

Lecture 4: annotation in bioconductor

©2008 VJ Carey PhD
Channing Lab

- commitments
 - platform annotations – SQLite basis
 - organism annotation – org.Hs.eg.db
 - web services: biomaRt
- ### Annotation concept review
- platforms: feature annotation
 - experiments: MIAME annotation
 - samples: phenotype, disease, protocol
 - genomes, biological processes: ontologies for sequence, gene products, etc.
 - resources: networks of databases and tables

1

Bioconductor commitments

- identifier annotation maps for prevalent platforms are created every three months
- open tools for annotating custom platforms are provided; advice given
- large-scale annotation sets for important organisms are also provided
 - org.Hs.eg.db
- harder problem: feature maps (e.g., CDF files for affy) for chips depend on manufacturer openness, tractability – see pd.mapping packages

3

2

Platforms: Classic examples

```
> library(hgfocus.db)
> objects("package:hgfocus.db")
 [1] "hgfocus"
 [4] "hgfocusCHR"
 [7] "hgfocusENSEMBL"
 [10] "hgfocusENZYME"
 [13] "hgfocusGO"
 [16] "hgfocusMAP"
 [19] "hgfocusORGANISM"
 [22] "hgfocusPFAM"
 [25] "hgfocusPROSITE"
 [28] "hgfocusUNIGENE"
 [31] "hgfocus_dbfile"
      "hgfocusACONUM"
      "hgfocusCHALLENGHS"
      "hgfocusENSEMBL2PROBE"
      "hgfocusENTREZID"
      "hgfocusGENENAME"
      "hgfocusGO2PROBE"
      "hgfocusOMIM"
      "hgfocusPATH2PROBE"
      "hgfocusPMD2PROBE"
      "hgfocusSYMBOL"
      "hgfocus_dbconn"
> library(annotate)
> nn = ls(hgfocusSYMBOL)[1:3]
> mget(nn, hgfocusSYMBOL)
$`1007_s_at`
 [1] "DDRL"
$`1053_at`
 [1] "RFC2"
$`117_at`
 [1] "HSPA6"
```

4

Platforms: Classic examples

```
> library(hgfocuscdf)
> hgfocuscdf
<environment: 0x2ebdf10>
> pid = ls(hgfocuscdf)[1:3]
> lapply(mget(pid, hgfocuscdf), "[", 1:3, 1:2)
$`1007_s_at`
      pm      mm
[1,] 52830 53278
[2,] 85725 86173
[3,] 158200 158648

$`1053_at`
      pm      mm
[1,] 179673 180121
[2,] 6092 6540
[3,] 151413 151861

$`117_at`
      pm      mm
[1,] 62627 63075
[2,] 75044 75492
[3,] 158420 158868
> library(affy)
> indics2xy(62627, cdf = "hgfocuscdf")
```

5

Platforms: Classic examples

```
> library(hgfocusprobe)
> hgfocusprobe
Object of class probetable.data.frame with 98149 rows and 6 columns.
> as.data.frame(hgfocusprobe)[1:3, ]
      sequence      x      y Probe.Set.Name Probe.Interrogation.Position
1 CACCCAGCTGGTCTGTGGATGGGA 413 117 1007_s_at 3330
2 GCCCAGCTGGACAACACTGATTCCT 156 191 1007_s_at 3443
3 TGGACCCAGCTGGCTGAGATCTGG 55 353 1007_s_at 3512
Target.Strandedness
1 Antisense
2 Antisense
3 Antisense
```

7

```
      x      y
[1,] 354 139
```

Platforms: new approach (oligo)

```
> library(pd.mapping50k.xba240)
> pd.mapping50k.xba240
An object of class "AffySNPPDInfo"
Slot "getdb":
function ()
{
  if (!is.null(globals$dbCon) && isIdCurrent(globals$dbCon))
    return(globals$dbCon)
  initDbConnection()
}
<environment: namespace:pd.mapping50k.xba240>

Slot "tableInfo":
NULL
<0 rows> (or 0-length row.names)

Slot "geometry":
[1] 1600 1600

Slot "manufacturer":
[1] "Affymetrix"

Slot "genomebuild":
[1] "NCBI Build 36"
```

8

6

pd.mapping approach

```

> xb = pd.mapping50k.xba240@getdb()
> dbListTables(xb)
[1] "featureSet" "mmfeature" "pm_jm" "pmfeature" "qcmfeature"
[6] "qcpm_qcmm" "qcpfeature" "sequence" "sqlite_stat1" "table_info"
> dbGetQuery(xb, "select * from featureSet limit 200,3")[, -11]
fsetid man_fsetid affy_snp_id dbsnp_rs_id chrom physical_pos strand
1 201 SNP_A-1676644 NA rs10483327 14 26493657 -
2 202 SNP_A-1673622 NA rs10520572 15 82382679 -
3 203 SNP_A-1723534 NA rs30058 5 122329602 -
cytoband allele_a allele_b fragment_length dbsnp
1 q12 A C 1149 0
2 q25.2 A G 814 1
3 q23.2 C T 1591 1
1
2
3 Variation_8213 // chr5:122325602-123314085 // Affymetrix 500K and 100K SNP Mapping Arrays // 176

```

9

```

"CCCTGCCATGTCCCTGGTGTACTGACCTCTCAAGGCTTCTCCAAATCTG" 0V5TtXrnHhd0KfXUN4
"CCCTGCCATGTCCCTGGTGTACTGACCTCTCAAGGCTTCTCCAAATCTG" 0V5eeeqi0e10.cqs9M
"CCCTGCCCTGCTGCTGGGGGAGATGCTGTCCATGTTTCTAGGGGTATTCAT" 0V6SLfektXhRRZRK1A
"CCCTGCCAGAGTCTTCTGAGGGATTACACTCACCCAGCCGACAGGGAGAA" 0V6gmRRRR0gnhqAoI
"CCCTGGGAAGCCGACATACACACCACATCGAAAGCTGACCGGAAAGGAAG" 0V7V1Ra7hdKfYX_neg
"CCCTGTCCCTCCACGCTCTGTGACCTCAGGCCACTAGGCTTTGGCTCTGGA"

```

illumina lumi twist

```

> library(LumiHumanV2)
> library(Lumi)
> kk = ls(LumiHumanV2REFSEQ)[1050:1054]
> kk
[1] "0V5TtXrnHhd0KfXUN4" "0V5eeeqi0e10.cqs9M" "0V6SLfektXhRRZRK1A"
[4] "0V6gmRRRR0gnhqAoI" "0V7V1Ra7hdKfYX_neg"
> nuID2targetID(kk, lib = "LumiHumanV2")
$ 0V5TtXrnHhd0KfXUN4
[1] "ILMN_1192"
$ 0V5eeeqi0e10.cqs9M
[1] "ILMN_20628"
$ 0V6SLfektXhRRZRK1A
[1] "ILMN_133840"
$ 0V6gmRRRR0gnhqAoI
[1] "ILMN_81927"
$ 0V7V1Ra7hdKfYX_neg
[1] "ILMN_17948"
> id2seq(kk)

```

10

illumina expression arrays: SQLite approach

```

> library(illuminaHumanv2.db)
> ilco = illuminaHumanv2.dbconn()
> dbListTables(ilco)
[1] "alias" "chromosomes" "chrlengths" "chromosome_locations"
[4] "chromosomes" "cytogenetic_locations" "ec"
[7] "ensembl" "gene_info" "genes"
[10] "go_bp" "go_bp_all" "go_cc"
[13] "go_cc_all" "go_jf" "go_jf_all"
[16] "kegg" "map_counts" "map_metadata"
[19] "metadata" "omim" "pfam"
[22] "probes" "prosite" "pubmed"
[25] "refseq" "sqlite_stat1" "unigene"

```

11

12

creating a hyperlinked page of annotation resolutions

```

> library(annaaffy)
> meLaFrag = c("EGF", "EGFR", "HRAS", "ARAF", "PIK3RS", "MAP2K1",
+ "NRAS", "MAPK1", "BRAF")
> library(GSEABase)
> gs = GeneSet(meLaFrag, geneIdType = SymbolIdentifier())
> geneIdType(gs) = AnnotationIdentifier("illuminaHumanv2.db")
> imf = geneIds(gs)
> tab = aafTableAnn(imf, "illuminaHumanv2.db")
> saveHTML(tab, file = "mftab.html")

```

13

queries

```

> library(org.Hs.eg.db)
> oc = org.Hs.eg.dbconn()
> dbListTables(oc)

[1] "accessions"
[4] "chromosome_locations"
[7] "ec"
[10] "gene_info"
[13] "go_bp_all"
[16] "go_mf"
[19] "map_counts"
[22] "onlin"
[25] "pubmed"
[28] "unigene"

"alias"
"chromosomes"
"ensembl"
"genes"
"go_cc"
"go_mf_all"
"map_metadata"
"pfam"
"refseq"

"chrlengths"
"cytogetic_locations"
"ensembl_prot"
"go_bp"
"go_cc_all"
"kegg"
"metadata"
"prosite"
"sqlite_stat1"

> dbGetQuery(oc, "select * from gene_info limit 199, 5")

_id      gene_name symbol
1 250 angiogenin, ribonuclease, RNase A family, 5 ANG
2 251 angiopoietin 1 ANGPT1
3 252 angiopoietin 2 ANGPT2
4 253 ankyrin 1, erythrocytic ANK1
5 254 ankyrin 2, neuronal ANK2

> dbGetQuery(oc, "select * from gene_info INNER JOIN go_MF
+ using(_id) where gene_info.symbol = 'TBX21' ")

```

15

organism-level annotation

org.Dm.eg.db	Biocore Data Team	Genome wide annotation for Fly
org.Hs.eg.db	Biocore Data Team	Genome wide annotation for Human
org.Mm.eg.db	Biocore Data Team	Genome wide annotation for Mouse
org.Rn.eg.db	Biocore Data Team	Genome wide annotation for Rat
org.Sc.spd.db	Biocore Data Team	Genome wide annotation for Yeast

14

```

_id gene_name symbol go_id evidence
1 12115 T-box 21 TBX21 GO:0003700 TAS

```

16

pathway and gene ontology resources

```
> library(KEGG.db)
> objects("package:KEGG.db")
[1] "KEGG" "KEGGGENZYMEID2GO" "KEGGEXTID2PATHID" "KEGGGO2ENZYMEID"
[5] "KEGGMAPCOUNTS" "KEGGPATHID2EXTID" "KEGGPATHID2NAME" "KEGGPATHNAME2ID"
[9] "KEGG_dbInfo" "KEGG_dbconn" "KEGG_dbfile" "KEGG_dbschema"
> library(GO.db)
> objects("package:GO.db")
[1] "GO" "GOBPCHILDREN" "GOBPCHILDREN" "GOBPOFFSPRING"
[5] "GOBPARENTS" "GOCCANCESTOR" "GOCCCHILDREN" "GOCCOFFSPRING"
[9] "GOMAPCOUNTS" "GOMFANCESTOR" "GOMFCHILDREN" "GOMFOFFSPRING"
[13] "GOMFOFFSPRING" "GOMFPARENTS" "GOMFSOLETE" "GOSYNONYM"
[17] "GOTERM" "GO_dbInfo" "GO_dbconn" "GO_dbfile"
[21] "GO_dbschema"
```

17

```
2 ENSEMBL_48_HOMOLOGY (SANGER)
3 ENSEMBL_48_PAIRWISE_ALIGNMENTS (SANGER)
4 ENSEMBL_48_MULTIPLE_ALIGNMENTS (SANGER)
5 ENSEMBL_48_VARIATION (SANGER)
6 ENSEMBL_48_GENOMIC_FEATURES (SANGER)
7 VEGA_29 (SANGER)
8 MSD_PROTOTYPE (EBI)
9 UNIPROT_PROTOTYPE (EBI)
10 GRAMENE (CSHL)
11 REACTOME (CSHL)
12 WORMBASE (CSHL)
13 RGD_IP1_MART (MCM)
14 RGD_IP1_MART (MCM)
15 RGD_MICROSATELLITE_MARKERS (MCM)
16 PRIDE (EBI)
17 EURATWART (EBI)
18 PEPSEEKER (UNIVERSITY OF MANCHESTER)
19 PANCREATIC_EXPRESSION_DATABASE (INSTITUTE OF CANCER)
```

19

web services: biomaRt

```
> library(biomaRt)
> allm = listMarts()
> allm
      name
1 ensembl
2 compara_mart_homology_48
3 compara_mart_pairwise_ga_48
4 compara_mart_multiple_ga_48
5 snp
6 genomic_features
7 vega
8 msd
9 uniprot
10 ENSEMBL_MART_ENSEMBL
11 REACTOME
12 wormbase_current
13 rgd_mart
14 ipi_rat_mart
15 SSIIP_mart
16 pride
17 ensembl_expressionmart_47
18 pepseekerGOLD_mart06
19 Pancreatic_Expression
      version
1 ENSEMBL_48_GENES (SANGER)
```

18

mart examples

```
> mm = useMart("snp")
> listDatasets(mm)[1:2, ]
      dataset
1 ptroglycetes_snp
2 hsapiens_snp
1
2 Homo sapiens SNPs (dbSNP127;HGvbase15;TSC1;ENSEMBL;Affy 100K Array;Affy 500K Array)
1
2 dbSNP127;HGvbase15;TSC1;ENSEMBL;Affy 100K Array;Affy 500K Array
> mm = useMart("snp", dataset = "hsapiens_snp")
Checking attributes and filters ... ok
> listFilters(mm)[1:3, ]
      name description
1 band_end <NA>
2 band_start <NA>
3 chr_name Chromosome name
> args(getEM)
function (attributes, filters = "", values = "", mart, curl = NULL,
output = "data.frame", list.names = NULL, na.value = NA,
checkFilters = TRUE, verbose = FALSE, uniqueRows = TRUE)
NULL
      description
1 Pan troglodytes SNPs (dbSNP125)
2 version
3 dbSNP125
```

20

Summary

- packages exist for platform, organism, pathway/ontology annotation maps
- SQLite database tables are fundamental entities
- annotate package lookUp function resolves vectors of identifiers
- AnnotationDbi get/mget; revmap facilities illustrated previously
- SQLite annotation packages used by annaffy – aafTableAnn – regardless of manufacturer
- sqlForge tools for creating .db packages in AnnotationDbi