Table 4. Power under Substantial and Moderate Stratification

Marker Type and Test Locus MAF	Known Strata	StratScore with 100 SNPs	StratScore with 200 SNPs	StratScore with 500 SNPs	StratScore with 800 SNPs
Highly Ancestry Informative					
0.1	0.691 (0.670)	0.67 (0.643)	0.619 (0.580)	0.403 (0.382)	0.243 (0.226)
0.25	0.914 (0.914)	0.902 (0.888)	0.871 (0.848)	0.648 (0.609)	0.412 (0.360)
0.4	0.953 (0.958)	0.940 (0.941)	0.911 (0.915)	0.702 (0.708)	0.437 (0.430)
Random					
0.1	0.678 (0.688)	0.739 (0.700)	0.650 (0.617)	0.404 (0.383)	0.230 (0.200)
0.25	0.914 (0.910)	0.932 (0.914)	0.883 (0.863)	0.634 (0.620)	0.376 (0.345)
0.4	0.959 (0.952)	0.967 (0.949)	0.937 (0.915)	0.719 (0.709)	0.430 (0.395)

Power results at nominal $\alpha = 0.05$ for 500 cases and 500 controls. The test locus has $F_{st} = 0.03$ and confers an odds ratio of 1.4 for each risk allele. Each entry shows the power under substantial stratification, followed by the power under moderate stratification in parentheses.

subpopulations. For $m = 100$ markers and substantial stratification, R^2 was ~0.19 when highly ancestry-informative markers were used, regardless of MAF, and 0.12 for random markers with $F_{st} = 0.03$. Under moderate stratification, the R^2 values were 0.07 for highly ancestry-informative markers, and 0.04 for random markers. As m increased, the R^2 values dropped even further. These relatively low values were apparently enough to provide error-control correction for the simulations reported in EAS, and other measures of correspondence than R^2 might be preferred. Nonetheless, these results further call into question the robustness of the PLS procedure, in which the stratification score does not strongly reflect the true stratification.

In summary, we conclude that aspects of the EAS method may be worthy of further exploration and development. However, in its present form, we have concerns about the routine use of StratScore, especially in the context of genome-wide scans. At the very least, the genomics community should be aware of the potential for power loss and sensitivity to the number of ancestry-informative markers employed. Additional, larger simulations in the context of whole-genome scans are necessary to provide convincing comparisons of the major approaches for controlling spurious association in case-control association studies.

Seunggeun Lee,[1] Patrick F. Sullivan,[2,3] Fei Zou,[1,3,4] and Fred A. Wright[1,3,4,*]
[1]Department of Biostatistics, [2]Department of Genetics, [3]Carolina Center for Genome Sciences, [4]Center for Environmental Bioinformatics, University of North Carolina at Chapel Hill, NC 27599, USA
*Correspondence: fwright@bios.unc.edu

Acknowledgments

The authors are supported in part by NIH grant R01 GM074175 and EPA RD-83272001. We thank the editors and reviewers for their comments.

References

1. Epstein, M.P., Allen, A.S., and Satten, G.A. (2007). A simple and improved correction for population stratification in case-control studies. Am. J. Hum. Genet. *80*, 921–930.
2. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
3. Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997–1004.
4. Abdi, H. (2003). Partial least squares regression (PLS-regression). In Encyclopedia for Research Methods for the Social Sciences, M. Lewis-Beck, A. Bryman, and T. Futing, eds. (Thousand Oaks, CA: Sage), pp. 792–795.
5. Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., Patterson, N., Gabriel, S.B., Topol, E.J., Smoller, J.W., Pato, C.N., et al. (2004). Assessing the impact of population stratification on genetic association studies. Nat. Genet. *36*, 388–393.

DOI 10.1016/j.ajhg.2007.10.014.

Response to Lee et al.

To the Editor: We thank Drs. Lee, Sullivan, Zou, and Wright (LSZW) for their letter, and for this opportunity to further discuss the use of stratification scores to control for confounding. We also take this opportunity to discuss the general question of model selection for stratification scores.

Although LSZW raise important points, we wish to start by objecting to their characterization of the stratification score as the output of partial least-squares regression (PLS). The stratification score defined by Epstein et al.[1] (EAS) is simply a model for $P[D|Z]$ where Z are markers (or potentially other covariates) used to control for confounding by population stratification and D is an indicator of disease status. We used a particular PLS-based procedure

for our calculations, but we stressed that any model, such as logistic regression or even random forests, can be used to calculate the stratification score.

With this in mind, we address criteria for determining what is a "good" model choice for a stratification score. As LSZW correctly demonstrate, prediction of D cannot be the goal, because the "best" model by this criterion would provide near-perfect prediction of D; poststratification use of such a score will result in most strata having only cases or only controls. This results in a loss of power when association is assessed because up to four-fifths of observations are in (nearly) uninformative strata. This raises the question: What is a good model for the stratification score?

Some guidance comes from distinguishing between population stratification and confounding by population stratification. Population stratification occurs whenever there is variation in allele frequencies that is explained by (typically unmeasured) covariates U. Confounding only occurs when U also accounts for some of the variability in D. Unfortunately, this means that a "good" stratification-score model is one that accounts for the variability in D that is caused by U, but not for any of the residual variation in D. Without knowing U, it is difficult to determine what variability in D the stratification score should explain. However, one clue is that variables used in the stratification-score model should explain variation in both D and the test-locus genotype G. Thus, we seek a stratification score that is a linear combination of marker genotypes Z and explains variability in both D and G.

After evaluating a wide range of possible stratification scores in simulated data, we propose the following approach: Use both D and G as the dependent variables in a PLS model (PLS allows multivariate dependent variables), and then use the first PLS component as the stratification score. We first confirmed that this proposal preserves size by using the simulated data from our original paper (results not shown). To evaluate the power of this proposal, as well as the effect of changing the number of markers, we simulated data from the following model. We assumed that genotype frequencies for both substructure markers (Z) and a test locus (G) were influenced by two continuous axes. Specifically, if A_{jk} is the maternal ($k = 1$) or paternal ($k = 2$) allele at the jth marker locus, we assumed that the probability of a "1" allele was given by

$$logit\{P[A_{jk} = 1 \mid r_1, r_2]\} = \gamma_{0j} + \gamma_1 r_{1j} + 0.2 \cdot r_{2j},$$
$$j = 1, \ldots, M \text{ and } k = 1, 2$$

with marker genotype at the jth locus given by $Z_j = A_{j1} + A_{j2}$, and where r_{1j} and r_{2j} are independent standard normal random variables. The values of γ_{0j} were chosen to mimic allele frequencies in the data of Akey et al.[2] We generated genotypes at a trait locus by using the same model, with $\gamma_0 = -0.4$ corresponding to a baseline minor-allele frequency of about 0.40. We then prospectively generated disease outcome for participants by using the model

$$ln \frac{\Pr[D = 1 \mid r_1, r_2]}{\Pr[D = 0 \mid r_1, r_2]} = -4.6 + 0 \cdot r_1 + 0.2 \cdot r_2 + ln(1.2)G,$$

which corresponds to a baseline disease prevalence of ~0.01. Notice that r_2 is a confounder but r_1 is not within our simulation model.

We simulated disease and marker data (assuming either 100 markers or 500 markers) by using $\gamma_1 = 1$, 1.5, 2 and generated data until 1000 case and 1000 control participants had been recruited. We then analyzed the data by using our original EAS approach, our joint (D, G) approach, principal components,[3] and a gold standard corresponding to the situation in which we knew the true r_2 confounder and adjusted for it appropriately within analysis. We repeated this procedure 1000 times. The estimated power results are given in Table 1.

Table 1. Power under the Stratification Model

	100 AIMs (γ_1)			500 AIMs (γ_1)		
Analysis	1.0	1.5	2.0	1.0	1.5	2.0
EAS	34.4	22.3	12.0	12.7	8.0	7.0
joint (D, G) model	48.5	43.1	37.7	43.0	43.4	37.6
Principal Components	47.7	40.4	37.3	41.9	41.9	35.7
analysis conditional on r_2	64.3	64.3	58.5	56.4	62.3	68.8

Estimated power at size $\alpha = 0.05$ for 1000 datasets generated with our simulation model for four analyses: the original stratification score of EAS, the new stratification score proposed here, principal components, and a logistic regression that conditions on the true confounder r_2.

The table shows that use of the first PLS component from a joint model for D and G clearly outperforms the stratification score used previously in EAS. In addition, the performance of the joint (D, G) approach does not degrade as the number of markers increase (some decrease in performance between 100 ancestry-informative markers (AIMs) and 500 AIMs is shared by all methods, even the approach that conditions on the usually unknown true confounder r_2). Although our approach slightly outperforms principal components, this difference is slight.

Finally, we confirmed that the stratification score obtained as the first PLS component in the joint (D, G) model controlled confounding in the association between height and the LCT (MIM 603202) single-nucleotide polymorphism (SNP) rs4988235 reported by Campbell et al.[4] We found that the p value for this analysis was 0.28, which compares favorably with that obtained in EAS. Recall that the p value obtained when principal components were used was 0.003.

We close with two discussion points. First, we have confined our discussion here to AIMs and have not considered random markers. It seems reasonable to us that, compared with AIMS, random markers would be more likely to explain variation in D that was not due to confounding. Thus, their use may be a threat to efficiency. Second, examination of Table 1 shows clearly that all currently available methods fall far short of the power available when the

true confounder r_2 is known. This indicates to us that there is much additional work that can be done to investigate the question of model selection for the stratification score.

Michael P. Epstein,[1,*] Andrew S. Allen,[2] and Glen A. Satten[3]
[1]Department of Human Genetics, Emory University, Atlanta, GA 30322, USA; [2]Department of Biostatistics and Bioinformatics and Duke Clinical Research Institute, Duke University, Durham, NC 27708, USA; [3]Centers for Disease Control and Prevention, Atlanta, GA 30341, USA
*Correspondence: mepstein@genetics.emory.edu

Acknowledgments

Any opinions expressed in this paper are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

References

1. Epstein, M.P., Allen, A.S., and Satten, G.A. (2007). A simple and improved correction for population stratification in case-control studies. Am. J. Hum. Genet. *80*, 921–930.
2. Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Res. *12*, 1805–1814.
3. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
4. Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., Altshuler, D., Ardlie, K.G., and Hirschhorn, J.N. (2005). Demonstrating stratification in a European American population. Nat. Genet. *37*, 868–872.

DOI 10.1016/j.ajhg.2007.11.010.

XMCPDT Does Have Correct Type I Error Rates

To the Editor: In the January 2007 issue of the *Journal*, Chung et al.[1] compared X-APL proposed by them to XMCPDT proposed by Ding et al.[2] Based on their simulation results, they stated that with use of allele frequencies estimated from observed parental genotypes, XMCPDT would give inflated type I error rates. Here we wish to point out that use of estimated allele frequencies is not the cause of inflated type I error rates. Rather, the actual cause was the severe violation of the XMCPDT assumption in their simulation settings, which was discussed at length in Ding et al.[2] As explicitly stated there, one assumption for XMCPDT to be a valid test for association under linkage is that "the pedigrees in a study are assumed to be drawn from a population of (extended) families, each of which has at least one affected offspring." They went on to say, "Otherwise, bias may exist, especially when all families have the same structure and affection pattern, which, fortunately, is not the case in a genetic study that collects pedigrees of all shapes and sizes and affection patterns." To study the robustness of the test statistic to departure from the assumption, Ding et al.[2] investigated trios as well as families with six children and concluded that "in a genetic study with pedigree data, bias should be negligible, and the proposed test statistic may be safely used." However, the simulation settings in Chung et al.,[1] which fixed the affection statuses of the offspring, severely violated the assumption, leading to appreciable bias.

A fuller dissection of the assumption of Ding et al.[2] is needed in order to facilitate understanding of why the settings in Chung et al.[1] constitute severe violations. The sampling assumption treats affection status of a given family structure as a random event, and as such, all sorts of affection patterns are permitted. For example, for nuclear families with three children (a setting in Table 4 of Chung et al.[1]), under the assumption, one would expect some families having one, some having two, and some having all three children being affected. However, Chung et al.[1] only allow exactly two of the three children in each of the nuclear families to be affected, thus severely violating the assumption. Such a restriction on the affection status appears to be rather unrealistic in a genetic epidemiological study, as it is unlikely that a family with three children would only be included in the study if exactly two of the three children were affected. With inclusion of one-affected and three-affected families, the power is expected to increase substantially. More importantly, as demonstrated below through simulations, it is in fact X-APL that gave inflated type I error rates when the XMCPDT assumption was roughly satisfied, especially when data from extended families were included.

Our first simulation setting made use of the same family structure, discussed above, as that of Chung et al.,[1] but ours allowed for one-affected and three-affected families to be included in addition to the two-affected ones. One hundred nuclear families, each with two parents and three offspring, were simulated in each replicate. Among those 100 families, 25 had three male offspring, 25 had two male and one female offspring, 25 had one male and two female offspring, and the remaining 25 had three female offspring. Furthermore, parents in 50 of the families had observed genotypes, and those in the other 50 families did not. The disease models were the same as those in Table 1 of Chung et al.[1] For each of the four family types, we simulated the data until we had 25 families, each with at least one affected offspring. The disease locus was used to calculate powers. In addition to the disease locus, a marker with the same allele frequencies and in complete linkage and linkage equilibrium was also simulated and used to calculate type I error rates. The second simulation setting had