American Journal of
EPIDEMIOLOGY

Volume 158

Number 5

September 1, 2003



Sponsored by the Society for Epidemiologic Research
Published by Oxford University Press

## ORIGINAL CONTRIBUTIONS

# Searching for Disease-Susceptibility Loci by Testing for Hardy-Weinberg Disequilibrium in a Gene Bank of Affected Individuals

**Wen-Chung Lee**

From the Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, Taipei, Taiwan.

*Received for publication June 4, 2002; accepted for publication February 11, 2003.*

The future of genetic studies of complex human diseases will rely more and more on the epidemiologic association paradigm. The author proposes to scan the genome for disease-susceptibility gene(s) by testing for deviation from Hardy-Weinberg equilibrium in a gene bank of affected individuals. A power formula is presented, which is very accurate as revealed by Monte Carlo simulations. If the disease-susceptibility gene is recessive with an allele frequency of ≤0.5 or dominant with an allele frequency of ≥0.5, the number of subjects needed by the present method is smaller than that needed by using a case-parents design (using either the transmission/disequilibrium test or the 2-df likelihood ratio test). However, the method cannot detect genes with a multiplicative mode of inheritance, and the validity of the method relies on the assumption that the source population from which the cases arise is in Hardy-Weinberg equilibrium. Thus, it is prone to produce false positive and false negative results. Nevertheless, the method enables rapid gene hunting in an existing gene bank of affected individuals with no extra effort beyond simple calculations.

disease susceptibility; epidemiologic methods; gene library; genetics; genome; Hardy-Weinberg equilibrium; Monte Carlo method; polymorphism, single nucleotide

Abbreviation: HWT, Hardy-Weinberg disequilibrium test.

***Editor's note:*** *An invited commentary on this article appears on page 401, and the author's response appears on page 404.*

Whereas linkage analysis has been successfully used to localize disease-causing genes for many monogenic diseases, it has been argued that genetic analysis of complex human diseases calls for a new approach—the epidemiologic association paradigm (1, 2). In particular, the application of the transmission/disequilibrium test in a case-parents study has received much attention (3, 4). However, a transmission/disequilibrium test analysis requires parental genotypes, which can pose serious problems in practice. Parents (serving as the control group) may live elsewhere and be hard to reach, may refuse to participate, or simply may have died already. This is particularly true when the disease under study has an age-at-onset in adulthood, such as non-insulin-dependent diabetes, cardiovascular diseases, Alzheimer's disease, many forms of cancers, and so on.

Reprint requests to Dr. Wen-Chung Lee, Graduate Institute of Epidemiology, National Taiwan University, No. 1, Jen-Ai Road, Section 1, Taipei 100, Taiwan (e-mail: wenchung@ha.mc.ntu.edu.tw).

Feder et al. (5) have suggested a control-free "case-only" approach to test for deviation from Hardy-Weinberg equilibrium among affected individuals. (A biallelic marker with alleles *A* and *a* is in Hardy-Weinberg equilibrium when its genotype frequencies are $q^2$ (*AA*), $2q(1-q)$ (*Aa*), and $(1-q)^2$ (*aa*), where $q$ is the frequency of allele *A* in the population. A population is in Hardy-Weinberg equilibrium when all the markers are in Hardy-Weinberg equilibrium.) They and subsequent researchers (6, 7) are concerned mainly with the problem of precise localization of a disease-susceptibility locus. Here, I propose to use the principle as a genome-wide screening tool. The method is especially suited for use in a large referral center, where genotyping is done routinely for affected individuals but where a control group, either the population control or the parental control, is difficult to obtain.

## TESTING FOR HARDY-WEINBERG DISEQUILIBRIUM IN A GENE BANK OF AFFECTED INDIVIDUALS

Suppose that a gene bank of marker genotypes across the whole genome for a total of $n$ affected individuals has been established for a disease in a particular population. For a particular marker (e.g., the *A* marker, with alleles *A* and *a*), the number of cases with genotype *AA* is denoted as $n_{11}$, the number with genotype *Aa* as $n_{12}$, and the number with genotype *aa* as $n_{22}$ ($n_{11} + n_{12} + n_{22} = n$). (This paper considers markers that are biallelic, because a dense map of biallelic single nucleotide polymorphisms will be ready for use in the very near future (8, 9).) The statistic to test for deviation from Hardy-Weinberg equilibrium using this marker, that is, the Hardy-Weinberg disequilibrium test (HWT), is the following (10):

$$\mathrm{HWT} = \frac{\sqrt{n}\times\hat{D}}{\hat{p}(1-\hat{p})},$$

where $\hat{p} = \hat{p}_{11} + \hat{p}_{12}/2 = n_{11}/n + n_{12}/(2n)$ is the allele frequency of *A* in the sample, and $D = -[\hat{p}_{12} - 2\hat{p}(1-\hat{p})]/2 = \hat{p}_{11}\hat{p}_{22} - \hat{p}_{12}^2/4$ is the sample estimate of the disequilibrium coefficient measuring the departure of the frequency of heterozygotes (*Aa*) from the Hardy-Weinberg expected frequency. Note that the square of HWT is the usual goodness-of-fit statistic, that is,

$$\mathrm{HWT}^2 = \frac{[n_{11} - n\hat{p}^2]^2}{n\hat{p}^2} + \frac{[n_{12} - 2n\hat{p}(1-\hat{p})]^2}{2n\hat{p}(1-\hat{p})} + \frac{[n_{22} - n(1-\hat{p})^2]^2}{n(1-\hat{p})^2}.$$

Under Hardy-Weinberg equilibrium (i.e., under $H_0$: $D = 0$), HWT is asymptotically distributed as a standard normal.

Assume that the source population from which the affected individuals arise is a population in Hardy-Weinberg equilibrium. The question here is whether the affected individuals themselves are also in Hardy-Weinberg equilibrium with respect to the *A* marker. Denote the genotype relative risks as $\Psi_1$ (*Aa/aa*) and $\Psi_2$ (*AA/aa*), respectively. (Here the *A* marker is assumed to be a disease-susceptibility gene.) Among the affected individuals, the expected genotypic frequencies of the *A* marker are

$$p_{11}(AA) = \frac{q^2\Psi_2}{R},$$
$$p_{12}(Aa) = \frac{2q(1-q)\Psi_1}{R},$$
$$p_{22}(aa) = \frac{(1-q)^2}{R},$$

and the expected allele frequency of *A* is

$$p = p_{11} + \frac{p_{12}}{2} = q\times\frac{q\Psi_2 + (1-q)\Psi_1}{R},$$

where $q$ is the allele frequency of *A* in the source population, and $R$ is a normalizing constant equal to $q^2\Psi_2 + 2q(1-q)\Psi_1 + (1-q)^2$, ensuring that the three probabilities sum to 1. The disequilibrium coefficient in the affected population is

$$D = p_{11}p_{22} - \frac{p_{12}^2}{4} = \left[\frac{q(1-q)}{R}\right]^2\times(\Psi_2 - \Psi_1^2).$$

It is clear that $D \neq 0$ when $\Psi_2 \neq \Psi_1^2$. Thus, by testing for deviation from Hardy-Weinberg equilibrium in a sample of affected individuals (using the above HWT) marker by marker with proper multiple-testing adjustment, one can screen for susceptibility gene(s) for the disease under study, provided that the gene(s) displays a nonmultiplicative mode of inheritance.

## POWER FORMULA AND POWER COMPARISON

The distribution of the HWT (under both $H_0$ and $H_1$) can be approximated by a normal distribution. Using the multivariate delta method (11), it can be shown that such a distribution has a mean of

$$\mu \approx \frac{\sqrt{n}\times D}{p(1-p)}$$

and a variance of

**TABLE 1. Sample size necessary to gain 80% power ($\alpha = 10^{-7}$) for the Hardy-Weinberg disequilibrium test (number of affected individuals needed), the transmission/disequilibrium test (affected individuals plus their parents), and the likelihood ratio test (affected individuals plus their parents) under various modes of inheritance**

| γ and $q$ | Additive MOI* | | | Recessive MOI | | | Dominant MOI | | |
|---|---|---|---|---|---|---|---|---|---|
| | HWT* | TDT* | LRT* | HWT | TDT | LRT | HWT | TDT | LRT |
| 4.0 | | | | | | | | | |
| 0.01 | 97,643 (0.80)† | 3,285 | 3,426 | 57,643 (0.80) | 13,032,216 | 248,589 | 40,191 (0.80) | 3,345 | 3,423 |
| 0.10 | 2,343 (0.80) | 582 | 564 | 958 (0.79) | 16,893 | 3,018 | 830 (0.80) | 693 | 570 |
| 0.50 | 2,096 (0.80) | 654 | 633 | 427 (0.80) | 621 | 444 | 412 (0.81) | 2,088 | 837 |
| 0.80 | 13,134 (0.80) | 1,659 | 1,704 | 2,187 (0.80) | 777 | 738 | 1,752 (0.80) | 28,152 | 4,371 |
| 2.0 | | | | | | | | | |
| 0.01 | | 17,265 | 18,642 | 439,088 (0.80) | 115,963,566 | 1,629,795 | 366,616 (0.80) | 17,841 | 18,939 |
| 0.10 | | 2,067 | 2,220 | 6,004 (0.80) | 135,213 | 18,606 | 5,436 (0.80) | 2,847 | 2,604 |
| 0.50 | | 1,014 | 1,083 | 1,369 (0.80) | 2,871 | 1,824 | 1,362 (0.81) | 5,517 | 2,559 |
| 0.80 | | 1,902 | 2,043 | 5,063 (0.80) | 2,553 | 2,370 | 4,621 (0.80) | 65,994 | 11,916 |
| 1.5 | | | | | | | | | |
| 0.01 | 1,663,335 (0.80) | 56,199 | 60,348 | 1,660,984 (0.80) | 462,524,670 | 5,810,394 | 1,494,425 (0.80) | 59,265 | 63,468 |
| 0.10 | 25,537 (0.80) | 5,265 | 5,334 | 21,431 (0.80) | 524,082 | 65,169 | 20,106 (0.80) | 8,691 | 8,223 |
| 0.50 | 8,581 (0.80) | 1,392 | 1,413 | 3,802 (0.80) | 9,297 | 5,553 | 3,798 (0.80) | 13,704 | 6,789 |
| 0.80 | 38,628 (0.80) | 2,094 | 2,208 | 11,950 (0.80) | 7,068 | 6,429 | 11,389 (0.80) | 152,478 | 29,637 |

* MOI, mode of inheritance; HWT, Hardy-Weinberg disequilibrium test; TDT, transmission/disequilibrium test; LRT, likelihood ratio test.
† Numbers in parentheses, empirical powers for the HWT based on 100,000 simulations.

$$\sigma^2 \approx 1 + \frac{4(p-0.5)^2}{p^2(1-p)^2} \times D + \frac{0.5 - 3p(1-p) - 8(p-0.5)^2}{p^3(1-p)^3} \times D^2 + \frac{-0.5 + 2p(1-p) + 4(p-0.5)^2}{p^4(1-p)^4} \times D^3.$$

Let $z_x$ denote the $x$-quantile of a standard normal distribution. Then

$$\text{Power of the HWT} \approx \Pr\left(Z < \frac{-z_{1-a/2} - \mu}{\sigma}\right) + \Pr\left(Z > \frac{z_{1-a/2} - \mu}{\sigma}\right),$$

with $Z$ being a standard normal-distributed random variable.

To compare the powers between the present method and the case-parents designs, we considered the same modes of inheritance as used by Knapp (12) (excluding the multiplicative mode of inheritance): 1) the "additive" model ($\Psi_1 = \gamma$ and $\Psi_2 = 2\gamma$, according to Camp's definition (13) of additive mode of inheritance for the sake of comparability), 2) the recessive model ($\Psi_1 = 1$ and $\Psi_2 = \gamma$), and 3) the dominant model ($\Psi_1 = \Psi_2 = \gamma$). The test was two sided with the $\alpha$-level set at $10^{-7}$. This corresponds to $\alpha = 5 \times 10^{-8}$ for the genome-wide, one-sided transmission/disequilibrium tests used by Risch and Merikangas (1). (If allele $A$ is positively associated with the disease and $\alpha$ is small, the power of the one-sided transmission/disequilibrium test with a type I error rate of $\alpha$ is very near the power of the two-sided transmission/disequilibrium test with a type I error rate of $2\alpha$.) For each combination of mode of inheritance, risk parameter $\gamma$ ($\gamma = 1.5, 2, 4$), and allele frequency of $A$ in the source population $q$ ($q = 0.01, 0.1, 0.5, 0.8$), the sample sizes necessary to gain 80 percent power for the (two-sided) HWT were calculated by solving the above power formula using a bisection method. (Note that Camp's additive model with $\gamma = 2$ was not considered here, because it is actually a multiplicative mode of inheritance.)

To check the precision of power approximation, 100,000 simulated data sets at the above-calculated sample sizes were generated. For each round of simulation, the HWT was calculated, and the true power was estimated as the proportion of simulations rejecting the null hypothesis at $\alpha = 10^{-7}$.

Table 1 presents the calculated sample sizes necessary to gain 80 percent power for an effect at a single locus by the HWT under various conditions. The empirical powers based on simulations (in parentheses) match very well with the expected value of 0.80, indicating that the power formula is quite accurate. The same table also presents the sample sizes needed by the case-parents design (using the conventional transmission/disequilibrium test as well as the 2-df likelihood ratio test that assumes Hardy-Weinberg equilibrium in the source population (14)). The sample sizes (numbers of study subjects) needed by the two-sided transmission/disequilibrium test were taken from table 3 of the article by Knapp (12) and were multiplied by three before presentation. (Knapp's paper presented the numbers of case-parents "triads" instead.) The numbers of study subjects needed by the likelihood ratio test were calculated using the method of Longmate (15). It is of particular interest to see that, if the disease-susceptibility gene is recessive with an allele-frequency of $\leq 0.5$ or dominant with an allele-frequency of $\geq 0.5$, the sample sizes needed by the HWT can be smaller than the corresponding sample sizes needed by the transmission/disequilibrium test and the likelihood ratio test (table 1).

## ASSUMPTIONS AND LIMITATIONS

Although this method makes a convenient gene-searching tool especially for the screening of dominant and recessive genes, it has no power at all to detect genes that display a multiplicative mode of inheritance. (The Hardy-Weinberg equilibrium is preserved under such a scenario.) The sample size requirement may further jeopardize its use to search for genes with an additive mode of inheritance. Therefore, even if a very dense array of markers were genotyped across the genome, one should not expect the method to produce a complete catalog of all the susceptibility genes of a disease.

The method assumes Hardy-Weinberg equilibrium in the source population from which the cases arise and also random sampling of cases (or at least that the missing cases are noninformatively missing). A population can deviate from the Hardy-Weinberg equilibrium for various reasons, biologic or nonbiologic (16). Differential survival of individuals with different genotypes is one of the biologic reasons. It causes the adult and elderly segments of a population to deviate from Hardy-Weinberg equilibrium, even if the newborns of that population are in Hardy-Weinberg equilibrium. Another biologic reason is assortative mating in the population, where the probability of mating between two individuals is related to their phenotypic similarity. A positive signal in a case-only HWT analysis should thus be interpreted with caution, for it could imply that the marker being tested is in linkage disequilibrium with a gene that contributes to disease susceptibility (true positive in the present context), with a gene that affects survival, or with a gene that is associated with the choice of mates.

Yet a more subtle nonbiologic reason for deviation from Hardy-Weinberg equilibrium is "population stratification," whereby the population has a mating substructure and mating is restricted to subjects in the same stratum. In this case, even if mating is random within each stratum, the population as a whole may deviate from Hardy-Weinberg equilibrium (16). A positive signal due to population stratification is the bona fide false alarm. Following the lead, a genomic search in the vicinity of the marker(s) with a significant HWT will most likely be ineffectual, with not a gene found that affects survival or mating choice, let alone a gene that contributes to disease susceptibility. If such a mating substructure can be delineated using variables such as ethnicity or race, one can perform a stratified analysis to adjust for the bias (e.g., by defining a "pooled" HWT statistic, such as

$$\sum_s n_s \hat{D}_s \Big/ \sqrt{\sum_s n_s \hat{p}_s^2 (1-\hat{p}_s)^2},$$

where $s$ is the stratum indicator (17)). If not, one can specify the HWT to be one sided, testing exclusively for heterozygote excess ($D < 0$). (A population substructure in itself can produce a deviation from Hardy-Weinberg equilibrium only in the direction of heterozygote deficiency, the Wahlund principle (18).) However, such a one-sided approach will fail to detect recessive genes or genes with $\Psi_2 \geq \Psi_1^2$.

## CONCLUSION

If a gene bank for a disease has been established in a particular population, one can scan the genome for possible disease-susceptibility gene(s) by testing marker by marker for deviation from Hardy-Weinberg equilibrium. This involves no extra effort beyond some simple calculations. The method thus enables rapid gene hunting. However, one should be aware of the potential for false positive and false negative results.

## ACKNOWLEDGMENTS

This study was partly supported by a grant from the National Science Council, Republic of China.

## REFERENCES

1. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516–17.
2. Khoury MJ, Yang Q. The future of genetic studies of complex human diseases: an epidemiologic perspective. Epidemiology 1998;9:350–4.
3. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–16.
4. Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision and admixture. Am J Hum Genet 1995;57:455–64.
5. Feder JN, Gnirke A, Thomas W, et al. A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet 1996;13:399–408.
6. Nielsen DM, Ehm MB, Weir BS. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Hum Genet 1999;63:1531–40.
7. Jiang R, Dong J, Wang D, et al. Fine-scale mapping using Hardy-Weinberg disequilibrium. Ann Hum Genet 2001;65:207–19.
8. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077–82.
9. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001;409:928–33.
10. Hernández JL, Weir BS. A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics 1989;45:53–70.
11. Agresti A. Categorical data analysis. New York, NY: John Wiley & Sons, 1990.
12. Knapp M. A note on power approximation for the transmission/disequilibrium test. Am J Hum Genet 1999;64:1177–85.
13. Camp NJ. Genomewide transmission/disequilibrium testing—consideration of the genotypic relative risks at disease loci. Am J Hum Genet 1997;61:1424–30.
14. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet 1998;62:969–78.
15. Longmate JA. Complexity and power in case-control association studies. Am J Hum Genet 2001;68:1229–37.
16. Sham P. Statistics in human genetics. New York, NY: Oxford University Press, 1998.
17. Nam JM. Testing a genetic equilibrium across strata. Ann Hum Genet 1997;61:163–70.
18. Hartl DL, Clark AG. Principles of population genetics. 3rd ed. Sunderland, MA: Sinauer Associates, Inc, 1997.