Presentations and posters

Tools for Reproducible Research

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

biostat.wisc.edu/~kbroman
github.com/kbroman
@kwbroman

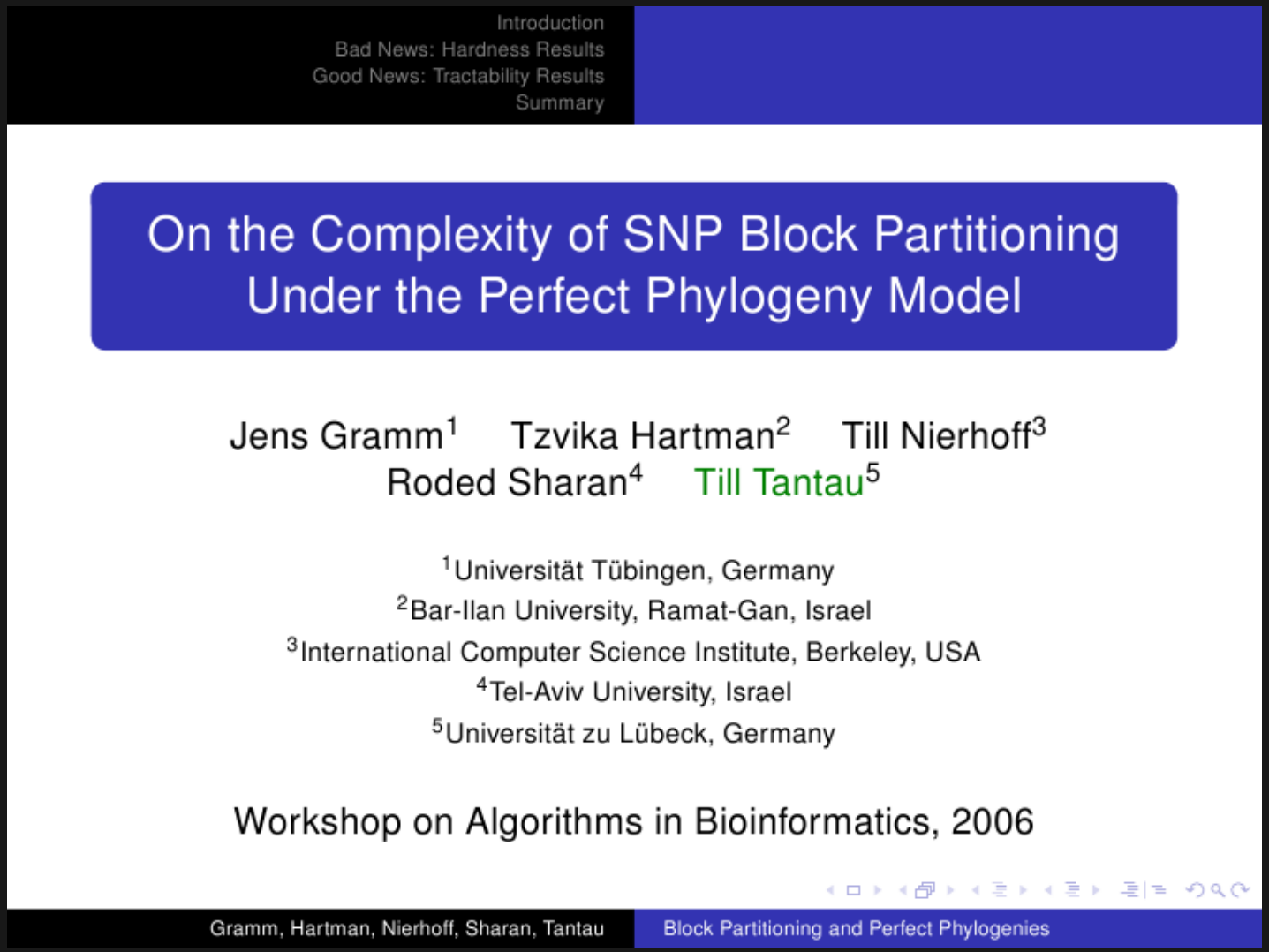
Course web: bit.ly/tools4rr

Powerpoint/Keynote

+ Standard
+ Easy to share slides
+ WYSIWYG (mostly)
+ Fancy animations

– Font problems
– Lots of copy-paste
– Hard to get equations
– Not reproducible

LaTeX Beamer package



Get rid of the junk

```
\usetheme{default}
\beamertemplatenavigationsymbolsempty

\definecolor{foreground}{RGB}{255,255,255}
\definecolor{background}{RGB}{24,24,24}
\definecolor{title}{RGB}{107,174,214}
\definecolor{subtitle}{RGB}{102,255,204}
\definecolor{hilit}{RGB}{102,255,204}
\definecolor{lolit}{RGB}{155,155,155}

\setbeamercolor{titlelike}{fg=title}
\setbeamercolor{subtitle}{fg=subtitle}
\setbeamercolor{institute}{fg=lolit}
\setbeamercolor{normal text}{fg=foreground,bg=background}
\setbeamercolor{item}{fg=foreground} % color of bullets
\setbeamercolor{subitem}{fg=lolit}
\setbeamercolor{itemize/enumerate subbody}{fg=lolit}
\setbeamertemplate{itemize subitem}{{\textendash}}
\setbeamerfont{itemize/enumerate subbody}{size=\footnotesize}
\setbeamerfont{itemize/enumerate subitem}{size=\footnotesize}

\newcommand{\hilit}{\color{hilit}}
\newcommand{\lolit}{\color{lolit}}
```

Also, slide numbers and fonts

```
% slide number
\setbeamertemplate{footline}{%
 \raisebox{5pt}{\makebox[\paperwidth]{\hfill\makebox[20pt]{\lolit
  \scriptsize\insertframenumber}}}\hspace*{5pt}}

% font
\usepackage{fontspec}
% http://www.gust.org.pl/projects/e-foundry/tex-gyre/
%      ...   heros/qhv2.004otf.zip
\setsansfont
  [ ExternalLocation = ../fonts/ ,
    UprightFont = *-regular ,
    BoldFont = *-bold ,
    ItalicFont = *-italic ,
    BoldItalicFont = *-bolditalic ]{texgyreheros}
% Palatino for notes
\setbeamerfont{note page}{family*=pplx,size=\footnotesize}
```

Title slide

```
\title{Put title here}
\subtitle{And maybe a subtitle}
\author{Author name}
\institute{Biostatistics \& Medical Informatics,
   UW{\textendash}Madison}
\date{\tt \scriptsize biostat.wisc.edu/{\textasciitilde}kbroman}

\begin{document}

{
\setbeamertemplate{footline}{} % no slide number here
\frame{
  \titlepage

\note{
  Summary of the talk, as a note.
}
} }
```

Typical slide

```
\begin{frame}{Title of slide}

\bbi
 \item Bullet 1
 \item Bullet 2
 \item Bullet 3
\ei

\note{
  Put a note here
}
\end{frame}
```

Typical slide

```
\begin{frame}{Title of slide}

\vspace{24pt} \begin{itemize} \itemsep8pt
 \item Bullet 1
 \item Bullet 2
 \item Bullet 3
\end{itemize}

\note{
  Put a note here
}
\end{frame}
```

Slide with a figure

```
\begin{frame}{Title of slide}

\figh{Figs/a_figure.png}{0.75}


\note{
  Put a note here
}
\end{frame}
```

Slide with a figure

```
\begin{frame}{Title of slide}

\centerline{\includegraphics[height=0.75\textheight]{%
            Figs/a_figure.png}}

\note{
  Put a note here
}
\end{frame}
```

Figures with KnitR

```
<<knitr_options, echo=FALSE>>=
opts_chunk$set(echo=FALSE, fig.height=7, fig.width=10)
change_colors <-
function(bg=rgb(24,24,24, maxColorValue=255), fg="white")
  par(bg=bg, fg=fg, col=fg, col.axis=fg, col.lab=fg,
      col.main=fg, col.sub=fg)
@

<<pdf_figure>>=
change_colors()
par(las=1)
n <- 100
x <- rnorm(n)
y <- 2*x + rnorm(n)
plot(x, y, pch=16, col="slateblue")
@
```

Figures with KnitR

```
% << >>= all on one line!
<<png_figure, dev="png", fig.align="center",
  dev.args=list(pointsize=30),
  fig.height=15, fig.width=15, out.height="0.75\\textheight",
  out.width="0.75\\textheight">>=
change_colors(bg=rgb(32,32,32,maxColorValue=255))
par(las=1)
n <- 251
x <- y <- seq(-pi, pi, len=n)
z <- matrix(ncol=n, nrow=n)
for(i in seq(along=x))
  for(j in seq(along=y))
    z[i,j] <- sin(x[i]) + cos(y[j])
image(x,y,z)
@
```

Slides with notes

```
\documentclass[12pt,t]{beamer}
\setbeameroption{hide notes}
\setbeamertemplate{note page}[plain]
```

```
\documentclass[12pt,t,handout]{beamer}
\setbeameroption{show notes}
\setbeamertemplate{note page}[plain]
\def\notescolors{1}
```

```
\ifx\notescolors\undefined % slides
  \definecolor{foreground}{RGB}{255,255,255}
  \definecolor{background}{RGB}{24,24,24}
\else % notes
  \definecolor{background}{RGB}{255,255,255}
  \definecolor{foreground}{RGB}{24,24,24}
\fi
```

Simple animations

```
\begin{frame}{Bullets entering one at a time}

\bbi
\item Bullet 1
\onslide<2->{\item Bullet 2}
\onslide<3->{\item Bullet 3}
\onslide<4->{\item Bullet 4}
\ei

\note{
  Do this sparingly.
}
\end{frame}
```

Simple animations

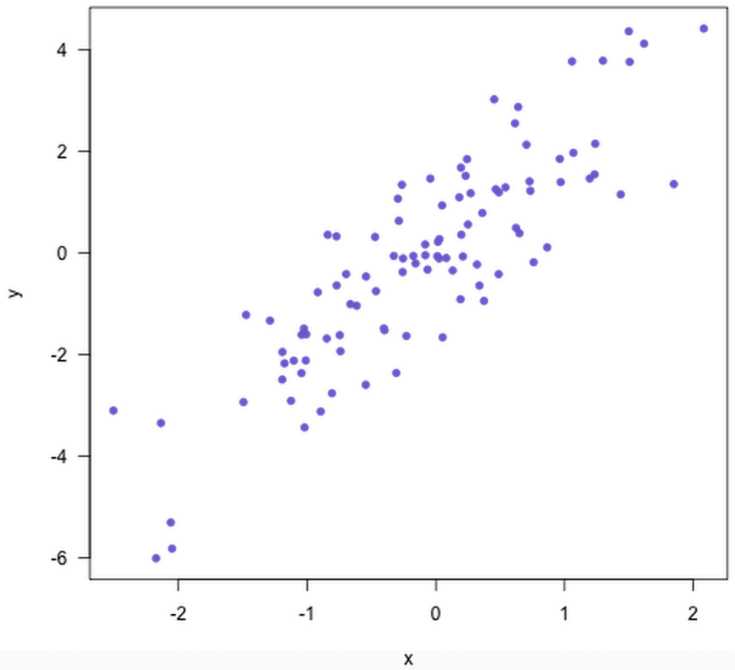
```
\begin{frame}{Bullets entering one at a time}

\bbi
\item {\lolit \only<1>{\color{foreground}} Bullet 1}
\item {\lolit \only<2>{\color{foreground}} Bullet 2}
\item {\lolit \only<3>{\color{foreground}} Bullet 3}
\item {\lolit \only<4>{\color{foreground}} Bullet 4}
\ei

\note{
  Do this sparingly.
}
\end{frame}
```

Slidify and R Markdown

A figure



5/7

Slidify and R Markdown

````
## Slide title

- Bullet 1
- Bullet 2
- Bullet 3
- Bullet 4

---

## A figure

```{r a_figure, echo=FALSE, fig.align="center"}
par(las=1)
n <- 100
x <- rnorm(n)
y <- 2*x + rnorm(n)
plot(x, y, pch=16, col="slateblue")
```
````

Using slidify

```
library(devtools)
install_github("slidify", "ramnathv")
install_github("slidifyLibraries", "ramnathv")

library(slidify)
author("slidify_example")

# edit slidify_example/index.Rmd

slidify("index.Rmd")
```

YAML header

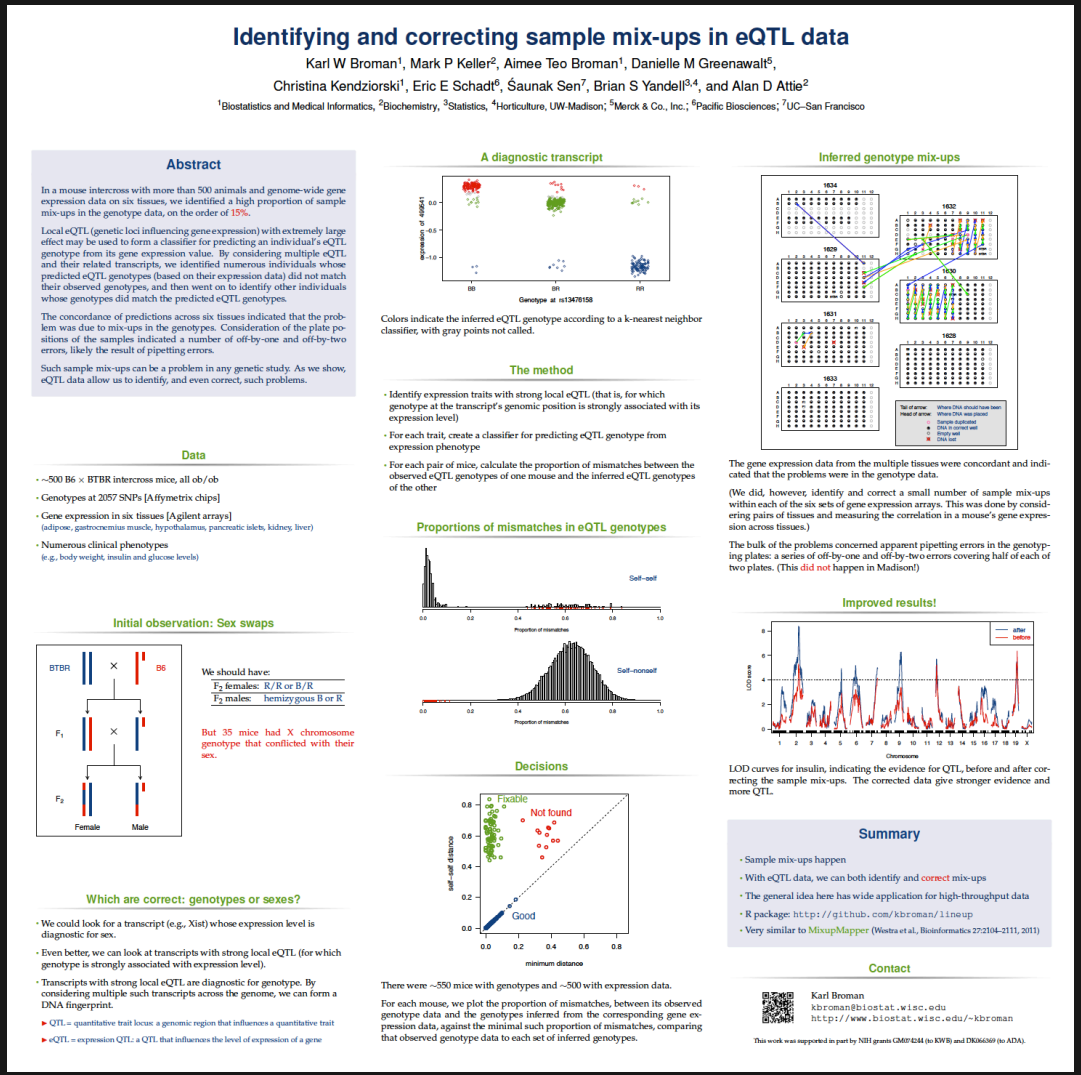
```
---
title        : Slidify example
subtitle     : Tools for reproducible research
author       : Karl Broman
job          : Biostatistics & Medical Informatics, UW-Madison
framework    : io2012        # {io2012, html5slides, shower, ...}
highlighter  : highlight.js  # {highlight.js, prettify, highlight}
hitheme      : tomorrow      #
widgets      : [mathjax]     # {mathjax, quiz, bootstrap}
mode         : standalone    # {selfcontained, standalone, draft}
---
```

Change the title slide colors

```
<style>
.title-slide {
  background-color: #EEE;
}

.title-slide hgroup > h1,
.title-slide hgroup > h2 {
  color: #005;
}
</style>
```

Beamer-based posters



Identifying and correcting sample mix-ups in eQTL data

Karl W Broman[1], Mark P Keller[2], Aimee Teo Broman[1], Danielle M Greenawalt[5], Christina Kendziorski[1], Eric E Schadt[6], Śaunak Sen[7], Brian S Yandell[3,4], and Alan D Attie[2]

[1]Biostatistics and Medical Informatics, [2]Biochemistry, [3]Statistics, [4]Horticulture, UW-Madison; [5]Merck & Co., Inc.; [6]Pacific Biosciences; [7]UC–San Francisco

Abstract

In a mouse intercross with more than 500 animals and genome-wide gene expression data on six tissues, we identified a high proportion of sample mix-ups in the genotype data, on the order of 15%.

Local eQTL (genetic loci influencing gene expression) with extremely large effect may be used to form a classifier for predicting an individual's eQTL genotype from its gene expression value. By considering multiple eQTL and their related transcripts, we identified numerous individuals whose predicted eQTL genotypes (based on their expression data) did not match their observed genotypes, and then went on to identify other individuals whose genotypes did match the predicted eQTL genotypes.

The concordance of predictions across six tissues indicated that the problem was due to mix-ups in the genotypes. Consideration of the plate positions of the samples indicated a number of off-by-one and off-by-two errors, likely the result of pipetting errors.

Such sample mix-ups can be a problem in any genetic study. As we show, eQTL data allow us to identify, and even correct, such problems.

Data

- ~500 B6 × BTBR intercross mice, all ob/ob
- Genotypes at 2057 SNPs [Affymetrix chips]
- Gene expression in six tissues [Agilent arrays] (adipose, gastrocnemius muscle, hypothalamus, pancreatic islets, kidney, liver)
- Numerous clinical phenotypes (e.g., body weight, insulin and glucose levels)

Initial observation: Sex swaps

We should have:

F_2 females:	R/R or B/R
F_2 males:	hemizygous B or R

But 35 mice had X chromosome genotype that conflicted with their sex.

Which are correct: genotypes or sexes?

- We could look for a transcript (e.g., Xist) whose expression level is diagnostic for sex.
- Even better, we can look at transcripts with strong local eQTL (for which genotype is strongly associated with expression level).
- Transcripts with strong local eQTL are diagnostic for genotype. By considering multiple such transcripts across the genome, we can form a DNA fingerprint.

► QTL = quantitative trait locus: a genomic region that influences a quantitative trait

► eQTL = expression QTL: a QTL that influences the level of expression of a gene

A diagnostic transcript

Colors indicate the inferred eQTL genotype according to a k-nearest neighbor classifier, with gray points not called.

The method

- Identify expression traits with strong local eQTL (that is, for which genotype at the transcript's genomic position is strongly associated with its expression level)
- For each trait, create a classifier for predicting eQTL genotype from expression phenotype
- For each pair of mice, calculate the proportion of mismatches between the observed eQTL genotypes of one mouse and the inferred eQTL genotypes of the other

Proportions of mismatches in eQTL genotypes

Decisions

There were ~550 mice with genotypes and ~500 with expression data.

For each mouse, we plot the proportion of mismatches, between its observed genotype data and the genotypes inferred from the corresponding gene expression data, against the minimal such proportion of mismatches, comparing that observed genotype data to each set of inferred genotypes.

Inferred genotype mix-ups

The gene expression data from the multiple tissues were concordant and indicated that the problems were in the genotype data.

(We did, however, identify and correct a small number of sample mix-ups within each of the six sets of gene expression arrays. This was done by considering pairs of tissues and measuring the correlation in a mouse's gene expression across tissues.)

The bulk of the problems concerned apparent pipetting errors in the genotyping plates: a series of off-by-one and off-by-two errors covering half of each of two plates. (This did not happen in Madison!)

Improved results!

LOD curves for insulin, indicating the evidence for QTL, before and after correcting the sample mix-ups. The corrected data give stronger evidence and more QTL.

Summary

- Sample mix-ups happen
- With eQTL data, we can both identify and correct mix-ups
- The general idea here has wide application for high-throughput data
- R package: `http://github.com/kbroman/lineup`
- Very similar to MixupMapper (Westra et al., Bioinformatics 27:2104–2111, 2011)

Contact

Karl Broman
kbroman@biostat.wisc.edu
http://www.biostat.wisc.edu/~kbroman

This work was supported in part by NIH grants GM074244 (to KWB) and DK066369 (to ADA).

github.com/kbroman/Poster_SampleMixups

Beamer-based posters

Data visualizations should be more interactive

Karl W Broman
Biostatistics & Medical Informatics, University of Wisconsin–Madison

Introduction

- High-dimensional data can be befuddling.
- With 3000 gene expression arrays, you'd think we'd make a lot of graphs, but we tend to make no graphs: We can't look at 3000 histograms, so why look at any?
- 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** (`ratudio.com/shiny`)
- **ggvis** (`ggvis.ratudio.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 (`github.com/kbroman/qtlcharts`)

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

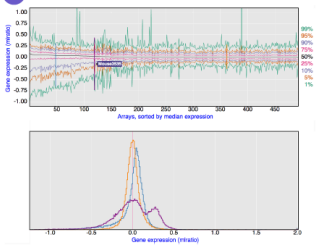
Karl Broman
kbroman@biostat.wisc.edu
@kwbroman
www.biostat.wisc.edu/~kbroman
github.com/kbroman

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

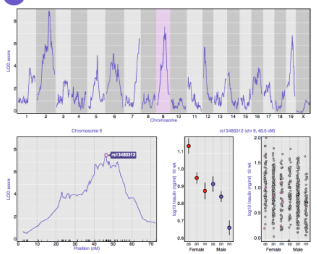


These are data from ~500 gene expression microarrays.

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.

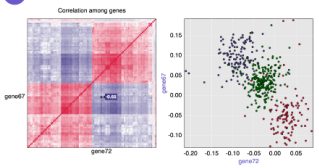
1b



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.

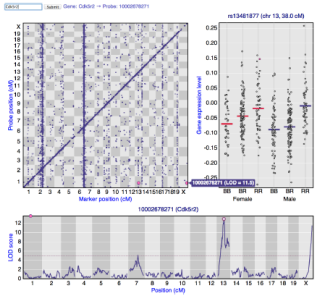
Click on a chromosome at the top and a detailed view of the LOD curve for that chromosome is shown on the bottom left. In the lower-left panel, hover over markers to see names; click to view an effect plot and and phenotype-vs-genotype plot to the right.

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.

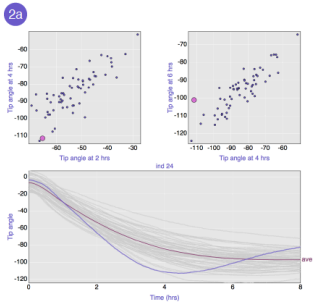
Hover over a point to see the probe ID and LOD score (measuring the strength of association); also highlighted are any other eQTL for that probe. Click on the point to see the LOD curves below.

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

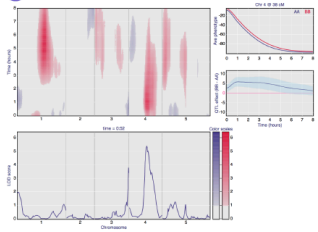
2a



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 of 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.

`github.com/kbroman/Poster_ENAR2014`

Beamer-based posters

```
\documentclass[final,plain]{beamer}
\usepackage[size=custom,width=152.4,height=91.44,scale=1.2]{%
    beamerposter}

\newlength{\sepwid}
\newlength{\onecolwid}
\newlength{\halfcolwid}
\newlength{\twocolwid}
\newlength{\threecolwid}

\setlength{\sepwid}{0.0192\paperwidth}
\setlength{\onecolwid}{0.176\paperwidth}
\setlength{\halfcolwid}{0.0784\paperwidth}
\setlength{\twocolwid}{0.3712\paperwidth}
\setlength{\threecolwid}{0.5664\paperwidth}
\setlength{\topmargin}{-0.5in}
\usetheme{confposter}
```

Basic code for a poster

```
\title{Data visualizations should be more interactive}
\author{Karl W Broman}
\institute{University of Wisconsin--Madison}

\begin{frame}[t]
\begin{columns}[t]
  \begin{column}{\sepwid}\end{column} % empty spacer column
  \begin{column}{\onecolwid}
    \begin{exampleblock}{\Large Introduction}{
      \begin{itemize} \itemsep18pt
        \item Bullet 1
        \item Bullet 2
      \end{itemize}
    }
  \colonevsep % between blocks
    \begin{block}{Barriers}{
    }
  \end{column}
\end{columns}
\end{frame}
```

Between-block spacing

```
\newcommand{\colonevsep}{\vspace{16mm}}
\newcommand{\coltwovsep}{\vspace{35.5mm}}
\newcommand{\colthreevsep}{\vspace{14mm}}
\newcommand{\colfourvsep}{\vspace{16mm}}
\newcommand{\colfivevsep}{\vspace{23mm}}
```

Summary

- Use LaTeX/Beamer or Slidify to create reproducible slides.
- Use LaTeX/Beamer to create reproducible posters.
- Include KnitR code chunks to create figures directly.
- Or keep the code for figures separate.