Bertrand Huguenin-Bizot

Master II de Bioinformatique

Université Claude Bernard Lyon I

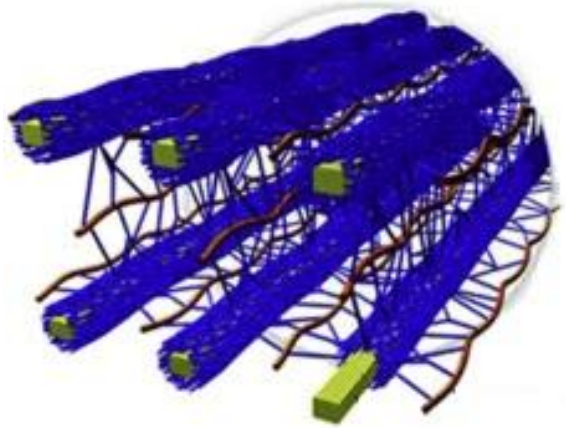
Intégration de données Omics issues de consortia microbiens impliqués dans la dégradation de la lignocellulose

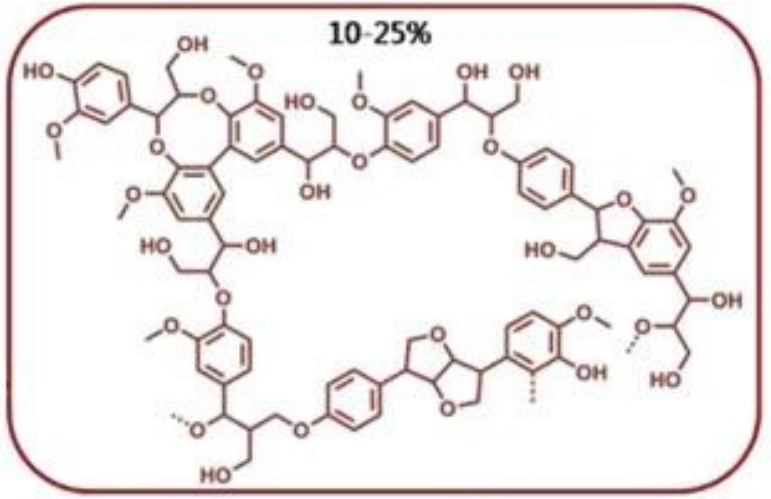
Toulouse Biotechnology Institute - INSA
Équipe Symbiose

Encadré par Guillermina Hernandez-Raquet, Sébastien Dejean et Melisande Albert

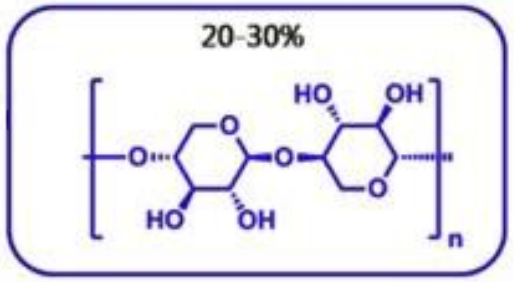
8 décembre 2022

Valorisation de la lignocellulose

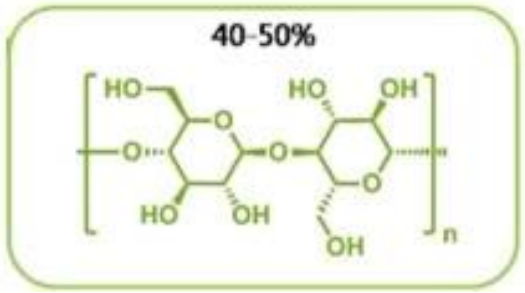




Lignine



Hémicellulose



Cellulose

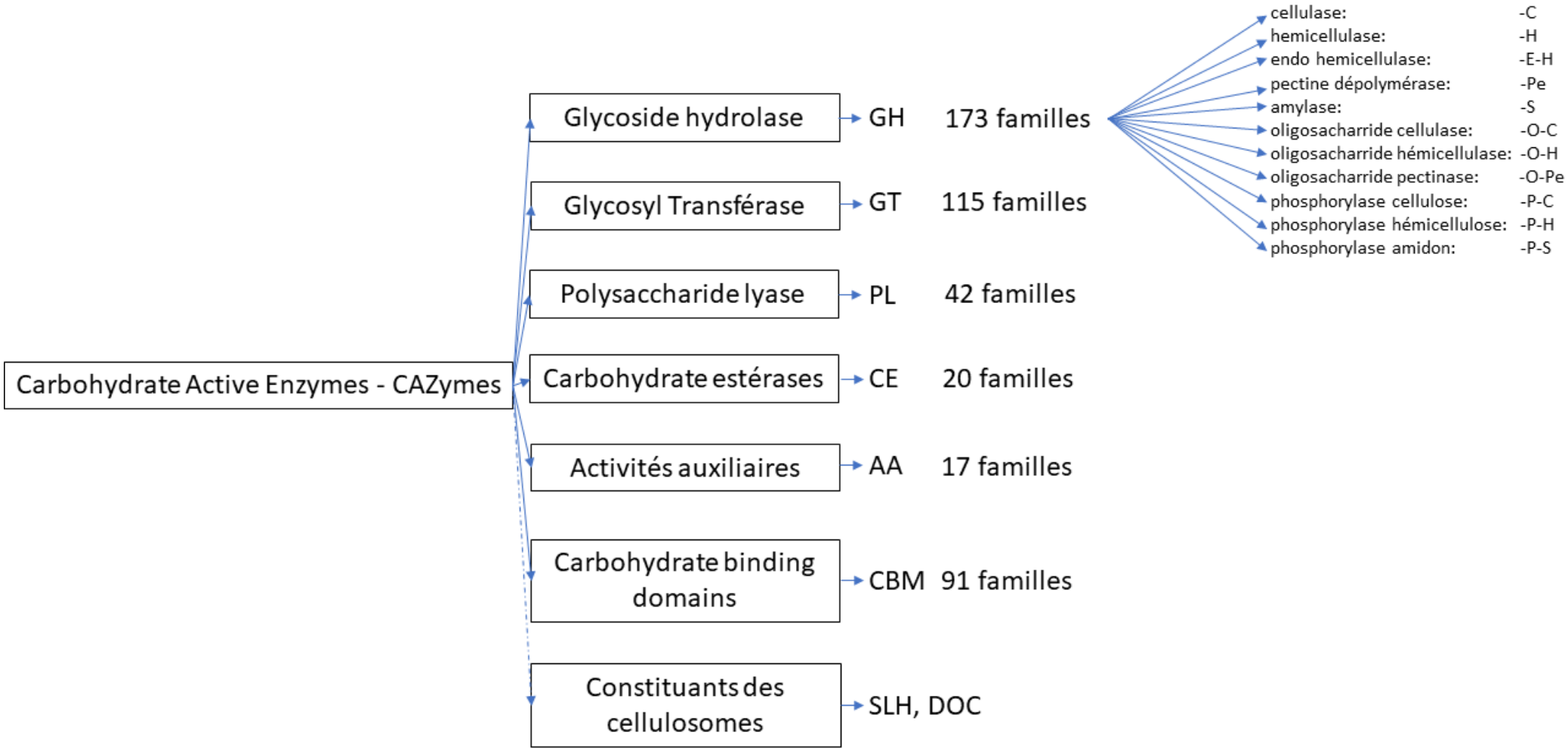
Objectifs

- Utiliser des communautés microbiennes complexes dégradant naturellement la lignocellulose.
  - Rumen
  - Termite

- Comprendre comment les communautés microbiennes évoluent.

- Comprendre comment le pool enzymatique évolue.

Dégradation microbienne - enzymes



Méthode: Digestion anaérobie en réacteurs fermés avec de la paille de blé comme seule source de carbone

Termite

4 temps de prélèvements

T1 T2 T3 T4 T1 T2 T3 T4

Rumen

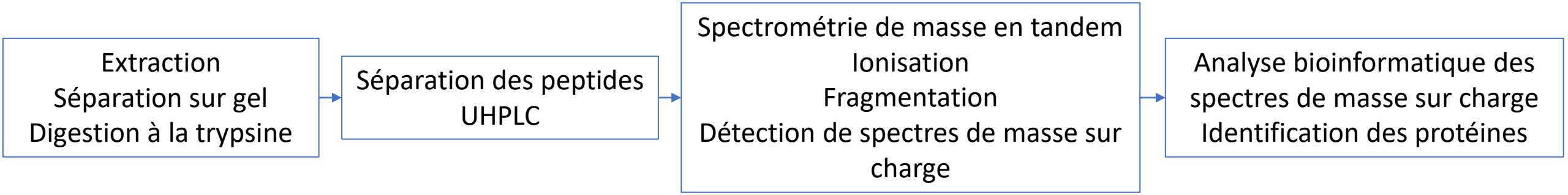
4 temps de prélèvements

T1 T2 T3 T4 T1 T2 T3 T4

2 réplicats biologiques - 3 réplicats techniques

2 x 2 x 4 x 3 = 48 échantillons

Métaprotéomique



Normalized Spectral Abundance Factor

SpC est le nombre de comptage de spectres attribués à une protéine et L la longueur de la protéine.

$$NSAF = \frac{(\frac{SpC}{L})_i}{\sum_{i=1}^{N}(\frac{SpC}{L})_i}$$

Pour un échantillon la somme des abondances NSAF est égale à 1.

$$\sum_i NSAF_i = 1$$

Nous sommes dans le domaine des données compositionnelles

Données compositionnelles

D abondances NSAF de protéines $x_1, x_2, \ldots, x_D$.

Simplex:

$$S^D = \{[x_1, x_2, \ldots, x_D] : x_i > 0 \ (i = 1, \ldots, D), x_1 + \cdots + x_D = 1\}$$

- Travailler dans le simplex: Géométrie d'Aitchison
- Sortir du simplex: Transformation centered log ratio

$$y = clr(x) = \left[\ln\left(\frac{x_1}{g(x)}\right), \ln\left(\frac{x_2}{g(x)}\right), \ldots, \ln\left(\frac{x_D}{g(x)}\right)\right]$$

Avec $g(x) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}$

Imputation des zéros

- 70 % de zéros dans les données initiales.
- Regroupement des protéines au niveau taxonomique ou de la famille de CAZymes: moins de zéros à remplacer
- Exemple de regroupement à la famille de CAZymes, 30% de zéros.

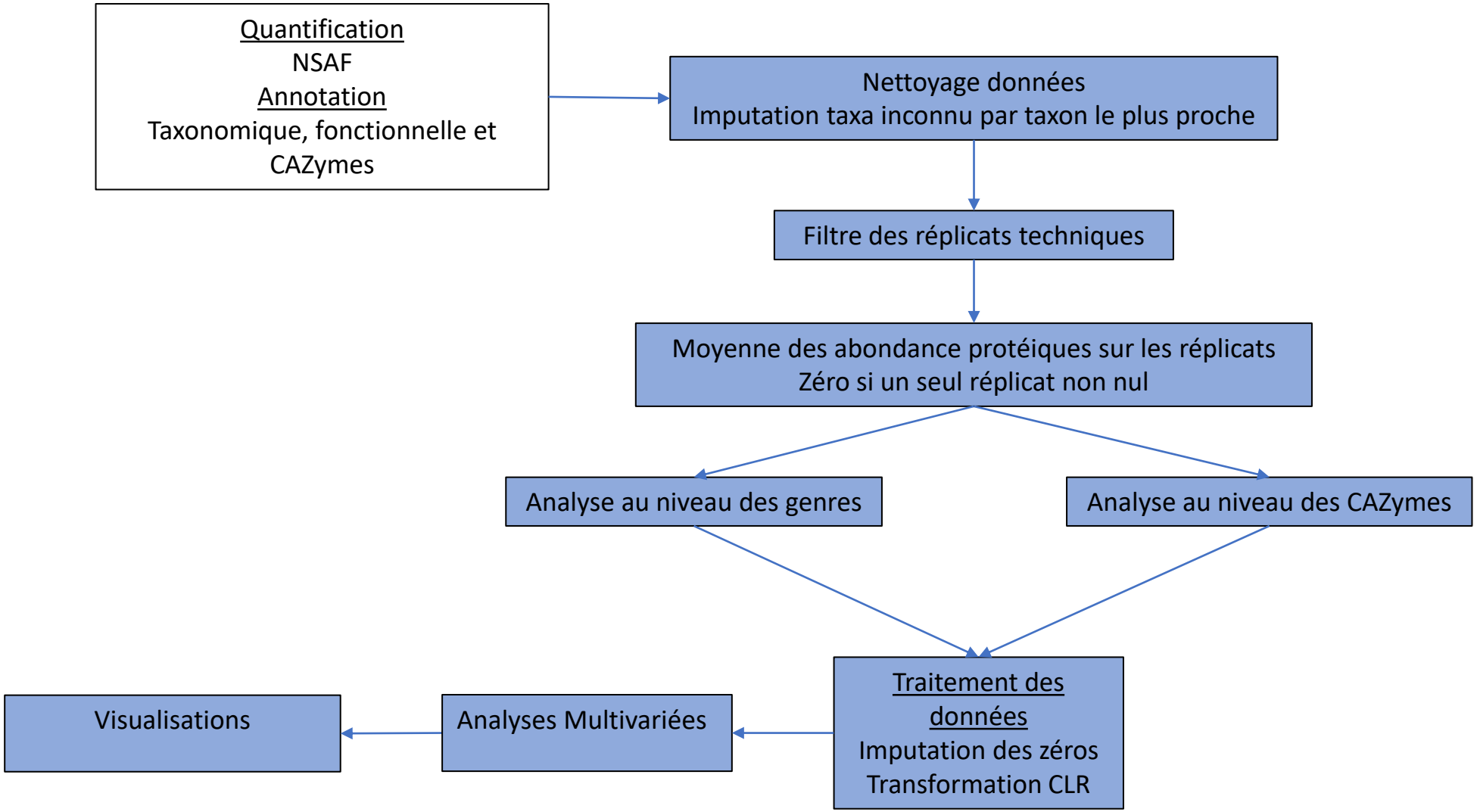
Imputation des zéros

- Imputation des valeurs trop faibles engendre une distorsion de la structure de covariance (Martin-Fernandez J.A. et al., 2003).

- Remplacement des zéros par 2/3 du seuil de détection (Lubbe S. et al, 2021; Martin-Fernandez J.A. et al., 2003).

- Adaptation en utilisant 2/3 de la valeur minimale de chaque variable.

Workflow


Quantification
NSAF
Annotation
Taxonomique, fonctionnelle et CAZymes
Nettoyage données
Imputation taxa inconnu par taxon le plus proche
Filtre des réplicats techniques
Moyenne des abondance protéiques sur les réplicats
Zéro si un seul réplicat non nul
Analyse au niveau des genres
Analyse au niveau des CAZymes
Traitement des données
Imputation des zéros
Transformation CLR
Analyses Multivariées
Visualisations

Analyses multivariées

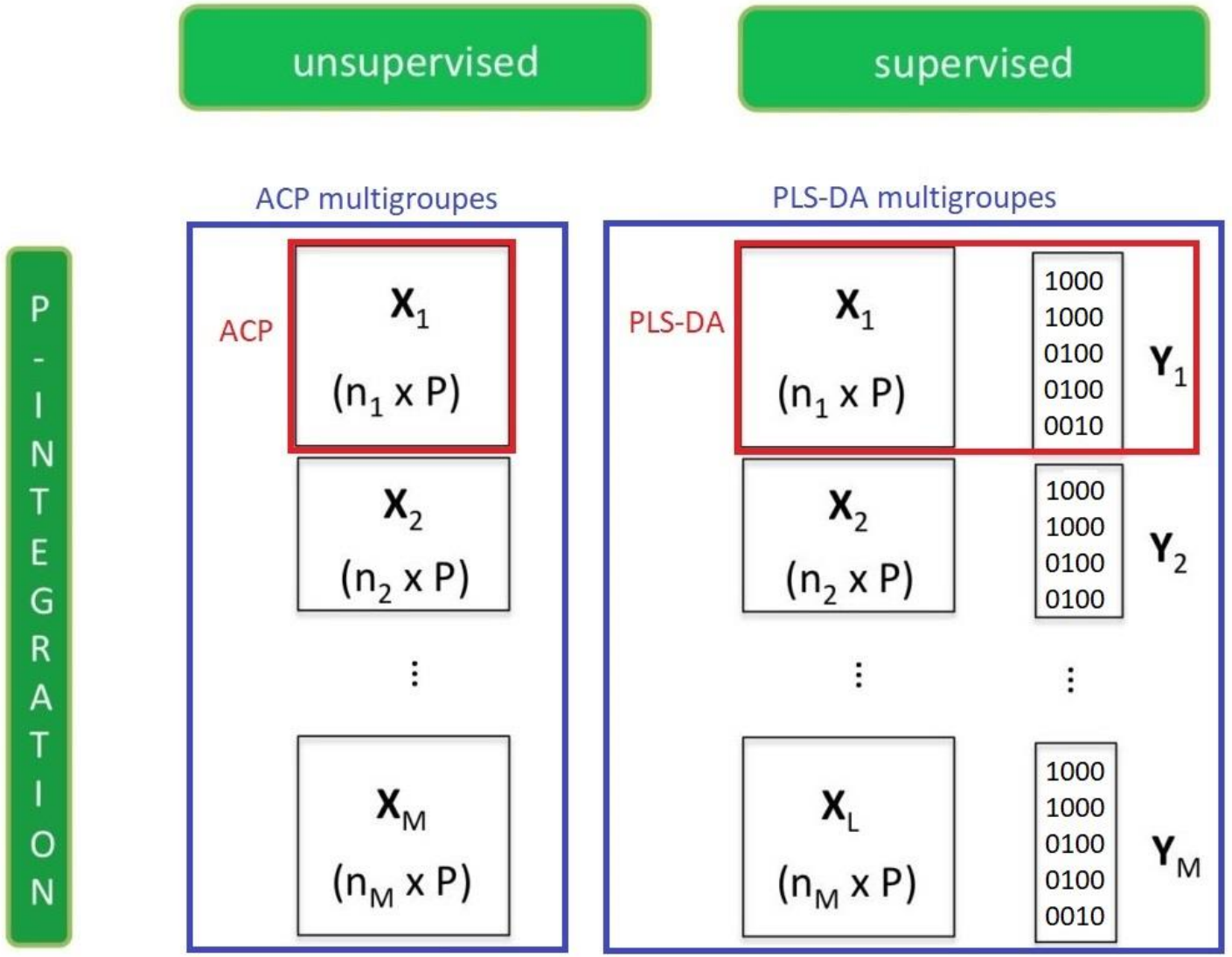
	Prot 1	Prot 2					Prot D
Echantillon 1							
Echantillon 2							
Echantillon n							

$$PC_1 = a_{11}prot_1 + a_{12}prot_2 + \cdots + a_{1D}prot_D$$

$$PC_2 = a_{21}prot_1 + a_{22}prot_2 + \cdots + a_{2D}prot_D$$

	PC1	PC2
Echantillon 1		
Echantillon 2		
Echantillon n		

Analyses multivariées



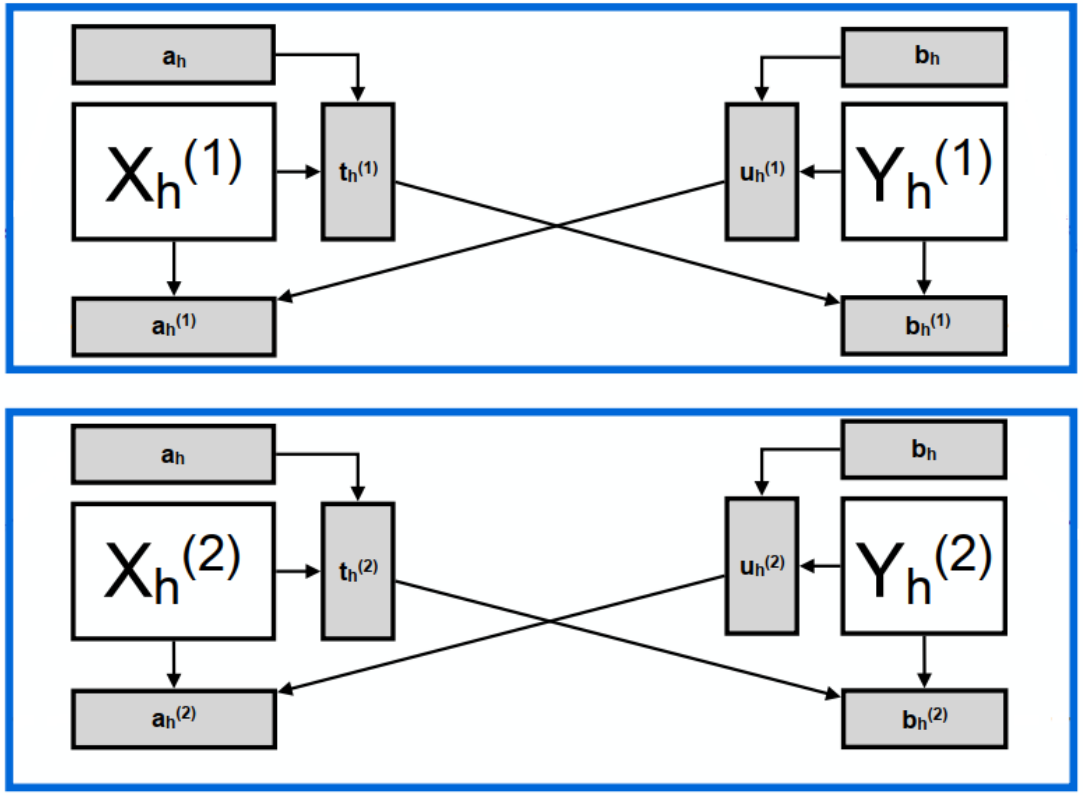
Multivariate INTegrative method - Partial Least Square- Discriminant Analysis
MINT-PLS-DA
Organisation des données

	Protéine 1	Protéine 2	Protéine 3	…	Protéine D		Discriminant Factor
Groupe 1 → Rumen1_T1							T1
Rumen1_T2							T2
Rumen1_T3							T3
Rumen1_T4							T4
Groupe 2 → Rumen2_T1							T1
Rumen2_T2							T2
Rumen2_T3							T3
Rumen2_T4							T4
Groupe 3 → Termite1_T1							T1
Termite1_T2							T2
Termite1_T3							T3
Termite1_T4							T4
Groupe 4 → Termite2_T1							T1
Termite2_T2							T2
Termite2_T3							T3
Termite2_T4							T4

Algorithme de la MINT-PLS-DA

$$t_h^{(m)} = X^{(m)} a_h \text{ et } u_h^{(m)} = Y^{(m)} b_h$$

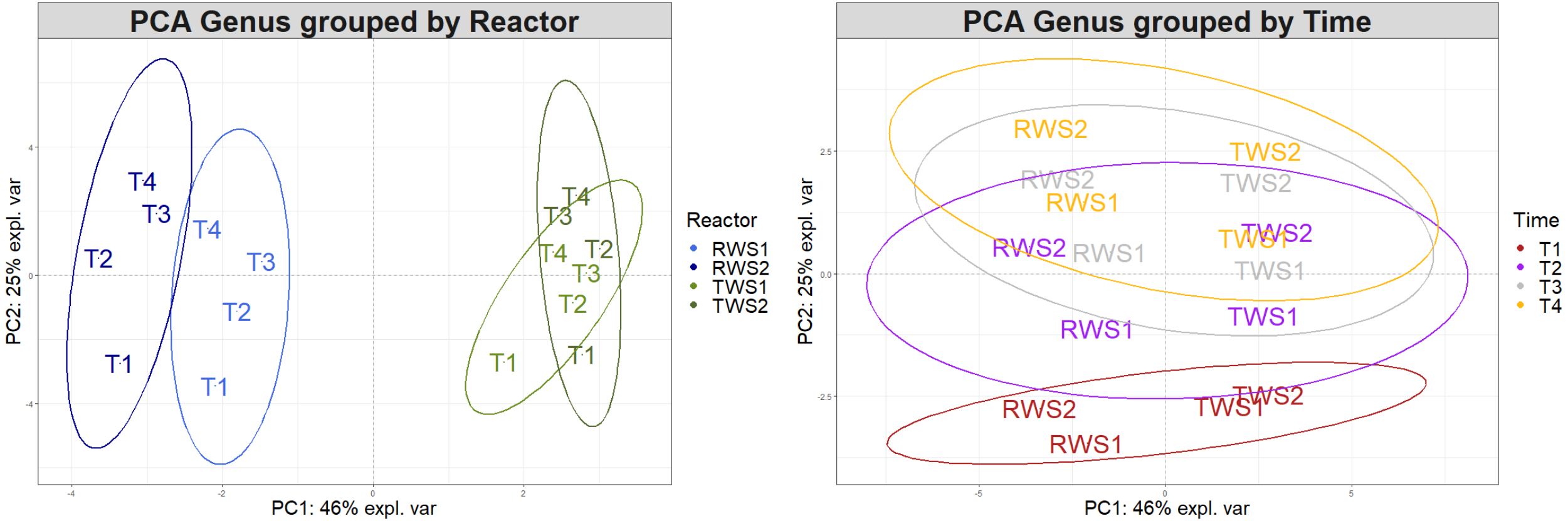
$$\max_{||a_h||_2 = ||b_h||_2 = 1} \sum_{m=1}^{M} n_m cov(X^{(m)} a_h , Y^{(m)} b_h)$$



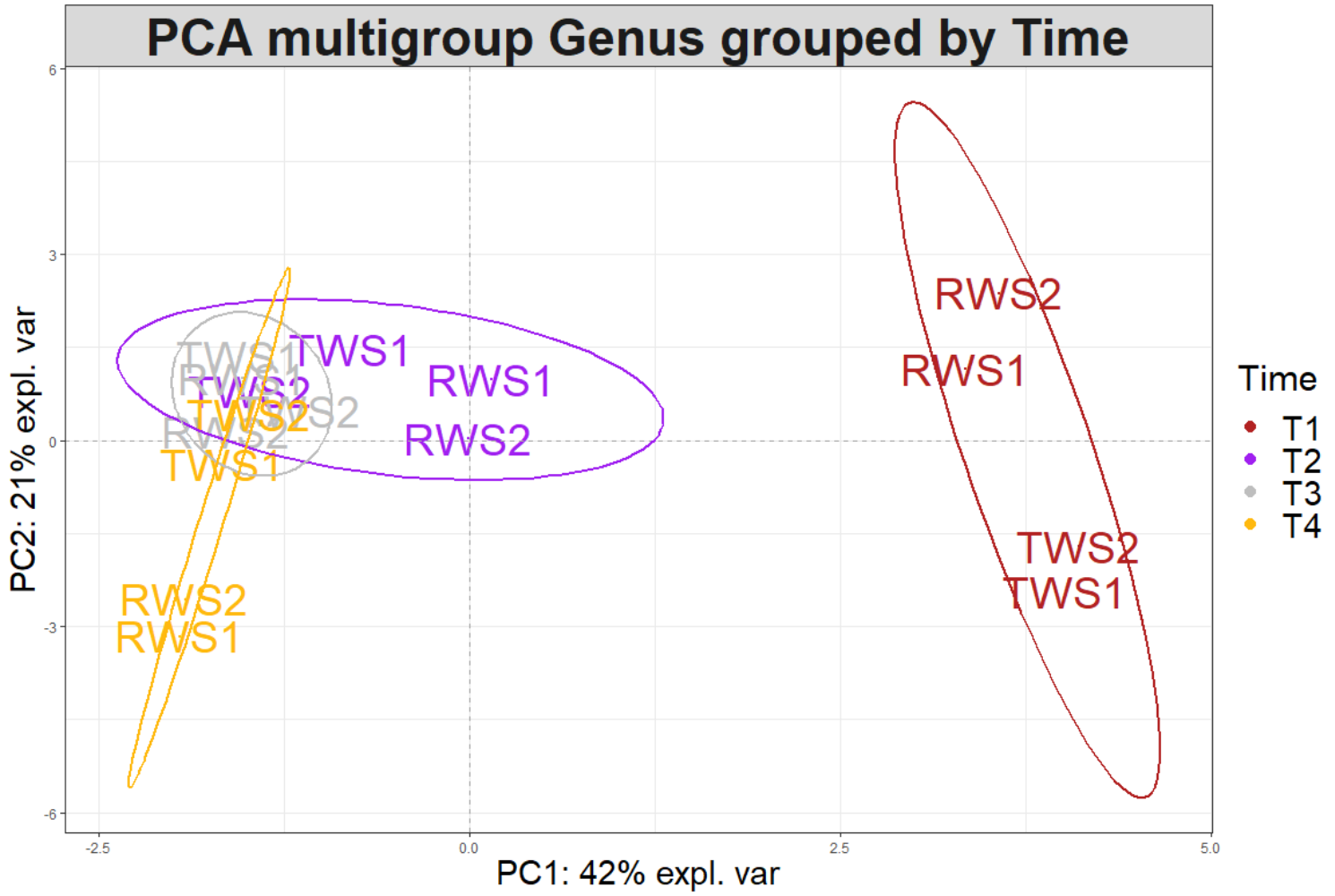
Résultats

- Analyse au niveau taxonomique et des familles de CAZymes
  - Différences entre Inocula: ACP suffisante
  - Différences communes entre inocula: Analyses multigroupes
    - Variabilité due aux temps de prélèvements

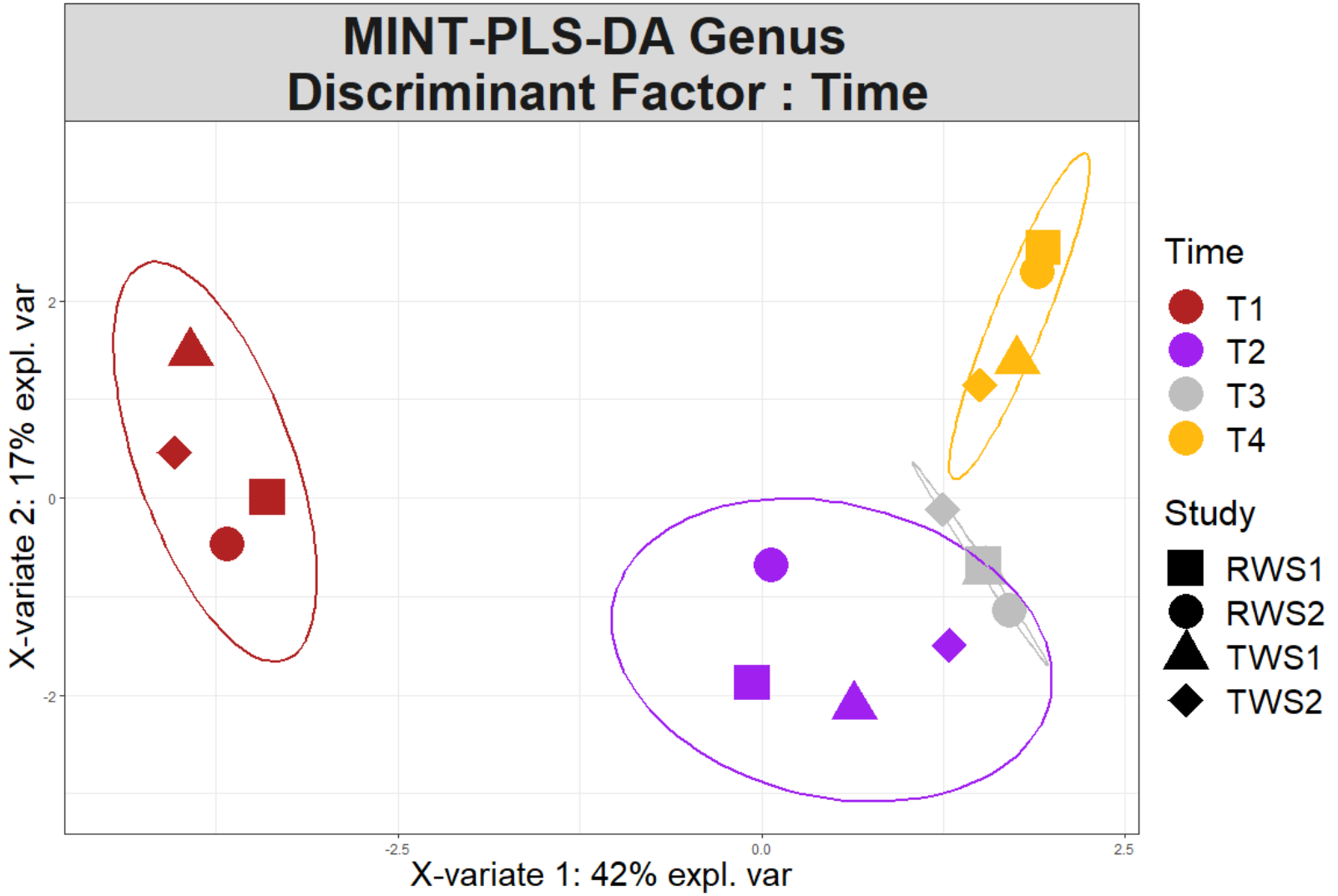
Analyse au niveau taxonomique : variance due aux Inocula



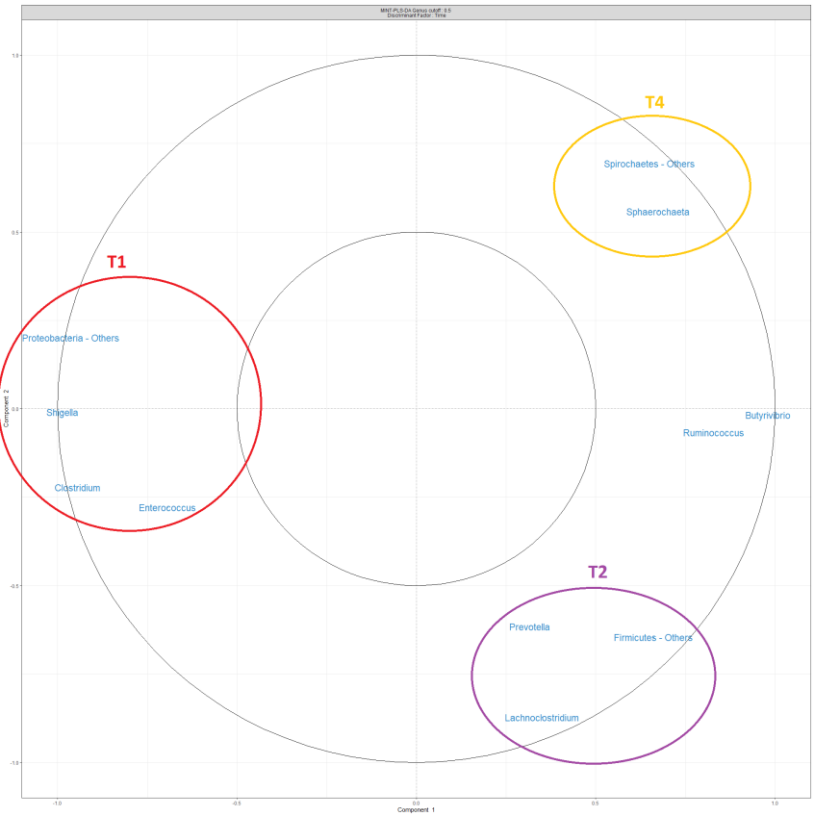
Analyse au niveau taxonomique : variance due aux temps

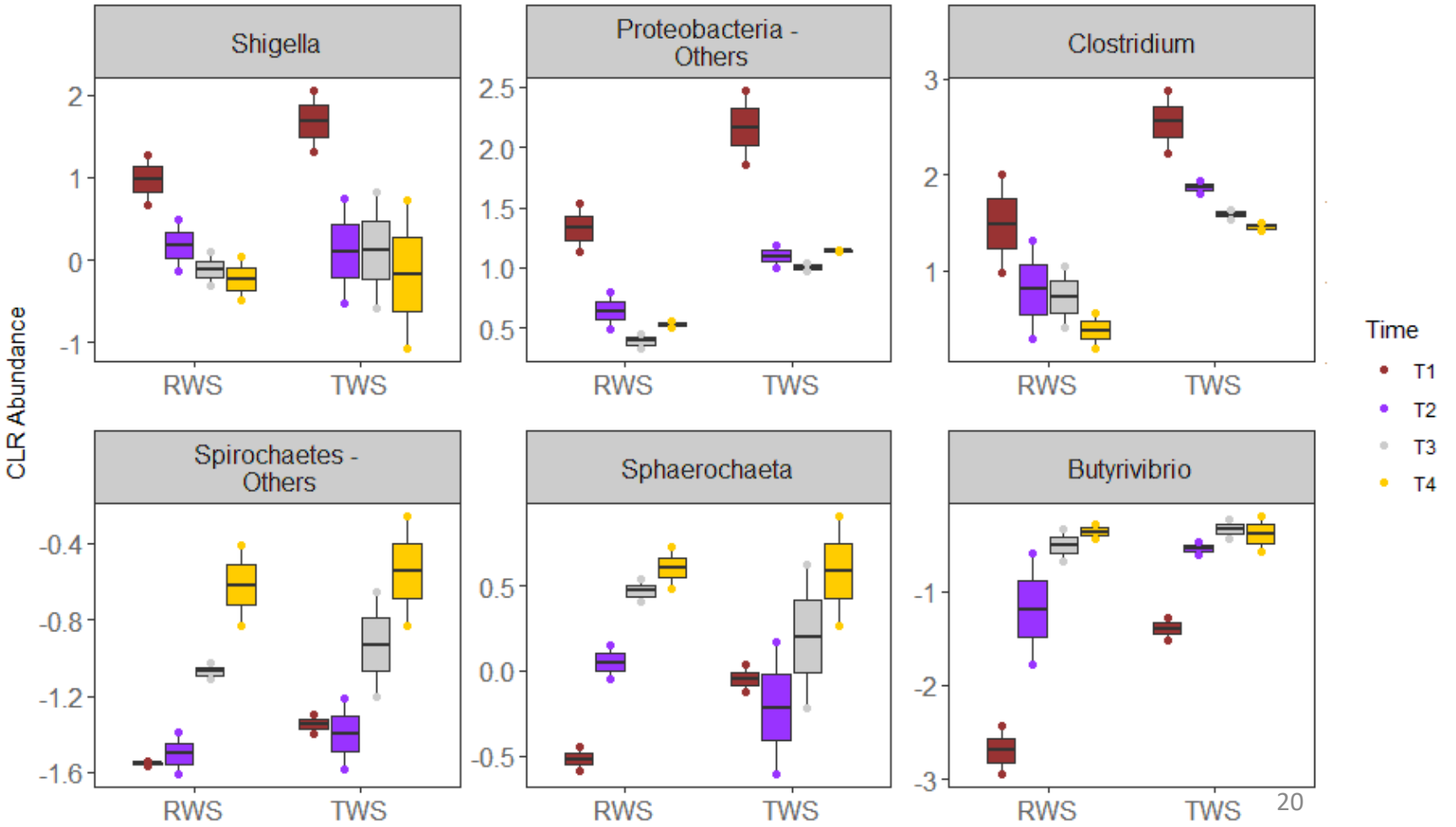


Analyse au niveau taxonomique : MINT-PLS-DA

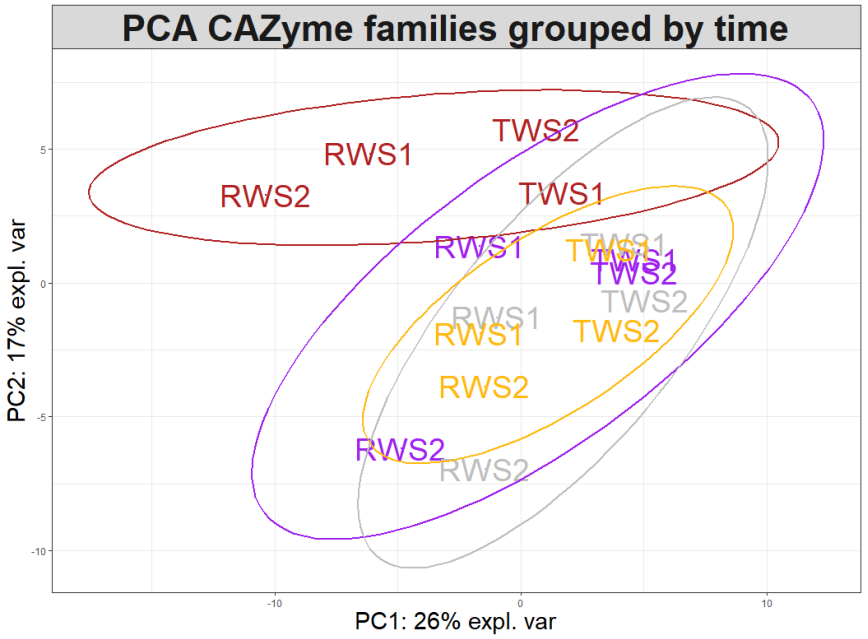


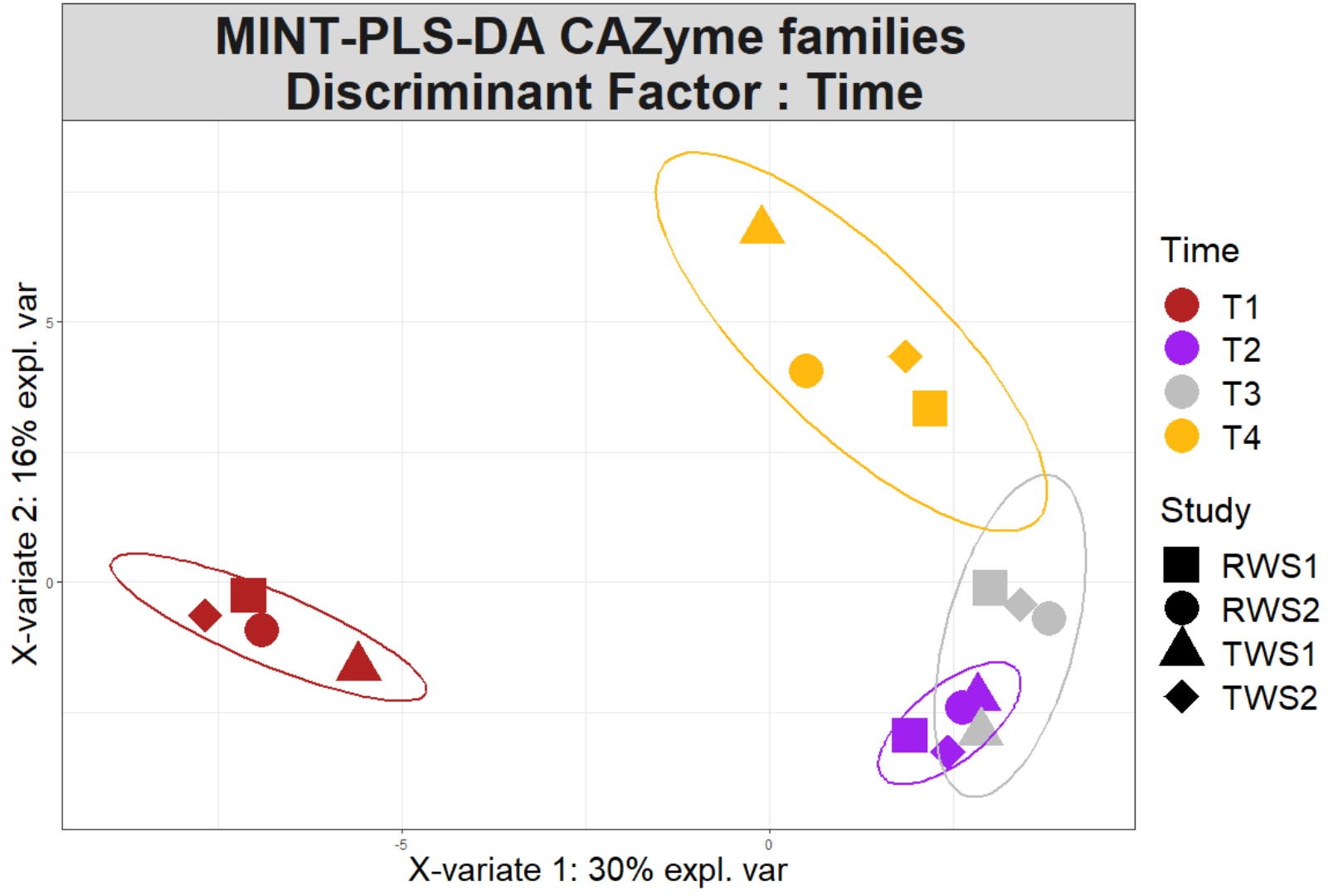
Evolution des taxa discriminants



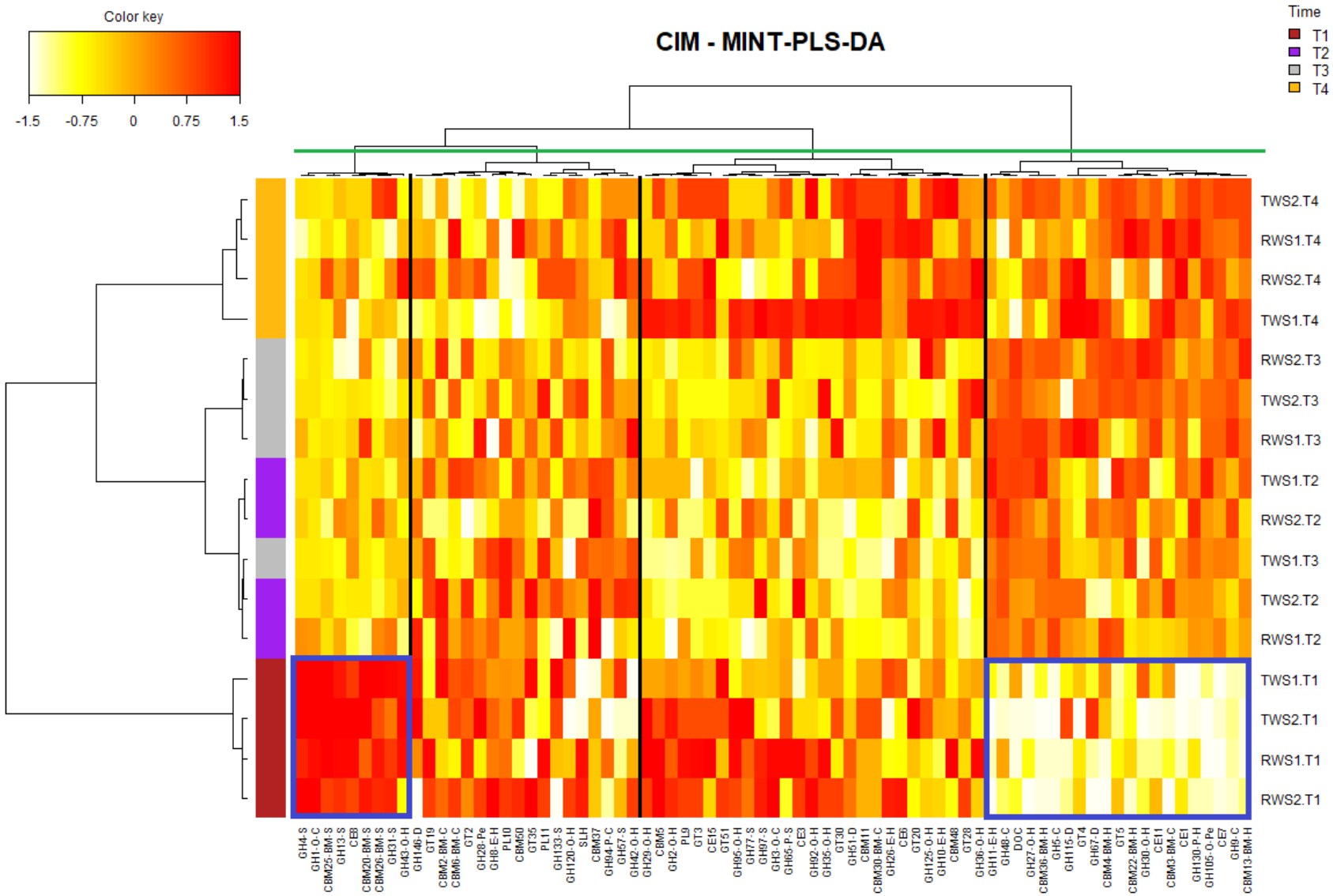


Analyses multivariées au niveau des CAZymes





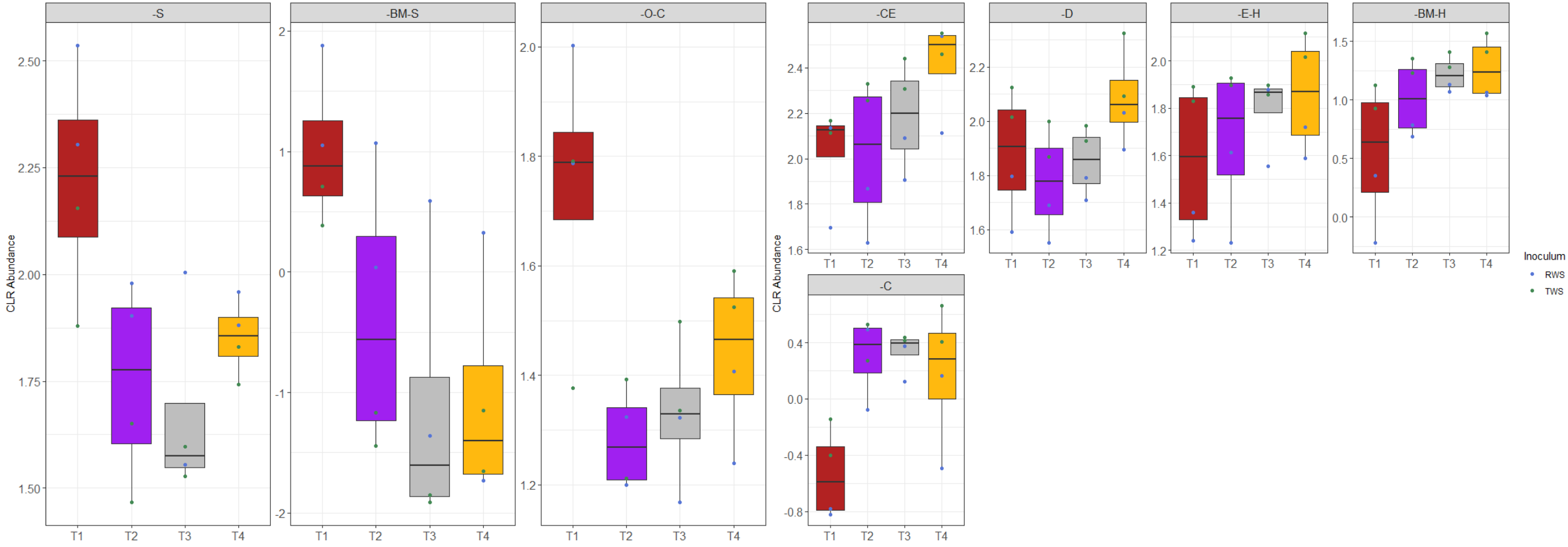
Analyse au niveau des familles de CAZymes: Clustered Image Map



Evolution des fonctions de CAZymes discriminantes

Fonctions décroissantes liées aux sucres libres

Fonctions croissantes liées à la cellulose et à l'hémicellulose



Conclusions et perspectives

L’ACP nous permet de connaître les différences entre inocula.

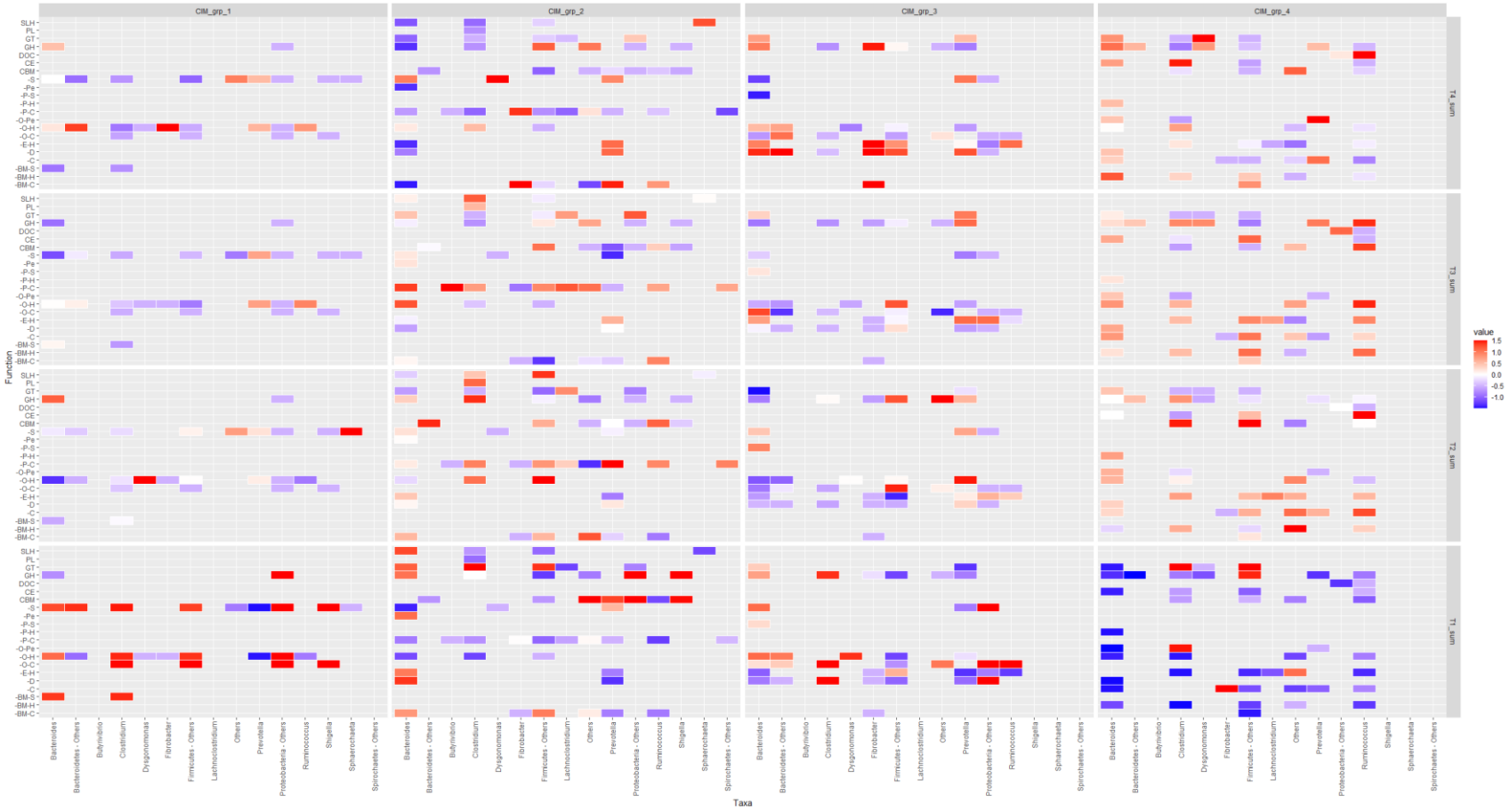
La MINT-PLS-DA: différences communes aux inocula entre les temps.

Nous identifions:

- L’évolution des taxa discriminants au cours du temps
- L’évolution des familles de CAZymes au cours du temps
- une diminution d'abondances des protéines impliquées dans la dégradation des sucres libres
- une augmentation d'abondances des protéines impliquées dans la dégradation de l'hémicellulose et de la cellulose.

Intégration de données issues de métagénomique sur les mêmes échantillons.

Heatmap des abondances des genres centrées réduites selon les fonctions, les groupes CIM et les temps



Heatmap des abondances des genres centrées réduites selon les fonctions au temps 1 et dans le groupe CIM1