LINK TO ORIGINAL ARTICLE
LINK TO INITIAL CORRESPONDENCE

Reply: On the value of haplotype-based genotype–phenotype analysis and on data transformation in pharmacogenetics and -genomics

David J. Balding

I welcome the comments of Vormfelde and Brockmöller, which I do not view as entirely contradicting me but as going further than I had space to do in my introductory article. However, I would like to put forward several caveats on their observations. Haplotype analyses can often convey advantages over analysing one SNP at a time, particularly when, as I noted on page 787 of my Review[1], there are multiple, tightly linked functional variants acting *in cis*. However, this potential power gain is offset to some extent by the additional computational cost and increased difficulty in dealing appropriately with multiple testing. There are many possible haplotype-based analyses; no single approach has gained widespread acceptance and naive approaches can lose power because of additional degrees of freedom. Moreover, if haplotypes have been inferred from genotype data, rather than being directly measured, then a haplotype-based analysis cannot be superior to an appropriate analysis that is based on the unphased multilocus genotype data. The range of possible multipoint methods for genotype data is almost as wide as for haplotype-based methods, so it might not be easy to identify the most appropriate multipoint analysis. However, if there is only a single underlying causal variant, then a multiple regression analysis is typically preferable to a haplotype-based analysis[2]: it is easier to implement, at least as powerful, and the task of phasing the genotype data is avoided.

Vormfelde and Brockmoller are correct in pointing out that, if a phenotype mean varies linearly with allele count, then a logarithmic (or any other non-linear) transformation will disturb this linear relationship. However, we are rarely in a position to know for certain that linearity holds for the untransformed data: the transformation may well improve linearity. In their Figure 2, Vormfelde and Brockmoller illustrate an extremely strong effect of allele count on phenotypic mean. We rarely enjoy observing such strong effects, and for the modest effects that are more common in practice a logarithmic transformation is close to linear. Furthermore, if the transformed phenotype is not linear in allele count, the researcher has the option of performing an ANOVA analysis rather than linear regression, in which genotype is treated as a categorical factor rather than a linear covariate (see the legend to Figure 3 of my Review[1]). Nevertheless, I accept that a logarithmic or any other transformation should not be uncritically applied, and can affect the validity of the assumed model.

Department of Epidemiology and Public Health, Imperial College, St Marys Campus, Norfolk Place, London W2 1PG, UK.

e-mail: d.balding@imperial.ac.uk

1. Balding, D. J. A tutorial on statistical methods for population association studies. *Nature Rev. Genet.* **7**, 781–791 (2006).
2. Clayton, D., Chapman, J. & Cooper, J. The use of unphased multilocus genotype data in indirect association studies. *Genet. Epidemiol.* **27**, 415–428 (2004).