

# 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](https://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 Gramm<sup>1</sup>   Tzvika Hartman<sup>2</sup>   Till Nierhoff<sup>3</sup>  
Roded Sharan<sup>4</sup>   **Till Tantau<sup>5</sup>**

<sup>1</sup>Universität Tübingen, Germany

<sup>2</sup>Bar-Ilan University, Ramat-Gan, Israel

<sup>3</sup>International Computer Science Institute, Berkeley, USA

<sup>4</sup>Tel-Aviv University, Israel

<sup>5</sup>Universität zu Lübeck, Germany

Workshop on Algorithms in Bioinformatics, 2006

# Get rid of the junk

```
\usetheme{default}
```

```
\beamertemplatenavigationsymbolsempy
```

# Change colors

```
\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}

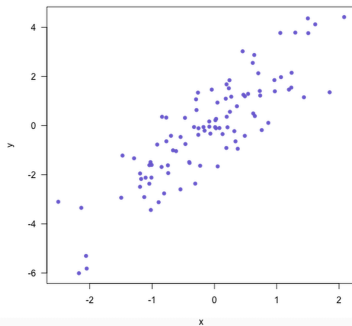
\bbi
\item {\lollit \only<1>{\color{foreground}} Bullet 1}
\item {\lollit \only<2>{\color{foreground}} Bullet 2}
\item {\lollit \only<3>{\color{foreground}} Bullet 3}
\item {\lollit \only<4>{\color{foreground}} Bullet 4}
\ei

\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 Broman<sup>1</sup>, Mark P Keller<sup>2</sup>, Ameer Teo Broman<sup>1</sup>, Danielle M Greenawald<sup>3</sup>,  
Christina Kendziora<sup>3</sup>, Eric E Schadt<sup>4</sup>, Šaunak Sen<sup>5</sup>, Brian S Yandell<sup>6,7</sup>, and Alan D Attie<sup>2</sup>

<sup>1</sup>Biostatistics and Medical Informatics, <sup>2</sup>Biochemistry, <sup>3</sup>Statistics, <sup>4</sup>Heritability, UW-Madison, <sup>5</sup>Merck & Co., Inc., <sup>6</sup>Pacific Biosciences, <sup>7</sup>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, at the order of 15%.

Local eQTLs, genetic loci influencing gene expression in an 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 used it to identify other individuals whose genotypes did match the predicted eQTL genotypes.

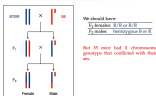
The concordance of predictions across six tissues indicated that the problems were due to mix-ups in the genotypes. Correlations of the gene expressions of the transcripts indicated a number of self by one and self by two errors, likely the result of pipetting errors.

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

### Data

- ~5000 tis. × 100000 markers data, all cis
- Genotypes at 257 SNPs (Autism chip)
- Gene expression in six tissues (Agilent arrays)
- Adverse, gene expression results: Hypertension, juvenile onset, history (new)
- Numerous clinical phenotypes (e.g. body weight, insulin and glucose levels)

### Initial observation: Sex swaps



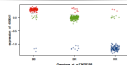
### Which are correct: genotypes or sexes?

- 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 QTL, eQTL that influences the level of expression of a gene

### A diagnostic transcript



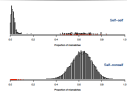
Colors indicate the inferred eQTL genotype according to a nearest neighbor classifier, with gray points not called.

### 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 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 ~500 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 estimated each proportion of mismatches, comparing that observed genotype data to each of inferred genotypes.

### Inferred genotype mix-ups

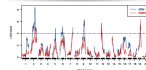


The gene expression data from the multiple tissues were concordant and both called that the problems were in the genotype data.

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 assuming the correlation in a mouse's gene expression data 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 Madhuri)

### Improved results



LSD curves for insulin, 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/kwibrom/1mixups>
- Very similar to <https://github.com/kwibrom/1mixups> (Phenostat, Bioinformatics 2016:201, 2017, 2018)

### Contacts

Karl Broman

karl.broman@wisc.edu, wisc.edu

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

This work was supported by grant R01HG004608 from NIGMS and 5R01HG004608.

# 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 statistics.
- All graphs could be improved with some interactivity.

Deviant: [bit.ly/uuu22014](http://bit.ly/uuu22014)

### Opportunities

- Exploration
  - Testing parameters
  - Identifying outliers
  - One fancy plot vs 1000 static plots
- Reports for collaboration
  - Living document
  - Allow deeper exploration of the results
  - Can you see sample questions?
- Big Data
  - Don't just rely on summary statistics
  - Compress information, but with access to the details
  - Zoom into data layers
  - More exploration, more connections
- Teaching
  - Can things be look at all play with
  - Animated dimensions of key concepts
  - Democratize data exploration
  - Enable users to explore 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)
- Incredible power of modern web browsers
- JavaScript-based web tools
- Reusable tools

### DS

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

### Other options

- Interact (j) and interact (j)
- ggplot (j), rpy and rconsole (j), rshiny (www.rstudio.com)
- Shiny (j) (rshiny.com)
- Acinony (j) (Acinony.com)
- ggplotViz (j) (ggplotviz.com)
- Shiny (j) (shiny.com)
- ggplot (j) (ggplot.com)
- ggplot (j) (ggplot.com)
- ggplot (j) (ggplot.com)

### simple ↔ flexible

Choose one, I choose flexible.

### Summary

- For high-dimensional data, good visualizations are critical.
- Interactive graphics, responsive, better
- Can be done with many tools
- Enable exploration with access to the details
- Visualizations must be tailored to the data and questions.
- Can be done with many tools
- Can be done with many tools
- Can be done with many tools
- Can be done with many tools

### Acknowledgments

Example 1: Alan Attie, Mark Keller, Alissa W. Broman, Christina Kozlowski, Brian Vander, Eric Schaff, Department of Biostatistics, Biostatistics & Medical Informatics, and Statistics, UW-Madison, Madison, WI.

Example 2: Candace Moore, Edgar Spalding, Logan Johnson, B. S. Broman, M. S. Broman, Department of Biostatistics, and Statistics, UW-Madison, Madison, WI.

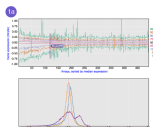
### Contact

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

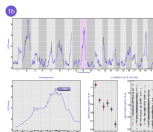
The web site was supported by grant by NSF grant DMS-1008000.

### Example 1: Expression genetics

- Gene expression microarrays in six tissues
- Genetic expression phenotypes



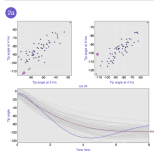
These are data from ~500 gene expression microarrays. The top panel is a heatmap of gene expression across six tissues. The bottom panel is a line plot showing the distribution of gene expression across six tissues.



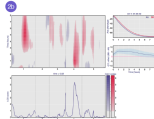
A genome scan for genetic loci (colored quantitative trait loci, QTLs) influencing muscle level. The LOD score is a log10 likelihood ratio measuring the strength of association between genotype and phenotype. Click on a chromosome at the top and a detailed view of the LOD score for that chromosome is shown in the bottom left. In the lower-left panel, hover over markers to see names; click to view an effect plot and phenotype vs genotype plot to the right.

### Example 2: Gravitropism

- Response to gravity in Arabidopsis seedlings
- Gravitropism: curvature of growth and velocity over time
- Measure the angle of the root tip every 2 min



Average root angle over time for 162 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 lines have larger phenotype. When you hover over a point in the top-left plot, the LOD curve for the corresponding time is shown below, and the phenotype average and estimated genetic effect (across time) are shown to the right.

An investigation of genetic loci (QTLs) influencing gene expression. In the top-left 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 QTL. Hover over a point to see the probe ID and LOD score (measuring the strength of association) also highlighted are any other QTLs for that probe. Click on the point to see the LOD curve 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.

[github.com/kbroman/Poster\\_ENAR2014](https://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.