Pipeline overview

The ENCODE RNA-seq pipeline for small RNAs can be used if your libraries are generated from rRNA-depleted total RNA libraries that are size-selected to be shorter than approximately 200 nucleotides. The pipeline takes as inputs both RNA-seq reads from single end stranded 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
- normalized RNA-seq signal for each strand (plus and minus) for unique reads and for unique and multimapping reads in bigwig file format
- raw gene quantifications as a tsv file (STAR output). The four columns of the file are as follows:
    - column 1: gene ID
    - column 2: counts for unstranded RNA-seq
    - column 3: counts for the 1st read strand aligned with RNA
    - column 4: counts for the second read strand aligned with RNA

References

These pipelines require both assembly information for the species of interest and a gene reference. The current reference files and indexes can be found on these links below:

- STAR Indexes: https://www.encodeproject.org/references/ENCSR999FPQ/
- Umodified Genome References and Chromosome Sizes: https://www.encodeproject.org/references/ENCSR425FOI/