Presenting the Results of an Analysis - annotating data and creating useful output for one-color arrays

James W. MacDonald
jmacdon@med.umich.edu

BioC 2008
July 28, 2008

Scenario

We assume that

- We are working with some sort of one-color arrays
  - Affymetrix
  - Nimblegen
  - Illumina
- Some of this is applicable to two-color arrays as well
- We want to map some identifier to gene information

Annotation Packages

- Chip - level packages
  - hgu133plus2.db
  - illuminaHumanv3ProbeID
- Species - level packages
  - org.Hs.eg.db
  - org.Mm.eg.db
- biomaRt
- GO.db
- KEGG.db

Chip - Level Data

What mappings can we do?

```
> ls("package:hgfocus.db")
```

```
 [1] "hgfocus"
 [2] "hgfocusACCNUM"
 [3] "hgfocusALIAS2PROBE"
 [4] "hgfocusCHR"
 [5] "hgfocusCHRLENGTHS"
 [6] "hgfocusCHRLOC"
 [7] "hgfocusENSEMBL"
 [8] "hgfocusENSEMBL2PROBE"
 [9] "hgfocusENTREZID"
[10] "hgfocusENZYME"
```

Annotation Table Contents

What is in a given table?

```
> ?hgfocusUNIGENE
```

Annotation Table Contents

What is in a given table?

```
> ?hgfocusUNIGENE
> head(toTable(hgfocusUNIGENE))
```

Annotation Table Contents

What is in a given table?

```
> ?hgfocusUNIGENE
> head(toTable(hgfocusUNIGENE))
```

```
   probe_id unigene_id
1 1007_s_at  Hs.631988
2   1053_at  Hs.647062
3    117_at  Hs.654614
4    121_at  Hs.469728
5 1255_g_at   Hs.92858
6   1294_at   Hs.16695
```

Gene Names

Gene name for one illumina probe:

```
> get("1010243", illuminaHumanv3ProbeIDGENENAME)
```

Gene Names

Gene name for one illumina probe:

```
> get("1010243", illuminaHumanv3ProbeIDGENENAME)
[1] "lipoprotein, Lp(a)-like 2"
```

Gene Names

Gene name for one illumina probe:

```
> get("1010243", illuminaHumanv3ProbeIDGENENAME)
```

```
[1] "lipoprotein, Lp(a)-like 2"
```

And for multiple probes:

```
> illumina <- c("10008", "10010", "10017", "10019")
> mget(illumina, illuminaHumanv3ProbeIDGENENAME)
```

Gene Names

Gene name for one illumina probe:

```
> get("1010243", illuminaHumanv3ProbeIDGENENAME)
[1] "lipoprotein, Lp(a)-like 2"
```

And for multiple probes:

```
> illumina <- c("10008", "10010", "10017", "10019")
> mget(illumina, illuminaHumanv3ProbeIDGENENAME)
$`10008`
[1] "arginine-glutamic acid dipeptide (RE) repeats"

$`10010`
[1] NA

$`10017`
[1] NA

$`10019`
```

Reverse Mapping

Say we want to know how many probesets
on the hgfocus chip map to the 'caffeine metabolism' pathway

```
> library(KEGG)
> get("Caffeine metabolism", revmap(KEGGPATHID2NAME))
```

Reverse Mapping

Say we want to know how many probesets
on the hgfocus chip map to the 'caffeine metabolism' pathway

```
> library(KEGG)
> get("Caffeine metabolism", revmap(KEGGPATHID2NAME))
[1] "00232"
```

Reverse Mapping

Say we want to know how many probesets
on the hgfocus chip map to the 'caffeine metabolism' pathway

```
> library(KEGG)
> get("Caffeine metabolism", revmap(KEGGPATHID2NAME))
[1] "00232"
> get("00232", revmap(hgfocusPATH))
```

Reverse Mapping

Say we want to know how many probesets
on the hgfocus chip map to the 'caffeine metabolism' pathway

```
> library(KEGG)
> get("Caffeine metabolism", revmap(KEGGPATHID2NAME))
[1] "00232"
> get("00232", revmap(hgfocusPATH))
[1] "214440_at"   "206797_at"   "207609_s_at"
[4] "1494_f_at"   "207718_x_at" "208327_at"
[7] "210301_at"
```

Practice

What is the gene symbol for Entrez Gene ID '1234'?
The gene name?
What Entrez Gene IDs map to the 'Caffeine metabolism' pathway?

Practice

What is the gene symbol for Entrez Gene ID '1234'?

```
> get("1234", org.Hs.egSYMBOL)
```

```
[1] "CCR5"
```

The gene name?
What Entrez Gene IDs map to the 'Caffeine metabolism' pathway?

Practice

What is the gene symbol for Entrez Gene ID '1234'?

```
> get("1234", org.Hs.egSYMBOL)
```

```
[1] "CCR5"
```

The gene name?

```
> get("1234", org.Hs.egGENENAME)
```

```
[1] "chemokine (C-C motif) receptor 5"
```

What Entrez Gene IDs map to the 'Caffeine metabolism' pathway?

Practice

What is the gene symbol for Entrez Gene ID '1234'?

```
> get("1234", org.Hs.egSYMBOL)
```

```
[1] "CCR5"
```

The gene name?

```
> get("1234", org.Hs.egGENENAME)
```

```
[1] "chemokine (C-C motif) receptor 5"
```

What Entrez Gene IDs map to the 'Caffeine metabolism' pathway?

```
> caff <- get("Caffeine metabolism",
+             revmap(KEGGPATHID2NAME))
> get(caff, revmap(org.Hs.egPATH))
```

```
[1] "9"    "10"   "1544" "1548" "1549" "1553"
[7] "7498"
```

What is biomaRt?

- Package for annotating using Biomart
- Multiple annotation sources
- Use webservice (Rcurl) or database (RMySQL)
- Species level
- Some manufacturer IDs
- Much greater amount of information

Basic Usage

Let's load the package and see what annotation choices we have

```
> library(biomaRt)
> listMarts()
```

Basic Usage

Let's load the package and see what annotation choices we have

```
> library(biomaRt)
> listMarts()
 [1] ENSEMBL 49 GENES (SANGER)
 [2] ENSEMBL 49 HOMOLOGY (SANGER)
 [3] ENSEMBL 49 PAIRWISE ALIGNMENTS (SANGER)
 [4] ENSEMBL 49 MULTIPLE ALIGNMENTS (SANGER)
 [5] ENSEMBL 49 VARIATION  (SANGER)
 [6] ENSEMBL 49 GENOMIC FEATURES (SANGER)
 [7] VEGA 30  (SANGER)
 [8] MSD PROTOTYPE (EBI)
 [9] UNIPROT PROTOTYPE (EBI)
[10] HIGH THROUGHPUT GENE TARGETING AND TRAPPING (SAN
[11] GRAMENE (CSHL)
[12] REACTOME (CSHL)
```

Ensembl

Available Datasets

We will use Ensembl annotations

```
> mart <- useMart("ensembl")
```

What datasets are available?

```
> listDatasets(mart)
```

Available Datasets

We will use Ensembl annotations

```
> mart <- useMart("ensembl")
```

What datasets are available?

```
> listDatasets(mart)
```

```
 [1] "oanatinus_gene_ensembl"
 [2] "gaculeatus_gene_ensembl"
 [3] "cporcellus_gene_ensembl"
 [4] "lafricana_gene_ensembl"
 [5] "stridecemlineatus_gene_ensembl"
 [6] "scerevisiae_gene_ensembl"
 [7] "eeuropaeus_gene_ensembl"
 [8] "etelfairi_gene_ensembl"
 [9] "ptroglodytes_gene_ensembl"
[10] "cintestinalis_gene_ensembl"
[11] "ppygmaeus_gene_ensembl"
[12] "ocuniculus_gene_ensembl"
```

Homo sapiens

We use the Homo sapiens dataset

```
> mart <- useMart("ensembl", "hsapiens_gene_ensembl")
```

```
Checking attributes and filters ... ok
```

Now some terminology

- attribute - thing(s) we want to get
- filter - input data type
- value - input values

Available attributes


```
> listAttributes(mart)
                      name          description
1          affy_hcg110          AFFY HCG110
2        affy_hg_focus        AFFY HG FOCUS
3 affy_hg_u133_plus_2 AFFY HG U133-PLUS-2
4        affy_hg_u133a        AFFY HG U133A
5     affy_hg_u133a_v2      AFFY HG U133A 2
6        affy_hg_u133b        AFFY HG U133B
```

Available filters


```
> listFilters(mart)

                        name
1           affy_hc_g110
2         affy_hc_g110-2
3          affy_hg_focus
4        affy_hg_focus-2
5    affy_hg_u133_plus_2
6 affy_hg_u133_plus_2-2
                  description
1       Affy hc g 110 ID(s)
2       Affy hc g 110 ID(s)
3       Affy hg focus ID(s)
4       Affy hg focus ID(s)
5 Affy hg u133 plus 2 ID(s)
6 Affy hg u133 plus 2 ID(s)
```

Simple query

We want the Entrez Gene ID for Affy ID 1007_s_at from HG-U133A chip

```
> getBM(attributes = c("affy_hg_u133a", "entrezgene")
+       filters = "affy_hg_u133a",
+       values = "1007_s_at",
+       mart = mart)
```

Simple query

We want the Entrez Gene ID for Affy ID 1007_s_at from HG-U133A chip

```
> getBM(attributes = c("affy_hg_u133a", "entrezgene")
+       filters = "affy_hg_u133a",
+       values = "1007_s_at",
+       mart = mart)

  affy_hg_u133a entrezgene
1     1007_s_at        780
```

Practice

What is the gene name for Ensembl ID ENSG00000112715?
What chromosome is it located on?
What are the start and stop coordinates?

Practice

What is the gene name for Ensembl ID ENSG00000112715?

```
> getBM(c("ensembl_gene_id", "description"),
+       "ensembl_gene_id", "ENSG00000112715",
+       mart)

  ensembl_gene_id
1 ENSG00000112715

1 Vascular endothelial growth factor A precursor (VEG
```

What chromosome is it located on?
What are the start and stop coordinates?

Practice

What is the gene name for Ensembl ID ENSG00000112715?
What chromosome is it located on?

```
> getBM(c("ensembl_gene_id","chromosome_name"),
+       "ensembl_gene_id", "ENSG00000112715",
+       mart)

  ensembl_gene_id chromosome_name
1 ENSG00000112715               6
```

What are the start and stop coordinates?

Practice

What is the gene name for Ensembl ID ENSG00000112715?
What chromosome is it located on?
What are the start and stop coordinates?

```
> getBM(c("sequence_gene_chrom_start",
+         "sequence_gene_chrom_end",
+         "ensembl_gene_id"),
+       "ensembl_gene_id", "ENSG00000112715",
+       mart)

  gene_chrom_start gene_chrom_end
1         43845926       43862202
  ensembl_gene_id
1 ENSG00000112715
```

Bulk Annotating

Scenario

- Data have been analyzed
- We want to present results to layperson
- Content-rich
- Easy to use
- Choices:
  - HTML
  - Text

Packages

- annaffy
- biomaRt and annotate
- affycoretools

annaffy

- HTML tables
- text tables
- aaf.handler()

```
> aaf.handler()

 [1] "Probe"
 [2] "Symbol"
 [3] "Description"
 [4] "Chromosome"
 [5] "Chromosome Location"
 [6] "GenBank"
 [7] "Gene"
 [8] "Cytoband"
 [9] "UniGene"
[10] "PubMed"
[11] "Gene Ontology"
[12] "Pathway"
```

affycoretools and annaffy

First a bit of setup...

```
> library(affycoretools)
> load(paste(.path.package("affycoretools"),
+                "/doc/exprSet.Rdata", sep=""))
> prbs <- featureNames(eset)[500:550]
```

Create tables


```
> probes2table(eset = eset, probids = prbs,
+                 lib = "hgfocus.db", text = TRUE,
+                 filename = "test")
```

affycoretools and biomaRt

Useful when data to be annotated does not have a chip-level annotation package.

```
> probes2tableBM(eset = eset, probids = prbs[1:10],
+                    species = "hsapiens",
+                    ann.source = "affy_hg_focus",
+                    filename = "test2",affyid = TRUE,
+                    mysql = FALSE)
```