Intro to the *HumanAffyData* experimental data package

Brad Nelms

October 19, 2016

Contents

1 Introduction 1
2 Dataset overview 1
3 Citation 4

1 Introduction

*HumanAffyData* is a re-analysis of human gene expression data generated on the Affymetrix HG_U133PlusV2 (EH176) and Affymetrix HG_U133A (EH177) platforms, provide as ExpressionSet objects. The original data were normalized using robust multiarray averaging (RMA) to obtain an integrated gene expression atlas across diverse biological sample types and conditions. The entire compendia comprise 9395 arrays for EH176 and 5372 arrays for EH177. It is intended to be used as a starting point for gene co-expression analysis, or as a resource to quickly examine where a gene is expressed from within the R environment.

EH176: the original data were gathered by [1] and normalized using robust multiarray averaging (RMA). The phenoData of the ExpressionSet object contains the title and description of the source entries on GEO.

EH177: the original data were gathered by [2] and normalized using robust multiarray averaging (RMA). [2] manually curated the dataset to establish uniform phenotypic information for each sample, which is available in the phenoData of the ExpressionSet object. This data is accesible on ArrayExpress under accession E-MTAB-62. The RMA-normalized expression values were then adjusted to reduce the influence of technical bias (i.e. variation in hybridization conditions or starting material) using the R package *bias 0.0.3* [3]. Finally, probesets were mapped to Entrez gene identifiers using the *Bioconductor* annotation package *hgu133a.db*, and values for probesets mapping to the same gene were averaged to produce a single expression measurement for each gene.

2 Dataset overview

First, access the HumanAffyData from ExperimentHub:

```r
> library(ExperimentHub)
> hub <- ExperimentHub()
> x <- query(hub, "HumanAffyData")
> x
```

ExperimentHub with 2 records
# snapshotDate(): 2016-08-08
# $dataprovier: ArrayExpress, GEO
Intro to the `HumanAffyData` experimental data package

# $species: Homo sapiens
# $rdataclass: ExpressionSet
# additional mcols(): taxonomyid, genome, description, coordinate.1_based,
# maintainer, rdatadateadded, preparerclass, tags, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["EH176"]]

<table>
<thead>
<tr>
<th>title</th>
</tr>
</thead>
<tbody>
<tr>
<td>EH176</td>
</tr>
<tr>
<td>EH177</td>
</tr>
</tbody>
</table>

data can then be extracted using:

```r
E.MTAB.62 <- x[["EH177"]]
```

This downloads the EH177 dataset, which contains an `ExpressionSet` object containing expression data from ArrayExpress accession E-MTAB-62:

```r
E.MTAB.62
```

ExpressionSet (storageMode: lockedEnvironment)

assayData: 12496 features, 5372 samples

- element names: exprs
- protocolData: none
- phenoData
  - sampleNames: GSM23227.CEL 1229968152.CEL ... 676426699.CEL (5372 total)
  - varLabels: OperatorVariation DataSource ... ArrayDataFile (16 total)
  - varMetadata: labelDescription
- featureData: none
- experimentData: use 'experimentData(object)'

Annotation: hgu133a

The experiment data can be extracted using the `exprs` function:

```r
data <- exprs(E.MTAB.62)
```

```r
> dim(data)
[1] 12496  5372
```

```r
> data[1:5,1:5]
   GSM23227.CEL 1229968152.CEL GSM133626.CEL GSM47465.CEL GSM124909.CEL
5982    8.055513     7.431500     8.222138     7.757324     7.660949
2978    5.460372     5.363735     5.454068     5.496320     5.272762
7318    6.293562     7.422237     7.540636     7.433086     6.893468
```

This results in a matrix of expression data with the column names indicating the Array Data File name of each sample, and the rownames providing the human Entrez IDs of each gene.

Similarly, the phenotype data can be extracted using the `pData` function:

```r
pDat <- pData(E.MTAB.62)
```

```r
> print(summary(pDat))
```

<table>
<thead>
<tr>
<th>OperatorVariation</th>
<th>DataSource</th>
<th>Groups.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Justin,,Lamb</td>
<td>324</td>
<td>GSE5258 : 324</td>
</tr>
<tr>
<td>Milton,,Taylor</td>
<td>308</td>
<td>GSE7123 : 308</td>
</tr>
<tr>
<td>Roel,,Verhaak</td>
<td>284</td>
<td>GSE1159 : 284</td>
</tr>
<tr>
<td>Benjamin,,Haibe-Kains</td>
<td>273</td>
<td>GSE4475 : 213</td>
</tr>
<tr>
<td>Michael,,Hummel</td>
<td>213</td>
<td>E-AFMX-6: 195</td>
</tr>
</tbody>
</table>

Similarly, the phenotype data can be extracted using the `pData` function:

```r
pDat <- pData(E.MTAB.62)
```
The phenotypic data contains several “meta groups”, labeled as “Groups_4”, “Groups_15”, and “Groups_369”. These are curated labels that group samples from a particular tissue, cell line, disease status, etc. The meta groups are explained further in [2]. [2] also discuss a “96 meta group” category, which is simply any members of the “369 meta groups” that contain at least 10 samples. The “96 meta groups” category can be re-created from the phenotypic data as follows:

```r
> Groups_96 <- as.character(pDat$Groups_369)
> Groups_96[Groups_96 %in% names(which(table(pDat$Groups_96) < 10))] <- ''
> pDat$Groups_96 <- as.factor(Groups_96)
```
3 Citation

> citation("HumanAffyData")

Please cite Engreitz, et al. (2010) for the EH176 dataset and Lukk, et al. (2010) for the EH177 dataset:


References

