Efficient approaches for large-scale GWAS with genotype uncertainty

Jørsboe, Emil; Albrechtsen, Anders

Published in:
G3: Genes, Genomes, Genetics (Bethesda)

DOI:
10.1093/g3journal/jkab385

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Jørsboe, E., & Albrechtsen, A. (2022). Efficient approaches for large-scale GWAS with genotype uncertainty. G3: Genes, Genomes, Genetics (Bethesda), 12(1), [jkab385]. https://doi.org/10.1093/g3journal/jkab385
Efficient approaches for large-scale GWAS with genotype uncertainty

Emil Joersbo1,2,* and Anders Albrechtsen1

1Department of Biology, The Bioinformatics Centre, University of Copenhagen, 2200 Copenhagen N, Denmark, and
2Faculty of Health and Medical Sciences, Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, 2200 Copenhagen N, Denmark

*Corresponding author: Email: emil.joersbo@sund.ku.dk

Abstract
Association studies using genetic data from SNP-chip-based imputation or low-depth sequencing data provide a cost-efficient design for large-scale association studies. We explore methods for performing association studies applicable to such genetic data and investigate how using different priors when estimating genotype probabilities affects the association results. Our proposed method, ANGSD-asso's latent model, models the unobserved genotype as a latent variable in a generalized linear model framework. The software is implemented in C/C++ and can be run multi-threaded. ANGSD-asso is based on genotype probabilities, which can be estimated using either the sample allele frequency or the individual allele frequencies as a prior. We explore through simulations how genotype probability-based methods compare with using genetic dosages. Our simulations show that in a structured population using the individual allele frequency prior has better power than the sample allele frequency. In scenarios with sequencing depth and phenotype correlation ANGSD-asso's latent model has higher statistical power and less bias than using dosages. Adding additional covariates to the linear model of ANGSD-asso's latent model has higher statistical power and less bias than other methods that accommodate genotype uncertainty, while also being much faster. This is shown with imputed data from UK Biobank and simulations.

Keywords: admixture; association mapping; case-control study; next-generation sequencing; quantitative traits

Introduction
Genome-wide association studies (GWASs) have classically been done to study genotype-phenotype associations. However, a slightly different design of GWAS is using low depth next-generation sequencing (NGS) data, because in such cases the genotype cannot be inferred accurately. Low depth sequencing provides a cost-efficient design, where the number of individuals studied can be increased many folds compared with high depth sequencing, since each individual will be a lot cheaper to sequence. The statistical power to detect associations increases with the number of individuals while only dropping slightly with lower depth and therefore this design provides good statistical power to detect associations (Pasaniuc et al. 2012). Another approach that is commonly used in GWAS is performing haplotype imputation from SNP-chips to infer missing genotypes which also generates genetic data with uncertainty on the inferred genotype.

A recent successful GWAS (Liu et al. 2018) with low depth NGS data has shown the viability of this approach. In Liu et al. (2018), around 140,000 individuals had a noninvasive prenatal test for fetal trisomy with low depth sequencing with an average sequencing depth of 0.1X. Haplotype imputation was performed on the low depth NGS data. For the association testing a score test approach using a linear model framework (Skotte et al. 2012) implemented in ANGSD (Korneliussen et al. 2014) was used, as this method takes genotype uncertainty into account. Despite the low sequencing depth several novel associations were discovered. This provides an example of a study where using methods that account for the genotype uncertainty in low depth NGS data, provides good statistical power for detecting associations.

In association studies, genotype uncertainty can be taken into account using a latent variable model that sums over the possible genotype states. Using latent variable models can have advantages compared with calling genotypes for low depth NGS data (Skotte et al. 2012). Skotte et al. (2012) implemented a score test where the coefficients are not estimated under the alternative hypothesis making the method computationally very fast; however, this means the effect size of the genotype is not estimated. In this article, we will introduce ANGSD-asso's latent model, it works in a generalized linear model (GLM) framework; it estimates the effect size of the unobserved genotype. This method can in practice be run almost as fast as the score test. ANGSD-asso's latent model uses a maximum likelihood approach; more specifically, we will make use of the EM algorithm to maximize the likelihood, treating the unobserved genotype G as a latent variable. Using a GLM framework enables us to include covariates thereby adjusting for possible confounders, such as population structure. We have implemented an EM algorithm that converges fast and that can be run multi-threaded, making the analysis of large data sets possible. We have also implemented a hybrid
model in ANGSD-assos’s latent model

We model the data using a maximum likelihood approach in a GLM framework. This enables us to test for an association without observing the genotype G directly. Rather, we observe our NGS data \( (x) \) from which we can calculate the genotype probability \( p(G(x)) \). We write the likelihood for our phenotype data \( (y) \) given our sequencing data \( (x) \) and covariates \( (z) \):

\[
p(y|x, z) = \prod_{i=1}^{N} p(y_i|x_i, z) = \prod_{i=1}^{N} \sum_{g(x)=g, z} p(y_i| G_i = g, z)p(G_i = g|x_i),
\]

where we use the law of total probabilities to introduce the genotype as a latent variable \( G \). \( N \) is the number of individuals, \( y = (y_1, y_2, \ldots, y_N) \) is a vector of our observed phenotype, \( x = (x_1, x_2, \ldots, x_N) \) is a vector of sequencing data and \( Z = (z_1, z_2, \ldots, z_N) \) is a \( N \times c \) matrix with the additional covariates. We see that the trait \( y \) is conditionally independent of the sequencing data given the genotype \( (meaning p(y|G_i = g, x_i, z_i) = p(y|G_i = g, z_i)) \). We can calculate the genotype probability from the sequence data, for example, by using the sample allele frequency as a prior, by assuming that the genotype is conditionally independent of the covariates, given the sequencing data and the frequency \( f \) (meaning \( p(G_i = g, x_i, f) = p(G_i = g, f) \)); however, for simplicity we omit \( f \) from the likelihood.

**Methods**

NGS produces short reads that are mapped to a reference genome. From the aligned reads the probability of observing these reads at a given site for a certain genotype can be calculated. This is known as the genotype likelihood (Nielsen et al. 2011). For more on the genotype likelihood and how to calculate it see the Supplementary Section S5.2. The genotype likelihoods together with a genotype prior can be used to calculate the probability of the genotype given the data which is referred to as the genotype probability. For an overview of the relationship between the different kinds of genetic data, and how they can be processed and analyzed in association studies see Figure 1.
This allows us to write the likelihood, also introducing the parameters of our GLM \( \theta = (x, \beta, \gamma) \), again we assume that the genotype is conditionally independent of the covariates given the sequencing data:

\[
L(\theta) \propto p(y|x, Z, \theta) = \prod_{i=1}^{N} p(y_i|x_i, z_i, \theta)
\]

\[
= \prod_{i=1}^{N} \sum_{g \in \{0,1,2\}} p(y_i|G_i = g, z_i, \theta)p(G_i = g|x_i).
\]

And the log-likelihood then becomes

\[
\log \left( \sum_{g \in \{0,1,2\}} p(y_i|G_i = g, z_i, \theta)p(G_i = g|x_i) \right).
\]

Equation (4) is the log-likelihood function that we want to maximize with regard to the parameters \( \theta \). For maximization, we use the EM algorithm where our latent variable is the unobserved genotype \( G \) and the variance among the \( G \) is distributed according to Equation (3) (Supplementary Methods). We therefore introduced sequencing depth and phenotype correlation in our simulations. Essentially, this means that having a large phenotype value or being a case will make it more likely that an individual is in the high depth group. This is described in more detail in Supplementary Sections S1.3 and S4.

**Simulation of phenotypes**

We simulate the phenotypes under a standard additive model with a normally distributed phenotype. The mean given is given by effect \( \beta \) from the genotype \( g \) and the ancestry/admixture proportions \( q \) with effect \( \gamma \) from being population 1 and SD 1. We simulate the phenotypes according to Equation (5). The mean for each simulation can also be seen in the last column of Table 1. We simulate different scenarios with and without the admixture proportions \( q \) and the ancestry correlation in our simulations. Essentially, this means that having a large phenotype value or being a case will make it more likely that an individual is in the high depth group. This is described in more detail in Supplementary Sections S1.3 and S4.

**Results and discussion**

We wanted to assess which prior performs best for generating the genotype probabilities when performing association studies with low depth sequencing data. We simulate different scenarios with and without population structure and with and without sequencing depth and phenotype correlation. For each of these scenarios, we both applied a sample allele frequency prior and an individual allele frequency prior and evaluated their performance...
Table 1: Overview of simulations

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Allele frequency</th>
<th>N</th>
<th>Population structure</th>
<th>Sequencing depth and phenotype</th>
<th>Simulated phenotype mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.45</td>
<td>1,000</td>
<td>No</td>
<td>Correlated</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.45</td>
<td>1,000</td>
<td>No</td>
<td>Not correlated</td>
<td>$\beta \cdot g$</td>
</tr>
<tr>
<td>3</td>
<td>(0.9, 0.1)</td>
<td>1,000</td>
<td>Yes</td>
<td>Correlated</td>
<td>$q \cdot \gamma$</td>
</tr>
<tr>
<td>4</td>
<td>(0.9, 0.1)</td>
<td>1,000</td>
<td>Yes</td>
<td>Not correlated</td>
<td>$\beta \cdot g + q \cdot \gamma$</td>
</tr>
<tr>
<td>5</td>
<td>0.45</td>
<td>1,000</td>
<td>No</td>
<td>Correlated</td>
<td>$\beta \cdot g$</td>
</tr>
<tr>
<td>6</td>
<td>(0.9, 0.1)</td>
<td>1,000</td>
<td>Yes</td>
<td>Correlated</td>
<td>$\beta \cdot g + q \cdot \gamma$</td>
</tr>
</tbody>
</table>

We simulate under a standard additive model with a normally distributed phenotype with a mean given in the last column, and SD 1, according to Equation (5). $N$ is the number of individuals, $g$ is the genotype with effect size $\beta$ and $q$ is the ancestry proportion with effect size $\gamma$. In scenarios 1 and 3, there is no effect of the genotype. In scenarios 4, 5, and 6, there is population structure with two ancestral populations with allele frequency of 0.9 and 0.1. The sequencing depth and phenotype correlation are simulated using (Equation 5, Supplementary Methods).

Evaluation of using different priors

Using different priors in a homogeneous population

In scenario 1, we simulate data without genotype effect, and population structure but with sequencing depth and phenotype correlation. We observe no inflation of the false positive rate for both priors (Supplementary Figure S1). This is expected since these priors become identical in the absence of population structure. In scenario 2, we add a genotype effect and again we observe no difference in the two priors in terms of statistical power (Supplementary Figure S2).

Using different priors in a structured population

In scenario 3, we simulate population structure and sequencing depth and phenotype correlation. As shown in Figure 2, using the sample allele frequency prior gives biased estimates of the effect size and leads to an increased false positive rate. The increased false positive rate is present even though we are adjusting for ancestry in the linear model, showing that this is not sufficient in this scenario. When using the individual allele frequency prior, we do not get biased estimates and have a false positive rate that is identical to using the true genotype. In scenario 4, we remove correlation between sequencing depth and phenotype and use a range of effect sizes for the genotype. Figure 3 shows that using the individual allele frequency prior has increased statistical power compared with the sample allele frequency prior. For example for an effect size of $\beta = 0.3$ the power is 0.74 compared with 0.61. Furthermore, in this scenario, we see that using dosages has similar statistical power to using genotype probabilities. When using the sample allele frequency prior the effect sizes are underestimated. This is due to the fact that using the individual allele frequency better describes the expected genotype in a structured population.

In scenario 5, we remove population structure but reintroduce correlation between sequencing depth and phenotype and use a range of effect sizes for the genotype. Supplementary Figures S3b and S4b show that in this scenario there is slightly increased statistical power when using the full genotype probabilities compared with using dosages, and that our estimated effect size is less biased. In scenario 6, we have correlation between sequencing depth and phenotype, and use a range of effect sizes for the genotype. Supplementary Figures S5b and S6b show that in this scenario we observe little reduction in power when using the full genotype probabilities compared with using dosages, and that our estimated effect size is slightly increased.

To evaluate the effect of estimating the admixture proportions from a limited number of genetic sites, we ran simulations with a varying number of sites used for estimating the admixture proportions. We observe a bias in effect size and a reduction in power when a small number (50 or 500) of sites is used. However, when the number of sites is increased (5000 or 50,000), there is very little reduction in power or bias (Supplementary Tables S3 and S4) compared with using the true admixture proportions.

Comparison with dosages in large-scale studies

To further explore the performance difference with dosages, we simulated a large case-control study with 100,000 individuals. All individuals have low depth sequencing data but the cases and controls have different average sequencing depths.

In Table 2, we show that using full genotype probabilities has increased statistical power compared with using genotype dosages. We have more power for small effect sizes, where we have a true positive rate that is almost 0.1 higher. We calculated the info measure for our dosages in cases and controls respectively, to make it comparable with haplotype imputation. To calculate the info measure, we used the ratio of observed variance to the expected binomial variance at Hardy-Weinberg equilibrium, as used in the imputation software MACH (Scott et al. 2007). When genotypes are predicted with high certainty the info measure will be close to 1. We see that the info measure is lower in controls, where we have a lower average sequencing depth. Supplementary Figures S3, a and b and S4, a and b show that in scenario 5, with a quantitative trait, there is also increased statistical power and less bias when using full genotype probabilities. It also shows that in these simulations when keeping individuals with 0 reads, there is less bias but similar statistical power. In Table 3, we run the analysis from Table 2, but including individuals with 0 reads. In this scenario, the difference between using dosages and genotype probabilities has been almost erased. However, it is worth noticing that in this scenario, we lose statistical power compared with when we remove individuals with 0 reads. To further investigate these scenarios, we looked at the bias of the estimated effect sizes. Supplementary Figures S9 and S10 show for a relative risk (RR) of 1.14 that the latent model gives less biased estimates of the effect sizes compared with dosages.
Supplementary Figures S5, a and b and S6, a and b, show that in scenario 6, with a quantitative trait there is increased statistical power and less bias, when using genotype probabilities compared with dosages. Last, if we do scenario 6, but for a binary phenotype, we show increased statistical power and a smaller bias of the effect size, when using genotype probabilities compared with dosages as shown in Supplementary Figures S7, a and b and S8, a and b. And that in these simulations removing individuals with 0 reads, leads to increased statistical power and less bias.

Comparison with SNPTEST

UK Biobank data

SNPTEST (Marchini et al. 2007) also implements a latent model for performing association (using the option -method em) with genotype probabilities also using a GLM framework. We applied both methods to the imputed data of UK Biobank (Bycroft et al. 2018); more specifically, we chose a 50 kb region of chromosome 2 (219,675–219,725 kb), which has the genetic variant rs78058190. This variant has been found to be associated with waist–hip ratio in Kichaev et al. (2019) and the association is even stronger when adjusted for body mass index (BMI) (Pulit et al. 2019). It is an imputed variant in the UK Biobank data (info/R² = 0.778797, minor allele frequency (MAF) = 0.049; Bycroft et al. 2018). We therefore ran both methods on the genotype probabilities from the imputation for this region, adjusting for gender, age, BMI, and top 10 genetic PCs (from UK Biobank), with waist–hip ratio as the phenotype but inverse quantile transformed to a standard normal distribution.

ANGSD-asso’s latent model runs this analysis of 292,432 individuals and 1647 genetic variants in 82.13 min (17.71 min with 20 threads), SNPTEST runs this in 175.64 min (no multi-threaded option). The estimated effect sizes from each method are compared in Figure 4, showing that the estimated effect sizes are very similar. However, SNPTEST adjusts for covariates by first regressing out the covariates (plus an intercept) on the phenotype doing ordinary least squares. SNPTEST then subsequently uses the EM algorithm to obtain estimates of the intercept and genotype effect. This approach can lead to bias in the estimated effect size of the tested variant when adjusting with covariates that are both correlated with the phenotype and the tested variant (Freckleton 2002; Vansteelandt et al., 2009). For example, when performing conditional analysis on other genetic variants, in order to determine if the tested variant might be the casual variant. We show using the UK Biobank data (Bycroft et al. 2018), that SNPTEST estimates lower effect sizes
We have performed 10,000 simulations for each effect size. The causal allele of the genetic variant has a frequency of 0.05 and the disease has a prevalence of 0.10 in the population. The phenotype is simulated as a binary trait. We have performed 10,000

<table>
<thead>
<tr>
<th>RR</th>
<th>RR = 1</th>
<th>RR = 1.1</th>
<th>RR = 1.12</th>
<th>RR = 1.14</th>
<th>RR = 1.16</th>
</tr>
</thead>
<tbody>
<tr>
<td>True genotype</td>
<td>0.0813</td>
<td>0.974</td>
<td>0.999</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Dosage</td>
<td>0.0914</td>
<td>0.262</td>
<td>0.523</td>
<td>0.772</td>
<td></td>
</tr>
<tr>
<td>Genotype probabilities</td>
<td>0.0974</td>
<td>0.273</td>
<td>0.538</td>
<td>0.783</td>
<td></td>
</tr>
<tr>
<td>R² cases/controls</td>
<td>0.91/</td>
<td>0.90/</td>
<td>0.90/</td>
<td>0.90/</td>
<td>0.84</td>
</tr>
<tr>
<td>0.77</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td></td>
</tr>
</tbody>
</table>

The phenotype is simulated as a binary trait. We have performed 10,000 simulations for each effect size. The causal allele of the genetic variant has a frequency of 0.05 and the disease has a prevalence of 0.10 in the population. We use a significance threshold of 10⁻⁵. The linear model is adjusted for ancestry. Each point is based on 10,000 simulations. (A) Statistical power to detect an association, using ANGSD-asso’s latent model and dosage model respectively with a sample frequency prior (f) and an individual allele frequency prior (π). (B) Bias of the estimated effect sizes.

Figure 3 Simulation scenario 4 with varying genotype effect size (β). We have a structured population with the same admixture proportions and mean sequencing depth as in Figure 2C. There is an effect of ancestry of population 1 (γ = 1). We use a significance threshold of 10⁻⁵. The linear model is adjusted for ancestry. Each point is based on 10,000 simulations. (A) Statistical power to detect an association, using ANGSD-asso’s latent model and dosage model respectively with a sample frequency prior (f) and an individual allele frequency prior (π). (B) Bias of the estimated effect sizes.

Table 2 The statistical power for different simulated effect sizes or RRs of the genotype

<table>
<thead>
<tr>
<th>RR</th>
<th>RR = 1</th>
<th>RR = 1.1</th>
<th>RR = 1.12</th>
<th>RR = 1.14</th>
<th>RR = 1.16</th>
</tr>
</thead>
<tbody>
<tr>
<td>True genotype</td>
<td>0.0813</td>
<td>0.974</td>
<td>0.999</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Dosage</td>
<td>0.0914</td>
<td>0.262</td>
<td>0.523</td>
<td>0.772</td>
<td></td>
</tr>
<tr>
<td>Genotype probabilities</td>
<td>0.0974</td>
<td>0.273</td>
<td>0.538</td>
<td>0.783</td>
<td></td>
</tr>
<tr>
<td>R² cases/controls</td>
<td>0.91/</td>
<td>0.90/</td>
<td>0.90/</td>
<td>0.90/</td>
<td>0.84</td>
</tr>
<tr>
<td>0.77</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td></td>
</tr>
</tbody>
</table>

The phenotype is simulated as a binary trait. We have performed 10,000 simulations for each effect size. The causal allele of the genetic variant has a frequency of 0.05 and the disease has a prevalence of 0.10 in the population. We use a significance threshold of 10⁻⁵. The linear model is adjusted for ancestry. Each point is based on 10,000 simulations. (A) Statistical power to detect an association, using ANGSD-asso’s latent model and dosage model respectively with a sample frequency prior (f) and an individual allele frequency prior (π). (B) Bias of the estimated effect sizes.

Table 3 This table is similar to Table 2; it is the same scenario also, but where we include individuals with 0 reads.

<table>
<thead>
<tr>
<th>RR</th>
<th>RR = 1</th>
<th>RR = 1.1</th>
<th>RR = 1.12</th>
<th>RR = 1.14</th>
<th>RR = 1.16</th>
</tr>
</thead>
<tbody>
<tr>
<td>True genotype</td>
<td>0.0813</td>
<td>0.974</td>
<td>0.999</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Dosage</td>
<td>0.0914</td>
<td>0.262</td>
<td>0.523</td>
<td>0.772</td>
<td></td>
</tr>
<tr>
<td>Genotype probabilities</td>
<td>0.0974</td>
<td>0.273</td>
<td>0.538</td>
<td>0.783</td>
<td></td>
</tr>
<tr>
<td>R² cases/controls</td>
<td>0.91/</td>
<td>0.90/</td>
<td>0.90/</td>
<td>0.90/</td>
<td>0.84</td>
</tr>
<tr>
<td>0.77</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td></td>
</tr>
</tbody>
</table>

SynpTest can also be shown to be biased in simulations. We simulated a scenario with population structure that is both correlated with the phenotype and the genotype. Supplementary Figure S12 shows that in this scenario SNPTest’s effect sizes are downward biased, whereas ANGSD-asso’s latent model has no bias, when using the individual allele frequency prior, and it also has increased statistical power compared with SNPTEST. We have used the most recent version of SNPTEST (v2.5.4-beta3). We ran all analyses with SNPTEST on the same data as ANGSD-asso, meaning disabling the option of transforming covariates and phenotype in SNPTEST.

As shown in Supplementary Figure S11 priming ANGSD-asso’s EM algorithm with the coefficients from regression on dosages drastically reduces the number of iterations needed for convergence of the EM algorithm.

We also compared ANGSD-asso’s latent model to SNPTEST in terms of computational speed and found that ANGSD-asso’s latent model is faster than SNPTEST, especially for binary data, as shown in Figure 5, and for quantitative data as shown in Supplementary Figure S13. ANGSD-asso’s latent model is capable of analyzing data sets of 100,000 individuals in <10h. Our hybrid approach can handle the analyses in <17 h unthreaded. SNPTEST will take days to run the largest data set, when running a logistic model.

Implementation of model

We have implemented ANGSD-asso’s latent model for taking genotype uncertainty into account when performing association studies. The advantage of this approach compared with the score test (Skotte et al. 2012), is that the effect size of the unobserved genotype is estimated. The effect size helps provide further insights into the relationship between genotype and phenotype. Furthermore, the estimated effect sizes also mean, we can make use of LD-score regression (Bulik-Sullivan et al. 2015). It is shown using UK Biobank data and simulations that ANGSD-asso’s latent model has increased statistical power and less bias compared with SNPTEST, when including covariates that are correlated with both phenotype and genotype in the model (Table 4 and Supplementary Figure S12). This is due to SNPTEST adjusting for
Biobank data that has been inverse quantile transformed to a standard normal distribution. These analyses have been adjusted for top 10 genetics PCs as provided by the UK Biobank data, age, gender, and BMI, the analyzed trait is waist–hip ratio from the UK Biobank data that has been inverse quantile transformed to a standard normal distribution.

**Table 4** Waist–hip ratio association of rs78058190 in UK biobank

<table>
<thead>
<tr>
<th>Covariates</th>
<th>SNPTEST P</th>
<th>SNPTEST β</th>
<th>ANGSD P</th>
<th>ANGSD β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, sex, BMI, and PC 1–10</td>
<td>8.02 × 10⁻¹²</td>
<td>0.030</td>
<td>7.62 × 10⁻¹²</td>
<td>0.030</td>
</tr>
<tr>
<td>Age, sex, BMI, and PC 1–10</td>
<td>0.00086</td>
<td>0.015</td>
<td>1.04 × 10⁻⁶</td>
<td>0.031427</td>
</tr>
<tr>
<td>rs113414093 (R² 0.58)</td>
<td>7.39 × 10⁻⁶</td>
<td>0.020</td>
<td>1.3 × 10⁻⁶</td>
<td>0.024</td>
</tr>
<tr>
<td>Age, sex, BMI, and PC 1–10</td>
<td>1.21 × 10⁻⁵</td>
<td>0.019</td>
<td>7.14 × 10⁻⁷</td>
<td>0.024</td>
</tr>
<tr>
<td>rs113414093 (R² 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs116204487</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs148358468</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This table shows the P-value and estimated effect size (β), for the association of rs78058190 with waist–hip ratio (inverse quantile transformed to a standard normal distribution) when conditioning on genetic variants in linkage disequilibrium LD with rs78058190. The R² values are shown in the table and are based on the LDproxy tool that is part of LDlink (Machiela and Chanock 2015). Results for rs113414093, rs116204487, and rs148358468 conditional on rs78058190 can be found in Supplementary Table S1 and S2.

Covariates by first regressing them on the phenotype and then running the EM algorithm on the residuals. Including covariates in the linear model is a common way to deal with confounders in association studies. Also, it is shown that ANGSD-asso’s latent model is much faster than SNPTEST.

We have chosen to compare our method to SNPTEST as it is the only other method commonly used that takes genotype uncertainty into account. Additionally, almost all GWAS that use dosages are based on a standard GLM which will give identical results regardless of the method used. We therefore felt, it was not necessary to compare the dosage method with other software implementations. There are some exceptions to using standard GLMs. One of the exceptions is the use of dosages in linear mixed models which we have not explored in this study.

**Different priors in structured and homogeneous populations**

We have shown how using the individual allele frequency prior, when estimating genotype probabilities, gives better statistical power to detect an association, when dealing with NGS data with population structure as shown in Figure 3. Also, it removes issues with an increased false positive rates when there is sequencing depth phenotype correlation as shown in Figure 2. This correlation might arise if the sequencing is not randomized, for
example, if cases and controls are being sequenced at different times or places thereby creating a systematic bias, or if different cohorts have been sequenced at different places, or if the trait of interest is much more prevalent at one place vs. another. The scenarios from Table 1 are most likely to arise when dealing with nonmodel species where imputation cannot be done. This leads us to recommend using the individual allele frequency prior when performing association studies with NGS data in structured populations, where imputation is not possible. Whether it is better to base the individual allele frequency prior from clustering using NGSadmix (Skotte et al. 2013) or using PCA using PCAngsd (Meisner and Albrechtsen 2018) depends on the individuals in the data. If the sampled individuals are recent descendants of fairly discrete populations, such as most African Americans, then a clustering approach will give the most accurate allele frequencies while if the structure is more continuous such as many Latino Americans, then PCA might be a better approach.

Comparison with dosages in large-scale studies

In Tables 2 and 3, we show through simulations increased statistical power when using genotype probabilities compared with dosages, with a larger gain in power for the scenario from Table 2. Since individuals with 0 reads are also removed from the true genotype in Table 2, there is higher statistical power for the true genotype in Table 3 as there are more individuals. In both instances, it is a case-control study with low depth sequencing data, where cases and controls have different average sequencing depths. A scenario like this, where there is better genotype information for some individuals, could also arise from haplotype imputation. As shown in Tables 2 and 3 with the info measure ($r^2$) for controls and cases, where cases have more informative genetic data. This could happen if a certain population is not being represented in the reference panel used for imputation or if different reference panels or SNP-chips are used for cases and controls. A systematic difference in imputation quality is roughly equivalent to having a different average sequencing depth. However, at the same time, we have to state that our simulations and our analyses have shown us that in many instances dosages perform just as well. However, as mentioned there are certain circumstances where they do not perform as well. We have tried to explore these circumstances in this article and what the impact of them is.

Another conclusion, we can draw from our simulations is that when there is sequencing depth and phenotype correlation, our estimates of the effect size will be biased when we do not know the true frequency as shown in Supplementary Figures S4, S6, and S8. Furthermore, when there is sequencing depth and phenotype correlation it seems that keeping individuals with 0 reads makes a difference; however, this is due to how the simulations were performed and might not generalize to all scenarios.

With ANGSD-asso implemented in ANGSD, we have made it possible to perform large association studies with low depth sequencing data retaining maximal statistical power, and also estimating effect sizes. SNPTEST is too slow for the analysis of large-scale data sets. The speed-up of ANGSD-asso’s latent model compared with SNPTEST is due to priming for faster convergence of the EM algorithm and threaded analysis using the ANGSD (Korneliussen et al. 2014) framework. ANGSD-asso makes the analysis of large-scale data possible as done in Liu et al. (2018) (141,431 individuals) while retaining maximal statistical power.

Data availability

The three methods in ANGSD-asso for association analysis are implemented in the ANGSD framework. ANGSD can be downloaded from its github page: https://github.com/ANGSD/angsd.
The R-scripts used for generating the simulations are available from https://github.com/e-jorsboe/ANGSD-asso-scripts. Some of the simulations use population frequencies from (Lazaridis et al. 2014), this data set is available at https://reich.hms.harvard.edu/datasets. The imputed genetic data and phenotypes that are used, are available with the permission of the UK Biobank (https://www.ukbiobank.ac.uk). We conducted the research using the UK Biobank resource under an approved data request (ref: 32683).

**Supplementary material** is available at G3 online.

**Acknowledgments**

E.J. thanks Thorfinn Sand Korneliussen for the help with implementing a BGEN file format reader.

**Funding**

The Lundbeck Foundation, Grant/Award Number: R215-2015-4174.

**Conflicts of interest**

The authors declare that there is no conflict of interest.

**Literature cited**


E.J. thanks Thorfinn Sand Korneliussen for the help with implementing a BGEN file format reader.

**Funding**

The Lundbeck Foundation, Grant/Award Number: R215-2015-4174.

**Conflicts of interest**

The authors declare that there is no conflict of interest.

**Literature cited**


