Estimation de taille d'étude dans le contexte d'analyse du microbiome

Roxane Couëron [a], Hélène Savel [a], Boris Hejblum [b]

[a] *CHU de Bordeaux, Pôle de Santé publique, Unité de Soutien Méthodologique à la Recherche clinique et epidémiologique (USMR) et CIC-EC 14-01, Bordeaux*

[b] *Univ. Bordeaux, ISPED, Inserm UMR U1219-Bordeaux Population Health Research Center, Inria SISTM, Bordeaux*

CONTEXTE

- **Objectif** : Proposer une méthode d'estimation de taille d'étude pour comparer la composition du microbiote entre deux conditions (ex : avant/après traitement)

- Méthodes classiques d'estimation de taille d'étude non applicables car :
  - Données de microbiote = données compositionnelles de grandes dimensions
  - Analyse statistique : analyse d' abondance différentielle
    identifier les micro-organismes dont l'abondance est différente entre deux conditions

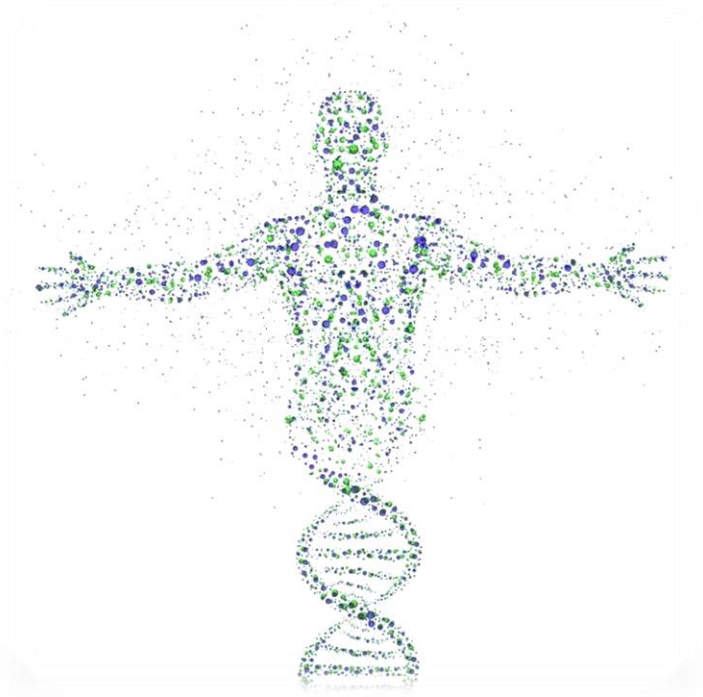
- **Application - étude MICAMU** :
  - Objectif : étudier l'impact de l'antibiothérapie intraveineuse (ATB IV) sur le microbiote intestinal chez les enfants atteints de mucoviscidose
  - Schéma : étude observationnelle, monocentrique, prospective
  - Critère de jugement principal : évolution du profil de microbiote fécal mesuré avant et après traitement

CONTEXTE : données de microbiote

- **Microbiote** : ensemble des micro-organismes vivant dans un environnement spécifique (**microbiome**) chez un hôte

- **Séquençage ciblé** d'un gène marqueur microbien spécifique
  - Bactéries : gène 16S
  - Champignons : gène ITS

- **OTU (Unité Taxonomique Opérationelle)** : ensemble de micro-organismes présentant un niveau de similarité de séquence d'ADN



http://riteshbawri.com/microbiome/

CONTEXTE : données de microbiote

Table d'abondances brutes de l'étude pilote MICAMU (table d'OTUs)

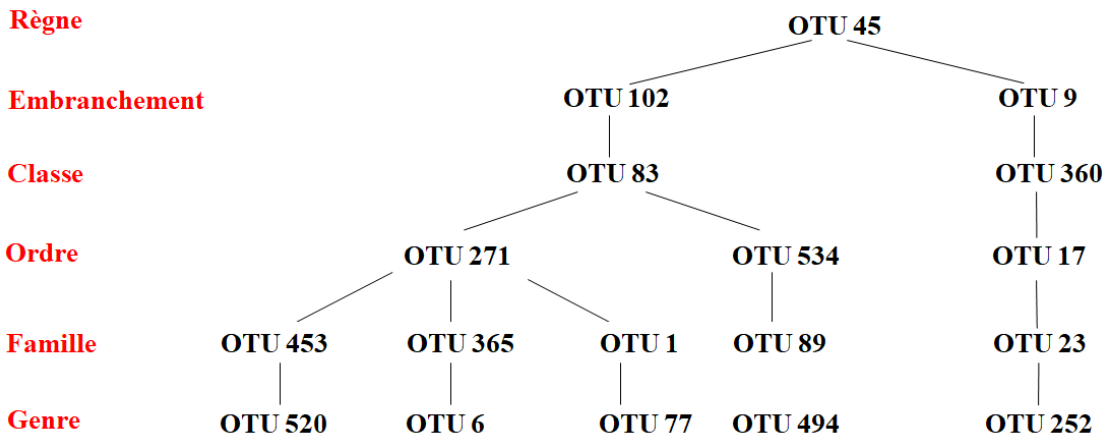
Sujets / OTUs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
OTU 1	0	0	0	0	2	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
OTU 2	0	0	0	0	0	0	6	0	0	0	0	0	0	0	4	0	0	0	0	0
OTU 3	11	0	2	7	5	35	72	69	0	121	70	168	52	2	62	23	2	4	14	0
OTU 4	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
OTU 5	447	232	117	49	22	480	43	177	4	98	305	333	176	9	163	98	0	7	2	0
OTU 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
OTU 7	0	0	0	0	0	0	0	0	0	0	0	0	364	98	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
OTU 576	12	7	0	0	0	0	0	9	0	0	0	1	0	0	0	0	0	0	0	0

CONTEXTE : données de microbiote

Des données spécifiques :

- Données compositionnelles

Arbre phylogénétique de la lignée taxonomique



Corrélations entre les différentes OTUs

CONTEXTE : données de microbiote

Des données spécifiques :

- Données compositionnelles
- Grande dimension

Table d'abondance brutes de l'étude pilote MICAMU (table d'OTUs)

Sujets / OTUs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
OTU 1	0	0	0	0	2	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
OTU 2	0	0	0	0	0	0	6	0	0	0	0	0	0	0	4	0	0	0	0	0
OTU 3	11	0	2	7	5	35	72	69	0	121	70	168	52	2	62	23	2	4	14	0
OTU 4	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
OTU 5	447	232	117	49	22	480	43	177	4	98	305	333	176	9	163	98	0	7	2	0
OTU 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
OTU 7	0	0	0	0	0	0	0	0	0	0	0	0	364	98	0	0	0	0	0	0
	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
OTU 576	12	7	0	0	0	0	0	9	0	0	0	1	0	0	0	0	0	0	0	0

CONTEXTE : données de microbiote

Des données spécifiques :

- Données compositionnelles
- Grande dimension
- Inflation de zéros
  - OTU absente
  - OTU présente < seuil

Table d'abondance brutes de l'étude pilote MICAMU (table d'OTUs)

Sujets / OTUs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
OTU 1	0	0	0	0	2	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
OTU 2	0	0	0	0	0	0	6	0	0	0	0	0	0	0	4	0	0	0	0	0
OTU 3	11	0	2	7	5	35	72	69	0	121	70	168	52	2	62	23	2	4	14	0
OTU 4	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
OTU 5	447	232	117	49	22	480	43	177	4	98	305	333	176	9	163	98	0	7	2	0
OTU 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
OTU 7	0	0	0	0	0	0	0	0	0	0	0	0	364	98	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
OTU 576	12	7	0	0	0	0	0	9	0	0	0	1	0	0	0	0	0	0	0	0

CONTEXTE : données de microbiote

Des données spécifiques :

- Données compositionnelles
- Grande dimension
- Inflation de zéros
  - OTU absente
  - OTU présente < seuil
- Normalisation

	Patient A
OTU 1	0
OTU 2	0
OTU 3	11
OTU 4	0
OTU 5	447
OTU 6	0
OTU 7	0
…	…
OTU q	82
Total	1788

	Patient B
OTU 1	0
OTU 2	0
OTU 3	3
OTU 4	0
OTU 5	447
OTU 6	0
OTU 7	0
…	…
OTU q	36
Total	4470

MÉTHODE

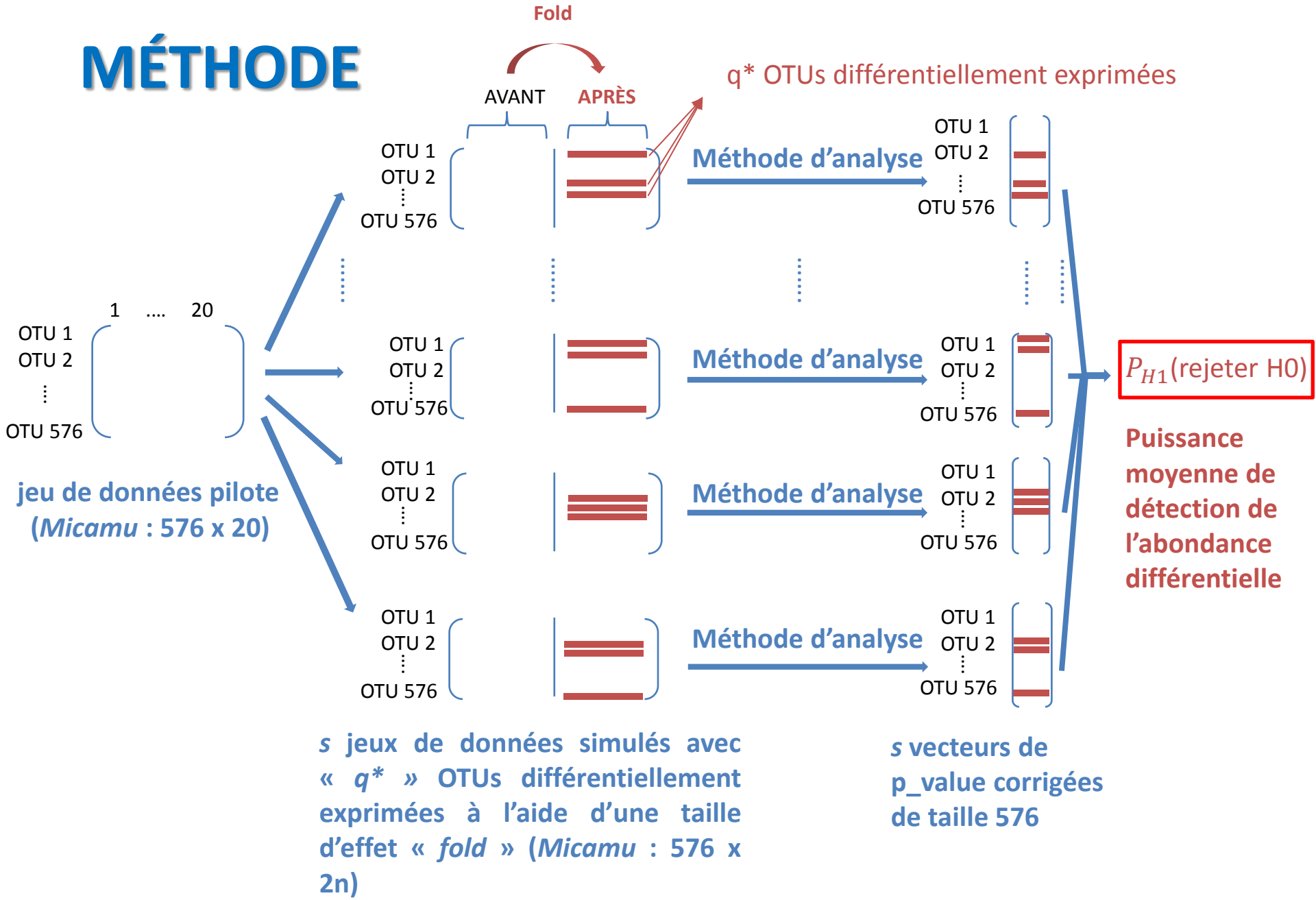
La puissance de détection de l'abondance différentielle est estimée à partir de :

- une étude pilote (difficulté à simuler des données compositionnelles de grande dimension)

- 3 paramètres fixés :
  - Une taille d'étude *« n »*
  - Une proportion d'OTUs différentiellement exprimées entre les deux conditions → appelé *« q* »*
  - Un effet minimum à mettre en évidence (multiplicatif) → appelé *« Fold »*

- Une méthode d'analyse adaptée aux données de microbiote (littérature)

MÉTHODE


Fold
AVANT
APRÈS
q* OTUs différentiellement exprimées
1 20
OTU 1
OTU 2
OTU 576
jeu de données pilote (*Micamu* : 576 x 20)
Méthode d'analyse
P_{H1}(rejeter H0)
Puissance moyenne de détection de l'abondance différentielle
s jeux de données simulés avec « *q** » OTUs différentiellement exprimées à l'aide d'une taille d'effet « *fold* » (*Micamu* : 576 x 2n)
s vecteurs de p_value corrigées de taille 576

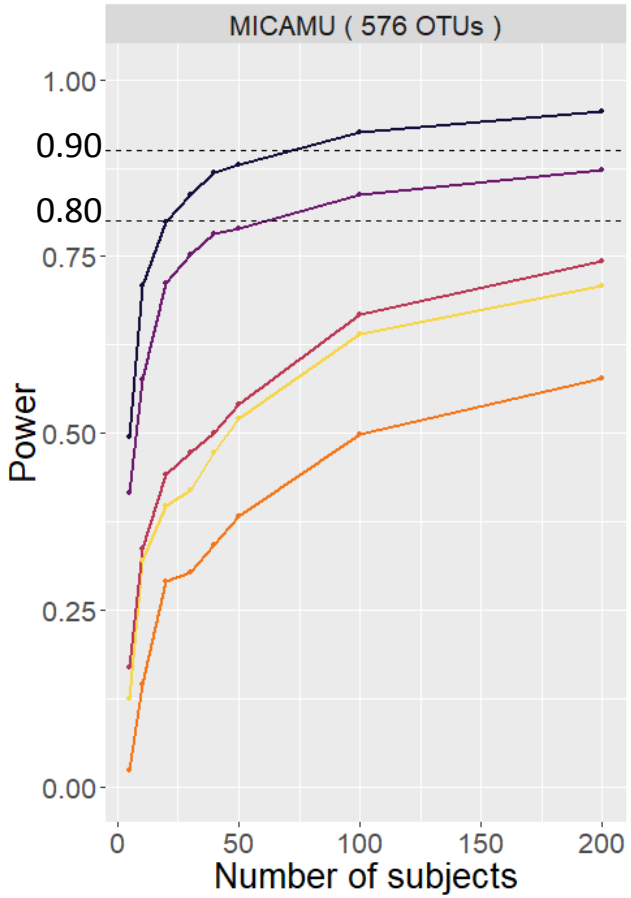
ALDEx 2

Fernandes A. D. et al. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2014; 2: 15

- Données compositionnelles
  - Transformation clr
- Normalisation
  - Loi de Dirichlet
- Inflation de zéros
  - Suppression des OTUs ayant des comptes à zéro dans tous les échantillons
  - Prior non informatif de ½ (statistique bayésiennes)
- Prise en compte de la multiplicité des tests
  - Correction de Benjamini-Hochberg (FDR)
- données appariées / indépendantes

RÉSULTATS

Evolution de la puissance de détection de l'abondance différentielle en fonction du nombre de sujets pour différentes valeurs de taille d'effet « fold »



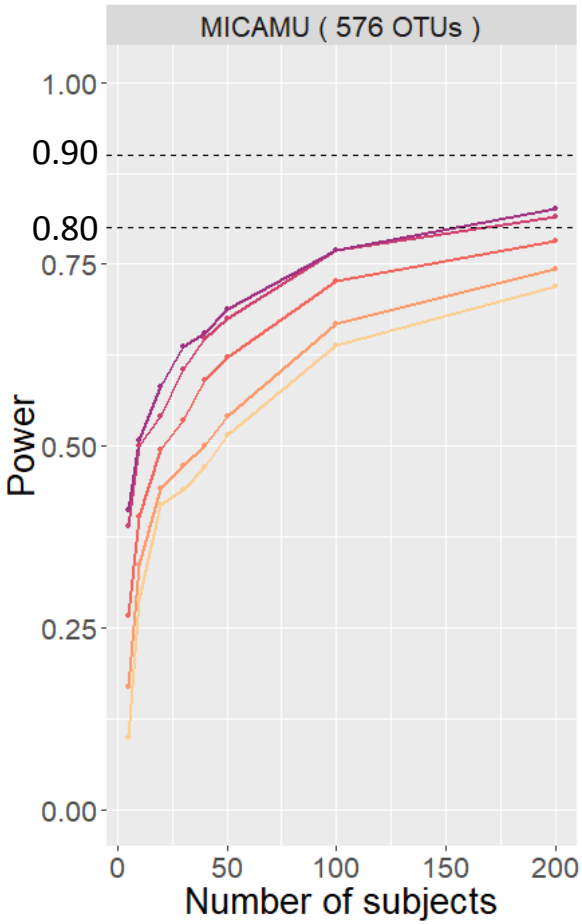
Effet minimum à mettre en évidence

fold = 10
fold = 5
fold = 2
fold = 0.5
fold = 1.5

- ALDEx2
- *100 simulations*
- $q^* = 10\% \times 576\ OTUs = 58\ OTUs$

RÉSULTATS

Evolution de la puissance de détection de l'abondance différentielle en fonction du nombre de sujets pour différentes proportions « q* » d'OTUs exprimées différentiellement entre les deux conditions



Proportion d'OTUs générées avec une abondance différente entre les deux conditions (parmi les 576 OTUs)

- q*= 50%
- q*= 30%
- q*= 20%
- q*= 10%
- q*= 5%

- *ALDEx2*
- *100 simulations*
- $\underline{fold = 2}$

Plus « $q*$ » est grand plus la puissance de détection est élevée

➔ Certains OTUs peuvent aider à détecter les autres (corrélations entre OTUs)

CONCLUSION

- Résultats conformes aux attentes
- Package R documenté, disponible sur demande → Prévision de le rendre accessible sur le CRAN

- Limites :
  - Disposer de données pilotes spécifiques à la question de recherche
  - Fiabilité des hypotheses concernant :
    - La taille d'effet
    - Le pourcentage d'OTUs qui vont modifier leur abondance entre les deux conditions

- Perspectives :
  - Créer une banque de données pilote (données disponibles publiquement)
  - Étudier la sensibilité aux données utilisées pour les simulations
  - Développer une application shiny (Interface web interactive qui fait tourner du code R)

CHU
Hôpitaux de
Bordeaux

MERCI DE VOTRE ATTENTION

Instituts thématiques
Inserm
Institut national
de la santé et de la recherche médicale

Inria
inventeurs du monde numérique

SUPPLÉMENTS

SUPPLEMENTAIRE 1
CONTEXTE

Données de séquençage

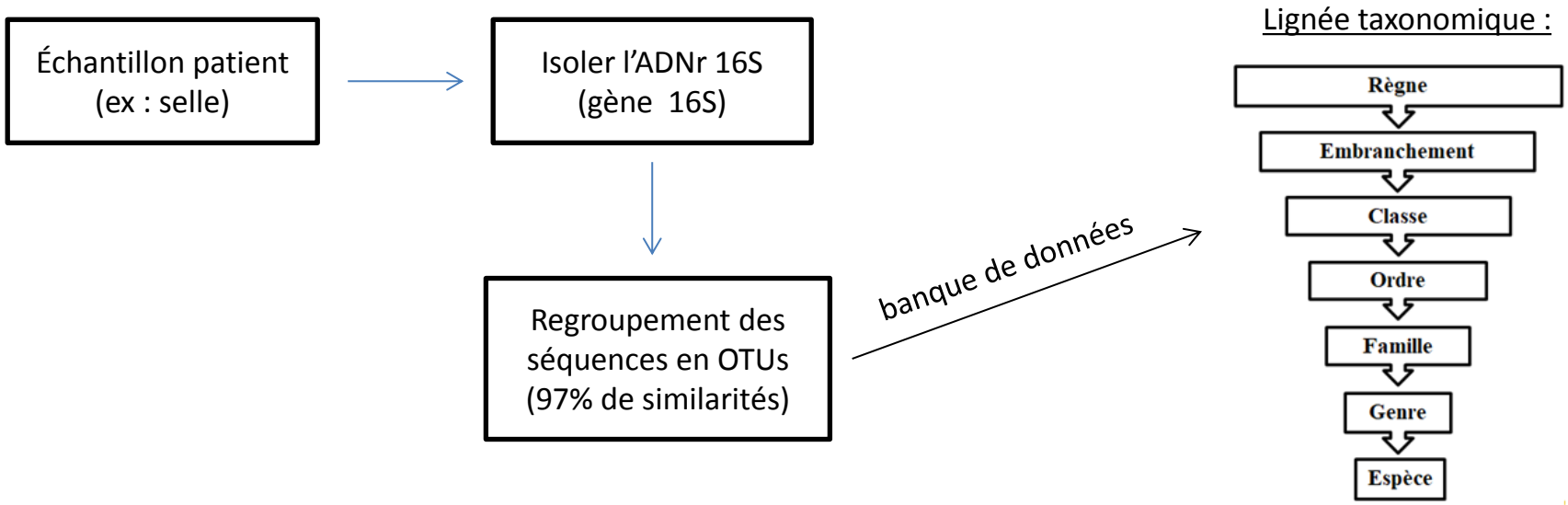
Vue d'ensemble et caractéristiques des données de deux jeux de données réels

Etudes	MucoFong	MICAMU
Niche écologique	Expectorations (crachats ou sécrétions)	selles
Population étudiée	Enfants et adultes atteints de mucoviscidose	Enfants de plus de 3 ans atteints de mucoviscidose
Nombre total des lectures	240828	55462160
Nombre d'échantillons	31	20
Nombre d'OTUs	71	576
Nombre total de lectures par échantillon médian [Q1 – Q3]	7330 [6743 - 8596]	2145706 [1525087 - 2987627]
Sparsité (pourcentage de zéros)	69.0%	60.2%

SUPPLEMENTAIRE 1
CONTEXTE : données de microbiote

- Séquençage d'un gène marqueur microbien spécifique (séquençage ciblé)

Données de séquençage



SUPPLEMENTAIRE 3
CONTEXTE : données de microbiote

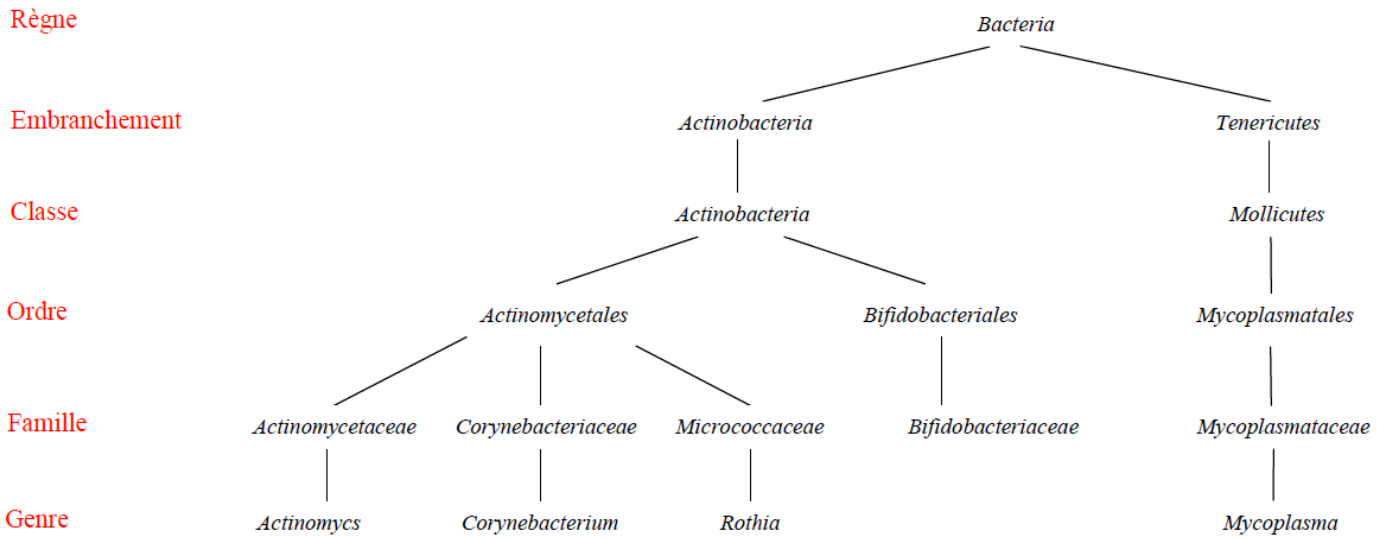
Particularités structurelles des données

- Données de grandes dimensions
- Les micro-organismes du microbiome sont organisés au sein d'un arbre phylogénique
- Données de compositions
  - Les méthodes statistiques standards ne sont pas directement applicables
  - La modélisation des comptes à l'aide de famille de modèles de probabilité peut ne pas convenir
  - Prendre des sous-compositions de données de compositions résulte souvent en une interprétation de la structure de corrélation complètement différente
- Grande proportion de zéros dans les données d'abondances

SUPPLEMENTAIRE 4
CONTEXTE : données de microbiote

Particularités structurelles des données

- Les micro-organismes du microbiome sont organisés au sein d'un arbre phylogénique



➔ Induit des corrélations entre les différents OTUs

SUPPLEMENTAIRE 5
CONTEXTE : données de microbiote

Particularités structurelles des données

- Données de composition

Simplex d' Aitchison

$$S^q = \left\{ x = (\pi_1, \dots, \pi_q) \in R^q \;\middle|\; \pi_i > 0, \sum_{i=1}^{q} \pi_i = 1 \right\}$$

- Les données ne sont pas indépendantes de l'unité
  ➔ **Les méthodes statistiques standards ne sont pas directement applicables**

- Les données du microbiome affichent à la fois des corrélations positives et négatives

➔ **La modélisation des comptes à l'aide de famille de modèles de probabilité peut ne pas convenir**

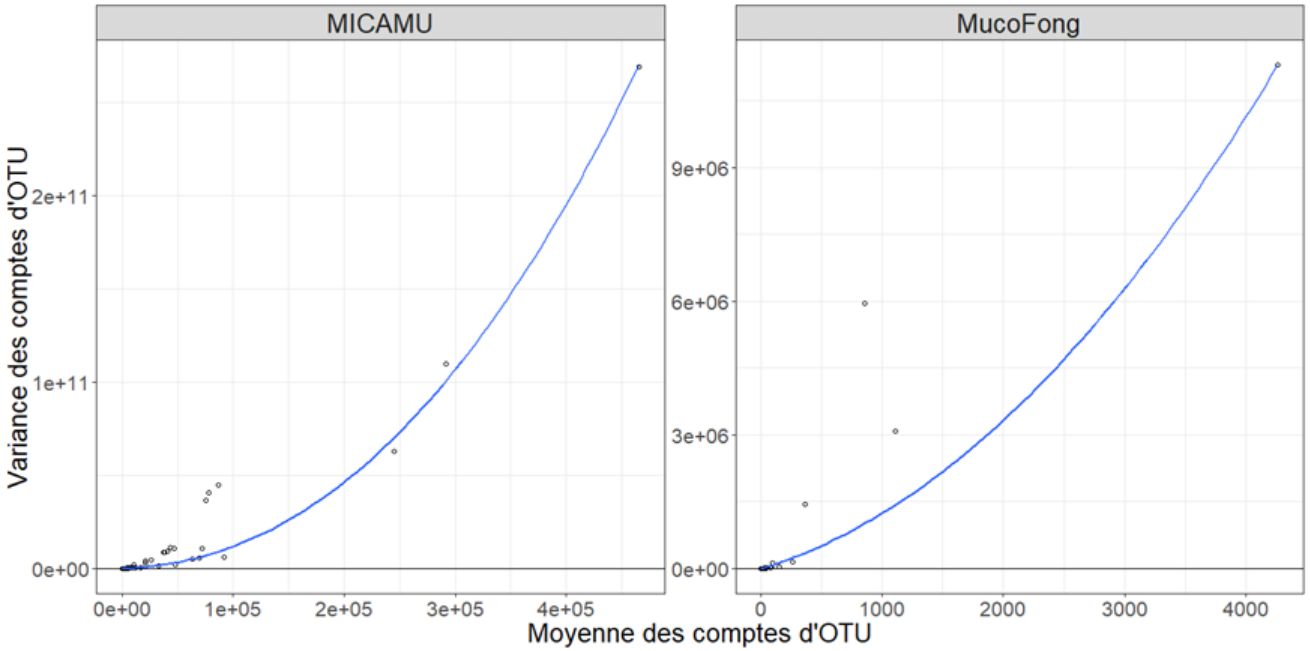
  - Distribution Dirichlet-multinomiale

- Prendre des sous-compositions de données de compositions résulte souvent en une interprétation de la structure de corrélation complètement différente

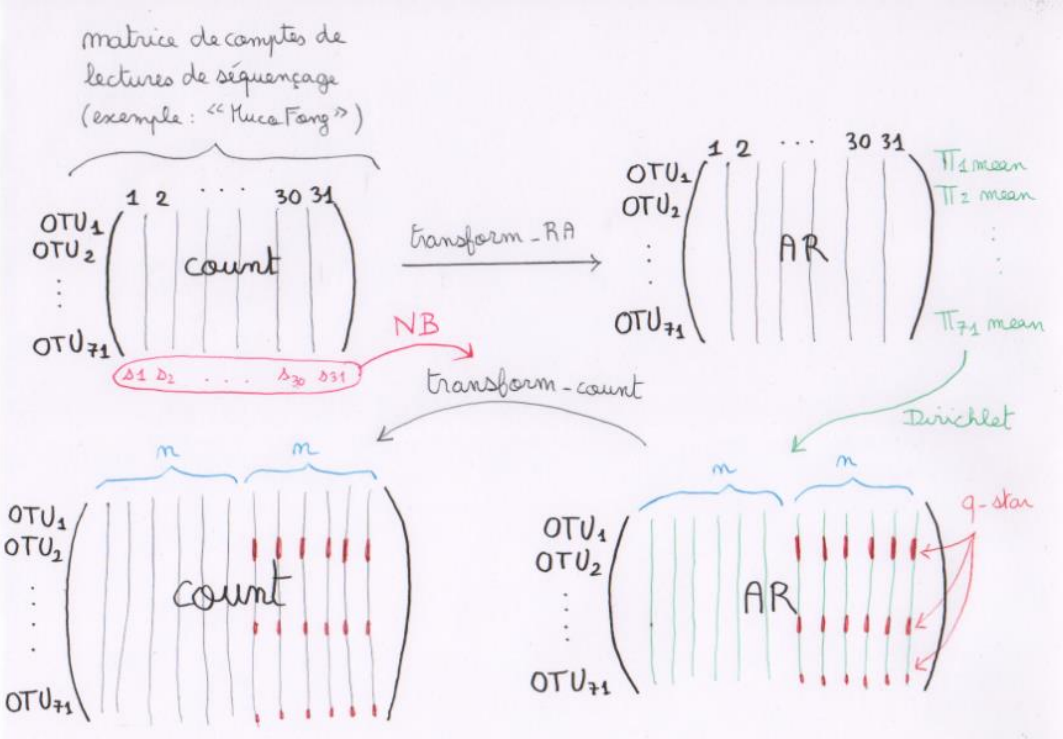
SUPPLEMENTAIRE 6
CONTEXTE : données de microbiote

Particularités structurelles des données

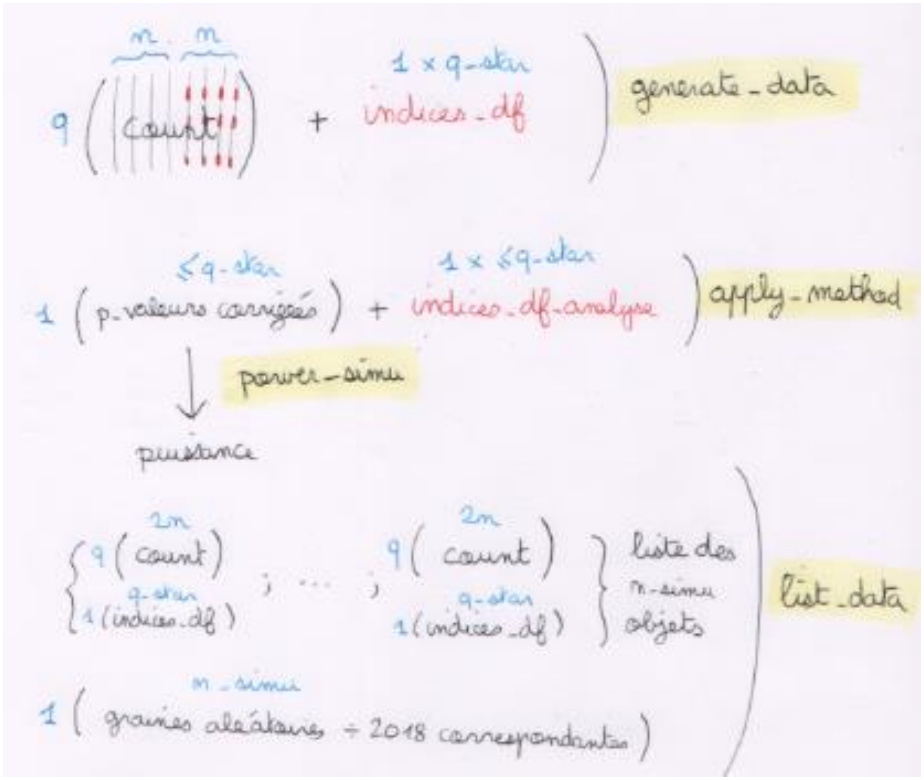
Relation de variance moyenne montrant une sur-dispersion, c'est-à-dire une variance plus élevée que la valeur moyenne d'une OTU donnée.

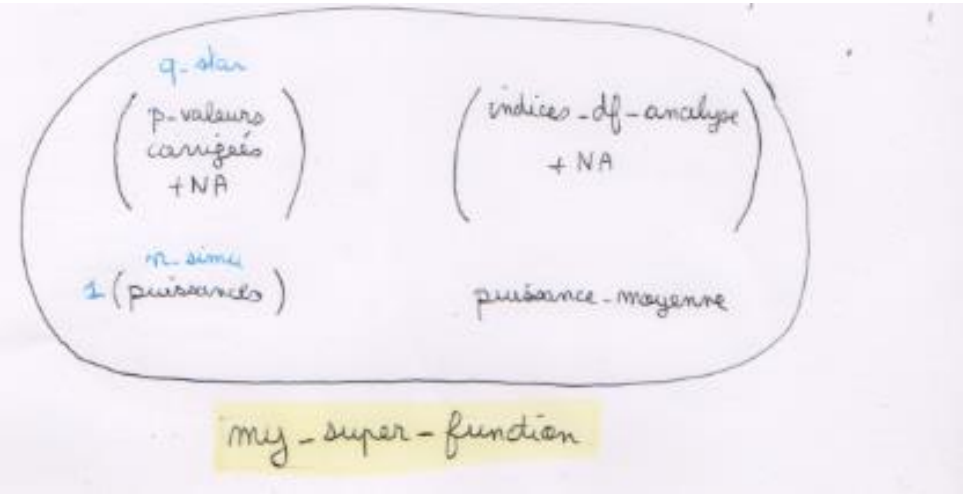


SUPPLEMENTAIRE 11
MÉTHODE



SUPPLEMENTAIRE 12
MÉTHODE





SUPPLEMENTAIRE 7
MÉTHODE : ALDEx 2

Tests d'hypothèses multiples utilisant l'approche du taux de fausse découverte (FDR) de Benjamini et Hochberg.

La technique de Benjamini et Hochberg (1995) contrôle le FDR :

	H_0 Vraies	H_0 Fausses	Total
H_0 non rejetée	U	T	W
H_0 rejetée	V	S	R
Total	m_0	$m - m_0$	m

Le taux de fausse découverte (**FDR**) : correspond à l'espérance de la proportion d'erreur de type I « $\alpha_{seuil} = P_{H_0}(rejet\ H_0)$ » prévue parmi l'ensemble des hypothèses rejetées :

$$FDR = E\left(\frac{V}{R} \mid R > 0\right)$$

SUPPLEMENTAIRE 8
MÉTHODE : ALDEX 2

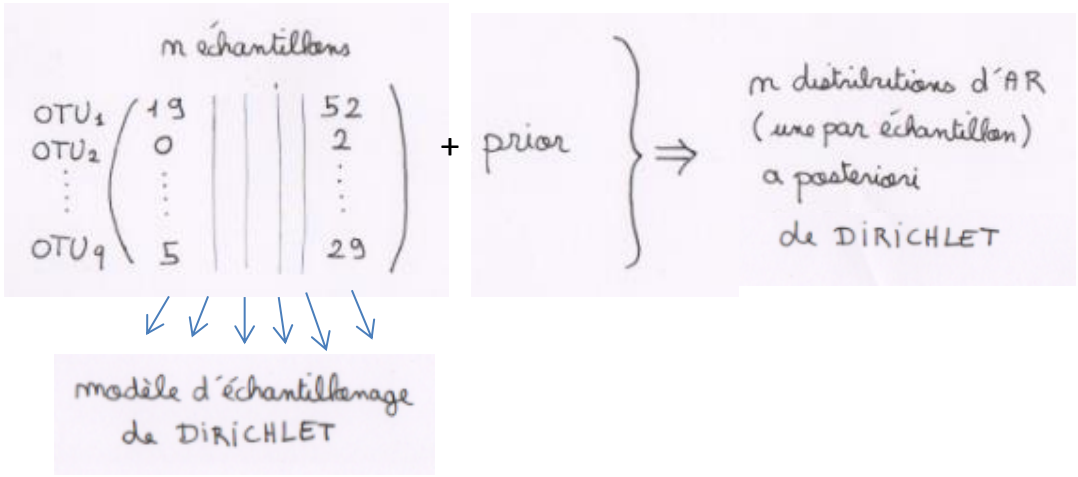
- Données de grandes dimensions
- Les micro-organismes du microbiome sont organisés au sein d'un arbre phylogénique
- Données de compositions
  - Les méthodes statistiques standards ne sont pas directement applicables
  - La modélisation des comptes à l'aide de famille de modèles de probabilité peut ne pas convenir
  - Prendre des sous-compositions de données de compositions résulte souvent en une interprétation de la structure de corrélation complètement différente
- Grande proportion de zéros dans les données d'abondances
- Les distributions de comptes ont une dépendance non triviale entre leur moyenne et leur variance (les variances des distributions de comptage sont plus grandes que leurs moyennes : sur-dispersion)

- Prend en compte l'échantillonnage aléatoire
- Calculs lents

SUPPLEMENTAIRE 9
MÉTHODE : ALDEX 2

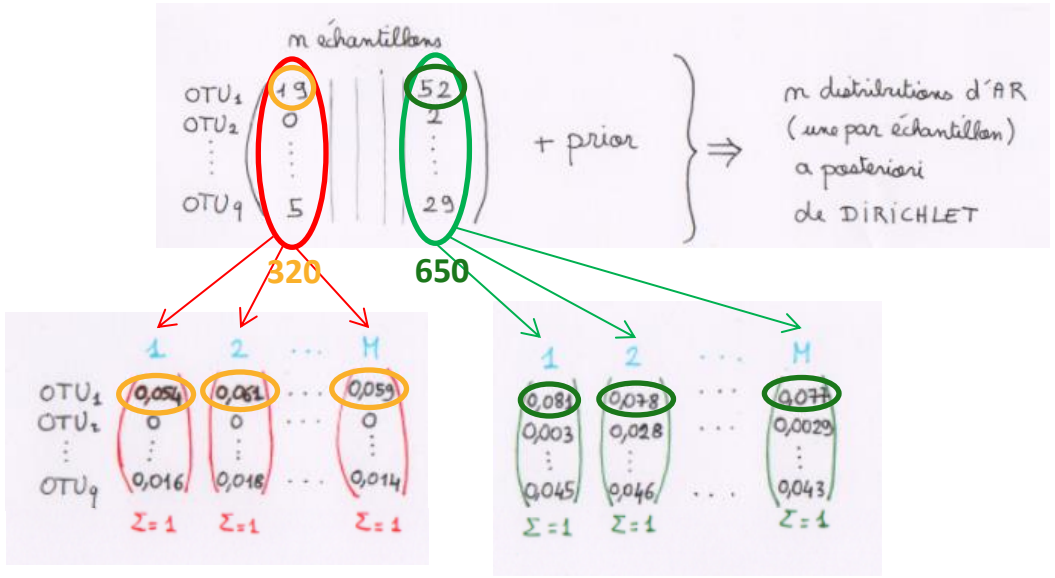
Etape 1 : Convertir la matrice de comptage en **n distributions d'abondance relative a posteriori** (une par échantillon)

- Dirichlet : données compositionnelles
- Inflation de zéros
  - suppression des OTUs présentant des comptes à zéro dans tous les échantillons
  - Prior non informatif de $\frac{1}{2}$



SUPPLEMENTAIRE 9
MÉTHODE : ALDEX 2

Etape 2 : Pour chacun des n échantillon, on génère **M échantillons de** Dirichlet **Monte-Carlo** avec la distribution a posteriori associée



Monte-Carlo

➔ Prise en compte de la variabilité des mesures

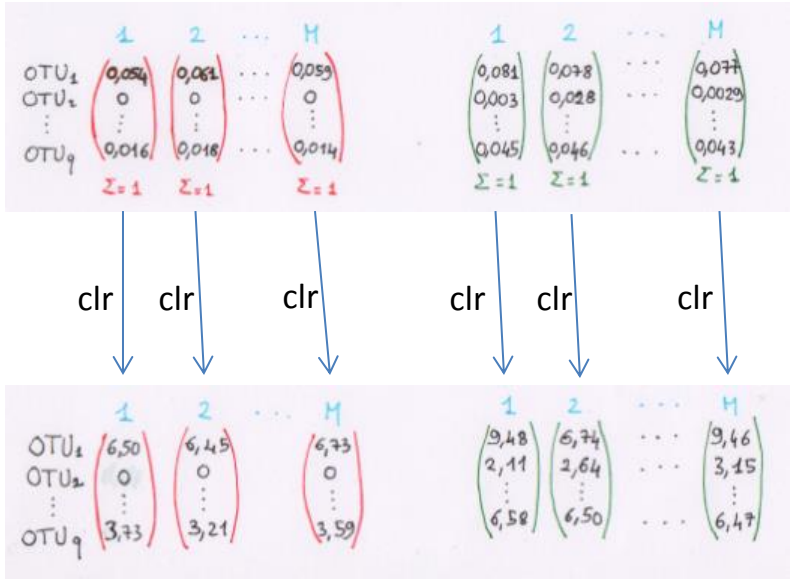
SUPPLEMENTAIRE 9
MÉTHODE : ALDEX 2

Etape 3 : Chaque vecteur $de\ DMC$ est normalisé avec la **transformation** centrée sur le log-ratio ($\boldsymbol{clr}$)

Déf : Soit $x = (x_1, x_2, \dots, x_q)$

$$clr(x) = (\log_2\left(\frac{x_1}{g(x)}\right), \log_2\left(\frac{x_2}{g(x)}\right), \dots, \log_2\left(\frac{x_q}{g(x)}\right))$$

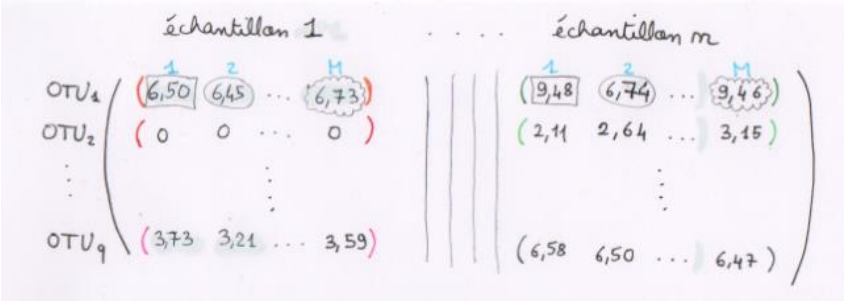
$$g(x) = \prod_{i=1}^{q} x_i^{1/q}$$



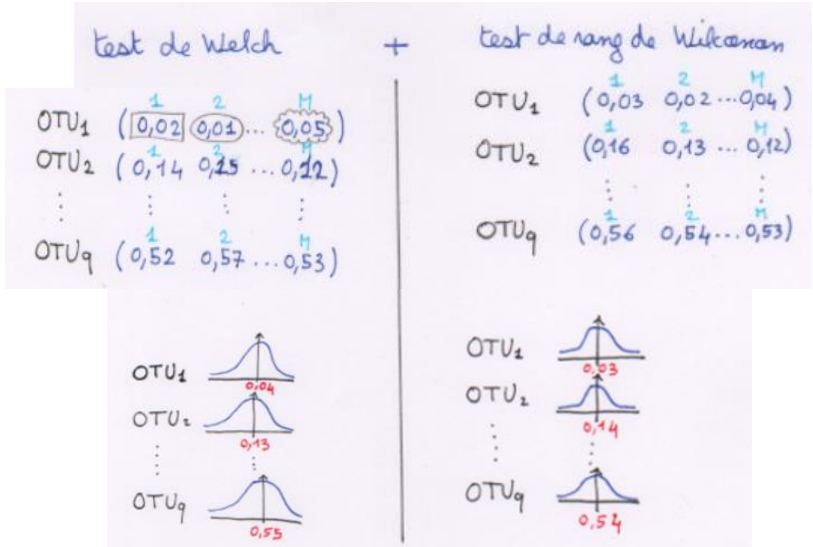
➔ **Cohérence sous-compositionnelle**

SUPPLEMENTAIRE 9
MÉTHODE : ALDEX 2

Etape 4 : Pour chaque OTU on teste la **différence de *clr* entre les deux conditions** pour chaque instance de Dirichlet Monte-Carlo + **correction pour les tests multiples** (Benjamini-Hochberg)



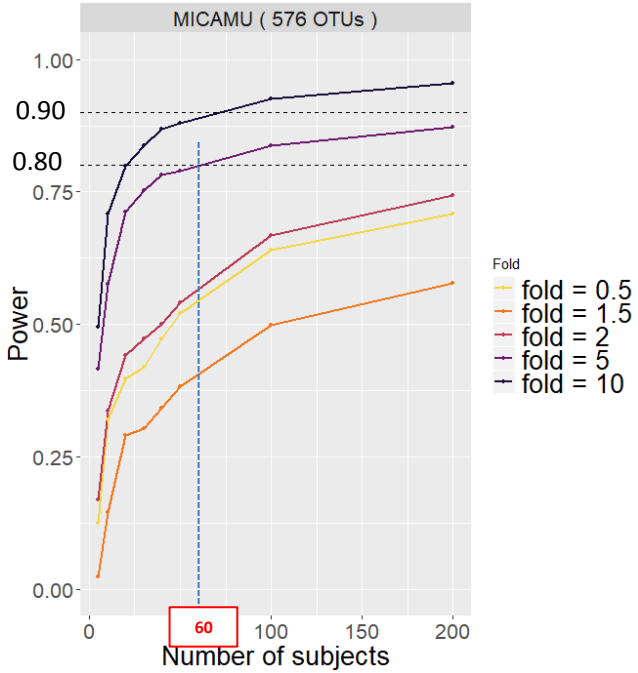
=> On obtient la distribution a posteriori des p-valeurs corrigées pour chaque OTU (tenant compte de la variabilité des mesures)



SUPPLEMENTAIRE 10
RÉSULTATS

Supposons que le clinicien prévoit une sur-expression de 10% des OTUs après l'initiation d'un traitement. Il souhaite pouvoir mettre en évidence une multiplication des abondances relatives par 5 de ces OTUs ($\text{fold} = 5$) avec une puissance statistique de 80%.

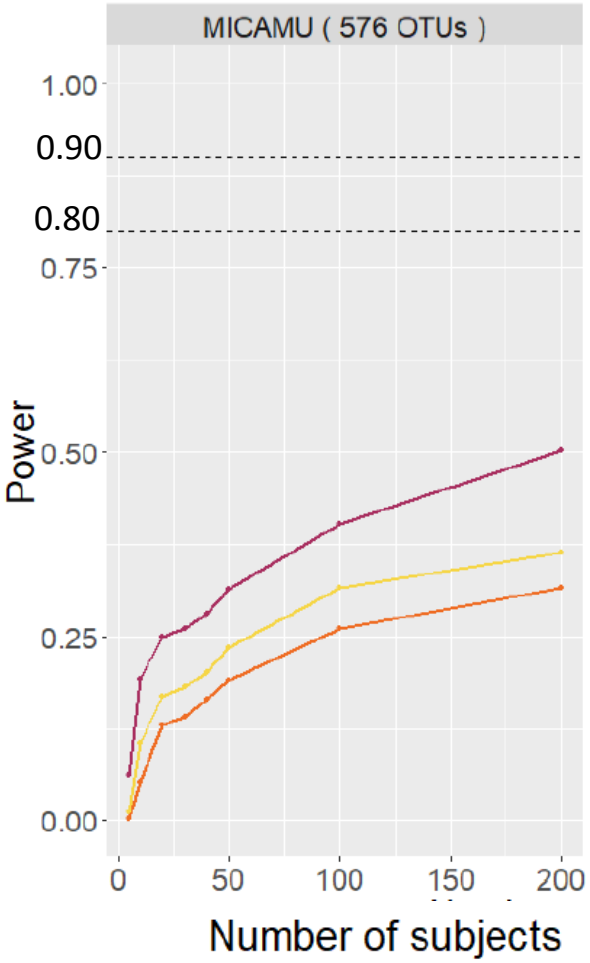
➔ Lorsqu'on considère que 576 OTUs sont détectés par séquençage il faut inclure environ 60 patients pour pouvoir répondre à la question du clinicien.



SUPPLEMENTAIRE 11 RÉSULTATS

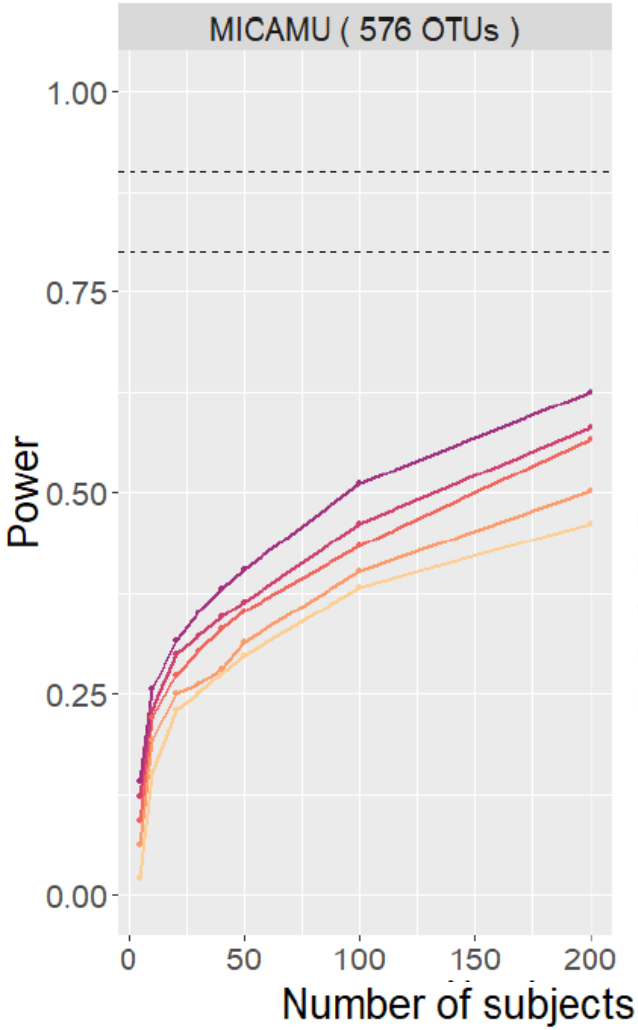
- ALDEx2
- *100 simulations*

Evolution de la puissance de détection de l'abondance différentielle en fonction du nombre de sujets pour différentes valeurs des paramètres



Effet minimum à mettre en évidence

fold = 2
fold = 0.5
fold = 1.5



Proportion d'OTUs générées avec une abondance différente entre les deux conditions (parmi les 576 OTUs)

q*= 50%
q*= 30%
q*= 20%
q*= 10%
q*= 5%

SUPPLEMENTAIRE 12
MÉTHODE : ANCOM

- Données de grandes dimensions
- Les micro-organismes du microbiome sont organisés au sein d'un arbre phylogénique
- Données de compositions
  - Les méthodes statistiques standards ne sont pas directement applicables
  - La modélisation des comptes à l'aide de famille de modèles de probabilité peut ne pas convenir
  - Prendre des sous-compositions de données de compositions résulte souvent en une interprétation de la structure de corrélation complètement différente
- Grande proportion de zéros dans les données d'abondances
- Les distributions de comptes ont une dépendance non triviale entre leur moyenne et leur variance (les variances des distributions de comptage sont plus grandes que leurs moyennes : sur-dispersion)

- Analyse longitudinale de la composition microbienne
- Calculs très lents