Genome-Wide Association analysis using `GenABEL`

Yurii Aulchenko, Najaf Amin

March 20, 2007

Abstract

In this exercise, you will become familiar with the `GenABEL` library, designed for GWA analysis. Compared to `dgc.genetics` package, it provides specific facilities for storage and manipulation of large amounts of data, very fast tests for GWA analysis, and special functions to analyse and graphically present the results of GWA analysis (thus "analysis of analysis").

`GenABEL` is rather new (first public release in mid-2006) and still developing (you can check http://mga.bionet.nsc.ru/nlru/GenABEL to see the history); the latest release was done in the beginning of March. This means there may still be (hopefully few) bugs or inconsistencies in the program. We will appreciate your suggestions on improving `GenABEL` .

In the first part of this exercise you will be guided through a GWA analysis of a small data set. In the second part you will investigate a larger data set by yourself, do a verification study and will answer the questions. All data sets used assume a study in a relatively homogeneous population. Try to finish the first part in the morning and the second part in the afternoon.

Though only few thousands of markers located at four chromosomes are used in the scan, we still going to call it Genome-Wide (GW), as the amount of data we will use is approaches the amount to be expected in a real experiment.

Contents

1 Example GWA session

Copy the file `ge03d2ex.RData` to your desktop and start `R` by double-clicking on it. Start `GenABEL` library by typing

```
> library(GenABEL)
```

You can read short overview of the package by asking for `help(GenABEL)`. Investigate the objects loaded by command

```
> ls()
```

```
[1] "ge03d2ex"
```

The `ge03d2ex` is a special data object of the class `gwaa.data`, as can be seeing from

```
> class(ge03d2ex)
```

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

As usual, if you are interested in details of this data type, you can get help by using command `help("gwaa.data-class")`; however, it is not strictly necessary to understand details of this data type to do GWA analysis in general and this exercise in particular. The only important thing to remember is that objects of this type contain a slot `phdata` which is a data frame with phenotypic information.

To check the variables in this data frame, you can use

```
> names(ge03d2ex@phdata)
```

```
[1] "id"     "sex"    "age"    "dm2"    "height" "weight" "diet"   "bmi"
```

Of cause, all standard R procedures will work on this data frame, e.g. we can check the summary for the age variable by

```
> summary(ge03d2ex@phdata$age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23.84   38.33   48.71   49.07   58.57   81.57
```

We can also attach this data frame to the `R` search path by

```
> attach(ge03d2ex@phdata)
```

1.1 Data descriptives and first round of GWA analysis

Let us investigate what are the traits presented in the data frame loaded and what are the characteristics of the distribution by using specific `GenABEL` function

```
> descriptives.trait(ge03d2ex)
```

```
         No    Mean     SD
id      136      NA     NA
sex     136   0.529  0.501
age     136  49.069 12.926
dm2     136   0.632  0.484
height  135 169.440  9.814
weight  135  87.397 25.510
diet    136   0.059  0.236
bmi     135  30.301  8.082
```

You can see that this frame contains the data on 136 people; the data on sex, age, height, weight, diet and body mass index (BMI) are available. Our trait of interest is `dm2` (type 2 diabetes). Note that every single piece of information in this data set is simulated; however, we tried to keep our simulations in a way we think the control of T2D works.

You can produce a summary for cases and controls separately and compare distributions of the traits by

```
> descriptives.trait(ge03d2ex, by = dm2)
```

```
       No(by=1)    Mean     SD No(by=0)    Mean     SD   Ptt   Pkw Pexact
id           86      NA     NA       50      NA     NA    NA    NA      NA
sex          86   0.593  0.494       50   0.420  0.499 0.053 0.052  0.074
age          86  50.250 12.206       50  47.038 13.971 0.179 0.205     NA
dm2          86      NA     NA       50      NA     NA    NA    NA      NA
height       86 170.448 10.362       49 167.671  8.586 0.097 0.141     NA
weight       86  93.587 27.337       49  76.534 17.441 0.000 0.000     NA
diet         86   0.058  0.235       50   0.060  0.240 0.965 0.965  1.000
bmi          86  32.008  8.441       49  27.304  6.463 0.000 0.001     NA
```

here, the `by` argument specifies the grouping variable. You can see that cases and controls are different in weight, which is expected, as T2D is associated with obesity.

Similarly, you can produce grand GW descriptives of the marker data by using

```
> descriptives.marker(ge03d2ex)
```

```
$`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   146.000      684.000     711.000    904.000 1555.000
Prop   0.036        0.171       0.178      0.226    0.389

$`Distribution of number of SNPs out of HWE, at different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 X>0.05
No     46.000   71.000 125.000 275.000   4000
Prop    0.011    0.018   0.031   0.069      1

$`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    1.000           0            0      135.000      0
```

```
Prop  0.007           0            0        0.993      0

$`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   37.000       6.000      996.000     1177.000 1784.000
Prop  0.009       0.002        0.249        0.294    0.446

$`Mean heterozygosity for a SNP`
[1] 0.2582298

$`Standard deviation of the mean heterozygosity for a SNP`
[1] 0.1592255

$`Mean heterozygosity for a person`
[1] 0.2476507

$`Standard deviation of mean heterozygosity for a person`
[1] 0.04291038
```

It is of note that we can see inflation of the proportion of the tests for HWE at particular threshold, as compared to expected. This may indicate poor genotyping quality and/or genetic stratification.

We can test the GW marker characteristics in controls by

```
> descriptives.marker(ge03d2ex, ids = (dm2 == 0))

$`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   233.000      676.000     671.000    898.000 1522.000
Prop   0.058        0.169       0.168      0.225    0.381

$`Distribution of number of SNPs out of HWE, at different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 X>0.05
No          0    3.000  14.000  98.000   4000
Prop        0    0.001   0.003   0.025      1

$`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           0            0           50      0
Prop      0           0            0            1      0

$`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   37.000      49.000     1523.000            0 2391.000
Prop  0.009       0.012        0.381            0    0.598

$`Mean heterozygosity for a SNP`
[1] 0.2555009

$`Standard deviation of the mean heterozygosity for a SNP`
[1] 0.1618707
```

```
$`Mean heterozygosity for a person`
[1] 0.2525720

$`Standard deviation of mean heterozygosity for a person`
[1] 0.04714886
```

Apparently, HWE distribution holds better in controls than in the total sample.

Let us check whether there are indications that deviation from HWE is due to cases. At this stage we are only interested in HWE distribution table, and therefore will ask to report only table two:

```
> descriptives.marker(ge03d2ex, ids = (dm2 == 1))[2]

$`Distribution of number of SNPs out of HWE, at different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 X>0.05
No     45.000    79.00 136.000 268.000   4000
Prop    0.011     0.02   0.034   0.067      1
```

It seems that indeed excessive number of markers are out of HWE in cases. If no laboratory procedure (e.g. DNA extraction, genotyping, calling) were done for cases and controls separately, this may indicate possible heterogeneity in cases. However, it is also possible that we detect more deviation from HWE in cases simply because the number of cases is larger than controls, yielding higher detection power.

It may be interesting to plot a $\chi^2 - \chi^2$ plot contrasting observed and expected distributions for the test for HWE in cases. First, we need to compute exact test for HWE by

```
> s <- summary(ge03d2ex@gtdata[(dm2 == 1), ])
```

Note the you have produced the summary for the `gtdata` slot of `ge03d2ex`; this is the slot which actually contain all genetic data in special compressed format.

You can see first 10 elements of this very long table by

```
> s[1:10, ]
```

	NoMeasured	CallRate	Q.2	P.11	P.12	P.22	Pexact	Chromosome
rs7435137	84	0.9767442	0.52380952	17	46	21	0.510978370	1
rs7725697	85	0.9883721	0.01176471	83	2	0	1.000000000	3
rs664063	86	1.0000000	0.08720930	71	15	0	1.000000000	2
rs4670072	60	0.6976744	0.11666667	53	0	7	0.001701645	X
rs546570	84	0.9767442	0.89880952	1	15	68	1.000000000	2
rs7908680	83	0.9651163	0.03012048	78	5	0	1.000000000	1
rs166732	83	0.9651163	0.04216867	76	7	0	1.000000000	1
rs4257079	86	1.0000000	0.07558140	73	13	0	1.000000000	1
rs5150804	84	0.9767442	0.39880952	31	39	14	0.820496827	2
rs3508821	83	0.9651163	0.20481928	52	28	3	1.000000000	2

Note that the column before the last provides P-exact we need. We can extract these to a separate vector by

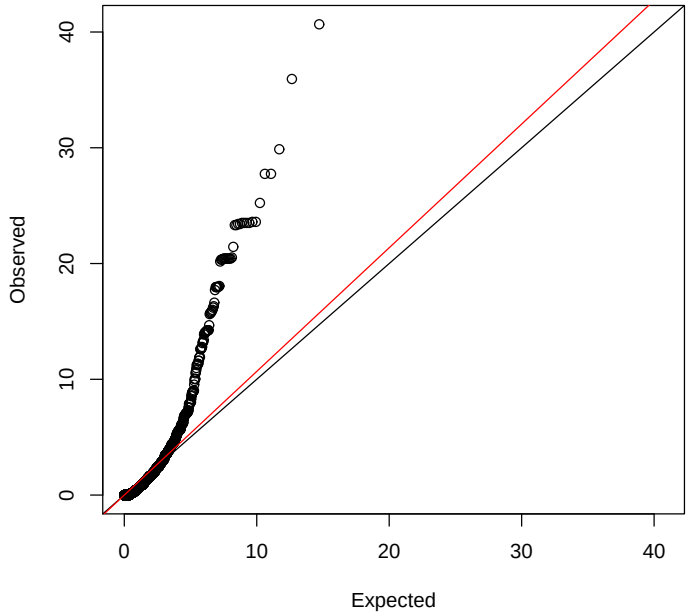

Figure 1: $\chi^2 - \chi^2$ plot for the exact test for HWE. Black line of slope 1: expected under no inflation; Red line: fitted slope.

```
> pexcas <- s[, "Pexact"]
```

and produce the $\chi^2 - \chi^2$ plot and estimate inflation factor by command `estlambda()`, which operates with a vector of P-values or χ^2s:

```
> estlambda(pexcas)

$estimate
[1] 1.068184

$se
[1] 0.02614764
```

By default, this function also produces a $\chi^2 - \chi^2$ plot, at which you can see some extreme deviation of observed from expected. The resulting plot (figure 1) shows extreme deviation for high values of the test. Looking at the λ estimate, we indeed see inflation of the test statistics.

You can repeat this test for the controls, if time permits.

Let us first try do GWA scan using raw (before quality control) data. We will use the score test, as implemented in the `qtscore()` funcrtion of `GenABEL` for testing:

```
> an0 <- qtscore("dm2~CRSNP", ge03d2ex)
```

The first argument used describes the model; here it is rather simple — the affection status, `dm2`, is supposed to depend on SNP genotype only (`CRSNP` term, which stands for CuRrent SNP, where current SNP is each of the SNPs typed in the study in turn).

You can see what objects are returned by this function by using

```
> names(an0)
```

```
 [1] "P1df"       "P2df"       "Pc1df"      "lambda"     "effB"
 [6] "effAB"      "effBB"      "snpnames"   "map"        "chromosome"
[11] "idnames"    "formula"    "family"
```

Here, `P1df`, `P2df` and `Pc1df` are most interesting; the first two are vectors of 1 and 2 d.f. P-values obtained in the GWA analysis, the last one is 1 d.f. P-value corrected for inflation factor λ (which is presented in `lambda` object).

Let us see if there is evidence for the inflation of the test statistics

```
> an0$lambda
```

```
$estimate
[1] 1.041173

$se
[1] 0.0007477997
```

The estimate of λ is 1.04, suggesting inflation of the test.

We can plot the results of analysis by

```
> plot(an0)
```

The resulting plot is presented in the figure 2. By default, $-log_{10}(P-value)$ on 1 d.f. are presented; see help to figure out how this behaviour can be changed.

You can also generate a descriptive table for the "top" (as ranked by P-value) results by

```
> descriptives.scan(an0)
```

```
          Chromosome Position      effB     P1df    Pc1df     effAB     effBB
rs1719133          1  4495479 -0.189730 0.000280 0.000369 -0.102941 -0.632353
rs2975760          3 10518480  0.182573 0.000298 0.000393  0.141182  0.274763
rs7418878          1  2808520  0.170464 0.000974 0.001229  0.154881  0.200980
rs5308595          3 10543128  0.223766 0.001054 0.001327  0.170057  0.375940
rs4804634          1  2807417 -0.079119 0.001197 0.001499  0.061353 -0.203788
rs3224311          2  6009769  0.142522 0.001329 0.001658  0.133082  0.170370
rs26325            3 10617781 -0.447811 0.001331 0.001661 -0.447811 -0.895623
rs8835506          2  6010852  0.142857 0.001532 0.001901  0.135566  0.163636
rs3925525          2  6008501  0.139601 0.001940 0.002387  0.128991  0.170370
rs2521089          3 10487652  0.108577 0.002052 0.002520  0.056511  0.170655
              P2df
rs1719133 0.000633
rs2975760 0.001143
rs7418878 0.002264
rs5308595 0.004593
```

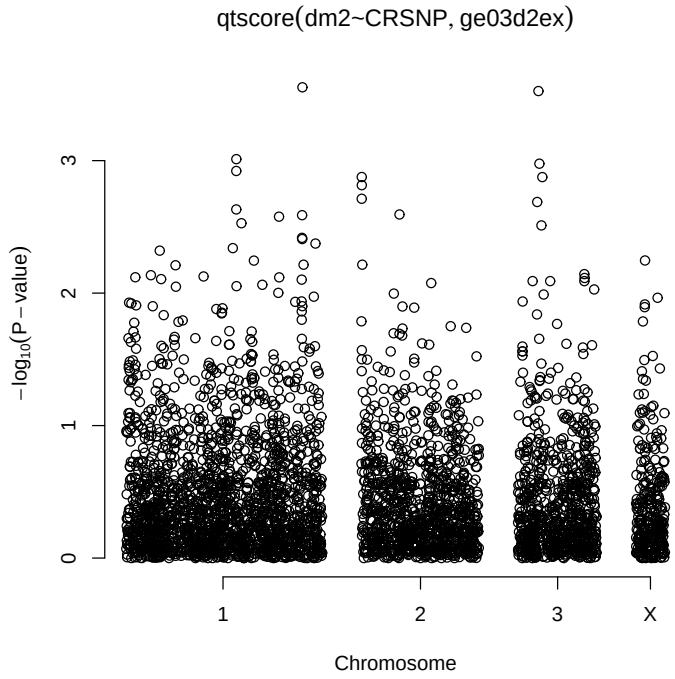

Figure 2: $-log_{10}(P - value)$ from the genome scan before QC procedure

```
rs4804634 0.003696
rs3224311 0.002941
rs26325   0.001331
rs8835506 0.003162
rs3925525 0.004555
rs2521089 0.006966
```

Here you see top 10 results, sorted by P-value with 1 d.f.. If you want to sort by the corrected P-value, you can use `descriptives.scan(an0,sort="Pc1df")`; to see more then 10 (e.g. 25) top results, use `descriptives.scan(an0,top=25)`. You can combine all these options.

Note: The `descriptives` family of functions was developed to facilitate production of tables which can be directly used in a manuscript — it is possible to save the output as a file, which can be open by Excel or Word. See e.g. `help(descriptives.trait)` for details.

Now let us apply `emp.qtscore()` function, which computes empirical GW (or experiment-wise) significance

```
> an0.e <- emp.qtscore("dm2~CRSNP", ge03d2ex)
```

```
100%

> descriptives.scan(an0.e, sort = "Pc1df")

          Chromosome Position      effB P1df Pc1df     effAB     effBB P2df
rs1719133          1  4495479 -0.189730 0.51  0.62 -0.102941 -0.632353 0.59
rs2975760          3 10518480  0.182573 0.54  0.62  0.141182  0.274763 0.86
rs7418878          1  2808520  0.170464 0.82  0.88  0.154881  0.200980 0.95
rs5308595          3 10543128  0.223766 0.83  0.92  0.170057  0.375940 1.00
rs4804634          1  2807417 -0.079119 0.87  0.94  0.061353 -0.203788 1.00
rs3224311          2  6009769  0.142522 0.92  0.96  0.133082  0.170370 0.97
rs26325            3 10617781 -0.447811 0.92  0.96 -0.447811 -0.895623 1.00
rs8835506          2  6010852  0.142857 0.95  0.97  0.135566  0.163636 0.99
rs3925525          2  6008501  0.139601 0.98  0.99  0.128991  0.170370 1.00
rs2521089          3 10487652  0.108577 0.98  0.99  0.056511  0.170655 1.00
```

None of the SNPs hits GW significance. If any did we could not trust the results, because the distribution of the HWE test and presence of inflation factor for the association test statistics suggest that the data may contain multiple errors (indeed they do). Therefore our first step should be rigorous Quality Control (QC).

1.2 Genetic data QC: simple checks

The major genetic data QC function of `GenABEL` is `check.marker()`. We will try to run it; the output is rather self-explaining. As it was detailed at the lecture, in the first round of the QC we do not want to check for HWE. This can be achieved by setting HWE P-value selection threshold to zero (`p.level=0`):

```
> qc1 <- check.marker(ge03d2ex, p.level = 0)

4000 markers and 136 people in total
Running sex (X-chromosome) checks...
Wrong male X genotypes (heterozygous) found for 198 genotypes
Error table is saved in Xerrtab
Marker rs4351348 is likely to be not an X marker (Odds> 100 )
Person id3374 is likely to be female (Odds> 100 )
Person id8410 is likely to be female (Odds> 100 )
If these people / snps are removed no errors are detected
RUN 1
4000 markers and 136 people in total
0 (0%) markers excluded as redundant (option = "no")
279 (6.975%) markers excluded as having low (<1.838235%) minor allele frequency
43 (1.075%) markers excluded because of low (<95%) call rate
0 (0%) markers excluded because they are out of HWE (P <0)
1 (0.7352941%) people excluded because of low (<95%) call rate
4 (2.941176%) people excluded because too high autosomal heterozygosity (FDR <1%)
Mean autosomal HET was 0.2742305 (s.e. 0.04145819), people excluded had HET >= 0.5019168
1 (0.7352941%) people excluded because of too high IBS (>=0.95)
Mean IBS was 0.7877711 (s.e. 0.02117194), as based on 2000 autosomal markers
In total, 3682 (92.05%) markers passed all criteria
```

```
In total, 130 (95.58824%) people passed all criteria
RUN 2
3681 markers and 129 people in total
0 (0%) markers excluded as redundant (option = "no")
101 (2.743820%) markers excluded as having low (<1.937984%) minor allele frequency
0 (0%) markers excluded because of low (<95%) call rate
0 (0%) markers excluded because they are out of HWE (P <0)
0 (0%) people excluded because of low (<95%) call rate
0 (0%) people excluded because too high autosomal heterozygosity (FDR <1%)
Mean autosomal HET was 0.2744468 (s.e. 0.01706392)
0 (0%) people excluded because of too high IBS (>=0.95)
Mean IBS was 0.77852 (s.e. 0.01691393), as based on 2000 autosomal markers
In total, 3580 (97.25618%) markers passed all criteria
In total, 129 (100%) people passed all criteria
RUN 3
3580 markers and 129 people in total
0 (0%) markers excluded as redundant (option = "no")
0 (0%) markers excluded as having low (<1.937984%) minor allele frequency
0 (0%) markers excluded because of low (<95%) call rate
0 (0%) markers excluded because they are out of HWE (P <0)
0 (0%) people excluded because of low (<95%) call rate
0 (0%) people excluded because too high autosomal heterozygosity (FDR <1%)
Mean autosomal HET was 0.2744468 (s.e. 0.01706392)
0 (0%) people excluded because of too high IBS (>=0.95)
Mean IBS was 0.772148 (s.e. 0.01841456), as based on 2000 autosomal markers
In total, 3580 (100%) markers passed all criteria
In total, 129 (100%) people passed all criteria
```

Note: The computation of all pairwise proportion of alleles identical-by-state (IBS) by `ibd()` function, which is also called by `check.markers()` may take quite some time, which is proportional to the square of the number of subjects. This is not a problem with the small number of people we use for this example or when modern computers are used. However, the computers in the Nihes computer room are very old. Therefore be prepared to wait for long time when you will do a self-exercise with 1,000 people.

From the output you can see that QC starts with checking the data for X-chromosome; it finds out that all errors are due to two people with wrong sex assigned and one marker, which looks like an autosomal one. This actually could be a marker from pseudoautosomal region, which should have been arranged as a separate "autosome".

Then, the procedure finds the markers with low call rate (≤ 0.95) across people, markers with low MAF (by default, low MAF is defined as 10 or less copies of the rare allele); people with low call rate (≤ 0.95) across SNPs, people with extreme heterozygosity (at FDR 0.01) and these who have GW IBS ≥ 0.95. These default parameters may be changed if you wish (consult help).

Because some of the people fail to pass the tests, the data set is not guaranteed to be really "clean" after single iteration, e.g. some marker may not pass the

call threshold after we exclude few informative (but apparently wrong) people. Therefore the QC is repeated iteratively until no further errors are found.

You can generate short summary of QC by marker and by person through

```
> summary(qc1)
```

```
$`Per-SNP fails statistics`
          NoCall NoMAF NoHWE Redundant Xsnpfail
NoCall        39     4     0         0        0
NoMAF         NA   376     0         0        0
NoHWE         NA    NA     0         0        0
Redundant     NA    NA    NA         0        0
Xsnpfail      NA    NA    NA        NA        1

$`Per-person fails statistics`
          IDnoCall HetFail IBSFail Xidfail
IDnoCall         1       0       0       0
HetFail         NA       3       0       1
IBSFail         NA      NA       1       0
Xidfail         NA      NA      NA       1
```

As you can see from the output, some markers and people fail to pass multiple criteria.

Note that the original data, `ge03d2ex`, are not modified during the procedure; rather, `check.markers()` generate a list of markers and people which pass or do not pass certain QC criteria. The objects returned by `check.markers()` are:

```
> names(qc1)
```

```
 [1] "nofreq"    "nocall"    "nohwe"     "snpok"     "idnocall"  "hetfail"
 [7] "ibsfail"   "idok"      "Pex.nohwe" "call"      "Xmrkfail"  "Xidfail"
[13] "Xerrtab"
```

The element `idok` provides the list of people who passed all QC criteria, and `snpok` provides the list of SNPs which passed all criteria. You can easily generate a new data set, which will consist only of these people and markers by

```
> data1 <- ge03d2ex[qc1$idok, qc1$snpok]
```

If there are any residual sporadic X-errors (male heterozygosity), these can be fixed by

```
> data1 <- Xfix(data1)
```

```
no X-errors to fix
```

Applying this function does not make any difference for the example data set, but you will need to use it for the bigger data set.

At this point, we are ready to work with the new, cleaned, data set `data1`. However, if we try

```
> table(dm2)
```

```
dm2
 0  1
50 86
```

we can see that the original phenotypic data are attached to the search path. Therefore we need to detach the data by

```
> detach(ge03d2ex@phdata)
```

At this stage, let us check if the first round of QC solves the problem of inflated test for HWE, which may be the case if this inflation is due to genotypic errors we managed to eliminate:

```
> descriptives.marker(data1)[2]
```

```
$`Distribution of number of SNPs out of HWE, at different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 X>0.05
No     44.000   65.000 121.000 240.000   3580
Prop    0.012    0.018   0.034   0.067      1
```

```
> descriptives.marker(data1[data1@phdata$dm2 == 1])[2]
```

```
$`Distribution of number of SNPs out of HWE, at different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 X>0.05
No     46.000    72.00 125.000 235.000   3580
Prop    0.013     0.02   0.035   0.066      1
```

```
> estlambda(summary(data1@gtdata[data1@phdata$dm2 == 1, ])[, "Pexact"])
```

```
$estimate
[1] 1.103741

$se
[1] 0.02787467
```

Apparently, the distribution (figure 3) looks better (note the scale difference between the graphs), but the test statistics is still quite inflated.

1.3 Finding genetic sub-structure

Now, we are ready for the second round of QC, detection of genetic outliers which may contaminate our data. We will detect genetic outliers using a technique, which resembles the one suggested by Price at al. (Nat Genet, 2006; to be discussed in the afternoon lecture session on March 20 and also very likely by David Evans). The difference is that we will not normalise genotypes in our analysis.

As a first step, we will compute a matrix of IBS between all pairs of people, using only autosomal markers by

```
> data1.ibs <- ibs(data1[, data1@gtdata@chromosome != "X"])
```

You can see the 5x5 upper left sub-matrix by

```
> data1.ibs[1:5, 1:5]
```

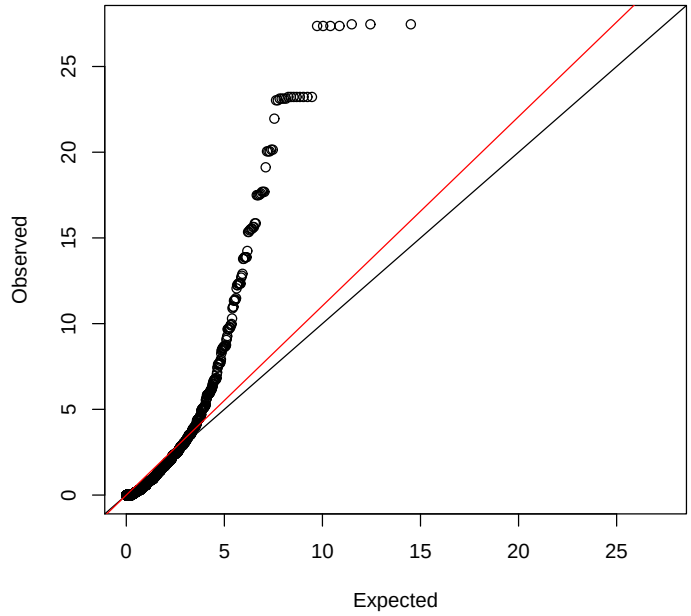

Figure 3: $\chi^2 - \chi^2$ plot for the exact test for HWE. Black line of slope 1: expected under no inflation; Red line: fitted slope.

```
          id199        id300        id403        id415 id666
id199        NA 3261.0000000 3260.0000000 3248.0000000  3264
id300 0.7733824           NA 3267.0000000 3256.0000000  3270
id403 0.7674847    0.7823691           NA 3254.0000000  3269
id415 0.7747845    0.7917690    0.7725876           NA  3258
id666 0.7552083    0.7614679    0.7766901    0.7579804    NA
```

Note: This may take few minutes on large data sets or when using old computers!

The numbers below the diagonal show IBS, the numbers above the diagonal tell how many SNPs were typed successfully for both (thus the IBS estimate is derived using this number of SNPs).

Second, we transform this matrix to a distance matrix using standard `R` command

```
> data1.dist <- as.dist(1 - data1.ibs)
```

Finally, we perform Classical Multidimensional Scaling by

```
> data1.mds <- cmdscale(data1.dist)
```

by default, the first two principal components are computed and returned.

Note: This may take few minutes on large data sets or when using old computers!

We can present the results graphically by

```
> plot(data1.mds)
```

The resulting plot is presented in figure 4. Each point on the plot corresponds to a person, and the 2D distances between points were fitted to be as close as possible to these presented in the original IBS matrix. You can see that study subjects clearly cluster in two groups.

You can identify the points belonging to clusters by

```
> km <- kmeans(data1.mds, centers = 2, nstart = 1000)
> cl1 <- names(which(km$cluster == 1))
> cl2 <- names(which(km$cluster == 2))
> cl1
```

```
  [1] "id199"  "id300"  "id403"  "id415"  "id666"  "id689"  "id765"  "id830"
  [9] "id908"  "id980"  "id994"  "id1193" "id1423" "id1505" "id1737" "id1827"
 [17] "id1841" "id2068" "id2094" "id2115" "id2151" "id2317" "id2618" "id2842"
 [25] "id2894" "id2985" "id3354" "id3368" "id3641" "id3831" "id3983" "id4097"
 [33] "id4328" "id4380" "id4395" "id4512" "id4552" "id4710" "id4717" "id4883"
 [41] "id4904" "id4934" "id4961" "id5014" "id5078" "id5274" "id5275" "id5454"
 [49] "id5853" "id5926" "id5969" "id6237" "id6278" "id6352" "id6501" "id6554"
 [57] "id6663" "id6723" "id7499" "id7514" "id7541" "id7598" "id7623" "id7949"
 [65] "id8059" "id8128" "id8281" "id8370" "id8400" "id8433" "id8772" "id8880"
 [73] "id8890" "id8957" "id8996" "id9082" "id9901" "id9930" "id1857" "id2528"
 [81] "id4862" "id9184" "id5677" "id6407" "id5472" "id2135" "id8545" "id4333"
 [89] "id1670" "id1536" "id6917" "id6424" "id3917" "id9628" "id9635" "id4729"
 [97] "id5190" "id6399" "id6062" "id620"  "id1116" "id6486" "id41"   "id677"
[105] "id4947" "id9749" "id6428" "id7488" "id5949" "id2924" "id5783" "id4096"
[113] "id903"  "id9049" "id185"  "id1002" "id362"  "id9014" "id5044" "id2749"
[121] "id2286" "id4743" "id4185" "id8330" "id6934"
```

```
> cl2
```

```
[1] "id2097" "id6954" "id2136" "id858"
```

Four outliers are presented in the smaller cluster.

Note: Now you will need to use the BIGGER cluster for to select study subjects. Whether this will be `cl1` or `cl2` in you case, is totally random.

We can form a data set which is free from outliers by using only people from the bigger cluster:

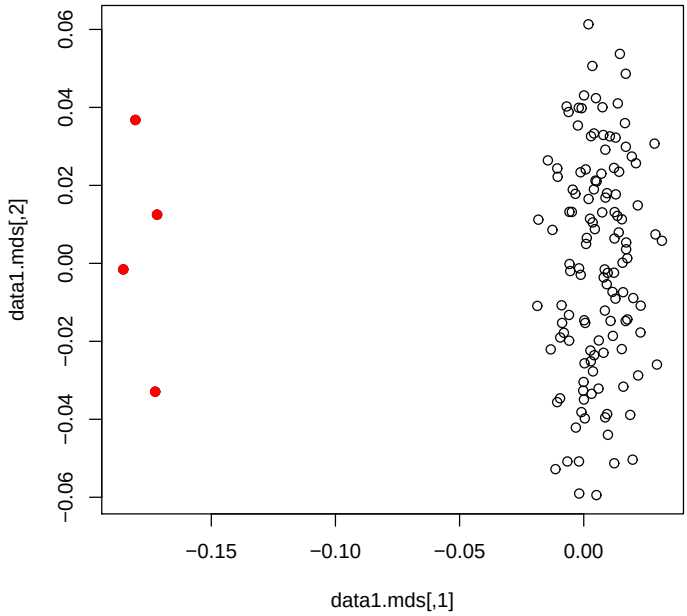

Figure 4: Mapping samples on the space of the first two Principle Components resulting from analysis of genomic IBS. Red dots identify genetic outliers

```
> data2 <- data1[cl2, ]
```

After we dropped the outliers, we need to repeat QC using `check.markers()`. At this stage, we want to allow for HWE checks (we will use only controls and exclude markers with FDR ≤ 0.2):

```
> qc2 <- check.marker(data2, hweids = (data2@phdata$dm2 == 0),
+     fdr = 0.2)

3580 markers and 125 people in total
Running sex (X-chromosome) checks...
No sex errors found
RUN 1
3580 markers and 125 people in total
0 (0%) markers excluded as redundant (option = "no")
40 (1.117318%) markers excluded as having low (<2%) minor allele frequency
0 (0%) markers excluded because of low (<95%) call rate
0 (0%) markers excluded because they are out of HWE (FDR <20%)
0 (0%) people excluded because of low (<95%) call rate
0 (0%) people excluded because too high autosomal heterozygosity (FDR <1%)
Mean autosomal HET was 0.2776398 (s.e. 0.01655241)
0 (0%) people excluded because of too high IBS (>=0.95)
Mean IBS was 0.7709423 (s.e. 0.01274011), as based on 2000 autosomal markers
```

```
In total, 3540 (98.88268%) markers passed all criteria
In total, 125 (100%) people passed all criteria
RUN 2
3540 markers and 125 people in total
0 (0%) markers excluded as redundant (option = "no")
0 (0%) markers excluded as having low (<2%) minor allele frequency
0 (0%) markers excluded because of low (<95%) call rate
0 (0%) markers excluded because they are out of HWE (FDR <20%)
0 (0%) people excluded because of low (<95%) call rate
0 (0%) people excluded because too high autosomal heterozygosity (FDR <1%)
Mean autosomal HET was 0.2776398 (s.e. 0.01655241)
0 (0%) people excluded because of too high IBS (>=0.95)
Mean IBS was 0.7721072 (s.e. 0.01230825), as based on 2000 autosomal markers
In total, 3540 (100%) markers passed all criteria
In total, 125 (100%) people passed all criteria
```

```
> summary(qc2)
```

```
$`Per-SNP fails statistics`
          NoCall NoMAF NoHWE Redundant Xsnpfail
NoCall         0     0     0         0        0
NoMAF         NA    40     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 Xidfail
IDnoCall        0       0       0       0
HetFail        NA       0       0       0
IBSFail        NA      NA       0       0
Xidfail        NA      NA      NA       0
```

Note: If the procedure did not run, check previous Note.

Indeed, in the updated data set few markers do not pass our QC criteria and we need to drop a few markers. This is done by

```
> data2 <- data2[qc2$idok, qc2$snpok]
```

This is going to be our final analysis data set, therefore let us attach the phenotypic data to the search path, then we do not need to type `data2@phdata$...` to access `dm2` status or other variables:

```
> attach(data2@phdata)
```

1.4 GWA association analysis

Let us start again with descriptives of the phenotypic and marker data

```
> descriptives.trait(data2, by = dm2)
```

```
       No(by=1)    Mean     SD No(by=0)    Mean     SD   Ptt   Pkw Pexact
id           78      NA     NA       47      NA     NA    NA    NA     NA
sex          78   0.603  0.493       47   0.426  0.500 0.057 0.056  0.065
age          78  50.508 12.406       47  45.752 13.313 0.050 0.066     NA
dm2          78      NA     NA       47      NA     NA    NA    NA     NA
height       78 170.454 10.580       46 167.911  8.689 0.150 0.203     NA
weight       78  94.047 26.806       46  77.015 17.528 0.000 0.000     NA
diet         78   0.064  0.247       47   0.064  0.247 0.995 0.995  1.000
bmi          78  32.188  8.291       46  27.424  6.598 0.001 0.001     NA
```

You can see that relation to weight is maintained in this smaller, but hopefully cleaner, data set; moreover, relation to age becomes boundary significant.

If you check descriptives of markers (only HWE part shown)

```
> descriptives.marker(data2)[2]
```

```
$`Distribution of number of SNPs out of HWE, at different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 X>0.05
No          1    2.000  22.000 108.000   3540
Prop        0    0.001   0.006   0.031      1
```

you can see that the problems with HWE are apparently fixed; we may guess that these were caused by the Wahlund's effect.

Run the score test on the cleaned data by

```
> data2.qt <- qtscore("dm2~CRSNP", data2)
```

and check lambda

```
> data2.qt$lambda
```

```
$estimate
[1] 1.029035

$se
[1] 0.0008736154
```

there is still some inflation, but it is in an acceptable range.

Produce the plot by

```
> plot(data2.qt)
```

(figure 5).

Produce the scan summary by

```
> descriptives.scan(data2.qt, sort = "Pc1df")
```

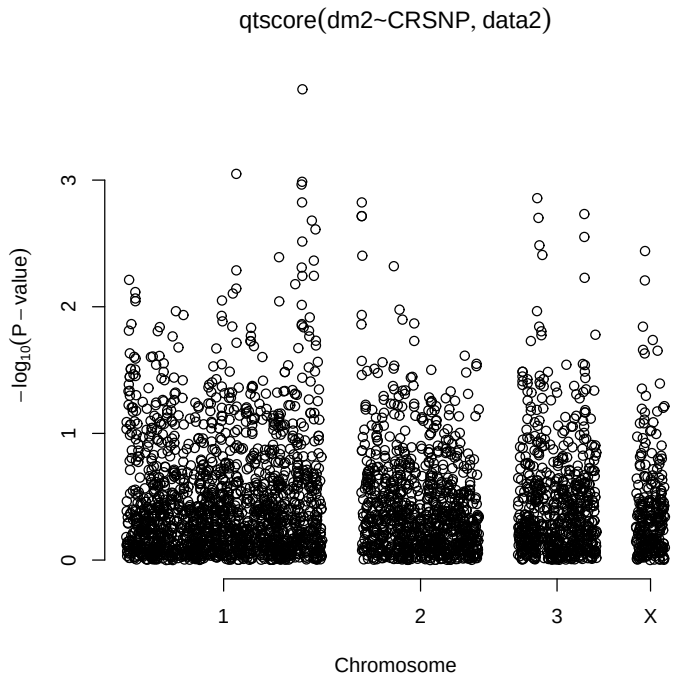

Figure 5: $-log_{10}(P-value)$ from the genome scan after the QC procedure

	Chromosome	Position	effB	P1df	Pc1df	effAB	effBB
rs1719133	1	4495479	-0.202947	0.000192	0.000237	-0.113362	-0.624000
rs4804634	1	2807417	-0.083944	0.000893	0.001057	0.084741	-0.228214
rs1013473	1	4487262	0.096930	0.001031	0.001216	0.026794	0.163879
rs4534929	1	4474374	-0.160678	0.001085	0.001277	-0.047637	-0.295699
rs2521089	3	10487652	0.120968	0.001389	0.001624	0.077864	0.180492
rs1048031	1	4485591	-0.107834	0.001500	0.001750	0.003016	-0.251016
rs8835506	2	6010852	0.146630	0.001501	0.001751	0.146630	0.146630
rs7522488	3	11689797	0.091153	0.001854	0.002151	-0.054301	0.212366
rs3925525	2	6008501	0.143123	0.001926	0.002233	0.139636	0.153778
rs3224311	2	6009769	0.143123	0.001926	0.002233	0.139636	0.153778

	P2df
rs1719133	0.000542
rs4804634	0.001785
rs1013473	0.003396
rs4534929	0.004386
rs2521089	0.004068
rs1048031	0.005530
rs8835506	0.002440
rs7522488	0.005067
rs3925525	0.003664
rs3224311	0.003664

Comparison with the top 10 from the scan before QC shows that results changed substantially with only few markers overlapping.

You can see similar results when accessing empirical GW significance:

```
> data2.qte <- emp.qtscore("dm2~CRSNP", data2)

100%

> descriptives.scan(data2.qte, sort = "Pc1df")

          Chromosome Position      effB P1df Pc1df     effAB     effBB P2df
rs1719133          1  4495479 -0.202947 0.25  0.30 -0.113362 -0.624000 0.50
rs4804634          1  2807417 -0.083944 0.81  0.89  0.084741 -0.228214 0.90
rs1013473          1  4487262  0.096930 0.87  0.92  0.026794  0.163879 1.00
rs4534929          1  4474374 -0.160678 0.90  0.94 -0.047637 -0.295699 1.00
rs1048031          1  4485591 -0.107834 0.97  0.97  0.003016 -0.251016 1.00
rs2521089          3 10487652  0.120968 0.96  0.97  0.077864  0.180492 1.00
rs8835506          2  6010852  0.146630 0.97  0.97  0.146630  0.146630 0.98
rs3925525          2  6008501  0.143123 0.98  0.99  0.139636  0.153778 1.00
rs8258863          1  4735725 -0.334915 0.99  0.99 -0.334915 -0.669829 1.00
rs2975760          3 10518480  0.169355 0.98  0.99  0.129032  0.267921 1.00
```

Again, none of the SNPs hits GW 5% significance. Still, you can see that after QC top markers achieve somewhat "better" significance.

In the last part, we will do several adjusted and stratified analyses. Only empirical P-values will be estimated to make the story shorter. To adjust for sex and age, we can

```
> data2.qtae <- emp.qtscore("dm2~sex+age+CRSNP", data2)

100%

> descriptives.scan(data2.qtae)

          Chromosome Position      effB P1df Pc1df     effAB     effBB P2df
rs1719133          1  4495479 -0.152503 0.31  0.31 -0.093299 -0.430760 0.64
rs4804634          1  2807417 -0.063060 0.85  0.85  0.064533 -0.172185 0.90
rs1013473          1  4487262  0.072049 0.92  0.92  0.018429  0.123231 0.98
rs7522488          3 11689797  0.071941 0.92  0.92 -0.041235  0.166254 0.98
rs4534929          1  4474374 -0.118323 0.94  0.94 -0.037428 -0.214947 1.00
rs1037237          3 11690145  0.069741 0.95  0.95 -0.043435  0.164054 0.98
rs1048031          1  4485591 -0.080579 0.95  0.95  0.004964 -0.191073 1.00
rs3925525          2  6008501  0.106498 0.96  0.96  0.107170  0.104445 0.98
rs2398949          1  4828375 -0.229306 0.96  0.96 -0.279048  0.268111 0.64
rs2521089          3 10487652  0.087149 0.96  0.96  0.064831  0.117971 1.00
```

You can see that there is little difference between adjusted and unadjusted analysis, but this is not always the case; adjustment may make your study much more powerful when covariates explain a large proportion of environmental trait variation.

Finally, let us do stratified (by BMI) analysis. We will contracts obese ($BMI \geq 30$) cases to all controls.

```
> data2.qtse <- emp.qtscore("dm2~sex+age+CRSNP", data2, ids = ((bmi >
+     30 & dm2 == 1) | dm2 == 0))

100%

> descriptives.scan(data2.qtse, sort = "Pc1df")

          Chromosome Position      effB P1df Pc1df     effAB     effBB P2df
rs1891586          1  2297398 -0.067937 0.87  0.87  0.067505 -0.158232    1
rs9630764          1  3897972  0.070521 0.96  0.96 -0.031769  0.127511    1
rs3215698          X 13559835 -0.284547 0.98  0.98 -0.240274 -0.317751    1
rs1037237          3 11690145  0.084565 0.98  0.98 -0.052101  0.212319    1
rs7522488          3 11689797  0.084565 0.98  0.98 -0.052101  0.212319    1
rs7435137          1  4259040  0.040260 1.00  1.00  0.012762  0.079080    1
rs664063           2  7288020 -0.028210 1.00  1.00  0.011410 -0.364971    1
rs546570           2  6120257 -0.005386 1.00  1.00 -0.005386 -0.010771    1
rs7908680          1  2311762  0.201118 1.00  1.00  0.201118  0.402235    1
rs166732           1  4716343 -0.018950 1.00  1.00 -0.018950 -0.037899    1
```

Again, noting interesting at GW significance level. If we would have had found something, naturally, we would not known if we mapped a T2D or obesity gene (or a gene for obesity in presence of T2D, or the one for T2D in presence of obesity).

At this point, you acquired the knowledge necessary for the self-exercise. Please close `R` by `q()` command and proceed to the next section.

2 GWA exercise

During the exercise, you will work with a larger data set (approximately 1,000 people and 7,000+ SNPs). You are to do complete three-round QC; perform GWA analysis with `dm2` as the outcome of interest and identify 10 SNPs which you would like to take to the confirmatory stage two scan. You will do confirmatory analysis using a confirmatory data set. If you did everything right, the SNPs which you identified as significant or replicated will be located in know T2D genes.

Please keep in mind that the data are simulated, and do not take your findings too seriously!

Start `R` by going to "Start -> Programs -> R -> R-2.4.1". Load `GenABEL` library by

```
> library(GenABEL)
```

The two data sets we will use in this exercise are part of the `GenABEL` distribution. The first one ("discovery" set) can be loaded by

```
> data(ge03d2)
```

Please move along the lines detailed in the guided exercise and try to answer following questions:

Question 1 *How many cases and controls are presented in the original data set?*

Question 2 *How many markers are presented in the original data set?*

Question 3 *Is there evidence for inflation of the HWE test staistics?*

Question 4 *Perform GWA analysis of the raw data, using assimptotic test and plot the results. Try to think how you can produce $\chi^2 - \chi^2$ plot for the P-values on 1 d.f.. What is the estimate of λ for the 1 d.f. test?*

Question 5 *Analyse empirical GW significance. How many SNPs pass genome-wide significance threshold, after correction for the inflation factor? Write down the names of these SNPs for further comparison.*

Perform complete three steps of the genetic data QC.

Question 6 *How many male turned apparently female?*

Question 7 *How many sporadic X errors do you still observe even when the female male and non-X X-markers are removed? (do not forget to `Xfix()` these!)*

Question 8 *How many "twin" DNAs did you discover?*

Question 9 *How many genetic outliers did you discover?*

After you have finished QC, answer the questions:

Question 10 *How many cases and controls are presented in the data after QC?*

Question 11 *How many markers are presented in the data after QC?*

Question 12 *Is there evidence for inflation of the HWE test staistics?*

Question 13 *Perform GWA analysis of the cleaned data, using assimptotic test and plot the results. What is the estimate of λ for the 1 d.f. test?*

Question 14 *Analyse empirical GW significance. How many SNPs pass genome-wide significance threshold, after correction for the inflation factor? Do these SNPs overlap much with the ones ranked at the top before the QC? If not, what could be the reason?*

If time permits, do analysis with adjustment for covariates and stratified analysis.

Select 10 SNPs which you would like to follow-up. Say, you've selected rs1646456, rs7950586, rs4785242, rs4435802, rs2847446, rs946364, rs299251, rs2456488, rs1292700, and rs8183220.

Make a vector of these SNPs with

```
> vec12 <- c("rs1646456", "rs7950586", "rs4785242", "rs4435802",
+     "rs2847446", "rs946364", "rs299251", "rs2456488", "rs1292700",
+     "rs8183220")
```

Load the confirmatory data set by

```
> data(ge03d2c)
```

and select the subset of SNPs you need by

```
> confdat <- ge03d2c[, vec12]
```

Analyse the `confdat` for association with `dm2`.

Question 15 *Given the two-stage design, and applying the puristic criteria specified in the lecture, for how many SNPs you can claim a significant finding?*

Question 16 *Using the same criteria, for how many SNPs you can claim a replicated finding?*

You can check if any of the SNPs you have identified as significant or replicated are the ones which were simulated to be associated with `dm2` by using the command

```
> show.ncbi(c("snpname1", "snpname2", "snpname3"))
```

where `snpnameX` stands for the name of your identified SNP. The "true" SNPs can be found on NCBI and are located in known T2D genes (just because we used these names to name the "significant" ones).

If time permits, characterise the mode of inheritance of the significant SNPs. You can convert data from `GenABEL` format to the format used by `dgc.genetics` and `genetics` libraries by using `as.genotype()` function. Consult help for details. Please do not attempt to convert more then few dozens SNPs: the format of `genetics` is not compressed, which means conversion may take long and your low-memory computer may even crash if you attempt to convert the whole data set.

If time permits, try to do first round of QC allowing for HWE checks (assume FDR of 0.1 for total sample). In this case, can you still detect stratification in the "cleaned" data?