necessarily LD itself, that give conditions in which a flip of allelic effects can occur. It is important, even if estimates of LD measures are the same, to examine the distribution of haplotype frequencies in different samples with apparent flip-flop effects.

As a second case, Zaykin and Shibata consider loci in linkage equilibrium. They show how certain configurations of haplotypes penetrances can give rise to a flip-flop when there is an unobserved variant whose allele frequency varies in different populations. This results when the effects at the observed locus (A) and unobserved locus (B) interact such that the effect of A_1 may be revessed depending on whether it is on the B_1 or B_2 background. This example highlights our point that failure to account for other interacting variants can produce ambiguous association results at the observed locus under question,[1] and it shows that this can happen even without LD.

Zaykin and Shibata's study and our study have given evidence-based explanations for the controversial phenomenon of flip-flop associations. They demonstrate that failure to account for multilocus differences in samples can lead to legitimate flip-flops in a variety of scenarios. However, neither of these two studies has attempted to provide a definitive explanation for the flip-flops because such a phenomenon can stem from various reasons, ranging from genotyping errors to genomic complexity. Still, the lesson is consistent: Genomic context is important. We need to interpret associations in the context of differences in haplotype structure that occur in different populations or as a result of sample heterogeneity. Furthermore, the effect of one locus on disease risk may be inconsistent or missed completely if we fail to examine it jointly in the context of other known disease variants. These examples help to emphasize the key point that "no gene is an island."

Ping-I Lin,[1,2] Jeffery M. Vance,[3]
Margaret A. Pericak-Vance,[3] and Eden R. Martin[3,*]
[1]Department of Medicine, Program in Genetics and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA; [2]Maryland Psychiatric Research Center, Department of Psychiatry, University of Maryland School of Medicine, Catonsville, MD 21228, USA; [3]Miami Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL 33101, USA
*Correspondence: emartin1@med.miami.edu

References

1. Lin, P.I., Vance, J.M., Pericak-Vance, M.A., and Martin, E.R. (2007). No gene is an island: The flip-flop phenomenon. Am. J. Hum. Genet. *80*, 531–538.

DOI 10.1016/j.ajhg.2008.02.004. ©2008 by The American Society of Human Genetics. All rights reserved.

Optimal Two-Stage Testing for Family-Based Genome-wide Association Studies

To the Editor: A recent paper[1] in the *Journal* addressed the important issue of hypothesis testing for family-based genome-wide association studies of quantitative traits. The authors discuss the optimal use of the two sources of information (between and within[2,3]) available with family-based samples and recommend the use of a "screening" step, followed by a "testing" step.[1,4,5] By drawing an analogy with two-stage studies, in which independent samples are used rather than between and within components, we show here that statistical power is always greater with a single ("total" or "joint") test than with a "screening" approach. Furthermore, Ionita-Laza et al.[1] propose a rank-based weighting scheme for use with the "screening" approach, but such an approach fails to take into account the magnitude of the evidence for association in the between-component test. An approach based on the total test (with the between component controlled for population stratification) should provide greater power than an approach simply based on ranks.

Ionita-Laza et al.[1] focus on the "conditional power," a statistic derived from simulations that use the parental genotypes and the offspring phenotypes but not the offspring genotypes.[4,5] It is worthwhile clarifying that the "conditional power" uses the same information as the between-family test—for the between component, the parental genotypes are used for calculating a coding that summarizes the information contained in the parents. In the simplest case, association is tested by regression of offspring quantitative trait on this coding. In Abecasis et al.,[3] the coding is based on a "genotype score," where for genotype 11, 12, or 22, the genotype score is -1, 0, or 1, respectively. The between coding, b_i, where i indexes each family in the data, equals the average of the genotype score of the parents. If the parents are unknown, coding based on the offspring can be used. The within component is based on the deviation of each offspring from the between component and by construction is orthogonal (independent) to the between component. Specifically, the within coding, w_{ij}, equals $g_{ij} - \mathrm{b_i}$ where g_{ij} is the genotype score of offspring j in family i. The information used for the within-component test is the offspring phenotype and the offspring genotype conditional on the parents genotype. Programs such as QTDT[3] and PLINK[6] offer a within-only test of association, as well as a total test of association (i.e., between plus within). An explicit between-only test is offered in PLINK.

Because the between and within components are independent, the question is then how best to use these two

Table 1. Power Values for "Total" Association and "Screening" Approaches

Analysis Type	Power (%)
Total	67
Screening: 10 SNPs	42
Screening: 100 SNPs	47
Screening: 1000 SNPs	39
Screening: 10000 SNPs	24

Power values are based on 100,000 SNPs; for other parameters, see main text.

sources of information to maximize power for association detection. Skol et al.[7] addressed this in a different but analogous situation, that of two-stage association studies using unrelated individuals. Skol et al. show that, given two separate samples, the best approach is to always combine both samples and perform a "joint" test of association on the markers of interest (i.e., the SNPs followed up in the second, or stage 2, sample). Perhaps counterintuitively, this "joint" test is always more powerful than a "replication" test in which only a (typically small) subset of stage 2 SNPs is tested in the stage 2 sample. This is despite the need to correct for many tests (typically hundreds of thousands) in the "joint" analysis but only relatively few tests in the "replication" type of analysis.

The above result was derived for the case in which the stage 2 sample only had information for association testing on a subset of SNPs. The result of course holds when the stage 2 sample has information for association testing for all SNPs, as is the case for family-based tests, for which the between test (or equivalently the "screening" step based on the conditional power) is treated as stage 1 and the within test is treated as stage 2 (i.e., the approach suggested by Van Steen et al.[4]). Some simple code for the statistical package R[8] demonstrates the extent of the loss of power (Appendix A). The power of the "total" test, for a noncentrality parameter (NCP) of 30 (e.g., the approximate NCP from testing 1000 informative trios for a QTL explaining 3% of the phenotypic variance), for 100,000 SNPs (assumed to be independent), at the alpha = 5% level is 67%. A "screening" approach that selects ten (out of 100,000) SNPs from the between test (i.e., stage 1) for testing in the within stage (i.e., stage 2) has only 42% power. If both the proportion of markers and the proportion of information coming from the "between" and "within" stages are varied across the full range of possible values, the power of the "screening" approach always remains lower than that of the "total" approach. The relative amount of information coming from the "between" and "within" stages will vary depending on the exact structure of the data; for example, if the families have multiple siblings, different sibling correlations will lead to different NCPs for each stage. The effect of varying the number of SNPs included in the screening step is shown in Table 1. The above choice of 100,000 SNPs is arbitrary, and the result holds for other values. This shows that a test combining both between and within components will consistently have the best power, even taking into account the increase in multiple testing implied by not having a "screening" step. It is of course possible that for a specific instance, chance factors will lead to the "screening" approach giving a more significant result than the "total" approach (for example, if the "within" test statistic is unusually high). However, the above power calculation shows that the "total" approach will be best in the long run.

The use of both between and within components together is also advantageous because it takes into account the actual test statistics in each component rather than just (as suggested by Ionita-Laza et al.[1]) the ranks from the first (between or conditional power) stage. Following the Ionita-Laza et al. approach, the exact test statistic of the most significant SNP is not used, and this SNP receives the same weight in stage 2 irrespective of whether its test statistic was only slightly higher than the second highest SNP or whether it was the most significant by a large margin. Furthermore, because this approach focuses only on the ranks, the direction of effect is ignored; if the allele increasing the trait in the between stage is in fact conferring a decrease in the trait when it is preferentially transmitted to offspring, then the case for focusing on this SNP will clearly not be as strong as when the effect directions are concordant.

The emphasis on the within component in family-based testing is because of the potential for incorrect type I error with the between component in the presence of population stratification. In many cases, the problem of population stratification can be effectively eliminated by use of methods that compute a corrected between test. First, for large stratification effects, because data on hundreds of thousands of SNPs are now routinely available for the samples of interest, a large number of markers can be used for construction of homogenous subpopulations (e.g., with Structure[9] or PLINK[6]) and the between tests conducted within each population. For subtle structure effects, the between test within each subpopulation can also be corrected with genomic control methods. The corrected between component can then be combined with the within component to provide a "total" association test that is robust to population stratification. Using this robust "total" approach, rather than an approach that uses the between component to screen SNPs for subsequent within-component-only analysis, will provide a uniformly more powerful approach in family-based association studies.

Appendix A

In R, the power of the "total" test with the parameters given above is

```
pchisq(qchisq(0.05/100000,df = 1,lower.tail = FALSE),
df = 1,ncp = 30,lower.tail = FALSE) = 0.67,
```

where pchisq and qchisq are the distribution and quantile function, respectively, of the $\chi_1{}^2$ distribution. The qchisq(0.05/100000,1,lower.tail = FALSE) part gives the critical value to be evaluated against the noncentral $\chi_1{}^2$ distribution function for a given NCP. The power of the "screening" test is

Pr(truly associated SNPs are in top 10 out of 100 ,000)
∗ Pr(second stage SNPs are significant after correction
for ten tests) = pchisq(qchisq(10/100000,df = 1,
lower.tail = FALSE),df = 1,ncp = 15,lower.tail = FALSE)
∗ pchisq(qchisq(0.05/10,1,lower.tail = FALSE),df = 1,
ncp = 15,lower.tail = FALSE) = 0.42

If both the proportion of markers and the proportion of information coming from the "between" and "within" stages are varied across the full range of possible values (by, for example, use of two nested loops in R), the power of the "screening" approach is always lower than for the "total" approach.

Stuart Macgregor[1,*]
[1]Genetic Epidemiology, Queensland Institute of Medical Research, Brisbane 4029, Australia
*Correspondence: stuart.macgregor@qimr.edu.au

Acknowledgments

The authors thank Dale Nyholt and Peter Visscher for helpful discussions on this topic. This work was supported by grants 389892 and 496674 from the Australian National Health and Medical Research Council.

References

1. Ionita-Laza, I., McQueen, M.B., Laird, N.M., and Lange, C. (2007). Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. Am. J. Hum. Genet. *81*, 607–614.
2. Fulker, D.W., Cherny, S.S., Sham, P.C., and Hewitt, J.K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. Am. J. Hum. Genet. *64*, 259–267.
3. Abecasis, G.R., Cardon, L.R., and Cookson, W.O. (2000). A general test of association for quantitative traits in nuclear families. Am. J. Hum. Genet. *66*, 279–292.
4. Van Steen, K., McQueen, M.B., Herbert, A., Raby, B., Lyon, H., Demeo, D.L., Murphy, A., Su, J., Datta, S., Rosenow, C., et al. (2005). Genomic screening and replication using the same data set in family-based association testing. Nat. Genet. *37*, 683–691.
5. Laird, N.M., and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. Nat. Rev. Genet. *7*, 385–394.
6. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., deBakker, P.I., Daly, M.J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
7. Skol, A.D., Scott, L.J., Abecasis, G.R., and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat. Genet. *38*, 209–213.
8. R Development Core Team (2004). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).
9. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.

DOI 10.1016/j.ajhg.2008.02.003. ©2008 by The American Society of Human Genetics. All rights reserved.

Response to Macgregor

To the Editor: We appreciate the opportunity to respond to the letter by Macgregor. Macgregor claims that a total test for family-based designs should be more powerful than a two-stage design of the kind we proposed,[1,2] by drawing an analogy to the population-based scenario illustrated in Skol et al. (2006).[3] It is difficult for us to verify this statement directly because we could not find a precise definition of a "total-family" test neither in Macgregor's letter nor in any of the cited papers.

In Ionita-Laza et al. (2007),[2] we compared our testing strategies directly to pure population-based tests; these define the upper limit in terms of statistical power. However, as shown in our paper, the power differences between our weighted Bonferroni approach and the population-based test are very small; intuitively, we would expect that no test can do better than the total population-based test from an efficiency point of view. Consequently, any "total-family" test can have only marginal improvements over the strategies we proposed.

We believe that the power differences between the total test and the two-stage test shown in Macgregor's letter are overestimated for two reasons. First, as we showed in Ionita-Laza et al. (2007),[2] the weighted Bonferroni offers significant power increases over the Top k approach,[1] which is the only two-stage approach assessed in the simulation studies by Macgregor. Second, in Macgregor's simulation studies, ranking is based on p values in the first stage of the testing strategy. Van Steen et al. (2005)[1] showed that ranking based on conditional power estimates provides greater overall power than ranking based on p values. Intuitively, one expects conditional power to be a better predictor for the FBAT. Besides the genetic effect-size estimate that is based on the between-family component, ranking on conditional power also takes into account important additional information: the number of informative transmissions in the subsequent FBAT statistic. On the other hand, screening based on p values for the between-family component is purely based on the between-family component and does not incorporate any information about the number of informative transmissions, which can