Bioconductor – MGED 2003

Sandrine Dudoit
Robert Gentleman
Outline

• reproducible research
• annotation and meta-data
• GO – more advanced usage
Reproducible Research

• A publication about scientific computing is not scholarship, it is merely an advertisement of scholarship, the scholarship lies elsewhere (Claerbout)

• Electronic journals are largely electronic only in their delivery mechanism. A few trees survive but for the author and the reader little has changed.
Reproducible Research

- most recipients of electronic documents have a computational engine available
- this suggests that we could in fact move (in a structured way) to navigable documents with dynamic content
- these documents would allow the reader to recreate (and modify) the results being reported
Early Work

- Claerbout's lab at Stanford
  - use of Makefiles
- Buckheit and Donoho (1995)
  - plots should be reproducible
- Vince Carey
  - Literate Programming
- Duncan Temple Lang
  - Literate programming
  - extensible dynamic docs
- Tony Rossini
  - Literate Data Analysis
- Fritz Leisch
  - Sweave
Compendiums

- we need to provide an entity that contains
  - text: the written content of the article(s)
  - code: computer code that will execute to provide outputs such as tables and graphics
  - data: on which the code operates and about which the text is reporting
Compendiums

- an amalgam of code, data, and text
- delivered as a single object that the user can transform into different outputs

- some outputs
  - papers suitable for publication
  - interim reports
  - long and short versions of articles
  - reports for clients etc.
Compendiums: Proof of Concept

- Sweave is a system for combining text and R code in alternating chunks
- the document looks like LaTeX but with code insterted in a special (but easy to use way)
- the document can be woven to produce a LaTeX document with all code chunks replaced by their outputs
\section{Data}

We see an interesting pattern in Figure~\ref{F1}

<<F1, fig=TRUE>>=
plot(data.x, data.y)
\@

And so we like it.

• on the left we see a section of an Sweave document

• first, standard LaTeX and then a small code chunk that is R code

• after weaving the code chunk will be replaced by the code to include the plot (which is in eps or pdf)
Compendiums: An Implementation

• the R package system provides a mechanism for both packaging together, data, code and Sweave documents and for distributing these

• with these two tools we have a proof of concept – one can carry out reproducible research with these tools

• I can give you a package that represents a paper and you can run it on your machine to reproduce that paper
Compendiums

- the concept is completely general
- given infrastructural tools (packages, distribution and transformation) any language (ie. Perl or Python) can provide these services
Annotation

• 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. LocusLink, 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
Important tasks

• Associate manufacturers or in-house probe identifiers to other available identifiers.
  E.g.

  Affymetrix IDs → LocusLink LocusID
  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).
annotate: Versioning

- it is important to keep all version information together with the mappings
- it is important to allow for new mappings to be used when they become available
- there are some interesting challenges and concerns that arise when comparing the strategies of on-line mappings versus compiled mappings
### annotate: matching IDs

<table>
<thead>
<tr>
<th>Affymetrix identifier</th>
<th>“41046_s_at”</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGU95A chips</td>
<td></td>
</tr>
<tr>
<td>LocusLink, LocusID</td>
<td>“9203”</td>
</tr>
<tr>
<td>GenBank accession #</td>
<td>“X95808”</td>
</tr>
<tr>
<td>Gene symbol</td>
<td>“ZNF261”</td>
</tr>
<tr>
<td>PubMed, PMID</td>
<td>“10486218”</td>
</tr>
<tr>
<td></td>
<td>“9205841”</td>
</tr>
<tr>
<td></td>
<td>“8817323”</td>
</tr>
<tr>
<td>Chromosomal location</td>
<td>“X”, “Xq13.1”</td>
</tr>
</tbody>
</table>
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, YG.
- 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.
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 `multiget` 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

```r
> 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"
```
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

> getS SYMBOL("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
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
  ```r
  browseURL("www.r-project.org")
  ```
  Other tools: `HTMLPage` class, `getTDRows`, `getQueryLink`, `getQuery4UG`, `getQuery4LL`, `makeAnchor`.
- The `XML` package is used to parse query results.
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**.
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**).
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)
pubmed(pmid, disp="browser")


absts <- pm.getabst("41046_s_at",
                    base="hgu95a")
pm.titles(absts)
pm.abstGrep("retardation", absts[[1]])
```
annotate: PubMed example
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")
```
<table>
<thead>
<tr>
<th>Article Title</th>
<th>Publication Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditional targeting of the DNA repair enzyme hOGG1 into mitochondria.</td>
<td>Nov 2002</td>
</tr>
<tr>
<td>Inter-individual variation, seasonal variation and close correlation of OGG1 and ERCC1 mRNA levels in full blood from healthy volunteers</td>
<td>Sep 2002</td>
</tr>
<tr>
<td>A limited association of OGG1 Ser326Cys polymorphism for adenocarcinoma of the lung</td>
<td>May 2002</td>
</tr>
<tr>
<td>Protection of human lung cells against hyperoxia using the DNA base excision repair genes hOGG1 and Fpg</td>
<td>Jul 2002</td>
</tr>
<tr>
<td>The human OGG1 DNA repair enzyme and its association with colonic cancer risk.</td>
<td>Jul 2002</td>
</tr>
<tr>
<td>Human OGG1 undergoes series phosphorylation and associates with the nuclear matrix and mitotic chromosomes in vivo.</td>
<td>Jun 2002</td>
</tr>
<tr>
<td>hOGG1 Ser(326)Cys polymorphism and modification by environmental factors of stomach cancer risk in Chinese.</td>
<td>Jul 2002</td>
</tr>
<tr>
<td>Association of the hOGG1 Ser326Cys polymorphism with lung cancer risk.</td>
<td>Jun 2002</td>
</tr>
<tr>
<td>Reciprocal &quot;shuffling&quot; underlies substrate recognition and catalytic activation by the human 8-oxo-guanine DNA glycosylase.</td>
<td>Apr 2002</td>
</tr>
<tr>
<td>Expression of 8-oxoguanine DNA glycosylase is reduced and associated with neurofibrillary tangles in Alzheimer's disease brain.</td>
<td>Mar 2002</td>
</tr>
<tr>
<td>Structure and chromosome location of human OGG1.</td>
<td>Jan 2002</td>
</tr>
<tr>
<td>Expression and differential intracellular localization of two major forms of human 8-oxoguanine DNA glycosylase encoded by alternatively spliced OGG1 mRNAs.</td>
<td>May 1999</td>
</tr>
<tr>
<td>Genetic polymorphisms and alternative splicing of the hOGG1 gene, that is involved in the repair of 8-hydroxy-2'-deoxyguanosine.</td>
<td>Jun 1993</td>
</tr>
<tr>
<td>Augmented expression of a human gene for 8-oxoguanine DNA glycosylase (MutM) in B lymphocytes of the dark zone in lymph node germinal centers.</td>
<td>Nov 1997</td>
</tr>
<tr>
<td>Opposite base-dependent reactions of a human base excision repair enzyme on DNA containing 7,8-dihydro-8-oxoguanine and abasic sites.</td>
<td>Oct 1997</td>
</tr>
<tr>
<td>Molecular cloning and functional expression of a human cDNA encoding the activator enzyme 8-hydroxyguanine-DNA glycosylase.</td>
<td>Jul 1997</td>
</tr>
<tr>
<td>Cloning and characterization of hOGG1, a human homolog of the OGG1 gene of Saccharomyces cerevisiae.</td>
<td>Jul 1997</td>
</tr>
</tbody>
</table>

**pm.html**

pmAbst2html function from annotate package
annotate: analysis reports

• A simple interface, ll.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 LocusLink.

• Entries can include various gene identifiers and statistics.
## BioConductor Gene Listing

Golub et al. data, genes with permutation maxT adjusted p-value < 0.01

### Locus Link Genes

<table>
<thead>
<tr>
<th>LocusID</th>
<th>Gene name</th>
<th>Chromosome</th>
<th>ALL mean</th>
<th>AML mean</th>
<th>t-statistic</th>
<th>raw p-value</th>
<th>adj p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>7791</td>
<td>X95735_at</td>
<td>7</td>
<td>0.295</td>
<td>1.59</td>
<td>-10.6</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>1471</td>
<td>E27891_at</td>
<td>20</td>
<td>-0.81</td>
<td>2.08</td>
<td>-9.78</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>1481</td>
<td>M55150_at</td>
<td>15</td>
<td>0.488</td>
<td>1.24</td>
<td>-8.03</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>4078</td>
<td>M10930_at</td>
<td>8</td>
<td>-0.284</td>
<td>1.11</td>
<td>7.98</td>
<td>2e-05</td>
<td>0.00016</td>
</tr>
<tr>
<td>5234</td>
<td>G02290_s_at</td>
<td>11</td>
<td>-0.162</td>
<td>1.36</td>
<td>-9.77</td>
<td>2e-05</td>
<td>2e-04</td>
</tr>
<tr>
<td>6929</td>
<td>M21233_at</td>
<td>19</td>
<td>0.855</td>
<td>0.391</td>
<td>7.59</td>
<td>2e-05</td>
<td>2e-04</td>
</tr>
<tr>
<td>5926</td>
<td>X74262_at</td>
<td>1</td>
<td>0.669</td>
<td>0.565</td>
<td>7.42</td>
<td>2e-05</td>
<td>0.000036</td>
</tr>
<tr>
<td>7153</td>
<td>Z1511S_at</td>
<td>3</td>
<td>1.94</td>
<td>3.945</td>
<td>7.28</td>
<td>2e-05</td>
<td>0.001</td>
</tr>
<tr>
<td>26999</td>
<td>L47779_at</td>
<td>5</td>
<td>0.784</td>
<td>0.779</td>
<td>7.21</td>
<td>2e-05</td>
<td>0.00014</td>
</tr>
<tr>
<td>4502</td>
<td>U23762_at</td>
<td>6</td>
<td>1.86</td>
<td>0.294</td>
<td>7.28</td>
<td>2e-05</td>
<td>0.00016</td>
</tr>
<tr>
<td>65108</td>
<td>H01612-H01612_at</td>
<td>1</td>
<td>1.91</td>
<td>0.889</td>
<td>7.11</td>
<td>2e-05</td>
<td>0.0017</td>
</tr>
<tr>
<td>58</td>
<td>M41432_at</td>
<td>1</td>
<td>0.431</td>
<td>0.771</td>
<td>7.08</td>
<td>2e-05</td>
<td>0.0018</td>
</tr>
<tr>
<td>5623</td>
<td>L41879_at</td>
<td>13</td>
<td>0.438</td>
<td>3.4</td>
<td>7.08</td>
<td>2e-05</td>
<td>0.0018</td>
</tr>
<tr>
<td>3548</td>
<td>U12988_at</td>
<td>1A</td>
<td>-0.097</td>
<td>-1.07</td>
<td>7.07</td>
<td>2e-05</td>
<td>0.0018</td>
</tr>
<tr>
<td>7939</td>
<td>X15214_at</td>
<td>5</td>
<td>1.92</td>
<td>1.07</td>
<td>7.06</td>
<td>2e-05</td>
<td>0.00186</td>
</tr>
<tr>
<td>4036</td>
<td>U50136_rna1_at</td>
<td>5</td>
<td>0.71</td>
<td>1.51</td>
<td>-6.97</td>
<td>2e-05</td>
<td>0.00232</td>
</tr>
<tr>
<td>5847</td>
<td>Y12670_at</td>
<td>1</td>
<td>-0.157</td>
<td>0.892</td>
<td>-6.96</td>
<td>2e-05</td>
<td>0.00238</td>
</tr>
<tr>
<td>3976</td>
<td>H00789_at</td>
<td>1</td>
<td>0.611</td>
<td>0.6183</td>
<td>5.85</td>
<td>2e-05</td>
<td>0.00238</td>
</tr>
<tr>
<td>3976</td>
<td>H00789_at</td>
<td>1</td>
<td>0.571</td>
<td>2.02</td>
<td>6.87</td>
<td>2e-05</td>
<td>0.00286</td>
</tr>
<tr>
<td>7576</td>
<td>H00789_at</td>
<td>1</td>
<td>0.413</td>
<td>0.982</td>
<td>6.66</td>
<td>2e-05</td>
<td>0.00286</td>
</tr>
<tr>
<td>1724</td>
<td>U20426_at</td>
<td>10</td>
<td>-0.229</td>
<td>-1.16</td>
<td>5.85</td>
<td>4e-05</td>
<td>0.00294</td>
</tr>
<tr>
<td>2202</td>
<td>U88278_at</td>
<td>7</td>
<td>0.64</td>
<td>0.501</td>
<td>6.92</td>
<td>2e-05</td>
<td>0.00306</td>
</tr>
<tr>
<td>1205</td>
<td>M23197_at</td>
<td>19</td>
<td>-0.991</td>
<td>0.354</td>
<td>6.79</td>
<td>2e-05</td>
<td>0.0033</td>
</tr>
<tr>
<td>1309</td>
<td>M58358_at</td>
<td>11</td>
<td>1.21</td>
<td>2.12</td>
<td>6.77</td>
<td>2e-05</td>
<td>0.00344</td>
</tr>
<tr>
<td>6923</td>
<td>M81299_at</td>
<td>14</td>
<td>1.13</td>
<td>1.32</td>
<td>6.78</td>
<td>2e-05</td>
<td>0.00352</td>
</tr>
<tr>
<td>957</td>
<td>X06254_rna1_at</td>
<td>12</td>
<td>0.6513</td>
<td>1.33</td>
<td>-6.76</td>
<td>2e-05</td>
<td>0.00352</td>
</tr>
<tr>
<td>5318</td>
<td>X07745_at</td>
<td>2</td>
<td>-0.939</td>
<td>0.553</td>
<td>-6.74</td>
<td>2e-05</td>
<td>0.00378</td>
</tr>
<tr>
<td>12343</td>
<td>M10321_at</td>
<td>12</td>
<td>0.109</td>
<td>0.935</td>
<td>6.71</td>
<td>2e-05</td>
<td>0.00404</td>
</tr>
<tr>
<td>7326</td>
<td>U62136_at</td>
<td>8</td>
<td>-0.153</td>
<td>0.92</td>
<td>6.68</td>
<td>2e-05</td>
<td>0.00428</td>
</tr>
<tr>
<td>3560</td>
<td>X15948_at</td>
<td>4</td>
<td>0.541</td>
<td>1.33</td>
<td>6.61</td>
<td>2e-05</td>
<td>0.00492</td>
</tr>
<tr>
<td>6924</td>
<td>H00244_at</td>
<td>14</td>
<td>0.863</td>
<td>0.365</td>
<td>6.14</td>
<td>2e-05</td>
<td>0.00492</td>
</tr>
</tbody>
</table>
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
• the relationship between gene products and BP, CC, MF are all many to many
• a child term may have one or more parent terms
• *transmembrane receptor protein-tyrosine kinase* is child of both *transmembrane receptor* and *protein tyrosine kinase*
GO Parent-Child

• the relationship between a parent and a child term can be either an *is-a* relationship or a *part-of* relationship

• a *mitotic chromosome* is a *chromosome*

• a *telomere* is a *part-of a chromosome*

• the child term is more specific than the parent term
GO Graphs

• GO itself has no reference to genes
• GO specifies a terminology and the relationships between terms
• each GO term is associated with a single node (so I will use the words term and node interchangeably) in the DAG
GO and Genes

• so GO as described above is a set of terms
• as such it can be used as the basis for searching relevant literature (McCray et al)
• but its real power comes from the annotation of specific genes and gene products at the different terms
• this is carried out by many organizations using criteria proposed by GO
GO and Genes

• a gene is annotated at one or more terms
• for each term the annotation must be supported by evidence and the evidence code is available (e.g)
  – TAS: traceable author statement
  – IEP: inferred from expression pattern
  – ISS: inferred from sequence similarity
• and many others
Data

• as part of Bioconductor we proved 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
GO Data

• for any gene obtain the most specific GO labels that gene is annotated at
• using these terms and the GO structure obtain the graph that has nodes representing those terms and all parents and edges for all child parent relationships
• this is called the *induced GO graph* or just the *GO graph*
• BP, CC and MF all induce different graphs
• ABL1 has Affymetrix identifier 1635_at
• this is annotated at
  GO:0004713 protein tyrosine kinase
  GO:0003677 DNA binding
• we then use the GO structure to produce the plot
Analysis: What Can We Do?

• we can use GO to provide annotations for lists or clusters of genes
• we can use GO to provide sets of genes with specific properties (or relationships)
• We can define distances between GO terms using the graph structure
• we can define distances between genes using GO and other data
ALL Example

• ALL experiment, 93 patients (courtesy Ritz, Foa, Chiaretti)

• selected genes that could differentiate three groups, ALL1/AF4, BCR/ABL, NEG

• this yielded 136 probes and 129 unique LocusLink ids of these 90 have GO MF annotation

• are there MF terms that are over represented in this list of genes?
ALL Example

• for the 129 genes there were a total of 192 MF terms in the induced graph
• each of these categories had probes annotated at it (spread from 1 to 9478; 37 had 10 or fewer probes)
ALL Example

• for each GO node the set of probes annotated at that node was determined
• for each probe the group (ALL1/AF4, BCR/ABL, NEG) with the highest mean was determined
• finally the group that had the most "highest means" was determined
ALL Example

- the induced MF graph was plotted
- nodes were colored as follows:
  - ALL1/AF4: red (66)
  - BCR/ABL: blue (91)
  - NEG: green (11)
  - no winner: white (24)
ALL Example
MF Graph
Relating Terms to Gene Lists

- suppose that we have a list of $n$ interesting genes (derived in any old way)
- for each GO term (in each ontology) we can ask whether the genes in the list are over-represented at that node
- this question can also be phrased in terms of a test of homogeneity (2-way table)
Terms to Gene Lists

- consider all genes assayed (or all genes expressed may be more relevant), $N$
- we have an urn with $N$ balls, $n$ of them are white (the interesting ones) and $N-n$ are black
- for a GO term we have $k$ genes annotated at that term
- this is like $k$ draws from the Urn and we ask whether we got more white balls than expected ($x=$number of white balls)
Terms to Gene Lists

- this is simply a Hypergeometric calculation

- issues:
  - multiple testing
  - lack of independence: genes are annotated at parents and children
  - can we (should we) take account of the GO hierarchy?
  - GO terms with too many genes (not specific)
  - GO terms with too few genes (not interesting)
  - shouldn't the genes all be interesting in the same way?
ALL Example

- for each MF category a Hypergeometric test was performed
- \( N=6422, \ n=90, \) for each term we found the number of unique LocusLink IDs annotated at that term were determined (this was \( k \))
- 8 nodes with \( p<0.01 \) and 30 nodes with \( p<0.05 \)
- we will explore the 8 nodes
## ALL: 8 GO Terms

<table>
<thead>
<tr>
<th>TERM</th>
<th>DESCRIPTION</th>
<th>k</th>
<th>x</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0005515</td>
<td>protein binding</td>
<td>800</td>
<td>22</td>
<td>0.0012</td>
</tr>
<tr>
<td>GO:0003821</td>
<td>class II major histocompatibility complex antigen</td>
<td>9</td>
<td>5</td>
<td>6e-8</td>
</tr>
<tr>
<td>GO:0003779</td>
<td>actin binding</td>
<td>111</td>
<td>7</td>
<td>9e-4</td>
</tr>
<tr>
<td>GO:0008092</td>
<td>cytoskeletal protein binding</td>
<td>155</td>
<td>8</td>
<td>0.0014</td>
</tr>
<tr>
<td>GO:0004601</td>
<td>peroxidase</td>
<td>20</td>
<td>3</td>
<td>0.0026</td>
</tr>
<tr>
<td>GO:0016684</td>
<td>oxidoreductase, acting on peroxide as acceptor</td>
<td>20</td>
<td>3</td>
<td>0.0026</td>
</tr>
<tr>
<td>GO:0045012</td>
<td>MHC class II receptor</td>
<td>4</td>
<td>2</td>
<td>0.0011</td>
</tr>
<tr>
<td>GO:0005095</td>
<td>GTPase inhibitor</td>
<td>6</td>
<td>2</td>
<td>0.0028</td>
</tr>
</tbody>
</table>
Using the GO Structure

• notice that the sequence
  3779->8092->5515
• has decreasing p-values
  .001 -> .002 -> .009
• evidence: 7/111; 8/155; 22/800
• how do we interpret this?
• set up as a series of nested 2 by 2 tables we might make some progress (log-rank)
Clustering and GO

- another way to view the previous test is as a two-way table and a test of homogeneity

<table>
<thead>
<tr>
<th>Node\Interesting</th>
<th>YES</th>
<th>NO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>5</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>NO</td>
<td>85</td>
<td>6328</td>
<td>6413</td>
</tr>
<tr>
<td>Total</td>
<td>90</td>
<td>6332</td>
<td>6422</td>
</tr>
</tbody>
</table>

- p-value=5e-8
The induced GO graph for all nodes with $p < 0.05$.

Nodes are colored:

- red: $p < 0.01$
- pink: $p < 0.05$
- white: all others
Using the GO Structure

• do we take that as stronger evidence in favor of an interesting effect than if there was no gradient?
• what about the child-parent relationships, are is-a and has-a important?
• are we happier if at least one of the is-a children show a similar effect?
Issues

• it will be important in some contexts to account for and adjust for the evidence on which an annotation was based
• for example if exploring sequence similarity as it relates to function all ISS based annotations should be excluded
Conclusions

• GO and the various collaborators have provided a very rich data set which has the potential to add meaning to data analyses

• there are a number of ways of using this data and it is not yet clear which will be most beneficial

• it is clear that we need better tools for working with the data
Acknowledgements

- Vincent Carey
- Steve North
- Emden Gansner
- Debby Swayne
- Duncan Temple Lang
- Sabina Chiaretti
- J. Ritz
- Jeff Gentry
- Jianhua Zhang
- Denise Scholtens
- Beiying Ding
- Elizabeth Whalen
- Cheng Li
- R. Foa