Icahn School of Medicine at Mount Sinai LINCS Center for Drug Toxicity Signatures

Standard Operating Procedure: Identification of Differentially Expressed Genes

DToxS SOP Index: CO-4.1
Last Revision: March 22, 2016
This version is a revision of CO-4.0. It contains changes to steps 5 and 7.
Written By: Yuguang Xiong and Eric Sobie

Approvals (Date): Joseph Goldfarb (3/22/16)
Marc Birtwistle (3/30/16)
Eric Sobie (3/22/16)
Ravi Iyengar (3/30/16)

Quality Assurance/Control (QA/QC) steps are indicated with green highlight.

Metadata recording is highlighted with yellow highlight and superscript indices.

---

1) Install the statistical software *R*[1] in either the Linux, Unix, or Mac OS X operating systems. We have used R version 3.2.2.
2) Install two R packages that are required for data analysis: *edgeR*[2] and *matrixStats*[3], using the following command in R:
   `install.packages(c("edgeR", "matrixStats"))`
   After installing the packages, use the following commands to check the package versions:
   `packageVersion("edgeR")`
   `packageVersion("matrixStats")`
   Above commands give the versions of edgeR and matrixStats packages. In general the most recent versions of the packages work well for the analysis (We have used edgeR version 3.10.5 and matrixStats version 0.50.1 QA/QC1).
   We have used edgeR version 3.10.5 and matrixStats version 0.50.1.
3) Download (upon request at DToxS website: see e-mails below under Correspondence) a gzip-compressed DEG dataset file `[deg-dataset-file]` of a particular plate that includes:
   a) The UMI read counts data files generated by the alignment of RNA sequencing dataset from the plate, *RNAseq_[date].unq.refseq.umi.dat*, as described in the *Step 13-b-i* of the **DToxS SOP CO-3.0 Generation of Gene Read Counts**.
   b) The experimental design file that describes the experimental settings of all the samples in the plate.
   c) A set of custom-built R programs for comparing the difference of gene expression levels based on DToxS mRNA sequencing datasets.

d) The parameter files for the drug groups and outlier-removal thresholds defined specifically for the dataset being analyzed:
   i) Drug groups are the groups to categorize each drug based on its toxicity and mitigation effects.
   ii) Outlier-removal thresholds are the cutoff values based on Pearson's correlation coefficient between replicate samples, such that a sample is removed if the closest distance between this sample and the cluster of the rest samples is greater than the cutoff value.

e) The configuration file for controlling various aspects of the analysis process by the R programs stated in *Step 3-c*.

4) Go to the directory containing the downloaded compressed DEG dataset file `[deg-dataset-file]` and extract it using the following command at a system shell terminal:
`tar –xzf [download-dir]/[deg-dataset-file]`
where:
`[download-dir]` is the directory folder where the downloaded `[deg-dataset-file]` is saved.
This command creates a new directory `[deg-dataset-dir]`, e.g. `LINCS.Dataset.Gene.LINCS.20150409.R20160310`, that contains all needed datasets and programs.

5) Go to the DEG dataset directory `[deg-dataset-dir]` generated at *Step 4*, and run the script `Setup-DEG-Env.sh` using the following command:
`[prog-dir]/Setup-DEG-Env.sh [top-deg-dir]`
where:
a) `[prog-dir]` is the program directory `Programs` under `[deg-dataset-dir]`.
b) `[top-deg-dir]` is the destination top directory for subsequent DEG analysis, e.g. `~/LINCS/RNAseq_20150409`.
This command creates a directory tree under `[top-deg-dir]` that contains all the data files and programs for DEG analysis

6) Go to the `Counts` directory under `[top-deg-dir]` that contains the UMI read counts data files and experimental design files, and execute the R program *Extract-Gene-Expression-Samples.R* included in the `Programs` directory using the following command:
`Rscript [prog-dir]/Extract-Gene-Expression-Samples.R [counts-dir]`
where:
a) `[prog-dir]` is the program directory `Programs` under `[top-deg-dir]`.
b) `[counts-dir]` is the `Counts` directory under `[top-deg-dir]`.
This command extracts the appropriate data for analysis.

7) Go to `[top-deg-dir]` and run the R program *Compare-Molecule-Expression.R* using the following command:
`Rscript [prog-dir]/Compare-Molecule-Expression.R [config-dir]/Configs.LINCS.Dataset.Gene.[type].[date].tsv [top-deg-dir] >& [top-deg-dir]/LINCS.Dataset.Gene.[type].[date].log`
where:
a) `[prog-dir]` is the program directory `Programs` under `[top-deg-dir]`.
b) `[config-dir]` is the configuration directory `Configs` under `[top-deg-dir]`.
c) `Configs.LINCS.Dataset.Gene.[type].[date].tsv` is a customizable configuration file included in the `Configs` directory under `[top-deg-dir]`, where `[type]` is the dataset type and `[date]` is the date tag of the dataset.
d) `LINCS.Dataset.Gene.[type].[date].log` is the log file that will be generated by *Compare-Molecule-Expression.R*. This file contains all the screen outputs from the program.

Note: this step usually takes a few minutes to finish.

8) When *Step 7* finishes, check the `Results` directory under `[top-deg-dir]` for the following data files generated by the program:
   a) Level-1 data:
      i) A single table of UMI read counts of all genes in all samples at all conditions: *ReadCounts-Merged.tsv*, which is used as the inputs of the differential comparison analysis.
      ii) Multiple tables of UMI read counts of all genes in all samples at individual conditions: *Human.[A/B/D/E]-Hour.48-Plate.[1/2/3/4]-ReadCounts-[Drug/Control].tsv*
   b) Level-2 data:
      i) Tables of normalized gene read counts in selected samples (by removing outlier samples with small correlation coefficients QA/QC2) for selected genes (by removing the genes with small read counts QA/QC3) at each pair of compared conditions: *Human.[A/B/D/E]-Hour.48-Plate.[1/2/3/4]-ReadCounts-Norm-[Control].[Drug].tsv*
   c) Level-3 data:
      i) List of top-40 genes based on their p-values for each pair of compared conditions: *Human.[A/B/D/E]-Hour.48-TOP.40-[Control].[Drug].tsv*
   d) Metadata:
      i) Tables of the metadata associated with each sample investigated under particular conditions in the plate: *Human.[A/B/D/E]-Hour.48-ExpConfigs-[Drug/Control].tsv*

Metadata

1. **R:**
   version 3.2.2,
   http://www.r-project.org.
2. **edgeR:**
   version 3.10.5,
   http://bioconductor.org/packages/release/bioc/html/edgeR.html.
3. **matrixStats:**
   version 0.50.1,
   http://cran.r-project.org/web/packages/matrixStats/index.html.

Quality Assurance/Control Steps (QA/QC)

QA/QC1: Verification that appropriate statistical packages have been correctly installed.

QA/QC2: Removal of outlier samples based on Pearson's correlation coefficients.

A divisive hierarchical clustering method using a distance metric based on Pearson's correlation coefficient is applied to the replicate samples of each control and drug-treated condition. Replicate samples will be removed from a condition if the coefficient is lower than a predefined threshold (determined by the DToxS experimental team).

QA/QC3: Removal of the genes with small read counts

In order to get rid of the influence of a large number of genes with very small read counts, a gene will be removed from all samples if the number of its CPM (count per million reads) counts that are greater than 1 is less than the minimum number of its replicates across all control and drug-treated conditions.

Correspondence

Yuguang Xiong
E-mail: yuguang.xiong@mssm.edu

Evren Azeloglu
E-mail: evren.azeloglu@mssm.edu