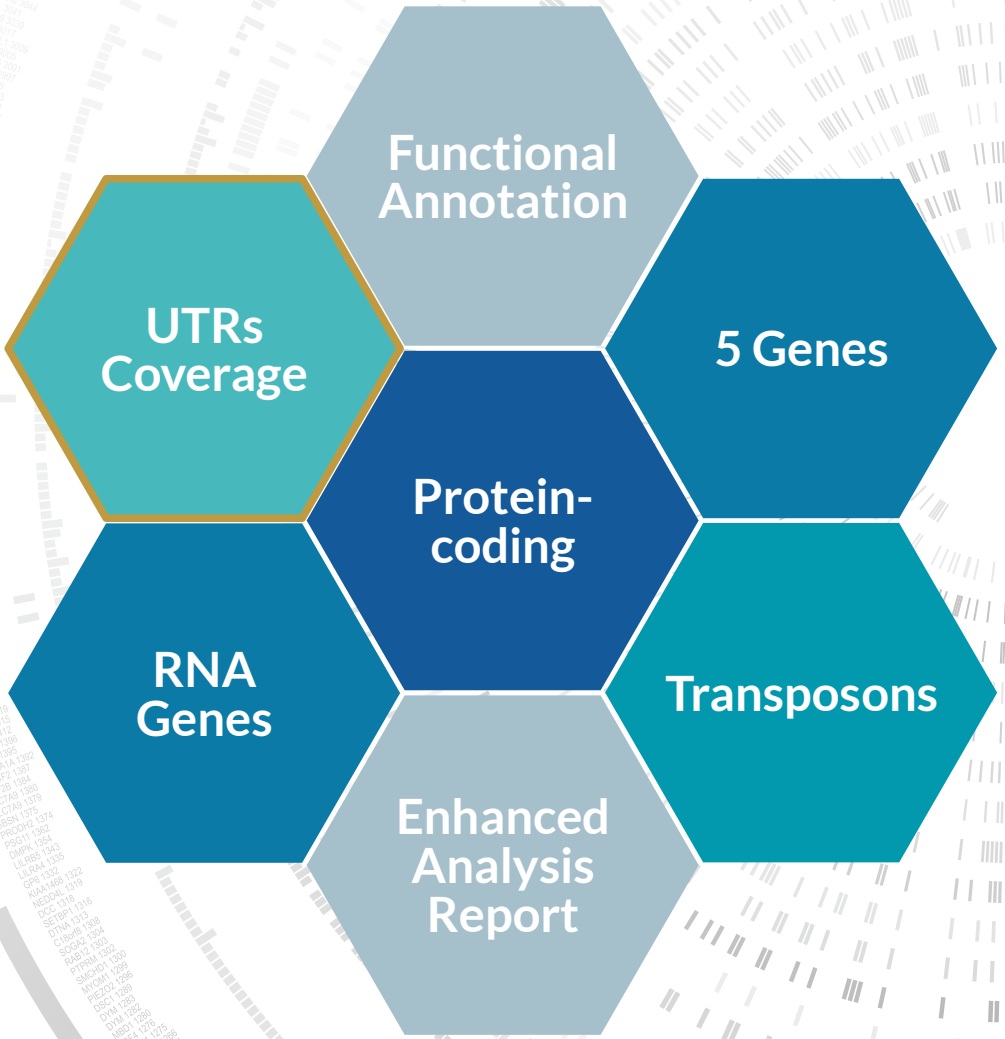
GOENOMICS

RNA-Seq data analysis

full-length and base-level coverage,
and annotation of UTR regions



RNA-Seq data analysis

full-length and base-level coverage, and annotation of UTR regions

Martin Kollmar, Dominic Simm
GOENOMICS GmbH

The use of transcriptome data for genome annotation offers several advantages. Transcriptome data can be used to predict protein-coding genes, transcriptome data provides evidence for annotated protein-coding genes, and transcriptome data is the only source for reconstructing the untranslated regions (UTR), i.e. the regions on either side of the coding sequence of a gene. Transcriptome data can be obtained by various methods, of which RNA-Seq and Isoseq are the most commonly used methods to support genome annotation. The length of reads generated by RNA-Seq depends on the sequencing technology used, but is usually between 50 and 300 base pairs. Isoseq is PacBio's RNA-Seq sequencing technology that can generate full-length transcript reads, but with a comparatively lower throughput, resulting in lower transcript coverage.

Coverage of gene structure features by RNA-Seq data

Coverage by RNA-Seq data can be defined in two ways: i) Full-length coverage, i.e. the support of the gene model from the beginning (5' end) to the end (3' end). In this case, the total length of the exons is defined as 100% and the coverage is the percentage of the mRNA covered by RNA-Seq data. ii) Base-level coverage, i.e. the support of each nucleotide by RNA-Seq reads. For simplicity, this coverage is usually averaged for each feature, CDS, exon and mRNA.

Approaches to determine coverage

To determine full-length coverage, the gene structure of a gene, including UTRs, must be known. Gene structures need to be confirmed by independent experiments, as high-throughput data are incomplete in terms of the percentage of genes covered and in terms of coverage of all gene features. As long as the prediction of gene structures is based on RNA-Seq data, full-length coverage is a self-fulfilling prophecy. Therefore, for the evaluation of genome annotations, it makes more sense to analyze the coverage at the base level. Base-level coverage in RNA-Seq data

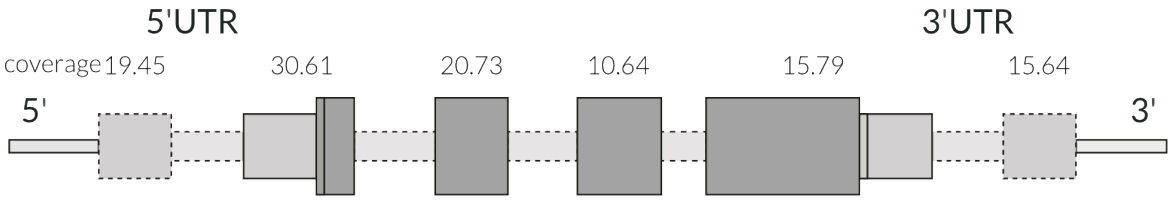


Figure 1: Gene structure with exons and introns from 5' to 3'. Small punctuated boxes represent introns removed during splicing. CDS features and non-coding exons are represented by large, dark-grey boxes and medium-sized boxes, respectively. UTR regions can consist of spliced and non-spliced exons.

Version 2024.11.11

is determined by mapping the sequencing reads to the genome and then quantifying the number of reads that overlap each nucleotide position. Modern software pipelines perform this process in several key steps: The first step is quality control of the reads, which is performed using tools such as FastQC or Trimmomatic to score and clean the reads by removing adapters, trimming low quality bases and filtering out poor quality reads to ensure high quality data for downstream analysis. The reads are then mapped to the genome using alignment tools such as HISAT2, STAR or Bowtie2. After alignment, the depth of coverage at each nucleotide position is calculated using tools such as Bedtools, Samtools or DeepTools.

mendle-analytics coverage annotation

RNA-Seq data can be provided by the customer or will be obtained from public databases. For mendle-analytics, we use STAR for the read alignment. Instead of the default parameters, we apply quite strict values to all settings that affect read mapping: i) Instead of the default ten mismatches, we only allow read alignments to be output if they contain less than five mismatches. ii) In contrast to the default, all reads with inconsistent and/or non-canonical introns are filtered out. iii) The maximum allowed intron length (default: 589,824 bp) is far above the longest intron lengths we have observed in genome annotation projects. The standard of the Encode project is even higher (1,000,000 bp). We calculate the maximum intron length based on the size of the genome assembly, resulting in about 800 bp for small yeast genomes and 200,000 bp for human-sized genomes. Similarly, the parameter for filtering the output of splice-junction reads depends on intron length. It is a function of the number of reads that support branching (maximum intron length supported by one read, maximum intron length supported by two reads, etc.), and we also calculate the respective maximum intron lengths per genome size.

The STAR BAM file is then entered into StringTie. StringTie examines the mapped reads and analyzes the alignment data to identify exons, splice junctions and the overall structure of the transcripts. First, overlapping reads are clustered to represent transcript fragments, which are then assembled into potential full-length transcripts. As part of the transcript assembly process, StringTie calculates base-level coverage by counting how many reads cover each nucleotide of the gene. It determines the depth per base for all positions along the transcript. This information is then used to calculate the coverage of exons, splice sites and full transcripts. Similar to STAR, we increase the rather weak default values to more restrictive values for coverage and abundance: i) The default minimum read coverage of one for transcripts is increased to two. ii) The default minimum isoform abundance of 0.01 is increased to 0.05.

The StringTie data is then merged into a single gff file using gffread. In this way, the source of the RNA-Seq data is lost, but since we only want to use the base-level coverage as evidence for our gene structure annotations, it is important that we have as many transcripts as possible for the mapping and that the transcripts are as long and complete as possible. The transcripts selected by gffread are not necessarily the ones with the highest coverage at the base level.

Important note: The coverage values at base level must not be used for differential expression analysis! They depend on the settings used for read mapping and are the result of merging multiple RNA-Seq datasets.

Annotation of UTR regions

The identification, prediction and annotation of untranslated regions (UTRs) of genes is crucial in genome annotation, as UTRs play an important role in the post-transcriptional regulation of gene expression. The current state of the art in UTR annotation involves the integration of multiple high-throughput data types and sophisticated computational tools. RNA-Seq is commonly used to identify 5' and 3' UTRs by mapping sequenced transcripts to the genome. However, standard short-read RNA-Seq does not always capture full-length transcripts, making it difficult to determine the exact boundaries of UTRs. Long-read sequencing technologies such as PacBio's Isoseq or Oxford Nanopore sequencing have become important tools for identifying full-length transcripts and provide more accurate UTR annotation by sequencing entire RNA

molecules, including UTRs, in a single read. 5'-RACE and 3'-RACE are experimental techniques used to determine the exact transcription start site (TSS) and polyadenylation site (PAS), respectively, especially in cases where RNA-Seq or long-read data cannot provide complete coverage. CAGE technology, Poly(A)-Seq and 3'-Seq are other powerful methods for precise mapping of TSS and PAS. AUGUSTUS and UTRannotator are computational tools for UTR prediction using RNA-Seq data for training, prediction and refinement of UTR regions.

mendle-analytics annotation of UTR regions

For UTR region annotation we use RNA-Seq and, if available, also Isoseq data. To avoid any bias by RNA-Seq mapping artifacts, we generate independent de novo transcriptome assemblies using the Trinity software. We use Trinity in genome-guided mode and provide the STAR BAM file for support. Trinity first assembles RNA-seq reads into the longest possible contiguous sequences (contigs) based on overlapping regions, capturing the most common sequences. These contigs are then clustered into sets of overlapping sequences that represent different isoforms or variations of the same gene. For each cluster a de Bruijn graph is built. Finally, the graphs are processed to reconstruct full-length transcripts, including alternative splicing isoforms and transcript variations. Trinity transcriptome assemblies and Isoseq data are mapped to the genome assemblies using GMAP. Similar to STAR and StringTie, we apply stricter values for the maximum length of internal introns and the total length of all introns. However, based on our experience with many genome annotations and manual annotation of thousands of genes, we allow the same maximum intron length for terminal introns where GMAP is restrictive. The GMAP data is finally merged with gffread. This data is used to add UTR regions to the coding regions of genes and to identify and annotate non-coding RNA genes.