bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 Genome-Wide Control of Population Structure and Relatedness in Genetic Association Studies
2 via Linear Mixed Models with Orthogonally Partitioned Structure

3

4 Matthew P. Conomos[1], Alex P. Reiner[2,3], Mary Sara McPeek[4,5], Timothy A. Thornton[1,6]

5

6 [1] Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
7 [2] Department of Epidemiology, University of Washington, Seattle, WA 98195, USA
8 [3]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109,
9 USA;
10 [4]Department of Statistics, University of Chicago, Chicago, IL 60637, USA
11 [5]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA
12 [6]Department of Statistics, University of Washington, Seattle, WA 98195, USA
13 Correspondence should be addressed to T.A.T (tathornt@uw.edu) or M.P.C.
14 (mconomos@uw.edu).

15
16
17
18
19
20
21
22
23
24

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

25

26 **Abstract**

27 Linear mixed models (LMMs) have become the standard approach for genetic association
28 testing in the presence of sample structure. However, the performance of LMMs has primarily
29 been evaluated in relatively homogeneous populations of European ancestry, despite many of
30 the recent genetic association studies including samples from worldwide populations with
31 diverse ancestries. In this paper, we demonstrate that existing LMM methods can have
32 systematic miscalibration of association test statistics genome-wide in samples with
33 heterogenous ancestry, resulting in both increased type-I error rates and a loss of power.
34 Furthermore, we show that this miscalibration arises due to varying allele frequency differences
35 across the genome among populations. To overcome this problem, we developed LMM-OPS, an
36 LMM approach which orthogonally partitions diverse genetic structure into two components:
37 distant population structure and recent genetic relatedness. In simulation studies with real and
38 simulated genotype data, we demonstrate that LMM-OPS is appropriately calibrated in the
39 presence of ancestry heterogeneity and outperforms existing LMM approaches, including
40 EMMAX, GCTA, and GEMMA. We conduct a GWAS of white blood cell (WBC) count in an
41 admixed sample of 3,551 Hispanic/Latino American women from the Women's Health Initiative
42 SNP Health Association Resource where LMM-OPS detects genome-wide significant associations
43 with corresponding p-values that are one or more orders of magnitude smaller than those from
44 competing LMM methods. We also identify a genome-wide significant association with
45 regulatory variant rs2814778 in the DARC gene on chromosome 1, which generalizes to
46 Hispanic/Latino Americans a previous association with reduced WBC count identified in African
47 Americans.

48

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

49 **Introduction**

50 The complete genealogy of individuals consists of recent genetic relatedness, such as pedigree
51 relationships of family members, as well as more distant genetic relatedness, such as that due to
52 population structure. In genetic association studies, it is well known that failure to appropriately
53 account for either recent or distant genetic relatedness among sampled individuals can result in
54 spurious association. To address this, linear mixed models (LMMs) have emerged as the
55 standard approach for genetic association testing in samples with population structure, family
56 structure, and/or cryptic relatedness[1-10]. Existing LMM implementations developed for GWAS
57 model the entire genealogy of sampled individuals as a random effect, with the covariance
58 structure of the phenotype specified by a single empirical genetic relationship matrix (GRM)[11-13].
59 This approach typically provides an acceptable genomic control inflation factor[14], which is
60 evaluated based on the median of the test statistics across all SNPs genome-wide. However, in
61 the presence of population stratification, previous studies[15,16] have shown that there may be
62 SNPs for which type-I error rates are not properly controlled, such as those SNPs with unusually
63 large allele frequency differences between populations.

64 Here, we utilize SNP genotyping data from release 3 of phase III of the International
65 Haplotype Map Project (HapMap)[17] to demonstrate that existing LMM approaches provide
66 miscalibrated association test statistics when phenotypes are correlated with ancestry. This
67 miscalibration arises due to variation across the genome in allele frequency differences between
68 the populations from which the sampled individuals descend, and we show that it impacts all
69 SNPs genome-wide, not only those with unusually large allele frequency differences. While
70 standard LMM approaches appropriately control type-I error rates at SNPs with typical allele
71 frequency differences, there is systematically inflated or deflated test statistics for SNPs with
72 greater or smaller differences, respectively. Interestingly, we demonstrate that this pattern of

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

73 test statistic inflation/deflation can occur not only in samples with continental ancestry
74 differences, but also in samples with subtle or fine-scale population structure. Furthermore, the
75 miscalibration of test statistics is observed for LMM methods that estimate variance
76 components once per genome screen, such as EMMAX[3] and GCTA[13], as well as those that re-
77 estimate variance components for every tested variant, such as GEMMA[8].

78 To address the shortfalls of existing LMM methods, we propose a linear mixed model
79 with orthogonally partitioned structure (LMM-OPS) method for genetic association testing of
80 quantitative traits in samples with diverse ancestries. LMM-OPS appropriately accounts for
81 variable population allele frequency differentiation across the genome to provide well-
82 calibrated association test statistics at *all* SNPs genome-wide. With LMM-OPS, genetic sample
83 structure is orthogonally partitioned into two separate components: a component for the
84 sharing of alleles inherited identical by descent (IBD) from recent common ancestors, which
85 represents familial relatedness, and another component for allele sharing due to more distant
86 common ancestry, which represents population structure. LMM-OPS models population
87 structure as a fixed effect by including vectors that are representative of genome-wide ancestry
88 (e.g. principal components (PCs) or admixture proportions calculated from genome-wide data)
89 as covariates, while recent genetic relatedness among individuals is modeled using a random
90 effect, with covariance structure specified by an ancestry-adjusted empirical GRM. An important
91 feature of the GRM used by LMM-OPS is that it is constructed to be orthogonal to the ancestry-
92 representative vectors that are included as fixed effects. This ancestry-adjusted GRM measures
93 the residual genetic covariance among sampled individuals, after adjusting for ancestry, as a
94 way of capturing only recent genetic relatedness. As a result, the ancestry-adjusted GRM and
95 the ancestry-representative vectors represent orthogonal information on sample structure, and

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

96 LMM-OPS avoids issues of double-fitting information in both the fixed and random effects,
97 which could lead to over-correction of sample structure and a loss of power[6,10].
98 We conduct simulation studies to demonstrate that LMM-OPS effectively accounts for
99 complex sample genealogy, including population stratification, ancestry admixture, and familial
100 relatedness, resulting in proper control of type-I error rates at all SNPs, as well as increased
101 power over existing LMM methods for detecting genetic association. We also apply LMM-OPS
102 and the LMM methods implemented in EMMAX[3], GEMMA[8], and GCTA[13] to a GWAS of white
103 blood cell (WBC) count in a sample of 3,551 Hispanic American postmenopausal women from
104 the Women's Health Initiative SNP Health Association Resource (WHI-SHARe) study[18,19]. The
105 WHI-SHARe Hispanics have complex sample structure, including continental and sub-continental
106 population structure as well as cryptic familial relatedness[20]. Consistent with our simulation
107 study results, the LMM-OPS *p*-values for genome-wide significant SNPs are one or more orders
108 of magnitude smaller than those from the competing LMM methods. Based on our analysis, we
109 replicate[21] and generalize to Hispanic/Latino Americans a genome-wide significant association
110 with regulatory variant rs2814778 in the Duffy Antigen Receptor for Chemokines (DARC) gene
111 that was previously found to associate with lower WBC count in African Americans[22,23].
112
113 **Materials and Methods**
114 **Standard Empirical GRM**
115 A genetic relationship matrix (GRM), Ψ, measures a weighted covariance of genotypes,
116 averaged over all SNPs across the genome, between each pair of individuals. Consider a set $\mathcal{N}$
117 of sampled individuals that have been genotyped at a set $\mathcal{S}$ of SNP genotype markers. A
118 standard empirical estimator[13] of a GRM that is widely used scales the contribution of each SNP
119 by the sample genotype variance under HWE and has $[i, j]^{th}$ element

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

120 $$\hat{\psi}_{ij} = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \frac{(x_{is}-2\bar{p}_s)(x_{js}-2\bar{p}_s)}{2\bar{p}_s(1-\bar{p}_s)}, \quad (1)$$

121 where $|\mathcal{S}|$ is the number of SNPs in the set $\mathcal{S}$, x_{is} is the genotype value for individual i at

122 SNP s, and $\bar{p}_s = \frac{1}{2}\widehat{\mathbb{E}}[x_{is}] = \frac{1}{2|\mathcal{N}|}\Sigma_{i\in\mathcal{N}} x_{is}$ is the sample average allele frequency at SNP s, as

123 $|\mathcal{N}|$ is the number of sampled individuals. The genotype covariance structure captured by $\hat{\Psi}$,

124 the empirical GRM constructed using the estimator $\hat{\psi}_{ij}$ in Equation (1), includes contributions

125 from both distant population structure and recent familial relatedness[24].

126

127 **Construction of an Empirical GRM Orthogonal to Genome-wide Ancestry**

128 Similar to the aforementioned standard GRM, an ancestry-adjusted GRM, Φ, also measures a

129 weighted covariance of genotypes, averaged over all SNPs across the genome, between each

130 pair of individuals, however with the covariance is obtained conditional on the genome-wide

131 ancestries of the sampled individuals. Let V be an $|\mathcal{N}|\times(k+1)$ matrix whose column vectors

132 include an intercept and k ancestry-representative vectors (e.g. principal components (PCs) or

133 admixture proportions calculated from genome-wide data). One empirical estimator of an

134 ancestry-adjusted GRM has $[i,j]^{th}$ element

135 $$\tilde{\phi}_{ij} = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} (x_{is}-2\hat{\mu}_{is})(x_{js}-2\hat{\mu}_{js}), \quad (2)$$

136 where $\hat{\mu}_{is}$ is the ith element of the vector $\hat{\mu}_s = \frac{1}{2}\widehat{\mathbb{E}}[\mathbf{x}_s|\mathrm{V}] = \frac{1}{2}[\mathrm{V}(\mathrm{V}^T\mathrm{V})^{-1}\mathrm{V}^T\mathbf{x}_s]$ of fitted

137 values from a linear regression of $\mathbf{x}_s$, the genotype values for all individuals at SNP s, on V.

138 To see that the ancestry-adjusted empirical GRM, $\tilde{\Phi}$, constructed using the estimator $\tilde{\phi}_{ij}$ in

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

139 Equation (2) is orthogonal to genome-wide ancestry, let R be an $|\mathcal{N}|\times|\mathcal{S}|$ matrix whose s^{th}
140 column vector is the residual vector from the linear regression of $\mathbf{x}_s$ on V; i.e. $(\mathbf{x}_s - 2\hat{\mu}_s)$.
141 Because the residuals from a linear regression are orthogonal to the predictors, $\mathsf{V}^T\mathsf{R} = \mathbf{0}$, and
142 since the ancestry adjusted empirical GRM can be written as $\tilde{\Phi} = \frac{1}{|\mathcal{S}|}\mathsf{R}\mathsf{R}^T$, we have that
143 $\mathsf{V}^T\tilde{\Phi} = \frac{1}{|\mathcal{S}|}\mathsf{V}^T\mathsf{R}\mathsf{R}^T = \mathbf{0}\mathsf{R}^T = \mathbf{0}$, indicating orthogonality of V and $\tilde{\Phi}$. If V fully captures the
144 population structure in the sample, then the genotype covariance structure represented by $\tilde{\Phi}$
145 only includes that due to the sharing of alleles IBD from recent common ancestors; i.e. recent
146 familial relatedness[24].

147 A potential limitation with $\tilde{\Phi}$, which we refer to as the 'centered only' ancestry-
148 adjusted empirical GRM, is that its elements have no meaningful biological interpretation
149 without scaling. To address this, an alternative ancestry-adjusted empirical GRM, $\hat{\Phi}$, can be
150 obtained using the PC-Relate method[24], where the $[i,j]^{th}$ element of this matrix is

151 $$\hat{\phi}_{ij} = \frac{\Sigma_{s\in\mathcal{S}}(x_{is} - 2\hat{\mu}_{is})(x_{js} - 2\hat{\mu}_{js})}{\Sigma_{s\in\mathcal{S}}[2\hat{\mu}_{is}(1-\hat{\mu}_{is})]^{1/2}[2\hat{\mu}_{js}(1-\hat{\mu}_{js})]^{1/2}}, \quad (3)$$

152 which is an estimator of twice the kinship coefficient for the pair of individuals i and j. We
153 refer to $\hat{\Phi}$ as either the 'PC-Relate' or the 'centered and standardized' ancestry-adjusted
154 empirical GRM. While $\hat{\Phi}$ has improved biological interpretability over $\tilde{\Phi}$, it is no longer
155 strictly orthogonal to the ancestry-representative vectors V because the scaling factor in the
156 denominator of Equation (3) depends on i and j. However, in practice we have found that the
157 scaling factors for each pair of individuals are generally similar, as they are computed as an

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

158 average across all SNPs in $\mathcal{S}$, and the elements of $\hat{\Phi}$ and $\tilde{\Phi}$ are very highly correlated (see
159 Results), indicating that $\hat{\Phi}$ is approximately orthogonal to V.
160

161 **Linear Mixed Models for GWAS**

162 A standard linear mixed model (LMM) used in GWAS to test for genetic association at SNP
163 $s' \in \mathcal{S}$ can be written as

164
$$\begin{aligned} \mathbf{Y} &= \mathrm{W}\alpha + \mathbf{x}_{s'}\beta_{s'} + \mathbf{g} + \epsilon \\ \mathbf{g} &\sim \mathrm{N}(\mathbf{0}, \sigma_A^2 \Psi) \\ \epsilon &\sim \mathrm{N}(\mathbf{0}, \sigma_\epsilon^2 \mathrm{I}_{|\mathcal{N}|}) \end{aligned}, \quad (4)$$

165 where $\mathbf{Y}$ is a vector of phenotype values for all individuals, W is a matrix of covariates
166 including an intercept, α is a corresponding vector of effect sizes, $\mathbf{x}_{s'}$ is the vector of genotype
167 values for all individuals at SNP s', $\beta_{s'}$ is the effect size of SNP s', $\mathbf{g}$ is a random effect that
168 captures the polygenic effect of other SNPs, σ_A^2 is a parameter that measures the additive
169 genetic variance of the phenotype, Ψ is the standard genetic relationship matrix (GRM)[25], ϵ is
170 a random effect that captures independent residual effects, σ_ϵ^2 is a parameter that measures
171 residual variance, and $\mathrm{I}_{|\mathcal{N}|}$ is an identity matrix. Generalized least squares (GLS) can be used to
172 fit the LMM in Equation (4) and test the null hypothesis that $\beta_{s'} = 0$; however, the overall
173 covariance structure of the phenotype, $\mathrm{Cov}[\mathbf{Y}] \equiv \sigma_A^2 \Psi + \sigma_\epsilon^2 \mathrm{I}_{|\mathcal{N}|}$, is unknown in practice and
174 must first be estimated. In order to do so, an empirical GRM, such as $\hat{\Psi}$ with $[i,j]^{th}$ element
175 given by Equation (1), is estimated from the available SNP data. Utilizing this empirical GRM in
176 the null model (i.e. the model with $\beta_{s'}$ fixed at 0), estimates of the variance components $\hat{\sigma}_A^2$

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

177 and $\hat{\sigma}^2_\epsilon$ are obtained, typically with restricted maximum likelihood (REML). GLS can then be

178 performed using the estimate of the overall phenotypic covariance structure,

179 $\widehat{\text{Cov}}[\mathbf{Y}] = \hat{\sigma}^2_A \hat{\Psi} + \hat{\sigma}^2_\epsilon \mathbf{I}_{|\mathcal{N}|}$.

180

181 **The LMM-OPS Model**

182 The LMM-OPS model that we propose has a similar form to the LMM presented in Equation (4),

183 but with the genealogical structure of the sample orthogonally partitioned into fixed and

184 random effects. Population structure is adjusted for as a fixed effect, and recent genetic

185 relatedness is accounted for as a random effect. The LMM-OPS model can be written as

186
$$\begin{aligned} \mathbf{Y} &= \mathrm{W}\alpha + \mathrm{V}\gamma + \mathbf{x}_{s'}\beta_{s'} + \mathbf{g} + \epsilon \\ \mathbf{g} &\sim \mathrm{N}(\mathbf{0}, \sigma^2_A \Phi) \\ \epsilon &\sim \mathrm{N}(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}_{|\mathcal{N}|}) \end{aligned} \quad . \qquad (5)$$

187 The differences in the LMM-OPS model in Equation (5) from the standard LMM model in

188 Equation (4) are that it includes V, the matrix of ancestry-representative vectors with

189 corresponding effect sizes γ, in the mean model to adjust for population structure, and it uses

190 an ancestry-adjusted GRM, Φ, that only measures recent familial relatedness, in place of the

191 standard GRM, Ψ. Therefore, with population structure modeled as a fixed effect in the mean,

192 the overall covariance structure of the phenotype in LMM-OPS model is given by

193 $\text{Cov}[\mathbf{Y}] \equiv \sigma^2_A \Phi + \sigma^2_\epsilon \mathbf{I}_{|\mathcal{N}|}$. As with the standard LMM in the previous section, GLS can be used to

194 fit the LMM-OPS model and test for genetic association. The procedure is identical, except that

195 $\hat{\sigma}^2_A$, $\hat{\sigma}^2_\epsilon$, and $\widehat{\text{Cov}}[\mathbf{Y}]$ are obtained utilizing an ancestry-adjusted empirical GRM estimated

196 from the available SNP data. Either the 'centered only' ancestry-adjusted empirical GRM, $\tilde{\Phi}$,

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

197 with $[i,j]^{th}$ element given by Equation (2), or the 'centered and standardized' ancestry-adjusted
198 empirical GRM, $\hat{\Phi}$, with $[i,j]^{th}$ element given by Equation (3), can be used for LMM-OPS.
199 Throughout the remainder of this manuscript, unless specified otherwise, we use the centered
200 and standardized ancestry-adjusted empirical GRM when presenting LMM-OPS results.
201
202 **Simulation Studies**
203 In all simulation studies, association testing was performed using LMM-OPS, EMMAX, GCTA,
204 GEMMA, and linear regression adjusted for PCs. LMM-OPS included the top PC from PC-AiR[26] as
205 a fixed effect to adjust for ancestry in the mean model, and it used an ancestry-adjusted
206 empirical GRM constructed with PC-Relate[24] to account for correlation among genotypes due to
207 recent genetic relatedness. All analyses with EMMAX, GCTA and GEMMA used the default
208 genetic relationship matrices implemented in their respective software to account for sample
209 structure. Details are provided in Appendix A. Throughout the simulation studies, EMMAX and
210 GCTA gave nearly identical results; therefore, only those from EMMAX are presented. Linear
211 regression adjusted for PCs used the top PC from PC-AiR, rather than EIGENSTRAT[27], to ensure
212 that ancestry was accurately captured and not confounded by pedigree structure[26].
213 Simulation studies were used to investigate the impact of variation across the genome
214 in allele frequency differences between populations on association test statistics at null SNPs.
215 Two simulation studies were conducted using samples from two different pairs of HapMap
216 populations: (1) the closely related CEU (Utah residents with Northern and Western European
217 ancestry from the CEPH collection; *n* = 165) and TSI (Toscans in Italy; *n* = 88) populations, which
218 are both European, and (2) the highly divergent CEU and YRI (Yoruba in Ibadan, Nigeria; *n* = 172)
219 populations, which are inter-continental. For each study, we simulated 1,000 replicates of a
220 heritable quantitative phenotype with a mean shift due to an individual's population

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

221 membership. To make each replicate of the phenotype 10% heritable, 100 SNPs from
222 chromosome 1 were randomly selected to be causal, each with an effect size chosen based on
223 allele frequency to account for 0.1% of the total phenotypic variability. The effects due to
224 population membership accounted for 18% of the phenotypic variability on average across
225 phenotype replicates in both studies. For each phenotype replicate, the SNPs on chromosomes
226 2-22, which had no direct causal link to the phenotype, were tested for association. Despite not
227 being causal, SNPs on these chromosomes could be indirectly correlated with the phenotype if
228 they had different allele frequencies in the two populations, resulting in inflated type-I error
229 rates if population stratification was not adequately accounted for.

230 Additional simulation studies were also carried out to assess the performance of the
231 association testing methods in the presence of ancestry admixture. We simulated genotypes for
232 three separate samples, each with two-way ancestry admixture, but each with a different choice
233 of F_{ST}[28] (0.01, 0.05, and 0.15) for the underlying populations. An F_{ST} of 0.01 is a typical value
234 between European populations, such as the CEU and TSI, while $F_{ST} = 0.15$ is representative of
235 divergent inter-continental populations, similar to what has previously been estimated between
236 the CEU and YRI populations[29,30]. To generate data sets under each choice of F_{ST}, allele
237 frequencies for the two underlying populations were generated for 200,000 independent SNPs
238 using the Balding-Nichols model[31], and genotype data at these SNPs were simulated for 3,000
239 individuals with admixed ancestry derived from the two populations. These SNPs were then
240 split into two disjoint sets of 100,000, which we refer to as Set 1 and Set 2. Each sample
241 included 2,160 unrelated individuals, 120 cousin pairs, and 30 four-generation, twenty-person
242 pedigrees (Figure S1). Individual ancestry proportions for unrelated individuals and pedigree
243 founders were randomly drawn from various beta distributions, and ancestry proportions for
244 pedigree descendants were calculated as the average of their parents'. For each of the three

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

245 admixed samples, we simulated 1,000 replicates of a quantitative phenotype with 10%
246 heritability and with a mean shift due to an individual's genome-wide ancestry that accounted
247 for 17% of the phenotypic variability on average. Causal SNPs for generating each phenotype
248 replicate were randomly selected from Set 2. The 100,000 SNPs in Set 1, which had no direct
249 causal link to the phenotype but could be indirectly correlated with it if they were associated
250 with genome-wide ancestry, were tested for association. For each association testing method,
251 sample genealogical structure was inferred using the SNPs in Set 1.
252 Finally, we performed simulation studies in the admixed setting with F_{ST} = 0.15 to assess
253 the power of each of the association testing methods to detect causal SNPs. Phenotypes were
254 simulated exactly as for the null SNP studies, but with an additional main effect due to a single
255 causal SNP of interest, *s'*, randomly selected from Set 2. The effect size for this causal SNP was
256 chosen based on allele frequency to account for a pre-specified percentage ($h^2_{s'}$ = 0.75%, 1.00%,
257 1.25%, or 1.50%) of the total phenotypic variability. For each choice of $h^2_{s'}$, a total of 10,000
258 phenotype-SNP pair replicates were generated and tested for association. Additional details on
259 how phenotypes and genotypes were generated for all simulations are provided in Appendix B.
260

261 **GWAS of WBC Count in WHI Hispanics.**

262 The Women's Health Initiative (WHI) is a long-term national health study focused on identifying
263 risk factors for common diseases in postmenopausal women. A total of 161,838 women aged
264 50–79 years were recruited from 40 clinical centers in the United States between 1993 and
265 1998. Detailed cohort characteristics and recruitment methods have been described
266 previously[18,19]. Approximately 17% of participants in this study are under-represented U.S.
267 minority women, and the WHI SNP Health Association Research (WHI-SHARe) minority cohort

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

268 includes 3,587 self-reported Hispanics who provided consent for DNA analysis. Affymetrix 6.0
269 genotyping and quality control filtering of these Hispanic-American samples was performed as
270 described previously[32]. Total circulating white blood cell (WBC) count was measured on a fresh
271 blood sample at local clinical laboratories using automated hematology cell counters and
272 standardized quality assurance procedures. Total WBC count was reported in millions of cells
273 per ml, and was log transformed prior to analysis to reduce skewness in the distributions of the
274 phenotypic data. A GWAS of the log-transformed WBC counts measured on women in the WHI-
275 SHARe Hispanic cohort was performed using LMM-OPS, EMMAX, GCTA, and GEMMA. For the
276 LMM-OPS analysis, the first 6 PCs generated with PC-AiR were included as fixed effects to adjust
277 for population stratification, and an ancestry-adjusted empirical GRM calculated conditionally
278 on these PCs with PC-Relate was used to account for recent familial relatedness. The other
279 LMM methods were run with their default settings, filters, and relationship matrices. A total of
280 616,556 autosomal SNPs were tested for association in the GWAS. Further details are provided
281 in Appendix A.
282
283 **Results**
284 **Impact of Variable Allele Frequency Differences on Association Test Statistics**
285 Using the two HapMap based simulation studies, we compared the test statistics obtained from
286 each association testing method for null SNPs. Penalized cubic regression splines were used to
287 find smoothed curves showing the relationship between the absolute value of the allele
288 frequency difference between the pair of populations at SNP *s*, denoted D_s, and the mean of the
289 test statistics from each method (Figure 1). Test statistics should follow a $\chi^2_{(1)}$ distribution at
290 null SNPs, so the mean of the test statistics for a well-calibrated method should be 1, regardless
291 of allele frequency differences between the two populations. However, in both HapMap

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

292 studies, the mean of the test statistics from EMMAX and GEMMA increased with increasing
293 values of D_s. Test statistics from these methods were substantially inflated (i.e. under-
294 corrected) at SNPs with the largest values of D_s and deflated (i.e. over-corrected) at SNPs with
295 the smallest values of D_s. In contrast, the mean of the test statistics from both LMM-OPS and
296 linear regression adjusted for the top PC from PC-AiR showed no relationship with the value of
297 D_s. This indicates that including the ancestry-representative PC as a fixed effect in the mean
298 model effectively accounted for the variable allele frequency differences across SNPs. However,
299 since linear regression adjusted for the top PC from PC-AiR did not account for the correlation of
300 phenotypes among relatives, its test statistics were equally inflated across all values of D_s (1.027
301 on average for the CEU/TSI sample, and 1.032 on average for the CEU/YRI sample). In both
302 studies, LMM-OPS was the only method that provided well-calibrated test statistics for all null
303 SNPs, with the mean of the test statistics near 1 for all values of D_s.
304 Comparing the results from EMMAX and GEMMA for the joint CEU/YRI sample (Figures
305 1A and 1B) to those for the joint CEU/TSI sample (Figures 1C and 1D), it is apparent that the
306 particular values of D_s for which the mean of the test statistics are either inflated or deflated
307 depends on the pair of populations being analyzed. To further understand this relationship, we
308 investigated the distribution of allele frequency differences across the genome for different
309 pairs of populations by estimating population specific allele frequencies at 1,423,833 autosomal
310 SNPs in the consensus data set for six HapMap populations: the previously mentioned CEU, TSI,
311 and YRI populations, as well as the LWK (Luhya in Webuye, Kenya; n = 90), CHB (Han Chinese in
312 Beijing, China ; n = 137), and JPT (Japanese in Tokyo, Japan; n = 86) populations. For each pair of
313 these populations, select quantiles (Table S1) of D_s and its cumulative distribution function
314 across all autosomal SNPs (Figure S2) were calculated. As expected, the distribution of D_s is
315 more concentrated at smaller values for pairs of populations from the same continent (i.e.

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

316 CEU/TSI, CHB/JPT, and LWK/YRI) as compared to pairs of populations from different continents.
317 Over 90% of SNPs have $D_s < 0.1$ for the three intra-continental pairs of populations, while at
318 least 49.8% of SNPs have $D_s \geq 0.1$ for each of the inter-continental pairs. When SNP *s* has a large
319 (small) D_s value relative to the other SNPs used to construct the GRM, both EMMAX and
320 GEMMA provide an inflated (deflated) test statistic for SNP *s* on average. Therefore, as a
321 consequence of the different distributions of D_s, inflation of EMMAX and GEMMA test statistics
322 is observed at smaller absolute values of D_s when jointly analyzing more closely related
323 populations, such as the CEU and TSI, compared to more divergent populations, such as the CEU
324 and YRI.
325
326 **Performance at Null SNPs in Admixed Populations**
327 We also compared the test statistics obtained from each association testing method for null
328 SNPs in the simulated admixed populations. Penalized cubic regression splines showing the
329 relationship between the local mean of the test statistics from each method and D_s showed the
330 same patterns as those from the HapMap simulations (Figures 2A-2C). Specifically, LMM-OPS
331 was the only method that provided well-calibrated test statistics for all D_s, test statistics from
332 linear regression adjusted for the top PC from PC-AiR were uniformly inflated for all values of D_s,
333 and EMMAX and GEMMA provided test statistics that were inflated at SNPs with the largest
334 values of D_s and deflated at SNPs with the smallest values of D_s. The distribution of D_s values
335 depended on the F_{ST} for the pair of populations contributing to the admixed sample (Figure 2D).
336 However, as demonstrated with the HapMap data, the qualitative patterns of test statistic
337 inflation and deflation for each method were the same across all choices of F_{ST}, regardless of the
338 range of D_s.

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

339 To further examine the performance of each method for null SNPs in the simulation
340 study with F_{ST} = 0.15, we defined three classes of SNPs based on the magnitude of their allele
341 frequency difference. SNPs were classified as weakly, moderately, or highly differentiated if they
342 were in the first (D_s < 0.07), second/third (0.07 < D_s < 0.28), or fourth (D_s > 0.28) quartile of the
343 distribution of the magnitude of allele frequency differences, respectively. Genomic inflation
344 factors, λ_{GC}, were calculated genome-wide, as well as in each of these three classes of SNPs, for
345 each of the 1,000 simulation replicates (Table 1). The genomic inflation factor is commonly used
346 in genetic association studies to evaluate confounding due to unaccounted for sample structure,
347 where $\lambda_{GC} \approx 1$ suggests appropriate correction, while $\lambda_{GC} > 1$ indicates an elevated type-I error
348 rate. LMM-OPS was the only method that obtained λ_{GC} values near 1 genome-wide as well as
349 within all three classes of SNPs. For linear regression adjusted for the top PC from PC-AiR, the
350 average genomic inflation factor was nearly the same ($\lambda_{GC} \approx 1.026$) genome-wide and within all
351 three classes of SNPs. Interestingly, both EMMAX and GEMMA obtained λ_{GC} values near 1
352 when calculated from the median of the test statistics for all SNPs genome-wide, but obtained
353 λ_{GC} values that were greater than 1 for highly differentiated SNPs and less than 1 for weakly
354 differentiated SNPs.
355 Additionally, modified QQ plots were generated for each LMM method using the *p*-
356 values for all 100,000,000 null SNPs pooled across the 1,000 phenotype replicates, as well as for
357 the subsets of these SNPs that were highly, moderately, or weakly differentiated (Figure 3). As
358 with the genomic-inflation factors, the QQ plots indicate that LMM-OPS is well calibrated
359 genome-wide as well as in all three classes of SNPs. In contrast, EMMAX and GEMMA appear
360 well calibrated when examining all SNPs genome-wide, but show deviation in the observed *p*-
361 values from those expected under the null when examining each of the three classes of SNPs
362 separately.

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

363 **Detection of Causal SNPs**

364 We also performed simulation studies in the setting with F_{ST} = 0.15 to assess the power of each
365 of the LMM methods. Linear regression with PCs was omitted from these comparisons because
366 it had consistent inflation of type-I error rates across all SNPs. Power to detect causal SNPs with
367 $h^2_{s'}$ = 0.75%, 1.00%, 1.25%, and 1.50% was computed at the genome-wide significance level α =
368 5 x 10^{-8} across all SNPs, as well as within the highly, moderately, and weakly differentiated
369 classes. When considering all causal SNPs, LMM-OPS had significantly higher power than
370 EMMAX and GEMMA by about 2-3% for each choice of $h^2_{s'}$ (Figure 4 and Table S2). This
371 difference in power corresponded with LMM-OPS detecting between 2% and 10% more causal
372 SNPs than the other LMM methods. Furthermore, LMM-OPS provided the highest power to
373 detect causal SNPs within each class of allele frequency differentiation. Perhaps surprisingly, this
374 included highly differentiated SNPs, for which EMMAX and GEMMA provide systematically
375 inflated test statistics at null SNPs and have inflated type-I error rates.

376

377 **GWAS of WBC Count in the WHI SHARe Hispanic Cohort**

378 LMM-OPS, EMMAX, GCTA, and GEMMA all obtained satisfactory genome-wide genomic
379 inflation λ_{GC} values (Table S3), but the QQ-plots of the $-\log_{10}$(p-values) from each method
380 appeared to show some early deviation from expectation (Figure 5A). The only SNPs
381 approaching genome-wide significance were on chromosome 1, so we recalculated λ_{GC} and
382 generated new QQ-plots with chromosome 1 excluded to investigate this deviation. The QQ-
383 plots excluding SNPs from chromosome 1 appeared well behaved for all four methods (Figure
384 5B), indicating that the early deviation in the genome-wide QQ-plots was due to a large number
385 of associated SNPs on chromosome 1. The corresponding λ_{GC} values were 1.005, 0.993, 0.993,
386 and 0.994 for LMM-OPS, EMMAX, GCTA, and GEMMA, respectively. We also conducted a GWAS

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

387 using standard linear regression including the top 6 PCs from PC-AiR as fixed effects. As
388 expected, the test statistics were inflated, giving λ_{GC} = 1.045 when excluding chromosome 1.
389 This inflation was most likely due to unaccounted for familial relatedness in the sample[20,24].
390 The genotype effect size estimates from all four LMM methods were similar on average,
391 however LMM-OPS consistently provided the smallest standard error estimates, and thus a
392 more efficient association test (Figure S3). The Manhattan plot of the LMM-OPS $-\log_{10}$(p-values)
393 shows a strong association signal for WBC count in a region on chromosome 1 (Figure 5C).
394 LMM-OPS attained the highest significance in this region, and all genome-wide significant p-
395 values with LMM-OPS were one or more orders of magnitude smaller than those from the
396 competing LMM methods (Figure 5D and Table 2). The most significant SNP on chromosome 1
397 was rs11265198 (LMM-OPS $p = 6.49 \times 10^{-13}$; EMMAX $p = 2.49 \times 10^{-10}$; GCTA $p = 2.77 \times 10^{-10}$;
398 GEMMA $p = 4.00 \times 10^{-11}$). In addition, there was one SNP in this region, rs6656586, which
399 attained genome-wide significance with LMM-OPS but did not with any of the other methods.
400 The most significant SNP, rs11265198, is near the Duffy Antigen Receptor for
401 Chemokines (DARC) gene. An African derived regulatory variant in DARC, rs2814778, was
402 previously found to associate with lower WBC count in African Americans[22,23]. Genotype
403 dosages for rs2814778 were imputed using MaCH-Admix[32] in the WHI-SHARe Hispanics and
404 tested for association using LMM-OPS. The p-value for rs2814778 was 1.89×10^{-18}, providing a
405 stronger signal than any of the directly genotyped SNPs tested for association with WBC count.
406 Additionally, we re-ran a GWAS with LMM-OPS conditional on rs2814778, and all of the
407 previously identified genome-wide significant SNPs on chromosome 1 became non-significant
408 (Figure S4), indicating that these SNPs were tagging this regulatory variant. Thus, we were able
409 replicate and generalize[21] to Hispanics/Latinos the previously identified association in African
410 Americans for WBC count and this regulatory variant in the DARC gene.

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

411

412

413 **Importance of Orthogonality**

414 The key feature of LMM-OPS that differentiates it from existing LMM methods for GWAS is the
415 partitioning of sample structure into separate orthogonal components due to recent and distant
416 genetic relatedness. For each of the LMM-OPS analyses presented here, the PC-Relate kinship
417 coefficient matrix, which is calculated conditionally on PCs, was used as the ancestry-adjusted
418 GRM to account for recent genetic relatedness. However, this matrix is actually only
419 approximately orthogonal to the PCs used to account for distant genetic relatedness (see
420 Methods). In order to construct a strictly orthogonal matrix, one could compute an ancestry-
421 adjusted GRM from genotype values that are centered conditional on PCs, without any scaling.
422 Nevertheless, we consistently find that the entries in either of these ancestry-adjusted GRMs
423 are very highly correlated across all pairs of individuals, and the resulting LMM-OPS association
424 test statistics and *p*-values are nearly identical when using either matrix (Figures S5 – S7 and
425 Table 2). Due to this observation, we typically recommend using the PC-Relate kinship
426 coefficient matrix in practice, as it has the advantage of biological interpretability of the matrix
427 elements as kinship coefficient estimates.

428 To further explore the impact of using an empirical GRM that is orthogonal to the PCs
429 included in the analysis as fixed effect covariates, we re-ran the EMMAX, GCTA, and GEMMA
430 analyses of WBC count, but included the same 6 PCs that were used in the LMM-OPS analysis as
431 fixed effect covariates, which corresponds to an approach that was previously proposed for
432 associating testing in structured samples with unusually differentiated SNPs[15,26]. In these
433 analyses, the PCs and the GRM were both adjusting for the population structure. The results
434 from these models were very similar to those from the corresponding models without PCs

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

435 (Figure S8 and Table 2). These results demonstrate that the efficiency gain achieved with LMM-
436 OPS cannot be replicated by simply including ancestry representative PCs in an LMM with a
437 standard empirical GRM. Partitioning sample structure due to population stratification and
438 recent genetic relatedness into separate, orthogonal components, as is done with LMM-OPS, is
439 essential for improved efficiency in association testing.
440
441 **Computation Time**
442 We compared the computation time required by LMM-OPS, EMMAX, GEMMA (v0.94), and GCTA
443 (v1.24.7) to perform the association analysis of WBC count at all 616,556 autosomal SNPs for
444 the 3,551 women in the WHI-SHARe sample. The computation times required to construct the
445 relationship matrices were not included in this assessment since all of the LMM methods can
446 use a pre-computed GRM, and each GRM has similar computational complexity[10]. Each method
447 was run using a single core 2.4 GHz Intel Xeon E5-2630L processor with 128 GB of RAM. LMM-
448 OPS is implemented in R, and was run using R v3.2.0 configured to use the BLAS and LAPACK
449 libraries within the Intel Math Kernel Library (MKL). LMM-OPS was substantially faster than the
450 other methods, which took at least twice as long, and up to over seven times as long to perform
451 the analysis. The computation time for LMM-OPS to complete the analysis was 20.4 minutes
452 (0.34 hours). In comparison, the computation time for GCTA was 44.6 minutes (0.74 hours), for
453 GEMMA was 1.84 hours, and for EMMAX was 2.43 hours.
454
455 **Discussion**
456 Genetic association studies involving ancestrally diverse populations from around the world
457 have recently become more common as there is increased interest in both identifying novel,
458 population specific variants that underlie phenotypic diversity and generalizing associations

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

459 across populations. Confounding due to ancestry is a serious concern for genetic association
460 studies since different ethnic groups often share distinct dietary habits and other lifestyle
461 characteristics that lead to many traits of interest being correlated with ancestry. Linear mixed
462 models (LMMs) have become the go-to approach for genome-wide association testing of
463 quantitative traits, as they are both computationally fast and statistically powerful[10].
464 Furthermore, it has been reported that LMMs are effective at controlling type I error rates in
465 samples with relatedness and population structure, as they tend to provide acceptable genomic
466 control inflation factors. However, we demonstrated through simulation studies that existing
467 implementations of LMMs for GWAS can provide systematically biased test statistics in samples
468 with population stratification when phenotypes are correlated with ancestry. Interestingly, we
469 found that the miscalibration of test statistics from widely used LMM approaches is a problem
470 that can occur when samples descend from either highly divergent inter-continental populations
471 or closely related intra-continental populations. Additionally, the problem manifests in the
472 presence of both discrete population substructure and ancestry admixture.

473 Incorrect calibration of association test statistics from existing LMM methods arises
474 because they use an empirical GRM calculated as a genome-wide average genotype covariance
475 to account for the entire sample genealogy, including both distant population structure and
476 recent familial relatedness. However, as we demonstrated using genotype data from release 3
477 of phase III of HapMap, allele frequency differences among human populations vary greatly by
478 SNP across the genome. As a consequence, the strength of the genetic covariance due to
479 ancestry also varies by SNP across the genome, and existing LMM methods can provide
480 inadequate correction for population stratification. The ancestry correction provided by these
481 methods is of appropriate size only for SNPs with typical allele frequency differences between
482 populations, while SNPs with relatively larger or smaller allele frequency differences receive an

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

483 under- or over-correction, respectively. This result is contrary to what has previously been
484 suggested[10], that SNPs with larger allele frequency differences between populations receive a
485 larger correction from these LMM methods. Notably, the deflation of test statistics for SNPs
486 with similar allele frequencies across populations, and the inflation of test statistics for SNPs
487 that are highly differentiated across populations, balances out across the genome, typically
488 leading to a genomic control inflation factor near 1 and QQ-plots that look acceptable when
489 considering all SNPs genome-wide, which likely contributed to why this phenomenon had not
490 been previously reported.

491 To address this issue, we developed LMM-OPS, which orthogonally partitions the
492 genealogical structure among sampled individuals into two separate components. Ancestry-
493 representative vectors are included as fixed effects in the mean model of LMM-OPS to account
494 for population stratification, and an ancestry-adjusted empirical GRM that is orthogonal to the
495 ancestry-representative vectors is used to model recent familial relatedness as a random effect.
496 In simulation studies with real and simulated genotype data, we demonstrated that the LMM-
497 OPS testing procedure provides well-calibrated association test statistics at *all* SNPs genome-
498 wide, regardless of the distribution of allele frequency differences among the underlying
499 populations. In addition to providing better protection against false positives, we also
500 demonstrated that, compared to existing LMM methods, LMM-OPS provides a more efficient
501 test with improved power to detect true SNP-phenotype associations. This increase in power
502 holds even at highly differentiated SNPs, for which existing LMM methods provide
503 systematically inflated test statistics resulting in an inflated type-I error rate.

504 We also compared the performance of LMM-OPS to existing implementations of LMMs
505 through a GWAS analysis of white blood cell (WBC) count in the Hispanic cohort of the WHI
506 SHARe study. This cohort contains multi-way continental ancestry admixture as well as cryptic

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

507 familial relatedness. All four methods gave similar genotype effect size estimates at SNPs for
508 WBC count, but LMM-OPS was the most efficient, providing consistently smaller standard errors
509 and genome-wide significant *p*-values that were more significant, by one or more orders of
510 magnitude, than EMMAX, GEMMA, and GCTA. Using LMM-OPS, we were able to replicate
511 generalize to this Hispanic American population ($p = 1.89 \times 10^{-18}$) the association at regulatory
512 variant rs2814778 in the Duffy Antigen Receptor for Chemokines (DARC) gene previously
513 identified in African Americans. Because of natural selection, rs2814778 is highly differentiated
514 between African and European ancestral populations, likely due to a protective effect against *P.*
515 *vivax* malaria[34,35]. Furthermore, through a conditional analysis including rs2814778 as a
516 covariate, we were able to demonstrate that other genome-wide significant associations in this
517 region on chromosome 1 could be explained by LD with this particular variant.
518 In the implementation of LMM-OPS presented here, we utilized ancestry-representative
519 principal components (PCs) from PC-AiR to adjust for population structure and construct an
520 orthogonal ancestry-adjusted empirical GRM to account for relatedness. Alternatively, vectors
521 of estimated individual admixture proportions from model-based methods such as
522 ADMIXTURE[36] or FRAPPE[37] could be used in place of PCs. We also performed the LMM-OPS
523 analysis of WBC count in the Hispanic American cohort of the WHI-SHARe using model-based
524 estimates of individual ancestry from a supervised ADMIXTURE analysis that included reference
525 population samples from the International Haplotype Map Project (HapMap) and the Human
526 Genome Diversity Project (HGDP)[38] for European, Native American, African, and East Asian
527 ancestry. The results were nearly identical to the analysis that used PCs (Figure S9). In general,
528 using model-based estimates of ancestry with LMM-OPS is expected to work well, provided that
529 prior assumptions regarding the underlying populations contributing ancestry to the sample are
530 accurate and suitable reference population panels representative of these populations are

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

531 available for reliable estimation of ancestry. For many genetic studies, however, the underlying
532 ancestral populations may not be completely known or well defined, and misspecification of the
533 ancestral populations will result in less efficient association tests due to biased estimates of
534 individual ancestry from model-based methods. For this reason, we recommend using PCs with
535 LMM-OPS, as they do not rely on strong modeling assumptions.

536 Numerous LMM approaches for GWAS have been proposed in the past few years, each
537 with small variations on the testing procedure. One variation among these methods lies in SNP
538 selection for constructing the empirical GRM. In each of the analyses performed here, the GRM
539 was constructed from all SNPs genome-wide with sample minor allele frequency (MAF) greater
540 than 1%. Others have suggested that including the SNP being tested, or SNPs in LD with the one
541 being tested, leads to proximal contamination and a loss of power[6,10]. This has led to the
542 development of automated methods for selecting subsets of SNPs to be used in the GRM[6], as
543 well as the leave-one-chromosome-out (LOCO) approach[10], where the GRM is constructed from
544 all autosomal SNPs not on the same chromosome as the SNP being tested. While not presented
545 here, LMM-OPS can easily incorporate using a subset of SNPs or a LOCO approach for
546 constructing the ancestry-adjusted GRM. However, it remains unclear as to what the optimal
547 set of SNPs is for obtaining the GRM. Using a subset of SNPs, as is done with FaST-LMM-Select[6],
548 may improve power, but it may also provide inadequate control of type-I error rates[39]. When
549 there is family relatedness among samples, this remains true even if PCs are included in the
550 model[39]. Similarly, using a LOCO approach, such as that implemented in GCTA-LOCO[10], may also
551 improve power, but may inadequately account for the effects of SNPs on the same chromosome
552 as the SNP being tested. A hybrid method that uses two GRMs, one constructed from all SNPs,
553 and the other from a selected subset of SNPs, has also been proposed[39]. However, this
554 approach remains susceptible to the systematic inflation/deflation issues illustrated in this work

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

555 for phenotypes correlated with ancestry. We have demonstrated that LMMs for GWAS should
556 certainly include ancestry-representative vectors as fixed effects to account for population
557 stratification, but exactly which SNPs should be used to construct the empirical GRM in order to
558 optimize power, while still adequately protecting against false positives due to family structure,
559 remains an open area of research.
560

561 **Appendix A**

562 *Sample Structure Inference*

563 Population structure inference for LMM-OPS and for linear regression adjusted with PCs was
564 performed using PC-AiR[25]. PC-AiR used SNPs LD-pruned with an r^2 = 0.1 threshold in a sliding
565 10Mb window for the HapMap and WHI-SHARe analyses, and it used all SNPs in Set 1 for the
566 simulations with admixture. Inference on recent genetic relatedness due to family structure for
567 LMM-OPS was performed using PC-Relate[23]; these kinship coefficient estimates were used to
568 construct the ancestry-adjusted empirical GRM. EMMAX, GEMMA, and GCTA each used the
569 default empirical GRM created by the respective software to infer all sample structure. The
570 GRM for each method was constructed from all of the autosomal SNPs with minor allele
571 frequency (MAF) > 1% in the HapMap and WHI-SHARe analyses, and from all SNPs in Set 1 in the
572 simulations with admixture.
573

574 **Appendix B**

575 *Simulated Genotype Data with Admixture*

576 For each of the three simulated data sets with admixture, allele frequencies for the two
577 populations at all 200,000 SNPs were generated using the Balding-Nichols model[30]. More
578 precisely, for each SNP *s*, the allele frequency p_s in the ancestral population was drawn from a

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

579 uniform distribution on [0.1, 0.9]. The allele frequency in each population was then drawn from
580 a beta distribution with parameters $p_s(1–F_{ST})/F_{ST}$ and $(1–p_s)(1–F_{ST})/F_{ST}$, where the quantity F_{ST} is
581 Wright's measure of genetic distance between populations[27]. The vector $\mathbf{p}_s$ contains the allele
582 frequencies at SNP *s* for each population. An individual's vector of ancestry proportions from
583 each of the two populations can be represented by $\mathbf{a}_i^T = (a, 1-a)$. The parameter *a* was
584 drawn from a beta distribution with mean 0.3 and s.d. 0.1 for one third of unrelated individuals
585 and pedigree founders, a beta distribution with mean 0.7 and s.d. 0.1 for another third, and a
586 uniform distribution on [0,1] for the remaining third. Within any given pedigree, every founder
587 had a drawn from the same distribution. As a result, individuals in a pedigree for which
588 founder ancestry proportions were drawn from either of the beta distributions had similar
589 ancestry to each other, which could be viewed as a type of ancestry-related assortative mating.
590 Ancestry proportions for pedigree descendants were calculated as the average of their parents'
591 ancestry proportions. Genotypes for unrelated individuals and pedigree founders were
592 randomly drawn from a $\mathrm{Bin}(2, \mathbf{a}_i^T \mathbf{p}_s)$ distribution, and alleles were passed down the pedigree
593 to generate genotypes for pedigree descendants (including the cousin pairs).
594
595 *Simulated Phenotypes*
596 For the simulations used to assess behavior at null SNPs, each replicate of the heritable
597 phenotype whose mean depended on genome-wide ancestry (or population membership) was
598 generated according to the model
599

$$\mathbf{Y} = \Sigma_{s \in \mathcal{S}_c} \mathbf{x}_s \beta_s + \gamma \mathbf{a} + \epsilon, \quad (6)$$

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

600 where $\mathcal{S}_c$ is a set of 100 randomly selected causal SNPs, $\mathbf{x}_s$ is the vector of genotype values for
601 all individuals at SNP $s \in \mathcal{S}_c$ with effect size β_s chosen using sample allele frequencies so that
602 SNP *s* explains 0.1% of the phenotypic variability, $\mathbf{a}$ is the vector of ancestry proportions for all
603 individuals with effect size γ , and $\epsilon_i \sim \mathrm{N}(0,1)$ is independent random noise. In the simulations
604 using the HapMap genotype data, SNPs for $\mathcal{S}_c$ were selected randomly from chromosome 1, $\mathbf{a}$
605 was 1 for an individual in the CEU population and 0 otherwise, and $\gamma = 1$. In the simulations
606 using the simulated genotype data with admixture, SNPs for $\mathcal{S}_c$ were randomly chosen from Set
607 2, $\mathbf{a}$ was an individual's ancestry proportion from population 1, and $\gamma = 2$.
608 For the simulations used to evaluate detection of causal SNPs, each of the causal SNP-
609 phenotype pairs was generated according to the model
610
$$\mathbf{Y} = \mathbf{x}_{s'}\beta_{s'} + \Sigma_{s \in \mathcal{S}_c} \mathbf{x}_s \beta_s + \gamma \mathbf{a} + \epsilon \,. \quad (7)$$
611 This is the same model as in Equation (6) but with one additional causal SNP of interest, *s'*. In
612 the simulations using the simulated genotype data with admixture, *s'* was selected at random
613 from Set 2. The effect size $\beta_{s'}$ for SNP *s'* was chosen to explain a pre-specified proportion of the
614 phenotypic variability, denoted $h^2_{s'}$, and 10,000 replicate SNP-phenotype pairs were simulated
615 for each choice of $h^2_{s'} \in$ {0.75%, 1.00%, 1.25%, 1.50%}.
616
617 **Supplemental Data**
618 Supplemental data include 9 Figures and 3 Tables.
619
620

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

621 **Acknowledgements**

622 This work was supported in part National Institutes of Health (NIH) grants P01 GM 099568 (to
623 T.A.T.), K01 CA148958 (to T.A.T), R01 HG001645 (to M.S.M.), and R01 HL116446 (to A.P.R. and
624 T.A.T.). The WHI program is funded by the National Heart, Lung, and Blood Institute, National
625 Institutes of Health, U.S. Department of Health and Human Services through contracts
626 HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C,
627 HHSN268201100004C, and HHSN271201100004C. Funding for WHI SNP Health Association
628 Resource (WHI-SHARe) genotyping was provided by NHLBI contract N02-HL-64278.

629

630 **Web Resources**

631 LMM-OPS is implemented in the R language and is freely available from Bioconductor as part of
632 the GENESIS package
633 (http://www.bioconductor.org/packages/release/bioc/html/GENESIS.html).

634
635
636
637
638
639
640
641
642
643
644

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B. and Kresovich, S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38, 203-208

2. Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics, 178, 1709-1723.

3. Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42, 348–354

4. Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., and Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42, 355

5. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods 8, 833–835

6. Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., and Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. Nat. Methods 9, 525–526

7. Svishcheva, G.R., Axenovich, T.I., Belonogova, N.M., van Duijn, C.M., and Aulchenko, Y.S. (2012). Rapid variance components-based method for whole-genome association analysis. Nat. Genet. 44, 1166-1170

8. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44, 821–824

9. Listgarten, J. Lippert, C., and Heckerman, D. (2013). FaST-LMM-Select for addressing confounding from spatial structure and rare variants. Nat. Genet. 45, 470-471

10. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. 46, 100-106

11. Hayes, B.J., Visscher, P.M., and Goddard, M.E (2009). Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91, 47-60

12. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Nat. Genet. 42, 565-569

13. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76-82

14. Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics 55, 997–1004

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

15. Price, A.L., Zaitlen, N.A., Reich, D. and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11, 459–463

16. Wu, C., DeWan, A., Hoh, J. & Wang, Z. (2011). A comparison of association methods correcting for population stratification in case–control studies. Ann. Hum. Genet. 75, 418-427

17. The International HapMap Consortium. (2005). A haplotype map of the human genome. Nature 437, 1299–1320

18. Hays, J., Hunt, J. R., Hubbell, F. A., Anderson, G. L., Limacher, M., Allen, C., and Rossouw, J. E. (2003). The women's health initiative recruitment methods and results. Ann. Epidemiol. 13, S18-S77

19. Anderson, G., Cummings, S., Freedman, L.S., Furberg, C., Henderson, M., Johnson, S.R., Kuller, L., Manson, J., Oberman, A., Prentice, R.L. and Rossouw, J.E. (1998). Design of the Women's Health Initiative clinical trial and observational study. Control. Clin. Trials 19, 61–109

20. Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., & Risch, N. (2012). Estimating kinship in admixed populations. Am. J. Hum. Genet. 91, 122-138

21. Jain, D., Hodonsky, C.J., Schick, U.M., Morrison, J.V., Minnerath, S., Brown, L., Schurmann, C., Liu, Y., Auer, P.L., Laurie, C.A., Taylor, K.D., et al. (2017). Genome-wide association of white blood cell counts in Hispanic/Latino Americans: the Hispanic Community Health Study/Study of Latinos. Hum. Mol Genet. 26, 1193-1204

22. Reich, D., Nalls, M.A., Kao, W.L., Akylbekova, E.L., Tandon, A., Patterson, N., Mullikin, J., Hsueh, W.C., Cheng, C.Y., Coresh, J., and Boerwinkle, E. (2009). Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. PLoS Genet. 5, e1000360

23. Reiner, A.P., Lettre, G., Nalls, M.A., Ganesh, S.K., Mathias, R., Austin, M.A., Dean, E., Arepalli, S., Britton, A., Chen, Z., and Couper, D. (2011). Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). PLoS Genet. 7, e1002108

24. Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton T.A. (2016). Model-free estimation of recent genetic relatedness. Am. J. Hum. Genet. 98, 127-148

25. Lynch, M., and Walsh, B. Genetics and Analysis of Quantitative Traits (Sinauer, Sunderland, Massachusetts, 1998).

26. Conomos, M.P., Miller, M., and Thornton, T. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet. Epidemiol. 39, 276-293

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

737 27. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006).
738 Principal components analysis corrects for stratification in genomewide association studies. Nat.
739 Genet. 38, 904–909
740
741 28. Wright, S. (1949). The genetical structure of populations. Ann. Eugen. 15, 323–354
742
743 29. Nelis, M., Esko, T., Mägi, R., Zimprich, F., Zimprich, A., Toncheva, D., Karachanak, S.,
744 Piskáčková, T., Balaščák, I., Peltonen, L., and Jakkula, E. (2009). Genetic structure of Europeans: a
745 view from the North–East. PLoS One 4, e5472
746
747 30. Bhatia, G., Patterson, N., Sankararaman, S. and Price, A.L. (2013). Estimating and
748 interpreting FST: the impact of rare variants. Genome Res. 23, 1514-1521
749
750 31. Balding, D.J., and Nichols, R.A. (1995). A method for quantifying differentiation between
751 populations at multi-allelic loci and its implications for investigating identify and paternity.
752 Genetica 96, 3–12
753
754 32. Reiner, A.P., Beleza, S., Franceschini, N., Auer, P.L., Robinson, J.G., Kooperberg, C., Peters,
755 U., and Tang, H. (2012). Genome-wide association and population genetic analysis of C-reactive
756 protein in African American and Hispanic American women. Am. J. Hum. Genet. 91, 502–512
757
758 33. Liu E.Y., Li M., Wang W., and Li Y. (2012). MaCH-Admix: genotype imputation for admixed
759 populations. Genet. Epidemiol. 37, 25-37
760
761 34. Miller, L.H., Mason, S.J., Clyde, D.F., and McGinniss, M.H. (1976). The resistance factor to
762 Plasmodium vivax in blacks: the Duffy-blood-group genotype, FyFy. New Engl. J. Med., 295, 302-
763 304
764
765 35. Horuk, R., Chitnis, C.E., Darbonne, W.C., Colby, T.J., Rybicki, A., Hadley, T.J,. and Miller, L.H.
766 (1993). A receptor for the malarial parasite Plasmodium vivax: the erythrocyte chemokine
767 receptor. Science 261, 1182-1184
768
769 36. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of
770 ancestry in unrelated individuals. Genome Res. 19, 1655-1664
771
772 37. Tang, H., Peng, J., Wang, P., and Risch, N.J. (2005). Estimation of individual admixture:
773 analytical and study design considerations. Genet. Epidemiol. 28, 289-301
774
775 38. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M.,
776 Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. (2008). Worldwide human
777 relationships inferred from genome-wide patterns of variation. Science 319, 1100-1104
778
779 39. Widmer, C., Lippert, C., Weissbrod, O., Fusi, N., Kadie, C., Davidson, R., Listgarten, J., and
780 Heckerman, D. (2014). Further improvements to linear mixed models for genome-wide
781 association studies. Sci. Rep. 4, 6874
782
783
784

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

785 **Figure Titles and Legends**

786 **Figure 1. Performance of association methods at null SNPs for a phenotype associated with**
787 **ancestry in the presence of population structure.**
788 Penalized cubic regression splines were used to fit smoothed curves showing the relationship
789 between the absolute value of the allele frequency difference between the two populations at
790 SNP s, D_s, and the local mean of the test statistics from each method for the simulations with (A-
791 B) the joint CEU/YRI sample, and (C-D) the joint CEU/TSI sample. The curves shown are the
792 average relationship across all 1,000 simulated replicates (individual points are omitted for
793 visual clarity). The shaded regions show estimated 95% confidence intervals. (A and C) The
794 curves are fit to all SNPs, and the range of D_s is held to [0,1]. (B) The curves are fit to the 98.6%
795 of SNPs with $D_s \leq 0.6$. (D) The curves are fit to the 99.9% of SNPs with $D_s \leq 0.2$.
796
797
798 **Figure 2. Performance of association methods at null SNPs for a phenotype associated with**
799 **ancestry in admixed populations.**
800 Penalized cubic regression splines were used to fit smoothed curves showing the relationship
801 between the absolute value of the allele frequency difference between the two populations at
802 SNP s, D_s, and the local mean of the test statistics from each method for the simulations with
803 admixture from a pair of populations with (A) $F_{ST} = 0.15$, (B) $F_{ST} = 0.05$, and (C) $F_{ST} = 0.01$. The
804 curves shown are the average relationship across all 1,000 simulated replicates (individual
805 points are omitted for visual clarity). The shaded regions show estimated 95% confidence
806 intervals. The range of D_s is kept the same in each panel to emphasize that EMMAX and GEMMA
807 provide inflated (deflated) test statistics at the largest (smallest) values of D_s, regardless of its
808 range. (D) Cumulative distribution functions showing the distribution of D_s values for each
809 choice of F_{ST}.
810
811
812 **Figure 3. Modified *p*-value QQ-plots for different classes of null SNPs.**
813 QQ-plots of *p*-values for (A) all SNPs, (B) highly differentiated SNPs, (C) moderately
814 differentiated SNPs, and (D) weakly differentiated SNPs are presented for LMM-OPS, EMMAX,
815 and GEMMA from the simulation with admixture from a pair of populations with $F_{ST} = 0.15$. To
816 more easily see deviation from the null, the y-axis is the difference between the observed and
817 expected $-\log_{10}$(*p*-values) rather than the observed. Points above the gray cone indicate
818 inflation, and points below indicate deflation.
819
820
821 **Figure 4. Comparison of power of LMM methods for a phenotype associated with ancestry in**
822 **the simulation study with admixture and $F_{ST} = 0.15$.**
823 The power of LMM-OPS, EMMAX, and GEMMA to detect causal SNPs with $h^2 = 0.75\%$, 1.00%,
824 1.25%, and 1.50% is shown across all SNPs as well as within the three classes of SNPs defined by
825 allele frequency differentiation. The points represent the power estimates, and the vertical bars
826 represent the 95% confidence intervals
827
828
829
830

bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

831 **Figure 5. Association testing results for white blood cell (WBC) count in the Hispanic cohort of**
832 **the WHI-SHARe study.**
833 QQ-plots for each of the LMM methods with (A) all autosomal SNPs, and (B) autosomal SNPs
834 excluding chromosome 1. (C) Manhattan plot of the $-\log_{10}(p\text{-values})$ from LMM-OPS. (D) Direct
835 comparison of $-\log_{10}(p\text{-values})$ for all autosomal SNPs from LMM-OPS to each of the other LMM
836 methods. The EMMAX and GCTA results are presented together, as they were nearly identical
837 and could not be distinguished in the figures.
838
839
840
841 **Tables**
842
843 **Table 1. Genomic inflation factors, λ_{GC}, at null SNPs for the simulation study with admixture**
844 **and F_{ST} = 0.15**

Method	Genome-Wide	Highly[a] Differentiated	Moderately[b] Differentiated	Weakly[c] Differentiated
LMM-OPS	1.000 (0.0002)	0.999 (0.0004)	1.001 (0.0003)	1.001 (0.0005)
EMMAX	1.001 (0.0002)	1.059 (0.0007)	0.989 (0.0003)	0.973 (0.0005)
GEMMA	1.004 (0.0002)	1.067 (0.0007)	0.990 (0.0003)	0.973 (0.0005)
Linear Reg. +PCs	1.026 (0.0006)	1.026 (0.0007)	1.027 (0.0007)	1.027 (0.0007)

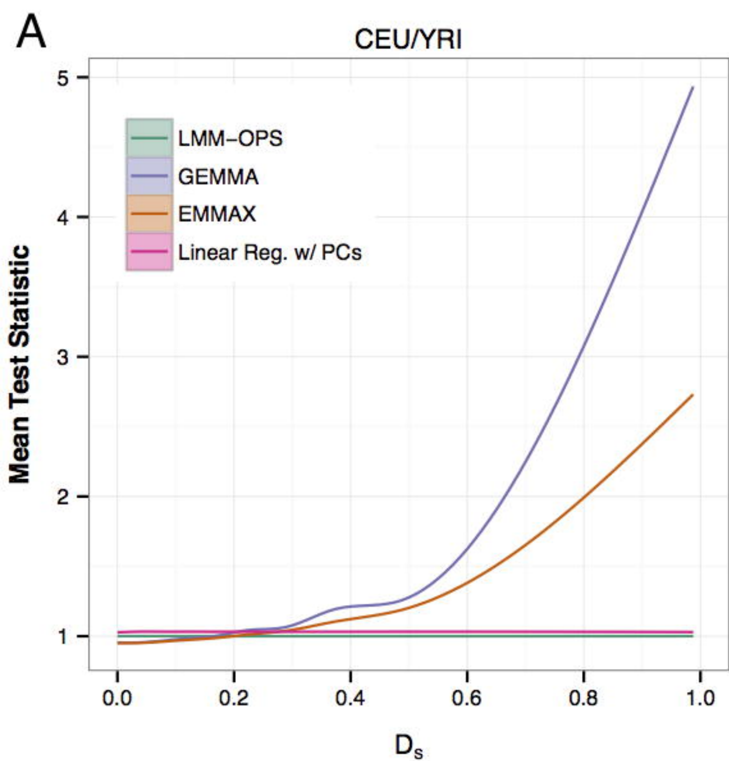
845 The values presented are the mean (s.e.) across all 1,000 phenotype replicates.
846 [a] Highly differentiated SNPs: $D_s > 0.28$ between the two populations
847 [b] Moderately differentiated SNPs: $0.07 < D_s < 0.28$ between the two populations
848 [c] Weakly differentiated SNPs: $D_s < 0.07$ between the two populations

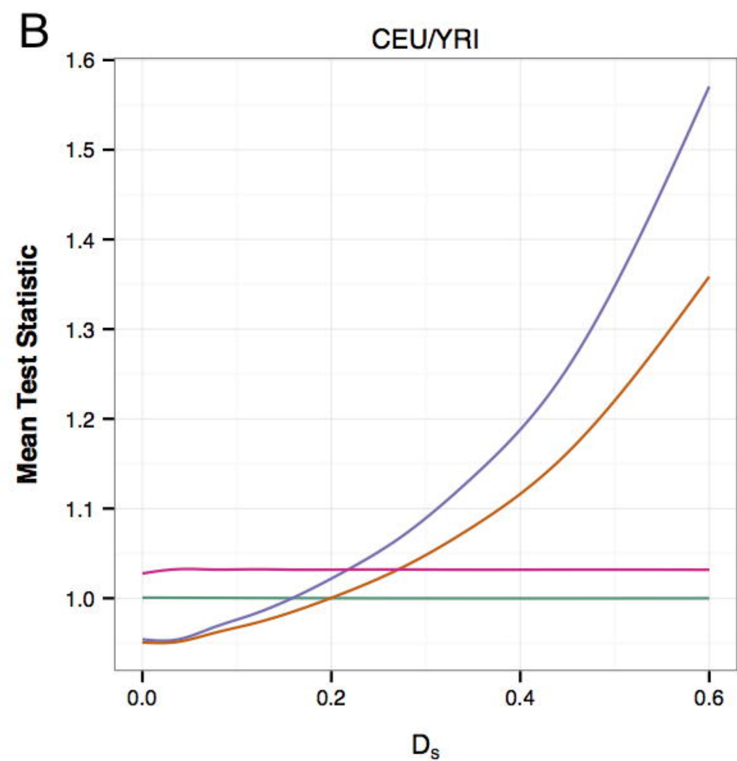
Table 2. Genome-wide significant SNPs on chromosome 1 for WBC count in WHI-SHARe Hispanics

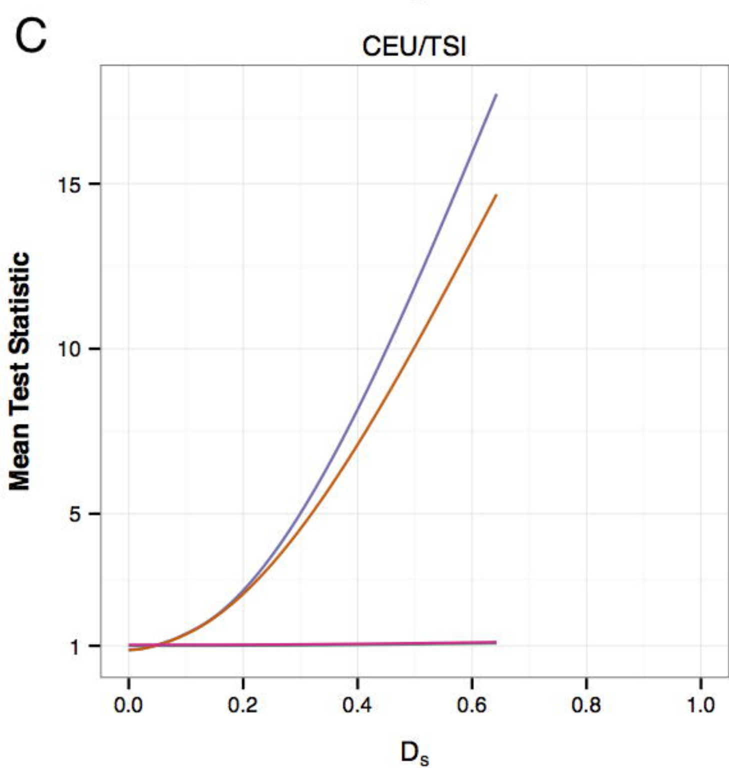
SNP ID	Position	*p*-value							
		LMM-OPS[a]	LMM-OPS[b]	EMMAX	EMMAX +PCs[c]	GCTA	GCTA +PCs[c]	GEMMA	GEMMA +PCs[c]
rs11265198	159450517	6.49×10^{-13}	5.75×10^{-13}	2.49×10^{-10}	2.45×10^{-10}	2.77×10^{-10}	3.39×10^{-10}	4.00×10^{-11}	4.14×10^{-11}
rs2808666	159591526	1.16×10^{-10}	1.07×10^{-10}	2.53×10^{-9}	3.65×10^{-9}	2.75×10^{-9}	4.44×10^{-9}	1.05×10^{-9}	1.70×10^{-9}
rs7534472	159500861	2.62×10^{-10}	2.45×10^{-10}	1.34×10^{-8}	1.36×10^{-8}	1.44×10^{-8}	1.63×10^{-8}	3.94×10^{-9}	4.21×10^{-9}
rs857682	158670244	7.92×10^{-10}	6.84×10^{-10}	2.14×10^{-8}	2.23×10^{-8}	2.28×10^{-8}	2.73×10^{-8}	9.53×10^{-9}	1.07×10^{-8}
rs856065	159013653	1.47×10^{-9}	1.44×10^{-9}	5.41×10^{-8}	6.55×10^{-8}	5.73×10^{-8}	7.56×10^{-8}	1.82×10^{-8}	2.45×10^{-8}
s6656586	159013653	2.40×10^{-8}	2.14×10^{-8}	3.53×10^{-7}	3.92×10^{-7}	3.69×10^{-7}	4.61×10^{-7}	1.91×10^{-7}	2.30×10^{-7}

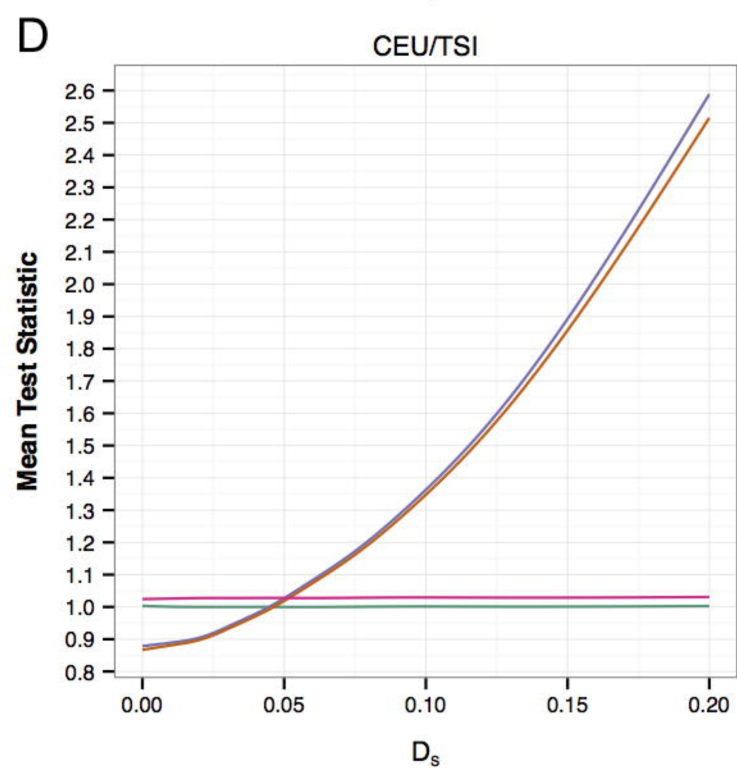
The *p*-values are presented for all six SNPs reaching genome-wide significance for WBC count with any of the LMM methods.
[a]LMM-OPS when using the PC-Relate ancestry-adjusted empirical GRM; [b]LMM-OPS when using the centered only ancestry-adjusted empirical GRM. [c]The results labeled "+PCs" are from each of the respective LMM methods when the top 6 PCs from PC-AiR are included as fixed effect covariates

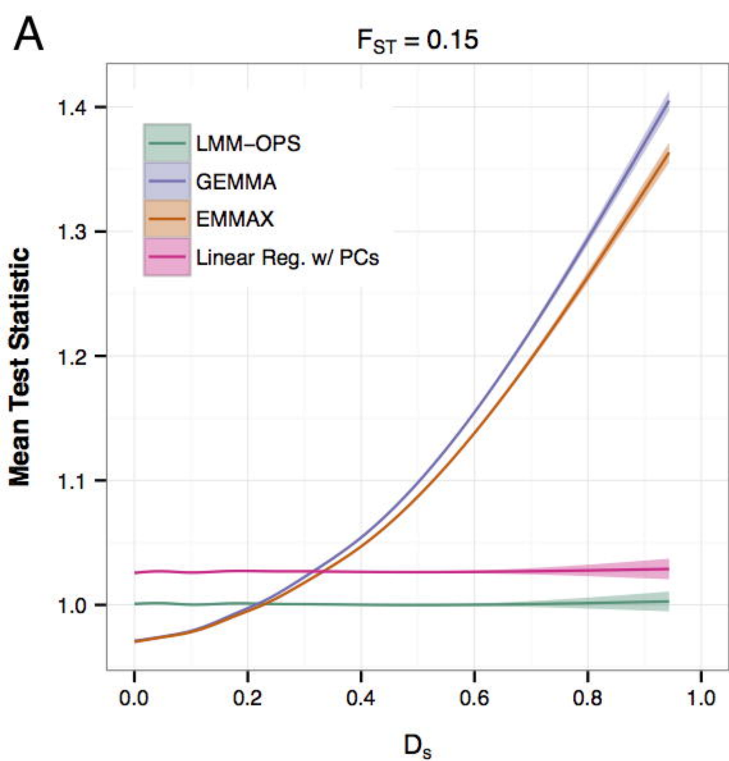
bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

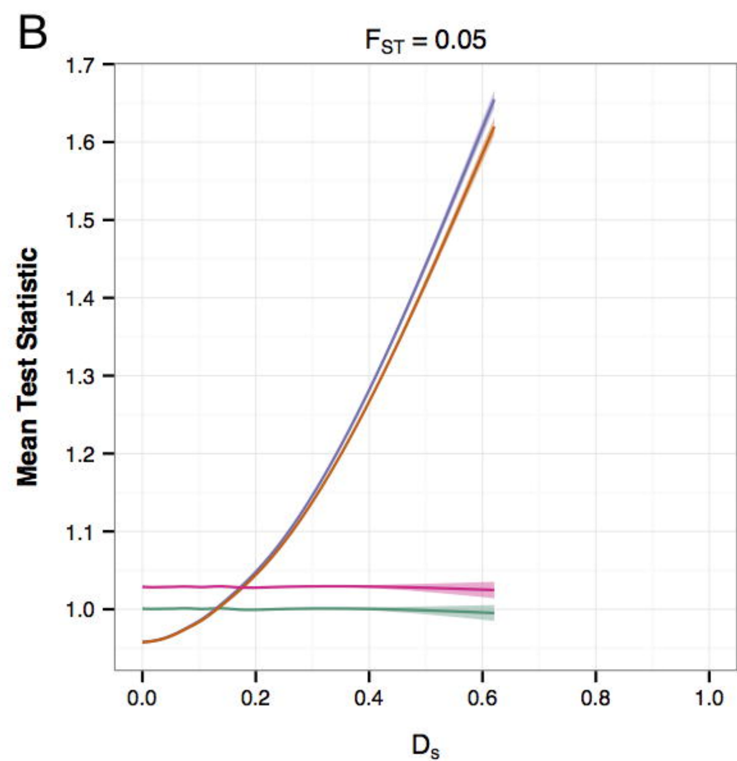

A
CEU/YRI
LMM-OPS
GEMMA
EMMAX
Linear Reg. w/ PCs
Mean Test Statistic
D_s

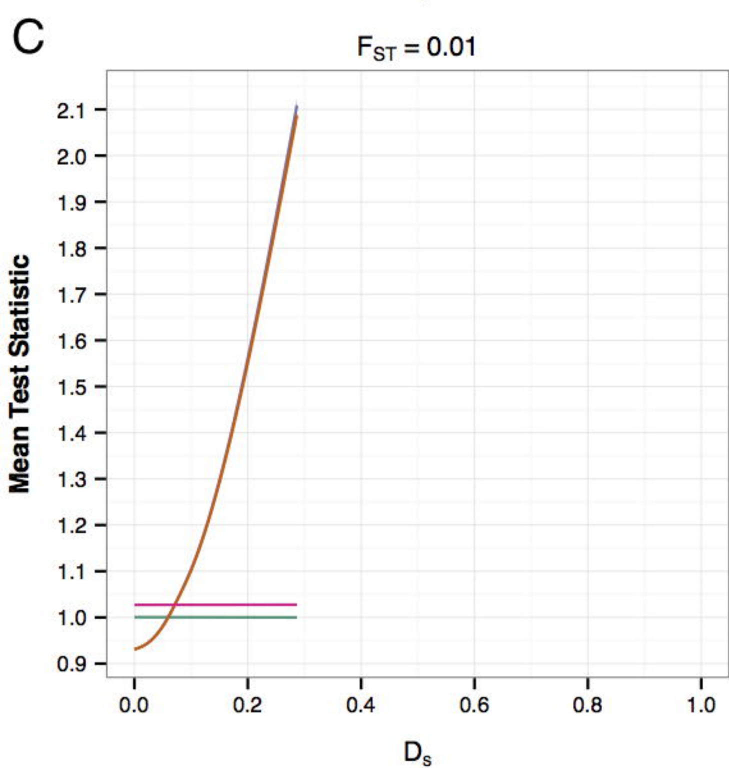

B
CEU/YRI
Mean Test Statistic
D_s

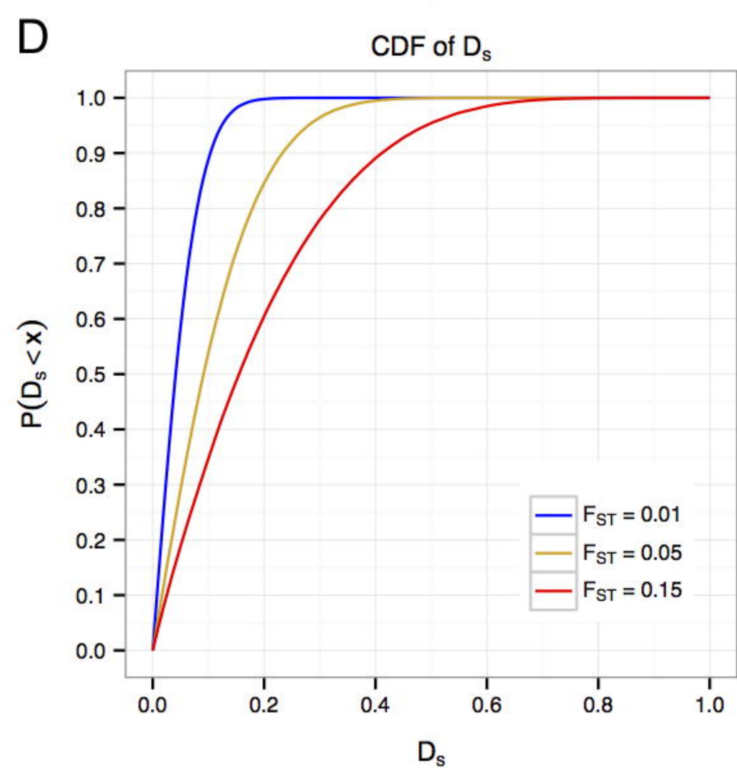

C
CEU/TSI
Mean Test Statistic
D_s

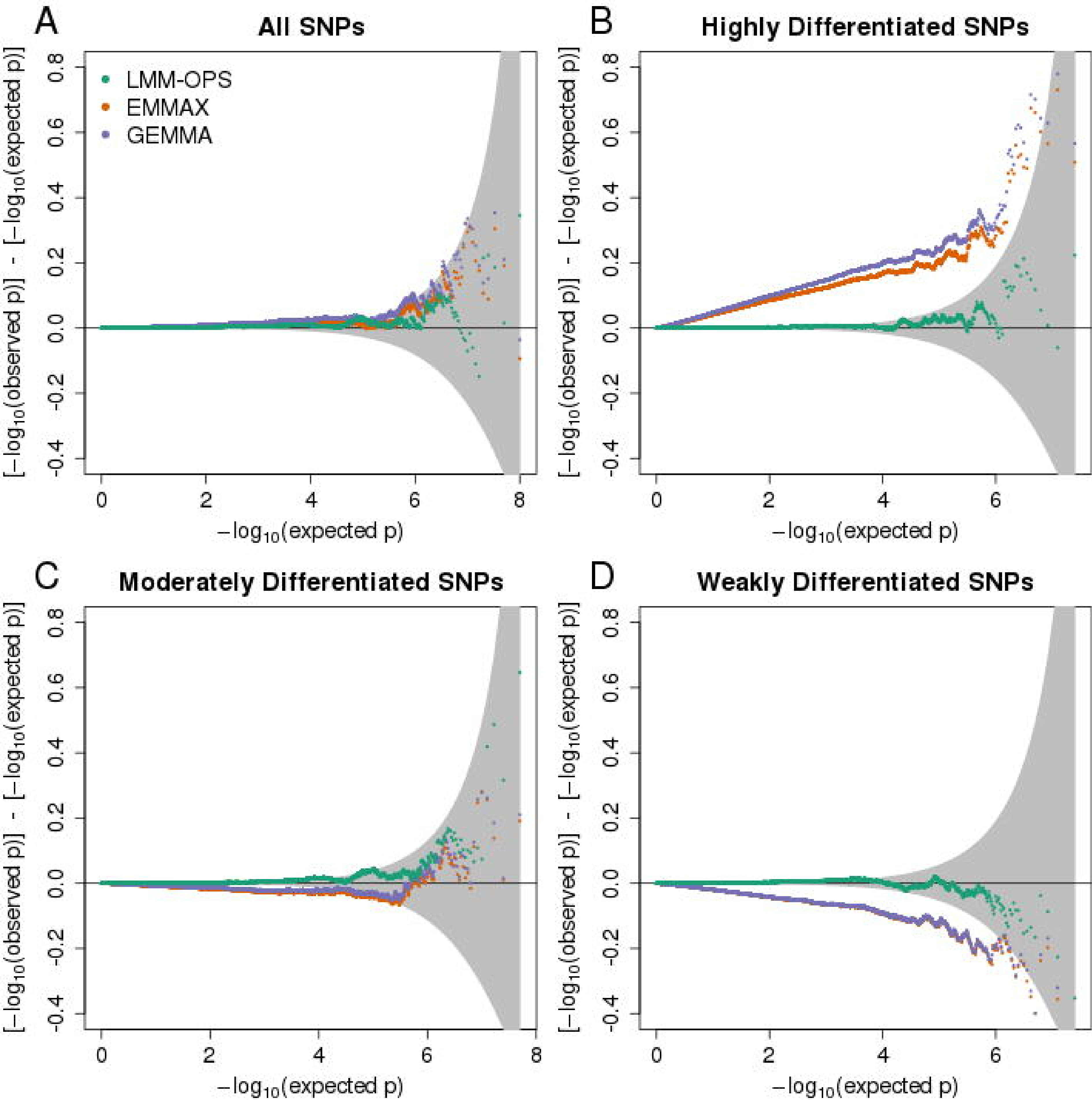

D
CEU/TSI
Mean Test Statistic
D_s


A
$F_{ST} = 0.15$
LMM-OPS
GEMMA
EMMAX
Linear Reg. w/ PCs
Mean Test Statistic
D_s

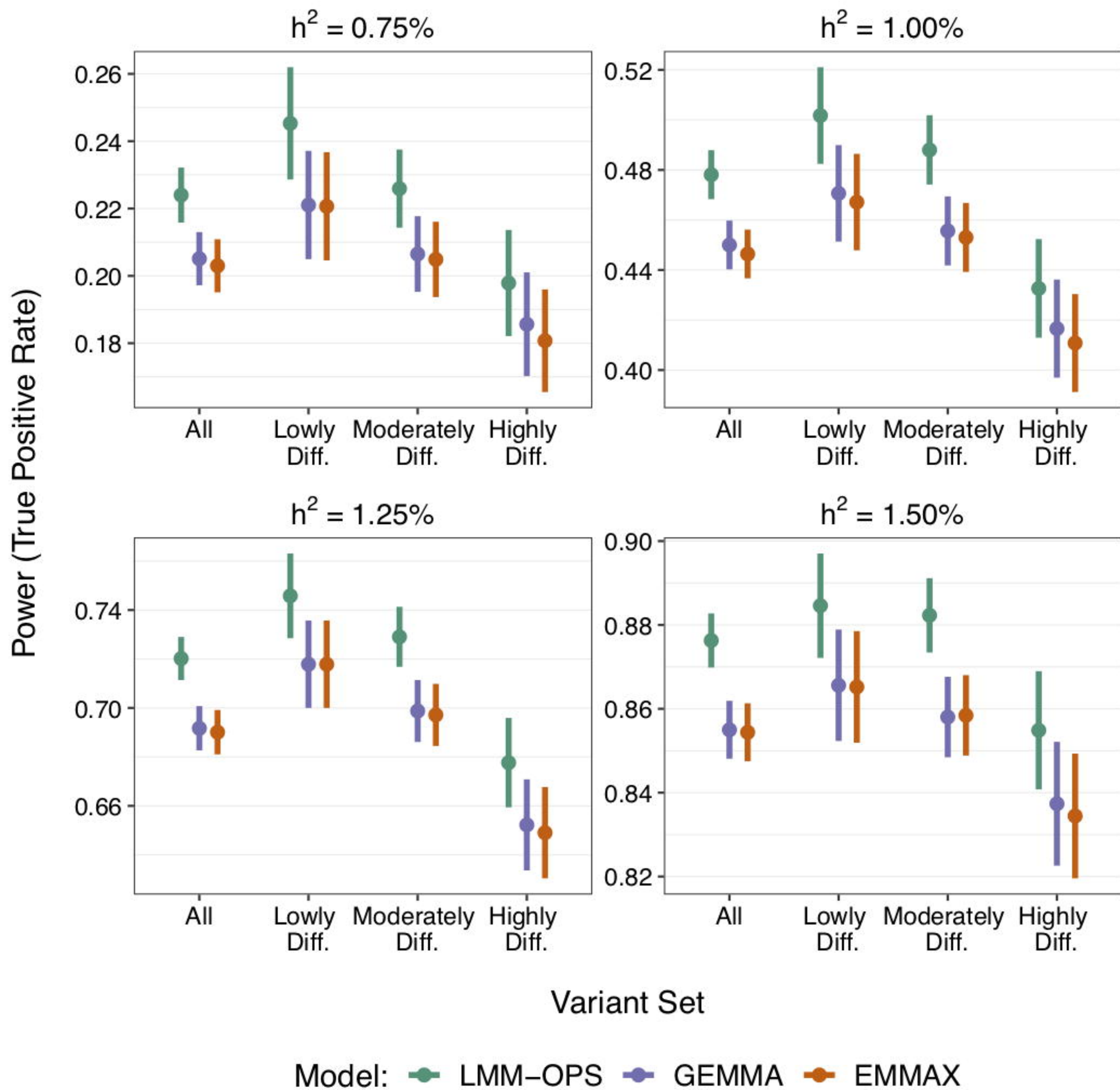

B
$F_{ST} = 0.05$
Mean Test Statistic
D_s

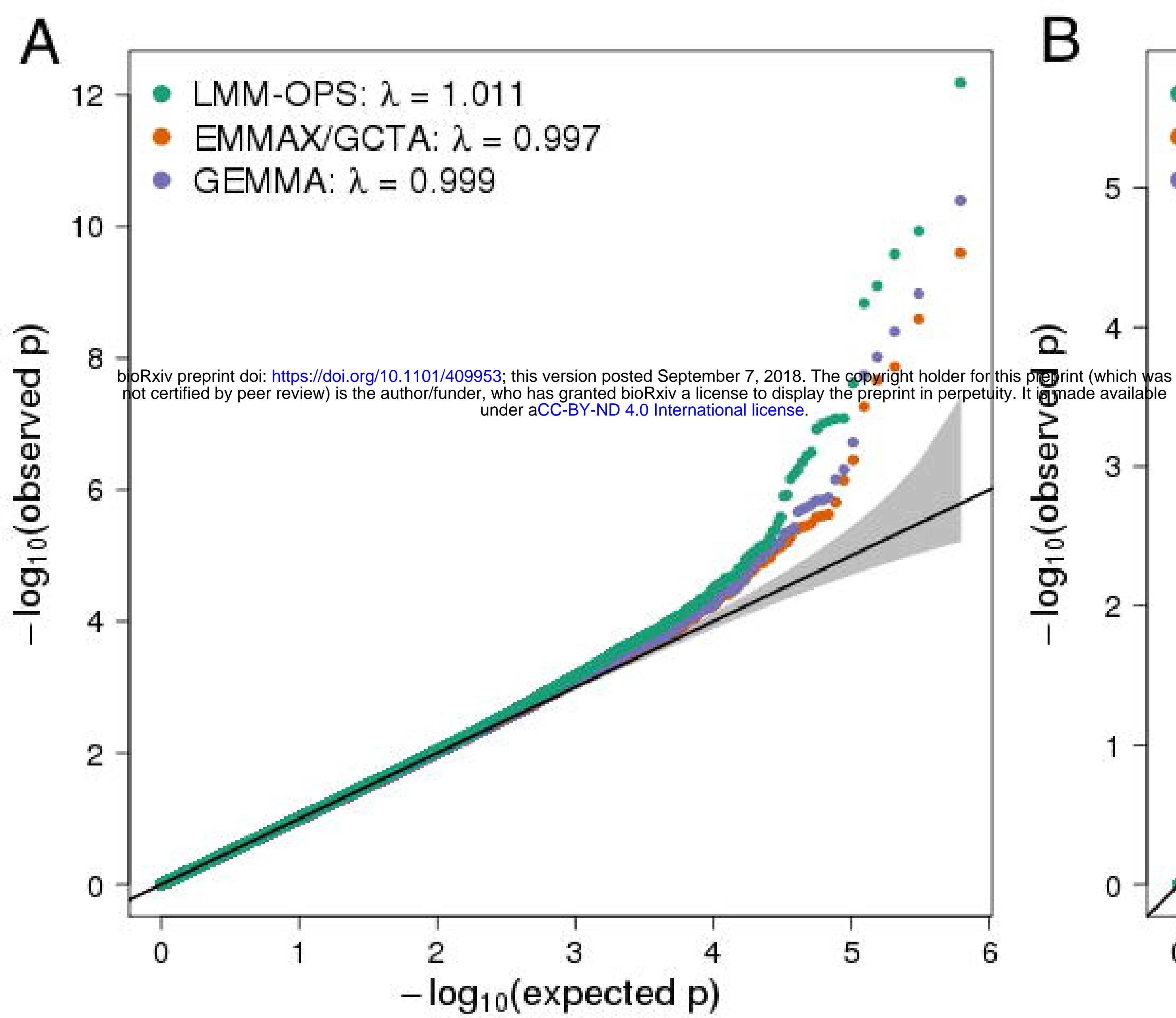

C
$F_{ST} = 0.01$
Mean Test Statistic
D_s


D
CDF of D_s
$P(D_s < x)$
$F_{ST} = 0.01$
$F_{ST} = 0.05$
$F_{ST} = 0.15$
D_s



A
All SNPs
B
Highly Differentiated SNPs
C
Moderately Differentiated SNPs
D
Weakly Differentiated SNPs
LMM-OPS
EMMAX
GEMMA
[−log10(observed p)] - [−log10(expected p)]
−log10(expected p)
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0
2
4
6
8


$h^2 = 0.75\%$
$h^2 = 1.00\%$
$h^2 = 1.25\%$
$h^2 = 1.50\%$
Power (True Positive Rate)
0.26
0.24
0.22
0.20
0.18
0.52
0.48
0.44
0.40
0.74
0.70
0.66
0.90
0.88
0.86
0.84
0.82
All
Lowly Diff.
Moderately Diff.
Highly Diff.
Variant Set
Model: LMM−OPS GEMMA EMMAX



bioRxiv preprint doi: https://doi.org/10.1101/409953; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.