

Projet long dans le cadre du Master 2 biologie-informatique

Etude de l'impact de MIF (macrophage migration inhibitory factor) sur la différenciation des progéniteurs myéloïdes, dans la Leucémie Myélomonocytaire Chronique (LMMC), par analyse single-cell RNAseq

Projet réalisé sous l'encadrement de

Marine AGLAVE

Préparé par

Maha GRAA

2022/2023



Sommaire

I - Introduction :	3
1. État de l'art :	3
2. Interaction TET2 et MIF :	3
3. But de l'étude :	3
II - Matériels et méthodes :	4
1. Matériels :	4
2. Méthodes :	4
III - Résultats et discussion :	6
1. Etude de la qualité des reads, Alignement et Génération de la table de comptage :	6
2. Etude de la qualité des Droplets :	9
3. Premiers filtrages et suite du contrôle-qualité:	12
4. Analyses individuelles :	14
5. Analyses groupées :	21
6. Analyses intégrées :	22
7. Étude des proportions cellulaires	27
8. Pour aller plus loin :	28
V - Lien GitHub Scripts R :	35
VI - Bibliographie :	36

I - Introduction :

1. État de l'art :

La leucémie myéloïde chronique (LMMC) est une pathologie rare qui affecte des sujets âgés et dont le pronostic est sombre. Elle atteint l'adulte avec une médiane vers 70 ans (moins de 10 % des patients ont moins de 60 ans), et touche deux fois plus les hommes que les femmes. Son incidence est de 1 à 2 pour 100 000 habitants par an. Elle est liée à une anomalie clonale acquise de la cellule souche hématopoïétique médullaire.

Elle est classée actuellement comme une hémopathie chronique parmi les syndromes dit « frontières », placée entre les syndromes myélodysplasiques (affecte les cellules souches myéloïdes) et les syndromes myéloprolifératifs (sur-prolifération de cellules myéloïdes), car elle réunit les symptômes de ces 2 syndromes. Cette pathologie n'a été considérée que récemment comme une entité à part entière.

Avant 2008, le diagnostic, les scores pronostiques clinico-biologiques et la prise en charge thérapeutique utilisés pour la LMMC étaient ceux des syndromes myélodysplasiques. Depuis 2008, la LMMC est reconnue comme une entité distincte et les recherches sur le diagnostic, le pronostic et le traitement de cette pathologie ont beaucoup progressé grâce à l'établissement de scores clinico-biologiques spécifiques, aux études en génétique moléculaire et en cytométrie en flux [1].

Malgré ces avancées, le diagnostic de LMMC demeure, aujourd'hui encore, un diagnostic d'exclusion, donc difficile à établir. En pratique, le diagnostic repose sur la chronicité de la monocytose (une monocytose supérieure à 1G/L de sang, supérieure ou égale à 10 % des globules blancs et persistant plus de 3 mois).

Les traitements habituels sont palliatifs et symptomatiques. Seule l'allogreffe de moelle a un potentiel curatif, mais sa réalisation est limitée compte tenu de l'âge des patients. [9]

2. Interaction TET2 et MIF :

La protéine "TET2 ten-eleven-translocation 2" codée par le gène TET2 (Tet Methylcytosine Dioxygenase 2), est une méthylcytosine dioxygénase qui catalyse la conversion de la méthylcytosine en 5-hydroxyméthylcytosine. Cette protéine est impliquée dans la myéloïèse (formation de la moelle osseuse). Des mutations au niveau du gène TET2 ont été associées à plusieurs troubles myéloprolifératifs. (RefSeq, Mar 2011) [2]

Le gène MIF (macrophage migration inhibitory factor) code pour une lymphokine impliquée dans l'immunité à médiation cellulaire, l'immunorégulation et l'inflammation. Il joue un rôle dans la régulation de la fonction des macrophages dans la défense de l'hôte, par la suppression des effets anti-inflammatoires des glucocorticoïdes. [provided by RefSeq, Jul 2008] [2].

Il est localisé à la fois dans les compartiments extracellulaires et intracellulaires et interagit avec plus d'une douzaine de protéines de surface cellulaire et intracellulaire différentes. MIF a été montrée pour jouer un rôle important dans la croissance de tumeurs et de métastases *in vivo*, par induction d'une sous-population immunosuppressive des cellules dérivées des myéloïdes. [10]

3. But de l'étude :

Les premiers résultats de l'équipe montrent que dans les CMMI avec une mutation du gène TET2, qui mène à une diminution de la production de la protéine TET2, on constate une augmentation de l'expression du gène MIF et une sécrétion excessive de sa protéine.

→ Les CMMI étant caractérisées par une monocytose, on souhaite étudier l'impact de l'expression excessive de MIF sur la différenciation des progéniteurs myéloïdes.

II - Matériels et méthodes :

1. Matériels :

On a utilisé deux échantillons provenant de la culture des cellules souches CD34+ du cordon ombilical (cellules collectées au 7ème jour de la culture):

- Échantillon CTRL (condition contrôle): cellules CD34+ infectées avec le lentivirus SCR-shRNA-GFP.
- Échantillon MIF (condition de répression de l'expression de TET2) : cellules CD34+ infectées avec le lentivirus TET2-shRNA-GFP.

Les cellules ont été cultivées dans un milieu contenant du stem cell factor (SCF), de l'interleukin-3 (IL-3), du Fmsrelated tyrosine kinase 3 ligand (FLT3L) et du granulocyte-colony stimulating factor (G-CSF).

2. Méthodes :

a- Introduction aux analyses Single-cell RNA-seq:

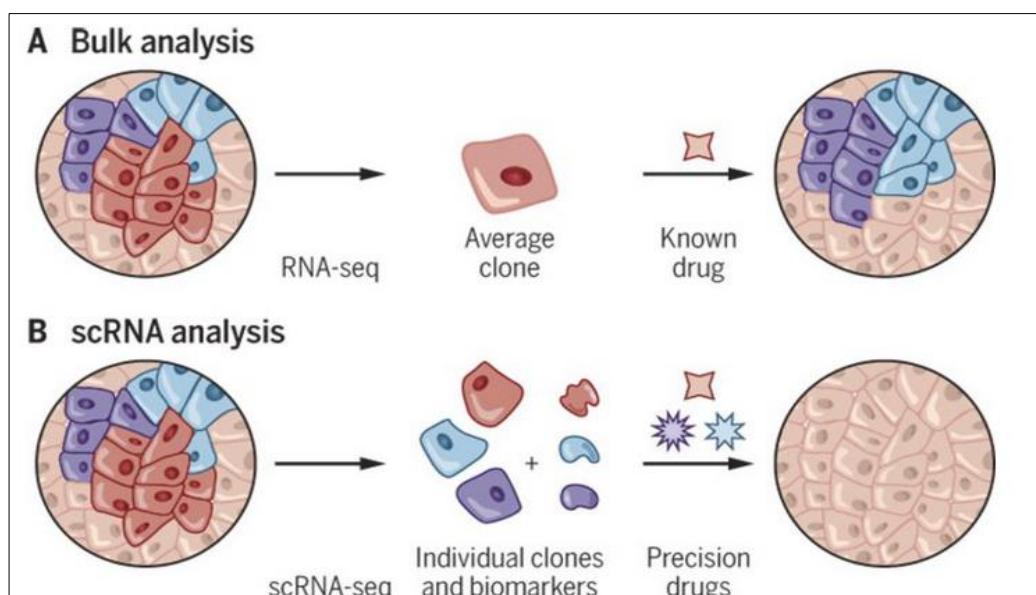
Les cellules représentent l'unité fonctionnelle minimale de la vie. Un objectif majeur de la biologie est de comprendre les mécanismes opérant dans cette unité minimale. De nos jours, l'analyse de la cellule unique peut être effectuée à une résolution sans précédent à l'aide de nouveaux dispositifs de laboratoire.

Le séquençage d'ARN unicellulaire (dit « single-cell RNA-seq » ou « scRNA-seq ») analyse les profils d'expression génique de cellules individuelles dérivées de populations homogènes et hétérogènes. Cette technique isole des cellules, généralement par encapsulation ou par cytométrie en flux, puis amplifie et séquence l'ARN de chaque cellule séparément. Cette approche à haute résolution permet aux chercheurs d'identifier les types de cellules, les états et les sous-populations. Le scRNA-seq peut également révéler une hétérogénéité cellulaire et des populations de cellules rares qui pourraient être masquées dans les données de séquençage d'ARN en vrac (dit « bulk RNA-seq ») [11].

Shalek et Benson [4] illustrent ce fait avec la figure ci-dessous :

(A) L'analyse en bulk d'une tumeur identifie le clone malin prédominant et suggère un médicament pour le cibler mais pas de médicaments pour les autres clones, ce qui va permettre leur développement.

(B) Le scRNA-seq permet d'identifier chaque clone dans la tumeur, ainsi que les biomarqueurs correspondants et les médicaments apparentés, permettant une thérapie réussie.



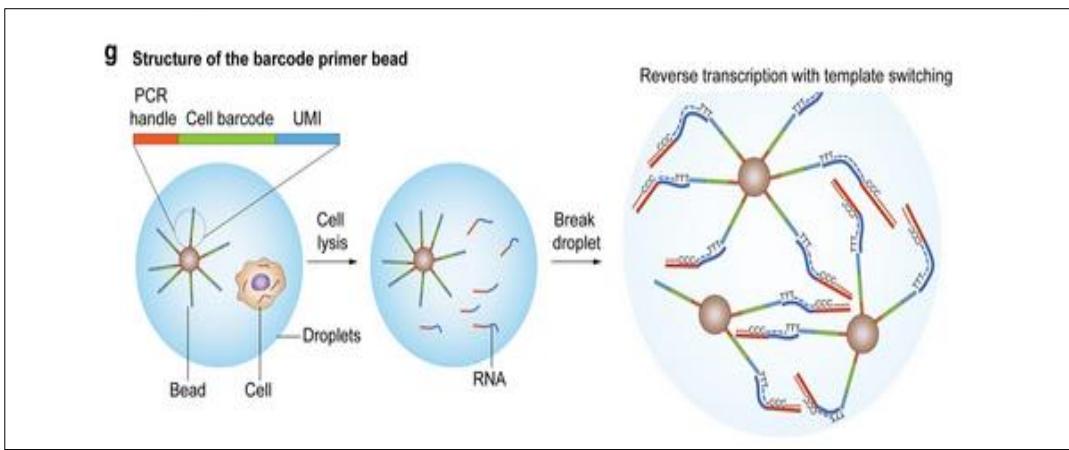
b- Les expériences biologiques :

Les échantillons ont été traités par la Plateforme de Génomique de Gustave Roussy.

Préparation d'une suspension cellulaire : [5]

Workflow du protocole	Description des bonnes pratiques
Dissection du tissu	Lavage avec du PBS.
Éminçage	Utilisez un scalpel, une lame ou des ciseaux.
Digestion enzymatique	<ul style="list-style-type: none">- Déterminer la concentration optimale des enzymes.- Déterminer la température optimale pour la digestion.- Inclure la DNase-I et l'EDTA dans le cocktail de digestion.
Incubation	Utiliser un agitateur pour aider à la dissociation mécanique.
Filtration	Retirer les morceaux des tissus non digérés.
Centrifugation	Cellules à centrifuger à 300–900 RCF (Relative Centrifugal Force).
Resuspension	Pipeter doucement et éviter les bulles.
Évaluation de la préparation single-cell	Vérifier la viabilité cellulaire et évaluer la suspension cellulaire.

Les suspensions unicellulaires ont été chargées dans le Chromium Single Cell Chip de 10x Genomics, conformément aux instructions du fabricant. L'encapsulation est réalisée grâce à un système microfluidique. Ainsi chaque cellule est isolée dans une gouttelette (appelé aussi « droplet ») avec une bille de gel barcodée (voir figure ci-dessous). Les bibliothèques ont été générées par lyse cellulaire, hybridation des ARNm aux barcodes cellulaires grâce à la queue poly-dT des billes de gel, transcription inverse en ADNc du premier brin, synthèse du deuxième brin et amplification de l'ADNc. Au cours de la transcription inverse, le barcode cellulaire est inclus à l'ADNc permettant de retrouver la cellule d'origine des séquences. De façon similaire, une séquence d'identification des molécules (nommé UMI pour « Unique Molecular Identifiers ») est également incluse. En utilisant cette stratégie, chaque lecture peut être attribuée à sa cellule d'origine en supprimant efficacement le biais de la PCR et en améliorant ainsi la précision. [6]



Ce processus a été réalisé à l'aide du kit Chromium Next GEM Single Cell 3 'GEM, Library & Gel Bead Kit v3.1 (10X Genomics) conformément aux instructions du fabricant, avec un taux de capture cible d'environ 10 000 cellules par échantillon. Tous les échantillons ont été traités et séquencés simultanément. Le séquençage a été effectué sur un séquenceur Illumina NovaSeq 6000.

c - Les analyses bio-informatiques :

Grâce au pipeline single-cell RNA-seq [12] établi par les bioinformaticiens au sein de la Plateforme de Bio-informatique de Gustave Roussy (BiGR), on va analyser les données scRNA-seq des deux échantillons CTRL et MIF obtenus. Le pipeline est principalement basé sur le package R Seurat [13], encapsulé dans des conteneurs Singularity [14] et géré par le gestionnaire de workflow Snakemake. [15]

Les étapes du pipeline Single-cell RNA-seq pour chaque échantillon sont:

- Contrôle de la qualité des Reads.
- Alignement et Génération de la table de comptage.
- Contrôle de la qualité des Droplets.
- Filtrage des Droplets de mauvaise qualité.
- Normalisation, réduction de dimensions et clustering.
- Annotation

Afin de comparer les échantillons, ces étapes sont suivies par une analyse groupée où les échantillons sont assemblés sous la même représentation graphique. Ainsi les étapes de réduction de dimensions, de clustering et d'annotation sont appliquées à nouveau.

Cependant, un effet batch entre échantillons peut apparaître lors de l'analyse groupée. Donc il faut ajouter une étape supplémentaire permettant d'appliquer une correction de ce biais technique si besoin. Cette étape s'appelle l'intégration. Ainsi les étapes d'une analyse intégrée sont : correction d'effet batch, réduction de dimensions, clustering et annotation.

III - Résultats et discussion :

1. Etude de la qualité des reads, Alignement et Génération de la table de comptage :

Le contrôle qualité des reads est effectué par les outils Fastqc et Fastq-screen et un rapport html est généré par Multiqc. L'alignement et la génération de la table de comptage ont été réalisés par la suite d'outils Kallisto-bustools. Lors de cette étape, on doit régler plusieurs paramètres dont :

- Sctech : "10xv3" technologie 10x Genomics utilisée pour l'identification des barcodes cellulaires et des UMIs (La version "10xv2" dispose de 8 bases par barcode, permettant d'identifier jusqu'à 8 000 cellules uniques, tandis que la version "10xv3" dispose de 16 bases par barcode, permettant d'identifier jusqu'à 120 000 cellules uniques).
- kindex.ge : chemin vers l'index de référence de l'espèce homo sapiens, nécessaire pour l'alignement des séquences.

Statistiques générales sur les reads :

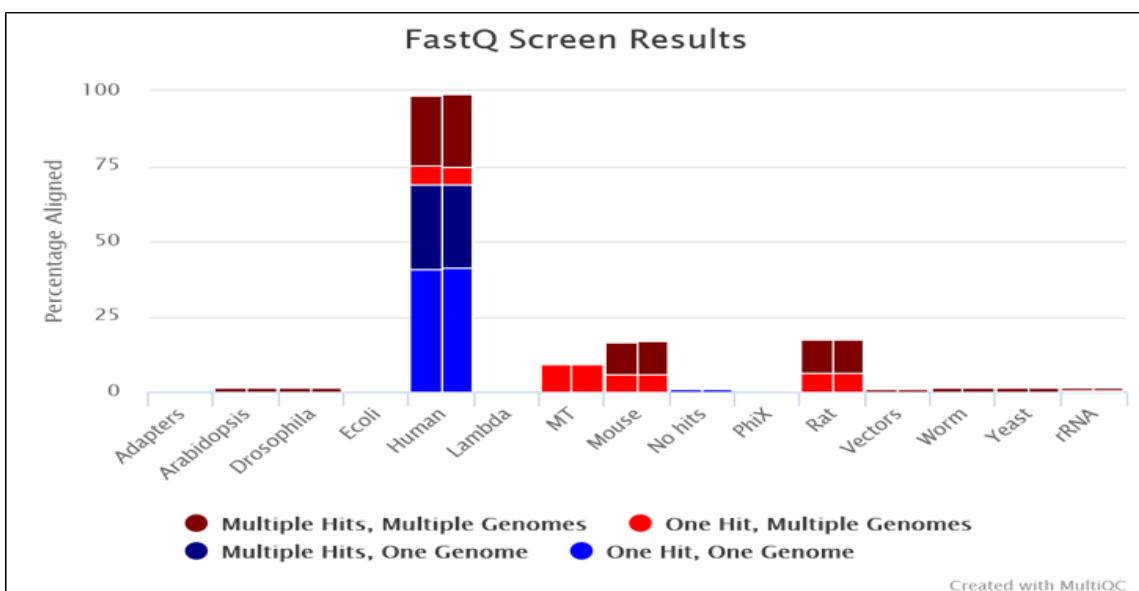
a- pour l'échantillon CTRL :

Résultats de Fastqc

General Statistics

Sample Name	% Dups	% GC	Length	M Seqs
CB50_CTRL_GE_S1_L001_R1_001	21.6%	47%	28 bp	123.6
CB50_CTRL_GE_S1_L001_R2_001	65.3%	47%	90 bp	123.6
CB50_CTRL_GE_S1_L002_R1_001	21.8%	47%	28 bp	123.7
CB50_CTRL_GE_S1_L002_R2_001	65.6%	47%	90 bp	123.7

Résultats de Fastq-screen



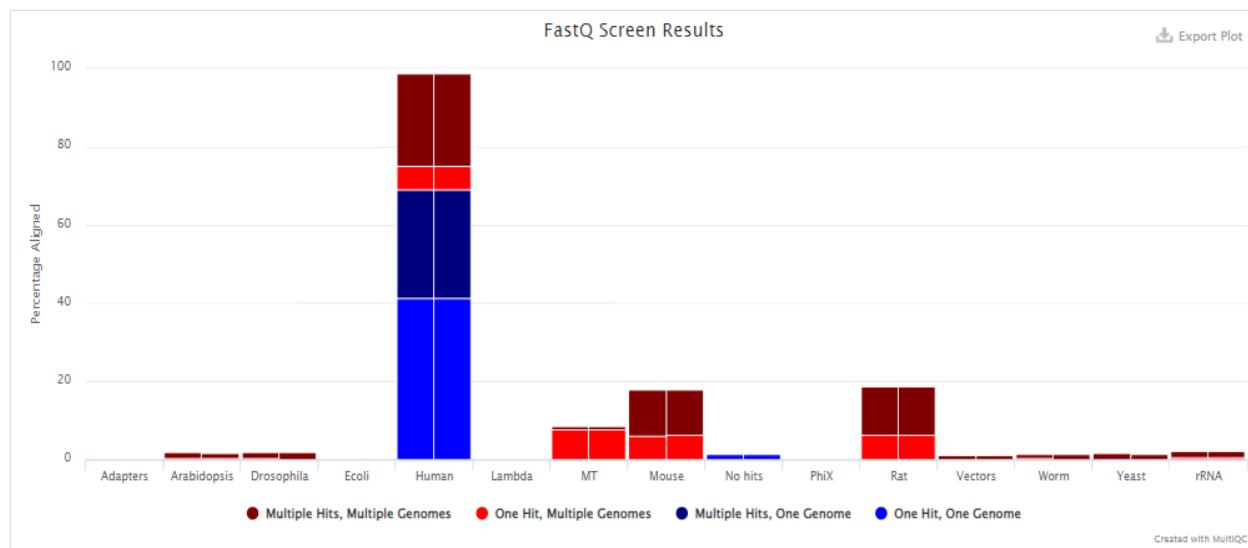
b- pour l'échantillon MIF :

Résultats de Fastqc

General Statistics

Sample Name	% Dups	% GC	Length	M Seqs
CB50_MIF_GE_S2_L001_R1_001	19.2%	48%	28 bp	118.8
CB50_MIF_GE_S2_L001_R2_001	65.8%	48%	90 bp	118.8
CB50_MIF_GE_S2_L002_R1_001	19.5%	48%	28 bp	118.9
CB50_MIF_GE_S2_L002_R2_001	66.0%	48%	90 bp	118.9

Résultats de Fastq-screen



Les pourcentages en GC (« % GC ») dans nos échantillons MIF et CTRL sont respectivement de 48 et 47 %, ce qui est cohérent avec des échantillons humains dont le pourcentage habituel est d'environ 50 %.

La taille des reads (« Length ») est identique selon les fichiers R1 et R2 et sont cohérents avec la technologie 10X utilisée.

Les nombres de séquences générées (« M Seqs ») sont similaires entre les échantillons.

Les pourcentages de séquences dupliquées sont de 21% pour les R1 et 66% pour les R2 pour l'échantillon CTRL, et de 19% pour les R1 et 66% pour les R2 pour l'échantillon MIF. Les échantillons sont donc très similaires. Un pourcentage élevé de séquence dupliqué n'est pas étonnant car la technique de 10X induit un biais lors de la PCR. En effet la polymérase présente plus d'affinité avec certains motifs et a tendance à dupliquer l'ADNc de façon non-homogène. Ainsi certaines séquences sont plus amplifiées que d'autres et les proportions d'ARNm initiales sont modifiées. Ce biais est corrigé lors de l'alignement grâce aux UMI qui ont été inclus avant la transcription inverse. Ainsi, comme la séquence en R1 correspond aux barcodes cellulaires et aux UMI, nous pouvons estimer qu'il y a environ 20 % des reads qui ont été dupliqués par ce biais, car l'association barcode-UMI est censé être unique. Les R2 correspondant aux ARNm d'intérêts, ils ont également 20 % des séquences qui sont dupliqués à cause du biais de PCR (puisque il s'agit du même ADNc que pour les barcodes-UMI) mais les 45 % restant sont issus des différentes copies d'ARNm pour un même gène, présentes naturellement dans une cellule.

De plus, fastqc a permis de vérifier la qualité des bases (proportion de chaque type de bases, nombre de base N etc), ainsi que l'absence des adaptateurs utilisés pour le séquençage par Illumina.

L'outil Fastq-screen permet de vérifier l'espèce d'un échantillon en alignant un échantillonnage des reads sur plusieurs références prédéfinies. Ces résultats nous confirment que nos échantillons sont humain.

Statistiques générales de l'alignement :

Noms	Définitions	Valeurs MIF	Valeurs CTRL
n_processed	nombre de fragments traités	237690185	247316149
n_pseudoaligned	nombre de fragments pseudo-alignés	159502996	164038967
n_unique	nombre de fragments qui ont été pseudo-alignés sur une séquence cible unique	45079563	49568927
p_pseudoaligned	pourcentage des fragments pseudo alignés	67.1	66.3
p_unique	pourcentage des fragments pseudo alignés sur une séquence cible unique	19.0	20.0

Ces statistiques sont dans les normes habituelles pour des échantillons de scRNA-seq.

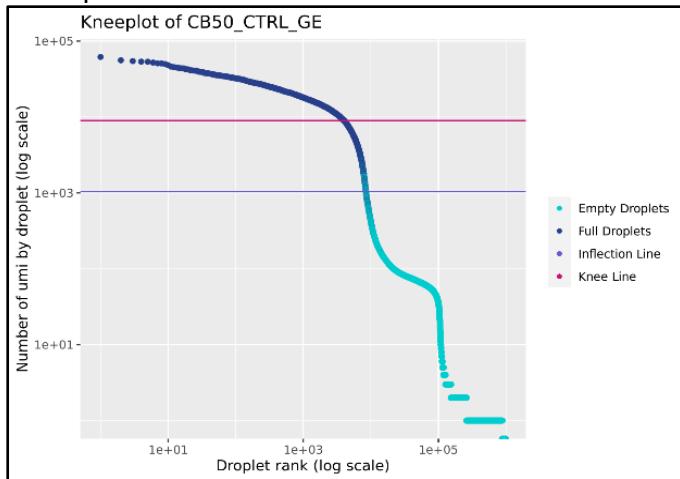
2. Etude de la qualité des Droplets :

Le but de cette étape est d'éliminer les droplets vides (contenant du bruit de fond) et de mesurer certaines statistiques de qualité des droplets.

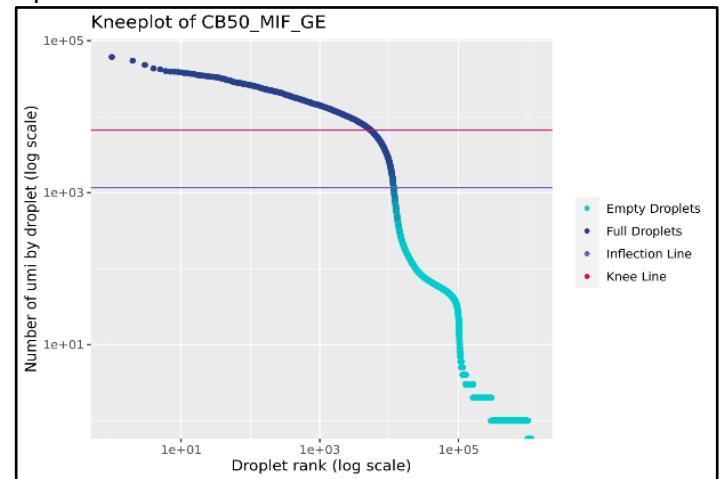
Droplets vides :

Pour éliminer les droplets vides, la bonne pratique est de tracer un kneeplot. Le kneeplot est un graphique représentant le nombre de molécules d'ARN (nombre d'UMI) en fonction du numéro de barcode cellulaire (trié par nombre de molécule d'ARN). Ainsi les droplets ayant le plus d'ARN se situent au niveau du plateau en haut sur la gauche du graphe, les droplets ayant peu d'ARN se situent sur et sous le plateau en bas à droite ; le graphe présentant une forme de double genou. La zone verticale entre les 2 plateaux correspond à des cellules de petite taille ou pas assez lysées.

a- pour l'échantillon CTRL :



b- pour l'échantillon MIF :

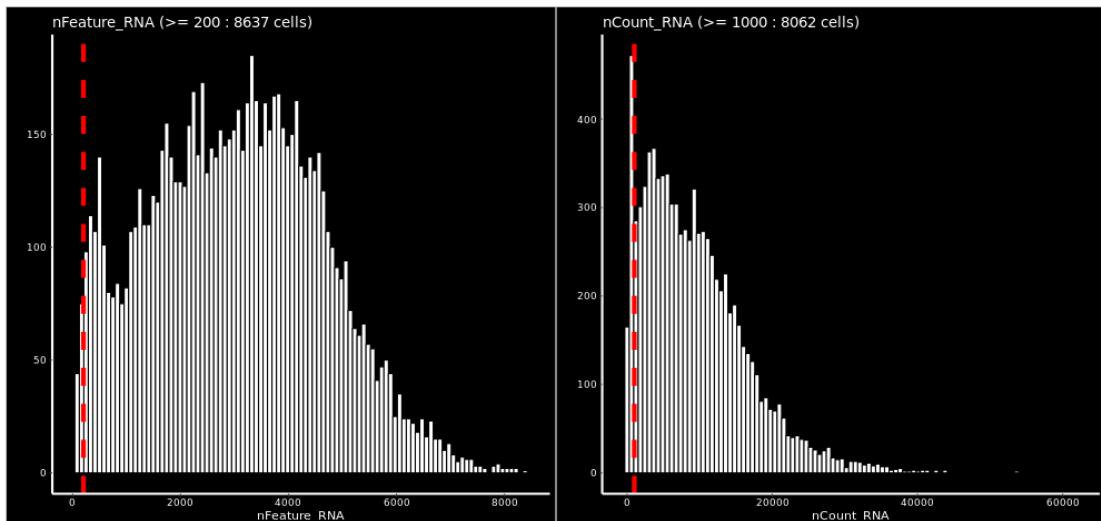


Le pipeline utilise l'outil EmptyDrops [16] pour identifier automatiquement les droplets à supprimer, mais nous pouvons choisir d'ajouter un seuil d'UMI minimum (paramètre min.counts, valeur par défaut de 1000 UMI), et un seuil de gènes minimum (paramètre min.features, valeur par défaut de 200 gènes), si on souhaite être plus stricte dans les filtres.

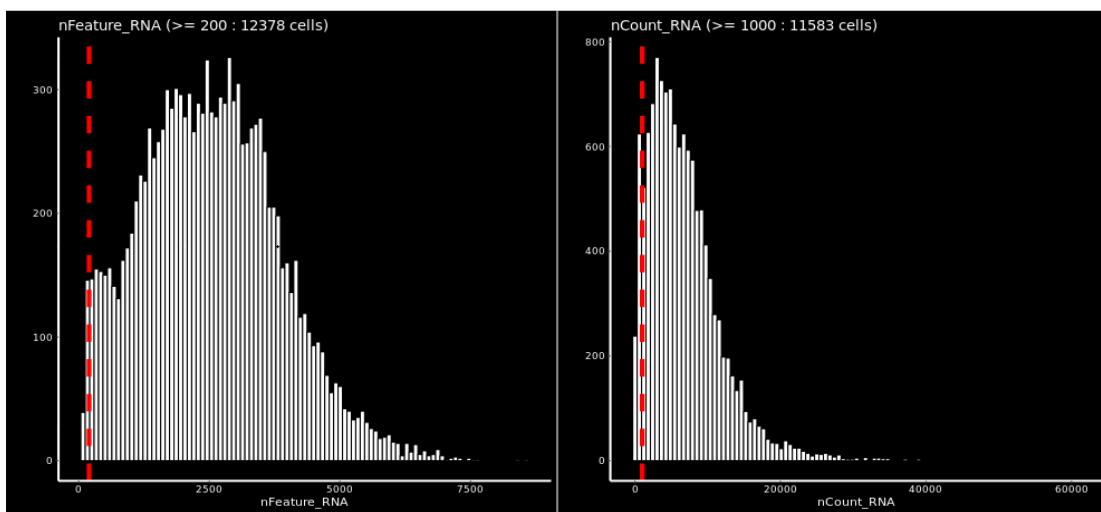
Pour les 2 échantillons, on constate que la frontière théorique des droplets pleines (genou du haut) est supérieur à ce qu'a identifié EmptyDrops ainsi qu'au seuil par défaut de 1000 UMI. C'est pour cette raison que j'ai décidé de filtrer à 1300 UMI, grâce au paramètre min.counts, afin de récupérer les droplets ayant un nombre d'UMI supérieur à 1300.

Les kneeplot étant en échelle log, afin de mieux visualiser le nombre de cellules qui seraient éliminées par les seuils par défaut, des histogrammes sont tracés :

a- pour l'échantillon CTRL :



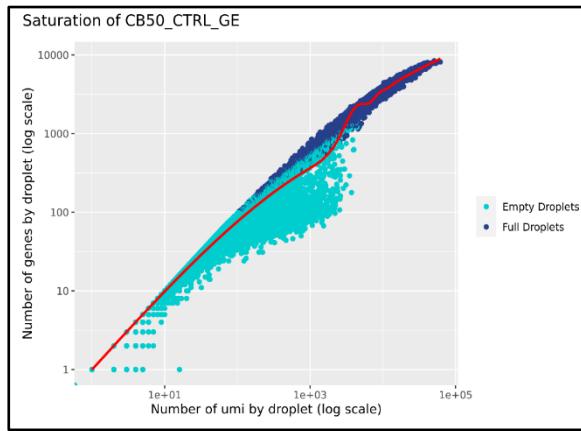
b- pour l'échantillon MIF :



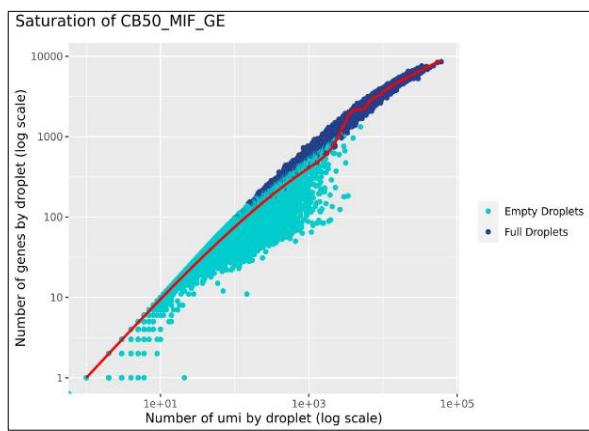
nFeature_RNA représente le nombre de gènes par droplet et nCount_RNA représente le nombre de d'UMI par droplet. L'axe des y représente le nombre de droplet pour chaque valeur de gènes ou d'UMI. Les lignes rouges en pointillées représentent les filtres par défaut qui seront appliqués lors de l'étape suivante dans le pipeline. Nous pouvons voir qu'avec les filtres par défaut de 200 gènes minimum ou 1000 UMI minimum, il resterait respectivement 8637 et 8062 droplets pour l'échantillon CTRL, et 12378 et 11583 droplets pour l'échantillon MIF. Ce nombre est cohérent avec les 10 000 cellules visées. De plus sur le kneeplot j'ai choisi d'augmenter le filtre sur les UMI à 1300 UMI, ce qui reste un seuil raisonnable d'après ces graphiques.

Saturation de séquençage :

a- pour l'échantillon CTRL :



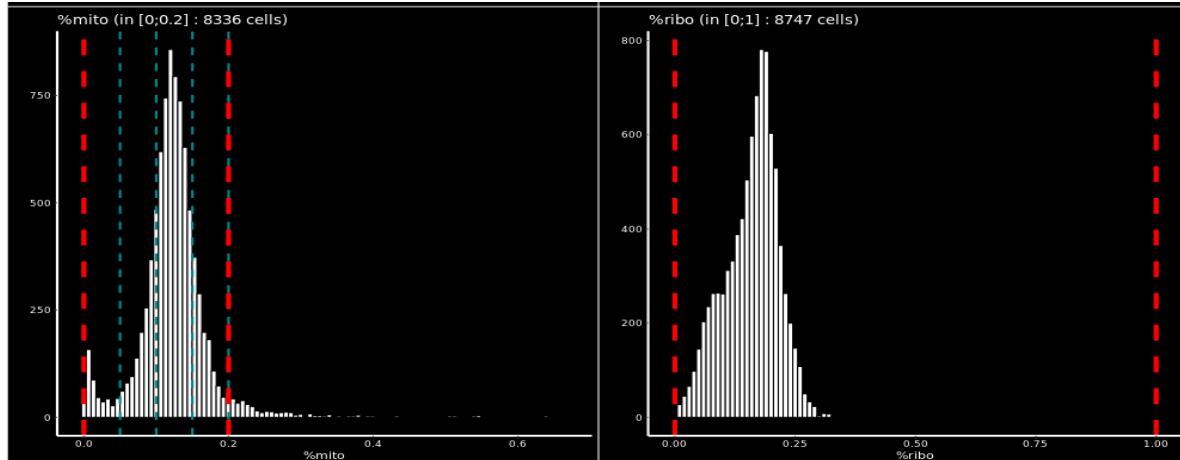
b- pour l'échantillon MIF :



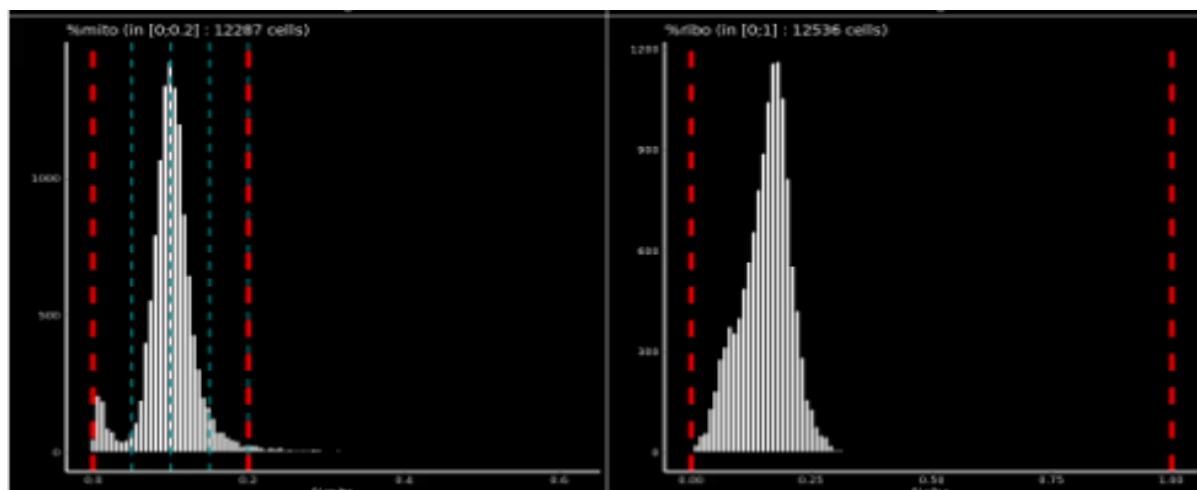
Ce graphique est une représentation du nombre de gènes par droplet en fonction du nombre d'UMI par droplet. Les droplets avec le moins de gènes et le moins d'UMI sont les droplets vides (contenant du bruit de fond). La ligne rouge est la courbe de tendance. On constate que le nombre de gènes et le nombre d'UMI sont corrélés. De plus, la courbe de tendance tend à former un plateau au sommet, ce qui montre que presque tous les gènes ont été découverts et donc que le séquençage est suffisamment profond.

Autres statistiques :Deux autres mesures sont habituellement faites dans le contrôle qualité des droplets. Il s'agit du pourcentage d'ARN mitochondriaux et du pourcentage d'ARN qui codent pour les protéines ribosomales. Comme pour le nombre de gènes et d'UMI, des histogrammes de ces mesures, pour chacun de nos échantillons, sont produits :

a- pour l'échantillon CTRL :



b- pour l'échantillon MIF :



D'après Daniel Osorio et James Cai [7], un nombre élevé de transcrits mitochondriaux est un indicateur de stress cellulaire, et donc le pourcentage d'ARN mitochondriaux est une mesure associée aux cellules apoptotiques, stressées et de mauvaise qualité. Le seuil habituellement recommandé est de 5% maximum, pourtant dans leur étude, ils démontrent que ce seuil devrait être adapté selon la situation. De plus, ce seuil est régulièrement établi à 20 % dans les publications.

Sur les histogrammes, nous pouvons voir que si nous plaçons ce seuil à 5 % nous éliminerons un nombre très important de droplets. C'est pour cette raison que pendant notre étude on a essayé de ne garder que les cellules qui ont un pourcentage de transcrits mitochondriaux <20%, pour ne pas limiter les analyses et ne se retrouver qu'avec un nombre très bas de cellules d'intérêt.

En ce qui concerne le pourcentage de transcrits qui codent pour les protéines ribosomales au sein des cellules, on n'applique pas de filtre au niveau du pipeline pour la même raison déjà citée ci-dessus qui est la limitation des analyses. De plus, un pourcentage élevé de transcrits des protéines ribosomales pourrait indiquer la présence d'une population de cellules avec une traduction active.

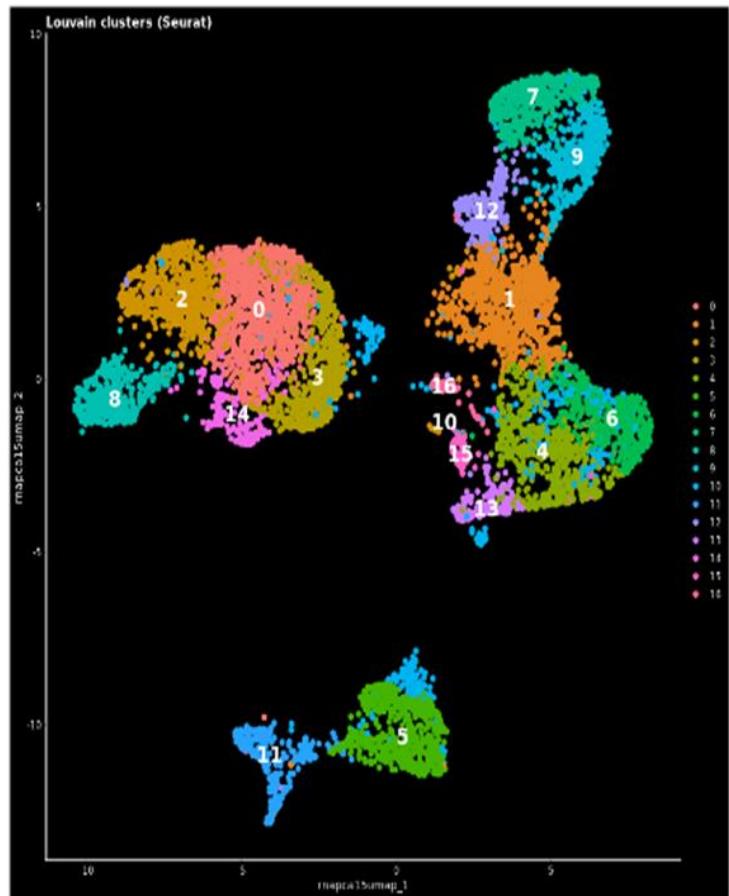
3. Premiers filtrages et suite du contrôle qualité:

Ainsi pour l'étape suivante du filtrage, on doit régler plusieurs paramètres, dont:

- min.counts = 1300 : en dessous de ce seuil, les droplets sont éliminées.
- pcmito.min = 0 : pas de pourcentage minimal de présence d'ARN mitochondrial exigé.
- pcmito.max = 20 % : au-delà de ce pourcentage de présence d'ARN mitochondrial, les droplets sont éliminées.

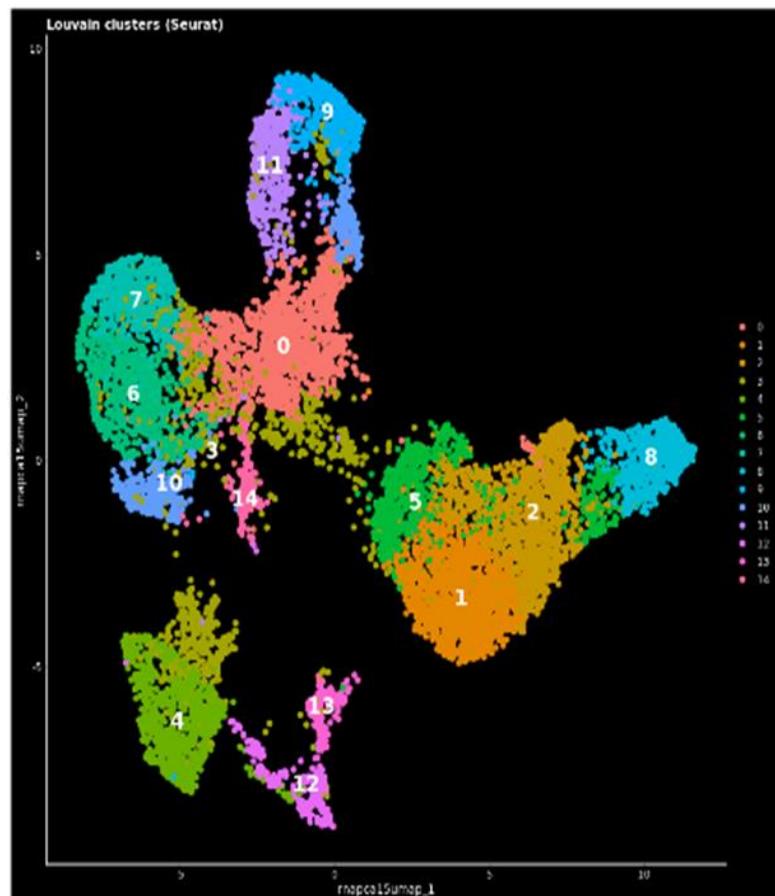
Ces premiers filtres ont permis de ne conserver que des droplets contenant des cellules de bonne qualité. Cependant il existe un phénomène inverse au droplets vide, nommé les doublets. Les doublets correspondent aux droplets ayant reçues 2 de cellules ou plus. Cela arrive dans toutes les manip' mais le nombre de doublets augmente avec le nombre de cellules injectées dans le Chromium selon la difficulté à séparer les cellules lors de la préparation de la suspension cellulaire. Cependant, les doublets contenant 2 cellules du même type cellulaire (contenant donc 2 fois plus de matériel génétique), peuvent être confondues avec des cellules en phase G2M du cycle cellulaire (qui ont multiplié par 2 leur matériel génétique). Ainsi il faut identifier les phases du cycle des droplets (par Seurat et Cyclone) et s'assurer que les outils permettant de détecter les doublets (scDoubletFinder[17]et scds [18]) n'identifient pas toutes les cellules en G2M comme étant des doublets.

a- Pour l'échantillon CTRL :

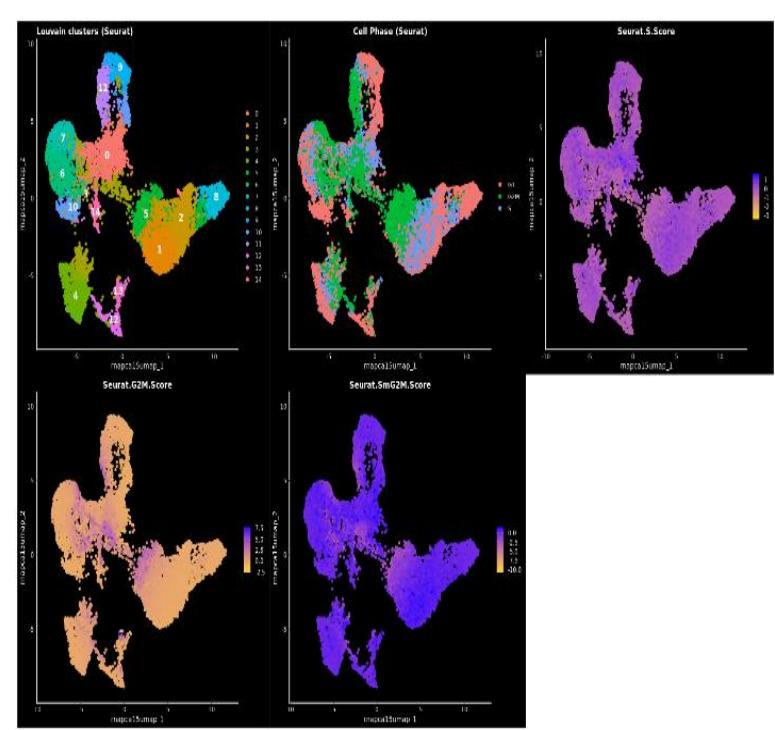
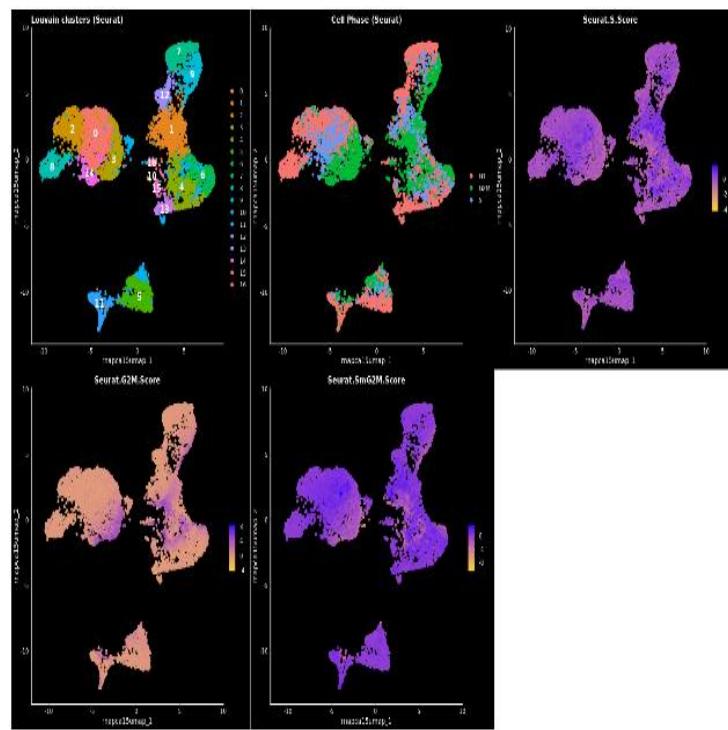


Cycle cellulaire Seurat

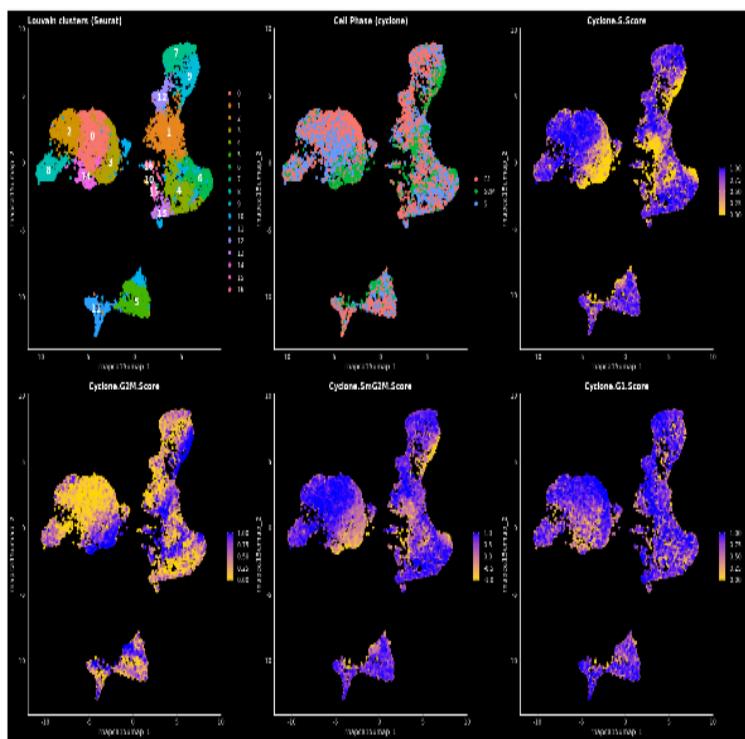
b- Pour l'échantillon MIF :



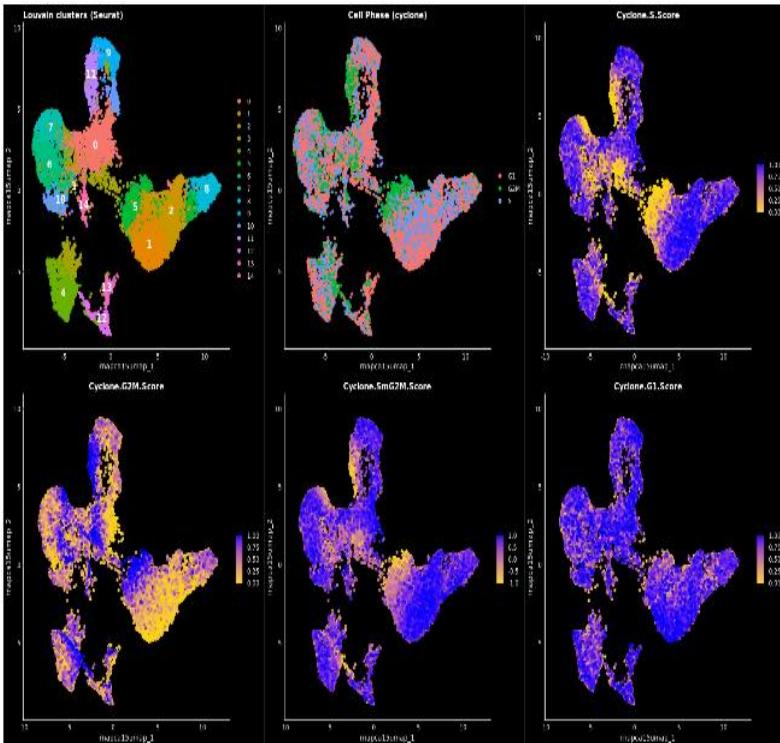
Cycle cellulaire Seurat



Cycle cellulaire cyclone



Cycle cellulaire cyclone



Le pipeline trace les cellules sur un graphe UMAP où chaque point correspond à une cellule. Le but est de vérifier que les cellules identifiées comme doubles ne correspondent pas entièrement aux cellules en phase G2M.

- Pour les 2 échantillons, les doubles ne correspondent pas entièrement aux cellules en phase G2M, donc nous pouvons les éliminer sans risque de perdre des cellules d'intérêt.

4. Analyses individuelles :

a- Normalisation, réduction de dimensions et clustering :

L'étape de la normalisation et de la réduction de dimensions est une étape cruciale de l'analyse de données transcriptomiques en single-cell.

Normalisation :

Les données transcriptomiques à cellule unique peuvent être biaisées en raison de différences dans la quantité d'ARN entre les cellules. La normalisation est une étape importante pour corriger ces biais et permettre une comparaison significative entre les cellules. La normalisation vise à ajuster les données pour réduire les effets des sources de variation non biologiques, tels que les biais dans la capture des molécules d'ARN ou les différences dans les taux d'expression des gènes.

La méthode la plus couramment utilisée pour normaliser les données transcriptomiques à cellule unique est proposée par l'outil Seurat. Cette méthode, nommée SCTransform, adapte un modèle binomial négatif régularisé, qui utilise la taille de la librairie comme seule variable explicative dans le modèle. Les résidus de Pearson de ce modèle sont utilisés comme valeurs d'expression normalisées et stabilisées par la variance.

Comme ce genre de calcul peut être long si on considère tous les gènes, Seurat passe par une étape de sélection de gènes les plus variables entre les cellules (sélection de 3000 gènes). En effet les gènes les plus variables permettent logiquement d'identifier les cellules qui se ressemblent le moins, et donc

de regrouper les cellules qui se ressemblent le plus, ce qui est le but principal d'une analyse scRNA-seq. De plus, nous pouvons donner à ce modèle des biais connu à corriger (comme le nombre de gènes exprimé, le pourcentage d'ARN mitochondriaux, les phases du cycle cellulaire etc).

Réduction de dimensions :

En scRNAseq, une cellule correspond à une dimension. Donc contrairement au RNAseq bulk (où un échantillon correspond à une dimension), ici nous travaillons en haute dimension (1 échantillon = 8000 cellules = 8000 dimensions). Chaque gène correspond également à une dimension, mais le nombre de gènes est identique dans les 2 méthodes. La haute dimension pose des problèmes de ressources nécessaires aux calculs, ainsi que des problèmes de visualisation des données. Ainsi il nous faut réduire ces données. Nous ne pouvons pas réduire les données sur les cellules car il s'agit de nos observations d'intérêt, mais nous pouvons nous concentrer sur un plus petit nombre de gènes. C'est ce qu'a commencé à faire SCTtransform lors de la normalisation avec la sélection des gènes les plus variables. Mais ces gènes les plus variables peuvent être combinés en eux pour former des composantes.

Pour cela on utilise l'Analyse en Composante Principale (abrégé en « ACP » ou « PCA » pour « Principal Component Analysis »), qui va nous permettre de combiner les gènes afin de former 50 composantes. Les premières composantes de l'ACP contiennent le plus d'informations biologiques pertinentes, et les dernières composantes contiennent majoritairement du bruit ou des variations d'expressions plus faibles. Ainsi nous pouvons choisir un nombre de composantes intermédiaires, permettant de récupérer un maximum d'information biologique utile et d'éliminer les composantes de bruit de fond.

Cependant même si on décide de ne garder que 20 composantes, on ne peut toujours pas les visualiser sur un graphe (l'humain étant limité à une vision en 3 dimensions). Donc il faut faire une réduction de dimensions sur les composantes de l'ACP afin d'obtenir que 2 dimensions. Pour cela on utilise une méthode nommé UMAP pour « Uniform Manifold Approximation and Projection ».

Clustering :

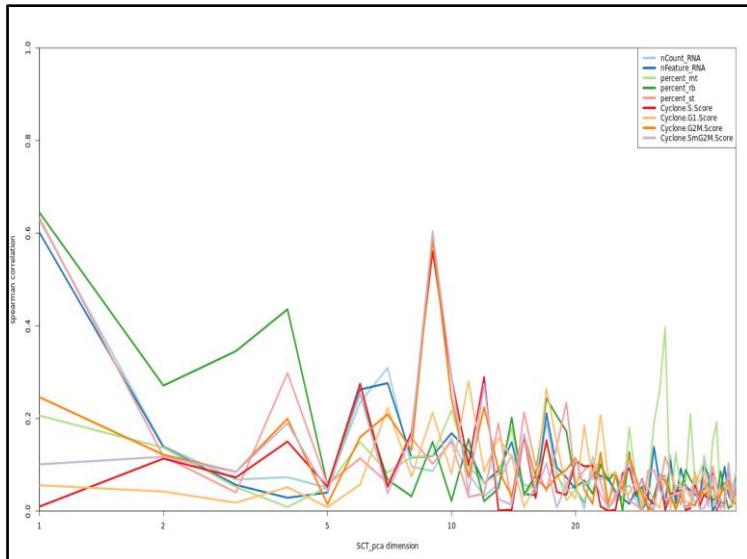
Une fois les cellules représentées sur un seul graphe, il faut pouvoir identifier les cellules similaires (appartenant au même type cellulaire) et les cellules différentes (appartenant à des types cellulaires différents). C'est le but du clustering. Le clustering est une méthode d'apprentissage non-supervisée qui est utilisée pour définir empiriquement des groupes de cellules avec des profils d'expression similaires. En single-cell, les méthodes utilisées sont les méthodes de graphes comme Louvain ou Leiden. Le pipeline utilise l'algorithme de Louvain qui est la méthode proposée par Seurat. Le paramètre principal de cette méthode est la résolution. Elle permet d'influencer le nombre de groupes de cellules à former. Plus la résolution est élevée, plus on forme de groupes. La formation de nouveaux groupes doit refléter les sous-types cellulaires. Ainsi les sous-types d'une résolution élevée, doivent être entièrement inclus dans le type cellulaire d'une résolution plus faible.

Résultats :

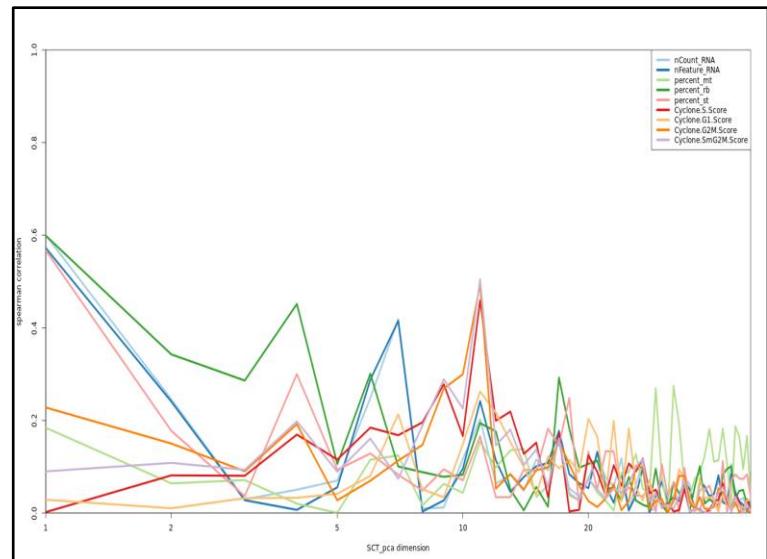
Comme vu précédemment, il est possible d'indiquer des biais techniques à corriger lors de la normalisation.

Pour identifier les biais à corriger, nous traçons l'impact de ces biais sur chaque composante de l'ACP. Les composantes devant être majoritairement corrélées avec l'expression de gènes permettant de séparer les cellules, il ne devrait pas y avoir de corrélation avec l'expression des gènes liés aux biais. Les biais mesurée sont le nombres de gènes par cellule (nFeature_RNA), le nombre de molécules d'ARN par cellule (nCounts_RNA), le pourcentage d'ARN mitochondriaux par cellule (percent_mt), le pourcentage d'ARN qui codent pour les protéines ribosomales par cellule (percent_rb), le pourcentage d'ARN lié au stresse mécanique du 10X par cellule (percent_st) et les différents scores des phases du cycle cellulaire par cellule (Cyclone.S.Score, Cyclone.G1.Score, Cyclone.G2M.Score, Cyclone.SmG2M.Score).

a- Pour l'échantillon CTRL :



b- Pour l'échantillon MIF :



Généralement, les biais qui dépassent 0.3 de corrélation doivent être corrigés. On constate dans ce cas que le pourcentage d'ARN qui codent pour les protéines ribosomales est bien corrélé avec les composantes, de même que le nombre de gènes, le nombre de molécules d'ARN et le pourcentage d'ARN liés au stress mécanique. Cependant les essais de corrections de ces biais ne semblent pas impacter les résultats. De manière générale, il vaut mieux éviter d'apporter des modifications aux données si elles ne sont pas pertinentes alors la suite des analyses est réalisée sans correction de biais.

Réduction des dimensions :

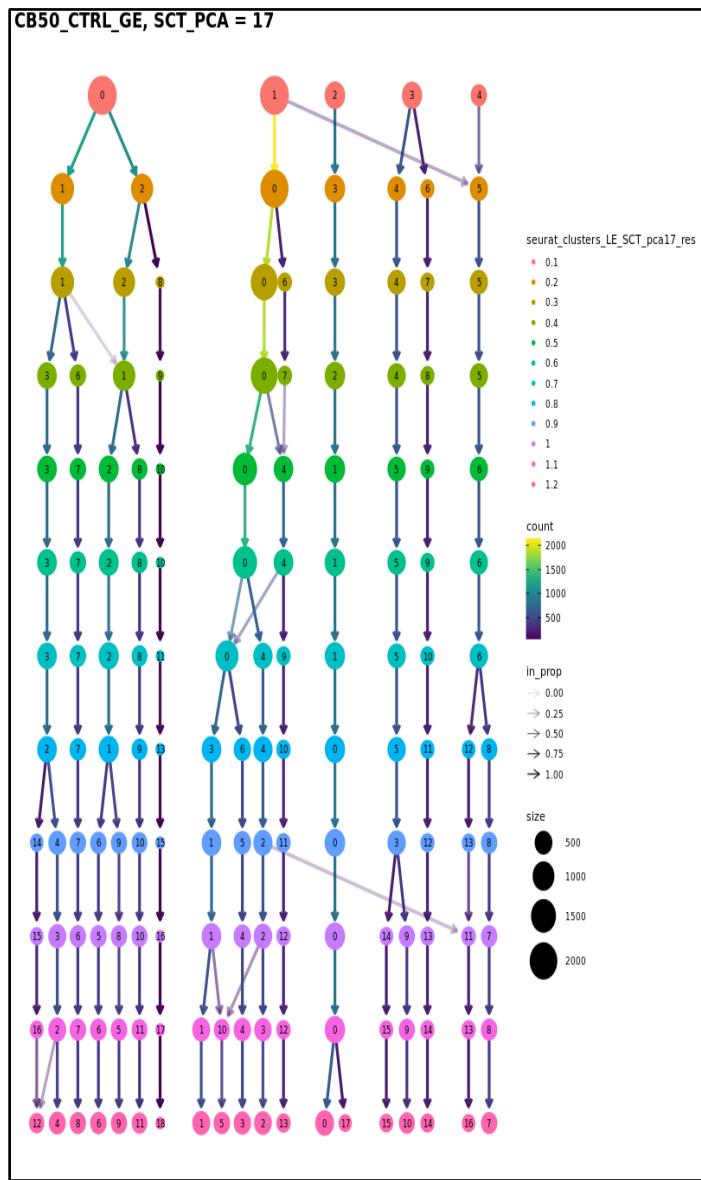
Ensuite, l'étape de la réduction de dimensions, nous permet de préciser le nombre de dimensions à garder pour le clustering. Le pipeline nous dessine toutes les UMAPs possibles selon les deux paramètres : résolutions et dimensions.

Il va tester entre 5 et 49 dimensions avec un pas de deux, et toutes les résolutions entre 0,1 et 1,2 avec un pas de 0,1. On doit vérifier tous les UMAP pour choisir celle qui contient les clusters les mieux isolés les uns des autres et qui est la plus cohérente avec le nombre de clusters estimé.

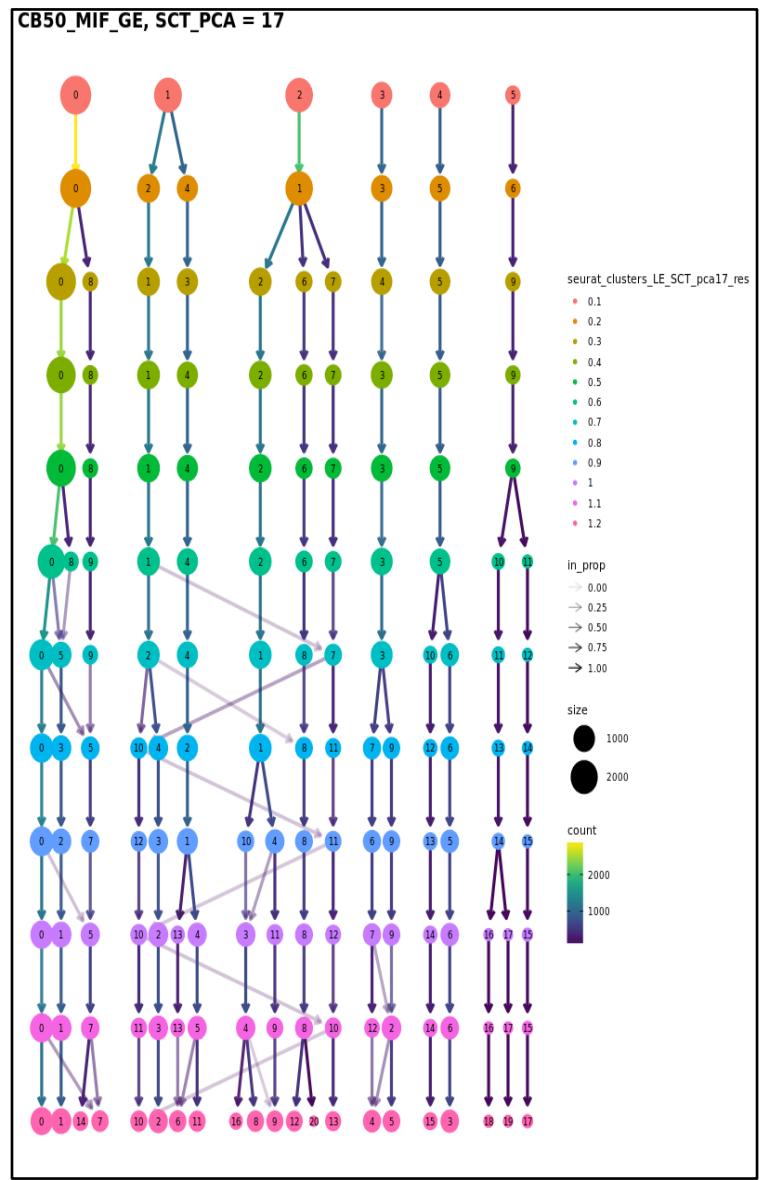
On a un ensemble de graphes clustree qui permettent de nous aider à préciser les dimensions les plus convenables. Le graphe met en évidence les changements de clusters des cellules qui auront lieu entre 2 résolutions successives.

Pour rappel, les sous-types d'une résolution élevée, doivent être entièrement inclus dans le type cellulaire d'une résolution faible. Ainsi sur ces graphes, il doit y avoir le moins de flèches transverses possible.

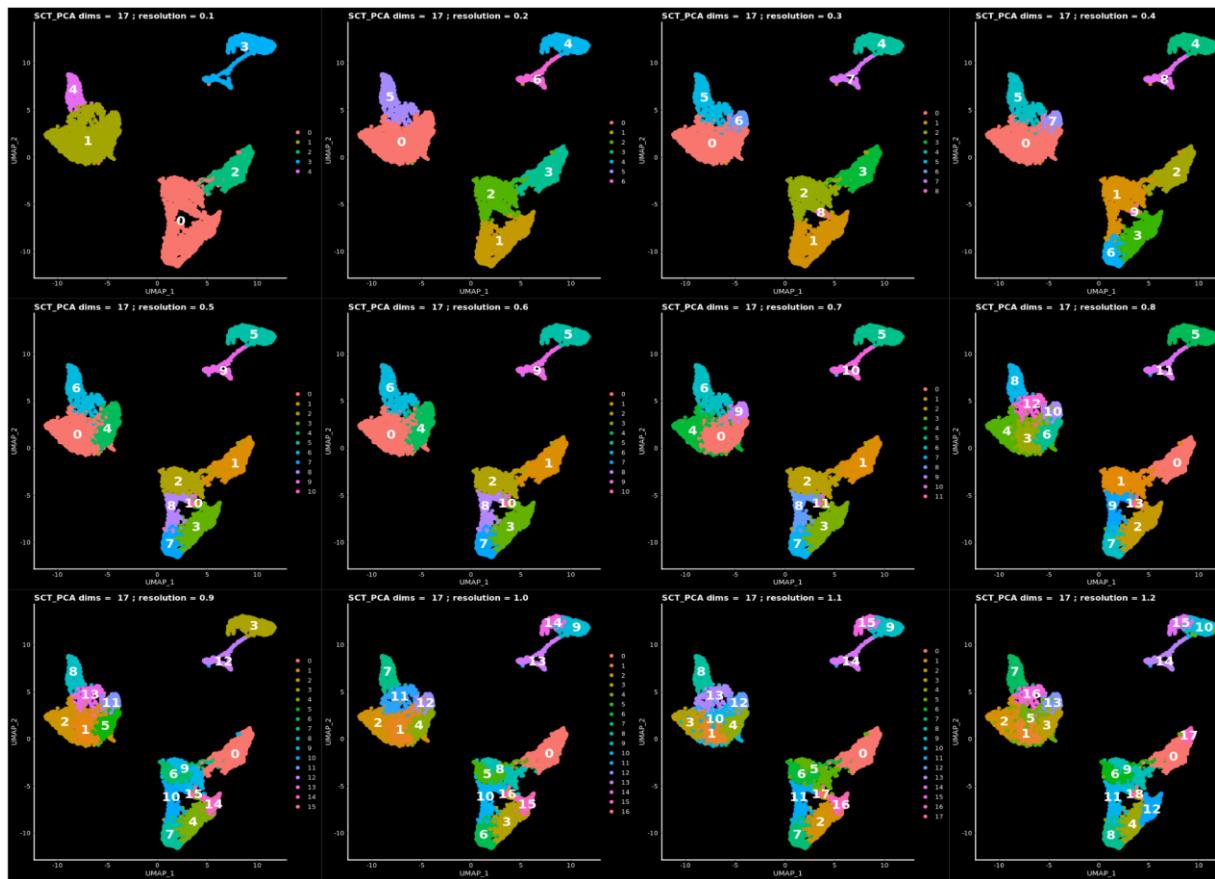
a- Pour l'échantillon CTRL :



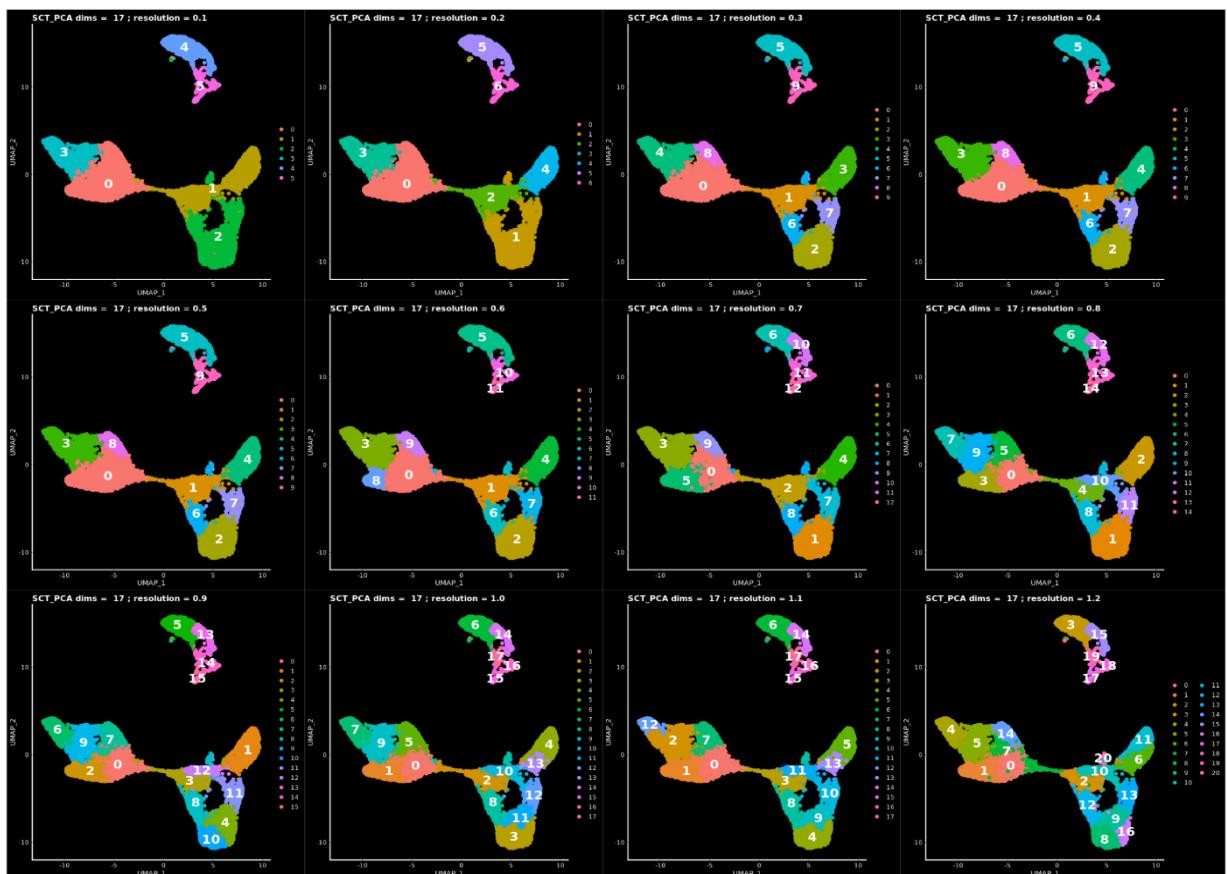
b- Pour l'échantillon MIF :



a- Pour l'échantillon CTRL :

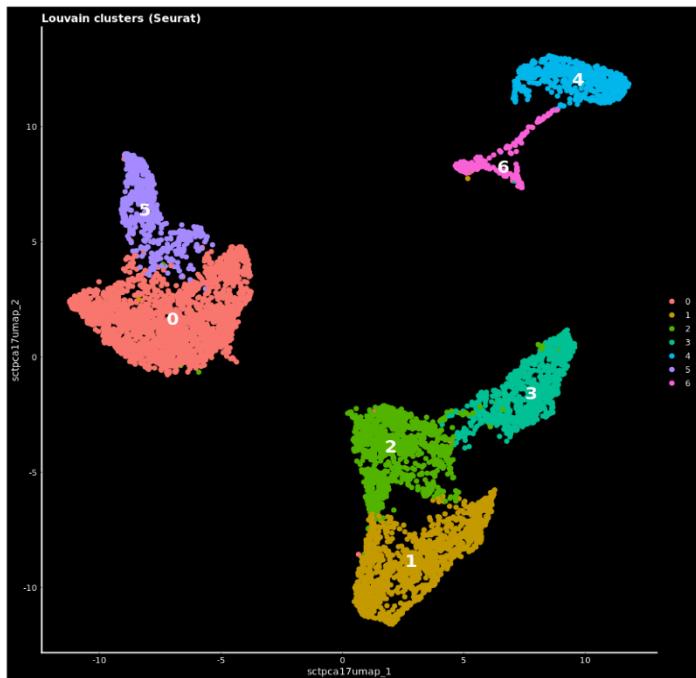


b- Pour l'échantillon MIF :

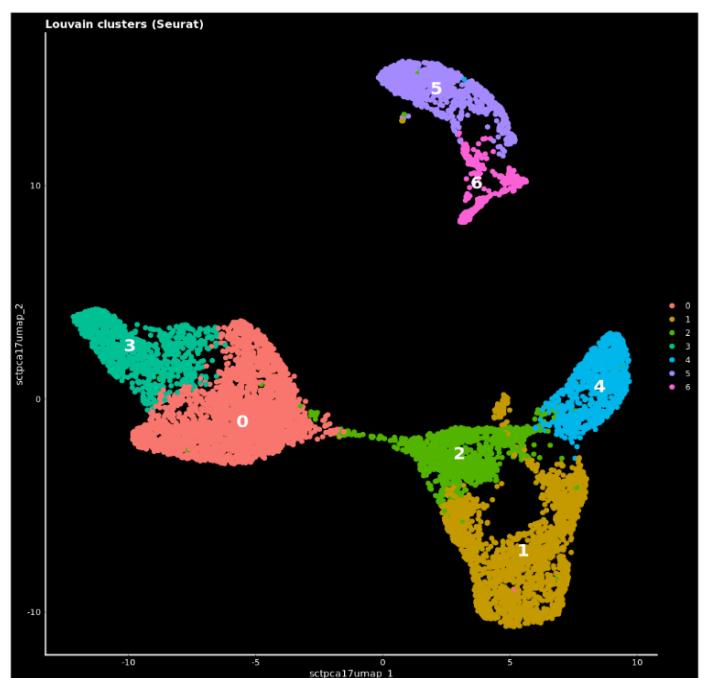


Pour les 17 premières dimensions, j'ai choisi une résolution de 0.2 qui semble correspondre au nombre de types cellulaires attendu, et permet de rester homogène entre les 2 échantillons. Je n'ai pas besoin d'une résolution très précise à ce stade. Le but étant de valider la qualité des échantillons et des paramètres des étapes précédentes. Une analyse plus poussée sera effectuée lors de l'analyse groupée et/ou intégrée, qui permettent vraiment de comparer les échantillons.

a- Pour l'échantillon CTRL :



a- Pour l'échantillon MIF :



Pour les 2 échantillons, j'ai choisi de conserver 17 dimensions avec une résolution de 0.2. Donc, il y a 7 clusters de cellules.

b- Annotation des clusters :

Ainsi pour cette étape, on doit régler les 2 paramètres identifiés précédemment:

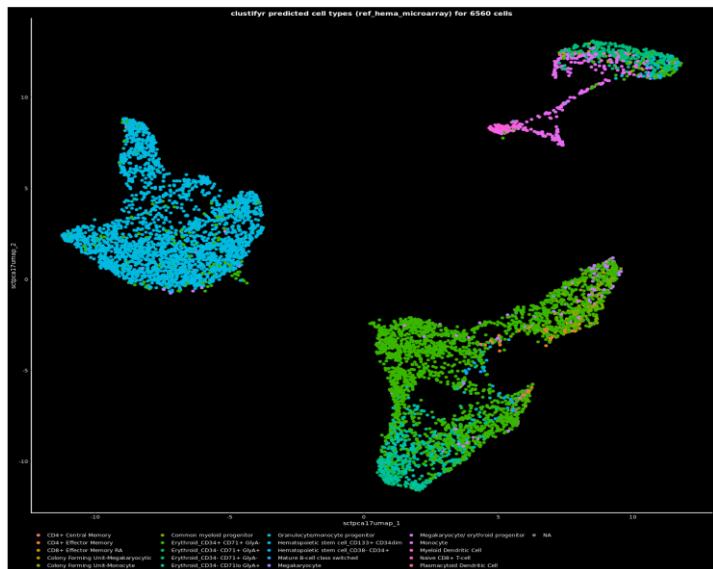
`dims = 17` : seules les 17 premières dimensions sont utilisées pour la UMAP et le clustering.

`res = 0.2` : résolution du clustering.

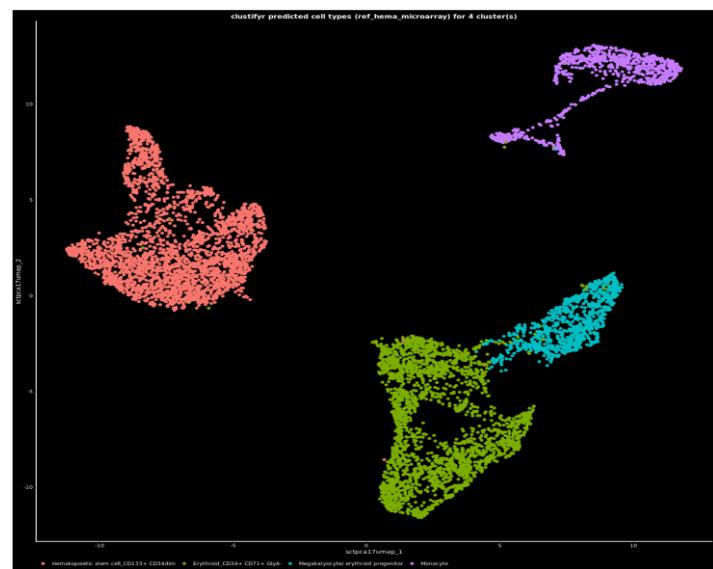
Puis, l'étape d'annotation est lancée. Durant cette étape du pipeline, l'annotation automatique est faite grâce à deux outils (Clustifyr et SingleR). Ces outils comparent l'expression des cellules avec l'expression de références. Toutes les références ne sont pas pertinentes car il peut y avoir un gap technologique entre celle utilisé pour la référence et celle des échantillons.

a- Pour l'échantillon CTRL

Clustifyr

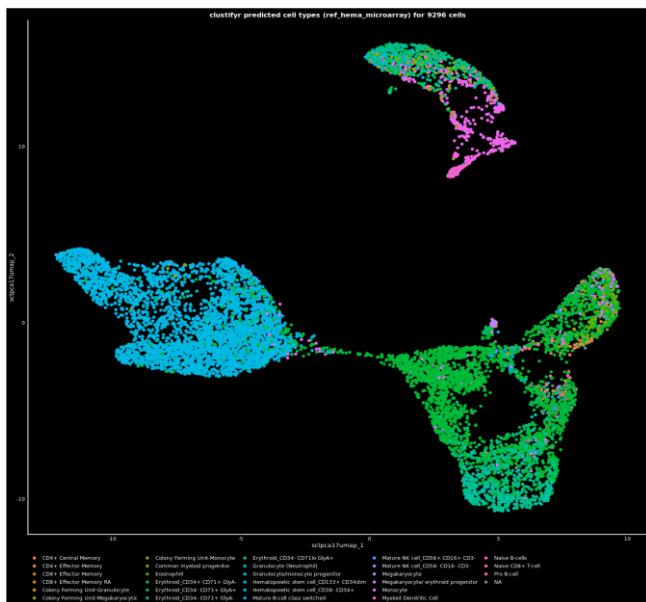


SingleR

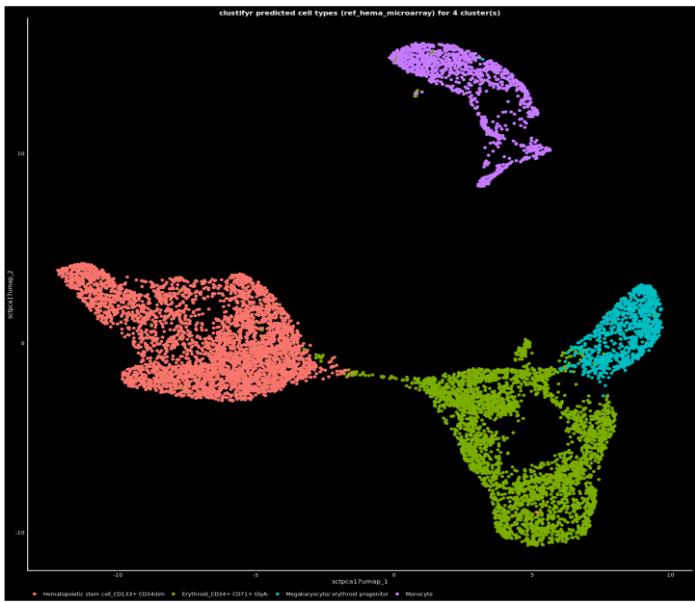


b- pour l'échantillon MIF :

Clustifyr



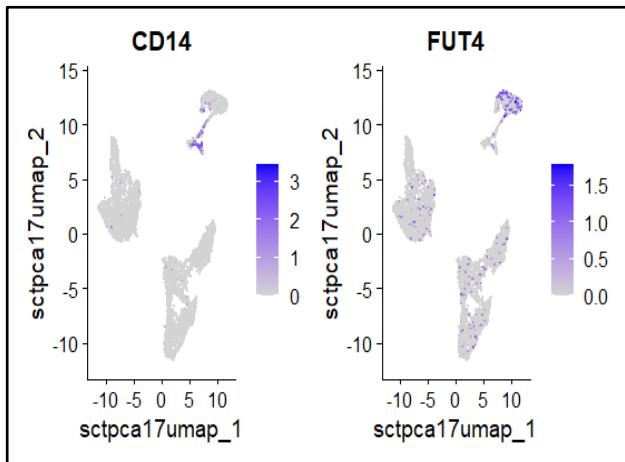
SingleR



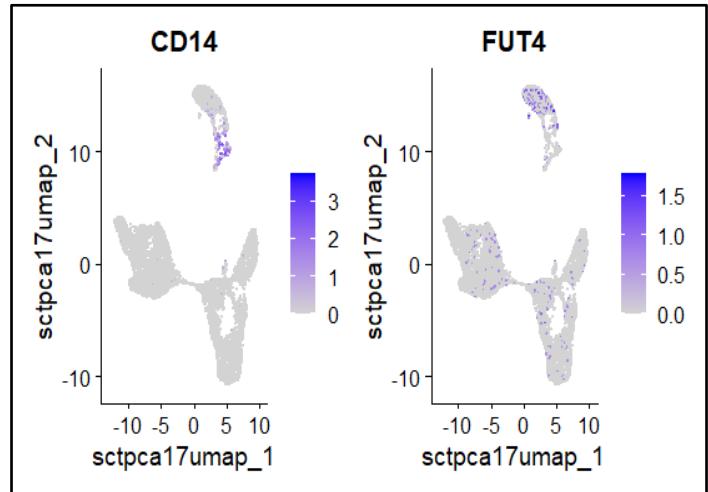
Suite aux différentes étapes réalisées lors de l'analyse individuelle des deux échantillons, on a obtenu un fichier rda par échantillon contenant un objet Seurat qu'on a pu exploiter pour confirmer l'annotation automatique des monocytes et des granulocytes. Pour cela j'ai tracé l'expression de 2 gènes marqueurs des monocytes de granulocytes sur les UMAP des 2 échantillons.

Les deux graphiques ci-dessous confirment cette annotation grâce à la présence du CD14 au niveau du cluster 6 pour l'échantillon CTRL et le cluster 6 pour l'échantillon MIF. Ainsi que la présence du gène FUT4 au niveau du cluster 4 pour l'échantillon CTRL et le cluster 5 pour l'échantillon MIF.

a- pour l'échantillon CTRL



b- pour l'échantillon MIF :



CD14 : marqueur monocytaire , FUT4 : marqueur granulocyttaire

Ainsi je peux déduire que :

- l'échantillon CTRL est composé de :

- Cluster 0 et Cluster 5 : cellules souches hématopoïétiques
- Cluster 1 + Cluster 2 : cellules érythroïdes
- Cluster 4 : granulocytes (neutrophiles)
- Cluster 6 : monocytes
- Cluster 3 : granulocytes (basophiles)

- l'échantillon MIF est composé de :

- Cluster 0 + 3 : cellules souches hématopoïétiques
- Cluster 1 + 2 : cellules érythroïdes
- Cluster 4 : granulocytes (basophiles)
- Cluster 5 : granulocytes (neutrophiles)
- Cluster 6 : monocytes

5. Analyses groupées :

L'objectif de l'analyse groupée des données est de s'assurer que les types de cellules d'une condition s'alignent sur les mêmes types de cellules de l'autre condition. Ces cellules doivent être entremêlées et indiscernables même si elles proviennent d'expériences différentes. En d'autres termes, les différences techniques entre les ensembles de données doivent être supprimées tandis que les principales variations biologiques doivent être préservées.

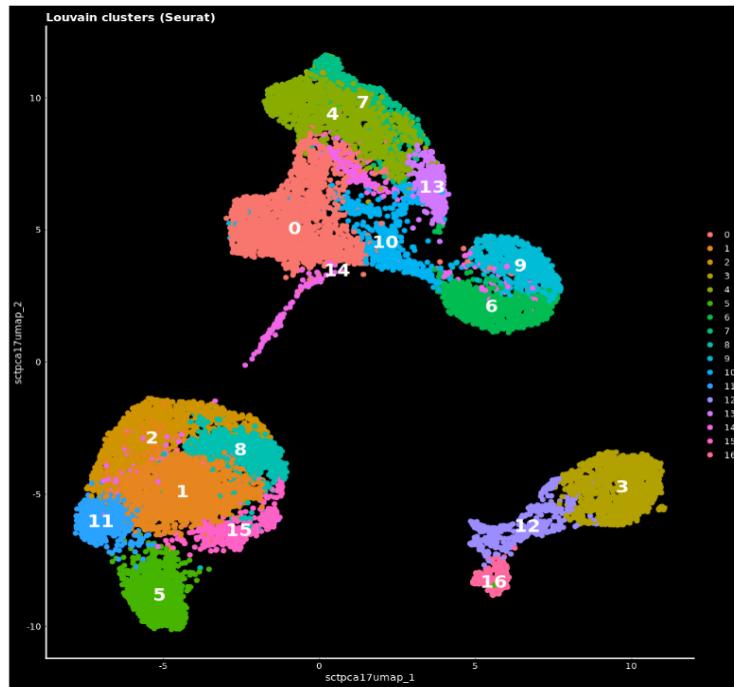
Pour réaliser cette analyse, on a utilisé les deux fichiers correspondant aux deux échantillons après analyses individuelles. Les normalisations ont été conservées, mais une réduction de dimensions commune a été réalisée. Ainsi on a recommencé les mêmes étapes que précédemment.

Réduction des dimensions et clustering:

Comme pour les analyses individuelles, les graphiques de clustree et les différentes UMAP m'ont permis de choisir les paramètres adéquats pour la résolution et le clustering.

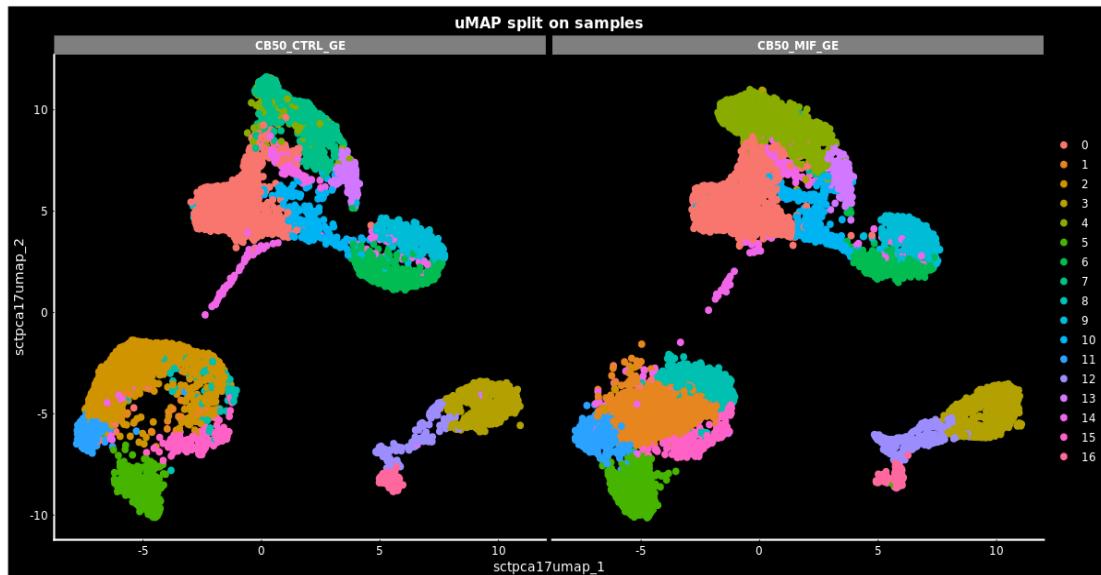
J'ai choisi la dimension 17 et la résolution 0,8.

UMAP finale du clustering de l'analyse groupée:



Comme vu précédemment, le but de l'analyse groupée est de s'assurer que les types de cellules d'une condition s'alignent sur les mêmes types de cellules de l'autre condition.

Pour cela le pipeline trace un graphique “split”, qui représente la UMAP commune mais pour chaque échantillon séparément. Ainsi il permet d'observer un potentiel effet batch.



Nous observons des clusters présents dans un seul des échantillons.

C'est le cas des clusters 2 et 7 pour l'échantillon CTRL, et 1, 4 et 8 pour l'échantillon MIF. Or les cellules de ces clusters ne sont pas censées avoir des propriétés différentes entre les deux échantillons, car les échantillons étaient similaires lors des analyses individuelles. De plus, nous observons ces différences pour toutes les dimensions testées, donc elles sont dues à un biais technique et doivent être corrigées.

6. Analyses intégrées :

Le but de l'analyse intégrée est le même que celui de l'analyse groupée. La seule différence est qu'on applique une correction de l'effet batch grâce à l'outil Seurat avant la réduction de dimension.

Intégration, réduction de dimensions et clustering:

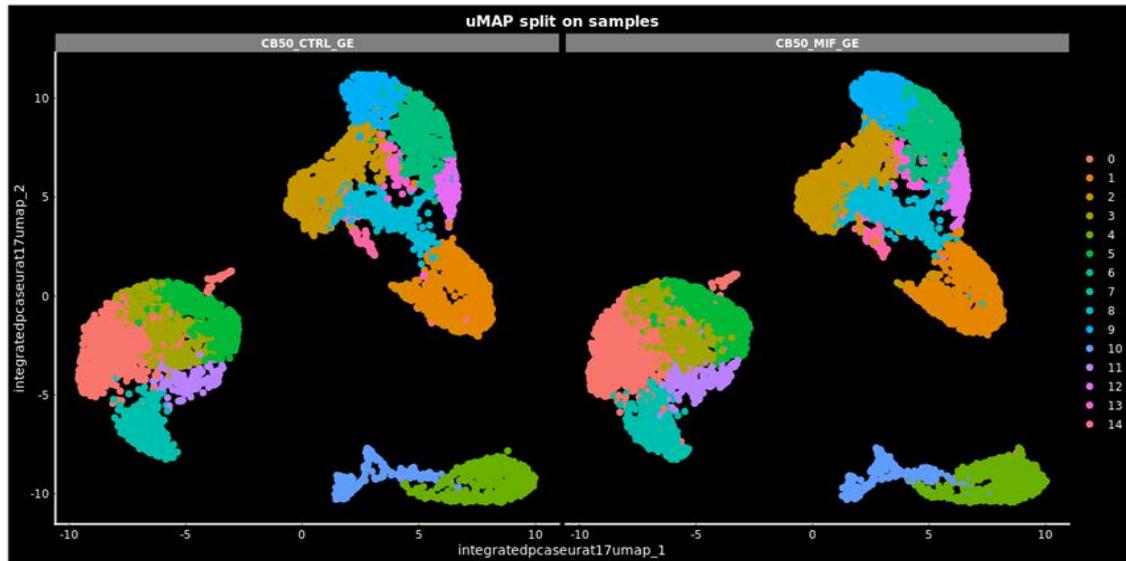
Pour réaliser cette analyse, on a utilisé les deux fichiers correspondant aux deux échantillons après analyses individuelles. Les normalisations ont été conservées, une correction de l'effet batch a été appliquée et une réduction de dimensions commune a été réalisée. Donc on a également recommencé

les mêmes étapes que précédemment avec la même stratégie d'identification des paramètres adéquat.

Entre 2 résolutions successives, j'observe systématiquement des changements de clusters pour certaines cellules. Cela indique une faible stabilité du clustering. Néanmoins, le graphe clustree avec les 17 premières dimensions semble être le plus stable. Puis je choisis une résolution à 0,7.

UMAP finale du clustering de l'analyse intégrée:

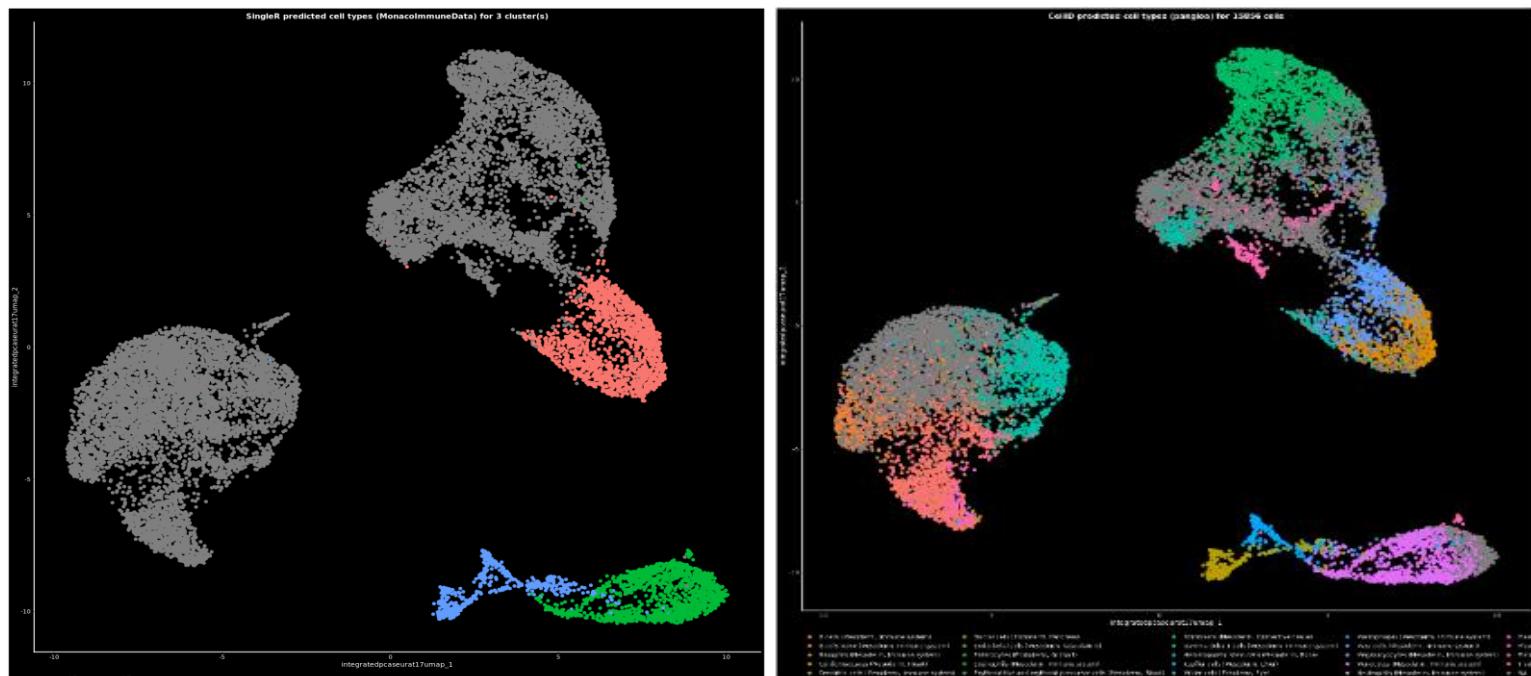
Similaire à précédemment, on vérifie la présence ou non d'effet batch grâce au graphe « split » :



Ici tous les clusters se superposent bien entre les conditions, donc l'effet batch a été corrigé.

Annotation des clusters :

Grâce à l'annotation SingleR du pipeline qui met en évidence que les monocytes (en bleu) sont plus proches des neutrophiles (granulocytes en vert) ce qui est concordant avec le système de différenciation immunitaire.



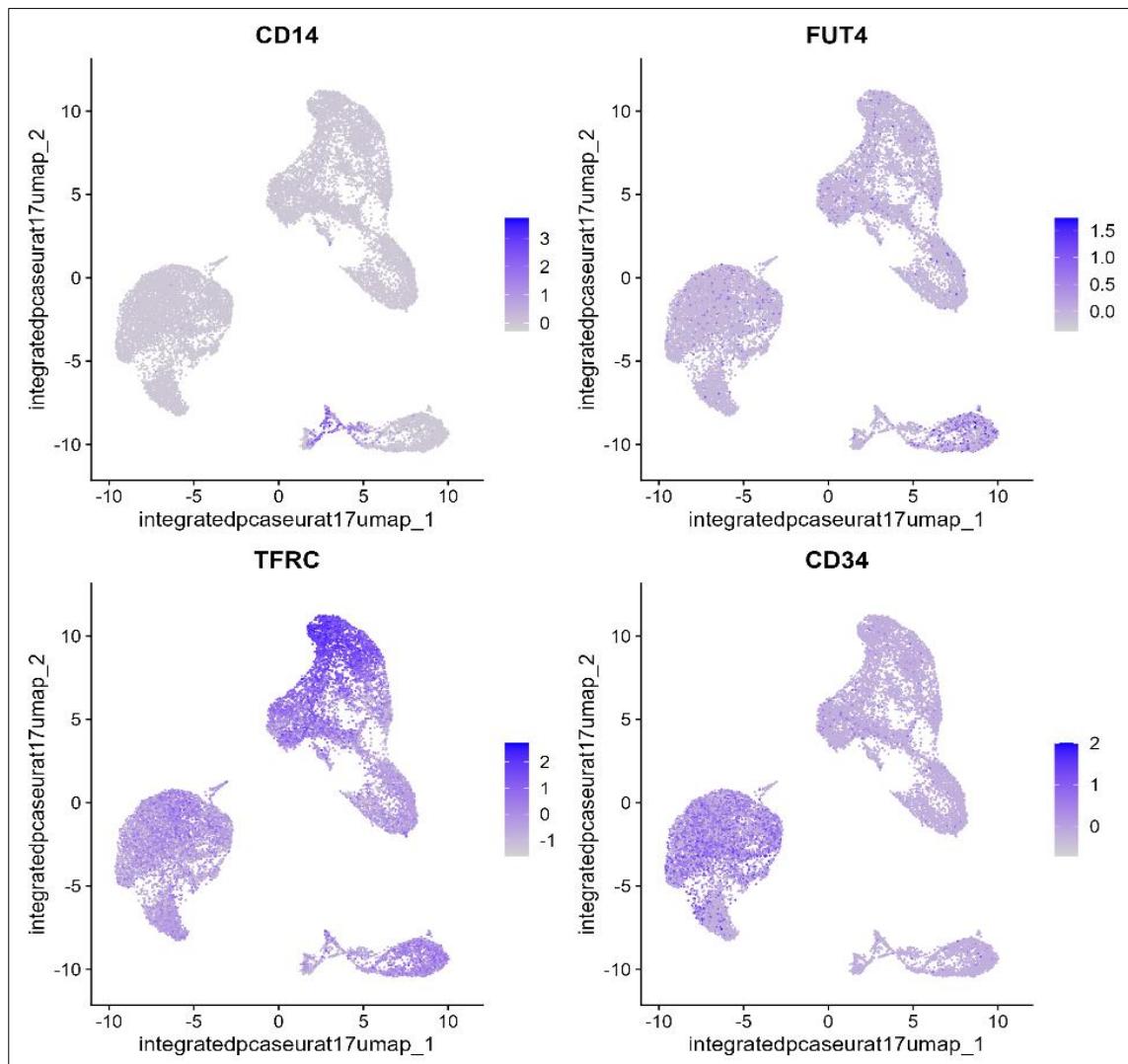
Ainsi je peux déduire :

- Cluster 10 : monocytes
- Cluster 4 : granulocytes (neutrophiles)
- Cluster 1 : granulocytes (Basophiles)
- Cluster 9+6+13+8+2+12+8 : érythrocytes
- Cluster 3 + Cluster 0 + Cluster 5 + Cluster 11 + Cluster 7 : cellules souches hématopoïétiques

Expression de certains gènes sur les UMAP :

Afin de confirmer l'annotation des clusters, on a tracé l'expression des gènes marqueurs sur le UMAP :

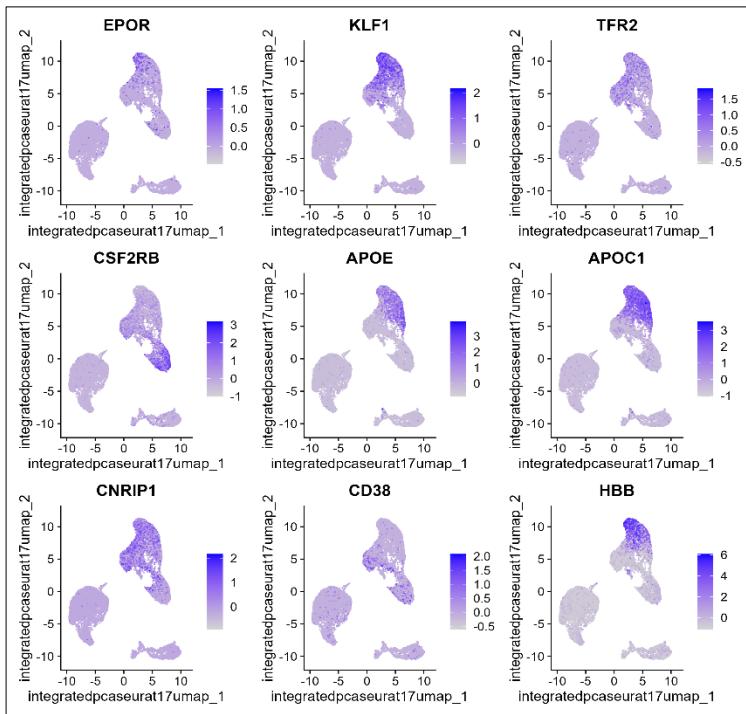
Echantillon MIF_CTRL



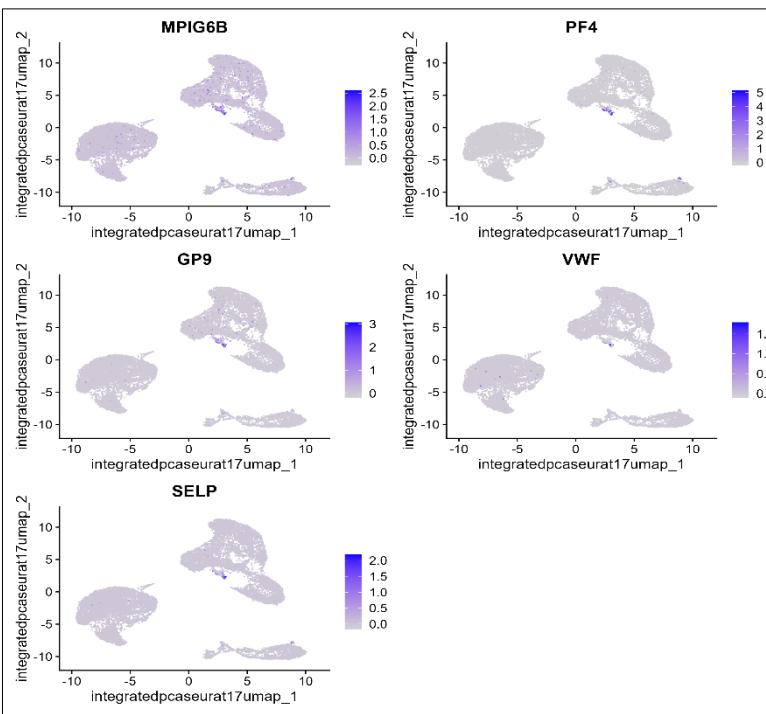
CD14 = Monocytes; FUT4 = Granulocytes; TFRC = Erythroïdes ; CD34 = cellules souches hématopoïétiques.

Autres listes de gènes :

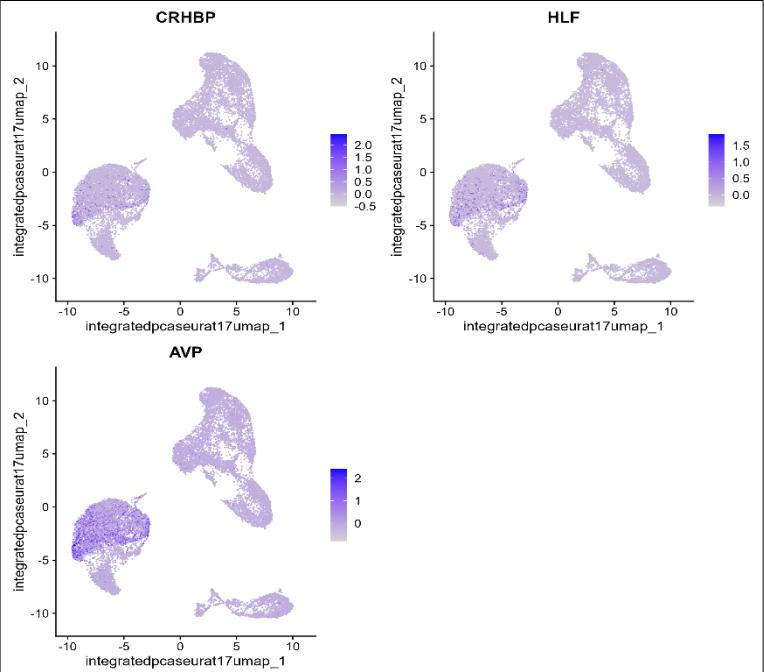
Erythroïdes



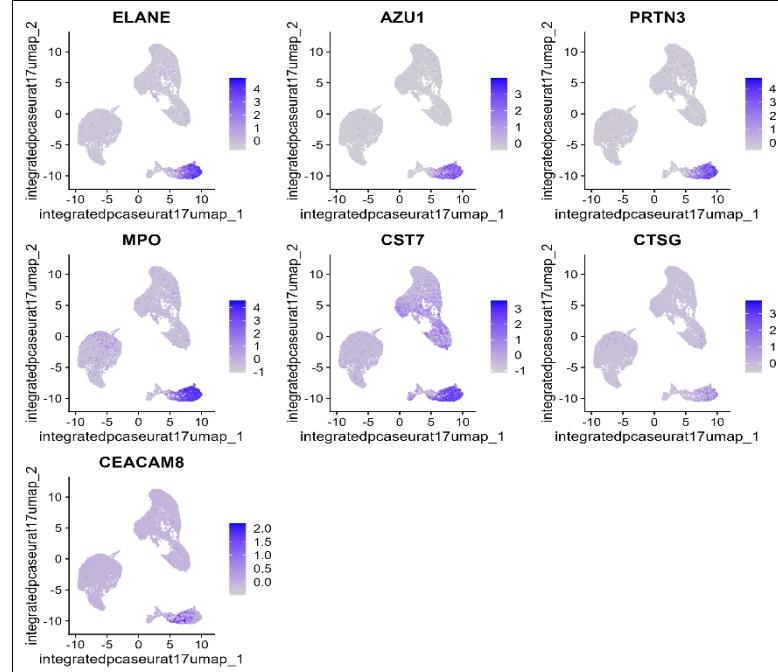
Mégacaryocytes



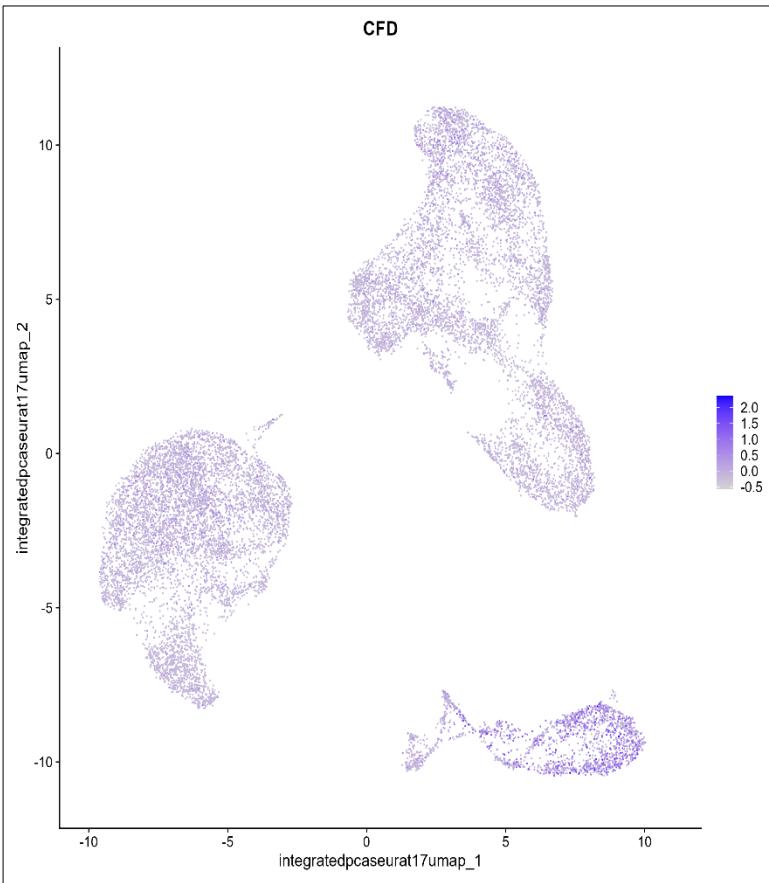
Cellules souches hématopoïétiques



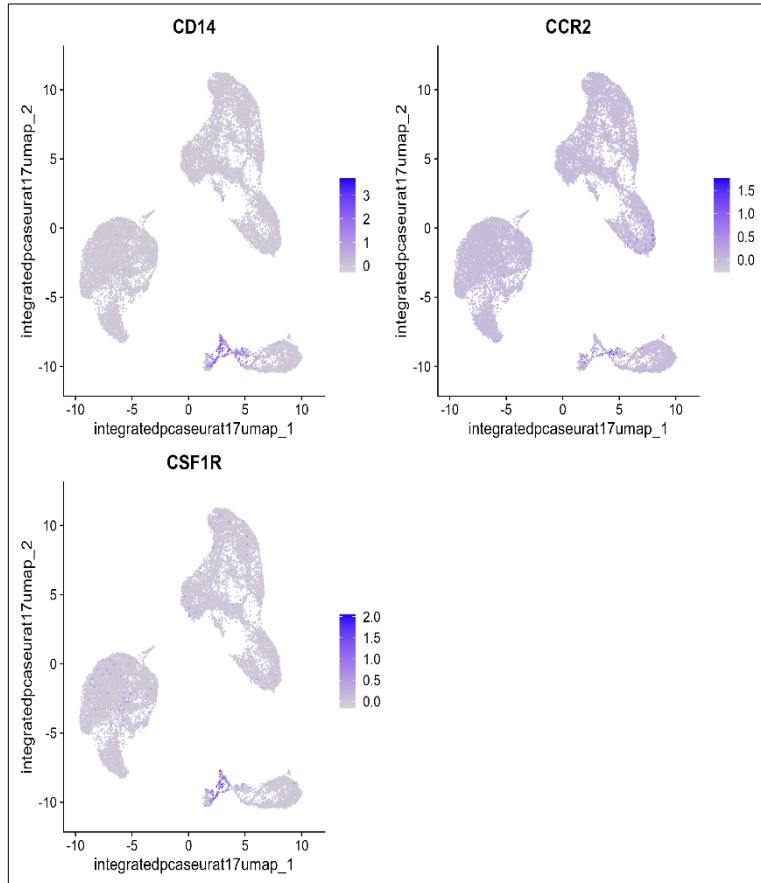
Neutrophiles



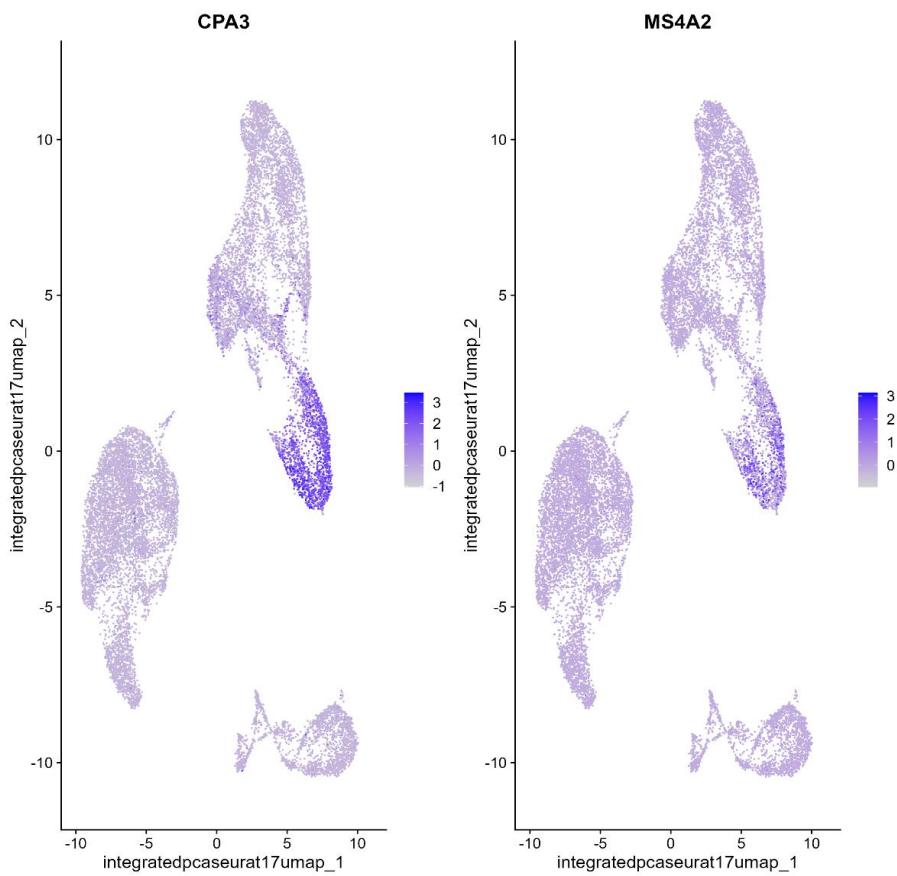
Granulocytes



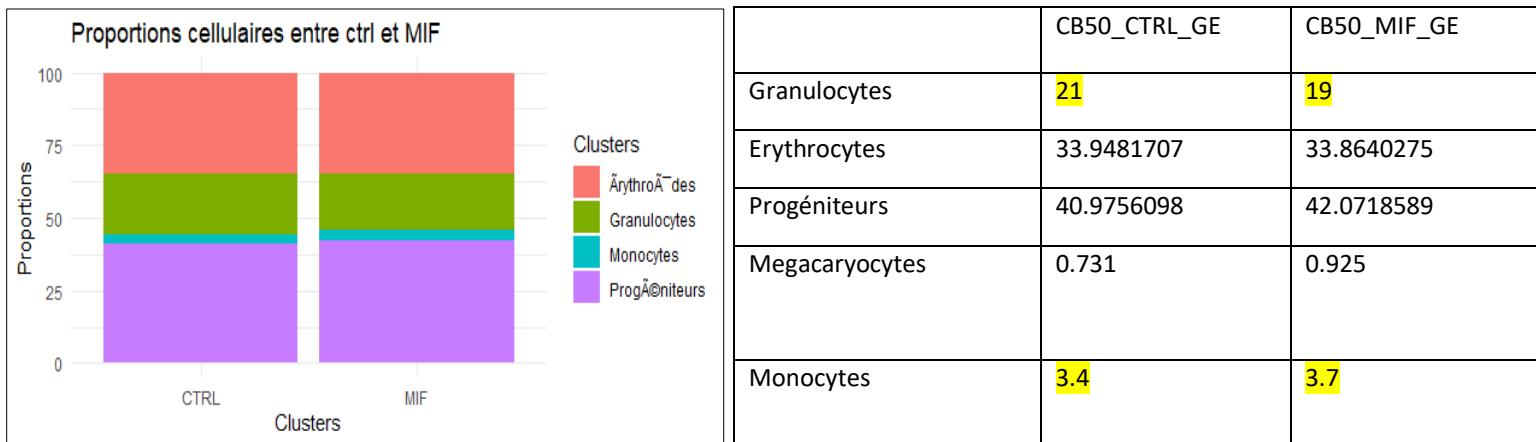
Monocytes



Basophiles



7. Étude des proportions cellulaires



Le graphique ci-dessus de comparaison des proportions cellulaires entre CTRL et MIF ainsi que le tableau montrent que les proportions monocytaire des deux échantillons sont presque égales avec une légère augmentation du nombre des monocytes chez l'échantillon MIF par rapport à l'échantillon CTRL ainsi qu'une légère augmentation du nombre des granulocytes chez l'échantillon CTRL par rapport à l'échantillon MIF.

Afin de vérifier si les deux conditions CTRL et MIF ont un impact sur la différenciation des progéniteurs myéloïdes, j'ai réalisé un test Chi2 de Pearson.

L'objectif principal d'un test Chi2 est de savoir si deux variables catégorielles sont susceptibles d'être liées ou pas. Ici les variables sont les types cellulaires et les 2 conditions d'échantillons.

Notre hypothèse d'indépendance est que les proportions cellulaires sont indépendantes de la répression ou non du gène TET2.

- Si p-valeur <5%, on rejette notre hypothèse => donc les proportions cellulaires sont liées aux caractéristiques des échantillons.
- Si p-valeur >5%, on ne peut pas rejeter l'hypothèse d'indépendance.

Test Chi2 pour toutes les proportions cellulaires :

Ce test s'applique sur le tableau de contingence ci-dessous qui est un tableau de croisement des deux conditions appliquées :

	CB50_CTRL_GE	CB50_MIF_GE
Granulocytes	1372	1811
Progéniteurs	2688	3911
Erythrocytes	2227	3148
Monocytes	225	340
Mégacaryocytes	48	86

On doit tout d'abord s'assurer que les effectifs du tableau de contingence sont supérieurs à 5, ce qui est vrai.

X-squared	df	p-value
7.3153	4	0.1201

df (degrés de liberté) : concept statistique qui représente le nombre de valeurs indépendantes qui peuvent varier lorsqu'une statistique est calculée à partir d'un ensemble de données. Dans le contexte du test chi2, les degrés de liberté correspondent au nombre de catégories moins un.

X-squared (valeur statistique calculée du test chi-square) : est obtenue en comparant les observations réelles avec les valeurs attendues selon une hypothèse nulle spécifique.

La p-valeur (résultat statistique) vaut 0.1201. C'est supérieur à 0.05, donc on ne peut pas rejeter l'hypothèse d'indépendance, donc on ne peut pas considérer que les différences des proportions entre les deux échantillons sont liées à nos conditions expérimentales CTRL et MIF.

Test de Chi2 sur les proportions cellulaires des cellules différencierées des progéniteurs myéloïdes (Granulocytes, Monocytes et Erythrocytes) :

	CB50_CTRL_GE	CB50_MIF_GE
Granulocytes	1372	1811
Erythrocytes	2227	3148
Monocytes	225	340

Les effectifs du tableau de contingence sont supérieurs à 5.

X-squared	df	p-value
3.37771.9802	2	0.1847

La p-valeur vaut 0.1847. C'est supérieur à 0.05, on ne peut pas considérer que l'augmentation du taux des monocytes de l'échantillon MIF est liée à la répression de l'expression du gène TET2.

8. Pour aller plus loin :

A défaut de présenter des proportions différentes entre nos conditions, nous souhaitons déterminer si les monocytes et les granulocytes présentent des différences d'expressions entre nos conditions.

Pour cela, j'ai réalisé une analyse d'expression différentielle.

L'analyse différentielle permet d'identifier les gènes qui sont différentiellement exprimés entre deux groupes de cellules. Le test proposé par Seurat, et qui est habituellement utilisé, est Wilcoxon.

Pour réaliser cette analyse, j'ai utilisé le fichier résultant de l'analyse intégrée contenant les deux échantillons.

a-Marqueurs différentiellement exprimés au niveau du cluster des Granulocytes :

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
AC040970.1	0	14,6105417	0,046	0,765	0
RGS9	2,46E-305	14,6536141	0,033	0,656	7,37E-302
PYROXD1	1,08E-303	5,23867235	0,068	0,629	3,23E-300
AC007336.1	1,89E-287	8,36420538	0,073	0,778	5,68E-284
NFAM1	5,41E-271	13,3332369	0,033	0,576	1,62E-267
ENC1	1,80E-263	6,95409605	0,044	0,531	5,41E-260
GADD45B	4,82E-257	-3,8080246	0,147	0,147	1,45E-253
RENBP	2,09E-246	-5,14589698	0,082	0,144	6,28E-243
ALS2	8,79E-241	7,10499725	0,139	0,732	2,64E-237
AC041005.1	5,71E-237	14,0288416	0,011	0,612	1,71E-233
AC098818.2	5,58E-234	4,44576156	0,056	0,356	1,67E-230
APBB2	1,44E-231	15,2529851	0,036	0,716	4,32E-228
AFF2	9,25E-226	11,0642432	0,056	0,64	2,77E-222
HYAL3	1,50E-220	13,0583964	0,157	0,743	4,51E-217
AC234772.2	9,15E-218	16,543026	0,019	0,649	2,75E-214
GTSF1	9,24E-216	4,79213372	0,039	0,175	2,77E-212
ZBTB20	3,22E-215	5,83135143	0,104	0,437	9,65E-212
RN7SL1	1,91E-213	-2,37293954	0,081	0,106	5,74E-210
ZNF490	4,55E-209	10,0834694	0,024	0,349	1,37E-205

Le tableau ci-dessus met en évidence la liste des gènes différentiellement exprimés entre les deux échantillons au niveau du cluster des Granulocytes selon un classement décroissant qui dépend de la valeur l'avg_log2FC et de P-val.

avg_logFC: log de l'expression moyenne entre les deux groupes de cellules comparés, appelé aussi log2FC (« log Fold Change »). Les valeurs positives indiquent que le gène est plus fortement exprimé dans le premier groupe. Par exemple, un logFC à 0.58 correspond à un FC de 1.5, ce qui signifie que le gène est 1,5 fois plus exprimé dans la condition testée par rapport au contrôle.

p_val : p-valeur brute.

p_val_adj : p-valeur ajustée, basée sur la correction de Bonferroni en utilisant tous les gènes de l'ensemble de données.

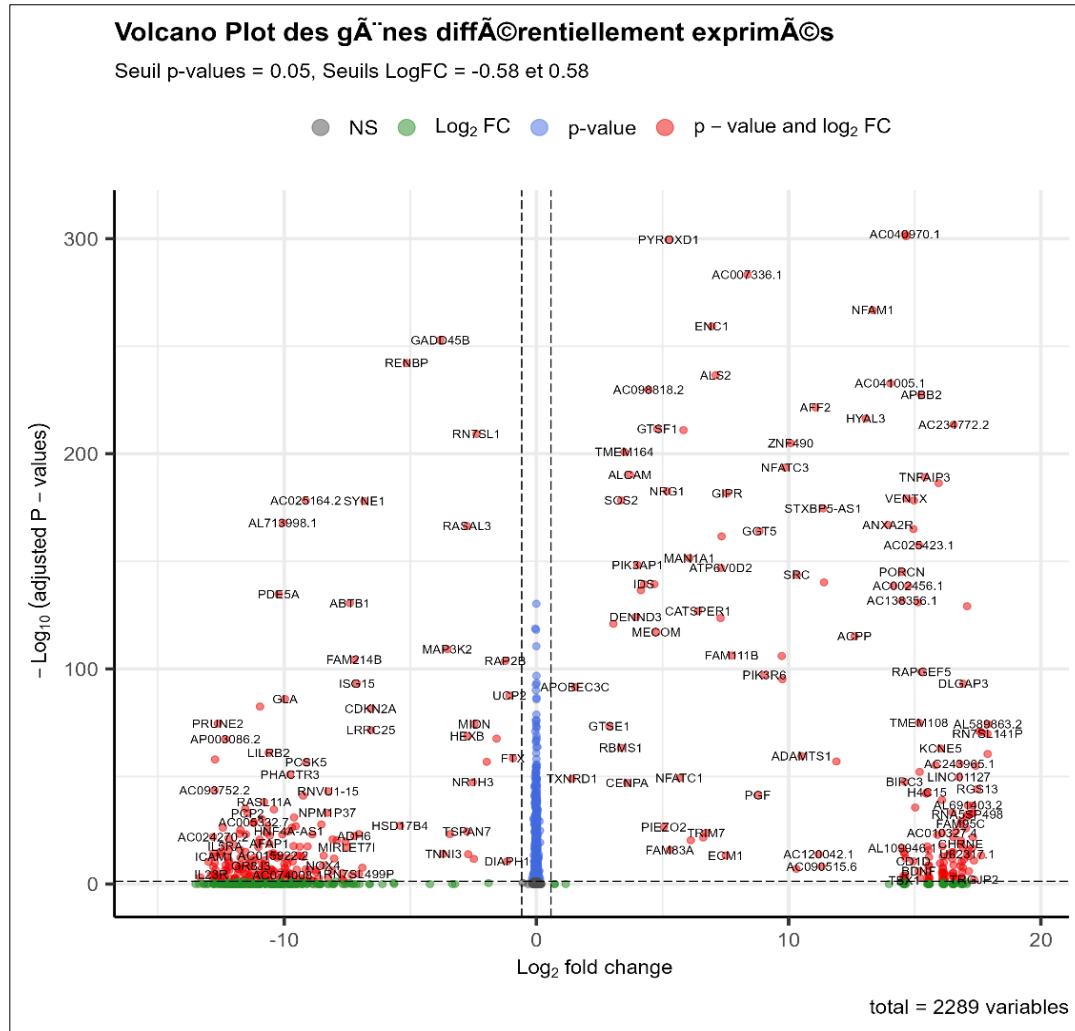
pct1 et pct2 : pourcentage de cellules exprimant le gène testé dans chaque groupe de cellules.

Au total, il y a 2289 gènes différentiellement exprimés (p-valeur ajustée < 5 %, et log2FC > 0.58 ou < -0.58).

On constate que parmi les gènes différentiellement exprimés entre les deux échantillons on trouve : **RGS9** : Ce gène code pour un membre de la famille RGS des protéines activatrices de la GTPase qui fonctionnent dans diverses voies de signalisation en accélérant la désactivation des protéines G. Cette protéine est ancrée aux membranes des photorécepteurs dans les cellules rétinienne et désactive les protéines G dans les cascades de phototransduction des bâtonnets et des cônes. Les mutations de ce gène entraînent une bradyopsie. [Fourni par RefSeq, septembre 2009]

PYROXD1 : Ce gène code pour pyridine nucleotide-disulphide oxidoreductase domain-containing protein 1(PNDR). Les PNDR sont des flavoprotéines qui catalysent la réduction dépendante des nucléotides pyridine des résidus thiol dans d'autres protéines. Il se localise dans le noyau et dans les

compartiments striés des sarcomères. Des mutations naturelles de ce gène provoquent une myopathie précoce avec des noyaux intérieurisés et une désorganisation myofibrillaire. [Fourni par RefSeq, avril 2017]



Interprétation gènes :

Le Volcanoplot obtenu présente les gènes différentiellement exprimés selon le log2FC et le log de la p-valeur. Les gènes différentiellement exprimés sont déjà cités dans le tableau ci-dessus.

Analyses d'enrichissement fonctionnelle :

Etudier les gènes un à un étant une étape fastidieuse, une analyse d'enrichissement fonctionnelle permet d'identifier les voies de signalisation différemment activées entre nos conditions. Il existe 2 méthodes : GSEA et Enricher.

GSEA évalue un profil d'expression à l'échelle du génome et détermine si des ensembles de gènes définis a priori présentent des changements cumulatifs statistiquement significatifs dans l'expression des gènes qui sont corrélés avec un phénotype. Le phénotype peut être catégoriel (par exemple : tumeur par rapport à la normale) ou continu (par exemple : un profil numérique sur tous les échantillons de l'ensemble de données d'expression). Elle attribue ensuite une p-valeur pour chaque jeu de gènes en fonction de son enrichissement relatif dans le classement. GSEA est particulièrement adaptée à l'analyse de données d'expression génique, comme les puces à ADN ou le séquençage ARN.

Enricher :

Enricher est une fonctionnalité plus générale utilisée dans diverses bibliothèques logicielles R (par exemple, clusterProfiler, enrichR) pour l'analyse d'enrichissement fonctionnel des gènes. Elle permet d'identifier les termes ou les annotations fonctionnelles qui sont enrichis dans un ensemble de gènes donné. L'analyse Enricher utilise généralement des tests statistiques tels que le test hypergéométrique ou le test du Chi2 pour évaluer l'enrichissement. Elle fournit des p-valeurs pour chaque terme, indiquant la significativité de l'enrichissement.

Les analyses d'enrichissement fonctionnelle sont réalisées grâce au package R clusterProfiler.

Pour réaliser ces analyses, j'ai utilisé la base de données des signatures moléculaires MSigDB (Molecular Signatures Database) qui est une source de dizaines de milliers d'ensembles de gènes annotés, divisés en collection humaine et souris.

Les jeux de gènes de la base de données MSigDB spécifiques à l'espèce "Homo sapiens" qui ont été utilisés sont :

- La catégorie "H" (Hallmark gene sets) :

Les ensembles de gènes Hallmark résument et représentent des états ou des processus biologiques spécifiques bien définis et affichent une expression cohérente. Ces ensembles de gènes ont été générés par une méthodologie de calcul basée sur l'identification des chevauchements entre les ensembles de gènes dans d'autres collections MSigDB et la conservation des gènes qui affichent une expression coordonnée.

- La catégorie "C2" (Curated gene sets) :

Les ensembles de gènes de cette collection sont sélectionnés à partir de diverses sources,(bases de données web et littérature biomédicale). La collection C2 est divisée en deux sous-collections : Perturbations chimiques et génétiques (CGP) et Voies canoniques (CP).

- La catégorie "C6" (Oncogenic signatures gène sets) :

Des ensembles de gènes qui représentent des signatures de voies cellulaires qui sont souvent dérégulées dans le cancer.

- La catégorie "C7" (Immunologic signature) :

Ensemble de gènes qui représentent les états cellulaires et les perturbations au sein du système immunitaire.

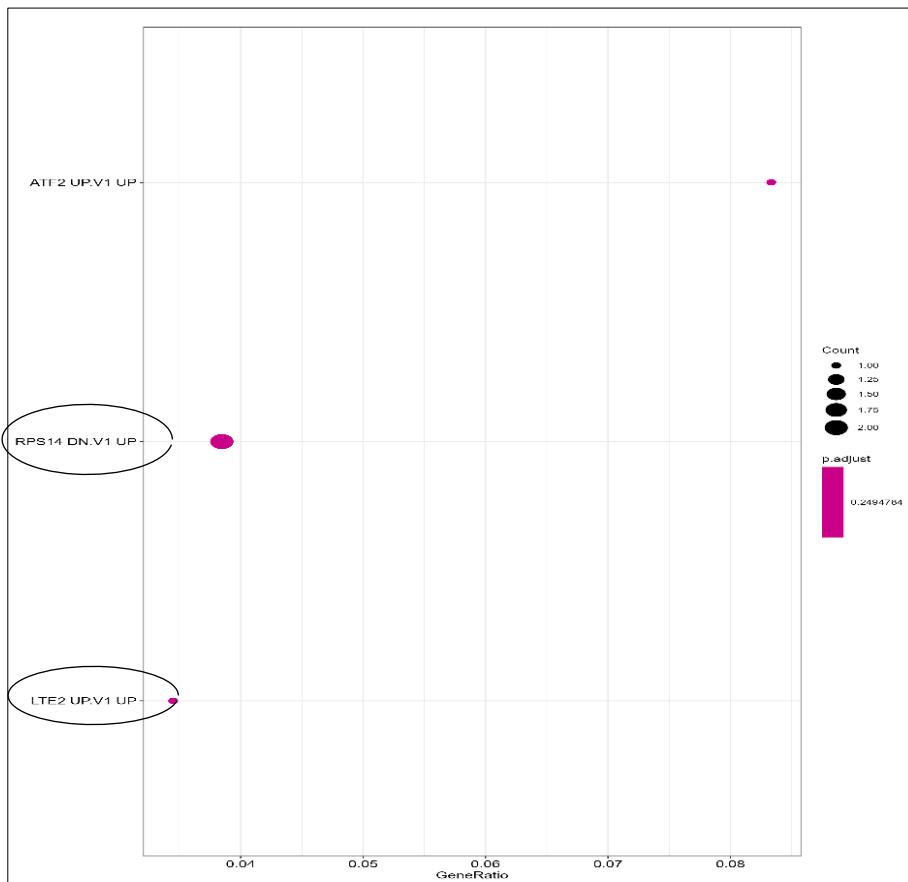
Durant notre étude, pour étudier les voies de signalisation liées aux gènes différentiellement exprimés, on a commencé par réaliser une analyse GSEA en chargeant les jeux de gènes de la base de données MSigDB cités ci-dessus.

Pour le cluster des granulocytes, en définissant un pvalueCutoff (seuil de significativité pour les p-valeurs associées aux enrichissements de jeux de gènes) à 0.25 :

Seul la catégorie "C6" (Oncogenic signatures gène sets) donne un résultat significatif:

RPS14_DN.V1_UP : Gènes surexprimés dans les cellules progénitrices hématopoïétiques CD34 +, ceci est dû à une activation de la protéine RPS14 .

LTE2_UP.V1_UP : Gènes régulés positivement dans les cellules MCF-7(cellules tumorales mammaires)



Par la suite, j'ai réalisé une analyse d'enrichissement.

Parmi l'ensemble des gènes différentiellement exprimés, aucune voie de signalisation enrichie n'a été identifiée pour le cluster des granulocytes.

b- Marqueurs différentiellement exprimés au niveau du cluster des Monocytes :

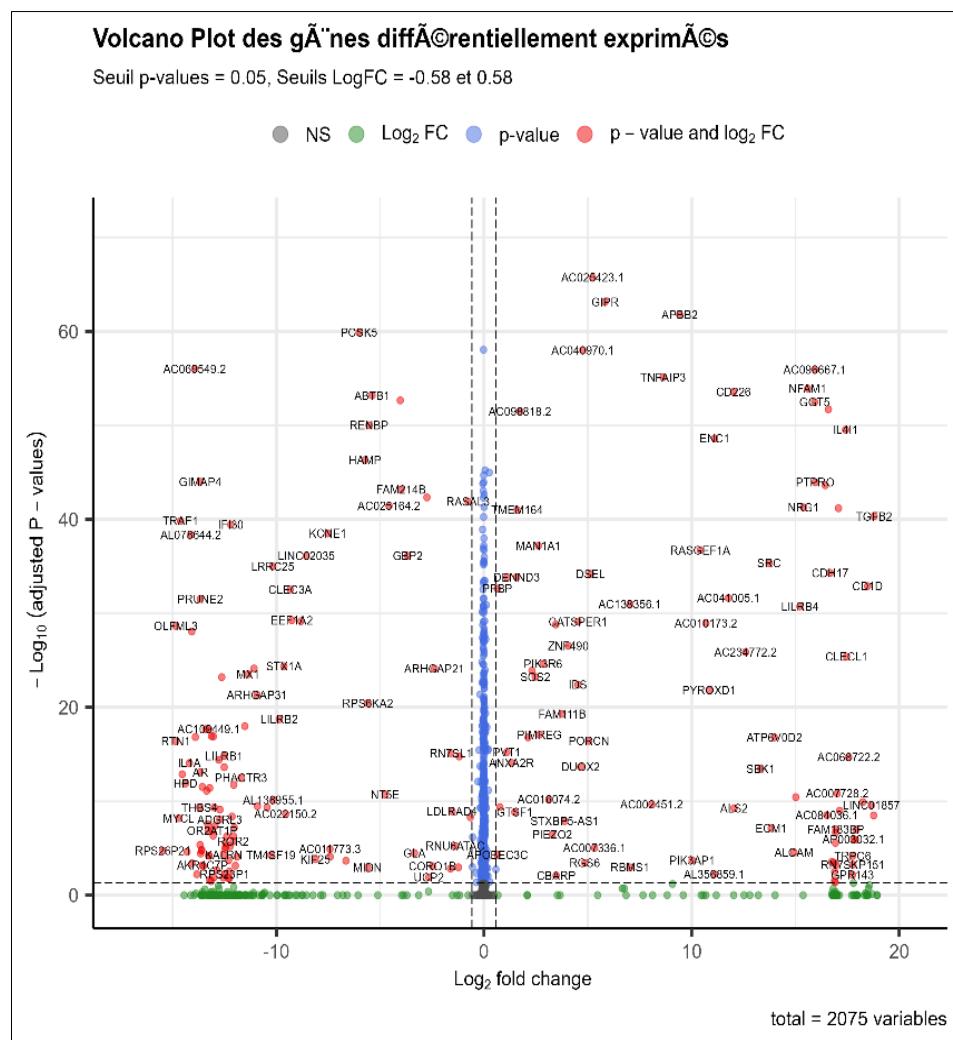
	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
AC025423.1	6,11E-70	5,22814779	0,026	0,662	1,83E-66
GIPR	2,32E-67	5,83166323	0,065	0,876	6,95E-64
APBB2	5,72E-66	9,42743349	0,015	0,702	1,72E-62
PCSK5	4,19E-64	-6,01472421	0,074	0,111	1,26E-60
LINC01191	2,93E-62	-0,02286366	0,009	0,764	8,79E-59
AC040970.1	3,36E-62	4,75836537	0,024	0,458	1,01E-58
AC069549.2	3,18E-60	-13,9359106	0,038	0,116	9,53E-57
AC096667.1	3,79E-60	15,9356153	0,088	0,982	1,14E-56
TNFAIP3	2,50E-59	8,62684772	0,094	0,867	7,49E-56
NFAM1	3,90E-58	15,5915284	0,091	0,898	1,17E-54
CD226	8,89E-58	12,0196515	0,065	0,84	2,67E-54
ABTB1	2,12E-57	-5,38787264	0,05	0,102	6,35E-54
SYNE1	7,18E-57	-4,03027786	0,085	0,116	2,15E-53
GGT5	1,11E-56	15,9247099	0,053	0,649	3,34E-53
ACPP	6,49E-56	16,5891022	0,109	1	1,95E-52
AC098818.2	1,07E-55	1,72845134	0,065	0,436	3,21E-52
RENBP	3,19E-54	-5,53555388	0,082	0,133	9,58E-51
IL4I1	9,67E-54	17,4171136	0,041	0,751	2,90E-50
ENC1	8,12E-53	11,1093641	0,091	0,804	2,44E-49

Le tableau ci-dessus met en évidence la liste des gènes différentiellement exprimés entre les deux échantillons au niveau du cluster des monocytes.

On constate que parmi les gènes différentiellement exprimés entre les deux échantillons on trouve :

GIPR : Ce gène code une protéine G couplée à un récepteur pour le polypeptide inhibiteur gastrique (GIP), qui a été identifié à l'origine comme ayant une activité dans les extraits intestinaux qui inhibe la sécrétion d'acide gastrique et la libération de gastrine, mais il a ensuite été démontré qu'il stimule la libération d'insuline en présence d'une glycémie élevée. Un défaut de ce gène peut contribuer à la pathogénèse du diabète. [Fourni par RefSeq, octobre 2011]

APBB2 : La protéine codée par ce gène interagit avec les domaines cytoplasmiques de la protéine précurseur de la bêta-amyloïde (A4) et de la protéine de type précurseur de la bêta-amyloïde (A4) 2. Cette protéine contient deux domaines de liaison à la phosphotyrosine (PTB), dont on pense qu'ils fonctionnent dans la transduction du signal. Des polymorphismes de ce gène ont été associés à la maladie d'Alzheimer. [Fourni par RefSeq, octobre 2009]



Le Volcanoplot obtenu présente les gènes différentiellement exprimés déjà cités ci-dessus au niveau du tableau.

GSEA (Gene Set Enrichment Analysis) :

Pour le cluster des Monocytes, en définissant un seuil de p-valeur à 0.25 : on a aucun résultat qui dépasse la p-valeur définie pour toutes les collections de gènes.

Enrichissement :

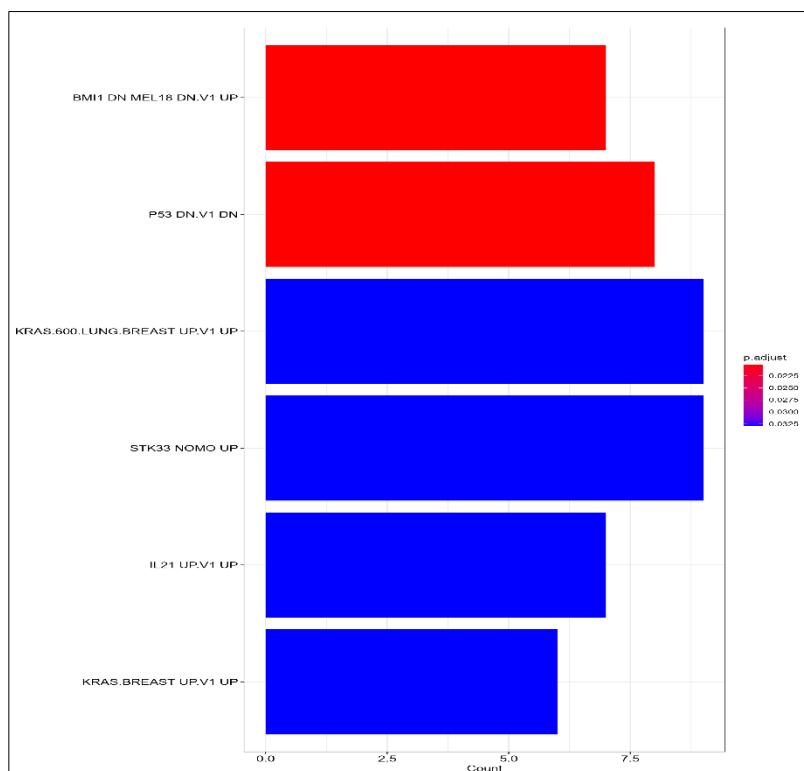
Parmi l'ensemble des gènes différentiellement exprimés, les voies de signalisation enrichies qui ont été identifiées grâce à leurs p-valeurs ajustées faible pour le cluster des monocytes sont :

Pour la catégorie "C6" (Oncogenic signatures gène sets) :

STK33_NOMO_UP : Gènes surexprimés dans les cellules NOMO-1 (AML : acute myeloid leukemia) après l'inactivation du gène STK33.

P53_DN.V1_DN : Gènes réprimés dans le panel NCI-60 de lignées cellulaires avec TP53 (gène suppresseur de tumeur) muté.

IL21_UP.V1_UP : Gènes régulés positivement dans les cellules Sez-4 (lymphocytes T) qui ont d'abord été privées d'IL2, puis stimulées avec IL21.



IV- Conclusion :

Durant cette étude, on a souhaité étudier l'impact de l'expression excessive de MIF sur la différenciation des progéniteurs myéloïdes. D'après les résultats obtenus en comparant les proportions cellulaires entre CTRL et MIF, on a constaté que les proportions monocytaire des deux échantillons sont presque égales avec une légère augmentation du nombre des monocytes chez l'échantillon MIF par rapport à l'échantillon CTRL, ainsi qu'une légère augmentation du nombre des granulocytes chez l'échantillon CTRL par rapport à l'échantillon MIF.

On a essayé de confirmer statistiquement ces résultats en appliquant un test Chi2 sur les proportions cellulaires obtenues. D'après les résultats obtenus, on ne peut pas confirmer que l'expression excessive de MIF a un impact significatif sur les différences des proportions cellulaires entre les deux échantillons CTRL et MIF.

Pour aller plus loin, on a réalisé une analyse différentielle pour déterminer si les monocytes et les granulocytes présentent des différences d'expressions entre nos conditions.

Une analyse GSEA et d'enrichissement a été appliquée aux gènes difféntiellement exprimés entre les granulocytes et les monocytes des deux conditions : plusieurs voies de signalisations ont été détectées.

Pour les granulocytes, on constate principalement une implication de certains gènes surexprimés dans les cellules progénitrices hématopoïétiques CD34+, suite à une activation de la protéine RPS14 ainsi qu'une régulation positive de certains gènes au niveau des cellules tumorales mammaires.

Pour les monocytes, on constate une implication de gènes surexprimés au niveau des cellules AML suite à une inactivation du gène STK33, de gènes réprimés dans le panel NCI-60 de lignées cellulaires suite à une mutation TP53 ainsi qu'une régulation positive au niveau des lymphocytes T.

Cependant ces voies de signalisation doivent être d'avantage étudiées.

V - Lien GitHub Scripts R :

Lien GitHub pour accéder au Scripts R :
https://github.com/maha652/Projet_long_singlecell_RNAseq_MG.git

VI - Bibliographie :

- [1] Place de l'étude des sous-populations monocytaires par cytométrie en flux dans le diagnostic de la LMMC, Anne Ducos 2017.
 - [2] Genecards.org .
 - [3] Chemical and biological single cell analysis, February 2010.
 - [4] Single-cell analyses to tailor treatments, Shalek and Benson (2017) Science Translational Medicine.
 - [5] Best Practices for Preparing a Single Cell Suspension from Solid Tissues for Flow Cytometry.
 - [6] Single-cell RNA sequencing technologies and bioinformatics pipelines ,pipelines, 2018.
 - [7] Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control.
 - [8] What are the applications of single-cell RNA sequencing in cancer research: a systematic review.
 - [9] Leucémie myélomonocytaire chronique : diagnostic et thérapeutique.
-
- [10] Macrophage migration inhibitory factor promotes tumor growth and metastasis by inducing myeloid-derived suppressor cells in the tumor microenvironment .
 - [11] Bulk RNA Sequencing vs. Single Cell RNA Sequencing – what's the difference between these powerful techniques? , March 15, 2023 .
 - [12] <https://github.com/gustaveroussy/single-cell>
 - [13] <https://satijalab.org/seurat/>
 - [14] https://docs.sylabs.io/guides/3.0/user-guide/quick_start.html
 - [15] <https://snakemake.readthedocs.io/en/stable/>
 - [16] EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data.
 - [17] DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors.
 - [18] scds: computational annotation of doublets in single-cell RNA sequencing data.