Creating reference datasets:
The Broad Connectivity Map (v1)

Thomas Sandmann

October 30, 2017

Contents

1 Introduction **2**

2 Analyzing the Broad Connectivity map (v.1) data **2**

1 Introduction

Public repositories, such as ArrayExpress or GEO provide access to many published expression profiling datasets, featuring perturbations in many different organisms, model systems and conditions.

The Atlas search engine offers a simple way to identify perturbation experiments of interest in the ArrayExpress repository.

This vignette shows how to obtain and process raw microarray data from a large-scale drug perturbation study performed in human cells, the `Connectivity Map` dataset (version 1), released by Lamb and co-workers in 2006. Similar workflows can be used to download an process many other publically available datasets.

In this study, the researchers treated multiple human cell lines with 164 distinct small molecules or matched controls. In total, 564 samples were generated, RNA extracted, labeled and hybridized either to A-AFFY-113 Affymetrix (HT_HG-U133A) or A-AFFY-33 Affymetrix (HG-U133A) microarrays.

The raw data for this study is available from ArrayExpress under accession `E-GEOD-5258` . The raw .cel files and the array annotations can be downloaded and compiled into a suitable `eSet` objects using the ArrayExpress Bioconductor package. Alternatively, the final RData object can be downloaded directly from ArrayExpress.

Please note that this is a large dataset and executing the following code will download more than 700 MB of data.

2 Analyzing the Broad Connectivity map (v.1) data

Data download and normalization

A call to the `ArrayExpress` function will retrieve the raw data for study E-GEOD-6907 from ArrayExpress. (As this is a large dataset, this might take while...)

```
> library(ArrayExpress)
> GEOD5258.batch <- ArrayExpress("E-GEOD-6907")
```

As this experiment was performed on two different array platforms, a list with two affyBatch objects is returned, one for each array platform.

We normalize each object separately using the `rma` function from the affy package.

```
> library( affy )
> length( GEOD5258.batch )
> GEOD5258.eSets <- lapply( GEOD5258.batch, rma )
```

The `mapNmerge` function from the gCMAP package averages the expression values different probes for the same gene by mapping them to Entrez ids. Alternatively, the `nsFilter` function from the genefilter package could be used.

```
> GEOD5258.eSets <- lapply( GEOD5258.eSets, mapNmerge)
```

Now that we have mapped the expression values to Entrez Ids, we can combine the two ExpressionSets into one

```
> GEOD5258.eSet <- mergeCMAPs( GEOD5258.eSets[[1]], GEOD5258.eSets[[2]] )
```

Defining perturbation experiments and performing differential expression analysis

The ArrayExpress dataset is associated with extensive sample annotation information, available in the `phenoData` slot of the `ExpressionSet`. Experimental factors are marked with the `Factor` prefix in the column name.

```
> head( pData(GEOD5258.eSet ))
> conditions <- grep("^Factor", varLabels( GEOD5258.eSet ), value=TRUE)
> conditions
```

In this case, we are interested in studying the effect of the different compounds, which are specified in the column of the `phenoData` slot. Controls are annotated with the Compound level `none` .

```
> unique( pData(  GEOD5258.eSet )$Factor.Value..Compound.)
```

To associate drug perturbation with their matched controls, we require that control experiments must have been performed in the same `CellLine` and with the same `Vehicle` . With this information, the `splitPerturbations` function from the gCMAP package can group treatment and perurbation samples into individual experiments of interest. Each of these experimental instances is returned in a separate `ExpressionSet`, grouped in the GEOD5258.list list.

```
> GEOD5258.list <- splitPerturbations( GEOD5258.eSet,
                       factor.of.interest="Compound",
                       control="none",
                       controlled.factors=c("CellLine", "Vehicle", "Time")
)
```

To track the experimental conditions assayed in each perturbation experiment, the first line (containing the perturbation) is extracted from each `phenoData` slot and deposited in a data.frame with one row for each perturbation / `ExpressionSet` in GEOD5258.list.

```
> anno <- t(sapply( GEOD5258.list, function(x) pData(x)[1,conditions]))
> anno <- apply( anno, 2, unlist)
> anno <- data.frame( anno )
> colnames( anno ) <- c("CellLine", "Vehicle", "Compound", "Time", "Dose")
```

The `generate_gCMAP_NChannelSet` function performs differential expression analysis (using limma) separately for each `ExpressionSet` in the list. It returns an `NChannelSet` object containing the log2 fold change, raw p-values and z-scores for all experiments.

```
> GEOD5258.ref <- generate_gCMAP_NChannelSet( GEOD5258.list,
                       uids=1:length( GEOD5258.list ),
                       sample.annotation=anno)
> pData(  GEOD5258.ref)[10:15,]
```

This object, containing the differential expression results for 12701 genes from 214 different perturbation experiments and sample-level annotations in its `phenoData` slot, is now ready to be used as a reference dataset by gCMAPWeb.

Inducing gene sets

If required, we can apply a threshold to one channel of the `NChannelSet` and define sets of differentially up- and down-regulated genes. For example, the following command applies a z-score cutoff of >3 or <-3 to each experiment and stores the results in a sparse-matrix within a `CMAPCollection`.

```
> GEOD5258.sets <- induceCMAPCollection( GEOD5258.ref, element="z", higher=3, lower=-3 )
> head( setSizes( GEOD5258.sets ) )
```

```
> sessionInfo()

R version 3.4.2 (2017-09-28)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.3 LTS

Matrix products: default
BLAS: /home/biocbuild/bbs-3.6-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.6-bioc/R/lib/libRlapack.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats4    parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] gCMAPWeb_1.18.0      Rook_1.1-1           gCMAP_1.22.0
 [4] limma_3.34.0         GSEABase_1.40.0      graph_1.56.0
 [7] annotate_1.56.0      XML_3.98-1.9         AnnotationDbi_1.40.0
[10] IRanges_2.12.0       S4Vectors_0.16.0     Biobase_2.38.0
[13] BiocGenerics_0.24.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.13       compiler_3.4.2     RColorBrewer_1.1-2 bitops_1.0-6
 [5] tools_3.4.2        digest_0.6.12      bit_1.1-12         RSQLite_2.0
 [9] memoise_1.1.0      tibble_1.3.4       lattice_0.20-35    rlang_0.1.2
[13] Matrix_1.2-11      DBI_0.7            Category_2.44.0    yaml_2.1.14
[17] GSEAlm_1.38.0      DESeq_1.30.0       hwriter_1.3.2      genefilter_1.60.0
[21] bit64_0.9-7        grid_3.4.2         RBGL_1.54.0        survival_2.41-3
[25] geneplotter_1.56.0 blob_1.1.0         splines_3.4.2      xtable_1.8-2
[29] brew_1.0-6         RCurl_1.95-4.8
```