CIPHOLD : Refonte de l'outils d'annotation automatique de clusters non-supervisé à partir de populations de référence SCAFFOLD

Quentin BARBIER[1,2,3,4], Calvin MATTEOLI [3], Pier Federico GHERARDINI[5], Samuel GRANJEAUD[6], Hervé LUCHE[1,2,3,4]

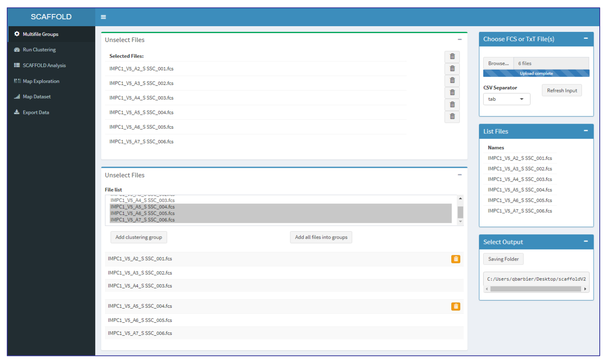
1. Centre d'Immunophenomique (CIPHE) - Phenomin
2. Institut National de la Santé et de la Recherche Médicale (INSERM) US012, Marseille
3. Aix-Marseille Université, UMS3367, Marseille, France
4. Centre National de la Recherche Scientifique (CNRS), UMS3367, Marseille
5. Parker Institute, San Francisco, USA
6. Centre de Recherche en Cancérologie de Marseille (CRCM), Marseille

Introduction

Avec l'émergence de nouvelles technologies de cytométrie permettant l'analyse de plus d'une vingtaine de marqueurs à l'échelle de la cellule unique, les jeux de données sont devenus complexes à analyser avec des outils conventionnels. L'analyse d'un grand nombre d'échantillons par expérience contenant plusieurs milliers d'évènements par échantillons peut-être fait par regroupement de cellules de phénotypes similaires sous forme de clusters. L'annotation des clusters générés par l'utilisateur est chronophage et subjective. Envisagée sur ce type de données, une approche non-supervisée est extrêmement couteuse en temps. SCAFFOLD (*P.F. GHERARDINI , 2015*) est une alternative pouvant répondre à ce type de problème. Il permet une annotation automatique de clusters non-supervisés avec des données issues d'une même analyse manuelle. En utilisant un pré-fenêtrage de populations d'intérêt, SCAFFOLD va créer une carte de référence en associant des clusters à ces populations d'intérêt. Les données provenant d'organes divers ou acquises à des dates éloignées dans le temps présentant potentiellement un effet batch, voir même acquises sur des machines différentes (Flux, Masse, Spectral, scRNAseq), peuvent être ainsi visualisées et comparées sur une même carte de référence. Les clusters non-supervisés sont obtenus à l'aide de CLARA. L'annotation se fait à partir d'une métrique de similarité cosinus. Cette similarité permet de ne pas prendre en compte les variations communes d'intensité dans l'espace multidimensionnel. Le graphe obtenu est ensuite ordonné selon l'algorithme de dessin des graphes forcés (*forced-directed graph*) pour optimiser leur distribution dans un espace à 2D. Afin d'améliorer l'ergonomie et la gestion des fichiers, nous avons travaillé sur CIPHOLD, une version avec du travail de Gherardini avec une refonte total de l'interface et avons ouvert les possibilité de format d'entrée https://github.com/cipheLab/CIPHOLD

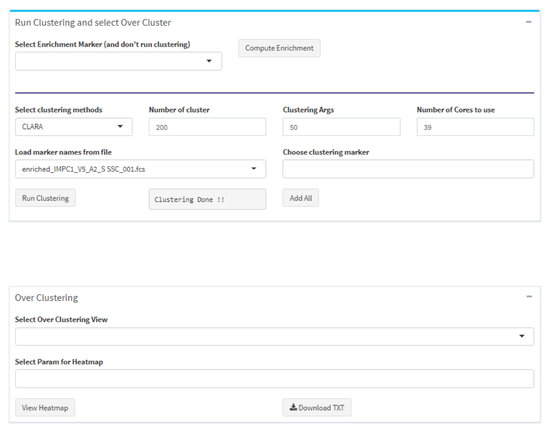
1° Preparation des données

- L'outil CIPHOLD peut prendre en entrée des fichiers FCS bruts, des fichiers FCS possédant déjà le résultat d'un *clustering* (fichier enrichi) ou des fichiers CSV d'une liste de cluster.
- Les fichiers FCS peuvent-être rassemblés par groupes expérimentaux pour permettre la visualisation de foldchange sur l'intensité des marqueurs pour chaque cluster
- Les fichiers peuvent subir une transformation différente de celle utilisée pour les fichier de références. Ceci permet d'utiliser dans une même analyse des données de technologie différentes
- La grosse différence avec l'ancienne version proposée par P.F Gherardini est d'intégrer des résultats produits à chaque étape dans une seule interface sans avoir besoin de gérer les dossiers et fichiers en dehors de l'application. Cette flexibilité accrue de l'outil permet de l'utiliser dans de multiples applications



2° Clustering

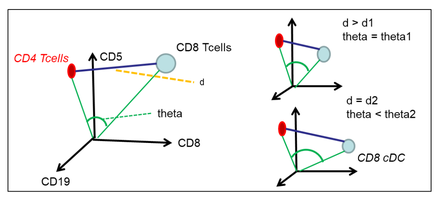
- Le *clustering* se fait sur plus de 100 clusters avec CLARA (overclustering). Il est possible de modifier les paramètres de clustering.
- Le but de ce partitionnement automatique des données est de regrouper les cellules ayant des profils d'expression de marqueurs similaires dans l'espace multidimensionnel.
- Il est possible de choisir un paramètre de cluster réalisé précédemment ou avec une autre méthode.
- Des fichiers .FCS sont ré-écrits à la fin de cette étape avec un nouveau paramètre « cluster » précisant le numéro de cluster qui a été attribué à une cellule donnée

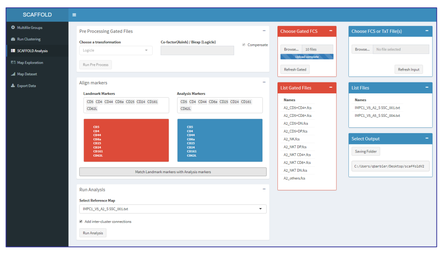


3° Mapping

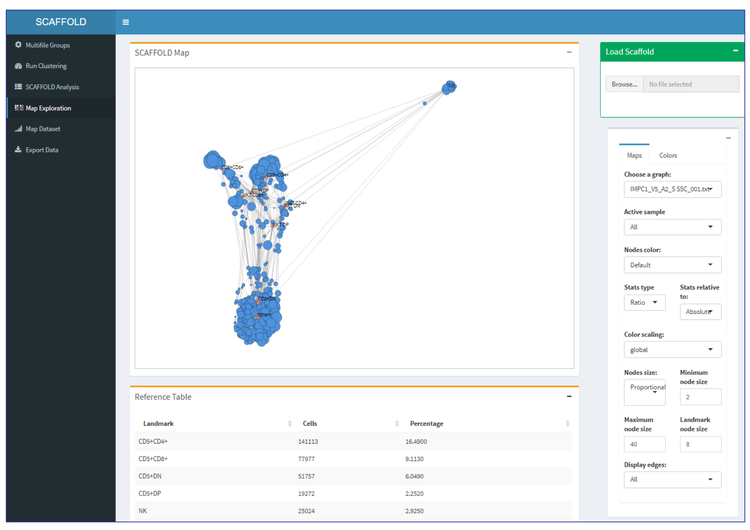
Une carte de référence est obtenue sur laquelle la position des « landmarks » ou balises est fixée à partir d'un fichier de référence choisi parmi ceux analysés grâce au *forced-directed-graph*.

La similarité cosinus diminue l'impact des différences d'intensité d'expression dans chaque dimension à l'étape de réalisation de la map. Cette métrique accorde plus d'importance aux variations de patrons d'expression multi dimensionnels. Les annotations des clusters peuvent être réalisées de manière robuste sur des jeux de données différents puisque les différences d'intensité d'expression observées pour un même marqueur sur des technologies différentes ne sont pas prises en compte pour établir les associations de clusters à aux populations/balises.de référence.





4° Visualisation et statistique



Les cercle bleus représentent les clusters et les rouges les populations de référence. La taille des cercles bleus est proportionnelle au nombre de cellules.

Les fonctions de base sur la visualisation de la map ont été conservées :

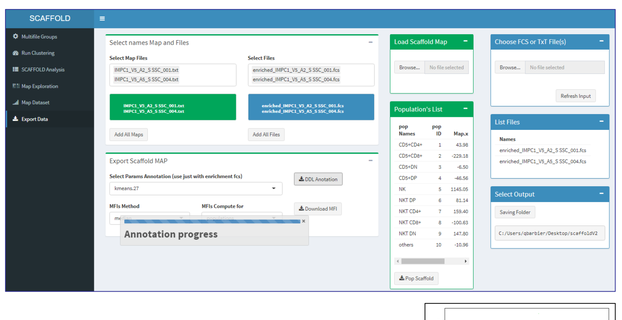
- Visualiser les différents échantillons
- Représenter l'intensité d'expression de chaque marqueur sur les clusters
- Afficher uniquement les meilleurs scores de similarité.

- Le nombre de cellules dans un cluster donné et la fréquence du nombre d'évènement total sont indiqués dans un tableau de synthèse sous le graphique. Il est téléchargeable pour en extraire les fréquences relatives de chaque population pour toute l'analyse.
- A cette étape il est possible de visualiser une carte réalisée précédemment et refaire un « mapping » ou un export des données.

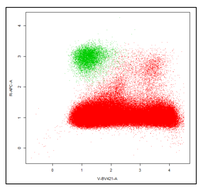
5° Export des données

En plus des fréquences, l'outil revisité permet de revenir à l'échelle de la cellule unique en écrivant un fichier FCS enrichi:

1. Il écrit l'annotation associée aux cellules d'une population donnée dans un nouveau fichier FCS
2. La MFI des populations peut-être calculée pour tous les marqueurs choisis d'après les cellules contenues dans les clusters regroupés autour d'une balise.



L'écriture dans un fichier FCS de l'appartenance d'une cellule à une annotation déterminée par SCAFFOLD permet de valider la qualité de l'annotation dans un autre logiciel d'analyse plus conventionnel (Flowjo, Kaluza, DIVA) ou utiliser d'autres outils bio-informatiques cherchant à établir des stratégies de gating minimales (hyperGate, GateFinder, MEM, C2G)



Perspectives

La première version de SCAFFOLD conçue par P.F. Gherardini requérait de disposer de tous les fichiers pour une analyse dans un dossier particulier. Les populations de référence devaient se trouver dans un sous-dossier et se conformer à une nomenclature particulière pour réaliser l'analyse. Lorsque l'on voulait utiliser la fonctionnalité de « Map Dataset » pour mapper d'autres jeux de donnée, l'outil devait être relancé. Il était impossible d'utiliser directement des listes de cluster provenant d'autres pipelines d'analyse (format .csv ou .fcs enrichis). Ces contraintes limitaient grandement l'utilisation de ce pipeline à différents types d'analyse réalisées par les biologistes. Avec cette version de SCAFFOLD améliorée par le CIPHE, il est maintenant possible d'intervenir sur chacune des étapes de la version originale de l'outil, d'intégrer facilement de nouvelles méthodes de clustering (FlowSOM ou k-means, métrique de distance euclidienne, etc…) et ne nécessite aucune disposition prédéfinie de dossiers ou répertoires. La possibilité de paralléliser les étape les plus longues du pipeline permet de réduire considérablement le temps de calcul utilisé pour l'étape d'over-clustering de CLARA. Il n'est plus nécessaire de relancer l'outil pour modifier ou changer entièrement une analyse. Cette version se différencie de celle publiée récemment par P.F. Gherardini où SCAFFOLD est maintenant séparé en trois outils distincts et indépendant les uns des autres (https://github.com/ParkerICI)

Centre d'Immunophénomique

INSERM US012
Parc Scientifique et Technologique de Luminy
163, Avenue de Luminy - Case 936
F-13288 Marseille Cedex 09

T: 0033 491.828.950
F: 0033 491.253.048