DupChecker: a package for checking high-throughput genomic data redundancy in meta-analysis

Quanhu Sheng*, Yu Shyr, Xi Chen

Center for Quantitative Sciences, Vanderbilt University, Nashville, USA

*shengqh (at) gmail.com

May 5, 2014

Abstract

In this vignette, we demonstrate the application of DupChecker as a package for checking high-throughput genomic data redundancy in meta-analysis. DupChecker can download the GEO/ArrayExpress raw data files from EBI/ncbi ftp server, extract individual CEL files, calculate MD5 fingerprint for each CEL file and validate the redundancy of those CEL files.

Contents

1 Introduction

Meta-analysis has become a popular approach for high-throughput genomic data analysis because it often can significantly increase power to detect biological signals or patterns in datasets. However, when using public-available databases for meta-analysis, duplication of samples is an often encountered problem, especially for gene expression data. Not removing duplicates would make study results questionable. We developed a package Dupchecker that efficiently identifies duplicated samples by generating MD5 fingerprints for raw data.

2 Standard workflow

2.1 Quick start

Here we show the most basic steps for a validation procedure. You need to create a target directory used to store the GEO data. Here, we assume the target directory is your work directory.

```
library(DupChecker)
geoDownload(datasets = c("GSE14333", "GSE13067",
                         "GSE17538"), targetDir=getwd())
datafile<-buildFileTable(rootDir=getwd())
result<-validateFile(datafile)
if(result$hasdup){
  duptable<-result$duptable
  write.csv(duptable, file="duptable.csv")
}
```

2.2 GEO/ArrayExpress data download

Firstly, function geoDownload/arrayExpressDownload will download raw data from ncbi/EBI ftp server based on datasets user provided. Once the compressed raw data is downloaded, CEL files will be extracted from compressed raw data.

If the download or decompress cost too much time in R environment, user may download the GEO/ArrayExpress raw data and decompress the data to individual CEL files using other tools. The reason that we expect the CEL file not compressed CEL file is the compressed files from same CEL file but by different compress softwares may have different MD5 fingerprint.

The following code will download two datasets from ArrayExpress system and three datasets from GEO system. It may cost a few minutes to a few hours based your network performance.

```
library(DupChecker)

#download from ArrayExpress system
datatable<-arrayExpress(datasets = c("E-TABM-158",
                        "E-TABM-43"), targetDir=getwd()))
datatable

#Or download from GEO system
datatable<-geoDownload(datasets = c("GSE14333", "GSE13067",
                       "GSE17538"), targetDir=getwd())
datatable
```

The datatable is a data frame containing dataset name and how many CEL files in that dataset.

2.3 Build file table

Secondly, function buildFileTable will try to find all files in the subdirectories under root directories user provided. The result data frame contains two columns, dataset and filename. Here, rootDir can also be an array of directories.

```
datafile <- buildFileTable(rootDir = getwd())
datafile
```

2.4 Validate CEL file redundancy

The function validateFile will calculate MD5 fingerprint for each file in table and then check to see if any two files have same MD5 fingerprint. The files

with same fingerprint will be treated as duplication. The function will return a table contains all duplicated files and datasets.

```
result <- validateFile(datafile)
if (result$hasdup) {
    duptable <- result$duptable
    write.csv(duptable, file = "duptable.csv")
}
```

3 Discussion

Althrough DupChecker package is purposed for GEO/ArrayExpress data redundancy validation, it can also be used for other file redundancy validation.

4 Session Info

- R version 3.1.0 (2014-04-10), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=English_United States.1252`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: knitr 1.5
- Loaded via a namespace (and not attached): evaluate 0.5.5, formatR 0.10, highr 0.3, stringr 0.6.2, tools 3.1.0