Projet de Biostatistiques

Licence 3

Lucie Gaspard-Boulinc
Pierre Vincens
Kévin Jean

Département de Biologie
École Normale Supérieure – Paris Sciences et Lettres

Modalités du projet

- Groupe de **4 personnes**
- **Séances:**
  - 07/02 (3h)
  - 12/02 (3h)
  - 14/02 (3h)
  - 21/02 -> Soutenances

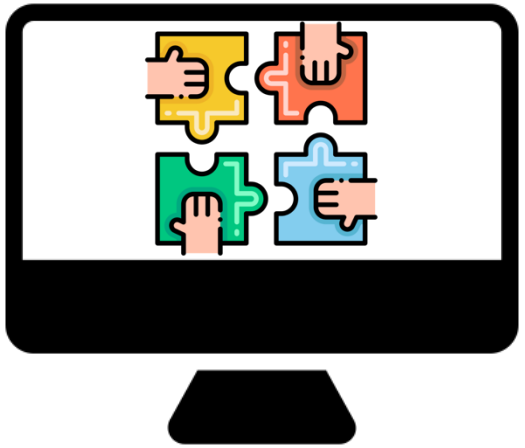
<u>Évaluation</u>:

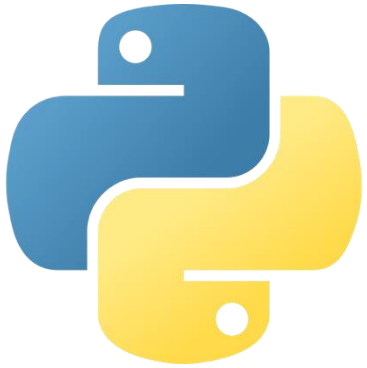
compte rendu écrit court de 2 pages maximum

et

présentation orale de 8 minutes par groupe

Attendus du projet





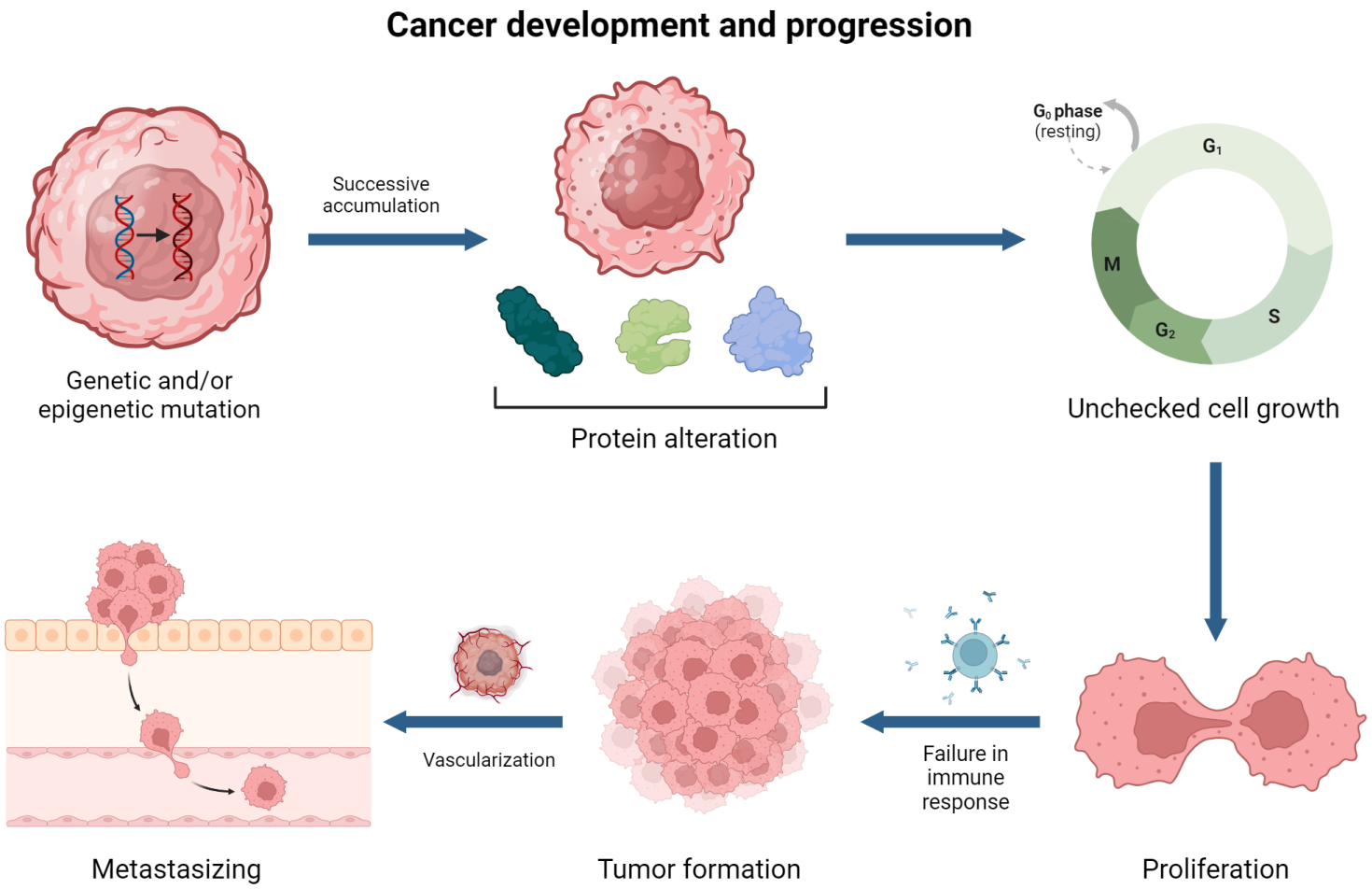
Analyse statistique d'un jeu de données en relation avec une question biologique

- Savoir faire une analyse descriptive d'un jeu de données
- Implémenter les tests statistiques vus en cours
- Utiliser un langage de programmation
- Être capable de poser une question biologique et d'implémenter une approche statistique pour y répondre

L’organotropisme métastatique

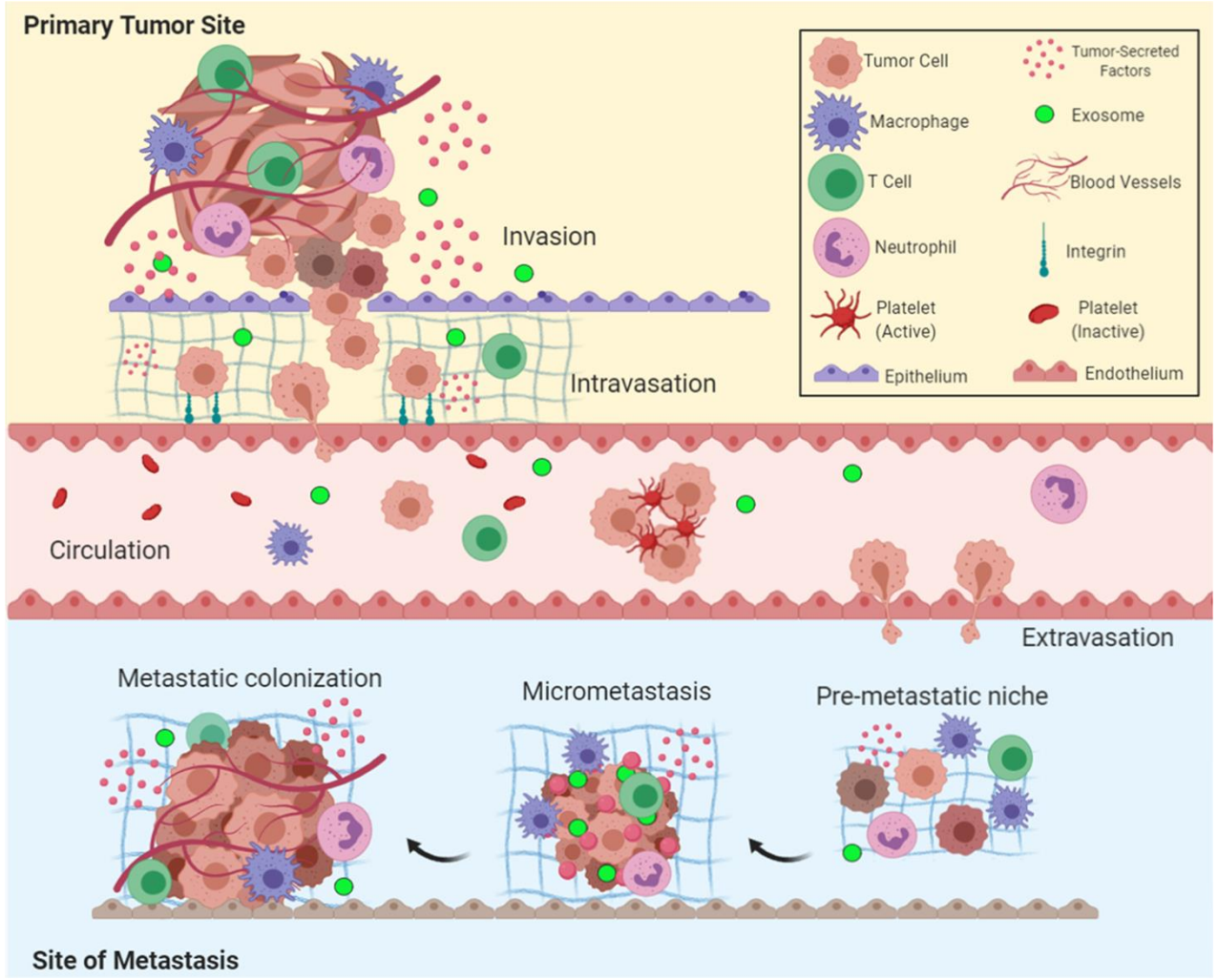
Contexte du projet de Biostatistiques

Les cancers



- Prolifération incontrôlée des cellules
- Échappement aux mécanismes de régulations usuels
- Destructions des tissus sains
- Envahissement d'autres organes

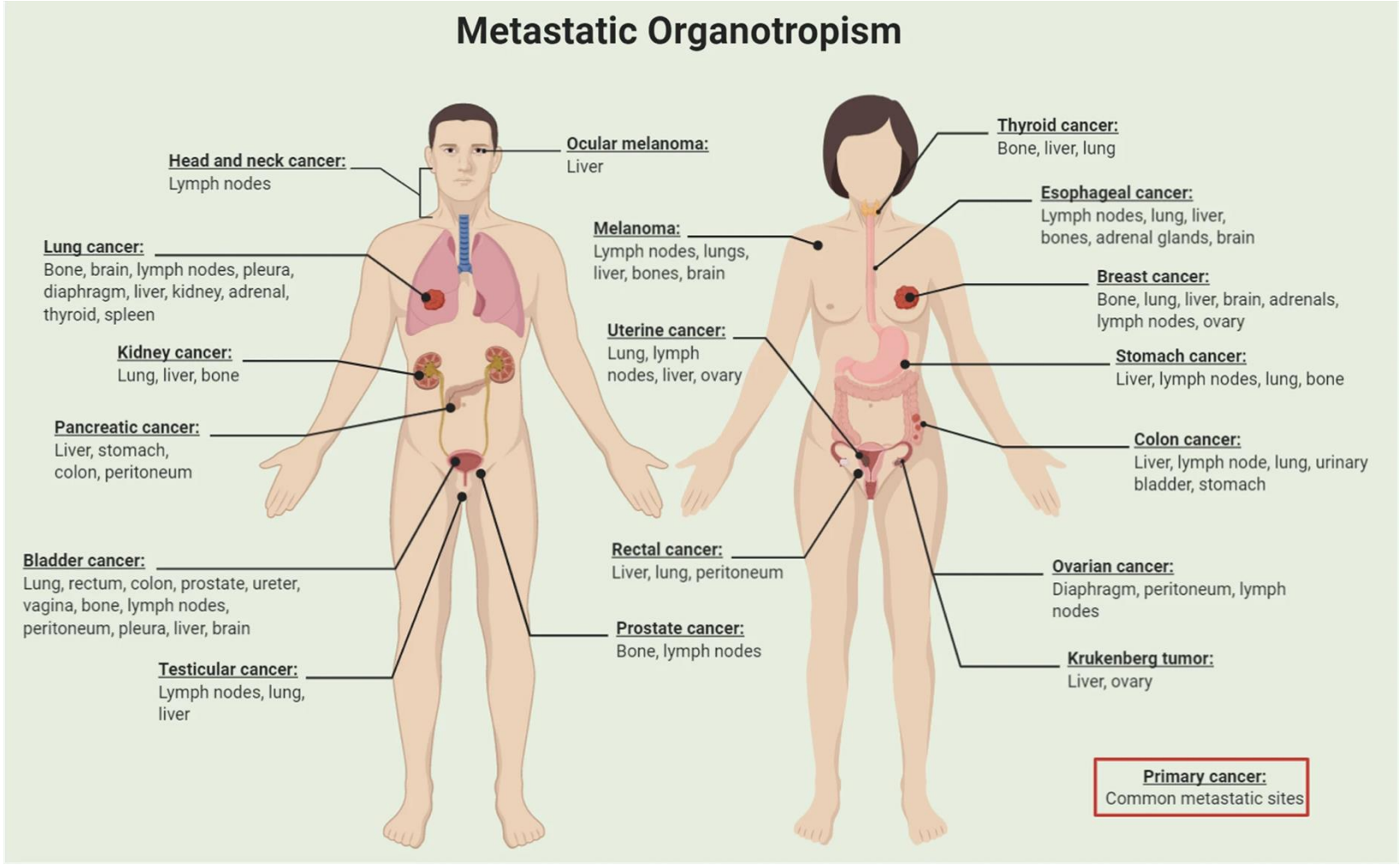
Les métastases



(Fares, J., Fares, M.Y., Khachfe, H.H. et al., Sig Transduct Target Ther ,2020)

- La plupart des morts liées au cancer sont dues aux métastases
- **Modèle simpliste de la semence et du sol:**
  1. Cellules avec capacité de migration (EMT)
  2. Circulation via les vaisseaux sanguins
  3. Colonisation dans un environnement propice au développement
  4. Formation de la métastase

L’organotropisme métastatique

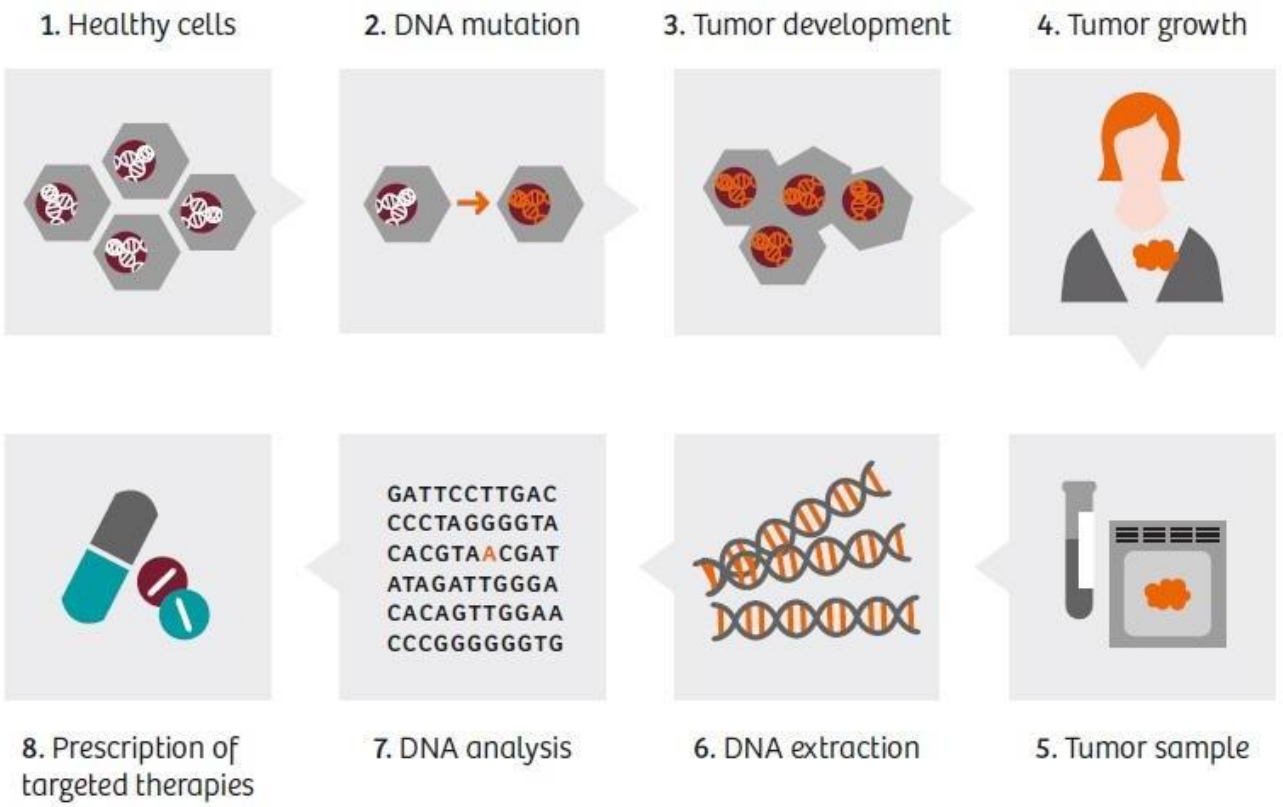


Wikipedia

Problématique(s)

- Comment caractériser l'organotropisme métastatique ?
- Quels sont les déterminants génétiques de l'organotropisme ?
- Les tumeurs primaires avec métastases présentent-elles des caractéristiques différentes que celles sans métastase ?
- Quels sont les liens entre organotropisme métastatique et le pronostic des patients?
- Et bien d'autres encore…

Un élément de réponse



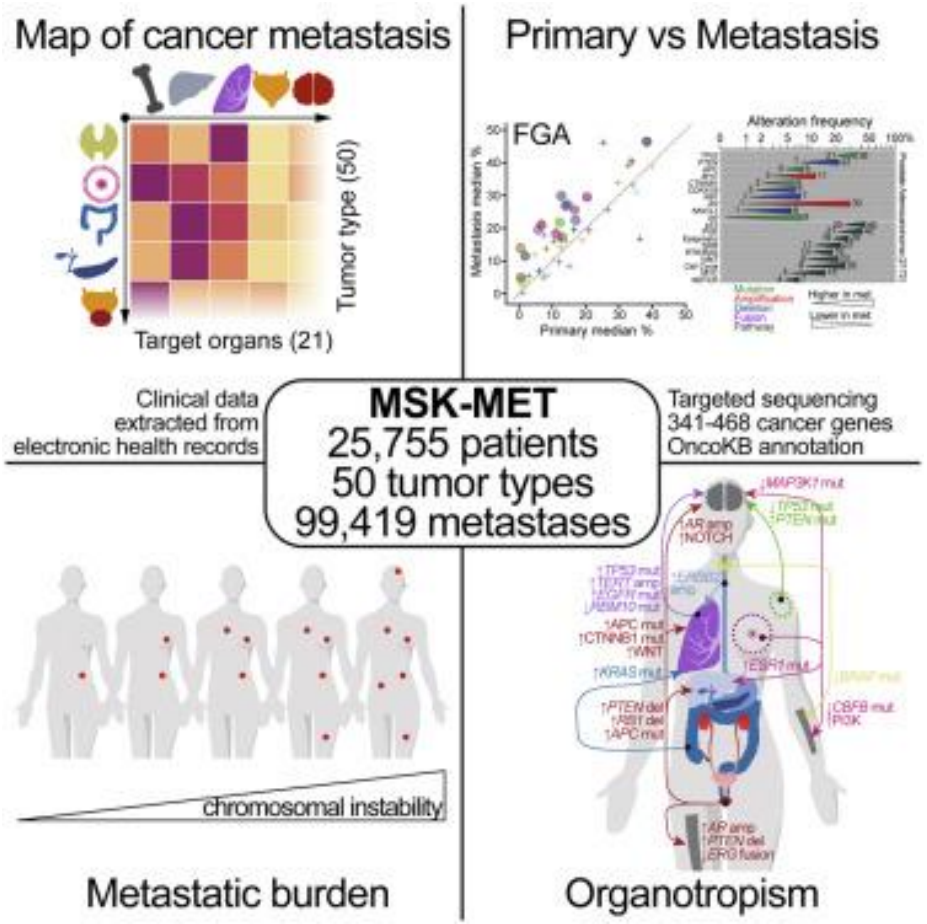
Siemens Healthineers

Le profilage moléculaire de l'ADN tumoral

« Méthode de laboratoire qui utilise un échantillon de tissu, de sang ou d'un autre liquide corporel pour rechercher certains gènes, protéines ou autres molécules qui peuvent être le signe d'une maladie ou d'un état pathologique, comme le cancer."

Mais les efforts de profilage moléculaire se sont pour le moment peu intéressé à l'organotropisme métastatique

Le jeu de données



Nguyen B, Fong C, Luthra A, et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. *Cell*. 2022;185(3):563-575.e11. doi:10.1016/j.cell.2022.01.003

- Analyse d'un jeu de données collectées aux États-Unis dans le cadre du programme MSK-MET (Memorial Sloan Kettering - Metastatic Events and Tropism)
- > 25,000 patients inclus
- Étude PAN-CANCER
- Profilage moléculaire de tumeurs primaires et de metastases
- 590 variables
- 2 fichiers:
  - MSKMET pan cancer.csv
  - MSKMET pancancer description.csv

Déroulé des séances (prévisionnel)

- Aujourd'hui: découverte du jeu de données, analyse descriptive guidée et début des analyses statistiques

- 12/02: Continuation des analyses statistiques et identification de votre question d'intérêt

- 14/02: Mise en place et réalisation de l'étude statistiques, rédaction du rapport court

RENDEZ-VOUS SUR

Projet de biostatistiques

Séance n°2 - 12/02/2025

Compte rendu de la séance n°1

- Chargement des fichiers et reconnaissance du nombre et type de variables
- Exploration du fichier et partie 1 des questions

```
# Données du projet
MSK_MET <- read.csv("./MSKMET_pan_cancer.csv", row.names = 1)
head(MSK_MET)
```

	SAMPLE_ID <chr>	PATIENT_ID <chr>	SEX <chr>	OS_MONTHS <dbl>	OS_STATUS <chr>	AGE_AT_EVIDENCE_OF_METS <dbl>
1	P-0000004-T01-IM3	P-0000004	Female	3.78	1:DECEASED	39.66
2	P-0000015-T01-IM3	P-0000015	Female	13.90	1:DECEASED	44.25
3	P-0000024-T01-IM3	P-0000024	Female	35.06	1:DECEASED	59.44
4	P-0000025-T02-IM5	P-0000025	Female	46.00	1:DECEASED	69.65
5	P-0000026-T01-IM3	P-0000026	Female	80.59	0:LIVING	68.41
6	P-0000027-T01-IM3	P-0000027	Female	12.32	1:DECEASED	79.50

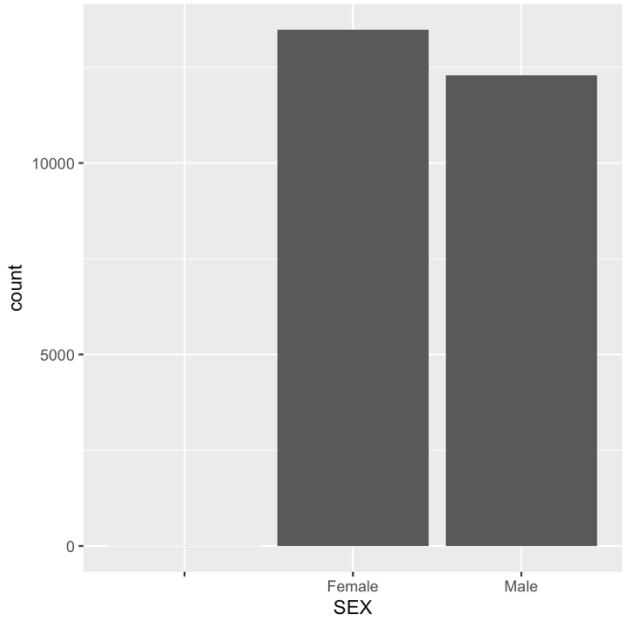
6 rows | 1-7 of 591 columns

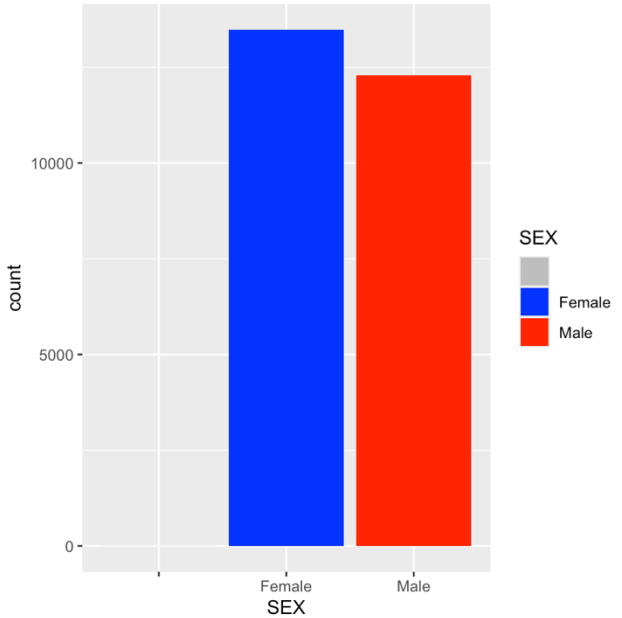
```
## Il y a 25775 patients dans le jeu de données et 590 variables.
```

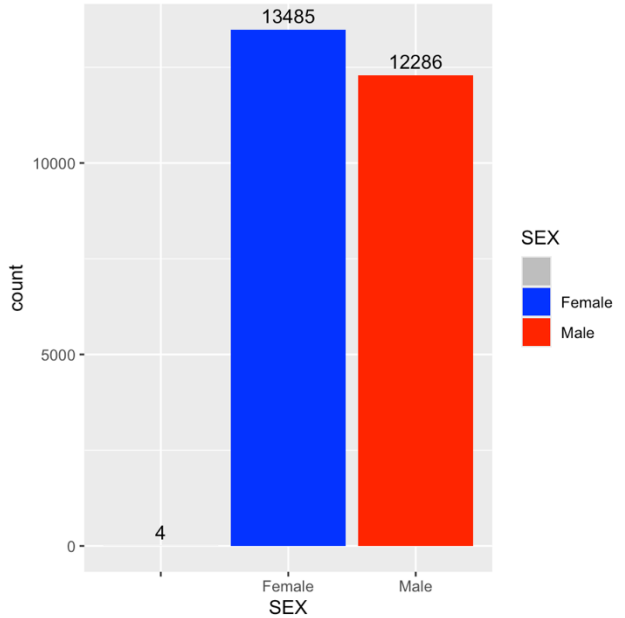
1. De quel type sont les variables du jeu de données (qualitatives, quantitatives, discrètes, continues...) ?
2. Lesquelles contiennent des valeurs manquantes ? En quelles proportions ?
3. Calculez les informations statistiques (moyenne, médiane, écart-type) que l'on peut obtenir sur ces variables.
4.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

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'échantillons de métastases ? Combien par sous-type de cancer ? Faire les représentations appropriées.
6. Combien d'échantillons de tumeurs primaires ont été étudie chez des patients presentant deja un cancer m métastatiques ? Combien par type et sous-type de cancer ?

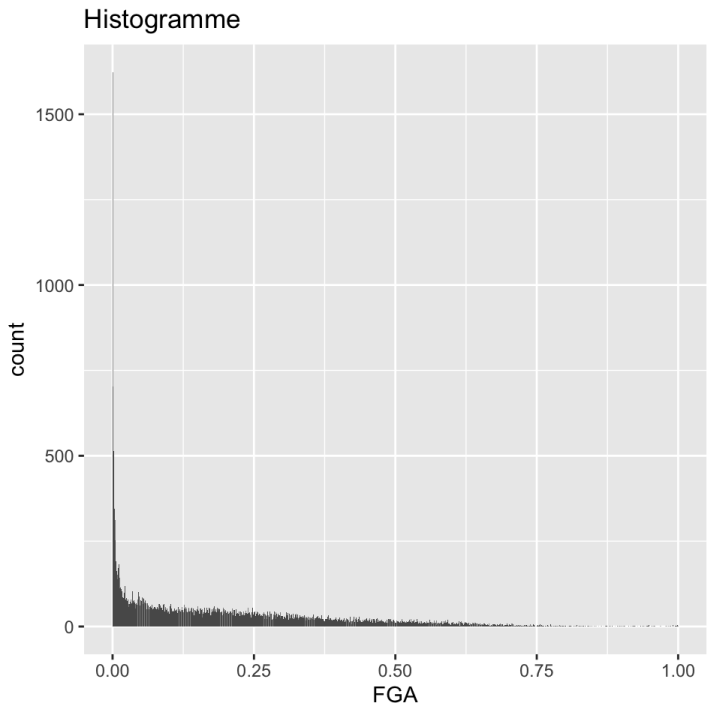
Exemple 1:
SEX

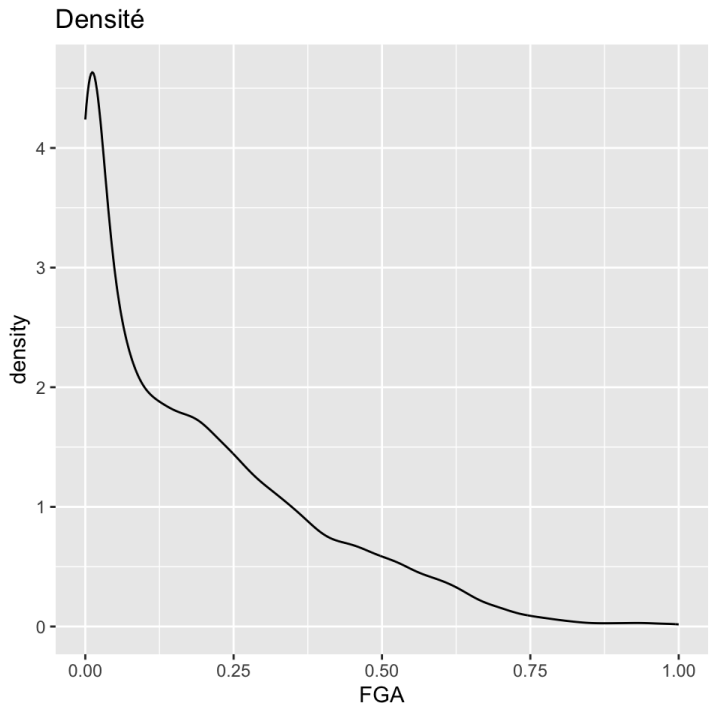


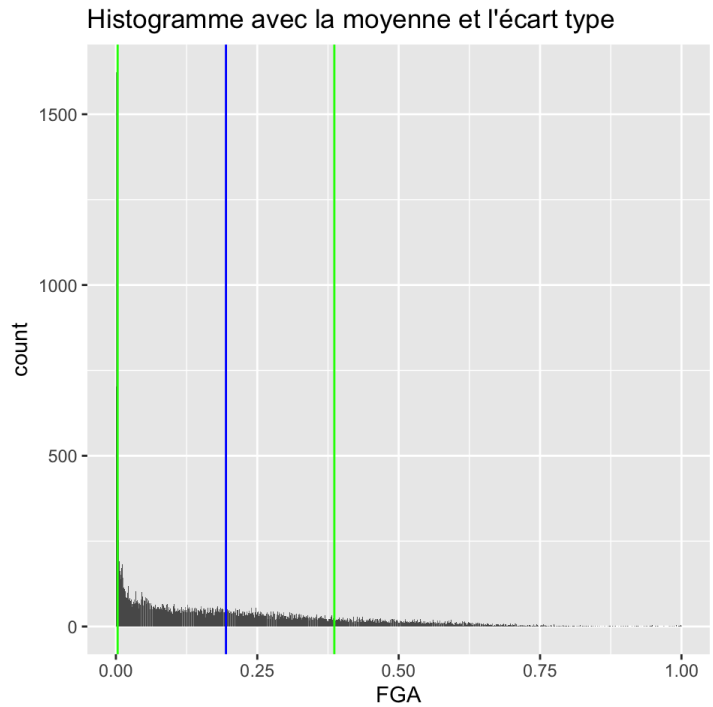




Exemple 2:
FGA

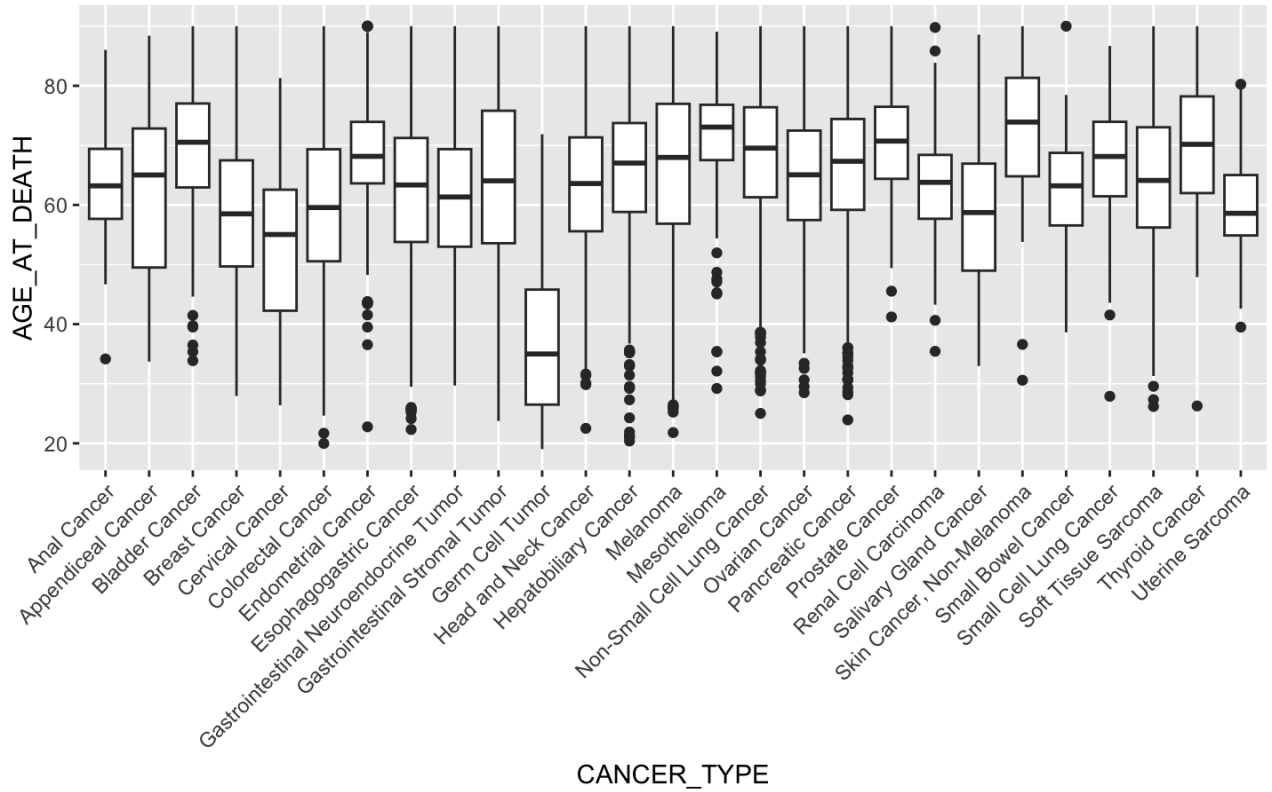




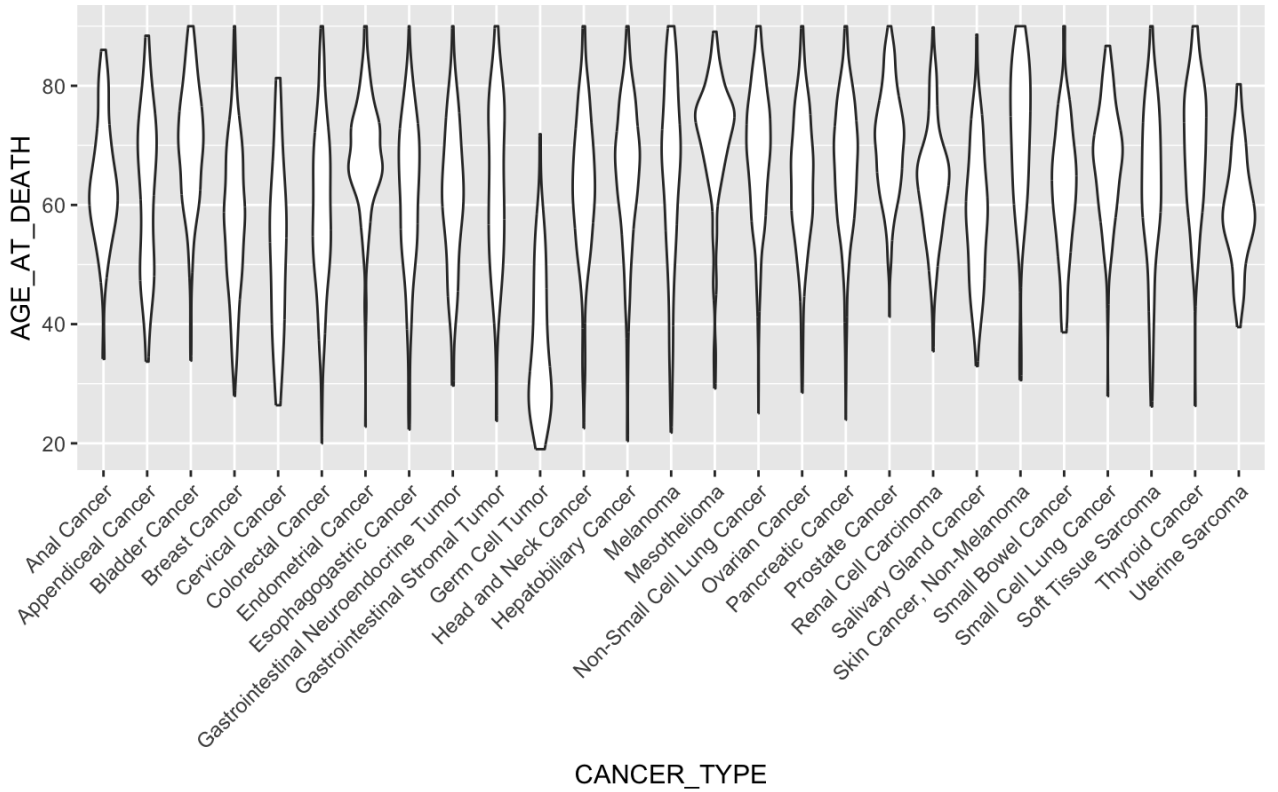


Exemple 3: AGE AT DEATH ~ CANCER TYPE

Version Boxplot



Version ViolinPlot



Et encore pleins d'autres graphiques et comparaisons possibles

590 variables (dont 9 variables numériques et 581 variables catégorielles)

> 25, 000 patients

590 variables (dont 9 variables numériques et 581 variables catégorielles)

> 25, 000 patients

MSK-MET dataset

Informations personnelles	Information Diagnostic/ Pronostic	Echantillon	Information sur les métastases	Résultats du profilage moléculaire
PATIENT_ID	CANCER TYPE	SAMPLE_ID	IS_DIST_MET_MAPPED	FGA
STATUS	SUBTYPE	**SAMPLE_TYPE**	MET_COUNT	MSI_SCORE
AGE_AT_DEATH	SURVIVAL	GENE_PANEL	MET_SITE_COUNT	MSI_TYPE
SEX	…	TUMOR_PURITY	DMETS_DX_*	TMB_NONSYNONYMOUS
AGE_AT_SURGERY		…	…	CNA_*
…				…

590 variables (dont 9 variables numériques et 581 variables catégorielles)

> 25, 000 patients

MSK-MET dataset

Informations personnelles	Information Diagnostic/ Pronostic	Echantillon	Information sur les métastases	Résultats du profilage moléculaire
PATIENT_ID STATUS AGE_AT_DEATH SEX AGE_AT_SURGERY …	CANCER TYPE SUBTYPE SURVIVAL …	SAMPLE_ID **SAMPLE_TYPE** GENE_PANEL TUMOR_PURITY …	IS_DIST_MET_MAPPED MET_COUNT MET_SITE_COUNT DMETS_DX_* …	FGA MSI_SCORE MSI_TYPE TMB_NONSYNONYMOUS CNA_* …

DIAGNOSTIC

Hiérarchie

- Organ
- Cancer type
- Cancer subtype

SAMPLE_TYPE

- Primary
- Metastasis

Faire et présenter une analyse statistiques

- Nous allons travailler sur la première question du sujet:

1. Les tumeurs primaires présentent-elles significativement plus d'altération génomiques que les métastases ? Est-ce dépendant du type de cancer ?

- Il s'agit bien ici d'une question de **BIOLOGIE**

Faire et présenter une analyse statistiques

- Nous allons travailler sur la première question du sujet:

1. Les tumeurs primaires présentent-elles significativement plus d'altération génomiques que les métastases ? Est-ce dépendant du type de cancer ?

- Il s'agit bien ici d'une question de **BIOLOGIE**
- Le jeu de donnée est adapté pour y répondre
- Nous avons besoin des statistiques pour faire le test

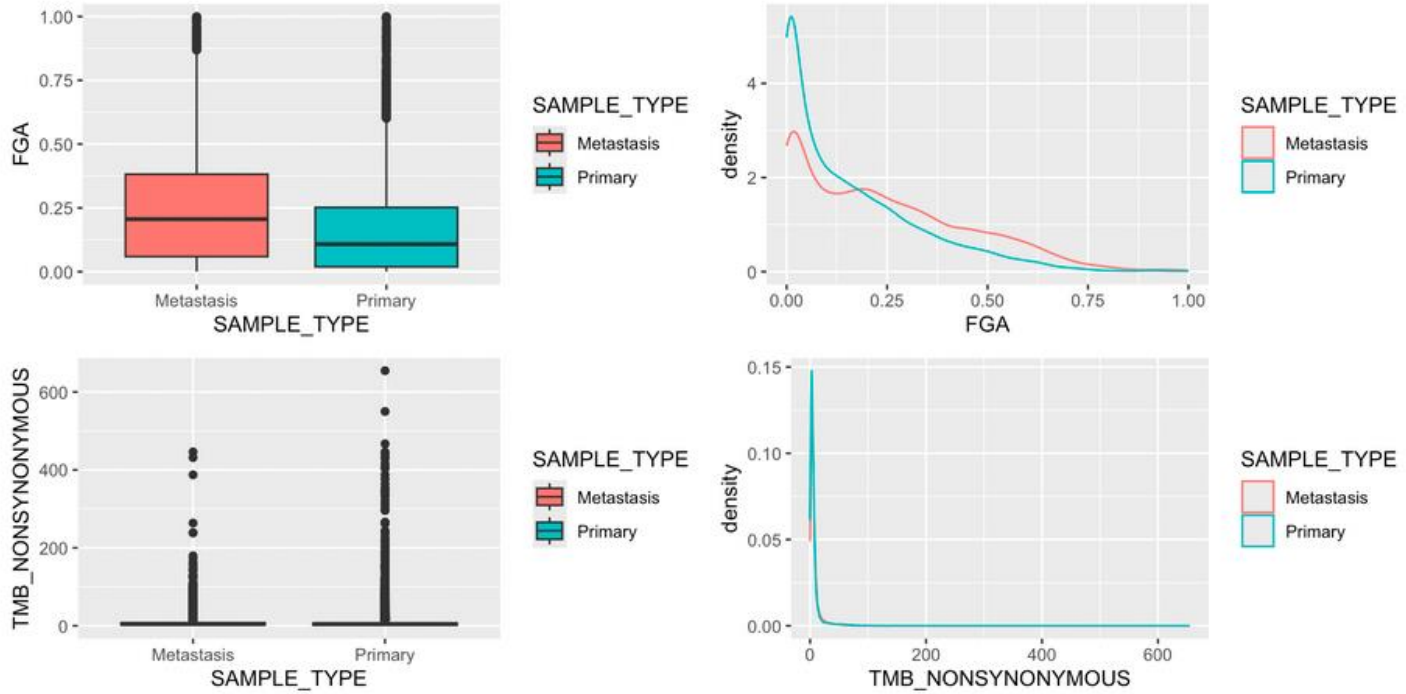
QUELLE EST LA PREMIÈRE ÉTAPE ?

ETAPE 1: FAIRE DES REPRESENTATIONS GRAPHIQUES

1. Les tumeurs primaires présentent-elles significativement plus d'altération génomiques que les métastases ? Est-ce dépendant du stype de cancer ?

Nous nous plaçons ici dans la cadre de la comparaison entre échantillons de tumeurs primaires et les métastases. Les altérations génomiques sont présentées de multiples façon: la fraction de génome alteré (FGA), la charge mutationelle (TMB_NONSYNONYMOUS) ainsi que le nombre de CNV (CNA_*). On peut donc effectuer plusieurs comparaisons. Je donne ici l'example avec le FGA.

```
# Visualisation du FGA
p1 <- ggplot(MSK_MET, aes(x=SAMPLE_TYPE, y=FGA, fill=SAMPLE_TYPE)) + geom_boxplot()
p2 <- ggplot(MSK_MET, aes(x=FGA, color=SAMPLE_TYPE)) + geom_density()
p3 <- ggplot(MSK_MET, aes(x=SAMPLE_TYPE, y=TMB_NONSYNONYMOUS, fill=SAMPLE_TYPE)) + geom_boxplot()
p4 <- ggplot(MSK_MET, aes(x=TMB_NONSYNONYMOUS, color=SAMPLE_TYPE)) + geom_density()
ggarrange(p1,p2,p3,p4, ncol = 2, nrow = 2)
```



C'est important car:

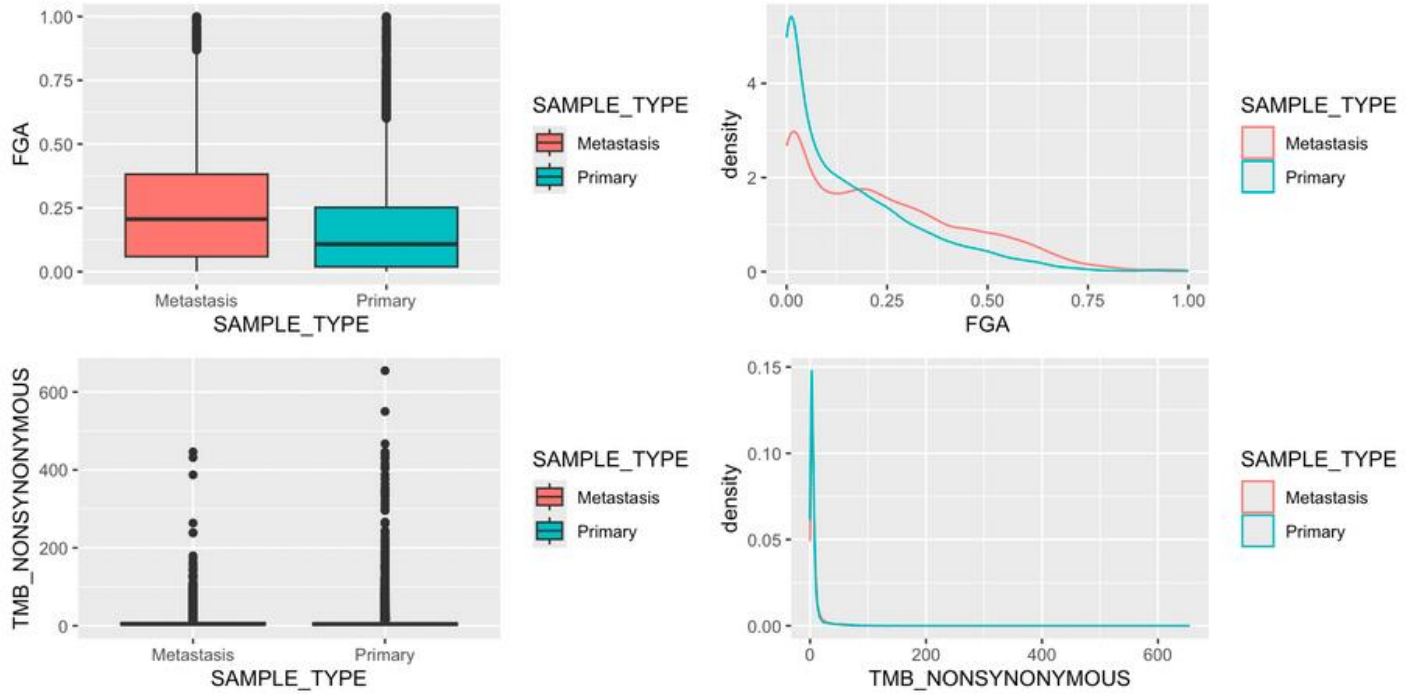
- Permet d'identifier quelles variables sont nécessaires pour répondre à la question
- Visualiser les données, et notamment leur **distribution**
- Permet d'orienter sur le traitement statistique

ETAPE 1: FAIRE DES REPRESENTATIONS GRAPHIQUES

1. Les tumeurs primaires présentent-elles significativement plus d'altération génomiques que les métastases ? Est-ce dépendant du stype de cancer ?

Nous nous plaçons ici dans la cadre de la comparaison entre échantillons de tumeurs primaires et les métastases. Les altérations génomiques sont présentées de multiples façon: la fraction de génome alteré (FGA), la charge mutationelle (TMB_NONSYNONYMOUS) ainsi que le nombre de CNV (CNA_*). On peut donc effectuer plusieurs comparaisons. Je donne ici l'example avec le FGA.

```
# Visualisation du FGA
p1 <- ggplot(MSK_MET, aes(x=SAMPLE_TYPE, y=FGA, fill=SAMPLE_TYPE)) + geom_boxplot()
p2 <- ggplot(MSK_MET, aes(x=FGA, color=SAMPLE_TYPE)) + geom_density()
p3 <- ggplot(MSK_MET, aes(x=SAMPLE_TYPE, y=TMB_NONSYNONYMOUS, fill=SAMPLE_TYPE)) + geom_boxplot()
p4 <- ggplot(MSK_MET, aes(x=TMB_NONSYNONYMOUS, color=SAMPLE_TYPE)) + geom_density()
ggarrange(p1,p2,p3,p4, ncol = 2, nrow = 2)
```

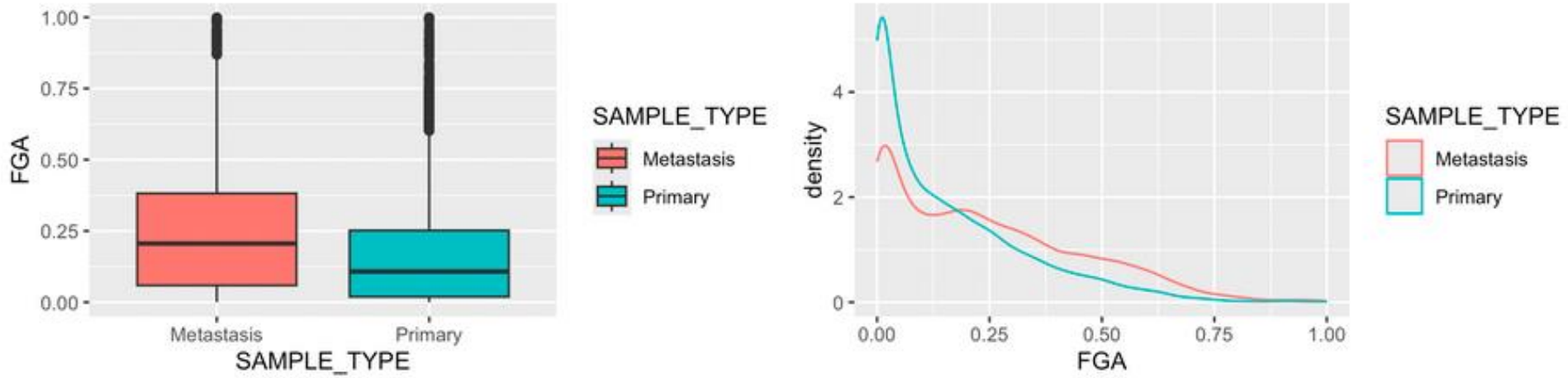


C'est important car:

- Permet d'identifier quelles variables sont nécessaires pour répondre à la question
- Visualiser les données, et notamment leur **distribution**
- Permet d'orienter sur le traitement statistiques

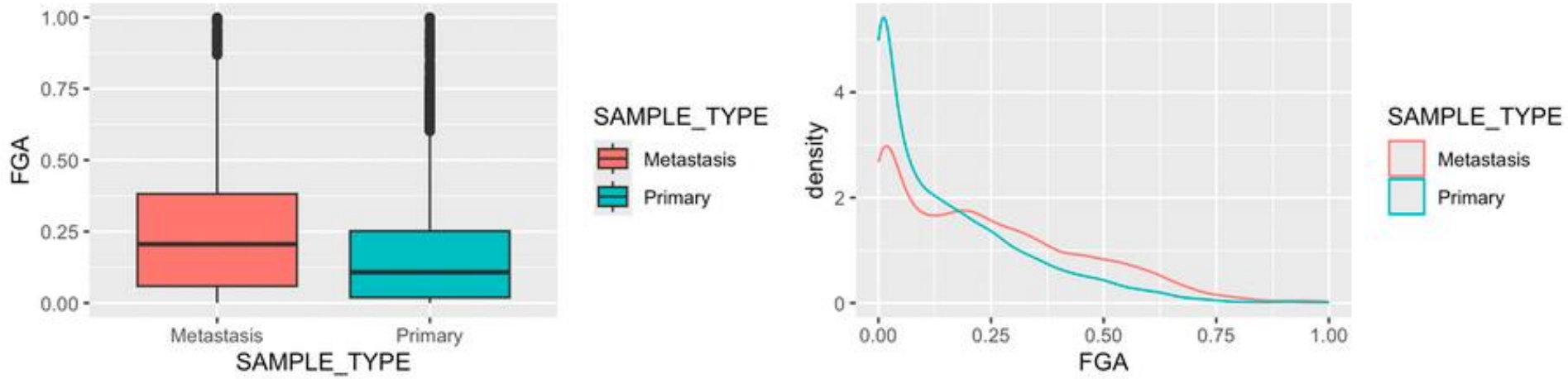
QUELLE EST LA DEUXIÈME ÉTAPE ?

ETAPE 2: POSER LES HYPOTHÈSE



Quelles sont les hypothèses ici ?

ETAPE 2: POSER LES HYPOTHÈSES



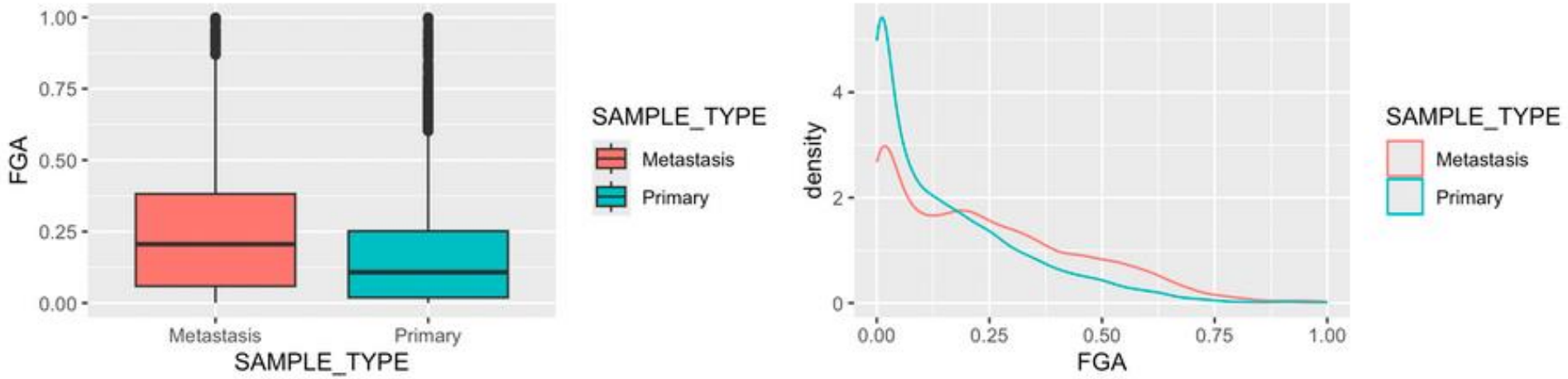
Quelles sont les hypothèses ici ?

Hypothèse nulle H0 : la différence observée est due au hasard de l'échantillonnage (variabilité naturelle)
Hypothèse nulle H0 : La différence observée de FGA est due au hasard

Hypothèse alternative H1 : la différence observée est trop grande pour n'être due qu'au hasard d'échantillonnage
Hypothèse alternative H1: Il y a une différence significative du FGA entre les tumeurs primaires et les métastases.

QUELLE EST LA TROISIÈME ÉTAPE ?

ETAPE 3: CHOISIR ET FAIRE LE TEST



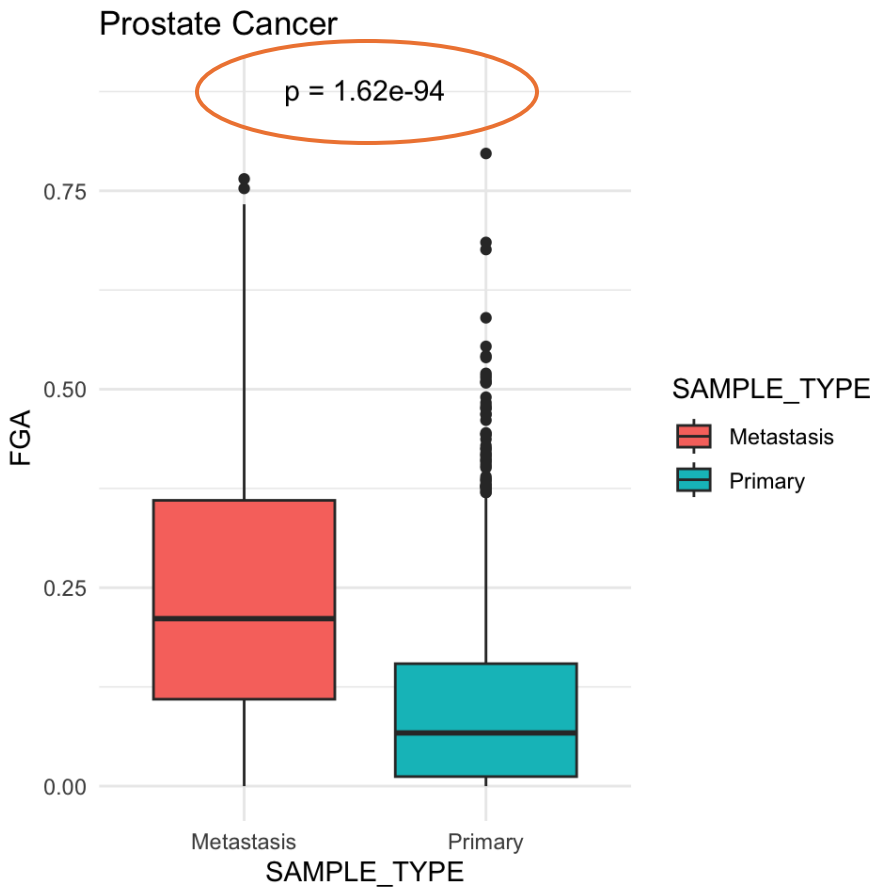
- On souhaite comparer **une différence de moyenne**
- On remarque grâce aux graphiques que la distribution n'est pas normale
- Les **échantillons sont indépendants**

↓

Test de Wilcoxon

wilcox.test(x=primary_group_FGA, y=metastasis_group_FGA)

ETAPE 4: PRÉSENTER SES RÉSULTATS (1)



- Faire la représentation graphique la plus appropriée
- Faire figurer sur la figure le résultat du test (avec la p-value ou des (*)

➔ attention à bien préciser la légende de la figure)

ETAPE 4: PRÉSENTER SES RÉSULTATS (2)

- **Bien formuler son texte**

« La moyenne de FGA dans les métastases de cancer de la prostate est de 0.2 alors qu'elle est de 0.07 dans les tumeurs primaires de cancers de la prostate. Un test de Wilcoxon a été réalisé pour comparer le FGA entre métasases et tumeurs primaires des cancers de la prostate. Les résultats montrent une difference statistiquement significative (statistic = XX, parameter = XX, p-value= XXX). Les metastases present une fraction de gènome alteré plus grande que les tumeurs primaires dans le cas des cancers de la prostate».

1) Mentionner les éléments de statistiques descriptives et ce que l'on cherche à comparer

ETAPE 4: PRÉSENTER SES RÉSULTATS (2)

- **Bien formuler son texte**

« La moyenne de FGA dans les métastases de cancer de la prostate est de 0.2 alors qu'elle est de 0.07 dans les tumeurs primaires de cancers de la prostate. Un test de Wilcoxon a été réalisé pour comparer le FGA entre métasases et tumeurs primaires des cancers de la prostate. Les résultats montrent une difference statistiquement significative (statistic = XX, parameter = XX, p-value= XXX). Les metastases present une fraction de gènome alteré plus grande que les tumeurs primaires dans le cas des cancers de la prostate».

1) Mentionner les éléments de statistiques descriptives et ce que l'on cherche à comparer

2) Expliquer la démarche effectuée avec le test choisi et la comparaison exacte

ETAPE 4: PRÉSENTER SES RÉSULTATS (2)

- **Bien formuler son texte**

« La moyenne de FGA dans les métastases de cancer de la prostate est de 0.2 alors qu'elle est de 0.07 dans les tumeurs primaires de cancers de la prostate. Un test de Wilcoxon a été réalisé pour comparer le FGA entre métasases et tumeurs primaires des cancers de la prostate. Les résultats montrent une difference statistiquement significative (statistic = XX, parameter = XX, p-value= XXX). Les metastases present une fraction de gènome alteré plus grande que les tumeurs primaires dans le cas des cancers de la prostate».

1) Mentionner les éléments de statistiques descriptives et ce que l'on cherche à comparer

2) Expliquer la démarche effectuée avec le test choisi et la comparaison exacte

3) Exprimer les résultats du test en indiquant les valeurs de la statistique, des parametres (ex: les degres de liberté si t.test) ainsi que la p-value

ETAPE 4: PRÉSENTER SES RÉSULTATS (2)

- **Bien formuler son texte**

« La moyenne de FGA dans les métastases de cancer de la prostate est de 0.2 alors qu'elle est de 0.07 dans les tumeurs primaires de cancers de la prostate. Un test de Wilcoxon a été réalisé pour comparer le FGA entre métasases et tumeurs primaires des cancers de la prostate. Les résultats montrent une difference statistiquement significative (statistic = XX, parameter = XX, p-value= XXX). Les metastases present une fraction de gènome alteré plus grande que les tumeurs primaires dans le cas des cancers de la prostate».

1) Mentionner les éléments de statistiques descriptives et ce que l'on cherche à comparer

2) Expliquer la démarche effectuée avec le test choisi et la comparaison exacte

3) Exprimer les résultats du test en indiquant les valeurs de la statistique, des parametres (ex: les degres de liberté si t.test) ainsi que la p-value

4) Conclusion du test en rapport avec la question biologique (+ Interpretation)

Résumé

Les 4 étapes pour réussir son analyse:

1) Faire des représentations graphiques
2) Poser les hypothèses
3) Choisir le bon test et le faire
4) Présenter ses résultats correctement

Consignes pour le compte rendu

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
- **poser les hypothèses,**
- **réaliser les tests statistiques nécessaires,**
- **produire les graphiques pertinents** illustrant votre question d'intérêt,
- **analyser les résultats et proposer une réponse**,
- effectuer une **synthèse**.