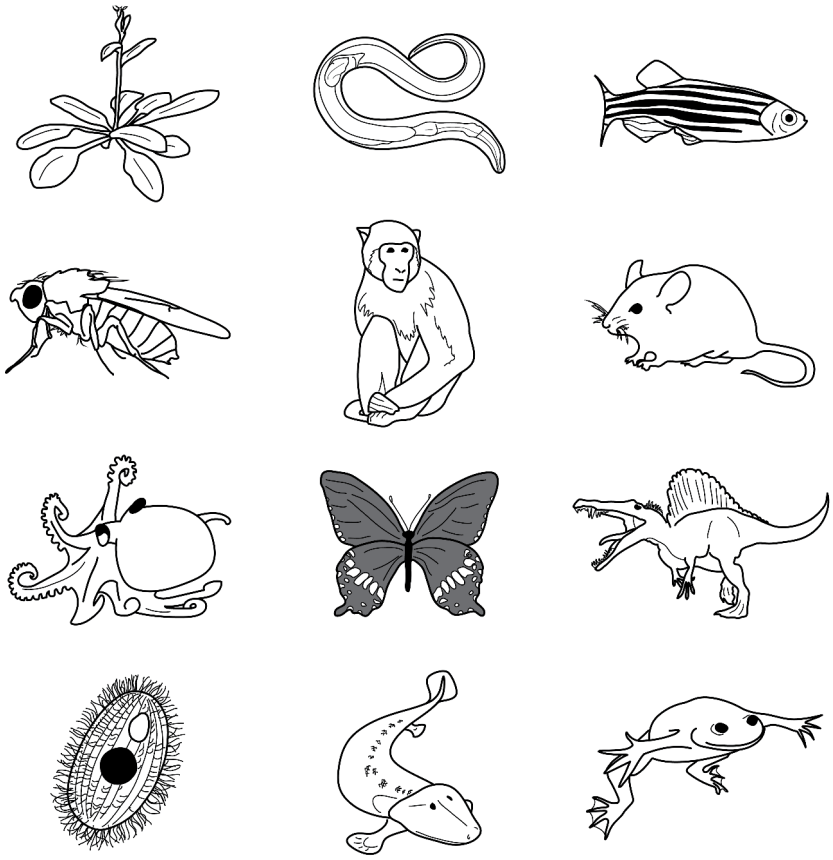
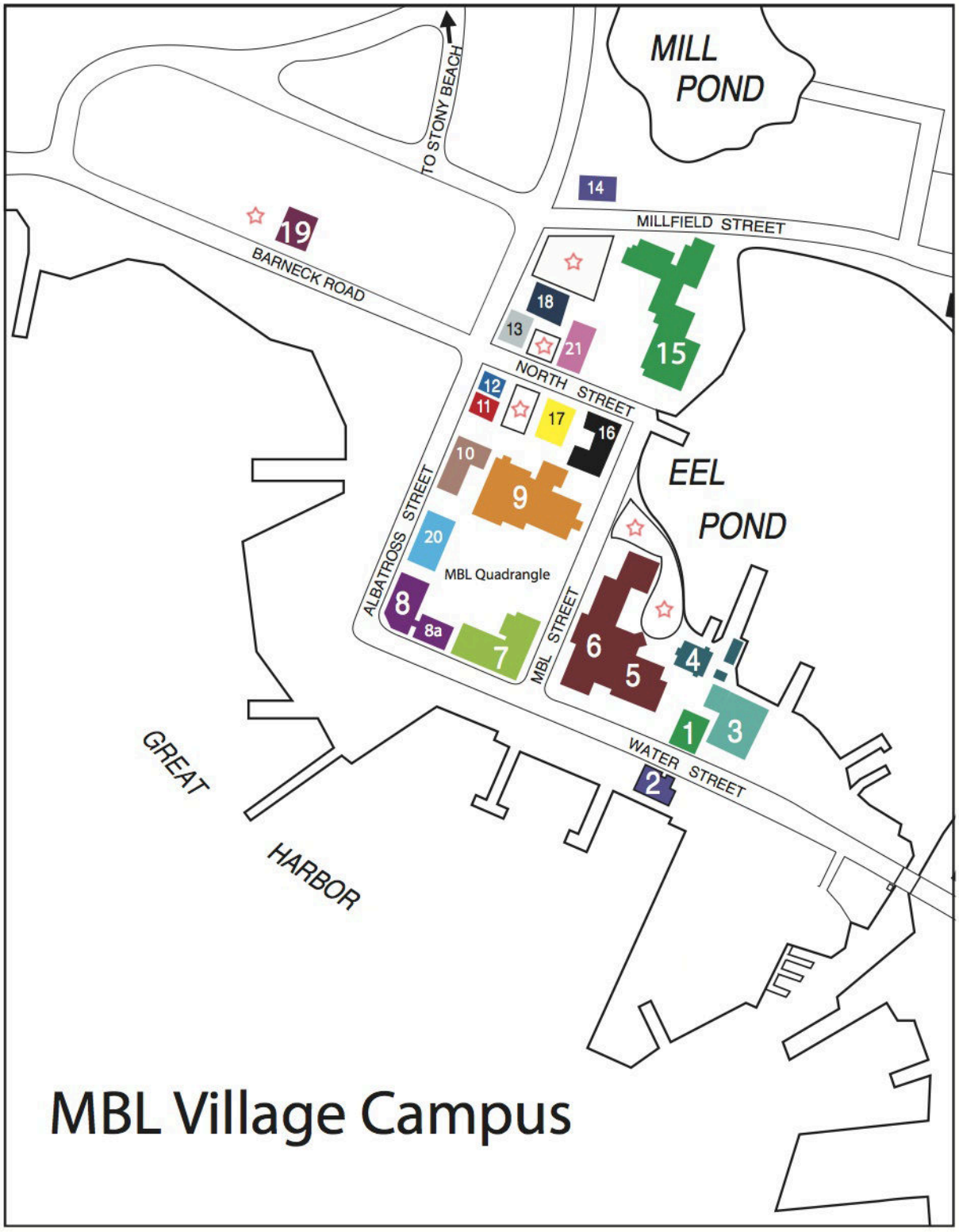
THE UNIVERSITY OF
CHICAGO BIOLOGICAL SCIENCES

BSD qBio^3 Boot Camp @ MBL



September 10–17, 2017
Marine Biological Laboratories
Woods Hole, MA



1 Candle House (Administration)

2 100 Water Street (Pierce Exhibit Center, Satellite Club, MBL Club)

3 Marine Resources Center

4 Collection Support Facility

5 Crane Wing (Labs, Shipping)

6 Lillie Laboratory (Labs, Service Shops, MBLWHOI Library)

7 Rowe Laboratory (Labs, Speck Auditorium)

8 Environmental Sciences Laboratory

8a Homestead Administration (Human Resources, Education Dept.)

9 Loeb Laboratory (Research and Teaching Labs, Lecture Rooms)

10 Brick Apartment House

11 Veeder House Dormitory

12 David House Dormitory

13 Broderick House (IT)

14 Crane House

15 Swope Center (Registration, Cafeteria, Dormitory, Meigs Room)

16 Ebert Hall Dormitory

17 Drew House Dormitory

18 15 North Street (Carpentry Shop)

19 Smith Cottage and Barneck Road Property

20 C.V. Starr Environmental Sciences Laboratory

21 11 North Street

☆ MBL Parking

Contact Information

Directors

- **Stefano Allesina**
  Professor, Ecology & Evolution
  sallesina@uchicago.edu
- **Stephanie Palmer**
  Assistant Professor, Organismal Biology and Anatomy
  sepalmer@uchicago.edu
- **Vicky Prince**
  Dean of Graduate Education/Professor, Organismal Biology and Anatomy
  vprince@uchicago.edu

Administrators

- **Diane Hall**
  Assistant Dean, OGPA
  djh8@uchicago.edu (773-383-4930)
- **Melissa Lindberg**
  Graduate Student Affairs Administrator, OGPA
  mlindber@bsd.uchicago.edu

Instructors

- **Peter Carbonetto**
  Computational scientist, Research Computing Center
  pcarbo@uchicago.edu
- **Sarah Cobey**
  Assistant Professor, Ecology & Evolution
  cobey@uchicago.edu
- **Christine Labno**
  Assistant Technical Director, Light Microscopy Core
  ccase@uchicago.edu
- **John Novembre**
  Professor, Human Genetics
  jnovembre@uchicago.edu
- **Lixing Yang**
  Assistant Professor, Ben May Department for Cancer Research
  lixingyang@uchicago.edu

Course Assistants

- **Arjun Biddanda**
  Novembre Lab, Human Genetics
  abiddanda@uchicago.edu
- **Marcos Costa Vieira**
  Cobey Lab, Ecology & Evolution
  mvieira@uchicago.edu
- **Lari DeWet**
  Vander Griend Lab, Cancer Biology
  ldewet@uchicago.edu.
- **Matthew Michalska-Smith**
  Allesina Lab, Ecology & Evolution
  mjsmith@uchicago.edu
- **Graham Smith**
  Cowan Lab, Computational Neuroscience
  grahams@uchicago.edu
- **Calvin VanOpstall**
  Vander Griend Lab, Cancer Biology
  cvanopstall@uchicago.edu
- **Frank Wen**
  Cobey Lab, Ecology & Evolution
  frankwen@uchicago.edu

BSD qBio^3 @ MBL

General Schedule

Sunday, September 10

2:45-5:00	Check-in
5:00-6:30	Dinner
6:30-7:00	Team building activities
7:15-8:30	Introduction – Allesina and Palmer

Monday, September 11

8:30-10:00	Basic comp. I
10:00-10:30	Coffee break
10:30-12:00	Basic comp. I
12:00-12:45	Lunch
1:00-2:30	Basic comp. II
2:30-2:45	Coffee break
2:45-4:30	Basic comp. II
5:00-6:30	Dinner
6:15-7:30	Welcome – Nishi & Prince
7:35-8:10	Life after PhD a) (*M.mulatta M.musculus*)
	Life after PhD b) (*S.aegyptiacus T.roseae*)
	Navigating first year a) (*D.rerio X.laevis*)
	Navigating first year b) (*C.elegans D.melanogaster*)
	Publishing primer a) (*O.bimaculoides P.polytes*)
	Publishing primer b) (*A.thaliana T.thermophila*)
8:15-8:50	Life after PhD a) (*D.rerio X.laevis*)
	Life after PhD b) (*C.elegans D.melanogaster*)
	Navigating first year a) (*O.bimaculoides P.polytes*)
	Navigating first year b) (*A.thaliana T.thermophila*)
	Publishing primer a) (*M.mulatta M.musculus*)
	Publishing primer b) (*S.aegyptiacus T.roseae*)
8:55-9:30	Life after PhD a) (*O.bimaculoides P.polytes*)
	Life after PhD b) (*A.thaliana T.thermophila*)
	Navigating first year a) (*M.mulatta M.musculus*)
	Navigating first year b) (*S.aegyptiacus T.roseae*)
	Publishing primer a) (*D.rerio X.laevis*)
	Publishing primer b) (*C.elegans D.melanogaster*)

Tuesday, September 12

8:30-10:00	Free time (*M.mulatta M.musculus O.bimaculoides P.polytes*)
	Imaging (*S.aegyptiacus T.roseae T.thermophila X.laevis*)
	Trip on the Gemma (*A.thaliana C.elegans*)
	Visit Marine Resources Center (*D.melanogaster D.rerio*)
10:00-10:30	Coffee break
10:30-12:00	Free time (*M.mulatta M.musculus O.bimaculoides P.polytes*)
	Imaging (*S.aegyptiacus T.roseae T.thermophila X.laevis*)
	Trip on the Gemma (*D.melanogaster D.rerio*)
	Visit Marine Resources Center (*A.thaliana C.elegans*)
12:00-1:00	Lunch
1:00-2:30	Data visualization (*D.rerio M.mulatta P.polytes T.thermophila*)
	Defensive programming (*D.melanogaster M.musculus O.bimaculoides S.aegyptiacus*)
	Reproducibility of data analysis (*A.thaliana C.elegans T.roseae X.laevis*)
2:30-2:45	Coffee break
2:45-4:30	Data visualization (*D.rerio M.mulatta P.polytes T.thermophila*)
	Defensive programming (*D.melanogaster M.musculus O.bimaculoides S.aegyptiacus*)
	Reproducibility of data analysis (*A.thaliana C.elegans T.roseae X.laevis*)
5:00-6:30	Dinner
7:15-8:00	MBL Lighting talks + Remsen

Wednesday, September 13

8:30-10:00	Free time (*S.aegyptiacus T.roseae T.thermophila X.laevis*)
	Imaging (*A.thaliana C.elegans D.melanogaster D.rerio*)
	Trip on the Gemma (*M.mulatta M.musculus*)
	Visit Marine Resources Center (*O.bimaculoides P.polytes*)
10:00-10:30	Coffee break
10:30-12:00	Free time (*S.aegyptiacus T.roseae T.thermophila X.laevis*)
	Imaging (*A.thaliana C.elegans D.melanogaster D.rerio*)
	Trip on the Gemma (*O.bimaculoides P.polytes*)
	Visit Marine Resources Center (*M.mulatta M.musculus*)
12:00-1:00	Lunch
1:00-2:30	Data visualization (*A.thaliana M.musculus O.bimaculoides X.laevis*)
	Defensive programming (*C.elegans M.mulatta P.polytes T.roseae*)
	Reproducibility of data analysis (*D.melanogaster D.rerio S.aegyptiacus T.thermophila*)
2:30-2:45	Coffee break
2:45-4:30	Data visualization (*A.thaliana M.musculus O.bimaculoides X.laevis*)
	Defensive programming (*C.elegans M.mulatta P.polytes T.roseae*)
	Reproducibility of data analysis (*D.melanogaster D.rerio S.aegyptiacus T.thermophila*)

5:00-6:30	Dinner
7:15-8:00	Free time

Thursday, September 14

8:30-10:00	Free time (*A.thaliana C.elegans D.melanogaster D.rerio*)
	Imaging (*M.mulatta M.musculus O.bimaculoides P.polytes*)
	Trip on the Gemma (*S.aegyptiacus T.thermophila*)
	Visit Marine Resources Center (*T.roseae X.laevis*)
10:00-10:30	Coffee break
10:30-12:00	Free time (*A.thaliana C.elegans D.melanogaster D.rerio*)
	Imaging (*M.mulatta M.musculus O.bimaculoides P.polytes*)
	Trip on the Gemma (*T.roseae X.laevis*)
	Visit Marine Resources Center (*S.aegyptiacus T.thermophila*)
12:00-1:00	Lunch
1:00-2:30	Data visualization (*C.elegans D.melanogaster S.aegyptiacus T.roseae*)
	Defensive programming (*A.thaliana D.rerio T.thermophila X.laevis*)
	Reproducibility of data analysis (*M.mulatta M.musculus O.bimaculoides P.polytes*)
2:30-2:45	Coffee break
2:45-4:30	Data visualization (*C.elegans D.melanogaster S.aegyptiacus T.roseae*)
	Defensive programming (*A.thaliana D.rerio T.thermophila X.laevis*)
	Reproducibility of data analysis (*M.mulatta M.musculus O.bimaculoides P.polytes*)
5:00-6:30	Dinner
7:15-8:00	MBL Talk – Mark Welch + Morgan

Friday, September 15

8:30-10:00	Statistics for a data rich world (*T.roseae T.thermophila X.laevis*)
	Workshop Cobey (*A.thaliana C.elegans D.rerio*)
	Workshop Novembre (*O.bimaculoides P.polytes S.aegyptiacus*)
	Workshop Yang (*D.melanogaster M.mulatta M.musculus*)
10:00-10:30	Coffee break
10:30-12:00	Statistics for a data rich world (*T.roseae T.thermophila X.laevis*)
	Workshop Cobey (*A.thaliana C.elegans D.rerio*)
	Workshop Novembre (*O.bimaculoides P.polytes S.aegyptiacus*)
	Workshop Yang (*D.melanogaster M.mulatta M.musculus*)
12:00-1:00	Lunch
1:00-2:30	Statistics for a data rich world (*D.rerio M.musculus S.aegyptiacus*)
	Workshop Cobey (*D.melanogaster O.bimaculoides T.thermophila*)
	Workshop Novembre (*C.elegans M.mulatta X.laevis*)

	Workshop Yang (*A.thaliana P.polytes T.roseae*)
2:30-2:45	Coffee break
2:45-4:30	Statistics for a data rich world (*D.rerio M.musculus S.aegyptiacus*)
	Workshop Cobey (*D.melanogaster O.bimaculoides T.thermophila*)
	Workshop Novembre (*C.elegans M.mulatta X.laevis*)
	Workshop Yang (*A.thaliana P.polytes T.roseae*)
5:00-6:30	Dinner
7:15-8:00	MBL Talk – Hanlon + Q & A

Saturday, September 16

8:30-10:00	Statistics for a data rich world (*C.elegans D.melanogaster P.polytes*)
	Workshop Cobey (*M.mulatta S.aegyptiacus T.roseae*)
	Workshop Novembre (*A.thaliana M.musculus T.thermophila*)
	Workshop Yang (*D.rerio O.bimaculoides X.laevis*)
10:00-10:30	Coffee break
10:30-12:00	Statistics for a data rich world (*C.elegans D.melanogaster P.polytes*)
	Workshop Cobey (*M.mulatta S.aegyptiacus T.roseae*)
	Workshop Novembre (*A.thaliana M.musculus T.thermophila*)
	Workshop Yang (*D.rerio O.bimaculoides X.laevis*)
12:00-1:00	Lunch
1:00-2:30	Statistics for a data rich world (*A.thaliana M.mulatta O.bimaculoides*)
	Workshop Cobey (*M.musculus P.polytes X.laevis*)
	Workshop Novembre (*D.melanogaster D.rerio T.roseae*)
	Workshop Yang (*C.elegans S.aegyptiacus T.thermophila*)
2:30-2:45	Coffee break
2:45-4:30	Statistics for a data rich world (*A.thaliana M.mulatta O.bimaculoides*)
	Workshop Cobey (*M.musculus P.polytes X.laevis*)
	Workshop Novembre (*D.melanogaster D.rerio T.roseae*)
	Workshop Yang (*C.elegans S.aegyptiacus T.thermophila*)
5:15-8:15	BBQ and Wrap up

Sunday, September 17

10:00-10:30 Departure for BOS

Tutorials

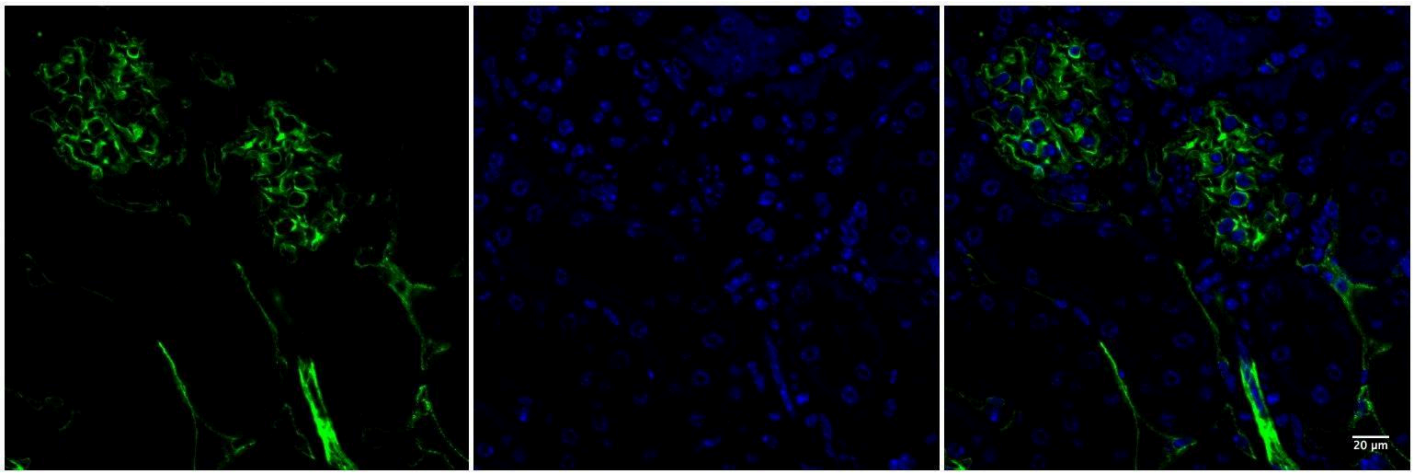
Imaging
Basic computing I
Basic computing II
Advanced computing I
Advanced computing II
Defensive programming
Data visualization
Reproducibility of data analysis
Statistics for a data-rich world

Image Processing with ImageJ Exercises

Using the information in the ImageJ Tutorials and the built-in ImageJ Command Finder tool (open ImageJ / Fiji and type L to get it to pop up), complete the following exercises. You may not be able to complete these exercises in the time you have, but do your best.

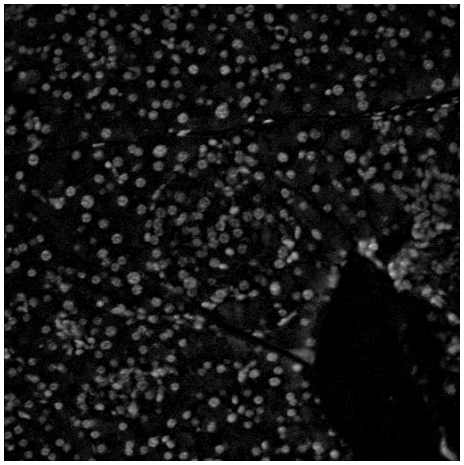
Exercise 1

Use the cd31 glomeruli.tif file to create the montage shown below. The image comes as two images grouped together in a stack, so your first job is to figure out how to separate them. The pixel size for the image is 0.25 microns. Don't forget the 20um scale bar in the lower right hand corner!



Exercise 2

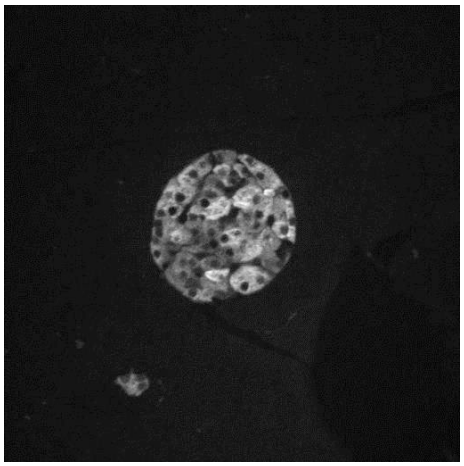
Use the pancreatic_islet.tif to answer the following questions:



a) How many cells are in the entire image? One nucleus per cell is marked by the DAPI stain in the first image.

b) What is the average nuclear area for all the cells in the image? Readout should be in um^2 NOT cm^2!

c) How would you create a data table that includes the mean gray value for every nucleus in the image? They are not all 255!



d) How many beta cells are in the image? Beta cells are represented by the insulin (green) stain in the second image. Hint: you can still count the nuclei if you find a way to restrict your count to nuclei in an islet.

Bonus Questions (if you have time)

Bonus 1: if there were multiple islets in the pancreatic_islet.tif image, how could you count the number of cells PER islet?

Bonus 2: Create a macro to automate the creation of the cd31 glomeruli montage.

Basic Computing 1 – Introduction to R

Stefano Allesina

Basic Computing 1

- **Goal:** Introduce the stastical software `R`, and show how it can be used to analyze biological data in an automated, replicable way. Showcase the `RStudio` development environment, illustrate the notion of assignment, present the main data structures available in `R`. Show how to read and write data, how to execute simple programs, and how to modify the stream of execution of a program through conditional branching and looping.
- **Audience:** Biologists with little or no background in programming.
- **Installation:** To complete the tutorial, we will need to install `R` and `RStudio`. To install `R`, follow instructions at `cran.rstudio.com`. Then install `Rstudio` following the instructions at `goo.gl/a42jYE`.

Motivation

When it comes to analyzing data, there are two competing paradigms. First, one could use point-and-click software with a graphical user interface, such as Excel, to perform calculations and draw graphs; second, one could write programs that can be run to perform the analysis of the data, the generation of tables and statistics, and the production of figures automatically.

This latter approach is to be preferred, because it allows for the automation of analysis, it requires a good documentation of the procedures, and is completely replicable.

A few motivating examples:

- You have written code to analyze your data. You receive from your collaborators a new batch of data. With simple modifications of your code, you can update your results, tables and figures automatically.
- A new student joins the laboratory. The new student can read the code and understand the analysis without the need of a lab mate showing the procedure step-by-step.
- The reviewers of your manuscript ask you to slightly alter the analysis. Rather than having to start over, you can modify a few lines of code and satisfy the reviewers.

Here we introduce `R`, which can help you write simple programs to analyze your data, perform statistical analysis, and draw beautiful figures.

What is R?

`R` is a statistical software that is completely programmable. This means that one can write a program (script) containing a series of commands for the analysis of data, and execute them automatically. This approach is especially good as it makes the analysis of data well-documented, and completely replicable.

`R` is free software: anyone can download its source code, modify it, and improve it. The `R` community of users is vast and very active. In particular, scientists have enthusiastically embraced the program, creating thousands of packages to perform specific types of analysis, and adding many new capabilities. You can find a list of official packages (which have been vetted by `R` core developers) at `goo.gl/SOSDWA`; many more are available on GitHub and other websites.

The main hurdle new users face when approaching R is that it is based on a command line interface: when you launch R, you simply open a console with the character > signaling that R is ready to accept an input. When you write a command and press Enter, the command is interpreted by R, and the result is printed immediately after the command. For example,

```
1 + 1
```

```
## [1] 2
```

A little history: R was modeled after the commercial statistical software S by Robert Gentleman and Ross Ihaka. The project was started in 1992, first released in 1994, and the first stable version appeared in 2000. Today, R is managed by the *R Core Team*.

RStudio

For this introduction, we're going to use RStudio, an Integrated Development Environment (IDE) for R. The main advantage is that the environment will look identical irrespective of your computer architecture (Linux, Windows, Mac). Also, RStudio makes writing code much easier by automatically completing commands and file names (simply type the beginning of the name and press Tab), and allowing you to easily inspect data and code.

Typically, an RStudio window contains four panels:

- **Console** This is a panel containing an instance of R. For this tutorial, we will work mainly in this panel.
- **Source code** In this panel, you can write a program, save it to a file pressing Ctrl + S and then execute it by pressing Ctrl + Shift + S.
- **Environment** This panel lists all the variables you created (more on this later); another tab shows you the history of the commands you typed.
- **Plots** This panel shows you all the plots you drew. Other tabs allow you to access the list of packages you have loaded, and the help page for commands (just type help(name_of_command) in the Console) and packages.

How to write an R program

An R program is simply a list of commands, which are executed one after the other. The commands are witten in a text file (with extension .R). When R executes the program, it will start from the beginning of the file and proceed toward the end of the file. Every time R encounters a command, it will execute it. Special commands can modify this basic flow of the program by, for example, executing a series of commands only when a condition is met, or repeating the execution of a series of commands multiple times.

Note that if you were to copy and paste (or type) the code into the **Console** you would obtain exactly the same result. Writing a program is advatageous, however, because it can be automated and shared with other researchers. Moreover, after a while you will have a large code base, so that you can recycle much of your code in several programs.

The most basic operation: assignment

The most basic operation in any programming language is the assignment. In R, assignment is marked by the operator <-. When you type a command in R, it is executed, and the output is printed in the **Console**. For example:

```
sqrt(9)
```

```
## [1] 3
```

If we want to save the result of this operation, we can assign it to a variable. For example:

```
x <- sqrt(9)
x
```

```
## [1] 3
```

What has happened? We wrote a command containing an assignment operator (`<-`). `R` has evaluated the right-hand-side of the command (`sqrt(9)`), and has stored the result (`3`) in a newly created variable called `x`. Now we can use `x` in our commands: every time the command needs to be evaluated, the program will look up which value is associated with the variable `x`, and substitute it. For example:

```
x * 2
```

```
## [1] 6
```

Types of data

`R` provides different types of data that can be used in your programs. The basic data types are:

- `logical`, taking only two possible values: `TRUE` and `FALSE`

```
v <- TRUE
class(v)
```

```
## [1] "logical"
```

- `numeric`, storing real numbers (actually, their approximations, as computers have limited memory and thus cannot store numbers like π, or even 0.2)

```
v <- 3.77
class(v)
```

```
## [1] "numeric"
```

- `integer`, storing whole numbers

```
v <- 23L # the L signals that this should be stored as integer
class(v)
```

```
## [1] "integer"
```

- `complex`, storing complex numbers (i.e., with a real and an imaginary part)

```
v <- 23 + 5i # the i marks the imaginary part
class(v)
```

```
## [1] "complex"
```

- `character`, for strings, characters and text

```
v <- 'a string' # you can use single or double quotes
class(v)
```

```
## [1] "character"
```

In `R`, the type of a variable is evaluated at runtime. This means that you can recycle the names of variables. This is very handy, but can make your programs more difficult to read and to debug (i.e., find mistakes). For example:

```r
x <- '2.3' # this is a string
x
```

```
## [1] "2.3"
```

```r
x <- 2.3 # this is numeric
x
```

```
## [1] 2.3
```

Operators and functions

Each data type supports a certain number of operators and functions. For example, numeric variables can be combined with `+` (addition), `-` (subtraction), `*` (multiplication), `/` (division), and `^` (exponentiation). A possibly unfamiliar operator is the modulo (`%%`), calculating the remainder of an integer division:

```r
5 %% 3
```

```
## [1] 2
```

meaning that `5 %/% 3` (5 integer divided by 3) is 1 with a remainder of 2. The modulo operator is useful to determine whether a number is divisible for another: if y is divisible by x, then `y %% x` is 0.

Numeric types also support many built-in functions, such as:

- `abs(x)` absolute value
- `sqrt(x)` square root
- `round(x, digits = 3)` round `x` to three decimal digits
- `cos(x)` cosinus (also supported are all the usual trigonometric functions)
- `log(x)` natural logarithm (use `log10` for base 10 logarithms)
- `exp(x)` calculating e^x

Similarly, `character` variables have their own set of functions, such as

- `toupper(x)` make uppercase
- `nchar(x)` count the number of characters in the string
- `paste(x, y, sep = "_")` concatenate strings, joining them using the separator `_`
- `strsplit(x, "_")` separate the string using the separator `_`

Calling a function meant for a certain data type on another will cause errors. If sensible, you can convert a type into another. For example:

```r
v <- "2.13"
class(v)
```

```
## [1] "character"
```

```r
# if we call v * 2, we get an error.
# to avoid it, we can convert v to numeric:
as.numeric(v) * 2
```

```
## [1] 4.26
```

If sensible, you can use the comparison operators `>` (greater), `<` (lower), `==` (equals), `!=` (differs), `>=` and `<=`, returning a logical value:

```r
2 == sqrt(4)
```

```
## [1] TRUE
```

```
2 < sqrt(4)
```

```
## [1] FALSE
```

```
2 <= sqrt(4)
```

```
## [1] TRUE
```

Similarly, you can concatenate several comparison and logical variables using `&` (and), `|` (or), and `!` (not):

```
(2 > 3) & (3 > 1)
```

```
## [1] FALSE
```

```
(2 > 3) | (3 > 1)
```

```
## [1] TRUE
```

Data structures

Besides these simple types, `R` provides structured data types, meant to collect and organize multiple values.

Vectors

The most basic data structure in `R` is the vector, which is an ordered collection of values of the same type. Vectors can be created by concatenating different values with the function `c()` (concatenate):

```
x <- c(2, 3, 5, 27, 31, 13, 17, 19)
x
```

```
## [1]  2  3  5 27 31 13 17 19
```

You can access the elements of a vector by their index: the first element is indexed at 1, the second at 2, etc.

```
x[3]
```

```
## [1] 5
```

```
x[8]
```

```
## [1] 19
```

```
x[9] # what if the element does not exist?
```

```
## [1] NA
```

`NA` stands for Not Available. Other special values are `NaN` (Not a Number, e.g., `0/0`), `Inf` (Infinity, e.g., `1/0`), and `NULL` (variable not set). You can test for special values using `is.na(x)`, `is.infinite(x)`, etc.

Note that in `R` a single number (string, logical) is a vector of length 1 by default. That's why if you type `3` in the console you see `[1] 3` in the output.

You can extract several elements at once (i.e., create another vector), using the colon (`:`) command, or by concatenating the indices:

```
x[1:3]
```

```
## [1] 2 3 5
```

```
x[4:7]
```

```
## [1] 27 31 13 17
```

```
x[c(1,3,5)]
```

```
## [1]  2  5 31
```

You can also use a vector of logical variables to extract values from vectors. For example, suppose we have two vectors:

```
sex <- c("M", "M", "F", "M", "F") # sex of Drosophila
weight <- c(0.230, 0.281, 0.228, 0.260, 0.231) # weigth in mg
```

and that we want to extract only the weights for the males.

```
sex == "M"
```

```
## [1]  TRUE  TRUE FALSE  TRUE FALSE
```

returns a vector of logical values, which we can use to subset the data:

```
weight[sex == "M"]
```

```
## [1] 0.230 0.281 0.260
```

Given that `R` was born for statistics, there are many statistical functions you can perform on vectors:

```
length(x)
```

```
## [1] 8
```

```
min(x)
```

```
## [1] 2
```

```
max(x)
```

```
## [1] 31
```

```
sum(x) # sum all elements
```

```
## [1] 117
```

```
prod(x) # multiply all elements
```

```
## [1] 105436890
```

```
median(x) # median value
```

```
## [1] 15
```

```
mean(x) # arithmetic mean
```

```
## [1] 14.625
```

```
var(x) # unbiased sample variance
```

```
## [1] 119.4107
```

```
mean(x ^ 2) - mean(x) ^ 2 # population variance
```

```
## [1] 104.4844
```

```
summary(x) # print a summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    4.50   15.00   14.62   21.00   31.00
```

You can generate vectors of sequential numbers using the colon command:

```
x <- 1:10
x
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

For more complex sequences, use `seq`:

```
seq(from = 1, to = 5, by = 0.5)
```

```
## [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
```

To repeat a value or a sequence several times, use `rep`:

```
rep("abc", 3)
```

```
## [1] "abc" "abc" "abc"
```

```
rep(c(1,2,3), 3)
```

```
## [1] 1 2 3 1 2 3 1 2 3
```

Exercise:

- Create a vector containing all the even numbers between 2 and 100 (inclusive) and store it in variable `z`.
- Extract all the elements of `z` that are divisible by 12. How many elements match this criterion?
- What is the sum of all the elements of `z`?
- Is it equal to $51 \cdot 50$?
- What is the product of elements 5, 10 and 15 of `z`?
- Does `seq(2, 100, by = 2)` produce the same vector as `(1:50) * 2`?
- What happens if you type `z ^ 2`?

Matrices

A matrix is a two-dimensional table of values. In case of numeric values, you can perform the usual operations on matrices (product, inverse, decomposition, etc.):

```
A <- matrix(c(1, 2, 3, 4), 2, 2) # values, nrows, ncols
A
```

```
##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4
```

```
A %*% A # matrix product
```

```
##      [,1] [,2]
## [1,]    7   15
## [2,]   10   22
```

```
solve(A) # matrix inverse
```

```
##      [,1] [,2]
## [1,]   -2  1.5
## [2,]    1 -0.5
```

```
A %*% solve(A) # this should return the identity matrix
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
```

```
B <- matrix(1, 3, 2) # you can fill the whole matrix with a single number (1)
B
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    1
## [3,]    1    1
```

```
B %*% t(B) # transpose
```

```
##      [,1] [,2] [,3]
## [1,]    2    2    2
## [2,]    2    2    2
## [3,]    2    2    2
```

```
Z <- matrix(1:9, 3, 3) # by default, matrices are filled by column
Z
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

To determine the dimensions of a matrix, use `dim`:

```
dim(B)
```

```
## [1] 3 2
```

```
dim(B)[1]
```

```
## [1] 3
```

```
nrow(B)
```

```
## [1] 3
```

```
dim(B)[2]
```

```
## [1] 2
```

```
ncol(B)
```

```
## [1] 2
```

Use indices to access a particular row/column of a matrix:

```
Z
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

```
Z[1, ] # first row
```

```
## [1] 1 4 7
```

```
Z[, 2] # second column
```

```
## [1] 4 5 6
```

```
Z [1:2, 2:3] # submatrix with coefficients in first two rows, and second and third column
```

```
##      [,1] [,2]
## [1,]    4    7
## [2,]    5    8
```

```
Z[c(1,3), c(1,3)] # indexing non-adjacent rows/columns
```

```
##      [,1] [,2]
## [1,]    1    7
## [2,]    3    9
```

Some operations use all the elements of the matrix:

```
sum(Z)
```

```
## [1] 45
```

```
mean(Z)
```

```
## [1] 5
```

Arrays

If you need tables with more than two dimensions, use arrays:

```
M <- array(1:24, c(4, 3, 2))
M
```

```
## , , 1
##
##      [,1] [,2] [,3]
## [1,]    1    5    9
## [2,]    2    6   10
## [3,]    3    7   11
## [4,]    4    8   12
##
## , , 2
##
##      [,1] [,2] [,3]
## [1,]   13   17   21
## [2,]   14   18   22
## [3,]   15   19   23
## [4,]   16   20   24
```

You can still determine the dimensions using:

```
dim(M)
```

```
## [1] 4 3 2
```

and access the elements as done for matrices. One thing you should be paying attention to: R drops dimensions that are not needed. So, if you access a "slice" of a 3-dimensional array:

```
M[,,1]
```

```
##      [,1] [,2] [,3]
## [1,]    1    5    9
```

```
## [2,]    2    6   10
## [3,]    3    7   11
## [4,]    4    8   12
```

you obtain a matrix:

```
dim(M[,,1])
```

```
## [1] 4 3
```

This can be problematic, for example, when your code expects an array and `R` turns your data into a matrix (or you expect a matrix but find a vector). To avoid this behavior, add `drop = FALSE` when subsetting:

```
dim(M[,,1, drop = FALSE])
```

```
## [1] 4 3 1
```

Lists

Vectors are good if each element is of the same type (e.g., numbers, strings). Lists are used when we want to store elements of different types, or more complex objects (e.g., vectors, matrices, even lists of lists). Each element of the list can be referenced either by its index, or by a label:

```
mylist <- list(Names = c("a", "b", "c", "d"), Values = c(1, 2, 3))
mylist
```

```
## $Names
## [1] "a" "b" "c" "d"
##
## $Values
## [1] 1 2 3
```

```
mylist[[1]] # access first element using index
```

```
## [1] "a" "b" "c" "d"
```

```
mylist[[2]] # access second element by index
```

```
## [1] 1 2 3
```

```
mylist$Names # access second element by label
```

```
## [1] "a" "b" "c" "d"
```

```
mylist[["Names"]] # another way to access by label
```

```
## [1] "a" "b" "c" "d"
```

```
mylist[["Values"]][3]  # access third element in second vector
```

```
## [1] 3
```

Data frames

Data frames contain data organized like in a spreadsheet. The columns (typically representing different measurements) can be of different types (e.g., a column could be the date of measurement, another the weight of the individual, or the volume of the cell, or the treatment of the sample), while the rows typically represent different samples.

When you read a spreadsheet file in R, it is automatically stored as a data frame. The difference between a matrix and a data frame is that in a matrix all the values are of the same type (e.g., all numeric), while in a data frame each column can be of a different type.

Because typing a data frame by hand would be tedious, let's use a data set that is already available in R:

```
data(trees) # Girth, height and volume of cherry trees
str(trees) # structure of data frame
```

```
## 'data.frame':    31 obs. of  3 variables:
##  $ Girth : num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
##  $ Height: num  70 65 63 72 81 83 66 75 80 75 ...
##  $ Volume: num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
```

```
ncol(trees)
```

```
## [1] 3
```

```
nrow(trees)
```

```
## [1] 31
```

```
head(trees) # print the first few rows
```

```
##   Girth Height Volume
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## 6  10.8     83   19.7
```

```
trees$Girth # select column by name
```

```
##  [1]  8.3  8.6  8.8 10.5 10.7 10.8 11.0 11.0 11.1 11.2 11.3 11.4 11.4 11.7
## [15] 12.0 12.9 12.9 13.3 13.7 13.8 14.0 14.2 14.5 16.0 16.3 17.3 17.5 17.9
## [29] 18.0 18.0 20.6
```

```
trees$Height[1:5] # select column by name; return first five elements
```

```
## [1] 70 65 63 72 81
```

```
trees[1:3, ] #select rows 1 through 3
```

```
##   Girth Height Volume
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
```

```
trees[1:3, ]$Volume # select rows 1 through 3; return column Volume
```

```
## [1] 10.3 10.3 10.2
```

```
trees <- rbind(trees, c(13.25, 76, 30.17)) # add a row
trees_double <- cbind(trees, trees) # combine columns
colnames(trees) <- c("Circumference", "Height", "Volume") # change column names
```

Exercise:

- What is the average height of the cherry trees?
- What is the average girth of those that are more than 75 ft tall?
- What is the maximum height of trees with a volume between 15 and 35 ft^3?

Reading and writing data

In most cases, you will not generate your data in `R`, but import it from a file. By far, the best option is to have your data in a comma separated value text file or in a tab separated file. Then, you can use the function `read.csv` (or `read.table`) to import your data. The syntax of the functions is as follows:

```
read.csv("MyFile.csv") # read the file MyFile.csv
read.csv("MyFile.csv", header = TRUE) # The file has a header.
read.csv("MyFile.csv", sep = ';') # Specify the column separator.
read.csv("MyFile.csv", skip = 5) # Skip the first 5 lines.
```

Note that columns containing strings are typically converted to *factors* (categorical values, useful when performing regressions). To avoid this behavior, you can specify `stringsAsFactors = FALSE` when calling the function.

Similarly, you can save your data frames using `write.table` or `write.csv`. Suppose you want to save the data frame `MyDF`:

```
write.csv(MyDF, "MyFile.csv")
write.csv(MyDF, "MyFile.csv", append = TRUE) # Append to the end of the file.
write.csv(MyDF, "MyFile.csv", row.names = TRUE) # Include the row names.
write.csv(MyDF, "MyFile.csv", col.names = FALSE) # Do not include column names.
```

Let's look at an example: Read a file containing data on the 6th chromosome for a number of Europeans (Data adapted from Stanford HGDP SNP Genotyping Data by John Novembre):

```
ch6 <- read.table("../data/H938_Euro_chr6.geno", header = TRUE)
```

where `header = TRUE` means that we want to take the first line to be a header containing the column names. How big is this table?

```
dim(ch6)
```

```
## [1] 43141     7
```

we have 7 columns, but more than 40k rows! Let's see the first few:

```
head(ch6)
```

```
##   CHR        SNP A1 A2 nA1A1 nA1A2 nA2A2
## 1   6  rs4959515  A  G     0    17   107
## 2   6   rs719065  A  G     0    26    98
## 3   6  rs6596790  C  T     0     4   119
## 4   6  rs6596796  A  G     0    22   102
## 5   6  rs1535053  G  A     5    39    80
## 6   6 rs12660307  C  T     0     3   121
```

and the last few:

```
tail(ch6)
```

```
##       CHR        SNP A1 A2 nA1A1 nA1A2 nA2A2
## 43136   6 rs10946282  C  T     0    16   108
## 43137   6  rs3734763  C  T    19    56    48
## 43138   6   rs960744  T  C    32    60    32
## 43139   6  rs4428484  A  G     1    11   112
## 43140   6  rs7775031  T  C    26    56    42
## 43141   6 rs12213906  C  T     1    11   112
```

The data contains the number of homozygotes (`nA1A1`, `nA2A2`) and heterozygotes (`nA1A2`), for 43,141 single nucleotide polymorphisms (SNPs) obtained by sequencing European individuals:

- `CHR` The chromosome (6 in this case)
- `SNP` The identifier of the Single Nucleotide Polymorphism
- `A1` One of the alleles
- `A2` The other allele
- `nA1A1` The number of individuals with the particular combination of alleles.

Exercise:

- How many individuals were sampled? Find the maximum of the sum `nA1A1 + nA1A2 + nA2A2`. Note: you can access the columns by index (e.g., `ch6[,5]`), or by name (e.g., `ch6$nA1A1`, or also `ch6[,"nA1A1"]`).
- Try using the function `rowSums` to obtain the same result.
- For how many SNPs do we have that all sampled individuals are homozygotes (i.e., all `A1A1` or all `A2A2`)?
- For how many SNPs, are more than 99% of the sampled individuals homozygous?

Conditional branching

Now we turn to writing actual programs in the **Source code** panel. To start a new `R` program, press `Ctrl + Shift + N`. This will open an `Untitled` script. Save the script by pressing `Ctrl + S`: save it as `conditional.R` in the directory `basic_computing_1/code/`. To make sure you're working in the directory where the script is contained, on the menu on the top choose `Session -> Set Working Directory -> To Source File Location`.

Now type the following script:

```
print("Hello world!")
x <- 4
print(x)
```

and execute the script by pressing `Ctrl + Shift + S`. You should see `Hello World!` and `4` printed in your console.

As you saw in this simple example, when `R` executes the program, it starts from the top and proceeds toward the end of the file. Every time it encounters a command (for example, `print(x)`, printing the value of `x` into the console), it executes it.

When we want a certain block of code to be executed only when a certain condition is met, we can write a conditional branching point. The syntax is as follows:

```
if (condition is met){
  # Execute this block of code
} else {
  # Execute this other block of code
}
```

For example, add these lines to the script `conditional.R`, and run it again:

```
print("Hello world!")
x <- 4
print(x)
if (x %% 2 == 0){
  my_message <- paste(x, "is even")
} else {
```

```
  my_message <- paste(x, "is odd")
}
print(my_message)
```

We have created a conditional branching point, so that the value of `my_message` changes depending on whether `x` is even (and thus the remainder of the integer division by 2 is 0), or odd. Change the line `x <- 4` to `x <- 131` and run it again.

Exercise: What does this do?

```
x <- 36
if (x > 20){
  x <- sqrt(x)
} else {
  x <- x ^ 2
}
if (x > 7) {
  print(x)
} else if (x %% 2 == 1){
  print(x + 1)
}
```

Looping

Another way to change the flow of the program is to write a loop. A loop is simply a series of commands that are repeated a number of times. For example, you want to run the same analysis on different data sets that you collected; you want to plot the results contained in a set of files; you want to test your simulation over a number of parameter sets; etc.

`R` provides you with two ways to loop over blocks of commands: the `for` loop, and the `while` loop. Let's start with the `for` loop, which is used to iterate over a vector (or a list): for each value of the vector, a series of commands will be run, as shown by the following example, which you can type in a new script called `forloop.R`.

```
myvec <- 1:10 # vector with numbers from 1 to 10

for (i in myvec) {
  a <- i ^ 2
  print(a)
}
```

In the code above, the variable `i` takes the value of each element of `myvec` in sequence. Inside the block defined by the `for` loop, you can use the variable `i` to perform operations.

The anatomy of the `for` statement:

```
for (variable in list_or_vector) {
  execute these commands
} # automatically moves to the next value
```

For loops are used when you know that you want to perform the analysis using a given set of values (e.g., run over all files of a directory, all samples in your data, all sequences of a fasta file, etc.).

The `while` loop is used when the commands need to be repeated while a certain condition is true, as shown by the following example, which you can type in a script called `whileloop.R`:

```
i <- 1

while (i <= 10) {
  a <- i ^ 2
  print(a)
  i <- i + 1
}
```

The script performs exactly the same operations we wrote for the `for` loop above. Note that you need to update the value of `i`, (using `i <- i + 1`), otherwise the loop will run forever (infinite loop—to terminate click on the stop button in the top-right corner of the console). The anatomy of the `while` statement:

```
while (condition is met) {
  execute these commands
} # beware of infinite loops: remember to update the condition!
```

You can break a loop using the command `break`. For example:

```
i <- 1

while (i <= 10) {
  if (i > 5) {
    break
  }
  a <- i ^ 2
  print(a)
  i <- i + 1
}
```

Exercise: What does this do? Try to guess what each loop does, and then create and run a script to confirm your intuition.

```
z <- seq(1, 1000, by = 3)
for (k in z) {
  if (k %% 4 == 0) {
     print(k)
  }
}
```

```
z <- readline(prompt = "Enter a number: ")
z <- as.numeric(z)
isthisspecial <- TRUE
i <- 2
while (i < z) {
  if (z %% i == 0) {
     isthisspecial <- FALSE
     break
  }
  i <- i + 1
}
if (isthisspecial == TRUE) {
  print(z)
}
```

Useful Functions

We conclude with a list of useful functions that will help you write your programs:

- `range(x)`: minimum and maximum of a vector `x`
- `sort(x)`: sort a vector `x`
- `unique(x)`: remove duplicate entries from vector `x`
- `which(x == a)`: returns a vector of the indices of `x` having value `a`
- `list.files("path_to_directory")`: list the files in a directory (current directory if not specified)
- `table(x)` build a table of frequencies

Exercises: What does this code do? For each snippet of code, first try to guess what will happen. Then, write a script and run it to confirm your intuition.

```
v <- c(1,3,5,5,3,1,2,4,6,4,2)
v <- sort(unique(v))
for (i in v){
  if (i > 2){
    print(i)
  }
  if (i > 4){
    break
  }
}
```

```
x <- 1:100
x <- x[which(x %% 7 == 0)]
```

```
my_files <- sort(list.files("../data/Saavedra2013/", full.names = TRUE))
for (f in my_files){
  M <- read.table(f)
  print(paste("The file", basename(f), "contains a matrix with", nrow(M),
  "rows and ", ncol(M), "columns. There are", sum(M == 1),
  "coefficients that are 1 and", sum(M == 0), "that are 0."))
}
```

```
my_amount <- 10
while (my_amount > 0){
  my_color <- NA
  while(is.na(my_color)){
    tmp <- readline(prompt="Do you want to bet on black or red? ")
    tmp <- tolower(tmp)
    if (tmp == "black") my_color <- "black"
    if (tmp == "red") my_color <- "red"
    if (is.na(my_color)) print("Please enter either red or black")
  }
  my_bet <- NA
  while(is.na(my_bet)){
    tmp <- readline(prompt="How much do you want to bet? ")
    tmp <- as.numeric(tmp)
    if (is.numeric(tmp) == FALSE){
      print("Please enter a number")
    } else {
      if (tmp > my_amount){
        print("You don't have enough money!")
      } else {
```

```
        my_bet <- tmp
        my_amount <- my_amount - tmp
      }
    }
  }
  lady_luck <- sample(c("red", "black"), 1)
  if (lady_luck == my_color){
    my_amount <- my_amount + 2 * my_bet
    print(paste("You won!! Now you have", my_amount, "gold dobloons"))
  } else {
    print(paste("You lost!! Now you have", my_amount, "gold dobloons"))
  }
}
print("I told you this was not a good idea...")
```

Programming Challenge

Instructions

You will work with your own group to solve the following exercise. When you have found the solution, go to `https://stefanoallesina.github.io/BSD-QBio3/` and follow the link `Submit solution to challenge 1` to submit your answer (alternatively, you can go directly to goo.gl/forms/dDJKvF0dOi7KUDqp1. At the end of the boot camp, the group with the highest number of correct solutions will be declared the winner. If you need extra time, you can work with your group during free time or in the breaks in the schedule.

Nobel nominations

The file `../data/nobel_nominations.csv` contains the nominations to the Nobel prize from 1901 to 1964. There are three columns (the file has no header): a) the field (e.g. `Phy` for physics), b) the year of the nomination, c) the id and name of the nominee.

1. Take Chemistry (`Che`). Who received most nominations?
2. Find all the researchers who received nominations in more than one field.
3. Take Physics (`Phy`). Which year had the largest number of nominees?
4. What is the average number of nominees for each field? Calculate the average number of nominee for each field across years.

Hints

- You will need to subset the data. To make operations clearer, you can give names to the columns. For example, suppose you stored the data in in a data frame called `nobel`. Then `colnames(nobel) <- c("Field", "Year", "Nominee")` will do the trick.
- The simplest way to obtain a count from a vector is to use the command `table`. The command `sort(table(my_vector))` produces a table of the occurrences in `my_vector` sorted from few to many counts.
- You can build a table using more than one vector. For example, store the Nobel nominations in the data frame `nobel`, and name the columns as suggested above. The command `head(table(nobel$Nominee, nobel$Field))` will build a table with the number of nominations in each field (column) for each nominee (rows).

Basic Computing 2 – Packages, Functions, Documenting code

Stefano Allesina

Basic Computing 2

- **Goal:** Show how to install, load and use the many freely available `R` packages. Illustrate how to write user-defined functions and how to organize code. Showcase basic plotting functions. Introduce the package `knitr` for writing beautiful reports.
- **Audience:** Biologists with basic knowledge of `R`.
- **Installation:** To produce well-documented code, we need to instal the package `knitr`. We will also use the package `MASS` for statistics.

Packages

`R` is the most popular statistical computing software among biologists due to its highly specialized packages, often written from biologists for biologists. You can contribute a package too! The `RStudio` support (`goo.gl/harVqF`) provides guidance on how to start developing `R` packages and Hadley Wickham's free online book (`r-pkgs.had.co.nz`) will make you a pro.

You can find highly specialized packages to address your research questions. Here are some suggestions for finding an appropriate package. The Comprehensive R Archive Network (CRAN) offers several ways to find specific packages for your task. You can either browse packages (`goo.gl/7oVyKC`) and their short description or select a scientific field of interest (`goo.gl/OWdIcu`) to browse through a compilation of packages related to each discipline.

From within your `R` terminal or `RStudio` you can also call the function `RSiteSearch("KEYWORD")`, which submits a search query to the website `search.r-project.org`. The website `rseek.org` casts an even wider net, as it not only includes package names and their documentation but also blogs and mailing lists related to `R`. If your research interests relate to high-throughput genomic data, you should have a look the packages provided by Bioconductor (`goo.gl/7dwQlq`).

Installing a package

To install a package type

```
install.packages("name_of_package")
```

in the Console, or choose the panel **Packages** and then click on *Install* in `RStudio`.

Loading a package

To load a package type

```
library(name_of_package)
```

or call the command into your script. If you want your script to automatically install a package in case it's missing, use this boilerplate:

```
if (!require(needed_package, character.only = TRUE, quietly = TRUE)) {
    install.packages(needed_package)
    library(needed_package, character.only = TRUE)
}
```

Example

For example, say we want to access the dataset `bacteria`, which reports the incidence of *H. influenzae* in Australian children. The dataset is contained in the package `MASS`.

First, we need to load the package:

```
library(MASS)
```

Now we can load the data:

```
data(bacteria)
bacteria[1:3,]
```

```
##   y ap hilo week  ID     trt
## 1 y  p   hi    0 X01 placebo
## 2 y  p   hi    2 X01 placebo
## 3 y  p   hi    4 X01 placebo
```

Do shorter titles lead to more citations?

To keep learning about `R`, we study a simple problem: do papers with shorter titles have more citations? This is what claimed by Letchford *et al.*, who in 2015 analyzed 140,000 papers (dx.doi.org/10.1098/rsos.150266) finding that shorter titles correlated with a larger number of citations.

In the `data/citations` folder, you find information on all the articles published between 2004 and 2013 by three top disciplinary journals (*Nature Neuroscience*, *Nature Genetics*, and *Ecology Letters*), which we are going to use to explore the robustness of these findings.

We start by reading the data in. This is a simple `csv` file, so that we can use

```
papers <- read.csv("../data/citations/nature_neuroscience.csv")
```

to load the data. However, running `str(papers)` shows that all the columns containing text have been automatically converted to `Factor` (categorical values, which is good when performing regressions). Because we want to manipulate these columns (for example, count how many characters are in a title), we want to avoid this automatic behavior. We can accomplish this by calling the function `read.csv` with an extra argument:

```
papers <- read.csv("../data/citations/nature_neuroscience.csv", stringsAsFactors = FALSE)
```

Next, we want to take a peek at the data. How large is it?

```
dim(papers)
```

```
## [1] 2000    7
```

Let's see the first few rows:

```
head(papers, 3)
```

```
##            Authors                                    Title Year
## 1 Logothetis, N.K.                Francis crick 1916-2004. 2004
## 2       Narain, C. Object-specific unconscious processing. 2005
## 3       Narain, C.                       Going down BOLDly. 2006
##           Source.title Page.start Page.end Cited.by
## 1  Nature neuroscience       1027     1028        0
## 2 Nature neuroscience.       1288       NA        0
## 3  Nature neuroscience        474       NA        0
```

Now, we want to test whether papers with longer titles do accrue fewer (or more) citations than those with shorter titles. The first step is therefore to add another column to the data, containing the length of the title for each paper:

```
papers$TitleLength <- nchar(papers$Title)
```

Basic statistics in R

In the original paper, Letchford *et al.* used rank-correlation: rank all the papers according to their title length and the number of citations. If the Kendall's Tau (rank correlation) is positive, then longer titles are associated with more citations; if Tau is negative, longer titles are associated with fewer citations. In R you can compute rank correlation using:

```
kendall_cor <- cor(papers$TitleLength, papers$Cited.by, method = "kendall")
kendall_cor
```

```
## [1] 0.04528715
```

To perform a significance test, use

```
cor.test(papers$TitleLength, papers$Cited.by, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  papers$TitleLength and papers$Cited.by
## z = 3.0023, p-value = 0.002679
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## 0.04528715
```

showing that the correlation between the ranks is positive and significant. We have found the opposite effect than Letchford *et al.*—longer titles are associated with **more** citations!

Now we are going to examine the data in a different way, to test whether these results are robust.

Basic plotting in R

To plot the title length vs. number of citations, we need to learn about plotting in R. To produce a simple scatterplot using the base functions for plotting, simply type:

```
plot(papers$TitleLength, papers$Cited.by)
```

It is hard to detect any trend in this plot, as there are a few papers with many more citations than the rest. We can transform the data by plotting on the y-axis the log10 of citations + 1 (so that papers with zero citations do not cause errors):

```
plot(papers$TitleLength, log10(papers$Cited.by + 1))
```

Again, it's hard to see any trend in here. Maybe we should plot the best fitting line and overlay it on top of the graph. To do so, we first need to learn about regressions in R.

Regressions in R

R was born for statistics — the fact that it's very easy to fit a linear regression is not surprising! To build a linear model, simply write

```
# model y = a + bx + error
my_model <- lm(y ~ x)
```

Because it's more convenient to call the code in this way, let's add a new column to the data frame with the log of citations:

```
papers$LogCits <- log10(papers$Cited.by + 1)
```

And perform a linear regression:

```
model_cits <- lm(papers$LogCits ~ papers$TitleLength)
# This is the best fitting line
model_cits
```

```
##
## Call:
## lm(formula = papers$LogCits ~ papers$TitleLength)
##
## Coefficients:
##        (Intercept)  papers$TitleLength
##           1.358195            0.006193
```

```
# This is a summary of all the statistics
summary(model_cits)
```

```
##
## Call:
## lm(formula = papers$LogCits ~ papers$TitleLength)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68022 -0.25261  0.02803  0.29188  1.82838
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.3581953  0.0445997   30.45   <2e-16 ***
## papers$TitleLength 0.0061927  0.0005339   11.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4522 on 1998 degrees of freedom
## Multiple R-squared:  0.06309,    Adjusted R-squared:  0.06263
## F-statistic: 134.6 on 1 and 1998 DF,  p-value: < 2.2e-16
```

And plotting

```r
# plot the points
plot(papers$TitleLength, log10(papers$Cited.by + 1))
# add the best fitting line
abline(model_cits, col = "red")
```

Again, we find a positive trend. One thing to consider, is that in the database we have papers spanning a decade. Naturally, older papers have had more time to accrue citations. In our models, we would like to control for this effect. First, let's plot the distribution of citations for a few years. To produce an histogram in R, use

```r
hist(papers$LogCits)
# increase the number of breaks
hist(papers$LogCits, breaks = 15)
```

Alternatively, estimate the density using

```r
plot(density(papers$LogCits))
```

Let's plot the density for years 2004, 2009, 2013:

```r
# plot the density for the older papers:
plot(density(papers$LogCits[papers$Year == 2004]))
lines(density(papers$LogCits[papers$Year == 2009]), col = "red")
lines(density(papers$LogCits[papers$Year == 2013]), col = "blue")
```

As expected, younger papers have fewer citations. We can account for this fact in our regression model:

```r
model_year_length <- lm(papers$LogCits ~ as.factor(papers$Year) + papers$TitleLength)
summary(model_year_length)
```

```
##
## Call:
## lm(formula = papers$LogCits ~ as.factor(papers$Year) + papers$TitleLength)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88880 -0.21178  0.01232  0.25186  1.51362
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.6953101  0.0499837  33.917  < 2e-16 ***
## as.factor(papers$Year)2005 -0.0277435  0.0424764  -0.653 0.513735
## as.factor(papers$Year)2006 -0.1137399  0.0431189  -2.638 0.008409 **
## as.factor(papers$Year)2007 -0.1603601  0.0435023  -3.686 0.000234 ***
## as.factor(papers$Year)2008 -0.1630806  0.0442859  -3.682 0.000237 ***
## as.factor(papers$Year)2009 -0.2373747  0.0430316  -5.516 3.91e-08 ***
## as.factor(papers$Year)2010 -0.2824792  0.0433906  -6.510 9.48e-11 ***
## as.factor(papers$Year)2011 -0.4927523  0.0425956 -11.568  < 2e-16 ***
## as.factor(papers$Year)2012 -0.6278381  0.0423156 -14.837  < 2e-16 ***
## as.factor(papers$Year)2013 -0.6539580  0.0419966 -15.572  < 2e-16 ***
## papers$TitleLength          0.0056728  0.0004645  12.213  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3912 on 1989 degrees of freedom
## Multiple R-squared:  0.302,  Adjusted R-squared:  0.2985
## F-statistic: 86.06 on 10 and 1989 DF,  p-value: < 2.2e-16
```

Using `as.factor(papers$Year)` we have fitted a model in which each year has a different baseline, and the title length influences this baseline. Again, we find that longer titles are associated with more citations.

Random numbers

As a reminder, the Kendall's τ takes as input two rankings x and y, both of length n. It calculates the number of "concordant pairs", in which if $x_i > x_j$ then $y_i > y_j$ and "discordant pairs". Then,

$$\tau = \frac{\text{num. concordant} - \text{num. discordant}}{\frac{n(n-1)}{2}}$$

If x and y were completely independent, we would expect τ to be distributed with a mean of 0. The variance of the null distribution of τ (and hence the p-value calculated above) depends on the data, and is typically approximated as a normal distribution. If you want to have a stronger result, you can use randomizations to approximate the p-value. Simply, compute τ for the actual data, and for many "fake" datasets obtained randomizing the data. Your p-value is well approximated by the proportion of τ values for the randomized sets that exceed the τ value for the actual data.

To perform this randomization, or any simulation, we typically need to draw random numbers. `R` has functions to sample random numbers from very many different statistical distributions. For example:

```
runif(5) # sample 5 numbers from the uniform distribution between 0 and 1
```

```
## [1] 0.5463838 0.1432251 0.4543838 0.7623244 0.1596632
```

```
runif(5, min = 1, max = 9) # set the limits of the uniform distribution
```

```
## [1] 5.369980 8.261499 5.501856 3.226635 1.194189
```

```
rnorm(3) # three values from standard normal
```

```
## [1]  0.5354295  2.3794487 -0.9501569
```

```
rnorm(3, mean = 5, sd = 4) # specify mean and standard deviation
```

```
## [1] 0.7711573 6.3066129 4.9966275
```

To sample from a set of values, use `sample`:

```
v <- c("a", "b", "c", "d")
sample(v, 2) # without replacement
```

```
## [1] "a" "b"
```

```
sample(v, 6, replace = TRUE) # with replacement
```

```
## [1] "d" "b" "b" "d" "c" "b"
```

```
sample(v) # simply shuffle the elements
```

```
## [1] "a" "b" "c" "d"
```

Let's try to write a randomization to calculate p-value associated with the τ observed for year 2006.

```
# first, we subset the data
papers_year <- papers[papers$Year == 2006, ] # get all rows matching the year
# compute original tau
tau_original <- cor(papers_year$TitleLength, papers_year$Cited.by, method = "kendall")
tau_original
```

```
## [1] 0.1140872
```

Now we want to calculate it on the "fake" data sets. To have confidence in the first two decimal digits, we should perform about ten thousand randomizations. This and similar randomization techniques are known as "bootstrapping".

```
num_randomizations <- 10000
pvalue <- 0 # initialize at 0
for (i in 1:num_randomizations){
  # calculate cor on shuffled data
  tau_shuffle <- cor(papers_year$TitleLength,
                     sample(papers_year$Cited.by), # scramble the citations at random
                     method = "kendall")
  if (tau_shuffle >= tau_original){
    pvalue <- pvalue + 1
  }
}
# calculate proportion
pvalue <- pvalue / num_randomizations
pvalue
```

```
## [1] 0.0081
```

Note that the p-value is different (and in fact smaller) than that calculated using the normal approximation:

```
cor.test(papers_year$TitleLength, papers_year$Cited.by, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  papers_year$TitleLength and papers_year$Cited.by
## z = 2.3902, p-value = 0.01684
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.1140872
```

Whenever possible, use randomizations, rather than relying on classical tests. They are more computationally expensive, but they allow you to avoid making assumptions about your data.

Writing functions

We have written code that analyzes one year of data. If we wanted to repeat the analysis on a different year, we would have to modify the code slightly. Instead of doing that, we can write a function that allows us to select a given year, and randomize the data. To do so, we need to learn about functions.

The R community provides about 7,000 packages. Still, sometimes there isn't an already made function capable of doing what you need. In these cases, you can write your own functions. In fact, it is in general a good idea to always divide your analysis into functions, and then write a small "master" program that calls the functions and performs the analysis. In this way, the code will be much more legible, and you will be able to recycle the functions for your other projects.

A function in R has this form:

```
my_function_name <- function(arguments of the function){
  # Body of the function
  # ...
  #
```

```
  return(return_value) # this is optional
}
```

A few examples:

```
sum_two_numbers <- function(a, b){
  apb <- a + b
  return(apb)
}
sum_two_numbers(5, 7.2)
```

```
## [1] 12.2
```

You can set a default value for some of the arguments: if not specified by the user, the function will use these defaults:

```
sum_two_numbers <- function(a = 1, b = 2){
  apb <- a + b
  return(apb)
}
sum_two_numbers()
```

```
## [1] 3
```

```
sum_two_numbers(3)
```

```
## [1] 5
```

```
sum_two_numbers(b = 9)
```

```
## [1] 10
```

The return value is optional:

```
my_factorial <- function(a = 6){
  if (as.integer(a) != a) {
    print("Please enter an integer!")
  } else {
    tmp <- 1
    for (i in 2:a){
      tmp <- tmp * i
    }
    print(paste(a, "! = ", tmp, sep = ""))
  }
}
my_factorial()
```

```
## [1] "6! = 720"
```

```
my_factorial(10)
```

```
## [1] "10! = 3628800"
```

You can return **only one** object. If you need to return multiple values, organize them into a vector/matrix/list and return that.

```
order_two_numbers <- function(a, b){
  if (a > b) return(c(a, b))
  return(c(b,a))
```

```
}
order_two_numbers(runif(1), runif(1))
```

```
## [1] 0.8116154 0.3387054
```

Having learned a little about functions, we want to write one that takes as input a vector of citations, a vector of title lengths, and a number of randomizations to perform. The function returns the value of τ as well as the associated p-value. In R, we can write:

```
tau_citations_titlelength <- function(citations, titlelength, num_randomizations = 1000){
  tau_original <- cor(titlelength, citations, method = "kendall")
  pvalue <- 0 # initialize at 0
  for (i in 1:num_randomizations){
    # calculate cor on shuffled data
    tau_shuffle <- cor(titlelength,
                       sample(citations), # scramble the citations at random
                       method = "kendall")
    if (tau_shuffle >= tau_original){
      pvalue <- pvalue + 1
    }
  }
  # calculate proportion
  pvalue <- pvalue / num_randomizations
  # return a list
  return(list(tau = tau_original,
              pvalue = pvalue))
}
```

We can write a loop that calls in turn the function for each year separately:

```
all_years <- sort(unique(papers$Year))
for (my_year in all_years){
  tmp <- tau_citations_titlelength(papers$Cited.by[papers$Year == my_year],
                                   papers$TitleLength[papers$Year == my_year],
                                   1000)
  print(paste(my_year, "-> Tau:", round(tmp$tau, 3), "pvalue:", tmp$pvalue))
}
```

```
## [1] "2004 -> Tau: 0.006 pvalue: 0.433"
## [1] "2005 -> Tau: 0.054 pvalue: 0.12"
## [1] "2006 -> Tau: 0.114 pvalue: 0.013"
## [1] "2007 -> Tau: 0.01 pvalue: 0.428"
## [1] "2008 -> Tau: 0.065 pvalue: 0.082"
## [1] "2009 -> Tau: 0.012 pvalue: 0.398"
## [1] "2010 -> Tau: -0.052 pvalue: 0.855"
## [1] "2011 -> Tau: 0.145 pvalue: 0.002"
## [1] "2012 -> Tau: 0.114 pvalue: 0.003"
## [1] "2013 -> Tau: 0.045 pvalue: 0.144"
```

Organizing and running code

Now we would like to be able to automate the analysis, such that we can repeat it for each journal. This is a good place to pause and introduce how to go about writing programs that are well-organized, easy to write, and easy to debug.

1. Take the problem, and divide it into its basic building blocks

2. Write the code for each building block separately, and test it thoroughly.
3. Extensively document the code, so that you can understand what you did, how you did it, and why.
4. Combine the building blocks into a master program.

For example, let's say we want to write a program that takes as input the name of a file containing the data on titles, years and citations for a given journal. The program should first run the linear model:

`log(citations + 1) ~ Year (categorical) + TitleLength`

And output the coefficient associated with `TitleLength` as well as its p-value.

Then, the program should run the Kendall's test for each year separately, again outputting τ and the p-value obtained with the normal approximation for each year.

Dividing it into blocks, we need to wite:

- code to load the data, calculate title lengths and log citations
- a function to perform the linear model
- a function to perform the Kendall's test
- a master code putting it all together

Our first function:

```
load_data <- function(filename){
  papers <- read.csv(filename, stringsAsFactors = FALSE)
  papers$TitleLength <- nchar(papers$Title)
  papers$LogCits <- log10(papers$Cited.by + 1)
  return(papers)
}
```

Make sure that everything is well by testing our function on the data:

```
for (my_file in list.files("../data/citations", full.names = TRUE)){
  print(my_file)
  print(basename(my_file))
  papers <- load_data(my_file)
}
```

```
## [1] "../data/citations/ecology_letters.csv"
## [1] "ecology_letters.csv"
## [1] "../data/citations/nature_genetics.csv"
## [1] "nature_genetics.csv"
## [1] "../data/citations/nature_neuroscience.csv"
## [1] "nature_neuroscience.csv"
```

Now the function to fit the linear model:

```
linear_model_year_length <- function(papers){
  my_model <- summary(lm(LogCits ~ as.factor(Year) + TitleLength, data = papers))
  # Extract the coefficient and the pvalue
  estimate <- my_model$coefficients["TitleLength","Estimate"]
  pvalue <- my_model$coefficients["TitleLength","Pr(>|t|)"]
  return(list(estimate = estimate,
              pvalue = pvalue))
}
```

Let's run this on all files:

```
for (my_file in list.files("../data/citations", full.names = TRUE)){
  print(basename(my_file))
  papers <- load_data(my_file)
```

```
  linear_model <- linear_model_year_length(papers)
  print(paste("Linear model -> coefficient",
        round(linear_model$estimate, 5),
        "pvalue", round(linear_model$pvalue, 5)))
}
```

```
## [1] "ecology_letters.csv"
## [1] "Linear model -> coefficient -3e-04 pvalue 0.45814"
## [1] "nature_genetics.csv"
## [1] "Linear model -> coefficient 0.00276 pvalue 0"
## [1] "nature_neuroscience.csv"
## [1] "Linear model -> coefficient 0.00567 pvalue 0"
```

Now the function that calls the Kendall's test for each year: we write two functions. One subsets the data, and the other simply runs the test.

```
Kendall_test <- function(a, b){
  my_test <- cor.test(a, b, method = "kendall")
  return(list(estimate = as.numeric(my_test$estimate),
              pvalue = my_test$p.value))
}

call_Kendall_by_year <- function(papers){
  all_years <- sort(unique(papers$Year))
  for (yr in all_years){
    my_test <- Kendall_test(papers$TitleLength[papers$Year == yr],
                  papers$Cited.by[papers$Year == yr])
    print(paste("Year", yr, "-> estimate", my_test$estimate, "pvalue", my_test$pvalue))
  }
}
```

Now a master function to test that the program is working:

```
analyze_journal <- function(my_file){
  # First, the linear model
  print(basename(my_file))
  papers <- load_data(my_file)
  linear_model <- linear_model_year_length(papers)
  print(paste("Linear model -> coefficient",
        round(linear_model$estimate, 5),
        "pvalue", round(linear_model$pvalue, 5)))
  # Then, Kendall year by year
  call_Kendall_by_year(papers)
}
analyze_journal("../data/citations/nature_genetics.csv")
```

```
## [1] "nature_genetics.csv"
## [1] "Linear model -> coefficient 0.00276 pvalue 0"
## [1] "Year 2004 -> estimate 0.045610559310653 pvalue 0.370327813824674"
## [1] "Year 2005 -> estimate -0.0550279205871904 pvalue 0.267581272430449"
## [1] "Year 2006 -> estimate 0.0596670158288517 pvalue 0.213567208478809"
## [1] "Year 2007 -> estimate -0.0390688441353209 pvalue 0.418061571895653"
## [1] "Year 2008 -> estimate 0.0674234973583218 pvalue 0.149431264169447"
## [1] "Year 2009 -> estimate 0.0297023932900025 pvalue 0.524960049608212"
## [1] "Year 2010 -> estimate -0.158944978390644 pvalue 0.00176578355758891"
## [1] "Year 2011 -> estimate 0.130352296718688 pvalue 0.00961339158713348"
```

```
## [1] "Year 2012 -> estimate 0.168763320248576 pvalue 0.000111079233389155"
## [1] "Year 2013 -> estimate 0.0792087074318355 pvalue 0.0880890674881281"
```

Finally, let's analyze all the journals!

```
for (my_file in list.files("../data/citations", full.names = TRUE)){
  analyze_journal(my_file)
}
```

```
## [1] "ecology_letters.csv"
## [1] "Linear model -> coefficient -3e-04 pvalue 0.45814"
## [1] "Year 2004 -> estimate -0.0330650126212166 pvalue 0.613150449110409"
## [1] "Year 2005 -> estimate 0.027544573550034 pvalue 0.671742234063287"
## [1] "Year 2006 -> estimate -0.0639728739951933 pvalue 0.334992423839896"
## [1] "Year 2007 -> estimate 0.130805775658087 pvalue 0.0792684296315682"
## [1] "Year 2008 -> estimate -0.0946065886996697 pvalue 0.185851799603582"
## [1] "Year 2009 -> estimate -0.00528034747952401 pvalue 0.932130040345834"
## [1] "Year 2010 -> estimate 0.0478717663229855 pvalue 0.429380872167305"
## [1] "Year 2011 -> estimate -0.103776346604215 pvalue 0.0989632669799205"
## [1] "Year 2012 -> estimate 0.0708774786316527 pvalue 0.205434298051716"
## [1] "Year 2013 -> estimate -0.0532355687009939 pvalue 0.326025326942473"
## [1] "nature_genetics.csv"
## [1] "Linear model -> coefficient 0.00276 pvalue 0"
## [1] "Year 2004 -> estimate 0.045610559310653 pvalue 0.370327813824674"
## [1] "Year 2005 -> estimate -0.0550279205871904 pvalue 0.267581272430449"
## [1] "Year 2006 -> estimate 0.0596670158288517 pvalue 0.213567208478809"
## [1] "Year 2007 -> estimate -0.0390688441353209 pvalue 0.418061571895653"
## [1] "Year 2008 -> estimate 0.0674234973583218 pvalue 0.149431264169447"
## [1] "Year 2009 -> estimate 0.0297023932900025 pvalue 0.524960049608212"
## [1] "Year 2010 -> estimate -0.158944978390644 pvalue 0.00176578355758891"
## [1] "Year 2011 -> estimate 0.130352296718688 pvalue 0.00961339158713348"
## [1] "Year 2012 -> estimate 0.168763320248576 pvalue 0.000111079233389155"
## [1] "Year 2013 -> estimate 0.0792087074318355 pvalue 0.0880890674881281"
## [1] "nature_neuroscience.csv"
## [1] "Linear model -> coefficient 0.00567 pvalue 0"
## [1] "Year 2004 -> estimate 0.00589142291037872 pvalue 0.918745470907139"
## [1] "Year 2005 -> estimate 0.0535867457573801 pvalue 0.243552261640925"
## [1] "Year 2006 -> estimate 0.114087173434659 pvalue 0.0168412989934459"
## [1] "Year 2007 -> estimate 0.00966171293776095 pvalue 0.843003041215137"
## [1] "Year 2008 -> estimate 0.0652257952013621 pvalue 0.200992557539214"
## [1] "Year 2009 -> estimate 0.012124492386758 pvalue 0.79906576984121"
## [1] "Year 2010 -> estimate -0.0518768931100408 pvalue 0.28616709714268"
## [1] "Year 2011 -> estimate 0.145346619264126 pvalue 0.00174156563082106"
## [1] "Year 2012 -> estimate 0.114380879075143 pvalue 0.0125050217924556"
## [1] "Year 2013 -> estimate 0.0446967410995934 pvalue 0.318724166924561"
```

Discussion: How many significant results we should expect, when citations and title lengths are completely independent?

Documenting the code using `knitr`

> *Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to humans what we want the computer to do.*

Donald E. Knuth, Literate Programming, 1984

When doing experiments, we typically keep track of everything we do in a laboratory notebook, so that when writing the manuscript, or responding to queries, we can go back to our documentation to find exactly what we did, how we did it, and possibly why we did it. The same should be true for computational work.

`RStudio` makes it very easy to build a computational laboratory notebook. First, create a new `R Markdown` file (choose `File` -> `New File` -> `R Markdown` from the menu).

The gist of it is that you write a text file (`.Rmd`). The file is then read by an iterpreter that transforms it into an `.html` or `.pdf` file, or even into a Word document. You can use special syntax to render the text in different ways. For example,

```
**********

*Test* **Test2**

# Very large header

## Large header

### Smaller header

## Unordered lists

* First
* Second
    + Second 1
    + Second 2

1. This is
2. A numbered list

You can insert `inline code`

----------
```

The code above yields:

Test **Test2**

Very large header

Large header

Smaller header

Unordered lists

- First
- Second
  - Second 1
  - Second 2

1. This is
2. A numbered list

You can insert `inline code`

The most important feature of `R Markdown`, however, is that you can include blocks of code, and they will be interpreted and executed by `R`. You can therefore combine effectively the code itself with the description of what you are doing.

For example, including

````
```r
print("hello world!")
```
````

will become

```r
print("hello world!")
```

```
## [1] "hello world!"
```

If you don't want to run the `R` code, but just display it, use `{r, eval = FALSE}`; if you want to show the output but not the code, use `{r, echo = FALSE}`.

You can include plots, tables, and even render equations using LaTeX. In summary, when exploring your data or writing the methods of your paper, give `R Markdown` a try!

You can find inspiration in the notes for the Boot Camp: both the notes for Basic and Advanced Computing are written in `R MarkDown`.

Programming Challenge

Instructions

You will work with your own group to solve the following exercise. When you have found the solution, go to `https://stefanoallesina.github.io/BSD-QBio3/` and follow the link `Submit solution to challenge 2` to submit your answer (alternatively, you can go directly to goo.gl/forms/QJhKmdGqRCIuGNPa2. At the end of the boot camp, the group with the highest number of correct solutions will be declared the winner. If you need extra time, you can work with your group during free time or in the breaks in the schedule.

Google Flu Trends

Google Flu started strong, with a paper in Nature (Ginsberg *et al.*, 2008) showing that using data on web queries, one could predict the number of physician visits for influenza-like symptoms. Over time, the quality of predictions degraded considerably, requiring many adjustments to the model. Now defunct, Google Flu Trends has been proposed as a poster child case of the *Big Data hubris* (Lanzer *et al.* Science 2014). In the folder `data/GoogleFlu` you can find the data used by Preis & Moat (2014, `dx.doi.org/10.1098/rsos.140095`) to show that, once accounted for some additional historical data, Google Flu Trends are correlated with outpatient visits due to influenza-like illness.

1. Read the data and plot number of visits vs. `GoogleFluTrends`
2. Calculate the (Pearson's) correlation using `cor`
3. The data spans 2010-2013. In Aug 2013 Google Flu changed their algorithm. Did this lead to improvements? Compare the data from Aug to Dec 2013 with the same months in 2010, 2011, and 2012. For each, calculate the correlation, and see whether the correlation is higher for 2013.

Hints You will need to extract the year from a string for each row. To do so, you can use `substr(google$WeekCommencing, 1,4)`.

Advanced Computing 1 – Data wrangling and visualization

Stefano Allesina

Data Wrangling and Visualization

- **Goal:** learn how to manipulate large data sets by writing efficient, consistent, and compact code. Introduce the use of `dplyr`, `tidyr`, and the "pipeline" operator `%>%`. Produce beautiful graphs and figures for scientific publications using `ggplot2`.
- **Audience:** experienced `R` users, familiar with the data type `data.frame`, loops, functions, and having some notions of data bases.
- **Installation:** the following packages need to be installed: `ggplot2`, `dplyr`, `tidyr`, `lubridate`, `ggthemes`

Data wrangling

As biologists living in the XXI century, we are often faced with tons of data, possibly replicated over several organisms, treatments, or locations. We would like to streamline and automate our analysis as much as possible, writing scripts that are easy to read, fast to run, and easy to debug. Base `R` can get the job done, but often the code contains complicated operations (think of the cases in which you used `lapply` only because of its speed), and a lot of `$` signs and brackets.

We're going to learn about the packages `dplyr` and `tidyr`, which can be used to manipulate large data frames in a simple and straightforward way. These tools are also much faster than the corresponding base `R` commands, are very compact, and can be concatenated into "pipelines".

To start, we need to import the libraries:

```
library(dplyr)
library(tidyr)
```

Then, we need a dataset to play with. We take a dataset containing all the Divvy bikes trips in Chicago in July 2014:

```
divvy <- read.csv("../data/Divvy_Trips_July_2014.csv")
```

A new data type, `tbl`

This is now a data frame:

```
is.data.frame(divvy)
```

`dplyr` ships with a new data type, called a `tbl`. To convert from data frame, use

```
divvy <- tbl_df(divvy)
divvy
```

The nice feature of `tbl` objects is that they will print only what fits on the screen, and also give you useful information on the size of the data, as well as the type of data in each column. Other than that, a `tbl` object behaves very much like a `data.frame`. In some rare cases, you want to transform the `tbl` back into a `data.frame`. For this, use the function `as.data.frame(tbl_object)`.

We can take a look at the data using one of several functions:

- `head(divvy)` shows the first few (10 by default) rows
- `tail(divvy)` shows the last few (10 by default) rows
- `glimpse(divvy)` a summary of the data (similar to `str` in base R)
- `View(divvy)` open in spreadsheet-like window

Selecting rows and columns

There are many ways to subset the data, either by row (subsetting the *observations*), or by column (subsetting the *variables*). For example, suppose we want to count how many trips (of the $>$ 410k) are very short. The column `tripduration` contains the length of the trip in seconds. Let's select only the trips that lasted less than 3 minutes:

```
filter(divvy, tripduration < 180)
```

You can see that "only" 11,099 trips lasted less than three minutes. We have used the command `filter(tbl, conditions)` to select certain observations. We can combine several conditions, by listing them side by side, possibly using logical operators.

> **Exercise:** what does this do?
>
> ```
> filter(divvy, gender == "Male", tripduration > 60, tripduration < 180)
> ```

We can also select particular variables using the function `select(tbl, cols to select)`. For example, select `from_station_name` and `from_station_id`:

```
select(divvy, from_station_name, from_station_id)
```

How many stations are represented in the data set? We can use the function `distinct(tbl)` to retain only the rows that differ from each other:

```
distinct(select(divvy, from_station_name, from_station_id))
```

Showing that there are 300 stations, once we removed the duplicates.

Other ways to subset observations:

- `sample_n(tbl, howmany, replace = TRUE)` sample `howmany` rows at random with replacement
- `sample_frac(tbl, proportion, replace = FALSE)` sample a certain proportion (e.g. `0.2` for 20%) of rows at random without replacement

- `slice(tbl, 50:100)` extract the rows between `50` and `100`
- `top_n(tbl, 10, tripduration)` extract the first `10` rows, once ordered by `tripduration`

More ways to select columns:

- `select(divvy, contains("station"))` select all columns containing the word `station`
- `select(divvy, -gender, -tripduration)` exclude the columns `gender` and `tripduration`
- `select(divvy, matches("year|time"))` select all columns whose names match a regular expression

Creating pipelines using %>%

We've been calling nested functions, such as `distinct(select(divvy, ...))`. If you have to add another layer or two, the code would become unreadable. `dplyr` allows you to "un-nest" these functions and create a "pipeline", in which you concatenate commands separated by the special operator `%>%`. For example:

```
divvy %>% # take a data table
  select(from_station_name, from_station_id) %>% # select two columns
  distinct() # remove duplicates
```

does exactly the same as the command above, but is much more readable. By concatenating many commands, you can create incredibly complex pipelines while retaining readability.

Producing summaries

Sometimes we need to calculate statistics on certain columns. For example, calculate the average trip duration. We can do this using `summarise`:

```
divvy %>% summarise(avg = mean(tripduration))
```

which returns a `tbl` object with just the average trip duration. You can combine multiple statistics (use `first`, `last`, `min`, `max`, `n` [count the number of rows], `n_distinct` [count the number of distinct rows], `mean`, `median`, `var`, `sd`, etc.):

```
divvy %>% summarise(avg = mean(tripduration),
                    sd = sd(tripduration),
                    median = median(tripduration))
```

Summaries by group

One of the most useful features of `dplyr` is the ability to produce statistics for the data once subsetted by *groups*. For example, we would like to measure whether men take longer trips than women. We can then group the data by `gender`, and calculate the mean `tripduration` once the data is split into groups:

```
divvy %>% group_by(gender) %>% summarise(mean = mean(tripduration))
```

showing that women tend to take longer trips than men.

> **Exercise:** count the number of trips for Male, Female, and unspecified gender.

Ordering the data

To order the data according to one or more variables, use `arrange()`:

```
divvy %>% select(trip_id, tripduration) %>% arrange(tripduration)
divvy %>% select(trip_id, tripduration) %>% arrange(desc(tripduration))
```

Renaming columns

To rename one or more columns, use `rename()`:

```
divvy %>% rename(tt = tripduration)
```

Adding new variables using mutate

If you want to add one or more new columns, use the function `mutate`:

```
divvy %>% select(from_station_id, to_station_id) %>%
  mutate(mylink = paste0(from_station_id, "->", to_station_id))
```

use the function `transmute()` to create a new column and drop the original columns. You can also use `mutate` and `transmute` on grouped data:

```
# A more complex pipeline
divvy %>%
  select(trip_id, gender, tripduration) %>%  # select only three columns
  rename(t = tripduration) %>% # rename a column
  group_by(gender) %>% # create a group for each gender value
  mutate(zscore = (t - mean(t) ) / sd(t)) %>% # compute z-score for t, according to gender
  ungroup() %>% # remove group information
  arrange(desc(t), zscore, gender) %>% # order by t (decreasing), zscore, and gender
  head(20) # display first 20 rows
```

Data visualization

The most salient feature of scientific graphs should be clarity. Each figure should make crystal-clear a) what is being plotted; b) what are the axes; c) what do colors, shapes, and sizes represent; d) the message the figure wants to convey. Each figure is accompanied by a (sometimes long) caption, where the details can be explained further, but the main message should be clear from glancing at the figure (often, figures are the first thing editors and referees look at).

Many scientific publications contain very poor graphics: labels are missing, scales are unintelligible, there is no explanation of some graphical elements. Moreover, some color graphs are impossible to understand if printed in black and white, or difficult to discern for color-blind people.

Given the effort that you put in your science, you want to ensure that it is well presented and accessible. The investment to master some plotting software will be rewarded by pleasing graphics that convey a clear message.

In this section, we introduce `ggplot2`, a plotting package for `R` This package was developed by Hadley Wickham who contributed many important packages to `R` (including `dplyr`). Unlike many other plotting

systems, `ggplot2` is deeply rooted in a "philosophical" vision. The goal is to conceive a grammar for all graphical representation of data. Leland Wilkinson and collaborators proposed The Grammar of Graphics. It follows the idea of a well-formed sentence that is composed of a subject, a predicate, and an object. The Grammar of Graphics likewise aims at describing a well-formed graph by a grammar that captures a very wide range of statistical and scientific graphics. This might be more clear with an example – Take a simple two-dimensional scatterplot. How can we describe it? We have:

- **Data** The data we want to plot.
- **Mapping** What part of the data is associated with a particular visual feature? For example: Which column is associated with the x-axis? Which with the y-axis? Which column corresponds to the shape or the color of the points? In `ggplot2` lingo, these are called *aesthetic mappings* (`aes`).
- **Geometry** Do we want to draw points? Lines? In `ggplot2` we speak of *geometries* (`geom`).
- **Scale** Do we want the sizes and shapes of the points to scale according to some value? Linearly? Logarithmically? Which palette of colors do we want to use?
- **Coordinate** We need to choose a coordinate system (e.g., Cartesian, polar).
- **Faceting** Do we want to produce different panels, partitioning the data according to one (or more) of the variables?

This basic grammar can be extended by adding statistical transformations of the data (e.g., regression, smoothing), multiple layers, adjustment of position (e.g., stack bars instead of plotting them side-by-side), annotations, and so on.

Exactly like in the grammar of a natural language, we can easily change the meaning of a "sentence" by adding or removing parts. Also, it is very easy to completely change the type of geometry if we are moving from say a histogram to a boxplot or a violin plot, as these types of plots are meant to describe one-dimensional distributions. Similarly, we can go from points to lines, changing one "word" in our code. Finally, the look and feel of the graphs is controlled by a theming system, separating the content from the presentation.

Basic `ggplot2`

`ggplot2` ships with a simplified graphing function, called `qplot`. In this introduction we are not going to use it, and we concentrate instead on the function `ggplot`, which gives you complete control over your plotting. First, we need to load the package. While we are at it, let's also load a package extending its theming system:

```
library(ggplot2)
library(ggthemes)
```

And then, let's get a small data set, containing the data on the Divvy stations:

```
divvy_stations <- read.csv("../data/Divvy_Stations_July_2014.csv")
```

A particularity of `ggplot2` is that it accepts exclusively data organized in tables (a `data.frame` or a `tbl` object). Thus, all of your data needs to be converted into a data frame format for plotting.

Let's look at the data:

```
head(divvy_stations)
```

For our first plot, we're going to plot the position of the stations, using the latitude (y axis) and longitude (x axis). First, we need to specify a dataset to use:

```
ggplot(data = divvy_stations)
```

As you can see, nothing is drawn: we need to specify what we would like to associate to the x axis, and what to the y axis (i.e., we want to set the *aesthetic mappings*):

```
ggplot(data = divvy_stations) + aes(x = longitude, y = latitude)
```

Note that we concatenate pieces of our "sentence" using the + sign! We've got the axes, but still no graph... we need to specify a geometry. Let's use points:

```
ggplot(data = divvy_stations) + aes(x = longitude, y = latitude) + geom_point()
```

You can now see the outline of Chicago, with the lake on the right (east), the river separating the Loop from the West Loop, etc. As you can see, we wrote a well-formed sentence, composed of **data** + **mapping** + **geometry**. We can add other mappings, for example, showing the capacity of the station using different point sizes:

```
ggplot(data = divvy_stations) +
  aes(x = longitude, y = latitude, size = dpcapacity) +
  geom_point()
```

or colors

```
ggplot(data = divvy_stations) +
  aes(x = longitude, y = latitude, colour = dpcapacity) +
  geom_point()
```

Scatterplots

Using `ggplot2`, one can produce very many types of graphs. The package works very well for 2D graphs (or 3D rendered in two dimensions), while it lack capabilities to draw proper 3D graphs, or networks.

The main feature of `ggplot2` is that you can tinker with your graph fairly easily, and with a common grammar. You don't have to settle on a certain presentation of the data until you're ready, and it is very easy to switch from one type of graph to another.

For example, let's calculate the median `tripduration` by `birthdate`, to see whether older people tend to take longer or shorter trips:

```
duration_byyr <- divvy %>%
  filter(is.na(birthyear) == FALSE) %>% # remove records without birthdate
  filter(birthyear > 1925) %>% # remove ultra centenarian people (probably, errors)
  group_by(birthyear) %>% # group by birth year
  summarise(median_duration  = median(tripduration)) # calculate median for each group

pl <- ggplot(data = duration_byyr) + # data
  aes(x = birthyear, y = median_duration) + # aesthetic mappings
  geom_point() # geometry

pl # or show(pl)
```

We can add a smoother by typing

```
pl + geom_smooth() # spline by default
pl + geom_smooth(method = "lm", se = FALSE) # linear model, no standard errors
```

> **Exercise:** repeat the plot of the median, but grouping the data by `gender` as well as `birthyear`. Set the aesthetic mapping `colour` to plot the results by gender.

Histograms, density and boxplots

How many trips did each bike take? We can plot a histogram showing the number of trips per bike:

```
ggplot(data = divvy, aes(x = bikeid)) + geom_histogram(binwidth = 50)
```

showing a quite uniform density. Speaking of which, we can draw a density plot:

```
ggplot(data = divvy, aes(x = bikeid)) + geom_density()
```

Similarly, we can produce boxplots, for example showing the tripduration for men and women (in `log10`, as the distribution is close to a lognormal):

```
ggplot(data = divvy, aes(x = gender, y = log10(tripduration))) + geom_boxplot()
```

It is very easy to change geometry, for example switching to a violin plot:

```
ggplot(data = divvy, aes(x = gender, y = log10(tripduration))) + geom_violin()
```

Duration by weekday

Now we're going to test whether the trip duration varies considerably by weekday. To do so, we load the package `lubridate`, which contains many excellent functions for manipulating dates and times.

```
library(lubridate)
```

we then create a new variable, `tripday` specifying the day of the week when the trip was initiated. First, we want to transform the string `starttime` into a date:

```
head(divvy) %>% mutate(tripday = mdy_hm(starttime)) #mdy_hm specifies the date format
```

then we can call `wday` with `label = TRUE` to have a label specifying the day of the week:

```
head(divvy) %>% mutate(tripday = wday(mdy_hm(starttime), label = TRUE))
```

Looks good! Let's perform this operation on the whole set:

```
divvy <- divvy %>% mutate(tripday = wday(mdy_hm(starttime), label = TRUE))
```

> **Exercises:**
>
> - Produce a barplot (`geom_bar`) showing the number of trips by day
> - Calculate the median trip duration per weekday. Then plot it with the command:

```
ggplot(medianbyweekday, aes(x = tripday, y = mediantrip)) +
       geom_bar(stat = "identity")
```

the command `stat = "identity"` tells `ggplot2` to interpret the `y` aesthetic mapping as the height of the barplot.

Scales

We can use scales to determine how the aesthetic mappings are displayed. For example, we could set the x axis to be in logarithmic scale, or we can choose how the colors, shapes and sizes are used. `ggplot2` uses two types of scales: `continuous` scales are used for continuos variables (e.g., real numbers); `discrete` scales for variables that can only take a certain number of values (e.g., colors, shapes, sizes).

For example, let's plot a histogram of `tripduration`:

```
ggplot(divvy, aes(x = tripduration)) + geom_histogram() # no transformation
ggplot(divvy, aes(x = tripduration)) + geom_histogram() +
  scale_x_continuous(trans = "log")
ggplot(divvy, aes(x = tripduration)) + geom_histogram() +
  scale_x_continuous(trans = "log10")
ggplot(divvy, aes(x = tripduration)) + geom_histogram() +
  scale_x_continuous(trans = "sqrt", name = "Duration in minutes")
ggplot(divvy, aes(x = tripduration)) + geom_histogram() + scale_x_log10() # shorthand
```

We can use different color scales. We can convert the capacity to a factor, to use discrete scales:

```
pl <- ggplot(data = divvy_stations) +
  aes(x = longitude, y = latitude, colour = as.factor(dpcapacity)) +
  geom_point()
pl + scale_colour_brewer()
pl + scale_colour_brewer(palette = "Spectral")
pl + scale_colour_brewer(palette = "Blues")
pl + scale_colour_brewer("Station Capacity", palette = "Paired")
```

Or use the capacity as a continuous variable:

```
pl <- ggplot(data = divvy_stations) +
  aes(x = longitude, y = latitude, colour = dpcapacity) +
  geom_point()
pl + scale_colour_gradient()
pl + scale_colour_gradient(low = "red", high = "green")
pl + scale_colour_gradientn(colours = c("blue", "white", "red"))
```

Similarly, you can use scales to modify the display of the shapes of the points (`scale_shape_contintuous`, `scale_shape_discrete`), their size (`scale_size_continuous`, `scale_size_discrete`), etc. To set values manually (useful typically for discrete scales of colors or shapes), use `scale_colour_manual`, `scale_shape_manual` etc.

Themes

Themes allow you to manipulate the look and feel of a graph with just one command. The package `ggthemes` extends the themes collection of `ggplot2` considerably. For example:

```
library(ggthemes)
pl <- ggplot(divvy, aes(x = tripduration)) +
  geom_histogram() +
  scale_x_continuous(trans = "log")
pl + theme_bw() # white background
pl + theme_economist() # like in the magazine "The Economist"
pl + theme_wsj() # like "The Wall Street Journal"
```

Faceting

In many cases, we would like to produce a multi-panel graph, in which each panel shows the data for a certain combination of parameters. In `ggplot` this is called *faceting*: the command `facet_grid` is used when you want to produce a grid of panels, in which all the panels in the same row (column) have axis-ranges in common; `facet_wrap` is used when the different panels do not have axis-ranges in common.

For example:

```
pl <- ggplot(data = divvy, aes(x = log10(tripduration))) + geom_histogram(binwidth = 0.1)
show(pl)
ggplot(data = divvy, aes(x = log10(tripduration), group = gender)) +
  geom_histogram(binwidth = 0.1) + facet_grid(~gender)
ggplot(data = divvy, aes(x = log10(tripduration), group = gender)) +
  geom_histogram(binwidth = 0.1) + facet_grid(gender~.)
```

Now faceting by `tripday` and `gender`

```
ggplot(data = divvy, aes(x = log10(tripduration),
                         colour = gender,
                         fill = gender,
                         group = tripday)) +
  geom_histogram(binwidth = 0.1) +
  facet_grid(tripday~gender)
```

Setting features

Often, you want to simply set a feature (e.g., the color of the points, or their shape), rather than using it to display information (i.e., mapping some aestethic). In such cases, simply declare the feature outside the `aes`:

```
pl <- ggplot(data = divvy_stations, aes(x = longitude, y = latitude))
pl + geom_point()
pl + geom_point(colour = "red")
pl + geom_point(shape = 3)
pl + geom_point(alpha = 0.5)
```

Saving graphs

You can either save graphs as done normally in `R`:

```
# save to pdf format
pdf("my_output.pdf", width = 6, height = 4)
print(my_plot)
dev.off()
# save to svg format
svg("my_output.svg", width = 6, height = 4)
print(my_plot)
dev.off()
```

or use the function `ggsave`

```
# save current graph
ggsave("my_output.pdf")
# save a graph stored in ggplot object
ggsave(plot = my_plot, filename = "my_output.svg")
```

Multiple layers

You can overlay different plots. To do so, however, they must share some of the aesthetic mappings. The simplest case is that in which you have only one dataset:

```
ggplot(data = divvy_stations, aes(x = longitude, y = latitude)) +
  geom_density2d() +
  geom_point()
```

in this case, the `geom_density2d` and `geom_point` shared the `aes`, and were taken from the same dataset.

Let's build a more complicated example:

```
# Capacity of stations in Michigan Avenue
data1 <- divvy_stations %>%
  filter(grepl("Michigan", as.character(name))) %>%
  select(name, dpcapacity) %>%
  rename(value = dpcapacity)
data1

# Number of trips leaving stations in Michigan Ave
data2 <- divvy %>%
  filter(grepl("Michigan", as.character(from_station_name))) %>%
  mutate(name = from_station_name) %>%
  select(name) %>%
  group_by(name) %>%
  summarise(value = n())
data2
```

Now we want to plot the capacity:

```
pl <- ggplot(data = data1, aes(x = name, y = value)) +
    geom_point() +
    scale_y_log10() +
    theme(axis.text.x=element_text(angle=90, hjust=1)) # rotate labels
```

And overlay the other data set:

```
pl + geom_point(data = data2, colour = "red")
```

which is allowed, as the two datasets have the same `aes`. Note that Divvy should increase the capacity of the station at Michigan & Oak!

Tidying up data

The best data to plot is the one in *tidy form*, meaning that a) each variable has its own column, and b) each observation has its own row. When data is not in tidy form, you can use the package `tidy` to reshape it.

For example, suppose we want to produce a table in which we have the number of trips departing a certain station by gender. First, we create a summary:

```
station_gender <- divvy %>%
  group_by(from_station_name, gender) %>%
  summarise(tot_trips = n()) %>%
  filter(gender == "Male" | gender == "Female") %>%
  ungroup()
```

Now we would like to create two columns (for `Male` and `Female`), containing the number of trips. To do so, we **spread** the column gender:

```
# Syntax:
# my_tbl %>% spread(COL_TO_SPREAD, WHAT_TO_USE_AS_VALUE, OPTIONAL: fill = NA)
station_gender <- station_gender %>% spread(gender, tot_trips)
station_gender
```

Having reshaped the data, we can see that the station with the highest proportion of women is in Hyde Park:

```
station_gender %>%
  mutate(proportion_female = Female / (Male + Female)) %>%
  arrange(desc(proportion_female))
```

In the data, we have a column from the station of departure and one for that of arrival. Suppose that for our analysis we would need only one column for the station name, and a separate column detailing whether this is the start or the end of the trip:

```
all_stations <- divvy %>% select(from_station_name, to_station_name, tripduration)
```

We can **gather** the two columns creating a column specifying whether it's a from/to station, and one containing the name of the station:

```
# Syntax:
# my_tbl %>% gather(NAME_NEW_COL, NAME_CONTENT, COLS_TO_GATHER)
all_stations %>% gather("FromTo", "StationName", 1:2)
```

Finally, sometimes we need to split the content of a column into several columns. We can use **separate** to do this quickly:

```
divvy %>% select(starttime) %>% separate(starttime, into = c("Day", "Time"), sep = " ")
```

Joining tables

If you have multiple data frames or `tbl` objects with columns in common, it is easy to join them (as in a database). To showcase this, we are going to create a map of all the trips in the data. First, we count the number of trips from/to each pair of stations:

```
num_trips <- divvy %>% group_by(from_station_id, to_station_id) %>% summarise(trips = n())
# remove trips starting and ending at the same point, for easier visualization
num_trips <- num_trips %>% filter(from_station_id != to_station_id)
```

Now we use `inner_join` to combine the data from `num_trips` and `divvy_stations`, creating the columns `x1` and `y1` containing the coordinates of the starting station. If we rename the columns so that their names match, the join is done automatically:

```
only_id_lat_long <- divvy_stations %>% select(id, latitude, longitude)

# Join the coordinates of the starting station
num_trips <- inner_join(num_trips,
                        only_id_lat_long %>%
                          rename(from_station_id = id,
                                 x1 = longitude,
                                 y1 = latitude))
# Join the coordinates of the ending station
num_trips <- inner_join(num_trips,
                        only_id_lat_long %>%
                          rename(to_station_id = id,
                                 x2 = longitude,
                                 y2 = latitude))

num_trips$trips <- as.numeric(num_trips$trips)

# Now we can plot all the trips!
ggplot(data = num_trips,
       aes(x = x1, y = y1, xend = x2, yend = y2,
           alpha = trips / max(trips)) )+
  geom_curve() + scale_alpha_identity() + theme_minimal()
```

Project: network analysis of Divvy data

Now that we have an overview of the methods available, we are going to perform some simple analysis on the data. First of all, we are going to create a matrix of station-to-station flows, where the rows are the starting stations, the columns the ending stations, and coefficients in the matrix measure the number of trips.

For this, we can use a combination of `dplyr` and `tidyr`:

```
flows <- divvy %>%
  select(from_station_id, to_station_id) %>%
  group_by(from_station_id, to_station_id) %>%
```

```
  summarise(trips = n())
# transform into a matrix
flows_mat <- flows %>% spread(to_station_id, trips, fill = 0) %>% as.matrix()
# remove the first col (use it for row name)
rownames(flows_mat) <- flows_mat[,1]
flows_mat <- flows_mat[,-1]
# see one corner of the matrix
flows_mat[1:10, 1:10]
```

Now we're going to rank stations according to their PageRank, the algorithm at the heart of Google's search engine. The idea of PageRank is to simulate a random walk on a set of web-pages: at each step, the random walker can follow a link (with a probability proportional its weight), or "teleport" to another page at random (with small probability). The walk therefore describes a Markov Chain, whose stationary distribution (Perron eigenvector) is the PageRank score for all the nodes. This value indicates how "central" and important a node in the network is.

Mathematically, we want to calculate the Perron eigenvector of the matrix:

$M' = (1 - \epsilon)M + \epsilon U$

Where M is a nonnegative matrix with columns summing to 1, and U is a matrix with all coefficients being 1. ϵ is the teleport probability.

First, we construct the matrix M, by dividing each row for the corresponding row sum, and transposing:

```
M <- t(flows_mat / rowSums(flows_mat))
```

Then, we choose a "teleport probability" (here $\epsilon = 0.01$), and build M':

```
U <- matrix(1, nrow(M), ncol(M))
epsilon <- 0.01
M_prime <- (1 - epsilon) * M + epsilon * U
```

and calculate the PageRank

```
ev <- eigen(M_prime)$vectors[,1]
# normalize ev
ev <- ev / sum(ev)
page_rank <- data.frame(station_id = as.integer(rownames(M_prime)), pagerank = Re(ev))
```

Which stations are the most "central" Divvy stations in Chicago? Let's plot them out:

```
st_pr <- inner_join(divvy_stations, page_rank, by = c("id" = "station_id"))
st_pr <- st_pr %>% mutate(lab = replace(name, pagerank < 0.0055, NA))
ggplot(st_pr,
       aes(x = longitude, y = latitude, colour = pagerank,
           size = pagerank, label = lab)) +
  geom_point() + geom_text(colour = "black",hjust=0, vjust=0)
```

Exercises in groups

The file `data/Chicago_Crimes_May2016.csv` contains a list of all the crimes reported in Chicago in May 2016. Form small groups and work on the following exercises:

- **Crime map** write a function that takes as input a crime's `Primary Type` (e.g., `ASSAULT`), and draws a map of all the occurrences. Mark a point for each occurrence using `Latitude` and `Longitude`. Set the `alpha` to something like 0.1 to show brighter colors in areas with many occurrences.
- **Crimes by community** write a function that takes as input a crime's `Primary Type`, and produces a barplot showing the number of crimes per `Community area`. The names of the community areas are found in the file `data/Chicago_Crimes_CommunityAreas.csv`. You will need to `join` the tables before plotting.
- **Violent crimes** add a new column to the dataset specifying whether the crime is considered violent (e.g., `HOMICIDE`, `ASSAULT`, `KIDNAPPING`, `BATTERY`, `CRIM SEXUAL ASSAULT`, etc.)
- **Crimes in time** plot the number of violent crimes against time, faceting by community areas.
- **Dangerous day** which day of the week is the most dangerous?
- **Dangerous time** which time of the day is the most dangerous (divide the data by hour of the day).
- **Correlation between crimes** which crimes tend to have the same pattern? Divide the crimes by day and type, and plot the correlation between crimes using `geom_tile` and colouring the cells according to the correlation (see `cor` for a function that computes the correlation between different columns).

Advanced Computing 2 – UNIX shell and Regular Expressions

Stefano Allesina

Advanced Computing 2

- **Goal:** Learn to write pipelines for data manipulation and analysis using the `UNIX` shell. Show how to interface the `UNIX` shell and `R`. Introduce the use of regular expressions.
- **Audience:** experienced `R` users, familiar with `dplyr` and `ggplot2`.
- **Installation:** Windows users need to install a UNIX shell emulator (e.g., `Git Bash`); for regular expressions, we are going to use the `R` package `stringr`.

The UNIX Shell

`UNIX` is an operating system (i.e., the software that lets you interface with the computer) developed in the 1970s by a group of programmers at the AT&T Bell laboratories. Among them were Brian Kernighan and Dennis Ritchie, who also developed the programming language `C`. The new operating system was an immediate success in academic circles, with many scientists writing new programs to extend its features. This mix of commercial and academic interest led to the many variants of `UNIX` available today (e.g., OpenBSD, Sun Solaris, Apple OS X), collectively denoted as *nix systems. Linux is the open source `UNIX` clone whose "engine" (kernel) was written from scratch by Linus Torvalds with the assistance of a loosely-knit team of hackers from across the internet.

All *nix systems are multi-user, network-oriented, and store data as plain text files that can be exchanged between interconnected computer systems. Another characteristic is the use of a strictly hierarchical file system.

Why use UNIX?

Many biologists are not familiar with coding in *nix systems and, given that the learning curve is initially fairly steep, we start by listing the main advantages of these systems over possible alternatives.

First, `UNIX` is an operating system written by programmers for programmers. This means that it is an ideal environment for developing your code and storing your data.

Second, hundreds of small programs are available to perform simple tasks. These small programs can be strung together efficiently so that a single line of `UNIX` commands can perform complex operations, which otherwise would require writing a long and complex program. The possibility of creating these pipelines for data analysis is especially important for biologists, as modern research groups produce large and complex data sets, whose analysis requires a level of automation that would be hard to achieve otherwise. For instance, imagine working with millions of files by having to open each one of them manually to perform an identical task, or try opening your single 80Gb whole-genome sequencing file in a software with a graphical user interface! In `UNIX`, you can string a number of small programs together, each performing a simple task, and create a complex pipeline that can be stored in a script (a text file containing all the commands). Then, you can let the computer analyze all of your data while you're having a cup of coffee.

Third, text is the rule: almost anything (including the screen, the mouse, etc.) in `UNIX` is represented as a text file. Using text files means that all of your data can be read and written by any machine, and without

the need for sophisticated (and expensive) proprietary software. Text files are (and always will be) supported by any operating system and you will still be able to access your data decades from today (while this is not the case for most commercial software). The text-based nature of `UNIX` might seem unusual at first, especially if you are used to graphical interfaces and proprietary software. However, remember that `UNIX` has been around since the early 1970s and will likely be around at the end of your career. Thus, the hard work you are putting into learning `UNIX` will pay off over a lifetime.

The long history of `UNIX` means that a large body of tutorials and support web sites are readily available online. Last but not least, `UNIX` is very stable, robust, secure, and—in the case of Linux—freely available.

In the end, entirely avoiding to work in a `UNIX` "shell" is almost impossible for a professional scientist: basically all resources for High-Performance Computing (computer clusters, large workstations, etc.) run a `UNIX` or Linux operating system. Similarly, the transfer of large files, websites, and data between machines is typically accomplished through command-line interfaces.

Directory structure

In `UNIX` we speak of "directories", while in a graphical environment the term "folder" is more common. These two terms are interchangeable and refer to a structure that may contain sub-directories and files. The `UNIX` directory structure is organized hierarchically in a tree. As a biologist, you can think of this structure as a phylogenetic tree. The common ancestor of all directories is also called the "root" directory and is denoted by an individual slash (`/`). From the root directory, several important directories branch:

- `/bin` contains several basic programs.
- `/etc` contains configuration files.
- `/dev` contains the files connecting to devices such as the keyboard, mouse and screen.
- `/home` contains the home directory of each user (e.g., `/home/yourname`; in OS X, your home directory is stored in `/Users/yourname`).
- `/tmp` contains temporary files.

You will typically work in your home directory. From there, you can access the Desktop, Downloads, Documents, and other directories you are likely familiar with. When you navigate the system you are in one directory and can move deeper in the tree or upward towards the root.

Using the terminal

Terminal refers to the interface that you use to communicate with the kernel (the core of the operating system). The terminal is also called *shell*, or *command-line interface* (CLI). It processes the commands you type, translates them for the kernel, and shows you the results of your operations. There are several shells available. Here, we concentrate on the most popular one, the `bash` shell, which is the default shell in both Ubuntu and OS X.

In Ubuntu, you can open a shell by pressing `Ctrl + Alt + t` or by opening the dash (press the `Super` key) and typing `Terminal`. In OS X, you want to open the application `Terminal.app`, which is located in the folder *Utilities* within *Applications*. Alternatively, you can type `Terminal` in *Spotlight*. In either system, the shell will automatically start in your home directory. Windows users can launch `Git Bash` or another terminal emulator.

The command line prompt ends with a "dollar" (`$`) sign. This means the terminal is ready to accept your commands. Here, a `$` sign at the beginning of a line of code signals that the command has to be executed in your terminal. You do not need to type the `$` sign in your terminal, just copy the command that follows it.

In `UNIX`, you can use the `Tab` key to reduce the amount you have to type, which in turn reduces errors caused by typos. When you press `Tab` in a (properly configured) shell, it will try to automatically complete your

command, directory or file name (if multiple completions are possible, you can display them all by hitting the `Tab` key twice). Additionally, you can navigate the history of commands you typed by using the up/down arrows (you do not need to re-type a command that you recently executed). There are also shortcuts that help when dealing with long lines of code:

- `Ctrl + A` Go to the beginning of the line
- `Ctrl + E` Go to the end of the line
- `Ctrl + L` Clear the screen
- `Ctrl + U` Clear the line before the cursor position
- `Ctrl + K` Clear the line after the cursor
- `Ctrl + C` Kill the command that is currently running
- `Ctrl + D` Exit the current shell
- `Alt + F` Move cursor forward one word (in OS X, `Esc + F`)
- `Alt + B` Move cursor backward one word (in OS X, `Esc + B`)

Mastering these and other keyboard shortcuts will save you a lot of time. You may want to print this list and keep it next to your keyboard—in a while you will have them all memorized and will start using them automatically.

Basic UNIX commands

Here we introduce some of the most basic (and most useful) `UNIX` commands. We write the commands in `fixed-width font` and specific, user-provided input is capitalized in square brackets. Again, the brackets and special formatting are not required to execute a command in your terminal.

Many commands require some arguments (e.g., copy which file to where), and all can be modified using the several options available. Typically, options are either written as a dash followed by a single letter (older style, e.g., `-f`) or two dashes followed by words (newer style, e.g., `--full-name`). A command, its options and arguments are separated by a space.

How to get help in UNIX

`UNIX` ships with hundreds of commands. As such, it is impossible to remember them all, let alone all their possible options. Fortunately, each command is described in detail in its manual page, which can be accessed directly from the shell by typing `man [COMMAND OF YOUR CHOICE]` (not available in `Git Bash`). Use arrows to scroll up and down and hit `q` to close the manual page. Checking the exact behavior of a command is especially important, given that the shell will execute any command you type without asking whether you know what you're doing (such that it will promptly remove all of your files, if that's the command you typed). You may be used to more forgiving (and slightly patronizing) operating systems in which a pop-up window will warn you whenever something you're doing is considered dangerous. In `UNIX`, it is always better to consult the manual rather than improvising.

If you want to interrupt the execution of a command, press `Ctrl + C` to halt any command that is currently running in your shell.

Navigating the directory system

You can navigate the hierarchical `UNIX` directory system using these commands:

- `pwd` print the path of the current working directory.
- `ls` list the files and sub-directories in the current directory. `ls -a` list all (including hidden) files. `ls -l` return the long list with detailed information. `ls -lh` provide file sizes with units (B, M, K, etc.).}

- `cd [NAMEOFDIR]` change directory. `cd ..` move one directory up; `cd /` move to the root directory; `cd ~` move to your home directory; `cd -` go back to the directory you visited previously (like "Back"" in a browser).

Handling directories and files

Create and delete files or directories using the following commands:

- `cp [FROM] [TO]` copy a file. The first argument is the file to copy. The second argument is where to copy it (either a directory or a file name).
- `mv [FROM] [TO]` move or rename a file. Move a file by specifying two arguments: the file, and the destination directory. Rename a file by specifying the old and the new file name in the same directory.
- `touch [FILENAME]` Update the date of last access to the file. Interestingly, if the file does not exist, this command will create an empty file.
- `rm [TOREMOVE]` remove a file. `rm -r` deletes the contents of a directory recursively (i.e., including all files and sub-directories in it; use with caution!). Similarly, `rm -f` removes the file and suppresses any prompt asking whether you are sure you want to remove the file.
- `mkdir [DIRECTORY]` make a directory. To create nested directories, use the option `-p` (e.g., `mkdir -p d1/d2/d3`).
- `rmdir [DIRECTORY]` remove an empty directory.

Printing and modulating files

`UNIX` was especially designed to handle text files, which is apparent when considering the multitude of commands dealing with text. Here are a few popular ones:

- `less [FILENAME]` progressively print a file on the screen (press `q` to exit). Funny fact: there is a command called `more` that does the same thing, but with less flexibility. Clearly, in `UNIX`, `less` is `more`.
- `cat [FILENAME]` concatenate and print files.
- `wc [FILENAME]` line, word, and byte (character) count of a file.
- `sort [FILENAME]` sort the lines of a file and print the result to the screen.
- `uniq [FILENAME]` show only unique lines of a file. The file needs to be sorted first for this to work properly.
- `file [FILENAME]` determine the type of a file.
- `head [FILENAME]` print the `head` (i.e., first few lines of a file).
- `tail [FILENAME]` print the tail (i.e., last few lines of a file).
- `diff [FILE1] [FILE2]` show the differences between two files.

Exercise To familiarize yourself with these commands, try the following:

- Go to the `data` directory for this tutorial.
- How many lines are in file `Marra2014_data.fasta`?
- Go back to the `code` directory.
- Create the empty file `toremove.txt`.
- List the content of the directory.
- Remove the file `toremove.txt`.

Miscellaneous commands

- `echo "[A STRING]"` print the string `[A STRING]`.
- `time` time the execution of a command.

- `wget [URL]` download the webpage at `[URL]`. (Available in Ubuntu; for OS X look at `curl`, or install `wget`).
- `history` list the last commands you executed.

Advanced UNIX commands

Redirection and pipes

So far, we have printed the output of each command (e.g., `ls`) directly to the screen. However, it is easy to direct the output to a file (*redirect*) or use it as the input of another command (*pipe*). Stringing commands together in pipes is the real power of `UNIX`—the ability to perform complex processing of large amounts of data in a single line of commands. First, we show how to redirect the output of a command into a file:

```
$ [COMMAND] > [FILENAME]
```

Note that if the file `[FILENAME]` exists, it will be overwritten. If instead we want to append to an existing file, we can use the `>>` symbol as in the following line:

```
$ [COMMAND] >> [FILENAME]
```

When the command is very long and complex, we might want to redirect the content of a file as input to a command, "reversing" the flow:

```
$ [COMMAND] < [FILENAME]
```

To run a few examples, let's start by moving to our `code` directory:

```
$ cd ~/BSD-QBio2/tutorials/advanced_computing2/code
```

The command `echo` can be used to print a string on the screen. Instead of printing to the screen, we redirect the output to a file, effectively creating a file containing the string we want to print:

```
$ echo "My first line" > test.txt
```

We can see the result of our operation by printing the file to the screen using the command `cat`:

```
$ cat test.txt
```

To append a second line to the file, we use `>>`:

```
$ echo "My second line" >> test.txt
$ cat test.txt
```

We can redirect the output of any command to a file. For example, it is quite common to have to determine how many files are in a directory. The files could have been created by an instrument or provided by a collaborator. Before analyzing the data, we want to get a sense of how many files will we need to process. If there are thousands of files, it is quite time consuming to count them by hand or even open a file browser that can do the counting for us. It is much simpler and faster to just type a command or two. To try this, let's create a file listing all the files contained in `data/Saavedra2013`:

```
$ ls ../data/Saavedra2013 >> filelist.txt
$ cat filelist.txt
```

Now we want to count how many lines are in the file. We can do so by calling the command `wc -l` (count only the lines):

```
$ wc -l filelist.txt
$ rm filelist.txt
```

However, we can skip the creation of the file by creating a short pipeline. The pipe symbol `|` tells the shell to take the output on the left of the pipe and use it as the input of the command on the right of the pipe. To take the output of the command `ls` and use it as the input of the command `wc` we can write:

```
$ ls  ../data/Saavedra2013 | wc -l
```

We have created our first, simple pipeline. In the following sections, we are going to build increasingly long and complex pipelines. The idea is always to start with a command and progressively add one piece after another to the pipeline, each time checking that the result is the desired one.

Selecting columns using `cut`

When dealing with tabular data, you will often encounter the Comma Separated Values (CSV) Standard File Format. The CSV format is platform and software independent, making it the standard output format of many experimental devices. The versatility of the file format should also make it your preferred choice when manually entering and storing data.

The main `UNIX` command you want to master for comma-, space-, tab-, or character-delimited text files is `cut`. To showcase its features, we work with data on generation time of mammals published by Pacifici *et al.*. First, let's make sure we are in the right directory (`advanced_computing2/data`). Then, we can print the header (the first line, specifying the content of each column) of the CSV file using the command `head`, which prints the first few lines of a file on the screen, with the option `-n 1`, specifying that we want to output only the first line:

```
$ head -n 1 Pacifici2013_data.csv
TaxID;Order;Family;Genus;Scientific_name;...
```

We now pipe the header to `cut`, specify the character to be used as delimiter (`-d ';'`), and use the `head` command to extract the name of the first column (`-f 1`), or the names of the first four columns (`-f 1-4`):

```
$ head -n 1 Pacifici2013_data.csv | cut -d ';' -f 1
TaxID

$ head -n 1 Pacifici2013_data.csv | cut -d ';' -f 1-4
TaxID;Order;Family;Genus
```

Remember to use the `Tab` key to auto-complete file names and the arrow keys to access your command history.

In the next example, we work with the file content. We specify a delimiter, extract specific columns, and pipe the result to the `head` command—to display only the first few elements:

```
$ cut -d ';' -f 2 Pacifici2013_data.csv  | head -n 5
Order
Rodentia
Rodentia
Rodentia
Macroscelidea

$ cut -d ';' -f 2,8 Pacifici2013_data.csv  | head -n 3
Order;Max_longevity_d
Rodentia;292
Rodentia;456.25
```

Now, we specify the delimiter, extract the second column, skip the first line (the header) using the `tail -n +2` command (i.e., return the whole file starting from the second line), and finally display the first five entries:

```
$ cut -d ';' -f 2 Pacifici2013_data.csv  | tail -n +2 | head -n 5
Rodentia
Rodentia
Rodentia
Macroscelidea
Rodentia
```

We pipe the result of the previous command to the `sort` command (which sorts the lines), and then again to `uniq`, (which takes only the elements that are not repeated). Effectively, we have created a pipeline to extract the names of all the Orders in the database, from Afrosoricida to Tubulidentata (a remarkable Order, which today contains only the aardvark).

```
$ cut -d ';' -f 2 Pacifici2013_data.csv  | tail -n +2 | sort | uniq
Afrosoricida
Carnivora
Cetartiodactyla
...
```

This type of manipulation of character-delimited files is very fast and effective. It is an excellent idea to master the `cut` command in order to start exploring large data sets without the need to open files in specialized programs (if you don't want to modify the content of a file, you should not open it in an editor!).

Exercise:

- If we order all species names (fifth column) of `Pacifici2013_data.csv` in alphabetical order, which is the first species? Which the last?
- How many families are represented in the database?

Substituting characters using `tr`

We often want to substitute or remove a specific character in a text file (e.g., to convert a comma-separated file into a tab-separated file). Such a one-by-one substitution can be accomplished with the command `tr`. Let's look at some examples in which we use a pipe to pass a string to `tr`, which then processes the text input according to the search term and specific options.

Substitute all characters `a` with `b`:

```
$ echo 'aaaabbb' | tr 'a' 'b'
bbbbbbb
```

Substitute every number in the range `1` through `5` with `0`:

```
$ echo '123456789' | tr 1-5 0
000006789
```

Substitute lower-case letters with upper-case (note the one-to-one mapping):

```
$ echo 'ACtGGcAaTT' | tr actg ACTG
ACTGGCAATT
```

We achieve the same result using bracket expressions that provide a predefined set of characters. Here, we use the set of all lower-case letters `[:lower:]` and translate into upper-case letters `[:upper:]`:

```
$ eecho 'ACtGGcAaTT' | tr [:lower:] [:upper:]
ACTGGCAATT
```

We can also indicate ranges of characters to substitute:

```
$ echo 'aabbccddee' | tr a-c 1-3
112233ddee
```

Delete all occurrences of `a`:

```
$ echo 'aaaaabbbb' | tr -d a
bbbb
```

"Squeeze" all consecutive occurrences of `a`:

```
$ echo 'aaaaabbbb' | tr -s a
abbbb
```

Note that the command `tr` reads standard input, and does not operate on files directly. However, we can use pipes in conjunction with `cat`, `head`, `cut`, etc. to create input for `tr`:

```
$  tr ' ' '\t' < [INPUTFILE] > [OUTPUTFILE]
```

In this example we input `[INPUTFILE]` to the `tr` command to replace all spaces with tabs. Note the use of quotes to specify the space character. The tab is indicated by `\t` and is called a "meta-character". We use the backslash to signal that the following character should not be interpreted literally, but rather is a special code referring to a character that is difficult to represent otherwise.

Now we can apply the command `tr` and the commands we have showcased earlier to create a new file containing a subset of the data contained in `Pacifici2013_data.csv`, which we are going to use in the next section.

First, we change directory to the `code`:

```
$ cd ../code/
```

Now, we want to create a version of `Pacifici2013_data.csv` containing only the `Order`, `Family`, `Genus`, `Scientific_name`, and `AdultBodyMass_g` (columns 2-6). Moreover, we want to remove the header, sort the lines according to body mass (with larger critters first), and have the values separated by spaces. This sounds like an awful lot of work, but we're going to see how this can be accomplished piping a few commands together.

First, let's remove the header:

```
$ tail -n +2 ../data/Pacifici2013_data.csv
```

Then, take only the columns 2-6:

```
$ tail -n +2 ../data/Pacifici2013_data.csv | cut -d ';' -f 2-6
```

Now, substitute the current delimiter (`;`) with a space:

```
$ tail -n +2 ../data/Pacifici2013_data.csv | cut -d ';' -f 2-6 | tr -s ';' ' '
```

To sort the lines according to body size, we need to exploit a few of the options for the command `sort`. First, we want to sort numbers (option `-n`); second, we want larger values first (option `-r`, reverse order); finally, we want to sort the data according to the sixth column (option `-k 6`):

```
$ tail -n +2 ../data/Pacifici2013_data.csv |cut -d ';' -f 2-6 | tr -s ';' ' ' | sort -n -r -k 6
```

That's it. We have created our first complex pipeline. To complete the task, we redirect the output of our pipeline to a new file called `BodyM.csv`.

```
$ tail -n +2 ../data/Pacifici2013_data.csv | cut -d ';' -f 2-6 | tr -s
  ';' ' ' | sort -r -n -k 6 > BodyM.csv
```

You might object that the same operations could have been accomplished with a few clicks by opening the file in a spreadsheet editor. However, suppose you have to repeat this task many times, e.g., to reformat every file that is produced by a laboratory device. Then it is convenient to automate this task such that it can be run with a single command.

Similarly, suppose you need to download a large CSV file from a server, but many of the columns are not needed. With `cut`, you can extract only the relevant columns, reducing download time and storage.

Selecting lines using `grep`

`grep` is a powerful command that finds all the lines of a file that match a given pattern. You can return or count all occurrences of the pattern in a large text file without ever opening it. `grep` is based on the concept of regular expressions, which we will cover just below (but using `R`, in which the syntax is slightly different).

We will test the basic features of `grep` using the file we just created. The file contains data on thousands of species:

```
$ wc -l BodyM.csv
5426 BodyM.csv
```

Let's see how many wombats (family Vombatidae) are contained in the data. First we display the lines that contain the term "Vombatidae":

```
$ grep Vombatidae BodyM.csv
Diprotodontia Vombatidae Lasiorhinus Lasiorhinus krefftii 31849.99
Diprotodontia Vombatidae Lasiorhinus Lasiorhinus latifrons 26163.8
Diprotodontia Vombatidae Vombatus Vombatus ursinus 26000
```

Now we add the option `-c` to count the lines:

```
$ grep -c Vombatidae BodyM.csv
3
```

Next, we have a look at the genus `Bos` in the data file:

```
$ grep Bos BodyM.csv
Cetartiodactyla Bovidae Bos Bos sauveli 791321.8
Cetartiodactyla Bovidae Bos Bos gaurus 721000
Cetartiodactyla Bovidae Bos Bos mutus 650000
Cetartiodactyla Bovidae Bos Bos javanicus 635974.3
Cetartiodactyla Bovidae Boselaphus Boselaphus tragocamelus 182253
```

Besides all the members of the `Bos` genus, we also match one member of the genus `Boselaphus`. To exclude it, we can use the option `-w`, which prompts `grep` to match only full words:

```
$ grep -w Bos BodyM.csv
Cetartiodactyla Bovidae Bos Bos sauveli 791321.8
Cetartiodactyla Bovidae Bos Bos gaurus 721000
Cetartiodactyla Bovidae Bos Bos mutus 650000
Cetartiodactyla Bovidae Bos Bos javanicus 635974.3
```

Using the option `-i` we can make the search case-insensitive (it will match both upper- and lower-case instances):

```
$ grep -i Bos BodyM.csv
Proboscidea Elephantidae Loxodonta Loxodonta africana 3824540
Proboscidea Elephantidae Elephas Elephas maximus 3269794
Cetartiodactyla Bovidae Bos Bos sauveli 791321.8
Cetartiodactyla Bovidae Bos Bos gaurus 721000
...
```

Sometimes, we want to know which lines precede or follow the one we want to match. For example, suppose we want to know which mammals have body weight most similar to the gorilla (*Gorilla gorilla*). The species are already ordered by size, thus we can simply print the two lines before the match using the option `-B 2` and the two lines after the match using `-A 2`:

```
$ grep -B 2 -A 2 "Gorilla gorilla" BodyM.csv
Cetartiodactyla Bovidae Ovis Ovis ammon 113998.7
Cetartiodactyla Delphinidae Lissodelphis Lissodelphis borealis 113000
Primates Hominidae Gorilla Gorilla gorilla 112589
Cetartiodactyla Cervidae Blastocerus Blastocerus dichotomus 112518.5
Cetartiodactyla Iniidae Lipotes Lipotes vexillifer 112138.3
```

Use option `-n` to show the line number of the match. For example, the gorilla is the 164th largest mammal in the database:

```
$ grep -n "Gorilla gorilla" BodyM.csv
164:Primates Hominidae Gorilla Gorilla gorilla 112589
```

To print all the lines that do not match a given pattern, use the option `-v`. For example, to get the other species of the genus *Gorilla* with the exception of *Gorilla gorilla*, we can use:

```
$ grep Gorilla BodyM.csv | grep -v gorilla
Primates Hominidae Gorilla Gorilla beringei 149325.2
```

To match one of several strings, use `grep "[STRING1]\|[STRING2]"`

```
$ grep -w "Gorilla\|Pan" BodyM.csv
Primates Hominidae Gorilla Gorilla beringei 149325.2
Primates Hominidae Gorilla Gorilla gorilla 112589
Primates Hominidae Pan Pan troglodytes 45000
Primates Hominidae Pan Pan paniscus 35119.95
```

You can use `grep` on multiple files at a time! Simply, list all the files to use instead of just one file.

Interfacing `R` and the `UNIX` shell

Note: this will work if you're using Mac OSX or UNIX; in Windows options are much more limited.

Calling `R` from the command line

In `R`, you typically work in "interactive" mode — you type a command, it gets executed, you type another command, and so on. Often, we want to be able to re-run a script on different data sets or with different parameters. For that purpose you can store all the commands in a text file (typically, with extension `.R`), and then re-run the analysis by typing in the `UNIX` command line

```
$ Rscript my_script_file.R
```

To properly automate our analysis and figure generation, however, we can additionally pass command-line arguments to `R`. This allows us for instance to perform the analysis using a specific input file, or save the figure using a specific file name.

`Rscript` accepts command-line arguments, that need to be parsed within `R`. The code at the beginning of the following script shows how this is accomplished:

```r
# Get all the command-line arguments
args <- commandArgs(TRUE)
# Assign each argument to a variable,
# making sure to convert it to the right
# type of variable (string by default)

# check the number of arguments
num.args <- length(args)
print(paste("Number of command-line arguments:", num.args))
# print all the arguments
if (num.args > 0) {
```

```
    for (i in 1:num.args) {
        print(paste(i, "->", args[i]))
    }
}

# We can initially set to default values
# (but pay attention to the order,
# the optional arguments should be at the end)
input.file <- "test.txt"
number.replicates <- 10
starting.point <- 3.14

if (num.args >= 1) {
    input.file <- args[1]
}
if (num.args >= 2) {
    number.replicates <- as.integer(args[2])
}
if (num.args >= 3) {
    starting.point <- as.double(args[3])
}

print(c(input.file, number.replicates, starting.point))

# Save this script as my_script.R
# Run the script in bash with different arguments
# Rscript my_script.R abc.txt 5 100.0
# Rscript my_script.R abc.txt 5
# Rscript my_script.R abc.txt
# Rscript my_script.R
```

Calling the command line from R

You can call the operating system from within R (assuming you're in /advanced_computing_2/code):

```
system("wc -l < ../../basic_computing_1/data/H938_Euro_chr6.geno")
```

You can also capture the output from the shell commands and save it into R. Everything is treated as text (convert to numeric if necessary):

```
numlines <- system("wc -l < ../../basic_computing_1/data/H938_Euro_chr6.geno",
                   intern = TRUE)
numlines
```

```
## [1] "43142"
```

You can also use a combination of shell commands and read.table to capture more complex output:

```
mydf <- system("grep rs125283 ../../basic_computing_1/data/H938_Euro_chr6.geno",
                     intern = TRUE)
mydf <- read.table(file = textConnection(mydf))
mydf
```

```
##   V1         V2 V3 V4 V5 V6  V7
## 1  6 rs12528302  G  A 26 59  39
## 2  6 rs12528322  G  A  0 21 103
## 3  6 rs12528313  G  T  1 25  98
## 4  6 rs12528341  C  T  3 31  90
```

Regular expressions in R

Sometimes data is hidden in free text. Think of citations in a manuscript, mentions of DNA motifs in tables, etc. You could copy and paste data from these unstructured texts yourself, but if you have much text, the task is very boring and error-prone. What you need is a way to describe a text pattern to a computer, and then have it extract the data automatically. Regular Expressions do exactly that.

Because you want to describe a text pattern using text, a level of abstraction is inevitable. What you want to do is to construct a pattern using **literal** characters and **metacharacters**.

For all our examples, we will use the package `stringr`, which makes the regular expression syntax consistent (there are many *dialects*), and provides a set of easy-to-use functions:

```
library(stringr)
```

All the functions have a common structure. For example, `str_extract` extracts text matching a pattern: `str_extract(text, pattern)`. The simplest possible expression is one in which the pattern is described literally (i.e., we want to find exactly the text we're typing):

```
str_extract("a string of text", "t")
```

```
## [1] "t"
```

```
str_extract_all("a string of text", "t")
```

```
## [[1]]
## [1] "t" "t" "t"
```

Of course, you need to be able to describe much more general patterns. Use the following metacharacters:

- `\d` Match a digit character (`0-9`)
- `\D` Match any character that is not a digit
- `\n` Match a newline
- `\s` Match a space
- `\t` Match a `Tab`
- `\b` Match a "word boundary"
- `\w` Match a "word" character (alphanumeric)
- `.` Match any character

Some examples (note that to escape characters, you want to use two backslashes — you need to escape the backslash itself!):

```
# find the first digit
str_extract("123.25 grams", "\\d")
```

```
## [1] "1"
```

```
# find word separator + word character + word separator
str_extract("Albert Einstein was a genius", "\\b\\w\\b")
```

```
## [1] "a"
```

```
# find all digits
str_extract_all("my cell is 773 345 6789", "\\d")
```

```
## [[1]]
##  [1] "7" "7" "3" "3" "4" "5" "6" "7" "8" "9"
```

```
# extract all characters
str_extract_all("for example, this and that", ".")
```

```
## [[1]]
##  [1] "f" "o" "r" " " "e" "x" "a" "m" "p" "l" "e" "," " " "t" "h" "i" "s"
## [18] " " "a" "n" "d" " " "t" "h" "a" "t"
```

Of course, you don't want to type `\\w` fifteen times, in case you are looking for a string that is 15 characters long! Rather, you can use quantifiers:

- `*` Match zero or more times. Match as many times as possible.
- `*?` Match zero or more times. Match as few times as possible.
- `+` Match one or more times. Match as many times as possible.
- `+?` Match one or more times. Match as few times as possible.
- `?` Match zero or one times. In case both zero and one time match, prefer one.
- `??` Match zero or one times, prefer zero.
- `{n}` Match exactly `n` times.
- `{n,}` Match at least `n` times. Match as many times as possible.
- `{n,m}` Match between `n` and `m` times.

Exercise

What does this do? Try to guess, and then type the command into `R`

```
str_extract_all("12.06+3.21i", "\\d+\\.?\\d+")
my_str <- "most beautiful and most wonderful have been, and are being, evolved."
str_extract(my_str, "\\b\\w{6,10}\\b")
str_extract(my_str, "\\b\\w+\\b")
str_extract(my_str, "w\\w*")
str_extract(my_str, "b\\w*")
str_extract(my_str, "b\\w+?")
str_extract(my_str, "\\s\\wn\\w+")
```

What if you want to match the characters `?`, `+`, `*`, `.`? You will need to escape them: for example, `\\.` matches the "dot" character.

You can specify anchors to signal that the match has to be in certain special positions in the text:

- `^` Match at the beginning of a line.
- `$` Match at the end of a line.

```
str_extract("Ah, ba ba ba ba Barbara Ann", "\\w{2,}$")
```

```
## [1] "Ann"
```

```
str_extract("Ah, ba ba ba ba Barbara Ann", "^\\w{2,}")
```

```
## [1] "Ah"
```

To match one of several characters, list them between brackets:

```
str_extract("01234567890", "[3120]+")
```

```
## [1] "0123"
```

```
str_extract("01234567890", "[3-5]+")
```

```
## [1] "345"
```

```
str_extract("supercalifragilisticexpialidocious", "[a-i]{3,}")
```

```
## [1] "agi"
```

To match either of two patterns, use alternations:

```
str_extract_all("The quick brown fox jumps over the lazy dog", "fox|dog")
```

```
## [[1]]
## [1] "fox" "dog"
```

If you need more complex alternations, use parentheses to separate the patterns.

Parentheses can also be used to define **groups**, which are used when you want to capture unknown text that is however flanked by known patterns. For example, suppose you want to save the user name of a UofC email:

```
str_match_all("sallesina@uchicago.edu mjsmith@uchicago.edu",
              "\\b([a-zA-Z0-9]*?)@uchicago.edu")
```

```
## [[1]]
##      [,1]                     [,2]
## [1,] "sallesina@uchicago.edu" "sallesina"
## [2,] "mjsmith@uchicago.edu"   "mjsmith"
```

Note that you have to use `str_match` or `str_match_all` to obtain information on the groups.

Useful functions

Many functions have a similar `str_***_all` version, returning all matches.

- `str_detect(strings, pattern)` do the `strings` contain the `pattern`? Returns a logical vector.
- `str_locate(strings, pattern)` find the character position of the pattern
- `str_extract(strings, pattern)` extracts the first match
- `str_match(strings, pattern)` like `extract` but capture groups defined by parentheses
- `str_replace(strings, pattern, newstring)` replaces the first matched `pattern` with `newstring`

Exercises in groups

Extract primers and polymorphic sites

In the file `data/Ptak_etal_2004.txt` you find a text version of the supplementary materials of Ptak *et al.* (PLoS Biology 2004, doi:10.1371/journal.pbio.0020155).

- Take a look at the file. You see that it first lists the primers used for the study (e.g., `TAP2-17-3' -> CTTGGATATAACACCAAACGCA`), and then the polymorphic sites for 24 chimpanzees (e.g., .
- Read the text as a single string, intervaled by new lines

```
my_txt <- paste(readLines("../data/Ptak_etal_2004.txt"), collapse="\n")
```

- Use regular expressions to produce a data frame containing the primers used in the study:

```
head(primers)
```

```
         ID               Sequence
1 TAP2-1-5' GAGAATCACTTGAACCTGGGAG
2 TAP2-2-5' TTGTCCACAGTGTACCACATGA
3 TAP2-3-5' TATTTCTTCCTGGGGTTTCCTT
4 TAP2-4-5' CATGATGTGTCATGCTGAATTG
5 TAP2-5-5' ATAGAACAAGAACCAAAGCCCA
6 TAP2-6-5' GGACAACAGATAAAGTTGCCCT
```

- Write another regular expression to extract the polymorphic region for each chimp. Use `str_replace_all` to remove extra spaces and newlines. The results should look like:

```
  Chimp                                                Sequence
1   311 ATACCCTGGAGGCAAGAATCTTCCGATAGACGCCAGTCCCTAGTTGT...
2   312 GCACCCTGGAGGCAAGAGTCTTCCGATAGACGCCAGTCCCTAGTTGC...
3   313 GCACCCTGGAGGCAAGAGTCTTCCGATAGACGCCAGTCCCCAGTTGT...
4   314 GCACCCTGGAGGCAAGAGTCTTCCGATAGACCCCAGTCCCCAGTTGC...
5   317 GCACCCTGGAGGTGGGGGCCCCCCAGGAGGCCCCAGCCCCCGGTCGT...
6   320 GCACCCTGGAGATAAGGGCCCCCCAGGAGGCCCCAGCCCCCGGTCGT...
```

A map of *Science*

Where does science come from? This question has fascinated researchers for decades and even led to the birth of the field of "Science of Science", where researchers use the same tools they invented to investigate nature to gain insights on the development of science itself. In this exercise, you will build a map of *Science*, showing where articles published in *Science* magazine have originated. You will find two files in the directory `data/MapOfScience`. The first, `pubmed_results.txt`, is the output of a query to PubMed listing all the papers published in Science in 2015. You will extract the US ZIP codes from this file, and then use the file `zipcodes_coordinates.txt` to extract the geographic coordinates for each ZIP code.

- Read the file `pubmed_results.txt`, and extract all the US ZIP codes.
- Count the number of occurrences of each ZIP code using `dplyr`.
- Join the table you've created with the data in `zipcodes_coordinates.txt`
- Plot the results using `ggplot2` (either use points with different colors/alphas, or render the density in two dimensions)

Defensive Programming in R

Sarah Cobey

Defensive Programming

- **Goal:** Convince new and existing programmers of the importance of defensive programming practices, introduce general programming principles, and provide specific tips for programming and debugging in `R`.
- **Audience:** Scientific researchers who use `R` and believe the accuracy of their code and/or efficiency of their programming could be improved.
- **Installation:** For people who have completed the other modules, there is nothing new to install. For others starting fresh, install `R` and `RStudio`. To install `R`, follow instructions at `cran.rstudio.com`. Then install `Rstudio` following the instructions at `goo.gl/a42jYE`. Download the `knitr` package.

Motivation

Defensive programming is the practice of anticipating errors in your code and handling them efficiently.

If you're new to programming, defensive programming might seem tedious at first. But if you've been programming long, you've probably experienced firsthand the stress from

- inexplicable, strange behavior by the code
- code that seems to work under some conditions but not others
- incorrect results or bugs that take days or weeks to fix
- a program that seems to produce the correct results but then, months or years later, gives you an answer that you know must be wrong... thereby putting all previous results in doubt
- the nagging feeling that maybe there's still a bug somewhere
- not getting others' code to run or run correctly, even though you're following their instructions

Defensive programming is thus also a set of practices for preserving sanity and conducting research efficiently. It is an art, in that the best methods vary from person to person and from project to project. As you will see, which techniques you use depend on the kind of mistakes you make, who else will use your code, and the project's requirements for accuracy and robustness. But that flexibility does not imply defensive programming is "optional": steady scientific progress depends on it. In general, we need scientific code to be perfectly accurate (or at least have well understood inaccuracies), but compared to other programmers, we are less concerned with security and ensuring that completely naive users can run our programs under diverse circumstances (although standards here are changing).

In the first part of this tutorial, we will review key principles of defensive programming for scientific researchers. These principles hold for all languages, not just `R`. In the second part, we will consider R-specific debugging practices in more depth.

Part 1: Principles

Part 1 focuses on defense. You saw a few of these principles in Basic Computing 2, but they are important enough to be repeated here.

1. Before writing code, draft your program as a series of modular functions with high-level documentation, and develop tests for it.

2. Write clearly and cautiously.
3. Develop one function at a time, and test it before writing another.
4. Document often.
5. Refactor often.
6. When you run your code, save each "experiment" in a separate directory with a complete copy of the code repository and all parameters.
7. Don't be *too* defensive.

In part 2, we will focus on what to do when tests (from Principle 3) indicate something is wrong.

Principle 1. Before writing code, draft your program as a series of modular functions with high-level documentation, and develop tests for it.

Many of us have had the experience of writing a long paper only to realize, several pages in, that we in fact need to say something slightly different. Restructuring a paper at this stage can be a real chore. Outlining your code as you would outline a paper avoids this problem. In fact, outlining your code can be even more helpful because it helps you think not just generally about how your algorithm will flow and where you want to end up, but also about what building blocks (functions and containers) you'll use to get there. Thinking about the "big picture" design will prevent you from coding yourself into a tight spot—when you realize hundreds of lines in that you should've been tracking past states of a variable, for instance, but your current containers are only storing the current state. Drafting also makes the actual writing much more easy and fun.

For complex programs, your draft may start as a diagram or flowchart showing how major functions and variables relate to one another or brief notes about each of step of the algorithm. These outlines are known as pseudocode, and there are myriad customs for pseudocode and programmers loyal to particular styles. For simple scripts, it is often sufficient to write pseudocode as placeholder comments. For instance, if we are simulating a population in which individuals can be born or die (and nothing else happens), we could write:

```
# initialize population size (N), birth rate (b), death rate (d),
# total simulation time (t_max), and time (t=0) while N > 0 and t < t_max
# ...generate random numbers to determine whether next event is
# birth or death, and time to event (dt) ...update N
# (increment if birth, decrement if death) and update time t to t+dt
```

Here, `b` and `d` are per capita rates. This is an example of the Gillespie algorithm. It was initially developed as an exact stochastic approach for simulating chemical reaction kinetics, and it is widely used in biology, chemistry, and physics.

Can you see some limitations of the pseudocode so far? First, it lacks obvious modularity, though this is partly due to its vagueness. The first step under the `while` loop could become its own function that is defined separately. Second, it is missing a critical feature, in that it's not obvious what is being output: do we want the population size at the end of the simulation, or the population size at each event? If the latter, we may need to initialize a container, such as a dataframe, in which we store the value of `N` and `t` at every event. After further thought, we might decide such a container would be too big—perhaps we only need to know the value of `N` at 1/1000th the typical rate of births, and so we might introduce an additional loop to store `N` only when the new time `(t+dt)` exceeds the most recent prescribed observation/sampling time. This sampling time would need to be stored in an extra variable, and we could also make the sampling procedure into its own function. And maybe we want a function to plot `N` over time at the end.

The next stage of drafting is to consider how the code might go wrong, and what it would take to convince ourselves that it is accurate. We'll spend more time on this later, but now is the time to think of every possible sanity check for the code. Sometimes this can lead to changes in code design.

Exercise

What sanity checks and tests would you include for the code above?

Some examples:

- Initial values of `b`, `d`, and `t_max` should be non-negative and not change
- The population size `N` should probably start as a positive whole number and never fall below zero
- If the birth and death rates are zero, `N` should not change
- If the birth rate equals the death rate, on average, `N` should not change when it is large
- The ratio of births to deaths should equal the ratio of the birth rate to the death rate, on average
- The population should, on average, increase exponentially at rate `b-d` (or decrease exponentially if `d>b`)

Some of these criteria arise from common sense or assumptions we want to build into the model (for instance, that the birth and death rates aren't negative), and others we know from the mathematics of the system. For instance, the population cannot change if there are no births and deaths, and it must increase on average if the birth rate exceeds the death rate. It helps that the Gillespie algorithm represents kinetics that can be written as simple differential equations. However, because the simulations are stochastic, we need to look at many realizations (randomizations, trajectories) to ensure there aren't consistent biases. We must use statistics to confirm that the distribution of `N` in 10,000 simulations after 100 time units is not significantly different from what we would predict mathematically.

The bottom line is that we have identified some tests for the (1) inputs, (2) intermediate variable states (like `N`), (3) final outputs to test that the program is running correctly. Note that one of our tests, the fraction of birth to death events, is not something we were originally tracking, and thus we might decide now to create a separate function and variables to handle these quantities. We will talk in more detail about how to implement these tests in Principle 3. Generally, tests of specific functions are known as **unit tests**, and tests of aggregate behavior (like the trajectories of 10,000 simulations) are known as **system tests**. As a general principle, we need to have both, and we need as much as possible to compare their outputs to analytic expectations. It is also very useful to identify what you want to see right away. For instance, you may want to write a function to plot the population size over time *before* you code anything else because having immediate visual feedback can be extremely helpful.

> "Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a *computer* what to do, let us concentrate rather on explaining to *human beings* what we want a computer to do." -Donald Knuth

This is also a good time to draft very high-level documentation for your code, for instance in a `readme.MD` file. What are the inputs, what does the code do, and what does it return?

> **Exercise**
>
> Assume there are two discrete populations. Each has nonoverlapping generations and the same generation time. Their per capita birth rates are `b1` and `b2`. Some of the newborns migrate between populations.
>
> - Write pseudocode to calculate the distribution of population frequencies after 100 generations.
> - Is your code optimally modular?
> - What are the inputs and outputs of the program? Of each function?
> - How could you test the inputs, functions, and overall program?
> - Discuss your approach with your neighbor.

Principle 2. Write clearly and cautiously.

You're already on your way to writing clearly and cautiously if you outline your program before you start writing it. Here we'll discuss some practices to follow as you write.

A general rule is that it is more important for scientific code to be readable and correct than it is for it to be fast. Do not obsess too much about the efficiency of the code when writing.

> "Premature optimization is the root of all evil." -Donald Knuth

Develop useful conventions for your variable and function names. There are many conventions, some followed more religiously than others. It's most important to be consistent within your own code.

- Don't make yourself and others guess at meaning by making names too short. It's generally better to write out `maxSubstitutionRatePerSitePerYear` or `max.sub.rate.per.site.yr` than `maxSR`.
- However, very common variables should have short names.
- When helpful, incorporate identifiers like `df` (data frame), `ctr` (counter), or `idx` (index) into variable names to remind you of their purpose or structure.
- Customarily, variables that start with capital letters indicate global scope in `R`, and function names also start with capital letters. People argue about conventions, and they vary from language to language. Here's Google's style guide.

Do not use magic numbers. "Magic numbers" refer to numbers hard-coded in your functions and main program. Any number that could possibly vary should be in a separate section for parameters. The following code is not robust:

```
while ((age >= 5) && (inSchool == TRUE)) {
  yearsInSchool = yearsInSchool + 1
}
```

We may decide the age cutoff of 5 is inappropriate, but even if we never do, being unable to change the cutoff limits our ability to test the code. Better:

```
while ((age >= AgeStartSchool) && (inSchool == TRUE)) {
  yearsInSchool = yearsInSchool + 1
}
```

Most of the time, the only numbers that should be hard-coded are 0, 1, and pi.

Use labels for column and row names, and load functions by argument.

Don't force (or trust) yourself to remember that the first column of your data frame contains the time, the second column contains the counts, and so on. When reviewing code later, it's harder to interpret `cellCounts[,1]` than `cellCounts$time`.

In the same vein, if you have a function taking multiple inputs, it is safest to pass them in with named arguments. For instance, the function

```
BirthdayGiftSuggestion <- function(age,budget) {
  # ...
}
```

could be called with

```
BirthdayGiftSuggestion(age=30,budget=20)
# or
BirthdayGiftSuggestion(30,20)
```

but the former is obviously safer.

Avoid repetitive code. If you ever find yourself copying and pasting a section of code, perhaps modifying it slightly each time, stop. It's almost certainly worth writing a function instead. The code will be easier to read, and if you find an error, it will be easier to debug.

Principle 3. Develop one function at a time, and test it before writing another.

The first part of this principle is easy for scientists to understand. When building code, we want to change one thing at a time. It makes it easier to understand what's going on. Thus, we start by writing just a single function. It might not do exactly what we want it to do in the final program (e.g., it might contain mostly

placeholders for functions it calls that we haven't written yet), but we want to be intimately familiar with how our code works in every stage of development.

The second part of this principle underscores one of the most important rules in defensive programming: **do not believe anything works until you have tested it thoroughly, and then keep your guard up.** *Expect* your code to contain bugs, and leave yourself time to play with the code (e.g., by trying to "break" it) until you can convince yourself they are gone. This involves an extra layer of defensive programming beyond the straightforward good practices discussed in Principle 2. Testing the code as you build it makes it much faster to find problems.

Unit tests. Unit tests are tests on small pieces of code, often functions. An intuitive and informal method of unit testing is to include `print()` statements in your code.

Here's a function to calculate the Simpson Index, a useful diversity index. It gives the probability that two randomly drawn individuals belong to the same species or type:

```
SimpsonIndex <- function(C) {
  print(paste(c("Passed species counts:", C),collapse=" "))
  fractions <- C/sum(C)
  print(paste(c("Fractions:", round(fractions, 3)), collapse=" "))
  S <- sum(fractions^2)
  print(paste("About to return S =",S))
  return(S)
}

# Simulate some data
numSpecies <- 10
maxCount <- 10^3
fakeCounts <- floor(runif(numSpecies,min=1,max=maxCount))

# Call function with simulated data
S <- SimpsonIndex(fakeCounts)
```

```
## [1] "Passed species counts: 368 836 677 544 694 281 41 873 191 995"
## [1] "Fractions: 0.067 0.152 0.123 0.099 0.126 0.051 0.007 0.159 0.035 0.181"
## [1] "About to return S = 0.130231338842975"
```

It's very useful to print values to screen when you are writing a function for the first time and testing that one function. When you've drafted your function, I recommend walking through the function with `print()` and comparing the computed values to calculations you perform by hand or some other way. It can also be useful to do this at a very high level (more on that later).

The problem with relying on `print()` is that it rapidly provides too much information for you to process, and hence errors can slip through.

Assertions

A more reliable way to catch errors is to use assertions. Assertions are automated tests embedded in the code. The built-in function for assertions in `R` is `stopifnot()`. It's very simple to use.

Let's remove the `print` statements and add a check to our input data:

```
SimpsonIndex <- function(C) {
  stopifnot(C>0)
  fractions <- C/sum(C)
  S <- sum(fractions^2)
  return(S)
}
```

If each element of our abundances vector `C` is positive, `stopifnot()` will be `TRUE`, and the program will continue. If any element does not satisfy the criterion, then `FALSE` will be returned, and execution will terminate. Explore for yourself:

```
# Simulate two sets of data
numSpecies <- 10
maxCount <- 10^3
goodCounts <- floor(runif(numSpecies,min=1,max=maxCount))
badCounts <- floor(runif(numSpecies,min=-maxCount,max=maxCount))

# Call function with each data set
S <- SimpsonIndex(goodCounts)
S <- SimpsonIndex(badCounts)
```

This gives a very literal error message, which is often enough when we are still developing the code. But what if the error might arise in the future, e.g., with future inputs? We can use the built-in function `stop()` to include a more informative message:

```
SimpsonIndex <- function(C) {
  if(any(C<0)) stop("Species counts should be positive.")
  fractions <- C/sum(C)
  S <- sum(fractions^2)
  return(S)
}
```

Now try it with `badCounts` again.

What about warnings? For instance, our calculation of the Simpson Index is an approximation: the index formally assumes we draw without replacement, but we compute $S = \sum p^2$, where p is the fraction of each species. It should be $S = \sum \frac{n(n-1)}{N(N-1)}$, where n is the abundance of each species n and N the total abundance. This simplification becomes important at small sample sizes. We could add a warning to alert users to this issue:

```
SimpsonIndex <- function(C) {
  if(any(C<0)) stop("Species counts should be positive.")
  if((mean(C)<20) || (min(C)<5)) {
    warning("Small sample size. Result will be biased. Consider corrected index.")
  }
  fractions <- C/sum(C)
  S <- sum(fractions^2)
  return(S)
}

smallCounts <- runif(10)
S <- SimpsonIndex(smallCounts)
```

```
## Warning in SimpsonIndex(smallCounts): Small sample size. Result will be
## biased. Consider corrected index.
```

The main advantage of `warning()` over `print()` is that the message is red and will not be confused with expected results, and warnings can be controlled (see `?warning`).

You could make a case that `warning()` should be `stop()`. In general, with defensive programming, you want to halt execution quickly to identify bugs and to limit misuse of the code.

Exercise

- What other input checks would make sense with `SimpsonIndex()`?

- You can see how the code would be more readable and organized if most of that function were dedicated to actually calculating the Simpson Index. Draft a separate function, `CheckInputs()`, and include all tests you think are reasonable.
- When you and a neighbor are done, propose a bad or dubious input for their function and see if it's caught.

There are many packages that produce more useful assertions and error messages than what is built into `R`. See, e.g., `assertthat` and `testit`.

Exception handling

Warnings and errors are considered "exceptions." Sometimes it is useful to have an automated method to handle them. `R` has two main functions for this: `try()` allows you to continue executing a function after an error, and `tryCatch()` allows you to decide how to handle the exception.

Here's an example:

```
UsefulFunction <- function(x) {
  value <- exp(x)
  otherStuff <- rnorm(1)
  return(list(value,otherStuff))
}

data <- "2"
results <- UsefulFunction(data)
print(results)
```

Now `results` is quite a disappointment: it could've at least returned a random number for you, right? You could instead try

```
UsefulFunction <- function(x){
  value <- NA
  try(value <- exp(x))
  otherStuff <- rnorm(1)
  return(list(value,otherStuff))
}
results <- UsefulFunction(data)
print(results)
```

```
## [[1]]
## [1] NA
##
## [[2]]
## [1] -0.1445207
```

Even though the function still can't exponentiate a string (`exp("2")` still fails), execution doesn't terminate. If we want to suppress the error message, we can use `try(..., silent=TRUE)`. This obviously carries some risk!

We could make this function even more useful by handling the error responsibly with `tryCatch()`:

```
UsefulFunction <- function(x){
  value <- NA
  tryCatch ({
      message("First attempt at exp()...")
      value <- exp(x)},
    error = function(err){
      message(paste("Darn:",err," Will convert to numeric."))
      value <<- exp(as.numeric(x))
```

```
      }
    )
  otherStuff <- rnorm(1)
  return(list(value,otherStuff))
}
results <- UsefulFunction(data)
print(results)
```

It is also possible to assign additional blocks for warnings (not just errors). The `<<-` is a way to assign to the value in the environment one level up (outside the `error=` block).

Exercise

The new package `ggjoy` works with package `ggplot2` to show multiple distributions in a superimposed but interpretable way. Let's say we want to run the following code:

```
library(ggplot2)
library(ggjoy)
ggplot(iris, aes(x = Sepal.Length, y = Species)) + geom_joy()
```

You probably don't have `ggjoy` installed yet, so you'll get an error. Use `tryCatch()` so that the package is installed if you do not have it and then loaded.

Test all the scales!

It's important to consider multiple scales on which to test as you develop. We've focused on unit tests (testing small functions and steps) and testing inputs, but it is easy to have correct subroutines and incorrect results. For instance, we can be excellent at the distinct activities of toasting bread, buttering bread, and eating bread, but we will fail to enjoy buttered toast for breakfast if we don't pay attention to the order.

With scientific programming, it is critical to simplify code to the point where results can be compared to analytic expectations. You saw this in Principle 1. It is important to add functions to check not only inputs and intermediate results but also larger results. For instance, when we set the birth rate equal to the death rate, does the code reliably produce a stable population? We can write functions to test for precisely such requirements. These are **system tests**. When you change something in your code, always rerun your system tests to make sure you've not messed something up. Often it's helpful to save multiple parameter sets or data files precisely for these tests.

It's hard to overstate the importance of taking a step back from the nitty-gritty of programming and asking, Are these results reasonable? Does the output make sense with different sets of extreme values? Schedule time to do this, and update your system tests when necessary.

Principle 4. Document often.

It is helpful to keep a running list of known "issues" with your code, which would include the functions left to implement, the tests left to run, any strange bugs/behavior, and features that might be nice to add later. Sites like GitHub and Bitbucket allow you to associate issues with your repositories and are thus very helpful for collaborative projects, but use whatever works for you. Having a formal list, however, is much safer than sprinkling to-do comments in your code (e.g., `# CHECK THIS!!!`). It's easy to miss comments.

Research code will always need a `readme` describing the software's purpose and implementation. It's easiest to develop it early and update as you go.

Principle 5. Refactor often.

To refactor code is to revise it to make it clearer, safer, or more efficient. Because it involves no changes in the scientific outputs (results) of the program, it might feel pointless, but it's usually not. Refactor when

you realize that your variable and function names no longer reflect their true content or purpose (rename things quickly with `Ctrl + Alt + Shift + M`), when certain functions are obsolete or should be split into two, when another data structure would work dramatically better, etc. Any code that you'll be working with for more than a week, or that others might ever use, should probably be refactored a few times. Debugging will be easier, and the code will smell better.

Important tip, repeated: Run unit and system tests after refactoring to make sure you haven't messed anything up. This happens more than you might think.

Principle 6. When you run your code, save each "experiment" in a separate directory with a complete copy of the code repository and all parameters.

When you're done developing the code and are using it for research, keep results organized by creating a separate directory for each execution of the code that includes not only the results but also the precise code used to generate the results. This way, you won't accidentally associate one set of parameters with results that were in fact generated by another set of parameters. Here's a sample workflow, assuming your repository is located remotely on GitHub, and you're in a UNIX terminal:

```
$ mkdir 2017-09-15_rho=0.5
$ cd 2017-09-15_rho=0.5
$ git clone git@github.com:MyName/my-repo
```

If we want, we can edit and execute our code from within `R` or `RStudio`, but we can also keep going with the command line. Here we are using a built-in UNIX text editor known as emacs. If you are a glutton for punishment, you could instead use vi(m).

```
$ cd my-repo
$ emacs parameters.json // (edit parameters, with rho=0.5)
$ Rscript mycode.R
```

Keeping your experiments separate is going to save your sanity for larger projects when you repeatedly revise your analyses. It also makes isolating bugs easier.

Principle 7. Don't be *too* defensive.

This is not necessary:

```
myNumbers <- seq(from=1,to=500,length.out=20)
stopifnot(length(myNumbers)==20)
```

It's fine to check stuff like this when you're getting the hang of a function, but it doesn't need to be in the code. Code with too many defensive checks becomes a pain to read (and thus debug). Try to find a balance between excess caution and naive optimism. Good luck!

Part 2: Debugging in R

Part 1 introduced principles that should minimize the need for aggressive debugging. You're in fact already debugging if you're regularly using input, unit, and system tests to make sure things are running properly. But what happens when despite your best efforts, you're not getting the right result?

We'll focus on more advanced debugging tools in this part. First, some general guidelines for fixing a bug:

- *Isolate the error and make it reproducible.* Try to strip the error down to its essential parts so you can reliably reproduce the bug. If a function doesn't work, copy the code, and keep removing pieces that are non-essential for reproducing the error. When you post for help on the website stackoverflow, for

instance, people will ask for a MRE or MWE—a minimum reproducible (working) example. You need to have not only the pared code but also the inputs (parameter values and the seeds of any random number generators) that cause the problem.
- *Use assertions and debugging tools to hone in on the problem.* We've already seen how to use `stop()` and `stopifnot()` to identify logic errors in unit tests. We'll cover more advanced debugging here.
- *Change/Test one thing at a time.* This is why we develop only one function at a time.

Tracing calls

If an error appears, a useful technique is to see the call stack, the sequence of functions immediately preceding the error. We can do this in R using `trackback()` or in RStudio.

The following example comes from a nice tutorial by Hadley Wickham:

```
f <- function(a) g(a)
g <- function(b) h(b)
h <- function(c) i(c)
i <- function(d) "a" + d
f(10)
```

You can see right away that running this code will create an error. Try it anyway in RStudio. Click on the "Show Traceback" to the right of the error message. What you're seeing is the stack, and it's helpfully numbered. At the bottom we have the most recent (proximate) call that produced the error, the preceding call above it, and so on. (If you're working from a separate R file that you sourced for this project, you'll also see the corresponding line numbers next to each item in the stack.) If you're in R, you can run `traceback()` immediately after the error.

Seeing the stack is useful for checking that the correct functions were indeed called, but it can suggest how to trace the error back in a logical sequence. But we can often debug faster with more information from RStudio's debugger.

Examining the environment

Let's pretend we have a group of people who need to be assigned random partners. These partnerships are directed (so person A may consider his partner person B, but B's partner is person F), and we'll allow the possibility of people partnering with themselves. Some code for this is in the file `BugFun.R`. (It's not terribly efficient code, but it is useful for this exercise.)

```
source("BugFun.R")
peopleIDs <- seq(1:10)
pairings <- AssignRandomPartners(peopleIDs)
```

Try running this a few times. We have an inconsistent bug.

We can use breakpoints to quickly examine what's happening at different points in the function. With breakpoints, execution stops on that line, and environmental states can be inspected. In RStudio, you can create breakpoints by clicking to the left of the line number in the `.R` file. A red dot appears. You can then examine the contents of different variables in debug mode. To do this, you have to make sure you have the right setting defined: Debug > On Error > Break in Code.

Exercise

Use breakpoints to identify the error(s) in the AssignRandomPartners() function. Go to `BugFun.R` and attempt to run the last line. Decide on a place to start examining the code. If you have adjusted your settings, the debug mode should start automatically once you define a breakpoint and try to run the code again. (If a message to source the file appears, follow it.) Your console should now have `Browse[1]>` where it previously had only `>`.

The IDE is now giving you lots of information. The green arrow shows you where you are in the code. The line that is about to be executed is in yellow. Anything you execute in the console shows states from your current environment. Test this for yourself by typing a few variables in the console. You can see these values in the Environment pane in the upper right, and you can also see the stack in the middle right.

If you hit enter at the console, it will advance you to the next step of the code. But it is good to explore the Console buttons (especially 'Next') to work through the code and watch the Environment (data and values) and Traceback as they change. By calling the function several times, you should be able to convince yourself of the cause of the error.

Much more detail about browsing in debugging mode is available at here.

When you are done, exit the debug mode by hitting the Stop button in the Console.

In R, you can insert the function `browser()` on some line of the code to enter debugging mode. This is also what you have to use if you want to debug directly in R Markdown.

Programming Challenge

Avian influenza cases in humans usually arise from two viral subtypes, H5N1 and H7N9. An interesting observation is that the age distributions for H5N1 and H7N9 cases differ: older people are more likely to get very sick and die from H7N9, and younger people from H5N1. There's no evidence for age-related differences in exposure. A recent paper showed that the risk of severe infection or death with avian influenza from 1997-2015 could be well explained by a simple model that correlated infection risk with the subtype of seasonal (non-avian) influenza a person was first exposed to in childhood. Different subtypes (H1N1, H2N2, and H3N2) have circulated in different years. Perhaps because H3N2 is more closely related to H7N9 than to H5N1, people with primary H3N2 infections seem protected from severe infections with H7N9. The complement is true for people first infected with H1N1 or H2N2 and later exposed to H5N1.

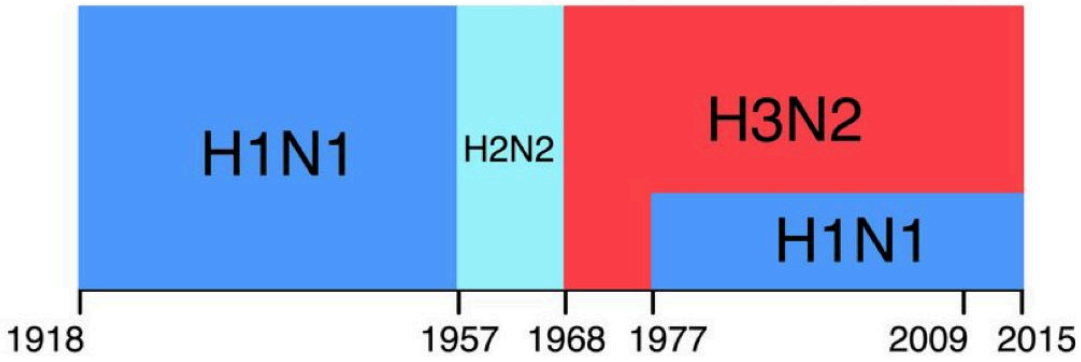


Figure 1: *Endemic influenza subtypes since 1918 (Gostic et al. 2016).*

Of course, we do not know the full infection history of any person who was hospitalized with avian influenza. We only know the person's age and the year of hospitalization or death. To perform their analysis, the authors needed to calculate the probability that each case had a primary infection with each subtype, i.e., the probability that a person born in a given year was first infected with each subtype. Your challenge is to calculate these probabilities.

The authors had to make some assumptions. First, they assumed that the risk of influenza infection is 28% in each year of life. Second, they assumed that the frequency of each circulating subtype could be inferred from the numbers of isolates sampled (primarily in hospitals) each year. These counts are given in `subtype_counts.csv`. [1]

The challenge: For every year between 1960 and 1996, calculate the probability that a person born in that year had primary infection with H1N1, H2N2, and H3N2. You must program defensively to pull this off.

When you have found the solution, go to stefanoallesina.github.io/BSD-QBio3 and follow the link `Submit solution to defensive programming challenge` to submit your answer (alternatively, you can go directly to goo.gl/forms/NjdZUyAjs9HBWMqL2).

[1] The counts are actually given for each influenza season in the U.S., which is slightly different from a calendar year, but you can ignore this. You'll notice that "1" and "0" are used where we know (or will assume) that only one subtype was circulating. The authors made several other assumptions, but this is good enough for now.

Data Wrangling and Visualization

Peter Carbonetto & John Novembre

August 30, 2017

Introduction

In this tutorial, we will use the ggplot2 package to create effective visualizations of biological data. The ggplot2 package is a powerful set of plotting functions that extend the base plotting fuctions in R.

ggplot2 will also introduce you to a different way of thinking about data visualization, based on the Grammar of Graphics. *Don't be discouraged if it takes time to get used to ggplot2 and the Grammar of Graphics. This is very common!*

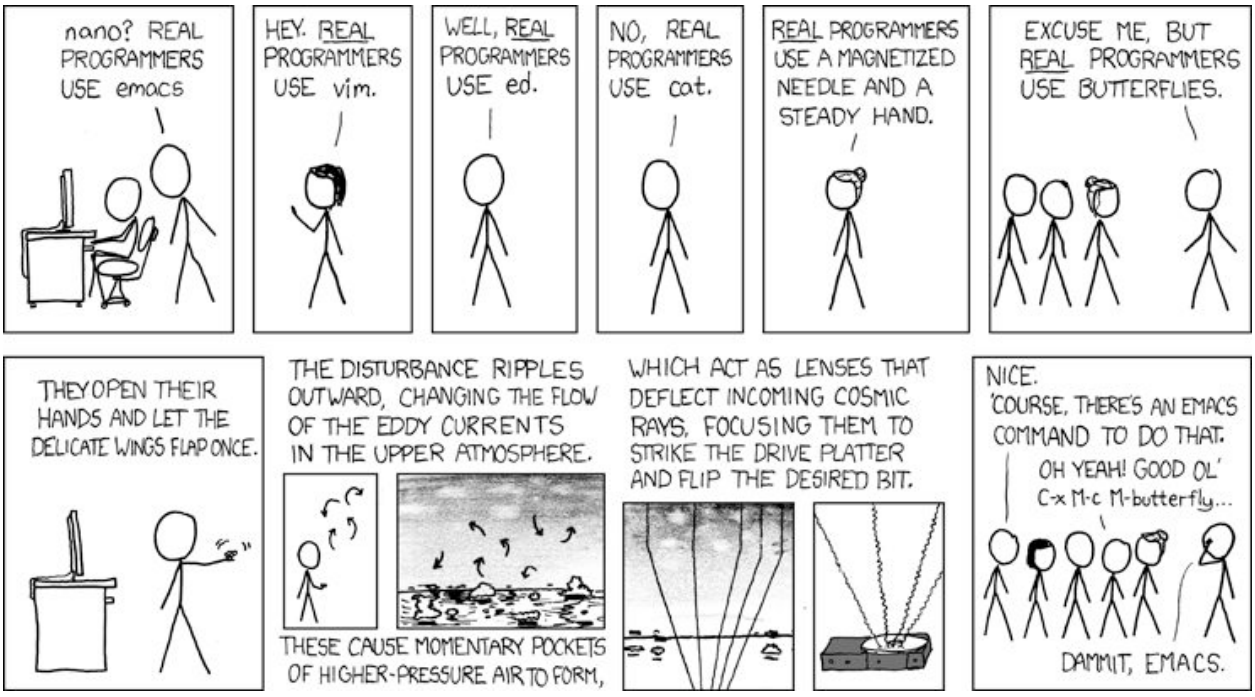
Another benefit of ggplot2 is that many help materials are available.

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 we can still illustrate some useful data preparation practices in this workshop.

Note

This tutorial demonstrates some plotting approaches that have worked well for our research. It is important to remember, however, that there is rarely a single best approach, and you may later discover other approaches that work better for you. The R landscape is continually evolving; learning from your peers is a great way to continue to improve your R programming practice along with your research practice.

Remember, *there is no best tool—use whatever works for you.*



Background

In this tutorial, we will investigate bone-mineral density (BMD) in a population of mice known as CFW mice. (CFW is short for "Carworth Farms White", the names of the scientists who bred the first CFW mice.) In particular, we will investigate results from a genome-wide association analysis of BMD, and create plots from these results to identify genetic contributors to variation in bone-minteral density.

Setup

To follow the examples in this tutorial, you should have a reasonably up-to-date copy of R or RStudio running on your computer.

You also need to install several packages:

```
install.packages(c("ggplot2","devtools","cowplot","ggcorrplot",
                   "htmlwidgets","plotly"))
```

Once you have successfully installed the packages, load ggplot2.

```
library(ggplot2)
```

Next, we load the phenotype data from the CFW study. These data have been compiled from the Data Dryad repository for distribution as a R package. If you have not done so already, install this R package using devtools:

```
library(devtools)
install_github("pcarbo/cfwlab")
```

It may take a few minutes to install this package.

Load the package and the phenotype data:

```
library(cfwlab)
data(cfw.pheno)
```

Take a look at the data set. Each row of the `cfw.pheno` data frame is an individual, and each column represents a measurement or experimental treatment on the individual.

```
head(cfw.pheno)
```

Check the size of the data set:

```
nrow(cfw.pheno)
ncol(cfw.pheno)
```

Today, we'll start by looking at a column in this table that is relevant to understanding human skeletal diseases such as osteoporosis—bone-mineral density.

The distribution of bone-mineral density in CFW mice

Our first step is to get a visual overview of bone-mineral density in the mice we are studying. The values are stored in the column `BMD` of the dataframe. The units of BMD are mg/cm^2. (Note that this is not volumetric density—"areal" BMD is easier to measure, and is considered an accurate approximation to the true BMD.)

We can easily get a *text-based* summary with the `summary` function:

```
summary(cfw.pheno$BMD)
```

The simplest way to create a *graphical* summary is using the `quickplot` function:

```
p <- quickplot(x = cfw.pheno$BMD,
               xlab = "Bone mineral density (mg/cm^2)",
               ylab = "number of mice") +
  theme(axis.line = element_blank())
print(p)
```

When provided with a single column of numbers in the first argument, `quickplot` automatically produces a histogram. The arguments `xlab` and `ylab` specify the x and y-axis labels.

Note that ggplot2 was thoughtful enough to give us a warning about rows (samples) that were not plotted because they contained missing data.

One way ggplot2 differs from other plotting systems is that the ggplot2 functions all have a return value. This return value is a "ggplot" object:

```
class(p)
```

The plot is drawn on the screen only after invoking `print(p)`, in which `p` is the return value. (Implicitly, the `print` function is invoked whenever you type a variable name on its own, so typing `p` and `print(p)` do the same thing.)

Also, in this example we added a *layer* to the plot—a "theme layer" that alters the the overall appearance of the plot (it removes the axis lines to make a cleaner plot). In ggplot2, *all plots are built up in this way by combining plotting layers, and all layers are combined using the + operator.*

The long right-hand "tail" in the histogram seems interesting. This is an example of a "skewed distribution"—specifically it is skewed to the right, sometimes known as a positive skew. In the next section, we will create a few plots in an attempt to identify some of the genetic determinants of this high BMD.

Mapping the genetic basis of osteopetrotic bones

To map loci for excessive mineralization, we created a binary trait called "abBMD" (short for "abnormal BMD") that signals whether an individual had abnormal, or *osteopetrotic*, bones. It takes a value of 1 when BMD falls on the "long tail" of the observed distribution (BMD value greater than 90), and 0 otherwise:

```
summary(cfw.pheno$abBMD)
```

We used GEMMA to carry out a "genome-wide association study" (GWAS) for this trait; that is, we computed association p-values from the genotype data available for 79,824 SNPs on chromosomes 1–19. Let's load these data and look at the top of the data frame.

```
data(cfw.gwscan)
head(cfw.gwscan)
```

Each row is a single genetic variant in the mouse genome (a "single nucleotide polymorphism", or "SNP"), and the columns give the chromosome ("chr"), position ("pos") and a series of p-values for a test of association between variant genotype and trait value, for 8 different traits.

Our first step is to get a overview of the association results for abnormal BMD. Instead of using `quickplot` like we did before, here we model a more powerful way of creating plots using ggplot2's layering approach.

First, we prepare the data. We create a small data frame (ggplot2 always works best with data frames) with the marker number and the the BMD p-values. These will be our plotting data.

```
n    <- nrow(cfw.gwscan)
pdat <- data.frame(chr = cfw.gwscan$chr,marker = 1:n,
                   pval = cfw.gwscan$abBMD)
```

Next, let's make a simple plot with these data, and later we will see if we can improve on it.

```
p <- ggplot(data = pdat,mapping = aes(x = marker,y = pval)) +
       geom_point(color = "darkblue",shape = 20)
print(p)
```

Although this example is small, it illustrates the basic elements of a plot in ggplot2:

1. A *data frame* containing the data we want to plot.
2. An *aesthetic mapping* describing how columns of the data frame are used draw the plotting features (axes, shapes, colors, *etc.*).
3. A *geom*, short for "geometric object," that defines how the combination of data and aesthetic mapping are rendered. In short, it controls the type of plot you create.

All the plots we will create consist of different combinations of these elements.

This is not a bad first attempt—the plot shows that there are two regions in the mouse genome exhibiting very strong support for association with abnormal BMD. Also notice that in each region there are *many* strongly associated SNPs. This is a common situation, and is known as linkage disequilibrium (LD). It arises as a natural consequence of low recombination rates between markers in small populations.

One limitation of this plot is that isn't clear *where* the association signals are located. It would be nice if we could add chromosome information to this plot (notice that we do have this information in the `chr` column of the `cfw.gwscan` data frame).

One very simple way to achieve this is to vary the colour of the points according to the chromosome number:

```
p <- ggplot(data = pdat,mapping = aes(x = marker,y = pval,color = chr)) +
       geom_point(shape = 20)
print(p)
```

Suddenly we have a much more colourful plot, but it requires quite a bit of squinting to find the chromosome numbers (usually a sign that the plot could be improved).

A better way to do this—and the approach most often used for plots in GWAS papers—is to show the chromosome numbers underneath the x-axis. First, we use the `tapply` function to find the mean position of each chromosome on the x-axis:

```
x.chr <- tapply(pdat$marker,pdat$chr,mean)
print(x.chr)
```

Next, we add a new layer, `scale_x_continuous`, to customize the x-axis:

```
p <- ggplot(data = pdat,mapping = aes(x = marker,y = pval)) +
       geom_point(color = "darkblue",shape = 20) +
       scale_x_continuous(breaks = x.chr,labels = 1:19)
print(p)
```

Now we have the chromosome numbers in our plot.

This example illustrates why using a *programmatic approach* to plotting is helpful—by creating plots within the R programming environment, we have a wide array of functions at our disposal to filter, manipulate, summarize and combine the data, allowing us to generate an endless variety of plots.

However, it isn't yet clear from the plot where one chromosome ends and the other begins. This was better achieved with the rainbow-like plot above. Instead of a rainbow, let's add a column which will allow us to plot the chromosomes in alternating colors:

```
pdat <- transform(pdat,odd.chr = (as.numeric(chr) %% 2) == 1)
```

Next, we map the `odd.chr` column to the colour aesthetic, just like we did above.

```
p <- ggplot(data = pdat,mapping = aes(x = marker,y = pval,color = odd.chr)) +
  geom_point(shape = 20) +
  scale_x_continuous(breaks = x.chr,labels = 1:19)
print(p)
```

ggplot2 is smart enough to figure out that only two colours are needed because we are plotting only two different values (even and odd).

This example also demonstrates that creating sophisticated plots in ggplot requires relatively little effort *provided the data are in the right form.*

We now clearly see that the two strongest association signals are on chromosomes 5 and 11.

To mimic the colors used in a typical GWAS paper, we add a `scale_color_manual` layer that customizes the color mapping. Also, ggplot2 adds legends by default, but we clearly don't need a legend in this case, so let's remove it.

```
p <- p + scale_color_manual(values = c("skyblue","darkblue"),guide = "none")
print(p)
```

This example illustrates another important feature of ggplot2: rather than re-type the plotting code from scratch, we were able to adjust an existing plot by adding a new layer to an existing ggplot object.

Finally, let's make a few more adjustments to our plot following Best Practices. Specifically, we add more informative axis labels, add a title, remove the axis lines and unneeded axis ticks, and change the theme. Again, let's make these adjustments by adding layers to the existing ggplot object. Here we use the default theme provided by the cowplot package, which extends ggplot2 (cowplot was developed by biologist Clause Wilke). *Note:* the cowplot theme becomes the default after the package is loaded, so the `theme_cowplot` call is only needed it you would like to change the font.

```
library(cowplot)
p <- p +
  scale_y_continuous(breaks = seq(0,14,2)) +
  labs(x = "chromosome",y = "-log10 p-value",
       title = "genome-wide scan for osteopetrotic BMD") +
  theme_cowplot(font_size = 11) +
  theme(axis.line = element_blank(),axis.ticks.x = element_blank())
print(p)
```

Without too much effort, we have reproduced a publication-quality "Manhattan plot" showing the results of our genome-wide association analysis.

How to save and share your plot

There are several ways to save a figure. The best approach will depend on the aim.

For exploratory analyses, GIF and PNG are great formats because the files are easy to attach to emails or webpages and they can be viewed nearly universally. Let's save our final Manhattan plot as a PNG using the `ggsave` function:

```
ggsave("gwscan-bmd.png",plot = p,dpi = 150)
```

The best resolution (dpi = dots per inch) usually requires some trial and error, but somewhere between 100 and 200 dpi is typically sufficient for screen viewing.

Please go ahead and share this plot with one of your neighbours (e.g., by sending them an email attachment, or sharing a link to a file on Dropbox or Google Drive).

For print or publication, a GIF or PNG is usually not going to cut it; it is almost always preferrable to save the plot in a vector graphics format. Currently the most widely adopted vector graphics format is PDF:

```
ggsave("gwscan-bmd.pdf",plot = p)
```

If you open this PDF in your favourite PDF viewer (e.g., Adobe Acrobat Reader), you will see that the edges remain sharp even at very high zoom levels, in contrast to the PNG file.

Please go ahead and share this PDF with one of your neighbours.

Examining the association signal on chromosome 11

Above, our plots gave us a *genome-wide* overview of the regions of the genome contributing to variation in BMD. Next, let's try to get a finer-scale view of the association signal on chromosome 11.

First, we prepare the data. Let's create a data frame with the association results on chromosome 11 only. This can be easily done using the `subset` function:

```
pdat <- subset(cfw.gwscan,chr == 11)
```

Next, convert the positions to Megabases (Mb) by dividing by 1 million (`1e6`). This will make the base-pair positions easier to read in the plots.

```
pdat <- transform(pdat,pos = pos/1e6)
```

Now we can easily plot the p-values for chromosome 11 only, like we did it before.

```
p <- ggplot(data = pdat,mapping = aes(x = pos,y = abBMD)) +
       geom_point(color = "darkblue",shape = 20)
print(p)
```

Let's make a few improvements to this plot following Best Practices:

```
p <- ggplot(data = pdat,mapping = aes(x = pos,y = abBMD)) +
       geom_point(color = "darkblue",shape = 20) +
       labs(x = "base-pair position on chromosome 11 (Mb)",
            y = "-log10 p-value") +
       theme(axis.line = element_blank()) +
       scale_x_continuous(breaks = seq(0,120,10)) +
       scale_y_continuous(limits = c(0,15),breaks = seq(0,15,5))
print(p)
```

The strongest association signal appears to be located somewhere between 90 and 100 Mb on chromosome 11. It would be nice if we could "zoom in" on this region to get a detailed view of the association signal.

This can be done by creating an *interactive* visualization of the same data. There are several very sophisticated R packages for creating interactive visualizations. A great place to start is the plotly package because it works well with ggplot2.

The `ggplotly` function converts a ggplot object to a "plotly" object, then we use the `saveWidget` function from the `htmlwidgets` package to embed the interactive plot in an HTML file.

```
library(plotly)
library(htmlwidgets)
p.plotly <- ggplotly(p,tooltip = c("pos","abBMD"))
saveWidget(p.plotly,"plotly.html",selfcontained = TRUE)
```

Go ahead and open the webpage in your favourite browser.

A key feature of plotly is the "tooltip"; it allows you to mouse over individual data points and quickly access more details. In this example, the base-pair position and p-value are displayed in the tooltip.

Alternatively, if you are creating an R Markdown notebook, invoking the variable name in a code chunk will automatically embed the interactive plot within any webpage generated from the R Markdown notebook.

```
p.plotly
```

Examining correlation patterns within the BMD locus

In this section, we will focus on the association results between 94 Mb and 98 Mb—roughly, the region spanning the strongest association signal on chromosome 11.

In addition to getting a higher resolution view of the association signal, typically an important analysis step is to study the correlation pattern ("linkage disequilibrium") among the markers. We will explore two different ways of doing this.

First, let's zoom in on this region and plot only the markers between 94 Mb and 98 Mb.

```
pdat <- subset(pdat,pos > 94 & pos < 98)
p <- ggplot(data = pdat,mapping = aes(x = pos,y = abBMD)) +
       geom_point(color = "darkblue",shape = 20,size = 2)
print(p)
```

Now we have a higher resolution view of the association signal within this region.

Next, calculating correlations between the markers is going to involve some more intensive programming because we will compare data in a data frame (`pdat`) against a large matrix (`cfw.geno`). So bear with us for a moment.

In this next chunk, we load the genotype data, when we compute the correlations between the top SNP (the SNP with the smallest p-value) and all the other SNPs in the region. Note that squared correlations are computed by convention in GWAS. Most of the work is done in the last line, which uses function `cor` to compute correlations between the top markers and all other markers in the 95–98 Mb region.

```
data(cfw.geno)
markers    <- pdat$id
top.marker <- markers[which.max(pdat$abBMD)]
pdat$r2    <- c(cor(cfw.geno[,top.marker],cfw.geno[,markers]))^2
```

We now have a new column in the data frame, `r2`, containing the squared correlations with the top SNP:

```
head(pdat)
summary(pdat$r2)
```

Having done the work to prepare the data in a data frame, adding the correlations to the plot is straightforward—all we need to do is adjust the aesthetic mapping.

```
p <- ggplot(data = pdat,mapping = aes(x = pos,y = abBMD,color = r2)) +
       geom_point(size = 2)
print(p)
```

This time, ggplot2 recognizes that the correlations are continuously-valued, and automatically draws the points using a colour *gradient* instead of discrete colours.

However, the default choice of colors is not great, so let's choose the colours carefully following these recommendations. Similar to before, we add a "scale layer" to adjust the colour mapping. Let's also make a few other improvements to the plot.

```
clrs <- c("midnightblue","darkviolet","darkorchid","maroon",
          "tomato","orange","gold","yellow")
p <- p + scale_color_gradientn(colors = clrs) +
       scale_x_continuous(breaks = 90:104) +
       scale_y_continuous(limits = c(0,15),breaks = seq(0,14,2)) +
       labs(x = "base-pair position on chromosome 11 (Mb)",
            y = "-log10 p-value") +
       theme(axis.line = element_blank())
print(p)
```

A careful choice of colors can make the difference between an unsatisfactory plot and a more effective one.

This sort of plot is very common in GWAS for visualizing fine-scale patterns of LD.

Here, we clearly see that almost all the strongest associations are highly correlated with the top SNP. It is also interesting that there are some less strongly correlated SNPs near 95 Mb that also have very small p-values; these observations suggest that another SNP in this region may be independently associated with abnormal BMD.

One question that is still unanswered is why the association p-values drop off very suddenly at around 95 Mb, and again around 97 Mb. To better understand what is happening in this region of the genome, let's visualize the correlations between pairs of SNPs, not just the top SNP. Since visualizing correlations between >500 SNPs is difficult, let's select a small subset of SNPs that will hopefully capture the broader patterns.

```
n       <- nrow(pdat)
rows    <- seq(1,n,length.out = 32)
markers <- pdat$id[rows]
R       <- cor(cfw.geno[,markers])
```

R is a 32 x 32 matrix containing the correlations between the 32 selected SNPs in the region.

The ggcorrplot package provides a very easy-to-use interface for plotting correlation matrices:

```
library(ggcorrplot)
ldplot <- ggcorrplot(R)
print(ldplot)
```

The plot is a bit messy so let's make a few adjustments to the plot:

```
ldplot <- ggcorrplot(R,colors = c("red","white","red")) +
  scale_x_discrete(breaks = NULL) +
  scale_y_discrete(breaks = NULL) +
  theme(axis.ticks = element_blank(),axis.line = element_blank())
print(ldplot)
```

The sign of the correlation doesn't matter for studying LD (it is arbitrary based on the genotype encoding), so we adjusted the colours so that correlations of -1 and 1 are both red.

The one problem with this plot is that it is difficult to relate to the p-values without any position information. So let's add this information to the plot in the "scale" layer for the x-axis. Because ggcorrplot is based on ggplot2, we can take the same layering to make adjustments to the plot's appearance.

```
pos <- round(pdat$pos[rows],digits = 2)
ldplot <- ggcorrplot(R,colors = c("red","white","red")) +
  scale_x_discrete(labels = pos) +
  scale_y_discrete(breaks = NULL) +
  theme(axis.text.x = element_text(size = 8,color = "black"),
        axis.ticks  = element_blank(),
        axis.line   = element_blank())
print(ldplot)
```

The striking result here is that the "block" of strong correlations aligns very closely with the p-value plot; in particular, there is a large block of strongly correlated SNPs between 95 Mb and 97 Mb. This explains why we see so many strong associations with BMD; many SNPs are strongly correlated with the variant (or variants) contributing to variation in BMD.

This "block pattern" is very common, and is due to recombination not occurring uniformly across the genome. (We will explore this further in one of the exercises.)

Identifying candidate BMD genes

In this section, we will attempt to identify promising BMD gene candidates from the genetic association results. One go-to resource for this is the UCSC Genome Browser. Here, we will instead incorporate the gene annotation data directly into our plot. In so doing, we will demonstrate how to combine multiple data sets into one plot.

First, we load the gene annotations for release 38 of the Mouse Genome Assembly. These annotations were compiled from a file downloaded from the RefSeq FTP site. The `start` and `end` columns give the base-pair positions where transcription starts and ends.

```
data(genes.m38)
head(genes.m38)
```

We are only interested in the genes overlapping the 94–98 Mb region on chromosome 11. Let's create a new data frame containing the annotations for these genes only. Also, let's adjust the units of the start and end positions to Megabases (Mb) like we did for the SNP positions.

```
pdat.genes <- subset(genes.m38,chr == 11 & start < 98e6 & end > 94e6)
pdat.genes <- transform(pdat.genes,
                        start = start/1e6,
                        end = end/1e6)
print(pdat.genes)
```

We now have a data frame containing the 105 genes that overlap the region of interest.

To create the "gene track", let's plot the gene symbols at the start positions using `geom_text`. To avoid gene names overlapping each other in a jumbled mess, let's vary the vertical position of the genes by creating a new column `vpos`, and setting the y-axis position to `vpos`:

```
n <- nrow(pdat.genes)
pdat.genes$vpos <- rep(1:12,length.out = n)
geneplot <- ggplot(pdat.genes,aes(x = start,y = vpos,label = gene.symbol)) +
  geom_text(size = 3,hjust = 0,fontface = "italic")
```

Notice that we added a "label" mapping.

The cowplot package has a useful function `plot_grid` for arranging multiple plots into a single figure.

```
print(plot_grid(p + theme(legend.position = "none"),geneplot,nrow = 2))
```

We have a nice visual of the genes in and nearby the BMD locus.

The two plots don't align perfectly, even after removing the legend in the first plot. This may work for sharing your results with your colleagues, but is likely not acceptable for a publication.

Let's try a different strategy: let's add the gene symbols to the plot we created earlier. This is possible because ggplot2 allows you to specify the data set and the aesthetic mapping within the geom function call. To make sure that the gene symbols appear below the SNP p-values, let's draw the genes below the y=0 line. There is a lot going on in this code chunk, so let's walk through it carefully.

```
combinedplot <- p + geom_text(data = pdat.genes,inherit.aes = FALSE,
                              aes(x = start,y = -vpos,label = gene.symbol),
                              hjust = 0) +
  scale_y_continuous(limits = c(-12,15))
print(combinedplot)
```

Now we have the SNP p-values and gene symbols combined into a single plot.

To make this work well, we had to fix a couple of technical issues: adjust the limits of the y axis (note that ggplot2 complains about this, but allows us to do it anyway); set `inherit.aes = FALSE` so that the aesthetic mapping used to draw the p-values isn't also used to draw the genes.

Let's go the extra mile and adjust some of the plotting settings to make this plot more clear: make the text a little smaller, draw the gene symbols in italics (which is the convention for genes), and remove the y-axis ticks below the y=0 line.

```
combinedplot <- p + geom_text(data = pdat.genes,inherit.aes = FALSE,
                              aes(x = start,y = -vpos,label = gene.symbol),
                              size = 3,hjust = 0,fontface = "italic") +
  scale_y_continuous(limits = c(-12,15),breaks = seq(0,14,2))
print(combinedplot)
```

What collagen-related gene lies in the region that might be involved in this genotype-phenotype association? (See the exercises below.)

Visualizing the relationship between genotype and phenotype

Above, we identified rs29477109 as the SNP most strongly associated with abnormal BMD. In this section, we will look closely at the relationship between BMD and the SNP genotype.

First, we prepare the data. We create a new table with two columns: one for the phenotype data, and one for the genotype data (more precisely, the genotype "dosage").

```
pdat <- data.frame(BMD = cfw.pheno$BMD,dosage = cfw.geno[,top.marker])
```

Our first plot is a simple scatterplot using the "point geom".

```
p <- ggplot(data = pdat,mapping = aes(x = dosage,y = BMD)) +
  geom_point(shape = 20,size = 2,color = "darkblue")
print(p)
```

Notice that ggplot2 gives us a warning about missing values in the data frame. These points are not plotted.

This plot suggests some trend, but it is hard to make out because many of the points overlap. Perhaps it would be helpful to calculate and plot the linear trend.

The `lm` function is the "swiss army knife" for linear regression in R. Here we'll use it to fit a linear regression of BMD given genotype dosage.

```
fit <- lm(BMD ~ dosage,pdat)
summary(fit)
```

Let's add the fitted regression line to the plot. We use the `coef` function to pull out the slope and intercept of the linear regression from the `lm` output, and the "abline geom" to draw an orange, dashed line.

```
p <- p + geom_abline(slope = coef(fit)["dosage"],
                     intercept = coef(fit)["(Intercept)"],
                     linetype = "dashed",color = "darkorange")
print(p)
```

It is now a little more apparent that larger dosage values correspond to lower BMD (on average). But the effect does not appear to be very large. Indeed, in the "lm" summary above we sw that only 7.3% of variance in BMD is explained by this SNP.

Let's take a different visualization strategy by treating the genotype dosage as a discrete (or "categorical") variable. From the `cfw.map` data frame, we see that the two alleles for this variant are T and C, so the possible genotypes are TT (dosage = 0), TC and CT (dosage = 1) and CC (dosage = 2).

```
data(cfw.map)
subset(cfw.map,id == top.marker)
```

To convert the dosages to a discrete variable, we round the dosages to the maximum-probability genotype, and convert the numbers 0, 1 and 2 to genotypes TT, TC and CC.

```
pdat <- transform(pdat,genotype = factor(round(dosage)))
levels(pdat$genotype) <- c("TT","CT","CC")
summary(pdat$genotype)
```

As expected (because it is a SNP with a high MAF), most of the the mice are heterozygous CT at this SNP.

Let's try a "box plot" first, which is the canonical plot type for comparing a continuous variable and a discrete variable. A box plot is easy to create in ggplot2 because `geom_boxplot` automatically computes the relevant statistics for you.

```
p <- ggplot(data = pdat,mapping = aes(x = genotype,y = BMD)) +
  geom_boxplot(width = 0.5,na.rm = TRUE)
print(p)
```

We prefer the "violin plot" because it is easier to interpret and shows the full density, rather than the box plot's arbitrary quantiles.

```
p <- ggplot(data = pdat,mapping = aes(x = genotype,y = BMD)) +
  geom_violin(na.rm = TRUE)
print(p)
```

Let's create a new plot combining the benefits of the violin and box plots. And let's make sure we are following Best Practices by giving informative labels and titles.

```
p <- ggplot(data = pdat,mapping = aes(x = genotype,y = BMD)) +
  geom_violin(na.rm = TRUE,fill = "skyblue",color = "skyblue") +
  geom_boxplot(width = 0.1,outlier.shape = NA,na.rm = TRUE)
p <- p + labs(y = "BMD (mg/cm^2)",title = top.marker)
print(p)
```

Finally, here is a "cleaner" alternative using the `stat_summary` function. This gives us more flexibility, but requires a slightly more advanced understanding of ggplot2 to use correctly.

```
p <- ggplot(data = pdat,mapping = aes(x = genotype,y = BMD)) +
  geom_violin(na.rm = TRUE,fill = "skyblue",color = "skyblue") +
  stat_summary(fun.y = median,fun.ymin = function (x) quantile(x,0.1),
               fun.ymax = function (x) quantile(x,0.9),
               color = "black",geom = "pointrange",na.rm = TRUE)
print(p)
```

One personal favourite is the box-percentile plot from the Hmisc package, but sadly this is not currently an option in ggplot2.

The box plot and the violin plot yield a richer picture of the relationship between phenotype and genotype. In particular, we see that the TT genotype is associated with a broader distribution of high bone-mineral densities. This relationship is more striking once we visualize the full distribution.

An important take-away message from this section is that it often requires experimentation with different plotting strategies to come up with a good visualization.

Programming challenges

Please go to https://goo.gl/forms/4UyqkBlvaulKosSZ2 to submit your solutions to the programming challenges.

Note: These challenges are ordered in increasing level of complexity. Do not be discouraged if you have difficulty completing all the exercises—please ask the instructors for advice if you get stuck.

Part 1: Follow-up on the BMD locus

1. Look up the BMD locus on chromosome 11 in the UCSC Genome Browser and identify candidate BMD genes based on the association signal. What collagen-related gene lies in the region that might be involved in this genotype-phenotype association?

Part 2: Explore and interpret other GWAS results

The `cfw.gwscan` data frame contains association results for abnormal BMD and 7 other traits.

2. Use the `help` function in R to read the cfwlab package documentation and identify all the mouse traits that might be useful for understanding musculoskeletal diseases in humans.
3. What p-value does a -log10(p-value) of 5 correspond to? a) 0.001 b) 0.0001 c) 0.00001 d) 0.000001
4. Out of all the traits (other than "abnormal BMD"), which trait has the strongest support for a genetic association?
5. How many genetic associations for soleus muscle weight have "strong" statistical support, specifically with log10(p-values) > 6? How many distinct regions of the genome are strongly associated with soleus muscle weight at this p-value threshold?
6. Using the plotting techniques presented above, identify a "quantitative trait locus" (QTL) with a strong association signal and identify gene(s) lying near the association signal that might explain the association.

Part 3: Devise a visualization to compare the distribution of BMD in CFW mice verus HMDP mice

7. Compare the distribution of BMD in the CFW population against BMD measured in the HMDP panel to better understand whether "osteopetrotic" BMD might be particular to CFW mice. Run `data(cfw.pheno)`and `data(hmdp.pheno)` to load the data from the cfwlab package for this exercise. Is the "long tail" of high BMD particular to the CFW population? *Advice:* There are many ways to compare histograms. Above, we arranged multiple plots in a single figure and combined multiple data sets into a single plot. Another strategy is to combine the data sets into a single data frame, and use define the aesthetic mapping in a clever way to draw the histograms in one plot.

Part 4: Add a "recombination track" to the BMD locus plot (most difficult)

8. In the *Identifying candidate BMD genes* section, we created a "gene track" to accompany the BMD locus plot. The objective of this exercise is to add recombination rate data below the BMD locus plot. Run `data(recomb.m38)` to load the the recombination rate data from the cfwlab package. To what extent to the estimated recombination rates align with the observed correlations? In particular, do recombination "hotspots" explain the sudden drop-offs in p-values around 95 Mb and 97 Mb? See Brunschwig *et al* for some interesting background on recombination hotspots in the mouse genome.

Statistics for a data rich world—some explorations

Stefano Allesina

Statistics for large data sets

- **Goal:** More and more often we need to analyze large and complex data sets. However, the statistical methods we've been taught in college have evolved in a data-poor world. Modern biology requires new tools, which can cope with the new questions and methods that arise in a data-rich world. Here we are going to discuss problems that often arise in the analysis of large data sets. We're going to review hypothesis testing (and what happens when we have many hypotheses) and discuss model selection. We're going to see the effects of selective reporting and p-hacking, and how they contribute to the *reproducibility crisis* in the sciences.
- **Audience:** Biologists with some programming background.
- **Installation:** To complete the tutorial, we will need to install `R` and `RStudio`. To install `R`, follow instructions at `cran.rstudio.com`. Then install `Rstudio` following the instructions at `goo.gl/a42jYE`.

Review of hypothesis testing

The basic idea of hypothesis testing is the following: we have devised an hypothesis on our system, which we call H_1 (the "alternative hypothesis"). We have collected our data, and we would like to test whether the data is consistent (or not) with the so-called "null-hypothesis" (H_0), which is associated with a contradiction to the hypothesis we would like to prove.

The simplest example is that of a bent coin: we believe that our coin favors heads (H_1: the coin is bent). We therefore toss the coin several times and check whether the number of heads we observe is consistent with the null hypothesis of a fair coin (H_0).

In `R` we can toss many coins in no time at all. Call `p` the probability of obtaining a head, and initially toss a fair coin (`p = 0.5`) a thousand times:

```
p <- 0.5 # probability of a head (fair coin)
flips <- 1000 # number of times we flip the coin
data <- sample(c("H", "T"),
                flips, prob = c(p, 1 - p),
                replace = TRUE)
heads <- sum(data == "H")
```

If the coin is fair, we expect approximately 500 heads, but of course we might have small variations due to the randomness of the process. We therefore need a way to distinguish between "bad luck" and an incorrect hypoythesis.

What is customarily done is to compute the probability of recovering the observed or a more extreme version of the pattern under the null hypothesis: if the probability is very small, we reject the null hypothesis (note that this does not guarantee that the alternative hypothesis is true). We call this probability a *p-value*.

For example, if the coin is fair, the number of heads should follow the binomial distribution. The probability of observing a larger number of heads than what we've got is therefore

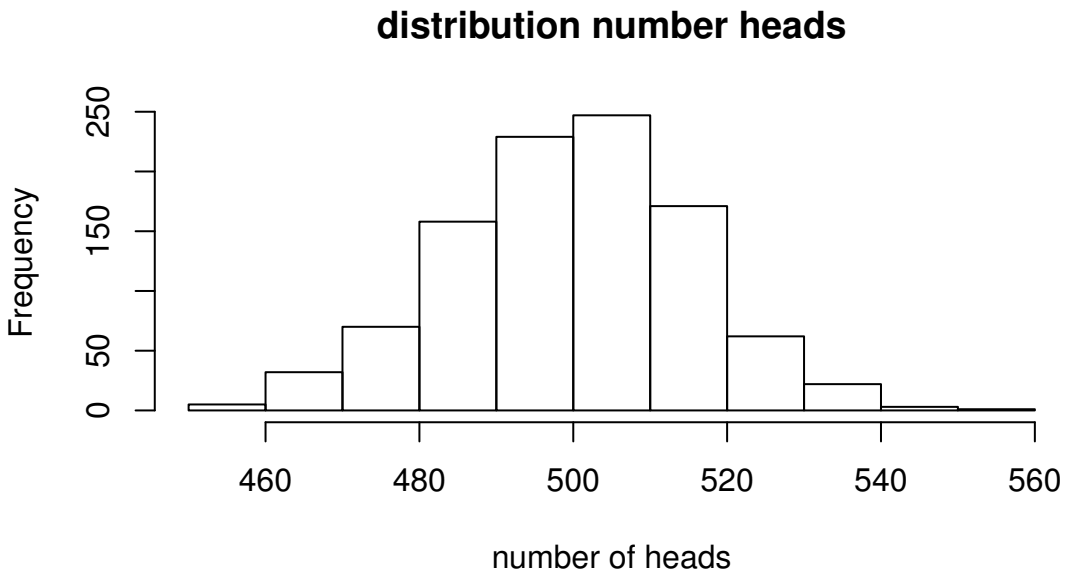
```
pvalue <- 1 - pbinom(heads, flips, 0.5)
```

which also suggests a faster way to flip the coins and count the heads:

```
heads <- rbinom(1, flips, p)
pvalue <- 1 - pbinom(heads, flips, 0.5)
```

What if we repeat the tossing many, many times?

```
# flip 1000 coins 1000 times
# produce histogram of number of heads
heads_distribution <- rbinom(1000, flips, p)
hist(heads_distribution, main = "distribution number heads", xlab = "number of heads")
```



You can see that it is very unlikely to get more than 540 (or less than 460) heads when flipping a fair coin 1000 times. Therefore, if we were to observe say 400 heads (or 800), we would tend to believe that the coin is biased (though of course this could have happened by chance!).

Errors

When testing an hypothesis, we can make two types of errors:

- **Type I error**: reject H_0 when it is in fact true
- **Type II error**: fail to reject H_0 when in fact it is not true

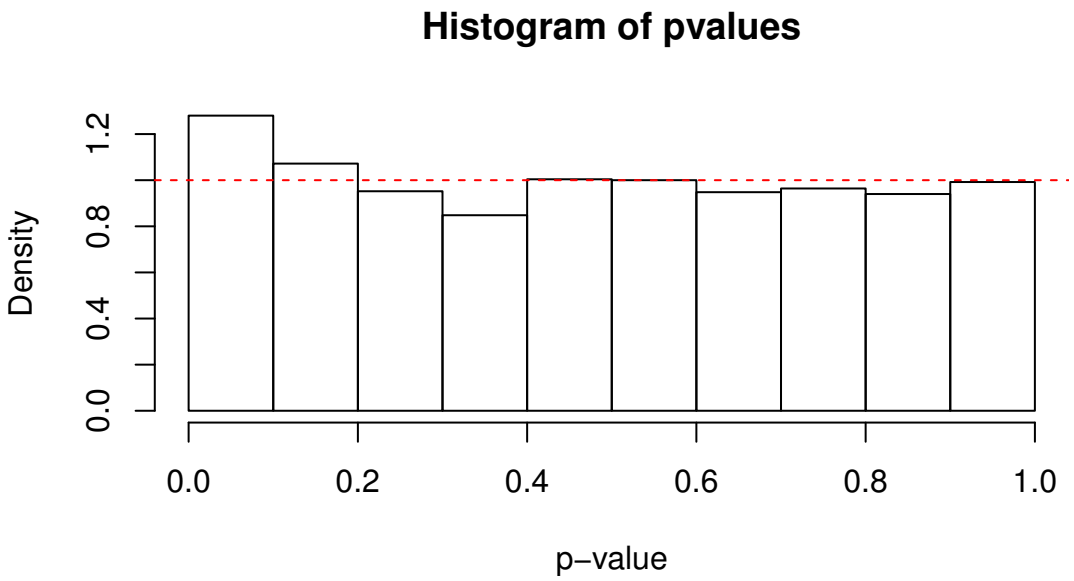
We call α the probability of making an error of type I, and β that of making a type II error. In practice, we tend to choose a quite stringent α (say, 0.01, 0.05), and then try to control for β as best as we can. What is typically reported is the smallest level of significance leading to the rejection of the null hypothesis (p-value). Therefore, the p-value quantifies how strongly the data contradicts the null hypothesis.

The distribution of p-values

Suppose that we are tossing each of several fair coins 1000 times. For each, we compute the corresponding p-value under the hypothesis `p = 0.5`. How are the p-values distributed?

```
ncoins <- 2500
heads <- rbinom(ncoins, flips, p)
pvalues <- 1 - pbinom(heads, flips, 0.5)
```

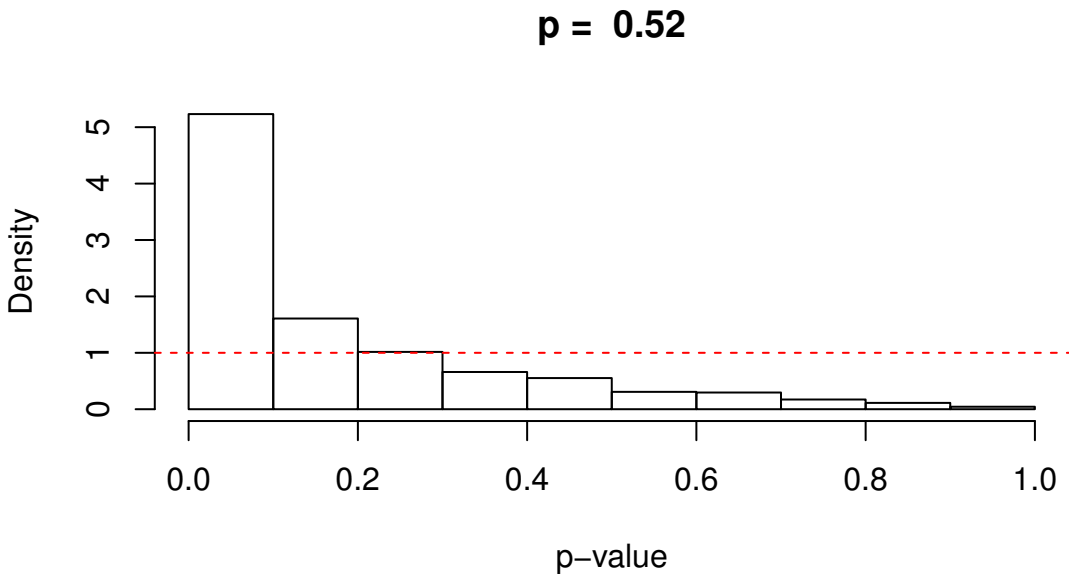
```r
hist(pvalues, xlab = "p-value", freq = FALSE)
abline(h = 1, col = "red", lty = 2)
```



As you can see, if the data were generated under the null hypothesis, the distribution of the p-values would be approximately uniform between 0 and 1. This means that if we set $\alpha = 0.05$, we would reject the null hypothesis 5% of the time (even though in this case we know the hypothesis is correct!).

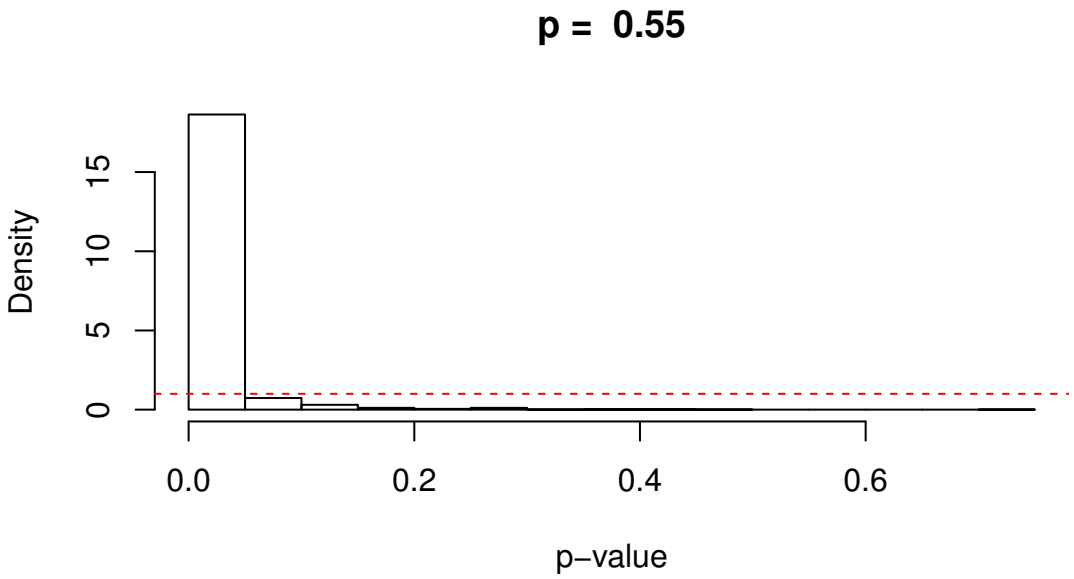
What is the distribution of the p-values if we are tossing biased coins? We will find an enrichment in small p-values, with stronger effects for larger biases:

```r
p <- 0.52 # the coin is biased
heads <- rbinom(ncoins, flips, p)
pvalues <- 1 - pbinom(heads, flips, 0.5)
hist(pvalues, xlab = "p-value", main = paste0("p =  ", p), freq = FALSE)
abline(h = 1, col = "red", lty = 2)
```




```r
p <- 0.55 # the coin is biased
heads <- rbinom(ncoins, flips, p)
```

```r
pvalues <- 1 - pbinom(heads, flips, 0.5)
hist(pvalues, xlab = "p-value", main = paste0("p =  ", p), freq = FALSE)
abline(h = 1, col = "red", lty = 2)
```



Problem: selective reporting

Articles reporting positive results are easier to publish than those containing negative results. Authors might have little incentive to publish negative results, which could go directly into the file-drawer.

This tendency is evidenced in the distribution of p-values in the literature: in many disciplines, one finds a sharp decrease in the number of tests with p-values just above 0.05 (which is customarily–and arbitrarily–chosen as a threshold for "significant results"). For example, we find many a sharp decrease in the number of reported p-values of 0.051 compared to 0.049–while we expect the p-value distribution to decrease smoothly.

Selective reporting leads to irreproducible results: we always have a (small) probability of finding a "positive" result by chance alone. For example, suppose we toss a fair coin many times, until we find a "signficant" result...

Problem: p-hacking

The problem is well-described by Simonsohn et al. (J. Experimental Psychology, 2014): "While collecting and analyzing data, researchers have many decisions to make, including whether to collect more data, which outliers to exclude, which measure(s) to analyze, which covariates to use, and so on. If these decisions are not made in advance but rather are made as the data are being analyzed, then researchers may make them in ways that self-servingly increase their odds of publishing. Thus, rather than placing entire studies in the file-drawer, researchers may file merely the subsets of analyses that produce nonsignificant results. We refer to such behavior as *p-hacking*."

The same authors showed that with careful p-hacking, almost anything can become significant (read their hylarious article in Psychological Science, where they show that listening to a song can change the listeners' age!).

Discussion on p-values

Selective reporting and p-hacking are only two of the problems associated with the widespread use and misuse of p-values. The discussion in the scientific community on this issue is extremely topical. I have collected some of the articles on this problem in the `readings` folder. Importantly, in 2016 the American Statistical Association released a statement on p-values every scientist should read.

Reproducibility crisis

P-values and hypothesis testing contribute considerably to the so-called *reproducibility crisis* in the sciences. A survey promoted by *Nature* magazine found that "More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments."

This problem is due to a number of factors, and addressing it will likely be one of the main goals of science in the next decade.

Exercise: p-hacking

Go to goo.gl/a3UOEF and try your hand at p-hacking, showing that your favorite party is good (bad) for the economy.

Multiple comparisons

The problem of multiple comparisons arises when we perform multiple statistical tests, each of which can (in principle) produce a significant result.

Suppose we perform our coin tossing exercise, flipping 1000 coins 1000 times each. For each coin, we determine whether our data differs significantly from what expected by contrasting our p-value with a significance level $\alpha = 0.05$.

Even if the coins are all perfectly fair, we would expect to find approximately $0.05 \cdot 1000 = 50$ coins that lead to the rejection of the null hypothesis.

In fact, we can calculate the probability of making at least one error of type I (reject the null when in fact it is true). This probability is called the Family-Wise Error Rate (FWER). It can be computed as 1 minus the probability of making no type I error at all. If we set $\alpha = 0.05$, and assume the tests to be independent, the probability of making no errors in m tests is $1 - (1 - 0.05)^m$. Therefore, if we perform 10 tests, we have about 40% probability of making at least a mistake; if we perform 100 tests, the probability grows to more than 99%. If the tests are not independent, we can still say that in general $FWER \leq m\alpha$.

This means that setting an α per test does not control for FWER.

Moving from tossing coins to biology, consider the following examples:

- **Gene expression** In a typical microarray experiment, we contrast the differential expression of tens of thousands of genes in treatment and control tissues.
- **GWAS** In Genomewide Association Studies we want to find SNPs associated with a given phenotype. It is common to test tens of thousands or even millions of SNPs for signficant associations.
- **Identifying binding sites** Identifying candidate binding sites for a transcriptional regulator requires scanning the whole genome, yielding tens of millions of tests.

Organizing the tests in a table

Suppose that we're testing m hypotheses. Of these, an unknown subset m_0 is true, while the remaining $m_1 = m - m_0$ are false. We would like to correctly call the true/false hypotheses (as much as possible). We can summarize the results of our tests in a table, of which the elements are unobservable:

	not rejected	rejected
H is True	U (true negatives)	V (false positives)
H is False	T (false negatives)	S (true positives)

What we would like to know is $m_1 = T + S$ and $m_0 = U + V$. Then V is the number of type I errors (rejected H when in fact it is true), and T is the number of type II errors (failed to reject a false H). However, we can only observe $V + S$ (the number of "discoveries"), and $U + T$ (number of "failures").

Our Per-Comparison Error Rate is $PCER = E[V]/m$ (where $E[X]$ stands for expectation), our Family-wise error rate is $P(V > 0)$. One quantity of interest is the False Discovery Rate (FDR), measured as the proportion of true discoveries $FDR = E[V/(V + S)]$ when $V + S > 0$. FDR measures the proportion of falsely rejected hypotheses.

Importantly $PCER \leq FDR \leq FWER$, meaning that when we control for FWER we're automatically controlling for the others, but not viceversa.

Bonferroni correction

One of the simplest and most widespread procedures to control for FWER is Bonferroni's correction. This procedure controls for FWER in the case of independent or dependent tests. It is typically quite conservative, especially when the tests are not independent (in practice, it becomes "too conservative" when the number of tests is moderate to high). Fundamentally, for a desired FWER α we choose as a (PCER) significance treshold α/m, where m is the number of tests we're performing. Equivalently, we can "adjust" the p-values as $q_i = \min(m \cdot p_i, 1)$, and call significant the values $q_i < \alpha$. In `R` it is easy to perform this correction:

```
original_pvals <- c(0.012, 0.06,  0.77, 0.001, 0.32)
adjusted_pvals <- p.adjust(original_pvals, method = "bonferroni")
print(adjusted_pvals)
```

```
## [1] 0.060 0.300 1.000 0.005 1.000
```

With these adjusted p-values, and an $\alpha = 0.05$, we would still single out as significant the fourth test, but not the first. The strength of Bonferroni is its simplicity, and the fact that we can perform the operation in a single step. Moreover, the order of the tests does not matter.

Other procedures

There are several refinements of Bonferroni's correction, some of which use the sequence of ordered p-values. For example, Holm's procedure starts by sorting the p-values in increasing order $p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq \dots p_{(m)}$. The hypothesis $H_{(i)}$ is rejected if $p_{(j)} \leq \alpha/(m - j + 1)$ for all $j = 1, \dots, i$. Equivalently, we can adjust the p-values as $q_{(i)} = \min(1, \max((m - i + 1)p_{(i)}, q_{(i-1)}))$. In this way, we use the most stringent threshold to determine whether the smallest p-value is significant, the next smallest p-value uses a slightly higher threshold and so on. For example, using the same p-values above:

```
original_pvals <- c(0.012, 0.06,  0.77, 0.001, 0.32)
adjusted_pvals <- p.adjust(original_pvals, method = "holm")
print(adjusted_pvals)
```

```
## [1] 0.048 0.180 0.770 0.005 0.640
```

We see that we would be calling the first test significant, contrary to what obtained with Bonferroni.

The function `p.adjust` offers several choices for p-value correction. Also, the package `multcomp` provides a quite comprehensive set of functions for multiple hypothesis testing.

Mix of coins

We're going to test these concepts by tossing repeatedly many coins. In particular, we're going to toss 1000 times 50 biased coins ($p = 0.55$) and 950 fair coins ($p = 0.5$). For each coin, we're going to compute a p-value, and count the number of type I, type II, etc. errors when using unadjusted p-values as well as when correcting using the Bonferroni or Holm procedure.

```
toss_coins <- function(p, flips){
  # toss a coin with probability p of landing on head several times
  # return a data frame with p, number of heads, pval and
  # H0 = TRUE if p = 0.5 and FALSE otherwise
  heads <- rbinom(1, flips, p)
  pvalue <- 1 - pbinom(heads, flips, 0.5)
  if (p == 0.5){
    return(data.frame(p = p, heads = heads, pval = pvalue, H0 = TRUE))
  } else {
    return(data.frame(p = p, heads = heads, pval = pvalue, H0 = FALSE))
  }
}

# To ensure everybody gets the same results, we're setting the seed
set.seed(8)
data <- data.frame()
# the biased coins
for (i in 1:50) data <- rbind(data, toss_coins(0.55, 1000))
# the fair coins
for (i in 1:950) data <- rbind(data, toss_coins(0.5, 1000))
# here's the data structure
head(data)
```

```
##      p heads         pval    H0
## 1 0.55   535 1.235282e-02 FALSE
## 2 0.55   558 1.061983e-04 FALSE
## 3 0.55   567 9.546428e-06 FALSE
## 4 0.55   532 1.988964e-02 FALSE
## 5 0.55   547 1.322765e-03 FALSE
## 6 0.55   574 1.178147e-06 FALSE
```

Now we write a function that adjusts the p-values and builds the table above

```
get_table <- function(data, adjust, alpha = 0.05){
  # produce a table counting U, V, T and S
  # after adjusting p-values for multiple comparisons
  data$pval.adj <- p.adjust(data$pval, method = adjust)
  data$reject <- FALSE
  data$reject[data$pval.adj < alpha] <- TRUE
  return(table(data[,c("reject","H0")]))
}
```

First, let's see what happens if we don't adjust the p-values:

```
no_adjustment <- get_table(data, adjust = "none", 0.05)
print(no_adjustment)
```

```
##        H0
## reject  FALSE TRUE
##   FALSE     2  905
##   TRUE     48   45
```

We correctly declared 48 of the biased coins "significant", but we also incorrectly called 2 biased coins "not significant" (Type II error). More worringly, we called 45 fair coins biased when they were not (Type I error). To control for the family-wise error rate, we can correct using Bonferroni:

```
bonferroni <- get_table(data, adjust = "bonferroni", 0.05)
print(bonferroni)
```

```
##        H0
## reject  FALSE TRUE
##   FALSE    40  950
##   TRUE     10    0
```

With this correction, we dramatically reduced the number of Type I errors (from 45 to 0), but at the cost of increasing Type II errors (from 2 to 40). In this way, we would make only 10 discoveries instead of 50.

In this case, Holm's procedure does not help:

```
holm <- get_table(data, adjust = "holm", 0.05)
print(holm)
```

```
##        H0
## reject  FALSE TRUE
##   FALSE    40  950
##   TRUE     10    0
```

More sophisticated methods, for example based on controlling FDR, can reduce the Type II errors, at the cost of a few Type I errors:

```
BH <- get_table(data, adjust = "BH", 0.05)
print(BH)
```

```
##        H0
## reject  FALSE TRUE
##   FALSE    17  946
##   TRUE     33    4
```

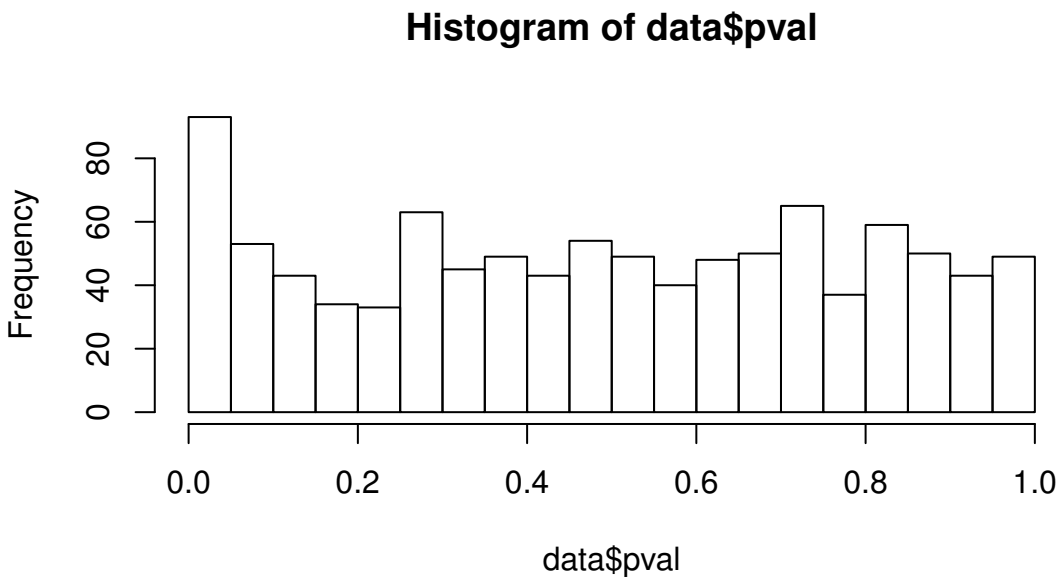
q-values

Inspired by the need for controlling for FDR in genomics, Storey and Tibshirani (PNAS 2003) have proposed the idea of a q-value, measuring the probability that a feature that we deemed significant turns out to be not significant after all.

One uses p-values to control for the false positive rate: when determining significant p-values we control for the rate at which null features in the data are called significant. The False Discovery Rate, on the other hand, measures the rate at which results that are deemed significant are truly null. While setting $PCER = 0.05$ we are stating that about 5% of the truly null features will be called significant, an $FDR = 0.05$ means that among the features that are called significant, about 5% will turn out to be null.

They proposed a method that uses the ensamble of p-values to determine the approximate (local) FDR. The idea is simple. If you plot your histogram of p-values when you have few true effect, and many nulls, you will see something like:

```
hist(data$pval, breaks = 25)
```



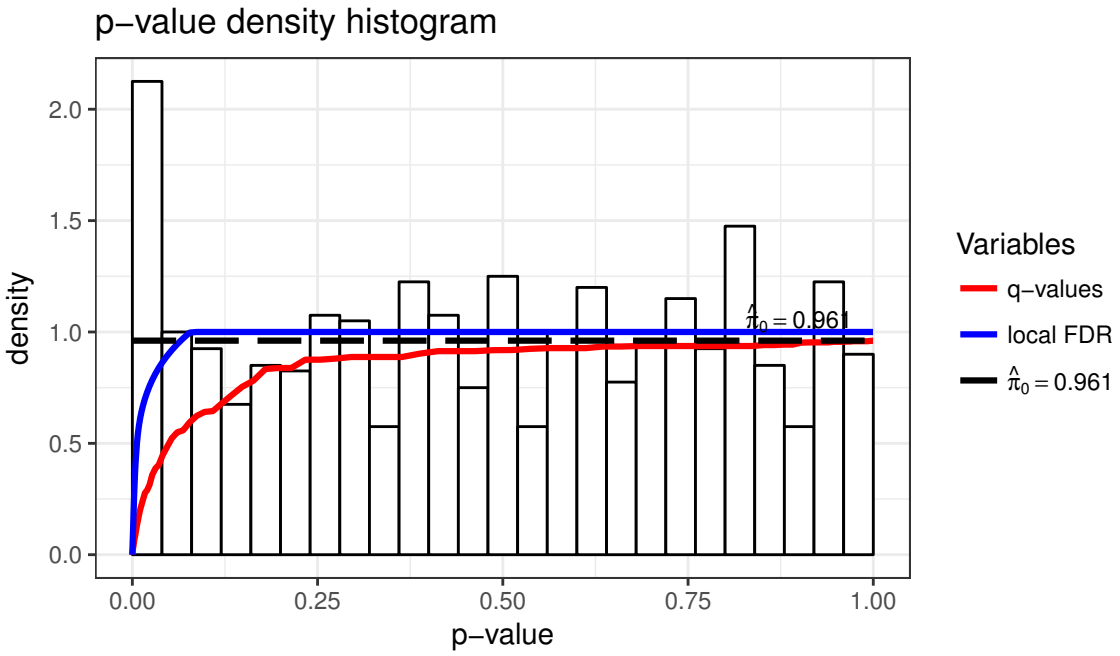
where the right side of the histogram is close to a uniform distribution. We could use the high p-values to find how tall the histogram would be if all effects were null, thereby estimating the proportion of truly null features $\pi_0 = m_0/m$.

Storey has built an R-package for this type of analysis:

```
# To install:
#install.packages("devtools")
#library("devtools")
#install_github("jdstorey/qvalue")
library("qvalue")
qobj <- qvalue(p = data$pval)
```

Here's the estimation of the π_0

```
hist(qobj)
```



which is quite good (in this case we know that $\pi_0 = 0.95$). The small p-values under the dashed line represent our false discoveries. Even better, through randomizations one can associate a q-value to each test, representing the probability of making a mistake when calling a result significant (formally, the q-value is the minimum FDR that can be attained when calling that test significant).

For example:

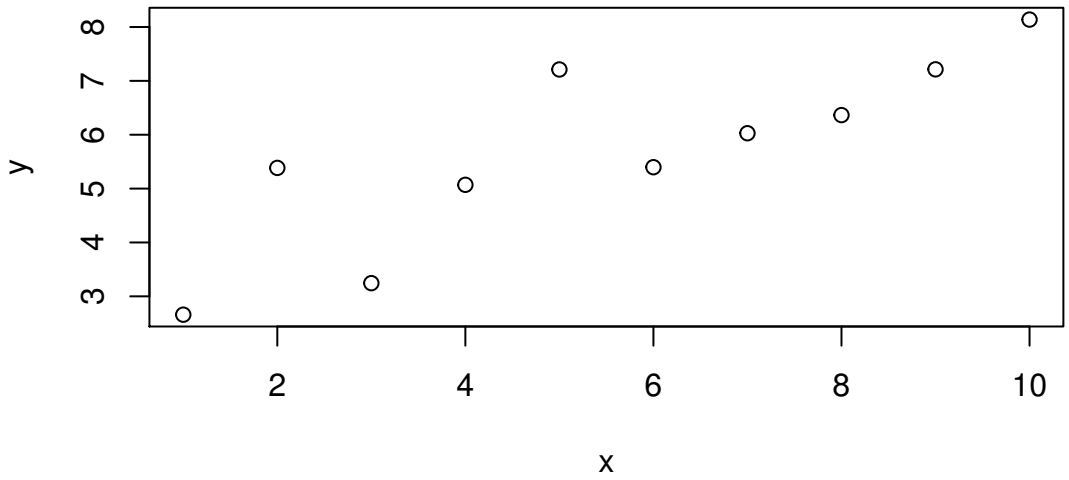
```
table((qobj$pvalues < 0.05) & (qobj$qvalues < 0.05), data$H0)
```

```
##
##         FALSE TRUE
##   FALSE    17  946
##   TRUE     33    4
```

Model selection

Often, we need to select a model out of a set of reasonable alternatives. However, we run the risk of overfitting the data (i.e., fitting the noise as well as the pattern). The simplest example is that of a regression:

```
# create fake data
set.seed(5)
x <- 1:10
y <- 3 + 0.5 * x + rnorm(10)
plot(y ~ x)
```



We can fit a linear regression to the data

```
model1 <- lm(y ~ x)
```

Or a more complex polynomial

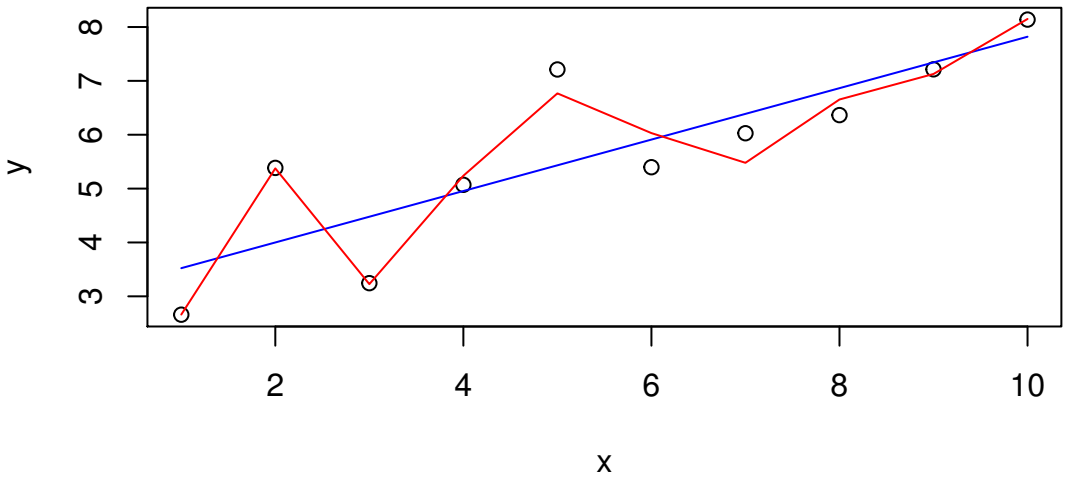
```
model2 <- lm(y ~ poly(x, 7))
```

Let's see the residuals etc.

```
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2330 -0.5096 -0.2437  0.2676  1.7790
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0452     0.6883   4.425  0.00221 **
## x             0.4774     0.1109   4.304  0.00260 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.007 on 8 degrees of freedom
## Multiple R-squared:  0.6984, Adjusted R-squared:  0.6607
## F-statistic: 18.53 on 1 and 8 DF,  p-value: 0.0026
```

```
#summary(model2)
plot(y~x)
points(model1$fitted.values~x, type = "l", col = "blue")
points(model2$fitted.values~x, type = "l", col = "red")
```



Our second model has a much greater R^2, but also many more parameters. Is the first model more parsimonious?

Model selection tries to address this and similar problems. Most model fitting and model selection procedures are based on likelihoods (e.g., Bayesian models, maximum likelihood, minimum description lenght). The likelihood $L(D|M, \theta)$ is (proportional to) the probability of observing the data D under the model M and parameters θ. Because likelihood can be very small when you have much data, typically one works with log-likelihoods. For example:

```
logLik(model1)
```

```
## 'log Lik.' -13.14838 (df=3)
```

```
logLik(model2)
```

```
## 'log Lik.' -2.773007 (df=9)
```

Typically, more complex models will yield better (less negative) log-likelihoods. We therefore want to penalize more complex models in some way.

Fox et al. (Research Integrity and Peer Review, 2017) analyzed the invitations to review for several scientific journals, and found that "The proportion of invitations that lead to a submitted review has been decreasing steadily over 13 years (2003–2015) for four of the six journals examined, with a cumulative effect that has been quite substantial". Their data is stored in `../data/FoxEtAl.csv`. We're going to build models trying to predict whether a reviewer will agree (or not) to review a manuscript.

```
# read the data
reviews <- read.csv("../data/FoxEtAl.csv", sep = "\t")
# take a peek
head(reviews)
```

```
##   Sort   Journal  msID Year ReviewerID ReviewerInvited ReviewerResponded
## 1    1 Evolution 34152 2007    8426852               1                 1
## 2    2 Evolution 34152 2007    8425970               1                 1
## 3    3 Evolution 34152 2007    8425116               1                 1
## 4    4 Evolution 34152 2007    8426128               1                 1
## 5    5 Evolution 34152 2007    9327585               1                 1
## 6    6 Evolution 34152 2007    8423528               1                 1
##   ReviewerAgreed ReviewerAssigned ReviewSubmitted
## 1              0                0              NA
## 2              0                0              NA
## 3              0                0              NA
```

```
## 4              0                0              NA
## 5              1                1               1
## 6              0                0              NA
```

```
# set NAs to 0
reviews[is.na(reviews)] <- 0
# how big is the data?
dim(reviews)
```

```
## [1] 113876     10
```

```
# that's a lot! Let's take 5000 review invitations for our explorations;
# we will fit the whole data set later
set.seed(101)
small <- reviews[order(runif(nrow(reviews))),][1:5000,]
```

Logistic regression

We will be playing with logistic regression. Call π_i the probability that a reviewer will agree to review manuscript i. We model $logit(\pi_i) = \log(\pi_i/(1-\pi_i))$ as a linear function. This type of regression (along with the probit) is often use to model binary response variables.

Constant rate

As a null model we build a model in which the probability to agree to review does not change in time/for journals:

```
# suppose the rate at which reviewers agree is a constant
mean(small$ReviewerAgreed)
```

```
## [1] 0.466
```

```
# fit a logistic regression
model_null <- glm(ReviewerAgreed~1, data = small, family = "binomial")
summary(model_null)
```

```
##
## Call:
## glm(formula = ReviewerAgreed ~ 1, family = "binomial", data = small)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.120  -1.120  -1.120   1.236   1.236
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.13621    0.02835  -4.805 1.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6908.3  on 4999  degrees of freedom
## Residual deviance: 6908.3  on 4999  degrees of freedom
## AIC: 6910.3
```

```
##
## Number of Fisher Scoring iterations: 3
```

```
# interpretation:
exp(model_null$coefficients[1]) / (1 + exp(model_null$coefficients[1]))
```

```
## (Intercept)
##       0.466
```

Declining trend

We now build a model in which the probability to review declines steadily from year to year:

```
# Take 2003 as baseline
model_year <- glm(ReviewerAgreed~I(Year - 2003), data = small, family = "binomial")
summary(model_year)
```

```
##
## Call:
## glm(formula = ReviewerAgreed ~ I(Year - 2003), family = "binomial",
##     data = small)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3441  -1.0879  -0.9686   1.2372   1.4017
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.383697   0.065482   5.860 4.64e-09 ***
## I(Year - 2003) -0.074753   0.008491  -8.804  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6908.3  on 4999  degrees of freedom
## Residual deviance: 6829.7  on 4998  degrees of freedom
## AIC: 6833.7
##
## Number of Fisher Scoring iterations: 4
```

Each year has its own parameter

What if the probability to agree were to vary from year to year, with no clear trend?

```
# Take 2003 as baseline
model_eachyr <- glm(ReviewerAgreed~as.factor(Year), data = small, family = "binomial")
#summary(model_eachyr)
```

Journal dependence

Reviewers might be more likely to agree for more prestigious journals:

```r
# Take the first journal as baseline
model_journal <- glm(ReviewerAgreed~Journal, data = small, family = "binomial")
summary(model_journal)
```

```
##
## Call:
## glm(formula = ReviewerAgreed ~ Journal, family = "binomial",
##     data = small)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.208  -1.110  -1.033   1.247   1.329
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.07187    0.06812   1.055  0.29141
## JournalFE    -0.27908    0.09411  -2.965  0.00302 **
## JournalJANIM -0.16544    0.09598  -1.724  0.08476 .
## JournalJAPPL -0.23319    0.09323  -2.501  0.01237 *
## JournalJECOL -0.26183    0.09403  -2.785  0.00536 **
## JournalMEE   -0.42223    0.12868  -3.281  0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6908.3  on 4999  degrees of freedom
## Residual deviance: 6892.7  on 4994  degrees of freedom
## AIC: 6904.7
##
## Number of Fisher Scoring iterations: 3
```

Model journal and year

Finally, we can build a model combining both features: we fit a parameter for each journal/year combination

```r
# Take the first journal as baseline
model_journal_yr <- glm(ReviewerAgreed~Journal:I(Year-2003),
                        data = small, family = "binomial")
#summary(model_journal_yr)
```

For completeness, the most complicated model, in which each journal and year combination has its own parameter.

```r
# Take the first journal as baseline
model_journal_eachyr <- glm(ReviewerAgreed~Journal:as.factor(Year),
                            data = small, family = "binomial")
#summary(model_journal_eachyr)
```

Likelihoods

In `R`, you can extract the log-likelihood from a model object calling the function `logLik`

```
logLik(model_null)
```

```
## 'log Lik.' -3454.167 (df=1)
```

```
logLik(model_year)
```

```
## 'log Lik.' -3414.845 (df=2)
```

```
logLik(model_eachyr)
```

```
## 'log Lik.' -3405.309 (df=13)
```

```
logLik(model_journal)
```

```
## 'log Lik.' -3446.335 (df=6)
```

```
logLik(model_journal_yr)
```

```
## 'log Lik.' -3395.365 (df=7)
```

```
logLik(model_journal_eachyr)
```

```
## 'log Lik.' -3365.024 (df=68)
```

Interpretation: because we're dealing with binary data, the likelihood is the probability of correctly predicting the agree/not agree for all the 5000 invitations considered. Therefore, the probability of guessing a (random) invitation correctly under the first model is:

```
exp(as.numeric(logLik(model_null)) / 5000)
```

```
## [1] 0.5011582
```

while the most complex model yields

```
exp(as.numeric(logLik(model_journal_eachyr))/ 5000)
```

```
## [1] 0.5101733
```

We didn't improve our guessing much by considering many parameters! This could be due to specific data points that are hard to predict, or mean that our explanatory variables are not sufficient to model our response variable.

AIC

One of the simplest methods to select among competing models is the Akaike Information Criterion (AIC). It penalizes models accoding to the **number of parameters**: $AIC = 2p - 2\log L(D|M, \theta)$, where p is the number of parameters. Note that **smaller** values of AIC stand for "better" models. In `R` you can compute AIC using:

```
AIC(model_null)
```

```
## [1] 6910.334
```

```
AIC(model_year)
```

```
## [1] 6833.691
```

```
AIC(model_eachyr)
```

```
## [1] 6836.618
```

```
AIC(model_journal)
```

```
## [1] 6904.669
```

```
AIC(model_journal_yr)
```

```
## [1] 6804.73
```

```
AIC(model_journal_eachyr)
```

```
## [1] 6866.049
```

As you can see, AIC would favor the `model_journal_yr` model.

AIC is rooted in information theory and measures (asymptotically) the loss of information when using the model instead of the data. There are several limitations of AIC: a) it only holds asymptotically (i.e., for very large data sets; for smaller data you need to correct it); it penalizes each parameter equally (i.e., parameters that have a large influence on the likelihood have the same weight as parameters that do no influence the likelihood much); it can lead to overfitting, favoring more complex models in simulated data generated by simpler models.

BIC

In a similar vein, BIC (Bayesian Information Criterion) uses a slightly different penalization: $BIC = \log(n)p - 2\log L(D|M,\theta)$, where n is the number of data points. Again, smaller values stand for "better" models:

```
BIC(model_null)
```

```
## [1] 6916.851
```

```
BIC(model_year)
```

```
## [1] 6846.725
```

```
BIC(model_eachyr)
```

```
## [1] 6921.341
```

```
BIC(model_journal)
```

```
## [1] 6943.772
```

```
BIC(model_journal_yr)
```

```
## [1] 6850.35
```

```
BIC(model_journal_eachyr)
```

```
## [1] 7309.218
```

Note that according to BIC, `model_year` is favored.

Cross validation

One very robust method to perform model selection, often used in machine learning, is cross-validation. The idea is simple: split the data in three parts: a small data set for exploring; a large set for fitting; a small set for testing (for example, 5%, 75%, 20%). You can use the first data set to explore freely and get inspired

for a good model. The data will be then discarded. You use the largest data set for accurately fitting your model(s). Finally, you validate your model or select over competing models using the last data set.

Because you haven't used the test data for fitting, this should dramatically reduce the risk of overfitting. The downside of this is that we're wasting precious data. There are less expensive methods for cross validation, but if you have much data, or data is cheap, then this has the virtue of being fairly robust.

Let's try our hand at cross-validation. First, we split the data into three parts:

```
reviews$cv <- sample(1:3, nrow(reviews), prob = c(0.05, 0.75, 0.2), replace = TRUE)
dataexplore <- reviews[reviews$cv == 1,]
datafit <- reviews[reviews$cv == 2,]
datatest <- reviews[reviews$cv == 3,]
# We've already done our exploration.
# Let's fit the data using model_journal
# and model_journal_yr, which seem to be the most promising
cv_model1 <- glm(ReviewerAgreed~I(Year-2003), data = datafit, family = "binomial")
cv_model2 <- glm(ReviewerAgreed~Journal:I(Year-2003), data = datafit, family = "binomial")
```

Now that we've fitted the models, we can use the function `predict` to find the fitted values for the `testdata`:

```
mymodel <- cv_model1
# compute probabilities
pi <- predict(mymodel, newdata = datatest, type = "resp")
# compute log likelihood
mylogLik <- sum(datatest$ReviewerAgreed * log(pi) +
                  (1 - datatest$ReviewerAgreed) * log(1 - pi))
print(mylogLik)
```

```
## [1] -15520.91
```

repeat for the other model

```
mymodel <- cv_model2
# compute probabilities
pi <- predict(mymodel, newdata = datatest, type = "resp")
# compute log likelihood
mylogLik <- sum(datatest$ReviewerAgreed * log(pi) +
                  (1 - datatest$ReviewerAgreed) * log(1 - pi))
print(mylogLik)
```

```
## [1] -15473.93
```

Cross validation supports the choice of the more complex model.

Other approaches

Bayesian models are gaining much traction in biology. The advantage of these models is that you can get a posterior distribution for the parameter values, reducing the need for p-values and AIC. The downside is that fitting these models is computationally much more expensive (you have to find a distribution of values instead of a single value).

There are three main ways to perform model selection in Bayesian models:

- **Reversible-jump MCMC** You build a Monte Carlo Markov Chain that is allowed to "jump" between models. You can choose a prior for the probability of being in each of the models; the posterior distribution gives you an estimate of how much the data supports each model. Upside: direct measure. Downside: difficult to implement in practice – you need to avoid being "trapped" in a model.

- **Bayes Factors** Ratio between the probability of two competing models. Can be computed analytically for simple models. Can also be interpreted as the average likelihood when parameters are chosen according to their prior (or posterior) distribution. Upside: straightforward interpretation — it follows from Bayes theorem; Downside: in most cases, one needs to approximate it; can be tricky to compute for complex models.
- **DIC** Similar to AIC and BIC, but using distributions instead of point estimates.

Another alternative paradigm for model selection is Minimum-Description Length. The spirit is that a model is a way to "compress" the data. Then you want to choose the model whose total description length (compressed data + description of the model) is minimized.

A word of caution

The "best" model you've selected could still be a terrible model (best among bad ones). Out-of-fit prediction (such as in the cross-validation above) can give you a sense of how well you're modeling the data.

When in doubt, remember the (in)famous paper in Nature (Tatem et al. 2004), which used some flavor of model selection to claim that, according to their linear regression, in the 2156 olympics the fastest woman would run faster than the fastest man. One of the many memorable letters that ensued reads:

> Sir — A. J. Tatem and colleagues calculate that women may out-sprint men by the middle of the twenty-second century (Nature 431,525; 2004). They omit to mention, however, that (according to their analysis) a far more interesting race should occur in about 2636, when times of less than zero seconds will be recorded.
>
> In the intervening 600 years, the authors may wish to address the obvious challenges raised for both time-keeping and the teaching of basic statistics.
>
> Kenneth Rice

Programming Challenge

Presidential tweets and crimes in Chicago

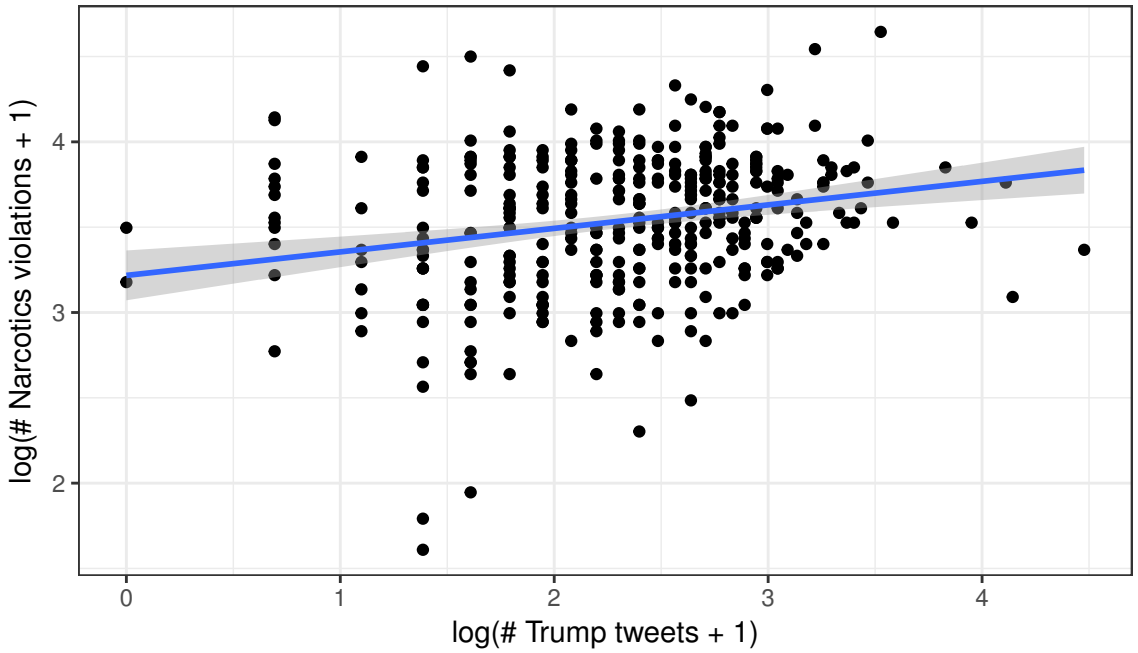
To try your hand at p-hacking and overfitting, and show how these practices can lead to completely inane results, you are going to show the strong correlation between the tweets by President Obama (or presidential candidate Donald Trump) in 2016 and the number of crimes in Chicago. For example, here's code showing that the number of tweets by Trump correlates with narcotics violations in Chicago.

```
library(dplyr)
library(ggplot2)
library(readr)
# read tweets
trump <- read_csv("../data/Trump_Tweets_2016.csv")
obama <- read_csv("../data/Obama_Tweets_2016.csv")
# read crimes
crimes <- read_csv("../data/Chicago_Crimes_2016.csv")

# count tweets by day
obama_all <- obama %>% group_by(day, month) %>% tally()
trump_all <- trump %>% group_by(day, month) %>% tally()

# join the data sets
crimes <- left_join(crimes, obama_all %>% rename(obama = n))
```

```
crimes <- left_join(crimes, trump_all %>% rename(trump = n))
# set crimes with 0 occurrences in a day
crimes[is.na(crimes)] <- 0

# take a look at the data
head(crimes)
```

```
## # A tibble: 6 × 6
##     day month PrimaryType     n obama trump
##   <int> <dbl>       <chr> <dbl> <dbl> <dbl>
## 1     1     1       ARSON     1     0    14
## 2     1     2       ARSON     0     0     4
## 3     1     3       ARSON     0     0    12
## 4     1     4       ARSON     2     0    10
## 5     1     5       ARSON     1     0     9
## 6     1     6       ARSON     1     0     3
```

```
# show that narcotics violations correlate with Trump's tweets
narco <- crimes %>% filter(PrimaryType == "NARCOTICS")
pl <- ggplot(narco) + aes(x = log(trump + 1), y = log(n + 1)) +
  theme_bw() + geom_point() + geom_smooth(method = "lm") +
  xlab("log(# Trump tweets + 1)") + ylab("log(# Narcotics violations + 1)")
show(pl)
```




```
# print correlation
print(cor.test(log(narco$trump + 1),log(narco$n + 1)))
```

```
##
##  Pearson's product-moment correlation
##
## data:  log(narco$trump + 1) and log(narco$n + 1)
## t = 4.4758, df = 364, p-value = 1.019e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.1289053 0.3233371
## sample estimates:
##       cor
## 0.2283974
```

You can see that the correlation is respectable (0.23), and that the p-value is minuscule (0.00001)! Try your hand at finding something even more striking. You can:

- transform the data (e.g., use logs, square roots, binning)
- combine multiple crimes
- use a specific date range
- summarize the data by week or by month
- use correlation (Pearson, Spearman, Kendall) or generalized linear models
- etc.

Once you've found something worth reporting in the front page of Chicago newspapers, post your answer at goo.gl/forms/ysqKxxcfogmUKmQq2

Workshops

Workshop Cobey
Workshop Novembre
Workshop Yang

Unraveling dynamics from time series

Instructors: Sarah Cobey and Marcos Costa Vieira

BSD Bootcamp on Quantitative Biology @ MBL

Goal

The aim of this workshop is to show how mathematical models can be used to investigate dynamical systems. We will begin by constructing and analyzing simple models of the spread of an infectious disease. We'll work with pencil and paper (or the whiteboard versions) before simulating the dynamics in R. We'll see that small nonlinearities can generate complex dynamics, which we can summarize with bifurcation diagrams. Finally, we'll discuss the use of likelihood in model fitting and review the criteria by which to judge a model.

At the end of the workshop, students should have facility analyzing simple models analytically and simulating systems of differential equations in R. They should also understand the steps involved in creating bifurcation diagrams. Conceptually, they should understand the spectrum of "mechanistic" to "statistical" models and the limitations of analyses based on correlations and linear assumptions. They should also be able to define the concept of the likelihood and rank factors to consider in judging the strength of a model.

Audience

This workshop is intended for biologists investigating nonlinear dynamical systems. The material is thus relevant to molecular and cell biologists, population geneticists, ecologists, immunologists, etc. The lecture assumes some familiarity with ordinary differential equations (ODEs), and the exercises assume familiarity with R. Although both ODEs and R will be covered to some extent in the tutorials, if you are eager to test your knowledge, you can inspect the code and suggested readings below.

Installation

Please have a recent version of R installed before the workshop. We will also use the deSolve package. It would be helpful if you could try installing this package before the workshop.

Workshop resources

- Slides
- Code
- Data: We will not be analyzing any real data as part of the workshop. However, if you have a time series in mind that you would like to focus on, please contact the instructors beforehand.

Readings and additional resources

These readings are not mandatory, but they review some of the techniques and models we will cover in the workshop.

- Lotka-Volterra Model (Wikipedia)
- Compartmental SIR Models (Wikipedia)
- Chaos theory (Wikipedia)
- May 1976: "Simple mathematical models with very complicated dynamics"
- Earn et al. 2000: "A simple model for complex dynamical transitions in epidemics"
- Shrestha et al. 2011: "Statistical inference for multi-pathogen systems"

Population genetics workshop

Instructor: **John Novembre**
Course Assistant: **Arjun Biddanda**

Welcome

This excercise is going to expose you to several basic ideas in probability and statistics as well as show you the utility of using R for basic statistical analyses. We'll do so in the context of a basic population genetic analysis.

The scenario

As a biologist, you will learn what are the major patterns that are expected when the data you work with is clean. Using that expertise will save you from the mistake of misinterpreting error-prone data. In population genetics, there are a number of patterns that we expect to see immediately in our datasets. In this exercise you will explore one of those major patterns. Rather than give it away — let's begin some analysis and see what we find. In the narative that follows, we'll refine our thinking as we go.

Introductory terminology

- Single-nucleotide polymorphism (SNP): A nucleotide basepair that is *polymorphic* (i.e. it has multiple types or *alleles* in the population)
- Allele: A particular variant form of DNA (e.g. A partciular SNP may have the "A-T" allele in one DNA copy and "C-G" in another; We typically define a reference strand of the DNA to read off of, and then denote the alleles according to the reference strand base - so for example, these might be called simply the "A" and "C" alleles. In many cases we don't care about the precise base, so we might call these simply the A_1 and A_2 alleles, or the A or a alleles, or the 0 and 1 alleles.)
- Minor allele: The allele that is more rare in a population
- Major allele: The allele that is more common in a population
- Genotype: The set of alleles carried by an individual (E.g. AA, AC, CC; or AA, AA, and aa; or 0/0, 1/1, 2/2)
- Genotyping array: A technology based on hybridization with probes and florensence that allows genotype calls to be made at 100s of thousands of SNPs per individual at an affordable cost.

The data-set and basic pre-processing

We will look at Illumina 650Y genotyping array data from the CEPH-Human Genome Diversity Panel. This sample is a global-scale sampling of human diversity with 52 populations in total.

The data were first described in Li et al (Science, 2008) and the raw files are available from the following link: http://hagsc.org/hgdp/files.html. These data have been used in numerous subsequent publications (e.g Pickrell et al, Genome Research, 2009) and are an important reference set. A few technical details are that the genotypes were filtered with a GenCall score cutoff of 0.25 (a quality score generated by the basic genotype calling software). Individuals with a genotype call rate <98.5% were removed, with the logic being that if a sample has many missing genotypes it may due to poor quality of the source DNA, and so none of the genotypes should be trusted. Beyond this, to prepare the data for the workshop, we have filtered down the indiviudals to a set of 938 unrelated individuals. (For those who are interested, the data are available as plink-formatted files `H938.bed` `H938.fam`, `H938.bim` from this link: http://bit.ly/1aluTln). We have also extracted the basic counts of three possible genotypes.

The files with these genotype frequencies are your starting points.

Note about logistics

We will use some of functions from the `dplyr` and `ggplot2` and `reshape2` libaries so first let's load them:

```
library(dplyr)
library(ggplot2)
library(reshape2)
```

If you get an error message saying there is no package titled `dplyr`,`ggplot2`, or `reshape2` you may need to first run `install.packages("dplyr")`,`install.packages("ggplot2")`, or`install.packages("reshape2")` to install the appropriate package.

We will not be outputting files - but you may want to set your working directory to the `sandbox` sub-directory in case you want to output some files.

The `MBL_WorkshopJN.Rmd` file has the R code that you can run. I provide code for most steps, but some you will need ot devise for yourselves to answer the questions that are part of the workshop narrative.

Initial view of the data

Read in the data table:

```
g <- read.table("../Data/H938_chr15.geno", header=TRUE)
```

It will be read in as a dataframe in R.

And use the "head" command to see the beginning of the dataframe:

```
head(g)
```

You should see that there are columns each with distinct names.

```
CHR         SNP A1 A2 nA1A1 nA1A2 nA2A2
```

- CHR: The chromosome number. In this case they are all from chromosome 2.
- SNP: The rsid of a SNP is a unique identifier for a SNP and you can use the rsid to look up information about a SNP using online resource such as dbSNP or SNPedia.
- A1: The minor allele at the SNP
- A2: The major allele
- nA1A1 : The number of A1/A1 homozygotes
- nA1A2 : The number of A1/A2 homozygotes
- nA2A2 : The number of A2/A2 homozygotes

Calculate the number of counts at each locus

Next compute the total number of observations by summing each of the three possible genotypes. Here we use the `mutate` function from the `dplyr` library to do the addition and add a new column to the dataframe in one nice step. (Note: You could also use the `colSums` function from the base R library).

```
g <- mutate(g, nObs = nA1A1 + nA1A2 + nA2A2)
```

Run `head(g)` and confirm your dataframe `g` has a new column called `nObs`.

Now use the `summary` function to print a simple summary of the distribution:

```
summary(g$nObs)
```

The `ggplot2` library has the ability to make "quick plots" with the command `qplot`. If we pass it a single column it will make a histogram of the data for that column. Let's try it:

```
qplot(nObs, data = g)
```

Our data are from 938 individuals. When the counts are less than this total, it's because some individuals had array data that was difficult to call a genotype for and so no genotype was reported.

Question: Do most of the SNPs have complete data?
Question: What is the lowest count observed? Is this number in rough agreement with what we know about how the genome-wide missingness rate filter was set to >98.5% of all SNPs

Calculating genotype and allele frequencies

Let's move on to calculating genotype and allele frequencies. For allele A_1 we will denote its frequency among all the samples as p_1, and likewise for A_2 we will use p_2.

```
# Compute genotype frequencies
g <- mutate(g, p11 = nA1A1/nObs , p12 = nA1A2/nObs, p22 = nA2A2/nObs )
# Compute allele frequencies from genotype frequencies
g <- mutate(g, p1 = p11 + 0.5*p12, p2 = p22 + 0.5*p12)
```

Question: With a partner or group member, discuss whether the equations in the code for p_1 and p_2 are correct and if so, why?

Run `head(g)` again and confirm `g` now has the extra columns for the genotype and allele frequencies.

And let's plot the frequency of the major allele (A2) vs the frequency of the minor allele (A1). The `ggplot2` library has the ability to make "quick plots" with the command `qplot`. Let's try it here:

```
qplot(p1, p2, data=g)
```

Notice that $p_2 > p_1$ (be careful to inspect the axes labels here) This makes sense because A_1 is supposed to be the minor (less frequent) allele. Note also that there is a linear relationship between p_2 and p_1

Question: What is the equation decribing this relationship?

The relationship exists because there are only two alleles - and so their proportions must sum to 1. The linear relationship you found exists because of this constraint. It also provides a nice check on our work (if p_1 and p_2 didn't sum to 1 it would suggest something is wrong with our code!).

Plotting genotype on allele frequencies

Let's look at an initial plot of genotype vs allele frequencies. We could use the base plotting functions, but the following uses the `ggplot2` commands. These are a little trickier, but end up being very compact (we need fewer lines of code overall to achieve our desired plot). To use `ggplot2` commands effectively our data need to be what statisticans call "tidy" (in this case, that means with one row per pair of points we will plot).

To do this, first we subset the data on the columns we'd like (using the `select` command and listing the set of columns we want), then we pass this (using the `%>%` operator) to the `melt` command which will reformat the data for us, and output it as `gTidy`:

```
gTidy <- select(g, c(p1,p11,p12,p22)) %>% melt(id='p1',value.name="Genotype.Proportion")
head(gTidy)
ggplot(gTidy) + geom_point(aes(x = p1,
                               y = Genotype.Proportion,
                               color = variable,
                               shape = variable))
```

Now let's look at the graph produced. There is some scatter in the relationship between genotype proportion and allele frequency for any given genotype, but at the same time there is a very regular underlying relationship between these variables.

Question: What are approximate relationships between p_{11} vs p_1, p_{12} vs p_1, and p_22 vs p_1? (Hint: These look like parabolas, which suggests are some very simple quadratic functions of p_1).

You might start to recognize that these are the classic relationships that are taught in introductory biology courses. If you recall, under assumptions that there is no mutation, no natural selection, infinite population size, no population substructure and no migration, then the genotype frequencies will take on a simple relationship with the allele frequencies. That is: $p_{11} = p_1^2$, $p_{12} = 2p_1(1-p_1)$ and $p_{22} = (1-p_1)^2$. In your basic texts, they typicaly use p and q for the frequencies of allele 1 and 2, and present these *Hardy-Wienberg proportions* as: p^2,$2pq$, and q^2.

Another way to think of the Hardy-Weinberg proportions is in the following way. If the state of an allele (A_1 vs A_2) is *independent* within a genotype, then the probability of a particular genotype state (such as A_1A_1) will be determined by taking the product of the alleles within it (so $p_{11} = p_1p_1$ or p_1^2).

Let's add to the plot lines that represent Hardy-Weinberg proportions:

```
ggplot(gTidy)+
  geom_point(aes(x=p1,y=Genotype.Proportion,color=variable,shape=variable))+
  stat_function(fun=function(p) p^2, geom="line", colour="red",size=2.5) +
  stat_function(fun=function(p) 2*p*(1-p), geom="line", colour="green",size=2.5) +
  stat_function(fun=function(p) (1-p)^2, geom="line", colour="blue",size=2.5)
```

On average, the data follow the classic theoretical expectations fairly well. It is pretty remarkable that such a simple theory has some bearing on reality!

By eye, we can see that the fit isn't perfect though. There is a systematic deficiency of heterozygotes and excess of homozygotes. Why?

Let's look at this more closely and more formally...

Testing Hardy Weinberg

Pearson's χ^2-test is a basic statistical test that can be used to see if count data conform to a particular expectation. It is based on the X^2-test statistic:

$$X^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

which follows a χ^2 distribution under the null hypothesis that the data are generated from a multinomial distribution with the expected counts given by e_i.

Here we compute the test statistic and obtain its associated p-value (using the `pchisq` function). We keep in mind that there is 1 degree of freedom (because we have 3 observations per SNP, but then they have to sum to a single total sample size, and we have to use the data once to get the estimated allele frequency, which reduces us down to 1 degree of freedom).

```
g <- mutate(g, X2 = (nA1A1-nObs*p1^2)^2 /(nObs*p1^2) +
            (nA1A2-nObs*2*p1*p2)^2 / (nObs*2*p1*p2) +
            (nA2A2-nObs*p2^2)^2 / (nObs*p2^2))
g <- mutate(g,pval = 1-pchisq(X2,1))
```

The problem of multiple testing

Let's look at the top few p-values:

```
head(g$pval)
```

How should we interpret these? A p-value gives us the frequency at which that the observed departure from expectations (or a more extreme departure) would occur if the null hypothesis is true. As an agreed upon standard (of the frequentist paradigm for statistical hypothesis testing), if the data are relatively rare under the null (e.g. p-value $< 5\%$), we reject the null hypothesi, and we would infer that the given SNP departs from Hardy-Weinberg expectations. This is problematic here though. The problem is that we are testing many, many SNPs (Use `dim(g)` to remind yourself how many rows/SNPs are in the dataset). Even if the null is universally true, 5% of our SNPs would be expected to be rejected using the standard frequentist paradigm. This is called the multiple testing problem. As an example, if we have 50,000 SNPs, that all obey the null hypothesis, we would on average naively reject the null for ~2500 SNPs based on the p-values < 0.05.

We clearly need some methods to deal with the "multiple testing problem". Two frameworks are the Bonferroni approach and false-discovery-rate (FDR) approaches. We will not say more about these here. Instead, we will do two simple checks to see though if our data are globally consistent with the null.

First, let's see how many tests have p-values less than 0.05. Is it much larger than the number we'd expect on average given the total number of SNPs and a 5% rate of rejection under the null?

```
sum(g$pval < 0.05, na.rm = TRUE)
```

Wow - we see many more. This is our first sign that though by eye these data show qualitiative similarities to HW, statistically they are not fitting Hardy-Weinberg well enough.

Let's look at this another way. A classic result from Fisher is that under the null hypothesis the p-values of a well-designed test should be distributed uniformly between 0 and 1. What do we see here?

```
qplot(pval, data = g)
```

The data show an enrichment for small p-values relative to a uniform distribution. Notice how the whole distribution is shifted towards small values - The data appear to systematically depart from Hardy-Weinberg.

Plotting expected vs observed heterozygosity

To understand this more clearly, let's make a quick plot of the expected vs observed heterozygosity (the proportion of heterozygotes):

```
qplot(2*p1*(1-p1), p12, data = g) + geom_abline(intercept = 0,
                                                slope=1,
                                                color="red",
                                                size=1.5)
```

Most of the points fall below the y=x line. That is, we see a systematic deficiency of heterozygotes (and this implies a concordant excess of homozygotes). This general pattern is contributing to the departure from HW seen in the X^2 statistics.

Discussion: Population subdivision and departures from Hardy-Weinberg expectations

We might wonder why the departure from Hardy-Weinberg proportional is directional, in that, on average, we are seeing a deficiency of heterozygotes (and excess of homozygotes). One enlightening way to understand this is by thinking about what Sewall Wright (a former eminent University of Chicago professor) called "the correlation of uniting gametes". To produce an A_1A_1 individual we need an A_1-bearing sperm and an A_1-bearing egg to unite. If these events were independent of each other, we would expect A_1A_1 individuals at the rate predicted by multiplying probabilites, that is, p_1^2 (an idea we introduced above). However, what if uniting gametes are positively correlated, in that an A-bearing sperm is more likely to join with an A-bearing egg? In this case we will have more A_1A_1 individuals than predicted by $2p_1^2$, and conversely fewer A_1A_2

individuals than predicted by $2p_1p_2$. If our population is structured somehow such that A_1 sperm are more likely to meet with A_1 eggs, then we will have such a positive correlation of uniting gametes, and the resulting excess of homozygotes and deficiency of heterozygotes.

Given the HGDP data is from 52 sub-populations from around the globe, and alleles have some some probability of clustering within populations, a good working hypothesis for the deficiency of heterozygotes in this dataset is the presence of some population structure.

While statistically significant, the population structure appears to be subtle in absolute terms — based on our plots, we have seen the genotype proportions are not wildly off from HW proportions.

Question: As an exercise, compute the average deficiency of heterozygotes relative to the expectated proportion. This is the average of

$$\frac{2p_1(1-p_1) - p_{12}}{2p_1(1-p_1)}$$

What is this number for this data-set? A common "rule-of-thumb" for this deficiency in a global sample of humans is approximately 10%. Do you find this to be true from the data?

A ~10% difference between expected and observed seems pretty remarkable given these samples are taken from across the globe. It is a reminder that human populations are not very deeply structured. Most of the alleles in the sample are globally widespread and not sufficiently geographically clustered to generate correlations among the uniting alleles. This is because all humans populations derived from an ancestral population in Africa around 100-150 thousand years ago, which is relatively small amount of time for variation across populations to accumulate.

Finding specific loci that are large departures from Hardy-Weinberg

Now, let's ask if we can find any loci that are wild departures from HW proportions. These might be loci that have erroneous genotypes, or loci that cluster geographically in dramatic ways (such that they have few heterozygotes relative to expectations).

To find these loci, we'll compute the same relative deficiency you computed above, but let's look at it per SNP. This number is referred to as F by Sewall Wright and has connections directly to correlation coefficients (advanced exercise: Try to work this out!). If we assume there is no inbreeding within populations, this number is an estimator of F_{ST} (a quantity that appears often in population genetics).

Let's add this value to our dataframe and plot how it's value changes across the chromosome from one end to another:

```
g <- mutate(g, F = (2*p1*(1-p1)-p12) / (2*p1*(1-p1)))
plot(g$F, xlab = "SNP number")
```

There are a few interesting SNPS that show either a very high or low F value.

Now, here's a trick. When a high or low F value is due to genotyping error, it likely only effects a single SNP. However, when there is some population genetic force acting on a region of the genome, it likely effects multiple SNPs in the region. So let's try to take a local average in a sliding window of SNPs across the genome, computing an average F over every 5 consecutive SNPs (in real data analysis we might use 100kb or 0.1cM windows).

The `stats::filter` command below calls the `filter` function from the `stats` library. The code above instructs the function to take 5 values centered on a focal SNP, weighting them each by 1/5 and then taking the sum. In this way it produces a local average in a sliding window of 5 SNPs. Let's define the `movingavg` function and then make a plot of its values:

```
movingavg <- function(x, n=5){stats::filter(x, rep(1/n,n), sides = 2)}
plot(movingavg(g$F), xlab="SNP number")
```

Wow — there appears to be one large spike where the average F is approximately 60% in the dataset!

Let's extract the SNP id for the largest value, and look at the dataframe:

```
outlier=which (movingavg(g$F) == max(movingavg(g$F),na.rm=TRUE))
g[outlier,]
```

Question: Which SNP is returned? By inserting the rs id into the UCSC genome browser (https://genome.ucsc.edu/), and following the links, find out what gene this SNP resides near. The gene names should start with "SLC.." What gene is it?

Question: Carry out a literature search on this gene using the term "positive selection" and see what you find. It's thought the high F value observed here is because natural selection led to a geographic clustering of alleles in this gene region. Discuss with your partners why this might or might not make sense.

Discussion: The outlier reigon

The region you've found is one of the most differentiated betwen human populations that is known. Notice in your literature search, how it is known to affect skin pigmentation and is thought to contribute to differences in skin pigmentation that are seen between human populations. Finding strong population structure for alleles that affect external morhological phenotypes is not uncommon when looking at other chromosomes. Some of the most differentiated genes that exist in humans are those that involve morphological phenotypes - such as skin pigmentation, hair color/thickness, and eye color (the genes OCA2/HERC2, SCL45A2, KITLG, EDAR all come to mind). Many of these are thought to have arisen due to direct or indirect effects of adaptation to local selective pressures (e.g. adaptation to varing levels of UV exposure, local pathogens, local diets, local mating preferences), though in most cases we still do not yet have a fully convincing understanding of their evolutionary histories. Regardless of the reasons, it is notable that many of the features that humans see externally in each other (i.e. the morphological differences) are controlled by genes that are outliers in the genome. At most variant SNPs, the patterns of variation are much closer to those of a single random mating populations than they are at variant sites like EDAR. Put another way, a genomic perspective shows us many of the differences people see in each other are in a sense, just skin-deep.

Wrap-up

Modern population genetics has a lot of additional tools on its workbench, but here using relatively simple and classical ideas combined with genomic-scale data, we have been able to observe and interpret some major features of human genetic diversity. We have also revisited some basic concepts of probablity and statistics such as indepedence vs correlation, the χ^2 test, and the problems of multiple testing. One remarkable thing we saw is that a very simple mathematical model based on assuming independence of alleles of genotypes can predict genotype proportions within ~10% of the true values. This gives us a hint of how simple mathematical models may be useful even in the face of biological complexity. Finally, we have gained more familiarity with R. We didn't discuss how genotyping errors that might create Hardy-Weinberg departures, but if we were doing additional analyses, we could use Hardy-Weinberg departures to filter them from our data. It's common practice to do so, but with a Bonferonni correction and using data from within populations to do the filtering.

Follow-up activities

In the 'addons' folder, we are including data files that you can explore to gain more experience. These include global data for other chromosomes (`H938_chr*.geno`) and the same data but limited to European populations (`H938_Euro_chr*.geno`). Here are a few suggested follow-up activities. It may be wise to split the activities across class members and reconvene after carrying them out.

Follow-up actvity: Look at a chromosome from the European-restricted data - is the global deficiency in heterozygosity as strong as it was on the global scale? Before you begin, what would you expect to see?

Follow-up activity Using the European data, do you find any regions of the genome that are outliers for F on chromosome 2? Using genome browsers and/or literature searches, can you find what is the likely locus under selection for that region?

Follow-up activity: Using the global data or the European data, analyze other chromosomes – do you find other loci that show high F values?

References

Li, Jun Z, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, et al. 2008. "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation." Science 319 (5866): 1100–1104.

Pickrell, Joseph K, Graham Coop, John Novembre, Sridhar Kudaravalli, Jun Z Li, Devin Absher, Balaji S Srinivasan, et al. 2009. "Signals of Recent Positive Selection in a Worldwide Sample of Human Populations." Genome Research 19 (5): 826–37.

NGS Data Analysis Workshop

Instructor: Lixing Yang, Ph.D.
Ben May Department of Cancer Research
lixingyang@uchicago.edu

Prerequisites:

Operating System: MacOS, Linux (Ubuntu, Fedora, etc.), or Windows with Cygwin installed to mimic a Linux environment.

Java: JRE v1.6 or above.

Linux skill: Familiar with entry level Linux commands: ls, cd, cat, grep, sort, uniq, wc, cut, more, less.

Statistics: Basic knowledge, R installed, basic knowledge of R.

Tutorial:

This tutorial is to demonstrate the analysis steps from getting the raw sequencing data from service provider to delivering actionable mutations to physicians. The sequencing data are from colorectal cancer patients.

0. Preparation

0.a. Data. All data and required software packages are in the ngs_workshop.zip file. Unzip the file. There is an empty folder called /results. Put all the output files there. Try to avoid putting the output files and input files in the same folder. There is a potential risk to remove or overwrite the raw data files. Always write to a different folder and create symbolic links (`ln -s`) pointing to the raw data if necessary.
PC users, copy the ngs_workshop folder under Cygwin installation folder. For example, if your Cygwin is installed at D:\cygwin, you can copy the ngs_workshop folder to D:\cygwin\home\username. Once you open the Cygwin terminal for the first time, a home folder will be created.

0.b. Terminal. Open a terminal, change your current path to ngs_workshop. The entire tutorial assumes your current path is ngs_workshop folder. Otherwise, you will need to change the paths of input, output, reference files in order to run the codes correctly. You can use `pwd` to find out your current path. All tasks in this tutorial are to be done in terminal by default.
PC users, you can use "Shift+Ins" to paste texts into the command line window.

1. Sample QC

1.a. Basic practice: sequencing quality (FastQC).
Exercise 1: Input data can be found in /data/1.QC. Are these two datasets of high quality?

1.a.1. Launch FastQC:
Mac/Linux users,

```
$ ./tools/FastQC/fastqc
```

PC users, in windows file explorer rather than Cygwin terminal, double click run_fastqc.bat in the FastQC folder.

1.a.2. Load both files and speculate the quality matrices.

1.b. Advanced practice: sample swapping (NGSCheckMate).

2. Read mapping

2.a. Basic practice: standard pipeline (bwa, hg19) and alignment QC.

Exercise 2: Align the reads by BWA. Input data can be found in /data/2.mapping.

2.a.1. Index the reference genome:

```
$ ./tools/bwa-0.7.15/bwa index ref/hg19.fa
```

2.a.2. Align the reads:

```
$ ./tools/bwa-0.7.15/bwa mem ref/hg19.fa data/2.mapping/TCGA-AA-A01T.tumor.1.fq.gz data/2.mapping/TCGA-AA-A01T.tumor.2.fq.gz | gzip -3 > results/TCGA-AA-A01T.tumor.sam.gz
```

2.a.3. Index the reference genome:

```
$ samtools faidx ref/hg19.fa
```

2.a.4. Convert sam to bam, sort bam and index bam:

```
$ samtools view -bt ref/hg19.fa.fai results/TCGA-AA-A01T.tumor.sam.gz > results/TCGA-AA-A01T.tumor.bam
$ samtools sort results/TCGA-AA-A01T.tumor.bam > results/TCGA-AA-A01T.tumor.sorted.bam
$ samtools index results/TCGA-AA-A01T.tumor.sorted.bam
```

The final sorted and indexed bam file should be the same as the one in /data/3.mutation folder.

2.a.5. Take a look at the alignment:

```
$ samtools view results/TCGA-AA-A01T.tumor.sorted.bam|less
```

Exercise 3: Alignment QC. What proportion of reads can be mapped to the reference genome?

2.a.6. Parse basic alignment statistics into a text file with samtools:

```
$ samtools stats results/TCGA-AA-A01T.tumor.sorted.bam > results/TCGA-AA-A01T.tumor.sorted.bam.stats
```

2.b. Advanced practice: aligner selection (novoalign), reference genome selection (with decoy sequences), and fine processing of bam files (Picard, GATK).

3. Mutation calling

3.a. Basic practice: standard pipeline (Varscan) and checking the mutation call in IGV.

Exercise 4: Call mutations with VarScan. Input data can be found in /data/3.mutation.

3.a.1. Mutation calling.

VarScan v.2.4.3 is provided under /tools. Run the following commands to call somatic mutations:

```
$ samtools mpileup -f ref/hg19.fa data/3.mutation/TCGA-AA-A01T.normal.sorted.bam data/3.mutation/TCGA-AA-A01T.tumor.sorted.bam > results/TCGA-AA-A01T.mpileup
$ java -jar tools/VarScan.v2.4.3.jar somatic results/TCGA-AA-A01T.mpileup results/TCGA-AA-A01T --mpileup 1 --output-vcf 1
$ java -jar tools/VarScan.v2.4.3.jar processSomatic results/TCGA-AA-A01T.snp.vcf
```

```
$ java -jar tools/VarScan.v2.4.3.jar somaticFilter
results/TCGA-AA-A01T.snp.Somatic.hc.vcf --output-file
results/TCGA-AA-A01T.snp.Somatic.hc.filtered.vcf
```

How many high quality somatic mutations did you find for this patient? (The final vcf file should be the same as the one in /data/4.annotation.)

Exercise 5: Visualize the mutations in IGV.

3.a.2. Launch IGV.

Mac/Linux users,

```
$ ./tools/IGV_2.3.97/igv.sh
```

PC users, in windows file explore rather than Cygwin terminal, double click igv.bat in the IGV_2.3.97 folder.

3.a.3. Load both tumor and normal bam files, type in the coordinate of one of the mutations called in the previous exercise in the search box on top as chr1:2345678. Does the somatic mutation look convincing?

3.b. Advanced practice: mutation caller selection.

4. Variant annotation

4.a. Basic practice: protein coding mutations (vcf2maf).

Exercise 6: Generate MAF (mutation annotation format) file. Input data can be found in /data/4.annotation. What genes are mutated? What genes might have altered function?

4.a.1. Run vcf2maf.pl:

```
$ perl tools/vcf2maf-master/vcf2maf.pl --input-vcf
data/4.annotation/TCGA-AA-A01T.snp.Somatic.hc.filtered.vcf --
output-maf results/TCGA-AA-A01T.somatic.snp.maf --vep-path
tools/ensembl-tools-release-
88/scripts/variant_effect_predictor/ --vep-data ref/ --
buffer-size 1000 --ref-fasta
ref/homo_sapiens/88_GRCh37/Homo_sapiens.GRCh37.75.dna.primary
_assembly.fa --cache-version 88 --filter-vcf
ref/homo_sapiens/88_GRCh37/ExAC_nonTCGA.r0.3.1.sites.vep.vcf.
gz
```

4.b. Advanced practice: non-coding mutations.

5. Results aggregation

5.a. Basic practice: summary and frequently mutated genes.

Exercise 7: Explore the text file /data/5.aggregation/CRC.somatic.SNVs.indels.maf. The relevant fields are:

Column 2: chromosome
Column 3: position
Column 7: patient id
Column 9: gene name
Column 13: predicted variant function
Column 14: variant type.

Use your favorite tool to answer the following questions:

How many patients there are? How many point mutations, insertions and deletions there are in total? Which patient has the most point mutations?

Hint, the following Unix command can give number of patients:

```
$ cat data/5.aggregation/CRC.somatic.SNVs.indels.maf |cut -
f7|sort|uniq|wc
```

Exercise 8: Answer the following questions using the data file /data/5.aggregation/CRC.somatic.SNVs.indels.maf:

How many genes are mutated (non-silent mutation)? What are the top three genes that are mutated in most number of patients? Hint, one patient may carry multiple mutations in the same gene. They should be counted as 1 at patient level.

5.b. Advanced practice: significantly mutated genes (MutSig), visualization.

6. Clinical relevance

6.a. Basic practice: actionable mutations.

Exercise 9: Answer the following questions using the data file /data/5.aggregation/CRC.somatic.SNVs.indels.maf:

9a. One of the available target therapies for colorectal cancer patients is anti-EGFR treatment. This treatment requires patients to have *EGFR* expressed and wild type *KRAS*, *NRAS* and *BRAF*. Assuming we have *EGFR* expression data, can you find out which patients are not suitable for anti-EGFR treatment based on mutation data?

9b. There are inhibitors available for BRAF. The most frequent *BRAF* mutation is V600E. How many patients carry *BRAF* V600E mutation? You need to find out the genomic coordinate of V600E mutation. These patients may benefit from combined therapy (anti-EGFR + anti-BRAF).

Exercise 10: Colorectal cancer patients characterized as micro-satellite instable (MSI) can respond to immunotherapy. Are there any patients in our cohort likely to respond to immunotherapy?

6.b. Advanced practice: natural language processing, machine learning (IBM Watson).