

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

On the Complexity of SNP Block Partitioning Under the Perfect Phylogeny Model

Jens Gramm¹ Tzvika Hartman² Till Nierhoff³
Roded Sharan⁴ **Till Tantau⁵**

¹Universität Tübingen, Germany

²Bar-Ilan University, Ramat-Gan, Israel

³International Computer Science Institute, Berkeley, USA

⁴Tel-Aviv University, Israel

⁵Universität zu Lübeck, Germany

Workshop on Algorithms in Bioinformatics, 2006

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\insertframenum}}} \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

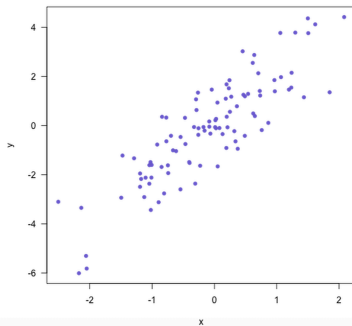
```
\begin{frame}{Bullets entering one at a time}

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

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

Slidify and R Markdown

A figure



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)
setwd("~/Docs/Talks/")
author("slidify_example")

# edit ~/Docs/Talks/slidify_example/index.Rmd

slidify("index.Rmd")
browseURL("index.html")
```

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 Broman¹, Mark P Keller², Ameer Teo Broman¹, Danielle M Greenawald¹,

Christina Kendziora³, Eric E Schadt⁴, Šaunak Sen⁵, Brian S Yandell^{6,7}, and Alan D Attie²

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

Abstract

In a recent interview with more than 500 individuals 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 eQTLs, genetic loci influencing gene expression with extremely large effect may be used to learn a classifier for predicting an individual's eQTL genotype from its gene expression values. By considering multiple eQTLs 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.

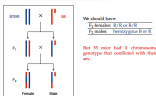
The concordance of predictions across six tissues indicated that the problems were due to mix-ups in the genotypes. Correlations of the gene expressions of the transcripts indicated a number of self-by-one and self-by-two errors, likely the result of pipetting errors.

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

Data

- ~500 lbs. x 100Kb reference data, all 4x4s
- Genotypes at 267 SNPs (Adaptive chips)
- Gene expression in six tissues (Agilent arrays)
- Self, gene expression results: Hypothetical, gene expression, history (text)
- Numerous clinical phenotypes (e.g. body weight, height and glucose levels)

Initial observation: Sex swaps



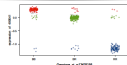
Which are correct: genotypes or sexes?

- 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 eQTLs, for which genotype is strongly associated with expression level.

- Transcripts with strong local eQTLs are diagnostic for the genotype. By considering multiple such transcripts across the genome, we can learn a DNA language.

- eQTL = quantitative trait locus: a genomic region that influences a quantitative trait
- eQTL = expression eQTL, eQTL that influences the level of expression of a gene

A diagnostic transcript



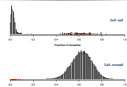
Colors indicate the inferred eQTL genotype according to a nearest neighbor classifier, with gray points not called.

The method

- Identify expression traits with strong local eQTLs (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 mixes, calculate the proportion of mismatches between the observed eQTL genotypes of one mix and the inferred eQTL genotypes of the other.

Proportions of mismatches in eQTL genotypes



Decisions



There were ~500 mixes with genotypes and ~500 with expression data. For each mix, we plot the proportion of mismatches between its observed genotype data and the genotypes inferred from the corresponding gene expression data, against the estimated each proportion of mismatches, comparing that observed genotype data to each of inferred genotypes.

Inferred genotype mix-ups

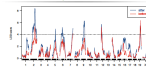


The gene expression data from the multiple tissues were concordant and both called that the problems were in the genotype data.

This 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 strong pairs of tissues and measuring the correlation in a woman's gene expression across tissues.

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

Improved results



LSD curves for (male), indicating the evidence for QTLs, before and after correcting the sample mix-ups. The corrected data give stronger evidence and more QTLs.

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: <https://github.com/kbroman/samplemixups>
- Very similar to [Pharmacia/Bioinformatics](https://github.com/Pharmacia/Bioinformatics)

Contacts



Karl Broman

karl.broman@wisc.edu

<https://www.biostat.wisc.edu/~kbroman/>

This work was supported by grant R01HG006048 to KBW and DMR000000 to ADA.

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.
- You come to the conclusion that...
- This visualization is often more important than formal inference.
- All graphs could be improved with some interactivity.

Deviant: bit.ly/wwa22014

Opportunities

- Exploration
 - Testing parameters
 - Identifying outliers
 - One fancy plot vs 1000 static plots
- Reports for collaborators
 - Living document
 - Allow deeper exploration of the results
 - Can you save sample questions?
- Big Data
 - Don't just rely on summary statistics
 - Compress compressed information, but with access to the details
 - Zoom into dense figures
 - More exploration, more connections
- Teaching
 - Can things be look at all play with
 - Animated dimensions of key concepts
 - Demonstrate data exploration
 - Enable interactive exploration of data

Barriers

- We never learned how
- It's a hassle
- No consistent platforms
- Journal articles are static (and obscure networks)
- Most statisticians are still creating terrible static plots (great work, otherwise telling)

But...many exciting new tools

- HTML5 + Scalable vector graphics (SVG)
- Unbeatable power of modern web browsers
- JavaScript-based web tools
- Reusable tools

D3

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

Other options

- Interact (1) and interact.js (2)
- ggplot (ggplot2, d3.js and canvas) (3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100)
- NodeInter (14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100)
- Acinquiry (14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100)
- ggplotViz (14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100)
- Shiny (14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100)
- ggplot (14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100)
- Recharts (14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100)

simple ↔ flexible

Choose one, I choose flexible.

Summary

- For high-dimensional data, good visualizations are critical.
- Interactive graphics, responsive, beautiful
- Bottom-up exploration
- Any good collaborative tool
- Enable navigation with access to the details
- Visualizations must be tailored to the data and questions.
- CDI is neither low level, but it
 - Is easily flexible (like it came graphics)
 - Provides layers of exploration
 - Can provide other forms of interaction
- It's lightweight package under development (github.com/kwbroman/cdi)

Acknowledgments

Example 1

Alan Attie, Mark Keller, Alissa W. Broman, Christina Kondratieva, Brian Vander, Eric Schaff, Department of Biostatistics, Biostatistics & Medical Informatics, and Statistics, UW-Madison, Madison, WI

Example 2

Caroline Moore, Edgar Spalding, Logan Johnson, B. S. Deep Kaul, M. S. Lacey, Department of History, Statistics, and Computer Science, UW-Madison

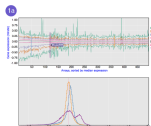
Contact

Karl Broman
kbroman@biostat.wisc.edu
broman@biostat.wisc.edu
biostat.wisc.edu
kbroman@biostat.wisc.edu

The web site supported by grant by NSF grant DMS-1000000

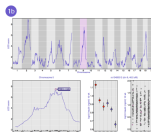
Example 1: Expression genetics

- Gene expression microarrays in six tissues
- Numerous clinical phenotypes

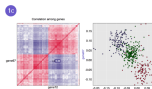


These are data from ~500 gene expression microarrays. The top panel is like 500 gene plots. These are data at the L1...99 percentiles for each of ~500 distributions. The distributions are sorted by their means.

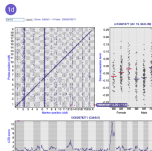
If you hover over a column in the top panel, the corresponding distribution is shown below. Click on the top panel for the distribution to zoom in, and click again to make it go away.



A genome scan for genetic loci (Loci) influencing gene expression. The top panel is like 500 gene plots. These are data at the L1...99 percentiles for each of ~500 distributions. The distributions are sorted by their means.



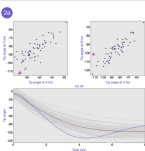
Association in gene expression among 300 genes that are influenced by a common genetic locus (Loci). The left panel is a heatmap of the correlation matrix, with blue = -1 and red = 1. Hover over points in the correlation matrix on the left to see the values; click to see the corresponding scatterplot on the right. Points in the scatterplot are colored by genotype of the underlying Loci.



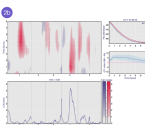
An investigation of genetic loci (Loci) influencing gene expression. In the top panel, the x-axis corresponds to marker location and the y-axis corresponds to the position of markers on a gene expression microarray. Each colored point is an individual Loci.

Example 2: Gravitropism

- Response to gravity in Arabidopsis seedlings
- Rotational orientation of gravity and values over time
- Measure the angle of the root tip every 2 min



Average root angle over time for 100 Arabidopsis lines. Hover over points in the top panel or curves in the bottom panel to highlight the corresponding line in the other panels.



The top-left panel is a heatmap of a measure of association (LOD score) between genotype at a fixed position and the phenotype at a fixed time. Red lines indicate that the Loci (Loci) lines have higher phenotype.

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.