Short Communication

doi: 10.1046/j.1529-8817.2003.00063.x

Testing the Genetic Relation Between Two Individuals Using a Panel of Frequency-unknown Single Nucleotide Polymorphisms

W.-C. Lee*

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

Summary

The author proposes a method to test the genetic relation between two individuals using a panel of SNPs. The method does not require information about the allele frequencies, and as such it can be used to test any pair of individuals from any population(s).

'Single nucleotide polymorphisms' (SNPs) are the most abundant type of human genetic markers (Wang *et al*. 1998; The International SNP Map Working Group, 2001). Here I propose a method to test the genetic relation between two individuals using a panel of SNPs. This method is intended to differentiate the following three possibilities: 1) the two individuals are from the same random mating population and are genetically unrelated (the H_0); 2) the two individuals are genetically related (the H_{a1}); and 3) the two individuals are from different random mating populations (the H_{a2}). The salient feature of the method is that it does not need a priori information about the allele frequencies of the SNPs. As such, it can be used to test any pair of individuals from any population(s). The testing of H_0 against H_{a1} may be useful in premarital genetic counselling to prevent consanguineous mating. It may also help reunite families separated by war or other acts. The testing of H_0 against H_{a2} may help determine ethnic affiliation. It can also monitor the admixture process in a structured population (by testing marrying couples consecutively over time). Additionally, one may perform a 'SNP matching' in a case-control association study, to ensure that the case and his/her matched control come from the same population, thus avoiding population-stratification bias (Ewens & Spielman, 1995; Witte *et al*. 1999).

*Correspondence: W.-C. Lee, Graduate Institute of Epidemiology, National Taiwan University, No. 1, Jen-Ai R.d., 1st Sec., Taipei,Taiwan. Fax: 886-2-235/1955, e-mail: wenchung@ha.mc.ntu.edu.tw

Assume that a total of n homologous SNPs have been genotyped. At a particular locus, two patterns are of interest, namely, the 'discordant homozygotes' (Dh) and the 'concordant heterozygotes' (Ch). The Dh is defined as the homologous SNP patterns (genotype of the first subject, genotype of the second subject): (11,00) or (00,11), whereas the Ch, the pattern: (10,10). Markers producing these two patterns are referred to as 'informative markers', which are indexed by i, $i = 1, \ldots, m (m \leq n)$. Let X_i be the indicator function with value of 1, if the ith marker is Ch, and value of 0, if Dh. The test statistic is $T_1 = m^{-1} \cdot \sum_{i=1}^{m} X_i$.

Let $p_i (q_i = 1 - p_i)$ denote the allele frequency of the ith informative locus. The conditional probabilities for concordance under H_0 are $\pi_i^{H_0} = \Pr(\text{Ch})/[\Pr(\text{Ch}) + \Pr(\text{Dh})] = 4p_i^2q_i^2/(4p_i^2q_i^2 + 2p_i^2q_i^2) = 2/3$. Note that the probabilities are equal for each and every locus and do not depend on p_i's, thus $E^{H_0}(T_1) = 2/3$. If the SNP markers are unlinked or in linkage equilibrium, the X_i's are independent of one another. Therefore, $\text{Var}^{H_0}(T_1) = m^{-2} \cdot \sum_{i=1}^{m} \text{Var}^{H_0}(X_i) = 2/(9m)$. And the statistic, $Z_1 = (T_1 - 2/3)/\sqrt{2/(9m)}$, is asymptotically the standard normal distribution under H_0.

Under H_{a1}, the conditional concordance probabilities are $\pi_i^{H_{a1}} = (4k_0p_i^2q_i^2 + k_1p_iq_i + 2k_2p_iq_i)/(4k_0p_i^2q_i^2 + k_1p_iq_i + 2k_2p_iq_i + 2k_0p_i^2q_i^2) = [2k_0 + 2\psi \cdot (p_iq_i)^{-1}]/[3k_0 + 2\psi \cdot (p_iq_i)^{-1}] \geq 2/3$, where k_0, k_1,

and k_2 denote, respectively, the probabilities of 0, 1, and 2 genes identical by decent (IBD) ($k_0 + k_1 + k_2 = 1$), and the $\psi = k_1/4 + k_2/2$ is the 'kinship coefficient' between the two individuals (Thompson, 1986). Since the probabilities are greater than or equal to the null value of 2/3 (equality holds when $k_0 = 1$ or equivalently $\psi = 0$), one can perform a one-sided test based on Z_1 for concordance excess to test H_0 against H_{a1}.

Under H_{a2}, we have $\pi_i^{H_{a2}} = 4p_i p_i' q_i q_i' / [4p_i p_i' q_i q_i' + p_i^2 (q_i')^2 + (p_i')^2 q_i^2] = 2/[3 + (p_i - p_i')^2 \cdot (2p_i p_i' q_i q_i')^{-1}] \leq 2/3$, where the p_i and the p_i' ($q_i = 1 - p_i$, $q_i' = 1 - p_i'$) represent the allele frequencies in the two source populations. This time, the probabilities are in the opposite direction from null (equality holds when $p_i = p_i'$). Thus, a one-sided test for concordance deficiency can be used to test H_0 against H_{a2}.

To relax the assumption of linkage equilibrium between the SNP markers, we first select a total of J widely spaced (and thus independent) 'localities' along the genome (indexed by j). Next, we type n_j SNP markers at/around each locality. The informative markers in the jth locality are indexed by i, $i = 1, \ldots, m_j (m_j \leq n_j)$. Let X_{ij} be the indicator function for the ith informative marker in the jth locality, with value defined as before. Let $D_j = \sum_{i=1}^{m_j} (X_{ij} - 2/3)$. As before, $E^{H_0}(D_j) = 0$. Since D_j's are independent of one another, we have $\mathrm{Var}^{H_0}(\sum_{j=1}^{J} D_j/J) = \sum_{j=1}^{J} \mathrm{Var}^{H_0}(D_j)/J^2 \approx \sum_{j=1}^{J} D_j^2/J^2$. Hence, the statistic, $Z_2 = \sum_{j=1}^{J} D_j / \sqrt{\sum_{j=1}^{J} D_j^2}$, is asymptotically the standard normal distribution under H_0.

Acknowledgement

This study was partly supported by a grant from the National Science Council, R.O.C.

References

Ewens, W.J. & Spielman, R.S. (1995) The transmission/disequilibrium test: history, subdivision and admixture. *Am J Hum Genet* **57**, 455–464.

The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933.

Thompson, E.A. (1986) Pedigree Analysis in Human Genetics. The Johns Hopkins University Press, Baltimore.

Wang, D.G. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082.

Witte, J.S., Gauderman, W.J. & Thomas, D.C. (1999) Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am J Epidemiol* **149**, 693–705.

Received: 24 June 2003
Accepted: 26 September 2003