bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ancestry-specific association mapping in admixed populations

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

1: Department of Epidemiology Research,
Statens Serum Institut, 2300 Copenhagen, Denmark

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

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

+: Contributed equally to this work.

*: Corresponding author: albrecht@binf.ku.dk, ida@binf.ku.dk

Keywords: Local ancestry, association mapping, admixture, power, GWAS.

Running title: Mapping ancestry specific effects.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

During the last decade genome–wide association studies have proven to be a powerful approach to identifying disease-causing variants. However, for admixed populations, most current methods for performing association testing are based on the assumption that the effect of a genetic variant is the same regardless of its ancestry. This is a reasonable assumption for a causal variant, but may not hold for the genetic variants that are tested in genome–wide association studies, which are usually not causal. The effects of non-causal genetic variants depend on how strongly their presence correlate with the presence of the causal variant, which 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 size is allowed to depend on the ancestry of a given 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 help determine if a SNP is causal. We demonstrate the usefulness of the method on data from the Greenlandic population.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

An individual's risk of developing common complex diseases, such as type 2 diabetes, is influenced by specific 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 homogeneous populations and this has led to important new findings (Morris et al. 2012, Voight B.F., et al. 2010, Scott et al. 2007). However, a few GWAS have now been performed in recently admixed populations (Moltke et al. 2014, The SIGMA Type 2 Diabetes Consortium 2013), and these studies have led to interesting new findings suggesting that such studies can be valuable to perform. This was recently shown very clearly in a study by Moltke et al. (2014) where a GWAS performed in the Greenlandic population, which has both Inuit and European ancestry, led to the identification of a variant that explains more than 10% of all cases of type 2 diabetes in Greenland. Notably, the variant had not been identified in earlier much larger studies of homogeneous populations like the European and East Asian populations, because it is very rare in these populations.

While being valuable, performing GWAS in recently admixed populations involves an important challenge: the admixture can bias the statistical test

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 linkage disequilibrium (LD) with it (see Figure 1). The effect size and strength of the association of a

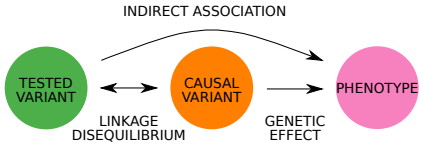

Figure 1: Indirect association between tested genetic variant and a phenotype.

variant tested in a GWAS will therefore depend on the allele frequencies of the causal and tested variants and the strength of the LD between them. 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 given variant may depend strongly on the ancestry of the chromosomal segment on which this variant is located. The extreme case

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

shown in Figure 2 provides a simple illustrative example. Here the tested variant is present in both populations, but the causal variant is only present in ancestral population 1, where it is in complete LD with the tested variant. In this example, the tested variant is clearly more strongly associated with the phenotype 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. This 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: the association with the tested variant in the example is much weaker than the association with the causal variant, which in this case equals the association with the tested variant inherited from population 1. This means that a GWAS can potentially gain power by allowing for 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, asaMap 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

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

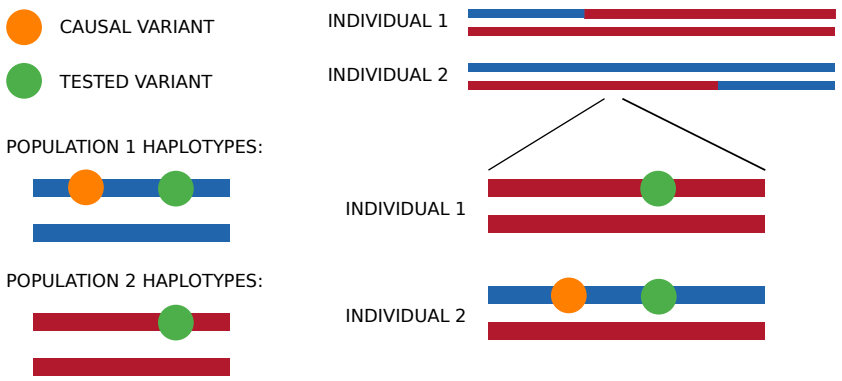

Figure 2: Extreme case of ancestry-specific effects in a population with ancestry from two populations: the tested variant is present in both ancestral populations, but the causal variant is only present in ancestral population 1, where it is in complete LD with the tested variant. The figure shows the 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.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mixture components are the phenotype distributions corresponding to given ancestries of the tested SNP and the mixture weights are the probabilities of these ancestries (for more details see Materials and Methods). This approach 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 weights. The mixture weights for a given SNP are in asaMap by default calculated from genome-wide admixture proportions, population specific allele frequencies and genotypes. However, asaMap also works with provided mixture weights and thus allows users to use more complex models such as hidden Markov models (Patterson et al. 2004, Price et al. 2009, Guan 2014) for obtaining these. The mixture components in asaMap are based on a generalized linear model (GLM) 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 GLM – 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 traits) as well as several different genetic effect types (additive and recessive effects). Third, it allows easy correction for any additional covariates such as sex or age.

We emphasize that asaMap is an association testing method and not a method for performing admixture mapping; a mapping approach where correlation between phenotype and inferred local ancestry is used to identify

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 and tested based on inferred knowledge about local ancestry. These methods focus on the potentially different effects of one specific allele (which we will refer to as "the assumed effect allele") in the ancestral populations and they assume that the other allele (which we will refer to as "the assumed non-effect allele") has the same mean phenotypic value in all ancestral populations after correcting for relevant covariates. Hence they ignore the possibility that it may be the assumed non-effect allele, and not the assumed effect allele, that mediates an ancestry-specific effect.

Importantly, although asaMap is similar to the methods of Pasaniuc et al. (2011) and Yorgov et al. (2014) in the sense that it allows for ancestry-specific effect, it differs from those method in several ways. First, as already described, asaMap does not require prior knowledge of the ancestry of each allele. Second, it enables correction for population structure. Third, although asaMap also focuses on the potentially different effects of a specific assumed effect allele, it offers an additional test that allows the user to assess if it may be beneficial to use the other allele as the assumed effect allele instead.

In the following section, we describe the model behind asaMap in detail. Then using simulated data, we 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 GLM, which is often used for association testing in GWAS. For example, we show that asaMap makes

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

it possible to test whether a specific allele has different effects in different ancestral populations. And since it is reasonable to assume that a disease-causing allele have the same effect regardless of its ancestry, such a test can be used to reject that a SNP identified in a GWAS is causal. Finally, using data from a GWAS in the admixed Greenlandic population (Moltke et al. 2014), we 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.

Materials and methods

Model

Our asaMap model framework is based on a GLM. It therefore applies to both quantitative traits and case-control traits (i.e. dichotomous traits) and allows for both additive and recessive genetic effects. Here we describe the quantitative trait model for additive genetic effects, while detailed descriptions of the model for case-control data as well as recessive genetic effects can be found in the Supporting Information.

As argued in the introduction, the strength of an association between genotype and phenotype is likely to depend on which of the ancestral populations a genetic allele has been inherited from. Thus, instead of estimating a single genetic effect of a given allele, we here allow for population-specific

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

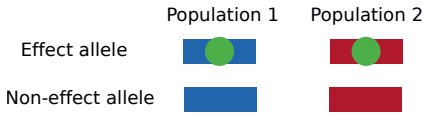

Figure 3: The four ancestry-specific allele types that can occur at a diallelic autosomal locus in an individual from a population that consists of two ancestral admixing populations.

genetic effect sizes, each of which we denote β_k for an allele inherited from ancestral population k. Below we describe the model that allows us to do this. The description is limited to two ancestral populations, but is easily extended to more than two populations.

Mixture model. We assume that we are analyzing data from N individuals from an admixed population that is a mixture of two ancestral populations. An individual from such a population will at any given diallelic autosomal locus have inherited each of its two chromosome copies from one of the two ancestral populations, and each of these will either carry the effect allele or not. Thus there are four possible ancestry-specific allele types, as illustrated in Figure 3. As a consequence, each individual will have one of 16 different ancestry-specific allele type combinations at any given diallelic autosomal locus. We will refer to these ancestry-specific allele type combinations as *locus states*, s. The 10 distinguishable locus states are shown in Figure 4. The locus state is unfortunately not directly observable from genotype data. It can sometimes be inferred, but local ancestry inference from genotype data is associated with uncertainty and ignoring this may lead to false positives.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

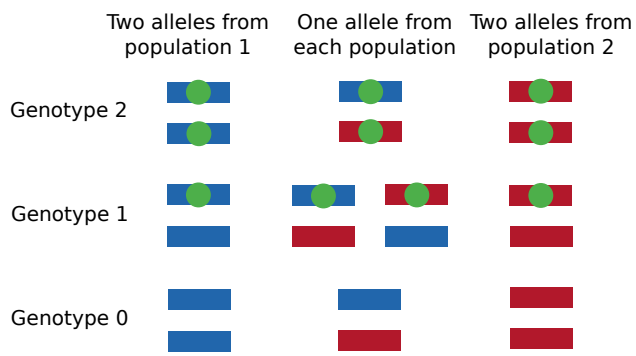

Figure 4: The 10 distinguishable locus states (combinations of two ancestry-specific allele types) that can occur at a diallelic autosomal locus in an individual from a population that consists of two ancestral admixing populations.

Therefore, we have chosen not to base asaMap on inferred locus states, but to instead use a model that allows us to take the uncertainty of the locus states 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 copies of a specified allele present at the tested locus in each individual, the likelihood function for the observed phenotypes, $Y = (y_1, y_2, ..., y_N)$, takes the form

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

Here G is a vector of all observed genotypes at the tested locus, Z is a matrix of appropriately chosen covariates, the product runs over all individuals $i = 1 \ldots N$ and the sum runs over all possible locus states s. Assuming that the trait is conditionally independent of the observed genotypes G given the

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 that we can model the probability of the observed phenotypes Y as a mixture of phenotype distributions, where each mixture component is the phenotype distribution given a specific locus state s, $p(y|s,Z)$, and the corresponding mixture weight is the probability $p(s|G)$ of that locus state given the observed genotypes. Notably, 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 weights. Furthermore, the above likelihood function is a function of our parameters of interest, namely the population-specific effects β_k, via the mixture components $p(y|s,Z)$. Hence we can use the model both for estimating the ancestry-specific effects β_k and for performing association testing. Specifically, in asaMap 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 weights in the likelihood function, followed by a detailed description of the parameter estimation and testing procedures.

Mixture components. For a given 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

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for a single individual i follows a normal distribution with mean given by the linear predictor

$$\eta_i = \alpha + \beta_1 x_1(s) + \beta_2 x_2(s) + \sum_c \gamma_c z_c^i. \tag{2}$$

Here α parameterizes the intercept (baseline) and additional covariates such as 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, assuming an additive model, $x_1(s)$ and $x_2(s)$ are the counts of the assumed effect allele from population 1 and 2, respectively, for a locus in locus state s. If one instead assumes a recessive model, the definition of the terms $x_1(s)$ and $x_2(s)$ is different and is described in more details in the Supporting Information.

Mixture weights. The simplest approach to calculating the mixture weights, i.e. the probabilities of the different possible locus states for individual i given i's genotype g, is to use i's genome-wide admixture proportions Q^i and the population specific allele frequencies f, both of which can be inferred using standard software tools such as ADMIXTURE (Alexander et al. 2009). In the case of two admixing populations the genome-wide admixture proportions are $Q^i = (q_1^i, q_2^i)$, where q_1^i and q_2^i is the fraction of is genome that has been inherited from population 1 and 2 respectively, and the population specific allele frequencies are $f = (f_1, f_2)$, where f_1 and f_2 is the frequency of assumed effect allele in population 1 and 2, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

$$p(t|g) = \begin{cases} 1 & \text{if} \quad g = 0 \\ \frac{1}{2} & \text{if} \quad g = 1 \\ 1 & \text{if} \quad g = 2 \end{cases} \cdot$$

In the the second step we consider the probability of the ordered allelic ancestry $a = (a_1, a_2)$ given the ordered allelic genotype $t = (t_1, t_2)$. Here we use the genome-wide admixture proportions Q^i to give the probability $p(a|f, Q^i) = q^i_{a_1} q^i_{a_2}$ of ancestry a assuming independent ancestry of alleles and we use the corresponding population specific allele frequencies f to calculate the probability of ordered allelic genotype given the ordered allelic ancestry $p(t_j|a_j, f, Q^i) = f^{t_j}_{a_j}(1 - f_{a_j})^{1-t_j}$. With this the desired probability of the ordered allelic ancestry a given the ordered allelic genotype t can be

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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, splitting the joint conditional probability into the two terms we have just derived:

$$p(s|g,f,Q) = p(a,t|g,f,Q) = p(a|t,g,f,Q)p(t|g,f,Q) = p(a|t,f,Q)p(t|g) \tag{3}$$

In the last derivation step we exploit the fact that a depends on g via t, and that t only depends on g. This three-step procedure is the default procedure used in asaMap. However, the user can also supply the distribution across locus states to asaMap and can therefore choose to for instance calculate it based on the output from local ancestry inference software instead.

Parameter estimation and hypothesis testing

Parameter estimation. The parameters of the model described above (and of the sub-models relevant for testing purposes described below) are

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 for this (see Supporting Information for details) that provides faster convergence than standard all-purpose numerical optimization algorithms such as BFGS.

Hypothesis testing. tandard GLM based methods that are commonly used for association testing typically assume an additive effect and make use of statistical tests comparing two models: a model where a given allele has a genetic effect versus a nested model where it 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. Specifically, for an additive genetic effect, then assuming 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 (see equation 2). The sub-model M2 assumes no effect in population 1 (i.e. $\beta_1 = 0$). The sub-model M3 assumes no effect in population 2 (i.e. $\beta_2 = 0$). The sub-model M4 assumes that the effect sizes are the same in both ancestral populations (i.e. $\beta_1 = \beta_2$), and finally the null sub-model M5 assumes that there is no effect in any of the populations (i.e. $\beta_1 = \beta_2 = 0$). An overview of these additive models is given in Table 1 and a detailed description can be found in the Supporting Information. For recessive genetic effects the standard GLM

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 five different possible additive models when there are two ancestral populations. β_1 is the effect size of alleles from population 1 and β_2 is the effect size of alleles from population 2. Comparing different pairs of nested models leads to the tests described in Table 2. For a description of recessive models, 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: Description of the possible tests comparing different nested models (described in table 1) assuming an additive genetic model.

based methods for association mapping tests a model where carrying two copies of a given allele has an effect on the individual's phenotype versus a model where carrying two copies of the allele has no effect. Here asaMap allows the effect size to be specific to the ancestry combination of the two allele copies. To do this seven (sub-)models R1-R7 described in the Supporting Information and Table S1 are implemented.

In asaMap hypothesis tests regarding the ancestry-specific effect sizes is carried out using likelihood ratio tests comparing nested models among the

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

models just described. The implemented models under the additive effect assumption allow 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). Furthermore, they allow us to test if there is an effect assuming that it is the same in the two populations (M4 vs. M5). Note that this latter test is equivalent of the standard test for association performed using a GLM and it has been implemented in asaMap to enable comparison of the other tests to the commonly used standard GLM based testing approach. In addition, the tests M1 vs. M2 and M1 vs. M3 are also implemented and may be used for performing model checks 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 Supporting Information and Table S2.

Implementation. The estimation and testing procedures described above have been implemented in the software asaMap available at `https://github.com/e-jorsboe/asaMap` (C^{++}) and `https://github.com/lineskotte/asamap` (R-package), making the method applicable to large scale genome wide association studies.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Testing if the assumed non-effect allele has a population specific effect

As described in the introduction, like similar methods by (Pasaniuc et al. 2011) and (Yorgov et al. 2014), asaMap focuses on the potentially different effects of one specific allele in the ancestral populations, the assumed effect allele, and ignores the possibility that it may in fact be the other allele, the assumed non-effect allele, that mediates an ancestry-specific effect. To test if this is the case, and thus that association testing focused on this other allele instead would be preferred, we introduce an additional model M0, which is an extension to model M1. In M0 the phenotype y_i for a single individual i follows a normal distribution with mean given by the linear predictor:

$$\eta_i = \alpha + \beta_1 x_1(s) + \beta_2 x_2(s) + \delta_1 u_1(s) + \sum_c \gamma_c z_c^i, \tag{4}$$

Here the additional predictor u_1 is the count of the assumed non-effect allele from population 1 and the parameter δ_1 is effect size of the assumed non-effect allele in population 1. The other symbols are the same as in equation 2. We use this model by comparing it to model M1 (Table 1), and test if we can reject the null hypothesis of δ_1=0. If so there is evidence that the assumed non-effect allele has an effect and that this allele may be beneficial to use as the assumed effect allele. Note that model M0 could equally well have been formulated using the assumed non-effect allele from population 2

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

instead of population 1. This would lead to the same test, because the total count of alleles across all allele types and ancestries for each individual has to sum to 2 and thus there are only three independent ancestry-specific allele count parameters.

Note also that if the assumed non-effect allele has an effect but this effect is the same in the two ancestral populations, the M0 vs. M1 test will not be significant. However, if this is the case it does not matter for the asaMap analyses which allele is chosen as the assumed effect allele and so there is no reason to analyze the assumed non-effect allele instead of the assumed effect allele.

The above test is for additive effects. We propose a similar test for recessive effects, in which an additional model, R0, is compared to Model R1. For details about R0 and this test, see Supporting Information.

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 in the context of the models proposed here due to the sum across locus states.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Simulated data

To be able to assess asaMap, we carried out analyses of 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. For all nine scenarios we simulated data from a total of 2500 individuals with admixture proportions from population 1, q_1^i, 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: for each individual we sampled the ordered allelic ancestry $a = (a_1, a_2)$ based on the individual admixture proportions q_1^i. Then based on this ancestry a and the allele frequencies f in the ancestral populations, we sampled the ordered allele types $t = (t_1, t_2)$ for each individual. Knowing a and t, the genotype and locus state for each individual is known and based on the latter the phenotype of each individual 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

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Scenario	Frequencies, f=(f_1,f_2)	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 scenarios have 2500 individuals and individual admixture proportions from population 1 in $\{0, 0.25, 0.5, 0.75, 1\}$. The allele frequencies $f = (f_1, f_2)$ in the two ancestral populations are shown in the "Frequencies" column. For each scenario we vary the effect size, however, they are restricted in some scenarios as shown in the "Effects" column. E.g. in scenario A1 there is only an effect in population 1, so here the effect size in population 2 is restricted to 0. When there is an effect in both populations, we assume that they are the same. The type of trait simulated is shown in the "Trait" column and the simulated effect model is shown in the "Model" column.

distribution.

These simulations gave us access to association data where the true ancestry-specific effects are known and therefore allowed us to assess the consistency and unbiasedness of the estimators in asaMap. The simulations also allowed us to assess the power of the tests for association implemented in asaMap. Finally, because we explicitly simulate the ancestry-specific allele type combinations (locus states), the simulations allowed 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.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Data for *TBC1D4* in a Greenlandic cohort

To illustrate that allowing for ancestry-specific effect sizes can be informative and appropriate for real data we applied asaMap to genotype data in combination with measurements of 2-hour (2-h) plasma glucose levels of 2575 individuals in the Inuit Health in Transition cohort from the admixed Greenlandic population (Moltke et al. 2014, Jorgensen et al. 2013). More specifically, we applied asaMap to five 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-h 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 that was in a later part of the study identified from sequencing data. The three remaining SNPs were also identified from sequencing data in the search for the causal variant.

The study received ethics approval from the Commission for Scientific Research in Greenland (project 2011-13, ref. no. 2011-056978; and project 2013-13, ref.no. 2013-090702) and was conducted in accordance with the ethical standards of the Declaration of Helsinki, second revision. Participants gave written consent after being informed about the study orally and in writing.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Results

To investigate the cost and benefits of asaMap compared to a standard GLM we first applied both methods to simulated data to compare their statistical power and to assess important statistical properties of asaMap, like bias and false positive rates. Then we applied both methods to real data to compare the range of their potential usage. In both cases we investigated populations that are mixtures between two populations.

Standard GLM based methods for association mapping makes use of statistical tests comparing two models: a model where a specific allele has a genetic effect is compared to a model where the allele 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. Briefly, 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 (by comparing M1 vs. M5), if there is an effect in population 1 (by comparing M3 vs. M5), if there is an effect in population 2 (by comparing M2 vs. M5), and if there is a different effect in the two populations (by comparing M1 vs. M4). Furthermore, it

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

allows us to test if there is an effect assuming that it is the same in the two populations by comparing 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 Supporting Information and Tables S1 and S2.

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

Simulation-based results

To assess the power and other statistical properties of asaMap we first simulated data for individuals with genetic ancestry from two admixing populations according to nine scenarios (scenarios A1-C3); six with quantitative traits and three with case-control traits (Table 3, see Materials and Methods for details).

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 an additive ef-

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

fect, 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. Furthermore, the tested variant was simulated to have a frequency 10% in ancestral population 1 and 30% in ancestral population 2. When applying asaMap to data from this scenario, the standard GLM based (linear regression) 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. Also, the test of whether 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 expected since it has only 1 degree of freedom and the full test has 2 degrees of freedom. Finally, asaMap could test whether there is a difference 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 GLM based test.

Second, we simulated a scenario (A2), where the tested variant has the same effect in both populations, which corresponds to a situation in which the causal allele is tested. In this scenario, the test of M1 vs. M5 was less powerful than the standard GLM based test (M4 vs. M5), which was expected because of the extra degree of freedom (scenario A2, Figure 5). However, interestingly the difference in power is very small. On the other hand, the test, M3 vs. M5, of whether there is an effect in population 1, where the tested variant has the lowest frequency is markedly less powerful.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

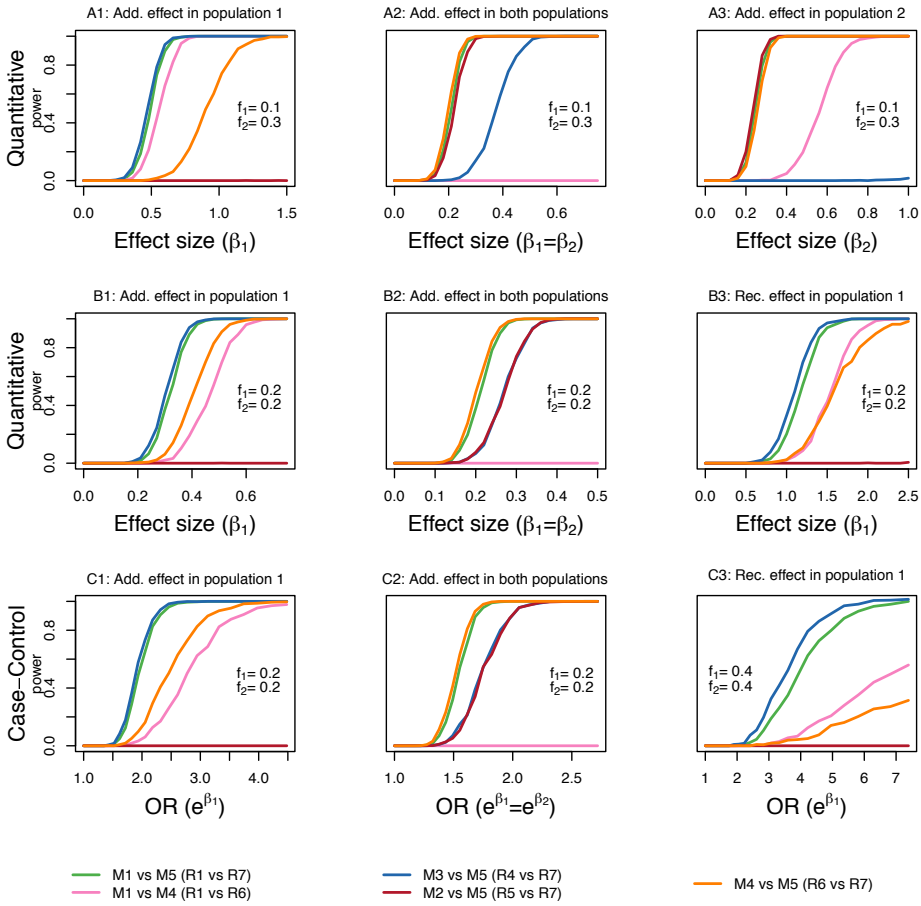

Figure 5: Power assessment results obtained by applying asaMap to data simulated from nine different. The curves show the fraction of simulated p-values that are smaller than 10^{-8}, based on 1000 simulations for each effect size for each scenario. The simulated population specific effects β_1, β_2 are additive, except for scenario B3 and C3. f_1, f_2 denotes the population allele frequencies.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Notably, in this scenario we do not reject that the effect is the same for the two populations (M1 vs M4).

Third, we simulated a scenario (A3) in which the tested variant only has an effect in population 2, where the variant allele occurs with the highest frequency. In this scenario, the statistical power of the test, M1 vs. M4, of whether there is a difference in effect sizes, was the same as for scenario A1, but, unlike in scenario A1 the standard GLM based test was almost as powerful as the remaining tests (scenario A3, Figure 5).

Next, we simulated a scenario (B1) where the tested variant has the same frequency of 20% in both ancestral populations. As was the case for scenario A1, the test with the best statistical power was the test of whether there is an effect in population 1 (M3 vs. M5) (scenario B1, Figure 5). Notably, this test was only slightly more powerful than the full test for effects (M1 vs. M5), but both these test provided remarkable improvements in power compared to the standard GLM based test of effect assuming same effect in both populations (M4 vs. M5).

We then simulated scenario B2, which was 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 expected loss of power of M1 vs. M5 due to the more complicated modeling compared to M4 vs. M5 was very small (scenario B2, Figure 5), 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

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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: the test of an effect in population 1 (R4 vs. R7) was slightly more powerful than the full test (R1 vs. R7) and both represent remarkable improvements compared to the standard GLM based test (R6 vs. R7) (Figure 5.

Power assessment for case-control study data. For case-control study data we simulated from a population of mixed ancestry where the tested variant has a frequency of 20% in both populations. The effect of the allele was simulated as log-additive in the logistic model and either only present in ancestral population 1 (scenario C1, Figure 5) or present in both ancestral populations (scenario C2, Figure 5). 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 GLM based (logistic regression) test for association. In scenario C2, where the effect is present in both ancestral populations, the standard GLM based 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

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

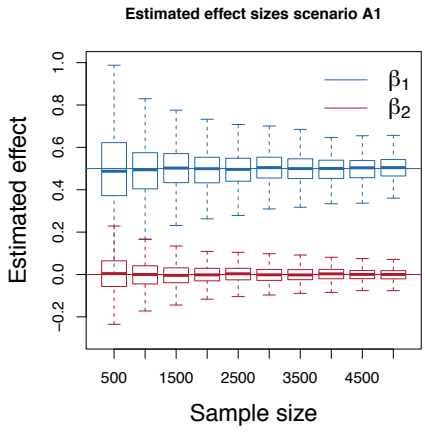

Figure 6: Consistency results obtained by applying asaMap to data from scenario A1 simulated with a fixed effect size of 0.5 in population 1 and 0 in population 2 and with sample sizes ranging from 500 to 5000. Each box corresponds to 1000 simulations.

the effect of the tested variant is recessive and present only in ancestral population 1 (scenario C3, Figure 5). Note that to reach any statistical power for the simulated odds ratios, we here allowed the frequency of the tested allele to be 40% in both ancestral populations. In this scenario, the standard GLM based test of whether 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 ratios (ORs). The tests that 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 ORs to reach full statistical power.

Bias, consistency and false positive rates. Besides using the simulated data for power comparisons, we also used it for assessing asaMap's estimators

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 S1). 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 effects and variants with recessive effects. The corresponding QQ-plots of the p-values achieved show that the false positive rate is indeed controlled for in all the tests (Figure S2).

Finally, 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 and that the decrease in variance with increasing sample size is consistent with the expected $1/n$ relation (Figure 6).

Assessment of potential power gain from knowing the locus states. In the process of simulating ancestry-specific association data for the above power assessment, we explicitly simulated the ancestry-specific allele type

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

combinations (locus states, s), which are not directly observable 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 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.

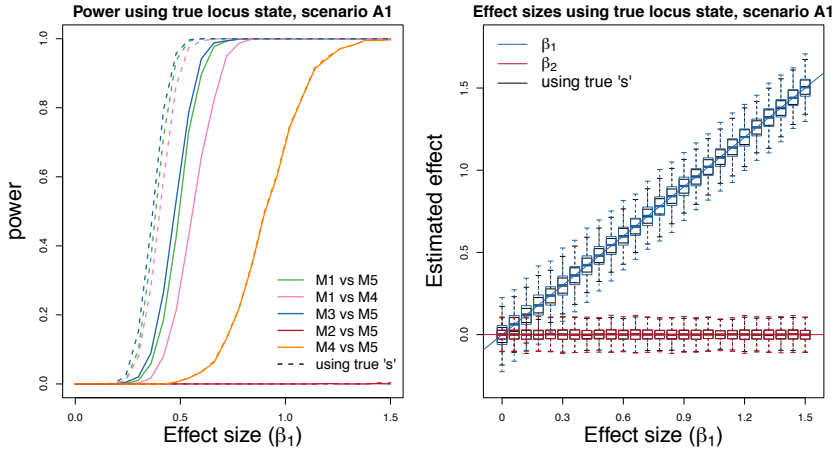

Figure 7: Comparison of power and bias results obtained by applying asaMap with or without known locus state (s) to simulated data from scenario A1. Left: Results of power comparison. The power obtained with the true (simulated) locus states are shown with dashed curves and the power obtained without are shown with solid curves. Right: Results of bias comparison. Box plots of estimates obtained with the true (simulated) locus states are shown in black, whereas estimates obtained without are shown in colors.

Assessment of the test for whether the assumed non-effect allele has an ancestry-specific effect. In the above assessments, we used the simulated effect allele as the assumed effect allele in all cases. To assess

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

how well our proposed tests for whether the assumed non-effect allele has an ancestry-specific effect and thus should be used as the assumed effect allele, we again simulated data from the nine scenarios from Table 3. However, this time we performed the asaMap analyses of the data with the simulated effect allele as the assumed non-effect allele in statistical models M1 (R1). We compared M0 vs. M1 (R0 vs. R1) to see if these tests correctly reject that $\delta_1 = 0$ (equation 4) and that $\delta_1 = \delta_2 = 0$ (equation 7) and thus reveal that the effect allele is misspecified. In most scenarios the tests have high power to correctly reject the choice of effect allele (Figure S3). The only exceptions are the scenarios where the effect sizes of the simulated effect allele is the same in the two populations (scenarios A2, B2 and C2) and where it therefore does not matter which allele is assumed to be the effect allele. To check that the false positive rate of the test for a misspecified effect allele is controlled, we also applied the M0 vs. M1 (R0 vs. R1) test with the simulated true effect allele as the assumed effect allele in the statistical model M1 (R1). Here, by simulation $\delta_1 = 0$ in the statistical model M0 and the corresponding hypothesis should only be rejected at rate α. The check indeed showed that the test has the expected false positive rate (Figure S3).

TBC1D4 gene in a Greenlandic cohort

To further assess asaMap, we also applied it to real data from the Greenlandic population. This population is an Inuit population which is highly admixed: more than 80% of Greenlanders have some recent European ancestry and the

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 high risk of type 2 diabetes and highly elevated 2-hour (2-h) plasma glucose levels. 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 (IHIT) cohort described in (Moltke et al. 2014), we tested the five above mentioned SNPs for ancestry-specific association with 2-h plasma glucose levels as a quantitative trait using a recessive model (for results see 8). For the four non-causal SNPs; rs7330796 (original lead SNP), rs1062087, rs2297206, and rs77685055, we observed a more significant recessive effect of the variant when both alleles are inherited from the Inuit population (R4 vs. R7) compared to the standard GLM based test (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 (R6 vs. R2) showed that the effect of carrying two effect alleles both inherited from the ancestral Inuit population is significantly different from the effect of carrying two effect alleles of which at least one is inherited from the ancestral European population. This suggests that

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

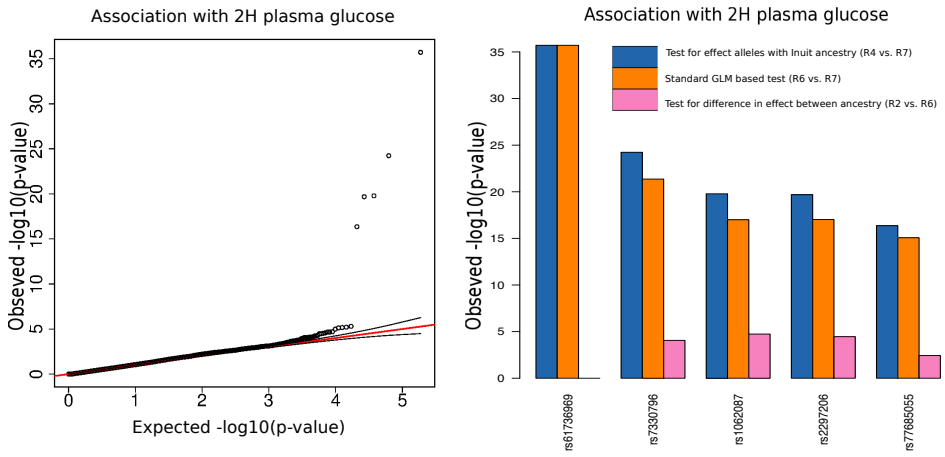

Figure 8: Association results for data from Greenland. Left: QQ-plot for 2-h plasma glucose in the Greenlandic IHIT cohort achieved using the asaMap recessive model for quantitative traits. Specifically, we tested for an effect in the ancestral Inuit population using the test R4 vs. R7. Right: Minus $\log_{10}$ of the p-values from three different statistical tests implemented in asaMap.

these four SNPs are not causal. On 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 is consistent with the conclusion from (Moltke et al. 2014) that it is casual.

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 tests, using the first ten principal components to correct for population structure as described in Materials and Methods.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

In this paper, we have presented asaMap, a flexible statistical testing 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 GLM based tests, which are 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. Notably, we have here demonstrated that when the effect does depend on ancestry, the full test in asaMap, which tests if there is an effect in any population while allowing for ancestry-specific effects (M1 vs. M5) will outperform the commonly used GLM based tests. However, the gain in power depends strongly on the allele frequencies in the ancestral populations. If there is a lower frequency in the population in which the effect is high-

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

est than in the other population the gain in power can be very substantial. Conversely, if the 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 ancestral 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 ancestral populations, 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, the former 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.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 effect alleles assuming there are two ancestral populations. However, we have here demonstrated that when an allele only has a recessive effect when both copies of it 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 ancestral population (R4 vs R7). We expect the same to be true in all cases where the frequency of the effect allele is high in the population where it does not have an effect, because in this case a lot of the individuals carrying two copies of the effect allele will not be affected, causing the standard GLM based 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 ancestral populations. Since we expect the effect of a causal allele to be similar in the two ancestral 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.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Finally, asaMap offers a test for whether the assumed non-effect allele, i.e. the allele that is not assumed to be the effect allele in the default asaMap model, has an ancestry-specific effect. The motivation for this is that asaMap, just like other similar methods (Pasaniuc et al. 2011, Yorgov et al. 2014), focuses its tests on the potentially different effects of a specific assumed effect allele. However, it could in principle be the other allele that has an ancestry specific effect. If this is the case it will not lead to false positives, but it may result in suboptimal power to detect an effect and particularly to detect difference in effects between populations. We have not experienced this as a problem in practice, but since it may occur, offering this additional test seems potentially useful to the users.

To test aseMap on real data, we applied it to genotype data from admixed individuals from Greenland for five SNPs and corresponding levels of 2-h plasma glucose. We demonstrated that the ancestry-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 identified causal loss of function variant being causal. Furthermore, asaMap correctly provided results, which support that the four remaining SNPs have ancestry-specific effects and thus are unlikely to be causal.

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 GLM based tests that are commonly used in GWAS. We have also shown the asaMap tests, at least

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the full test, are almost as powerful as the standard GLM based test in all other investigated 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.

Funding

L.S. and T.S.K. was supported by grants from the Carlsberg Foundation (CF15-0899, CF16-0913), I.M. was supported by two grants from Independent Research Fund Denmark (DFF-11-116283, DFF-4090-00244). E.J. and A.A. was supported by a grant from the Lundbeck Foundation (R215-2015-4174) to A.A.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

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

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,

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.,

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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', *Ge-*

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

netics **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.,

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

Morris, A., Voight, B., Teslovich, T., Ferreira, T., Segre, A., Steinthorsdottir, V., Strawbridge, R., Khan, H., Grallert, H., Mahajan, A. & et al. (2012), 'Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes', *Nat. Genet.* **44**, 981–990.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., Prokunina-Olsson, L., Ding, C.-J., Swift, A. J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X.-Y., Conneely, K. N., Riebow, N. L., Sprau, A. G., Tong, M., White, P. P., Hetrick, K. N., Barnhart, M. W., Bark, C. W., Goldstein, J. L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T. A., Watanabe, R. M., Valle, T. T., Kinnunen, L., Abecasis, G. R., Pugh, E. W., Doheny, K. F., Bergman, R. N., Tuomilehto, J., Collins, F. S. & Boehnke, M. (2007), 'A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants', *Science* **316**(5829), 1341–1345.

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.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The SIGMA Type 2 Diabetes Consortium (2013), 'Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico', *Nature* **506**, 97–101.

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.

Voight B.F., et al. (2010), 'Twelve type 2 diabetes susceptibility loci identified

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

through large-scale association analysis', *Nat. Genet.* **42**, 579–589.

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.

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

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supporting Information

Supplementary figures

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

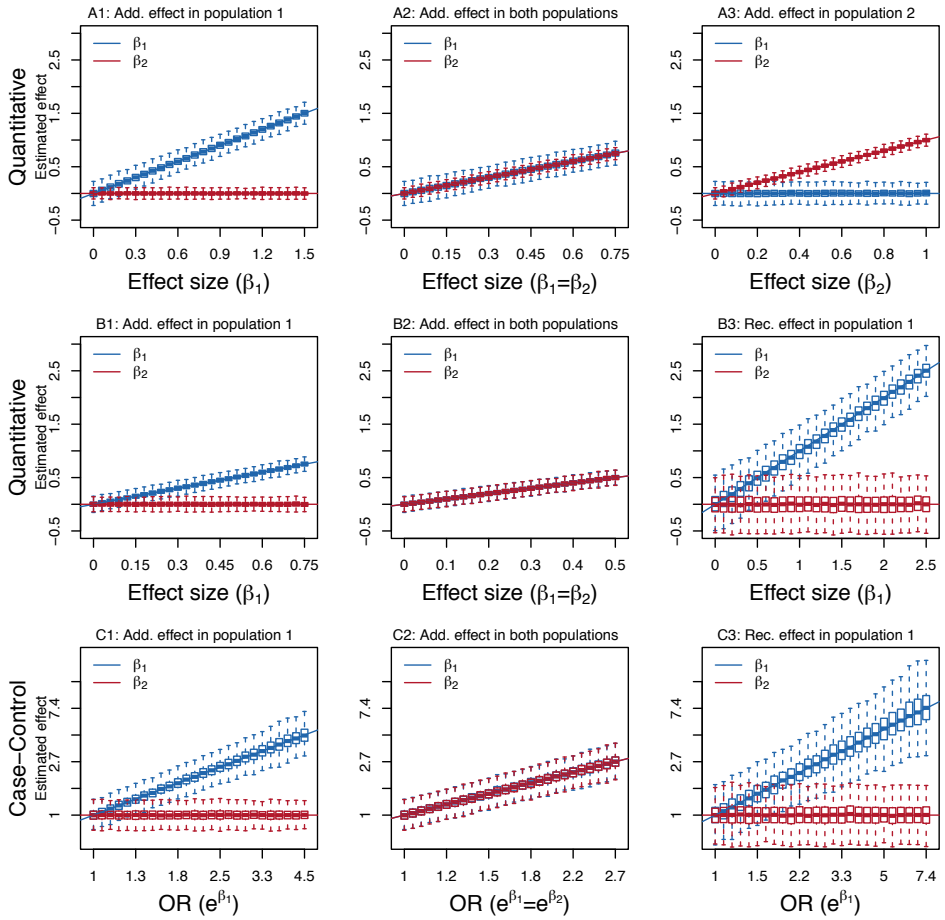

Figure S1: Results of bias simulations for all nine scenarios (A1-C3). Each box show the results of 1000 simulations. The lines show the true effect sizes. β_1 and β_2 denotes the effect sizes in population 1 and 2, respectively. The simulated scenarios are described in Table 3 and the tests are described in Table 2 and Table S2.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

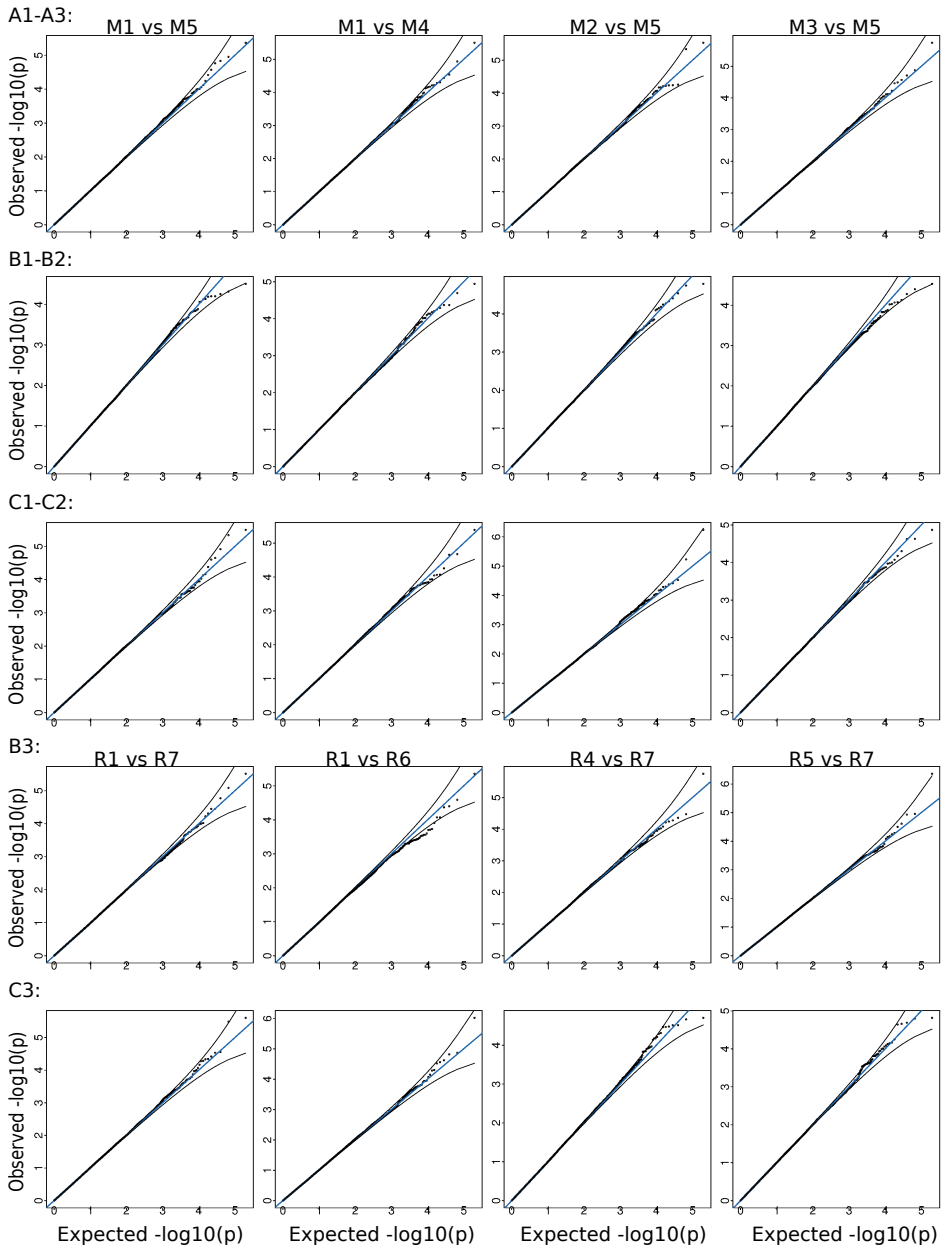

Figure S2: QQ-plots of the p-values obtained by applying asaMap to simulations from the null model of all simulation scenarios (A1-C3). Each QQ-plot is based on 100000 simulations. The simulation scenarios are described in Table 3 and the tests are described in Table 2 and Table S2. The three top rows show results for the asaMap tests that assume an additive genetic effect, while the two bottom rows show results for the asaMap tests that assume a recessive genetic effect.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

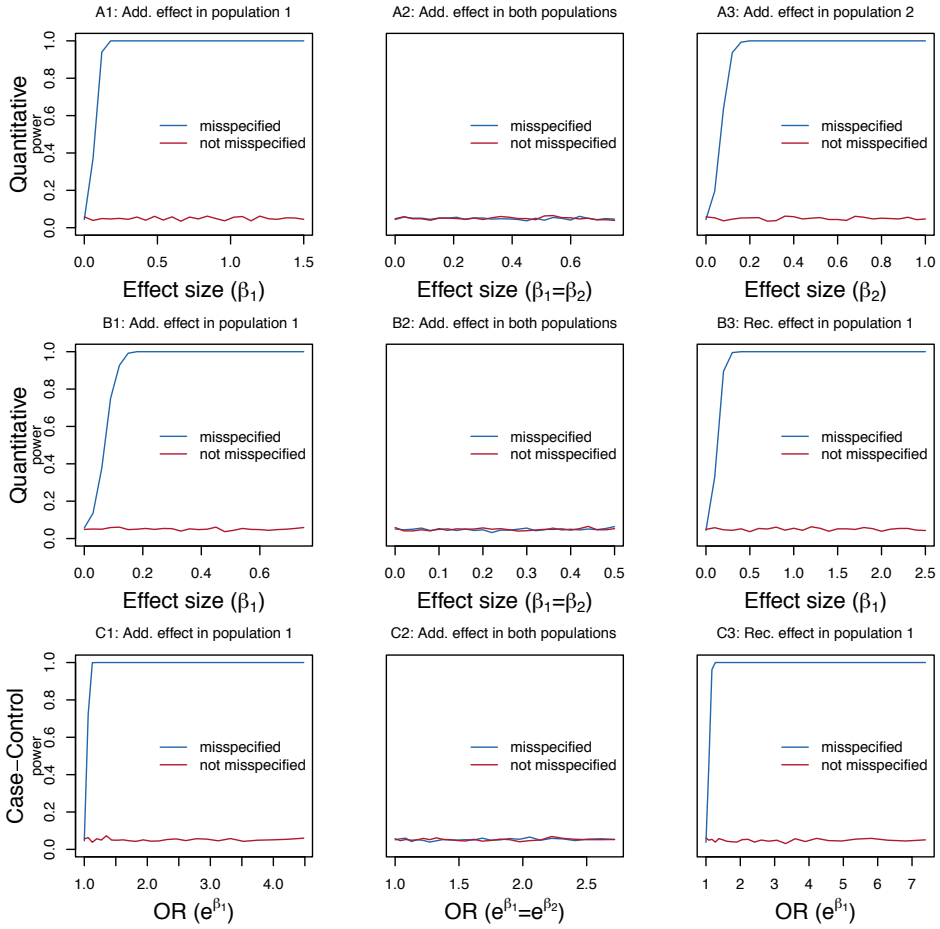

Figure S3: Results for the tests of whether the assumed non-effect allele has an effect applied to simulated data from nine scenarios (A1-C3). The simulation scenarios are described in Table 3. The tests are based on comparing M0 vs. M1 and R0 vs. R1. M0 and R0 are described in equations 4 and 7. All the curves based on 1000 simulations for each effect size for each scenario and show the fraction of the tests that led to p-values that are smaller than 0.05. The curves labeled "misspecified" show the results of applying the tests with the assumed effect allele misspecified. The curves labeled "not misspecified" show the results of applying the tests with the assumed effect allele correctly specified.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary tables

Model	Hypothesis	The model assumes
R1	$(\beta_1, \beta_m, \beta_2) \in R^3$	Ancestry-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	$\beta_1 = \beta_m = \beta_2 = 0$	No effect

Table S1: Recessive ancestry-specific genetic effects models. β_1 is the effect of carrying two copies of the effect allele when both are inherited from population 1, β_2 is the effect carrying two copies of the effect allele when both are inherited from population 2 and β_m is the effect of carrying two copies of the effect allele when one of the copies is inherited from population 1 and the other from population 2.

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: Relevant tests comparing the different nested recessive ancestry-specific genetic effects models described in Table S1.

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary text

Below we describe the models and methods that were not described in detail in the main text.

Case-control study modification

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

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \alpha + \beta_1 x_1(s) + \beta_2 x_2(s) + \sum_c \gamma_c z_c^i. \tag{5}$$

where $x_1(s)$ and $x_2(s)$ are the design matrix entries for the two ancestry-specific effects and z_c^i is the value of covariate c for individual i.

Overview of recessive models and tests

In addition to the models that assume an additive genetic effect, models that assume a recessive effect have also been implemented in asaMap. Their design is slightly more complicated than the additive models. 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. Notably, due to this more complex mod-

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

elling, the recessive genetic model is complicated for more than two ancestral populations.

A recessive model in general assumes that there only is an effect, when an individual carries two copies of the effect allele. The full ancestry-specific recessive model, R1, includes three different effects: β_1 is the effect of carrying two copies of the effect allele when both are inherited from population 1, β_2 is the effect of carrying two copies of the effect allele when both are inherited from population 2 and β_m is the effect of carrying two copies of the effect allele when one of the copies is inherited from population 1 and the other from population 2. The linear predictor for this model is given by

$$\eta_i = \alpha + \beta_1 r_1(s) + \beta_2 r_2(s) + \beta_m r_m(s) + \sum_c \gamma_c z_c^i, \tag{6}$$

where $r_1 = 1$ if the individual carries two copies of the effect allele that are both inherited from population 1 and $r_1 = 0$ otherwise, $r_2 = 1$ if the individual carries two copies of the effect allele that are both inherited from population 2 and $r_2 = 0$ otherwise and where $r_m = 1$ if the individual carries two copies of the assumed effect-allele and has inherited one copy from each of population 1 and 2.

This allows us to fit both R1 and a range of sub-models (see Table S1 for an overview and later text for a more detailed description). The models R4 and R5 are of most interest. In R4 it is assumed that there is no effect of carrying two copies of the effect allele - unless both alleles are inherited

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

from population 1. This model can then be compared against R7 - where it is assumed that there is no effect - to test if there is a significant recessive effect of the effect allele when inherited from population 1. To test if the model assumption for model R4 is appropriate we can compare against R1 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 a recessive effect when both variants are inherited from population 2 and R1 vs. R5 can be used to check the assumptions of this test.

We have also implemented an expanded model, R0, which also models the effects of carrying two copies of the assumed non-effect allele that are both inherited from population 1 or population 2:

$$\eta_i = \alpha + \beta_1 r_1(s) + \beta_2 r_2(s) + \beta_m r_m(s) + \delta_1 ru_1(s) + \delta_2 ru_2(s) + \sum_c \gamma_c z_c^i \quad (7)$$

Here, $r_1, r_2, r_m, \beta_1, \beta_2$ and β_m are defined as above, ru_1, ru_2 are indicators of whether an individual carries two copies of the assumed non-effect, which are both inherited from population 1 or population 2, respectively and δ_1, δ_2 are the corresponding effect sizes. This model allows the user to check if the assumed effect allele is miss-specified. This check is performed by testing model R0 against model R1, to see if we reject the null hypothesis of $\delta_1 = \delta_2$=0. Note that, for this model we assume that δ_{m1} meaning being recessive

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for the assumed non-effect allele but with two different ancestries will not have an effect that δ_1 or δ_2 does not detect.

Detailed equations for the additive models

Below we provide a more detailed description of the different nested models used under the assumption of an additive genetic effect outlined in Equation 2, Equation 4 and Table 1. More specifically, we list the linear predictors for all the models. In this list, the covariates $x_1(s)$ and $x_2(s)$ denote the counts of the assumed effect allele from population 1 and 2, respectively for a locus in locus state s. In addition, α parameterizes the intercept (baseline) and additional covariates enters the model in the z_c^i with effects γ_c. Finally, we use δ_1 to parameterize the effect of the assumed non-effect allele and u_1 is the observed number of assumed non-effect alleles from population 1. With this notations the linear predictors are:

M0:

$$\eta_i = \alpha + \beta_1 x_1(s) + \beta_2 x_2(s) + \delta_1 u_1(s) + \sum_c \gamma_c z_c^i, \tag{8}$$

M1:

$$\eta_i = \alpha + \beta_1 x_1(s) + \beta_2 x_2(s) + \sum_c \gamma_c z_c^i, \tag{9}$$

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

M2:

$$\eta_i = \alpha + \beta_2 x_2(s) + \sum_c \gamma_c z_c^i, \tag{10}$$

M3:

$$\eta_i = \alpha + \beta_1 x_1(s) + \sum_c \gamma_c z_c^i, \tag{11}$$

M4:

$$\eta_i = \alpha + \beta(x_1(s) + x_2(s)) + \sum_c \gamma_c z_c^i, \tag{12}$$

M5:

$$\eta_i = \alpha + \sum_c \gamma_c z_c^i, \tag{13}$$

Detailed equations for the recessive models

Below we provide a more detailed description of the different nested models used under the assumption of a recessive genetic effect outlined in Equation 6, Equation 7 and Table S1. More specifically, we list the linear predictors for all the models.

In this list, r_1 and r_2 indicates if the individual carries two copies of the assumed effect allele and both are inherited from population 1 or 2, respectively. Similarly, r_m indicates if the individual carries two copies of the

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

assumed effect allele and they are inherited from both populations. The corresponding effects are denoted by β_1, β_2 and β_m. In addition, α parameterizes the intercept (baseline) and additional covariates enters the model in the z_c^i with effects γ_c. Finally, ru_1, ru_2 indicates if the individual carries two copies of the assumed non-effect allele and both alleles are inherited from population 1 or 2, respectively and δ_1, δ_2 are the corresponding effects. With this notation the predictors are:

R0:

$$\eta_i = \alpha + \beta_1 r_1(s) + \beta_2 r_2(s) + \beta_m r_m(s) + \delta_1 ru_1(s) + \delta_2 ru_2(s) + \sum_c \gamma_c z_c^i \quad (14)$$

R1:

$$\eta_i = \alpha + \beta_1 r_1(s) + \beta_2 r_2(s) + \beta_m r_m(s) + \sum_c \gamma_c z_c^i \quad (15)$$

R2:

$$\eta_i = \alpha + \beta_1 r_1(s) + \beta(r_2(s) + r_m(s)) + \sum_c \gamma_c z_c^i \quad (16)$$

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

R3:

$$\eta_i = \alpha + \beta(r_1(s) + r_m(s)) + \beta_2 r_2(s) + \sum_c \gamma_c z_c^i \tag{17}$$

R4:

$$\eta_i = \alpha + \beta_1 r_1(s) + \sum_c \gamma_c z_c^i \tag{18}$$

R5:

$$\eta_i = \alpha + \beta_2 r_2(s) + \sum_c \gamma_c z_c^i \tag{19}$$

R6:

$$\eta_i = \alpha + \beta(r_1(s) + r_2(s) + r_m(s)) + \sum_c \gamma_c z_c^i \tag{20}$$

R7:

$$\eta_i = \alpha + \sum_c \gamma_c z_c^i \tag{21}$$

EM algorithm

Below we will go through the details of the EM-algorithm used in asaMap.

Notation. We will use the following notation:

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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, $Q^i = \{q_1^i, q_2^i\}$.

f $= \{f_1, f_2\}$ population specific allele frequencies

ϕ vector of effect sizes e.g. (β_1, β_2) and other regression parameters e.g. (α, γ, σ)

s locus state, $s = \{a, t\}$ where $a = (a_1, a_2)$ is the ordered information on ancestry and $t = (t_1, t_2)$ is the ordered allelic genotype.

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

$$p(Y|G, Q, f, \phi) = \prod_i p(y^i|g^i, Q, f, \phi) \tag{22}$$

and splitting the probabilities according to locus state (see Figure 4) gives

$$p(y_i|g_i, Q, f, \phi) = \sum_s p(y_i, s|g_i, Q, f, \phi) \tag{23}$$

$$= \sum_s p(y_i|s, \phi)p(s|g_i, f, Q) \tag{24}$$

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

$$p(y_i|s,\phi) \sim N(\eta^i(s,\alpha,\beta,\gamma),\sigma^2) \tag{25}$$

with

$$\eta_i(s,\alpha,\beta,\gamma) = \alpha + \beta_1 x_1(s) + \beta_2 x_2(s) + \sum_c \gamma_c z_c^i \tag{26}$$

where we use that for population k $x_k(s) = t_1 1_{a_1=k} + t_2 1_{a_2=k}$ 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

$$\begin{aligned} f(y|\eta,\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(y-\eta)^2}{2\sigma^2} \\ &= \exp\left(\frac{y\eta - \eta^2/2}{\sigma^2} - y^2/2\sigma^2 - \log(2\pi\sigma^2)/2\right) \\ &= \exp\left(\frac{y\eta - b(\eta)}{a(\sigma)} + c(y,\sigma)\right) \end{aligned} \tag{27}$$

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:

$$E_{S|Y,G,\phi^*,Q,f}[\log p(Y,S|G,\phi,Q,f)] = \sum_i E_{s_i|y_i,G,\phi^*,Q,f}[\log p(y_i,s_i|G,\phi,Q,f)] \tag{28}$$

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

as a function of all regression parameters ϕ, where ϕ^* is fixed to the parameter values 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 the parameters to be optimized, this is equivalent to maximizing

$$\sum_i E_{s_i|y_i,G,\phi^*,Q,f}[\log p(y_i|s_i, \phi)] = E_{S|Y,G,\phi^*,Q,f}[\log p(Y|S, \phi)] \tag{29}$$

Following (Lake et al. 2003) we take the derivative with respect to the vector of ancestry-specific effect sizes and get:

$$\begin{aligned}
\frac{\partial}{\partial\beta} E_{s_i|y_i,G,\phi^*,Q,P}[\log p(y_i|s_i, \phi)] &= E_{s_i|y_i,G,\phi^*,Q,f}[\frac{\partial}{\partial\beta} \log p(y_i|s_i, \phi)] \\
&= E_{s_i|y_i,G,\phi^*,Q,f}[\frac{\partial\eta_i}{\partial\beta}\frac{\partial}{\partial\eta_i} \log p(y_i|s_i, \phi)] \\
&= E_{s_i|y_i,G,\phi^*,Q,f}[x^i \frac{y_i - b'(\eta_i)}{a(\sigma)}] \\
&= \sum_{s_i} x^i \frac{y_i - \eta_i}{\sigma^2} p(s_i|y_i, G, \phi^*, Q, f)
\end{aligned} \tag{30}$$

where $x^i = (x_1(s_i), x_2(s_i))$ 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

$$\frac{\partial}{\partial\phi} E_{s|y,G,\phi^*,Q,P}[\log p(y|s, \phi)] = \sum_i \sum_{s_i} (x^i, z^i) \frac{y^i - \eta^i(\phi)}{\sigma^2} p(s_i|y_i, G, \phi^*, Q, f) \tag{31}$$

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

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 ϕ 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

$$p(s|y_i, G, \phi^*, Q, f) = \frac{p(y_i|s, \phi^*)p(s|g_i, Q, f)}{\sum_{s'} p(y_i|s, \phi^*)p(s'|g_i, Q, f)} \tag{32}$$

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 distributed trait. First a rough initial guess of the standard deviation is calculated by

$$\sigma = \mathrm{sd}[y] \tag{33}$$

bioRxiv preprint doi: https://doi.org/10.1101/014001; this version posted December 6, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and randomly chosen start values for the regression parameters are sampled from a uniform distribution e.g.

$$\alpha, \beta, \gamma \sim \text{unif}(-1, 1) \tag{34}$$

Then regression weights are calculated according to (32) and a weighted regression according to the score function in (31) 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 the number of individuals and p is the number of effect parameters in the linear predictor e.g. $(p = 3 + C)$ for M1 model with C being the number of γ parameters.