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

Peter Carbonetto *University of Chicago*

In this tutorial, we will use ggplot2 to create effective visualizations of biological data. The ggplot2 package is a powerful set of plotting functions that extend the base plotting functions in R. In this lesson, we will also see that creating effective visualizations in R *hinges on good data preparation.* In reality, good data preparation in the lab can take days or weeks, but here we can still illustrate some useful data preparation practices. The main difference with Advanced Computing 1 is that we take a more careful look at ggplot2 and strategies for data visualization.

Hands-on exercise: "Rising Dough, Rising Neighbourhoods"

The scenario

You have just begun a summer internship at the Mansueto Institute for Urban Innovation. The institute has formed a partnership with the City of Chicago to develop new ways of quickly gaining insight into local economic activity throughout the city.

A computer scientist working at the institute recently has been exploring ways to measure economic health of neighbourhoods by aggregating publicly available data on pizza restaurants throughout the city. The city is excited about the potential for this project, and would like to present a report to the Mayor. Unfortunately, the computer scientist left a few weeks ago to start a research team at Amazon's new Headquarters in the West Loop. (*Disclaimer:* In case you haven't figured this out already, much of this story is fictional.) So, as the new student intern, you have been tasked with building on the computer scientist's work to put together a report.

The good news is that this computer scientist was diligent about keeping track of the code and data files she used in her analyses. She even went so far as to provide detailed comments in her code explaining what the code does. (This is perhaps one of the less realistic parts of this scenario.) Your advisor has provided you with the R code and data files that were used to generate the plots that will go in the report. Therefore, although your only experience in R is from a statistics course you took during your undergrad, your initial sense of dread has turned into cautious optimism.

Instructions

- Make sure you have downloaded the tutorial packet.
- Locate the files for this exercise on your computer (see "Materials" below).
- To run the code, you will need to have the following R packages installed on your computer: **ggplot2**, **cowplot** and **readr**.
- Open the the R source file, **pizzaplots.R**, in RStudio, or in your favourite editor (e.g., emacs).
- Make sure your R working directory is the same directory containing the tutorial materials; use `getwd()` to check this.

*This document is included as part of the Data Visualization tutorial packet for the BSD qBio Bootcamp, MBL, 2018. **Current version**: August 07, 2018; **Corresponding author**: pcarbo@uchicago.edu. Thanks to John Novembre and Matthew Stephens for their support and guidance.

- Follow the instructor for additional instructions (see also the slides PDF included with the tutorial packet).

Materials

- **pizzaplots.R:** R code you will use to reproduce the plots.
- **Food_Inspections.csv.gz**: Compressed text file in CSV Format containing the food inspection data that were downloaded from the Chicago Data Portal.

Follow-up programming challenges

1. Our first plot was a bar chart showing an estimate of the number of new pizza restaurants per year in Chicago. Sometimes lines are more effective than bars for showing trends.
    - If you wanted to show the trend as a line plot, how would you modify the plotting code in **pizzaplots.R** to do this? *Hint:* `geom_line` might be useful.
    - What change do you need to make to the `counts` data frame so that counts can be plotted as lines?
    - Which do you find more effective, the line plot or the bar chart? Identify one benefit of one plot over the other.
    - To submit your response, you will need to upload a file containing your final plot. Use `ggsave` to save your plot as a file.
2. Your advisor has asked you to colour the bars in the bar chart so that they align with the colours used in the map.
    - What change do you need to make to the `counts` data frame so that the years can be mapped to colors? *Hint:* The code used for creating the line plot may be useful here, too.
    - Submit the *combined* plot (bar chart & map). When submitting your plot, save it as a PDF using `ggsave`.
    - Why did your advisor request a PDF? Why is a PNG file less suitable for a final report or publication?

Main 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 have finished your summer internship at the Mansueto Institute, and you are now embarking on your first graduate research project in a lab studying 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 known as "CFW" (short for "Carworth Farms White", the names of the scientists who bred the first CFW mice). The ultimate 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 all the exercises. Also, do not hesitate to ask the instructors for advice if you get stuck.

Instructions

- Locate the files for this exercise on your computer (see "Materials" below).
- Make sure your R working directory is the same directory containing the tutorial materials; use `getwd()` to check this.
- Follow the instructor for additional guidance (see also the slides included in the tutorial packet).
- Some of the programming challenges require uploading an image file containing a plot. You can use `ggsave` to save your plot as a file; any standard image format is acceptable.
- No additional R packages are needed beyond what you used in the hands-on exercise above.

Materials

- **pheno.csv:** Text file containing physiological and behavioral phenotype data on 1,219 male mice from the CFW outbred mouse stock. The data are stored in comma-delimited (CSV) format, with one sample per line. Data are from Parker *et al*, 2016. Use **readpheno.R** to read the phenotype data from the CSV file into a data frame. After discarding some of the samples, this script will create a data frame, `pheno`, containing phenotype data on 1,092 samples, with one row per sample.
- **hmdp.csv:** Text file in CSV format containing bone-mineral density measurements taken in 878 mice from the Hybrid Mouse Diversity Panel (HMDP). Data are from Farber *et al*, 2011. To load the data into your R environment, run the following code.

```
hmdp <- read.csv("hmdp.csv",stringsAsFactors = FALSE)
hmdp <- transform(hmdp,sex = factor(sex,c("M","F")))
```

  This will create a data frame, `hmdp`, containing BMD data on 878 mice, with one mouse per row.

- **gwscan.csv:** Text file in CSV format containing results of a "genome-wide scan" for abnormal BMD. 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 single genetic variant in the mouse genome (a single nucleotide polymorphism, or "SNP"). The columns give the 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 association test p-value.

- **geno_rs29477109.csv:** Text file in CSV format 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, one for each sample (mouse). The genotypes are encoded as "dosages"; specifically, the expected number of times the alternative allele is observed in the genotype. This will either be an integer (0, 1 or 2), or a

real number between 0 and 2 when there is some uncertainty in the estimate of the genotype. In this case, 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 (note genotypes CT and TC are equivalent).

- **wtccc.png:** Example genome-wide scan ("Manhattan plot") taken 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.

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

Your first task is to create plots to explore of some of the phenotype data collected for the CFW study.

1. A basic initial step in an exploratory analysis is to visualize the empirical distribution of the data. For several reasons (e.g., to ensure validity of statistical tests used), it is convenient if the empirical 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 function `geom_histogram`.
   - Is the distribution of TA weight roughly normal? Are there mice with unusually large or unusually small TA muscles (*i.e.*, "outliers")? If so, how many "outliers" are there? It is sometimes important to identify outliers, since unusually small or large values can lead to misleading results in standard statistical tests.
2. It is also often important to understand the relationships among the measured quantities to be analyzed. 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. Units of tibia length are mm. Based on this plot, what can you say about the relationship between TA weight and tibia length? Also, quantify this relationship by fitting a linear model, before and after removing the outlying TA values. *Hint*: Use the `lm` and `summary` functions for this. If you are unsure how to quantify the relationship, see the description of the "r.squared" output in `help(summary.lm)`.
3. The "AvToneD3" column contains data collected from a behavioral test called the "Conditioned Fear" test. Specifically, AvToneD3 is the average proportion of time freezing on the third day of testing during the presentation of tones (the conditioned stimulus). The shorthand for this behavioral phenotype is "freezing to cue".
   - Visualize the empirical distribution of 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-behaving proportion 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 devices used in the tests 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 freezing to cue (the transformed version) and the device used (column "FCbox"). *Hint:* Try creating a boxplot using

[1] An R implementation of the logit function: `logit <- function(x) log((x + 0.001)/(1 - x + 0.001))`

`geom_boxplot`. Based on this plot, would you say that the apparatus used affected the measurements in this behavioral test?

Part B: Exploratory analyses of bone-mineral density data

In this part, you will examine data on bone-mineral density in mice. This is a trait that is important for studying human diseases such as osteoporosis. The units of BMD are $\mathrm{mg/cm^2}$. (Strictly speaking, this is not a density—it is "areal" BMD, which is easier to measure, and is considered a good approximation to the true BMD.)

- Plot the distribution of BMD in CFW mice (see column "BMD"). What feature stands out from the histogram?
- To investigate whether this feature is particular to the CFW mouse population, compare these BMD data against BMD measurements taken in a "reference" mouse population. As reference, we will use the Hybrid Mouse Diversity Panel. To provide a direct visual comparison, create two histograms, and draw them one on top of the other. What difference do you observe in the BMD distributions? Note that, in the CFW study, BMD was measured in the femurs of male mice (whereas the HMDP study included males and females). Further note that BMD in the HMDP data set is recorded in $\mathrm{g/cm^2}$. *Tips:* Functions `xlim` and `labs` from the ggplot2 package, and `plot_grid` from the cowplot package, might be useful for creating the plots. The `binwidth` argument in `geom_histogram` may also be useful.

Part C: Mapping the genetic basis of osteopetrotic bones

Based on the exploratory analyses of BMD in CFW and HMDP mice, we defined a binary trait, "abnormal BMD", 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 observed distribution (BMD greater than 90 $\mathrm{mg/cm^2}$), and 0 otherwise.

We used GEMMA to carry out a "genome-wide association study" (GWAS) for this trait; that is, we estimated support for association between abnormal BMD and 79,824 genetic variants (single nucleotide polymorphisms, or "SNPs") on chromosomes 1–19. At each SNP, we computed a *p*-value to assess the support for association with abnormal BMD.

1. After running the *p*-value computations, the next step in any GWAS is to get an overview of the association results. Your task here is to create a "Manhattan plot" summarizing the results. Follow as closely as possible the provided prototype, **wtccc.png**, which shows a genome-wide scan for Crohn's disease. This prototype plot is from an influential paper that set many of the standards for conducting genome-wide association studies. (Note you do not need to worry about highlighting the strongest *p*-values in green since the threshold for choosing the "strongest" *p*-values is not clearly defined in this study.) *A few tips:* 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 `gwscan` data frame. Functions from the ggplot2 package that you may find useful for this exercise 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 a common situation, and is due to a genetic phenomenon known as linkage disequilibrium (LD). It arises as a natural 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 "quantitative trait locus" (QTL) with the strongest association signal (this is the region where the strongest associations cluster). What is, roughly, the size of the QTL in Megabases (Mb), if we define the QTL by the 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. Within this QTL, Parker *et al*, 2016 identified *Col1a1* as a compelling candidate for being the causal BMD gene. Was this gene one of the genes included in your count? *Note:* All SNP positions are based on Mouse Genome Assembly 38 from the NCBI database (mm10, December 2011).

2. In this last exercise, your task is to visualize the relationship between genotype and phenotype. From the genome-wide scan of abnormal BMD, we identified rs29477109 as 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 the following:
   - 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` is 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 plot of BMD vs. dosage. But ultimately it is more directly interpretable if the genotypes (CC, CT and TT) are plotted instead. *Hint:* 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 new `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.

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

Notes

Useful online resources

- ggplot2 reference, where you will also find a ggplot2 cheat sheet. (This cheat sheet is also included in the tutorial packet, and you may have seen it in a previous tutorial.)
- Fundamentals of Data Visualization by Claus Wilke. If you are interested, I've also generated a PDF version of this book from Claus's source code. You can access the PDF here on the UChicago Box. This is a book I wish I had when I started my Ph.D!

License

Session info

This next code chunk gives information about the computing environment, including the version of R and the R packages, that were used to test the tutorial examples.

```
sessionInfo()
# R version 3.4.3 (2017-11-30)
# Platform: x86_64-apple-darwin15.6.0 (64-bit)
# Running under: macOS High Sierra 10.13.6
#
# Matrix products: default
# BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
# LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib
#
# locale:
# [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
#
# attached base packages:
# [1] stats     graphics  grDevices utils     datasets  base
#
# other attached packages:
# [1] cowplot_0.9.2.9900 ggplot2_3.0.0.9000 readr_1.1.1
#
# loaded via a namespace (and not attached):
#  [1] Rcpp_0.12.17     bindr_0.1.1      knitr_1.20       magrittr_1.5
#  [5] hms_0.4.0        tidyselect_0.2.4 munsell_0.4.3    colorspace_1.4-0
#  [9] R6_2.2.2         rlang_0.2.1      dplyr_0.7.5      stringr_1.3.0
# [13] plyr_1.8.4       tools_3.4.3      grid_3.4.3       gtable_0.2.0
# [17] withr_2.1.2      htmltools_0.3.6  assertthat_0.2.0 lazyeval_0.2.1
# [21] yaml_2.1.19      rprojroot_1.3-2  digest_0.6.15    tibble_1.4.2
# [25] bindrcpp_0.2.2   purrr_0.2.5      glue_1.2.0       evaluate_0.10.1
# [29] rmarkdown_1.9    stringi_1.1.7    compiler_3.4.3   pillar_1.2.1
# [33] methods_3.4.3    scales_0.5.0     backports_1.1.2  pkgconfig_2.0.1
```

Other notes

- To generate this PDF, run `rmarkdown::render("handout.Rmd")` in R.
- The Latex template used to generate the PDF from R Markdown was modified from a template created by Steven Miller.
- The CFW phenotype and genotype data were downloaded from the Data Dryad repository.
- The City of Chicago food inspection data were downloaded from the Chicago Data Portal. Specifically, the data were downloaded in CSV format from here on August 3, 2018.
- The Chicago community map was downloaded from here.
- For background on the using fast food data to assess the economy, read this.