Ancestry specific association mapping in admixed populations

Line Skotte[1*], Thorfinn Sand Korneliussen[2], Ida Moltke[3]
and Anders Albrechtsen[1]

January 20, 2015

1: The Bioinformatics Centre, Department of Biology,
University of Copenhagen, 2200 Copenhagen, Denmark

2: Centre for GeoGenetics, Natural History Museum of Denmark,
University of Copenhagen, 1350 Copenhagen, Denmark

3: Department of Human Genetics,
University of Chicago, Chicago, IL 60637, USA

*: Corresponding author: line@lineskotte.dk

Abstract

Genetic association studies have during the last decade proven to be a powerful approach to identifying disease-causing variants. However, many populations, like the Greenlandic Inuit population, have recently experienced substantial admixture with other populations, which can complicate association studies. One important complication is that most current methods for performing association testing are based on the assumption that the effect of the tested genetic marker is the same regardless of ancestry. This is a reasonable assumption for a causal variant, but may not hold for the genetic markers that are tested in association studies, which are usually not causal. The effects of non-causal genetic markers depend on how strongly their presence correlate with the presence of the causal marker, and this may vary between ancestral populations because of different linkage disequilibrium patterns and allele frequencies.

Motivated by this, we here introduce a new statistical method for association testing in recently admixed populations, where the effect sizes are allowed to depend on the ancestry of the allele. Our method does not rely on accurate inference of local ancestry, yet using simulations we show that in some scenarios it gives a dramatic increase in statistical power to detect associations. In addition, the method allows for testing for difference in effect size between ancestral populations, which can be used to determine if a SNP is causal. We demonstrate the usefulness of the method on data from the Greenlandic population.

Introduction

An individual's risk of developing common complex diseases, such as type 2 diabetes, is believed to be influenced by genetic variants and identifying such variants using genome–wide association mapping studies (GWAS) has been a rapidly growing research field the last decade (Klein et al. 2005, Duerr et al. 2006, Burton et al. 2007, Unoki et al. 2008, Thorleifsson et al. 2009, Sparso et al. 2009, Holm et al. 2011). So far, most GWAS have been performed in large populations, like the European and this has led to important new findings (refs?). However, recently a few GWAS have also been performed in historically small and isolated populations. The idea behind this approach is that in such populations substantial genetic drift over many generations has increased the probability that disease-causing variants have overcome their selective disadvantage and now occur at higher frequencies, making them easier to discover in these populations. Additionally, historically small and isolated populations have extended linkage disequilibrium (LD) compared to large populations, which means that more variants can be tested indirectly using the same amount of SNPs. Hence performing GWAS in historically small and isolated populations constitutes a powerful approach to discovering novel disease-causing variants (Zeggini 2014), which compliments GWAS in large populations well. This was recently shown very clearly in a study by Moltke et al. (2014) where a GWAS performed in the historically small and isolated population of Greenland led to the identification of a variant that explains more than 10% of all cases of type 2 diabetes in Greenland, but that had not been identified in previous much larger studies of large populations like the European and East Asian populations, because it is rare in these. However, while being a powerful approach both in large and in historically small and isolated populations, performing GWAS often involves an important challenge: many populations, like the Greenlandic, have experienced substantial amounts of recent admixture, which can bias the statistical test in the association mapping and lead to false discoveries.

Statistical methods for association mapping that solves this challenge exists (Devlin & Roeder 1999, Price et al. 2006, Zhou & Stephens 2012), but these methods all share one essential limitation: they are based on the assumption that the tested genetic variant has the same effect regardless of which ancestral population it is inherited from. This assumption is reasonable for a disease-causing variant. However, in GWAS the disease-causing variant is often not tested directly. Instead a small fraction of common single nucleotide polymorphisms (SNPs) are genotyped and tested and the aim is to identify the subset of these SNPs, if any, that are indirectly associated with the disease, because they are located close to the causal SNP and therefore in LD with it (see Figure 1). The effect sizes and strength of associations of the variants tested in a GWAS will therefore depend on the allele frequencies of the causal and tested variants and the strength of the LD between the tested variants and the causal variant. And importantly, since allele frequencies and LD patterns will often be different between different populations, this means that the effect size and the strength of association of a variant that is tested in a GWAS performed in an admixed population may depend strongly on the ancestry of the chromosomal segment which the genetic variant is located on. The extreme case shown in Figure 2 provides a

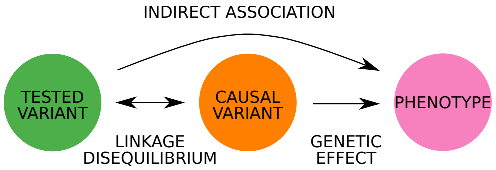

Figure 1: Indirect association between tested genetic variant and phenotype.

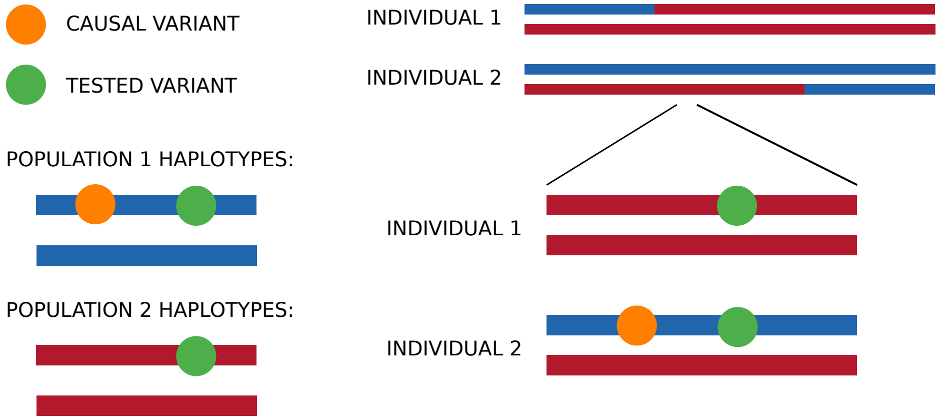

Figure 2: Extreme case of ancestry specific effects: The causal variant exists only in population 1 and is here in complete LD with the tested variant, which exists in both populations. The figure shows the homologous chromosomes of two admixed individuals with chromosomal segments colored according to which population the segments have been inherited from (population 1 is blue and population 2 is red). Both individuals carry one copy of the tested variant, however they have inherited them from different populations and only individual 2, who inherited the tested variant from population 1, carries the causal variant.

simple illustrative example. Here the causal variant is only present in ancestral population 1 and the LD between the tested and causal variant is complete in this population. The tested variant is present in both populations. In this example the tested variant is clearly stronger associated with the disease when inherited from ancestral population 1 than if inherited from population 2 and thus the effect of the tested variant will depend strongly on which ancestral population it has been inherited from. Hence this example illustrates that the assumption of ancestry-independent effects, which most methods for association testing in admixed populations are based on, does not always hold in the context of GWAS in admixed populations. The example also illustrates another important point: it is clear that the disease association with the tested variant in the example is much weaker than the disease association with the causal variant, which in this case equals the disease association with tested variant inherited from population 1. This means that a GWAS can potentially gain power by allowing ancestry-specific effect sizes.

Motivated by this, we here propose a statistical method for performing association mapping in admixed populations, named asaMap, that allows estimation and significance testing of ancestry specific effect sizes. In individuals from admixed populations the local ancestry of an allele (corresponding to the red/blue color in Figure 2) is not directly observable, but can sometimes be inferred (Sankararaman et al. 2008, Price et al. 2009, Maples et al. 2013, Guan 2014). However, this proposed method for ancestry specific association mapping does not rely on inferred local allelic ancestry because such inference can be prone to errors. Instead asaMap is based on a mixture model, where the mixture components are the phenotype distributions corresponding to given ancestries of the tested SNP and the mixture proportions are the probabilities of these ancestries (for more details see Materials and Methods). This mixture model allows us to take the uncertainty of the ancestry of the individual alleles into account by allowing for all possible ancestries and weighting each possible ancestry according to its probability of being the true ancestry; the mixture proportions. The mixture proportions for a given SNP are in asaMap by default calculated from global admixture proportions, population specific allele frequencies and genotypes. However, asaMap also allows the users to provide the mixture proportions and thus allows them to use more complex models such as hidden Markov models (Patterson et al. 2004, Price et al. 2009, Guan 2014) for obtaining these proportions. The mixture components in asaMap are based on a general linear model framework. This has at least three advantages. First, it means we can correct for population structure by simply including principal components as covariates. Second, it makes asaMap very flexible, since it means that it – like a general linear model – can be used to perform tests in a wide range of settings: asaMap can be applied to several different trait types (quantitative traits and case-control information) as well as several different genetic effect types (additive, dominant and recessive effects). Third, it allows easy correction for any additional covariates such as sex or age. Note that asaMap is an association testing method and not a method for performing admixture mapping where correlation between phenotype and inferred local ancestry is used to identify candidate regions (Patterson et al. 2004). asaMap is more similar to the methods of Pasaniuc et al. (2011) and Yorgov et al. (2014), where ancestry specific effects are estimated. But unlike these methods, asaMap does not require prior knowledge of the ancestry of each allele and furthermore enable correction for population structure.

In the following section we will describe the model behind asaMap in detail. Then using simulated data we will show that asaMap in some cases provides a substantial increase in power for association testing and that asaMap provides a framework that is even more flexible than the general linear model, which is often used for association testing in GWAS. For example, asaMap makes it possible to test whether a variant has ancestry-specific effects that differs between populations. It is reasonable to assume that a disease-causing allele have the same effect regardless of its ancestry. Therefore, this test can potentially be used to reveal if a SNP identified in a GWAS is causal. Finally, using data from a Greenlandic GWAS study (Moltke et al. 2014), we will show that asaMap can provide increased power for SNPs that are in strong LD with the causal SNP and that asaMap can be used to discriminate between causal and non-causal variants, not only in simulated data but also in real data.

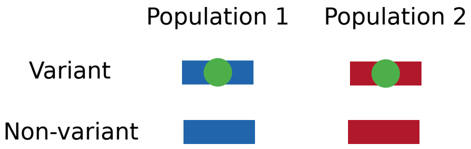

Figure 3: Ancestry specific allele types. When the population consist of two ancestral admixing populations, there are four possible ancestry specific allele types for the tested variant.

Materials and methods

Model

Our model framework is based on generalized linear regression models and thus applies to both quantitative traits and case-control studies (or dichotomous traits) and allows additive as well as dominant and recessive genetic effects. Here we describe the quantitative trait model for additive genetic effects, while detailed descriptions of case-control data as well as recessive genetic models can be found in the appendix.

As argued in the introduction, the strength of an association between genotype and phenotype is likely to depend on which of the ancestral populations the genetic variant has been inherited from. Thus, instead of estimating a single genetic effect of the variant, we here allow for population-specific genetic effect sizes, each of which we denote β_k for variants inherited from ancestral population k. Below we describe the model that allows us to do this.

Mixture model

We assume that we are analyzing data for N individuals from an admixed population that is a mixture of K ancestral populations. An individual from such a population will in any given diallelic autosomal locus have inherited each of its two allele copies from one of the K ancestral populations, and each of its two allele copies will either carry the minor or the major variant, which means each allele copy can be of $2 * K$ ancestry-specific allele types. The 4 types for $K = 2$ are shown in Figure 3. As a consequence an individual's two allele copies combined can be of $(2 * K)^2$ different ancestry-specific allele type combinations. We will here refer to these ancestry-specific allele type combinations as locus states, s. The 10 distinguishable locus states for $K = 2$ are shown in Figure 4.

The state of a locus is unfortunately not directly observable. It can sometimes be inferred, but local ancestry inference from genotype data is associated with uncertainty and ignoring this may lead to false positives. We have therefore chosen not to base asaMap on inferred states, but to instead use a model that allows us to take the uncertainty into account. This model is based on the observation that when all that can be observed are the genotypes i.e. the total number of variant copies present at the tested locus in each individual, the likelihood function for observing the phenotypes, $Y = (y_1, y_2, ..., y_N)$, takes

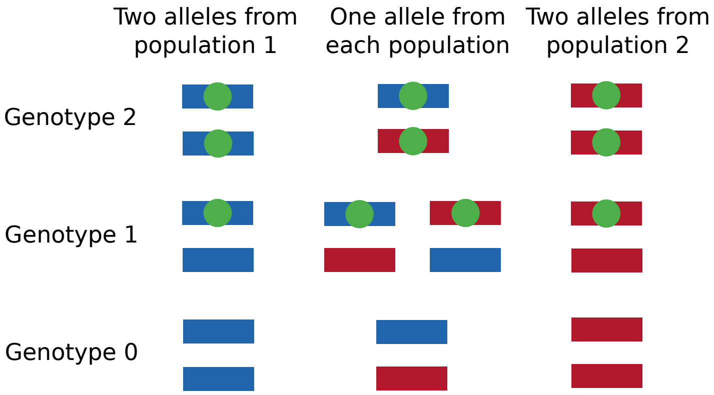
Two alleles from population 1
One allele from each population
Two alleles from population 2
Genotype 2
Genotype 1
Genotype 0

Figure 4: Locus states (ancestry specific allele type combinations) for a diallelic autosomal locus.

the form

$$p(Y|G,Z) = \prod_i p(y_i|G,Z) = \prod_i \sum_s p(y_i|s,G,Z)p(s|G,Z)$$

where G is a vector of all observed genotypes at a specific locus and Z are appropriately chosen covariates and where the product runs over all individuals $i = 1 \dots N$ and the sum runs over all possible locus states s. Assuming that the trait is conditionally independent on the observed genotypes G given the latent variable s and the covariates Z, this likelihood also takes the form

$$p(Y|G,Z) = \prod_i \sum_s p(y_i|s,Z)p(s|G). \tag{1}$$

This means we can model the probability of the observed phenotypes Y as a mixture of phenotype distributions, where each mixture component is the phenotype distribution $p(y|s,Z)$ that corresponds to a given locus state, s, and the mixture proportions are the probability $p(s|G)$ of that state given the observed genotypes. Importantly, this modeling approach makes it very easy to take the uncertainty of the unobserved ancestry into account, since this uncertainty is explicitly included in the model in the form of the mixture proportions. Furthermore, the above likelihood function is a function of our parameters of interest, namely the population-specific effects β_k, via the mixture components $p(y|s,Z)$. This means that we can use the model both for estimating the effects β_k and for performing association testing, which is what is done in asaMap. More specifically, based on the above likelihood function the β_ks are estimated using maximum likelihood estimation and testing for association is performed using likelihood ratio tests.

Below is a detailed description of how we model the mixture components and the mixture proportions in the likelihood function.

Mixture components

For a quantitative trait, Y, the mixture component, i.e. the phenotype distribution $p(y|s, Z)$, is based on a linear regression model. We assume that given the locus state, s, the phenotype y_i for a single individual i follows a normal distribution with mean given by the linear predictor

$$\eta_i = \alpha + \sum_k \beta_k x_k(s) + \sum_c \gamma_c z_c^i, \tag{2}$$

where α parameterizes intercept (baseline) and where additional covariates including principal components for correcting for potential confounding by population structure (Price et al. 2006) enters the model in the z_c^i with effects γ_c. Finally, $x_k(s)$ is the number of risk alleles from population k for a locus in state s assuming an additive model. The definition of x_k is different for the recessive model and is described in more details in the appendix.

Mixture proportions

The simplest approach to calculating the mixture proportions, i.e. the probabilities of all different possible locus states for individual i given i's genotype g, is to use i's global admixture proportions Q^i and the population specific allele frequencies f, which can both be inferred using standard software tools such as ADMIXTURE (Alexander et al. 2009). In the general case of K admixing populations the global admixture proportions are $Q^i = (q_1^i, q_2^i, \ldots, q_K^i)$, where q_k^i is the fraction of is genome that has been inherited from population k and the population specific allele frequencies are $f = (f_1, f_2, \ldots, f_K)$, where f_k is the frequency of the tested variant in population k.

We can use Q^i and f to calculate the probability of the locus state s given genotype g in three steps after introducing the notation $s = (a, t)$, where $a = (a_1, a_2)$ is the allelic ancestry and $t = (t_1, t_2)$ is the allelic genotype (with $t_1 + t_2 = g$). In the first step we consider the conditional distribution of ordered allelic genotype $t = (t_1, t_2)$ given genotype $g = t_1 + t_2$ which takes the form:

$$p(t|g) = \frac{1_{t_1+t_2=g}}{\sum_{t_1',t_2'} 1_{t_1'+t_2'=g}}.$$

The second step concerns the probability of allelic ancestry $a = (a_1, a_2)$ given allelic genotype $t = (t_1, t_2)$. We use the global admixture proportions Q^i to give the probability $p(a|P, Q^i) = q_{a_1}^i q_{a_2}^i$ of ancestry a assuming independent ancestry of alleles and we use the corresponding population specific allele frequencies f to calculate the probability of allelic genotype given allelic ancestry $p(t_j|a_j, f, Q^i) = f_{a_j}^{t_j}(1 - f_{a_j})^{1-t_j}$. Then the desired probability of allelic ancestry a given allelic genotype t can be calculated by

$$\begin{aligned}
p(a|t,f,Q^i) &= \frac{p(a,t|f,Q^i)}{p(t|f,Q^i)} = \frac{p(t|a,f,Q^i)p(a|f,Q^i)}{p(t|f,Q^i)} \\
&= \frac{q_{a_1}q_{a_2}p(t|a,f,Q^i)}{\sum_{a'\in\{1,2\}^2} q_{a'_1}q_{a'_2}p(t|a',f,Q^i)} \\
&= \frac{q^i_{a_1}q^i_{a_2}f^{t_1}_{a_1}(1-f_{a_1})^{1-t_1}f^{t_2}_{a_2}(1-f_{a_2})^{1-t_2}}{\sum_{a'\in\{1,2\}^2} q^i_{a'_1}q^i_{a'_2}f^{t_1}_{a'_1}(1-f_{a'_1})^{1-t_1}f^{t_2}_{a'_2}(1-f_{a'_2})^{1-t_2}}.
\end{aligned}$$

The third step combines the results of the first two steps to calculate the conditional distribution of locus states given the observed genotype using Q^i and f, since this distribution can be written as a combination of the conditional probabilities calculated above:

$$p(s|g,f,Q) = \frac{p(a|t,f,Q)p(t|g)}{\sum_{s'} p(a'|t',f,Q)p(t'|g)} \tag{3}$$

Alternatively, this distribution across locus states can be supplied by the user and may for instance be calculated based on the output from local ancestry inference software.

Parameter estimation and hypothesis testing

The parameters of the model (and of the sub-models relevant for testing purposes) are estimated using maximum likelihood based on the likelihood function given in equation 1 (using the details provided in equation 2 and equation 3). Optimization of this likelihood function must be done numerically and we have developed an EM algorithm (see appendix for details) that provides faster convergence than standard all-purpose numerical optimization algorithms such as BFGS.

Standard generalized linear model based methods for association mapping makes use of statistical tests comparing two models: a model where the tested variant has a genetic effect versus a model where the variant has no effect. In asaMap we allow the effect sizes to be specific to ancestral populations and therefore several more nested models can be compared. For example, for an additive genetic effect in the case of two admixing populations five models M1-M5 are available for comparison. The full model M1 allows separate genetic effects for each of the two ancestral population: β_1 and β_2. The sub-model M2 assumes no effect in population 1. The sub-model M3 assumes no effect in population 2. The model M4 assumes that the effect sizes are the same in both ancestral populations, and finally the null model M5 which assumes that the variant has no effect in any population. An overview of these additive models is given in Table 1. For recessive genetic effects the standard generalized linear model based methods for association mapping tests a model where carrying two copies of the variant allele has an effect on the individuals phenotype versus a model where the variant has no effect. In this context asaMap allows the effect size to be specific to the ancestry combination of the two variant alleles and seven (sub-)models R1-R7 described in the Appendix and Table S1 are implemented.

Model	Hypothesis	The model assumes
M1	$(\beta_1, \beta_2) \in R^2$	population specific effects
M2	$\beta_1 = 0, \beta_2 \in R$	no effect in population 1
M3	$\beta_1 \in R, \beta_2 = 0$	no effect in population 2
M4	$\beta_1 = \beta_2 \in R$	same effect in both populations
M5	$\beta_1 = \beta_2 = 0$	no effect in any population

Table 1: Description of the different possible additive models for two ancestral populations. Comparing two nested models will lead the tests described in table 2. For a description of recessive model see table S1.

Models	Tests if there is
M1 vs. M5	an effect in any population
M1 vs. M2	an effect in population 1
M1 vs. M3	an effect in population 2
M1 vs. M4	a different effect in the two populations
M2 vs. M5	an effect in population 2 assuming no effect in population 1
M3 vs. M5	an effect in population 1 assuming no effect in population 2
M4 vs. M5	an effect assuming it is the same in both populations

Table 2: Possible tests assuming additive models as described in table 1 for two ancestral populations.

Hypotheses regarding the ancestry specific effect sizes are carried out using likelihood ratio tests comparing nested models. The implemented models under the additive assumption allows us to test if there is an effect in any population (M1 vs. M5), an effect in population 1 (M3 vs. M5), an effect in population 2 (M2 vs. M5), and a difference in the effect specific to the two ancestral populations (M1 vs. M4). And last but not least we can test if there is an effect assuming that it is the same in the two populations, i.e. M4 vs. M5. Note that this latter test is equivalent of the standard test for association performed using a generalized linear model and has been implemented to enable comparison of the other tests in asaMap to the standard generalized linear model approach. In addition the tests M1 vs. M2 and M1 vs. M3 are also implemented and may be used for model check of the tests based on models M2 and M3. An overview of the implemented tests for additive genetic effects is given in Table 2. The corresponding tests comparing nested recessive models are described in the Appendix and Table S2.

The estimation and testing procedures described above has been implemented in the software asaMap available at `http://popgen.dk/software`, making the method applicable to large scale genome wide association studies.

Correcting for population structure

To correct for population structure in the real data (described below), we include as covariates the first 10 principal components calculated from a genotype-based covariance matrix (Price et al. 2006). We are aware that a more powerful approach would be a mixed effects approach similar to Kang et al. (2008) or Zhou & Stephens (2012), but we have not succeeded in implementing this in a computationally tractable way due to the sum across locus states.

Simulated data

We carried out analysis based on simulated samples with genetic ancestry from two admixing populations. We simulated data from a total of nine scenarios. In each of these scenarios we simulated data from a SNP locus, which is assumed not to be causal, but to be in LD with a causal variant. We assumed that the variant has an effect in one or both of the ancestral populations. For all nine scenarios we simulated data for a total of 2500 individuals with admixture proportions, Q, from population 1 in the set $\{0, 0.25, 0.5, 0.75, 1\}$ (500 individuals for each value). What varies between the scenarios is the frequency of the tested variant in the two populations, $f = (f_1, f_2)$, the effect sizes in the two populations (β_1 and β_2), the type of trait (quantitative or case-control) and the underlying genetic effect model (additive or recessive). For a description of the nine scenarios see Table 3.

For all scenarios, we followed the same simulation procedure: based on the individual admixture proportions Q we sampled the ancestry $a = (a_1, a_2)$ for each allele copy for all individuals. Then based on this ancestry a and the allele frequencies f in the ancestral populations, we sampled the allele types $t = (t_1, t_2)$ of each allele. Knowing a and t the number of risk allele copies inherited from each ancestral population is known and based on this the phenotype is simulated using the relevant phenotype distribution (quantitative or case-control), the relevant genetic model (additive or recessive) and scenario specific effect sizes (β_1 and β_2). For quantitative traits we generated the phenotype value using a normal distribution with variance 1 and for case-control studies we generated the disease status using a binomial distribution.

The simulations thus give us access to association data where the true ancestry specific effects are known and therefore allows us to assess the consistency and unbiasedness of the estimators in asaMap. The simulations also allows us to assess the power of the tests for assocation implemented in asaMap. Finally, because we explicitly simulate the ancestry specific allele type combinations (locus states), the simulations allow us to compare the power of the tests in asaMap to the hypothetical power of a test where the true locus states, which in reality are unobservable, are known.

Data for *TBC1D4* in a Greenlandic cohort

To demonstrate that allowing effect sizes to be specific to ancestral populations can be informative and appropriate for real data we apply the methods implemented in asaMap

Scenario	Frequencies	Effects	Trait	Model
A1	$(0.1, 0.3)$	Population 1	Quantitative	Additive
A2	$(0.1, 0.3)$	Both	Quantitative	Additive
A3	$(0.1, 0.3)$	Population 2	Quantitative	Additive
B1	$(0.2, 0.2)$	Population 1	Quantitative	Additive
B2	$(0.2, 0.2)$	Both	Quantitative	Additive
B3	$(0.2, 0.2)$	Population 1	Quantitative	Recessive
C1	$(0.2, 0.2)$	Population 1	Case-Control	Additive
C2	$(0.2, 0.2)$	Both	Case-Control	Additive
C3	$(0.4, 0.4)$	Population 1	Case-Control	Recessive

Table 3: Simulated scenarios. All have 2500 individuals and individual admixture proportions from population 1 in $\{0, 0.25, 0.5, 0.75, 1\}$. For each scenario we vary the effect size and when there is an effect in both population we assume that they are the same

to genotype data in combination with measurements of 2 hour plasma glucose levels of 2575 individuals in the Inuit Health in Transition cohort (Moltke et al. 2014, Jorgensen et al. 2013). More specifically we applied the methods in asaMap to genotyped SNPs in the *TBC1D4* gene (rs61736969, rs7330796, rs1062087, rs2297206 and rs77685055). In Moltke et al. (2014) all five SNPs were found to be strongly associated with an increase in 2 hour plasma glucose levels. rs7330796 was the lead SNP in the discovery part of the study, which was based on SNP chip data and rs61736969 is the causal SNP and was identified from sequencing data. The three remaining SNPs were also identified from sequencing data in the search for the causal variant.

Results

To investigate the cost and benefits of asaMap compared to a standard generalized linear model we first applied both methods to simulated data to compare their statistical power and to assess important statistic properties of asaMap. Then we applied both methods to real data to compare the range of their potential usage. In all cases we investigated populations that are mixtures between two populations, however we note that asaMap can be applied to populations that are mixtures of any number of populations, see Materials and Methods.

Standard generalized linear model based methods for association mapping makes use of statistical tests comparing two models: a model where the tested variant has a genetic effect versus a model where the variant has no effect. In asaMap where we allow the effect sizes to be specific to ancestral populations, several more models can be compared (for an overview see Table 1 and Table 2), a detailed description is provided in Materials and Methods. In the context of an additive genetic effect, the full model (M1) allows separate genetic effects for each ancestral population, in the case of two ancestral populations: β_1 and β_2. The sub-models then assumes no effect in population 1 (M2), no effect in population 2 (M3), the same effect in both populations (M4), and no effect in any population (M5). This allows us to test if there is an effect in any population (M1 vs. M5), an effect in population 1 (M3 vs. M5), an effect in population 2 (M2 vs. M5), and a different effect in the two populations (M1 vs. M4). And last but not least it allows us to test if there is an effect assuming that it is the same in the two populations, i.e. M4 vs. M5. In the context of a recessive genetic effect, corresponding hypothesis can be tested by comparing the models R1-R7, further details are given in the Appendix and Table S1-S2.

Note that test comparing M4 and M5 (R6 vs. R7) is equivalent of the standard test for association performed using a generalized linear model. In the following we will therefore perform the comparison of asaMap and the standard generalized linear model by comparing the M4 vs. M5 (R6 vs. R7) test with the remaining tests in asaMap. In the following we will therefore perform the comparison of asaMap and the standard generalized linear model by comparing the M4 vs. M5 (R6 vs. R7) test with the remaining tests in asaMap.

Simulation-based results

To assess asaMap we first simulated data for individuals with genetic ancestry from two admixing populations according to nine scenarios; six with quantitative traits and three with case-control traits (Table 3, see Materials and Methods for details). We used this data to assess how powerful the different tests in asaMap are in different settings. Since asaMap's test of M4 vs. M5 is equivalent of the standard test for association performed using a generalized linear model, this power assessment includes a power comparison between asaMap and a generalized linear model. We also used the simulated data to assess other important statistical properties of asaMap, including whether it provides unbiased and consistent estimates of the population specific estimates. Below is a description of all the simulation-based results.

Power assessment for quantitative traits

First we simulated a scenario (scenario A1) with a causal variant that is only present in one of the ancestral populations, but with the tested variant present in both ancestral populations, causing the tested variant to have population-specific effects. More specifically, the tested variant was simulated to have a frequency 10% in ancestral population 1 and 30% in ancestral population 2. Furthermore, the tested variant was simulated to have an additive effect, with an effect size in population 1, β_1, that varied in the range $[0, 1.5]$, and with no effect in population 2 since the causal variant is not present in this population. When applying asaMap to data from this scenario, the standard test (M4 vs. M5), where the effect size is assumed to be the same in both populations, required much larger effect sizes for full statistical power than the full test where the effect size is not assumed to be the same (M1 vs. M5), see Figure 5:A1. Further, the test if there is an effect in population 1 (M1 vs. M2 and M3 vs. M5) was slightly more powerful than the full test (M1 vs. M5), which is anticipated since it has only 1 degree of freedom and the full test has 2 degrees of freedom. Finally, asaMap could address the interesting question if there is a difference

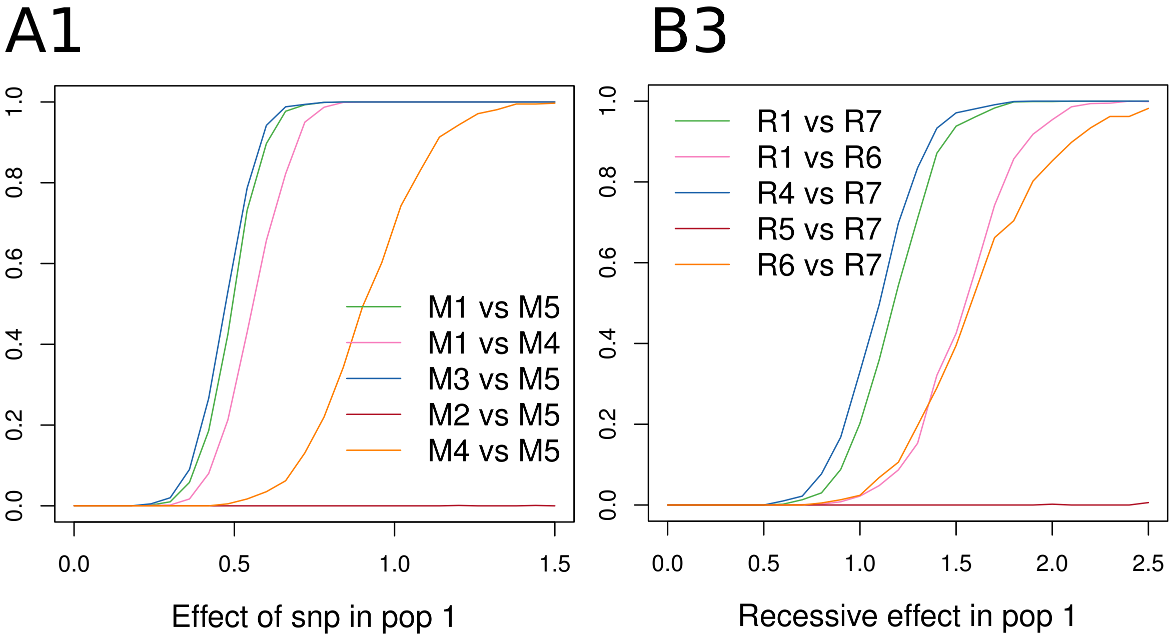

Figure 5: Results of power simulations for the additive scenario A1 and recessive scenario B3. For scenario A1, the allele frequencies in the admixing populations are 0.1 and 0.3, the genetic effect is simulated to be additive on a quantitative trait and only present when the variant is inherited from population 1. For scenario B3, the allele frequencies in the admixing populations are both 0.2, the effect is simulated to be recessive on a quantitative trait and only present when the variant is inherited from population 1. The curves show the fraction of simulated p-values that are smaller than 10^{-8}, based on 1000 simulations for each effect size.

in the effect sizes between ancestral populations (M1 vs. M4) with good statistical power, even for variants with effect sizes lower that those detectable using the standard test.

Second, we simulated a scenario (scenario A2), which only differs from scenario A1 in one way: in scenario A2 the tested variant has the same effect in both populations. For this scenario the test of M1 vs. M5 was less powerful than the standard test (M4 vs. M5), which was anticipated because the modeling underlying the former test is more complicated (scenario A2, Figure S1). However the difference in power is very small. The test, M3 vs. M5, if there is an effect in population 1, where the tested variant has the lowest frequency on the other hand, is markedly less powerful.

Third, we simulated one more scenario (scenario A3) that only differs from scenario 1A in one way, this time by switching which of the two ancestral populations the tested variant had an effect in, so now the tested variant was simulated to only have an effect in population 2, where the allele occurs with the highest frequency. In this scenario, the statistical power of the test, M1 vs. M4, a test if there is a difference in effect sizes, was the same as for scenario A1, but, unlike in scenario A1 all other relevant tests (M1 vs. M5, M2 vs. M5, M4 vs. M5) had almost identical statistical power (scenario A3, Figure S1). Hence in this scenario the standard test was just as powerful as the remaining tests.

Next, we simulated yet another scenario (scenario B1) like scenario A1. This time the only difference was that the tested variant was simulated to have a frequency of 20% in both ancestral populations. As was the case for scenario A1 the test with the best statistical power was the test if there is an effect in population 1 (M3 vs. M5) (scenario B1, Figure S1). However, this test was only slightly more powerful that the full test for effects (M1 vs. M5), whereas both these test provided remarkable improvements in power compared to the standard test of effect assuming same effect in both populations (M4 vs. M5).

FiXme Note: B1: M1 vs M4?

We also simulated a scenario (scenario B2) identical to scenario B1 with the exception that the tested variant was simulated to have the same effect in both ancestral populations. Again, as was the case for scenario A2, the anticipated loss of power of M1 vs. M5 due to the more complicated modeling compared to M4 vs. M5 was very small (scenario B2, Figure S1), while the single population tests (M1 vs. M2 and M1 vs. M3) were less powerful.

Finally, to compare the power of the different tests for a variant with a recessive effect, we simulated a scenario (scenario B3), where the frequency of the tested variant was 20% in both ancestral populations and it had a recessive effect in population 1 and no effect in population 2. The results for this scenario were similar to the results for scenario A1 and B1, with the test of an effect in population 1 (R4 vs. R7) being slightly more powerful than the full test (R1 vs. R7) and both represent remarkable improvements compared to the standard test (R6 vs. R7), see Figure 5.

Power assessment for case-control study data

For case-control study data we similarly simulated from a population of mixed ancestry where the tested variant has a frequency of 20% in both populations. The effect of the allele is simulated as log-additive in the logistic model and either only present in ancestral population 1 (scenario C1, Figure S1) or present in both ancestral populations (scenario C2, Figure S1). The results are similar to the results for the quantitative trait versions of these scenarios (scenarios B1 and B2). In scenario C1 the asaMap test for an effect in ancestral population 1 (M3 vs M5) is the most powerful test, slightly more powerful than the asaMap test for an effect in any population (M1 vs. M5), and both these tests outperform the standard logistic regression test for association. In scenario C2, where the effect is present in both ancestral populations, the standard test is as expected the most powerful, but only slightly better than the asaMap test for an effect in any population (M1 vs. M5).

We also simulated a similar case-control scenario (scenario C3), where the effect of the tested variant is recessive and present only in ancestral population 1 (scenario C3, Figure S1). Note that to reach any statistical power for the simulated odds, we allowed the frequency of the tested allele to be 40% in both ancestral populations. In this scenario, the standard test if there is an effect assuming it is the same in both populations (R6 vs. R7), does not reach satisfactory statistical power for any realistic odds. The tests that

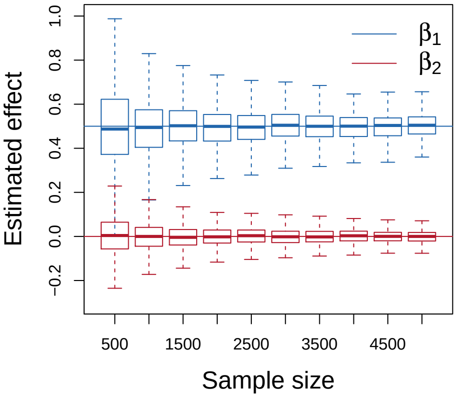

Figure 6: Results of consistency simulations for scenario A1 with a fixed population specific effect of size 0.5 in population 1 and 0 in population 2 and with increasing sample size in the range from 500 to 5000. Each box corresponds to 1000 simulations.

allow for population-specific effects (R1 vs. R7 and R4 vs. R7) on the other hand perform much better, although they also require quite high odds to reach full statistical power.

Bias, consistency and false positive rate

Besides using the simulated data for power comparisons we also used it for assessing asaMap's estimators for population specific effects. We did this for all nine simulated scenarios (A1-3,B1-3 and C1-3). This showed that asaMap's estimators are unbiased for these scenarios(Figure S2). For all simulation setups, we also simulated data under the null, i.e. without any effect in any of the populations, and applied all tests available in asaMap to the data to assess if asaMap has a controlled false positive rate. This was done to ensure that the uncertainty in ancestry does not lead to inflated test statistics. More specifically we did this for the shared null model of scenarios A1-A3, the shared null of scenarios B1-B2, the null of scenario B3, the shared null of scenario C1-C2 and the null of scenario C3, which means that we performed the assessment both in the context of quantitative traits and of case-control data and both for variants with additive and variants with recessive effects, corresponding to the null models for all nine simulated scenarios. The corresponding QQ-plots of the p-values achieved show that the false positive rate is indeed controlled for in all the tests (Figure S3).

To assess consistency of the estimators, we next re-simulated scenario A1 with increasing sample sizes and a fixed population specific effect size of 0.5 in population 1 and 0 in population 2. This showed that asaMap's estimators are consistent (Figure 6 and that the decrease in variance with increasing sample size is consistent with the expected $1/n$ relation. Finally, in the process of simulating ancestry specific association data, we explicitly simulated the ancestry specific allele type combinations (locus states), which are not di-

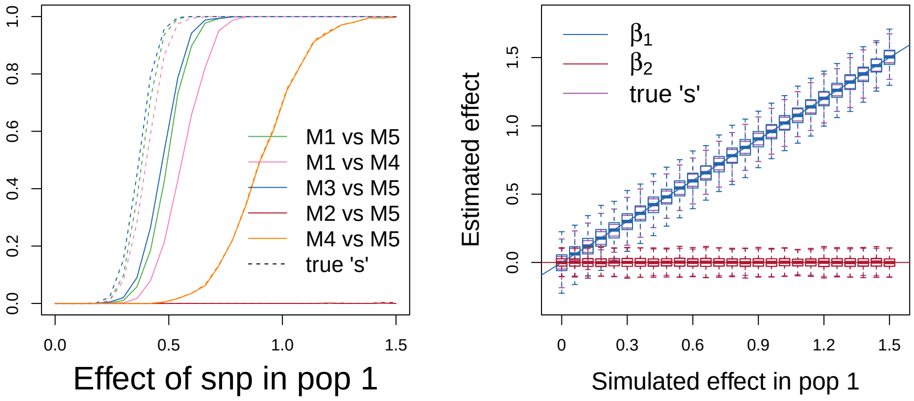

Figure 7: Results for scenario A1. Left: Results of power simulations for scenario A1. The power of the tests based on true (simulated) locus states are shown with dashed curves. Right: Results of bias simulations for scenario A1. Boxplots of estimates based on true (simulated) locus states are shown in purple.

rectly obervable in real data. This allowed us to compare the tests implemented in asaMap with hypothetical tests of equivalent models based on known locus states. As expected, the tests based on correctly known locus states is shown to be slightly more powerful and the variance of the estimators is a bit smaller (Figure 7, results only shown for scenario A1), which shows that if the ancestry of each allele copy was known without error there is potential for an even larger increase in statistical power.

TBC1D4 gene in a Greenlandic cohort

To further assess asaMap, we also applied it to real data from the Greenlandic population. This populations is a historically small and isolated Inuit population and recent investigations of the its genetic history have shown that it is highly admixed: more than 80% of Greenlanders have some recent European ancestry and the Greenlanders have on average approximately 25% European ancestry (Moltke et al. 2015). A recent GWAS in the Greenlandic population (Moltke et al. 2014) led to the identification of a variant in the gene TDB1D4 which confers type 2 diabetes. The lead SNP in the discovery part of this study was rs7330796 and to locate the causal variation four coding SNPs in high LD was identified using exome sequencing and subsequently genotyped. Among these SNPs, rs61736969 located in *TBC1D4* was identified as the causative variant and shown to have a recessive effect.

Based on genotype and phenotype data from the Greenlandic Inuit Health in Transition cohort described in Moltke et al. (2014), we tested the five above mentioned SNPs for ancestry specific association with 2 hour plasma glucose levels as a quantitative trait using a recessive model (see appendix). For the four non-causal SNPs (rs7330796 (original

	rs61736969[a]	rs7330796[b]	rs1062087	rs2297206	rs77685055
R4 vs. R7	1.944e-36	5.779e-25	1.648e-20	2.088e-20	4.417e-17
R6 vs. R7	1.943e-36	4.326e-22	1.004e-17	9.499e-18	8.556e-16
R2 vs. R6	1	8.673e-05	1.825e-05	3.499e-05	3.703e-03

Table 4: P-values 2-h plasma glucose for SNPs in the *TBC1D4* gene under a recessive model.[a] The causal SNP identified from sequencing data. [b] the lead SNP in the discovery part of the study based on SNP chip data.

lead SNP), rs1062087, rs2297206, rs77685055) we saw that the ancestry specific test for a recessive effect of the variant when both alleles are inherited from the Inuit population (R4 vs. R7) is more significant than the standard test for a recessive effect (R6 vs. R7), supporting our simulation based observation that asaMap can increase the power to detect associations when the causal SNPs remains untyped. Furthermore, for all four non-causal SNPs asaMap (R4 vs. R7) showed that the effect of carrying two risk variant alleles both inherited from the ancestral Inuit population is significantly different from the effect of carrying two risk variant alleles of which at least one is inherited from the ancestral European population. This suggests that these four SNPs are not causal. One the contrary, for rs61736969 the p-value for the population specific test for an recessive effect in the Greenlandic population is identical to the p-value of the standard test (R6 vs R7), which suggests that this SNP is causal. These results are all in line with the conclusions about causality drawn in Moltke et al. (2014). Finally, we note that the QQ plot in Figure 8 show that the ancestry specific association test is not more inflated than regular association mapping, using the first ten principal components to correct for population structure as described in Materials and Methods.

Discussion

In this paper, we have presented asaMap, a flexible new statistical test framework for association mapping in admixed populations, which allows for the possibility that a tested variant can have different effects in the different ancestral populations. asaMap does this by modeling the local allelic ancestry as a latent variable.

Using simulated data we have demonstrated that asaMap provides ancestry-specific effect estimates that are unbiased and consistent. Furthermore, we have assessed how powerful asaMap's association tests are compared to the standard tests, which are most commonly used for performing association mapping in admixed populations. Unlike asaMap, these commonly used tests do not allow for the possibility that a tested variant can have different effects in the different ancestral populations. On the contrary, they are based on the assumption that the effect of the tested variant is the same regardless of its ancestry. This assumption is reasonable for a causal variant, but may not hold for the SNPs tested in a GWAS which are usually not causal. Importantly, we have here demonstrated that when the effect does depend on ancestry, the full test in asaMap, which tests if there is

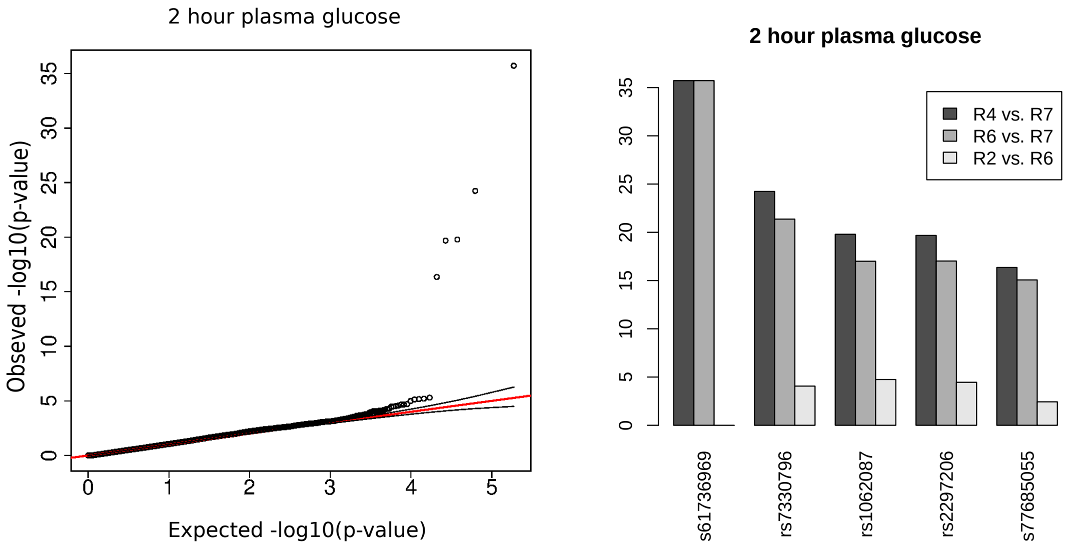

Figure 8: Association results for data from Greenland. Left: QQ plot for on 2-hour plasma glucose in the Greenlandic IHIT cohort achieved using the asaMap recessive model for quantitative trait, while testing for a population specific effect in the ancestral Inuit population using the test R4 vs. R7. Right: Minus $\log_{10}$ of the p-values shown in Table 4.

an effect in any population while allowing for ancestry-specific effects (M1 vs. M5) will outperform the commonly used tests. However, the gain in power depends strongly on the frequency of the tested variant in the ancestral populations. If the frequency in the population where the effect is highest is lower than in the other population the gain in power can be very substantial. Conversely if the allele frequency is higher, then gain in power can be negligible. When the effect is not ancestry-specific, the full test in asaMap is less powerful than the commonly used test, which could be expected since the full test is based on a more complicated model. However, the difference in power is small.

In addition to the full test, asaMap also allows for testing if a variant has an effect in a specific population - both with or without assuming that there is an effect in the other population. Testing if a variant has an effect in one of the population, while allowing for an effect in the other population (M1 vs. M2 or M1 vs. M3) will be less powerful for identifying new associations than testing if a variant has an effect in one of the populations, while assuming no effect in the other population (M2 vs. M5 and M3 vs. M5). However, they are useful tests for establishing if the causal variant is present and in LD with the tested variant in a specific population. Also, if one of the populations has already been extensively studied for a very large number of individuals, and no effect has been detected here, then it can be practical to assume that there is no effect in that population: our results show that the tests M2 vs. M5 and M3 vs. M5 are the most powerful if the effect is actually absent in one population. We therefore recommend using the tests M2 vs. M5 for association testing in datasets where very large-scale association tests have already been applied to one of the ancestral populations.

All the power results described above were based on simulations of variants with additive effects. For variants with recessive effects it is a bit more complicated since there are three possible effects for individuals carrying two risk alleles assuming there are two populations. However, we have here demonstrated that when a variant only has a recessive effect when both alleles are inherited from one of the two populations there is potential to gain a great amount of power by allowing for ancestry-specific effects. More specifically we observed a large gain in power when the tested variant was in high frequency in both populations, both for the full test (R1 vs. R7) and even more so for the test for an effect in a specific population (R4 vs R7). We expect the same to be true in all cases where the frequency of the tested allele is high in the population without any effect, because in this case a lot of the individuals carrying two copies of the risk variant will not be affected, causing the standard recessive association test to have low statistical power.

Another useful test is M1 vs. M4, which tests whether the effect sizes are different in the two populations. Since we expect the effect of a causal allele to be similar in the two populations a significant test is an indication that a variant is not causal. However, a non-significant test for different effect sizes can clearly not be taken as evidence that the variant is causal, since two fairly different populations can have different amounts of LD between the causal site and the tested variant, but this may not always be true.

Using genotype data for the Inuit Health in Transition cohort(Moltke et al. 2014, Jorgensen et al. 2013) for five SNPs and 2 hour plasma glucose levels for the same individuals we demonstrated that the population specific tests in asaMap can increase the statistical power of the GWAS when the causal variant remains untyped. Also, asaMap correctly provided results that were consistent with the causal SNP being causal. Furthermore, asaMap correctly provided results which support that the four remaining SNPs have ancestry-specific effects and thus are not causal. And asaMap did this despite the fact that all four SNPs showed strong evidence of association.

In summary, we have shown, using both simulated and real data, that asaMap by allowing for ancestry-specific effects provides tests that in some cases are much more powerful than the standard tests that are commonly used in GWAS. We have also shown the same tests, at least the full test, are almost as powerful as the standard test in all other cases. Finally, we have shown that asaMap can be used to test if a variant has an ancestry dependent effect, which can be helpful for assessing if a tested SNP is causal. This suggests that asaMap is a powerful and flexible complement to the standard tests commonly used when carrying out a GWAS in admixed populations.

As future work, we consider extending the model to account for the genotype uncertainty present when working with next generation sequencing data (Nielsen et al. 2011, Skotte et al. 2012) or imputation of genotypes (refs). In the case of next generation sequencing data, admixture proportions and population specific effect sizes can be estimated taking the genotype uncertainty into account using NGSadmix (Skotte et al. 2013). The uncertain genotypes from next generation sequencing as well as from imputation can easily be accounted for by expanding the state space of the latent variable in the mixture model (see Materials and Methods) to also include all possible genotypes while conditioning on the observed data.

Additionally, since some of the power simulations indicated that asaMap would be even more powerful if the true ancestry of the allele copies were known, another potential future direction we consider is to see if we can reduce the uncertainty of ancestry of the allele copies and thereby make asaMap even more powerful. A first step in this direction has already been taken by allowing the user to provide a probability distribution across locus states, however further work is needed to determine the optimal strategy to obtain these probabilities.

References

Alexander, D. H., Novembre, J. & Lange, K. (2009), 'Fast model-based estimation of ancestry in unrelated individuals', *Genome Res.* **19**(9), 1655–1664.

Balding, D. J. (2006), 'A tutorial on statistical methods for population association studies', *Nat. Rev. Genet.* **7**(10), 781–791.

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., Todd, J. A., Donnelly, P., Barrett, J. C., Burton, P. R., Davison, D., Donnelly, P., Easton, D., Evans, D., Leung, H. T., Marchini, J. L., Morris, A. P., Spencer, C. C., Tobin, M. D., Cardon, L. R., Clayton, D. G., Attwood, A. P., Boorman, J. P., Cant, B., Everson, U., Hussey, J. M., Jolley, J. D., Knight, A. S., Koch, K., Meech, E., Nutland, S., Prowse, C. V., Stevens, H. E., Taylor, N. C., Walters, G. R., Walker, N. M., Watkins, N. A., Winzer, T., Todd, J. A., Ouwehand, W. H., Jones, R. W., McArdle, W. L., Ring, S. M., Strachan, D. P., Pembrey, M., Breen, G., St Clair, D., Caesar, S., Gordon-Smith, K., Jones, L., Fraser, C., Green, E. K., Grozeva, D., Hamshere, M. L., Holmans, P. A., Jones, I. R., Kirov, G., Moskvina, V., Nikolov, I., O'Donovan, M. C., Owen, M. J., Craddock, N., Collier, D. A., Elkin, A., Farmer, A., Williamson, R., McGuffin, P., Young, A. H., Ferrier, I. N., Ball, S. G., Balmforth, A. J., Barrett, J. H., Bishop, D. T., Iles, M. M., Maqbool, A., Yuldasheva, N., Hall, A. S., Braund, P. S., Burton, P. R., Dixon, R. J., Mangino, M., Suzanne, S., Tobin, M. D., Thompson, J. R., Samani, N. J., Bredin, F., Tremelling, M., Parkes, M., Drummond, H., Lees, C. W., Nimmo, E. R., Satsangi, J., Fisher, S. A., Forbes, A., Lewis, C. M., Onnie, C. M., Prescott, N. J., Sanderson, J., Mathew, C. G., Barbour, J., Mohiuddin, M. K., Todhunter, C. E., Mansfield, J. C., Ahmad, T., Cummings, F. R., Jewell, D. P., Webster, J., Brown, M. J., Clayton, D. G., Lathrop, G. M., Connell, J., Dominczak, A., Samani, N. J., Marcano, C. A., Burke, B., Dobson, R., Gungadoo, J., Lee, K. L., Munroe, P. B., Newhouse, S. J., Onipinla, A., Wallace, C., Xue, M., Caulfield, M., Farrall, M., Barton, A., Bruce, I. N., Donovan, H., Eyre, S., Gilbert, P. D., Hider, S. L., Hinks, A. M., John, S. L., Potter, C., Silman, A. J., Symmmons, D. P., Thomson, W., Worthington, J., Clayton, D. G., Dunger, D. B., Nutland, S., Stevens, H. E., Walker, N. M., Widmer, B., Todd, J. A., Frayling, T. A., Freathy, R. M., Lango, H., Perry, J. R., Shields, B. M., Weedon, M. N., Hattersley, A. T., Hitman, G. A., Walker, M., Elliott, K. S., Groves, C. J., Lindgren, C. M., Rayner, N. W., Timpson, N. J., Zeggini, E., McCarthy, M. I., Newport, M., Sirugo, G., Lyons, E., Vannberg, F., Hill, A. V., Bradbury, L. A., Farrar, C., Pointon, J. J., Wordsworth, P., Brown, M. A., Franklyn, J. A., Heward, J. M., Simmonds, M. J., Gough, S. C., Seal, S., Stratton, M. R., Rahman, N., Ban, M., Goris, A., Sawcer, S. J., Compston, A., Conway, D., Jallow, M., Newport, M., Sirugo, G., Rockett, K. A., Kwiatowski, D. P., Bumpstead, S. J., Chaney, A., Downes, K., Ghori, M. J., Gwilliam, R., Hunt, S. E., Inouye, M., Keniry, A., King, E., McGinnis, R., Potter, S., Ravindrarajah, R., Whittaker, P., Widden, C., Withers, D., Deloukas, P., Leung, H. T., Nutland, S., Stevens, H. E., Walker, N. M., Todd, J. A., Easton, D., Clayton, D. G., Burton, P. R., Tobin, M. D., Barrett, J. C., Evans, D., Morris, A. P., Cardon, L. R., Cardin, N. J., Davison, D., Ferreira, T., Pereira-Gale, J., Hallgrimsdottir, I. B., Howie, B. N., Marchini, J. L., Spencer, C. C., Su, Z., Teo, Y. Y., Vukcevic, D., Donnelly, P., Bentley, D., Brown, M. A., Gordon, L. R., Caulfield, M., Clayton, D. G., Compston, A., Craddock, N., Deloukas, P., Donnelly, P., Farrall, M., Gough, S. C., Hall, A. S., Hattersley, A. T., Hill, A. V., Kwiatkowski, D. P., Mathew, C., McCarthy, M. I., Ouwehand, W. H., Parkes, M., Pembrey, M., Rahman, N., Samani, N. J., Stratton, M. R., Todd, J. A. & Worthington, J. (2007), 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature* **447**(7145), 661–678.

Devlin, B. & Roeder, K. (1999), 'Genomic control for association studies', *Biometrics* **55**(4), 997–1004.

Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., Steinhart, A. H., Abraham, C., Regueiro, M., Griffiths, A., Dassopoulos, T., Bitton, A., Yang, H., Targan, S., Datta, L. W., Kistner, E. O., Schumm, L. P., Lee, A. T., Gregersen, P. K., Barmada, M. M., Rotter, J. I., Nicolae, D. L. & Cho, J. H. (2006), 'A genome-wide association study identifies IL23R as an inflammatory bowel disease gene', *Science* **314**(5804), 1461–1463.

Guan, Y. (2014), 'Detecting structure of haplotypes and local ancestry', *Genetics* **196**(3), 625–642.

Holm, H., Gudbjartsson, D. F., Sulem, P., Masson, G., Helgadottir, H. T., Zanon, C., Magnusson, O. T., Helgason, A., Saemundsdottir, J., Gylfason, A., Stefansdottir, H., Gretarsdottir, S., Matthiasson, S. E., Thorgeirsson, G. M., Jonasdottir, A., Sigurdsson, A., Stefansson, H., Werge, T., Rafnar, T., Kiemeney, L. A., Parvez, B., Muhammad, R., Roden, D. M., Darbar, D., Thorleifsson, G., Walters, G. B., Kong, A., Thorsteinsdottir, U., Arnar, D. O. & Stefansson, K. (2011), 'A rare variant in MYH6 is associated with high risk of sick sinus syndrome', *Nat. Genet.* **43**(4), 316–320.

Jorgensen, M. E., Borch-Johnsen, K., Stolk, R. & Bjerregaard, P. (2013), 'Fat distribution and glucose intolerance among Greenland Inuit', *Diabetes Care* **36**(10), 2988–2994.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J. & Eskin, E. (2008), 'Efficient control of population structure in model organism association mapping', *Genetics* **178**(3), 1709–1723.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C. & Hoh, J. (2005), 'Complement factor H polymorphism in age-related macular degeneration', *Science* **308**(5720), 385–389.

Lake, S. L., Lyon, H., Tantisira, K., Silverman, E. K., Weiss, S. T., Laird, N. M. & Schaid, D. J. (2003), 'Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous', *Hum. Hered.* **55**(1), 56–65.

Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. (2013), 'RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference', *Am. J. Hum. Genet.* **93**(2), 278–288.

Moltke, I., Fumagalli, M., Korneliussen, T. S., Crawford, J. E., Bjerregaard, P., Jorgensen, M. E., Grarup, N., Gullov, H. C., Linneberg, A., Pedersen, O., Hansen, T., Nielsen, R. & Albrechtsen, A. (2015), 'Uncovering the genetic history of the present-day greenlandic population', *Am. J. Hum. Genet.* **96**(1), 54–69.

Moltke, I., Grarup, N., Jorgensen, M. E., Bjerregaard, P., Treebak, J. T., Fumagalli, M., Korneliussen, T. S., Andersen, M. A., Nielsen, T. S., Krarup, N. T., Gjesing, A. P., Zierath, J. R., Linneberg, A., Wu, X., Sun, G., Jin, X., Al-Aama, J., Wang, J., Borch-Johnsen, K., Pedersen, O., Nielsen, R., Albrechtsen, A. & Hansen, T. (2014), 'A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes', *Nature* **512**(7513), 190–193.

Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. (2011), 'Genotype and SNP calling from next-generation sequencing data', *Nat. Rev. Genet.* **12**(6), 443–451.

Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G. K., Tandon, A., Kao, W. H., Ruczinski, I., Fornage, M., Siscovick, D. S., Zhu, X., Larkin, E., Lange, L. A., Cupples, L. A., Yang, Q., Akylbekova, E. L., Musani, S. K., Divers, J., Mychaleckyj, J., Li, M., Papanicolaou, G. J., Millikan, R. C., Ambrosone, C. B., John, E. M., Bernstein, L., Zheng, W., Hu, J. J., Ziegler, R. G., Nyante, S. J., Bandera, E. V., Ingles, S. A., Press, M. F., Chanock, S. J., Deming, S. L., Rodriguez-Gil, J. L., Palmer, C. D., Buxbaum, S., Ekunwe, L., Hirschhorn, J. N., Henderson, B. E., Myers, S., Haiman, C. A., Reich, D., Patterson, N., Wilson, J. G. & Price, A. L. (2011), 'Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium', *PLoS Genet.* **7**(4), e1001371.

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., O'Brien, S. J., Altshuler, D., Daly, M. J. & Reich, D. (2004), 'Methods for high-density admixture mapping of disease genes', *Am. J. Hum. Genet.* **74**(5), 979–1000.

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', *Nat. Genet.* **38**(8), 904–909.

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D. & Myers, S. (2009), 'Sensitive detection of chromosomal segments of distinct ancestry in admixed populations', *PLoS Genet.* **5**(6), e1000519.

Sankararaman, S., Sridhar, S., Kimmel, G. & Halperin, E. (2008), 'Estimating local ancestry in admixed populations', *Am. J. Hum. Genet.* **82**(2), 290–303.

Skotte, L., Korneliussen, T. S. & Albrechtsen, A. (2012), 'Association testing for next-generation sequencing data using score statistics', *Genet. Epidemiol.* **36**(5), 430–437.

Skotte, L., Korneliussen, T. S. & Albrechtsen, A. (2013), 'Estimating individual admixture proportions from next generation sequencing data', *Genetics* **195**(3), 693–702.

Sparso, T., Grarup, N., Andreasen, C., Albrechtsen, A., Holmkvist, J., Andersen, G., Jorgensen, T., Borch-Johnsen, K., Sandbaek, A., Lauritzen, T., Madsbad, S., Hansen, T. & Pedersen, O. (2009), 'Combined analysis of 19 common validated type 2 diabetes susceptibility gene variants shows moderate discriminative value and no evidence of gene-gene interaction', *Diabetologia* **52**(7), 1308–1314.

Thorleifsson, G., Walters, G. B., Gudbjartsson, D. F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., Styrkarsdottir, U., Gretarsdottir, S., Thorlacius, S., Jonsdottir, I., Jonsdottir, T., Olafsdottir, E. J., Olafsdottir, G. H., Jonsson, T., Jonsson, F., Borch-Johnsen, K., Hansen, T., Andersen, G., Jorgensen, T., Lauritzen, T., Aben, K. K., Verbeek, A. L., Roeleveld, N., Kampman, E., Yanek, L. R., Becker, L. C., Tryggvadottir, L., Rafnar, T., Becker, D. M., Gulcher, J., Kiemeney, L. A., Pedersen, O., Kong, A., Thorsteinsdottir, U. & Stefansson, K. (2009), 'Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity', *Nat. Genet.* **41**(1), 18–24.

Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., Ng, D. P., Holmkvist, J., Borch-Johnsen, K., Jorgensen, T., Sandbaek, A., Lauritzen, T., Hansen, T., Nurbaya, S., Tsunoda, T., Kubo, M., Babazono, T., Hirose, H., Hayashi, M., Iwamoto, Y., Kashiwagi, A., Kaku, K., Kawamori, R., Tai, E. S., Pedersen, O., Kamatani, N., Kadowaki, T., Kikkawa, R., Nakamura, Y. & Maeda, S. (2008), 'SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations', *Nat. Genet.* **40**(9), 1098–1102.

Yorgov, D., Edwards, K. L. & Santorico, S. A. (2014), 'Use of admixture and association for detection of quantitative trait loci in the Type 2 Diabetes Genetic Exploration by Next-Generation Sequencing in Ethnic Samples (T2D-GENES) study', *BMC Proceedings* **8**(Suppl 1), S6.

Zeggini, E. (2014), 'Using genetically isolated populations to understand the genomic basis of disease', *Genome Med* **6**(10), 83.

Zhou, X. & Stephens, M. (2012), 'Genome-wide efficient mixed-model analysis for association studies', *Nat. Genet.* **44**(7), 821–824.

Appendix

Figures

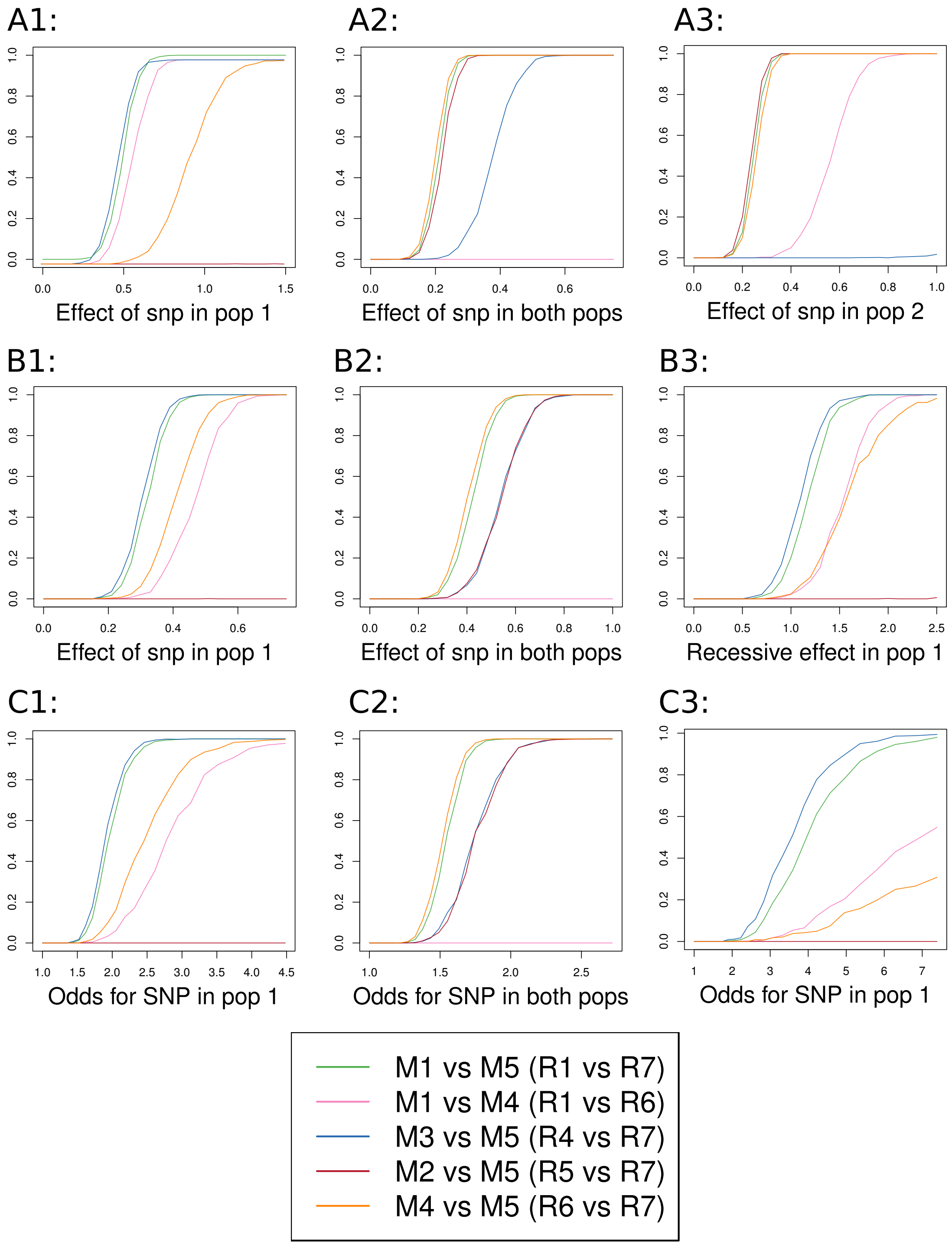

Figure S1: Results of power simulations for all nine scenarios (A1-C3). Curves show the fraction of simulated p-values that are smaller than 10^{-8} based on 1000 simulations for each effect size for each scenario. The simulated scenarios are described in Table 3 and the tests are described in Table 2 and Table S2.

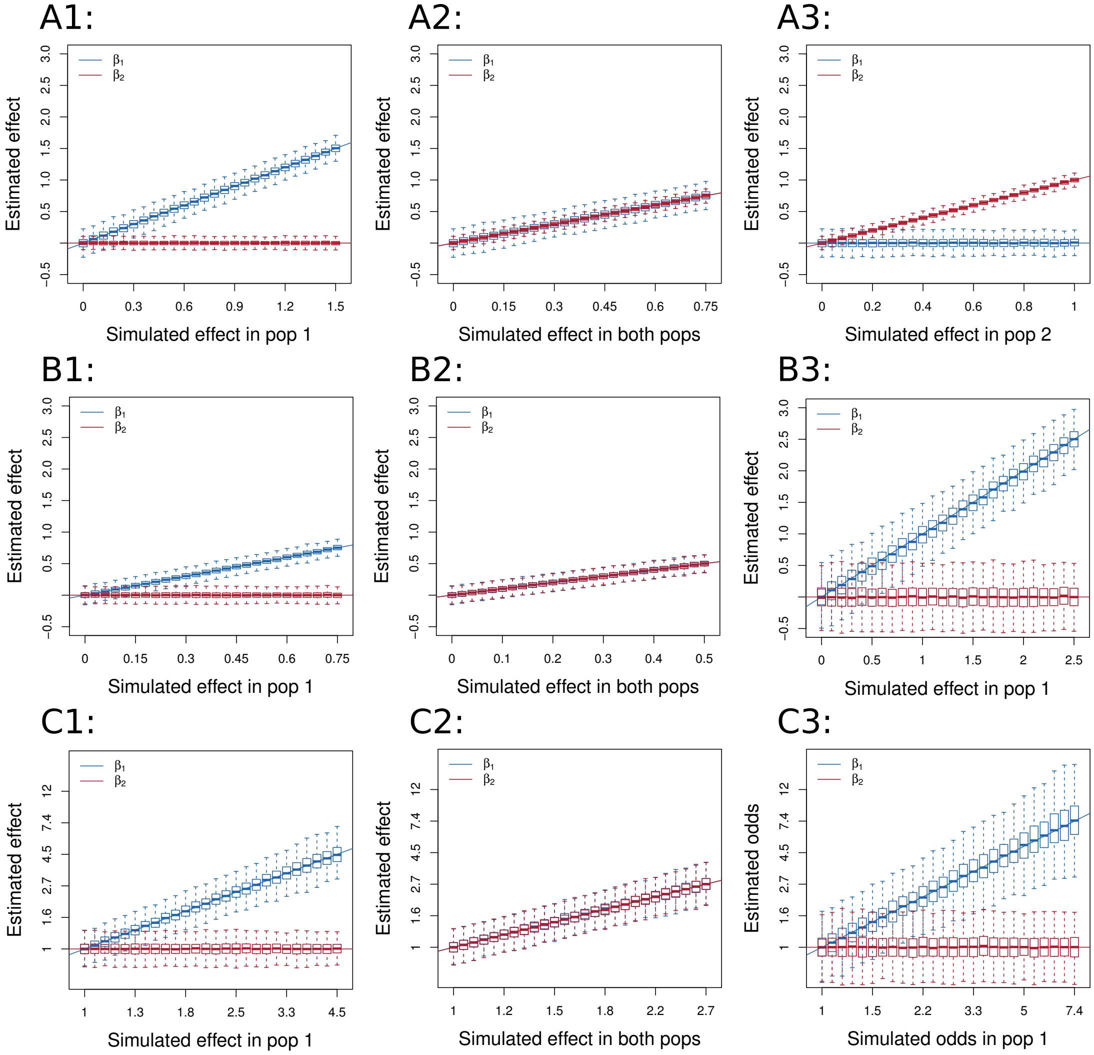

Figure S2: Results of bias simulations for all nine scenarios (A1-C3). Each boxplot is based on 1000 simulations. The simulated scenarios are described in Table 3 and the tests are described in Table 2 and Table S2

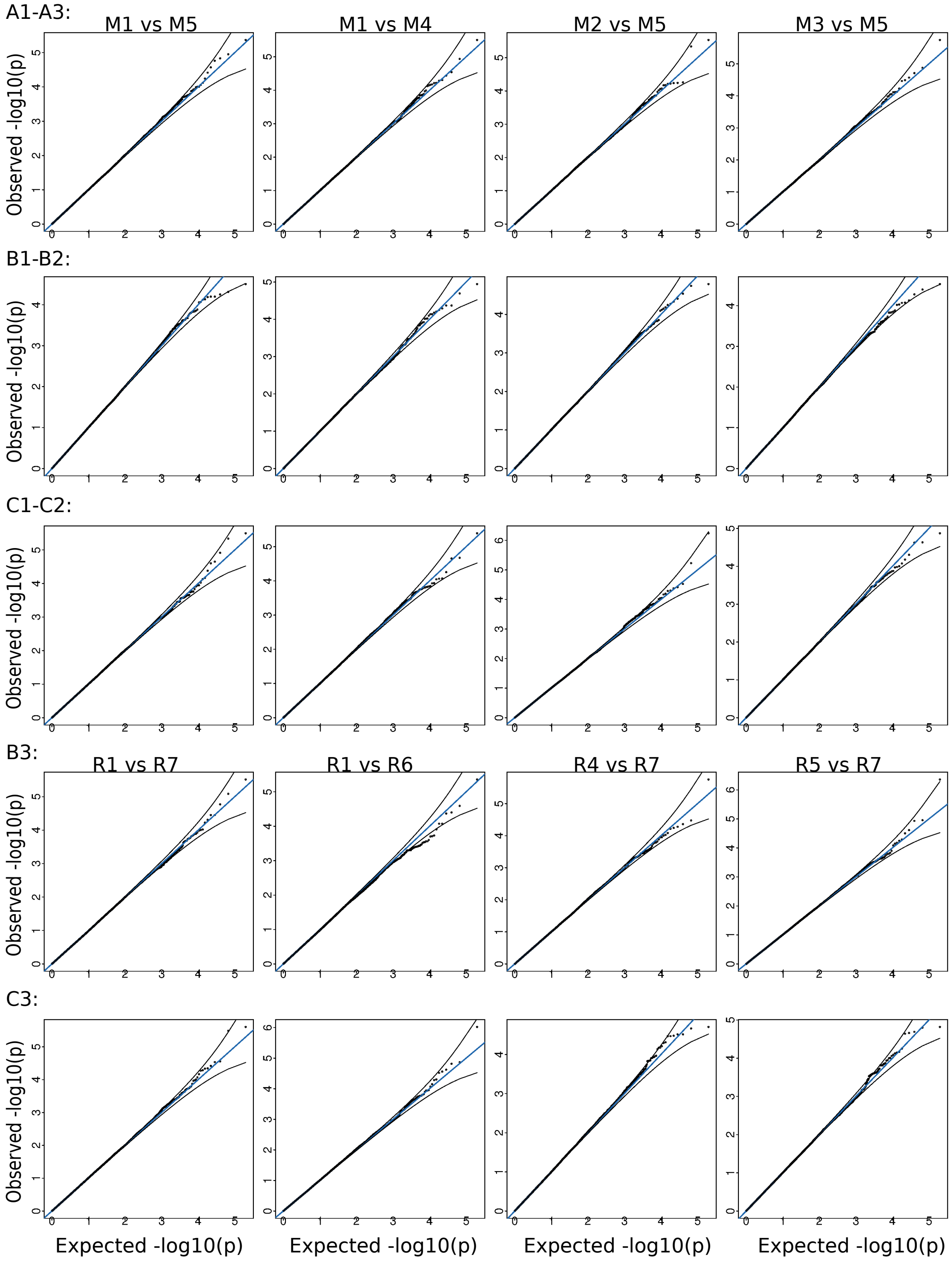

Figure S3: Results of P-value simulations for the null model of all scenarios (A1-C3). Each QQ-plot is based on 100000 simulations. The simulated scenarios are described in Table 3 and the tests are described in Table 2 and Table S2. The 3 top rows show results for the additive asaMap tests, while the 2 bottom rows show results for the reccessive asaMap tests.

Case-control study modification

For case-control studies or dichotomous traits the analysis is based on a logistic regression analysis, since this allows the inclusion of additional covariates (such as principal components) in the model. Given the locus state s, the probability π_i that individual i is affected is modeled by

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \alpha + \sum_{k=1}^{K} \beta_k x_k(s) + \sum_c \gamma_c z_c^i. \tag{4}$$

where x_k are the design matrix entries for the population specific effects and z_c^i is the value of covariate c for individual i.

Recessive genetic model

In addition to the additive genetic model, a recessive model has been implemented in asaMap. Its design is slightly more complicated. This is because we wish to have the best possible power to detect the genetic effect of a recessive disease causing variant that may remain untyped and only be present in one of the ancestral populations. Due to this more complex modelling, the recessive genetic model has only been implemented for $K = 2$ ancestral populations.

The recessive model assumes that there only is an effect, when two variant alleles are present. The full ancestry specific recessive model estimates three different effects: β_1 is the effect of having two variant alleles inherited from population 1, β_2 is the effect of having two variant alleles inherited from population 2 and β_m is the effect of having two variant alleles, when one is inherited from each of the ancestral populations.

This allows us to fit range of submodels (see table S1). The models R4 and R5 are of most interest. In R4 it is assumed that there is no effect of being homozygous for the variant allele - unless both variant alleles are inherited from population 1. This model can then be compared against R7 - where it is assumed that there is no effect of the variant to test if there is a sigificant recessive effect of the variant when inherited from population 1. This test (R4 vs. R7) is particularly powerful when the tested variant is only in LD with the causal recessive variant in population 1, particularly if the causal variant is extremely rare in population 2. To test if the model assumption for model R4 is appropriate we can compare against R1 in the test R1 vs. R4 and to test if there is a different recessive effect when both alleles are inherited from population 1 than otherwise, we can compare R2 to R6, where the effect is assumed to be independent of ancestry. In the exact same way R5 vs. R7 can be used to test for an recessive effect when both variants are inherited from population 2 and R1 vs. R5 can be used to check the assumptions of this test.

Model	Hypothesis	The model assumes
R1	$(\beta_1, \beta_m, \beta_2) \in R^3$	population specific effects
R2	$\beta_1 \in R, \beta_m = \beta_2 \in R$	same effect when one or both variant alleles are from pop 2
R3	$\beta_1 = \beta_m \in R, \beta_2 \in R$	same effect when one or both variant alleles are from pop 1
R4	$\beta_1 \in R, \beta_m = \beta_2 = 0$	only an effect when both variant alleles are from pop 1
R5	$\beta_1 = \beta_m = 0, \beta_2 \in R$	only an effect when both variant alleles are from pop 2
R6	$\beta_1 = \beta_m = \beta_2 \in R$	same effect regardless of ancestry
R7	$\beta_1 = \beta_m = \beta_2 \in R$	no effect

Table S1: Recessive ancestry specific genetic effects models

Models	Tests if there is
R1 vs. R7	a recessive effect for some combination of ancestry
R1 vs. R4	only an effect when both alleles are from pop 1
R1 vs. R5	only an effect when both alleles are from pop 2
R1 vs. R6	any ancestry dependence of the effect
R2 vs. R6	a different effect when both alleles are from pop 1
R3 vs. R6	a different effect when both alleles are from pop 2
R4 vs. R7	a recessive effect in population 1
R5 vs. R7	a recessive effect in population 2
R6 vs. R7	a recessive effect assuming it is independent of ancestry

Table S2: Recessive ancestry specific genetic effects models

EM algorithm

Notation

n individuals
y_i phenotype of ind $i \in \{1, \ldots, n\}$
g_i genotype of ind i, $g \in \{0, 1, 2\}$
Q admixture proportions, for individual i, $Q^i = \{q_1^i, \ldots, q_K^i\}$.
$f = \{f_1, \ldots, f_K\}$ population specific allele frequencies
ϕ vector of effect sizes (β_1, β_2) and other regression parameters (α, γ, σ)
s locus state, $s = \{a, t\}$ where $a = (a_1, a_2)$ is the information on ancestry and $t = (t_1, t_2)$ is allelic genotypes.

Likelihood functions

The likelihood for the observed data, assuming that individuals are independent given genotypes, admixture proportion and population specific allele frequencies, is given by:

$$p(Y|G, Q, f, \phi) = \prod_i p(y^i|g^i, Q, f, \phi) \tag{5}$$

and splitting the probabilities according to locus state gives

$$\begin{aligned} p(y_i|g_i, Q, f, \phi) &= \sum_s p(y_i, s|g_i, Q, f, \phi) &(6)\\ &= \sum_s p(y_i|s, \phi)p(s|g_i, f, Q) &(7) \end{aligned}$$

Conditional on locus state, the phenotype follows a normal distribution.

$$p(y_i|s, \phi) \sim N(\eta^i(s, \alpha, \beta, \gamma), \sigma^2) \tag{8}$$

with

$$\eta_i(s, \alpha, \beta, \gamma) = \alpha + \sum_k \beta_k x_k(s) + \sum_c \gamma_c z_c^i \tag{9}$$

where we use $x_k(s) = t_1 1_{a_1=k} + t_2 1_{a_2=k}$ for the additive genetic model. The normal distribution is part of the exponential family. The density of a normal with mean η and standard deviation σ is

$$\begin{aligned} f(y|\eta, \sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(y-\eta)^2}{2\sigma^2} \\ &= \exp\left(\frac{y\eta - \eta^2/2}{\sigma^2} - y^2/2\sigma^2 - \log(2\pi\sigma^2)/2\right) \\ &= \exp\left(\frac{y\eta - b(\eta)}{a(\sigma)} + c(y, \sigma)\right) &(10) \end{aligned}$$

with $b(\eta) = \eta^2/2$, $a(\sigma) = \sigma^2$ and $c(y, \sigma) = -y^2/2\sigma^2 - \log(2\pi\sigma^2)/2$.

Derivation of EM algorithm

The expression that must be maximized in a single EM algorithm step is:

$$E_{S|Y,G,\phi^*,Q,f}[\log p(Y,S|G,\phi,Q,f)] = \sum_i E_{s_i|y_i,G,\phi^*,Q,f}[\log p(y_i,s_i|G,\phi,Q,f)] \tag{11}$$

as a function of all regression parameters ϕ, where ϕ^* is fixed to the value from previous iteration. Using $p(y_i,s_i|G,\phi,Q,f) = p(y_i|s_i,\phi)p(s_i|Q,f)$ where the second term does not depend on parameters to be optimized, this is equivalent to maximizing

$$\sum_i E_{s_i|y_i,G,\phi^*,Q,f}[\log p(y_i|s_i,\phi)] = E_{S|Y,G,\phi^*,Q,f}[\log p(Y|S,\phi)] \tag{12}$$

Following Lake et al. (2003) we get by taking the derivative with respect to the vector of population specific effect sizes:

$$\begin{aligned}\frac{\partial}{\partial\beta}E_{s_i|y_i,G,\phi^*,Q,P}[\log p(y_i|s_i,\phi)] &= E_{s_i|y_i,G,\phi^*,Q,f}[\frac{\partial}{\partial\beta}\log p(y_i|s_i,\phi)]\\ &= E_{s_i|y_i,G,\phi^*,Q,f}[\frac{\partial\eta_i}{\partial\beta}\frac{\partial}{\partial\eta_i}\log p(y_i|s_i,\phi)]\\ &= E_{s_i|y_i,G,\phi^*,Q,f}[x^i\frac{y_i-b'(\eta_i)}{a(\sigma)}]\\ &= \sum_{s_i} x^i\frac{y_i-\eta_i}{\sigma^2}p(s_i|y_i,G,\phi^*,Q,f)\end{aligned} \tag{13}$$

where $x^i = (x_1(s_i),\ldots,x_k(s_i))$ and $\eta_i = \eta(x^i,z^i,\alpha,\beta,\gamma)$. The equivalent formula holds for the derivative with respect to α and γ. We therefore get

$$\frac{\partial}{\partial\phi}E_{s|y,G,\phi^*,Q,P}[\log p(y|s,\phi)] = \sum_i\sum_{s_i}(x^i,z^i)\frac{y^i-\eta^i(\phi)}{\sigma^2}p(s_i|y_i,G,\phi^*,Q,f) \tag{14}$$

which is recognized as the score function for a weighted regression where each individual, i, contributes one observation per possible state s_i and where the weights, $p(s_i|y_i,g_i,\phi^*,Q,f)$, are the posterior distribution of states given all the observed data and based on the previously fitted parameters (see below for details). The same formula holds for the logistic regression. The updated regression parameters α,β and γ can therefore be estimated by fitting a weighted linear regression in case of a quantitative trait and a weighted logistic regression for case/control data.

Posterior distribution of locus state

The conditional distribution of locus state given previous parameters, observed data and genotype is found using

$$p(s|y_i,G,\phi^*,Q,f) = \frac{p(y_i|s,\phi^*)p(s|g_i,Q,f)}{\sum_{s'}p(y_i|s,\phi^*)p(s'|g_i,Q,f)} \tag{15}$$

where $p(s|g_i,Q,f)$ is given in Equation 3 and $p(y_i|s,\phi^*)$ is the phenotype distribution given locus state and previous parameters.

Optimization strategy for normal dist trait

First a rough initial guess of the standard variation is calculated by

$$\sigma = \mathrm{var}[y] \tag{16}$$

and randomly chosen start values for the regression parameters are sampled from

$$\alpha, \beta, \gamma \sim \mathrm{runif}(-1, 1) \tag{17}$$

Then regression weights are calculated according to (15) and a weighted regression according to the score function in (14) is carried out to update β, γ. This is followed by an update of σ using the weighted sum of squared residuals from the weighted regression and $n - p$ df, where n is number of individuals and p is the number of effect parameters in the linear predictor ($p = 1 + K + C$).