*Cahier des charges*

Biomarqueurs pronostiques dans le cancer du côlon

Ekaterina Flin

21/10/2024

**Attention !**

**Les figures dans ce document sont données à titre d’exemple pour illustrer une représentation possible des résultats. Ces figures peuvent venir d’autres projets, ce ne sont pas de « vrais » résultats attendus.**

Table des matières

[Données 2](#_Toc180502281)

[Données d’expression de gènes 2](#_Toc180502282)

[Annotations biocliniques 3](#_Toc180502283)

[Annotations du GEC 3](#_Toc180502284)

[Méthodologie 4](#_Toc180502285)

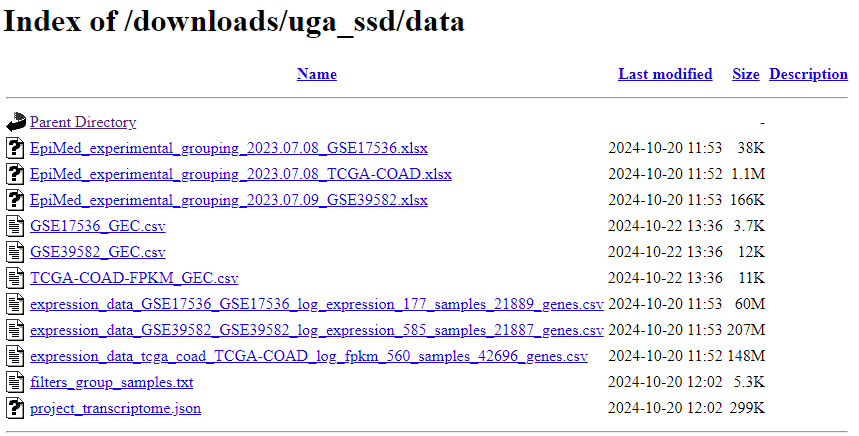
[Identifier les échantillons du cancer du côlon (T-Colon) dans les données 4](#_Toc180502286)

[Construire des courbes de survie pour le Gene Expression Classifier (GEC) 6](#_Toc180502287)

# 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 |

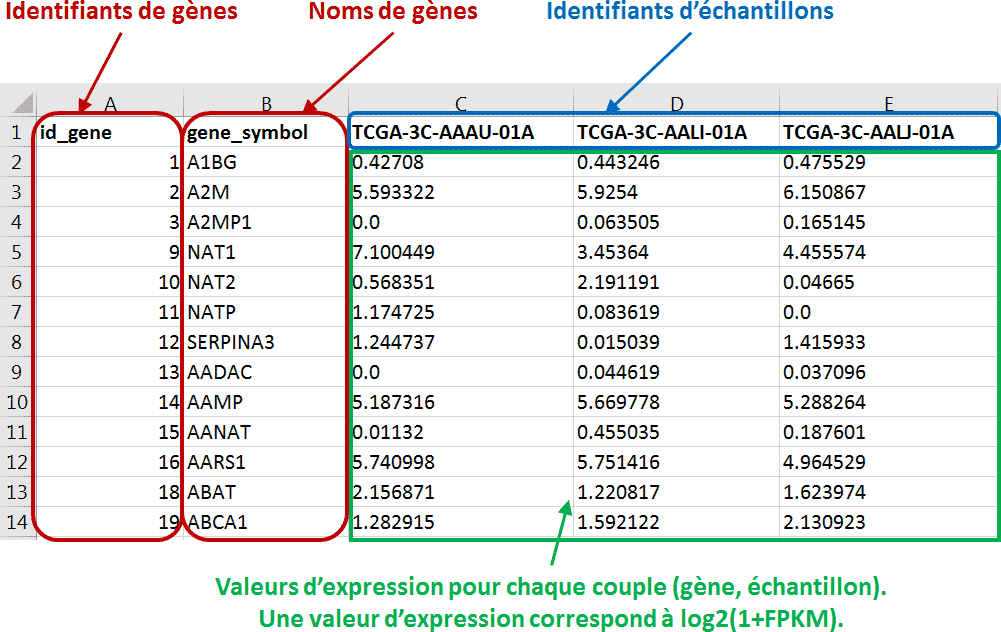
## Données d’expression de gènes

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 ».

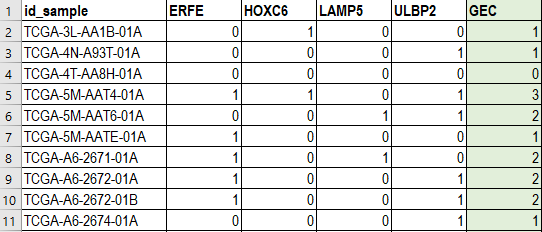
|  |  |
| --- | --- |
| **Colonne** | **Signification** |
| os\_month | La durée de la survie globale OS (overall survival) en mois. |
| os\_censor | La censure de la survie globale.  0 – donnée censurée, c’est-à-dire, l’évènement (décès) n’a pas été observé pendant la durée de suivi indiqué dans « os\_months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

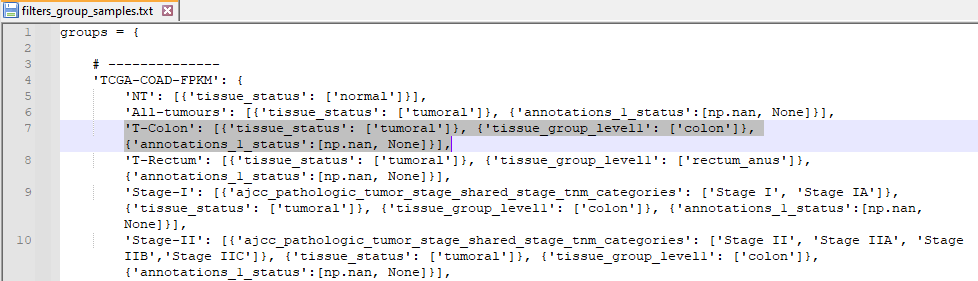
Il y a deux façons de le faire :

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).

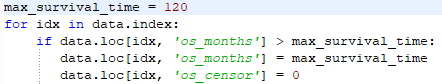
Voici l’algorithme :

- Définir une variable 

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

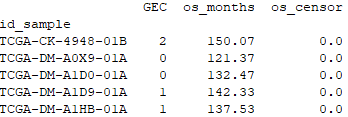
- Si os\_months > max\_survival\_time, alors os\_months = max\_survival\_time ET os\_censor = 0

En Python pandas :

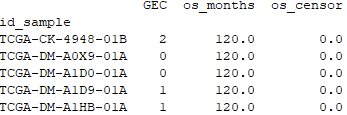


Résultat :

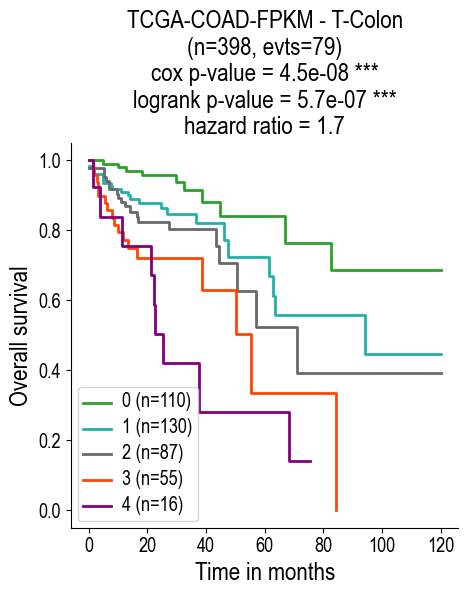
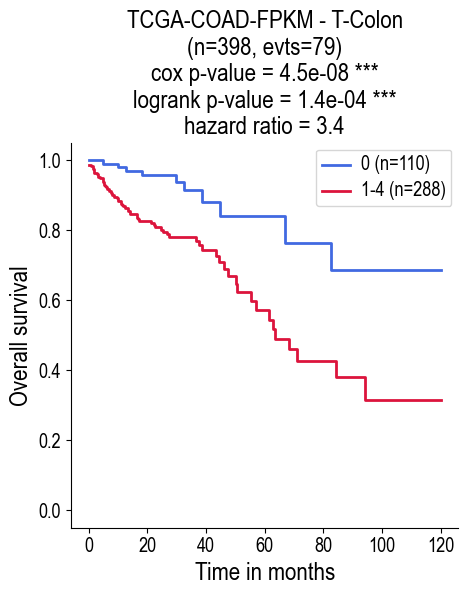
Avant



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").