Développement d'un pipeline automatisé d'analyse de données de cytométrie de masse pour l'identification de populations immunitaires

Juliette Maes

AGROCAMPUS OUEST

☐ **CFR Angers**

☑ **CFR Rennes**

Année universitaire : 2018-2019 Spécialité : Cursus Ingénieur Agronome Spécialisation (et option éventuelle) : Master BMC - Biologie Moléculaire et Cellulaire	**Mémoire de fin d'études** ☑ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☐ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☐ d'un autre établissement (étudiant arrivé en M2)

Développement d'un pipeline automatisé d'analyse de données de cytométrie de masse pour l'identification de populations immunitaires

Par : Juliette MAES

Soutenu à Rennes le 11 juin 2019

Devant le jury composé de :

Frédéric Lecerf

Claire Piquet-Pellorce

Jean-Pierre Tassan

Résumé long

Immune cell subpopulations identification and characterization is essential for the understanding of immune diseases and the development of immunotherapies against cancer. For a long time, flow cytometry has been the most widely used technique to study immune cell heterogeneity. Today, mass cytometry represents a great opportunity to increase the number of features that can be surveyed to a level never reached before. By coupling flow cytometry and mass spectrometry, isotope-tagged antibodies can be used to measure the expression of over 40 cellular markers simultaneously with a high resolution. This technology has led to the generation of high dimensional data hard to analyze manually. Indeed, such analysis is time-consuming, hardly reproducible and subjective. This is why an automated approach for data analysis appears necessary. In this article, we present a new pipeline which enables a dependable, fast and objective way to analyze mass cytometry data while limiting the intervention of the user.

The pipeline is divided into several parts: data preprocessing, quality control, subpopulations identification and finally the display of identified populations. We proposed two different approaches for subpopulation identification which can be used separately or in association to compare the obtained results.

The first one is based on the *OpenCyto* framework. It involves the automation of manual gating usually conducted for flow cytometry data. This supervised technique rests on the choice of a gating strategy by the user, containing the populations to identify within the data set and the markers needed to segregate the cells. Then, the gating strategy is automatically and objectively applied to the data. The framework has been enriched with two new gating functions developed and presented in this paper. The first one enables the gating of dense cell regions whereas the second one makes possible the split of a sample into two populations based on their expression profile for a marker. This method is particularly interesting for routine experiments and multi-site analysis.

The second approach for cell type identification relies on the use of machine learning techniques. This is an unsupervised clustering method organized in two levels. First, an artificial neural network called self-organizing map is built with the FlowSOM algorithm in order to gather cells into 100 nodes. Secondly, a consensus clustering is performed to distribute those nodes to a smaller number of clusters. Finally, clusters are manually assigned to a cell population.

The last step of our pipeline is data representation. Because of the structure and the high dimensionality of mass cytometry data, they cannot be displayed with classical linear dimension reduction techniques such as Principal Component Analysis. Therefore, we compared the representation quality of two non-linear embedding techniques: t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). We found that UMAP algorithm better highlighted the structure of the data. We also noticed that clusters identified by unsupervised clustering matched the ones build with UMAP very well.

The pipeline has been tested on a pilot study led by Sanofi and Bioaster in which 45 surface and intra-cellular markers were surveyed. The aim of this experiment was to assess the effect of several mono-specific antibodies on T cell activation. Both supervised and unsupervised techniques succeed in identifying the major T cell populations. They were also able to identify a drop in the number of naïve T cells and an enrichment in memory CD4 and CD8 T cells when comparing samples activated by different antibodies to control ones.

In a short time, a graphical user interface will be developed in order to facilitate the access and the implementation of the pipeline.

Remerciements

Je remercie tout d'abord Franck Augé pour m'avoir accueilli au sein de son équipe de bioinformatique chez Sanofi.

Je remercie chaleureusement mon maître de stage Charles Bettembourg pour son encadrement bienveillant et sa disponibilité qui m'ont permis d'aborder avec confiance un nouveau domaine de la biologie et de mener à bien le projet qui m'a été confié.

Je remercie également vivement Dorine Chassin, Céline Lefranc et Nicole Meritet pour leurs judicieux conseils et l'intérêt qu'elles ont porté à mon projet.

Enfin, merci à toute l'équipe pour leur convivialité et l'ambiance de travail chaleureuse.

Table des matières

Liste des abréviations

ADN : acide désoxyribonucléique

CC : Consensus Clustering

CD : cluster de différentiation

CMF : cytométrie en flux

CyTOF: cytométrie de masse (ou Cytometry by Time Of Flight)

FlowSOM : Flow Self Organizing Map

ICP : torche à plasma (ou Inductively Coupled Plasma)

LT : lymphocytes T

MDS : positionnement multidimensionnel (ou multi-dimensional plot)

MST : arbre couvrant de poids minimal (ou Minimum Spanning Tree)

NK : cellules tueuses naturelles (ou natural killer)

Treg : lymphocytes T régulateurs

tSNE : t-distributed Stochastic Neighbor Embedding

Uma : unité de masse atomique

UMAP : approximation et projection uniforme de variétés (Uniform Manifold Approximation and Projection)

PBMC : cellule mononuclée du sang périphérique

RPMI: Roswell Park Memorial Institute medium

1. Introduction

La cytométrie de masse (CyTOF pour Cytometry by Time Of Flight) est une technologie développée dans la fin des années 2000 par Bandura et al. (2009), permettant de mesurer l'expression de plus d'une quarantaine de marqueurs membranaires et intracellulaires simultanément, à une résolution dite « single cell ». Elle repose sur l'association de la cytométrie en flux et du spectromètre de masse. Des isotopes stables de métaux principalement de la classe des lanthanides sont couplés à des anticorps ciblant des antigènes d'intérêts (Han et al., 2018). Une fois marquées, les cellules passent individuellement dans le cytomètre de masse via un dispositif microfluidique pour y être nébulisées en fines gouttelettes, fragmentées puis ionisées sous forme mono chargée grâce à une torche à plasma (ICP). Un quadripôle permet de ne filtrer que les ions dont le rapport masse/charge (m/z) est supérieur à 80 uma. Ces ions sont ensuite séparés dans un analyseur à temps de vol et leur intensité est détectée. Grâce aux spectres générés, il est possible d'extraire les niveaux d'expression des marqueurs étudiés pour chaque cellule.

La cytométrie en flux (CMF) est la méthode la plus répandue pour étudier l'expression de marqueurs dans des cellules sanguines. Cependant, les phénomènes d'autofluorescence et de chevauchement des spectres d'émissions des fluorochromes limitent aujourd'hui l'augmentation du nombre de marqueurs mesurables simultanément (Bendall et al., 2012). La CyTOF présente l'avantage considérable de pouvoir quantifier un nombre quatre à cinq plus important de paramètres sans perdre en qualité de mesure. Si aujourd'hui la plupart des études sont réalisées avec une quarantaine d'anticorps différents, il est théoriquement possible d'en utiliser jusqu'à cent différents par cellules (Bandura et al., 2009). Cette technologie permet d'envisager une approche plus exploratoire de l'identification de sous-populations cellulaires et des mécanismes de différentiation ainsi que la mise en évidence des variations de fréquences de populations immunitaires en conditions pathologiques.

L'analyse des données de CyTOF représente aujourd'hui un véritable enjeu. En effet, bien que de nombreuses méthodes aient été proposées, aucune n'a encore fait consensus au sein de la communauté scientifique. Il est possible de traiter les données de CyTOF manuellement, de la même façon qu'en CMF, d'autant plus que la majorité des outils informatiques d'analyse comme FlowJO ou Cytobank sont aussi utilisables en CyTOF. Dans ce cas, les cellules sont représentées sous forme d'un nuage de points en fonction de deux marqueurs. A l'aide d'outils graphiques de sélection, elles sont ensuite manuellement regroupées en sous-populations selon leurs expressions relatives pour ces marqueurs : c'est le « gating ».

Cependant, l'étude de 40 marqueurs conduit théoriquement à générer 780 nuages de points par échantillon. Une analyse manuelle semble donc fastidieuse, d'autant plus qu'elle est très dépendante de l'analyste et peu reproductible à l'échelle d'études multi-sites. Il existe donc un véritable besoin de développement d'une méthode d'identification de sous-populations cellulaires qui soit à la fois rapide, objective, reproductible et automatisée. Pour tenter de répondre à cette question, différentes méthodes ont été publiées ces dernières années. Elles reposent sur deux stratégies distinctes : la semi-automatisation du gating et l'identification de sous-populations cellulaires par clustering.

Le gating semi-automatique est une méthode avec à priori qui repose sur le choix par l'expérimentateur d'une stratégie de gating. Il s'agit d'une suite de combinaisons de deux marqueurs utilisée pour générer des nuages de points et identifier des sous-populations. Une fois définie, cette stratégie est appliquée automatiquement à chaque échantillon d'une même expérience. Des packages R comme *Opencyto* développé par Finak et al. (2014) permettent de standardiser l'identification des sous-populations cellulaires objectivement et sans intervention de l'expérimentateur, une fois la stratégie de gating définie. Cependant, cette stratégie rend difficile l'identification de nouveaux types cellulaires ou de cellules en cours de différentiation.

Les méthodes de clustering, quant à elles, sont des approches sans à priori qui permettent de partitionner un échantillon en différents groupes de cellules. En considérant l'intégralité des marqueurs d'intérêts, les cellules sont regroupées selon leurs degrés de proximité. Une dizaine d'algorithmes tels que Flow Self-Organizing Map (FlowSOM) (Van Gassen et al., 2015), X-shift (Samusik et al., 2016), Spanning-tree Progression Analysis of Density-normalized Events (SPADE) (Qiu et al., 2011) et CITRUS (Bruggner et al., 2014) ont été développés pour former des clusters correspondant théoriquement aux différents types cellulaires au sein des échantillons. D'autres méthodes ont aussi été adaptées à partir d'autres technologies, comme par exemple ConsensusClusterPlus (Wilkerson and Hayes, 2010), issu du Consensus Clustering (Monti, 2003), initialement utilisé pour l'analyse de séquençage d'ARN single cell. La difficulté réside aujourd'hui dans la capacité de ces algorithmes à définir automatiquement le nombre optimal de clusters à identifier au sein d'un échantillon puis à leur associer la sous-population cellulaire à laquelle ils appartiennent. En effet, le nombre optimal défini algébriquement ne correspond que rarement à la réalité biologique. Aujourd'hui, aucune méthode n'est assez performante pour associer automatiquement et avec exactitude chaque cluster à un type cellulaire, notamment en ce qui concerne les populations peu abondantes ; cette étape nécessitant toujours la supervision d'un expert.

Après avoir identifié différents clusters, il est d'usage de les représenter sur un graphique en deux ou trois dimensions. Cependant, les visualiser sur un graphique en fonction de deux marqueurs seulement entraine une perte non négligeable d'information sur la structure des données. Ce sont donc des algorithmes non linéaires de visualisation et de réduction de dimensions qui sont utilisés. Ces derniers doivent permettre de représenter des données à plus de quarante dimensions, dans un espace à deux ou trois dimensions seulement, tout en conservant leur structure. De cette façon, l'expérimentateur peut valider la cohérence des résultats mais aussi mettre en lumière des différences entre les échantillons selon leurs conditions de traitement. Certains des algorithmes évoqués précédemment comme SPADE et FlowSOM permettent aussi de réduire les dimensions d'un jeu de données et de visualiser la hiérarchie des clusters. Ce sont les algorithmes d'apprentissage d'intégration stochastique du voisinage t-distribué (t-distributed stochastic neighbor embedding ou tSNE) (Van der Maaten and Hinton, 2008) et d'approximation et projection uniforme de variétés (UMAP ou Uniform Manifold Approximation and Projection) (McInnes et al., 2018) qui sont les plus utilisés en analyse de CyTOF. Ces méthodes non supervisées et plus exploratoires pourraient permettre la mise en évidence de nouveaux marqueurs impliqués dans des processus biologiques, des types cellulaires rares mais aussi les différenciations cellulaires, grâce à la hiérarchie existante entre les clusters.

Devant le nombre considérable de méthodes disponibles pour traiter des données de CyTOF, il est nécessaire de les confronter et de déterminer la meilleure méthode d'identification des sous-populations cellulaires. L'objectif de cet article est donc de définir un pipeline d'analyse le plus objectif, rapide et automatisé possible, qui limite l'intervention de l'expérimentateur tout en étant accessible à un public non bio-informaticien. Le choix de la méthode d'analyse (supervisée ou non) sera laissé à l'utilisateur et il lui sera possible de comparer les résultats obtenus. Les expressions de marqueurs cibles et les modulations de populations immunitaires pourront alors être étudiées chez les sous-populations identifiées entre différentes conditions (état pathologique, traitement, etc.). Bien que développé pour des données de CyTOF, ce pipeline pourra aussi être implémenté en CMF.

Le pipeline a été mis en place à partir d'une étude pilote menée par Sanofi et Bioaster. Celle-ci s'inscrit dans le cadre de la recherche d'anticorps mono ou multi-spécifiques inhibiteurs de check-point, permettant de lever la tolérance immunitaire envers les cellules tumorales. Pour cette étude, les cellules mononuclées sanguines périphériques (PBMC) de deux donneurs sains ont été cultivées et traitées avec 5 anticorps différents afin d'observer leur effet activateur des lymphocytes T. Les cellules ont été marquées avec 45 anticorps

repartis en deux panels différents et ciblant des protéines membranaires, des cytokines ou des facteurs de transcription. Elles ont ensuite été analysées au cytomètre de masse. Grâce à cette étude, on souhaite savoir si les données de CyTOF mettent en évidence des variations de populations cellulaires entre les différentes conditions d'activation et les témoins. A terme, on cherchera également à identifier des biomarqueurs d'activité clinique lorsque les cellules sont activées par différents anticorps monoclonaux.

2. Matériels et Méthodes

Origine et culture des PBMC

Les couches leuco-plaquettaires fraiches de deux donneurs conservées dans une solution anticoagulante de citrate, phosphate et dextrose ont été fournies par l'Etablissement Français du Sang de Rungis. Les PBMC de ces échantillons ont été isolées par gradient Ficoll puis cultivées dans du milieu Roswell Park Memorial Institute medium (RPMI) (10 % de sérum de veau fœtal et 1 % de pénicilline/streptomycine), à 37 °C et dans une atmosphère contenant 5 % de CO_2. Pendant 6 jours, les cellules ont été stimulées avec un anticorps anti-CD3 à 1 µg/ml seul ou en combinaison avec un anti-CD28 à 0,7 nM, du CD137L à 70 nM, de l'Urelumab (anti 4-1BB ou CD137, BMS-663513) à 70 nM, ou un isotype de l'Urelumab (IgG4 S228P) à 70 nM. Une partie des PBMC n'a pas été stimulée afin de servir de témoin.

Marquages cellulaires et panels d'anticorps

Après six jours de stimulation, les cellules mortes ont été marquées avec un intercalant de l'ADN enrichi en ^{103}Rh (Cell-ID^{TM} Intercalator-103Rh) selon le protocole Fluidigm *PRD004 Version 7*. Les récepteurs Fc ont été bloqués par incubation avec du réactif de blocage de Miltenyi Biotech pendant 10 min à 4 °C afin d'éviter un marquage aspécifique. Les protéines membranaires ont été marquées par le cocktail d'anticorps de surface des panels 1 et 2 séparément (cf. Tableau 1) en suivant le protocole Fluidigm *PN 400276 A4*. Les cellules ont ensuite été fixées par incubation à 4 °C durant 15 min avec du tampon MaxPar® Fix I (Fluidigm) pour le panel 1 ou durant 40 min avec du tampon FoxP3 Fixation/Perméabilisation (eBioscience). Elles ont alors été lavées avec du tampon de perméabilisation MaxPar® Perm-S (Fluidigm). Les cytokines et facteurs de transcription ont été marqués pendant 30 min avec les cocktails d'anticorps intracellulaires du panel 1 et 2 séparément. Enfin, un agent intercalant enrichi en $^{191/193}$Ir (MaxPar® Intercalator-Ir, Fluidigm) a été ajouté suivant le protocole *PRD006 Version 6* de Fluidigm. Les cellules ont été reprises dans un tampon de marquage (Fuidigm) à une concentration de 5.10^5 cellules/ml et conservées à 4 °C.

Tableau 1 Panel des anticorps utilisés.

Les anticorps spécifiques au panel 1 ou 2 sont indiqués en bleu et orange respectivement. F/L : fonctionnel/lignée

Localisation	Métal	Isotope	Cluster de différenciation et autres marqueurs	Rôle	Antigène cible	Fournisseur
cytoplasme	Dy	163	CD107a	F	LAMP-1 (Lysosome-associated membrane glycoprotein 1)	Ozyme
Membrane plasmique	Y	89	CD45	L	PTPRC (protein tyrosine phosphatase receptor type C)	Fluidigm
	Cd	112	CD38	F	ADP-ribosyl cyclase	Qdot655
	In	115	CD85j	F	LIR-1 (Leukocyte immunoglobulin-like receptor subfamily B)	Ozyme-Biolegend
	Pr	141	CD39	F	ENTPD1 (Ectonucleoside triphosphate diphosphohydrolase 1)	Ozyme-Biolegend
	Nd	142	CD223	F	LAG3 (Lymphocyte-activation gene 3)	Ozyme-Biolegend
	Nd	143	CD183	F	CXCR3 (C-X-C Motif Chemokine Receptor 3)	Ozyme
	Nd	144	CD69	F	CLEC2C (C-type lectin domain family 2 member C)	Ozyme
	Nd	145	CD4	L	T-cell surface glycoprotein CD4	Miltenyi
	Nd	146	CD8	L	T-cell surface glycoprotein CD8	-
	Nd	148	TIGIT	F	T cell immunoreceptor with Ig and ITIM domains	-
	Nd	148	CD155	F	PVR (poliovirus receptor)	Ozyme-Biolegend
	Sm	149	CD25	L	IL2RA (Interleukin-2 receptor alpha chain)	Miltenyi
	Nd	150	CD27	L	TNFRSF7 (Tumor necrosis factor receptor superfamily 7)	Ozyme
	Eu	151	CD197	L	CCR7 (C-C chemokine receptor type 7)	Miltenyi
	Eu	153	CD28	F	T-cell-specific surface glycoprotein CD28	Ozyme-Biolegend
	Sm	154	CD196	F	CCR6 (C-C chemokine receptor type 6)	Miltenyi
	Gd	156	CD279	F	PD-1 (Programmed cell death protein 1)	Miltenyi
	Gd	158	CD95	F	TNFRSF6 (Tumor necrosis factor receptor superfamily member 6)	Ozyme-Biolegend
	Gd	161	CD357	F	TNFRSF18, GITR (Tumor necrosis factor receptor superfamily member 18)	Ozyme-Biolegend
	Dy	162	CD274	F	PDL-1 (Programmed cell death 1 ligand 1)	Ozyme-Biolegend
	Dy	164	CD137	F	4-1BB TNFSF9 (Tumor necrosis factor receptor superfamily member 9)	Ozyme-Biolegend
	Ho	165	CD185	F	CXCR5 (C-X-C chemokine receptor type 5)	Miltenyi
	Er	167	CD134	F	OX40 TNFSF4 (Tumor necrosis factor receptor superfamily member 4)	Ozyme-Biolegend
	Er	168	CD56	L	NCAM-1 (Neural cell adhesion molecule 1)	Ozyme
	Tm	169	CD45RA	L	isoforme de CD45 (protein tyrosine phosphatase receptor type C)	-
	Er	170	CD3	L	T-cell surface glycoprotein CD3	Miltenyi
	Yb	171	CD366	F	TIM3 HAVCR2 (Hepatitis A virus cellular receptor 2)	Ozyme-Biolegend
	Yb	173	CD294	F	CRTH2 (Prostaglandin D2 receptor 2)	Miltenyi
	Lu	175	HLADR	L	human leukocyte antigen DR isotype	Miltenyi
	Yb	176	CD127	L	Interleukin-7 receptor	Miltenyi
	Pt	194	CD152	F	CTLA-4 (Cytotoxic T-lymphocyte protein 4)	Ozyme-Biolegend
	Pt	198	CD278	F	ICOS (Inducible T-cell costimulator)	Miltenyi
	Bi	209	CD16	L	FCGR3A (Low affinity immunoglobulin γ Fc region receptor III-A)	Fluidigm
Nucléaire	In	115	RORγ	L	Nuclear receptor ROR-γ	Miltenyi
	Ce	140	Ki67	F	Proliferation marker protein Ki-67	Miltenyi
	Gd	155	Tbet	L	T-box expressed in T cells	Miltenyi
	Er	166	FoxP3	L	Forkhead box protein P3	BD pharmingen
	Yb	172	Eomes	L	Eomesodermin	R&D systems
Sécrété	Sm	147	GRANZB	F	Granzyme B	Miltenyi
	Sm	152	PERFORIN	F	Perforin	Ozyme-Biolegend
	Gd	155	IL4	F	Interleukine 4	Miltenyi
	Tb	159	IL5	F	Interleukine 5	R&D systems
	Gd	160	CD101	F	IgSF2 (Immunoglobulin superfamily member 2)	NOVUS-Biotechne
	Dy	163	IL10	F	Interleukine 10	Miltenyi
	Er	166	TNFα	F	Tumor Necrosis Factor α	Miltenyi
	Yb	172	IFNγ	F	Interferon γ	Miltenyi
	Yb	173	IL2	F	Interleukine 2	Miltenyi
	Yb	174	CD40L	F	Cytokine CD40 Ligand	-
	Rh	103	VIABILITE	-	Agent intercalant	Fluidigm
	Ir	191	CELLS1	-	Agent intercalant	Fluidigm
	Ir	193	CELLS2	-	Agent intercalant	Fluidigm

Acquisition des données par cytométrie de masse

Le lendemain du marquage, un volume de 500 µl de chaque échantillon, contenant des cellules à 5.10^5 cellules/ml ont été analysés successivement par un CyTOF2. Le marquage cellulaire et l'acquisition ont été réalisés par l'entreprise Bioaster.

Contrôle qualité des données

Les graphiques de positionnements multidimensionnels (MDS) et de densité d'expressions de marqueurs dans les différents échantillons ont été générés comme présentés par Nowicka et al. (2017).

Identification de sous-types cellulaires par gating semi-automatique

Import des données et gating

L'ensemble des fichiers fcs a été converti en flowFrame et rassemblés dans un flowSet grâce à la fonction *read.flowSet* du package *flowCore*. Les objets ayant été générés, sont de type S4 et contiennent notamment une matrice d'expression pour chaque marqueur, le nom de l'échantillon et le nom des marqueurs associé aux isotopes utilisés, le tout pour chaque échantillon de l'expérience. La fonction *GatingSet* a été utilisée pour convertir le flowSet en GatingSet, objet compatible avec le package o*penCyto*. La stratégie de gating a ensuite été appliquée automatiquement à tous les échantillons grâce à la fonction *gating* du package *openCyto*. La fonction prend en argument un gatingSet ainsi qu'un template. Cette fonction est parallélisable pour augmenter sa vitesse d'exécution. Dans le cadre de cette analyse, elle a tourné sur 10 cœurs de processeur. Les résultats du gating ont été visualisés grâce aux packages *ggplot2* et *ggcyto*. Les statistiques des différentes populations ont été générées par le biais de la fonction *getPopStat* du package *flowWorkspace* puis exportées en format csv. La fonction *getData* crée un flowSet regroupant uniquement les mesures d'expression des cellules appartenant à une population précise, pour chaque échantillon testé. Ce fichier a été enregistré sous forme de fichiers fcs indépendants pour chaque échantillon du flowSet avec la fonction *write.flowSet* du package *flowCore*.

Utilisation de template

Les tableaux aux formats csv, txt ou xlsx ont été importés et convertis en templates grâce à la fonction *gatingTemplate* du package *openCyto*. Ils doivent respecter l'ordre et le nom des colonnes présentés dans la documentation R de la fonction.

Architecture des fonctions de gating supplémentaires

Pour compléter les méthodes de gating de populations cellulaires, les fonctions *densityGate* et *separationGate*, nouvellement développées, ont été ajoutées au package *openCyto*. Leur architecture est similaire. Elles prennent en paramètres d'entrée un flowframe, un ou deux marqueurs ainsi que d'autres variables facultatives permettant l'ajustement du gating effectué. *DensityGate* est une fonction d'emballage (« wrapping ») qui fait appel à une sous-fonction

search.minimumDensity. Cette dernière calcule pour un marqueur et un flowFrame donnés, le positionnement optimal d'une gate pour qu'elle sépare au mieux deux régions de cellules denses. Cette gate est utilisée par la fonction d'emballage pour découper l'échantillon en deux populations positives ou négatives par rapport à ce marqueur et ainsi y répartir les cellules du flowFrame. Pour que la fonction de gating soit reconnue et utilisable dans les templates, elle a été enregistrée grâce à la fonction *registerPlugins* d'*openCyto*.

Identification de sous types cellulaires par clustering non supervisé

Le clustering non supervisé des échantillons a été réalisé en quatre étapes. D'abord, les fichiers fcs (préalablement transformés par $arcsinh\left(\frac{x}{5}\right)$) ont été convertis en objet flowSOM grâce à la fonction *ReadInput*. Un Self Organizing Map a ensuite été entrainé sur la totalité des cellules, uniquement à partir des 12 marqueurs « lignée ». La fonction *BuildSOM* du package *FlowSOM* a été utilisée, avec une grille de neurones de taille 10x10 et un taux d'apprentissage variant de 0,05 à 0,01. L'apprentissage a été réalisé 10 fois avec initialisations aléatoires de chaque neurone, définie par *set.seed(100)*. Les cellules ont alors été réparties dans 100 nœuds différents. Ces coordonnées de dimensions 100x12, ont ensuite été utilisées comme paramètres d'entrée pour le Consensus Clustering. Il consiste en 100 répétitions d'un clustering hiérarchique ascendant avec un échantillonnage aléatoire de 90 % des nœuds. La fonction du package *ConsensusClusterPlus* a généré le consensus entre ces 100 répétitions et a été réalisé 23 fois pour un nombre de clusters défini allant de 2 à 24. La graine aléatoire a également été déterminée par *set.seet(1420)*. Le nombre optimal de clusters a ensuite été déterminé comme décrit par Van Gassen et al. (2015) et ajusté manuellement en fonction du profil d'expression des marqueurs « lignées » des différents clusters.

Visualisation des données

La réduction de dimension a été réalisée après mise en commun des cellules de tous les échantillons. Seuls les marqueurs annotés « lignée » ont été utilisés pour réaliser la réduction de dimension après avoir été transformés. Ces cellules ont permis de générer des t-SNE grâce à la fonction R*tsne.multicore* du package du même nom, avec 1000 itérations, une perplexité de 30 et en parallèle sur 2 cœurs de processeur. Les représentations UMAP ont été construites à l'aide de la fonction *umap* du package du même nom. La durée d'exécution des algorithmes a été mesurée avec la fonction *system.time*. Les coordonnées issues de la réduction de dimension déterminées par les algorithmes ont ensuite été visualisées sur un graphique en deux dimensions avec le package *ggplot2*. La représentation a été faite en affichant 10 % de la totalité des cellules échantillonnées (pour des questions de lisibilité) ou en affichant

séparément les cellules issues des différents donneurs et conditions d'activation. Les cellules ont ensuite été colorées selon leur appartenance aux clusters identifiés précédemment par FlowSOM puis ConsensusClustering. Les représentations t-SNE et UMAP ont été générés plusieurs fois avec des valeurs des graines aléatoires définies par la fonction *set.seed(1234)* de manière à vérifier la stabilité des groupes formés et de pouvoir reproduire les résultats. L'arbre couvrant poids minimal (MST) a été construit avec la fonction *BuiltMST* du package *FlowSOM*.

Environnement informatique et accessibilité

Le système d'exploitation utilisé est Red Hat Enterprise Linux Server 7.3 Sur ce serveur, 32 cœurs de processeur et 64 Go de mémoire vive sont disponibles. Les algorithmes ont été écrits et exécutés sur la version 3.4.3 de R. La version des packages utilisés est la plus récente compatible avec cette version de R.

3. Résultats

Le pipeline d'analyse est divisé en plusieurs parties : le prétraitement des données, le contrôle qualité, l'identification de sous-populations par gating semi-automatique puis par clustering, la visualisation des sous-populations puis l'analyse différentielle. Les deux méthodes d'identification de sous-populations peuvent être utilisées seules ou en combinaisons. Le pipeline est résumé dans la Figure 1.

Prétraitement des données

Les données brutes sont exportées du cytomètre de masse sous forme de fichiers fcs. Ils comprennent un tableau rassemblant tous les « évènements » mesurés par cytomètre en ligne et les marqueurs d'intérêts en colonne. Un pourcentage non négligeable des évènements correspond à des débris cellulaires, des agrégats ou bien des cellules mortes, qu'il est nécessaire de retirer de l'analyse. Pour se faire, les évènements sont représentés sur un nuage de points en fonction du temps d'acquisition et de l'agent intercalant enrichi en ^{191}Ir. Utilisé comme marquage après la fixation et la perméabilisation des cellules, cet agent permet de quantifier l'ADN présents dans un évènement. Un évènement trop faiblement marqué est associé à un débris cellulaire tandis qu'un trop fort signal correspondrait à un amas cellulaire. En sélectionnant la zone la plus dense, et avec un signal ^{191}Ir moyen, on élimine une partie des évènements non désirés. De la même façon, les cellules sont sélectionnées temps d'acquisition et de la « durée de l'évènement ». Ce paramètre correspond à la durée pendant laquelle le détecteur capte les ions provenant d'un même évènement. Si celui-ci est trop long,

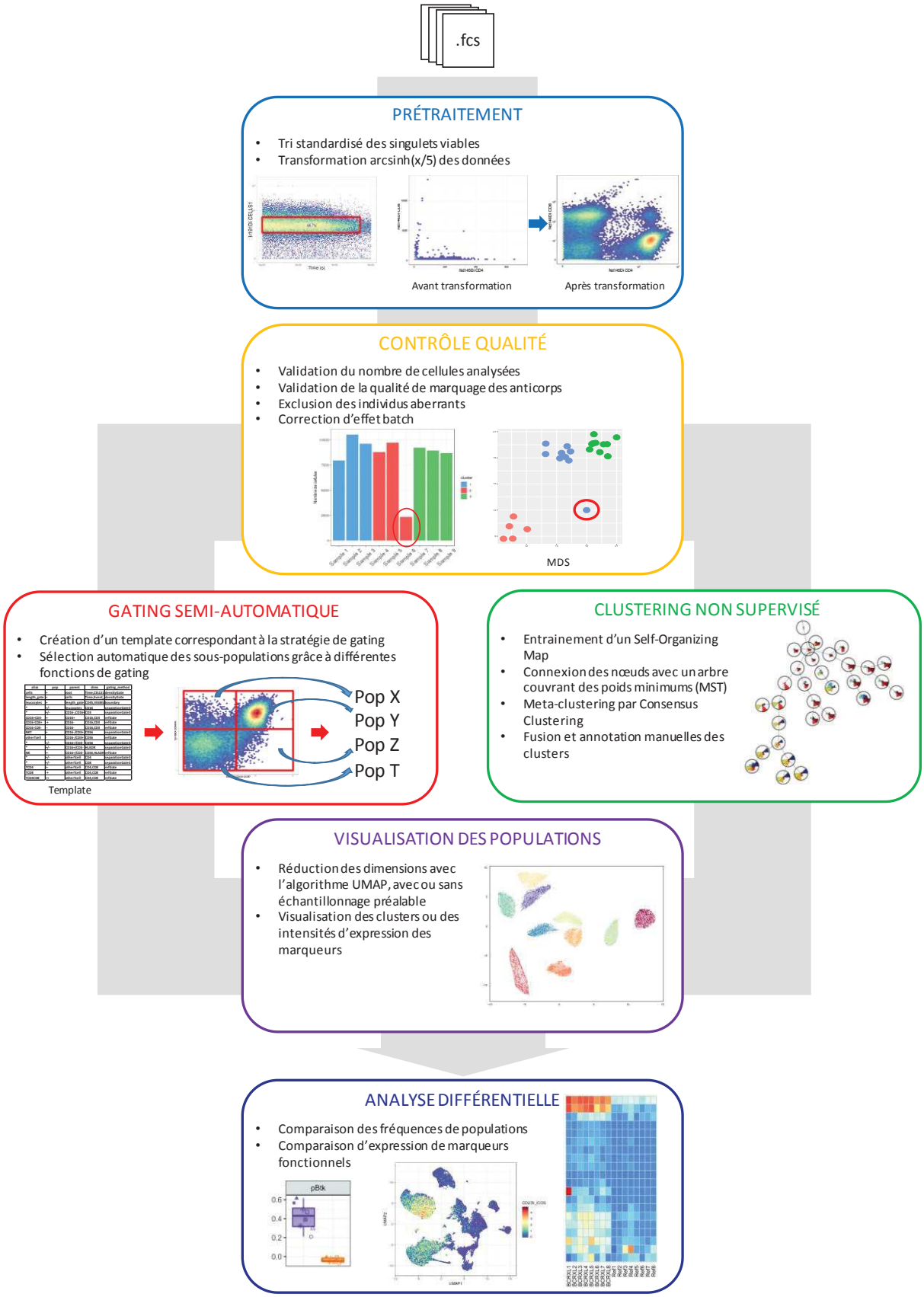


Figure 1 Schéma bilan du pipeline d'analyse de données de cytométrie de masse.

Certaines images sont extraites des articles de Nowicka et al. (2017) et Van Gassen et al. (2015).

il est supposé que plusieurs cellules ont été détectées à des intervalles trop proches et que leurs spectres se sont chevauchés. Enfin, les leucocytes vivants, seules cellules nous intéressant ici, sont sélectionnés grâce au marqueur de viabilité 103Rh et au marqueur CD45. Les mesures d'intensités des marqueurs et des intercalants sont transformées par la fonction $arcsinh(\frac{x}{co-facteur})$, avec un co-facteur égal à 5 pour des données de CyTOF et 150 en CMF (Nowicka et al., 2017). La filtration des données s'effectue automatiquement par le biais du gating semi-automatique présenté plus bas.

Contrôle qualité

Avant de procéder aux analyses, il est nécessaire de contrôler la qualité des données collectées. Certain de ces contrôles ont été présentés par Nowicka et al. (2017). Premièrement, il est important de déterminer si le nombre de cellules est suffisant pour être analysé et du même ordre de grandeur entre les différents échantillons. Ensuite, il est conseillé de tracer un graphique des positionnements multidimensionnels (MDS) des différents échantillons de manière à observer leur répartition pour identifier de potentiels mesures aberrantes à exclure du jeu de données. Enfin, tracer les densités d'expression de chaque marqueur et chaque condition permet de mettre en lumière des effets « batch » entre plusieurs séries d'expérimentations, une anomalie sur un anticorps, ou bien une décroissance du signal au cours de l'acquisition (Finck et al., 2013). Dès lors, une covariance ou un coefficient pourra être appliqué aux données pour les corriger. Les données du projet pilote ont été soumises au contrôle qualité et les résultats sont visibles sur la Figure 2. Le pourcentage de cellules analysables varie jusqu'à un facteur 2 d'un échantillon à l'autre, ce qui pourrait induire un biais au cours de la comparaison entre les différentes conditions. Il serait conseillé de retirer l'échantillon anti-CD3+isotype de l'Urelumab du donneur 2 de l'analyse. Par ailleurs, la Figure 2 B met en évidence une variabilité entre les profils des donneurs aussi importante que celle existant entre la condition contrôle et les différents traitements. Le graphique des densités d'expression des marqueurs (Figure 2 C) permet de vérifier qu'il n'y pas eu de décroissance du signal au cours de l'acquisition. Une fois ces contrôles qualité effectués, il est possible de passer à l'analyse des données.

Identification de sous-populations par gating semi-automatique

Le gating semi-automatique repose sur l'utilisation de l'environnement *OpenCyto* (Finak et al., 2014). L'atout du gating semi-automatique est de remplacer le positionnement manuel des « gates » par un positionnement automatique et objectif. Pour cela, l'utilisateur doit créer un template, c'est-à-dire un tableau dont chaque ligne correspond au gating d'un groupe de

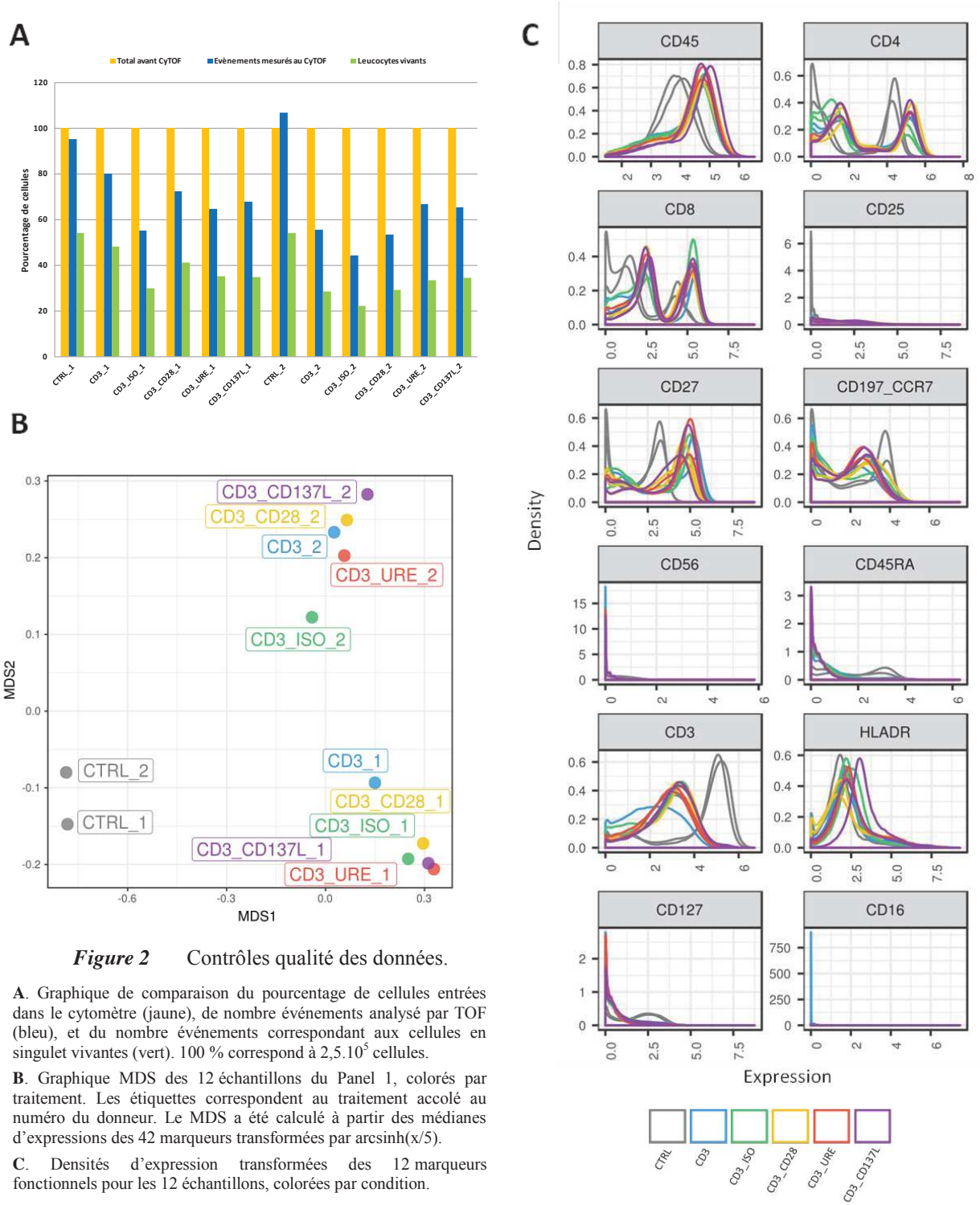


Figure 2 Contrôles qualité des données.

A. Graphique de comparaison du pourcentage de cellules entrées dans le cytomètre (jaune), de nombre événements analysé par TOF (bleu), et du nombre événements correspondant aux cellules en singulet vivantes (vert). 100 % correspond à $2{,}5.10^5$ cellules.

B. Graphique MDS des 12 échantillons du Panel 1, colorés par traitement. Les étiquettes correspondent au traitement accolé au numéro du donneur. Le MDS a été calculé à partir des médianes d'expressions des 42 marqueurs transformées par arcsinh(x/5).

C. Densités d'expression transformées des 12 marqueurs fonctionnels pour les 12 échantillons, colorées par condition.

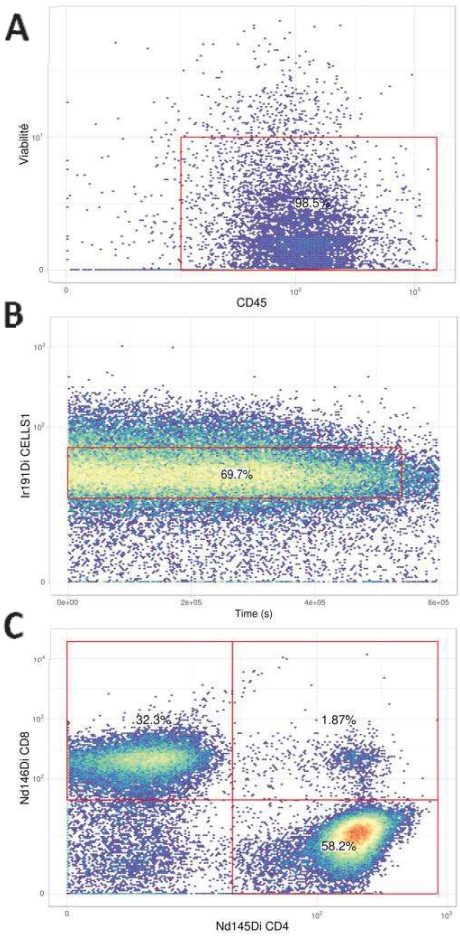
cellules, l'ensemble du tableau constituant la stratégie de gating. Celui-ci est lu par le programme qui l'applique automatiquement aux échantillons et génère ainsi les différentes gates. Pour plus de lisibilité, la stratégie de gating peut être visualisée sous la forme d'une arborescence comme présentée en Figure 3 D. Un template peut être appliqué à tout échantillon tant que les marqueurs qui y sont utilisés sont mesurés dans l'échantillon. Cette méthode est très avantageuse si les expérimentations sont toujours réalisées avec le même panel d'anticorps. De plus, il permet de supprimer la variabilité inter-expérimentateurs. Une

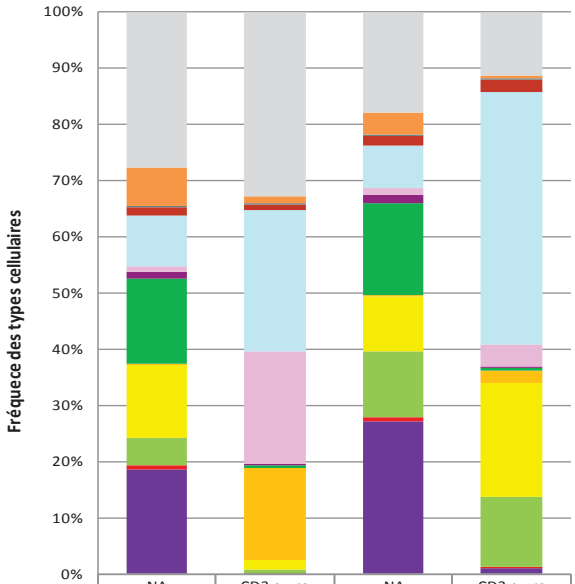
fois le gating terminé, il est possible de visualiser la position des gates et les différentes sous-populations identifiées sous forme de nuages de points (Figure 3 A-C). Enfin, les pourcentages de cellules de chaque sous-population ainsi que leurs mesures d'expressions peuvent être exportés sous forme de tableau au format csv ou de fichier fcs pour des analyses ultérieures.

Dans cet article sont présentées deux fonctions de gating supplémentaires qui viennent enrichir celles disponibles dans le package *openCyto*. La fonction *densityGate* permet d'effectuer une sélection des cellules en forme de rectangle sur la base de leur densité sur le nuage de point. Elle est notamment utilisée pour les deux premières étapes de prétraitement des données (Figure 3 B). La fonction *separationGate* permet quant à elle de distinguer deux populations au sein d'un échantillon. A partir de la courbe de répartition des cellules en fonction de l'expression d'un marqueur, elle détermine la valeur d'expression pour laquelle à la densité de cellules est minimale (Figure 3 C). En d'autres termes, elle cherche la position optimale séparant deux régions denses. Si une cellule a un niveau d'expression supérieur à la limite déterminée, elle sera considérée comme positive pour ce marqueur, et inversement. Cette fonction dispose de nombreux paramètres ajustables tels que l'intervalle de recherche de la position de la gate, la valeur par défaut à renvoyer dans le cas où aucune limite n'est identifiée, etc.

L'algorithme a été appliqué au jeu de données avec le template illustré dans la Figure 3 D et a permis d'identifier plusieurs populations. Les trois premières étapes de la stratégie de gating correspondent à la sélection des leucocytes vivants telles que décrites dans la partie « prétraitement des données ». Le gating permet ensuite d'identifier des populations telles que les lymphocytes T (LT) CD4 et LT CD8 naïfs, effecteurs, de la mémoire centrale et effecteurs à mémoire, les LT doubles positifs, les LT régulateurs (Treg), les cellules tueuses naturelles (NK) et les lymphocytes NKT. Le pourcentage que représente chaque type cellulaire par rapport au nombre total de leucocytes par échantillon est représenté dans la Figure 3 E. Seuls les résultats des PBMC du donneur 1 non traitées et traitées à l'Urelumab sont présentés. Il est mis en évidence une forte augmentation de fréquence des LT CD8 mémoires et une quasi disparition des LT CD4 et CD8 naïfs, dans les conditions d'activation des lymphocytes T. En revanche, il y a une forte variabilité de fréquence entre les deux panels mais aussi entre les donneurs. Cela empêche de faire ressortir une quelconque différence entre les anticorps utilisés pour l'activation de lymphocytes T. La variabilité observée peut être d'origine expérimentale : utilisation d'anticorps aspécifiques, toxicité des protocoles de marquage, etc. Ces éléments peuvent alors modifier les profils d'expression et influencer le positionnement

automatique des gates conduisant à une mauvaise répartition des cellules dans les populations.





	PANEL 1		PANEL 2	
	NA	CD3 + ure	NA	CD3 + ure
Autres	27,7	32,7	17,9	11,2
Cellules NK	6,9	1,3	4,0	0,7
Cellules NKT	0,1	0,3	0,1	0,2
LT CD4+ CD8+	1,5	0,9	1,8	2,1
TCD8 EM	9,1	25,2	7,5	45,0
TCD8 CM	0,9	20,1	1,3	4,0
TCD8 Effecteurs	1,2	0,1	1,4	0,3
TCD8 Naifs	15,2	0,6	16,3	0,4
Treg	0,3	16,3	0,3	2,3
TCD4 EM	12,9	1,8	9,7	20,2
TCD4 CM	5,0	0,5	11,8	12,5
TCD4 Effecteurs	0,8	0,0	0,6	0,3
TCD4 Naifs	18,5	0,1	27,2	1,0

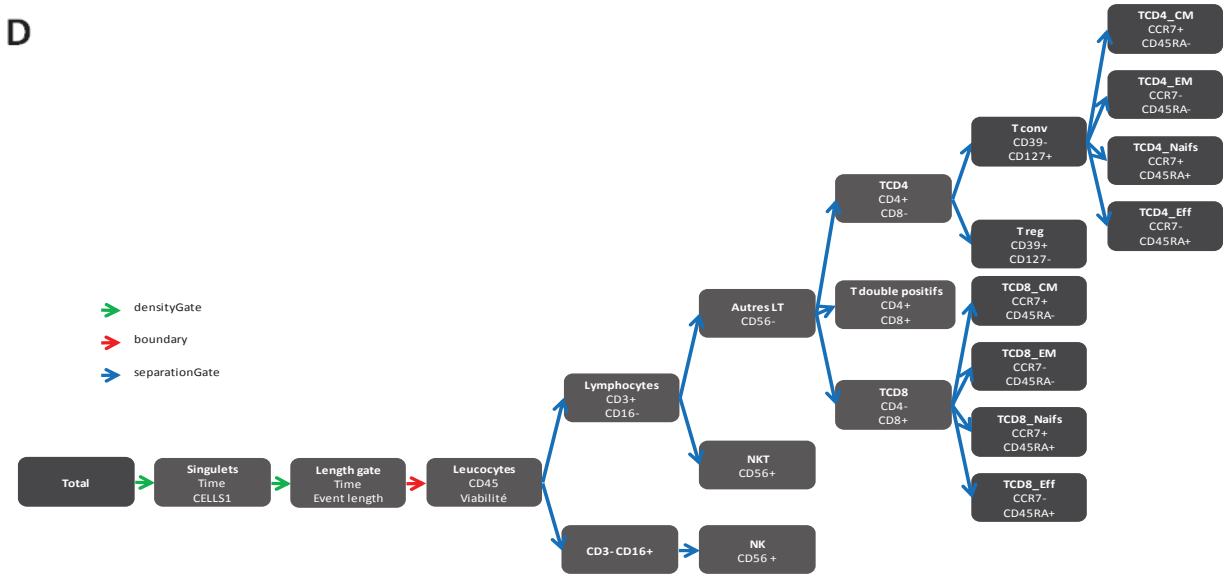


Figure 3 Gating semi-automatique grâce à la structure Opencyto.

A-C. Exemples de gates placées sur des nuages des cellules du donneur 1 en conditions non activées. (**A**) Gating des leucocytes vivants déterminé automatiquement grâce à la fonction boundary, disponible dans le package OpenCyto. Gating des singulets (**B**) et des lymphocytes T CD4, T CD8 et T doubles positifs (**C**) grâce aux fonctions densityGate et separationgate respectivement, développées dans cet article. Les différents rectangles de gating sont attribués à une sous-population tel qu'indiqué dans le template.

D. Illustration de la stratégie de gating du template, utilisée pour identifier les populations cellulaires présentes dans les échantillons du panel 1. Dans chaque bulle, la première ligne correspond au nom de la sous-population et les autres aux noms des marqueurs utilisés. Les quatre premières bulles correspondent au prétraitement des données.

E. Résultats du gating semi-automatique des panels 1 et 2, pour les conditions non activées (NA) et activées avec un cocktail d'anti-CD3 et d'Urelumab (CD3+Ure). Les répartitions des populations cellulaires sont indiquées en pourcentage du nombre de leucocytes vivants totaux.

Identification de sous-population par clustering et la visualisation des données

La suite de ce pipeline propose d'identifier les différentes populations cellulaires grâce à des algorithmes de clustering. Le clustering est partagé en quatre étapes. Premièrement, l'algorithme FlowSOM (Van Gassen et al., 2015) est entrainé sur l'ensemble des données confondues, préalablement arcsinh-transformées et réduites. Il utilise un réseau de neurones artificiel : le Self-Organizing Map introduit par Kohonen (1990) pour répartir la totalité des cellules dans 100 nœuds différents. D'après Weber and Robinson (2016), FlowSOM est l'algorithme de clustering dont le résultat est le plus proche de celui d'un gating manuel réalisé par un expert. Son temps d'exécution est également le plus court de toutes les méthodes testées. Deuxièmement, un arbre couvrant des poids minimums (MST) est tracé afin de visualiser la proximité entre les nœuds. Générer un grand nombre de nœuds permet d'augmenter les chances d'isoler les cellules appartenant à un type cellulaire rare, sans qu'elles ne soient absorbées par un type plus abondant. Troisièmement, un méta-clustering est effectué grâce à l'algorithme ConsensusClustering (CC) (Monti, 2003). Grâce à cet algorithme, les 100 nœuds sont réduits en k clusters, pour être plus proches de la réalité biologique. Le CC repose sur plusieurs itérations d'un clustering hiérarchique réalisé sur un pourcentage de nœuds échantillonnés aléatoirement. Du fait du tirage aléatoire, deux nœuds peuvent ne pas toujours être regroupés au sein du même cluster. On considère que plus des nœuds sont maintenus dans un même cluster au cours des différentes itérations, plus celui-ci est stable. Le résultat du ConsensusClustering sera donc composé des clusters les plus stables. Le nombre idéal k de clusters à former peut être imposé par l'expérimentateur ou bien déterminé mathématiquement grâce au « critère du coude ». Cela correspond à étudier la stabilité globale de partitionnements en x clusters, pour différentes valeurs de x. On étudie ensuite les variations de stabilité lorsqu'on passe d'un partitionnement en x à $x+1$ clusters. Le nombre optimal k de clusters sera atteint lorsque passer de x à $x+1$ clusters n'augmentera plus la stabilité globale du partionnement. En d'autres mots, lorsqu'on aura atteint un plateau sur la Figure 4 C. Cependant, il est conseillé d'effectuer le méta-clustering en surestimant un peu le nombre attendu et de l'ajuster avec une fusion manuelle, de sorte à conserver les clusters peu abondants aux phénotypes marginaux. La dernière étape consiste en l'assignement des clusters formés avec un type cellulaire particulier. Pour cela, aucune méthode ne propose actuellement une assignation automatique. C'est pourquoi cette étape doit être réalisée par l'expérimentateur, en comparant la co-expression des marqueurs « lignée ». Les 100 nœuds formés par le SOM et leurs proximités sont représentés sur la Figure 4 A. Les nœuds sont

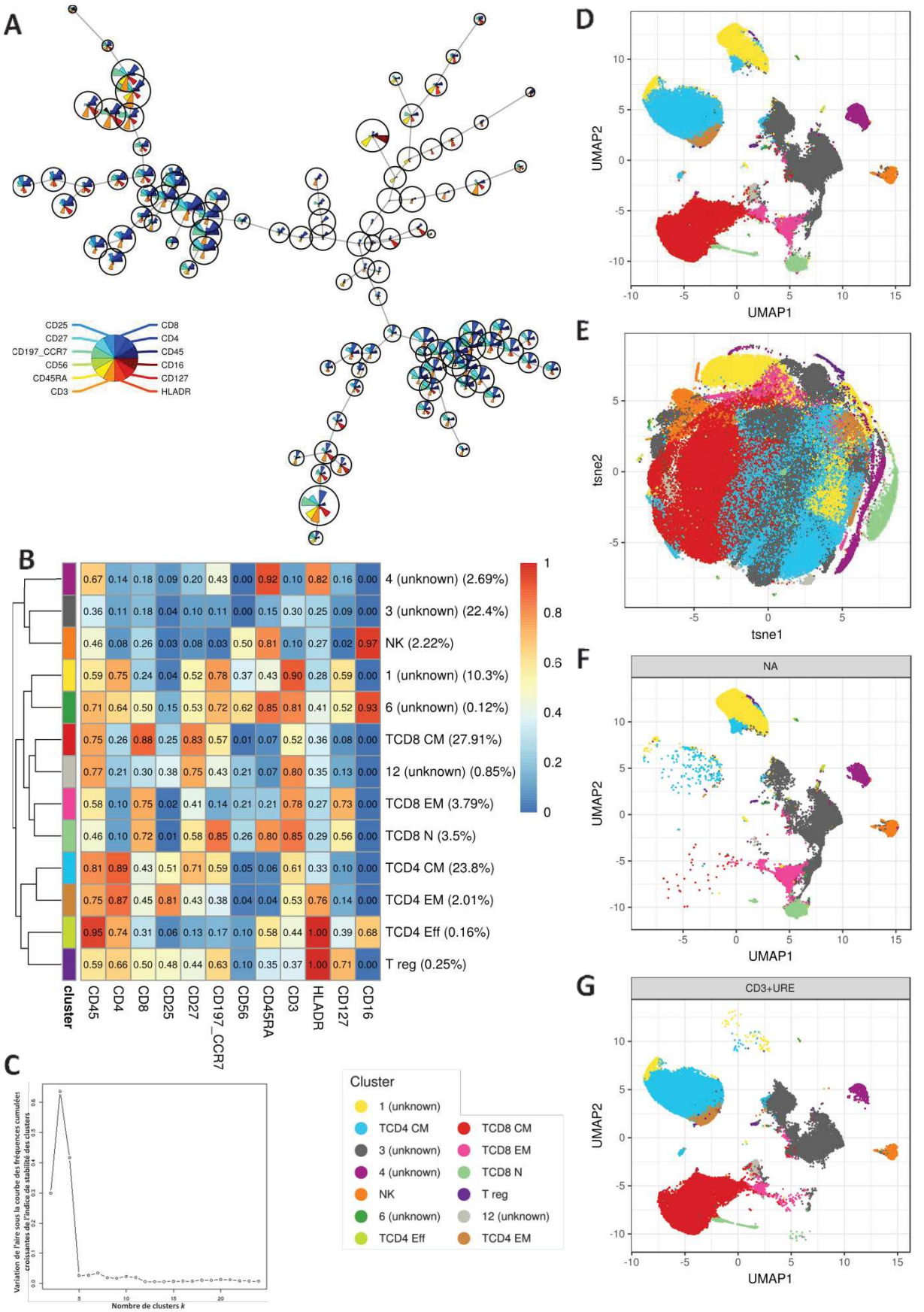


Figure 4 Identification de populations cellulaires par clustering automatique.

A. MST représentant le résultat du clustering par FlowSOM, réalisé sur l'ensemble des échantillons et à partir des 12 marqueurs de lignées transformées (arcsinh(x/5)). Les nœuds sont représentés par des cercles dont la taille est proportionnelle au nombre de cellules qui le composent. Les cercles sont divisés en portions dans lesquelles on peut lire la médiane de l'intensité d'expression des marqueurs.

B. Heat-map de la médiane d'expression centrée et réduite des marqueurs lignées au sein des différents clusters obtenus par Consensus Clustering. Les 13 clusters ont été déterminés après estimation mathématique du nombre de clusters associés à la plus forte stabilité, suivi d'un ajustement manuel. Les clusters ont été annotés manuellement à partir des données d'expression lisibles dans la heat-map.

C. A l'issue du clustering par FlowSOM, les 100 nœuds sont partitionnés grâce à l'algorithme Consensus Clustering, en un nombre k de clusters variant de 2 à 24. Les différents indices de stabilité sont calculés pour chaque clusters formés et pour chaque Consensus Clustering réalisés (pour k=2, k=3,…, k=24). La variation du gain de stabilité lorsqu'on passe d'un partitionnement de k à k+1 clusters est représenté. Le nombre optimal de cluster correspond au point à partir duquel le gain de stabilité est nul.

D-E. Visualisations des graphiques UMAP (**D**) et tSNE (**E**) générés à partir des données d'expressions transformées des marqueurs de lignées de l'ensemble des cellules. Seul 10 % des cellules sont représentées et sont colorées par clusters.

F-G. Graphique UMAP ne représentant que les cellules appartenant au donneur 1 et à la condition non activée (NA) (**F**) ou activée avec un anti-CD3 et de l'Urelumab (**G**) EM : effecteur à mémoire, CM : central à mémoire, Eff : effecteur, N : Naïf, NK : cellule tueuse naturelle, Treg : lymphocyte T régulateur

relativement bien repartis en fonction de leur profil d'expression. Par exemple, les trois branches principales sont différenciables entre autres par leurs profils d'expressions de CD4 et CD8. Il semble que plus les clusters sont éloignés du centre de l'arbre, plus ils ont des profils différenciés. Le méta-clustering a ensuite été appliqué et a permis de partitionner l'ensemble les 100 nœuds en 13 clusters. Ce nombre a été déterminé en combinant l'analyse de la stabilité des clusters (Figure 4 C) et la fusion manuelle. Seulement 8 des 13 clusters ont pu être annotés manuellement. Leur profil d'expression est présenté dans la Figure 4 B. Le pourcentage de chaque population cellulaire identifiée semble cohérent avec les résultats obtenus par gating semi-automatique.

Afin de faciliter la compréhension et l'analyse des résultats, les données sont visualisées sur des graphiques en deux dimensions. Il est primordial de conserver la structure locale des données pour représenter convenablement les clusters : les cellules proches dans un espace à 40 dimensions doivent être proches sur un graphique en 2D. La qualité de représentation de deux algorithmes de réduction de dimensions non linéaires, UMAP (McInnes et al., 2018) et t-SNE (Van der Maaten and Hinton, 2008), ont donc été comparés. La réduction de dimension a été lancée sur la totalité des cellules des différents échantillons. La Figure 4 D-E montre qu'UMAP discrimine mieux les différents clusters, c'est donc l'algorithme qui sera choisi dans le pipeline pour représenter les données. Il est à noter qu'à cause de la nature aléatoire du processus d'initialisation des algorithmes, ces derniers doivent être exécutés plusieurs fois afin de sélectionner la meilleure représentation. Les représentations UMAP sont ensuite utilisées pour comparer les fréquences de populations immunitaires entre différentes conditions d'activation, comme présenté Figure 4 F-G. Pour des raisons de visibilité, seules deux des conditions testées ont été représentées. Ces graphiques mettent en évidence un enrichissement en LT CD4 et CD8 à mémoire centrale et effecteurs dans les conditions activées par rapport à la condition contrôle. Ils soulignent aussi la disparition du cluster 1 lorsque les cellules sont activées. Cela pourrait sous-entendre que ce cluster non identifié à l'issu du clustering correspondrait aux LT CD8 naïfs. Ici encore, aucune différence n'a été observée entre les différentes conditions d'activations.

4. Discussion

Le pipeline proposé dans cet article permet d'appréhender les données à hautes dimensions générées par cytométrie de masse ou de flux. Il se décompose en plusieurs parties : le prétraitement des données, le contrôle qualité, l'identification de sous-populations cellulaires et leur visualisation. En combinant approches supervisée et non supervisée, ce pipeline rend possible aussi bien l'étude de mécanismes connus que la découverte de nouveaux marqueurs impliqués dans la modulation immunitaire en réponse aux pathologies. Le pipeline est développé pour limiter l'intervention de l'expérimentateur et augmenter la vitesse et la reproductibilité de l'analyse. Avec le gating semi-automatique, l'utilisateur décide de la stratégie de gating mais laisse l'algorithme l'appliquer de façon impartiale. Son principe de fonctionnement relativement facile à appréhender constitue une bonne alternative pour les sceptiques vis-à-vis des méthodes non supervisées. Cette méthode est cependant assez sensible à la qualité de l'acquisition, notamment lorsque celle-ci conduit à la génération de profils de cellules aux distributions non binaires sur lesquelles une sélection positive/négative est impossible. Elle a tout de même permis d'identifier plusieurs populations de lymphocytes T dans les échantillons et notamment un enrichissement en lymphocytes T mémoires suite à l'ajout d'agents activateurs.

L'identification de sous-populations par clustering non supervisé semble quant à elle prometteuse car capable de partitionner un échantillon à partir d'un grand nombre de dimensions. En ce qui concerne les données de l'étude pilote, le clustering par FlowSOM puis par Consensus Clustering a permis de partitionner les échantillons en 13 clusters, dont 8 ont pu être manuellement associés à une population cellulaire. Ces résultats sont cohérents avec ceux issus du gating semi-automatique, malgré la difficulté à déterminer le nombre optimal de clusters. L'association automatique d'un cluster avec un type cellulaire reste l'étape finale qui n'a pas encore été atteinte. En effet, un échantillon se compose d'un continuum de cellules à des stades de différentiation différents. Il peut conduire à la formation de clusters aux phénotypes très différents mais surtout de clusters aux profils d'expression moyen et difficile à attribuer. Pour faciliter leur identification, une approche de quantification de l'enrichissement des différents marqueurs dans les différents clusters (Diggins et al., 2017) pourrait être ajoutée à ce pipeline. Ces résultats mettent en évidence l'importance du design du panel d'anticorps utilisés. En effet, certains marqueurs d'exclusion comme CD19 ou CD14 auraient pu être ajoutés pour permettre une meilleure discrimination des différents groupes de cellule et ainsi réduire la part non négligeable de cellules non associée à un type cellulaire. Bien que ce soit FlowSOM qui soit utilisé ici, d'autres algorithmes performants tels que

FlowMeans (Aghaeepour et al., 2011) ou Xshift (Samusik et al., 2016) pourraient être implémentés dans ce pipeline.

De par le design de l'expérience, il est impossible de confirmer statistiquement les tendances observées et de déterminer à quel point elles sont biaisées par les protocoles expérimentaux ou l'état immunologique du donneur. De plus, la comparaison des différentes conditions d'activations est très délicate car aucune normalisation n'est réalisable ici. Pour permettre cette comparaison, il est possible de réunir les échantillons et d'ajouter des billes de calibration. Dans ce cas, les différents échantillons sont marqués à l'aide d'un code barre composé d'une combinaison unique d'isotopes du palladium (Zunder et al., 2015), avant d'être rassemblés dans un unique tube. Cela permet de diminuer d'une part de la variabilité expérimentale, d'autre part celle induite par le cytomètre de masse au cours de l'acquisition. En effet, Finck et al. (2013) ont observé qu'au cours d'une acquisition de 2 h, le signal pouvait perdre jusqu'à 30 % d'intensité. Ce barcoding peut être associé à l'ajout de billes de calibration mono-marquées avec une quantité constante d'anticorps (Finck et al., 2013). Elles permettent d'évaluer la contamination d'un canal à l'autre induite par l'oxydation des métaux et par l'utilisation d'anticorps impurs. Même si la CyTOF ne présente pas de chevauchement comme en CMF, la contamination entre les canaux induit un biais non négligeable qui peut être compensé comme présenté par Chevrier et al. (2018). Ces corrections et normalisation n'ont pas pu être réalisées avec les données du projet pilote et peuvent expliquer la variabilité des résultats observés entre les échantillons, les panels et entre les deux méthodes d'identification de sous-population. Les étapes de déconvolution du signal et de compensation des données seront ultérieurement ajoutées au pipeline car elles sont essentielles pour affiner l'identification des sous-populations et mener des analyses différentielles des profils d'expressions des marqueurs fonctionnels entre différentes conditions. Enfin, le pipeline pourra être amélioré pour des donnés acquises sur Helios (Fluidigm), la troisième génération cytomètre. Ce dernier mesure de nouveaux paramètres permettant l'optimisation du nettoyage des données. Une méthode développée par Bagwel (2017) est basée sur l'analyse de la forme de la courbe représentant l'intensité du nombre de « push » mesurés par cellule par cytomètre de masse. La sélection des cellules vivantes en singulet sur la base d'agents intercalant et de la « durée de de l'évènement » est couramment utilisée mais est de plus en plus remise en cause. Elle éliminerait notamment les cellules en division et les cellules dont la chromatine est fortement condensée.

A terme, une interface graphique sera développée avec R Shiny afin de garantir l'accessibilité du pipeline aux chercheurs moins familiarisés avec la bio-informatique.

Bibliographie

Aghaeepour, N., Nikolic, R., Hoos, H.H., and Brinkman, R.R. (2011). Rapid cell population identification in flow cytometry data. Cytometry A *79A*, 6–13.

Bagwel, B. (2017). A New Analytic Approach for Live Singlet Identification (6th Annual Mass Cytometry Summit).

Bandura, D.R., Baranov, V.I., Ornatsky, O.I., Antonov, A., Kinach, R., Lou, X., Pavlov, S., Vorobiev, S., Dick, J.E., and Tanner, S.D. (2009). Mass Cytometry: Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-of-Flight Mass Spectrometry. Anal. Chem. *81*, 6813–6822.

Bendall, S.C., Nolan, G.P., Roederer, M., and Chattopadhyay, P.K. (2012). A deep profiler's guide to cytometry. Trends Immunol. *33*, 323–332.

Bruggner, R.V., Bodenmiller, B., Dill, D.L., Tibshirani, R.J., and Nolan, G.P. (2014). Automated identification of stratifying signatures in cellular subpopulations. Proc. Natl. Acad. Sci. *111*, E2770–E2777.

Chevrier, S., Crowell, H.L., Zanotelli, V.R.T., Engler, S., Robinson, M.D., and Bodenmiller, B. (2018). Compensation of Signal Spillover in Suspension and Imaging Mass Cytometry. Cell Syst. *6*, 612-620.e5.

Diggins, K.E., Greenplate, A.R., Leelatian, N., Wogsland, C.E., and Irish, J.M. (2017). Characterizing cell subsets using marker enrichment modeling. Nat. Methods *14*, 275–278.

Finak, G., Frelinger, J., Jiang, W., Newell, E.W., Ramey, J., Davis, M.M., Kalams, S.A., De Rosa, S.C., and Gottardo, R. (2014). OpenCyto: An Open Source Infrastructure for Scalable, Robust, Reproducible, and Automated, End-to-End Flow Cytometry Data Analysis. PLoS Comput. Biol. *10*, e1003806.

Finck, R., Simonds, E.F., Jager, A., Krishnaswamy, S., Sachs, K., Fantl, W., Pe'er, D., Nolan, G.P., and Bendall, S.C. (2013). Normalization of mass cytometry data with bead standards. Cytometry A *83A*, 483–494.

Han, G., Spitzer, M.H., Bendall, S.C., Fantl, W.J., and Nolan, G.P. (2018). Metal-isotope-tagged monoclonal antibodies for high-dimensional mass cytometry. Nat. Protoc. *13*, 2121–2148.

Kohonen, T. (1990). The self-organizing map. Proc. IEEE *78*, 1464–1480.

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

Monti, S. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Mach. Learn. *52*, 91–118.

Nowicka, M., Krieg, C., Weber, L.M., Hartmann, F.J., Guglietta, S., Becher, B., Levesque, M.P., and Robinson, M.D. (2017). CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. F1000Research *6*, 748.

Qiu, P., Simonds, E.F., Bendall, S.C., Gibbs, K.D., Bruggner, R.V., Linderman, M.D., Sachs, K., Nolan, G.P., and Plevritis, S.K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat. Biotechnol. *29*, 886–891.

Samusik, N., Good, Z., Spitzer, M.H., Davis, K.L., and Nolan, G.P. (2016). Automated mapping of phenotype space with single-cell data. Nat. Methods *13*, 493–496.

Van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. 2579–2605.

Van Gassen, S., Callebaut, B., Van Helden, M.J., Lambrecht, B.N., Demeester, P., Dhaene, T., and Saeys, Y. (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data: FlowSOM. Cytometry A *87*, 636–645.

Weber, L.M., and Robinson, M.D. (2016). Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data: Comparison of High-Dim. Cytometry Clustering Methods. Cytometry A *89*, 1084–1096.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573.

Zunder, E.R., Finck, R., Behbehani, G.K., Amir, E.D., Krishnaswamy, S., Gonzalez, V.D., Lorang, C.G., Bjornson, Z., Spitzer, M.H., Bodenmiller, B., et al. (2015). Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. Nat. Protoc. *10*, 316–333.

AGRO CAMPUS OUEST	Diplôme : Ingénieur Agronome Spécialité : Cursus ingénieur Agronome Spécialisation / option : Master BMC – Biologie Moléculaire et Cellulaire Enseignant référent : Frédéric Lecerf

Auteur(s) : Juliette MAES Date de naissance* : 11 juin 1996	Organisme d'accueil : Sanofi Aventis R & D Unité de Sciences Translationnelles Adresse : 1 avenue Pierre Brossolette 91385 Chilly-Mazarin cedex France Maître de stage : Charles Bettembourg
Nb pages : 27 Annexe(s) : 0	
Année de soutenance : 2019	

Titre français : Développement d'un pipeline automatisé d'analyse de données de cytométrie de masse pour l'identification de populations immunitaires

Titre anglais : Development of an automated mass cytometry data analysis pipeline for immune population identification

Résumé (1600 caractères maximum) :

La cytométrie en flux a été pendant longtemps la méthode de choix pour étudier le phénotype des cellules immunitaires. Aujourd'hui, la nouvelle technologie de cytométrie de masse offre une alternative puissante car elle permet de mesurer l'expression de trois fois plus de marqueurs simultanément à plus haute résolution. En augmentant le nombre de marqueurs mesurés par cellules, l'analyse manuelle des données est devenue de plus en plus fastidieuse en plus d'être dépendante de l'analyste. Dans cet article est présenté un pipeline d'analyse automatisé, rapide et objectif, développé pour les données de grandes dimensions de cytométrie de masse. Il se divise en plusieurs parties : le nettoyage des données brutes, le contrôle qualité, l'identification de sous-populations immunitaires présentes et leur visualisation. Deux approches différentes sont proposées pour l'identification des types-cellulaires et peuvent être utilisées seules ou en parallèle. La première est une méthode supervisée de gating semi-automatique basée sur l'environnement OpenCyto. La seconde utilise les algorithmes de clustering non supervisés FlowSOM et Consensus Clustering pour partitionner des cellules en différentes populations. Celles-ci sont ensuite visualisées grâce à l'algorithme de réduction de dimension UMAP. Ce pipeline implémenté sur R sera bientôt complété d'une interface graphique R Shiny pour faciliter son utilisation.

Abstract (1600 caractères maximum) :

Until recently, flow cytometry has been the most widely used method to study immune cells phenotype. Mass cytometry is a new promising technology which enables the measurement of over forty markers per cell simultaneously with a higher resolution. However, the increase in the number of features entails the need of new analysis methods. Indeed, manual analysis is very fastidious, time-consuming and analyst-dependant. In this article, we present a pipeline for automated, fast and objective mass cytometry data analysis. It is divided into four parts: raw data pre-processing, quality control, identification of immune subpopulations, and their visualizations. Two different approaches for cell-type identification are available. They can be used alone or in parallel to compare the results. The first one is a supervised method based on the *OpenCyto* framework. It consists in the automation of manual gating usually conducted for flow cytometry data. The second one uses FlowSOM and Consensus Clustering, two unsupervised embedding algorithms, in order to cluster cell into different populations. Afterward, identified populations can be displayed through the dimension reduction algorithm UMAP. In the near future, a graphical user interface will be developed with R Shiny in order to facilitate the access and the implementation of the pipeline.

Mots-clés : clustering non supervisé, gating automatique, données à hautes dimensions, réduction de dimension, visualisation, couplage anticorps isotope

Key Words: unsupervised clustering, automated gating, high dimensional data, dimension reduction, visualization, isotope-tagged antibodies

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*