

Module Entrepôt et fouille de données

18 Decembre 2017

Master 2 Informatique Biomédicale

YIMEN KAMGANG MICHELE LOUNICI-ALI SELMA

1. **Introduction**

Le cancer du poumon représente la première cause de mortalité chez l'homme, ils sont les plus répandus dans le monde avec 900 000 nouveaux cas par an chez l'homme et 330 000 chez la femme, Ainsi le cancer du poumon est l'une des formes les plus fréquentes de cancer avec près de 40000 nouveaux cas par an. Plus de 4 cancers du poumon sur 5 sont liés au tabac. Les hommes sont actuellement plus touchés que les femmes. Néanmoins, cette tendance tend à disparaître compte tenu de l'augmentation du tabagisme féminin.

Le tabagisme représente la principale étiologie du cancer du poumon. Selon les estimations, près de 92 % des décès par cancer des poumons chez l'homme résultent d'une consommation de tabac. Le risque s'accroît en fonction de plusieurs paramètres : dose journalière de tabac, durée du tabagisme… La durée pendant laquelle on fume semble plus importante que la quantité de cigarettes fumées. Néanmoins, d'autres facteurs extérieurs peuvent être impliqués dans la survenue d'un cancer bronchique. C'est le cas, notamment, d'une exposition prolongée à des substances radioactives, à de l'amiante ou à d'autres toxiques (arsenic, nickel, chrome…). On peut également citer parmi les facteurs de risque la consommation de cannabis, les maladies inflammatoires chroniques des bronches ou encore la pollution atmosphérique.

Le cancer du poumon, appelé également cancer bronchique ou cancer broncho-pulmonaire, est une tumeur maligne développée à partir des cellules du poumon. Il existe différentes sortes de cancers pulmonaires, comme le cancer à petites cellules, très agressif, ou les adénocarcinomes, les carcinomes épidermoïdes entre autres...

L’objectif du projet est de faire une analyse descriptive des données cliniques issues d’une cohorte de patients atteints d’un cancer du poumon et d’implémenter des arbres de décision interactifs pour prédire les différents sous-types de cancer à partir des données d’expression génétique en utilisant les packages cart et randomforest.

1. **Matériels et méthodes**

Cette partie vise à décrire les outils utilisés pour atteindre l’objectif assigné à nos travaux ainsi que les différentes étapes suivies.

* 1. **Matériel :**
* RStudio version 3.3.2  est un environnement de développement multiplateforme pour R, un langage de programmation utilisé pour le traitement de données et l’analyse statistique. Il est disponible sous la licence libre AGPLv3, ou bien sous une licence commerciale, soumise à un abonnement annuel.
* GEO2R est utilisé pour comparer deux ou plusieurs groupes d'échantillons afin d'identifier les gènes qui sont différentiellement exprimés dans les conditions expérimentales. Les résultats sont présentés sous la forme d'un tableau de gènes classés par ordre de signification.
* Jeux de données « Lung3.metadata »  a été utilisé dans le cadre de ce projet, il renseigne sur certains aspects cliniques de 89 patients atteints du cancer de poumon.

# Résultat : question1

L’interface presenté ci-dessous permet de visualiser les données cliniques de chaque patient. Lorsque l’utilisateur selectionne un patient dans la liste deroulante, il obtient directement un tableau contenant les données cliniques du patient sélectionné, ce tableau est situé dans l’onglet table. Grâce au bouton « voir tous les paients » l’interface permet également de representer tous les patients et leurs données cliniques, de plus elle permet aussi de presenter la repartition et le grade du cancer chez les femmes et chez les hommes à l’aide dun graphe.

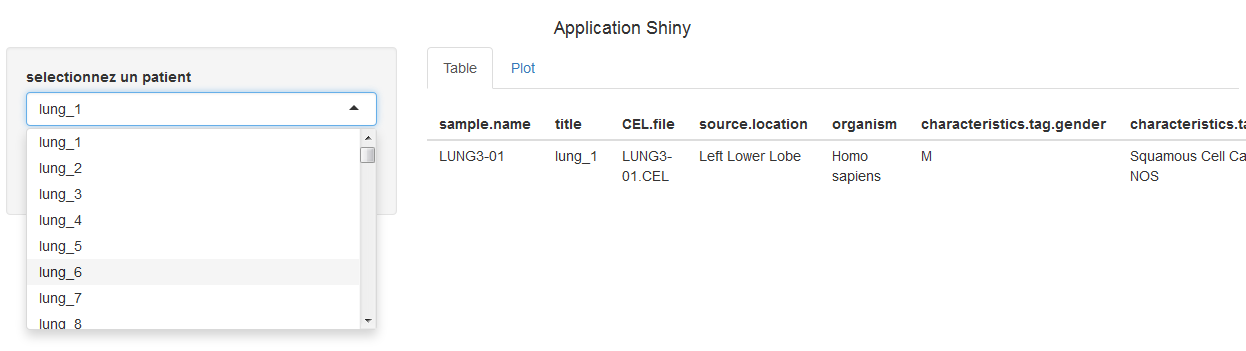


Figure 1: Patient Lung\_1 sélectionné à droite données du patient

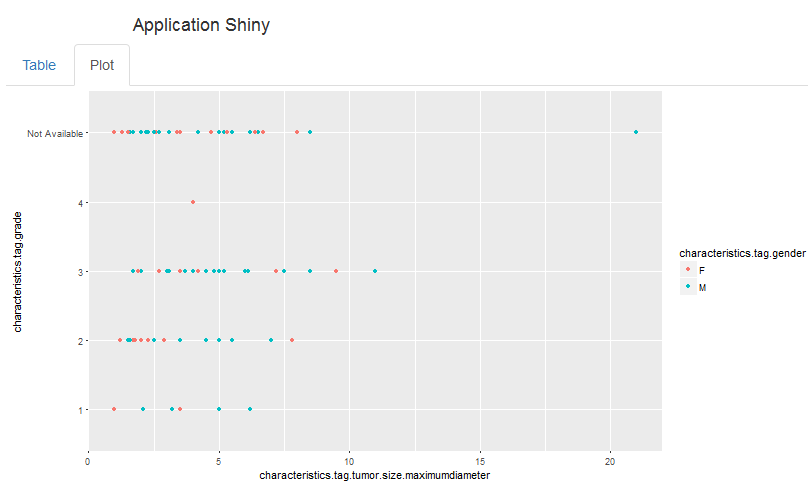
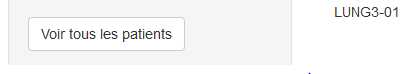


Figure 2:Plot présentant la taille et le grade de la tumeur chez les femmes et les hommes

En cliquant sur le bouton « voir tous les patients » on obtient un tableau avec tous les patients et l’interface donne la possibilité de choisir le nombre de patients qu’on souhaite avoiret donne également la possibilité de faire des recherches sur les patients.



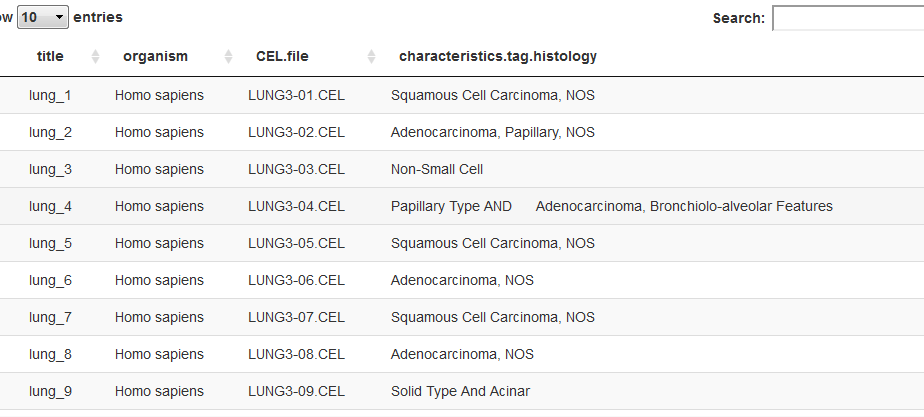


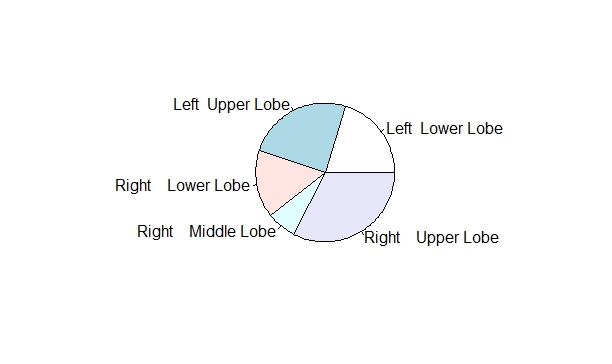
Figure : Tableau de tous les patients

1. **L’analyse descriptive des données cliniques des patients**

A l’aide du logiciel R version 3.3.3 (R est un logiciel libre de traitement des données et d'analyse statistiques mettant en œuvre le langage de programmation S) des statistiques descriptives ont été effectuées sur l’ensemble des variables constituants les données des patients atteints du cancer du poumon. Cela on passant par différentes étapes comme l’importation du fichier « lung3 » la sélection des variables pertinentes en utilisant des commandes propre au langage R ( voir Script)

* 1. **La variable : Localisation**

C’est une variable qualitative qui représente la localisation de la tumeur nous avons pu la décrire avec une représentation graphique en camembére (pie(table(dttp$Localisation))

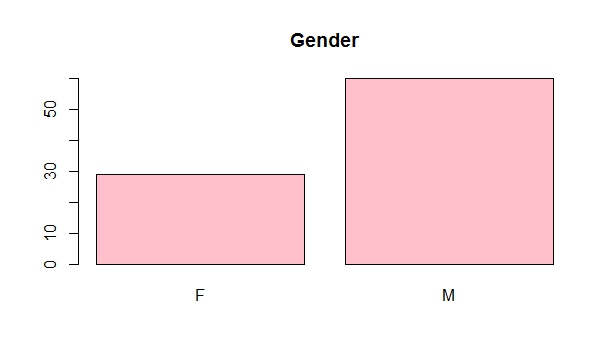


|  |  |
| --- | --- |
| Localisation | Patient |
| Left Lower Lobe | 18 |
| Left Upper Lobe | 22 |
| Right Lower Lobe | 14 |
| Right Middle Lobe | 6 |
| Right Upper Lobe | 29 |

*Figure4 : Source.location*

* 1. **La variable Gender**

Les patients sont constitués de 29 femmes et 60 hommes



*Figure5 : Gender*

* 1. **La variable : caractéristiques histologique**

Il existe 17 sous types de cancer qui

# Résultat :

# question3 : analyses univariées entre les données d’expression et les sous-types de cancer

On a utilisé GEO2R pour comparer deux groupes d'échantillons : Adenocarcinoma et Cell Carcinoma afin d'identifier les gènes qui sont différentiellement exprimés dans les conditions expérimentales. Les résultats sont présentés sous la forme d'un tableau de gènes classés par ordre de signification. Les échantillons utilisés proviennent de la serie GSE58661.

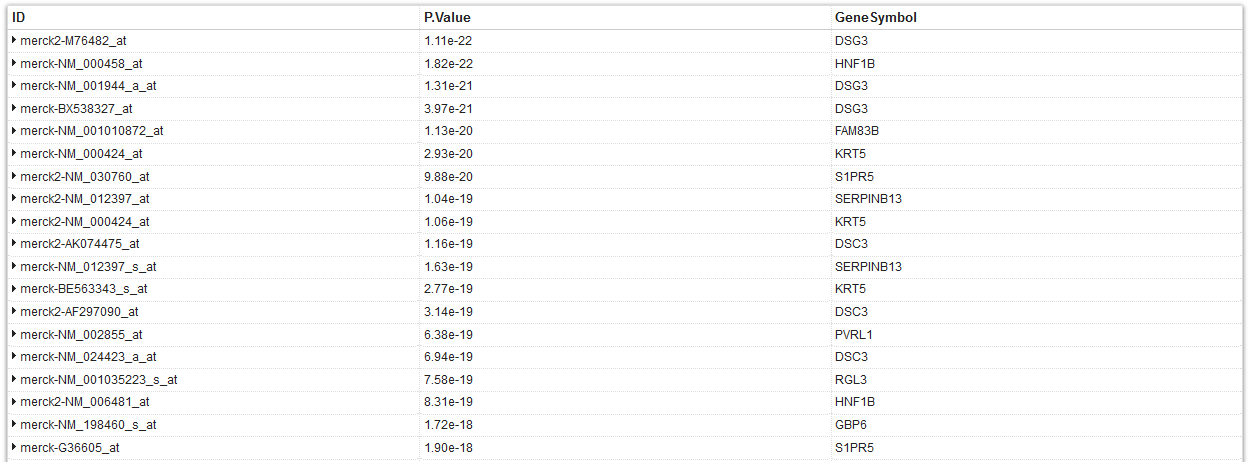
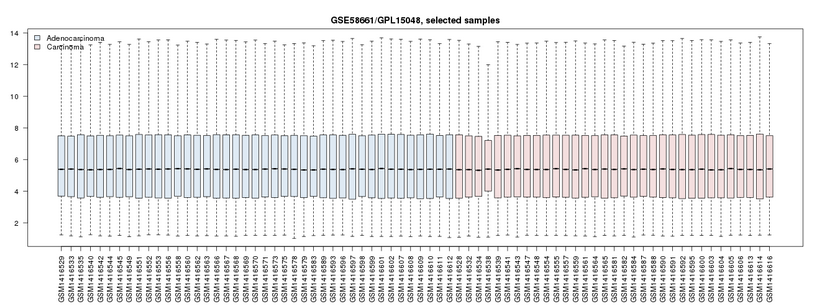
 Lorsque l’on réalise une analyse d’expression différentielle, on obtient le tableau et l’ensemble de Boxplot ci-dessous (par GEO2R et R) :

Figure 4:Tableau de gènes exprimés

On note qu’il y a une différence d’expression entre les patients atteints du type de tumeur Adenocarcinoma et ceux atteints de Carcinoma.