Hum Genet (2001) 109 : 564–565
DOI 10.1007/ s00439-001-0611-4

FOOD FOR THOUGHT

Dale R. Nyholt

Genetic case-control association studies – correcting for multiple testing

Received: 5 July 2001 / Accepted: 10 September 2001 / Published online: 19 October 2001

Sir, in a recent paper, Boehringer et al. (2000) commented on a paper by Unoki et al. (2000) who had studied 33 single nucleotide polymorphisms (SNPs) in 14 candidate genes in a search of a possible association with bronchial asthma (BA). Although Boehringer and colleagues (2000) correctly suggest a need to correct for multiple testing, their recommendations are flawed and require rectification.

First, the claim by Boehringer et al. (2000) that "association studies are not normally performed in the spirit of a statistical test, i.e., they do not involve a decision-making process", is incorrect. Fundamentally, a chi-square (χ^2) test of homogeneity is used to test hypotheses concerning the allele frequency distribution of populations (Weiss and Hassett 1986). By specifying that the magnitude of a calculated χ^2 test statistic has an associated probability of 5% or less (α=0.05), its occurrence is considered too unlikely to be attributable to random sampling alone, and it is reasonable to conclude that the null hypothesis (H_0: the population allele frequency distributions do not differ) is false (Zar 1996; Weiss and Hassett 1986). Therefore, association studies are clearly performed in the spirit of a statistical test and involve a decision-making process.

Second, basic statistical theory dictates that when multiple comparisons are performed, the probability of wrongly concluding that the two samples differ (type I error) increases. For example, Utoki et al.'s (2000) statistical analysis of 33 SNPs, yielded only one χ^2 test with a nominal significance (P=0.03) less than the conventional α of 0.05. In view of the large number of markers tested (n=33), the probability of correctly declining to reject all of them is 0.97^{33}=0.37. In other words, there is an 63% probability that such an association simply occurred by chance! Therefore, a Bonferroni-corrected α level of 0.0015 (0.05/33) should be used to give a 95% probability of correctly concluding not to reject H_0. Consequently, the claim by Boehringer et al. (2000) that there are good reasons why such "study-wide adjustments are not sensible in the context of genetic association analyses" could not be further from the truth. Moreover, Boehringer et al. (2000) cites Perneger (1998), in support of questioning the validity of correcting for such multiple testing. It would seem however, that Boehringer et al. (2000) is unaware of the somewhat scathing reply to Perneger (1998), by Bender and Lange (1999), which stated Perneger's "main arguments against multiplicity adjustments are based on misunderstanding of and a lack of knowledge about simultaneous statistical significance".

Third, Boehringer et al. (2000) suggest a correction for multple testing is required "when several (independent) markers cover a given gene region", but not when markers from different regions are tested. However, this belief is a common missconception, where, in fact, quite the opposite is true, i.e., the problem is most severe when the various tests are independent, whereas if tests are correlated, then the multiple testing problem is less critical because each new test does not provide a completely independent opportunity for a type I error (Ott 1999). Therefore, closely linked markers in linkage disequilibrium (LD) with each other may not be independent, and a Bonferroni-type correction is likely to be too conservative. To this end, nonindependence of multiple markers, covering the same region, may be examined by using the EH program (Xie and Ott 1993; Terwilliger and Ott 1994). The EH program provides a method of testing for allelic association (LD) between two markers. If significant LD exists between markers, then their disease-marker association results will be correlated and a Bonferroni-type correction will be too conservative. In this case, it may be best simply to report the disease-marker association results without correction, stating the presence of marker-marker association. For example, if all markers covering the same regions in the Utoki et al. (2000) study were in complete LD with each other, then 14 independent comparisons were performed and a standard Bonferroni-cor-

D.R. Nyholt (✉)
Queensland Instituite of Medical Research,
Post Office Royal Brisbane Hospital, Brisbane QLD 4029,
Australia
e-mail: daleN@qimr.edu.au,
Tel.: +61-7-33620258, Fax: +61-7-33620101

rected α level of 0.0036 (0.05/14) is required to give a 95% probability of correctly concluding not to reject H_0.

Other common forms of multiple testing involve multiallelic markers and sub-group analyses. In the case of multiallelic markers (e.g., microsatellites), researchers may be tempted to examine whether certain classes (alleles) are associated by comparing one allele with the remaining (grouped) allele frequencies. Although a researcher may be able to elucidate which allele to compare with the remaining alleles simply by observing the frequency data or individual χ^2 values for each data class (allele), it is not however proper to test such statistical hypotheses (without correcting for multiple comparisons) developed after examining the data to be tested (i.e., post-hoc), because essentially all possible comparisons are performed mentally. In the case of sub-group analyses, additional to analyzing the total sample, researchers stratify the sample based on sex, age, disease severity, and/or any other factor deemed suitable. However, such sub-group analyses also increase the chance of type I error and need to be corrected for appropriately.

Finally, it should be noted that significance has only been discussed in terms of multiple comparisons and type I error per report. However, because a researcher will continue to test new candidate loci for association, regardless of whether a significant association has been obtained (i.e., because multiple genes are suspected to influence most complex traits), the concept of "genome-wide" significance becomes important. To this end, more stringent genome-wide significance thresholds are required, with some researchers suggesting thresholds at least as stringent as genome-wide linkage studies (i.e., $\alpha=4.8\times10^{-5}$). However, this would be equivalent to performing over 1000 independent case-control comparisons, which most researchers will never accomplish. Therefore, perhaps a compromise lies in estimating the total number of candidate loci one may (eventually) investigate and developing an appropriate genome-wide significance threshold. However, this may not be practical, both in relation to the identification of new candidate gene families and with regard to the power of most sample sizes. Consequently, the scientific community, at the very least, should insist on correcting for multiple testing per report and (continue to) rely on replication as a means of verification.

Acknowledgement Dr. Nyholt is supported by an Australian National Health and Medical Research Council (NHMRC) Peter Doherty Postdoctoral Training Fellowship.

References

Bender R, Lange S (1999) Multiple test procedures other than Bonferroni's deserve wider use. BMJ 318:600–601

Boehringer S, Epplen JT, Krawczak M (2000) Genetic association studies of bronchial asthma – a need for Bonferroni correction? Hum Genet 107:197

Ott J (1999) Analysis of human genetic linkage, 3rd edn. Johns Hopkins University Press, Baltimore

Perneger TV (1998) What's wrong with Bonferroni adjustments. BMJ 316:1236–1238

Terwilliger JD, Ott J (1994) Handbook of genetic linkage. Johns Hopkins University Press, Baltimore

Unoki M, Furuta S, Onouchi Y, Watanabe O, Doi S, Fujiwara H, Miyatake A, Fujita K, Tamari M, Nakamura Y (2000) Association studies of 33 single nucleotide polymorphisms (SNPs) in 29 candidate genes for bronchial asthma: positive association a T924C polymorphism in the thromboxane A2 receptor gene. Hum Genet 106:440–446

Weiss NA, Hassett MJ (1986) Introductory statistics, 2nd edn. Addison-Wesley, Sydney

Xie X, Ott J (1993) Testing linkage disequilibrium between a disease gene and marker loci. Am J Hum Genet 53:1107

Zar JH (1996) Biostatistical analysis. Prentice-Hall, Upper Saddle River, NJ