

Data visualizations should be more interactive

Karl W Broman

Biostatistics & Medical Informatics, University of Wisconsin–Madison

Abstract

The value of interactive graphics for making sense of high-dimensional data has long been appreciated but is still not in routine use. Interactive graphics facilitate data exploration, are great collaborative tools, allow compressed summary plots to be linked to the underlying details, and can be fabulous teaching tools. New web-browser-based tools, such as the JavaScript library D3, have greatly simplified the development of interactive visualizations. A number of R packages, including Shiny and rCharts, provide customizable interactive visualizations directly from R. I provide a number of examples to illustrate the value of interactive graphics, with live demonstrations on a laptop. (I'd initially planned to use handheld devices, but the resolution is too poor for what I am trying to do.)

Live 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

Options

- **locator()** and **identify()**
- **ggobi** (ggobi.org) and **cranvas** (github.com/ggobi/cranvas/wiki)
- **Mondrian** (rosuda.org/software/Mondrian)
- **gWidgets** (gwidgets.r-forge.r-project.org)
- **rpanel** (www.stats.gla.ac.uk/~adrian/rpanel)
- **Acinoynx** (aka **iPlots eXtreme**) (rforge.net/Acinonynx)
- **googleVis** (code.google.com/p/google-motion-charts-with-r)
- **gridSVG** (r-forge.r-project.org/projects/gridsvg)
- **RGL** (rgl.neoscientists.org)
- RStudio's **manipulate()** (www.rstudio.com/ide/docs/advanced/manipulate)
- RStudio's **Shiny** (www.rstudio.com/shiny)
- RStudio's **ggvis** (ggvis.rstudio.com)
- **Rcharts** (rcharts.io)
- **D3** (d3js.org)

simple ↔ flexible

Choose one.
I choose **flexible**.

D3

- Javascript library for manipulating HTML and SVG elements
- Connects data to elements
- Low level, but flexible
- Used at the NY Times
- I use CoffeeScript rather than JavaScript

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

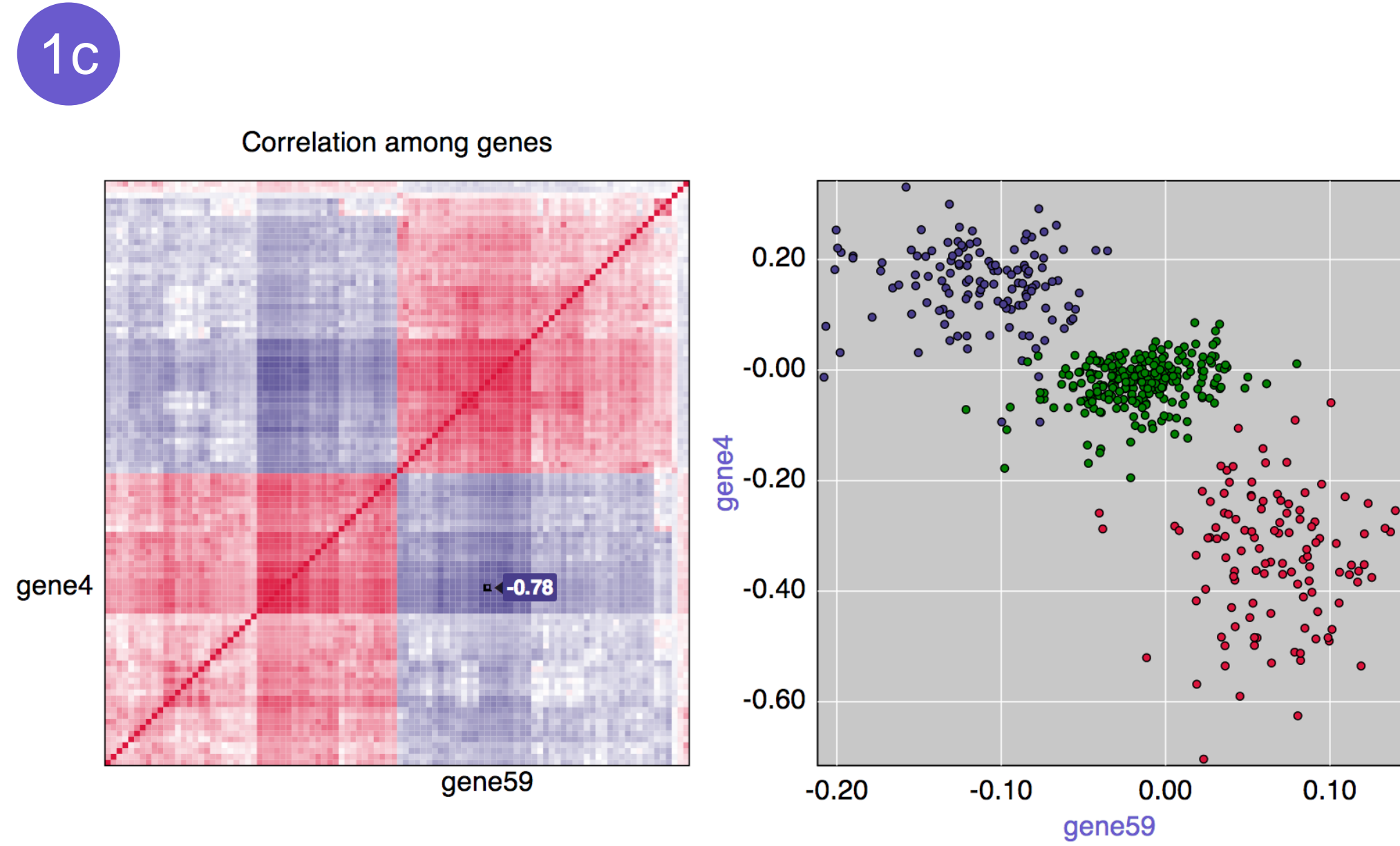
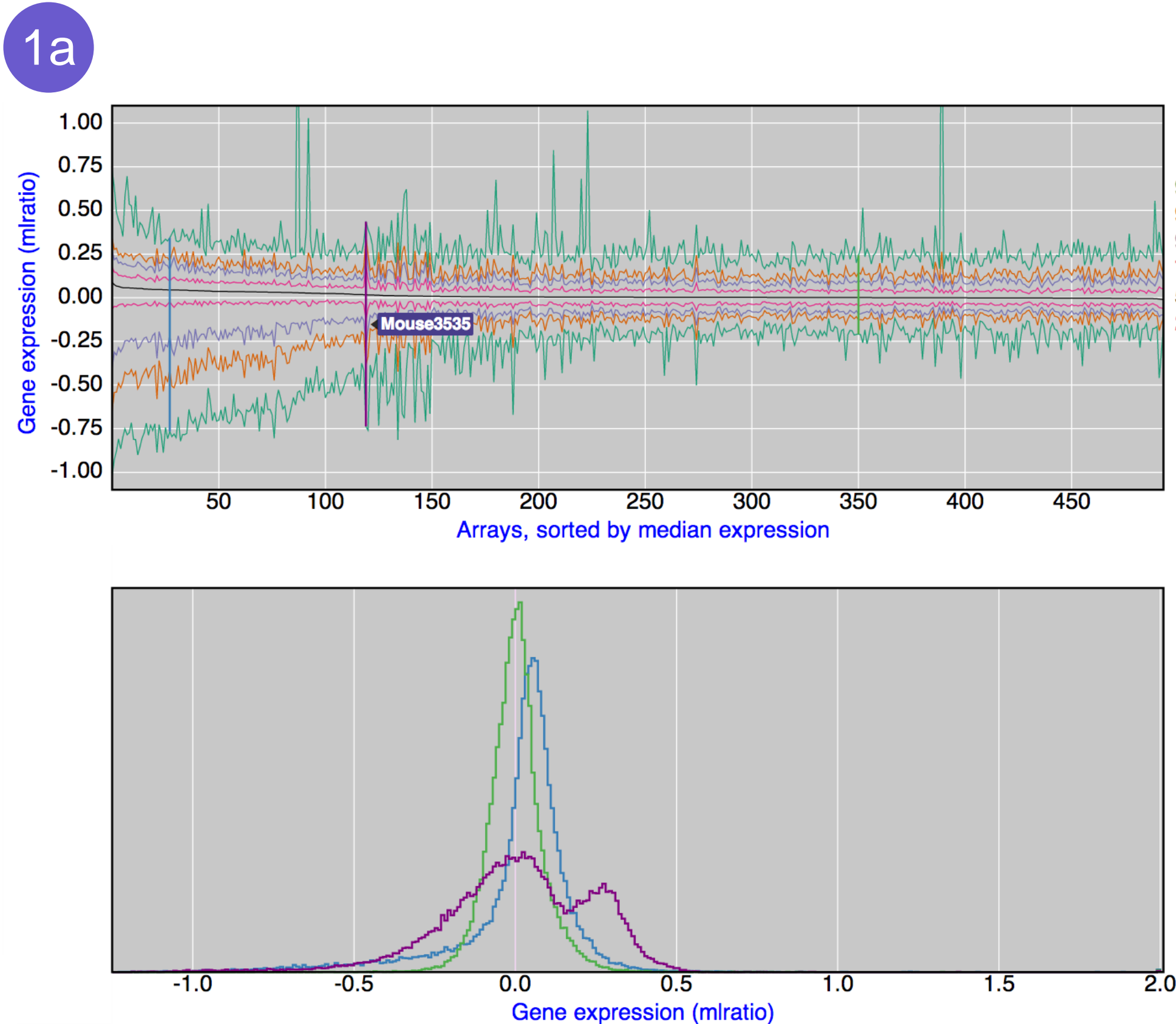
Contact

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

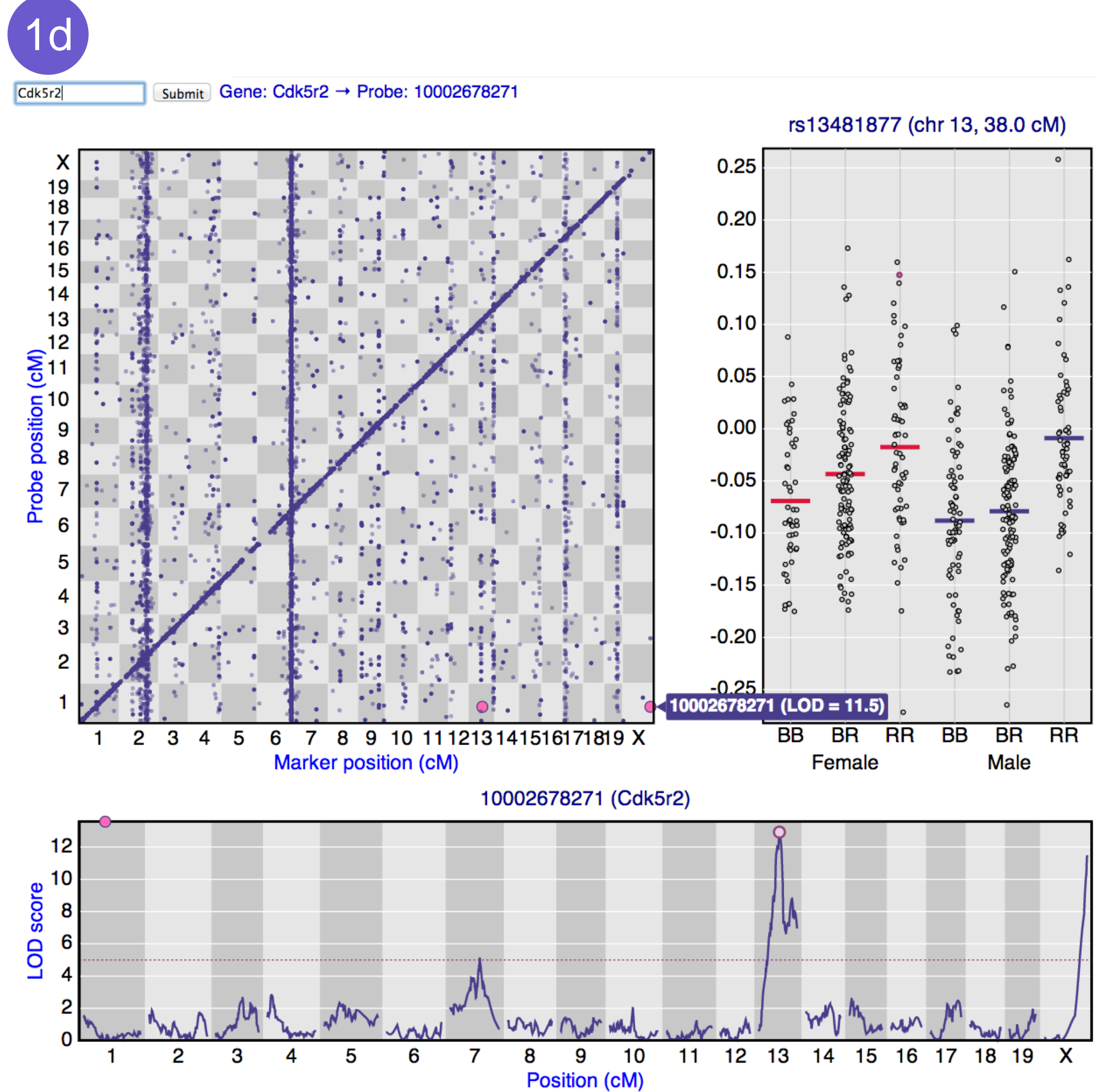
This work was supported in part by NIH grant GM074244.

Example 1: Expression genetics

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



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.



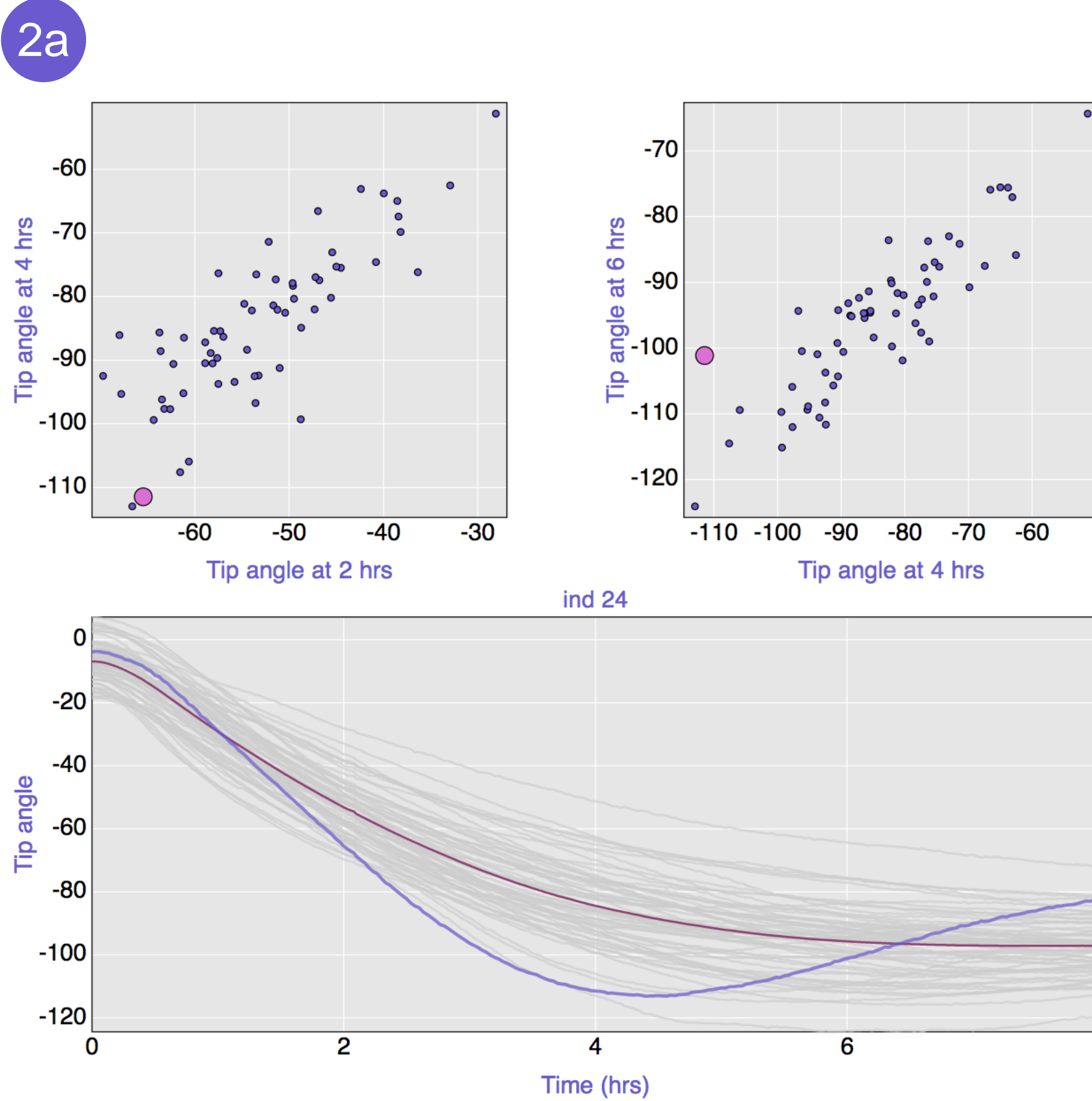
An investigation of genetic loci (eQTL) influencing gene expression. In the top-left panel, the x-axis corresponds to marker location and the y-axis corresponds to the position of probes on a gene expression microarray. Each plotted point is an inferred eQTL.

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.

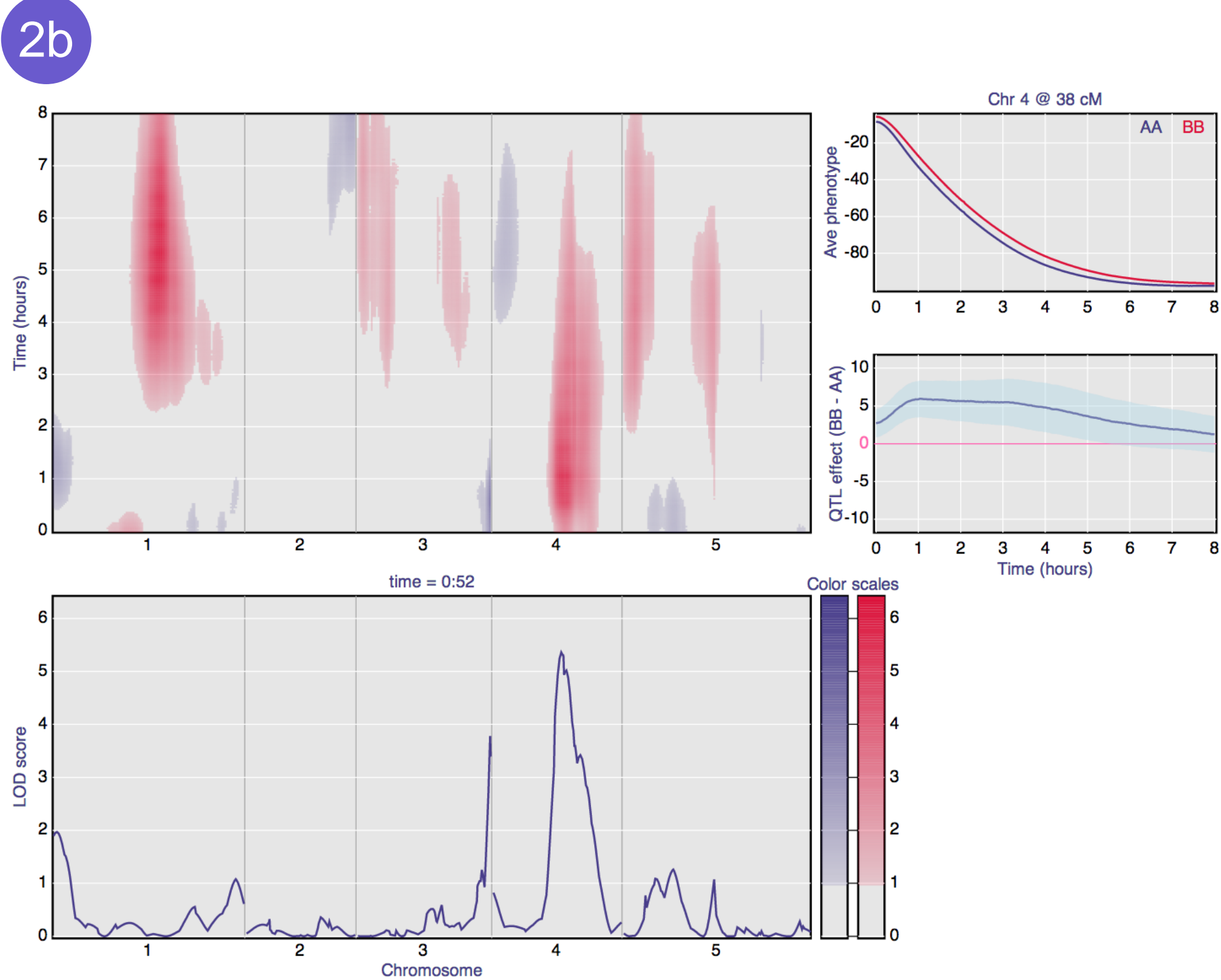
Example 2: Gravitropism

- 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



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.



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.