NOV 30, 2018

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.s24eggw

Protocol Citation: Andreas Schüttler 2018. Select, load, annotate, normalize, and process toxicogenomic raw data from GEO and ArrayExpress. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.s24eggw>

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: In development
 We are still developing and optimizing this protocol

Created: Aug 28, 2018

Last Modified: Nov 30, 2018

PROTOCOL integer ID:
 15164

Select, load, annotate, normalize, and process toxicogenomic raw data from GEO and ArrayExpress

Andreas Schüttler¹

¹Helmholtz Centre for Environmental Research - UFZ

Andreas Schüttler

ABSTRACT

Gene expression databases like Gene Expression Omnibus or ArrayExpress by now contain a wealth of toxicogenomic datasets. This offers great possibilities for advanced data analyses, like meta-analyses, co-expression studies, etc.

However, automated retrieval of this data is still a challenge.

With this computational pipeline we retrieve toxicogenomic data of the *Danio rerio* (zebrafish) embryo from GEO and ArrayExpress.

To make the data as comparable as possible, we download the raw datasets, and re-map the probes or probesets to the most recent version.

In the end a matrix with logFC in response to different chemical treatments is compiled.

BEFORE START INSTRUCTIONS

The pipeline makes use of the custom R-package "toxprofileR".

This package is accessible via <https://git.ufz.de/schuettl/toxprofileR>

Select data from Gene expression databases

1. GEO: The first step is to retrieve metadata from Gene Expression Omnibus. This is achieved with the help of the R-package 'GEOmetadb'.
From the metadata and from manually curated information, datasets are selected and list for downloading data are created.

Command

```
rm(list = ls())  
# load libraries -----  
library(
```

2. ArrayExpress: The same as for GEO is done for ArrayExpress.

Command

```
rm(list = ls())  
# load libraries -----  
library(
```

Download Array and Platform data

- 3 Next step is to download the array and platform data selected in step 1.

We create a "data" directory where all data is downloaded to.

Command

```
#!/bin/bash  
# download GEO data  
cat ./data/download_lists/ftp_download_list_geo.txt | parallel --gnu
```

Probe mapping

- 4 Since microarrays are designed for different genome version, it is necessary to re-map the probes to the recent genome version (here dRer11).

For probe mapping, as a first step, fasta-files have to be created from the downloaded platform-files from GEO and ArrayExpress.

Command

```
rm(list = ls())  
# load platform information geo-----  
load(
```

- 5 Perform Blat

Command

```
#!/bin/bash  
skriptdir=$(pwd)  
cd ./data/PlatformData/fastas/  
find $directory -type f -name
```

- 6 map with gene annotation

Command

```
rm(list = ls())  
# load libraries -----  
library(
```

Create target file and table of comparisons

- 7 Before loading the data, we compile a targets data frame and a table with all comparisons for logFC calculation

Command

```
rm(list = ls())  
# combine sample metadata from databases with manual annotation -----  
## Array Express -----  
# manual annotation  
zfe_tox_ae_cure <- read.csv(
```

Read raw data

- 8 Based on the R-packages "limma" and "oligo", data is loaded into R.

Command

```
rm(list = ls())  
library(
```

Create logFC matrix

- 9 Last but not least a logFC matrix is created from all normalized data.

Command

```
rm(list = ls())  
library(
```