

## Oral UC2 Big Data et données “-omiques”

ALLYNDREE J., CLERC T., LACOSTE L., LORTHIOS T.

- 1 Méthode de sélection
- 2 Sélection des métabolites
- 3 Sélection des protéines
- 4 Métabolites KEGG
- 5 Protéines ThaleMine, UniProt

Le chargement a nécessité le package : glmnet

Loaded glmnet 4.1-3

Le chargement a nécessité le package : mvtnorm

Le chargement a nécessité le package : survival

Le chargement a nécessité le package : TH.data

Le chargement a nécessité le package : MASS

Attachement du package : 'TH.data'

L'objet suivant est masqué depuis 'package:MASS':

geyser

## Section 1

### Méthode de sélection

La méthode utilisée pour sélectionner les métabolites et protéines repose sur le package `MultiVarSel` et utilise la *méthode du Lasso*.

## Section 2

## Sélection des métabolites





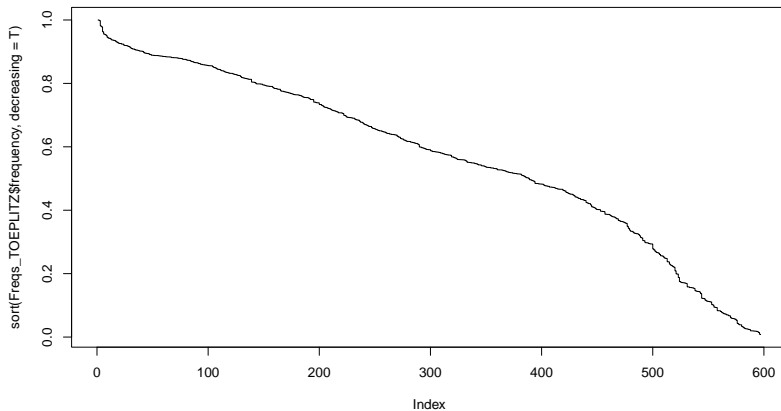


## Matrices $Y$ et $X$

Lignes	Valeur	Colonnes	Valeur
X	9	X	3
Y	9	Y	199



On voit donc que lorsque l'on applique la structure non paramétrique la P-valeur de 0.741 nous indique que parmi les fonctions testées c'est la structure non paramétrique qui permet de blanchir la matrice.

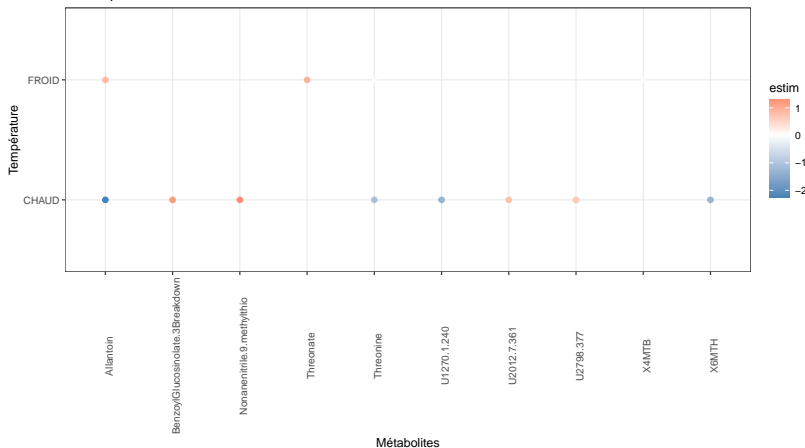


1, 0.998, 0.9812, 0.9784, 0.9628, 0.9554, 0.9544, 0.9516, 0.9458,  
0.9428, 0.9422, 0.9402, 0.9372, 0.9366, 0.9356, 0.9346, 0.9312,  
0.9298, 0.9272, 0.9258, 0.9256, 0.924, 0.9238, 0.9206, 0.9196,  
0.9194, 0.9184, 0.9166, 0.9152, 0.912, 0.9098, 0.9084, 0.9082,  
0.9062, 0.905, 0.9048, 0.9036, 0.9022, 0.9018, 0.9016, 0.9008,  
0.897, 0.8954, 0.895, 0.8942, 0.894, 0.8914, 0.8906, 0.8888, 0.8888

Alanine, Asparagine, Leucine, Threonine, Threonine, Tryptophan, Tryptophan, Valine, gamma.Tocopherol, Eicosanoate, Quercetin, Quercitrin, X4MTB, X4MTB, X5MTP, X6MTH, X6MTH, X7MTH, BenzoylGlucosinolate.3Breakdown, Hexanenitrile.6methylthio, Nonanenitrile.9.methylthio, U2609.4.361, U3122.4.202.I3M., U3122.4.202.I3M., UGlucosinolate140.1, UGlucosinolate140.1, Pentonate.4, Threonate, Allantoin, Allantoin, Xylitol, Galactinol, Glucopyranose..H2O., U1093.6.147, U1270.1.240, U1270.1.240, U1767.3.243, U1852.0.217, U2012.7.361, U2197.2.494, U2315.2.245, U2529.8.361, U2688.5.333, U2692.9.361, U2798.377, U2839.3.312, U2882.5.297, U3080.7.361, U3218.5.297, U3279.7.361

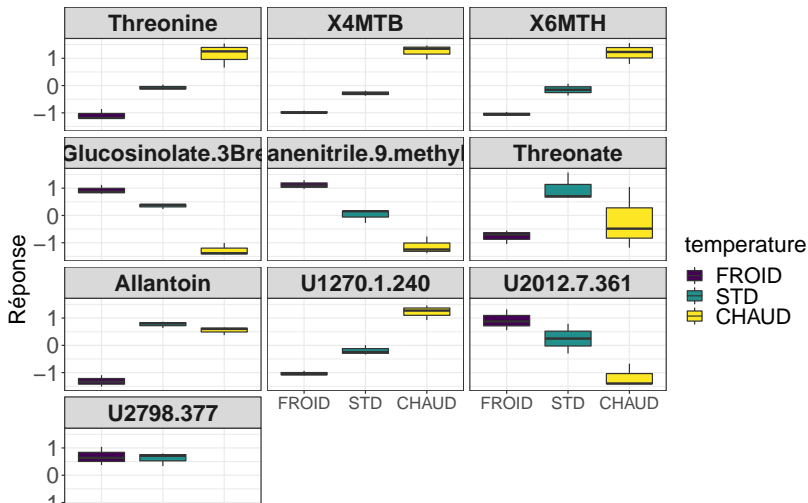
# Réponses des métabolites dépassant le seuil 0.94

Réponse des Métabolites sélectionnés pour les conditions de températures, au seuil 0.94



# Boxplots des réponses des métabolites dépassant le seuil 0.94

Using temperature as id variables





# Métabolites retenus au seuil 0.94

## Table des métabolites retenues

---

Metabolites
-------------

---

Threonine
-----------

X4MTB
-------

X6MTH
-------

BenzoylGlucosinolate.3Breakdown
---------------------------------

Nonanenitrile.9.methylthio
----------------------------

Threonate
-----------

Allantoin
-----------

U1270.1.240
-------------

U2012.7.361
-------------

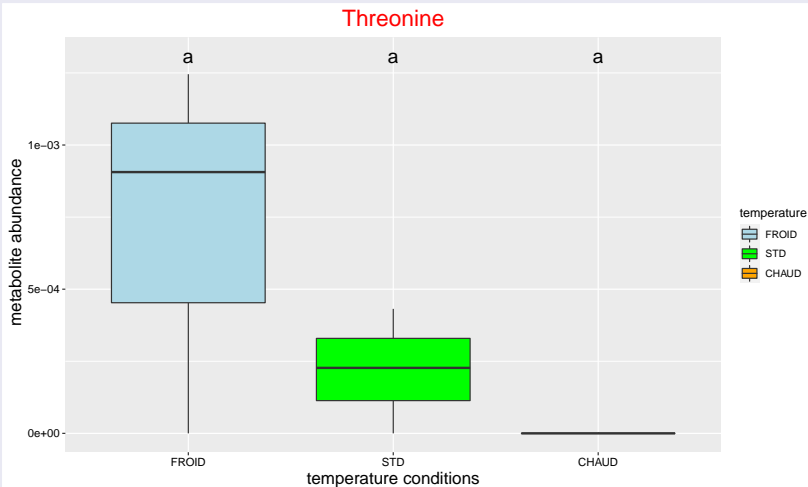
U2798.377
-----------

---

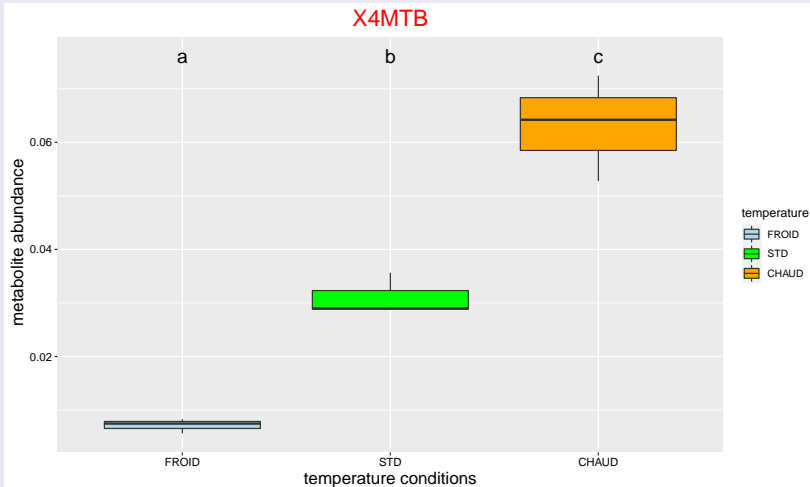
## Boxplots de Tukey (avec cld)

Les graphiques suivants présentent les boxplots pour les métabolites sélectionnés par la méthode du lasso

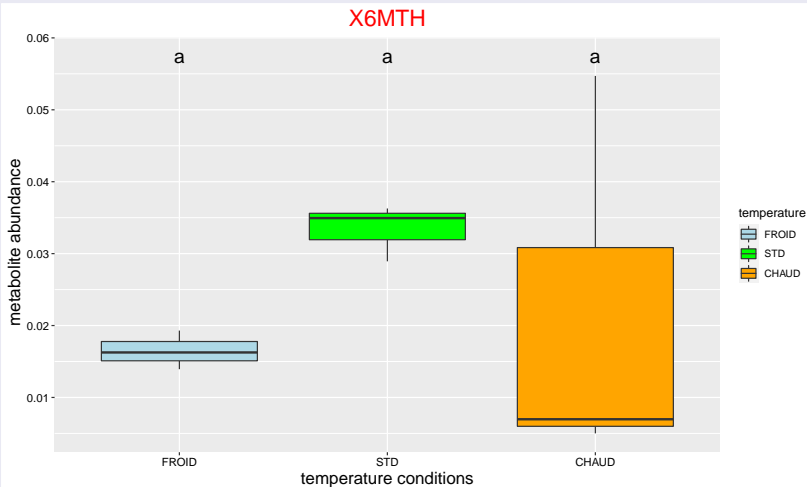
# Threonine



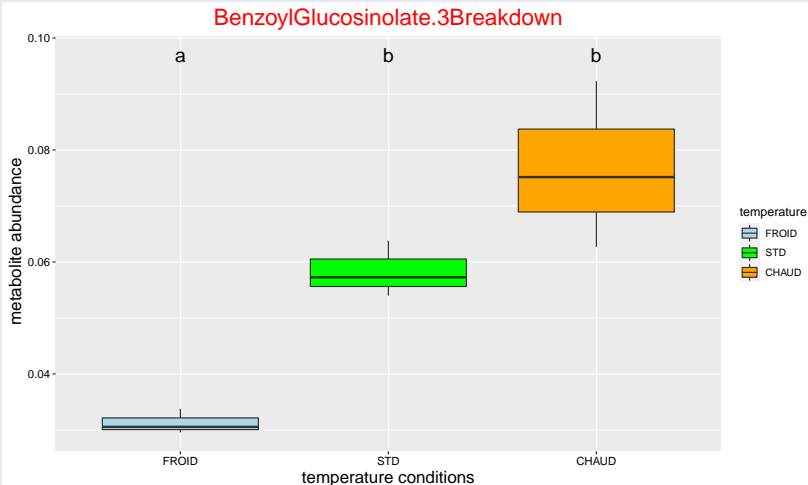
## X4MTB



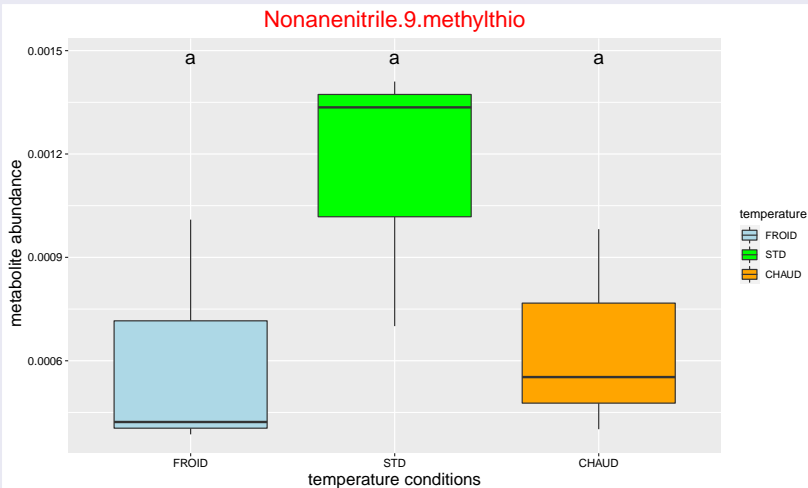
## X6MTH



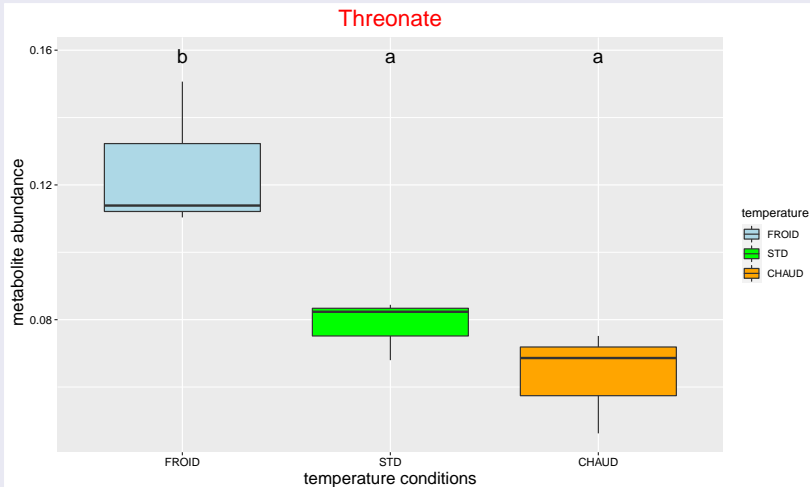
## BenzoylGlucosinolate.3Breakdown



# Nonanenitrile.9.methylthio

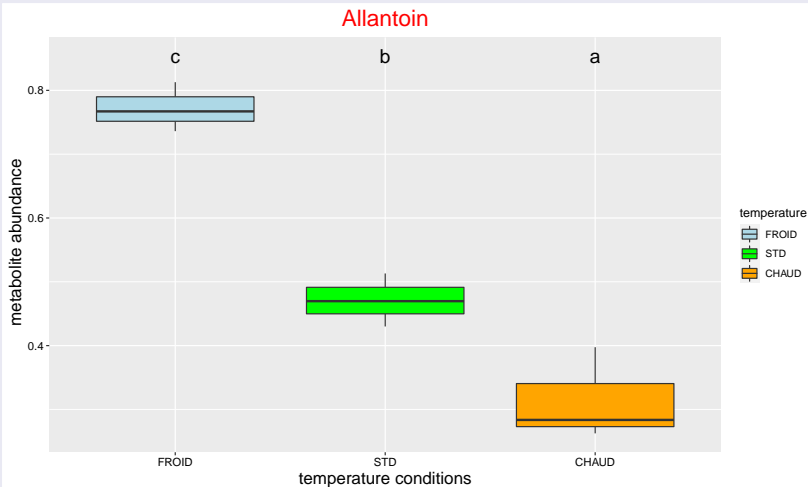


# Threonate

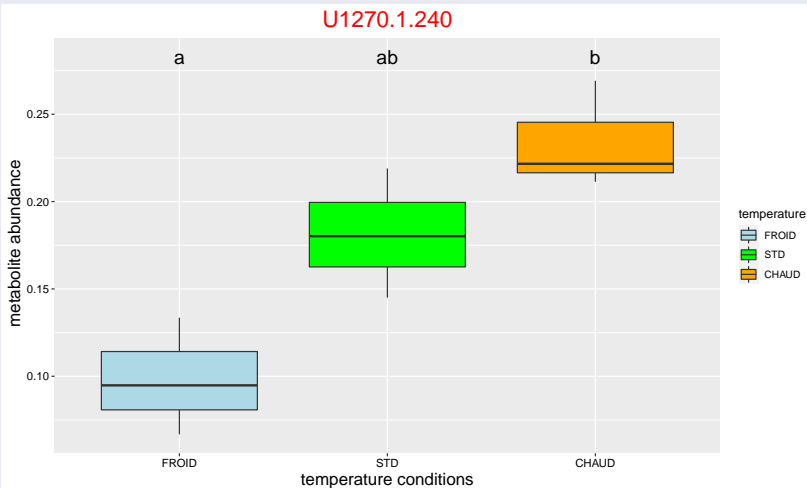




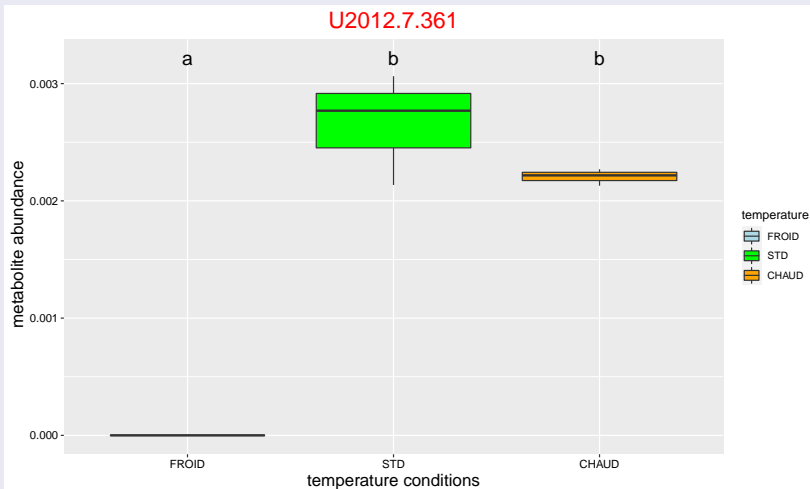
## Allantoin



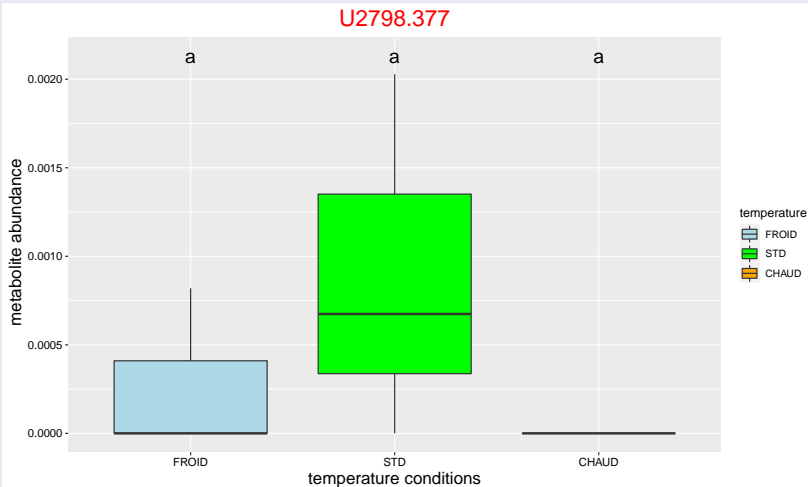
U1270.1.240



U2012.7.361



U2798.377



Le fichier **metabolites\_selection\_lasso\_0.94.csv** existe déjà.

## Sélection des protéines

Seuil défini à 0.95

- Nombre de colonne dont la moyenne est nulle : 0
- Nombre de colonne où la variable est constante : 0
- Nombre de NAs dans le dataframe : 0

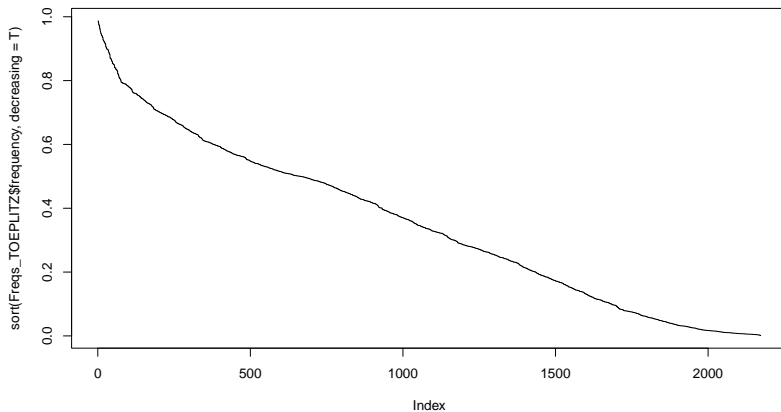


# Matrices Y et X

Lignes	Valeur	Colonnes	Valeur
X	9	X	3
Y	9	Y	724

P-valeur du test de blancheur : 0.0624035. Donc on rejette  $H_0$  et  $E$  ne suit pas un bruit blanc, les colonnes ne sont pas indépendantes et  $\Sigma \neq Id$ .

On voit donc que lorsque l'on applique la structure non paramétrique la P-valeur de 1 nous indique que parmi les fonctions testées c'est la structure non paramétrique qui permet de blanchir la matrice.



# Les 50 Protéines les plus fréquents

## Liste des fréquences

0.9868, 0.981, 0.976, 0.9734, 0.9712, 0.9636, 0.9612, 0.9544, 0.952,  
0.9474, 0.9452, 0.9422, 0.9418, 0.9388, 0.936, 0.9348, 0.9282,  
0.926, 0.925, 0.9228, 0.9212, 0.9192, 0.9168, 0.9162, 0.9134,  
0.9058, 0.9054, 0.9016, 0.9006, 0.898, 0.8976, 0.8972, 0.8966,  
0.895, 0.891, 0.8882, 0.8856, 0.8842, 0.8814, 0.8708, 0.8704,  
0.8692, 0.868, 0.865, 0.863, 0.86, 0.8592, 0.8538, 0.8532, 0.8528

## Liste des noms de protéines

AT1G03030.1, AT1G04480.1, AT1G07985.1, AT1G07985.1,  
AT1G08110.1, AT1G08830.1, AT1G14950.1, AT1G18500.1,  
AT1G18500.1, AT1G20260.1, AT1G20620.1, AT1G22300.1,  
AT1G43170.1, AT1G53750.1, AT1G64110.1, AT1G68010.1,  
AT1G68010.1, AT1G69800.1, AT1G72680.1, AT1G77090.1,  
AT1G77120.1, AT2G02930.1, AT2G25970.1, AT2G32120.1,  
AT2G37660.1, AT2G38380.1, AT3G02530.1, AT3G02560.1,  
AT3G05190.1, AT3G13300.1, AT3G21720.1, AT3G26060.1,  
AT3G53990.1, AT3G54400.1, AT4G02340.1, AT4G16155.1,  
AT4G16210.1, AT4G16760.1, AT4G27520.1, AT4G34670.1,  
AT4G36360.1, AT5G04885.1, AT5G04885.1, AT5G11520.1,  
AT5G20950.1, AT5G26830.1, AT5G42740.1, AT5G52840.1,  
AT5G66190.1, AT5G66190.1

Scatter plot showing the relationship between protein names (X-axis) and temperature categories (Y-axis). The Y-axis is labeled 'Température' and has three categories: CHAUD, FROID, and STD. The X-axis is labeled 'Protéines' and lists seven protein names: AT1G07985.1, AT1G08830.1, AT1G18500.1, AT1G22300.1, AT3G02530.1, AT3G53980.1, and AT4G27520.1. A color scale on the right indicates the 'estim' value, ranging from -1 (blue) to 1 (red).

Protéines	Température	estim
AT1G07985.1	CHAUD	1.0
AT1G08830.1	CHAUD	0.0
AT1G18500.1	CHAUD	-0.5
AT1G22300.1	STD	-0.5
AT3G02530.1	STD	-0.5
AT3G53980.1	STD	-0.5
AT4G27520.1	FROID	-0.5





# Protéines retenues au seuil 0.95

## Table des protéines retenues

---

### Proteines

---

AT1G07985.1

AT1G08830.1

AT1G18500.1

AT1G22300.1

AT3G02530.1

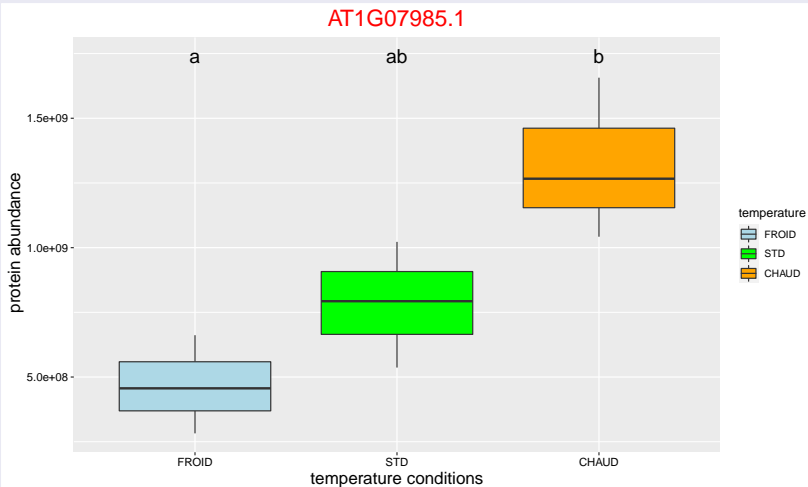
AT3G53990.1

AT4G27520.1

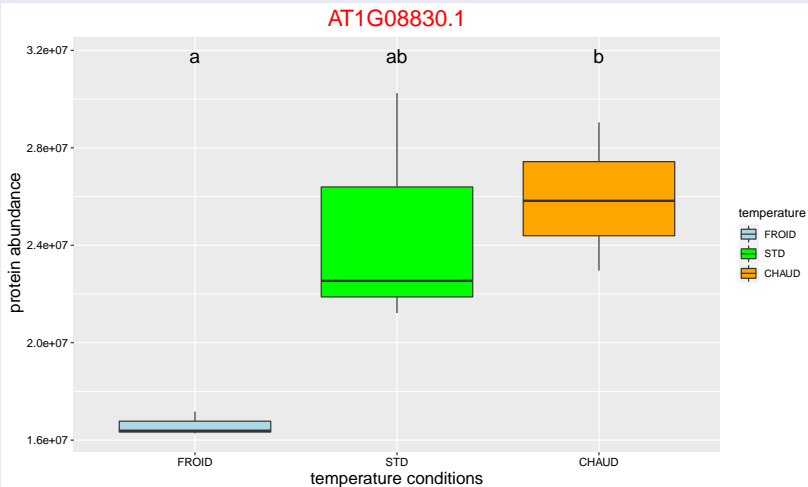
## Boxplots de Tukey (avec cld)

Les graphiques suivants présentent les boxplots pour les protéines sélectionnées par la méthode du lasso

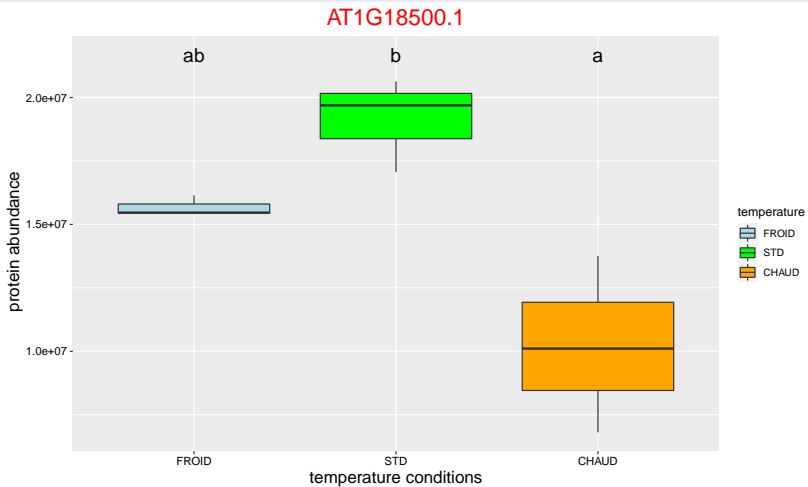
AT1G07985.1



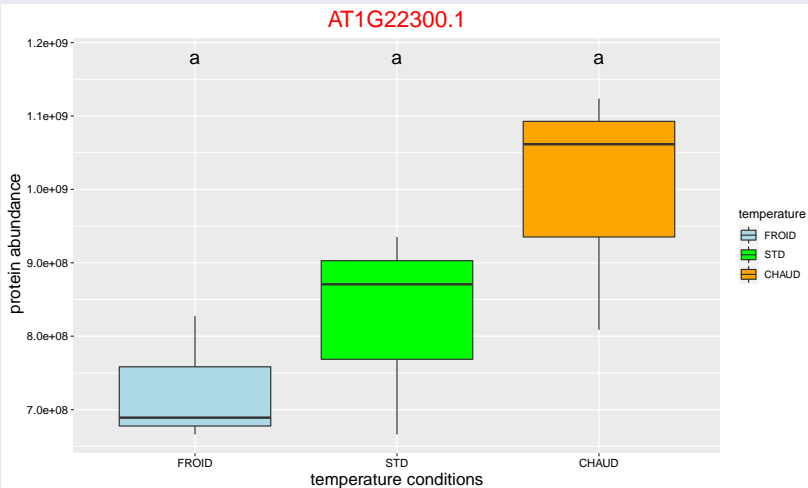
AT1G08830.1



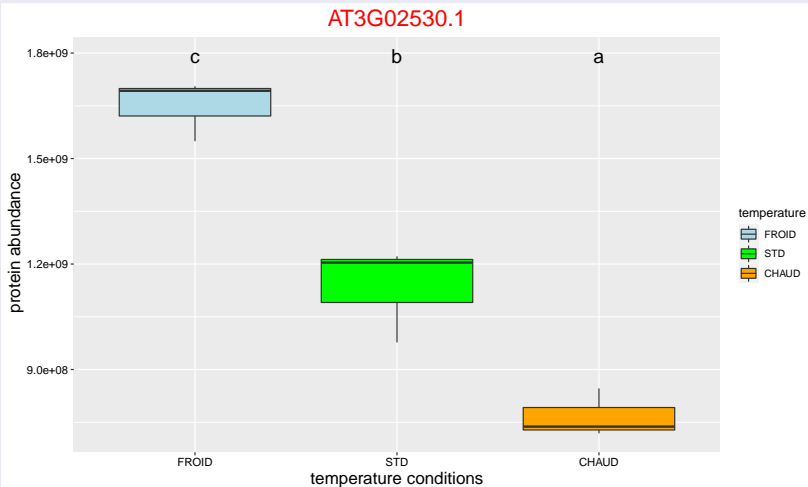
AT1G18500.1



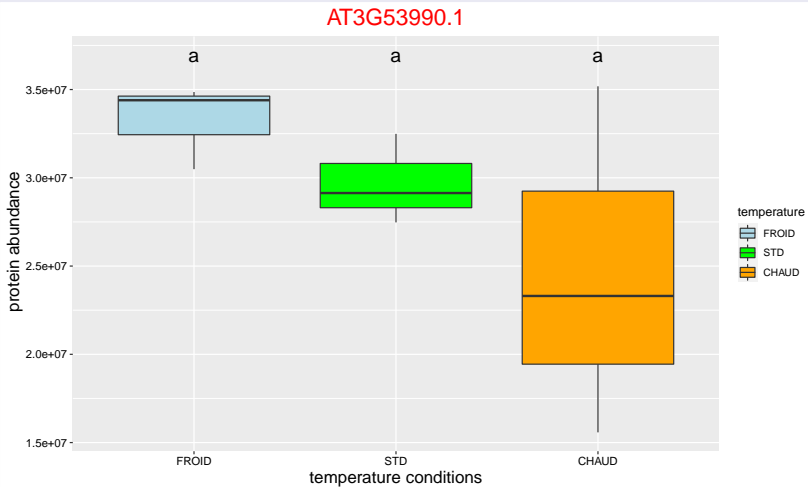
## AT1G22300.1



AT3G02530.1

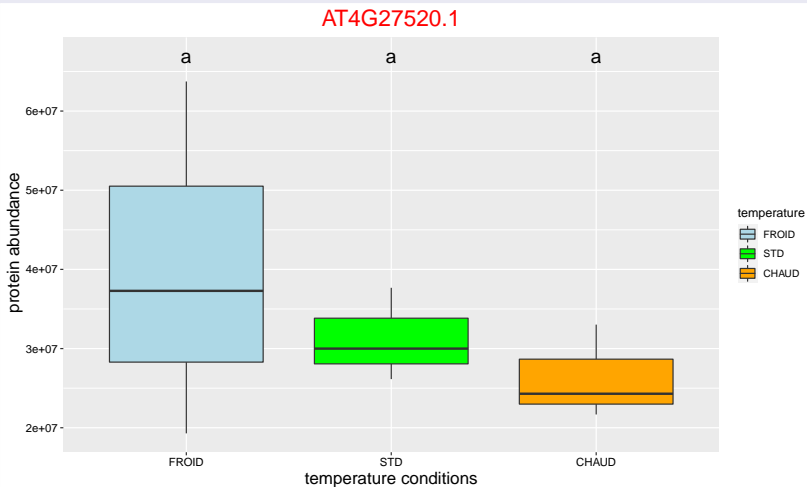


## AT3G53990.1





AT4G27520.1



## Exportation des Protéines sélectionnés par GLM Lasso

Le fichier **proteines\_selection\_lasso\_0.95.csv** existe déjà.

## Métabolites KEGG

## Section 5

Protéines ThaleMine, UniProt