Hardy-Weinberg Equilibrium and the Ternary Plot

J. Graffelman[1] and J. Morales Camarena[2]
[1] Universitat Politècnica de Catalunya, Barcelona, Spain; *jan.graffelman@upc.edu*
[2] Universitat Politècnica de Catalunya, Barcelona, Spain; *jair.morales@upc.edu*

Abstract

The Hardy-Weinberg law, formulated about 100 years ago, states that under certain assumptions, the three genotypes AA, AB and BB at a bi-allelic locus are expected to occur in the proportions p^2, $2pq$, and q^2 respectively, where p is the allele frequency of A, and $q = 1 - p$. There are many statistical tests being used to check whether empirical marker data is in agreement with the Hardy-Weinberg principle. Among these are the classical χ^2 test (with or without continuity correction), the likelihood ratio test, Fisher's exact test, and exact tests in combination with MCMC algorithms. Tests for Hardy-Weinberg equilibrium (HWE) are numerical in nature, typically requiring the computation of a test statistic and a p-value. There is however, ample space for the use of graphics in HWE tests, in particular for the ternary plot. Nowadays, many genetical studies are using genetical markers known as Single Nucleotide Polymorphisms (SNPs). SNP data comes in the form of counts, but from the counts one typically computes genotype frequencies and allele frequencies. These frequencies satisfy the unit-sum constraint, and their analysis therefore falls within the realm of compositional data analysis (Aitchison, 1986). SNPs are usually bi-allelic, which implies that the genotype frequencies can be adequately represented in a ternary plot. Compositions that are in exact HWE describe a parabola in the ternary plot. Compositions for which HWE cannot be rejected in a statistical test are typically "close" to the parabola, whereas compositions that differ significantly from HWE are "far". By rewriting the statistics used to test for HWE in terms of heterozygote frequencies, acceptance regions for HWE can be obtained that can be depicted in the ternary plot. This way, compositions can be tested for HWE purely on the basis of their position in the ternary plot (Graffelman & Morales, 2008). This leads to nice graphical representations where large numbers of SNPs can be tested for HWE in a single graph. We show examples of graphical tests for HWE using SNP data from human chromosome 22.

Key words: χ^2-test, genotypic compositions, continuity correction, single nucleotide polymorphism.

1 Introduction

In many modern genetical studies a set of individuals is typed with respect to one or more genetical markers. In particular, modern genotyping technology allows determination of the genotype for a large number of popular markers known as SNPs (Single Nucleotide Polymorphisms). From these data sets, geneticists typically compute allele and genotype frequencies. These frequencies are subject to a unit-sum constraint. Tools from the field of compositional data analysis could therefore prove to be very useful in genetics.

Most SNPs are bi-allelic markers that give rise to 3 genotypes such as AA, AT and TT, if the polymorphism is an A-T polymorphism or AA, GA and GG if the polymorphism is an A-G polymorphism. Such markers should in principle be in Hardy-Weinberg equilibrium if a series of assumptions is met. These assumptions, usually discussed in detail in textbooks on population genetics (Hartl, 1980; Hedrick, 2005) are:

- The organism under study is diploid.
- There is sexual reproduction.
- Non-overlapping generations.
- Random mating (w.r.t. the trait under study).
- Population size is very large.
- Migration is negligible.
- Mutation can be ignored.
- Natural selection does not affect the trait under study.

Let p be the allele frequency of A, and q the allele frequency of B in a population. When all the assumptions above are met, the random union of gametes implies that in one generation the genotype frequencies f_{AA}, f_{AB} and f_{BB} will achieve the Hardy-Weinberg equilibrium frequencies $f_{AA} = p^2$, $f_{AB} = 2pq$, and $f_{BB} = q^2$. Both allele frequencies and genotype frequencies satisfy a unit-sum constraint:

$$p + q = 1 \quad \text{and} \quad f_{AA} + f_{AB} + f_{BB} = 1.$$

If we square the heterozygote frequency, then the Hardy-Weinberg law can also be expressed as

$$f_{AB}^2 = 4 f_{AA} f_{BB}. \tag{1}$$

If there are multiple alleles $A_1, A_2, \ldots A_i$ at a locus, then the Hardy-Weinberg equilibrium condition corresponds to homozygote genotypes $A_1A_1, A_2A_2, \ldots, A_iA_i$ occurring with frequencies $p_1^2, p_2^2, \ldots, p_i^2$ respectively, and heterozygote genotypes $A_1A_2, A_1A_3, \ldots, A_iA_j$ occurring with frequencies $2p_1p_2$, $2p_1p_3$, $\ldots$, $2p_ip_j$ respectively.

If a population is in HWE, then this does not necessarily imply that all the assumptions above are met. Conversely, if some assumptions are not met, the population is not necessarily out of HWE, as the effects of violation of several assumptions simultaneously may cancel out. Genotype data is not error free, e.g. AB individuals may be misclassified as homozygotes. Probably the most common cause for rejection of HWE is genotyping error, rather than violation of some of the assumptions mentioned above. SNP data is extensively tested for HWE prior to subsequent statistical analysis, precisely to detect genotyping error (Hosking, 2004; Wigginton, 2005). Testing

for HWE is part of a quality control procedure for genetical markers.

Genotypic compositions obtained from SNP data are three-way compositions that can be adequately represented in a ternary plot. The latter plot, well-known in compositional data analysis, is known in genetics as a de Finetti diagram (Cannings & Edwards, 1968, Edwards, 1977). An example of a ternary plot is shown in Figure 1. In genetics, the base line of the triangle (AA-BB) has biological meaning: it is a linear 0-1 axis (from left to right) for the frequency of the B allele. The condition of Hardy-Weinberg equilibrium implies a constraint for the compositions. Compositions that are in exact HWE fall on a parabola in the plot.

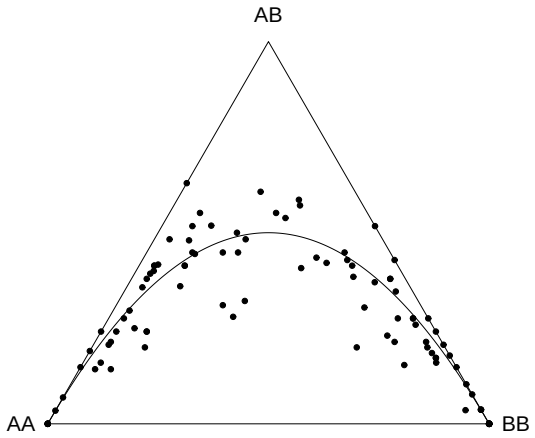

Figure 1: Ternary plot of 107 SNPs from chromosome 22 for a Japanese population

When genotypic compositions are represented in a ternary plot, several situations can be distinguished, depending on the type of data under study. It is possible that only one genetical marker has been typed in several samples (e.g. different biological populations). In this case, the position of each sample in the ternary plot is unambiguous, as for each sample the counts AA, AB and BB are registered and uniquely determine the position of the sample in the plot. However, it also possible to type one sample for many markers simultaneously. If a generic notation is used such as AA, AB, and BB for every marker, then the position of each marker is not uniquely determined. Only the coordinate with respect to the heterozygote axis (AB) is known, and the coordinates with respect to the axis for the homozygotes can be interchanged. This gives a ternary plot where each composition has two possible positions, each composition can be mirrored in the AB axis to obtain its alternative position. When we have many markers, the data constitute in fact a multiway table of high dimensionality, and we only represent the marginals of this table in the ternary plot. The most complicated situation arises if many markers are typed over several samples. The latter situation is better handled by making several ternary plots.

The structure of the remainder of this paper is as follows. In Section 2 we summarize some of the classical statistical tests for Hardy-Weinberg equilibrium. In Section 3 we derive expressions for the acceptance regions of these tests that can be represented in a ternary plot. Section 4 shows an example of graphical significance testing with data from human chromosome 22. We finish the paper with a Discussion (Section 5) and some comments on software for graphical significance testing (Section 6).

2 Statistical tests for HWE

There are several statistical procedures available that test the null hypothesis of Hardy-Weinberg equilibrium. We discuss the χ^2-test, the χ^2-test with continuity correction and Fisher's exact test. A discussion of these tests is also given by Weir (1996).

2.1 The χ^2-test

The classical χ^2-test is probably the most popular test that is used to investigate whether a genetical marker is in equilibrium or not. Let e_{AA}, e_{AB} and e_{BB} be the expected counts of the genotypes under HWE. Let f_{AA}, f_{AB} and f_{BB} be the observed relative frequencies of the genotypes. Let n be the sample size, and D the deviation from independence for the heterozygote count given by $\frac{1}{2}(nf_{AB} - e_{AB})$. Then it is straightforward to show that the sample χ^2-statistic, indicated by X^2, for testing for HWE is related to D, p, q and n by

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

This statistic has a χ^2_1 distribution with one degree of freedom under the null hypothesis (HWE).

2.2 The χ^2-test with continuity correction

When the sample size is small, Yates' continuity correction is often applied, and the corrected χ^2 statistic, denoted as X^2_c, is computed as

$$X^2_c = \sum_{i \geq j} \frac{(|n_{ij} - e_{ij}| - c)^2}{e_{ij}}, \tag{3}$$

where c represents the continuity correction (usually $c = 0.5$) and n_{ij} and e_{ij} are observed and expected counts for the three categories as laid out in a twoway table. The corrected χ^2 statistic is related to D by

$$X^2_c = \frac{D^2 - 2c|D|(1 - pq) + c^2(1 - 3/2pq)}{np^2q^2}. \tag{4}$$

The value of X^2_c is also compared with a χ^2_1 reference distribution. If $c = 0$ then Equation (4) reduces to Equation (2).

2.3 Fisher's exact test

When the sample size is small, Fisher's exact test is often used to test for HWE. A description of this test is given by Agresti (2002). The exact test computes the probabilities of a particular number of heterozygotes given the allele frequencies, under the null hypothesis of HWE. The obtained probabilities are sorted in ascending order. The total probability of obtaining the observed number of heterozygotes or a more extreme result is calculated, and this is the p-value of the test. A detailed example of a Fisher exact test for HWE can be found in Weir (1996).

3 Graphical significance testing

The statistical tests discussed in the previous section can be performed in a graphical manner. This is particularly useful when large numbers of markers or samples are studied. Instead of computing a set of χ^2 statistics and their corresponding p-values, one can represent the genotypic compositions in a ternary plot, jointly with the limits of the acceptance region for HWE. Significant markers are then quickly singled out by mere visual inspection.

3.1 χ^2-test

By substituting $D = \frac{1}{2}(nf_{AB} - e_{AB})$, we can rewrite Equation (2) and express the relative sample frequency of heterozygotes f_{AB} in terms of the allele frequency p to obtain the parabolas

$$f_{AB} = 2pq \pm 2pq\sqrt{X^2/n}. \tag{5}$$

Sample heterozygote frequencies lie on the upper parabola if $D > 0$ (heterozygote excess) and on the lower parabola if $D < 0$ (heterozygote dearth). The heterozygote frequency differs from exact HWE by a quantity of $2pq\sqrt{X^2/n}$. Exact HWE is achieved when $X^2 = 0$. HWE will be rejected if the statistic exceeds a prespecified critical value $\chi^2_1(\alpha)$. From Equation (5) we see that this is equivalent to rejecting HWE whenever the sample frequency of the heterozygote is too large or too small. When there is an excess of heterozygotes ($D > 0$), we will reject HWE when $f_{AB} > 2pq + 2pq\sqrt{\chi^2_1(\alpha)/n}$, and when there is a dearth of heterozygotes ($D < 0$), we will reject HWE when $f_{AB} < 2pq - 2pq\sqrt{\chi^2_1(\alpha)/n}$. We can thus write the acceptance region for HWE as:

$$\left(2pq - 2pq\sqrt{\chi^2_1(\alpha)/n} \leq f_{AB} \leq 2pq + 2pq\sqrt{\chi^2_1(\alpha)/n}\right). \tag{6}$$

The upper and the lower limit of this region are quadratic equations in p that can be represented in the ternary plot for a given sample size and significance level. The upper parabola from Equation (6) cuts the vertices of the triangle for $p = 0$ and $p = 1$, and also cuts the AA-AB and BB-AB edges at the points $p_1 = \sqrt{\chi^2_1(\alpha)/n}/(1 + \sqrt{\chi^2_1(\alpha)/n})$ and $p_2 = 1/(1 + \sqrt{\chi^2_1(\alpha)/n})$.

In Figure 2 we show the acceptance region for HWE for different values of n and α. With these reference graphs in mind, testing for HWE can be done approximately by eye. For small samples, the acceptance region occupies nearly the whole triangle, and HWE is only rejected for extreme heterozygote frequencies.

3.2 χ^2-test with continuity correction

By substituting $D = \frac{1}{2}(nf_{AB} - e_{AB})$, Equation (4) can also be rewritten in terms of the sample frequency of the heterozygotes as

$$f_{AB} = 2pq + \text{sign}(D) \cdot \left(2c(1 - pq)/n \pm 2\sqrt{c^2pq(pq - 0.5)/n^2 + p^2q^2X_c^2/n}\right). \tag{7}$$

This equation describes two sets of two equations that express the heterozygote frequency in terms of the allele frequencies, the continuity correction c and the value of X_c^2. For $c = 0$ Equation (7) reduces to Equation (6).

In Figure 3 we show the acceptance region for HWE for a χ^2-test with continuity correction, for different values of n. In comparison with the left panel of Figure 2 the limits for $D > 0$ have moved upwards, whereas for $D < 0$ the limits have moved downwards. This illustrates the fact

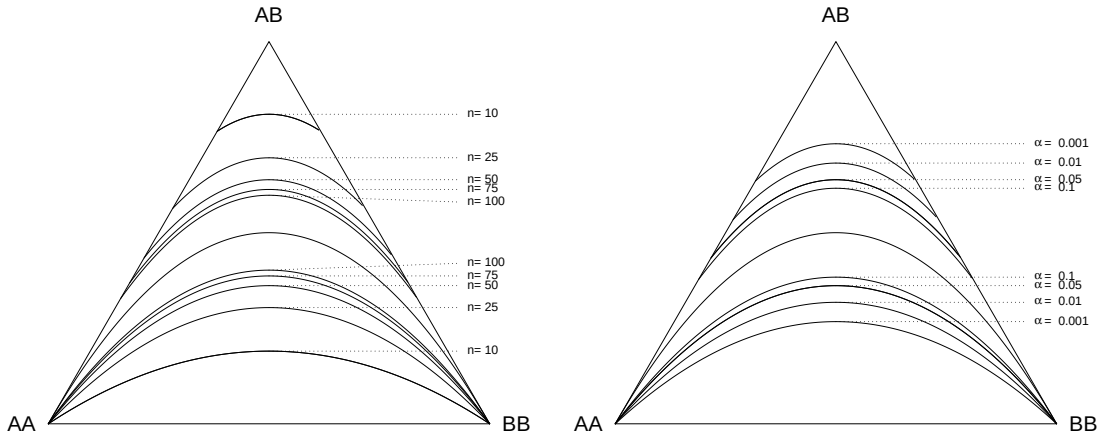

Figure 2: Ternary plots with the acceptance region of a χ^2 test. Left panel: acceptance region as a function of the sample size n given $\alpha = 0.05$. Right panel: acceptance region as a function of α given $n = 50$.

that the test with continuity correction is more conservative, and has a larger acceptance region. For small sample sizes, the upper curves for $D < 0$ cut the edge of the ternary plot twice. This implies that HWE can be rejected for markers with extreme allele frequencies in small samples (see Section 4). For intermediate allele frequencies the region given by Equation (7) can be simplified by considering the outer curves only, and and acceptance region, irrespective of the value of D, can be given as:

$$(2pq - 2c(1-pq)/n - 2\delta \leq f_{AB} \leq 2pq + 2c(1-pq)/n + 2\delta)\,, \tag{8}$$

where $\delta = \sqrt{c^2 pq(pq - 0.5)/n^2 + p^2q^2\chi_1^2(\alpha)/n}$.

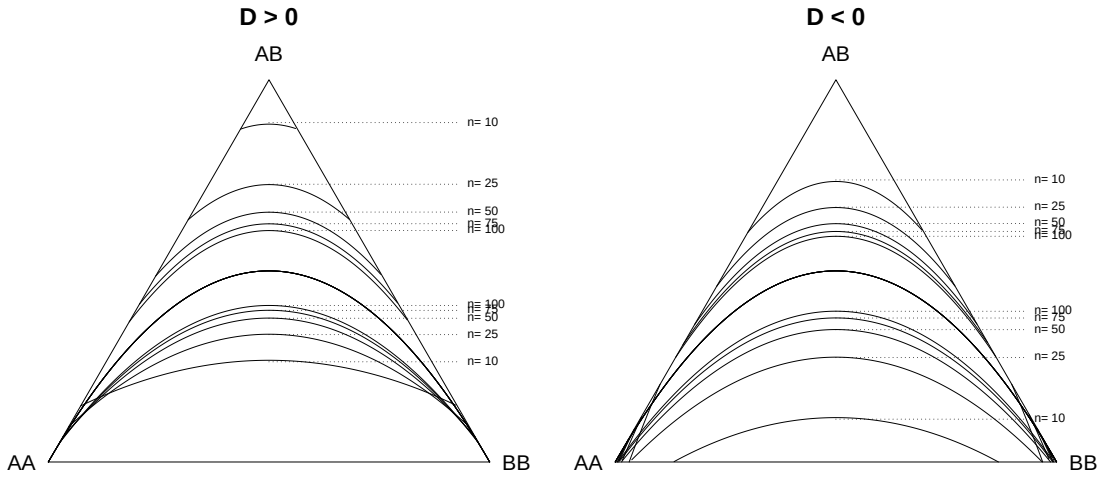

Figure 3: Ternary plots with the acceptance region of a χ^2 test with continuity correction. Left panel: acceptance region as a function of the sample size n given $\alpha = 0.05$ for $D > 0$. Right panel: the acceptance region for $D < 0$ as a function of the sample size n given $\alpha = 0.05$.

3.3 Fisher's exact test

If the number of possible genotypic compositions is not too large, then the p-value of the Fisher exact test of each possible composition for the given sample size can be computed. For a sample of n individuals, there are $\frac{1}{2}(n+1)(n+2)$ possible genotypic compositions. A picture of the acceptance region can then be obtained by colouring the compositions in the ternary plot according to their p-value. By connecting those compositions that are just significant at given allele frequencies by line segments, an approximate picture of the acceptance region is obtained. The so obtained acceptance region is usually not very different from the one obtained by the χ^2-test with continuity correction. Figure 4 shows the acceptance region of a Fisher exact test ($\alpha \leq 0.05$) as a function of the sample size.

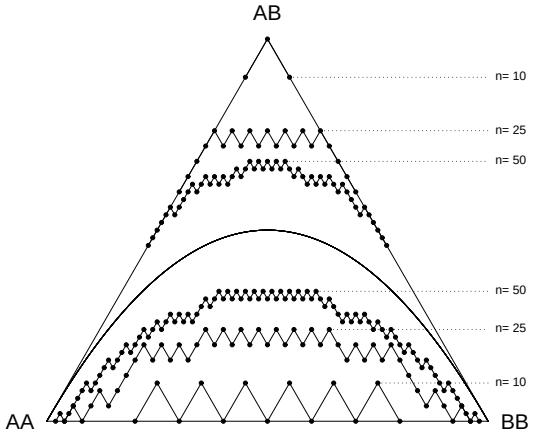

Figure 4: Ternary plots with the acceptance region of a Fisher exact test. Acceptance region as a function of the sample size n given $\alpha \leq 0.05$.

4 Examples

We analyse the set of 107 SNPs on chromosome 22 for 29 individuals of a Japanese population, already introduced in Section 1. Figure 5 shows three ternary plots with the limits of the acceptance region for a χ^2 test, a χ^2 test with continuity correction and a Fisher exact test. The upper left panel shows 7 significant SNPs falling outside the χ^2 acceptance region ($\alpha = 0.05$). This is close to what we would expect to find by chance ($7/107 = 0.065$). Most of the significant SNPs are below the lower bound of the acceptance region. This is an indication of a dearth of heterozygotes for the genomic region studied. Overall 45 SNPs have $D < 0$, and 41 SNPs have $D > 0$. There is a considerable number of SNPs with $D = 0$, typically corresponding to multiple monomorphic markers in the AA and BB vertices of the plot.

The upper right panel of Figure 5 shows 8 significant SNPs falling outside the acceptance region for a χ^2 test with continuity correction. Even though the acceptance region is enlarged with respect to the ordinary χ^2 test, the number of significant SNPs has increased. This is due to the fact that the upper curve of the acceptance region (for $D < 0$) cuts the AA-AB and BB-AB edges twice. Some SNPs with very extreme allele frequencies, close to the AA and BB vertices, are now significant, whereas in the ordinary χ^2 test these were not significant. For extreme allele frequencies the χ^2 test is not recommended, as the expected counts are lower than 5. Markers

with such extreme allele frequencies are close to fixed, and are not as informative as markers with an intermediate allele frequency. The significance of the SNPs with very extreme allele frequencies can be considered an "edge effect" of the continuity correction. We also note that two SNPs (34 and 36) are not significant, but fall outside the acceptance region. The reason is that for these two SNPs the sample size is smaller, due to the presence of missing values. The limiting curves of the acceptance region are based on complete data and are valid for markers with $n = 29$.

The lower left panel of Figure 5 shows that 3 SNPs are significant if a Fisher exact test is used. Due to rounding and small fluctuations in the sample size from marker to marker (due to missing data) the zig-zag line does not perfectly separate significant from non-significant markers. SNPs 46 and 95 are significant in all tests and deserve therefore to be screened in more detail for an excess of missing values or possible genotyping errors.

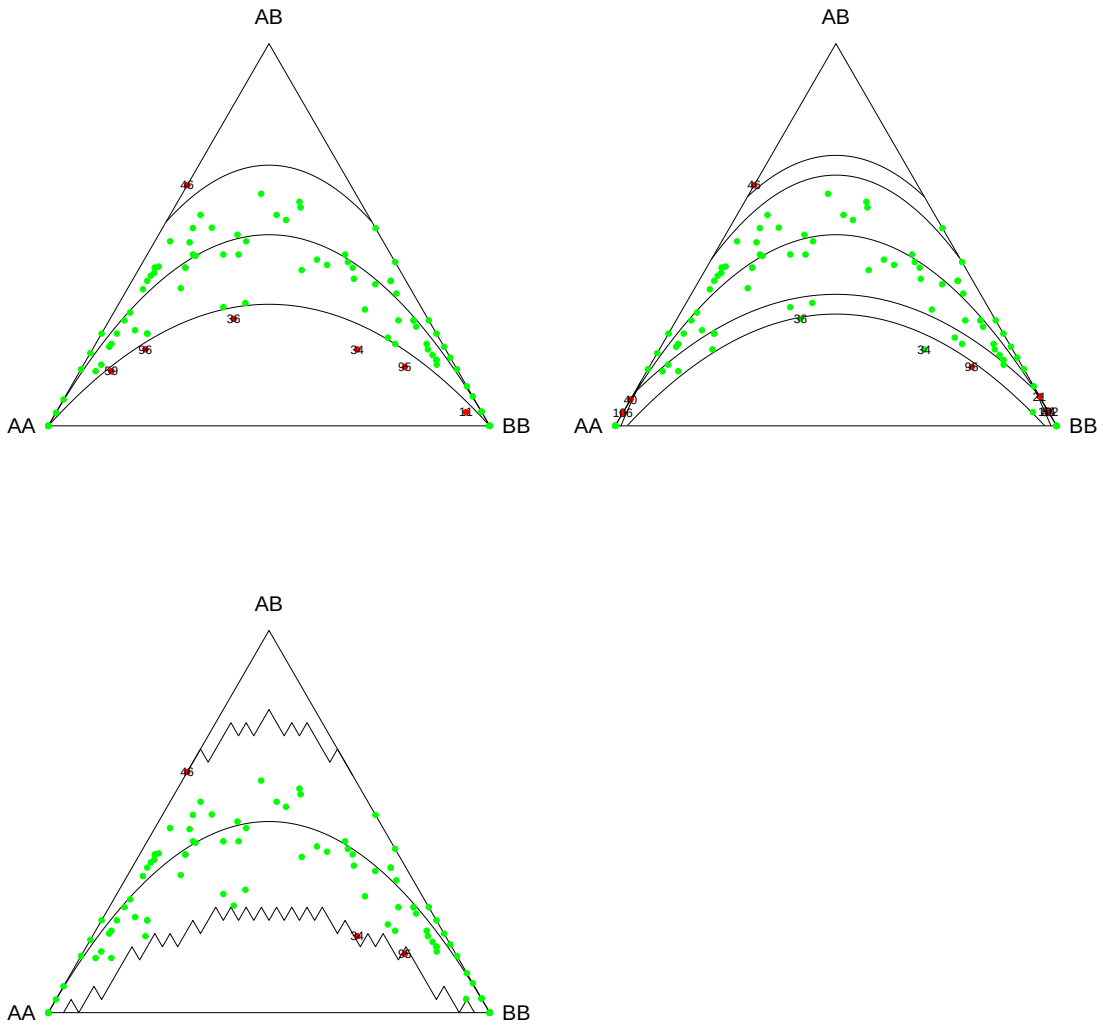

Figure 5: Ternary plots with the acceptance region ($\alpha = 0.05$) of a χ^2 test (upper left panel), a χ^2 test with continuity correction (upper right panel) and a Fisher exact test (lower left panel).

The results in Figure 5 show, for intermediate allele frequencies, that the χ^2 test is less conservative, followed by the Fisher exact test and the χ^2 test with continuity correction.

5 Discussion

We have derived expressions for curves that delimit the acceptance region for HWE in a ternary plot for different statistical tests. This allows for a quick and easy graphical assessment of HWE. The plots proposed in this paper therefore form an excellent tool for the graphical exploration of bi-allelic marker data. The plots in this paper have shown that the three HW tests studied can be ranked in order of increasing conservativeness as χ^2 , FE and χ^2 with continuity correction. This is in agreement with numerical studies from other scholars (Elston and Forthofer, 1977; Hernández and Weir, 1989).

Results from a large-scale genotyping study often yield markers with a considerable percentage of missing values. The number of missing values can vary from marker to marker. This should be taken into account when performing graphical tests for HWE. If the limiting curves are drawn for complete data, the correct curves for markers with missing data will be farther outwards with respect to the HW parabola. We propose however, to keep the curves for the complete data as a reference. This implies that we are more critical for markers with many missing values and more easily "reject" HWE for these markers. As a result, it will be less likely that markers with an excessive amounts of missing data go unnoticed.

Genetic marker data often contains a considerable number of monomorphic markers. Graphically these markers are represented by two single points in the AA and BB vertices of the diagram. It remains necessary to complement the plot with some numerical statistics regarding the percentage of fixed markers or markers with MAF$<$ 0.05 in order to detect the large number of markers possibly piling up in the edges of the diagram.

The proposed plots have shown useful for exploring differences between the various test procedures, and can help to gain more insight in the relative merits of the different tests for HWE. This has been demonstrated in particular by pointing out differences between the χ^2 test, the χ^2 test with continuity correction and the FE test.

6 Software

The R package `HardyWeinberg`, available on CRAN (`http://www.r-project.org`), can be used to construct the type of plots discussed in this paper. It includes a routine to generate data under the null hypothesis of HWE (`HWData`), several routines to test genetical compositions for HWE (`HWChisq, HWLratio`) and a graphical routine (`HWTernaryPlot`) that can make ternary plots with representations of the HW parabola and the different acceptance regions discussed above.

Acknowledgements

This study was supported by grant SEC2003-04476 from the Ministerio de Ciencia y Tecnología from the Spanish Government.

References

Agresti, A. (2002). *Categorical data analysis.* Second edition. John Wiley & Sons, New York.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman & Hall.

Cannings, C. and Edwards, A. W. F. (1968) Natural selection and the de Finetti diagram. *Annals of Human Genetics 31*, pp. 421–428.

Edwards, A. W. F. (1977). *Foundations of mathematical genetics.* Cambridge University Press.

Elston, R. C. and Forthofer, R. (1977). Testing for Hardy-Weinberg equilibrium in small samples. *Biometrics 33* pp. 536–542.

Emigh, T. H. (1980) A Comparison of Tests for Hardy-Weinberg Equilibrium. *Biometrics 36*, pp. 627–642.

Graffelman, J. and Morales-Camarena, J. (2008). Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Human Heredity 65*(2), pp. 77–84. (DOI: 10.1159/000108939).

Hartl, D. L. (1980). *Principles of Population Genetics.* Sinauer Associates, Sunderland, Massachusetts.

Hedrick, P. W. (2005). *Genetics of Populations.* Third edition. Jones and Bartlett Publishers, Sudbury, Massachusetts.

Hernández, J. L., Weir, B. S. (1989). A Disequilibrium coefficient approach to Hardy-Weinberg testing. *Biometrics 45*, pp. 53–70.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Purvis, I. and Xu, C. (2004). Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European Journal of Human Genetics 12*(5), pp. 395–399.

Weir, B. S. (1996). *Genetic Data Analysis II.* Sinauer Associates, Massachusetts.

Wigginton, J. E., Cutler, D. J. and Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics 76*, pp. 887–893.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society [Supplement] 1*, pp. 217–235.