Nov 30, 2018

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

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

Andreas Schüttler[1]

[1]Helmholtz Centre for Environmental Research - UFZ

Andreas Schüttler

OPEN ACCESS

DOI: https://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

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

Created: August 28, 2018

Last Modified: November 30, 2018

Protocol Integer ID: 15164

Keywords: toxicogenomic dataset, toxicogenomic raw data, wealth of toxicogenomic dataset, arrayexpress gene expression database, toxicogenomic data, process toxicogenomic raw data from geo, arrayexpress gene expression databases like gene expression omnibus, toxicogenomic data of the danio rerio, gene expression omnibus, gene expression, raw dataset, embryo from geo, gene, offers great possibilities for advanced data analysis, advanced data analysis, embryo, retrieval of this data, automated retrieval

Abstract

Gene expression databases like Gene Expression Omnibus or ArrayExpress by now contain a wealth of toxicogenomic datasets. This is a 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.

Troubleshooting

Before start

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 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 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/fasta/
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(
```