

Data visualization tutorial: exploring data and telling stories using ggplot2*

Peter Carbonetto *University of Chicago*

Programming challenge: Mapping the genetic basis of physiological and behavioral traits in outbred mice

In this programming challenge, you will use simple visualizations to gain insight into biological data.

You are working in a lab studying the genetics of physiological and behavioral traits in mice. The lab has just completed a large study of mice from an outbred mouse population, “CFW” (“Carworth Farms White”). The aim of the study is to identify genetic contributors to variation in behaviour and musculoskeletal traits.

Note these challenges are roughly ordered in increasing level of complexity. *Do not be discouraged if you have difficulty completing every one.*

Collaboration strategy

Before diving into the problems, first agree on a collaboration strategy with your teammates. Important aspects include communication and co-ordination practices, and setting goals and deadlines. How will your team collaborate on code, and share solutions? (Consider online resources such as Etherpad or the UofC-hosted Google Drive.) The aim is not just to tackle the challenges, but also to do so collaboratively.

Instructions

- Locate the files for this exercise on your computer (see “Materials” below).
- Make sure your R working directory is set to the same directory containing the tutorial materials; use `getwd()` to check this. (If you need to change your R working directory, you can either use the `setwd()` function or, in RStudio, you can select **Session > Set Working Directory > Choose Directory...**)
- Some of the programming challenges require uploading an image file containing a plot. Use `ggsave` to save your plot as a file. Any standard image format (e.g., PDF, PNG) is fine.
- No additional R packages are needed beyond what we used in the examples above.

Materials

- **pheno.csv:** CSV file containing physiological and behavioral phenotype data on 1,219 male mice from the CFW outbred mouse stock. Data are from [Parker *et al*, 2016](#). Use `readpheno.R` to read the phenotype data from the CSV file into a data frame. After filtering out some of

*This document is included as part of the Data Visualization tutorial packet for the BSD qBio9 Bootcamp, University of Chicago, 2023. **Current version:** August 11, 2023; **Corresponding author:** pcarbo@uchicago.edu.

the samples, this script should create a new data frame, `pheno`, containing phenotype data on 1,092 samples (these are the rows of the data frame).

- **hmdp.csv:** CSV file containing bone-mineral density measurements taken in 878 male mice from the Hybrid Mouse Diversity Panel (HMDP). Data are from [Farber et al, 2011](#). To load the data into your R environment, run this code:

```
hmdp <- read.csv("hmdp.csv", stringsAsFactors = FALSE)
```

This will create a data frame, `hmdp`, containing BMD data on 878 mice (these are the rows of the data frame).

- **gwscan.csv:** CSV file containing results of a “genome-wide scan” for abnormal BMD. (The association p -values were computed using GEMMA 0.96.) To read the results of the genome-wide scan, run the following code:

```
gwscan <- read.csv("gwscan.csv", stringsAsFactors = FALSE)
gwscan <- transform(gwscan, chr = factor(chr, 1:19))
```

This will create a data frame, `gwscan`. Each row of the data frame is a genetic variant (a single nucleotide polymorphism, or “SNP”). The columns are chromosome (“chr”), base-pair position on the chromosome (“pos”), and the p -value for a test of association between variant genotype and trait value (“abnormalBMD”). The value stored in the “abnormalBMD” column is $-\log_{10}(P)$, where P is the p -value.

- **geno_rs29477109.csv:** CSV file containing estimated genotypes at one SNP (rs29477109) for 1,038 CFW mice. Use the following code to read the genotype data into your R environment:

```
geno <- read.csv("geno_rs29477109.csv", stringsAsFactors = FALSE)
geno <- transform(geno, id = as.character(id))
```

This will create a new data frame, `geno`, with 1,038 rows (samples). The genotypes are encoded as “dosages”—that is, the expected number of times the alternative allele is observed in the genotype. This will be an integer (0, 1, 2), or a real number between 0 and 2 when there is some uncertainty in the estimate of the genotype. For this SNP, the reference allele is T and the alternative allele is C. Therefore, dosages 0, 1 and 2 correspond to genotypes TT, CT and CC, respectively (genotypes CT and TC are equivalent).

- **wtccc.png:** Example genome-wide scan (“Manhattan plot”) from Fig. 4 of the [WTCCC paper](#). The p -values highlighted in green show the regions of the human genome most strongly associated with Crohn’s disease risk.

A couple tips

- Some “geoms” you may find useful: `geom_point`, `geom_histogram`, `geom_boxplot`.
- In some cases it may be useful to convert to a *factor*.

Part A: Exploratory analysis of muscle development and conditioned fear data

Your first task is to create plots to explore the data.

1. A basic initial step in an exploratory analysis is to visualize the distribution of the data. It is often convenient if the distribution is normal, or “bell shaped”.

- Visualize the empirical distribution of tibialis anterior (TA) muscle weight (column “TA”) with a histogram. Units are mg. *Hint:* Try using function `geom_histogram`.
 - Is the distribution of TA weight roughly normal? Are there mice with unusually large or unusually small values (“outliers”)? If so, how many “outliers” are there?
2. It is also important to understand relationships among measured quantities. For example, the development of the tibia bone (column “tibia”) could influence TA muscle weight. Create a scatterplot (`geom_point`) to visualize the relationship between TA weight and tibia length. (Tibia length units are mm.) Based on this plot, what can you say about the relationship between TA weight and tibia length? Quantify this relationship by fitting a linear model, before and after removing the outlying TA values. (*Hint:* Use the `lm` and `summary` functions. See also the “`r.squared`” return value in `help(summary.lm)`.)
 3. The “AvToneD3” column contains data collected from a behavioral test called the “Conditioned Fear” test.
 - Visualize the empirical distribution of AvToneD3 (“freezing to cue”) with a histogram. Is the distribution of AvToneD3 approximately normal?
 - Freezing to cue is a proportion (a number between 0 and 1). A common way to obtain a more “normal” quantity is to transform it using the “logit” function.¹ Visualize the empirical distribution of the logit-transformed phenotype. Is the transformed phenotype more “bell shaped”? After the transformation, do you observe unusually small or unusually large values?
 - A common concern with behavioral tests is that the testing devices can lead to measurement error. It is especially a concern when multiple devices are used, as the devices can give slightly different measurements, even after careful calibration. Create a plot to visualize the relationship between (transformed) freezing to cue and the device used (“FCbox” column). *Hint:* Try a boxplot (`geom_boxplot`). Based on this plot, does the apparatus used affect these behavioral test measurements?

Part B: Exploratory analysis of bone-mineral density data

Now you will examine data on bone-mineral density (BMD) in mice. This is a trait that is important for studying human diseases such as osteoporosis (units are mg/cm^2).

- Plot the distribution of BMD in CFW mice (see column “BMD”). What is most notable about the distribution?
- Compare these data against BMD measurements taken in a “reference” mouse population, the Hybrid Mouse Diversity Panel. To compare, create two histograms, and draw one on top of the other. What difference do you observe in the BMD distributions? For a correct comparison, consider that: (1) BMD in CFW mice was measured in the femurs of male mice only; (2) BMD in HMDP mice was recorded in g/cm^2 . *Hints:* Potentially useful functions include `xlim` and `labs` from the `ggplot2` package, and `plot_grid` from the `cowplot` package. The `binwidth` argument in `geom_histogram` may also be useful.

¹R code: `logit <- function(x) log((x + 0.001) / (1 - x + 0.001))`

Part C: Mapping the genetic basis of osteopetrotic bones

A binary trait, “abnormal BMD”, was defined that signals whether an individual mouse had “abnormal”, or osteopetrotic, bones. It takes a value of 1 when BMD falls on the “long tail” of the distribution (BMD greater than 90 mg/cm²), otherwise zero.

GEMMA was used to carry out a “genome-wide association study” (GWAS) for this trait; that is, support for association with abnormal BMD was evaluated at 79,824 genetic variants (single nucleotide polymorphisms, or “SNPs”) on chromosomes 1–19. At each SNP, a p -value quantifies the support for an association with abnormal BMD.

1. Your first task is to get an overview of the association results by creating a “Manhattan plot”. Follow as closely as possible the provided prototype, **wtccc.png**, which shows a genome-wide scan for Crohn’s disease. (Don’t worry about highlighting the strongest p -values in green.) *Hints:* Replicating some elements of this plot may be more challenging than others, so start with a simple plot, and try to improve on it. Recall the adage that creating plots requires relatively little effort *provided the data are in the right form*—consider adding appropriate columns to the `gwscore` data frame before writing your `ggplot` code. Functions from the `ggplot2` package that you may find useful include `geom_point`, `scale_color_manual` and `scale_x_continuous`.
 - In your plot, you should observe that the most strongly associated SNPs cluster closely together in small regions of the genome. This is common—it is due to a genetic phenomenon known as linkage disequilibrium (LD). It is a consequence of low recombination rates between markers in small populations. How many SNPs have “strong” statistical support for association with abnormal BMD, specifically with a $-\log_{10} p$ -value > 6 ? How many “distinct” regions of the genome are strongly associated with abnormal BMD at this p -value threshold?
 - What p -value does a $-\log_{10} p$ -value of 6 correspond to?
 - Using your plot, identify the “distinct region” (this is called a “quantitative trait locus”, or QTL) with the strongest association signal. What is, roughly, the size of the QTL in Megabases (Mb) if we define the QTL by base-pair positions of the SNPs with $-\log_{10} p$ -value > 6 ? Using the [UCSC Genome Browser](#), get a rough count of the number of genes that are *transcribed* in this region. (Parker *et al*, 2016 identified *Col1a1* as a candidate BMD gene.) *Hint:* All SNP positions are based on *NCBI Mouse Genome Assembly 38* (mm10, December 2011).
2. Your next task is to visualize the relationship between genotype and phenotype. From the genome-wide scan of abnormal BMD, you should find that rs29477109 is the SNP most strongly associated with abnormal BMD. Here you will look closely at the relationship between BMD and the genotype at this SNP. In developing your visualization, consider that:
 - The samples listed in the phenotype and genotype tables are not the same. So you will need to align the two tables to properly show analyze the relationship. *Hint:* Function `match` could be useful for this.
 - The genotypes, stored in file **geno_rs29477109.csv**, are encoded as “dosages” (numbers between 0 and 2). You could start with a scatterplot of BMD vs. dosage. But ultimately it is more effective if the genotypes (CC, CT and TT) are plotted instead. *Hints:* In effect, what you need to do is convert from a continuous variable (dosage) to a discrete variable (genotype). One approach is to create a factor column from the “dosage”

column. (For dosages that are not exactly 0, 1 or 2, you could simply round to the nearest whole number.) A boxplot is recommended; see function `geom_boxplot`.

Based on your plot, how would describe (in plain language) the relationship between the genotype and BMD?