Breast Cancer Treatment Signaling Pathways

El Hijjawi, Cherkaoui, Lo, Belkhiter, Ayrivié (**Group 5).**
+ Sophie Lèbre (UPVM & IMAG).

Institut de Recherche en Cancérologie
I R C Montpellier

INTRODUCTION

- Understanding Cancer Resistance.
- Breast Cancer Research at Montpellier.
- Signaling Pathways and Resistance Mechanisms.
- Gene Analysis and Feature Selection.

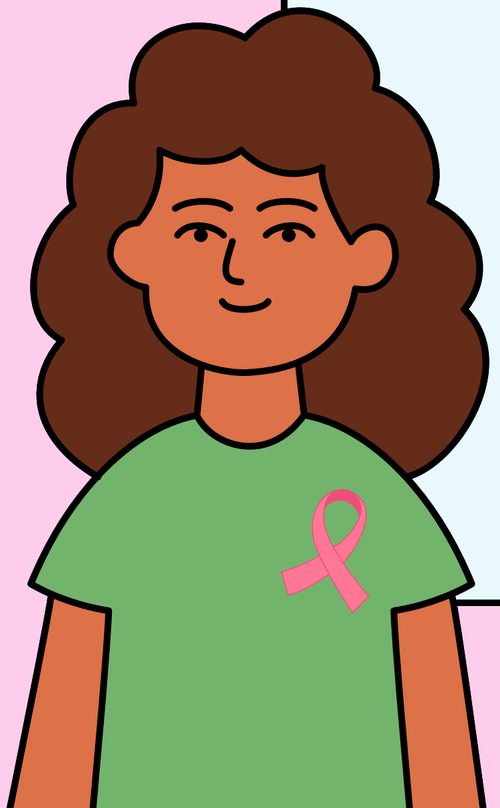
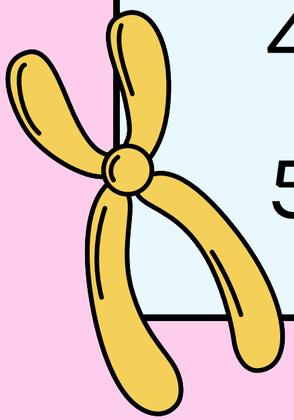
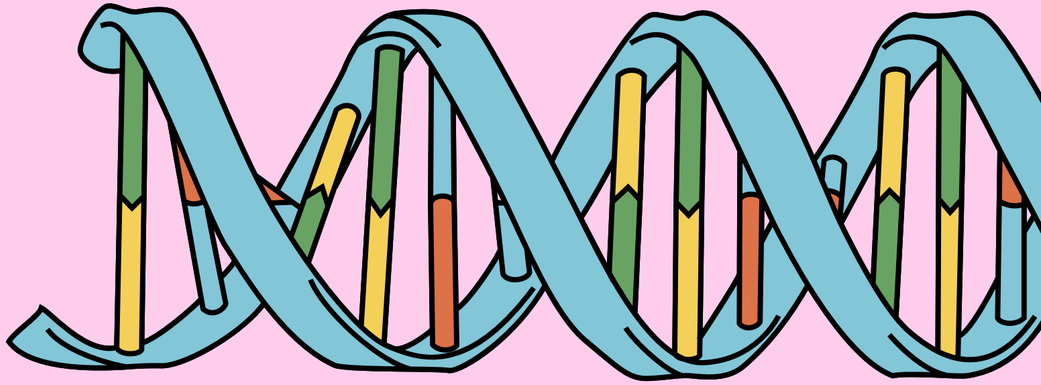


Table of contents

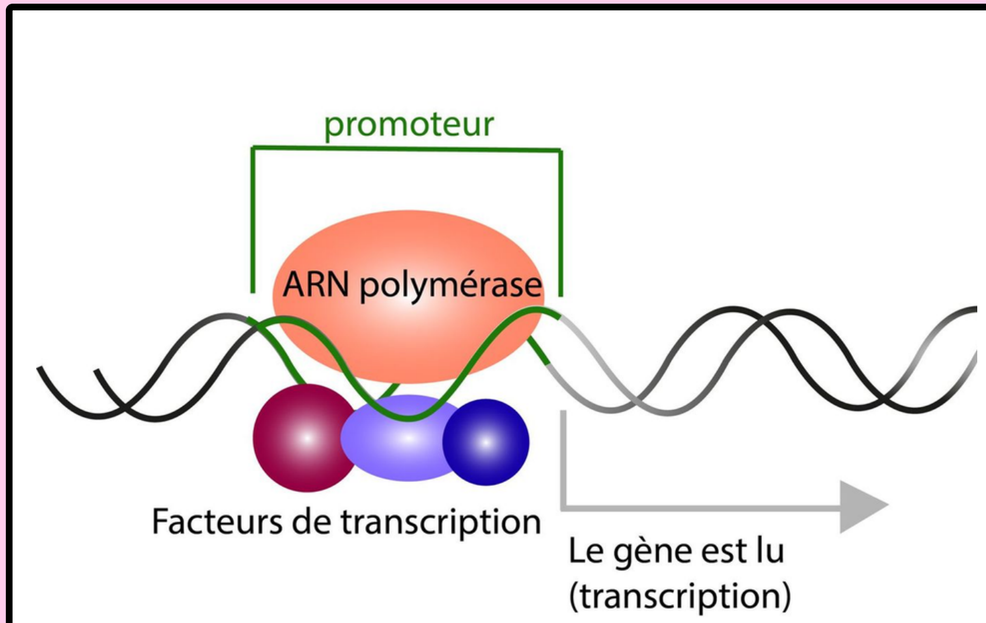
1. INTRODUCTION OF THE SUBJECT
2. PROJECT MANAGEMENT
3. CREATION OF THE DATASETS : FEATURE ENGINEERING
4. STATISTICS AND VISUALIZATIONS
5. NEXT STEPS

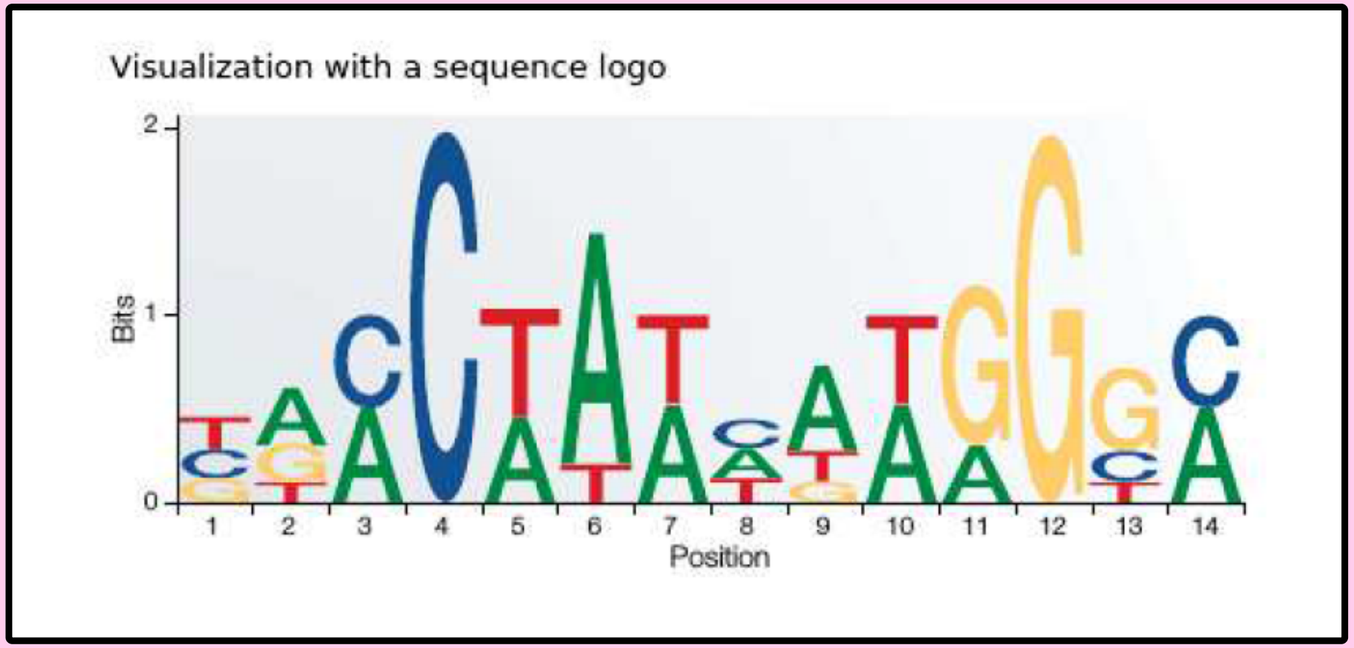


Context



- Motifs, Gene, Gene Sequences.

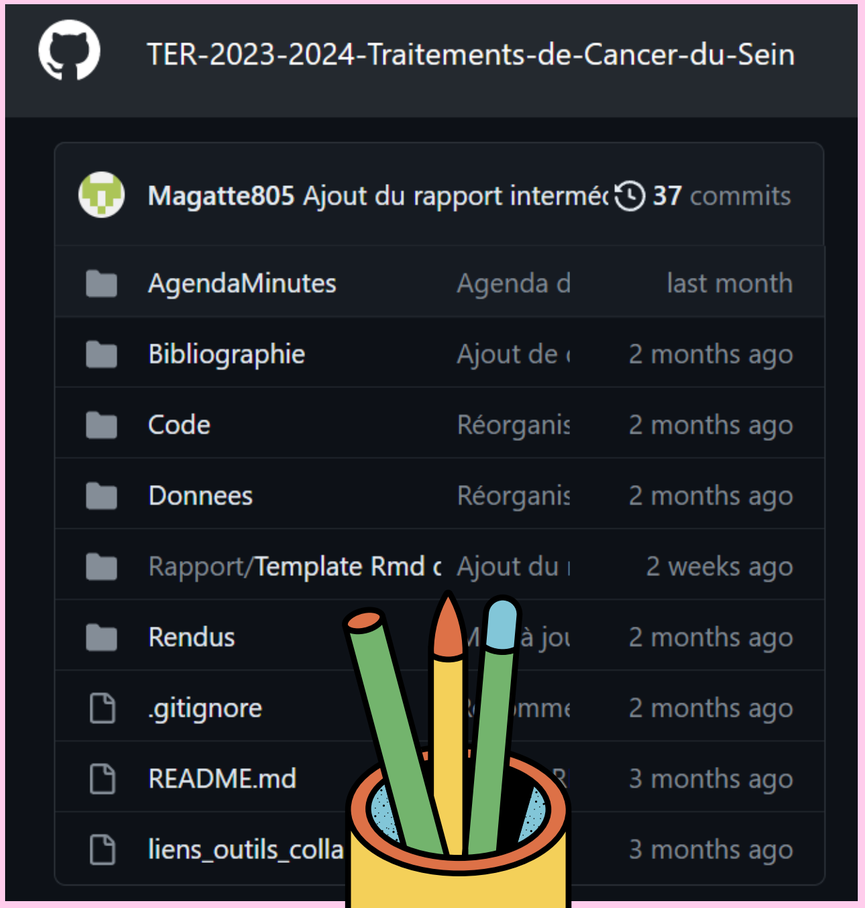


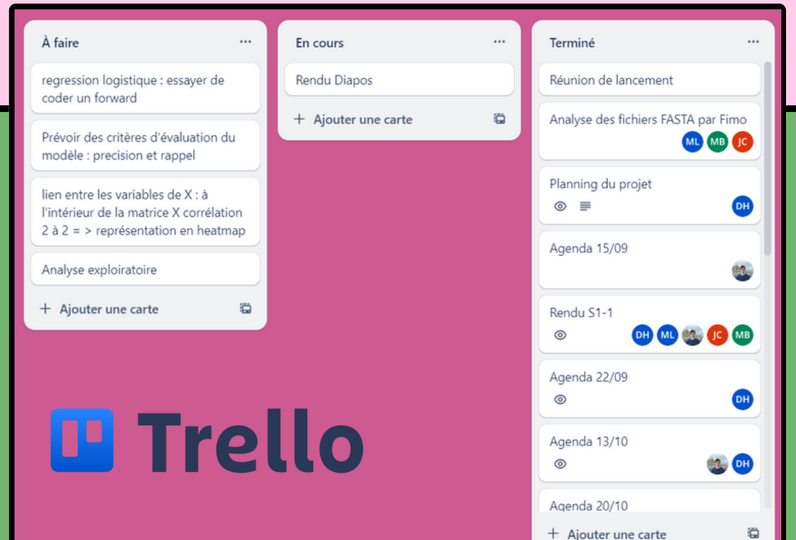


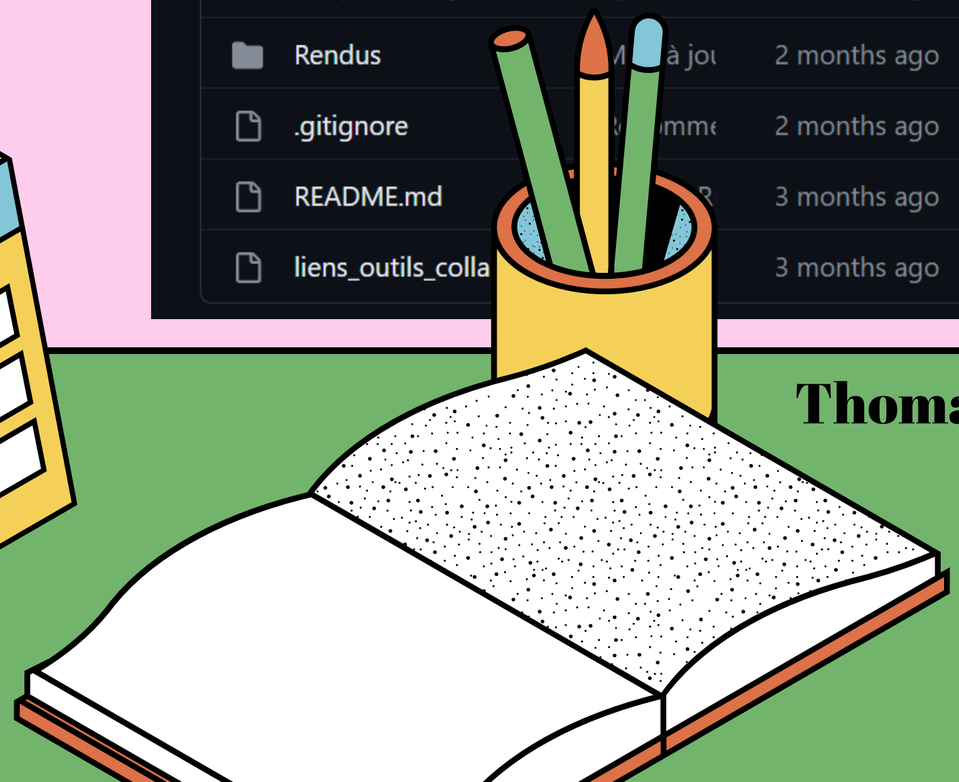
Dina

PROJECT MANAGEMENT

- Used tools : Python, R, Spark, FIMO.
- Discord, Trello.
- Shared deposit : GitHub.







Thomas

CONSTRUCTION OF THE DATASET

1

FIMO : Find Individual Motifs Occurrences

2

Creation X Matrix - DNA Sequence

3

Creation Y Matrix - Gene activity

4

Joining of X Y Matrixs

CONSTRUCTION OF THE DATASET

1

FIMO : Find Individual Motifs Occurrences

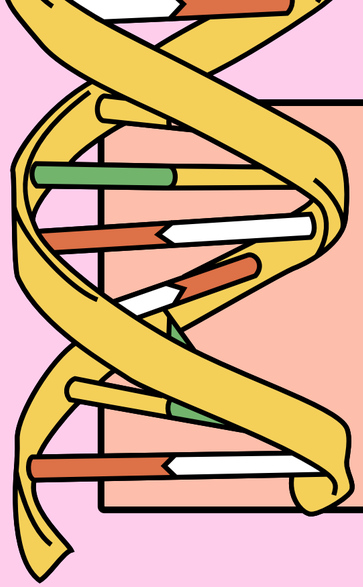
START OF THE SEQUENCE “CATG00000000002” :

TATTCTCTTATCTGGCCCCACCCACATCCTGCTGA
TGGGTAGAGCCTAGTGGTCTGTTTTGACAGGGCG
CTGATTGGTGCATTTACAATTC...

23557 sequences
each of them containing 1001 nucleotides

A	C	G	T
5339197	6462521	6448600	5329850
22,6%	27,4%	27,3%	22,6%

Jamila



CONSTRUCTION OF THE DATASET

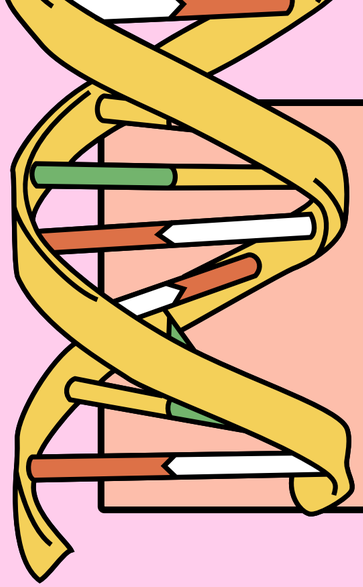
2

Creation X Matrix - DNA Sequence

23557 sequences
1205 variables (motif scores + DNA sequence composition)

	RUNX1	TFAP2A
CATG0002	p-values and frequences	
CATG0010		

Mehdi



CONSTRUCTION OF THE DATASET

3

Creation Y Matrix - Gene activity

89 patients

17014 gene expression measurements

	Patient1	patient2
gene_id1	Fold Change	
gene_id2		

Magatte

CONSTRUCTION OF THE DATASET

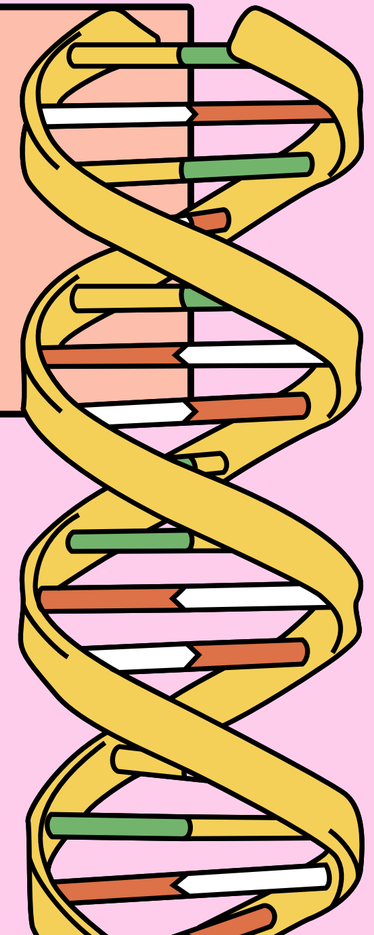
4

Joining of X Y Matrixs

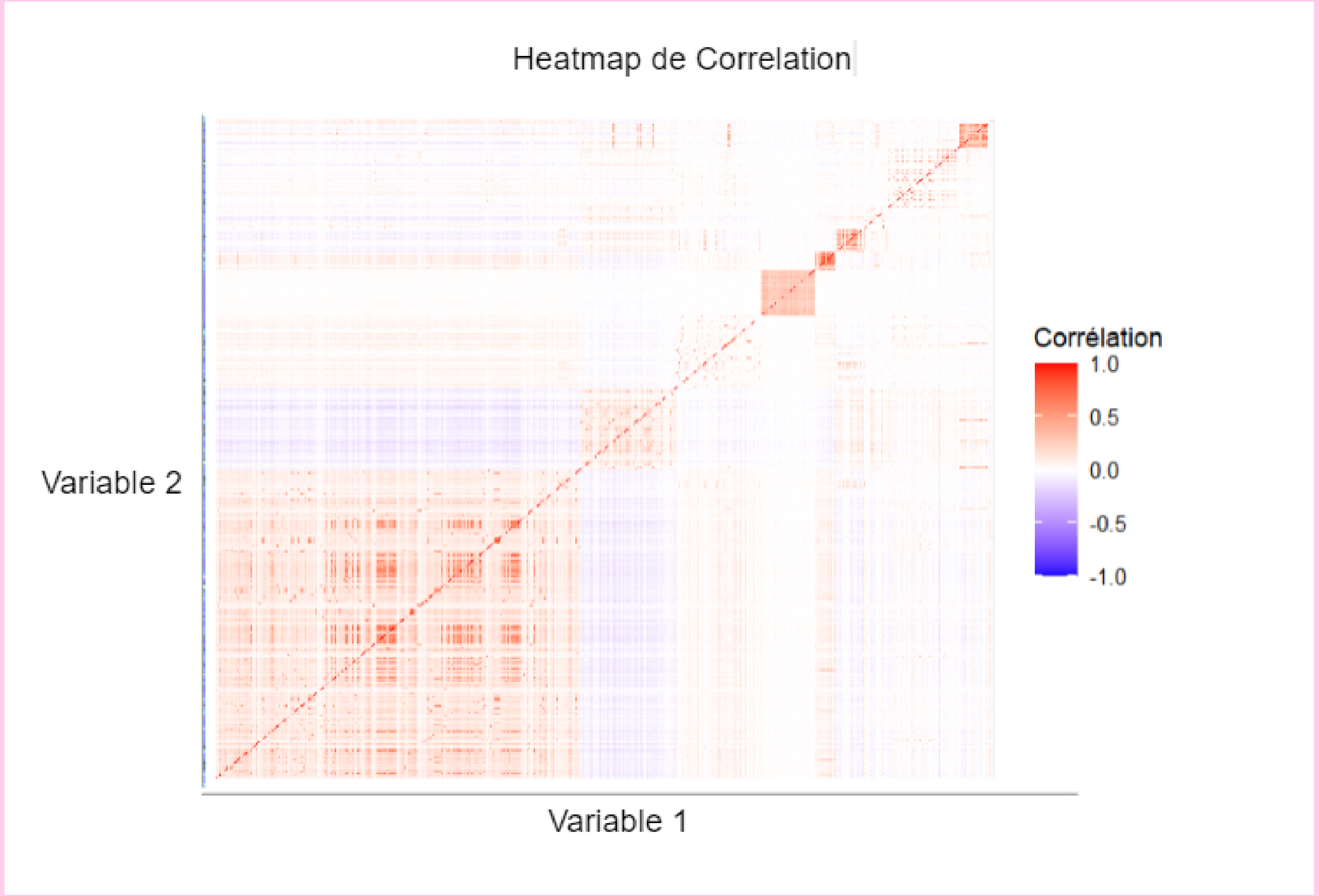
17014 sequences

1294 variables
Out off 1205 explicative variables
and 89 reponse variables

STATISTICS AND VISUALIZATIONS



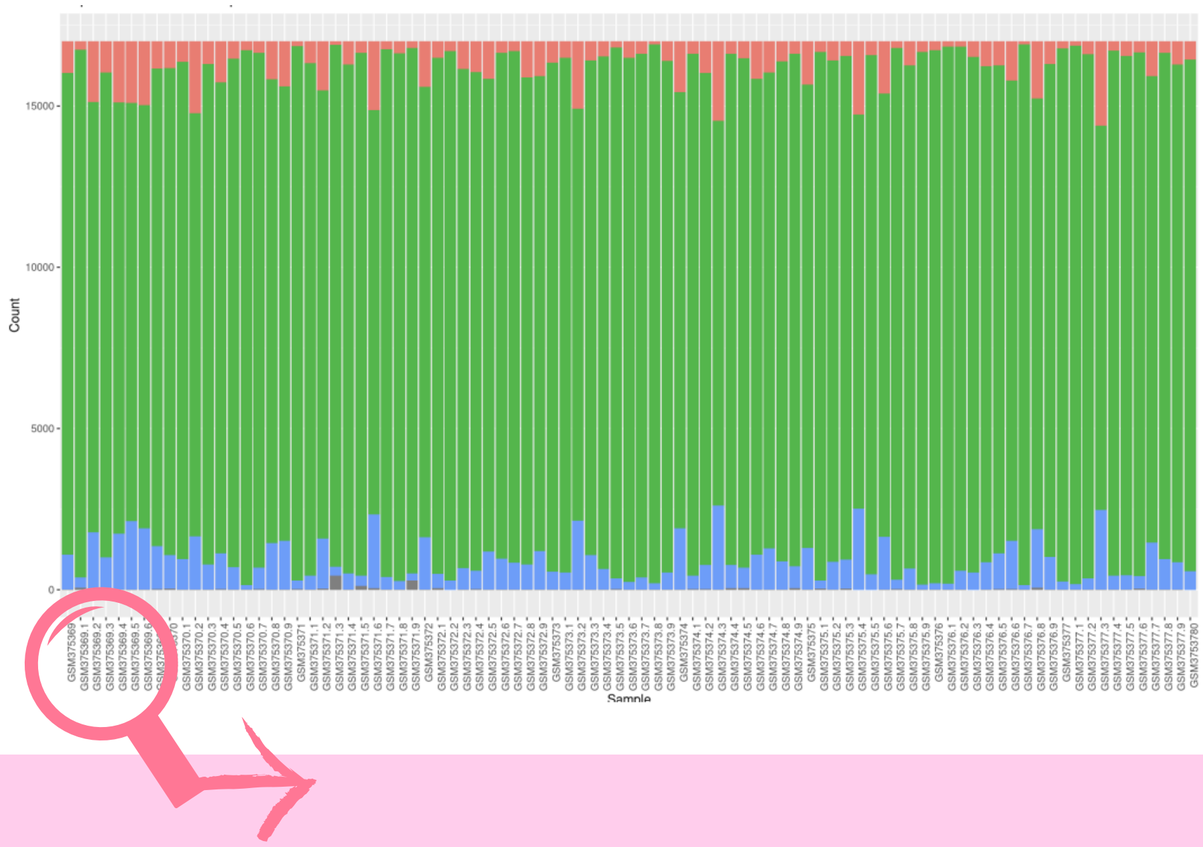
Correlation HeatMap Between Motifs

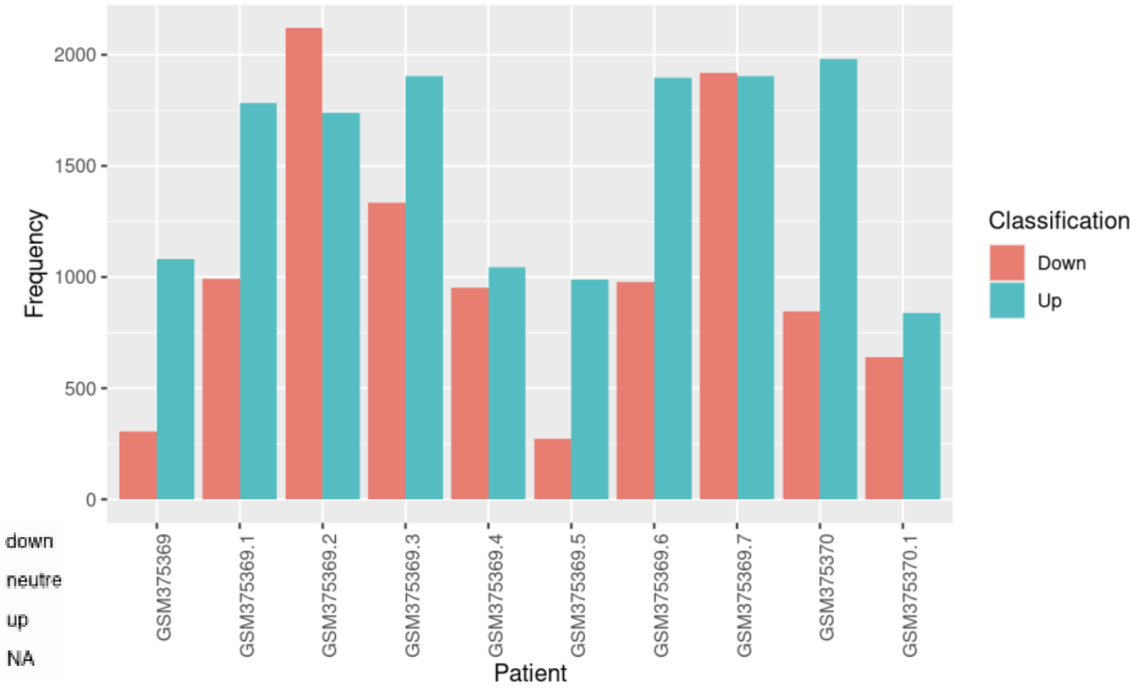


Dina

Global class distribution

An extract repartition of ups and downs

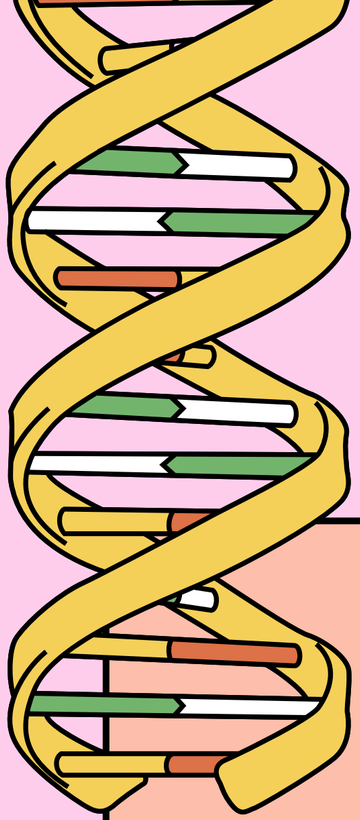




Thomas / Jamila

Next Steps

- Further Data Analysis
- Main difficulty : highly correlated explicative variables => selection of groups of variables with the same importance (hierarchical clustering, MLGL R package)
- Machine learning : Supervised clustering : logistic regression, Y = f (motifs, DNA relative frequencies), feature selection (LASSO, glmnet)



Thank you for your attention