Approche non supervisée pour le nettoyage de données de cytométrie en flux

Stage de recherche (10 semaines)

Réalisé par **Félicia DOSSOU**
Etudiante en Master 1 Cursus Master Ingénierie de la Statistique et Informatique (CMI ISI), Université de Bordeaux

Supervisé par **Christèle ETCHEGARAY**
Chargée de Recherche Inria, Institut de Mathématiques de Bordeaux (IMB)

En collaboration avec **Aguirre MIMOUN**
Médecin biologiste et responsable du laboratoire de cytométrie en flux au CHU de Bordeaux

Stage hébergé par **l'Institut de Mathématiques de Bordeaux**
et financé par **le Réseau Thématique Math Bio Santé**

Résumé — Ce rapport présente le développement d'une méthode automatique de nettoyage des données de cytométrie en flux, basée sur des techniques de clustering non supervisé. L'étude s'appuie sur des échantillons de patients atteints de leucémie aiguë myéloblastique issus du registre régional DatAML.

Table des matières

Remerciements

Je tiens à remercier toutes les personnes ayant contribué à la réussite de mon stage et à l'élaboration de ce rapport.

Tout d'abord, je remercie chaleureusement mon encadrante Mme Christèle Etchegaray, chargée de recherche à l'Inria Bordeaux, pour son accueil au sein de l'équipe-projet MONC, ses nombreux suivis tout au long du stage, ainsi que pour son expertise précieuse et ses conseils éclairés, qui ont guidé l'ensemble de ce travail.

Je remercie également M. Aguirre Mimoun, médecin biologiste et responsable du secteur de cytométrie en flux au CHU de Bordeaux, pour ses retours réguliers et ses remarques pointues qui ont permis de faire avancer et approfondir l'analyse.

Un grand merci à Jonathan Legrand pour sa disponibilité, sa patience et ses réponses toujours claires à mes nombreuses questions techniques.

Je remercie aussi l'ensemble de l'équipe MONC pour son accueil chaleureux, sa bienveillance et l'environnement de travail motivant dont j'ai bénéficié durant ces dix semaines.

Enfin, je souhaite remercier toutes les personnes, de près ou de loin, qui m'ont accompagnée, soutenue ou encouragée durant ce stage.

1 Introduction

1.1 Leucémie aiguë myéloblastique

La leucémie aiguë myéloblastique (LAM) est une forme agressive de cancer du sang, caractérisée par la prolifération anarchique de cellules immatures appelées *blastes* dans la moelle osseuse. Pour mieux comprendre ce phénomène, il est utile de revenir brièvement sur le processus physiologique de formation des cellules sanguines, appelé hématopoïèse.

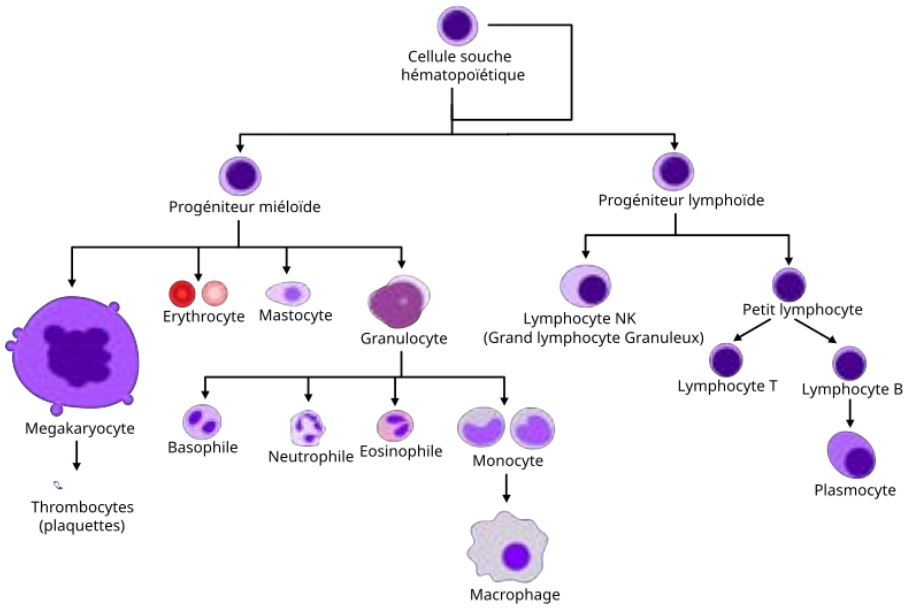

FIGURE 1 – Modèle simplifié de l'hématopoïèse. [5]

Le processus d'hématopoïèse est un mécanisme de fabrication des cellules sanguines au cours duquel les cellules souches hématopoïétiques situées dans la moelle osseuse se différencient progressivement en différentes lignées cellulaires spécialisées. [1]

Ce processus donne naissance aux principaux types de cellules sanguines : globules rouges, plaquettes et globules blancs (lymphocytes, monocytes, granulocytes). En temps normal, la transformation des cellules souches en cellules sanguines matures est un processus bien contrôlé, où chaque type cellulaire est produit en quantité adaptée. Cette régulation empêche l'accumulation de cellules immatures dans le sang ou la moelle osseuse. Dans le cas de la leucémie aiguë myéloïde (LAM), cette maturation est interrompue au stade des myéloblastes (entre le progéniteur myéloide et le granulocyte sur le schéma), entraînant leur accumulation anormale dans la moelle osseuse et le sang.

Le traitement repose principalement sur la chimiothérapie, qui permet d'obtenir une rémission chez environ 80 % des patients. Cependant, toutes les leucémies ne réagissent pas de la même manière aux traitements, car il existe de nombreux sous-types moléculaires liés à des combinaisons spécifiques de mutations génétiques. [3]

Il existe plusieurs approches thérapeutiques, comme le ciblage d'une mutation, l'administration d'une chimiothérapie moins intensive, la mise en place de soins palliatifs ou encore la cytométrie en flux. Le choix du traitement repose sur l'évaluation réalisée à partir des données cliniques recueillies au moment du diagnostic.

Le diagnostic de leucémie débute par une ponction de moelle osseuse. L'échantillon est d'abord observé au microscope pour évaluer la proportion de blastes. Si elle dépasse 20 %, un diagnostic de leucémie est posé.

1.2 La cytométrie en flux

La cytométrie en flux joue un rôle essentiel dans le diagnostic, car elle fournit des informations biologiques précises sur les cellules leucémiques.

Il s'agit d'une technique d'analyse multiparamétrique qui permet de mesurer simultanément plusieurs caractéristiques physiques et biologiques de millions de cellules en suspension.

Lors d'une analyse, les cellules sont introduites dans un flux hydrodynamique qui les aligne une à une. Elles passent alors devant un ou plusieurs lasers, ce qui provoque la diffraction de la lumière. (voir Figure 2)

Deux composantes principales de diffusion sont mesurées :

- le **Forward Scatter** (FS), proportionnel à la taille de la cellule ;
- le **Side Scatter** (SS), lié à sa complexité interne (granularité).

(voir Figure 3)
Ces signaux permettent une première distinction des populations cellulaires sur la base de leurs propriétés morphologiques.

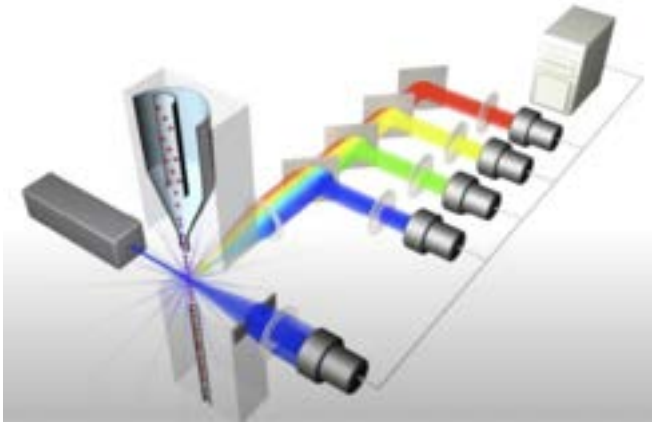

FIGURE 2 – Schéma d'un dispositif de cytométrie en flux [1]

En complément, les cellules peuvent être marquées par des anticorps conjugués à des fluorochromes. Lorsqu'un laser excite ces fluorochromes, ceux-ci réémettent une lumière de longueur d'onde spécifique. Chaque marqueur peut ainsi être détecté via un canal fluorescent, et plusieurs lasers permettent d'exciter simultanément des fluorochromes différents. Cela autorise la mesure conjointe de nombreux paramètres biologiques.

Tous les signaux lumineux sont ensuite dirigés vers différents photodétecteurs, qui convertissent la lumière reçue en signaux électriques. Ces impulsions sont numérisées et associées à chaque cellule, produisant un vecteur de mesures par événement.

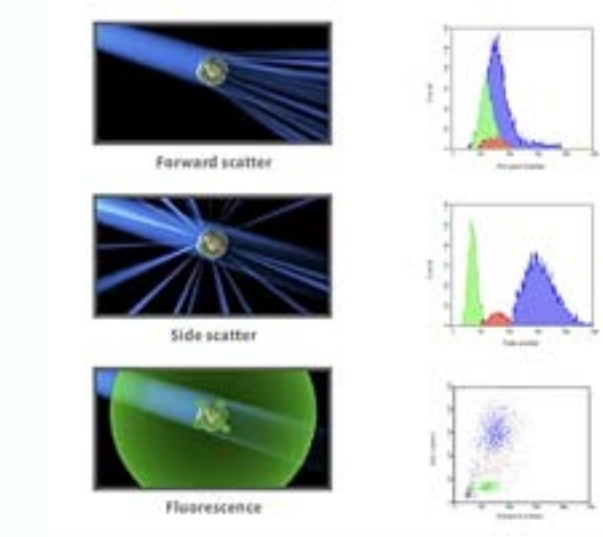

FIGURE 3 – Répartition des prinipaux signaux lumineux détectés [1]

Les signaux peuvent être quantifiés selon différentes méthodes d'intégration du pic lumineux : sa hauteur (FF-H), son aire (FS-A) ou sa largeur (FS-W).

Ces valeurs, une fois converties numériquement, forment une matrice de données où chaque ligne correspond à une cellule, et chaque colonne à un paramètre mesuré (canal). Ces données sont sauvegardées dans des fichiers standardisés au format `.fcs`, exploitables par des logiciels spécialisés ou des scripts d'analyse. Elles peuvent ensuite être visualisées sous forme de graphes univariés (histogrammes Figure **??**) ou bivariés (scatter plots Figure **??**).

À partir de ces visualisations, les cliniciens peuvent identifier différentes sous-populations cellulaires en se basant sur la taille des cellules, leur granularité ou l'intensité de fluorescence.

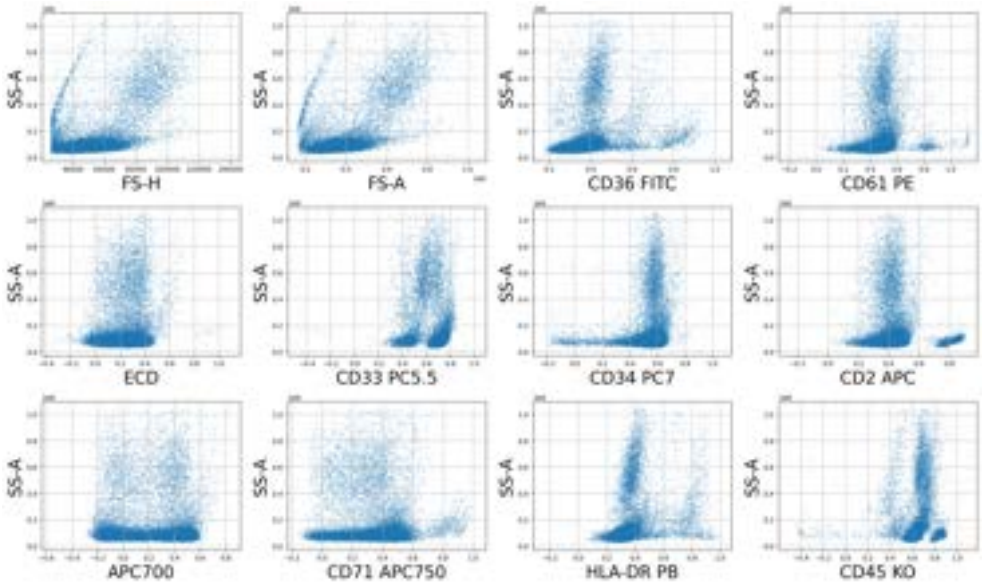

FIGURE 4 – Scatter plots des cellules sur différents marqueurs du patient 199 Tube C

Cependant, les données de cytométrie sont sensibles à de nombreux biais : autofluorescence, bruit biologique, interférences spectrales, artefacts liés aux cellules mortes ou aux débris, ou encore variabilité inter-individuelle. Il est donc crucial d'appliquer des étapes de prétraitement.

1.3 Preprocessing et nettoyage

Le prétraitement des données est essentiel pour :

1. Éliminer les artéfacts techniques tels que les événements marginaux (valeurs aberrantes en SS et en FS) et les doublets, c'est-à-dire deux cellules qui traversent en même temps le faisceau lumineux et dont le signal est interprété comme celui d'une seule cellule. Dans notre cas, nous supposons qu'ils sont absents ;
2. Appliquer la compensation des chevauchements spectraux via une matrice de compensation pré-enregistrée ;
3. Transformer les données (logicle, arcsinh) pour faciliter leur interprétation ;
4. Sélectionner les cellules viables en excluant les débris et cellules mortes.

Les trois premières étapes peuvent être réalisées automatiquement, bien que la compensation soit parfois corrigée par les cytométristes. En revanche, la quatrième — la sélection des cellules viables — repose encore sur une expertise manuelle des cytométristes, ce qui rend le processus coûteux en temps et peu reproductible.

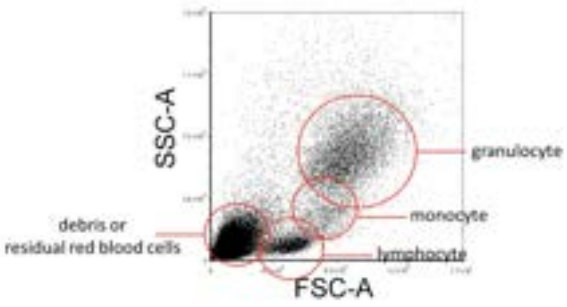

FIGURE 5 – Exemple d'identification manuelle de débris [8]

Débris et cellules mortes

En cytométrie en flux, les débris cellulaires et les cellules mortes ne présentent pas toujours des caractéristiques fixes et bien définies. Leur identification automatique est complexe, car leurs propriétés varient selon l'échantillon, les conditions de préparation, ou encore le type de cellules analysées.

Cependant, certaines tendances générales sont souvent observées :
— **Faible FS** : les débris sont généralement plus petits que les cellules intactes, ce qui se traduit par un signal FS réduit.
— **SS variable** : leur granularité peut être faible ou modérée selon leur état de dégradation, ce qui rend le signal SS peu fiable seul.
— **Fluorescence non spécifique** : la fixation des anticorps peut être imparfaite, ce qui génère du bruit ou des signaux aberrants sur plusieurs canaux. Le niveau de fluorescence est souvent plus faible chez les débris, mais peut également apparaître normal chez des cellules mortes.

Ces éléments constituent des indices utiles, mais ne peuvent pas être considérés comme des critères rigoureux. De plus, le risque d'éliminer des cellules importantes, comme les blastes, est élevé, ce qui rend l'expertise des cytométristes indispensable. Une approche prudente, fondée sur la combinaison de plusieurs critères, est donc nécessaire afin d'exclure les artefacts sans compromettre l'analyse des cellules viables.

Typiquement, les cytométristes appliquent un *gating* qui exclut les événements présentant un FS faible et un SS moyen à élevé. Cette stratégie permet d'éliminer les objets de petite taille mais granuleux, caractéristiques des débris cellulaires. La Figure 6 illustre l'identification de cellules CD133+ ; on voit que l'étape A correspond à l'exclusion des débris et consiste à exclure la zone de FS bas dans le graphe FS-SS. On observe néanmoins en D la persistance de certaines cellules mortes, DAPI étant un marqueur de viabilité des cellules qui n'est en général pas disponible en routine.

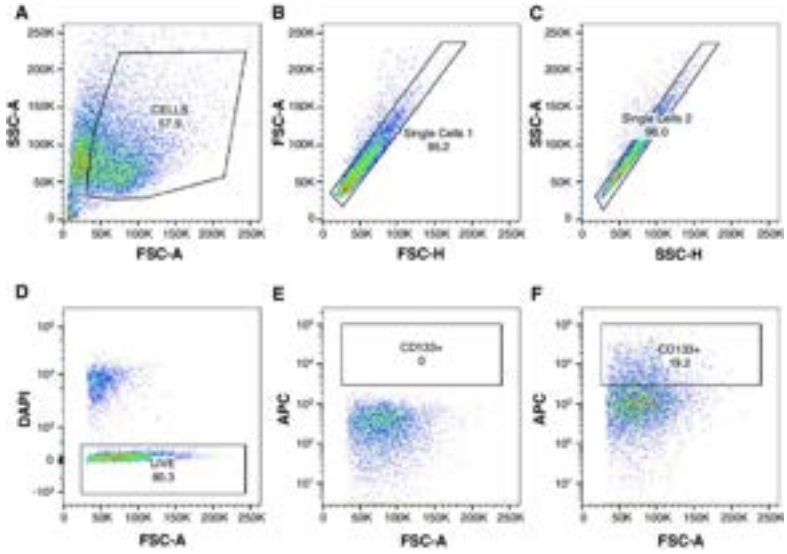

FIGURE 6 – Stratégie de gating séquentiel pour l'exclusion des artefacts et l'enrichissement en cellules CD133+ [9]

La présence de débris ou de cellules mortes peut fortement biaiser les résultats, en particulier lors de l'étude de populations rares ou de l'évaluation précise de l'expression de marqueurs de surface. Leur élimination constitue donc une étape essentielle du prétraitement des données en cytométrie. Toutefois, cette procédure repose largement sur l'expertise de l'opérateur : elle est peu reproductible et requiert un temps d'analyse non négligeable.

Comme l'illustre la Figure 7, les cytométristes s'appuient principalement sur les représentations **FS-A vs SS-A** et **CD45 KO vs SS-A**. Dans la suite, nous nous concentrerons également sur ces deux graphes, car ils permettent une détection plus intuitive des débris : sur le graphe FS-SS, ceux-ci apparaissent sous la forme d'un croissant situé sur le côté gauche, tandis que sur le graphe CD45-SS, les cellules de gauche quasi-invisibles à l'oeil nu, se séparent de la masse principale.

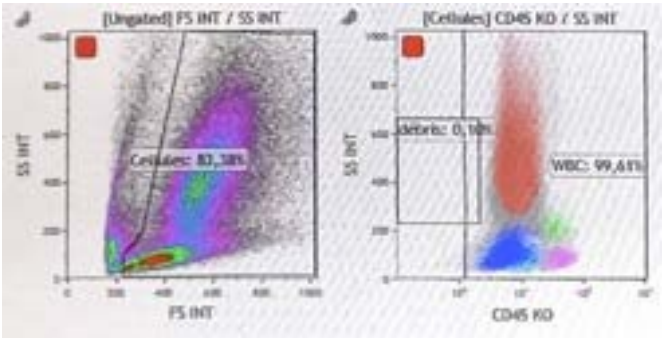

FIGURE 7 – Exemple de gating des cellules débris réalisé à la main par Agguire Mimoun sur les graphes FS-SS et CD45-SS.

Il convient toutefois de rester attentif à la présence des blastes, qu'il est essentiel de préserver en raison de leur importance biologique. Sur le graphe FS-SS, ils correspondent aux points rouges situés en bas à gauche. Leur localisation proche des débris explique la difficulté portée par le nettoyage des données.

Au laboratoire du CHU, plusieurs dizaines voire une centaine d'échantillons peuvent être traités chaque jour. Même si chaque nettoyage est relativement rapide, le volume global représente une charge importante. C'est pourquoi l'objectif de ce stage est de développer une méthode automatique permettant de détecter les débris.

1.4 Objectifs du stage

Ce projet, réalisé en collaboraton avec CHU de Bordeaux à partir des données du registre *DatAML*, a pour objectif de développer un pipeline automatisé de détection des débris et des cellules mortes, en s'appuyant sur des méthodes non supervisées.

Deux grandes sous-problématiques structurent ce travail :

1. Définir une méthode de clustering qui segmente correctement les populations cellulaires, en assurant une bonne séparation entre les cellules viables et les éléments à exclure (débris, cellules mortes) ;
2. Identifier automatiquement, au sein de ce découpage, les clusters à supprimer, sans intervention manuelle, et malgré la variabilité inter-patients.

1.4.1 Données disponibles

L'ensemble du jeu de données regroupe 199 patients, chacun ayant fait l'objet de trois acquisitions correspondant à trois panels d'anticorps différents (tubes A, B et C). Chaque panel contient une dizaine de marqueurs, définissant ainsi un sous-espace propre d'analyse. Cela permet de couvrir différentes facettes de l'immunophénotypage, tout en introduisant une variabilité dans les canaux mesurés selon le tube. Dans le cadre du stage, on utilisera uniquement les données du Tube A.

	FS-A	SS-A	CD33 PC5.5	CD34 PC7	CD45 KO	CD65 FITC	CD64 FITC	CD36 FITC	CD14 PE	CD10 PE	CD61 PE	CD13 ECD	CD4 ECD	CD117 APC	CD123 APC	CD2 APC	CD7 APC700	CD56 APC700	CD71 APC750	CD110 APC750	CD19 APC750	CD38 PB	HLA-DR PB	CD16 PB
Panel A	x	x	x	x	x	x			x			x		x			x			x				x
Panel B	x	x	x	x	x		x			x			x		x		x			x				
Panel C	x	x	x	x	x			x			x					x		x	x		x	x	x	

Les données utilisées dans ce projet proviennent de fichiers `.fcs`, format standard en cytométrie en flux. Chaque fichier contient les intensités mesurées pour un ensemble de canaux. Des informations supplémentaires (métadonnées) sont également disponibles dans l'entête du fichier, telles que l'identifiant patient, la date, ou les paramètres d'acquisition.

Tous les fichiers ont été préalablement anonymisés afin de garantir la confidentialité des données cliniques.

1.4.2 Organisation méthodologique

Dans le cadre de ce projet, je réutilise les étapes de prétraitement mises en place par Camilla Paleari [4], dans le cadre de son stage de Master 2.

Ces étapes incluent :
— le chargement des fichiers `.fcs` ;
— le retrait des événements marginaux ;
— la compensation des chevauchements spectraux entre canaux fluorescents ;
— l'application d'une transformation logicle avec les paramètres suivants : $w = 0{,}75$, $t = 275\,000$, $m = 4{,}5$, $a = 0$.

Dans la suite de mon travail, je travaillerai sur des données pré-traitées de la sorte.

Le rapport s'organise en quatre grandes parties. La première est consacrée à l'introduction et présente successivement le contexte biologique de la leucémie aiguë myéloblastique, les principes de la cytométrie en flux, les étapes de prétraitement et de nettoyage des données, ainsi que les objectifs du stage. La deuxième partie traite de la problématique centrale, à savoir la séparation des cellules viables et des débris, en explorant notamment une segmentation basée sur la norme d'expression membranaire et l'utilisation du clustering par K-means. La troisième partie est dédiée à l'automatisation de la sélection des clusters de débris. Enfin, la quatrième partie propose un récapitulatif des principaux résultats obtenus.

2 Séparer cellules viables et débris pour chaque patient

Dans ce chapitre, nous proposons différentes stratégies de clustering afin de distinguer cellules viables et débris cellulaires. Ici, nous travaillons à l'échelle individuelle : chaque patient est analysé séparément, en utilisant ses propres données. Nous illustrerons les résultats obtenus sur plusieurs cas-patients.

La sous-section 2.1 présentera une méthode basée sur le calcul d'une norme d'expression en protéines membranaires. La sous-section 2.2 présentera une approche basée sur la méthode des kmeans. D'autres méthodes, telles que *FlowSOM*, *DBScan* ou encore le clustering hiérarchique agglomératif, ont également été testées, mais elles se sont révélées moins satisfaisantes que celles présentées dans la suite.

2.1 Segmentation par la norme d'expression membranaire

Dans cette approche, nous nous appuyons sur le travail de Léa Comin [7], qui a proposé une méthode pour discriminer les cellules viables des débris à partir de leur profil d'expression dans les différents marqueurs considérés.

Soit N_{cell} le nombre de cellules dans l'échantillon, $1 \leq k \leq N_{\text{cell}}$ l'indice d'une cellule, et n le nombre total de marqueurs. On note $(m_{jk})_{1 \leq j \leq n}$ le vecteur d'expression de la cellule k où m_{jk} représente l'intensité du marqueur j pour la cellule k, et n est le nombre total de marqueurs de fluorescence.
Sa norme euclidienne s'écrit alors :

$$\text{Norme}_k = \sqrt{\sum_{j=1}^{n} m_{jk}^2}$$

Cette quantité permet d'obtenir une mesure globale de l'intensité d'expression, potentiellement indicative de la viabilité cellulaire.

2.1.1 Norme classique

Nous calculons la norme d'expression chez toutes les cellules de 13 patients. La Figure 8 présente les histogrammes correspondants.

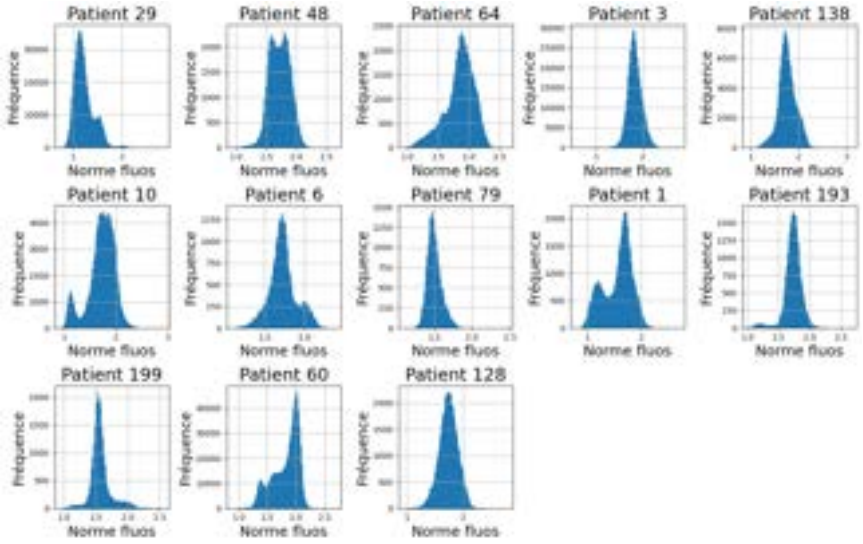

FIGURE 8 – Histogrammes des normes d'expression (fluorescence) sur treize patients.

Les distributions observées présentent un à plusieurs modes, jusqu'à quatre. L'approche de Léa Comin proposait de filtrer les données en dessous d'un seuil. Ici, il semblerait que filtrer sous le seuil 1.4 puisse enlever des sous-populations de cellules d'expression faible. Cependant, la procédure reste arbitraire et la variabilité inter-patients reste très importante.

Nous proposons de calibrer un modèle de mélange de trois gaussiennes sur l'histogramme global obtenu à partir du même sous échantillon de treize patients.

La Figure 9 illustre le résultat. L'ajustement est satisfaisant pour la première composante, qui nous intéresse le plus, tandis que les suivantes s'ajustent moins bien, ce qui reste secondaire pour notre analyse.

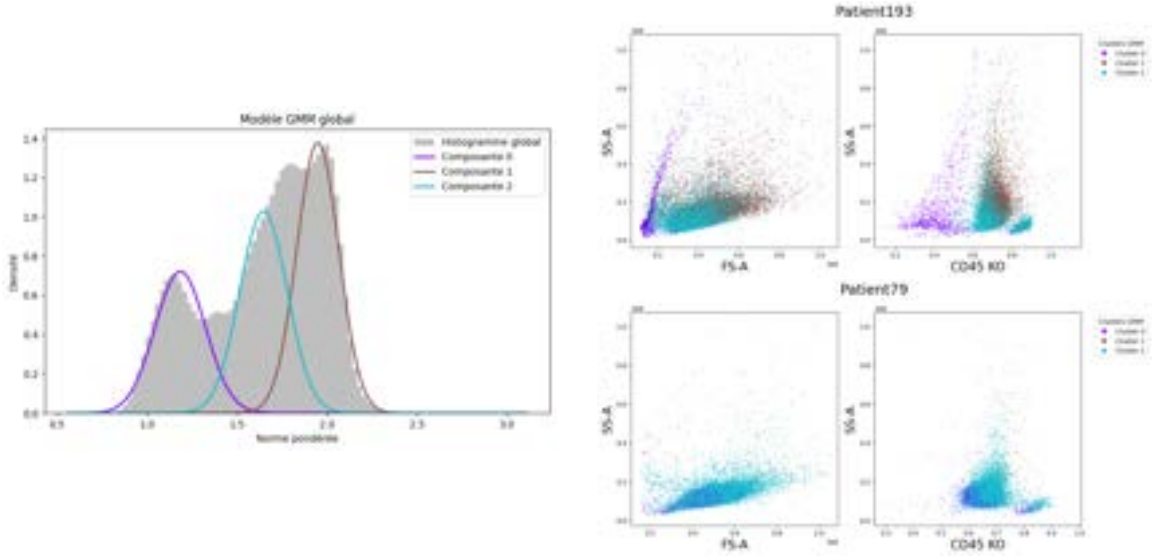

FIGURE 9 – Ajustement d'un modèle GMM à 3 composantes (gauche) sur l'ensemble des normes de 13 patients : 1, 3, 6, 10, 29, 48, 60, 64, 79, 128, 193, 199 avec graphes sur les plans FS-SS et CD45-SS (droite) des cellules des patients 193 (haut) et 79 (bas) dans les plans FS-SS (gauche) et CD45-SS (droite). La couleur des points correspond aux clusters obtenus par calibration d'un mélange de 3 gaussiennes.

Les trois composantes identifiées présentent respectivement des moyennes de 1.18, 1.95 et 1.65, des écarts-types de 0.14, 0.12 et 0.13, et des poids relatifs de 25% 40% et 34%.

Nous supposons que la composante de plus faible intensité correspond aux cellules à exclure.

Pour le patient 193, on observe des cellules bleues et marron dans le croissant du graphe FS–SS, signe que certaines cellules sont conservées alors qu'elles auraient probablement été exclues par les cytométristes ; à l'inverse, quelques cellules violettes éparses pourraient correspondre à des débris. Dans l'ensemble, on ne semble pas supprimer de cellules CD45+. Le patient 79 présente moins de cellules, et donc moins de débris. Dans le graphe FS-SS, on retrouve le même écueil que chez le patient 193 : toutes les cellules du croissant ne sont pas captées par le cluster 0. Dans le graphe CD45-SS, les cellules du cluster 0 sont co-localisées avec celles du cluster 2, notamment en SS bas avec CD45 autour de 0,6 (zone des blastes) et vers SS $\approx 0{,}8$ (zone des lymphocytes) ; ces cellules pourraient correspondre à des cellules mortes d'expression plus faible, mais leur suppression ferait courir le risque d'éliminer des cellules à garder.

Plus généralement, ces observations suggèrent que des cellules présentant une morphologie de débris ou de cellules mortes peuvent conserver une expression « normale ». À l'inverse, des cellules morphologiquement « normales » peuvent exhiber une norme d'expression basse, compatible avec un état de mort cellulaire, mais dont l'exclusion ferait peser un risque sur la conservation de cellules viables, sans moyen direct de vérifier leur viabilité dans nos données.

Afin d'éviter de sélectionner à tort des cellules $CD45^+$ à exclure, nous proposons d'introduire une norme pondérée accordant un poids spécifique au marqueur CD45.

Au vu de la forme de l'ajustement final, il semblerait que le modèle s'ajusterait peut-être mieux sur 4 composantes. Nous procédons à la même opération en ajustant un modèle GMM à 4 composantes.

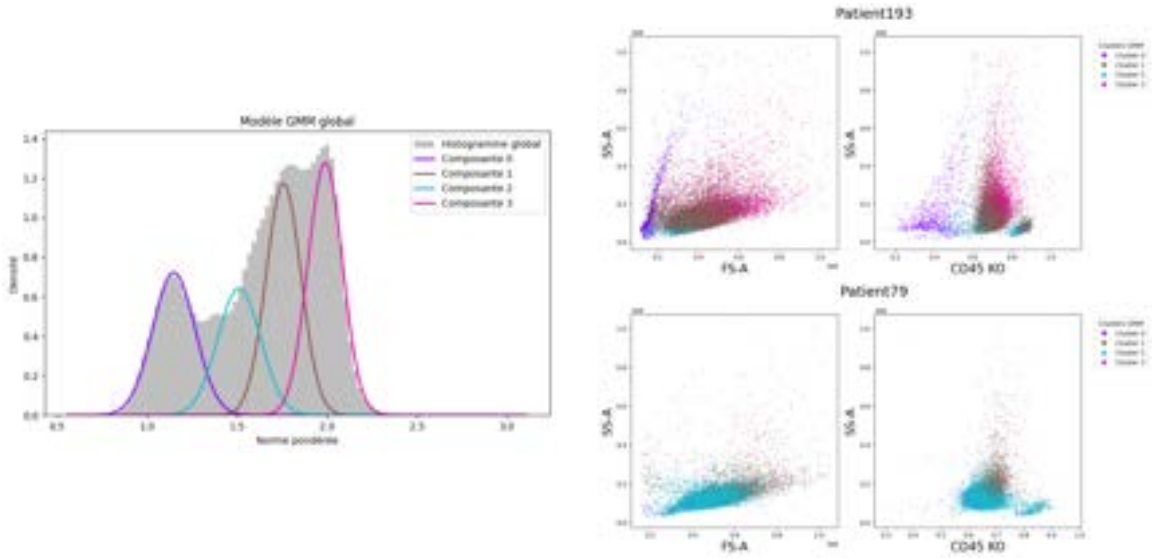

FIGURE 10 – Ajustement d'un modèle GMM à 4 composantes (gauche) sur l'ensemble des normes de 13 patients : 1, 3, 6, 10, 29, 48, 60, 64, 79, 128, 193, 199 avec graphes (droite) des cellules des patients 193 (haut) et 79 (bas) dans les plans FS-SS (gauche) et CD45-SS (droite). La couleur des points correspond aux clusters obtenus par calibration d'un mélange de 4 gaussiennes.

Les trois composantes présentent des moyennes respectives de 1.15, 1.76, 1.51 et 1.99 ; des écarts-types de 0.12, 0.1, 0.12 et 0.1 ; et des poids relatifs de 21 %, 29 % et, 19 % et 31 %.

COMMENTER

Par ailleurs, nous savons que les débris peuvent être caractérisés par une faible expression en CD45. Dans la sous-partie suivante, nous proposons donc d'utiliser cette caractéristique.

2.1.2 Norme pondérée sur CD45

Dans cette partie, nous explorons une approche basée sur une **norme pondérée**, en attribuant un poids plus important au canal CD45 KO. Nous supposons que cela peut favoriser la discrimination des cellules CD45.

Les Figures 11 et 12 présentent les histogrammes obtenus pour différents patients, avec une pondération du marqueur **CD45 KO** fixée à 2 et à 10. On observe que l'augmentation du poids accentue les pics et rend les populations plus distinctes.

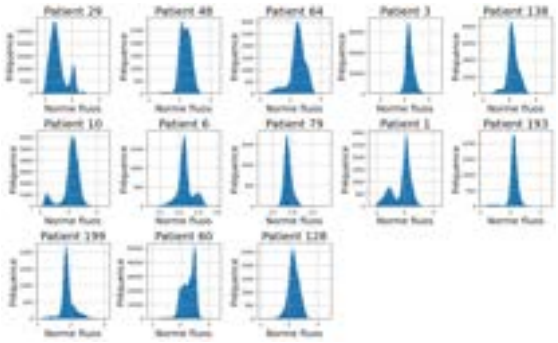

FIGURE 11 – CD45 pondéré par un facteur 2

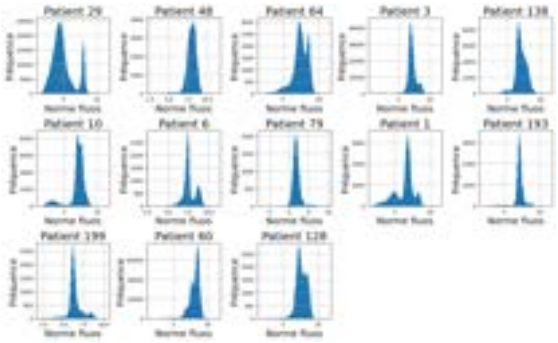

FIGURE 12 – CD45 pondéré par un facteur 10

De même, nous avons ajusté un modèle de mélange de 3 gaussiennes sur l'ensemble des normes pondérées, des mêmes 13 patients. Le résultat est présenté en Figure 13.

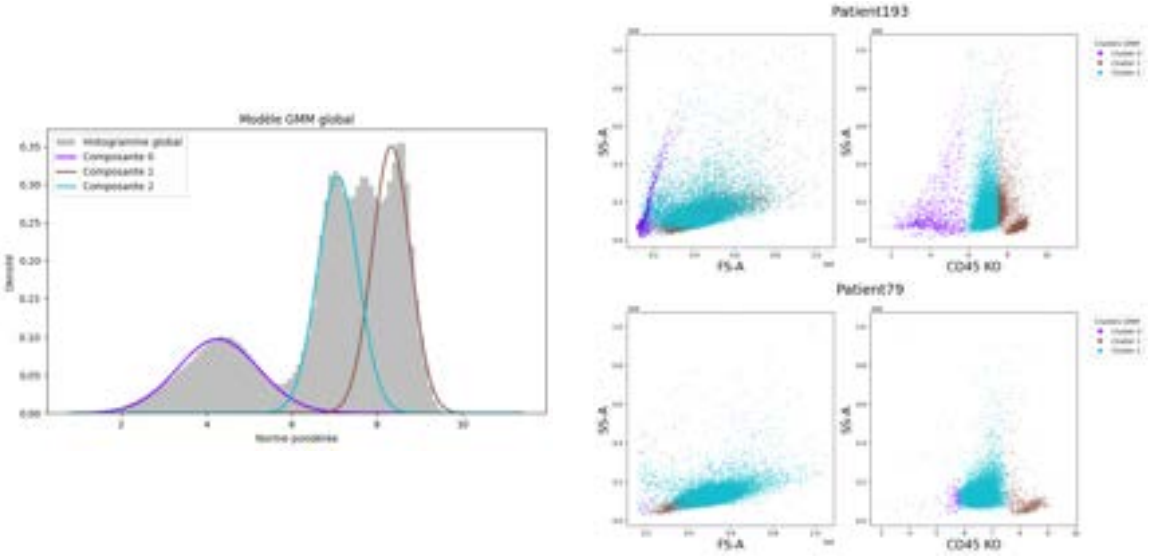

FIGURE 13 – Ajustement d'un modèle GMM à 4 composantes (gauche) sur l'ensemble des normes pondérées sur CD45 par 10 de 13 patients : 1, 3, 6, 10, 29, 48, 60, 64, 79, 128, 193, 199 avec graphes (droite) des cellules des patients 193 (haut) et 79 (bas) dans les plans FS-SS (gauche) et CD45-SS (droite). La couleur des points correspond aux clusters obtenus par calibration d'un mélange de 3 gaussiennes.

Les quatre composantes présentent des moyennes respectives de 4.27, 8.33 et 7.08 ; des écarts-types de 0.95, 0.45 et 0.48 ; et des poids relatifs de 23 %, 39 % et 37 %.

Comparée à la version non pondérée, cette pondération accentue la séparation entre les groupes, en particulier pour la première composante. Comme dans l'autre cas, on fait l'hypothèse que la première composante correspond à des cellules à enlever.

Pour le patient 193, la situation est assez similaire : on observe également des cellules bleues et marron dans le croissant du graphe FS-SS. Les cellules marron se situent toutefois en partie au niveau inférieur, correspondant aux blastes, tandis que le reste apparaît plus dispersé. Dans le graphe CD45-SS, les frontières entre clusters sont très nettes, autour des valeurs 6 et 7.5.

Chez le patient 79, dans le graphe CD45-SS, les cellules du cluster 1 sont plus nombreuses, alors que le cluster 0 reste presque absent. Les débris correspondent à la sélection attendue, comme dans la Figure 7. Là encore, le graphe CD45-SS présente une séparation bien marquée aux mêmes niveaux. On note cependant un mélange entre débris et blastes, ce qui rend l'interprétation plus confuse.

Les résultats obtenus sont globalement satisfaisants, mais il convient de rester vigilant, car certaines leucémies présentent une très faible expression des marqueurs. Bien que ces situations soient rares, elles peuvent compliquer l'analyse. Dans de tels cas, le cytométriste s'appuie davantage sur les paramètres morphologiques SS et FS, et peut compléter l'évaluation en ajoutant d'autres marqueurs de fluorescence, voire en recourant à l'observation microscopique directe.

La distribution suggère encore 4 modes, on on réitère.

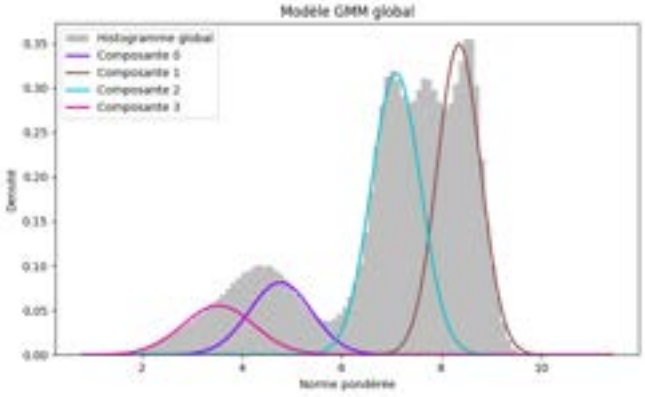

FIGURE 14 –

Les quatre composantes présentent des moyennes respectives de 4.77, 8.35, 7.09 et 3.53 ; des écarts-types de 0.63, 0.44, 0.49 et 0.72 ; et des poids relatifs de 13 %, 38 % et, 39 % et 10 %.

On constate que l'ajustement n'est pas satisfaisant, car les courbes ne suivent pas correctement la forme des histogrammes et ne reflètent pas fidèlement la distribution de la norme. Dans ce cas, le fit n'apporte pas d'information pertinente.

Conclusion. Ainsi, une approche fondée uniquement sur l'expression globale ne permet pas d'identifier de façon spécifique les cellules à exclure. La norme d'expression constitue un indicateur intéressant : les cellules présentant les plus petites valeurs correspondent généralement à des événements à éliminer. Toutefois, cela ne couvre pas l'ensemble des cas, car certaines cellules ayant l'apparence morphologique de débris conservent des niveaux d'expression membranaire « normaux ». Par ailleurs, la variabilité de la proportion de débris entre patients rend inadapté un filtrage basé sur un simple seuil ou pourcentage fixe.

2.1.3 Lien entre expression membranaire et morphologie

La norme d'expression est calculée à partir de la fluorescence associée aux marqueurs membranaires. Elle fournit une indication globale de l'intensité d'expression de ces marqueurs pour chaque cellule. Cependant, dans la pratique, la distinction entre cellules viables et débris repose aussi largement sur les paramètres morphologiques mesurés par cytométrie, notamment l'intensité en diffusion avant (FS) et latérale (SS), qui reflètent respectivement la taille et la granularité de la cellule.

Une question centrale est donc la suivante : *quelle est la relation entre ces deux types de quantités ?* Autrement dit, les cellules présentant une faible norme d'expression correspondent-elles systématiquement à une morphologie de type « débris » ?

Visualisations 2D. Pour explorer ce lien, nous avons représenté des histogrammes 2D croisant la norme d'expression avec les paramètres SS-A, FS-A et CD45 pour trois patients choisis aléatoirement. L'objectif est d'identifier des tendances communes ou des différences marquées entre individus.

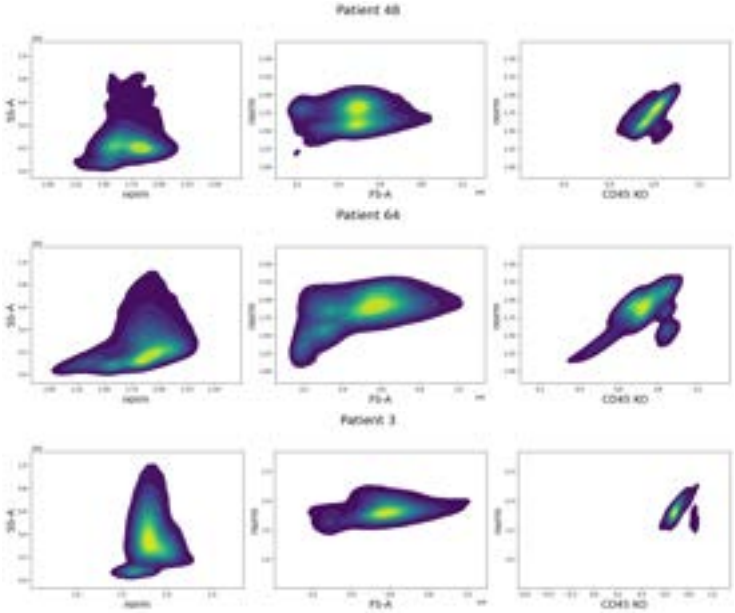

FIGURE 15 – Histogrammes 2D de la norme d'expression croisée avec SS-A, FS-A et CD45 pour 3 patients : 48, 64 et 3

Pour le patient 48, on observe que les cellules de norme faible présentent de petits FS, formant une tâche séparée. En SS, ces mêmes cellules apparaissent également avec des valeurs plus basses. En revanche, dans le graphe CD45, elles ne se situent pas dans les valeurs faibles ; au contraire, elles tendent vers les lymphocytes. Il est d'ailleurs surprenant que l'on ne retrouve pas dans ce graphe les cellules de norme autour de 1.25, qui devraient pourtant y figurer.

Chez le patient 64, le nombre de cellules est plus important. On retrouve la même tendance en FS : bien qu'il n'y ait pas de séparation nette avec les autres cellules, on observe une corrélation entre une norme faible et un FS faible. En SS, la tendance est similaire. En CD45, cette fois, les cellules de norme faible apparaissent comme CD45 faibles, ce qui correspond aux populations qui nous intéressent. Toutefois, ce résultat souligne aussi le risque de sélectionner par erreur des lymphocytes si l'on applique un seuil uniquement basé sur la norme, un problème déjà constaté chez un autre patient.

Enfin, le patient 3 illustre le cas d'un échantillon avec peu de cellules. On y retrouve néanmoins la même tendance que chez le patient 48.

Cela peut servir à dire que les cellules de faible norme sont généralement dans la zone morphologique ciblée par les médecins. Cependant, il y a un risque d'inclure des cellules viables dans les zones frontières : on a vu que des lymphocytes peuvent être de faible norme. *Faut-il les enlever ?*

Par ailleurs, dans les trois cas, on voit un amas de cellules de normes juste au dessus de 1.5, et avec SS bas. Cela correspond sans doute aux blastes, donc les cellules à ne pas enlever. Déterminer la frontière uniquement avec la norme parait difficile.

Par contre, pour ces deux difficultés, CD45 peut être intéressant : on voit que chez les patients 48 et 3, on ne distingue pas de population différente et de CD45 faible, ce qui suggère de ne rien enlever. Chez le patient 64, on voit une population supplémentaire qui correspond aux cellules à enlever.

Dans l'optique d'optimiser la spécificité du nettoyage en n'enlevant pas des cellules viables à tort,

on perçoit donc l'importance de CD45 pour aider à les distinguer des blastes et des lymphocytes.

2.2 K-means

Les algorithmes de clustering (ou classification non supervisée) visent à regrouper automatiquement des observations similaires en groupes homogènes, sans avoir besoin d'étiquettes préalables. Ils sont utilisés pour explorer les structures cachées dans les données, en réduire la complexité, ou encore en tant qu'étape de prétraitement avant d'autres analyses.

Dans notre projet, cette méthode est appliquée aux données de cytométrie en flux, où chaque événement correspond à une cellule caractérisée par des paramètres morphologiques et biologiques.

2.2.1 Principe de la méthode

L'algorithme *K-means* regroupe n points en K clusters en minimisant la **variance intra-classe**, c'est-à-dire la somme des distances au carré entre chaque point et le centroïde de son groupe. L'algorithme fonctionne de manière itérative :

1. Initialiser aléatoirement K centroïdes.
2. Affecter chaque point au centroïde le plus proche selon la distance euclidienne.
3. Recalculer les centroïdes comme moyenne des points de chaque cluster :

$$\mu_k^{(t+1)} = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

4. Répéter les étapes 2 et 3 jusqu'à convergence (centroïdes stables).

Standardisation des données Avant d'appliquer un algorithme de clustering, il est essentiel de standardiser les données, notamment si les variables n'ont pas la même échelle.

On transforme chaque variable X_j correspondant à un marqueur, en une variable centrée réduite :

$$X_j^{\text{scaled}} = \frac{X_j - \mu_j}{\sigma_j}$$

où μ_j est la moyenne de la variable X_j sur l'ensemble des cellules patients et σ_j est son écart-type.

Cela permet de donner le même poids, en termes d'ordres de grandeur, à toutes les variables dans les calculs de distances.

Choix du k L'analyse de silhouette permet d'estimer le nombre optimal de clusters en évaluant la compacité interne des groupes et leur séparation mutuelle. Dans notre cas, les meilleurs résultats sont obtenus pour des valeurs de k comprises entre 2 et 4. Toutefois, ce critère n'est pas nécessairement adapté à l'identification spécifique des débris. C'est pourquoi, dans la suite, nous comparerons différents choix de k en fonction de leur capacité à isoler efficacement les débris.

2.2.2 Choix des caractéristiques considérées

Dans un premier temps, nous cherchons à déterminer quels marqueurs inclure dans le clustering.

Afin de faciliter l'analyse comparative des algorithmes, nous restreignons l'étude au même sous-échantillon de 13 patients : 1, 3, 6, 10, 29, 48, 60, 64, 79, 128, 193 et 199 que précédemment.

Apprentissage sur les 12 marqueurs disponibles

Dans un premier temps, nous appliquons l'algorithme *k-means* en utilisant l'ensemble des marqueurs morphologiques et de fluorescence disponibles. Cette approche vise à exploiter la totalité de l'information biologique.

La Figure 16 représente les résultats du clustering obtenus pour différents nombre de clusters.

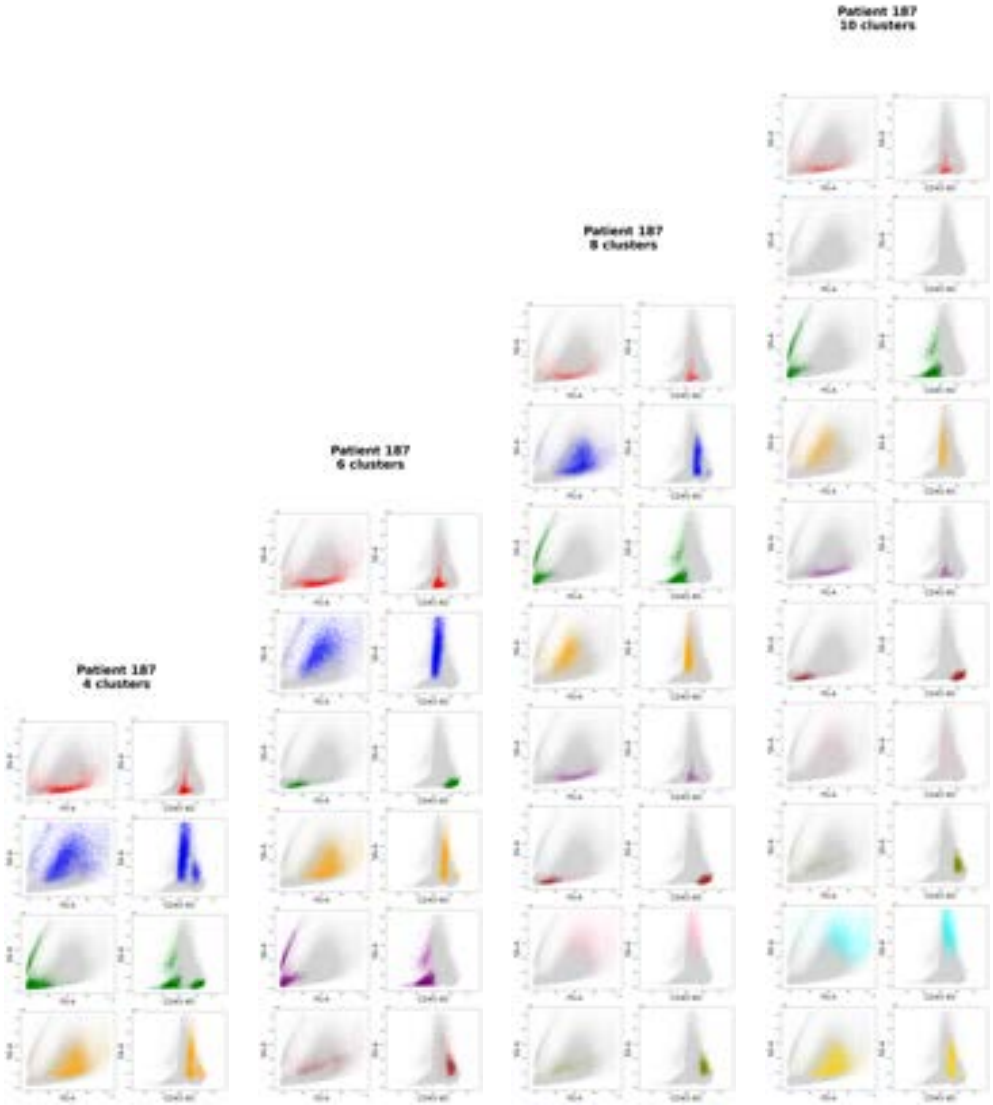

FIGURE 16 – Graphes FS-SS et CD45-SS des données du patient 187 en séparant les clusters obtenus par clustering sur les 12 marqueurs. Les colonnes correspondent respectivement aux cas de 4, 6, 8 et 10 clusters, de gauche à droite. Les couleurs correspondent aux clusters.

Dans le cas de 4 clusters (colonne de gauche), on voit notamment que le troisième cluster (troisième ligne en partant du haut) comporte à la fois des cellules ayant une morphologie de débris et des lymphocytes. Cela montre que ce cas n'est pas satisfaisant, et donc que le résultat de l'analyse de silhouette n'est pas pertinent dans notre cas.

Dans les trois autres cas, on voit que les lymphocytes sont bien séparés (colonne 2 ligne 3, colonne 3 ligne 6, colonne 4 ligne 6).

Par contre, on voit aussi qu'on garde ensemble des cellules de la zone de débris et des cellules à garder, même lorsque k augmente et qu'on obtient des clusters plus fins (par exemple colonne 2 ligne 2, puis colonne 3 ligne 4 et colonne 4 ligne 4).

Cela montre que les données d'expression prennent trop le pas sur les données morphologiques, ce qui rejoint l'écueil de la partie précédente.

Apprentissage sur les 3 marqueurs spécifiques

Nous restreignons maintenant l'analyse à trois marqueurs : SS-A (granularité), FS-A (taille) et CD45 (marqueur pan-leucocytaire) [6].

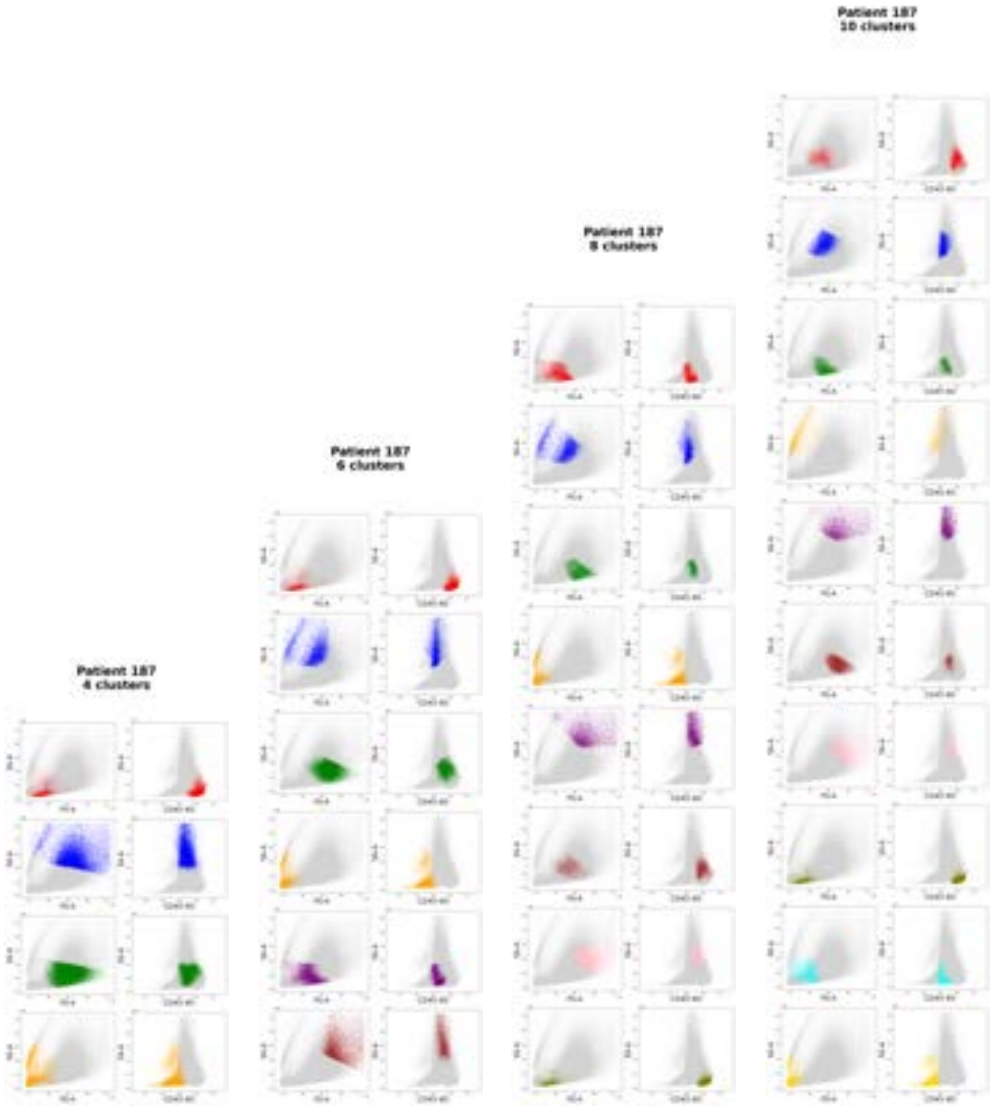

FIGURE 17 – Graphes S FS-SS et CD45-SS des données du patient 187 en séparant les clusters obtenus par clustering sur les marqueurs sur SS, FS et CD45. Les colonnes correspondent respectivement aux cas de 4, 6, 8 et 10 clusters, de gauche à droite.

La Figure 27 montre que cette configuration réduit les ambiguïtés observées précédemment. Les clusters apparaissent mieux définis, en particulier dans les zones correspondant au croissant FS-SS caractéristique des débris. L'ajout ciblé de CD45 semble renforcer la capacité de séparation entre les

populations.

Dans les cas $k = 4$, $k = 6$ et $k = 8$, les résultats ne sont pas satisfaisants, en particulier pour les clusters de la ligne 2 où des cellules très différentes en morphologies sont regroupées.

Pour $k = 10$, certains clusters correspondent clairement à la zone morphologique des débris (colonne 4, lignes 4 et 10). Une légère confusion subsiste toutefois, comme en colonne 4, ligne 8, où les cellules identifiées sont en réalité des lymphocytes. On note d'ailleurs que certains lymphocytes présentent des morphologies proches de débris, mais leur nombre reste limité et leur conservation peut être tolérée.

Conclusion. Bien que les 12 marqueurs portent davantage d'information biologique, leur utilisation ne permet pas de séparer les débris de manière satisfaisante. En revanche, la configuration à trois marqueurs (SS-FS-CD45) présente des clusters séparant débris des cellules viables de manière plus satisfaisante, en particulier lorsque le nombre de clusters est assez grand. Nous choisissons donc cette configuration pour la suite de l'analyse.

Impact de la norme d'expression

Afin d'évaluer l'impact de l'intégration de la norme d'expression dans le processus de clustering, nous avons appliqué un K-means avec $k = 8$ en faisant varier les variables incluses dans l'analyse.

La Figure 18 illustre les résultats obtenus. Chaque ligne de la Figure correspond à un cluster, représenté à gauche dans l'espace FS-SS et à droite dans l'espace CD45-SS. En bas de la Figure, un boxplot synthétise la distribution de la norme d'expression pour chacun des clusters. Cette représentation permet ainsi de comparer directement l'effet du choix des variables sur la segmentation des cellules et sur la séparation des débris.

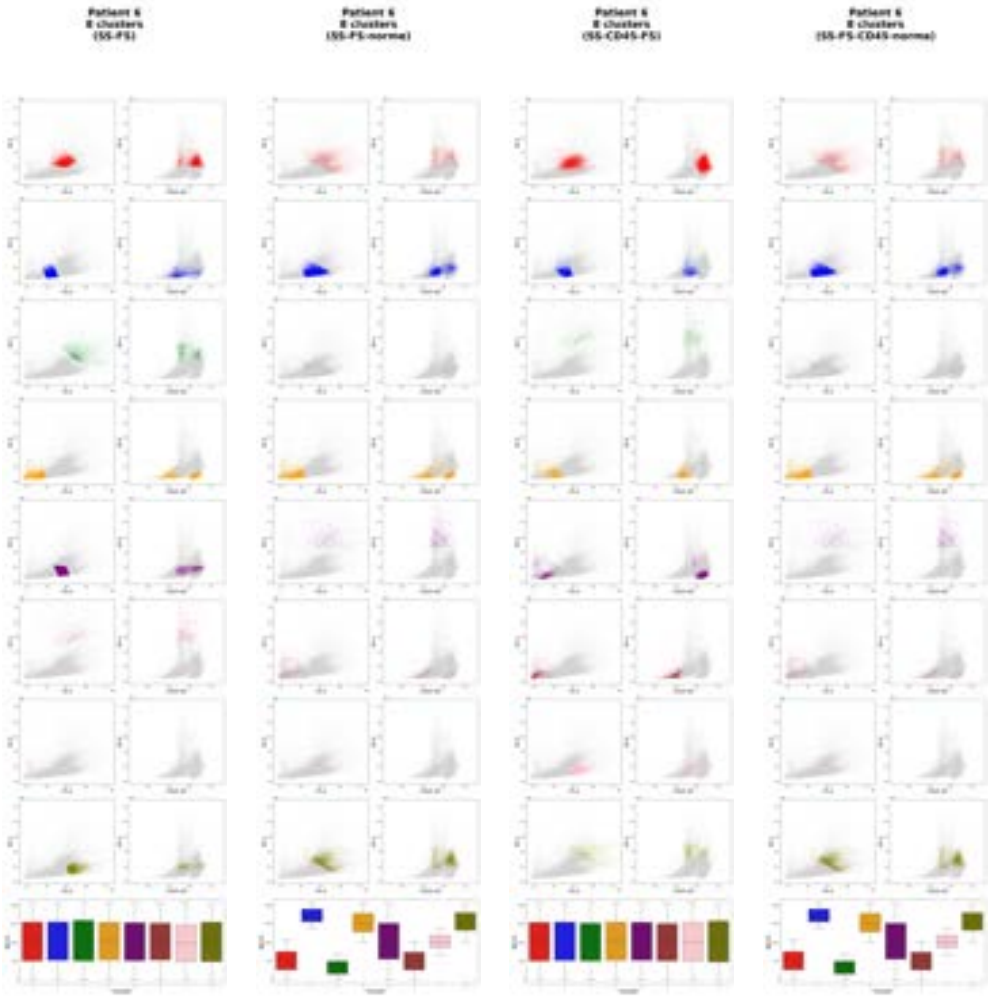

FIGURE 18 – Graphes FS-SS et CD45-SS de comparaison de 4 variantes de clustering 8-means sur le patient 6 sur différents marqueurs, avec les boxplots des normes d'epression dans les différents clusters

Pour le clustering basé sur FS-SS, on perçoit des limites, comme le cluster 5 en partant du haut qui mélange des lymphocytes, et des débris, voire des blastes. C'est normal puisqu'on ne tient pas compte de l'information du CD45.

Lorsqu'on ajoute la norme au clustering, cela ne change pas cette limite. Cependant, les boxplots de norme montrent que les nouveaux clusters distinguent les cellules de normes plus basses (clusters 1, 3 et 6 en partant du haut). On remarque en particulier que les clusters 3 et 6 présentent des cellules qui n'ont pas nécessairement des morphologies de débris.

Comme déjà remarqué, le clustering basé sur SS-FS-CD45 est plus satisfaisant, bien que les clusters soient homogènes en termes de normes d'expression.

En ajoutant la norme d'expression au clustering les frontières des clusters sont plus nettes et les contours sont plus perceptibles (cluster 6 en partant du haut).

Ainsi, considérer la norme d'expression dans le clustering peut aider à identifier des clusters de

norme basse, qui pourraient correspondre à de jeunes cellules mortes. Cependant, vérifier cette hypothèse nécessiterait d'utiliser des données portant un marqueur de viabilité. Il s'agit donc d'une perspective pour des travaux futurs.

C'est intéressant car pour FS-SS la norme semble distinguer des cellules un peu éparses qui ont des normes plus petites. Cela pourrait correspondre à des cellules mortes qui ne sont pas encore dans la zone morphologique bien identifiée. Ce serait intéressant de vérifier cette hypothèse avec des données comportant un marqueur de viabilité, mais nous n'avons pas ce type de données, donc cela reste une perspective.

On dirait que la norme a séparé certains groupes, qu'on retrouve éparpillés dans le plan. On se demande sil y a une grande différence entre les cellules en périphérie et celles du centre

La norme peut être un paramètre synthétique supplémentaire. L'ajout de la norme ne suffit donc pas toujours à isoler clairement les débris : certaines cellules mortes ou intermédiaires conservent une norme modérée, probablement en raison d'une hétérogénéité d'expression. Ainsi, bien que prometteuse, la norme d'expression gagne à être combinée à d'autres indicateurs comme CD45 pour améliorer la robustesse du clustering.

Chez certains patients, on observe un ou plusieurs clusters bien séparés dans la partie gauche ou inférieure du nuage de points — ce qui suggère une segmentation pertinente et potentiellement exploitable pour une sélection automatique.

2.2.3 Conclusions

En résumé, le triplet SS–FS–CD45 semble offrir un bon compromis pour discriminer les débris des autres populations cellulaires. Après comparaison de différents nombres de clusters, nous avons choisi de considérer k=10 clusters dans la suite de l'analyse.

Nous avons vu que la norme d'expression peut être utilisée conjointement aux marqueurs morphologiques, et semble capable d'identifier des cellules éparses de normes basses. Celles-ci pourraient être des cellules mortes. Cependant, il faudrait tester cette hypothèse avec des marqueurs de viabilité qui ne sont pas présents dans ces données. Cela représente donc une perspective pour des travaux ultérieurs.

Ce chapitre nous a permis de proposer une approche de clustering par 10-means basée sur les marqueurs SS-FS-CD45, qui semble former des clusters homogènes de cellules à enlever à l'échelle d'un patient.

Nous abordons donc dans le chapitre suivant les stratégies d'**automatisation de la sélection des clusters de débris**. L'objectif est de développer un pipeline capable d'identifier ces événements de manière fiable, sans supervision manuelle, tout en s'adaptant à la variabilité inter-patient.

3 Automatisation de la sélection des clusters débris

Dans ce chapitre, nous cherchons à définir une stratégie robuste de sélection des débris et cellules mortes, applicable à l'ensemble des patients. L'un des principaux défis de la cytométrie en flux est la forte variabilité inter-patients, liée à la fois à des facteurs biologiques, comme la quantité de débris ou l'hétérogénéité cellulaire, et à des facteurs techniques, tels que les réglages, la compensation ou le bruit. Il est donc essentiel de concevoir un cadre de clustering capable d'identifier automatiquement les groupes de débris, tout en s'adaptant à cette diversité sans supervision manuelle.

La pratique des cytométristes repose sur une zone morphologique définie dans l'espace FS-SS par une valeur faible de FS, et qui présente souvent une forme courbe (voir Figure 7). Dans la suite, on appellera cette zone la *zone morphologique* des débris.

Dans un second temps, les cytométristes s'assurent de supprimer les cellules restantes qui seraient CD45-. En parallèle, les cellules CD45- sont systématiquement éliminées. La question est alors de savoir comment traduire cette pratique manuelle en une méthode automatisée, suffisamment robuste pour être généralisée.

Notre proposition consiste à exploiter le clustering local réalisé patient par patient, puis à opérer un méta-clustering sur les centroïdes des clusters locaux. Cette approche permet d'agréger l'information issue d'un ensemble de patients afin de former un cadre commun, et ouvre ensuite la possibilité de définir une stratégie de décision directement au niveau des méta-clusters.

La Figure 19 illustre la variabilité inter-patients. Elle présente la position des centroïdes obtenus sur cinq patients avec un découpage en huit clusters sur trois marqueurs (SS-A, FS-A, CD45 KO). Chaque couleur correspond à un patient et les huit points de cette couleur représentent ses centroïdes. On observe que, bien que certaines positions se retrouvent dans des zones typiques de débris ou de cellules classiques, d'autres se situent en frontière ou en dehors de ces zones. Cette dispersion complique la correspondance directe entre clusters d'un patient à l'autre et justifie le recours à une méthode en population capable d'intégrer cette variabilité.

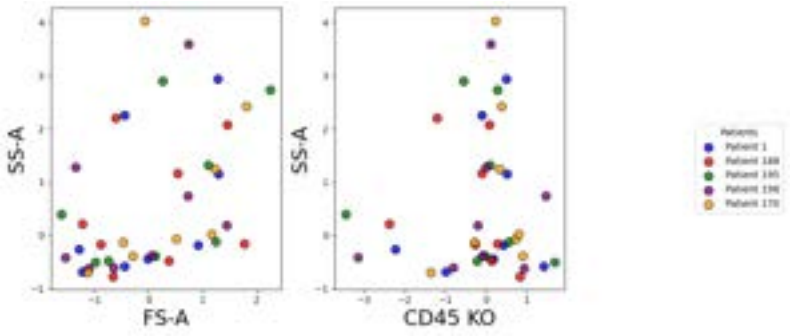

FIGURE 19 – Représentation sur les plans FS-SS et CD45-SS des centroïdes obtenus sur 5 patients (1, 170, 188, 195 et 198) avec les clusters résultants d'un 8means sur les 3 marqueurs : SS, FS et CD45.

Dans la suite, nous proposons deux approches. La première vise à détecter automatiquement la zone morphologique dans l'espace (FS,SS). La seconde cherche à identifier directement les populations à faible expression en CD45. Ces deux approches seront illustrées sur un sous-échantillon représentatif de cent patients, en s'appuyant systématiquement sur un clustering local par 10-means appliqué aux marqueurs (SS, FS, CD45), comme présenté au chapitre précédent.

3.1 Détecter la zone morphologique de débris dans une population

Dans un premier temps, nous cherchons à isoler la zone morphologique des débris définie dans le plan (FS-SS). Nous commençons par regrouper les clusters locaux via un nouveau clustering en population.

3.1.1 Clustering en population

L'approche consiste à regrouper les centroïdes des clusters individuels de plusieurs patients dans un même espace de référence, puis à appliquer un second niveau de clustering sur ces centroïdes. Cette stratégie permet de dégager des structures communes sur un jeu de données élargi, et non plus limité à l'échelle individuelle. Dans la suite, nous appelons ce nouveau clustering le meta-clustering. Après différents essais, nous décidons de considérer autant de meta-clusters que de clusters dans l'étape locale. Nous considérons d'abord un meta-clustering basé uniquement sur les paramètres morphologiques.

10-means sur (FS,SS) La Figure 20 représente les clusters locaux des 100 patients considérés dans les plans (FS,SS) et (CD45,SS). Les couleurs correspondent aux meta-clusters obtenus.

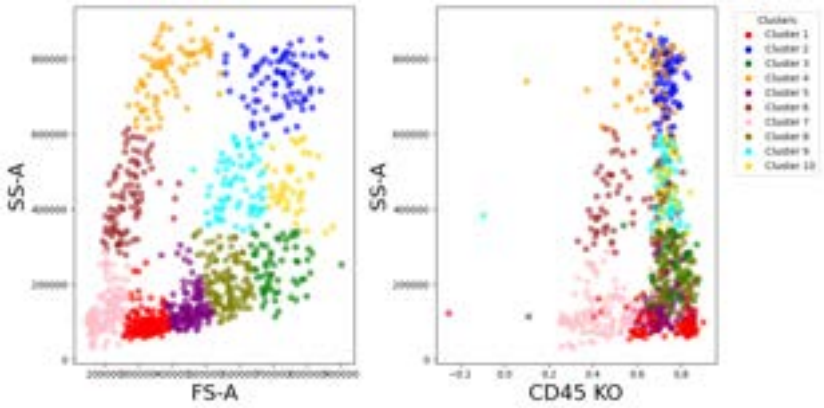

FIGURE 20 – Centroïdes des clusters locaux des 100 individus considérés, représentés dans les plans (FS,SS) et (CD45,SS). Les couleurs corespondent aux méta-clusters obtenus par 10-means sur (SS,FS).

On observe des meta-clusters bien définis en (SS,FS), dont 3 qui correspondent à la zone morphologique d'intérêt (meta-clusters 4,6,7). La distinction entre les meta-clusters dans le plan (SS,CD45) est plutôt encourageante. On peut cependant remarquer deux limites. D'une part, certains clusters forment des outliers en CD45, et sont tous associés à des meta-clusters différents. Or, ils sont CD45-, et sont donc clairement à supprimer lors du nettoyage.

3.1.2 Rapprochement des méta-clusters d'intérêt à retirer

Dans cette partie, nous proposons une stratégie de sélection automatique des méta-clusters correspondant aux cellules à exclure (méta-clusters 4, 6 et 7 de la Figure 20). La difficulté est que se baser uniquement sur les plus faibles valeurs de `FS` conduit à inclure des populations indésirables, comme les blastes (par exemple le cluster 1 en rouge), alors que l'on souhaite cibler plutôt des zones de `FS` légèrement plus élevées (par exemple le cluster 4 en jaune foncé). Pour contourner ce problème, nous exploitons la forme caractéristique de la zone morphologique des débris en introduisant un changement de repère, de manière à mieux isoler ces celluleset à redresser la zone des débris.

Pour cela, nous définissons une droite D passant par les deux méta-centroïdes des méta-clusters associés aux débris (vert et jaune dans la Figure 20).

Soient p_1 et p_2 les coordonnées de ces deux méta-centroïdes dans le plan (SS, FS).

Le vecteur $p_2\vec{p}_1 = p_2 - p_1$ est dirigé du premier centroïde vers le second. On définit alors un vecteur directeur unitaire de D :

$$\vec{e}_D = \frac{p_2 - p_1}{\| p_2 - p_1 \|}, \quad \text{avec } \vec{e}_D = (u_1, u_2).$$

Pour un point M de coordonnées (x, y) dans le repère (e_{FS}, e_{SS}), on cherche ses nouvelles coordonnées $(\bar{x}, \bar{y})$ dans le repère (e_{FS}, e_D). On a l'égalité vectorielle :

$$x\, e_{FS} + y\, e_{SS} = \bar{x}\, e_{FS} + \bar{y}\, e_D.$$

En remplaçant $\vec{e}_D = u_1 e_{FS} + u_2 e_{SS}$, on obtient :

$$x\, e_{FS} + y\, e_{SS} = (\bar{x} + u_1 \bar{y})\, e_{FS} + (u_2 \bar{y})\, e_{SS}.$$

Par identification des coefficients, on en déduit :

$$\bar{y} = \frac{y}{u_2}, \qquad \bar{x} = x - \frac{u_1}{u_2} y.$$

Ainsi, les nouvelles coordonnées de M dans le repère (e_{FS}, e_D) sont données par :

$$(\bar{x}, \bar{y}) = \left(x - \tfrac{u_1}{u_2} y,\ \tfrac{y}{u_2} \right).$$

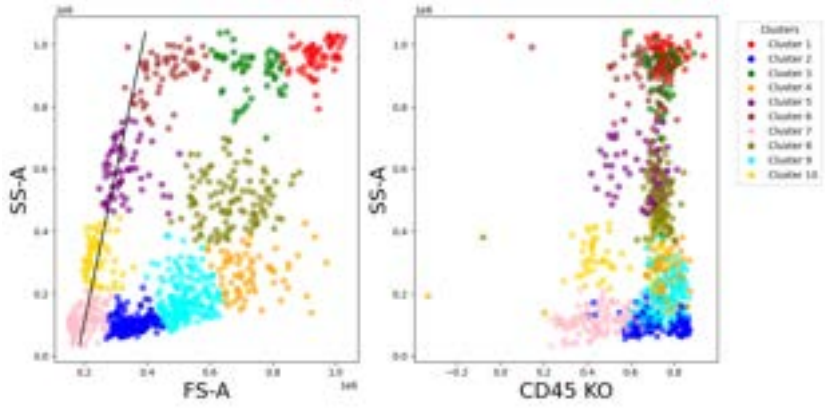

FIGURE 21 – Centroïdes des 100 patients considérés dans les plans (FS,SS) et (CD45,SS), où les couleurs correspondent aux meta-centroïdes obtenus par 10-meta-clustering sur les marqueurs SS-FS-CD45, et où la droite passe par les meta-centroïdes des deux clusters de plus bas FS.

La représentation dans ce nouvel espace met clairement en évidence un allongement marqué sur le plan (FS-SS) des événements associés aux débris :

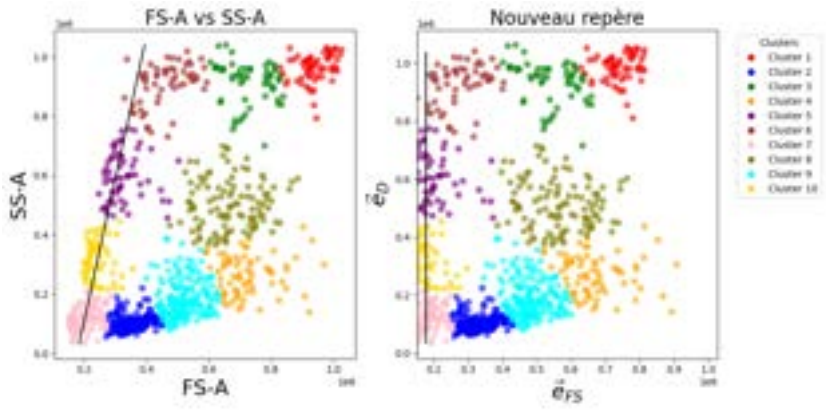

FIGURE 22 – Projection des événements dans le nouveau repère.

Les débris apparaissent désormais alignés le long de l'axe $\vec{e}_D$. Le redressement des données améliore leur lisibilité et facilite la séparation algorithmique. Cette transformation ouvre ainsi la voie à un nouveau clustering K-Means pour tester si les débris peuvent être regroupés dans une seule composante.

3.1.3 Sélection automatique des clusters à retirer

Nous appliquons d'abord un k-means à quatre classes uniquement sur la nouvelle coordonnée en `FS` des centroïdes dans le repère (FS,D). Le résultat (Figure 23) montre des méta-clusters allongés et obliques dans le plan (FS,SS), ce qui traduit bien l'effet du changement de repère. Le cluster de plus faible `FS` moyen (cluster 3 en vert) correspond à la zone morphologique recherchée. Toutefois, certains regroupements en `SS` faible peuvent inclure des populations de blastes, notamment celles à `CD45` intermédiaire.

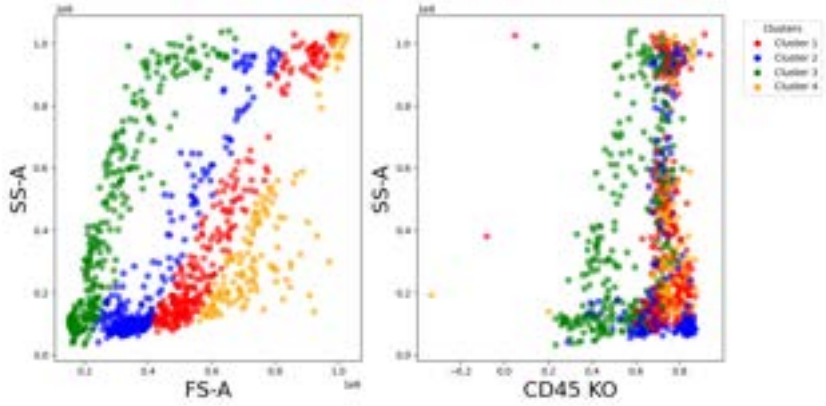

FIGURE 23 – Clustering en quatre classes appliqué sur `NewFS`.

Une alternative consiste à ajouter une dimension immunologique en combinant `NewFS` avec `CD45` dans un k-means à quatre classes. Comme illustré en Figure 24, cette approche ne permet pas de lever les confusions : les clusters de blastes potentiels restent mélangés et la zone morphologique ciblée n'est plus portée par un méta-cluster homogène.

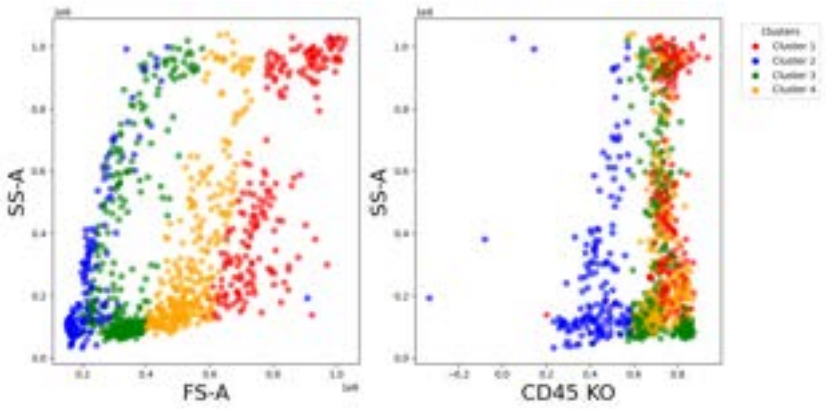

FIGURE 24 – Clustering en quatre classes appliqué sur `NewFS` et `CD45`.

Nous retenons donc la première approche, fondée uniquement sur `NewFS`, et proposons d'en examiner les résultats directement au niveau des clusters locaux (patient par patient).

3.1.4 Résultats

Nous visualisons ensuite quelques patients afin de vérifier que la zone ciblée correspond bien au méta-cluster retenu (le cluster vert de la Figure 24, le plus fiable en FS). Les clusters locaux sont colorés selon leur appartenance à un méta-cluster, et ceux associés à des moyennes particulièrement faibles en `FS` ou en `CD45` — interprétés comme des débris — sont mis en évidence par un encadré noir.

Cette représentation permet d'évaluer, patient par patient, comment le méta-cluster des débris se manifeste dans les données individuelles.

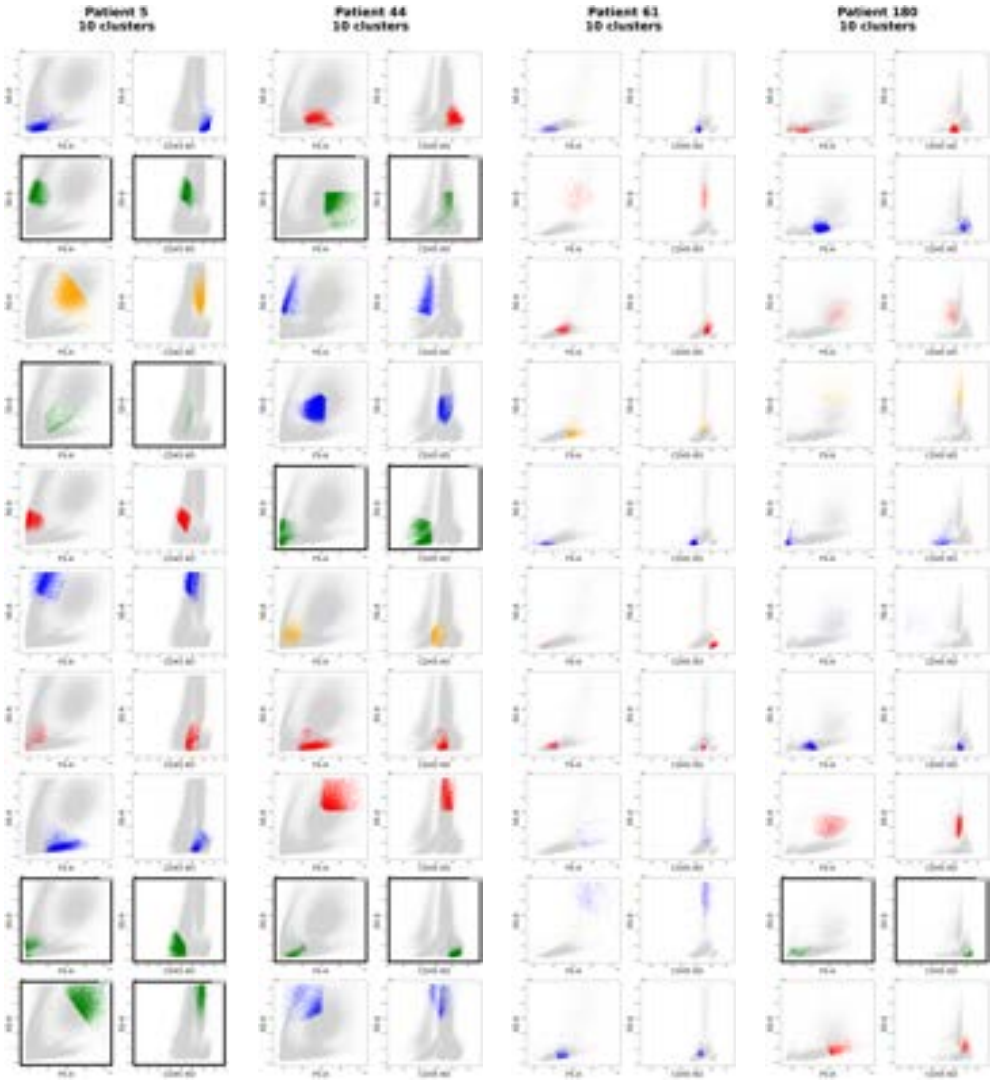

FIGURE 25 – Représentation des clusters locaux dans les plans FS-SS et CD45-SS pour quatre patients (5, 44, 61 et 180, de gauche à droite). Les clusters associés au métacluster identifié comme débris sont encadrés en noir.

Conformément à la pratique des cytométristes, le méta-méta-cluster vert correspondent bien à des groupes à supprimer. Toutefois, on constate que certains d'entre eux regroupent également des cellules `CD45+`. Par exemple, le cluster 2 du patient 44 et le cluster 10 du patient 5 contiennent de nombreux blastes. Cela illustre la limite de l'approche : en zone de `SS` faible, certains clusters à petit `FS` mais à `CD45` intermédiaire peuvent en réalité correspondre à des blastes.

À l'inverse, une sélection manuelle aurait parfois conduit à retenir des clusters mieux adaptés. C'est le cas, par exemple, du cluster 5 du patient 180, qui, bien que rattaché au deuxième plus faible niveau de `FS`, semble morphologiquement pertinent à conserver.

La sélection fondée uniquement sur des critères morphologiques présente une limite importante : elle peut conduire à exclure certaines cellules `CD45+`. L'objectif est donc désormais de cibler spécifiquement les cellules `CD45-` ou faiblement exprimées.

3.2 Détecter la zone de CD45 faible au sein d'une population

On choisit désormais de détecter les débris sur la base du marqueur `CD45`. Pour cela, nous réalisons un second niveau de clustering sur les centroïdes, cette fois en incluant `CD45` en plus de `SS` et `FS`.

10-means sur SS, FS et CD45 La Figure 26 présente les clusters locaux des 100 patients dans les plans (FS,SS) et (CD45,SS), colorés selon les méta-clusters obtenus par un 10-means appliqué aux trois marqueurs. La séparation est moins visible dans le plan morphologique (FS–SS), mais beaucoup plus marquée dans le plan (CD45–SS).

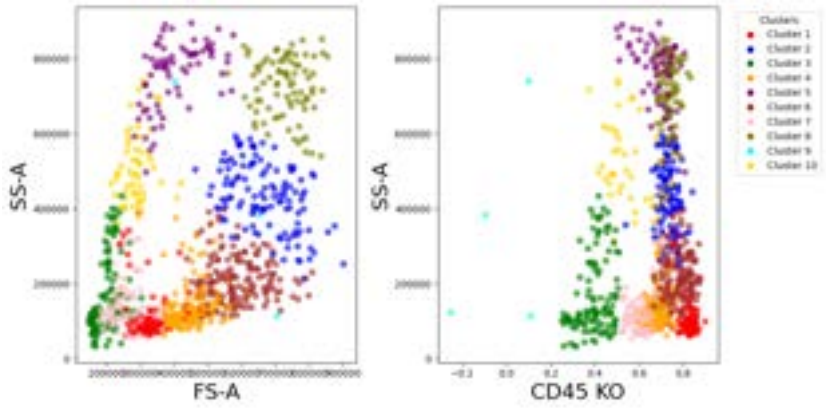

FIGURE 26 – Centroïdes des clusters locaux des 100 individus considérés, représentés dans les plans (FS,SS) et (CD45,SS). Les couleurs correspondent aux méta-clusters obtenus par 10-means sur (SS,FS,CD45).

Avec ce découpage, les séparations apparaissent nettement dans le plan (CD45–SS). Sur recommandation d'Aguirre, le cluster 7 (rose clair) est écarté. Nous retenons donc les trois méta-clusters présentant les intensités de `CD45` les plus faibles. Les outliers en `CD45` sont par ailleurs regroupés dans un méta-cluster distinct, plus cohérent. En revanche, certains méta-clusters restent mélangés dans le plan (FS–SS) — notamment les 3, 7 et 1 — mais leur distinction est bien capturée par `CD45`.

L'étape suivante consiste à définir une stratégie automatique pour sélectionner les méta-clusters d'intérêt (3 en vert, 10 en jaune et 9 en cyan). Pour cela, nous utilisons un méta-clustering fondé directement sur les 3 valeurs les plus faibles de `CD45` des méta-centroïdes.

3.2.1 Résultats

La Figure 27 illustre la représentation des clusters locaux pour quatre patients (5, 44, 61 et 180). Les clusters correspondant au méta-cluster identifié comme débris sont encadrés en noir. On observe que cette approche permet déjà une séparation plus nette que les méthodes précédentes, notamment dans le plan (CD45–SS).

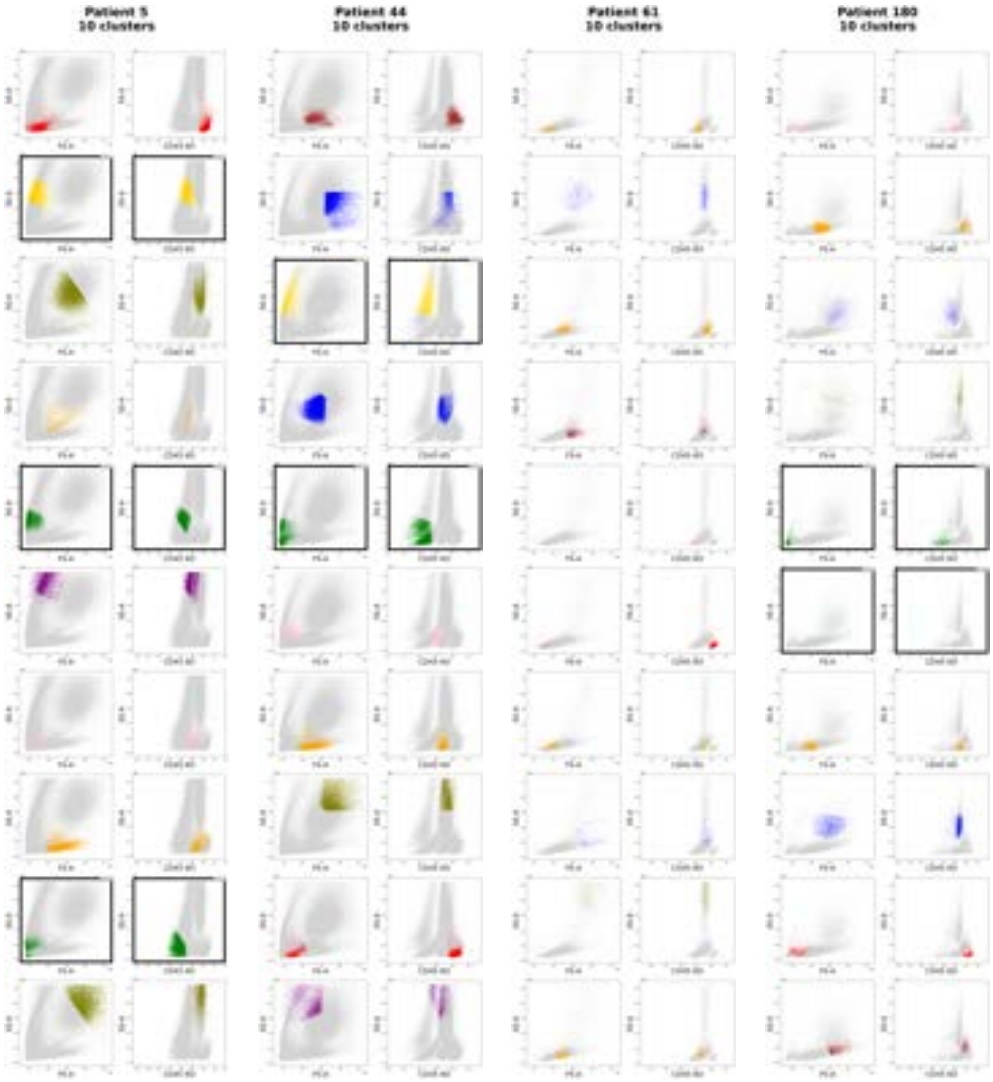

FIGURE 27 – Représentation des clusters locaux dans les plans (FS–SS) et (CD45–SS) pour quatre patients (5, 44, 61 et 180, de gauche à droite). Les clusters associés au méta-cluster identifié comme débris sont encadrés en noir.

Cependant, une limite subsiste : dans le plan (FS–SS), les clusters associés aux valeurs les plus élevées de SS restent mal isolés. Ces événements seraient en principe capturés par le cluster 5 (violet) de la Figure 26, mais ce méta-cluster est difficile à cibler directement. En effet, si l'on retient simplement les quatre valeurs les plus faibles de CD45, on risque de sélectionner par erreur le cluster 7 (rose clair), qui contient de nombreux blastes et qu'il convient donc d'écarter.

Plusieurs pistes peuvent être envisagées pour surmonter cette difficulté. La première consisterait à définir un nouveau changement de repère, cette fois dans le plan (CD45–SS), afin de mieux isoler la zone des hautes valeurs de SS. La seconde serait de supprimer au préalable les outliers extrêmes (par exemple à l'aide d'un filtrage basé sur le z-score), puis de réappliquer un k-means avec un nombre réduit de classes (3, 4 ou 5). Cette itération pourrait permettre de capturer plus efficacement cette zone morphologique, tout en évitant l'inclusion des blastes.

4 Conclusion

Ce travail s'inscrit dans un contexte où la cytométrie en flux joue un rôle central dans le diagnostic de la leucémie aiguë myéloblastique. L'un des principaux enjeux réside dans l'identification et l'exclusion des débris cellulaires et cellules mortes. Aujourd'hui, cette étape repose encore sur l'expertise manuelle des cytométristes, ce qui soulève des questions de temps et de reproductibilité. L'objectif de ce stage était donc de proposer des pistes méthodologiques pour automatiser ce processus, en s'appuyant sur des techniques non supervisées.

Une première approche reposait sur une norme d'expression membranaire. L'hypothèse était que les débris et cellules mortes présentaient un degré d'expression plus faible. À l'aide d'un ajustement par des mélanges de gaussiennes, il a été possible d'isoler une partie de ces événements. Toutefois, cette méthode reste insuffisante, car certaines cellules viables sont détectées comme étant des débris. Néanmoins, cette piste apparaît intéressante.

Une secondes approche a été explorée via des méthodes de clustering par K-means. Dans un premier temps, un découpage a été réalisé patient par patient, en se basant sur les deux marqueurs morphologiques et le CD45, afin de préserver les blastes. Ces clusters locaux se sont révélés globalement cohérents. Dans un second temps, une méthode de regroupement des centroïdes au niveau global a été proposée pour traiter la variabilité inter-patients. Deux variantes ont été étudiées. La première s'appuie sur l'identification de la zone morphologique des débris dans un repère redressé, où un nouveau K-means permet de sélectionner automatiquement les clusters à éliminer. Cette approche fonctionne globalement bien, mais présente encore des risques de sélectionner à tort des populations de CD45 positifs à faible SS, ce qui nécessitera des validations supplémentaires, notamment en cas de changement de jeu de données. La seconde variante, centrée sur le marqueur CD45, propose une sélection plus stricte des clusters de faible intensité en CD45, ce qui apparaît plus rassurant pour préserver les blastes. Elle repose toutefois sur un seuil arbitraire (les trois plus faibles clusters), qui devrait être généralisé, par exemple via la gestion des valeurs extrêmes en amont (z-score) ou un redressement analogue sur CD45.

Ce travail constitue une première étape vers une automatisation fiable du nettoyage des données en cytométrie en flux. Les résultats obtenus montrent que les approches testées sont prometteuses mais encore perfectibles. La combinaison des approches basées sur la norme et sur le clustering apparaît comme une piste particulièrement intéressante pour améliorer le pipeline.

De plus, une validation plus rigoureuse devra être entreprise, par exemple à travers des procédures de validation croisée sur un nombre plus large de patients.

Références

[1] *Understanding Flow Cytometry*, https://www.youtube.com/watch?v=sfWWxFBltpQ

[2] *Principles of Cytometry*, https://www.youtube.com/watch?v=ccR5snuCE80

[3] Fondation sur la recherche contre le cancer *Cancer : les traitements et les soins de support* https://www.fondation-arc.org/cancer/cancer-les-traitements-et-les-soins-de-support

[4] Camilla Paleari (2023). *Deep learning approaches for FLT3 gene mutation prediction in acute myeloid leukemia*. Thèse de Master.

[5] *Wikipedia Hématopoiese* https://fr.wikipedia.org/wiki/HÃľmatopoÃírÃÍse

[6] Laboratoire ADMED (2021) *Immunophénotypage des Lymphocytes* https://admed.ch/wp-content/uploads/admedINFO-labo-112.pdf

[7] Léa Comin (2022) *Analyse de données de cytométrie en flux pour la leucémie aiguë myéloblastique*. Rapport de stage Master 1

[8] Stemcell Technologies *Considerations for Flow Cytometry Gating* https://www.stemcell.com/considerations-for-facs-gating.html

[9] Joshua D Frenster, Dimitris Placantonatis (2018) *Bioluminescent In Vivo Imaging of Orthotopic Glioblastoma Xenografts in Mice*. Research Gate https://www.researchgate.net/figure/Flow-cytometric-isolation-of-GSCs-a-Exclusion-of-debris-b-c-Gating-for-single_fig10_322879250