Improving the power of GWAS and avoiding confounding from population stratification with PC-Select

George Tucker*, Alkes L Price†, Bonnie Berger*,†

April 28, 2014

*Department of Mathematics and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139
†Department of Epidemiology and Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115
†E-mail: bab@mit.edu

Copyright 2014.
Running Head: Avoiding confounding in GWAS with PC-Select

Key Words: Mixed Models, Population Stratification, GWAS

Corresponding Author:

Bonnie Berger

Department of Mathematics 2-373
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139

Tel: (617) 253-1827
Fax: (617) 258-5429
bab@mit.edu
Abstract

Using a reduced subset of SNPs in a linear mixed model can improve power for genome-wide association studies, yet this can result in insufficient correction for population stratification. We propose a hybrid approach using principal components that does not inflate statistics in the presence of population stratification and improves power over standard linear mixed models.
In recent years, there has been extensive research on linear mixed models (LMM) to calculate genome-wide association study (GWAS) association statistics (KANG et al. 2008; KANG et al. 2010; SEGURA et al. 2012; ZHOU and STEPHENS 2012; SVISHCHEVA et al. 2012; YANG et al. 2014). While linear mixed models implicitly assume that all SNPs have an effect on the phenotype (an infinitesimal genetic architecture), it is widely believed that disease phenotypes do not follow an infinitesimal model and that modeling a genetic architecture where most SNPs have negligible effect and some have modest effect (a non-infinitesimal genetic architecture) would increase power. As a step in that direction, Listgarten et al. (LISTGARTEN et al. 2012; LIPPERT et al. 2013) recently developed the state-of-the-art FaST-LMM Select method, which constructs a genetic relationship matrix (GRM) from a subset of top associated SNPs that are more likely to be causal. However, as a recent Perspective paper (YANG et al. 2014) shows, limiting the GRM to a subset of SNPs can result in insufficient correction for population stratification, leading to significantly inflated statistics and false positive associations (Tables 1, 2 and File S1: Supporting Text).

As a solution to this problem, we propose PC-Select, a novel hybrid approach that includes the principal components (PCs) of the genotype matrix as fixed effects in FaST-LMM Select. PC-Select leverages the advantages of the FaST-LMM Select framework while correcting for population stratification. The two main steps of FaST-LMM Select are ranking SNPs by linear regression p-values to form the GRM with the top ranked SNPs and then calculating association statistics in a mixed model framework using this GRM. We used the top 5 PCs* as fixed effects in both of these steps (See Methods). As a result, PC-Select yields non-inflated test statistics in the presence of population stratification and maintains high power to detect causal SNPs.

Specifically, to examine inflation and power, we followed the simulation procedure in YANG

---

*We follow the recommendations in the literature (PRICE et al. 2006) and use a fixed number of PCs. We have found that 5 PCs is generally sufficient to correct for stratification in simulated and real data sets. Alternatively, the number of PCs may be selected through cross validation or Tracy-Widom statistics (PATTERSON et al. 2006).
et al. (2014) and generated data sets each containing 10,000 SNPs for 1,000 individuals. To avoid a loss in power for LMM that can occur when candidate SNPs are included in the GRM† (Listgarten et al. 2012; Yang et al. 2014), we separately simulated a set of candidate SNPs to compute test statistics. We sampled individuals from two populations with $F_{st} = 0.05$, ancestral minor allele frequencies uniform in [0.1, 0.5], and mean phenotypic difference 0.25 standard deviations. To simulate causal SNPs in the GRM, we selected a fraction $p = 0.05$ or 0.005 of the SNPs at random and sampled Gaussian effect sizes (variance equal to 0.5 divided by the number of casual SNPs in the GRM) for these SNPs. We generated 500 candidate test null SNPs that were not causal, and to measure inflation we calculated $\lambda_{GC}$, the median Wald statistic on these SNPs divided by the theoretical median under the null distribution (Devlin and Roeder 1999). To investigate power, we generated 50 causal candidate SNPs with normally distributed effect sizes (variance equal to 0.5 divided by the number of causal candidate SNPs) and measured mean Wald statistic on these SNPs. We split the variability from causal SNPs evenly between the GRM and the causal candidate SNPs. We repeated all simulations 100 times and report the mean and standard error.

We found that when few SNPs were causal ($p = 0.005$), FaST-LMM Select inflated null statistics in the presence of population stratification ($\lambda_{GC} = 1.26 \pm 0.03$), whereas PC-Select was properly calibrated ($\lambda_{GC} = 1.01 \pm 0.01$) (Table 1). Moreover, FaST-LMM Select lost power in the presence of population stratification (as measured by the mean Wald statistic on causal SNPs: 14.3 ± 0.2 with stratification versus 16.4 ± 0.1 without), whereas PC-Select’s power in simulations with and without population stratification was not significantly different (16.3 ± 0.1 versus 16.3 ± 0.1) (Figure 1). Thus, even though PC-Select corrected for stratification, this advantage did not come at the expense of power. This gain is likely because the PCs reduce noise in selecting subsets of SNPs for the GRM in the presence of population stratification. In addition,

†Both PC-Select and FaST-LMM Select avoid this by removing the candidate SNP and nearby SNPs from the GRM when computing the association statistic.
Table 1. Extent of null statistic inflation as measured by $\lambda_{GC}$ (median Wald statistic on test null SNPs divided by the theoretical median under the null distribution (DEVLIN et al. 1999)). We tabulate $\lambda_{GC}$ for linear regression, linear regression with PCs, standard LMM, FaST-LMM Select, and PC-Select on simulated genotypes and phenotypes with and without population stratification as the fraction of casual SNPs ($p = 0.05$, 0.005) varies. Values shown are mean $\lambda_{GC}$ over 100 simulations with standard error in parenthesis. FaST-LMM Select inflates statistics in the presence of population stratification when few SNPs are causal ($p = 0.005$), which may result in false positives.

PC-Select chose fewer SNPs than FaST-LMM Select to include in the GRM (over 100 simulations, mean SNPs chosen: $\sim$20 versus $\sim$240, Figure S1), yielding potential computational savings. When many SNPs were causal ($p = 0.05$), both methods used nearly all SNPs in the GRM (over 100 simulations, mean SNPs chosen: $\sim$9, 400 and $\sim$8, 800 out of 10, 000, respectively), achieving similar performance to standard LMM.

We also investigated a recent extension of FaST-LMM Select, the genard method (HOFFMAN 2013) that fits a data-adaptive low-rank GRM; however, we found that it did not have increased power over LMM in our simulations (Figure S5), which is consistent with previous simulations in a similar context (HOFFMAN 2013).

Next, we evaluated inflation and power on real genotypes with simulated phenotypes in a similar manner. We analyzed 5,000 individuals randomly subsampled from a multiple sclerosis (MS) study genotyped on Illumina arrays (SAWCER et al. 2011) made available via Welcome Trust Case Control Consortium 2 (WTCCC2) (See Methods). As before, we separated GRM SNPs...
Figure 1. Comparison of power for linear regression, linear regression with PCs, standard LMM, FaST-LMM Select, and PC-Select on simulated genotypes and phenotypes (A) and real genotypes and simulated phenotypes (B) with and without population stratification as the fraction of casual SNPs ($p = 0.05, 0.005$) varies. To measure power, we plot the mean Wald statistic on test causal SNPs. In all cases, PC-Select has the highest power of the methods that do not inflate statistics.
and candidate SNPs to avoid proximal contamination and provide a fair comparison of methods. We randomly sampled 50,000 SNPs for the GRM from chromosomes 3 to 22, 250 causal SNPs from chromosome 1, and 500 null SNPs from chromosome 2. To simulate environmental variance aligned with population structure, we added 0.25 times the first PC (after the PC had been normalized to variance 1) to each individual’s phenotype. Otherwise, we generated phenotypes as before and report simulations over 200 randomly generated phenotypes.

We again found that when few SNPs were causal \((p = 0.005)\), FaST-LMM Select inflated null statistics in the presence of population stratification \((\lambda_{GC} = 1.06 \pm 0.01)\), whereas PC-Select was properly calibrated \((\lambda_{GC} = 1.01 \pm 0.01)\) (Table 2). Moreover, FaST-LMM Select lost power in the presence of population stratification (as measured by the mean Wald statistic on causal SNPs: \(14.64 \pm 0.05\) with stratification versus \(16.02 \pm 0.05\) without); in contrast, PC-Select’s power in simulations with and without population stratification was not significantly different \((16.02 \pm 0.05\) versus \(16.08 \pm 0.05)\) (Figure 1). In all of our simulations, PC-Select produced non-inflated statistics and high power.

<table>
<thead>
<tr>
<th></th>
<th>pop. strat.</th>
<th>pop. strat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression</td>
<td>(1.58 (0.02))</td>
<td>(1.55 (0.02))</td>
</tr>
<tr>
<td>Linear reg. with PCs</td>
<td>(1.01 (0.01))</td>
<td>(1.00 (0.01))</td>
</tr>
<tr>
<td>LMM</td>
<td>(1.02 (0.01))</td>
<td>(1.01 (0.01))</td>
</tr>
<tr>
<td>FaST-LMM Select</td>
<td>(1.02 (0.01))</td>
<td>(1.06 (0.01))</td>
</tr>
<tr>
<td>PC-Select</td>
<td>(1.01 (0.01))</td>
<td>(1.01 (0.01))</td>
</tr>
</tbody>
</table>

**Table 2.** Extent of null statistic inflation as measured by \(\lambda_{GC}\). We tabulate \(\lambda_{GC}\) for linear regression, linear regression with PCs, standard LMM, FaST-LMM Select, and PC-Select on real genotypes and simulated phenotypes with and without population stratification as the fraction of casual SNPs \((p = 0.05, 0.005)\) varies. Values shown are mean \(\lambda_{GC}\) over 200 simulations with standard error in parenthesis. FaST-LMM Select inflates statistics in the presence of population stratification when few SNPs are causal \((p = 0.005)\), which may result in false positives.
Finally, we analyzed data from 10,204 MS cases and 5,429 controls genotyped on Illumina arrays (SAWCER et al. 2011) made available via WTCCC2 (See Methods). The cases and controls were not matched for ancestry and thus exhibited substantial population stratification. Evaluated over all SNPs, PC-Select had $\lambda_{GC} = 1.24$ and FaST-LMM Select had $\lambda_{GC} = 1.20$. Due to polygenicity, we expect $\lambda_{GC}$ on all markers to be larger than 1. On the same data, YANG et al. (2014) report $\lambda_{GC} = 1.23$ and 1.20 for linear regression with PCs and LMM, respectively, which they show is consistent with polygenicity. To evaluate power, we considered Wald statistics at 75 known associated SNPs (See Methods, Table S1 for Wald statistics). PC-Select consistently gave larger Wald statistics than FaST-LMM Select (63 of 75 markers; $P = 2 \times 10^{-9}$, mean Wald statistic 12.07 versus 11.30). Based on cross-validation, both PC-Select and FaST-LMM Select chose to use all markers. This may indicate that the disease is not caused by a small number of loci with large effects or that our sample size is too small to capture this effect. Although, PC-Select and FaST-LMM Select chose to use all SNPs and thus neither method inflated statistics, we emphasize that without a priori knowledge about the genetic architecture, PC-Select automatically tunes the number of SNPs to include in the GRM to optimize power and simultaneously protects against population stratification at no cost to power.

Janss et al. caution against using PCs as fixed effects in combination with a random effect derived from the GRM when estimating heritability (JANSS et al. 2012). This may result in an ill-posed model because the PCs enter both as fixed effects and implicitly through the random effect. We avoid this issue when estimating variance components by using the PCs as fixed effects in a restricted maximum likelihood (REML) approach, which projects the genotype matrix into a subspace orthogonal to the PCs, effectively removing them from the random effect. We also note that population structure and PCs have previously been used successfully as fixed effects (or separate random effects) in mixed model settings to address confounding from population structure and from unusually differentiated markers (YU et al. 2006; ZHAO et al. 2007; PRICE et al. 2010; SUL and ESKIN 2013; PRICE et al. 2013).

Using PCs in a linear model does not correct for family relatedness and cryptic relatedness
(PRICE et al. 2010). As suggested by YANG et al. (2014), due to the large length of segments shared identical-by-descent, using a subset of SNPs may correct for cryptic relatedness. LISTGARTEN et al. (2012) show that using a subset of SNPs in the GRM does not inflate statistics on the WTCCC data, where inflation is likely primarily due to cryptic relatedness. We expect that PC-Select will not be inflated by cryptic relatedness for the same reasons. In most human data sets with unrelated individuals, family relatedness is not an issue; however, for data sets with strong family relatedness, we suspect there may be cases where both PC-Select and FaST-LMM Select inflate statistics.

PC-Select has the same asymptotic run-time as FaST-LMM Select, quadratic in the number of individuals and linear in the number of markers. In practice, the run-time for the additional step of computing the PCs for the genotype matrix is minimal because both methods require several spectral decompositions of matrices of nearly the same size for the cross-validation step. It should be noted that while the asymptotic run-time of PC-Select and FaST-LMM Select is the same as previously published exact LMM methods (LIPPERT et al. 2011; ZHOU and STEPHENS 2012), the actual run-time of both methods is ostensibly longer by a factor of 10 due to the cross-validation step. The cross-validation step is parallelizable, so in practice this is not a significant limitation.

Including PCs as fixed effects allows PC-Select to infer ancestry from all SNPs simultaneously, while at the same time maintaining the benefits of using a statistically-chosen subset of the SNPs to estimate the GRM (LISTGARTEN et al. 2012; LIPPERT et al. 2013). As we have shown, using a combination of PCs and a subset of SNPs in the GRM gives the best of both worlds.

METHODS

MS dataset: We analyzed data from 10,204 MS cases and 5,429 controls (from NBS and 1958BC) genotyped on Illumina arrays made available to researchers via WTCCC2 (http://wtccc.org.uk/ccc2/). We follow the quality control standards in YANG et al. (2014). Although SAWCER et al. (2011) analyzed UK and non-UK samples separately followed by meta-analysis in most of their analyses, the data made available to researchers includes both UK and non-UK cases.
but only UK controls. We retained all samples in order to maximize sample size. We considered markers that were present in each of MS, NBS and 1958 BC datasets and removed markers with > 0.5% missing data, $P < 0.01$ for allele frequency difference between NBS and 1958BC, $P < 0.05$ for deviation from Hardy-Weinberg equilibrium, $P < 0.05$ for differential missingness between cases and controls, or $MAF < 0.1\%$ in any dataset, leaving 360,557 markers. The 75 known associated markers were defined by including, for each MS-associated marker listed in the NHGRI GWAS catalogue (http://genome.gov/gwastudies/), a single best tag at $r^2 > 0.4$ from the set of 360,557 markers if available.

**Statistical methods:** PC-Select follows a similar framework as FaST-LMM Select (Lipper et al. 2011; Listgarten et al. 2012; Lipper et al. 2013). For completeness, we list the steps and equations we used.

First, we describe a method for computing association statistics, then in subsequent sections we describe the steps of PC-Select.

**Association statistics:** The phenotype $y$, covariates $X$, and genotypes $W$ are mean centered. Additionally, each genotype is divided by $\sqrt{2\hat{p}(1-\hat{p})}$ where $\hat{p}$ is the estimated minor allele frequency. Then the phenotype is modeled as

$$y = X\alpha + u + \epsilon,$$

where $u \sim N(0, \sigma_u^2 K)$, $\epsilon \sim N(0, \sigma_e^2 I)$, $\alpha$ is a vector of weights for the covariates, and $K$ is the GRM. This model naturally leads to an association statistic based on the Wald statistic.

To calculate the association statistic for SNP $w$, we add $w$ as a fixed effect covariate to the previous model and test whether its coefficient is significantly different than 0. Specifically, consider the model

$$y = w\beta + X\alpha + u + \epsilon,$$

where $\beta$ is the coefficient for the test SNP. We estimate $\sigma_u^2$ and $\sigma_e^2$ by REML. The fixed effect coefficients $(\beta, \alpha)$ are estimated by maximum likelihood.
It is straightforward to construct the Wald statistic to test whether $\beta \neq 0$. Let $V = \sigma_g^2 K + \sigma_e^2 I$ and $Q = [w; X]$. Then $\hat{\beta}$ is equal to the first entry of $(Q^T V^{-1} Q)^{-1} Q^T V^{-1} y$ and $\text{var}(\hat{\beta})$ is equal to the first entry of $(Q^T V^{-1} Q)^{-1}$. The test statistic is

$$\frac{\hat{\beta}^2}{\text{var}(\hat{\beta})},$$

which is asymptotically $\chi^2$ distributed with 1 degree of freedom.

Now we describe the PC-Select method:

**Step 1: Extracting PCs:** We extract the top 5 PCs from a GRM formed using all of the genotype data, $WW^T$, to use as fixed effect covariates. We use $X$ to denote the matrix of user specified covariates and the top 5 PCs.

**Step 2: Ranking SNPs by linear regression:** Second, we rank the SNPs by a linear regression test statistic. Linear regression test statistics are calculated by fixing $\sigma_g^2$ to 0 and using the procedure described above to calculate Wald statistics.

**Step 3: Determining the GRM:** As in FaST-LMM Select, PC-Select uses a subset of the SNPs that are likely to be causal. In this step, we determine $k$, the number of top SNPs (as ranked in Step 2) to include in the GRM. We use 10-fold cross-validation on predictive log-likelihood to choose the number of top SNPs.

We choose $k$ from a list of user defined possibilities (e.g., $k \in \{100, 1000, 3000, 10,000, 30,000, \ldots\}$). First, we randomly divide individuals into 10 equal groups or folds. For each fold $i$, we form a test set from the individuals in fold $i$ and use the rest of the individuals as a training set. For each choice of $k$, we consider a subset of the genotype matrix consisting only of the top $k$ SNPs (the ranking of the SNPs is recomputed per fold using the training data). For notational simplicity, we will also refer to the reduced genotype matrix by $W$, and it will be clear from context if this refers to the full genotype matrix or a subset. Let $W_i$ denote the genotypes from fold $i$ and $W_{-i}$ represent the genotypes from the rest of the folds (similarly for $y$ and $X$). We wish to evaluate the
predictive log-likelihood of $y_i$ given the training information $(y_{-i}, X_{-i}, X_i)$ to assess the predictive power of using only the top $k$ SNPs in the GRM.

Specifically, to evaluate the predictive log-likelihood, we start by forming a GRM from the training set $W_{-i} W_{-i}^T$. Then we estimate $\sigma_g^2$ and $\sigma_e^2$ from the training set by REML. We estimate $\alpha$ by ML with these variance parameters fixed. Then under the model

$$y = X\alpha + u + \epsilon$$

where $u \sim N(0, \sigma_g^2 W W^T)$ and $\epsilon \sim N(0, \sigma_e^2 I)$, the predictive distribution of the phenotypes given the training parameters, $y_i | y_{-i}, W, \alpha, \sigma_g^2, \sigma_e^2$, is normally distributed with mean

$$\sigma_g^2 W_i W_{-i}^T (W_{-i} W_{-i}^T \sigma_g^2 + \sigma_e^2 I)^{-1} (y_{-i} - X_{-i} \alpha) + X_i \alpha$$

and covariance

$$W_i W_{-i}^T \sigma_g^2 + \sigma_e^2 I - \sigma_g^2 W_i W_{-i}^T (W_{-i} W_{-i}^T \sigma_g^2 + \sigma_e^2 I)^{-1} W_{-i} W_{-i}^T \sigma_g^2.$$

This can be evaluated efficiently using the spectral decompositions computed in the REML step (LIPPERT et al. 2011; LISTGARTEN et al. 2012). We average the predictive log-likelihood over each of the 10 folds and choose the $k$ that gives the highest average log-likelihood.

**Step 4: Calculating association statistics:** Finally, with the number of top SNPs to use in the GRM fixed, we calculate association statistics for each SNP. Let $W$ be the genotype matrix using the top $k$ SNPs chosen in the previous step. To avoid proximal contamination (LISTGARTEN et al. 2012), we use a leave-one-chromosome-out procedure (YANG et al. 2014). For each test SNP $w$ (which is not necessarily in $W$), we exclude the chromosome including that SNP from the GRM and calculate the Wald statistic for $w$ with this GRM. We do this efficiently by precomputing and storing the GRM excluding each chromosome in turn.

**ACKNOWLEDGMENTS**

We thank Po-Ru Loh, Sean Simmons, and Jian Peng for helpful discussions. This study makes use of data generated by the Wellcome Trust Case Control Consortium. This research was funded by
National Institute of Health (NIH) grants R01 GM108348 and R01 HG006399.

LITERATURE CITED


PRICE, A. L., N. J. PATTERSON, R. M. PLENGE, M. E. WEINBLATT, N. A. SHADICK, and
D. REICH, 2006 Principal components analysis corrects for stratification in genome-wide

PRICE, A. L., N. A. ZAITLEN, D. REICH, and N. PATTERSON, 2010 New approaches to popu-
463.


SAWCER, S., G. HELLENTHAL, M. PIRINEN, C. SPENCER, N. PATSOPoulos, L. MOUT-
SIANAS, A. DILTHEY, Z. SU, C. FREEMAN, S. HUNT, and OTHERS, 2011 Genetic risk and
a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476(7359):

SEGURA, V., B. J. VILHJ ´ALMSSON, A. PLATT, A. KORTE, ¨U. SEREN, Q. LONG, and M. NORD-
BORG, 2012 An efficient multi-locus mixed-model approach for genome-wide association

SUL, J. H. and E. ESKIN, 2013 Mixed models can correct for population structure for genomic

SVISHCHEVA, G. R., T. I. AXENOVICH, N. M. BELONOGOVA, C. M. VAN DUIJN, and Y. S.
AULCHENKO, 2012 Rapid variance components-based method for whole-genome association

YANG, J., S. H. LEE, M. E. GODDARD, and P. M. VISSCHER, 2011 GCTA: a tool for genome-

YANG, J., N. A. ZAITLEN, M. E. GODDARD, P. M. VISSCHER, and A. L. PRICE, 2014 Ad-
vantages and pitfalls in the application of mixed-model association methods. Nature Genet-
ics 46(2): 100–106.
