Journal of Genetic Syndromes & Gene Therapy
ISSN: 2157-7412

Hartwig, J Genet Syndr Gene Ther 2014, 5:4
DOI: 10.4172/2157-7412.1000231

Short Communication Open Access

Considerations to Calculate Expected Genotypic Frequencies and Formal Statistical Testing of Hardy-Weinberg Assumptions for non-pseudoautosomal X chromosome SNPs

Fernando Pires Hartwig*

Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, RS, Brazil

Abstract

HWE is a popular concept from population genetics that provides the investigator with an expectation regarding how genotypes are distributed in a given population. Although relying on a set of assumptions, HWE has been proven useful for genetic studies and, among its several applications, has been used as a quality control metric in GWAS. In this correspondence, the calculation to obtain expected genotypic frequencies of non-pseudoautosomal X chromosome SNPs (assuming HWE holds) by weighting within-sex expected genotypic frequencies by the prevalence of the respective sex in the sample is shown. It is also described that, under the assumption that each sex represents 50% of the sample, a one degree of freedom test can be derived. The calculation is simple and intuitive and, therefore, straightforward to incorporate in routine quality control analyses. This correspondence fills an existing gap in GWAS and may contribute to future investigations.

Keywords: Hardy-Weinberg Equilibrium; Data cleaning; X chromosome SNPs; Expected frequencies

Rapid Communication

Hardy-Weinberg Equilibrium (HWE) is a popular population genetics concept which states that, in a given population, genotypic and allelic frequencies will remain constant across generations given a set of assumptions [1,2]. Although some of these are impossible (e.g., infinite population size), HWE proportions tend to approximate observed genotypic proportions well in several contexts (unless there is inbreeding, population stratification and/or genotyping errors). HWE assumptions can be formally tested by comparing expected genotypic counts with observed genotypic counts in a statistical framework (e.g., Chi-squared test with one degree of freedom and Fisher's Exact Test) [3]. Although such testing is straightforward for autosomal SNPs and has been applied extensively for these, there is, to date, no standard approach to test HWE assumptions for X chromosome SNPs.

For a given diallelic non-pseudoautosomal X chromosome SNP (located in the non-pseudoautosomal region) which alleles are X^A and X^a (denoting the major and minor alleles, respectively), the following genotypes can be observed: X^A/X^A, X^A/X^a, X^a/X^a (in females), X^A/Y and X^a/Y (in males). Assuming that HWE holds, the minor allele frequency (MAF) within males (MAF♂) and females (MAF♀) is expected to be the same in the underlying population (i.e., MAF♂ = MAF♀), with differences caused by chance due to sampling variation. In females, the genotypic distribution follows HWE proportions for autosomal variants: $X^A/X^A = (1 - MAF)^2$; $X^A/X^a = 2 \times (1 - MAF) \times (MAF)$; $X^a/X^a = (MAF)^2$. In males, the proportion of each genotype is simply the frequency of each allele: $X^A/Y = (1 - MAF)$; $X^a/Y = MAF$.

Since the expected genotypic proportions under HWE sum to 1 (or 100%) within each sex, the overall expected proportions of each genotype can be calculated as the within-sex proportion multiplied by the prevalence of the respective sex in the sample. Therefore, expected proportions of female genotypes in the sample are: $X^A/X^A = (1 - MAF)^2 \times P♀$; $X^A/X^a = 2 \times (1 - MAF) \times (MAF) \times P♀$; $X^a/X^a = (MAF)^2 \times P♀$, where P♀ = prevalence of females. For male genotypes, the calculation is: $XA/Y = (1 - MAF) \times P♂$; $X^a/Y = MAF \times P♂$, where P♂ = prevalence of males.

In spite of the simplicity of the above calculations, the fact that they provide overall genotypic proportions under HWE can be explored to statistically test HWE assumptions for non-pseudoautosomal X chromosome SNPs. In a common scenario where males are called as homozygous (i.e., X^A/Y individuals are called as X^A/X^A and X^a/Y individuals are called as X^a/X^a), treating a non-pseudoautosomal X chromosome SNP in HWE as an autosomal SNP would result in an excess of homozygotes and too few heterozygotes (as would occur if there were inbreeding), therefore increase the false-positive error rates of the HWE test. A simple modification of the test can be used to make such excess of homozygotes expected, allowing proper HWE testing.

To actually perform the test, the expected genotypic proportions can be used to calculate expected genotypic counts based on observed MAF, prevalence of each sex and number of non-missing observations for the particular SNP (Table 1). Since the genotypes of non-pseudoautosomal X chromosome SNPs for male individuals are normally called as homozygous (wild or variant), the expected counts can be grouped as shown in Table 2. This table displays two distinct situations: one where the prevalence of each sex is estimated from the data and another where it is assumed that each sex represents 50% of the sample. In the former, MAF and prevalence of one of the sexes allow the genotypic frequencies to vary, so this would be a test with two degrees of freedom. In the later, MAF fully determines the frequency of each genotype, so there is only one degree of freedom associated with the test. Since assuming that the prevalence of each sex is 50% may be a good approximation in several situations (more specifically, when

***Corresponding author:** Fernando Pires Hartwig, Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, RS, Brazil, Tel: (5553) 81347172; E-mail: fernandophartwig@gmail.com

Received April 17, 2014; **Accepted** July 28, 2014; **Published** July 31, 2014

Citation: Hartwig FP (2014) Considerations to Calculate Expected Genotypic Frequencies and Formal Statistical Testing of Hardy-Weinberg Assumptions for non-pseudoautosomal X chromosome SNPs. J Genet Syndr Gene Ther 5: 231. doi:10.4172/2157-7412.1000231

Table 1: Calculations of expected counts of non-pseudoautosomal X chromosome SNPs under HWE assumptions.

Sex	Genotypes	5x2 table	
		Expected counts	**Observed counts**
Females	X^A/X^A	$(1 - MAF)^2 \times P_{♀} \times N$	Nº of X^A/X^A
	X^A/X^a	$2 \times (1 - MAF) \times (MAF) \times P_{♀} \times N$	Nº of X^A/X^a
	X^a/X^a	$(MAF)^2 \times P_{♀} \times N$	Nº of X^a/X^a
Males	X^A/Y	$(1 - MAF) \times P_{♂} \times N$	Nº of X^A/Y
	X^a/Y	$MAF \times P_{♂} \times N$	Nº of X^a/Y

N: Number of non-missing observations for the particular SNP.
Nº: Number.
MAF: Minor allele (X^a) frequency in the sample.
$P_{♀}$: Prevalence of females in the sample.
$P_{♂}$: Prevalence of males in the sample.

Table 2: Calculating observed and expected counts for Chi-squared HWE tests with 1 and 2 degrees of freedom.

Counts	Called genotypes		
	Homozygous wild	**Heterozygous**	**Homozygous variant**
Observed	Nº of X^A/X^A + Nº of X^A/Y	Nº of X^A/X^a	Nº of X^a/X^a + Nº of X^a/Y
Expected			
2 df	$[(1 - MAF)^2 \times P_{♀} + (1 - MAF) \times P_{♂}] \times N$	$2 \times (1 - MAF) \times (MAF) \times P_{♀} \times N$	$[(MAF)^2 \times P_{♀} + MAF \times P_{♂}] \times N$
1 df	$\{[(1 - MAF)^2 + (1 - MAF)] \times 0.5\} \times N$	$(1 - MAF) \times (MAF) \times N$	$\{[(MAF)^2 + MAF] \times 0.5\} \times N$

df: Degrees of freedom.
N: Number of non-missing observations for the particular SNP.
Nº: Number.
MAF: Minor allele (X^a) frequency in the sample.
$P_{♀}$: Prevalence of females in the sample.
$P_{♂}$: Prevalence of males in the sample.

there is no expectation of enrichment of one of the sexes in the sample), it could be considered the standard approach because it results in a more powerful test. However, when this assumption does not hold, the two degrees of freedom test should be used at the cost of reduced statistical power.

The rationale described above is particularly important since HWE is widely used as a quality control metric in genetic epidemiology studies, including genome-wide association studies [4], which are currently considered the gold-standard for population-level investigations of the genetic basis of complex traits [5,6]. Regarding HWE tests in comprehensive programs for genome-wide data management and analysis as GenABEL [7], PLINK [8] and SNP and Variation Suite (Golden Helix), there is no description in the documentation on how non-pseudoautosomal X chromosome SNPs are treated in the first two, while the third simply warns the user if there are non-autosomal markers included in the analyses. Of note, for HWE-based filtering, the proposed method should be applied after dealing with heterozygous genotypes in males (e.g., setting such genotypes as missing). In conclusion, the proposed HWE test for non-pseudoautosomal X chromosome SNPs is a straightforward method that could be used in genome-wide data quality control as a biologically meaningful approach.

References

1. Stern C (1943) The HARDY-WEINBERG LAW. See comment in PubMed Commons below Science 97: 137-138.
2. Hartl DL, Clarck AG (2007) Principles of Population Genetics. (4edn) Sinauer.
3. Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. See comment in PubMed Commons below Am J Hum Genet 76: 887-893.
4. Weale ME (2010) Quality control for genome-wide association studies. See comment in PubMed Commons below Methods Mol Biol 628: 341-372.
5. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. See comment in PubMed Commons below Proc Natl Acad Sci U S A 106: 9362-9367.
6. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. See comment in PubMed Commons below Am J Hum Genet 90: 7-24.
7. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. See comment in PubMed Commons below Bioinformatics 23: 1294-1296.
8. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. See comment in PubMed Commons below Am J Hum Genet 81: 559-575.