Interpreting a genetic case-control finding: What can be said, what cannot be said and implications in Indian populations

Saurabh Ghosh
Human Genetics Unit, Indian Statistical Institute, Kolkata, India

Identification of genetic variants responsible for complex disorders using association mapping is an active area of research. There are two broad classes of association methodologies: population-based case-control studies and family-based transmission analyses. While case-control analyses are more popular and in general, more powerful than family-based analyses, they suffer from some inherent limitations. Thus, it is of importance, to understand the implications of an association finding obtained from a case-control study design. This article discusses the relative advantages and disadvantages of the two classes of association analyses, particularly in the context of genetic diversity in Indian populations.

Key words: Family-based transmission, genetic association, population stratification

Introduction

With the development of dense maps of highly polymorphic microsatellite markers and single nucleotide polymorphisms (SNPs), there have been considerable successes in mapping of susceptible genes for heritable traits. Most of these genes pertain to Mendelian disorders [for example, the CFTR gene for Cystic Fibrosis[1] which are believed to be controlled by a single major locus with a known mode of inheritance. However, there has also been successful mapping of genes involved in complex disorders like the 5q31 cytokine gene cluster for Crohn’s Disease.[2] Unlike simple Mendelian traits, complex traits are believed to be controlled by multiple loci, some with minor gene effects and genetic variation at any one locus does not completely determine the trait. These traits tend to be more common than Mendelian traits and environmental factors often modify the genetic risk to develop the disease.

The two major statistical tools used in the mapping of genes are linkage and association. Linkage analysis[3] is a statistical method to identify chromosomal regions where related individuals exhibit similar patterns of inheritance of a particular trait. Association studies[4] are designed to narrow down the regions identified by linkage analyses to localize the position of a mutant allele responsible for the underlying trait.

Studies have shown that association studies are statistically more powerful than linkage studies in gene-mapping of complex traits.[5] This is because association, which is measured by a parameter called linkage disequilibrium exists over small distances on the genome, while linkage exists over larger distances. Thus, a positive association finding gives a more precise location of a locus responsible for the trait. This article concentrates on the different statistical methods for genetic association analyses and discusses their relative advantages and disadvantages.

Population-based case-control studies

The most popular and intuitively simple statistical method for testing allelic association is case-control analysis. When studying a particular disorder, a random set of individuals who are affected with the disorder (cases) and an unrelated set of unaffected individuals (controls) are sampled from the population and genotyped at candidate loci for the disease. Thus, the case-control design is population-based comprising unrelated individuals. A significant difference in the distributions
of allele frequencies between cases and controls at any locus would provide evidence of allelic association between the candidate locus and a true disease locus. The statistical significance is evaluated by a standard two-sample test for equality of binomial proportions and the test statistic has an asymptotic standard normal distribution (the square of the test statistic is distributed asymptotically as \( \chi^2 \) with 1 degree of freedom) under the null hypothesis of no association. An alternative to the allele-based test is a genotype-based test where the null hypothesis of no association is equivalent to the genotypic distributions at a candidate locus being the same for cases and controls. The test statistic is asymptotically distributed as \( \chi^2 \) with 2 degrees of freedom for a biallelic locus. While the candidates can themselves be true disease loci, the statistical test would only indicate that each locus showing significant allelic differences between cases and controls is in linkage disequilibrium with one disease locus. Thus, a positive association finding may identify only functional polymorphisms that may or may not be the mutant alleles responsible for the disease.

While data for case-control studies are much easier to collect than family-based data and the test is statistically very powerful for detecting allelic association, case-control studies suffer from some inherent limitations and require very careful designs to circumvent these limitations. The two major concerns regarding case-control studies are issues pertaining to case-control matching and presence of population stratification. Epidemiological studies have shown that allelic distributions may vary greatly with different ethnic populations. Thus, if the cases are selected from one ethnic group and the controls from a separate group, it is quite possible that due to different allelic compositions, spurious association results may be obtained. However, it is not sufficient to simply match the case group and the control group by ethnic backgrounds to avoid spurious association. If the study population is a pooled set of genetically diverse sub-populations, false positive association can result. To illustrate this phenomenon, let us consider the following example as provided in Table 1. Suppose we perform a case-control study at a biallelic locus with alleles \( A \) and \( a \). Based on 120 cases and 120 controls (denoted by “pooled population” in Table 1), we can see that the frequency of allele \( A \) among cases is 0.6, while that among controls is 0.3. This indicates a statistically significant difference at level 0.05 and would lead to an inference that allele \( A \) predisposes to high risk of the disease. However, it may happen in reality that the population considered is a combination of two sub-populations (denoted by “Population 1” and “Population 2” in the Table). In “Population 1”, the frequency of \( A \) among both cases and controls is 0.8, while that in “Population 2” is 0.2 among both cases and controls. Thus, allele \( A \) provides no evidence of association with the disease in either of the sub-populations. The spurious association is generated by pooling the two genetically heterogeneous populations (as is evident from the different allele frequencies of \( A \) in the two populations).

While there is no efficient procedure to test for population stratification, there are some intelligent methods to correct the test statistic for possible population stratification effect. The most popular approach is known as “genetic controls”.\(^6\),\(^7\) A large set of marker loci dispersed throughout the different chromosomes on the genome, preferably neutral (like \textit{Alu} insertion/deletion polymorphisms), unlinked to the polymorphism under study and known to have varying allelic distributions across populations are genotyped and case-control tests are performed at all these loci using the same sample. Since these loci are expected to be unrelated to the disease, the value of the case-control \( \chi^2 \) trend statistic\(^8\) at the candidate locus should be higher than most of the values at the “genomic control” loci in the absence of population stratification. However, in the presence of population stratification, these \( \chi^2 \) values are inflated and the test statistic at the candidate

| Genotype | Population 1 | Population 2 | Pooled population |
|----------|--------------|--------------|-------------------|
|          | Cases        | Controls     | Cases            | Controls       | Case | Controls |
| AA       | 56           | 14           | 4                | 10             | 60   | 24        |
| Aa       | 16           | 4            | 8                | 20             | 24   | 24        |
| aa       | 8            | 2            | 28               | 70             | 36   | 72        |
| Total    | 80           | 20           | 40               | 100            | 120  | 120       |

Table 1: Case-control data in the presence of population stratification
locus needs to be normalized by a factor depending on the median of the Chi-squares values obtained for the “genomic control” loci. However, one requires a very large number of “genomic control” loci for efficient estimation of this normalizing factor and hence the strategy may not be very cost-effective, particularly in view of the fact that these loci do not provide any information on genetic association with the disease under study. Pritchard et al.\[9\] proposed a likelihood-based approach to adjust for population stratification, which involves estimation of allele frequencies for different sub-populations within the data and the ancestry of these alleles. The major limitation of this method is that it depends on apriori classification of individuals into sub-populations. Thus, the problem of population stratification in genetic case-control studies has not yet been circumvented satisfactorily and continues to be an active area of research.

**The alternative**

The alternative to population-based case-control studies is analyses-based on family-based designs. The most popular of them is the transmission disequilibrium test (TDT) proposed by Spielman et al.\[10\] The design comprises “trios”: a nuclear family ascertained by an affected proband and his/her parents. The statistical method requires at least one of the parents to be heterozygous at the polymorphism of interest. The test statistic measures the degree of preferential transmission of a particular marker allele from a heterozygous parent to the affected offspring over the other alleles and is distributed as Chi-squares with 1 degree of freedom in the absence of allelic association. Moreover, from a statistical perspective, the distribution of this statistic is the same even in the absence of linkage between the marker locus and the true putative disease locus. Thus, a statistical significant TDT result implies the presence of both linkage and allelic association. In other words, the TDT is protected against population stratification: spurious association cannot result due to admixture of genetically different sub-populations in the data set. This is the greatest advantage of the TDT over case-control studies. On the other hand, the TDT design is disadvantageous in many respects compared to case-control designs. For late onset diseases like Alzheimers disease, it is often very difficult to obtain data on parents. Sham and Curtis\[11\] showed that even if one heterozygous parent is available and it can be deciphered which allele the parent transmitted to the offspring (for example, if one parent has genotype $Aa$, the other parent is missing and the offspring has genotype $aa$), the TDT can be biased if the heterozygosity at the polymorphism is very low. Trios where both parents are homozygous at the polymorphism under study need to be excluded from the statistical analysis after genotyping, making the design less cost-effective. Moreover, some simulation studies\[12\] have shown that the number of required TDT families is virtually equal to the number of cases required for similar power in case-control studies with an equal number of cases and controls. This implies that one needs to genotype 1.5 times more individuals under a TDT design to achieve similar power as a case-control design.

**The Indian context**

Recent studies\[13-15\] have shown that there is a significant genetic diversity among different ethnic groups of India. This diversity can be observed between different linguistic groups residing in different geographical locations and different caste and tribal populations. Thus, genetic case-control studies in Indian populations need to be designed with great caution. The common belief that there are genetic differences only between “North” and “South” Indian populations but not within them is contrary to the recent empirical findings. Positive case-control findings need to be validated by the use of genomic controls as described earlier or a family-based design like the TDT, failing which there is a high probability of reporting a false positive result.

**Acknowledgments**

This work was supported by the grant R01-TW-6604-4 from the Fogarty International Center of NIH.

**References**

1. Dork T, Fislage R, Neumann T, Wulf B, Tummler B. Exon 9 of the CFTR gene: splice site haplotypes and cystic fibrosis mutations. Hum Genet 1994;93:67-73.
2. Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhardt H, Cohen Z, et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat
3. Ott J. Analysis of human genetic linkage. 3rd ed. Johns Hopkins University Press: Baltimore; 1999.
4. Weir BS. Genetic data analysis II. Sinauer: Sunderland, MA; 1996.
5. Risch N, Merikangas K. The future of genetic studies of complex human. Dis Sci 1996;273:1516-7.
6. Bacanu SA, Devlin B, Roeder K. The power of genomic control. Am J Hum Genet 2000;66:1933-44.
7. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 2001;60:155-66.
8. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association in structured populations. Am J Hum Genet 2000;67:170-81.
9. Armitage P. Tests for linear trends in proportions and frequencies. Biometrics 1955;11:375-86.
10. 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.
11. Curtis D, Sham PC. A note on the application of the transmission disequilibrium test when a parent is missing. Am J Hum Genet 1995;56:811-2.
12. McGinnis R, Shifman S, Darvasi A. Power and efficiency of the TDT and case-control design for association scans. Behav Genet 2002;32:135-44.
13. Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, et al. Ethnic India: A genomic view, with special reference to peopling and structure. Gen Res 2003;13:2277-90.
14. Langstieh BT, Reddy BM, Thangaraj K, Kumar V, Singh L. Genetic diversity and relationships among the tribes of Meghalaya compared to other Indian and Continental populations. Hum Biol 2004;76:569-90.
15. Thangaraj K, Sridhar V, Kivisild T, Reddy AG, Chaubey G, Singh VK, et al. Different population histories of the Mundari- and Mon-Khmer-speaking Austro-Asiatic tribes inferred from the mtDNA 9-bp deletion/insertion polymorphism in Indian populations. Hum Genet 2005;116:507-17.

Source of Support: Fogarty International Center of NIH.
Conflict of Interest: None declared.