# RnaSeqSampleSizeData: Read counts and dispersion distribution from real data for sample size estimation of RNA-seq experiment

Shilin Zhao

November 18, 2014

## 1 Introduction

Sample size estimation is the most important issue in the design of RNA sequencing experiments. We developed a sample size estimation method based on the distributions of gene read counts and dispersions from real data. The uses can use their own prior data to do sample size estimation, and we also provide RnaSeqSampleSizeData package, which contains the read counts and dispersion distribution from some real datasets and can be used to do sample size estimation.

## 2 Data source

### 2.1 13 RNA-seq datasets from TCGA database

The read counts data of RNA-seq experiments for 13 different cancer types were extracted from the level 3 data of TCGA database with the following steps:

1. For each cancer type, downloading the .genes.results file for every sample from Open-Access HTTP Directory of Cancer Genome Atlas website `https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp`. For example, the files for all BRCA samples can be downloaded from `https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/brca/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/`;
2. For each cancer type, combining the raw_count column in each file into a single file, so that the columns represented samples with this cancer type, and the rows represented genes;
3. Loading the file to R and using *est_count_dispersion* function in RnaSeqSampleSize package to estimate the read counts and dispersion distribution for each cancer type;

As a result, the distribution data for 13 cancer types was packaged in RnaSeqSampleSizeData package and can be used with following names:

```
##  [1] "TCGA_BLCA" "TCGA_BRCA" "TCGA_CESC" "TCGA_COAD" "TCGA_HNSC" "TCGA_KIRC" "TCGA_LGG"
##  [8] "TCGA_LUAD" "TCGA_LUSC" "TCGA_PRAD" "TCGA_READ" "TCGA_THCA" "TCGA_UCEC"
```

# 3 Usage

Please refer to Section 3 (Estimation of sample size or power by prior real data) in vignette of RnaSeqSampleSize package to see how to estimate sample size or power with the datasets in RnaSeqSampleSizeData package.