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

Standard Operating Procedure: Identification of Differentially Expressed Proteins

DToxS SOP Index: CO-2.0
Last Revision: March 30, 2016
Written By: Yuguang Xiong and Eric Sobie

Approvals (Date): Joseph Goldfarb (03/31/16)
Marc Birtwistle (04/01/16)
Eric Sobie (03/31/16)
Ravi Iyengar (04/27/16)

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 DEP read counts dataset file `[dep-dataset-file]` from a particular proteomic mass spectrometric experiment that includes:
   a) The read counts data files generated by the alignment of proteomic sequencing data, *PRTseq-Read-Counts.LINCS.[date].[number].tsv*, as provided by our collaborators (see DToXS SOP CO-1.0)
   b) A set of custom-built R programs for comparing the difference of protein expression levels based on DToxS proteomic read counts datasets.
   c) 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.

d) 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 DEP dataset file `[dep-dataset-file]` and extract it using the following command at a system shell terminal:
`tar –xzf [download-dir]/[dep-dataset-file]`
where:
`[download-dir]` is the directory folder where the downloaded `[dep-dataset-file]` is saved.
This command creates a new directory `[dep-dataset-dir]`, e.g. `LINCS.Dataset.Protein.LINCS.20151015.R20160310`, that contains all needed datasets and programs.

5) Go to the DEP dataset directory `[dep-dataset-dir]` generated at *Step 4*, and run the script `Setup-DEP-Env.sh` using the following command:
`[prog-dir]/Setup-DEP-Env.sh [top-dep-dir]`
where:
a) `[prog-dir]` is the program directory `Programs` under `[dep-dataset-dir]`.
b) `[top-dep-dir]` is the destination top directory for subsequent DEP analysis, e.g. `~/LINCS/PRTseq_20151015`.
This command creates a directory tree under `[top-dep-dir]` that contains all the data files and programs for DEP analysis

6) Go to `[top-dep-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.Protein.[type].[date].tsv [top-dep-dir] >& [top-dep-dir]/LINCS.Dataset.Protein.[type].[date].log`
where:
a) `[prog-dir]` is the program directory `Programs` under `[top-dep-dir]`.
b) `[config-dir]` is the configuration directory `Configs` under `[top-dep-dir]`.
c) `Configs.LINCS.Dataset.Protein.[type].[date].tsv` is a customizable configuration file included in the `Configs` directory under `[top-dep-dir]`, where `[type]` is the dataset type and `[date]` is the date tag of the dataset.
d) `LINCS.Dataset.Protein.[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.

7) When *Step 6* finishes, check the `Results` directory under `[top-dep-dir]` for the following data files generated by the program:
a) Level-1 data:
i) A single table of read counts of all proteins in all samples at all conditions: *ReadCounts-Merged.tsv*, which is used as the inputs of the differential comparison analysis.
ii) Multiple tables of read counts of all proteins in all samples at individual conditions: *Human.[A/B/D/E]-Hour.48-Plate.0-ReadCounts-[Drug/Control].tsv*
b) Level-2 data:

i) Tables of normalized protein read counts in selected samples (by removing outlier samples with small correlation coefficients QA/QC2) at each pair of compared conditions:
Human.[A/B/D/E]-Hour.48-Plate.0-ReadCounts-Norm-[Control].[Drug].tsv

c) Level-3 data:
   i) List of top-40 proteins 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).

Correspondence

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

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