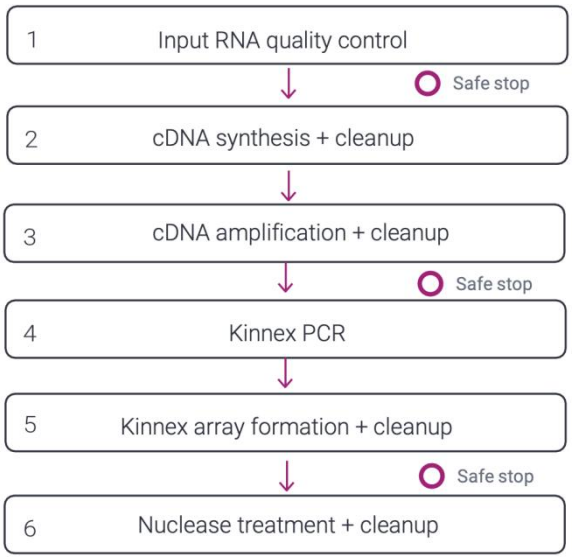
Novogene

For reference only.

Library construction and sequencing process

From RNA samples to final data acquisition, each link of sample testing, library construction, and sequencing will affect the quality and quantity of data, and the data quality will directly affect the results of subsequent information analysis. In order to ensure the accuracy and reliability of sequencing data from the source, Novogene strictly controls each production step of sample testing, library construction, and sequencing, fundamentally ensuring the output of high-quality data. The Kinnex full-length transcriptome library uses Iso-Seq express 2.0 kit to synthesize cDNA chains and Kinnex full-length RNA kit to build the library. The library construction process is as follows:

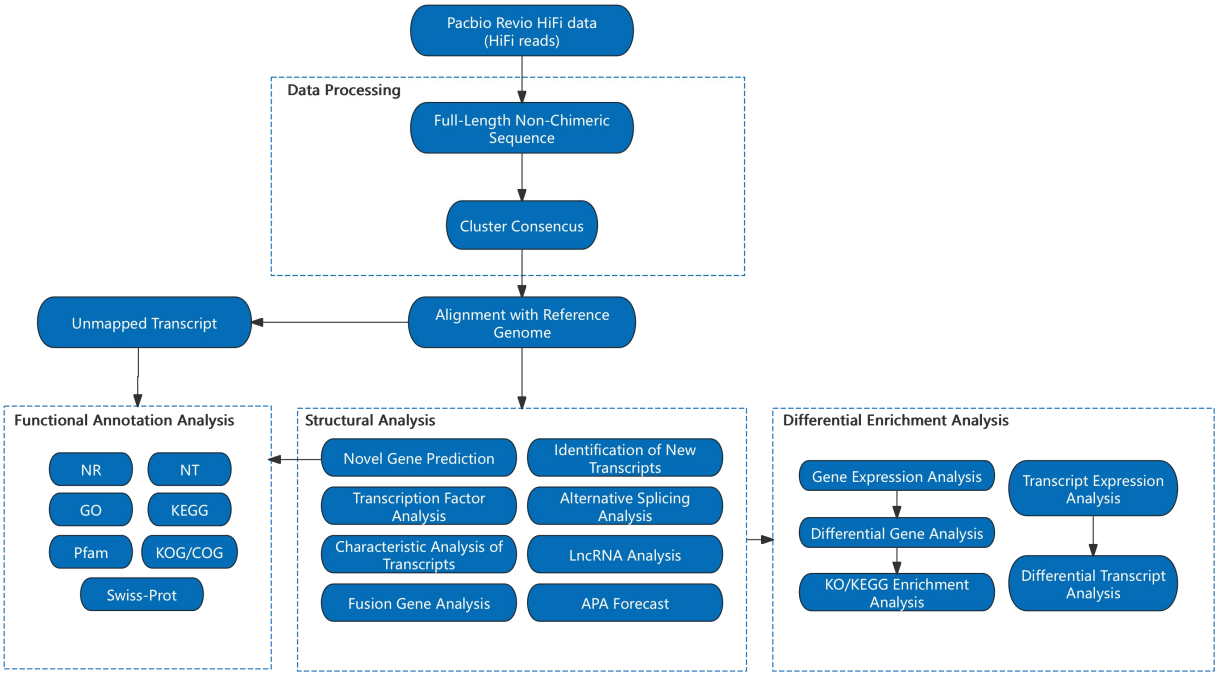


Official protocol: https://www.pacb.com/wp-content/uploads/Procedure-checklist-Preparing-Kinnex-libraries-using-the-Kinnex-full-length-RNA-kit.pdf

Data analysis

Novogene

1. Information Analysis Process



2. Full-length transcript analysis

After sequencing is completed, the original data will be removed from the machine for adapters and low-quality reads. The data will contain data measured with 0, 1 or 2 molecules as templates, but the information of 0 or 2 molecules will cause great interference to subsequent information analysis. This part of the data will be discarded and only the data measured with 1 molecule will be retained.

Revio platform is hifi reads after downloading, and then use software SMRTlink v 13.0 Filter and process hifi reads :

1. CCS is classified by detecting whether it contains 5' -primer, 3'-primer, and poly-A to find out FLNC (full-length non chimera) sequence and nFL (Non-Full-Length: non-full-length non-chimera sequence) sequence;

2. Cluster the FLNC sequences of the same transcript using the hierarchical n*log(n) algorithm to obtain the consensus sequence .

3. Reference genome analysis

The consensus sequence was aligned to the reference genome using pbmm2 (--preset ISOSEQ) software . pbmm2 is a C++ program rewritten by SMRT based on minimap2 software. Its principle is the same as minimap2 [1]

4. Transcript database annotation

In order to obtain comprehensive annotation information, transcripts that were not aligned to the

reference genome and transcripts with structural annotation were compared with seven major databases: NR [2], NT , Pfam, KOG/COG [3], Swiss-prot [4], KEGG [5], and GO [6]. The characteristics, URLs, annotation methods, and parameters of these seven databases are as follows:

database	Introduction	More Information	Software and parameters
Nr	NCBI non-redundant protein sequences, NCBI official protein sequence database	It includes protein coding sequences of GenBank genes, PDB (Protein Data Bank) protein database, SwissProt protein sequences, and protein sequences from databases such as PIR (Protein Information esource) and PRF (Protein Research Foundation). Its characteristics are that the content is relatively comprehensive, and the annotation results will contain species information, which can be used for species classification.	diamond v0.8.36; e-value = 1e-5, --more-sensitive
Nt	NCBI nucleotide sequences, NCBI official nucleic acid sequence database	Includes nucleotide sequences from GenBank, EMBL and DDBJ (but not EST, STS, GSS, WGS, TSA, PAT, HTG sequences).	ncbi-blast-2.7.1+; e-value = 1e-5
Pfam	Protein family, the most comprehensive classification system for protein domain annotation	Proteins are composed of domains, and the protein sequence of each specific domain is conservative to a certain extent. PFAM divides the domains of proteins into different protein families, and establishes the HMM statistical model of the amino acid sequence of each family by comparing the protein sequences. PFAM families are divided into two categories according to the reliability of the annotation results: the Pfam-A family with high reliability of manual annotation and the Pfam-B family automatically generated by the program. Through the HMMER3 program, you can search for the HMM model of the established protein domain, so as to annotate the protein family of Gene. For details, see http://pfam.sanger.ac.uk/.	HMMER 3.1 package, hmmscan; --acc
KOG/COG	COG: Clusters of Orthologous Groups of proteins;KOG: euKaryotic Ortholog Groups	Both KOG and COG are based on gene orthologous relationships in NCBI, of which COG is for prokaryotes and KOG is for eukaryotes. COG/KOG divides homologous genes from different species into different ortholog clusters based on evolutionary relationships. Currently, COG has 4873 classifications and KOG has 4852 classifications. Genes from the same ortholog have the same function, so the functional annotations can be directly inherited to other members of the same COG/KOG cluster. For details, see http://www.ncbi.nlm.nih.gov/COG/.	diamond v0.8.36; e-value = 1e-5, --more-sensitive
Swiss-Prot	A manually annotated and reviewed protein	A collection of protein sequences that have been collated and studied by experienced biologists. For details, see	diamond v0.8.36; e-value = 1e-5, --more-sensitive

	sequence database	http://www.ebi.ac.uk/uniprot/.	
KEGG	Kyoto Encyclopedia of Genes and Genomes	A database that systematically analyzes the metabolic pathways of gene products and compounds in cells and the functions of these gene products. It integrates data on genomes, chemical molecules, and biochemical systems, including metabolic pathways (KEGG PATHWAY), drugs (KEGG DRUG), diseases (KEGG DISEASE), functional models (KEGG MODULE), gene sequences (KEGG GENES), and genomes (KEGG GENOME). The KO (KEGG ORTHOLOG) system links various KEGG annotation systems together. KEGG has established a complete KO annotation system that can complete the functional annotation of the genome or transcriptome of newly sequenced species. For details, see http://www.genome.jp/kegg/ .	diamond v0.8.36; e-value = 1e-5, --more-sensitive
GO	Gene Ontology, an internationally standardized classification system for describing gene function	GO is divided into three categories of ontologies: biological process, molecular function and cellular component, which are used to describe the biological process, molecular function and cellular environment of the products encoded by genes. The basic unit of GO is term, each term has a unique identifier (consisting of "GO:" plus 7 numbers, such as GO:0072669); the terms of each category of ontology form a directed acyclic topological structure through the connections between them (is_a, part_of, regulate). For details, please see http://www.geneontology.org/.	Self-written script based on protein annotation results of Pfam database

5. Gene structure analysis

6.1 Classification of full-length transcripts and polyadenylation analysis

Based on the pbmm2 alignment results, isoseq collapse was used to remove transcript redundancy with the following parameters: --do-not-collapse-extra-5exons --min-aln-coverage 0.95 --min-aln-identity 0.95 --max-fuzzy-junction 5. Pigeon was officially packaged by SMRT based on SQANTI3 [7]. It classified and analyzed the features of non-redundant full-length transcripts , and finally classified the transcripts according to the alignment results and splicing sites:

Category	Description
FSM (Full Splice Match)	The reference and query isoform have the same number of exons and each internal junction matches the positions of the reference. The exact 5' start and 3' end can differ within the first/last exons.
ISM (Incomplete Splice Match)	The query isoform has fewer external exons than the reference, but each internal junction matches the positions of the reference. The exact 5' start and 3' end can differ within the first/last exons.
NIC (Novel In Catalog)	The query isoform does not have a FSM or ISM match, but is using a combination of known donor/acceptor sites.
NNC (Novel Not in Catalog)	The query isoform does not have a FSM or ISM match, and has at least one donor or acceptor site that is not annotated.

Category	Description
Antisense	The query isoform does not have overlap a same-strand reference gene but is anti-sense to an annotated gene.
Genic Intron	The query isoform is completely contained within a reference intron.
Genic Genomic	The query isoform overlaps with introns and exons.
Intergenic	The query isoform is in the intergenic region.

6.2 Alternative splicing analysis

Alternative splicing analysis was performed using SUPPA [8] software (default parameters).

6.3 Transcription factor analysis

Animal transcription factors were identified using the animal transcription factor database [9] - animalTFDB 2.0. For species included in the database, if the gene is from Ensembl geneid, transcription factors were directly screened. For genes that are not from Ensembl geneid, transcription factors were compared with the gene of the species in the database. The known transcription factor protein sequences were screened by BLASTX; for species not included in the database, hmmsearch was used to identify them based on the pfam file of the transcription factor family. Plant transcription factors were identified using iTAK 1.7a (-f 3F).

6.4 LncRNA Analysis

Due to the limitation of library construction principle, we can only obtain lncRNAs with poly A tails. We used CNCI [10], PLEK [11], CPC [12] software and Pfam [13] database to predict the coding potential of PacBio sequencing data, and then carried out subsequent analysis on the obtained lncRNAs.

The steps of LncRNA analysis are as follows:

1. PLEK and CNCI software predict coding potential based on the sequence characteristics of transcripts. The PLEK support vector machine classifier uses the optimized K-mer method to construct the best classifier to evaluate coding potential, which is suitable for species that lack high-quality genome sequences and annotation information. CNCI effectively distinguishes coding and non-coding sequences based on the spectrum of adjacent trinucleotides, and can effectively predict the coding potential of incomplete transcripts and antisense transcripts. In order to accurately predict lncRNA, PLEK and CNCI software are used to predict the coding potential of transcripts;

2. The transcript sequences predicted by PLEK and CNCI software are then compared with the known protein database by BLAST using CPC software. CPC evaluates the coding potential of transcripts based on the biological sequence characteristics of each coding frame of the transcripts through the support vector machine classifier;

3. The transcript sequences predicted by PLEK, CNCI and CPC software were subjected to hmmscan homology search with Pfam-A and Pfam-B databases. The Pfam-A database records the high-quality structural domains of most known proteins, while the Pfam-B database covers the

structural domain family more comprehensively. After database comparison, the coding potential was predicted more accurately, and the lncRNA sequence was finally obtained.

6.5 Fusion gene analysis

Fusion gene identification was performed using GMAP v2017-06-20 (--max-intronlength-ends 50000; -f 4; -z sense_force; -n 0).

6. Quantitative analysis (based on third-generation sequencing data)

6.1 Results of three-generation comparison

Use minimap2 to align FLNC data to the reference genome and count the mapping of each sample

6.2 Gene expression level analysis

IsoQuant [14] software was used for quantitative analysis at the gene and transcript levels to directly obtain read count values and TPM values.

7. Variance Analysis

The input data for differentially expressed gene analysis is the readcount data obtained from the gene expression level analysis. The analysis is mainly divided into three parts:

1. First, normalize the readcount;
2. Then the hypothesis test probability (pvalue) is calculated based on the model;
3. Finally, multiple hypothesis testing correction was performed to obtain the FDR value (false discovery rate).

Different software will be used to analyze differentially expressed genes in different situations. The analysis methods are as follows:

type	software	Standardized methods	pvalue calculation model	FDR calculation method	Differential gene screening criteria
Biological replication	DESeq2 (Anders et al, 2014)	DESeq	Negative binomial distribution	BH	|log2(FoldChange)| > 0 & padj < 0.05
No biological replication	DEGseq (Wang et al, 2010)	TMM	Poisson distribution	BH	|log2(FoldChange)| > 1 & padj < 0.005

8. GO enrichment analysis of differentially expressed genes

We used GOseq software for GO enrichment analysis. The software is based on the Wallenius non-central hyper-geometric distribution. Compared with the ordinary hyper-geometric distribution, the characteristic of this distribution is that the probability of extracting an individual from a certain category is different from the probability of extracting an individual from outside a certain category. This difference in probability is obtained by estimating the preference for gene length, so that the probability of GO terms being enriched by differential genes can be calculated more accurately.

Novogene

9. KEGG enrichment analysis of differentially expressed genes

KOBAS v3.0 (Corrected P-Value < 0.05) was used for KEGG enrichment analysis. The integrated metabolic pathway query provided by KEGG is excellent, including the metabolism of carbohydrates, nucleosides, amino acids, etc. and the biodegradation of organic matter. It not only provides all possible metabolic pathways, but also comprehensively annotates the enzymes that catalyze each step of the reaction, including amino acid sequences, links to PDB libraries, etc. It is a powerful tool for in vivo metabolic analysis and metabolic network research.

10. How to decompress the result file

Compression format	User Type	method
Compressed file in the form of file.tar.gz	Unix/Linux/Mac users	Use tar -zxvf file.tar.gz command
	Windows Users	Use decompression software such as WinRAR, 7-Zip, etc.
Compressed file in file.gz format	Unix/Linux/Mac users	Use gzip -d file.gz command
	Windows Users	Use decompression software such as WinRAR, 7-Zip, etc.
Compressed file in file.zip format	Unix/Linux/Mac users	Use unzip file.zip command
	Windows Users	Use decompression software such as WinRAR, 7-Zip, etc.

11. Result file format description

File Type	File Description	Open
file.fa/fasta	Sequence file, fasta format, usually gene sequence or genome sequence. Because the file is generally large, it is difficult to open	Unix/Linux/Mac users use the less or more command
		Windows users use advanced text editors such as Editplus/Notepad++
file.fq/fastq	Sequence file, fastq format, usually reads sequence; because the file is generally large, it is difficult to open	Unix/Linux/Mac users use the less or more command
		Windows users use advanced text editors such as Editplus/Notepad++
file.txt/xls	Result data table file; the file is separated by tabs	Unix/Linux/Mac users use the less or more command
		Windows users can use advanced text editors such as Editplus/Notepad++, or open with Microsoft Excel
file.pdf/svg	Result image file; vector image, can be enlarged or reduced without distortion, convenient for users to view and edit, can be edited using Adobe Illustrator, used for article publishing, etc.	Windows/Mac users can use Adobe Reader/Foxit Reader/web browser to open
		Unix/Linux users use the evince command to open
file.png	Resulting image file; bitmap, lossless compression	Unix/Linux/Mac users use the display command to open
		Windows users can use image browsers to open, such as Photoshop, etc.

Novogene

12. Analysis software list and versions

Analysis.content	Software and Version	Parameter	Remark	Software Link
Raw data processing and analysis	SMRT-Link V13.0	Default parameters	FLNC Identification and Transcript Deduplication	https://www.pacb.com/support/software-downloads
Database Function Notes	Diamond blastx v0.8.36	--more-sensitive -k 10；-e 1e-5；-f 6；-p 4	NR、KOG/COG、Swiss-prot、KEGG	https://github.com/bbuchfink/diamond
	ncbi-blast-2.7.1+ blastn V2.7.1	-outfmt 6；-evalue 1e-5；-max_target_seqs 10；-num_threads 4	NT Annotation	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
	Hmmscan v3.1b2	--acc；--domtblout	Pfam Annotation	http://hmmer.org/download.html
Gene Structure Analysis	pbmm2 v1.13.1(with minimap2 v2.26)	--preset ISOSEQ；--sort	Reference Genome Alignment	https://github.com/PacificBiosciences/pbmm2.git
	pigeon v1.2.0	Default parameters	Transcript Structure Identification	https://isoseq.how/classification/pigeon.html
	SUPPA V2.3	Default parameters	Alternative Splicing Identification	https://bitbucket.org/regulatorygenomicsupf/suppa
Transcription Factor Identification	Plant：iTAK iTAK：1.7a	-f 3F	Transcription Factor Identification	https://github.com/kentnf/iTAK/
	Animal：AnimalTFDB AnimalTFDB：2.0	Default parameters	Transcription Factor Identification	http://bioinfo.life.hust.edu.cn/AnimalTFDB/
lncRNA Analysis	CPC v0.9	Default parameters	Coding Potential Prediction	http://cpc.cbi.pku.edu.cn/

Novogene

Analysis.content	Software and Version	Parameter	Remark	Software Link
	CNCI V2	Default parameters	Coding Potential Prediction	https://github.com/www-bioinfo-org/CNCI
	PLEK v1.2	Default parameters	Coding Potential Prediction	https://sourceforge.net/projects/plek/
	PfamScan V1.6	Default parameters	Protein Domain Detection	https://www.ebi.ac.uk/seqdb/confluence/display/THD/PfamScan
Gene Fusion Identification	GMAP v2017-06-20	--max-intronlength-ends 50000; -f 4; -z sense_force; -n 0	Gene Fusion Identification	http://research-pub.gene.com/gmap/
Alignment Quantification	IsoQuant V3.3	Default parameters	Gene-Level Quantification	https://github.com/ablab/IsoQuant
Differential Expression Analysis	DEseq v1.10.1	qvalue < 0.05	With Biological Replicates	http://www.bioconductor.org/packages/release/bioc/html/DESeq.html
	DEGseq v1.12.0	|log2(FoldChange)| > 1 & qvalue < 0.005	Without Biological Replicates	http://www.bioconductor.org/packages/release/bioc/html/DEGseq.html
GO Enrichment	GOSeq V1.10.0	Corrected P-Value<0.05	GO Enrichment Analysis	http://www.bioconductor.org/packages/release/bioc/html/goseq.html
KEGG Enrichment	KOBAS v3.0	Corrected P-Value<0.05	KEGG Enrichment Analysis	http://kobas.cbi.pku.edu.cn/download.php

References

[1]Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094-3100. doi: 10.1093/bioinformatics/bty191. PMID: 29750242; PMCID: PMC6137996.

[2] LI W, LUKASZ J, ADAM G. Tolerating Some Redundancy Significantly Speeds up Clustering of Large Protein Databases.[J]. Bioinformatics, 2002(1): 77–82.

[3] TATUSOV RL, FEDOROVA ND, JACKSON JD, et al. The Cog Database: An Updated Version Includes Eukaryotes[J]. Bmc Bioinformatics, 2003, 4(1): 41–41.

[4] AMOS B, ROLF A. The Swiss-Prot Protein Sequence Database and Its Supplement Trembl in 2000[J]. Nucleic Acids Research, 2000(1): 45.

[5] MINORU K, SUSUMU G, SHUICHI K, et al. The Kegg Resource for Deciphering the Genome[J]. Nucleic Acids Research, 2004, 32(Database issue): D277.

[6] ASHBURNER M, BALL CA, BLAKE JA, et al. Gene Ontology: Tool for the Unification of Biology. the Gene Ontology Consortium.[J]. Nature Genetics, 2000, 25(1): 25–9.

[7]Francisco J. Pardo-Palacios, Angeles Arzalluz-Luque, Liudmyla Kondratova, Pedro Salguero, Jorge Mestre-Tomás, Rocío Amorín, Eva Estevan-Morió, Tianyuan Liu, Adalena Nanni, Lauren McIntyre, Elizabeth Tseng, Ana Conesa.bioRxiv 2023.05.17.541248; doi: https://doi.org/10.1101/2023.05.17.541248

[8]Alamancos GP, Pagès A, Trincado JL, et al. Leveraging transcript quantification for fast computation of alternative splicing profiles[J]. Rna, 2015, 21(9): 1521-1531.

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

[10] 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.

[11] 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

[12]L. Kong, Y. Zhang, ZQ Ye, XQ Liu, SQ 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.

[1 3]RDFinn, P. Coggill, RY Eberhardt, SR Eddy, J. Mistry, AL Mitchell, SC Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, GA 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.

[14]Prjibelski, AD, Mikheenko, A., Joglekar, A. et al. Accurate isoform discovery with

IsoQuant using long reads. Nat Biotechnol 41, 915–918 (2023).

https://doi.org/10.1038/s41587 -022-01565-y