illumina®

RNA-Seq Data Processing

An overview of the RNA sequencing pipeline in the BaseSpace® Correlation Engine

Introduction

The RNA sequencing (RNA-Seq) pipeline ("the pipeline") for the BaseSpace Correlation Engine processes sequencing data from mRNA to estimate transcript abundance and identify differentially expressed transcripts across samples. This technical note provides details of individual steps in the RNA-Seq workflow (Figure 1).

Extraction of Raw Sequences

The RNA-Seq pipeline supports the input of raw next-generation sequencing (NGS) data in the FASTQ format. Many published NGS studies in the Gene Expression Omnibus (GEO)[1] provide direct links to the raw sequence data stored at the Sequence Read Archive (SRA).[2] The sequence data from SRA normally requires decompression and, sometimes, proper splitting to generate the right FASTQ files. When the input FASTQ files are from private sources, it is expected that the sequences have been properly trimmed to remove adapter sequences and low-quality tails.

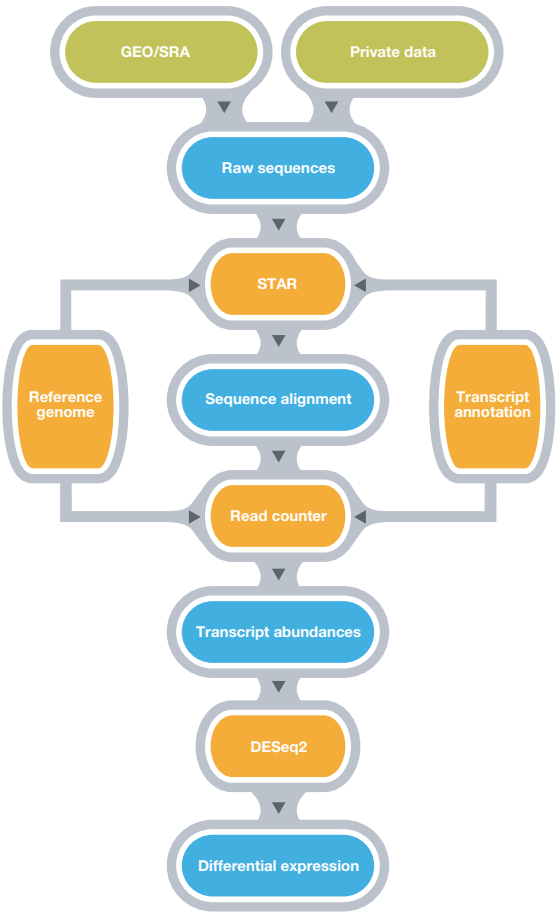

Figure 1: RNA-Seq Workflow

Sequence Alignment

Input NGS sequences are aligned against a reference genome using STAR2.3 and RefSeq annotations.[3] STAR aligns noncontiguous sequences directly to the genome and detects splice junctions in a single alignment pass without the need for a reference database of splice junctions. The alignment process of STAR involves 2 major steps:

1. Seed search

A sequential search for the Maximum Mappable Prefix (MMP) begins with the first base of a read until it cannot be mapped contiguously to the genome (for example, when a splice junction is encountered). The MMP search is repeated for the unmapped portions of the read until all contiguous portions have been successfully mapped. The search is performed in both forward and reverse directions of the read sequence to improve mapping sensitivity.

2. Clustering, stitching, and scoring

The mapped sequence seeds are merged by clustering around a selected set of anchor seeds within a defined genomic window and then stitched together. When available, the seeds from the mated paired-end reads are merged concurrently. If the entirety of a read is not found within one genomic window, STAR will try to find two or more windows that cover the entire read, resulting in a chimeric alignment.

The stitching is guided by a scoring scheme that applies scores and penalties for matches, mismatches, insertions, deletions, and junction gaps. The stitched combination with the highest score is reported as the alignment of the read. All the mapped sequences are merged and reported in the BAM format.

Transcript Abundance Estimation

After alignment, gene expression is estimated using an internally developed RNA read counter that is similar to htseq-count.[4] The number of aligned reads that overlap each gene in the annotations are counted. Reads are assigned to a gene if the read (or both reads in a pair) unambiguously map to the exons of one gene.

The read counts are used as input for differential expression analysis between test and control groups using R and DESeq2.[5] The base-mean read count, fold change, p-value, and q-value (Benjamini-Hochberg adjusted) are derived from this analysis. The median value of fragments per kilobase of transcript per million mapped reads (FPKM) per group are calculated separately based on normalized read counts, number of aligned reads and the full gene length.

Bioset Generation

The output of the differential expression analysis is consolidated into a single text file and converted into an expression bioset with the following fields:

- Transcript ID
- Fold Change
- Test Expression (FPKM)
- Control Expression (FPKM)
- P-Value
- Q-Value

References

1. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002;30(1):207–210.
2. Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res.* 2011;39:D19–D21. doi:10.1093/nar/gkq1019.
3. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15-21.
4. Anders S, Pyl PT, Hube, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics.* 2015;31(2):166-169.
5. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.

Illumina • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

illumina®