中文版 methods

1 实验流程

1.1 样品检测

详见样本检测报告。

1.2 文库构建和上机测序

取样本的 1 μg 基因组 DNA，用 Covaris 超声波破碎仪随机打断成长度约为 350 bp 的片段后进行文库的构建，经末端修复、加 A 尾、加测序接头、纯化、PCR 扩增等步骤完成整个文库制备。文库构建完成后，先使用 AATI 检测文库片段的完整性及插入片段大小，符合预期后，使用 Q-PCR 方法对文库的有效浓度进行准确定量(文库有效浓度>3 nM)，以保证文库质量。库检合格后，把不同文库按照有效浓度及目标下机数据量的需求 pooling 后进行 PE150 测序。

2 生物信息分析

2.1 测序结果预处理

使用 fastp (https://github.com/OpenGene/fastp)对 NovaSeq 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)当任一测序 read 中含有接头序列，去除此 paired reads; b)当任一测序 read 中含有的低质量(Q<=5)碱基数超过该条 read 碱基数的 50%时，去除此 paired reads; c)当任一测序 read 中 N 含量超过该 read 碱基数的 10%时，去除此 paired reads。

如果样品存在宿主污染，需与宿主序列进行比对，过滤掉可能来源于宿主的 reads，默认采用 Bowtie2 软件(http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)，参数设置--end-to-end, --sensitive, - I 200, -X 400 (Karlsson FH et al., 2012; Karlsson FH et al., 2013; Scher JU et al., 2013)。

2.2 Metagenome 组装

使用 MEGAHIT 软件对 clean data 进行组装分析，组装参数设置:-- presets meta-large (--end-to-end, --sensitive, -I 200, -X 400 (Karlsson FH et al., 2013; Nielsen HB et al., 2014)，然后将组装得到的 scaffolds 从 N 连接处打断，得到不含 N 的 scaftigs(Qin J et al., 2010; Li D et al., 2015)。

2.3 基因预测及丰度分析

使用 MetaGeneMark (http://topaz.gatech.edu/GeneMark/)对各样品的 scaftigs (>=500bp)进行 ORF 预测(Karlsson FH et al., 2012; Mende DR et al.,2012; Li J et al., 2014; Oh J et al., 2014; Qin N et al., 2014)，并过滤掉预测结果中长度小于 100 nt 的信息(Qin J et al., 2010; Zhu W et al., 2010; Nielsen HB et al., 2014; Zeller G et al., 2014; Sunagawa S et al., 2015)，均采用默认参数。对 ORF 预测结果，采用 CD-HIT 软件(http://www.bioinformatics.org/cd-hit/)进行去冗余(Li W et al., 2006; Fu L et al., 2012)，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 (Zeller G et al., 2014)，参数设置: -c 0.95,-G 0,-aS 0.9,-g 1,-d 0 (Li J et al., 2014; Qin N et al., 2014)。使用 Bowtie2 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，比对参数：--end-to-end, --sensitive, -I 200,- X 400 (Qin J et al., 2010; Li J et al., 2014)。过滤掉各个样品中 reads 数目<=2 的基因(Zeller G et al., 2014)，获得最终用于后续分析的 gene catalogue (unigenes)。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，如计算公式所示，r 为比对上基因的 reads 数目，L 为基因的长度(Cotillard A et al., 2013; Buchfink B et al., 2015; Villar E et al., 2015)。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。

2.4 物种注释

使用 DIAMOND 软件(https://github.com/bbuchfink/diamond/) (Buchfink B et al., 2015)，将 unigenes 与 Micro_NR 进行比对，参数设置：blastp, -e 1e-5(Karlsson FH et al., 2013)。Micro_NR 是从 NCBI 的 NR 数据库(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。

对于每一条序列的比对结果，选取 evalue <=最小 evalue*10 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 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(Huson DH et al., 2011)。

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(Karlsson FH et al., 2012; Li J et al., 2014; Feng Q et al., 2015); 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。

从各个分类层级上的丰度表出发，进行 Krona 分析(Ondov BD et al., 2011)，相对丰度概况展示，丰度聚类热图展示。并进行 PCA (R ade4 package)(Rao C R et al., 1964)，PCoA(R ade4 package)和 NMDS (R vegan package)降维分析(Legendre P, 1998)；使用 Anosim 分析(R vegan package)检验组间的差异情况；然后使用 MetaGenomeSeq 和 LEfSe 分析寻找组间差异物种，MetaGenomeSeq 分析对各个分类层级做组间的假设检验得到 p 值与 Q 值，LEfSe 分析使用 LEfSe 软件(LDA Score 默认为 4) (Segata N et al., 2011)；最后应用随机森林(RandomForest) (R pROC and randomForest packages,Version 2.15.3) (Breiman L,2001)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。

2.5 常用功能数据库注释

使用 DIAMOND 软件(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 (Li J et al., 2014; Feng Q et al., 2015)。功能数据库包括 KEGG 数据库 (http://www.kegg.jp/kegg/) (Kanehisa M et al., 2006; Kanehisa M et al., 2017)，eggNOG 数据库(http://eggnogdb.embl.de/#/app/home) (Jaime Huerta-Cepas et al., 2016)，CAZy 数据库(http://www.cazy.org/) (Cantarel BL et al., 2009)，VFDB 数据库(http://www.mgc.ac.cn/VFs/main.htm)，PHI 数据库(http://www.phi-base.org/index.jsp)对于每一条序列的比对结果，选取 score 最高的比对结果进行后续分析(Qin J et al., 2012; Li J et al., 2014; Qin N et al., 2014; Bä ckhed F et al., 2015)。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(Karlsson FH et al., 2012; Li J et al., 2014)。

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。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 MetaGenomeSeq 和 LEfSe 分析。

2.6 抗性基因注释

使用CARD 数据库(Martí nez JL et al., 2015)提供的Resistance Gene Identifier (RGI)软件(Jia B et al., 2017)将 Unigenes 与 CARD 数据库(https://card.mcmaster.ca/)进行比对(RGI 内置 blastp，默认 evalue < 1e-30) (McArthur AG et al., 2013)；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;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)。

可移动遗传元件 mobile genetic elements (MGEs)，将 unigenes 分别与插入序列(isfinder)、整合子(integrall)和质粒(plasmid)数据库进行比对，得到丰度信息。将注释得到的丰度信息进一步可视化，进行丰度柱状图与相对丰度热图结果展示。

3 参考文献

Bä ckhed F, Roswall J, Peng Y, et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe. 2015;17(6):852. doi:10.1016/j.chom.2015.05.012

Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59-60. doi:10.1038/nmeth.3176

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. 2009;37(Database issue):D233-D238. doi:10.1093/nar/gkn663

Chen K, Pachter L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol. 2005;1(2):106-112. doi:10.1371/journal.pcbi.0010024

Cotillard A, Kennedy SP, Kong LC, et al. Dietary intervention impact on gut microbial gene richness [published correction appears in Nature. 2013 Oct 24;502(7472)580]. Nature. 2013;500(7464):585-588. doi:10.1038/nature12480

Feng Q, Liang S, Jia H, et al. Gut microbiome development along the colorectal adenoma- carcinoma sequence. Nat Commun. 2015;6:6528. Published 2015 Mar 11. doi:10.1038/ncomms7528

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. doi:10.1093/bioinformatics/bts565

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998;5(10):R245-R249. doi:10.1016/s1074-5521(98)90108-9

Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 2011;21(9):1552-1560. doi:10.1101/gr.120618.111

Jaime Huerta-Cepas, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, Daniel R. Mende, Shinichi Sunagawa, Michael Kuhn, Lars Juhl Jensen, Christian von Mering, Peer Bork; eggNOG 4.5: a hierarchical orthology

framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences, Nucleic Acids Research, Volume 44, Issue D1, 4 January 2016, Pages D286–D293

Jia B, Raphenya AR, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 2017;45(D1):D566-D573. doi:10.1093/nar/gkw1004

Karlsson FH, Få k F, Nookaew I, et al. Symptomatic atherosclerosis is associated with an altered gut metagenome. Nat Commun. 2012;3:1245. doi:10.1038/ncomms2266

Karlsson FH, Tremaroli V, Nookaew I, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. 2013;498(7452):99-103. doi:10.1038/nature12198

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353-D361. doi:10.1093/nar/gkw1092

Kanehisa M, Goto S, Hattori M, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006;34(Database issue):D354-D357. doi:10.1093/nar/gkj102

Klipper-Aurbach Y, Wasserman M, Braunspiegel-Weintrob N, et al. Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus. Med Hypotheses. 1995;45(5):486-

490. doi:10.1016/0306-9877(95)90228-7

Le Chatelier E, Nielsen T, Qin J, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013;500(7464):541-546. doi:10.1038/nature12506 Legendre P, Legendre L. Numerical ecology, 2nd edition[J].1998.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674-1676. doi:10.1093/bioinformatics/btv033

Li J, Jia H, Cai X, et al. An integrated catalog of reference genes in the human gut microbiome.

Nat Biotechnol. 2014;32(8):834-841. doi:10.1038/nbt.2942

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. doi:10.1093/bioinformatics/btl158

Martí nez JL, Coque TM, Baquero F. What is a resistance gene? Ranking risk in resistomes. Nat Rev Microbiol. 2015;13(2):116-123. doi:10.1038/nrmicro3399

McArthur AG, Waglechner N, Nizam F, et al. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother. 2013;57(7):3348-3357. doi:10.1128/AAC.00419-13

Mende DR, Waller AS, Sunagawa S, et al. Assessment of metagenomic assembly using simulated next generation sequencing data [published correction appears in PLoS One. 2014;9(11):e114063]. PLoS One. 2012;7(2):e31386. doi:10.1371/journal.pone.0031386

Nielsen HB, Almeida M, Juncker AS, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32(8):822-828. doi:10.1038/nbt.2939

Rao C R. The Use and Interpretation of Principal Component Analysis in Applied Research[J]. Sankhyā : The Indian Journal of Statistics, Series A (1961-2002), 1964, 26(4):329-358.

Oh J, Byrd AL, Deming C, et al. Biogeography and individuality shape function in the human skin metagenome. Nature. 2014;514(7520):59-64. doi:10.1038/nature13786

Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;12:385. Published 2011 Sep 30. doi:10.1186/1471-

2105-12-385

Raes J, Foerstner KU, Bork P. Get the most out of your metagenome: computational analysis of environmental sequence data. Curr Opin Microbiol. 2007;10(5):490-498. doi:10.1016/j.mib.2007.09.001

Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic

sequencing. Nature. 2010;464(7285):59-65. doi:10.1038/nature08821

Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490(7418):55-60. doi:10.1038/nature11450

Qin N, Yang F, Li A, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature. 2014;513(7516):59-64. doi:10.1038/nature13568

Scher JU, Sczesnak A, Longman RS, et al. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. Elife. 2013;2:e01202. Published 2013 Nov 5. doi:10.7554/eLife.01202

Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12(6):R60. Published 2011 Jun 24. doi:10.1186/gb-2011-12-6-r60

Sunagawa S, Coelho LP, Chaffron S, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science. 2015;348(6237):1261359.doi:10.1126/science.1261359

Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet. 2005;6(11):805-814. doi:10.1038/nrg1709

Tringe SG, von Mering C, Kobayashi A, et al. Comparative metagenomics of microbial communities. Science. 2005;308(5721):554-557. doi:10.1126/science.1107851

Villar E, Farrant GK, Follows M, et al. Ocean plankton. Environmental characteristics of Agulhas rings affect interocean plankton transport. Science. 2015;348(6237):1261447. doi:10.1126/science.1261447

Zeller G, Tap J, Voigt AY, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014;10(11):766. Published 2014 Nov 28. doi:10.15252/msb.20145645

Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010;38(12):e132. doi:10.1093/nar/gkq275