European Journal of Human Genetics (2006) 14, 1238–1239

www.nature.com/ejhg

LETTER

An utter refutation of the 'Fundamental Theorem of the HapMap' by Terwilliger and Hiekkalinna

European Journal of Human Genetics (2006) **14,** 1238–1239. doi:10.1038/sj.ejhg.5201697; published online 2 August 2006

Terwilliger and Hiekkalinna[1] (T&H) have provided a thought-provoking discussion of some fundamental issues underlying the HapMap Project and its use in genome-wide association (GWA) studies with the new generation of high-density SNP genotyping technologies. They make three important points: first, that estimates of r^2 will be upwardly biased in small samples and also in case–control studies; second, that (unlike recombination fraction estimates in linkage analysis) pairwise r^2's do not necessarily have a multiplicative relationship across sets of three or more loci in the presence of three-way interactions; and third, that a causal association of a particular SNP with disease does not necessarily imply the absence of a three-way interaction involving disease, the causal locus, and some other marker locus in LD with it. In particular, they provide an insightful example of the latter situation involving genetic heterogeneity, and discuss in general terms how population stratification or gene–environment interactions could produce a similar phenomenon. Taken together, these three observations lead to their conclusion that the expected sample size required to demonstrate an association between disease and some marker in LD with a causal variant can be underestimated, possibly considerably so – their extreme example illustrates how it would be theoretically possible for an association with a marker in strong LD with a causal locus to be undetectable with even an infinite sample size.

We agree with these general principles, but question their relevance to the HapMap and its application to GWA studies. Regarding the bias in r^2 estimates, our own simulations indicate that for sample sizes typically used by HapMap and for reasonably large r^2's that are of interest as potential markers, this upwards bias is generally modest. Of greater concern is the bias in the maximum r^2 over a set of markers in a region, which could be quite substantial. As it is this maximum (pairwise or multivariate) that is typically used in one-way or another by most tag SNP selection algorithms, it is quite possible that the ability of a tag SNP panel to predict an unobserved variant in a new sample, based on estimates from small samples, could be exaggerated. Although T&H are correct that estimation of r^2 from a pooled case–control sample would be biased by the over-representation of cases, most practitioners do not in fact use this approach. More typically, study of haplotype and LD structure and the selection of tag SNPs are carried out in controls only or completely independent samples from populations similar to those under study. We have analyzed the bias in haplotype relative risk estimates that can arise from case–control samples when the population haplotype frequencies are estimated from the combined data[2] and have provided a simple correction for this ascertainment bias using an appropriate ascertainment-corrected prospective likelihood. See also Epstein and Satten[3,4] for an alternative approach based on the retrospective likelihood.

With regard to T&H's second contention, we question its relevance in the case of the association of a disease (C) with a causal variant (B) and a marker (A) in LD with it. If B were truly *the* causal variable, then we would expect A and C to be conditionally independent, given B. The bounds on the possible r^2_{AC} given r^2_{AB} and r^2_{BC} are then irrelevant, because $r^2_{AC} = r^2_{AB} r^2_{BC}$ owing to the absence of three-way interaction. T&H admit this argument at the beginning of their Discussion, but counter with their worked example illustrating what can go wrong in the case of genetic heterogeneity. In this case, B is not the *sole* cause of C, but there is another causal locus D, also in LD with A, which just happens to counteract the association in such a way as lead to no association between C and A. In this situation, we would respond that the correct analysis would be to identify additional markers that would effectively tag the real causal B–D haplotype. If we denote this expanded set of markers as A^*, then we contend that conditional on B–D, A^* and D would be conditionally independent, so again the conditions for multiplicativity would be met and the sample size inflation that would be required would be simply proportion to the inverse of the multivariate r^2 for the A^* by B–D haplotype association.

Racial heterogeneity is a well-known concern about disease association studies using unrelated individuals,[5] but can be addressed in the usual ways by matching on self-reported ethnic origins, genomic control, or use of family studies. Gene–environment interactions are an issue only if the gene and environmental factor are correlated in the source population. Although there certainly are examples

where gene–environment independence might be questionable in candidate gene association studies, this seems unlikely for most situations involving genome-wide scans with SNPs for which there is no prior hypothesis about environmental modifiers.

Going beyond the example of T&H, it is important to remember that balanced interactions between variants in their effects on risk can obscure main effects entirely, even without any added complications brought about by the 'measurement error' problem discussed by T&H (ie even if all possible variants were directly genotyped) and whether or not the markers are in LD. Does this mean that we should despair of ever finding any single-marker main effects? Of course not, but it does raise important design and analysis questions about what data-mining approaches to unearth effects involving interactions can be effectively applied at the whole genome scale, and how large sample sizes should be to have a hope of winnowing out the false from true-positive associations from such undertakings. Our contention is that it ultimately will be this question that will produce the greatest challenges in the future, as the r^2 problem is likely to be rapidly overcome by increasingly better knowledge of haplotype structure of the human genome and by increasingly more sophisticated SNP chips.

Duncan C Thomas*[,1] and Daniel O Stram[1]
[1]*Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA*

**Correspondence: Dr DC Thomas, Department of Preventive Medicine, University of Southern California, 1540 Alcazar St. CHP-220, Los Angeles, CA 90089-9011, USA.*
Tel: +1 323 442 1218; Fax: +1 323 442 2349;
E-mail: dthomas@rcf.usc.edu

References

1 Terwilliger JD, Hiekkalinna T: An utter refutation of the 'Fundamental Theorem of the HapMap'. *Eur J Hum Genet*, (2006/02/15/ online 2006; e-pub ahead of print 15 February 2006).
2 Stram DO, Pearce CL, Bretsky P *et al*: Modeling and E–M estimation of haplotype-specific relative risks from genotype data for a case–control study of unrelated individuals. *Hum Hered* 2003; **55**: 179–190.
3 Epstein MP, Satten GA: Inference on haplotype effects in case–control studies using unphased genotype data. *Am J Hum Genet* 2003; **73**: 1316–1329.
4 Satten GA, Epstein MP: Comparison of prospective and retrospective methods for haplotype inference in case–control studies. *Genet Epidemiol* 2004; **27**: 192–201.
5 Thomas DC, Witte JS: Point: population stratification: a problem for case–control studies of candidate–gene associations? *Cancer Epidemiol Biomark Prev* 2002; **11**: 505–512.