GWAS Pipeline Full

Michael Beyeler

2018-05-13

# Setup

## Loading packages, Reproducibility

**Concerning reproducibility**: In order to guarantee reproducibility, keep the .BuildReproducibleEnvironment(...) FUNCTION active.

if(!require('checkpoint', character.only=T))  
 install.packages('checkpoint')  
  
source('Helper\_Scripts/Environment\_Manipulation\_and\_Reproducibility.R')  
  
# In order to make the script 100% reproducible, keep the next line active:  
.BuildReproducibleEnvironment(PROJECT.SNAPSHOT.DATE = '2017-12-31',  
 PROJECT.VERSION = '3.4.3',  
 SCAN.FOR.PACKAGES = FALSE)  
  
.LIST.OF.PACKAGES <- c(  
 'data.table', #  
 'icesTAF', # dos2unix function  
 'lintr', # good debugging tool  
 'lme4', #   
 'tictoc' #   
)  
.LoadPackages(.LIST.OF.PACKAGES)

sessionInfo()

## R version 3.4.3 (2017-11-30)  
## Platform: x86\_64-w64-mingw32/x64 (64-bit)  
## Running under: Windows 10 x64 (build 17134)  
##   
## Matrix products: default  
##   
## locale:  
## [1] LC\_COLLATE=English\_United States.1252   
## [2] LC\_CTYPE=English\_United States.1252   
## [3] LC\_MONETARY=English\_United States.1252  
## [4] LC\_NUMERIC=C   
## [5] LC\_TIME=English\_United States.1252   
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] tictoc\_1.0 lme4\_1.1-15 Matrix\_1.2-12   
## [4] lintr\_1.0.2 icesTAF\_1.4-1 data.table\_1.10.4-3   
## [7] checkpoint\_0.4.3 RevoUtils\_10.0.7 RevoUtilsMath\_10.0.1  
##   
## loaded via a namespace (and not attached):  
## [1] rex\_1.1.2 Rcpp\_0.12.14 knitr\_1.18 magrittr\_1.5   
## [5] MASS\_7.3-47 splines\_3.4.3 lattice\_0.20-35 R6\_2.2.2   
## [9] minqa\_1.2.4 stringr\_1.2.0 httr\_1.3.1 tools\_3.4.3   
## [13] grid\_3.4.3 nlme\_3.1-131 htmltools\_0.3.6 yaml\_2.1.16   
## [17] lazyeval\_0.2.1 rprojroot\_1.3-1 digest\_0.6.13 nloptr\_1.0.4   
## [21] evaluate\_0.10.1 rmarkdown\_1.6 stringi\_1.1.6 compiler\_3.4.3   
## [25] backports\_1.1.2

## Setting up Anaconda python

Path is set up depending on what OS is used

To-do: check if it also works for Mac

if(.Platform$OS.type == "unix") {  
 knitr::opts\_chunk$set(engine.path = list(python = '/anaconda/bin/python'))  
} else {  
 knitr::opts\_chunk$set(engine.path = list(python = file.path(Sys.getenv("USERPROFILE"),"Anaconda2\\python.exe", fsep='\\')))  
}

## Functions

Functions that were specifically programmed for this script

source('Functions/NormalityHistogram.R')  
source('Functions/ChisqForNormality.R')  
source('Functions/WriteBare.R')

# Filtering Lines and Low Minor Allele Frequencies

At the moment, bash is causing some trouble in Windows R Markdown. Thus, in the meantime, I'm using a workaround to write a bash shell script using R code, and then executing it in an unix environment.

cat("#!/bin/bash  
  
PHENOTYPE\_NAME='Mass'  
MAF=0.05  
  
cd plink2\_linux\_x86\_64  
./plink2 --bfile ../Data/dgrp2 --keep ../Outputs/Plink-Lines-$PHENOTYPE\_NAME.txt --maf $MAF --make-bed --out ../Outputs/MassVariants\_MAF5  
",  
file='Scripts/Plink2\_Filtering\_Alleles.sh')  
  
# The following command is necessary to   
dos2unix('Scripts/Plink2\_Filtering\_Alleles.sh')

./Scripts/Plink2\_Filtering\_Alleles.sh

## PLINK v2.00a1LM 64-bit Intel (11 Feb 2018) www.cog-genomics.org/plink/2.0/  
## (C) 2005-2018 Shaun Purcell, Christopher Chang GNU General Public License v3  
## Logging to ../Outputs/MassVariants\_MAF5.log.  
## Options in effect:  
## --bfile ../Data/dgrp2  
## --keep ../Outputs/Plink-Lines-Mass.txt  
## --maf 0.05  
## --make-bed  
## --out ../Outputs/MassVariants\_MAF5  
##   
## Start time: Mon May 14 13:09:14 2018  
## 16221 MB RAM detected; reserving 8110 MB for main workspace.  
## Using up to 8 compute threads.  
## 205 samples (205 females, 0 males; 205 founders) loaded from ../Data/dgrp2.fam.  
## 4438427 variants loaded from ../Data/dgrp2.bim.  
## Note: No phenotype data present.  
## --keep: 157 samples remaining.  
## 157 samples (157 females, 0 males; 157 founders) remaining after main filters.  
## Calculating allele frequencies... 0%1%2%4%5%7%8%10%11%13%14%16%17%19%20%22%23%25%26%28%29%31%32%33%35%36%38%39%41%42%44%45%47%48%50%51%53%54%56%57%59%60%62%63%64%66%67%69%70%72%73%75%76%78%79%81%82%84%85%87%88%90%91%93%94%95%97%98%done.  
## 2443905 variants removed due to minor allele threshold(s)  
## (--maf/--max-maf/--mac/--max-mac).  
## 1994522 variants remaining after main filters.  
## Writing ../Outputs/MassVariants\_MAF5.bed ... 0%1%3%4%6%7%9%10%12%13%15%16%18%19%21%22%24%25%26%28%30%31%33%34%36%37%38%40%42%43%45%46%48%49%51%52%54%55%57%59%60%61%63%64%65%66%67%69%70%72%73%75%76%78%79%80%82%83%85%86%88%89%91%92%94%95%97%98%done.  
## Writing ../Outputs/MassVariants\_MAF5.bim ... done.  
## Writing ../Outputs/MassVariants\_MAF5.fam ... done.  
## End time: Mon May 14 13:09:15 2018

# Phenotype Adjustment

## Constants

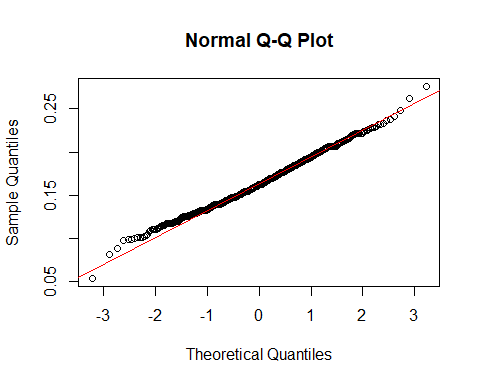
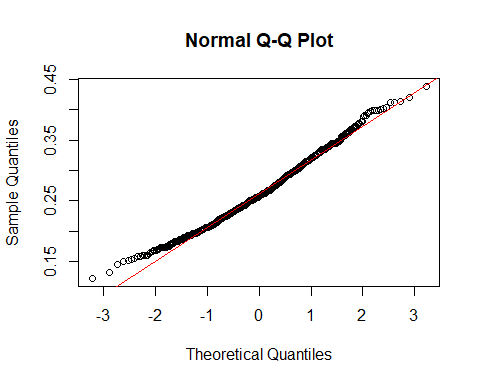
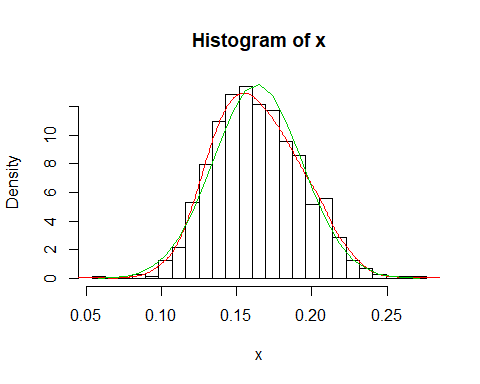
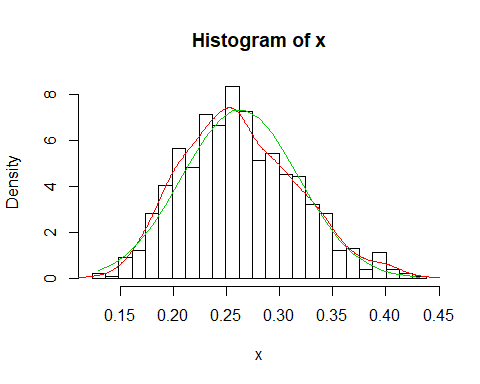
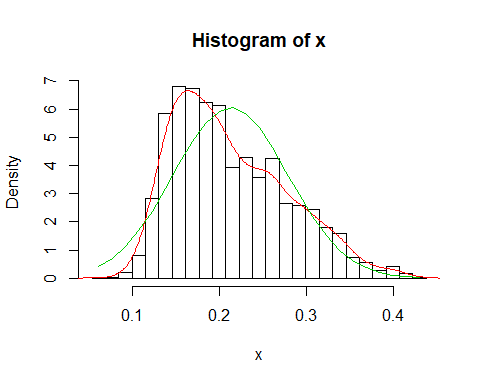
SEXUAL.DIMORPHISM <- TRUE  
PHENOTYPE.NAME <- 'Mass'  
NORMALITY.SIGNIFICANCE.LEVEL <- 0.05  
INVERSIONS.CONSIDERED <- c('In.2L.t', 'In.2R.NS', 'In.3R.P', 'In.3R.K', 'In.3R.Mo')

## Data

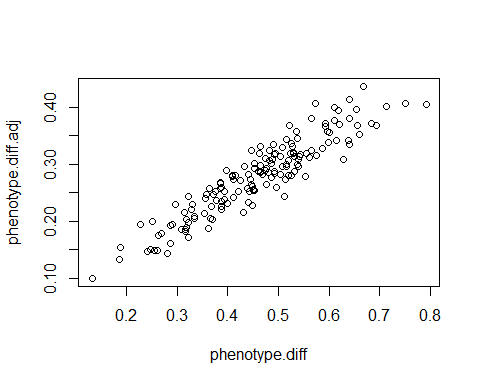
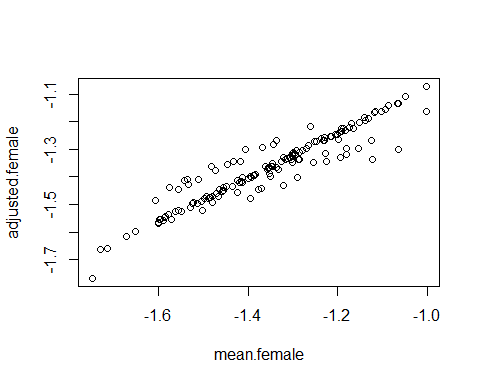
# Main data  
  
# Phenotype\_Raw <- read.delim('Data/Food-Intake-Garlapow.txt', header=F)  
# Phenotype\_Raw <- read.delim('Data/Vonesch2016-IOD-Raw.txt', header=F)  
Phenotype\_Raw <- read.delim('Data/dgrp\_mass.csv', header=T)  
  
# Supporting data  
  
Dgrp2\_Inversions <- read.csv('Data/inversion.csv', header=T)  
Dgrp2\_Infection <- read.csv('Data/wolbachia.csv')  
  
# Some phenotypes actually use the flystock ID instead of the DGRP ID.  
# For these cases, this data frame will come in handy.  
Dgrp\_Flystock\_Ids <- read.delim('Data/Dgrp-Flystocks-Ids.txt',  
 comment.char='#')

## Adjustment procedure

## [1] "Unique to Female: "  
## [1] "Unique to Male: "



## [1] "p-value total 5.2630561996935e-21"  
## [1] 1.025024e-05  
## [1] 0.008176191  
## [1] "At least one of the phenotypes was not normally distributed. Log-transformation performed."



# FaST-LMM GWAS

Coming soon...