仅供客户写英文文章时参考，分析内容和方法请以结题报告为准

——三代测序业务线

Sample collection and preparation

Library preparation and sequencing

The Iso-Seq library was prepared according to the Isoform Sequencing protocol (Iso-Seq) using the Clontech SMARTer PCR cDNA Synthesis Kit and the BluePippin Size Selection System protocol as described by Pacific Biosciences (PN 100-092-800-03).

Data Analysis

Data processing

Sequence data were processed using the SMRTlink 5.0 software. Circular consensus sequence (CCS) was generated from subread BAM files, parameters: min_length 200, max_drop_fraction 0.8, no_polish TRUE, min_zscore -9999, min_passes 1, min_predicted_accuracy 0.8, max_length 18000. CCS.BAM files were output, which were then classified into full length and non-full length reads using pbclassify.py script, ignore polyA false, minSeq Length 200. Non-full length and full-length fasta files produced were then fed into the cluster step, which does isoform-level clustering (ICE), followed by final Arrow polishing, hq_quiver_min_accuracy 0.99, bin_by_primer false, bin_size_kb 1, qv_trim_5p 100, qv_trim_3p 30.

Error correction using Illumina reads

Additional nucleotide errors in consensus reads were corrected using the Illumina RNA-seq data with the software LoRDEC.

Mapping to the reference genome

Aligning consensus reads to reference using GMAP with parameters --no-chimeras --cross-species --expand-offsets 1 -B 5 -K 50000 -f samse -n 1 against reference genome.

Gene structure analysis

Gene structure analysis was performed using TAPIS pipeline. The GMAP output bam format file and gff/gtf format genome annotation file were used for gene and transcript determination. Alternative splicing events and Alternative polyadenylation events were then analyzed. Fusion transcripts were determined as transcripts mapping to two or more long-distance range genes and was validated by at least two Illumina reads.

Unmapped transcripts and Novel gene transcripts functional annotation

Unmapped transcripts and Novel gene transcripts function were annotated based on the following databases:

NR (NCBI non-redundant protein sequences);

NT (NCBI non-redundant nucleotide sequences);

Pfam (Protein family);

KOG/COG (Clusters of Orthologous Groups of proteins);

Swiss-Prot (A manually annotated and reviewed protein sequence database);

KO (KEGG Ortholog database);
GO (Gene Ontology).
We used the software of BLAST and set the e-value ‘1e-10’ in NT database analysis;
We used the software of Diamond BLASTX and set the e-value ‘1e-10’ in NR KOG Swiss-Prot KEGG databases analysis;
We used the software of Hmmscan in Pfam database analysis.

TF analysis

Plant transcription factors were predicted using iTAK software, animal was performed using the animalTFDB 2.0 database.

LncRNA analysis

We used CNCI CPC Pfam-scan PLEK four tools to predict the coding potential of transcripts.

- CNCI

CNCI (Coding-Non-Coding-Index) profiles adjoining nucleotide triplets to effectively distinguish protein-coding and non-coding sequences independent of known annotations. We use CNCI with default parameters.

- CPC

CPC (Coding Potential Calculator) mainly through assess the extent and quality of the ORF in a transcript and search the sequences with known protein sequence database to clarify the coding and non-coding transcripts. We used the NCBI eukaryotes' protein database and set the e-value ‘1e-10’ in our analysis.

- Pfam-scan

We translated each transcript in all three possible frames and used Pfam Scan to identify occurrence of any of the known protein family domains documented in the Pfam database. Any transcript with a Pfam hit would be excluded in following steps. Pfam searches use default parameters of -E 0.001 --domE 0.001.

- PLEK

The PLEK SVM classifier uses an optimized K-mer approach to construct the best classifier to assess coding potential for species that lack high-quality genomic sequences and annotations.PLEK used default parameters of -minlength 200.
Transcripts predicted with coding potential by either/all of the three tools above were filtered out, and those without coding potential were our candidate set of lncRNAs.

Quantification of transcript expression

Cuffdiff(v2.1.1) was used to calculate FPKMs of all transcripts in each sample. Isoforms FPKMs were computed by summing the FPKMs of transcripts in each gene group. FPKM means fragments per kilo-base of exon per million fragments mapped, calculated based on the length of the fragments and reads count mapped to this fragment.

Differential Alternative Splice

Suppa was used to calculate expression weight(Psi) of alternative splice based on transcript TPM values.Differential alternative splice of two conditions was performed using significance test of Psi. The dpsi value was adjusted using the Mann-Whitney U test method. The absolute dpsi value of 0.1 and p-value of 0.05 were set as the threshold for significantly differential alternative splice.

Transcript differential expression analysis

Cuffdiff provides statistical routines for determining differential expressionin digital transcript or gene expression data using a model based on the negative binomial distribution. Transcripts with an P-adjust<0.05 were assigned as differentially expressed.

Reads mapping to the reference genome

Reference genome and gene model annotation files were downloaded from genome website directly. Index of the reference genome was built using Hisat2(v2.1.0) and paired-end clean reads were aligned to the reference genome using Hisat2. We selected Hisat2 as the mapping tool for that Hisat2 can generate a database of splice junctions based on the gene model annotation file and thus a better mapping result than other non-splice mapping tools.

Quantification of gene expression level

HTSeq v0.6.1 was used to count the reads numbers mapped to each gene. And then FPKM of each gene was calculated based on the length of the gene and reads count mapped to this gene. FPKM, expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced, considers the effect of sequencing depth and gene length for the reads count at the same time, and is currently the most commonly used method for estimating gene expression levels.

Differential expression analysis

(For DESeq with biological replicates) Differential expression analysis of two conditions/groups (two biological replicates per condition) was performed using the DESeq R package (1.18.0). DESeq provide statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. The resulting P-values were adjusted using the Benjamini and Hochberg's approach for controlling the false discoveryrate . Genes with an adjusted P-value <0.05 found by DESeq were assigned as differentially expressed.

(For DEGSeq without biological replicates) Prior to differential gene expression analysis, for each sequenced library, the read counts were adjusted by edgeR program package through one scaling normalized factor. Differential expression analysis of two conditions was performed using the DEGSeq R package (1.20.0). The P values were adjusted using the Benjamini & Hochberg method. Corrected P-value of 0.005 and log2(Fold change) of 1 were set as the threshold for significantly differential expression.

GO and KEGG enrichment analysis

Gene Ontology (GO) enrichment analysis of differentially expressed genes or lncRNA target genes were implemented by the GOseq R package, in which gene length bias was corrected. GO terms with corrected Pvalue less than 0.05 were considered significantly enriched by differential expressed genes.

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (http://www.genome.jp/kegg/). We used KOBAS software to test the statistical enrichment of differential expression genes or lncRNA target genes in KEGG pathways.

References

Salmela, Leena, and Eric Rivals. "LoRDEC: accurate and efficient long read error correction." Bioinformatics 30.24 (2014): 3506-3514.

Wu T D, Watanabe C K. GMAP:a genomic mapping and alignment program for mRNA and EST sequences[J].Bioinformatics,2005,21(9):1859-1875.

Wang B,Tseng E,Regulski M,et al.Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing[J]. Nature communications,2016,7.

Abdel-Ghany S E, Hamilton M, Jacobi J L, et al. A survey of the sorghum transcriptome using single-molecule long reads[J]. Nature communications, 2016, 7.

Rogers M F, Thomas J, Reddy A S N, et al. SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data[J].Genome biology, 2012,13(1):R4.

Weirather,J. L.,et al.(2015)."Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing." Nucleic Acids Res 43(18):e116.

Shimizu K, Adachi J, Muraoka Y. ANGLE: a sequencing errors resistant program for predicting protein coding regions in unfinished cDNA[J]. Journal of bioinformatics and computational biology,2006,4(03):649-664.

Zhang H M,Liu T,Liu C J,et al. AnimalTFDB 2.0: A resource for expression, prediction and functional study of animal transcription factors[J]. Nucleic acids research,2014:gku887.

Liang Sun, Haitao Luo, Dechao Bu, Guoguang Zhao, Kuntao Yu, Changhai Zhang, Yuanning Liu, RunSheng Chen and Yi Zhao*, Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Research (2013), doi: 10.1093/nar/gkt646.

L. Kong, Y. Zhang, Z.Q. Ye, X.Q. Liu, S.Q. Zhao, L. Wei, and G. Gao. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine[J].Nucleic Acids Res 36: W345-349.

Aimin Li, Junying Zhang and Zhongyin Zhou. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme[J].BMC Bioinformatics 2014, 15:311

R.D.Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, S.C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G.A. Salazar, J. Tate, A. Bateman, The Pfam protein families database: towards a more sustainable future.[J].Nucleic Acids Research (2016) Database Issue 44:D279-D285.

Gael P. Alamancos, Amadís Pagès, Juan L. Trincado, Nicolás Bellora, and Eduardo Eyras, Leveraging transcript quantification for fast computation of alternative splicing profiles[J].RNA. 2015 Sep; 21(9): 1521–1531.

Alamancos, Gael P., et al. "Leveraging transcript quantification for fast computation of alternative splicing profiles." Rna 21.9 (2015): 1521-1531.