Received: 25 July 2016 | Revised: 1 April 2017 | Accepted: 25 April 2017

DOI: 10.1002/gepi.22056

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Joint genotype- and ancestry-based genome-wide association studies in admixed populations

Piotr Szulc[1] | Malgorzata Bogdan[2] | Florian Frommlet[3] | Hua Tang[4]

[1]Faculty of Mathematics, Wroclaw University of Technology, Wroclaw, Poland

[2]Faculty of Mathematics and Computer Science, University of Wroclaw, Wroclaw, Poland

[3]Department of Medical Statistics, CEMSIIS, Medical University of Vienna, Vienna, Austria

[4]Departments of Genetics and Statistics, Stanford University, Stanford, California, United States of America

Correspondence
Małgorzata Bogdan, Faculty of Mathematics and Computer Science, University of Wroclaw, Plac Grunwaldzki 2/4, 50-384 Wroclaw, Poland.
Email: malgorzata.bogdan@uwr.edu.edu

ABSTRACT

In genome-wide association studies (GWAS) genetic loci that influence complex traits are localized by inspecting associations between genotypes of genetic markers and the values of the trait of interest. On the other hand, admixture mapping, which is performed in case of populations consisting of a recent mix of two ancestral groups, relies on the ancestry information at each locus (locus-specific ancestry). Recently it has been proposed to jointly model genotype and locus-specific ancestry within the framework of single marker tests. Here, we extend this approach for population-based GWAS in the direction of multimarker models. A modified version of the Bayesian information criterion is developed for building a multilocus model that accounts for the differential correlation structure due to linkage disequilibrium (LD) and admixture LD. Simulation studies and a real data example illustrate the advantages of this new approach compared to single-marker analysis or modern model selection strategies based on separately analyzing genotype and ancestry data, as well as to single-marker analysis combining genotypic and ancestry information. Depending on the signal strength, our procedure automatically chooses whether genotypic or locus-specific ancestry markers are added to the model. This results in a good compromise between the power to detect causal mutations and the precision of their localization. The proposed method has been implemented in R and is available at `http://www.math.uni.wroc.pl/~mbogdan/admixtures/`.

KEYWORDS

admixture mapping, model selection, multiple regression, quantitative trait

1 | INTRODUCTION

Genome-wide association studies (GWAS) have proven to be a powerful approach for mapping loci that underly complex traits. GWAS data are most commonly analyzed by testing each marker individually. Given the large number of markers involved, it is crucial to carefully address the problem of multiple testing. Classical approaches include the Bonferroni correction, which controls the family wise error rate (FWER), or the Benjamini-Hochberg procedure (BH; Benjamini & Hochberg, 1995), aimed at controlling the false discovery rate (FDR). In these methods *P* values for single-marker tests are adjusted, taking into account the total number of tests.

When a trait is polygenic, the power of single-marker testing can be improved by adjusting for variation at known trait loci. This partially explains the improved power of the popular mixed effects approaches, such as EMMAX (Kang et al., 2010), which model polygenic background as random effects. Alternatively one can build multilocus models, which have the additional advantage of opening the possibility to incorporate gene-gene or gene-environment interaction terms. A fairly large number of model-based algorithms for complex trait mapping are now available (see, e.g., Dolejsi, Bodenstorfer, & Frommlet, 2014; Hoffman, Logsdon, & Mezey, 2013, and references given there), but not all of them treat the resulting model selection problem rigorously. Regularization-based methods, such as the Least

Absolute Shrinkage and Selection Operator (LASSO, Tibshirani, 1996), focus on prediction, and postselection statistical inference with respect to each selected variable is still an open question. In contrast the methods developed in Frommlet, Ruhaltinger, Twaróg, & Bogdan (2012) and Dolejsi et al. (2014) for GWAS are based on modifications of the Bayesian information criterion called mBIC and mBIC2. These were specifically designed to control FWER and FDR, respectively, in a high dimensional setting. Here, we want to adapt the FDR controlling criterion mBIC2 to the problem of admixture mapping.

Mapping complex traits in populations that have experienced recent admixture, such as the African American and Hispanic populations, are particularly challenging for two reasons. First, recent genetic admixture creates linkage disequilibrium (LD) between unlinked loci, giving rise to spurious association (Cardon & Palmer, 2003; Halder & Shriver, 2003). In GWAS, various methods have been developed, which use high-density genotype data to infer individual-level ancestry proportions; adjusting these ancestry proportions offers an effective solution to eliminate confounding due to population stratification (Price et al., 2006; Tang, Coram, Wang, Zhu, & Risch, 2006). A second challenge, which has no satisfactory solution, is that minority cohorts are much smaller than the available European cohorts. Under a polygenic genetic architecture there is a large number of loci each of which contributes some moderate effects; as a result, for a trait variant with comparable allele frequency and allelic size across populations, the statistical power of detecting this variant is much lower among African Americans or Hispanics than in the European population.

Improving the power of mapping complex traits in recently admixed minority populations is an important goal for several reasons. First, admixed individuals derive their genome from multiple ancestral populations, and therefore offer opportunities to uncover trait variants that are not polymorphic in a single population. Second, even when a trait variant is shared across populations, its allelic effect may vary. Therefore, for genetic risk assessment, it is important to characterize the effect of each trait variant in the relevant population. Third, admixed populations enable mapping of trait loci that underlie population-level trait difference. Admixture mapping, which seeks genomic regions where the phenotype is statistically associated with the ancestry origin of the chromosomal segment, is particularly powerful to map trait variants with disparate allele frequencies in the ancestral populations (Hoggart, Shriver, Kittles, Clayton, & McKeigue, 2004; Winkler, Nelson, & Smith, 2010; Zhu, Tang, & Risch, 2008). Using high-density genotype data, the ancestry information at each locus along the different chromosomes, referred to as local ancestry, can be accurately inferred using a number of computational approaches (Price et al., 2009; Sankararaman, Kimmel, Halperin, & Jordan, 2008; Sundquist, Fratkin, Do, & Batzoglou, 2008; Tang et al., 2006); in this study, we assume that local ancestry is known without error.

It was first proposed in Tang, Siegmund, Johnson, Romieu, and London (2010) to combine genotype- and ancestry-based tests. There it was shown that the two tests provide complementary information and that adding admixture mapping to the genotype-based tests does not significantly increase the burden of multiple testing. Although the approach of Tang et al. (2010) was developed specifically for the parent-trio Transmission Disequilibrium Test (TDT, Spielman, McGinnis and Ewens, 1993), the idea of combining ancestry and genotype information was further explored in Pasaniuc et al. (2011) where a new powerful test for case-control GWAS was proposed. This test is based on the so-called MIX score, which combines association and admixture signals using admixture scores only from cases. However, this interesting idea does not generalize to GWAS for quantitative traits. Alternatively Pasaniuc et al. (2011) combined ancestry and genotype information in a procedure called QSUM, whose test statistic is a sum of two chi-square test statistics: the classical admixture association test and the SNP association test conditioned on a local ancestry.

Subsequently, Shriner, Adeyemo, and Rotimi (2011) proposed a different testing procedure for quantitative traits, which combines genotype and ancestry information using an ad hoc "Bayesian approach." Their procedure called BMIX consists of two stages. First, admixture mapping P values are used to calculate a posterior probability of association, which is then taken as a prior probability for the genotype-based analysis. Now P values from genotype mapping are transformed to obtain a final "posterior probability" which is used to declare association whenever this probability exceeds 0.5. "Bayesian" calculations presented in Shriner et al. (2011) rely on choices of prior probabilities under no association and of the distribution of P values under association that appear to be not entirely justified.

In this study, we develop a rigorous procedure for building multilocus models for complex traits in population-based GWAS, which combines genotype and ancestry information in admixed populations. To this end, we consider regression models with two sets of candidate explanatory variables: one set (X) representing the genotype of each SNP and a second set (Z) representing the local ancestry at the location of the genotyped SNPs. When building models it is important to be aware of the extremely different correlation structure between X variables and between Z variables, respectively. Correlation between neighboring genotype markers is governed by LD, which generally decays rapidly due to historic recombination. In contrast, correlation between local ancestry depends on recombination *after admixing;* in recently admixed populations, such as African Americans or Hispanics, correlation in local ancestry decays much more slowly compared to LD between genotype tests.

We introduce an FDR-controlling modification of BIC, which properly accounts for the differential correlation structure by introducing separate penalty terms for the ancestry and genotype variables. Specifically, we will make use of an "effective number" of ancestry state variables to specify the penalty for the Z variables. After formally introducing the new selection criterion, we compare its performance with the following competing procedures: single-marker tests, multilocus models using only genotype information or only local ancestry, respectively, the QSUM test of Pasaniuc et al. (2011), and the BMIX procedure (Shriner et al., 2011). To this end, we perform a comprehensive simulation study under complex genetic models and also reanalyze GWAS data of HDL cholesterol in an African American cohort. Our methodology has been implemented in R and is available at http://www.math.uni.wroc.pl/~mbogdan/admixtures/.

2 | METHODS

2.1 | Genotype and admixture mapping

Our goal is the identification of DNA regions harboring "causal" mutations based on a sample of n unrelated individuals from the admixture of two distinct populations. We will focus on GWAS with quantitative traits, where the measurement of the trait for the ith individual is denoted as y_i, $i \in \{1, \dots, n\}$. Furthermore, we assume that for each individual the genotypes of p SNPs as well as the corresponding locus-specific ancestry are known. The genotype for the jth SNP of the ith individual is coded as:

$$x_{ij} = \begin{cases} -1 & \text{for } aa \\ 0 & \text{for } aA \\ 1 & \text{for } AA, \end{cases} \tag{2.1}$$

where a and A generically denote the two variants of each SNP. For our purposes, we do not have to know which one is the wild type and which one the mutation. Similarly, the ancestry status of the ith individual at the jth SNP location is coded as

$$z_{ij} = \begin{cases} -1 & \text{for } bb \\ 0 & \text{for } bB \\ 1 & \text{for } BB, \end{cases} \tag{2.2}$$

where b and B refer to the two different ancestral populations. We will use the notation X_j and Z_j for the underlying random variables of genotype and ancestry state at location j.

The simplest way to perform GWAS is to test each genotype or ancestry marker individually for association with the trait in question. To eliminate spurious associations due to population stratification, the genome-wide ancestry, $q_i = \frac{1}{2p}\sum_j (z_{ij} + 1)$, is included in the model as a covariate (Redden et al., 2006). Thus, the standard single-marker genotype-phenotype association test uses the model:

$$y_i = \mu + a_0 q_i + \beta_j x_{ij} + \epsilon_i \tag{2.3}$$

to test the hypotheses $H_{xj} : \beta_j = 0$, for $j = 1, \dots, p$. Likewise, admixture mapping uses the model:

$$y_i = \mu + a_0 q_i + \gamma_j z_{ij} + \epsilon_i, \tag{2.4}$$

to test the hypotheses $H_{zj} : \gamma_j = 0$. We will use for all models the generic notation ϵ_i for error terms and will always assume that they are independent and normally distributed, $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$.

Multilocus models including only genotypic effects can be written as extensions of (2.3) in the form:

$$y_i = \mu + a_0 q_i + \sum_{j \in G} \beta_j x_{ij} + \epsilon_i, \tag{2.5}$$

where G specifies the model by indexing the subset of markers that might influence the trait. We will also write X_G for the submatrix of X that includes only the columns corresponding to the index set G. Important SNPs can then be localized by looking for that model which minimizes some model selection criterion that balances the complexity of the model and its fit to the data. One of the most popular tools for this task is the BIC (Schwarz, 1978), which recommends selecting the model for which

$$\mathrm{BIC}(X_G) = n \log \mathrm{RSS} + k \log n \tag{2.6}$$

obtains a minimal value. Here RSS is the residual sum of squares under least squares regression, and $k := |G|$ refers to the model size. However, in a series of papers (Bogdan et al., 2004; Broman & Speed, 2002; Zak-Szatkowska & Bogdan, 2011) it was shown that in the context of gene mapping, where p is much larger than n and the true model is assumed to be relatively small, this criterion leads to a large number of false discoveries. To solve this problem, various modifications of the BIC were introduced, such as mBIC2 (Frommlet et al., 2012; Zak-Szatkowska & Bogdan, 2011; Frommlet, Bogdan and Ramsey, 2016):

$$\begin{aligned} \mathrm{mBIC2}(X_G) = {} & n \log \mathrm{RSS} + k \log n \\ & + 2k \log(p/C) - 2\log(k!), \end{aligned} \tag{2.7}$$

which was designed to control the FDR of wrongly detected SNPs when both n and p are large but the true number of causal SNPs is relatively moderate. Compared with BIC (Schwarz, 1978), mBIC2 contains the additional penalty term $2k\log(p/C) - 2\log(k!)$, which corrects for "multiple testing" and allows us to keep the fraction of false discoveries under control in the context of GWAS (see Frommlet et al., 2012).

The criterion is consistent (see, e.g. Szulc, 2012), thus its FDR converges to zero and the power converges to 1 when the sample size increases. The choice $C = 4$, recommended, for example, in Frommlet et al. (2012), allows us to keep FDR below 8% for sample sizes $n > 200$ (see Frommlet, Chakrabarti, Murawska & Bogdan, 2010, for detailed calculations).

Similarly, we will consider linear models for the local ancestry state variables Z_j of the form:

$$y_i = \mu + a_0 q_i + \sum_{j \in A} \gamma_j z_{ij} + \epsilon_i, \tag{2.8}$$

with A specifying the set of ancestry markers of the model. We will denote the corresponding design matrix as Z_A. Due to the long-range correlation structure of ancestry state variables, the corresponding penalty for Z variables can be relaxed. A closely related problem occurs for densely spaced markers in experimental crosses discussed in Bogdan et al. (2008), where the selection criterion mBIC was adapted by using an effective number p^{eff} of markers instead of the total number p. Similarly, we will here modify mBIC2 for the ancestry state variables:

$$\begin{aligned} \mathrm{mBIC2}(Z_A) = {} & n \log \mathrm{RSS} + k \log n \\ & + 2k \log(p^{eff}/C) - 2\log(k!), \end{aligned} \tag{2.9}$$

where now $k = |A|$ is the number of ancestry state variables in the model. The effective number of markers p^{eff} can be either calculated using a permutation approach or, if the average admixture time is known, based on the theoretical calculations presented in Appendix B.

To fully exploit the potential of admixture mapping , a new test statistic was proposed in Tang et al. (2010), which combines the genotype and the ancestry information in family-based association studies. Here, we extend this idea to the case of GWAS in admixed populations by combining the multiple regression models (2.5) and (2.8) to include both genotypic and ancestry state variables,

$$y_i = \mu + a_0 q_i + \sum_{j \in G} \beta_j x_{ij} + \sum_{j \in A} \gamma_j z_{ij} + \epsilon_i \ , \tag{2.10}$$

where G denotes the set of genotype variables and A the set of ancestry state variables included in the model. Accordingly, we adapt our model selection criterion to take the form:

$$\begin{aligned} \mathrm{mBIC2}(X_M, Z_A) = {} & n \log \mathrm{RSS} + (k_1 + k_2) \log n \\ & + 2k_1 \log(p/C) + 2k_2 \log(p^{eff}/C) \\ & - 2\log(k_1!) - 2\log(k_2!), \end{aligned} \tag{2.11}$$

where k_1 is the number of genotype variables included in the model and k_2 is the number of ancestry state variables.

Due to the large number of SNPs considered in GWAS, it is not feasible to calculate mBIC2 for all possible regression models. Instead, we apply a modification of the classical greedy step-wise approach to search for a model that minimizes the criterion. As a preliminary step, we perform single marker tests according to the models (2.3) and (2.4). Step-wise search is then only performed for those markers with P values smaller than 0.15. Starting from the null model, we perform the sequence of forward-backward steps. At first, we search through all explanatory variables not yet in the model to identify the one whose inclusion leads to the largest decrease of the mBIC2 criterion. If this "best" marker allows us to improve mBIC2, it is included in the model. When there are no more markers left, which allow to decrease mBIC2, then backward elimination is performed, where in each step that marker is eliminated from the model that gives the largest decrease in mBIC2. Backward elimination is terminated if the removal of any of the remaining markers does not lead to a further decrease of mBIC2. Forward search and backward elimination are performed alternately till convergence. Although this simple step-wise procedure does not guarantee identification of the optimal regression model, our extensive simulations illustrate that it performs very well (i.e., identifies models close to the optimal one) if the number of true causal mutations is small or moderately large. In our simulations, we acknowledge the possibility of slight mislocation of detected SNPs by considering the whole neighborhood of an identified SNP with 30% correlation as the genomic region of potential interest.

2.2 | Simulation study

To investigate the performance of our model selection strategy, we generated data in computer simulations for 1,000 individuals from an admixture of the West African (YRI) and European (CEU) populations as discussed in Price et al. (2008). The ancestry state variables were simulated, based on the hybrid isolation model of Long (1991), for 482,906 autosomal SNPs from the Illumina 650K microarray. The average admixing time was equal to 10 generations, and the average proportion of the genome inherited from YRI was 0.7. The genotype data for the blocks of a given ancestry were obtained by a random selection of individuals of this ancestry from The International HapMap Consortium (2007) genotype data. Because the average admixture time for the simulated data is known, we can avoid the computationally intensive permutation approach and use instead the theory described in the Appendix to derive the effective number of ancestry markers. The resulting number of $p^{eff} = 4,722$ tests for ancestry state variables is approximately 100 times smaller than the total number of SNPs.

We consider three different simulation scenarios. In the first scenario, trait data were generated under the total null

TABLE 1 Selected SNPs and their characteristics

	SNP's Name	MAF	AF	LD
1	ch01_27796	0.455	0.000	0.994
2	ch03_10846	0.418	0.000	0.990
3	ch05_07371	0.414	0.000	0.991
4	ch10_00444	0.488	0.000	0.990
5	ch02_39189	0.432	0.000	0.943
6	ch17_04306	0.495	0.000	0.942
7	ch19_06378	0.466	0.000	0.991
8	ch22_00033	0.485	0.000	0.947
9	ch01_32763	0.430	0.803	0.872
10	ch04_05127	0.461	0.765	0.993
11	ch06_25838	0.428	0.743	0.895
12	ch11_12611	0.491	0.719	0.807
13	ch12_03421	0.419	0.808	0.977
14	ch14_06999	0.414	0.821	0.996
15	ch15_03859	0.401	0.785	0.932
16	ch16_04525	0.426	0.720	0.868
17	ch01_19810	0.497	0.715	0.363
18	ch08_15190	0.400	0.583	0.377
19	ch02_22034	0.456	0.634	0.379
20	ch10_08265	0.492	0.646	0.377
21	ch11_20057	0.447	0.718	0.358
22	ch18_01031	0.431	0.650	0.382
23	ch19_01377	0.499	0.656	0.376
24	ch03_02703	0.497	0.654	0.460

MAF is the minor allelic frequency in the admixed population, AF is the difference in allelic frequencies between both admixing populations and LD is defined as the maximum of genotypic correlation with the 100 nearest SNPs (50 on each side).

hypothesis according to $y_i \sim \mathcal{N}(0,1)$. For the two scenarios imitating complex traits, we determined 24 SNPs to be "causal" (see Table 1). The 24 selected SNPs all have relatively large minor allele frequencies ($MAF \geq 0.4$) and are almost unlinked: the maximal pair-wise correlation coefficient between their genotype state variables does not exceed 0.17 and the maximal correlation between their ancestry state variables is below 0.28. They can be divided into three groups, depending on LD with neighboring markers and the difference in the minor allele frequencies in parental populations. Eight of them were strongly correlated with some neighboring SNPs (for each of them the maximum genotypic correlation with 50 neighboring SNPs in each direction exceeded 0.94) and had the same allelic frequencies in both parental populations. Such a strong correlation in the admixture population implies that these SNPs are also strongly correlated to their neighbors in each of the ancestral population. The second group consisted of eight SNPs that were again strongly correlated with neighboring SNPs but now had substantially different allelic frequencies in the two parental populations (difference in frequency of a given allele exceeded 0.7). The last group contained SNPs whose genotypes were only slightly correlated with genotypes of neighboring SNPs, while the allelic frequencies in both parental populations were substantially different.

For the second and third scenario trait values for each individual, $i = 1, \dots, n$ were simulated using the SNPs from Table 1 as causal according to the following two multiple regression models:

$$\text{Scenario 2:} \quad y_i = 0.5 \sum_{j=1}^{24} x_{ij} + \epsilon_i, \;\; \epsilon_i \sim \mathcal{N}(0,1), \quad (2.12)$$

$$\text{Scenario 3:} \quad y_i = 0.5 \sum_{j \in S_1} x_{ij} + \sum_{j \in S_2} (0.33 x_{ij} + 0.17 z_{ij}) + \epsilon_i, \;\; \epsilon_i \sim \mathcal{N}(0,1). \quad (2.13)$$

The sets S_1 and S_2 are defined as $S_1 = \{1,2,3,4,9,10,11,12,17,18,19,20\}$ and $S_2 = \{1,\dots,24\} \setminus S_1$. Simulations according to the last model (2.13) represent the situation when both SNP and ancestry state have an impact on the trait. This allows us to model the situation when the gene effect is population specific. The coefficients are selected in such a way that the effects of ancestry only modify the genotypic effect at a given location. In each group of eight, we choose half of the SNPs to be population specific. After generating the trait, we eliminated the "causal" SNPs from the design matrix to imitate the common situation where the causal mutation has not been genotyped and can only be detected through its neighboring SNPs. The proportion of phenotypic variance explained by all true causal variants is equal to 0.75 in Scenario 2 and 0.73 in Scenario 3.

In our simulation study, we investigate the performance of mBIC2 when applied only to genotype variables (2.5), only to ancestry variables (2.8), or when combining both types (2.10). In all cases, we use the constant $C = 4$, recommended in Frommlet et al. (2012). For the classical single-marker tests (2.3) and (2.4), we applied both Bonferroni correction and the more liberal BH (Benjamini & Hochberg, 1995) multiple testing procedure. For tests of genotype variables x_{ij} and of ancestry variables z_{ij}, we considered Bonferroni-adjusted significance levels of $0.05/p$ and $0.05/p^{eff}$, respectively. BH was performed at the same nominal levels, which means that the kth smallest P-value was compared to $0.05k/p$ for genotype variables and to $0.05k/p^{eff}$ for ancestry state variables, respectively. Finally, we compared our approach to $BMIX$ (Shriner et al., 2011) and $QSUM$ (Pasaniuc et al., 2011), which also combine genotype and ancestry information. In case of $QSUM$, we used the BH multiple testing procedure, adjusted to the total number of markers (i.e., we compare kth smallest P-value to $0.05k/p$). We decided to use BH rather than the more conservative Bonferroni correction for the fair comparison with mBIC2, which is calibrated to control the FDR.

In the first experiment, we investigated the performance of different methods in the situation when the trait has no genetic component. For this purpose, the trait values were independently generated from the standard normal distribution and 1,000 replicates were simulated to estimate the FWER and the average number of false discoveries. In the second and third experiment, we simulated 250 independent data sets from the complex models (2.12) and (2.13), respectively, and calculated the average number of true and false discoveries (TP and FP) as well as the empirical FDR. In case of mBIC2, we define a detected SNP to be a true positive if the correlation between the variable representing this SNP in the identified model and the respective "causal" variable exceeded 0.3. Here, we always compare variables of the same type (i.e., either ancestry state or genotype variables) with each other. When an ancestry variable in the model is strongly correlated with the ancestry variable at the location of the causal SNP, we count this detection as a true positive, even when the data generating model included only the genotype variable of that location. Two or more detections corresponding to the same causal SNP are counted as just one true positive, all other detections are classified as false positives.

For the multiple testing procedures, one typically observes that detections appear in "clumps" of correlated SNPs, marking the potentially interesting regions of the chromosome. In this case, we use the concept of scan statistics to define false and true discoveries (Siegmund, Yakir, & Zhang, 2011). For each of the detected variables (region seed), we form a detection region consisting of other detected variables of the same type whose correlation with the seed exceeds 0.3. When the detection regions of different SNPs intersect, we combine them to form a larger clump. The clump is considered a true discovery if at least one of its members is strongly correlated with the respective causal variable ($\rho > 0.3$), otherwise it is classified as a false discovery.

To analyze the dependency of our procedure on the sample size, n, and the choice of the scaling constant, C, we performed additional simulations under Scenario 2. We consider n in the range between 650 and 1,000, where smaller samples were obtained by random elimination of individuals from our design matrix. The reported results are based on 250 independent replicates of each experiment.

2.3 | Analysis of HDL in an African American Cohort

We applied the joint genotype-ancestry model to analyze the concentration of high-density lipoprotein (HDL) cholesterol in African Americans, using genotype and phenotype data from the Women's Health Initiative SNP Health Association Resource (WHI-SHARe). The WHI is a U.S.-based study focusing on common health issues in postmenopausal women. Individual characteristics of the participants and genotyping quality assessment analyses are described in Coram et al. (2013); in total, $656,852$ SNPs passed all QC criteria. Here, we reanalyze the log-transformed HDL phenotype in $8,153$ individuals. Genome-wide European ancestry proportions of the individuals are estimated using the program `frappe` (Tang, Peng, Wang, & Risch, 2005), while locus-specific ancestries are estimated using `SABER+` (Johnson et al., 2011). The number of effective ancestry tests $p^{eff} = 6,000$ was calculated based on the permutation approach.

3 | RESULTS

3.1 | Simulation Study

3.1.1 | Scenario 1: Weak sense family wise error rate

The estimated FWER (probability of detecting at least one signal) based on 1,000 replicates for different procedures are presented in Table 2. Note that FWER of the Bonferroni procedure for the ancestry single marker tests slightly exceeds the nominal level of 5%. This may be explained by the fact that markers are not uniformly spaced and admixture times vary between different individuals. Both of these distributional effects might not be captured adequately by using the respective average values in our theoretical formulas. However, our model selection approach controls FWER at the desired level. Due to the consistency of mBIC2 and the relatively large sample size, the FWER for search over ancestry dummy variables with mBIC2 is as low as 2.3%. Enlarging the design matrix by including genotype state variables leads to an increased FWER of 3.4%, which is still substantially below 5%. The $QSUM$ test controls FWER at the level of 2%, which suggests that the applied adjustment to the total number of markers is slightly too conservative. The largest FWER is produced by BMIX at approximately 8%. This is still reasonably small given that the "effective number of tests" used by BMIX is equal to 370, which is much smaller than our own estimate of 4,722. This is probably because BMIX uses the effective number of tests only in a rather informal manner when constructing a prior probability for the expected number of causal mutations.

3.1.2 | Scenario 2 and 3: Complex traits

Table 3 summarizes the simulation results for the two scenarios with complex traits. We start with looking at the results of single-marker tests. Here, as expected, BH has in both scenarios much larger power than the Bonferroni procedure that comes at the price of a substantially larger number of false positives. Comparing mBIC2 with BH for the same type of marker, respectively, one observes that generally speaking

TABLE 2 Comparison of family-wise error rate (FWER) and expected number of false positives (FP) under the global null hypothesis

	Bonferroni		Ben-Hoch		mBIC2			BMIX	QSUM
	X	*Z*	*X*	*Z*	*X*	*Z*	*X+Z*	*X+Z*	*X+Z*
FWER	0.044	0.069	0.044	0.069	0.010	0.023	0.034	0.080	0.031
FP	0.045	0.071	0.046	0.074	0.011	0.024	0.036	0.086	0.032

Bonferroni and Ben-Hoch refer to single-marker tests with Bonferroni and BH procedure at nominal levels 0.05. mBIC2 refers to the model selection approach. Methods with X use only genotypic markers, with Z only ancestral markers, and with $X + Z$ a combination of both types of markers.

TABLE 3 Summary of results for scenarios 2 and 3 in terms of expected true positives (TP), false positives (FP), and false discovery rate (FDR)

	Bonferroni			Benjamini-Hochberg			mBIC2				BMIX	QSUM
	X	*Z*	(*X+Z*)	*X*	*Z*	(*X+Z*)	*X*	*Z*	(*X+Z*)	*X+Z*	*X+Z*	*X+Z*
Scenario 2												
TP	7.93	4.74	10.71	11.78	8.21	15.80	16.33	8.21	19.08	21.01	7.54	11.38
FP	0.07	0.05	0.12	0.96	0.39	1.35	1.09	0.12	1.21	0.77	0.11	0.21
FDR	0.01	0.01	0.01	0.08	0.04	0.08	0.06	0.02	0.06	0.03	0.02	0.02
Scenario 3												
TP	6.61	6.75	11.62	8.92	10.38	15.56	11.60	11.64	17.94	19.76	9.67	12.81
FP	0.10	0.10	0.19	1.04	0.66	1.59	0.80	0.12	0.92	0.62	0.10	0.31
FDR	0.01	0.01	0.02	0.09	0.06	0.09	0.06	0.01	0.05	0.03	0.01	0.02

The columns ($X + Z$) present results obtained by combining the separate analysis over genotype- and ancestry-based GWAS, while the last column $X + Z$ refers to our new model selection approach based on (2.11).

TABLE 4 Power to detect individual SNPs for scenario 2

	Bonferroni		Ben-Hoch		mBIC2			BMIX	QSUM
	X	*Z*	*X*	*Z*	*X*	*Z*	*X+Z*	*X+Z*	*X+Z*
1	0.98	0.00	1.00	0.00	1.00	0.00	1.00 (*Z*: 0.00)	0.56	0.99
2	0.73	0.00	0.92	0.00	0.99	0.00	1.00 (*Z*: 0.00)	0.19	0.79
3	1.00	0.00	1.00	0.00	1.00	0.00	1.00 (*Z*: 0.00)	0.91	1.00
4	0.48	0.00	0.76	0.00	0.99	0.00	0.97 (*Z*: 0.00)	0.04	0.54
5	1.00	0.00	1.00	0.00	1.00	0.00	1.00 (*Z*: 0.00)	0.93	1.00
6	0.40	0.00	0.68	0.00	1.00	0.00	1.00 (*Z*: 0.00)	0.00	0.44
7	0.64	0.00	0.91	0.00	1.00	0.00	1.00 (*Z*: 0.00)	0.39	0.80
8	0.26	0.00	0.66	0.00	0.99	0.00	1.00 (*Z*: 0.00)	0.01	0.34
9	0.22	0.50	0.54	0.84	0.74	0.85	0.98 (*Z*: 0.74)	0.42	0.54
10	0.62	0.53	0.92	0.83	0.99	0.67	1.00 (*Z*: 0.10)	0.67	0.73
11	0.21	0.19	0.59	0.49	0.96	0.47	0.96 (*Z*: 0.20)	0.34	0.35
12	0.00	0.01	0.02	0.11	0.96	0.02	0.87 (*Z*: 0.15)	0.04	0.04
13	0.58	0.76	0.88	0.94	1.00	0.82	1.00 (*Z*: 0.30)	0.71	0.72
14	0.10	0.31	0.39	0.71	0.97	0.82	1.00 (*Z*: 0.06)	0.45	0.57
15	0.17	0.09	0.55	0.40	0.88	0.53	0.98 (*Z*: 0.24)	0.34	0.30
16	0.54	0.85	0.88	0.97	1.00	0.94	1.00 (*Z*: 0.08)	0.62	0.83
17	0.00	0.00	0.00	0.00	0.00	0.00	0.35 (*Z*: 0.35)	0.00	0.01
18	0.00	0.00	0.00	0.04	0.01	0.03	0.25 (*Z*: 0.23)	0.05	0.08
19	0.00	0.00	0.00	0.02	0.01	0.25	0.57 (*Z*: 0.56)	0.00	0.01
20	0.00	0.62	0.04	0.88	0.32	0.69	0.92 (*Z*: 0.91)	0.46	0.60
21	0.00	0.32	0.02	0.72	0.22	0.78	0.97 (*Z*: 0.95)	0.27	0.40
22	0.00	0.19	0.00	0.52	0.03	0.67	0.90 (*Z*: 0.90)	0.14	0.19
23	0.00	0.38	0.00	0.74	0.01	0.65	0.68 (*Z*: 0.68)	0.03	0.12
24	0.00	0.00	0.01	0.00	0.26	0.00	0.62 (*Z*: 0.51)	0.00	0.00

Large power is marked by red and small power by blue. In the last column $X + Z$, the values in brackets give the percentage of cases for which detection resulted from ancestry dummy variables.

mBIC2 has larger or comparable power and at the same time lower FDR. Compared to single-marker tests operating on individual variables, $BMIX$ does not seem to provide any gain in power. Specifically, it has a similar FDR, but a substantially smaller power than the Bonferroni correction applied to both marker types ($X + Z$). In our simulation study, $BMIX$ is substantially outperformed by $QSUM$, which keeps FDR at a similar level but offers a significantly increased power. In both simulation scenarios, the power of $QSUM$ exceeds the power of the combined Bonferroni test. This result differs from the conclusion of the simulation study in Pasaniuc et al. (2011), where $QSUM$ had systematically lower power than the genotype-based Armitrage test. We believe that this is due to the fact that in our simulations we removed "causal" mutations from the list of available markers and consequently $QSUM$ gains from the advantage of admixture mapping for detecting variants in regions of low LD. In our simulation study all single marker approaches are decisively outperformed by the new mBIC2 procedure that combines genotype and local ancestry information. Specifically, mBIC2 keeps the FDR at the level of 3% and detects much more "causal" variants than any of the single marker approaches.

We will next discuss the power of different procedures to detect individual SNPs as summarized in Tables 4 and 5, which provide the percentage of simulation runs for which a corresponding neighborhood was identified. For genotypic markers, the selection based on mBIC2 applied separately to X has in the majority of cases larger power than BH. An interesting example is provided by SNP number 12 in Scenario 2. This SNP has a relatively large LD (0.807) and is detected with a power of 96% by mBIC2 applied to the genotype data matrix X. On the other hand, both single-marker tests have very small power to detect this SNP. This phenomenon can be easily explained by the fact that the estimator of the regression coefficient in a single-marker test does not necessarily represent the strength of the effect of a given SNP, but depends also on the correlations between this SNP and the genome-wide ancestry q and other causal variants (see Frommlet et al.,

TABLE 5 Power to detect individual SNPs for scenario 3

	Bonferroni		Ben-Hoch		mBIC2			BMIX	QSUM
	X	*Z*	*X*	*Z*	*X*	*Z*	*X+Z*	*X+Z*	*X+Z*
1	1.00	0.00	1.00	0.00	1.00	0.00	1.00 (*Z*: 0.00)	0.84	0.99
2	0.89	0.00	0.98	0.00	0.98	0.00	0.98 (*Z*: 0.00)	0.34	0.91
3	1.00	0.00	1.00	0.00	1.00	0.00	1.00 (*Z*: 0.00)	0.80	0.99
4	0.80	0.00	0.95	0.00	0.98	0.00	0.95 (*Z*: 0.00)	0.33	0.86
5	0.82	0.00	0.95	0.04	0.71	0.01	0.90 (*Z*: 0.00)	0.88	0.97
6	0.00	0.00	0.02	0.00	0.39	0.03	0.60 (*Z*: 0.16)	0.00	0.03
7	0.00	0.04	0.06	0.22	0.36	0.20	0.54 (*Z*: 0.18)	0.46	0.50
8	0.00	0.00	0.01	0.04	0.12	0.10	0.53 (*Z*: 0.10)	0.07	0.11
9	0.58	0.72	0.86	0.95	0.64	0.97	0.98 (*Z*: 0.73)	0.72	0.80
10	0.73	0.68	0.92	0.92	0.97	0.89	1.00 (*Z*: 0.28)	0.77	0.84
11	0.30	0.27	0.62	0.63	0.90	0.82	0.96 (*Z*: 0.30)	0.42	0.54
12	0.04	0.09	0.13	0.37	0.81	0.30	0.86 (*Z*: 0.40)	0.11	0.16
13	0.34	0.98	0.66	1.00	0.95	0.99	1.00 (*Z*: 0.94)	0.74	0.91
14	0.05	0.49	0.25	0.84	0.58	0.98	1.00 (*Z*: 0.91)	0.53	0.65
15	0.04	0.46	0.21	0.84	0.62	0.96	0.99 (*Z*: 0.81)	0.28	0.45
16	0.00	0.02	0.00	0.11	0.02	0.15	0.15 (*Z*: 0.09)	0.01	0.05
17	0.00	0.00	0.00	0.00	0.00	0.03	0.45 (*Z*: 0.45)	0.00	0.01
18	0.00	0.04	0.01	0.28	0.02	0.23	0.45 (*Z*: 0.42)	0.43	0.46
19	0.00	0.00	0.00	0.08	0.00	0.66	0.77 (*Z*: 0.76)	0.00	0.01
20	0.02	0.93	0.15	0.99	0.16	0.89	0.82 (*Z*: 0.82)	0.82	0.90
21	0.01	0.60	0.11	0.92	0.26	0.99	0.99 (*Z*: 0.99)	0.42	0.57
22	0.00	0.56	0.01	0.88	0.02	0.94	0.99 (*Z*: 0.99)	0.31	0.45
23	0.00	0.85	0.00	0.96	0.03	0.94	0.97 (*Z*: 0.97)	0.20	0.56
24	0.00	0.03	0.01	0.32	0.09	0.57	0.88 (*Z*: 0.87)	0.04	0.11

2012, for more details). In fact, the power of detecting a given SNP by single-marker tests depends on the so-called noncentrality parameter that captures these inter-SNP correlations (see Fig. 1) and is calculated according to the formula:

$$(X\beta)'\left(\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}' - \frac{1}{n}E\right)(X\beta).$$

Here X denotes the matrix containing genotypes of all causal SNPs, β is the vector of true regression coefficients, $\tilde{X}$ contains all variables included in the single-marker test model (2.3; i.e., the columns of ones, q_i and x_{ij}), where the causal mutation is replaced by the most correlated SNP that was genotyped and $E_{n\times n}$ is the matrix for which all elements are equal to 1.

Figure 1 shows that for SNP number 12, the noncentrality parameter takes a rather small value, and thus it cannot be detected by single-marker tests. The main reason for this phenomenon seems to be a relatively large correlation between the genotype state variable of SNP12 and the genome-wide ancestry q, which is equal to 0.31. Thus, a large part of the effect of SNP12 is taken over by q. This side effect of conditioning on q diminishes when the multiple regression model is used and the effects of other causal SNPs are estimated and removed from the background noise.

In Figure 1, we also highlight SNP number 17, which has high AF (0.715), but when we use only the matrix Z, even mBIC2 has no power to detect it (the noncentrality parameter for this SNP is the lowest of all). However, it was found in 35% of the simulation runs when both matrices are used.

Looking at the results specifically in terms of the three different types of causal SNPs used in the simulation study, then Table 4 indicates that combining genotype and ancestry data increases power particularly for mutations in regions of low LD. In the combined analysis of model (2.10), SNPs of the group 17–24 were almost exclusively detected by ancestry state variables, and in almost all cases the power to detect these SNPs was substantially lower when using model (2.8) with the Z matrix alone. The explanation for this is that in the combined model the effect of detected genotype variables is removed from the residual error and thus it becomes easier to detect further ancestry state variables. For Scenario 3, the gain in power for the last eight SNPs by combining genotype and ancestry data is smaller somewhat than in Scenario 2, but there is additional gain in power obtained for SNPs 5–8. These SNPs of the first group are not easily detected when searching over X or Z variables separately, whereas the combined approach yields substantial power for all four SNPs (especially for SNP 8). The reason for this gain in power is the very same as in Scenario 2. By including many other causal SNPs in the model the residual sum of squares is reduced and the chance is increased to detect these SNPs. Comparing Scenario 2 with Scenario 3 furthermore shows that ancestry variables play a substantial role in identifying mutations whose effect is population specific. In Scenario 2, SNPs 13–16 are mainly detected by X variables, whereas in Scenario 3, they are identified mainly with Z variables.

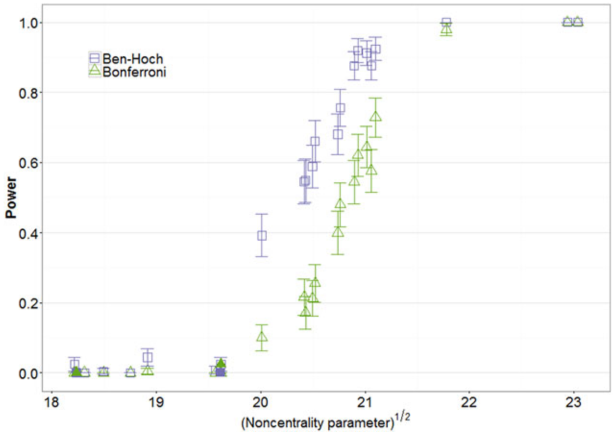

FIGURE 1 Power for individual tests (matrix X) versus noncentrality parameter for Scenario 2. Green triangles and violet squares mark two SNPs discussed in the text: SNP12 (low power for Bonf and BH, but high power for mBIC2) is near the middle of the picture, SNP17 (high AF, but seen only for X+Z) on the left. Error bars represent 95% confidence intervals for means

The final example of this section will discuss the working principle of our new approach to combine genotypic and ancestry state data. Figure 2 illustrates the role of X and Z in detecting SNP 14 in Scenario 2 as a function of the sample size. For smaller sample sizes, this SNP is detected mainly by the ancestry state variable Z, whereas for larger n it is more frequently identified by the genotype variable X. This is due to the fact that for small sample sizes, the reduced

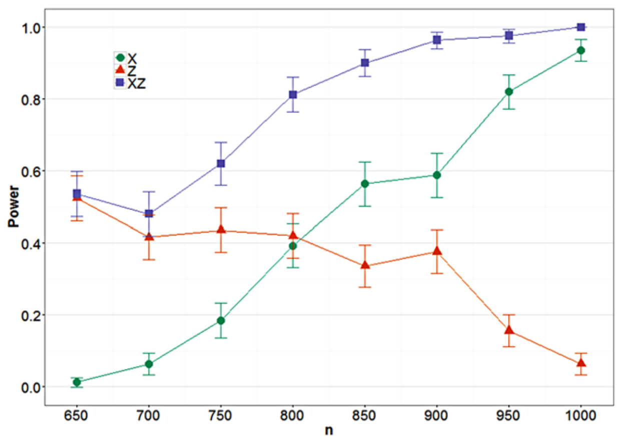

FIGURE 2 Power for SNP 14 in Scenario 2 for different sample size n based on model selection with mBIC2 using both genetic and ancestry markers ($X + Z$). Here, the lines denoted by X and Z illustrate the proportion of detections due to genetic markers and ancestry markers, respectively, for the combined approach. Error bars represent 95% confidence intervals for means

multiple testing correction used in admixture mapping substantially enhances the power. However, when the sample size is large enough, the causal gene can be detected by the genotype variable, which provides a more precise localization. Note that in this argument the sample size can be substituted by the magnitude of the gene effect. Consequently in our admixture mapping approach, the ancestry state variables are helpful for detecting genes with small effects, which however comes at the price of rather imprecise estimation of their location. The major advantage of our combined mBIC2 criterion (2.11) is that it allows us to decide automatically for every SNP whether selection is based on X or Z variables, depending on the magnitude of its effect.

3.2 | WHI-SHARe HDL analysis

The joint genotype-ancestry analysis produces a multivariate model that includes seven genotype variables and two locus-specific ancestry variables. Table 6 lists these variables in the order they enter the model, as well as the single-marker P values from either genotype-based or admixture mapping analysis. The first seven terms to enter the multivariate model, including six SNP genotype variables and the locus-specific ancestry on chromosome 11, coincided with the genome-wide significant findings through single-marker analyses (Coram et al., 2013). Interestingly, the next term to enter the multivariate model was the local ancestry on chromosome 17q, which does not meet genome-wide significance in a single-marker admixture mapping analysis ($P = 5.80 \times 10^{-5}$). In contrast, the local ancestry at 9q22, which is significant in admixture mapping ($P = 5.58 \times 10^{-7}$) was not selected by the multivariate model. We verified that given the first seven variables entered in the model, adding local ancestry on 17q indeed led to a greater improvement in model fitting than adding the local ancestry of 9q22: the multiple R^2 statistics were 0.04884 and 0.04854, respectively. Of course, without an independent validation data, we cannot say which model will have better predictive value; however, these results illustrate that a multivariate model can prioritize variables according to a different order than a single-variable approach. Finally, the last variable to enter the model, SNP rs7249565 has a P value of 1.13×10^{-5} in a single-marker test; after including all the other eight variables, its P value (corresponding to the main effect of this SNP in the multivariate model) decreased to 1.10×10^{-7}, presumably because the other variables reduced the estimated residual variance.

4 | DISCUSSION

The presented simulation study and real data analysis demonstrate that our model selection approach efficiently integrates the genetic and ancestry information. As expected, in comparison to classical GWAS inclusion of ancestry state variables results in a substantial increase of power to detect causal

TABLE 6 Results of real data analysis using mBIC2

	Single Genotype Test	Single Ancestry Test	Multivariate Model
CH16rs247617	1.48E-44		< 2E-16
CH7rs6963015	7.53E-10		7.96E-11
CH8rs326	1.23E-08		8.92E-09
CH8rs1461729	7.39E-09		1.04E-08
CH21rs13046373	2.26E-08		3.35E-08
CH19rs12979813	1.99E-09		1.79E-09
CH11rs531964anc		2.82E-07	2.05E-06
CH17rs11867417anc		5.80E-05	4.38E-06
CH19rs7249565	1.13E-05		1.10E-07

SNPs are presented in the order they enter the multivariate model. SNPs identified by the ancestry dummy variables are marked by suffix “anc” after the SNP name. Last column contains P values in the multiple regression model including all SNPs identified by mBIC2, while first and second columns contain P values from single marker genotype or ancestry tests, depending on the type of variable used to select a given SNP.

mutations in the regions of low LD. Also, in comparison to admixture mapping, our approach yields a substantially larger power to detect "admixture effects," due to reducing the residual error by including detected genotype state variables in the model. Interestingly, inclusion of admixture state variables usually does not deflate the power of detecting genes in regions of high LD, because the long range correlations between admixture variables make it possible to use a relatively small additional correction for multiple testing. The corresponding increase in the penalty of the selection criterion is most often counterbalanced by the reduction of the residual error due to detected admixture variables. Furthermore, our simulation study shows that the admixture state variables help to detect causal mutations in regions of high LD, if the gene signal is weak or the sample size is small. The presented model selection approach provides an automatic choice between genotype and ancestry state variables, which yields high power of gene detection if the sample size is small (choice of admixture variable) and high precision of gene localization if the sample size is large enough (choice of genotype variable).

Our model selection approach creates a general framework for GWAS in admixture populations, which, apart from other advantages, automatically incorporates the dependency of gene effects on the population-specific genetic background. The search procedure gives a high power of gene detection while keeping the number of false discoveries under control. The approach can be easily extended for case-control studies and any other trait distribution that can be modeled by generalized linear models (see Zak-Szatkowska & Bogdan, 2011, or Dolejsi et al., 2014). According to Zak, Baierl, Bogdan, and Futschik (2007) one can also expect good performance of a rank-based version of our criterion in case when the trait distribution is heavy tailed or the data contains some proportion of outliers. These assertions will be investigated in some follow-up research.

ACKNOWLEDGMENTS

The authors thank two anonymous referees, whose comments helped to improve the presentation. They also thank David O. Siegmund for discussions and helpful suggestions. PS and MB were supported by the European Union's 7th Framework Programme for research, technological development, and demonstration under grant agreement no. 602552, cofinanced by the Polish Ministry of Science and Higher Education under grant agreement 2932/7.PR/2013/2. HT was supported by the US National Institutes of Health grants GM073059. Large part of this research was performed during MB visit at Stanford University, financed by the Advanced Researchers Fellowship from the Fulbright Association.

REFERENCES

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *57*, 289–300.

Bogdan, M., Frommlet, F., Biecek, P., Cheng, R., Ghosh, J. K., & Doerge, R. W. (2008). Extending the modified Bayesian information criterion (mBIC) to dense markers and multiple interval mapping. *Biometrics*, *64*, 1162–1169.

Bogdan, M., Ghosh, J. K., & Doerge, R. W. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, *167*, 989–999.

Broman, K. W., & Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *64*, 641–656.

Cardon, L. R., & Palmer, L. J. (2003). Population stratification and spurious allelic association. *Lancet*, *361*, 598-604.

Coram, M. A., Duan, Q., Hoffmann, T. J., Thornton, T., Knowles, J. W., Johnson, N. A., ... Tang, H. (2013). Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. *American Journal of Human Genetics*, *92*(6), 904–916.

Dolejsi, E., Bodenstorfer, B., & Frommlet, F. (2014). Analyzing genome-wide association studies with an FDR controlling modification of the Bayesian information criterion. *PLOS ONE*, *9*(7), e103322.

Feingold, E., Brown, P. O., & Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high resolution maps of identity-by-descent. *American Journal of Human Genetics*, *53*, 234–251.

Frommlet, F., Bogdan, M., & Ramsey, D. (2016). *Phenotypes and Genotypes: the Search for Influential Genes*. Springer.

Frommlet, F., Chakrabarti, A., Murawska, M., & Bogdan, M. (2010). Asymptotic Bayes optimality under sparsity for generally distributed effect sizes under the alternative. *Technical Report*, https://doi.org/arxiv.org/abs/1005.4753

Frommlet, F., Ruhaltinger, F., Twaróg, P., & Bogdan, M. (2012). A model selection approach to genome wide association studies. *Computational Statistics and Data Analysis*, *56*, 1038–1051.

Halder, I., & Shriver, M. (2003). Measuring and using admixture to study the genetics of complex diseases. *Human Genetics*, *1*, 52–62.

Hoffman, G. E., Logsdon, B. A., & Mezey, J. G. (2013). PUMA: A unified framework for penalized multiple regression analysis of GWAS data. *Plos Computational Biology*, *9*(6), e1003101.

Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G., & McKeigue, P. M. (2004). Design and analysis of admixture mapping studies. *American Journal of Human Genetics*, *274*(5), 965–978.

Johnson, N. A., Coram, M. A., Shriver, M. D., Romieu, I., Barsh, G. S., London, S. J., & Tang, H. (2011). Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genetics*, *7*(12), e1002410.

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

Long, J. C. (1991). The genetic structure of admixed populations. *Genetics*, *127*, 417–428.

Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G. K., Tandon, A., Kao, W. H. L., (…) Price, A. L. (2011). Enhanced statistical tests for GWAS in admixed populations: Assessment using African Americans from CARe and a breast cancer consortium. *PLoS Genetics*, *7*(4), e1001371. https://doi.org/10.1371/journal.pgen.1001371

Price, A. L., Patterson, N. J., Hancks, D. C., Myers, S., Reich, D., Cheung, V. G. & Spielman, R. S. (2008). Effects of *cis* and *trans* genetic ancestry on gene expression in African Americans. *PLOS Genetics*, *4*(12), e1000294.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909.

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., ... Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, *5*, e1000519.

Redden, D. T., Divers, J., Vaughan, L. K., Tiwari, H. K., Beasley, T. M., Fernandez, J. R., ... Allison, D. B. (2006). Regional admixture mapping and structured association testing: Conceptual unification and an extensible general linear model. *PLoS Genetics*, *2*, e137.

Sankararaman, S., Kimmel, G., Halperin, E., & Jordan, M. I. (2008). On the inference of ancestries in admixed populations. *Genome Research*, *18*, 668–675.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* *6*(2), 461–464.

Shriner, D., Adeyemo, A., & Rotimi, C. N. (2011). Joint ancestry and association testing in admixed individuals. *PLoS Computational Biology*, *7*(12), e1002325. https://doi.org/10.1371/journal.pcbi.1002325

Siegmund, D., & Yakir, B. (2007). *The Statistics of Gene Mapping*. Springer Science & Business Media.

Siegmund, D., Yakir, B., & Zhang, N. R. (2011). The false discovery rate for scan statistics. *Biometrika*, *98*(4), 979–985.

Spielman, R. S., McGinnis, R.E., & Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. *American Journal of Human Genetics*, *52*, 506–516.

Sundquist, A., Fratkin, E., Do, C. B., & Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, *18*, 676–682.

Szulc, P. (2012). Weak consistency of modified versions of Bayesian information criterion in a sparse linear regression. *Probability and Mathematical Statistics*, *32*, 47–55.

Tang, H., Coram, M., Wang, P., Zhu, X., & Risch, N. J. (2006). Reconstructing genetic ancestry blocks in admixed populations. *American Journal of Human Genetics*, *79*, 1–12.

Tang, H., Peng, J., Wang, P., & Risch, N. J. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, *28*(4), 289–301.

Tang, H., Siegmund, D. O., Johnson, N. A., Romieu, I., & London, S. J. (2010). Joint testing of genotype and ancestry association in admixed families. *Genetic Epidemiology*, *34*, 783–791.

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*, 851–861.

Tibshirani, R. (1996). Regression shrinkage and selection via lasso. *Journal of the Royal Statistical Society B*, *58*, 267–288.

Winkler, C. A., Nelson, G. W., & Smith, M. W. (2010). Admixture mapping comes of age. *Annual Review of Genomics and Human Genetics*, *11*, 65–89.

Zak, M., Baierl, A., Bogdan, M., & Futschik, A. (2007). Locating multiple interacting quantitative trait loci using rank-based model selection. *Genetics*, *176*, 1845–1854.

Zak-Szatkowska, M., & Bogdan, M. (2011). Modified versions of Bayesian information criterion for sparse generalized linear models. *Computational Statistics and Data Analysis*, *55*, 2908–2924.

Zhu, X., Tang, H., & Risch, N. (2008). Admixture mapping and the role of population structure for localizing disease genes. *Advances in Genetics*, *60*, 547–569.

How to cite this article: Szulc P, Bogdan M, Frommlet F, Tang H. Joint genotype- and ancestry-based genome-wide association studies in admixed populations. *Genet Epidemiol*. 2017;41:555–566. https://doi.org/10.1002/gepi.22056

APPENDIX A: CORRELATIONS BETWEEN ANCESTRY AND GENOTYPE STATE VARIABLES OF MARKERS USED IN THE SIMULATION STUDY

In this section, we provide correlations between ancestry and genotype state variables of markers used in the simulation study, which are located on the same chromosome.

For two SNPs on chromosome 11: CH11_12611 and CH11_20057—the correlation between their genotype state variables X is equal to 0.13 and between their ancestry state variables Z is equal to 0.23.

TABLE A1 Chromosome 1: Correlations between genotype state variables X

	CH01_19810	CH01_27796	CH01_32763
CH01_19810	1.00	0.01	0.10
CH01_27796	0.01	1.00	0.01
CH01_32763	0.10	0.01	1.00

TABLE A2 Chromosome 1: Correlations between ancestry state variables Z

	CH01_19810	CH01_27796	CH01_32763
CH01_19810	1.00	0.16	0.16
CH01_27796	0.16	1.00	0.20
CH01_32763	0.16	0.20	1.00

APPENDIX B: CALCULATION OF THE EFFECTIVE NUMBER OF ANCESTRY TESTS IN THE SIMULATION STUDY

First, assume that we perform a genome scan based on ancestry markers that are equally spaced at a distance of L Morgan. To assess the necessary multiple testing correction for the ancestry markers, we consider the simple multiple regression models (2.4) where we are simultaneously testing the null hypotheses $H_{0j} : \gamma_j = 0$. Consider an individual with admixing time t and a genome-wide ancestry value q_i. Elementary calculations show that the conditional correlation between ancestry state variables at the neighboring loci does not depend on the specific ancestry value, but only on the individual admixing time and is given by:

$$\rho := \mathrm{Corr}(z_{ij}, z_{i(j+1)}|q_i) = \exp(-tL).$$

Moreover, based on the arguments presented in [Feingold, Brown and Siegmund (1993)] (see also [Siegmund and Yakir (2007)]), the sequence of test statistics at consecutive locations can be approximated by the square of an Ornstein-Uhlenbeck process. One then can show that the FWER α of such a search is approximately:

$$\begin{aligned}\alpha &= P_{H_0}\left(max_{j\in\{1,\dots,p\}} LRT_j > c\right)\\ &\approx 1 - exp\left(-2\left[1-\Phi\left(\sqrt{c}\right)\right]\right) - 0.02\varphi\left(\sqrt{c}\right) ptL\sqrt{c}\nu\left(\sqrt{0.02tLc}\right),\end{aligned}\tag{B.1}$$

where $\Phi(\cdot)$ and $\varphi(\cdot)$ denote the cumulative distribution function and the density of the standard normal distribution and

$$\nu(t) \approx \frac{(2/t)(\Phi(t/2) - 0.5)}{(t/2)\Phi(t/2) + \varphi(t/2)}.\tag{B.2}$$

On the other hand, the FWER resulting from performing p^{eff} independent tests is equal to

$$\alpha = P_{H_0}\left(\max_{i\in\{1,\dots,p^{eff}\}} LRT_j > c\right) \approx 1 - \left[1 - 2\left(1 - \Phi\left(\sqrt{c}\right)\right)\right]^{p^{eff}}.\tag{B.3}$$

Comparing (B.1) and (B.3) results in the following effective number of tests for the ancestry state variables

$$p^{eff} = \log(1-\alpha)/\log\left(2\Phi\left(\sqrt{c}\right) - 1\right).$$

As observed in Bogdan et al. (2008), the dependency of p^{eff} on α is rather weak and the value of p^{eff} calculated for $\alpha = 0.05$ can be used as a good approximation for p^{eff} corresponding to any $\alpha \in (0, 0.1]$. In the simulation study, we calculated the effective number of tests separately for each chromosome. Because the average admixture time is equal to 10, for this analysis, we replaced tL in equation (B.1) with $10 \times \bar{L}_j$, where $\bar{L}_j$ is the average distance between neighboring markers on a given chromosome. Finally, the effective number of SNPs for mBIC2 was obtained by adding the effective number of tests on each chromosome, which results in $p^{eff} = 4,722$.