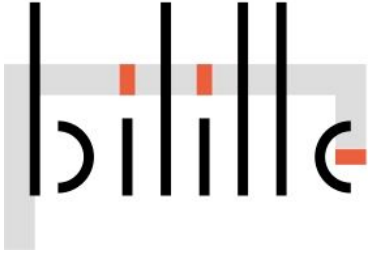
Analyse secondaire

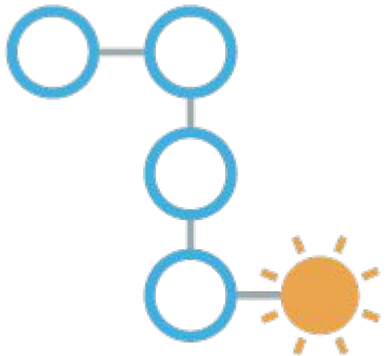
formation métagénomique
Bilille - 4-5-6 avril 2023



OTU ou ASV le cheminement de l’analyse secondaire reste le même

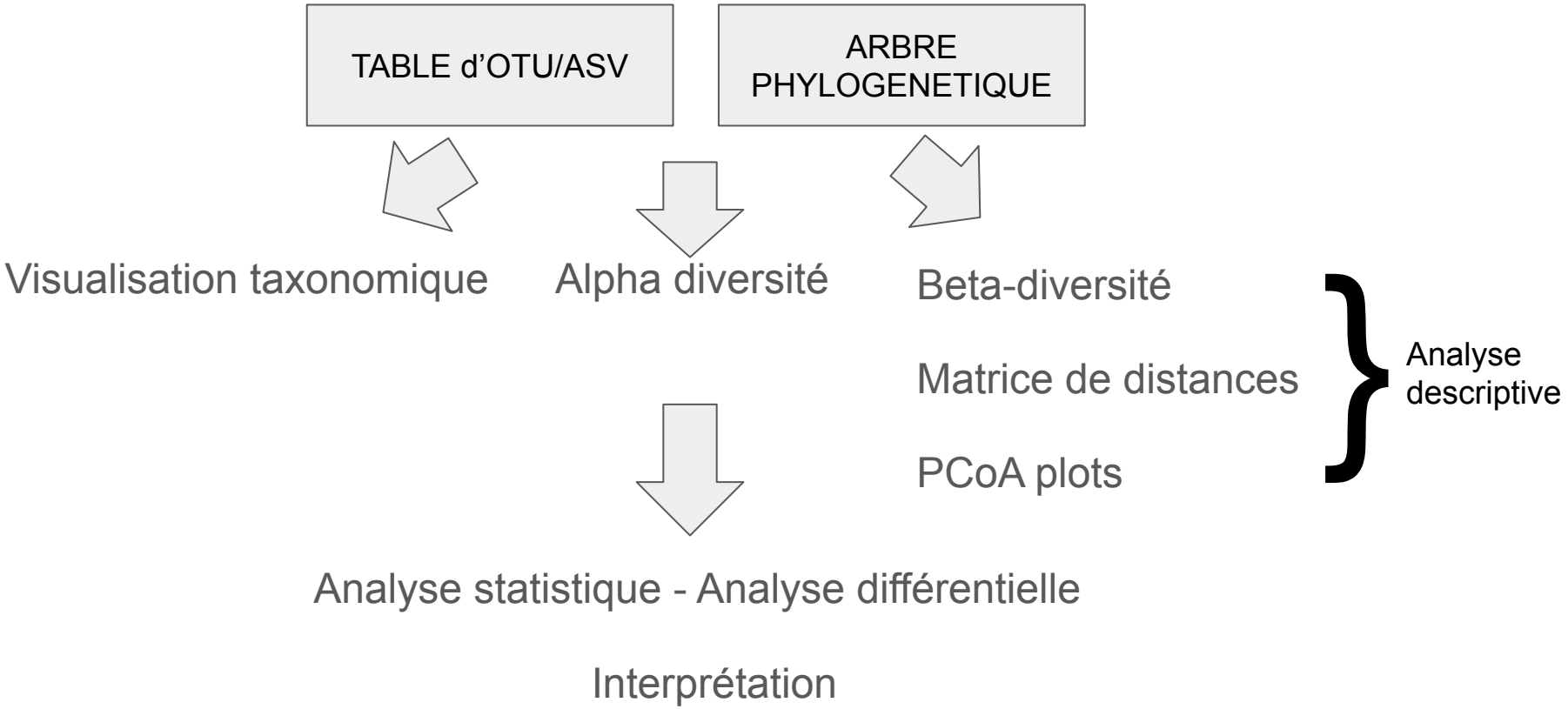
Objectif : valoriser nos données

Extraire de la connaissance

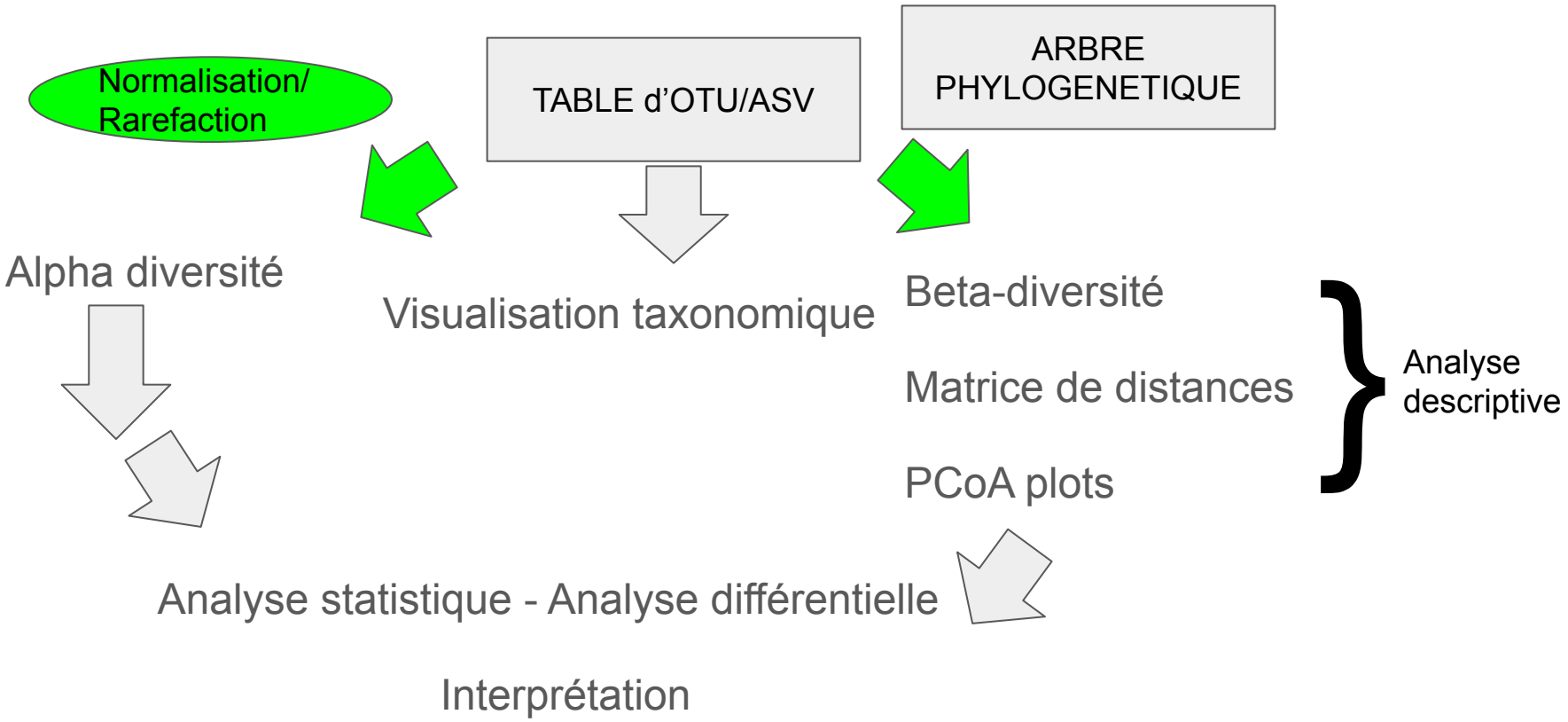


question biologique

De la table de comptage à l'extraction de connaissance



Normalisation des données



Pourquoi normaliser les données?

- Table de comptage des OTU “brutes”
- Nombre de reads ≠ selon les échantillons
- Différences entres les échantillons :
  - Profondeur de séquençage
  - Sans réalité biologique

⇨ Nécessaire pour la comparaison (beta diversité et analyse différentielle)

De l’importance de la normalisation

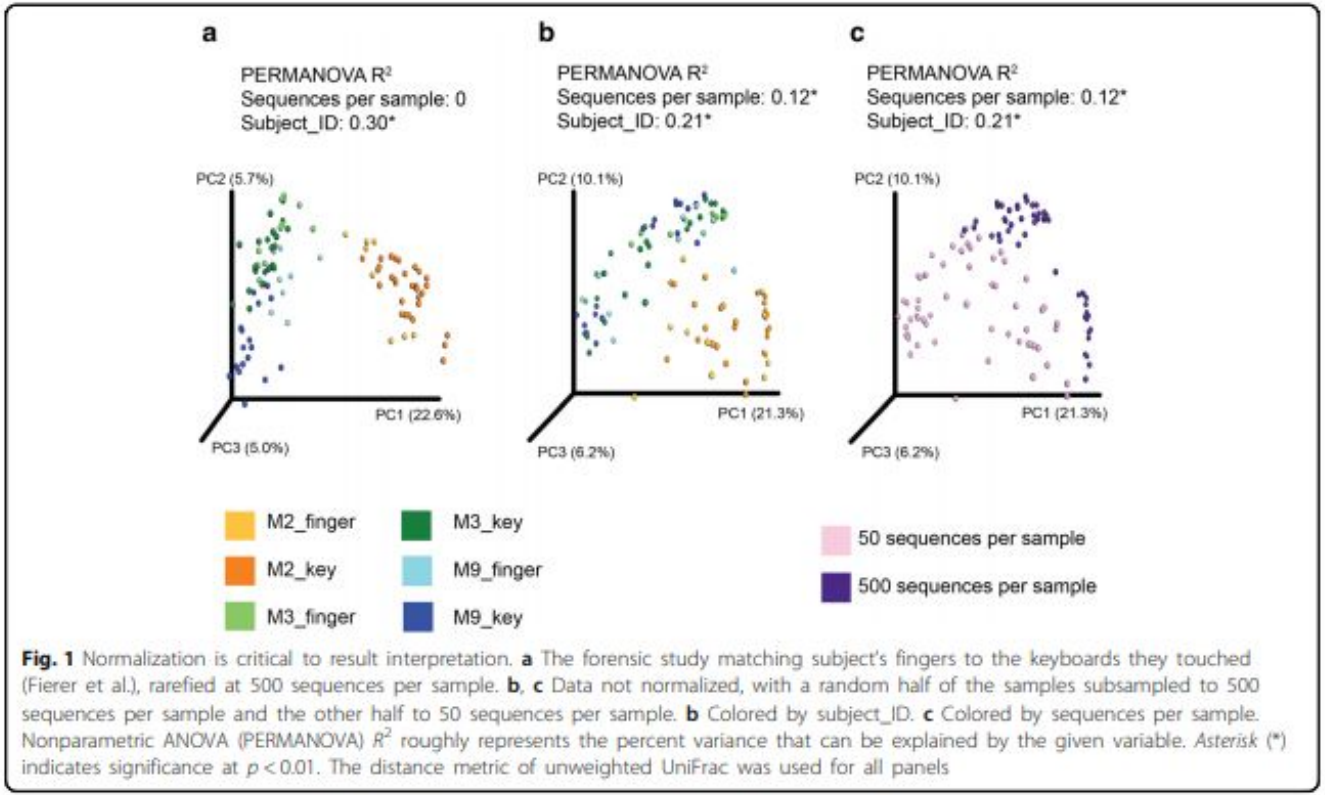
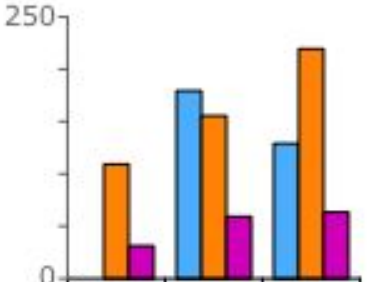


Fig. 1 Normalization is critical to result interpretation. **a** The forensic study matching subject's fingers to the keyboards they touched (Fierer et al.), rarefied at 500 sequences per sample. **b**, **c** Data not normalized, with a random half of the samples subsampled to 500 sequences per sample and the other half to 50 sequences per sample. **b** Colored by subject_ID. **c** Colored by sequences per sample. Nonparametric ANOVA (PERMANOVA) R^2 roughly represents the percent variance that can be explained by the given variable. *Asterisk* (*) indicates significance at $p < 0.01$. The distance metric of unweighted UniFrac was used for all panels

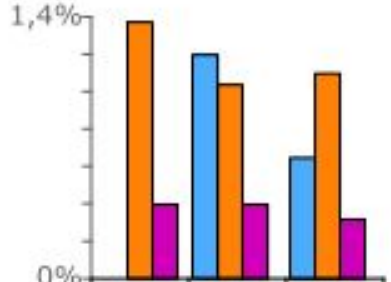
Normalisation: méthode “classique”

a) Table initiale d'observations



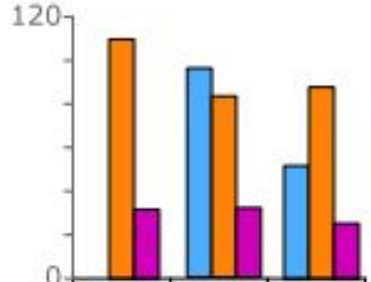
	A	B	C
OTU1	0	180	130
OTU2	110	156	220
OTU3	32	60	64
⋮	⋮	⋮	⋮
Total	8k	15k	20k

b) Normalisation TSS



	A	B	C
OTU1	0	1,2	0,65
OTU2	1,375	1,04	1,1
OTU3	0,4	0,4	0,32
⋮	⋮	⋮	⋮
Total	100%	100%	100%

c) Normalisation par raréfaction



	A	B	C
OTU1	0	96	52
OTU2	110	83,2	88
OTU3	32	32	25,6
⋮	⋮	⋮	⋮
Total	8k	8k	8k

source graphique : L. Siegwald - 2017

Normalisation : 2 références 2 messages

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie,Susan Holmes
Published: April 3, 2014
https://doi.org/10.1371/journal.pcbi.1003531

Normalization and microbial differential abundance strategies depend upon data characteristics

Sophie Weiss et al.
Published: Mars 3, 2017
https://dx.doi.org/10.1186%2Fs40168-017-0237-y

Discussion autour de la normalisation

2014, article qui fait référence : raréfier les données est inadmissible !!!

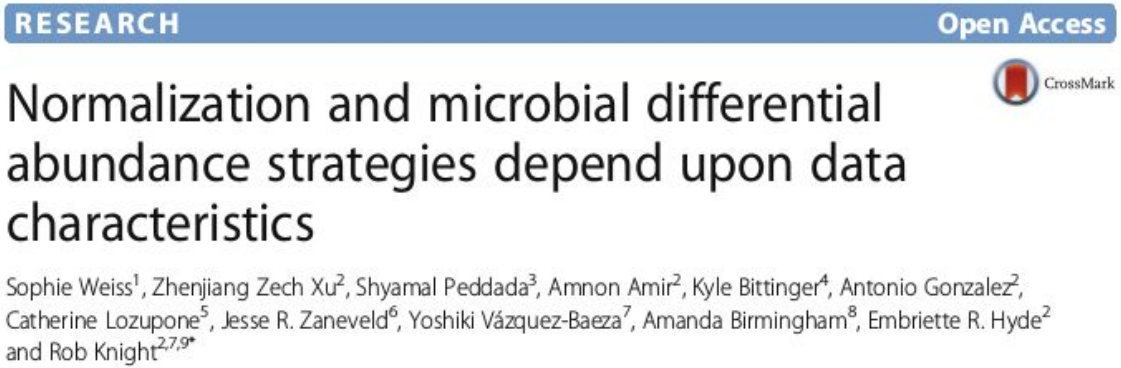
“these approaches are **inappropriate** for detection of differentially abundant species”

“Result in **a high rate of false positives** in tests for species that are differentially abundant across sample classes.”

“Regarding microbiome sample-wise clustering, we also show that the rarefying procedure often discards samples that can be accurately clustered **by alternative methods**”

Discussion autour de la normalisation

Publication de 2017



RESEARCH Open Access

Normalization and microbial differential abundance strategies depend upon data characteristics

Sophie Weiss[1], Zhenjiang Zech Xu[2], Shyamal Peddada[3], Amnon Amir[2], Kyle Bittinger[4], Antonio Gonzalez[2], Catherine Lozupone[5], Jesse R. Zaneveld[6], Yoshiki Vázquez-Baeza[7], Amanda Birmingham[8], Embriette R. Hyde[2] and Rob Knight[2,7,9*]

Finalement la raréfaction n’est pas inadmissible....

La « bonne » normalisation dépend des données...

Normalisation : Qu’est ce que la raréfaction?

- sous-échantillonner le même nombre de séquences de chaque échantillon
- NB : les échantillons sans ce nombre de séquences sont rejetés.

- Préoccupations :
- Trop bas : ignorer beaucoup d'informations sur les échantillons
- Trop élevé : ignorer beaucoup d'échantillons
    - Toujours un bon choix pour la normalisation (Weiss S, et al. Microbiome. 2017) :

> “Rarefying more clearly clusters samples according to biological origin than other normalization techniques do for ordination metrics based on presence or absence”
>
> “Alternate normalization measures are potentially vulnerable to artifacts due to library size”

- Le chercheur doit choisir la profondeur d'échantillonnage, mais comment ?

Type de normalisation : méthode alternative

- DESeq : méthode dérivée de la transcriptomique : Calcul d'un facteur d'échelle pour chaque échantillon qui permettra de multiplier chaque observation

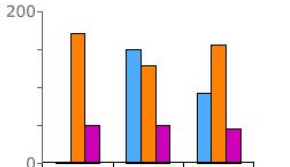
a) Table initiale d'observations

	A	B	C	Échantillon moyen
OTU1	1	180	130	28,60
OTU2	110	156	220	155,71
OTU3	32	60	64	49,72
⋮	⋮	⋮	⋮	⋮

b) Ratios des observations sur l'échantillon moyen

	A	B	C
OTU1	0,03	6,29	4,55
OTU2	0,71	1,00	1,41
OTU3	0,64	1,21	1,27
⋮	⋮	⋮	⋮
Médiane	0,64	1,21	1,41

c) Mise à l'échelle des observations par la médiane des ratios



	A	B	C
OTU1	1,55	149,15	92,01
OTU2	170,90	129,26	155,71
OTU3	49,72	49,72	45,30
⋮	⋮	⋮	⋮

Figure 1.16 : Normalisation par DESeq.

besoin de modifier les données pour prendre en compte les valeurs nulles

source graphique : L. Siegwald - 2017

Type de normalisation : la normalisation CLR

Centered-Log-Ratio

=> **transformer les abondances relatives en coordonnées logarithmiques**

Désormais largement utilisée en analyse métagénomique

Basée sur une transformation log-ratio des abondances relatives des différentes espèces ou taxons présents dans un échantillon

=> centrée autour de zéro

$$x_i^{CLR} = \ln\left(\frac{x_i}{g(x)}\right)$$

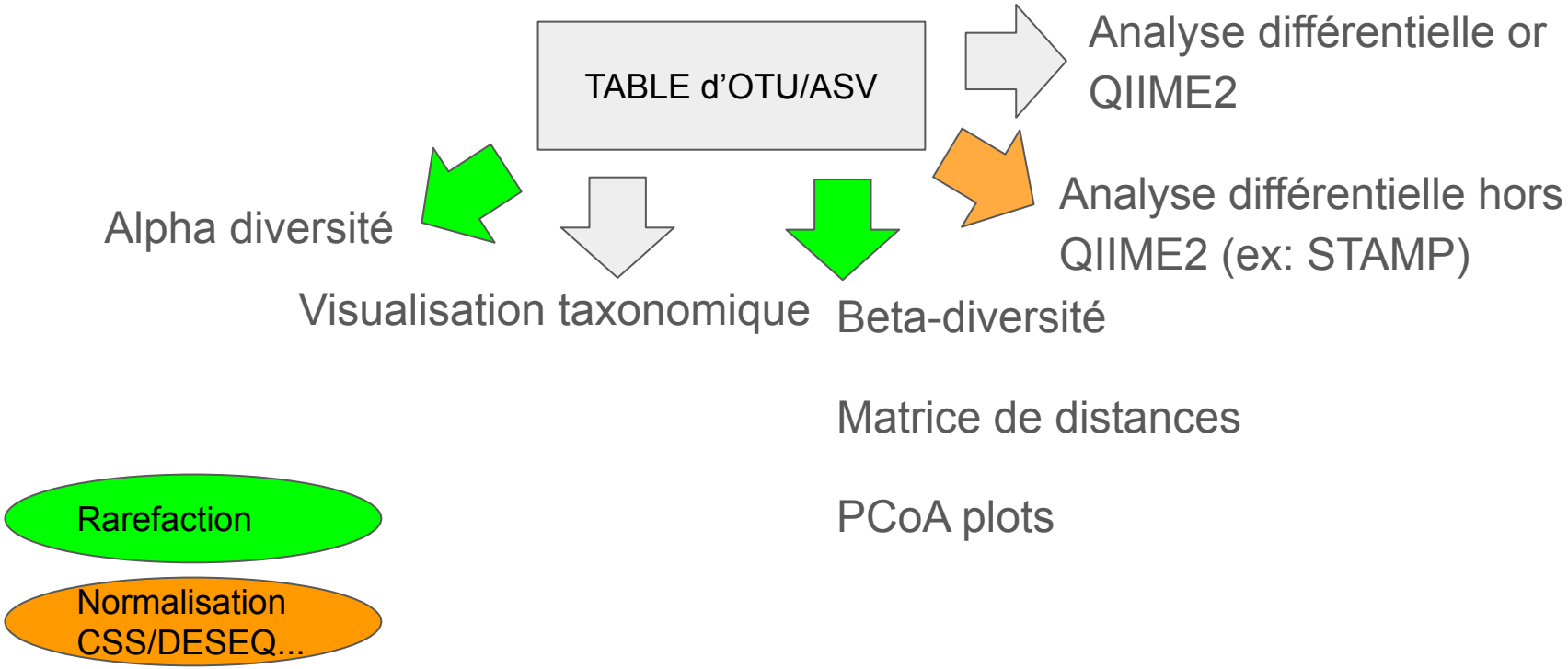
X : abondance relative de l'espèce
G : moyenne géométrique des abondances relatives

=> Utilisé par défaut dans certaines méthodes d'analyse différencielle (ex: ANCOM)

Normalisation: les bonnes pratiques

- Dans le plupart des cas, les méthodes de normalisation montrent des résultats assez similaires
- Test différentiel : La raréfaction n'augmente pas la proportion de faux positifs et peut être adaptée en cas de forte variation de la profondeur de séquençage entres échantillons (à condition que la profondeur de séquençage soit assez élevée).
- Test différentiel : La raréfaction réduit la sensibilité par rapport aux autres techniques (plus de faux négatifs).
- Attention à la profondeur d'échantillonnage quand on utilise la raréfaction compromis à trouver (voir pratique)
- DEseq2, CSS, TMM... : méthodes intéressantes mais récentes et pas forcément adaptées aux données métagénomiques.

Bonne pratique : notre vision avec QIIME2



Alpha-diversité : qu’est ce que c’est?

Elle est utilisée pour mesurer la diversité au sein d’un échantillon.

- Une valeur par échantillon
- Une multitude de métriques pour estimer la diversité de façon différentes.

La richesse (‘Richness’)- basé sur le comptage de la table d’OTU sans prendre en compte l’abondance relative

= nombre d’espèces présent dans l’échantillon (métriques chao1, ACE,...)

La diversité (‘Eveness’) : comptage OTU mais en prenant en compte l’abondance relative

Alpha-diversité : métrique

QIIME 2 calcule un ensemble de métriques pour vous avec une seule commande

- L'indice de diversité de **Shannon** (une mesure quantitative de la richesse communautaire)
- **OTU observées** “observed OTUs” (une mesure qualitative de la richesse communautaire)
- **Faith’s Phylogenetic Diversity** (une mesure qualitative de la richesse communautaire qui intègre les relations phylogénétiques entre les échantillons).deux OTU proche phylogénétiquement auront moins de poids dans le calcul de la diversité => proche de la composition biologique du microbiote
- **Evenness** (ou Evenness de Pielou ; une mesure de la diversité de la communauté)

Alpha-diversité : en pratique

Un seul outil pour générer les diversités (alpha et beta)

qiime2 diversity core-metrics-phylogenetic Core diversity metrics (phylogenetic and non-phylogenetic)

Alpha-diversité : les courbes de raréfaction

Comment savoir si notre richesse estimée correspond à la richesse réelle de notre microbiote?

ex : profondeur de séquençage trop faible pour capter les organismes en faible proportion

⇨ **Observer les courbes de raréfaction**

Principe : Compter le nombre d'OTU/ASVs pour un ensemble de sous-échantillons à différents intervalles de profondeur.

A visualiser lorsqu'on a supprimé des échantillons en raréfiant.

Alpha-diversité : les courbes de raréfaction

Recherche de l'asymptote

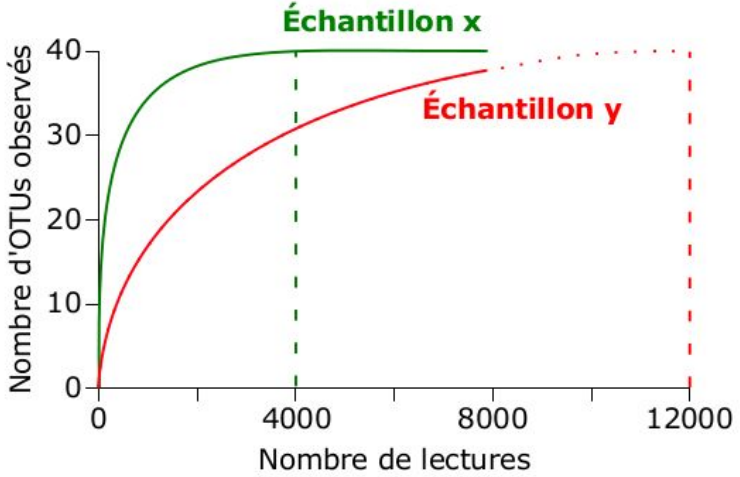


Figure 1.19 : Courbes de raréfaction de deux échantillons. L'échantillon x atteint une asymptote, et a de ce fait une profondeur de séquençage suffisante. L'échantillon y n'atteint pas d'asymptote même à profondeur maximale (8 000 lectures).

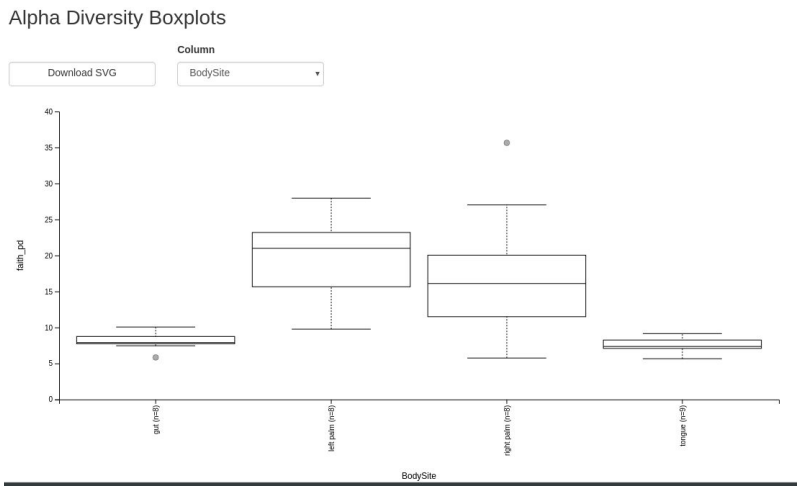
source graphique : L. Siegwald - 2017

Alpha-raréfaction : en pratique

Alpha-diversité / analyse différentielle

Association entre les métadonnées et la diversité alpha.

Test de kruskall-wallis : Test non-paramétrique sur k échantillons indépendants



Alpha-diversité / analyse différentielle: en pratique

qiime2 diversity alpha-group-significance Alpha diversity comparisons

Alpha-diversité : notre vision avec QIIME2

Utiliser les alpha-diversité générés par l'outil *core-metric-phylogenetic*

NE pas appliquer l'alpha diversité sur données normalisées avec des techniques alternatives (DESEQ2/CSS…), utiliser la raréfaction.

Beta-diversité : qu'est ce que c'est?

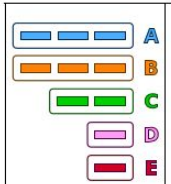
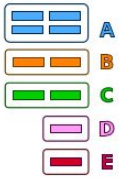
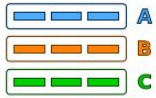
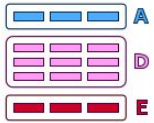
Permet d'estimer la différence de diversité INTER-échantillons

=

Diversité des espèces entres les échantillons

Si plus de deux échantillons on calcule une matrice de distance (ou matrice de dissimilarité) ou chaque "case" représente un score de béta-diversité entre deux échantillons

a) Indices de dissimilarité de Bray-Curtis

	 Échantillon 1	 Échantillon 2	 Échantillon 3	 Échantillon 4
Échantillon 1	0			
Échantillon 2	0,100	0		
Échantillon 3	0,158	0,263	0	
Échantillon 4	0,600	0,600	0,750	0

source graphique : L. Siegwald - 2017

Beta-diversité : Quelques métriques

Théorie :

Indice de Jaccard : d(jaccard) = b + c / a + b + c

a : nombre d’OTUs partagés

b : nombre d’OTUs spécifiques au premier échantillon

c : nombre d’OTUs spécifiques au deuxième échantillon

Pas de prise en compte des proportions

Beta-diversité : Quelques métriques

Bray-curtis (prise en compte des proportions) - distance non-euclidienne :

$$d_{Bray-Curtis} = \frac{\sum_{i=1}^{N} |p_{iA} - p_{iB}|}{\sum_{i=1}^{N} (p_{iA} + p_{iB})}$$

où p iA et p iB sont les abondances relatives de l'OTU i dans l'échantillon A et B respectivement.

L'indice de dissimilarité de Bray-Curtis est compris entre :

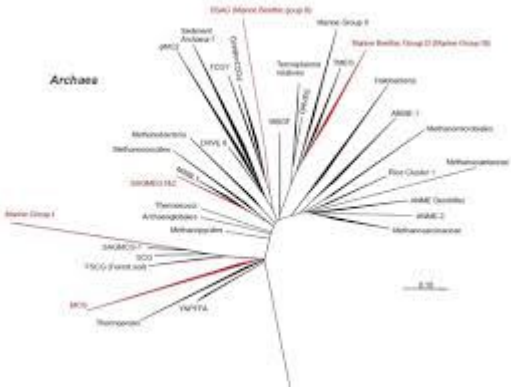
0 (les deux échantillons ont la même composition) et

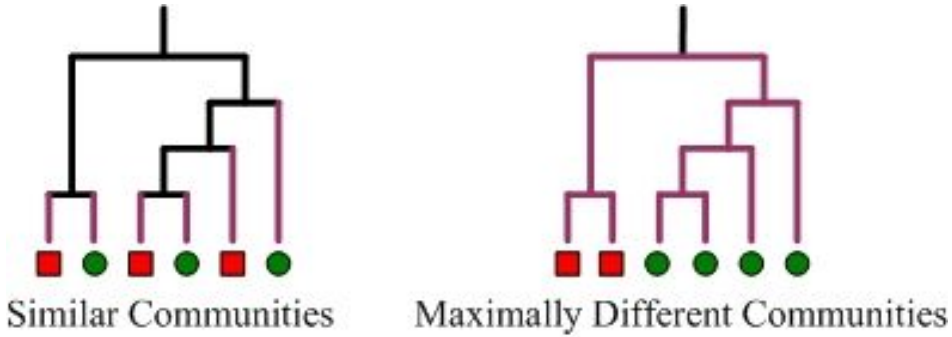
1 (les échantillons sont totalement dissemblables)

Beta-diversité : Quelques métriques

Distance unifrac :

Prise en compte de l'arbre phylogénétique



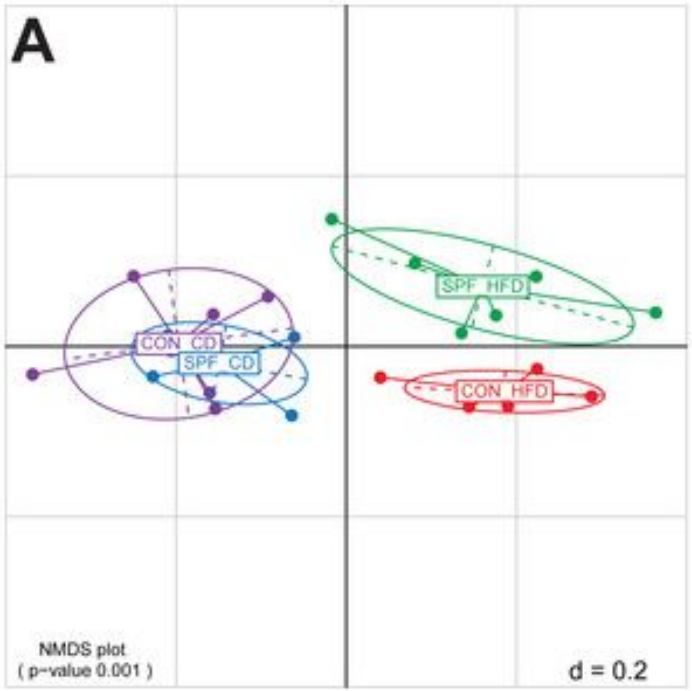


Unweighted : uniquement présence / absence d'OTU

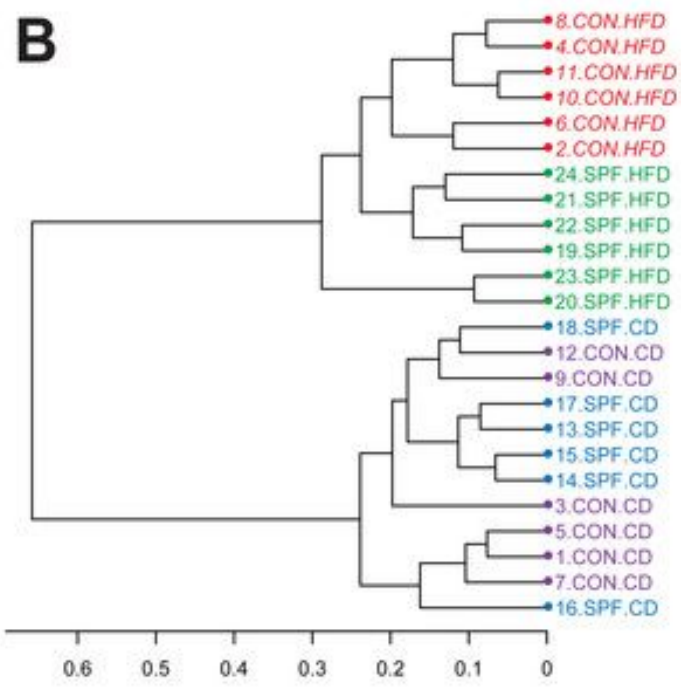
Weighted : prise en compte des proportions dans les échantillons

source graphique : https://www.mothur.org

Beta-diversité : Visualisation



PcOa



Clustering Hiérarchique

source graphique : Rhea : https://peerj.com/articles/2836/

Beta-diversité : en pratique

QIIME 2 calcule un ensemble de métriques pour vous avec une seule commande

Diversité bêta :

- Distance de Jaccard (une mesure qualitative de la dissimilitude communautaire)
- Distance de Bray-Curtis (mesure quantitative de la dissimilitude communautaire)
- la distance non pondérée de l'UniFrac (une mesure qualitative de la dissimilitude communautaire qui intègre les relations phylogénétiques entre les échantillons)
- la distance pondérée de l'UniFrac (une mesure quantitative de la dissimilitude communautaire qui intègre les relations phylogénétiques entre les échantillons)

Beta-diversité : comparaison statistique

Comment évaluer statistiquement si les bêta-diversités (ex : UniFrac pondérées) diffèrent d'un groupe à l'autre (par rapport à vos métadonnées)?

Vous pouvez effectuer une analyse **PERMANOVA**

Analyse de variance multivariée par permutation (non-paramétrique) **(Anderson 2005)**

PERMANOVA plus robuste pour les données métagénomiques **(Metagenomics, Diana Marco, 2017)**

Ces méthodes permettent de voir si il y a des différences significatives de diversité entres deux groupes donc de tester la dissimilarité entre deux communautés

Attention les résultats dépendent de la normalisation utilisée et du type de distance

Beta-diversité comparaison : en pratique

qiime2 diversity beta-group-significance Beta diversity group significance

Beta-diversité : visualisation dans le temps

qiime2 emperor plot Visualize and Interact with Principal Coordinates Analysis Plots

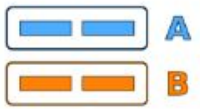
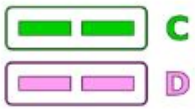
Diversités : résumé

Alpha-diversité ***α*** : permet d'estimer la diversité à l'échelle d'un échantillon

Beta-diversité **β** : permet d'estimer la différence de diversité à l'échelle d'un groupe d'échantillon

Gamma-diversité ***γ*** : alpha-diversité totale sur l'union des échantillons

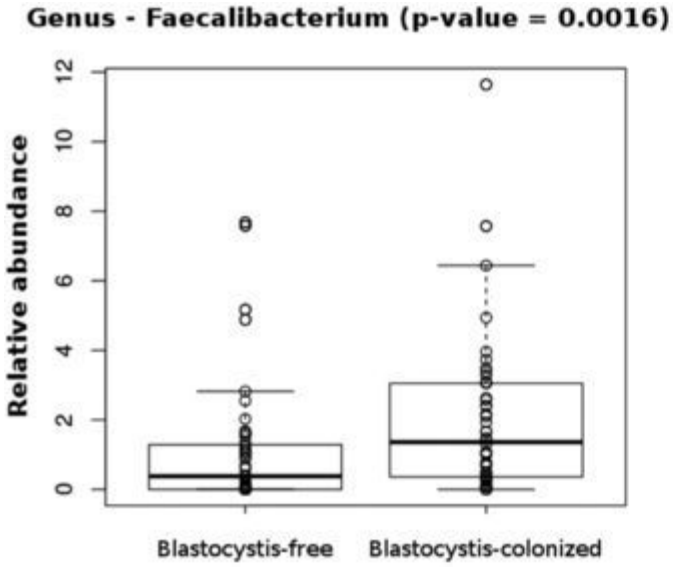
β = *γ* / *α*

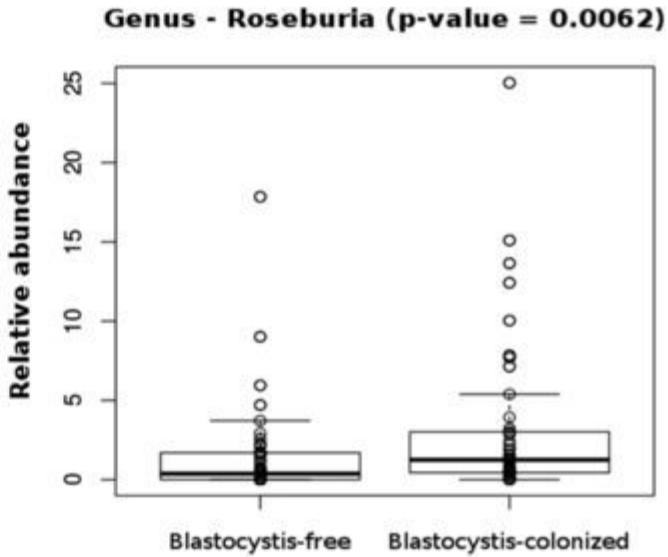
		
Échantillon 1	**Échantillon 2**	**Échantillon 3**
α-diversité = 2	**α-diversité** = 2	**α-diversité** = 1
γ-diversité = 4		
β-diversité = (4/2 + 4/2 + 4/1) / 3 = 2,67		

source graphique : L. Siegwald - 2017

Analyses différentielle - Tests statistiques

Quels sont les ASVs significativement exprimés pour une condition donnée?





Analyses différentielle : quelques tests

Tests non paramétriques : type Wilcoxon.

Ce sont ces tests qui ont longtemps été utilisés (et qui le sont toujours)

Problème : ils génèrent trop de faux négatifs

Tests utilisés avec QIIME2 : ANCOM / GNEISS

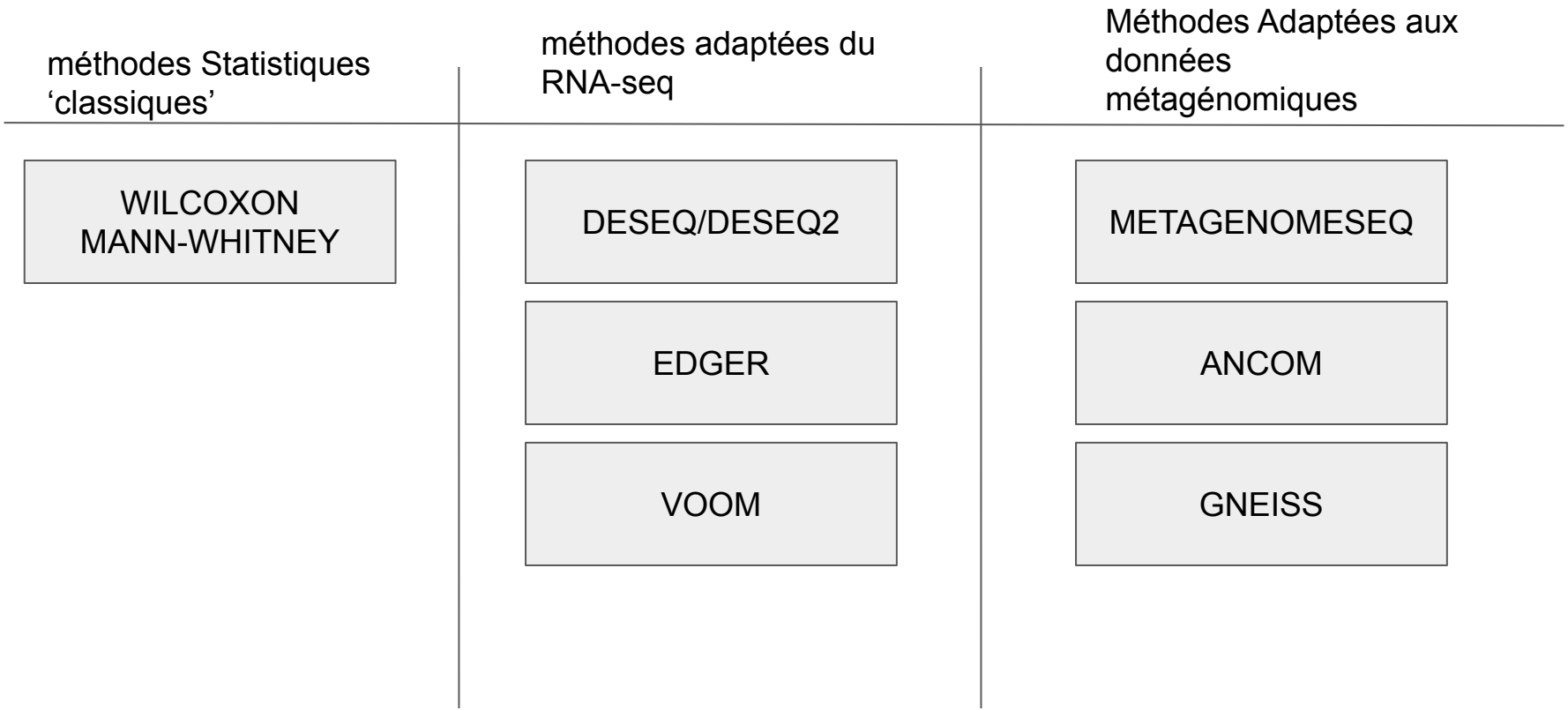
Tests alternatifs : DESeq2

De nombreuses méthodes statistiques ont été proposées dans la littérature pour comparer l'abondance (relative) des taxons entre deux groupes (p. ex. cas vs témoins). Certaines méthodes statistiques développées spécifiquement pour les données RNA-Seq, telles que DESeq, DESeq2, DESeq2, edgeR et Voom, ont été proposées pour utilisation sur les données métagénomiques. D'autre part, métagénomeSeq et l'analyse de la composition des microbiomes (ANCOM) ont été développés spécifiquement pour les ensembles de données microbiennes, qui contiennent généralement **beaucoup plus de zéros** que les données du RNA-Seq.

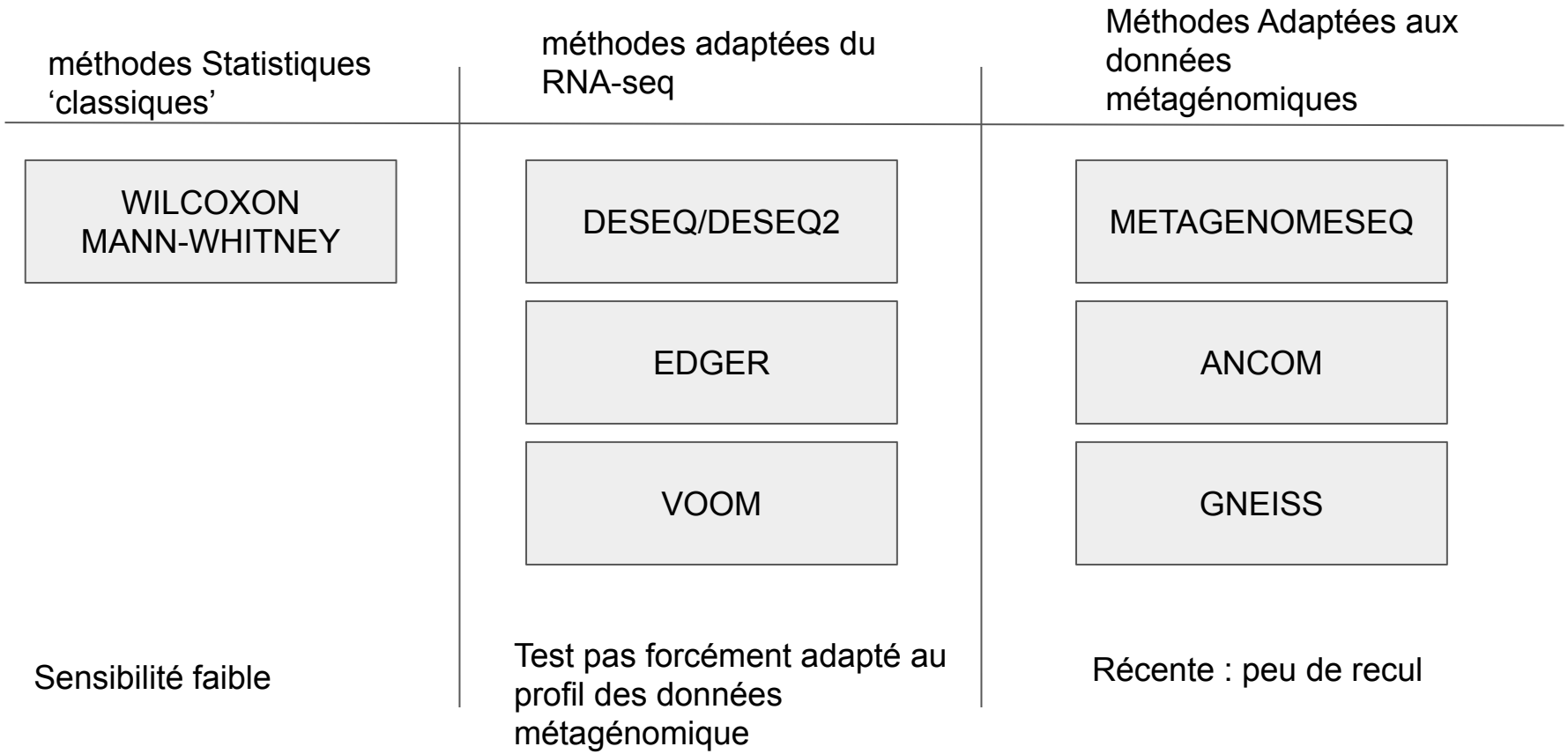
Table 2 Differential abundance methods investigated in this study

Method	Description
Wilcoxon rank-sum test	Also called the Mann-Whitney *U* test. A non-parametric rank test, which is used on the un-normalized ("None"), proportion normalized, and rarefied matrices
DESeq	nbinom Test—a negative binomial model conditioned test. More conservative shrinkage estimates compared to DESeq2, resulting in stricter type I error control
DESeq2	nbinomWald Test—The negative binomial GLM is used to obtain maximum likelihood estimates for an OTU's log-fold change between two conditions. Then Bayesian shrinkage, using a zero-centered normal distribution as a prior, is used to shrink the log-fold change towards zero for those OTUs of lower mean count and/or with higher dispersion in their count distribution. These shrunken long fold changes are then used with the Wald test for significance
edgeR	exact Test—The same normalization method (in *R*, method = RLE) as DESeq is utilized, and for differential abundance testing also assumes the NB model. The main difference is in the estimation of the dispersion, or variance, term. DESeq estimates a higher variance than edgeR, making it more conservative in calling differentially expressed OTUs
Voom	Variance modeling at the observational level—library sizes are scaled using the edgeR log counts per million (cpm) normalization factors. Then LOWESS (locally weighted regression) is applied to incorporate the mean-variance trend into precision weights for each OTU
metagenomeSeq	fitZIG—a zero-inflated Gaussian (ZIG) where the count distribution is modeled as a mixture of two distributions: a point mass at zero and a normal distribution. Since OTUs are usually sparse, the zero counts are modeled with the former, and the rest of the log transformed counts are modeled as the latter distribution. The parameters for the mixture model are estimated with an expectation-maximization algorithm, which is coupled with a moderated *t* statistic
	fitFeatureModel—a feature-specific zero-inflated lognormal model with empirical Bayes shrinkage of parameter estimates
ANCOM	Analysis of composition of microbiomes—compares the log ratio of the abundance of each taxon to the abundance of all the remaining taxa one at a time. The Mann-Whitney *U* is then calculated on each log ratio

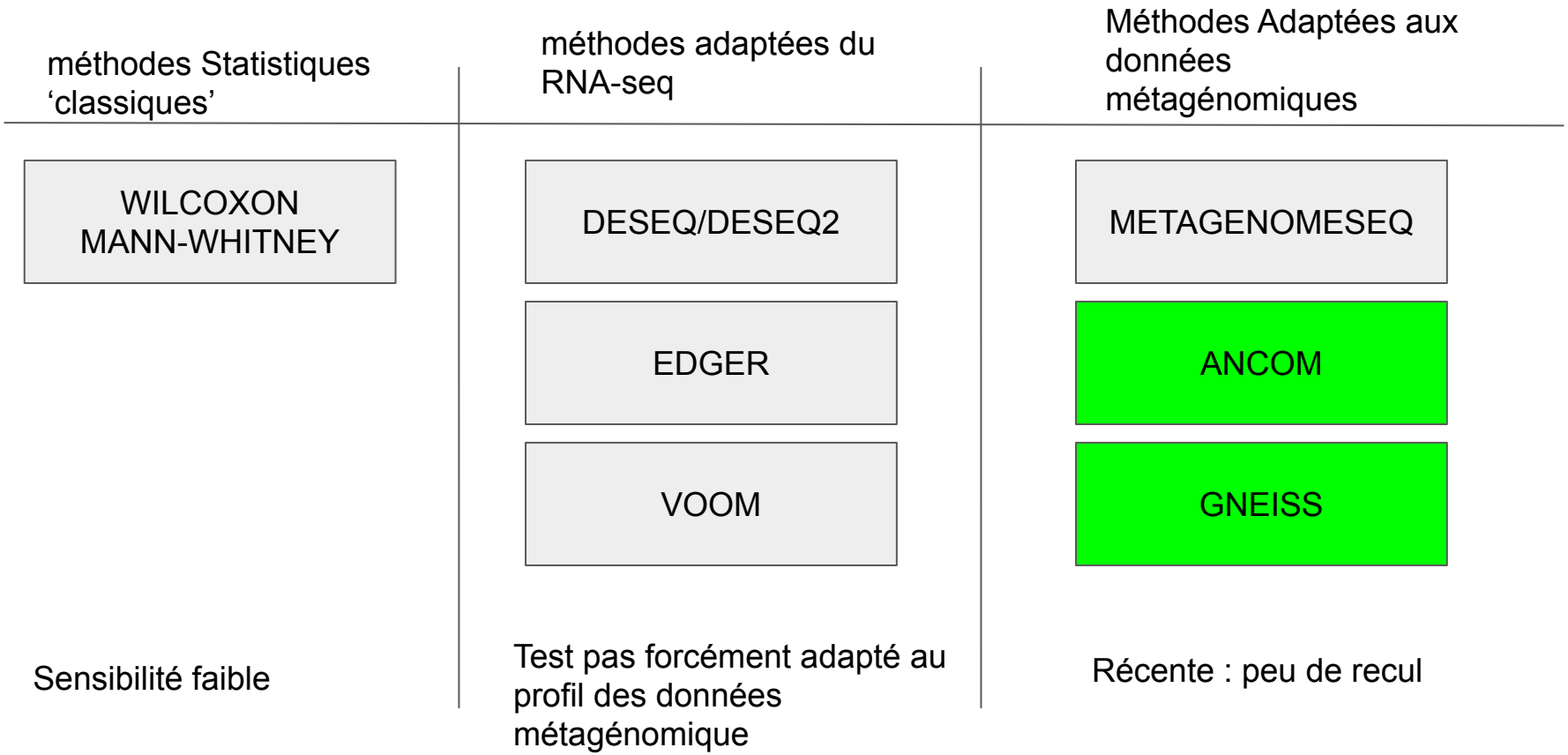
Analyses différentielles : etat de l'art



Analyses différentielles : etat de l'art



Analyses différentielles : etat de l'art



ANCOM : principe

Etape 1 : compare le rapport logarithmique de l'abondance de chaque taxon à l'abondance de tous les taxons restants deux à deux

$$H_{0ri} : E[\log(\mu_i^{(1)}/\mu_r^{(1)})] = E[\log(\mu_i^{(2)}/\mu_r^{(2)})],$$
$$\text{against } H_{ari} : E[\log(\mu_i^{(1)}/\mu_r^{(1)})] \neq E[\log(\mu_i^{(2)}/\mu_r^{(2)})]. \quad (7)$$

Ainsi, s'il existe M taxons, il effectue pour chaque taxon M - 1 tests pour chaque taxons

Etape 2 : comptage du nombre de test pour lesquels l'hypothèse nulle est rejetée
=> W
=> **A retenir : Plus W est élevé plus on rejette H0, plus le résultat est différentiellement exprimé**
Etape 3 : Estimation du cutoff pour déterminer le seuil au-delà duquel W est considéré comme différentiellement exprimés

Analyse différentielle QIIME2 : ANCOM

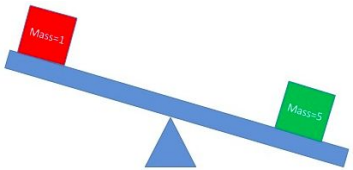
Analyse intégrée dans QIIME2

- Analyse “stringente” : les résultats sont très “sur”.
- Attention toutefois à la comparaison sur des microbiotes très différents (augmentation du FDR) (exemple : comparaison microbiote GUT et MAIN DROITE) ou avec beaucoup de 0.

Pour plus d’information lire la publication sur ANCOM
https://www.ncbi.nlm.nih.gov/pubmed/26028277

Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease* 26, 10.3402/mehd.v3426.27663, doi:10.3402/mehd.v26.27663 (2015).

Analyse différentielle QIIME2 : GNEISS

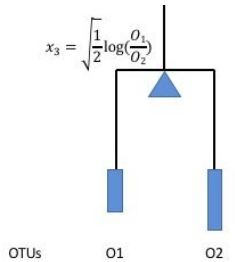


Differential abundance analysis using balances

- Rend plus de résultats que ANCOM similitude avec STAMP mais tous les taxons ne sont pas identiques
- Méthode récente, peu de recul sur son utilisation
- Peut prendre en compte beaucoup de covariables (age, sexe, poids, valeurs….)

Publication :

Morton JT, Sanders J, Quinn RA, McDonald D, Gonzalez A, Vázquez-Baeza Y, Navas-Molina JA, Song SJ, Metcalf JL, Hyde ER, Lladser M, Dorrestein PC, Knight R. 2017. Balance trees reveal microbial niche differentiation. mSystems 2:e00162-16. https://doi.org/10.1128/mSystems.00162-16.



Analyses Différentielles : conclusion

Ces techniques complexes pour l'abondance différentielle ont de l'intérêt

Les techniques adaptées du RNA-seq sont prometteuses mais elles ont tendances à **augmenter le FDR (proportion de faux positifs parmis l'ensemble des positifs)** quand la taille des échantillons varie (~10X)

Dans le cadre général, **la raréfaction est à éviter pour les tests d'analyse différentielles** (chute de sensibilité, non détection des OTUs rares)

ANCOM maintient un FDR bas pour toutes les tailles d'échantillons

Cependant, avec l'ANCOM, **la sensibilité est réduite** sur de petits ensembles de données (<20 échantillons par groupes).

Analyses Différentielles : conclusion

● La normalisation est nécessaire (Eviter donc de ne pas normaliser ou de transformer en proportion (%))

> la raréfaction est également déconseillée mais peut être appliquée si les tailles de librairies sont très variables

● **Attention à bien contrôler la taille de librairie**, les tests RNA-seq, ne fonctionnent pas correctement si la différence entre les librairies est grande.

● ANCOM constitue **un bon compromis** et son utilisation est facilitée dans QIIME2

Complément : En dehors de QIIME2 il existe ANCOM-BC2 (pacakge R) qui constitue une alternative intéressante à ANCOM

Analyse Différentielle : En pratique

TP ANCOM directement dans QIIME2

Exportation : BIOM

Qiime2 est en perpétuelle amélioration et de nouvelles méthodes analytiques et de visualisation sont régulièrement ajoutées (ex : `q2-corncob`(2020) pour l'analyse différentielle).

Si besoin, un format standard existe : BIOM qui permet d'exporter les données vers d'autres solutions (STAMP, package R phyloseq …)

Utiliser :

qiime2 tools export Export data from a QIIME 2 artifact

Explication Moyenne géométrique

Le principe de calcul de la moyenne géométrique des abondances relatives pour un échantillon est le suivant :

pour chaque espèce $i dans l'échantillon, on calcule la moyenne géométrique de toutes les abondances relatives de l'échantillon, à l'exception de l'abondance relative de l'espèce $i elle-même. Cette moyenne géométrique est donnée par :

$$g_i = \left(\prod_{j \neq i} x_j \right)^{\frac{1}{n-1}}$$

où $n est le nombre total d'espèces dans l'échantillon.