Information Analysis of Metagenomic Project

Metagenomic analysis can more truly reflect the microbial composition and interaction in the sample, and study the metabolic pathway and gene function at the molecular level [1-4].

1.Sequencing results pretreatment

1) Preprocessing the Raw Data obtained from the Illumina HiSeq sequencing platform using Readfq （V8, https://github.com/cjfields/readfq ）was conducted to acquire the Clean Data for subsequent analysis . The specific processing steps are as follows: a) remove the reads which contain low quality bases (default quality threshold value ≤ 38)above a certain portion (default length of 40 bp); b) remove the reads in which the N base has reached a certain percentage (default length of 10 bp);c) remove reads which shared the overlap above a certain portion with Adapter (default length of 15 bp).

2) Considering the possibility of host pollution may exist in samples, Clean Data need to be blast to the host database which default using Bowtie2.2.4 software (Bowtie2.2.4, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) to filter the reads that are of host origin, the parameters[14,15] are as follows : --end-to-end, --sensitive, -I 200, -X 400.

2.Metagenome Assembly

1) Single sample assembly :

To the samples taken from non-complex environment, such as intestine, faeces and so on, the Clean Data is assembled and analyst [6] by SOAPdenovo software (V2.04, http://soap.genomics.org.cn/soapdenovo.html),

the parameters[7-10] are as follows: -d 1, -M 3, -R, -u, -F, -K 55; To the samples taken from complex environment, such as water, soil and so on, MEGAHIT software (v1.0.4-beta) could be used to assemble the Clean Data and the parameters[11] are -presets meta-large（-- min-count 2 --k-min 27 --k-max 87 --k-step 10）; then interrupted the assembled Scaftigs from N connection and leave the Scaftigs without N [7,12,13] . All samples' Clean Data are compared to each Scaffolds respectively by Bowtie2.2.4 software to acquire the PE reads not used and the parameters [7] are: --end-to-end, --sensitive, -I 200, -X 400.

2) Mixed assembly:

all the reads not used in the forward step of all samples are combined and then use the software of SOAPdenovo (V2.04) / MEGAHIT (v1.0.4-beta) for mixed assembly with the parameters same as single assembly; Break the mixed assembled Scaffolds from N connection and obtained the Scaftigs. Filter the fragment shorter than 500 bp in all of Scaftigs for statistical analysis both generated from single or mixed assembly.

3.Gene prediction and abundance analysis

1）The Scaftigs (≥ 500 bp) assembled from both single and mixed are all predicted the ORF by MetaGeneMark (V2.10, http://topaz.gatech.edu/GeneMark/) software, and filtered the length information shorter than 100 nt [7,13,16,17,19]from the predicted result with default parameters.

2) For ORF predicted, CD-HIT[22,23] software (V4.5.8, http://www.bioinformatics.org/cd-hit) is adopted to redundancy and obtain the unique initial gene catalogue (the genes here refers to the nucleotide sequences coded by unique and continuous genes[18]), the parameters option [17,18] are -c 0.95, -G 0, -aS 0.9, -g 1, -d 0.

3) The Clean Data of each sample is mapped to initial gene catalogue using Bowtie2.2.4 and get the number of reads to which genes mapped in each sample with the parameter setting [7,19] are --end-to-end, --sensitive, -I 200, -X 400. Filter the gene which the number of reads ≤ 2 [19,24] in each sample and obtain the gene catalogue (Unigenes) eventually used for subsequently analysis.

4)Based on the number of mapped reads and the length of gene, statistic the abundance information of each gene in each sample. The format is as follow, r represents the number of reads mapped to the genes and L represents gene's length[15-17, 25-27].

5)The basic information statistic, core-pan gene analysis, correlation analysis of samples and venn figure analysis of number of genes are all based on the abundance of each gene in each sample in gene catalogue.

4.Taxonomy prediction

1) DIAMOND[28] software (V0.9.9, https://github.com/bbuchfink/diamond/) is used to blast the Unigenes to the sequences of Bacteria, Fungi, Archaea and Viruses which are all extracted from the NR database (Version: 2018-01-02, https://www.ncbi.nlm.nih.gov/) of NCBI with the parameter setting are blastp，-e 1e-5.

2)For the finally aligned results of each sequence, as each sequence may have multiple aligned results, choose the result of which the e value ≤ the smallest e value * 10 [20] to take the LCA algorithm which is applied to system classification of MEGAN [29] software to make sure the species annotation information of sequences.

3) The table containing the number of genes and the abundance information of each sample in each taxonomy hierarchy (kingdom, phylum, class, order, family, genus, species) are obtained based on the LCA annotation result and the gene abundance table. The abundance of a specie in one sample equal the sum of the gene abundance annotated for the specie; the gene number of a specie in a sample equal the number of genes whose abundance are nonzero.

4) Krona analysis, the exhibition of generation situation of relative abundance, the exhibition of abundance cluster heat map, PCA [31] (R ade4 package, Version 2.15.3) and NMDS[32](R vegan package, Version 2.15.3) decrease-dimension analysis are based on the abundance table of each taxonomic hierarchy. The difference between groups is tested by Anosim analysis (R vegan package, Version 2.15.3). Metastats and LEfSe analysis are used to look for the different species between groups. Permutation test between groups is used in Metastats analysis for each taxonomy and get the P value, then use Benjamini and Hochberg False Discovery Rate to correct P value and acquire q value[33]. LEfSe analysis is conducted by LEfSe software (the default LDA score is 3) [34] ; Finally, random forest (RandoForest) [43] (R pROC and randomForest packages, Version 2.15.3) was used to construct a random forest model. Screen out important species by MeanDecreaseAccuracy and MeanDecreaseGin, then cross-validate each model (default 10 times) and plot the ROC curve.

5.Common functional database annotations

1) Adopt DIAMOND software (V0.9.9) to blast Unigenes to functional database with the parameter setting of blastp, -e 1e-5 [19,8]. Functional database exclude KEGG [35,36] database (Version 2018-01-01, http://www.kegg.jp/kegg/), eggNOG [37] database (Version 4.5, http://eggnogdb.embl.de/#/app/home), CAZy [38] database (Version 201801, http://www.cazy.org/). For each sequence's blast result, the best Blast Hit is used for subsequent analysis [8,19,39].

2) Statistic of the relative abundance of different functional hierarchy, the relative abundance of each functional hierarchy equal the sum of relative abundance annotated to that functional level.

3) Based on the function annotation result and gene abundance table, the gene number table of each sample in each taxonomy hierarchy is obtained. The gene number of a function in a sample equal the gene number that annotated to this function and the abundance is nonzero.

4) Based on the abundance table of each taxonomy hierarchy, not only the counting of annotated gene numbers, the exhibition of the general relative abundance situation, the exhibition of abundance cluster heat map and the decrease-dimension analysis of PCA and NMDS are conducted, but also the Anosim analysis of the difference between groups (inside) based on functional abundance, comparative analysis of metabolic pathways, the Metatat and LEfSe analysis of functional difference between groups are performed.

6.Resistance gene annotation

1）Use Resistance Gene Identifier (RGI) software to align the Unigenes to CARD database(https://card.mcmaster.ca/) [40-42]with the parameter settting are blastp, evalue ≤ 1e-30.

2）Based on the aligned result, count the relative abundance of ARO.

3）Based on the abundance of ARO, the abundance bar charts, the abundance cluster heatmap and the resistance genes’ number difference between groups are displayed. In the same way, The resistance genes’ abundance distribution in each samples, the species attribution analysis of resistance genes and the resistance mechanism of resistance genes analysis are also conducted.

References

[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.

[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.