Data visualizations should be more interactive

Karl W Broman

Biostatistics & Medical Informatics, University of Wisconsin–Madison

Introduction

- I've found high-dimensional data befuddling.
- With 3000 gene expression arrays, you'd think we'd make a lot of graphs, but we tend to make no graphs.
- Interactive graphics provide a solution to this problem.
- I've come to the conclusions that
  - Data visualization is often more important than formal inference
  - All graphs could be improved with some interactivity

Demos: **`bit.ly/enar2014`**

Opportunities

- **Exploration**
  - **Tuning parameters**
  - **Identifying outliers**
  - **One fancy plot vs 1000 static plots**
- **Reports for collaborators**
  - **Living documents!**
  - **Allow deeper exploration of the results**
  - **Cut down on simple questions?**
- **Big Data**
  - **Don't just rely on summary statistics**
  - **Greatly compressed information, but with access to the details**
  - **Zoom into dense figures**
  - **More exploration, more connections**
- **Teaching**
  - **Cool things to look at and play with**
  - **Animated illustrations of key concepts**
  - **Demonstrate data exploration**
  - **Enable intro students to explore data**

Barriers

- We never learned how
- It's a hassle
- No consistent platform
- Journal articles are static
  (and what else matters?)
- Most statisticians are still creating terrible static plots
  (even worse, obnoxious tables)

But: many exciting new tools

- HTML5 + Scalable vector graphics (SVG)
- Incredible power of modern web browsers
- Javascript-based web tools
- RStudio's tools

D3

- Javascript library for manipulating HTML and SVG elements
- Connects data to elements
- Low level, but flexible

Other options

- **`locator()`** and **`identify()`**
- **ggobi** (`ggobi.org`) and **cranvas** (`github.com/ggobi/cranvas/wiki`)
- **Mondrian** (`rosuda.org/software/Mondrian`)
- **Acinoynx** (aka iPlots eXtreme) (`rforge.net/Acinonyx`)
- **googleVis** (`code.google.com/p/google-motion-charts-with-r`)
- **Shiny** (`www.rstudio.com/shiny`)
- **ggvis** (`ggvis.rstudio.com`)
- **Rcharts** (`rcharts.io`)

simple ⟷ **flexible**

Choose one.
I choose **flexible**.

Summary

- For high-dimensional data, good visualizations are critical
- Interactive graphics require effort, but they
  - Facilitate exploration
  - Are great collaborative tools
  - Enable summaries with access to the details
- Visualizations must be tailored to the data and questions
- D3 is rather low level, but it
  - Is totally flexible (like R's static graphics)
  - Provides hours of enjoyment
  - Can provide other hours of frustration
- R/qtlcharts package under development

Acknowledgments

Example 1

Alan Attie[1], Mark Keller[1], Aimee Teo Broman[2], Christina Kendziorski[2], Brian Yandell[3], Eric Schadt[4]; Departments of [1]Biochemistry, [2]Biostatistics & Medical Informatics, and [3]Statistics, UW-Madison; [4]Mount Sinai

Example 2

Candace Moore[1], Edgar Spalding[1], Logan Johnson[1], Il-Youp Kwak[2], Miron Livny[3]; Departments of [1]Botany, [2]Statistics, and [3]Computer Sciences, UW-Madison

Contact

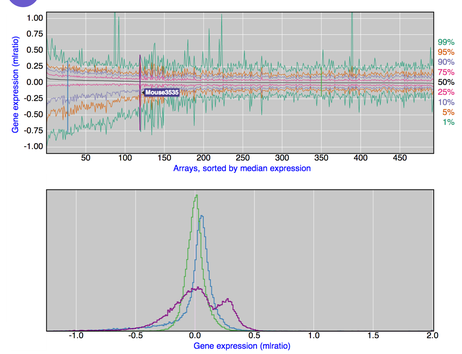
Karl Broman
`kbroman@biostat.wisc.edu`
`@kwbroman`
`www.biostat.wisc.edu/~kbroman`
`github.com/kbroman`

This work was supported in part by NIH grant GM074244.

Example 1: Expression genetics

- Mouse intercross, B6 × BTBR
- ~500 mice
- Genotypes at 2057 SNPs
- Gene expression microarrays in six tissues
- Numerous clinical phenotypes

1a

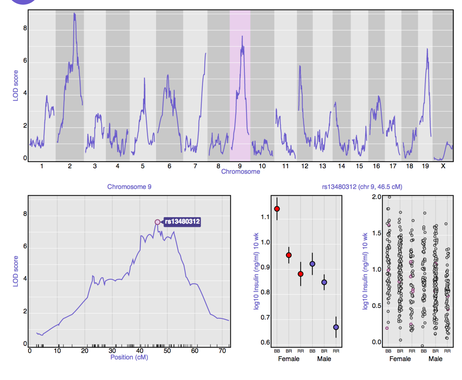


These are data from ~500 gene expression microarrays.

The top panel is like 500 box plots: lines are drawn at the 1, 5, . . . , 99 percentiles for each of ~500 distributions. The distibutions are sorted by their medians.

If one hovers over a column in the top panel, the corresponding distribution is shown below. Click in the top panel for the distribution to persist, and click again to make it go away.

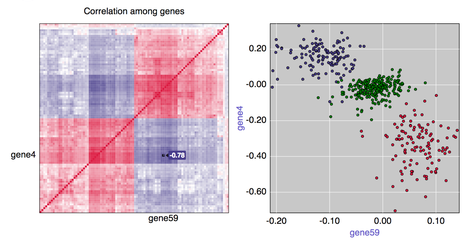
1b



A genome scan for genetic loci (called quantitative trait loci, QTL) influencing insulin level. The LOD score is a $\log_{10}$ likelihood ratio measuring the strength of association between genotype and phenotype.

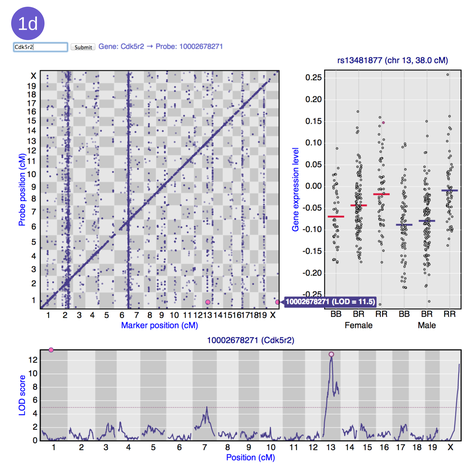
Click on a chromosome at the top and a detailed view of the LOD curve for that chromosome is shown on the bottom left. In the lower-left panel, hover over markers to see names; click to view an effect plot and and phenotype-vs-genotype plot to the right.

1c



Association in gene expression among 100 genes that are influenced by a common genetic locus (QTL). The left panel is a heat map of the correlation matrix, with blue = –1 and red = +1. Hover over pixels in the correlation matrix on the left to see the values; click to see the corresponding scatterplot on the right. Points in the scatterplot are colored by genotype at the underlying QTL.

1d



An investigation of genetic loci (eQTL) influencing gene expression. In the top-left panel, the x-axis corresponds to marker location and the y-axis corresponds to the position of probes on a gene expression microarray. Each plotted point is an inferred eQTL.

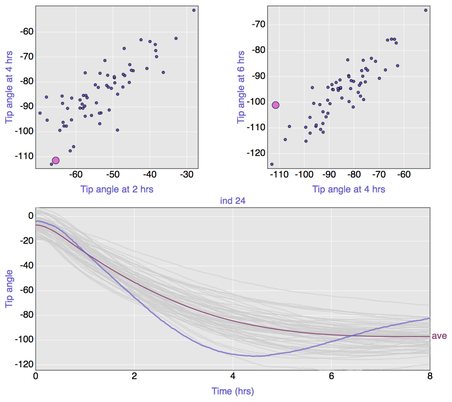
Hover over a point to see the probe ID and LOD score (measuring the strength of association); also highlighted are any other eQTL for that probe. Click on the point to see the LOD curves below.

Hover over markers in the LOD curve plot to view marker names; click on a marker to see the phenotype-vs-genotype plot to the right.

Example 2: Gravitropism

- Response to gravity in Arabidopsis seedlings
- Rotate orientation of gravity and video over 8 hrs
- Measure the angle of the root tip every 2 min

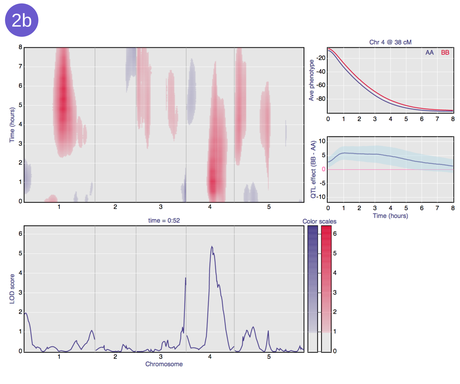
2a



Average tip angle over time for 162 Arabidopsis lines.

Hover over points in the top panels or curves in the bottom panel to highlight the corresponding line in the other panels.

2b



The top-left panel is a heat map of a measure association (LOD score) between genotype at a fixed position and the phenotype at a fixed time. Red (blue) indicates that BB (AA) lines have larger phenotype.

When you hover over a point in the top-left plot, the LOD curves for the corresponding time are shown below, and the phenotype averages and estimated genetic effect (across time) are shown to the right.