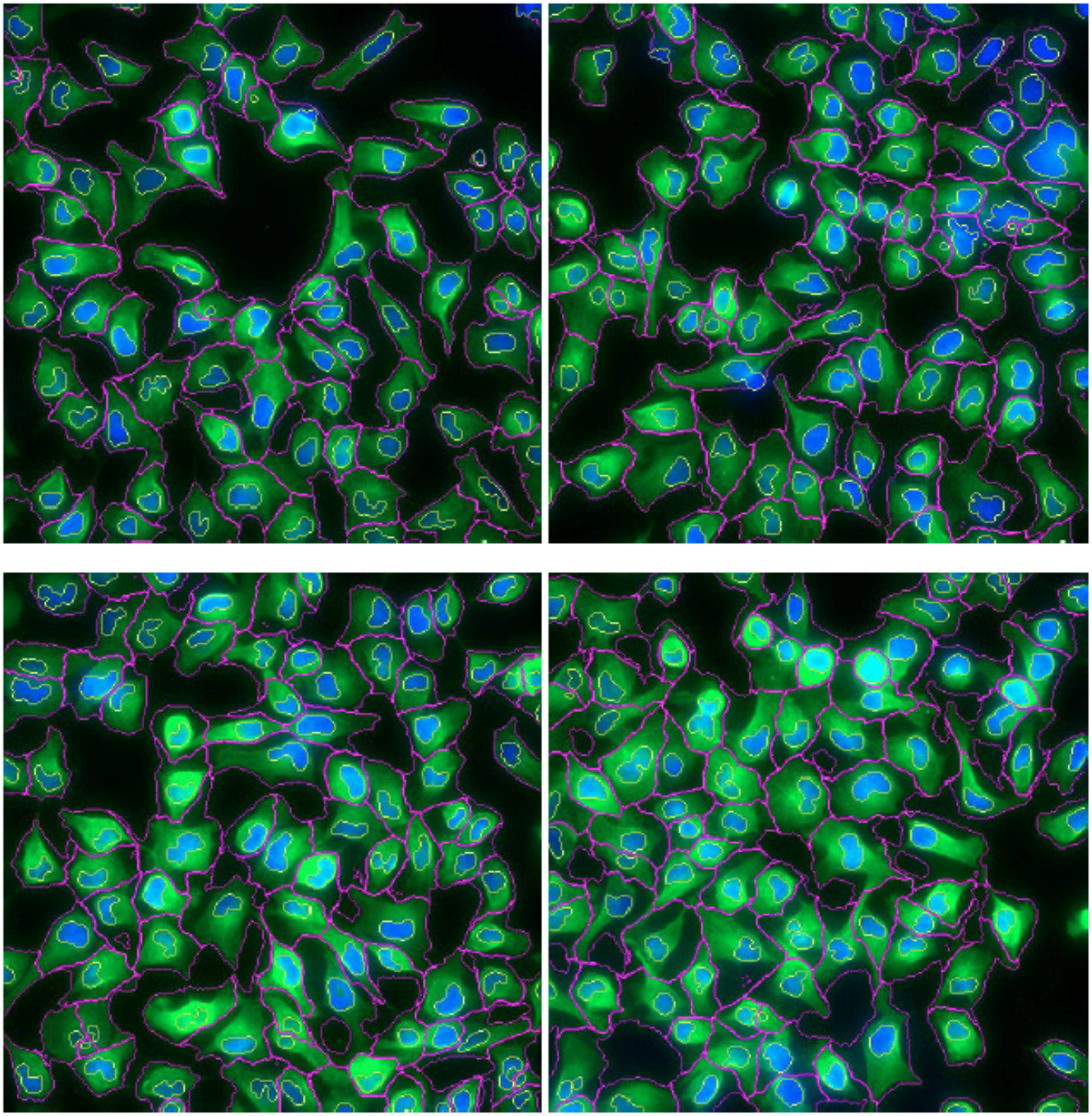
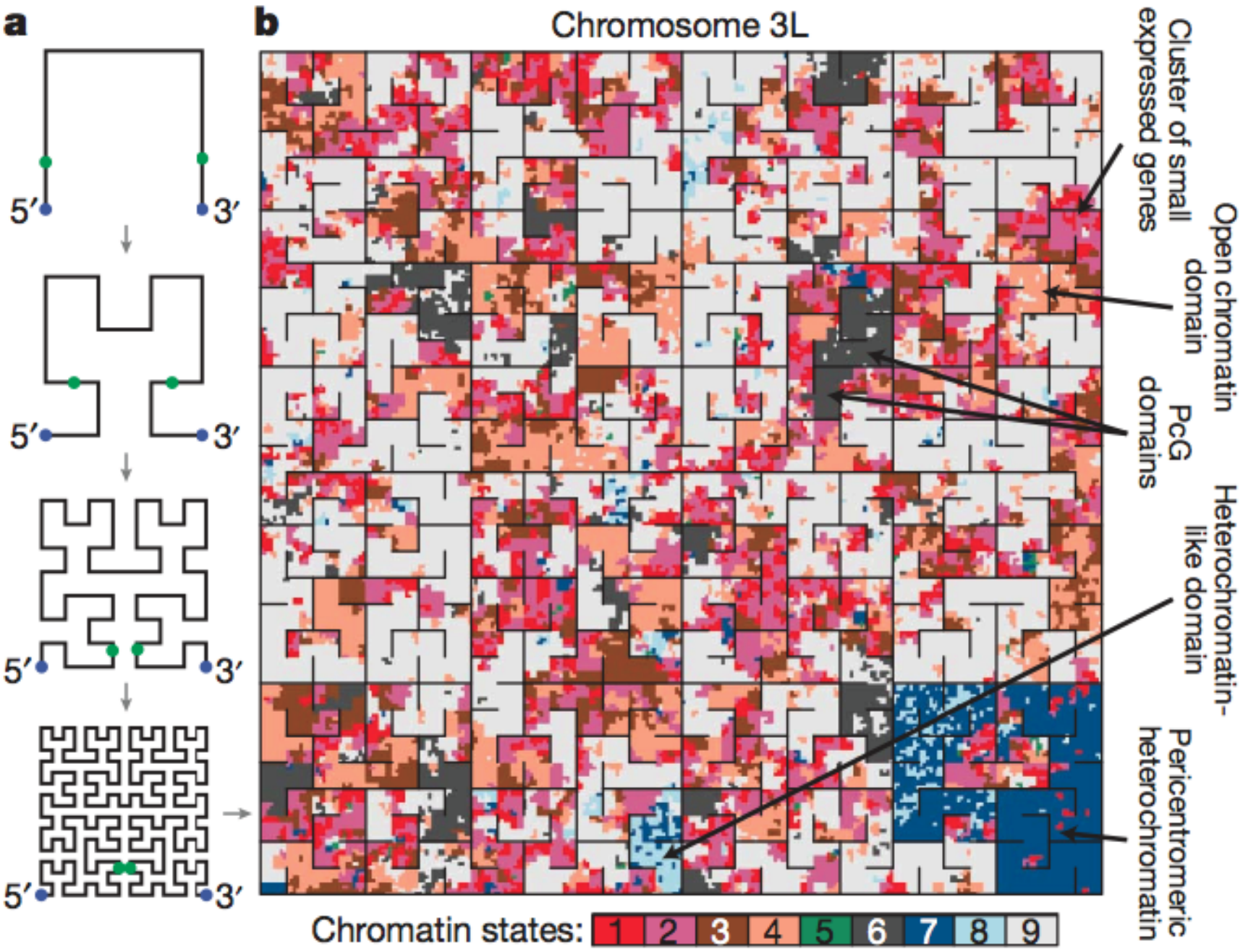
CSAMA 2013: Computational visualization in genomic data analysis

Vince Carey PhD

Harvard Medical School




a
b
Chromosome 3L
5′
3′
5′
3′
5′
3′
5′
3′
Cluster of small expressed genes
Open chromatin domain
PcG domains
Heterochromatin-like domain
Pericentromeric heterochromatin
Chromatin states:
1
2
3
4
5
6
7
8
9

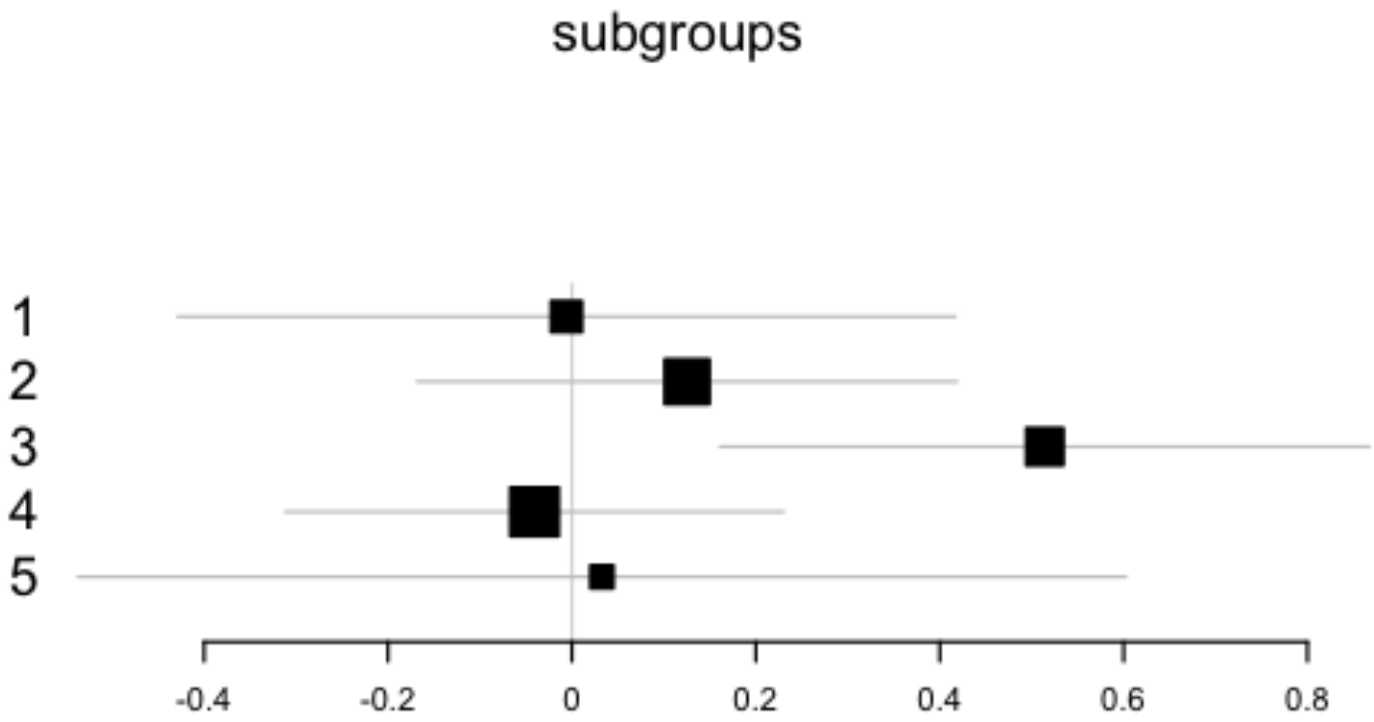
Road map of the talk

- Motivations for computational visualization
  - Concrete accompaniment to more abstract statistical tabulations/inferences
  - Exploring an observational frontier
- Plotting and basic statistics (design, univariate, multivariate, multiple comparisons)
- Grammar of graphics concepts, deployment against Yeast Cell Cycle expression archive
- ggbio applications to GWAS
- Object designs for CCLE and a cancer regulatory network

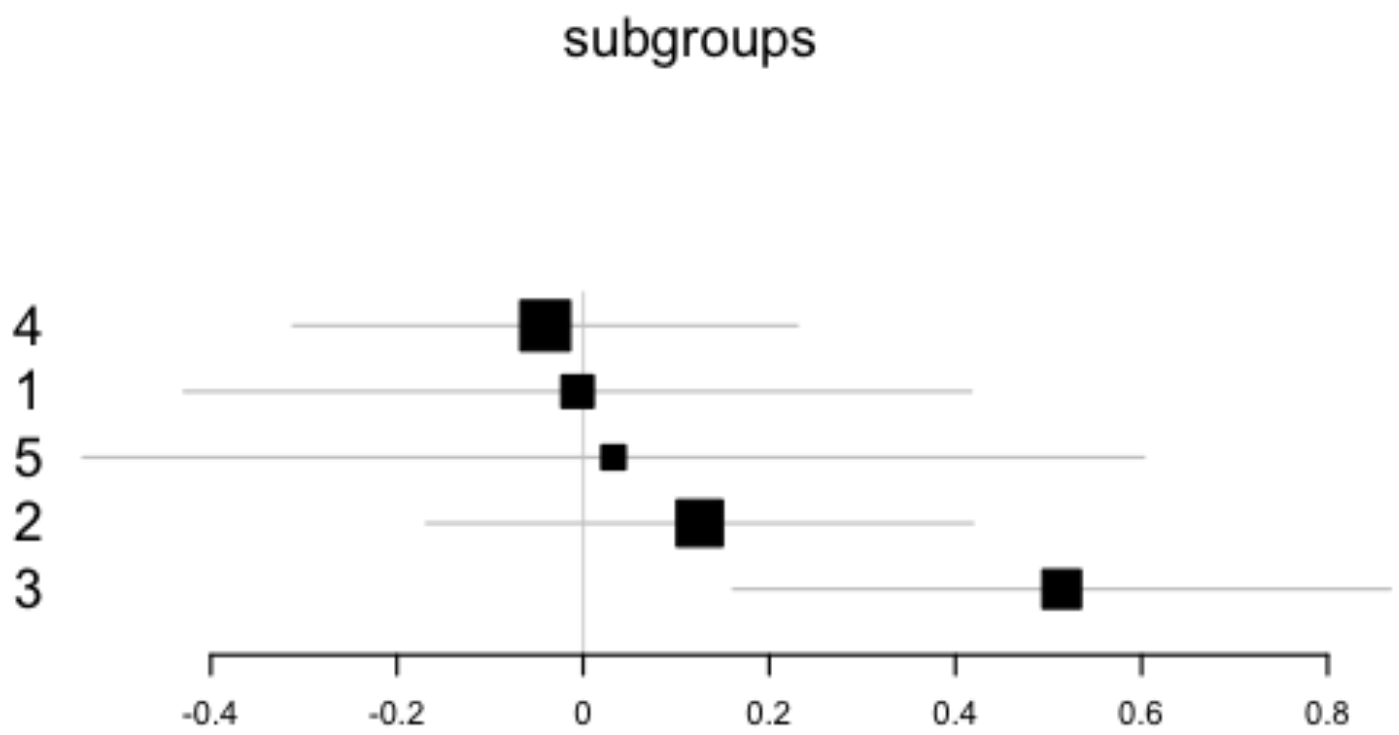
Three principles

- Visualizations should be obtained using a program: no photoshop
  - Steps by which snapshots of interactive visualizations are obtained should have a textual representation
- The process by which a given visualization was selected (from among relevant variations) should be disclosed
- Uncertainty and variability can be hard to depict but attempts should be made to indicate:
  - Roles of modeling assumptions
  - Scope and sources of measurement variation

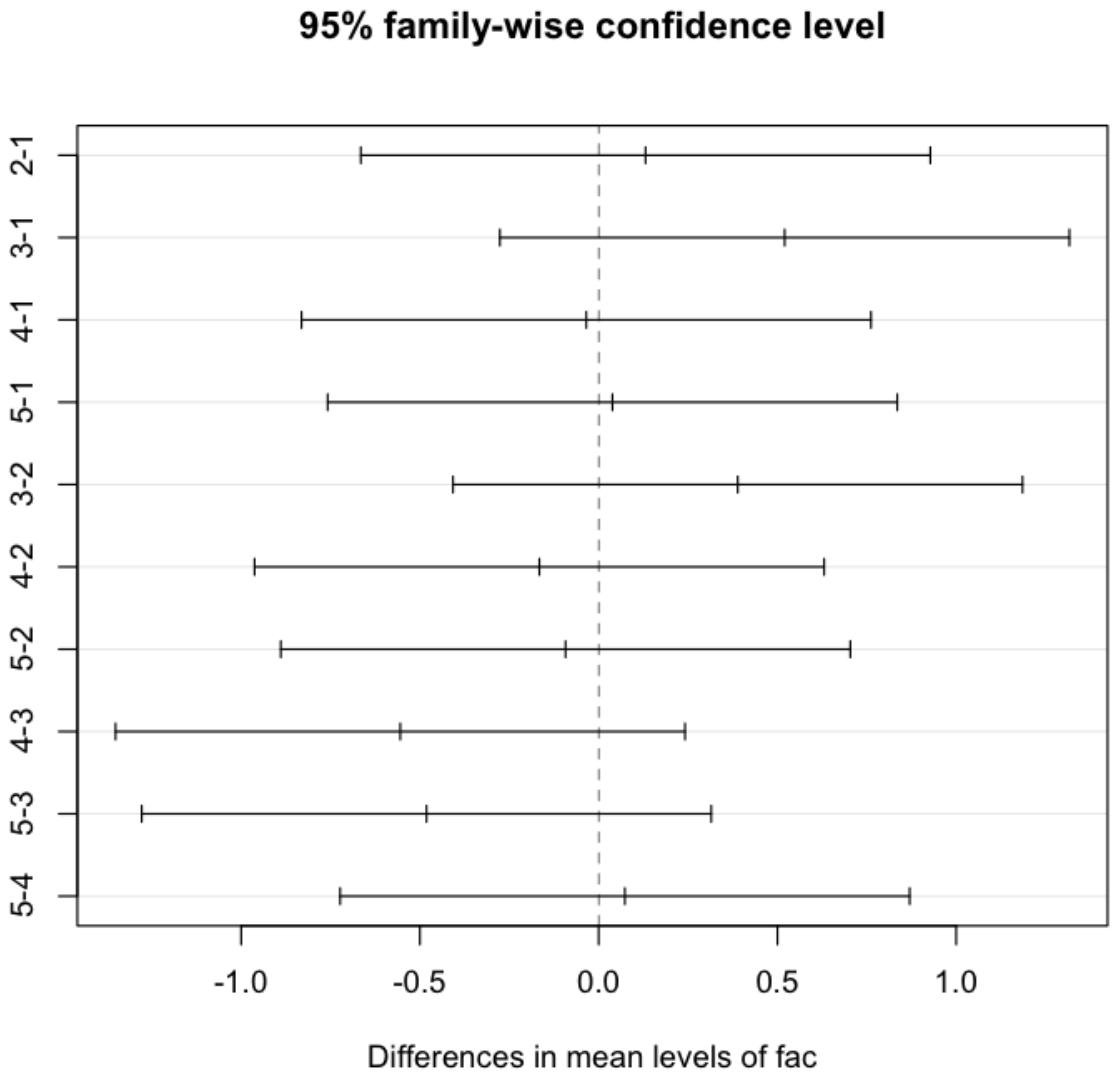
Report to an academy



“looks like there is something there”



Honest pairwise comparisons



A basic dilemma

- Use visualization to discover/communicate relationships in data
- Avoid choosing visualizations that overstate the strength of relationship
  - Are there trustworthy practices?
  - Can we guide the viewer to results of a principled statistical analysis? Relationships that are highly likely to "independently replicate"?

Consolidated Standards of Reporting Trials

From Wikipedia, the free encyclopedia

CONSORT (**Consolidated Standards Of Reporting Trials**) encompasses various initiatives developed by the CONSORT Group to alleviate the problems arising from inadequate reporting of randomized controlled trials.

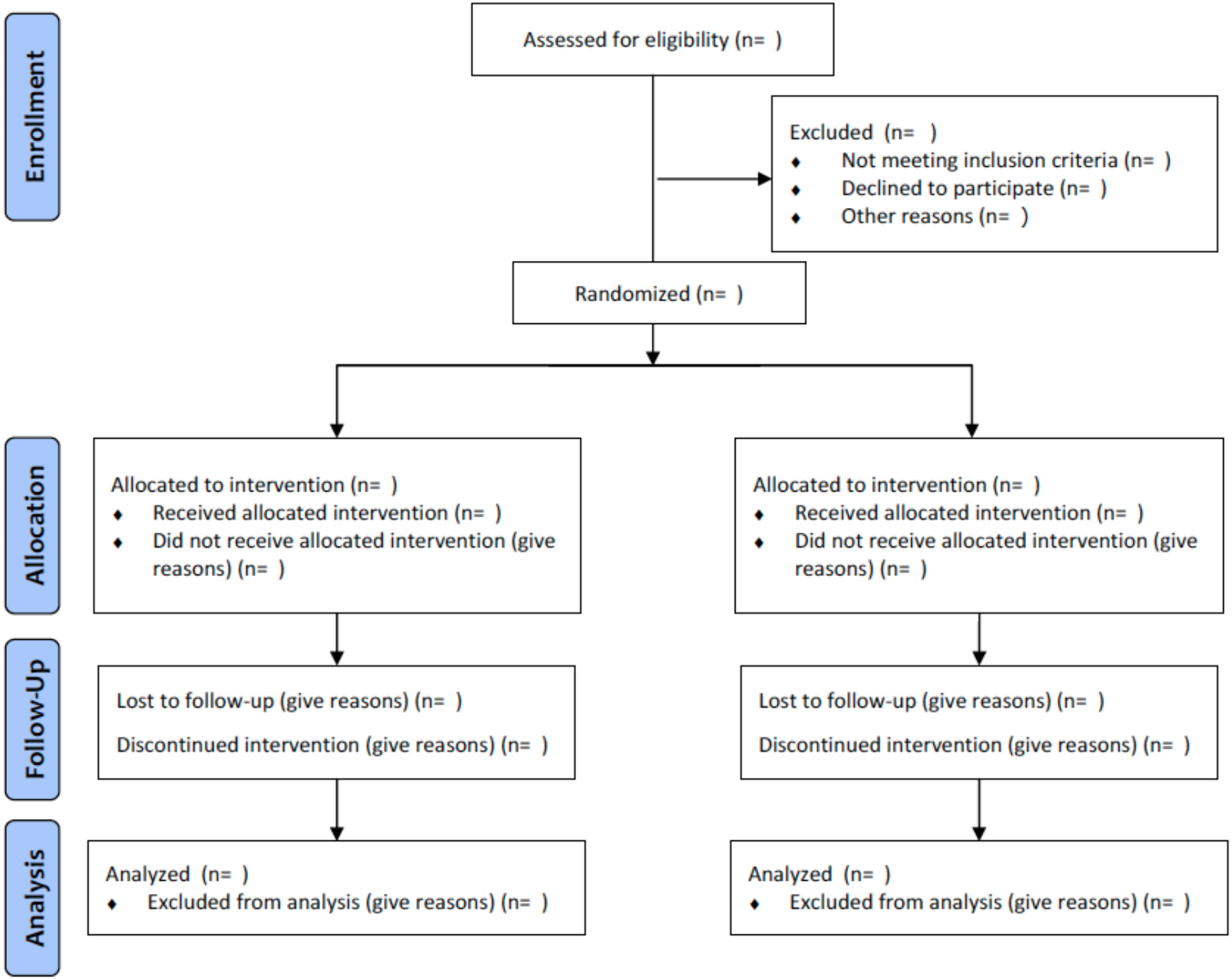
Contents [hide]

The CONSORT Statement [edit]

The main product of the CONSORT Group is the CONSORT Statement,[1] which is an evidence-based, minimum set of recommendations for reporting randomized trials. It offers a standard way for authors to prepare reports of trial findings, facilitating their complete and transparent reporting, reducing the influence of bias on their results, and aiding their critical appraisal and interpretation.

Communicate about the experimental design

- Objectives
- Demonstration of validity and satisfactory power of proposed test procedure
- Illustration of development of the dataset from screening to analysis


Enrollment
Assessed for eligibility (n=)
Excluded (n=)
Not meeting inclusion criteria (n=)
Declined to participate (n=)
Other reasons (n=)
Randomized (n=)
Allocation
Allocated to intervention (n=)
Received allocated intervention (n=)
Did not receive allocated intervention (give reasons) (n=)
Allocated to intervention (n=)
Received allocated intervention (n=)
Did not receive allocated intervention (give reasons) (n=)
Follow-Up
Lost to follow-up (give reasons) (n=)
Discontinued intervention (give reasons) (n=)
Lost to follow-up (give reasons) (n=)
Discontinued intervention (give reasons) (n=)
Analysis
Analyzed (n=)
Excluded from analysis (give reasons) (n=)
Analyzed (n=)
Excluded from analysis (give reasons) (n=)

R/Bioconductor and scientific visualization

- Integrative, self-describing containers
  - `ExpressionSet, SummarizedExperiment, GRanges`
- Packages that define workflow components
  - *affy, oligo, limma, ShortRead, DESeq, GSEAlm* – these often include their own visualization funcs.
- Packages that specialize in visualization
  - *geneplotter, ggbio* (focused *ggplot2*), *Gviz, Rgraphviz, HilbertVis*

By the end of the course

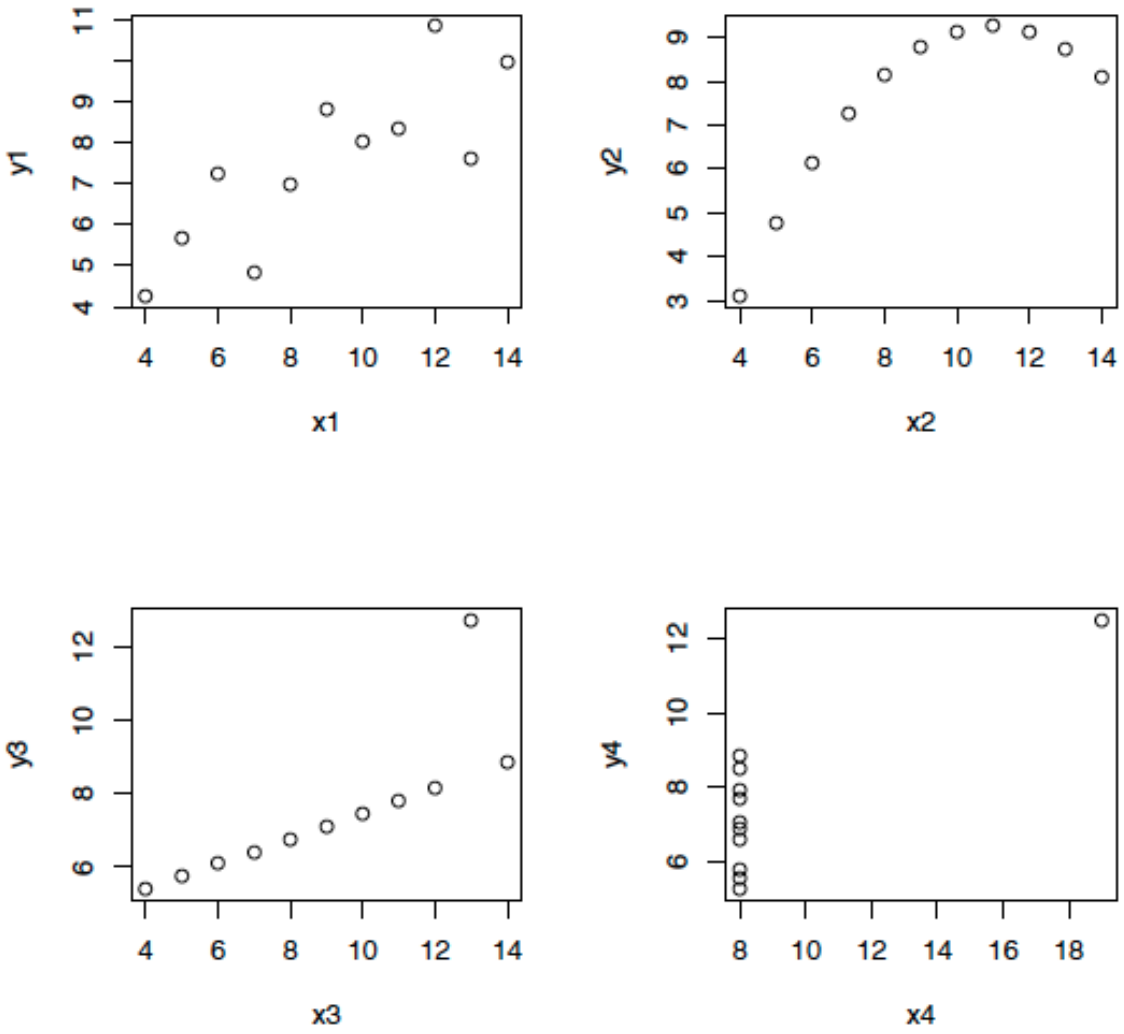
- Understand the rationale for the container designs, and the ones you will actually use
- Understand why functions have been packaged up as they have been
- Get a sense of the discipline required to use the containers and packages vs. files and scripts
- Sharpen your statistical acumen … recognize the good arguments, improve the flawed ones

A reproducible experiment? Four labs assay a sample of blinded origin on two quantities – as regulator, do you declare them consistent?

```
> cor.test(x1, y1)

        Pearson's product-moment correlation

data:  x1 and y1
t = 4.2415, df = 9, p-value = 0.00217
alternative hypothesis: true correlation is not
95 percent confidence interval:
 0.4243912 0.9506933
sample estimates:
      cor
0.8164205

> cor.test(x2, y2)

        Pearson's product-moment correlation

data:  x2 and y2
t = 4.2386, df = 9, p-value = 0.002179
alternative hypothesis: true correlation is not
95 percent confidence interval:
 0.4239389 0.9506402
sample estimates:
      cor
0.8162365
```

```
> attach(anscombe)
> cor.test(x3, y3)

        Pearson's product-moment correlation

data:  x3 and y3
t = 4.2394, df = 9, p-value = 0.002176
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4240623 0.9506547
sample estimates:
      cor
0.8162867

> cor.test(x4, y4)

        Pearson's product-moment correlation

data:  x4 and y4
t = 4.243, df = 9, p-value = 0.002165
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4246394 0.9507224
sample estimates:
      cor
0.8165214
```

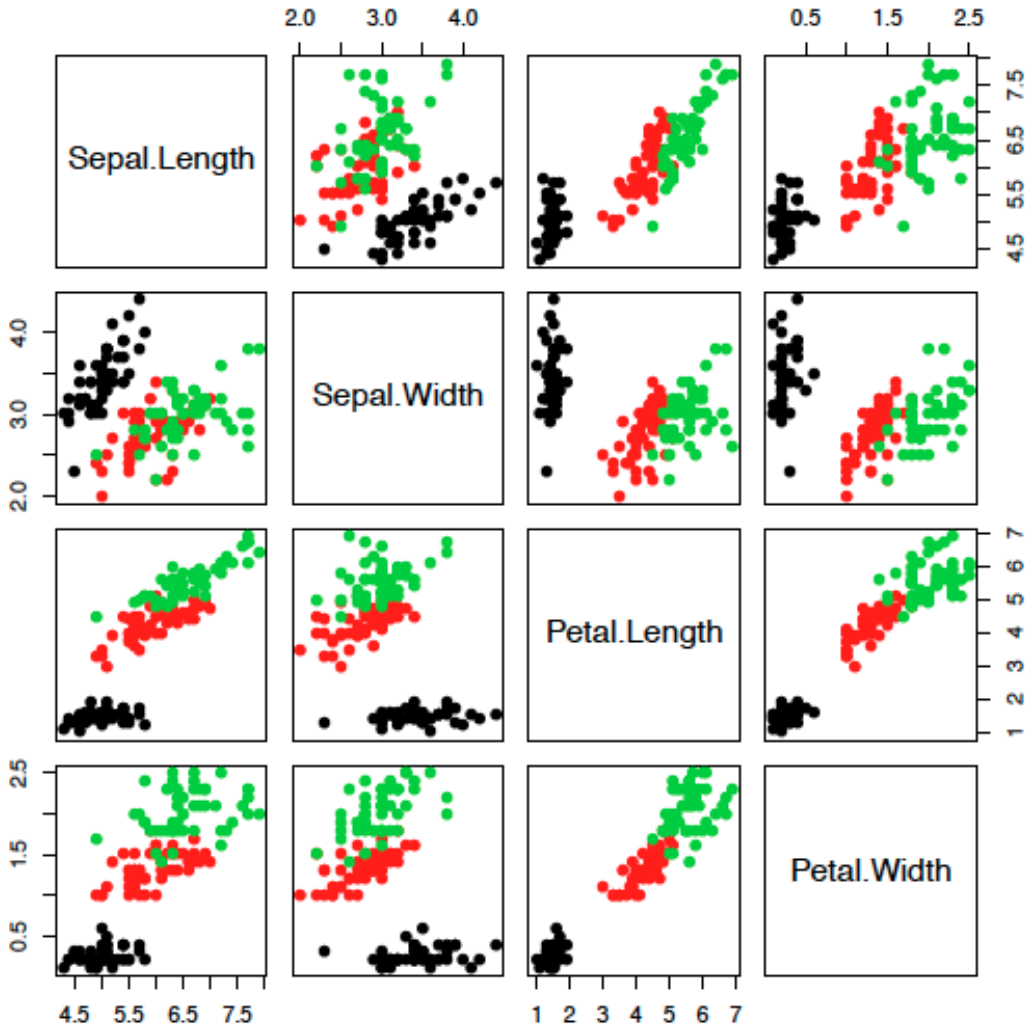
When we look at the data….



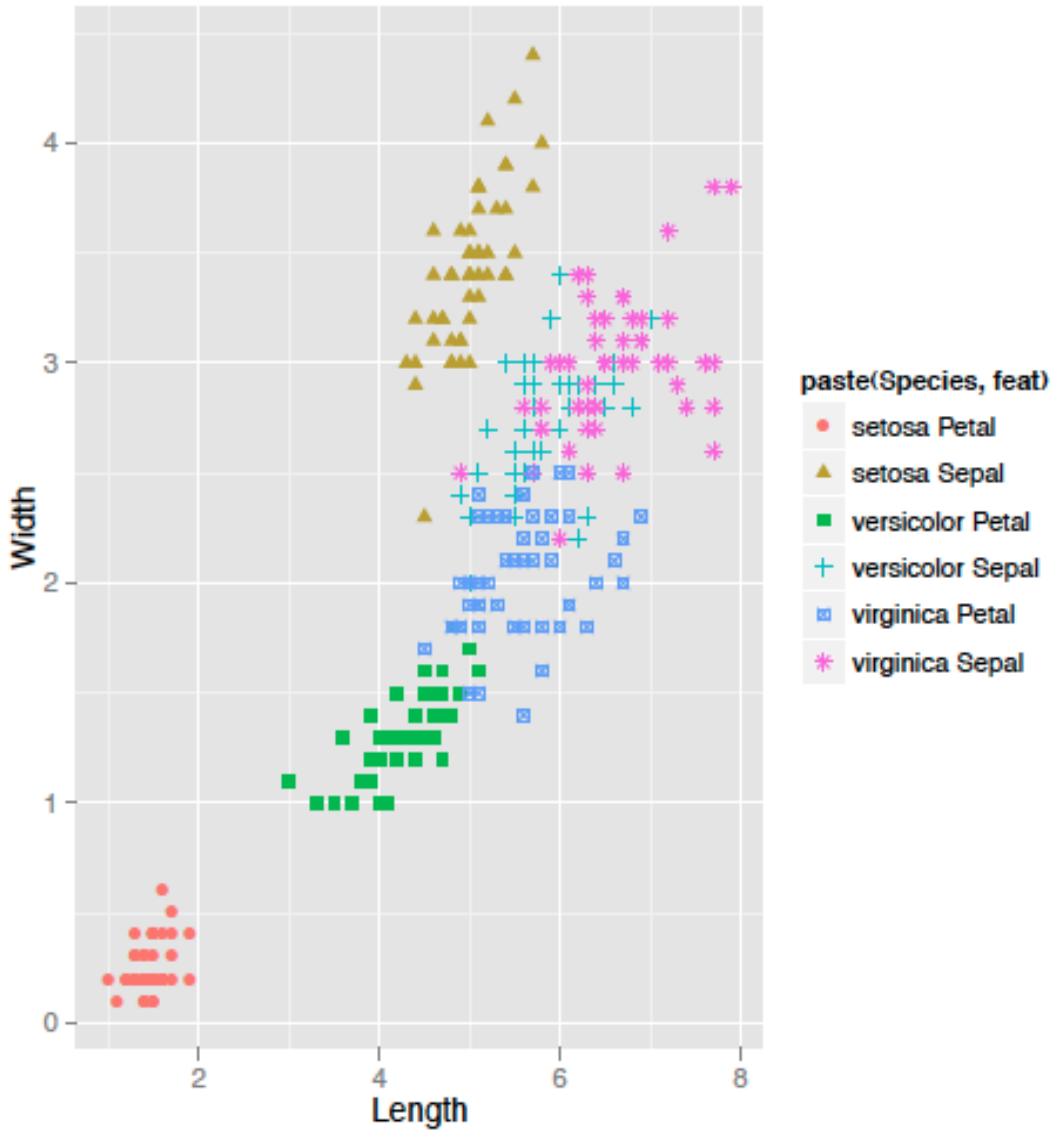
Moral of the Anscombe data?

- "High-quality" summary statistics (e.g., estimators that are unbiased, efficient under mild regularity conditions …) can hide a lot of scientifically relevant information
  - Sometimes a good visualization will reveal important structures
- **Always** create an opportunity to **look**
- Fold visualizations at various scales into your reports

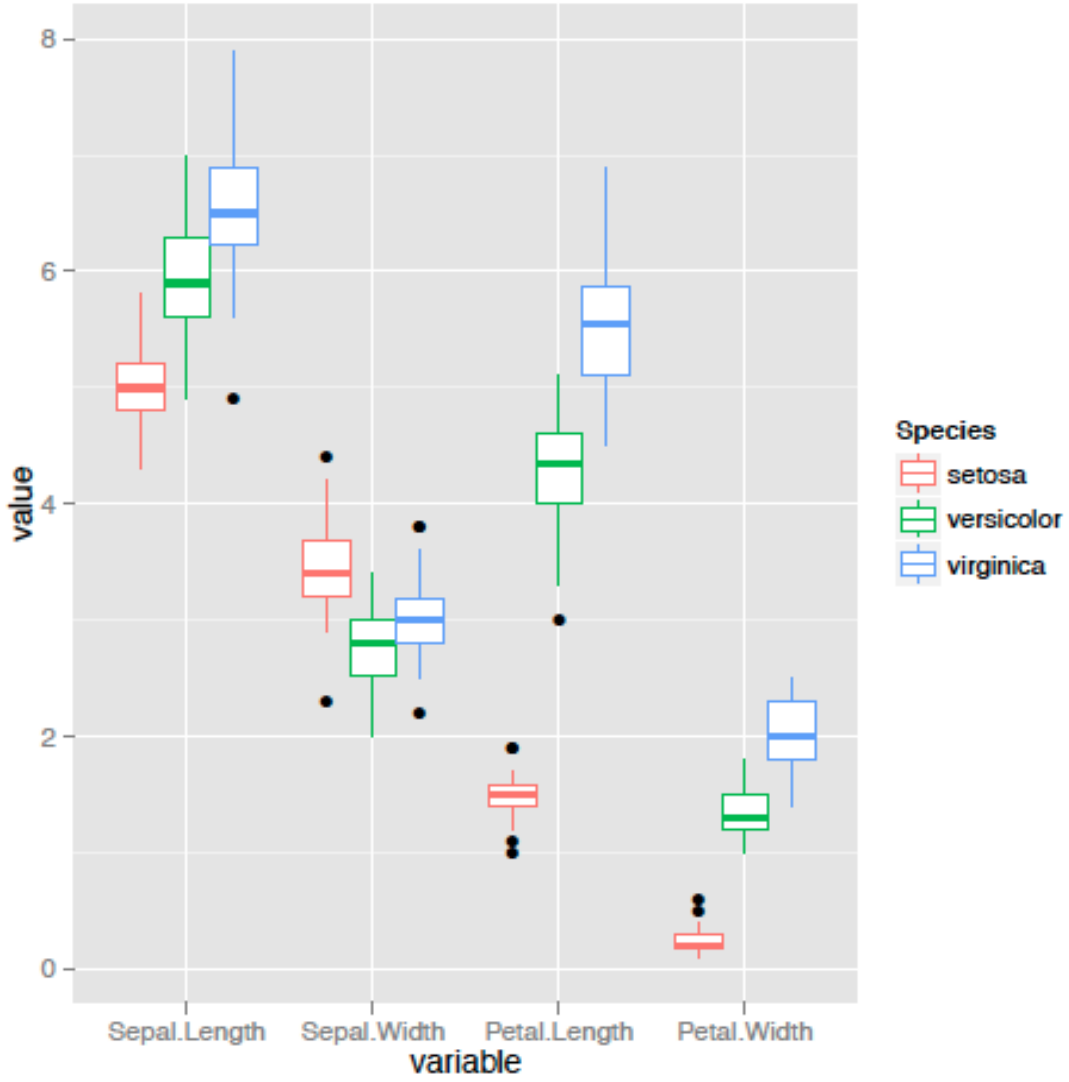
Displaying 150x4 iris measurements with `pairs()`



A unified view with *ggplot2*

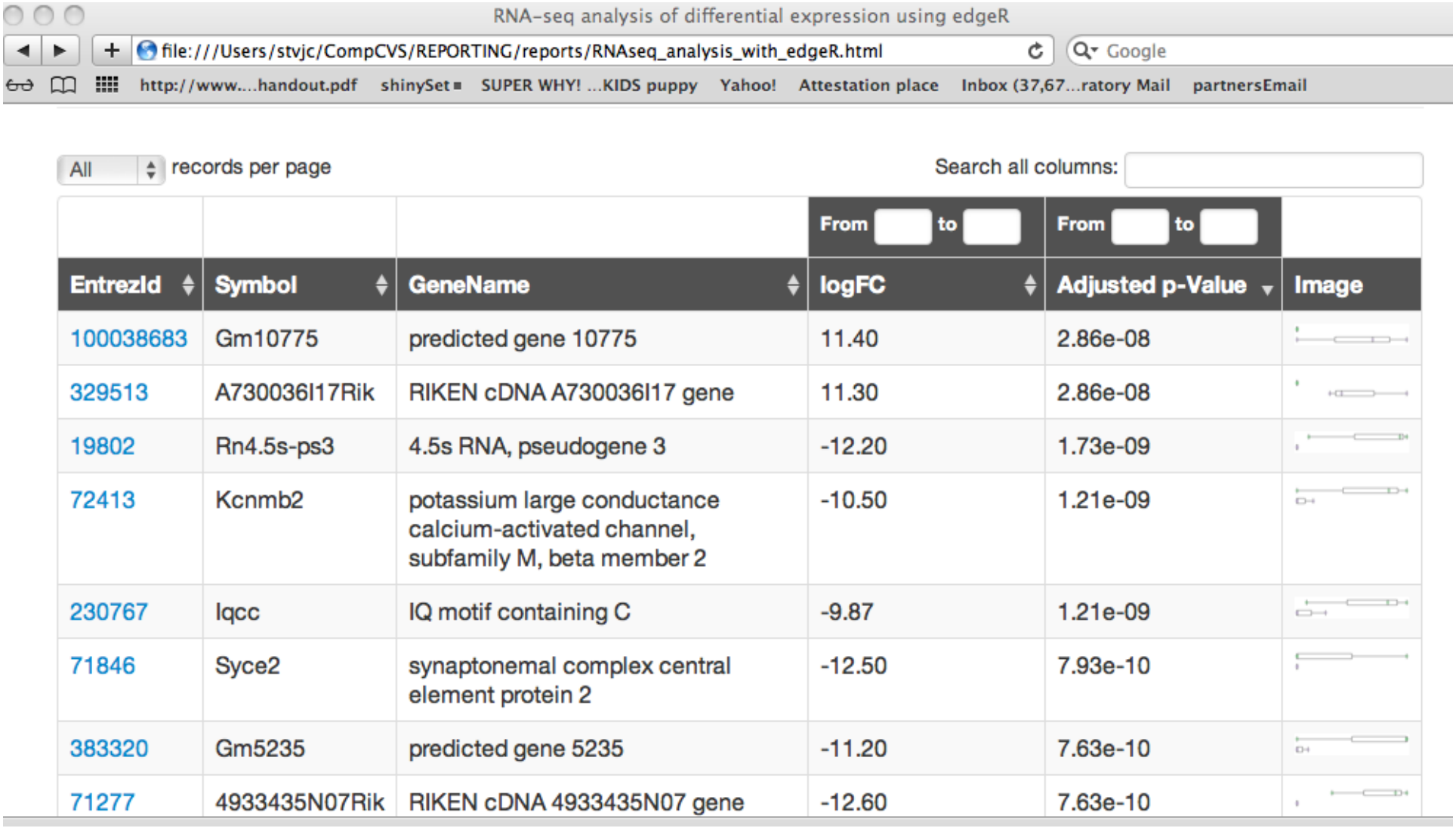


Marginal distributions for species discrimination



A detour: large-scale use of boxplots

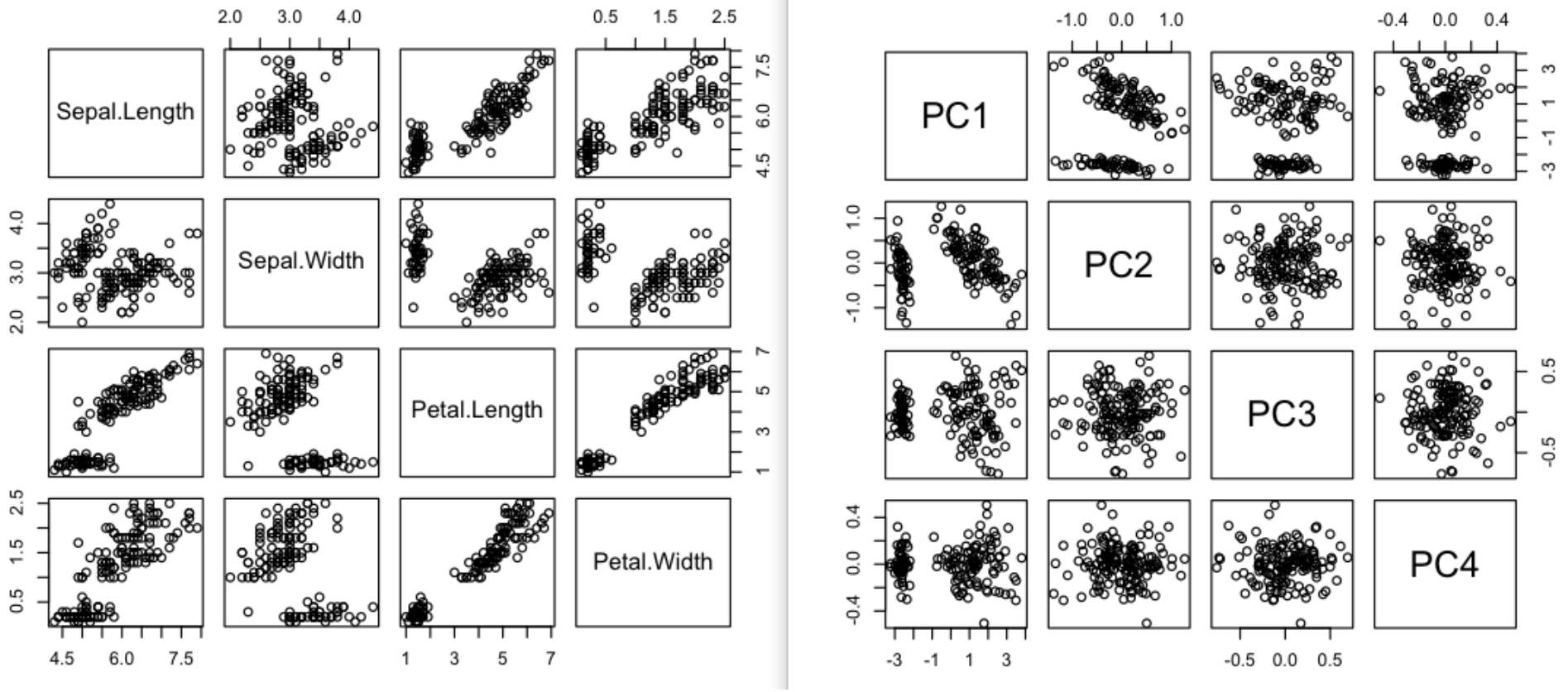
- The *ReportingTools* package is an important new contribution that allows developers to synthesize approaches to analysis and visualization for high-level communications
- Target medium is a modern browser with significant client-side capabilities permitting search, sorting, etc.


RNA-seq analysis of differential expression using edgeR
file:///Users/stvjc/CompCVS/REPORTING/reports/RNAseq_analysis_with_edgeR.html
Google
http://www....handout.pdf
shinySet
SUPER WHY! ...KIDS puppy
Yahoo!
Attestation place
Inbox (37,67...ratory Mail
partnersEmail
All
records per page
Search all columns:
From
to
From
to
EntrezId
Symbol
GeneName
logFC
Adjusted p-Value
Image
100038683 Gm10775 predicted gene 10775 11.40 2.86e-08
329513 A730036I17Rik RIKEN cDNA A730036I17 gene 11.30 2.86e-08
19802 Rn4.5s-ps3 4.5s RNA, pseudogene 3 -12.20 1.73e-09
72413 Kcnmb2 potassium large conductance calcium-activated channel, subfamily M, beta member 2 -10.50 1.21e-09
230767 Iqcc IQ motif containing C -9.87 1.21e-09
71846 Syce2 synaptonemal complex central element protein 2 -12.50 7.93e-10
383320 Gm5235 predicted gene 5235 -11.20 7.63e-10
71277 4933435N07Rik RIKEN cDNA 4933435N07 gene -12.60 7.63e-10

Summary thus far

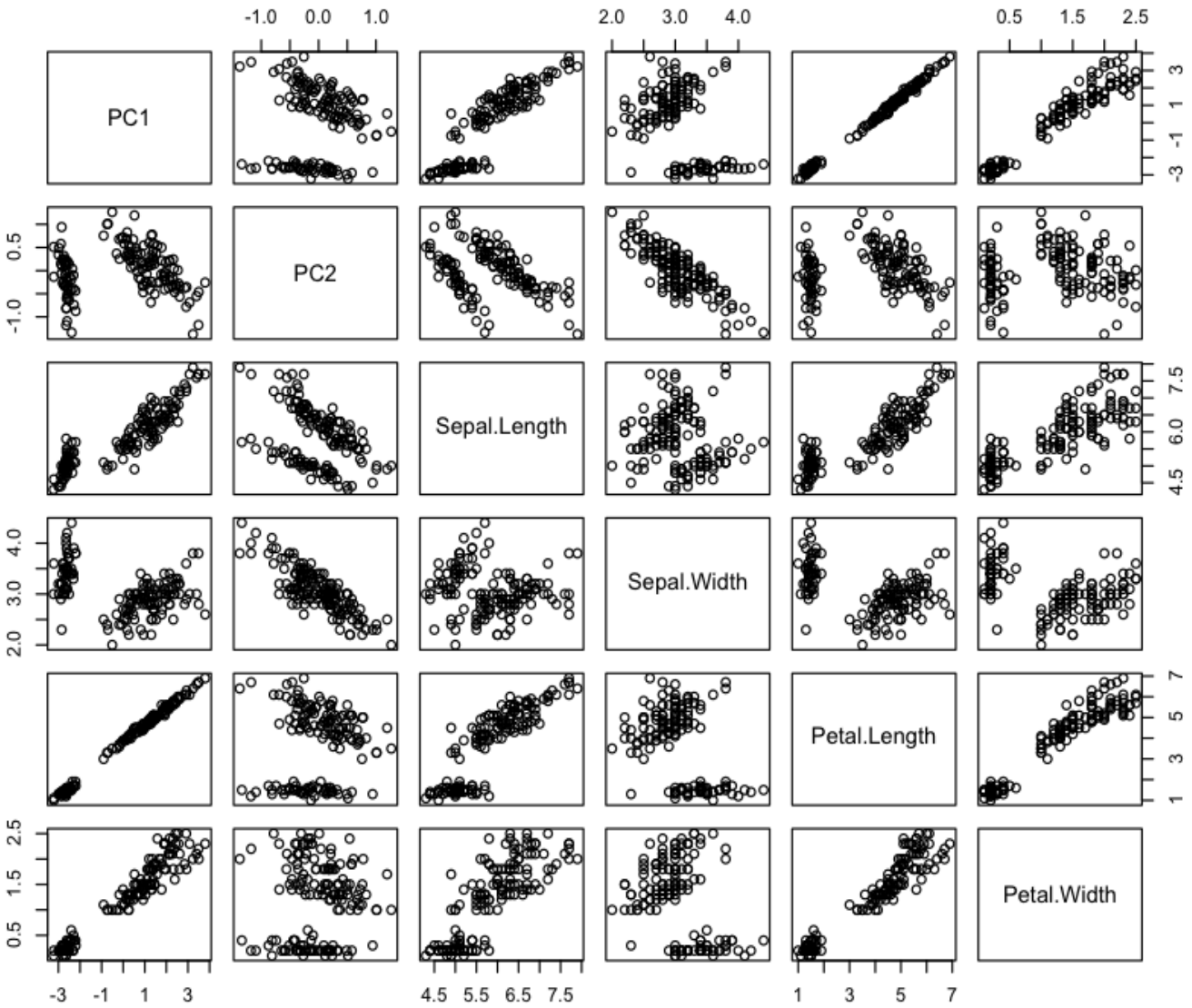
- Visuals are to help with interpretation
  - Interpretation is difficult without clear sense of objectives of underlying experiment
  - Experiments are often messy … data filtering diagram like CONSORT should be standard practice
- Marginal and pairwise joint distributions are easy to visualize
  - Marginal and pairwise statistics/tests can obscure important patterns
  - Efficient enhancements to tables (ReportingTools)

Principal components reexpression of a multivariate dataset

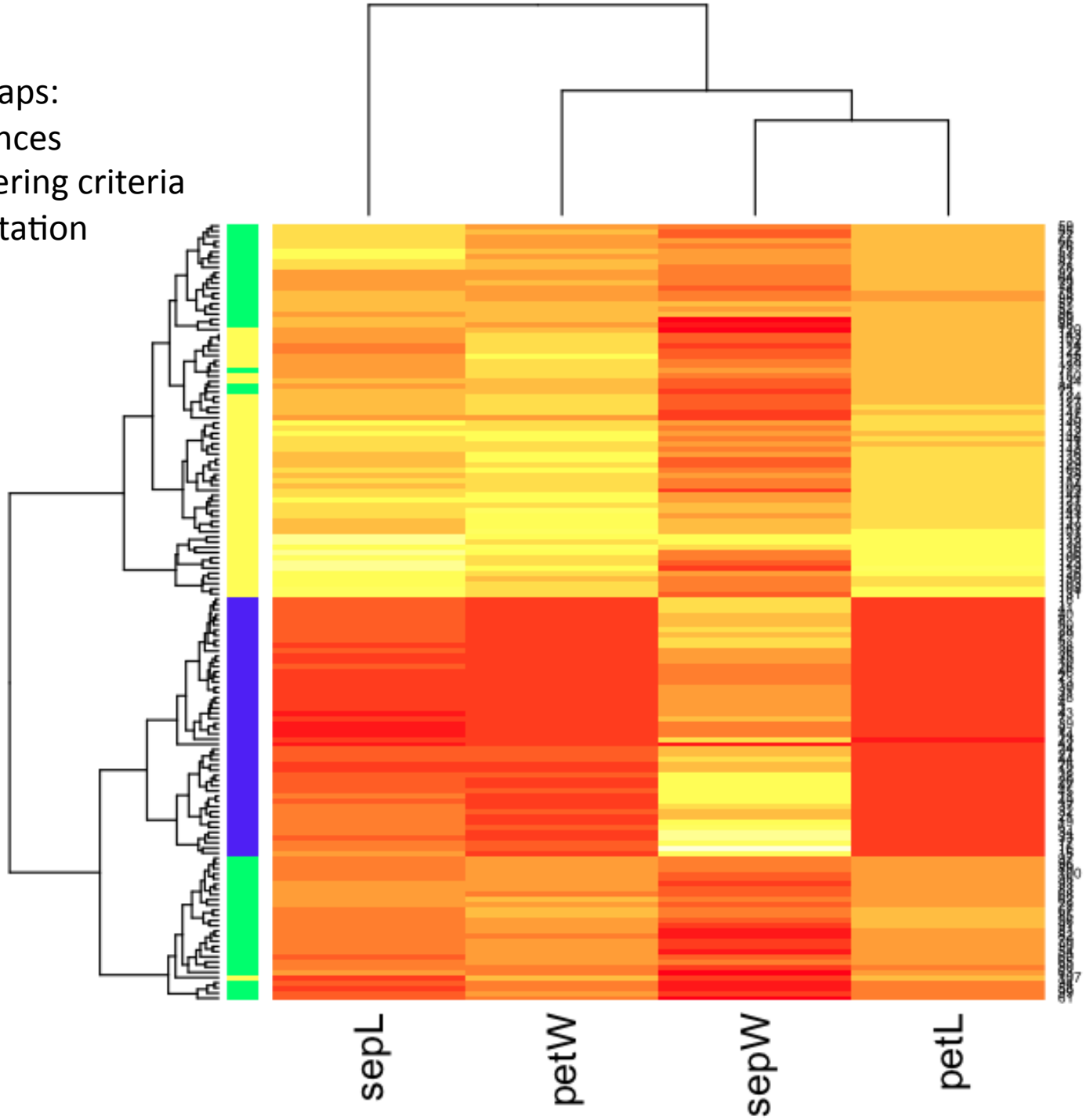


PC1 is the linear combination of the original variables that has maximum variance among all linear combinations of the original variables; PC2 is MVLC among all LC uncorrelated with PC1, …

Interpreting the “loading” on PCs



Heatmaps:
- distances
- clustering criteria
- annotation



Under the hood with dendrograms: c1 = hclust(dist(iris[,1:4]))

hclust (*, "complete")

Choice of features and distance for comparing and clustering objects are key determinants of results of cluster analyses. The exercises involve assessment of distances used in this simple hierarchical clustering. Consider how to assess the sensitivity of the cluster assignments to choice of feature set and distance function.

Exercise: Evaluate `names(c1)`. Explain the value of `c1$merge[1,]` (consult the help page for `hclust`).

Explain:

```
> which(c1$height>.25)[1]
[1] 44
> c1$merge[44,]
[1] -69 -88
> dist(iris[c(69,88),-5])
          69
88 0.2645751
```

Summary

- PCA is widely used for “dimension reduction”, “eigengene” reexpression, QA, removal of extraneous variation
- Cluster analyses are commonly used and underlie many prominent displays/analyses
  - Highly tunable
  - R and other tools hide complexity: “don’t believe in magic”, know how to open the box

“Grammar of graphics”

4.1 Layered grammar of graphics

Briefly, layers of data graphics consist of

- data (variables and observations, in tabular form) and aesthetic mappings (which variable will be used as x, which as y, which to choose glyphs or colors)
- statistics (transformations of variables such as binnings, smooths, boxplot quantities)
- geometric objects (choice of points, lines, polygons) to communicate aspects of data
- position adjustments (jittering, dodging)

Layers are brought to view with selections of scales, coordinate systems, and facets that reflect groupings.

In R, there is a strong connection between convenience of visualization or specification of desired specification and underlying data structure. Data reshaping is a high-level activity particularly when dealing with measurements over time. We’ll contrast two approaches to visualizing expression trajectories in yeast colonies.

Deploying grammar of graphics directly on an expression archive

4.2 The ggplot2 approach

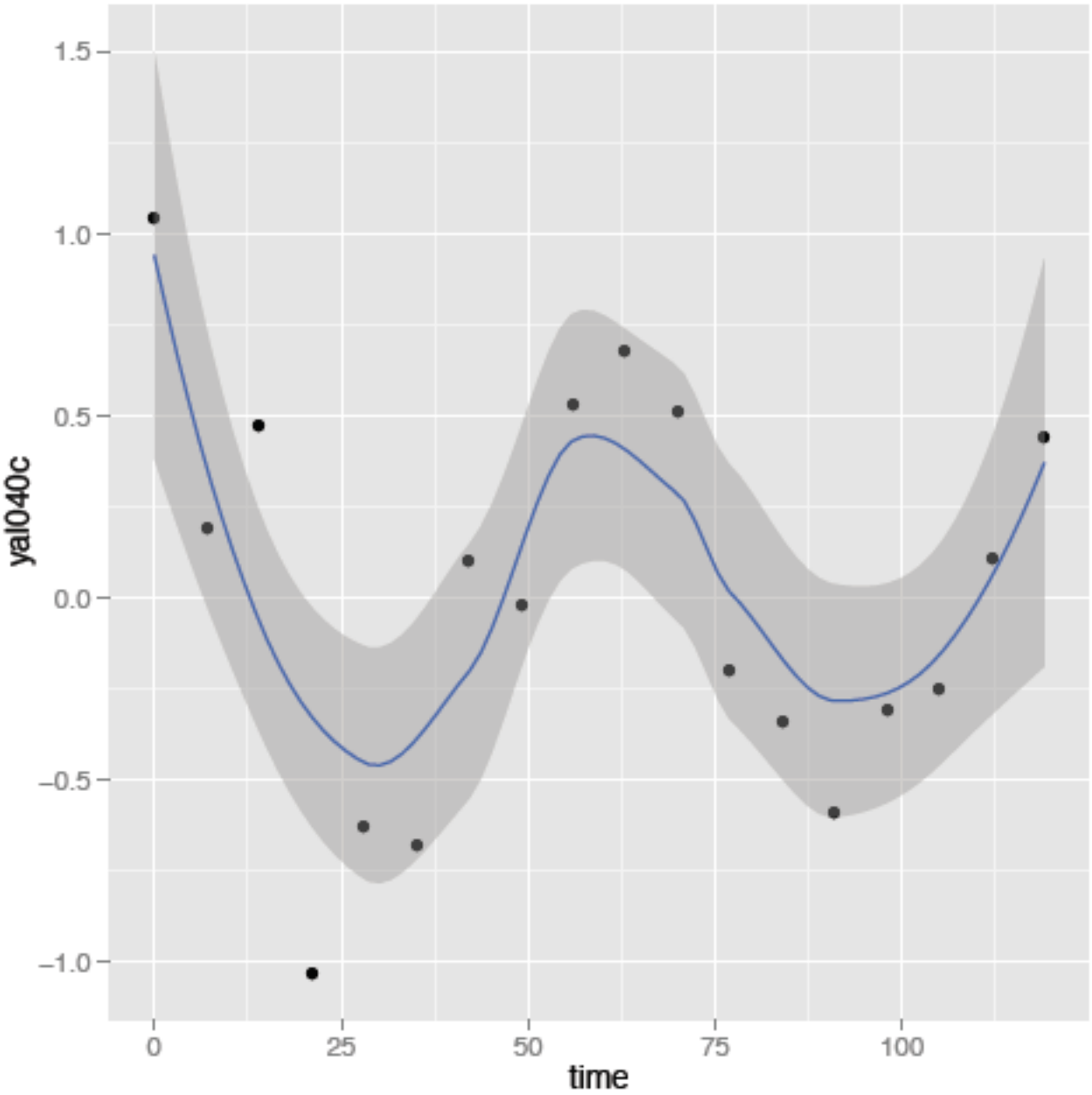
There are two key phases to plotting with ggplot2.
We initialize:

```
'ggplot()' initializes a ggplot object. It can be used to declare
   the input data frame for a graphic and to specify the set of plot
   aesthetics intended to be common throughout all subsequent layers
   unless specifically overridden.
```

and then we specify how to render by building up layers of representations of information in the data.

4.2.1 A smoothed enhancement to a trajectory scatterplot

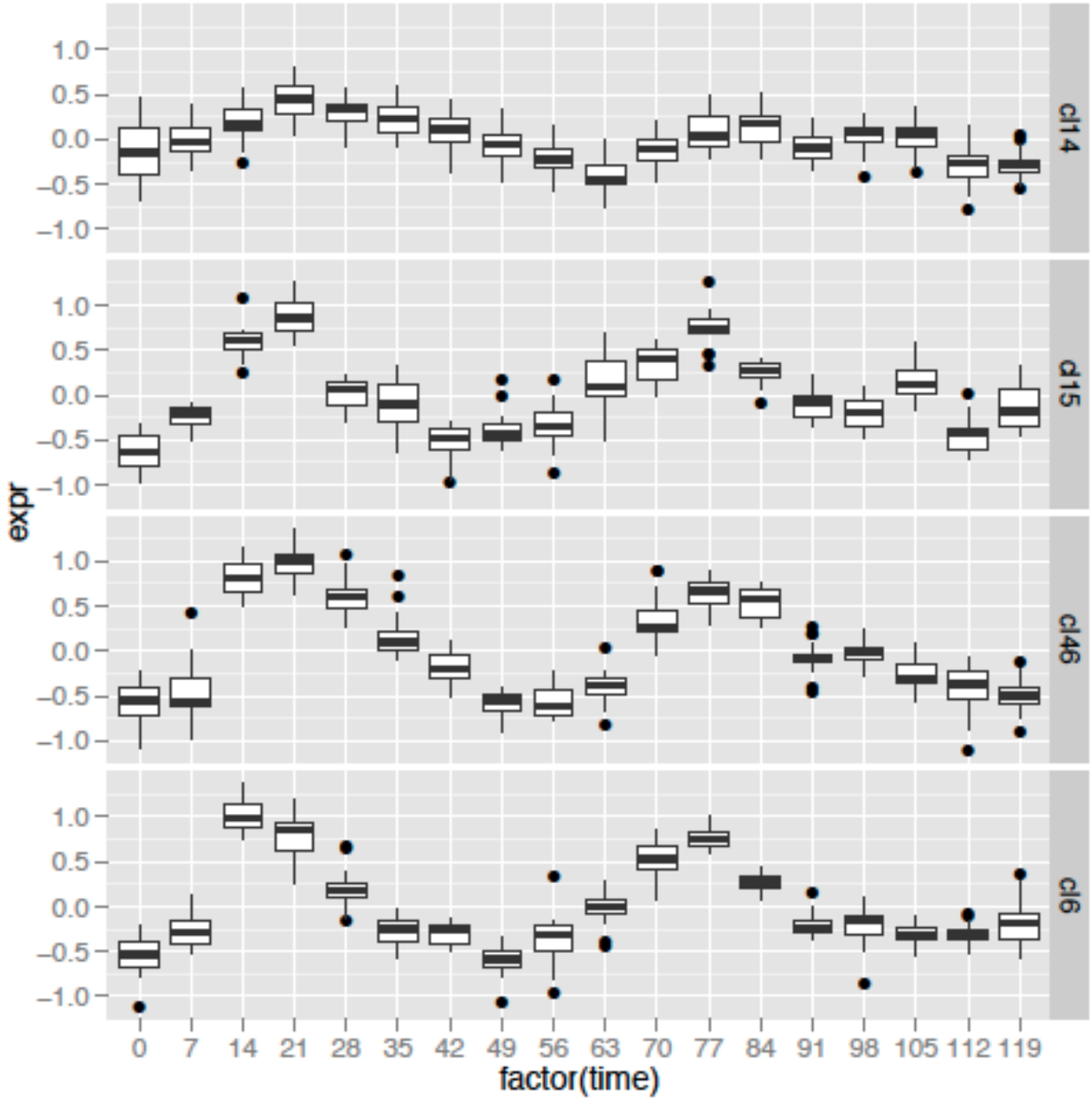
```
> library(ggplot2)
> df = data.frame(yal040c = exprs(alp)["YAL040C", ], time = alp$time)
> p1 = ggplot(data = df, aes(y = yal040c, x = time))
> p1 + geom_point() + stat_smooth()
```


1.5
1.0
0.5
0.0
−0.5
−1.0
yal040c
0
25
50
75
100
time

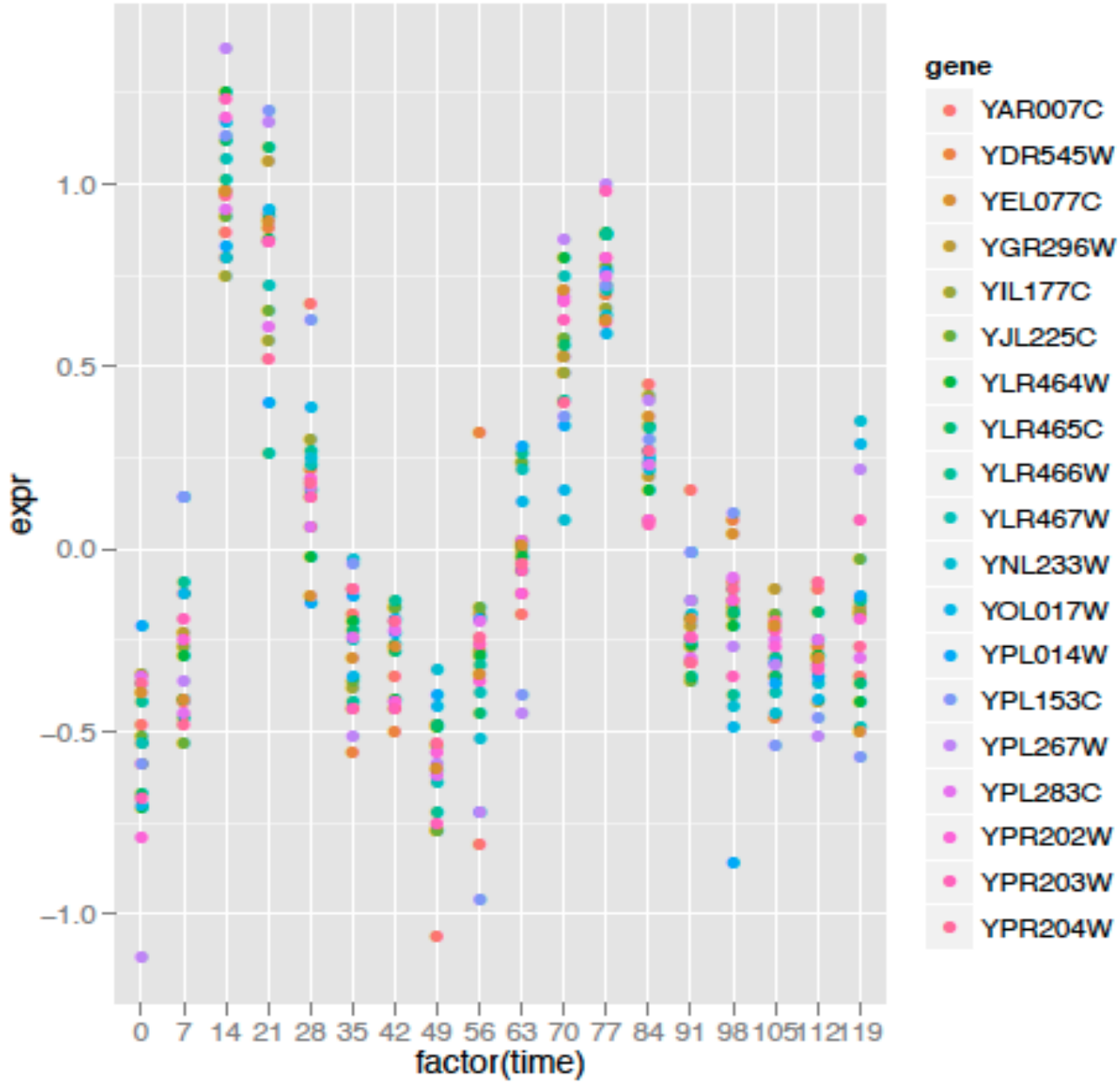
In the brixvis2013 lab (optional)

- Function `clquad()` reexpresses a piece of a cluster analysis of cell-cycle expression trajectories
- Function `clpts()` breaks up a cluster into its constituent trajectories
- Neither is written for general use, but help to illustrate exposure of variability at different scales for a “holistic” workflow step

```
> clquad(c(6, 46, 15, 14))
```



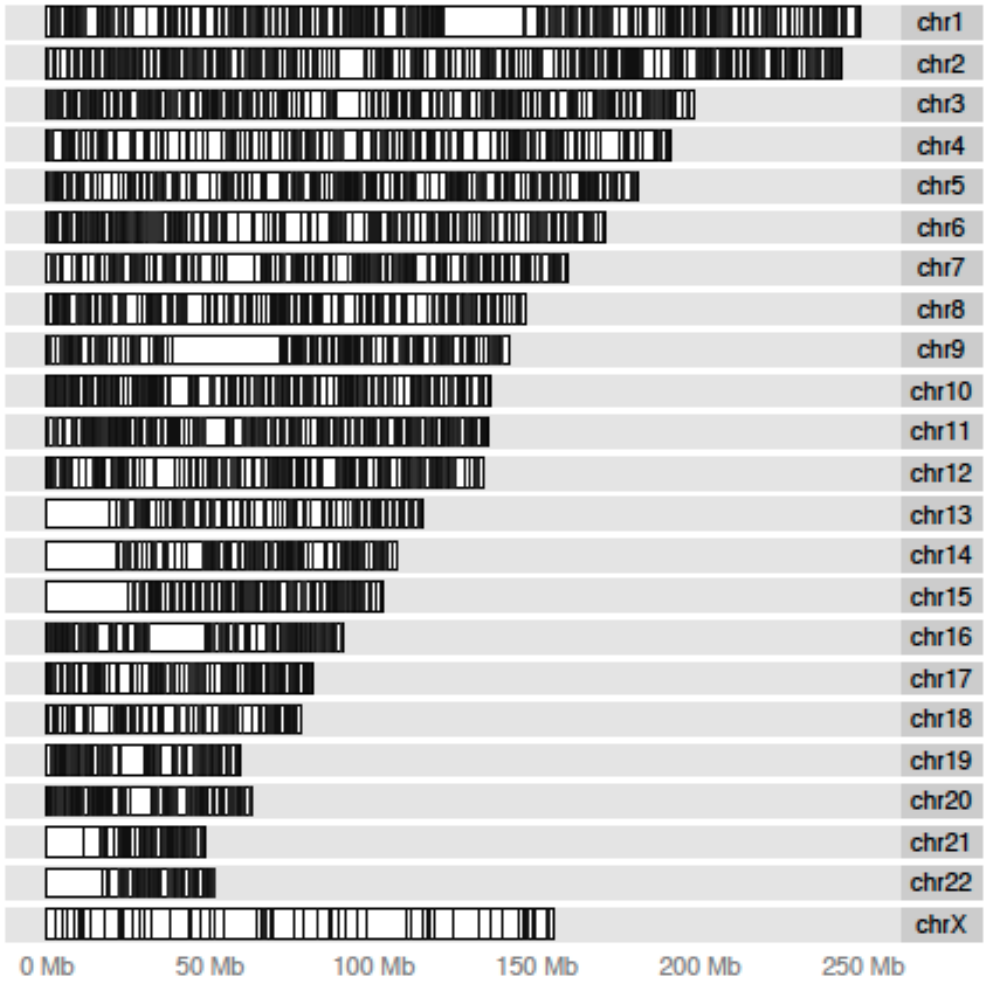
```
> clpts(6)
```


expr
1.0
0.5
0.0
−0.5
−1.0
0 7 14 21 28 35 42 49 56 63 70 77 84 91 98 105 112 119
factor(time)
gene
YAR007C
YDR545W
YEL077C
YGR296W
YIL177C
YJL225C
YLR464W
YLR465C
YLR466W
YLR467W
YNL233W
YOL017W
YPL014W
YPL153C
YPL267W
YPL283C
YPR202W
YPR203W
YPR204W

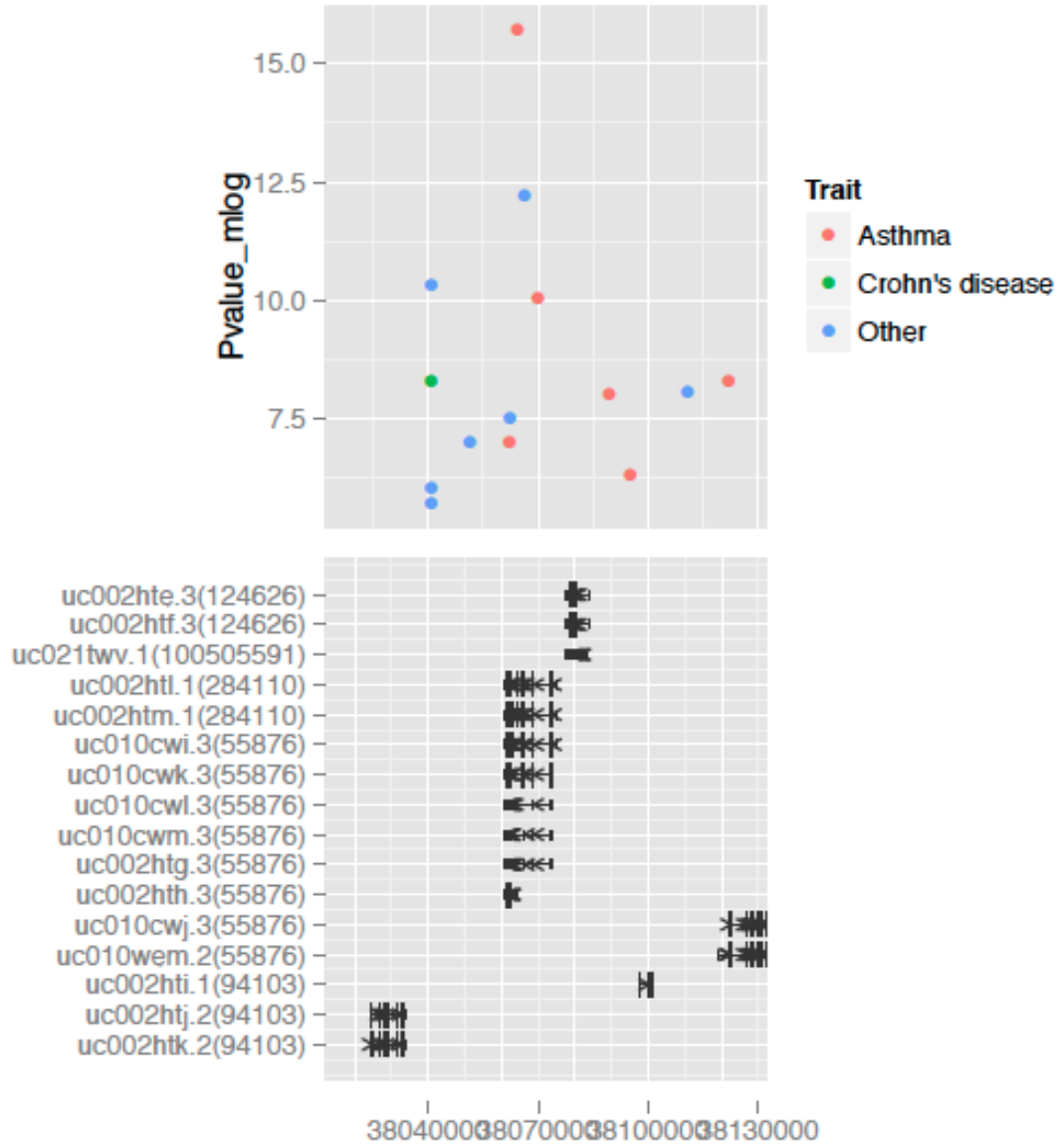
Some examples with *ggbio* package: autoplot method sensitive to input class

```
> showMethods("autoplot")
Function: autoplot (package ggplot2)
object="ANY"
object="BamFile"
object="BSgenome"
object="character"
object="ExpressionSet"
object="GAlignments"
object="GRanges"
object="GRangesList"
object="IRanges"
object="matrix"
object="Rle"
object="RleList"
object="Seqinfo"
object="SummarizedExperiment"
object="TranscriptDb"
object="VCF"
object="Views"
```

```
> library(gwascat)
> library(ggbio)

> gwr = as(gwrngs, "GRanges")
> ap1 = autoplot(gwr, layout = "karyogram")

> ap1
```



```
> txdb = TxDb.Hsapiens.UCSC.hg19.knownGene
> selr3 = GRanges("chr17", IRanges(38022000, width = 1e+05))
> ap4 = autoplot(txdb, which = selr3)
> mp = traitsManh(gwrngs, selr = selr3)

> tracks(mp, ap4)
```



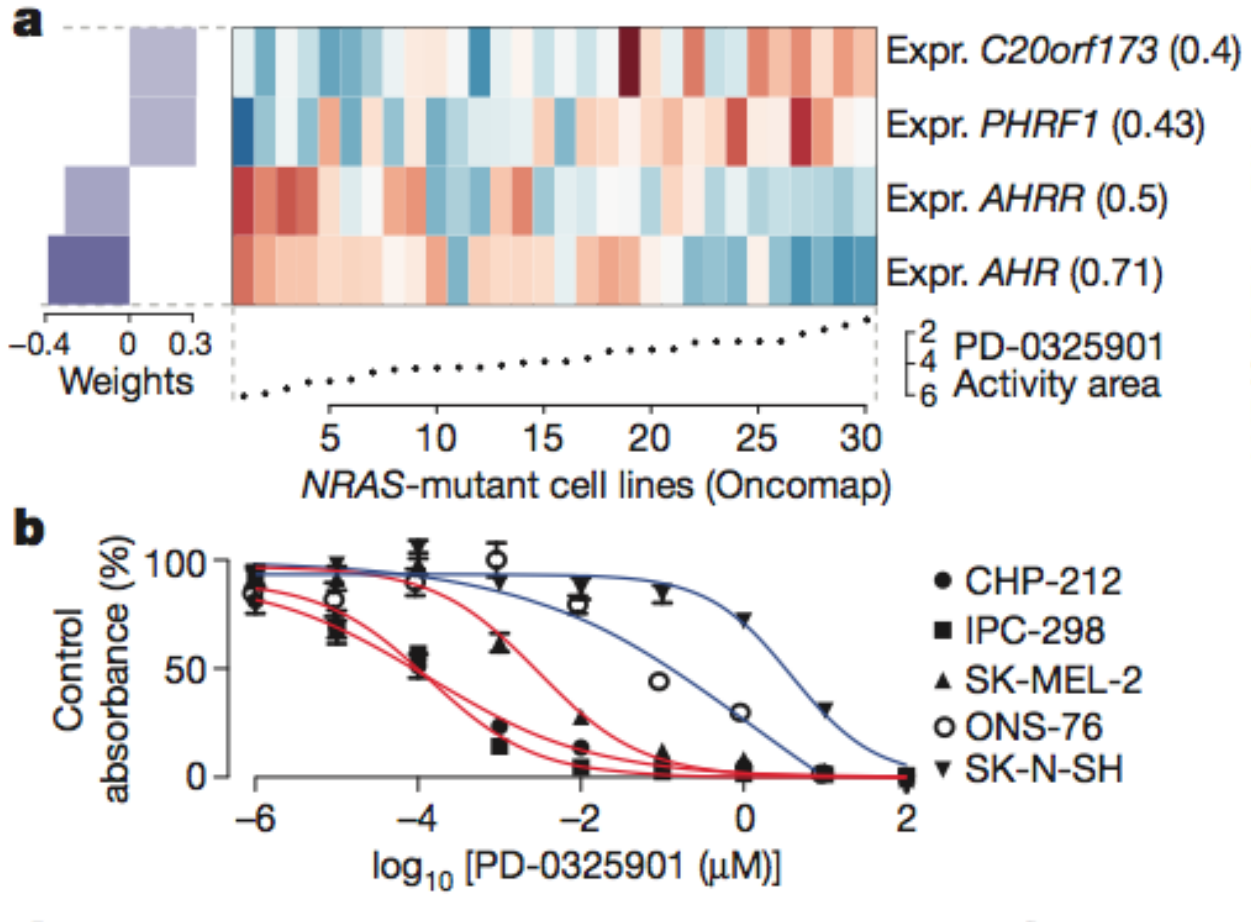
ggbio comments

- Containers for genomic annotation and experimental results have `autoplot` methods
- *ggplot2* factorization of visualization tasks allows programmatically efficient embellishments
- You can go your own way

LETTER

doi:10.1038/nature11003

The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity



Lightweight containers for substantial experimental data

2.4 Pharmacologic profiling data

A special container has been used to manage the profiling data.

```
> data(ccleRx)
> ccleRx

Broad/Novartis Cancer Cell Line Encyclopedia data.
There are  11670  lines/experiments represented.
Use '[', '[[', organ(), compound(), ... to obtain more information.

> ccleRx[[100]]

Cancer Cell Line Encyclopedia experiment data for line GI-1
 organ CENTRAL_NERVOUS_SYSTEM
 compound 17-AAG
 target HSP90
```

Simple syntax for CCLE surveys

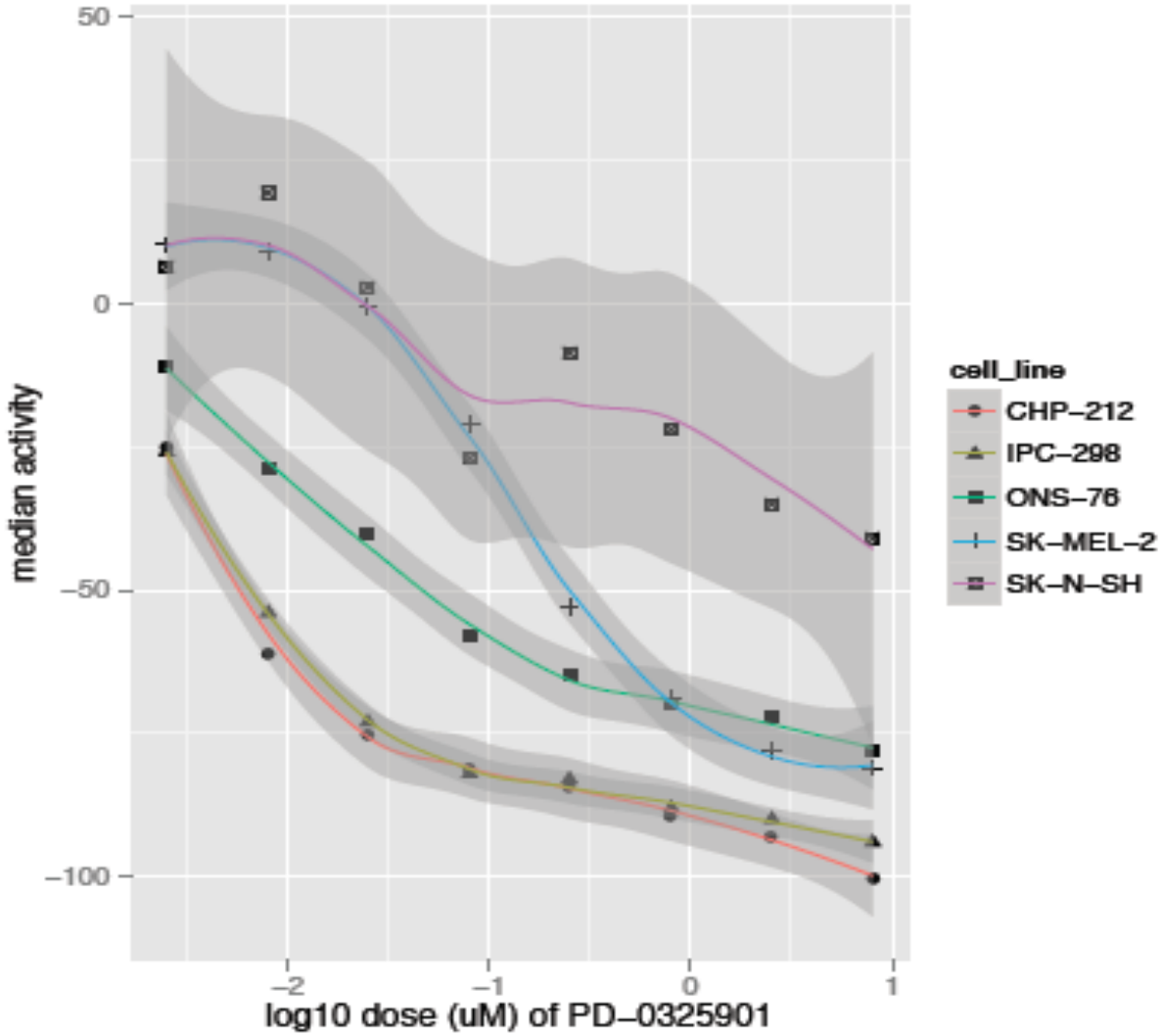
```
> table(organ(ccleRx))[1:10]
```

```
                AUTONOMIC_GANGLIA                             BILIARY_TRACT
                              223                                        24
                             BONE                                    BREAST
                              260                                       701
          CENTRAL_NERVOUS_SYSTEM                                ENDOMETRIUM
                              669                                       458
HAEMATOPOIETIC_AND_LYMPHOID_TISSUE                                   KIDNEY
                             1677                                       209
                LARGE_INTESTINE                                      LIVER
                              535                                       434
```

```
> table(compound(ccleRx))[1:10]
```

```
    17-AAG      AEW541    AZD0530    AZD6244  Erlotinib Irinotecan   L-685458
       503         503         504         503         503         317         491
 Lapatinib     LBW242  Nilotinib
       504         503         420
```

```
> fig3braw = ccleRx[which(compound(ccleRx) == "PD-0325901" & line(ccleRx) %in%
+     c("CHP-212", "IPC-298", "SK-MEL-2", "ONS-76", "SK-N-SH"))]
> plot(fig3braw)
```



Genetics of topotecan sensitivity

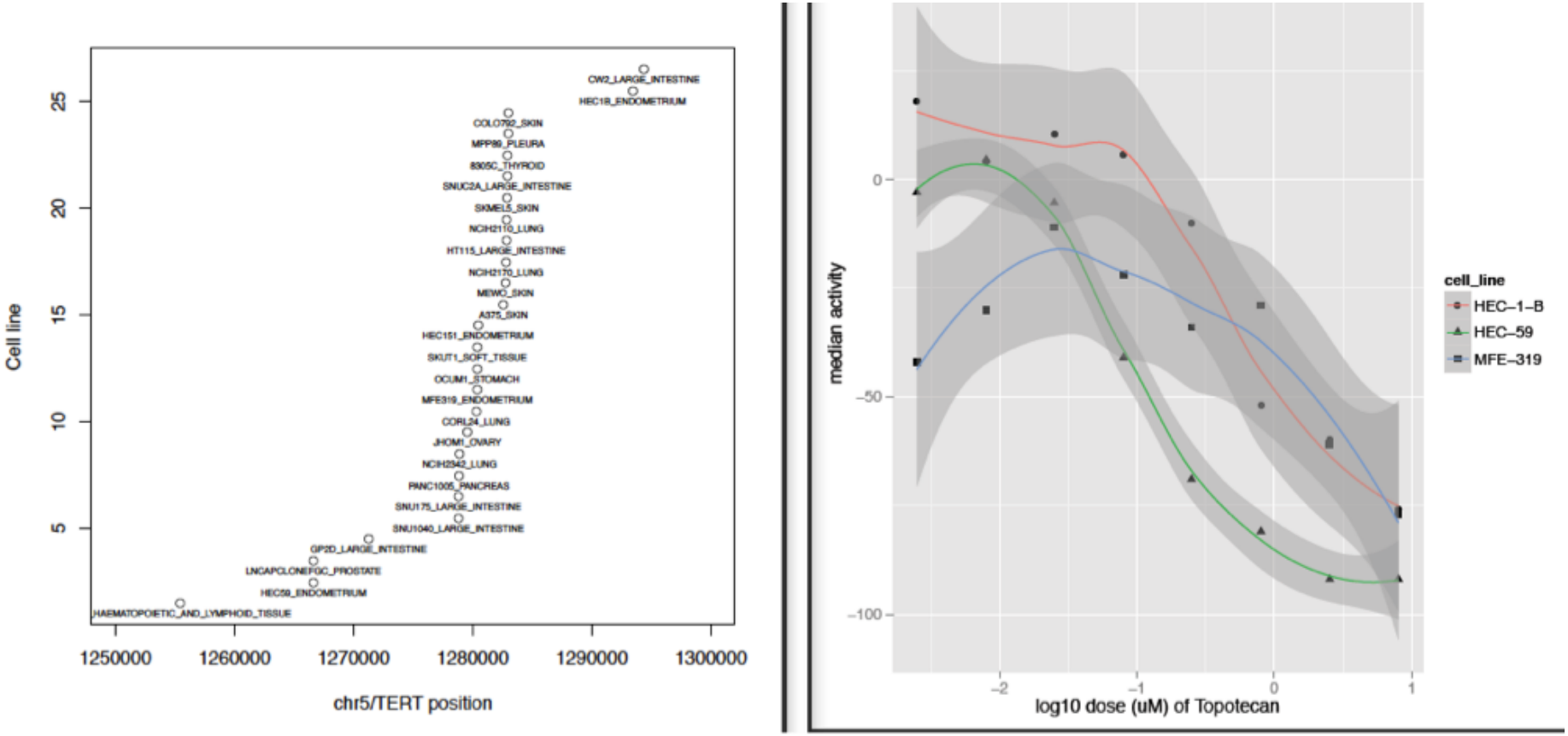


Figure 3. Two views of information on coding variations in TERT in tumor cell line data distributed at Broad/Novartis CCLE.

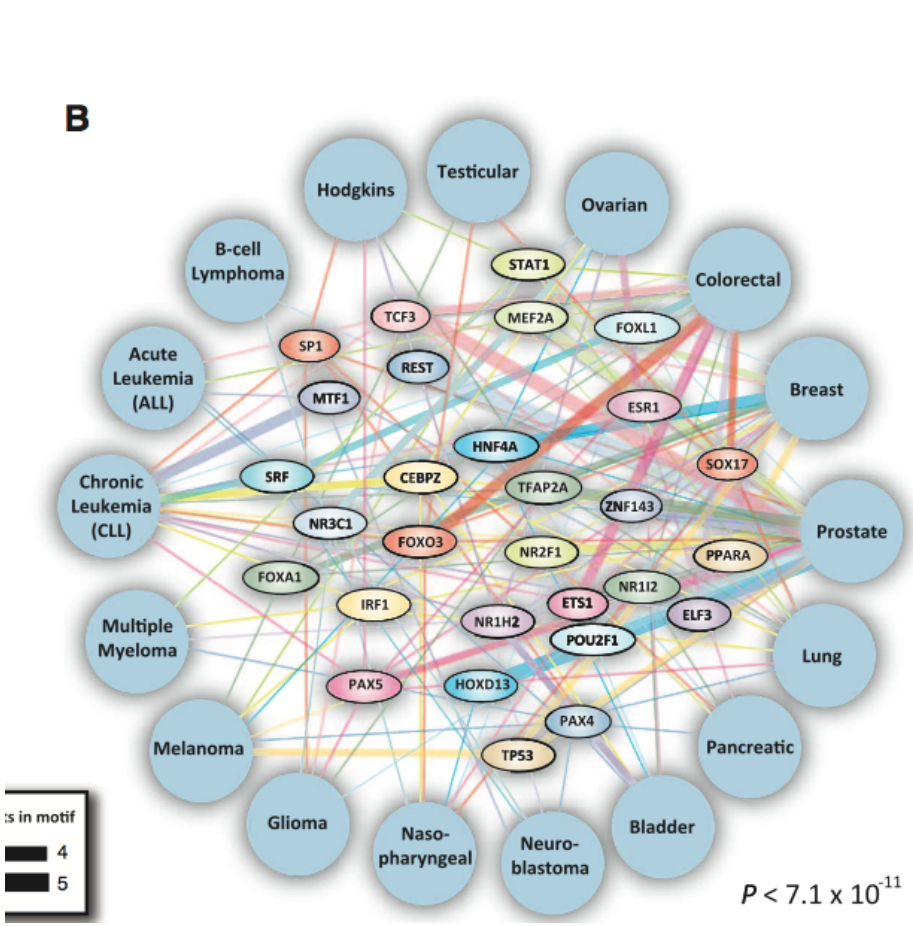
Summary of CCLE visualization support

- Regularities in data structure across cell lines identified for retention in hierarchical object structure
- `plot()` methods defined for inputs at different levels of the hierarchy
- Use of *ggplot2* infrastructure permits immediate use of tunable statistical visualization patterns, factorization of embellishments

Two levels of encyclopedia structure

```
> getClass("ccleSet")
Class "ccleSet" [package "ccleWrap"]

Slots:

Name:        expts dateCreated     csvname csvhash.md5
Class:        list   character   character   character
> getClass("ccleExpt")
Class "ccleExpt" [package "ccleWrap"]

Slots:

Name:            line          organ       compound         target
Class:      character      character      character      character

Name:        doses_uM activityMedian     activitySD        fitType
Class:        numeric        numeric        numeric      character

Name:         EC50_uM        IC50_uM           Amax        ActArea
Class:        numeric        numeric        numeric        numeric
```

Envoi: Variant-TF-Phenotype network structures



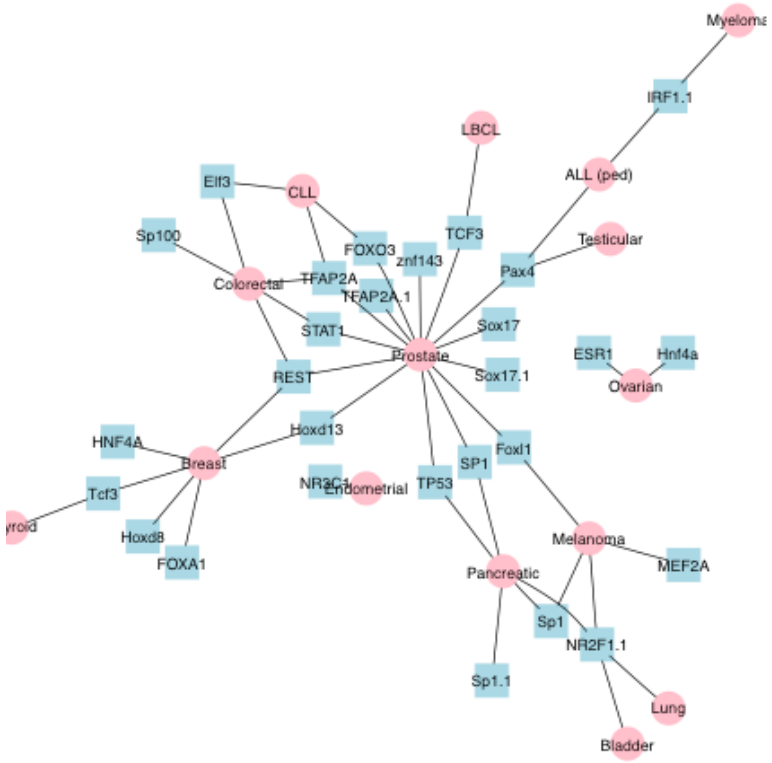


Figure 3. Variant-TF-Phenotype network after Maurano et al. (2012), generated using NHGRI GWAS hits and Bioconductor MotifDb models with FIMO.

Conclusions

- R/bioconductor provide substantial infrastructure for flexible approaches to visualization
- Emphasis: statistical integrity, reproducibility, acknowledgment of variability, uncertainty
- Productive approach: separately conceptualize the underlying data structure, visualization objectives, and code to render the *information* in a tunable, extensible way