Influence du choix de l'arbre dans les études d'abondance différentielle

Antoine Bichat[1], Mahendra Mariadassou[2], Jonathan Plassais[3] & Christophe Ambroise[4]

[1]*LaMME, UEVE, Enterome – antoine.bichat@univ-evry.fr*
[2]*MaiAGE, INRA – mahendra.mariadassou@inra.fr*
[3]*Enterome – jplassais@enterome.com*
[4]*LaMME, UEVE, CNRS – christophe.ambroise@univ-evry.fr*

Résumé. En métagénomique, il est courant de réaliser des études dites d'abondance différentielle pour identifier les bactéries dont l'abondance est associée à une variable donnée, comme par exemple l'état d'un patient. De telles études s'intéressent à des centaines, voire quelques milliers de bactéries différentes, et font autant de tests. Il est donc nécessaire de corriger pour la multiplicité des tests. Après correction par des méthodes classiques (Bonferroni, Benjamini-Hochberg), quasiment aucun taxon n'est détecté, du fait de la faible puissance statistique des études (faible nombre d'échantillons, faible taille d'effet). Pour contourner ce problème, certains ont proposé d'exploiter la structure hiérarchique fournie par la taxonomie, qui constitue un *a priori* biologique sur la structure des données, pour augmenter la puissance des tests. Bien que la structure taxonomique semble naturelle, et justifiée biologiquement, elle présente néanmoins des désavantages et il peut être justifié de chercher d'autres structures hiérarchiques. Le présent travail met en évidence certains de ces désavantages et étudie l'impact du remplacement de l'arbre taxonomique par un arbre créé à partir des corrélations entre taxons.

Mots-clefs. Biostatistique, Métagénomique, Tests multiples, FDR, Analyse d'abondance différentielle, Arbres.

Abstract. In metagenomics, differential abundances studies are commonly used to identify taxa whose abundances are associated to a covariate of interest, *e.g.* healthy versus diseased. However, those studies consider hundreds to a few thousands of bacteria simultaneously and perform one test per bacteria. It is therefore necessary to correct for the multiplicity of tests using a multiple testing procedure. However, after correction with classical methods (Bonferroni, Benjamini-Hochberg), almost no association is detected due to the small statistical power in many microbiome studies (few samples, small effect size). To circumvent this problem, a hierarchical procedure has been proposed to increase the power by leveraging some a priori structure. Although the taxonomy is a natural candidate thanks to its biological grounding and has been used in the literature, it suffers from some drawbacks. In this work, we illustrate those drawbacks and show how substituting the taxonomy with a correlation-based tree can improve the detection power.

Keywords. Biostatistics, Metagenomics, Multiple testing, FDR, Differential abundance study, Trees.

1 Introduction

Le microbiote peut se définir comme l'ensemble des micro-organismes présents dans un environnement donné. Depuis plusieurs années, la recherche scientifique dans le domaine est très active sur ce sujet et découvre de nouvelles corrélations entre microbiote et maladies [8, 11, 14] ou microbiote et comportements spécifiques [4, 13] à un rythme élevé.

Les données métagénomiques apportent des informations sur les taxons et les échantillons considérés dans l'étude. Elles comportent une table d'abondance qui représente le comptage (ou la proportion) de chaque taxon dans chaque échantillon, les informations relatives aux échantillons (maladie, environnement, antibiotiques, ...) ainsi qu'une taxonomie, c'est à dire une organisation hiérarchique des espèces, pouvant être représentée sous la forme d'un arbre.

Ici, nous nous intéressons à des problématiques de type « abondance différentielle », dans lequel les échantillons ou patients sont divisés en plusieurs groupes, dont l'objectif est d'identifier les bactéries dont l'abondance varie entre les groupes.

Différentes méthodes d'abondance différentielle sur les données métagénomiques sont couramment utilisées : analyse de la variance [12], modèle linéaires généralisés [7] ou test des rangs de Wilcoxon [15] par exemple.

Dans une cohorte, plusieurs centaines d'espèces sont testées simultanément et l'utilisation d'une procédure de correction pour tests multiples s'impose. La correction de Bonferroni qui contrôle le *Family Wise Error Rate* (FWER) étant très conservative, nous lui préférons une procédure qui contrôle le *False Discovery Rate* (FDR) [1]. La procédure de contrôle la plus utilisée est celle de Benjamini-Hochberg [1]. Elle requiert néanmoins que les tests soient indépendants (ou que leur distribution jointe soit PRDS [2]) ce qui n'est pas forcément le cas pour des taxa fortement anticorrélés. La variante dite de Benjamini-Yekutieli [2], s'affranchit de cette hypothèse, mais au prix d'une très forte perte de puissance.

Note. Dans ce résumé, nous utilisons le mot taxon pour désigner l'unité biologique considérée, qui peut être une unité soit taxonomique comme l'espèce, le genre, etc soit opérationnelle comme l'OTU (*Operational Taxonomic Unit* [6]), l'ASV (*Amplicon Sequence Variant* [5]), la MSP (*Metagenomic Species Pangenome* [10]). Nous utilisons également le mot taxonomie de façon abusive pour désigner à la fois l'arbre taxonomique et la phylogénie des taxons, sauf mention explicite du contraire.

2 Procédure de contrôle du FDR

Nous rappelons ici la procédure de contrôle du FDR au niveau q [1].

Algorithme 1 Procédure de Benjamini-Hochberg au niveau q

Soit $P_{(1)} \leq \cdots \leq P_{(m)}$ les m valeurs critiques considérées ordonnées.
$r \leftarrow \max\left\{i \text{ tels que } P_{(i)} \leq i\frac{q}{m}\right\}$.
Si $r > 0$ on rejette les hypothèses correspondant aux r plus petites valeurs critiques : $P_{(1)} \leq \cdots \leq P_{(r)}$.

Dans le cas où les tests sont organisés au sein d'une hierarchie, Yekutieli [16] a proposé une variante pour augmenter la puissance de la procédure : le FDR hiérarchique.

Pour un arbre donné, nous notons $\mathcal{T}_n$ l'ensemble des fils du nœud n. Dans l'exemple figure 1, $\mathcal{T}_1 = \{3, 4\}$.

$\mathcal{T}_0$, l'ensemble des fils de la racine est testé en premier avec un FDR au niveau q. Récursivement, à chaque fois qu'un nœud t est rejeté, ses enfants directs $\mathcal{T}_t$ sont testés, avec un FDR au niveau q, jusqu'à ce qu'il n'y ait plus de nœuds à rejeter. Dans l'exemple de la figure 1,

— on rejette les nœuds 1 et 2,
— on teste les nœuds 3 et 4 (dans $\mathcal{T}_1$) puis 5 et 6 (dans $\mathcal{T}_2$) avant de rejeter $\mathbf{H}_3$ et $\mathbf{H}_5$,
— on teste les nœuds 7 et 10 à 12 (mais pas 8 et 9 puisque $\mathbf{H}_4$ n'a pas été rejetée) avant de rejeter $\mathbf{H}_{10}$ et $\mathbf{H}_{12}$.

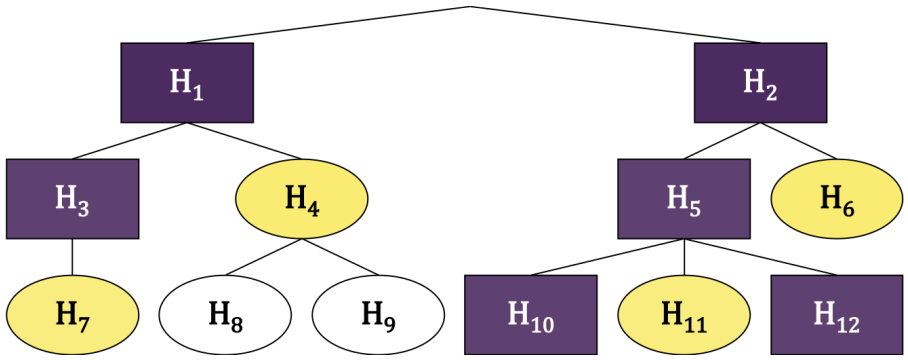

FIGURE 1 – Exemple de FDR hiérarchique . Les hypothèses sont organisées au sein d'une hiérarchie, l'hypothèse H_i est associée au nœud i. Les nœuds blancs ne sont pas testés, les nœuds jaunes sont testés mais pas rejetés, les nœuds violets sont testés et rejetés.

Cette procédure garantit un contrôle *a posteriori* du FDR au sein des feuilles au niveau

$$1.44 \times q \times \frac{\#\text{découvertes} + \#\text{nœuds testés}}{\#\text{découvertes} + 1}.$$

3 Choix de l'arbre et comparaisons

Le FDR hiérarchique nous fournit donc une procédure pour contrôler le FDR au sein de familles d'hypothèses organisées sous forme d'arbre. Dans [12], l'arbre utilisé pour la procédure hiérarchique est l'arbre phylogénétique. La phylogénie, et son pendant la taxonomie, sont en effet souvent considérées comme un bon reflet de la structure des données, et [9] a montré qu'elles sont cohérentes avec les grandes niches écologiques.

Cependant, la taxonomie et la phylogénie capturent nettement moins bien la structure des données au sein d'une niche ou de niches similaires. De plus, ces deux hiérarchies ne sont pas toujours estimables, par exemple la phylogénie lorsque l'unité biologique considérée ne correspond pas à l'espèce mais au gène ou la taxonomie lorsque l'on a affaire à des taxons provenant d'espèces inconnues.

Afin d'obtenir une information sur la structure des données, nous proposons de remplacer la taxonomie par un arbre construit à partir de la matrice de corrélation des

abondances des taxons. Cet arbre présente de plus l'avantage d'être toujours disponible.

Nous pouvons alors nous demander à quel point la taxonomie et l'arbre des corrélations sont différents. L'arbre des corrélations étant obtenu à partir des données, il est sensible aux perturbations et a une variabilité propre. Nous pouvons donc reformuler cette questions sous la forme de deux tests :

— H_0 La taxonomie est dans l'ensemble de confiance de l'arbre des corrélations.
— H_0' La distance entre la taxonomie et l'arbre des corrélations est inférieure à la distance moyenne entre l'arbre des corrélations et un arbre aléatoire.

Pour répondre à ces questions, nous utilisons la distance de Billera-Holmes-Vogtmann [3], définie comme la longueur du plus court chemin entre arbres dans l'espace des arbres $\mathscr{T}$. Elle peut aussi s'interpréter comme la longueur minimale des branches qu'il est nécessaire de contracter ou d'étendre pour transformer le premier arbre en le second.

Nous avons construit l'ensemble de confiance autour de l'arbre des corrélations en rééchantillonnant par *bootstrap* les échantillons du jeu de données. Pour chaque jeu de données *bootstrapé*, nous avons calculé un arbre des corrélations pour obtenir *in fine* une forêt d'arbres *bootstrapés*.

Pour la deuxième question, nous avons effectué une permutation aléatoire des noms des feuilles de la taxonomie et de l'arbre des corrélations. Cela permet d'échantillonner convenablement des arbres aléatoires dans $\mathscr{T}$.

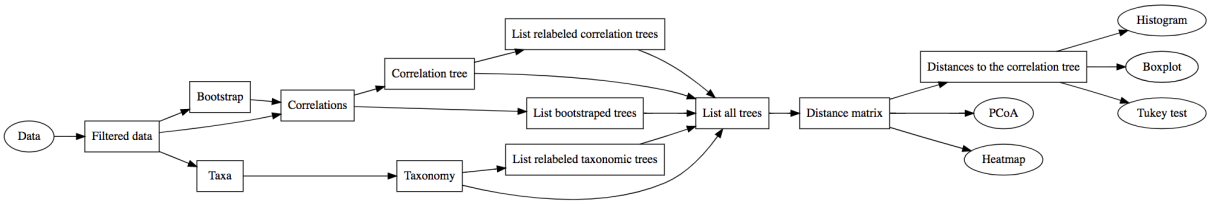

FIGURE 2 – Processus de comparaison des arbres

La construction de ces arbres supplémentaires est résumée dans la figure 2. Une fois la forêt construite, nous calculons les distances deux-à-deux au sein de celle-ci pour calculer la distribution des distances sous H_0 et H_0'.

Nous avons appliqué cette procédure aux données de [11]. La figure 3 montre que les arbres des corrélations obtenus par rééchantillonnage sont plus proches de l'arbre de corrélation initial que tous les autres arbres, en particulier celui de la taxonomie (ligne horizontale rouge). De même, l'arbre des corrélations n'est pas plus proche de la taxonomie que d'un arbre aléatoire.

Ces résultats montrent que l'arbre des corrélations est bien différent de la taxonomie et que le choix de l'un ou l'autre de ces arbres peut impacter les conclusions d'une analyse d'abondance différentielle.

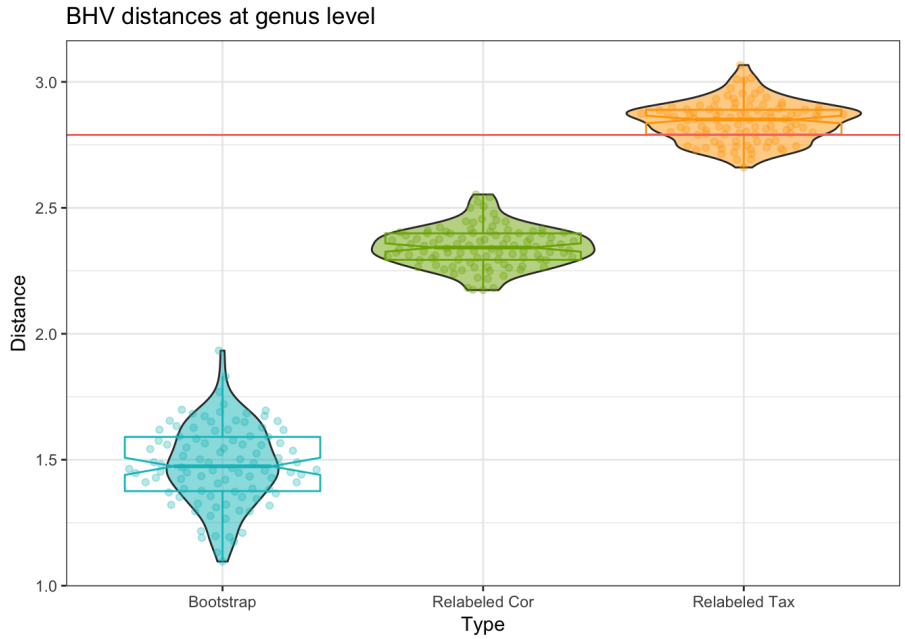

FIGURE 3 – Distance à l'arbre des corrélations

En effet, en remplaçant la phylogénie par l'arbre des corrélations dans la procédure hiérarchique [16] appliquée aux données d'abondance bactériennes de [12], nous détectons plus d'espèces significatives, avec un taux de contrôle *a posteriori* du FDR similaire au niveau des feuilles.

En conclusion, il nous paraît donc intéressant de bien réflechir à la structure hiérarchique utilisée comme *a priori* sur les abondances pour maximiser la puissance tout en contrôlant le FDR à un niveau donné. Le même travail concernnat le choix d'un arbre adéquat est en cours sur d'autres procédures de tests hiérarchiques [15].

Références

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal statistical society : series B (Methodological)*, 57(1) :289–300, 1995.

[2] Yoav Benjamini, Daniel Yekutieli, et al. The control of the false discovery rate in multiple testing under dependency. *The annals of statistics*, 29(4) :1165–1188, 2001.

[3] Louis J Billera, Susan P Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4) :733–767, 2001.

[4] Nicholas A Bokulich, Jennifer Chung, Thomas Battaglia, Nora Henderson, Melanie Jay, Huilin Li, Arnon D Lieber, Fen Wu, Guillermo I Perez-Perez, Yu Chen, et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science translational medicine*, 8(343) :343ra82–343ra82, 2016.

[5] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. Dada2 : high-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7) :581, 2016.

[6] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5) :335, 2010.

[7] Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data–the deseq2 package. *Genome Biol*, 15(550) :10–1186, 2014.

[8] Xochitl C Morgan, Timothy L Tickle, Harry Sokol, Dirk Gevers, Kathryn L Devaney, Doyle V Ward, Joshua A Reyes, Samir A Shah, Neal LeLeiko, Scott B Snapper, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome biology*, 13(9) :R79, 2012.

[9] Laurent Philippot, Siv GE Andersson, Tom J Battin, James I Prosser, Joshua P Schimel, William B Whitman, and Sara Hallin. The ecological coherence of high bacterial taxonomic ranks. *Nature Reviews Microbiology*, 8(7) :523, 2010.

[10] Florian Plaza Oñate, Emmanuelle Le Chatelier, Mathieu Almeida, Alessandra CL Cervino, Franck Gauthier, Frédéric Magoulès, S Dusko Ehrlich, Matthieu Pichaud, and Jonathan Wren. Mspminer : abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics*, 2018.

[11] Jacques Ravel, Pawel Gajer, Zaid Abdo, G Maria Schneider, Sara SK Koenig, Stacey L McCulle, Shara Karlebach, Reshma Gorle, Jennifer Russell, Carol O Tacket, et al. Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, 108(Supplement 1) :4680–4687, 2011.

[12] Kris Sankaran and Susan Holmes. structssi : simultaneous and selective inference for grouped or hierarchically structured data. *Journal of statistical software*, 59(13) :1, 2014.

[13] Gil Sharon, Daniel Segal, John M Ringo, Abraham Hefetz, Ilana Zilber-Rosenberg, and Eugene Rosenberg. Commensal bacteria play a role in mating preference of drosophila melanogaster. *Proceedings of the National Academy of Sciences*, page 201009906, 2010.

[14] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *nature*, 457(7228) :480, 2009.

[15] Jian Xiao, Hongyuan Cao, and Jun Chen. False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics*, 33(18) :2873–2881, 2017.

[16] Daniel Yekutieli. Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, 103(481) :309–316, 2008.