

PROJET TER PAR L'ÉQUIPE 5

PROJET TER PAR L'ÉQUIPE 5 RENDU
INTERMÉDIAIRE 3 : LES DONNÉES ET LEUR
EXPLORATION

Thomas AYRIVIÉ, Mehdi BELKHITER, Jamila CHERKAoui, Dina EL
HIJJAWI, Magatte LO.



Département MIASHS, UFR 6 Informatique, Mathématique et Statistique
Université Paul Valéry, Montpellier 3

Novembre 2023

SOUmis COMME CONTRIBUTION PARTIELLE
POUR LE COURS TER TV15MI - TV25MI

Déclaration de non plagiat

Nous déclarons que ce rapport est le fruit de notre seul travail, à part lorsque cela est indiqué explicitement.

Nous acceptons que la personne évaluant ce rapport puisse, pour les besoins de cette évaluation:

- la reproduire et en fournir une copie à un autre membre de l'université; et/ou,
- en communiquer une copie à un service en ligne de détection de plagiat (qui pourra en retenir une copie pour les besoins d'évaluation future).

Nous certifions que nous avons lu et compris les règles ci-dessus.

En signant cette déclaration, nous acceptons ce qui précède.

Signature : **Thomas AYRIVIÉ**, n°22000580, Date : 17/11/2023.

Signature : **Mehdi BELKHITER**, n°21813356, Date : 17/11/2023.

Signature : **Jamila CHERKAoui**, n°22309204, Date : 17/11/2023.

Signature : **Dina EL HIJJAWI**, n°22310171, Date : 17/11/2023.

Signature : **Magatte LO**, n°22311161, Date : 17/11/2023.

Préface

Dans le cadre de notre master 1 Miashs, les étudiants réalisent un **projet de Travaux d'Études et de Recherche** (TER) en lien avec un commanditaire. Ce présent document est un rapport intermédiaire dans lequel nous présenterons la définition de notre projet ainsi que les données utilisées et leur exploration.

Sujet 6 : « Identification de voies de signalisation activées par des traitements de cancer du sein ».

Encadrement pédagogique :

Mme Sophie Lèbre, sophie.lebre@univ-montp3.fr (Institut Montpelliérain Alexander Grothendieck) avec Mathilde Robin (Institut de Recherche en Cancérologie de Montpellier, LIRMM Laboratoire d'informatique, de robotique et de microélectronique de Montpellier), Charles Lecellier (LIRMM) et Laurent Bréhélin (LIRMM).

Composition du groupe :

Dina EL HIJJAWI, n°22310171, dina.el-hijjawi@etu.univ-montp3.fr. Coordinatrice.

Thomas AYRIVIE, n°22000580, thomas.ayrivie@etu.univ-montp3.fr.

Mehdi BELKHITER, n°21813356, mehidi.belkhiter@etu.univ-montp3.fr.

Jamila CHERKAOUI, n°22309204, jamila.cherkaoui@etu.univ-montp3.fr.

Magatte LO, n°22311161, magatte.lo@etu.univ-montp3.fr.

Table des matières

Chapitre 1	Sources des données	1
Chapitre 2	Création de notre jeu de données et traitements	2
Chapitre 3	Résumés graphiques et numériques:	3
Chapitre 4	Annexe	11

CHAPITRE 1

Sources des données

Pour mener à bien notre recherche, nous utilisons la base de données Jaspar¹, une ressource spécialisée dans les matrices de séquences d'ADN liées aux sites de liaison des facteurs de transcription. Ces matrices, également appelées matrices de poids de position (PWM), offrent une représentation numérique des motifs de liaison des facteurs de transcription qui sont des protéines codées par des gènes ADN qui sont responsables de l'activation ou l'inhibition d'autres gènes. Jaspar fournit des informations détaillées sur la fréquence de chaque base (A, C, G, T)², à chaque position au sein de ces motifs de liaison. Cette base de données est essentielle pour développer des modèles plus précis dans le but de mieux comprendre la régulation génique et les réponses cellulaires aux traitements du cancer du sein et elle est aussi essentielle pour identifier des sites de liaisons potentiels au sein des séquences et construire un ensemble de variables explicatives à partir des données existantes afin d'améliorer la performance et la capacité prédictive du modèle.

En complément, nous utilisons pareillement la base de données Series GSE130787 GEO³ (Gene Expression Omnibus) qui est chargée dans RStudio grâce à GEOquery⁴. GEO stocke une grande variété de données d'expression génique provenant de diverses expériences et technologies, incluant les puces à ADN et les séquences ARN. Les chercheurs déposent leurs données d'expression génique dans GEO, permettant ainsi à d'autres scientifiques d'y accéder, de les partager et de les analyser. Cette base de données facilite l'exploration et la comparaison des profils d'expression génique dans différents contextes biologiques et expérimentaux.

Jaspar nous aide à analyser les motifs de liaison des facteurs de transcription, tandis que Geoquery nous permet d'accéder à une vaste gamme de données d'expression génique pour une exploration approfondie des mécanismes biologiques et des réponses aux traitements.

¹Wasserman Lab <http://www.cmmt.ubc.ca/wasserman-lab/>

²Les bases azotées symbolisées par A, C, G, T renvoient dans le mémoire à A pour Adénine, C pour Cytosine, G pour Guanine, T pour Thymine.

³Hsiaowang Chen, Series GSE130787 de NCBI. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130787>

⁴Davis S, Meltzer P (2007). "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." *Bioinformatics*, 14, 1846–1847. <https://bioconductor.org/packages/release/bioc/html/GEOquery.html>

CHAPITRE 2

Création de notre jeu de données et traitements

La population utilisée est des patientes atteintes de cancer du sein. Les unités statistiques utilisées sont les p-valeurs des motifs et séquences ADN.

Un fichier FASTA est un format de fichier texte couramment utilisé pour représenter des séquences d'acides nucléiques (comme l'ADN ou l'ARN) ou des séquences de protéines. Chaque fichier FASTA est composé d'une description de la composition et de la structure. Ligne d'En-tête : Chaque séquence dans un fichier FASTA commence par une ligne d'en-tête, qui est précédée du symbole ">". Cette ligne d'en-tête est souvent utilisée pour stocker des identifiants et/ou des descriptions. Notre fichier contient 23557 lignes composées d'un identifiant suivi d'une séquence.

JASPAR est une base de données de référence pour les motifs de liaison des facteurs de transcription (TF). Elle fournit des matrices de position de poids (PWM) ou des profils qui représentent les motifs de séquence préférés des facteurs de transcription. Notre fichier contient 1205 motifs.

Matrices de Position de Poids (PWM) : Les PWM dans JASPAR sont des représentations mathématiques des motifs de liaison des facteurs de transcription. Chaque colonne de la matrice correspond à un site de la séquence (par exemple, une base d'ADN) et chaque ligne représente un nucléotide (A, T, C, G). Les valeurs dans la matrice indiquent la probabilité relative de chaque nucléotide à chaque position du motif.

Utilisation en Bioinformatique : Les PWM de JASPAR sont largement utilisés pour identifier et prédire les sites de liaison des facteurs de transcription dans les séquences génomiques. Elles sont utilisées dans diverses analyses bioinformatiques, notamment l'analyse de l'expression génique, la régulation génétique et l'annotation des génomes.

À partir de là, nous faisons une matrice de correspondance des p-valeurs entre les noms de séquences et les noms de motifs. Les p-valeurs nulles sont remplacées par 2×10^{-3} . Il est à noter que la jointure a supprimé des données. Notre matrice Y contient 89 patientes en colonnes et 17014 séquences en lignes.

CHAPITRE 3

Résumés graphiques et numériques:

Initialement après avoir réalisé la Jointure, certaines colonnes présentaient des valeurs manquantes (figure3.1). Pour remédier à cela , nous avons substitué ces valeurs manquantes par 2×10^{-3} . Ci-dessous, vous trouverez un aperçu de notre base de données finale(figure3.2).

sequence_name	MA0002.1::RUNX1	MA0002.2::Runx1	MA0003.1::TFAP2A	MA0003.2::TFAP2A	MA0003.3::TFAP2A
ENSG00000000003	NULL	7.13E-4	NULL	1.78E-4	NULL
ENSG00000000457	3.63E-4	3.093E-4	8.65E-4	NULL	1.585E-
ENSG00000000460	6.299999999999999E-4	6.291333333333333E-4	NULL	5.826666666666667E-4	NULL
ENSG00000000971	3.471333333333333E-4	2.604200000000000...	NULL	NULL	NULL
ENSG00000001461	7.36E-4	1.903333333333333...	4.593333333333333...	5.679666666666667E-4	5.332E-
ENSG00000001561	6.724999999999999E-4	5.82E-4	3.320125E-4	4.166238888888889E-4	4.703999999999999...
ENSG00000001617	8.41E-5	3.18675E-4	9.33E-4	4.250000000000000...	1.585E-
ENSG00000001629	1.93E-4	6.478E-4	9.915E-5	3.85E-5	3.781E-
ENSG00000002587	5.62E-4	3.7825E-4	8.37E-4	7.920000000000001E-4	8.41E-
ENSG00000002822	7.635000000000001E-4	5.05E-4	7.325000000000001E-4	3.05E-4	9.11E-
ENSG00000002919	1.26E-4	2.420033333333333...	4.98E-4	5.22E-4	5.5775E-
ENSG00000003147	4.120000000000000...	2.735E-4	5.16E-4	7.143333333333333E-4	2.9475E-
ENSG00000003402	NULL	5.836666666666667E-4	6.665E-4	2.39645E-4	6.5225E-
ENSG00000003436	3.81E-4	3.683333333333333...	NULL	7.09E-4	NULL
ENSG00000003509	5.3275E-4	4.052E-4	NULL	NULL	NULL
ENSG00000003987	1.84E-4	4.680000000000000...	5.52E-4	NULL	4.430000000000000...
ENSG00000003989	7.66E-4	NULL	6.774999999999999E-4	2.379999999999999...	6.84E-
ENSG00000004059	8.405E-4	3.605000000000000...	3.81E-4	NULL	8.67E-
ENSG00000004142	NULL	9.6E-4	7.09E-4	NULL	NULL
ENSG00000004478	1.525000000000000...	8.53E-4	NULL	NULL	7.83E-

only showing top 20 rows

Figure 3.1: Jeu de données avec les NAN

sequence_name	MA0002.1::RUNX1	MA0002.2::Runx1	MA0003.1::TFAP2A	MA0003.2::TFAP2A	MA0003.3::TFAP2A	MA0003.4::TFAP2A	MA0003.5::TFAP2A
ENSG00000000003	2.6989700043360187	3.1469104701481343	2.6989700043360187	3.7495799976911106	2.6989700043360187	2.6989700043360187	3.540607
ENSG00000000457	3.4400933749638876	3.509620079996821	3.0629838925351858	2.6989700043360187	2.6989700043360187	3.3851027839668655	2.6989700
ENSG00000000460	3.2006594505464183	3.2012573040071413	2.6989700043360187	3.234579826421278	2.6989700043360187	3.1313555616051745	2.6989700
ENSG00000000971	3.4595036813799899	3.5843256654337394	2.6989700043360187	2.6989700043360187	2.6989700043360187	2.6989700043360187	2.6989700
ENSG00000001461	3.133122185662501	3.7204851464738145	3.3378720371408552	3.2456771518022314	3.2731098592581878	3.51154329680891347	3.279840
ENSG00000001561	3.472307711325544	3.2350770153501114	3.47804556515491	3.380258315674	3.327532606931918	3.966110030390975	3.279840
ENSG00000001617	4.075204004202088	3.4966520055187855	3.0301183562535	3.3716110699496884	3.7099707334462296	3.5843590201038458	2.6989700
ENSG00000001629	3.714442690092226	3.188559056325842	4.003707281458679	4.414539270491499	4.422393322637465	3.6677635845085566	2.6989700
ENSG00000002587	3.250263684430939	4.222210633047756	3.07727454200674	3.1012748184105066	3.075204004202088	3.167810538931487	3.316052
ENSG00000002822	3.11719095860756	3.2967086218813386	3.135192370973853	3.515700160653214	3.0404816230270018	3.260902553882525	2.6989700
ENSG00000002919	3.899629454882437	3.61617865204311	3.3027706572402824	3.282329496997738	3.2535604210441247	3.708853282681145	2.6989700
ENSG00000003147	3.3851027839668655	3.5630426693305504	3.2873502983727887	3.146099083677834	3.5305461862328733	3.1677040289415226	3.540607
ENSG00000003402	2.6989700043360187	3.2338351086362165	3.176199846250122	3.6204316277824056	3.1855859122277415	3.6452514800439161	2.6989700
ENSG00000003436	3.4190750243243806	3.433758976698533	2.6989700043360187	3.1493537648169334	2.6989700043360187	3.5243288116755704	2.6989700
ENSG00000003509	3.2734765416137606	3.392330563311757	2.6989700043360187	2.6989700043360187	2.6989700043360187	2.6989700043360187	3.540607
ENSG00000003987	3.7351821769904636	3.329754146925876	3.258060922270801	2.6989700043360187	3.3535962737769305	3.4424927980943423	2.6989700
ENSG00000003989	3.115771230367396	2.6989700043360187	3.1690907004535567	3.623423042943488	3.164943898279884	3.329290404776203	3.279840
ENSG00000004059	3.075462282245103	3.4430947309445523	3.4190750243243806	2.6989700043360187	3.0619809025237896	2.6989700043360187	2.6989700
ENSG00000004142	2.6989700043360187	3.0177287669604316	3.1493537648169334	2.6989700043360187	2.6989700043360187	3.4485500020271247	2.6989700
ENSG00000004478	3.8167301563171954	3.0690509688324767	2.6989700043360187	2.6989700043360187	3.1062382379420566	2.6989700043360187	3.540607

only showing top 20 rows

Nombre total de valeurs NULL dans le DataFrame après traitement étendu : 0

Figure 3.2: Jeu de données final

La jointure entre les deux matrices a été réalisée en utilisant Spark en raison de la taille importante et du volume conséquent de la base de données.

```
[10] !pip install pyspark
      from pyspark.sql import SparkSession
      spark = SparkSession.builder.master("local[*]").getOrCreate()
```

Figure 3.3: Importation de Spark

```
# On souhaite faire la jointure avec le fichier txt
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, count, sum as sql_sum

#session Spark
spark = SparkSession.builder.master("local[*]").appName("Data_Join").getOrCreate()

#data frame précédent déjà charger

# Charger le fichier texte

df_txt = spark.read.csv('/content/drive/My Drive/exprs_data.txt', sep='\t', header=True, inferSchema=True)

|
df_joined = df.join(df_txt, df.sequence_name == df_txt.gene_id)

# Afficher le résultat pour vérification
df_joined.show()

### On a des valeurs nulles donc on doit les traiter
```

Figure 3.4: Code de la jointure entre les deux matrices

Voici aussi une représentation visuelle de la matrice Y regroupé en classe(up, down et neutre)

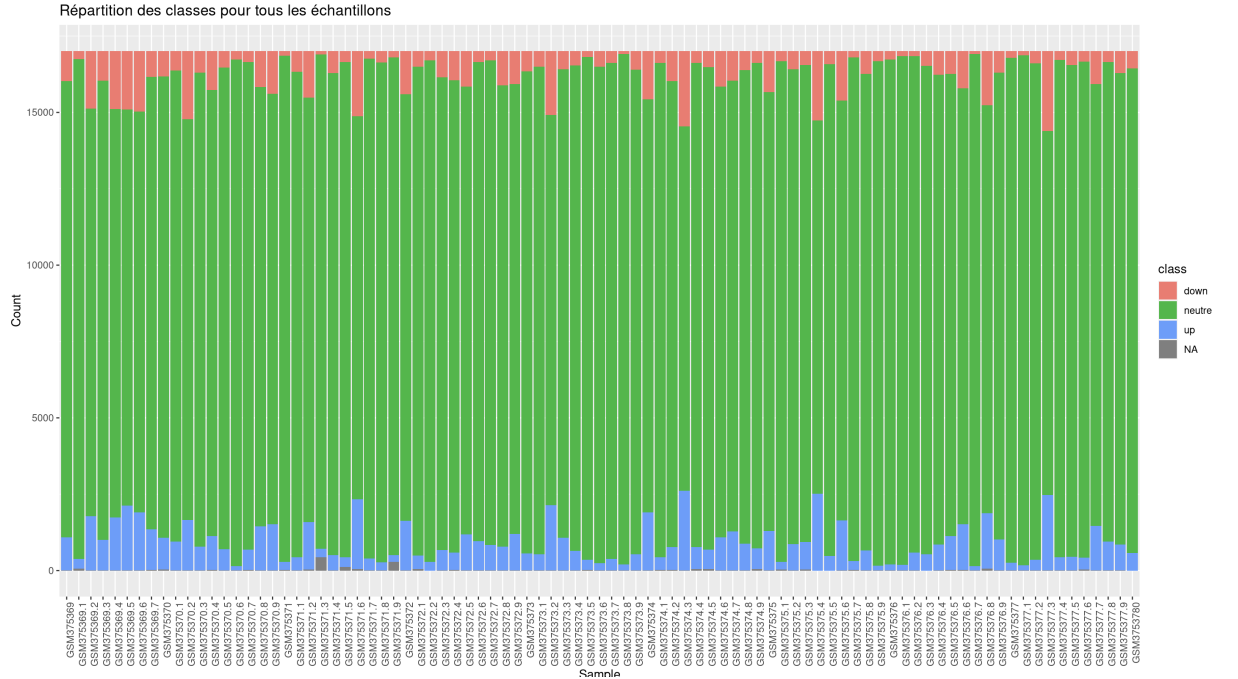


Figure 3.5: les classes de la matrice Y

Les barres vertes représentent le nombre de séquences d'ADN classées comme "neutres" pour chaque patiente. Cela signifie que le $\log_{10}(\text{fold change})$ de ces séquences se situe entre $-\log_{10}(2)$ et $\log_{10}(2)$, indiquant que le changement d'expression n'est pas assez significatif pour être considéré comme "up" ou "down".

Les barres rouges indiquent les séquences d'ADN classées comme "down" pour chaque patiente. Cela signifie que leur $\log_{10}(\text{fold change})$ est inférieur à $-\log_{10}(2)$, ce qui indique une diminution de l'expression d'au moins moitié par rapport à l'état de référence.

Les barres bleues montrent les séquences d'ADN classées comme "up". Cela indique que leur $\log_{10}(\text{fold change})$ est supérieur à $\log_{10}(2)$, signalant un doublement ou plus de l'expression par rapport à l'état de référence.

Pour clarifier, $\log_{10}(2)$ est approximativement 0.3010, donc les seuils utilisés ici pour classer les séquences sont :

"up" si le $\log_{10}(\text{fold change}) > 0.3010$

"down" si le $\log_{10}(\text{fold change}) < -0.3010$

"neutre" si $-0.3010 \leq \log_{10}(\text{fold change}) \leq 0.3010$

Chaque colonne du graphique représente une patiente différente, et la hauteur de chaque barre indique le nombre de séquences classées dans chaque catégorie pour cette patiente.

Pour approfondir notre compréhension des données, nous avons réalisé une ACP (Analyse en Composante Principale). Les résultats capturés dans le summary de l'ACP (figure 3.6) et le cercle de corrélation interactif (figure 3.7) ci-dessous offrent un aperçu visuel essentiel de la structure et des relations au sein de notre ensemble de données

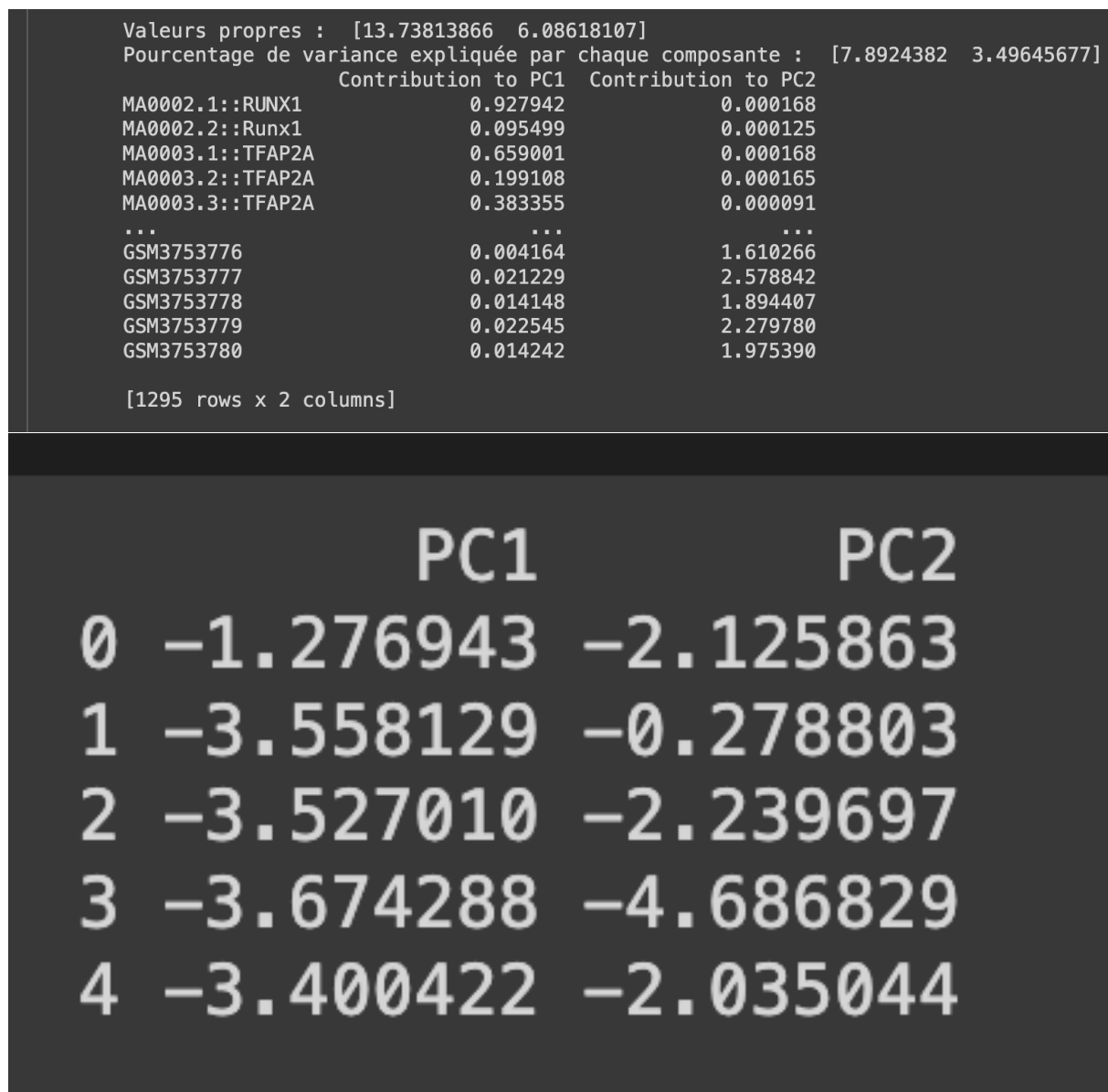


Figure 3.6: Aperçu du summary de l'acp

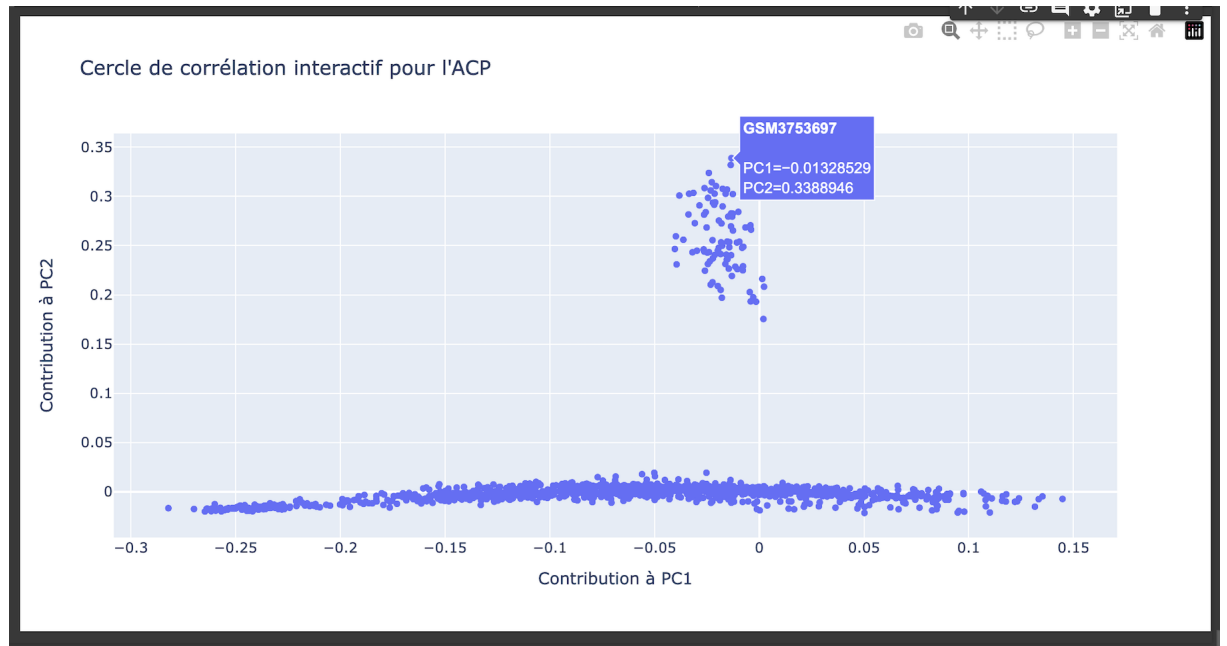


Figure 3.7: Cercle de corrélation interactif pour l'ACP

L'ACP réalisée a réduit la dimensionnalité de jeu de données en mettant en évidence les composantes principales qui expliquent la plus grande variance dans les données. Les premières composantes principales (PC1 et PC2 dans notre cas) capturent les aspects les plus significatifs des données. Les contributions de chaque variable aux composantes principales aident à comprendre quelles variables ont le plus d'impact sur ces composantes. une contribution élevée d'une variable à PC1 indique qu'elle joue un rôle important dans la variance capturée par cette composante.

Les valeurs propres indiquent la quantité de variance expliquée par chaque composante principale. la première composante principale (PC1) a une valeur propre de 13.738, et la deuxième (PC2) de 6.086. Cela suggère que PC1 explique une part significative de la variance dans les données, plus que PC2.

Pourcentage de variance expliquée : PC1 explique environ 7.89% de la variance totale, tandis que PC2 en explique environ 3.50%. Cela signifie que ces deux composantes, ensemble, capturent un peu plus de 11% de la variance totale des données.

Contributions des variables : Des variables telles que MA0002.1::RUNX1 ont une contribution élevée à PC1, ce qui signifie qu'elles influencent fortement cette composante. Cela peut indiquer que ces variables jouent un rôle crucial dans les variations observées dans l'ensemble des données.

Pour creuser un peu plus et mieux comprendre nos données, on a réalisé d'autres graphiques mais aussi des calculs descriptifs.

1. Analyse descriptive sur les variables

Pour une exploration approfondie de nos données, nous avons entrepris une analyse descriptive complète sur chaque variable. Les graphiques ci dessous illustrent visuellement les caractéristiques essentielles telles que la moyenne , l'écart type , le minimum ect. Ces représentations fournissent un aperçu détaillé de la distribution et de la variabilité de nos données. De plus on a fait des graphiques pour représenter la moyenne et l'écart type pour chaque variable pour fournir un aperçu clair des tendances centrales et de la dispersion au sein de nos données.

58 s	MA0002.1::RUNX1	MA0002.2::Runx1	MA0003.1::TFAP2A	MA0003.2::TFAP2A	\
count	17014.000000	17014.000000	17014.000000	17014.000000	
mean	3.203247	3.306100	3.226897	3.250650	
std	0.413615	0.307423	0.322542	0.295193	
min	2.698970	2.698970	2.698970	2.698970	
25%	2.698970	3.175417	3.117475	3.149354	
50%	3.216096	3.291296	3.280376	3.280973	
75%	3.404797	3.424235	3.399510	3.399027	
max	6.551294	6.424812	5.062984	6.241088	
	MA0003.3::TFAP2A	MA0003.4::TFAP2A	MA0004.1::Arnt	\	
count	17014.000000	17014.000000	17014.000000		
mean	3.221931	3.295051	2.963394		
std	0.298360	0.269682	0.308079		
min	2.698970	2.698970	2.698970		
25%	3.135192	3.196088	2.698970		
50%	3.249402	3.307237	2.698970		
75%	3.365859	3.424260	3.279841		
max	5.679854	5.612610	3.540608		
	MA0006.1::Ahr::Arnt	MA0007.1::Ar	MA0007.2::AR	...	GSM3753771
count	17014.000000	17014.000000	17014.000000	...	17014.000000
mean	2.989994	3.163837	3.287973	...	1.283484
std	0.305371	0.391715	0.340066	...	0.457047
min	2.698970	2.698970	2.698970	...	0.002000
25%	2.698970	2.698970	3.142179	...	0.985963
50%	3.034798	3.183096	3.281498	...	1.207854
75%	3.196543	3.367543	3.425969	...	1.490764
max	3.540608	6.248721	7.009217	...	4.698970
	GSM3753772	GSM3753773	GSM3753774	GSM3753775	GSM3753776
count	17014.000000	17014.000000	17014.000000	17014.000000	17014.000000
mean	1.135745	0.836701	1.244617	1.167807	1.162125
std	0.472822	0.522787	0.481807	0.490000	0.479592
min	-0.144241	-0.413965	-0.114434	-0.117914	-0.027023
25%	0.812634	0.468547	0.940796	0.839869	0.831797
50%	1.052640	0.744378	1.188157	1.091247	1.083678
75%	1.359693	1.117106	1.477556	1.412822	1.400444
max	5.170053	4.154902	5.440093	5.545155	5.060481
	GSM3753777	GSM3753778	GSM3753779	GSM3753780	
count	17014.000000	17014.000000	17014.000000	17014.000000	
mean	0.992521	1.101732	1.086015	1.111845	
std	0.510109	0.483849	0.491717	0.491864	
min	-0.190391	-0.213887	-0.236177	-0.114574	
25%	0.642165	0.784369	0.764427	0.781839	
50%	0.895069	1.032429	1.013923	1.027311	
75%	1.251890	1.337478	1.332734	1.353670	
max	5.344862	4.698970	4.301030	5.978811	
[8 rows x 1294 columns]					
Nombre de valeurs uniques par colonne					

Figure 3.8: Analyse descriptive sur les variables

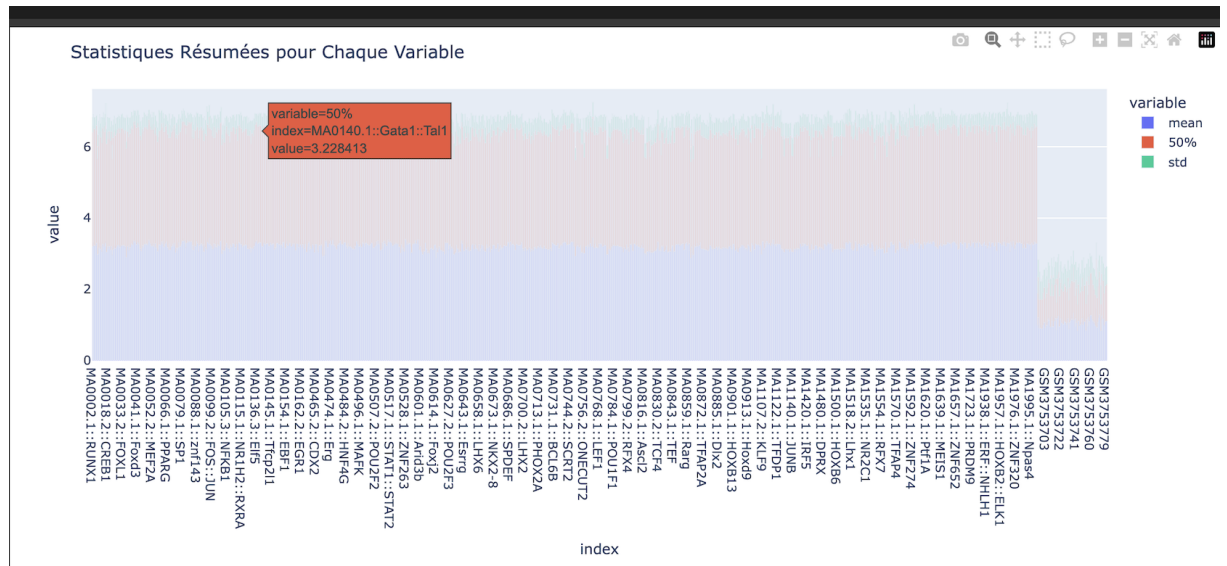


Figure 3.9: Statistiques résumées pour chaque variable

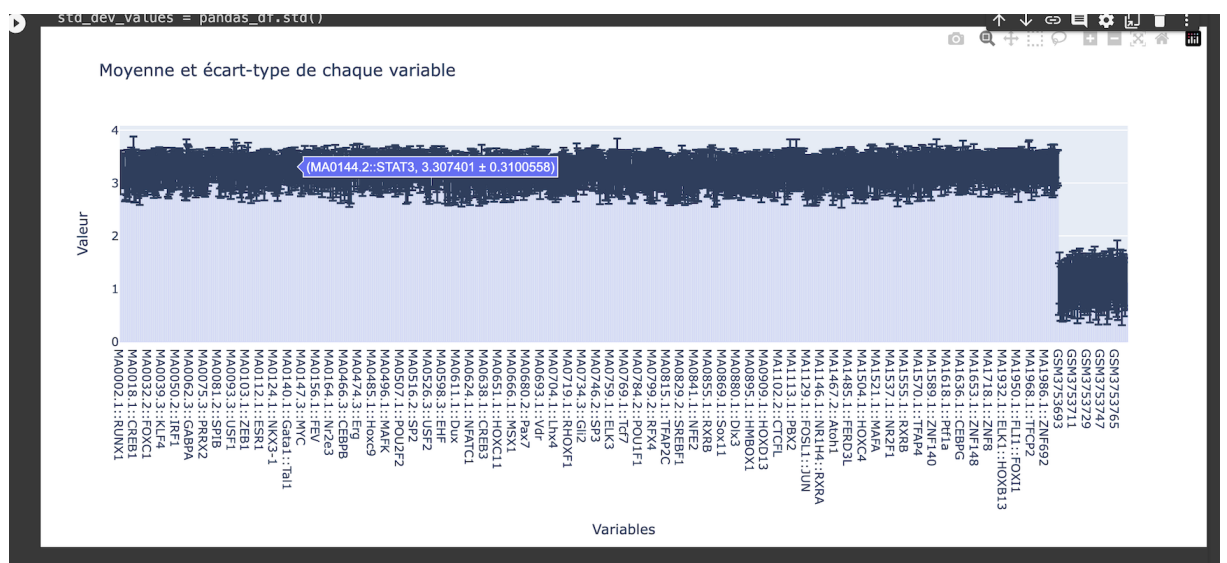
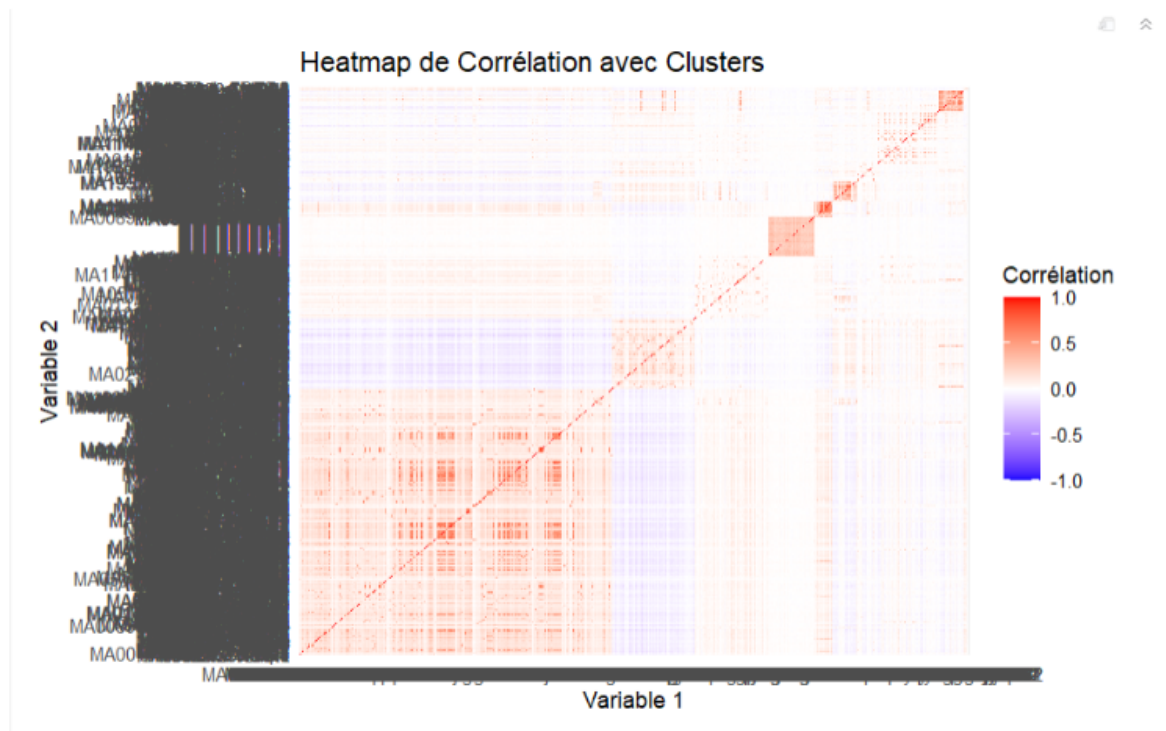


Figure 3.10: Graphe des moyennes et écart-type

2.Heatmap de corrélation



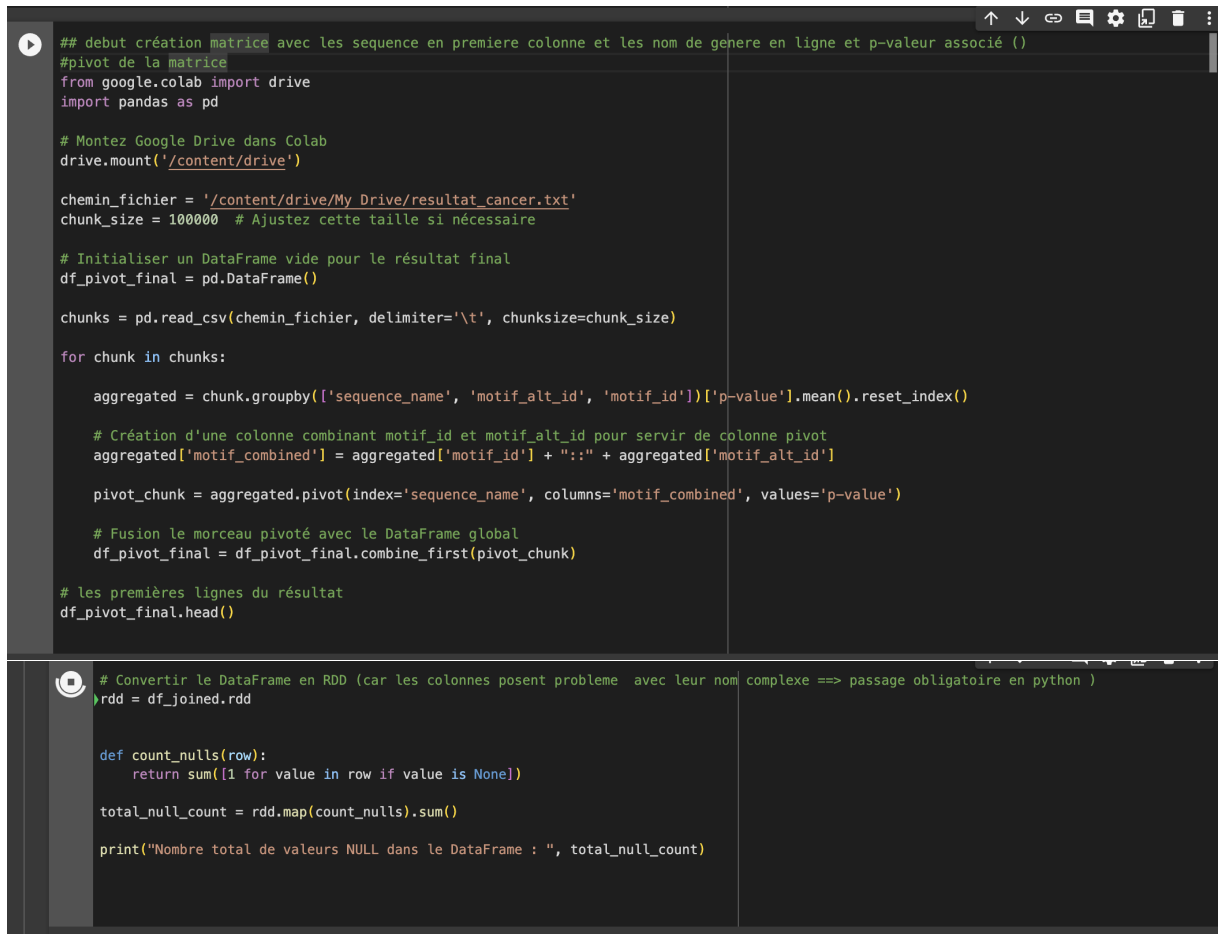
La heatmap de corrélation, qui est une représentation visuelle des coefficients de corrélation entre les variables est surchargée de données et donc difficile à interpréter.

La palette de couleurs passe du bleu (corrélation négative) au blanc (corrélation neutre) au rouge (corrélation positive). Les couleurs sont censées indiquer la force et la direction de la corrélation, mais avec autant de petites cellules, il est difficile de déterminer les valeurs exactes.

Cependant, on peut encore distinguer une concentration de rouge ce qui indique une forte corrélation entre les variables

CHAPITRE 4

Annexe



```
## debut création matrice avec les sequence en premiere colonne et les nom de genere en ligne et p-valeur associé ()
#pivot de la matrice
from google.colab import drive
import pandas as pd

# Montez Google Drive dans Colab
drive.mount('/content/drive')

chemin_fichier = '/content/drive/My Drive/resultat_cancer.txt'
chunk_size = 100000 # Ajustez cette taille si nécessaire

# Initialiser un DataFrame vide pour le résultat final
df_pivot_final = pd.DataFrame()

chunks = pd.read_csv(chemin_fichier, delimiter='\t', chunksize=chunk_size)

for chunk in chunks:

    aggregated = chunk.groupby(['sequence_name', 'motif_alt_id', 'motif_id'])['p-value'].mean().reset_index()

    # Création d'une colonne combinant motif_id et motif_alt_id pour servir de colonne pivot
    aggregated['motif_combined'] = aggregated['motif_id'] + ":" + aggregated['motif_alt_id']

    pivot_chunk = aggregated.pivot(index='sequence_name', columns='motif_combined', values='p-value')

    # Fusion le morceau pivoté avec le DataFrame global
    df_pivot_final = df_pivot_final.combine_first(pivot_chunk)

# les premières lignes du résultat
df_pivot_final.head()
```

```
# Convertir le DataFrame en RDD (car les colonnes posent probleme avec leur nom complexe ==> passage obligatoire en python )
rdd = df_pivot_final.to_spark(rdd)

def count_nulls(row):
    return sum([1 for value in row if value is None])

total_null_count = rdd.map(count_nulls).sum()

print("Nombre total de valeurs NULL dans le DataFrame : ", total_null_count)
```

Figure 4.1: Création de la matrice X