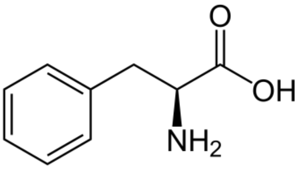
Hardy-Weinberg

Hans R Siegismund

Exercise 1

In Denmark, each year 6 children are born that are homozygotes for deleterious alleles at the phenylkenuria (PKU) locus (Følling[1] disease). These children are given a diet with a low concentration of phenylalanine to prevent them from developing a severe mental disease. In total, 60,000 children are born.



Phenylalanine

1) What is the frequency of the deleterious allele? (Assume that there are Hardy-Weinberg proportions.)

   We know that 6 out of 60.000 children are homozygous which means the genotype frequency for the homozygous carriers of the deleterious allele is 6/60,000=1/10,000. And if we assume Hardy-Weinberg proportions we know that this genotype frequency is equal to the allele frequency squared. Thus we have:

   $$q^2 = \frac{1}{10{,}000} \Rightarrow q = \sqrt{\frac{1}{10{,}000}} = 0.01$$

2) How many carriers (healthy heterozygotes) are born each year?

   If we assume Hardy-Weinberg proportions we know the frequency heterozygous is $2pq$. And from 1) we know that q=0.01 and thus that p=1-0.01=0.99. So we get $2pq = 2 \times 0.99 \times 0.01 = 0.0198$. Therefore, 60,000 × 0.0198 = 1,188 carriers are born each year. This illustrates nicely that the largest frequency (99%) of the rare allele is carried by heterozygote carriers in the population

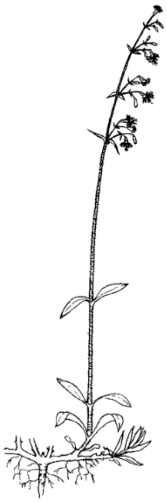
3) What is the fraction of affected children where both parents are healthy?

   The probability that both parents are healthy carriers is $(2pq)^2$. One quarter of their children will get the disease. Therefore, they contribute with $(2pq)^2/4$ to diseased children who in total constitute q^2 of all children Hence the relative contribution of the affected children with both parents healthy is $p^2q^2/q^2 = p^2 = 0.99^2 \approx 0.98$.

[1] Named after Asbjørn Følling, a Norwegian that discovered PKU in 1934.

Exercise 2

Silene nutans is a hermaphroditic plant that is self-compatible[2]. In a study of this plant, genetic variation at two life stages was recorded in a population: seedlings and adult plants. In the table below the genotype distributions at an enzyme locus are given



	11	12	22	Sum
Seedlings O	79	43	21	143
Seedlings E	70.63	59.74	12.63	
Adults O	70	60	13	143
Adults E	69.93	60.14	12.93	

1) Estimate the allele frequencies in the two groups.

p_1 in seedlings $= (2\times79+43)/(2\times143) \approx 0.703$

p_2 in seedlings $= 1- p_1 \approx 0.297$

p_1 in adults $= (2\times70+60)/(2\times143) \approx 0.699$

p_2 in adults $= 1-p_1 \approx 0.301$

2) Do the allele frequencies differ between the groups?

No. If you *really* want to do a test, you can do a 2 × 2 χ^2 homogeneity test on the allele counts in the two groups.

3) Do the genotype distributions differ from Hardy-Weinberg proportions?

To assess this we first calculate the expected proportions based on the results in 1)

For seedlings:

$E_{11} = p_1^2 \times 143 \approx 70.63$

$E_{12} = 2p_1p_2 \times 143 \approx 59.74$

$E_{22} = 143 - E_{11} - E_{12} \approx 12.63$

For adults:

$E_{11} = p_1^2 \times 143 \approx 70.93$

$E_{12} = 2p_1p_2 \times 143 = 60.14$

$E_{22} = 143 - E_{11} - E_{12} \approx 12.93$

[2] It can fertilize itself.

Then we use these numbers along with the observed counts to get the test statistics:

χ^2 for seedlings = $\Sigma(O_i - E_i)^2/E_i$

$= (79-70.63)^2/70.63+(43-59.74)^2/59.74+(21-12.63)^2/12.63$

≈ 11.23

χ^2 for adults = $\Sigma(O_i - E_i)^2/E_i$

$\approx (70-70.93)^2/70.93+(60-60.14)^2/60.14+(13-12.93)^2/12.93$

≈ 0.001

For the seedlings, the test statistic, 11.23, is much higher than 3.84 so we reject the null hypothesis of Hardy Weinberg proportions. For the adults, the test statistic is very small because there is an almost perfect match among observed and expected genotype distributions and we do not reject the null hypothesis of Hardy Weinberg proportions.

4) Estimate the inbreeding coefficient F for both groups. [$F = (H_E - H_O)/H_E$, where H_E and H_O are the expected and observed frequencies of heterozygotes.]

We see that F is high in seedlings:

$F_{\text{seedlings}} = (H_E - H_O)/ H_E = (59.74 - 43)/ 59.74 = 0.280$

And that this is due to a marked deficiency of heterozygotes ($H_E = 59.74$ and H_O=43) among the seedlings.

But this is not the case in the adults

$F_{\text{adults}} = (H_E - H_O)/ H_E = (60.14 - 60)/ 60.14 = 0.002$

since here the the observed number of heterozygotes is very close to the expected number.

5) What could have caused the differences in HW proportions between the groups?

Since *Silene nutans* is a hermaphroditic plant that is self-compatible, a considerable fraction of the seedlings seems to have been produced through selfing, which produces highly inbred offspring ($F = 0.5$) resulting in a high inbreeding coefficient in the seedling population.

6) What is happening among the seedling stage and the adult life stage?

Inbreeding exposes deleterious recessive genes in homozygotes, which results in inbreeding depression, i.e. reduced fitness of the offspring. We assume that the adults have started with a similar genotypic distribution when they were seedlings. In the case of *Silene nutans*, the inbreeding is extremely severe: all inbred individuals are eliminated from the population before they become adults.

Note on how to do this exercise in R on the server:
You can use Jan Graffelman's HardyWeinberg R package described in the paper "Exploring Diallelic Genetic Markers: The HardyWeinberg Package", which is available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=HardyWeinberg. A part of the paper has been copied further below in this exercise.

To use this open R and copy the following in the command line

```
library("HardyWeinberg")
Seedlings <- c(AA = 79, AB = 43, BB  = 21)
HW.test.Seedlings <- HWChisq(Seedlings, cc = 0, verbose = TRUE)
HW.test.Seedlings
Adults<-  c(AA = 70, AB = 60, BB  = 13)
HW.test.Adults <- HWChisq(Adults, cc = 0, verbose = TRUE)
HW.test.Adults
```

Exercise 3
A young unexperienced biologist has collected data for the human SNP rs16891982 for Europeans (EUR) and Africans (AFR). He found the following variation at this SNP:

	CC	CG	GG	Sum
AFR O	617	41	3	661
AFR E	614.84	45.33	0.84	
EUR O	4	54	445	503
EUR E	1.91	58.18	442.91	

1) Estimate the allele frequencies in the two groups.
p(C, AFR) = (617 × 2+41)/(2 ×661) ≈ 0.964
p(C, EUR) = (4 × 2+54)/(2 ×503) ≈ 0.062

2) Do the allele frequencies differ between the groups? (A statistical test should not be necessary.)

Yes

3) Do the genotype distributions differ from Hardy-Weinberg proportions?

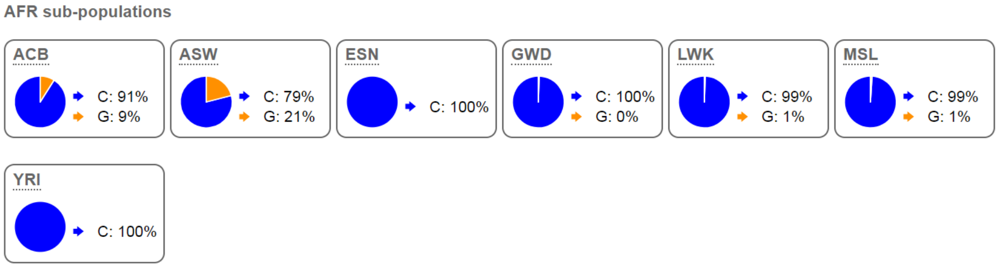
As in exercise 2 we calculate the expected number of each genotype category and based these and the observed numbers we calculate χ^2 :

AFR: $\chi^2 = 6.03$;
EUR: $\chi^2 = 2.60$;

The African population deviates significantly from Hardy-Weinberg proportions (χ^2>3.84) because there is an excess of homozygotes observed compared to the expected number. This is also the case for the European sample; but it is not significant (χ^2<3.84).

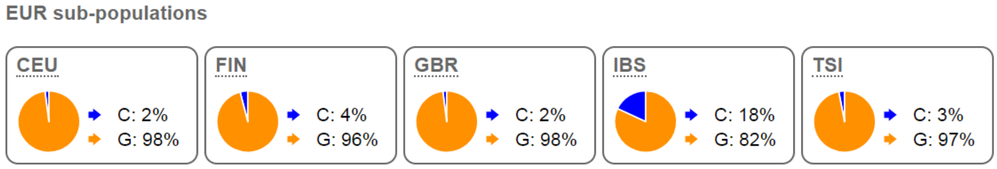
4) What could have caused a possible deviation from Hardy-Weinberg proportions?

The student has taken the data from the Ensembl database. Here, it turns out that the African population consists of the following subpopulations:



ACB	African Caribbean in Barbados
ASW	African Ancestry in Southwest US
ESN	Esan in Nigeria
GWD	Gambian in Western Division, The Gambia - Mandinka
LWK	Luhya in Webuye, Kenya
MSL	Mende in Sierra Leone
YRI	Yoruba in Ibadan, Nigeria

And the European population consists of the following subpopulations



CEU	Utah residents (CEPH) with Northern and Western European ancestry CEPH : Centre d'Etude du Polymorphism Humain
FIN	Finnish in Finland
GBR	British in England and Scotland

IBS	Iberian populations in Spain
TSI	Toscani in Italy

Hence both samples consist of samples from several subpopulations. Pooling samples from several populations results in an excess of homozygotes relative to the expected Hardy-Weinberg proportions. This effect has been called the Wahlund[3] effect. It is easy to illustrate the most extreme case for two populations that have been fixed for two different alleles. A sample from both populations consists only of the two different homozygotes and has an inbreeding coefficient of $F = 1$. Since $H_O = 0$, we have $F = (H_E - H_O)/H_E = H_E/H_E = 1$.

5) In which gene is the SNP found?

   It is in SLC45A2 (solute carrier family 45 member 2), which is involved in the production of melanin.

6) Does the variation have an effect on the phenotype?

   Yes, the mutation at position 1122 where Africans have a C and Europeans have the derived G results in the substitution of phenylalanine with leucine in the protein. The substitution causes light skin in Europeans.

7) Among the AFR subpopulations, two (ACB, ASW) seem to differ from the others, which seem to be homogeneous. This is also the case for one European subpopulation (IBS). What could the reason for this be?

   The ACB and ASW are probably admixed with Europeans. The IBS population is probably admixed with Africans.

[3] Named after Sten Wahlund, a Swedish geneticist that documented it in 1928.
Wahlund, S. (1928). Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. Hereditas 11: 65–106.

Exercise 4

In this exercise, we will analyze datasets that consist of a larger number of loci. For this purpose we will use the HardyWeinberg R package. There are two datasets: CEU_500.hw and CEU_500.sim. The first dataset consists of 500 diallelic SNPs from a sample of 60 humans of CEU[4] origin. The second dataset is a simulated dataset of 500 SNPs where one of the assumptions for having Hardy-Weinberg proportions is violated. The simulations have been carried out for 500 separate populations. There is no admixture among populations. The sample size is 60 for both datasets.

Get the two files:
In the Linux terminal: Change the directory to the exercise directory, make a HardyWeinberg directory and copy your data (*We assume that the directory ~/exercises exists. Otherwise, create it first.*)*:*

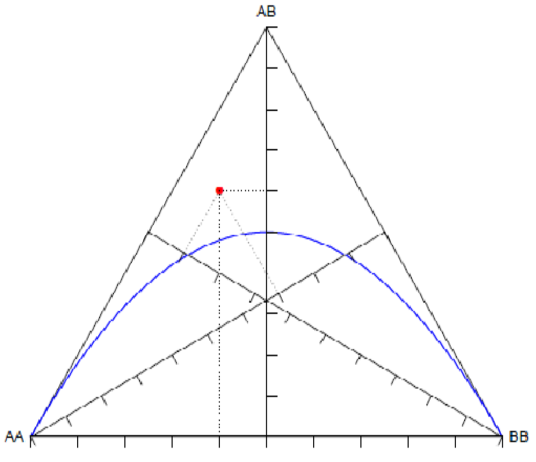
```
cd ~/exercises
mkdir HardyWeinberg
cd HardyWeinberg
cp ~/groupdirs/SCIENCE-BIO-Popgen_Course/exercises/HardyWeinberg/CE* .
```

Analyze (*in the ~/exercises/HardyWeinberg folder*) *the two data sets:*
Open R and copy the following in the command line

```
library("HardyWeinberg")
SNP_data <-as.matrix(read.table("CEU_500.hw", header=TRUE))
SNP_sim_data <-as.matrix(read.table("CEU_500.sim", header=TRUE))
#How does the data look like?
head(SNP_data)
#Plot the genotype frequencies for each of the 500 populations
#in a deFinetti diagram and indicate populations that differ
#signicantly from Hardy-Weinberg proportions.
par(mfrow=c(1,2))
HWTernaryPlot(SNP_data,   region = 1,
              curvecols=c("black", "red","green","black","purple"),
              vbounds = FALSE, main ="Original CEU data",
              cex = 0.5,  cex.main=1.5, font.main = 1)
HWTernaryPlot(SNP_sim_data, region = 1,
              curvecols=c("black", "red","green","black","purple"),
              vbounds = FALSE, main ="Simulated data",
              cex = 0.5, cex.main=1.5, font.main = 1)
```

Note

The plots are explained in “Graphical test for multiple diallelic markers”; see below. In short, each dot indicates the

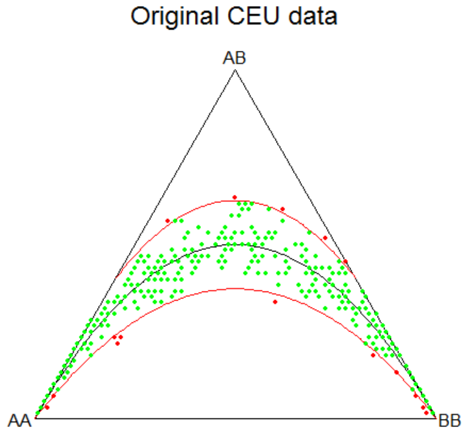


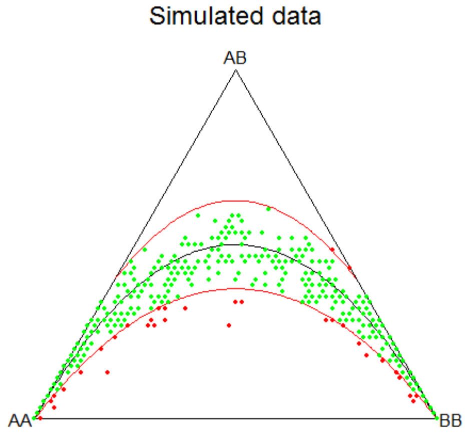
[4] Utah residents (CEPH) with Northern anc
CEPH : Centre d'Etude du Polymorphism

genotypic proportion at a locus. The distance of a dot perpendicular to the *X*-axis indicates the frequency of heterozygotes at the locus. Similarly, the distances of the dot perpendicular to the other two sides of the triangle indicate the frequency of the two homozygotes. In the example to the right, the genotype frequencies are

AA: 0.3
AB: 0.6
BB: 0.1

In the two plots produced in R the parabola in the middle indicates Hardy-Weinberg proportions. The other two curves indicate the acceptance regions for a chi-square test with one degree of freedom at a significance level of 0.05. Loci that deviate significantly from Hardy-Weinberg proportions are indicated in red, whereas loci that do not deviate significantly are indicated in green. Values above the upper curve indicate an excess of heterozygotes and those below the lower curve indicate a deficiency of heterozygotes compared to Hardy-Weinberg proportions.


Original CEU data


Simulated data

1) Are there any indications that the original SNP data has genotypic proportions that deviate significantly from Hardy-Weinberg proportions? (How many significant tests at a significance level of 0.05 do you expect to see?)

Copy the following in the command line in R to count the number of significant tests

```
#First, we make a vector with chi test values
#using the function HWChisqStats
chitest<-HWChisqStats(SNP_data,pvalues=FALSE)
#Then we make a vector with significant tests
sigchitest <-chitest[chitest>3.84]
length(sigchitest)
```

No. We can see some deviations from Hardy-Weinberg proportions; some have an excess of homozygotes and some have a deficiency relative to the expected HW proportions. We expect 500/20 = 25 significant deviations from Hardy-Weinberg proportions. There are 24.

2) Do the simulated dataset overall agree with Hardy-Weinberg proportions?

No. We can now see that there are few deviations with an excess of heterozygotes and substantially more deviations with a deficiency of heterozygotes. In total, there are now 60 significant tests. (We used a similar procedure on `SNP_sim_data.)`

3) What could the reason for a deviation from Hardy-Weinberg proportions be?

   As observed, there seems to be an excess of homozygotes compared to the expected Hardy-Weinberg proportions. There are (at least) two things that can cause this effect: population admixture (illustrated in exercise 3) and inbreeding (illustrated in exercise 2). We can rule out population admixture. All 500 populations were simulated independently. Therefore, we suggest inbreeding. Actually, the simulated dataset was generated with the following command in R using the function `HWData` in the package HardyWeinberg:

```
set.seed (16)
SNP_sim_data <- HWData(nm = 500, n = 60,
                       p = af(SNP_data),
                       f = rep(0.0625,500) )
#set.seed(16) sets the random number seed to 16. This
#guarantees that we get exact the same results in the
#simulations if we repeat them.
#nm number of markers
#n sample size
#p allele frequencies for the simulated population
#af() allele frequency function, estimated from the SNP_data
#f inbreeding coefficient, f = 0.0625 for all 500 markers
```

   The simulated markers are based on the allele frequencies from the observed dataset. It was assumed that the inbreeding coefficient was 0.0625 for each marker. The inbreeding coefficient is what we expect for offspring of parents that are cousins (like Darwin and his wife).

We can use a Q–Q (quantile-quantile) plot to check whether the data perform as expected. In this case, we check whether the genotypic proportions are the expected Hardy-Weinberg proportions. We use the HardyWeinberg package to do this. See **6.4. Q-Q plots** below in this exercise.

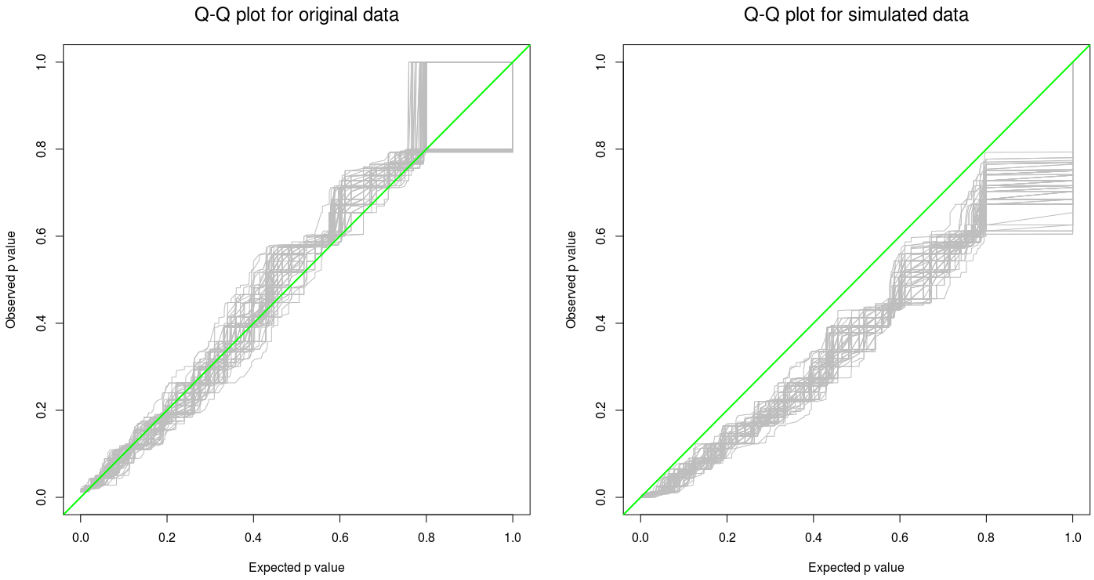
"The function HWQqplot of the package plots the p values against samples from the null distribution." It uses 100 simulations to compare observed and expected *p* values. If they perform as expected from the null model (Hardy-Weinberg), they will be distributed around the green line where the observed and expected values are equal.

In R copy the following in the command line

```
par(mfrow=c(1,2))
HWQqplot(SNP_data,
         main ="Q-Q plot for original data",
         cex.main=1.5, font.main = 1)
HWQqplot(SNP_sim_data,
         main ="Q-Q plot for simulated data",
         cex.main=1.5, font.main = 1)
```

Use these simulations to support your answers to the previous three questions.

The QQ plot for the observed data seems to come from a population in Hardy-Weinberg proportions. The QQ plot for the simulated data shows that the observed *p* values consistently are smaller than the expected *p* values for a population in Hardy-Weinberg proportions.



Graphical test for multiple diallelic markers

The following has been copied from Jan Graffelman's HardyWeinberg R package described in the paper "Exploring Diallelic Genetic Markers: The HardyWeinberg Package", which is available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=HardyWeinberg.

6. Graphics for Hardy-Weinberg equilibrium

6.2. The ternary plot

The Italian statistician Bruno De Finetti (1926) represented genotype frequencies in a ternary diagram. This diagram is known as a de Finetti diagram in the genetics literature (Canning sand Edwards 1968). The HWE condition defines a parabola in the ternary plot. A ternary plot of the genotype frequencies with the HWE parabola is an information-rich graphical display. From this plot one can recover genotype frequencies, allele frequencies, and infer the equilibrium status of a genetic marker at a glance (see Figure 2). The ternary plot is most useful for plotting data consisting of multiple samples that have all been genotyped for the same genetic marker. In that case the three vertices of the display are fully identified. An example is shown in Figure 3 where the genotype counts for the MN blood group locus are shown for 216 samples of various human populations from different geographical origin (Mourant et al. 1976, Table 2.5). The plot shows relatively higher allele frequencies for the N allele for samples from Oceania, and lower allele frequencies for this allele for the Eskimo samples. African, American, European and Asian populations have intermediate allele frequencies. Most samples clearly cluster around the HWE parabola though there are several deviating samples as well. The ternary plot may also be used to represent multiple markers, though this is a bit tricky because the obtained display is no longer uniquely determined. In this case, one vertex, usually the top vertex, is chosen to represent the heterozygote frequency of each marker. The two bottom vertices are used for one of the two homozygote frequencies. It is arbitrary to place aa on the right and bb on the left or the other way round. Representing multiple markers amounts to overplotting all ternary diagrams for each individual marker in such a way that the axes for the heterozygotes always coincide. Despite the indeterminacy of the homozygote vertices, the plot remains highly informative, as now minor allele frequency, genotype frequencies and equilibrium status are visualized simultaneously for many markers in just one plot. Graffelman and Morales-Camarena (2008) amplified the ternary plot by representing the acceptance regions of chi-square and exact tests inside the plot. An example with multiple markers is shown in the right panel of Figure 3. This figure shows 225 SNPs of the dataset HapMapCHBChr1. The function HWTernaryPlot of the package allows the construction of ternary plots with the equilibrium parabola and various acceptance regions.

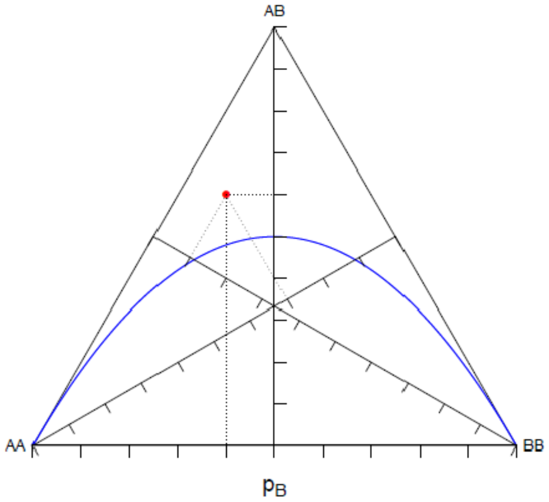


Figure 2: Ternary plot of a genetic marker, showing the recovery of genotype frequencies ($f_{AA} = 0.30, f_{AB} = 0.60$ and $f_{BB} = 0.10$) and allele frequencies ($p_B = 0.40$). The parabola represents Hardy-Weinberg equilibrium.

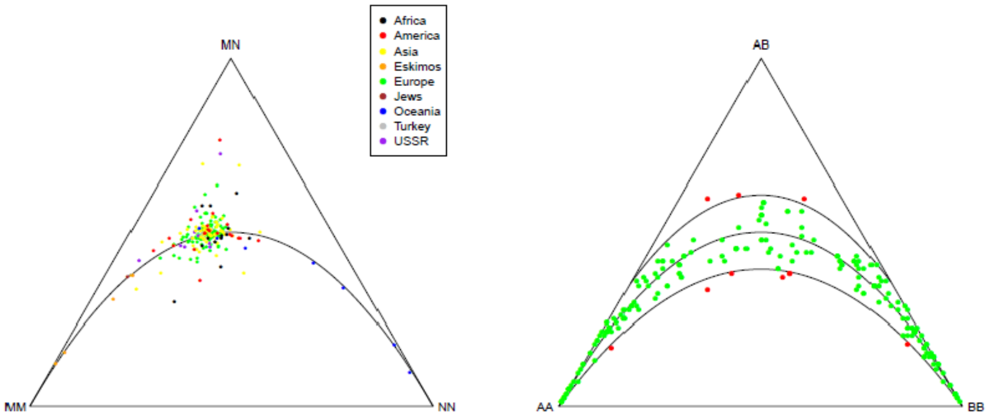


Figure 3: Left panel: ternary plot for one marker: MN blood group genotype frequencies for 216 samples from different human populations. Right panel: ternary plot for multiple markers: 225 SNPs on chromosome 1 of a sample of 84 individuals from the Han Chinese population. HWE parabola and acceptance region for a chi-square test are shown in the latter plot.

…

6.4. Q-Q plots

Genetic association studies nowadays investigate many markers for their possible relation with diseases. The equilibrium status of the markers is important, since deviation from HWE may be indicative of genotyping error. Moreover, disequilibrium for cases in a case-control study is indicative for disease association. Given that so many markers are tested, it is cumbersome to do this all in a numerical manner only, and it is known beforehand that false positives will arise. Even if we find that 5% of the markers is significant when we use a significance level of α = 0:05, this does not imply that the database as a whole can be considered to be "in equilibrium". The distribution of the test results (chi-square statistics or p values) then becomes interesting to look at. One way to do this is to compare the sample percentiles of the chi-square statistics of all markers with the theoretical percentiles of a χ_1^2 distribution in a chi-square quantile-quantile plot (Q-Q plot). For exact tests, Q-Q plots of the p values are used. Often the uniform distribution is chosen as the reference distribution. However, with discrete data the p value distribution under the null is not uniform. The function HWQqplot of the package plots the p values against samples from the null distribution rather than the uniform distribution. The function takes into account that sample size and allele frequency can vary over markers. Figure 5 shows Q-Q plots for the HapMap data (left panel) and also for simulated data under moderate inbreeding (right panel, f = 0.05). The green line is the reference line passing through the origin with slope 1. Each grey line plots a sample from the null distribution against the empirical quantiles of the p values. Deviation of the green line from the grey zone is taken as evidence that HWE is violated. The HapMap data set is seen to be in good agreement with what is expected under the null. This is not surprising, as the markers of the project undergo a quality control filter, and markers that strongly deviate from HWE (p value of an exact test < 0.001) are discarded from the project. For the dataset simulated under inbreeding, a manifest deviation from HWE is found. Q-Q plots assume independent observations. We note that this assumption will be violated if the markers under study are closely neighboring markers from the same region of a single chromosome.

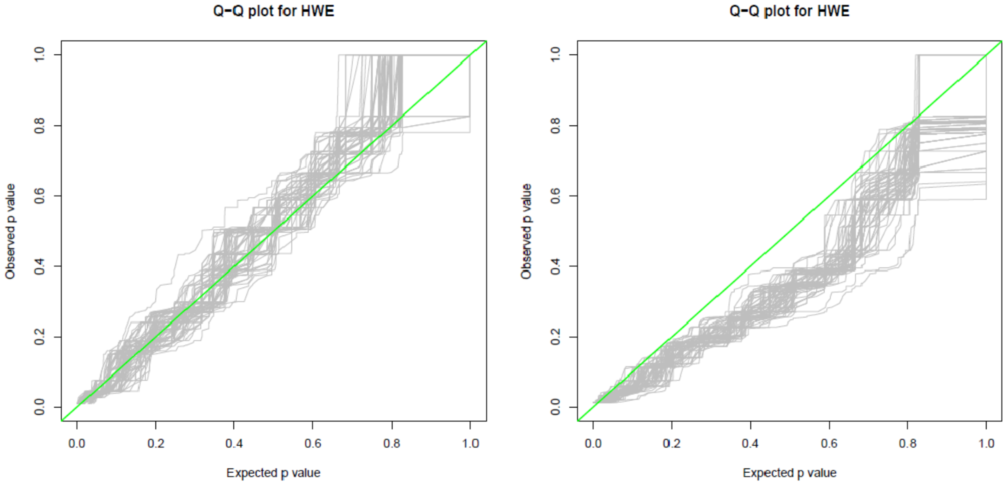


Figure 5: Left panel: Q-Q plot for 225 SNPs on chromosome 1 of a sample of 84 individuals from the Han Chinese population. Right panel: Q-Q plot for simulated data (225 SNPs, 84 individuals) with inbreeding ($f = 0.05$).