Association Studies for Quantitative Traits in Structured Populations

Silviu-Alin Bacanu, B Devlin and Kathryn Roeder

Department of Psychiatry (S-AB, BD)
University of Pittsburgh
Pittsburgh,PA
Department of Statistics (KR)
Carnegie Mellon University
Pittsburgh,PA

Running Title: Association Studies for Quantitative Traits

Address for correspondence and reprints:

Kathryn Roeder, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: roeder@stat.cmu.edu

Abstract

Association between disease and genetic polymorphisms often contributes critical information in our search for the genetic components of common diseases. Devlin and Roeder (1999) introduced genomic control, a statistical method that overcomes a drawback to the use of population-based samples for tests of association, namely spurious associations induced by population structure. In essence, genomic control (GC) uses markers throughout the genome to adjust for any inflation in test statistics due to substructure. To date genomic control (GC) has been developed for binary traits and bi- or multiallelic markers. Tests of association using GC have been limited to single genes. In this report, we generalize GC to quantitative traits (QT) and multilocus models. Using statistical analysis and simulations, we show that GC controls spurious associations in reasonable settings of population substructure for QT models, including gene-gene interaction. Through simulations we explore GC power for both random and selected samples, assuming the QT locus tested is causal and its specific heritability is 2.5 - 5%. We find that GC, combined with either random or selected samples, has good power in this setting, and that more complex models induce smaller GC corrections. The latter suggests greater power can be achieved by specifying more complex genetic models, but this observation only follows when such models are largely correct and specified *a priori*.

Key Words: Genomic Control, population structure, polymorphisms, complex disease, overdispersion

Introduction

Quantitative traits can be a direct expression of human disease, such as extreme blood pressure and obesity, an indirect expression, such as the degree of obsessionality for certain psychiatric disorders, or even a latent measure of disease liability. Therefore locating the polymorphisms affecting quantitative traits is a major goal of genetic epidemiology (Risch 2000; Blangero et al. 2001). Finding these polymorphisms has proven challenging, however, because the diseases and the quantitative traits (QT) underlying them generally have a complex genetic and environmental basis (Risch 2000). This complexity dampens the power of any linkage analysis. Magnifying the challenge, the polymorphisms themselves can have only a subtle impact on the QT.

One approach that could increase power is to assess candidate QT loci (QTL) for association using population-based samples and regression methods. For example, for a candidate Single Nucleotide Polymorphism (SNP), the measured value of the quantitative trait Y can be regressed on the count of a specific allele and a test performed to determine if $\hat{\beta}$, the least squares estimate of association between the phenotype and the allele count, differs significantly from zero. By using population-based samples, however, a significant association between disease and SNP alleles could arise from three different sources: by chance; by tight linkage to a causal polymorphism; or, spuriously, by the impact of population structure. The latter is counter to family-based studies, which are robust to the impact of population substructure (Allison 1997; Rabinowitz 1997; Zhu and Elston 2001).

With respect to confounding generated by population substructure, two factors perturb the distribution of $\hat{\beta}$ from that expected in the typical regression setting: (i) the phenotypes of samples drawn from the same subpopulation are positively correlated, increasing the variance of $\hat{\beta}$ over that expected under the independence model and leading to an over-dispersed test statistic; and (ii) the $E[\hat{\beta}]$ is not zero, even under the null hypothesis of no linkage. Regarding (ii), the statistic may be biased in either direction, depending upon the nature of the population substructure. Although the bias can be sizable for rare Mendelian disorders, it is typically dominated by over-dispersion for complex disorders (Devlin et al., 2001a).

Concerns about confounding due to population substructure have increased the popularity of family-based studies at the expense of population-based studies. Yet family-based methods are not without their own drawbacks. Recruiting family members can be difficult, and thus the sample sizes for family-based studies tend to be relatively small. The connection between

diminished sample size and statistical power is obvious. Moreover, in some settings, family-based designs are less powerful than population-based methods (Risch and Teng 1998; Bacanu et al. 2000) even for the same number of individuals sampled.

Recently, Devlin and Roeder (1999) put forth an alternative approach that utilizes population-based samples, but protects against confounding. We call it genomic control (GC) because GC uses genotypes obtained from across the genome to control for violations from independence present in the sample [Devlin and Roeder 1999; Bacanu et al., 2000; Devlin et al. 2001a]. "Null" loci, defined as loci unlikely to have a functional effect on the trait under investigation, are used to estimate the impact of the confounding observed in population-based samples. Provided the same individuals are genotyped at all the loci, this approach corrects for confounding due to population stratification even for the "convenience samples" often encountered in clinical trials and epidemiological samples.

Here we evaluate GC to test for association between QT and genetic markers in several settings: tests of association with a single SNP; a more general model with two SNP and interaction; for omnibus F-tests; and, finally, for selection of individuals based on their QT values. For power analysis, we assume the SNP under study has a direct impact on the QT.

Quantitative Traits Model and GC

For association studies, confounding occurs when individuals are drawn from a substructured population (sensu Wright 1951). Due to the substructure, the multilocus genotype probabilities for unlinked loci are not equal to a product of the allele probabilities. Population substructure creates correlation among alleles sampled from the same subpopulation, which is measured by F_{st} (Wright 1969).

Consider a biallelic locus with alleles labeled 0 and 1 with the probability of a 1 equal to $p_k = 1 - q_k$. For alleles (W_1, W_2) drawn from the same subpopulation, the proportion of genotypes (1,1), (1,0) and (0,0) are $F_{st}p_k + (1 - F_{st})p_k^2$, $2(1 - F_{st})p_kq_k$ and $F_{st}q_k + (1 - F_{st})q_k^2$ respectively (Wright 1951). As a consequence $\mathrm{Cov}(W_1, W_2) = F_{st}p_k(1 - p_k)$ if the two alleles are sampled from the same subpopulation, and 0 otherwise; likewise $\mathrm{Var}(W_1 + W_2) = 2(1 + F_{st})p_k(1 - p_k)$.

In terms of experimental design, we draw a population-based sample of N individuals, each measured for a QT Y. Assume Y could be influenced by the genotypes at c unlinked, biallelic loci of interest and n_h other hidden biallelic loci, which are not under investigation.

Define X_k, $k = 1, \ldots, c + n_h$, to be the individual's centered genotype at locus k, given by the number of 1 alleles minus two times the proportion of 1 alleles in the sample. A working model for Y is an additive linear model with interactions between loci:

$$Y_i = \beta_0 + \sum_{k=1}^{c+n_h} \beta_k X_{ik} + \sum_{1 \le k < l \le c+n_h} \beta_{kl} X_{ik} X_{il} + \zeta_i. \quad (1)$$

We assume $\mathrm{Var}(\zeta_i)$ is a constant, $\mathrm{Cov}(\zeta_i, \zeta_j) = 0$, and only require the working model to approximately hold for the terms indexed by $k = 1, \ldots, c$. These targeted loci can be tested for association with the phenotype by fitting the model,

$$E[Y_i | X_1, \ldots, X_c] = \beta_0 + \sum_{k=1}^{c} \beta_k X_{ik} + \sum_{1 \le k < l \le c} \beta_{kl} X_{ik} X_{il}, \quad (2)$$

To test the effect of locus k, we assume under the null hypothesis that any β with a subscript involving $k \le c$ is zero.

Given model (2), we investigate the variance, defined to be $\mathrm{Var}[Y_i] = \sigma^2$, and the covariance, defined to be $\mathrm{Cov}[Y_i, Y_j] = \rho$, when i and j are drawn from the same subpopulation. An assumption of the GC approach is that neither σ^2 nor ρ depend strongly upon the allele frequencies for any loci indexed by $k \le c$ (see Devlin et al. 2001a for discussion and analysis). By definition

$$\sigma^2 = \mathrm{Var}\left[\sum_{k=c+1}^{c+n_h} \beta_k X_{ik} + \sum_{c+1 \le k < l \le c+n_h} \beta_{kl} X_{ik} X_{il} + \zeta_i \right],$$

in which there are no restrictions on the β_k's. Because $\mathrm{Var}[X_{ik}] = 2(1 + F_{st}) p_k (1 - p_k)$, it follows that σ^2 is a function of $\{p_k, k = c + 1, \ldots, n_h + c\}$. If (i, j) are not in the same subpopulation, $\mathrm{Cov}[Y_i, Y_j] = 0$; this quantity is defined to be ρ otherwise. Clearly ρ is also a function of $\{p_k, k = c + 1, \ldots, n_h + c\}$ because $\mathrm{Cov}[X_{ik}, X_{jk}] = 4 F_{st} p_k (1 - p_k)$ if (i, j) are in the same subpopulation. However, neither σ^2 nor ρ depend upon the allele frequencies of the loci under investigation because we condition upon $\{X_1, \ldots, X_c\}$.

Let $\hat{\beta}$ be the least squares estimate of $\beta = (\beta_0, \beta_1, \ldots, \beta_{c-1,c})$ obtained from model (2), ignoring the effects of population substructure. Bias and over-dispersion perturb the distribution of $\hat{\beta}$ from that expected in the typical regression setting. To construct a valid test for association between Y and a SNP genotype, beyond that expected due to population substructure, we extend GC to the quantitative traits setting by estimating and controlling for any inflation in the variance over that expected for independent data. We focus on over-dispersion. When bias is also present, it results in additional over-dispersion, which is automatically corrected for by GC (Devlin et al. 2001a).

Here we outline how the GC approach works, assuming $c = 1$. In the remainder of the report, we generalize the model and develop these ideas further. Estimate $\hat{\beta}_1$ at M loci. Ideally a subset of loci can be designated as null loci, but any set of loci for which the majority are not associated with the QT will suffice. Compute $T_k = \hat{\beta}_1 / SE_{ind}[\hat{\beta}_1]$, in which the denominator is the standard error of the numerator, ignoring population substructure. Under the null hypothesis, and for large sample sizes, T_k is approximately distributed $N(0, \lambda)$, $\lambda = \eta^2 + \tau^2$, in which η^2 is proportional to the square of the expected bias of the test statistic and τ^2 is the increase in the variance due to correlation among subjects. Consequently, T_k^2/λ is distributed as χ_1^2 (Devlin et al. 2001a). The inflation factor λ can be robustly estimated using $\hat{\lambda} = \{\text{median}(T_1^2, T_2^2, \ldots, T_M^2)/.456\}$ or any other of the robust methods described in Devlin et al. [2001b]. Then compare $T_k^2/\hat{\lambda}$ with $\chi_1^2(1)$ to determine whether the locus is significantly associated with the QT. To test L candidate loci, one at a time, compare $T_k^2/\hat{\lambda}$ with $\chi_1^2(1 - \alpha/L)$ to determine which are significantly associated with the QT.

Inferences for quantitative traits using GC

Testing for a single locus effect

To test if a single locus, say $k = 1$, is associated with the phenotype we work with the model

$$E[Y|X_{i1}] = \beta_0 + \beta_1 X_{i1}, \tag{3}$$

and test whether the slope is different from zero. To simplify exposition, we drop the subscript k. The usual estimator of the parameter of interest is $\hat{\beta}_1 = \sum_i X_{i1} Y_i / \sum_i X_{i1}^2 = \sum_i b_i Y_i$. We need to compute $\text{Var}[\hat{\beta}_1] = \sum_i b_i^2 \text{Var}[Y_i] + 2 \sum_{i<j} b_i b_j \text{Cov}[Y_i, Y_j]$. Note that $\frac{1}{N} \sum_i X_i^2 \approx \text{Var}(X_i) = 2(1 + F_{st})p(1-p)$. Consequently $\sum_i b_i^2 \approx 1/\{2N(1 + F_{st})p(1-p)\}$. Recall that $E[X_i X_j] = 4F_{st}p(1-p)$ if i and j are drawn from the same subpopulation, and 0 otherwise. It follows that for the N_a pairs of observations from subpopulation S_a,

$$\frac{1}{N_a} \sum_{i<j \in S_a} b_i b_j \approx \text{Cov}[b_i, b_j] = F_{st}/\{N^2[2(1 + F_{st})^2 p(1-p)]\}.$$

Let R be a count of the number of covariance terms not equal to zero, which is the sum of $N_j(N_j - 1)/2$ over each of the j subpopulations. Define $SE_{ind}[\hat{\beta}_1]$ as the usual standard error term when the Y_i's are independent. $SE_{ind}[\hat{\beta}_1]$ is approximately equal to $\{\sigma^2/2N(1 + F_{st})p(1-p)\}^{1/2}$, the same term obtained from any statistical regression package. It is not assumed here that the pair of alleles within an individual, W_{1i} and W_{2i}, are independent.

Putting all the pieces together we find that

$$\begin{aligned} \mathrm{Var}[\hat{\beta}_1] &\approx \frac{\sigma^2}{2N(1+F_{st})p_1(1-p_1)}\left[1+\frac{4RF_{st}\rho}{N(1+F_{st})\sigma^2}\right] \\ &\approx SE^2_{ind}[\hat{\beta}_1]\times\tau^2. \end{aligned}$$

The adjustment to the usual variance term, τ^2, is the inflation factor due to correlation among the subjects in the study. If $F_{st}=0$, $\tau^2=1$ and the variance reduces to $\sigma^2/\{2Np(1-p)\}$. In general, we anticipate $\tau^2>1$. If R and F_{st} are large, then the second term in τ^2 can be large. Alternatively, if there are many small subpopulations, then R will be small and the impact of population substructure on the variance will be small.

Throughout this derivation we assume a constant correlation structure within the sample defined by F_{st}; however, the GC principle applies even when the correlation between sampled individuals is not constant across pairs of individuals within a subpopulation. It suffices that the correlation between individuals is approximately constant across the genome. For instance, the method is still valid if some subpopulations are more differentiated than others, or the sample includes some pairs of individuals who are close relatives. Furthermore, in practice, even if the fundamental assumption of constant correlation between individuals across the genome is violated in a minor way the test tends to be conservative.

Confounding due to population substructure also introduces a bias into the problem. In expectation the bias squared is also proportional to the standard error: $\mathrm{bias}^2=\eta^2\,SE^2_{ind}[\hat{\beta}_1]$ (Devlin et al., 2001a). For any single locus, neither η^2 nor τ^2 can be directly estimated because they depend upon unknown quantities, namely the allele frequencies and the allelic effects at QT loci that determine the heritability of the trait. However, both η^2 and τ^2 are approximately constant regardless of which loci we are studying, assuming the null hypothesis holds (Roeder and Devlin 1999; Bacanu et al. 2000). Consequently $\lambda=\eta^2+\tau^2$ can be estimated from the null loci as described previously.

Testing for effects at two loci

Consider fitting two more complex models

$$E[Y|X_1,X_2]=\beta_0+\beta_1X_1+\beta_2X_2, \tag{4}$$

$$E[Y|X_1,X_2]=\beta_0+\beta_1X_1+\beta_2X_2+\beta_{12}X_1X_2. \tag{5}$$

For reference we label the models in (3), (4), and (5) as models I, II and III, respectively.

Applying the same type of argument as used for the single locus test to model II, our analysis (Appendix), shows $\mathrm{Var}[\hat{\beta}_1] = SE^2_{ind}[\hat{\beta}_1][\tau^2 - H]$, in which $SE_{ind}[\hat{\beta}_1]$ is the standard error of $\hat{\beta}_1$ computed when fitting model II and assuming independent observations, τ^2 is the inflation factor obtained previously when fitting model I, and H is a small positive term that accounts for a slight decrease in the inflation term obtained when fitting model II rather than model I. In principle H can depend upon p_A and p_B; in practice we found $\tau^2 - H$ to be approximately constant across the genome in our simulations. Based on these results it follows that GC applies when fitting model II, but the inflation factor is now $\tau^2 - H$, which we estimate as described previously.

Incorporating the interaction term as in model III greatly increases the complexity of the problem, making it difficult to obtain a transparent analytical picture. The additional complexity occurs because the level of correlation between alleles at different loci in substructured populations is not determined by F_{st}; in fact, there is no simple analytical representation of these correlations. Because analytical representations of the variance are not available for the main effects or the interaction effect, we primarily pursue this model further with simulations. Nevertheless we determine some properties of GC under model III in the Appendix.

To test for multiple effects simultaneously, a common approach involves an F-test based upon the extra sums of squares:

$$F = \frac{SSE(reduced) - SSE(full)}{df \times MSE(full)}, \tag{6}$$

where df is the difference in degrees of freedom between the full and reduced models. For large N, $MSE(full) \approx \sigma^2$. Consequently, for an independent identically distributed sample, $df \times F$ is approximately distributed as a χ^2_{df}. In the numerator of the omnibus test (6), the extra sums of squares attributable to two main effects and an interaction can be partitioned as $SSR(X_1) + SSR(X_2|X_1) + SSR(X_{12}|X_1, X_2)$. If a single main effect is fit

$$\frac{SSR(X_1)}{\lambda \times MSE} = \frac{\hat{\beta}_1^2}{\lambda\, SE^2_{ind}(\hat{\beta}_1)},$$

where λ denotes the single main effect inflation factor. If a second main effect is included

$$\frac{SSR(X_2|X_1)}{\lambda \times MSE} = \frac{\hat{\beta}_2^2}{\lambda\, SE^2_{ind}(\hat{\beta}_2)},$$

where λ now denotes the main effect inflation factor, given that one main effect is already in

the model. Finally, if the interaction term is included along with two main effects

$$\frac{SSR(X_{12}|X_1, X_2)}{\lambda_{int} \times MSE} = \frac{\hat{\beta}_{12}^2}{\lambda_{int}\, SE_{ind}^2(\hat{\beta}_{12})},$$

where λ_{int} denotes the interaction effect inflation factor. In the previous section we showed that, for a substructured population, each of the three quantities above is approximately distributed as a χ_1^2. Consequently, the GC omnibus test can be obtained by sequentially computing these three terms and then summing them to obtain a χ^2 test with three degrees of freedom. In the upcoming simulation we show that a test with proper Type I error rate and excellent power is obtained by simply computing the F-test indicated in (6) for the loci under investigation, as well as a set of null loci; estimating λ_{tot} based upon the median of the tests obtained from the null loci; and then adjusting the F-test for the effect of population substructure by dividing by $\hat{\lambda}_{tot}$. For $df = d$, $\hat{\lambda}_{tot} = \{median(F_1, \ldots, F_M)/v_d\}$ in which $v_2 = 0.70$ and $v_3 = 0.79$.

Simulations

Simulation algorithm

We generate data using our working model (1) with $\mathrm{Var}(\zeta_i) = 1$ and $\beta_{kl} = 0$ for $\{c < k < l\}$. To compute heritability in our simulation, X_i is assumed to be independent of X_j (unlinked loci) but not independent of $X_i X_j$ for $i \neq j$. Without population substructure, the heritability due to the c loci of interest is:

$$\begin{aligned}
& \sum_{i=1}^{c} \beta_i^2 \mathrm{Var}(X_i) + \sum_{1 \le i < j \le c} \beta_{ij}^2 \mathrm{Var}(X_i X_j) \\
+ \;& 2 \sum_{1 \le i < j \le c} [\beta_i \beta_{ij} \mathrm{Cov}(X_i, X_i X_j) + \beta_j \beta_{ij} \mathrm{Cov}(X_j, X_i X_j)] \\
= \;& \sum_{i=1}^{c} \beta_i^2 2 p_i q_i + \sum_{1 \le i,j \le c} \beta_{ij}^2 4 p_i p_j (p_i + p_j + 1 - 3 p_i p_j) \\
+ \;& 2 \sum_{1 \le i < j \le c} (\beta_i \beta_{ij} 4 p_i p_j q_i + \beta_j \beta_{ij} 4 p_i p_j q_j) \\
= \;& 2 \sum_{i=1}^{c} \beta_i^2 p_i q_i + 4 \sum_{1 \le i,j \le c} \beta_{ij}^2 p_i p_j (p_i + p_j + 1 - 3 p_i p_j) \\
+ \;& 8 \sum_{1 \le i < j \le c} p_i p_j \beta_{ij} (\beta_i q_i + \beta_j q_j);
\end{aligned}$$

in this calculation the X_i's are not centered. The heritability of all the $c + n_h$ loci is similar.

To achieve a fixed level of heritability attributable to a set of loci, either the parameters for the slopes of the main effects and interactions can be held constant and solve for allele frequencies that satisfy the constraints, or vice versa. For all of our simulations the total heritability is set at 50%.

Using the algorithm described in Bacanu et al. (2000) with some small changes to incorporate quantitative traits, a sample of size N can be generated from a collection of n_{strata} subpopulations characterized by allelic correlation within a subpopulation equal to F_{st}. Each sample contains a phenotype and genotypes at n loci consisting of c candidate loci, n_h hidden loci and $n - c - n_h$ null loci. For simulations under the null hypothesis only hidden loci $k = 1, \ldots, 50$ are generated; each locus contributes 1% to achieve a total heritability of 50%. Under the alternative hypothesis, h^2 denotes the heritability attributable to each locus under study or their interaction; in addition hidden loci are generated to achieve a total heritability of 50%.

Simulation Results

In our simulations we investigate the conjecture made previously that the over-dispersion parameter is approximately constant as a function of the allele frequencies of the loci being tested. Null loci are sampled from a stratified population with $n_{strata} = 2$, $F_{st} = 0.03$, and allele frequencies p ranging between 0.08 and 0.92. Then model III is fit to these data. The mean of χ^2 tests, pooled over common levels of p, are plotted as a function of p in Figure 1. For both main and interaction effects, the test statistic clearly does not depend upon p. Consequently the GC approach is valid for model III even under an extreme case of population stratification. The only pattern apparent in these results is that the test statistic for the interaction has greater variability when both alleles are relatively uncommon. This is not surprising because, under these conditions, sample sizes may not be sufficient for $\hat{\beta}_{12}$ to achieve normality.

Data are simulated under the null hypothesis with $c = 2$ to investigate the Type I error rate of the test and other features of the GC procedure using the following conditions for the candidate loci in each of 4000 populations: $p_1 = p_2 = 0.2$; $n_h = 50$; and $N = 2000$. Two levels of stratification (low and high) are evaluated, $F_{st} = 0.003$ and $F_{st} = 0.03$, with $n_{strata} = 6$. For each sample, models I, II and III are fit to the full data ($N = 2000$: unselected) and a selected fraction of the sample (selected). Our selection strategy took individuals if their QT value fell in the lower or upper quartile.

From the simulations using either unselected or selected samples, the following results (Table I) obtain: (i) the test achieves the appropriate (or conservative) Type I error rate; (ii) the parameter estimates for β are unbiased on average; (iii) for $F_{st} = 0.03$, λ is slightly smaller if model II is fit instead of I, consistent with our algebraic calculations; (iv) if model III is fit λ is reduced even further; and (v) λ_{int} is smaller than λ. Result (iv) follows because fitting model III is roughly equivalent to conditioning on some of the confounding. Thus, fitting the interaction removes some of the variability in the main effect estimates, which was induced by substructure. Result (v) obtains because it is difficult to have the right configuration of subpopulations to create a strong interaction effect due to confounding alone. More so than in main effects models, there is a natural averaging effect across subpopulations. Hence there simply is less confounding to remove from the calculations. These trends are less apparent when $F_{st} = 0.003$ presumably because this is a low level of stratification and the effects of confounding are modest. Finally, note that λ_{tot} is generally a compromise between λ and λ_{int}.

To investigate the power of the test for one and two trait loci ($c = 1$ and 2), we simulate 200 populations, and set other parameters as defined previously. In Tables II and III the data were generated with model I, while in Tables III and IV the data were generated with model III. When a locus accounts for 2.5% (5%) of the heritability variation, then the coefficient equals 0.280 (0.395). For model III $\beta_{12} = 0.349$ or $\beta_{12} = 0.494$ respectively. Models I, II and III are fit to the simulated data regardless of whether only one locus accounts for the heritability or both loci and their interaction account for the heritability.

From these simulations and unselected samples the following results obtain: (i) if the true model is fitted, the estimated parameters are unbiased on average (Tables II and IV); (ii) if the true model has two main effects and an interaction, then fitting a single one of the two main effects will yield an upwardly biased estimate of this main effect (Table IV); (iii) the power decreases as F_{st} increases, but not dramatically (Tables III and V); (iv) if the true model has one main effect and no interaction, then fitting the appropriate main effect using model I leads to an increase in power compared to fitting model II or III (Table III); (v) if the true model has two main effects and an interaction, then fitting a single one of the two main effects using model I will substantially decrease the power to detect the effect relative to fitting model II or III (Table V), using the omnibus test.

In general, λ is smaller when a more complex model is fitted and hence it is not surprising that the power is often greater when model III is fitted rather than model I; however, there is a tradeoff. When exploring L candidate loci there are $L(L-1)/2$ pairwise interaction

models to test and hence a substantial Bonferroni correction is required. For models such as the ones simulated here, our simulations suggest that it is better to explore one-factor main effects sequentially unless targeted comparisons are of interest. Of course these results do not necessarily extend to other interaction models. When the interaction dominates the main effects, then the tradeoff is likely to favor fitting interaction models. Clearly, in the extreme case where there is a gene-gene interaction with no main effects the optimal form of analysis involves fitting the model $E[Y] = \beta_0 + \beta_{12} X_1 X_2$. The inferences will be much more powerful with this approach than would be obtained fitting model I.

For the selected samples, parameter estimates are biased, increasing to almost twofold their true value even when the true model is fit to the data (Tables II and IV); yet the test achieves the targeted Type I error rate (Table I). Moreover, by sampling from the upper and lower quartiles of the trait distribution, one achieves equivalent power genotyping 50 to 60% of the N required for random sampling (data not shown), regardless of the model.

Discussion

Genomic Control (GC) is a new approach to analysis of population-based samples to find associations between disease and marker alleles (Devlin and Roeder 1999; Devlin et al. 2001a). GC permits valid inference from population-based samples drawn from substructured and therefore heterogeneous human populations. By assessing multiple loci across the genome, only a subset of which could have a meaningful impact on the trait of interest, GC eliminates the confounding effects of population substructure.

In their original article, Devlin and Roeder (1999) outline methods for discrete outcomes, principally through a case-control design. We extend GC to the search for quantitative trait loci (QTL). As predicted by Bacanu et al. (2000), we find that GC extends naturally to the analysis of quantitative traits (GC-QT). Our results go further than a simple extension by exploring GC-QT for models with multiple QTL. We find that the principles of GC extend to this more complicated setting also. In fact, while the results are not shown here, we have also explored log-linear models for discrete (case-control) outcomes. We find that the results shown for GC-QT and the results for log-linear models are very similar. Thus it appears a whole class of multilocus models are amenable to GC analysis.

Various quantitative traits and samples could be the target of GC-QT analyses. One source of attractive samples would be those drawn for clinical trials. Such studies usually produce a

quantitative outcome of drug efficacy, either for ameliorating disease symptoms or the disease itself, based on a large, population-based sample. Often the drug response is heterogeneous within the sample, presumably due to the impact of genetic variability. It is of considerable interest, therefore, to understand the heterogeneous response based upon the pharmacological properties of the drug and its targets within the body. As noted in Devlin and Roeder (1999) and elsewhere (Bacanu et al. 2000; Devlin et al. 2001a), the most powerful implementation of GC in this setting is to account for any obvious sources of population substructure in the statistical model *a priori*.

An interesting issue for GC is the need for "null" loci. Throughout this article, we allude to the use of such loci to develop the GC-QT correction. When many candidate genes are evaluated, such as in a search for liability loci, null loci are extraneous – assuming, reasonably, that most candidate genes will not have important effects on disease outcome. Instead, researchers can apply the Bayesian outlier model developed by Devlin and Roeder (1999). In essence, this mixture model looks for genes that have unusually large associations relative to the bulk of the genes evaluated. Simple frequentist methods for estimating λ from a candidate gene study are also available (Devlin et al. 2001b; Tzeng et al. 2001). With the recent determination of human DNA sequence and the development of large SNP databases, we anticipate analyses with this structure will become commonplace.

Our power analyses for GC-QT explored the use of one and two-locus models, two-locus models with interactions, and omnibus F-tests for various true models. These simulations assume 50% heritability for the trait, but modest heritability for any particular locus. Population heterogeneities are taken to be similar to that for samples of Caucasian or African Americans and mixtures of the two ethnic groups. For all of our reported parameter values and even for selected sampling, we find the Type I error rate of GC-QT test to be less than or equal to its nominal value (Table I); even for extreme substructure, near nominal values are obtained (unpublished data). Furthermore, we find excellent power for reasonable sample sizes (Tables III and V).

Power, of course, depends on the congruence between the true model generating the data and the model fit to the data (Tables III and V). If our earlier conjecture is true – that studies will soon be assessing a large number of genes simultaneously in their search for liability genes – then our results are particularly interesting. When multiple genes are tested, there will be great temptation to test for gene-gene interaction. Including interactions in the model can lead to more powerful inference, but only when the interaction is large relative to the main

effects; otherwise substantial power is lost. Thus, tests for interaction must be used sparingly. These results are reminiscent of the results of Dupuis et al. (1995), who show that whole-genome two-liability-locus linkage is only powerful when the gene-gene interactions are large. In the future, what will be required are models that incorporate prior biological information and thereby delineate appropriate interactions *a priori*.

In our simulation study the effect of population stratification appears to be quite small for low levels of stratification, as evidenced by the magnitude of λ when $F_{st} = 0.003$ (Table I). This simulation was designed to mimic a sample drawn from a fairly homogeneous ethnic group such as a population of European decent. However, these results should not be taken to indicate that population stratification can always be safely ignored when the sample is drawn from a fairly homogeneous sample because the effect of confounding grows as the sample size increases.

The GC approach has now been extended to a number of data structures obtained from population-based samples, and it seems likely that this general idea can be applied even more broadly. For instance, two methods designed for the analysis of population-based samples recently appeared which may be amenable to correction using a GC approach: threshold-defined case control analysis for QT [Schork et al. 2000] and case-control studies including relatives [Slager and Schaid 2001].

Acknowledgements

This research was supported by National Institute of Health grants MH57881 and MH56193 and National Science Foundation grant DMS-9803433.

Appendix

For model II, we wish to find the approximate distribution of the estimates of β_1 and β_2. For notational convenience let $A = X_1$ and $B = X_2$. To facilitate algebraic manipulations, redefine the quantitative trait as the centered analog, $Y_i = Y_i - \bar{Y}$, and set $\beta_0 = 0$. From the least squares calculations it follows that

$$\hat{\beta}_1 = \left[\sum_i A_i^2 \sum_i B_i^2 - (\sum_i A_i B_i)^2\right]^{-1} \left[\sum_i A_i^2 \sum_i A_i Y_i - \sum_i A_i B_i \sum_i B_i Y_i\right]. \tag{7}$$

Define $\frac{1}{N}\sum_i A_i^2 = \sigma_A^2$, $\frac{1}{N}\sum_i B_i^2 = \sigma_B^2$, and $\frac{1}{N}\sum_i A_i B_i = \sigma_{AB}$. Let $I_{ij} = 1$ if i and j are in the same subpopulation and zero otherwise. Re-express (7) as

$$\left[\sigma_A^2\sigma_B^2 - \sigma_{AB}^2\right]^{-1} \left[\sigma_B^2 \frac{1}{N}\sum_i A_i Y_i - \sigma_{AB}\frac{1}{N}\sum_i B_i Y_i\right].$$

It follows that

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}_1] &= \left(\frac{\sigma_B^2}{\sigma_A^2\sigma_B^2 - \sigma_{AB}^2}\right)^2 \left[\frac{\sigma^2\sigma_A^2}{N} + \frac{2\rho}{N^2}\sum_{i<j} I_{ij}A_iA_j\right] \\
&+ \left(\frac{\sigma_{AB}}{\sigma_A^2\sigma_B^2 - \sigma_{AB}^2}\right)^2 \left[\frac{\sigma^2\sigma_B^2}{N} + \frac{2\rho}{N^2}\sum_{i<j} I_{ij}B_iB_j\right] \\
&- \frac{2\sigma_B^2\sigma_{AB}}{(\sigma_A^2\sigma_B^2 - \sigma_{AB}^2)^2} \left[\frac{\sigma^2\sigma_{AB}^2}{N} + \frac{2\rho}{N^2}\sum_{i\neq j} I_{ij}A_iB_j\right].
\end{aligned} \tag{8}$$

Claim

$$\begin{aligned}
\sum_{i<j} I_{ij}A_iA_j &= \frac{2RF\sigma_A^2}{1+F} \\
\sum_{i<j} I_{ij}B_iB_j &= \frac{2RF\sigma_B^2}{1+F} \\
\sum_{i\neq j} I_{ij}A_iB_j &= 2R\sigma_{AB}.
\end{aligned}$$

Plugging these quantities into (8), and after some algebra, we obtain

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2\sigma_A^2}{N(\sigma_A^2\sigma_B^2 - \sigma_{AB}^2)}\left[1 + \frac{4R\rho F}{N(1+F)\sigma^2} - H\right],$$

in which

$$H = \frac{4R\rho\sigma_{AB}^2}{N\sigma^2(\sigma_A^2\sigma_B^2 - \sigma_{AB}^2)}\left[1 - \frac{2F}{1+F}\right].$$

But this is $SE_{ind}^2[\hat{\beta}_1][\tau^2 - H]$.

Incorporating the interaction term as in model III greatly increases the complexity of the problem, making it difficult to obtain a transparent analytical picture. The additional complexity occurs because the level of correlation among alleles across loci in substructured populations is not determined by F_{st}; in fact, there is no simple analytical representation of these correlations. Nevertheless we can determine some properties of GC under model III.

For model III define

$$\begin{aligned} c &= \sum_i A_i^2 B_i^2 \times \sum_i B_i^2 - (\sum_i A_i B_i^2)^2 \\ d &= \sum_i A_i^2 B_i \times \sum_i A_i B_i^2 - \sum_i A_i^2 B_i^2 \times \sum_i A_i B_i \\ e &= \sum_i A_i B_i \times \sum_i A_i B_i^2 - \sum_i A_i^2 B_i \times \sum_i B_i^2 \end{aligned}$$

In terms of these quantities

$$\hat{\beta}_1 = \left[c \sum_i A_i^2 + d \sum_i A_i B_i + e \sum_i A_i^2 D_i \right]^{-1} \left[c \sum_i A_i Y_i + d \sum_i B_i Y_i + e \sum_i A_i B_i Y_i \right].$$

By the Cauchy-Schwarz inequality $c > 0$ and, from simulations, we find that, in expectation d and e are near zero. Consequently, asymptotically $\hat{\beta}_1$ behaves as when we fit model I, $E[\hat{\beta}_1] \approx \sum_i A_i Y_i / \sum_i A_i^2$.

The variance of $\hat{\beta}_1$ is determined by the variances and covariances of the terms in the second set of square brackets. For any particular dataset d and e deviate from zero and hence $\mathrm{Var}[\hat{\beta}_1]$ clearly differs when we fit model III rather than model I. Because analytical representations of the variance are not available for the main effects or the interaction effect, we pursue this model further with simulations.

References

Allison DB. 1997. Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet 60:676-690.

Bacanu S-A, Devlin B, Roeder K. 2000. The power of genomic control. Am J Hum Genet 66:1933-1944.

Blangero J, Williams JT, Almasy L. 2001. Variance component methods for detecting domplex trait loci. Adv Genet 42:151-181.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics 55:997-1004.

Devlin B, Roeder K, Wasserman L. 2001a. Genomic control, a new approach to genetic-based association studies. Theoretical Population Biology (in press).

Devlin B, Roeder K, Bacanu S-A. 2001b. Unbiased methods for population-based association studies. (In preparation.)

Dupuis J, Brown PO, Siegmund D. 1995. Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. Genetics 140:843-856.

Longmate JA. 2001. Complexity and power in case-control association studies. Am J Hum Genet 68:1229-1237.

Rabinowitz D. 1997. A transmission disequilibrium test for quantitative trait loci. Hum Hered 47:342-350.

Risch NJ. 2000. Searching for genetic determinants in the new millennium. Nature 405:847-856.

Risch N, Teng J. 1998. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human disease. I. DNA pooling. Genome Res 8:1273-1288.

Schork NJ, Nath SK, Fallin D, Chakravarti A 2000. Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control subjects. Am J Hum Genet. 67 :1208-18.

Slager SL, Schaid DJ. 2001. Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects. Am J Hum Genet 68, 1457-1462. .

Tzeng J-Y, Byerley W, Devlin B, Roeder K., Wasserman L 2001. Outlier detection and False Discovery Rates for whole-genome DNA matching. J Amer Stat Assoc, (submitted).

Wright S. 1951. The genetical structure of populations. Ann Eugen 15:323-354.

Wright S. 1969. Evolution and the genetics of populations. Vol 2: The theory of gene frequencies. Chicago: University of Chicago Press.

Zhu X, Elston RC. 2001. Transmission/disequilibrium tests for quantitative traits. Genet Epidemiol 20:57-74.

Table I. Performance of GC under the null model. For each of 6000 repetitions, $p_1 = p_2 = 0.2$, $n_h = 50$, $N = 2000$, $n_{strata} = 6$ for $F_{st} = 0.003$ and 2 for $F_{st} = 0.03$, and each hidden locus contributes 1% to the total heritability of 50%. Three models were fit, a single main effect ($k = 1$) (I), two main effects (II) and two main effects with an interaction (III).

unselected		coefs		multiplier			level ($\alpha = 0.05$)		
F_{st}	model	$\hat{\beta}_k$	$\hat{\beta}_{12}$	λ	λ_{int}	λ_{tot}	main	interaction	omnibus
	I	-0.0006	-	1.09	-	-	0.033	-	-
0.003	II	-0.0006	-	1.09	-	1.03	0.033	-	0.033
	III	-0.0009	0.007	1.05	1.00	1.02	0.037	0.038	0.034
	I	0.0006	-	1.84	-	-	0.047	-	-
0.03	II	0.0005	-	1.81	-	1.70	0.047	-	0.052
	III	0.0017	0.009	1.30	1.10	1.50	0.050	0.037	0.049
selected									
	I	-0.0010	-	1.03	-	-	0.041	-	-
0.003	II	-0.0009	-	1.03	-	1.01	0.041	-	0.043
	III	0.0017	-0.0035	1.00	1.03	1.04	0.040	0.038	0.38
	I	0.0007	-	1.71	-	-	0.050	-	-
0.03	II	0.0007	-	1.69	-	1.61	0.050	-	0.047
	III	0.0010	-0.0013	1.44	1.12	1.43	0.046	0.038	0.048

Table II. Performance of GC under the alternative model. For each of 200 repetitions, $p_1 = p_2 = 0.2, N = 2000$, and $n_{strata} = 6$. When $h^2 = 2.5\%$ $\beta_1 = 0.280$; when $h^2 = 5\%$ $\beta_1 = 0.395$; and for all conditions $\beta_2 = \beta_{12} = 0$. Each hidden locus k, $k = 3, \ldots, n_h + 2$, contributes 1% to achieve a total heritability of 50%. Three models were fit, a single main effect ($k = 1$) (I), two main effects (II) and two main effects with an interaction (III).

			unselected			selected		
h^2	F_{st}	model	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_{12}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_{12}$
		I	0.285	-	-	0.517	-	-
2.5	0.003	II	0.285	0.005	-	0.516	0.004	-
		III	0.289	0.007	-0.005	0.511	0.001	0.007
		I	0.283	-	-	0.511	-	-
2.5	0.03	II	0.283	0.002	-	0.511	-0.005	-
		III	0.282	0.004	-0.005	0.528	0.000	-0.012
		I	0.399	-	-	0.714	-	-
5	0.003	II	0.399	-0.004	-	0.714	-0.008	-
		III	0.398	-0.006	0.005	0.698	-0.019	0.027
		I	0.402	-	-	0.729	-	-
5	0.03	II	0.402	0.014	-	0.730	-0.010	-
		III	0.414	0.018	-0.012	0.748	-0.007	-0.009

Table III. Estimated sample size required to attain 80% power for the model described in Table II and $\alpha = 0.001$. Estimates were obtained using exemplary data methods (Longmate 2001). Entries for the main effect refer to the sample size required for the marginal one degree of freedom test for a main effect at locus 1, after accounting for other terms already in the model.

			unselected	
h^2	F_{st}	model	main	omnibus
2.5	0.003	I	569	-
		II	1371	1590
		III	3213	1615
2.5	0.03	I	869	-
		II	2228	2604
		III	3554	2494
5	0.003	I	427	-
		II	1123	1249
		III	2951	1376
5	0.03	I	553	-
		II	1742	2178
		III	3092	2050

Table IV. Performance of GC under the alternative model. For each of 200 repetitions, $p_1 = p_2 = 0.2$, $N = 2000$, and $n_{strata} = 6$. When $h^2 = 2.5\%$, $\beta_1 = \beta_2 = 0.280$, $\beta_{12} = 0.349$, and when $h^2 = 5\%$ $\beta_1 = \beta_2 = 0.395$, $\beta_{12} = 0.494$. Each hidden locus k, $k = 3, \ldots, n_h + 2$, contributes 1% to achieve a total heritability of 50%. Three models were fit, a single main effect ($k = 1$) (I), two main effects (II) and two main effects with an interaction (III).

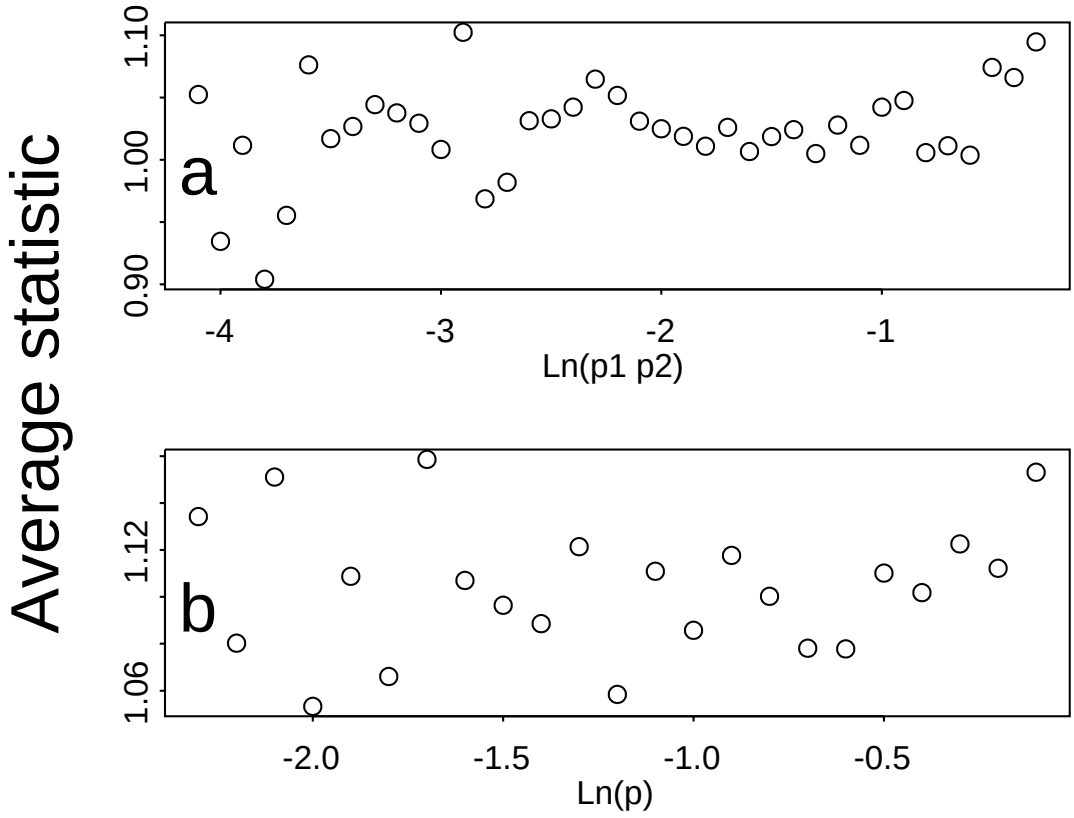
			unselected		selected	
h^2	F_{st}	model	$\hat{\beta}_k$	$\hat{\beta}_{12}$	$\hat{\beta}_k$	$\hat{\beta}_{12}$
2.5	0.003	I	0.419	-	0.762	-
		II	0.413	-	0.714	-
		III	0.270	0.367	0.545	0.382
2.5	0.03	I	0.413	-	0.745	-
		II	0.413	-	0.709	-
		III	0.275	0.350	0.537	0.392
5	0.003	I	0.593	-	1.048	-
		II	0.565	-	0.990	-
		III	0.401	0.490	0.765	0.494
5	0.03	I	0.585	-	1.056	-
		II	0.596	-	0.983	-
		III	0.400	0.500	0.760	0.484

Table V. Estimated sample size required to attain 80% power for the model described in Table IV and $\alpha = 0.001$. Entries for the main effect refer to the sample size required for the marginal one degree of freedom test for a main effect at either locus 1 or 2, after accounting for other terms already in the model. Likewise, entries for the interaction refer to the marginal one degree of freedom test for an interaction.

			unselected		
h^2	F_{st}	model	main	interaction	omnibus
2.5	0.003	I	569	-	-
		II	542	-	313
		III	1795	2093	306
2.5	0.03	I	869	-	-
		II	810	-	345
		III	2043	2246	328
5	0.003	I	427	-	-
		II	398	-	198
		III	1344	1669	144
5	0.03	I	553	-	-
		II	483	-	230
		III	1557	1718	230

Figure Caption

Figure 1. Average test statistic plotted as a function of the allele frequencies of the markers. (a) $\left[\hat{\beta}_{12}/SE_{ind}(\hat{\beta}_{12})\right]^2$ vs. $\ln(p_1 \times p_2)$; and (b) $\left[\hat{\beta}/SE_{ind}(\hat{\beta})\right]^2$ vs. $\ln(p)$.


Average statistic
a
1.10
1.00
0.90
-4
-3
-2
-1
Ln(p1 p2)
b
1.12
1.06
-2.0
-1.5
-1.0
-0.5
Ln(p)