宏基因组项目信息分析

目　录

宏基因组学分析能够更真实的反应样本中微生物组成、互作情况，同时在分子水平对其代谢通路、基因功能进行研究[1-4]。

1. 测序结果预处理

1）使用 Readfq（V8 ,https://github.com/cjfields/readfq）对 Illumina HiSeq 测序平台获得的原始数据(Raw Data)进行预处理，获取用于后续分析的有效数据(Clean Data)。具体处理步骤如下：a）去除所含低质量碱基（默认质量阈值为≤38）超过一定比例（默认长度值为 40bp）的 reads；b）去除 N 碱基达到一定比例的 reads（默认长度值为 10bp）；c）去除与 Adapter 之间 overlap 超过一定阈值（默认长度值为 15bp）的 reads。

2）如果样品存在宿主污染，需与宿主数据库进行比对，过滤掉可能来源于宿主的 reads，默认采用 Bowtie2 软件(version2.2.4 , http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)，参数设置[14,15]: --end-to-end, --sensitive, -I 200, -X 400。

2. Metagenome 组装

1）单个样品组装：

对于肠道或粪便等非复杂环境样本使用 SOAPdenovo 软件（V2.04, http://soap.genomics.org.cn/soapdenovo.html）对 Clean Data 进行组装分析[6]，参数设置[7,8,9,10]：-d 1, -M 3, -R, -u, -F, -K 55；对于复杂环境样本如水体，土壤等，使用 MEGAHIT 软件（v1.0.4-beta）进行组装，组装参数设置[11]：--presets meta-large（--min-count 2 --k-min 27 --k-max 87 --k-step 10）；然后将组装得到的 Scaffolds 从 N 连接处打断，得到不含 N 的 Scaftigs[7,12,13]。使用 Bowtie2 软件（Bowtie2.2.4）将各样品的 Clean Data 分别比对至其 Scaftigs 上，获取未被利用上的 PE reads，参数设置[7]：--end-to-end, --sensitive, -I 200, -X 400。

2）混合组装：

将各样品未被利用上的 reads 合并后，使用 SOAPdenovo（V2.04）/MEGAHIT（v1.0.4-beta）进行混合组装[14,15,16]，参数设置同单样本组装；将混合组装的 Scaffolds 从 N 连接处打断，得到 Scaftigs。将单样品和混合组装生成的 Scaftigs，过滤掉 500bp 以下的片段[7,17,18,19]，并进行统计分析。

3. 基因预测及丰度分析

1) 使用 MetaGeneMark（V2.10, http://topaz.gatech.edu/GeneMark/）对各样品及混合组装的 Scaftigs（>=500bp）进行 ORF 预测[12,13,14,15,20,21]，并过滤掉预测结果中长度小于 100nt[19,7,13,16,17] 的信息，均采用默认参数。

2) 对 ORF 预测结果，采用 CD-HIT[22,23] 软件（V4.5.8, http://www.bioinformatics.org/cd-hit/）进行去冗余，以获得非冗余的初始 gene catalogue

（此处将非冗余的连续基因编码的核酸序列称之为 genes[18]），参数设置[17,18]：-c 0.95, -G 0, -aS 0.9, -g 1, -d 0。

3）使用 Bowtie2（Bowtie2.2.4）将各样品的 Clean Data 比对至初始 gene catalogue，计算得到基因在各样品中比对上的 reads 数目，参数设置[19,7]：--end-to-end, --sensitive, -I 200, -X 400。过滤掉各个样品中 reads 数目<=2[19,24] 的基因，获得最终用于后续分析的 gene catalogue（Unigenes）。

4）从比对上的 reads 数目及基因长度出发，计算得到各基因在各样品中的丰度信息，计算公式如下所示，r 为比对上基因的 reads 数目，L 为基因的长度[15,16,17,25,26,27]：

$$G_k = \frac{r_k}{L_k} \cdot \frac{1}{\sum_{i=1}^{n} \frac{r_i}{L_i}}$$

5）基于 gene catalogue 中各基因在各样品中的丰度信息，进行基本信息统计，core-pan 基因分析，样品间相关性分析，及基因数目韦恩图分析。

4.物种注释

1）使用 DIAMOND[28] 软件（v0.9.9.110, https://github.com/bbuchfink/diamond/ ），将 Unigenes 与从 NCBI 的 NR 数据库(Version 2018-01-02 , https://www.ncbi.nlm.nih.gov/)中抽提出的细菌(Bacteria)、真菌(Fungi)、古菌(Archaea)和病毒(Viruses)序列进行比对，参数设置：blastp，-e 1e-5。

2）对于每一条序列的比对结果，选取 evalue <= 最小 evalue*10[20] 的结果，由于每一条序列可能有多个比对结果，采取 LCA 算法(应用于 MEGAN[29] 软件的系统分类，（https://en.wikipedia.org/wiki/Lowest_common_ancestor）来确定该序列的物种注释信息。

3）从 LCA 注释结果及基因丰度表出发，获得各个样品在各个分类层级（界门纲目科属种）上的丰度信息及基因数目表，对于某个物种在某个样品中的丰度，等于注释为该物种的基因丰度的加和[14,15,16]；对于某个物种在某个样品中的基因数目，等于在注释为该物种的基因中，丰度不为 0 的基因数目。

4）从各个分类层级上的丰度表出发，进行 Krona 分析[30]，相对丰度概况展示，丰度聚类热图展示。并进行 PCA[31](R ade4 package, Version 2.15.3)和 NMDS[32](R vegan package, Version 2.15.3)降维分析；使用 Anosim 分析(R vegan package, Version 2.15.3)检验组间的差异情况；然后使用 Metastats 和 LEfSe 分析寻找组间差异物种，Metastats 分析对各个分类层级做组间的 permutation test，得到 p 值，然后利用 Benjamini and Hochberg False Discovery Rate 方法对于 p 值进行矫正，得到 q 值[33]， LEfSe 分析使用 LEfSe 软件（LDA Score 默认为 3）[34]；最后应用随机森林（RandomForest）[43]（R pROC and randomForest packages,Version 2.15.3）对种水平物种按梯度选取，构建随机森林模型。通过 MeanDecreaseAccuracy 和 MeanDecreaseGin 筛选出重要的物种，之后对每个模型做交叉验证（默认 10-fold）并绘制 ROC 曲线。

5.常用功能数据库注释

1）使用 DIAMOND 软件（v0.9.9.110, https://github.com/bbuchfink/diamond/）将 Unigenes 与功能数据库进行比对，参数设置：blastp， -e 1e-5 [19,8]。功能数据库包括 KEGG[35,36]数据库 (Version 2018-01-01, http://www.kegg.jp/kegg/), eggNOG[37] 数据库 (Version 4.5, http://eggnogdb.embl.de/#/app/home), CAZy[38] 数据库 (Version 201801, http://www.cazy.org/)。对于每一条序列的比对结果，选取 Best Blast Hit 结果进行后续分析[19,8,39]。

2）从比对结果出发，统计不同功能层级的相对丰度（各功能层级的相对丰度等于注释为该功能层级的基因的相对丰度之和[24,14,15]）。

3）从功能注释结果及基因丰度表出发，获得各个样品在各个分类层级上的基因数目表，对于某个功能在某个样品中的基因数目，等于在注释为该功能的基因中，丰度不为 0 的基因数目。

4）从各个分类层级上的丰度表出发，进行注释基因数目统计，相对丰度概况展示，丰度聚类热图展示，PCA 和 NMDS 降维分析，基于功能丰度的 Anosim 组间（内）差异分析，代谢通路比较分析，组间功能差异的 Metastat 和 LEfSe 分析。

6.抗性基因注释

1）使用 CARD 数据库提供的 Resistance Gene Identifier (RGI)软件将 Unigenes 与 CARD 数据库（https://card.mcmaster.ca/）进行比对（RGI 内置 blastp，默认 evalue ≤ 1e-30）[40-42]；

2）根据 RGI 的比对结果，结合 Unigenes 的丰度信息，统计出各 ARO 的相对丰度；

3）从 ARO 的丰度出发，进行丰度柱形图展示，丰度聚类热图展示，丰度分布圈图展示，组间 ARO 差异分析，抗性基因（注释到 ARO 的 unigenes）及抗性机制物种归属分析等（对部分名称较长的 ARO，用其前三个单词与下划线缩写的形式展示）。

参考文献

[1] Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & biology, 5(10), R245-R249.

[2] Chen, K., & Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS computational biology, 1(2), e24.

[3] Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. Nature reviews genetics, 6(11), 805-814.

[4] Tringe, S. G., Von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., ... & Rubin, E. M. (2005). Comparative metagenomics of microbial communities. Science, 308(5721), 554-557.

[5] Law J, Jovel J, Patterson J, Ford G, O'keefe S, Wang W, et al. (2013) Identification of Hepatotropic Viruses from Plasma Using Deep Sequencing: A Next Generation Diagnostic Tool. PLoS ONE 8(4): e60595.

[6] Luo et al.: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 2012 1:18.

[7] Qin N, Yang F, Li A, et al. Alterations of the human gut microbiome in liver cirrhosis[J]. Nature, 2014.

[8] Feng Q, Liang S, Jia H, et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence[J]. Nature communications, 2015, 6.

[9] Scher J U, Sczesnak A, Longman R S, et al. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis[J]. Elife, 2013, 2: e01202.

[10] Brum J R, Ignacio-Espinoza J C, Roux S, et al. Patterns and ecological drivers of ocean viral communities[J]. Science, 2015, 348(6237): 1261498.

[11] Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph[J]. Bioinformatics,2015, 31:1674
–1676.

[12] Mende D R, Waller A S, Sunagawa S, et al. Assessment of metagenomic assembly using simulated next generation sequencing data[J]. PloS one, 2012, 7(2): e31386.

[13] Nielsen H B, Almeida M, Juncker A S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes[J]. Nature biotechnology, 2014, 32(8): 822-828.

[14] Karlsson FH, Tremaroli V, Nookaew I, Bergstrom G, Behre CJ, Fagerberg B, Nielsen J, Backhed F: Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature 2013, 498(7452):99-103.

[15] Karlsson F H, Fåk F, Nookaew I, et al. Symptomatic atherosclerosis is associated with an altered gut metagenome[J]. Nature communications, 2012, 3: 1245.

[16] Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing[J]. nature, 2010, 464(7285): 59-65.

[17] Zeller G, Tap J, Voigt A Y, et al. Potential of fecal microbiota for early‐stage detection of colorectal cancer[J]. Molecular systems biology, 2014, 10(11): 766.

[18] Sunagawa S, Coelho L P, Chaffron S, et al. Structure and function of the global ocean microbiome[J]. Science, 2015, 348(6237): 1261359.

[19] Li J, Jia H, Cai X, et al. An integrated catalog of reference genes in the human gut microbiome[J]. Nature biotechnology, 2014, 32(8): 834-841.

[20] Oh J, Byrd A L, Deming C, et al. Biogeography and individuality shape function in the human skin metagenome[J]. Nature, 2014, 514(7520): 59-64.

[21] Zhu, Wenhan, Alexandre Lomsadze, and Mark Borodovsky. "Ab initio gene identification in metagenomic sequences." Nucleic acids research 38.12 (2010): e132-e132

[22] Li W, Godzik A: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006, 22(13):1658-1659.

[23] Fu L, Niu B, Zhu Z, Wu S, Li W: CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012, 28(23):3150-3152.

[24] Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes[J]. Nature, 2012, 490(7418): 55-60.
[25] Villar E, Farrant G K, Follows M, et al. Environmental characteristics of Agulhas rings affect interocean plankton transport[J]. Science, 2015, 348(6237): 1261447.
[26] Cotillard A, Kennedy S P, Kong L C, et al. Dietary intervention impact on gut microbial gene richness[J]. Nature, 2013, 500(7464): 585-588.
[27] Le Chatelier E, Nielsen T, Qin J, et al. Richness of human gut microbiome correlates with metabolic markers[J]. Nature, 2013, 500(7464): 541-546.
[28] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015;12:59-60.
[29] Huson, Daniel H., et al. "Integrative analysis of environmental sequences using MEGAN4." Genome research 21.9 (2011): 1552-1560.
[30] Ondov B D, Bergman N H, Phillippy A M. Interactive metagenomic visualization in a Web browser[J]. BMC bioinformatics, 2011, 12(1): 385.
[31] Avershina, Ekaterina, Trine Frisli, and Knut Rudi. De novo Semi-alignment of 16S rRNA Gene Sequences for Deep Phylogenetic Characterization of Next Generation Sequencing Data. Microbes and Environments 28.2 (2013): 211-216.
[32] Rivas M N, Burton O T, Wise P, et al. A microbiota signature associated with experimental food allergy promotes allergic sensitization and anaphylaxis[J]. Journal of Allergy & Clinical Immunology, 2013, 131(1):201-212.
[33] White J R, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples[J]. PLoS Comput Biol, 2009, 5(4): e1000352.
[34] Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation[J]. Genome Biology, 2011, 12(6):1-18.
[35] Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. (2006). From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34(Database issue): D354–7.
[36] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M.; Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199–D205 (2014).
[37] Powell S, Forslund K, Szklarczyk D, et al. eggNOG v4. 0: nested orthology inference across 3686 organisms[J]. Nucleic acids research, 2013: gkt1253.
[38] Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) .The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37:D233-238.
[39] Bäckhed F, Roswall J, Peng Y, et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life[J]. Cell host & microbe, 2015, 17(5): 690-703.
[40] Martínez J L, Coque T M, Baquero F. What is a resistance gene? Ranking risk in resistomes[J]. Nature Reviews Microbiology, 2014, 13(2):116-23.
[41] Jia B, Raphenya A R, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database[J]. Nucleic Acids Research, 2017, 45(D1):D566.
[42] Mcarthur A G, Waglechner N, Nizam F, et al. The Comprehensive Antibiotic Resistance Database[J]. Antimicrobial Agents & Chemotherapy, 2013, 57(7):3348.
[43] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1):5-32.