Pipeline overview

The ENCODE RNA-seq pipeline for long RNAs can be used if your libraries are generated from mRNAs (poly-A(+)), rRNA-depleted total RNA, or poly-A(-) RNA populations that are size-selected to be longer than approximately 200 bp. The pipeline takes as inputs both RNA-seq reads (from paired-end stranded or single end unstranded libraries) and a gene annotation file (by default GENCODE), and outputs several products:

- mapping of the reads to the genome creates an alignment file in bam file format
- mapping of the reads to the transcriptome creates a transcriptome alignment file in bam file format
- normalized RNA-seq signal for each strand (plus and minus) for unique reads and for unique and multimapping reads in bigwig file format
- gene quantifications (including the spike-ins quantifications) as a tsv file
- transcript quantifications (including the spike-ins quantifications) as a tsv file

The mapping of the reads is done using the STAR program and the quantification of genes and transcripts is done with the RSEM program. Although there is general agreement between the mappings and the gene quantifications produced by different RNA-seq pipelines, quantifications of individual transcript isoforms, being much more complex, can differ substantially depending on the processing pipeline employed, and are of unknown accuracy. Therefore, mapping and gene quantifications can be used confidently, while transcript quantifications should be used with care.

References

These pipelines require both assembly information for the species of interest and a gene reference. Each of the main programs, TopHat, STAR, and RSEM create an index for use in subsequent steps. The current reference files and indexes can be found on these links below:

- STAR Indexes: www.encodeproject.org/references/ENCSR314WMD/
- TopHat Indexes: www.encodeproject.org/references/ENCSR641UDW/
- RSEM Indexes: www.encodeproject.org/references/ENCSR219BJA/
- Unmodified Genome References and Chromosome Sizes: www.encodeproject.org/references/ENCSR425FOI/
- GENCODE References: www.encodeproject.org/references/ENCSR884DHJ/

Exogenous RNA spike-in controls

Exogeneous RNA spike-in controls are added to samples to create a standard baseline for the quantification of RNA expression (PMC3166838). The ENCODE consortium is standardizing on the use of the Ambion Mix 1 (ThermoFisher: 4456740) commercially available spike-ins at a dilution of ~2% of final mapped reads. However, there is a mixture of older data and imported data. Therefore, to track the spike-ins used in a given library, there is a dataset associated with the

library. That dataset will contain the spike-ins sequence file in fasta format and information on the concentrations. These spike-in sequences are expected to be found in the genome index used in the mapping step(s) and in the subsequently generated bam. The quantifications of the sequences can be found in the RSEM transcript and gene quantification.

Spike-in datasets:
www.encodeproject.org/search/?searchTerm=spike&type=Reference&lab.title=Barbara+Wold%2C+Caltech&lab.title=Brenton+Graveley%2C+UConn&lab.title=Thomas+Gingeras%2C+CSHL

Certificate of analysis for ERCC spike-ins:
www-s.nist.gov/srmors/view_cert.cfm?srm=2374

ERCC dash board: www.nist.gov/programs-projects/erccdashboard