

Thesis, FRANZEN Rachelle

Auteur : Choulli Rifi, Nada

Promoteur(s) : Ehx, Grégory

Faculté : Faculté de Médecine

Diplôme : Master en sciences biomédicales, à finalité spécialisée en biomédical data management

Année académique : 2025-2026

URI/URL : <http://hdl.handle.net/2268.2/25982>

Avertissement à l'attention des usagers :

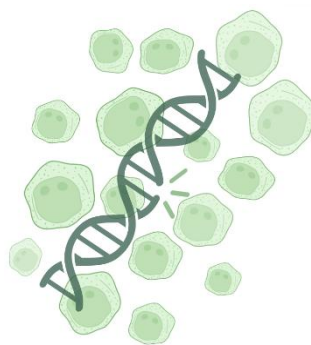
Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

Développement d'une approche d'annotation des cellules leucémiques basée sur la charge mutationnelle unicellulaire dans la leucémie myéloïde aiguë

*Mémoire présenté en vue de l'obtention du grade académique
de Master en Sciences Biomédicales à finalité spécialisée en
Data Management*

Présenté par
Nada Choulli Rifi



Promoteur
Dr. Grégory Ehx

Superviseur
Dr. Sébastien Charles

Laboratoire d'hématologie, Institut GIGA, ULiège
Immunologie Translationnelle et Leucémie

Remerciements

Avant de présenter ce mémoire, je tiens à remercier sincèrement Grégory et Sébastien, qui m'ont permis de réaliser ce travail au sein de leur équipe. Je leur suis profondément reconnaissante pour leur encadrement, leur disponibilité et leur investissement tout au long de cette expérience, aussi bien pendant la réalisation du mémoire que lors de sa rédaction.

À aucun moment je ne me suis sentie seule face à ce projet. Je savais que je pouvais toujours me tourner vers eux pour poser mes questions, partager mes doutes ou simplement demander un avis. Ils ont toujours pris le temps de m'expliquer les choses, malgré leurs horaires chargés et leurs nombreuses occupations. Même lorsque je leur envoyais des messages à toute heure, ils répondaient toujours présents. Je les remercie également pour leurs encouragements et leurs conseils, qui ont dépassé le cadre du mémoire et m'ont apporté de véritables conseils de vie. Grâce à eux, cette expérience, qui peut être assez stressante, a été rendue aussi agréable que possible.

Je souhaite également remercier mes proches, ma famille et mes amis, qui comptent pour moi comme une seconde famille. Merci pour votre soutien, votre patience et votre implication dans ce mémoire.

J'aimerais enfin remercier tout particulièrement mes parents, pour tous les sacrifices qu'ils ont faits afin que nous puissions avoir accès à des opportunités auxquelles eux-mêmes n'ont pas eu accès.

Utilisation de l'intelligence artificielle

Dans le cadre de la rédaction de ce mémoire, l'intelligence artificielle a été utilisée comme outil d'aide à certaines traductions, à la compréhension d'articles scientifiques et à l'aide à la constitution de la liste des abréviations. Elle a également servi à la correction linguistique, notamment pour vérifier l'orthographe, la grammaire et la syntaxe, ainsi qu'à l'amélioration de certaines tournures de phrase.

Ces outils n'ont en aucun cas été utilisés pour la conception, l'analyse ou l'interprétation scientifique des résultats. L'auteur et l'équipe restent entièrement responsables du contenu, de son exactitude et de son intégrité scientifique.

Résumé

La leucémie myéloïde aiguë est une hémopathie maligne caractérisée par la prolifération incontrôlée de blastes myéloïdes immatures dans la moelle osseuse et/ou le sang périphérique. Elle est associée à un taux de rechute élevé malgré une rémission complète fréquemment obtenue après chimiothérapie. Cette persistance leucémique est en partie attribuée à des cellules capables de survivre au traitement sans dépendre nécessairement de l'acquisition de nouvelles mutations de résistance, désignées sous le terme de *Drug-Tolerant Persisters*. L'identification de ces cellules à partir de données de séquençage d'ARN à cellule unique (scRNA-seq) reste limitée par les approches actuelles, qui reposent souvent sur la détection d'un transcrit muté unique ou d'une altération génétique spécifique, et sont difficilement généralisables entre patients.

Ce mémoire propose une stratégie d'annotation des cellules leucémiques fondée sur la charge mutationnelle totale normalisée par le nombre de gènes détectés par cellule, calculée à partir des variants identifiés dans les données de scRNA-seq. Une classification non supervisée par *k-means* est appliquée pour distinguer les populations *AML-like*, définies ici comme des populations compatibles avec une origine leucémique, des populations *non-AML-like*, avec l'introduction d'une marge d'incertitude afin de limiter les classifications ambiguës. Cette approche a été évaluée sur six patients issus de la cohorte de Naldini *et al.*, porteurs d'une mutation du gène *NPM1* ou d'une délétion du chromosome 7, à plusieurs points temporels, incluant le diagnostic et la période suivant la chimiothérapie d'induction.

Les résultats montrent une concordance partielle avec les annotations de référence, variable selon les patients, les altérations étudiées et les points temporels analysés. Notre approche permet notamment d'identifier des populations *AML-like* après traitement dans certains échantillons, y compris lorsque les annotations de référence deviennent moins informatives. Des analyses transcriptomiques complémentaires soutiennent la cohérence biologique de ces annotations. Ce travail constitue une première étape méthodologique vers un suivi clonal des populations leucémiques persistantes à partir de données scRNA-seq, afin de mieux caractériser les mécanismes cellulaires impliqués dans la rechute.

Abstract

Acute myeloid leukemia is a hematological malignancy characterized by the uncontrolled proliferation of immature myeloid blasts in the bone marrow and/or peripheral blood. It is associated with a high relapse rate despite the frequent achievement of complete remission after chemotherapy. This leukemic persistence is partly attributed to cells capable of surviving treatment without necessarily relying on the acquisition of new resistance mutations, referred to as Drug-Tolerant Persisters. The identification of these cells from single-cell RNA sequencing (scRNA-seq) data remains limited by current approaches, which often rely on the detection of a single mutated transcript or a specific genetic alteration and are difficult to generalize across patients.

This thesis proposes a strategy for annotating leukemic cells based on the total mutation burden normalized by the number of genes detected per cell, calculated from variants identified in scRNA-seq data. Unsupervised k-means clustering is applied to distinguish AML-like populations, defined here as populations compatible with a leukemic origin, from non-AML-like populations, with the introduction of an uncertainty margin to limit ambiguous classifications. This approach was evaluated on six patients from the Naldini *et al.* cohort, carrying a mutation in the *NPM1* gene or a chromosome 7 deletion, at several time points, including diagnosis and post-induction chemotherapy.

The results show partial agreement with the reference annotations, varying according to the patients, the alterations studied, and the time points analyzed. Our approach notably enables the identification of AML-like populations after treatment in some samples, including when the reference annotations become less informative. Additional transcriptomic analyses support the biological consistency of these annotations. This work represents a first methodological step toward clonal tracking of persistent leukemic populations from scRNA-seq data, in order to better characterize the cellular mechanisms involved in relapse.

Table des matières

1. Introduction	1
1.1 Contexte clinique de la LMA	1
1.1.1 Diagnostic	1
1.1.2 Traitement	2
1.1.3 Réponse au traitement et rémission	3
1.1.4 Rechute et maladie réfractaire	4
1.2 Bases biologiques de la persistance leucémique	5
1.2.1 Hématopoïèse physiologique	5
1.2.2 Transformation leucémique et hématopoïèse	6
1.2.3 Altérations génétiques d'intérêt	7
1.2.4 Émergence des cellules souches leucémiques (CSL)	8
1.2.5 Drug-Tolerant Persisters	10
1.3 Défis liés à l'étude de la LMA	12
1.3.1 Limites des approches classiques	13
1.3.2 Apport du scRNA-seq dans la LMA	14
1.3.3 Limites de la discrimination <i>AML-like</i> / <i>non-AML-like</i> en scRNA-seq	15
2. Objectif et hypothèse	17
3. Matériel et méthodes	18
3.1 Source des données et cadre de l'étude	18
3.2 Génération des données de scRNA-seq	19
3.3 Analyse mutationnelle	19
3.4 Analyse transcriptomique	22
4. Résultats	31
4.1 Normalisation et corrélation entre les charges mutationnelles	31
4.2 Analyse détaillée d'un patient <i>NPM1^{mut}</i>	33
4.3 Analyse détaillée d'un patient del(7)	35
4.4 Analyse transcriptomique complémentaire des cellules <i>AML-like</i>	36
4.4.1 L'expression de signatures transcriptomiques des DTPs soutient nos annotations	36
4.4.2 Une signature transcriptomique LMA soutient nos annotations	37
5. Discussion	40
5.1 Pertinence du signal d'annotation cellulaire	40
5.2 Comparaison avec les annotations de référence dérivées de Naldini <i>et al.</i>	42

5.3 Cohérence transcriptomique des populations <i>AML-like</i> identifiées	43
5.4 Limitations et perspectives	45
6. Bibliographie	47

Liste des abréviations

A8	Cytarabine administrée du jour 1 au jour 4
AD	Profondeur allélique (<i>allele depth</i>)
ADAR	<i>Adenosine deaminase acting on RNA</i>
ADAR1	<i>Adenosine deaminase acting on RNA 1</i>
AML	<i>Acute myeloid leukemia</i> , équivalent anglais de la LMA
AML-like	Cellules présentant un profil compatible avec une origine leucémique
AraC	Cytarabine
BAM	<i>Binary Alignment Map</i>
BCL-2	<i>B-cell lymphoma 2</i>
BED	<i>Browser Extensible Data</i>
CMF	Cytométrie en flux multiparamétrique
COSMIC	<i>Catalogue of Somatic Mutations in Cancer</i>
CSH	Cellules souches hématopoïétiques
CSL	Cellules souches leucémiques
D4	Jour 4
D14	Jour 14
D30	Jour 30
dbSNP	<i>Database of Single Nucleotide Polymorphisms</i>
del(7)	Délétion du chromosome 7
DTPs	<i>Drug-Tolerant Persisters</i>
DX	Diagnostic
ELN	<i>European Leukemia Network</i>
FLA-Ida	Fludarabine, cytarabine et idarubicine
FLAG-Ida	Fludarabine, cytarabine, G-CSF et idarubicine
FLT3	<i>Fms-like tyrosine kinase 3</i>
G-CSF	<i>Granulocyte colony-stimulating factor</i>
GEO	<i>Gene Expression Omnibus</i>
HMA/Ven	Agent hypométhylant associé au vénétoclax
HSCT	Transplantation de cellules souches hématopoïétiques
IFN-γ	Interféron gamma
IFN_DTP	Ensemble de gènes correspondant à l'intersection entre les gènes surexprimés dans les cellules DTPs et les gènes induits par l'IFN-γ
INDEL	Insertion/délétion
LMA	Leucémie myéloïde aiguë
MAMAC3	Amsacrine et cytarabine administrées du jour 1 au jour 3
MAMAC5	Amsacrine et cytarabine administrées du jour 1 au jour 5
MO	Moelle osseuse
MRD	Maladie résiduelle mesurable

NA	Non applicable
non-AML-like	Cellules ne présentant pas un profil compatible avec une origine leucémique selon la méthode utilisée
NPM1	<i>Nucleophosmin 1</i>
NR	Absence de réponse
PC	Composante principale
PCA	Analyse en composantes principales
PI	Investigateur principal (<i>principal investigator</i>)
PMID	Identifiant PubMed (<i>PubMed Identifier</i>)
RC	Rémission complète
Rel1/Rel2	Première/deuxième rechute
scRNA-seq	Séquençage d'ARN à cellule unique
SNV	Variant nucléotidique simple (<i>single nucleotide variant</i>)
UMAP	<i>Uniform Manifold Approximation and Projection</i>
UMI	Identifiant moléculaire unique (<i>unique molecular identifier</i>)
UTR	Région non traduite (<i>Untranslated region</i>)
VAF	Fréquence allélique
VCF	<i>Variant Call Format</i>
VEP	<i>Variant Effect Predictor</i>

1. Introduction

1.1 Contexte clinique de la LMA

La leucémie myéloïde aiguë (LMA) est une hémopathie maligne clonale à progression rapide, caractérisée par la prolifération incontrôlée de blastes myéloïdes bloqués à un stade immature dans la moelle osseuse (MO) et/ou le sang périphérique¹⁻⁴.

Ces cellules anormales envahissent progressivement la MO et remplacent les cellules hématopoïétiques normales, entraînant une pancytopénie se manifestant notamment par une anémie, des saignements liés à la thrombopénie, ainsi qu'une susceptibilité accrue aux infections⁵.

La LMA est une pathologie relativement rare, avec environ 601 nouveaux cas rapportés en 2023 selon le *Belgian Cancer Registry*⁶. Elle représente néanmoins un enjeu majeur de santé publique en raison de son pronostic souvent défavorable et de la population qu'elle touche préférentiellement : l'âge médian au diagnostic se situe, selon les études, entre 67 et 70 ans^{2,6,7}. Sur le plan étiologique, la majorité des cas surviennent de manière sporadique, en l'absence de cause identifiée (LMA *de novo*). Une proportion plus limitée survient dans un contexte prédisposant : dans une étude suédoise, environ 10 % des cas correspondaient à une LMA liée à un traitement antérieur et 19 % étaient associés à des antécédents hématologiques⁸. Ces contextes incluent notamment un syndrome myélodysplasique, une chimiothérapie antérieure ou une exposition à des agents environnementaux tels que le benzène⁹⁻¹¹.

Le développement du séquençage à haut débit a permis de mieux comprendre la LMA et de mettre en évidence une importante hétérogénéité génétique⁹. Les recommandations de l'*European Leukemia Network* (ELN) 2022 permettent ainsi de classer les patients en trois catégories : favorable, intermédiaire et défavorable, sur la base de leurs anomalies cytogénétiques et mutationnelles, afin d'estimer la probabilité de réponse au traitement, le risque de rechute et la survie globale. Cette hétérogénéité clinique et moléculaire explique en partie la diversité des réponses thérapeutiques observées et constitue l'un des principaux défis dans la prise en charge de la maladie¹⁰.

1.1.1 Diagnostic

Le diagnostic de la LMA repose sur la combinaison des données cliniques, morphologiques, immunophénotypiques, cytogénétiques et moléculaires⁹. La maladie est généralement suspectée chez les patients présentant des anomalies hématologiques inexplicables telles qu'une cytopénie, une hyperleucocytose ou la présence de blastes circulants¹⁰. La confirmation diagnostique repose en grande partie sur l'examen de la MO, permettant de quantifier les blastes et de réaliser les analyses nécessaires à la caractérisation complète de la maladie^{7,10}.

La caractérisation moléculaire de la LMA met en évidence des anomalies récurrentes impliquées dans le développement, la classification et le pronostic de la maladie^{12,13}. Parmi ces altérations, certaines mutations, notamment celle du gène *Nucleophosmin 1 (NPM1)*, ainsi que certaines anomalies cytogénétiques récurrentes, comme la monosomie du chromosome 7, constituent des éléments importants pour l'identification de sous-groupes biologiques distincts. Leur présence influence l'évaluation du risque de rechute et la stratégie thérapeutique¹³. Ces deux altérations occupent une place centrale dans ce travail et seront détaillées ultérieurement.

1.1.2 Traitement

Le traitement de la LMA repose sur deux phases successives : une phase d'induction visant à éliminer la majorité des cellules leucémiques, suivie d'une phase de consolidation destinée à éradiquer les cellules résiduelles et à prévenir les rechutes⁷.

L'éligibilité à un traitement intensif repose sur plusieurs critères, notamment l'âge, l'état général du patient, ainsi que le profil moléculaire et cytogénétique de la maladie^{10,14}. Chez les patients âgés, la LMA est fréquemment associée à des anomalies cytogénétiques défavorables et à une résistance accrue aux traitements intensifs^{15,16}.

Chez les patients éligibles, le traitement d'induction repose généralement sur le protocole 7+3, associant cytarabine (AraC) et une anthracycline, administrées respectivement pendant sept et trois jours^{1,15}. L'identification de mutations spécifiques au diagnostic permet également d'associer certaines thérapies ciblées au traitement standard, notamment les inhibiteurs de *FLT3 (Fms-like tyrosine kinase 3)*, tels que la midostaurine, utilisés chez les patients porteurs d'une mutation activatrice de *FLT3*^{9,17}. Chez les patients inéligibles, des approches thérapeutiques moins agressives ont été développées¹⁴. L'association de vénétoclax, un inhibiteur sélectif de *BCL-2 (B-cell lymphoma 2)* favorisant l'apoptose des cellules leucémiques, à un agent hypométhylant, comme l'azacitidine ou la décitabine, constitue une option thérapeutique pour cette population^{14,16,18}.

La stratégie de consolidation est guidée par le profil de risque ELN : elle peut reposer sur une chimiothérapie à haute dose ou, chez les patients présentant un risque élevé de rechute, notamment en présence d'anomalies cytogénétiques ou moléculaires défavorables, d'une maladie résiduelle mesurable (MRD) positive après induction ou d'une LMA secondaire, sur une allogreffe de cellules souches hématopoïétiques (CSH)^{15,19,20}. Malgré ces avancées, la rechute reste l'un des principaux défis thérapeutiques de la LMA et demeure la principale cause de mortalité²¹. Elle peut concerner jusqu'à environ deux tiers des patients, avec une survie post-rechute qui reste limitée, souvent estimée entre 10 et 40 %²².

1.1.3 Réponse au traitement et rémission

L'objectif initial du traitement de la LMA est d'induire une rémission complète (RC)²³. Celle-ci est définie par des critères morphologiques et hématologiques stricts : moins de 5 % de blastes dans la MO, l'absence de blastes circulants dans le sang périphérique et une récupération hématologique, correspondant à un taux de polynucléaires neutrophiles supérieur à 1 000/ μ L et à un taux de plaquettes supérieur à 100 000/ μ L^{19,20}. Elle implique également une indépendance transfusionnelle, c'est-à-dire l'absence de besoin de transfusion pour maintenir des concentrations adéquates en globules rouges ou en plaquettes^{23,24}.

Il est désormais établi que la RC, même lorsqu'elle remplit tous les critères morphologiques et hématologiques, reste un indicateur imparfait du devenir à long terme des patients^{7,23}. En effet, de nombreux patients en RC morphologique conservent une MRD, détectable uniquement par des techniques très sensibles, notamment la cytométrie en flux multiparamétrique (CMF)^{7,23}.

La MRD constitue un paramètre pronostique majeur, permettant d'évaluer la profondeur de la réponse thérapeutique en détectant des cellules leucémiques à des niveaux bien inférieurs au seuil de détection morphologique^{13,25}. En CMF, une MRD est généralement considérée comme positive lorsque des cellules leucémiques résiduelles sont détectées au-dessus du seuil de sensibilité de la méthode, souvent autour de 10^{-3} , soit environ 0,1 %²⁶. Une MRD positive post-traitement est associée à un risque accru de rechute et peut justifier une stratégie thérapeutique plus intensive, notamment le recours à une allogreffe de CSH^{27,28}. À l'inverse, une MRD indétectable après l'induction ou la consolidation, généralement définie par l'absence de détection de cellules leucémiques au seuil de sensibilité mentionné précédemment, est associée à une diminution significative du risque de rechute et à une amélioration de la survie globale^{4,26}. Cependant, elle ne garantit pas l'absence totale de cellules leucémiques résiduelles, mais indique que celles-ci ne sont pas détectées au-dessus du seuil défini par la technique utilisée^{28,29}.

Au-delà de leur valeur pronostique, les approches utilisées en routine pour évaluer la MRD permettent de détecter et de caractériser certains profils immunophénotypiques aberrants²⁶. Cependant, elles restent limitées par les panels d'anticorps utilisés et ne permettent pas toujours de caractériser finement l'identité transcriptionnelle, mutationnelle ou l'hétérogénéité de ces cellules au niveau unicellulaire^{26,28,29}. La réponse au traitement ne peut donc pas être comprise uniquement comme une réduction globale de la masse leucémique. Elle soulève également une question centrale : quelles cellules survivent à la chimiothérapie, et dans quelle mesure participent-elles à la rechute ou à la maladie réfractaire ?

1.1.4 Rechute et maladie réfractaire

Malgré l'obtention fréquente d'une RC, la rechute demeure l'une des principales causes de mortalité dans la LMA. Elle survient chez environ 40 % des patients jeunes et chez une proportion plus élevée des patients âgés²¹. La rechute correspond à la réapparition de la maladie après obtention d'une RC, se manifestant généralement par une réémergence des blastes dans la MO ou par une atteinte extramédullaire²³.

La survie globale médiane post-rechute est de l'ordre de 5 à 6 mois³⁰. Le pronostic est particulièrement défavorable chez les patients présentant une rechute précoce, survenant moins de 6 mois après la RC, chez les patients âgés et chez ceux présentant des anomalies cytogénétiques défavorables, telles que la monosomie 7, la délétion 5q, les anomalies du chromosome 17 ou un caryotype complexe, ainsi que certaines mutations de mauvais pronostic comme *TP53*^{4,31-33}.

La maladie réfractaire, définie par l'absence de RC après deux cycles d'induction, représente une situation encore plus critique³⁴. Ces patients, considérés en échec primaire, ont un pronostic très défavorable et disposent d'options thérapeutiques limitées malgré l'émergence de nouveaux traitements^{21,25}. Bien que la rechute et la maladie réfractaire correspondent à des situations cliniques différentes, elles partagent une problématique biologique commune : comment certaines cellules leucémiques parviennent-elles à survivre malgré la pression thérapeutique ?

Cette persistance suggère que certaines cellules possèdent, ou acquièrent au cours du temps et du traitement, des propriétés leur permettant d'échapper à la chimiothérapie^{21,25,35}. Parmi les populations susceptibles de contribuer à ce phénomène, deux sont particulièrement pertinentes : les cellules souches leucémiques (CSL) et les *Drug-Tolerant Persisters* (DTPs). Les CSL correspondent à des cellules leucémiques capables d'auto-renouvellement et de régénération de la maladie, tandis que les DTPs désignent des cellules capables de survivre transitoirement au traitement sans nécessairement dépendre de nouvelles mutations de résistance^{21,25,35}. Ces deux concepts constituent le cœur biologique de ce travail.

Les principales trajectoires cliniques possibles après traitement sont résumées dans la Figure 1. Celle-ci illustre l'évolution possible de la charge leucémique après l'induction, depuis l'obtention d'une RC avec persistance éventuelle d'une MRD, jusqu'à la rechute ou la maladie réfractaire^{21,23,29,34}. Ainsi, même lorsqu'une réponse initiale favorable est obtenue, la persistance de cellules leucémiques peut permettre la reprise de la maladie après traitement. Ces observations renforcent l'idée que l'échec thérapeutique ne dépend pas uniquement de la réponse initiale au traitement, mais également de la capacité de certaines cellules leucémiques à survivre à la pression exercée par la chimiothérapie^{28,36,37}.

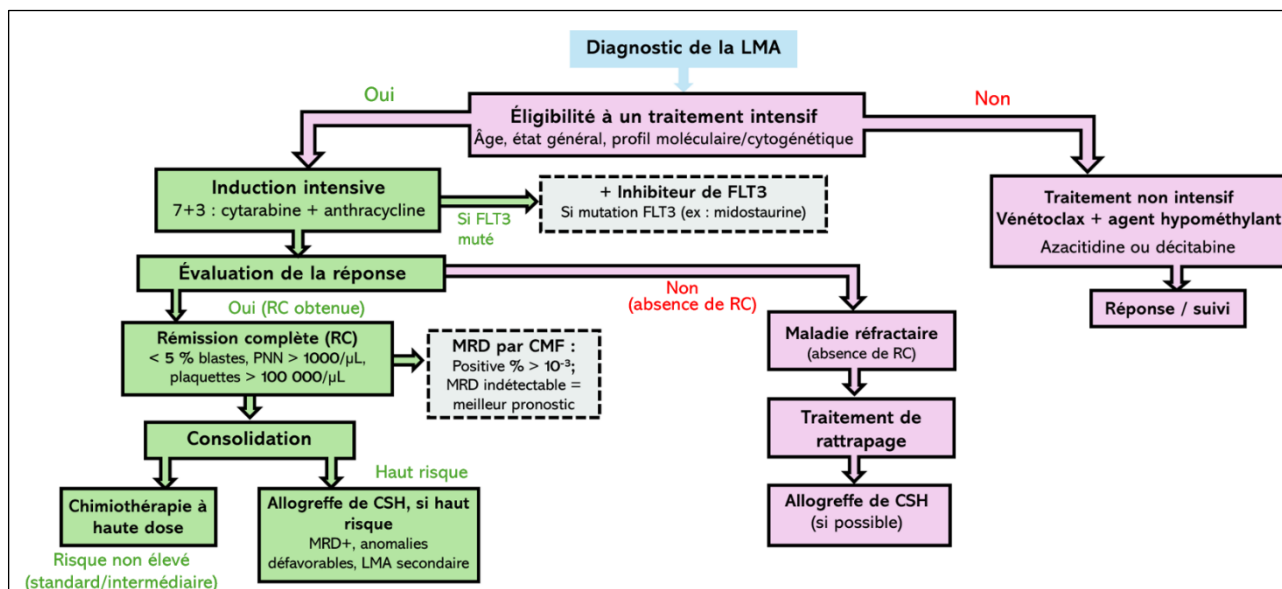


Figure 1. Principales trajectoires thérapeutiques dans la LMA.

Schéma simplifié des principales stratégies thérapeutiques après diagnostic de LMA, incluant l'induction intensive, l'évaluation de la réponse, la consolidation, l'allogreffe et les traitements non intensifs.

Schéma original réalisé dans le cadre de ce mémoire, adapté des recommandations ELN⁴.

1.2 Bases biologiques de la persistance leucémique

Les trajectoires cliniques décrites précédemment montrent que la rechute et la maladie réfractaire ne peuvent pas être expliquées uniquement par l'efficacité initiale du traitement^{21,25,28,36}. Elles impliquent également la survie de cellules leucémiques capables de persister malgré la pression thérapeutique^{28,36,37}. Pour comprendre l'origine de cette capacité de persistance, il est nécessaire de revenir à l'organisation de l'hématopoïèse physiologique et aux propriétés des CSH^{38,39}. Certaines de ces propriétés, telles que l'auto-renouvellement, la quiescence et les interactions avec la niche médullaire, peuvent être détournées au cours de la transformation leucémique et contribuer à la persistance des cellules leucémiques^{7,40}.

1.2.1 Hématopoïèse physiologique

L'hématopoïèse physiologique repose sur des CSH localisées dans la MO, caractérisées par leur capacité d'auto-renouvellement et leur multipotence, leur permettant de générer l'ensemble des cellules sanguines (Figure 2)^{38,39,41}. Ces propriétés sont essentielles au maintien de l'hématopoïèse physiologique. Les CSH donnent naissance à des progéniteurs progressivement engagés vers les lignées myéloïdes et lymphoïdes, à l'origine notamment des globules rouges, des plaquettes, des granulocytes, des monocytes/macrophages et des lymphocytes^{41,42}. Dans le contexte de la LMA, ces propriétés peuvent être réutilisées ou dérégulées par des cellules leucémiques, contribuant à leur survie et à leur résistance au traitement^{2,33,34}.

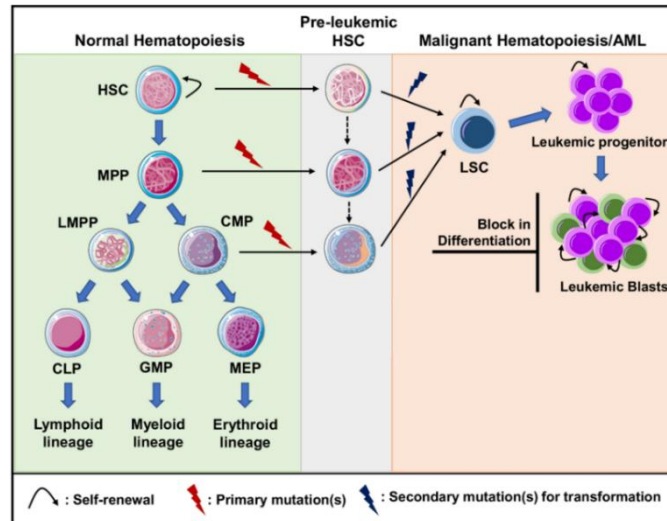


Figure 2. Progression de l'hématopoïèse normale à la LMA.

Les mutations de classe I (éclair rouge) s'accumulent dans les cellules souches hématopoïétiques (CSH) générant des cellules pré-leucémiques qui acquièrent des mutations de classe II (éclair bleu) ce qui entraîne l'émergence de cellules souches leucémiques (CSL). Ces mutations bloquent la différenciation, aboutissant à l'accumulation de cellules immatures (blastes).

Figure adaptée de Long *et al.*²

Dans la MO, les CSH interagissent avec le microenvironnement médullaire, qui constitue une niche régulant leur prolifération, leur différenciation et leur quiescence, par l'intermédiaire de signaux moléculaires tels que CXCL12, SCF ou IL-7, ainsi que d'interactions cellulaires d'adhésion^{38,39,43,44}. Ce microenvironnement peut également contribuer à la protection des cellules leucémiques face aux traitements, en fournissant des signaux de survie et de quiescence, en favorisant leur adhésion aux cellules stromales et en permettant aux cellules leucémiques de détourner des voies normalement utilisées par les CSH^{44,45}.

1.2.2 Transformation leucémique et hématopoïèse

Le passage de l'hématopoïèse physiologique à la LMA est un phénomène progressif et en plusieurs étapes, trouvant le plus souvent son origine dans les CSH. Ce processus implique l'accumulation d'altérations génétiques et épigénétiques, ainsi qu'un dérèglement du microenvironnement médullaire, et aboutit à l'émergence de cellules souches leucémiques (CSL)^{2,3,11}.

Dans un premier temps, des altérations génétiques s'accumulent progressivement dans des CSH encore fonctionnelles, touchant notamment des régulateurs épigénétiques tels que *DNMT3A*, *TET2* ou *ASXL1*^{44,46,47}. Ces CSH mutées restent fonctionnelles, mais acquièrent un avantage compétitif, c'est-à-dire une meilleure capacité à survivre, à se maintenir et à s'imposer progressivement dans la MO par rapport aux cellules non mutées⁴⁶. Elles peuvent notamment mieux survivre au stress environnemental et conserver une capacité accrue d'auto-renouvellement^{47,48}. Progressivement, ces clones s'imposent et représentent une proportion croissante des cellules de la MO. Ce phénomène, appelé hématopoïèse clonale, peut se maintenir silencieusement pendant plusieurs années sans provoquer de LMA déclarée⁴⁹.

L'expansion des clones pré-leucémiques est favorisée par un environnement inflammatoire, notamment dans un contexte de vieillissement⁵⁰. Certaines mutations pré-leucémiques, comme celles touchant *DNMT3A* ou *TET2*, peuvent modifier la réponse inflammatoire des cellules hématopoïétiques et favoriser la production de cytokines pro-inflammatoires, telles que l'IL-1 β , l'IL-6 ou le TNF- α . Ces signaux contribuent à créer un environnement défavorable aux CSH physiologiques, mais plus permissif pour l'expansion des clones mutés^{48,50,51}.

Parallèlement, la niche médullaire peut être modifiée dans les états pré-leucémiques et leucémiques. Les cellules leucémiques peuvent remodeler la vascularisation médullaire, en particulier au niveau des niches endostéales et centrales, et produire des facteurs favorisant l'angiogenèse^{44,45}. Ces modifications perturbent le soutien apporté aux CSH physiologiques, favorisent la compétition pour l'occupation de la niche et permettent aux cellules leucémiques de détourner des voies normalement utilisées par les CSH^{44,45,52}. La transformation leucémique ne résulte pas uniquement de l'accumulation de mutations dans les cellules hématopoïétiques, mais aussi d'interactions avec un microenvironnement capable de favoriser la survie et l'expansion de certains clones^{44,45,52}.

Ces altérations génétiques, épigénétiques et microenvironnementales créent des conditions favorables à l'émergence de cellules leucémiques. Parmi elles, certaines altérations récurrentes jouent un rôle important dans ce mémoire, car elles permettent l'identification des cellules dites *AML-like*, dont le profil mutationnel ou transcriptionnel est compatible avec une origine leucémique⁴.

1.2.3 Altérations génétiques d'intérêt

La LMA est caractérisée par une grande diversité d'altérations cytogénétiques et moléculaires, impliquées dans la classification diagnostique, la stratification pronostique et, pour certaines, dans l'orientation thérapeutique. Parmi les altérations fréquemment prises en compte figurent notamment des mutations de *NPM1*, *FLT3*, *CEBPA*, *IDH1/2*, *TP53*, *RUNX1* ou *ASXL1*, ainsi que des anomalies cytogénétiques telles que la perte complète ou partielle du chromosome 5 (-5/del(5q)) ou la perte complète ou partielle du chromosome 7 (-7/del(7q))^{1,53,54}.

Dans ce mémoire, l'attention sera portée sur la mutation du gène *NPM1* et sur la monosomie du chromosome 7, notamment parce que ces altérations font partie des anomalies récurrentes importantes dans la LMA. Les mutations de *NPM1* représentent l'un des sous-groupes moléculaires les plus fréquents, tandis que la perte du chromosome 7 fait partie des anomalies cytogénétiques récurrentes associées aux formes défavorables de la maladie⁵⁵. Par ailleurs, les données analysées proviennent de patients présentant l'une ou l'autre de ces altérations. Ces deux altérations servent également de base à l'identification des cellules *AML-like* dans les annotations de référence associées à ces données⁵⁶.

1.2.3.1 *NPM1*

La mutation de *NPM1* (*NPM1^{mut}*) est retrouvée chez approximativement 30 % des patients adultes atteints de LMA⁵³. Elle est fréquemment associée aux LMA *de novo*, apparaissant généralement en l'absence de maladie myéloïde antérieure connue⁵⁷. Cette mutation est considérée comme une mutation driver, dont l'apparition constitue une étape importante dans l'installation de la LMA^{53,57,58}.

Le gène *NPM1*, localisé sur le chromosome 5, code pour une protéine chaperonne principalement localisée dans le nucléole. Les mutations de *NPM1* dans la LMA touchent principalement l'exon 12⁵⁹. Plusieurs types de mutations ont été décrits, mais la plus fréquente est la mutation dite de type A, correspondant à une insertion/duplication de quatre paires de bases, généralement TCTG, et représentant environ 75 à 80 % des mutations de *NPM1*⁵⁹. Cette altération provoque un décalage du cadre de lecture au niveau de l'extrémité C-terminale et entraîne une localisation cytoplasmique aberrante de la protéine *NPM1* mutée^{53,57}. La délocalisation de la protéine perturbe certaines fonctions nucléaires de *NPM1* et contribue à la dérégulation du programme transcriptionnel leucémique⁴⁹.

La mutation de *NPM1* favorise le maintien d'un programme transcriptionnel immature, notamment par l'activation des gènes *HOX* et de leur cofacteur *MEIS1*⁵³. Elle participe ainsi au blocage de la différenciation myéloïde et au maintien de l'auto-renouvellement, deux caractéristiques centrales de la transformation leucémique^{53,57,60}.

1.2.3.2 Monosomie du chromosome 7

La monosomie du chromosome 7 (del(7)), correspondant à la perte complète d'un exemplaire du chromosome 7, est l'anomalie chromosomique autosomique la plus fréquente chez les patients atteints de LMA à haut risque⁵⁴. Contrairement aux mutations de *NPM1*, plus souvent associées aux LMA *de novo*, elle est davantage associée à des LMA secondaires, survenant après un syndrome myélodysplasique, ainsi qu'aux LMA liées aux traitements, notamment après exposition à des agents alkylants ou à des radiations^{57,58,61}.

Sur le plan moléculaire, cette perte chromosomique entraîne une haplo-insuffisance, c'est-à-dire une réduction de l'expression de plusieurs gènes localisés sur le chromosome 7⁵⁴. Elle affecte notamment des gènes impliqués dans la régulation transcriptionnelle ou épigénétique, tels que *EZH2*, *CUX1*, *SAMD9* et *SAMD9L*^{54,58,62}. Cette perte génétique est associée à une LMA particulièrement agressive, en partie en raison de son impact sur l'homéostasie des CSH^{54,62}.

1.2.4 Émergence des cellules souches leucémiques (CSL)

L'accumulation progressive de ces altérations aboutit à l'émergence de populations cellulaires aux propriétés dérégulées. Pour comprendre comment certaines de ces cellules peuvent persister après

traitement, il est utile de distinguer deux types de mutations impliquées dans la leucémogénèse : les mutations germinales et les mutations somatiques^{58,61,62}.

Les altérations impliquées dans la LMA peuvent être d'origine germinale ou somatique⁶². Les mutations germinales, héréditaires et présentes dans l'ensemble des cellules de l'organisme, peuvent prédisposer au développement de la maladie^{61,62}. Elles ne sont toutefois pas suffisantes à elles seules pour déclencher une LMA, car elles créent surtout un terrain de prédisposition nécessitant l'acquisition d'altérations somatiques supplémentaires^{62,63}. À l'inverse, les mutations somatiques sont acquises au cours de la vie cellulaire. Elles peuvent s'accumuler dans les cellules pré-leucémiques ou leucémiques et être partagées par des cellules issues d'un même clone^{58,61}. Lorsqu'elles sont partagées par plusieurs sous-populations, elles peuvent servir de marqueurs pour reconstruire leur trajectoire clonale^{57,61}. Cette propriété est utile dans la LMA, où la maladie évolue selon une architecture clonale complexe impliquant plusieurs sous-clones pouvant diverger au cours du temps (Figure 3)^{1,58,64}.

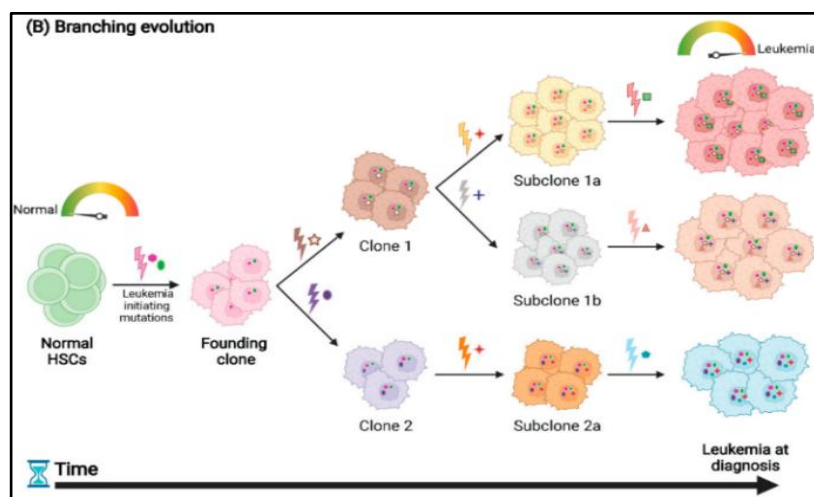


Figure 3. Évolution clonale ramifiée dans la LMA.

Représentation schématique de l'émergence progressive de sous-clones leucémiques à partir d'un clone fondateur. L'accumulation successive de mutations au sein des différents clones contribue à augmenter l'hétérogénéité intratumorale.

Figure issue de Desai et al., 2022⁷⁷.

Dans le modèle dit « *two-hit* », ou hypothèse des deux coups, un premier évènement initiateur, souvent une mutation précoce touchant des régulateurs épigénétiques tels que *DNMT3A*, *TET2* ou *ASXL1*, favorise l'expansion clonale, tandis que l'acquisition d'altérations somatiques additionnelles aboutit à la transformation leucémique complète^{49,54,57}. Cette progression aboutit à l'émergence de cellules souches leucémiques (CSL), caractérisées par une capacité d'auto-renouvellement dérégulée et un blocage de la différenciation myéloïde (Figure 2)^{2,61}.

En raison de leurs propriétés de type cellule souche, notamment l'auto-renouvellement, la quiescence et les interactions avec la niche médullaire, les CSL ont été historiquement considérées comme des cellules capables de reconstituer la LMA humaine après transplantation dans des modèles murins

immunodéficients^{65,66}. Sur cette base, elles ont longtemps été proposées comme des cellules responsables de la rechute, en raison de leur capacité à régénérer la maladie⁶⁷. Cependant, cette hypothèse a été largement nuancée par des études récentes montrant que les CSL ne sont pas enrichies dans la population des cellules leucémiques survivant à la chimiothérapie^{68,69}. Les cellules survivantes peuvent plutôt correspondre à des états transitoires de tolérance au traitement, appelés DTPs, capables de survivre à la pression thérapeutique et de contribuer à la reprise de la maladie après traitement^{68,69}.

1.2.5 Drug-Tolerant Persisters

Au diagnostic, plusieurs types cellulaires coexistent dans la LMA, notamment des CSH saines, des CSL et des blastes leucémiques, ces derniers constituant l'essentiel de la masse tumorale. Après traitement, cette masse tumorale diminue fortement, jusqu'à un nadir correspondant à la phase où la maladie résiduelle est minimale. Une fraction des cellules persiste néanmoins à ce stade et peut constituer un réservoir cellulaire à partir duquel une reprise leucémique peut émerger, contribuant ainsi à la rechute^{70,71}. Cette dynamique est illustrée dans la Figure 4.

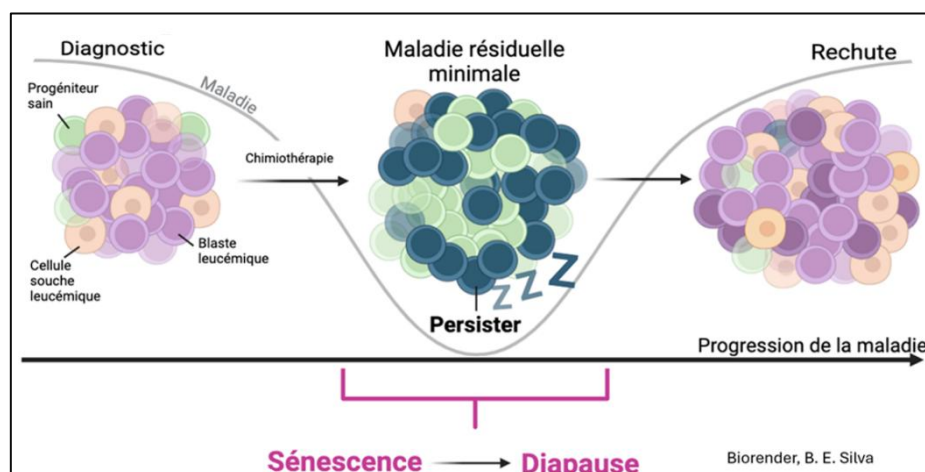


Figure 4. Dynamique des DTPs au cours du traitement de la LMA

Schéma illustrant la diminution de la masse tumorale leucémique après traitement. Une fraction de cellules persiste au nadir et peut ensuite contribuer à la régénération de la maladie lors de la rechute.

Figure réalisée par Bianca E. Silva

Contrairement à la résistance génétique, qui repose sur l'acquisition de mutations conférant un avantage sélectif, les DTPs sont définies par leur capacité à survivre au traitement sans dépendre de nouvelles mutations⁶⁴. Elles adoptent plutôt un état phénotypique transitoire et potentiellement réversible^{71,72}. Leur survie repose sur un ralentissement de la prolifération et une adaptation au stress thérapeutique. Ces mécanismes leur permettent de persister durant le traitement, notamment pendant l'induction ou la consolidation, puis de reprendre un programme prolifératif lorsque la pression thérapeutique diminue^{64,70,71}.

Sous l'effet du stress génotoxique, les DTPs peuvent adopter des états qualifiés de *senescence-like* et *diapause-like*^{64,69}. Ces états ne correspondent pas à une sénescence ou à une diapause strictes, mais plutôt à des programmes de ralentissement de la prolifération, de tolérance au stress et d'économie énergétique⁶⁹. Le terme *diapause-like* fait référence, par analogie, à la diapause, c'est-à-dire à un état réversible d'arrêt temporaire du développement permettant la survie en conditions défavorables^{61,64-67}. Dans ce contexte, les cellules quiescentes échappent plus facilement à l'action des agents cytotoxiques, qui ciblent préférentiellement les cellules en division.

Ces mécanismes de tolérance impliquent plusieurs axes complémentaires. Sur le plan épigénétique, les DTPs présentent des modifications de l'état de la chromatine et des marques d'histones^{71,72,76}. Leur métabolisme peut également s'appuyer davantage sur la phosphorylation oxydative, ce qui peut soutenir leur survie dans un contexte de stress thérapeutique^{64,71,77}. Par ailleurs, des travaux récents ont montré que les DTPs peuvent augmenter transitoirement la rigidité de leur membrane plasmique par remodelage du lipidome, ce qui réduit l'accumulation intracellulaire des agents chimiothérapeutiques et renforce leur protection face au traitement^{64,71}.

À ces mécanismes intrinsèques s'ajoute la niche médullaire, qui joue un rôle complémentaire en fournissant des signaux de survie, en favorisant la quiescence et en modulant la réponse au stress⁷¹. Cette combinaison de facteurs intrinsèques et extrinsèques confère aux DTPs une capacité de survie accrue face aux agents cytotoxiques⁷¹.

La relation entre les DTPs et les CSL reste encore mal comprise. Plusieurs modèles ont été proposés dans la littérature pour expliquer l'origine des cellules persistantes après traitement. Un premier modèle suggère que ces cellules correspondent à une sous-population préexistante, déjà dotée de propriétés de résistance⁵⁶. Un second modèle propose qu'elles résultent plutôt d'un état transitoire induit par la pression thérapeutique⁶⁹. Ces deux modèles ne sont pas nécessairement exclusifs, mais ils posent une question centrale : les DTPs correspondent-elles principalement à une population déjà présente avant traitement ou à un état adaptatif induit par la chimiothérapie ?^{56,69}

Les travaux de Naldini *et al.* s'inscrivent plutôt dans le premier modèle. Ils suggèrent que les cellules à l'origine de la rechute après chimiothérapie pourraient correspondre à des CSL préexistantes, caractérisées par une signature transcriptomique *miR-126^{high}*. Dans cette étude, les cellules *miR-126^{high}* correspondent à une sous-population de blastes majoritairement quiescents, enrichis en signatures de type CSL et dotés d'une capacité de résistance intrinsèque⁵⁶.

À l'inverse, les travaux de Duy *et al.* nuancent cette interprétation. Ils montrent que les cellules survivant au nadir ne présentent pas nécessairement un enrichissement initial en marqueurs classiques de CSL, mais adoptent surtout des états *senescence-like* et *diapause-like*. Les signatures de CSL apparaissent

davantage au moment de la rechute, suggérant que le phénotype CSL pourrait être acquis ou renforcé après le traitement⁶⁹.

Ainsi, la persistance leucémique pourrait résulter à la fois de propriétés préexistantes présentes dans certaines cellules leucémiques au diagnostic et de programmes adaptatifs induits par la chimiothérapie. Cette incertitude rend nécessaire l'étude des cellules leucémiques à haute résolution, dès le diagnostic et au cours du traitement, afin de déterminer si certaines sous-populations présentent déjà des caractéristiques compatibles avec une survie post-thérapeutique.

Dans ce mémoire, nous faisons l'hypothèse qu'une sous-population de cellules leucémiques, déjà présente au diagnostic, pourrait posséder des caractéristiques la rendant plus apte à survivre à la chimiothérapie. En exploitant des données de séquençage d'ARN à cellule unique (scRNA-seq) longitudinales et une approche fondée sur la charge mutationnelle à l'échelle unicellulaire, nous cherchons à évaluer l'existence et les caractéristiques de sous-populations *AML-like* associées à la persistance leucémique chez les patients présentant une mutation du gène *NPM1* ou une monosomie du chromosome 7.

1.3 Défis liés à l'étude de la LMA

Tester cette hypothèse nécessite de pouvoir identifier, à l'échelle unicellulaire, les cellules leucémiques susceptibles de persister, tout en les distinguant des cellules normales ou pré-leucémiques présentes dans la MO. Cette identification est complexe, car elle implique de prendre en compte l'hétérogénéité cellulaire, la plasticité induite par le traitement et la proximité transcriptionnelle entre cellules leucémiques immatures et cellules hématopoïétiques saines^{35,74,78,79}.

L'identification précise des populations impliquées dans la persistance et la rechute constitue un défi biologique et technologique majeur dans l'étude de la LMA^{35,64}. En effet, la LMA ne correspond pas à une population homogène de blastes, mais à un ensemble hétérogène de cellules leucémiques présentant différents états de différenciation^{35,78}. Au sein d'un même patient, ces cellules peuvent présenter des capacités variables de prolifération, d'auto-renouvellement, de résistance au traitement et de contribution à la rechute^{80,81}. Dans ce contexte, certaines sous-populations rares, parfois peu représentées dans la masse tumorale, peuvent néanmoins jouer un rôle important dans la persistance après traitement^{64,81}.

À cette hétérogénéité s'ajoute une plasticité cellulaire importante. Certaines cellules leucémiques peuvent changer de programme cellulaire en réponse au stress thérapeutique, sans acquérir de nouvelles mutations, comme décrit précédemment pour les cellules DTPs^{71,78}. Cette plasticité complique l'analyse et l'interprétation des échantillons, car les populations observées à un temps

donné ne reflètent pas nécessairement des états fixes, mais peuvent correspondre à des états transitoires influencés par le traitement^{82,83}. L'analyse de la LMA nécessite donc des approches capables de résoudre cette diversité à haute résolution, afin d'identifier les populations rares et de mieux comprendre leur dynamique au cours du traitement^{82,83}.

Un autre défi majeur réside dans la proximité biologique entre certaines cellules leucémiques et les cellules souches ou progéniteurs hématopoïétiques sains³⁵. En effet, les cellules leucémiques peuvent adopter ou conserver des programmes transcriptionnels proches de ceux de l'hématopoïèse normale, ce qui complique leur distinction avec les CSH ou les progéniteurs sains³⁵. Cette difficulté s'accroît après chimiothérapie, lorsque les cellules leucémiques résiduelles, souvent peu abondantes, peuvent coexister avec des cellules normales ou pré-leucémiques dans une moelle en cours de régénération⁷⁸. Certaines cellules normales en régénération peuvent alors partager des marqueurs de surface, tels que CD34, ou des états transcriptionnels proches des cellules leucémiques, tels que les programmes associés à la quiescence, au stress ou à l'inflammation⁸⁴.

1.3.1 Limites des approches classiques

L'approche de référence dans l'étude de la LMA repose sur des analyses de CMF^{56,81}. Bien que cette technique reste cruciale pour le diagnostic et le suivi clinique de la MRD, elle présente certaines limitations dans un contexte de recherche visant à comprendre l'hétérogénéité cellulaire, transcriptionnelle et clonale de la maladie⁸¹. La principale limitation est sa dépendance à un panel d'antigènes prédéfini et limité. Par exemple, le panel recommandé par l'ELN pour le suivi de la MRD inclut notamment : CD45, CD33, CD13, CD7, CD56, CD34, CD117, et HLA-DR⁶⁹. Or, la LMA présente une hétérogénéité immunophénotypique importante, et le phénotype des cellules leucémiques peut évoluer en dehors des antigènes étudiés. L'expression du CD36 a notamment été décrite comme un marqueur clé des DTPs, illustrant le fait que certains états cellulaires pertinents peuvent ne pas être capturés par les panels classiques⁸⁵. Cette dépendance à un nombre restreint de marqueurs peut limiter la caractérisation fine et le suivi longitudinal de la maladie^{35,81,86}. De plus, il existe une hétérogénéité phénotypique importante au sein de la population de blastes, avec des marqueurs de référence, comme CD34 ou CD117, qui ne sont pas toujours exprimés à des niveaux détectables par la totalité de la population des cellules malignes^{35,80}. Se baser uniquement sur un nombre limité de marqueurs phénotypiques reste donc insuffisant pour identifier l'ensemble des cellules leucémiques^{35,80}.

Les approches omiques représentent une solution complémentaire aux limites de la CMF, puisqu'elles permettent l'étude de l'expression d'un nombre important de protéines ou de gènes sans dépendre d'un ensemble limité de cibles prédéfinies^{87,88}. Cependant, les approches omiques de type *bulk*, notamment

le séquençage d'ARN en population cellulaire (*bulk RNA sequencing*, *bulk RNA-seq*), restent également limitées par leur résolution car elles ne permettent de capturer que l'expression moyenne des protéines ou des gènes à l'échelle de l'échantillon^{88,89}.

Pour pallier cette limitation, les approches omiques à cellule unique (*single-cell*), et en particulier le scRNA-seq, ont été développées. Le scRNA-seq fournit une mesure de l'expression des gènes détectables dans chaque cellule individuelle analysée^{88,89}. Ce type d'approche permet donc une analyse fine de l'hétérogénéité intra-tumorale de la LMA, ainsi qu'un suivi longitudinal des populations cellulaires au cours du traitement^{63,71,88-90}. Cette technique a donc été choisie comme principale source de données pour le présent mémoire.

1.3.2 Apport du scRNA-seq dans la LMA

Face aux limitations des approches décrites précédemment, le scRNA-seq constitue un outil pertinent pour l'étude de la LMA⁸⁹. Sur le plan technique, les approches *droplet-based* de type 10x Genomics reposent sur l'encapsulation de cellules individuelles dans des gouttelettes, permettant d'associer à chaque cellule un code-barres cellulaire spécifique. Les transcrits capturés sont également marqués par des identifiants moléculaires uniques (*unique molecular identifiers*, UMIs), permettant de distinguer les molécules d'ARN réellement capturées des duplicats d'amplification^{89,91}. Après séquençage, les données sont organisées sous forme d'une matrice d'expression gènes-cellules, constituant la base des analyses bioinformatiques ultérieures^{89,91}. Le principe général de ce *pipeline* est schématisé dans la Figure 5.

Dans le cadre de la LMA, cette technologie permet d'étudier la coexistence de différents états cellulaires au sein d'un même échantillon, incluant des cellules normales ou potentiellement pré-leucémiques et des populations associées à la persistance après traitement⁷⁸. L'étude de van Galen *et al.*, l'une des premières études majeures de scRNA-seq dans la LMA, a montré que les cellules leucémiques peuvent être organisées en plusieurs états transcriptionnels le long d'un axe de différenciation myéloïde, révélant une hétérogénéité cellulaire qui ne peut pas être résolue par les approches *bulk*⁸⁹. Petti *et al.* ont ensuite montré qu'il était possible de détecter certaines mutations exprimées directement dans les données de scRNA-seq, permettant de distinguer les cellules leucémiques des cellules non leucémiques à l'échelle unicellulaire⁷⁹.

Plus récemment, Naldini *et al.* ont appliqué le scRNA-seq à des échantillons longitudinaux de patients atteints de LMA, notamment des patients présentant une mutation de *NPM1* ou une monosomie du chromosome 7. Cette étude a permis de suivre les cellules leucémiques avant et après chimiothérapie, en distinguant les cellules malignes de l'hématopoïèse régénérative grâce à la détection de la mutation *NPM1* ou à l'analyse de l'expression des gènes du chromosome 7⁵⁶. Les auteurs ont ainsi mis en évidence des réponses transcriptionnelles hétérogènes au traitement⁵⁶.

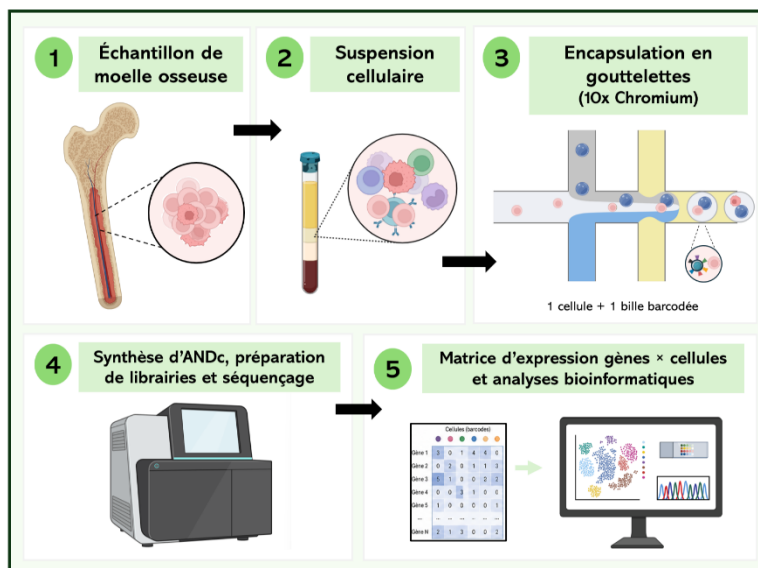


Figure 5. Principe général du scRNA-seq basé sur la technologie 10x Genomics.

Schéma simplifié des principales étapes du *pipeline* scRNA-seq, depuis la préparation d'un échantillon de moelle osseuse jusqu'à la génération d'une matrice d'expression gènes × cellules et aux analyses bioinformatiques ultérieures.

Figure réalisée avec BioRender

1.3.3 Limites de la discrimination *AML-like* / *non-AML-like* en scRNA-seq

Une limitation majeure des études de scRNA-seq dans la LMA est la distinction entre les cellules malignes, dites *AML-like*, et les cellules saines, dites *non-AML-like*. En effet, il n'existe pas de marqueur transcriptionnel unique, ni de combinaison universelle de marqueurs transcriptionnels uniques, permettant d'identifier les blastes leucémiques chez tous les patients^{79,89}. Les approches actuelles reposent souvent sur la détection d'altérations génétiques informatives, telles que des variants nucléotidiques simples (*Single Nucleotide Variant*, SNV), des insertions, des délétions ou des anomalies chromosomiques, afin de distinguer les cellules leucémiques des cellules non leucémiques⁷⁹.

Par exemple, van Galen *et al.* ont utilisé des mutations dans des gènes comme *NPM1*, *DNMT3A*, *RUNX1*, *FLT3* ou *TP53* pour contribuer à la discrimination entre cellules leucémiques et non leucémiques⁸⁹. Dans l'étude de Naldini *et al.*, dont les données brutes ont été réutilisées dans ce mémoire, la discrimination des cellules leucémiques chez les patients *NPM1^{mut}* repose sur la détection de transcrits porteurs d'insertions dans le gène *NPM1* au niveau unicellulaire, via une approche informatique

développée spécifiquement⁵⁶. Pour les patients atteints d'une LMA del(7), la discrimination repose sur le niveau d'expression des gènes du chromosome 7. Techniquement, cette séparation était réalisée à l'aide d'un algorithme *k-means* avec $k = 2$, assignant chaque cellule à l'un des deux groupes, *AML-like* ou *non-AML-like*, selon son score d'expression des gènes du chromosome 7 (Figure 6)⁵⁶.

Bien qu'efficaces dans les contextes pour lesquels elles ont été développées, ces approches présentent plusieurs limitations. Leur utilisation reste d'abord restreinte aux patients pour lesquels une altération moléculaire ou chromosomique liée à la LMA est présente^{4,55}. Elles nécessitent également de connaître à l'avance les altérations présentes dans les cellules leucémiques afin de pouvoir les rechercher dans les données de scRNA-seq^{79,91}. De plus, lorsque des technologies de type 10x 3' sont utilisées, la détection des mutations dépend fortement de leur position dans le transcrit, les mutations proches de l'extrémité 3' étant plus facilement capturées que celles situées dans des régions moins couvertes^{79,91}. Cette contrainte s'ajoute à la dépendance vis-à-vis de l'expression du gène cible : une mutation ne peut être détectée que si le transcrit correspondant est effectivement exprimé et capturé dans la cellule analysée⁷⁹. Par ailleurs, ces stratégies reposent souvent sur un seul signal discriminant, comme une mutation donnée ou un score chromosomique, ce qui peut limiter leur robustesse en cas de faible couverture, de *dropout* ou d'hétérogénéité clonale^{56,79}. Enfin, elles nécessitent des méthodes de détection différentes selon le type d'altération : les SNV peuvent être recherchées par appel de variants (*variant calling*), tandis que les insertions, délétions ou pertes chromosomiques requièrent des approches spécifiques⁷⁹. Leur application reste donc fortement dépendante du patient, du type d'altération présente et de sa détectabilité dans les données disponibles⁷⁹.

Ces limitations ont motivé, dans ce travail, le développement d'une approche fondée sur la charge mutationnelle totale normalisée au niveau unicellulaire. Cette approche combine l'ensemble des variants détectés dans chaque cellule afin de proposer un signal de classification, moins dépendant de la détection d'un transcrit muté unique ou d'un score d'expression chromosomique.

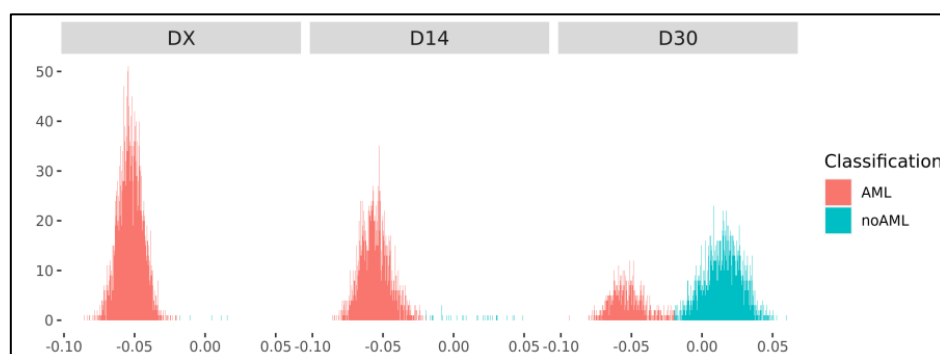


Figure 6. Distribution du score d'expression des gènes du chromosome 7 chez un patient del(7).
Les cellules annotées *AML-like* sont représentées en rouge et les cellules *non-AML-like* en turquoise.
Figure générée à partir des données supplémentaires de Naldini *et al.*⁵⁶

2. Objectif et hypothèse

Face aux limites des stratégies d'annotation des cellules leucémiques dans les données de scRNA-seq discutées précédemment, ce travail vise à développer une approche d'annotation des cellules leucémiques moins dépendante de la détection d'un transcrit muté unique et plus facilement transposable d'un patient à l'autre. Cette approche repose sur l'hypothèse que les cellules leucémiques devraient présenter, dans leur transcriptome, une charge mutationnelle détectable plus élevée que les cellules non leucémiques.

La LMA est généralement considérée comme un cancer à faible charge mutationnelle. Cependant, cette observation repose majoritairement sur des analyses réalisées en *bulk*, qui détectent surtout les variants partagés par une proportion importante de cellules dans un même échantillon. À l'échelle unicellulaire, nous faisons l'hypothèse que les cellules leucémiques peuvent présenter une accumulation plus importante de variants transcrits que les cellules non leucémiques, même si la nature exacte de ces variants peut varier d'une cellule à l'autre.

L'approche développée consiste donc à utiliser la charge mutationnelle détectée dans le transcriptome comme signal de discrimination entre cellules *AML-like*, supposées présenter la charge mutationnelle la plus élevée dans leur transcriptome, et cellules *non-AML-like*. La stratégie adoptée comprend : la quantification de la charge mutationnelle, sa normalisation par le nombre de gènes détectés par cellule, puis une classification non supervisée par *k-means*, intégrant une zone d'incertitude destinée à limiter les cas de classification ambiguë.

Cette classification a ensuite été évaluée par comparaison avec les annotations de référence de Naldini *et al.*⁵⁶, ainsi qu'avec des signatures biologiques indépendantes, notamment par le chevauchement avec une liste de gènes associés à la LMA issue de GeneCards⁹². L'objectif est d'évaluer si cette approche permet d'identifier de manière robuste les cellules *AML-like* au diagnostic et après traitement, chez des patients présentant une mutation de *NPM1* ou une monosomie du chromosome 7.

Enfin, ce travail s'inscrit également dans une volonté de développer une stratégie de suivi clonal (*clonal tracking*) des cellules leucémiques persistantes, notamment des DTPs, dans la MO des patients, à partir des données scRNA-seq disponibles. L'approche développée dans ce mémoire constitue une première étape méthodologique vers cette stratégie de suivi clonal, poursuivie à la suite de ce mémoire au laboratoire d'hématologie.

3. Matériel et méthodes

3.1 Source des données et cadre de l'étude

Les données de scRNA-seq proviennent des travaux de Naldini *et al.*⁵⁶ qui étudient le suivi longitudinal de la réponse à la chimiothérapie dans la LMA. Nous avons utilisé les données issues d'une cohorte de treize patients atteints de LMA, dont six patients ont été retenus pour notre analyse : trois patients étaient porteurs d'une mutation du gène *NPM1*, et trois étaient porteurs d'une délétion du chromosome 7. Les échantillons de MO ont été prélevés à différents moments de l'évolution de la maladie, systématiquement au diagnostic, puis selon les patients : après chimiothérapie d'induction et/ou à la rechute⁵⁶. Dans cette cohorte, les traitements reposaient principalement sur le protocole 7+3, associant l'AraC à la daunorubicine. Selon les patients et l'évolution de la maladie, d'autres schémas thérapeutiques ont été utilisés, notamment FLA-Ida, FLAG-Ida ou encore l'AraC en monothérapie (Figure 7)⁵⁶.

Les aspects éthiques, incluant les approbations par le comité d'éthique et le consentement éclairé des patients, sont détaillés dans l'article de Naldini *et al.*⁵⁶. Les données brutes de scRNA-seq sont accessibles au public via la base de données *Gene Expression Omnibus* (GEO) sous le numéro d'accès GSE185993.

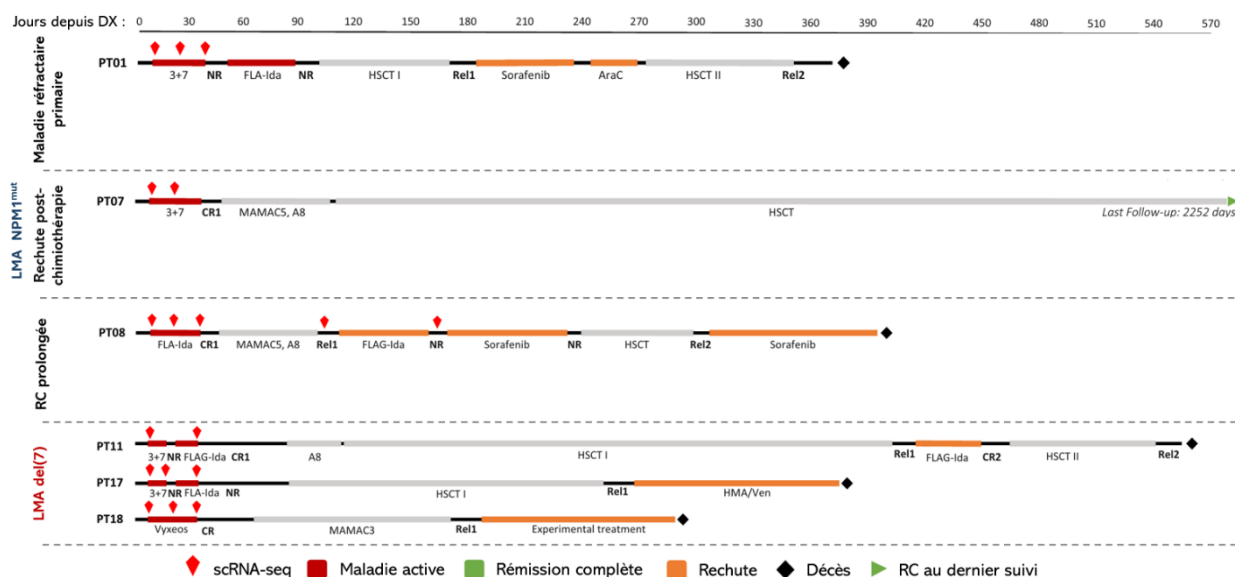


Figure 7. Patients et temps d'échantillonnage retenus pour l'analyse.

Schéma simplifié des trajectoires cliniques des patients issus de la cohorte de Naldini *et al.* retenus dans ce mémoire. Les patients sont regroupés selon leur altération principale, *NPM1*^{mut} ou del(7). Les losanges rouges indiquent les temps d'échantillonnage disponibles pour le scRNA-seq, tandis que les barres colorées représentent l'état clinique au cours du suivi.

Abréviations : DX, diagnostic ; scRNA-seq, séquençage d'ARN à cellule unique ; NR, absence de réponse ; RC/CR, rémission complète ; Rel1/Rel2, première/deuxième rechute ; 7+3, cytarabine pendant 7 jours associée à une anthracycline pendant 3 jours ; FLA-Ida, fludarabine, cytarabine et idarubicine ; FLAG-Ida, fludarabine, cytarabine, G-CSF et idarubicine ; MAMAC5, amsacrine et cytarabine administrées du jour 1 au jour 5 ; MAMAC3, même schéma que MAMAC5 mais administré du jour 1 au jour 3 ; A8, cytarabine administrée du jour 1 au jour 4 ; AraC, cytarabine ; HMA/Ven, agent hypométhylant associé au vénétoclax ; HSCT, transplantation de cellules souches hématopoïétiques ; Vyxeos, formulation liposomale associant daunorubicine et cytarabine.

Figure adaptée de Naldini *et al.*⁵⁶

3.2 Génération des données de scRNA-seq

La préparation des échantillons et la génération des données brutes de scRNA-seq ont été réalisées par l'équipe de Naldini *et al.* Brièvement, les cellules de MO enrichies en blastes leucémiques ont été isolées et préparées pour le séquençage unicellulaire selon une technique *droplet-based* de type 10x Genomics. Les librairies scRNA-seq ont été construites et séquencées avec une profondeur permettant la détection de l'expression des gènes et de certaines altérations génétiques au niveau unicellulaire.

3.3 Analyse mutationnelle

I. Appel de variants et filtrage

À partir des fichiers *Binary Alignment Map* (BAM), qui correspondent à la représentation binaire des *reads* de scRNA-seq alignés sur le génome de référence, un *pipeline* bioinformatique a été développé par notre équipe afin d'identifier les variants présents dans ces données. Ce *pipeline* comprend plusieurs étapes basées sur la littérature⁹³. Le déroulement général de cette analyse mutationnelle est résumé dans la Figure 8.

Une première étape d'appel de variants a été réalisée avec cellSNP-lite (mode 2a)⁹⁴. Cet outil permet d'identifier des positions candidates de variants nucléotidiques simples (SNV) à partir des fichiers BAM et du génome de référence humain (GRCh38)^{95,96}. CellSNP-lite génère deux types de sorties essentielles. D'abord, un fichier *Variant Call Format* (VCF), qui liste toutes les positions où des variants ont été détectés avec leurs caractéristiques : la position chromosomique, l'allèle de référence, l'allèle alternatif et la profondeur de séquençage. La seconde sortie est une matrice associant chaque variant à chaque cellule. Cette association variant-cellule est établie à l'aide des codes-barres cellulaires, qui permettent d'attribuer les *reads* à leur cellule d'origine⁹⁴.

Au départ, cellSNP-lite détecte un grand nombre de positions candidates, dont une proportion importante peut correspondre soit à des faux positifs techniques, soit à des variants non informatifs, par exemple des variants fréquemment présents dans la population générale. Afin d'assurer l'obtention de résultats pertinents, les étapes de filtrage sont critiques^{95,97}. Dans un premier temps, nous avons filtré les variants selon la profondeur allélique (AD). Celle-ci correspond au nombre de *reads* soutenant l'allèle alternatif à une position donnée. Afin de limiter le nombre de faux positifs techniques, seuls les variants présentant une AD supérieure ou égale à 2 dans au moins deux cellules distinctes ont été retenus^{79,95}. Un filtrage complémentaire a ensuite été appliqué sur la fréquence allélique (VAF), définie comme la proportion de lectures alternatives par rapport au nombre total de lectures, alternatives et de référence, à une position donnée. Les variants avec une VAF supérieure à 50 % ont été écartés, une VAF élevée étant le plus souvent compatible avec des variants germinaux ou homozygotes⁹⁸.

Les variants dont la VAF était inférieure ou égale à 1 % ont également été exclus, cette faible VAF étant généralement associée à des artefacts techniques⁹⁷.

Les variants localisés dans les régions répétées du génome ont été identifiés puis exclus à l'aide des annotations *RepeatMasker*, téléchargées depuis l' UCSC Genome Browser, afin d'éviter les faux positifs liés à des erreurs d'alignement¹⁰⁰. De même, les variants localisés dans les régions non traduites (5'UTR et 3'UTR) ont été écartés, la technique d'amplification scRNA-seq 3' présentant un biais de couverture vers l'extrémité 3' des transcrits¹⁰¹. L'exclusion de ces variants a été réalisée à partir de fichiers *Browser Extensible Data* (BED) contenant les coordonnées génomiques des régions UTR, issus de l'UCSC Genome Browser. L'outil *bedtools intersect* a été utilisé pour comparer ces coordonnées aux positions des variants préalablement retenus et éliminer ceux chevauchant les régions non traduites¹⁰². Une fois les variants filtrés, ils ont été soumis à une étape d'annotation fonctionnelle.

II. Annotations fonctionnelles avec VEP

L'objectif de cette étape est d'identifier le type de conséquence que les variants retenus peuvent avoir, les gènes potentiellement affectés ainsi que leur localisation précise dans le génome. Pour cela, deux outils d'annotation ont été utilisés : *Variant Effect Predictor* (VEP)¹⁰³ d'Ensembl pour l'annotation fonctionnelle des variants, et des annotations génomiques issues de l'UCSC *Genome Browser*¹⁰⁴.

L'annotation VEP a été utilisée pour l'annotation fonctionnelle, c'est-à-dire pour déterminer l'effet potentiel du variant sur un gène ou un transcrit. Concrètement, le fichier VCF généré précédemment par cellSNP-lite contenait la liste des variants détectés avec leurs informations principales (chromosome, position, allèles de référence et alternatifs, etc.). Ce fichier a ensuite été utilisé comme entrée pour VEP, qui associe à chaque variant des annotations issues de bases de données de référence du génome. Les annotations produites comprenaient le type de variant : SNV ou INDEL (insertion/délétion), la conséquence fonctionnelle (faux-sens, non-sens, synonyme, etc.), ainsi que le gène affecté¹⁰⁴.

En parallèle, l'annotation basée sur l'UCSC Genome Browser et *bedtools* a permis de localiser les variants dans le génome. Pour cela, un fichier BED issu du portail UCSC contenant les coordonnées de différentes régions génomiques a été utilisé^{102,104}. Les positions des variants, préalablement converties au format BED, ont été croisées avec ces annotations à l'aide de la fonction *bedtools intersect*. Chaque variant a ensuite été annoté selon son biotype afin de déterminer s'il est localisé dans : un exon codant, un exon non-codant, par exemple dans un gène de type ARN long non codant, un intron, un élément transposable ou une région intergénique.

Une stratégie de priorisation hiérarchique a été implémentée afin d'attribuer une annotation unique aux variants chevauchant plusieurs régions. L'ordre de priorité était le suivant : *exon* > région non traduite

(UTR) > *élément transposable* > *intron* > *intergénique*. Cet ordre a été établi sur la base de l'importance décroissante de ces régions dans le contexte de l'analyse, visant à privilégier les régions considérées comme plus informatives.

Les exons ont été priorisés car ils correspondent aux régions codantes, dont les variations sont les plus susceptibles d'avoir un impact fonctionnel sur la protéine produite^{105,106}. Les UTR, bien qu'elles ne soient pas traduites, jouent un rôle important dans la régulation post-transcriptionnelle et peuvent influencer la stabilité, la localisation ou la traduction de l'ARN messager¹⁰⁷. Les éléments transposables ont été priorisés avant les introns dans cette stratégie, car ils peuvent être associés à des effets régulateurs ou à des perturbations de l'expression génique¹⁰⁸. Enfin, les variations introniques ont été retenues avant les variations intergéniques, ces dernières étant considérées comme moins informatives dans le cadre de cette analyse¹⁰⁹.

À l'issue de ce *pipeline*, deux matrices ont été générées : une matrice d'expression des gènes par cellule et une matrice de variants par cellule. Ces résultats ont ensuite été utilisés pour générer les fichiers d'entrée nécessaires à l'analyse sous Seurat, à savoir les matrices de codes-barres et de *features*.

Pipeline d'analyse mutationnelle à partir des données scRNA-seq

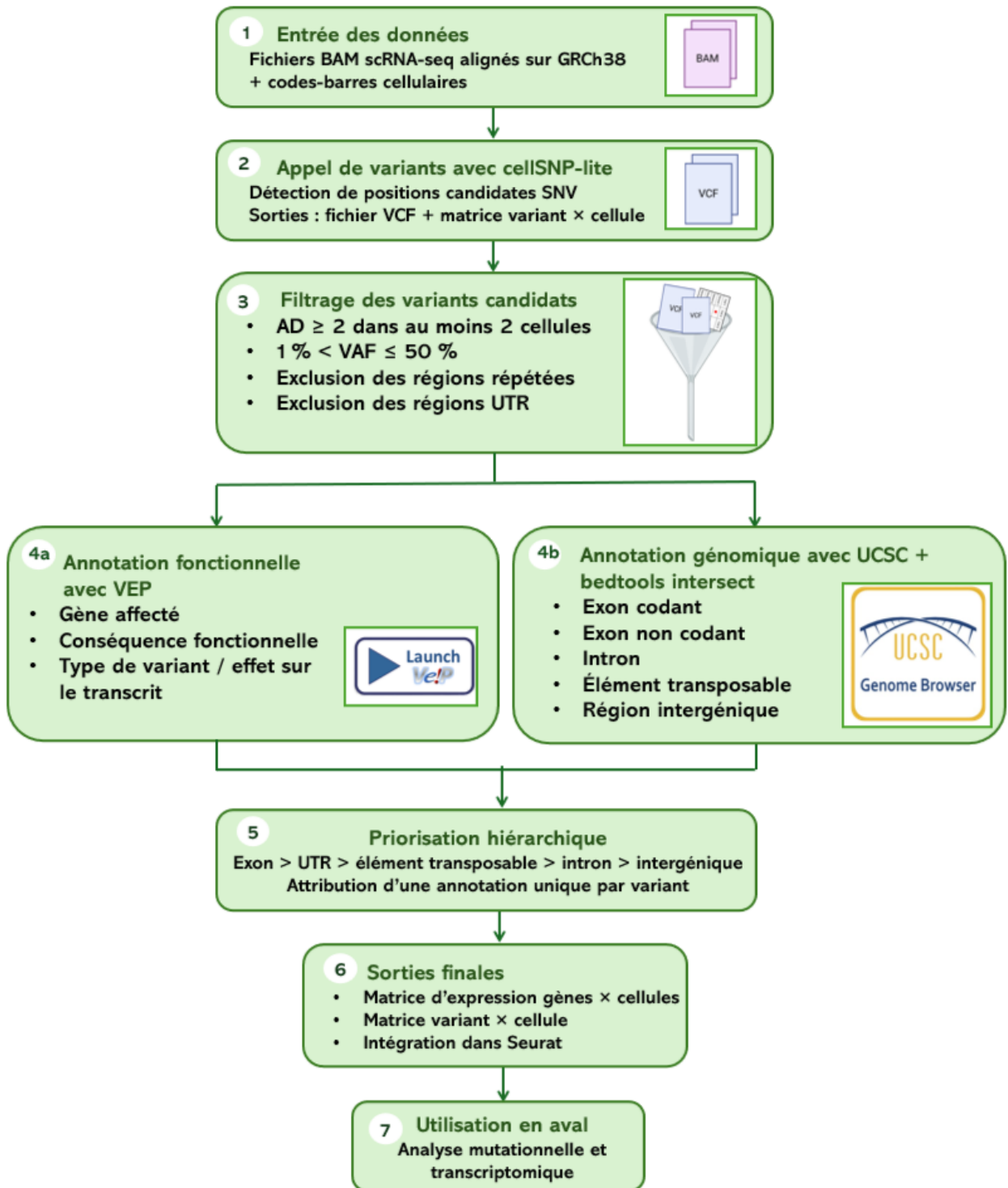


Figure 8. Pipeline d'analyse mutationnelle à partir des données scRNA-seq

Schéma résumant les principales étapes utilisées pour détecter, filtrer et annoter les variants à partir des fichiers BAM scRNA-seq alignés sur GRCh38. Les variants candidats sont identifiés avec cellSNP-lite, puis filtrés selon la profondeur allélique, la VAF, les régions répétées et les régions UTR. Les variants retenus sont ensuite annotés avec VEP et les annotations UCSC/bedtools, avant la génération des matrices utilisées pour les analyses mutationnelles et transcriptomiques.

Schéma original réalisé dans le cadre de ce mémoire.

3.4 Analyse transcriptomique

I. Environnement informatique

L'analyse transcriptomique a été réalisée sur le cluster du GIGA. Les calculs ont été exécutés via le gestionnaire de tâches SLURM, avec une allocation de 12 cœurs de CPU et 48 Go (12 × 4 000 Mo) de mémoire RAM. Les scripts ont été écrits et exécutés sous la forme d'un *notebook* JupyterLab (Jupyter Server v2.17.0) lancé depuis un nœud disponible via un environnement Conda (v25.7.0 ; Python 3.10). Le noyau R utilisé était l'IRkernel (v1.3.2), sous R (v4.4.2).

Nous avons utilisé principalement le *package* Seurat (v5.3.0) pour le traitement et l'analyse des données scRNA-seq. La création de l'objet Seurat, les contrôles de qualité, la normalisation, la réduction dimensionnelle (PCA, UMAP) et le *clustering* non supervisé sont expliqués plus en détail dans les points suivants. La gestion des matrices a été réalisée avec le *package* Matrix (v1.7.4). La gestion des données tabulaires a été assurée via l'écosystème tidyverse, comprenant dplyr (v1.1.4), tibble (v3.2.1), stringr (v1.5.0) et tidyr (v1.3.1). La visualisation graphique a été réalisée avec ggplot2 (v3.5.1). La reproductibilité des analyses a été renforcée par l'utilisation d'une graine aléatoire commune (set.seed(1234)).

II. Création de l'objet Seurat et contrôle qualité

Les données scRNA-seq ont été importées à l'aide de la fonction Read10X(), compatible avec les sorties 10x Genomics (codes-barres cellulaires, gènes et matrice d'expression). Un objet Seurat a été créé à partir de la matrice d'expression génique à l'aide de la fonction CreateSeuratObject(). En parallèle, les données relatives aux variants ont été importées de façon similaire. La matrice résultante a ensuite été intégrée à l'objet Seurat dans un *assay* distinct nommé « Mutation », permettant de réaliser des analyses transcriptomiques et mutationnelles au sein d'un même objet.

Les contrôles de qualité ont été effectués sur la base des critères utilisés par Naldini *et al.* Les filtres ont été appliqués selon trois paramètres : le nombre de gènes détectés (nFeature_RNA), le nombre total de transcrits détectés (nCount_RNA), et le pourcentage de transcrits mitochondriaux (percent.mt). Ces paramètres ont été visualisés à l'aide de *violin plot* avant et après filtrage. Les cellules présentant entre 100 et 7 000 gènes détectés, moins de 10 000 transcrits totaux, et un pourcentage de transcrits mitochondriaux inférieur à 20 % ont été conservées.

III. Normalisation, réduction dimensionnelle et clustering

Après l'élimination des cellules ne respectant pas les critères de qualité, une étape de normalisation a été réalisée. Cette étape est essentielle car elle permet de rendre comparables les cellules présentant des profondeurs de séquençage différentes en réduisant le bruit technique lié au nombre total de transcrits détectés. En effet, la détection des transcrits varie d'une cellule à l'autre, ce qui peut

artificiellement donner l'impression qu'un gène est davantage exprimé dans une cellule plus profondément séquencée.

La normalisation a été effectuée à l'aide de la fonction `NormalizeData()` de Seurat, avec la méthode `LogNormalize`. Concrètement, les comptages bruts d'un gène sont divisés par le total de comptages de la cellule correspondante, puis multipliés par un facteur d'échelle (fixé par défaut à 10 000) afin d'éviter de manipuler des valeurs trop faibles. Une transformation logarithmique est ensuite appliquée afin de réduire l'asymétrie de distribution des niveaux d'expression et de limiter l'influence des valeurs extrêmes. Cette étape permet de passer de comptages bruts à des valeurs d'expression comparables entre cellules.

Avant de passer aux étapes de réduction dimensionnelle et de *clustering*, deux fonctions de Seurat ont été utilisées : `FindVariableFeatures()` et `ScaleData()`. La première a permis d'identifier les gènes les plus variables, c'est-à-dire ceux dont l'expression transcriptomique varie le plus d'une cellule à l'autre. Cette variabilité est particulièrement informative pour la suite de l'analyse, car elle permet de distinguer les différentes populations cellulaires. Dans cette analyse, nous avons adopté le paramètre par défaut de Seurat, fixé à 2 000 gènes variables.

Ensuite, `ScaleData()` a été appliquée pour ramener l'expression de tous les gènes sur une même échelle. Pour chaque gène, la moyenne d'expression est soustraite, puis le résultat est divisé par l'écart-type. Une valeur positive indique que la cellule exprime ce gène davantage que la moyenne, une valeur négative indique une expression inférieure à la moyenne, et une valeur nulle correspond à une expression égale à la moyenne.

Après le *scaling*, nous avons réalisé une analyse en composantes principales (PCA) afin de réduire la dimensionnalité du jeu de données tout en conservant les variations transcriptomiques les plus informatives. Le nombre de composantes principales (PC) à retenir a été déterminé à l'aide d'un *elbow plot*. Les dix premières PC ont été retenues pour construire un graphe de plus proches voisins à l'aide de la fonction `FindNeighbors()`. Cette étape permet d'identifier les cellules présentant des profils transcriptomiques similaires.

Le *clustering* non supervisé a été appliqué avec la fonction `FindClusters()`, avec une résolution de 0.8, conformément aux recommandations de la documentation Seurat¹¹⁰, afin de générer des clusters biologiquement interprétables regroupant des cellules aux similarités transcriptionnelles. L'identification préalable des gènes les plus variables à l'aide de la fonction `FindVariableFeatures()` est une étape importante, car elle permet de sélectionner les gènes les plus informatifs pour la PCA et, par conséquent, pour le *clustering*. Enfin, afin de visualiser les cellules dans un espace bidimensionnel et

de représenter la distribution des clusters, nous avons appliqué la fonction RunUMAP(), produisant une projection *Uniform Manifold Approximation and Projection* (UMAP).

IV. Classification Naldini et al. *NPM1*^{mut} / del(7)

Comme indiqué précédemment, l’objectif de ce mémoire n’est pas de reproduire à l’identique l’analyse de Naldini et al., mais d’utiliser leur classification comme référence externe afin d’évaluer la cohérence de notre propre approche. Pour les patients porteurs d’une mutation du gène *NPM1*, nous avons utilisé directement les annotations fournies dans l’étude de Naldini et al.⁵⁶. Ces annotations correspondent à une duplication TCTG. Plusieurs motifs contenant cette duplication ont été recherchés dans les *reads* afin d’identifier les cellules exprimant le transcrit muté (Tableau 1).

Motifs recherchés dans les <i>reads</i>				
Duplication TCTG	^TCTCTGTCTGGC GATCTCTGTCTGG\$	GATCTCTGTCTG\$ GATCTCTGTCT\$	GATCTCTGTC\$ ^CTCTGTCTGGCAG	^TCTGTCTGGCAG ^CTGTCTGGCAG

Tableau 1. Motifs de séquence utilisés pour la détection de la mutation *NPM1* dans les *reads* scRNA-seq.

Pour les patients del(7), ces annotations n’étant pas disponibles, nous avons reproduit leur approche afin d’isoler les codes-barres cellulaires présentant un profil compatible avec une perte du chromosome 7.

Dans un premier temps, nous sommes partis du fichier d’annotation génomique *Homo_sapiens.GRCh38.113.gtf*, à partir duquel nous avons construit une liste de gènes localisés sur le chromosome 7. Dans un second temps, cette liste a été appliquée aux objets Seurat des différents points temporels pour tous les patients del(7). Nous avons utilisé la fonction AddModuleScore() afin de calculer, pour chaque cellule, un score d’expression des gènes du chromosome 7. Enfin, un algorithme de *k-means* a été appliqué pour séparer les cellules en deux populations selon le niveau d’expression de cette signature. Les cellules présentant un score bas, c’est-à-dire négatif, compatible avec une diminution de l’expression des gènes du chromosome 7, ont été considérées comme *AML-like*, tandis que celles présentant un score plus élevé, proche de zéro ou positif, ont été considérées comme *non-AML-like* (Figure 6).

Afin de visualiser la distribution des cellules, les annotations *AML-like* et *non-AML-like* ont été projetées sur un graphe UMAP. Nous avons ensuite calculé, pour chaque cluster, la proportion de cellules appartenant à chacune de ces deux catégories, afin d’effectuer une annotation au niveau du cluster plutôt qu’au niveau unicellulaire, tout en suivant la logique de l’étude. Un cluster était annoté comme *AML-like* ou *non-AML-like* selon la majorité des cellules qu’il contenait, tandis qu’une répartition équilibrée 50/50 conduisait à une annotation « mixte ». Ce graphe UMAP a ensuite été comparé à celui obtenu à partir des annotations issues de notre propre stratégie.

V. Charge mutationnelle normalisée

Nous avons commencé l'analyse mutationnelle par la binarisation de la matrice de mutations. La présence d'une mutation dans une cellule était codée par « 1 » et son absence par « 0 ». Cette étape a permis de quantifier, pour chaque cellule, le nombre total de mutations en additionnant l'ensemble des mutations détectées.

Afin de caractériser les profils mutationnels, nous avons comparé les mutations détectées dans chaque cellule à des bases de données externes. Les annotations des variants ont été harmonisées au préalable afin de correspondre à la nomenclature utilisée dans ces bases de données. Les mutations chevauchant la *Database of Single Nucleotide Polymorphisms* (dbSNP, v157)⁹⁹ ont été considérées comme compatibles avec une origine germinale, tandis que celles chevauchant la base de données *Catalogue of Somatic Mutations in Cancer* (COSMIC) ont été considérées comme probablement somatiques. Cette étape ne constitue pas une validation absolue, mais fournit une approximation des mutations de type *germinal-like* et *somatique-like* sur la base de leurs correspondances connues avec ces bases de données.

Étant donné que la quantité de mutations détectées peut être influencée par la couverture transcriptomique variable d'une cellule à l'autre, une étape de normalisation a été nécessaire. Pour cela, nous avons divisé le nombre total de mutations somatiques et germinales détectées dans chaque cellule par le nombre de gènes détectés dans cette cellule, estimé à l'aide de `nFeature_RNA`. Cette normalisation a été appliquée séparément aux charges mutationnelles germinale et somatique. Elle avait pour but de réduire le biais introduit par l'inégalité de détection des gènes entre cellules.

Dans un premier temps, notre analyse était principalement centrée sur la charge mutationnelle somatique normalisée, définie à partir du chevauchement avec COSMIC. Cependant, étant donné que l'utilisation de bases de données externes introduit une dépendance à ces ressources, nous avons exploré l'hypothèse selon laquelle la charge mutationnelle totale normalisée pourrait fournir un signal équivalent. Afin d'évaluer cette hypothèse, la charge mutationnelle somatique normalisée et la charge mutationnelle totale normalisée ont été comparées à l'aide de corrélations de Pearson et de Spearman. Un test de significativité a également été réalisé pour la corrélation de Pearson. Nous avons ainsi obtenu une corrélation positive entre ces deux charges, indiquant que les cellules présentant une charge mutationnelle somatique normalisée élevée tendent également à présenter une charge mutationnelle totale normalisée élevée. Sur cette base, nous avons choisi d'utiliser la charge totale normalisée pour la suite des analyses, celle-ci représentant une alternative indépendante des bases de données externes tout en restant représentative de la charge mutationnelle.

VI. K-means

L'étape suivante consistait à visualiser la distribution des charges mutationnelles totales normalisées à travers les différents clusters à l'aide de *violin plot*, puis à calculer la médiane de ces charges pour chaque cluster. Ces médianes ont ensuite servi d'entrée à une approche de *k-means* avec $k = 2$, afin de séparer les clusters en deux groupes correspondant à une charge mutationnelle faible ou élevée. Le déroulement global de cette stratégie de classification est résumé dans la Figure 11.

Le choix de cette approche reposait sur le fait que le *k-means* constitue une méthode de partitionnement non supervisé simple, largement utilisée et facilement interprétable, particulièrement adaptée lorsqu'un nombre fixe de groupes est prédéfini¹¹¹. Dans notre cas, le choix de $k = 2$ était directement lié à la nature de notre problématique, qui vise à distinguer deux populations cellulaires présentant des profils compatibles soit avec une origine leucémique, soit avec une origine non leucémique¹¹². Le principe général de l'algorithme *k-means*, reposant sur l'initialisation de centres, l'assignation des observations au centre le plus proche puis la mise à jour itérative des centres jusqu'à convergence, est illustré dans la Figure 9.

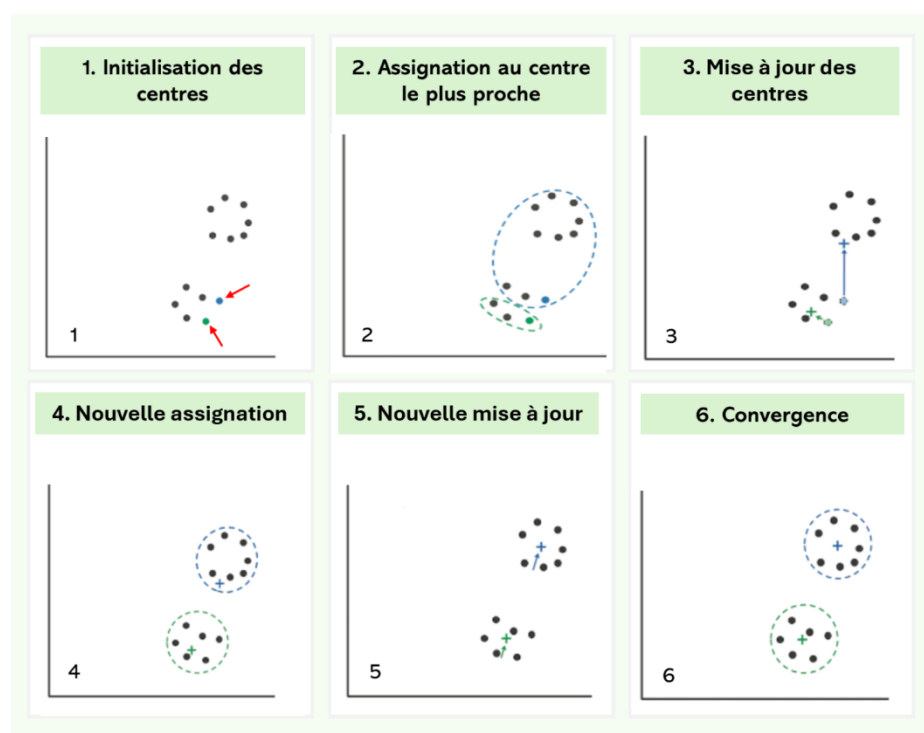


Figure 9. Principe général de l'algorithme k-means.

Schéma illustrant le fonctionnement général de l'algorithme *k-means*, utilisé ici avec $k = 2$. Dans un premier temps, deux centres sont initialisés de manière aléatoire. Chaque point est ensuite assigné au centre le plus proche, ce qui permet de former deux groupes provisoires. Les centres sont alors recalculés à partir de la position moyenne des points appartenant à chaque groupe. Ces étapes d'assignation et de mise à jour sont répétées de manière itérative jusqu'à convergence, c'est-à-dire jusqu'à stabilisation de la position des centres et de l'appartenance des observations aux groupes. Dans ce mémoire, ce principe a été appliqué aux médianes de charge mutationnelle totale normalisée calculées pour chaque cluster Seurat, afin de séparer les clusters en deux groupes correspondant à une charge mutationnelle faible ou élevée.

Schéma original réalisé dans le cadre de ce mémoire.

Lorsque le jeu de données ne contenait que deux clusters, le *k-means* n'était pas appliqué. Les médianes des deux clusters étaient alors utilisées directement, et le seuil de classification était défini comme le point médian entre elles. Ce seuil était ensuite appliqué à la charge mutationnelle totale normalisée au niveau unicellulaire. Les cellules situées au-dessus du seuil ont été considérées comme *AML-like*, tandis que celles situées en dessous ont été attribuées à l'étiquette *non-AML-like*.

Lorsque le jeu de données contenait au moins trois clusters, un *k-means* à deux centres était appliqué aux médianes. L'algorithme repose sur une initialisation aléatoire des centres, puis attribue chaque médiane au centre le plus proche avant de mettre à jour la position des centres jusqu'à convergence. Dans notre implémentation, 50 initialisations étaient réalisées afin de limiter l'effet du hasard lié à l'initialisation et de retenir la meilleure répartition. Une fois les deux centres identifiés, le seuil de classification était défini comme le point médian entre eux. Ce seuil était ensuite utilisé pour classer les cellules au niveau unicellulaire. En revanche, lorsqu'un seul cluster était présent, aucune classification n'était possible.

Une étape importante consistait ensuite à introduire une marge d'incertitude de part et d'autre du seuil central, puis à sélectionner automatiquement la marge optimale. L'idée était que les cellules proches du seuil sont plus difficiles à classer de manière fiable. Plutôt que de les assigner arbitrairement à une catégorie, le script introduisait une troisième annotation, *uncertain*¹¹³.

Dans un premier temps, le script testait une série de marges comprises entre 1 % et 30 % autour du seuil central. Pour chaque marge, deux seuils étaient définis :

- Un seuil inférieur = seuil $\times$ (1 – marge)
- Un seuil supérieur = seuil $\times$ (1 + marge)

Les cellules situées en dessous du seuil inférieur étaient classées comme *non-AML-like*, celles situées au-dessus du seuil supérieur étaient classées comme *AML-like*, et celles comprises entre les deux étaient classées comme *uncertain*. Cette approche permettait de définir une zone grise autour du seuil, dans laquelle la classification était considérée comme moins fiable¹¹³. Le principe de définition du seuil central et de la zone d'incertitude utilisée pour la classification est illustré dans la Figure 10.

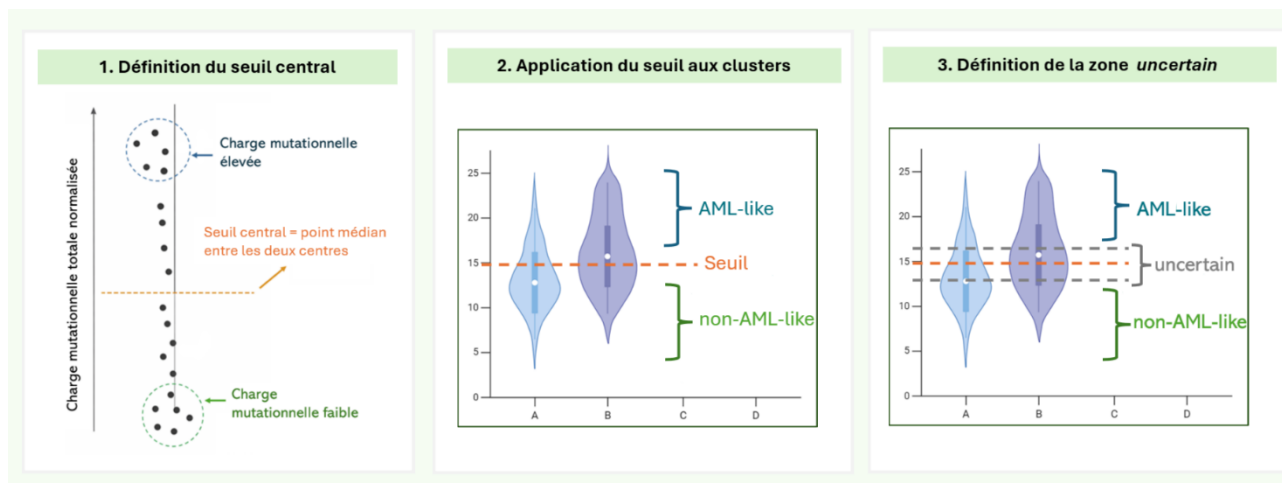


Figure 10. Définition du seuil central et de la zone d'incertitude pour la classification AML-like/non-AML-like.

Schéma illustrant la définition du seuil central à partir des deux centres de charge mutationnelle faible et élevée, puis son application aux distributions de charge mutationnelle totale normalisée par cluster. Une marge d'incertitude est ajoutée autour du seuil afin d'identifier les cellules proches de la zone de séparation, annotées comme *uncertain*.

Schéma original réalisé dans le cadre de ce mémoire.

Pour chaque marge testée, trois indicateurs étaient calculés : le pourcentage global de cellules *uncertain*, le pourcentage maximal de cellules *uncertain* observé dans un cluster, et un score de séparation entre les groupes *AML-like* et *non-AML-like*. Ce dernier correspondait à la différence entre les médianes des charges mutationnelles totales normalisées des cellules classées comme *AML-like* et celles classées comme *non-AML-like*. Une valeur élevée indiquait une meilleure séparation entre les deux groupes.

La sélection de la marge optimale reposait d'abord sur un filtre appliqué au pourcentage maximal de cellules *uncertain* observé dans chaque cluster, qui ne devait pas dépasser 20 %. Ce pourcentage d'incertitude a été retenu après avoir testé plusieurs marges comprises entre 1 % et 30 %. Pour les marges générant moins de 20 % de cellules incertaines, la quasi-totalité des cellules était assignée à l'une des deux catégories, ce qui réduisait la prise en compte des profils intermédiaires attendus dans les données leucémiques. Pour les marges générant plus de 20 % de cellules incertaines, la proportion de cellules non classifiées devenait trop importante pour permettre une comparaison robuste entre les deux groupes. Le seuil de 20 % a ainsi été identifié comme la valeur à partir de laquelle la classification produisait des groupes biologiquement interprétables tout en maintenant un niveau d'incertitude raisonnable.

Parmi les marges respectant ce critère, celles présentant le score de séparation le plus élevé étaient retenues. Une fois la marge optimale sélectionnée, les seuils inférieur et supérieur étaient recalculés, les cellules étaient classées en *AML-like*, *non-AML-like* ou *uncertain*, puis ces étiquettes étaient stockées dans l'objet Seurat.

En suivant la même logique que celle appliquée précédemment aux graphes UMAP générés à partir des annotations de Naldini *et al.*, nous avons attribué à chaque cluster l'annotation correspondant à la majorité des cellules *AML-like* ou *non-AML-like* qu'il contenait. Une catégorie « mixte » a été définie en cas de répartition équilibrée (50/50), puis ces annotations ont été projetées sur un graphe UMAP.

VII. Expression IFN- γ et gènes LMA

Un ensemble de gènes développé au sein du laboratoire de l'équipe de Grégory Ehx (PI, *principal investigator*) a été utilisé dans cette analyse. Cet ensemble de gènes, élaboré par Grégory Ehx et Bianca E. Silva (doctorante), correspond à un programme transcriptionnel associé à la réponse à l'interféron gamma (IFN- γ) dans un contexte de persistance après traitement.

L'ensemble de gènes IFN_DTP (signature interféron associée aux *Drug-Tolerant Persisters*) a été défini à partir de données de séquençage d'ARN obtenues sur des lignées cellulaires leucémiques traitées par AraC en présence ou en l'absence d'IFN- γ . Il correspond à l'intersection entre les gènes surexprimés dans les cellules DTPs, c'est-à-dire les cellules survivantes au traitement, et les gènes induits par l'IFN- γ . Cette signature n'est pas utilisée ici pour identifier, au diagnostic, des DTPs déjà pleinement constituées, mais plutôt pour rechercher de rares sous-populations cellulaires présentant dès ce stade un état compatible avec une prédisposition à la persistance et partageant des similitudes avec les DTPs détectées après traitement.

Afin de quantifier l'activité de ce programme au niveau unicellulaire, un score d'expression a été calculé à l'aide de la fonction `AddModuleScore()` du package Seurat. Cette approche calcule, pour chaque cellule, un score basé sur l'expression moyenne des gènes composant l'ensemble de gènes, puis ajuste cette valeur en soustrayant l'expression moyenne de gènes dits de contrôle. Ces gènes de contrôle ne sont pas sélectionnés aléatoirement, mais choisis de manière à présenter un niveau d'expression moyen similaire à celui des gènes d'intérêt, ce qui permet de limiter les biais liés aux différences d'expression génique entre cellules.

Ensuite, nous avons analysé les différences d'expression génique entre les catégories *AML-like* et *non-AML-like* à l'aide de la fonction `FindMarkers()` du package Seurat. La comparaison a été effectuée par un test de Wilcoxon. Nous avons appliqué un seuil minimal d'expression dans 10 % des cellules (`min.pct = 0.1`) afin d'exclure les gènes détectés dans une faible fraction des cellules, qui peut refléter une expression sporadique ou des artefacts. En complément, un seuil minimal de *log-fold change* a été fixé à 0.25 (`logfc.threshold = 0.25`) afin de ne retenir que les gènes présentant une différence d'expression suffisamment marquée entre les deux groupes. Cette analyse a permis d'identifier des gènes préférentiellement exprimés dans les cellules *AML-like* ou *non-AML-like*.

En parallèle, un ensemble de 600 gènes associés à la LMA a été utilisé afin d'estimer, au niveau unicellulaire, dans quelle mesure chaque cellule exprimait une signature fréquemment retrouvée dans les cellules de LMA. Ces gènes ont été extraits de la base de données GeneCards⁹². Un score a été calculé pour chaque cellule, selon le même principe que celui décrit pour la signature IFN_DTP. Le score obtenu a ensuite été projeté sur le graphe UMAP à l'aide de FeaturePlot(), permettant de visualiser la distribution de cette signature LMA au sein des différentes sous-populations cellulaires.

Pipeline de classification AML-like / non-AML-like basé sur la charge mutationnelle totale normalisée

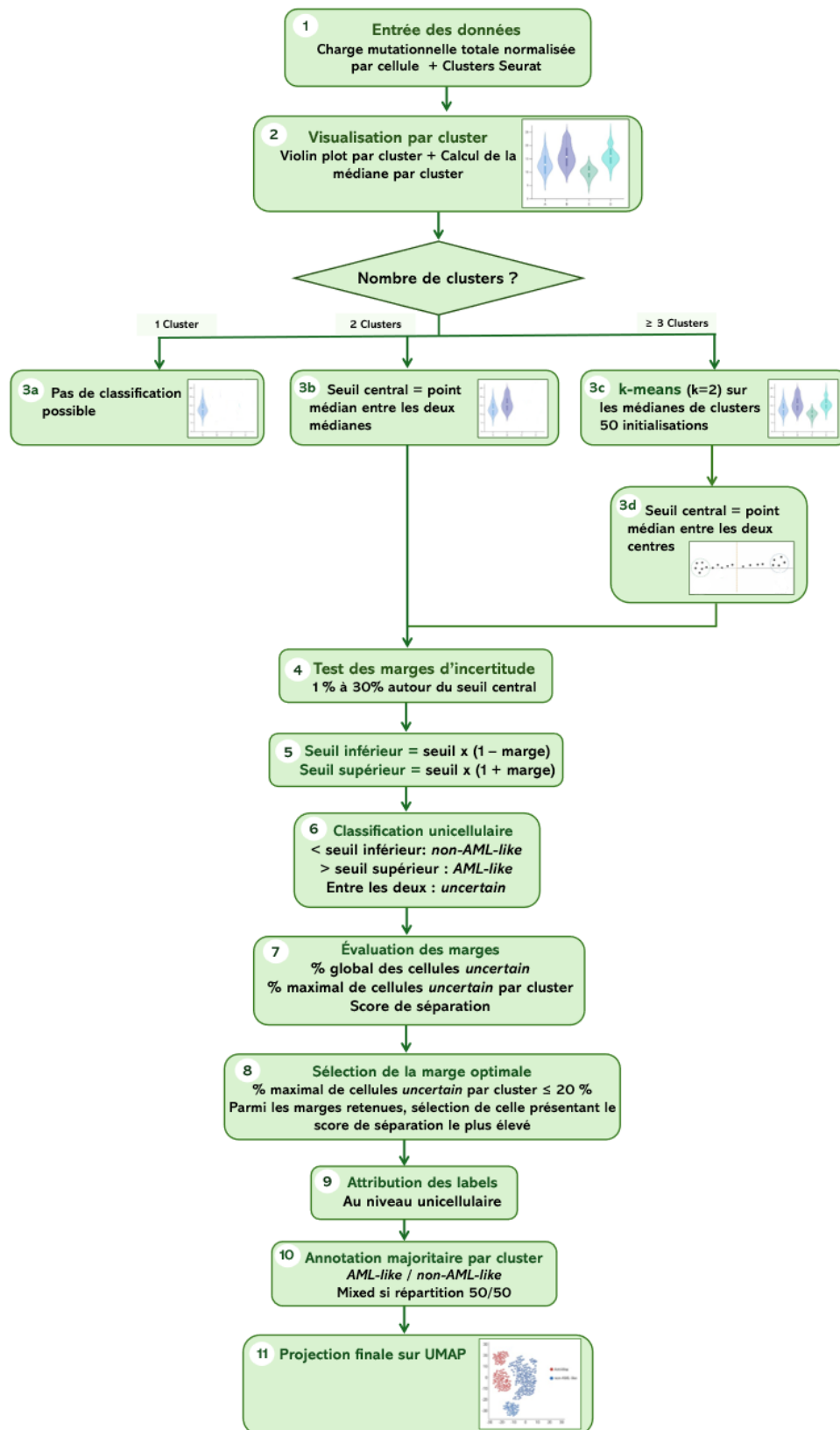


Figure 11. Pipeline de classification AML-like / non-AML-like basé sur la charge mutationnelle totale normalisée.

Schéma simplifié des étapes de classification utilisées dans ce mémoire. Les charges mutationnelles normalisées sont résumées par cluster, puis utilisées pour définir un seuil de classification par *k-means* ou par point médian selon le nombre de clusters. Une marge d'incertitude est ensuite appliquée afin de distinguer les cellules AML-like, non-AML-like et *uncertain*, avant annotation majoritaire par cluster et projection finale sur UMAP.

Schéma original réalisé dans le cadre de ce mémoire.

4. Résultats

Afin d'évaluer si la charge mutationnelle unicellulaire permet d'identifier des populations compatibles avec une origine leucémique, les analyses ont été menées en plusieurs étapes sur six patients issus de l'étude de Naldini *et al*⁵⁶. Ces patients ont été sélectionnés afin de couvrir deux altérations d'intérêt, *NPM1*^{mut} et del(7), ainsi que différents contextes cliniques de réponse au traitement.

Les trois patients *NPM1*^{mut} analysés correspondent à trois trajectoires distinctes : PT01 présente une maladie réfractaire, PT07 présente une réponse complète prolongée, et PT08 présente une rechute précoce après RC. Les trois patients del(7) comprennent deux cas de maladie réfractaire, PT11 et PT17, ainsi qu'un cas de rechute précoce, PT18. Cette diversité clinique permet d'évaluer l'approche fondée sur la charge mutationnelle dans différents contextes de réponse au traitement (Tableau 2).

4.1 Normalisation et corrélation entre les charges mutationnelles

Avant d'évaluer la relation entre la charge mutationnelle totale normalisée et la charge mutationnelle somatique normalisée, l'impact de la complexité transcriptomique sur le nombre brut de mutations détectées a d'abord été analysé. Le nombre brut de mutations détectées par cellule a été comparé au nombre de gènes détectés par cellule, estimé par nFeature_RNA. Cette analyse montre une corrélation positive et statistiquement significative pour l'ensemble des patients (Figure 12). Les coefficients de corrélation de Pearson sont élevés, avec des valeurs comprises entre 0.818 et 0.968 ($p < 0.001$) pour l'ensemble des patients.

Au vu de la corrélation entre la charge mutationnelle et le nombre de gènes détectés par cellule, un score de charge mutationnelle normalisé a été calculé en divisant, pour chaque cellule, le nombre de variants détectés par le nombre de gènes exprimés. Ensuite, nous avons évalué si la charge mutationnelle totale normalisée ainsi calculée reflète ou non la charge mutationnelle somatique normalisée. Après annotation des variants avec la base de données COSMIC, permettant d'identifier les variants répertoriés comme somatiques dans des cancers, nous avons observé des corrélations positives et significatives chez l'ensemble des patients analysés entre la charge mutationnelle totale et la charge mutationnelle somatique (Figure 13A). Les corrélations les plus fortes sont observées chez les patients PT01, PT08, PT11, PT17 et PT18, avec des coefficients de corrélation de Pearson compris entre 0.669 et 0.888 ($p < 0.001$).

Le patient PT07 présente une corrélation plus faible lorsque l'ensemble des cellules est inclus dans l'analyse (Pearson $r = 0.223$; $p < 0.001$). L'analyse visuelle du nuage de points révèle une sous-population cellulaire caractérisée par une charge mutationnelle totale normalisée élevée, mais une charge mutationnelle somatique normalisée relativement faible. Après exclusion de ces cellules atypiques, définies par une charge mutationnelle totale normalisée supérieure à 2, la corrélation

augmente de manière notable (Pearson $r = 0.634$; $p < 0.001$). L'analyse de la distribution de ces cellules au sein des clusters Seurat montre que la majorité d'entre elles appartient au cluster 5, représentant 66 % des cellules exclues (Figure 13B).

Sur base de ces corrélations positives, la charge mutationnelle totale normalisée a été retenue pour la suite des analyses. Ce choix repose sur le fait qu'elle intègre un plus grand nombre de variants détectés que la charge somatique, tout en présentant une bonne corrélation avec celle-ci chez la majorité des patients. Elle permet ainsi de réduire la dépendance vis-à-vis d'un nombre limité de variants annotés comme somatiques dans les bases de données de référence.

Patient	Altération	Contexte clinique	Première ligne de traitement	Greffe reçue	Points temporels analysés
PT01	<i>NPM1</i> ^{mut}	Maladie réfractaire	7+3	Oui	DX, D14, D30
PT07	<i>NPM1</i> ^{mut}	Réponse complète prolongée	7+3	Oui	DX, D14
PT08	<i>NPM1</i> ^{mut}	Rechute précoce après RC	FLA-Ida	Oui	DX, D14, D30
PT11	<i>del(7)</i>	Maladie réfractaire	7+3	Oui	DX, D30
PT17	<i>del(7)</i>	Maladie réfractaire	7+3	Oui	DX, D14, D30
PT18	<i>del(7)</i>	Rechute précoce après RC	Vyxeos	Non	DX, D14

Tableau 2. Contexte clinique, première ligne de traitement, greffe reçue et points temporels analysés chez les six patients inclus dans ce mémoire, d'après Naldini et al⁵⁶.

Les patients ont été sélectionnés afin de représenter deux altérations d'intérêt, *NPM1*^{mut} et *del(7)*, ainsi que différents contextes cliniques de réponse au traitement.

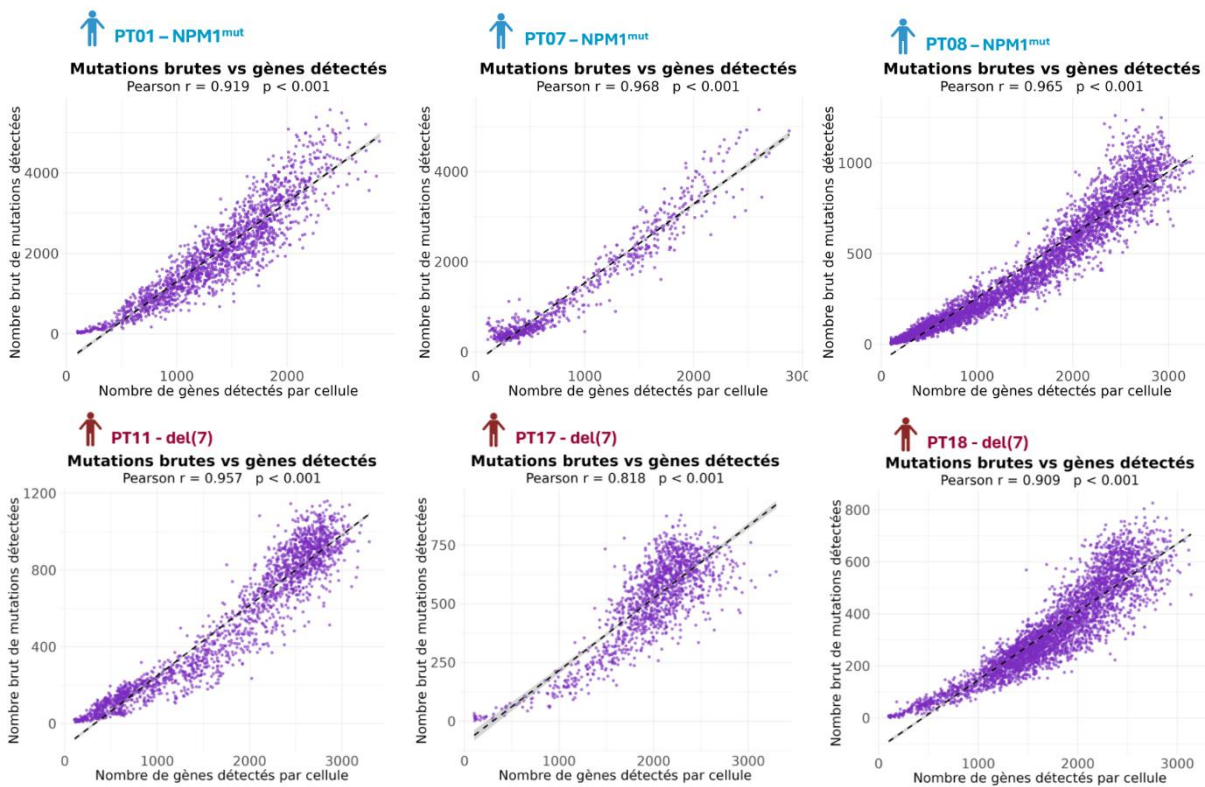


Figure 12. Relation entre le nombre brut de mutations détectées et le nombre de gènes détectés par cellule chez les six patients analysés au diagnostic.

Pour chaque patient le nombre brut de mutations détectées par cellule a été comparé au nombre de gènes détectés par cellule, estimé par nFeature_RNA. Une corrélation positive est observée chez l'ensemble des patients.

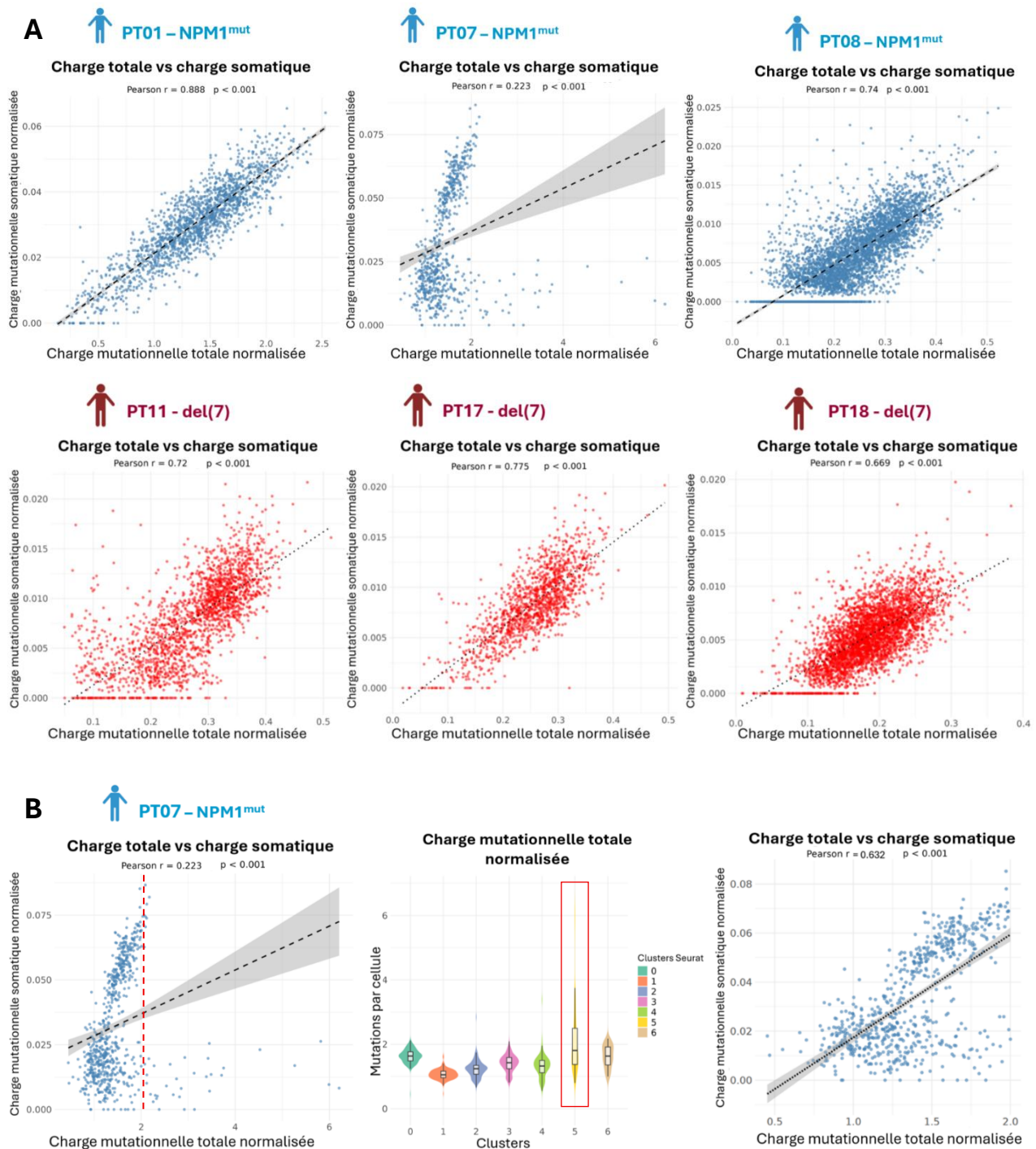


Figure 13. Corrélation entre la charge mutationnelle totale normalisée et la charge mutationnelle somatique normalisée chez les six patients analysés au diagnostic.

A. Nuages de points représentant la relation entre la charge mutationnelle totale normalisée et la charge mutationnelle somatique normalisée chez trois patients *NPM1*^{mut} (PT01, PT07 et PT08) et trois patients del(7) (PT11, PT17, PT18). Chaque point correspond à une cellule. La ligne pointillée représente la régression linéaire, et le coefficient de corrélation de Pearson est indiqué pour chaque patient.

B. Analyse détaillée du patient PT07. Le nuage de points de gauche montre la corrélation lorsque toutes les cellules sont incluses. La ligne rouge pointillée indique le seuil utilisé pour identifier les cellules présentant une charge mutationnelle totale normalisée élevée (> 2). Le *violin plot* central représente la distribution de la charge mutationnelle totale normalisée entre les clusters Seurat, montrant que la majorité des cellules à forte charge mutationnelle totale est assignée au cluster 5. Le nuage de points de droite montre la corrélation après exclusion des cellules présentant une charge mutationnelle totale normalisée > 2 .

4.2 Analyse détaillée d'un patient *NPM1^{mut}*

Afin d'illustrer l'application de cette stratégie chez les patients *NPM1^{mut}*, le patient PT01 au diagnostic a été sélectionné comme exemple représentatif. Tout d'abord, les cellules ont été visualisées selon le *clustering* Seurat afin d'observer la structure transcriptomique générale de l'échantillon (Figure 14A). La distribution de la charge mutationnelle totale normalisée au sein des différents clusters a ensuite été analysée (Figure 14B).

La méthode *k-means* a ensuite été appliquée aux médianes de la charge mutationnelle totale normalisée calculées pour chaque cluster Seurat (voir Figure 9 pour le principe général de l'algorithme). Cette approche non supervisée a permis de séparer les clusters en deux groupes, correspondant à une charge mutationnelle faible ou élevée, et de définir un seuil central de classification à 0.975 (Figure 14C). Afin de limiter les classifications trop proches de ce seuil, plusieurs marges d'incertitude comprises entre 1 % et 30 % ont été testées autour du seuil central (Tableau 3).

Pour chaque marge, trois indicateurs ont été calculés : le pourcentage global de cellules classées comme *uncertain*, le pourcentage maximal de cellules *uncertain* observé dans un cluster, ainsi qu'un score de séparation entre les cellules *AML-like* et *non-AML-like*. Le pourcentage global de cellules *uncertain* correspond à la proportion de cellules non attribuées à l'une des deux catégories principales dans l'ensemble de l'échantillon. Le score de séparation correspond à la différence entre les médianes de charge mutationnelle totale normalisée des cellules classées comme *AML-like* et celles classées comme *non-AML-like*. Une valeur plus élevée indique donc une meilleure séparation entre les deux groupes.

La marge de 9 % a été retenue car elle présente un score de séparation égal à 0.85 ainsi qu'un pourcentage maximal de cellules *uncertain* égal à 16.76 % (Figure 14D). Cette marge a conduit à un seuil inférieur égal à 0.888 et à un seuil supérieur de 1.063 et à la classification de 1 372 cellules comme *AML-like*, 225 cellules comme *non-AML-like* et 150 cellules comme incertaines (Figure 14E). Après application d'une annotation majoritaire au niveau des *clusters* Seurat, plutôt qu'à l'échelle unicellulaire, 1 621 cellules sont classées comme *AML-like* et 126 cellules comme *non-AML-like* (Figure 14F). Le cluster 7 est ainsi annoté comme *non-AML-like*, car la majorité (> 50 %) de ses cellules présente une charge mutationnelle totale normalisée inférieure au seuil de classification, en accord avec la faible charge observée dans la Figure 14B. Les autres clusters sont classés comme *AML-like*.

Cette classification a ensuite été comparée aux annotations de référence utilisées dans l'étude de Naldini *et al.* (Figure 14G). Leur approche montre que les clusters 3, 4 et 7 sont annotés comme *non-AML-like*, tandis que les autres clusters sont annotés comme *AML-like*.

Marge testée	Cellules <i>uncertain</i> globales (%)	Cellules <i>uncertain</i> max. par cluster (%)	Score de séparation
0.06	5.61	9.83	0.826
0.07	6.70	11.56	0.834
0.08	7.56	13.97	0.845
0.09	8.70	16.76	0.851
0.10	9.67	20.23	0.863
0.11	10.59	21.97	0.876

Tableau 3. Sélection de la marge d'incertitude autour du seuil *k-means* chez le patient PT01 au diagnostic.

Plusieurs marges ont été testées afin d'évaluer leur impact sur le pourcentage de cellules classées comme *uncertain* et sur le score de séparation entre les groupes *AML-like* et *non-AML-like*. La marge de 9 % a été retenue comme compromis entre séparation des groupes et proportion limitée de cellules incertaines.

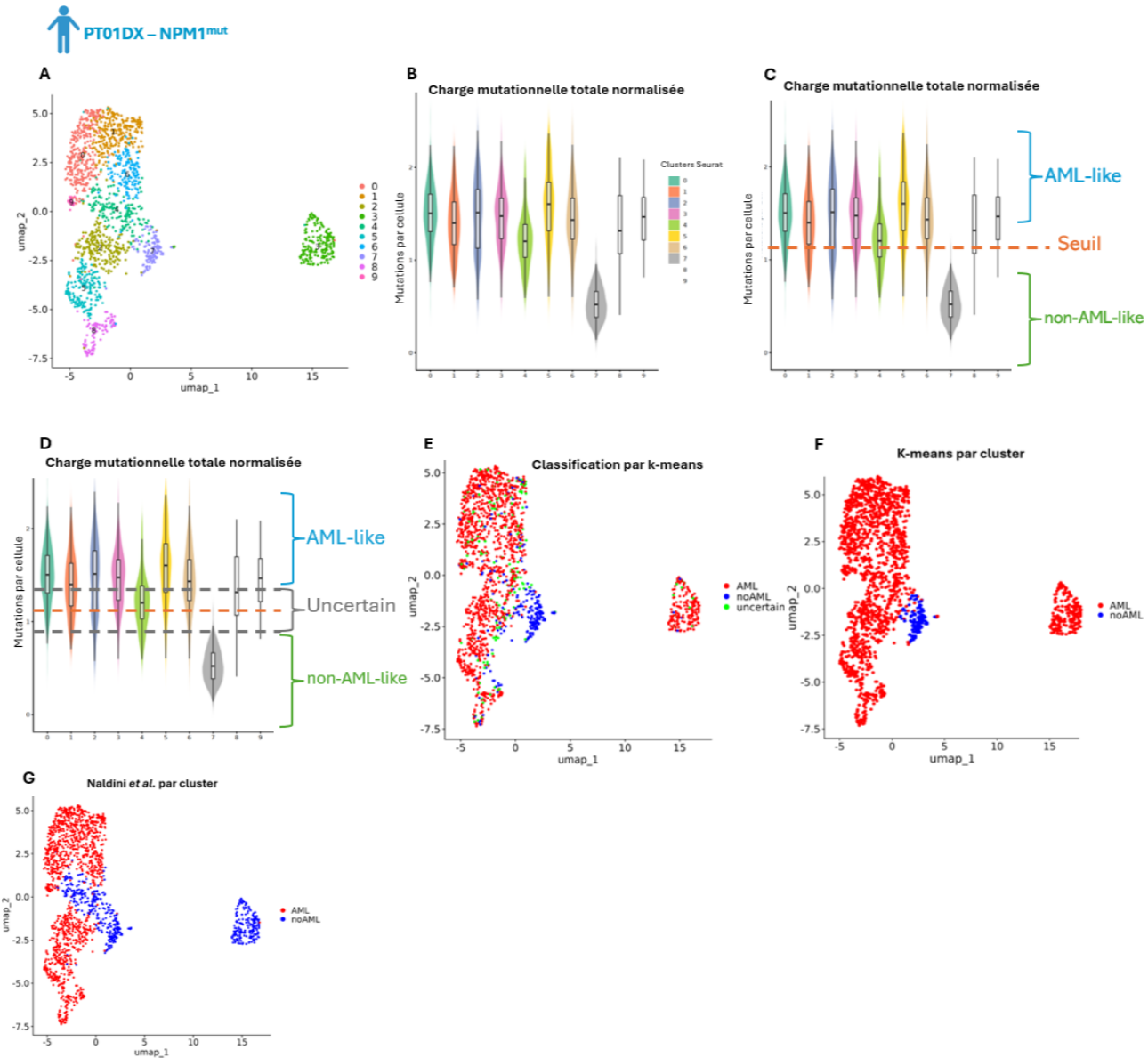


Figure 14. Application de la stratégie de classification basée sur la charge mutationnelle totale normalisée chez le patient PT01 au diagnostic.

A. UMAP des clusters Seurat **B.** Distribution de la charge mutationnelle totale normalisée par cluster. **C-D.** Visualisation du seuil *k-means* et de la marge d'incertitude (0.09). **E.** Classification unicellulaire en *AML-like*, *non-AML-like* et *uncertain*. **F.** Annotation finale par cluster selon la majorité des cellules. **G.** Annotation dérivée de Naldini *et al.*

Les classifications obtenues pour les autres points temporels du patient PT01 ainsi que pour les autres patients *NPM1^{mut}* ont été résumées au niveau cellulaire et au niveau des clusters Seurat, puis comparées aux annotations de référence de Naldini *et al.* (Tableau 4).

Afin de comprendre la divergence observée entre nos résultats et ceux de Naldini *et al.*, l'expression du gène *NPM1* a été analysée chez le patient PT01, présenté ici comme exemple détaillé des patients *NPM1^{mut}*. Chez ce patient, une diminution de l'expression de *NPM1* est observée entre le diagnostic et les points temporels post-thérapeutiques, avec des médianes de 2.84 à DX, 0.00 à D14 et 1.70 à D30 ($p < 0.001$) (Figure 15). Les données internes de notre laboratoire montrent également une diminution de l'expression de *NPM1* dans plusieurs lignées de LMA traitées par cytarabine, par rapport aux conditions contrôles (Annexe 1, résultats fournis par Bianca E. Silva). Ces résultats illustrent bien la limitation de l'approche standard de détection des cellules leucémiques : la perte ou la diminution d'expression du gène contenant la mutation d'intérêt peut limiter la détection de la mutation elle-même.

Afin d'évaluer la concordance avec les annotations de référence, les annotations de Naldini *et al.* ont été comparées à celles obtenues avec notre approche au niveau cellulaire (Tableau 4) et après attribution majoritaire au niveau des clusters Seurat (Tableau 5), pour l'ensemble des patients et points temporels analysés. Ainsi, certains échantillons peuvent contenir un faible nombre de cellules *AML-like* selon Naldini *et al.* au niveau cellulaire, sans qu'aucun cluster ne soit finalement annoté comme *AML-like* lorsque ces cellules représentent moins de 50 % du cluster. Le recouvrement cellulaire a été évalué à partir des codes-barres cellulaires communs entre les deux classifications, afin de déterminer quelles cellules étaient classées *AML-like* par les deux approches.

Au niveau cellulaire, un recouvrement partiel est observé au diagnostic chez les patients PT01, PT07 et PT08. Les cellules *AML-like* communes aux deux approches représentent 883 cellules chez PT01DX, soit 54.5 % des cellules *AML-like* identifiées par notre approche, 24 cellules chez PT07DX, soit moins de 10 %, et 547 cellules chez PT08DX, soit 28 %. Aux points temporels post-chimiothérapie, l'approche de Naldini *et al.* n'identifie aucune cellule *AML-like* chez l'ensemble des patients. À l'inverse, notre approche identifie encore des cellules *AML-like* pour ces mêmes points temporels (Tableau 4).

Au niveau des clusters Seurat, les clusters *AML-like* identifiés par l'approche de Naldini *et al.* au diagnostic sont inclus dans les clusters *AML-like* identifiés par notre approche pour l'ensemble des patients au diagnostic. En revanche, aucun cluster *AML-like* n'est identifié par l'approche de Naldini *et al.* aux points temporels post-chimiothérapie, alors que notre approche identifie encore plusieurs clusters *AML-like*. Cette différence est particulièrement notable chez les patients PT01 et PT08, correspondant respectivement à une maladie réfractaire et à une rechute précoce après RC (Tableau 5).

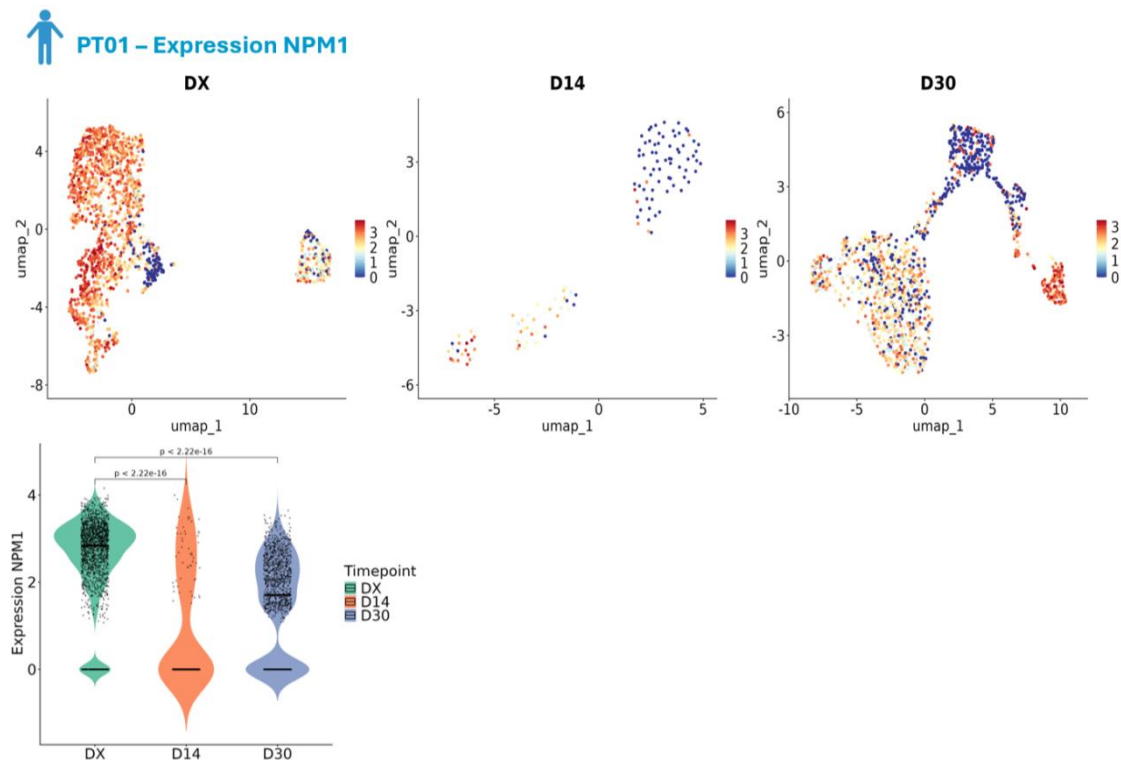


Figure 15. Expression de *NPM1* chez le patient *NPM1*^{mut} PT01 aux différents points temporels.

L'expression de *NPM1* est projetée sur UMAP pour les trois points temporels analysés (DX, D14, D30) et comparée par *violin plot*. Une diminution significative de l'expression est observée après traitement (comparaison par paires à l'aide de tests de Wilcoxon, $p < 0.001$).

Patient <i>NPM1</i>	Timepoint	AML-like	Non-AML-like	AML-like Naldini et al.	Non-AML-like Naldini et al.	Overlap Naldini/ means n(%)	k-
PT01	DX	1 619	126	1 243	504	883 (54.5 %)	
	D14	52	73	0	125	NA	
	D30	1 010	405	0	1 412	NA	
PT07	DX	278	426	40	664	24 (<10 %)	
	D14	146	98	0	238	NA	
PT08	DX	1 932	2 472	603	3 801	547 (28 %)	
	D14	37	23	0	60	NA	
	D30	1 937	579	0	2 516	NA	

Tableau 4. Répartition cellulaire des cellules AML-like et non-AML-like chez les patients *NPM1*^{mut}.

Le tableau présente, pour chaque patient et point temporel, le nombre de cellules classées comme AML-like ou non-AML-like par notre approche et par l'approche de Naldini *et al.*, ainsi que le recouvrement entre les deux classifications après annotation majoritaire au niveau des clusters Seurat (NA, non applicable).

Patients	Altération	Clusters AML-like – Naldini et al.	Clusters AML-like – K-means
PT01 - DX	<i>NPM1</i> ^{mut}	0, 1, 2, 5, 6, 8, 9	Tous sauf : 7
PT01 - D14		Aucun	1
PT01 - D30		Aucun	0, 1, 3, 4, 5
PT07 - DX	<i>NPM1</i> ^{mut}	6	0, 5, 6
PT07 - D14		Aucun	0, 1
PT08 - DX	<i>NPM1</i> ^{mut}	0, 8	0, 2, 8, 9, 10, 13, 14
PT08 - D14		Aucun	0
PT08 - D30		Aucun	0, 1, 2, 4, 5, 6, 9, 10, 11, 12, 14, 15

Tableau 5. Comparaison de clusters AML-like entre les annotations de Naldini *et al.* et l'approche k-means.

Une approche similaire à celle de Naldini *et al.* a été adoptée : si un cluster Seurat contient plus de 50 % de cellules annotées comme AML-like, tout le cluster est considéré comme AML-like.

4.3 Analyse détaillée d'un patient del(7)

Après l'analyse des patients *NPM1^{mut}*, la même stratégie de classification a été appliquée aux patients del(7), afin d'évaluer son application dans un contexte cytogénétique différent. Le patient PT17 au jour 14 après traitement a été sélectionné pour illustrer la classification dans un contexte post-thérapeutique précoce, après chimiothérapie d'induction (Figure 16).

Comme décrit pour le patient PT01DX, la classification a été réalisée à partir de la charge mutationnelle totale normalisée. Un seuil central a été défini par *k-means*, puis une marge d'incertitude a été appliquée autour de ce seuil afin de limiter les classifications trop proches de la zone de séparation entre les groupes *AML-like* et *non-AML-like*.

Chez PT17D14, cette approche conduit à la classification de 1 761 cellules comme *AML-like*, 316 cellules comme *non-AML-like* et 214 cellules comme *uncertain* avant annotation majoritaire (Figure 16A). Après attribution majoritaire au niveau des clusters Seurat, 2 121 cellules sont classées comme *AML-like* et 170 cellules comme *non-AML-like*. Les clusters 7 et 8 sont annotés comme *non-AML-like*, tandis que les autres clusters sont classés comme *AML-like* (Figure 16B). Cette classification a ensuite été comparée aux annotations de référence dérivées de l'étude de Naldini *et al.*, dans lesquelles les clusters 8 et 11 sont annotés comme *non-AML-like*, tandis que les autres clusters sont annotés comme *AML-like* (Figure 16C).

Les classifications obtenues pour les patients PT17, PT11 et PT18, ont ensuite été résumées afin d'évaluer la concordance entre notre approche et les annotations de Naldini *et al.* Cette comparaison a été réalisée au niveau cellulaire et après annotation majoritaire au niveau des clusters Seurat (Tableaux 6 et 7).

Chez les patients del(7), contrairement aux patients *NPM1^{mut}* post-chimiothérapie, des populations *AML-like* sont identifiées par les deux approches à l'ensemble des points temporels analysés, y compris après traitement. Au niveau cellulaire, plusieurs échantillons présentent un recouvrement important entre les cellules *AML-like* identifiées par l'approche de Naldini *et al.* et celles retrouvées par notre approche, notamment PT11DX (93.1 %), PT17DX (94 %), PT17D14 (97.8 %) et PT18D14 (98.3 %). Ce recouvrement reste toutefois variable selon les patients et les points temporels, avec des valeurs plus faibles pour PT11D30 (62.3 %), PT18DX (64.7 %) et surtout PT17D30 (32.5 %) (Tableau 6). Le pourcentage de recouvrement indique la proportion de cellules *AML-like* identifiées par notre approche qui sont également classées *AML-like* par l'approche de Naldini *et al.*

Au niveau des clusters Seurat, les deux approches identifient également des clusters *AML-like* chez les patients del(7), mais les clusters concernés ne sont pas toujours strictement identiques (Tableau 7).

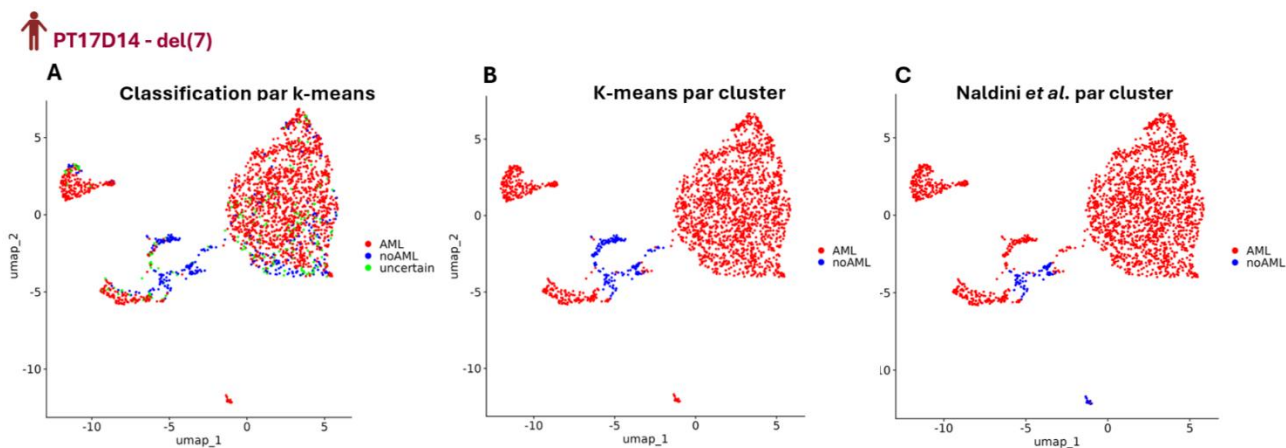


Figure 16. Classification AML-like/non-AML-like du patient PT17D14 basée sur la charge mutationnelle totale normalisée.

A. Projection UMAP de la classification cellule par cellule obtenue par k-means à partir de la charge mutationnelle totale normalisée. Les cellules sont classées en AML-like, non-AML-like ou incertaines (*uncertain*). **B.** Projection UMAP de la classification finale obtenue après annotation majoritaire au niveau des clusters Seurat. Après cette annotation par cluster, 2 121 cellules sont assignées au groupe AML-like et 170 cellules au groupe non-AML-like. **C.** Projection UMAP des annotations AML-like/non-AML-like dérivées de Naldini et al. après annotation au niveau des clusters Seurat.

Patient del(7)	Timepoint	AML-like	Non-AML-like	AML-like Naldini et al.	Non-AML-like Naldini et al.	Overlap Naldini/k-means n(%)
PT11	DX	1 216	127	1 225	793	1133 (93.1 %)
	D30	2 102	4 415	1 504	3 608	1 311 (62.3 %)
PT17	DX	1 172	846	1 310	33	1 102 (94 %)
	D14	2 121	170	2 194	97	2 075 (97.8 %)
PT18	D30	2 254	3 720	1 011	5 506	732 (32.5 %)
	DX	2 727	934	2 020	1 641	1 765 (64.7 %)
PT18	D14	1 457	491	1 501	447	1 432 (98.3 %)

Tableau 6. Répartition des cellules AML-like et non-AML-like chez les patients del(7) après annotation majoritaire au niveau des clusters Seurat.

Le tableau compare les annotations obtenues avec notre approche *k-means* aux annotations de référence de Naldini et al. pour chaque patient et point temporel. Le recouvrement correspond au nombre et au pourcentage de cellules AML-like identifiées par notre approche qui sont également classées comme AML-like par Naldini et al.

Patients	Altération	Clusters AML-like – Naldini et al.	Clusters AML-like – K-means
PT11 - DX	del(7)	0, 1, 2, 5, 9	0, 1, 2, 8
PT11 – D30		0, 3	Tous sauf : 4, 11, 14
PT17 – DX	del(7)	Tous sauf : 8	Tous sauf : 6, 7, 8
PT17 – D14		Tous sauf : 8, 11	Tous sauf : 7, 8
PT17 – D30		4, 8, 14	4, 5, 8, 10, 13, 14
PT18 - DX	del(7)	Tous sauf : 0, 3, 4, 8, 9	Tous sauf : 3, 8, 9, 11
PT18 – D14		Tous sauf : 2, 6, 8	Tous sauf : 2, 6, 8, 9

Tableau 7. Comparaison des clusters AML-like entre les annotations de l'approche *k-means* et les annotations de Naldini et al. chez les patients del(7).

4.4 Analyse transcriptomique complémentaire des cellules *AML-like*

Après la classification des cellules en *AML-like* et *non-AML-like*, une analyse transcriptomique a été réalisée afin d'évaluer si les cellules *AML-like* identifiées présentaient des signatures compatibles avec un état de persistance post-thérapeutique.

4.4.1 L'expression de signatures transcriptomiques des DTPs soutient nos annotations

Dans cette analyse, le score IFN- γ a été utilisé comme signature transcriptomique associée à un état de persistance post-thérapeutique. Cette signature provient d'une analyse interne réalisée au laboratoire sur des lignées cellulaires de LMA traitées par cytarabine pendant trois jours, puis lavées et séquencées par *bulk* RNA-seq 24 heures après le retrait du traitement. Elle correspond à un programme transcriptionnel associé à la réponse à l'IFN- γ , observé dans une population de cellules de LMA ayant survécu à une exposition à la cytarabine.

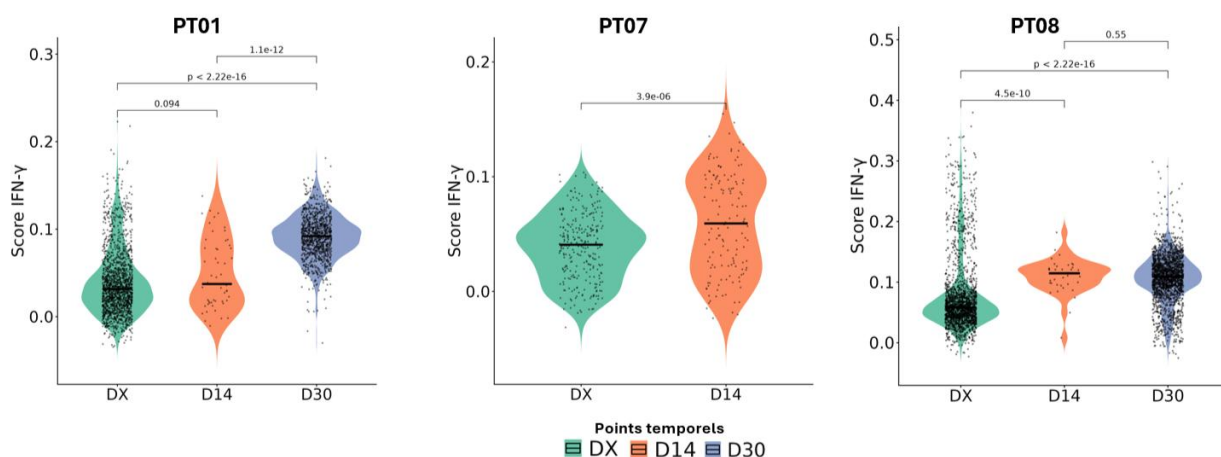
Le score IFN- γ a d'abord été comparé entre les différents points temporels au sein des cellules *AML-like* des patients *NPM1^{mut}* (Figure 17A). Chez PT01, une augmentation progressive du score est observée entre DX, D14 et D30. Chez PT07, le score est plus élevé à D14 qu'au diagnostic. Chez PT08, le score augmente à D14 par rapport au diagnostic, puis reste élevé à D30, bien que la comparaison entre D14 et D30 ne montre pas de différence significative. Une surexpression de cette signature post-chimiothérapie dans les cellules annotées *AML-like* est compatible avec l'idée que notre approche capture des cellules leucémiques survivant à la chimiothérapie.

Le score IFN- γ a ensuite été comparé entre les cellules *AML-like* et *non-AML-like* pour chaque point temporel (Figure 17B). Chez PT01, le score est significativement plus élevé dans les cellules *AML-like* au diagnostic et à D30, tandis qu'aucune différence significative n'est observée à D14. Chez PT07, une différence significative est observée à DX et D14, avec un score plus élevé dans les cellules *AML-like* à D14. Chez PT08, des différences significatives sont observées aux trois points temporels analysés.

Chez les patients *del(7)*, le score IFN- γ a également été analysé au cours du temps dans les cellules *AML-like* (Figure 18A). Chez PT11, une différence significative est observée entre DX et D30. Chez PT17, le score augmente fortement à D14 et D30. Chez PT18, une augmentation significative du score est observée à D14 par rapport au diagnostic.

La comparaison entre cellules *AML-like* et *non-AML-like* chez les patients *del(7)* montre également des différences selon les patients et les points temporels (Figure 18B). Chez PT11, le score IFN- γ est significativement différent entre les deux populations à DX et D30. Chez PT17, aucune différence significative n'est observée au diagnostic, tandis que des différences significatives apparaissent à D14 et D30. Chez PT18, des différences significatives sont observées à DX et D14.

A Évolution du score IFN-γ dans les cellules AML-like



B Comparaison de score IFN-γ entre cellules AML-like et non-AML-like

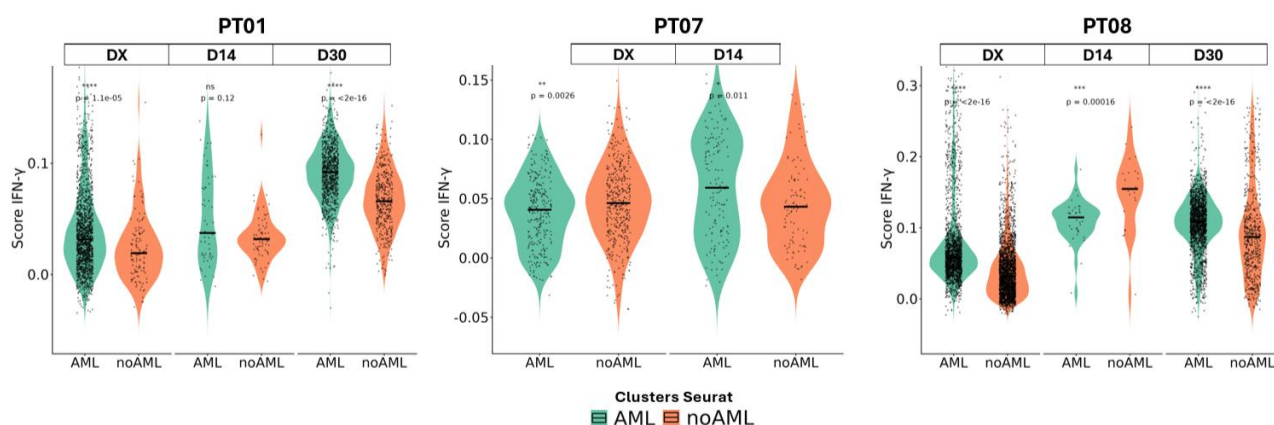
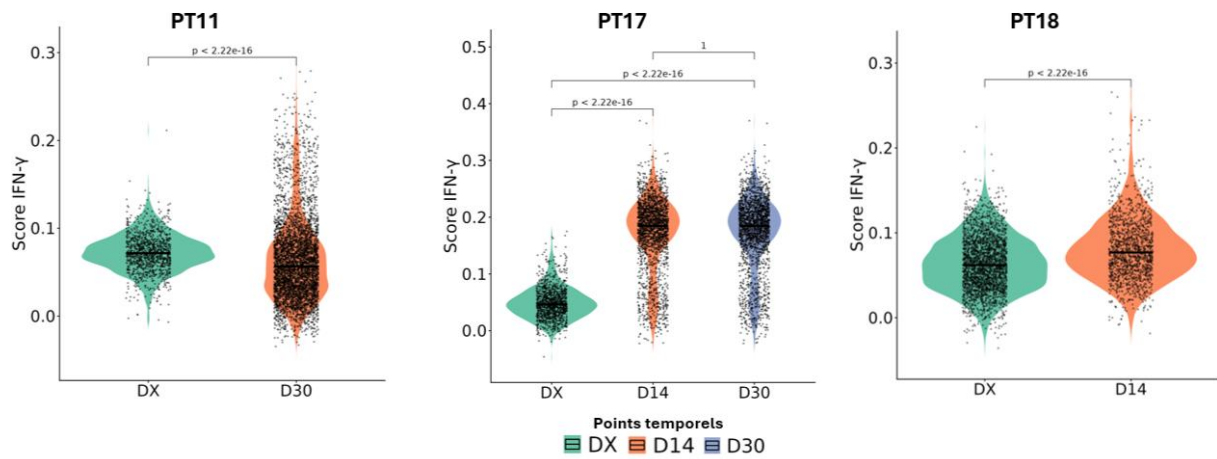


Figure 17. Analyse du score IFN-γ chez les patients *NPM1*^{mut}.

A. Évolution du score IFN-γ dans les cellules AML-like entre les différents points temporels. Chez PT01, le score augmente significativement entre DX et D30 ($p < 2.22 \times 10^{-16}$) ainsi qu'entre D14 et D30 ($p = 1.1 \times 10^{-12}$), tandis que la différence entre DX et D14 n'est pas significative ($p = 0.094$). Chez PT07, le score augmente significativement entre DX et D14 ($p = 3.9 \times 10^{-6}$). Chez PT08, le score augmente significativement entre DX et D14 ($p = 4.5 \times 10^{-10}$) et entre DX et D30 ($p < 2.22 \times 10^{-16}$), tandis que la différence entre D14 et D30 n'est pas significative ($p = 0.55$).

B. Comparaison du score IFN-γ entre cellules AML-like et non-AML-like à chaque point temporel. Des différences significatives sont observées chez PT01 à DX ($p = 1.1 \times 10^{-5}$) et D30 ($p < 2 \times 10^{-16}$), mais pas à D14 ($p = 0.12$). Chez PT07, les différences sont significatives à DX ($p = 0.0026$) et D14 ($p = 0.011$). Chez PT08, les différences sont significatives à DX ($p < 2 \times 10^{-16}$), D14 ($p = 0.00016$) et D30 ($p < 2 \times 10^{-16}$). Les comparaisons statistiques ont été réalisées par tests de Wilcoxon.

A Évolution du score IFN-γ dans les cellules AML-like



B Comparaison de score IFN-γ entre cellules AML-like et non-AML-like

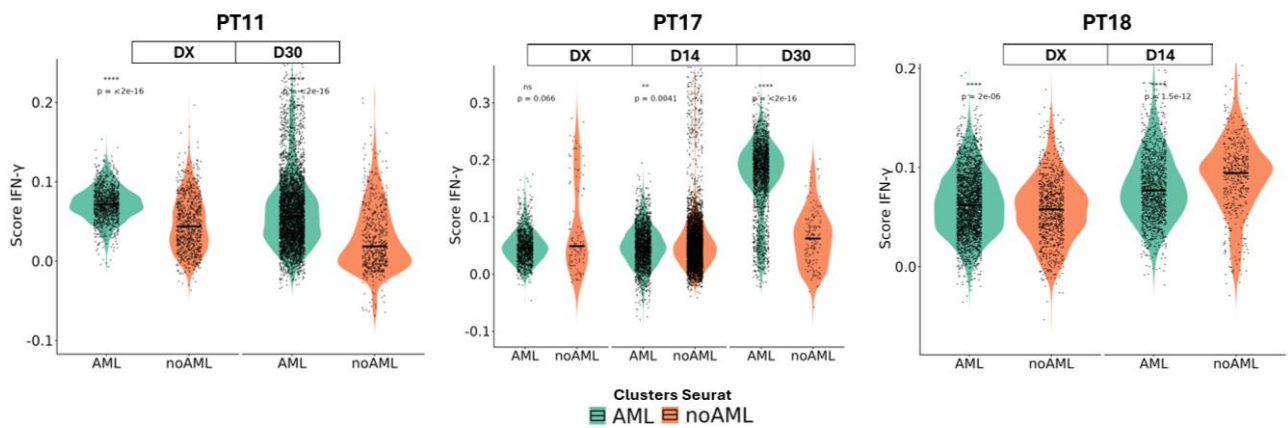


Figure 18. Analyse du score IFN-γ chez les patients del(7).

A. Évolution du score IFN-γ dans les cellules AML-like entre les différents points temporels. Chez PT11, le score diffère significativement entre DX et D30 ($p < 2.22 \times 10^{-16}$). Chez PT17, le score augmente significativement entre DX et D14 ($p < 2.22 \times 10^{-16}$) ainsi qu'entre DX et D30 ($p < 2.22 \times 10^{-16}$), tandis qu'aucune différence n'est observée entre D14 et D30 ($p = 1$). Chez PT18 le score augmente significativement entre DX et D14 ($p < 2.22 \times 10^{-16}$).

B. Comparaison du score IFN-γ entre cellules AML-like et non-AML-like à chaque point temporel. Chez PT11, des différences significatives sont observées à DX et D30 ($p < 2 \times 10^{-16}$). Chez PT17, aucune différence significative n'est observée à DX ($p = 0.066$), tandis que des différences significatives sont observées à D14 ($p = 0.0041$) et D30 ($p < 2 \times 10^{-16}$). Chez PT18 des différences significatives sont observées à DX ($p = 2 \times 10^{-6}$) et D14 ($p = 1.5 \times 10^{-12}$). Les comparaisons statistiques ont été réalisées par tests de Wilcoxon.

4.4.2 Une signature transcriptomique LMA soutient nos annotations

Afin de valider nos annotations AML-like / non-AML-like, nous avons utilisé un ensemble de gènes composé des 600 gènes associés à la LMA selon la base de données GeneCards⁹². Cette signature a ensuite été utilisée afin de calculer un score LMA pour les cellules des patients *NPM1*^{mut} et del(7).

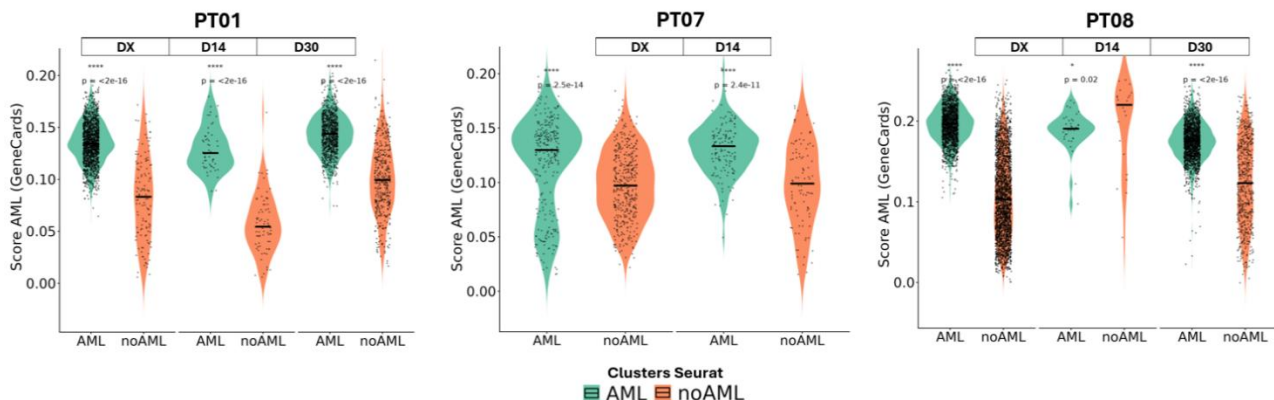
Dans un premier temps, ce score a été comparé entre les cellules AML-like et non-AML-like définies par notre approche *k-means* chez les patients *NPM1*^{mut} (Figure 19A). Une analyse similaire a ensuite été réalisée à partir des annotations de référence de Naldini *et al.*, afin d'évaluer si celles-ci présentaient un contraste comparable pour le score GeneCards (Figure 19B).

Avec notre approche, le score LMA GeneCards est globalement plus élevé dans les cellules AML-like que dans les cellules non-AML-like, avec des différences significatives dans la majorité des comparaisons. Les annotations de Naldini *et al.* montrent également un contraste dans certains échantillons, notamment au diagnostic. Cette comparaison est plus limitée aux points temporels après traitement chez les patients *NPM1*^{mut}, en raison du faible nombre ou de l'absence de cellules AML-like identifiées par cette approche après traitement.

La même analyse a ensuite été réalisée chez les patients del(7) (Figure 20). Avec l'approche *k-means*, le score LMA GeneCards est globalement plus élevé dans les cellules AML-like que dans les cellules non-AML-like, avec des différences significatives dans la majorité des comparaisons. Chez PT17DX, la différence n'est pas significative, ce qui indique que le contraste transcriptomique entre les deux groupes peut varier selon les patients et les points temporels.

Les annotations de Naldini *et al.* montrent également un contraste significatif entre cellules AML-like et non-AML-like dans plusieurs échantillons del(7). Contrairement aux patients *NPM1*^{mut} post-chimiothérapie, les annotations de Naldini *et al.* identifient encore des cellules AML-like après traitement, ce qui permet une comparaison plus complète entre les deux approches.

A Score GeneCards selon l'approche k-means



B Score GeneCards selon les annotations de Naldini *et al.*

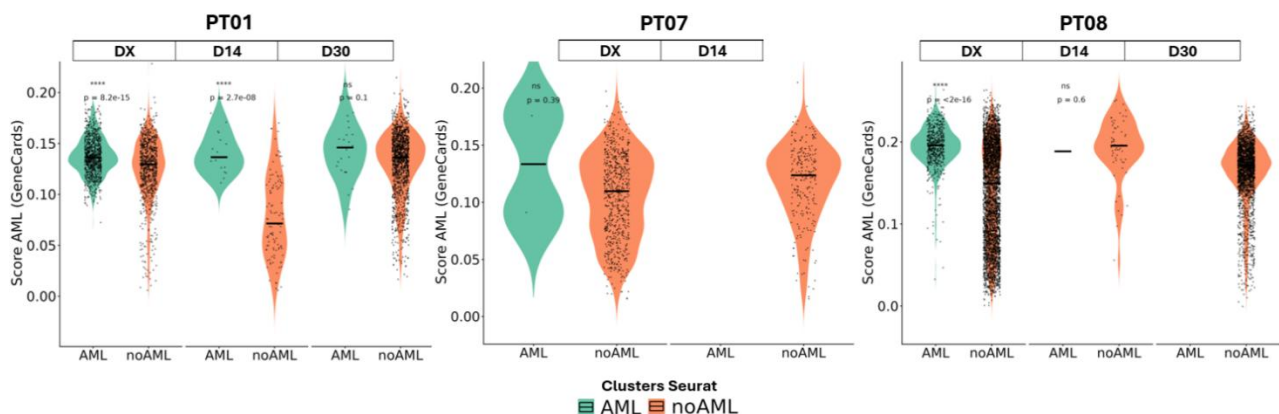


Figure 19. Comparaison du score LMA GeneCards chez les patients *NPM1*^{mut}.

A. Comparaison du score LMA GeneCards entre cellules AML-like et non-AML-like selon l'approche k-means.

Des différences significatives sont observées dans la majorité des comparaisons, notamment chez PT01 à DX, D14 et D30 ($p < 2 \times 10^{-16}$), chez PT07 à DX ($p = 2.5 \times 10^{-14}$) et D14 ($p = 2.4 \times 10^{-11}$), ainsi que chez PT08 à DX et D30 ($p < 2 \times 10^{-16}$) et à D14 ($p = 0.02$).

B. Comparaison du score LMA GeneCards entre cellules AML-like et non-AML-like selon les annotations de Naldini *et al.*

Chez PT01, des différences significatives sont observées à DX ($p = 8.2 \times 10^{-15}$) et D14 ($p = 2.7 \times 10^{-8}$) mais pas à D30 ($p = 0.02$). Chez PT07, aucune différence significative n'est observée à DX ($p = 0.39$). Chez PT08, des différences significatives sont observées à DX ($p = 2 \times 10^{-16}$). Les comparaisons statistiques ont été réalisées par tests de Wilcoxon.

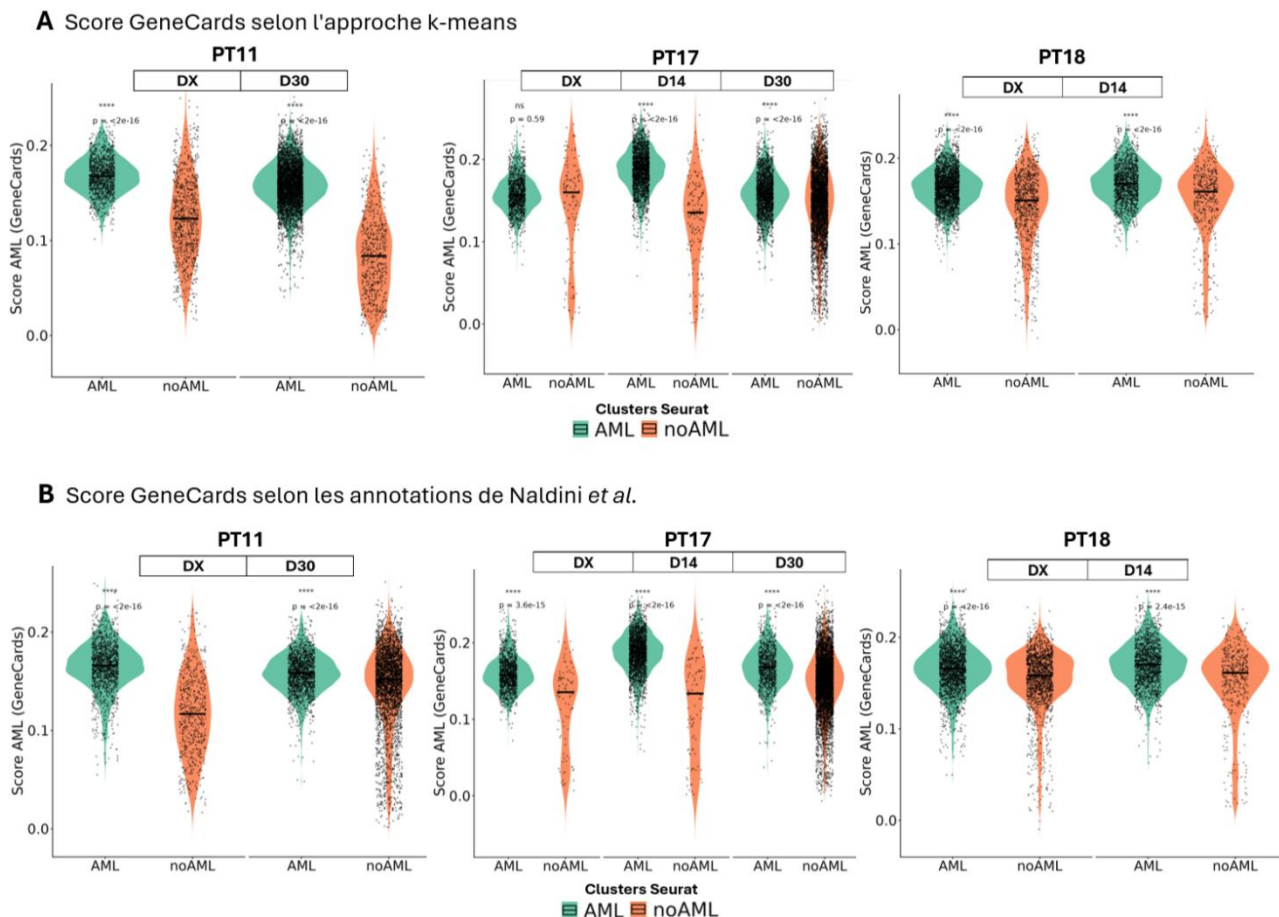


Figure 20. Comparaison du score LMA GeneCards chez les patients del(7).

A. Comparaison du score LMA GeneCards entre cellules AML-like et non-AML-like selon l'approche k-means.

Des différences significatives sont observées dans la majorité des comparaisons, notamment chez PT11 à DX et D30 ($p < 2 \times 10^{-16}$), chez PT17 à D14 et D30 ($p < 2 \times 10^{-16}$), ainsi que chez PT18 à DX et D14 ($p < 2 \times 10^{-16}$). La comparaison chez PT17-DX n'est pas significative ($p = 0.59$).

B. Comparaison du score LMA GeneCards entre cellules AML-like et non-AML-like selon les annotations de Naldini *et al.*

Des différences significatives sont observées chez PT11 à DX et D30 ($p < 2 \times 10^{-16}$), chez PT17 à DX ($p = 3.6 \times 10^{-15}$), D14 et D30 ($p < 2 \times 10^{-16}$), ainsi que chez PT18 à DX ($p < 2 \times 10^{-16}$) et D14 ($p = 2.4 \times 10^{-15}$). Les comparaisons statistiques ont été réalisées par tests de Wilcoxon.

Afin de compléter cette comparaison quantitative, le score LMA GeneCards a ensuite été projeté sur les UMAP des patients *NPM1*^{mut} et del(7), en parallèle des annotations AML-like et non-AML-like obtenues par notre approche k-means (Figure 21 et 22).

Cette visualisation permet d'évaluer si les zones présentant un score LMA élevé recouvrent spatialement les populations annotées AML-like. Chez les patients *NPM1*^{mut} comme chez les patients del(7), les projections UMAP montrent un recouvrement partiel.

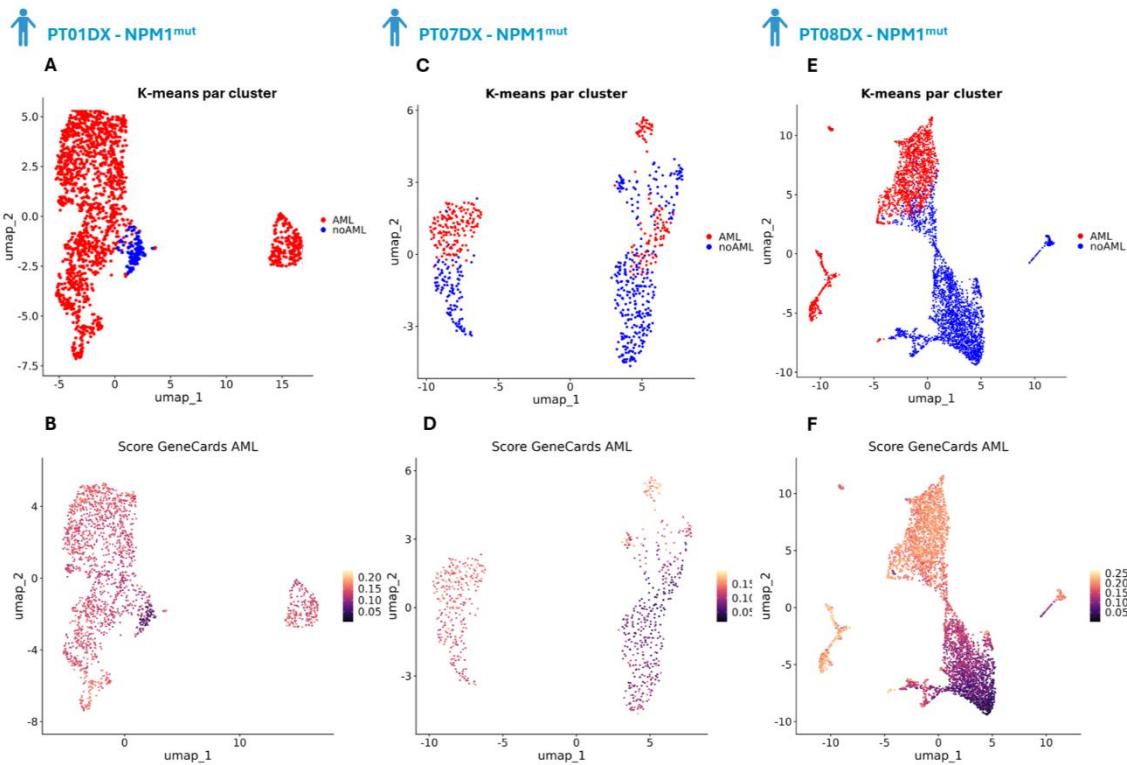


Figure 21. Projection des scores LMA GeneCards chez les patients *NPM1*^{mut}.

A, C, E. Projection UMAP de la classification finale obtenue après annotation majoritaire au niveau des clusters Seurat à partir de la classification par k-means. **B, D, F.** Projection UMAP du score LMA calculé à partir de la signature LMA GeneCards de 600 gènes.

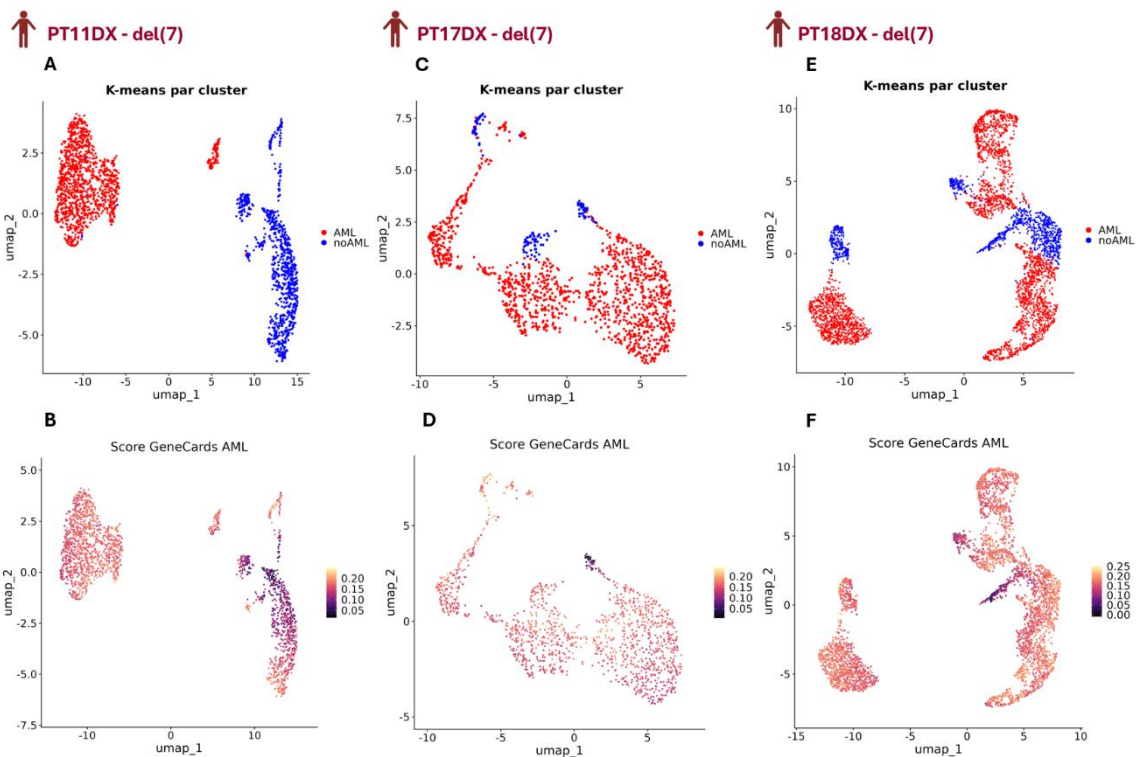


Figure 22. Projection des scores LMA GeneCards chez les patients *del(7)*.

A, C, E. Projection UMAP de la classification finale obtenue après annotation majoritaire au niveau des clusters Seurat à partir de la classification par k-means. **B, D, F.** Projection UMAP du score LMA calculé à partir de la signature LMA GeneCards de 600 gènes.

5. Discussion

L'objectif de ce travail était de développer une stratégie permettant d'identifier, à partir de données scRNA-seq, des populations cellulaires compatibles avec une origine leucémique dans les échantillons de LMA analysés à différents points temporels. Cette caractérisation constitue une étape importante vers le suivi clonal, car il est d'abord nécessaire de distinguer les cellules potentiellement *AML-like* des cellules *non-AML-like* avant de pouvoir étudier leur évolution au cours du traitement.

Cette approche répond aux limites des annotations reposant sur la détection d'un seul transcrit muté, comme *NPM1*, ou sur des signatures propres à une altération donnée, comme la perte du chromosome 7. Les résultats obtenus seront donc discutés selon quatre axes principaux : la pertinence de la charge mutationnelle totale normalisée comme signal d'annotation, la comparaison avec les annotations de référence dérivées de Naldini *et al.*, la cohérence transcriptomique des populations identifiées comme *AML-like* à travers le score IFN- γ et la signature GeneCards associée à la LMA, ainsi que les limites et perspectives de cette stratégie dans une optique de suivi clonal.

5.1 Pertinence du signal d'annotation cellulaire

L'un des premiers enjeux méthodologiques de ce travail était de déterminer si la charge mutationnelle détectée à partir des données de scRNA-seq pouvait constituer un signal exploitable pour distinguer des populations cellulaires compatibles avec une origine leucémique. Cette hypothèse repose sur l'idée que, même si la LMA est généralement considérée comme une pathologie à faible charge mutationnelle en analyse *bulk*, certaines cellules pourraient présenter à l'échelle unicellulaire un signal mutationnel plus élevé que les cellules non leucémiques⁹⁵. Cette idée reste cohérente avec des travaux récents montrant que les variants détectés dans les données scRNA-seq peuvent contenir une information utile pour identifier des populations cancéreuses ou clonales^{95,114}.

La première étape consistait à évaluer l'influence de la complexité transcriptomique sur le nombre brut de variants détectés. En effet, une cellule dans laquelle un plus grand nombre de gènes est détecté présente potentiellement plus de variants dans les transcrits séquencés⁹⁵. Dans ce cas, une charge mutationnelle brute élevée pourrait résulter d'une quantité d'information transcriptomique plus importante, plutôt que d'une réelle différence biologique liée au statut leucémique⁸⁹.

Nos résultats montrent une corrélation positive et significative entre le nombre brut de variants détectés et le nombre de gènes détectés par cellule, estimé par `nFeature_RNA`, chez l'ensemble des patients analysés au diagnostic (Figure 12). Cette observation justifie l'utilisation d'une charge mutationnelle normalisée par `nFeature_RNA` dans les analyses suivantes, afin de réduire l'effet de la complexité transcriptomique sur le signal mutationnel calculé.

Après normalisation, la charge mutationnelle totale normalisée a été comparée à la charge mutationnelle somatique normalisée. Cette comparaison montre une corrélation positive et significative chez l'ensemble des patients analysés, avec des intensités variables selon les échantillons (Figure 13). Ces résultats suggèrent que la charge mutationnelle totale normalisée, bien qu'elle intègre l'ensemble des variants détectés sans distinction stricte de leur origine, reflète au moins partiellement la distribution du signal somatique à l'échelle cellulaire.

L'utilisation de ce signal répond à une limite importante des approches reposant uniquement sur la transcriptomique, car les cellules cancéreuses et non cancéreuses peuvent être difficiles à distinguer sur la seule base de leur profil d'expression^{56,95,114}. Le signal mutationnel apporte alors une information complémentaire : deux cellules proches sur le plan transcriptomique peuvent potentiellement être distinguées par l'analyse des variants qu'elles portent^{79,114}.

Plusieurs travaux montrent l'intérêt d'intégrer une information mutationnelle aux données scRNA-seq. Petti *et al.* ont montré que des mutations somatiques, préalablement identifiées par séquençage du génome entier, pouvaient être recherchées dans les données scRNA-seq afin d'identifier des cellules porteuses de mutations spécifiques de la tumeur et de distinguer les cellules leucémiques des cellules normales⁷⁹. De manière complémentaire, Velten *et al.* ont montré que la combinaison de la transcriptomique unicellulaire avec un suivi clonal basé sur des SNV nucléaires et des variants mitochondriaux permettait d'identifier et de profiler des CSL, des cellules pré-leucémiques et des cellules hématopoïétiques normales¹¹⁴. Plus récemment, Marot-Lassauzaie *et al.* ont proposé une approche permettant d'exploiter des SNV appelés directement à partir de données scRNA-seq afin d'identifier des clones cancéreux⁹⁵. Contrairement aux approches reposant sur un variant connu au préalable, cette approche cherche des profils de variants retrouvés ensemble dans les mêmes cellules afin d'identifier des populations clonales. Cette étude soutient donc l'idée que l'information mutationnelle contenue dans les données scRNA-seq peut contribuer à distinguer des cellules cancéreuses des cellules non cancéreuses⁹⁵.

L'ensemble de ces travaux soutient l'idée que l'information mutationnelle extraite de données unicellulaires peut apporter une information complémentaire à la transcriptomique pour identifier des populations leucémiques^{79,114}. À la différence de ces approches, notre stratégie utilise une charge mutationnelle totale normalisée. Ce choix vise à capturer un signal global moins dépendant de la détection d'un seul variant, d'un seul transcrit muté ou de bases d'annotation externes telles que COSMIC et dbSNP, dans lesquelles certaines mutations peuvent ne pas être répertoriées ou certains variants rares peuvent être mal classés¹¹⁵.

5.2 Comparaison avec les annotations de référence dérivées de Naldini et al.

La classification obtenue par notre approche a ensuite été comparée aux annotations de référence dérivées de Naldini et al. Il est cependant important de préciser que ces deux approches ne reposent pas sur le même signal.

Chez les patients *NPM1^{mut}*, une concordance partielle est observée au diagnostic. Les clusters *AML-like* identifiés par Naldini et al. sont inclus dans les clusters annotés *AML-like* par notre approche pour les trois patients analysés (Tableau 5). Ce résultat constitue un contrôle positif important, car il indique que la charge mutationnelle totale normalisée permet de retrouver des populations considérées comme *AML-like* par l'approche de référence au diagnostic.

La différence principale apparaît après chimiothérapie. Chez les trois patients *NPM1^{mut}*, les cellules *AML-like* détectées par l'approche dérivée de Naldini et al. ne sont pas suffisamment représentées au sein des clusters Seurat pour conduire à l'annotation de clusters *AML-like* après attribution majoritaire (Tableau 5, Annexe 2A). À l'inverse, notre approche identifie encore des populations *AML-like* à D14 et D30 après annotation majoritaire par cluster (Tableaux 4 et 5). Cette observation est principalement retrouvée chez les patients PT01 et PT08, qui correspondent respectivement à une maladie réfractaire et à une rechute précoce après RC (Tableau 2). Dans ces deux contextes cliniques défavorables, la persistance de populations *AML-like* détectées par notre approche est cohérente avec l'hypothèse d'une persistance leucémique après traitement.

Cette divergence peut s'expliquer par la dépendance de l'approche de Naldini et al. à la détection du transcrit *NPM1* muté. Nos résultats montrent une diminution de l'expression de *NPM1* après traitement chez le patient PT01, avec une médiane passant de 2.84 au diagnostic à 0.00 à D14, puis 1.70 à D30 (Figure 15). De plus, cette observation est cohérente avec les résultats internes obtenus dans notre laboratoire, montrant une diminution de l'expression de *NPM1* dans les lignées de LMA traitées par cytarabine (Annexe 1)¹¹⁶. Ces éléments soutiennent l'hypothèse qu'une diminution de l'expression de *NPM1* après traitement pourrait limiter la détection des cellules leucémiques par une approche qui repose fortement sur l'identification de la mutation *NPM1* dans les transcrits détectés.

Cette limite est également renforcée par le fait que le profil mutationnel des cellules leucémiques peut évoluer sous pression thérapeutique. Ding et al. ont montré, par séquençage du génome entier d'échantillons prélevés chez les mêmes patients au diagnostic et à la rechute, que la rechute est associée à une évolution clonale. Cette évolution peut impliquer soit l'acquisition de nouvelles mutations par le clone fondateur, soit la survie puis l'expansion d'un sous-clone déjà présent au diagnostic¹¹⁷. Dans ce cas, une approche reposant principalement sur la mutation du gène *NPM1* peut

devenir moins informative après traitement, tandis qu'une approche tenant compte d'un signal mutationnel global pourrait continuer à détecter des populations *AML-like* persistantes.

Chez les patients *del(7)*, la concordance entre les deux approches est globalement meilleure. Les populations *AML-like* sont détectées par les deux méthodes à plusieurs points temporels, y compris après traitement (Tableaux 6 et 7). Cette différence avec les patients *NPM1^{mut}* peut s'expliquer par la nature du signal utilisé par Naldini *et al.* En effet, chez les patients *del(7)*, l'annotation repose sur un score d'expression calculé à partir des gènes localisés sur le chromosome 7. L'hypothèse est qu'une cellule portant cette altération devrait présenter une diminution globale de l'expression de ces gènes, ce qui permettrait de distinguer les cellules *AML-like* des cellules *non-AML-like*⁵⁶. Ce signal, dépendant de plusieurs gènes, est probablement moins dépendant de l'expression d'un seul transcrit que l'approche utilisée chez les patients *NPM1^{mut}*. Cela peut expliquer pourquoi les annotations de référence restent informatives chez les patients *del(7)* après traitement, alors qu'elles deviennent moins informatives chez les patients *NPM1^{mut}* après chimiothérapie. Cette interprétation est cohérente avec le principe selon lequel des signatures multi-gènes peuvent fournir un signal discriminant plus robuste pour distinguer des populations cellulaires en scRNA-seq¹¹⁸.

Cette différence entre patients *NPM1^{mut}* et *del(7)* souligne une limitation importante de l'approche de référence : selon l'altération considérée, la classification ne repose pas sur un signal de même nature. Dans ce contexte, notre approche présente l'intérêt d'appliquer une logique commune à l'ensemble des patients, en utilisant la charge mutationnelle totale normalisée comme signal d'annotation. Elle pourrait ainsi fournir un signal plus homogène entre différents contextes génétiques pour annoter des populations *AML-like* à partir de données scRNA-seq.

5.3 Cohérence transcriptomique des populations *AML-like* identifiées

Après la comparaison avec les annotations dérivées de Naldini *et al.*, une analyse transcriptomique complémentaire a été réalisée afin de déterminer si les populations *AML-like* identifiées par la charge mutationnelle totale normalisée présentaient également des caractéristiques compatibles avec une origine leucémique. Deux approches ont été utilisées : un score IFN- γ , associé à une réponse transcriptionnelle observée après exposition à la cytarabine, et une signature GeneCards associée à la LMA⁹².

Le score IFN- γ correspond à un programme transcriptionnel observé dans des cellules leucémiques ayant survécu à une exposition à la cytarabine. Dans nos résultats, la plupart des patients présentent une augmentation du score IFN- γ dans les cellules *AML-like* après traitement, à l'exception du patient PT11 (Figures 17 et 18). Chez ce dernier, le score IFN- γ ne présente pas d'augmentation entre DX et D30, la différence observée reflète plutôt une modification de la distribution, avec une médiane à D30

légèrement inférieure à celle du diagnostic ($p < 0.001$). Cette absence d'augmentation claire ne contredit pas nécessairement l'annotation *AML-like*, car le score IFN- γ reflète un état transcriptionnel particulier et non un marqueur universel de toutes les cellules leucémiques persistantes. Cette variabilité peut être liée à l'hétérogénéité inter-patient, au contexte thérapeutique, notamment au profil clinique de PT11, marqué par une maladie réfractaire et deux HSCT, ainsi qu'aux différences de signal IFN- γ selon les sous-types de LMA et les états cellulaires (Tableau 2)¹¹⁹.

De plus, dans plusieurs échantillons, ce score est significativement plus élevé dans les cellules *AML-like* que dans les cellules *non-AML-like* (Figures 17 et 18). Ces observations suggèrent que les cellules *AML-like* identifiées par notre approche ne sont pas uniquement définies par une charge mutationnelle élevée, mais présentent aussi un état transcriptionnel compatible avec une réponse post-thérapeutique. Cette interprétation est cohérente avec les travaux montrant que la signalisation IFN- γ peut jouer un rôle antitumoral, mais peut également être associée à des phénomènes de résistance thérapeutique^{119,120}. Il est cependant important de préciser que l'activation d'un programme IFN- γ ne constitue pas une preuve directe de la persistance leucémique ou d'une identité de DTP, mais reste un élément de cohérence biologique supplémentaire.

En complément, la signature GeneCards associée à la LMA a été utilisée afin d'évaluer si les cellules annotées *AML-like* présentaient un score transcriptionnel globalement associé à la LMA⁹². Chez les patients *NPM1*^{mut} comme chez les patients del(7), le score GeneCards associé à la LMA est globalement plus élevé dans les cellules *AML-like* que dans les cellules *non-AML-like*, avec des différences significatives dans la majorité des comparaisons (Figures 19 et 20). Les projections UMAP montrent également que les zones enrichies en score LMA recouvrent en partie les populations annotées *AML-like* par notre approche (Figures 21 et 22). Ces résultats soutiennent la cohérence des annotations obtenues par la charge mutationnelle totale normalisée, tout en montrant que les deux approches ne capturent pas nécessairement exactement les mêmes sous-populations cellulaires.

Ainsi, les analyses transcriptomiques complémentaires renforcent l'interprétation des populations *AML-like* comme biologiquement cohérentes. L'association entre un signal mutationnel élevé, une augmentation du score IFN- γ après traitement et un enrichissement en signature GeneCards associée à la LMA suggère que la charge mutationnelle totale normalisée peut constituer un signal d'annotation pertinent pour identifier des populations compatibles avec une origine leucémique dans les données scRNA-seq.

5.4 Limitations et perspectives

Une première limitation concerne la nature des variants détectés à partir des données scRNA-seq, contrairement à un séquençage génomique. À la différence d'un séquençage génomique, le scRNA-seq ne permet d'observer que les variants présents dans des régions transcrites et suffisamment couvertes. Le signal obtenu peut donc être influencé par la profondeur de séquençage, le dropout allélique, la qualité d'alignement et les biais de capture des transcrits. De plus, les variants détectés ne correspondent pas nécessairement tous à des mutations somatiques, ils peuvent inclure des variants germinaux, des erreurs techniques ou des événements d'édition de l'ARN⁹⁵.

Cette limite est particulièrement importante dans le cadre de notre approche. En effet, plusieurs études ont montré qu'une partie des SNV détectés à partir de données de scRNA-seq peut en réalité correspondre à des événements d'édition de l'ARN plutôt qu'à de véritables mutations somatiques. Cette édition de l'ARN peut notamment se traduire par des substitutions de type A/G ou C/T dans les transcrits, notamment via l'activité d'enzymes de la famille ADAR, dont *ADAR1*, surexprimée dans la LMA¹²¹. Cela pourrait expliquer le faible recouvrement avec des bases de données de variants somatiques comme COSMIC. Toutefois, l'analyse de la distribution des types de substitutions dans nos données montre que les transitions A>G et C>T ne représentent que 29.3 % des SNV, tandis que les autres substitutions représentent 70.7 % (Annexe 4)¹²². De plus, les événements d'édition de l'ARN sont principalement localisés dans des régions non codantes, notamment les régions 3'UTR et les séquences répétées, qui sont exclues de notre *pipeline* d'analyse¹²². Cependant, le bruit résiduel associé à l'édition de l'ARN ne peut pas être totalement exclu.

Une autre limite concerne l'utilisation de la charge mutationnelle totale normalisée comme signal de classification. Cette approche repose sur un signal global qui peut entraîner une diminution de la spécificité. La normalisation par *nFeature_RNA* réduit l'effet de la complexité transcriptomique, mais ne corrige pas l'ensemble des biais liés à la couverture, à l'expression des gènes ou à la quantité de *reads*. De plus, l'annotation majoritaire au niveau des clusters Seurat peut masquer une hétérogénéité intra-cluster.

Une limite supplémentaire importante est liée à l'utilisation d'une classification non supervisée en deux groupes ($k = 2$). Cette stratégie suppose implicitement que chaque échantillon contient à la fois des cellules *AML-like* et des cellules *non-AML-like*. Dans un échantillon contenant très peu de cellules leucémiques, voire uniquement des cellules non leucémiques, la méthode pourrait identifier artificiellement un groupe à plus forte charge mutationnelle et l'annoter comme *AML-like*. Ce biais impose donc d'interpréter les annotations obtenues en tenant compte du contexte clinique, du point temporel analysé et des validations transcriptomiques complémentaires.

Enfin, la comparaison avec les annotations dérivées de Naldini *et al.* comporte des limites techniques. Les codes-barres des cellules annotées comme *AML-like* ou *non-AML-like* ne sont pas publiés sous une forme directement exploitable, et certaines étapes ou objets intermédiaires nécessaires à une reproduction exacte du *pipeline* ne sont pas disponibles. Des éléments comme les objets Seurat intermédiaires, les fichiers BAM ou les filtres appliqués peuvent influencer les annotations obtenues. Ces limites ne remettent pas en cause la comparaison, mais imposent de l'interpréter comme une comparaison indirecte plutôt que comme une reproduction stricte de l'analyse d'origine⁵⁶.

Malgré ces limites, l'approche développée répond à l'objectif initial de proposer une annotation moins dépendante d'un transcrit muté unique ou d'une altération spécifique. En appliquant une logique commune aux différents patients, la charge mutationnelle totale normalisée pourrait constituer une première étape vers un suivi clonal des populations *AML-like*, notamment pour étudier leur persistance après traitement et leur éventuelle réémergence à la rechute.

Une perspective importante de ce travail concerne l'utilisation de cette annotation *AML-like/non-AML-like* comme point de départ pour une stratégie de suivi clonal. L'objectif serait d'exploiter le profil mutationnel des cellules *AML-like* détectées après traitement afin de déterminer si elles dérivent de sous-populations déjà présentes au diagnostic. Cette approche permettrait d'identifier, parmi les cellules leucémiques initiales, celles qui possèdent une capacité à persister sous pression thérapeutique et à contribuer à la rechute.

À terme, la comparaison des caractéristiques mutationnelles et transcriptomiques de ces sous-populations pourrait aider à mieux définir les états cellulaires associés à la persistance leucémique et à proposer de nouvelles pistes pour cibler plus spécifiquement les cellules à l'origine de la rechute. Le principe général de cette perspective est illustré dans la Figure 23.

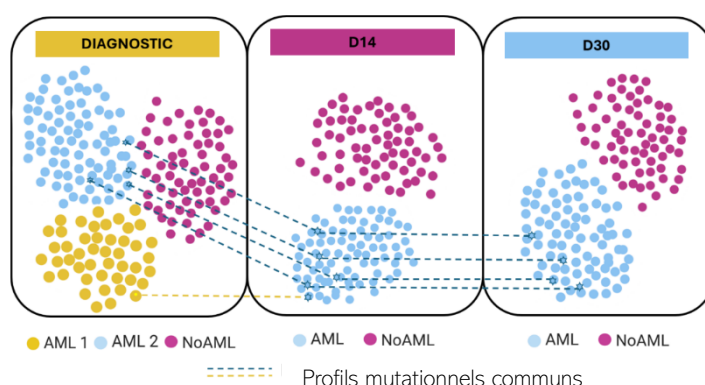


Figure 23. Principe du suivi clonal des cellules *AML-like* persistantes après traitement.

Ce schéma illustre l'utilisation potentielle des profils mutationnels pour relier les cellules *AML-like* détectées après traitement aux sous-populations leucémiques présentes au diagnostic, afin d'identifier celles qui pourraient contribuer à la persistance après traitement et à la rechute.

Schéma original réalisé dans le cadre de ce mémoire.

6. Bibliographie

- DiNardo, C. D., Erba, H. P., Freeman, S. D. & Wei, A. H. Acute myeloid leukaemia. *The Lancet* **401**, 2073–2086 (2023).
- Long, N. A., Golla, U., Sharma, A. & Claxton, D. F. Acute Myeloid Leukemia Stem Cells: Origin, Characteristics, and Clinical Implications. *Stem Cell Rev. Rep.* **18**, 1211–1226 (2022).
- Wachter, F. & Pikman, Y. Pathophysiology of Acute Myeloid Leukemia. *Acta Haematol.* **147**, 229–246 (2024).
- Döhner, H. et al. Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood* **140**, 1345–1377 (2022).
- Leucémie myéloïde aiguë (LMA) - Troubles du sang. *Manuels MSD pour le grand public* <https://www.msmanuals.com/fr/accueil/troubles-du-sang/leucémies/leucémie-myéloïde-aiguë-lma>.
- Volume-outcome relationship in cancer care: Acute Myeloid Leukemia – Literature overview | Belgian Cancer Registry. <https://kankerregister.org/fr/publicaties/volume-outcome-relationship-cancer-care-acute-myeloid-leukemia-literature-overview>.
- Acute Myeloid Leukemia Treatment (PDQ®) - NCI. <https://www.cancer.gov/types/leukemia/hp/adult-aml-treatment-pdq> (2026).
- Nilsson, C. et al. Characterization of therapy-related acute myeloid leukemia: increasing incidence and prognostic implications. *Haematologica* **108**, 1015–1025 (2023).
- El Chaer, F., Hourigan, C. S. & Zeidan, A. M. How I treat AML incorporating the updated classifications and guidelines. *Blood* **141**, 2813–2823 (2023).
- Vakiti, A., Reynolds, S. B. & Mewawalla, P. Acute Myeloid Leukemia. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2026).
- Bao, J., Freund, O., Sund, L. & Du, W. Inside the Battle Against Acute Myeloid Leukemia: Biology, Breakthroughs, and Hope. *Cells* **15**, 338 (2026).
- Arber, D. A. et al. International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. *Blood* **140**, 1200–1228 (2022).
- Chaudhary, S., Chaudhary, P., Ahmad, F. & Arora, N. Acute Myeloid Leukemia and Next-Generation Sequencing Panels for Diagnosis: A Comprehensive Review. *J. Pediatr. Hematol. Oncol.* **46**, 125–137 (2024).
- Gangat, N. & Dinardo, C. D. Newly diagnosed acute myeloid leukemia in unfit patients: 2026 treatment algorithms. *Blood Cancer J.* **15**, 139 (2025).
- Roman Diaz, J. L., Vazquez Martinez, M. & Khimani, F. New Approaches for the Treatment of AML beyond the 7+3 Regimen: Current Concepts and New Approaches. *Cancers* **16**, 677 (2024).
- Cortes, J. E. & Mehta, P. Determination of fitness and therapeutic options in older patients with acute myeloid leukemia. *Am. J. Hematol.* **96**, 493–507 (2021).
- Hieber, C. et al. Inhibitors of the tyrosine kinases FMS-like tyrosine kinase-3 and WEE1 induce apoptosis and DNA damage synergistically in acute myeloid leukemia cells. *Biomed. Pharmacother.* **177**, 117076 (2024).
- DiNardo, C. D. et al. Azacitidine and Venetoclax in Previously Untreated Acute Myeloid Leukemia. *N. Engl. J. Med.* **383**, 617–629 (2020).
- Jimenez-Chillon, C., Dillon, R. & Russell, N. Optimal Post-Remission Consolidation Therapy in Patients with AML. *Acta Haematol.* **147**, 148–159 (2024).
- Döhner, H. et al. Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood* **140**, 1345–1377 (2022).
- Thol, F. & Ganser, A. Treatment of Relapsed Acute Myeloid Leukemia. *Curr. Treat. Options Oncol.* **21**, 66 (2020).
- Biology and management of relapsed acute myeloid leukaemia. https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1365-2141.2004.05318.x?saml_referrer.
- Shallis, R. M., Pollyea, D. A. & Zeidan, A. M. The complete story of less than complete responses: The evolution and application of acute myeloid leukemia clinical responses. *Blood Rev.* **48**, 100806 (2021).
- Pabon, C. M. et al. Morphologic leukemia-free state in acute myeloid leukemia is sufficient for successful allogeneic hematopoietic stem cell transplant. *Blood Cancer J.* **11**, 92 (2021).
- Shimony, S., Stahl, M. & Stone, R. M. Acute Myeloid Leukemia: 2025 Update on Diagnosis, Risk-Stratification, and Management. *Am. J. Hematol.* **100**, 860–891 (2025).
- Badaoui, B., Sloma, I. & Tarfi, S. Suivi de la maladie résiduelle dans les leucémies aiguës myéloïdes. *Rev. Francoph. Lab.* **2023**, 64–74 (2023).
- Paras, G. et al. Conditioning intensity and peritransplant flow cytometric MRD dynamics in adult AML. *Blood* **139**, 1694–1706 (2022).
- Jongen-Lavrencic, M. et al. Molecular Minimal Residual Disease in Acute Myeloid Leukemia. *N. Engl. J. Med.* **378**, 1189–1199 (2018).
- Short, N. J. et al. Association of Measurable Residual Disease With Survival Outcomes in Patients With Acute

Myeloid Leukemia: A Systematic Review and Meta-analysis. *JAMA Oncol.* **6**, 1890–1899 (2020).

30. van der Maas, N. G. *et al.* A revised prognostic model for patients with acute myeloid leukemia and first relapse. *Blood Adv.* **9**, 3853–3864 (2025).

31. Ganzel, C. *et al.* Very poor long-term survival in past and more recent studies for relapsed AML patients: The ECOG-ACRIN experience. *Am. J. Hematol.* **93**, 1074–1081 (2018).

32. Bataller, A. *et al.* Outcomes and genetic dynamics of acute myeloid leukemia at first relapse. *Haematologica* **109**, 3543–3556 (2024).

33. Evaluation of Trends and Prognosis Over Time in Patients with AML Relapsing After Allogeneic Hematopoietic Cell Transplant Reveals Improved Survival for Young Patients in Recent Years | Clinical Cancer Research | American Association for Cancer Research. <https://aacrjournals.org/clincancerres/article/26/24/6475/82911/Evaluation-of-Trends-and-Prognosis-Over-Time-in>.

34. Shahswar, R. *et al.* A novel prognostic risk model for patients with refractory/relapsed acute myeloid leukemia receiving venetoclax plus hypomethylating agents. *Leukemia* **39**, 614–622 (2025).

35. Arnone, M. *et al.* Acute Myeloid Leukemia Stem Cells: The Challenges of Phenotypic Heterogeneity. *Cancers* **12**, 3742 (2020).

36. Minimal Residual Disease in Acute Myeloid Leukemia: Old and New Concepts. <https://www.mdpi.com/1422-0067/25/4/2150>.

37. Lim, J. J. *et al.* Time independent factors that predict relapse in adults with acute myeloid leukemia. *Blood Cancer J.* **14**, 5 (2024).

38. Pinho, S. & Frenette, P. S. Haematopoietic stem cell activity and interactions with the niche. *Nat. Rev. Mol. Cell Biol.* **20**, 303–320 (2019).

39. Wei, Q. & Frenette, P. S. Niches for Hematopoietic Stem Cells and Their Progeny. *Immunity* **48**, 632–648 (2018).

40. Ally, F. & Chen, X. Acute Myeloid Leukemia: Diagnosis and Evaluation by Flow Cytometry. *Cancers* **16**, 3855 (2024).

41. Mann, Z., Sengar, M., Verma, Y. K., Rajalingam, R. & Raghav, P. K. Hematopoietic Stem Cell Factors: Their Functional Role in Self-Renewal and Clinical Aspects. *Front. Cell Dev. Biol.* **10**, (2022).

42. Karamitros, D. *et al.* Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. *Nat. Immunol.* **19**, 85–97 (2018).

43. Boulais, P. E. & Frenette, P. S. Making sense of hematopoietic stem cell niches. *Blood* **125**, 2621–2629 (2015).

44. Méndez-Ferrer, S. *et al.* Bone marrow niches in haematological malignancies. *Nat. Rev. Cancer* **20**, 285–298 (2020).

45. Duarte, D. *et al.* Inhibition of Endosteal Vascular Niche Remodeling Rescues Hematopoietic Stem Cell Loss in AML. *Cell Stem Cell* **22**, 64–77.e6 (2018).

46. Asada, S. & Kitamura, T. Clonal hematopoiesis and associated diseases: A review of recent findings. *Cancer Sci.* **112**, 3962–3971 (2021).

47. Chow, R. D. *et al.* Persistent postremission clonal hematopoiesis shapes the relapse trajectories of acute myeloid leukemia. *Blood Adv.* **9**, 1888–1899 (2025).

48. Gaulin, C., Kelemen, K. & Arana Yi, C. Molecular Pathways in Clonal Hematopoiesis: From the Acquisition of Somatic Mutations to Transformation into Hematologic Neoplasm. *Life* **12**, 1135 (2022).

49. Corces-Zimmerman, M. R. & Majeti, R. Pre-leukemic evolution of hematopoietic stem cells: the importance of early mutations in leukemogenesis. *Leukemia* **28**, 2276–2282 (2014).

50. Hemmati, S., Haque, T. & Gritsman, K. Inflammatory Signaling Pathways in Preleukemic and Leukemic Stem Cells. *Front. Oncol.* **7**, (2017).

51. Naji, N. S., Sathish, M. & Karantanos, T. Inflammation and Related Signaling Pathways in Acute Myeloid Leukemia. *Cancers* **16**, 3974 (2024).

52. Hematopoietic and leukemic stem cells homeostasis: the role of bone marrow niche - PMC. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11438561/>.

53. Sharma, N. & Liesveld, J. L. NPM 1 Mutations in AML- The Landscape in 2023. *Cancers* **15**, 1177 (2023).

54. Eisfeld, A.-K. *et al.* Mutational Landscape and Gene Expression Patterns in Adult Acute Myeloid Leukemias with Monosomy 7 as a Sole Abnormality. *Cancer Res.* **77**, 207–218 (2017).

55. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).

56. Naldini, M. M. *et al.* Longitudinal single-cell profiling of chemotherapy response in acute myeloid leukemia. *Nat. Commun.* **14**, 1285 (2023).

57. Falini, B. NPM1-mutated acute myeloid leukemia: New pathogenetic and therapeutic insights and open questions. *Am. J. Hematol.* **98**, 1452–1464 (2023).

58. Kishtagari, A. & Levine, R. L. The Role of Somatic Mutations in Acute Myeloid Leukemia Pathogenesis. *Cold Spring Harb. Perspect. Med.* **11**, a034975 (2021).

59. Hindley, A., Catherwood, M. A., McMullin, M. F. & Mills, K. I. Significance of NPM1 Gene Mutations in AML. *Int. J. Mol. Sci.* **22**, 10040 (2021).

60. Patel, S. S. NPM1-Mutated Acute Myeloid Leukemia: Recent Developments and Open Questions. *Pathobiology* **91**, 18–29 (2023).

61. Lindsley, R. C. *et al.* Acute myeloid leukemia ontogeny is defined by distinct somatic mutations. *Blood* **125**, 1367–1376 (2015).

62. Klcó, J. M. & Mullighan, C. G. Advances in germline predisposition to acute leukaemias and myeloid neoplasms. *Nat. Rev. Cancer* **21**, 122–137 (2021).
63. Smith, S. C. & Zhang, L. Germline and Somatic Changes Associated with the Development of Inherited and De Novo Pediatric Acute Myeloid Leukemia. *Genes* **16**, 798 (2025).
64. Morgenstern, Y. *et al.* Acute myeloid leukemia drug-tolerant persister cells survive chemotherapy by transiently increasing plasma membrane rigidity, that also increases their sensitivity to immune cell killing. *Haematologica* **110**, 893–903 (2025).
65. A cell initiating human acute myeloid leukaemia after transplantation into SCID mice | Nature. <https://www.nature.com/articles/367645a0>.
66. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* **3**, 730–737 (1997).
67. Thomas, D. & Majeti, R. Biology and relevance of human acute myeloid leukemia stem cells. *Blood* **129**, 1577–1585 (2017).
68. Identification of Chemotherapy-Induced Leukemic-Regenerating Cells Reveals a Transient Vulnerability of Human AML Recurrence: Cancer Cell. [https://www.cell.com/cancer-cell/fulltext/S1535-6108\(18\)30367-2?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1535610818303672%3Fshowall%3Dtrue](https://www.cell.com/cancer-cell/fulltext/S1535-6108(18)30367-2?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1535610818303672%3Fshowall%3Dtrue).
69. Duy, C. *et al.* Chemotherapy Induces Senescence-Like Resilient Cells Capable of Initiating AML Recurrence. *Cancer Discov.* **11**, 1542–1561 (2021).
70. He, J. *et al.* Drug tolerant persister cell plasticity in cancer: A revolutionary strategy for more effective anticancer therapies. *Signal Transduct. Target. Ther.* **9**, 209 (2024).
71. Li, M., Wang, X., He, W. & Zhou, H. Drug-tolerant persister cells in acute myeloid leukemia: pressing challenge and promising new strategies for treatment. *Front. Med.* **12**, (2025).
72. Mikubo, M., Inoue, Y., Liu, G. & Tsao, M.-S. Mechanism of Drug Tolerant Persister Cancer Cells: The Landscape and Clinical Implication for Therapy. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **16**, 1798–1809 (2021).
73. Dhanyamraju, P. K., Schell, T. D., Amin, S. & Robertson, G. P. Drug-Tolerant Persister Cells in Cancer Therapy Resistance. *Cancer Res.* **82**, 2503–2514 (2022).
74. Miller, D. *et al.* Heterogeneity in leukemia cells that escape drug-induced senescence-like state. *Cell Death Dis.* **14**, 503 (2023).
75. Chen, H.-L. & Jin, W.-L. Diapause-like Drug-Tolerant Persister State: The Key to Nirvana Rebirth. *Medicina (Mex.)* **60**, 228 (2024).
76. Cancer drug-tolerant persister cells: from biological questions to clinical opportunities | Nature Reviews Cancer. <https://www.nature.com/articles/s41568-024-00737-z>.
77. van Gils, N., Denkers, F. & Smit, L. Escape From Treatment; the Different Faces of Leukemic Stem Cells and Therapy Resistance in Acute Myeloid Leukemia. *Front. Oncol.* **11**, 659253 (2021).
78. Huang, F. Y., Trumpp, A. & Stelmach, P. Resolving leukemic stem cell heterogeneity and plasticity with single-cell multiomics. *Semin. Hematol.* **62**, 218–225 (2025).
79. Petti, A. A. *et al.* A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat. Commun.* **10**, 3660 (2019).
80. Thomas, D. & Majeti, R. Biology and relevance of human acute myeloid leukemia stem cells. *Blood* **129**, 1577–1585 (2017).
81. Acute Myeloid Leukemia Stem Cells in Minimal/Measurable Residual Disease Detection. <https://www.mdpi.com/2072-6694/15/10/2866>.
82. Leukemic stem cells and therapy resistance in acute myeloid leukemia | Haematologica. <https://haematologica.org/article/view/10948>.
83. Li, K. *et al.* Single-cell analysis reveals the chemotherapy-induced cellular reprogramming and novel therapeutic targets in relapsed/refractory acute myeloid leukemia. *Leukemia* **37**, 308–325 (2023).
84. Zhang, Y. *et al.* Single-cell transcriptomics reveals multiple chemoresistant properties in leukemic stem and progenitor cells in pediatric AML. *Genome Biol.* **24**, 199 (2023).
85. Farge, T. *et al.* CD36 Drives Metastasis and Relapse in Acute Myeloid Leukemia. *Cancer Res.* **83**, 2824–2838 (2023).
86. Thakral, D., Gupta, R. & Khan, A. Leukemic stem cell signatures in Acute myeloid leukemia- targeting the Guardians with novel approaches. *Stem Cell Rev. Rep.* **18**, 1756–1773 (2022).
87. Desai, R. H., Zandvakili, N. & Bohlander, S. K. Dissecting the Genetic and Non-Genetic Heterogeneity of Acute Myeloid Leukemia Using Next-Generation Sequencing and In Vivo Models. *Cancers* **14**, 2182 (2022).
88. Zhou, J. & Chng, W.-J. Unveiling novel insights in acute myeloid leukemia through single-cell RNA sequencing. *Front. Oncol.* **14**, 1365330 (2024).
89. Galen, P. van *et al.* Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265–1281.e24 (2019).
90. Lambo, S. *et al.* A longitudinal single-cell atlas of treatment response in pediatric AML. *Cancer Cell* **41**, 2117–2135.e12 (2023).
91. Jovic, D. *et al.* Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.* **12**, e694 (2022).

92. aml related genes - GeneCards Search Results. <https://www.genecards.org/Search/Keyword?startPage=0&queryString=aml&pageSize=-1>.
93. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
94. Huang, X. & Huang, Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* **37**, 4569–4571 (2021).
95. Marot-Lassauzaie, V. *et al.* Identifying cancer cells from calling single-nucleotide variants in scRNA-seq data. *Bioinformatics* **40**, btac512 (2024).
96. Homo sapiens genome assembly GRCh38.p14. *NCBI* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/.
97. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
98. Korthauer, K. *et al.* A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.* **20**, 118 (2019).
99. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
100. Enge, M. *et al.* Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* **171**, 321–330.e14 (2017).
101. Gasper, W., Rossi, F., Ligorio, M. & Gherzi, D. Variant calling enhances the identification of cancer cells in single-cell RNA sequencing data. *PLoS Comput. Biol.* **18**, e1010576 (2022).
102. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
103. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
104. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
105. Kreimer, A. & Pe'er, I. Variants in exons and in transcription factors affect gene expression in trans. *Genome Biol.* **14**, R71 (2013).
106. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
107. Mayr, C. Regulation by 3'-Untranslated Regions. *Annu. Rev. Genet.* **51**, 171–194 (2017).
108. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
109. Rose, A. B. Introns as Gene Regulators: A Brick on the Accelerator. *Front. Genet.* **9**, 672 (2019).
110. Getting Started with Seurat • Seurat. https://satijalab.org/seurat/articles/get_started_v5_new.
111. Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **31**, 651–666 (2010).
112. Zhang, X. *et al.* Single-cell RNA sequencing and spatial transcriptomics reveal the heterogeneity and intercellular communication of cancer-associated fibroblasts in gastric cancer. *J. Transl. Med.* **23**, 344 (2025).
113. Ferri, C., Hernández-Orallo, J. & Flach, P. Setting decision thresholds when operating conditions are uncertain. *Data Min. Knowl. Discov.* **33**, 805–847 (2019).
114. Velten, L. *et al.* Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics. *Nat. Commun.* **12**, 1366 (2021).
115. Hiltemann, S., Jenster, G., Trapman, J., van der Spek, P. & Stubbs, A. Discriminating somatic and germline mutations in tumor DNA samples without matching normals. *Genome Res.* **25**, 1382–1390 (2015).
116. Kövy, P. *et al.* Nucleophosmin1 and isocitrate dehydrogenase 1 and 2 as measurable residual disease markers in acute myeloid leukemia. *PLoS ONE* **16**, e0253386 (2021).
117. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukemia revealed by whole genome sequencing. *Nature* **481**, 506–510 (2012).
118. Dumitrascu, B., Villar, S., Mixon, D. G. & Engelhardt, B. E. Optimal marker gene selection for cell type discrimination in single cell analyses. *Nat. Commun.* **12**, 1186 (2021).
119. Wang, B. *et al.* Comprehensive characterization of IFN γ signaling in acute myeloid leukemia reveals prognostic and therapeutic strategies. *Nat. Commun.* **15**, 1821 (2024).
120. Vadakekolathu, J. *et al.* Immune landscapes predict chemotherapy resistance and immunotherapy response in acute myeloid leukemia. *Sci. Transl. Med.* **12**, eaaz0463 (2020).
121. Shi, Z. *et al.* ADAR1 is required for acute myeloid leukemia cell survival by modulating post-transcriptional Wnt signaling through impairing miRNA biogenesis. *Leukemia* **39**, 599–613 (2025).
122. Ding, J., Lin, C. & Bar-Joseph, Z. Cell lineage inference from SNP and scRNA-Seq data. *Nucleic Acids Res.* **47**, e56 (2019).

Annexe

Lignée	Moyenne CT	Moyenne ARA	Diminution approximative
HL60	1 719	504	-71 %
KG1	2 047	1 334	-34 %
ML2	795	419	-47 %
MOLM13	1 365	685	-50 %
OCIAML3	1 092	520	-52 %
SKM1	1 369	585	-57 %
THPA	780	409	-48 %

Annexe 1. Expression de *NPM1* dans des lignées de LMA traitées par cytarabine.

Les valeurs correspondent à l'expression de *NPM1* mesurée dans différentes lignées cellulaires de LMA en condition contrôle (CT) et après traitement par cytarabine (ARA). Dans l'ensemble des lignées présentées, l'expression de *NPM1* est diminuée après traitement par cytarabine par rapport à la condition contrôle.

Patient <i>NPM1</i>	Timepoint	AML-like Naldini et al. avant annotation majoritaire	Non-AML-like Naldini et al. avant annotation majoritaire
PT01	DX	889	858
	D14	18	107
	D30	22	1 390
PT07	DX	72	632
	D14	0	238
PT08	DX	1 027	3 377
	D14	1	59
	D30	0	2 516

Annexe 2A. Distribution des annotations dérivées de Naldini et al. avant attribution majoritaire par cluster chez les patients *NPM1^{mut}*.

Ce tableau présente le nombre de cellules annotées AML-like et non-AML-like selon l'approche dérivée de Naldini et al., avant l'attribution majoritaire au niveau des clusters Seurat.

Patient del(7)	Timepoint	AML-like Naldini et al. avant annotation majoritaire	Non-AML-like Naldini et al. avant annotation majoritaire
PT11	DX	1 388	630
	D30	1 642	3 470
PT17	DX	1 247	96
	D14	2 189	102
	D30	1 010	5 507
PT18	DX	2 084	1 577
	D14	1 503	445

Annexe 2B. Distribution des annotations dérivées de Naldini et al. avant attribution majoritaire par cluster chez les patients *del(7)*.

Ce tableau présente le nombre de cellules annotées AML-like, non-AML-like selon l'approche dérivée de Naldini et al., avant l'attribution majoritaire au niveau des clusters Seurat.

Patient <i>NPM1</i>	Timepoint	<i>AML-like k-means</i> avant annotation majoritaire	<i>Non-AML-like k-means</i> avant annotation majoritaire	<i>Uncertain k-means</i>
PT01	DX	1 370	223	152
	D14	40	69	16
	D30	825	413	177
PT07	DX	262	382	60
	D14	101	119	24
PT08	DX	1 759	2 205	440
	D14	34	18	8
	D30	1 696	629	191

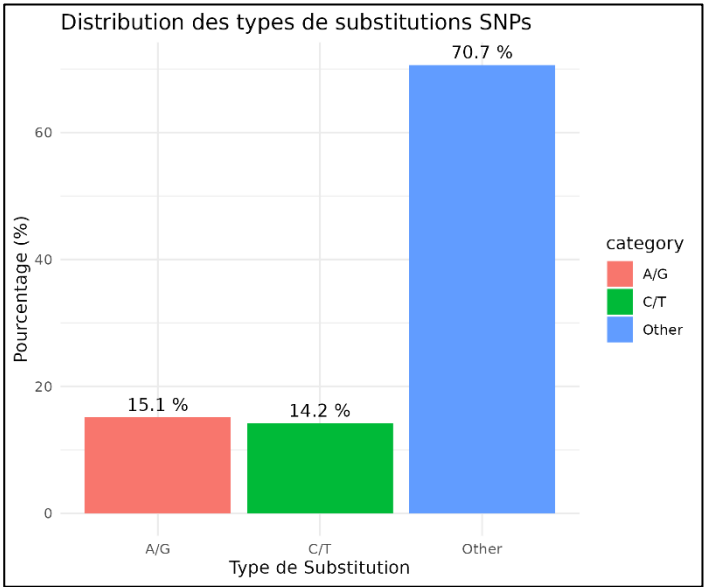
Annexe 3A. Distribution des annotations dérivées de l'approche k-means avant attribution majoritaire par cluster chez les patients *NPM1^{mut}*.

Ce tableau présente le nombre de cellules annotées *AML-like*, *non-AML-like* et *uncertain* selon l'approche k-means, avant l'attribution majoritaire au niveau des clusters Seurat.

Patient <i>del(7)</i>	Timepoint	<i>AML-like k-means</i> avant annotation majoritaire	<i>Non-AML-like k-means</i> avant annotation majoritaire	<i>Uncertain k-means</i>
PT11	DX	825	410	108
	D30	3 983	2 209	325
PT17	DX	1 323	558	137
	D14	1 761	316	214
	D30	2 254	3 720	543
PT18	DX	2 144	1 213	304
	D14	1 168	619	161

Annexe 3B. Distribution des annotations dérivées de l'approche k-means avant attribution majoritaire par cluster chez les patients *del(7)*.

Ce tableau présente le nombre de cellules annotées *AML-like*, *non-AML-like* et *uncertain* selon l'approche k-means, avant l'attribution majoritaire au niveau des clusters Seurat.



Annexe 4. Distribution des types de substitutions des SNV détectés.

Le graphique présente la proportion des substitutions A>G, C>T et des autres substitutions parmi les SNV détectés dans les données analysées. Les substitutions A>G et C>T, compatibles avec des événements d'édition de l'ARN, représentent respectivement 15.1 % et 14.2 % des SNV, soit 29.3 % au total. Les autres types de substitutions représentent 70.7 % des SNV détectés.