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

Shilin Zhao*

December 3, 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 TCGA datasets in RnaSeqSampleSizeData were downloaded by *TCGA-Assembler* at Nov 25 2014. They can be downloaded and processed with the following steps:

2.1.1 Illustration of file URLs on TCGA database

The TCGA datasets can be downloaded manually from Open-Access HTTP Directory of Cancer Genome Atlas website https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp. For example, the lastes RNA-seq dataset for all BRCA samples can be found at https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/brca/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.1.10.0/. And the .rsem.genes.results files contain the raw counts data. The download link structure includes five parts

*zhaoshilin@gmail.com

and an illustration example can be found in the Supplementary Figure 1 from http://www.nature.com/nmeth/journal/v11/n6/extref/nmeth.2956-S1.pdf. Here is the explanation for the link of latest BRCA samples:

1. URL of the root directory including public TCGA data: https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/;
2. Cancer type: brca;
3. Institution that generated the data: unc.edu;
4. Assay platform: illuminahiseq_rnaseqv2;
5. Version number: 3.1.10.0;

2.1.2 Downloading datasets from TCGA database by TCGA-Assembler

The *TCGA-Assembler* is an open-source, freely available tool that automatically downloads, assembles and processes public The Cancer Genome Atlas (TCGA) data. The paper can be found at dx.doi.org/doi:10.1038/nmeth.295 and the software can be found at http://www.compgenome.org/TCGA-Assembler/. For example, if we need to download RNA-seq data for all READ samples, we can use the following R codes:

```
RNASeqRawData = DownloadRNASeqData(
  traverseResultFile = "./DirectoryTraverseResult_Nov-25-2014.rda",
  saveFolderName = ".", cancerType = "READ",assayPlatform = "RNASeqV2",
  dataType = "rsem.genes.results");
```

2.1.3 Processing downloaded data

1. After downloading the datasets, a file named by "cancerType"_"institution"_"platform"_rsem.genes.results_"date".txt will be generated. For the READ example, the file will be READ__unc.edu__illuminaga_rnaseqv2__rsem.genes.results__Nov-25-2014.txt. The columns represented samples with this cancer type, and the rows represented genes. For each sample, "raw_count" and "scaled_estimate" will be stored.
2. We will select all the cancer samples (barcode 01A), and the "raw_count" columns will be extracted and rounded to integer to make a new expression matrix file for further analysis.
3. *RnaSeqSampleSize* package will be loaded and the `est_count_dispersion` function will be used 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"
```

2.1.4 More details about the workflow of RNA-Seq quantification

The workflow of RNA-Seq quantification for TCGA data can be acessed at the datasets download link. For example, the workflow for BRCA can be found at https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/brca/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.1.10.0/DESCRIPTION.txt.

From the workflow, we can find quantification results of TCGA datasets were obtained using the RSEM method (Paper at http://www.biomedcentral.com/1471-2105/12/323. Software at http://deweylab.biostat.wisc.edu/rsem/). And the "raw_count" in the TCGA result was "expected_count" in RESM method, which is the sum of the posterior probability of each read comes from this gene over all reads. It is generally a non-integer value and we will round it to integer for further analysis.

3 Usage

Please refer to Section 3.2 (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.