Journal of Genetics and Molecular Biology
Vol. 14, No. 3, 159-163, September 1, 2003

Testing for Hardy-Weinberg Disequilibrium Based on Cases Only: A General Approach for Sample Size Calculation

Wen-Chung Lee

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

A genomewide testing for Hardy-Weinberg disequilibrium based on cases only has recently been proposed as a low-cost approach to mapping disease-susceptibility gene(s). Sample size calculation for such an approach had been carried out assuming one of the typed markers is the disease-susceptibility locus itself. In this paper, the author presents a general approach. It is found that the sample size needed increases strikingly when (1) the allele frequency of the marker deviates increasingly from that of the disease-susceptibility locus; and (2) the degree of linkage disequilibrium deviates increasingly from its maximum.

Key words: case-only design, case-parents triads, epidemiologic methods, genetic epidemiology, single nucleotide polymorphism.

Introduction

To fine map a disease-susceptibility locus, Feder et al. [1] proposed an innovative approach to test for deviation from Hardy-Weinberg equilibrium (HWE) among affected individuals. (A bi-allelic marker with alleles A and a is in HWE when its genotype frequencies are q^2[AA], $2q(1-q)$ [Aa], and $(1-q)^2$ [aa] where, q is the frequency of A-allele in the population. A population is in HWE when all the markers are in HWE.) In a recent paper [2], I adopted the method as a genome-wide screening tool. The method is a 'case-only' approach that can dispense with a control group entirely. It was pointed out that "[t]he 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" [2].

Sample size calculation for such a case-only Hardy-Weinberg disequilibrium test (HWT) had been carried out in a genome-wide scan scenario, assuming one of the typed markers is the disease-susceptibility locus itself [2]. In the future when the cost of genotyping single nucleotide polymorphisms (SNPs) [3, 4] goes down, we may use a very dense set of markers to

*Corresponding author: Prof Wen-Chung Lee, Graduate Institute of Epidemiology, National Taiwan University, No. 1, Jen-Ai Rd, 1st Sec, Taipei, Taiwan.
Tel: 886-2-2312-3456x8357 Fax: 886-2-2351-1955 e-Mail:wenchung@ha.mc.ntu.edu.tw

scan the genome. This will guarantee at least one of the typed markers to be very close to the disease-susceptibility locus and hence to be in strong linkage disequilibrium with it. However, the closest marker is in general not the disease-susceptibility locus itself. Therefore, the previous calculation is a special case at best. In this paper, I will present a general approach for sample size calculation.

Methods and Results

Hardy-Weinberg disequilibrium test

Supposed that a gene bank for a disease has been established in a particular population, which consists of marker genotypes across the whole genome for a total of n affected individuals. Denote the marker that is closest to the disease-susceptibility locus as the 'M' marker, with alleles 'M' and 'm'. The number of cases with genotype MM is denoted as n_{11}, number with genotype Mm as n_{12}, and number with genotype mm as n_{22} ($n_{11}+n_{12}+n_{22}=n$). The HWT statistic is as followed [2]:

$$\mathrm{HWT} = \frac{\sqrt{n}\cdot \hat{D}_M}{\hat{P}_M(1-\hat{P}_M)},$$

where $\hat{p}_M = \hat{p}_{11} + \hat{p}_{12}/2 = n_{11}/n + n_{12}/(2n)$ is the allele frequency of M in the sample, and $\hat{D}_M = -[\hat{p}_{12} - 2\hat{p}_M(1-\hat{p}_M)]/2 = \hat{p}_{11}\hat{p}_{22} - \hat{p}_{12}^2/4$ is the sample estimate of the Hardy-Weinberg disequilibrium coefficient regarding the M marker. Under HWE (that is, under H_0: $D_M = 0$), HWT is asymptotically distributed as a standard normal.

Power formula

Let the disease-susceptibility locus be denoted as the 'S' locus, with alleles 'S' and 's'. The locus has penetrances of f_2(SS), f_1(Ss), and f_0(ss), respectively. Assume the source population from which the affected individuals arise to be a population in HWE. The frequencies of the S allele and the M allele in the source population are denoted as q_S and q_M, respectively. In the affected population, the genotypic frequencies of the M marker are [5]:

$$p_{11}[\mathrm{MM}] = q_M^2 + \frac{2q_M\eta\delta + \tau\delta^2}{\phi},$$

$$p_{12}[\mathrm{Mm}] = 2q_M(1-q_M) + \frac{-2[q_M-(1-q_M)]\eta\delta - 2\tau\delta^2}{\phi},$$

$$p_{22}[\mathrm{mm}] = q_M^2 + \frac{-2(1-q_M)\eta\delta + \tau\delta^2}{\phi},$$

where the $\eta = q_s(f_2-f_1)+(1-q_s)(f_1-f_0)$ is the allelic effect (of the S-allele over the s-allele) in the source population, the $\tau = f_2 - 2f_1 + f_0$ is a parameter measuring the deviation from an additive relation in the penetrances, the δ is the linkage disequilibrium parameter between the S locus and the M marker, and the $\phi = q_s^2 f_2 + 2q_s(1-q_s)f_1 + (1-q_s)^2$ is the disease prevalence in the source population. From these equations, the frequency of the M allele in the affected population is

$$p_M = p_{11} + \frac{p_{12}}{2} = q_M + \frac{\eta\delta}{\phi},$$

and the Hardy-Weinberg disequilibrium coefficient in the affected population is

$$D_M = p_{11}p_{22} - \frac{p_{12}^2}{4} = \frac{\delta^2(f_2f_0 - f_1^2)}{\phi^2}.$$

Using the multivariate delta method [6], it can be shown that the distribution of the HWT has a mean of

$$\mu \approx \frac{\sqrt{n}\cdot D_M}{p_M(1-p_M)},$$

and a variance of

$$\sigma^2 \approx 1 + \frac{4(p_M-0.5)^2}{p_M^2(1-p_M)^2}\cdot D_M$$

$$+\frac{0.5-3p_M(1-p_M)-8(p_M-0.5)^2}{p_M{}^3(1-p_M)^3}\cdot D_M{}^2$$
$$+\frac{-0.5+2p_M(1-p_M)+4(p_M-0.5)^2}{p_M{}^4(1-p_M)^4}\cdot D_M{}^3.$$

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

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

with Z being a standard normal-distributed random variable.

Sample size calculation

To calculate sample size, one can solve the above power formula using a bisection method (a root-finding method) [7]. Note that the sample size depends on the genotype relative risks: $\psi_1 = f_1/f_0$, $\psi_2 = f_2/f_0$, the allele frequencies in the source population: q_s, q_M, and the linkage disequilibrium parameter δ, in addition to the α-level and power.

To illustrate, I consider the following modes of inheritance: (1) the 'additive' model ($\psi_1 = 4$ and $\psi_2 = 4 + 4$, according to Camp's definition [8]), (2) the recessive model ($\psi_1 = 1$ and $\psi_2 = 4$), and (3) the dominant model ($\psi_1 = \psi_2 = 4$). Note that the HWT cannot detect gene that displays a multiplicative mode of inheritance ($\psi_2 = \psi_1^2$) [2]. The test was two-sided with α-level set at 10^{-7} (for a genomewide scan). The q_s was fitted at 0.4. The q_M was set at 0.4, 0.3, 0.2, and 0.1, respectively. The proportion of maximum δ between the disease and the marker loci was set at 1.0, 0.9, 0.8, and 0.7, respectively. The sample sizes necessary to achieve 80% power for the (two-sided) HWT were calculated. To check the precision of the approach, 100,000 simulated datasets at the 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 $\alpha = 10^{-7}$.

The table presents the calculated sample

Table. Sample size necessary (number of affected individuals needed) to achieve 80% power ($\alpha = 10^{-7}$) for the Hardy-Weinberg disequilibrium test under various modes of inheritances. Shown in parenthesis are the empirical powers for the HWT based on 100,000 simulations.

$\lvert q_M - q_s \rvert$	Proportion of maximum δ			
	1.0	0.9	0.8	0.7
Additive model				
0.0	1550(.80)	2485(.80)	4139(.80)	7265(.80)
0.1	5644(.80)	8516(.80)	13439(.80)	22473(.80)
0.2	21323(.80)	31202(.80)	47823(.80)	77789(.80)
0.3	128349(.80)	184297(.80)	277303(.80)	442972(.80)
Recessive model				
0.0	352(.79)	553(.80)	906(.80)	1571(.80)
0.1	1176(.80)	1769(.80)	2787(.80)	4655(.80)
0.2	4301(.80)	6304(.80)	9682(.80)	15788(.80)
0.3	25537(.80)	36761(.80)	55476(.80)	88928(.80)
Dominant model				
0.0	349(.81)	541(.81)	876(.80)	1502(.80)
0.1	1063(.81)	1594(.80)	2507(.81)	4189(.80)
0.2	3598(.80)	5312(.80)	8828(.80)	13546(.80)
0.3	20106(.80)	29396(.80)	45086(.80)	73504(.80)

sizes necessary to achieve 80% power for an effect at a single locus by the HWT under various conditions. The empirical powers based on simulations (in parenthesis) match very well with the expected value of 0.80, indicating that the present method is quite accurate. From the table, we also see that the sample size needed increases strikingly when (1) the allele frequency of the marker deviates increasingly from that of the disease-susceptibility locus; and (2) the degree of linkage disequilibrium deviates increasingly from its maximum.

Discussion

The square of HWT is the usual goodness-of-fit statistics [2]. Under the null, it is distributed as a chi-square distribution with one degree of freedom. To approximate its power, one may turn to a noncentral chi-square distribution. However, simulation studies showed that power approximation using such an alternative approach is less than perfect (results not shown). The reason is simple: the noncentral chis-quare distribution has only one parameter—the noncentrality parameter, whereas the normal distribution, upon which the present approach is based, can have two—the μ and the σ^2.

Those gene hunters searching for disease-susceptibility genes must not be discouraged by the present finding that the required sample size becomes considerably larger as the degree of linkage disequilibrium becomes lower. Recently, the human genome has been found to have a block-like structure of linkage disequilibrium [9-12]. Within a haplotype block, it is found that meiotic recombinations are suppressed and that markers display nearly complete linkage disequilibrium with one another. Based on this recent discovery, Weinberg and Morris [13] provided guidelines for choosing genomewide SNP markers to reduce the cost of genotyping and the sample size needed.

Acknowledgement

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

References

1. 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.
2. Lee WC: Searching for disease-susceptibility loci by testing for Hardy-Weinberg disequilibrium in a gene bank of affected individuals. Am J Epidemiol, 2003; 158: 397-400.
3. 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.
4. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 2001; 409: 928-933.
5. 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.
6. Agresti A: Categorical data analysis. New York, NY: John Wiley & Sons, 1990.
7. Press WH, Flannery BP, Teukolsky SA, Vetterling WT: Numerical recipes in C: the art of scientific computing. New York, NY: Cambridge University Press, 1988, p261-263.
8. Camp NJ: Genomewide transmission/ disequilibrium testing—consideration of the genotypic relative risks at disease loci. Am J Hum Genet, 1997; 61: 1424-30.
9. Goldstein DB. Islands of linkage disequilibrium. Nat Genet, 2001; 29: 109-11.
10. Jeffreys AJ, Kauppi L, Neumann R: Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet, 2001; 29: 217-22.
11. Daly MJ, Rioux JD, Schaffner SE, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. Nat Genet, 2001; 29: 229-32.
12. Johnson GCL, Esposito L, Barratt BJ, Smith

AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RCJ, Payne F, Hughes W, Nutland S, Stevens H, Carr P Tuomilehto-Wolf E, Tuomilehto J, Gough SCL, Clayton DG, Todd JA: Haplotype tagging for the identification of common disease genes. Nat Genet, 2001; 29: 233-237.

13. Weinberg CR, Morris RW: Invited commentary: testing for Hardy-Weinberg disequilibrium using a genome single-nucleotide polymorphism scan based on cases only. Am J Epidemiol, 2003; 158: 401-403.