methods

The method is only used as a reference for publication purpose. Customers are responsible for the related risks of duplicate checking.

1.1 Sample Quality Control

Please refer to Novogene's QC report for methods of sample quality control.

1.2 Library Construction, Quality Control and Sequencing

Sampling 1 μg of genomic DNA, the sample is randomly fragmented into segments of approximately 350 bp using a Covaris ultrasonic disruptor to construct the library. The entire library preparation is completed through steps including end repair, addition of A-tails, ligation of sequencing adapters, purification, and PCR amplification. After library construction, the integrity of the library fragments and the size of the inserted fragments are assessed using AATI analysis. If the insert size meets expectations, the accurate concentration of the effective library is quantified using Q-PCR (effective library concentration > 3 nM) to ensure the library quality. After the library passes the quality check, different libraries are pooled according to their effective concentrations and target data output requirements, and then subjected to PE150 sequencing.

2 Bioinformatics Analysis Pipeline

2.1 Preprocessing of sequencing results

Fastp (https://github.com/OpenGene/fastp) is used for preprocessing raw data from the Illumina sequencing platform to obtain clean data for subsequent analysis. We will discard the paired reads in the following situation: when either one read contains adapter contamination; when either one read contains more than 10 percent uncertain nucleotides; when either one read contains more than 50 percent low quality nucleotides (base quality less than 5).

Considering the possibility of host contamination in samples, clean data needs to be blasted to the host database to filter out reads that may come from host origin. Bowtie2 software (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) is used by default, with the following parameter settings: --end-to-end, --sensitive, -I 200, and -X 400 (Karlsson FH et al., 2012; Karlsson FH et al., 2013; Scher JU et al., 2013).

2.2 Assembly of Metagenome

MEGAHIT software is used for assembly analysis of clean data, with assembly parameter settings: --presets meta-large (--end-to-end, --sensitive, -I 200, -X 400) (Karlsson FH et al., 2013; Nielsen HB et al., 2014), and scaftigs without N is obtained by breaking the resulted scaffolds from the N junction (Qin J et al., 2010; Li D et al., 2015).

2.3 Gene prediction and abundance analysis

With the default parameters, MetaGeneMark (http://topaz.gatech.edu/GeneMark/) is used to perform ORF prediction for scaftigs (>= 500 bp) of each sample (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), and the information with a length less than 100 nt in the prediction results is filtered out (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).

For the ORF prediction results, CD-HIT software (http://www.bioinformatics.org/cd-hit/) is used to eliminate redundancy (Li W et al., 2006; Fu L et al., 2012) and obtain the non- redundant initial gene catalogue (the nucleic acid sequences encoded by successive non- redundant genes are called genes) (Zeller G et al., 2014), with parameter settings: -c 0.95,-G 0,-aS 0.9,-g 1,-d 0 (Li J et al., 2014; Qin N et al., 2014).

Clean data of each sample is aligned to the initial gene catalogue by using Bowtie2 to calculate the number of reads of the genes on each sample alignment, with parameter settings: --end-to-end, --sensitive, -I 200, -x 400 (Qin J et al., 2010; Li J et al., 2014). Genes with reads <= 2 in each sample are filtered out to finally determine the gene catalogue (Unigenes) for subsequent analysis (Zeller G et al., 2014).

Based on the number of reads aligned and the length of gene, the abundance of each gene in each sample is calculated by the following formula, in which r is the number of gene reads on alignment, and L is the length of gene (Cotillard A et al., 2013; Buchfink B et al., 2015; Villar E et al., 2015):

Based on the abundance of each gene in the gene catalogue in each sample, basic information statistics, core-pan gene analysis, correlation analysis between samples, and Venn diagram analysis of gene number are performed.

2.4 Species annotation

DIAMOND software (https://github.com/bbuchfink/diamond/) (Buchfink B et al., 2015) is used for alignment of unigenes sequences with Micro_NR database, which includes sequences from bacteria, fungi, archaea, and viruses extracted from NCBI's NR database (https://www.ncbi.nlm.nih.gov/). The alignment is performed useing the blastp algorithm with a parameter setting of 1e-5 (Karlsson FH et al., 2013).

From the alignment results of each sequence, the one with evalue <= min. evalue *10 is selected. Since each sequence may have multiple alignment results, LCA algorithm (applied to systematic taxonomy of MEGAN software (https://en.wikipedia.org/wiki/Lowest_common_ancestor) is adopted to determine the species annotation information of the sequence (Huson DH et al., 2011).

Out of the results of LCA annotation and gene abundance table, the abundance of each sample at each taxonomy (kingdom, phylum, class, order, family, genus, or species) and the corresponding gene abundance tables are acquired. The abundance of a species in a sample

is equal to the sum of the abundance of those genes annotated as that species (Karlsson FH et al., 2012; Li J et al., 2014; Feng Q et al., 2015). The number of genes of a species in a sample is equal to the number of genes whose abundance is non-zero among the genes annotated as that species.

On the basis of the abundance tables at each taxonomy level, Krona analysis (Ondov BD et al., 2011), relative abundance overview, and abundance clustering heatmap are performed, combined with PCA (R ade4 package) (Rao C R et al., 1964), PCoA (R ade4 package) ,and NMDS (R vegan package) analysis of dimension reduction (Legendre P, 1998). Anosim analysis (R vegan package) is used to test the differences between groups. MetaGenomeSeq and LEfSe analysis are used to search for species differences between groups. MetaGenomeSeq analysis is used to perform permutation test between groups on each taxonomy level and get a p-value. LEfSe software is used for LEfSe analysis (LDA Score is 4 by default) (Segata N et al., 2011). Finally, Random forest analysis (R pROC and randomForest packages) (Breiman L, 2001) is applied to select the species at species level by gradient and build a RandomForest model. Important species are screened out by MeanDecreaseAccuracy and MeanDecreaseGin, and then cross-validation (default 10-fold) is performed for each model and ROC curves are drawn.

2.5 Annotations of common functional database

DIAMOND software (https://github.com/bbuchfink/diamond/) is used to align unigenes with those in the functional database, with parameter settings: blastp, -e 1e-5 (Li J et al., 2014; Feng Q et al., 2015). Functional databases include KEGG database (http://www.kegg.jp/kegg/) (Kanehisa M et al., 2006; Kanehisa M et al., 2017), eggNOG database (http://eggnogdb.embl.de/#/app/home) (Jaime Huerta-Cepas et al., 2016), CAZy database (http://www.cazy.org/)(Cantarel BL et al., 2009), VFDB database (http://www.mgc.ac.cn/VFs/main.htm) and PHI database (http://www.phi-base.org/index.jsp). From the alignment results of each sequence, the best blast hit results are selected for subsequent analysis (Qin J et al., 2012; Li J et al., 2014; Qin N et al., 2014; Bä ckhed F et al., 2015).

According to the alignment results, the relative abundance at different functional levels is calculated (the relative abundance at each functional level is equal to the sum of the relative abundance of genes annotated as that functional level) (Karlsson FH et al., 2012; Li J et al., 2014).

The gene number table of each sample at each taxonomy level is derived from the result of functional annotation and gene abundance table. The number of genes with a certain function in a sample is equal to the number of genes whose abundance is non-zero among the genes annotated with this function.

Based on the abundance table at each taxonomy level, annotated genes statistics, relative abundance overview, and abundance clustering heat map are carried out, combined with PCA and NMDS analysis of dimension reduction, Anosim analysis of inter-/intra-group

differences on the basis of functional abundance, metabolic pathway comparative analysis, as well as MetaGenomeSeq and LEfSe analysis on inter-group functional difference.

2.6 Annotations of resistance gene

Unigenes are aligned to the CARD database (https://card.mcmaster.ca/) (Martí nez JL et al., 2015) using the Resistance Gene Identifier (RGI) software (Jia B et al., 2017) provided by the CARD database (RGI built-in blastp, default evalue < 1e-30) (McArthur AG et al., 2013).

According to the RGI alignment result and unigenes abundance information, the relative abundance of each ARO is calculated.

Based on the abundance of ARO, the abundance histogram, abundance clustering heat map, abundance distribution circle map, ARO difference analysis between groups, resistance genes (unigenes annotated as ARO) and species attribution analysis of resistance mechanism are carried out (some AROs with long names are abbreviated as the first three words plus underlines.)

Mobile genetic elements (MGEs), unigenes were compared with the insertion sequence (isfinder), integrall and plasmid databases, respectively, to obtain abundance information. The annotated abundance information was further visualized, and the results of abundance histogram and relative abundance heatmap were displayed.

3 Reference

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