Intro to the *HumanAffyData* experimental data package

Brad Nelms

October 31, 2024

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 comprisee 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:

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

ExperimentHub with 2 records
# snapshotDate(): 2024-10-24
# $dataprovider: GEO, ArrayExpress
# $species: Homo sapiens
# $rdataclass: ExpressionSet
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["EH176"]]'

          title
  EH176 | GEO accession data GSE64985 as an ExpressionSet
  EH177 | ArrayExpress accession data E-MTAB-62 as an ExpressionSet
```

Data can then be extracted using:

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

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

```
> 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:

```
> data <- exprs(E.MTAB.62)
> dim(data)

[1] 12496  5372

> 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
3310     6.444028       6.639300      6.652987     6.716288      6.509133
7849     6.403596       6.447042      7.294512     6.506119      6.309392
```

```
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:

```
> pDat <- pData(E.MTAB.62)
> print(summary(pDat))
             OperatorVariation    DataSource        Groups_4
 Justin,,Lamb         : 324    GSE5258 : 324    cell line:1259
 Milton,W,Taylor      : 308    GSE7123 : 308    disease  : 765
 Roel,,Verhaak        : 284    GSE1159 : 284    neoplasm :2315
 Benjamin,,Haibe-Kains: 273    GSE4475 : 213    normal   :1033
 Michael,,Hummel      : 213    E-AFMX-6: 195
 Angela,,Hodges       : 195    GSE2990 : 167
 (Other)              :3775    (Other) :3881
                           Groups_15
 solid tissue neoplasm cell line: 831
 breast cancer                  : 672
 leukemia                       : 567
 normal solid tissue            : 566
 normal blood                   : 467
 blood non neoplastic disease   : 388
 (Other)                        :1881
                                 Groups_369   BloodNonBloodmetagroups
 breast cancer                        : 672   blood    :1922
 mononuclear cell infection           : 314   non blood:3450
 acute myeloid leukemia               : 295
 B-cell lymphoma                      : 213
 MCF7 breast epithelial adenocarcinoma: 213
 mononuclear cell                     : 143
 (Other)                              :3522
         Organism           OrganismPart
 Homo sapiens:5369   blood        :1089
 Mus musculus:   3   mammary gland:1033
                     bone marrow  : 733
                                  : 287
                     lung         : 286
                     brain        : 166
                     (Other)      :1778
                              CellType        CellLine
                                  :3333           :4112
 peripheral blood mononuclear cell: 452   mcf7    : 213
 blast cell, mononuclear cell     : 284   cultured:  88
 CD138+ plasma cell               : 142   pc3     :  64
 Leukocyte                        : 107   k562    :  48
 lymphocyte                       :  88   a549    :  30
 (Other)                          : 966   (Other) : 817
                        DiseaseState  DevelopmentalStage
```

```
                              :1274         :4816
 breast cancer                : 686   adult : 404
 acute myeloid leukemia       : 322   embryo: 110
 hepatitis c                  : 192   fetus :  42
 diffuse large B-cell lymphoma: 160
 breast tumor                 : 154
 (Other)                      :2584
                DiseaseStage             Sex                       Age
                      :4236                :3037                     :4681
 primary              : 500   female       :1016   10 days to 12 days:  23
 aggressive           : 141   hermaphrodite:   4   69                :  18
 grade 2              :  74   male         :1272   62                :  17
 lymph node metastasis:  59   mixed sex    :   9   65                :  17
 grade 1              :  39   unknown sex  :  34   61                :  15
 (Other)              : 323                        (Other)           : 601
        ArrayDataFile
 1102960533.CEL:   1
 1102960569.CEL:   1
 1102960602.CEL:   1
 1102960632.CEL:   1
 1102960664.CEL:   1
 1102960695.CEL:   1
 (Other)       :5366
```

The pheontypic data contains several "meta groups", labed 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:

```
> 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:

  Engreitz JM, Daigle BJ Jr, Marshall JJ, Altman RB. Independent
  component analysis: mining microarray data for fundamental human gene
  expression modules. J Biomed Inform 2010, 43(6):932-44.

  Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, Huber W,
  Ukkonen E, Brazma A. A global map of human gene expression. Nat
  Biotechnol 2010, 28(4):322-324.
  Brad Nelms (2016). _HumanAffyData experimental data package_. R
  package version 1.32.0,
  <https://www.bioconductor.org/packages/release/data/experiment/html/HumanAffyData.html>.

To see these entries in BibTeX format, use 'print(<citation>,
bibtex=TRUE)', 'toBibtex(.)', or set
'options(citation.bibtex.max=999)'.
```

References

[1] Jesse M. Engreitz, Bernie J. Daigle, Jonathan J. Marshall, and Russ B. Altman. Independent component analysis: Mining microarray data for fundamental human gene expression modules. *Journal of Biomedical Informatics*, 43(6):932–944, dec 2010. URL: http://dx.doi.org/10.1016/j.jbi.2010.07.001, doi:10.1016/j.jbi.2010.07.001.

[2] Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen, and Alvis Brazma. A global map of human gene expression. *Nature Biotechnology*, 28(4):322–324, 2010. URL: http://dx.doi.org/10.1038/nbt0410-322, doi:10.1038/nbt0410-322.

[3] Aron C Eklund and Zoltan Szallasi. Correction of technical bias in clinical microarray data improves concordance with known biological information. *Genome Biology*, 9(2):R26, 2008. URL: http://dx.doi.org/10.1186/gb-2008-9-2-r26, doi:10.1186/gb-2008-9-2-r26.