*Cahier des charges*

Biomarqueurs pronostiques dans le cancer du côlon

Ekaterina Flin

06/11/2024

Table des matières

[Données 2](#_Toc181812135)

[Données d’expression de gènes 2](#_Toc181812136)

[Annotations biocliniques 3](#_Toc181812137)

[Annotations du GEC 3](#_Toc181812138)

[Méthodologie 4](#_Toc181812139)

[Identifier les échantillons du cancer du côlon (T-Colon) dans les données 4](#_Toc181812140)

[Construire des courbes de survie pour le Gene Expression Classifier (GEC) 6](#_Toc181812141)

[Tâches à réliser pour le projet tutoré 7](#_Toc181812142)

[Calculer l’impact pronostique du panel de 3 gènes GEC (3-GEC) 7](#_Toc181812143)

[Identifier le statut MSI/MSS pour les échantillons du TCGA-COAD-FPKM 8](#_Toc181812144)

[Analyse des données de méthylation 9](#_Toc181812145)

[Analyse des données de mutation 10](#_Toc181812146)

[Analyse des données de CNA 11](#_Toc181812147)

[Créer un modèle d’apprentissage pour le statut MSI/MSS à partir des données d’expression de gènes TCGA-COAD-FPKM 12](#_Toc181812148)

[Identifier le statut MSI/MSS dans les deux datasets restants GSE39582 et GSE17536 12](#_Toc181812149)

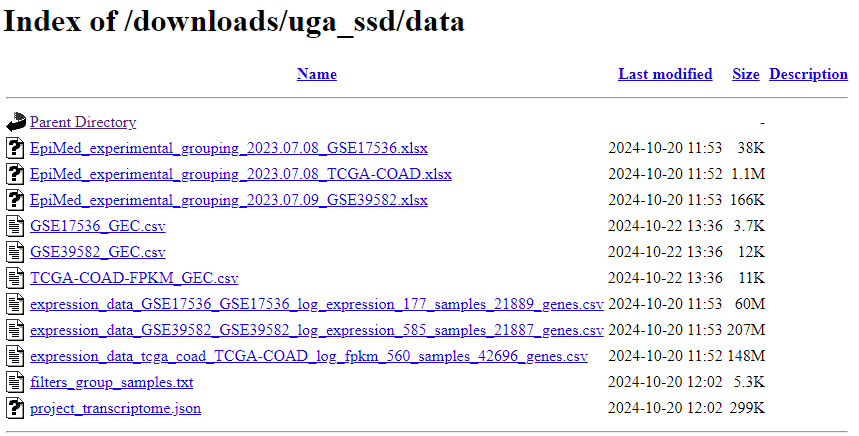
[Réaliser une analyse de survie multivariée pour 4-GEC et 3-GEC en intégrant le statut MSI/MSS 12](#_Toc181812150)

[Réaliser une analyse de survie multivariée pour 4-GEC et 3-GEC en intégrant d’autres facteurs de risque connus 12](#_Toc181812151)

# Données

Les données sont disponibles à l’adresse suivante (voir le dossier « data ») :

http://epimed.univ-grenoble-alpes.fr/downloads/uga\_ssd/



Les données contiennent 3 datasets du cancer du côlon.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Technologie** | **Source** |
| TCGA-COAD | RNA-seq | GDC portal: https://portal.gdc.cancer.gov/ |
| GSE39582 | Microarrays | NCBI GEO: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39582 |
| GSE17536 | Microarrays | NCBI GEO: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17536 |

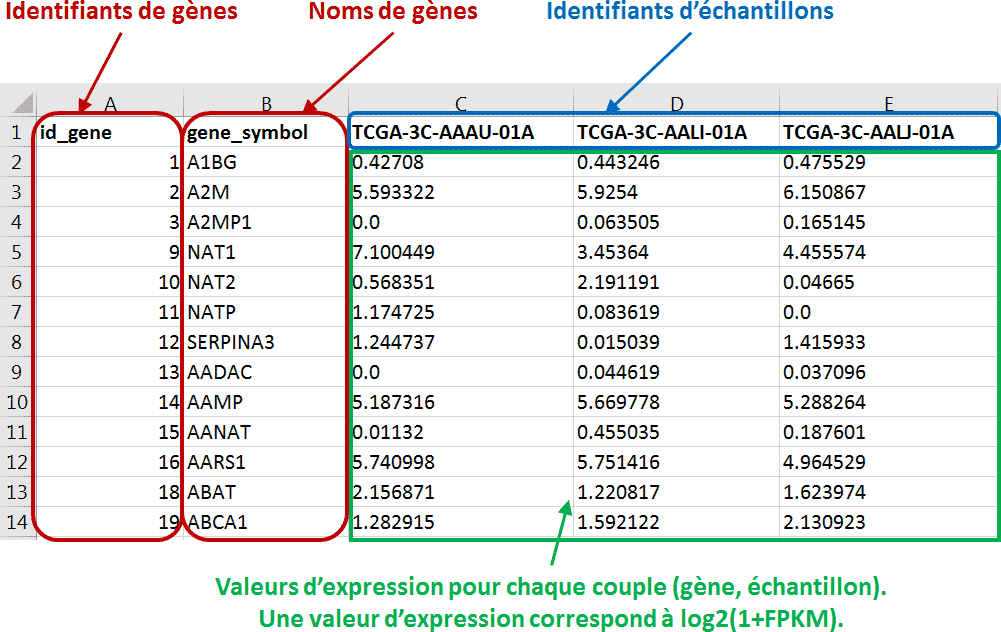
## Données d’expression de gènes

Pour chaque dataset, plusieurs fichiers sont disponibles :

* Les niveaux d’expression de gènes  
  (fichier nommé expression\_data\*.csv, séparateur « ; »)
* Les annotations biocliniques des échantillons  
  (fichier nommé EpiMed\_experimental\_grouping\*.xlsx)
* Le score GEC : \*\_GEC.csv

(expression\_data\*.csv)

Les données transcriptomiques sont disponibles sous la forme d’une matrice de valeurs d’expression de gènes, normalisées et log-transformées pour chaque paire gène-échantillon.



En fonction du dataset, les valeurs peuvent correspondre aux données normalisées FPKM (pour les données RNA-seq) ou à d’autres valeurs d’expression normalisées (pour les données Microarrays). Dans les deux cas, les données sont directement exploitables pour faire des analyses statistiques.

## Annotations biocliniques

(EpiMed\_experimental\_grouping\*.xlsx)

Les annotations biocliniques dans les fichiers Excel sont présentées avec deux onglets : « standard exp\_group » et « original parameters ». Les données de survie sont disponibles dans les colonnes suivantes dans l’onglet « standard exp\_group ».

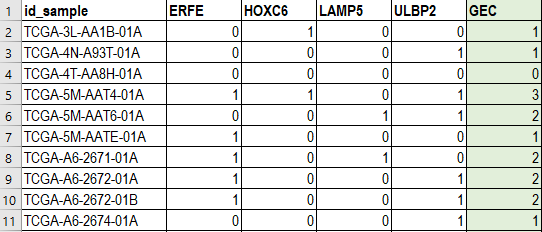
|  |  |
| --- | --- |
| **Colonne** | **Signification** |
| os\_month | La durée de la survie globale OS (overall survival) en mois. |
| os\_censor | La censure de la survie globale.  0 – donnée censurée, c’est-à-dire, l’évènement (décès) n’a pas été observé pendant la durée de suivi indiqué dans « os\_months ».  1 – l’évènement a été observé après la durée indiqué dans « os\_months ». |

## Annotations du GEC

Les annotations du GEC pour les échantillons du cancer du côlon sont disponibles dans les fichiers \*\_GEC.csv, individuellement pour chaque jeu de données.



Exemple pour TCGA-COAD-FPKM\_GEC.csv :



**id\_sample** : l’identifiant des échantillons du cancer du côlon

**[ERFE, HOXC6, LAMP5, ULBP2]** : 0 – non activé, 1 – activé dans l’échantillon

**GEC** : le score GEC final, de 0 à 4, égal à la somme ERFE + HOXC6 + LAMP5 + ULBP2

# Méthodologie

## Identifier les échantillons du cancer du côlon (T-Colon) dans les données

La première tâche à faire consiste à identifier les échantillons du cancer du côlon dans les données d’expression et dans les annotations cliniques. Les fichiers d’origine EpiMed\_experimental\_grouping\*.xlsx et expression\_data\*.csv ne sont pas nettoyés. Ils contiennent différents types échantillons, normaux et tumoraux, certains échantillons ne correspondent pas au tissu côlon, pour d’autres il n’existe pas de données d’expression.

Tout d’abord, il faut retenir uniquement les échantillons du cancer du côlon pour lesquels les données d’expression sont disponibles.

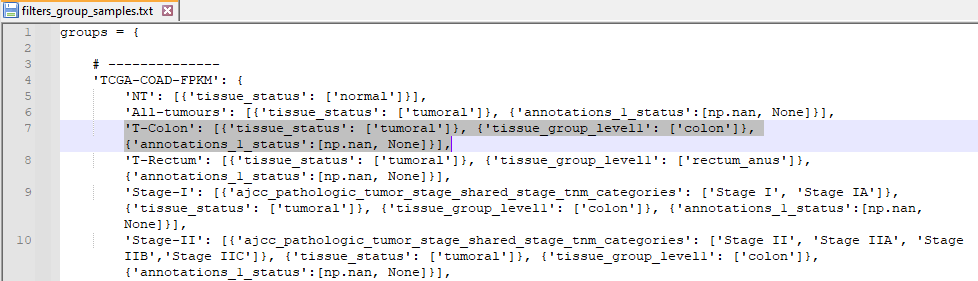
Il y a deux façons de le faire :

1) Le plus simples est d’importer les groupes d’échantillons déjà définis dans le fichier project\_transcriptome.json. Le nom du groupe qui contient tous les échantillons du cancer du côlon s’appelle « T-Colon » (tumor colon).



Le fichier contient également d’autres groupes qui peuvent être utilisés plus tard dans le projet.

2) Les échantillons du groupe « T-Colon », listés ci-dessus, ont été initialement identifiés à partir des fichiers d’origine (EpiMed\_experimental\_grouping\*.xlsx et expression\_data\*.csv) en appliquant les filtres définis dans filters\_group\_samples.txt. Il est possible de réappliquer ces filtres pour définir les groupes. On doit normalement obtenir la même liste d’échantillons.



Par exemple, le filtre pour le groupe T-Colon impose les conditions suivantes : la colonne « tissus status » = « tumoral » ET la colonne « tissue\_group\_level1 » = « colon » ET la colonne « annotations\_1\_status » est vide dans le fichier EpiMed\_experimental\_grouping\*.xlsx. En plus, les échantillons correspondants doivent être disponibles dans le fichier de données d’expression expression\_data\*.csv.

Vous devez normalement obtenir les effectifs suivants :

|  |  |
| --- | --- |
| **Dataset** | **T-Colon (sample size)** |
| TCGA-COAD | 398 |
| GSE39582 | 566 |
| GSE17536 | 177 |

## Construire des courbes de survie pour le Gene Expression Classifier (GEC)

Faire séparément pour les 3 datasets.

Exemple pour le dataset TCGA-COAD :

1) Utiliser le fichier d’annotation GEC pour savoir quel score GEC correspond à quel échantillon.



2) Utiliser le fichier d’annotations cliniques pour extraire les données de survie pour les mêmes échantillons. Prendre les colonnes « os\_months » et « os\_censor » dans le premier onglet de ce fichier.



3) Recalculer les données de survie, en les arrêtant à 120 mois. On considère qu’à partir de 120 mois (10 ans) les données de survie sont moins fiables. Si le patient décède 10 ans après son cancer il est possible que la cause de décès ne soit pas son cancer mais une autre maladie ou une autre cause (accident).

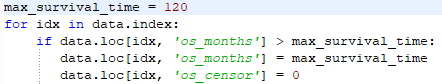
Voici l’algorithme :

- Définir une variable 

- Pour chaque échantillon, regarder le « os\_months ».

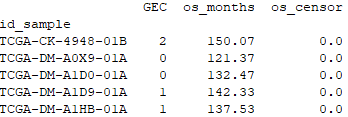
- Si os\_months > max\_survival\_time, alors os\_months = max\_survival\_time ET os\_censor = 0

En Python pandas :

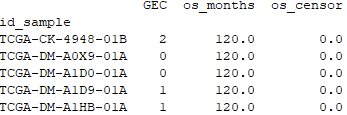


Résultat :

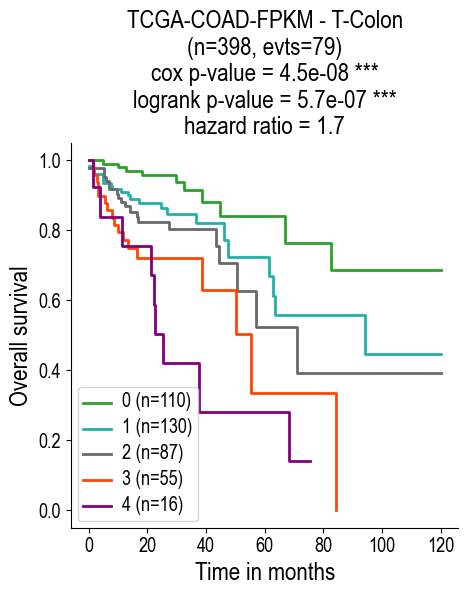
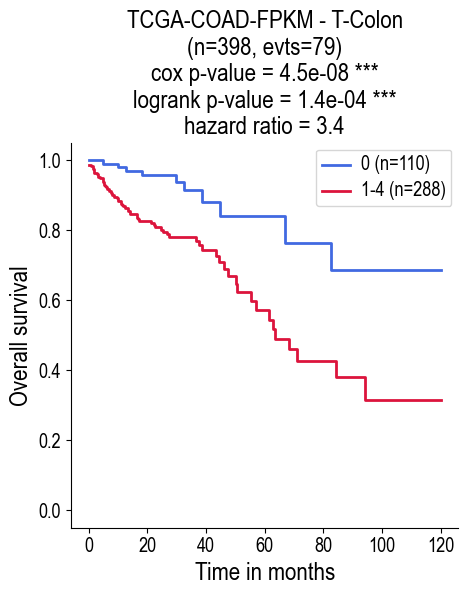
Avant



Après



4) Tracer la courbe de Kaplan-Meier pour les différents groupes GEC. Utiliser, par exemple, le package Python « lifelines » (voir KaplanMeierFitter).

5) Les p-valeurs peuvent être calculées avec le test statistique du logrank (voir multivariate\_logrank\_test dans « lifelines ») et avec le modèle de Cox (voir CoxPHFitter dans « lifelines").

# Tâches à réliser pour le projet tutoré

## Calculer l’impact pronostique du panel de 3 gènes GEC (3-GEC)

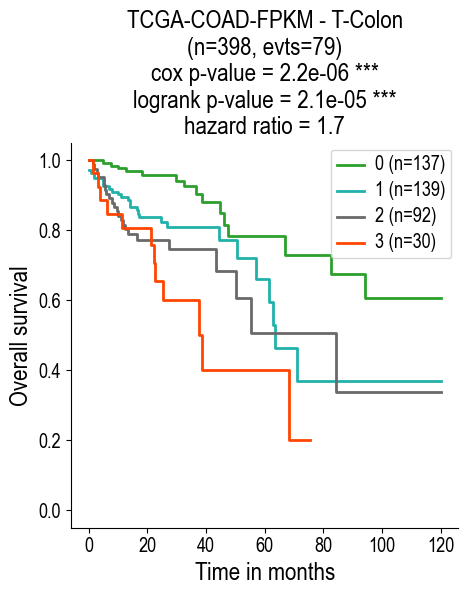
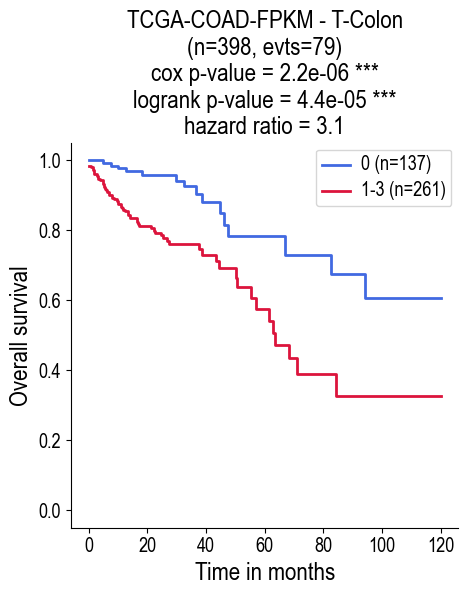
Le panel initial GEC est composé de 4 gènes (4-GEC) : ERFE, HOXC6, LAMP5, ULBP2.

Les médecins souhaitent développer un test pronostique simplifié basé sur la signature GEC, utilisable en routine dans les hôpitaux. Pour développer avec un tel test en immunohistochimie (technologie déjà disponible dans les hôpitaux), il est nécessaire de posséder des anticorps pour les 4 gènes ERFE, HOXC6, LAMP5, ULBP2. Or, il se trouve que les anticorps industriels existent seulement pour 3 gènes : HOXC6, ULPB2 et LAMP5.

La question est de savoir si la signature pronostique initiale 4-GEC peut être réduite à ces 3 gènes 3-GEC : HOXC6, ULPB2 et LAMP5. Est-ce que l’impact sur la survie reste significatif si on utilise 3 gènes au lieu de 4.

Pour répondre à cette question, il faut régénérer les courbes de survie pour la signature 3-GEC dans tous les datasets et calculer les p-valeurs correspondantes.

Exemple pour le dataset TCGA-COAD-FPKM :

## Identifier le statut MSI/MSS pour les échantillons du TCGA-COAD-FPKM

MSI = microsattelite instability

MSS = microsattelite stability

Le statut de l’échantillon tumoral est MSI si :

* Soit la région promotrice du gène MLH1 est méthylée et son expression est faible.
* Soit l’un des gènes MLH1, MSH2, MSH6, PMS2 a une mutation qui empêche sa fonction biologique normale. On dit que la mutation est de type « perte de fonction » ou « loss of function » en anglais ou « LOF ».
* Soit une région entière de l’ADN qui contient l’un des gènes MLH1, MSH2, MSH6, PMS2 est perdue. On appelle cela en anglais « somatic deletion » ou « deleterious copy number alteration » ou « deleterious CNA ».

Autrement, le statut de l’échantillon est MSS.

Pour identifier le statut MSI/MSS de tous les échantillons du dataset TCGA-COAD-FPKM, vous devez analyser les trois causes possibles, en utilisant trois types de données (méthylation, mutation et copy number alteration) pour chaque échantillon.

A l’issue des trois analyses, vous pouvez, par exemple, présenter les résultats dans un tableau comme ceci :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample of the dataset TCGA-COAD-FPKM | MLH1 metylated | Mutation LOF | Deleterious CNA | Status MSI/MSS |
| Sample 1 | no | MSH2, PMS2 | no | MSI |
| Sample 2 | no |  | no | MSS |
| Sample 3 | yes |  | no | MSI |
| Sample 4 | no |  | yes | MSI |
| Sample 5 | NO DATA | NO DATA | NO DATA | Unknown |

### Analyse des données de méthylation

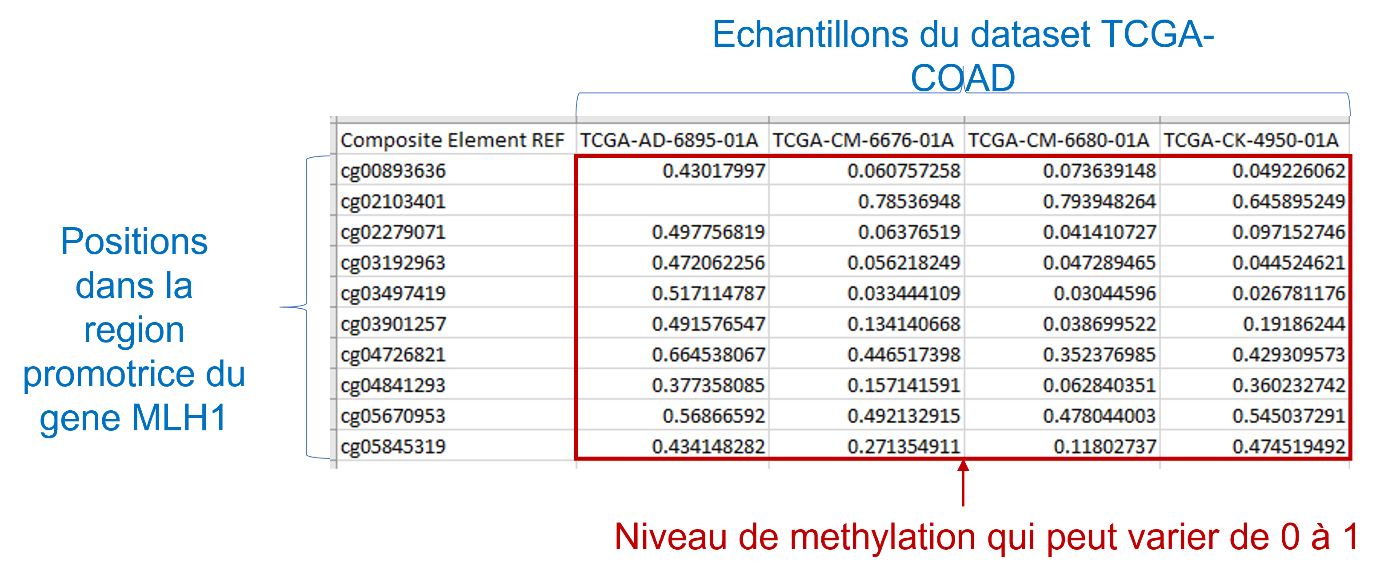
Cette analyse est à faire uniquement pour le gène MLH1.

Les données de méthylation sont disponibles à l’adresse :

http://epimed.univ-grenoble-alpes.fr/downloads/uga\_ssd/data/methylation\_data/



Il faut télécharger le fichier « methylation\_data\_TCGA-COAD\_MLH1.csv ». Il contient les niveaux de méthylation pour le gène MLH1 uniquement. Ce gène possède plusieurs positions dans sa région promotrice qui peuvent être méthylées. Ces positions sont annotées « cg[numéro] » dans le fichier. Nous allons nous intéresser uniquement à la méthylation moyenne a travers toutes ces positions dans chaque échantillon.



Position in the promoter region of the MLH1 gene

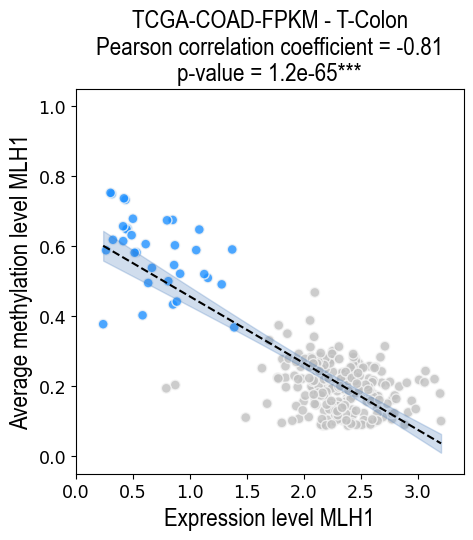
TCGA-COAD Dataset samples

**Methylation level (from 0 to 1)**

Le calcul se fait de la manière suivante :

* Prendre les échantillons tumoraux du groupe « T-Colon » uniquement. Pour certains échantillons « T-Colon », il n’y a pas de données de méthylation, c’est normal. Restreindre l’analyse aux échantillons disponibles.
* Pour chaque échantillon tumoral, calculer **le niveau moyen de méthylation** à travers toutes les positions « cg ». Cela revient à calculer la moyenne pour chaque colonne dans le tableau de données.
* Calculer une corrélation entre le niveau moyen de méthylation pour le gène MLH1 et son niveau d’expression (voir MLH1 dans les données d’expression de gènes). Calculer le coefficient de corrélation de Pearson et la p-valeur associée. Vous pouvez utiliser la fonction « pearsonr » dans « scipy.stats » de Python.
* Identifier deux clusters (à l’œil c’est OK dans notre cas) d’échantillons avec une forte méthylation et une faible expression, et avec une faible méthylation et une forte expression.

Exemple :



* Les échantillons avec une forte méthylation et une faible expression sont MSI (en bleu sur la figure). On peut considérer que le critère de sélection des échantillons MSI est la suivante : méthylation>0.3, expression<1.5.

A titre d’information, dans mon exemple, j’ai obtenu 34 échantillons MSI sur la totalité de 271 échantillons dans le dataset TCGA-COAD-FPKM dans le groupe « T-Colon » pour lesquels les données de méthylation et d’expression étaient disponibles simultanément. Cela fait 12.5% d’échantillons MSI par le gène MLH1 méthylé.

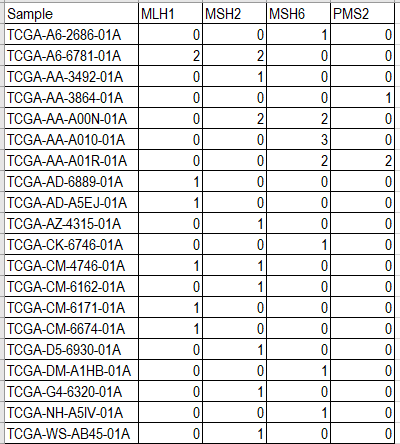
### Analyse des données de mutation

Les données de mutation sont disponibles à l’adresse :

http://epimed.univ-grenoble-alpes.fr/downloads/uga\_ssd/data/mutation\_data/



Le fichier indique le nombre de mutations LOF dans les gènes MLH1, MSH2, MSH6, PMS2 pour chaque échantillon. Les informations sur les mutations ne sont pas disponibles pour tous les échantillons « T-Colon », c’est normal.

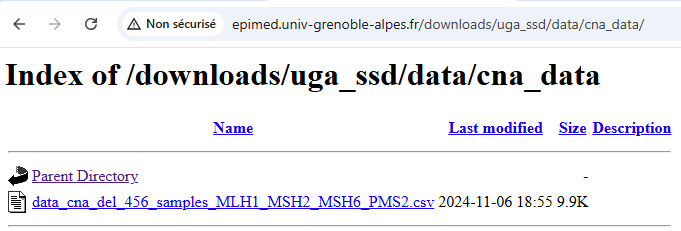


Le chiffre indique le nombre de mutations de chaque gène. Si au moins une mutation est détectée pour l’un des gènes MLH1, MSH2, MSH6, PMS2, alors l’échantillon est MSI.

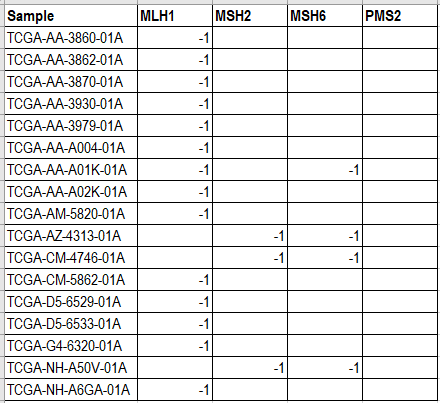
### Analyse des données de CNA

Les données de mutation sont disponibles à l’adresse :

http://epimed.univ-grenoble-alpes.fr/downloads/uga\_ssd/data/cna\_data/



Dans les données, le chiffre « -1 » ou « -2 » indique une délétion CNA.



On va considérer que les échantillons pour lesquels il y a « -1 » ou « -2 » pour l’un des gènes MLH1, MSH2, MSH6, PMS2 sont des échantillons MSI.

**Combien d’échantillons MSI, MSS et « unknown » vous trouvez dans le dataset TCGA-COAD-FPKM ?**

## Créer un modèle d’apprentissage pour le statut MSI/MSS à partir des données d’expression de gènes TCGA-COAD-FPKM

A décrire

## Identifier le statut MSI/MSS dans les deux datasets restants GSE39582 et GSE17536

A décrire

## Réaliser une analyse de survie multivariée pour 4-GEC et 3-GEC en intégrant le statut MSI/MSS

A décrire

## Réaliser une analyse de survie multivariée pour 4-GEC et 3-GEC en intégrant d’autres facteurs de risque connus

A décrire