Genetic structure advanced practical

Yurii Aulchenko

February 18, 2013

Contents

1 Loading the libraries and data

Start R, load library GenABEL

```
library(GenABEL)
```

```
Loading required package: MASS
GenABEL v. 1.7-4 (February 03, 2013) loaded

Installed GenABEL version (1.7-4) is not the same as stable
version available from CRAN (1.7-3). Unless used intentionally,
consider updating to the latest CRAN version. For that, use
'install.packages("GenABEL")', or ask your system administrator
to update the package.
```

Load the data set

```
load("data/hm.RData")
ls()
class(df)
```

```
 [1] "df"  "old"
 [1] "gwaa.data"
 attr(,"package")
 [1] "GenABEL"
```

Characterize the data set:

Ex. 1 — How many people and SNPs are there in the data set?

Ex. 2 — Explore the phenotypic data. How many males/females are in the data? Look at the variable 'pop' – how many people from each population are present in the data set?

2 GWAS with genomic control

Use 'qtscore' function for GWAS of the phenotype 'phe'.

Answer to the following questions:

Ex. 3 — What is the mean and standard deviation of 'phe'?

Ex. 4 — Do you need to adjust for sex?

Ex. 5 — Perform GWAS of 'phe'. If you ignore the fact that you did not account for the population structure and look at un-corrected P-values, do you get GW-significant results?

Ex. 6 — What is the genomic control λ for naive GWAS?

Ex. 7 — Are there any significant results left after GC correction?

Ex. 8 — **Advanced** Estimate λ using other estimators, such as mean, median, and trimmed mean. How much these vary?

Ex. 9 — **Challenge** Try to judge which λ estimator is better.

3 Analyses accounting for population structure

Perform analyses accounting for population structure in different ways.

Ex. 10 — Perform stratified analysis using the 'qtscore' function. What is the GC λ? Do you get any GW-significant hits?

Ex. 11 — Estimate genomic kinship matrix (use 'ibs' function with default options to run IBS estimator; then do not forget to set the diagonal to 1 by 'diag(myIbsMatrix)<-1'). Plot the first two PCs. Perform PC-adjusted GWAS using the 'mlreg' function for analysis. If time permits, 1) vary the number of PCA included and see what happens to λ 2) try analysis with 'qtscore'. Explain the difference in results. with λ.

Ex. 12 — Use kinship matrix to estimate pseudo-heritability of 'phe' with 'polygenic', and then run mixed model with 'mmscore'. What is th estimate of pseudo-h^2? What is the resulting λ? Are results better/worse than these before? Why? If time permits, use correlation kinship for the same analysis.

Ex. 13 — What analysis method is the 'best' for this data set?

4 Appendix A: Answers to exercises

Answer (Ex. 1) — Run R commnds

```
nids(df)
nsnps(df)
 [1] 210
 [1] 529279
```

Answer (Ex. 2) — Run R commnds

```
table(phdata(df)$sex)
table(phdata(df)$pop)

   0   1
 105 105

 CEU CHB JPT YRI
  60  45  45  60
```

Answer (Ex. 3) — They are

```
mean(phdata(df)$phe)
```

```
sd(phdata(df)$phe)
 [1] 166.0977
 [1] 16.0122
```

Answer (Ex. 4) — Yes, you need to adjust for sex:

```
summary(lm(phe~sex,data=phdata(df)))

Call:
lm(formula = phe ~ sex, data = phdata(df))

Residuals:
    Min      1Q  Median      3Q     Max
-37.547 -11.887   0.114  11.912  31.952

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  160.313      1.460 109.798  < 2e-16 ***
sex           11.569      2.065   5.603 6.63e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.96 on 208 degrees of freedom
Multiple R-squared:  0.1311,    Adjusted R-squared:  0.127
F-statistic: 31.39 on 1 and 208 DF,  p-value: 6.633e-08
```

Answer (Ex. 5) — Yes, you get ~50000 GW-significant hits. Too many!

```
qts0 <- qtscore(phe~sex,data=df)
summary(qts0)
table(qts0[,"P1df"]<5e-8)
Summary for top 10 results, sorted by P1df
          Chromosome   Position Strand A1 A2   N      effB  se_effB chi2.1df
rs7168440         15 3053736555      +  T  C 210 -12.85670 1.155925 123.7086
rs7166206         15 3053755920      +  G  A 205 -12.80976 1.162289 121.4657
rs7183892         15 3053654692      +  T  C 210 -12.64348 1.147217 121.4625
rs6494411         15 3053611621      +  T  C 208 -12.47627 1.132840 121.2922
rs289818          15 3053571597      +  A  G 209 -12.60420 1.144676 121.2455
rs289816          15 3053570721      -  T  C 210 -12.52065 1.137800 121.0938
rs4411464         15 3053771183      +  T  C 210 -12.61157 1.152967 119.6476
```

```
rs4721415          7 1543802799      +  A  G 209  13.03773 1.198578 118.3236
rs7178111         15 3053851271      +  T  C 209 -12.37196 1.163472 113.0746
rs7178104         15 3053881343      +  T  C 209 -12.45988 1.175327 112.3854
                  P1df     effAB     effBB chi2.2df         P2df        Pc1df
rs7168440 9.757159e-29 -14.41471 -25.65463 124.0966 1.129204e-27 0.0003787244
rs7166206 3.021810e-28 -14.49570 -25.53945 121.8890 3.405146e-27 0.0004281220
rs7183892 3.026646e-28 -14.06609 -25.22785 121.7699 3.614195e-27 0.0004281963
rs6494411 3.297976e-28 -14.52451 -24.85527 121.8113 3.540073e-27 0.0004322039
rs289818  3.376500e-28 -13.76605 -25.14802 121.4349 4.273124e-27 0.0004333089
rs289816  3.644851e-28 -13.76605 -24.97999 121.3121 4.543623e-27 0.0004369199
rs4411464 7.555834e-28 -13.29581 -25.19636 119.7212 1.006617e-26 0.0004729014
rs4721415 1.472851e-27  15.05171  25.79182 118.9972 1.445717e-26 0.0005084603
rs7178111 2.078104e-26 -11.81512 -24.75983 113.1243 2.725011e-25 0.0006781148
rs7178104 2.941829e-26 -14.90122 -24.78222 113.4409 2.326075e-25 0.0007042824

 FALSE   TRUE
479302  49977
```

Answer (Ex. 6) — The λ as repoted by standard GenABEL procedure is huge:

```
lambda(qts0)$est
 [1] 9.791453
```

Answer (Ex. 7) — No (see Pc1df in previous output)

Answer (Ex. 8) — The estimates of λ vary strongly depending on the method used:

```
# regression 100%
c2 <- qts0[,"chi2.1df"]
estlambda(c2)$est
# regression 95%
estlambda(c2,prop=0.95)$est
# mean
mean(c2)
# median
median(c2)/qchisq(.5,1)
# trimmed mean at 95%
```

```
mean(sort(c2)[1:round(0.95*length(c2))])/((1/.95)*pchisq(qchisq(.95,df=1),df=3))
 [1] 9.791453
 [1] 11.04772
 [1] 10.99943
 [1] 13.64433
 [1] 11.72667
```

Answer (Ex. 9) — One of the ways to answer this question is to look how close is the fit of the corrected test statistic to the null distribution. From the tests below it follows that the trimmed mean option is the best. Note, however, that you are working with the data *without any association*, and your results, strictly speaking, can not be generalized to a real situation when true associations are present.

```
sW <- getOption("warn")
options(warn=-1)
# regression 100%
c2 <- qts0[,"chi2.1df"]
lambda <- estlambda(c2)$est
c2c <- c2/lambda
ks.test(c2c,"pchisq",df=1)$stat
# regression 95%
lambda <- estlambda(c2,prop=0.95)$est
c2c <- c2/lambda
ks.test(c2c,"pchisq",df=1)$stat
# mean
lambda <- mean(c2)
c2c <- c2/lambda
ks.test(c2c,"pchisq",df=1)$stat
# median
lambda <- median(c2)/qchisq(.5,1)
c2c <- c2/lambda
ks.test(c2c,"pchisq",df=1)$stat
# trimmed mean at 95%
lambda <- mean(sort(c2)[1:round(0.95*length(c2))])/((1/.95)*pchisq(qchisq(.95,df=1),df=
c2c <- c2/lambda
ks.test(c2c,"pchisq",df=1)$stat
options(warn=sW)
         D
0.07522724
```

```
         D
0.04939044
         D
0.05027084
         D
0.04587066
         D
0.03788888
```

Answer (Ex. 10) — Lambda is very reasonable; there is GW-significant hit. Run the R code

```
qts <- qtscore(phe~sex,data=df,strata=phdata(df)$pop)
lambda(qts)$est
summary(qts)
[1] 1.006406
Summary for top 10 results, sorted by P1df
           Chromosome   Position Strand A1 A2   N      effB   se_effB chi2.1df
rs10518394          1   72214606      +  C  T 210  6.579754 1.1440942 33.07468
rs10518395          1   72255233      +  T  C 210  6.409265 1.3215342 23.52121
rs7732608           5 1233959521      +  A  G 210 -2.743810 0.6037035 20.65668
rs4888646          16 3213669549      +  G  A 208  3.156958 0.7249879 18.96164
rs7513441           1   72196705      +  G  A 208  3.479265 0.8013470 18.85097
rs960529           12 2584310280      +  C  T 206  3.504380 0.8154756 18.46717
rs17721635          4  844039774      +  A  G 207  6.585803 1.5437516 18.19963
rs6751506           2  336097537      +  G  A 208  6.756108 1.6032181 17.75857
rs1860393          12 2494213167      +  G  A 208  3.064326 0.7385016 17.21738
rs3889228          16 3228175489      -  G  T 185 -2.574434 0.6205887 17.20901
                    P1df     effAB      effBB chi2.2df         P2df        Pc1df
rs10518394 8.868594e-09  7.286720  12.331776 36.36228 1.270667e-08 9.882894e-09
rs10518395 1.235445e-06  3.863802  21.732415 23.84868 6.627107e-06 1.335435e-06
rs7732608  5.494514e-06  9.549322  11.524601 21.18435 2.511170e-05 5.885152e-06
rs4888646  1.333728e-05  5.792755  14.970399 19.17415 6.860995e-05 1.420838e-05
rs7513441  1.413392e-05  9.081618  11.912046 23.25607 8.912677e-06 1.505173e-05
rs960529   1.728564e-05  3.318340   7.383783 20.68956 3.216015e-05 1.838556e-05
rs17721635 1.989170e-05 24.301819  29.796347 18.32933 1.046732e-04 2.113937e-05
rs6751506  2.507854e-05 11.706539  17.498339 17.89261 1.302172e-04 2.661401e-05
rs1860393  3.333723e-05  7.522871  15.054994 18.67697 8.797244e-05 3.531720e-05
rs3889228  3.348441e-05 -5.874787 -19.095723 17.30602 1.746003e-04 3.547218e-05
```

Answer (Ex. 11) — Run the code

```
gIbs <- ibs(df)
diag(gIbs) <- 1
pcIbs <- cmdscale(as.dist(1-gIbs),k=10)
mlr10 <-mlreg(phe~sex+pcIbs[,1:10],data=df)
estlambda(mlr10[,"chi2.1df"])$est
summary(mlr10)
[1] 0.9936808
Summary for top 10 results, sorted by P1df
           Chromosome   Position Strand A1 A2   N      effB   se_effB chi2.1df
rs10518394          1   72214606      +  C  T 210  6.690857 1.0809712 38.31204
rs10518395          1   72255233      +  T  C 210  6.860952 1.2829639 28.59828
rs960529           12 2584310280      +  C  T 206  3.984842 0.8018764 24.69491
rs6515824          20 3646777846      +  C  T 209  5.133286 1.1248409 20.82613
rs11077325         16 3144966713      +  G  A 210  5.767299 1.2699835 20.62285
rs3845906           3  655408823      -  T  C 204 -5.479251 1.2069263 20.61014
rs10050474          5 1103211963      +  A  G 209  7.882775 1.7407183 20.50696
rs4666808           2  479290594      +  T  C 210  6.730852 1.4882892 20.45340
rs4888646          16 3213669549      +  G  A 208  3.949398 0.8802285 20.13126
rs7732608           5 1233959521      +  A  G 210 -4.073474 0.9123340 19.93528
                   P1df        Pc1df effAB effBB chi2.2df P2df
rs10518394 6.028952e-10 5.321370e-10    NA    NA       NA   NA
rs10518395 8.906122e-08 8.107796e-08    NA    NA       NA   NA
rs960529   6.716145e-07 6.190641e-07    NA    NA       NA   NA
rs6515824  5.029198e-06 4.693223e-06    NA    NA       NA   NA
rs11077325 5.592464e-06 5.222245e-06    NA    NA       NA   NA
rs3845906  5.629728e-06 5.257255e-06    NA    NA       NA   NA
rs10050474 5.941490e-06 5.550217e-06    NA    NA       NA   NA
rs4666808  6.110102e-06 5.708700e-06    NA    NA       NA   NA
rs4888646  7.230521e-06 6.762459e-06    NA    NA       NA   NA
rs7732608  8.010849e-06 7.496959e-06    NA    NA       NA   NA
```

Answer (Ex. 12) — `h2 <- polygenic(phe~sex,data=df,kin=(gIbs/2),quiet=TRUE)`

```
h2$est
mmsIbs <- mmscore(h2,data=df)
lambda(mmsIbs)
summary(mmsIbs)
[1] 0.9999991
 $estimate
```

```
[1] 1.116662

$se
[1] 3.659768e-05
Summary for top 10 results, sorted by P1df
           Chromosome   Position Strand A1 A2   N      effB   se_effB chi2.1df
rs17675813         16 3206116502      +  G  A 188 -6.705347 1.2303213 29.70335
rs4772447          13 2774249004      +  G  A 180 -5.772128 1.0643268 29.41182
rs12441236         15 3053517961      +  G  A 185 -6.175968 1.1398250 29.35849
rs10976178          9 1939841622      +  C  T 185 -5.298327 0.9795148 29.25874
rs1863188           2  514484026      -  C  T 175 -5.490856 1.0260496 28.63804
rs9988693          10 2126599674      +  G  T 181  6.156263 1.1504974 28.63275
rs10518394          1   72214606      +  C  T 210  6.847941 1.2800644 28.61912
rs17615522          4  875799875      +  A  C 195 -5.695996 1.0816857 27.72919
rs749873            2  433156296      +  C  T 195  7.221541 1.4278989 25.57790
rs10988617          9 2064525746      +  C  T 192 -5.310638 1.0511805 25.52341
                   P1df        Pc1df effAB effBB chi2.2df P2df
rs17675813 5.034782e-08 2.502263e-07    NA    NA        0   NA
rs4772447  5.851958e-08 2.864359e-07    NA    NA        0   NA
rs12441236 6.015244e-08 2.936076e-07    NA    NA        0   NA
rs10976178 6.332997e-08 3.075077e-07    NA    NA        0   NA
rs1863188  8.725107e-08 4.101178e-07    NA    NA        0   NA
rs9988693  8.749009e-08 4.111273e-07    NA    NA        0   NA
rs10518394 8.810769e-08 4.137344e-07    NA    NA        0   NA
rs17615522 1.395409e-07 6.254298e-07    NA    NA        0   NA
rs749873   4.248779e-07 1.701477e-06    NA    NA        0   NA
rs10988617 4.370478e-07 1.745226e-06    NA    NA        0   NA
```

Answer (Ex. 13) — The best analysis method is stratified analysis and/or the PCA-based analysis. These do reflect the design of the study and produces reasonable GC λ.

5 Appendix B: Generation of the data set

```
library(GenABEL)
load("data/HapMap_r21_550K.RData")
chr <- chromosome(hapmap550k)
qc <- check.marker(hapmap550k[,(chr=="21" | chr=="X" | chr=="Y")])
phdata(hapmap550k)$sex[which(idnames(hapmap550k) %in% qc$ismale)] <- 1
hapmap550k@gtdata@male <- phdata(hapmap550k)$sex
```

```
df <- hapmap550k[,autosomal(hapmap550k)]
gIbs <- ibs(df,weight="no")
gIbd <- ibs(df,weight="freq")
attach(phdata(df))

nPolySnps <- 10000
popMu=list(CHB=160,JPT=165,YRI=170,CEU=180)
popVar=list(CHB=49,JPT=25,YRI=64,CEU=25)
betaSex <- 12
varG <- 2*25
set.seed(9)
polySnps <- sample(snpnames(df),nPolySnps)
polyX <- scale(as.numeric(gtdata(df[,polySnps])))
any(is.na(polyX))
polyX[is.na(polyX)] <- 0
any(is.na(polyX))
G <- as.vector( polyX %*% rep(0.1,nPolySnps) )
var(G)
G <- G + as.vector( polyX[,1:2] %*% c(25,30) )
colnames(polyX)[1:2]
var(G)
G <- G/sd(G)
G <- G*sqrt(varG)
phe <- sex*betaSex + G - 10
for (cPop in names(popMu)) {
  cIds <- which(pop == cPop)
  phe[cIds] <- phe[cIds] + popMu[[cPop]] + rnorm(length(cIds),sd=sqrt(popVar[[cPop]]))
}
phdata(df)$phe <- phe
summary(lm(phe~sex,data=phdata(df)))
summary(lm(phe~sex+pop,data=phdata(df)))
save(df,file="data/hm.RData")
```

```
Loading required package: MASS
GenABEL v. 1.7-4 (February 03, 2013) loaded

Installed GenABEL version (1.7-4) is not the same as stable
version available from CRAN (1.7-3). Unless used intentionally,
consider updating to the latest CRAN version. For that, use
'install.packages("GenABEL")', or ask your system administrator
```

```
to update the package.
 Excluding people/markers with extremely low call rate...
21182 markers and 210 people in total
0 people excluded because of call rate < 0.1
0 markers excluded because of call rate < 0.1
Passed: 21182 markers and 210 people

Running sex chromosome checks...
0 heterozygous X-linked male genotypes found
0 X-linked markers are likely to be autosomal (odds > 1000 )
0 male are likely to be female (odds > 1000 )
105 female are likely to be male (odds > 1000 )
0 people have intermediate X-chromosome inbreeding (0.5 > F > 0.5)
If these people/markers are removed, 0 heterozygous male genotypes are left
Warning in check.marker(hapmap550k[, (chr == "21" | chr == "X" | chr ==  :
  The number of Y-chromosome SNPs is low (6). Consider ignoring Y-checks by setting 'XX

 0 possibly female Y genotypes identified
None of these people excluded based on Y-threshold of 0.8
Passed: 21182 markers and 105 people

Checking Y-chromsome heterozygous genotypes... 0 ( NaN %) found.

no X/Y/mtDNA-errors to fix


RUN 1
21182 markers and 105 people in total
303 (1.43046%) markers excluded as having low (<2.380952%) minor allele frequency
635 (2.997828%) markers excluded because of low (<95%) call rate
5962 (28.14654%) markers excluded because they are out of HWE (FDR <0.2)
2 (1.904762%) people excluded because of low (<95%) call rate
Mean autosomal HET is 0.3194573 (s.e. 0.01776116)
0 people excluded because too high autosomal heterozygosity (FDR <1%)
Mean IBS is 0.7203834 (s.e. 0.02325525), as based on 2000 autosomal markers
0 (0%) people excluded because of too high IBS (>=0.95)
In total, 14481 (68.36465%) markers passed all criteria
In total, 103 (98.09524%) people passed all criteria

RUN 2
```

```
14481 markers and 103 people in total
26 (0.1795456%) markers excluded as having low (<2.427184%) minor allele frequency
0 (0%) markers excluded because of low (<95%) call rate
0 (0%) markers excluded because they are out of HWE (FDR <0.2)
0 (0%) people excluded because of low (<95%) call rate
Mean autosomal HET is 0.3203484 (s.e. 0.01720416)
0 people excluded because too high autosomal heterozygosity (FDR <1%)
Mean IBS is 0.7197077 (s.e. 0.02371458), as based on 2000 autosomal markers
0 (0%) people excluded because of too high IBS (>=0.95)
In total, 14455 (99.82045%) markers passed all criteria
In total, 103 (100%) people passed all criteria

RUN 3
14455 markers and 103 people in total
0 (0%) markers excluded as having low (<2.427184%) minor allele frequency
0 (0%) markers excluded because of low (<95%) call rate
0 (0%) markers excluded because they are out of HWE (FDR <0.2)
0 (0%) people excluded because of low (<95%) call rate
Mean autosomal HET is 0.3203484 (s.e. 0.01720416)
0 people excluded because too high autosomal heterozygosity (FDR <1%)
Mean IBS is 0.7234528 (s.e. 0.02411495), as based on 2000 autosomal markers
0 (0%) people excluded because of too high IBS (>=0.95)
In total, 14455 (100%) markers passed all criteria
In total, 103 (100%) people passed all criteria
[1] TRUE
[1] FALSE
[1] 2985.701
[1] "rs1377826"  "rs10518394"
[1] 4994.655

Call:
lm(formula = phe ~ sex, data = phdata(df))

Residuals:
    Min      1Q  Median      3Q     Max
-37.547 -11.887   0.114  11.912  31.952

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  160.313      1.460 109.798  < 2e-16 ***
```

```
sex           11.569      2.065   5.603 6.63e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.96 on 208 degrees of freedom
Multiple R-squared:  0.1311,    Adjusted R-squared:  0.127
F-statistic: 31.39 on 1 and 208 DF,  p-value: 6.633e-08

Call:
lm(formula = phe ~ sex + pop, data = phdata(df))

Residuals:
     Min       1Q   Median       3Q      Max
-21.3562  -4.7761  -0.0142   5.3910  21.3648

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  176.352      1.170 150.686   <2e-16 ***
sex           11.517      1.103  10.437   <2e-16 ***
popCHB       -32.230      1.577 -20.443   <2e-16 ***
popJPT       -26.742      1.577 -16.961   <2e-16 ***
popYRI       -11.816      1.460  -8.095    5e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.995 on 205 degrees of freedom
Multiple R-squared:  0.7555,    Adjusted R-squared:  0.7507
F-statistic: 158.3 on 4 and 205 DF,  p-value: < 2.2e-16
```