

Illustration of GWA data handling

Arne Schillert & Andreas Ziegler
Institut für Medizinische Biometrie und Statistik,
Universität zu Lübeck, Germany

Contents

1	Brief introduction into the R package GenABEL	1
2	Quality control	3
2.1	Sample level	3
2.2	Marker level	4
3	Association analysis	8

1 Brief introduction into the R package GenABEL

- GenABEL [1] is an R package designed for the analysis of genome-wide data.
- A comprehensive tutorial can be found on the web site <http://mga.bionet.nsc.ru/~yurii/ABEL/GenABEL/>
- Phenotypic and genotypic data kept separately.
- Phenotype data (`pheFile`) is a text file and genotype data (`rawFile`) is a binary file in which the genotype data is stored efficiently.
- We use an example data set provided by the developer of another GWA-toolchain termed PLINK [2].
- The original data is available at <http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml>
- The converted files are available on the course web page.
- After loading the GenABEL package with `library()` the data can be imported:

```
R> library(GenABEL)
R> pheFile <- "wgDat.phe"
R> rawFile <- "wgDat.raw"
R> wgDat <- load.gwaa.data(pheFile, rawFile)
```

```
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!
```

Using `str()` to show the structure, reveals that `wgDat` is of class `gwaa.data`. Details about this class can be found in the GenABEL internals section of the GenABEL tutorial.

```
R> str(wgDat)
```

```
Formal class 'gwaa.data' [package "GenABEL"] with 2 slots
..@ phdata:'data.frame':      90 obs. of  6 variables:
.. ..$ id      : chr [1:90] "NA18526" "NA18524" "NA18529" "NA18558" ...
.. ..$ sex     : int [1:90] 0 1 0 1 0 1 1 0 1 0 ...
.. ..$ fid     : chr [1:90] "CH18526" "CH18524" "CH18529" "CH18558" ...
.. ..$ mid     : int [1:90] 0 0 0 0 0 0 0 0 0 0 ...
.. ..$ pid     : int [1:90] 0 0 0 0 0 0 0 0 0 0 ...
.. ..$ affection: int [1:90] 1 1 1 1 1 1 1 2 2 1 ...
..@ gtdata:Formal class 'snp.data' [package "GenABEL"] with 11 slots
.. .. ..@ nbytes      : num 23
.. .. ..@ nids        : int 90
.. .. ..@ nsnpns      : int 228694
.. .. ..@ idnames     : chr [1:90] "NA18526" "NA18524" "NA18529" "NA18558" ...
.. .. ..@ snpnames    : chr [1:228694] "rs3094315" "rs6672353" "rs4040617" "rs2905036" ..
.. .. ..@ chromosome: Factor w/ 22 levels "1","10","11",...: 1 1 1 1 1 1 1 1 1 1 ...
.. .. .. ..- attr(*, "names")= chr [1:228694] "rs3094315" "rs6672353" "rs4040617" "rs29
.. .. ..@ map         : Named num [1:228694] 792429 817376 819185 832343 839326 ...
.. .. .. ..- attr(*, "names")= chr [1:228694] "rs3094315" "rs6672353" "rs4040617" "rs29
.. .. ..@ coding      :Formal class 'snp.coding' [package "GenABEL"] with 1 slots
.. .. .. ..@ .Data: raw [1:228694] 04 0b 04 01 ...
.. .. ..@ strand      :Formal class 'snp.strand' [package "GenABEL"] with 1 slots
.. .. .. ..@ .Data: raw [1:228694] 01 01 01 01 ...
.. .. ..@ male        : Named int [1:90] 0 1 0 1 0 1 1 0 1 0 ...
.. .. .. ..- attr(*, "names")= chr [1:90] "NA18526" "NA18524" "NA18529" "NA18558" ...
.. .. ..@ gtps        :Formal class 'snp.mx' [package "GenABEL"] with 1 slots
.. .. .. ..@ .Data: raw [1:23, 1:228694] 59 5a 56 55 ...
```

To get an idea of the data, we query for the number of cases and controls:

```
R> table(wgDat@phdata$affection)
```

```
 1  2
41 49
```

Here, controls are coded as 1 and cases as 2. We create a new phenotype variable which codes controls as 0 and cases as 1.

```
R> wgDat@phdata$aff.01 <- wgDat@phdata$affection - 1
R> with(wgDat@phdata, table(affection, aff.01))
```

```
      aff.01
affection 0  1
          1 41 0
          2  0 49
```

2 Quality control

The first step of quality control deals with the identification of failed arrays and/or subjects which show significant genetic differences, i.e., belong to a different ethnical group. These subjects are removed afterwards.

2.1 Quality control on the sample level

Call fraction and proportion of heterozygosity

For simplicity we exclude samples if

- The call fraction is $< 97\%$, or if
- The proportion of heterozygosity (poh) exceeds an interval of $3 \cdot s$ around the mean poh .

```
R> idSummar <- perid.summary(wgDat)
R> head(idSummar)
```

	NoMeasured	NoPoly	Hom	E(Hom)	Var
NA18526	227637	187689	0.7500450	0.7469145	0.5001890
NA18524	227727	187759	0.7512592	0.7469606	0.5232234
NA18529	228056	188040	0.7487284	0.7469029	0.4923326
NA18558	226352	186635	0.7496333	0.7466043	0.5211830
NA18532	227286	187376	0.7473712	0.7467768	0.4944136
NA18561	228215	188173	0.7488509	0.7468856	0.5199134
	F	CallPP	Het		
NA18526	0.012369340	0.9953781	0.2499550		
NA18524	0.016987728	0.9957716	0.2487408		
NA18529	0.007212723	0.9972102	0.2512716		
NA18558	0.011953545	0.9897592	0.2503667		
NA18532	0.002347274	0.9938433	0.2526288		
NA18561	0.007764444	0.9979055	0.2511491		

Columns CallPP and Het contain the call fraction (1- percentage of missing genotypes) and proportion of heterozygosity respectively.

After computing the thresholds for the proportion of heterozygosity the samples are determined which have to be excluded.

```
R> hetMean <- mean(idSummar$Het)
R> hetSd <- sd(idSummar$Het)
R> hetThreshLow <- hetMean - 3 * hetSd
R> hetThreshUpp <- hetMean + 3 * hetSd
R> removeIdx <- with(idSummar, which(CallPP < 0.97 |
      Het < hetThreshLow | Het > hetThreshUpp))
R> idSummar$keep <- TRUE
R> idSummar$keep[removeIdx] <- FALSE
R> keepIDs <- row.names(idSummar[idSummar$keep, ])
R> wgDatIdClean <- wgDat[keepIDs, ]
```

As can be seen from Figure 1, 4 samples are removed.

Genetic substructure

Aim of these steps is to detect population stratification. Therefore, the pairwise similarity based on IBS is computed. It is important to make sure that only autosomal SNPs are used in this step. Next, principal component analysis (PCA) or multidimensional scaling (MDS) is applied to group samples. For simplicity, we use MDS in our example although PCA is requested by several reviewers of papers submitted to journals with high impact factor.

Figure 2 shows the results of the multidimensional scaling. No obvious population substructure can be identified.

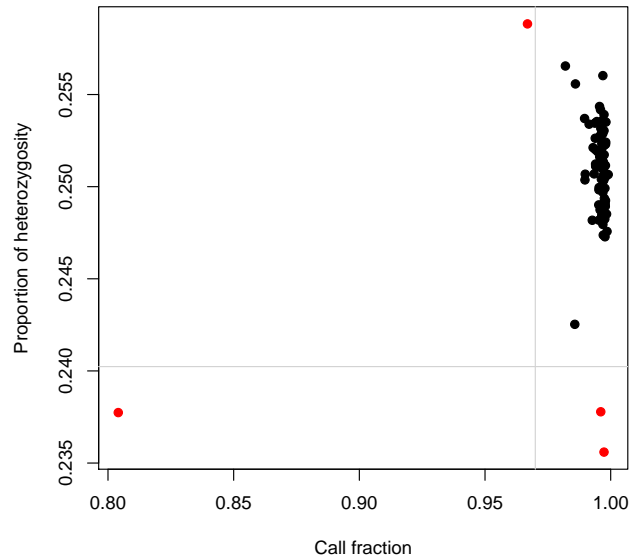


Figure 1: Call rate vs. rate of heterozygosity. The horizontal grey lines indicate the thresholds applied for filtering with rate of heterozygosity. The horizontal line indicates the call rate threshold of 97%. Samples which fail the criteria are marked in red.

2.2 Quality control on the marker level

On the marker level, filtering for minor allele frequency (MAF), call fraction (CF) and departure from Hardy-Weinberg equilibrium (HWE) belong to the standard quality control. To protect against informative missingness one computes the call rate for cases and controls separately. We exclude SNPs with

- $MAF < 0.01$, or
- $P(HWE) < 0.0001$, or
- $CF_{cases} < 0.98$ or $CF_{controls} < 0.98$.

As the departure from HWE is computed in controls only we have to run the function `summary` for cases, controls and all data. We combine the results to filter SNPs conveniently.

```
R> casesIDs <- subset(wgDatIdClean@phdata, affection ==
  2, id, drop = TRUE)
R> controlsIDs <- subset(wgDatIdClean@phdata, affection ==
  1, id, drop = TRUE)
R> sumMaf <- summary(wgDat)$Q.2
R> maf <- data.frame(maf = pmin(sumMaf, 1 - sumMaf))
```

```

R> pwSim <- ibs(wgDatIdClean)
R> pwDist <- as.dist(0.5 - pwSim)
R> mdsDat <- cmdscale(pwDist)
R> plot(mdsDat[, 1], mdsDat[, 2], xlab = "Component 1",
       ylab = "Component 2", pch = 19)

```

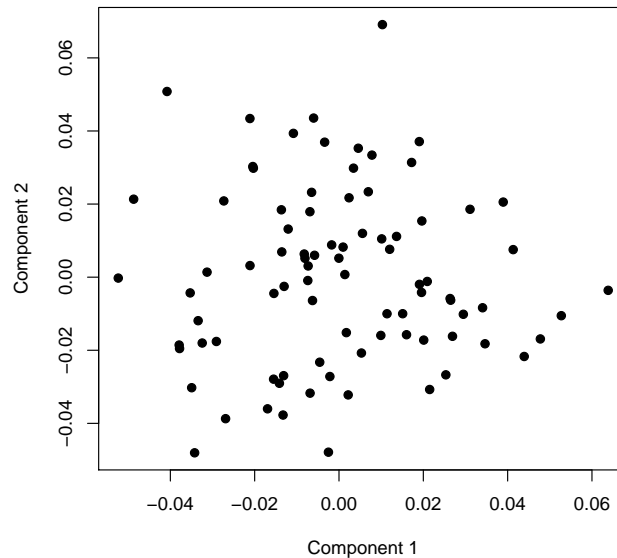


Figure 2: Multidimensional scaling of the pairwise IBS distances.

```

R> sumControls <- summary(wgDatIdClean[controlsIDs,
  ]) [, c("Pexact", "CallRate")]
R> names(sumControls) <- c("pHWE", "cfControls")
R> sumCases <- summary(wgDatIdClean[casesIDs, ]) [, "CallRate",
  drop = FALSE]
R> names(sumCases) <- "cfCases"
R> snpSummar <- do.call(cbind, list(maf, sumControls,
  sumCases))
R> snpRemoveIdx <- with(snpSummar, which(maf < 0.01 |
  pHWE < 1e-04 | cfCases < 0.98 | cfControls <
  0.98))
R> snpSummar$keep <- TRUE
R> snpSummar$keep[snpRemoveIdx] <- FALSE
R> keepSNPs <- row.names(snpSummar[snpSummar$keep, ])
R> wgDatClean <- wgDatIdClean[, keepSNPs]

```

We illustrate the effect of the various filters in a Venn diagram. We use the function `qcVenn` which is provided in the file `plot-VennDiagram.R`

```
R> source("plot-VennDiagram.R")
```

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

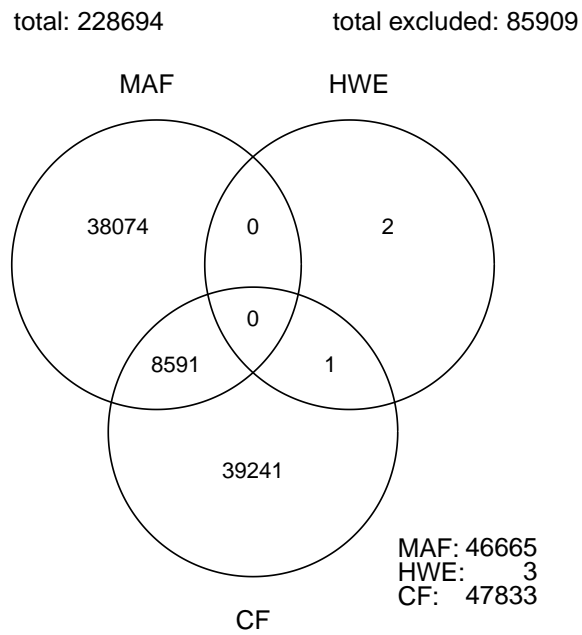


Figure 3: Venn diagram illustrating the effects of the SNP filters.

3 Association analysis

We test the hypothesis of association between the affection status and the genotype distribution per SNP using a logistic regression model.

```
R> assocRes <- mlreg(aff.01 ~ 1, data = wgDatClean,  
                    gtmode = "additive", trait.type = "binomial")  
R> plot(assocRes, main = "", ystart = 2)
```

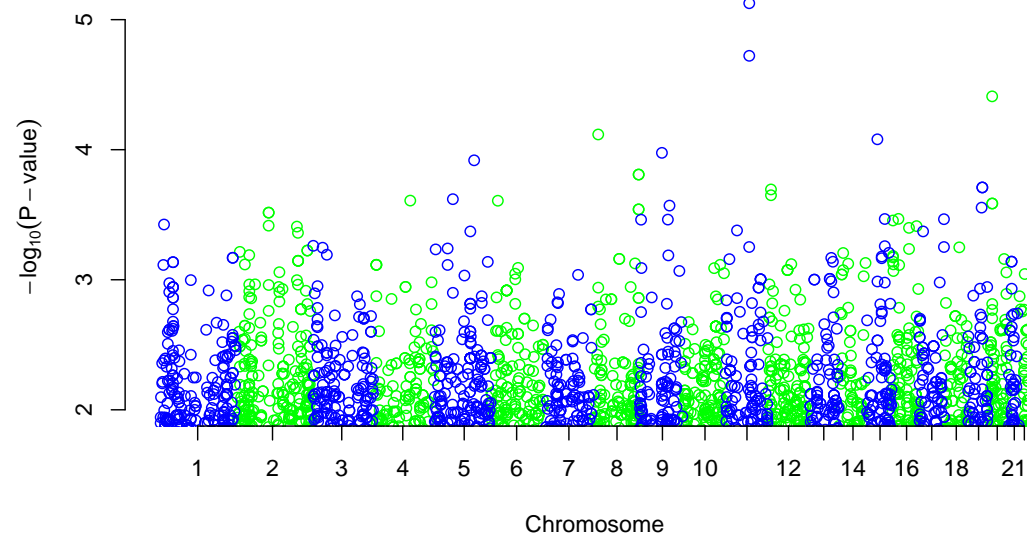


Figure 4: Manhattan plot for the single SNP analysis. For simplicity only SNPs with $-\log_{10}(p) > 2$ are plotted. (pdf or postscript graphics with $> 10,000$ points tend to become very large and slow to load.)

The result `assocRes` is of class `scan.gwaa-class`. See the documentation for the available extractor functions.

```
R> help("scan.gwaa-class")
```

Resources

This document was written using the following resources:

```
R> toLatex(sessionInfo(), locale = FALSE)
```

- R version 2.12.0 (2010-10-15), i486-pc-linux-gnu
- Base packages: base, datasets, graphics, grDevices, methods, stats, tools, utils
- Other packages: GenABEL 1.6-4, MASS 7.3-7

References

- [1] YS Aulchenko, S Ripke, A Isaacs, and CM van Duijn. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–6, 2007.
- [2] S Purcell, B Neale, K Todd-Brown, L Thomas, MAR Ferreira, D Bender, J Maller, P Sklar, PIW de Bakker, MJ Daly, and PC Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3): 559–575, 2007.