Genomic Control to the extreme

To the editor:
In the study of complex disease, separating causal from confounded factors is a challenge for genetic epidemiologists. One tool useful for separating these factors is Genomic Control (GC). In this communication we clarify how and when to use GC. We also describe a refined approach to GC, which should be used when GC is applied to extreme settings.

Population-based studies, such as case-control studies, are common designs used to determine the genetic and environmental bases of disease. To avoid false positive associations, design and analysis of population-based studies should account for population stratification, which can inflate association test statistics. One analytic method used to control the false positive rate is GC. In our original paper[1], we investigated two scenarios with two corresponding analytic methods. GC is the version similar to the typical approach to hypothesis testing, and GCB is the version that uses Bayesian inference. GC is suitable when a modest number of candidate genes are assessed and L supplementary loci are included for control. The supplementary loci, called null loci, are used to correct any inflation, λ, in association test statistic(s) by estimating λ from the null test statistics. GC produces average rejection rates close to the targeted 0.05 significance level[2–4].

We also considered population-based studies when large numbers of markers are tested[1]. GCB is designed for this scenario. Rather than preselecting null loci, GCB delineates loci associated with disease as 'outliers' relative to most of the loci tested.

In our original papers[1,2], we argued that a population-based study should attempt to remove the effect of stratification by experimental design and analysis, such as by matching cases and controls for ethnicity and environmental covariates. GC then adjusts for the residual effects of stratification. Careful study design and implementation pay off in statistical power[5,6]; even small stratification can have considerable consequences for large samples[1–5].

Marchini *et al.*[7] explored the efficacy of GC for population-based studies under less ideal conditions, using subjects that originate from different populations and including environmental effects that induce geographically distinct prevalences; both of these possibilities were ignored in the design and analysis. Because they genotyped a large number of loci, they required an extremely small significance level (α) for P values. They found that GC could be anticonservative when α is small. Their results are sensible because GC treats λ as a known constant[8]. For small values of α, variability in the estimate of λ matters. The population-based studies explored by Marchini *et al.*[7] can produce highly inflated test statistics (**Fig. 1**), and, because these population-based studies involve a large number of candidate loci, they are more

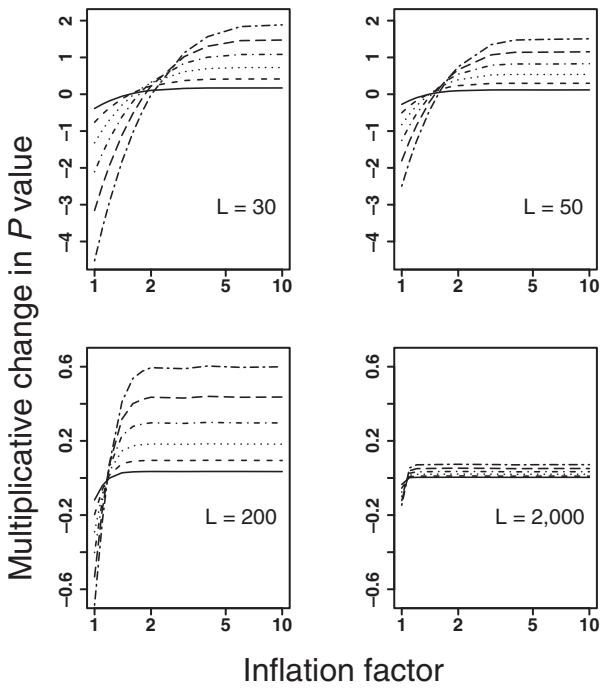

Figure 1 Performance of GC as a function of the targeted significant P value (α), the effect of stratification (λ) and the number of null loci included (L). For the solid line, $\alpha = 10^{-2}$; at $\lambda = 10$, α decreases by an order of magnitude for each consecutive line thereafter. Note the different scales in the top panels versus the bottom panels. Marchini *et al.*[7] generated their data by using a beta-binomial model. We avoided generating individual loci by working with a summary statistic for the values, thereby obtaining a good approximation to their simulations. The tests are distributed as a scaled χ^2 statistic, $\lambda\chi^2$. A sketch of our procedure, for a single choice of λ, α and L, requires several steps: generate L copies of x, each x distributed χ^2, and multiply each x by λ; use GC to compute λ_ε; draw another random realization x from a χ^2 and compute the GC test statistic as $y = \lambda x/\lambda_\varepsilon$. Carrying out these steps many times produces p_m, the expected fraction of times the P value exceeds α for a given λ and L. Then $\log_{10}(p_m/\alpha)$ is calculated. Carrying out this procedure for a large number of settings for λ, α and L produces these results, which capture the essence of the results that Marchini *et al.*[7] obtained by using their simulation techniques. For $n = 1{,}000$, models A1, A2 and B1–B5 of Marchini *et al.*[7] are inflated by $\lambda \approx 18.8$, 11.0, 1.1, 1.2, 1.7, 1.6 and 4.1, respectively. See **Supplementary Note** online for more information.

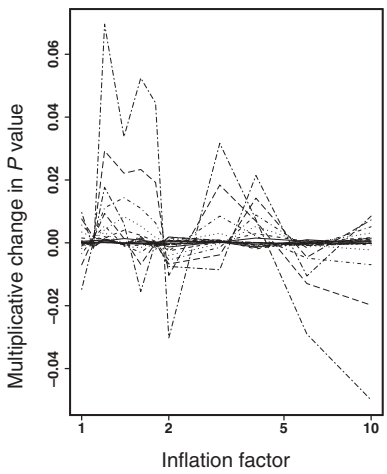

Figure 2 Performance of GCF as a function of α, λ_m and L. Simulations were done as described for **Figure 1**, with two exceptions. Instead of using the robust estimate for λ, λ_e, we used the mean λ_m. And instead of determining the P value of y from a χ^2 distribution, y was assumed to be distributed according to F(1,L) and the P values were calculated from that distribution. Note the compressed vertical scale (relative to **Fig. 1**), reflecting the miniscule error for all settings. The greatest error was observed for L = 30 and $\alpha = 10^{-7}$. See **Supplementary Note** online for more information.

appropriately analyzed by using GCB rather than GC.

Because λ is determined by sample size, stratification and differential prevalence[1], we can generate and compactly represent the general findings of Marchini *et al.*[7] (**Fig. 1**). Four features stand out from our results: (i) GC works well in situations for which it was originally intended, namely larger values of α and/or smaller values of λ (refs. 1–4); (ii) GC becomes increasingly anticonservative as α decreases or as λ increases; (iii) bias is also a function of L (refs. 1,2); and (iv) even minor stratification can have a substantial impact on population-based studies with large sample sizes[1,7,9]. Of these results, the second feature is new to Marchini *et al.*[7]; the fourth was shown mathematically[1] before it was demonstrated empirically[7,9].

Is there a way to adjust the procedure if a researcher wishes to apply the logic of GC and use an extremely small α value?

Correcting the bias in GC is straightforward by simple modification of the test statistic (GCF). For GCF, estimate λ using the mean (λ_m) of the null test statistics and account for the variability of λ_m by using an F test to determine the P values. Notably, GCF is accurate throughout the parameter space, even for only 30 null loci (**Fig. 2**).

Our means of validating the results of Marchini *et al.*[7] and our own results use a shortcut method that Marchini *et al.*[7] did not use. Our results are also supported by using the simulation methods of Marchini *et al.*[7]. When we used their methods and analyzed the data using GCF, we again found that GCF yielded an excellent approximation for small values of α (**Table 1**), even when λ is inflated substantially by large sample size or geographically distinct prevalences.

In summary, when a large number of candidate loci are genotyped or when α is small, application of GC produces misleading results (**Fig. 1**), as Marchini *et al.*[7] show. Because GCF corrects this bias for small values of α, and does so in a range of settings (**Fig. 2** and **Table 1**), we conclude that the bias is largely due to the uncertainty in λ. GCF accounts for this uncertainty in its degrees of freedom. Thus, GCF provides a simple alternative to recently suggested methods based on the confidence interval for λ (ref. 9).

As we have pointed out before[1,10], for experiments involving a large number of tests of genetic markers, one should analyze the entire distribution of test statistics. In this setting different statistical paradigms should be considered, such as methods based on the false discovery rate principle[11], which has great promise for this setting (refs. 10,12 and S.-A.B., B.D., K.R. and L. Wasserman, unpublished data).

Note: Supplementary information is available on the Nature Genetics website.

B Devlin[1], Silviu-Alin Bacanu[1] & Kathryn Roeder[2]

[1]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA. [2]Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, USA. Correspondence should be addressed to B.D. (devlinbj@upmc.edu).

1. Devlin, B. & Roeder, K. *Biometrics* **55**, 997–1004 (1999).
2. Bacanu S.-A., Devlin, B. & Roeder K. *Am. J. Hum. Genet.* **66**, 1933–1944 (2000).
3. Devlin, B., Roeder, K. & Bacanu S.A. *Genet. Epidemiol.* **21**, 273–284 (2001).
4. Pritchard, J.K. & Donnelly, P. *Theor. Pop. Biol.* **60**, 227–237 (2001).
5. Lee, W.-C. *Genet. Epidemiol.* **27**, 1–13 (2004).
6. Hinds, D.A. *et al. Am. J. Hum. Genet.* **74**, 317–325 (2004).
7. Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly P. *Nat. Genet.* **36**, 512–728 (2004).
8. Reich, D.E. & Goldstein, D.B. *Genet. Epidemiol.* **20**, 4–16 (2001).
9. Freedman, M.L. *et al. Nat. Genet.* **36**, 388–393 (2004).
10. Tzeng, J-Y., Byerley, W., Devlin, B., Roeder, K. & Wasserman, L. *J. Amer. Statist. Assoc.* **98**, 236–247 (2003).
11. Benjamini, Y. & Hochberg, Y. *J. R. Statist. Soc. B* **57**, 289–300 (1995).
12. Devlin, B., Roeder, K. & Wasserman L. *Genet. Epidemiol.* **25**, 36–47 (2003).

Table 1 Targeted significance levels of GCF compared with the realized values produced by simulations using a beta-binomial model

			Targeted significance level[b]				
RR[a]	L	n	0.050	0.010	0.0010	0.00010	0.000010
1:2	100	1000	0.050	0.0097	0.000092	0.000084	0.0000076
1:19	100	1000	0.050	0.0098	0.00099	0.00010	0.000010
1:2	100	10,000	0.049	0.0092	0.00083	0.000072	0.0000062
1:19	100	10,000	0.049	0.010	0.0012	0.00014	0.000018
1:2	1,000	1,000	0.050	0.0097	0.00091	0.000083	0.0000073
1:19	1,000	1,000	0.049	0.0095	0.00085	0.000071	0.0000060
1:2	1,000	10,000	0.050	0.0098	0.00091	0.000082	0.0000071
1:19	1,000	10,000	0.049	0.010	0.0012	.000015	0.000019
Average			0.0495	0.0098	0.00099	0.000010	0.000011

[a]For RR = 1:2 and n = 1,000 (10,000), λ = 1.7 (7.8). For RR = 1:19 and n = 1,000 (10,000), λ = 5.9 (49.8). [b]The beta-binomial model was used to generate data for single-nucleotide polymorphisms drawn from different populations, with structure identical to that measured by Marchini *et al.*[7] from the Chinese and Japanese samples. n cases are sampled from these simulated populations according to the relative risk (RR) specified, n controls are sampled at random and then a test statistic is generated. The procedure is repeated many times. The fraction of statistics exceeding the targeted level, determined by the F-distribution ($F_{1,L}$), is then the realized significance level for the beta-binomial model. See **Supplementary Note** online for more information.

In reply:

The main point of our original paper[1] was that even the relatively small levels of structure in large populations cannot be ignored in the coming generation of association studies, effectively because of the sizes of these studies (both sample size and numbers of loci). We continue to believe, however, that association studies have a central role in unraveling the genetic basis of common human diseases, provided that population structure is handled appropriately. One published method for dealing with population structure is Genomic Control (GC)[2]. Our paper showed that GC typically performs well but that there are some previously unrecognized problems in certain settings.

We are delighted that our work prompted Devlin and his colleagues to correct this aspect of GC. Their new procedure, GCF, represents an important advance and should be used in place of the original method. We also agree that this approach to handling uncertainty in the estimation of the correction factor λ is better than the use of confidence limits[3].

But whether the settings in which GC had problems should be dismissed as 'extreme' is less clear. Of course the design and analysis of studies should attempt to control for stratification. This is not simple to do in practice. First, there are important unresolved empirical questions about the levels and nature of such structure in population groups (*e.g.*, people of European descent in a particular country or African Americans) and unresolved statistical issues about how best to use this kind of information in study design and analysis. Second, in the real world many studies will not meet these worthy objectives, in some cases because relevant confounding factors are not known or not easily measured and in other cases because investigators apportion their limited resources in other directions. Finally, as our paper noted[1], even with the best design and analysis, there is likely to be a level of residual structure after allowing for known confounders. At present there is limited relevant data to determine the probable levels of residual structure, but the simulations in our paper deliberately included plausible scenarios for these. Notably, in their original paper[2], Devlin and Roeder described the level of population structure that we considered in ref. 1 (F = 0.01 in their notation) as "realistic". Further, as noted in ref. 2, cryptic relatedness poses as much of a threat to association studies as does geographic population structure and is much more difficult to reduce by experimental design. Preliminary analysis of a large UK case-control study (886 cases, 878 controls, 8,000 markers) showed substantial inflation of χ^2 statistics even after accounting for broad geographical region, with a portion of this inflation plausibly due to population structure (D. Clayton, personal communication).

Although we are positive in general about Bayesian statistical methods, we urge caution against viewing the Bayesian mixture approach (GCB), and more generally false discovery rates[4–6], as a simple panacea to multiple testing issues. There are not often free lunches. The idea of GCB is to partition loci into two groups: those associated with the disease (outlier loci) and those not associated with the disease, using a sensible statistical model, and method, to assign loci to each group. Informally, this will be easy if the test statistics of outlier loci look very different from those of nonassociated loci, which would be the case if the genetic effects were large and if there were moderate numbers of loci in each category. On the other hand, for the small effects appropriate to complex diseases, genome scans with massive numbers of nonassociated loci and a small relative number of true disease loci, the tail of the null distribution (after GC) of test statistics may well overlap, or possibly even bury, the few values from associated loci, and no statistical procedure will reliably separate the two. These kinds of settings have not been extensively explored.

We conclude with two points of detail. It is false that our original paper[1] assumed "subjects that originate from different populations". Much of our focus (*e.g.*, Fig. 4c–e and Fig. 6 in ref. 1) deliberately (and explicitly) concerned structure plausible within current populations. Finally, there are two different ways in which GC (or GCB or GCF) could fail in practice: (i) the null distribution of the test statistic may not behave as a simple multiple of a χ^2 distribution, or (ii) the statistical allowance for the inflation factor may not be effective. The 'short cut' simulations given by Devlin *et al.* above presuppose that the first point is not a problem. In the absence of a formal mathematical proof, and with abundant computing resources, it would seem better to check routinely both aspects of GC, as in their Table 1, rather than only the second, as in their Figures 1 and 2.

Jonathan Marchini[1], Lon R Cardon[2], Michael S Phillips[3] & Peter Donnelly[1]

[1]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. [2]Wellcome Trust Center for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK. [3]Genome Quebec and McGill University Genome Center, Montreal H3A 1A4, Canada. Correspondence should be addressed to P.D. (donnelly@stats.ox.ac.uk).

1. Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly P. *Nat. Genet.* **36**, 512–728 (2004).
2. Devlin, B. & Roeder, K. *Biometrics* **55**, 997–1004 (1999).
3. Freedman, M.L. *et al. Nat. Genet.* **36**, 388–393 (2004).
4. Benjamini, Y. & Hochberg, Y. J. R. Stat. Soc. B **57**, 289–300 (1995).
5. Storey, J.D. J. R. Stat. Soc. B **64**, 479–498 (2002)
6. Storey, J.D. & Tibshirani, R. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).