Genetics HW12

Jessica Murphy

May 15, 2019

```r
sessionInfo()

## R version 3.5.1 (2018-07-02)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: CentOS release 6.10 (Final)
##
## Matrix products: default
## BLAS: /usr/lib64/R/lib/libRblas.so
## LAPACK: /usr/lib64/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] knitr_1.21
##
## loaded via a namespace (and not attached):
## [1] compiler_3.5.1 magrittr_1.5   tools_3.5.1    stringi_1.3.1
## [5] stringr_1.4.0  xfun_0.5       evaluate_0.13
```

1 Read data

```r
ceu_tmp = read.table("/home/murphjes/BIOS6660/Genetics/ceu.frq",
                     stringsAsFactors = F, header=T)
ceu = ceu_tmp[!is.na(ceu_tmp$MAF), ]

yri_tmp = read.table("/home/murphjes/BIOS6660/Genetics/yri.frq",
                     stringsAsFactors = F, header=T)
yri = yri_tmp[!is.na(yri_tmp$MAF), ]

chb_jpt_tmp = read.table("/home/murphjes/BIOS6660/Genetics/chb_jpt.frq",
                         stringsAsFactors = F, header=T)
chb_jpt = chb_jpt_tmp[!is.na(chb_jpt_tmp$MAF), ]

#identify SNPs that appear in all three datasets
common_SNPs = intersect(intersect(ceu$SNP, yri$SNP), chb_jpt$SNP)

#create new CEU dataset only having common SNPs and sorted by SNP names
ceu_common = ceu[ceu$SNP %in% common_SNPs, ]
ceu_order = ceu_common[order(ceu_common$SNP), ]

#create new YRI dataset only having common SNPs and sorted by SNP names
yri_common = yri[yri$SNP %in% common_SNPs, ]
yri_order = yri_common[order(yri_common$SNP), ]
names(yri_order) = c( "CHRy", "SNPy", "A1y","A2y", "MAFy", "NCHROBSy")

#create new chb jpt dataset only having common SNPs and sorted by SNP names
chb_jpt_common = chb_jpt[chb_jpt$SNP %in% common_SNPs, ]
chb_jpt_order = chb_jpt_common[order(chb_jpt_common$SNP), ]
names(chb_jpt_order) = c( "CHRa", "SNPa", "A1a","A2a", "MAFa", "NCHROBSa")

#merge three datasets
alldata = cbind(ceu_order, yri_order, chb_jpt_order)

#check if merge was correct
all.equal(alldata$SNP, alldata$SNPy, alldata$SNPa)

## [1] TRUE
```

2 CEU vs YRI

```r
#CEU and YRI have some SNPs with different minor allele
alldata[alldata$A1==alldata$A2y, "MAF"] = 1 - alldata[alldata$A1==alldata$A2y, "MAF"]

#calculate absolute value of the MAF difference between each SNP
```

```
ceu_yri_diff = abs(alldata$MAF - alldata$MAFy)
hist(ceu_yri_diff)
```

Histogram of ceu_yri_diff