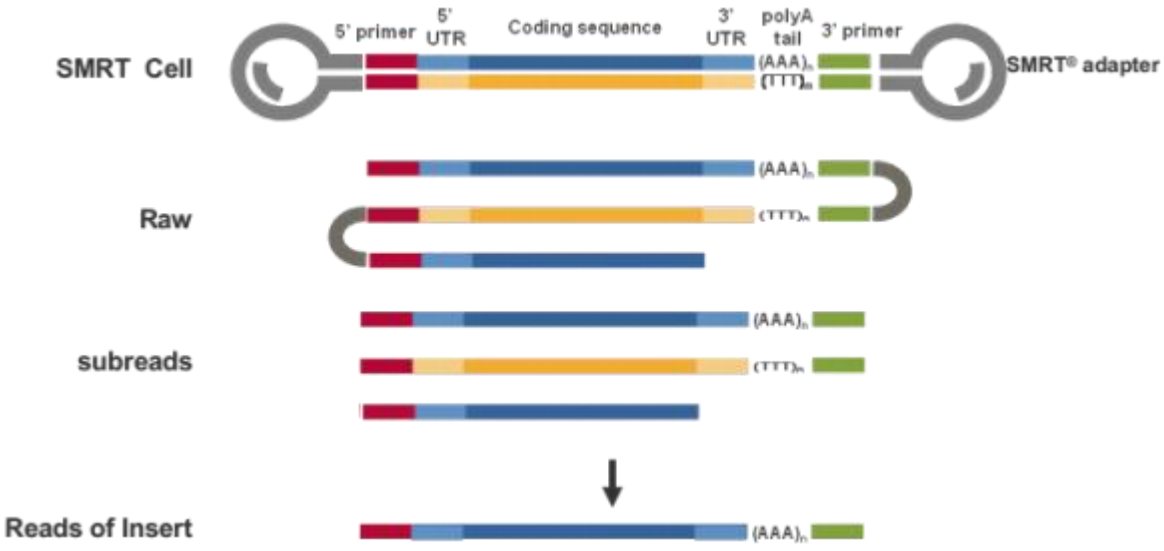
Novogene | Advancing Genomics, Improving Life

Full-length transcriptome common Q&A

2023

Q1: Explanation of related terms



ZMW:

Zero-mode waveguide hole (ZMW): Zero-mode waveguide hole is the smallest sequencing unit of the Sequel sequencing platform, and each SMRTcell contains 1 million zero-mode waveguide holes. During sequencing, the sequence in the well will be read repeatedly, and one subread is obtained for each read, that is, all Subreads in the same ZMW come from the same transcript.

Polymerase Read:

Enzyme polymerase read: The nucleic acid sequence synthesized by DNA polymerase with SMRTbellTM circular template chain, which can be used for quality control of each round (run) in the sequencing process. Polymerase reads are filtered and only high-quality fragments remain, including adapter sequences and copies of multiple sequences synthesized through circular template strands, as shown in the figure'raw'.

Subread:

Each polymerase sequence (polymerase read) can be divided into one or more subsequences (Subread), and the subread is synthesized by the polymerase through a round of (passes) with a template strand of SMRTbellTM, excluding the linker sequence. Each Subreads contains quality values and associated enzyme activity parameters. There are 2 and a half Subreads in the figure, and the number of fullpasses (complete subreads) is 2.

number of full passes:

Refers to the existence of subsequences in the original sequence that contain

SMRTbell™ adapters (adapters, gray areas in the figure) at both ends (the number of sequences between adapters, in the figure, the number of fullpasses is 2.

Circular Consensus (CCS) Read:

The CCS sequence is a consensus sequence obtained from the (Subreads) subsequences in each ZMW well, and there is no need to compare reference sequences. Different from ROI sequence, CCS sequence requires at least 1 complete (full-pass) Subreads in each insert sequence.

Full-Length (FL) Read versus Non-Full-Length (nFL) Read:

SMRTAnalysis software defines that both ends contain 3 primers and 5 primers, and the sequence containing polyA tail before the 3 primers is called full-length sequence (FulL-Length (FL) Read). 5 or 3 primers can be Clontech or other full-length cDNA library construction primers, or gene-specific RT-PCR primers, and vice versa, non-full-length reads.

Full-Length non-chimericRead (FLNC):

The chimeric sequence generated by the direct connection of two cDNA template strands due to the low adapter concentration or SMRTbell concentration during library construction is called artificial chimeric sequence, as shown in the figure below. Non-chimeric sequences in the full-length sequence are called full-length non-chimeric sequences.

Q2: What is the definition of full transcript?

A: Detect the CCS formed by self-error correction of single-molecule multiple sequencing sequences. If the CCS contains 5-primer, 3'primer, and poly-A, the CCS sequence is defined as the full-length transcript sequence.

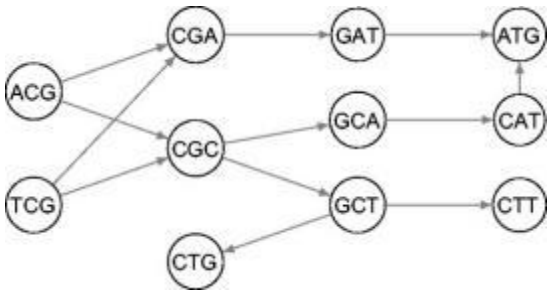
Q3: What is the meaning of Consensus sequence ID? What does the Consensus sequence ID mean?

A: For example, i0_HQ_HepS_c0/f3p0/922) i: Divide Subreads into multiple sections with a length of 1k; i0 represents the section whose length of Subreads is in the range of 0-1k); HQ/LQ: Indicated by HQ It is high-quality reads, that is, Accuracy299% and supported by at least 2 full-length sequences, and LQ means low-quality reads, that is, Accuracy<99%. ;c: FLNC is divided into multiple clusters according to the sequence similarity (c0 represents the first cluster group); f3p0:f represents the number of full-length non-chimeric sequences, and p represents the number of non-full-length non-chimeric sequences number; 992: represents the base

number of the sequence.

Q4: What is a DBG graph?

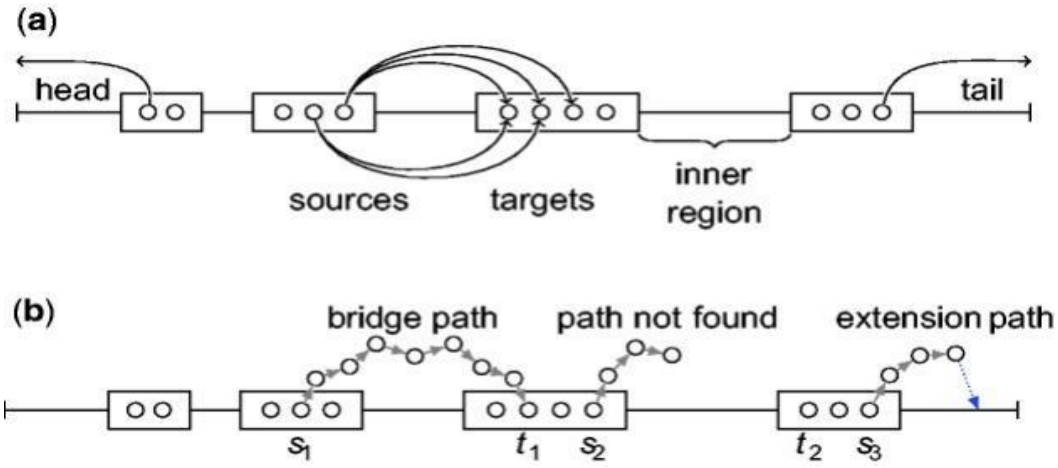
A: First, short reads are broken into K-length nucleic acid fragments, that is, Kmers, and then the DBG graph is constructed using the overlap relationship between Kmers.



Note: The picture above is the DBG map of kmer=3.

Q5: What is LoDERC's correction method for Long read?

A: (a) According to the DBG of short reads, long reads are divided into weak and solid regions (respectively, straight lines and rectangles). The start and end weak regions of long reads are called head and tail, while other weak regions are called inner regions. The circles in the rectangles represent the kmers, and the kmers around the weak areas are used as sources and targets to find paths in the DBG. Each internal area will look for multiple paths. (b) A DBG path is found from s1 to t1 to correct the inner region, but no suitable path can be found between s2 and t2 for correction. As for the tail, the extension path is searched starting from s3, and once found, the corrected sequence is aligned with the tail (dotted arrow in the figure).



Q6: How to count common and unique transcripts among samples?

A: Combine two or more three-generation sample data to remove redundancy, and use CD-HIT

software again (parameters: -c 0.95 -T 6 -G 0 -aL 0.90 -AL 100 -aS 0.99 -AS 30) Transcript clustering analysis was performed, and the unique and common transcripts among all samples were analyzed according to the clustering results. The comparative analysis of transcripts here is based on sequence similarity (>95%) and is not related to the expression level of the transcripts. Therefore, a sample-specific transcript here does not refer to a specifically expressed transcript.

Q7: What is the explanation of the ID and its description information in the CDS forecast file?

A: Take <seq_id> type:<tag>-<completeness> len:<CDS length (aa)> strand:<strand> pos:<CDS range (bp)> as an example:

seq_id : Transcript id plus index number (eg m.1)

type: sequence type, tag attribute includes confident (the transcript predicts only one CDS sequence), likely (the transcript predicts multiple CDS sequences, but only one exceeds the threshold of 100aa), suspicious (there are multiple transcripts exceeding the threshold CDS sequence), dumb (CDS obtained by fuzzy prediction); the completeness attribute can be divided into complete (complete), 5partial (5' part), 3partial (3' part), internal (internal), NA (unconfirmed: corresponding The tag attribute is the CDS sequence of likely and suspicious)

len: the length of the CDS sequence (aa)

strand : Positive and negative strands

pos: the start and end region of the CDS sequence on the transcript (nt)

Q8: Why is the predicted CDS number in the result file greater than the Unigene number?

A: When using the officially recommended Angel software for CDS prediction, it will start according to the three coding reading frames in the forward direction. There are three possibilities, that is, starting from the first to third bases in the forward direction (marked as 0, 1, 2) for reading and translation. According to the prediction results, multiple CDS sequences (amino acid length ≥ 50) may be output in the end, and the CDS sequence with the longest transcript can be selected as the best prediction result.

Q9: Why some transcript ids have corresponding annotation results, but there is no corresponding prediction result in CDS prediction?

A: The CDS prediction does not depend on the annotation results of the database. For each transcript, ANGEL software uses the trained model to predict the most

probable codons in each of the three ORFs, which are then taken as the most probable ORFs. If only one ORF is predicted in a transcript, it will be marked as confident; if there are multiple ORFs, but only one exceeds the minimum amino acid length threshold (length ≥ 50aa), then output an ORF and mark it as likely; if there are multiple ORFs meeting the threshold, output all ORF sequences and mark them as suspicious.

The reason why the CDS sequence is not predicted by analyzing the transcript may be: 1. The training model does not support the prediction of the ORF of this transcript; 2. The CDS predicted by this transcript does not meet the threshold of 50aa and is filtered out.

Q10: How to calculate gene expression level?

A: In RNA-seq technology, FPKM (expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced) is the number of fragments per kilobase of a gene in every million fragments, which also takes into account sequencing The effect of depth and gene length on fragment counts is currently the most commonly used method for estimating gene expression levels.

Q11: What are the differences and significance of H-clster and K-means cluster analysis?

A: Cluster analysis is used to judge the expression patterns of differential genes under different experimental conditions. By clustering genes with the same or similar expression patterns into clusters, the functions of unknown genes or unknown functions of known genes can be identified. These homogeneous genes may have similar functions, or participate in the same metabolic process or cellular pathway. The number of subclusters in each clustering method is not artificially set. H-cluster and K-means cluster use corresponding distance algorithms to calculate the distance between each gene, and calculate the relative distance between genes through repeated iterations. Finally, According to the relative distance of genes, it is divided into different subclusters. Due to different clustering algorithms (H-clster adopts hierarchical clustering algorithm, K-means cluster adopts K-means clustering algorithm), the number of subclusters obtained by the two clustering methods will be different.

Q12: A gene has a large difference in expression between two samples, but it does not exist in the list of genes with significant differences. Why?

The screening of differential genes is based on statistical significance, and the presence or absence of differential genes cannot be judged intuitively by the magnitude of the two values. First of all: Affected by the sequencing depth, some samples have a deeper sequencing depth, which may lead to a higher readcount value of the sample. The first step in differential analysis is to eliminate the impact of sequencing depth and standardize the original data (we are in For repeated items, use the standardization method that comes with DESeq; for non-repetitive items, use the TMM standardization method); secondly: in the process of difference

analysis, the distribution of readcount needs to be estimated. Experience shows that readcount obeys the negative binomial distribution. In projects with duplication, the quality of the duplication will also have an impact on whether there are differential genes. If the replication is poor, the differences within groups will mask some of the differences between groups. After estimating the parameters, it is necessary to use a specific test method to determine whether there are differential genes; again: after calculating the pvalue, it is necessary to correct the pvalue for multiple hypothesis testing to reduce false positives. This process will make padj greater than the original pvalue, so that some genes that pass the pvalue threshold cannot pass the padj threshold.

Q13: The readcount of a gene is 0, but it also has foldchange, pvalue, and qvalue? The readcount value of a certain gene is 0, but it also has foldchange, pvalue, and qvalue values?

A: In DESeq, if the corrected readcount of a gene is 0 in one sample but not 0 in another sample, foldchange will be INF or -INF; if both values are 0, log2foldchange and The pvalue and qvalue values are both NA; in DEGseq, if the corrected readcount of a gene in a sample is 0, the software will make a slight correction to 0 by default, and correct it to a value close to 0, but not 0 value, it will generate foldchange and pvalue, qvalue value.

Q14: Why do multiple hypothesis testing to calculate padj value instead of directly using pvalue to screen differential genes?

A: There is no problem with using pvalue for a single hypothesis test, but in the process of difference analysis, a hypothesis test must be performed for each gene. A species often has tens of thousands of genes, and tens of thousands of hypothesis tests must be performed, so false positives are reduced. greatly increase. Assuming that the pvalue is 0.05 (only 5 of the 100 differential genes are false positives), this accuracy is sufficient for a gene that has been subjected to a hypothesis test, but for tens of thousands of genes as a whole, the accuracy is far from Not enough, because for every 10,000 genes tested, 500 will be false positives. In order to control reasonably, it is necessary to introduce a more stringent indicator, that is, the padj value. Of course, if there are too few differential genes, you can also use pvalue as the next best thing, which is biologically meaningful and can be verified by experiments.

Q15: What is the meaning of the nodes and background colors in the KEGG Pathway picture?

A: There are 5 types of KEGG Pathway, namely:

map : Reference pathway

ko: Reference pathway (KO)

ec: Reference pathway (EC)

rn : Reference pathway (Reaction)

org: Organism-specific pathway map

1. "map" pathway: The node represents a gene, the enzyme encoded by the gene, and the reaction that the enzyme participates in. Hover the mouse over a person node to see the above information, and the background color is colorless.

2. "ko/ec/rn" pathway: the nodes in the ko pathway only represent genes; the nodes in the ec pathway only represent related enzymes; the nodes in the pathway "ko/ec/rn" pathway: only represent the genes involved in this point A reaction, reactants, and reaction type. The background color is represented by blue.

3. "org" pathway: species-specific pathway map. Node has the same meaning as map pathway. The background color of the species-related nodes is green, indicating that each pathway of the data path has a corresponding unique number, such as map00010, which can be queried on the official website of the kegg database. The pathway diagrams obtained from the reference genome project are all Organism-specific pathways.

Q16: What are the meanings of n, N, i, M in the schematic diagram of enrichment analysis?

A: In the enrichment analysis results, i represents the number of differential genes annotated to a pathway, n represents the number of differential genes annotated to all pathways, M is the number of background genes annotated to a pathway, N is the number of genes annotated to all pathways number of background genes