Scripps Research

Center for Computational Biology and Bioinformatics (CCBB)
2023

RNA-seq data is valuable as it allows the measure of RNA expression levels as a transcriptional readout and the study of RNA structures in order to understand how RNA-based mechanisms impact gene regulation and thus disease and phenotypic variation (https://www.encodeproject.org/rna-seq/)

rRNA databases: https://github.com/biocore/sortmerna/tree/master/data/rRNA_databases
Reference build: https://support.illumina.com/sequencing/sequencing_software/igenome.html

Steps in data processing:

Assess data quality

Software/module	fastqc
Input	*.fastq.gz files
Output	websummary.html

Adapter and quality trimming of reads

Software/module	TrimGalore!
Input	*.fastq.gz files
Output	Trimmed *.fastq.gz files

Removal of ribosomal RNA

Software/module	SortMeRNA
Input	Trimmed *.fastq.gz files; rRNA databases
Output	Ribosomal RNA removed and trimmed *.fastq.gz files

Alignment to the genome

Software/module	STAR
Input	Ribosomal RNA removed and trimmed *.fastq.gz files; reference build
Output	*.bam files

Sort and index alignment

Software/module	SAMTools
Input	*.bam files
Output	Sorted *.bam files and *.bai files

Duplicate read marking

Software/module	Picard markDuplicates
Input	Sorted *.bam files and *.bai files
Output	*.markDups.bam files and *.markDups.bai files

Quality control

Software/module	MultiQC
Input	Output summaries from RSeQC, Qualimap, dupRadar, Preseq, edgeR
Output	websummary.html

Expression quantification

Software/module	featureCounts
Input	*.markDups.bam files and *.markDups.bai files; reference build
Output	*.featureCounts.txt files

Differential expression

Software/module	DESeq2
Input	*.featureCounts.txt files; sample metadata (names, groups, contrasts)
Output	* deseq2.results.txt files and *.deseq2.plots.pdf files