

Sujet de stage de master M2 : Comparaisons de différentes stratégies analytiques pour les données de métabarcoding et de métagenomique dans le cadre de deux projets de recherche.

Unité : UMR CIRAD/Université de Montpellier QUALISUD – Montpellier, France.

Encadrement : Florentin Constancias (Chercheur CIRAD) et Julie Reveillaud (C.R. INRA)

Contexte & problématique :

Le développement des outils de séquençage de seconde génération (*i.e.*, NGS) a révolutionné l'écologie microbienne en permettant l'obtention massive de séquences provenant des communautés microbiennes d'environnements variés (*e.g.*, sol, aliments, compartiments du corps humain) via des approches de métabarcoding (*a.k.a.*, *métagenomique* ciblée) ou de métagenomique *shot-gun*. En revanche, les stratégies analytiques, bio-informatiques et statistiques développées pour l'analyse de ces données ont connues des développements plus récents.

Par exemple, la normalisation de données issues des approches de métabarcoding fait l'objet de débats, tout comme les métriques de diversité adaptées à ce type de données. Similairement, de nouveaux outils de profilage taxonomique de données de métagenomique *shot-gun* sont fréquemment publiés sans réel information sur la résolution taxonomique à laquelle ils permettent d'accéder.

Dans ce contexte, le but de ce stage est de (i) comparer différents pipelines bio-informatiques adaptés aux données métabarcoding et d'évaluer leurs apports vis à vis des approches classiques, de (ii) statuer sur le potentiel de métriques de diversité (alpha/beta) récemment développées, et enfin (iii) d'évaluer le potentiel de différents outils bio-informatiques adaptés aux données de métagenomique *shot-gun* en terme de caractérisation taxonomique et fonctionnelle.

L'ensemble de ces analyses comparatives sera réalisée dans le cadre de projets de recherche concrets au travers des données de métabarcoding et métagenomique *shot-gun*.

Objectifs détaillés :

Les objectifs de ce stage sont :

- Pour les données de type métabarcoding, de :
 - o Comparer les différents pipeline bio-informatiques (*e.g.*, [swarm2](#), [dada2](#), [deblur](#)), en terme de diversité (alpha), de structure de communautés et taxonomie.
 - o Explorer les différentes métriques de diversité récemment développées (*i.e.*, [TINA & PINA](#); [Breakaway](#), [beta & divnet](#), ...) et comparer les *pattern* obtenus avec des métriques classiques (*e.g.*, bray-curtis, unifrac).

- Pour les données métagenomique *shot-gun*, d'explorer le potentiel de caractérisation taxonomique et fonctionnelle des outils :
 - o de (i) profilage, taxonomique (*i.e.*, [Metataxa2](#), [kraken](#), [centrifuge](#), [metaphlan2](#), [SLIMM](#)), intra-spécifique (*i.e.*, [panphlan](#), [strainphlan](#)) ainsi que fonctionnel (*i.e.*, [humann2](#), [superfocus](#))
 - o complémentaires (ii), basées sur la reconstruction et l'analyses de MAGs (*i.e.*, Genomes assemblés à partir de metagenomes), via la plateforme [Anvi'o](#).

Projets - Données disponibles :

Afin de répondre à l'ensemble des problématiques, le(a) candidat(e) sera amené(e) à travailler dans le cadre de deux projets de recherche, sur les deux jeux de données suivants (déjà séquencés et non publiés) :

- Un jeu de données comprenant 375 échantillons de pommes caractérisées par métabarcoding 16S et ITS (marqueur taxonomiques bactériens et fongiques, respectivement). Ce projet vise à inventorier les communautés bactériennes et fongiques à la surface des fruits et à comprendre l'influence de différents facteurs (*i.e.*, variété, pratiques culturelles, parcelle) sur ces communautés.

- Un second jeu de données comprenant 50 échantillons issus de différents sites de la cavité buccale de patients caractérisés par métagenomique *shot-gun*. Un des objectifs de ce projet est de caractériser la diversité microbienne spécifique et intraspécifique entre les patients et les sites de la cavité buccale. N.B. : Pour ce second projet, certaines étapes nécessitant des temps de calculs importants ont déjà été effectuées.

Des communautés *mock communities* séquencés, publiés et disponibles seront également analysés en tant que contrôles pour les différents tests/comparaisons.

Profil recherché :

Nous recherchons un(e) candidat(e) motivé(e), curieux(se) et ambitieux(se) afin de contribuer activement à ces projets de recherche. Un fort attrait pour l'écologie microbienne, de solides compétences en linux, une maîtrise du langage R ainsi que des connaissances en analyses statistiques des communautés microbiennes sont nécessaires. L'ensemble des analyses devant être facilement lisibles et reproductibles, une maîtrise des outils de bonnes pratiques de reproductibilité scientifique serait considérée comme un plus (*e.g.*, github, Rmarkdown).

Le (la) stagiaire travaillera étroitement avec d'autres doctorants impliqués dans ces projets ainsi qu'avec des membres du Collectif de Bioinformatique de Baillarguet (C2B).

Il (elle) aura pour missions : (i) l'installation des outils et dépendances manquantes sur le cluster de calcul – la plupart des outils étant déjà installés, (ii) le traitement bioinformatique (cluster linux, SGE), (iii) la visualisation (R : ggplot2 – Anvi'o)) et (iv) l'analyse biostatistique (R : phyloseq, vegan, mvabund) des résultats.

Ce stage n'aboutira - *a priori* – pas sur un sujet de thèse mais à l'issue de ce travail le (la) candidat(e) disposera d'une expertise dans l'analyse de données métabarcoding et de métagenomique facilement valorisable par la suite.

Gratification :

Mous ne pouvons pas proposer plus que le minimum légal, soit ~525 – 577 €/ mois.

Lieu du stage :

Le (la) stagiaire sera localisé au CIRAD de Montpellier sur les sites de Lavalette et de Baillarguet.

Contact :

Pour toute information concernant ce stage de M2, merci de contacter [Florentin Constancias](mailto:florentin.constancias@cirad.fr) et [Julie Reveillaud](mailto:julie.reveillaud@cirad.fr) par email (florentin.constancias@cirad.fr julie.reveillaud@cirad.fr) en renseignant comme objet [STAGE Bioinfo].

Pour candidater, merci de joindre un *curriculum vitae* ainsi qu'une lettre de motivation succincte.

Mots clefs :

Microbiome, 16S/ITS Métabarcoding, Benchmarking, NGS, métagenomique shot-gun, Ecologie microbienne.

Références :

McMurdie, P.J. & Holmes, S. (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol 10(4). <https://doi.org/10.1371/journal.pcbi.1003531>

Eren, A.M. et al. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319 <https://doi.org/10.7717/peerj.1319>

Callahan, B.J., et al. (2016). DADA2: High-resolution sample inference from Illumina amplicon data Nat. Methods, 13 pp. 581-583

Schmidt, T.S. B. et al. (2017). A family of interaction-adjusted indices of community similarity. ISME Journal, 11(3), 791–807. <https://doi.org/10.1038/ismej.2016.139>

Weiss, S. et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5, 27

Walsh, A.M. et al. (2018). Species classifier choice is a key consideration when analysing low-complexity food microbiome data. Microbiome, 6(1), 50. <https://doi.org/10.1186/s40168-018-0437-0>