Partie Analyse de Survie

Le rôle de la méthylation des protéines dans le cancer du sein peut être étudier en réaliser des analyses de survie. L’objectif exploratoire est le même qu’avec l’analyse d’expression des gènes : identifier des agents de méthylations qui peuvent jouer un rôle dans le développement et/ou l’agressivité du cancer du sein.

De la même manière que pour la première partie sur l’analyse d’expression des gènes, nous avons dans un premier temps travail sur un ensemble de plus de 20 000 gènes présentent dans 8 jeux de données (TCGA, Naderi-Caldas…). Puis, nous avons finalement affiner notre analyse une base d’environ 600 gènes qui sont les plus susceptibles d’être des agents de méthylations impliqués dans le développement du cancer du sein.

Dans cette partie, objectif est de construire des courbes de survie :

* Pour chaque gène
* Dans chaque jeu de données
* Dans chaque sous-type (sauf « Non-tumours »)
* Pour la survie global (Overall Survival) et la survie sans rechute (Disease Free Survival)

Les critères de sélections des gènes les plus influent sur la survie sont basés sur 2 modèles statistiques, qui sont les plus fréquemment utilisés en analyse de survie :

**1 – Le test du Logrank**

Le test du Logrank évalue si deux populations de patients ont une probabilité de survie similaire (hypothèse H0) ou différente (hypothèse H1). Le modèle de Cox ressemble à une régression linéaire, adaptée à des données de survie. Il permet de savoir si une variable explicative est corrélée à la probabilité de survie des patients.

**2 – Le modèle de Cox**

Le modèle de Cox estime l’impact de la valeur numérique du niveau d’expression du gène sur la probabilité de survie. Il permet de répondre à la question suivante : est-ce que l’augmentation du niveau d’expression est associée à l’augmentation du risque de décès ? Pour cela, on calcule la p-valeur associée et le hasard ratio du modèle.

**Les critères de significativité**

Pour identifier les gènes qui impactent le plus la survie des patients, nous utiliserons les critères de significativité suivant :

* Soit les deux p-valeurs (ajustées par la méthode de Bonferonni) de Cox et Logrank < 0.10 à la fois
* Soit l’une ou l’autre des p-valeurs ajustées < 0.10

Nous pouvons également considérer le Hazard Ratio, cette quantité est largement rependue pour analyse des courbes de survie et pour choisir les gènes qui impactent le plus la survie. Au niveau de l’interprétation, un Hazard Ratio supérieur à 1 signifie que le couple (gène, sous-type moléculaire) donné impacte significativement la survie. A l’inverse, un Hazard Ratio inférieur à 1 indique que le couple n’impacte pas la survie de manière significative. Dans ce rapport, nous considérons le logarithme du Hazard Ratio, ainsi le seuil de significativité est 0.

Avant de réaliser l’analyse de survie, nous cherchons à répondre à quelques questions que nous pouvons nous poser pour identifier des gènes qui peut être impliqué dans le développement et l’agressivité du cancer du sein.

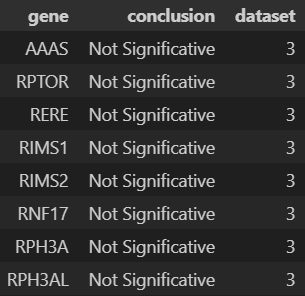
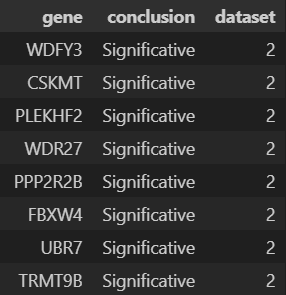
**Quels sont les gènes dont le niveau d’expression est associé à la survie de façon robuste, c’est-à-dire, avec validation dans plusieurs jeux de données ?**

La réponse a cette première question peut permettre d’identifier une base réduite de gènes dont les conclusions peuvent être prises au sérieux car des conclusions similaires se répètent dans plusieurs jeux de données.

Pour répondre à cette question, nous pouvons extraire un tableau qui donne la répétition d’une même conclusion pour un gène dans plusieurs jeux de données. Nous considérons 2 conclusions possibles :

* « All significatives » : Lorsque la pvaleur ajustée du Logrank et du modèle de Cox sont tout 2 inférieurs à 10%.
* « Not significatives » : Lorsque qu’au moins une des 2 pvaleurs ajustées est supérieure à 10%.

Nous allons distinguer les résultats pour les données de survie sans rechute (DFS) et les données de survie globale. Voici un extrait des tableaux obtenus :



A gauche, nous avons dans l’ordre décroissant, les gènes dont la conclusion « Significative » se répète dans le plus grand nombre de jeux de données. A droite, nous avons à l’inverse, les gènes dont la conclusion « Not significative » se répète dans le plus grand nombre de jeux de données.

Les résultats pour les données de survie sans rechute (DFS) sont nettement plus intéressants que les données en survie globale (OS) car dans les données à disposition, nous avons beaucoup plus de DFS que d’OS. Cela s’explique par le fait que le cancer du sein est un cancer beaucoup moins mortel que la moyenne.

**Matrice des gènes candidats dans le plus grand nombre de jeux de données**

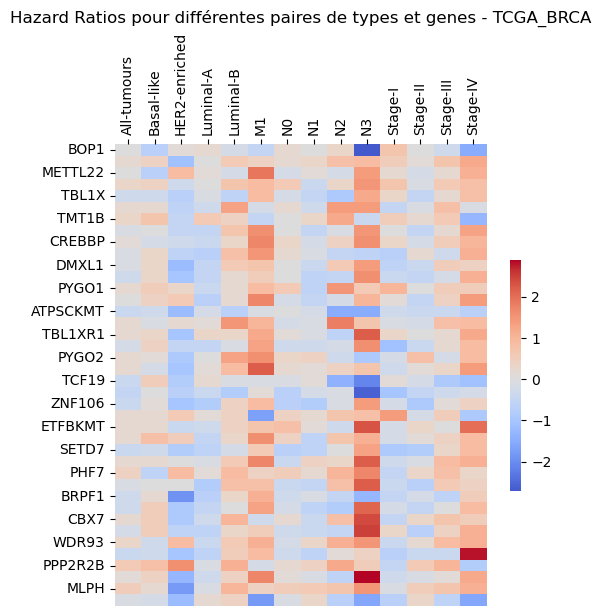
Pour visualiser les gènes que nous retrouvons dans le plus grand nombre de jeux de données, nous pouvons réaliser une matrice. Ce visuel propose les jeux de données en ligne et les gènes en colonnes. Les cellules de cette matrice sont colorées selon leur hasard ratio. En bleu, nous avons les croisements pour lesquels le hasard ratio est négatif et en rouge celle pour lesquels le hasard ratio est positif.

METTRE LE VISUEL

Avec ce visuel, il est intéressant de repérer les gènes pour lesquels la même conclusion se répète dans plusieurs jeux de données. En effet, plus une même conclusion se répète plus, celle-ci peut être considérée comment fiable.

**Matrice des gènes candidats croisés avec les différents sous-types**

Un autre visuel qu’il est intéressant de représenter, c’est une matrice qui croise les gènes les plus susceptibles d’être impliquées dans le développement du cancer du sein avec les différents sous-types. Dans cette matrice, nous ne nous limitons pas aux sous-types moléculaires présentés tout au long de l’analyse de survie, mais nous analyserons également les stages (I, II, III, IV) et la classification TMN (M1, N0, N1, N2, N3). Dans nos analyses, nous réalisons une matrice par jeu de données (8 au total). Voici un exemple avec le jeu de données TCGA-BRCA :



Dans la matrice ci-dessus, la quantité qui définie l’échelle de couleur est le Logarithme du Hazard Ratio. Il est également important de noter qu’il y a 2 lignes pour chaque gène dans la matrice, la première ligne représente les données de survie globale (Overall Survival) et la seconde ligne représente les données de survie sans rechute (Disease Free Survival).

DIRE QUE EN DFS, on a bcp plus de données car, le cancer du sein est un des cancers les moins mortels. Dons c’est plus rare d’avoir des données OS que des DFS.