Comment on a Simple and Improved Correction for Population Stratification

To the Editor: In the May 2007 issue of the *American Journal of Human Genetics*, Epstein, Allen, and Satten[1] (hereafter referred to as EAS) introduced a new method for controlling population stratification in case-control association studies. The method computes a stratification score by performing partial least-squares regression (PLS) of phenotypes (case-control status) on a matrix of genotypes at markers used to correct for ancestry. The quantitative stratification score is then used to divide subjects into a number of strata, so that a stratified test of case-control association may be performed at any test locus not in linkage disequilibrium with the ancestry-informative markers. The stratification and testing procedure are implemented in the program StratScore, available as SAS code from the authors.

EAS described a retrospective case-control model involving the latent true stratification variable and provided practical recommendations for dividing the estimated stratification score into a number of strata. The PLS procedure, however, was presented in less detail, although it is key to the performance of the overall approach. A primary motivation was the claim that stratified analysis based on principal components[2] or genomic control[3] cannot fully control for population ancestry. The authors cited an example and provided simulations in which stratification resulted in inflated type I errors when using these methods for 100 ancestry-informative markers. An immediate concern is whether these results reflect current practice—in a modern whole-genome scan, hundreds of thousands of markers are available for ancestry control. The results of Price et al.[2] suggest that, with the availability of thousands of markers, principal components do provide effective ancestry control, and indeed a large number of markers may be necessary for correcting stratification within continental-level populations.[5] Moreover, the use of principal components does not require predefined ancestry-informative markers and thus may potentially control for unanticipated strata, including technical phenomena unrelated to ancestry.[2] In terms of statistical power, the principal-components-based approach appeared to fare quite well in EAS.[1]

To better understand the issues and how the EAS approach might be best applied, we examined the PLS procedure more closely. Here, PLS finds linear combinations T of the matrix of ancestry-informative markers X such that the covariance between phenotypes Y and T is maximized (see [4] for details on partial least-squares regression). Predictions of case status from a logistic-regression model (Y on T) are then used as the stratification score. A risk of PLS is the potential for finding spurious relationships, although EAS employed a variable selection technique to control the number of T variables used. If spurious apparent stratification arises from PLS, it has the potential to greatly reduce statistical power because the stratification variable could account for phenotype variation caused by a true disease gene. Moreover, although the inclusion of a large number of ancestry-informative markers should be desirable for ancestry prediction, the resulting increased flexibility in the PLS factors might produce even stronger spurious stratification, thereby resulting in decreased power as the number of such markers increases.

To further investigate the utility of StratScore and to test our predictions about the method, we performed simulations under no stratification, for random unlinked markers with minor allele frequencies (MAF) ranging uniformly from 0.1 to 0.5. Table 1 shows the results from representative simulations analyzed by StratScore, with $m = 100$, 200, 500, and 800 markers used to infer ancestry. Note that the stratification score has a very high correlation with case-control status, although no true correlation exists between the markers and phenotype because no stratification exists. As the number of markers increases, the spurious correlation increases, and the case-control numbers for many of the strata become highly imbalanced. Such strata cannot meaningfully contribute to detection of case-control association.

We further performed simulations of case-control association data, by following the conditions and terminology described in EAS. For each setup, 5000 simulations were

Table 1. Illustrative Simulations of Case-Control Status versus StratScore Inferred Strata

Number of Markers and Case Status	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5	Total
$m = 100$						
Case	65	80	110	106	139	500
Control	135	120	90	94	61	500
Total	200	200	200	200	200	1000
$m = 200$						
Case	46	70	105	124	155	500
Control	154	130	95	76	45	500
Total	200	200	200	200	200	1000
$m = 500$						
Case	5	41	103	157	194	500
Control	195	159	97	43	6	500
Total	200	200	200	200	200	1000
$m = 800$						
Case	0	1	100	199	200	500
Control	200	199	100	1	0	500
Total	200	200	200	200	200	1000

This table shows case-control status versus Stratscore inferred strata, based on 500 cases and 500 controls. m is the number of markers used for computation of stratification score.

Table 2. Type I Error under Substantial and Moderate Stratification

Marker Type and Test Locus MAF	No Adjustment	Known Strata	StratScore with 100 SNPs	StratScore with 200 SNPs	StratScore with 500 SNPs	StratScore with 800 SNPs
Highly Ancestry Informative						
0.1	0.155 (0.079)	0.051 (0.047)	0.049 (0.046)	0.049 (0.047)	0.042 (0.046)	0.057 (0.058)
0.25	0.220 (0.097)	0.051 (0.047)	0.051 (0.041)	0.048 (0.039)	0.046 (0.041)	0.056 (0.057)
0.4	0.178 (0.090)	0.053 (0.053)	0.046 (0.049)	0.049 (0.048)	0.048 (0.045)	0.054 (0.057)
Random						
0.1	0.160 (0.085)	0.049 (0.055)	0.050 (0.057)	0.048 (0.055)	0.041 (0.046)	0.054 (0.052)
0.25	0.223 (0.097)	0.047 (0.045)	0.059 (0.048)	0.049 (0.043)	0.049 (0.040)	0.059 (0.049)
0.4	0.166 (0.089)	0.054 (0.047)	0.059 (0.053)	0.047 (0.045)	0.044 (0.044)	0.047 (0.049)

Type I error results at nominal $\alpha = 0.05$ for 500 cases and 500 controls, when a test locus with $F_{st} = 0.03$ is used. Each entry shows the type I error under substantial stratification, followed by the type I error under moderate stratification in parentheses. Simulation conditions are described in the text.

performed for 500 cases and 500 controls, with three underlying populations of equal size. We simulated substantial stratification by sampling cases in the proportions 0.45, 0.33, and 0.22 from subpopulations 1, 2, and 3. Moderate stratification was achieved by sampling in the proportions 0.40, 0.33, and 0.27. The alternative hypothesis was simulated with odds of disease increasing by a factor 1.4 for each copy of the risk allele for the test locus, which had F_{st} values of 0.03 and 0.15 in various simulation setups. EAS simulated ancestry markers on the basis of F_{st} selection criteria applied to SNPs from a real data set. To reproduce their results and to better control the simulation conditions, we simulated marker SNPs following the method in Price et al.[1] For each of MAF values 0.1, 0.25, and 0.4, sets of random marker SNPs were simulated with $F_{st} = 0.03$, and highly ancestry-informative markers with F_{st} values were drawn uniformly from 0.5 to 0.8. Although EAS reported results for sets of $m = 100$ ancestry markers,[1] we also performed simulations for sets of $m = 200$, 500, and 800 markers.

With a significance threshold of $\alpha = 0.05$ and a test locus with $F_{st} = 0.03$, we found approximately correct type I error control by using the StratScore approach for all choices of m markers (Table 2, effectively an expanded version of Tables 2 and 3 in EAS). However, when the test locus had $F_{st} = 0.15$, we found type I errors ranging from 0.02 to 0.098 (Table 3), depending on the ancestry marker setup and degree of stratification. EAS had reported correct StratScore error control for some of these same setups (see Table 4 in EAS). We are unsure of the reason for the discrepancy, although minor variation in generalized PLS[1] versus the standard PLS implemented in StratScore is a possibility. To investigate whether the results might be specific to our use of the simulation approach of Price et al.[1] (beta sampling of minor allele frequencies, followed by rejection sampling of F_{st} values), we also employed a deterministic approach. We set allele frequencies for the three populations (order determined randomly) as p/a, p, and pa, where a and p were determined to achieve specified F_{st} and MAF values. Our conclusions under this scheme were unchanged. Although our main focus is on the power of StratScore, these results suggest a lack of robustness that may be problematic in StratScore error control and deserves further inquiry.

We next investigated power for StratScore as the number of markers increases. Table 4 presents the power under the alternative hypothesis for Cochran Mantel Haenszel (CMH) tests under moderate and substantial true stratification. Here, the best-case scenario of known strata is compared to the StratScore approach for various numbers of ancestry markers. As predicted, the power drops dramatically as the number of ancestry markers increases, thereby restricting the number of markers that can be used. Note that this restriction depends in an essential way on the case-control sample size. Studies in which the true stratification is subtle may require a larger number of markers for ancestry control and therefore limit the utility of StratScore.

Another aspect of EAS that was unclear was the degree of correspondence between the stratification score and the true subpopulations. For the alternative-hypothesis simulation setups, we computed average ANOVA R^2 values for the stratification score versus the three true

Table 3. Type I Error, Test Locus $F_{st} = 0.15$

Marker Type and Test Locus MAF	No Adjustment	Known Strata	StratScore with 100 SNPs
Highly Ancestry Informative			
0.1	0.433 (0.150)	0.051 (0.053)	0.040 (0.028)
0.25	0.751 (0.264)	0.046 (0.050)	0.020 (0.020)
0.4	0.757 (0.270)	0.051 (0.049)	0.028 (0.024)
Random			
0.1	0.446 (0.155)	0.050 (0.048)	0.078 (0.049)
0.25	0.759 (0.271)	0.053 (0.049)	0.096 (0.054)
0.4	0.757 (0.267)	0.048 (0.050)	0.098 (0.051)

Type I error results at nominal $\alpha = 0.05$ for 500 cases and 500 controls, when a test locus with $F_{st} = 0.15$ is used. Each entry shows the type I error under substantial stratification, followed by the type I error under moderate stratification in parentheses.

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