Introduction à l'analyse des données compositionnelles

Paul Dou & Simon Roques

Équipe Digestion, Nutrition, Aliments, Métabolisme, Microbes (DINAMIC)
Unité Mixte de Recherche sur les Herbivores (UMRH)
INRAE Centre Clermont-Auvergne-Rhône-Alpes

2025-11-26

Objectifs et organisation de la présentation

1. Reconnaitre une donnée compositionnelle
2. Probléme en analyse statistique liés aux données compositionnelle.
3. Transformation des données pour l'analyse statistique
4. Cas des analyses NGS d'abondance differentielle (NGS):

Définition d'une donnée compositionnelle[1]

Définition (Aitchison)

Une **donnée compositionnelle** est un vecteur $\mathbf{x} = (x_1, \dots, x_D)$ de valeurs positives,
dont seule l'information **relative** entre les composantes est pertinente.
Le vecteur est **contraint** à une somme constante κ et vit dans le *simplexe* :

$$S^D = \left\{ \mathbf{x} \in \mathbb{R}^D \;\middle|\; x_i > 0,\ \sum_{i=1}^{D} x_i = \kappa \right\}.$$

[1] J.Aitchison (1982). "The Statistical Analysis of Compositional Data". Journal of the Royal Statistical Society.

Exemples selon la constante de fermeture κ

Type de données	Constante (κ)	Exemple d'unité
Proportions	1	fractions (ex. 0.2, 0.3, 0.5)
Pourcentages	100	20 %, 30 %, 50 %
CPM (*Counts Per Million*)	1 000 000	données normalisées de séquençage (RNA-seq, microbiome)

Exemple:

Animal	Species1	Species2	Species3	Total
A	0.1	0.4	0.5	1
B	0.2	0.3	0.5	1
C	0.4	0.4	0.2	1
D	0.5	0.3	0.2	1

En colonnes:

- Taxon
- OTU : Operational Taxonomic Unit
- ASV : Amplicon Sequence Variants

Reconnaitre une donnée compositionnelle

Considérations pratiques

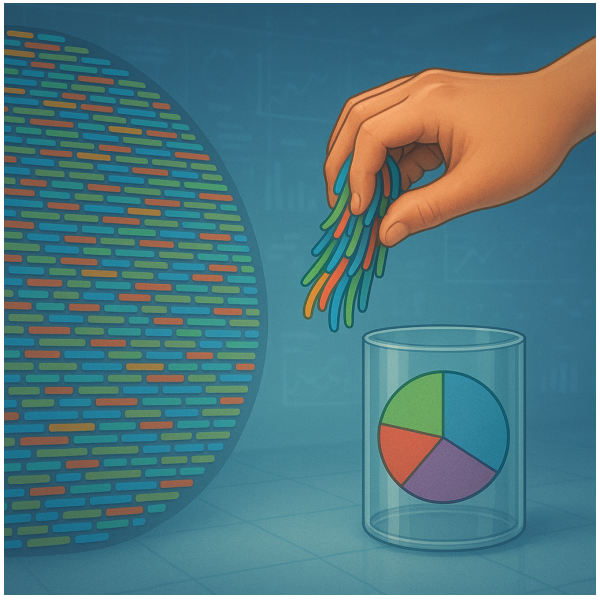
Contexte / Question posée	Total important ?	Unités pertinentes ?	Type d'échelle recommandée	Conclusion à en tirer
Comparer la propreté de l'eau	Oui	Oui	**Échelle absolue réelle positive**	Travailler sur les **concentrations absolues** (mg/L, mol/L) pour évaluer la charge ionique.
Microbiome (NGS, comptages de séquences)	**Non (profondeur $\neq$ abondance réelle)**	Non	**Échelle de comptages compositionnels** ou **composition continue normalisée**	Les comptages NGS sont des **données compositionnelles** : seule l'information relative entre taxons est exploitable.

Sequençage NGS: Cas du metabarcoding et de la metagenomique

Protocol:

Séquençage NGS: Cas du metabarcoding et de la métagénomique

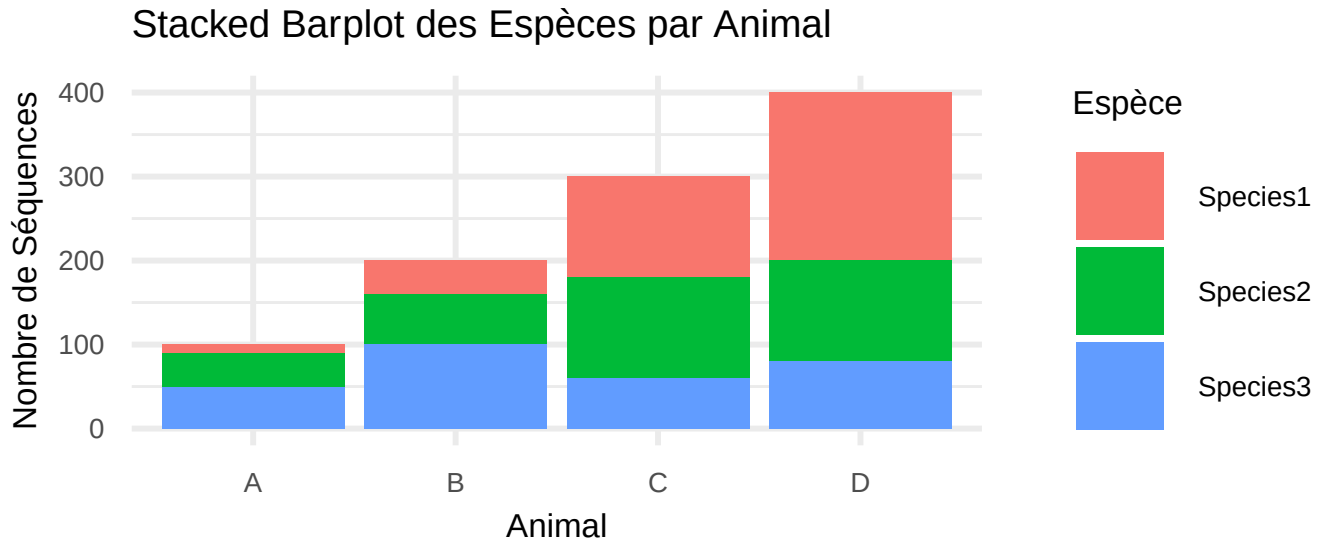
Protocol:



Le comptage total par échantillon peut varier :

- l'efficacité de la PCR,
- la qualité des réactifs,
- ou des variations dans l'instrumentation.

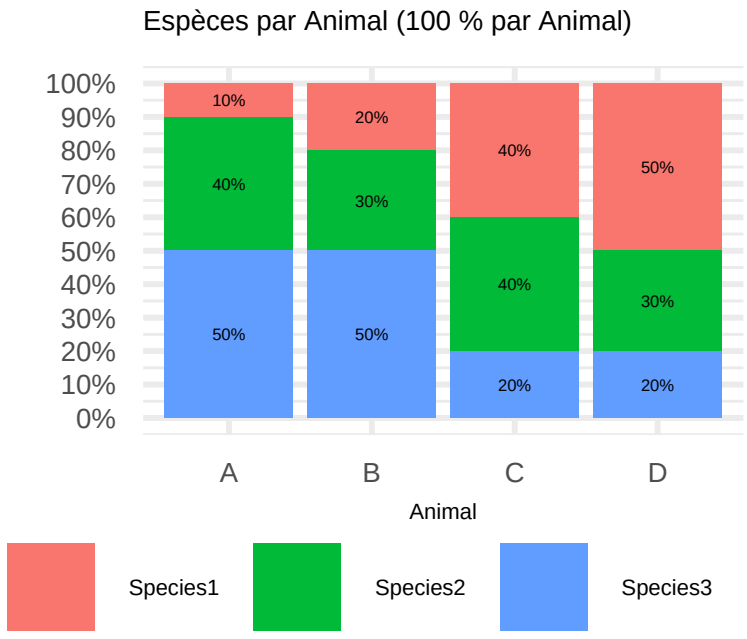
Problèmes liées au metabarcoding



Comparaison directe des comptages bruts impossible..

Proportions

- Chaque barre représente un échantillon (A, B, C, D).
- L'axe des y est en pourcentages



- **Application: Regression, PLS, ACP ?**

Problèmes avec les proportions

La distance euclidienne

Animal	Species1	Species2	Species3	Total
A	0.1	0.4	0.5	1
B	0.2	0.3	0.5	1
C	0.4	0.4	0.2	1
D	0.5	0.3	0.2	1

- Utilisation classique de la distance euclidienne :

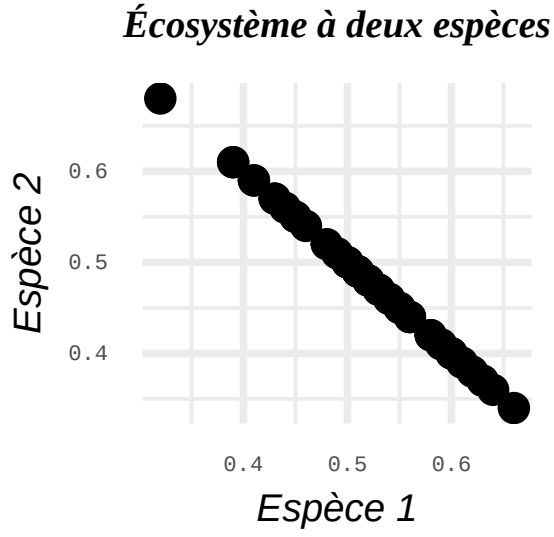
$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{D} (x_i - y_i)^2}$$

$$d(\mathrm{A}, \mathrm{B}) = \sqrt{(0.1-0.2)^2 + (0.4-0.3)^2 + (0.5-0.5)^2} = \sqrt{0.01+0.01+0} = \sqrt{0.02} \approx 0.141$$

$$d(\mathrm{C}, \mathrm{D}) = \sqrt{(0.4-0.5)^2 + (0.4-0.3)^2 + (0.2-0.2)^2} = \sqrt{0.01+0.01+0} = \sqrt{0.02} \approx 0.141$$

Biais de négativité

- Exemple: Ecosysteme à 2 especes. Quand la proportion de l'une augmente l'autre diminue.



Prenons la covariance de x_1 avec la contrainte de somme unitaire :

$$\mathrm{Cov}(x_1, x_1 + x_2 + \cdots + x_D) = \mathrm{Cov}(x_1, 1).$$

$$\mathrm{Cov}(x_1, 1) = 0,$$

$$\mathrm{Cov}(x_1, x_1 + x_2 + \cdots + x_D) = \mathrm{Cov}(x_1, x_1) + \sum_{j=2}^{D} \mathrm{Cov}(x_1, x_j).$$

$$\mathrm{Var}(x_1) + \sum_{j=2}^{D} \mathrm{Cov}(x_1, x_j) = 0.$$

Interprétation des corrélations [2]

Composition or subcompostion	*Crude correlation between parts* AB	AD	AE	BD	BE	DE
ABCDE	0.51	−0.05	−0.16	−0.50	−0.56	−0.22
ABDE	−0.92	0.65	0.61	−0.78	−0.79	0.30
ABD	−0.94	0.65		−0.87		
ADE		−0.60	−0.67			−0.20
BDE				−0.86	−0.85	0.46

Attention : Dans ce tableau, A, B, C, D, E représentent des variables (espèces, OTU, ASV...). Par contre, dans la slide 3, ces lettres représentent les échantillons.

[2] J.Aitchison (1982). "The Statistical Analysis of Compositional Data". Journal of the Royal Statistical Society.

Conclusion

- Les méthodes statistiques usuelles sont inapplicables.
- Comment représenter les données compositionnelles dans un espace euclidien ?
  - Transformation des données pour sortir de l'espace du simplexe.

Simplexe

Le simplexe comme espace vectoriel

Une composition vit dans le **simplexe** :

$$\mathcal{S}_D = \{x_i > 0, \sum x_i = \kappa\}$$

On définit :

$$x \oplus y = \mathcal{C}[x_i y_i], \quad \alpha \odot x = \mathcal{C}[x_i^\alpha]$$

où :

- $\oplus$ perturbation (analogue de l'addition)
- $\odot$ puissance (analogue de la multiplication par un scalaire)
- $\mathcal{C}$ fermeture (normalisation telle que $\sum x_i = 1$)

- Produit scalaire d'Aitchison :

$$\langle x, y \rangle_a = \frac{1}{2D} \sum_{i,j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$$

Transformations log-ratio et propriétés

Transformation	Isométrie	Formule	Remarques
ALR (Additive log-ratio)	Non	$\text{alr}(x) = \ln(x_i/x_D)$	dépend du dénominateur choisi
CLR (Centered log-ratio)	Oui	$\text{clr}(x) = \ln(x_i/g(x))$	somme nulle, dimension D
ILR (Isometric log-ratio)	Oui	$\text{ilr}(x) = \text{clr}(x)\Phi^T$	base orthonormale dans $\mathbb{R}^{D-1}$

Distance d'Aitchison

La **distance d'Aitchison** :

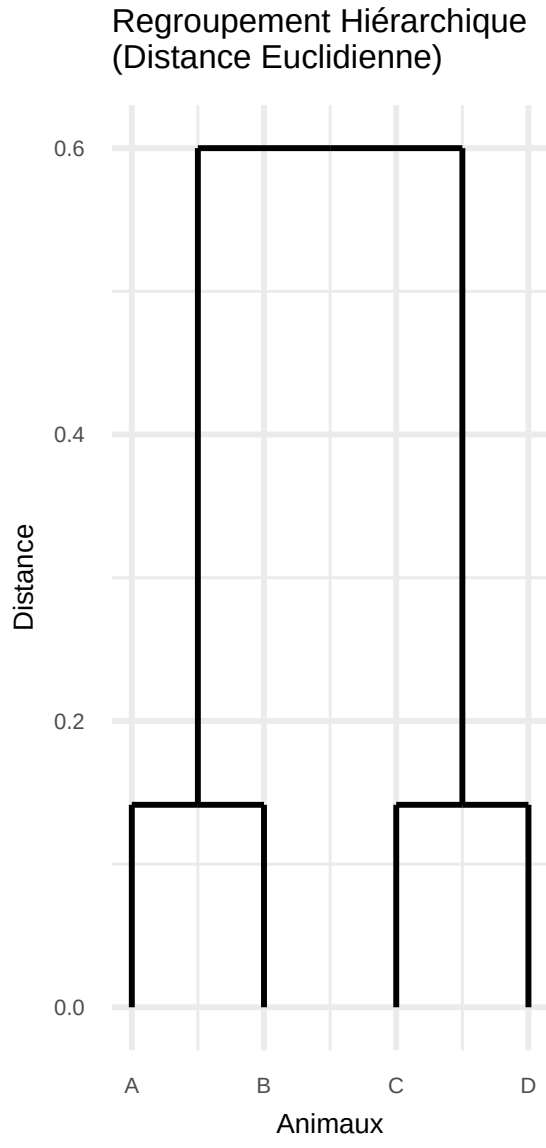
$$d_a(x,y) = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left[\ln\left(\frac{x_i}{x_j}\right) - \ln\left(\frac{y_i}{y_j}\right)\right]^2}$$

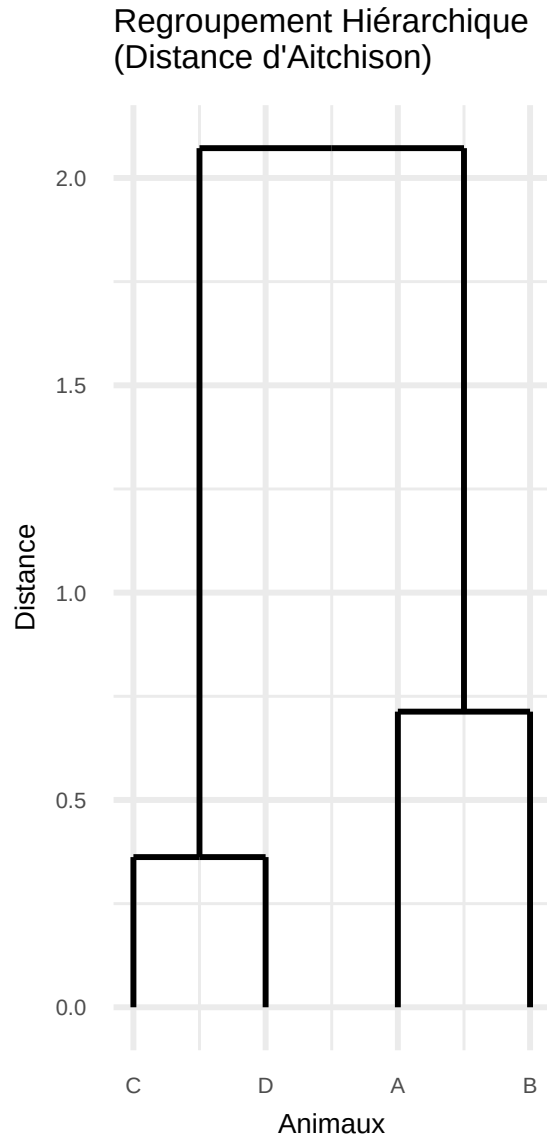
où $x, y \in \mathcal{S}^D$.

Équivalence euclidienne

La distance d'Aitchison correspond à la distance euclidienne dans l'espace transformé :

$$d_a(x, y) = \|\mathsf{clr}(x) - \mathsf{clr}(y)\|_2 = \|\mathsf{ilr}(x) - \mathsf{ilr}(y)\|_2$$


Regroupement Hiérarchique
(Distance Euclidienne)
Distance
0.6
0.4
0.2
0.0
A
B
C
D
Animaux

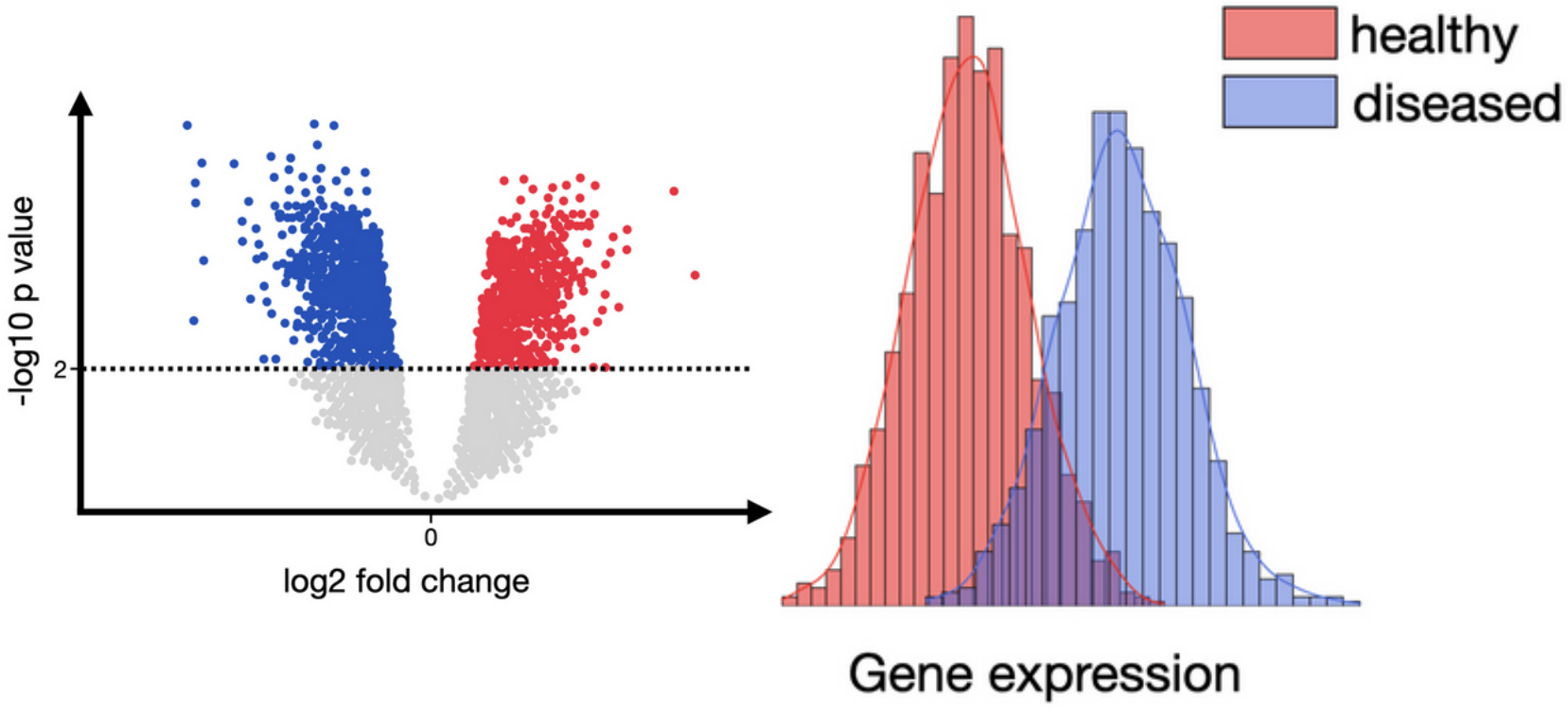

Regroupement Hiérarchique
(Distance d'Aitchison)
Distance
2.0
1.5
1.0
0.5
0.0
C
D
A
B
Animaux

Gestion des zéros en CoDA

Empêchent les transformations log-ratio (CLR, ALR, ILR)
Stratégies d'imputation :

- **Simple** : valeur aléatoire selon distribution triangulaire entre 0 et $x_{\min}$
  - $E(x) = \frac{2}{3} x_{\min}$
- **Méthodes itératives**
  - Implémentées dans **zCompositions** et **robCompositions**

Analyse différentielle : Cas de la variable dépendante



Comment expliquer la variance des composantes des données compositionnelles à partir d'une ou plusieurs covariables ?

Analyse différentielle

Objectifs de l'analyse

- Identifier les principaux facteurs influençant la variable (Log-ratio de taxon/ espece /genre ...).
- Vérifier la significativité statistique à l'aide de tests statistiques (Wilcoxon, Kruskal–Wallis, régression, etc.).

$$\text{Approche fréquente:} Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i$$

- Y_i : variable dépendante (Log-ratio de taxon/ espece /genre ...).
- β_0 : Intercept.
- β_j : Coefficients de régression associés aux variables explicatives X_{ij}.
- ϵ_i : Terme d'erreur aléatoire.

Log-Ratio :
ANCOM
(Analysis of Composition of Microbiomes)
Mandal et al. (2015)
Transformation ALR
(Additive Log Ratio)

(Log-Ratio) Analyse différentielle : ANCOM[3]

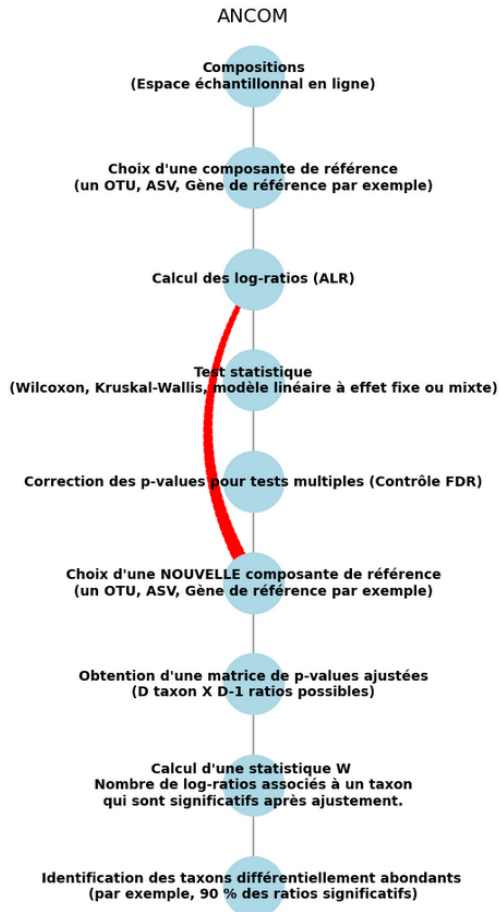
- Développée pour analyser les données compositionnelles issues du microbiome.
- Comparaison de plusieurs classes.
- Ajustement des résultats en fonction de covariables
- Gestion des contraintes compositionnelles (ALR : Additive Log Ratio)

La transformation ALR (Additive Log Ratio) consiste à prendre le logarithme du ratio d'une variable d'intérêt par rapport à une variable de référence.

$$\text{ALR}(x_i) = \log\left(\frac{x_i}{x_j}\right) \quad \text{où} \quad x_j \text{ est une variable (OTU...) de référence}$$

[3]Siddhartha Mandal et al. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial Ecology in Health and Disease.

(Log-Ratio) Analyses differentielles : ANCOM[4]



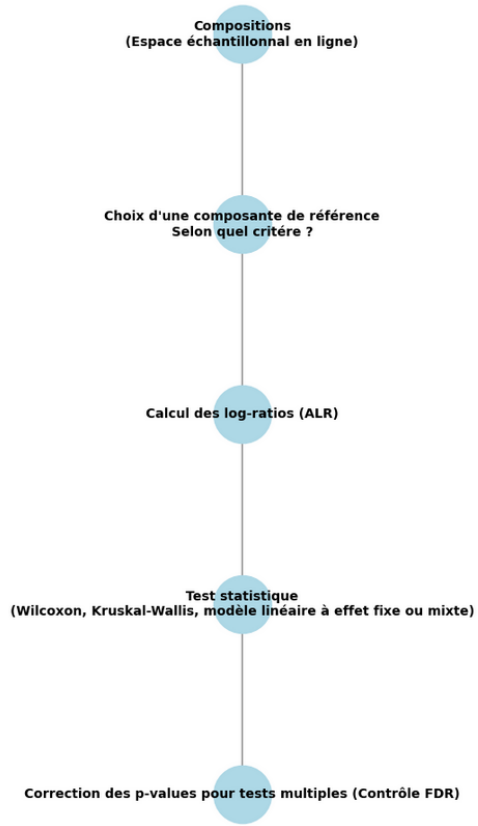
[4]Siddhartha Mandal et al (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial Ecology in Health and Disease.

ANCOM

- Sensibilité réduite sur de petits ensembles de données (< 20 échantillons par groupe).
- **Coût computationnel élevé (temps de calcul long)** :
  - La transformation ALR est appliquée avec chaque taxon comme référence, ce qui peut allonger le temps de calcul.
- Absence de p-valeurs spécifiques à chaque taxon:
  - Aucune erreur standard ni intervalle de confiance disponible pour chaque variable.

Autre possibilité pour détecter des ALR différentielles : choisir une seule composante de référence.
Mais comment la déterminer ?

Régression linéaire (Log-Ratio) après transformation ALR : composante de référence unique[5]



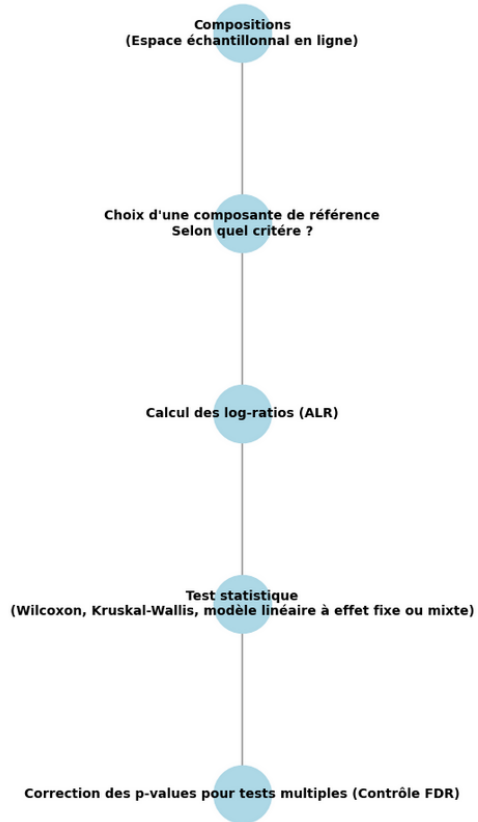
[5] Michael Greenacre et al. (2021). Compositional Data Analysis of Microbiome and Any-Omics Datasets: A Validation of the Additive Logratio Transformation. Microbial Ecology in Health and Disease. Frontiers in Microbiology

Régression linéaire (Log-Ratio) après transformation ALR : composante de référence unique[6]

- Comment choisir la référence ?
  - Rappel :

$$\mathrm{ALR}(x_i) = \log\left(\frac{x_i}{x_{\text{ref}}}\right)$$

  - Minimisation de la variance de $\log(x_{\text{ref}})$.
  - Faible corrélation ou dépendance de x_{ref} avec les variables explicatives ou covariables.
  - Éviter les valeurs proches de zéro ou nulles.
  - Respect de la géométrie exacte des log-ratios.



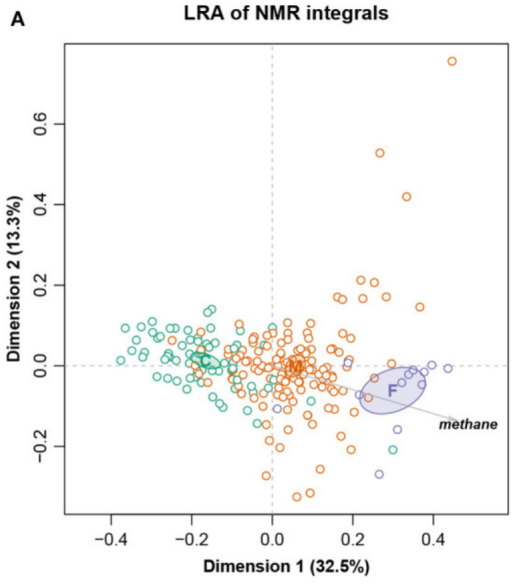
[6] Michael Greenacre et al. (2021). Compositional Data Analysis of Microbiome and Any-Omics Datasets: A Validation of the Additive Logratio Transformation. Microbial Ecology in Health and Disease. Frontiers in Microbiology

(Log-Ratio) Sélection d'une référence

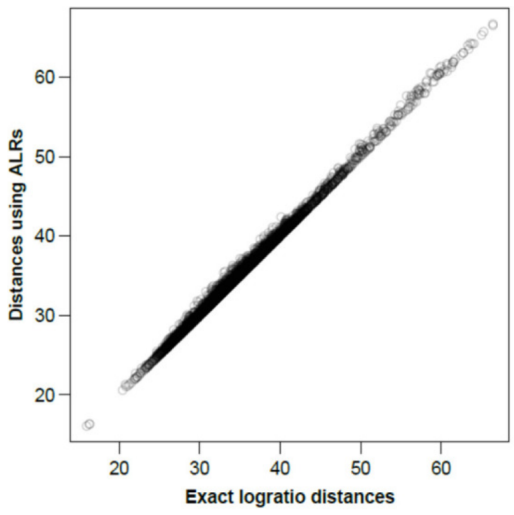
Quelques points à savoir :

- La transformation ALR est un sous-ensemble des log-ratios possibles. Il existe $\frac{1}{2}J(J-1)$ combinaisons possibles, où J est le nombre de composantes.
- CLR: fidèle à la géométrie exacte des log-ratios.
- ALR, en tant que sous-ensemble des log-ratios possibles, ne peut qu'approcher la géométrie exacte des log-ratios.

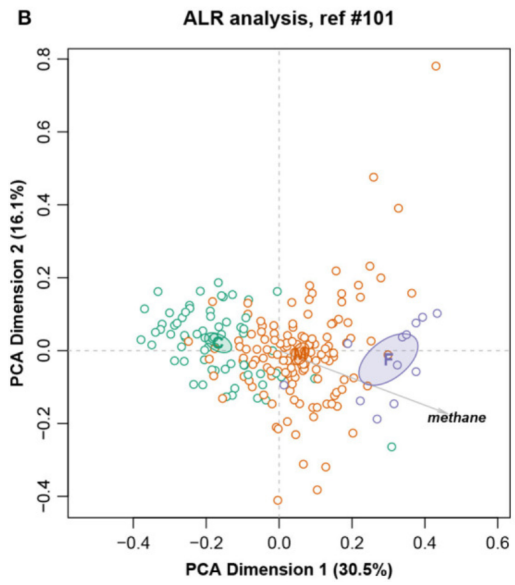
(Log-Ratio) Sélection d'une référence



Calcul des distances **log-ratio exactes** entre échantillons à partir d'une analyse log-ratio (ACP dans l'espace CLR)



Corrélation des distances euclidiennes entre échantillons



Calcul des distances **log-ratio** entre échantillons à partir d'une analyse ALR (ACP dans l'espace ALR)

Autre stratégie de choix d'une réference. Cas log ratio explicatif d'un phénotype.

Interprétation des coefficients selon la transformation. Contexte biologique : communautés pro- et anti-méthanogènes

Pôle	Groupe fonctionnel	Genre / Famille typique	Rôle métabolique	Effet sur CH_4
Pro-méthanogène	Méthanogènes	*Methanobrevibacter smithii* (*Methanobacteriaceae*)	Utilisent $H_2 + CO_2 \rightarrow CH_4$	$+++$
	Producteurs d'H_2	*Ruminococcus albus*	Fermentation des glucides $\rightarrow$ production de H_2	$++$
	Producteurs d'H_2	*Protozoaire*		$+++$
Anti-méthanogène	Acétogènes	*Acetobacterium woodii*	$H_2 + CO_2 \rightarrow$ acétate (voie concurrente)	$\downarrow$
	Producteurs de propionate	*Prevotella*	Voie propionate consommatrice de H_2	$\downarrow\downarrow\downarrow$

Modélisation CoDA de la production de CH_4 (g CH_4 / kg ingéré)

Cas ALR (Additive Log-Ratio)

- Référence : *Ruminococcus* (producteur d'H_2, neutre à légèrement favorable)

$$CH_4 = \beta_0 + \beta_1 \log\left(\frac{x_{\text{Methanobacteriaceae}}}{x_{\text{Ruminococcus}}}\right) + \beta_2 \log\left(\frac{x_{\text{Prevotella}}}{x_{\text{Ruminococcus}}}\right) + \beta_3 \log\left(\frac{x_{\text{Acetobacterium}}}{x_{\text{Ruminococcus}}}\right) + \varepsilon$$

Référence biologique : Ruminococcus = producteur d'H_2 $\rightarrow$ tous les effets sont relatifs à sa contribution à la méthanogenèse.

Interprétation des effets via le log-contraste du modèle. Voir[7]

[7] Michael Greenacre et al (2021). Compositional Data Analysis. Annual Review of Statistics and Its Application.

Amalgamation

Cas SLR (Summed Log-Ratio)

Amalgamation par fonction biologique :

$$z_{\text{SLR}} = \log\left(\frac{x_{\text{Methanobacteriaceae}} + x_{\text{Ruminococcus}} + x_{\text{Fibrobacter}}}{x_{\text{Prevotella}} + x_{\text{Acetobacterium}} + x_{\text{Blautia}}}\right)$$

$$CH_4 = \beta_0 + \beta_1 z_{\text{SLR}} + \varepsilon$$

Coefficient	Interprétation écologique	Signe attendu
β_1	Si le rapport "pro-CH_4 / anti-CH_4" augmente, CH_4 augmente	**positif fort**

Un $\beta_1 > 0 \rightarrow$ l'écosystème bascule vers une dominance pro-méthanogène (forte production de CH_4).

Data-driven amalgamation (package : amalgam)[8]

Principe de l'Amalgamation

$$Y = XA$$

$$Y = XA = [0.2 \quad 0.3 \quad 0.5] \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = [0.5 \quad 0.5]$$

- X: matrice des compositions $(n \times D)$
  $\rightarrow$ n échantillons, D parties (taxons, métabolites, etc.)
- A: **matrice d'amalgamation** $(D \times D')$
  $\rightarrow$ chaque ligne = une partie, chaque colonne = un groupe
  $\rightarrow$ $A_{ij} = 1$ si la partie *i* appartient à l'amalgam *j*
- Y: **nouvelles compositions réduites** $(n \times D')$
  $\rightarrow$ chaque colonne est une somme de parties = un *amalgam*

Objectif :
Réduire la dimension du simplex tout en conservant l'information relative.
Une amalgamation = regroupement de composantes en blocs interprétables.

[8] Quinn et al (2020). "Amalgams: data-driven amalgamation for the dimensionality reduction of compositional data". NAR Genomics and Bioinformatics.

Fonction Objectif : Trouver la "meilleure" matrice A[9]

Deux stratégies selon le but de l'analyse :

1. Objectif non supervisé

Préserver la géométrie du simplex.

$$A_d = \arg\max_A \rho(d(X), d(XA))$$

- Corrélation entre distances originales et distances réduites.

[9] Quinn et al (2020). "Amalgams: data-driven amalgamation for the dimensionality reduction of compositional data". NAR Genomics and Bioinformatics.

2. Objectif supervisé[10]

Maximiser la séparation entre groupes.

$$A_c = \arg\max_A RDA(\mathit{ilr}(XA) \sim L)$$

- L = variable de groupe (ex. Groupe A vs Groupe B)
- Recherche d'amalgams **discriminants**.

Dans les deux cas, la fonction objectif évalue la qualité d'une amalgamation donnée.

[10] Quinn et al (2020). "Amalgams: data-driven amalgamation for the dimensionality reduction of compositional data". NAR Genomics and Bioinformatics.

Merci pour votre attention