Genome-wide association analysis with GenABEL : quck start guide for the impatient

Yurii Aulchenko

August 24, 2009

Contents

1 What is GenABEL ?

GenABEL is an R library developed to facilitate Genome-Wide Association (GWA) analysis of binary and quantitative traits. R is a free, open source language and environment for general-purpose statistical analysis (available at R-project website[1]). It implements powerful data management and analysis tools. Though it is not strictly necessary to learn everything about R to run GenABEL, it is highly recommended as this knowledge will improve flexibility and quality of your analysis.

Originally GenABEL was developed to facilitate GWA analysis of quantitative traits using data coming from extended families and/or collected form genetically isolated populations. At the same time GenABEL implements a large number of procedures used in analysis of population-based data; it supports analysis of binary and quantitative tarits, and of survival (time-till-event) data. Most up-to-date information about GenABEL can be found at the web-site http://mga.bionet.nsc.ru/nlru/GenABEL/.

GenABEL is a part of more extensive ABEL collection (http://mga.bionet.nsc.ru/~yurii/ABEL/) of software supporting different kinds of GWA analyses.

[1] http://www.r-project.org/

GenABEL is easy to install and keep updated: you need to install R (http://www.r-project.org/), start it and install the GenABEL package (which can be done by the command install.packages("GenABEL")).

In this overview, we will

- load and explore GWA data
- perform basic quality control
- investigate the phenotypes and covariates
- perform GWA analysis

2 Loading and exploring the data

Copy the data file ge03d1p3.RData to the desktop. This file contains an R "workspace" – a collection of R objects. Double-click on the file to start R and load the data. After R started and loaded the data, you will see R command prompt ">". Next, load the GenABEL library by typing on the R command line

```
> library(GenABEL)
```

You can check what data objects has been loaded by typing ls():

```
> ls()

[1] "gwadat0"
```

There is a single data object, "gwadat0", which contains data on GWA study performed in a small number of individuals. To check the number of people in the study, use

```
> gwadat0@gtdata@nids

[1] 176
```

and to check the number of SNPs which were typed, use

```
> gwadat0@gtdata@nsnps

[1] 309470
```

A more detailed description of the structure of GWA data as implemented in GenABEL can be found in Figure 1, page 4, or in the "ABEL-tutorial"[2])

A general summary of genotypic data can be generated with

```
> descriptives.marker(gwadat0)
```

[2]to be bound at http://mga.bionet.nsc.ru/nlru/GenABEL/

```
$`Minor allele frequency distribution`
     X<=0.01 0.01<X<=0.05 0.05<X<=0.1 0.1<X<=0.2      X>0.2
No   465.000     12478.00    33980.00  71530.000 191017.000
Prop   0.002         0.04        0.11      0.231      0.617

$`Cumulative distr. of number of SNPs out of HWE, at different alpha`
     X<=1e-04 X<=0.001  X<=0.01   X<=0.05  all X
No         27  204.000 2308.000 11843.000 309470
Prop        0    0.001    0.007     0.038      1

$`Distribution of porportion of successful genotypes (per person)`
     X<=0.9 0.9<X<=0.95 0.95<X<=0.98 0.98<X<=0.99  X>0.99
No    2.000           0        1.000        8.000 165.000
Prop  0.011           0        0.006        0.045   0.938

$`Distribution of porportion of successful genotypes (per SNP)`
     X<=0.9 0.9<X<=0.95 0.95<X<=0.98 0.98<X<=0.99     X>0.99
No        0    2436.000     8733.000     40080.00 258221.000
Prop      0       0.008        0.028         0.13      0.834

$`Mean heterozygosity for a SNP`
[1] 0.3472016

$`Standard deviation of the mean heterozygosity for a SNP`
[1] 0.1321091

$`Mean heterozygosity for a person`
[1] 0.3440069

$`Standard deviation of mean heterozygosity for a person`
[1] 0.006399047
```

Here, a number of important statistics are present. From the "Minor allele frequency distribution" section you can see that the markers presented on the panel generally have rather high minor allele frequency, e.g. 62% of markers have MAF>0.2 and only 4% have MAF lower than 0.05.

From "Distribution of proportion of successful genotypes (per SNP)" you can see that most SNPs worked pretty well (96% of SNPs had a call rate of 98% or more).

From the table "Distribution of proportion of successful genotypes (per person)" you can see that most of the arrays worked well and had a call rate over 98%, though for two samples the call rate was indeed low (<90%).

Table provided under the caption "Cumulative distribution of number of SNPs out of HWE, at different alpha" shows how many markers deviate from Hardy-Weinberg equilibrium at certain α. You can see, for example, that 11843 (3.8%) markers demonstrate HWE p-value of less than 5%. In general, the proportion of markers showing deviation from HWE is in good agreement with the α threshold. If the observed proportions were much higher then the thresholds, this could have indicated a potential problems with genotyping or calling pro-

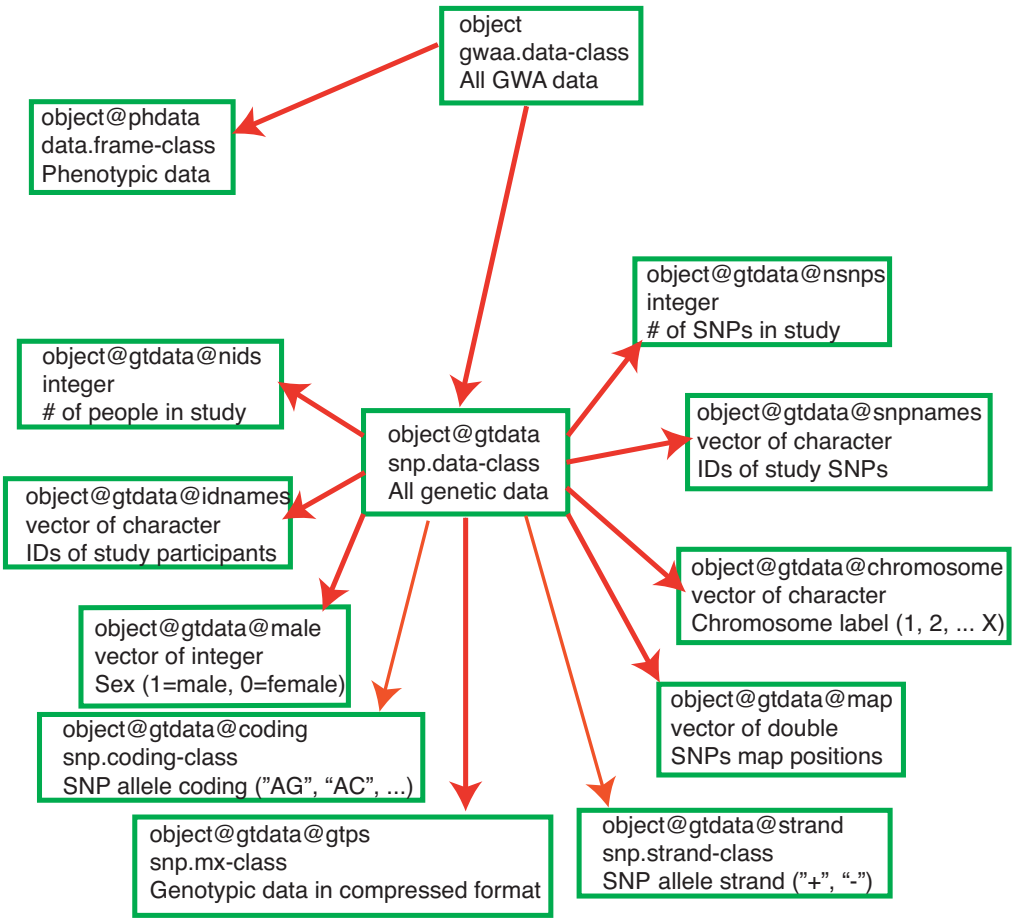

Figure 1: Structure of gwaa.data-class. In every box, first line contains the object and slot names, second line describes the class of this object, and third line describes what information is contained.

cedure, or indicate a population stratification (see "ABEL-tutorial"[3] for more details).

Inspect the output of "`descriptives.marker`" function and answer the following questions:

Ex. 1 — How many samples have a call rate less than 98%?

Ex. 2 — How many SNPs deviate from HWE with p-value $< 10^{-4}$?

Ex. 3 — How many rare SNPs (MAF $< 1\%$) are presented in the study?

From observing the output of the "`descriptives.marker`" function you can draw a conclusion that at the first glance the GWA data under consideration have no severe quality problems.

However, before proceeding with GWA we need to do QC and remove some SNPs and samples from our consideration.

3 Basic Quality Control

The basic procedure of `GenABEL` which performs genetic data quality control (QC) is "`check.marker`". This procedure identifies samples and SNPs which should be removed from the data according to a number of different criteria. For individual samples the checks performed include call rate across all SNPs, excess heterozygosity, identity between pairs of samples, mismatch between reported and genetically determined sex, etc. For individual SNPs, the checks include call rate across all individuals, number of copies of rare allele in the sample (or minor allele frequency, MAF), test fo Hardy-Weinberg equilibrium (HWE) p-value, etc.

For most thresholds used in QC, the default values specified in "`check.marker`" can be used in most of the studies. Few parameters, like sample and SNP call rate thresholds, however, depend on specific GWA chips used. Next to it, there is no conventional "accepted" threshold for filtering based on deviation from HWE.

In this example, Illumina data are used. Therefore we will set the SNP and sample call rate thresholds to 98%. We will also use somewhat arbitrary HWE p-value threshold of 10^{-8}. Such low p-value would usually suggest severe problems with SNP quality, for example, presence of a third allele.

To run QC with these thresholds, use the command

```
> qc0 <- check.marker(gwadat0, call = 0.98, perid.call = 0.98,
+     maf = 1e-08, p.lev = 1e-08)

Excluding people/markers with extremely low call rate...
309470 markers and 176 people in total
0 people excluded because of call rate < 0.1
0 markers excluded because of call rate < 0.1
```

[3] to be bound at `http://mga.bionet.nsc.ru/nlru/GenABEL/`

```
Passed: 309470 markers and 176 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 )
0 female are likely to be male (odds > 1000 )
If these people/markers are removed, 0 heterozygous male genotypes are left
Passed: 309470 markers and 176 people

no X/Y/mtDNA-errors to fix


RUN 1
309470 markers and 176 people in total
24 (0.007755194%) markers excluded as having low (<1e-06%) minor allele frequency
11169 (3.609074%) markers excluded because of low (<98%) call rate
0 (0%) markers excluded because they are out of HWE (P <1e-08)
3 (1.704545%) people excluded because of low (<98%) call rate
Mean autosomal HET is 0.3463363 (s.e. 0.004523418)
0 people excluded because too high autosomal heterozygosity (FDR <1%)
Mean IBS is 0.719174 (s.e. 0.008416355), as based on 2000 autosomal markers
0 (0%) people excluded because of too high IBS (>=0.95)
In total, 298279 (96.38382%) markers passed all criteria
In total, 173 (98.29545%) people passed all criteria

RUN 2
298279 markers and 173 people in total
0 (0%) markers excluded as having low (<1e-06%) minor allele frequency
0 (0%) markers excluded because of low (<98%) call rate
0 (0%) markers excluded because they are out of HWE (P <1e-08)
0 (0%) people excluded because of low (<98%) call rate
Mean autosomal HET is 0.3463762 (s.e. 0.004471527)
0 people excluded because too high autosomal heterozygosity (FDR <1%)
Mean IBS is 0.7201139 (s.e. 0.008561709), as based on 2000 autosomal markers
0 (0%) people excluded because of too high IBS (>=0.95)
In total, 298279 (100%) markers passed all criteria
In total, 173 (100%) people passed all criteria
```

Inspect the output and answer:

Ex. 4 — How many SNPs and samples do pass QC?

Ex. 5 — How many SNPs had HWE p-value $< 10^{-8}$?

Summary of excluded samples and SNPs can be generated with the command

```
> summary(qc0)

$`Per-SNP fails statistics`
          NoCall NoMAF NoHWE Redundant Xsnpfail
```

```
NoCall     11167     2     0         0        0
NoMAF         NA    22     0         0        0
NoHWE         NA    NA     0         0        0
Redundant     NA    NA    NA         0        0
Xsnpfail      NA    NA    NA        NA        0

$`Per-person fails statistics`
         IDnoCall HetFail IBSFail isfemale ismale isXXY
IDnoCall        3       0       0        0      0     0
HetFail        NA       0       0        0      0     0
IBSFail        NA      NA       0        0      0     0
isfemale       NA      NA      NA        0      0     0
ismale         NA      NA      NA       NA      0     0
isXXY          NA      NA      NA       NA     NA     0
```

These tables provide a summary of SNPs and samples which failed to pass certain QC thresholds; the view is pair-wise because some SNPs/samples may fail on several tests.

Ex. 6 — How many SNPs had call rate $< 98\%$ ("NoCall") and, at the same time, had very low MAF ("NoMAF")?

Ex. 7 — How many SNPs had call rate $< 98\%$?

It is important to know that "`check.marker`" function **does not** modify the data, it rather collects summary statistics and identifies samples and SNPs which do pass QC thresholds. To generate a new data set including only the QCed samples and SNPs, use

```
> gwadat1 <- gwadat0[qc0$idok, qc0$snpok]
> save(gwadat1, file = "clean.RData")
```

Here, "`qc0$idok`" is a vector containing IDs of samples which do pass QC thresholds, and "`qc0$snpok`" contains the list of the SNPs. The object "`gwadat1`" now contains QCed data.

Ex. 8 — Generate summary of QCed data using "`descriptives.marker`" command and tell what proportion of SNPs had HWE p-value $< 5\%$.

At this point, we have QCed GWA data. However, before doing GWA we need to inspect the phenotypes we will analyse.

4 Investigation of the phenotype

First of all, let us check what variables are presented in the data frame "`gwadat1@phdata`" by asking their names with

```
> names(gwadat1@phdata)

[1] "id"   "sex"  "age"  "quat" "bint"
```

There are 5 variables present: ”id” corresponds to the unique subject identification string, ”sex” is gender, ”age” contains age of the subject at the time of the study. The two variables of interest are names ”bint” (BINary Trait) and ”quat” (QUAntitative Trait). In this example, we will investigate the binary trait ”bint”. This trait describes the case/control status (for cases, the value of ”bint” is ”1” and for controls it is ”0”).

Simple summary over all variables in the data frame can be generated by

```
> summary(gwadat1@phdata)
```

```
      id                 sex              age              quat
 Length:173         Min.   :0.0000   Min.   :18.44   Min.   :-3.55855
 Class :character   1st Qu.:0.0000   1st Qu.:39.14   1st Qu.:-0.03143
 Mode  :character   Median :0.0000   Median :49.69   Median : 1.40402
                    Mean   :0.3931   Mean   :49.23   Mean   : 1.57407
                    3rd Qu.:1.0000   3rd Qu.:60.95   3rd Qu.: 3.15077
                    Max.   :1.0000   Max.   :84.38   Max.   : 8.93472
                                                     NA's   : 2.00000
      bint
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.4035
 3rd Qu.:1.0000
 Max.   :1.0000
 NA's   :2.0000
```

Ex. 9 — What is the mean age of study participants?

Ex. 10 — How many people do not have case/control status?

Ex. 11 — What is control-to-case ratio in the study?

Now, we will ”attach”[4] the data frame containing the data

```
> attach(gwadat1@phdata)
```

Now, let us inspect the number of cases and controls in the sample by generating a table

```
> table(bint)
```

```
bint
  0   1
102  69
```

[4]attaching the data allows accessing the variables by direct reference to their names, without typing the name of the object (gwadat1@phdata) they are contained within

There are 102 controls and 69 cases in the study.

As usual, you need to check the relation between the study trait and important covariates, such as sex and age. Such, and other "environmental" covariates need to be included to analysis because that may increase the power and obtain adjusted estimates of the genetic effects. Let us investigate the relation between "bint" and sex by generating 2 x 2 table:

```
> t <- table(bint, sex)
> t

    sex
bint  0  1
   0 72 30
   1 32 37
```

Here, we generated the 2 x 2 table and saved it to the object named "t", which was then displayed. There is a strong relation between the sex and binary trait under the study, which can be statistically tested using the Fisher's Exact Test

```
> fisher.test(t)

        Fisher's Exact Test for Count Data

data:  t
p-value = 0.002293
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.398791 5.509975
sample estimates:
odds ratio
  2.757664
```

Ex. 12 — Is relation between sex and bint significant? What is p-value?

Ex. 13 — What is the Odds Ratio?

Next, let us investigate the relation between "bint" and age by visualising that with a box-plot by:

```
> boxplot(age ~ bint)
```

(the box-plot is presented in Figure 2) and testing the significance of age differences between cases and controls using the T-test:

```
> t.test(age ~ bint)

        Welch Two Sample t-test

data:  age by bint
t = 1.0346, df = 166.81, p-value = 0.3024
alternative hypothesis: true difference in means is not equal to 0
```

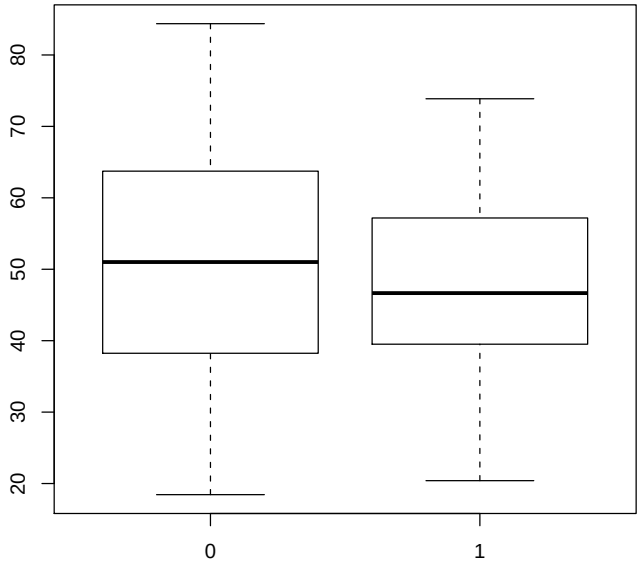

Figure 2: Relation between case-control status and age

```
95 percent confidence interval:
 -2.030867  6.502543
sample estimates:
mean in group 0 mean in group 1
       50.23806        48.00223
```

Ex. 14 — Is relation between age and bint significant? What is p-value?

From our analysis we can conclude that it is necessary to include sex as covariate into further analysis of the binary trait.

We are not going to access the data in `gwadat1@phdata` directly any more, therefore we will now "detach" the data with

```
> detach(gwadat1@phdata)
```

5 Genome-wide association analysis

Genome-wide association analysis of the binary trait, pre-adjusted for sex, can be run in `GenABEL` using the score test:

```
> qts <- qtscore(bint ~ sex, gwadat1, trait = "binomial")
```

A genome-wide manhattan plot for the data can be produced with "`plot(qts)`" (try it).

A summary of "top" hits can be obtained with

```
> descriptives.scan(qts, dig = 20)
```

```
           Chromosome  Position   N      effB         P1df         Pc1df
rs10111989          8 128971174 171 3.7859033 5.341616e-12 1.495810e-11
rs4733579           8 128967351 171 2.7956261 1.694318e-08 3.396388e-08
rs598296            9  14333670 170 2.7793316 1.507373e-06 2.512607e-06
rs2297646          10  28265903 171 0.3765958 6.054547e-06 9.535274e-06
rs1632673          17  75560331 170 3.5770861 6.082894e-06 9.578094e-06
rs2344843          15  84173904 171 0.3972320 1.575991e-05 2.387297e-05
rs837230            8 131055045 171 0.4413160 1.708659e-05 2.579782e-05
rs11195943         10 114144805 171 0.2489879 1.994463e-05 2.992456e-05
rs6770825           3 196168620 171 2.1192931 2.086939e-05 3.125453e-05
rs1356774          12  88034998 169 2.0708702 2.170690e-05 3.245699e-05
               effAB      effBB         P2df
rs10111989 4.0248674 19.9656533 4.642146e-11
rs4733579  2.3850313 12.9212223 5.739041e-08
rs598296   3.7465974  4.3204618 2.775543e-06
rs2297646  0.4123700  0.1480807 3.460039e-05
rs1632673  3.3964513        Inf 3.526287e-05
rs2344843  0.2952316  0.2544411 9.326743e-06
rs837230   0.4640669  0.2121837 9.331538e-05
rs11195943 0.2746943  0.1772771 8.558426e-05
rs6770825  2.9163433  5.1256084 8.961915e-05
rs1356774  2.6760870  4.4529373 9.078105e-05
```

In this summary table, the chromosome, genomic position is base pairs, and number of study participants for which gneotypes were available ("N") is listed. "P1df" corresponds to the p-value from the score test, while "Pc1df" provides corrected (by genomic control) p-value. Inflation is small for this study, and for simplicity, we will concentrate of non-corrected p-values. "effB" corresponds to the additive effect of the tested allele. When "binomial" option used, the effects are reported on logit scale – so to get Odds Ratio, you need to exponentiate the "effB" value. However, it is advised to estimate effects using logistic regression, either using "`mlreg`"[5] function of `GenABEL` or logistic regression implemented in base R (see the end of this exercise).

Ex. 15 — According to the threshold of $p<5 \cdot 10^{-8}$, are there genome-wide significant results in the scan?

Ex. 16 — What is the most significantly associated SNP name?

Ex. 17 — What is the chromosome and location of the most significantly associated SNP?

The sample under the study is relatively small. Therefore asymptotic p-values are likely to be wrong. In such situation, it is very important to access

[5] see `help(mlreg)`

empirical significance. Empirical genome-wide significance can be obtained by running the "qtscore" command with "times" argument, which tells how many permutations should be performed. Let us run this procedure using 200 permutations:

```
> qts.emp <- qtscore(bint ~ sex, gwadat1, trait = "binomial", times = 200)
```

here, for every SNP genome-wide significance was estimated in 200 permutation experiments. Summary of empirical results can be generated with

```
> descriptives.scan(qts.emp)
```

```
           Chromosome  Position   N      effB        P1df       Pc1df     effAB
rs4733579           8 128967351 171 2.7956261 0.004975124 0.004975124 2.3850313
rs10111989          8 128971174 171 3.7859033 0.004975124 0.004975124 4.0248674
rs598296            9  14333670 170 2.7793316 0.155000000 0.250000000 3.7465974
rs2297646          10  28265903 171 0.3765958 0.575000000 0.715000000 0.4123700
rs1632673          17  75560331 170 3.5770861 0.575000000 0.715000000 3.3964513
rs2344843          15  84173904 171 0.3972320 0.900000000 0.950000000 0.2952316
rs837230            8 131055045 171 0.4413160 0.910000000 0.975000000 0.4640669
rs6770825           3 196168620 171 2.1192931 0.940000000 0.990000000 2.9163433
rs11195943         10 114144805 171 0.2489879 0.940000000 0.990000000 0.2746943
rs1356774          12  88034998 169 2.0708702 0.940000000 0.995000000 2.6760870
               effBB        P2df
rs4733579  12.9212223 0.010000000
rs10111989 19.9656533 0.004975124
rs598296    4.3204618 0.250000000
rs2297646   0.1480807 0.990000000
rs1632673         Inf 0.990000000
rs2344843   0.2544411 0.665000000
rs837230    0.2121837 1.000000000
rs6770825   5.1256084 1.000000000
rs11195943  0.1772771 1.000000000
rs1356774   4.4529373 1.000000000
```

Please note that because the empirical procedure is based on random sampling, your results may deviate from these presented in this manual a little. when "times" argument is used, "P1df" shows not the nominal, but the empirical genome-wide p-values obtained in specified number of permutations.

Ex. 18 — Do you observe any SNPs which are significant at genome-wide p of less than 5%?

We can investigate the best associated SNP, rs10111989, in more details using the "summary" function, which reports the frequency of "B" allele ("Q.2"), genotypic distribution, HWE p-value ("Pexact"), and other details:

```
> summary(gwadat1@gtdata[, "rs10111989"])
```

```
           NoMeasured CallRate       Q.2 P.11 P.12 P.22    Pexact      Fmax
rs10111989        173        1 0.2485549  101   58   14 0.2195623 0.1025045
               Plrt Chromosome
rs10111989 0.1860247          8
```

Ex. 19 — What is the call rate for the SNP rs10111989?

Ex. 20 — Does this SNP significantly deviate from HWE?

Check genomic context around the best SNP using

```
> show.ncbi("rs10111989")
```

and answer the follwoing questions:

Ex. 21 — In what gene is the SNP located?

Ex. 22 — Where in the gene it is located (intron. exon, ...)?

If time permits, proceed to the next section.

6 Additional questions

We can see association to the region on chromosome 8 in more details. For this, let us first select the SNPs in the region

```
> reg <- gwadat1@gtdata@snpnames[gwadat1@gtdata@chromosome == "8" &
+     gwadat1@gtdata@map > (128971174 - 250000) & gwadat1@gtdata@map <
+     (128971174 + 250000)]
```

here, we have selected all SNPs, which map to chromosome 8, and their map position deviates from 128,971,174 (position of rs10111989) by no more than 250 kbp. The number of SNPs in this region is the length of the vector which contains the SNP names:

```
> length(reg)

[1] 75
```

Thus, 75 SNPs are located in 500 kbp region surrounding rs10111989.

We can perform analysis of this selected region now:

```
> qts.reg <- qtscore(bint ~ sex, gwadat1[, reg], trait = "binomial")
```

and depict that graphically:

```
> plot(qts.reg)
```

The resulting graph is shown in figure 3. You can see that association is supported by multiple, though not reaching genome-wide significance, hits presented in the region. It is important because hits which are generated by a genotyping error usually lack such a support.

Another point worth investigation is effect estimates. This is especially important for this small study, where big effects are observed, – a situation when the score test implemented by "`qtscore`" is likely to generate results different from these generated by logistic regression.

Let us check what effect estimates and p-values are obtained if logistic regression is used. For that, let us first convert the genotypic data for rs10111989 to numeric format, which can be utilised by standard R functions:

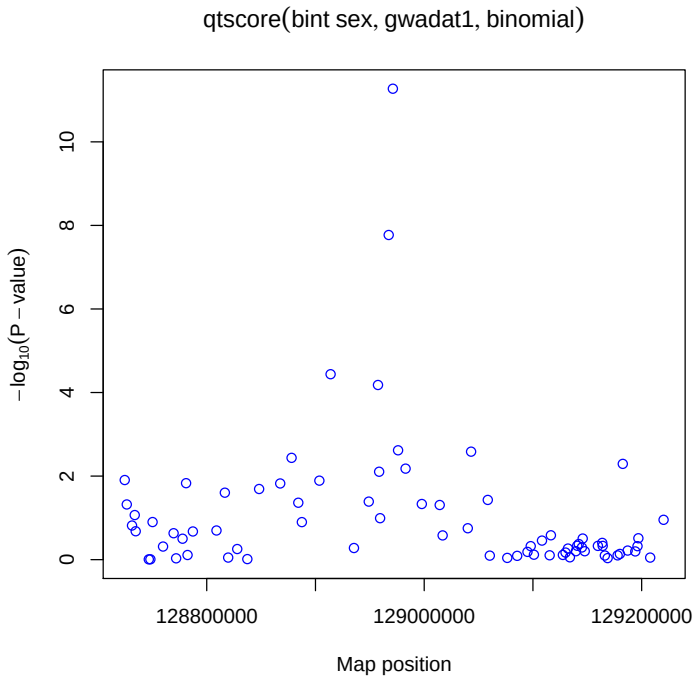

Figure 3: Details of association to the 500kb region on chromosome 8

```
> gt <- as.numeric(gwadat1@gtdata[, "rs10111989"])
```

Now we can run logistic regression with

```
> summary(glm(bint ~ sex + gt, data = gwadat1@phdata, family = binomial))

Call:
glm(formula = bint ~ sex + gt, family = binomial, data = gwadat1@phdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9365  -0.9546  -0.4496   0.8736   2.1642

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.2408     0.3787  -5.918 3.27e-09 ***
sex           1.6913     0.4322   3.914 9.09e-05 ***
gt            2.2580     0.3908   5.777 7.59e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 230.65  on 170  degrees of freedom
```

```
Residual deviance: 168.22  on 168  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 174.22

Number of Fisher Scoring iterations: 5
```

You can see that significance estimated using logistic regression is lower than that estimated using the score test; also the estimate of the effect is lower compared to the score test.

Ex. 23 — Perform GWA analysis without adjustment for sex. What is the difference compared to adjusted analysis?

Ex. 24 — What is possible explanation for that?

Advanced questions:

Ex. 25 — Characterise statistically relation between traits "bint" and "quat". Is there a significant relation?

Ex. 26 — Perform GWA analysis of the quantitative trait "quat". Is there a SNP which is genome-wide significantly associated with the trait?

Ex. 27 — Perform analysis of association between "bint" and rs10111989, adjusting for "quat". Is there still significant association? What is possible explanation of the results?

7 Answers to questions

Answer (ex. 1) — Three sample have call rate <98%

Answer (ex. 2) — Twenty-seven SNPs deviate from HWE with p-value $< 10^{-4}$

Answer (ex. 3) — 465 SNPs have MAF<1%

Answer (ex. 4) — 173 samples and 298279 SNPs pass the QC

Answer (ex. 5) — Unexpectedly, zero SNPs had HWE p-value $< 10^{-8}$ (actually just because you have received partly "cleaned" data set)

Answer (ex. 6) — Two

Answer (ex. 7) — 11167 + 2 + ... = 11169

Answer (ex. 9) — Mean age is 49.23

Answer (ex. 10) — 2 people have case/control status missing (NA)

Answer (ex. 11) — The proportion of cases is the same as mean value of the "bint" variable. Therefore control-to-case ratio is $\frac{1-mean(bint)}{mean(bint)}$, which is 1.48

Answer (ex. 12) — It is significant: p-value of 0.002 is < 0.05

Answer (ex. 13) — Odds ratio is equal to 2.76

Answer (ex. 14) — It is not significant: p-value of 0.3 is > 0.05

Answer (ex. 15) — Yes, two SNPs generate genome-wide significant association

Answer (ex. 16) — The "top" associated SNP is rs10111989

Answer (ex. 17) — SNP rs10111989 is located on chromosome 8, at the position 128,971,174 base pairs

Answer (ex. 18) — Yes, two SNPs located on chromosome 8 show empirical genome-wide significance with $p < 5\%$

Answer (ex. 19) — Call rate for the SNP rs10111989 is 100%

Answer (ex. 20) — The HWE p-value is 0.22, therefore genotypic proportions are in good agreement with HWE

Answer (ex. 21) — SNP rs10111989 is located in a gene called *PVT1*

Answer (ex. 22) — SNP rs10111989 is located in an intron of *PVT1*

Answer (ex. 23) — If analysis is not adjusted for sex, the results become less significant:

```
> tmp <- qtscore(bint, gwadat1)
> descriptives.scan(tmp)
```

	Chromosome	Position	N	effB	P1df	Pc1df
rs10111989	8	128971174	171	0.3774178	1.00000e-10	3.00000e-10
rs4733579	8	128967351	171	0.2830056	5.91500e-07	1.09210e-06
rs11195943	10	114144805	171	-0.3599644	4.13730e-06	6.98910e-06
rs598296	9	14333670	170	0.3012061	5.01010e-06	8.39010e-06
rs3745205	19	45640042	171	-0.3171021	5.74570e-06	9.56220e-06
rs1356774	12	88034998	169	0.2421640	7.46040e-06	1.22694e-05
rs288740	13	106299212	171	0.2647059	8.61590e-06	1.40773e-05
rs2297646	10	28265903	171	-0.2526763	9.02790e-06	1.47191e-05
rs2000600	18	932860	171	0.3579114	9.20120e-06	1.49887e-05
rs599367	1	20306989	171	-0.2440512	1.02672e-05	1.66422e-05

	effAB	effBB	P2df
rs10111989	0.3589474	0.7800000	8.00000e-10
rs4733579	0.1702564	0.6880342	4.88900e-07
rs11195943	-0.3817603	-0.5121951	1.99841e-05
rs598296	0.3623932	0.4679487	1.61346e-05
rs3745205	-0.3452685	-0.5217391	2.77662e-05
rs1356774	0.3123663	0.4578947	2.45756e-05
rs288740	0.1646825	0.6694444	7.71400e-06
rs2297646	-0.2953930	-0.4555237	3.89324e-05
rs2000600	0.3587615	0.6991870	5.34973e-05
rs599367	-0.3449367	-0.3949367	1.28887e-05

Answer (ex. 24) — Possible explanation for loosing some significance is that

sex is an important risk factor in the trait under investigation and not adjusting for it we loose much power.

Answer (ex. 25) — Yes, there is a very strong significant relation between "bint" and "quat":

```
> attach(gwadat1@phdata)
> t.test(quat ~ bint)

        Welch Two Sample t-test

data:  quat by bint
t = -13.9588, df = 131.002, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.808709 -2.863174
sample estimates:
mean in group 0 mean in group 1
      0.2279858       3.5639270

> detach(gwadat1@phdata)
```

Answer (ex. 26) — Yes, rs10111989 is GW-significantly associated with "quat". This is exactly the same SNP which came out of the adjusted analysis of "bint".

Answer (ex. 27) — If adjusted for "quant", there is no significant relation between "bint" and SNP rs10111989:

```
> summary(glm(bint ~ sex + quat + gt, data = gwadat1@phdata, family = binomial))

Call:
glm(formula = bint ~ sex + quat + gt, family = binomial, data = gwadat1@phdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.05864  -0.24436  -0.04557   0.16961   2.50604

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.5140     0.8249  -5.472 4.44e-08 ***
sex          -1.4409     0.8413  -1.713   0.0868 .
quat          2.2042     0.4133   5.333 9.68e-08 ***
gt            1.1162     0.5824   1.917   0.0553 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 230.648  on 170  degrees of freedom
Residual deviance:  78.419  on 167  degrees of freedom
```

```
  (2 observations deleted due to missingness)
AIC: 86.419

Number of Fisher Scoring iterations: 7
```

One of the possible explanations is that ”quat” is an endophenotype for ”bint”, that is the region on chromosome 8 is controlling the levels of ”quat”, which is in turn a very strong risk factor for ”bint”