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 Analsysis
- 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

Sweave

```
\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

annotate: matching IDs

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 import 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

Affymetrix identifier HGU95A chips	`"41046_s_at"`
LocusLink, LocusID	`"9203"`
GenBank accession #	`"X95808"`
Gene symbol	`"ZNF261"`

PubMed, PMID	`"10486218"`
	`"9205841"`
	`"8817323"`

Chromosomal location	`"X", "Xq13.1"`

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`**, **`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.

`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 `multiget` functions.

annotate: matching IDs

```
> 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

www.ncbi.nlm.nih.gov/Genbank/index.html

- 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")
```

http://www.ncbi.nih.gov/entrez/query.fcgi?tool=bioconductor&cmd=Search&db=Nucleotide&term=X95808

```
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")
```

http://www.ncbi.nih.gov/LocusLink/LocRpt.cgi?l=9203

- **locuslinkQuery**: given a search string, the results of the LocusLink query are displayed in the browser.

```
locuslinkQuery("zinc finger")
```

http://www.ncbi.nih.gov/LocusLink/list.cgi?Q=zinc finger&ORG=Hs&V=0

- **getQuery4LL**.

annotate: querying PubMed

www.ncbi.nlm.nih.gov

- 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

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

http://www.ncbi.nih.gov/entrez/query.fcgi?tool=bioconductor&cmd=Retrieve&db=PubMed&list_uids=10486218%2c9205841%2c8817323

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

annotate: PubMed example

RGui - [R Console]

File Edit Misc Packages Windows Help

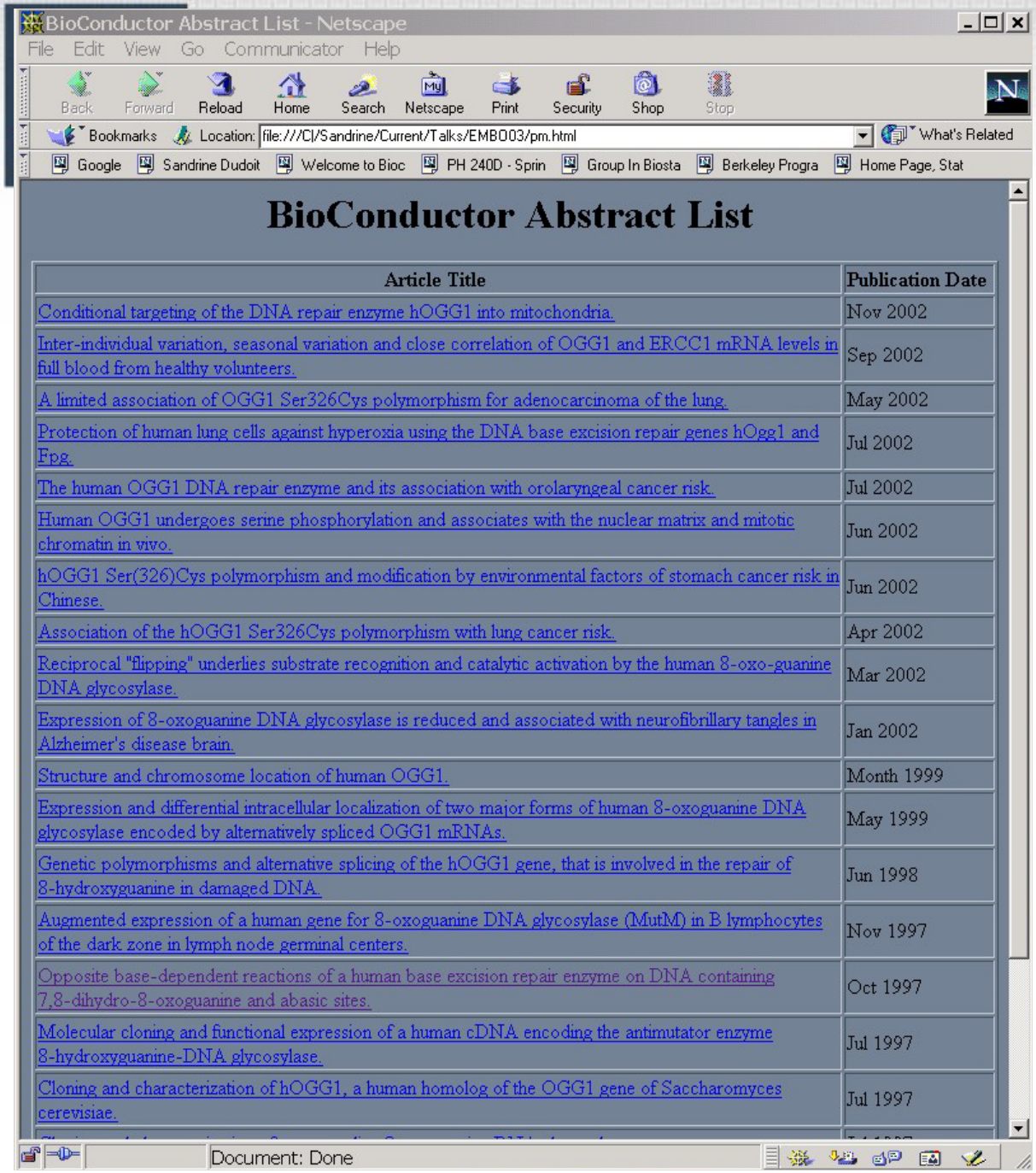
```
Slot "articleTitle":
[1] "Prediction of the coding sequences of unidentified human genes. VII. The complete sequences of 100 new cDNA clones from brain which can$

Slot "journal":
[1] "DNA Res"

Slot "pubDate":
[1] "Apr 1997"

Slot "abstUrl":
[1] "No URL Provided"


[[3]]
An object of class "pubMedAbst"
Slot "authors":
[1] "S M SM van der Maarel" "I H IH Scholten"       "I I Huber"             "C C Philippe"          "R F RF Suijkerbuijk"
[6] "S S Gilgenkrantz"      "J J Kere"              "F P FP Cremers"        "H H HH Ropers"

Slot "abstText":
[1] "In several families with non-specific X-linked mental retardation (XLMR) linkage analyses have assigned the underlying gene defect to t$

Slot "articleTitle":
[1] "Cloning and characterization of DXS6673E, a candidate gene for X-linked mental retardation in Xq13.1."

Slot "journal":
[1] "Hum Mol Genet"

Slot "pubDate":
[1] "Jul 1996"

Slot "abstUrl":
[1] "No URL Provided"


> pm.titles(absts)
[[1]]
[1] "Cloning and mapping of members of the MYM family."                                                                                $
[2] "Prediction of the coding sequences of unidentified human genes. VII. The complete sequences of 100 new cDNA clones from brain which can$
[3] "Cloning and characterization of DXS6673E, a candidate gene for X-linked mental retardation in Xq13.1."                           $

> pm.abstGrep("retardation",absts[[1]])
[1]  TRUE FALSE  TRUE
>
```

R 1.5.1 - A Language and Environment

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.

```
pmAbst2HTML(absts[[1]],
    filename="pm.html")
```

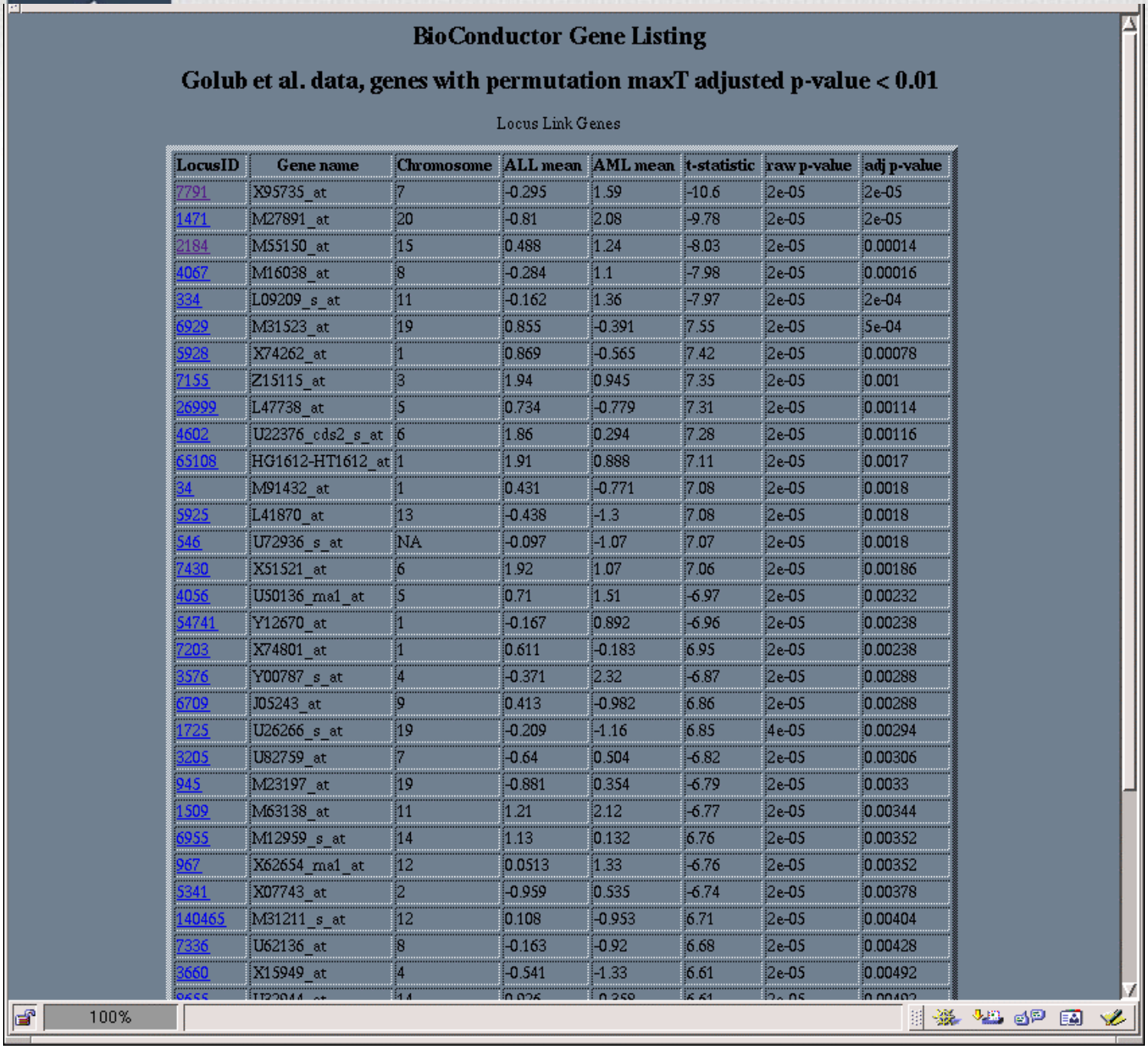


`pmAbst2html` function from **`annotate`** package

pm.html

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

LocusID	Gene name	Chromosome	ALL mean	AML mean	t-statistic	raw p-value	adj p-value
7791	X95735_at	7	-0.295	1.59	-10.6	2e-05	2e-05
1471	M27891_at	20	-0.81	2.08	-9.78	2e-05	2e-05
2184	M55150_at	15	0.488	1.24	-8.03	2e-05	0.00014
4067	M16038_at	8	-0.284	1.1	-7.98	2e-05	0.00016
334	L09209_s_at	11	-0.162	1.36	-7.97	2e-05	2e-04
6929	M31523_at	19	0.855	-0.391	7.55	2e-05	5e-04
5928	X74262_at	1	0.869	-0.565	7.42	2e-05	0.00078
7155	Z15115_at	3	1.94	0.945	7.35	2e-05	0.001
26999	L47738_at	5	0.734	-0.779	7.31	2e-05	0.00114
4602	U22376_cds2_s_at	6	1.86	0.294	7.28	2e-05	0.00116
65108	HG1612-HT1612_at	1	1.91	0.888	7.11	2e-05	0.0017
34	M91432_at	1	0.431	-0.771	7.08	2e-05	0.0018
5925	L41870_at	13	-0.438	-1.3	7.08	2e-05	0.0018
546	U72936_s_at	NA	-0.097	-1.07	7.07	2e-05	0.0018
7430	X51521_at	6	1.92	1.07	7.06	2e-05	0.00186
4056	U50136_rna1_at	5	0.71	1.51	-6.97	2e-05	0.00232
54741	Y12670_at	1	-0.167	0.892	-6.96	2e-05	0.00238
7203	X74801_at	1	0.611	-0.183	6.95	2e-05	0.00238
3576	Y00787_s_at	4	-0.371	2.32	-6.87	2e-05	0.00288
6709	J05243_at	9	0.413	-0.982	6.86	2e-05	0.00288
1725	U26266_s_at	19	-0.209	-1.16	6.85	4e-05	0.00294
3205	U82759_at	7	-0.64	0.504	-6.82	2e-05	0.00306
945	M23197_at	19	-0.881	0.354	-6.79	2e-05	0.0033
1509	M63138_at	11	1.21	2.12	-6.77	2e-05	0.00344
6955	M12959_s_at	14	1.13	0.132	6.76	2e-05	0.00352
967	X62654_rna1_at	12	0.0513	1.33	-6.76	2e-05	0.00352
5341	X07743_at	2	-0.959	0.535	-6.74	2e-05	0.00378
140465	M31211_s_at	12	0.108	-0.953	6.71	2e-05	0.00404
7336	U62136_at	8	-0.163	-0.92	6.68	2e-05	0.00428
3660	X15949_at	4	-0.541	-1.33	6.61	2e-05	0.00492
9655	U22944_at	14	0.926	0.269	6.61	2e-05	0.00492

100%

`ll.htmlpage` function from `annotate` package

genelist.html

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

GO

- 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

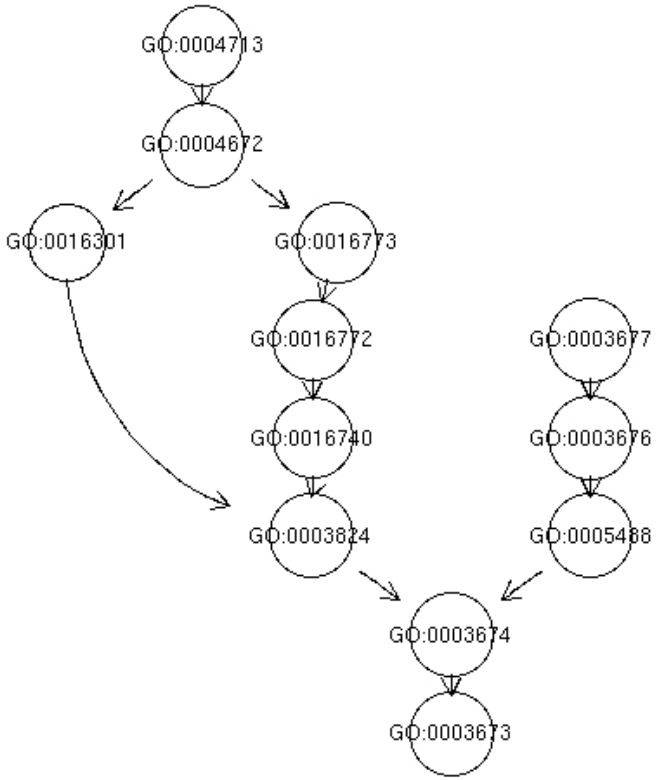
ABL 1

- 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

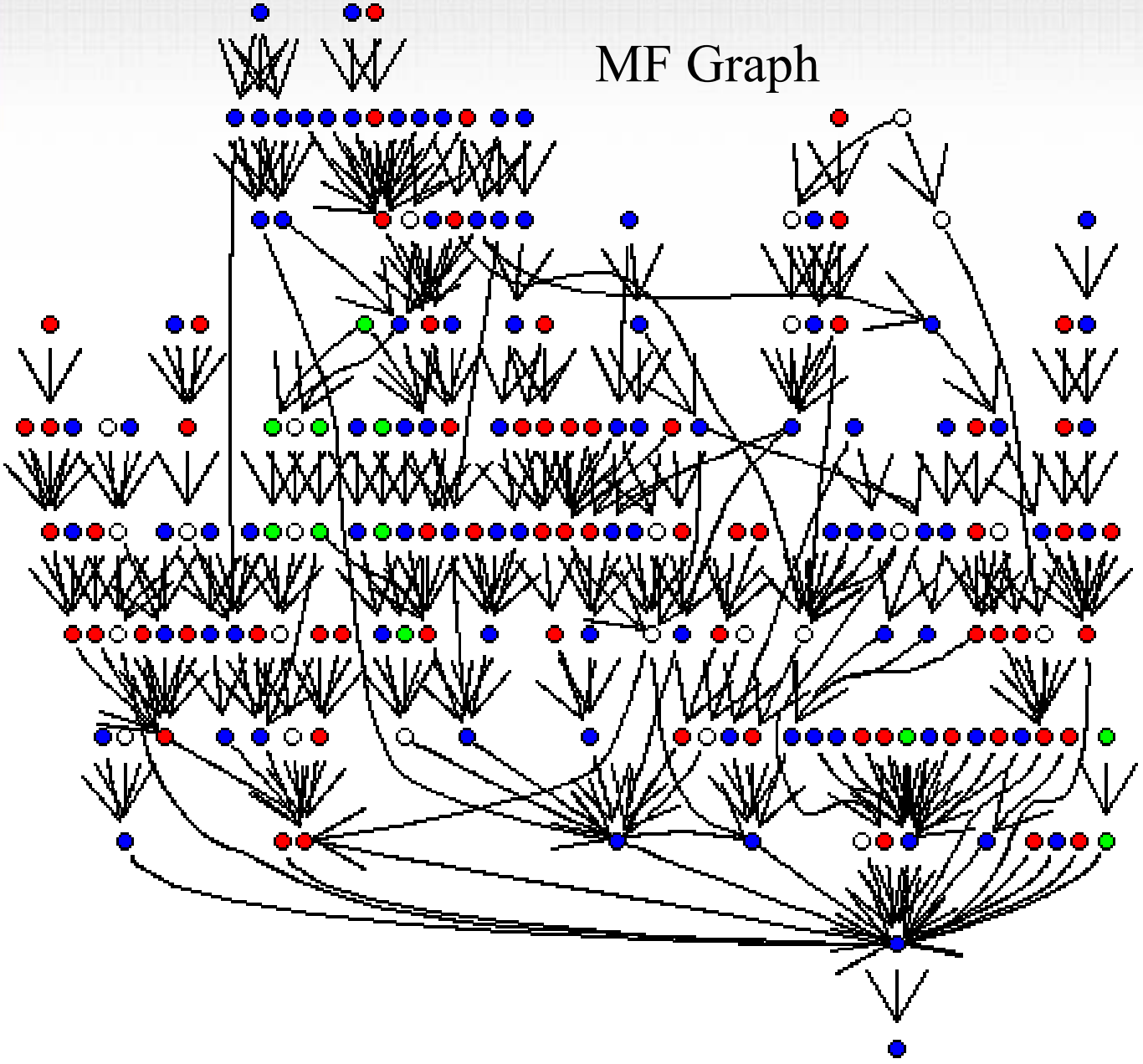
- for each GO node the set of probes annotated at that node was determined
- for each probe the group (ALL1/AF4, BCR/ABL, NEG) with the highest mean was determined
- finally the group that had the most "highest means" was determined

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?

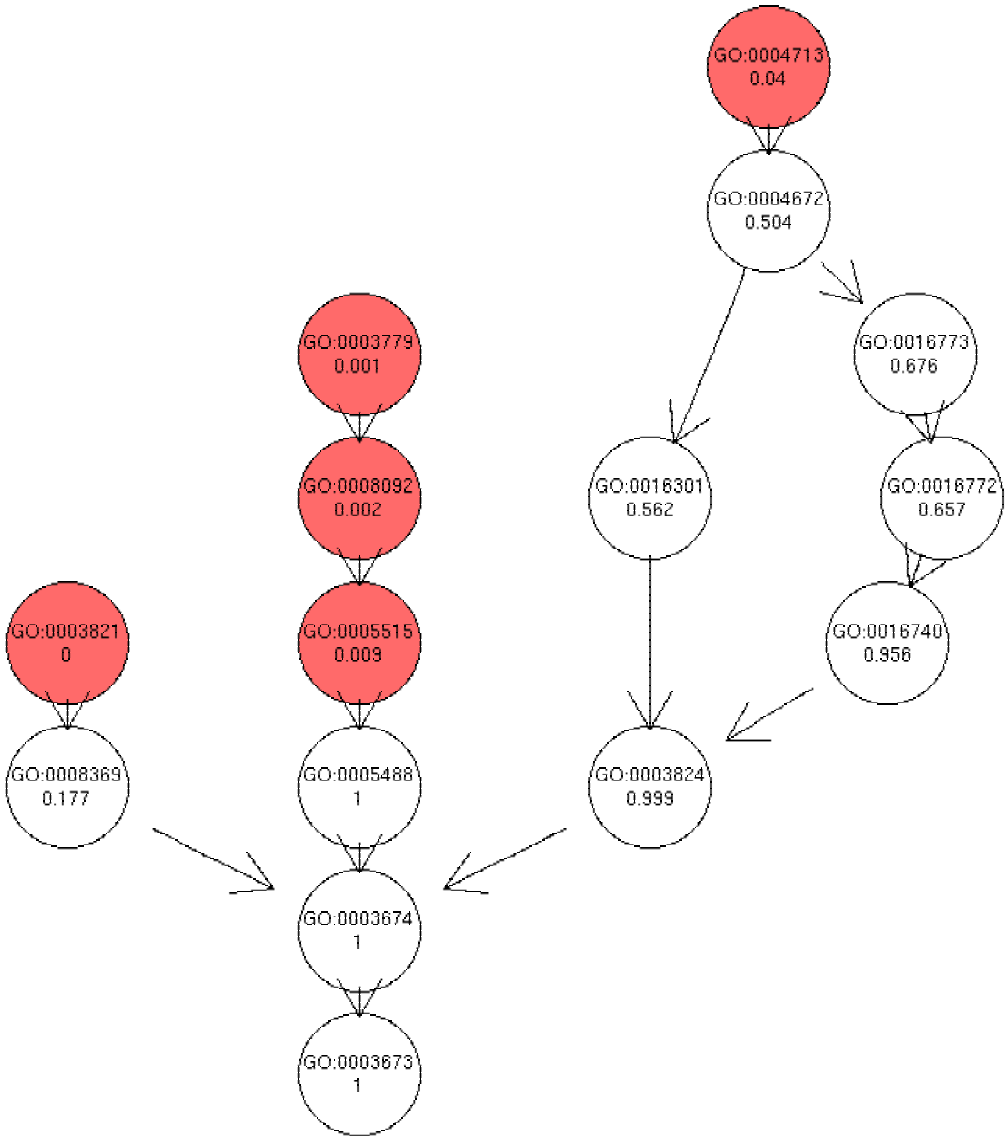
ALL Example

- 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

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

GO graph for ALL Example



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

Node\Interesting	YES	NO	Total
YES	5	4	9
NO	85	6328	6413
Total	90	6332	6422

- p-value=5e-8

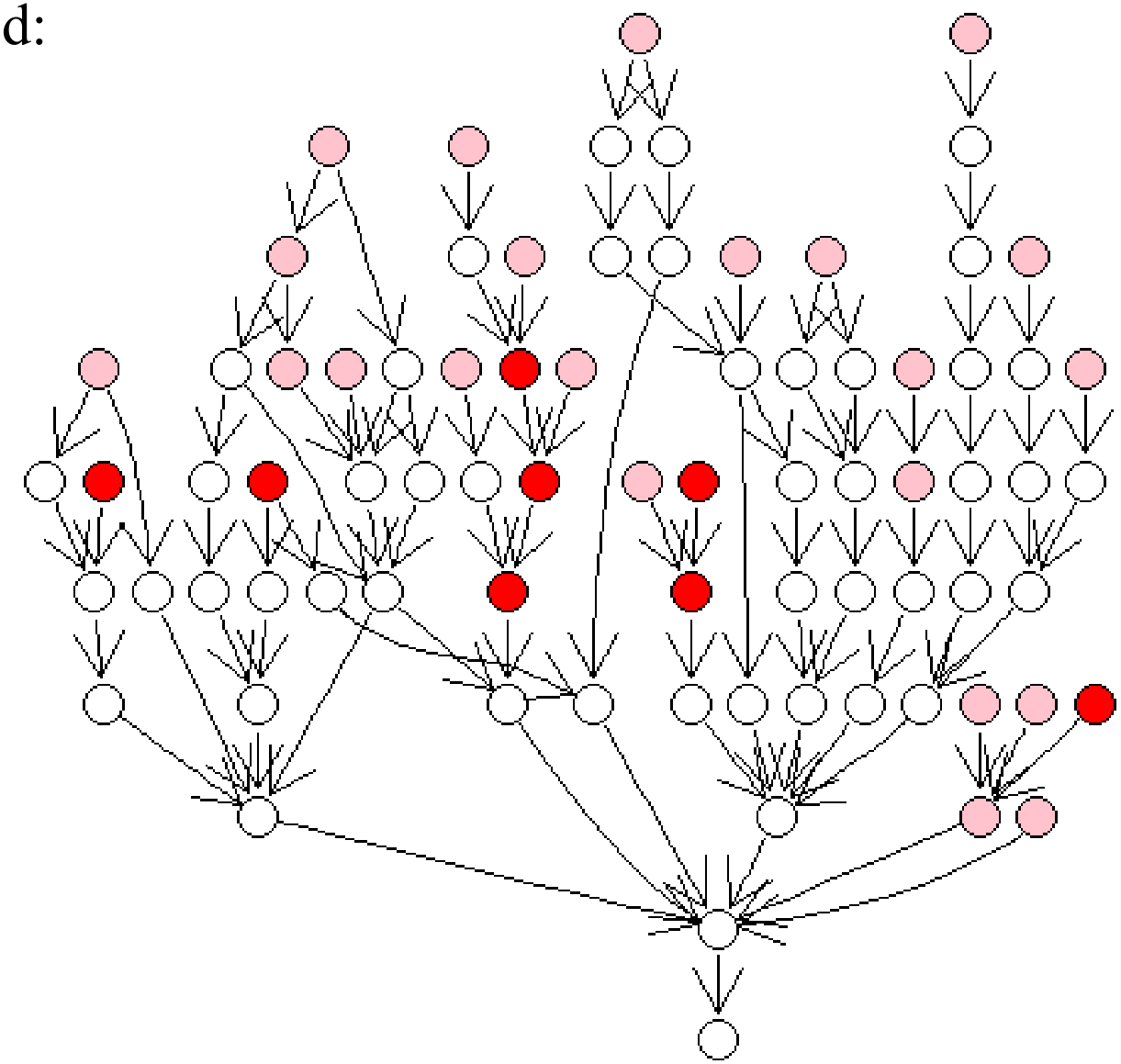
The induced GO graph for all nodes with $p < 0.05$.

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