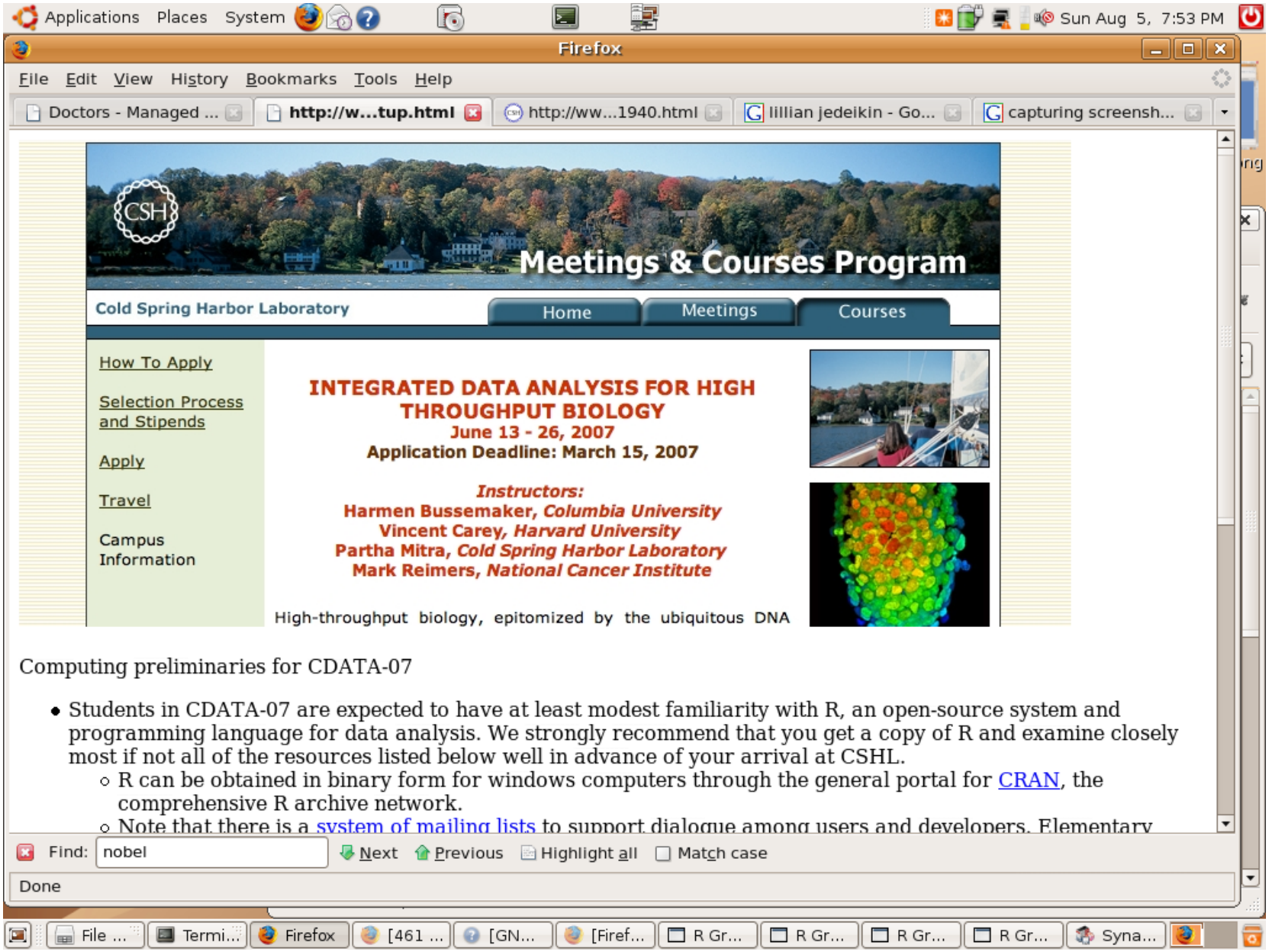
Data architecture and workflow for high-throughput genomics

VJ Carey, Channing Lab;

- views on outreach
- four experimental paradigms
- interface contracts and compliance

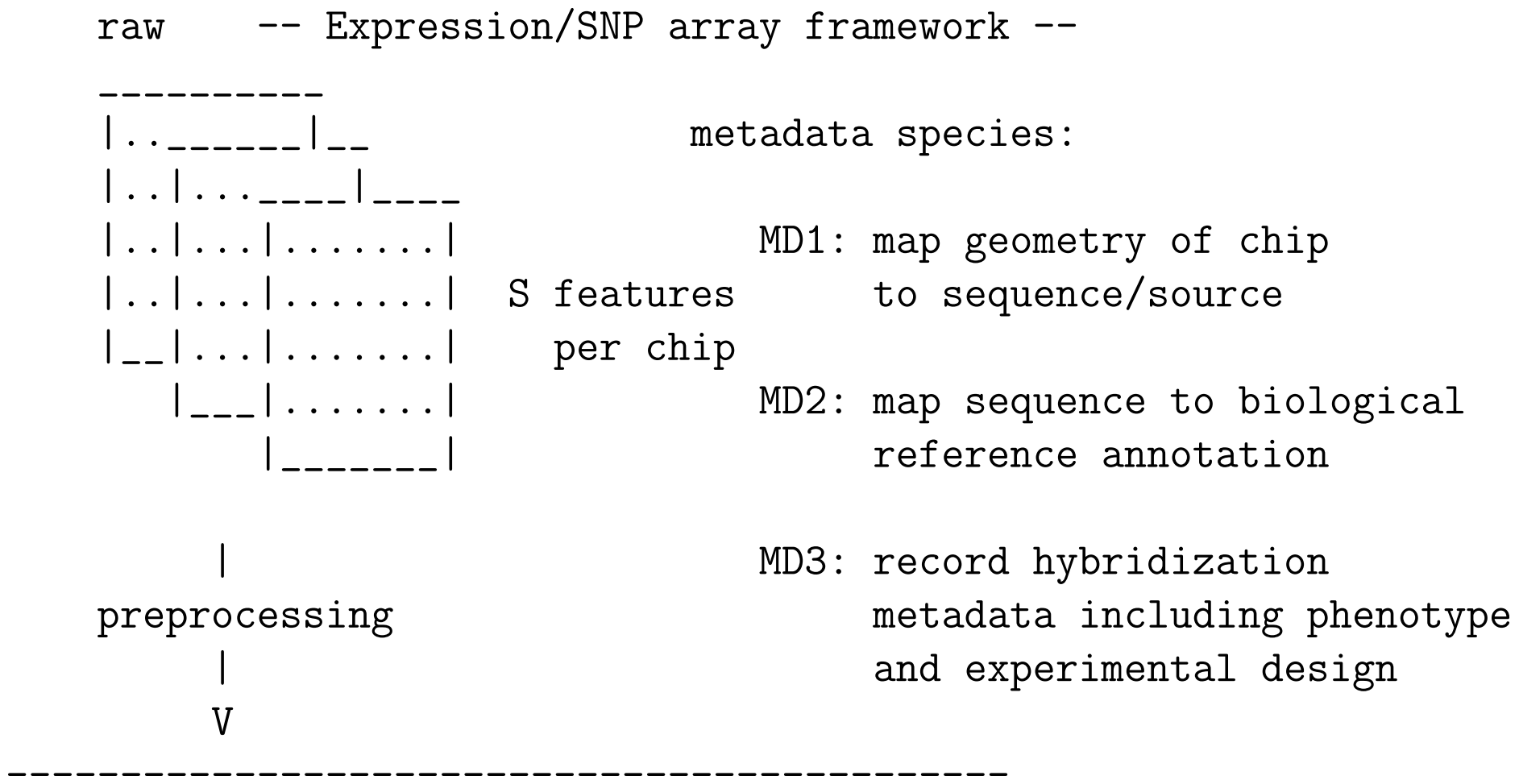

Applications Places System
Sun Aug 5, 7:38 PM
Firefox
File Edit View History Bookmarks Tools Help
http://www.cshl.edu/genetics100/1940.html
ing screenshots in linux
Latest BBC Headlines
bioc
biocTests
nyt
slshd
Doctors - Managed ...
The Fermi Paradox ...
http://w...940.html
lillian jedeikin - Go...
capturing screensh...
Cold Spring Harbor Laboratory
Celebrating 100 Years of Genetics, 1904 - 2004
CSH
student
s
tivated
n
attend
tended
Find: carey
Next
Previous
Highlight all
Match case
Done
R Gr...
Termi...
Firefox
[461 ...
[GN...
[Firef...
stvjc ...
stvjc ...
stvjc ...
stvjc ...
Syna...


Applications Places System
Sun Aug 5, 7:53 PM
Firefox
File Edit View History Bookmarks Tools Help
Doctors - Managed ...
http://w...tup.html
http://ww...1940.html
lillian jedeikin - Go...
capturing screensh...
Meetings & Courses Program
Cold Spring Harbor Laboratory
Home
Meetings
Courses
How To Apply
Selection Process and Stipends
Apply
Travel
Campus Information
INTEGRATED DATA ANALYSIS FOR HIGH THROUGHPUT BIOLOGY
June 13 - 26, 2007
Application Deadline: March 15, 2007
Instructors:
Harmen Bussemaker, Columbia University
Vincent Carey, Harvard University
Partha Mitra, Cold Spring Harbor Laboratory
Mark Reimers, National Cancer Institute
High-throughput biology, epitomized by the ubiquitous DNA
Computing preliminaries for CDATA-07
Students in CDATA-07 are expected to have at least modest familiarity with R, an open-source system and programming language for data analysis. We strongly recommend that you get a copy of R and examine closely most if not all of the resources listed below well in advance of your arrival at CSHL.
R can be obtained in binary form for windows computers through the general portal for CRAN, the comprehensive R archive network.
Note that there is a system of mailing lists to support dialogue among users and developers. Elementary
Find: nobel
Next
Previous
Highlight all
Match case
Done
File ...
Termi...
Firefox
[461 ...
[GN...
[Firef...
R Gr...
R Gr...
R Gr...
R Gr...
Syna...

Cold Spring Harbor CDATA-07 [Integrative data analysis for high-throughput biology]

- Strategy – 2 week course (!)
  - prominent experimentalists/methodologists lecture most days (e.g., Futcher, Lieb, Wigler, Cheung, Spielman, Baggerly, Irizarry, Carvalho)
  - dataset/packages associated with each technique form basis for stat-oriented lectures and labs
  - most students are experimentalists (have had NIH and FDA project officers as well); many not very happy with statistics curriculum
- will target advertising at stats and comp-bio departments for 08;

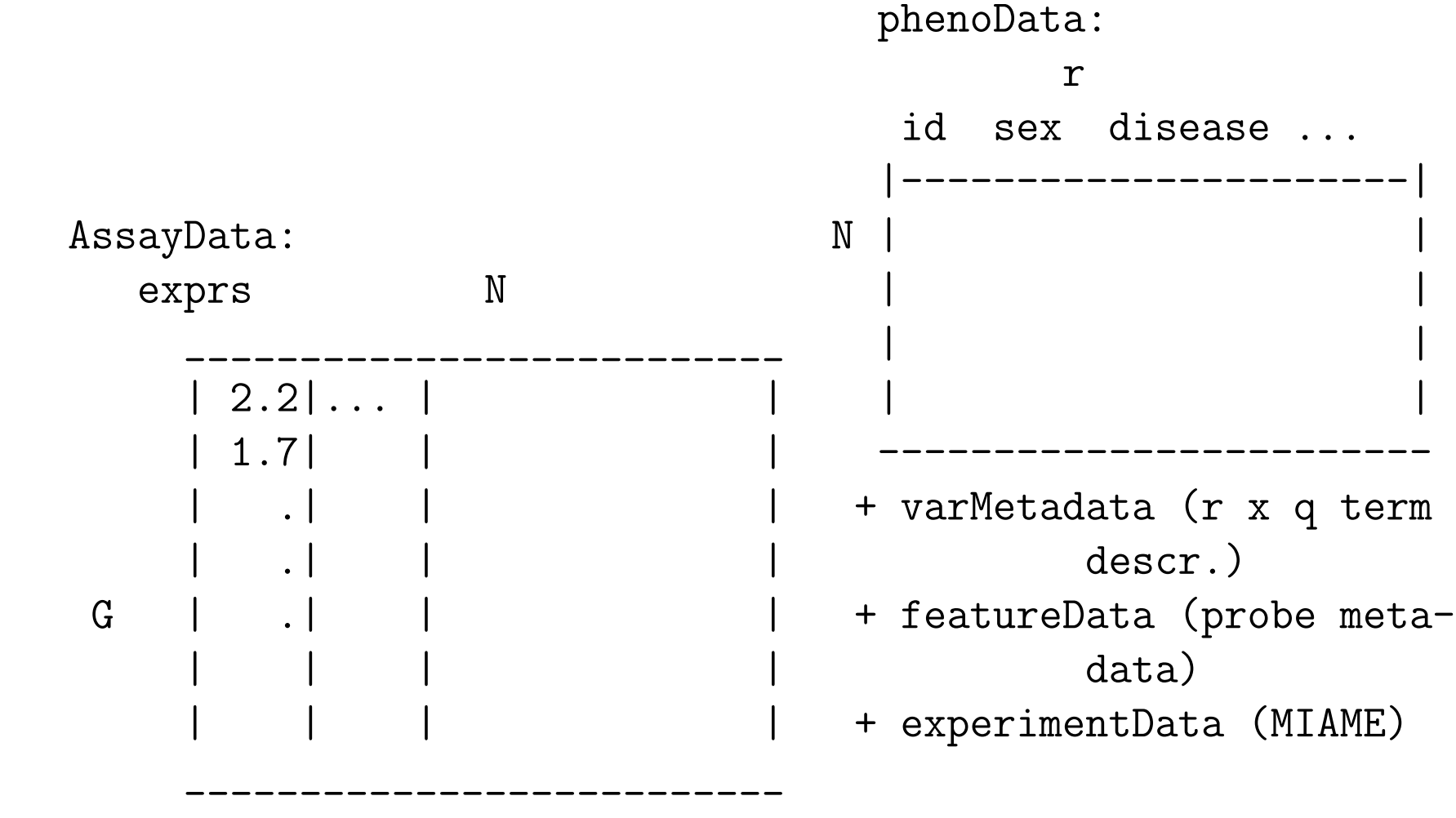
Upstream data structures



Downstream data structures

preprocessing... removal of nonbiologic variation

N chips -> ExpressionSet instance



Basic outreach approach

- make it easy for practitioners to touch, manipulate, interrogate digital products of exemplary experiments
- big motivation to learn a little R

Integrative container design: Four experimental paradigms

- Useful aim: *All relevant information from an experiment or family of related experiments should be contained in a single, aptly named, R variable*
- Moderately successful examples:
  - `Golub_Merge` (but developers still include their own matrix representations in `data`?)
  - `yeastCC` – same
  - `chr20GGceuRMA` (hgfocus + hapmap 700K for Utah CEPH founders)
  - `harbChIP` (abuse of ExpressionSet structure; featureData includes intergenic sequence data)
  - `neveExCGH` (array CGH + u133a on 50 breast cancer cell lines)

Paradigm 1: expression time series

```
> library(yeastCC)
> data(spYCCES)
> spYCCES
ExpressionSet (storageMode: environment)
assayData: 6178 features, 77 samples
  element names: exprs
phenoData
  sampleNames: cln3_40, cln3_30, ..., elu_390  (77 total)
  varLabels and varMetadata description:
    syncmeth: experimental method of synchronization or cyclin induction
    time: in minutes
    phase: Phase of the cell cycle. M: mitosis, G1: gap 1, S: DNA synthesis, G
   gap 2.
featureData
  featureNames: YAL001C, YAL002W, ..., YPR204W  (6178 total)
  fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
  pubMedIds: 9843569
Annotation: YEAST
```

```
> experimentData(spYCCES)
Experiment data
  Experimenter name: Spellman PT
  Laboratory: Department of Genetics, Stanford University Medical Center, Stan
  Contact information:
  Title: Comprehensive identification of cell cycle-regulated genes of the yea
  URL:
  PMIDs: 9843569

  Abstract: A 150 word abstract is available. Use 'abstract' method.
```

```
> abstract(spYCCES)
[1] "We sought to create a comprehensive catalog of yeast genes whose
transcript levels vary periodically within the cell cycle. To this end,
we used DNA microarrays and samples from yeast cultures synchronized
by three independent methods: alpha factor arrest, elutriation, and
arrest of a cdc15 temperature-sensitive mutant. Using periodicity
and correlation algorithms, we identified 800 genes that meet an objective
minimum criterion for cell cycle regulation. In separate experiments, designed
to examine the effects of inducing either the G1 cyclin Cln3p or the
B-type cyclin Clb2p, we found that the mRNA levels of more than half
of these 800 genes respond to one or both of these cyclins....
```

Design: sync meth vs sampling times

```
> table(spYCCES$sync, spYCCES$time)[, 1:15]
       0 7 10 14 20 21 28 30 35 40 42 49 50 56 60
alpha  1 1  0  1  0  1  1  0  1  0  1  1  0  1  0
cdc15  0 0  1  0  0  0  0  1  0  0  0  0  1  0  0
cdc28  1 0  1  0  1  0  0  1  0  1  0  0  1  0  1
clb2   0 0  0  0  0  0  0  0  0  2  0  0  0  0  0
cln3   0 0  0  0  0  0  0  1  0  1  0  0  0  0  0
elu    1 0  0  0  0  0  0  1  0  0  0  0  0  0  1
```

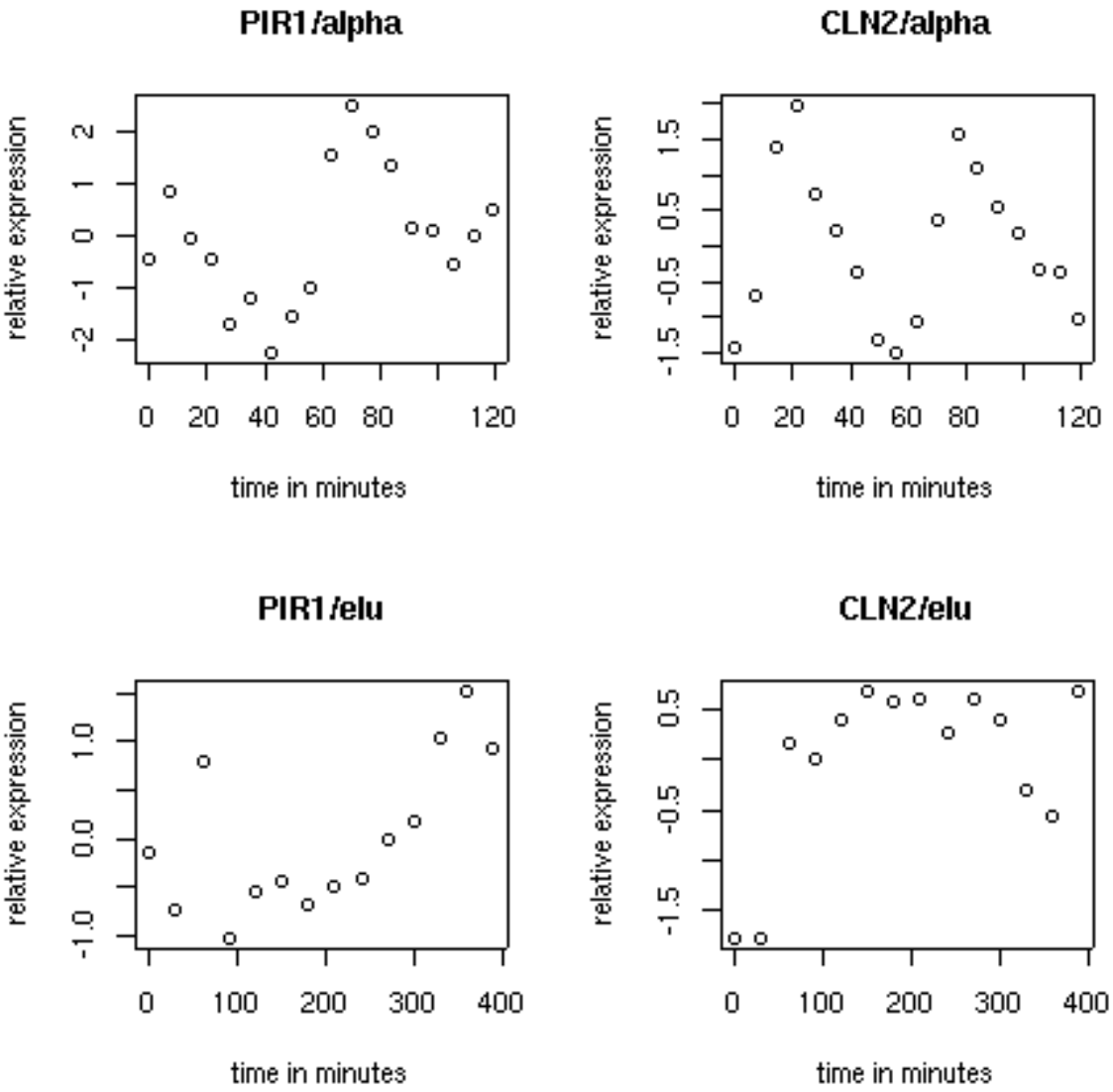
Biology: (declared) phase vs sampling times

```
> table(spYCCES$phase, spYCCES$time)[, 1:15]
      0 7 10 14 20 21 28 30 35 40 42 49 50 56 60
G1    0 0  1  1  1  1  0  0  0  0  0  0  1  0  1
G2    0 0  0  0  0  0  0  0  0  0  1  0  1  0  0
M     0 0  0  0  0  0  0  0  0  0  0  1  0  1  1
M/G1  3 1  1  0  0  0  0  2  0  0  0  0  0  0  0
S     0 0  0  0  0  0  1  1  1  1  0  0  0  0  0
```

programming: filtering to see (declared) phase vs sampling times

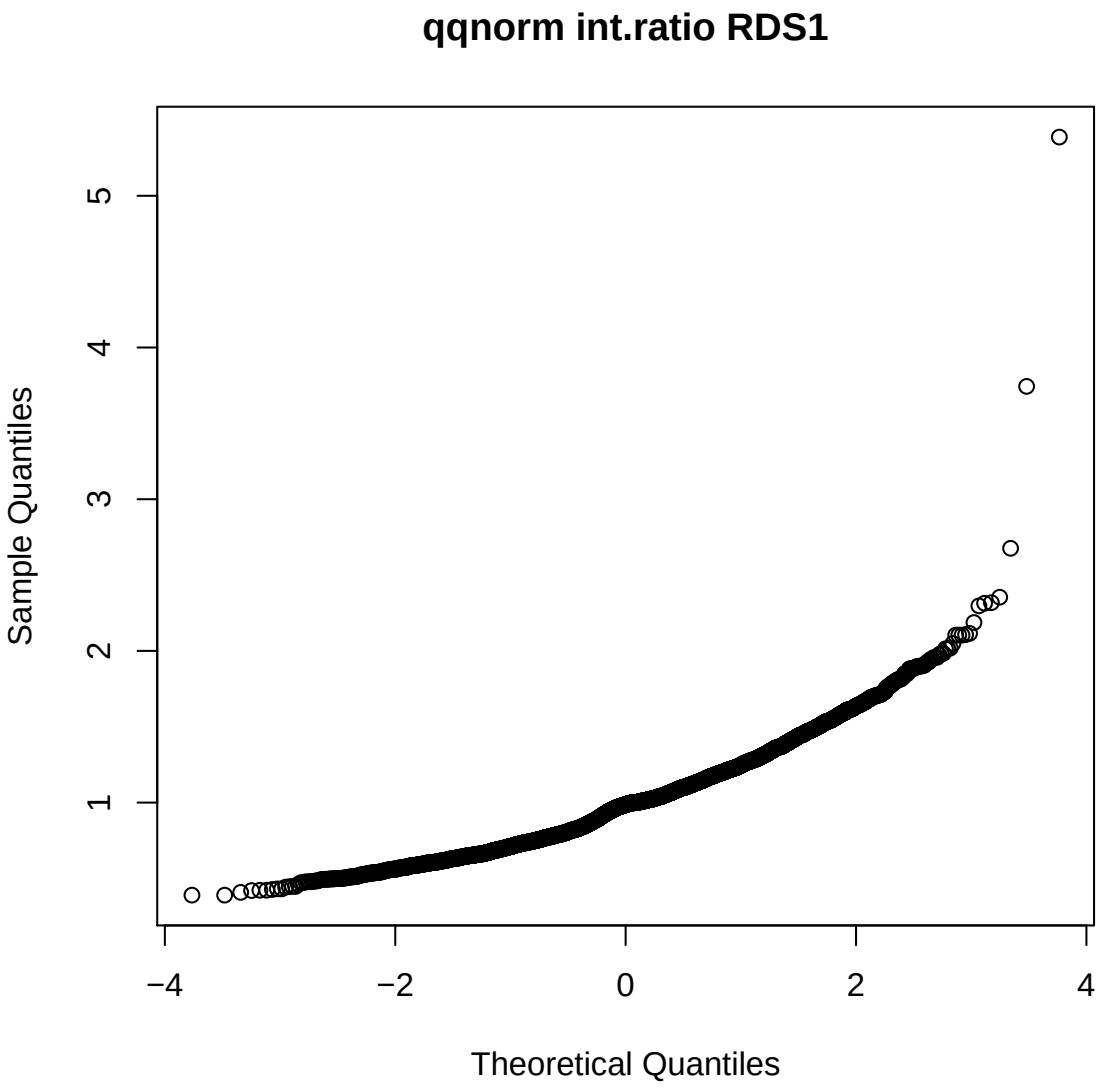
```
> CDC15 = which(spYCCES$sync=="cdc15")
> table(spYCCES$phase[CDC15], spYCCES$time[CDC15])[, 1:15]
      10 30 50 70 80 90 100 110 120 130 140 150 160 170 180
  G1   0  0  1  1  0  0   0   0   0   0   1   1   1   0   0
  S    0  0  0  0  1  0   0   0   0   0   0   0   0   1   1
  G2   0  0  0  0  0  1   0   0   0   0   0   0   0   0   0
  M    0  0  0  0  0  0   1   1   0   0   0   0   0   0   0
  M/G1 1  1  0  0  0  0   0   0   1   1   0   0   0   0   0
```

- Exercise: compare samples on a given gene obtained with different synchronization methods with respect to periodicity of expression


PIR1/alpha
relative expression
-2
-1
0
1
2
0
20
40
60
80
120
time in minutes
CLN2/alpha
relative expression
-1.5
-0.5
0.5
1.5
0
20
40
60
80
120
time in minutes
PIR1/elu
relative expression
-1.0
0.0
1.0
0
100
200
300
400
time in minutes
CLN2/elu
relative expression
-1.5
-0.5
0.5
0
100
200
300
400
time in minutes

paradigm 2: ChIP-chip, harbChIP in package:harbChIP

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 6230 features, 204 samples
  element names: exprs, se.exprs
phenoData
  rowNames: A1 (MATA1), ABF1, ..., ZMS1  (204 total)
  varLabels and varMetadata description:
    txFac: transcription factor symbol from Harbison website CSV fil
  s
featureData
  featureNames: YAL001C, YAL002W, ..., MRH1  (6230 total)
  fvarLabels and fvarMetadata description:
    ID: NA
    PLATE: NA
    ...: ...
    REV_SEQ: NA
    (12 total)
experimentData: use 'experimentData(object)'
  pubMedIds: 15343339
```


qqnorm int.ratio RDS1
Sample Quantiles
Theoretical Quantiles
1
2
3
4
5
−4
−2
0
2
4

- Exercise: verify that the motif for RDS1 discovered by Harbison et al, CGGCCG, is present in the most highly bound intergenic regions for that TF.
- Is it present in modestly or weakly bound regions?

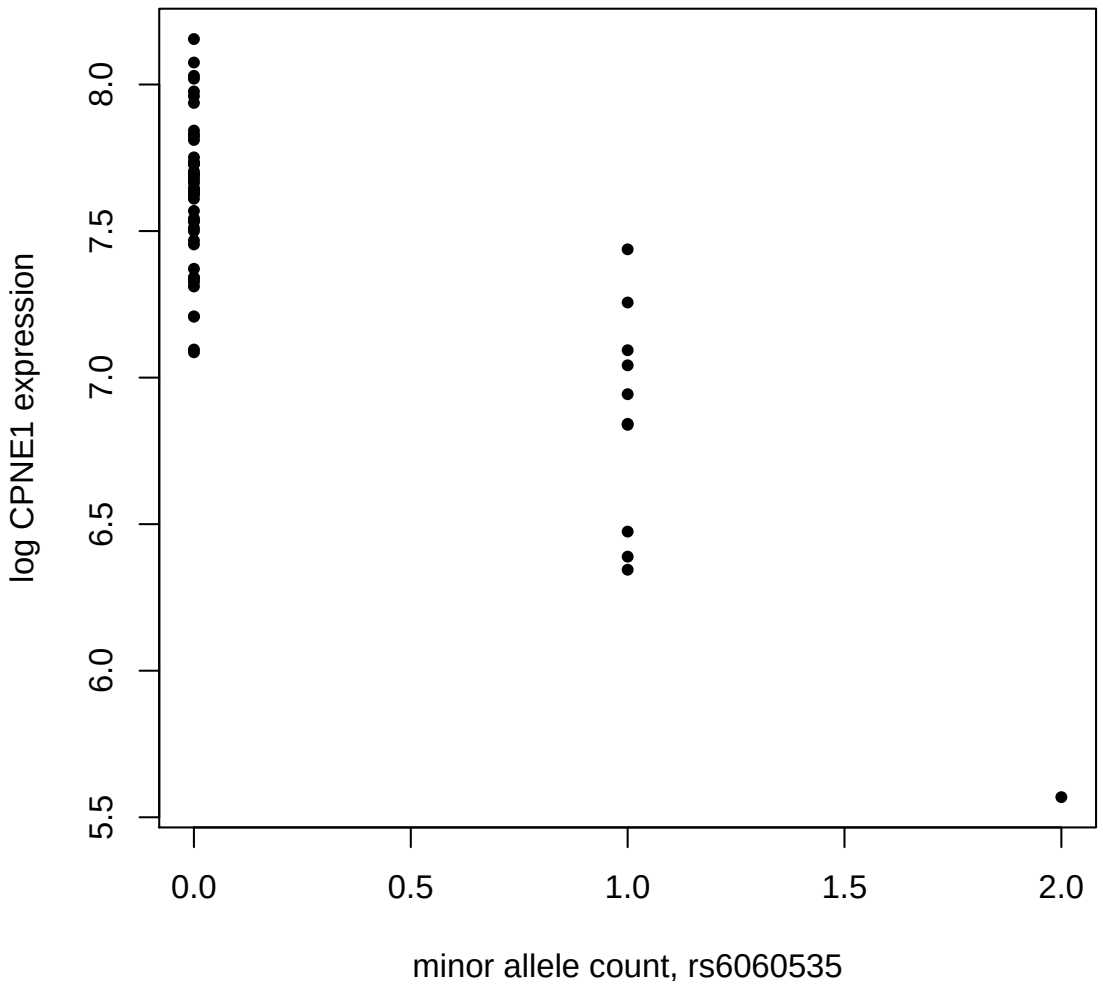
```
> bigRDS1 = sort(exprs(harbChIP)[, "RDS1"], decr = TRUE)[1:5]
> bigRDS1
 YCR107W  YCR105W  YOR001W  YOR258W  YGL157W
5.387419 3.743832 2.676375 2.354235 2.317927
> pData(featureData(harbChIP))[names(.Last.value),8:10]
                                                          FOR_SEQ REV_SEQ_NO
YCR107W                                                      <NA>       <NA>
YCR105W      CCGCTGCTAGGCGCGCCGTGAAAATGCATGTCAAATCTCGGA   YGP24747
YOR001W  CCGCTGCTAGGCGCGCCGTGTAAAAGGTGATTATGTAAAACAAGCG   YGP34235
YOR258W       CCGCTGCTAGGCGCGCCGTGCAAGCTTTCTCGCATTTCTTT   YGP34755
YGL157W CCGCTGCTAGGCGCGCCGTGGTATCACGCTAATTGAAGTTTTTTTTG   YGP27527
                                                          REV_SEQ
YCR107W                                                      <NA>
YCR105W GCAGGGATGCGGCCGCTGACTTCATTTGTTTATCTACCGCTTACATT
YOR001W     GCAGGGATGCGGCCGCTGACATTTTCTATGCGAAGCCTGATGT
YOR258W GCAGGGATGCGGCCGCTGACTTATGATGTTAAAAAGACATGTGTATG
YGL157W GCAGGGATGCGGCCGCTGACTAATTATTTTTGAAACTCTTTTGCAGC
```

Paradigm 3: genetics of gene expression

```
> library(GGtools)
> chr20GGdem
racExSet instance (SNP rare allele count + expression)
rare allele count assayData:
  Storage mode: lockedEnvironment
  featureNames: rs4814683, rs6076506, ..., rs6062370, rs6090120 (117
  Dimensions:
          racs
Features 117417
Samples      58

expression assayData
  Storage mode: lockedEnvironment
  featureNames: 1007_s_at, 1053_at, ..., AFFX-r2-P1-cre-3_at, AFFX-r
  Dimensions:
        exprs
Features  8793
Samples     58
```

```
> exprs(chr20GGdem)[1:5,1:5]
           NA06985  NA06993  NA06994  NA07000  NA07022
1007_s_at 6.236674 5.631134 5.883270 5.791671 5.995744
1053_at   6.535133 6.680420 6.860158 6.298467 6.503476
117_at    4.660155 5.006104 5.018725 4.952051 6.156085
121_at    7.694798 7.331357 7.163441 6.941026 7.361222
1255_g_at 2.831350 2.709704 2.729904 2.723419 3.032477
> snps(chr20GGdem)[1:5,1:5]
          NA06985 NA06993 NA06994 NA07000 NA07022
rs4814683       2       0       0       2       1
rs6076506       0       0       0       0      NA
rs6139074       2       0       0       2       1
rs1418258       2       0       0       2       1
rs7274499       0       0       0       0      NA
```

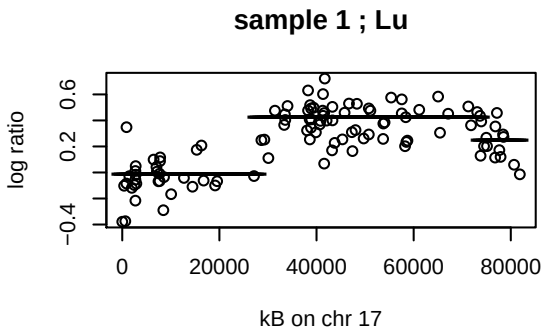

log CPNE1 expression
5.5
6.0
6.5
7.0
7.5
8.0
0.0
0.5
1.0
1.5
2.0
minor allele count, rs6060535

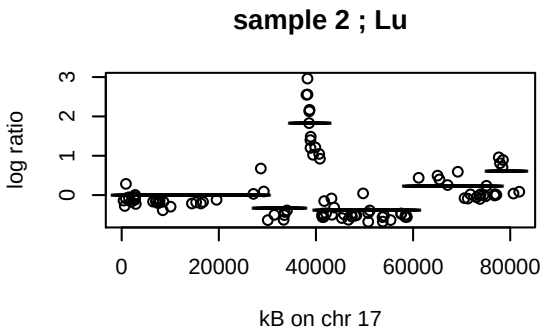
Genetics of gene expression: analyses, mechanics

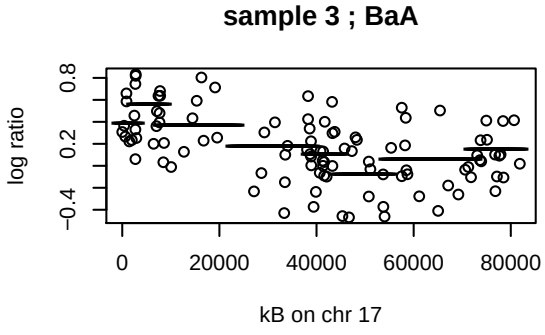
- snpScreen is well-documented, can help with focused tests of association between expression and genotypes
- messy, because gene and SNP location data are in various places
- SQLite annotation paradigm will greatly simplify storage and querying of metadata
- other genotype data representations (e.g., Clayton's snpMatrix) will be interfaced

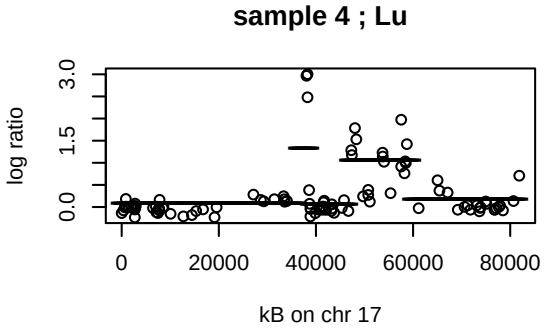
paradigm 4: expression + aCGH

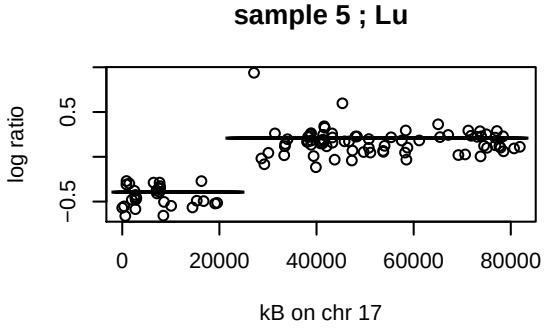
- Neve2006 documents and exemplifies cghExSet class
- ML lab will work with neveCGHmatch; in 2.6 neveExCGH is unified
- logRatios() and exprs(); cloneNames(), cloneMeta()
- NB: the contract for a modeling method – segmentation
- Consider rpart as a device for fitting piecewise constant model
- for free: predict on any compliant data frame, x-validated complexity diagnostic

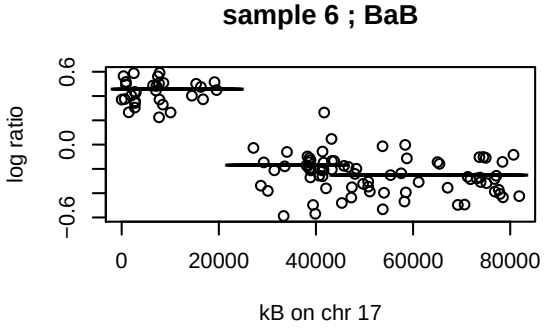

sample 1 ; Lu
log ratio
−0.4
0.2
0.6
0
20000
40000
60000
80000
kB on chr 17


sample 2 ; Lu
log ratio
0
1
2
3
0
20000
40000
60000
80000
kB on chr 17


sample 3 ; BaA
log ratio
−0.4
0.2
0.8
0
20000
40000
60000
80000
kB on chr 17


sample 4 ; Lu
log ratio
0.0
1.5
3.0
0
20000
40000
60000
80000
kB on chr 17


sample 5 ; Lu
log ratio
−0.5
0.5
0
20000
40000
60000
80000
kB on chr 17


sample 6 ; BaB
log ratio
−0.6
0.0
0.6
0
20000
40000
60000
80000
kB on chr 17

Integrative container behavior

- Useful idiom: `X[G, S]` is a selection
  - `G` is a reporter selection predicate
  - `S` is a sample selection predicate
- can we entertain:

  ```
  X[ clonesNear("CPNE1"), homRare("rs6060535") ]
  ```

  and would we want to? This would select copy number data near a certain gene for samples that have a certain genotype. The string parameters might work, or we might need `hugo("CPNE1")`, e.g., to specify semantics
- awaited maturity of SQLite annotation, now can experiment

contracts of statistical modeling procedures

- formula interface is useful
- predict(), plot(), coef(), residuals(), should exist and make sense (comply with reasonable expectations)
- many things that are fitting models (e.g., normalization functions) do not attempt this
- self-discipline is hard; software helpers (e.g., stub generators) are in wide use in other languages

transparency and agnosticism

- DR Cox, AAS 2007

  It is interesting and perhaps surprising that J. W. Tukey, who had an extraordinarily wide-ranging knowledge of the natural sciences down to fine detail, favored largely ignoring that knowledge in the main phases of analysis, introducing it only in the final stages of interpretation.

transparency and agnosticism

- JD Watson, The Double Helix, ch 28.

  Maurice, in a lab devoid of structural chemists, did not have anyone to tell him that all the textbook pictures were wrong.

conclusions

- main jobs: data capture, removal of irrelevant noise, interrogation/modeling, reporting
- multiassay containers coordinate and provide access to diverse measurements
- concise interrogators: generic methods that capitalize on simultaneous access to assay, phenotype, and biological metadata
- modeling functions should respect a well-established tradition of input and return capabilities
- tremendous development of container infrastructure: S. Falcon, M. Morgan, others