Teaching with Bioconductor for statistical analysis of genome-scale data: Software, Documents, Experiments

VJ Carey, PhD, Channing Lab
Harvard Medical School

- Context: Some ambitions of genome-scale experimentalists
- Bioconductor's tasks:
  - Reduce barriers to entry for statisticians
  - Foster software reliability, analysis reproducibility
  - Ease biologists' use of modern statistical tools
- Vehicles for teaching, developing

Ambitions I: structural theories of gene function

RESEARCH

A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules

Joshua M. Stuart,[1*†] Eran Segal,[2*] Daphne Koller,[2‡]
Stuart K. Kim[3‡]

To elucidate gene function on a global scale, we identified pairs of genes that are coexpressed over 3182 DNA microarrays from humans, flies, worms, and yeast. We found 22,163 such coexpression relationships, each of which has been conserved across evolution. This conservation implies that the coexpression of these gene pairs confers a selective advantage and therefore that these genes are functionally related. Many of these relationships provide strong evidence for the involvement of new genes in core biological functions such as the cell cycle, secretion, and protein expression. We experimentally confirmed the predictions implied by some

Design: large scale integration of published arrays

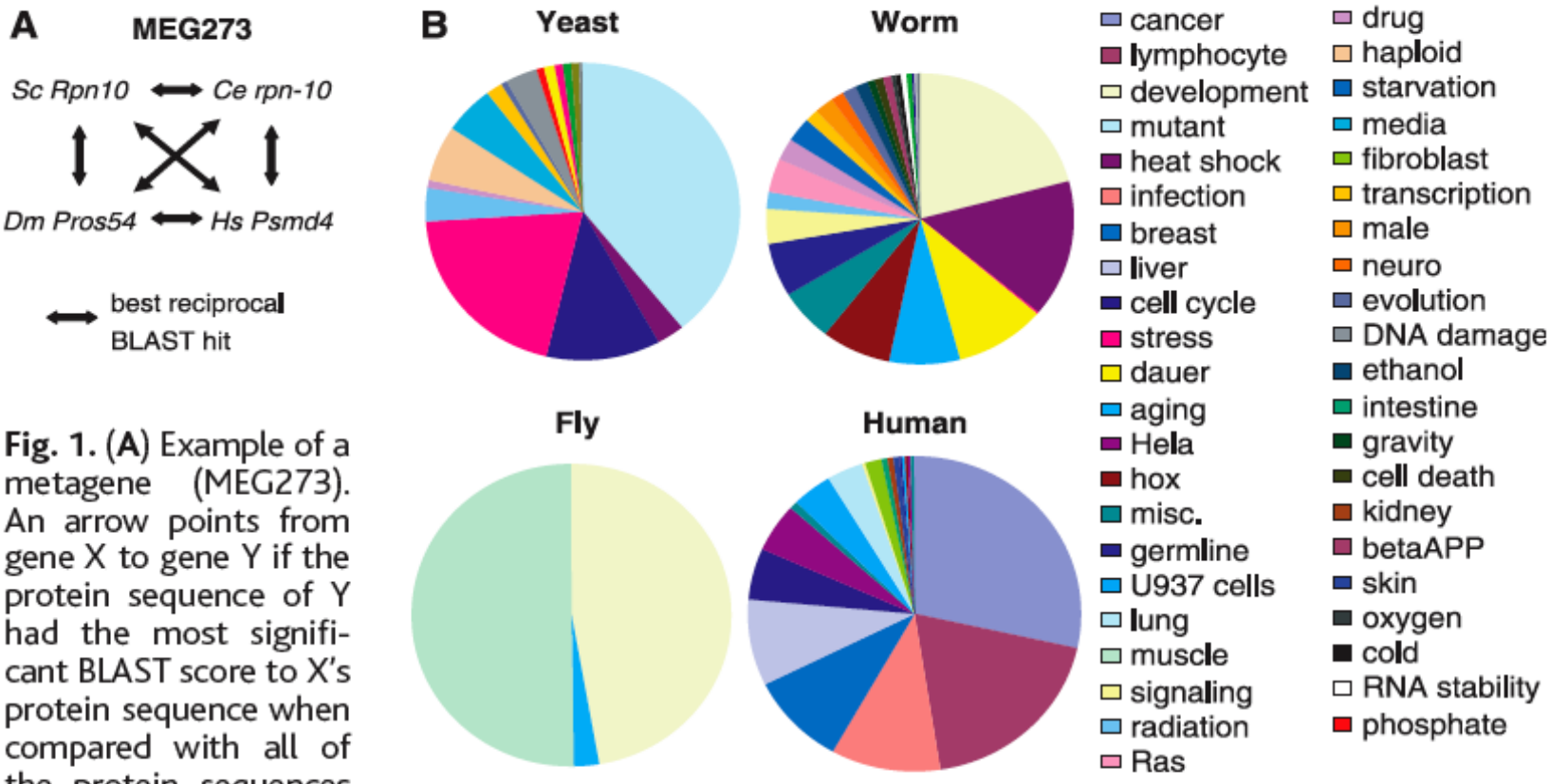
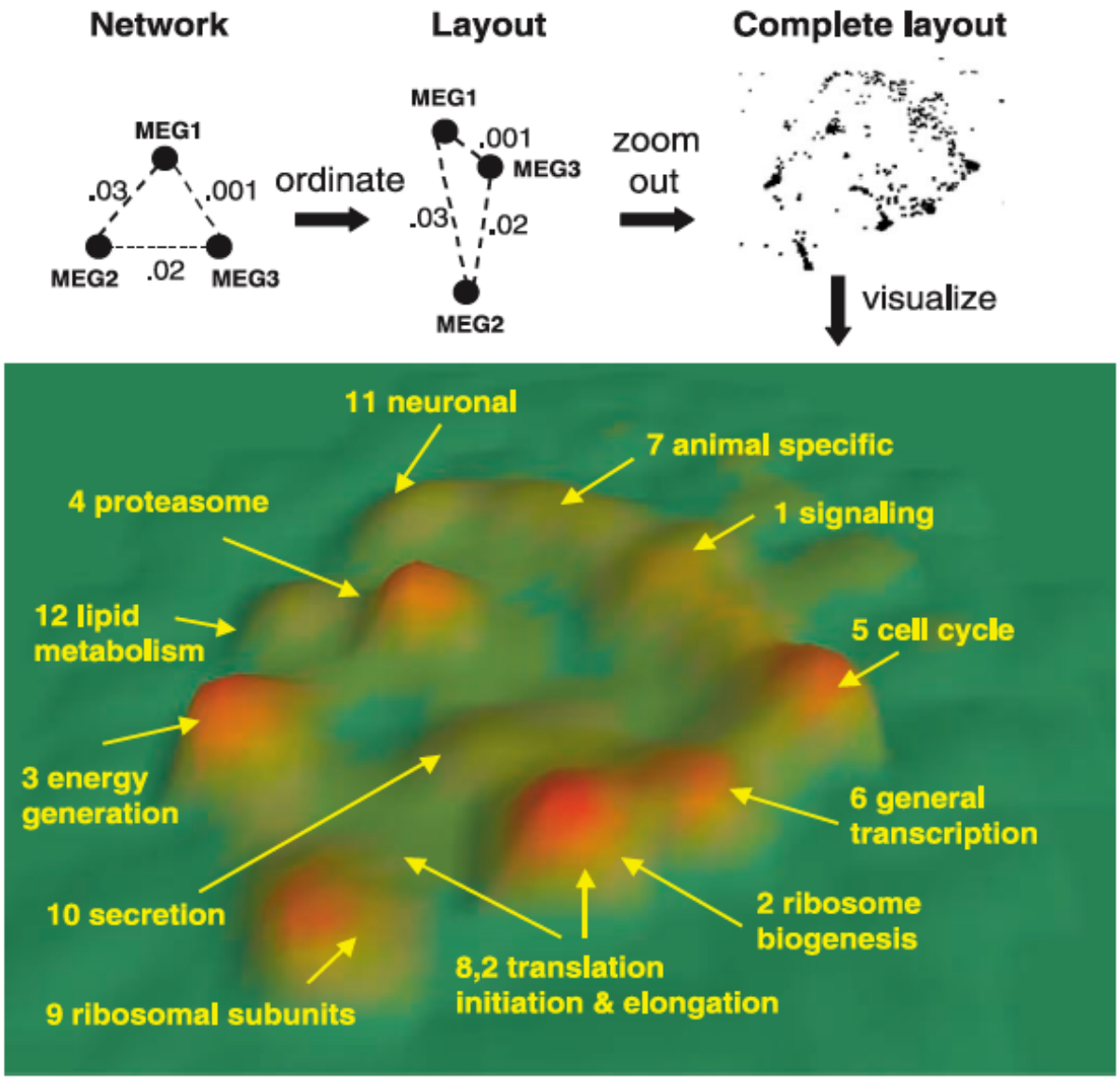


Fig. 1. (**A**) Example of a metagene (MEG273). An arrow points from gene X to gene Y if the protein sequence of Y had the most significant BLAST score to X's protein sequence when compared with all of the protein sequences in Y's database. (**B**) Compendiums of microarray expression data from four organisms included in the analysis. Color shows the type of DNA microarray experiment.

Product: coexpression implies shared role

- Cross-organism integration: metagenes formed by reciprocal BLAST
- Cross-experiment integration: metagene constituent expression level vectors are obtained
- Measure of coexpression of two metagenes is inversely proportional to the p-value for correlation tests computed for their constituents over experiments
- An ordination is constructed so that highly coexpressed metagenes are at small distances on the plane
- Clusters of genes in plane assessed for functional characteristics


Network
Layout
Complete layout
MEG1
.03
.001
MEG2
.02
MEG3
ordinate
MEG1
.001
MEG3
.03
.02
MEG2
zoom
out
visualize
11 neuronal
7 animal specific
4 proteasome
1 signaling
12 lipid
metabolism
5 cell cycle
3 energy
generation
6 general
transcription
10 secretion
2 ribosome
biogenesis
8,2 translation
initiation & elongation
9 ribosomal subunits

Issues for statistical educators

- Has bias been avoided? Hard to formalize.
  - Risk of (some) circularity – annotation at hump derived from experiments used to establish the GO annotation?
- Has information been used efficiently?
  - Arbitrary thresholds employed to distinguish coexpressed gene pairs
  - Proximities based on 'p-values'
  - Single predominant category assigned to each cluster in the coexpression graph
- Inspiring analysis, statistical rigor modest at best
- Question: How would you check the work?

Ambitions II: Reliable methodology

- MAQC, Nat Biotech Sept 2006

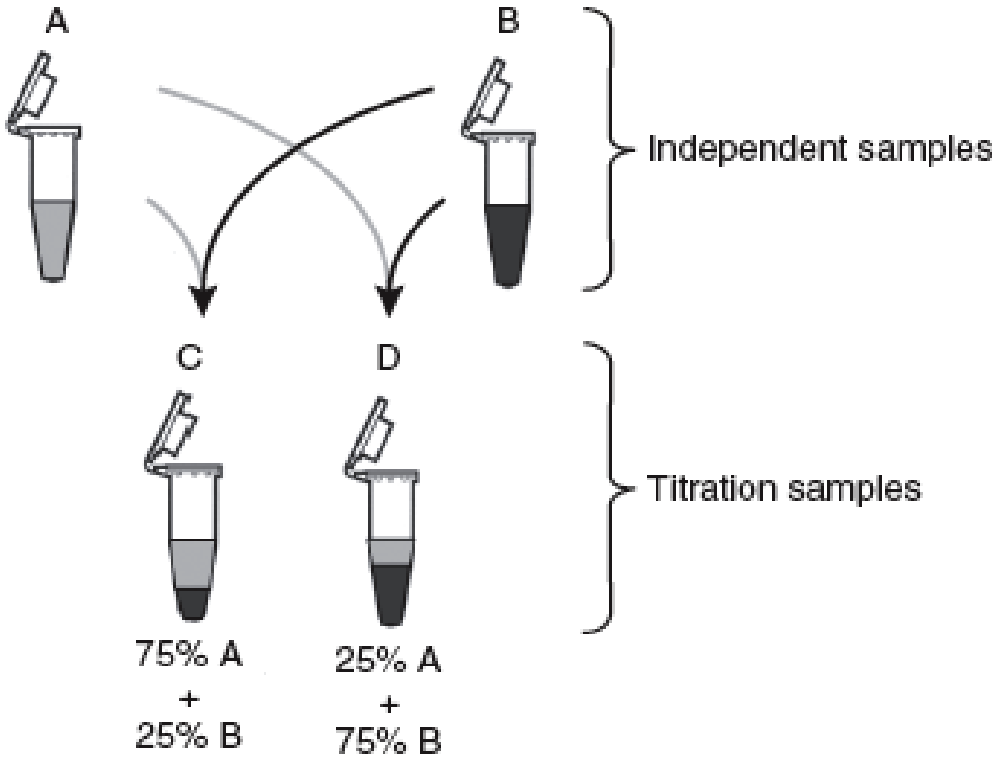
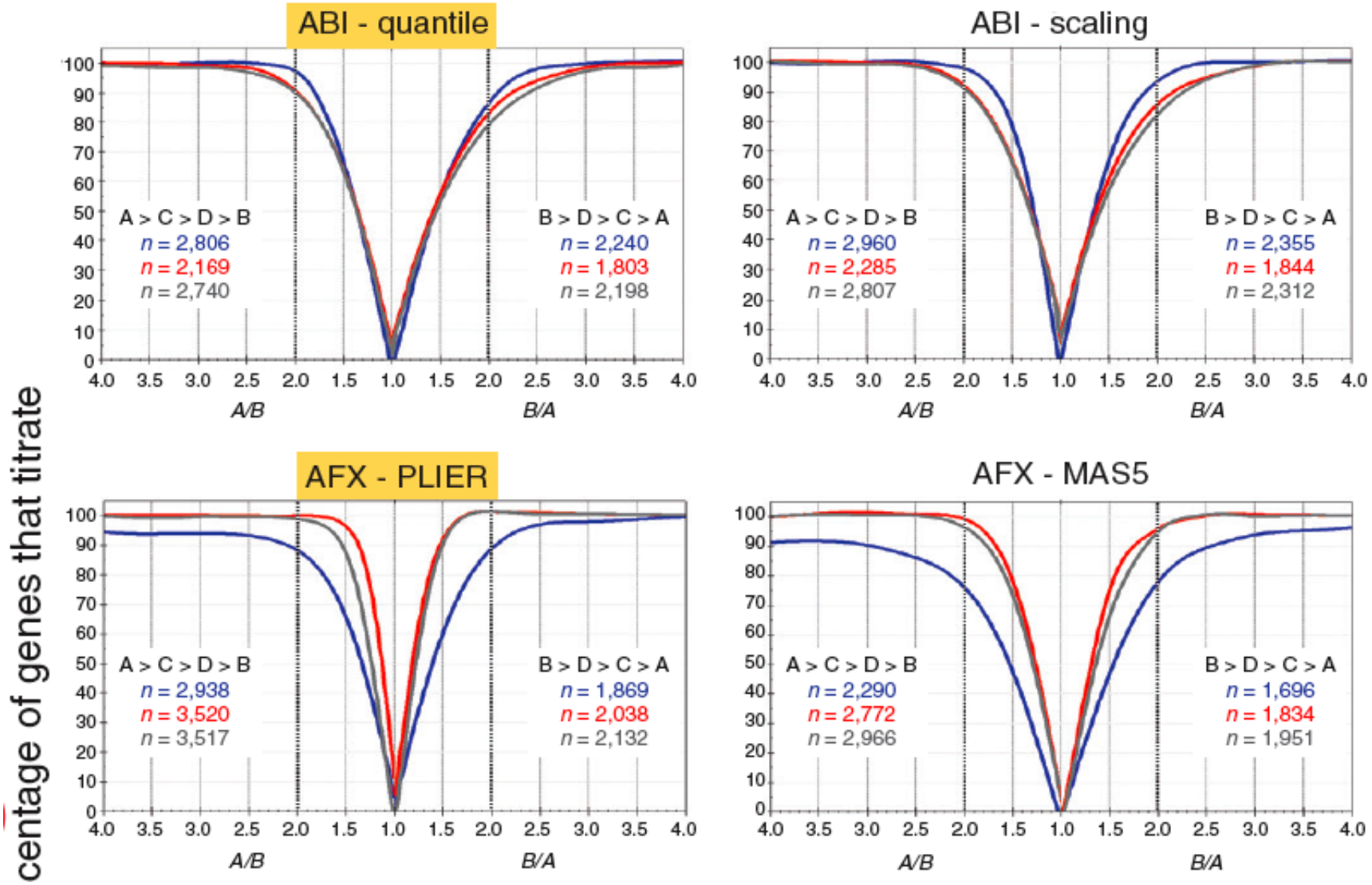


Figure 1 RNA samples. We used expression measurements from two independent total RNA samples, A and B, and mixtures of these two samples at defined ratios of 3:1 (C) and 1:3 (D). The titration mixtures were generated once for all experiments, with samples A and B at equal total RNA concentrations as determined by A_{260}.

Self-consistent monotone titration frequencies



Other criteria for platform performance

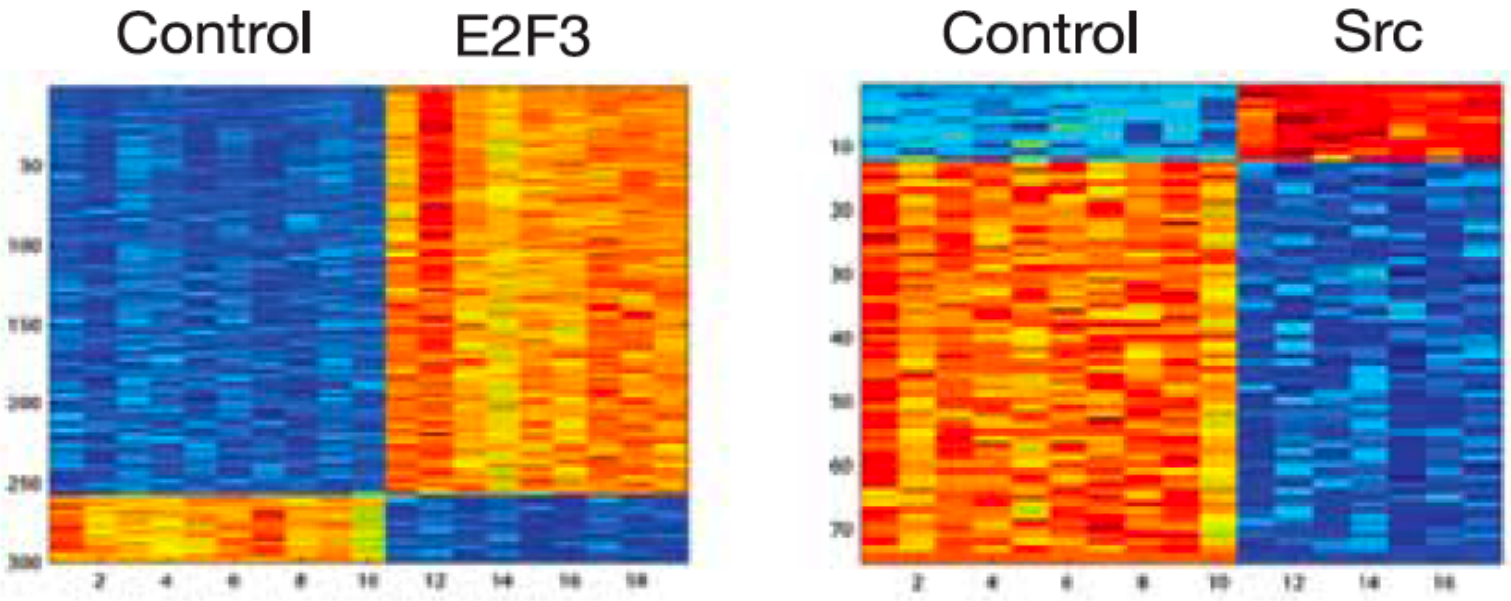
Table 1 Gene counts for AFX and ABI (top) and AG1, GEH and ILM (bottom) for each no

		Quantile			Scaling			PLIER			
Row	Condition	ABI_1	ABI_2	ABI_3	ABI_1	ABI_2	ABI_3	AFX_1	AFX_2	AFX_3	AFX_1
1	Detected in A · B · C · D	8,049	7,863	8,550	8,049	7,863	8,550	7,359	7,006	7,424	7,359
2	A > B	4,284	4,191	4,509	4,308	4,219	4,424	4,423	4,291	4,557	4,244
3	B > A	3,765	3,672	4,041	3,741	3,644	4,126	2,936	2,715	2,867	3,115
4	A > B and $P < 0.001$	3,144	2,298	3,046	3,143	2,376	3,037	3,723	3,632	3,848	2,982
5	B > A and $P < 0.001$	2,572	1,886	2,436	2,571	1,930	2,494	2,356	2,176	2,306	2,074
6	A > C > D > B	3,063	2,924	3,159	3,296	3,104	3,256	3,042	3,751	3,616	2,493
7	B > D > C > A	2,471	2,424	2,622	2,670	2,487	2,772	1,924	2,154	2,222	1,873
8	A > C > D > B and $P < 0.001$	2,806	2,169	2,740	2,960	2,285	2,807	2,938	3,520	3,517	2,290
9	B > D > C > A and $P < 0.001$	2,240	1,803	2,198	2,355	1,844	2,312	1,869	2,038	2,132	1,696
10	(A > C > D > B) / (A > B)	0.71	0.70	0.70	0.77	0.74	0.74	0.69	0.87	0.79	0.59
11	(B > D > C > A) / (B > A)	0.66	0.66	0.65	0.71	0.68	0.67	0.66	0.79	0.78	0.60

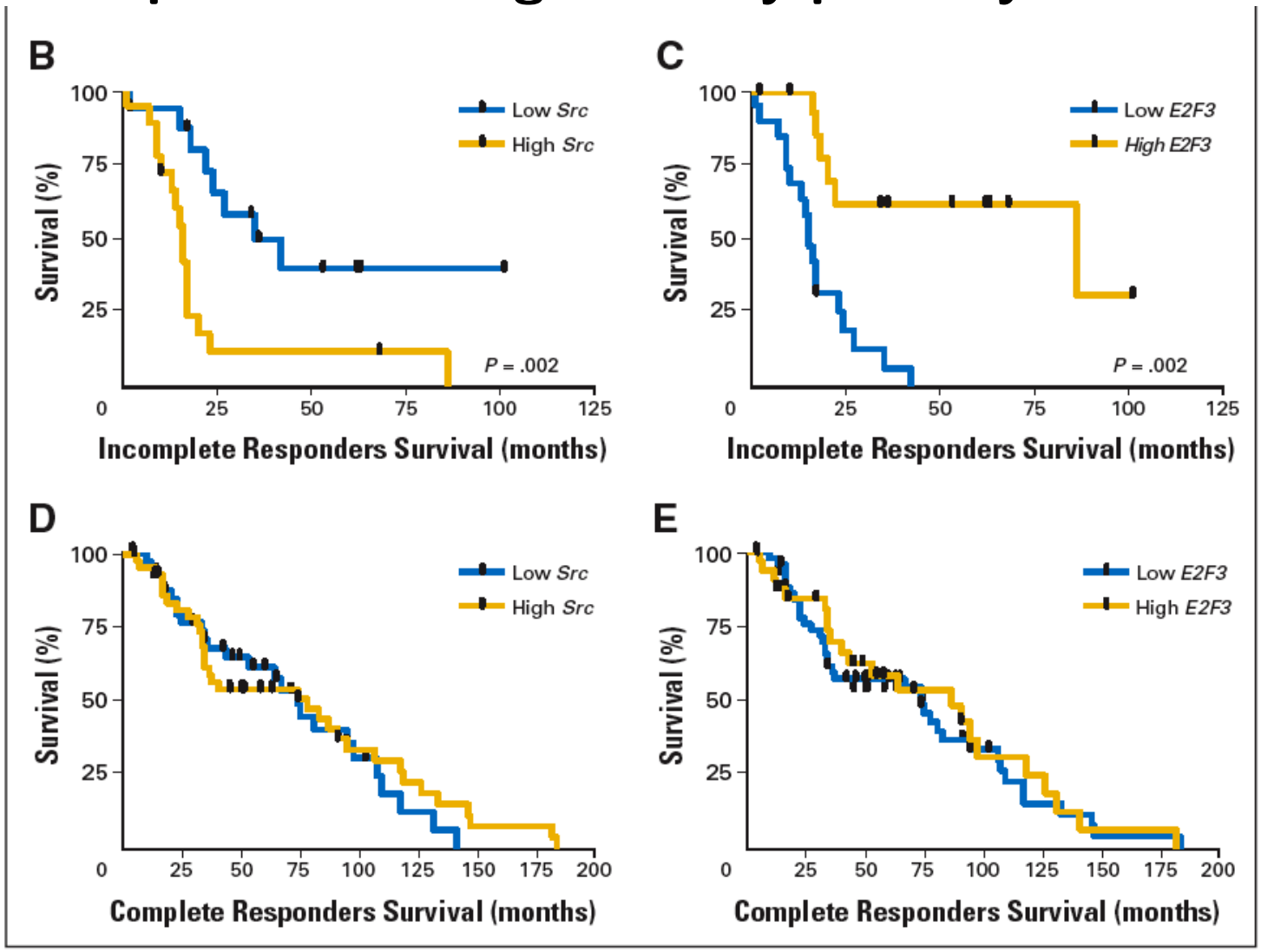
Ambitions III: treatment algorithms based on integrative transcript profiling

- Signature/score for drug responsiveness via standard transcript profiling
- Pathway activation signatures/scores based on profiling cell lines that have been transfected with pathway-specific vectors
- Illustrate survival differences for drug-non-responders who differ with respect to pathway activation

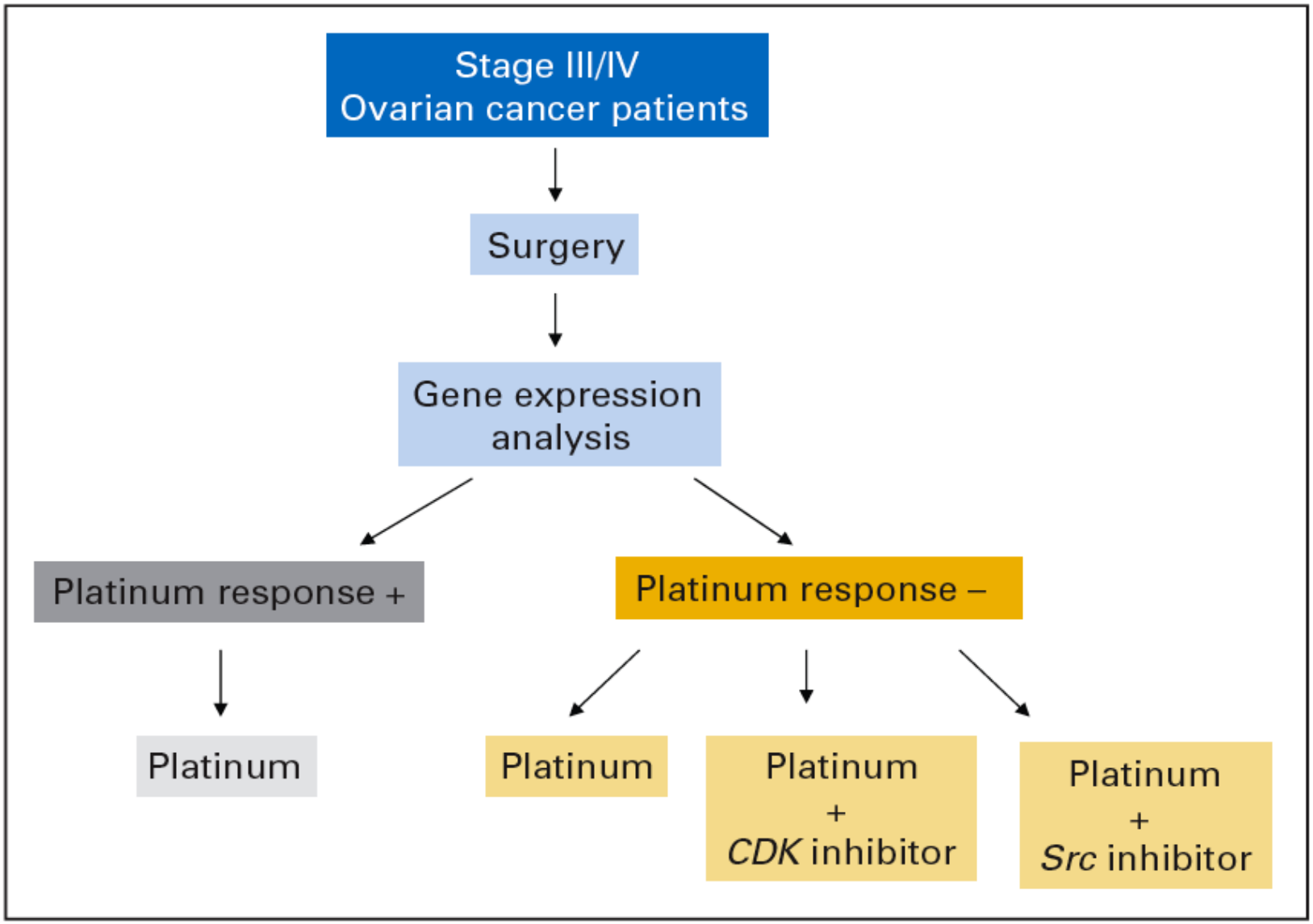
Bild et al Nature 2006: Genes manifesting pathway dysregulation



Dressman et al JCO 2007: Survival among platinum non-responders distinguished by pathway activation



Algorithm


Stage III/IV
Ovarian cancer patients
Surgery
Gene expression analysis
Platinum response +
Platinum response –
Platinum
Platinum
Platinum + *CDK* inhibitor
Platinum + *Src* inhibitor

Reproducibility challenge: Baggerly et al. JCO letter 2008

Run Batch Effects Potentially Compromise the Usefulness of Genomic Signatures for Ovarian Cancer

TO THE EDITOR: A major goal of personalized medicine is to predict, before administering a treatment, whether the patient will respond to it. Recently, Dressman et al[1] presented an approach that appeared to move us toward this goal in the context of ovarian cancer. Using microarray expression profiles, they first identified a set of genes that could differentiate between patients who did (CR) and did not (NR) respond to primary platinum-based chemotherapy. Then, following Bild et al,[2] they scored each tumor for the levels of five different oncogenic pathways. They reported that three pathways (Src, E2F3, and Myc) stratify the NRs into subgroups with significantly different survival characteristics, suggesting how further therapies might be targeted for these patients.

We examined these data in order to help investigators at our institution make better use of this approach. We were unable to reproduce the results reported, and the structure that we did find appears driven far more by run date than by clinical response. Our findings are outlined here; supplementary reports (ovca01-ovca07) provide details.

(1) The posted mapping of numbers to samples is scrambled (as

(3) We identified 107 Affymetrix (Santa Clara, CA) probese corresponding to the "best" 100 genes reported by Dressman e ambiguities in annotation led to some duplication (ovca03).

(4) The CEL files can be grouped into clearly separated ba on the basis of run date. Response and survival are confounded run date, particularly with the samples processed earliest (ovca0

(5) We contrasted the CR and NR samples, gene by gene, two-sample *t* tests. *P* values from the reported "best 100" gene uniformly distributed, suggesting results no better than chance. tering based on these genes fails to separate CRs from NRs. Th some evidence of differential expression in the set of all genes. by-gene analyses of variance, however, suggest strong batch effec almost every gene. After correcting for these batch effects, separ between CRs and NRs drops to low levels (ovca05).

(6) Using data from Bild et al,[2] we computed our own pat scores for each tumor sample. Our pathway gene lists differ slightly those of Bild et al[2] due to differences in array processing (Affym Microarray Analysis Suite, v.5.0 in Bild et al,[2] robust multi-array an here). These scores are relatively robust with respect to the precise ge selected, but they show clear confounding with run batch. After corre for batch, the scores change substantially (ovca06).

(7) Finally, we looked for differences in survival as a functi dichotomized (high/low) pathway scores. For each pathway looked at results for three patient subgroups (NR, CR, and all) usi

Summary of some ambitions

- surveys of array collections still problematic, underdocumented, hard to reproduce (Ioannidis Nat Biotech 09)
- methodologies for interpreting transcript profiling experiments not harmonized – GSEA, GSA, SAFE, ...
- experimental toolchest exploding, statisticians can barely keep up
- 'qualitative' findings and presentations are common
  - MAQC monotone titration frequencies
  - 'pathway activation'
  - null distributions derived by label permutation, observed result is 'in the tail' – but model underlying permutations may not be justified

Bioconductor's tasks:

- Reduce barriers to entry for statisticians
- Foster software reliability, analysis reproducibility
- Ease biologists' use of modern statistical tools

Teaching case study: Problem 4-72 of Alberts, MBoC

Does loss of histone H4 increase expression of a greater fraction of genes near telomeres than in the rest of the genome? Explain.

Solution: 'statistical analysis'

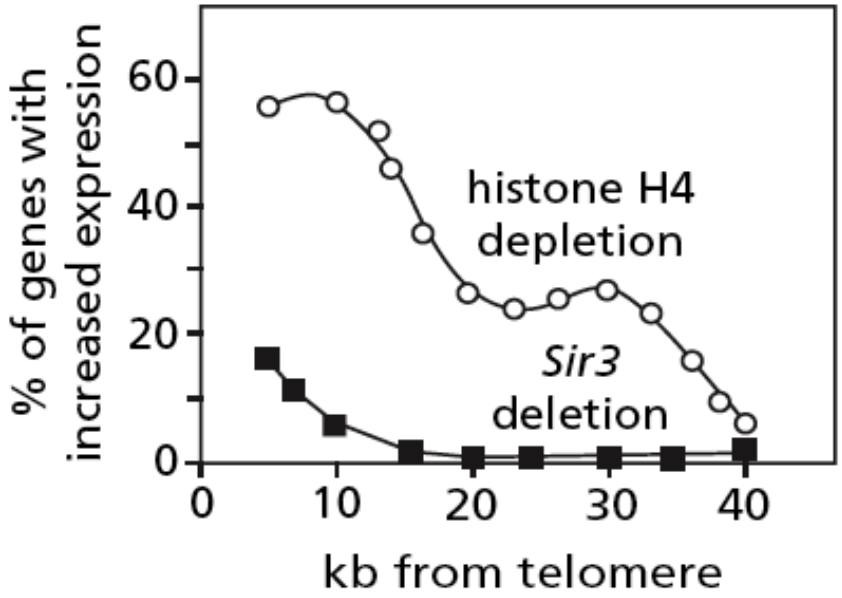


Figure 4–46 Extent of telomeric gene activation after depletion of histone H4 or deletion of the *Sir3* gene (Answer 4–72). For this analysis all the

The student is told that 50-gene windows were formed and proportion of expressed genes is plotted relative to distance of window midpoint to telomere

Caveats on the Alberts volume presentation

- Student is not given the source of the chromosome display (Wyrick et al 1999; it is in the bibliography with a back-reference, but no direct reference)
- Graphics in the text are worse than in my slide
- The display is hard to parse; it is not clear that all chromosomes exhibit telomere-proximal expression enrichment
- There are 2 DVDs supplied with the book, but no numerical data – all figures and text
- Getting students to perform informal visual interpretation/explanation is nice, but mechanics of justifying the interpretation must also be mastered

An approach using Bioconductor concepts

- The computer programming language R is used
- It includes built-in documentation and can be extended by users without special privileges
- A self-describing data structure is used to manage and support interrogation of the relevant data
- The student explores the numerical patterns in the data directly
- This serves to model both
  - the administrative tasks of structuring and managing experimental results
  - the analytic tasks of discovering and interpreting patterns in quantitative data

Example for Wyrick data

```
> H4dep6h

ExpressionSet (storageMode: lockedEnvironment)
assayData: 6365 features, 4 samples
  element names: exprs
phenoData
  sampleNames: W1, M1, W2, M2
  varLabels and varMetadata description: none
featureData
  featureNames: YAL069W, YAL067C, ..., YPR132W  (6365 total)
  fvarLabels and fvarMetadata description:
    chrloc: NA
    chrnum: NA
experimentData: use 'experimentData(object)'
  pubMedIds: 10586882
Annotation: org.Sc.sgd.db
```

MIAME

```
> experimentData(H4dep6h)

Experiment data
  Experimenter name: Wyrick JJ
  Laboratory: Whitehead Institute for Biomedical Research
  Contact information:
  Title: Chromosomal landscape of nucleosome-dependent ge
  URL: http://jura.wi.mit.edu/young_public/chromatin/
  PMIDs: 10586882

  Abstract: A 169 word abstract is available. Use 'abstra
```

```
> abstract(H4dep6h)
[1] "Eukaryotic genomes are packaged into nucleosomes, which are
thought to repress gene expression generally. Repression is particularly
evident at yeast telomeres, where genes within the telomeric
heterochromatin appear to be silenced by the histone-binding silent
information regulator (SIR) complex (Sir2, Sir3, Sir4) and Rap1
(refs 4-10). Here, to investigate how nucleosomes and silencing
factors influence global gene expression, we use high-density arrays
to study the effects of depleting nucleosomal histones and
silencing factors in yeast. Reducing nucleosome content by depleting
histone H4 caused increased expression of 15% of genes
and reduced expression of 10% of genes, but it
had little effect on expression of the majority (75%)
of yeast genes. Telomere-proximal genes were found to be
de-repressed over regions extending 20 kilobases from the telomeres,
well beyond the extent of Sir protein binding and
the effects of loss of Sir function. These results
indicate that histones make Sir-independent contributions to telomeric silencing,
and that the role of histones located elsewhere in
chromosomes is gene specific rather than generally repressive."
```

from `help(org.Sc.sgdCHRLOC)`

```
Each ORF identifier maps to a named vector of chromosomal
locations, where the name indicates the chromosome.

Chromosomal locations on both the sense and antisense strands are
measured as the number of base pairs from the p (5' end of the
sense strand) to q (3' end of the sense strand) arms. Chromosomal
locations on the antisense strand have a leading "-" sign (e. g.
-1234567).

Since some genes have multiple start sites, this field can map to
multiple locations.
```

`X[G,S]` for filtering; other functions

```
> exprs(H4dep6h[1:3,])   # dump
```

```
        W1  M1 W2  M2
YAL069W  5  54  5  37
YAL067C 82 123 22 129
YAL066W 52  23 28  26
```

```
> cloc = function(x) abs(fData(x)$chrloc)  # helper
> H4dep6hL = H4dep6h[ !is.na(cloc(H4dep6h)), ]
> erat = exprs(H4dep6hL[,2])/exprs(H4dep6hL[,1])
> mean(erat[cloc(H4dep6hL)<40000]>3)
```

```
[1] 0.3525424
```

```
> mean(erat[cloc(H4dep6hL)>=40000]>3)
```

```
[1] 0.1325854
```

What happened?

- We examined Wyrick's data as distributed at the URL encoded in `H4dep6h` (`exprs()` dump)
- We simplified filtering of expression values by chromosomal location using a new function `cloc`
- We got rid of all genes with missing locations (SGD)
- We computed expression ratios (MT to WT)
- We compared frequency of three-fold increase with H4 depletion in two regions: within 40kb of telomere and beyond 40kb
- We seem to have confirmed the basic assertions in Wyrick

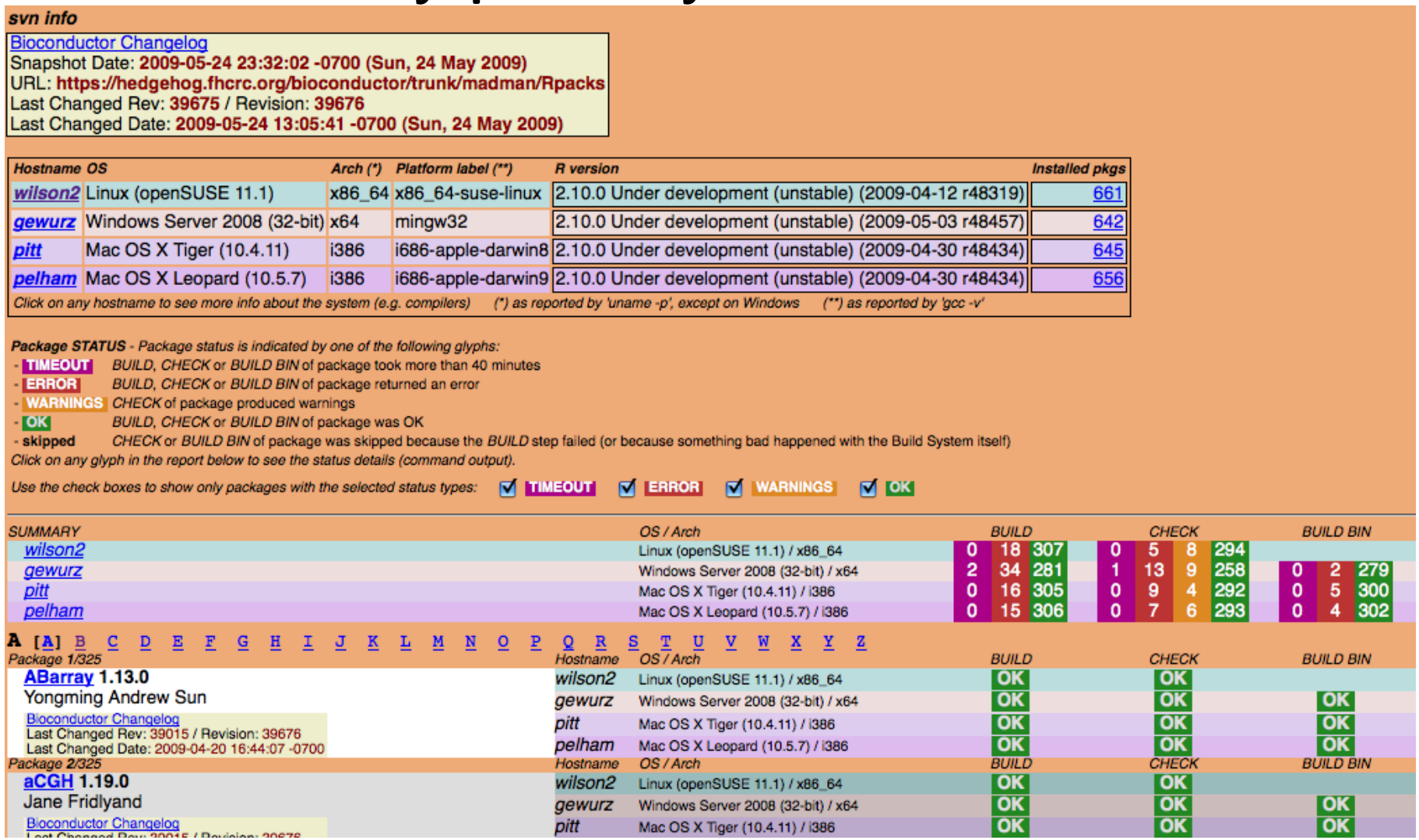
Additional potentials

- Which genes are not de-repressed when histone H4 is depleted? What is different about them?
- Is the rate of decline of de-repression over base-pairs constant across chromosomes?
- Are there other biologically interesting 'bumps' of de-repression away from telomeres?
- With the data in hand, much can be done
- With data from other experiments (e.g., heat shock, cell cycle) along with annotation resources and sequence, broader integrative inquiries can be made

Barriers to entry: reduced?

- with an R package encoding the Wyrick expression data,
  - a statistician can jump right in to methods consideration
  - a biologist can, with proper training, reproduce published analyses and perturb them to illuminate unaddressed concerns
    - * – is the inference sensitive to the three-fold de-repression threshold?
    - * – how does de-repression localize in detail?
- for de novo inquiries, some programming will be needed; patterns that are reused become packaged/turnkey

Software reliability, portability



svn info

Bioconductor Changelog
Snapshot Date: **2009-05-24 23:32:02 -0700 (Sun, 24 May 2009)**
URL: **https://hedgehog.fhcrc.org/bioconductor/trunk/madman/Rpacks**
Last Changed Rev: **39675** / Revision: **39676**
Last Changed Date: **2009-05-24 13:05:41 -0700 (Sun, 24 May 2009)**

Hostname	OS	Arch (*)	Platform label (**)	R version	Installed pkgs
wilson2	Linux (openSUSE 11.1)	x86_64	x86_64-suse-linux	2.10.0 Under development (unstable) (2009-04-12 r48319)	661
gewurz	Windows Server 2008 (32-bit)	x64	mingw32	2.10.0 Under development (unstable) (2009-05-03 r48457)	642
pitt	Mac OS X Tiger (10.4.11)	i386	i686-apple-darwin8	2.10.0 Under development (unstable) (2009-04-30 r48434)	645
pelham	Mac OS X Leopard (10.5.7)	i386	i686-apple-darwin9	2.10.0 Under development (unstable) (2009-04-30 r48434)	656

Click on any hostname to see more info about the system (e.g. compilers) *(*) as reported by 'uname -p', except on Windows* *(**) as reported by 'gcc -v'*

***Package STATUS** - Package status is indicated by one of the following glyphs:*
- TIMEOUT *BUILD, CHECK or BUILD BIN of package took more than 40 minutes*
- ERROR *BUILD, CHECK or BUILD BIN of package returned an error*
- WARNINGS *CHECK of package produced warnings*
- OK *BUILD, CHECK or BUILD BIN of package was OK*
- **skipped** *CHECK or BUILD BIN of package was skipped because the BUILD step failed (or because something bad happened with the Build System itself)*

Click on any glyph in the report below to see the status details (command output).

Use the check boxes to show only packages with the selected status types: ☑ TIMEOUT ☑ ERROR ☑ WARNINGS ☑ OK

SUMMARY	OS / Arch	BUILD	CHECK	BUILD BIN
wilson2	Linux (openSUSE 11.1) / x86_64	0 18 307	0 5 8 294	
gewurz	Windows Server 2008 (32-bit) / x64	2 34 281	1 13 9 258	0 2 279
pitt	Mac OS X Tiger (10.4.11) / i386	0 16 305	0 9 4 292	0 5 300
pelham	Mac OS X Leopard (10.5.7) / i386	0 15 306	0 7 6 293	0 4 302

A [A] B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Package 1/325	Hostname	OS / Arch	BUILD	CHECK	BUILD BIN
ABarray 1.13.0 Yongming Andrew Sun Bioconductor Changelog Last Changed Rev: 39015 / Revision: 39676 Last Changed Date: 2009-04-20 16:44:07 -0700	wilson2	Linux (openSUSE 11.1) / x86_64	OK	OK	
	gewurz	Windows Server 2008 (32-bit) / x64	OK	OK	OK
	pitt	Mac OS X Tiger (10.4.11) / i386	OK	OK	OK
	pelham	Mac OS X Leopard (10.5.7) / i386	OK	OK	OK
Package 2/325	**Hostname**	**OS / Arch**	**BUILD**	**CHECK**	**BUILD BIN**
aCGH 1.19.0 Jane Fridlyand Bioconductor Changelog	wilson2	Linux (openSUSE 11.1) / x86_64	OK	OK	
	gewurz	Windows Server 2008 (32-bit) / x64	OK	OK	OK
	pitt	Mac OS X Tiger (10.4.11) / i386	OK	OK	OK

Training vehicles

- Short courses – 5-6x yearly
- Project conference, tutorials, developer day
- Vignettes – computable documents illustrating workflow patterns
- Books (three monographs since 2005)
- Cold Spring Harbor Lab Summer course (integrative statistical analysis of genome scale data)

Conclusions

- Numerical/computational competencies of biologists must increase
- Reliability of inferences cannot falter in the face of volume/complexity of experimental results
- R has growing acceptance in many domains; Excel should not be used for nontrivial tasks involving statistics/visualization
- R packages of experimental data are important teaching resources
  - self-documenting
  - can include scripts to completely reproduce intepretive analyses
- in progress: tracking the quantitative components of a well-accepted biology text with an R/bioconductor companion