One of the largest challenges in analyzing genomic data is associating the experimental data with the available biological metadata, e.g., sequence, gene annotation, chromosomal maps, literature. And making that data available for computation.

Bioconductor provides three main packages for this purpose:
- annotate (end-user);
- AnnBuilder (developer);
- annaffy (end-user – will see a name change)
WWW resources

- Nucleotide databases: e.g. GenBank.
- Gene databases: e.g. Entrez Gene, UniGene.
- Protein sequence and structure databases: e.g. SwissProt, Protein DataBank (PDB).
- Literature databases: e.g. PubMed, OMIM.
- Chromosome maps: e.g. NCBI Map Viewer.
- Pathways: e.g. KEGG.
- Entrez is a search and retrieval system that integrates information from databases at NCBI (National Center for Biotechnology Information).
- if you know of some we should be using – please let us know

annotate: matching IDs

Important tasks
- Associate manufacturers or in-house probe identifiers to other available identifiers. E.g.
  - Affymetrix IDs → Entrez Gene IDs
  - Affymetrix IDs → GenBank accession number.
- Associate probes with biological data such as chromosomal position, pathways.
- Associate probes with published literature data via PubMed (need PMID).

<table>
<thead>
<tr>
<th>Affymetrix identifier</th>
<th>“41046_s_at”</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGU95A chips</td>
<td>“9203”</td>
</tr>
<tr>
<td>Entrez Gene ID</td>
<td>“X95808”</td>
</tr>
<tr>
<td>GenBank accession #</td>
<td>“ZNF261”</td>
</tr>
<tr>
<td>Gene symbol</td>
<td>“10486218” “9205841” “8817336”</td>
</tr>
<tr>
<td>PubMed, PMID</td>
<td>“X”, “Xq13.1”</td>
</tr>
</tbody>
</table>
Annotation data packages

- The Bioconductor project provides annotation data packages, that contain many different mappings to interesting data
  - Mappings between Affy IDs and other probe IDs: hgu95av2 for HGU95Av2 GeneChip series, also, hgu133a, hu6800, mgu74a, rgu34a, etc.
  - Affy CDF data packages.
  - Probe sequence data packages.
- These packages are updated and expanded regularly as new data become available.
- They can be downloaded from the Bioconductor website and also using installDataPackage.
- DPExplorer: a widget for interacting with data packages.
- AnnBuilder: tools for building annotation data packages.

annotate: matching IDs

- Much of what annotate does relies on matching symbols.
- This is basically the role of a hash table in most programming languages.
- In R, we rely on environments.
- The annotation data packages provide R environment objects containing key and value pairs for the mappings between two sets of probe identifiers.
- Keys can be accessed using the R ls function.
- Matching values in different environments can be accessed using the get or multi get functions.

annotate: matching IDs

```R
> library(hgu95av2)
> get("41046_s_at", env = hgu95av2ACCNUM)
[1] "X95808"
> get("41046_s_at", env = hgu95av2LOCUSID)
[1] "9203"
> get("41046_s_at", env = hgu95av2SYMBOL)
[1] "ZNF261"
> get("41046_s_at", env = hgu95av2GENENAME)
[1] "zinc finger protein 261"
> get("41046_s_at", env = hgu95av2SUMFUNC)
[1] "Contains a putative zinc-binding motif (MYM)|Proteome"
> get("41046_s_at", env = hgu95av2UNIGENE)
[1] "Hs.9568"
```
annotate: matching IDs

> get("41046_s_at", env = hgu95av2CHR)
[1] "X"
> get("41046_s_at", env = hgu95av2CHRLOC)
X
-68692698
> get("41046_s_at", env = hgu95av2MAP)
[1] "Xq13.1"
> get("41046_s_at", env = hgu95av2PMID)
[1] 10486218 9205841 8817323
> get("41046_s_at", env = hgu95av2GO)
TAS          TAS          IEA
"GO:0003677" "GO:0007275" "GO:0016021"

annotate: matching IDs

- Instead of relying on the general R functions for environments, new user-friendly functions have been written for accessing and working with specific identifiers.
- E.g. getGO, getGOdesc, getLL, getPMID, getSYMBOL.

annotate: matching IDs

> getSYMBOL("41046_s_at", data="hgu95av2")
41046_s_at
"ZNF261"
> gg<- getGO("41046_s_at",data="hgu95av2")
> getGOdesc(gg[[1]], "MF")
"GO:0003677"
"DNA binding activity"
> getLL("41046_s_at",data="hgu95av2")
41046_s_at
9203
> getPMID("41046_s_at",data="hgu95av2")
"41046_s_at"
[1] 10486218 9205841 8817323
annotate: querying databases

The **annotate** package provides tools for
• Searching and processing information from various WWW biological databases
  – GenBank,
  – LocusLink,
  – PubMed.
• Regular expression searching of PubMed abstracts.
• Generating nice HTML reports of analyses, with links to biological databases.

annotate: WWW queries

• Functions for querying WWW databases from R rely on the `browseURL` function
  `browseURL("www.r-project.org")`
  Other tools: HTMLPage class, `getTDRows`, `getQueryLink`, `getQuery4UG`, `getQuery4LL`, `makeAnchor`.
• The XML package is used to parse query results.

annotate: querying GenBank


• Given a vector of GenBank accession numbers or NCBI UIDs, the `genbank` function
  – opens a browser at the URLs for the corresponding GenBank queries;
  – returns an XMLdoc object with the same data.

  `genbank("X95808", disp="browser")`
  `genbank(1430782, disp="data", type="uid")`
annotate: querying LocusLink
www.ncbi.nlm.nih.gov/LocusLink/

- locuslinkByID: given one or more LocusIDs, the browser is opened at the URL corresponding to the first gene.
  locuslinkByID("9203")

- locuslinkQuery: given a search string, the results of the LocusLink query are displayed in the browser.
  locuslinkQuery("zinc finger")
  http://www.ncbi.nlm.nih.gov/LocusLink/list.cgi?Q=zinc finger&ORG=Hs&V=0

- getQuery4LL

annotate: querying PubMed

- For any gene there is often a large amount of data available from PubMed.
- The annotate package provides the following tools for interacting with PubMed
  - pubMedAbst: a class structure for PubMed abstracts in R.
  - pubmed: the basic engine for talking to PubMed (pmidQuery).

annotate: pubMedAbst class

Class structure for storing and processing PubMed abstracts in R

- pmid
- authors
- abstText
- articleTitle
- journal
- pubDate
- abstUrl
annotate: high-level tools for querying PubMed

- `pm.getabst`: download the specified PubMed abstracts (stored in XML) and create a list of `pubMedAbst` objects.
- `pm.titles`: extract the titles from a list of PubMed abstracts.
- `pm.abstGrep`: regular expression matching on the abstracts.

annotate: PubMed example

```r
pmid <- get("41046_s_at", env=hgu95aPMID)
pumed(pm, disp="browser")

absts <- pm.getabst("41046_s_at", base="hgu95a")
pm.titles(absts)
pm.abstGrep("retardation", absts[[1]])
```

annotate: PubMed HTML report

- The new function `pmAbst2HTML` takes a list of `pubMedAbst` objects and generates an HTML report with the titles of the abstracts and links to their full page on PubMed.

  ```r
  pmAbst2HTML(absts[[1]],
               filename="pm.html")
  ```
annotate: analysis reports

• A simple interface, htmlpage, can be used to generate an HTML report of analysis results.
• The page consists of a table with one row per gene, with links to Entrez Gene, Affymetrix, SwissProt, UniGene, or OMIM.
• Entries can include various gene identifiers and statistics.
annaffy

- Provides simplified mappings between Affymetrix IDs and annotation data
- Relies on chip-level annotation packages created by AnnBuilder
- Supplies functions to produce mappings for almost all environments in a given annotation package

annaffy:Interactive

```r
> symbol <- aafSymbol(probids, "hgu95av2")
> getText(symbol)
[1] "COL11A2" "FLT3" "BDNF" "CD19" "GSTT2" "FGFR2" "IL18"
[8] "IFNB1" "RAB5B" "TAF11"

> gos <- aafGO(probids, "hgu95av2")
> gos[3]
An object of class "aafGO"
[1] An object of class "aafGOItem"
@id   "GO:0007399"
@name "neurogenesis"
@type "Biological Process"
@evid "TAS"

annaffy:Interactive

```r
> gbs <- aafGenBank(probids, "hgu95av2")
> getURL(gbs[3])
> browseURL(getURL(gbs[3]))
```
This will open a browser pointing to this particular GenBank ID
annaffy: Non-interactive

- Primary function of annaffy is to produce very nice HTML or text tables
- These tables can contain:
  - Links to databases
  - Statistics
  - Expression measures
    - Color-coded to intensity for easy viewing

annaffy: HTML Table

```r
> aaf.handler()
[1] "Probe"   "Symbol"   "Description"
[4] "Function" "Chromosome" "Chromosome Location"
[7] "GenBank"  "LocusLink"  "Cytoband"
[10] "UniGene"  "PubMed"    "Gene Ontology"

> annatable <- aafTableAnn(probids[1:10], "hgud9av2", aaf.handler()[c(1:3, 10)])
> stattable <- aafTable("t-stat" = rnorm(10), "p-value" = runif(10))
> table <- merge(annatable, stattable)
> table <- merge(table, exprtable)
> saveHTML(table, "faketable.HTML", title="Some Fake Results")
```
What is GO?

- The Gene Ontology Consortium coordinates the development and refinement of GO
- GO is a set of three ontologies for gene products
  - molecular function
  - cellular component
  - biological process

Data

- as part of Bioconductor we provide a GO package which has all the GO specific data
  - terms and relationships
  - some whole species data
- for each instrument (chip) we provide chip specific data
  - maps from the probes to GO terms
  - counts of probes per GO term + children
- constantly evolving and being updated

KEGG

- Kyoto Encyclopedia of Genes and Genomes
- Database of pathway information
- Not many genes are annotated to pathways yet
Statistics on Annotation

- Very little pathway information
- Use GO terms as a surrogate (primarily the Molecular Function and Biological Process terms)
- Use Fisher’s Exact test
- GOstats package

Fisher’s Exact

<table>
<thead>
<tr>
<th></th>
<th>Significant</th>
<th>Not Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apoptosis</td>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>Not Apoptosis</td>
<td>215</td>
<td>12383</td>
</tr>
</tbody>
</table>

GOstats:graphing

- The GOstats package can also be used to produce graphics to visualize various pathways
  - GO directed acyclic graph (DAG)
  - KEGG pathway
CoCiteStats

- Infer interrelatedness between genes using PubMed
- Again, based on the idea of a two way table

CoCiteStats

<table>
<thead>
<tr>
<th>Cite gene 1</th>
<th>Don't cite gene 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>14</td>
<td>12584</td>
</tr>
</tbody>
</table>

Cite gene 2
Don't cite gene 2
CoCiteStats: statistics

<table>
<thead>
<tr>
<th>Cite gene 1</th>
<th>Don't cite gene 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>n11</td>
<td>n12</td>
</tr>
<tr>
<td>n21</td>
<td>n22</td>
</tr>
</tbody>
</table>

Cite gene 2

Don't cite gene 2

- Concordance (n11)
- Jaccard Index (n11/n11 + n12 + n21)
- Hubert’s gamma
- p-values by permutation for all statistics