

UMR Inserm 1052 - CNRS 5286

Centre Léon Bérard

Directeur : Dr Patrick Mehlen

Dir adjoints : Dr Véronique Maguer-Satta, Pr Charles Dumontet

Équipes : (co-encadrement) Lyon Immunotherapy for Cancer Laboratory (LIICL) et Pole data

Structure : Centre Léon Bérard, Centre Léon Bérard -28 Rue Laennec 69373 LYON CEDEX 08

Co-encadrants et contacts : Youenn DROUET et Uzma HASAN

Youenn.drouet@lyon.unicancer.fr; Uzma.hasan@lyon.unicancer.fr

Profil recherché : Étudiant de master 2 avec de solides connaissances en bio-informatique et biostatistiques, notamment avec les langages R/Python.

Période : 6 mois, dès que possible

Titre du projet : Etude du profil génétique et immunologique des tumeurs malignes solides et hématologiques de tous types histologiques à un stade avancé de la cohorte ProfilER

Contexte général du projet : Le diagnostic et le traitement des cancers reposent principalement sur l'analyse histopathologique. Les avancées récentes en génomique et les outils bio-informatiques, permettent d'établir le profil génétique de chaque tumeur ce qui ouvre la voie à des traitements ciblant les altérations génétiques responsables du développement tumoral.

Par ailleurs, le système immunitaire du patient a un impact majeur sur le développement et l'évolution du cancer. Par exemple, des biomarqueurs tels que la lymphopénie la diversité des TCR en périphérie ou la proportion accrue de lymphocytes T régulateurs et de macrophages dans l'environnement tumoral, ainsi que la surexpression de PD-L1, influencent la réponse aux traitements anti-tumoraux. Le programme ProfilER vise établir le profil génique à partir de prélèvements tumoraux archivés et immunologique sanguin de patients atteints de tumeurs avancées, solides ou hématologiques permettant d'orienter le traitement des patients vers des thérapies ciblées ou des immunothérapies recommandées.

Sets de données disponibles : La cohorte ImmunoProfiler intègre 492 patients de l'étude Profiler, qui ont eu un immunophénotypage des cellules du sang avec 8 panels de 8 marqueurs constituant 300 colonnes de variables phénotypiques. L'ensemble de ces 492 patients ont des données cliniques et génétiques associées. On note une hétérogénéité importante des localisations tumorales évaluées. Les types tumoraux les plus représentés sont les cancers digestifs (n=104), ORL (n=68), cancers du système nerveux central (n=44), sarcomes (n=42) et cancer du pancréas (n=41).

Résultats préliminaires (Figure 1) : Des analyses préliminaires ont été réalisées avec les données quantitatives des sous-types de cellules immunitaires. On retrouve des clusters de fortes corrélations entre les populations de polynucléaires, les populations de lymphocytes TCD8, TCD4, de monocytes, ou de cellules dendritiques (DC). Les patients du cluster 3 sont enrichis en populations TCD4 et de manière un peu moins homogène, TCD8. Les patients du cluster 2 ont globalement peu de cellules immunitaires circulantes, et les patients du cluster 1 sont caractérisés par une hétérogénéité de comptages. Ces clusters sont répartis de manière relativement homogène dans les différentes localisations tumorales : Ce résultat suggère que les clusters ainsi définis pourraient avoir un sens « pancancer ». On observe cependant une légère sous-représentation des cancers ORL dans le cluster 3 (3% vs. 14% sur l'ensemble des patients) et une sur-représentation des cancers digestifs dans le cluster 1 (29% vs 21.5% sur l'ensemble des patients).

Objectif du projet: L'objectif de ce projet est de définir les liens existants entre les quantités des populations immunes et/ou la fréquence d'expression de marqueurs et 1) les altérations moléculaires (eg : profil mutationnel), et 2) les types tumoraux (histotypes, stade, données cliniques).

Méthodologie envisagée : Il s'agit d'un problème de « fouille de données en grande dimension ». On pourra considérer une approche agnostique, consistant à comparer les résultats obtenus par plusieurs algorithmes de machine learning (notamment Svm, Lasso, Ridge, Elasticnet, Random Forest etc.). La taille limitée de l'échantillon (n=492) ne permet pas une séparation des données en sous-échantillons « entraînement » et « validation ». La grande dimension (nb de variables > 300) oblige donc à utiliser des approches de correction pour les tests multiples du type « Benjamini–Hochberg » dans le cas des analyses univariées, conjuguées à des techniques de cross-validation ou bootstrap pour corriger l'optimisme des modèles en analyses multivariées.

Analyses attendues :

A- Description des données dans cette cohorte de 492 patients :

- Consolider les analyses avec les données cliniques issues de la base de données ConSoRe.
- Obtenir des waterfall plot mutationnels sur l'ensemble de la cohorte après génération de la matrice de mutations,
- identifier si des profils mutationnels sont spécifiques de certaines pathologies tumorales.

B- Analyser les données de fréquence de marqueurs sur les différentes populations immunitaires en utilisant la même stratégie que l'analyse des données quantitatives en groupant tous les échantillons (analyse « pan cancer »)

- Travailler sur la normalisation des données afin de définir s'il est plus important de conserver l'échelle d'origine (sens biologique) ou de normaliser les différentes variables (perte du sens biologique au profit du signal statistique) ?
- Compléter l'analyse « pan-cancer » par une analyse détaillée par localisation tumorale avec deux axes :
 - a) mieux caractériser les profils immunologiques circulant des sous-types de cancer digestifs, enrichis dans les clusters 1 et 3, et au vu du possible signal sur les monocytes dans ces clusters,
 - b) préciser et décrire la déplétion immunitaire dans le sang observée dans les cancers ORL (cluster 2).

C- Explorer les liens entre anomalies moléculaires et les données de quantité et de fréquence des différentes populations immunitaires circulantes.

- Intégrer les trois jeux de données « quantitatives », « fréquence » et « mutations » avec dans un premier temps la sélection de deux entités tumorales les plus représentées (e.g. digestif, ORL, l'étudiant en master sera accompagné par un étudiant en médecine spécialiste en ORL).
- Explorer les corrélations existantes et la possibilité de classer les patients (clustering non supervisé) en fonction de ces paramètres.

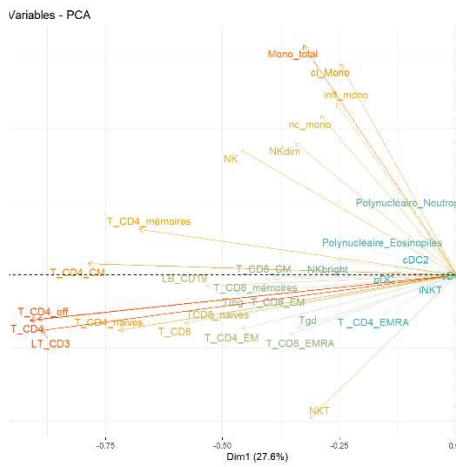
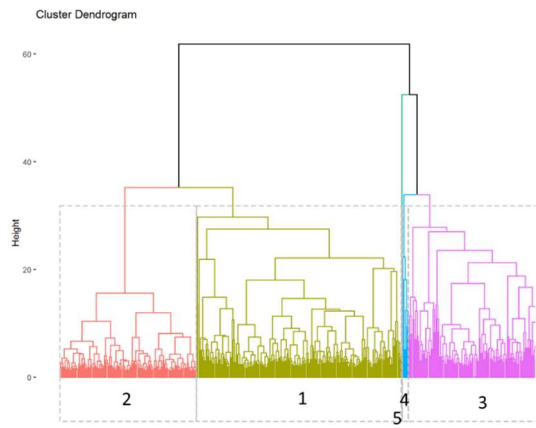
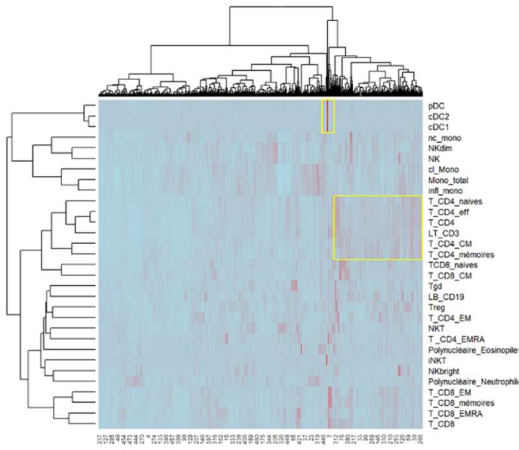
Environnement

Uzma HASAN coordonne le LICL, qui fournit une expertise en immunologie des tumeurs et des outils avancés ainsi que des compétences pour analyser les paramètres immunitaires dans le sang et les tumeurs des patients atteints de cancer dans le cadre de projets de recherche fondamentale, translationnelle et clinique. Le LICL est composé de 5 scientifiques, 10 ingénieurs et 2 chercheurs post-doctoraux. Le Pole Data est un jeune groupe composé de séniors Data Scientist, Data Manager, Bioinformaticiens et Médecins au CLB-CRCL.

Output :

Publication d'un article scientifique, Presentations

Figure 1. Analyses préliminaires réalisées sur la cohorte ImmunoProfiler



Organ	1, N = 204	2, N = 136	3, N = 135	4, N = 6	5, N = 1
Autres	26 (13%)	19 (14%)	19 (15%)	1 (20%)	0 (0%)
Digestif	59 (29%)	14 (10%)	29 (22%)	1 (20%)	0 (0%)
Systeme_nerveux_central	12 (5.9%)	19 (14%)	13 (10%)	0 (0%)	0 (0%)
ORL	26 (13%)	33 (25%)	4 (3.1%)	1 (20%)	0 (0%)
Genital_femme	18 (8.9%)	11 (8.2%)	13 (10%)	0 (0%)	0 (0%)
Pancréas	18 (8.9%)	8 (6.0%)	15 (12%)	0 (0%)	0 (0%)
Poumon	6 (3.0%)	3 (2.2%)	6 (4.6%)	0 (0%)	1 (100%)
Genital_homme	5 (2.5%)	2 (1.5%)	1 (0.8%)	2 (40%)	0 (0%)
Voies_urinaires	8 (3.9%)	7 (5.2%)	6 (4.6%)	0 (0%)	0 (0%)
Sarcome	14 (6.9%)	11 (8.2%)	15 (12%)	0 (0%)	0 (0%)
Sein	11 (5.4%)	7 (5.2%)	9 (6.9%)	0 (0%)	0 (0%)
Unknown	1	2	5	1	0