Introduction to R: the Power of the Data Frame

Peter Carbonetto
pcarbo@uchicago.edu

Research Computing Center
and the Dept. of Human Genetics
University of Chicago

All materials presented today are available at
http://github.com/rcc-uchicago/R-intro

http://github.com/rcc-uchicago/R-intro

Structure of today's workshop

1. Introduction & motivation.
2. Introduce yourself to your neighbours.
3. Setup…
    - Your R programming environment.
    - Download the code & data.
    - Install & load packages.
4. Execute the code.
5. Walking through the code…
    - Loading the data.
    - Inspecting the data.
    - Manipulating the data.
    - Visualizing the data.
6. Feedback.

I need help,
or I am stuck.

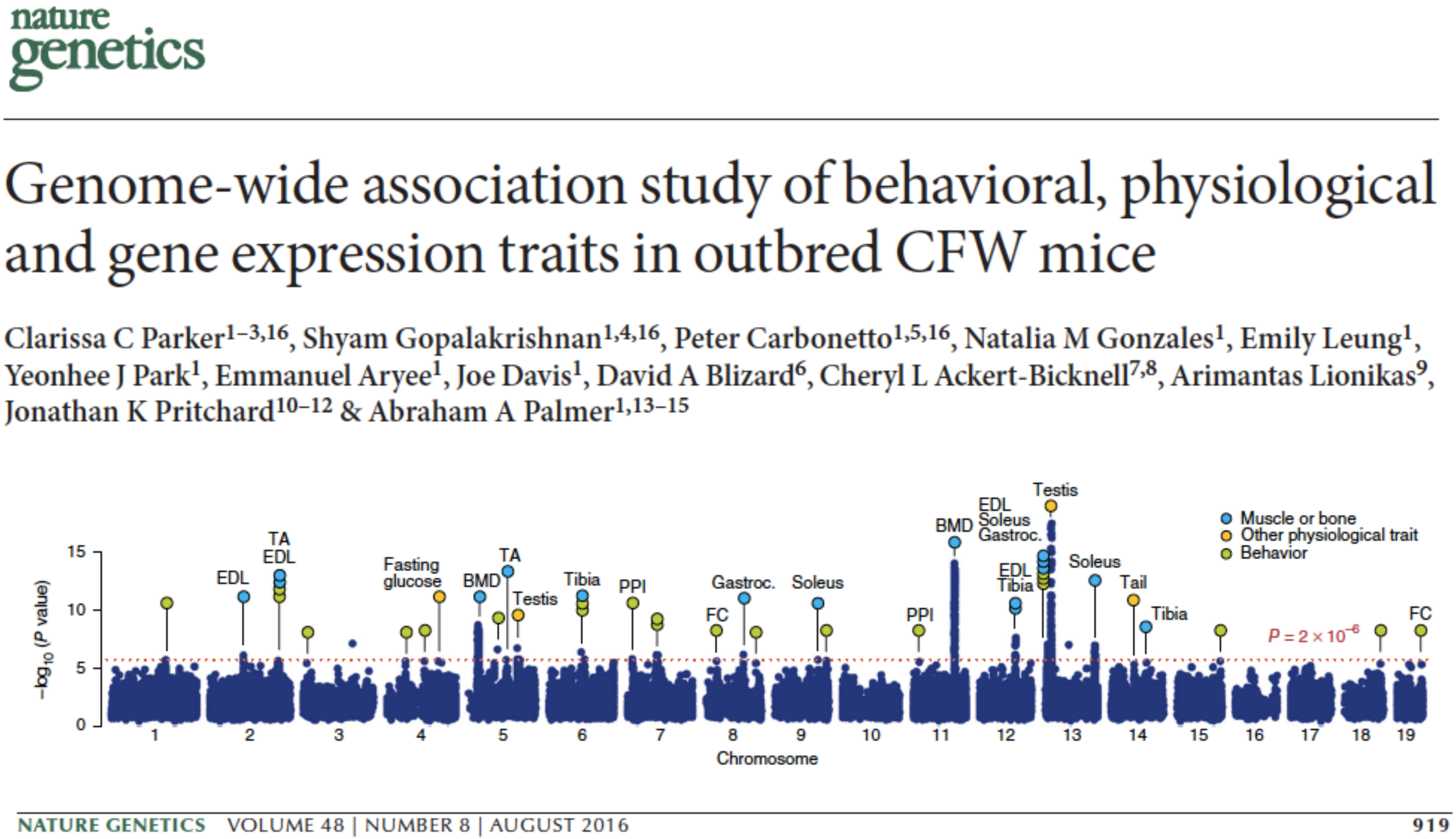
Downloads

R: **cran.r-project.org**

RStudio: **www.rstudio.com/products/rstudio**

ThinLinc: **www.cendio.com/thinlinc/download**

Data analysis in R



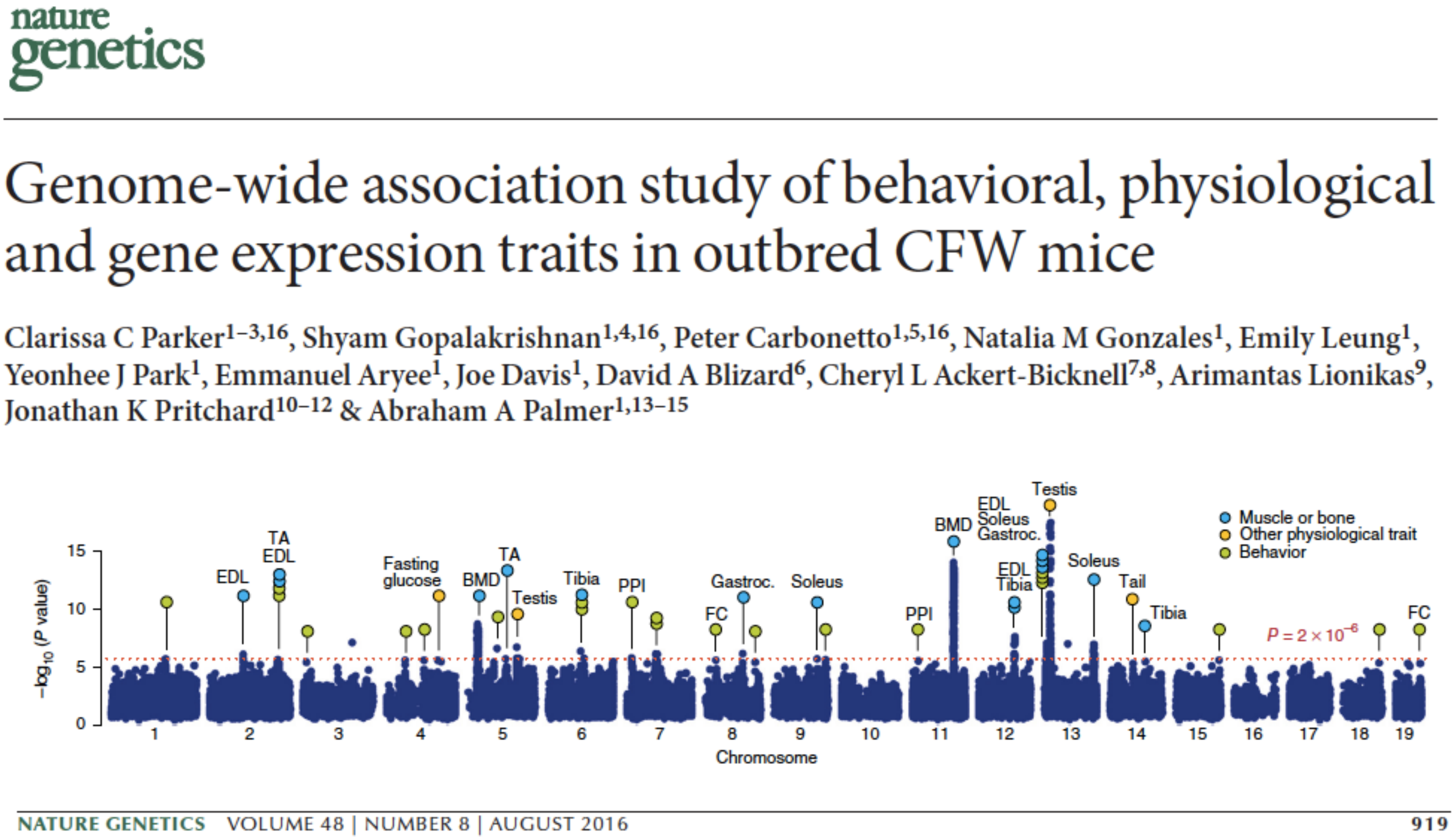
nature
genetics

Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice

Clarissa C Parker[1–3,16], Shyam Gopalakrishnan[1,4,16], Peter Carbonetto[1,5,16], Natalia M Gonzales[1], Emily Leung[1], Yeonhee J Park[1], Emmanuel Aryee[1], Joe Davis[1], David A Blizard[6], Cheryl L Ackert-Bicknell[7,8], Arimantas Lionikas[9], Jonathan K Pritchard[10–12] & Abraham A Palmer[1,13–15]

NATURE GENETICS VOLUME 48 | NUMBER 8 | AUGUST 2016 919

Data analysis in R

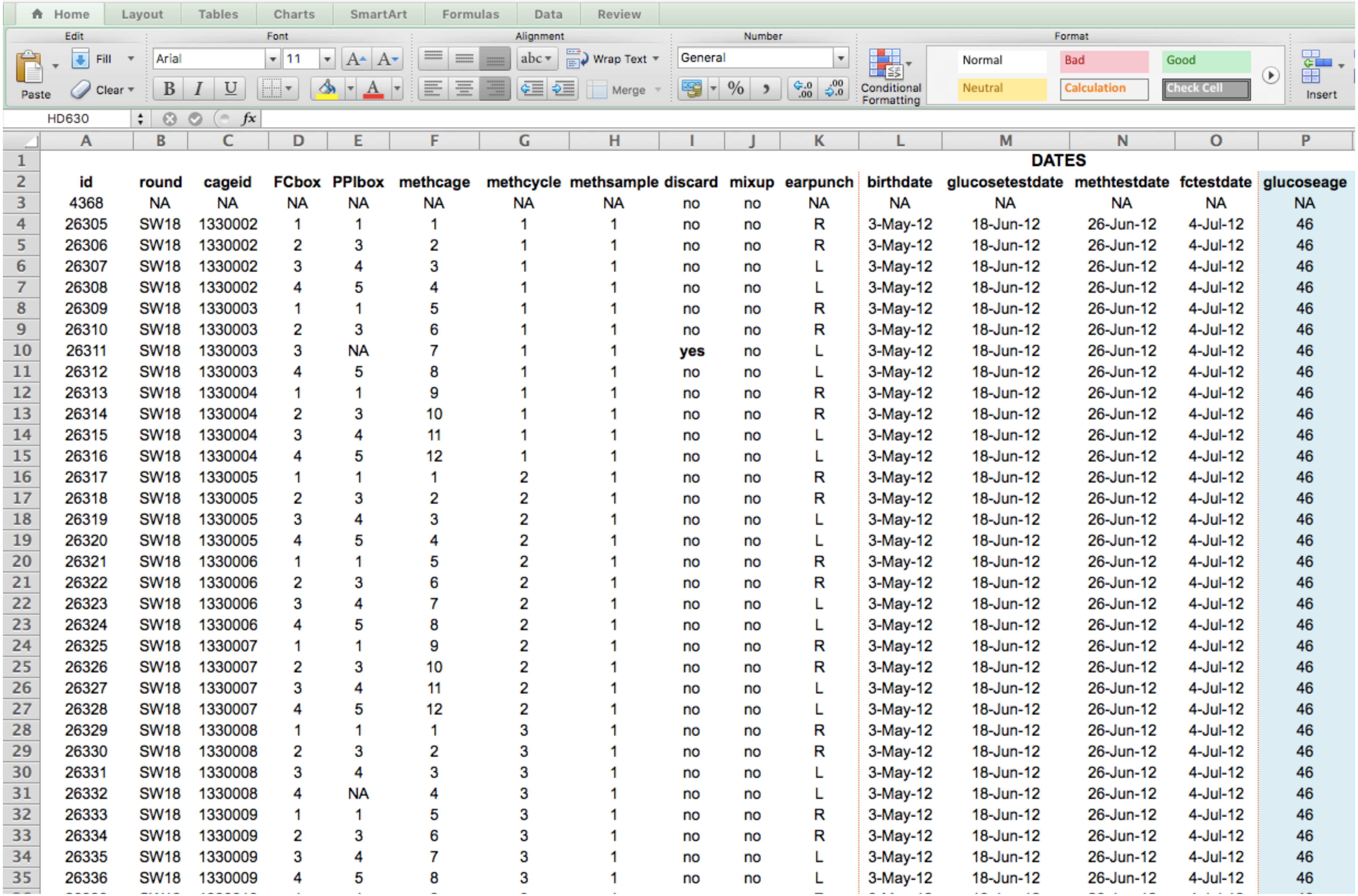


nature genetics

Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice

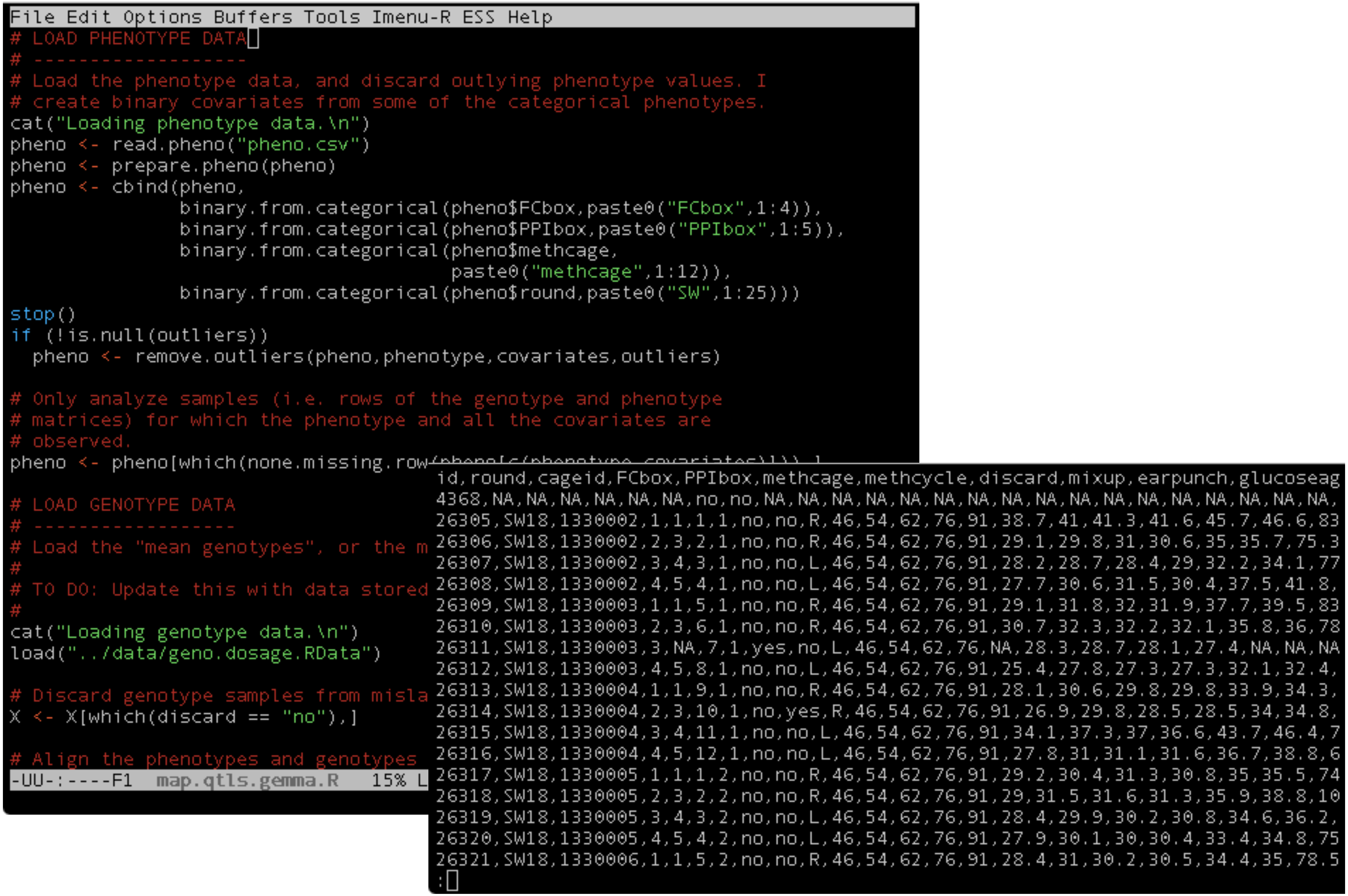
Clarissa C Parker[1–3,16], Shyam Gopalakrishnan[1,4,16], Peter Carbonetto[1,5,16], Natalia M Gonzales[1], Emily Leung[1], Yeonhee J Park[1], Emmanuel Aryee[1], Joe Davis[1], David A Blizard[6], Cheryl L Ackert-Bicknell[7,8], Arimantas Lionikas[9], Jonathan K Pritchard[10–12] & Abraham A Palmer[1,13–15]

Clarissa’s data in Excel



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1												DATES				
2	id	round	cageid	FCbox	PPIbox	methcage	methcycle	methsample	discard	mixup	earpunch	birthdate	glucosetestdate	methtestdate	fctestdate	glucoseage
3	4368	NA	NA	NA	NA	NA	NA	NA	no	no	NA	NA	NA	NA	NA	NA
4	26305	SW18	1330002	1	1	1	1	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
5	26306	SW18	1330002	2	3	2	1	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
6	26307	SW18	1330002	3	4	3	1	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
7	26308	SW18	1330002	4	5	4	1	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
8	26309	SW18	1330003	1	1	5	1	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
9	26310	SW18	1330003	2	3	6	1	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
10	26311	SW18	1330003	3	NA	7	1	1	yes	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
11	26312	SW18	1330003	4	5	8	1	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
12	26313	SW18	1330004	1	1	9	1	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
13	26314	SW18	1330004	2	3	10	1	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
14	26315	SW18	1330004	3	4	11	1	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
15	26316	SW18	1330004	4	5	12	1	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
16	26317	SW18	1330005	1	1	1	2	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
17	26318	SW18	1330005	2	3	2	2	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
18	26319	SW18	1330005	3	4	3	2	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
19	26320	SW18	1330005	4	5	4	2	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
20	26321	SW18	1330006	1	1	5	2	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
21	26322	SW18	1330006	2	3	6	2	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
22	26323	SW18	1330006	3	4	7	2	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
23	26324	SW18	1330006	4	5	8	2	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
24	26325	SW18	1330007	1	1	9	2	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
25	26326	SW18	1330007	2	3	10	2	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
26	26327	SW18	1330007	3	4	11	2	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
27	26328	SW18	1330007	4	5	12	2	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
28	26329	SW18	1330008	1	1	1	3	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
29	26330	SW18	1330008	2	3	2	3	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
30	26331	SW18	1330008	3	4	3	3	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
31	26332	SW18	1330008	4	NA	4	3	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
32	26333	SW18	1330009	1	1	5	3	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
33	26334	SW18	1330009	2	3	6	3	1	no	no	R	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
34	26335	SW18	1330009	3	4	7	3	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46
35	26336	SW18	1330009	4	5	8	3	1	no	no	L	3-May-12	18-Jun-12	26-Jun-12	4-Jul-12	46

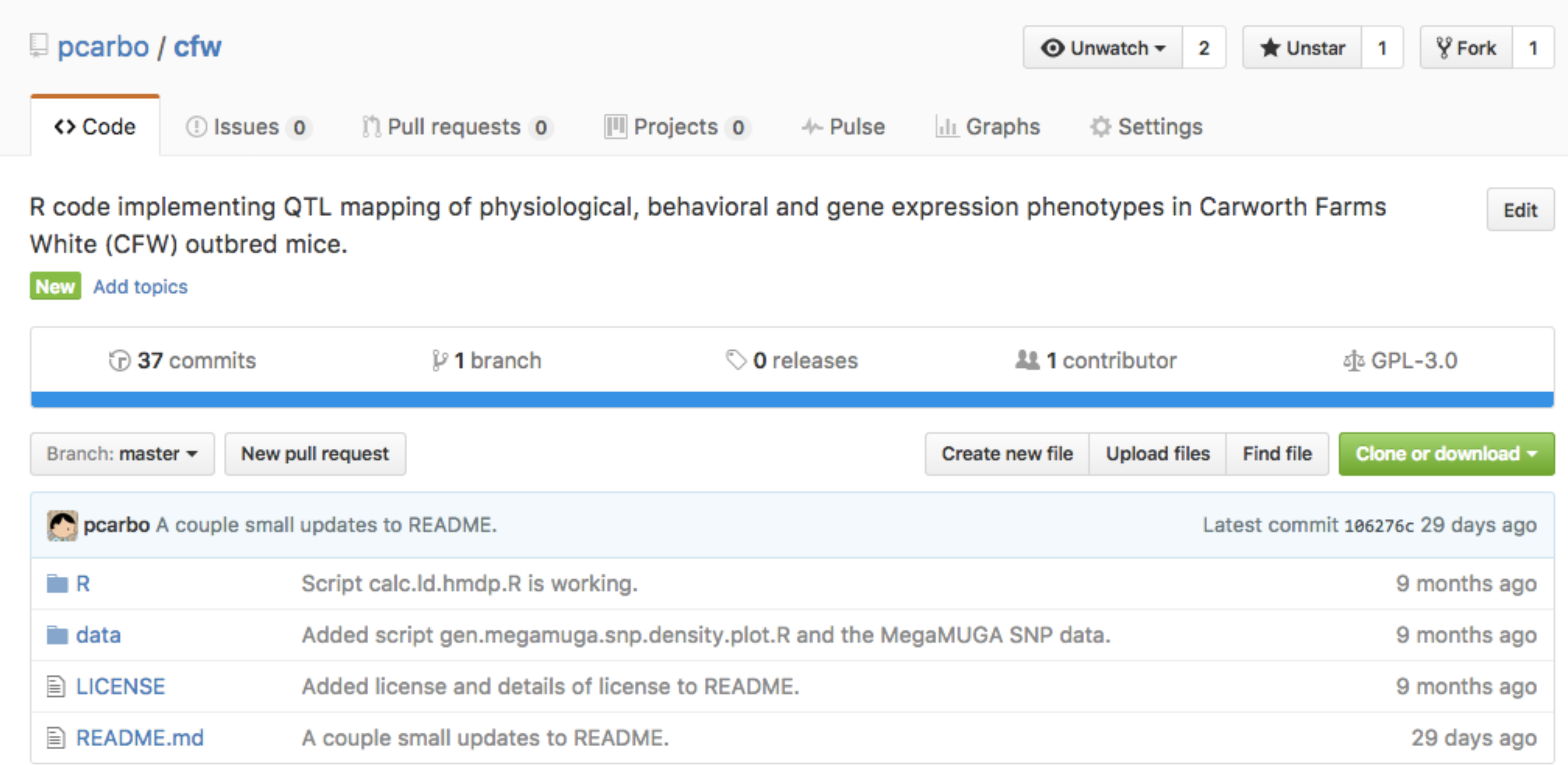
Clarissa’s data in R


File Edit Options Buffers Tools Imenu-R ESS Help
-UU-:----F1 map.qtls.gemma.R   15% L

```
# LOAD PHENOTYPE DATA
# -------------------
# Load the phenotype data, and discard outlying phenotype values. I
# create binary covariates from some of the categorical phenotypes.
cat("Loading phenotype data.\n")
pheno <- read.pheno("pheno.csv")
pheno <- prepare.pheno(pheno)
pheno <- cbind(pheno,
               binary.from.categorical(pheno$FCbox,paste0("FCbox",1:4)),
               binary.from.categorical(pheno$PPIbox,paste0("PPIbox",1:5)),
               binary.from.categorical(pheno$methcage,
                                       paste0("methcage",1:12)),
               binary.from.categorical(pheno$round,paste0("SW",1:25)))
stop()
if (!is.null(outliers))
  pheno <- remove.outliers(pheno,phenotype,covariates,outliers)

# Only analyze samples (i.e. rows of the genotype and phenotype
# matrices) for which the phenotype and all the covariates are
# observed.
pheno <- pheno[which(none.missing.row

# LOAD GENOTYPE DATA
# ------------------
# Load the "mean genotypes", or the m
#
# TO DO: Update this with data stored
#
cat("Loading genotype data.\n")
load("../data/geno.dosage.RData")

# Discard genotype samples from misla
X <- X[which(discard == "no"),]

# Align the phenotypes and genotypes
```

```
id,round,cageid,FCbox,PPIbox,methcage,methcycle,discard,mixup,earpunch,glucoseag
4368,NA,NA,NA,NA,NA,NA,no,no,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
26305,SW18,1330002,1,1,1,1,no,no,R,46,54,62,76,91,38.7,41,41.3,41.6,45.7,46.6,83
26306,SW18,1330002,2,3,2,1,no,no,R,46,54,62,76,91,29.1,29.8,31,30.6,35,35.7,75.3
26307,SW18,1330002,3,4,3,1,no,no,L,46,54,62,76,91,28.2,28.7,28.4,29,32.2,34.1,77
26308,SW18,1330002,4,5,4,1,no,no,L,46,54,62,76,91,27.7,30.6,31.5,30.4,37.5,41.8,
26309,SW18,1330003,1,1,5,1,no,no,R,46,54,62,76,91,29.1,31.8,32,31.9,37.7,39.5,83
26310,SW18,1330003,2,3,6,1,no,no,R,46,54,62,76,91,30.7,32.3,32.2,32.1,35.8,36,78
26311,SW18,1330003,3,NA,7,1,yes,no,L,46,54,62,76,NA,28.3,28.7,28.1,27.4,NA,NA,NA
26312,SW18,1330003,4,5,8,1,no,no,L,46,54,62,76,91,25.4,27.8,27.3,27.3,32.1,32.4,
26313,SW18,1330004,1,1,9,1,no,no,R,46,54,62,76,91,28.1,30.6,29.8,29.8,33.9,34.3,
26314,SW18,1330004,2,3,10,1,no,yes,R,46,54,62,76,91,26.9,29.8,28.5,28.5,34,34.8,
26315,SW18,1330004,3,4,11,1,no,no,L,46,54,62,76,91,34.1,37.3,37,36.6,43.7,46.4,7
26316,SW18,1330004,4,5,12,1,no,no,L,46,54,62,76,91,27.8,31,31.1,31.6,36.7,38.8,6
26317,SW18,1330005,1,1,1,2,no,no,R,46,54,62,76,91,29.2,30.4,31.3,30.8,35,35.5,74
26318,SW18,1330005,2,3,2,2,no,no,R,46,54,62,76,91,29,31.5,31.6,31.3,35.9,38.8,10
26319,SW18,1330005,3,4,3,2,no,no,L,46,54,62,76,91,28.4,29.9,30.2,30.8,34.6,36.2,
26320,SW18,1330005,4,5,4,2,no,no,L,46,54,62,76,91,27.9,30.1,30,30.4,33.4,34.8,75
26321,SW18,1330006,1,1,5,2,no,no,R,46,54,62,76,91,28.4,31,30.2,30.5,34.4,35,78.5
:
```

Sharable data analysis

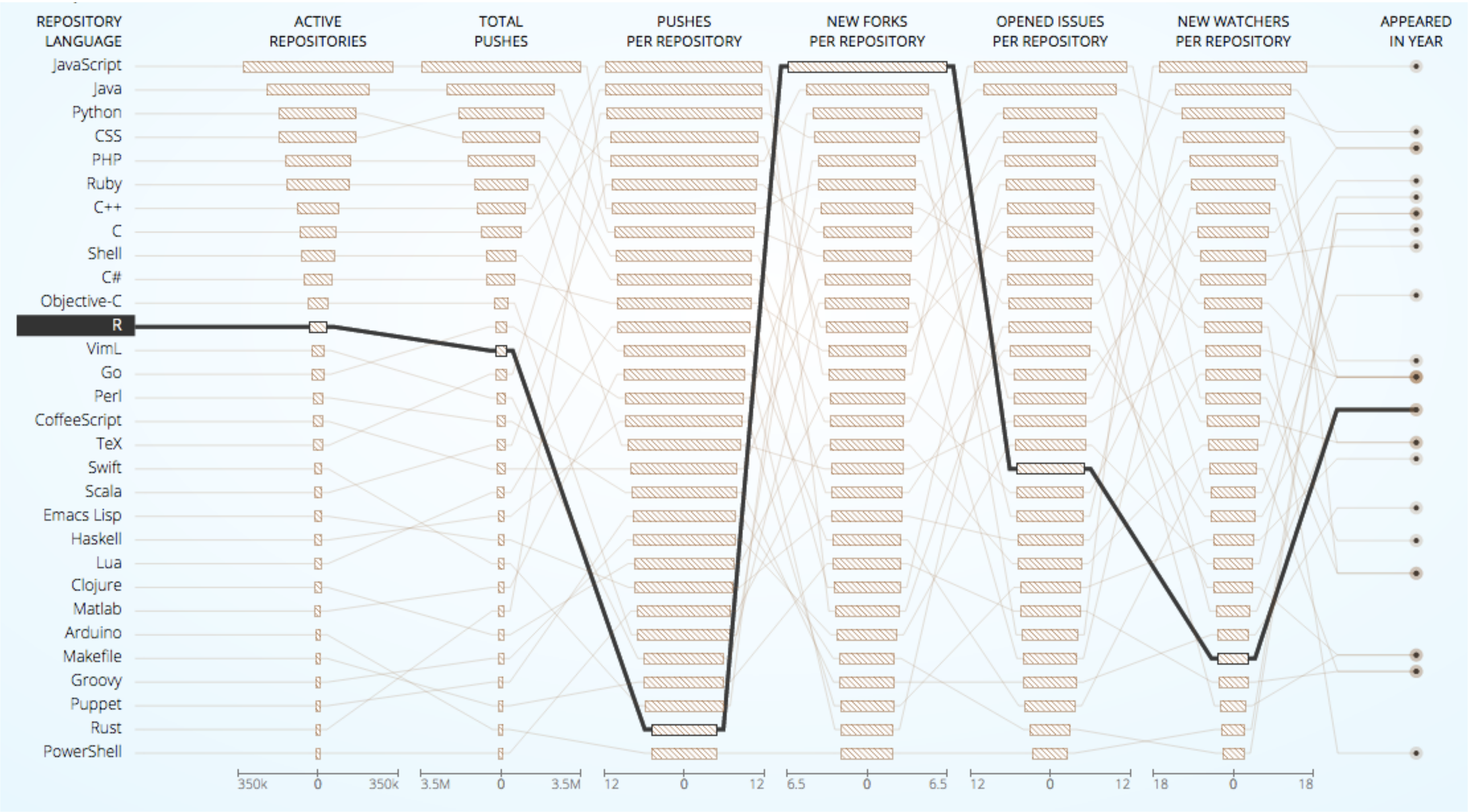


R code implementing QTL mapping of physiological, behavioral and gene expression phenotypes in Carworth Farms White (CFW) outbred mice.

pcarbo A couple small updates to README. Latest commit 106276c 29 days ago

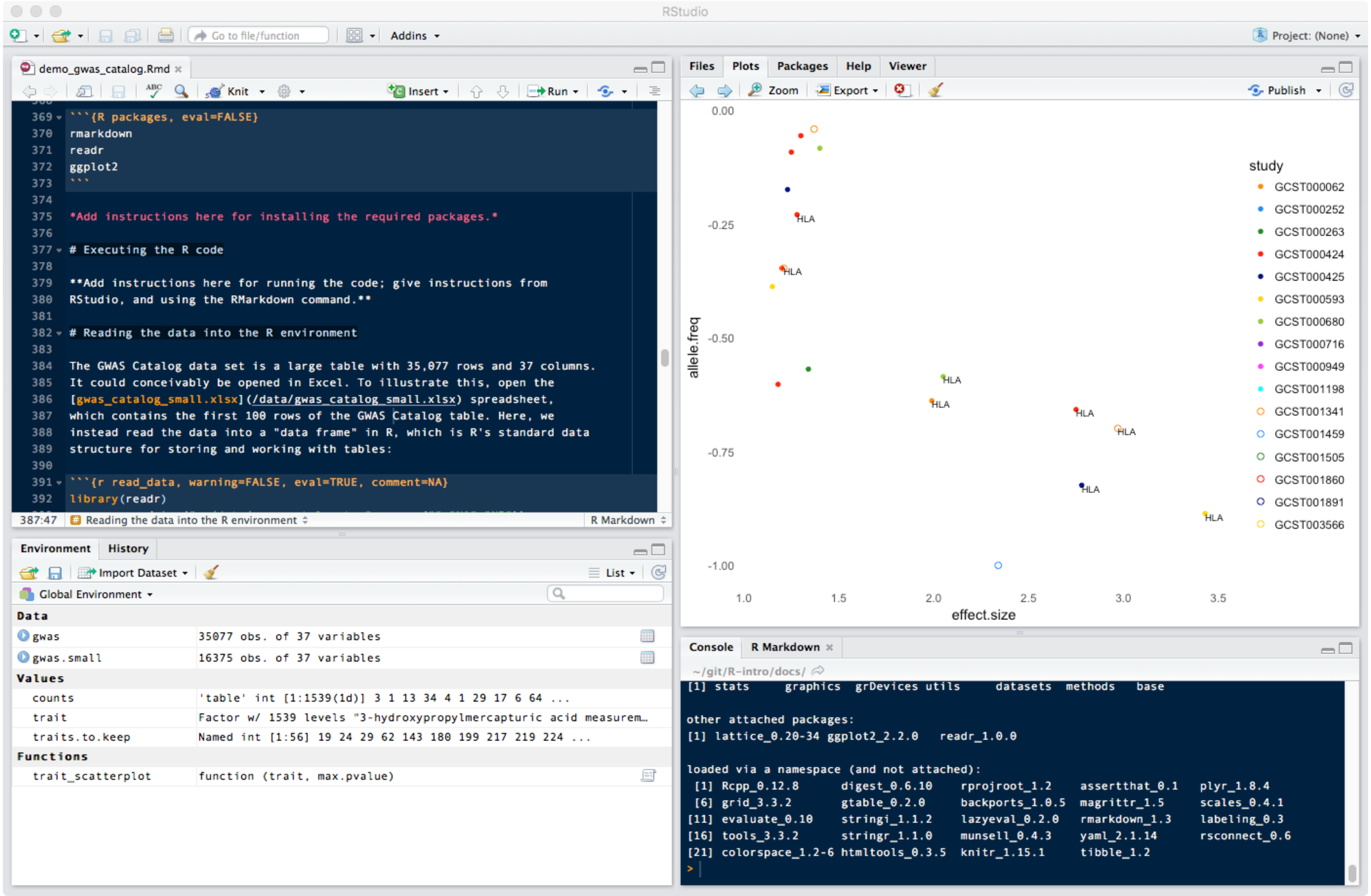
R	Script calc.ld.hmdp.R is working.	9 months ago
data	Added script gen.megamuga.snp.density.plot.R and the MegaMUGA SNP data.	9 months ago
LICENSE	Added license and details of license to README.	9 months ago
README.md	A couple small updates to README.	29 days ago

Why R?



source: GitHut.info (Carlo Zapponi, 2014)

IDE = integrated development environment



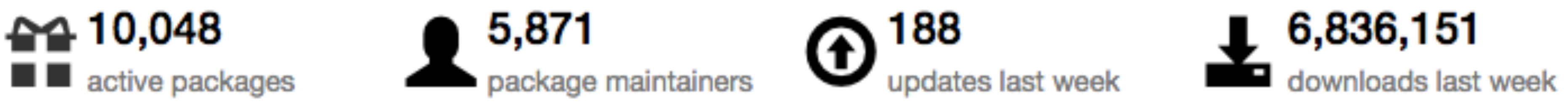
~~~
```{R packages, eval=FALSE}
rmarkdown
readr
ggplot2
```

*Add instructions here for installing the required packages.*

# Executing the R code

**Add instructions here for running the code; give instructions from
RStudio, and using the RMarkdown command.**

# Reading the data into the R environment

The GWAS Catalog data set is a large table with 35,077 rows and 37 columns.
It could conceivably be opened in Excel. To illustrate this, open the
[gwas_catalog_small.xlsx](/data/gwas_catalog_small.xlsx) spreadsheet,
which contains the first 100 rows of the GWAS Catalog table. Here, we
instead read the data into a "data frame" in R, which is R's standard data
structure for storing and working with tables:

```{r read_data, warning=FALSE, eval=TRUE, comment=NA}
library(readr)
~~~

387:47 Reading the data into the R environment ÷ R Markdown ÷

```
Data
gwas               35077 obs. of 37 variables
gwas.small         16375 obs. of 37 variables
Values
counts             'table' int [1:1539(1d)] 3 1 13 34 4 1 29 17 6 64 ...
trait              Factor w/ 1539 levels "3-hydroxypropylmercapturic acid measurem…
traits.to.keep     Named int [1:56] 19 24 29 62 143 180 199 217 219 224 ...
Functions
trait_scatterplot  function (trait, max.pvalue)
```

```
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] lattice_0.20-34 ggplot2_2.2.0   readr_1.0.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.8      digest_0.6.10    rprojroot_1.2    assertthat_0.1   plyr_1.8.4
 [6] grid_3.3.2       gtable_0.2.0     backports_1.0.5  magrittr_1.5     scales_0.4.1
[11] evaluate_0.10    stringi_1.1.2    lazyeval_0.2.0   rmarkdown_1.3    labeling_0.3
[16] tools_3.3.2      stringr_1.1.0    munsell_0.4.3    yaml_2.1.14      rsconnect_0.6
[21] colorspace_1.2-6 htmltools_0.3.5  knitr_1.15.1     tibble_1.2
>
```

R is a community-driven resource

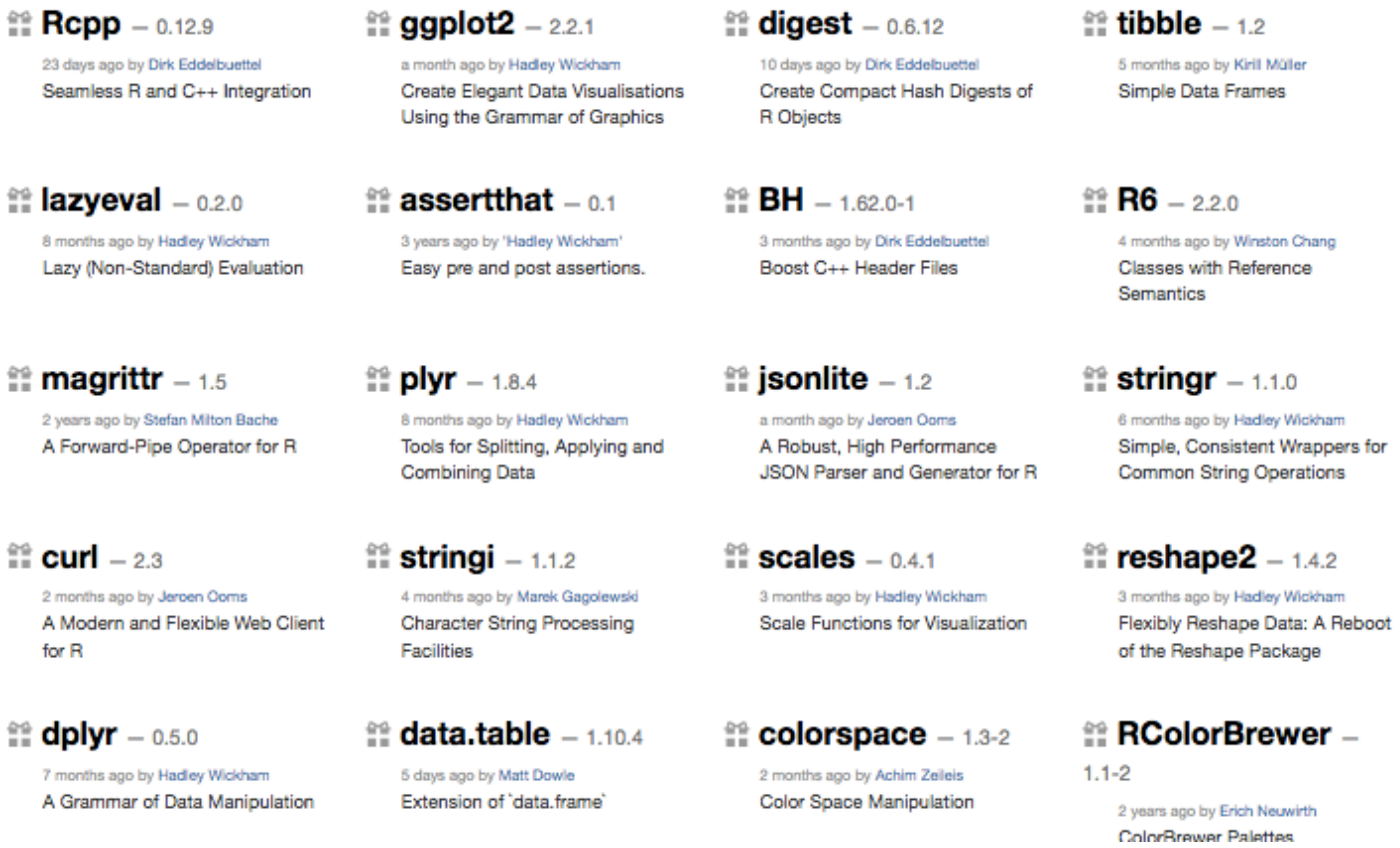


10,048 active packages

5,871 package maintainers

188 updates last week

6,836,151 downloads last week



Rcpp – 0.12.9
23 days ago by Dirk Eddelbuettel
Seamless R and C++ Integration

ggplot2 – 2.2.1
a month ago by Hadley Wickham
Create Elegant Data Visualisations Using the Grammar of Graphics

digest – 0.6.12
10 days ago by Dirk Eddelbuettel
Create Compact Hash Digests of R Objects

tibble – 1.2
5 months ago by Kirill Müller
Simple Data Frames

lazyeval – 0.2.0
8 months ago by Hadley Wickham
Lazy (Non-Standard) Evaluation

assertthat – 0.1
3 years ago by 'Hadley Wickham'
Easy pre and post assertions.

BH – 1.62.0-1
3 months ago by Dirk Eddelbuettel
Boost C++ Header Files

R6 – 2.2.0
4 months ago by Winston Chang
Classes with Reference Semantics

magrittr – 1.5
2 years ago by Stefan Milton Bache
A Forward-Pipe Operator for R

plyr – 1.8.4
8 months ago by Hadley Wickham
Tools for Splitting, Applying and Combining Data

jsonlite – 1.2
a month ago by Jeroen Ooms
A Robust, High Performance JSON Parser and Generator for R

stringr – 1.1.0
6 months ago by Hadley Wickham
Simple, Consistent Wrappers for Common String Operations

curl – 2.3
2 months ago by Jeroen Ooms
A Modern and Flexible Web Client for R

stringi – 1.1.2
4 months ago by Marek Gagolewski
Character String Processing Facilities

scales – 0.4.1
3 months ago by Hadley Wickham
Scale Functions for Visualization

reshape2 – 1.4.2
3 months ago by Hadley Wickham
Flexibly Reshape Data: A Reboot of the Reshape Package

dplyr – 0.5.0
7 months ago by Hadley Wickham
A Grammar of Data Manipulation

data.table – 1.10.4
5 days ago by Matt Dowle
Extension of `data.frame`

colorspace – 1.3-2
2 months ago by Achim Zeileis
Color Space Manipulation

RColorBrewer – 1.1-2
2 years ago by Erich Neuwirth
ColorBrewer Palettes

source: www.r-pkg.org (accessed Feb. 6, 2017)

Key features of R

1. R is based on the statistical programming language **S**.
2. R is **open source** (GPL).
3. R is **high-level**.
4. R is **interpreted** (rather than compiled).
5. R supports some aspets of object-oriented programming.
6. R is a **programming environment**.
7. RStudio provides a **free IDE** (integrated development interface).
8. R evolution is **community driven** through development of **packages**.

Some general advice

1. Use **midway2**.
2. There is probably a package for you (don't reinvent the wheel).
3. If you have trouble installing an R package, email help@rcc.uchicago.edu.
4. Use `help(some_function)` and stackoverflow.com.
5. Learn to avoid loops as much as possible; e.g., use `apply()`, `lapply()`, `tapply()`, `do.call()`.
6. The "defaults" in R are often not what you want—check the function outputs carefully.
7. Document your setup—start with `sessionInfo()`.
8. Someone else's R code is difficult to understand (and your code months later)—please add comments to your code to explain what it does!

What we will do today

- You will get exposed to…
  - Setting up your laptop and/or the cluster to do interactive programming in R.
  - Installing and using packages.
  - Working with “R notebooks”.
  - Executing R code and build notebooks into sharable documents.
  - Implementing simple data analysis steps by example.

What we will ***not*** cover today

- How to program.
- Syntax and grammar of R.
- High-performance computing in R.

Set up your R environment

1. R on your laptop:
   1. Install R (text-based).
   2. Install RStudio (IDE).
2. R on midway—no graphics:
   1. Connect to midway2.
   2. R module.
   3. Rscript.
3. R on midway—with graphics:

   a. Connect via ThinLinc.

   b. RStudio module.

   c. RStudio Server (*limited availability for now*)

Feedback

You will receive an email announce@rcc.uchicago.edu requesting feedback on this workshop. **Please complete this survey!**