Exploring Diallelic Genetic Markers: The HardyWeinberg Package

Jan Graffelman
Universitat Politècnica de Catalunya
version 1.6.1
May 29, 2018

Abstract

Testing genetic markers for Hardy-Weinberg equilibrium is an important issue in genetic association studies. The HardyWeinberg package offers the classical autosomal tests for equilibrium, functions for power computation and for the simulation of marker data under equilibrium and disequilibrium. Recently, specific frequentist and Bayesian tests for X-chromosomal markers have been developed and included in the package. Functions for testing equilibrium in the presence of missing data by using multiple imputation are provided. The package also supplies various graphical tools such as ternary plots with acceptance regions, log-ratio plots and Q-Q plots for exploring the equilibrium status of a large set of diallelic markers. Classical tests for equilibrium and graphical representations for diallelic marker data are reviewed. Several data sets illustrate the use of the package.

Keywords: ternary plot, Q-Q plot, chi-square test, exact test, permutation test, power, log-ratio.

1. Introduction

The HardyWeinberg package (Graffelman 2015) consists of a set of tools for analyzing diallelic genetic markers in the R environment (R Core Team 2014), and is particularly focused on the graphical representation of their (dis)equilibrium condition in various ways. The package is mainly aimed at researchers working in the fields of genetics, statistics, epidemiology, bioinformatics and bio-statistics and is available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=HardyWeinberg. This paper describes the state of the art of version 1.6.0 of the package. If you appreciate this software and wish to cite it, please cite the corresponding paper in the Journal of Statistical Software (Graffelman 2015). The structure of this paper is as follows. In Section 2 we briefly introduce Hardy-Weinberg equilibrium. Section 3 reviews the classical statistical tests and power computation for Hardy-Weinberg equilibrium. Section 4 briefly presents the X-chromosomal tests for equilibrium. Section 6 treats graphical representations of Hardy-Weinberg equilibrium for sets of markers. Section 7 is an example session showing how to analyze genetic markers with the functions of the package. Finally, a discussion (Section 8) with some comments on related packages completes the paper.
2. Hardy-Weinberg equilibrium

A diallelic genetic marker with alleles A and B with respective frequencies \( p \) and \( q \) \((p + q = 1)\) is said to be in Hardy-Weinberg equilibrium if the relative genotype frequencies \( f_{AA} \), \( f_{AB} \) and \( f_{BB} \) are given by \( p^2 \), \( 2pq \) and \( q^2 \) respectively. This law, independently formulated by Hardy (1908) and Weinberg (1908), is a fundamental principle of modern genetics (Crow 1988). The term “Hardy-Weinberg equilibrium” was proposed by Stern (1943). The law is easily extended to a system with multiple alleles \( A_1, \ldots, A_k \) with frequencies \( p_1, \ldots, p_k \), giving genotype frequencies \( p_i^2 \) for homozygotes and \( 2p_ip_j \) for heterozygotes. An alternative formulation of the law for the diallelic case is obtained by squaring the heterozygote frequency:

\[
f_{AB}^2 = 4f_{AA}f_{BB}.
\] (1)

Hardy-Weinberg equilibrium (HWE) is achieved in one generation of random mating. In the absence of disturbing forces (migration, mutation, selection, among other possibilities) the law predicts that genotype and allele frequencies will remain in their equilibrium state over the generations. We refer to genetic textbooks (Crow 1988; Hartl 1980) for a more detailed treatment of the long list of assumptions that underlie HWE. The law plays an important role in the context of genetic association studies for various reasons. Disequilibrium may be the result of genotyping error, most typically the confusion of heterozygotes and homozygotes. Tests for HWE may thus help to detect (gross) genotyping error. On the other hand, disequilibrium among cases in a case-control study may be indicative of disease association. Thus, tests for HWE may also provide clues in marker-disease association studies.

3. Classical autosomal tests for Hardy-Weinberg equilibrium

There are several statistical tests available for investigating whether a genetic marker can be considered to be in equilibrium or not. The classical chi-square test for goodness-of-fit has been the most popular test for HWE for decades, though nowadays exact procedures are more and more often employed. A likelihood ratio test is also available. A description of the different tests is given by Weir (1996, Chapter 3). Bayesian inference for HWE (Lindley 1988; Ayres and Balding 1998; Shoemaker, Painter, and Weir 1998; Wakefield 2010; Consonni, Moreno, and Venturini 2010) is not considered here. In the following sections we summarize the chi-square test (Section 3.1), the likelihood ratio test (Section 3.2), the exact test (Section 3.3) and the permutation test (3.4), and also describe the computation of power for HWE tests (Section 3.5). Testing for HWE with missing genotype data is addressed in Section 3.6.

3.1. Chi-square test

The chi-square test is the classical test for HWE and is typically explained in genetic textbooks (Hedrick 2005; Hartl 1980). Let \( n_{AA} \), \( n_{AB} \) and \( n_{BB} \) represent the observed genotype counts, and \( e_{AA} = np^2 \), \( e_{AB} = 2npq \) and \( e_{BB} = nq^2 \) the expected genotype counts under HWE. The chi-square statistic \( X^2 \) can be computed as

\[
X^2 = \frac{(n_{AA} - e_{AA})^2}{e_{AA}} + \frac{(n_{AB} - e_{AB})^2}{e_{AB}} + \frac{(n_{BB} - e_{BB})^2}{e_{BB}},
\] (2)

and compared with a \( \chi^2_1 \) reference distribution. Alternatively, the chi-square statistic may be
Table 1: Three genotype counts $n_{AA}$, $n_{AB}$ and $n_{BB}$ represented in a two-way table.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$n_{AA}$</td>
<td>$\frac{1}{2} n_{AB}$</td>
</tr>
<tr>
<td>B</td>
<td>$\frac{1}{2} n_{AB}$</td>
<td>$n_{BB}$</td>
</tr>
<tr>
<td></td>
<td>$\frac{1}{2} n_{A}$</td>
<td>$\frac{1}{2} n_{B}$</td>
</tr>
</tbody>
</table>

Table 2: Allele counts represented in a two-way table.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$2 n_{AA}$</td>
<td>$n_{AB}$</td>
</tr>
<tr>
<td>B</td>
<td>$n_{AB}$</td>
<td>$2 n_{BB}$</td>
</tr>
<tr>
<td></td>
<td>$n_A$</td>
<td>$n_B$</td>
</tr>
</tbody>
</table>

expressed as

$$X^2 = \frac{D^2}{p^2q^2n},$$

where $D = \frac{1}{2}(n_{AB} - e_{AB})$ indicates the deviation from independence for the heterozygote. The computation of $D$ (or other disequilibrium statistics) is recommended because $X^2$ itself is not informative about the nature of disequilibrium (excess or lack of heterozygotes). A chi-square test for HWE can be carried out by using function `HWChisq` of the package, and supplying the vector of the three genotype counts. However, in R standard chi-square tests for independence are typically carried out on tables or matrices. If the genotype counts are re-organized in a two-way layout given in Table 1, then a standard chi-square test for independence (function `chisq.test` in R) applied to this table is the same as a chi-square test for HWE.

The total of Table 1 is the number of individuals, and the margins are half the allele counts. If the table is multiplied by 2 then the margins of the table have a more substantive interpretation as allele counts $n_A = 2n_{AA} + n_{AB}$ and $n_B = 2n_{BB} + n_{AB}$, and the total of the table is the total number of alleles, as shown in Table 2.

We note that, due to the multiplication by 2, the latter table has a chi-square statistic that doubles the chi-square statistic of Table 1. It is well known that the chi-square statistic is related to the sample correlation coefficient ($r$) between two indicator variables for the row and column categories by the expression

$$X^2 = nr^2.$$
Hardy-Weinberg Equilibrium in R

<table>
<thead>
<tr>
<th>Individual</th>
<th>Maternal Allele</th>
<th>Paternal Allele</th>
<th>$I_{SB}$</th>
<th>$I_{S'B}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>A</td>
<td>A</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AB</td>
<td>A</td>
<td>B</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BB</td>
<td>B</td>
<td>B</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3: Coding of genotype data by indicator variables.

be estimated by maximum likelihood (ML) as

$$\hat{f} = \frac{4n_{AA}n_{BB} - n_{AB}^2}{n_An_B},$$

and this is identical to the aforementioned sample correlation coefficient $r$ in Equation 4. Function `HWf` of the package computes this statistic.

### 3.2. Likelihood ratio test

In general, the likelihood of a sample of genotype counts is given by the multinomial distribution

$$L(P_{AA}, P_{AB}, P_{BB}) = \binom{n}{n_{AA}, n_{AB}, n_{BB}} P_{AA}^{n_{AA}} P_{AB}^{n_{AB}} P_{BB}^{n_{BB}},$$

and the ML estimator is given by the relative sample genotype frequencies. We thus obtain

$$L_1 = \binom{n}{n_{AA}, n_{AB}, n_{BB}} (\frac{n_{AA}}{n})^{n_{AA}} (\frac{n_{AB}}{n})^{n_{AB}} (\frac{n_{BB}}{n})^{n_{BB}}.$$

Under the assumption of HWE, the likelihood is

$$L_0 = \binom{n}{n_{AA}, n_{AB}, n_{BB}} (\frac{n_A}{2n})^{2n_{AA}} (\frac{n_B}{2n})^{2n_{BB}}.$$

The logarithm of the likelihood ratio of the latter two is given by

$$\ln \left( \frac{L_0}{L_1} \right) = -2n \ln (2) - n \ln (n) + n_{AB} \ln (2) + n_A \ln (n_A) + n_B \ln (n_B)$$

$$- n_{AA} \ln (n_{AA}) - n_{AB} \ln (n_{AB}) - n_{BB} \ln (n_{BB}),$$

and the statistic $G^2 = -2 \ln \left( \frac{L_0}{L_1} \right)$ has, asymptotically, a $\chi^2_1$ distribution. The likelihood ratio test for HWE can be carried out using the function `HWLratio` of the package. Asymptotically, the likelihood ratio test is equivalent to a chi-square test for HWE.

### 3.3. Exact test
Exact test procedures for HWE are based on the conditional distribution of the number of heterozygotes \((N_{AB})\) given the minor allele count \((N_A)\). This distribution was derived by Levene (1949) and Haldane (1954) and is given by Equation 7.

\[
P(N_{AB}|N_A) = \frac{n_A!n_B!2^{n_{AB}}}{2(n_A - n_{AB})!n_{AB}!(n_B - n_{AB})!(2n)!}. \tag{7}
\]

The standard way to compute the \(p\) value of an exact test is to sum probabilities according to Equation 7 for all samples that are as likely or less likely than the observed sample. This way to compute the \(p\) value has been termed the SELOME \(p\) value (select equally likely or more extreme samples). The function \texttt{HWExact} provides the standard exact test for HWE, even though it also implements alternative definitions of the \(p\) value. In particular, the function also offers the possibility to do a one-sided test, or to use the mid \(p\) value (Lancaster 1961). The mid \(p\) value is defined as half the probability of the observed sample plus the probabilities of all possible samples that are less likely than the observed sample. The mid \(p\) value is less conservative, has a type I error rate that is closer to the nominal level, and has been shown to have better power (Graffelman and Moreno 2013).

The exact test for HWE is often confused with Fisher’s exact test for a two-way table. Whereas the chi-square test on the two-way Table 1 is equivalent to a chi-square test for HWE, Fisher’s exact test (implemented in the \texttt{R} function \texttt{fisher.test}) applied to Table 1 or 2 is not equivalent to an exact test for HWE. We note in this respect that the off-diagonal element in Table 1 may be non-integer for an odd number of heterozygotes, and that the exact test is thus not applicable to this table. With regard to Table 2 we note that in Fisher’s exact test, \(n_{AB}\) would be allowed to take on any integer value in the range \(0, \ldots, \min(n_A, n_B)\), since all tables with the same marginal counts are considered. However, in the exact test for HWE, \(n_{AB}\) can only take the values \((0, 2, \ldots, n_{AB})\) if \(n_A\) is even, or \((1, 3, \ldots, n_{AB})\) if \(n_A\) is odd, and thus the results differ from Fisher’s test on a two-way table.

3.4. Permutation test

Hardy-Weinberg equilibrium refers to the statistical independence of alleles within individuals. This independence can also be assessed by a permutation test, where all \(2n\) alleles of all individuals are written out as a single sequence (E.g. AAAAABABBBA...). This sequence is then permuted many times, and for each permuted sequence pairs of successive alleles are taken as individuals. For each permutation a test statistic (the pseudo-statistic) for disequilibrium is computed. The test statistic for the original observed sample is compared against the distribution of the pseudo-statistic, where the latter was generated under the null hypothesis. The \(p\) value of the test is calculated as the fraction of permuted samples for which the pseudo-statistic is equal to or exceeds the test statistic. Such a test is computer intensive but has the advantage that it does not rely on asymptotic assumptions. Function \texttt{HWPerm} performs this test.

3.5. Power calculations

The power of the chi-square test or of an exact test can be calculated if the sample size, minor allele count and significance level \((\alpha)\) are known, and if the degree of deviation from equilibrium (the effect size) is specified. The effect size can be specified by providing a
disequilibrium parameter $\theta$, given by

$$\theta = \frac{P_{AB}^2}{P_{AA}P_{BB}}.$$  \hspace{1cm} (8)

When there is exact equilibrium $\theta = 4$. The situation $\theta > 4$ refers to heterozygote excess, and the situation $\theta < 4$ refers to heterozygote dearth. Alternatively, the degree of disequilibrium may also be parametrized by using the inbreeding coefficient $f$. Under inbreeding, the population genotype frequencies are given by

$$P_{AA} = p_A^2 + p_A p_B f,$$
$$P_{AB} = 2p_A p_B (1 - f),$$
$$P_{BB} = p_B^2 + p_A p_B f,$$

with $-\frac{p_m}{1-p_m} \leq f \leq 1$, and $p_m$ is the minor allele frequency $\min(p_A, p_B)$. If $f = 0$ then the genotype frequencies correspond to the Hardy-Weinberg proportions. Both specifications of disequilibrium are interrelated (Rohlf and Weir 2008). Power calculations are made possible by the HWPower function of the package.

### 3.6. Missing data

Genotype data often have missing values. If missing values are not missing completely at random, inference with respect to HWE may be biased (Graffelman, Sánchez, Cook, and Moreno 2013). Multiple imputation (Little and Rubin 2002) of missing values, taking information from allele intensities and/or neighboring markers and into account, can improve inference for HWE. Function HWMissing of the package does inference for HWE in the presence of missing data. The multiple imputation part is resolved by the package mice (van Buuren and Groothuis-Oudshoorn 2011). In brief, HWMissing computes the inbreeding coefficient (see Equation 5) for each imputed data set, and combines all estimates according to Rubin’s pooling rules. A confidence interval for $f$ and a $p$ value for a test for HWE can then be computed. Alternatively, exact inference for equilibrium when there are missings is also possible by combining the exact $p$ values of the imputed data sets (Graffelman, Nelson, Gogarten, and Weir 2015). An example of inference for HWE with missing values is given in Section 7.

### 4. X-chromosomal tests for Hardy-Weinberg equilibrium

Recently, Graffelman and Weir (2016) have proposed specific tests for HWE for bi-allelic markers on the X-chromosome. These tests take both males and females into account. The X-chromosomal tests can be carried out by the same functions mentioned in the previous Section (HWChisq, HWRatio, HWEnera, HWPerm) and adding the argument x.linked=TRUE to the function call. For a detailed treatment of frequentist X-chromosomal tests, see Graffelman and Weir (2016). The frequentist X-chromosomal procedures are omnibus tests that simultaneously test equality of allele frequencies in males and females and Hardy-Weinberg proportions in females. Recently, a Bayesian method for testing bi-allelic X-chromosomal variants has been proposed by Puig, Ginebra and Graffelman (2017). Examples of frequentist and Bayesian testing of X-chromosomal markers for HWE are given below in Section 7.
5. Hardy-Weinberg equilibrium and allele frequencies of the sexes

Testing HWP for a genetic variant is contingent on the assumption of equality of allele frequencies in the sexes. Likewise, a chi-square of exact test for equality of allele frequencies assumes HWP. Recently, Graffelman and Weir (2017) proposed exact and likelihood ratio procedures that can test HWP and equality of allele frequencies (EAF) jointly or independently, using six different scenarios for a bi-allelic variant. Function `HWLRtest` can compare the different scenarios with a likelihood ratio test (LRT). Function `HWAIC` calculates Akaike’s information criterion (AIC) for the different scenarios, in order to assist the decision which model best fits the data. Examples are given in Section 7.

6. Graphics for Hardy-Weinberg equilibrium

Several graphics can complement statistical tests for HWE, in particular if many markers are tested simultaneously. The package `HardyWeinberg` provides several graphical routines which are briefly discussed in the following subsections, where we consider scatter plots (Section 6.1), ternary plots (Section 6.2), log-ratio plots (Section 6.3) and Q-Q plots (Section 6.4). We will use two data sets to illustrate the different graphics. The first data set, `HapMapCHBChr1`, concerns 225 single nucleotide polymorphisms (SNPs) with no missing data from chromosome 1 for a sample of 84 individuals from the Han Chinese population in Beijing, compiled from the publicly available datasets of the HapMap project (http://hapmap.ncbi.nlm.nih.gov/, The International HapMap Consortium 2007). The second data set, `Mourant`, consists of the genotype counts for the MN blood group locus for 216 samples from different human populations. This data set was compiled by Mourant, Kopeć, and Domaniewska-Sobczak (1976, Table 2.5). We will refer to these data sets as the HapMap and the Mourant data set respectively.

6.1. Scatter plots of genotype frequencies

Relationships between the genotype frequencies can be explored by making scatter plots of the frequencies, such as $f_{AB}$ versus $f_{AA}$ or $f_{BB}$ versus $f_{AA}$. In these scatter plots, genetic markers tend to follow a particular curve described by the Hardy-Weinberg law. Any scatter plot of two of the three genotype frequencies will reveal structure if the law holds. In a plot of $f_{AB}$ versus $f_{AA}$, the Hardy-Weinberg law is given by Equation 10,

$$f_{AB} = 2(\sqrt{f_{AA}} - f_{AA})$$

(10)

and in a plot of $f_{BB}$ versus $f_{AA}$ the law is described by Equation 11,

$$f_{BB} = \left(1 - \sqrt{f_{AA}}\right)^2$$

(11)

These relationships are easily derived from Equation 1. Examples of both plots using the `HapMapCHBChr1` data set are shown in Figure 1. Both graphs show that all samples cluster closely around the HWE curve. The function `HWGenotypePlot` can be used to create these plots.

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 (Cannings and 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 fre-
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.

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 mul-
multiple markers is shown in the right panel of Figure 3. This figure shows 225 SNPs of the dataset HapMapCHBChr1. The function \texttt{HWTernaryPlot} of the package allows the construction of ternary plots with the equilibrium parabola and various acceptance regions.

### 6.3. Log-ratio plots

A vector of genotype counts \((AA, AB, BB)\) can be seen as a composition, where these counts form parts of a whole. Compositional data analysis (Aitchison 1986) is a branch of statistics dedicated to the analysis of compositions. Some of the tools employed in compositional data analysis such as ternary diagrams and log-ratio transformations can be useful for the analysis of genotype counts. Currently, three types of log-ratio transformations are in use: the additive log-ratio (alr) transformation, the centered log-ratio (clr) transformation and the isometric log-ratio (ilr) transformation (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, and Barceló-Vidal 2003). Starting with a vector of genotype counts \((x = (n_{AA}, n_{AB}, n_{BB}))\), the log-ratio transformations for diallelic markers were given by Graffelman and Egozcue (2011), and are also detailed below:

\[
alr(x) = \begin{cases} 
\left( \ln \frac{f_{AA}}{f_{AB}}, \ln \frac{f_{BB}}{f_{AB}} \right), \\
\left( \ln \frac{f_{AB}}{f_{AA}}, \ln \frac{f_{BB}}{f_{AA}} \right), \\
\left( \ln \frac{f_{AA}}{f_{BB}}, \ln \frac{f_{BB}}{f_{BB}} \right), 
\end{cases}
\]

\[
clr(x) = \left( \ln \frac{f_{AA}}{g_m(x)}, \ln \frac{f_{AB}}{g_m(x)}, \ln \frac{f_{BB}}{g_m(x)} \right),
\]

\[
ilr(x) = \begin{cases} 
\left( \frac{1}{\sqrt{2}} \ln \frac{f_{AA}}{f_{BB}}, \frac{1}{\sqrt{6}} \ln \frac{f_{AA} f_{BB}}{f_{AB}^2} \right), \\
\left( \frac{1}{\sqrt{2}} \ln \frac{f_{AB}}{f_{BB}}, \frac{1}{\sqrt{6}} \ln \frac{f_{AB} f_{BB}}{f_{AA}^2} \right), \\
\left( \frac{1}{\sqrt{2}} \ln \frac{f_{AA}}{f_{AB}}, \frac{1}{\sqrt{6}} \ln \frac{f_{AA} f_{AB}}{f_{BB}^2} \right), 
\end{cases}
\]

where \(g_m(\cdot)\) denotes the geometric mean of its argument. Note that there exist 3 alr and 3 ilr transformations depending on which genotype count is used as the divisor in the log-ratios. We will use the first of the three ilr transformations, because the isometric log-ratio transformation yields a space with an orthonormal basis, and because HWE is in these coordinates represented by a simple horizontal line. With this transformation, HWE implies that the second ilr coordinate is constant \((-\sqrt{2/3} \ln (2))\) and the first coordinate is \(\sqrt{2}\) times the log odds of the allele frequency. The package includes some standard routines for computing additive, centered and isometric log-ratio coordinates for vectors of genotype counts (\texttt{HWAlr}, \texttt{HWClr} and \texttt{HWIlr}), and also graphical routines that display markers in log-ratio coordinates (\texttt{HWAlrPlot}, \texttt{HWClrPlot} and \texttt{HWIlrPlot}). HWE is represented in log-ratio coordinates by a perfect linear relationship between the first and second log-ratio coordinate. Examples of the log-ratio plots for the Mourant data and the HapMap data are given in Figure 4.

### 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
Figure 4: Left panel: ilr plot of MN blood group genotype frequencies for 216 samples from different human populations. Right panel: ilr plot for 225 SNPs on chromosome 1 of a sample of 84 individuals from a Han Chinese population. HWE is represented by the horizontal line with ordinate $-\sqrt{2/3} \ln \left(2\right) = -0.57$. Markers are colored according to a chi-square test for HWE (red points significant, green points not significant).

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 ($f = 0.05$).

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 $\alpha = 0.05$, 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 $\chi^2$ 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.

7. An example session

This section shows the basic use of the package for testing and plotting genetic markers. We consider installation (Section 7.1), testing of markers (Section 7.2), power computations (Section 7.5), simulation of marker data (Section 7.6) and graphics for HWE (Section 7.7).

7.1. Installation

The `HardyWeinberg` package can be installed as usual via the command line or graphical user interfaces, e.g., the package can be installed and loaded by:

```r
R> install.packages("HardyWeinberg")
R> library("HardyWeinberg")
```

This will make, among others, the functions `HWChisq, HWData, HWExact, HWLratio, HWMissing, HWPowere, HWQqplot, and HWternaryplot` available. The document describing the package (this paper) can be consulted from inside R by typing:

```r
R> vignette("HardyWeinberg")
```

7.2. Testing autosomal markers for HWE

We show how to perform several classical tests for Hardy-Weinberg equilibrium. As an example we use a sample of 1000 individuals genotyped for the MN blood group locus described by Hedrick (2005, Table 2.4). We store the genotype counts (298, 489 and 213 for MM, MN and NN respectively) in a vector $x$:
```r
R> library("HardyWeinberg")
R> x <- c(MM = 298, MN = 489, NN = 213)
R> HW.test <- HWChisq(x, verbose = TRUE)

Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
Chi2 = 0.1789563 DF = 1 p-value = 0.6722717 D = -3.69375 f = 0.01488253
```

This shows that the chi-square statistic has value 0.179, and that the corresponding \( p \) value for the test is 0.6723. Taking Taking a significance level of \( \alpha = 0.05 \), we do not reject HWE for the MN locus. When \texttt{verbose} is set to \texttt{FALSE} (default) the test is silent, and \texttt{HW.test} is a list object containing the results of the test (chi-square statistic, the \( p \) value of the test, half the deviation from HWE (\( D \)) for the heterozygote (\( D = \frac{1}{2}(f_{AB} - e_{AB}) \)), the minor allele frequency (\( p \)) and the inbreeding coefficient \( f \)). By default, \texttt{HWChisq} applies a continuity correction. This is not recommended for low minor allele frequencies. In order to perform a chi-square test without Yates’ continuity correction, it is necessary to set the \texttt{cc} parameter to zero:

```r
R> HW.test <- HWChisq(x, cc = 0, verbose = TRUE)

Chi-square test for Hardy-Weinberg equilibrium (autosomal)
Chi2 = 0.2214896 DF = 1 p-value = 0.6379073 D = -3.69375 f = 0.01488253
```

The test with correction gives a smaller \( \chi^2 \)-statistic and a larger \( p \) value in comparison with the ordinary \( \chi^2 \) test. The likelihood ratio test for HWE can be performed by typing

```r
R> HW.lrtest <- HWLratio(x, verbose = TRUE)

Likelihood ratio test for Hardy-Weinberg equilibrium
G2 = 0.2214663 DF = 1 p-value = 0.637925
```

Note that the \( G^2 \)-statistic and the \( p \) value obtained are very close to the chi-square statistic and its \( p \) value. An exact test for HWE can be performed by using routine \texttt{HWExact}.

```r
R> HW.exacttest <- HWExact(x, verbose = TRUE)

Haldane Exact test for Hardy-Weinberg equilibrium (autosomal)
using SELOME p-value
sample counts: nMM = 298 nMN = 489 nNN = 213
H0: HWE (D==0), H1: D <> 0
D = -3.69375 p-value = 0.6556635
```

The exact test leads to the same conclusion, we do not reject HWE (\( p \) value = 0.6557). Both one-sided and two-sided exact tests are possible by using the argument \texttt{alternative}, which can be set to \texttt{"two.sided"}, \texttt{"greater"}, or \texttt{"less"}. Three different ways of computing the \( p \) value of an exact test are implemented, and can be specified by the \texttt{pvaluetype} argument, which can be set to \texttt{dost} (double one-sided tail probability), \texttt{selome} (sum equally likely or
HardyWeinberg Equilibrium in R

more extreme) or midp (the mid p value). The exact test is based on a recursive algorithm. For very large samples, R may give an error message "evaluation nested too deeply: infinite recursion". This can usually be resolved by increasing R's limit on the number of nested expressions with options(expressions = 10000) prior to calling HWExact. See ?HWExact for more information on this issue. The permutation test for HWE is activated by:

R> set.seed(123)
R> HW.permutationtest <- HWPerm(x, verbose = TRUE)

Permutation test for Hardy-Weinberg equilibrium
Observed statistic: 0.2214896 17000 permutations. p-value: 0.6606471

and the number of permutations can be specified via the nperm argument. By default the chi-square statistic will be used as the test statistic, but alternative statistics may be supplied by the FUN argument.

All routines HWChisq, HWExact, HWLratio and HWPerm assume that the data are supplied as a vector of genotype counts listed in order (AA, AB, BB). The genotype counts may be specified in a different order, but in that case the elements of the count vector must be appropriately labeled. E.g., the HWChisq function may also be called with:

R> x <- c(MN = 489, NN = 213, MM = 298)
R> HW.test <- HWChisq(x, verbose = TRUE)

Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
Chi2 = 0.1789563 DF = 1 p-value = 0.6722717 D = -3.69375 f = 0.01488253

Often many markers are tested for HWE. If the genotype counts AA, AB, BB are collected in a $m \times 3$ matrix, with each row representing a marker, then HWE tests can be run over each row in the matrix by the routines HWChisqMat and HWExactMat. These routines return a list with the p values and test statistics for each marker.

If, for some reason, the equilibrium status of a particular marker is at stake, you may wish to perform all tests to see to what extent they do agree or disagree. You can use HWAlltests in order to perform all tests with one call and obtain a table of all p values.

R> HW.results <- HWAlltests(x, verbose = TRUE, include.permutation.test = TRUE)

<table>
<thead>
<tr>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square test:</td>
<td>0.2214896 0.6379073</td>
</tr>
<tr>
<td>Chi-square test with continuity correction:</td>
<td>0.1789563 0.6722717</td>
</tr>
<tr>
<td>Likelihood-ratio test:</td>
<td>0.2214663 0.6379250</td>
</tr>
<tr>
<td>Exact test with selome p-value:</td>
<td>NA 0.6556635</td>
</tr>
<tr>
<td>Exact test with dost p-value:</td>
<td>NA 0.6723356</td>
</tr>
<tr>
<td>Exact test with mid p-value:</td>
<td>NA 0.6330965</td>
</tr>
<tr>
<td>Permutation test:</td>
<td>0.2214896 0.6462353</td>
</tr>
</tbody>
</table>
The MN data concern a large sample \((n = 1000)\) with an intermediate allele frequency \((p = 0.4575)\), for which all test results closely agree. For smaller samples and more extreme allele frequencies, larger differences between the tests are typically observed.

We also indicate how to test for HWE when there is missing genotype data. We use the data set \texttt{Markers} for that purpose.

\begin{verbatim}
R> data(Markers)
R> Markers[1:12,]

SNP1  iG  iT SNP2 SNP3
1   TT  641 1037 AA  GG
2   GT 1207  957 AC  AG
3   TT 1058 1686 AA  GG
4   GG 1348  466 CC  AA
5   GT 1176  948 AC  AG
6   GG 1906  912 CC  AA
7   GG 1844  705 CC  AA
8   GG 2007  599 CC  AA
9   GT 1369 1018 AC  AG
10  GG 1936  953 CC  AA
11  GG 1952  632 AC  AG
12 <NA>  947  920 AC  AG
\end{verbatim}

Note that this data is at the level of each individual. Dataframe \texttt{Markers} contains one SNP with missings (SNP1), the two allele intensities of that SNP (iG and iT) and two covariate markers (SNP2 and SNP3). Here, the covariates have no missing values. We first test SNP1 for HWE using a chi-square test and ignoring the missing genotypes:

\begin{verbatim}
R> Xt <- table(Markers[,1])
R> Xv <- as.vector(Xt)
R> names(Xv) <- names(Xt)
R> HW.test <- HWChisq(Xv, cc=0, verbose=TRUE)

Chi-square test for Hardy-Weinberg equilibrium (autosomal)
Chi2 = 8.67309  DF = 1  p-value = 0.003229431  D = -6.77551  f = 0.297491
\end{verbatim}

This gives a significant result \((p < 0.0032)\). If data can be assumed to be missing completely at random (MCAR), then we may impute missings by randomly sampling the observed data. This can be done by supplying the \texttt{method = "sample"} argument, and we create 50 imputed data sets \((m = 50)\).

\begin{verbatim}
R> set.seed(123)
R> Results <- HWMissing(Markers[,1], m = 50, method = "sample", verbose=TRUE)
\end{verbatim}
Test for Hardy-Weinberg equilibrium in the presence of missing values
Inbreeding coefficient $f = 0.3007$
95% Confidence interval (0.1051, 0.4963)
p-value = 0.0027
Relative increase in variance of $f$ due to missings: $r = 0.4471$
Fraction of missing information about $f$: $\lambda = 0.3116$

As could be expected, the conclusion is the same: there is significant deviation from HWE ($p = 0.0027$). It will make more sense to take advantage of variables that are correlated with SNP1, and use multiple imputation of the missings of SNP1 using a multinomial logit model. The multinomial logit model will be used when we set method = "polyreg" or leave the method argument out, since "polyreg" is the default for imputation of factor variables by means of a multinomial logit model used by package mice. We test SNP1 (with missings) for HWE, using a multinomial logit model to impute SNP1 using information from the allele intensities iG and iT and the neighboring markers SNP2 and SNP3.

```R
R> set.seed(123)
R> Results <- HWMissing(Markers[, 1:5], m = 50, verbose = TRUE)
```

Test for Hardy-Weinberg equilibrium in the presence of missing values
Inbreeding coefficient $f = 0.0596$
95% Confidence interval (-0.1079, 0.2271)
p-value = 0.4857
Relative increase in variance of $f$ due to missings: $r = 0.066$
Fraction of missing information about $f$: $\lambda = 0.0621$

Note the sharp drop of the inbreeding coefficient, and the missing data statistics $\lambda$ and $r$. The test is now not significant ($p$-value = 0.4857). Exact inference for HWE with missings is possible by setting the argument statistic="exact". This gives the result

```R
R> set.seed(123)
R> Results <- HWMissing(Markers[, 1:5], m = 50, statistic = "exact", verbose = TRUE)
```

Two-sided Exact test for Hardy-Weinberg equilibrium in the presence of missing values
p-value = 0.4517989

and a similar $p$-value is obtained. See Graffelman et al. (2013) for more details on testing for HWE with missing data.

### 7.3. Testing X-chromosomal markers for HWE

We show here how to perform HWE tests for X-chromosomal markers. We use a vector of 5 elements, containing male and female genotype counts.

```R
R> SNP1 <- c(A=399,B=205,AA=230,AB=314,BB=107)
R> HWChisq(SNP1,cc=0,x.linked=TRUE,verbose=TRUE)
```
Chi-square test for Hardy-Weinberg equilibrium (X-chromosomal)
Chi2 = 7.624175 DF = 2 p-value = 0.022102 D = NA f = -0.0003817242

When males are excluded from the test we get:

```r
R> HWChisq(SNP1[3:5],cc=0)
```

Chi-square test for Hardy-Weinberg equilibrium (autosomal)
Chi2 = 9.485941e-05 DF = 1 p-value = 0.9922291 D = 0.05990783 f = -0.0003817242

Note that the test including males is significant, whereas the test excluding males is not.
The exact test for HWE for an X-chromosomal marker can be performed by adding the `x.linked=TRUE` option:

```r
R> HWExact(SNP1,x.linked=TRUE)
```

Graffelman-Weir exact test for Hardy-Weinberg equilibrium on the X-chromosome using SELOME p-value
Sample probability 5.682963e-05 p-value = 0.02085798

which gives a p-value similar to the \( \chi^2 \) test. When the mid p-value is used we obtain

```r
R> HWExact(SNP1,x.linked=TRUE,pvaluetype="midp")
```

Graffelman-Weir exact test for Hardy-Weinberg equilibrium on the X-chromosome using MID p-value
Sample probability 5.682963e-05 p-value = 0.02082957

These exact tests show that the joint null of Hardy-Weinberg proportions and equality of allele frequencies has to be rejected. An exact test using the females only gives again a non-significant result:

```r
R> HWExact(SNP1[3:5])
```

Haldane Exact test for Hardy-Weinberg equilibrium (autosomal) using SELOME p-value
sample counts: nAA = 230 nAB = 314 nBB = 107
H0: HWE (D==0), H1: D <> 0
D = 0.05990783 p-value = 1

The permutation test for X-linked markers gives

```r
R> HWPerm(SNP1,x.linked=TRUE)
```

Permutation test for Hardy-Weinberg equilibrium of an X-linked marker
Observed statistic: 7.624175 17000 permutations. p-value: 0.02264706
And an X-chromosomal likelihood ratio test gives

\[
R > \text{HWLratio}(\text{SNP1}, \text{x.linked}=\text{TRUE})
\]

Likelihood ratio test for Hardy-Weinberg equilibrium for an X-linked marker
\[
G^2 = 7.693436 \quad \text{DF} = 2 \quad \text{p-value} = 0.02134969
\]

Finally, a summary of all frequentist X-chromosomal tests is obtained by

\[
R > \text{HWAlltests}(\text{SNP1}, \text{x.linked}=\text{TRUE}, \text{include.permutation.test}=\text{TRUE})
\]

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square test</td>
<td>7.624175</td>
<td>0.02210200</td>
</tr>
<tr>
<td>Chi-square test with continuity correction</td>
<td>7.242011</td>
<td>0.02675576</td>
</tr>
<tr>
<td>Likelihood-ratio test</td>
<td>7.693436</td>
<td>0.02134969</td>
</tr>
<tr>
<td>Exact test with selome p-value</td>
<td>NA</td>
<td>0.02085798</td>
</tr>
<tr>
<td>Exact test with dost p-value</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Exact test with mid p-value</td>
<td>NA</td>
<td>0.02082957</td>
</tr>
<tr>
<td>Permutation test</td>
<td>7.624175</td>
<td>0.02211765</td>
</tr>
</tbody>
</table>

Results of all tests are similar. Finally we test equality of allele frequencies in males and females with:

\[
R > \text{AFtest}(\text{SNP1})
\]

Fisher Exact test for equality of allele frequencies for males and females.

Table of allele counts:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>399</td>
</tr>
<tr>
<td>F</td>
<td>774</td>
</tr>
</tbody>
</table>

Sample of 1255 individuals with 1906 alleles. p-value = 0.006268363

For this SNP, there is a significant difference in allele frequency between males and females. Puig, Ginebra and Graffelman (2017) have proposed a Bayesian test for HWE for variants on the X-chromosome which is implemented in the function \text{HWPosterior}.

A Bayesian analysis of the same SNP is obtained by:

\[
R > \text{HWPosterior}(\text{SNP1})
\]

Bayesian test for Hardy-Weinberg equilibrium of X-chromosomal variants.

<table>
<thead>
<tr>
<th>Posterior_Prob</th>
<th>log10(Bayes Factor)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0 (HWE)</td>
<td>0.3384</td>
</tr>
<tr>
<td>M1 (f!=0)</td>
<td>0.0138</td>
</tr>
<tr>
<td>M2 (d!=1)</td>
<td>0.6222</td>
</tr>
<tr>
<td>M3 (f!=0 &amp; d!=1)</td>
<td>0.0256</td>
</tr>
</tbody>
</table>
and shows that a model with Hardy-Weinberg proportions for females and different allele frequencies for both sexes has the largest posterior probability, and the largest Bayes factor.

7.4. Testing sets of markers for HWE

Functions `HWCHisq`, `HWLratio`, `HWExact`, `HWPerm` test a single diallelic marker for HWE. Large sets of markers can be tested most efficiently with the functions `HWCHisqStats` for the chi-square test, and with `HWExactStats` for the exact tests. Both these functions allow for X-linked markers via the `x.linked` argument. Exact tests that rely on exhaustive enumeration are slow in R, and `HWExactStats` now uses by default faster C++ code generously shared by Christopher Chang. The same C++ code is used in the current version (2.0) of Plink (Purcell et al. (2007)).

7.5. Power computation

Tests for HWE have low power for small samples with a low minor allele frequency, or samples that deviate only moderately from HWE. It is therefore important to be able to compute power. The function `HWPower` can be used to compute the power of a test for HWE. If its argument $\theta$ is set to 4 (the default value), then the function computes the type I error rate for the test. Function `mac` is used to compute the minor allele count. E.g.,:

```R
R> x <- c(MM = 298, MN = 489, NN = 213)
R> n <- sum(x)
R> nM <- mac(x)
R> pw4 <- HWPower(n, nM, alpha = 0.05, test = "exact", theta = 4,
+     pvaluetype = "selome")
R> print(pw4)
[1] 0.04822774

R> pw8 <- HWPower(n, nM, alpha = 0.05, test = "exact", theta = 8,
+     pvaluetype = "selome")
R> print(pw8)
[1] 0.9996853
```

These computations show that for a large sample like this one, the type I error rate (0.0482) is very close to the nominal rate, 0.05, and that the standard exact test has good power (0.9997) for detecting deviations as large $\theta = 8$, which is a doubling of the number of heterozygotes with respect to HWE. Type I error rate and power for the chi-square test can be calculated by setting `test="chisq"`. With the allele frequency of this sample (0.5425), $\theta = 8$ amounts to an inbreeding coefficient of -0.1698.

7.6. Simulating data

The package `HardyWeinberg` allows for the simulation of genetic markers under equilibrium and disequilibrium conditions. This enables the user to create simulated data sets that match
Figure 6: Ternary plots for markers simulated under different conditions. (a) multinomial sampling with \( p = 0.5 \). (b) multinomial sampling with a random uniform allele frequency. (c) multinomial sampling with \( p = 0.5 \) and with inbreeding \( (f = 0.5) \). (d) multinomial sampling with a random allele frequency with inbreeding \( (f = 0.5) \). (e) sampling from the Levene-Haldane distribution with fixed allele frequencies. (f) a data set in exact equilibrium with a uniform allele frequency. Red points represent markers that are significant in a chi-square test for HWE, green points represent non-significant markers.
the observed data set in sample size and allele frequency. The comparison of graphics and statistics for observed and simulated datasets is helpful when assessing the extent of HWE for a large set of markers. We simulate \( m = 100 \) markers for \( n = 100 \) individuals by taking random samples from a multinomial distribution with \( \theta_{AA} = p^2 \), \( \theta_{AB} = 2pq \), and \( \theta_{BB} = q^2 \). This is done by routine \texttt{HWData}, which can generate data sets that are in or out of Hardy-Weinberg equilibrium. Routine \texttt{HWData} can generate data that are in exact equilibrium (\texttt{exactequilibrium = TRUE}) or that are generated from a multinomial distribution (default). The markers generated by \texttt{HWData} are independent (there is no linkage disequilibrium). \texttt{HWData} returns a list with both the matrix of genotype counts \( X_t \) and the matrix with genotype compositions \( X_c \) with the relative frequencies of \( AA \), \( AB \) and \( BB \). Routine \texttt{HWData} can simulate genotype counts under several conditions. A fixed allele frequency can be specified by setting \texttt{pfixed = TRUE}, and setting \( p \) to a vector with the desired allele frequencies. Sampling is then according to Levene-Haldane’s exact distribution in Equation 7. If \texttt{pfixed} is \texttt{FALSE}, the given vector \( p \) of allele frequencies will be used in sampling from the multinomial distribution. If \( p \) is not specified, \( p \) will be drawn from a uniform distribution, and genotypes are drawn from a multinomial distribution with probabilities \( p^2, 2pq \) and \( q^2 \) for \( AA \), \( AB \) and \( BB \) respectively. It is also possible to generate data under inbreeding, by specifying a vector of inbreeding coefficients \( f \). We illustrate the use of \texttt{HWData} by simulating several data sets as shown below. Each simulated dataset is plotted in a ternary diagram in Figure 7 in order to show the effect of the different simulation options. We subsequently simulate 100 markers under HWE with allele frequency 0.5 (\( X_1 \)), 100 markers under HWE with a random uniform allele frequency (\( X_2 \)), 100 markers under inbreeding (\( f = 0.5 \)) with allele frequency 0.5 (\( X_3 \)), 100 markers under inbreeding (\( f = 0.5 \)) with a random uniform allele frequency (\( X_4 \)), 100 markers with fixed allele frequencies of 0.2, 0.4, 0.6 and 0.8 (25 each, \( X_5 \)) and 100 markers in exact equilibrium with a random uniform allele frequency (\( X_6 \)).

\begin{verbatim}
R> set.seed(123)
R> n <- 100
R> m <- 100
R> X1 <- HWData(m, n, p = rep(0.5, m))
R> X2 <- HWData(m, n)
R> X3 <- HWData(m, n, p = rep(0.5, m), f = rep(0.5, m))
R> X4 <- HWData(m, n, f = rep(0.5, m))
R> X5 <- HWData(m, n, p = rep(c(0.2, 0.4, 0.6, 0.8), 25), pfixed = TRUE)
R> X6 <- HWData(m, n, exactequilibrium = TRUE)
R> opar <- par(mfrow = c(3, 2),mar = c(1, 0, 3, 0) + 0.1)
R> par(mfg = c(1, 1))
R> HWTernaryPlot(X1, main = "(a)", vbounds = FALSE)
R> par(mfg = c(1, 2))
R> HWTernaryPlot(X2, main = "(b)", vbounds = FALSE)
R> par(mfg = c(2, 1))
R> HWTernaryPlot(X3, main = "(c)", vbounds = FALSE)
R> par(mfg = c(2, 2))
R> HWTernaryPlot(X4, main = "(d)", vbounds = FALSE)
R> par(mfg = c(3, 1))
R> HWTernaryPlot(X5, main = "(e)", vbounds = FALSE)
R> par(mfg = c(3, 2))
\end{verbatim}
\texttt{R}\texttt{> HWTernaryPlot(X6, main = "(f)", vbounds = FALSE)}
\texttt{R}\texttt{> par(opar)}
Figure 7: Ternary plots for markers simulated under different conditions. (a) multinomial sampling with $p = 0.5$. (b) multinomial sampling with a random uniform allele frequency. (c) multinomial sampling with $p = 0.5$ and with inbreeding ($f = 0.5$). (d) multinomial sampling with a random allele frequency with inbreeding ($f = 0.5$). (e) sampling from the Levene-Haldane distribution with fixed allele frequencies, (f) a data set in exact equilibrium with a uniform allele frequency. Red points represent markers that are significant in a chi-square test for HWE, green points represent non-significant markers.
7.7. Graphics for HWE

Genetic association studies, genome-wide association studies in particular, use many genetic markers. In this context, graphics such as ternary plots, log-ratio plots, and Q-Q plots become particularly useful, because they can reveal whether HWE is a reasonable assumption for the whole data set. We begin to explore the Han Chinese HapMap data set by making a ternary plot shown in Figure 8.

```r
R> data("HapMapCHBChr1", package = "HardyWeinberg")
R> HWternaryPlot(HapMapCHBChr1, region = 1, vbounds = FALSE)
R> HWternaryPlot(HapMapCHBChr1, region = 7, vbounds = FALSE)
```

For large databases of SNPs, drawing the ternary plot can be time consuming. Usually, the matrix with genotype counts contains several rows with the same counts. The ternary plot can be constructed faster by plotting only the unique rows of the count matrix. Function `UniqueGenotypeCounts` extracts the unique rows of the count matrix and also counts their frequency. Figure 8 shows 10 significant SNPs (two significant markers overlap). A ternary plot with the acceptance region of the exact test is shown in the right panel of Figure 8. This plot only shows 4 significant markers, and illustrates that the exact test is more conservative. A log-ratio plot of the same data was already shown in Figure 4, and can be created with `HWIlrPlot(HapMapCHBChr1)`. We proceed to make a Q-Q plot of the exact p values. At the same time, we construct a simulated database that matches the `HapMapCHBChr1` database in allele frequency distribution. This is achieved by setting argument `p` of `HWDAta` equal to the allele frequencies of the observed data, where the latter are computed with function `af`.

```r
R> set.seed(123)
R> data("HapMapCHBChr1", package = "HardyWeinberg")
```

![Ternary plots](image)

Figure 8: Ternary plots of 225 SNPs on chromosome 1 of a sample of 84 individuals from a Han Chinese population. Left panel: ternary plot with the acceptance region of a chi-square test. Right panel: ternary plot with the acceptance region of an exact test.
Figure 9: 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, matched in allele frequency).

R> HWQqplot(HapMapCHBChr1)
R> dev.off()
R> set.seed(123)
R> SimulatedData <- HWData(nm = 225, n = 84, p = af(HapMapCHBChr1))$Xt
R> HWQqplot(SimulatedData)

The Q-Q plots in Figure 9 show that both the HapMap dataset and its simulated counterpart are in agreement with HWE.

8. Discussion

The package **HardyWeinberg** offers functions and graphics for analyzing the Hardy-Weinberg equilibrium status of diallelic genetic markers. There are several other packages for the R environment that implement functionality for investigating genetic markers for HWE. The package **genetics** by Warnes (2011) offers data structures for genetic markers, and also includes several functions for testing markers for HWE and for linkage equilibrium. Bayesian tests for HWE are implemented in the package **HWEBayes** of Wakefield (2010) and the package **HWEintrinsic** of Venturini (2011). A loglinear modeling approach to HWE is available in the package **hwde** from Maindonald and Johnson (2011). The **PLINK** software by Purcell et al. (2007) is a standard in genetic data analysis, and can interact with R by means of the package **Rserve** (Urbanek 2013).

We briefly enumerate and comment some features of the **HardyWeinberg** package not provided by the aforementioned R packages: the package provides several graphics for HWE (ternary plots with acceptance regions, log-ratio plots and Q-Q plots against the truly null
distribution). These graphics are useful for analyzing datasets of multiple markers (e.g., a set of markers used in a candidate gene study, or the study of a specific genomic region), and can shed light on the question if the HWE assumption is tenable for the dataset as a whole. The functions provided for the simulation of marker data under HWE (and under disequilibrium) are also useful in this respect. They allow to create datasets that are similar to the observed data in terms of sample size and allele frequency distribution. The comparison of HWE graphics for simulated and observed data can help to rule out or confirm the HWE assumption. Functions for power calculation make it possible to compute the power to detect deviation from HWE for the data at hand. The exact tests of the package are apparently the only ones available that implement several types of p values, and HardyWeinberg is apparently the only software package that performs inference for HWE with missing genotype information using multiple imputation. Future versions of the package may incorporate functions for testing for HWE with multiple alleles. All tests of the package assume homogeneous samples of individuals from one population. Testing for HWE with individuals from different populations (stratification) may also be addressed in future versions of the package.

Acknowledgments

This work was partially supported by grant 2014SGR551 from the Agència de Gestió d’Ajuts Universitaris i de Recerca (AGAUR) of the Generalitat de Catalunya and by grant MTM2012-33236 of the Spanish Ministry of Economy and Competitiveness. This document was generated using Sweave (Leisch 2002). I thank two anonymous referees for their comments that have helped to improve the package and the paper. I also thank professor Steve Marron from the University of North Carolina for his comments on sampling fluctuations in Q-Q plots.

References


**Affiliation:**

Jan Graffelman  
Department of Statistics and Operations Research  
Universitat Politècnica de Catalunya  
Barcelona, Spain  
E-mail: jan.graffelman@upc.edu  
URL: http://www-eio.upc.es/~jan/