

Pathway annotation case study
intro insert, March 2008 bioconductor course
©2008, VJ Carey Ph D

- "Gene sets" are popular tools for analysis
- rapid survey of a large family of gene sets is facilitated by programming
- conversion of moderately conventional annotation for genes/gene sets to operators on Bioconductor data structures is illustrated

Is my gene in any pathways?

- case of TBX21
- KEGG – nothing
- NCBI – nothing
- What about the Broad GSEA-related gene sets?
- Bioconductor package GSEABase helps navigate these

broadsets.rda

```
> library(GSEABase)
> if (!exists("broadsets")) load("broadsets.rda")
> broadsets

GeneSetCollection
names: chr16q, chr5q23, ..., GNF2_ZAP70 (3337 total)
unique identifiers: CMAR, USP10, ..., ZNF302 (38975 total)
types in collection:
  geneIdType: SymbolIdentifier (1 total)
  collectionType: BroadCollection (1 total)

> class(broadsets)

[1] "GeneSetCollection"
attr(,"package")
[1] "GSEABase"

> getClass(class(broadsets))

Slots:

Name: .Data
Class: list

Extends:
Class "list", from data part
Class "vector", by class "list", distance 2
Class "AssayData", by class "list", distance 2
```

information on a set

```
> broadsets[[1]]
```

```
setName: chr16q  
geneIds: CMAR, USP10, ..., SLI1 (total: 5)  
geneIdType: Symbol  
collectionType: Broad  
  bcCategory: c1 (Positional)  
  bcSubCategory: NA  
details: use 'details(object)'
```

```
> details(broadsets[[1]])
```

```
setName: chr16q  
geneIds: CMAR, USP10, ..., SLI1 (total: 5)  
geneIdType: Symbol  
collectionType: Broad  
  bcCategory: c1 (Positional)  
  bcSubCategory: NA  
setIdentifier: c1:100  
description: Genes in cytogenetic band chr16q  
organism: Human  
pubMedIds:  
urls: ftp://ftp.broad.mit.edu/pub/gsea/msigdb_v2.1.xml  
      http://genome.ucsc.edu/cgi-bin/hgTracks?position=16q  
contributor: Broad Institute  
setVersion: 0.0.1  
creationDate: Wed Feb 27 21:36:52 2008
```

GeneSetCollection operations

- a GeneSetCollection instance is an R list of GeneSets
- iteration over list elements is relatively easy in R
- need to know how to operate usefully on a GeneSet
- poked at one above with 'details' method
- another method of interest: `geneIds`

```
> geneIds(broadsets[[1]])
```

```
[1] "CMAR" "USP10" "PSORS8" "WT3" "SLI1"
```

```
> allids = lapply(broadsets, geneIds)
```

```
> tbxchk = sapply(allids, function(x) any(x == "TBX21"))
```

```
> sum(tbxchk)
```

```
[1] 12
```

```
> hastbx = which(tbxchk)
```

```
> sapply(broadsets[hastbx], setName)
```

```
[1] "chr17q21" "MATSUDA_VALPHAINKT_DIFF"
```

```
[3] "CATRRAGC_UNKNOWN" "CTTTGA_V$LEF1_Q2"
```

```
[5] "CCCNNGGAR_V$OLF1_01" "V$LYF1_01"
```

```
[7] "V$P53_DECAMER_Q2" "V$ISRE_01"
```

```
[9] "V$OLF1_01" "GNF2_CD7"
```

```
[11] "GNF2_IL2RB" "GNF2_PTPN4"
```

More info

- we have identified 12 'gene sets' that include TBX21
- what are they? we can see that one is just a cytoband

```
> broadsets[["V$LYF1_01"]]
```

```
setName: V$LYF1_01
```

```
geneIds: RTN3, IMPA2, ..., ZNRF1 (total: 273)
```

```
geneIdType: Symbol
```

```
collectionType: Broad
```

```
bcCategory: c3 (Motif)
```

```
bcSubCategory: NA
```

```
details: use 'details(object)'
```

More info

```
> details(broadsets[["V$LYF1_01"]])
```

```
setName: V$LYF1_01
```

```
geneIds: RTN3, IMPA2, ..., ZNRF1 (total: 273)
```

```
geneIdType: Symbol
```

```
collectionType: Broad
```

```
  bcCategory: c3 (Motif)
```

```
  bcSubCategory: NA
```

```
setIdentifier: c3:1105
```

```
description: Genes with promoter regions [-2kb,2kb] around transcription start site containing motif  
(longDescription available)
```

```
organism: Human,Mouse,Rat,Dog
```

```
pubMedIds:
```

```
urls: ftp://ftp.broad.mit.edu/pub/gsea/msigdb_v2.1.xml
```

```
contributor: Xiaohui Xie
```

```
setVersion: 0.0.1
```

```
creationDate: Wed Feb 27 21:37:41 2008
```

```
> longDescription(broadsets[["V$LYF1_01"]])
```

```
"XX<br> XX<br> XX<br> FA LyF-1<br> XX<br> SY LyF-1; Ikaros; lymphoid transcription factor; Lyf-1
```

- This longDescription result is not very useful ... it is a long string of HTML. If we write it to a file, we can run a browser. Use writeLines

SY LyF-1; Ikaros; lymphoid transcription factor; Lyf-1.

XX

OS mouse, *Mus musculus*

OC eukaryota; animalia; metazoa; chordata; vertebrata; tetrapoda; mammalia;

OC eutheria; rodentia; myomorpha; muridae; murinae

XX

XX

CL C0001; CH; 2.3.2.2.9.

XX

SZ 50 kDa (SDS)

XX

SF at least three proteins generated by alternative splicing are known, Ik-1

SF , Ik-2 , Ik-5 [2]

XX

CP B, T cells

XX

FF essential regulator for lymphoid lineage specification and subsequent

FF proliferation in the T lineage [1];

FF lack of Ikarus (LyF-1) activity at the late stages of thymocyte maturation

FF leads to uncontrolled lymphoproliferation and to the rapid development of

FF malignant T cell leukemia and lymphoma [1];

XX

∫N [1]; RE0006503.

∫X MEDLINE; 96028103.

∫A Winandy S., Wu P., Georgopoulos K.

∫T A dominant mutation in the Ikaros gene leads to rapid development of
∫T leukemia and lymphoma

∫L Cell 83:289-299 (1995).

∫N [2]; RE0006501.

∫X MEDLINE; 95021239.

∫A Hahn K., Ernst P., Lo K., Kim G. S., Turck C., Smale S. T.

∫T The lymphoid transcription factor LyF-1 is encoded by specific,
∫T alternatively spliced mRNAs derived from the Ikaros gene

∫L Mol. Cell. Biol. 14:7111-7123 (1994).

∫N [3]; RE0001480.

∫X MEDLINE; 92017799.

∫A Lo K., Landau N. R., Smale S. T.

∫T LyF-1, a transcriptional regulator that interacts with a novel class of
∫T promoters for lymphocytes-specific genes

∫L Mol. Cell. Biol. 11:5229-5243 (1991).

∫X

Upshots

- A collection of over 3000 sets of genes is bound to a single R variable name (broadsets)
- Each set is self-documenting and includes a list of HUGO identifiers (as given by Broad)
- methods `geneIds`, `details`, `longDescription` provide uniform information on each set
- programming expertise useful
 - general string matching (`x == 'TBX21'`) or pattern matching (`grep`, `caseconversion` etc) available directly to constituents
 - shortcuts `broadsets[["V$LYF1_01"]]`
- exploit sets and their structures for thorough statistical analysis

Application – note number of features retained

```
> library(Biobase)
> library(ALL)
> data(ALL)
> keep = broadsets[["V$LYF1_01"]]
> geneIdType(keep) = AnnotationIdentifier(annotation(ALL))
> ALL[geneIds(keep), ]
```

ExpressionSet (storageMode: lockedEnvironment)

assayData: 193 features, 128 samples

element names: exprs

phenoData

sampleNames: 01005, 01010, ..., LAL4 (128 total)

varLabels and varMetadata description:

cod: Patient ID

diagnosis: Date of diagnosis

...: ...

date last seen: date patient was last seen

(21 total)

featureData

featureNames: 36496_at, 31395_i_at, ..., 40730_at (193 total)

fvarLabels and fvarMetadata description: none

experimentData: use 'experimentData(object)'

pubMedIds: 14684422 16243790

Annotation: hgu95av2