Methods (Metagenome Analysis)

Experimental procedure

1. DNA extraction

Genomic DNA was extracted using HiPure Bacterial DNA Kits (Magen, Guangzhou, China) according to the manufacturer's instructions. The DNA quality was detected using Qubit (Thermo Fisher Scientific, Waltham, MA) and Nanodrop (Thermo Fisher Scientific, Waltham, MA) accordingly.

3. Illumina sequencing

Qualified genomic DNA was firstly fragmented by sonication to a size of 350bp, and then end-repaired, A-tailed, and adaptor ligated using NEBNext® MLtra™ DNA Library Prep Kit for Illumina (NEB, USA) according to the preparation protocol. DNA fragments with length of 300-400 bp were enriched by PCR. At last, PCR products were purified using AMPure XP system (Beckman Coulter, Brea, CA, USA) and libraries were analysed for size distribution by 2100 Bioanalyzer (Agilent, Santa Clara, CA) and quantified using real-time PCR. Genome sequencing was performed on the Illumina Novaseq 6000 sequencer using the pair-end technology (PE 150).

Bioinformatic analysis

1. Quality control

Raw data from Illumina platform were filtered using FASTP (version 0.18.0) [1] by the folowing standards,1) removing reads with ≥10 % unidentified nucleotides (N); 2) removing reads with ≥ 50 % bases having phred quality scores ≤ 20; 3) removing reads aligned to the barcode adapter. After filtering, resulted clean reads were used for genome assembly.

2. Assembly, gene prediction and gene catalogue

Clean reads of each sample were assembled individually using MEGAHIT(version 1.1.2) [2] stepping over a k-mer range of 21 to 99 to generate sample-derived assembly. Genes were predicted based on the final assembly contigs (>500bp) using MetaGeneMark (version 3.38) [3]. The predicted genes ≥ 300 bp in length from all samples were pooled and combined based on ≥ 95% identity and 90% reads coverage using CD-HIT (version 4.6) [4] in order to reduce the number of redundant genes for the downstream assembly step. The reads was re-align to predicted gene using Bowtie (version 2.2.5) [5] to count reads numbers. Final gene catalogue was obtained from non-redundant genes with gene reads count >2.

3. Function annotations

We utilized several complementary approaches to annotate the assembled sequences. The unigenes were annotated using DIAMOND (version 0.9.24) [6] by aligning with the deposited ones in diverse protein databases including National Center for Biotechnology Information (NCBI) non-redundant protein (Nr) database, Kyoto Encyclopedia of Genes and Genomes (KEGG), evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG). Additional annotation was carried out basing on the following databases: Carbohydrate-Active enZYmes (CAZy), Pathogen Host Interactions (PHI), Virulence Factors of Pathogenic Bacteria (VFDB), CARD (Comprehensive Antibiotic Resistance Database).

4. Taxonomic profiling

Clean reads were used to generate taxonomic profile using Kaiju (version 1.6.3) [7].

5. Comparative analysis

Venn graph was plotted using VennDiagram package [8] in R project. Bray-curtis distance matrix based on gene/taxon/function gene abundance was generated by R Vegan package [9]. Multivariate statistical techniques including PCA (principal component analysis), PCoA (principal coordinates analysis) and NMDS (non-metric multi-dimensional scaling) of Bray- curtis distances were calculated using R vegan package [9]and plotted using R ggplot2 package [10]. Violin plot and box plot were graphed using R ggplot2 package [10]. Statistic analysis of Welch's t-test, ANOVA (analysis of variance), Adonis (also called Permanova) and Anosim test was calculated using R project Vegan package [9]. Heatmap graph were plotted using R Pheatmap package [11]. Ternary plot of species was plotted using R ggtern package [12]. Circular layout representations of species or functional gene abundance were graphed using circos (version 0.69-3) [13]. Biomarker features in each group were screened by Metastats (version 20090414) [14] and LEfSe software (version 1.0) [15].

References

[1] Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor[J]. bioRxiv, 2018: 274100.

[2] Li D, Liu C M, Luo R, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph[J]. Bioinformatics, 2015, 31(10): 1674-1676.

[3] Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences[J]. Nucleic acids research, 2010, 38(12): e132-e132.

[4] Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation

sequencing data[J]. Bioinformatics, 2012, 28(23): 3150-3152.

[5] Langmead B, Salzberg S L. Fast gapped-read alignment with Bowtie 2[J]. Nature methods, 2012, 9(4): 357.

[6] Buchfink B, Xie C, Huson D H. Fast and sensitive protein alignment using DIAMOND[J]. Nature methods, 2015, 12(1): 59.

[7] Menzel P, Ng K L, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju[J]. Nature communications, 2016, 7: 11257.

[8] Chen H, Boutros P C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R[J]. BMC bioinformatics, 2011, 12(1): 35.

[9] Oksanen J, Blanchet F G, Kindt R, et al. Vegan: community ecology package. R package version 1.17-4[J]. http://cran. r-project. org>. Acesso em, 2010, 23: 2010.

[10] Wickham H, Chang W. ggplot2: An implementation of the Grammar of Graphics[J]. R package version 0.7, URL: http://CRAN. R-project. org/package= ggplot2, 2008, 3.

[11] Kolde R, Kolde M R. Package 'pheatmap'[J]. R Package, 2015, 1(7).

[12] Hamilton N E, Ferry M. ggtern: Ternary diagrams using ggplot2[J]. Journal of Statistical Software, 2018, 87(1): 1-17.

[13] Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics[J]. Genome research, 2009, 19(9): 1639-1645.

[14] White, James Robert, Niranjan Nagarajan, and Mihai Pop. "Statistical methods for detecting differentially abundant features in clinical metagenomic samples." PLoS Comput Biol 5.4 (2009): e1000352.

[15] Segata, Nicola, et al. "Metagenomic biomarker discovery and explanation." Genome biology 12.6 (2011): 1.