Méthodes

Séquençage

1. Extraction d'ADN du génome

L'ADN du génome total des échantillons a été extrait par la méthode CTAB/FDS. La concentration et la pureté de l'ADN ont été contrôlées sur des gels d'agarose. à 1%. Selon la concentration, l'ADN a été dilué à 1ng/μL en utilisant de l'eau stérile.

2. Génération d'amplicons

Les gènes ARNr 16S/ARNr 18S/ITS de différentes régions (16SV4/16SV3/16SV3-V4/16SV4-V5, 18S V4/18S V9, ITS1/ITS2, Arc V4) ont été amplifiés à l'aide d'une amorce spécifique (par exemple, 16S V4 : 515F -806R, 18S V4 : 528F-706R, 18S V9 : 1380F-1510R, et al) avec le code-barres. Toutes les réactions PCR ont été réalisées avec le mélange maître PCR haute fidélité Phusion® (New England Biolabs).

3. Quantification et qualification des produits PCR

Mélanger le même volume de tampon de chargement 1X (contenant du vert SYB) avec des produits PCR et exécuter une électrophorèse sur gel d'agarose à 2 % pour la détection. Des échantillons avec une bande principale brillante entre 400 et 450 pb ont été choisis pour d'autres expériences.

4. Mélange et purification des produits PCR

Les produits de PCR ont été mélangés dans des proportions d'équidensité. Le mélange de produits de PCR a ensuite été purifié avec le kit d'extraction de gel Qiagen (Qiagen, Allemagne).

5. Bibliothèque de préparation et de séquençage.

Des bibliothèques de séquençage ont été générées à l'aide du kit NEBNext Ultra DNA Library Pre® pour Illumina, conformément aux recommandations du fabricant,

et des codes d'index ont été ajoutés. La qualité de la bibliothèque a été évaluée sur le fluorimètre Qubit® 2.0 (Thermo Scientific) et le système Agilent Bioanalyzer 2100. Enfin, la bibliothèque a été séquencée sur une plate-forme Illumina et des lectures appariées de 250 pb ont été générées.

L'analyse des données

1. Assemblage de lectures appariées et contrôle qualité.

a. Division des données.

Les échantillons ont été assignés à des lectures appariées en fonction de leur code-barres unique et tronqués en coupant le code-barres et la séquence d'amorce.

b. Montage de séquence.

Les lectures correspondantes ont été fusionnées à l'aide de FLASH (V1.2.7, http://ccb.jhu.edu/software/FLASH/)[1] , un outil d'analyse très rapide et précis, conçu pour fusionner les lectures correspondantes lorsqu'au moins certaines des lectures chevauchent la lecture générée à partir de l'extrémité opposée du même fragment d'ADN, et les séquences d'épissage sont appelées étiquettes brutes.

c. Fuite de données.

Le filtrage de qualité sur les balises brutes a été effectué dans des conditions de filtrage spécifiques pour obtenir des balises propres de haute qualité[2] selon le processus de qualité contrôlée QIIME (V1.7.0, http://qiime.org/index.html)[3].

d. Élimination des Chimères.

Les balises ont été comparées à la base de données de référence (Gold Database, http://drive5.com/uchime/uchime_download.html) à l'aide de l'algorithme UCHIME (UCHIME Algorithm, http://www.drive5.com/usearch/manua/uchime_something.html)[4] pour détecter les séquences chimères, puis les séquences chimères[5] ont été supprimées. Ainsi, les balises effectives ont finalement été obtenues.

2. Groupe OTU et annotation d'espèces.

a. Fabrication OTU.

L'analyse de séquence a été réalisée avec le logiciel Uparse (Uparse v7.0.1001, http://drive5.com/uparse/)[6]. Les séquences présentant une similarité ≥ 97 % ont été attribuées à la même OTU. La séquence représentative de chaque OTU a été évaluée pour obtenir plus d'informations.

b. Annotation des espèces.

Pour chaque séquence représentative, la base de données GreenGene (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi)[7] basée sur le classificateur RDP 3 (Version 2.2, http://sourceforge.net/projects/rdp-classifier/)[8] algorithme d'annotation des informations taxonomiques.

c. Construction de la relation phylogénétique.

Pour étudier la relation phylogénétique des différentes OTU et la différence des espèces dominantes dans différents échantillons (groupes), un alignement de séquences multiples a été réalisé à l'aide du logiciel MUSCLE (Version 3.8.31, http://www.drive5.com/muscle/)[9].

d. Normalisation des données.

Les informations d'abondance OTU sont normalisées à l'aide d'un standard de numéro de séquence correspondant à l'échantillon avec moins de séquences. Toutes les analyses ultérieures de la diversité alpha et de la diversité bêta ont été effectuées sur la base de ces données de sortie normalisées.

3. Alpha Diversité.

La diversité alpha est appliquée dans l'analyse de la complexité de la diversité des espèces pour un échantillon à travers 6 indices, y compris les espèces observées, Chao1, Shannon, Simpson, ACE, Bonne couverture. Tous ces indices dans nos échantillons ont été calculés avec QIIME (Version 1.7.0) et affichés avec le logiciel R (Version 2.15.3).

Deux indices ont été sélectionnés pour identifier la richesse communautaire :
Chao - the Chao1 estimator (http://www.mothur.org/wiki/Chao)
ACE - the ACE estimator (http://www.mothur.org/wiki/Ace)

Deux indices ont été sélectionnés pour identifier la diversité communautaire :
Shannon - the Shannon index (http://www.mothur.org/wiki/Shannon)
Simpson - the Simpson index (http://www.mothur.org/wiki/Simpson)

Un index pour caractériser la profondeur de séquençage:
Coverage - the Good's coverage (http://www.mothur.org/wiki/Coverage).

4. Diversité bêta.

L'analyse de la diversité bêta a été utilisée pour évaluer les différences d'échantillon dans la complexité des espèces. La diversité bêta en unifrac pondéré et non pondéré a été calculée à l'aide du logiciel QIIME (version 1.7.0).

L'analyse typologique a été précédée d'une analyse en composantes principales (ACP), qui a été appliquée pour réduire la dimension des variables d'origine à l'aide du package FactoMineR et du logiciel ggplot2 park in R (version 2.15.3).

L'analyse des coordonnées principales (PCoA) a été effectuée pour obtenir les coordonnées principales et visualiser des données multidimensionnelles complexes. Une matrice de distance unifrac ou pondérée non pondérée entre les échantillons obtenus ci-dessus a été transformée en un nouvel ensemble d'axes orthogonaux, dans lequel le facteur de variation maximal est représenté par la première coordonnée principale, et le deuxième maximum par la coordonnée principale, et ainsi de suite. L'analyse PCoA a été présentée à l'aide du package WGCNA, des packages stat et du package ggplot2 dans le logiciel R (version 2.15.3).

Méthode de groupe de paires non pondérées avec moyennes arithmétiques (UPGMA). Le regroupement a été effectué comme un type de méthode de regroupement hiérarchique pour interpréter la matrice de distance à l'aide de la moyenne des liens et a été effectué par le logiciel QIIME (version 1.7.0).

Référence

[1] Magoč T, Salzberg S L. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27.21 (2011): 2957-2963.

[2] Bokulich, Nicholas A., et al. Quality-filtering vastly improves diversity estimates from Illuminaamplicon sequencing. Nature methods 10.1 (2013): 57-59.

[3] Caporaso, J. Gregory, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods 7.5 (2010): 335-336.

[4] Edgar, Robert C., et al. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27.16 (2011): 2194-2200.

[5] Haas, Brian J., et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons.Genome research 21.3 (2011): 494-504.

[6] Edgar, Robert C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature methods 10.10 (2013): 996-998.

[7] DeSantis, Todd Z., et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and environmental microbiology 72.7 (2006): 5069-5072.

[8] Wang, Qiong, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology 73.16 (2007): 5261-5267.

[9] Edgar R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research32.5(2004): 1792-1797.