NCBI sequence read archive (SRA) data, available through multiple cloud providers. Through the largest publicly available datasets, RNA-sequencing (RNA-seq) data are provided. Here, we will present first the analysis using download open data of RNA-seq from fastq data and analyze them in Linux environment. At the end of this chapter, we will provide R commands to analyze the data matrix from RNA-seq and differentially express genes using DESeq2 library in R environment.

1. Set up the R studio working environment: R version 4.1.2 (2011-11-01).

R version 4.1.2 (2011-11-01)

2. Set up Linux environment: Linux version 3.10.0-1160.71.1.e17.x86\_64

3. In order to import NCBI SRA data in Linux, users need SRA IDs for interested data and download fastq file on File Type selection. In addition, Platform, Strategy, and Source of interested data have been shown on the website.

4. Gene expression matrix

To quantify abundances of transcripts from bulk and single-cell RNA-seq data, there have two popular alignment programs: Hisat2 and Kallisto for mapping next-generation sequencing reads. Here, this study presents steps for using Hisat2

- Create a genomic region index, construct an index for reference genome by using Hisat2-build.

hisat2-build <references file> <reference\_index\_file>

- Align the sequencing reads of samples in fastq format to the reference genome using default parameters of hisat2. At the same time, using pipes to sort the aligned reads by genomic position and save outputs in BAM format.

hisat-2 -q -x <reference\_index\_file> -1 <fastq> -S | samtools sort -s SAM -o BAM | samtools index -b BAM

The outputs of commands contain: BAM and BAI files of samples. Therefore, users could see the visual exploration of genomic data through Integrative genomics viewer (IGV) software.

- Count the number of aligned reads overlapping a region in genome, users could use bedtools command. Bedtools command is a powerful toolset for genome arithmetic, version v2.30.0.

bedtools multicov -bams <BAM files of samples> -bed < reference.bed> > count.tsv

The data matrix of RNA-seq counts is count.tsv. Users could load tsv files to R environment for reading and calculate differentially expression genes.