Master Biologie-Santé cursus Santé
Groupe Binôme :
- Huynh Tram Anh
- Meza Verastegui Monica Lizeth

TP Bio-informatique

Q1 : Déterminer la taille du jeu des données :

La taille du jeu des données est le nombre total des gènes7126 multiplié par le nombre d'échantillons réalisés 50 ce qui représente un total de 356300 points de mesure.

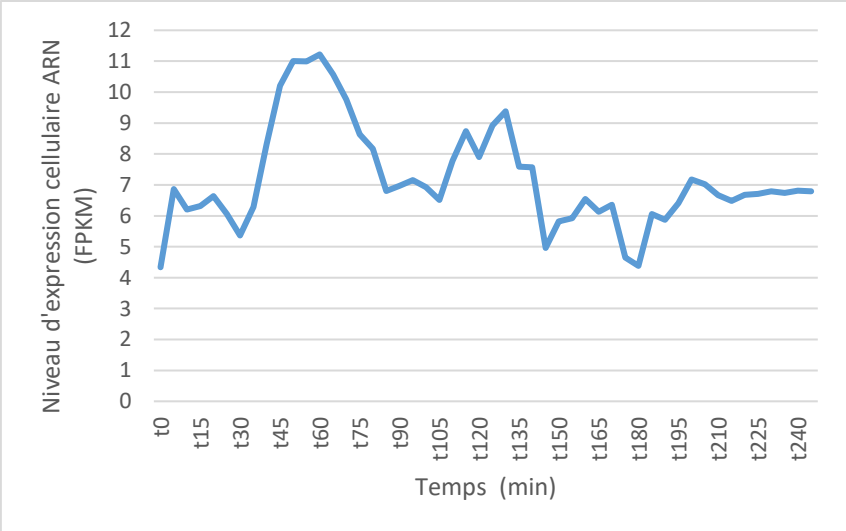
Q2 : Rappelez succinctement la fonction de chacun de ces gènes chez S. cerevisiae :

APC1 (Anaphase Promoting Complex subunit) : Est une ubiquitine-protéine ligase requise pour la dégradation des inhibiteurs de l'anaphase, y compris les cyclines mitotiques, pendant la transition métaphase-anaphase.

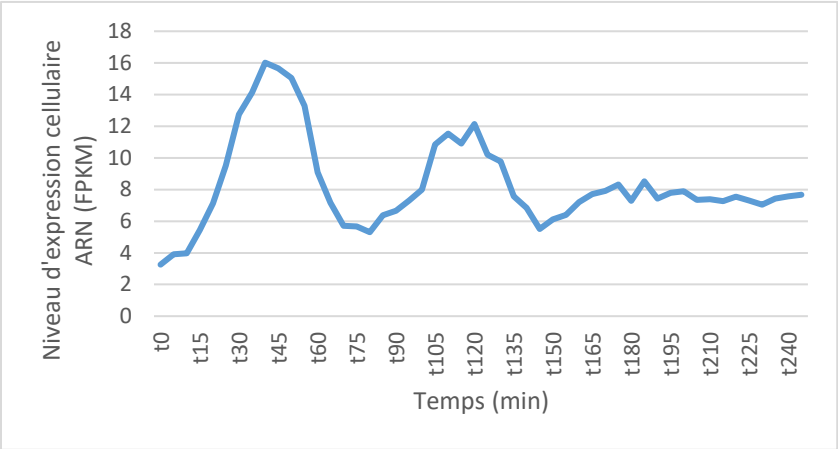
KIP1 (Kinesin related protein) : Protéine motrice liée à la kinésine, requis pour l'assemblage du fuseau mitotique, la ségrégation chromosomique et le partitionnement plasmidique de 2 microns.

UBC9 (Ubiquitin-Conjugating) : Est une enzyme de conjugaison SUMO impliquée dans la voie de conjugaison Smt3p. Protéine nucléaire nécessaire à la dégradation de la cycline en phase S et M et au contrôle mitotique.

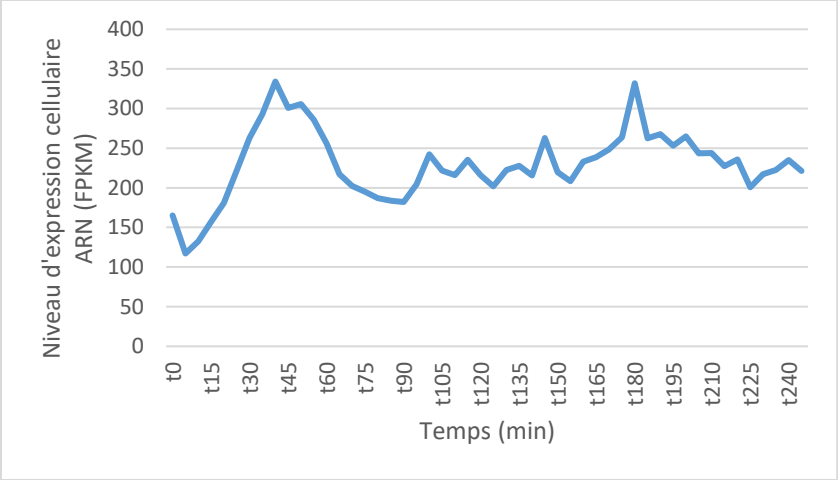
Q3 : Récupérez les données correspondant à ces gènes et tracez, pour chacun d'eux, un graphique représentant leur niveau d'expression en fonction du temps.



Graphique 1 : Expression cellulaire ARN du gène APC1 chez *S. cerevisiae.* La variation périodique de l'expression du gène au cours du cycle cellulaire en FPKM (Fragments Per Kilobase of transcript per Million mapped reads) pendant l'intervalle de 245 minutes.



Graphique 2 : Expression cellulaire ARN du gène KIP1 chez *S. cerevisiae.* La variation périodique de l'expression du gène au cours du cycle cellulaire en FPKM (Fragments Per Kilobase of transcript per Million mapped reads) pendant l'intervalle de 245 minutes.

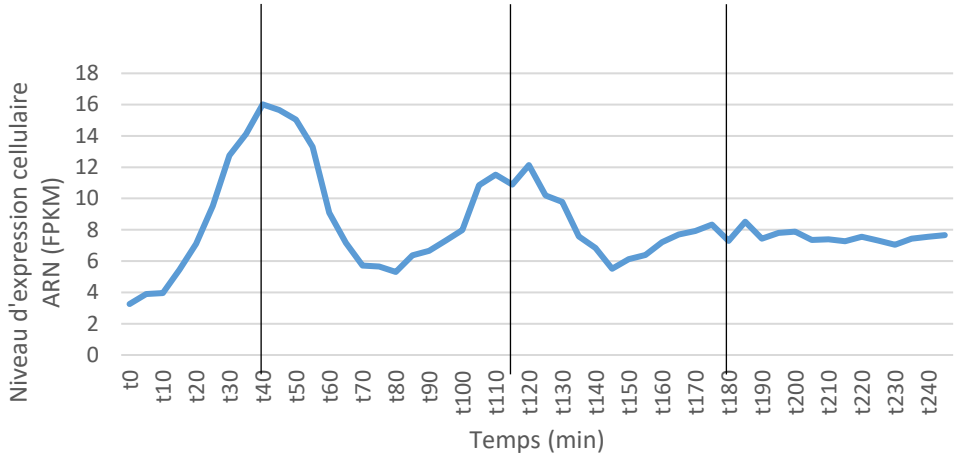


Graphique 3 : Expression cellulaire ARN du gène UBC9 chez *S. cerevisiae.* La variation périodique de l'expression du gène au cours du cycle cellulaire en FPKM (Fragments Per Kilobase of transcript per Million mapped reads) pendant l'intervalle de 245 minutes.

Q4 : Quelle est la durée moyenne d'un cycle ? Combien de cycles sont analysés au cours de cette cinétique de 245 minutes ?

D'après les graphiques précédents, nous observons qu'il y a une expression périodique de ces gènes et nous en déduisons que la durée moyenne d'un cycle est de 70 minutes. Ce temps est établi entre deux niveaux maximum d'expression. Par conséquent il y a deux cycles cellulaires au cours des 245 minutes.

Gène	APC1	KIP1	UBC9
La durée moyenne d'un cycle	65 min	70 min	70 min
Nombre des cycles	3	2	2
Explication	Cycle 1 = t60 – t5 Cycle 2 = t130 - t60 Cycle 3 = t200 - t130	Cycle 1 = t115 – t40 Cycle 2 = t180 - t115	Cycle 1 = t110 - t40 Cycle 2 =t180 - t110



Graphique 4 : Expression cellulaire ARN du gène KIP1 chez *S. cerevisiae.* Visualisation des deux cycles cellulaires avec une durée moyenne de 70 min.

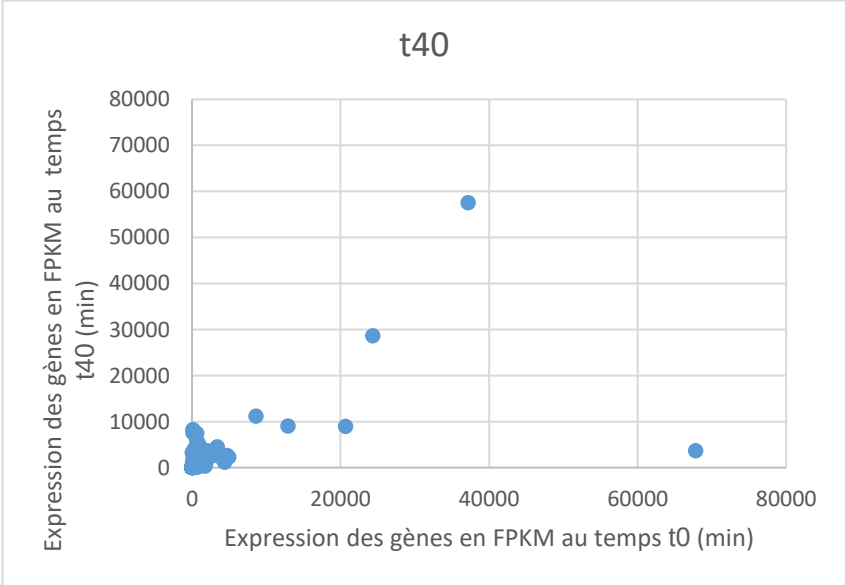
Q5 : Que constatez-vous concernant la modulation de l'expression de ces gènes au cours du temps ?

Au cours du temps, nous pouvons voir que pour les gènes KIP1 et APC1 le niveau maximum d'expression va diminuer progressivement. Mais pour le gène UBC9 le niveau maximum d'expression va fluctuer. Nous constatons que les gènes ont une expression périodique et ceci peut être relationnée aux facteurs propres de la cellule. Toutes les cellules ont été synchronisées au début de phase G1 mais au cours du temps chaque cellule va développer l'expression de ces propres gènes dans un temps précis selon divers conditions et besoins. Le gène UBC9 est le plus exprimé parmi ces 3 gènes.

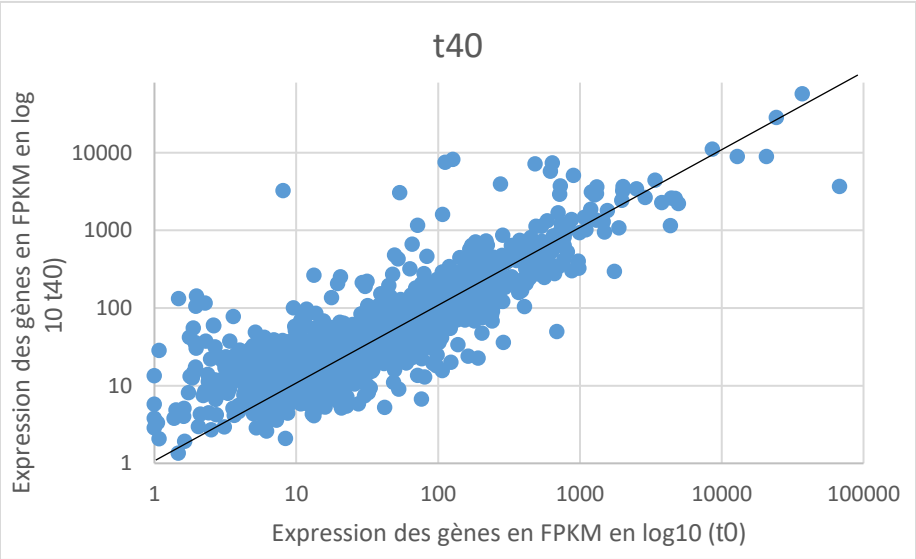
Q6 : À quel point de la cinétique, chacun de ces gènes est-il le plus fortement exprimé ?

D'après les graphiques réalisés, nous observons que le gène APC1 est plus fortement exprimé dans le T60, le gène KIP1 dans le T40 et le gène UBC9 dans le T40. Pour cette raison, nous allons considérer le temps T40 en tant que le temps où il y a plus d'expression pour les trois gènes.

Q7 : Réaliser deux graphiques :



Graphique 5 : Expression des gènes au T40 versus T0 chez S. cerevisiae. Distribution des niveaux d'expression des gènes en FPKM au temps 0 et au temps 45.

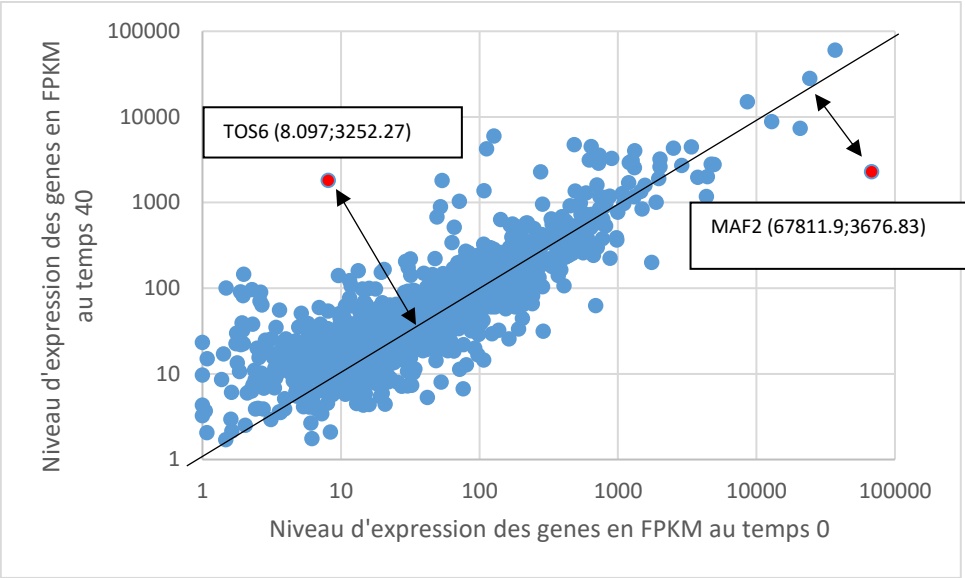


Graphique 6 : Expression des gènes en échelle logarithmique au T0 versus T40 chez S. cerevisiae. Distribution des niveaux d'expression des gènes en FPKM au temps 0 et au temps 45 en échelle logarithmique base 10.

Q8 : Quel est l'impact de la transformation des données en log ? Interprétez la distribution globale des données ?

Le graphique 4 (Nuage des points) permet de déterminer la différenciation d'expression cellulaire des gènes au temps T40 par rapport au temps T0. Néanmoins, la transformation des données en échelle logarithmique (graphique 5) permet de rendre plus visible les variations des niveaux d'expressions, voir les différences les plus importants. Nous allons voir que les gènes qui traversent la ligne droite ont le même niveau d'expression dans le temps T40 et T0, soit fortement exprimés ou faiblement exprimés. Tandis que les gènes qui sont dispersés vont indiquer quels gènes sont surexprimés dans le T40 par rapport au T0, en haut de la ligne droite, ou sous-exprimés dans le temps T40 par rapport au T0, en bas de la ligne droite.

Q9 : Sélectionnez sur le graphique un gène parmi les plus surexprimés et un parmi les plus sous-exprimés au temps T45 par rapport au temps T0. Explicitez sur quels critères vous avez choisi ces gènes.



Graphique 7 : Le gène le plus surexprimé et le gène le plus sous-exprimé en échelle logarithmique au T0 et T40 chez S. cerevisiae.
Présentation du gène TOS6 comme le plus surexprimé et le gène MAF2 comme le plus sous-exprimés au temps 40 versus temps0.

Le gène TOS6 a été choisi en tant que le gène le plus surexprimés au temps 40 par rapport au temps 0 et le gène MAF2 e été choisi en tant que le gène le plus sous-exprimés au temps 40 par rapport au temps 0. Nous avons choisi ces deux gènes sont qu'ils sont les plus éloignés de la droite parce que ceci montre une différence de valeur entre le T40 et T0. Nous avons aussi considéré la différence numérique notable, le gène TOS6 au temps 0 présent 8.097 FPKM et au temps 40 présent 3252.27 FPKM, tandis que le gène MAF2 au temps 0 présent 67811.9 FPKM et au temps 40 présent 3676.83 FPKM.

Q10 : Sélectionner le gène le plus surexprimé et le gène le plus sous-exprimé basé sur les valeurs de modulation de l'expression des gènes au temps T par rapport au temps de référence de T0. Ce résultat est-il concordant avec la question précédente ?

Les résultats sont concordants avec les gènes que nous avons sélectionnés précédemment.

Gène TOS6 : T0= 8.097 T40= 3252.27 → Modulation (t40/t0) = 401.6635791

Gène MFA2 : T0=67811.9 T40=3676.83 → Modulation (t40/t0) = 0.054221014

Q11 : Indiquez succinctement la fonction de chacun de ces gènes.

Gène TOS6 : Protéine de paroi cellulaire dépendante du glycosylphosphatidylinositol ; l'expression est périodique et diminue en réponse à la perturbation de l'ergostérol ou à l'entrée en phase stationnaire ; la déplétion augmente la résistance à l'acide lactique.

Gène MFA2 (Mating factor A) : Facteur A de phéromone d'accouplement, interagit avec les cellules alpha pour induire l'arrêt du cycle cellulaire et d'autres réponses menant à l'accouplement. La biogenèse implique la modification C-terminale, la protéolyse N-terminale et l'exportation ; également encodé par MFA1.

Q12 : D'après le tableau obtenu, dans quels processus cellulaires sont impliqués ces 30 gènes ? Quelle est la catégorie fonctionnelle la plus enrichie ? Quel est le facteur d'enrichissement de cette catégorie ? La sur-représentation de cette catégorie fonctionnelle dans le top 30 est-elle statistiquement significative ?

Ces 30 gènes sont impliqués dans les suivantes processus cellulaires :

- Cell cycle (Cluster frequency: 16 of 30 genes, 53.3%, 53.3%)
- Regulation of macromolecule metabolic process (16 of 30 genes, 53.3%)
- regulation of metabolic process (16 of 30 genes, 53.3%)

La catégorie fonctionnelle la plus enriche est « cell cycle » (Cluster frequency: 16 of 30 genes, 53.3%, Genome frequency 791 of 7166 genes, 11.0%, P-value: 3.45e-06).La sur-représentation de cette catégorie fonctionnelle dans le top 30 est statistiquement significative (P-value < 0.01). Le facteur d'enrichissement de cette catégorie est 4.85, obtenu par le ratio entre 53,3% (pourcentage des gènes de ce groupe de 30 gènes qui appartient à ce groupe fonctionnelle) et 11,0% (pourcentage des gènes du génome de Saccharomyces cerevisiae qui appartient à ce groupe fonctionnelle).

Result Table

Terms from the Process Ontology of gene_association.sgd with p-value <= 0.01

Gene Ontology term	Cluster frequency	Genome frequency	Corrected P-value	FDR	False Positives	Genes annotated to the term
cell cycle	16 of 30 genes, 53.3%	791 of 7166 genes, 11.0%	3.45e-06	0.00%	0.00	CLB6, SPO16, MCD1, YHP1, NRM1, YOX1, AIM20, GIC1, HHT1, GIN4, CLB2, CLB1, HHF2, HHF1, PCL1, CLN1
cell cycle process	14 of 30 genes, 46.7%	647 of 7166 genes, 9.0%	1.94e-05	0.00%	0.00	[illegible], GIN4, CLB2, CLB6, SPO16, CLN1, HHT1, HHF1, GIC1, HHF2, NRM1, YOX1
regulation of cyclin-dependent protein serine/threonine kinase activity	5 of 30 genes, 16.7%	34 of 7166 genes, 0.5%	5.83e-05	0.00%	0.00	CLB1, PCL1, CLN1, CLB6, CLB2
regulation of cyclin-dependent protein kinase activity	5 of 30 genes, 16.7%	35 of 7166 genes, 0.5%	6.79e-05	0.00%	0.00	CLB6, CLB2, PCL1, CLB1, CLN1
mitotic cell cycle phase transition	8 of 30 genes, 26.7%	183 of 7166 genes, 2.6%	0.00014	0.00%	0.00	CLN1, GIC1, NRM1, YOX1, CLB1, YHP1, CLB2, CLB6

Figure 1 : Principales implications fonctionnelles des 30 gènes les plus surexprimés. Le 53.3% de ces 30 gènes sont impliqués dans le processus du cycle cellulaire avec un p-value de 3.45 e-06.

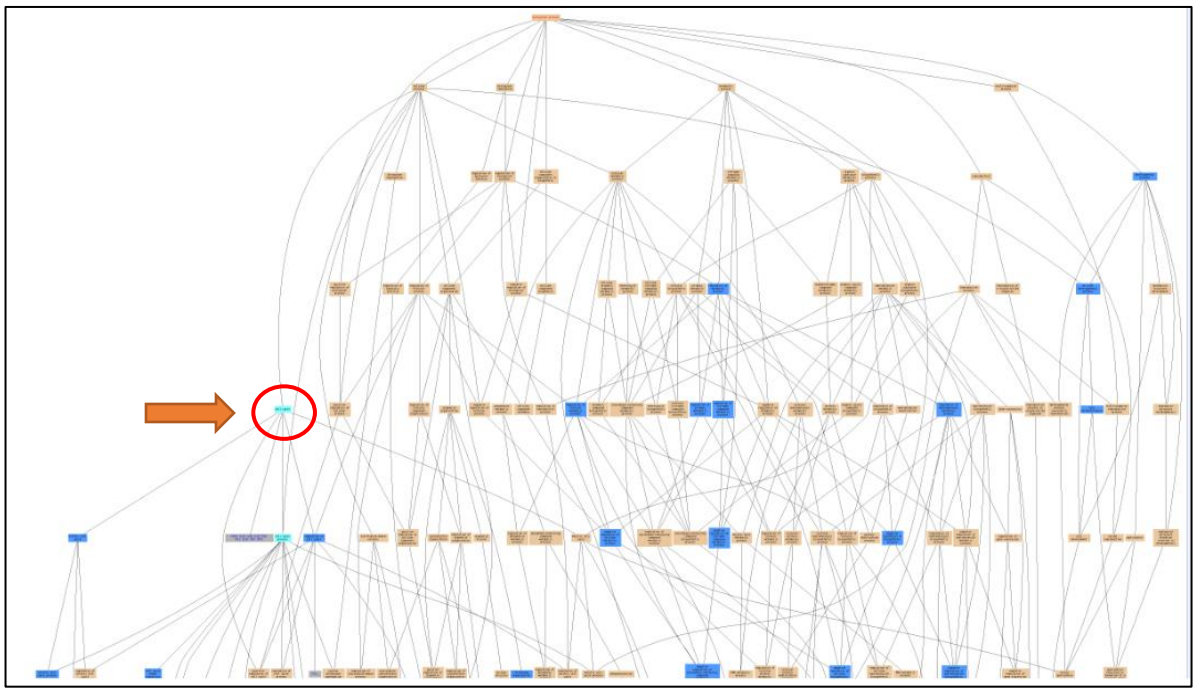


Figure 2 : L'arbre de l'ontologie GO des 30 gènes les plus surexprimés. Position de la catégorie fonctionnelle la plus enrichie.

Q13 : Combien de groupes de gènes identifiez-vous ? Décrivez la modulation des gènes au sein de chacun de sous-groupes.

D'après le clustering que nous avons obtenu, nous allons voir les différences d'expressions de ces gènes au cycle cellulaire pendant les 245 minutes. La couleur bleue représente la surexpression des gènes, tandis que la couleur rouge représente une sous-expression. La couleur blanche est neutre. Nous avons identifié deux groupes des gènes. Le premier groupe a été sous-exprimé au début du cycle cellulaire jusqu'à le temps 15, mais après il y a eu une fluctuation sans avoir un niveau important. Par contre, le deuxième group a eu une expression neutre jusqu'au temps 30 et après il y a une surexpression presque constante tout au long de ce temps d'évaluation.

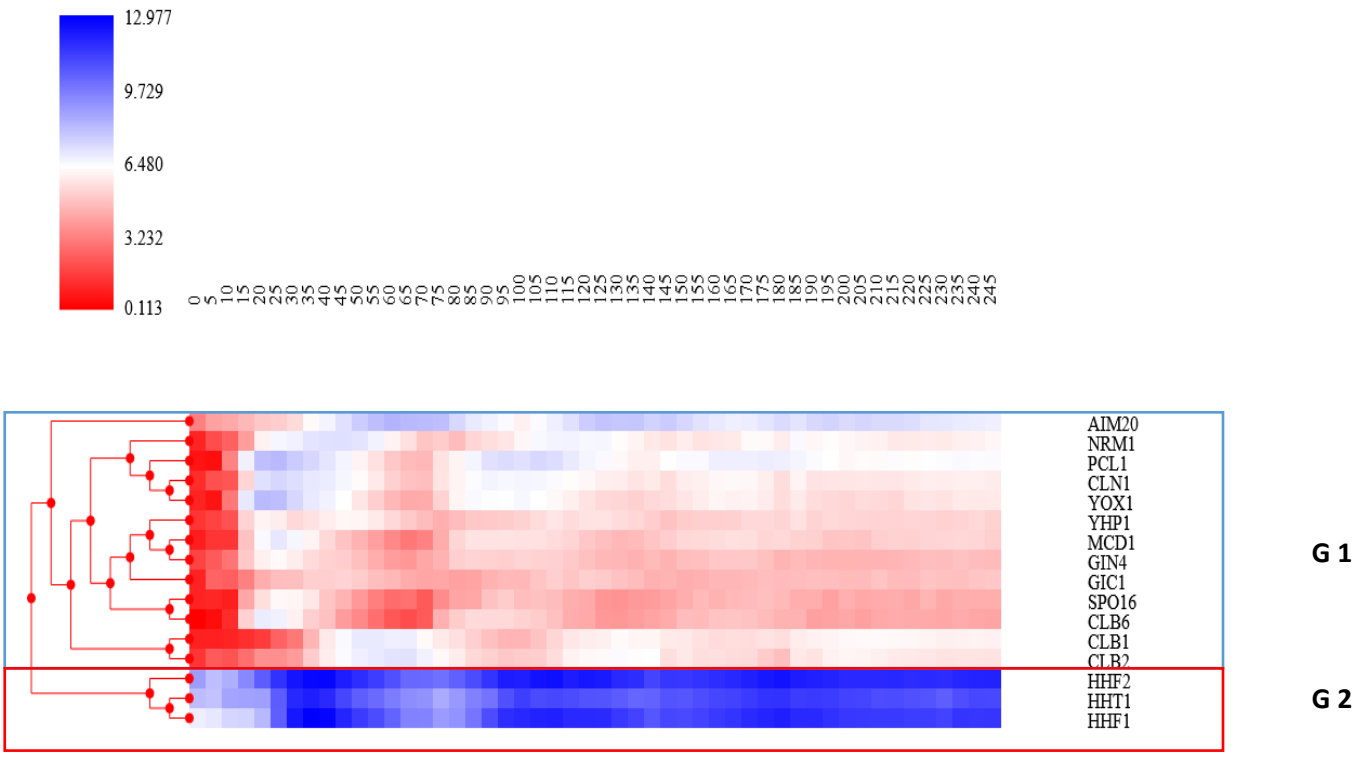


Figure 3 : Clustering *Cell Cycle* des gènes. Différences des niveaux d'expression au cycle cellulaire de ces gènes au cours de 245 minutes. La couleur bleue exprime surexpression et la couleur rouge sous-expression.

Q14 : Représentez-en boxplot l'expression des gènes des différentes sous-groupes en fonction du temps. Tous les gènes présentent-ils une expression périodique ? Que vous apporte cette représentation par rapport aux informations fournies par le groupement hiérarchique ?

D'après les bloxplots représentés pour chaque groupe, nous voyons que pour le groupe 1 il y a une expression périodique mais pas très notable au cours du temps (neutre et sous-expression). Tandis que pour le groupe 2 il y a une expression périodique plus remarquable. La boite de Boxplot représente le 50% des valeurs obtenu dans le temps T signalé et la droite noire horizontale indique la médiane. Cette représentation des niveaux d'expression des gènes nous indique que tous les gènes présent une expression importante périodique qui peut être déterminé par divers conditions et besoins de la cellule à un moment donné. Quelques gènes peuvent avoir une expression plus notable par rapport aux autres et ceci peut être relaitonné a l'importance de son activité précise, par exemple la préservation de l'intégralité du génome ou assemblage de la chromatine réalisé par le gène HHF1, un des plus surexprimés. Ceci explique pourquoi le groupe le plus enrichie dans le question précèdent est le cycle cellulaire.

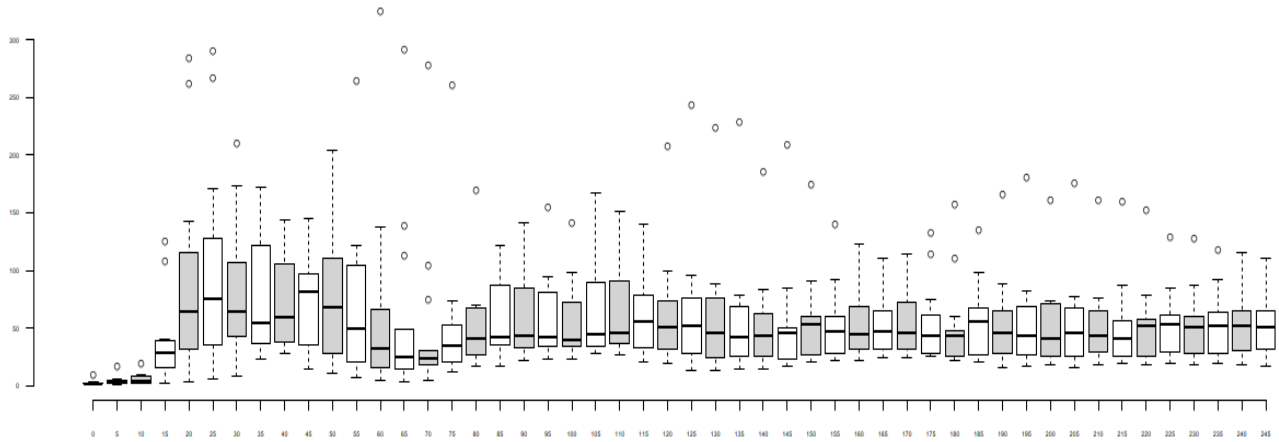


Figure 4 : Boxplot des gènes du groupe 1. Représentation quantitative de l'expression des gènes AIM20, NRM1, PCL1, CLN1, YOX1, YHP1, MCD1, GIN4, GIC1, SPO16, CLB6, CLB1, CLB2 au cours du temps (245 minutes).

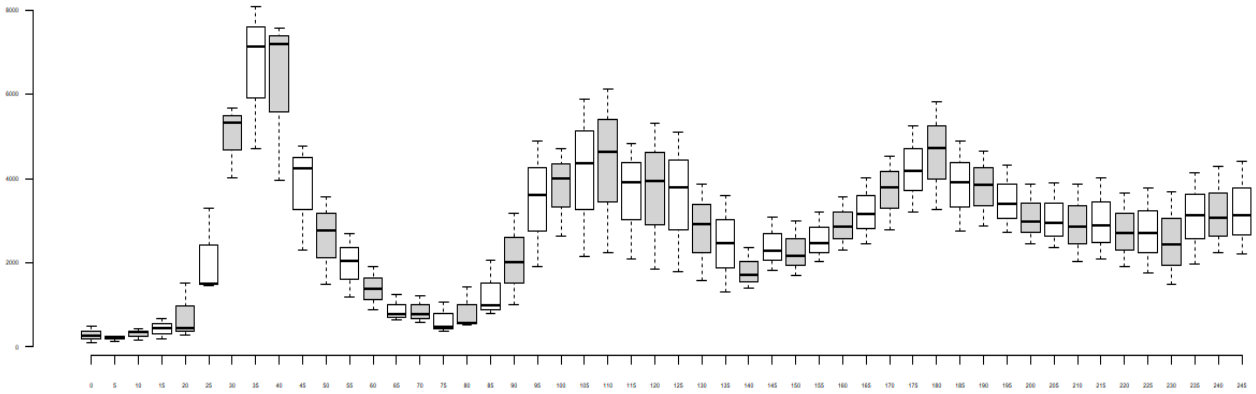


Figure 5 : Boxplot des gènes du groupe 2. Représentation quantitative de l'expression des gènes HHF1, HHF2, HHT1 au cours du temps (245 minutes).

Q15 : Que pensez-vous du schéma expérimental d'étude ? Que suggérez-vous comme amélioration ?

Nous pensons que le schéma expérimental d'étude nous a aidé à travailler en équipe et améliorer notre capacité de recherche par nous-même. Aussi ce schéma nous a servi d'instrument pour avoir la base de connaissances d'utilisation du web et d'analyse comme boxplot, mev, yeastgenome, etc.

Nous suggérons une séance qui soit réalisé et expliqué pour les personnes qui ont un ordinateur MAC parce que les logiciels sont complètement différents et nous avons eu quelques difficultés pour la réalisation des graphiques.