Front. Biol. 2010, 5(4): 348–353
DOI 10.1007/s11515-010-0580-x

REVIEW

The Hardy-Weinberg principle and its applications in modern population genetics

John J. CHEN

Department of Preventive Medicine, State University of New York at Stony Brook, Stony Brook, NY 11794-8036, USA

Abstract The discovery of the Hardy-Weinberg principle marked the beginning of the field of population genetics. Over the past hundred years, it has provided a starting point for many population genetic investigations. In this review, the Hardy-Weinberg principle, its statistical testing, and several of its applications in various modern population genetic research areas, including allelic variability and selection in the human leukocyte antigen region, microsatellite genotyping error detection, and accuracy of haplotype estimation, are discussed.

Keywords the Hardy-Weinberg principle, statistical testing, population genetics, human leukocyte antigens, microsatellite genotyping, haplotype estimation

1 Introduction

With the rediscovery of Gregor Mendel's Laws of segregation and independent assortment in 1900, many researchers started to concentrate their research efforts in the establishment of properties of Mendel's unit of inheritance, i.e., Mendelian genes. In 1908, British mathematician Godfrey H. Hardy and German physician Wilhelm Weinberg independently discovered the relationship between gene and genotype frequencies, known as the Hardy-Weinberg principle, or Hardy-Weinberg equilibrium (Hardy, 1908; Weinberg, 1908). Since its discovery, the Hardy-Weinberg principle has become a powerful research tool in both theoretical and applied research in population and quantitative genetics (Crow, 1988).

The Hardy-Weinberg principle states that under the condition of large population size, diploid organisms with non-overlapping generations and random mating, the genotype frequencies at a locus are determined by the allele frequencies, and both the genotype and the allele frequencies will stay constant in future generations when the conditions of no mutation, no migration and no selection hold. Hardy-Weinberg equilibrium is not very sensitive to certain kinds of departures from these assumptions and the effects of the deviations from several assumptions can cancel each other out. Therefore, violation from these assumptions may not necessarily result in an observable deviation from Hardy-Weinberg proportions. But deviation from the Hardy-Weinberg equilibrium itself strongly suggests that at least one of the assumptions is violated. There are many possible reasons for a significant deviation from Hardy-Weinberg equilibrium. For instance, population stratification could result in non-random mating. Alternatively, the researcher might have not correctly specified the underlying genetic basis for the trait of interest. The realization of significant deviation from Hardy-Weinberg equilibrium can potentially lead to a better genetic model and establishing interesting alternative hypotheses for further investigation. Because of its elegance and theoretical importance, the Hardy-Weinberg principle has become an important starting point for population genetic investigations.

In the last decade or so, the Hardy-Weinberg principle has again become a subject of intensive study. There are several reasons for this. First, increasing amounts of molecular genetic data are now available to biomedical researchers. Checking genotyping quality has become an important component of genetic analysis. One population genetic tool that has been widely used to detect potential typing error is testing for Hardy-Weinberg equilibrium. It is routinely used to check both single nucleotide polymorphism (SNP) and microsatellite genotype data (e.g., Hosking et al., 2004). Second, it is very common for researchers in population genetics and genetic epidemiology to utilize the Hardy-Weinberg principle as a crucial assumption in many of their disease models. For example, in haplotype analysis, neighboring SNP data are usually utilized to construct the haplotypes through an

Received April 30, 2010; accepted June 8, 2010

E-mail: John.Chen@stonybrook.edu

Expectation-Maximization (EM) statistical algorithm, which assumes that the haplotypes are in Hardy-Weinberg equilibrium (Excoffier and Slatkin, 1995); in case-control molecular epidemiologic studies, associations between disease and genetic markers are usually evaluated under the assumption of the Hardy-Weinberg principle, at least among controls. Evaluation of the impact of deviation from Hardy-Weinberg equilibrium becomes a crucial step of establishing the robustness of the disease models and their results. The study of deviation from Hardy-Weinberg equilibrium has also been utilized in the investigation of allelic variability and selection of a specific genetic region, such as the human leukocyte antigen (HLA) region (e.g, Chen et al., 1999; Solberg et al., 2008).

2 Statistical testing of the Hardy-Weinberg principle

Several statistical methods of testing overall deviation from Hardy-Weinberg equilibrium have been proposed. The traditional χ^2 goodness-of-fit (GOF) test and several corrections to improve the test when small expectations exist have been developed and widely used (Li, 1955; Elston and Forthofer, 1977; Emigh, 1980; Smith, 1986). But it has been recognized that the asymptotic goodness-of-fit tests can sometimes be misleading especially when the sample size and/or some genotypic frequencies are small. Louis and Dempster (1987) proposed an algorithm to generate the exact distribution of a sample drawn from a population in Hardy-Weinberg equilibrium. To avoid the complete enumeration, an overall Hardy-Weinberg test using a Markov Chain Monte Carlo (MCMC) approach was developed (Guo and Thompson, 1992) and has been shown to be more powerful than other overall tests. Bayesian approaches have also been attempted to evaluate the overall deviation from Hardy-Weinberg equilibrium (Shoemaker et al., 1998). Most of the tests mentioned above have been implemented in population genetics software packages, e.g., Arlequin (Excoffier et al., 2005), and PyPop (Lancaster et al., 2007).

Besides checking for overall fit to Hardy-Weinberg equilibrium, researchers are interested in testing specific alternative hypotheses, e.g., heterozygote excess/deficiency, or whether one or several particular genotypes are over or under represented in the data. Rousset and Raymond (1995) provided test procedures for situations of selection or a disturbance to panmixia. When the alternative hypothesis is correctly specified, the tests can be close to optimal. Using a Hardy-Weinberg deviation coefficient, Hernández and Weir (1989) proposed an individual χ^2 test for deviation of an individual heterozygote, and Chen and Thomson (1999) provided the appropriate variance of this test under the null hypothesis. To evaluate the deviation of an individual homozygote, a one degree of freedom χ^2 testing procedure was also recently developed (Chen et al., 2005). Simulation studies show that the individual genotype test not only has reasonable type I error rate, but also superior statistical power, especially when *a priori* hypothesis exists with regard to individual genotype of interest. The results were independent of number of alleles and the overall allele frequency distribution, and only directly related to the allele frequency of the genotype tested and the sample size. When no *a priori* hypothesis exists, researchers need to be careful about the problem of multiple testing when applying individual test to multiple genotypes in a dataset. Although one can utilize such individual test results to generate hypotheses for future studies, it is important to properly interpret the significance of multiple tests, and adjustment of *P*-values, using either the Bonferroni correction or permutation test approach (Westfall and Young, 1993), should be considered.

3 Some applications of the Hardy-Weinberg principle

The Hardy-Weinberg principle has many useful applications. In population genetic studies, especially when family data involving parents and offspring are not feasible, demonstration of Hardy-Weinberg equilibrium provides some support to a Mendelian genetic basis for the trait under investigation.

On the other hand, given a known genetic basis for a trait, significant deviations from Hardy-Weinberg equilibrium provide researchers with strong evidence that some of the assumptions behind the Hardy-Weinberg principle are not satisfied and interesting alternative hypotheses could be generated. In the following several recent applications of the Hardy-Weinberg principle in population genetic research are reviewed.

3.1 Allelic variability and selection in the HLA region

The human major histocompatibility complex (MHC), or the human leukocyte antigen (HLA) super-locus, is located on the chromosomal position 6p21. The ~4 Mb HLA region is one of the most well-defined regions of human genome and contains over 250 coding and non-coding genes. The best-known genes in the HLA region are the subgroup that encode antigen-presenting proteins on the cell surface. Over 100 human diseases, including diabetes, rheumatoid arthritis, asthma and many other autoimmune and infectious disorders have been found to be associated with these genes (Parham, 2005). The HLA region is regarded as the prototype for many current investigative approaches in genomic research on population diversity and human disease association studies (Traherne, 2008).

The HLA is divided into three regions: class I, II, and III. The most intensely researched genes in this region are the nine so-called classical HLA genes: HLA-A, HLA-B,

HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, and HLA-DRB1. The A, B, C, E, F, and G genes belong to class I, whereas the six D genes belong to class II. HLA genes, especially the above classical genes are highly polymorphic. It is not uncommon for an HLA locus to have several hundreds of known alleles (Robinson et al., 2009).

The significant allelic variability of these genes, especially the high level of nonsynonymous coding changes in peptide binding domains, compared with other regions of the genome, strongly suggests that selection might be at work. Balance selection involving heterozygote advantage has been considered as the main mechanism for the high levels of allelic diversity (Takahata et al., 1992). It is argued that individuals who are heterozygous could present a wider range of foreign peptides to T cells and a large sample from an adult population might therefore reveal a significant excess of heterozygotes, given decent selection pressure exists. Besides balancing selection, other mechanisms, e.g., demographic effects, linkage disequilibrium effects, or other forms of selection, are also possible for the generation and maintenance of the high allelic variability of HLA genes and should also be considered. Recently, an alternative hypothesis of host-pathogen co-evolution through frequency-dependent selection has been proposed (Trachtenberg et al., 2003). It hypothesizes that the most common allele will be under the strongest pathogenic pressure, i.e., being selected against. As a result, the less common alleles will tend to be positively selected for.

Evaluating the patterns of deviation from the Hardy-Weinberg principle for HLA class II loci (DRB1, DQA1, DQB1, and DPB1) in data from 26 human ethnic groups, Chen et al. (1999) performed both overall and individual Hardy-Weinberg testing. They reported that for DRB1 locus, seven of 19 aboriginal human groups studied showed significant deviations from Hardy-Weinberg equilibrium, but none of the seven urban populations included indicated any significant deviation. This seems to suggest that historical selective pressures that might be operating on some of the aboriginal populations may not be present in the urban populations because of improved public health in recent history. The analysis concluded that the Hardy-Weinberg deviations were both locus and ethnic group specific. Though other mechanisms might also be at work, recent admixture seemed to be at least partially responsible for the variability observed. This study and other large investigations of MHC genetic data in non-human organisms, e.g., sockeye salmon (Miller et al., 2001), suggest that selection, if it exists, may not be strong enough to be detectable in a single generation, or it may operate in some geographic locations, but not in others.

Using high-resolution HLA allele frequency data, Solberg et al. (2008) conducted a meta-analysis of over 66800 individuals from 497 population samples throughout the world. They investigated several measures of variability and selection, including Hardy-Weinberg proportions, at each locus for each population to investigate the broad global patterns of allelic differentiation. Strong evidence for balancing selection was reported at DQA1, HLA-C, and DQB1 loci. The study confirmed a role for balancing selection in maintaining much of the allelic variation, at least in some HLA loci.

3.2 Genotyping error detection in microsatellite data

With the development of automated sequencing tools, increasing amount of molecular data has become available and microsatellite polymorphisms are among the most commonly used markers in genetic mapping, population genetics, and forensic studies (Hearne et al., 1992; Ellegren, 2000, 2004).

Microsatellites, also called short tandem repeats (STRs), are interspersed throughout the human genome. Unlike the usually bi-allelic SNPs, microsatellite polymorphisms consist of repeating di-, tri-, tetra-, or penta-nucleotide sequences, and tend to be highly polymorphic. These polymorphisms are generated and maintained by a combination of replication slippage mutations and nucleotide substitutions leading to length expansion and contraction events, respectively (Ellegren, 2000, 2004). They can be efficiently typed in a multiplex PCR using fluorescently-labeled primers.

While microsatellites provide an abundant and cost effective source of genetic markers, there are aspects of their typing that are important to consider when using them in genetic analyses. Inaccurate genotype results can occur in microsatellite genotyping, especially when low concentration or low quality DNA samples and loci with multiple alleles were used. For microsatellite heterozygotes, the shorter fragment size generally amplifies better than the larger fragment. The greater the difference in allele fragment sizes, the greater the difference in the degree of amplification. When the preferential amplification is extreme it can be difficult to distinguish the longer allele from background noise leading to the overestimation of homozygotes for the shorter alleles (Demers et al., 1995).

Another related microsatellite genotyping problem is allele dropout (Rodriguez et al., 2001), in which a certain allele simply does not amplify irrespective of allele length. This could be caused by the variability in primer sequence or low concentration or low quality of template DNA. Allele dropout can also result in an overestimation of individual homozygotes.

Without correction, these non-random microsatellite genotyping errors will affect the subsequent analyses using these genotyping results. This is particularly relevant when results from microsatellite typing are used in genetic analyses that assume Hardy-Weinberg equilibrium for the markers such as haplotype frequency estimation (Single et al., 2002). In order to detect and correct potential extreme preferential amplification and allele dropout

problems, properly checking the Hardy-Weinberg principle, especially individual homozygote tests, becomes crucial, especially when excesses of homozygotes are observed (Gomes et al., 1999). An excess of homozygotes primarily for the shorter alleles and a deficiency of heterozygotes between long and short alleles suggest possible extreme preferential amplification problems. Researchers can manually examine the genotyping traces or retype the microsatellite for the suspected individual DNA samples using a modified PCR program designed to reduce the amplification difference between long and short alleles (Rodriguez et al., 2001). Individual homozygote tests could be applied to the datasets, both before and after retyping, and results compared to detect potential sources of errors (Chen et al., 2005).

In a study of six microsatellite markers from 272 whale samples and 33 microsatellites from 213 bowhead whales, the effects of sample and marker specific genotyping errors were investigated through Hardy-Weinberg proportion comparison (Morin et al., 2009). It was found that seven whales, all homozygous for a rare allele, were highly influential on estimates of Hardy-Weinberg proportions for six different microsatellite markers. This result demonstrated that Hardy-Weinberg testing can be very sensitive to homozygotes in low frequency alleles. Indeed more than half of these individuals had genotype errors possibly due to low DNA quality. The study raises the possibility that even small, normal levels of genotyping errors can result in deviation from Hardy-Weinberg equilibrium and hence overestimate the population structure. To alleviate such problems, routine checking on influential individuals and using multiple replications of those samples in microsatellite genotyping are strongly recommended.

3.3 Accuracy of haplotype estimation

SNPs occur frequently throughout the human genome and provide a very useful tool in disease linkage and association investigations. Detecting any effect from a single SNP can be difficult if the SNP under investigation is only partially associated with the disease locus causally affecting the phenotype of interest, or if there are multiple neighboring SNPs that, only when combined together, influence the phenotype. Constructing haplotype using several or many closely linked SNPs to extract multilocus information has become an important component of genetic investigations. After the haplotypes have been inferred, genetic disease risks can then be evaluated through haplotype-based association analysis. For a detailed review of haplotype analysis, see Schaid (2004).

To properly construct haplotypes, one needs to determine the phase of the double heterozygotes. The two haplotypes present in an individual who is heterozygous for both loci cannot be deduced directly from the unphased genotype. If genotype data from parents or relatives are available, the phased genotypes, i.e., which alleles at each locus were transmitted from which parent, can sometimes be uniquely determined. Alternatively, molecular laboratory techniques can be utilized to sequence a specific multilocus allele and as a result, determine the individual haplotypes. However, this process can be very costly and time-consuming.

When investigating relatively large samples of unrelated subjects, statistical approaches in deriving the haplotypes and estimating their frequencies are efficient and irreplaceable. The most widely used iterative method for such purpose is based on the Expectation-Maximization (EM) algorithm (Excoffier and Slatkin, 1995).

The EM algorithm is a general numerical method of finding maximum likelihood estimates especially in situations when incomplete data exist (Dempster et al., 1977). The iterative algorithm contains the following steps: First, the initial values for the unknown parameters are assigned; second, the expected values of the missing data are imputed, i.e., the Expectation (E)-step; third, the maximum likelihood estimates are obtained using the completed data by combining the imputed data and the available data, i.e., the Maximization (M)-step; fourth, the estimated parameter values are then used in a new round of E-step to provide improved imputed values for the missing data. The E- and M-steps can be iteratively applied until convergence is achieved for the estimates.

An important aspect of the EM algorithm for haplotype estimation is that it assumes Hardy-Weinberg equilibrium, at the point when genotype frequencies are replaced by the product of haplotype frequencies (Excoffier and Slatkin, 1995). As a result, this affects the accuracy of EM-based haplotype frequency estimation when data deviate from Hardy-Weinberg equilibrium. This can become even more problematic if the population under investigation is disease- or response-based, e.g., patients in a case-control study, in which the Hardy-Weinberg principle is known to be violated.

Through simulation studies, Fallin and Schork (2000) investigated the relative importance of sampling error (between population and sample frequencies) versus haplotype frequency estimation error (between sample and estimate frequencies) in the haplotype estimation process. Overall, their simulations showed that the magnitude of sampling error is dominant over that of the haplotype estimation error. Even though the paper concluded that the reliability of haplotype frequency estimates tend to be good, departure from Hardy-Weinberg equilibrium did result in meaningful haplotype estimation errors, with excessive heterozygosity increasing estimation errors while excessive homozygosity improving estimation accuracy.

Several approaches have been attempted to take into consideration the possible Hardy-Weinberg deviations in haplotype-based analysis. To improve haplotype estimates, Single et al. (2002) recommended removing from haplotype analysis individual SNPs which showed

significant SNP-level deviation from Hardy-Weinberg equilibrium. An Expectation-Conditional Maximization algorithm (Meng and Rubin, 1993) was implemented for the inference of haplotype effect in an association study, which allows certain patterns of Hardy-Weinberg deviation by incorporating a single common fixation index (Epstein and Satten, 2003). Further research is still needed in this area to properly incorporate diverse patterns of deviations from Hardy-Weinberg equilibrium into the haplotype estimation and analysis process.

References

Chen J J, Duan T, Single R, Mather K, Thomson G (2005). Hardy-Weinberg testing of a single homozygous genotype. Genetics, 170 (3): 1439–1442

Chen J J, Hollenbach J A, Trachtenberg E A, Just J J, Carrington M, Rønningen K S, Begovich A, King M C, McWeeney S, Mack S J, Erlich H A, Thomson G (1999). Hardy-Weinberg testing for HLA class II (DRB1, DQA1, DQB1, and DPB1) loci in 26 human ethnic groups. Tissue Antigens, 54(6): 533–542

Chen J J, Thomson G (1999). The variance for the disequilibrium coefficient in the individual Hardy-Weinberg test. Biometrics, 55(4): 1269–1272

Crow J F (1988). Eighty years ago: the beginnings of population genetics. Genetics, 119(3): 473–476

Demers D B, Curry E T, Egholm M, Sozer A C (1995). Enhanced PCR amplification of VNTR locus D1S80 using peptide nucleic acid (PNA). Nucl Acids Res, 23(15): 3050–3055

Dempster A P, Laird N M, Rubin D B (1977). Maximum likelihood from incomplete data via EM algorithm. J R Stat Soc, B, 39: 1–38

Ellegren H (2000). Heterogeneous mutation processes in human microsatellite DNA sequences. Nat Genet, 24(4): 400–402

Ellegren H (2004). Microsatellites: simple sequences with complex evolution. Nat Rev Genet, 5(6): 435–445

Elston R C, Forthofer R (1977). Testing for Hardy-Weinberg equilibrium in small samples. Biometrics, 33(3): 536–542

Emigh T H (1980). A comparison of tests for Hardy-Weinberg equilibrium. Biometrics, 36(4): 627–642

Epstein M P, Satten G A (2003). Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet, 73 (6): 1316–1329

Excoffier L, Laval G, Schneider S (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online, 1: 47–50

Excoffier L, Slatkin M (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol, 12(5): 921–927

Fallin D, Schork N J (2000). Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet, 67(4): 947–959

Gomes I, Collins A, Lonjou C, Thomas N S, Wilkinson J, Watson M, Morton N (1999). Hardy-Weinberg quality control. Ann Hum Genet, 63(Pt 6): 535–538

Guo S W, Thompson E A (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics, 48(2): 361–372

Hardy G H (1908). Mendelian proportions in a mixed population. Science, 28(706): 49–50

Hearne C M, Ghosh S, Todd J A (1992). Microsatellites for linkage analysis of genetic traits. Trends Genet, 8(8): 288–294

Hernández J L, Weir B S (1989). A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics, 45(1): 53–70

Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu C F (2004). Detection of genotyping errors by Hardy-Weinberg equilibrium testing. Eur J Hum Genet, 12(5): 395–399

Lancaster A K, Single R M, Solberg O D, Nelson M P, Thomson G (2007). PyPop update—a software pipeline for large-scale multilocus population genomics. Tissue Antigens, 69(Suppl 1): 192–197

Li C C (1955). Population Genetics. Chicago: University of Chicago Press

Louis E J, Dempster E R (1987). An exact test for Hardy-Weinberg and multiple alleles. Biometrics, 43(4): 805–811

Meng X L, Rubin D B (1993). Mximum-likelihood estimation via the ECM algorithm – a general framework. Biometrika, 80(2): 267–278

Miller K M, Kaukinen K H, Beacham T D, Withler R E (2001). Geographic heterogeneity in natural selection on an MHC locus in sockeye salmon. Genetica, 111(1–3): 237–257

Morin P A, Leduc R G, Archer F I, Martien K K, Huebinger R, Bickham J W, Taylor B L (2009). Significant deviations from Hardy–Weinberg equilibrium caused by low levels of microsatellite genotyping errors. Molecular Ecology Resources, 9(2): 498–504

Parham P (2005). MHC class I molecules and KIRs in human history, health and survival. Nat Rev Immunol, 5(3): 201–214

Robinson J, Waller M J, Fail S C, McWilliam H, Lopez R, Parham P, Marsh S G E (2009). The IMGT/HLA database. Nucl Acids Res, 37 (SI): D1013–D1017

Rodríguez S, Visedo G, Zapata C (2001). Detection of errors in dinucleotide repeat typing by nondenaturing electrophoresis. Electrophoresis, 22(13): 2656–2664

Rousset F, Raymond M (1995). Testing heterozygote excess and deficiency. Genetics, 140(4): 1413–1419

Schaid D J (2004). Evaluating associations of haplotypes with traits. Genet Epidemiol, 27(4): 348–364

Shoemaker J, Painter I, Weir B S (1998). A Bayesian characterization of Hardy-Weinberg disequilibrium. Genetics, 149(4): 2079–2088

Single R M, Meyer D, Hollenbach J, Nelson M, Noble J A, Erlich H A, Thomson G (2002). Haplotype frequency estimation in patient populations: an example from the HLA region. Genet Epidemiol, 22: 186–195

Smith C A B (1986). Chi-squared tests with small numbers. Ann Hum Genet, 50(Pt 2): 163–167

Solberg O D, Mack S J, Lancaster A K, Single R M, Tsai Y, Sanchez-Mazas A, Thomson G (2008). Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. Hum Immunol, 69(7): 443–464

Takahata N, Satta Y, Klein J (1992). Polymorphism and balancing selection at major histocompatibility complex loci. Genetics, 130(4): 925–938

Trachtenberg E, Korber B, Sollars C, Kepler T B, Hraber P T, Hayes E, Funkhouser R, Fugate M, Theiler J, Hsu Y S, Kunstman K, Wu S,

Phair J, Erlich H, Wolinsky S (2003). Advantage of rare HLA supertype in HIV disease progression. Nat Med, 9(7): 928–935

Traherne J A (2008). Human MHC architecture and evolution: implications for disease association studies. Int J Immunogenet, 35 (3): 179–192

Weinberg W (1908). Über den Nachweis der Vererbung beim Menschen. Jahresh. Ver. Vaterl. Naturkd. Württemb, 64: 369–382. On the demonstration of heredity in man (Translated by S.H. Boyer, 1963)

Westfall P H, Young S S (1993). Resampling-based multiple testing. New York: Wiley