锐翌基因
REALGENE

ITS rDNA Amplicon 测序

方法描述

上海锐翌基因科技有限公司

目录

方法.....4
1 测序与数据质控.....4
1.1 序列优化统计.....4
2 OTU 分析.....4
2.1 OTU 聚类.....4
2.2 抽平处理.....5
2.3 Core Microbiome 分析.....5
2.4 OTU Venn 分析.....5
2.5 OTU PCA 分析.....5
2.6 Specaccum 物种累积曲线.....5
3 物种分类和丰度分析.....6
3.1 物种注释分析.....6
3.2 物种注释结果统计.....6
3.3 物种丰度分析.....6
3.4 物种热图分析.....6
3.5 Rank Abundance 曲线.....7
4 Alpha 多样性分析.....7
4.1 单个样品多样性分析.....7
5 Beta 多样性分析.....7
6 显著性差异分析.....8
6.1 LEfSe 分析.....8

6.2 组间差异分析.....8

方法

1 测序与数据质控

1.1 序列优化统计

对 ITS rDNA 高变区测序序列进行测序，测序为 ITS2 区；

通过 Pandaseq[1]软件利用重叠关系将双末端测序得到的成对 Reads 拼接成一条序列，得到高变区的长 Reads。然后使用内部撰写的程序对拼接后的 Reads 进行如下处理，获取 Clean Reads：

1）去除平均质量值低于 20 的 Reads；

2）去除 Reads 含 N 的碱基数超过 3 个的 Reads；

3）Reads 长度范围为 220~500 nt。

统计 Clean Reads 的长度分布和数量。

参考文献：

[1] Andre P Masella, Andrea K Bartram, Jakub M Truszkowski, Daniel G Brown and Josh D Neufeld. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics 2012, 13:31.

2 OTU 分析

2.1 OTU 聚类

为便于下游物种多样性分析，将长 Reads 聚类为 OTUs(Operational Taxonomic Units)。首先把拼接的长 Reads 中的 singletons（对应 reads 只有一条的序列）过滤掉，因为 singletons 可能由于测序错误造成，故将这部分序列去除，不加入聚类分析，利用 Usearch 在 0.97 相似度下进行聚类，对聚类后的序列进行嵌合体过滤后，得到用于物种分类的 OTU，每个 OTU 被认为可代表一个物种。

参考文献：

[1] Edgar, R.C. (2013) UPARSE: Highly accurate OTU sequences from microbial amplicon reads, Nature Methods.

2.2 抽平处理

为避免因样品数据大小不同而造成分析时的偏差，我们在样品达到足够的测序深度的情况下，对每个样品进行随机抽平处理。测序深度用 Alpha 多样性指数来衡量。抽平的参数必须在保证测序深度足够的前提下去选取。

2.3 Core Microbiome 分析

根据样品的共有 OTU 以及 OTU 所代表的物种，可以找到 Core microbiome（覆盖 100%样品的微生物组）。

2.4 OTU Venn 分析

Venn 图可以很好的反应组间共有以及组内特有的 OTU 数目。利用 R 语言的 VennDiagram 包的 venn.diagram 函数实现。

2.5 OTU PCA 分析

PCA 可以初步的反映出不同处理或不同环境间的样品可能表现出分散和聚集的分布情况，从而可以判断相同条件的样品组成是否具有相似性。利用 R 语言 ade4 包里的 dudi.pca 函数实现。

2.6 Specaccum 物种累积曲线

在生物多样性和群落调查中，物种累积曲线被广泛用于抽样量充分性的判断以及物种丰富度估计。

利用物种累积曲线判断抽样量是否充分是根据曲线的特征来判断：如果曲线一直急剧上升，几为直线，表明抽样量不足，需要增加抽样量；如果曲线在急剧上升后变为一渐近线，上升舒缓，则表明抽样充分。

3 物种分类和丰度分析

3.1 物种注释分析

从各个 OTU 中挑选出一条序列，作为该 OTU 的代表序列。将该代表序列，用 RDP 方法，与已知物种的 ITS 数据库（RDP，http://rdp.cme.msu.edu/）进行比对，从而对每个 OTU 进行物种归类。归类后，根据每个 OTU 中序列的条数，从而得到 OTU 丰度表。

参考文献：

[1] Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysisNucl. Acids Res. 42(Database issue):D633-D642; doi: 10.1093/nar/gkt1244 [PMID: 24288368]

[2] Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 73(16):5261-5267; doi:10.1128/AEM.00062-07 [PMID: 17586664]

3.2 物种注释结果统计

根据物种注释情况，统计每个样品注释到各分类水平(Kingdom, Phylum, Class, Order, Family, Genus)上的相对丰度，由此可轻松了解注释到各分类水平的整体情况。

3.3 物种丰度分析

在门、纲、目、科、属水平，将每个注释上的物种或 OTU 在不同样品中的序列数整理在一张表格，形成 profiling 柱状图、星形图及统计表。

3.4 物种热图分析

物种热图利用颜色梯度可以很好反应出样品在不同物种下的丰度大小以及物种聚类、样品聚类信息。利用 R 语言的 gplots 包的 heatmap.2 函数实现。

3.5 Rank Abundance 曲线

用于同时解释样品多样性的两个方面，即样品所含物种的丰富程度和均匀程度。物种的丰富程度由曲线在横轴上的长度来反映，曲线越宽，表示物种的组成越丰富；物种组成的均匀程度由曲线的形状来反映，曲线越平坦，表示物种组成的均匀程度越高。

4 Alpha 多样性分析

4.1 单个样品多样性分析

Alpha 多样性（Alpha diversity）是对单个样品中物种多样性的分析，包括 observed species 指数、chao1 指数、shannon 指数、simpson 指数以及 PD_whole_tree 指数。利用 QIIME 软件计算样品的 Alpha 多样性指数的值，并做出相应的稀释曲线。

稀释曲线是利用已测得 16S rDNA 序列中已知的各种 OTUs 的相对比例，来计算抽取 n 个（n 小于测得 Reads 序列总数）Reads 时各 Alpha 多样性指数的期望值，然后根据一组 n 值（一般为一组小于总序列数的等差数列）与其相对应的 Alpha 多样性指数的期望值做出曲线来。并作出 Alpha 多样性指数的统计表格。

参考文献：

[1] Paul FK, Josephine Y A. (2004). Bacterial diversity in aquatic and other environments:what 16S rDNA libraries can tell us. *FEMS Microbiol.Ecol* 47:161-177.

4.2 单个样品差异分析

分别对 Alpha diversity 的各个指数进行秩和检验分析(若两组样品比较则使用 R 中的 wilcox.test 函数，若两组以上的样品比较则使用 R 中的 kruskal.test 函数)，通过秩和检验筛选不同条件下的显著差异的 Alpha diversity 指数。

5 Beta 多样性分析

5.1 UniFrac 热图分析

Beta 多样性分析反映了不同样品在物种多样性方面存在的差异大小。分析各类群在样品中的含量，进而计算出不同样品间的 Beta 多样性值。本分析中通过 QIIME 软件，采用迭

代算法，分别在加权物种分类丰度信息和不加权物种分类丰度信息的情况下进行差异计算，得到最终的统计分析结果表并做出组间的距离 box 图及 PCoA 展示图。

5.2 PCoA 分析

为了进一步展示样品间物种多样性差异，使用主坐标分析（Principal coordinates analysis，PCoA）的方法展示各个样品间的差异大小

6 显著性差异分析

6.1 LEfSe 分析

LDA EffectSize （LEfSe 分析）：LEfSe 采用线性判别分析（LDA）来估算每个组分（物种）丰度对差异效果影响的大小，找出对样品划分产生显著性差异性影响的群落或物种。本分析采用 LEfSe Tools 进行。

参考文献：

[1] Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. Genome Biol. 12, R60.

6.2 组间差异分析

使用秩和检验的方法对不同分组之间的进行显著性差异分析，以找出对组间划分产生显著性差异影响的物种。本分析对于两组间的差异分析采用 R 语言 stats 包的 wilcox.test 函数，对于两组以上的组间差异分析采用 R 语言 stats 包的 kruskal.test 函数。