Comment le prétraitement des données façonne les conclusions en métagénomique ? Un benchmark reproductible pour guider l'analyse du microbiome

Emile MARDOC[a], Maxence KLOCK[b], Xavier RAFFOUX[b], Julie Aubert [c], Christelle HENNEQUET-ANTIER[d,e], Marie COURBARIAUX[f], Mathilde SOLA[b], Emmanuelle LE CHATELIER[b], Nicolas MAZIERS[b], Florence THIRION[b], Florian PLAZA OÑATE[b], Giacomo VITALI[b], Lindsay GOULET[b], Mahendra MARIADASSOU[d,e], Mathieu ALMEIDA[b], Magali BERLAND[g]

Résumé (max 300 mots)

Le microbiome intestinal, impliqué dans la digestion, l'immunité et la synthèse de métabolites bioactifs, est étudié via la métagénomique shotgun pour identifier ses liens avec les maladies humaines [1]. Cependant, l'analyse des données métagénomiques requiert une série de prétraitements visant à corriger des biais techniques (profondeur de séquençage, séquençage pairé) ou des spécificités statistiques des données (compositionnalité, distributions non gaussiennes, forte variabilité des données). Le seuil de raréfaction choisi, la normalisation ou transformation des données d'abondance : chacun de ces choix constitue des « degrés de liberté » du chercheur et peut impacter plus ou moins fortement les résultats des analyses biostatistiques menées par la suite [2], et donc les interprétations biologiques qui en sont faites.

Nous proposons un benchmark reproductible, entièrement réalisé en R, afin d'évaluer l'intérêt et l'impact des prétraitements en s'appuyant sur six grands jeux de données publics (n > 300) dans des environnements cliniques et géographiques différents [3]. Nous étudions le choix du prétraitement de façon différenciée selon le type d'analyse mené ensuite : (i) abondance différentielle, (ii) étude de la diversité alpha, (iii) analyses multivariées (PCoA, NMDS), (iv) classification supervisée (random forests) et (v) inférence de réseaux d'interactions microbiennes. Le critère d'évaluation est la stabilité des signaux statistiques, estimée lors de plusieurs sous-échantillonnages. Ce benchmark se démarque de ceux déjà publiés [4], [5], [6], [7] en se concentrant sur les spécificités des données métagénomiques shotgun, et en ne se limitant pas aux analyses d'abondance différentielle. Nos résultats permettent de proposer des recommandations pratiques pour choisir les combinaisons prétraitement-analyse les plus adaptées selon le contexte biologique et le type d'analyse. Ce travail s'inscrit dans une démarche de science ouverte et reproductible — un pas vers une meilleure standardisation des analyses des données du microbiome.

[a] Université Paris-Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France, emile.mardoc@inrae.fr

[b] Université Paris-Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France

[c] Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France

[d] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

[e] Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350, Jouy-en-Josas, France

[f] Sorbonne Université, Maison des Modélisations Ingénieries et Technologies (SUMMIT), 75005 Paris, France

[g] Université Paris-Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France, magali.berland@inrae.fr

Mots-clefs (3 à 5) : Benchmark, Microbiome intestinal, Métagénomique shotgun, Science reproductible, Prétraitements des données

Références

[1] Y. Fan et O. Pedersen, « Gut microbiota in human metabolic health and disease », *Nat Rev Microbiol*, vol. 19, n° 1, p. 55-71, janv. 2021, doi: 10.1038/s41579-020-0433-9.

[2] R. Aghdam, X. Tang, S. Shan, R. Lankau, et C. Solís-Lemus, « Human limits in machine learning: prediction of potato yield and disease using soil microbiome data », *BMC Bioinformatics*, vol. 25, n° 1, p. 366, nov. 2024, doi: 10.1186/s12859-024-05977-2.

[3] E. Le Chatelier, M. Sola, et F. Plaza Oñate, « Large-scale shotgun metagenomic cohorts of gut microbiota samples with standardized minimal metadata from adults in industrialized countries ». Recherche Data Gouv, 2025. doi: 10.57745/UPITJ0.

[4] M. B. Pereira, M. Wallroth, V. Jonsson, et E. Kristiansson, « Comparison of normalization methods for the analysis of metagenomic gene abundance data », *BMC Genomics*, vol. 19, n° 1, p. 274, déc. 2018, doi: 10.1186/s12864-018-4637-6.

[5] L. Yang et J. Chen, « A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions », *Microbiome*, vol. 10, n° 1, p. 130, août 2022, doi: 10.1186/s40168-022-01320-0.

[6] J. Pelto, K. Auranen, J. V. Kujala, et L. Lahti, « Elementary methods provide more replicable results in microbial differential abundance analysis », *Briefings in Bioinformatics*, vol. 26, n° 2, p. bbaf130, mars 2025, doi: 10.1093/bib/bbaf130.

[7] S. D. Gamboa-Tuz *et al.*, « Commonly used compositional data analysis implementations are not advantageous in microbial differential abundance analyses benchmarked against biological ground truth ». Bioinformatics, février 2025. doi: 10.1101/2025.02.13.638109.