Annotating data with affycoretools and biomaRt

James W. MacDonald

June 15, 2014

1 Introduction

Most of the functions in *affycoretools* have been designed to annotate Affymetrix probesets using the annotation packages produced by the Biocore data team, in concert with the *annaffy* package. This paradigm works quite well for those Affymetrix chips that are popular enough to have an annotation package, but if there isn't a package for a given chip, one is left to either try to build a package using *AnnBuilder* (which can be a daunting task), or to try to annotate their probesets using *biomaRt*.

The *biomaRt* package is getting to be much more user friendly, but there still exists a gap between getting the annotation for a set of probesets and producing a finished product that can be presented to someone (e.g., an HTML table). Steffen Durinck (the maintainer of *biomaRt*) was kind enough to add functionality to his package that will allow one to easily take the output from his package straight into the `htmlpage` function of the *annotate* package to create HTML tables. I have since written some functions that are analogous to the existing *affycoretools* functions, but that use *biomaRt* and `htmlpage`.

For any first time users of *affycoretools*, I would direct you to the 'Using affycoretools' vignette first, which is designed to be an introductory text for this package.

2 Getting Started

First, a note about the *biomaRt* package. The functions in this package are designed to interface with an online BioMart database, usually one hosted at `http://www.biomart.org`. There are two interfaces; the default interface uses the *RCurl* package to connect to the database, but there is another interface that uses the *RMySQL* package to connect. The default interface is nice for doing interactive annotation of a small number of probesets, but can get *exceedingly* slow for anything more than say, 30 or 40 probesets. Therefore, I strongly recommend using the *RMySQL* interface if at all possible. Note that the kind folks in Seattle usually have a compiled version of *RMySQL* that can be installed using `biocLite`.

Given that most *affycoretools* functions are designed to help make an analysis more efficient, the most reasonable way to show how things work is to emulate an analysis. For this, we will be using the data in R_Home/library/affycoretools/examples. We can assume that these data are four different samples (which we will call A – D), that were run in triplicate (biological replicates). These data come from the Affymetrix HG-focus array, which has a BioC annotation package. However, instead of using the normal probeset mapping provided by Affy, we can use a re-mapped cdf where the mappings are based on Entrez Gene. This re-mapped cdf doesn't have an annotation package, so we will have to use *biomaRt* to annotate.

First, we compute expression values.

```
library(affycoretools)
eset <- justRMA(cdfname = "hsfocushsentrezgcdf")
```

Now we model the data using *limma*. Using a cell means model (without an intercept, denoted by the ~0 in the call to model.matrix) is necessary for most of the functions in the *affycoretools* package, as the names for the output are in general extracted from the contrasts matrix.

```
library(limma)
design <- model.matrix(~ 0 + factor(rep(1:4, each = 3)))
colnames(design) <- LETTERS[1:4]
contrast <- makeContrasts(A-B, C-D, levels = design)
fit <- lmFit(eset, design)
fit2 <- contrasts.fit(fit, contrast)
fit2 <- eBayes(fit2)
```

Here we are using A-B and C-D as names for the contrasts matrix. I usually give more descriptive names that will help somebody not involved in the analysis figure out what is being compared. After fitting the model and computing contrasts, we can output all significant probesets with an adjusted p-value less than 0.05 using limma2biomaRt.

```
limma2biomaRt(eset, fit2, design, contrast,
              species = "hsapiens", pfilt = 0.05,
              interactive = FALSE)
```

There are many more arguments for this function, but the defaults should be good for most uses (I hope). This will create HTML tables for the two comparisons we made.

Note that if there is no replication and one is just selecting probesets based on fold change, the foldFiltBM function can be used. In addition, if one has a vector of probe IDs to annotate, the probes2tableBM function will output HTML tables as well.

One might ask what probesets are differentially expressed in common between the two comparisons (or one might be interested in those probesets that are not in common). We can visualize this using a Venn diagram.

```
dt <- decideTests(fit2, method = "nestedF")
rslt <- vennCounts2(dt)
vennDiagram(rslt)
```

Figure 1 shows the Venn diagram. Now if we want to output tables containing the probesets in each cell of the Venn diagram, we can use vennSelectBM.

```
vennSelectBM(eset, design, dt,
             contrast, fit2, species = "hsapiens")
```

Note that there are two helper functions being used by these functions; linksBM, and annBM. These functions have two purposes; first, if called with no arguments, they list the type of annotation that is possible to use to create hyperlinks (linksBM), or that can be used for annotating without hyperlinking (annBM). For linksBM, this list includes all the annotation sources that htmlpage is able to hyperlink. This list can always be increased, and I am open to suggestions for other databases that one might want to link to. For annBM, the list is comprised of things that I normally use, and could also be extended to other annotation data.

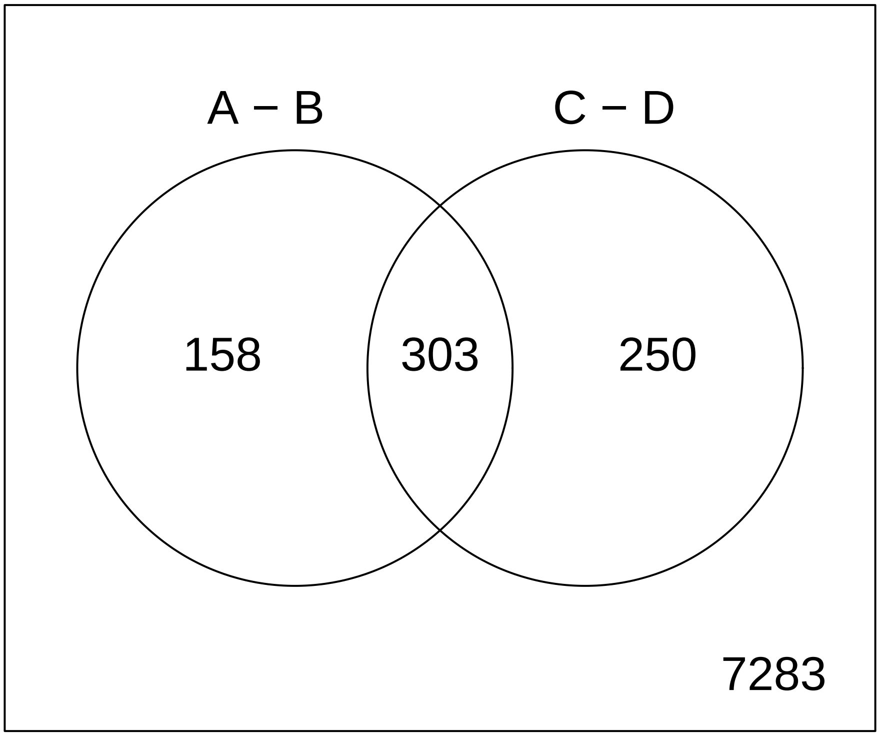

Figure 1: Venn diagram

The second purpose for these two functions is to list which of the *possible* annotation sources actually exist at a particular BioMart. This is done by passing a `mart` object to either function. This purpose is necessary because not all BioMart databases contain the same information. For instance, some of the BioMart databases don't have gene symbols.

```
library(biomaRt)
annBM()
```

```
## [1] "Symbol"      "Description" "GO"          "GOID"        "Chromosome"  "ChromLoc"
```

```
mart <- useMart("ensembl",
                "hsapiens_gene_ensembl")
annBM(mart, species="hsapiens")
```

```
## [1] "Symbol"      "Description" "Chromosome"  "ChromLoc"
```

```
mart <- useMart("ensembl",
                "celegans_gene_ensembl")
annBM(mart, species="celegans")
```

```
## [1] "Symbol"      "Description" "Chromosome"
```

3 Session Information

The version number of R and packages loaded for generating this vignette were:

- R version 3.1.0 (2014-04-10), x86_64-unknown-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=C, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils
- Other packages: AnnotationDbi 1.26.0, Biobase 2.24.0, BiocGenerics 0.10.0, DBI 0.2-7, GO.db 2.14.0, GenomeInfoDb 1.0.2, RSQLite 0.11.4, affy 1.42.2, affycoretools 1.36.1, biomaRt 2.20.0, genefilter 1.46.1, hgfocuscdf 2.14.0, limma 3.20.4
- Loaded via a namespace (and not attached): AnnotationForge 1.6.1, BBmisc 1.6, BSgenome 1.32.0, BatchJobs 1.2, BiocInstaller 1.14.2, BiocParallel 0.6.1, BiocStyle 1.2.0, Biostrings 2.32.0, Category 2.30.0, DESeq2 1.4.5, Formula 1.1-1, GOstats 2.30.0, GSEABase 1.26.0, GenomicAlignments 1.0.1, GenomicFeatures 1.16.2, GenomicRanges 1.16.3, Hmisc 3.14-4, IRanges 1.22.9, KernSmooth 2.23-12, MASS 7.3-33, Matrix 1.1-3, PFAM.db 2.14.0, R.methodsS3 1.6.1, R.oo 1.18.0, R.utils 1.32.4, R2HTML 2.2.1, RBGL 1.40.0, RColorBrewer 1.0-5, RCurl 1.95-4.1, Rcpp 0.11.2, RcppArmadillo 0.4.300.8.0, ReportingTools 2.4.0, Rsamtools 1.16.1, VariantAnnotation 1.10.2, XML 3.98-1.1, XVector 0.4.0, affyio 1.32.0, annaffy 1.36.0, annotate 1.42.0, biovizBase 1.12.1, bit 1.1-12, bitops 1.0-6, brew 1.0-6, caTools 1.17, cluster 1.15.2, codetools 0.2-8, colorspace 1.2-4, dichromat 2.0-0, digest 0.6.4, edgeR 3.6.2, evaluate 0.5.5, fail 1.2, ff 2.2-13, foreach 1.4.2, formatR 0.10, gcrma 2.36.0, gdata 2.13.3, geneplotter 1.42.0, ggbio 1.12.5, ggplot2 1.0.0, gplots 2.13.0, graph 1.42.0, grid 3.1.0, gridExtra 0.9.1, gtable 0.1.2, gtools 3.4.1, highr 0.3, hwriter 1.3, iterators 1.0.7, knitr 1.6, lattice 0.20-29, latticeExtra 0.6-26, locfit 1.5-9.1, munsell 0.4.2, oligoClasses 1.26.0, plyr 1.8.1, preprocessCore 1.26.1, proto 0.3-10, reshape2 1.4, rtracklayer 1.24.2, scales 0.2.4, sendmailR 1.1-2, splines 3.1.0, stats4 3.1.0, stringr 0.6.2, survival 2.37-7, tools 3.1.0, xtable 1.7-3, zlibbioc 1.10.0