Novogene 诺禾致源 | 基因科技守护生命健康 Advancing Genomics, Improving life

Fungi draftmap Method

1 Genome sequencing

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

The genomic DNA sample was fragmented into short fragments. These DNA fragments were then end-repaired, A-tailed, and ligated with full-length adapters for Illumina sequencing before further size selection. PCR amplification was then conducted unless specified as PCR-free. Purification was then conducted through the AMPure XP system (Beverly). The resulting library was assessed on the Agilent Fragment Analyzer System (Agilent) and quantified to 1.5nM through Qubit (Thermo Fisher Scientific) and qPCR.

The qualified libraries were pooled and sequenced on Illumina platforms, according to the effective library concentration and data amount required.

2 Genome assembly

2.1 Data processing

The raw data obtained by sequencing (Raw Data) had a certain proportion of low-quality data. In order to ensure the accuracy and reliability of the subsequent information analysis results, the original data must be filtered to obtain valid data (Clean Data).

The specific processing steps included as follows:
1)The reads containing low-quality bases (mass value ≤20) over a certain percentage (the default was 40%) were removed;
2) The number of N in reads beyond a certain proportion (the default was 10%) were removed;
3) Some reads, the overlap between them and the Adapter, which exceeded a certain threshold (the default was 15bp) and had less then 3 mismatches between them, were removed;
4) If the sample such as minigenome was contaminated by host, it needs to be blast with the host data to filter out reads that may originate from the host.

2.2 assembly

The specific processing steps for genome assembly with Clean Data included as follows:
1) Assembled with SOAP denovo software:
Different K-mers (default selections 95, 107, 119) were selected for assembly, According to the project type, based on the optimal kmer and the least scaffolds choosed the best result, further adjusted other parameters (-d -u -R - F, etc.) to obtain preliminary assembly results;
2)The gap close software was used to fill the gap in preliminary assembly results and

removed the same lane pollution by filtering the reads with low sequencing depth (less than 0.35 of the average depth) to obtain the final assembly result;
3) Fragments below 500 bp were filtered out and the final result was counted for gene prediction.

3 Genome Component prediction

Genome component prediction included the prediction of the coding gene, repetitive sequences and non-coding RNA. The available steps were proceeded as follows:

For Fungi, by default, we used the Augustus 2.7 program to retrieve the related coding gene. If homology reference gene sequences and transcript sequencing data were provided， a complete annotation pipeline, PASA, as implemented at the Broad Institute, involves the following steps:
(A) ab initio gene finding using a selection of the following software tools: GeneMarkHMM, FGENESH, Augustus, and SNAP, GlimmerHMM.
(B) protein homology detection and intron resolution using the GeneWise software and the uniref90 non-redundant protein database.
(C) alignment of known ESTs, full-length cDNAs, and most recently, Trinity RNA-Seq assemblies to the genome.
(D) PASA alignment assemblies based on overlapping transcript alignments from step (C).
(E) use of EVidenceModeler (EVM) to compute weighted consensus gene structure annotations based on the above (A, B, C, D).
(F) use of PASA to update the EVM consensus predictions, adding UTR annotations and models for alternatively spliced isoforms (leveraging D and E).

The interspersed repetitive sequences were predicted using the Repeat Masker[7] (http://www.repeatmasker.org/). The tandem Repeats were analyzed by the TRF[8] (Tandem repeats finder).

Transfer RNA (tRNA) genes were predicted by the tRNAscan-SE[9]. Ribosome RNA (rRNA) genes were analyzed by the rRNAmmer[10]. sRNA, snRNA and miRNA were predicted by BLAST against the Rfam[11][12]database.

4 Gene function

We used seven databases to predict gene functions. They were respective GO[16] (Gene Ontology), KEGG[17][18](Kyoto Encyclopedia of Genes and Genomes), KOG[19](Clusters of Orthologous Groups), NR[20] (Non-Redundant Protein Database databases), TCDB[21](Transporter Classification Database), P450 and Swiss-Prot[22]. A whole genome Blast[23] search (E-value less than 1e-5, minimal alignment length percentage larger than 40%) was performed against above seven databases. The secretory proteins were predicted by the Signal P[24]database. Meanwhile, we analyzed the secondary metabolism gene clusters by the antiSMASH[25]. For pathogenic fungi, we added the pathogenicity and drug resistance analyses. We used the PHI[26](Pathogen Host Interactions), DFVF (database of fungal

virulence factors) to perform the above analyses. Carbohydrate-Active enzymes were predicted by the Carbohydrate-Active enZYmes Database[27].

Reference

[1]Lim H J , Lee, E.-H, Yoon Y , et al. Portable lysis apparatus for rapid single-step DNA extraction of Bacillus subtilis[J]. Journal of Applied Microbiology, 2016, 120(2):379-387.

[2]Li R, Zhu H, Ruan J, et al. De novo assembly of human genomes with massively parallel short read sequencing[J]. Genome research, 2010, 20(2): 265-272.

[3]Li R, Li Y, Kristiansen K, et al. SOAP: short oligonucleotide alignment program[J]. Bioinformatics, 2008, 24(5): 713-714.

[4]Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing[J]. Journal of Computational Biology, 2012, 19(5): 455-477.

[5]Simpson JT, Wong K, Jackman SD, et al. ABySS: a parallel assembler for short read sequence data.[J]. Genome Research, 2009, 19(6):1117.

[6]Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions[J]. Nucleic Acids Research, 2001, 29(12): 2607-2618.

[7]Saha S, Bridges S, Magbanua Z V, et al. Empirical comparison of ab initio repeat finding programs[J]. Nucleic acids research, 2008, 36(7): 2284-2294.

[8]Benson G. Tandem repeats finder: a program to analyze DNA sequences[J]. Nucleic acids research, 1999, 27(2): 573.

[9]Lowe T M, Eddy S R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence[J]. Nucleic acids research, 1997, 25(5): 0955-964.

[10]Lagesen K, Hallin P, Rødland E A, et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes[J]. Nucleic acids research, 2007, 35(9): 3100-3108.

[11]Gardner P P, Daub J, Tate J G, et al. Rfam: updates to the RNA families database[J]. Nucleic acids research, 2009, 37(suppl 1): D136-D140.

[12]Nawrocki EP, Kolbe DL, Eddy SR: Infernal 1.0: inference of RNA alignments. Bioinformatics 2009, 25(10):1335-1337.

[13]Hsiao W, Wan I, Jones S J, et al. IslandPath: aiding detection of genomic islands in prokaryotes[J]. Bioinformatics, 2003, 19(3): 418-420.

[14]You Z, YJ Liang, Karlene L, et al. “PHAST: A Fast Phage Search Tool” Nucl. Acids Res, 2011 doi:10.1093/nar/gkr485.

[15]Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats[J]. Nucleic acids research, 2007, 35(suppl 2): W52-W57.

[16]Ashburner M, Ball C A, Blake J A, et al. Gene Ontology: tool for the unification of biology[J]. Nature genetics, 2000, 25(1): 25-29.

[17]Kanehisa M, Goto S, Kawashima S, et al. The KEGG resource for deciphering the genome[J]. Nucleic acids research, 2004, 32(suppl 1): D277-D280.

[18]Kanehisa M, Goto S, Hattori M, et al. From genomics to chemical genomics: new developments in KEGG[J]. Nucleic acids research, 2006, 34(suppl 1): D354-D357.

[19]Galperin MY, Makarova KS, Wolf YI, et al. Expanded microbial genome coverage and improved protein family annotation in the COG database[J]. Nucleic Acids Research, 2015, 43(Database issue):261- 9.

[20]Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases[J]. Bioinformatics, 2002, 18(1): 77-82.

[21]Milton SJ, Vamsee SR, Dorjee GT, et al. The Transporter Classification Database. Nucleic Acids Research, 2014, doi:10.1093/nar/gkt1097.

[22]Amos B, Rolf A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000, 28(1): 45-48.

[23]Mount DW: Using the Basic Local Alignment Search Tool (BLAST). CSH Protoc 2007, 2007:pdb top17.

[24]Petersen T N, Brunak S, von Heijne G, et al. SignalP 4.0: discriminating signal peptides from transmembrane regions[J]. Nature methods. 2011, 8(10): 785-786.

[25]Medema M H, Blin K, Cimermancic P , et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences[J].

[26]Martin U, Rashmi P, Arathi R et al. The Pathogen-Host Interactions database (PHI-base): additions and future developments. Nucleic Acids Research. 2015, doi: 10.1093/nar/gku1165.

[27]Cantarel B L, Coutinho P M, Rancurel C, et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics[J]. Nucleic acids research, 2009, 37(suppl 1):D233-D238.