Mixed models can correct for population structure for genomic regions under selection

Jae Hoon Sul and Eleazar Eskin

An article in this journal by Price *et al.* (New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459–463 (2010))¹ showed by simulations that mixed models² may be susceptible to spurious associations on markers with unusual allele frequency differences between populations, such as markers in regions under selection. They stated that the reason for the spurious associations or inflation of test statistics is because mixed models model population structure as a random effect, although it is a fixed effect.

After investigating this problem further, we found that modelling population structure as a random effect is not the cause of inflation and that it is a kinship matrix that determines the performance of mixed models. The kinship matrix defines pairwise genetic relatedness among individuals and is usually estimated by using all genotyped markers. Because most markers in the simulations carried out by Price *et al.*¹ have small allele frequency differences between two populations, a kinship matrix estimated from all markers does not effectively capture the population structure. However, when a kinship

matrix is computed from the first principal component or from only the unusually differentiated markers (UDMs), we find that mixed models achieve almost the same performance as EIGENSTRAT³, which is a method that incorporates population structure as a fixed effect (TABLE 1). A kinship matrix that captures the same information as the first principal component vector can be obtained by computing the outer product of the vector: $\mathbf{w}\mathbf{w}^T$, where $\mathbf{w}$ is the first principal component vector. Use of this matrix in mixed models is closely related to including the first principal component as a covariate in EIGENSTRAT. As for the kinship matrix generated from UDMs, because UDMs are not known in advance, one may try to detect these using methods to identify markers under selection, one of which was developed by our group. This method, which is called spatial ancestry analysis (SPA)⁴, correctly identifies UDMs in the simulations, and mixed models with kinship estimated from those markers detected by SPA have almost the same inflation as kinship from the true UDMs.

Although the approach of using kinship matrices from UDMs is effective in capturing broad differences among individuals, it

may not capture narrow sample structure, such as family structure. One approach to solving this problem is to include an additional kinship matrix estimated from markers other than UDMs. This means that we have two kinship matrices in mixed models: one that is computed from UDMs and the other that is computed from the rest of markers. This would effectively remove inflation by population structure and other sample structure. We apply this approach to the simulations and show that this approach removes inflation on UDMs (TABLE 1).

We also investigated whether UDMs cause the inflation of statistics in real genome-wide association studies (from the 1966 Northern Finland Birth Cohort (NFBBC66)⁵ and the Wellcome Trust Case Control Consortium (WTCCC)⁶). We observed inflation on a few phenotypes, but no inflation was statistically significant (data not shown).

In summary, mixed models are equivalent to methods that consider population structure as a fixed effect when the appropriate kinship matrix is applied. Mixed models can easily be extended to correct for inflation caused by UDMs, although our results failed to identify a case in which the phenomenon reported in Price *et al.* occurs in practice.

Jae Hoon Sul is at the Computer Science Department, University of California, Los Angeles, California 90095, USA.

Eleazar Eskin is at the Department of Human Genetics, University of California, Los Angeles, California 90095, USA.

Correspondence to E.E.
e-mail: eeskin@cs.ucla.edu

doi:10.1038/nrg2813-c1
Published online 26 February 2013

Table 1 | Genomic control inflation factor of mixed models with various kinship matrices

	Normally differentiated makers		Unusually differentiated markers	
	Genomic control (λ)	Correlation to EIGENSTRAT	Genomic control (λ)	Correlation to EIGENSTRAT
Uncorrected	1.396239	0.6933167	48.59404	0.1063447
EIGENSTRAT	1.000579	1	1.001239	1
MM (IBS)	0.9981917	0.9914998	1.534722	0.6763116
MM (first PC)	0.9992498	0.9997281	1.001937	0.9713147
MM (UDMs)	0.9995067	0.998002	1.015939	0.9270086
MM (SPA)	0.9995421	0.998009	1.017233	0.9257633
MM (UDMs and other)	0.9973654	0.9971357	0.9996444	0.9397491
MM (SPA and other)	0.9973461	0.9971259	0.9999022	0.9382122

We computed the genomic control parameters, λ , on two sets of markers: a group of normally differentiated markers ($F_{ST} = 0.01$) and a group of unusually differentiated markers (UDMs), the allele frequency difference between two populations of which is 0.6. We also computed the correlation between χ^2 statistics from EIGENSTRAT and χ^2 statistics from mixed models and uncorrected analysis. MM (IBS) involves mixed models with an identity by state of kinship matrix using all makers. MM (first PC) uses kinship estimated from the first principal component of principal component analysis (PCA). MM (UDMs) uses kinship estimated only from unusually differentiated markers (UDMs), and MM (SPA) uses kinship estimated from markers with spatial ancestry analysis (SPA) scores that are 3 standard deviations from the average SPA scores. MM (UDMs and other) and MM (SPA and other) use two kinship matrices: one estimated either from UDMs or SPA, and the other estimated from all other markers.

1. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Rev. Genet.* **11**, 459–463 (2010).
2. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genet.* **42**, 348–354 (2010).
3. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–909 (2006).
4. Yang, W.-Y. Y., Novembre, J., Eskin, E. & Halperin, E. A model-based approach for analysis of spatial structure in genetic data. *Nature Genet.* **44**, 725–731 (2012).
5. Sabatti, C. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genet.* **41**, 35–46 (2008).
6. WTCC Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

Acknowledgements

J.H.S. and E.E. are supported by US National Science Foundation grants 0513612, 0731455, 0729049, 0916676 and 1065276, and US National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568 and P01-HL28481.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Eleazar Eskin's homepage: <http://www.cs.ucla.edu/~eeskin>

ALL LINKS ARE ACTIVE IN THE ONLINE PDF