Projet de Biostatistiques

Le projet proposé porte sur l'analyse d'un jeu de données collectées aux États-Unis dans le cadre du programme MSK-MET (Memorial Sloan Kettering - Metastatic Events and Tropism). **Ce projet sera réalisé en groupe de quatre personnes**. Dans une première partie, il est demandé de faire une analyse descriptive des données fournies. À partir de cette analyse, vous devrez définir une question d'intérêt sur cette problématique pouvant être résolue par une approche biostatistique et tenter d'y répondre. L'attendu est un **compte-rendu écrit court (2 pages maximum) et une présentation orale de 8 minutes par groupe** de travail, dans lesquels vous détaillerez la question d'intérêt que vous aurez choisie, votre démarche et vos résultats.

Contexte

Cancers et métastases

Le cancer est une maladie caractérisée par la prolifération incontrôlée de cellules, liée à un échappement aux mécanismes de régulation qui assure le développement harmonieux de notre organisme [1].En se multipliant de façon anarchique, les cellules cancéreuses donnent naissance à des **tumeurs** de plus en plus grosses qui se développent en envahissant puis détruisant les zones qui les entourent. La cellule cancéreuse qui se détache de la tumeur, peut migrer dans une autre partie du corps voisine voire un autre organe pour créer une nouvelle tumeur, appelés alors **métastase**.
La plupart des morts liées aux cancers sont dues au **developpement de métastase**. Cependant, très peu de connaissance sont disponibles sur les déterminants génétique des métastases. Particulièrement, la distribution des sites de métastase pour un type de cancer donné est généralement non aléatoire et dépend de multiples facteurs tels que la localisation anotomique du cancer primaire, la cellule d'origine ou le sous-type moléculaire[2].

Organotropisme métastatique

Ce phénomème, appelé organotropisme métastatique, caractérise l'affinité des métastases d'un cancer donné pour un tissu. **L'hypothèse classique de la semence et du sol**, selon laquelle les cellules cancéreuses disséminées colonisent de préférence les organes qui permettent leur propre croissance et sont compatibles avec celle-ci, est étudiée depuis plus d'un siècle. Pourtant, il reste beaucoup à apprendre sur **l'interaction entre les caractéristiques génomiques des tumeurs et le potentiel métastatique**, ainsi que sur les schémas de métastases spécifiques aux organes.

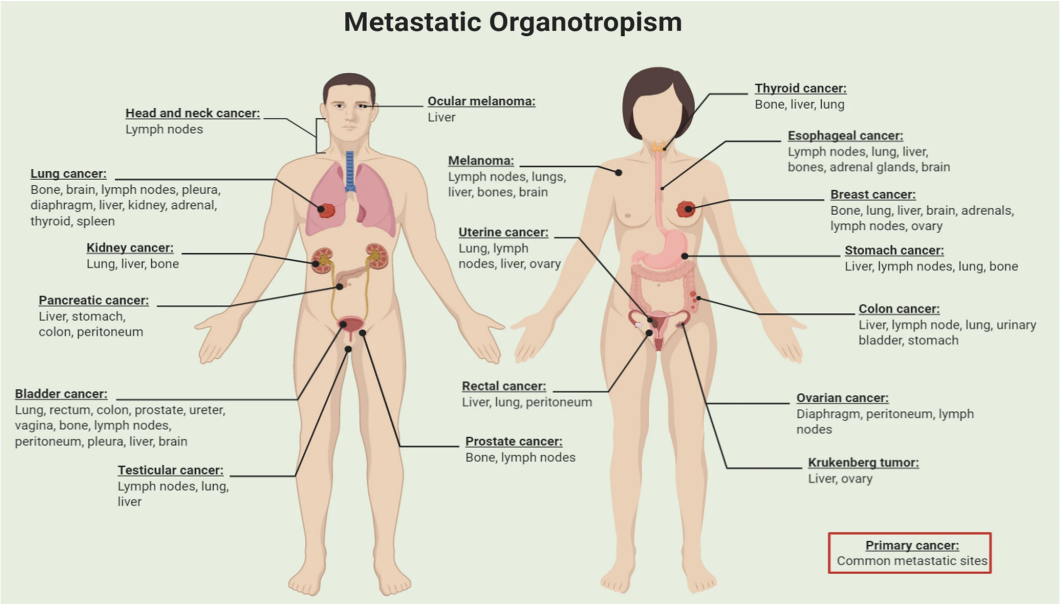

Figure 1: Current landscape of Metastastatic Organotropism for multiple cancer types. *From Fares, J., Fares, M.Y., Khachfe, H.H. et al., Sig Transduct Target Ther 5, 28 (2020)*

Le profilage moléculaire des tumeurs couplé à l'annotation clinique des événements métastatiques pourrait contribuer à éclairer cette question. Les avancées dans les méthodes de séquençage permettent désormais d'étudier la présence de mutation dans un grand ensemble de gènes à partir d'échantillon de tumeur[3]. Cependant, les efforts de séquençage du cancer à grande échelle se sont jusqu'à présent concentrés sur les tumeurs primaires non traitées (par exemple, The Cancer Genome Atlas [4]), ont caractérisé le paysage génomique global de la maladie métastatique sans interroger explicitement l'organotropisme métastatique spécifique ou se concentrent sur un unique type de cancer.

Les données disponibles

Le jeu de données MSK-MET contient les informations de 25,000 patients atteints de multiples cancers. Ce type de cohorte est appelée une **cohorte pan-cancer**. Avec les informations cliniques (sexe, age, données diagnostics), ce jeu de donnée contient des informations sur les caractéristiques génomiques des cancers primaires et des métastases telles la présence de mutations[5], l'instabilité microsatellite[6] ou la fraction de génome alteré. Au total, 590 variables sont disponibles dans ce jeu de données dont certaines sont des identifiants comme 'SAMPLE_ID' ou 'PATIENT_ID.

Les variables sont rapportées dans le fichier (MSKMET_pan_cancer.csv). Vous trouverez la description de chaque variable dans le fichier MSKMET_pancancer_description.csv. **Vous ferez particulièrement attention dans votre analyse à selectionner les colonnes pertinentes pour la question biologique que vous souhaitez résoudre.**

Étude des données

Lecture des fichiers

Le fichier MSKMET_pan_cancer.csv est au format CSV, comportant en première ligne le nom des colonnes. Le séparateur de colonne est la virgule. Les valeurs manquantes sont codées par une absence de caractères. Après lecture, vous devez vérifier que les données lues sont cohérentes avec celles présentes dans le fichier initial (nombre de lignes et colonnes lues, identité des valeurs dans les premières et dernières lignes, type ou classe des variables lues, ...).

Le fichier MSKMET_pancancer_description.csv, également au format CSV, comprend en première colonne le nom des variables (i.e. les colonnes du fichier SEER_breast_cancer.csv) et en deuxième colonne leur description. Certaines variables ont un suffixe commun (exemple: CNA_XXX). Une seule description pour toutes les variables avec le même suffixe est proposée.

Les questions suivantes ont pour objectif de vous guider dans l'analyse des données et de vous donner des exemples de questions à résoudre, mais l'évaluation portera sur **une question en particulier que vous devrez définir et à laquelle vous devrez répondre à l'aide des données.**

Description des variables

1. De quel type sont les variables du jeu de données (qualitatives, quantitatives, discrètes, continues...) ?
2. Lesquelles contiennent des valeurs manquantes ? En quelles proportions ?
3. Calculez les informations statistiques (moyenne, médiane, écart-type) que l'on peut obtenir sur ces variables.
4. Représentez graphiquement la distribution de ces variables (e.g. histogrammes, boxplots)
5. Combien y'a-t-il d'echantillons de métastases ? Combien par sous-type de cancer ? Faire les représentations appropriées.
6. Combien d'echantillons de tumeurs primaires ont été etudié chez des patients présentant déja un cancer métastatiques ? Combien par type et sous-type de cancer ?

La description peut s'étendre bien au dela en regardant les relations des différentes variables les unes avec les autres et les facteurs confondants comme l'age et le sexe.

Analyse

1. Les tumeurs primaires présentent-elles significativement plus d'altération génomiques que les métastases ? Est-ce significatif pour tous les types de cancers ?
2. Y-a-t-il des évènements génétiques plus fréquents dans les tumeurs primaires que les métastases? Faire le test avec comme example le gène AR.
3. Evaluez le lien entre la charge métastatique et la survie globale.
4. Les tumeurs primaires de patients métastatiques possèdent-elles d'avantage d'altérations génétique que les tumeurs primaires de patients non metastatique ? Y-a-t-il d'autres différences entre ces deux populations?
5. Certains cancers colorectaux ont la particularité de porter une mutation dans le système de réparation de l'ADN (MMR), donnant alors un phénotype d'hypermutabilité de l'ADN appelé instabilité microsatellite (MSI) et qui est souvent de meilleur prognostic. Dans ce cancer, évaluer les différences entre les patients MSI stable et MSI instables. Vous pourez aussi utiliser le score MSI.
6. Evaluer, pour les métastases, le lien entre la fraction de génome altéré des métastases du cerveau en fonction du type de la tumeur primaire.
7. Y-a-t-il un organotropisme métastatique dans les cancers du sein ?
8. Quels facteurs de confusion peuvent influencer les liens étudiés ? Quel niveau d'analyse vous semble le plus pertinent entre le niveau pan-cancer et le niveau type ou sous-type de cancer ?

Evaluation

L'évaluation de votre travail se fera sur une seule analyse, vous devrez donc :

- choisir une question d'intérêt à laquelle il est possible de répondre via le jeu de données disponibles (vous pouvez vous inspirer de celles posées dans ce document, ou en proposer une nouvelle),
- poser les hypothèses,
- réaliser les tests statistiques nécessaires,
- produire les graphique pertinents illustant votre question d'intérêt,
- analyser les résultats et proposer une réponse,
- effectuer une synthèse.

Notes

1. Definition de la Ligue Contre le Cancer, https://www.ligue-cancer.net/article/26088-quest-ce-que-le-cancer
2. Gao Y, Bado I, Wang H, Zhang W, Rosen JM, Zhang XH. Metastasis Organotropism: Redefining the Congenial Soil. Dev Cell. 2019 May 6;49(3):375-391. doi: 10.1016/j.devcel.2019.04.012.
3. Définition du profilage moléculaire: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/molecular-profiling
4. Sanchez-Vega F, Mina M, Armenia J, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018;173(2):321-337.e10. doi:10.1016/j.cell.2018.03.035
5. Définition de mutation: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/mutation
6. Définition de l'instabilité microsatellite: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/microsatellite-instability