Bioinformatics - Homework 1 – Q&A style

Instructions: in this assignment you will test your understanding of basic GWAS concepts and GenABEL functions. The materials needed for the homework (two datasets and R function) are all available on the course homepage. You are asked to:

*Provide a **single report per group** with the (concise) answers to the following questions (include Introduction, Discussion and Results and Conclusion sections).*
Provide any additional code you used (if any) in Appendix or a separate file

---

Part 1: Preprocessing the genotype data

Preprocessing bioinformatics-related data is a mandatory step if one wants to perform a relevant analysis and highlight interesting relationships in the data. In this homework we will 1st try to get familiar with some GWA techniques and apply them on a real data. We will then focus on the preprocessing steps (i.e. quality control) at the sample level and at the marker level and we will finish with a brief association analysis using the cleaned data. To do so we will use the GenABEL R package for which a full tutorial is available at the http://mga.bionet.nsc.ru/~yurii/ABEL/GenABEL/.

After loading the GenABEL library and downloading data, you should load the data by running the following commands:

```
library(GenABEL);
load(HW1_WS.Rdata);
```

If everything is loaded correctly you should get the following messages:

```
ids loaded...
marker names loaded...
chromosome data loaded.
map data loaded...
allele coding data loaded...
strand data loaded...
genotype data loaded...
snp.data object created...
assignment of gwaa.data object FORCED; X-errors were not checked!
```

Before going through further data preprocessing steps, it is important to get familiar with the data we are dealing with. Using some basic GenABEL commands such as *str()* or *perid.summary()* you should be able to get some information about both genotypic and phenotypic data.

Question 1 :

- How many patients are represented in this study? Tip: run `dim(wgDat@phdata)`

- Here `controls` are coded with `affection = 1` and `cases` with `affection = 2`. How many controls are there in the sample? Tip: run `length(which(wgDat@phdata[,"affection"]==1))` and `length(which(wgDat@phdata[,"affection"]==2))` commands
- What is the sex of the patient with id "*NA18945*"? Tip: run `wgDat@phdata["NA18572",]`
- On which chromosome is the SNP "*rs4075116*" located? Tip: run descriptives.marker(wgDat, "rs4075116")
- Are there missing samples for this SNP (tip: check the call rate value)? Tip: run `descriptives.marker(wgDat, "rs4075116")`
- What is the minor allele for *rs4075116* bi-allelic SNP (tip: convert genotypes to allele counts)? Tip run `summary(wgDat@gtdata)["rs4075116",]` and look at documentation
- Are patients with ids "*NA18529*" and *"NA18968"* homozygous or heterozygous for that SNP? Tip: run `as.character(wgDat[c("NA18529", "NA18968"),"rs4075116"])`

Now that you are familiar with basic GenABEL functions we will go through the quality control steps. We will use the Travemünde criteria both at the sample and at the SNP (marker) level.

1.1 Sample (phenotypic) level

First let us focus on the two criteria about call fraction (proportion of correctly identified genotypes) and heterozygosity.

Question 2:

- What are the constraints (i.e. thresholds) on the "call fraction" and "heterozygosity" according to the Travemünde criteria (see your lecture notes)?
- What are the main factors that could impact heterozygosity of a given population? Discuss in the context of Hardy-Weinberg equilibrium (HWE) and allele segregation
- Why quality control is important in GWAS? Give several reasons and explain.

1.2 Gentotypes (SNP) level

The second part of the preprocessing consists of removing the SNPs that does not meet the Travemünde criteria.

Question 4 :

- What are the constraints/thresholds on MAF, MiF by group and HWE according to Travemünde criteria?
  - Note: MiF stands for Missing Frequency, i.e. the proportion of missing genotypes
- Explain those three criteria in your own words. What are the negative effects that could occur if those criteria are not met? What will be a possible impact on the final results interpretation?

```
idSummar <- perid.summary(wgDat);
hetMean <- mean(idSummar$Het);
hetSd <- sd(idSummar$Het);
hetThreshUpp <- (hetMean + 3*hetSd);
hetThreshLow <- (hetMean - 3*hetSd);
```

```
removeIdx <- with(idSummar, which(hetThreshLow>idSummar$Het,
hetThreshUpp<idSummar$Het, 0.97>idSummar$CallPP) );
idSummar$keep <- TRUE; idSummar$keep[removeIdx] <- FALSE;
keepIDs <-row.names(idSummar[idSummar$keep, ]);
wgDatIdClean <- wgDat[keepIDs, ];

casesIDs <- subset(wgDatIdClean@phdata, affection == 2, id, drop = TRUE);
controlsIDs <- subset(wgDatIdClean@phdata, affection == 1, id, drop = TRUE);
sumMaf <- summary(wgDat)$Q.2;
maf <- data.frame(maf=sumMaf);
sumControls <- summary(wgDatIdClean[controlsIDs, ])[,
                                          c("Pexact", "CallRate")];
names(sumControls) <- c("pHWE", "cfControls");

sumCases <- summary(wgDatIdClean[casesIDs, ])[, "CallRate", drop = FALSE];
names(sumCases) <- "cfCases";

snpSummar <- do.call(cbind, list(maf, sumControls, sumCases));

snpRemoveIdx <- with(snpSummar, which(snpSummar$maf<0.01,
snpSummar$cfControls < 0.98 & snpSummar$cfCases < 0.98 ,
snpSummar$pHWE < 1e-4));
snpSummar$keep <- TRUE;
snpSummar$keep[snpRemoveIdx] <- FALSE;

keepSNPs <- row.names(snpSummar[snpSummar$keep, ]);
wgDatClean <- wgDatIdClean[, keepSNPs];

mafSNPs <- row.names(subset(snpSummar, maf < 0.01));
hweSNPs <- row.names(subset(snpSummar, pHWE < 1e-04));
crSNPs <- row.names(subset(snpSummar, cfCases < 0.98 | cfControls < 0.98));
qcVenn(mafSNPs, hweSNPs, crSNPs, labels = c("MAF", "HWE", "CF"), numberSnps =
      nrow(snpSummar));
```

Question 4b:

- Look at the plot and find out what criteria were used? Please name them
- How many SNP and individuals left after cleaning? Tip: run `dim(wgDatClean@gtdata)`
- Looking at the Venn plot, how many SNPs were excluded according to both the MAF and CF criteria?
- Are the three criteria were optimally selected for our dataset?

Part 2: Introduction to association analysis

Now that the data is cleaned we can perform an association analysis. Run the following code to transform the affection status from 1/2 to 1/0 coding:

```
wgDatClean@phdata$aff.01 <- wgDatClean@phdata$affection-1;
with(wgDatClean@phdata, table(affection, aff.01));
```

We then test the hypothesis of a link between the trait and every remaining SNP using a linear regression model. Note that regression model eliminates intercept and considers y-intercept at the origin:

```
assocRes <- mlreg(aff.01  ~ 1, data = wgDatClean, gtmode = "additive",
      trait.type = "binomial");
plot(assocRes, main = "", ystart = 0, df=1, sort=T);
```

Question 5:

- What is the name of the resulting graph? What y and x axis refer to in this graph? If we consider an association being statistically significant at log(*p*-value) > 4.5, are there SNPs significantly associated to the trait according to the chosen threshold? If any, where are they located, what are their names (SNP ids)? Tip: run `summary(assocRes)[,1:10]` and look at "Pc1df" column