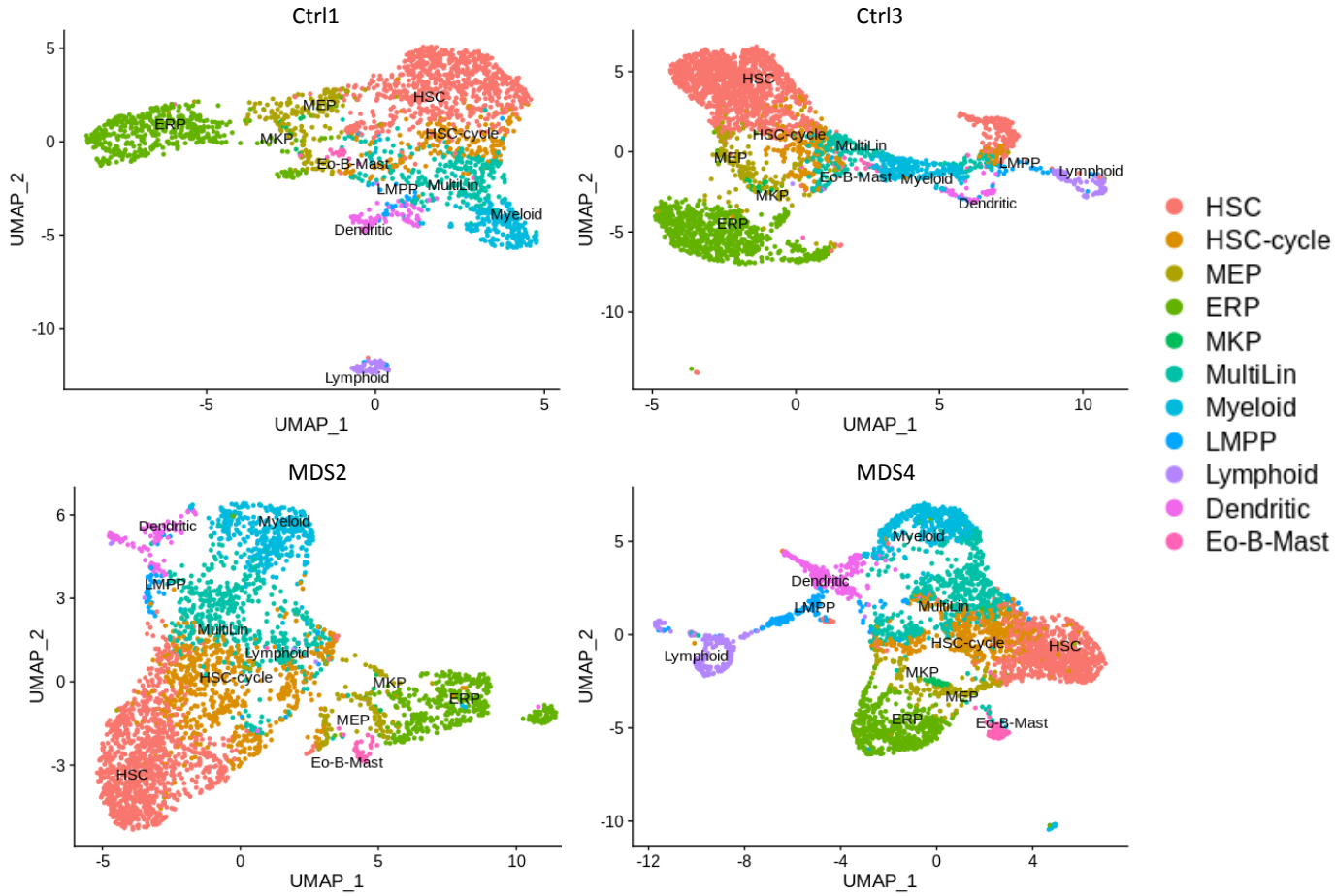


Figure 43 : Représentation par UMAP des HSPC des sujets âgés et des SMD analysés individuellement et réparties en 11 populations simplifiées.

Les cellules ont été annotées individuellement par SingleR, puis regroupées par type cellulaire pour une visualisation simplifiée. Chaque sous type cellulaire simplifié est représenté par une couleur différente.

Nous avons ensuite pour chaque patient et pour chaque voie de différenciation calculé l'entropie moyenne avec la même méthode que celle que nous avons utilisé pour l'échantillon de moelle normale de la publication de Granja et al[190]. Pour l'érythropoïèse, dans les 4 échantillons étudiés, on observe une augmentation de l'entropie au cours de la différenciation, mais celle-ci ne redescend pas. Ceci est probablement dû à l'absence d'une quantité suffisante de cellules érythroides matures **(Figure 44).** Concernant la granulopoïèse, l'évolution de l'entropie au cours de la différenciation est comparable avec celle observée dans la moelle normale. Un pic d'entropie apparait pour les 4 échantillons au niveau de la population des progéniteurs multilignés (CD34+ MultiLin), puis redescend chez la population des progéniteurs granuleux (CD34+ Gran) comme observé précédemment dans la moelle normale **(Figure 45).** A propos de la maturation dendritique, et de la lymphopoïèse B, on observe la même tendance que pour la granulopoïèse. En effet, l'entropie atteint un pic au niveau de la population des progéniteurs multilignés, puis redescends dans les populations plus matures **(Figure 46 et Figure 47).**

Dans les données obtenues sur la moelle totale d'un sujet sain jeune, on observe dans les populations les plus matures, une redescente de l'entropie en dessous du niveau basal (celui des CSH). Dans nos données chez les sujets âgés et les SMD, l'entropie ne redescend pas en dessous du niveau de base car notre analyse se limite au compartiment HSPC. En effet les populations les plus mature de ce compartiment sont encore des populations progénitrices et non des cellules matures dont la différenciation est terminée.

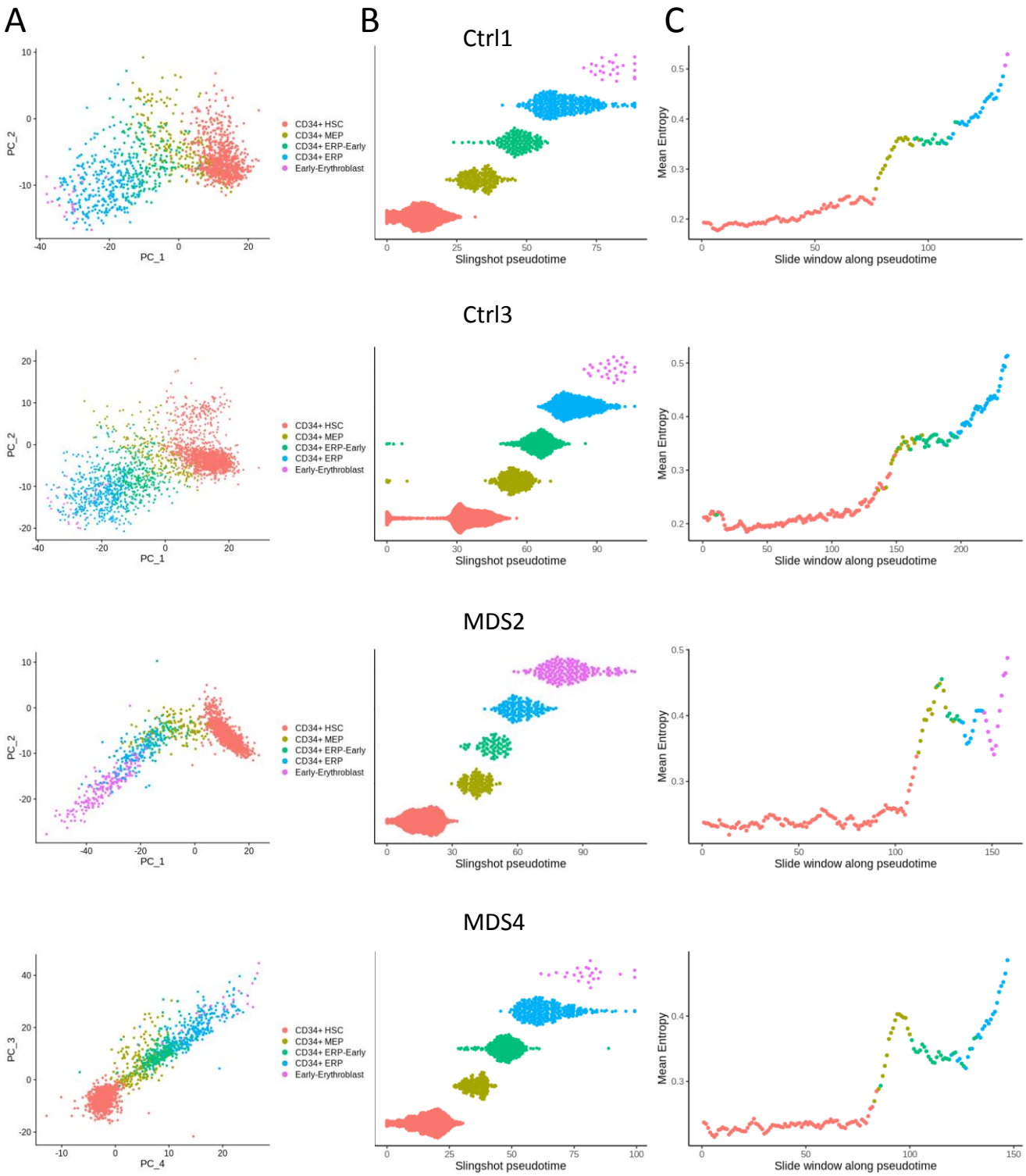


Figure 44 : Evolution de la moyenne d'entropie des populations HSPC au cours de l'érythropoïèse chez les SMD et les sujets âgés

A) Les populations cellulaires spécifiques de l'érythropoïèse sont sélectionnées et représentées en deux dimensions selon les coordonnées de l'ACP). **B)** Les cellules sont ordonnées selon le pseudotemps calculé par Slingshot. **C)** Chaque point correspond à la moyenne d'entropie de tous les gènes calculée sur une fenêtre de 50 cellules. La fenêtre avance le long du pseudotemps avec un pas de 10 cellules. La couleur de chaque point sur le graphique correspond à la nature de la première cellule de la fenêtre correspondante.

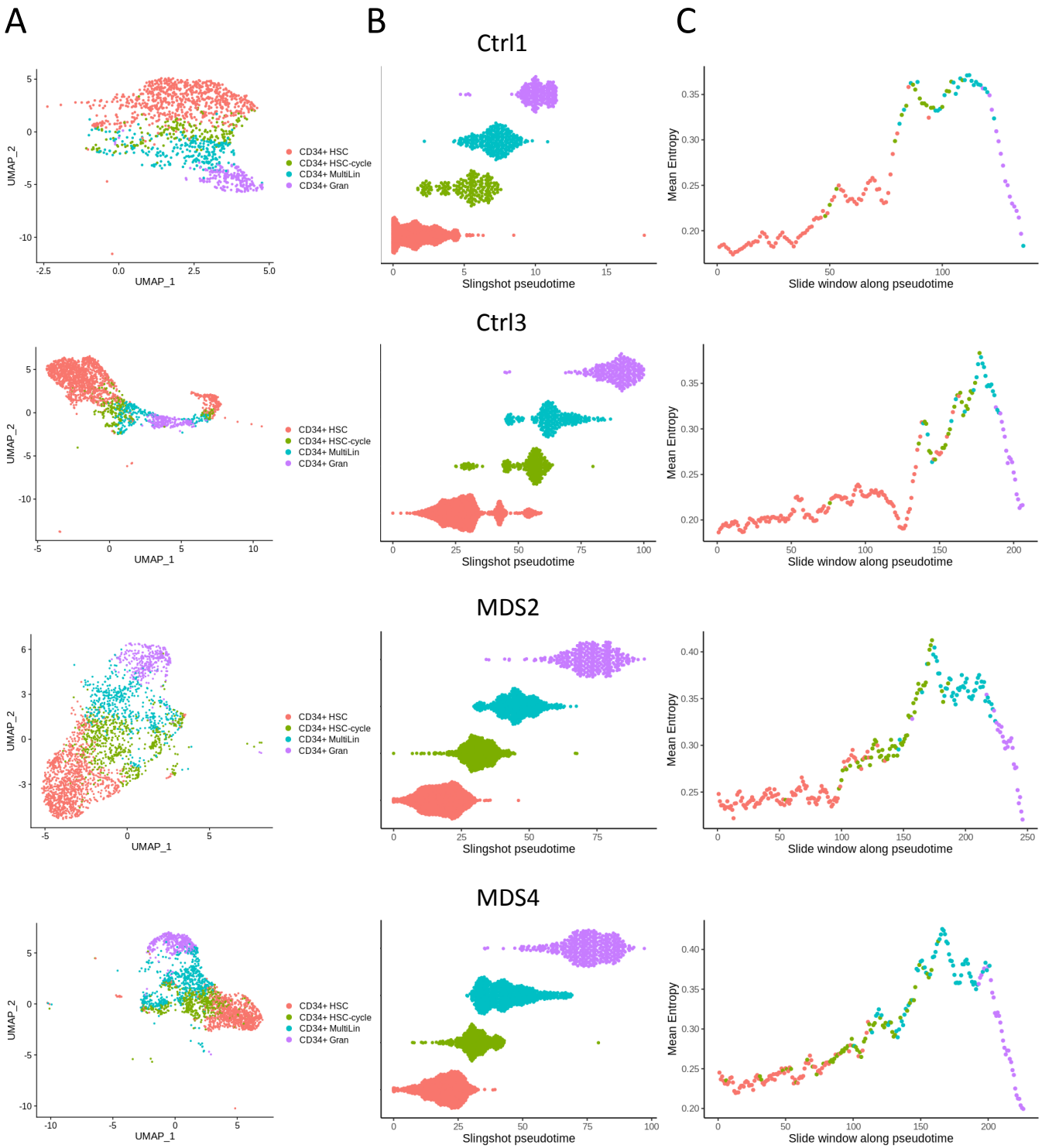


Figure 45 : Evolution de la moyenne d'entropie des populations HSPC au cours de la granulopoïèse chez les SMD et les sujets âgés

A) Les populations cellulaires spécifiques de la granulopoïèse sont sélectionnées et représentées en deux dimensions selon les coordonnées UMAP. **B)** Les cellules sont ordonnées selon le pseudotemps calculé par Slingshot. **C)** Chaque point correspond à la moyenne d'entropie de tous les gènes calculée sur une fenêtre de 50 cellules. La fenêtre avance le long du pseudotemps avec un pas de 10 cellules. La couleur de chaque point sur le graphique correspond à la nature de la première cellule de la fenêtre correspondante.

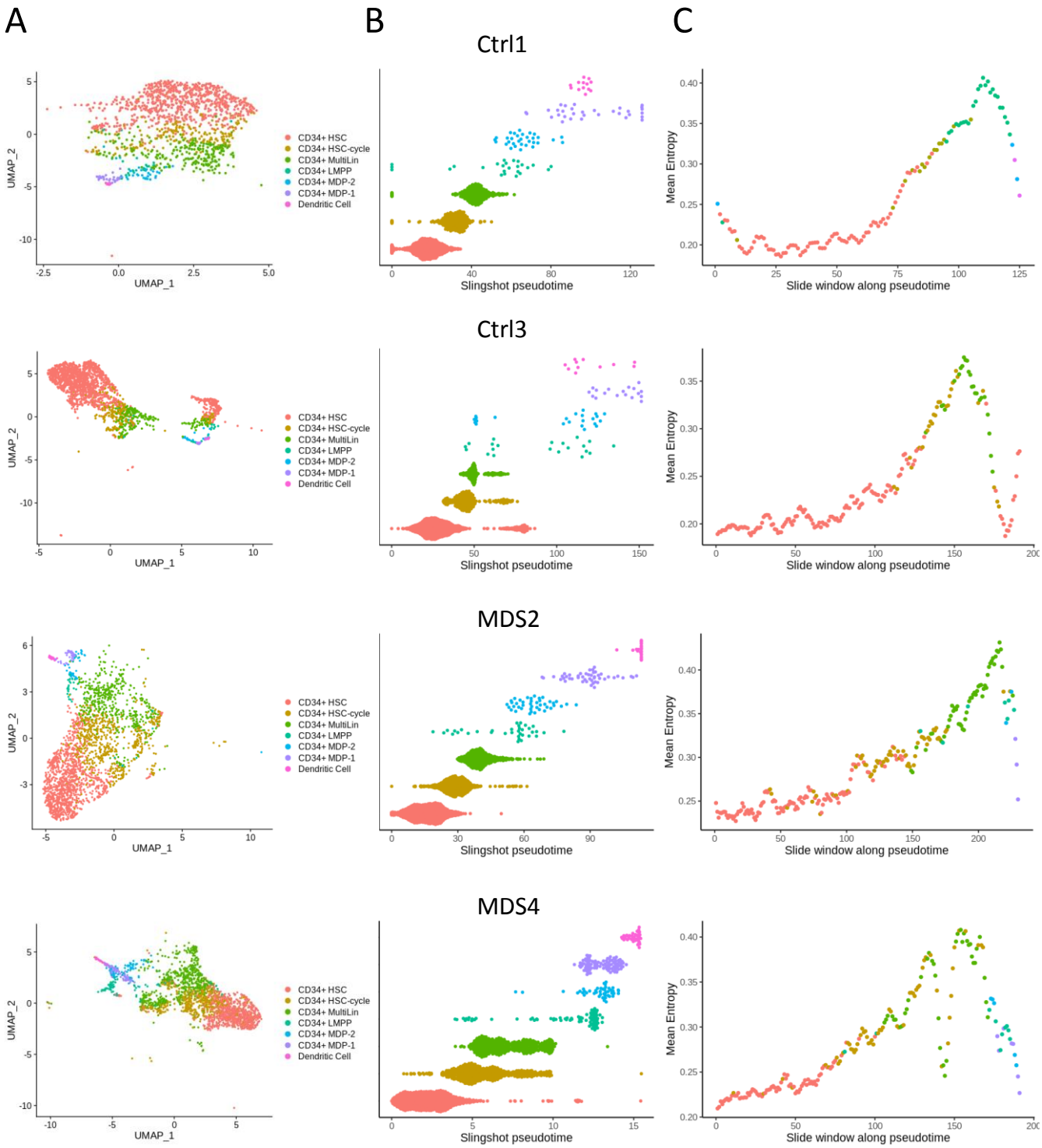


Figure 46 : Evolution de la moyenne d'entropie des populations HSPC au cours de la maturation dendritique chez les SMD et les sujets âgés

A) Les populations cellulaires spécifiques de la maturation dendritique sont sélectionnées et représentées en deux dimensions selon les coordonnées UMAP. **B)** Les cellules sont ordonnées selon le pseudotemps calculé par Slingshot. **C)** Chaque point correspond à la moyenne d'entropie de tous les gènes calculée sur une fenêtre de 50 cellules. La fenêtre avance le long du pseudotemps avec un pas de 10 cellules. La couleur de chaque point sur le graphique correspond à la nature de la première cellule de la fenêtre correspondante.

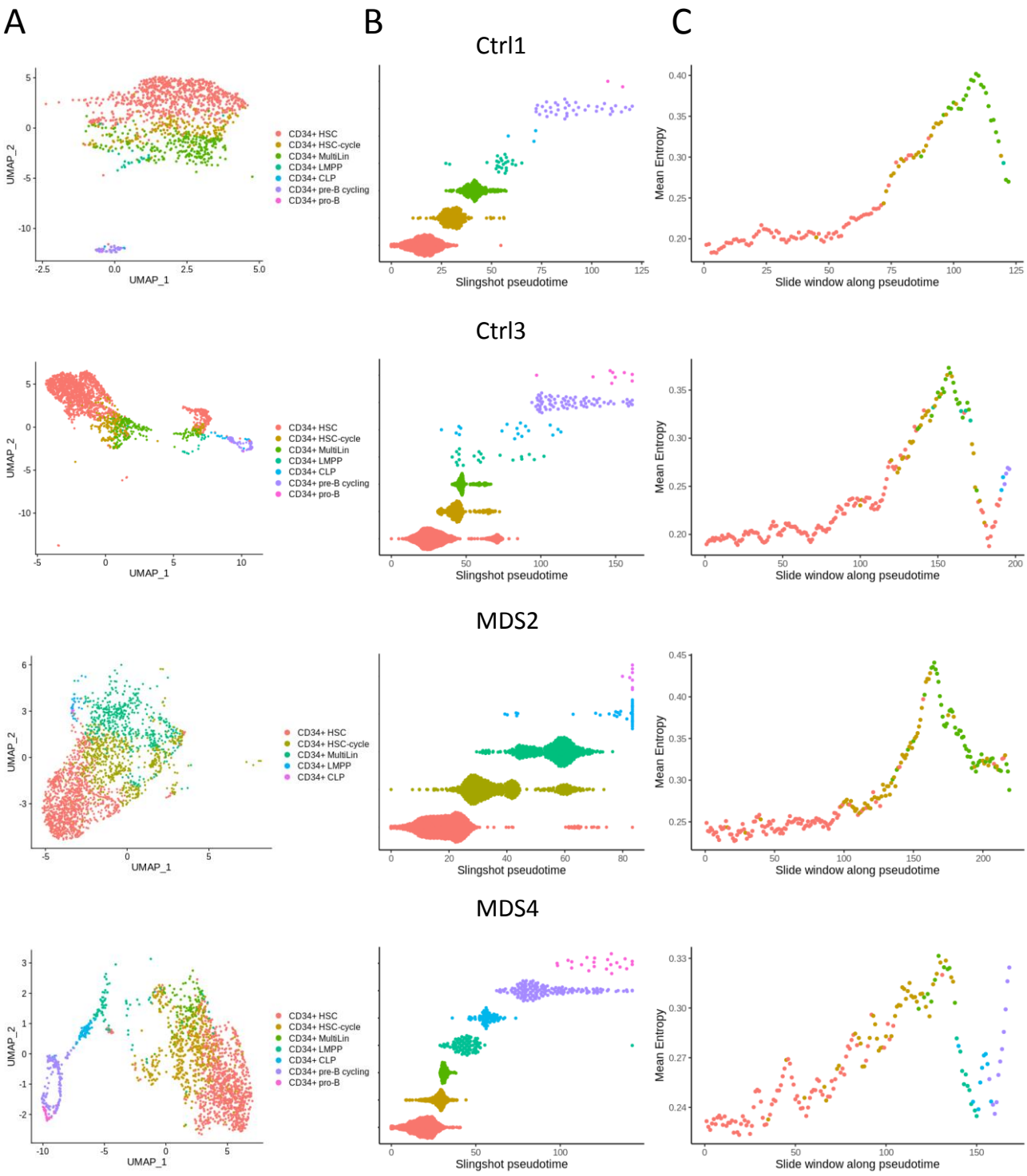


Figure 47 : Évolution de la moyenne d'entropie des populations HSPC au cours de la lymphopoïèse B chez les SMD et les sujets âgés.

A) Les populations cellulaires spécifiques de la maturation dendritique sont sélectionnées et représentées en deux dimensions selon les coordonnées UMAP. **B)** Les cellules sont ordonnées selon le pseudotemps calculé par Slingshot. **C)** Chaque point correspond à la moyenne d'entropie de tous les gènes calculée sur une fenêtre de 50 cellules. La fenêtre avance le long du pseudotemps avec un pas de 10 cellules. La couleur de chaque point sur le graphique correspond à la nature de la première cellule de la fenêtre correspondante.

Afin de représenter les variations d'entropie de manière comparable entre les sujets âgés et les SMD, nous avons utilisé la matrice gènes cellules intégrée des 4 échantillons afin de calculer un pseudotemps commun. Pour chaque voie de différenciation, nous avons réalisé un sous échantillonnage afin d'avoir un nombre de cellules identique pour chaque patient. La moyenne d'entropie est alors calculée pour chaque patient sur une fenêtre glissante de 50 cellules avançant avec un pas de 10 sur le pseudotemps commun **(Figure 48).** Les profils d'évolution de l'entropie au cours des 4 voies de différenciation sont comparables. On observe un pic d'entropie chez les sujets sains et les SMD pour toutes les voies de différenciation.

Ces données montrent qu'il existe chez les sujets âgés et les SMD mutés SF3B1 une augmentation de la variabilité de l'expression des gènes au cours de la différenciation hématopoïétique comparable à ce qui est observé dans la moelle normale.

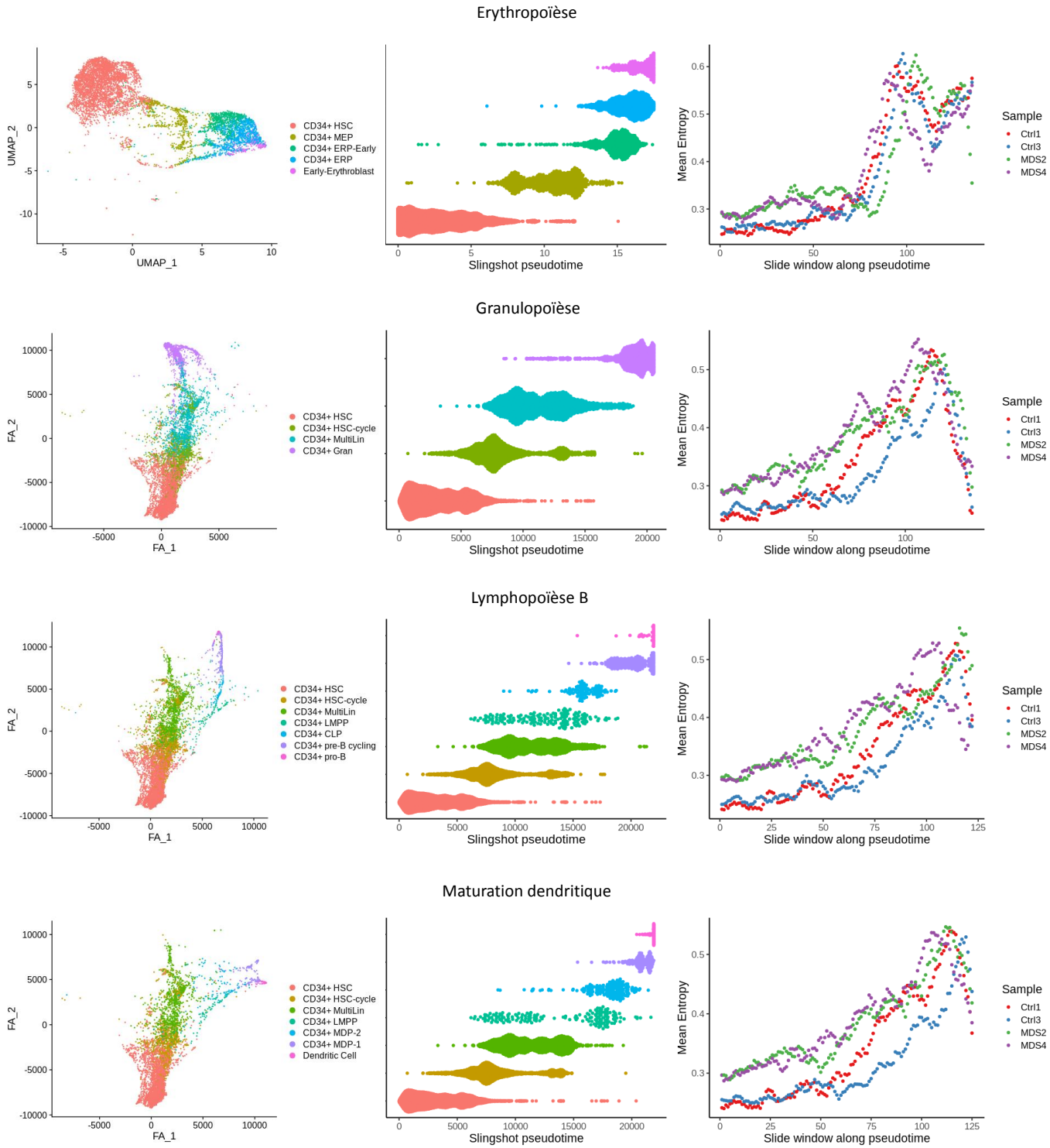


Figure 48 : Comparaison chez les sujets âgés et les SMD de l'évolution de l'entropie au cours des 4 principales voies de différenciation hématopoïétiques.

Pour chaque voie de différenciation, un pseudotemps commun est calculé sur la matrice gènes cellules intégrée. Un sous échantillonnage est effectué pour avoir un nombre de cellules identique par échantillon. La moyenne d'entropie est alors calculée individuellement pour chaque patient sur une fenêtre glissante de 50 cellules qui avance avec un pas de 10 cellules sur le pseudotemps commun.

3.5. La variabilité de l'expression génique est augmentée dans les CSH de SMD de bas risque.

En observant l'évolution de l'entropie comparée des sujets âgés et des SMD, on remarque au début de chaque voie de différenciation que l'entropie est plus élevée chez les SMD que chez les sujets âgés sains **(Figure 48).** Pour confirmer cette observation, nous avons calculé l'entropie des gènes ainsi que la moyenne spécifiquement sur la population des cellules souches hématopoïétiques (CD34+ HSC). Etant donné que la valeur d'entropie qui reflète la variabilité de l'expression génique dépend du nombre de cellules, nous avons effectué un sous échantillonnage afin de calculer l'entropie sur le même nombre de CSH par échantillon **(Figure 49).**

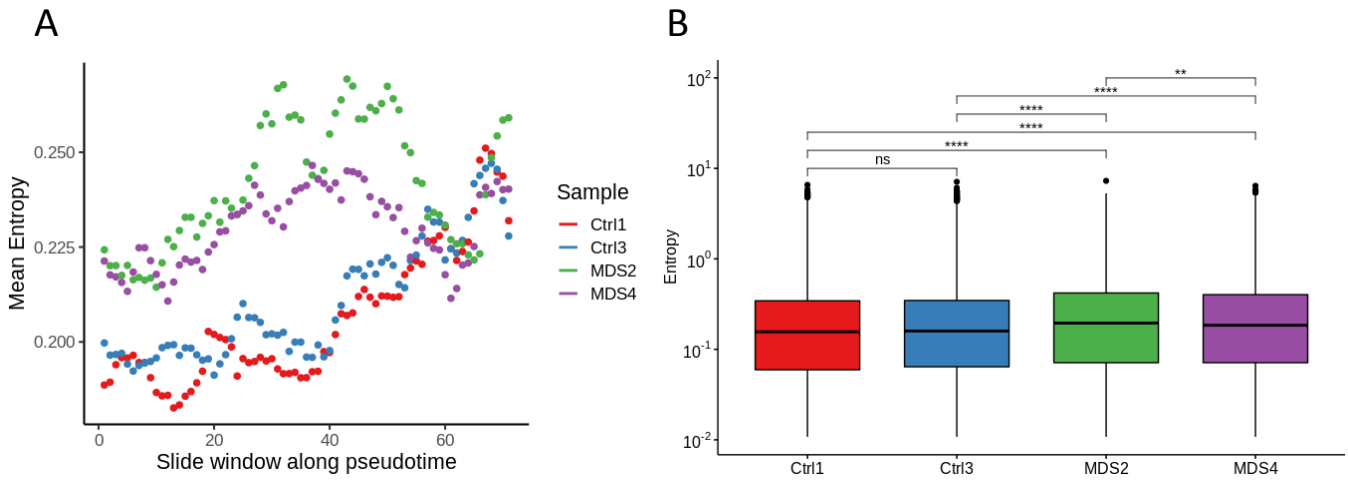


Figure 49 : Comparaison de l'entropie des CSH entre les sujets âgés et les SMD.
A) Moyenne d'entropie calculée individuellement dans la population des CSH pour chaque patient sur une fenêtre glissante de 50 cellules qui avance avec un pas de 10 cellules sur le pseudotemps érythroïde commun. **B)** Boxplot de l'entropie de tous les gènes calculée sur les CSH de chaque échantillon. Un test de wilcoxon a été utilisé pour comparer les moyennes d'entropie entre les échantillons. (* : p<0.05 ; ** : p<0,01 ; *** : p<0,001 ; **** : p<0.0001)

On ne retrouve pas de différence significative entre la moyenne d'entropie des CSH des deux sujets sains âgés. Par contre l'entropie des CSH des patients SMD est significativement supérieure à l'entropie des sujets sains âgés. De plus, l'entropie du patient MDS2 est significativement supérieur à l'entropie du patient MDS4. Ces données suggèrent que la variabilité de l'expression génique est augmentée dans les CSH de SMD. Cette augmentation pourrait être en lien avec la physiopathologie de la maladie et mérite d'être confirmée sur un plus grand nombre d'échantillons.

3.6. Un pic de variabilité de l'expression génique est observé au cours de l'hématopoïèse chez les SMD de haut risque et après traitement par azacytidine.

Après avoir observé l'évolution de la variabilité de l'expression génique (mesurée par l'entropie) chez les sujets âgés et les SMD de bas risque, nous avons voulu observer son évolution chez des patients SMD de haut risque traités par azacytidine. Pour cela, nous avons extrait les cellules souches et progénitrices CD34+ (HSPC) de deux patients avant et après traitement au long cours par azacytidine **(Tableau 3).** Nous avons réalisé cette expérience pour faire la preuve du concept de l'impact d'un traitement par agent hypométhylant sur les variations d'entropie identifiées précédemment. L'évaluation de la réponse au traitement (correction des cytopénies et diminution de la blastose médullaire) après respectivement 12 et 13 cures a permis d'objectiver une réponse chez le patient R et une absence de réponse chez le patient NR **(Figure 50).**

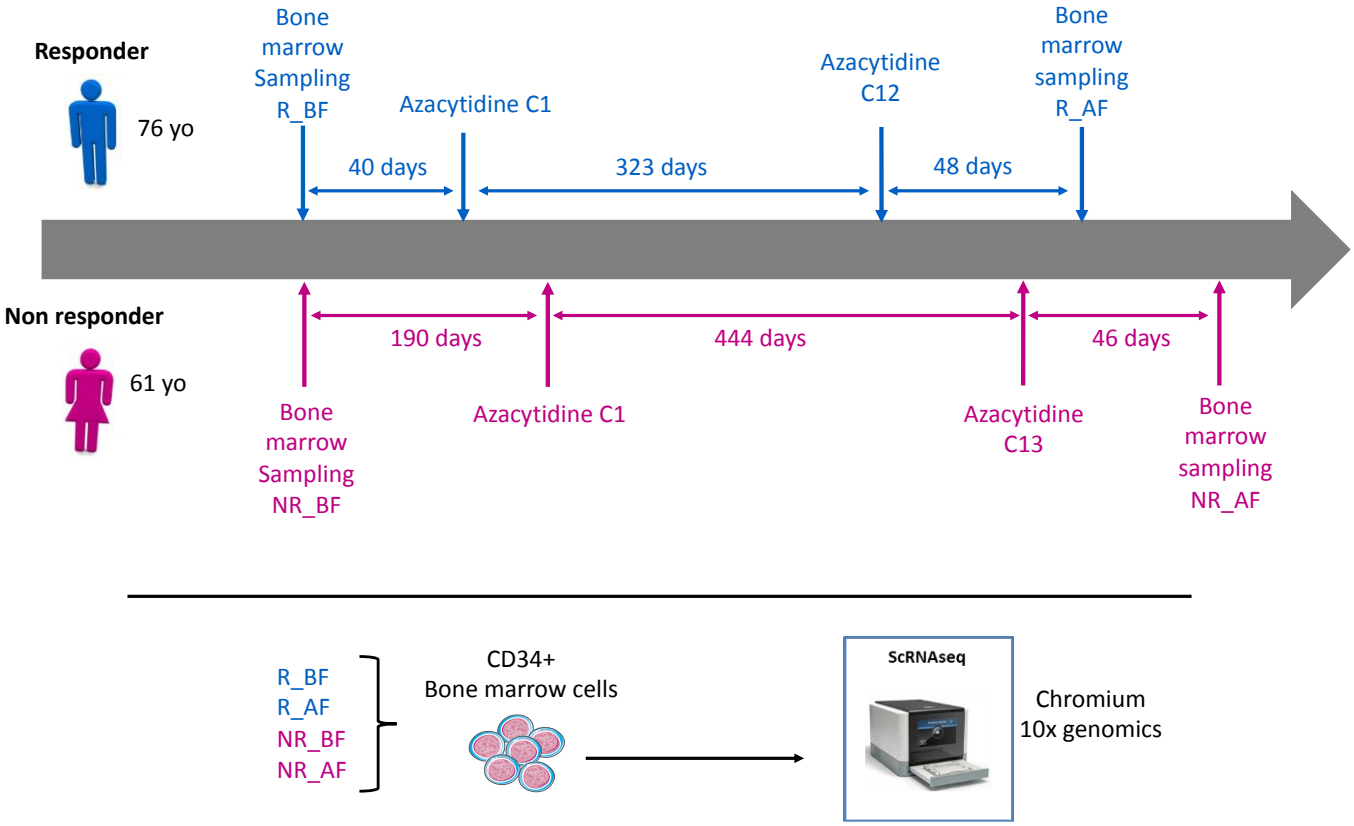


Figure 50 : Description de l'approche expérimentale pour l'évaluation de l'effet d'un traitement par azacytidine sur le transcriptome des HSPC.

Les cellules souches et progénitrices CD34+ des patients SMD de haut risque (l'un répondeur et l'autre non répondeur au traitement par azacytidine) sont isolées de la moelle osseuse par tri immunomagnétique. L'analyse par scRNA-Seq est effectuée selon la technologie chromium 10x genomics.

Nous avons ensuite appliqué à ces échantillons la technologie chromium (10x genomics), ce qui nous a permis de disposer de données transcriptomiques à l'échelle unicellulaire d'un total de 13381 HSPC réparties sur les 4 échantillons (3685 provenant de R_BF, 2684 provenant de R_AF, 2803 provenant de NR_BF, et 4209 provenant de NR_AF) **(Figure 50).** Pour générer une carte de référence de chaque patient, nous avons combiné avec la méthode d'intégration implémenté dans Seurat[194], d'un côté les HSPC avant et après traitement du patient répondeur (R_BF et R_AF) et de l'autre, celles du patient non répondeur (NR_BF et NR_AF)Figure 51. Après annotation des cellules par SingleR à partir des profils d'expression génique des populations hématopoïétiques décrites par Hay et al[93], nous avons retiré les types cellulaires dont le nombre total était inférieur à 10 cellules par échantillon. Pour le patient répondeur et non répondeur, les UMAP qui en résulte distingue respectivement 20 et 21 sous

types cellulaires différents qui sont organisés selon les voies de différenciations hématopoïétiques principales (érythropoïèse, granulopoïèse, et maturation dendritique) **(Figure 51).** Les progéniteurs lymphoides en aval des LMPP et CLP sont quasi absents dans les 4 échantillons étudiés et ne pourront pas être analysés. Les marqueurs spécifiques des sous populations distinguées par SingleR ne sont pas complètement en accord avec les études précédemment réalisées sur le compartiment HSPC médullaire[89,93,195,198], avec par exemple l'expression de *MAP3K8* qui caractérise les LMPP du patient non répondeur et *FAM30A* qui caractérise les CSH en cycle du patient répondeur **(Figure 52)**. Cela suggère que chez les SMD de haut risque, le transcriptome du compartiment HSPC est fortement altéré. Nous avons comparé pour les deux patients, l'annotation par SingleR des clusters déterminés par Seurat avec l'annotation cellule par cellule. Les deux méthodes sont cohérentes, mais l'annotation cellule par cellule étant plus précise nous avons choisi de l'utiliser pour les analyses réalisées en aval (**Figure 53 et Figure 54).** Ces données suggèrent que la hiérarchie hématopoïétique du compartiment HSPC de nos deux patients SMD traités par azacytidine est globalement conservée par rapport à l'hématopoïèse normale. En effet, on observe dans nos échantillons les populations souches et progénitrices retrouvées dans l'hématopoïèse normale même si leur transcriptome semble en partie altéré.

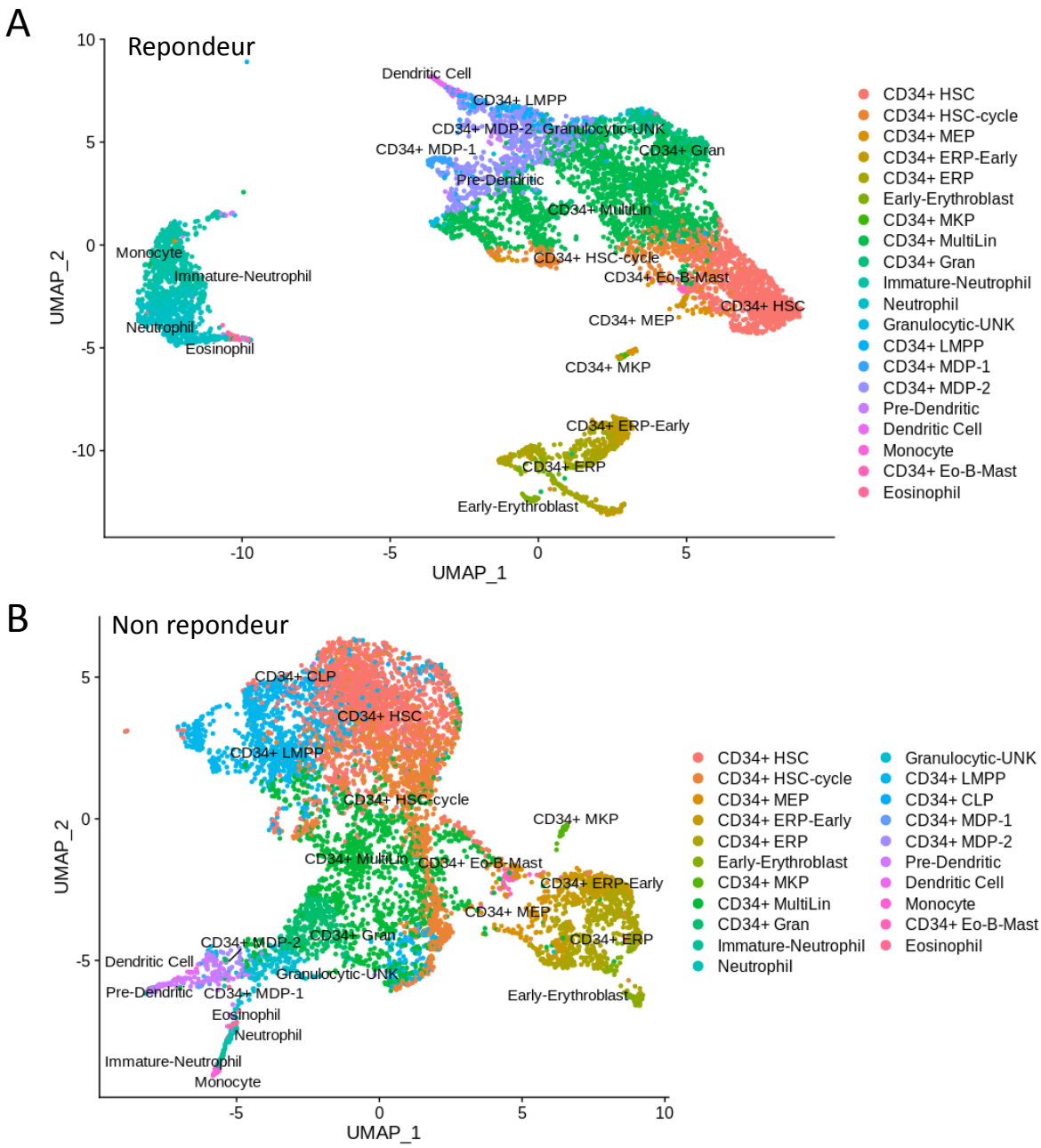
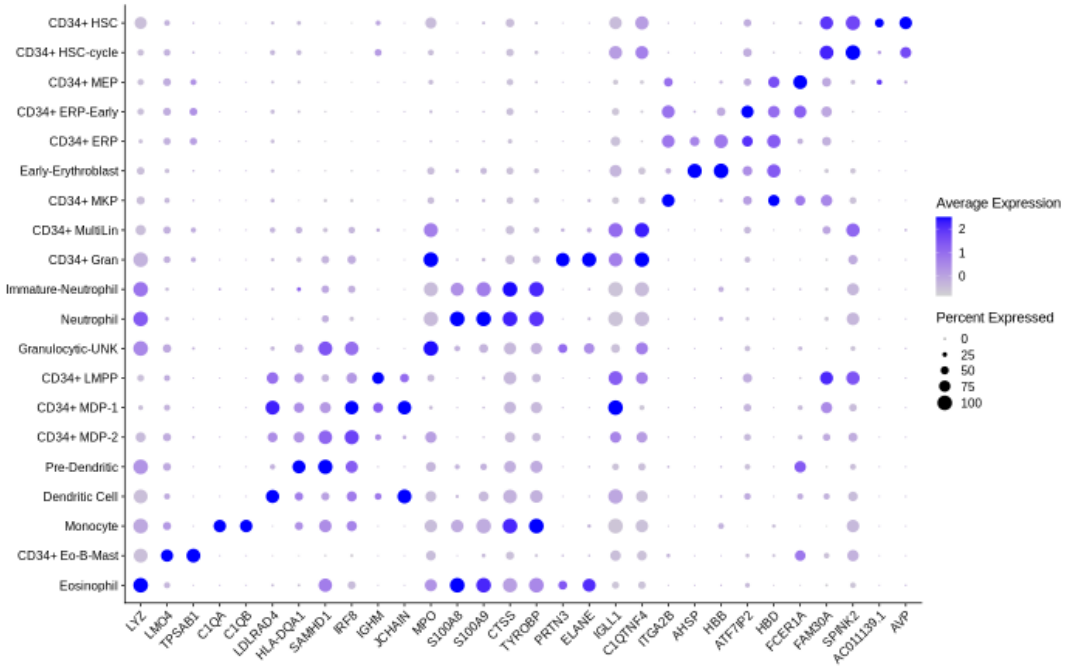


Figure 51 : Paysage trancriptionnel du compartiment HSPC d'un SMD répondeur (A) et d'un SMD non répondeur (B) avant et après traitement par azacytidine.

A) Représentation UMAP en deux dimensions de la matrice gènes-cellules (6369 cellules, 20951 gènes) intégrée des échantillons du patient SMD répondeur avant et après traitement par l'azacytidine. Les cellules annotées par SingleR sont classées en 20 sous types différents représenté chacun par une couleur différente. **B)** idem pour le patient SMD non répondeur (7012 cellules, 20951 gènes).

Répondeur



Non répondeur

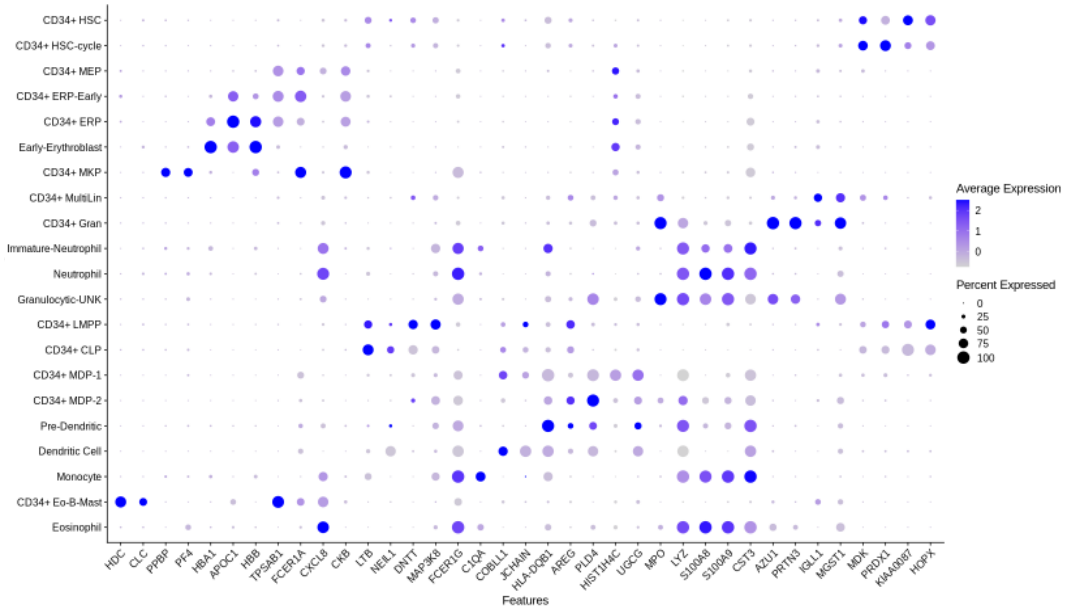


Figure 52 : Expression transcriptionnelle des marqueurs les plus spécifiques des sous populations HSPC annotées par SingleR des patients SMD répondeur et non répondeur avant et après traitement par azacytidine.

Les gènes les plus différentiellement exprimés de chaque sous population annotées par SingleR en comparaison avec toutes les autres ont été déterminés par la fonction FindMarkers de Seurat. Les deux gènes les plus spécifiques de chaque sous population sont représentés sur ce graphique. La couleur du cercle indique les valeurs d'expression génique et la taille du cercle représente la proportion de cellules exprimant le marqueur dans la sous population.

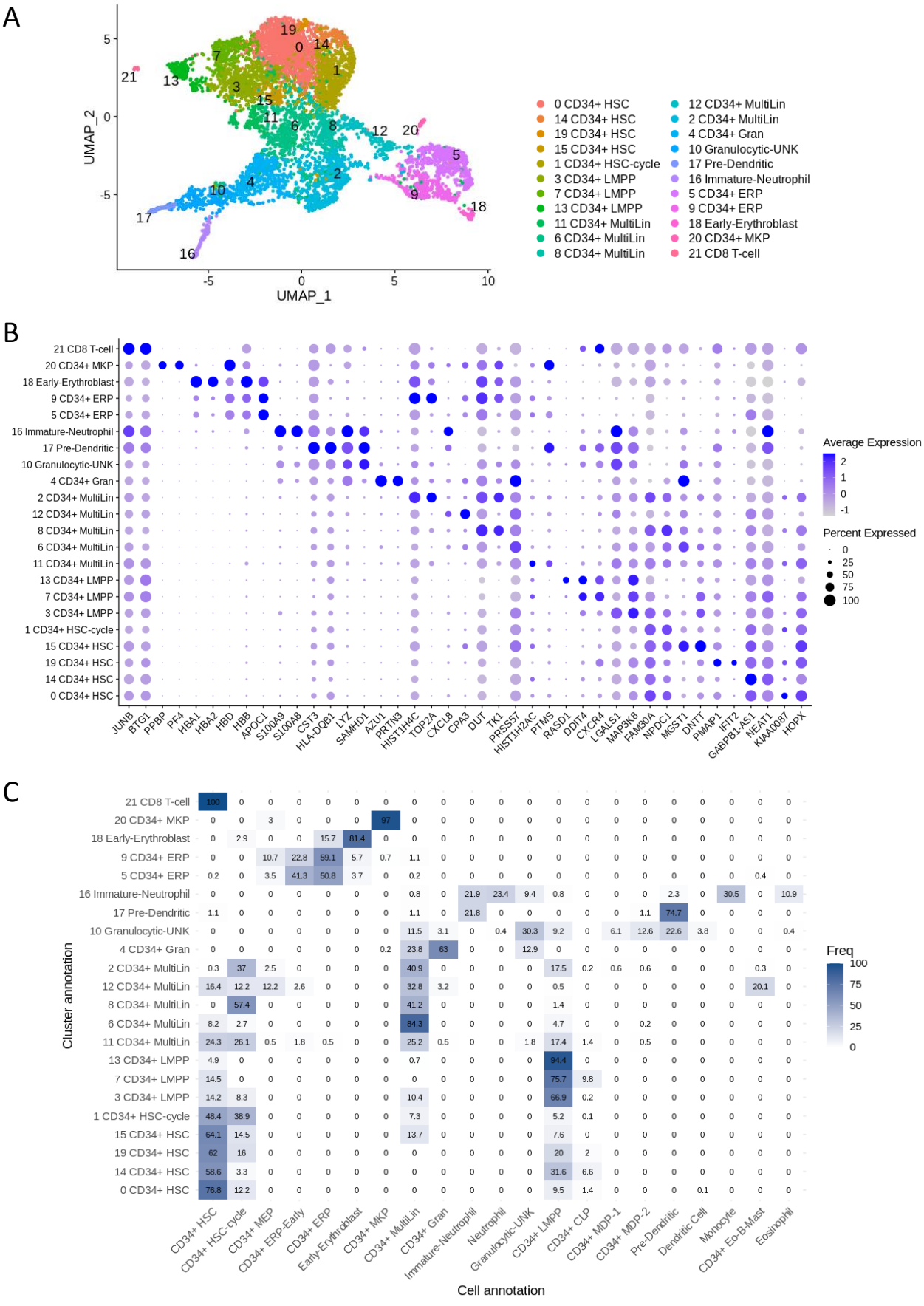

A
UMAP_2
UMAP_1
0 CD34+ HSC
14 CD34+ HSC
19 CD34+ HSC
15 CD34+ HSC
1 CD34+ HSC-cycle
3 CD34+ LMPP
7 CD34+ LMPP
13 CD34+ LMPP
11 CD34+ MultiLin
6 CD34+ MultiLin
8 CD34+ MultiLin
12 CD34+ MultiLin
2 CD34+ MultiLin
4 CD34+ Gran
10 Granulocytic-UNK
17 Pre-Dendritic
16 Immature-Neutrophil
5 CD34+ ERP
9 CD34+ ERP
18 Early-Erythroblast
20 CD34+ MKP
21 CD8 T-cell
B
Average Expression
Percent Expressed
C
Cluster annotation
Cell annotation
Freq

Figure 53 : Annotation par clusters du compartiment HSPC du patient non répondeur à l'azacytidine avant et après traitement.

A) Représentation par UMAP du compartiment HSPC intégré des échantillons du patient SMD non répondeur à l'azacytidine avant et après traitement. Les 22 clusters distingués par Seurat et annotées par SingleR sont représentés chacun par une couleur différente**. B)** Les gènes les plus différentiellement exprimés de chaque cluster annoté par SingleR en comparaison avec toutes les autres ont été déterminés par la fonction FindMarkers de Seurat. Les deux gènes les plus spécifiques de chaque sous population sont représentés sur ce graphique. La couleur du cercle indique les valeurs d'expression génique et la taille du cercle représente la proportion de cellules exprimant le marqueur dans la sous population. **C)** Comparaison de l'annotation par cluster avec l'annotation par cellule, chaque case représente le pourcentage de cellule du cluster correspondant à l'annotation cellule par cellule.

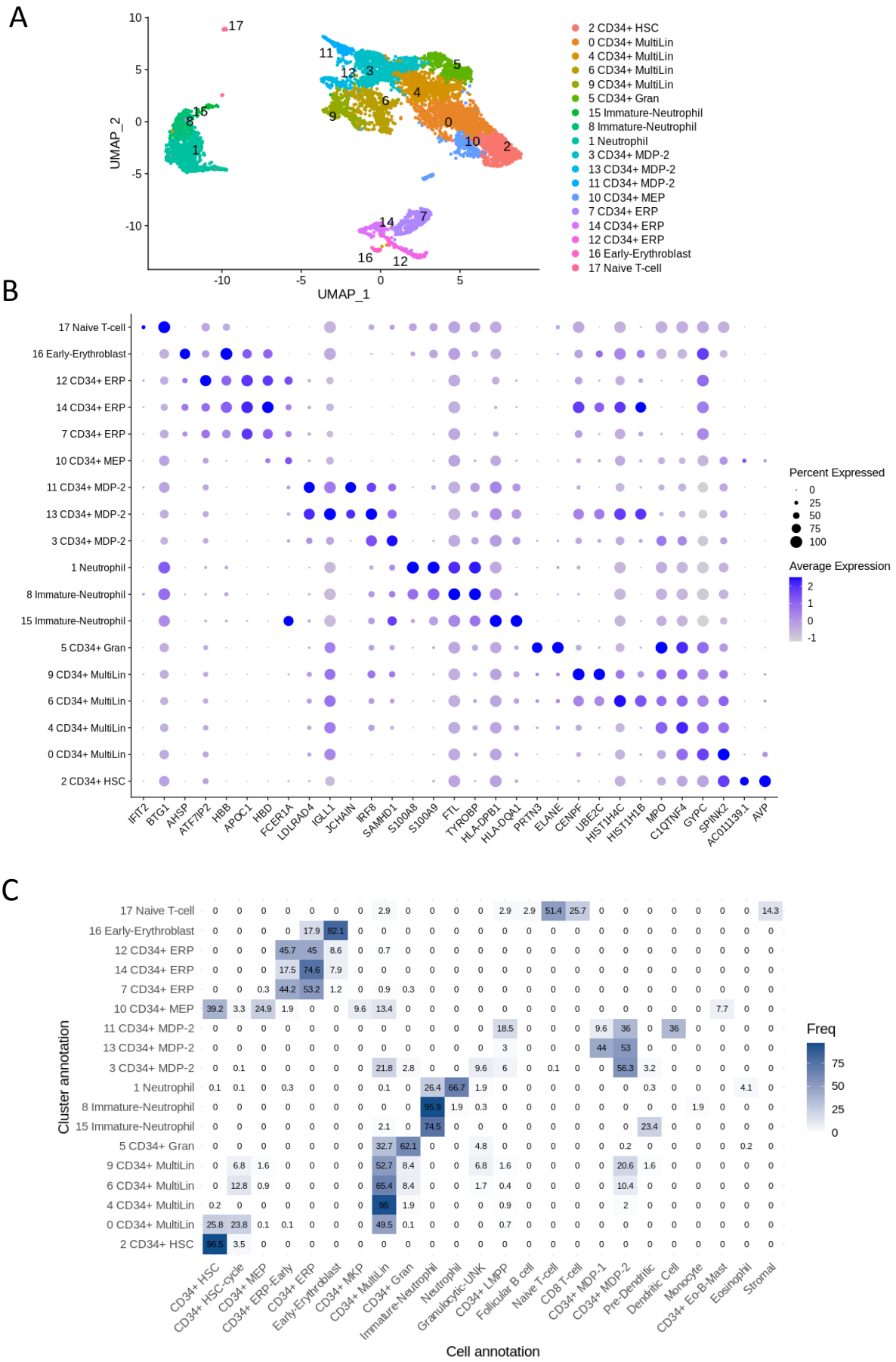

A
UMAP_2
UMAP_1
2 CD34+ HSC
0 CD34+ MultiLin
4 CD34+ MultiLin
6 CD34+ MultiLin
9 CD34+ MultiLin
5 CD34+ Gran
15 Immature-Neutrophil
8 Immature-Neutrophil
1 Neutrophil
3 CD34+ MDP-2
13 CD34+ MDP-2
11 CD34+ MDP-2
10 CD34+ MEP
7 CD34+ ERP
14 CD34+ ERP
12 CD34+ ERP
16 Early-Erythroblast
17 Naive T-cell
B
Percent Expressed
Average Expression
C
Cluster annotation
Cell annotation
Freq

Figure 54 : Annotation par clusters du compartiment HSPC du patient répondeur à l'azacytidine avant et après traitement.

A) Représentation par UMAP du compartiment HSPC intégré des échantillons du patient SMD répondeur à l'azacytidine avant et après traitement. Les 18 clusters distingués par Seurat et annotées par SingleR sont représentés chacun par une couleur différente**. B)** Les gènes les plus différentiellement exprimés de chaque cluster annoté par SingleR en comparaison avec toutes les autres ont été déterminés par la fonction FindMarkers de Seurat. Les deux gènes les plus spécifiques de chaque sous population sont représentés sur ce graphique. La couleur du cercle indique les valeurs d'expression génique et la taille du cercle représente la proportion de cellules exprimant le marqueur dans la sous population. **C)** Comparaison de l'annotation par cluster avec l'annotation par cellule, chaque case représente le pourcentage de cellule du cluster correspondant à l'annotation cellule par cellule.

Par la suite nous avons exploré chez ces patients les variations d'entropie au cours de l'hématopoïèse afin d'observer l'effet de l'azacytidine chez le patient répondeur et chez le patient non répondeur. Pour cela, nous avons dans un premier temps analysé les 4 échantillons individuellement **(Figure 55).** Puis nous avons pour chaque patient et pour chaque voie de différenciation (érythropoïèse, granulopoïèse, maturation dendritique) calculé l'entropie moyenne avec la même méthode décrite précédemment **(Figure 56, Figure 57 et Figure 58).** Concernant l'érythropoïèse, pour les 4 échantillons, on observe une augmentation de l'entropie des CSH (CD34+ HSC) jusqu'aux progéniteurs érythroïdes (CD34+ ERP). L'entropie redescend seulement pour l'échantillon NR_AF qui contient un nombre suffisant d'érythroblastes immatures (Early-erythroblast). Pour la granulopoïèse, l'entropie augmente régulièrement à partir des CSH pour atteindre un pic au niveau des progéniteurs granuleux (CD34+ Gran). L'entropie redescend ensuite dans les échantillons qui contiennent des cellules plus matures tel que les neutrophiles immatures (Immature-neutrophil). Au cours de la maturation des cellules dendritiques, l'entropie augmente des CSH jusqu'aux progéniteurs multilignés (CD34+ Multilin), puis redescend à partir des progéniteurs mono-dendritiques (CD34+ MDP-2). Ces données préliminaires suggèrent que dans les SMD de haut risque, il existe également un pic de variabilité de l'expression génique au cours de 3 des voies de différenciation de l'hématopoïèse et que ce pic est également présent après traitement par azacytidine que l'on soit répondeur ou non répondeur.

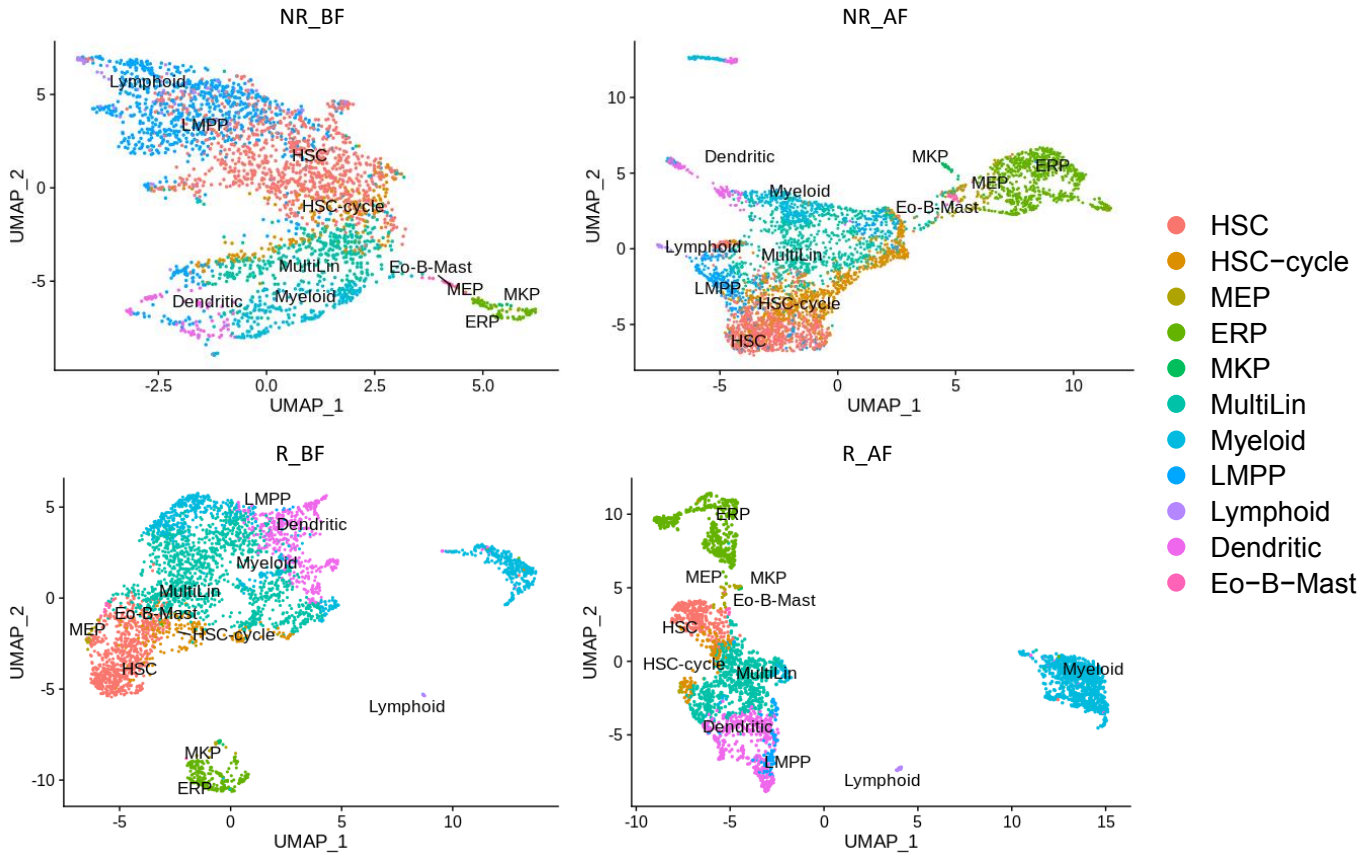


Figure 55 : Représentation individuelle par UMAP des HSPC du patient non répondeur et du patient répondeur avant et après traitement par azacytidine.

Les cellules ont été annotées individuellement par SingleR, puis regroupés par type cellulaire pour une visualisation simplifiée. Chaque sous type cellulaire simplifié est représenté par une couleur différente**.**

NR_BF : patient non répondeur avant traitement. NR_AF : patient non répondeur après traitement. R-BF : patient répondeur avant traitement. R_AF : patient répondeur après traitement.

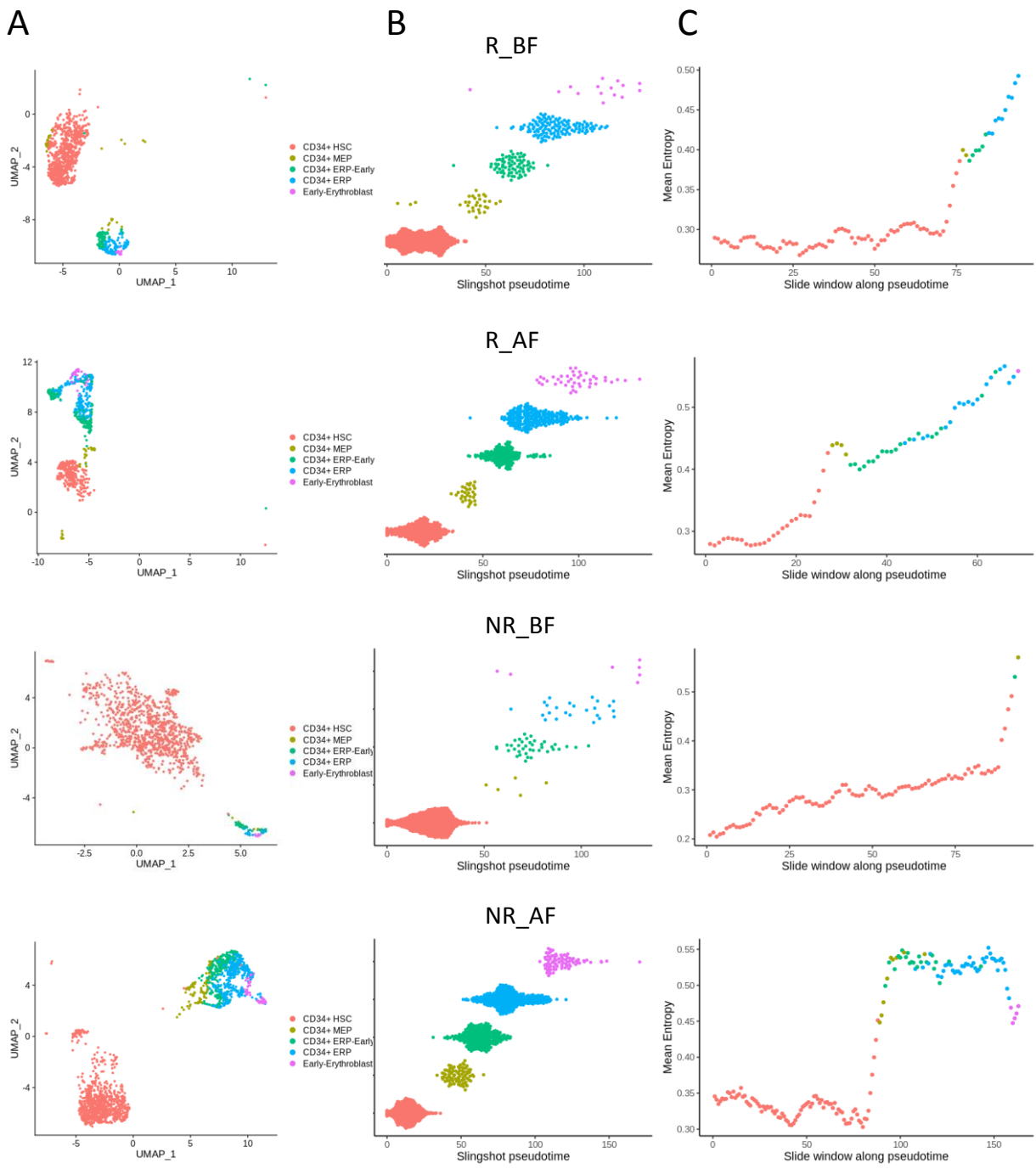


Figure 56 : Évolution de la moyenne d'entropie des populations HSPC au cours de l'érythropoïèse chez un patient non répondeur et un patient répondeur avant et après traitement par azacytidine.

A) Les populations cellulaires spécifiques de l'érythropoïèse sont sélectionnées et représentées en deux dimensions selon les coordonnées UMAP. **B)** Les cellules sont ordonnées selon le pseudotemps calculé par Slingshot. **C)** Chaque point correspond à la moyenne d'entropie de tous les gènes calculée sur une fenêtre de 50 cellules. La fenêtre avance le long du pseudotemps avec un pas de 10 cellules. La couleur de chaque point sur le graphique correspond à la nature de la première cellule de la fenêtre correspondante. NR_BF : patient non répondeur avant traitement. NR_AF : patient non répondeur après traitement.R-BF : patient répondeur avant traitement.R_AF : patient répondeur après traitement.

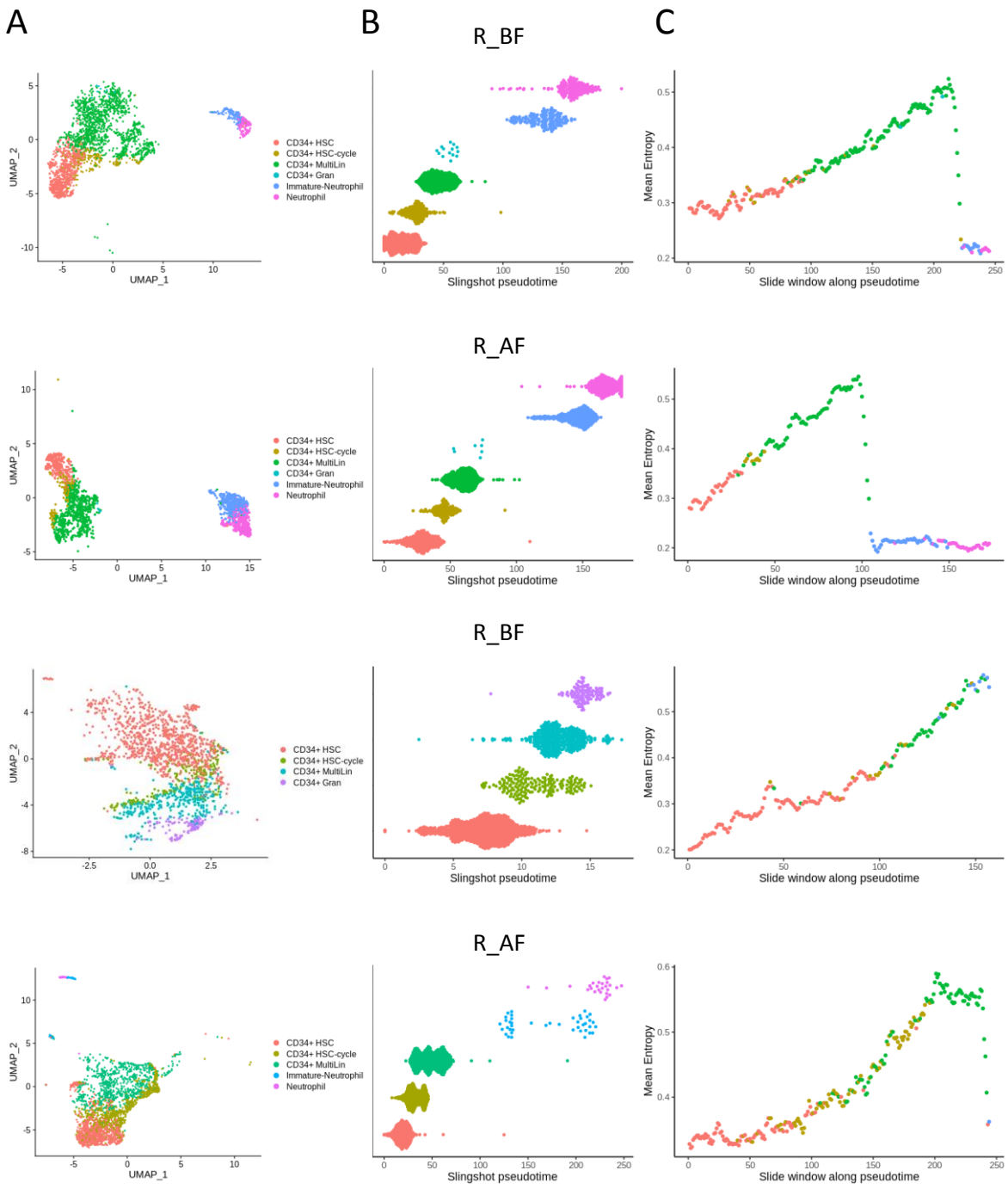


Figure 57 : Évolution de la moyenne d'entropie des populations HSPC au cours de la granulopoïèse chez un patient non répondeurs et un patient répondeur avant et après traitement par azacytidine.

A) Les populations cellulaires spécifiques de l'érythropoïèse sont sélectionnées et représentées en deux dimensions selon les coordonnées UMAP. **B)** Les cellules sont ordonnées selon le pseudotemps calculé par Slingshot. **C)** Chaque point correspond à la moyenne d'entropie de tous les gènes calculée sur une fenêtre de 50 cellules. La fenêtre avance le long du pseudotemps avec un pas de 10 cellules. La couleur de chaque point sur le graphique correspond à la nature de la première cellule de la fenêtre correspondante. NR_BF : patient non répondeur avant traitement. NR_AF : patient non répondeur après traitement.R-BF : patient répondeur avant traitement.R_AF : patient répondeur après traitement

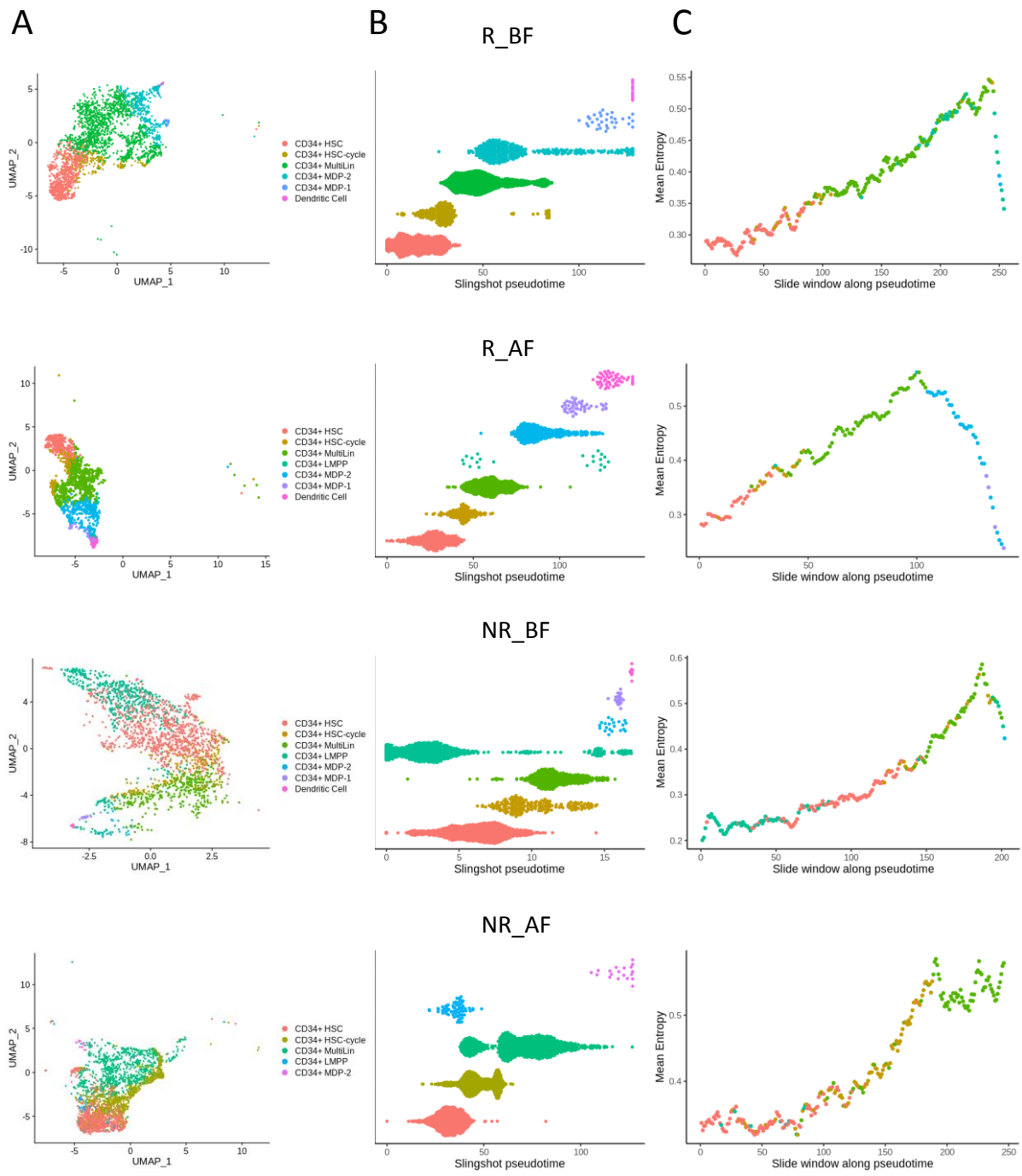


Figure 58 : Évolution de la moyenne d'entropie des populations HSPC au cours de la différenciation dendritique chez un patient non répondeur et un patient répondeur avant et après traitement par azacytidine.

A) Les populations cellulaires spécifiques de l'érythropoïèse sont sélectionnées et représentées en deux dimensions selon les coordonnées UMAP. **B)** Les cellules sont ordonnées selon le pseudotemps calculé par Slingshot. **C)** Chaque point correspond à la moyenne d'entropie de tous les gènes calculée sur une fenêtre de 50 cellules. La fenêtre avance le long du pseudotemps avec un pas de 10 cellules. La couleur de chaque point sur le graphique correspond à la nature de la première cellule de la fenêtre correspondante. NR_BF : patient non répondeur avant traitement. NR_AF : patient non répondeur après traitement.R-BF : patient répondeur avant traitement.R_AF : patient répondeur après traitement

Afin de représenter les variations d'entropie de manière comparable, nous avons utilisé pour le patient répondeur et le patient non répondeur, la matrice gènes cellules intégrée des échantillons avant et après traitement par azacytidine. Pour chaque patient, nous avons pu calculer un pseudotemps commun avant/après traitement. Comme réalisé précédemment et pour chaque voie de différenciation, nous avons réalisé un sous échantillonnage afin d'avoir un nombre de cellules identique pour chaque patient. La moyenne d'entropie est alors calculée pour chaque patient sur une fenêtre glissante de 50 cellules avançant avec un pas de 10 sur le pseudotemps commun **(Figure 59 et Figure 60).** Chez le patient répondeur, l'évolution de l'entropie est comparable dans les 3 voies de différenciation étudiées. Par contre, chez le patient non répondeur, il semble que l'entropie au début de chaque voie de différenciation est plus élevée après le traitement (courbe rouge).

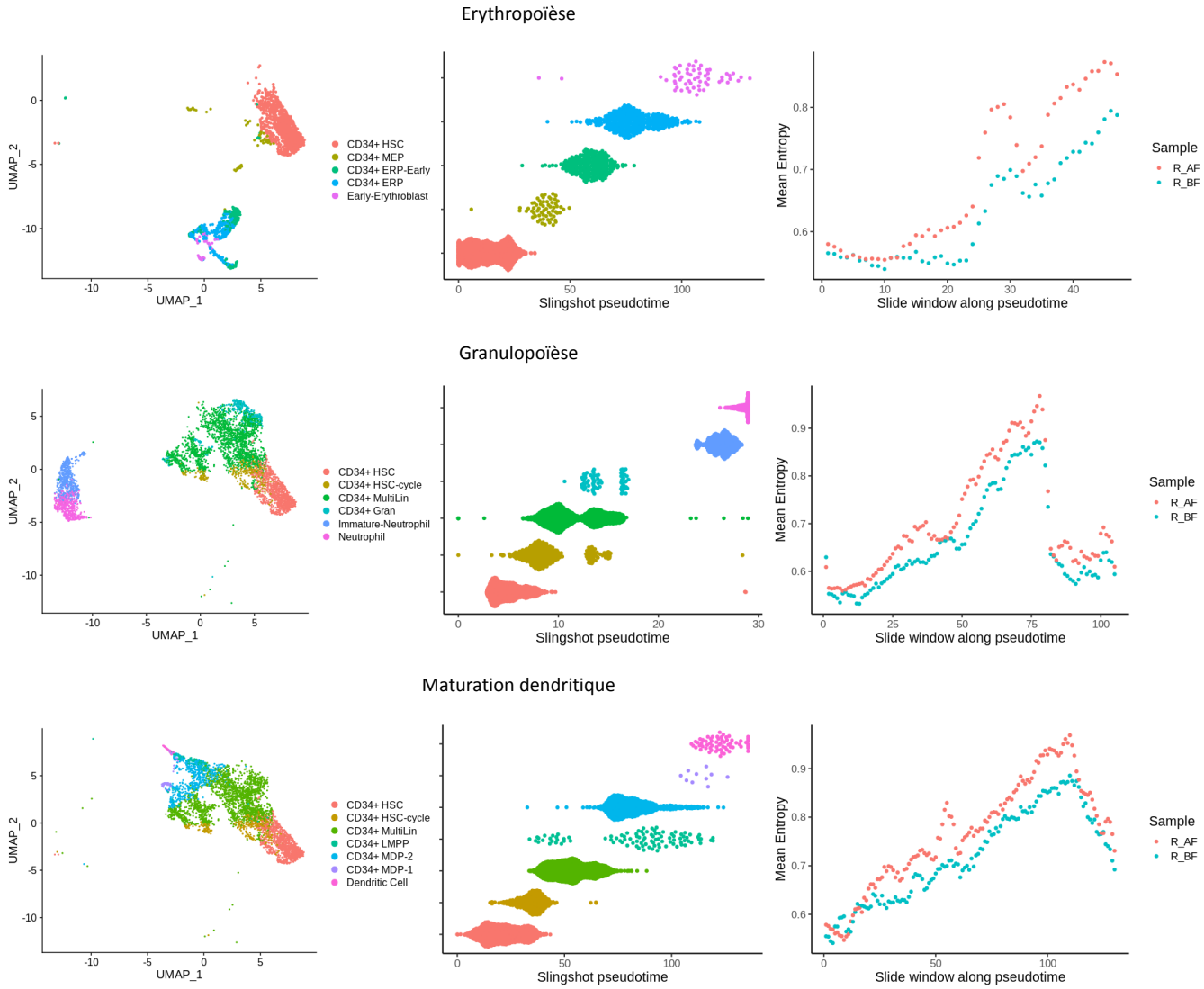


Figure 59 : Comparaison chez un patient SMD répondeur à l'azacytidine avant et après traitement de l'évolution de l'entropie au cours de 3 des principales voies de différenciation hématopoïétiques.

Pour chaque voie de différenciation, un pseudotemps commun est calculé sur la matrice gènes cellules intégrée. Un sous échantillonnage est effectué pour avoir un nombre de cellules identique par échantillon. La moyenne d'entropie est alors calculée individuellement pour chaque patient sur une fenêtre glissante de 50 cellules qui avance avec un pas de 10 cellules sur le pseudotemps commun.

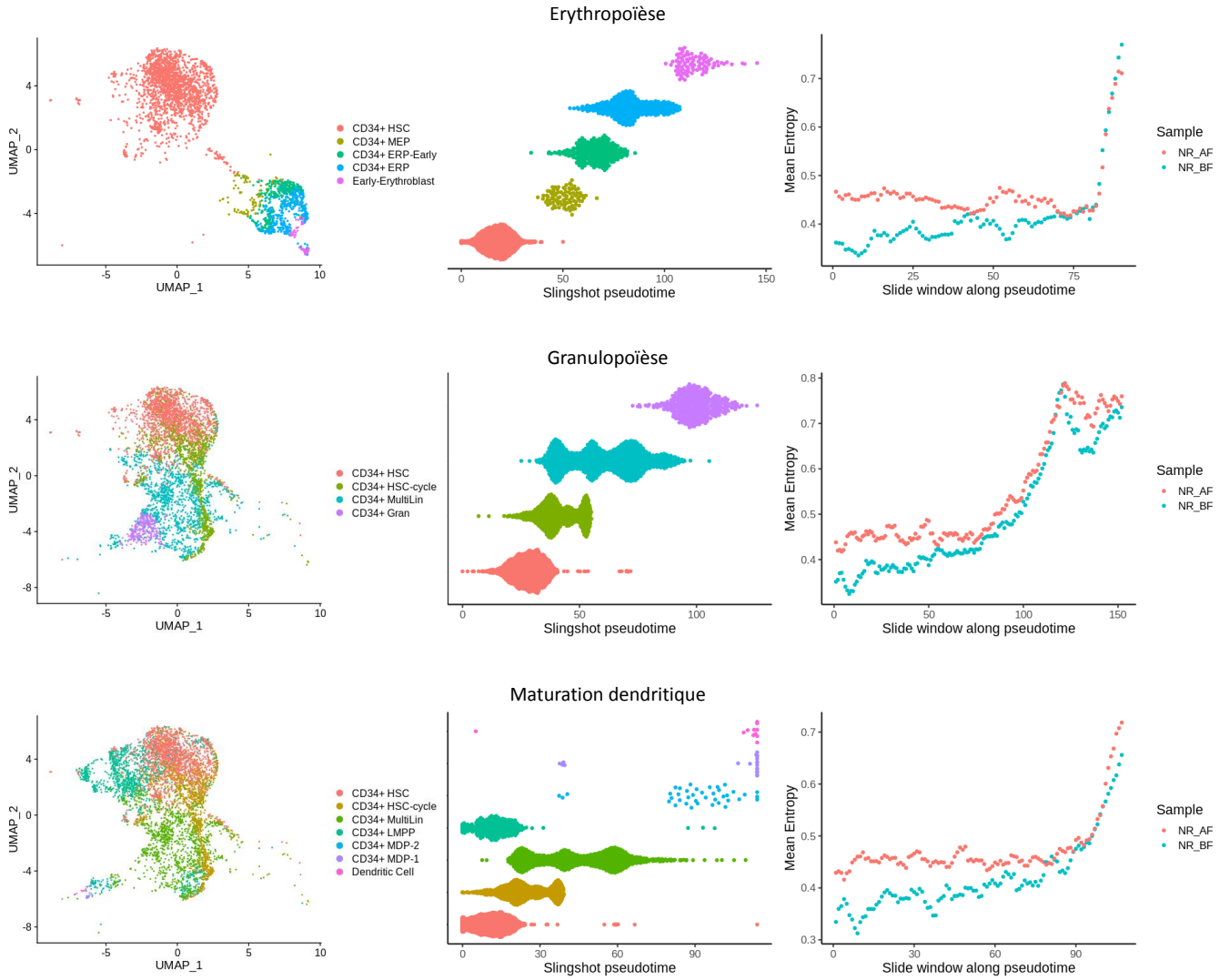


Figure 60 : Comparaison chez un patient SMD non répondeur à l'azacytidine avant et après traitement de l'évolution de l'entropie au cours de 3 des principales voies de différenciation hématopoïétiques.

Pour chaque voie de différenciation, un pseudotemps commun est calculé sur la matrice gènes cellules intégrée. Un sous échantillonnage est effectué pour avoir un nombre de cellules identique par échantillon. La moyenne d'entropie est alors calculée individuellement pour chaque patient sur une fenêtre glissante de 50 cellules qui avance avec un pas de 10 cellules sur le pseudotemps commun.

3.7. La variabilité de l'expression génique des CSH augmente au cours de l'évolution des SMD mais se stabilise chez les patients répondeurs aux agents déméthylants.

Pour vérifier si la différence d'entropie en début de différenciation est modifiée par le traitement, nous avons calculé l'entropie moyenne spécifiquement sur la population des CSH **(Figure 61).** Avant et après traitement chez le patient répondeur à l'azacytidine, nous n'observe pas de différence significative. En revanche chez le patient non répondeur, l'entropie moyenne des CSH est significativement augmenté après le traitement (p<0.0001).

Ainsi nos données suggèrent que lorsque la maladie n'est pas contrôlée par le traitement,, l'entropie des CSH augmente. Au contraire, lorsque la maladie est stabilisée par l'azacytidine, l'entropie des CSH semble rester stable.

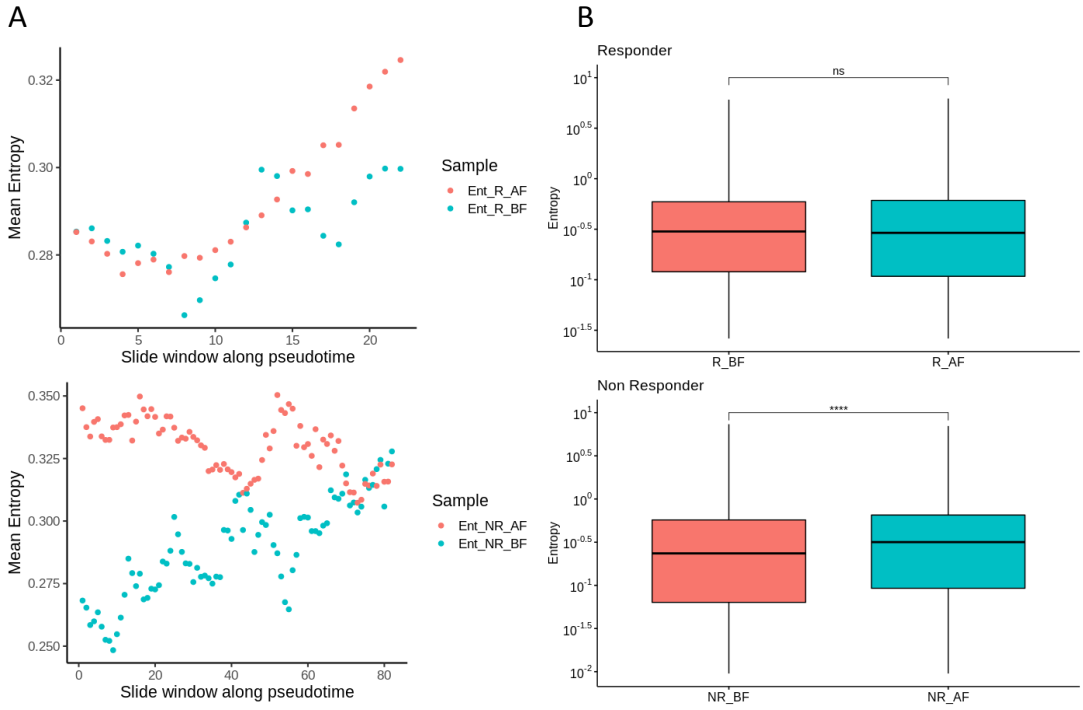


Figure 61 : Comparaison de l'entropie des CSH avant et après traitement par azacytidine chez un sujet répondeur et un sujet non répondeur.

A) Moyenne d'entropie calculée individuellement dans la population des CSH pour chaque échantillon sur une fenêtre glissante de 50 cellules qui avance avec un pas de 10 cellules sur le pseudotemps érythroïde commun.**B)** Boxplot de l'entropie de tous les gènes calculée sur les CSH de chaque échantillon. Un test de wilcoxon a été utilisé pour comparer les moyennes d'entropie entre les échantillons. (**** : p < 0.0001)

4. Conclusion

Notre étude montre que la variabilité de l'expression génique (estimée par l'entropie de Shannon) est un phénomène important dans la différenciation hématopoïétique normale. Au sein des voies de différenciation étudiées, nous avons mis en évidence un pic de variabilité de l'expression des gènes. Ce pic de variabilité apparait dans les cellules progénitrices, qui sont décrites pour représenter l'embranchement de différentes voies de différenciations dans la vision actuelle de l'hématopoïèse. La majeure partie des gènes qui sont responsables de ce pic de variabilité sont spécifiques des voies de différenciation étudiées, et ont pour la plupart d'entre eux été décrits dans la littérature comme jouant un rôle spécifique dans ces voies de différenciation. De manière intéressante, les gènes dont la variation d'expression au cours de la différenciation est la plus importante (à ne pas confondre avec la variabilité de l'expression mesurée par l'entropie) sont communs aux 4 voies de différenciations étudiées, sont impliqués dans la prolifération et le métabolisme cellulaire, et présentent une variation d'entropie particulièrement faible en comparaison avec tous les autres gènes étudiés. Ces données suggèrent qu'au cours de la différenciation cellulaire, la cellule met en place des processus tel que la synthèse protéique et la respiration cellulaire de manière équivalente dans toutes les cellules. Ces processus ne seraient alors pas stochastiques mais plutôt déterministe. A l'inverse, une grande partie des gènes dont l'entropie varie le plus au cours de la différenciation sont spécifiques de chaque voie de différenciation. Si le processus de différenciation en lui-même est plutôt déterministe, le destin final de la cellule serait plutôt un processus stochastique.

Nos observations montrent que le pic de variabilité de l'expression génique est également présent au cours de l'hématopoïèse des sujets âgés, des SMD de bas risque, des SMD de haut risque, ainsi que chez les sujets traités par azacytidine. Cette caractéristique est donc conservée même dans l'hématopoïèse pathologique des sujets myélodysplasiques et dans l'hématopoïèse modifiée par des traitements connus pour altérer le destin cellulaire.

Il semble également que la variabilité de l'expression génique des cellules souches de SMD soit plus élevée que chez les sujets sains d'âge comparable. De plus lors de l'évolution de la maladie, l'entropie des CSH semble augmenter alors qu'elle reste stable lorsque l'évolution de la maladie est stabilisée par l'azacytidine. Ainsi les cellules souches des patients

SMD présenteraient une caractéristique anormale supplémentaire qui pourrait jouer un nouveau rôle dans la physiopathologie de la maladie et possiblement permettre de réfléchir à de nouvelles approches thérapeutiques.

Discussion et perspectives

1. Immunophénotypage du compartiment CD34+ des SMD

Mon travail de thèse a permis de montrer que dans les SMD de haut risque, il existe une augmentation spécifique des populations GMP et LMPP identifiées par une approche de cytométrie en flux. Dans la moelle normale, la population LMPP est très faiblement représentée. En complément des stratégies actuellement utilisées pour le diagnostic des SMD, la présence au sein du compartiment CD34+ d'un contingent augmenté de LMPP pourrait servir de marqueur additionnel dans le cadre de la prise en charge des SMD de haut risque. Dans notre étude nous avons eu à disposition les échantillons d'un patient au stade SMD de bas risque puis au stade d'évolution de sa maladie en SMD de haut risque. Il est intéressant de remarquer que la population LMPP de ce patient était déjà augmentée dans l'échantillon au diagnostic lorsque le patient était classé en SMD de bas risque. Ainsi la présence d'une quantité augmentée de LMPP pourrait être un signe précoce de gravité, de risque d'aggravation de la maladie et d'évolution en LAM. Il a été démontré que dans 80% des LAM, les cellules souches leucémique (CSL) sont ordonnées hiérarchiquement. Les CSL dont l'immunophénotype correspond aux LMPP donnent naissance à des CSL plus matures correspondant aux GMP [44]. Ainsi, dans les SMD, il se pourrait que les premiers évènements oncogéniques dans la CSH, donne naissance dans un premier temps à des CSL leucémiques de phénotype LMPP, puis les LMPP donneraient naissance à une autre population de cellules souches leucémiques de type GMP. Une perspective intéressante serait d'étudier l'architecture clonale de ces populations LMPP et GMP pour déterminer comment se répartissent les mutations au sein de ses sous populations. Des techniques de génotypage à l'échelle unicellulaire réalisées au diagnostic, lors du suivi de la maladie jusqu'à l'évolution en LAM pourrait permettre de suivre l'évolution de sous clones[201]. Par ailleurs, étudier le comportement de ces sous populations dans des modèles murins de xénogreffes permettrait de voir si elles sont capables de reconstituer efficacement la maladie. Les tentatives de modéliser les SMD par xénotransplantation restent encore à ce jour décevantes. En effet, les cellules souches de SMD ont une capacité limitée de prise de greffe dans les modèles murins

actuels. Des études récentes ont permis leur amélioration : soit en utilisant des souris immunodéficientes modifiées génétiquement pour produire des cytokines humaines favorisant la prise de greffe[202], soit en créant une niche médullaire humanisé sous-cutané appelée « ossicle » formée par différenciation *in situ* de cellules stromales mésenchymateuses humaines. L'équipe de Françoise Pflumio au CEA a perfectionné ce dernier modèle et une collaboration va permettre prochainement d'étudier la prise de greffe des sous populations LMPP et GMP des SMD de haut risque dans ce nouveau modèle, ainsi que leurs relations avec les cellules souches stromales mésenchymateuses. Il sera intéressant de voir si une meilleure sélection des cellules du compartiment des HSPC de SMD pourra aider à une amélioration du taux de prise de greffe chez l'animal.

L'entropie de la répartition des cellules du compartiment HSPC est diminuée dans nos échantillons de SMD de haut risque. Cette diminution pourrait correspondre à l'expansion d'un clone pathologique au détriment de la diversité de l'hématopoïèse normale. Cette diminution d'entropie pourrait être utilisée pour prédire la maladie, certains patients SMD de bas risque dont l'entropie de répartition des cellules est faible pourrait avoir un risque augmenté de se transformer en SMD de haut risque voire en LAM.

L'immunophénotypage des populations souches et progénitrices de la moelle osseuse tel qu'il a été réalisé dans notre étude nécessite cependant une quantité importante de cellules et un tri immunomagnétique des cellules CD34+ positives (au moins 20 000 cellules CD34+). Ceci est un facteur limitant car nous n'avons pu réaliser l'immunophénotypage que sur un nombre restreint de prélèvements, en effet la plupart du temps le nombre de cellules issues de la ponction de moelle osseuse était trop faible (inférieur à 10 millions de CMN après Ficoll). La quantité de cellules issues de l'aspiration médullaire n'est pas seulement dépendante de la qualité technique du prélèvement. Ainsi les SMD dont la moelle est pauvre ou difficilement aspirable ne sont pas représentés dans notre étude ce qui entraine un biais de sélection dont nous avons conscience. C'est pourquoi j'ai participé au laboratoire d'hématologie biologique de l'hôpital Cochin à la mise en place d'un panel permettant de détecter les populations progénitrices LMPP, B/NK et CMPMEP3 dans les échantilllons de moelle totale sans tri préalable des cellules CD34+. Cette technique demande peu de cellules et peut être réalisé sur les prélèvements de la quasi-totalité des patients pour lesquels on suspecte SMD. Une

cohorte est actuellement en cours de constitution, ce qui devrait nous permettre de confirmer les observations faites sur les cellules CD34+ triées en colonne. A partir des résultats obtenus sur cette cohorte, nous allons également tenter de prédire l'évolution en leucémie ainsi que la réponse à l'azacytidine. Pour cela, nous allons appliquer des techniques d'apprentissage automatisé (Machine learning) aux données de cytométrie combinées avec les données de génomiques, l'hémogramme, le caryotype et l'entropie de la répartition des cellules.

2. Variabilité de l'expression génique, entropie et différenciation hématopoïétique.

L'analyse des données de scRNA-Seq de la moelle d'un sujet sain montre clairement pour la première fois qu'il existe un pic de variabilité de l'expression génique (que l'on mesure par l'entropie de Shannon) au cours des principales voies de différenciation de l'hématopoïèse.

Pour arriver à cette conclusion nous avons dû faire des choix et des assomptions que je vais tenter d'expliquer. Nous avons premièrement choisi de ne pas annoter manuellement les clusters de cellules déterminés par Seurat. Nous avons utilisé les profils d'expression génique des populations médullaires décrites par Hay et al[93] pour annoter les cellules une par une. Cette étude ayant analysé en scRNA-Seq 280000 cellules issues de la moelle osseuse de 8 donneurs sains, c'est à notre connaissance la base de données la plus complète existante à ce jour. Après avoir appliqué différentes stratégies, cette manière d'annoter les cellules nous a semblé être la plus précise possible. Après annotation des cellules, nous avons dû choisir les cellules spécifiques de chacune des voies de différenciation, pour cela nous avons pris en compte la répartition des cellules sur la UMAP **(Figure 33)** et nos connaissances de l'hématopoïèse. Les annotations de Hay et al ne sont pas forcément à prendre au pied de la lettre, en effet pour la différenciation lymphoïde B, les cellules correspondantes aux CD34+ pre-B cycling sont situées « avant » les cellules CD34+ pro-B, ce qui ne correspond pas à ce qui est connu dans la littérature. Nous aurions pu ensuite calculer l'entropie de chaque sous populations, mais si l'annotation cellule par cellule est ce qu'on peut faire de plus précis, la différenciation hématopoïétique correspond plus à un continuum de cellules qu'à des sous populations clairement distinctes. Ainsi, nous avons pensé qu'il serait plus élégant de calculer

cette entropie sur une fenêtre glissante de 50 cellules qui avance avec un pas de 10 cellules le long de la voie de différenciation étudiée. Pour cela il faut ordonner les cellules des plus immatures aux plus matures le long de la trajectoire de différenciation. Nous avons choisi après avoir testé plusieurs méthodes (Pseudotime, vélocité), de calculer le pseudotemps de chaque voie de différenciation avec le package R Slingshot[34] qui nous a donné les résultats les plus reproductibles et les plus cohérents vis-à-vis de nos connaissances de l'hématopoïèse. Une fois les cellules ordonnées, nous avons dû choisir une taille de fenêtre et un pas. En effet la variabilité de l'expression génique est mesurée par l'entropie de Shannon dans une population de cellules définie. Nous avons fait varier la taille des fenêtres de 10 à 100 cellules, et avons choisi 50 cellules qui était la valeur nous donnant les résultats les plus reproductibles et qui permettait un temps de calcul raisonnable. Le choix du pas avec lequel se déplace la fenêtre s'est arrêté, après différents tests, à une valeur de 10 cellules afin d'allier précision et temps de calcul accessible à nos ressources actuelles.

La courbe d'entropie le long des différentes voies de différenciation forme une cloche **(Figure 35)**. Au départ, les valeurs d'entropie sont faibles ce qui implique une faible variabilité de l'expression génique au sein des CSH. On pourrait penser que pour pouvoir donner naissance à tous les types de cellules hématopoïétiques, la variabilité de l'expression génique et donc l'entropie devrait être élevée. Mais les cellules souches ont des propriétés intrinsèques spécifiques au niveau de leur métabolisme[203], et de leur capacité à réparer les dommages à l'ADN[204]. Nous pouvons émettre l'hypothèse que les gènes nécessaires à la mise en place de ces caractères spécifiques ont une expression relativement homogène dans la population des CSH, en accord avec une faible valeur de l'entropie en début de différenciation. Ces données suggèrent que les CSH sont dans un état relativement stable entropiquement parlant en comparaison avec les progéniteurs multipotents auxquelles elles donnent naissance. Si l'on met à part l'érythropoïèse, dans la différenciation granuleuse, dendritique et lymphoïde, les CSH vont donner naissance aux CSH en cycle (CD34+ HSC-cycle) puis aux progéniteurs multilignés (CD34+ MultiLin). Nous avons identifié un pic d'entropie au niveau de ces sous populations ce qui suggère une variabilité intercellulaire importante, ce qui va leur permettre de donner naissance aux lignées granuleuses, dendritiques et lymphoïdes. Ces sous populations sont dans un état instable et tendent à évoluer vers la stabilité. Ainsi dans les

populations en aval des progéniteurs myéloïdes et qui sont spécifiques des voies de différenciation, l'entropie diminue pour atteindre un minimum correspondant à un état stable. Il est intéressant de noter que l'entropie en fin de différenciation est plus basse que l'entropie en début de différenciation. Cela suggère que l'état cellule différenciée est plus stable que l'état cellule souche. En effet, on peut penser que les CSH en plus de leurs fonctions intrinsèques de cellule souche, doivent pouvoir donner naissance à plusieurs types cellulaires nécessitant une certaine variabilité dans l'expression génique, alors que les cellules les plus mature d'une voie de différenciation donnée ont toutes la même fonction et donc une variabilité ide l'expression génique très faible. Pour l'érythropoïèse, l'entropie augmente après les CSH pour atteindre son pic dans la population MEP (progéniteur érythro-mégacaryocytaire), ainsi cette population est instable car elle va pouvoir donner naissance aux lignées érythroïdes et mégacaryocytaires. Dans les données analysées, l'entropie redescend ensuite pour être minimale dans la population des érythroblastes matures dont la fonction est homogène impliquant une faible variabilité de l'expression génique. A noter que nous n'avons pas pu calculer l'entropie au cours de la mégacaryopoïèse du fait du nombre trop réduit de cellules de cette lignée dans les données étudiées. On peut penser que les sous populations cellulaires pour lesquelles l'entropie est maximale correspondent au moment critique de la différenciation ou le destin cellulaire va basculer vers l'une ou l'autre des voies de différenciation hématopoïétique. On peut résumer ainsi le processus de différenciation cellulaire par la figure suivante :

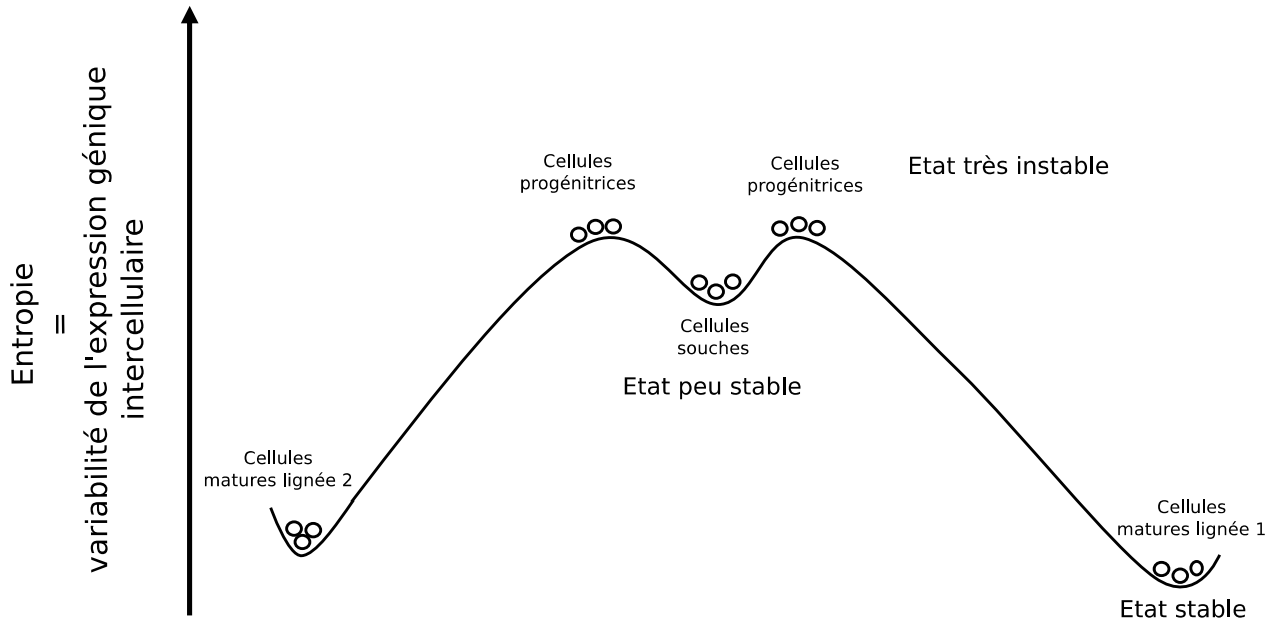


Figure 62 : Modèle théorique de la différenciation cellulaire.

Au stade cellule souche, les cellules sont dans un état de stabilité modéré et la variabilité de l'expression génique est basale. L'augmentation de la stochasticité de l'expression des gènes va permettre aux cellules souches d'atteindre le stade de cellule progénitrice, les cellules sont alors dans un état très instable où la variabilité de l'expression génique est maximale. C'est à ce moment-là que va se faire le choix du destin cellulaire. Ainsi, chaque cellule progénitrice va basculer vers un état cellulaire stable correspondant à une cellule mature où la variabilité de l'expression génique est minimale.

Si la position du pic d'entropie au cours de la granulopoïèse, de la maturation dendritique et de la lymphopoïèse B intervient à peu près toujours au niveau de la population des progéniteurs multilignés dans tous les jeux de données étudiés **(Figure 35, Figure 45, Figure 46, Figure 47, Figure 57, Figure 58), ce** n'est pas le cas pour l'érythropoïèse. En effet si dans la moelle du sujet sain jeune, le pic d'entropie se situe au niveau des MEP, dans nos données chez les sujets âgés et les sujets SMD, l'entropie continue de monter dans les progéniteurs érythroblastiques les plus matures et ne redescend que pour certains patients au niveau de la population des érythroblastes immatures (early-erythroblast) **(Figure 35,Figure 44,Figure 56).** On peut en partie expliquer ses différences par la manière dont on calcul l'entropie. Les valeurs d'entropie représentées sur les graphiques dépendent des types cellulaires présents dans la fenêtre de 50 cellules. Ainsi, l'allure de la courbe d'entropie va dépendre de la répartition des types cellulaires le long du pseudotemps. Si les dernières fenêtres ne contiennent pas une population homogène d'érythroblastes immatures, l'entropie ne redescend pas. Ce qui ne nous permet pas toutefois d'expliquer pourquoi le pic d'entropie ne

redescend pas après la population MEP. Quoiqu'il en soit, nos données indiquent pour la première fois, qu'il existe au cours de l'hématopoïèse humaine normale, un phénomène d'augmentation de la variabilité de l'expression des gènes qui est une caractéristique commune aux différentes voies de différenciation. Ceci suggère que l'augmentation de la stochasticité de l'expression des gènes pourrait être nécessaire au processus de différenciation hématopoïétique. Ces résultats supportent l'hypothèse que l'hématopoïèse serait un processus stochastique plutôt que déterministe.

Dans les données publiques de scRNA-Seq de moelle osseuse de sujet sain jeune, nous avons mis en évidence des gènes que l'on a désigné comme étant les gènes les plus « variablement entropiques ». Ce sont en fait les gènes dont la variabilité de l'expression varie le plus au cours de la différenciation. En pratique, un gène dont la variabilité de l'expression génique (mesurée par l'entropie) est faible dans la population la plus mature de la différenciation et très forte dans une population progénitrice sera un gène très variablement entropique (par exemple les gènes de l'hémoglobine ont une expression très variable dans les progéniteurs érythroïdes et très homogène dans les érythroblastes matures). C'est la différence entre le maximum et le minimum de variabilité de l'expression génique qui détermine les gènes les plus variablement entropiques. Ainsi nous avons mis en évidence que les gènes les plus variablement entropiques au cours des principales voies de différenciation hématopoïétiques leurs sont spécifiques **(Figure 37).** Nos données suggèrent que les gènes spécifiques de chaque voie de différenciation ont une variabilité d'expression maximale dans les populations progénitrices qui correspondent au pic d'entropie. Ainsi, au sein de la population cellulaire où semble se décider le destin cellulaire, la variabilité de l'expression des gènes spécifiques des différents choix possibles est maximale. Parmi les gènes les plus variablement entropiques, certains n'ont pas de rôle connu dans la différenciation hématopoïétique, il pourrait être intéressant d'étudier le rôle de ces gènes dans le contexte de la différenciation (par exemple en inactivant par CRISPR un de ces gènes dans un modèle de différenciation humaine in vitro). On pourrait ainsi utiliser les variations d'entropie pour détecter les gènes qui ont un rôle important dans la différenciation.

Nous avons également montré que les gènes les plus variablement entropiques ne sont pas les mêmes que les gènes les plus variablement exprimés au cours de la

différenciation. En effet, les gènes dont la variation d'expression est la plus importante sont communs aux 4 voies de différenciation. Ce sont des gènes impliqués dans la prolifération et le métabolisme cellulaire. Ces gènes par rapport à tous les autres gènes détectés ont une variation d'entropie au cours de la différenciation très faible. Ainsi ces gènes sont exprimés de manière homogène dans les sous populations cellulaires. Nous avons regardé à titre d'exemple l'évolution de l'expression de ces gènes sur le pseudotemps de la différenciation lymphoïde B **(Figure 63)**, ce sont des gènes très exprimés dans les cellules souches et les progéniteurs immatures puis leur expression décroit pour être quasi nulle dans les cellules les plus matures. Les cellules souches et progénitrices ont donc besoin d'exprimer ces gènes pour proliférer et se différencier, mais une fois la différenciation terminée, ces gènes ne sont pas utiles au fonctionnement des cellules les plus matures. Ces données suggèrent qu'il existe un processus commun à toutes les voies de différenciation cellulaire (prolifération, métabolisme cellulaire). Ainsi, dans l'hématopoïèse, la différenciation cellulaire pourrait être un processus déterministe, tandis que le destin cellulaire serait un processus stochastique.

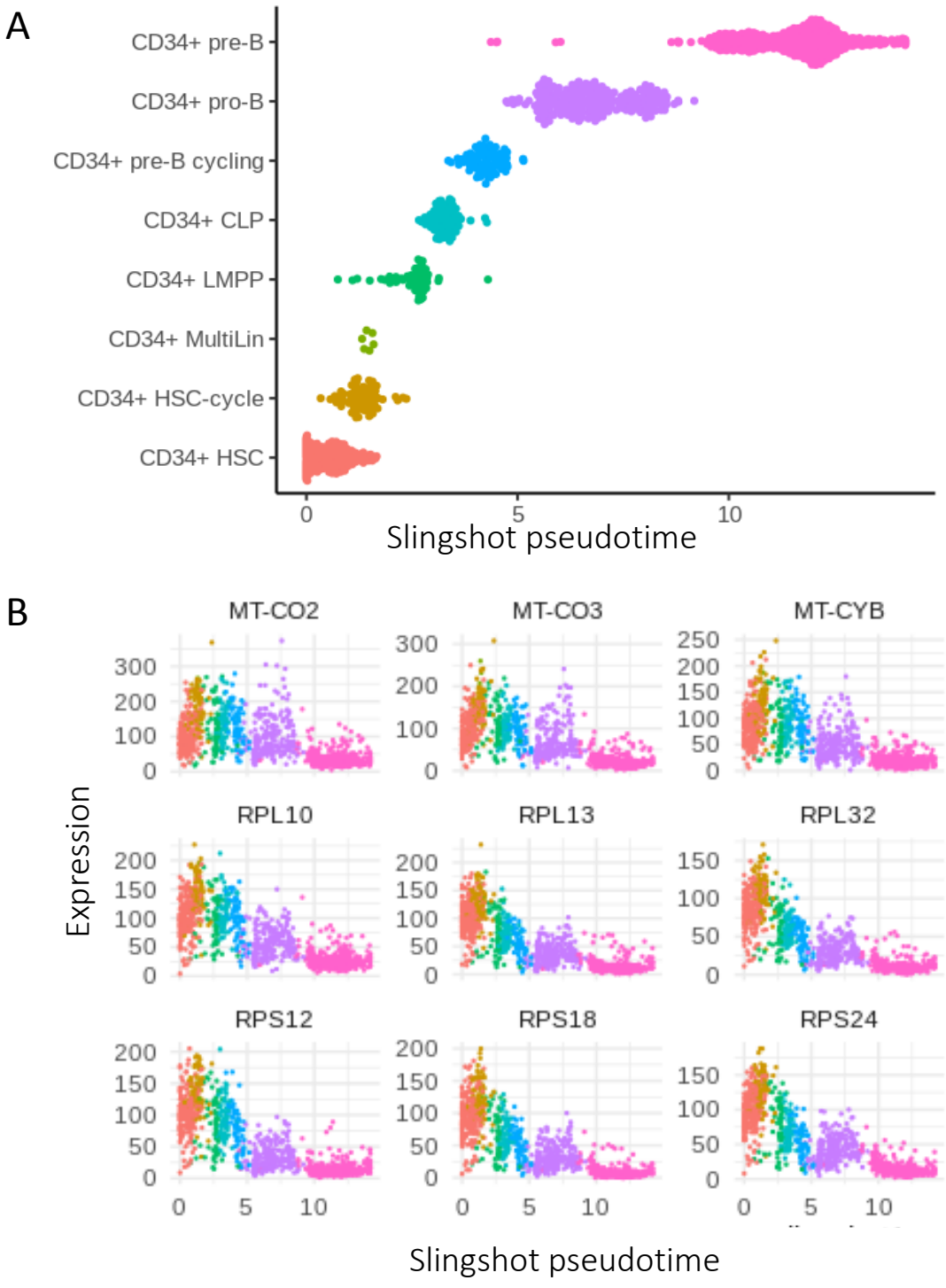


Figure 63 : Evolution de l'expression des gènes dont la variation d'expression est la plus importante au cours de la différenciation lymphoïde B médullaire chez un sujet sain.

A) Position des cellules des différentes sous population de la lymphopoïèse B sur le pseudotemps déterminé par Slingshot. **B)** L'expression génique est réprésentée en ordonnée et le pseudotemps en abscisse, chaque point correspond à une cellule, la couleur correspond à la sous population cellulaire

3. SMD : une maladie de l'entropie ?

Dans ce travail de thèse, j'ai mis en évidence que l'entropie (reflet de la variabilité de l'expression génique) des CSH est augmentée dans nos échantillons de SMD par rapport aux sujets sains âgés. Nous sommes confortés dans cette idée par le fait que l'entropie augmente dans les CSH d'un patient au cours de l'évolution de la maladie chez un patient non répondeur au traitement par azacytidine, alors que l'entropie est stable chez un patient répondeur. On sait qu'il n'existe à l'heure actuelle pas de traitement curatif de la maladie en dehors de l'allogreffe de moelle osseuse. L'azacytidine ne fait que stabiliser la maladie en empêchant son évolution. On peut supposer qu'en mesurant l'entropie des CSH d'un patient SMD avant et après greffe de moelle osseuse, on devrait observer une diminution de celle-ci, mais cela reste à prouver.

Chez nos patients SMD mutés pour le gène *SF3B1*, nous avons observé une anomalie qualitative de l'érythropoïèse avec une diminution de l'amorçage des progéniteurs vers la lignée érythroblastique. Cette anomalie est encore plus prononcée chez le patient MDS2 par rapport au patient MDS4. L'entropie des CSH du patient MDS2 est également supérieur significativement à celle des CSH du patient MDS4, et il est tentant de penser que ces deux anomalies sont liées.

D'autres études sur un panel d'échantillons de SMD plus nombreux seront nécessaires pour confirmer nos observations.

Si ces observations se confirment, il faudra tenter de savoir si l'augmentation de la variabilité de l'expression génique dans les CSH est une cause ou une conséquence des SMD. L'équipe d'Olivier Gandrillon a montré dans un modèle in vitro d'érythropoïèse aviaire, qu'il était possible de diminuer le niveau de variabilité de l'expression génique en traitant les cellules avec l'artémisinine ou l'indométhacine, et inversement de l'augmenter avec le MB-3. Ainsi l'artémisinine et l'indométhacine ont diminué la quantité de cellules différenciées au contraire du MB-3 qui l'a augmenté. Pour montrer que la variabilité de l'expression génique est la cause du développement des SMD, on pourrait moduler celle-ci in vivo dans les CSH murines pour voir les effets sur l'hématopoïèse. Il parait difficile de prouver la relation de causalité entre variabilité de l'expression génique et apparition d'un SMD chez l'homme. Néanmoins, on pourrait étudier l'effet de la variabilité de l'expression génique sur les CSH

humaines dans des modèles in vitro et in vivo de SMD[205]. Ainsi, en utilisant un modèle murin de xénogreffe de SMD, on peut imaginer injecter des cellules CD34+ de SMD prétraités avec de l'indométhacine, de l'artémisine et du MB-3 pour voir si cela favorise ou diminue la prise de greffe et l'agressivité de la maladie. Pour prouver que la variabilité de l'expression génique est une conséquence des SMD, il faudrait transformer des cellules souches hématopoïétiques saines en cellules de SMD et voir si la variabilité de l'expression génique augmente. Pour cela, on peut imaginer introduire artificiellement dans ces cellules des mutations et/ou les placer dans un environnement propice au développement de la maladie. Cela nous permettrait de savoir par exemple si l'apparition de certaines mutations dans les CSH ou bien le contact avec le microenvironnement médullaire des SMD peut entrainer une augmentation de la variabilité de l'expression génique.

Nos données montrent que la variabilité de l'expression génique des CSH du patient MDS2 est supérieur à celle du patient MDS4, associé au fait que cette variabilité augmente au cours de l'évolution de la maladie, on peut alors penser que la valeur basale de la variabilité de l'expression génique dans les CSH pourrait être prédictive de la survie ou de la progression de la maladie.

Nous avons identifié les gènes les plus différentiellement entropiques et les gènes dont l'expression varie le plus au cours de la différenciation dans des données publiques de scRNA-Seq de moelle osseuse de sujet sain jeune. Nous allons faire de même dans nos données de moelle osseuse de sujet âgés, de SMD et au cours du traitement. Ainsi nous allons pouvoir savoir si les gènes responsables de la variabilité de l'expression génique sont les mêmes ou bien sont différents dans les conditions que nous avons étudiées. Si ces gènes sont différents, il serait intéressant d'étudier leur rôle dans la physiopathologie des SMD.

4. L'entropie : futures utilisations.

Dans le futur, il faudrait étendre l'analyse de la variabilité de l'expression génique à tous les types de différenciation cellulaire connu. Cela permettrait de savoir si le pic de variabilité de l'expression génique est un phénomène ubiquitaire nécessaire à la différenciation cellulaire.

Des expérience de scRNA-Seq successives ont récemment été réalisées au cours de la dédifférenciation de fibroblastes murins en cellules souches pluripotentes induites[206]. Les auteurs ont montré qu'au cours de la reprogrammation, une partie des cellules deviennent pluripotentes, tandis qu'une autre partie bifurque vers un chemin alternatif ou les cellules n'acquière pas cette capacité de pluripotence. On peut faire l'hypothèse que la variabilité de l'expression génique devrait augmenter au cours du processus, puis redescendre lorsque le stade cellule souche pluripotente induite est atteint. Quant aux cellules qui n'acquièrent pas la capacité de pluripotence au cours de la reprogrammation, peut être que la variabilité de l'expression des gènes ne diminue pas et reste trop importante ce qui expliquerait l'absence de pluripotence.

Nous nous sommes limités à comparer l'entropie d'une population cellulaire entre deux conditions car nous n'avons à l'heure actuelle pas de test statistique fiable à notre disposition pour comparer les courbes d'entropie. En effet, le calcul que l'on utilise pour calculer l'entropie dépend du nombre de cellules constituant la population cellulaire étudiée. Ainsi pour comparer deux conditions entre elles, il faut qu'elles contiennent un nombre de cellules identiques. Pour comparer visuellement les courbe d'entropie sur un même graphique nous avons dû choisir un nombre identique de cellules de chaque sous type cellulaire pour les organiser sur un pseudotemps commun, ce qui nous oblige à retirer des cellules pour un des deux échantillons. L'idéal serait d'avoir une méthode statistique pour comparer les courbes d'entropie sans avoir à sous échantillonner l'une ou l'autre condition.

Il sera très utile dans le futur de savoir s'il est possible de comparer l'entropie de groupes cellulaires issus d'expériences de scRNA-Seq provenant de différents batch ou de différentes techniques. Pouvoir comparer de manière fiable l'entropie dans des données de transcriptomiques issues d'expériences et de techniques différentes serait un avantage certain pour comparer entre elles un maximum de conditions différentes. C'est pourquoi nous n'avons pour l'instant pas comparé nos deux expériences entre elles et avec des jeux de données publiques.

De nombreuses données de scRNA-Seq ont été publiées dans d'autres pathologies hématologiques acquises[207–209] et constitutionnelles[210]. Il sera intéressant d'explorer la variabilité de l'expression génique dans ces maladies. Il est probable que la variabilité de

l'expression génique joue un rôle dans les pathologies où il existe un blocage de différenciation comme les leucémies aiguës myéloïdes (LAM) et lymphoblastiques (LAL). Dans ces maladies, on peut imaginer que l'absence de pic dans la variabilité de l'expression génique pourrait être la cause, où le reflet de l'absence de différenciation. Dans les maladies caractérisées par une prolifération de cellules matures telles que les lymphomes et les syndromes myéloprolifératifs, une modification de la variabilité de l'expression génique semble intuitivement moins évidente.

Il existe depuis peu de nouvelles techniques permettant d'associer les données transcriptomiques avec les données de génotypage à l'échelle unicellulaire. Soit les mutations sont détectées directement dans les données de scRNA-Seq[211], soit la transcriptomique est couplé avec la génomique sur la même cellule[212]. Ces techniques permettraient de comparer la variabilité de l'expression génique dans les cellules mutées et non mutées et ainsi de montrer le potentiel impact de différentes mutations fréquemment mises en évidences dans les hémopathies myéloïdes. Enfin, des données récentes de scRNA-Seq couplé au Single-Cell ATAC-seq (Assay for Transposase-Accessible Chromatin) sur l'hématopoïèse fœtale hépatique et médullaire, permettrait de comparer la variabilité de l'expression génique avec l'état d'ouverture de la chromatine[213].

La variabilité de l'expression génique semble jouer un rôle important dans les processus de différenciation, il serait intéressant d'observer si cette variabilité est aussi présente au niveau de l'expression protéique. Il existe des jeux de données de mesure de l'expression protéique à l'échelle unicellulaire sur des échantillons de moelle osseuse mesurée par cytométrie en flux et cytométrie de masse[196]. Malheureusement, ces données ne permettent pas forcément d'ordonner les cellules le long d'une voie de différenciation au vu du nombre limité de marqueurs protéiques analysés. En revanche, des données de CITE-Seq publiées[194] sur la moelle osseuse saine pourrait nous permettre d'organiser les cellules le long des voies de différenciation avec les données transcriptomiques pour ensuite mesurer la variabilité de l'expression des marqueurs protéiques. Ces techniques permettent de mesurer l'expression d'une centaine de protéine, ce qui est peu en comparaison avec les presque 20000 gènes détectés dans les expériences de scRNA-Seq. De plus les mesures d'expression protéiques de ces méthodes sont basées sur des réactions Antigène-Anticorps et sont donc

limitées par la perméabilité cellulaire, l'encombrement stérique, l'accessibilité des épitopes et la disponibilité d'anticorps hautement spécifiques qui se lient stoechiométriquement à leurs protéines apparentées[214]. L'idéal serait d'utiliser une technique de protéomique à l'échelle unicellulaire récemment développée[215], beaucoup plus spécifique et permettant de détecter jusqu'à 4000 protéines par cellule.

Conclusion

Les technologies développées au cours de la dernière décennie et notamment le scRNA-Seq ont permis de mieux décrire et de changer significativement notre point de vue sur l'hématopoïèse normale et pathologique. Ces nouvelles technologies produisent des quantités de données très importantes dont l'interprétation ne peut se faire que grâce à une collaboration étroite entre les biologistes, les bioinformaticiens et les mathématiciens. C'est grâce à ce type de collaboration que nous avons pu mettre en évidence l'importance de la variabilité de l'expression génique au cours de l'hématopoïèse et dans les syndromes myélodysplasiques. Ce travail décrit un phénomène que seule la biologie des systèmes à l'échelle unicellulaire aura permis de mettre en évidence, ouvrant ainsi la porte à de nombreux projets actuels et futurs qui auront, à terme, des impacts sur la prise en charge des pathologies cancéreuses entre autres domaines.

Bibliographie

1. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
2. Fan, H. C., Fu, G. K. & Fodor, S. P. A. Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, 1258367 (2015).
3. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The Technology and Biology of Single-Cell RNA Sequencing. *Mol. Cell* **58**, 610–620 (2015).
4. Training Modules - 10x Genomics. https://www.10xgenomics.com/videos/training-modules/.
5. *Illumina Sequencing by Synthesis*. https://www.youtube.com/watch?v=fCd6B5HRaZ8
6. scRNA-tools. https://www.scrna-tools.org/.
7. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
8. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).
9. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
10. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nat. Methods* **14**, 565–571 (2017).
11. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostat. Oxf. Engl.* **19**, 562–578 (2018).
12. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 1–15 (2019).
13. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
14. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst.* **2**, 239–250 (2016).
15. F.R.S, K. P. LIII. On lines and planes of closest fit to systems of points in space. *Lond.*

Edinb. Dublin Philos. Mag. J. Sci. **2**, 559–572 (1901).

16. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2018).

17. Weinreb, C., Wolock, S. & Klein, A. M. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinforma. Oxf. Engl.* **34**, 1246–1248 (2018).

18. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).

19. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

20. What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism? *Cell Syst.* **4**, 255–259 (2017).

21. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, (2017).

22. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).

23. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).

24. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).

25. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).

26. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinforma. Oxf. Engl.* **27**, 1739–1740 (2011).

27. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

28. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).

29. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

30. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).

31. Tarca, A. L., Bhatti, G. & Romero, R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS One* **8**, e79217 (2013).

32. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).

33. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

34. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

35. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).

36. Alpert, A., Moore, L. S., Dubovik, T. & Shen-Orr, S. S. Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat. Methods* **15**, 267–270 (2018).

37. Van den Berge, K. *et al.* Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).

38. Till, J. E. & McCULLOCH, E. A. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat. Res.* **14**, 213–222 (1961).

39. Spangrude, G. J., Heimfeld, S. & Weissman, I. L. Purification and characterization of mouse hematopoietic stem cells. *Science* **241**, 58–62 (1988).

40. Seita, J. & Weissman, I. L. Hematopoietic stem cell: self-renewal versus differentiation. *WIREs Syst. Biol. Med.* **2**, 640–653 (2010).

41. Majeti, R., Park, C. Y. & Weissman, I. L. Identification of a hierarchy of multipotent hematopoietic progenitors in human cord blood. *Cell Stem Cell* **1**, 635–645 (2007).

42. Manz, M. G., Miyamoto, T., Akashi, K. & Weissman, I. L. Prospective isolation of human clonogenic common myeloid progenitors. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 11872–11877 (2002).

43. Doulatov, S. *et al.* Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. *Nat. Immunol.* **11**, 585–593 (2010).

44. Goardon, N. *et al.* Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia. *Cancer Cell* **19**, 138–152 (2011).

45. Doulatov, S., Notta, F., Laurenti, E. & Dick, J. E. Hematopoiesis: a human perspective. *Cell Stem Cell* **10**, 120–136 (2012).

46. Hoogenkamp, M. *et al.* Early chromatin unfolding by RUNX1: a molecular explanation for differential requirements during specification versus maintenance of the hematopoietic gene expression program. *Blood* **114**, 299–309 (2009).

47. Metcalf, D. Hematopoietic cytokines. *Blood* **111**, 485–491 (2008).

48. Novershtern, N. *et al.* Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis. *Cell* **144**, 296–309 (2011).

49. Nerlov, C., Querfurth, E., Kulessa, H. & Graf, T. GATA-1 interacts with the myeloid PU.1 transcription factor and represses PU.1-dependent transcription. *Blood* **95**, 2543–2551 (2000).

50. Dahl, R., Iyer, S. R., Owens, K. S., Cuylear, D. D. & Simon, M. C. The transcriptional repressor GFI-1 antagonizes PU.1 activity through protein-protein interaction. *J. Biol. Chem.* **282**, 6473–6483 (2007).

51. Starck, J. *et al.* Functional cross-antagonism between transcription factors FLI-1 and EKLF. *Mol. Cell. Biol.* **23**, 1390–1402 (2003).

52. Bouilloux, F. *et al.* EKLF restricts megakaryocytic differentiation at the benefit of erythrocytic differentiation. *Blood* **112**, 576–584 (2008).

53. Ji, H. *et al.* Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467**, 338–342 (2010).

54. Trowbridge, J. J., Snow, J. W., Kim, J. & Orkin, S. H. DNA methyltransferase 1 is essential for and uniquely regulates hematopoietic stem and progenitor cells. *Cell Stem Cell* **5**, 442–449 (2009).

55. Wilting, R. H. *et al.* Overlapping functions of Hdac1 and Hdac2 in cell cycle regulation and haematopoiesis. *EMBO J.* **29**, 2586–2597 (2010).

56. Kluiver, J., Kroesen, B.-J., Poppema, S. & van den Berg, A. The role of microRNAs in normal hematopoiesis and hematopoietic malignancies. *Leukemia* **20**, 1931–1936 (2006).

57. Schofield, R. The relationship between the spleen colony-forming cell and the haemopoietic stem cell. *Blood Cells* **4**, 7–25 (1978).

58. Yu, V. W. C. & Scadden, D. T. Chapter Two - Hematopoietic Stem Cell and Its Bone Marrow Niche. in *Current Topics in Developmental Biology* (ed. Bresnick, E. H.) vol. 118 21–44 (Academic Press, 2016).

59. Acar, M. *et al.* Deep imaging of bone marrow shows non-dividing stem cells are mainly perisinusoidal. *Nature* **526**, 126–130 (2015).

60. Kunisaki, Y. *et al.* Arteriolar niches maintain haematopoietic stem cell quiescence. *Nature* **502**, 637–643 (2013).

61. Asada, N. *et al.* Differential cytokine contributions of perivascular haematopoietic stem cell niches. *Nat. Cell Biol.* **19**, 214–223 (2017).

62. Comazzetto, S. *et al.* Restricted Hematopoietic Progenitors and Erythropoiesis Require SCF from Leptin Receptor+ Niche Cells in the Bone Marrow. *Cell Stem Cell* **24**, 477-486.e6 (2019).

63. Beerman, I., Luis, T. C., Singbrant, S., Lo Celso, C. & Méndez-Ferrer, S. The evolving view of the hematopoietic stem cell niche. *Exp. Hematol.* **50**, 22–26 (2017).

64. Kusumbe, A. P., Ramasamy, S. K. & Adams, R. H. Coupling of angiogenesis and osteogenesis by a specific vessel subtype in bone. *Nature* **507**, 323–328 (2014).

65. Itkin, T. *et al.* Distinct bone marrow blood vessels differentially regulate haematopoiesis. *Nature* **532**, 323–328 (2016).

66. Hooper, A. T. *et al.* Engraftment and reconstitution of hematopoiesis is dependent on VEGFR2-mediated regeneration of sinusoidal endothelial cells. *Cell Stem Cell* **4**, 263–274 (2009).

67. Méndez-Ferrer, S., Lucas, D., Battista, M. & Frenette, P. S. Haematopoietic stem cell release is regulated by circadian oscillations. *Nature* **452**, 442–447 (2008).

68. García-García, A. *et al.* Dual cholinergic signals regulate daily migration of hematopoietic stem cells and leukocytes. *Blood* **133**, 224–236 (2019).

69. Zhao, M. *et al.* N-Cadherin-Expressing Bone and Marrow Stromal Progenitor Cells Maintain Reserve Hematopoietic Stem Cells. *Cell Rep.* **26**, 652-669.e6 (2019).

70. Sugiyama, T., Kohara, H., Noda, M. & Nagasawa, T. Maintenance of the hematopoietic stem cell pool by CXCL12-CXCR4 chemokine signaling in bone marrow stromal cell niches. *Immunity* **25**, 977–988 (2006).

71. Kiel, M. J. *et al.* SLAM Family Receptors Distinguish Hematopoietic Stem and Progenitor Cells and Reveal Endothelial Niches for Stem Cells. *Cell* **121**, 1109–1121 (2005).

72. Sacchetti, B. *et al.* Self-renewing osteoprogenitors in bone marrow sinusoids can organize a hematopoietic microenvironment. *Cell* **131**, 324–336 (2007).

73. Ho, Y.-H. *et al.* Remodeling of Bone Marrow Hematopoietic Stem Cell Niches Promotes Myeloid Cell Expansion during Premature or Physiological Aging. *Cell Stem Cell* **25**, 407-418.e6 (2019).

74. Ding, L., Saunders, T. L., Enikolopov, G. & Morrison, S. J. Endothelial and perivascular cells maintain haematopoietic stem cells. *Nature* **481**, 457–462 (2012).

75. Omatsu, Y. *et al.* The essential functions of adipo-osteogenic progenitors as the hematopoietic stem and progenitor cell niche. *Immunity* **33**, 387–399 (2010).

76. Li, H. *et al.* Low/negative expression of PDGFR-α identifies the candidate primary mesenchymal stromal cells in adult human bone marrow. *Stem Cell Rep.* **3**, 965–974 (2014).

77. Tormin, A. *et al.* CD146 expression on primary nonhematopoietic bone marrow stem cells is correlated with in situ localization. *Blood* **117**, 5067–5077 (2011).

78. Zhao, M. *et al.* Megakaryocytes maintain homeostatic quiescence and promote post-injury regeneration of hematopoietic stem cells. *Nat. Med.* **20**, 1321–1326 (2014).

79. Nakamura-Ishizu, A., Takubo, K., Fujioka, M. & Suda, T. Megakaryocytes are essential for HSC quiescence through the production of thrombopoietin. *Biochem. Biophys. Res. Commun.* **454**, 353–357 (2014).

80. Yamazaki, S. *et al.* Nonmyelinating Schwann cells maintain hematopoietic stem cell hibernation in the bone marrow niche. *Cell* **147**, 1146–1158 (2011).

81. Méndez-Ferrer, S. *et al.* Bone marrow niches in haematological malignancies. *Nat. Rev. Cancer* 1–14 (2020) doi:10.1038/s41568-020-0245-2.

82. Notta, F. *et al.* Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116 (2016).

83. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

84. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).

85. Velten, L. *et al.* Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19**, 271–281 (2017).

86. Karamitros, D. *et al.* Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. *Nat. Immunol.* **19**, 85 (2018).

87. Psaila, B. *et al.* Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways. *Genome Biol.* **17**, 83 (2016).

88. Miyawaki, K. *et al.* Identification of unipotent megakaryocyte progenitors in human hematopoiesis. *Blood* blood-2016-09-741611 (2017) doi:10.1182/blood-2016-09-741611.

89. Pellin, D. *et al.* A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat. Commun.* **10**, 2395 (2019).

90. Hoppe, P. S. *et al.* Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature* **535**, 299–302 (2016).

91. Strasser, M. K. *et al.* Lineage marker synchrony in hematopoietic genealogies refutes the PU.1/GATA1 toggle switch paradigm. *Nat. Commun.* **9**, 2697 (2018).

92. Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418–426 (2018).

93. Hay, S. B., Ferchen, K., Chetal, K., Grimes, H. L. & Salomonis, N. The Human Cell Atlas bone marrow single-cell interactive web portal. *Exp. Hematol.* **68**, 51–61 (2018).

94. Watcham, S., Kucinski, I. & Gottgens, B. New Insights into Haematopoietic Differentiation Landscapes from scRNA-seq. *Blood* blood-2018-08-835355 (2019) doi:10.1182/blood-2018-08-835355.

95. Cabezas-Wallscheid, N. *et al.* Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy. *Cell* **169**, 807-823.e19 (2017).

96. Rodriguez-Fraticelli, A. E. *et al.* Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212–216 (2018).

97. Tusi, B. K. *et al.* Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).

98. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).

99. Hayashi, T. *et al.* Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.* **9**, 619 (2018).

100. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).

101. Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35**, 936–939 (2017).

102. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* (2018) doi:10.1016/j.cell.2018.03.074.

103. Clark, S. J. *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).

104. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, (2018).

105. Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 928-943.e22 (2019).

106. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).

107. Spanjaard, B. *et al.* Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).

108. Biasco, L. *et al.* In Vivo Tracking of Human Hematopoiesis Reveals Patterns of Clonal Dynamics during Early and Steady-State Reconstitution Phases. *Cell Stem Cell* **19**, 107–119 (2016).

109. Biezuner, T. *et al.* A generic, cost-effective, and scalable cell lineage analysis platform. *Genome Res.* **26**, 1588–1599 (2016).

110. Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325-1339.e22 (2019).

111. Giladi, A. *et al.* Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat. Cell Biol.* **20**, 836 (2018).

112. Jacobsen, S. E. W. & Nerlov, C. Haematopoiesis in the era of advanced single-cell technologies. *Nat. Cell Biol.* **21**, 2 (2019).

113. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).

114. Gandrillon, O., Kolesnik-Antoine, D., Kupiec, J.-J. & Beslon, G. Chance at the heart of the cell. *Prog. Biophys. Mol. Biol.* **110**, 1–4 (2012).

115. Richard, A. Analyse de la variabilité de l'expression génique et du métabolisme glycolytique au cours du processus de différenciation érythrocytaire : de l'analyse à grande échelle aux questions mécanistiques. (Université de Lyon, 2018).

116. Sharp, K. & Matschinsky, F. Translation of Ludwig Boltzmann's Paper "On the Relationship between the Second Fundamental Theorem of the Mechanical Theory of Heat and Probability Calculations Regarding the Conditions for Thermal Equilibrium" Sitzungberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissen Classe. Abt. II, LXXVI 1877, pp 373-435 (Wien. Ber. 1877, 76:373-435). Reprinted in Wiss. Abhandlungen, Vol. II, reprint 42, p. 164-223, Barth, Leipzig, 1909. *Entropy* **17**, 1971–2009 (2015).

117. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

118. Richard, A. *et al.* Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. *PLoS Biol.* **14**, e1002585 (2016).

119. Stumpf, P. S. *et al.* Stem Cell Differentiation as a Non-Markov Stochastic Process. *Cell Syst.* **5**, 268-282.e7 (2017).

120. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).

121. Rué, P. & Martinez Arias, A. Cell dynamics and gene expression control in tissue homeostasis and development. *Mol. Syst. Biol.* **11**, 792 (2015).

122. Haas, S. *et al.* Inflammation-Induced Emergency Megakaryopoiesis Driven by Hematopoietic Stem Cell-like Megakaryocyte Progenitors. *Cell Stem Cell* **17**, 422–434 (2015).

123. Pina, C. *et al.* Inferring rules of lineage commitment in haematopoiesis. *Nat. Cell Biol.* **14**, 287–294 (2012).

124. Kouno, T. *et al.* Temporal dynamics and transcriptional control using single-cell gene expression analysis. *Genome Biol.* **14**, R118 (2013).

125. Kupiec, J. J. A Darwinian theory for the origin of cellular differentiation. *Mol. Gen. Genet. MGG* **255**, 201–208 (1997).

126. Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* **6**, 451–464 (2005).

127. Huang, S. Systems biology of stem cells: three useful perspectives to help overcome the paradigm of linear pathways. *Philos. Trans. R. Soc. B Biol. Sci.* **366**, 2247–2259 (2011).

128. Yvert, G. 'Particle genetics': treating every cell as unique. *Trends Genet. TIG* **30**, 49–56 (2014).

129. Rebhahn, J. A. *et al.* An animated landscape representation of CD4+ T-cell differentiation, variability, and plasticity: insights into the behavior of populations versus cells. *Eur. J. Immunol.* **44**, 2216–2229 (2014).

130. Mojtahedi, M. *et al.* Cell Fate Decision as High-Dimensional Critical State Transition. *PLoS Biol.* **14**, e2000640 (2016).

131. Semrau, S. *et al.* Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat. Commun.* **8**, 1096 (2017).

132. Guillemin, A., Duchesne, R., Crauste, F., Gonin-Giraud, S. & Gandrillon, O. Drugs modulating stochastic gene expression affect the erythroid differentiation process. *PloS One* **14**, e0225166 (2019).

133. Ma, X. Epidemiology of myelodysplastic syndromes. *Am. J. Med.* **125**, S2-5 (2012).

134. Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).

135. Montalban-Bravo, G. & Garcia-Manero, G. Myelodysplastic syndromes: 2018 update on diagnosis, risk-stratification and management. *Am. J. Hematol.* **93**, 129–147 (2018).

136. Haase, D. *et al.* New insights into the prognostic impact of the karyotype in MDS and correlation with subtypes: evidence from a core dataset of 2124 patients. *Blood* **110**, 4385–4395 (2007).

137. Eclache, V. *et al.* Cytogenetic place in managing myelodysplastic syndromes: an update by the Groupe francophone de cytogénétique hématologique (GFCH). *Ann. Biol. Clin. (Paris)* **74**, 525–534 (2016).

138. Malcovati, L. *et al.* Diagnosis and treatment of primary myelodysplastic syndromes in adults: recommendations from the European LeukemiaNet. *Blood* **122**, 2943–2964 (2013).

139. Porwit, A. *et al.* Revisiting guidelines for integration of flow cytometry results in the WHO classification of myelodysplastic syndromes-proposal from the International/European LeukemiaNet Working Group for Flow Cytometry in MDS. *Leukemia* **28**, 1793–1798 (2014).

140. Ogata, K. *et al.* Diagnostic utility of flow cytometry in low-grade myelodysplastic syndromes: a prospective validation study. *Haematologica* **94**, 1066–1074 (2009).

141. Mathis, S. *et al.* Flow cytometric detection of dyserythropoiesis: a sensitive and powerful diagnostic tool for myelodysplastic syndromes. *Leukemia* **27**, 1981–1987 (2013).

142. Haferlach, T. *et al.* Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia* **28**, 241–247 (2014).

143. Papaemmanuil, E. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122**, 3616–3627 (2013).

144. Malcovati, L. *et al.* Driver somatic mutations identify distinct disease entities within myeloid neoplasms with myelodysplasia. *Blood* **124**, 1513–1521 (2014).

145. Greenberg, P. *et al.* International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood* **89**, 2079–2088 (1997).

146. Greenberg, P. L. *et al.* Revised international prognostic scoring system for

myelodysplastic syndromes. *Blood* **120**, 2454–2465 (2012).

147. Fenaux, P., Platzbecker, U. & Ades, L. How we manage adults with myelodysplastic syndrome. *Br. J. Haematol.* **n/a**,.

148. Hellström-Lindberg, E., Tobiasson, M. & Greenberg, P. Myelodysplastic syndromes: moving towards personalized management. *Haematologica* (2020) doi:10.3324/haematol.2020.248955.

149. Ogawa, S. Genetics of MDS. *Blood* **133**, 1049–1059 (2019).

150. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).

151. Gerstung, M. *et al.* Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat. Commun.* **6**, 5901 (2015).

152. Chesnais, V. *et al.* Architectural and functional heterogeneity of hematopoietic stem/progenitor cells in non-del(5q) myelodysplastic syndromes. *Blood* blood-2016-03-707745 (2016) doi:10.1182/blood-2016-03-707745.

153. Aleshin, A. & Greenberg, P. L. Molecular pathophysiology of the myelodysplastic syndromes: insights for targeted therapy. *Blood Adv.* **2**, 2787–2797 (2018).

154. Aleshin, A. *et al.* Abstract 3004: Single-cell mutational profiling of clonal evolution in myelodysplastic syndromes (MDS) during therapy and disease progression. *Cancer Res.* **78**, 3004–3004 (2018).

155. Chesnais, V. *et al.* Effect of lenalidomide treatment on clonal architecture of myelodysplastic syndromes without 5q deletion. *Blood* **127**, 749–760 (2016).

156. Nagata, Y. *et al.* Invariant patterns of clonal succession determine specific clinical features of myelodysplastic syndromes. *Nat. Commun.* **10**, 5386 (2019).

157. Papaemmanuil, E. *et al.* Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N. Engl. J. Med.* **365**, 1384–1395 (2011).

158. Malcovati, L. *et al.* Clinical significance of SF3B1 mutations in myelodysplastic syndromes and myelodysplastic/myeloproliferative neoplasms. *Blood* **118**, 6239–6246 (2011).

159. Malcovati, L. *et al.* SF3B1 mutation identifies a distinct subset of myelodysplastic syndrome with ring sideroblasts. *Blood* **126**, 233–241 (2015).

160. Mortera-Blanco, T. *et al.* SF3B1-initiating mutations in MDS-RSs target lymphomyeloid hematopoietic stem cells. *Blood* **130**, 881–890 (2017).

161. Shallis, R. M., Ahmad, R. & Zeidan, A. M. The genetic and molecular pathogenesis of

myelodysplastic syndromes. *Eur. J. Haematol.* **101**, 260–271 (2018).

162. Bondu, S. *et al.* A variant erythroferrone disrupts iron homeostasis in SF3B1-mutated myelodysplastic syndrome. *Sci. Transl. Med.* **11**, (2019).

163. Delhommeau, F. *et al.* Mutation in TET2 in myeloid cancers. *N. Engl. J. Med.* **360**, 2289–2301 (2009).

164. Langemeijer, S. M. C. *et al.* Acquired mutations in TET2 are common in myelodysplastic syndromes. *Nat. Genet.* **41**, 838–842 (2009).

165. Bejar, R., Levine, R. & Ebert, B. L. Unraveling the Molecular Pathophysiology of Myelodysplastic Syndromes. *J. Clin. Oncol.* **29**, 504–515 (2011).

166. Bejar, R. *et al.* Clinical effect of point mutations in myelodysplastic syndromes. *N. Engl. J. Med.* **364**, 2496–2506 (2011).

167. Bejar, R. *et al.* TET2 mutations predict response to hypomethylating agents in myelodysplastic syndrome patients. *Blood* **124**, 2705–2712 (2014).

168. Chao, M. P., Seita, J. & Weissman, I. L. Establishment of a normal hematopoietic and leukemia stem cell hierarchy. *Cold Spring Harb. Symp. Quant. Biol.* **73**, 439–449 (2008).

169. Nimer, S. D. MDS: a stem cell disorder--but what exactly is wrong with the primitive hematopoietic cells in this disease? *Hematol. Am. Soc. Hematol. Educ. Program* 43–51 (2008) doi:10.1182/asheducation-2008.1.43.

170. Adès, L., Itzykson, R. & Fenaux, P. Myelodysplastic syndromes. *Lancet Lond. Engl.* **383**, 2239–2252 (2014).

171. Park, S. *et al.* Predictive factors of response and survival in myelodysplastic syndrome treated with erythropoietin and G-CSF: the GFM experience. *Blood* **111**, 574–582 (2008).

172. Santini, V. *et al.* Randomized Phase III Study of Lenalidomide Versus Placebo in RBC Transfusion-Dependent Patients With Lower-Risk Non-del(5q) Myelodysplastic Syndromes and Ineligible for or Refractory to Erythropoiesis-Stimulating Agents. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **34**, 2988–2996 (2016).

173. Stahl, M. *et al.* The use of immunosuppressive therapy in MDS: clinical outcomes and their predictors in a large international patient cohort. *Blood Adv.* **2**, 1765–1772 (2018).

174. Park, S. *et al.* Outcome of Lower-Risk Patients With Myelodysplastic Syndromes Without 5q Deletion After Failure of Erythropoiesis-Stimulating Agents. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **35**, 1591–1597 (2017).

175. Alessandrino, E. P. *et al.* Prognostic impact of pre-transplantation transfusion history

and secondary iron overload in patients with myelodysplastic syndrome undergoing allogeneic stem cell transplantation: a GITMO study. *Haematologica* **95**, 476–484 (2010).

176. Fenaux, P. *et al.* Luspatercept in Patients with Lower-Risk Myelodysplastic Syndromes. *N. Engl. J. Med.* **382**, 140–151 (2020).

177. Oliva, E. N. *et al.* Eltrombopag versus placebo for low-risk myelodysplastic syndromes with thrombocytopenia (EQoL-MDS): phase 1 results of a single-blind, randomised, controlled, phase 2 superiority trial. *Lancet Haematol.* **4**, e127–e136 (2017).

178. de Witte, T. *et al.* Allogeneic hematopoietic stem cell transplantation for MDS and CMML: recommendations from an international expert panel. *Blood* **129**, 1753–1762 (2017).

179. Fenaux, P. *et al.* Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase III study. *Lancet Oncol.* **10**, 223–232 (2009).

180. Itzykson, R. *et al.* Prognostic factors for response and overall survival in 282 patients with higher-risk myelodysplastic syndromes treated with azacitidine. *Blood* **117**, 403–411 (2011).

181. Itzykson, R. *et al.* Impact of TET2 mutations on response rate to azacitidine in myelodysplastic syndromes and low blast count acute myeloid leukemias. *Leukemia* **25**, 1147–1152 (2011).

182. DiNardo, C. D. *et al.* Durable Remissions with Ivosidenib in IDH1-Mutated Relapsed or Refractory AML. *N. Engl. J. Med.* **378**, 2386–2398 (2018).

183. Welch, J. S. *et al.* TP53 and Decitabine in Acute Myeloid Leukemia and Myelodysplastic Syndromes. *N. Engl. J. Med.* **375**, 2023–2036 (2016).

184. Sallman, D. A. *et al.* Phase 2 Results of APR-246 and Azacitidine (AZA) in Patients with TP53 mutant Myelodysplastic Syndromes (MDS) and Oligoblastic Acute Myeloid Leukemia (AML). *Blood* **134**, 676–676 (2019).

185. Dussiau, C. & Fontenay, M. Mechanisms underlying the heterogeneity of myelodysplastic syndromes. *Exp. Hematol.* **58**, 17–26 (2018).

186. Meunier, M. *et al.* Molecular dissection of engraftment in a xenograft model of myelodysplastic syndromes. *Oncotarget* **9**, 14993–15000 (2018).

187. Silvin, A. *et al.* Elevated calprotectin and abnormal myeloid cell subsets discriminate severe from mild COVID-19. *Cell* **0**, (2020).

188. Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and

interpretation of cytometry data. *Cytom. Part J. Int. Soc. Anal. Cytol.* **87**, 636–645 (2015).

189. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

190. Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).

191. Petukhov, V. *et al.* dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biol.* **19**, 78 (2018).

192. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

193. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).

194. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

195. Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).

196. Oetjen, K. A. *et al.* Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. *JCI Insight* **3**, (2018).

197. Sun, Y. & Ma, L. New Insights into Long Non-Coding RNA MALAT1 in Cancer and Metastasis. *Cancers* **11**, (2019).

198. Mende, N. *et al.* Quantitative and molecular differences distinguish adult human medullary and extramedullary haematopoietic stem and progenitor cell landscapes. *bioRxiv* 2020.01.26.919753 (2020) doi:10.1101/2020.01.26.919753.

199. Morinishi, L., Kochanowski, K., Levine, R. L., Wu, L. F. & Altschuler, S. J. Loss of TET2 Affects Proliferation and Drug Sensitivity through Altered Dynamics of Cell-State Transitions. *Cell Syst.* **11**, 86-94.e5 (2020).

200. Mende, N. *et al.* Quantitative and molecular differences distinguish adult human medullary and extramedullary haematopoietic stem and progenitor cell landscapes. *bioRxiv* 2020.01.26.919753 (2020) doi:10.1101/2020.01.26.919753.

201. Pellegrino, M. *et al.* High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. *Genome Res.* **28**, 1345–1352 (2018).

202. Song, Y. *et al.* A highly efficient and faithful MDS patient-derived xenotransplantation model for pre-clinical studies. *Nat. Commun.* **10**, 366 (2019).

203. Shyh-Chang, N. & Ng, H.-H. The metabolic programming of stem cells. *Genes Dev.* **31**, 336–346 (2017).

204. Vitale, I., Manic, G., De Maria, R., Kroemer, G. & Galluzzi, L. DNA Damage in Stem Cells. *Mol. Cell* **66**, 306–319 (2017).

205. Rouault-Pierre, K. *et al.* Preclinical modeling of myelodysplastic syndromes. *Leukemia* (2017) doi:10.1038/leu.2017.172.

206. Guo, L. *et al.* Resolving Cell Fate Decisions during Somatic Cell Reprogramming by Single-Cell RNA-Seq. *Mol. Cell* **73**, 815-829.e7 (2019).

207. van Galen, P. *et al.* Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265-1281.e24 (2019).

208. Caron, M. *et al.* Single-cell analysis of childhood leukemia reveals a link between developmental states and ribosomal protein expression as a source of intra-individual heterogeneity. *Sci. Rep.* **10**, 8079 (2020).

209. Giustacchini, A. *et al.* Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat. Med.* **23**, 692–702 (2017).

210. Wu, Z. *et al.* Sequencing of RNA in single cells reveals a distinct transcriptome signature of hematopoiesis in GATA2 deficiency. *Blood Adv.* **4**, 2702–2716 (2020).

211. Petti, A. A. *et al.* A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat. Commun.* **10**, 3660 (2019).

212. Nam, A. S. *et al.* Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature* **571**, 355–360 (2019).

213. Ranzoni, A. M. *et al.* Integrative Single-cell RNA-Seq and ATAC-Seq Analysis of Human Foetal Liver and Bone Marrow Haematopoiesis. *bioRxiv* 2020.05.06.080259 (2020) doi:10.1101/2020.05.06.080259.

214. Levy, E. & Slavov, N. Single cell protein analysis for systems biology. *Essays Biochem.* **62**, 595–605 (2018).

215. Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* **19**, 161 (2018).

Abstract :

Myelodysplastic syndromes (MDS) are clonal hematopoietic stem cell (HSC) malignancies. The accumulation of mutations and and/or cytogenetic abnormalities in bone marrow HSC compartment confers a proliferative advantage to pathological cells. These diseases are characterized by dysplastic and ineffective hematopoiesis and evolve in 40% of cases into acute myeloid leukemia with poor prognosis. In this work, we mapped the MDS CD34+ compartment with flow cytometry, and also with single-cell RNA-Seq that allow to characterize the transcriptome at single cell level.

We demonstrated that the characterization of CD34+ stem and progenitor cells of MDS by flow cytometry can differentiate them from healthy age-matched controls. Indeed, some MDS appear to have a diminished proportion of CMPMEP3 populations as compared to controls. In addition, a subgroup of MDS appears to have an increased number of B/NK progenitors relative to the other SMD and controls. We also showed that CMP and LMPP populations are over-represented in a portion of high-risk MDS. In addition, we determined that the entropy of cell distribution within the different highlighted clusters, which reflects the cell diversity of the CD34+ compartment, is significantly decreased in high-risk MDS consistent with the expansion of a pathological clone to the detriment of the diversity of normal hematopoiesis.

From publicly available scRNA-Seq healthy donor bone marrow data, we have established that during hematopoietic differentiation there is a peak in intercellular gene expression variability as measured by Shannon entropy. This entropy peak is observed in all the processes studied including erythropoiesis, granulopoiesis, B lymphopoiesis as well as in the maturation process of dendritic cells. Our work showed that this entropy peak is also present in MDS hematopoiesis. In addition, we showed that MDS HSCs had significantly higher entropy than age-matched healthy controls HSCs Our data also allowed us to observe significant variations in HSC entropy associated with the response to one-year azacytidine therapy in MDS patients. There are therefore variabilities in intercellular gene expression objectified by Shannon entropy in the HSCs of MDS patients that should be compared with the dysplasia and intramedullary abortion of hematopoietic progenitors leading to cytopenias observed in these diseases.

Keywords : myelodysplastic syndromes, hematopoiesis, gene expression variability, entropy, differentiation, stem and progenitors cells, scRNA-Seq.

Charles Dussiau **Université de Paris, année 2020**

Etude à l'échelle unicellulaire du compartiment des cellules souches et progénitrices des syndromes myélodysplasiques.

Résumé de la thèse :

Les syndromes myélodysplasiques (SMD) sont des maladies clonales de la cellule souche hématopoïétique (CSH). L'accumulation de mutations et/ou d'anomalies cytogénétiques dès le compartiment des CSH de la moelle osseuse des malades confère un avantage prolifératif aux cellules pathologiques au détriment de l'hématopoïèse normale. Ces maladies sont caractérisées par une hématopoïèse dysplasique et inefficace, et évoluent dans 40% des cas en leucémie aiguë myéloïde de pronostic sombre. Nous avons cherché à cartographier le compartiment des cellules CD34+ de SMD par cytométrie en flux, mais aussi en utilisant la technologie single-cell RNA-Seq (scRNA-Seq) qui permet la caractérisation du transcriptome à l'échelle unicellulaire.

Nous avons démontré que la caractérisation des cellules souches et progénitrices CD34+ des SMD par cytométrie en flux peut permettre de les différencier des sujets sains d'âge comparable. Une partie des SMD semble avoir une diminution des populations CMPMEP3 par rapport aux témoins. De plus un sous-groupe de SMD semble présenter un nombre augmenté de progéniteurs B/NK par rapport aux autres SMD et aux témoins. Nous montrons également que les populations CMP et LMPP sont surreprésentés dans une partie des SMD de haut risque. De plus, nous avons déterminé que l'entropie de la répartition des cellules au sein des différents clusters mis en évidence et qui reflète la diversité cellulaire du compartiment CD34+, est significativement diminué dans les SMD de haut risque en accord avec l'expansion d'un clone pathologique au détriment de la diversité de l'hématopoïèse normale.

A partir de données publiques de scRNA-Seq de moelle osseuse de donneur sain, nous avons établi qu'au cours de la différenciation hématopoïétique il existe un pic de variabilité de l'expression génique intercellulaire mesuré par l'entropie de Shannon. Ce pic d'entropie est observé dans tous les processus étudiés dont l'érythropoïèse, la granulopoïèse, la lymphopoïèse B ainsi que dans le processus de maturation des cellules dendritiques. Nos travaux démontrent que ce pic d'entropie est également présent dans l'hématopoïèse des SMD. Par ailleurs, nous avons montré que les CSH de SMD présentent une entropie significativement supérieure aux CSH de témoins sains appariés en âge. Nos données nous ont également permis d'observer des variations significatives de l'entropie des CSH associées à la réponse à un traitement au long court par azacytidine chez les patients SMD. Il existe donc des variabilités de l'expression génique intercellulaire objectivée par l'entropie de Shannon au sein des CSH de patients SMD qu'il convient de mettre en regard de la dysplasie et de l'avortement intramédullaire des progéniteurs hématopoïétiques aboutissant aux cytopénies observées dans ces maladies.

Mots clés : syndromes myélodysplasiques, hématopoïèse, variabilité de l'expression génique, entropie, différenciation, cellules souches et progénitrices, scRNA-Seq.