Université de Montréal

Étude de l'hétérogénéité génétique de la leucémie myéloïde aigue par analyse scRNA-seq

Par

Azer Farah

Département de Biochimie, Université de Montréal, Faculté de Médecine

Mémoire présenté en vue de l'obtention du grade de Maitrise en bio-informatique

Octobre 2021

Université de Montréal

Unité académique : Département de Biochimie, Université de Montréal, Faculté de Médecine

---

Ce mémoire (ou cette thèse) intitulé(e)

Étude de l'hétérogénéité génétique de la leucémie myéloïde aigue par analyse scRNA-seq

Présenté par

Azer Farah

A été évalué(e) par un jury composé des personnes suivantes

Morgan Craig

Président-rapporteur

Vincent Philippe Lavallée

Directeur de recherche

Sébastien Lemieux

Codirecteur

Julie Lessard

Membre du jury

Résumé

Les leucémies myéloïdes aiguës (LMA) sont un groupe de cancers résultant de la différenciation anormale et incomplète des cellules souches et progénitrices hématopoïétiques (HSPC), suite à l'acquisition séquentielle de diverses anomalies génétiques et cytogénétiques. Ce processus se reflète probablement dans l'hétérogénéité cellulaire de la LMA mais reste mal caractérisé. Les technologies de séquençage de l'ARN sur cellule unique (scRNA-seq) ont permis d'explorer l'hétérogénéité phénotypique. Cependant, déduire l'hétérogénéité génotypique telle que les variantes sous-clonales d'un seul nucléotide (SNV) et les variations du nombre de copies (CNV) est très difficile en partie à cause de la rareté des données.

Pour résoudre ce problème, nous avons développé un classificateur de forêt aléatoire pour annoter les cellules LMA. Nous avons développé un pipeline pour identifier les mutations liées à la LMA qui peuvent être détectées dans scRNA-seq. Nous avons combiné les données scRNA-seq avec les données de séquençage en « Bulk » d'exome appariées tumoraux et sains des mêmes échantillons pour définir la sous-structure clonale dans ces échantillons.

Nous avons appliqué notre classificateur à plus de 130K cellules obtenues à partir de 20 patients LMA en utilisant le système *10X Genomics Chromium*. Nous avons identifié 35 types cellulaires distincts, y compris un grand nombre de cellules de type HSPC. Dans cette cohorte, nous avons remarqué que des mutations dans les gènes *NPM1, U2AF1, SMC3, EZH2, RAD21* et *KRAS* peuvent être détectées dans les données scRNA-seq à des occurrences allant de 0,02 % à 75 % de cellules mutées par échantillon. Dans huit échantillons, nous avons identifié des sous-populations de cellules tumorales portant de grandes CNV telles que les aneuploïdies des chromosomes 5 et 7. Ces aneuploïdies sont récurrentes et pertinentes sur le plan pronostique dans la LMA. Notre travail fournit un outil de recherche unique pour étudier la relation entre la diversité phénotypique et génotypique ; offrant de nouvelles perspectives sur le développement de la leucémie.

Mots-clés : leucémie myéloïde aiguë, scRNA-seq, SNV, CNV et annotation cellulaire.

Abstract

Acute myeloid leukemias (AML) are a group of cancers resulting from the abnormal and incomplete differentiation of hematopoietic stem and progenitor cells (HSPC), following the sequential acquisition of various genetic and cytogenetic abnormalities. This process is likely reflected in the AML cellular heterogeneity but it remains poorly characterized. Single-cell RNA sequencing (scRNA-seq) technologies enabled the exploration of phenotypic heterogeneity. However, inferring the genotypic heterogeneity such as subclonal single nucleotide variants (SNV) and copy number variations (CNV) is highly challenging partly because of data sparsity.

To address this, we developed a random forest classifier to annotate AML cells. We developed a pipeline to identify which of the known AML driver mutations can be detected in scRNA-seq. We combined scRNA-seq data with bulk tumoral and germline exomes data from the same samples to define the clonal substructure in these samples.

We applied our classifier to over 130K cells obtained from 20 AML patients using the *10X Genomics* Chromium system. We identified 35 distinct cell types including large numbers of HSPC-like. In this cohort, we noticed that mutations in *NPM1, U2AF1, SMC3, EZH2, RAD21* and *KRAS* genes can be detected in scRNA-seq data at occurrences ranging from 0.02% to 75% of mutated cells per sample. In eight samples, we identified sub-populations of tumor cells carrying large CNVs such as aneuploidies of chromosomes 5 and 7. These aneuploidies are recurrent and prognostically relevant in AML.

Our work provides a unique research tool to investigate the relationship between phenotypic and genotypic diversity; offering novel insights into leukemia development.

Keywords : Acute myeloid leukemia, scRNA-seq, CNV, SNV, cell type annotation.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

ACP : Analyse en composantes principales

AEP : Analyse d'enrichissement des voies

AML : Acute myeloid leunemia

BC : Code-barres

BCLQ : Banque de cellules leucémiques du Québec

CH : Hématopoïèse Clonale

CHIP : Hématopoïèse clonale à potentiel indéterminé

CK : Complexe karyotype

CLP : Progéniteur lymphoïde commun

CMP : Progéniteur myéloïde commun

CNV : Copy Number Variation

DT : Arbre de décision

ELN : European LeukemiaNet

FAB : French-American-British

FC : Fold change

FDR : False discovery rate

GDE : Gènes différentiellement exprimés

GMP : Progéniteur des granulocytes/macrophages

HCA : Human cell Atlas

HSC : Hematopetic stem cell

HSPC : Cellules souches et progénitrices hématopoïétiques

KNN : K plus proches voisins

KS : Kolmogorov-Smirnov

LMA : Leucemie Myeloide Aigue

LMA : Leucémie myéloïde aigue

LT-HSC : Cellule souche hématopoïétique à long terme

MCC : Coefficient de corrélation de Matthews

MCMC : Markov Chain Monte Carlo

MEP : Progéniteur mégacaryocyte/érythroïde

MLP : Progéniteur multi-lymphoïde

MPP : Cellule Progénitrice multipotente

NK : Natural Killer

NK : Normal karyotype

OMS : Organisation mondiale de la santé

OOB : Out of bag

PC : Composantes principales

RCBT : Réseau canadien de banques de tissus

RD : Réduction dimensionnelle

RF : Random forest

scRNA-seq : single cell RNA sequencing

SMD : Syndromes myélodysplasiques

SNV : Single Nucletide Variation

ST-HSC, Cellule souche hématopoïétique à court terme

t-SNE : t-Stochastic Neighbor Embedding

UMAP : Uniform Manifold Approximation and Projection

UMI : Unique molecular identifier

WHO : World Health Organization

À la mémoire de mon père,

À ma famille,

À mes amis…

Remerciements

Je remercie les Dr Morgan Craig et Dr Julie Lessard pour avoir accepté de juger mon travail et pour l'honneur qu'elles m'ont fait en participant à mon jury.

Cette partie est l'occasion pour moi de remercier les personnes qui ont participé de près ou de loin à ma maitrise ainsi qu'à la création d'un environnement idéal pour que je puisse la réaliser dans les meilleures conditions possibles. Il n'y a pas de mots pour exprimer la gratitude que j'ai envers mes directeurs de recherche, Dr Vincent Philippe Lavallée et Dr Sébastien Lemieux.

Je remercie Sébastien Lemieux pour la confiance qu'il m'a accordé en acceptant de co-encadrer ce travail. Sa compétence scientifique, ses conseils et ses commentaires ont été très précieux pour mener à bien ce travail. Je suis infiniment heureux et honoré d'avoir fait ma maitrise sous sa codirection.

Bien que je fusse son premier étudiant, Vincent Philippe Lavallée a été un directeur de recherche exemplaire. Sa démarche scientifique, son esprit critique et sa capacité de vulgarisation sont et resteront des modèles pour moi. Je le remercie sincèrement pour sa présence, sa disponibilité et sa patience. Aucune expression de gratitude ne sera suffisante pour lui exprimer mon respect et ma reconnaissance. Il a également su me laisser l'indépendance et l'autonomie dont j'avais besoin et qui m'ont permis à terme de m'épanouir complètement au sein de l'équipe. Pour cela, je remercie tous les membres de l'équipe, anciens et actuels : Anissa Djedid, qui a eu la patience de m'accompagner lors de mon premier stage, Véronique Lisi, avec qui j'ai beaucoup échangé et appris, Nehme El-hachem, qui a apporté sa bonne humeur. Ça a été un plaisir de travailler avec eux et rien n'aurait été pareil sans leurs implications.

Je remercie chaleureusement tous les membres du groupe Leucégène et plus particulierement Dr Guy Sauvageau qui par leur aide, leurs conseils ou leurs critiques, ont contribué à la réalisation de ce travail.

Tout cela n’aurait pas été possible sans la présence de mes amis et de ma famille. Je remercie mes parents qui ont toujours cru en moi, depuis que je suis enfant et qui m’ont permis de réaliser ce que j’ai toujours rêvé d’être. C’est grâce à eux que je suis là. Merci …

Chapitre 1 – Introduction

Le sang est un tissu en constante régénération avec des milliards de cellules produites chaque jour dans une moelle osseuse adulte humaine normale. Le processus de formation du sang se nomme hématopoïèse et se déroule dans la cavité médullaire de certains os chez les mammifères. L'hématopoïèse est un processus continue qui assure la régénération permanente des cellules sanguines à courte durée de vie. Tous les éléments du sang sont issus d'un type cellulaire unique : les cellules souches hématopoïétiques (HSC). Plusieurs étapes successives de prolifération et de différenciation de ces cellules souches multipotentes génèrent différents types de progéniteurs dont la différenciation terminale génère plusieurs types de lignées cellulaires matures détenant des fonctions spécifiques dans l'organisme. Une HSC est capable de produire deux principales populations de progéniteurs. D'une part, des cellules progénitrices myéloïdes commune pour les lignées érythrocytaires, granulocytaires et mégacaryocytaires et d'autre part, les cellules progénitrices lymphoïdes qui donneront naissance aux lymphocytes responsables des réponses spécifiques immunitaires [1].

1.1 Hématopoïèse et tumeurs malignes hématologiques

1.1.1 Définition du concept de HSC et principales caractéristiques

Les cellules souches hématopoïétiques constituent une population aux propriétés uniques et spécifiques qui les rendent différentes des autres progéniteurs ou des cellules plus matures. Par exemple, la capacité d'auto-renouvellement où les HSC peuvent maintenir leur statut initial après la division cellulaire ou suivre le chemin de la différenciation vers une cellule hématopoïétique. La division des HSC peut suivre soit une division symétrique qui donne lieu à une génération de deux HSC filles identiques ou une division asymétrique pour donner naissance à d'autres types de cellules sanguines plus matures. De plus, la migration est une autre caractéristique importante des HSC non seulement pendant les stades de développement mais aussi à l'âge adulte.

Enfin, l'apoptose est un autre élément important capable de réguler le nombre de HSC à maintenir l'homéostasie [1,2].

L'auto-renouvellement est la capacité des cellules à générer une copie d'elles-mêmes qui présente un potentiel identique ou très similaire. C'est une caractéristique principale des HSC, car d'autres populations comme les cellules progénitrices hématopoïétiques peuvent donner naissance à d'autres cellules sanguines matures pendant une période longue mais limitée. Par conséquent, la présence de HSC est cruciale pour soutenir le long terme de la génération des lignées hématopoïétiques [3]. Une autre propriété des HSC est la capacité de différenciation en lignées progénitrices et différents types de cellules hématopoïétiques matures. Une fois que les HSC se sont engagés vers la différenciation, cet état sera maintenu et l'auto-renouvellement ne pourra plus être atteint [1].

1.1.2 Hiérarchie de l'hématopoïèse

Bien qu'en 1961 la HSC a été décrite et son potentiel identifié, ce n'est qu'en 1994 que pour la première fois, un travail de recherche mené par Spangrude et collègues a montré la présence de deux de population multipotente; identifiées comme Long-Term (LT)-HSC et Short-Term (ST)-HSC. De plus, une autre population a été identifiée Multi-Potent Progenitor (MPP) cette dernière n'ayant pas la capacité d'auto-renouvellement présentée par les deux autres types [4,5].

Comme décrit précédemment, la HSC a été reconnue comme le chef d'une hiérarchie dans laquelle repose la production des différentes lignées hématopoïétiques [6]. Un schéma (Figure 1) décrit la présence des rares LT- HSC à la tête de la hiérarchie, cette population est connue pour son état de repos mais capable d'entrer dans le cycle cellulaire face à différents stimuli [7]. Au prochain échelon de la hiérarchie, les ST-HSC détenant la capacité de reconstruire le système hématopoïétique à court terme [8,9]. Cette population est à l'origine des MMP [10]. Bien que ces dernières ne présentent pas la capacité d'auto-renouvellement, sont capables de donner naissance aux progéniteurs myéloïdes communs (CMP) et aux progéniteurs lymphoïdes communs (CLP).

Les CLP sont identifiés comme des progéniteurs avec une capacité de différenciation lymphoïde limitée, tandis que les CMP peuvent se différencier en progéniteurs de mégacaryocytes/érythrocytes (MEP) et en progéniteurs de granulocytes/macrophages (GMP) [10] (Figure 1).

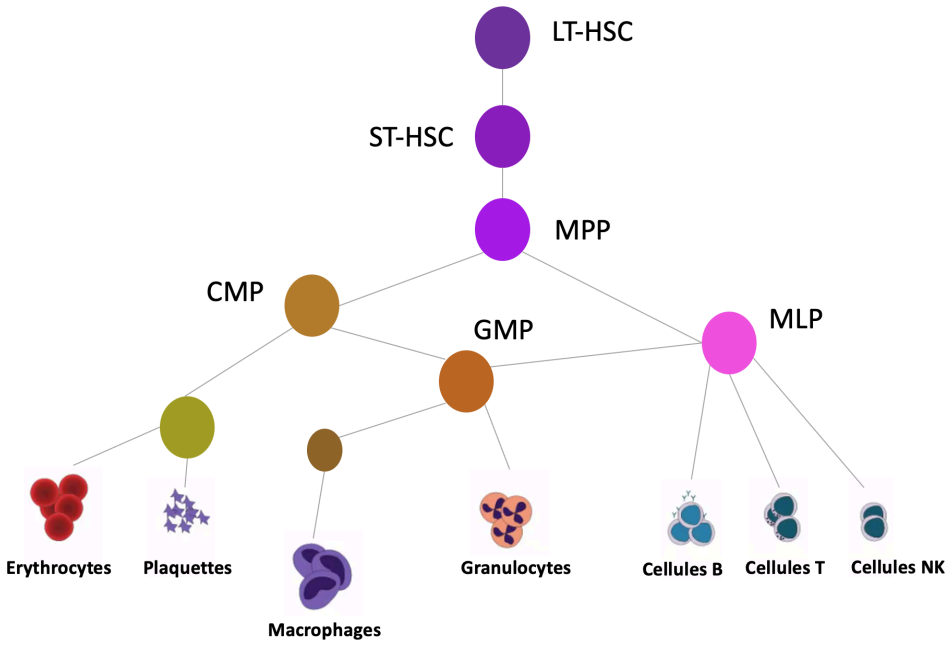

Figure 1. – Hiérarchie hématopoïétique.

Les cellules souches hématopoïétiques sont responsables de la formation de cellules sanguines fonctionnelles entièrement différenciées de manière hiérarchique. LT-HSC, Cellule souche hématopoïétique à long terme; ST-HSC, Cellule souche hématopoïétique à court terme; MPP, cellule progénitrice multipotente ; CLP, progéniteur lymphoïde commun ; CMP, progéniteur myéloïde commun ; GMP, progéniteur des granulocytes/macrophages ; MLP, progéniteur multi-lymphoïde; MEP, progéniteur mégacaryocyte/érythroïde ; NK, cellule tueuse naturelle. Figure adaptée de [11].

1.1.3 Les tumeurs malignes hématologiques

Les cancers hématologiques sont un type de cancer du sang, résultant du tissu hématogène comme la moelle osseuse ou les cellules appartenant au système immunitaire.

Ces tumeurs malignes comprennent les différentes formes de leucémie, lymphome et myélome. Chaque malignité peut être liée à plusieurs facteurs tels que l'acquisition d'une ou plusieurs mutations dans les oncogènes, les gènes suppresseurs de tumeur ou les gènes de réparation de l'ADN. En outre, des facteurs externes jouent également un rôle important dans les hémopathies malignes en tant qu'exposition à certaines substances chimiques ou pathogènes [12].

Le concept de leucémie a été établi à partir des mots grecs «leukos» et «hémie» qui indiquent la quantité élevée de globules blancs dans le corps. La leucémie est un terme général pour définir les cancers du sang. Ces désordres malins affectent la production normale de cellules sanguines saines et fonctionnelles impliqués dans l'hémostase. La population anormale de cellules sanguines est appelée « cellules leucémiques », une fois qu'elles s'accumulent dans la moelle osseuse, elles pourront conduire à une défaillance de moelle osseuse et des altérations du système immunitaire. Les cellules leucémiques présentent un état plus immature par rapport à une cellule sanguine normale, due au manque d'une activité fonctionnelle appropriée. Une classification des leucémies a été établie en fonction de la lignée touchée et de l'état de maturation de la cellule [13].

1.1.4 Leucémie myéloïde aiguë (LMA)

1.1.4.1 Généralités

Les leucémies aiguës myéloïdes (LMA) sont un groupe de cancers résultant de la différenciation anormale et incomplète des cellules souches et progénitrices hématopoïétiques (HSPC), suite à l'acquisition séquentielle de diverses anomalies génétiques et cytogénétiques. La LMA est caractérisée par une expansion clonale anormale des blastes dans la moelle osseuse et dans le sang périphérique [14,15]. La LMA est le type de leucémie aiguë le plus courant chez les adultes avec environ 80% de cas de leucémie sous cette forme. De nos jours, le taux de survie chez les jeunes s'est amélioré alors que malheureusement le pronostic chez les 65 ans reste plus faible avec un taux de mortalité de 70% dans l'année qui suit le diagnostic [15,16].

Par conséquent, le développement de thérapies plus efficaces chez les patients âgés atteints de LMA représente un défi majeur qui peut être atteint par une meilleure compréhension des mécanismes résultant de cette maladie [17].

1.1.4.2 Épidémiologie

L'incidence de LMA la plus élevée au monde peut être observée aux États-Unis, en Australie et en Europe occidentale [18]. Aux États-Unis, le nombre de cas est estimé entre trois à cinq malades pour chaque 100 000 habitants. En 2015, 20 830 nouveaux cas de LMA ont été diagnostiqué [17]. L'âge médian de la LMA est d'environ 68 ans et la prévalence est plus élevée chez les hommes que chez les femmes avec 3 cas masculins pour 2 cas féminins. En 2019, le nombre estimé de nouveaux cas de LMA aux États-Unis semble maintenir une constance autour 21 450 cas, qui ne reflète pas l'évolution observée en 2015. Au sein de ces nouveaux cas, le nombre de décès estimé approche le 50% des cas avec environ 10 920 décès. Les statistiques sont claires et reflètent le manque de méthodes efficaces de diagnostic et de traitement des patients atteints de LMA [17].

1.1.4.3 Classification

L'idée de classifier les cas de LMA est basée sur l'importance de les organiser en des groupes partageant les mêmes caractéristiques. Cette classification permet d'identifier la meilleure approche thérapeutique pour chaque pronostique et ce en fonction du type de LMA [19]. Pour la première fois en 1976, un système de classification international a été défini afin de pouvoir organiser les différents sous-types de LMA. Ce nouveau système reconnu comme la classification franco-américano-britannique (FAB), catégorise la LMA en six sous-types différents de M1 à M6, en utilisant comme référence les caractéristiques morphologiques et cytochimiques des cellules leucémiques (Tableau 1). En 2016, l'organisation mondiale de la santé (OMS) a mis à jour le nouveau système de catégorisation des différents types de leucémie. De nombreux aspects comme l'information génétique cellulaire et moléculaire, l'immunophénotype, la morphologie et les données cliniques ont été utilisés pour déterminer les principales catégories de LMA avec leurs anomalies, leurs caractéristiques liées à la myélodysplasie et leurs traitements [20,21] (Tableau 2).

Sous-type FAB	Nom
M0	LMA avec différentiation minimale
M1	LMA sans maturation
M2	LMA avec maturation
M3	Leucémie promyélocytaire aiguë (LPA)
M4	Leucémie myélocytaire aigue
M4Eo	Leucémie myélocytaire aiguë avec Éosinophilie
M5	Leucémie monocytaire aiguë
M6	Leucémie erythroblastique aiguë
M7	Leucémie megacaryocytaire aiguë

Tableau 1. – Classification FAB des LMA [22].

Description
◊ **LMA avec anomalies cytogénétiques récurrentes** LMA avec [t(8;21)(q22;q22.1)] *RUNX1-RUNX1T1*; LMA avec [inv(16)(p13.1q22)] ou [t(16;16)(p13.1;q22);] *CBFB-MYH11*; Leucémie aigüe promyélocytaire avec *PML-RARA*; LMA avec [t(9;11)(p21.3;q23.3)] *MLLT3-KMT2A*; LMA avec [t(6;9)(p23;q34.1);] *DEK-NUP214*; LMA avec [inv(3)(q21.3q26.2)] ou [t(3;3)(q21;q26.2);] *GATA2, MECOM*; LMA (mégacaryoblastique) avec [t(1;22)(p13.3;q13.3);] *RBM15-MKL1*; LMA avec BCR-ABL LMA avec NPM1 muté ; LMA avec mutation biallélique de CEBPA. LMA avec RUNX1
◊ **LMA avec anomalies associées aux myélodysplasies** • **LMA avec caryotype complexe** **Anomalies déséquilibrées** LMA avec [-7/del(7q)] LMA avec [del(5q)/t(5q)] LMA avec [i(17q)/t(17p)] LMA avec [-13/del(13q)] LMA avec [del(11q)] LMA avec [del(12p)/t(12p)] LMA avec [idic(X)(q13)] **Anomalies équilibrées** LMA avec [t(11;16)(q23;q13.3)] LMA avec [t(3;21)(q26.2;q22.1)] LMA avec [t(1;3)(p36.3;q21.2)] LMA avec [t(2;11)(p21;q23.3)] LMA avec [t(5;12)(q32;p12);] LMA avec [t(5;7)(q32;q11.2);] LMA avec [t(5;17)(q32;p13.2)] LMA avec [t(5;10)(q32;q21.1)] LMA avec [t(3;5)(q25.3;q35.1)] • **LMA, sans autre indication** LMA avec différentiation minimale LMA sans maturation LMA avec maturation LA myélomonocytaire LA monoblastique / monocytaire LA érythroïde pure LA mégacaryoblastique LMA à composante basophile LA avec myélofibrose
◊ **Leucémies aigües de lignée ambigüe** Leucémie aigüe indifférenciée Leucémie aigüe à phénotype mixte avec t(9;22)(q34.1;q11.2); *BCR-ABL1* Leucémie aigüe à phénotype mixte avec t(v;11q23.3); *KMT2A* réarrangé Leucémie aigüe à phénotype mixte B/myéloïde Leucémie aigüe à phénotype mixte T/myéloïde

Tableau 2. – Système de classification de la LMA de l'OMS (2016) [23].

1.1.4.4 Pathogénèse

Grâce aux progrès réalisés dans le domaine du séquençage de nouvelle génération, l'hétérogénéité moléculaire de la LMA a été mise en évidence et les anomalies génétiques ont suscité un intérêt croissant non seulement comme marqueurs diagnostiques, mais aussi comme facteurs pronostiques. En effet, selon les recommandations de l'European LeukemiaNet (ELN), la LMA est classée en trois groupes de risque pronostique qui prennent en compte à la fois les anomalies du caryotype et les mutations génétiques : favorable, intermédiaire et défavorable (Tabableau 3) [24].

Le dépistage des mutations *NPM1, FLT3 et CEBPA* a été introduit dans la pratique clinique de routine [25]. En effet, les mutations bialléliques du *CEBPA* et le *NPM1* muté, en présence de *FLT3-ITD* de type sauvage, sont considérés comme des facteurs de risque pronostiques favorables indépendamment des autres anomalies coexistantes [26,27], alors que les patients présentant des mutations dans les deux gènes *NPM1* et *FL3-ITD* sont classés dans le groupe à risque intermédiaire. En revanche, les mutations FLT3-ITD en l'absence de mutations *NPM1* confèrent un risque indésirable [24]. Ainsi, l'impact pronostique d'un facteur peut dépendre fortement de la présence ou de l'absence d'autres anomalies. Environ 45 % des patients atteints de LMA ont un caryotype normal avec des mutations somatiques et/ou des changements dans l'expression des gènes [28]. Par contre, la majorité des cas LMA présente des aberrations cytogénétiques comprenant des délétions, des insertions et des aneuploïdies [24]. Ces altérations peuvent entraîner un blocage de la différenciation, une apoptose altérée, un auto-renouvellement accru et une prolifération de précurseurs hématopoïétiques [29]. De plus, ces aberrations cytogénétiques identifiées dans les néoplasmes myéloïdes ont fait des progrès dans le diagnostic et le pronostic au cours des dernières décennies pour les patients atteints de LMA (Tableau 3) [5]. Parmi toutes les aberrations chromosomiques (tableau 3), les translocations t(8;21) (q22;q22.1), t(16;16)(p13.1;q22) ou inv(16)(p13.1q22) et t(15;17) sont considérés des marqueurs de bon pronostic, indépendamment de la présence d'anomalies supplémentaires.

Par contre, les aberrations chromosomiques comme t(6;9)(p23;q34.1), t(v;11q23.3), t(9;22)(q34.1;q11.2), inv(3)(q21.3q26.2) ou t(3;3)(q21.3;q26.2), la monosomie 5 ou del(5q), ainsi que la monosomie 7 et 17 ou abn(17p) sont inclus dans le groupe à risque défavorable avec les caryotypes complexes et monosomiques [30]. Selon les critères ELN, un caryotype complexe est défini par la présence de trois ou plusieurs aberrations chromosomiques, en absence de translocations ou d'inversions récurrentes, y compris t(8;21), inv(16) ou t(16;16), t(15;17), t(9;11), t(6;9), t(v;11q23.3), t(9;22), inv(3) ou t(3;3) [31]. Environ 10 à 14 % de tous les patients atteints de LMA et 23 % des patients âgés présentent un caryotype complexe avec un mauvais pronostic malgré un traitement intensif [32].

Risk Group	Abnormalities
Favorable	t(8;21)(q22;q22.1); *RUNX1-RUNX1T1* inv(16)(p13.1q22) or t(16;16)(p13.1;q22); *CBFB-MYH11* Mutated *NPM1* without *FLT3-ITD* or with *FLT3-ITD^{low}* Biallelic mutated *CEBPA*
Intermediate	Mutated *NPM1* and *FLT3-ITD^{high}* Wild-type *NPM1* without *FLT3-ITD* or with *FLT3-ITD^{low}* (without adverse-risk genetic lesions) t(9;11)(p21.3;q23.3); *MLLT3-KMT2A* Cytogenetic abnormalities not classified as favorable or adverse
Adverse	t(6;9)(p23;q34.1); *DEK-NUP214* t(v;11q23.3); *KMT2A* rearranged t(9;22)(q34.1;q11.2); *BCR-ABL1* inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); *GATA2,MECOM(EVI1)* -5 or del(5q); -7; -17/abn(17p) Complex karyotype, monosomal karyotype Wild-type *NPM1* and *FLT3-ITD^{high}* Mutated *RUNX1* Mutated *ASXL1* Mutated *TP53*

Tableau 3. – Stratification des risques de la LMA par European LeukemiaNet (ELN) 2017 [31].

t: translocation; inv: inversion; del: deletion; abn: anormale.

1.1.4.5 L’hématopoïèse clonale

Les HSC peuvent à la fois s'auto-renouveler, ce qui signifie qu'elles donnent naissance à une autre HSC au cours de la division cellulaire, et se différencier, ce qui signifie qu'elles progressent plus loin dans les lignées sanguines et s'éloignent de la multipotence [33]. Comme les HSC se divisent ou s'auto-renouvellent, des mutations somatiques normales peuvent survenir à partir d'erreurs dans la réplication de l'ADN, comme elles le font pour toutes les cellules en division. Les mutations somatiques s'accumulent à mesure qu'une personne vieillit [34]. Si ces mutations se produisent dans des gènes qui donnent à une cellule souche sanguine un avantage sur d'autres cellules souches, cette cellule souche continuera à contribuer de manière disproportionnée à la production de cellules sanguines par rapport à d'autres cellules souches sanguines et ont une plus grande taille de clone. Ce processus est connu sous le nom d'hématopoïèse clonale et augmente avec l'âge [35]. Dans les cancers hématologiques, les mutations acquises séquentielles qui sélectionnent l'expansion clonale dans le sang donneront naissance à des clones cancéreux qui peuvent finir par dominer le sang d'une personne. L'hématopoïèse clonale des HSC portant ces mutations somatiques est la première étape dans le développement de cancers hématologiques et pourrait être utilisée pour dépister les risques futurs. Dans ce cas, l'hématopoïèse clonale peut spécifiquement servir de précurseur à la LMA [17].

La progression vers la LMA a été liée aux HSC préleucémiques et aux clones fondateurs qui hébergent certaines des mutations trouvées dans les cellules leucémiques [36]. Ces mutations préleucémiques se trouvent dans des gènes qui sont également couramment mutés dans l'hématopoïèse clonale, notamment *DNMT3A, TET2 et ASXL1* [35]. Ces gènes fréquemment mutés sont impliqués dans le remodelage de la chromatine et les modifications épigénétiques [37]. Ces résultats mettent en évidence le lien entre les mutations couramment trouvées dans l'hématopoïèse clonale et la LMA agressive. L'hématopoïèse clonale est également liée à un risque plus élevé de syndromes myélodysplasiques (SMD), un précurseur courant de la LMA. Les SMD sont un groupe hétérogène de troubles hématopoïétiques clonaux caractérisés par une hématopoïèse et une cytopénie inefficace, ou une réduction du nombre de cellules sanguines matures. Environ 30 % des patients atteints de SMD développeront une LMA [38].

Cependant, il est important de noter que bien que l'hématopoïèse clonale confère un risque accru de SMD et de LMA, la plupart des personnes qui développent une hématopoïèse clonale au cours du vieillissement ne développeront jamais de SMD ou de LMA [39]. C'est pourquoi l'hématopoïèse clonale liée à l'âge est communément appelée hématopoïèse clonale à potentiel indéterminé (CHIP), car le potentiel de développement de ces clones HSC étendus en maladie n'a pas encore été clairement défini. Alors que les mutations CHIP peuvent servir de marqueur pronostique important pour le risque de maladies hématologiques comme le SMD et la LMA, il est important de comprendre ce qui motive les clones HSC porteurs de ces mutations pour se développer en maladie. Cela pourrait être la clé pour cibler ces clones étendus avant qu'ils ne se développent en cancers hématologiques et cibler les clones fondateurs responsables de la rechute.

1.2 Séquençage d'ARN sur cellule individuelle

1.2.1 La technologie RNA-seq

Avec l'introduction du séquençage à haut débit, une nouvelle méthode appelée séquençage d'ARN « bulk » (RNA-seq) a surmonté les limitations des tests de quantification d'expression précédents et a été utilisé pour la première fois en 2006 [40]. Cette technique repose sur la combinaison d'un séquençage à haut débit de librairies d'ADNc et d'outils de calcul correspondants pour analyser et quantifier l'expression génique d'un échantillon d'ARN à la fois [41]. Bien qu'il comporte ses propres défis, le RNA-seq fournit plusieurs avantages clés par rapport aux méthodes précédentes telles que les puces à ADN. Premièrement, il n'est pas limité par la connaissance a priori nécessaire des séquences [42]. En outre, le RNA-seq est non seulement capable de quantifier l'expression des gènes, mais aussi de trouver des mutations telles que les polymorphisme/variants nucléotidiques simples (SNP/SNV), les indels et même les gènes de fusion en raison de sa résolution. De plus, sa précision, sa plage de quantification et sa reproductibilité surpassent les autres méthodes a moindre coûts [43].

Avec des améliorations constantes, le RNA-seq est devenu la méthode transcriptomique dominante en 2015 [44]. L'un des désavantages du RNA-seq est que la variabilité entre cellule ne peut être détectées.

1.2.2 Les technologies scRNA-seq

Récemment, il est devenu possible d'effectuer un séquençage d'ARN sur cellule unique (scRNA-seq) et d'étudier le transcriptome de chaque cellule. Il existe de nombreuses situations où il est important de comprendre comment des types cellulaires spécifiques réagissent au développement ou aux perturbations [45]. Ceci est souvent entravé dans les analyses en « bulk » qui peuvent être affectées par les proportions inconnues de types de cellules dans un échantillon (Figure 2). Les études sur l'expression des gènes dans des types cellulaires spécifiques nécessitaient auparavant de sélectionner et d'isoler les cellules d'intérêt, ce qui les sépare des autres types cellulaires auxquels elles sont généralement associées et rend impossible l'étude des interactions entre elles [46].

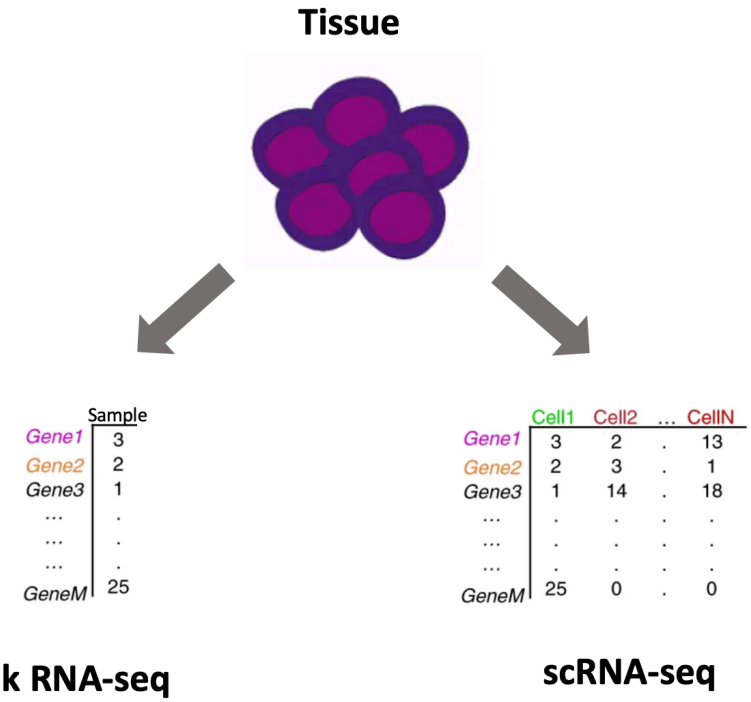

Figure 2. – La technologie scRNA-seq révèle une hétérogénéité cellulaire masquée par RNA-seq.

Avec les technologies scRNA-seq, il est désormais possible d'observer simultanément le transcriptome de tous les types cellulaires d'un tissu, ce qui a permis de mieux comprendre ce qui distingue les types cellulaires et de découvrir des types cellulaires auparavant inconnus [46].
Le premier protocole scRNA-seq a été publié en 2009 [45], juste un an après la première publication de RNA-seq (Figure 3) [47]. Alors que cette approche permettait des mesures du transcriptome dans des cellules individuelles, elle nécessitait une manipulation manuelle et se limitait à l'inspection de quelques cellules. Depuis, de nombreux protocoles scRNA-seq ont été développés (dont CEL-Seq [48], CEL-Seq2 [49], Quartz-Seq [50], Quartz-Seq2 [51] et Smart-seq2 [52]) et le nombre de cellules dans les expériences scRNA-seq a augmenté de façon exponentielle [53]. La première plate-forme de capture cellulaire disponible commercialement était le Fluidigm C1. Ce système utilise la microfluidique pour séparer facilement les cellules dans des puits individuels sur une plaque où elles sont lysées, transcrites en ADNc collecté puis, amplifié par PCR.

Après cette étape, le produit est extrait de la plaque et des librairies préparées pour le séquençage Illumina. La plupart des données Fluidigm C1 ont été produites à l'aide d'une plaque à 96 puits, mais plus récemment, une plaque à 800 puits est devenue disponible, augmentant considérablement le nombre de cellules pouvant être capturées à la fois. L'un des inconvénients des technologies de capture cellulaire à base de plaques est que les puces utilisées ont une fenêtre de taille fixe, ce qui signifie que seules les cellules d'une taille particulière peuvent être capturées en un seul passage. Cependant, comme les cellules sont capturées dans des puits individuels, elles peuvent être imagées avant la lyse [54].

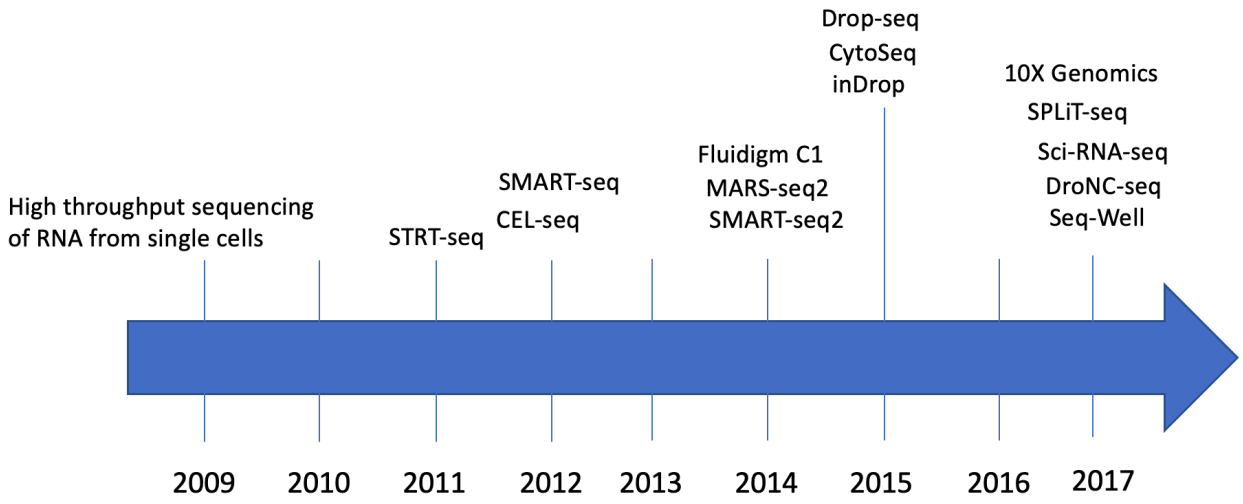

Figure 3. – Évolution des technologies scRNA-seq adaptée de [53].

1.2.3 Capture cellulaire à base de gouttelettes

Une alternative à l'utilisation des plaques de capture de cellules consiste à les capturer dans des nano-gouttelettes. Au cours de ce processus, un mélange cellulaire dissocié est mis dans un dispositif microfluidique, tandis que les billes revêtues d'amorces entrent par un autre canal. Le dispositif est conçu pour former des gouttelettes aqueuses dans l'huile minérale. Tous les composants sont introduits de manière à ce que les cellules et les billes puissent être simultanément capturées dans une gouttelette. Ensuite, les réactifs transportés avec la bille lysent la cellule et toutes les molécules d'ARN ayant une queue poly(A) peuvent se lier aux sondes de capture sur la bille. Suite à la transcription inverse et l'amplification PCR, une librairie d'ADNc individuelle est produite pour chaque cellule, étiquetée avec la séquence de code-barres unique présente sur la bille. Le principal avantage de cette technologie est la capacité de capturer beaucoup plus de cellules à la fois, plusieurs milliers. Cette technologie est également moins sélective quant à la taille des cellules et produit moins de doublets.

De plus, les goulettes contiennent un petit volume ce qui réduit le cout par cellule par rapport aux approches précédentes. La capture basée sur les gouttelettes a été popularisée par la publication des plateformes Drop-seq [54] et InDrop [55] en 2015 et la mise à jour d'InDrops en 2017 [56]. Une plate-forme semblable est commercialisée sous le nom *Chromium* par la compagnie *10X Genomics* qui automatise une grande partie du processus (Figure 4) [57].

Cette machine utilise des technologies basées sur les gouttelettes pour une gamme d'applications, y compris la capture de cellules pour scRNA-seq.

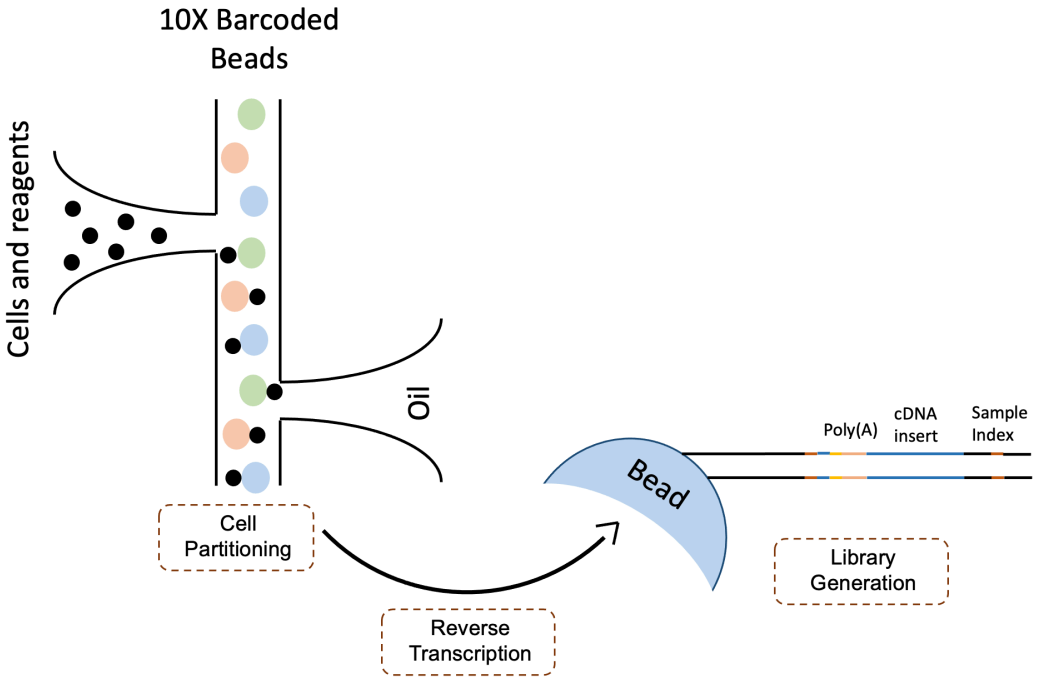

Figure 4. – Schéma du processus de capture cellulaire *10X Genomics* adapté de [58].

Les billes à code-barres circulent dans un dispositif microfluidique avec les cellules dissociées où elles sont capturées dans des gouttelettes aqueuses d'une solution d'huile. Les cellules sont lysées dans les gouttelettes et l'ARNm est transcrit en sens inverse en ADNc. Les gouttelettes sont ensuite brisées et l'ADNc est récolté pour le séquençage.

1.2.4 Identifiants moléculaires uniques

Les méthodes de capture basées sur des gouttelettes utilisent généralement des protocoles qui incluent de courtes séquences nucléotidiques aléatoires connues sous le nom d'identifiants moléculaires uniques (UMI) [59]. Pour obtenir suffisamment d'ADNc pour le séquençage, une étape d'amplification par PCR est nécessaire [59]. En fonction de leur séquence, différents transcrits sont amplifiés à des vitesses différentes, ce qui peut fausser leurs proportions relatives. Les UMI améliorent la quantification de l'expression des gènes en éliminant des doublons PCR produits lors de l'amplification (Figure 5) [60]. Les sondes nucléotidiques utilisées dans les protocoles de capture à base de gouttelettes comprennent : une séquence poly(T) qui se lie aux

molécules d'ARNm matures, une séquence code-barres qui est la même pour chaque sonde et une séquence UMI unique à chaque sonde. Les séquences UMI sont suffisamment longues (8 à 10 bases) pour que la probabilité de capturer deux copies d'un transcrit sur deux sondes avec le même UMI soit extrêmement faible. Après la transcription inverse, l'amplification, le séquençage et l'alignement, la déduplication est effectuée en identifiant les « reads » avec le même UMI qui devraient être des doublons PCR plutôt que des copies exprimées d'un transcrit [59].

La majorité des méthodes scRNA-seq utilisent 3'-tag RNA-seq, qui amplifie et séquence uniquement un fragment de la région 3' de chaque transcrit [61,62]. Le séquençage sur cellule unique 3'-tag réduit le nombre de « reads » requis par échantillon et sert ainsi d'alternative peu coûteuse par rapport au scRNA-seq complet [63,62]. Les techniques scRNA-seq actuelles sont toujours confrontées à de multiples limitations dans la détection de transcrits de faible abondance [63,61]. La technologie *10X Genomics Chromium* peut détecter en moyenne 4 500 gènes et environ 20 000 transcrits dans leur chimie V2, ce qui représente environ 14 à 15 % de tous les transcrits dans une cellule. La chimie V3, de la même technologie, détecte jusqu'à 32 % de tous les transcrits par cellule, ce qui devrait améliorer la sensibilité de ces méthodes [58,64].

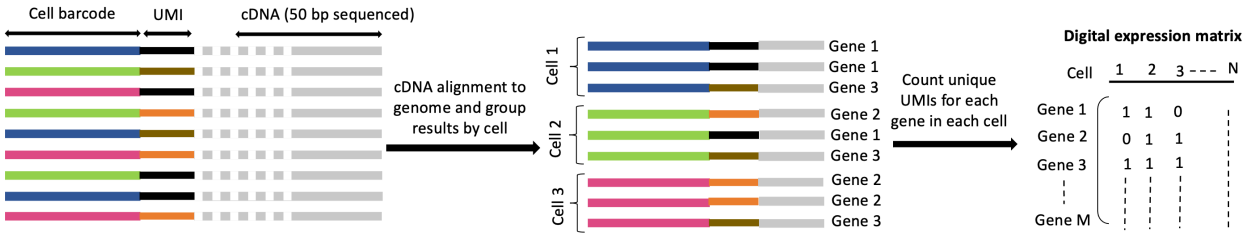

Figure 5. – Les identifiants moléculaires uniques (UMI).

Les UMI sont des séquences aléatoires de 8 à 10 pb incluses dans la sonde de capture d'ARNm avec le code-barres cellulaireet. La séquence d'ARNm est alignée sur un génome de référence. Chaque UMI n'est compté qu'une seule fois à chaque emplacement.

1.2.5 Analyse bioinformatique des données scRNA-seq

1.2.5.1 Aperçu de l'analyse

L'analyse bioinformatique des données scRNA-seq se fait en deux étapes principales. La première partie de l'analyse consiste à l'étape de prétraitement des données. Cette étape commence par l'association des « reads » avec leurs cellules d'origine, l'alignement des séquences sur le génome de référence et la quantification des transcrits. Suite à un contrôle qualité, les cellules et les « reads » de bonne qualité sont comptabilisées dans la matrice de compte en UMI. Quelques méthodes de calcul sont développées pour gérer la matrice de compte UMI à partir des données scRNA-seq brutes [65]. Parmi ces méthodes, Cell Ranger [57] est la méthode la plus couramment utilisé. Par la suite, d'autres analyses en aval sont effectuées pour répondre à des questions biologiques spécifiques [62]. Ces analyses secondaires comportent une autre étape de contrôle qualité, suivis de la normalisation, la réduction des dimensions, le regroupement et en fin l'analyse différentielle de l'expression génique [62,65].

Cependant, en raison du nombre important de données et d'un niveau de variance plus élevé, plusieurs packages d'analyse ont été développées tel que Seurat [66], Scater [67] et Scanpy [68].

1.2.5.2 Contrôle qualité

Plusieurs étapes de filtrage sont d'abord appliquées aux données de comptage afin de garantir que les « reads » de mauvaise qualité et les cellules endommagées sont éliminées des analyses ultérieures. Lors du comptage des UMI, Cell Ranger ne prend en compte que les « reads » qui ont un UMI et un code-barre valides. De plus, Cell Ranger prend en compte les « reads » qui correspondent exactement à un gène.

Dans la première étape du contrôle qualité, les cellules avec une fraction d'ARN mitochondrial anormalement élevée sont filtrés, ceci indique que ces cellules sont endommagées [69]. Les doublets, qui sont des échantillons résultant de la capture et du séquençage de deux cellules ou plus dans la même gouttelette, sont recherchées [69]. Ces cellules aberrantes doivent être éliminées pour éviter le biais des analyses en aval. La fréquence des doublets augmente avec le nombre de cellules encapsulées. Ces cellules peuvent être aussi identifiées avec certains algorithmes.

1.2.5.3 Réduction de la dimensionnalité

Les données scRNA-seq sont de grande dimension, chaque gène représente une dimension le long de laquelle les cellules peuvent varier. Il est difficile de visualiser de manière significative tous les points de données dans un espace de dimension aussi élevée. Ainsi, une méthode de réduction dimensionnelle (RD) est souvent nécessaire pour visualiser et analyser les données scRNA-seq dans un espace de dimension inférieure [46]. L'analyse en composantes principales (ACP), est l'une des méthodes RD les plus populaires, basée sur des combinaisons linéaires de variables (les gènes), qui expliquent la plus grande quantité de variance des données [70]. Chaque combinaison linéaire représente une composante principale sur laquelle les cellules peuvent être projetées. Même si l'ACP réussit bien à identifier les dimensions qui contribuent le plus à la variance, en tant qu'outil de transformation non supervisée et linéaire, cette approche ignore les informations des *clusters* et ne parvient pas à détecter la relation non linéaire entre les cellules [46].

D'autres méthodes RD ont été développées pour surmonter ce problème, notamment le t-Stochastic Neighbor Embedding (t-SNE) [71] et l'Approximation et la cartographie uniforme (UMAP) [72]. Le t-SNE cartographie les points de données, c'est-à-dire les cellules dans les données scRNA-seq, d'un espace de grande dimension à un espace à 2 dimensions afin que la proximité des points de données dans les deux espaces soit préservée [71,72]. La préservation de la structure locale rend tSNE utile pour visualiser les *clusters* de cellules, même si l'algorithme lui-même n'est pas conçu pour le regroupement. tSNE est toujours limité par sa capacité à cartographier un grand nombre de cellules (> 30 000 cellules) et son compromis sur la préservation de la structure globale [71]. UMAP est apparue comme une nouvelle méthode de RD capable de préserver à la fois la structure de données globale et locale dans la visualisation en 2D pour déduire la similitude entre les *clusters* par leur distance [72]. Ces méthodes ont permis d'améliorer considérablement la visualisation des ensembles de données scRNA-seq (Figure 6).

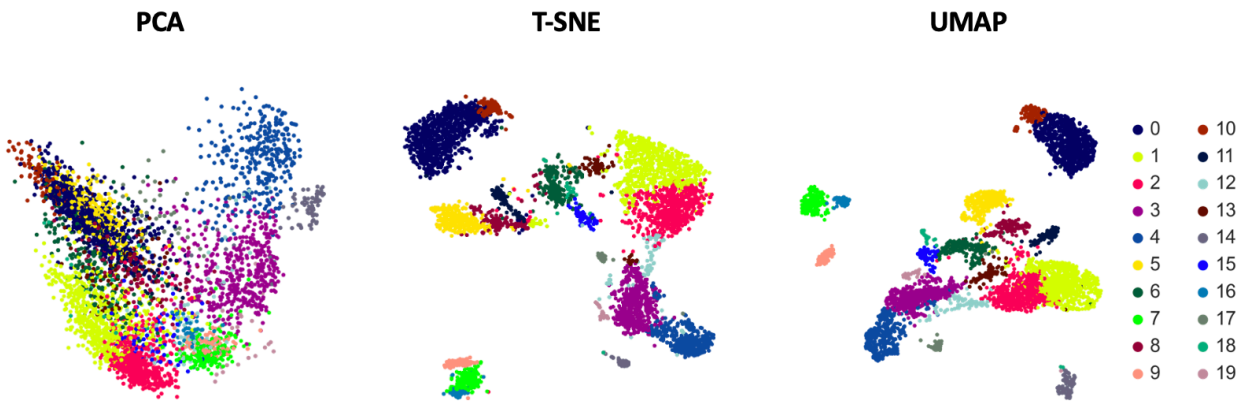

Figure 6. – Visualisation de 5 000 cellules.

Regroupement en utilisant les deux premières composantes d'ACP (ou PCA en anglais) (à gauche), t-SNE (au milieu) et UMAP (à droite). Chaque point représente un point de données. L'ACP ne discerne pas les *clusters* de points dans l'ensemble de données contrairement au tSNE et UMAP. Les *clusters* dans le graphique t-SNE sont proches les uns des autres et les distances par paires entre eux ne signifient pas nécessairement à quel point ils sont différents. UMAP préserve mieux la structure globale dans l'ensemble de données mais il reste encore difficile à interpréter les distances.

1.2.5.4 Normalisation

Les données scRNA-seq sont différentes des données RNA-seq en « bulk » en terme de la rareté et de la variabilité élevées [63]. Une grande proportion du nombre de « reads » scRNA-seq est nulle, ce qui se produit à la fois en raison de la faible sensibilité des techniques de séquençage et du manque d'expression génique dans certaines populations cellulaires [61]. De plus, les mesures d'expression génique sont affectées par des biais systématiques, tels que l'efficacité de la capture, la transcription inverse et le nombre de «reads». Les comptes en UMI bruts ne sont donc pas à la même échelle entre les cellules et ne peuvent pas être directement comparés. L'objectif des méthodes de normalisation est de ramener toutes les mesures d'expression génique à une échelle commune en supprimant les biais spécifiques à la cellule [73]. Ce processus consiste à diviser les UMI de chaque gène par le nombre d'UMI totale dans cette cellule. Cette méthodes surpassent les autres pour estimer les vrais facteurs d'échelle spécifiques aux cellules dans les ensembles de données scRNA-seq [68].

1.2.5.5 Regroupement

L'une des principales applications du scRNA-seq est de caractériser l'hétérogénéité dans un échantillon biologique, ce qui implique la détection des sous-populations cellulaires distinctes. Ces sous-populations peuvent représenter des types cellulaires qui ont été précédemment étudiés dans l'ontologie cellulaire [74]. Le regroupement de cellules similaires en sous-populations est connu sous le nom « *clustering* » et a été résolu avec un groupe diversifié d'algorithmes, notamment le *clustering* k-means, hiérarchique et basé sur des graphes [75]. Pour réduire les ressources de calcul requises, ces méthodes de *clustering* sont généralement exécutées en aval de l'ACP, de sorte que toutes les distances calculées sont dans un espace de dimension réduite. Le *clustering* K-means, implémenté dans SC3 [76] et RaceID [77], initie k centres de *cluster* et attribue de manière itérative les cellules au centre de *cluster* le plus proche jusqu'à ce que les distances entre les centres de *cluster* et leurs membres de *cluster* respectifs soient minimisées. Au final les centres sont mis à jour par la moyenne.
Le *clustering* hiérarchique, implémenté dans Mpath [78] et BackSPIN [79], divise séquentiellement les cellules en *clusters* plus petits ou ajoute des cellules pour former des *clusters* plus grands en fonction des distances entre les cellules.

Ces deux algorithmes ont une complexité de calcul élevée et ne sont donc pas adéquats pour les grands ensembles de données scRNA-seq. Cependant, les algorithmes de *clustering* basés sur des graphes fonctionnent souvent mieux en termes de vitesse. PhenoGraph [80] fournit une méthode, basée sur des graphes, efficace en termes de calcul pour identifier des sous-populations dans des données scRNA-seq de grande dimension. L'algorithme représente la population cellulaire sous-jacente par un réseau dans lequel chaque cellule est connectée à d'autres cellules qui sont phénotypiquement similaires. Ces graphes sont ensuite regroupés en communautés phénotypiques distinctes, en utilisant l'optimisation de la modularité [80]. PhenoGraph est particulièrement puissant pour les grands échantillons où il produit des résultats de haute qualité sans sous-échantillonnage.

1.2.5.6 Analyse de l'expression différentielle des gènes

Une analyse typique des données scRNA-seq consiste à identifier les gènes qui sont exprimés de manière différentielle (GDE) dans chaque sous-population, ce qui peut expliquer des processus biologiques dans certains états cellulaires ou déterminer le devenir de certains types cellulaires. Des méthodes GDE, développées pour analyser les données RNA-seq (ex. edgeR [81] et DESeq [82]) ont été adapté scRNA-seq. Au même temps, de nombreux outils spécifiques à cette technologie ont été développés pour tenir compte des fonctionnalités scRNA-seq, notamment SCDE [83], MAST [84] et autres [85]. Ces méthodes partagent le cadre commun de la modélisation du nombre de « reads » ou des valeurs d'expression avec une distribution de probabilité, puis l'utilisation d'un test statistique avec ajustement de la limite de signifiance pour identifier les GDEs et leur signification [86]. EdgeR et DESeq modélisent l'expression des gènes en utilisant la distribution binomiale négative et utilisent un test exact pour déterminer les GDEs [81]. SCDE [83] tient compte des abandons (lorsqu'un gène est observé à un niveau d'expression faible ou modéré dans une cellule mais n'est pas détecté dans une autre cellule du même type cellulaire) dans les données scRNA-seq en modélisant l'expression génique à l'aide d'un modèle probabiliste mixte. MAST [84] adapte un modèle linéaire généralisé en deux parties pour modéliser le taux et les niveaux d'expression d'un gène individuel, puis utilise un test de probabilité pour les GDEs.

Outre ces méthodes, des tests statistiques non paramétriques généraux, tels que le test de somme des rangs de Wilcoxon, ont également été utilisés pour tester les GDEs, avec une précision comparable aux méthodes mentionnées, mais avec une complexité de calcul plus faible et une parallélisation plus facile, ce qui entraîne une vitesse de calcul plus élevée [86].

1.2.5.7 Analyse d'enrichissement des voies

Avec la longue liste de gènes générés par l'analyse GDEs vient le défi d'interpréter leurs fonctions. L'analyse d'enrichissement des voies (AEP) apporte une solution à ce problème en identifiant des voies enrichies, qui sont des ensembles de gènes qui fonctionnent ensemble pour réaliser certaines fonctions biologiques dans les cellules, à partir d'une liste donnée de GDEs [87]. L'AEP nécessite une base de données de voies composée d'ensembles de gènes annotés qui ont été

sélectionnés à partir de la littérature, par ex. Gene Ontology [88], Reactome [89] et KEGG [90]. L'enrichissement des voies peut être évalué par l'analyse de surreprésentation, qui identifie les ensembles de gènes contenant plus de gènes dans une liste de GDEs que prévu par hasard et teste leur signification à l'aide de tests statistiques tels que le chi carré ou le test exact de Fisher. Dans certains outils, la puissance statistique repose sur la notation de classe fonctionnelle, qui calcule un score d'enrichissement pour chaque ensemble de gènes basé sur la position du gène dans une liste classée de GDEs [91]. Cette approche s'est avérée plus robuste pour détecter des changements plus subtils dans l'enrichissement des voies et est mise en œuvre dans deux méthodes AEP populaires, Gene Set Enrichment Analysis (GSEA) [92] et Gene Set Variation Analysis (GSVA) [93]. Le résultat de ces méthodes est une liste de voies enrichies et leurs statistiques associées. Les résultats de l'enrichissement peuvent ensuite être visualisés sous forme de réseaux avec des logiciels tels que Cytoscape [87] et EnrichmentMap [94], ce qui est utile pour identifier des voies étroitement liées qui sont enrichies dans une population et déterminer des thèmes communs dans les voies enrichies [95,87].

1.3.6 HCA

L'Atlas des cellules humaines ou Human Cell Atlas (HCA) est une base de données qui catalogue des cartes de référence moléculaires détaillées de toutes les populations cellulaires saines du corps humain [96]. Avec l'émergence de la technologie scRNA-seq, les efforts actuels se sont concentrés sur les analyses transcriptomiques en cellules uniques des dizaines de milliers de cellules provenant de plusieurs donneurs sains. En 2018, HCA a publié le premier jeu de données scRNA-seq de 100 000 cellules provenant des moelles osseuses saines de huit donneurs sains [97]. Ces données présentent 35 populations de cellules transcriptionnellement cohérentes associées à diverses lignées cellulaires préalablement définies, des progéniteurs CD34+, des états de maturation intermédiaires et des types cellulaires différenciés. Ces données présentent un exemple de la trajectoire cellulaire de l'hématopoïèse normale. L'exploration de ces données est essentielle dans le contexte de LMA afin de permettre des découvertes translationnelles.

1.3.7 Le scRNA-seq dans la LMA

La technologie de séquençage en cellules uniques a trouvé de nombreuses applications dans l'étude de l'hétérogénéité cellulaire de plusieurs cancers. Dans le contexte de LMA, L'hétérogénéité intratumorale est appréciée depuis les années 1960, mais ce n'est que récemment qu'il est devenu possible d'étudier cette complexité à l'aide des analyses en cellules uniques [80,57,98,99]. L'hétérogénéité transcriptionnelle des échantillons LMA proviennent clairement des états de différenciation des cellules normales et tumorales, des états du cycle cellulaire et des anomalies génétiques. Petti et collègues ont pu intégrer la détection des variants dans l'analyse scRNA-seq pour révéler des corrélations entre l'hétérogénéité mutationnelle et transcriptionnelle [100]. L'équipe a montré que des SNV étaient détectables dans 22,7% des cellules de leurs échantillons. Avec cette approche, ils étaient capables de distinguer les cellules tumorales des cellules normales dans des données scRNA-seq de LMA. De plus, Une autre étude a démontré que les données scRNA-seq a transcription complète sont déterminantes pour détecter les CNV dans des cellules individuelles.

Cependant, la plupart de leurs échantillons LMA ne présentait pas des CNV et leur jeu de données et ne représente pas la diversité génétique des LMA [100]. Alors que les études sur la LMA se concentrent souvent sur les types cellulaires immatures, il a été démontré que les cellules malignes différenciées contribuent également à la biologie de la LMA [101]. Des résultats récents donnent un aperçu des programmes régulateurs aberrants des cellules LMA primitives, révèlent une correspondance frappante entre les hiérarchies de développement et la génétique tumorale. Ces analyses ont permis d'identifier des cellules LMA différenciées avec des propriétés immunosuppressives [101].

Aucune étude n'a exploré l'hétérogénéité génétique de la LMA à partir des données scRNA-seq en combinant l'information relative à l'expression avec les CNV et les SNV. Des études supplémentaires de scRNA-seq d'échantillons de tumeurs primaires de LMA sont nécessaires pour étudier cette hétérogénéité complexe et identifier des sous-clones LMA. La caractérisation de ces derniers est une étape importante pour la réponse au traitement, la rechute et le développement de thérapies de précision spécifiques au génotype [102].

L'identification des types cellulaires à partir de matrices en cellules uniques d'expression de gènes est un problème particulièrement difficile, étant donné que les jeux de données ne sont pas bien annotés. Le *clustering* non-supervisé est donc largement appliqué dans l'analyse des données scRNA-seq afin de découvrir des modèles ayant une signification biologique et des groupements possibles de cellules dans des *clusters* spécifiques qui partagent des profils d'expression similaires [103].

La dernière décennie a été témoin d'une augmentation exponentielle de la taille de l'ensemble de données scRNA-seq, alors que le coût par cellule continue de diminuer. Par exemple, le Human Cell Project (HCA) [104] vise à caractériser la carte en cellules uniques de toutes les cellules humaines, et son ordre de grandeur atteindra des milliards. Face à une telle croissance explosive des données, l'un des principaux défis est l'identification fiable et rapide du type de cellule à partir d'une cellule nouvellement séquencée.

L'annotation de type de cellule supervisée des données nouvellement générées à l'aide d'étiquettes annotées est devenue plus souhaitable que les approches non supervisées, car les approches non supervisées ont tendance à être beaucoup plus laborieuses et intensives en calculs. Les méthodes de classification telles que le classificateur de forêt aléatoire (RF) [105] ont montré un potentiel pour une identification précise, rapide et robuste d'une seule cellule. Une telle pratique peut devenir inefficace si le jeu de données de référence n'est pas adéquat. Dans cette section, je présente une brève revue de l'application des méthodes d'apprentissage automatique qui pourraient être utiliser dans l'annotation des données scRNA-seq.

1.4 Apprentissage automatique

L'apprentissage automatique est une branche de l'intelligence artificielle qui fournit diverses méthodes et algorithmes entraînés sur des exemples (soit un couple (entrée, sortie)), et un modèle en est extrait. Par la suite, ce modèle est testé sur un ensemble différent d'exemples, puis les performances de l'algorithme sont mesurées [106]. Dans ce domaine, les tâches d'apprentissage sont souvent caractérisées en termes de retour d'information au modèle d'apprentissage. Il existe trois types d'apprentissage: supervisé, non supervisé et semi-supervisé [106].

Dans l'apprentissage non supervisé, seuls les échantillons sont donnés, sans les étiquettes de classe. L'apprentissage semi-supervisé utilise la connaissance des étiquettes de classe d'apprentissage supervisé ainsi qu'une méthode non supervisée pour regrouper des données similaires. Dans l'apprentissage supervisé, les échantillons étiquetés sont transmis à l'algorithme de classification, qui crée le modèle prédictif [106]. L'apprentissage supervisé est le processus d'apprentissage qui peut être envisagé comme un enseignant supervisant le processus d'apprentissage car les bonnes réponses sont bien connues [106]. L'algorithme fait de manière itérative des prédictions sur les données d'entraînement et est corrigé par l'enseignant. L'apprentissage s'arrête lorsque l'algorithme atteint un niveau de performance acceptable. Pour former l'algorithme d'apprentissage, un ensemble de données d'entrainement étiquetés est utilisé dans l'apprentissage supervisé. L'objectif principal d'une approche d'apprentissage supervisé est de prédire une variable de sortie pour un ensemble de données test sur la base des connaissances acquises auprès de l'ensemble d'apprentissage pour lequel la valeur de sortie est fournie (Figure 7) [106].

Grâce à la famille de méthodes d'apprentissage supervisé, nous pouvons davantage différencier les méthodes de classification, qui se concentrent sur la prédiction de sorties discrètes, et les méthodes de régression, qui prédisent des sorties continues. Les méthodes de régression dépassent le cadre de notre étude.

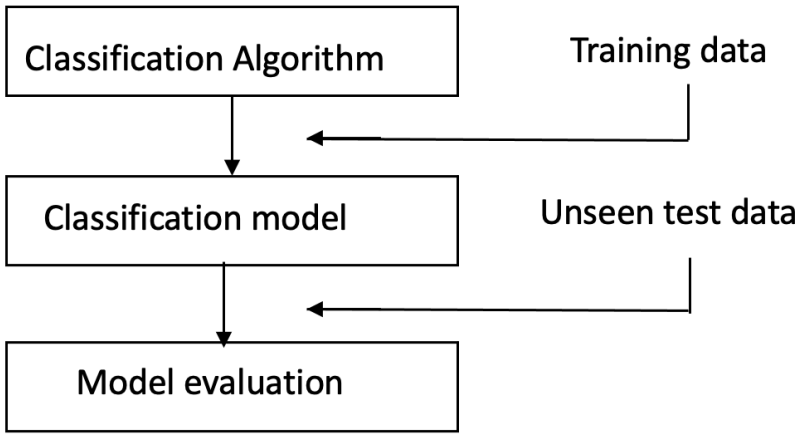

Figure 7. – Un processus général d'apprentissage supervisé pour la classification.

Il existe de nombreux algorithmes qui ont été conçus pour travailler sur des problèmes de classification. Dans cette thèse, nous utilisons des algorithmes de classification multi- classe: forêt aléatoire (RF), arbre de décision et KNeighbors.

1.4.1 Classification multi-classe

Dans la classification multi-classes, il y a plus de deux classes. Il y a plusieurs façons de résoudre ce problème. L'approche courante est : le un contre tous [107]. Dans ce cas, chaque classificateur est entraîné et testé sur une classe par rapport aux autres classes. Si l'ensemble de données a *r* classes, alors *r* classificateurs sont construits dans ce cas. Une classe est attribuée aux nouveaux échantillons par le classificateur qui produit le score de confiance le plus élevé [107]. Tous les classificateurs résolvent le problème d'un contre tous de cette manière [107]. Dans notre cas, les classes *r* correspondent aux différents types cellulaires. Les échantillons sont des cellules et les l'expression de chaque gène est considérée comme une caractéristique ou un attribut.

1.4.1.1 Méthode des k plus proches voisins

L'algorithme des k plus proches voisins (KNN) est le plus simple parmi tous les algorithmes d'apprentissage automatique. Cover et Hart [108] montrent que l'erreur de la règle du plus proche voisin est bornée au-dessus par le double de l'erreur de Bayes (taux d'erreur de classificateur connaissant le modèle qui a généré des données) sous certaines hypothèses raisonnables. De plus, l'erreur de la méthode générale KNN se rapproche asymptotiquement de l'erreur de Bayes et peut être utilisé pour l'approximer. Donc, leur idée est de mémoriser l'apprentissage puis prédire l'étiquette de toute nouvelle instance sur la base des étiquettes de ses plus proches voisins dans l'ensemble de formation [108]. La justification de cette méthode repose sur l'hypothèse que les caractéristiques, utilisées pour décrire les points, sont pertinents pour leur étiquetage d'une manière qui rend la proximité des points susceptibles d'avoir la même étiquette [108]. De plus, dans certaines situations, même lorsque l'ensemble d'apprentissage est immense, trouver un voisin le plus proche peut se faire extrêmement rapidement [109]. L'algorithme calcule la distance entre les points où k est le nombre de points de données les plus proches du point qui doit être affecté à une classe [109].

1.4.1.2 Arbre de décision

Un arbre de décision (DT) est un algorithme d'apprentissage supervisé basé sur l'algorithme de Quinlan utilisé pour la classification. L'algorithme d'arbre de décision construit un arbre avec un nœud racine et des branches [110]. Le nœud racine est sélectionné sur la base de la valeur de gain d'informations. Tout d'abord, les entropies des classes sont calculées, puis l'entropie de chaque caractéristique est calculée. Le gain d'information est la différence entre l'entropie des classes et des caractéristiques. L'attribut avec le gain d'informations le plus élevé agit comme un nœud racine, et chaque nœud est construit sur la base de la valeur de gain d'informations [110]. L'arbre est autorisé à croître de cette manière. Enfin, les motifs sont induits en partant de la racine, en prenant une décision à chaque nœud, en suivant une branche à chaque étape, se terminant par un nœud feuille qui correspond à une certaine classe. L'un des avantages de l'arbre de décision est qu'il est très facile à comprendre [110].

1.4.1.3 Forêts aléatoires

Les forêts aléatoires (RF) représentent un algorithme d'apprentissage automatique qui sert principalement à résoudre des problèmes de classification et de régression. Leur principe consiste à construire une multitude d'arbres de décision lors de leur formation dépendamment du paramétrage indiqué. Pour les tâches de classification, le résultat proposé par RF est la classe choisie par la majorité des arbres. Les RF corrigent l'habitude des arbres de décision de s'adapter à leur ensemble d'apprentissage [107]. Le premier algorithme pour les forêts de décision aléatoire a été créé en 1995 par Tin Kam Ho [111]. Une extension de l'algorithme a été développée par Leo Breiman et Adele Cutler qui ont enregistré «Random Forests» en tant que marque en 2006 [112]. L'extension combine l'idée de «*bagging*» et la sélection aléatoire de caractéristiques afin de construire une collection d'arbres de décision à variance contrôlée. Le RF est un moyen de faire la moyenne de plusieurs arbres de décision profonds, entraînés sur différentes parties du même ensemble d'entraînement, dans le but de réduire la variance. Cela se fait au détriment d'une légère augmentation du biais et d'une certaine perte d'interprétabilité, mais augmente considérablement les performances du modèle final [111].

Il existe trois types d'ensembles de données dans l'approche de «*bagging*» [113]:

- L'ensemble de données d'origine est toute information donnée à l'algorithme pendant la phase d'entrainement;
- L'ensemble de données *«bootstrap»* est créé en choisissant au hasard des objets dans l'ensemble de données d'origine. En outre, il doit être de la même taille que l'ensemble de données d'origine. Cependant, la différence est que l'ensemble de données *«bootstrap»* peut avoir des objets en double;
- L'ensemble de données *out-of-bag* représente les objets restants qui ne sont pas dans l'ensemble de données *«bootstrap»*. Il peut être calculé en prenant la différence entre les ensembles de données d'origine et *«bootstrap»*.

La création de ces ensembles de données est cruciale car elle peut être utilisée pour tester la précision d'un algorithme RF. Étant donné que l'algorithme génère plusieurs arbres (entre 100 et 1000) et donc plusieurs ensembles de données, le risque qu'un objet soit exclu de l'ensemble de données *«bootstrap»* est faible [113].

L'étape suivante de l'algorithme implique la génération d'arbres de décision à partir de l'ensemble de données *«bootstrap»* [114]. Pour y parvenir, le processus examine chaque caractéristique et détermine pour combien d'échantillon la présence ou l'absence de cette caractéristique donne un résultat positif ou négatif. Ces informations sont ensuite utilisées pour calculer une matrice de confusion, qui répertorie les vrais positifs, les faux positifs, les vrais négatifs et les faux négatifs. Ces caractéristiques sont ensuite classées selon diverses métriques de classification en fonction de leurs matrices de confusion. Ensuite, elles sont utilisées pour diviser les échantillons en deux ensembles : ceux qui contiennent la caractéristique et ceux qui ne la contiennent pas. Ce processus est répété de manière récursive pour les niveaux successifs de l'arbre jusqu'à ce que la profondeur souhaitée soit atteinte. Tout en bas de l'arbre, les échantillons dont le test est positif pour la caractéristique finale sont généralement classés comme positifs. Ces arbres sont ensuite utilisés comme prédicteurs pour classer les nouvelles données [114].

La prochaine étape de l'algorithme consiste à introduire un autre élément de variabilité parmi les arbres *«bootstrap»*. En plus de chaque arbre examinant uniquement un ensemble d'échantillons *«bootstrap»*, seul un nombre restreint de caractéristique est pris en compte lors de leur classement en tant que classificateurs. Cela signifie que chaque arbre ne connaît que les données relatives à un petit nombre constant d'entités et à un nombre variable d'échantillon inférieur ou égal à celui de l'ensemble de données d'origine [114]. Par conséquent, les arbres sont plus susceptibles de renvoyer plus de réponses, dérivées de connaissances plus diverses. La procédure de formation pour les forêts aléatoires permet naturellement de calculer certaines métriques utiles tout au long du processus de l'entrainement [115] :

- L'erreur «*out of bag»* (OOB) est l'erreur de prédiction moyenne de chaque ensemble d'apprentissage qui ne figure pas dans l'ensemble *«bootstrap»*. Cela permet au RF d'être adapté et validé à l'ajout de chaque nouvel arbre pendant l'entraînement.
- L'importance des caractéristiques : ou la mesure de l'importance de chaque caractéristique. Les valeurs de chaque caractéristique peuvent être aléatoirement permutées. Lorsque l'erreur OOB est recalculée, la différence de l'erreur OOB résultante des permutations est calculée.

La moyenne de toutes les différences pour la caractéristique, normalisée par l'écart type, forme le classement final [112]. Cela donne une indication globale de l'importance d'une caractéristique en question [116].

La figure 8 montre comment fonctionne le modèle RF. Chaque arbre de décision de la forêt aléatoire prédit la classe indépendamment. Chaque arbre vote à quelle classe appartient chaque échantillon [107]. Le nombre total de votes est calculé et la classe majoritaire est attribuée à cette cellule. Dans la figure 8, l'arbre de décision 1 et l'arbre de décision 2 ont voté pour la classe + ; par conséquent, la classe + est attribuée à l'échantillon.

L’algorithme de forêt aléatoire est très rapide et atteint généralement une très bonne précision pour les grands ensembles de données [107]. Pour ces raisons, la forêt aléatoire a été sélectionnée comme l'un de nos algorithmes de classification.

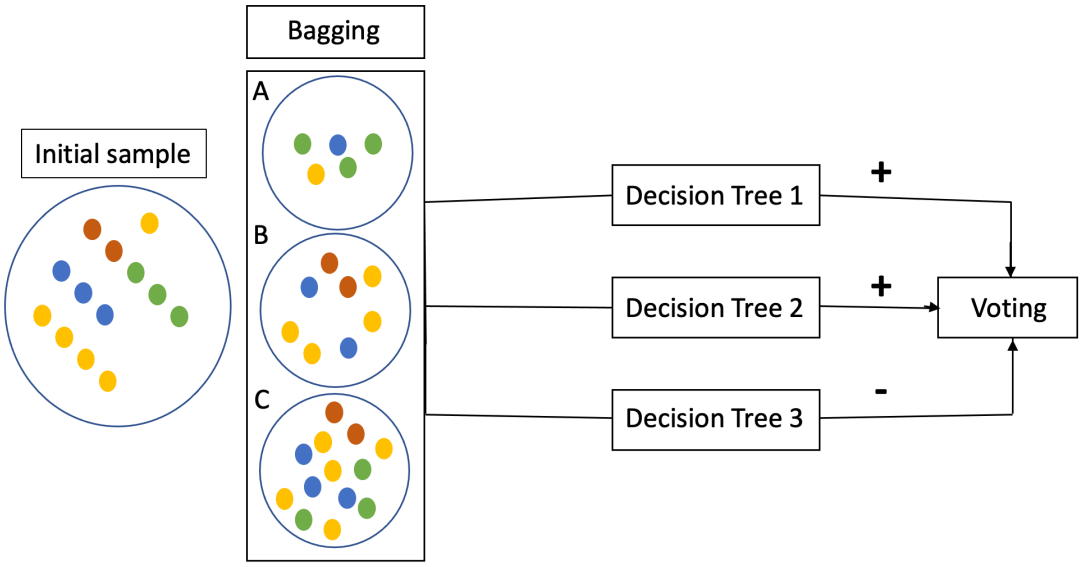

Figure 8. – Exemple de forêt aléatoire.

A : l'ensemble de données *«bootstrap»* ; B : l'ensemble de données *out-of-bag et* C : l’ensemble de données d'origine.

1.4.2 Sélection des caractéristiques

La sélection des caractéristiques ou des gènes est un moyen de sélectionner un sous-ensemble d’informations à partir des données pour éliminer les caractéristiques bruyantes et redondantes, réduisant ainsi la dimensionnalité des données [117]. L’objectif de la sélection des gènes est de réduire la complexité du classificateur et d'augmenter autant que possible la précision de la classification [117]. La plupart des techniques de sélection analysent l'ensemble des gènes pour le sous-ensemble de gènes le plus optimal [117]. Parmi ces techniques, le seuil de variance est une approche de base simple, qui pourrait être efficace pour la sélection de caractéristiques. Par défaut, cette approche élimine tous les gènes à variance nulle, c'est-à-dire les gènes qui ont la même valeur dans toutes les cellules [118]. Les caractéristiques ou les gènes avec une variance plus élevée peuvent contenir des informations plus utiles, mais cette approche ne prend pas en compte la relation entre les différents gènes, ce qui est l'un des inconvénients des méthodes de filtrage [119].

1.4.3 Validation croisée

Dans ce travail, la validation croisée k-Fold est utilisée pour la validation du classificateur. Cette méthode de validation fonctionne comme suit. Initialement, les données d'entrée sont divisées en k sous-ensembles égaux. Le classifieur est ensuite entraîné sur k-1 sous-ensembles et testé sur la partie restante. La figure 9 illustre la validation croisée en 4 fois. L'ensemble de données est divisé en quatre sous-ensembles égaux. Trois sous-ensembles sont donnés au classificateur pour l'apprentissage, et une partie est utilisée pour tester le modèle. Ce processus est itéré 4 fois. Enfin, la moyenne de la mesure de performance souhaitée est calculée pour évaluer le classificateur [120].

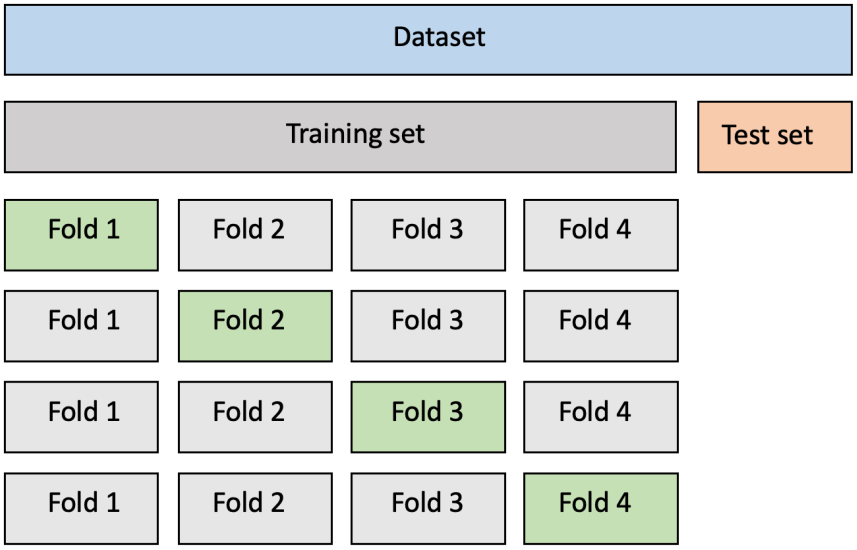

Figure 9. – Illustration du processus de validation croisée en 4 fois.

1.4.4 Mesures de rendement

Des mesures de performance sont nécessaires pour comparer les performances des classificateurs sur les ensembles de données. Chaque classificateur rapporte une matrice de confusion, qui aide à évaluer les performances du classificateur. De nombreuses mesures de performance peuvent être utilisées pour comparer les classificateurs : la précision, la mesure F et le coefficient de corrélation de Matthews (MCC) [121]. Le tableau 5 représente les formules de calcul des différentes mesures de performance. Les classes réelles sont les étiquettes associées à chaque cellule, tandis que les classes prédites sont les classes données par le classificateur. Un résultat est dit vrai lorsqu'un échantillon est correctement détecté par le classificateur, tandis

qu'un faux se produit lorsqu'un échantillon est mal prédit [121]. En général, la précision est une bonne mesure de performance dans le cas d'ensembles de données équilibrés. Plus la précision est élevée, meilleures sont les performances du classificateur (Tableau 5). La précision est la probabilité qu'un échantillon soit positif et qu'il soit effectivement prédit comme positif.

La précision et le rappel se réfèrent aux échantillons positifs. Ils se concentrent sur la façon dont le classificateur classe uniquement les échantillons positifs. Le rappel est également connu sous le nom de sensibilité ou taux de vrais positifs. La précision et le rappel sont inversement proportionnels l'un à l'autre. La mesure F est la moyenne harmonique de la précision et du rappel. Plus la moyenne harmonique est élevée, meilleur est le classificateur considéré [121]. Comme le montre le tableau 5, le coefficient de corrélation de Matthews (MCC) est considéré comme une mesure de performance équilibrée pour évaluer un classificateur. MCC traite tous les points positifs et négatifs de la matrice de confusion. La valeur MCC varie de -1 à +1. Si la valeur est proche de -1, alors le classificateur contredit les classes réelles et prédites. Si la valeur est +1, alors le classificateur est considéré comme le meilleur classificateur. Si la valeur est 0, le classificateur a effectué une prédiction aléatoire [121].

	Predicted Negative Class	Predicted Positive Class
Actual Negative Class	TN (True negatif)	FP (False positif)
Actual Positive Class	FN (False negatif)	TP (True positif)

Tableau 4. – Matrice de confusion.

Total	N = TN + FR + FN + TP
Accuracy	(TP + TN) / N
Precision	TP / (TP + FP)
Recall	TP/Precision
F-measure	2 * ((Precision * Recall) / (Precision + Recall))
MCC	TP * TN – FP * FN / (√(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))

Tableau 5. – Mesures de performance utilisées pour évaluer l'efficacité d'un classificateur.

1.5 Problématique et hypothèse

La LMA est la forme de leucémie la plus courante chez l'adulte toujours associée à une mortalité élevée. Cette maladie représente un groupe de cancers agressifs qui touchent les cellules souches et progénitrices de la moelle osseuse, bloquées à des stades de différenciation immatures suite à l'acquisition de mutations et anomalies cytogénétiques diverses. Les mutations présentes dans les LMA sont souvent sous-clonales, c'est-à-dire qu'elles ne sont présentes que dans une portion des cellules. L'hétérogénéité cellulaire de la LMA reste faiblement caractérisée car peu d'études se sont intéressées à l'explorer. Nous croyons que cette hétérogénéité cellulaire pourrait avoir une grande importance dans la sensibilité de diverses cellules leucémiques à divers traitements. Dans la dernière décennie le séquençage de nouvelle génération a permis de mieux comprendre la diversité entre différents patients leucémiques. Cependant, ces technologies analysent toutes les cellules à la fois et ne peuvent pas éclairer sur la diversité cellulaire. La technique de séquençage de l'ARN de cellules individuelles a récemment été développée pour étudier le profil de chaque cellule qui composent un tissu. Ces innovations récentes en transcriptomique ont permis aussi de retracer les hiérarchies cellulaires des tissus complexes à l'aide des mutations somatiques naturelles tel que les variations mononucléotidiques (SNV) et les variations de nombre de copies (CNV) des gènes. Ces mutations se propagent à travers les divisions cellulaires et permettent de reconstruire la hiérarchie clonale. De telles variations génétiques peuvent être détectées à travers des données de séquençage en ARN de cellules individuelles (single cell RNA sequencing - scRNA-seq). Notre hypothèse est qu'il est possible d'utiliser ces variations présentes dans les données de scRNA-seq en tant que codes-barres génétiques endogènes pour génotyper et caractériser les différents sous clones de LMA. Ceci nous permettrait de déterminer quel est l'impact de la sous-clonalité sur l'hétérogénéité cellulaire de LMA et de déterminer quel est la conséquence précise des anomalies génétiques propres à différent clone sur le profil transcriptomique.

Le but général de mon projet de maîtrise est d'utiliser les données de scRNA-seq d'une grande cohorte de LMA afin d'identifier différents sous-clones leucémiques et de caractériser les différences de leur profil transcriptomique.

Nos objectifs spécifiques sont donc de :

- Annoter les populations cellulaires des échantillons LMA de notre cohorte de 20 patients.
- Identifier quelles mutations somatiques de la LMA, pourront être facilement détectées dans les données de scRNA-seq.
- Détecter les variations de nombre de copies d'ADN dans cette même cohorte et de détecter les sous-populations qui diffèrent à cet égard chez un même patient.
- Caractériser les sous-clones identifiés par les approches précédentes et d'identifier les aberrations d'expression spécifiques à chaque clone.

Chapitre 2 – Méthodologie

2.1 Description de la cohorte

Cette étude fait partie du projet Leucégène (http://Leucegene.ca), une initiative approuvée par les comités d'éthique de la recherche de l'Université de Montréal et de l'Hôpital Maisonneuve-Rosemont. La cohorte Leucégène d'échantillons primaires de LMA humaine provenant de la Banque de cellules leucémiques du Québec (BCLQ) a été collectée entre 2001 et 2015 dans 5 hôpitaux universitaires et 4 hôpitaux régionaux de la province de Québec, Canada, selon les procédures de la BCLQ et la Déclaration d'Helsinki. Ces échantillons sont prélevés et cryoconservés conformément aux Procédures opératoires normalisées du Réseau canadien de banques de tissus (RCBT). Dans ce mémoire, nous nous sommes intéressés à 20 échantillons de la cohorte Leucégène (Tableau 6). Les cellules mononucléées des patients séparées par centrifugation en gradient Ficoll-Paque des aspirats de sang ou de moelle osseuse prélevés au moment du diagnostic, ont fait l'objet d'un scRNA-seq par la technique *10X Genomics Chromium*.

Sample	Cytogenetic group	Tissue	FAB classification	% de blast
11H103	Complex	Blood	AML-M1	70
10H130	Complex	Blood	Not classifiable by FAB	61
12H106	Complex	Blood	Not classifiable by FAB	87
12H138	Complex	Blood	AML-M2	30
06H074	Complex	Blood	AML-M5A	50
10H031	MLL translocations	Blood	AML-M5B	73
09H032	MLL translocations	Bone marrow	AML-M5A	94
09H010	MLL translocations	Bone marrow	AML-M5A	93
05H066	MLL translocations	Bone marrow	AML-M4	75
09H060	Monosomy 7	Bone marrow	AML-M4	26
17H065	Monosomy 5	Bone marrow	Not classifiable by FAB	66

11H097	Monosomy 5	Blood	AML-M1	75
05H193	Monosomy 7	Bone marrow	Not classifiable by FAB	58
05H034	Monosomy 5	Bone marrow	AML-M1	73
04H096	Monosomy 7	Bone marrow	AML-M0	92
14H007	Normal karyotype	Blood	AML-M2	80
07H134	Normal karyotype	Blood	AML-M5	81
08H087	Normal karyotype	Bone marrow	AML-M2	60
12H010	Normal karyotype	Bone marrow	AML-M1	85
09H070	Normal karyotype	Bone marrow	AML-M1	80

Tableau 6. – Description de la cohorte LMA.

2.2 Le scRNA-seq de 10X Genomics Chromium

2.2.1 Génération des données scRNA-seq

Les données scRNA-seq pour toute la cohorte LMA ont été obtenues par la technique du *10X Genomics Chromium*. Les suspensions cellulaires triées ont été chargées sur des puces microfluidiques à usage unique. Les librairies d'ARN-seq à cellule unique ont été préparées à l'aide de la technologie GemCodeTM et du kit de la librairie Single Cell 3' v2 selon les spécifications du fabricant. Les librairies de séquençage à code-barres ont été chargées sur une plate forme de séquençage Illumina NextSeq ou NovaSeq. Pour la génération d'une matrice d'expression des gènes de toutes les cellules, plusieurs étapes d'analyse bioinformatique ont été nécessaires. Les fichiers FASTQ ont été téléchargés à partir de la plate forme de partage de données et ont été traités sur des programmes basés sur UNIX. Tous les programmes utilisés étaient gratuits et disponibles en ligne, certains ont été spécialement développés pour l'analyse des données de cellules individuelles *chromium 10X*. Les étapes détaillées seront expliquées dans ce chapitre. Des exemples de commandes seront expliqués pour une compréhension plus approfondie. Avant de commencer l'analyse, un génome humain de référence pour l'alignement a dû être créé.

Par conséquent, nous avons téléchargé les fichiers de référence du génome à partir du serveur UCSC. La version du génome humain hg38 a été utilisée pour les alignements de chaque échantillon. L'alignement des transcrits, le comptage et la normalisation inter-librairies ont été effectués à l'aide du pipeline Cell Ranger [58] (*10X Genomics*, paramètres par défaut, version 5.0). Parmi les fichiers qui sont générés automatiquement par Cell Ranger pour chaque échantillon, deux fichiers sont indispensables pour la suite de l'analyse :

- Un fichier Bam : « possorted_genome_bam.bam » qui contient les « reads » alignées sur le génome et le transcriptome annotées avec des informations de code-barres.
- Un fichier HDF5 : « filtered_feature_bc_matrix.h5 » qui contient une matrice de compte filtrées en nombre de molécules ou UMIs.

2.2.2 Analyse bio-informartique

Pour la suite de l'analyse, nous avons utilisé la librairie Scanpy v1.4 [68] implémenté en Python qui traite efficacement les ensembles de données de plus d'un million de cellules. Scanpy est une boîte à outils évolutive pour l'analyse des données scRNA-seq. Elle comprend le prétraitement, la visualisation, le regroupement et les tests d'expression différentielle.

2.2.2.1 Contrôle de qualité et normalisation

En raison des variations biologiques et techniques, il a été nécessaire d'établir une variété de paramètres de contrôle de la qualité afin que chaque ensemble de données soit comparable et que les erreurs possibles soient détectées et éliminées. Les cellules contenant plus de 20 % (seuil basé sur la distribution) de transcrits mitochondriaux ont été supprimées. Ces cellules sont généralement non informatives en voie d'apoptose avec un faible pourcentage de gènes exprimés. Les gènes exprimés dans moins d'une cellule ont également été retirés. Pour chaque cellule, l'expression de chaque gène a été normalisée à la profondeur de séquençage de la cellule, mise à l'échelle à une profondeur constante (10 000) et transformée en log (log(X=1)).

2.2.2.2 Réduction dimensionnelle par analyse en composantes principales

Une analyse en composantes principales (ACP) a été réalisée sur les données normalisées.
Cette fonction au sein de Scanpy [68] évalue chaque gène en fonction de la corrélation des composantes calculées et créé un ensemble de variables linéaires non corrélées appelées composantes principales. Pour l'identification de marqueurs génétiques ayant une forte corrélation avec l'hétérogénéité cellulaire, cette méthode est généralement le premier choix. Il est donc important d'évaluer le nombre de PC lors de l'analyse en aval. Dans cette analyse, nous nous sommes concentrés sur l'écart-type de chaque PC en les traçant. Nous avons utilisé les 100 premières PC.

2.2.2.3 Regroupement des cellules

Dans cette approche, Scanpy utilise les informations de l'analyse des composantes principales pour regrouper les cellules avec une expression génique similaire proches les unes des autres. En appliquant l'algorithme PhenoGraph [80], une méthode de regroupement conçue pour les données scRNA-seq de grande dimension, les cellules sont regroupées sur la base de leurs similitudes phénotypiques en formant des communautés sous forme de réseau. Dans Scanpy, il est possible de définir la granularité pour définir le *nombre* de *clusters* résultants.

Les deux outils les plus utiliser pour visualiser et explorer de nouveaux ensembles de données sont tSNE [71] et UMAP [72] qui ont été présenté dans le premier chapitre. Avec cette réduction dimensionnelle non linéaire, les cellules avec des voisinages locaux similaires se regroupent dans un espace de faible dimension. En conséquence, la visualisation de cellules groupées peut aider à fournir des informations sur les proportions de cellules au sein de chaque groupe et à voir les différences transcriptionnelles et les similitudes dans un graphique. Comme entrée pour cette réduction dimensionnelle non linéaire, des PC générés précédemment ont été utilisés. Ces graphiques peuvent également donner un excellent aperçu du moment où les données des patients ont été combinées.

2.2.2.4 Gènes et biomarqueurs différentiellement exprimés

Jusqu'à présent, nous n'avions pas visualisé l'expression des gènes. Pour obtenir une liste de gènes définissant l'expression génique de chaque groupe spécifiquement, le package SciPy [122] implémenté en Python fournit une adaptation de la procédure de test semi-paramétrique basée sur la distance Wasserstein qui est utilisé pour identifier les distributions différentielles dans les données de scRNA-seq. Les dix premiers marqueurs positifs et négatifs de chaque groupe ont été comparés à toutes les autres cellules dans cette approche. La visualisation de ces marqueurs est utile lorsqu'il s'agit de comparer les *clusters*. Cela pourrait aider à mettre en évidence la spécificité d'un gène a un groupe de cellules et même la qualité du regroupement. Une représentation en «heatmap» a été utilisée pour examiner l'hétérogénéité au sein et entre les *clusters*. Un certain nombre de marqueurs spécifiques aux groupes ont été choisis pour être tracés. Tous ces outils nous ont aidés à définir l'hétérogénéité des échantillons et les expressions des biomarqueurs. En examinant ces biomarqueurs, nous avons était capable de définir les types cellulaires majoritaires de certain *cluster* et l'hétérogénéité au sein d'un échantillon a pu être explorée.

2.3 Annotation des types cellulaires

2.3.1 Prédiction de type cellulaire

La plupart des analyses scRNA-seq dépendent des connaissances des experts pour attribuer manuellement les cellules à une liste de types de cellules attendus, sur la base d'hypothèses biologiques qui sont presque toujours spécifiques à un ensemble de données. Le processus d'annotation de type de cellule peut être fastidieux et inefficace, ce qui entraîne de grandes variations entre les annotations, ce qui empêche la reproductibilité des résultats dans différents ensembles de données et laboratoires de recherche. Ce problème est exacerbé à mesure que le nombre de cellule augmente de manière exponentielle, ce qui nécessite plus de temps et de ressources humaines pour terminer l'annotation d'un seul ensemble de données.

Dans le contexte d'un cancer donnée, l'expression de certains gènes marqueurs peut être aberrante d'où l'avantage d'utiliser le transcriptome complet pour annoter les cellules. De plus, nous établissons l'hypothèse que les cellules leucémiques partageront une certaine similitude avec des cellules de différentiation similaire de l'hématopoïèse normale. D'où l'idée de développer un classificateur en utilisant cette information, et une grande multitude de gènes dans notre approche d'annotation. Nous suivons une approche supervisée, où nous utilisons le jeu de données de « Human Cell Atlas » décrit ci-dessous (Figure 10) pour entrainer et valider les classificateurs. Ceci nous permettra de sélectionner le classificateur le plus performant qu'on utilisera pour annoter les types cellulaires des échantillons LMA. En suivant le raisonnement ci-dessus, nous sélectionnons trois classificateurs : Random Forest (RF), Decision Tree (DT) et KNN pour effectuer une classification multi-classe pour l'annotation des données scRNA-seq des LMA (Figure 10).

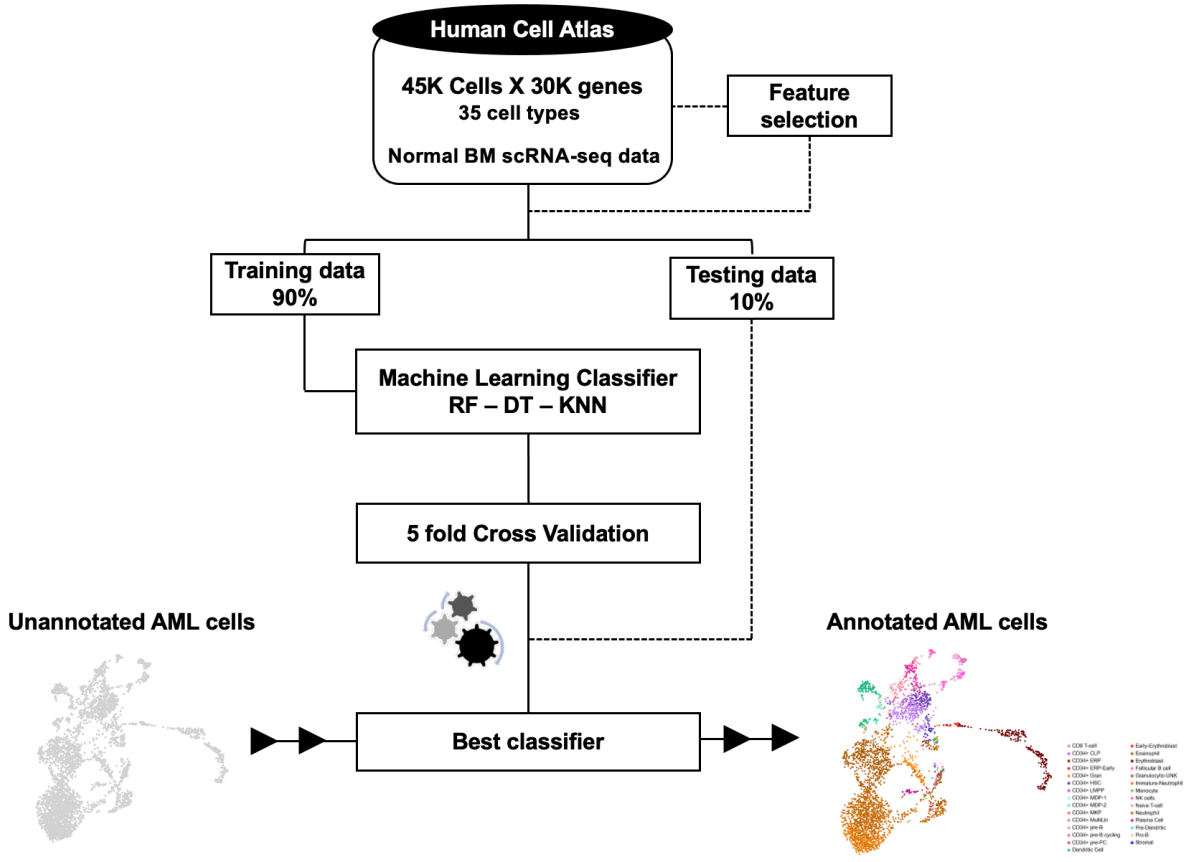

Figure 10. – Classificateur d'apprentissage automatique pour l'annotation de type cellulaire de LMA.

2.3.2 Collecte et pré-traitement des données

Les données de scRNA-seq de la moelle osseuse de huit donneurs sains ont été obtenues à partir du portail de données HCA [104] (https://preview.data.humancellatlas.org). La matrice en UMI a été obtenu avec le logiciel Cell Ranger V2 de la plateforme *10X Genomics*. Bien que, collectivement, ces données incluent des données pour plus de 100 000 codes-barres cellulaires, nous avons limité le jeu de données pour éviter la surreprésentation de certain type cellulaire par rapport aux autres. Nous avons sélectionné un maximum de 3 000 cellules par type cellulaire pour un total de 43 668 cellules. Ensuite, nous avons appliqué les mêmes méthodes de prétraitement utilisées précédemment avec nos échantillons LMA. De plus, nous avons utilisé l'annotation originale des types cellulaires fournie par le Consortium HCA.

2.3.3 Sélection des attributs

Le processus de sélection des caractéristiques identifie les gènes qui présentent une variation élevée de cellule à cellule en raison de différences biologiques. Ce processus suppose que les grandes différences dans l'expression des gènes entre les cellules individuelles peuvent être attribuées à d'importantes différences biologiques entre les cellules, plutôt qu'au bruit technique. Pour se faire, un sous-ensemble de 10 000 gènes les plus variables ont été sélectionné en se basant sur le calcul de la déviation standard.

2.3.4 Entrainement des classificateurs

Nous avons utilisé la classification pour prédire la similitude de cellules individuelles de LMA avec les 35 types différents détectés dans les moelles osseuses normales HCA [104]. Dans le cas de notre classification, les échantillons représentent les cellules, les caractéristiques représentent des gènes et les classes représentent différents types de cellules. Pour notre analyse, nous avons utilisé la librairie « scikit-learn » [123] de l'apprentissage automatique implémenté en Python.
Une fois la matrice de données de HCA est prête (43 668 cellules X 10 000 gènes), nous avons partagé le jeu de données en trois parties : jeu de données d'entrainement 70%, jeu de données de validation 20 % et jeu de données de test 10%. Les classificateurs ont a été évalué en effectuant une validation croisée 5 fois en divisant l'ensemble de données d'entrainement en cinq parties de taille égale.

À chaque itération de la validation croisée, quatre de ces parties ont été utilisées pour générer un classificateur qui a ensuite été utilisé pour prédire les probabilités de classe de la partie restante.

2.3.5 Évaluations des classificateurs

Nous avons implémenté le score de précision (fonction metrics.accuracy_score dans le package sklearn) comme mesure de performance par défaut, qui correspond au rapport entre le nombre total de cellules correctement attribuées et le nombre total de toutes les cellules testées. Nous avons également implémenté le score de précision équilibré (fonction metrics.balanced_accuracy_score dans le package sklearn), qui peut être calculé comme suit : pour chaque type de cellule de l'ensemble de test, nous calculons le ratio correctement attribué (rappel), puis obtenons la moyenne de rappels sur chaque type de cellule. L'évaluation du classificateur a été effectué sur une partie de l'ensemble du jeu de données de cellules normales de HCA en comparant la prédiction du classificateur avec l'étiquette de référence de chaque cellule.

2.3.6 Validations de l'annotation des données de LMA

Une fois le classificateur le plus performant est sélectionné, le modèle ainsi que la liste des attributs ont été enregistré pour nous servir à annoter la cohorte de cellules individuelles de LMA. Une inspection manuelle des gènes marqueurs pour chaque type cellulaire a été effectué afin de valider la précision de notre classificateur à prédire correctement des cellules malignes.

2.3.6.1 Validations basée sur l'expression des gènes marqueurs

Pour l'exploration et la validation de l'annotation des échantillons LMA, nous avons utilisé un échantillon de notre cohorte qui a une bonne granularité pour comparer le profil d'expression des gènes marqueurs par type cellulaire au niveau des cellules de HCA et de notre échantillon LMA (09H060) avant d'annoter toute la cohorte.

2.3.6.2 Validations basée sur l'analyse d'enrichissement des gènes

Dans cette étape, nous avons eu recours à une analyse d'enrichissement des gènes pour confirmer l'annotation des échantillons LMA par le classificateur.

Une moyenne de l'expression des gènes par type cellulaire a été calculée pour notre échantillons ainsi pour le jeu de données de HCA. Ces matrices d'expression ont fait l'objet de l'analyse d'enrichissement d'une liste des gènes (C8) téléchargée à partir de la base de données MSigDB [124]. Cet ensemble de gènes contient des gènes marqueurs de *cluster* pour les types cellulaires identifiés dans des études de séquençage à cellule unique de tissus humains. Ces ensembles de gènes ont été sélectionnés à partir de la littérature et représentent des gènes de signature et des identifications de types cellulaires tels que représentés dans leurs publications d'origine respectives. Les ensembles de gènes présents dans cette collection couvrent un certain nombre de types de cellules du cœur, du tractus gastro-intestinal, du pancréas, des reins, du foie, du système immunitaire, de la rétine, des tissus olfactifs et du cerveau. Le développement de ces ensembles de gènes a été fourni par le programme « Collaborative Computational Tools for the Human Cell Atlas » [104]. Le principe de la méthode est présenté à la fin du chapitre des méthodes.

2.4 Couverture et performance pour la détection des variants dans les données scRNA-seq

La capacité de génotyper des cellules à partir des données scRNA-seq est un attribut idéal pour déduire des populations de LMA sous clonales. Ici, nous évaluons l'utilité des données scRNA-seq pour la détection de variantes somatiques, autre que ceux situés seulement dans la région 3', dans des échantillons de moelle osseuse LMA cryoconservés. Pour répondre à cette question nous avons sélectionnée une liste de 50 gènes (Annexes) avec les mutations récurrentes dans la pathologie de LMA [125]. Par la suite, nous avons créé un fichier « .bed » qui comporte les cordonnées de ces gènes. Le fichier comporte trois colonnes : le chromosome, la position de début et de fin du gène. Ces informations ont été obtenu à partir du fichier GTFfile du génome humain hg38. La librairie « samtools » avec l'option « coverage » a été utilisé pour interroger les fichier « bam » des 20 échantillons LMA. Cette commande nous permet d'avoir la couverture de chaque position des 50 gènes sélectionnés. Au final, les différents fichiers générés ont été combinés. Des figures ont été générées en combinant la couverture des gènes ainsi les positions des mutations somatiques spécifiques pour chaque gène.

2.5 Détection des sous-clones LMA

2.5.1 Basée sur les SNP

Une fois les données de scRNA-seq annotées, nous avons sélectionnés quelques mutations somatiques qui sont détectables dans notre jeu de données. Les coordonnés de ces mutations ont été enregistrées dans un fichier «.bed». Pour faire le génotypage des cellules LMA, les fichiers BAM des échantillons sont interrogés avec freebayes 1.2.0 [126], un outil d'appel de variant largement utilisé. Freebayes est un détecteur de variant génétique bayésien conçu pour trouver de petits polymorphismes, en particulier des SNP (polymorphismes mononucléotidiques), des indels (insertions et suppressions), des MNP (polymorphismes multinucléotidiques) et des événements complexes (événements composites d'insertion et de substitution). Freebayes utilise des alignements en « reads » (BAM) d'un ou plusieurs échantillons et un génome de référence (au format FASTA) pour déterminer la combinaison de génotypes la plus probable pour la population à chaque position dans la référence.

Freebayes signale les positions qu'il trouve potentiellement polymorphes au format de fichier d'appel variant (VCF). L'expression spécifique d'un allèle dans des cellules de tumeur peut conduire à de fortes corrélations entre la présence d'un variant exprimé spécifique et le regroupement basé sur l'expression. La superposition des informations sur les variantes avec le regroupement d'expressions peut conduire à de nouvelles informations sur la LMA et à l'accumulation de mutations pouvant conduire à différents phénotypes tels que la rechute ou la résistance aux médicaments. Pour évaluer l'hétérogénéité au sein de chaque échantillon, nous avons utilisé l'outils Vartrix [100] pour associer à chaque cellule son génotype. Cet outil été développé et publié sur des données scRNA-seq de LMA. Cependant, la cohorte analysée ne reflète pas une aussi grande diversité génétique que celle de notre cohorte. VarTrix est un outil permettant d'extraire des informations sur les variants à partir de données scRNA-seq *10X Genomics*.

VarTrix prendra un ensemble de variants précédemment définis dans le fichier VCF généré par Freebayes et l'utilisera pour identifier ces variants dans les données scRNA-seq à partir du fichier BAM. VarTrix utilise l'alignement Smith-Waterman pour évaluer les « reads » qui correspondent à chaque locus de variant d'entrée connu et attribuer ces variants à chaque cellule. VarTrix nécessite également un fichier de codes-barres cellulaires produit par Cell Ranger. VarTrix produit une matrice qui contiendra des informations sur chaque variant pour chaque code-barres de cellule [100]. Quatre cas de figures sont possibles : la cellule possède l'allèle de référence (sauvage), l'allèle alternative (muté), les deux allèles ou aucun des deux. Comme la majorité des mutations sont hétérozygote, nous avons considéré les cellules ayant à la fois l'allèle de référence et alternatif comme des cellules mutées. Au final ces données sont présentées sur des figures en UMAP pour une meilleure compréhension de la distribution des mutations dans chaque échantillon.

2.5.2 Basée sur les CNV

Un défi majeur pour le scRNA-seq de la LMA, est de distinguer les cellules cancéreuses des types de cellules non malignes, ainsi que la présence de plusieurs sous-clones tumoraux. Par exemple, il pourrait y avoir des HSPC normaux coexistant avec des cellules blastiques dans le microenvironnement leucémique.

Dans cette optique, nous avons optimisé un algorithme publié récemment, CopyKAT [127] qui combine une approche bayésienne avec un regroupement hiérarchique pour calculer les profils de nombre de copies génomiques de cellules individuelles et définir la sous-structure clonale à partir de données 10X à haut débit. La logique sous-jacente du calcul des événements du nombre de copies d'ADN, à partir des données scRNA-seq, est que les niveaux d'expression génique de nombreux gènes adjacents peuvent fournir des informations approfondies pour déduire le nombre de copies génomiques dans cette région. Pour cela nous avons utilisé, la matrice d'expression génique en UMI comme entrée pour les calculs. Cette approche commence par l'annotation des gènes pour les ordonner par leurs coordonnées génomiques.

Une transformation est effectuée pour stabiliser la variance, suivie d'une modélisation linéaire dynamique polynomiale pour lisser les valeurs aberrantes dans les dénombrements UMI. L'étape suivante consiste à déduire les valeurs de base du nombre de copies des cellules diploïdes normales. Pour détecter les points de rupture chromosomiques, l'algorithme intégre un modèle Poisson-gamma et des itérations de Markov Chain Monte Carlo (MCMC) pour générer des moyennes postérieures par fenêtre de gêne, puis applique des tests de Kolmogorov-Smirnov (KS) pour joindre des fenêtres adjacentes qui n'ont pas de différences significatives dans leur signifie. Les valeurs finales du nombre de copies pour chaque fenêtre sont ensuite calculées comme les moyennes postérieures de tous les gènes couvrant les points de rupture chromosomiques adjacents dans chaque cellule [127]. Par la suite, les valeurs du nombre de copies résultantes de l'espace géniques sont converties en positions génomiques en réarrangeant les gènes en bacs génomiques variables de 220 kb pour obtenir un profil de nombre de copies à l'échelle du génome pour chaque cellule à une résolution approximative de 5Mb. Enfin, les données sur le nombre de copies monocellulaires sont regroupées pour identifier les sous-populations clonales et calculer les profils de consensus représentant les génotypes sous-clonaux pour une analyse plus approfondie de leurs différences d'expression génique [127].

2.5.2.1 Optimisation de l'approche

CopyKAT a eu des difficultés à prédire automatiquement les cellules tumorales et normales dans les cas de tumeurs liquides comme la LMA. CopyKAT fournit deux façons de contourner cela : utiliser un jeu de données de cellules normales connues à partir du même ensemble de données par exemple des cellules T ou choisir un ensemble de cellule de référence pour mieux déduire les valeurs de base du nombre de copies des cellules.

Pendant l'étape d'optimisation, nous avons sélectionné un échantillon avec une aberration chromosomique dans la majorité de ses cellules déjà connue à travers le caryotype de cet échantillon pour valider tous les paramètres de l'algorithme avec les différents modes. Concernant les cellules de référence, nous avons utilisé cellules immatures des échantillons a caryotype normal.

Nous avons par la suite fixé au moins 15 gènes par segment dans chaque chromosome pour calculer le nombre de copies d'ADN. Pour le reste des paramètres, nous avons conservé ceux utilisés par défaut par l'algorithme. CopyKAT est implémenté en R 4.0. Les lignes de code utilisés seront jointes en annexe. De plus, nous avons utilisé les données de séquençage en « Bulk » d'exome appariées tumoraux et sains des mêmes échantillons pour définir les segments chromosomiques associés aux aneuploïdies. Ces données ont été utilisées pour délimiter les fenêtres d'expression estimées avec CopyKAT. Ces données ont été déjà générées par l'équipe du laboratoire grâce à l'algorithme Sequenza [128]. C'est un outil pour analyser les données de séquençage génomique à partir d'échantillons de tumeurs et tissus sains appariés, y compris l'estimation de la ploïdie ; détection, quantification et visualisation du nombre de copies.
Au final, pour chaque aberration chromosomique, la moyenne du nombre de copies de toutes les fenêtres des données de scRNA-seq estimées par CopyKAT situées entre les intervalles, estimés par sequenza, sont présentées sur des figures en UMAP pour une meilleure visualisation.

2.6 Caractérisations des sous-clones LMA

La combinaison des informations du caryotype et les résultats de la variation du nombre des copies, nous a permis de détecter des sous clones spécifiques pour chaque échantillon. Par la suite, nous avons pensé à caractériser les variations transcriptomiques entre les populations avec différents aberrations chromosomiques. Pour cela nous avons suivi l'analyse en parallèle, pour caractériser les deux groupes de LMA : les monosomies 5 ou 7 et les échantillons avec un caryotype complexe.

2.6.1 Analyse de l'expression différentielle des gènes

L'analyse des gènes d'expression différentielle a été réalisée à l'aide de Python et de la librairie Scipy. Un test t de student a été effectué pour chaque gène entre les cellules des deux populations comparées. Un taux « false discovery rate » FDR a été calculé pour une limite de signifiance de (p=0.05). Un FDR de 5% signifie que, parmi tous les gènes significatifs, 5% d'entre eux sont vraiment nulles. Une transformation logarithmique de base 10 a été appliquée. Les ratios d'expressions (FC) de chaque gène ont été calculé pour chaque population par rapport à l'autre. Une transformation logarithmique binaire de ces ratios a été effectuée.

A la fin seulement les gènes significativement surexprimés ou sous-exprimés ont été définis par | log2(FC) | > 1 et un FDR < 0.05.

2.6.2 Analyse d'enrichissement des gènes

Par la suite, nous avons utilisé la libraire GSEAPY [129] en Python pour effectuer une analyse d'enrichissement de l'ensemble de gènes des ensembles des données LMA. Notre hypothèse est que les voies pertinentes devraient contenir certains gènes avec des niveaux d'expression qui varient entre les cellules de différents types cellulaires.

Nous avons choisi l'analyse d'enrichissement des voies KEGG, HallMark, GO et Reactome pour comparer les voies enrichies dans la LMA. Les différents ensembles de gènes ont été téléchargées à partir de la base de données MSigDB [124] pour effectuer cette analyse. Dans cette étude, la distance Wasserstein été calculée pour détecter des gènes différentiellement exprimés entre les populations à caractériser. La moyenne statistique t des gènes a été calculée dans chaque voie en utilisant un test de permutation avec 1 000 réplications. Les voies surexprimées ont été définies par un score d'enrichissement normalisé (NES) > 0.40 et les sous-exprimées ont été définies par un NES < 0.4. Les voies avec une valeur FDR-P 0.05 ont été choisies comme significativement enrichies. Nous avons utilisé le diagramme de Venn pour montrer les voies enrichies parmi ces groupes.

Chapitre 3 – Résultats

3.1 Analyse bio-informartique des données scRNA-seq

3.1.1 La cohorte LMA

Pendant l'étape de quantification de l'expression des gènes et de contrôle qualité de la matrice en UMI, les 20 échantillons scRNA-seq de LMA ont été analyses un par un. Par la suite les données ont été regroupées ensemble. Au final, ~15K cellules et ~15K gènes ont été filtrés pour avoir une matrice finale de 115K cellules X 18K gènes. Un aperçu de tous les transcriptomes LMA de haute qualité est présenté sur la figure 11. L'ensemble des cellules a été regroupé en 46 groupes lors de la visualisation des cellules sous la forme d'un UMAP (Figure 11 A). Nous remarquons que chaque échantillon LMA forme un groupe unique séparé des autres échantillons (Figure 11 B). Ceci peut être expliquer par le fait que les échantillons ont des mutations et/ou aberrations cytogénétiques différentes ce qui pourrait expliquer l'origine de cette hétérogénéité des LMA. Seulement quelques groupes sont composées des cellules a l'origine de plusieurs échantillons. Ces cellules pourraient être des cellules normales avec des profils transcriptomiques similaires. Le regroupement des cellules selon la classification FAB ainsi leur groupe génétique, montre aussi que malgré les différences, les échantillons appartenant à la même classe FAB (Figure 11 C) ou au même groupe sont plus proche (Figure 11 D).

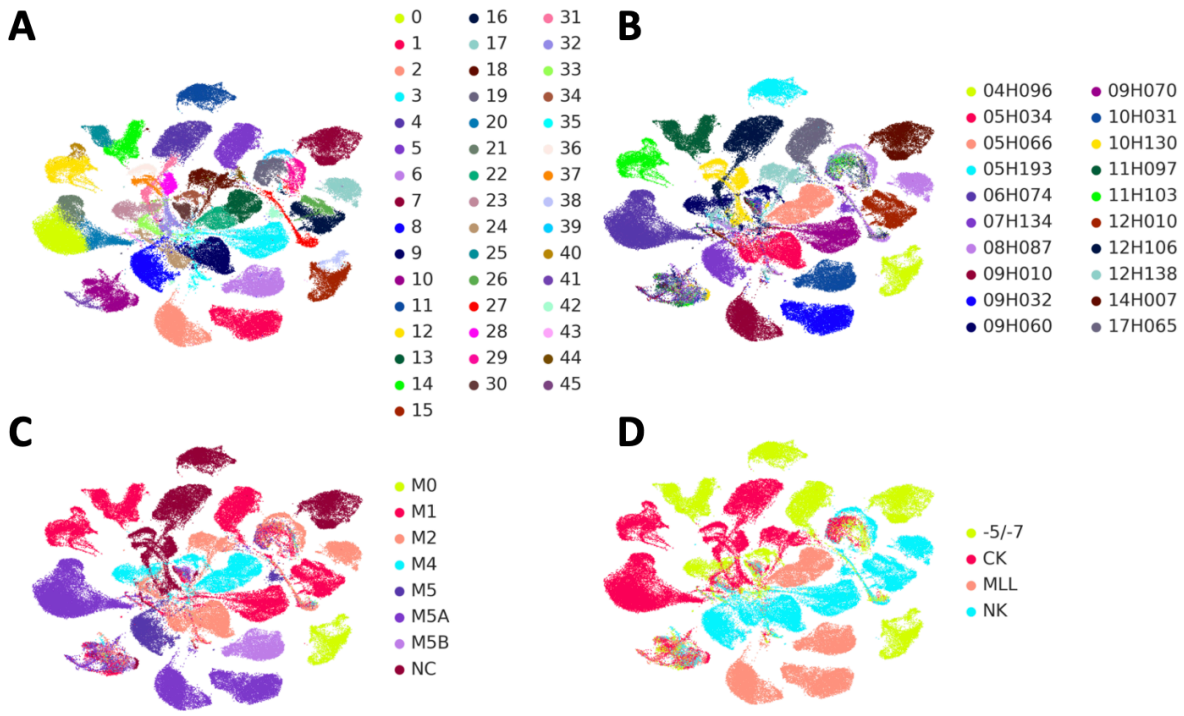

Figure 11. – Visualisation UMAP des données de 115K scRNA-seq des échantillons LMA.

Les cellules ont été marquées en fonction de (A) leur regroupement « Phenograph », (B) échantillon d'origine, (C) classification FAB et (D) groupe génétique : CK; caryotype complexe, MLL ; fusion du gène *MLL* (*KMT2A*), -5/-7 ; monosomies 5 et 7 et NK : caryotype normal.

3.1.2 Les données HCA

Les données HCA sont aussi analysées et filtrées avec la même méthode utilisée pour la cohorte LMA. La figure ci-dessous présente le regroupement des 50K cellules en plusieurs groupes annotées selon leur étiquette de type cellulaire publiée dans l'article. Comme le montre la figure 12, on peut distinguer 35 types cellulaires différents représentant les cellules souches et progénitrices hématopoïétiques, les cellules différenciées en phase terminale et les états intermédiaires. Cet atlas nous aidera pour entrainer les modèles d'apprentissage automatique afin d'annoter les données LMA et aussi servira pour une carte transcriptionnelle des états des cellules immunitaires dans une condition normale pour pouvoir comprendre la dérégulation dans le contexte de LMA.

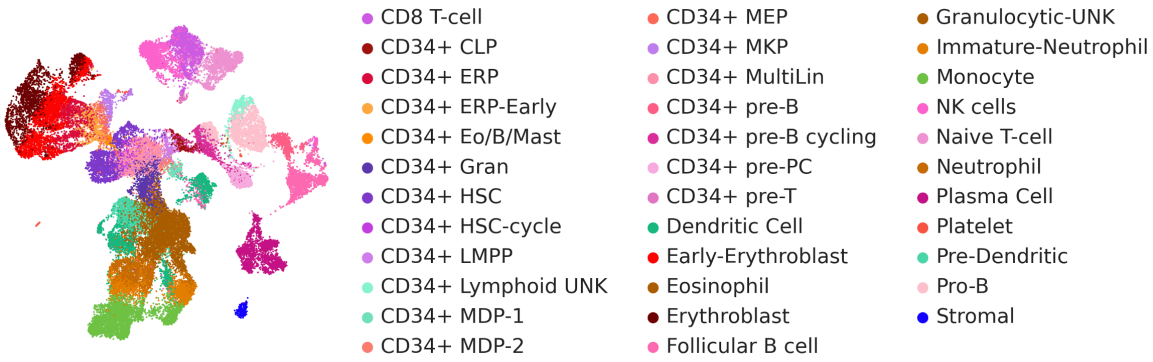

Figure 12. – Visualisation UMAP des données de 50K scRNA-seq de HCA annotées selon leur type cellulaire.

3.2 Annotation des types cellulaires

3.2.1 Évaluations des classificateurs

Le jeu de données de HCA a été ensuite utilisé pour l'entrainement des trois modèles forêt aléatoire (RF), arbre de décision (DT) et KNeighbors (KNN). Afin de prédire le type cellulaire des données LMA. Pendant la phase de validation, la précision a été mesurée pour les différents classificateurs. Ces résultats illustrés dans le tableau ci-dessous montre que le classificateur de foret aléatoire RF performe plus que le DT et RF avec une précision moyenne (validation croisée de 5 fois) de 0.90 (Tableau 7).

	KNN	DT	RF
Average Accuracy	0.26	0.83	0.90

Tableau 7. – Mesure de précision des classificateurs.

Selon la mesure de précision nous avons sélectionnés le classificateur RF pour poursuivre les étapes de validation du model et annoter la cohorte LMA. A cette étape nous avons calculé la mesure F1 score pour chaque type cellulaire afin d'évaluer la performance du classificateur RF de chaque classe.

Selon la figure 13, nous remarquons que la majorité des classes (21/35) ont des score F1 qui dépassent 80%, quelques classes (8/35) avec des scores F1 entre 60% et 80% et le score de la dernière fraction (6/35) est de moins de 60%. En se basant sur ces données, nous pouvons conclure que les classes avec un score F1 les plus faibles correspondent aux stades progéniteurs intermédiaires. Ces lignées reflètent des états de transition moins bien défini dans un système continu ; soit entre les HSC et les cellules de différentiation terminale. Ces cellules ont des profils transcriptomiques très similaires ce qui peut expliquer la faible performance du classificateur RF à prédire ces classes.

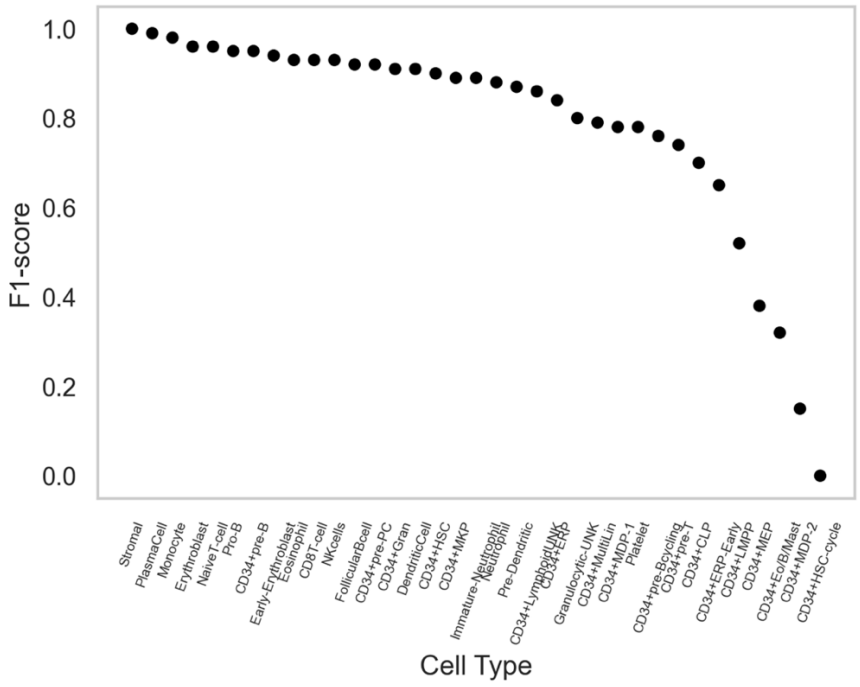

Figure 13. – Mesure de performance du classificateur RF par type cellulaire de HCA.

3.2.2 Validations de l'annotation des données de LMA

3.2.2.1 Validations basée sur l'expression des gènes marqueurs

Pour la validation de notre modèle RF sur les données LMA. Nous avons commencé par annoter un premier échantillon de la cohorte (09H060) (Figure 14 A). Une inspection manuelle des gènes marqueurs pour chaque type cellulaire a été effectué afin de valider la précision de notre classificateur à prédire correctement les cellules malignes (Figure 14 B).

Les types cellulaires identifiés avec le classificateur RF concorde avec l'expression des gènes marqueurs spécifiques. Dans cet échantillon, nous avons identifiées 4 branches majoritaires de l'hématopoïèse anormale : les Progéniteurs-like CD34+ et HOPX+, une différenciation mégacaryocytaire (Mega-like) avec un gradient d'expression des gènes *GATA1* et *HBB*, une population myéloïde (monocytes/granulocytes-like) qui exprime bien des marqueurs comme *LYZ*, *CD14* et *MPO* spécifiques aux monocytes/granulocytes, *CD1C* et *IL3RA* spécifiques aux cellules dendritiques et une population lymphoïde mature composées de plusieurs groupes de cellules : les lymphocytes T (*CD3A*+), lymphocytes T (*CD8A*+), les lymphocytes B (*CD79A*+) et les plasmocytes (*JCHAIN*+).

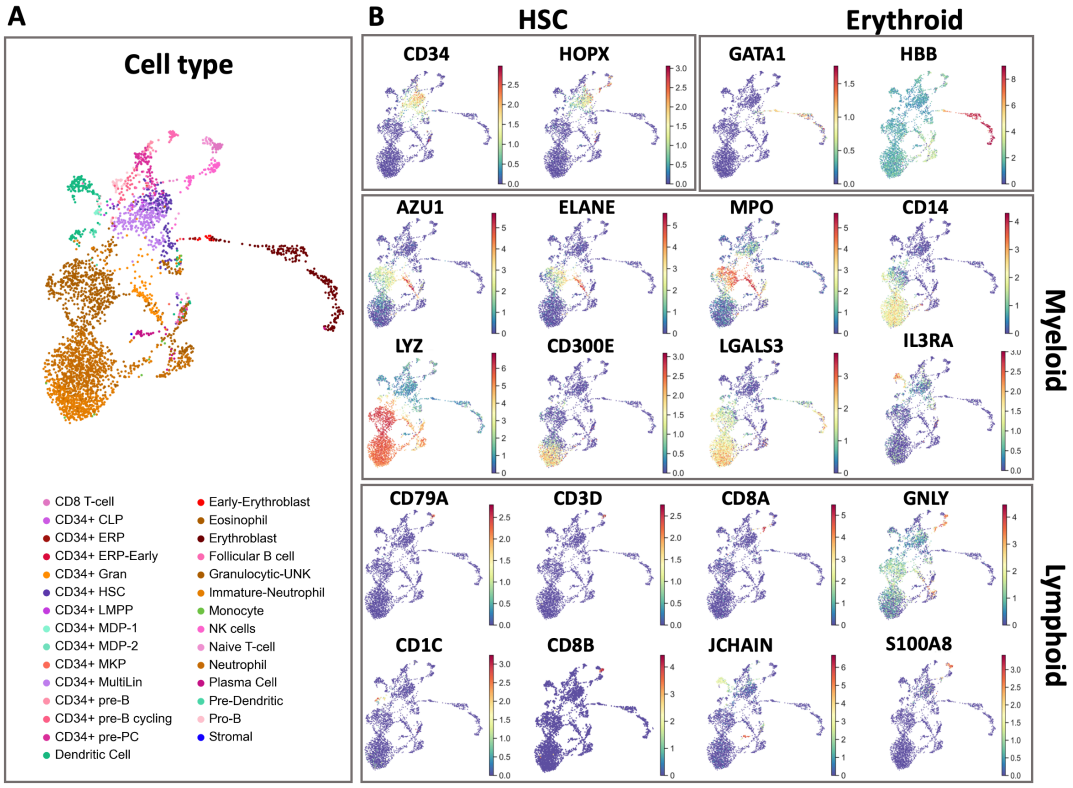

Figure 14. – Validation de l'annotation d'un échantillon LMA annotées par le classificateur RF.

(A) Visualisation UMAP des données scRNA-seq de 3500 cellules LMA annotées par type cellulaire.

(B) Visualisation UMAP de l'expression des gènes marqueurs.

Dans cette optique, nous avons comparé le profil d'expression d'une liste de marqueurs dans chaque type cellulaire des cellules normales de HCA (Figure 15 A) et de l'échantillon 09H060 figure (Figure 15 B) pour mieux comprendre comment notre classificateur RF performe avec les cellules LMA annotées. Comme le montre la figure ci-dessous, le profil d'expression des gènes ordonnés selon le niveau de différenciation définit 5 populations majoritaires similaires de HCA et l'échantillon 09H060 malgré l'expression aberrante de certains marqueurs. A ce stade, nous avons validé l'annotation avec une première approche visuelle en se basant sur les gènes marqueurs.

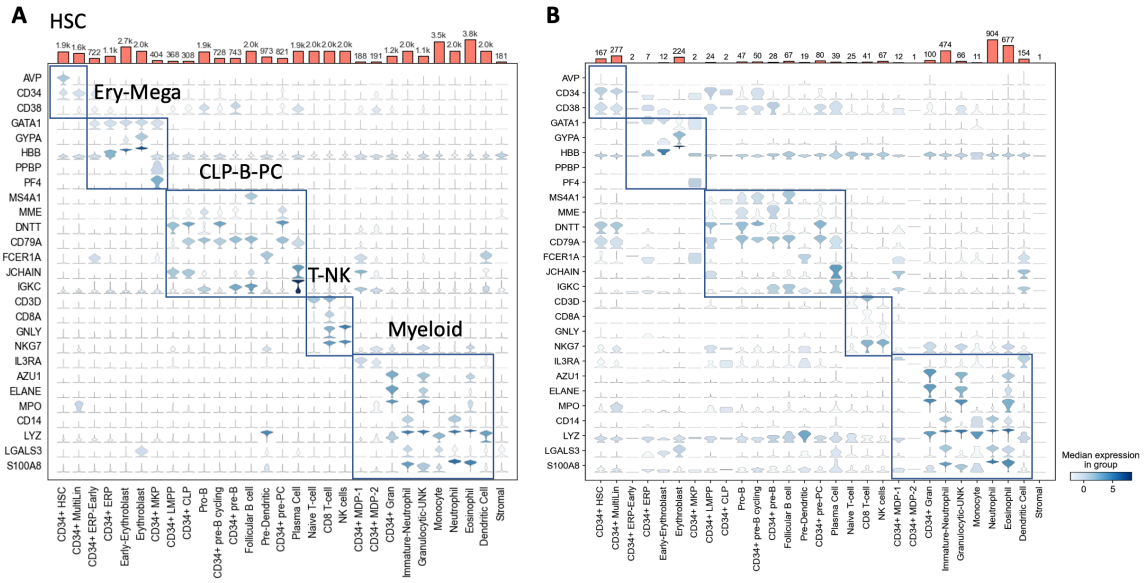

Figure 15. – « Stacked violin plot » de la différenciation hématopoïétique caractérisant les cinq populations hématopoïétiques principales.

L'expression moyenne des gènes marqueurs est calculée par type cellulaire dans le contexte des données HCA (A) et de l'échantillon LMA 09H060 (B). Une expression élevée est indiquée en bleu foncée et une expression faible en bleu clair. Le nombre de cellules par chaque type cellulaire est présenté avec une barre.

3.2.2.2 Validations basée sur l'analyse d'enrichissement des gènes

Pour une meilleure validation de l'annotation, nous avons utilisé une deuxième approche basée sur les scores d'enrichissent des types cellulaires (C8) de la base données MsigDB. Dans les deux cas de la figure 16 A (HCA) et B (LMA 09H060) la diagonale de score d'enrichissement est conservée avec quelques différences mineures. Les deux diagonales correspondent au maximum score d'enrichissement entre chaque deux type cellulaires semblables. Ces données appuient nos résultats obtenus lors de la première validation basée sur les profils d'expression des gènes marqueurs. L'ensemble de ces résultats augmentent notre degré de confiance pour annoter toute la cohorte avec le classificateur RF.

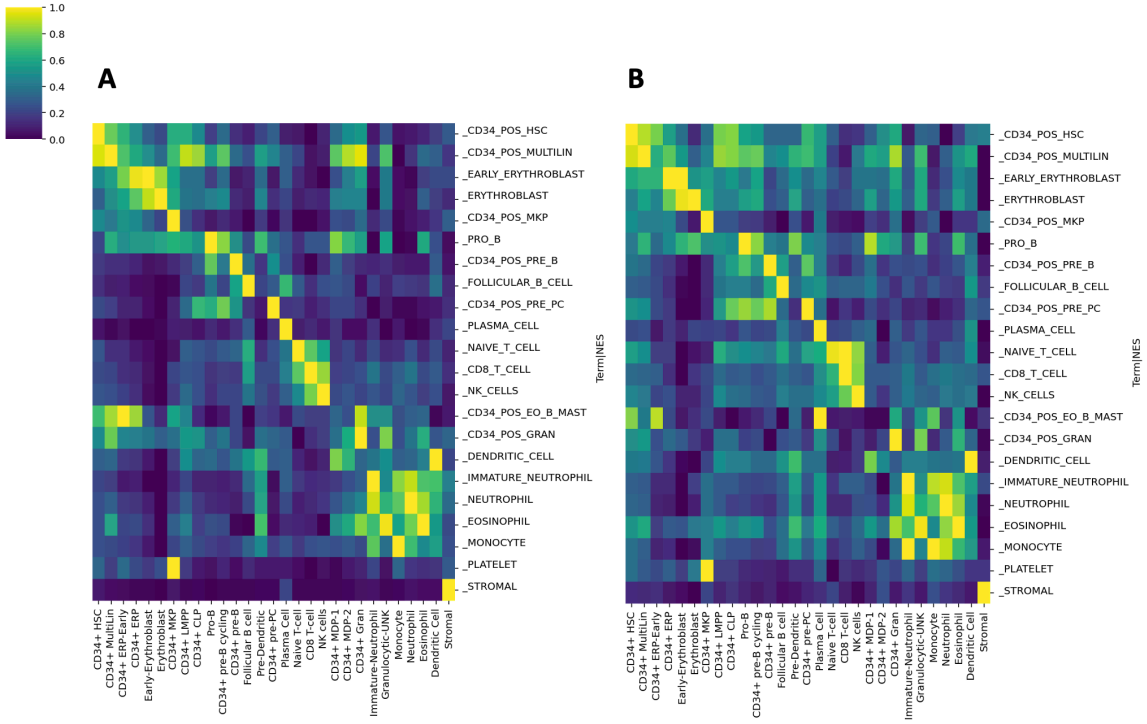

Figure 16. – Score d'enrichissement des types cellulaires.

Les types cellulaires présentés sur l'axe des x correspondent aux HCA (A), aux LMA 09H060 (B) et l'axe des y correspond aux types cellulaires (MsigDB-C8). Tous les types cellulaires sont ordonnées selon leurs dégrées de différentiation.

3.2.2 Annotation de la cohorte totale

Par la suite, toute la cohorte scRNA-seq de LMA a été annotée avec le classificateur RF. En termes d'efficacité, l'annotation de plus que 100K cellules a été effectué en moins d'une minute. Nous présentons ci-dessous une visualisation finale des cellules LMA annotées selon leur type cellulaire (Figure 17 A). Pour résumer l'hétérogénéité d'expression dans chaque cas et mieux comprendre la composition de chaque échantillon, nous avons générée un histogramme empilé ordonné selon le pourcentage de cellules immatures (CD34+ HSC et MultiLin).

Cela a indiqué que la distribution des types cellulaire est une source majeure d'hétérogénéité d'expression et varie selon les échantillons, comme prévu. La composition des échantillons varie considérablement entre les sujets, en particulier en ce qui concerne la fraction de cellules définies par les deux lignées majoritaires myéloïde et progénitrice. Ceci reflète clairement un gradient de différenciation ou de maturation des cellules immatures vers la lignée myéloïde mature (Figure 17 B). Nous pouvons constater que les échantillons appartenant à la classe FAB M5 sont les plus matures. De plus, nous avons remarqué que le groupe des monosomies 5/7 sont les plus immatures, par contre le groupe MLL est le groupe avec le plus de maturation.

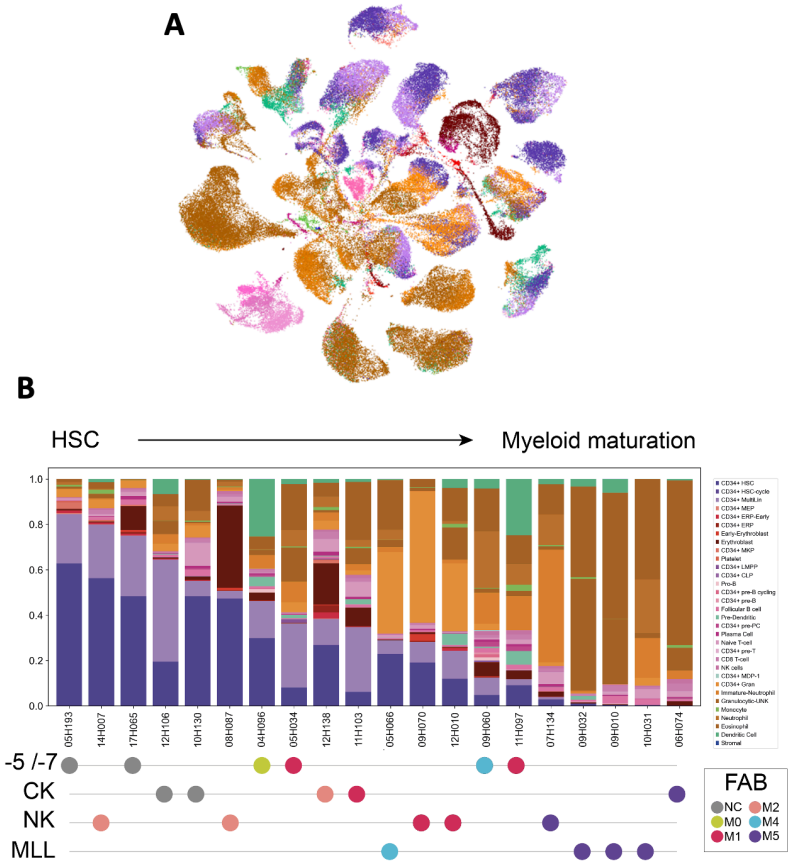

Figure 17. – Annotation des données scRNA-seq de la cohorte totale de LMA.

(A) Visualisation UMAP des données de 115K cellules annotées selon leur type cellulaire. (B) Histogramme empilé présentant les proportions de type cellulaire par échantillon. Le groupe génétique ainsi la classification FAB est indiquée sous chaque patient.

3.3 Performance de détection des variations de petites tailles dans les données scRNA-seq

Maintenant que la cohorte scRNA-seq de LMA a été annotée, nous avons évalué la possibilité de détecter des mutations somatiques dans nos données scRNA-seq de LMA. La capacité de détecter des mutations dans les données scRNA-seq dépend d'un certain nombre de variables, notamment la fréquence allélique du variant VAF, le niveau d'expression du gène muté, la position de la mutation dans le transcrit, la profondeur de séquençage et le nombre de cellules séquencées.

Au total, nous avons analysées la couverture de séquençage de 42 gènes impliquées dans la LMA (Annexe).

La figure 18 est un exemple qui illustre bien le résultat obtenu. Ici, nous avons observé une couverture de séquence loin des extrémités 3' des gènes qui peut dépendre de plusieurs facteurs (la longueur des exons, la longueur du gène …). Prenons l'exemple des mutations du gènes *KRAS* : les mutations situées dans l'exon 4 sont plus couvertes (entre 100 à 1000 fois) que celles de l'exon 2 et 6 (< 10 fois). Les gènes *NPM1* et *CALR* sont bien couverts dans la majorité de leurs exons (Figure 18). La moyenne de couverture de séquençage de tous les gènes varie entre 0 et 5000 fois avec une variation moyenne entre les 20 échantillons de 100 fois.

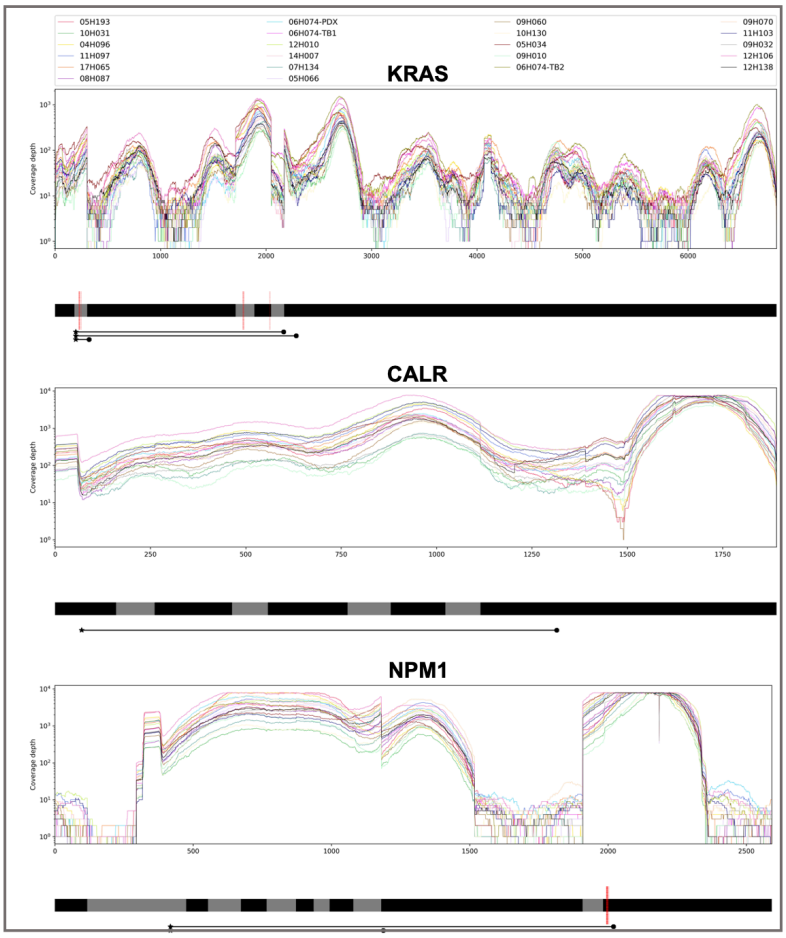

Figure 18. – Couverture de séquençage de certains gènes dans les données scRNA-seq de LMA.

Chaque ligne colorée représente un échantillon. Les boites noires et grises définissent les exons des gènes. Les lignes noires continuent en bas correspondent aux transcrits les plus long des gènes avec une étoile pour définir le début et la fin. Les barres verticales rouges présentent les positions spécifiques des mutations somatiques.

Une vue d'ensemble de la couverture de 500 mutations somatiques situées sur les 42 gènes sélectionnés est présentée dans la figure 19. L'analyse de ces données montre que les mutations des gènes *NPM1, U2AF1, SMC3, EZH2, RAD21, KRAS et PTPN11* sont les plus couvertes (plus que 100 fois) et qui peuvent être facilement détectées dans les données scRNA-seq même avec une chimie V3. La mutation NPM1 située dans la région 3' du gène est la plus couverte parmi toutes les mutations étudiées.

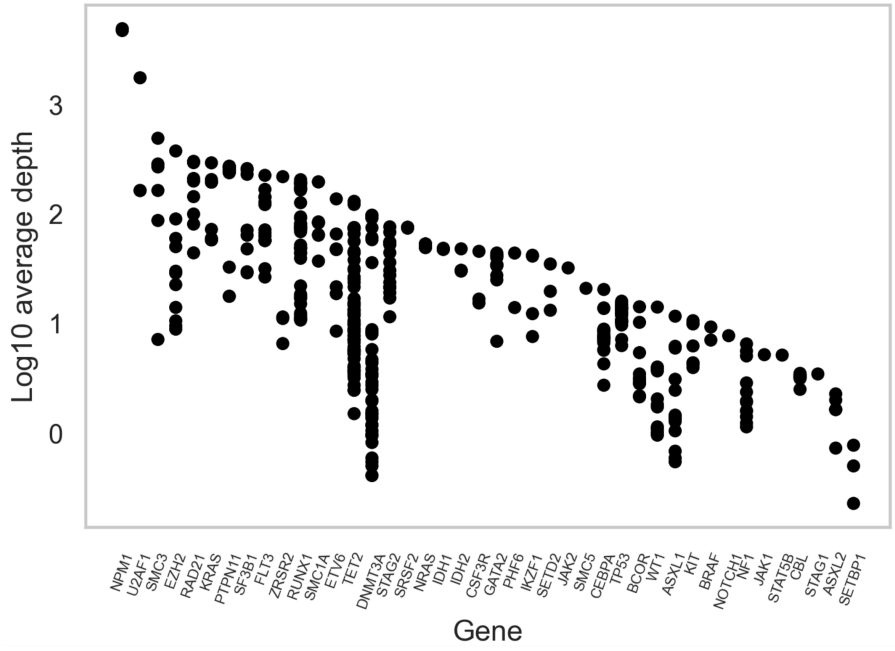

Figure 19. – Couverture de séquençage des mutations somatiques des gènes impliquées dans la LMA.

Chaque point correspond à une mutation somatique. Les gènes sont ordonnés selon la couverture maximale de leur mutation.

3.4 Détection des sous-clones LMA

3.4.1 Utilisation des variations de petite taille pour distinguer les cellules tumorales des cellules normales

De ce fait, nous nous sommes intéressés à explorer ces variants pour quelques échantillons pour voir si leur distribution est spécifique à un type cellulaire bien déterminé, sous-clonale. Une approche directe impliquerait de combiner la détection de mutations dans cellules uniques avec le regroupement basé sur l'expression.

Nous avons superposé les données de mutation sur les projections UMAP en mettant en évidence les cellules mutantes et les cellules un allèle sauvage (Figure 20). Pour une mutation hétérozygote, il y a 50% de chance que le transcrit observé soit muté, et 50% de chance qu'il soit de type sauvage, conduisant au phénomène connu sous le nom de décrochage allélique. Ceci a deux conséquences principales : premièrement, il est impossible de conclure qu'une cellule est de type sauvage ; d'autre part, la sensibilité de détection des mutations est réduite d'un facteur de 2 [100]. Ainsi, sauf erreur de séquençage, on peut en principe classer le génotype d'une cellule comme « mutant » s'il contient un ou plusieurs transcrits mutants, et « sauvage » s'il n'en contient pas. Nous avons donc étiqueté une cellule « mutante » si elle contenait au moins une « reads » contenant un variant, et « sauvage » si seules des « reads » de type sauvage ou aucune « reads » n'étaient détectées. A l'aide cette approche, nous avons détecté des cellules exprimant des mutations dans plusieurs gènes, dont *NPM1, CLAR et KRAS* (Figure 20). Prenons l'exemple de la mutation somatique hétérozygote Ala146Val du gène *KRAS* détectée dans l'échantillon 09H060 (Figure 20 A). Malgré sa couverture moyenne, nous avons constaté que, la mutation est distribuée de part et d'autre dans les groupes de cellule progénitrice immature (CD34+ HSC et MultiLin) et de cellule myéloïde mature (Granulocyte/Neutrophile). Concernant la délétion de la mutation L367fs*46 du gène *CALR* était détectable dans l'échantillon 12H138 (Figure 20 B). En se basant sur l'annotation de type cellulaire de cet échantillon, nous avons identifié deux populations de cellules progénitrices immatures différentes. Ces populations contiennent des cellules mutantes du gène *CALR*. Cependant, la présence de la mutation *CALR* n'explique pas cette structure. Nous avons aussi analysé les données de l'échantillon 07H134 avec une mutation hétérozygote Trp288fs du gène *NPM1* (Figure 20 C). La majorité des cellules de cet échantillon avaient la mutation. Aucune différence entre les types cellulaires n'a été détectée. Cela a démontré trois points clés : premièrement, les groupes de cellules qui présentent uniquement l'allèle sauvage (généralement les lymphocytes) sont des cellules normales non leucémiques. Deuxièmement, les cellules des *clusters* enrichis en mutations sont également susceptibles d'être des cellules LMA, même si elles ne contiennent aucune mutation détectable. Troisièmement, les mutations somatiques n'étaient pas concentrées dans des *clusters* cellulaires spécifiques.

Globalement, cette approche de l'identification des cellules LMA en combinant les données d'expression et de mutation, nous a permis d'identifier les cellules normales des cellules LMA. De plus, nous avons conclu que les mutations somatiques arrivent très tôt, touchent les progénitures et se propagent dans la lignée myéloïde anormalement différenciées. Nous avons pu détecter dans une certaine mesure certaines mutation chez les 20 LMA étudiées. Mais, en aucun cas nous avons trouvé une mutation qui affectait seulement un sous ensemble des cellules leucémiques. Ceci ne présume pas que les échantillons ne possèdent pas une sous-clonalité mais probablement l'information génétique peut être dérivée d'un grand nombre de cellules.

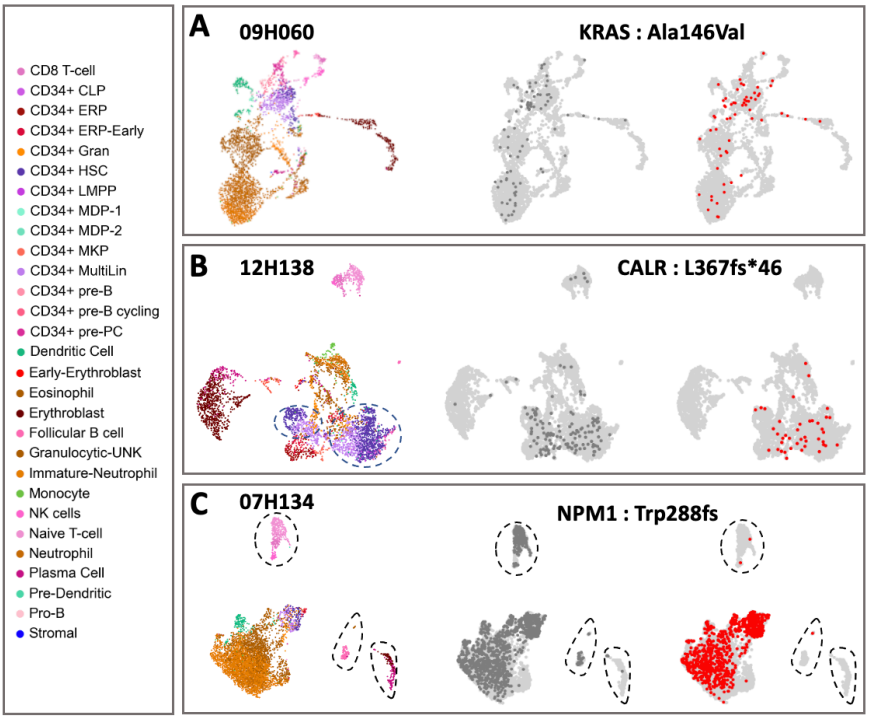

Figure 20. – Détection et interprétation des mutations dans trois échantillons scRNA-seq de LMA.

Les UMAPs de gauche présentent les cellules colorées selon les types cellulaires. Les UMAPs de droites et de milieu illustrent les cellules colorées selon le génotype à l'échelle unicellulaire sur les sites des mutations (A) *KRAS* Ala146Val (09H060), (B) *CLAR* L367fs*46 (12H138) et (C) *NPM1* Trp288fs (07H134). Gris clair, pas de couverture. Rouge, cellule avec au moins un allèle muté détecté ; gris foncé, cellule présente seulement l'allèle sauvage. Les lignes bleues discontinues (B) définissent les deux populations de cellules progénitrices immatures. Les lignes noires discontinues (C) définissent les cellules normales non leucémiques.

3.4.2 Utilisation des CNV pour détecter les sous-clones LMA

3.4.2.1 Résultats de l'optimisation de l'approche

Une approche complémentaire à la détection des sous-clones basée sur les SNV consiste à détecter les CNV en se basant sur l'expression des gènes pour identifier les cellules tumorales aneuploïdes et délimiter la sous-structure clonale de différentes sous-populations qui coexistent au sein du même échantillon. Pour relever ce défi nous avons testé CopyKAT. Avant d'entamer l'analyse de la cohorte avec l'outil CopyKAT, un échantillon, sélectionné selon son caryotype, a fait l'objet de plusieurs essais afin de déterminer quelles cellules nous allons utiliser comme une référence pour la normalisation et le calcul des CNV.

Pour se faire, nous avons testé trois approches : le mode automatique, les lymphocytes de chaque échantillon et 1500 cellules immatures des échantillons LMA à caryotype normal comme une référence pour prédire les CNV dans les données scRNA-seq. L'échantillon 17H065 qui est utilisé pour l'optimisation, appartient au groupe des monosomies 5. Selon le caryotype, 21 des cellules analysées de cet échantillon sont des cellules leucémiques ayant une perte de chromosome 5 et un gain du chromosome 8. La figure 21 illustre les résultats obtenus au cours de l'optimisation.

Ces résultats montrent clairement que les l'algorithme CopyKAT n'était pas capable de prédire les aberrations chromosomiques automatiquement (Figure 21 A) ou encore avec les lymphocytes (Figure 21 B). Cependant, grâce aux cellules immatures LMA avec un caryotype normal, nous avons réussi à avoir un bon signal au niveau du chromosome 5 et un signal au niveau du chromosome 8 qui correspondent à la perte de chromosome 5 et le gain du chromosome 8 prévu (Figure 21 C). Les cellules diploïdes (en vert) dans la figure 21 C correspondent aux cellules immatures LMA à caryotype normal qui ont été utilisées comme une référence pour prédire les CNV dans les cellules de cet échantillon (les cellules aneuploïdes en orange). Ce résultat est en parfaite adéquation à ce que nous observons au caryotype de cet échantillon. Par contre, l'inefficacité des modes automatique et lymphocyte de CopyKAT peut être expliquée d'une part par le fait que certains échantillons ne contiennent pas de cellules normales diploïdes. Ces

derniers seront détectés automatiquement par CopyKAT. D'autre part, il y a une différence dans Les niveaux d'expressions des gènes entre les lymphocytes normaux et les blastes leucémiques.

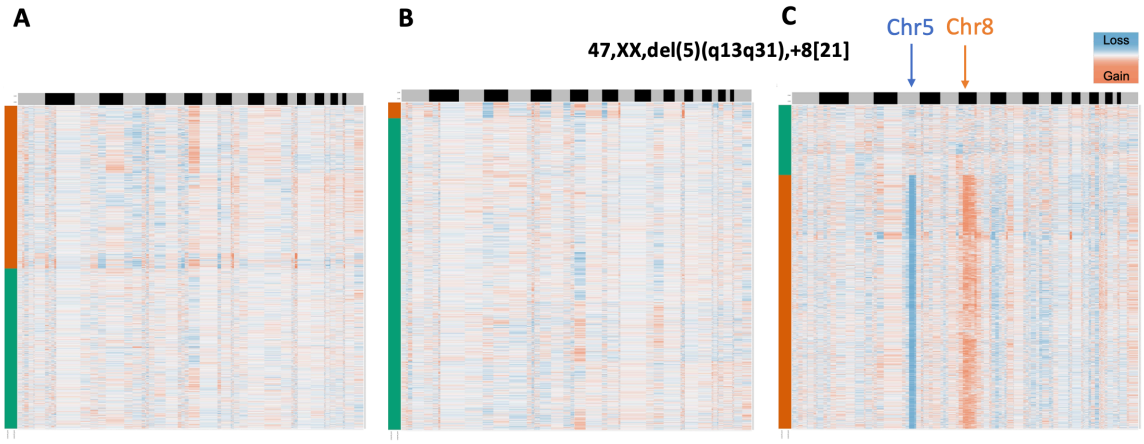

Figure 21. – Optimisation de la détection des Profils de nombre de copies estimés à partir des données scRNA-seq d'un échantillon LMA à l'aide de CopyKAT.

« Heatmap » groupée de 5000 profils de nombre de copies scRNA-seq de l'échantillon 17H065, estimée avec le mode automatique (A), les lymphocytes de l'échantillon (B) et les cellules immatures NK-LMA (C). Les boites noires et grises correspondent aux chromosomes. L'axe des y correspond aux cellules ; les cellules diploïdes sont colorées en vert et les cellules aneuploïdes sont colorées en orange. Le caryotype est rapporté en gras.

3.4.2.2 Résultats de la validation de l'approche

Par la suite, nous avons utilisé les données scRNA-seq de l'échantillon 09H060 afin de valider l'étape d'optimisation de l'approche afin de détecter les variations du nombre de copies à partir des données scRNA-seq d'un échantillon avec un caryotype caractérisé par une monosomie 7 et un gain du chromosome X. Comme le montre la figure 22, la perte du chromosome 7 et le gain du chromosome X se distinguent parfaitement sur le « heatmap » avec un signal bleu et rouge respectivement. Les deux aberrations chromosomiques touchent la même population ce qui est en concordance avec le caryotype : 6/21 cellules étaient aneuploïdes, l'équivalent de ~30% des cellules de l'échantillon.

Selon la figure 22 (A et B), cette population de cellule aneuploïde correspond aux cellules immatures, notamment les HSC CD34+ et les multiLin CD34+ ainsi les progénitures myéloïdes et les cellules dendritiques. Ces résultats ont été ensuite confirmés avec une analyse d'enrichissement des voies positionnelles qui montre clairement que les segments 7q36 ainsi Xq21 et Xq22 sont les plus enrichies. Dans ce cas, nous pouvons conclure que les CNV sont spécifiques aux cellules immatures seulement, ceci peut être un bon exemple de la sous-clonalité des CNV dans la LMA. Comme déjà présenté dans la partie précédente, cet échantillon présente le variant Ala146Val du gène *KRAS* détectée dans les populations immatures et myéloïdes matures. Donc, nous pouvons déduire que la mutation est survenue avant les aberrations cytogénétiques dans cet échantillon.

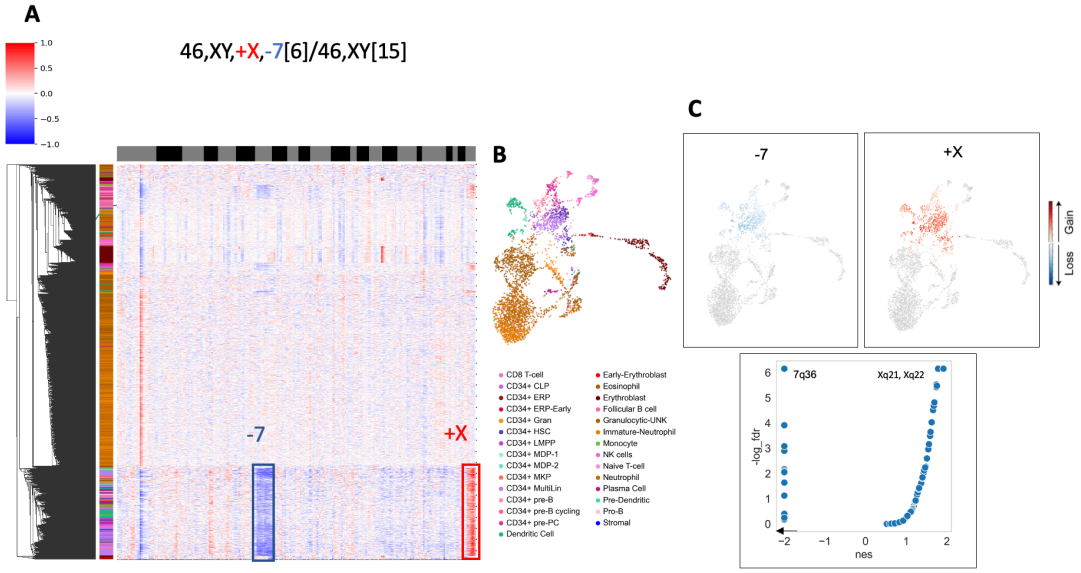

Figure 22. – Détection des Profils de nombre de copies estimés à partir des données scRNA-seq d'un échantillon LMA à l'aide de CopyKAT.

(A) « heatmap » groupée de 3500 profils de nombre de copies scRNA-seq de l'échantillon 09H060. Les boites noires et grises correspondent aux chromosomes. L'axe des y correspond aux cellules colorées selon les types cellulaires. Le caryotype est signalé en gras. (B) Visualisation en UMAP des cellules colorées selon les types cellulaires. (C) Les visualisations UMAP présentent la moyenne du nombre de copies de toutes les fenêtres des données de scRNA-seq estimé par CopyKAT. Les nuages de points présentent l'enrichissement des voies positionnelles par GSEA du vecteur de corrélation des transcriptomes avec le nombre de copie estimé par CopyKAT.

3.4.2.3 Détection des CNV dans la cohorte

3.4.2.3.1 Dans le groupe monosomie 5/7

Suite à cette validation de l'approche, nous avons analysé les dix échantillons de la cohorte avec une aberration du nombre de copies : 8 échantillons avec une monosomie 5 ou 7 et 2 échantillons avec un caryotype complexe. Nous avons commencé par l'analyse de 40 000 transcriptomes de cellule unique de huit échantillons de LMA : 4 avec une monosomie 7 et 4 avec une monosomie 5. Nous avons identifié avec succès des sous-populations de cellules tumorales aneuploïdes chez tous les individus (Figure 23 et 24). Les cellules tumorales prédites avaient des CNV à l'échelle du génome, y compris des pertes de 5q, 5p, 7p, 7q qui sont couramment signalées dans la LMA.

Toute la lignée des cellules lymphoïdes matures (lymphocytes B, lymphocytes T et NK) n'avait pas de profils de CNV récurrents. Ceci confirme encore notre approche. La projection UMAP des profils CNV (Figure 23 et 24) montrent clairement que certains échantillons (17H065, 11H103, 05H193, 12H106 et 04H096) présentent l'aberration de CNV dans la majorité de leurs cellules. Alors que les profils de CNV de l'échantillon 09H060 sont spécifiques à la population immature HSC. Ces résultats peuvent être expliquée par le fait que le premier groupe d'échantillon LMA appartient aux classes FAB les moins différenciés (NC, M0 et M1) et le deuxième groupe est classé entre M1 et M4 (Figure 17).

De plus, nous avons remarqué que dans certains cas (11H097, 05H034), le signal de la perte 5q n'était pas clair. Ceci peut être explique soit par un nombre limité de gène exprimé ou capturé dans les données scRNA-seq, ou encore la petite taille de la région perdu (ex. 5q22-5q33) comparant à la perte de tout le chromosome 7 chez les échantillons (09H060, 05H193, 12H106 et 04H096).

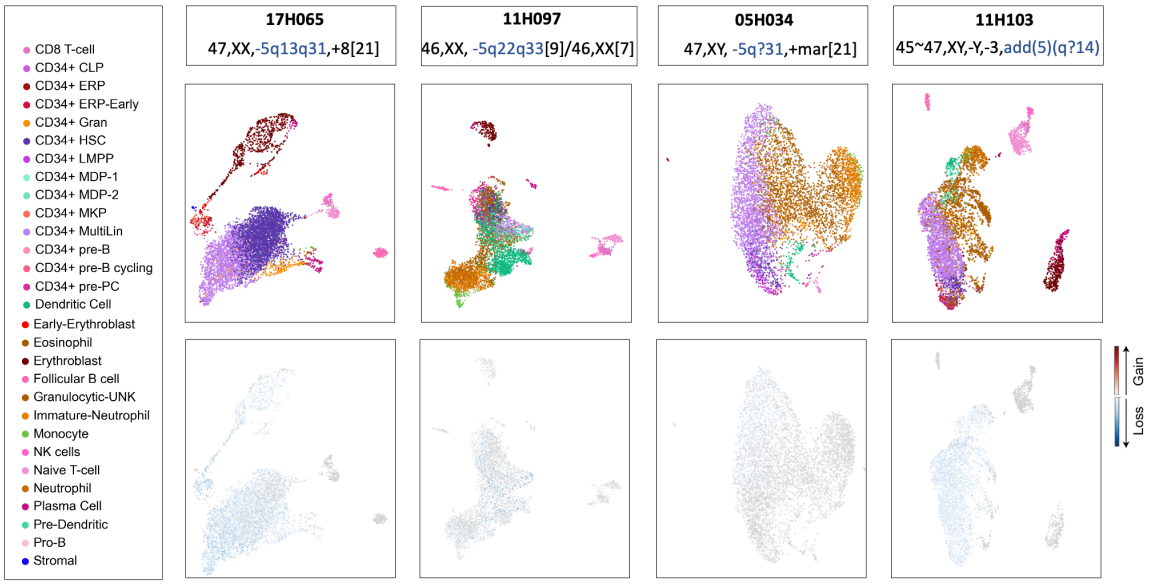

Figure 23. – Détection des profils de nombre de copies estimés à partir des données scRNA-seq des échantillons LMA avec monosomies 5 à l'aide de CopyKAT.

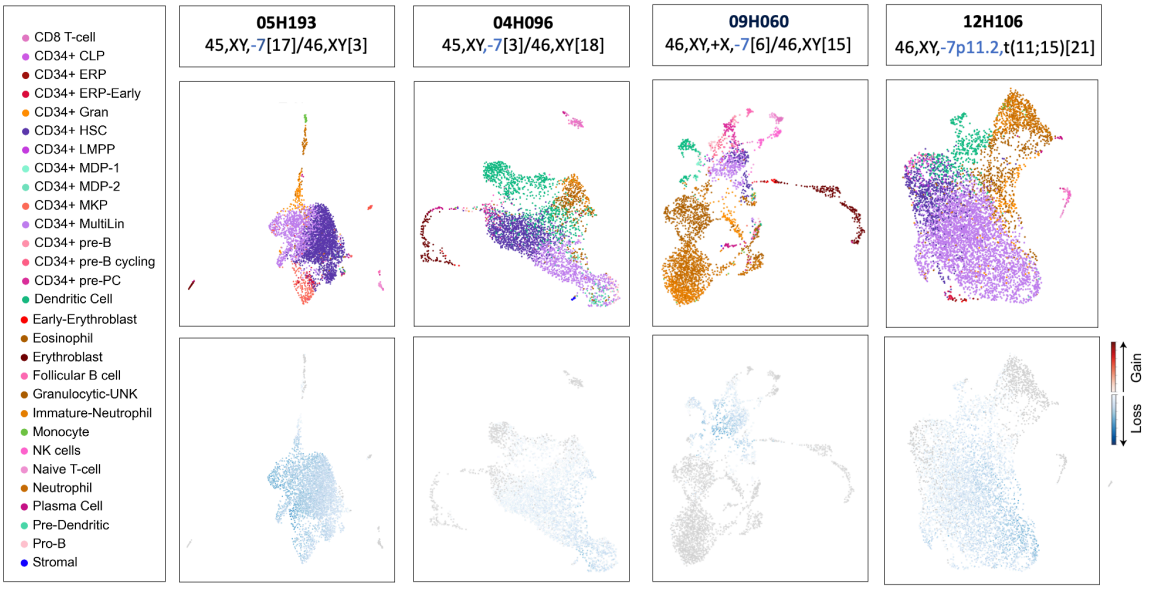

Figure 24. – Détection des Profils de nombre de copies estimés à partir des données scRNA-seq des échantillons LMA avec monosomies 7 à l'aide de CopyKAT.

Le caryotype de chaque échantillon est signalé en gras. Le panel en haut montre une visualisation en UMAP des cellules colorées selon les types cellulaires. Le panel en bas présente les visualisations UMAP des profils CNV estimée par CopyKAT.

3.4.2.3.1 Complexite clonale dans les LMA à caryotype complexe

Par la suite, nous avons analysé 10 000 transcriptomes de cellule unique de deux échantillons LMA avec un caryotype complexe. Pour l'échantillon 12H138 (Figure 25), en se basant sur l'annotation de type cellulaire de cet échantillon, nous avons identifié deux populations de cellules progénitrices immature différente (HSC) (Figure 25 A). Dans cet échantillon hébergeant une complexité clonale (voir caryotype, Figure 25 B), les données d'exome ont identifié des CNV pertinents, y compris un gain cytogénétiquement cryptique de chr3q. Les profils CNV déduits par CopyKAT ont montré que les cellules leucémiques de cet échantillon comprennent deux sous-clones génétiques majeurs, l'un comportant un gain du 1q et l'autre un gain 3q et une perte 7q (Figure 25 B).

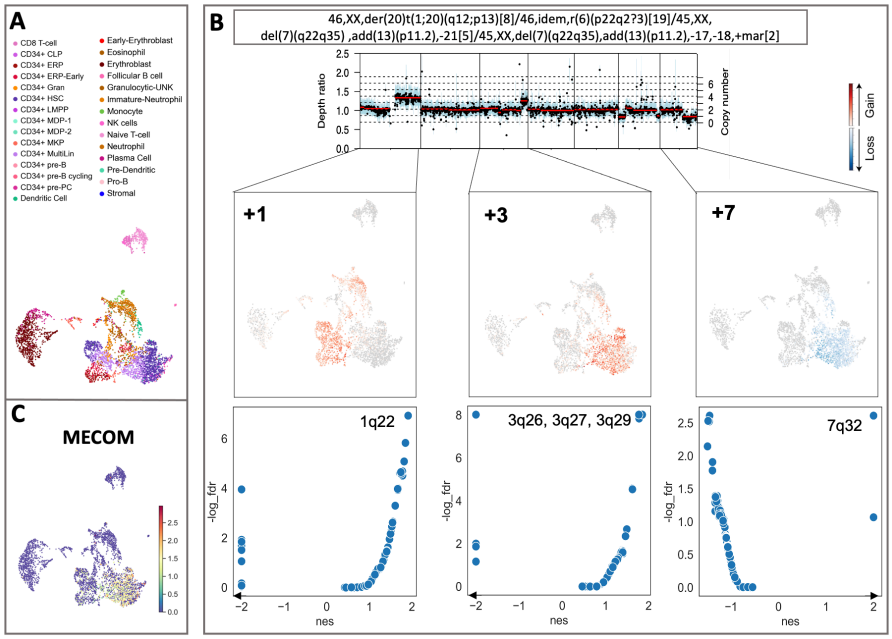

Figure 25. – Détection des profils de nombre de copies estimés à partir des données scRNA-seq de l'échantillon 12H138 à l'aide de CopyKAT.

(A) Visualisation en UMAP des cellules colorées selon les types cellulaires. (B) Le caryotype de l'échantillon est signalé. CNV détectés à l'aide des données WES. Les visualisations UMAP des profils CNV estimée par CopyKAT. Les nuages de points présentent l'enrichissement des voies positionnelles par GSEA du vecteur de corrélation des transcriptomes avec le nombre de copie estimé par CopyKAT. (C) UMAP colorée par l'expression *MECOM*.

Ces résultats ont été ensuite confirmés avec une analyse d'enrichissement des voies positionnelles qui montre clairement que les segments 1q22, 3q26, 3q27, 3q29 et 7q32 sont les plus enrichies. En tant qu'exemple, le gain du signal chr3q est fortement corrélé à l'expression du gène *MECOM*, un oncogène connu situé sur la bande chromosomique 3q26.2 (16éme gène le plus corrélé, figure 25 C).

Dans un autre échantillon (10H130), l'annotation de type cellulaire a identifié cinq populations de cellules progénitrices immatures différentes (HSC) (Figure 26 A). Dans cet échantillon hébergeant une complexité clonale (voir caryotype, Figure 26 B), les données d'exome ont identifié des CNV pertinents. Les profils CNV des cellules uniques déduits par CopyKAT ont montré que les cellules leucémiques de cet échantillon comprennent aussi cinq sous-clones génétiques majeures, y compris +3q, -5q, +8q, +11q et +13 (Figure 26 B).

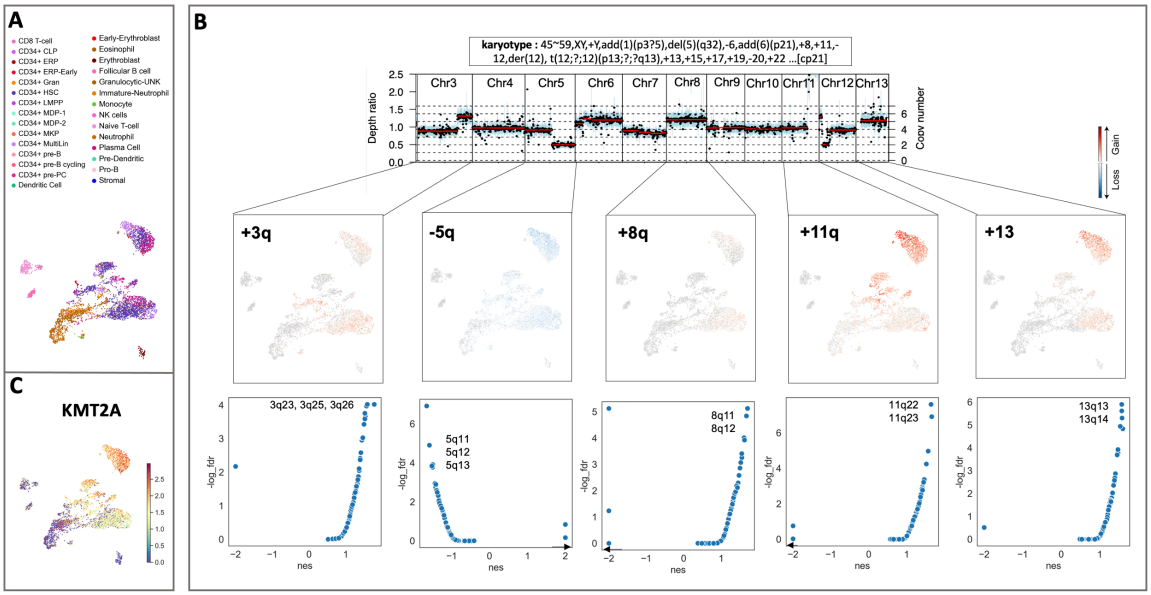

Figure 26. – Détection des Profils de nombre de copies estimés à partir des données scRNA-seq de l'échantillon 10H130 à l'aide de CopyKAT.

(A) Visualisation en UMAP des cellules colorées selon les types cellulaires. (B) Le caryotype de l'échantillon est rapporté. CNV détectés à l'aide des données WES. Les visualisations UMAP des profils CNV estimée par CopyKAT. Les nuages de points présentent l'enrichissement des voies positionnelles par GSEA du vecteur de corrélation des transcriptomes avec le nombre de copie estimé par CopyKAT. (C) UMAP colorée par l'expression *KMT2A*.

Ces résultats ont été ensuite confirmés avec une analyse d'enrichissement des voies positionnelles qui montre clairement que les segments 3q23, 3q25, 3q26, 5q11, 5q12, 5q13, 8q11, 8q12, 11q22, 11q23, 13q13 et 13q14 sont les plus enrichies. De plus, le gain du signal chr11q est corrélé à l'expression du gène *KMT2A*, un régulateur majeur de l'hématopoïèse réarrangé dans 10 % des LMA, situé sur la bande chromosomique 11q23 (Figure 26 C).

3.5 Caractérisations des sous-clones LMA

3.5.1 Analyse de l'expression différentielle des gènes

Grace à cette approche, nous avons comparé les profils d'expression des deux populations HSC-like dans deux contextes d'aneuploïdies distincts (-5 et -7). Pour se faire, nous avons effectué une analyse par expression différentielle entre le groupe des cellules HSC-like (-5) (G1) et le groupe des cellules HSC-like (-7) (G2). L'expression différentielle permet de trouver les gènes qui ont un taux d'expression différent et significatifs entre deux sous-groupes. Dans le contexte de nos deux sous-groupes G1 et G2, nous avons effectué l'analyse d'expression différentielle en comparant les taux d'expression génique de G1 comparés à G2.

Au final, parmi les 18000 gènes testés, nous avons sélectionné les gènes significativement surexprimés ou sous-exprimés ont été définis par $| \log2(FC) | > 1$ et un $FDR < 0.05$. Ces résultats, présentées dans la figure 27, montrent que 5 gènes : *LGALS1, PDLIM1, S100A4* (sous-exprimés), *MTRNR2L8, SPINK2* (surexprimés) dépassent les seuils.

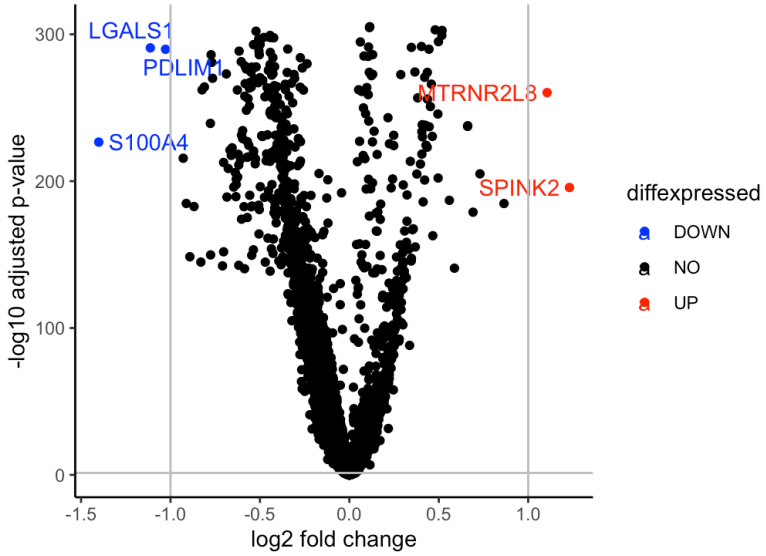

Figure 27. – Les gènes différentiellement exprimés entre les populations G1 et G2.

Graphique de volcan ou chaque point correspond à un gène. Les points rouges correspondent aux gènes surexprimés et les points bleues correspondent aux gènes sous-exprimes.

3.5.2 Analyse d'enrichissement des gènes

Finalement nous avons voulu regarder les processus biologique associés à la comparaison des gènes entre les deux groupes G1 et G2 testés. La liste de tous les gènes a été soumise pour les analyses d'enrichissement des voies KEGG, HallMark et GO pour comparer les voies enrichies dans la LMA. Pour le G1 des cellules HSC avec une monosomie 5, les GDE surexprimés suggèrent un enrichissement significatif dans les protéines du complexe CMH, l'activité des récepteurs chimo-attractants couplés à la protéine G, la liaison à l'antigène peptide, la liaison des chemokines, la régulation de la cytotoxicité à médiation des cellules T, la membrane du réticulum endoplasmique, la régulation du développement des cellules endothéliales, la régulation de la protéolyse de l'ecto-domaine des protéines membranaires, la fixation du récepteur de chemokine, l'angiogenèse, la présentation de l'antigène, la migration trans endothéliale des leucocytes, la réponse de l'interféron alpha, molécules d'adhésion cellulaire et complément.
De plus, les segment 5q15, 5q22 et 5q31 ont été enrichis suite à un enrichissement des voies positionnelles (Figure 28).

Pour le G2 des cellules HSC avec une monosomie 7, les GDE étaient enrichis en phosphorylation oxydative, la localisation des protéines sur la membrane, l'initiation de la traduction, le ribosome, les processus cataboliques de l'ARNm nucléaire et le ciblage des protéines sur la membrane (Figure 28). Avec cette approche, nous étions capables de comparer les différences transcriptomiques des deux populations HSC-like dans deux contextes d'aneuploïdies distincts (-5 et -7). Ceci est un exemple d'analyse qui peut être effectué suite à l'annotation et la détection des CNV à partir des données scRNA-seq.

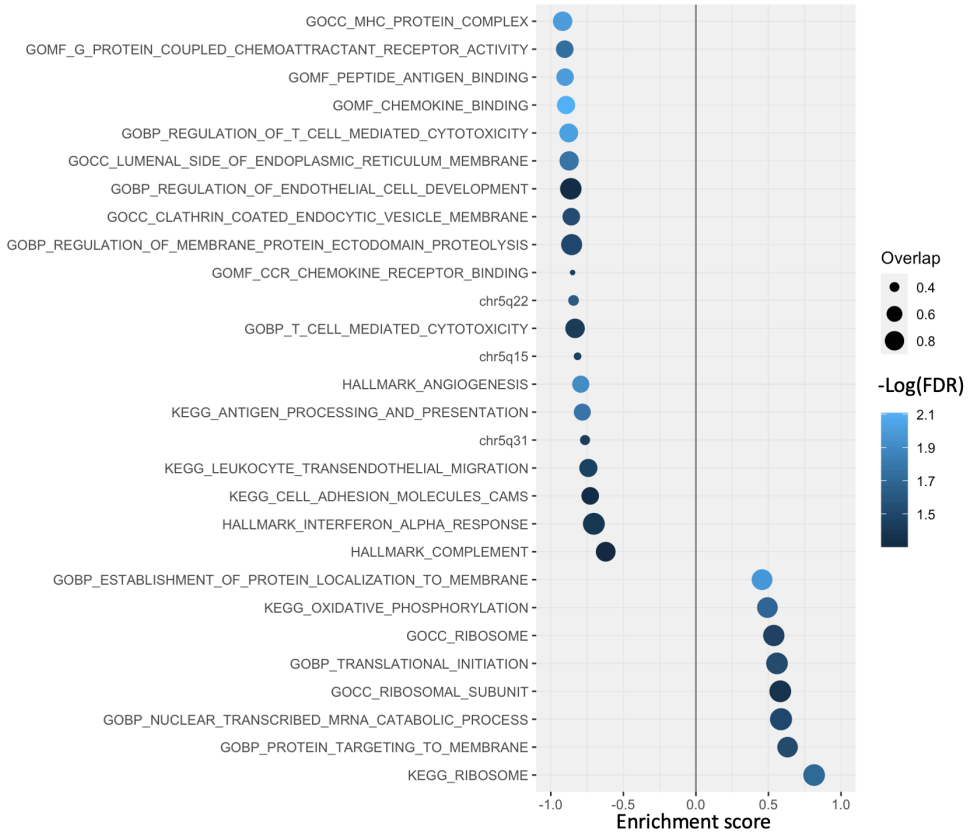

Figure 28. – Les voies d'enrichissement des gènes différentiellement exprimés entre G1 et G2.

GO : ontologie génétique ; KEGG : Encyclopédie de Kyoto des gènes et des génomes ; MEC : matrice extracellulaire

Chapitre 4 -Discussion

4.1 Une approche d'identification de sous-groupes moléculaires

4.1.1 Annotation des types cellulaires

Le premier objectif de cette étude était d'annoter les populations cellulaires des échantillons LMA de notre cohorte de 20 patients.

Actuellement, l'annotation des types cellulaires des données scRNA-seq, après un regroupement non supervisé est principalement effectuée manuellement. La limitation de la procédure manuelle rend impossible la génération de résultats d'annotation de haute qualité, reproductibles et standardisés pour le nombre croissant d'ensembles de données scRNA-seq. De plus, l'annotation manuelle est limitée par les connaissances des biologistes, et peut être déroutées par l'expression aberrante de gènes si seulement un petit nombre de marqueurs est utilisé. Pour relever ce défi, nous avons entrainé trois modèles de classification supervisée en utilisant les données scRNA-Seq provenant de la moelle osseuse de huit donneurs sains du portail de données HCA. Le classificateur avec la meilleur précision (RF) a été sélectionné pour annoter les 115 000 cellules LMA après les étapes de prétraitement, nettoyage, normalisation et regroupement. Dans l'annotation de type cellulaire, une stratégie consiste à utiliser des gènes spécifiquement exprimés dans un *cluster* de cellules pour marquer le type de cellule. Cependant, l'utilisation de quelques gènes marqueurs n'est souvent pas suffisante pour distinguer un *cluster* cellulaire des autres. De plus, l'utilisation de l'ensemble des ensembles de gènes exprimés peut diminuer la capacité de trouver les vrais modèles au sein de chaque groupe de cellules. Par conséquent, nous avons utilisé les gènes les plus variables pour éviter l'influence des gènes exprimés de manière ubiquitaire et sélectionner les gènes appropriés pour calculer le score optimal dans le modèle d'annotation. Il existe encore certaines limitations, qui peuvent influencer la précision de l'annotation du type de cellule à l'aide de cette méthode.

Premièrement, le nombre de gènes a eu un impact considérable sur les résultats de l'annotation des types cellulaires vu que certains types de cellules soient inclassables en raison du manque de marqueurs appropriés. Une sélection des GDE de chaque type cellulaire pourrait être effectuée pour résoudre ce problème. Deuxièmement, la précision de l'annotation des cellules LMA repose fortement sur un entrainement sur un jeu de données de cellules saines. Cependant, l'expression des gènes dans le contexte de LMA peut être aberrante ce qui cause un étiquetage erroné. Des efforts supplémentaires pourraient être faits pour améliorer la capacité d'annotation de RF en prenant en compte plus d'informations (par exemple, utiliser l'information du transcriptome au complet au lieu des GDE et entrainer un nouveau modèle RF sur un jeu de données LMA déjà annoté par le premier classificateur après validation manuelle). Mais un tel jeu de données n'existe pas pour le moment. Le seul jeu de données disponible est celui de Peter van Galen [130] avec (16 patients, 30 712 cellules) et ne représente pas la diversité génétique des LMA. De plus, l'équipe de Peter van Galen [130] a utilisé la classification aléatoire basée sur les forêts à deux fins différentes : pour prédire la similitude de cellules individuelles aux 15 types cellulaires différents détectés dans la BM saine (classificateur 1), et pour prédire si une seule cellule d'un échantillon tumoral est maligne ou normale (classificateur 2). Pour former le premier classificateur, ils ont d'abord effectué une étape de sélection de caractéristiques pour sélectionner les gènes les plus informatifs parmi tous les 14 554 gènes exprimés dans l'ensemble de données (expression moyenne > 0,01). La sélection des fonctionnalités a été effectué en entraînant un classificateur de forêt aléatoire « extérieur » sur tous les gènes exprimés. Ils ont formé 1 000 arbres, en utilisant un sous-ensemble aléatoire de 50 cellules de chaque type de cellule pour chaque arbre. Sur la base du rapport d'importance globale des gènes dans le classificateur « externe », ils ont ensuite sélectionné uniquement les 1 000 gènes les plus informatifs pour l'entraînement du classificateur « interne ». L'erreur de classification erronée des cellules qui n'ont pas été utilisées pour l'apprentissage d'un arbre donné était inférieure de 20 % pour le classificateur « interne » que pour le classificateur « externe », justifiant l'utilisation d'une étape de sélection initiale des caractéristiques [130].

Pour évaluer indépendamment si le classificateur de forêt aléatoire était un choix approprié pour classer les types de cellules, ils ont aussi comparé les performances du premier classificateur de forêt aléatoire (RF) à un classificateur indépendant de machine à vecteurs de support (SVM).

Bien que le classificateur SVM ait généré des résultats raisonnables, il n'a pas été aussi performant que le classificateur de forêt aléatoire en validation croisée. De ce fait, ils ont conclu que l'algorithme de forêt aléatoire est un choix approprié pour classer les types de cellules dans les données scRNA-seq [130]. Nous pensons aussi que RF est un ajout important à la boîte à outils utilisée pour les études monocellulaires et améliorera considérablement notre efficacité et notre capacité à explorer les types de cellulaires de LMA. Mais, une amélioration du RF serait nécessaire.

4.1.2 Détection des sous-clones LMA

4.1.2.1 Utilisation des SNV pour distinguer les cellules tumorales des cellules normales

4.1.2.1.1 Performance de détection des variants dans les données scRNA-seq

L'objectif subséquent était d'évaluer l'utilité des données scRNA-seq 10x pour la détection de variants somatiques dans des échantillons de moelle osseuse LMA cryoconservés. Nous avons approché cet objectif en utilisant l'algorithme Freebayes pour l'appel des variants et l'outil Vartrix pour le genotypage des cellules.

Nous avons remarqué que la technologie *10X Genomics Chromium* Single Cell 3' produit une couverture de transcription étonnamment élevée loin des extrémités 3' des transcriptions. Bien que cette couverture distale soit clairsemée, elle est suffisante pour la détection des variants dans des cellules individuelles : les SNVs étaient détectables dans 20 % des cellules de nos échantillons, en moyenne, mais souvent dans très peu de cellules. Cependant, un certain nombre de cellules mutées est indispensable pour affirmer avec confiance qu'il s'agit bien d'une population mutée.

Basé sur les profils de couverture à travers l'ensemble des gène étudiés, la mutation la mieux couverte *NPM1* pourrait être regardée dans un plus grand nombre d'échantillon pour voir si cette couverture est suffisante pour confirmer la sous clonalité avec confiance. Pour les autres mutations, les analyses dépendront réellement des nouvelles avancées technologiques pour y parvenir. Des études antérieures ont démontré que les SNV peuvent également être identifiés avec une sensibilité élevée dans des transcrits complets de dizaines à des centaines de cellules uniques à l'aide de techniques de plusieurs plate-formes : Smartseq et chromium [127].

Une application courante de la détection de variants dans les données scRNA-seq consiste à distinguer les cellules tumorales des cellules normales dans des échantillons hétérogènes. Cependant, étant donné que les cellules malignes peuvent avoir des profils d'expression qui imitent des cellules normales plus fortement différenciées, les données d'expression génique seules ne sont pas suffisantes pour identifier les cellules LMA. De plus, les cellules LMA affichent parfois une infidélité de lignée, où certaines cellules LMA affichent les caractéristiques de types cellulaires différenciés d'autres lignées. L'hétérogénéité transcriptionnelle dans les échantillons de LMA provient clairement de plusieurs sources, y compris les états de différenciation de cellules normales et tumorales, états du cycle cellulaire et des mutations présentes dans des sous-ensembles de cellules. L'intégration de la détection de variants dans l'analyse scRNA-seq permet de distinguer ces sources en facilitant la distinction entre les cellules tumorales et normales, et en révélant des corrélations entre l'hétérogénéité mutationnelle et transcriptionnelle [100].

La détection de cellules avec des mutations exprimées dans les données scRNA-seq est soumise à plusieurs limitations. L'abandon (y compris l'abandon de la transcription et l'abandon allélique) se produit avec la plupart des plates-formes scRNA-seq. En conséquence, il est impossible de déterminer si une cellule est vraiment de type sauvage pour une mutation donnée. De plus, le décrochage réduit la sensibilité de détection des mutations par un facteur de deux. La couverture partielle des transcriptions est spécifique aux plates-formes biaisées telles que la plate-forme Chromium, et limite également la sensibilité de la détection des variants.

De plus, la couverture diminue de manière non linéaire sur toute la longueur de la transcription, de sorte que certaines variants sont beaucoup plus facilement détectables que d'autres. L'utilité de cette approche dépend donc de la composition mutationnelle spécifique de l'échantillon en question, et sera probablement plus performante pour d'autres types de tumeurs, qui ont presque toutes des charges de mutation plus élevées que la LMA [100]. Un certain nombre d'autres approches pour identifier les mutations exprimées dans les données de scRNA-seq ont été décrites [131–133,130].

Chaque méthode a des forces et des faiblesses différentes qui devraient influencer le choix de la plate-forme pour une question expérimentale spécifique. Les variables clés incluent la taille d'insertion de la librairie, le biais de fin et la complexité, la profondeur de séquençage et la longueur de « reads », le taux d'abandon et le débit. De plus, les technologies qui permettent le séquençage simultané de l'ADN et de l'ARN de cellules individuelles, telles que G&T-seq [134], peuvent devenir très puissantes avec un débit accru. Le rythme rapide des progrès technologiques dans ce domaine augmentera probablement la puissance de scRNA-seq pour identifier et distinguer les différentes sources d'hétérogénéité transcriptionnelle dans les échantillons de tumeurs primaires.

4.1.2.2 Utilisation des CNV pour détecter les sous-clones LMA

Une approche alternative à la détection des sous-clones basée sur les SNV, est celle qui consiste à détecter les CNV en se basant sur l'expression des gènes pour identifier les cellules tumorales aneuploïdes et délimiter la sous-structure clonale de différentes sous-populations qui coexistent au sein du même échantillon. Pour relever ce défi nous avons testé CopyKAT. C'est une approche de segmentation bayésienne intégrative pour quantifier les profils de nombre de copies génomiques à partir de données scRNA-seq à haut débit. Cependant les paramètres par défaut de CopyKAT ne parvenaient pas analyser et prédire les CNV dans la LMA. Suite à l'optimisation et la validation de l'approche, nous avons été capable les dix échantillons de la cohorte avec une aberration du nombre de copies : 8 échantillons avec une monosomie 5 ou 7 et 2 échantillons avec un caryotype complexe. Un total de 50 000 transcriptomes de cellule uniques a été analysé.

Nous avons identifié avec succès des sous-populations de cellules tumorales aneuploïdes chez tous les individus. Les cellules tumorales prédites avaient des CNV à l'échelle du génome, y compris des pertes ou des gains de 1q, 3q, 5q, 5p, 7p, 7q, 8q, 11q et 13 qui sont couramment signalées dans la LMA.

Deux méthodes antérieures ont également été développées pour estimer les altérations du nombre de copies [135,133]. InferCNV [135] utilise une fenêtre mobile moyenne d'expression génique après avoir exclu les gènes fortement et faiblement exprimés ; cependant, inferCNV a une capacité limitée à résoudre avec précision les points de rupture chromosomiques. Une autre méthode, HoneyBadger [139], a été conçue pour prédire les variations du nombre de copies à partir des données scRNA-seq en analysant conjointement le déséquilibre allélique de nombreux sites variants dans des *clusters* regroupées de cellules individuelles, mais elle dépend fortement de l'obtention de données de couverture complète du gène. Ainsi, une limitation de ces méthodes précédentes est qu'elles ne sont pas compatibles avec les méthodes scRNA-seq 3' (10X Genomics, Drop-Seq, InDrop) [54,55,57] qui sont maintenant largement utilisées dans le domaine de la génomique mais étaient à la place développée pour les méthodes scRNA-seq de première génération, telles que Fluidigm [136] et SMART-seq2 [52]. En revanche, CopyKAT est compatible avec les méthodes scRNA-seq à haut débit qui génèrent des données éparses (100 000 « reads » par cellule) sur des milliers de cellules qui sont séquencées en parallèle et est également compatible avec les données des scRNA-seq de première génération [127].

Par contre, CopyKAT est principalement limité à la détection d'événements CNV et ne peut pas détecter d'autres événements génomiques qui contribuent à la diversité génomique, y compris les réarrangements structurels chromosomiques [127]. De plus, CopyKAT ne peut pas fournir d'informations fiables sur le nombre de copies sur les génomes de cellules individuelles avec des génotypes uniques en raison de la variabilité technique des données scRNA-seq 3'. Pour résoudre ce problème, nous avons dû utiliser les données de séquençage « bulk » d'exome appariées tumoraux et sains des mêmes échantillons, générées grâce à l'outils Sequenza, pour débruiter les différences subtiles et délimiter les fenêtres d'expression estimées avec CopyKAT.

Une autre application de cette approche est la délimitation de la sous-structure clonale dans la LMA sur la base des différences dans les CNV. Dans le cas des deux échantillons à caryotype complexe, nous avons identifié des sous-populations majeures dans chaque tumeur qui différaient par des événements CNV distincts. De plus, nous montrons qu'à partir de ces données, nous pouvons lier les génotypes des sous-clones à leurs phénotypes (programmes transcriptionnels) pour comprendre comment les altérations génomiques ont influencé différentes propriétés tumorales. Fait intéressant, dans ces deux échantillons LMA, nous avons identifié deux sous-clones avec une amplification des oncogènes *MECOM* et *KMT2A* qui sont corrélées au gain du 3q et 11q, respectivement. Une surexpression de ces oncogènes est observée chez environ 10 % des patients atteints de LMA et est associée à la chimiorésistance et de mauvais pronostic. Ces analyses montrent que la surexpression de ces oncogènes s'accompagne souvent de modifications caryotypiques supplémentaires [137].

Ces sous-clones nécessitent plus d'exploration s'ils s'avèrent être des sous-populations courantes dans de plus grandes cohortes d'individus atteints de LMA dans les études futures. En résumé, cette approche fournit une méthode automatisée puissante pour classer les cellules tumorales et normales dans des données scRNA-seq. Cette approche nous permet d'identifier les populations qui possèdent différents CNV, s'il y a lieu, dans un échantillon LMA. De plus, Ceci nous permet maintenant d'étudier les profils d'expression propres à ces différents CNV, notamment entre 2 populations phénotypiquement apparenté (ex HSC-like).

4.1.3 Caractérisation des sous-groupes LMA

L'objectif de caractériser les sous-groupes potentiels identifiés par la méthode décrite précédemment est fait grâce à l'analyse par expression différentielle des différents sous-groupes. Pour cela, nous avons comparé l'expression entre les deux sous-groupes G1 (cellules immatures HSC avec une monosomie 5) et G2 (cellules immatures HSC avec une monosomie -7). On remarque que seulement 5 gènes différentiellement exprimés dépassent la limite de signifiance : *LGALS1, PDLIM1, S100A4, MTRNR2L8* et *SPINK2*.

Concernant le gène *SPINK2*, également connu sous le nom d'inhibiteur de plasma séminal humain II, appartient à la famille SPINK. L'expression de *SPINK2* est étroitement associée au développement du cancer. *SPINK2* sert de marqueur de classification pour le lymphome, ainsi que de marqueur prédictif de la réponse au traitement contre le cancer [138].
Chen et collègues ont rapporté que *SPINK2* était significativement élevé dans la majorité des lignées cellulaires leucémiques étudiées et jouait un rôle important dans la progression tumorale et la réponse au traitement [139]. Ainsi, nous émettons l'hypothèse que *SPINK2* peut jouer un rôle important dans le processus de LMA. Cependant, le mécanisme moléculaire de *SPINK2* affectant le processus tumorigène reste incertain. Ainsi, d'autres études fonctionnelles sont nécessaires pour étudier en profondeur comment *SPINK2* affecte le développement de la LMA. De plus, *MTRNR2L8* est un gène qui inhibe l'apoptose et favorise la progression tumorale dans le cancer du sein triple négatif [140].

La famille des gènes *PDLIM* joue un rôle crucial dans de nombreux processus biologiques fondamentaux tels que la polarité cellulaire, les jonctions intercellulaires, la reconnaissance de cellules immunitaires, le contrôle de la prolifération et de la migration cellulaire et des activités réduites observées dans certains processus pathologiques, notamment l'oncogenèse [141]. Une étude récente a indiqué qu'une expression élevée de *PDLIM2* et *PDLIM7* étaient des facteurs de mauvais pronostic pour la LMA et elle peut favoriser cette malignité en activant la voie Ras-ERK [142].

S100A4 est un gène qui appartient à la famille multigénique S100 des protéines de liaison au calcium. Cette famille est impliquée dans divers processus cellulaires, notamment la régulation de la prolifération, la progression du cycle cellulaire, l'apoptose, la différenciation, l'homéostasie du Ca2+, la migration, l'adhésion et la transcription. *S100A4* a déjà été associé à un mauvais pronostic dans plusieurs tumeurs solides et dans la leucémie.
Ces données suggèrent également que *S100A4* est essentiel pour la survie de la LMA et pourrait être une cible thérapeutique dans la LMA [143].

La galectine 1 (LGALS1) participe à diverses voies de survie qui soutiennent de nombreuses molécules pro-tumorales, en particulier celles régulées par RAS. *LGALS1* régule la différenciation des cellules hématopoïétiques, il n'est donc pas surprenant que *LGALS1* joue un rôle dans de nombreuses hémopathies malignes telles que la leucémie lymphoblastique aiguë.
LGALS1 joue un rôle important dans diverses voies de survie, y compris la voie p53, et *LGALS1* régule la mobilisation des cellules leucémiques de la niche de la leucémie. Cependant, le rôle de *LGALS1* dans la LMA n'est pas bien défini. Par conséquent, une stratégie ciblant *LGALS1* peut bénéficier aux patients atteints de LMA [144].

Ces observations nous ont menés à regarder les processus biologiques associés à la comparaison des gènes entre les deux groupes G1 et G2 testés. Cependant, les analyses de l'enrichissement et la caractérisation des sous-clones restent préliminaires pour bâtir des conclusions. Nous présentons ici un premier exemple d'une approche qui deviendra une stratégie globale effectuée à travers un grand nombre d'échantillon. En effet, la capacité de relier les informations génétiques et transcriptomiques dans des cellules individuelles a des implications importantes pour l'étude des populations cellulaires hétérogènes de LMA. Cependant, l'importance de ces sous-groupes reste à être démontrée. Par exemple, des analyses sur les sous-clones retrouvées dans les échantillons de caryotype complexe pourraient être faites pour confirmer les différences transcriptomiques entre les différents sous-clones.

De plus, de multiples éléments de preuve suggèrent maintenant que la LMA implique un processus d'évolution de ramification et que ces points de ramification peuvent être délimités sur la base de mutations génomiques partagées au sein de chaque sous-clone. À ce jour, les translocations équilibrées (par exemple t(15;17), t(8;21), t(16;16), inv(16) et les réarrangements MLL) et les variantes nucléotidiques dans *DNMT3A* et *TET2* apparaissent presque universellement dans le clone fondateur, et sont susceptibles d'être des événements initiatiques. En revanche, +8, +22, -X, -Y et les variantes de *FLT3*, *NRAS/KRAS, WT1* et *KIT* apparaissent fréquemment dans les sous-clones et sont donc susceptibles d'être des événements de progression [145].

Les progrès récents des techniques de génomique ont mis au jour l'hétérogénéité moléculaire de la LMA et contribuent à affiner la stratification et le pronostic du risque. Les patients atteints de LMA à risque indésirable nécessitent un traitement plus agressif, y compris une greffe de cellules souches hématopoïétiques allogéniques dans la première rémission complète et éventuellement de nouveaux agents ciblés, pour améliorer le pronostic. Cependant, le modèle complexe de coopérativité et d'exclusivité mutuelle entre les différentes mutations reste un défi clinique [146].

Depuis 2017, il y a eu une explosion d'options de traitement nouvellement approuvées pour adapter le traitement personnalisé de la LMA. Chacune de ces thérapies ciblées a un calendrier de traitement, un dosage, une efficacité et des effets indésirables uniques et une gestion appropriée est cruciale pour le succès du traitement. D'autres combinaisons de thérapies moléculaires ciblées et de chimiothérapie cytotoxique standard ou d'autres nouveaux agents pour améliorer l'efficacité sont toujours à l'étude [146]. Nous pensons que la position de ces mutations au sein de l'architecture sous-clonale de la LMA, a des implications importantes pour l'administration et l'interprétation de la réponse aux traitements ciblés, en particulier parce que bon nombre de petites molécules les plus prometteuses en développement ciblent des mutations qui peuvent se présenter dans les sous-clones plutôt que dans le clone fondateur.

4.2 Limitations

La plus grande limitation de mon approche d'annotation des types cellulaires est la dépendance de celle-ci à la sélection des gènes grâce à la variance. En effet, l'utilisation du transcriptome complet pour annoter les cellules serait une piste qu'on va explorer pour améliorer la précision du classificateur. Cette limitation est amplifiée aussi par le débalancement du nombre de cellules par type cellulaire, utilisée pendant la phase d'entrainement du classificateur. Nous avons noté, que des cellules HSC-like sont parfois confondues avec les plasmocytes. L'utilisation d'une démarche de « *super sampling* » pour éviter la surreprésentation des différents types cellulaires pourrait atténuer ce problème. On pourrait aussi entrainer un algorithme de réseaux de neurone et comparer sa précision avec celle obtenu avec le RF. De plus, nous avons remarqué un très haut pourcentage d'éosinophile dans le jeu de données de HCA ce qui ne reflètent peut-être pas la réalité. Ceci pourrait être remplacée avec une autre étiquette qui répond plus à l'ordre de l'hématopoïèse normale.

Une autre limitation de la méthode utilisée pour comparer les gènes differentiellement exprimés entre les sous-groupes, est le choix du test statistique. Tout d'abord, étant donné que les comptes sont basés sur le nombre, ils ne peuvent pas être distribuées normalement. Deux distributions pour les données basées sur le nombre sont poisson (qui suppose que la variance et la moyenne sont égales) ou binomiale négative (ce qui n'est pas le cas). C'est particulièrement un problème technique pour certains gènes lorsque le nombre de comptes est faible car il est difficile de modéliser avec précision la variance des données basées sur le nombre si on ne regarde que ce gène et on fait des hypothèses de données continues normalement distribuées (c'est-à-dire un test t). Une bonne estimation de la variance pour chaque gène est essentielle pour déterminer si les changements sont dus au hasard. Dans ce cas, on pourrait essayer un autre outil pour la robustesse de l'approche. DESeq2 est un outil populaire pour l'analyse de l'expression différentielle au niveau des gènes. Il utilise la distribution binomiale négative, employant une

approche légèrement plus stricte par rapport à certaines méthodes tout en ayant un bon équilibre entre sensibilité et spécificité (réduction à la fois des faux positifs et des faux négatifs).

4.3 Conclusion et perspective

Pour conclure, nous avons développé une méthode pour annoter systématiquement les cellules leucémiques par rapport à l'hématopoïèse normale. Ceci a révélé une grande diversité de maturation cellulaire dans les différents échantillons étudiés. Par la suite, nous avons supposé qu'une partie de la diversité phénotypique était expliquée par les mutations ou CNVs. Nous avons conclu que les mutations autres que *NPM1* sont bien difficile à génotyper en raison de contraintes techniques. Mais, nous sommes parvenus à assez bien inférer les CNVs en optimisant une méthode CopyKAT grâce à l'intégration multiOmics utilisant le « bulk » exome pairé des mêmes échantillons. Ceci ouvre la voie entre autres à l'exploration des différences transcriptomiques pour un type cellulaire entre échantillons d'un sous-groupe génétique donné, ou même entre différents sous clones CNVs d'un même échantillon.

Références bibliographiques

[1] Self-renewal, differentiation or death: regulation and manipulation of hematopoietic stem cell fate - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/10322312/ (accessed November 5, 2020).

[2] Bone Marrow (Hematopoietic) Stem Cells | stemcells.nih.gov n.d. https://stemcells.nih.gov/info/Regenerative_Medicine/2006Chapter2.htm (accessed November 5, 2020).

[3] Allsopp RC, Morin GB, Horner JW, DePinho R, Harley CB, Weissman IL. Effect of TERT over-expression on the long-term transplantation capacity of hematopoietic stem cells. Nat Med 2003;9:369–71. https://doi.org/10.1038/nm0403-369.

[4] Purification and characterization of mouse hematopoietic stem cells - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/2898810/ (accessed November 5, 2020).

[5] Seita J, Weissman IL. Hematopoietic stem cell: self-renewal versus differentiation. Wiley Interdiscip Rev Syst Biol Med 2010;2:640–53. https://doi.org/10.1002/wsbm.86.

[6] Stem cells, cancer, and cancer stem cells - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/11689955/ (accessed November 5, 2020).

[7] Long-Term Lymphohematopoietic Reconstitution by a Single CD34-Low/Negative Hematopoietic Stem Cell | Science n.d. https://science.sciencemag.org/content/273/5272/242.abstract (accessed November 5, 2020).

[8] The bulk of the hematopoietic stem cell population is dispensable for murine steady-state and stress hematopoiesis - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/27357698/ (accessed November 5, 2020).

[9] Yang L, Bryder D, Adolfsson J, Nygren J, Månsson R, Sigvardsson M, et al. Identification of Lin(-)Sca1(+)kit(+)CD34(+)Flt3- short-term hematopoietic stem cells capable of rapidly reconstituting and rescuing myeloablated transplant recipients. Blood 2005;105:2717–23. https://doi.org/10.1182/blood-2004-06-2159.

[10] Pietras EM, Reynaud D, Kang Y-A, Carlin D, Calero-Nieto FJ, Leavitt AD, et al. Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions. Cell Stem Cell 2015;17:35–46. https://doi.org/10.1016/j.stem.2015.05.003.

[11] Blank U, Karlsson S. TGF-β signaling in the control of hematopoietic stem cells. Blood 2015;125:3542–50. https://doi.org/10.1182/blood-2014-12-618090.

[12] The World Health Organization classification of hematological malignancies report of the Clinical Advisory Committee Meeting, Airlie House, Virginia, November 1997 - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/10697278/ (accessed November 5, 2020).

[13] Uribesalgo I, Di Croce L. Dynamics of epigenetic modifications in leukemia. Brief Funct Genomics 2011;10:18–29. https://doi.org/10.1093/bfgp/elr002.

[14] Olsen M. Overview of Hematologic Malignancies n.d.:17.

[15] Acute Myeloid Leukemia - Cancer Stat Facts. SEER n.d. https://seer.cancer.gov/statfacts/html/amyl.html (accessed November 4, 2020).

[16] Grove CS, Vassiliou GS. Acute myeloid leukaemia: a paradigm for the clonal evolution of

cancer? Dis Model Mech 2014;7:941–51. https://doi.org/10.1242/dmm.015974.
[17] Deschler B, Lübbert M. Acute myeloid leukemia: epidemiology and etiology. Cancer 2006;107:2099–107. https://doi.org/10.1002/cncr.22233.
[18] Jemal A, Thomas A, Murray T, Thun M. Cancer statistics, 2002. CA Cancer J Clin 2002;52:23–47. https://doi.org/10.3322/canjclin.52.1.23.
[19] Estey E, Döhner H. Acute myeloid leukaemia. Lancet 2006;368:1894–907. https://doi.org/10.1016/S0140-6736(06)69780-8.
[20] De Kouchkovsky I, Abdul-Hay M. "Acute myeloid leukemia: a comprehensive review and 2016 update." Blood Cancer J 2016;6:e441. https://doi.org/10.1038/bcj.2016.50.
[21] The 2016 WHO classification of acute myeloid leukemia: What the practicing clinician needs to know - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/30926096/ (accessed November 4, 2020).
[22] Bennett JM, Catovsky D, Daniel MT, Flandrin G, Galton DA, Gralnick HR, et al. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. Br J Haematol 1976;33:451–8. https://doi.org/10.1111/j.1365-2141.1976.tb03563.x.
[23] Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood 2016;127:2391–405. https://doi.org/10.1182/blood-2016-03-643544.
[24] Yang JJ, Park TS, Wan TSK. Recurrent Cytogenetic Abnormalities in Acute Myeloid Leukemia. Methods Mol Biol 2017;1541:223–45. https://doi.org/10.1007/978-1-4939-6703-2_19.
[25] Döhner H, Estey EH, Amadori S, Appelbaum FR, Büchner T, Burnett AK, et al. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. Blood 2010;115:453–74. https://doi.org/10.1182/blood-2009-07-235358.
[26] Preudhomme C, Sagot C, Boissel N, Cayuela J-M, Tigaud I, de Botton S, et al. Favorable prognostic significance of CEBPA mutations in patients with de novo acute myeloid leukemia: a study from the Acute Leukemia French Association (ALFA). Blood 2002;100:2717–23. https://doi.org/10.1182/blood-2002-03-0990.
[27] Döhner K, Schlenk RF, Habdank M, Scholl C, Rücker FG, Corbacioglu A, et al. Mutant nucleophosmin (NPM1) predicts favorable prognosis in younger adults with acute myeloid leukemia and normal cytogenetics: interaction with other gene mutations. Blood 2005;106:3740–6. https://doi.org/10.1182/blood-2005-05-2164.
[28] Bacher U, Schnittger S, Haferlach T. Molecular genetics in acute myeloid leukemia. Curr Opin Oncol 2010;22:646–55. https://doi.org/10.1097/CCO.0b013e32833ed806.
[29] Zheng X, Beissert T, Kukoc-Zivojnov N, Puccetti E, Altschmied J, Strolz C, et al. Gamma-catenin contributes to leukemogenesis induced by AML-associated translocation products by increasing the self-renewal of very primitive progenitor cells. Blood 2004;103:3535–43. https://doi.org/10.1182/blood-2003-09-3335.
[30] Grimwade D, Hills RK, Moorman AV, Walker H, Chatters S, Goldstone AH, et al. Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials. Blood 2010;116:354–65. https://doi.org/10.1182/blood-2009-11-254441.

[31] Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood 2017;129:424–47. https://doi.org/10.1182/blood-2016-08-733196.
[32] Stölzel F, Mohr B, Kramer M, Oelschlägel U, Bochtler T, Berdel WE, et al. Karyotype complexity and prognosis in acute myeloid leukemia. Blood Cancer J 2016;6:e386. https://doi.org/10.1038/bcj.2015.114.
[33] Weissmann S, Alpermann T, Grossmann V, Kowarsch A, Nadarajah N, Eder C, et al. Landscape of TET2 mutations in acute myeloid leukemia. Leukemia 2012;26:934–42. https://doi.org/10.1038/leu.2011.326.
[34] Risques RA, Kennedy SR. Aging and the rise of somatic cancer-associated mutations in normal tissues. PLoS Genet 2018;14:e1007108. https://doi.org/10.1371/journal.pgen.1007108.
[35] Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG, et al. Age-related clonal hematopoiesis associated with adverse outcomes. N Engl J Med 2014;371:2488–98. https://doi.org/10.1056/NEJMoa1408617.
[36] Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt DC, et al. The Origin and Evolution of Mutations in Acute Myeloid Leukemia. Cell 2012;150:264–78. https://doi.org/10.1016/j.cell.2012.06.023.
[37] Challen GA, Sun D, Jeong M, Luo M, Jelinek J, Berg JS, et al. Dnmt3a is essential for hematopoietic stem cell differentiation. Nat Genet 2011;44:23–31. https://doi.org/10.1038/ng.1009.
[38] Klepin HD. Myelodysplastic Syndromes and Acute Myeloid Leukemia in the Elderly. Clinics in Geriatric Medicine 2016;32:155–73. https://doi.org/10.1016/j.cger.2015.08.010.
[39] Steensma DP, Bejar R, Jaiswal S, Lindsley RC, Sekeres MA, Hasserjian RP, et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. Blood 2015;126:9–16. https://doi.org/10.1182/blood-2015-03-631747.
[40] Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, et al. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics 2006;7:246. https://doi.org/10.1186/1471-2164-7-246.
[41] Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 2008;453:1239–43. https://doi.org/10.1038/nature07002.
[42] Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, et al. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol 2008;17:1636–47. https://doi.org/10.1111/j.1365-294X.2008.03666.x.
[43] Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 2008;5:613–9. https://doi.org/10.1038/nmeth.1223.
[44] Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLOS Computational Biology 2017;13:e1005457. https://doi.org/10.1371/journal.pcbi.1005457.
[45] F T, C B, Y W, E N, C L, N X, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 2009;6:377–82. https://doi.org/10.1038/nmeth.1315.
[46] Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front Genet 2019;10:317. https://doi.org/10.3389/fgene.2019.00317.
[47] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying

mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621–8. https://doi.org/10.1038/nmeth.1226.
[48] Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell Reports 2012;2:666–73. https://doi.org/10.1016/j.celrep.2012.08.003.
[49] Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol 2016;17:77. https://doi.org/10.1186/s13059-016-0938-8.
[50] Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol 2013;14:R31. https://doi.org/10.1186/gb-2013-14-4-r31.
[51] Sasagawa Y, Danno H, Takada H, Ebisawa M, Tanaka K, Hayashi T, et al. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. Genome Biol 2018;19:29. https://doi.org/10.1186/s13059-018-1407-3.
[52] Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods 2013;10:1096–8. https://doi.org/10.1038/nmeth.2639.
[53] Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc 2018;13:599–604. https://doi.org/10.1038/nprot.2017.149.
[54] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 2015;161:1202–14. https://doi.org/10.1016/j.cell.2015.05.002.
[55] Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 2015;161:1187–201. https://doi.org/10.1016/j.cell.2015.04.044.
[56] Zilionis R, Nainys J, Veres A, Savova V, Zemmour D, Klein AM, et al. Single-cell barcoding and sequencing using droplet microfluidics. Nat Protoc 2017;12:44–73. https://doi.org/10.1038/nprot.2016.154.
[57] Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8:14049. https://doi.org/10.1038/ncomms14049.
[58] Single Cell Gene Expression – 10X Genomics n.d. https://kb.10xgenomics.com/hc/en-us/categories/360000149952-Single-Cell-Gene-Expression (accessed July 31, 2021).
[59] Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods 2011;9:72–4. https://doi.org/10.1038/nmeth.1778.
[60] Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 2014;11:163–6. https://doi.org/10.1038/nmeth.2772.
[61] Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell 2015;58:610–20. https://doi.org/10.1016/j.molcel.2015.04.005.
[62] Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell 2017;65:631-643.e4. https://doi.org/10.1016/j.molcel.2017.01.023.

[63] Saliba A-E, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res 2014;42:8845–60. https://doi.org/10.1093/nar/gku555.
[64] Baran-Gale J, Chandra T, Kirschner K. Experimental design for single-cell RNA sequencing. Brief Funct Genomics 2018;17:233–9. https://doi.org/10.1093/bfgp/elx035.
[65] Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. Nat Commun 2019;10:4667. https://doi.org/10.1038/s41467-019-12266-7.
[66] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–20. https://doi.org/10.1038/nbt.4096.
[67] McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics 2017;33:1179–86. https://doi.org/10.1093/bioinformatics/btw777.
[68] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:15. https://doi.org/10.1186/s13059-017-1382-0.
[69] Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol 2016;17:29. https://doi.org/10.1186/s13059-016-0888-1.
[70] Principal Component Analysis | I.T. Jolliffe | Springer n.d. https://www.springer.com/gp/book/9780387954424 (accessed July 31, 2021).
[71] Maaten L van der, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research 2008;9:2579–605.
[72] McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv:180203426 [Cs, Stat] 2020.
[73] Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods 2017;14:565–71. https://doi.org/10.1038/nmeth.4292.
[74] Bard J, Rhee SY, Ashburner M. An ontology for cell types. Genome Biol 2005;6:R21. https://doi.org/10.1186/gb-2005-6-2-r21.
[75] Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 2019;20:273–82. https://doi.org/10.1038/s41576-018-0088-9.
[76] Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods 2017;14:483–6. https://doi.org/10.1038/nmeth.4236.
[77] Grün D, Muraro MJ, Boisset J-C, Wiebrands K, Lyubimova A, Dharmadhikari G, et al. De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. Cell Stem Cell 2016;19:266–77. https://doi.org/10.1016/j.stem.2016.05.010.
[78] Chen J, Schlitzer A, Chakarov S, Ginhoux F, Poidinger M. Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. Nat Commun 2016;7:11988. https://doi.org/10.1038/ncomms11988.
[79] Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 2015;347:1138–42. https://doi.org/10.1126/science.aaa1934.
[80] Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, et al. Data-Driven

Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell 2015;162:184–97. https://doi.org/10.1016/j.cell.2015.05.047.
[81] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–40. https://doi.org/10.1093/bioinformatics/btp616.
[82] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology 2010;11:R106. https://doi.org/10.1186/gb-2010-11-10-r106.
[83] Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods 2014;11:740–2. https://doi.org/10.1038/nmeth.2967.
[84] Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 2015;16:278. https://doi.org/10.1186/s13059-015-0844-5.
[85] Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics 2019;20:40. https://doi.org/10.1186/s12859-019-2599-6.
[86] Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods 2018;15:255–61. https://doi.org/10.1038/nmeth.4612.
[87] Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc 2019;14:482–517. https://doi.org/10.1038/s41596-018-0103-9.
[88] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–9. https://doi.org/10.1038/75556.
[89] Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res 2014;42:D472-477. https://doi.org/10.1093/nar/gkt1102.
[90] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30. https://doi.org/10.1093/nar/28.1.27.
[91] García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway Analysis: State of the Art. Front Physiol 2015;6:383. https://doi.org/10.3389/fphys.2015.00383.
[92] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102.
[93] Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics 2013;14:7. https://doi.org/10.1186/1471-2105-14-7.
[94] Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. PLOS ONE 2010;5:e13984. https://doi.org/10.1371/journal.pone.0013984.
[95] Kucera M, Isserlin R, Arkhangorodsky A, Bader GD. AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. F1000Res 2016;5:1717. https://doi.org/10.12688/f1000research.9090.1.
[96] Hon C-C, Shin JW, Carninci P, Stubbington MJT. The Human Cell Atlas: Technical approaches and challenges. Brief Funct Genomics 2017;17:283–94. https://doi.org/10.1093/bfgp/elx029.
[97] Hay SB, Ferchen K, Chetal K, Grimes HL, Salomonis N. The Human Cell Atlas bone marrow single-cell interactive web portal. Exp Hematol 2018;68:51–61.

https://doi.org/10.1016/j.exphem.2018.09.004.
[98] Giustacchini A, Thongjuea S, Barkas N, Woll PS, Povinelli BJ, Booth CAG, et al. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. Nat Med 2017;23:692–702. https://doi.org/10.1038/nm.4336.
[99] Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature 2017;541:331–8. https://doi.org/10.1038/nature21350.
[100] Petti AA, Williams SR, Miller CA, Fiddes IT, Srivatsan SN, Chen DY, et al. A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. Nat Commun 2019;10:3660. https://doi.org/10.1038/s41467-019-11591-1.
[101] Austin R, Smyth MJ, Lane SW. Harnessing the immune system in acute myeloid leukaemia. Crit Rev Oncol Hematol 2016;103:62–77. https://doi.org/10.1016/j.critrevonc.2016.04.020.
[102] Leick MB, Levis MJ. The Future of Targeting FLT3 Activation in AML. Curr Hematol Malig Rep 2017;12:153–67. https://doi.org/10.1007/s11899-017-0381-2.
[103] SciBet as a portable and fast single cell type identifier | Nature Communications n.d. https://www.nature.com/articles/s41467-020-15523-2 (accessed August 1, 2021).
[104] Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. Elife 2017;6:e27041. https://doi.org/10.7554/eLife.27041.
[105] Vig L. Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset. Open Access Library Journal 2014;1:1–7. https://doi.org/10.4236/oalib.1100660.
[106] Vidyasagar M. Machine learning methods in the computational biology of cancer. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 2014;470:20140081. https://doi.org/10.1098/rspa.2014.0081.
[107] Classification and Regression by randomForest | BibSonomy n.d. https://www.bibsonomy.org/bibtex/2ba2e49a65786a6ff232994289edb42f3/lukasbeckmann (accessed August 1, 2021).
[108] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 1967;13:21–7. https://doi.org/10.1109/TIT.1967.1053964.
[109] Shalev-Shwartz S, Ben-David S. Understanding Machine Learning: From Theory to Algorithms. USA: Cambridge University Press; 2014.
[110] Quinlan JR. Induction of Decision Trees. Mach Learn 1986;1:81–106.
[111] Ho TK. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, 1995, p. 278–82 vol.1. https://doi.org/10.1109/ICDAR.1995.598994.
[112] Breiman L. Random Forests. Machine Learning 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
[113] Ho TK. A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors. Pattern Anal Appl 2002;5:102–12. https://doi.org/10.1007/s100440200009.
[114] Sohil F, Sohali MU, Shabbir J. An introduction to statistical learning with applications in R: by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7. Statistical Theory and Related Fields 2021:1–1. https://doi.org/10.1080/24754269.2021.1980261.
[115] Mitchell MW. Bias of the Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters. Open Journal of Statistics 2011;1:205–11. https://doi.org/10.4236/ojs.2011.13024.
[116] Deng H, Runger G, Tuv E. Bias of Importance Measures for Multi-valued Attributes and

Solutions. In: Honkela T, Duch W, Girolami M, Kaski S, editors. Artificial Neural Networks and Machine Learning – ICANN 2011, Berlin, Heidelberg: Springer; 2011, p. 293–300. https://doi.org/10.1007/978-3-642-21738-8_38.
[117] Chandrashekar G, Sahin F. A survey on feature selection methods. Computers & Electrical Engineering 2014;40:16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.
[118] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27:1226–38. https://doi.org/10.1109/TPAMI.2005.159.
[119] Liu H, Setiono R. Chi2: Feature Selection and Discretization of Numeric Attributes. In Proceedings of the Seventh International Conference on Tools with Artificial Intelligence, 1995, p. 388–91.
[120] Refaeilzadeh P, Tang L, Liu H. Cross-Validation. In: LIU L, ÖZSU MT, editors. Encyclopedia of Database Systems, Boston, MA: Springer US; 2009, p. 532–8. https://doi.org/10.1007/978-0-387-39940-9_565.
[121] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020;21:6. https://doi.org/10.1186/s12864-019-6413-7.
[122] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.
[123] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–30.
[124] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27:1739–40. https://doi.org/10.1093/bioinformatics/btr260.
[125] Rouette A, Trofimov A, Haberl D, Boucher G, Lavallée V-P, D'Angelo G, et al. Expression of immunoproteasome genes is regulated by cell-intrinsic and –extrinsic factors in human cancers. Sci Rep 2016;6:34019. https://doi.org/10.1038/srep34019.
[126] Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, et al. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. Sci Rep 2017;7:43169. https://doi.org/10.1038/srep43169.
[127] Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol 2021;39:599–608. https://doi.org/10.1038/s41587-020-00795-2.
[128] Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol 2015;26:64–70. https://doi.org/10.1093/annonc/mdu479.
[129] Fang Z. gseapy Documentation n.d.:49.
[130] van Galen P, Hovestadt V, Wadsworth Ii MH, Hughes TK, Griffin GK, Battaglia S, et al. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell 2019;176:1265-1281.e24. https://doi.org/10.1016/j.cell.2019.01.031.
[131] Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature 2016;539:309–13. https://doi.org/10.1038/nature20123.

[132] Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science 2017;355:eaai8478. https://doi.org/10.1126/science.aai8478.
[133] Fan J, Lee H-O, Lee S, Ryu D-E, Lee S, Xue C, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res 2018;28:1217–27. https://doi.org/10.1101/gr.228080.117.
[134] Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods 2015;12:519–22. https://doi.org/10.1038/nmeth.3370.
[135] Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 2014.
[136] Durruthy-Durruthy R, Ray M. Using Fluidigm C1 to Generate Single-Cell Full-Length cDNA Libraries for mRNA Sequencing. Methods in molecular biology (Clifton, N.J.), vol. 1706, 2018, p. 199–221. https://doi.org/10.1007/978-1-4939-7471-9_11.
[137] Annals of Leukemia Research | Somato Publications n.d. https://www.somatopublications.com/annals_leukemia_research/volume_issue.php?vid=Mg==&sid=MQ== (accessed December 3, 2021).
[138] Xue C, Zhang J, Zhang G, Xue Y, Zhang G, Wu X. Elevated SPINK2 gene expression is a predictor of poor prognosis in acute myeloid leukemia. Oncol Lett 2019;18:2877–84. https://doi.org/10.3892/ol.2019.10665.
[139] Chen T, Lee T-R, Liang W-G, Chang W-SW, Lyu P-C. Identification of trypsin-inhibitory site and structure determination of human SPINK2 serine proteinase inhibitor. Proteins 2009;77:209–19. https://doi.org/10.1002/prot.22432.
[140] Chan J, Quintanal-Villalonga Á, Gao V, Xie Y, Allaj V, Chaudhary O, et al. Single cell profiling reveals novel tumor and myeloid subpopulations in small cell lung cance 2020. https://doi.org/10.1101/2020.12.01.406363.
[141] Hunter CS, Rhodes SJ. LIM-homeodomain genes in mammalian development and human disease. Mol Biol Rep 2005;32:67–77. https://doi.org/10.1007/s11033-004-7657-z.
[142] Cui L, Cheng Z, Hu K, Pang Y, Liu Y, Qian T, et al. Prognostic value of the PDLIM family in acute myeloid leukemia. Am J Transl Res 2019;11:6124–31.
[143] Alanazi B, Munje CR, Rastogi N, Williamson AJK, Taylor S, Hole PS, et al. Integrated nuclear proteomics and transcriptomics identifies S100A4 as a therapeutic target in acute myeloid leukemia. Leukemia 2020;34:427–40. https://doi.org/10.1038/s41375-019-0596-4.
[144] Ruvolo P, Ma H, Ruvolo V, Zhang X, Post S, Andreeff M. AML-044: LGALS1 Acts as a Pro-Survival Molecule in AML. Clinical Lymphoma, Myeloma and Leukemia 2020;20:S176–7. https://doi.org/10.1016/S2152-2650(20)30707-2.
[145] Welch JS. Subclonal architecture in acute myeloid leukemia 2013:7.
[146] Hou H-A, Tien H-F. Genomic landscape in acute myeloid leukemia and its implications in risk classification and targeted therapies. Journal of Biomedical Science 2020;27:81. https://doi.org/10.1186/s12929-020-00674-7.

Annexes

Liste des gènes associés à la LMA :

BRAF, ASXL1, ASXL2, BCOR, BIRC3, CALR, CBL, CEBPA, CSF3R, DNMT3A, ETV6, EZH2, FBXW7, FLT3, GATA2, HRAS, IDH1, IDH2, IKZF1, JAK1, JAK2, JAK3, KIT, KRAS, MPL, MYD88, NF1, NOTCH1, NPM1, NRAS, PHF6, PTEN, PTPN11, RAD21, RUNX1, SETBP1, SETD2, SF3B1, SMC1A, SMC3, SMC5, SRSF2, STAG1, STAG2, STAT3, STAT5A, STAT5B, TET2, TP53, U2AF1, WT1, ZRSR2.