A Genotype Likelihood Framework for GWAS with Low Depth Sequencing Data from Admixed Individuals

Emil Jørsboe, Anders Albrechtsen

September 28, 2019

The Bioinformatics Centre, Department of Biology, University of Copenhagen, 2200 Copenhagen N, Denmark.

Correspondence to: emil.jorsboe@bio.ku.dk

1 Abstract

Introduction: Association studies using low depth NGS data provide a cost efficient design. Here we introduce an association method that works for low depth NGS data where the genotype is not directly observed. We will investigate how using different priors when calculating genotype probabilities will affect association analysis, and how this approach is affected by population structure. Doing association studies with genetic dosages is a widely used method for taking genotype uncertainty into account. We will investigate how our genotype probability based method compares to using dosages in large association studies with low depth NGS data. **Methods:** Our association method for low depth NGS data works by modelling the unobserved genotype as a latent variable. Our implementation is in a generalised linear model framework, using a maximum likelihood approach. We use the EM algorithm for maximising the likelihood. **Results & Discussion:** Our simulations using different priors in low depth NGS data in a structured population, show that using an individual allele frequency prior has better statistical power for association analysis. When there is a correlation between sequencing depth and phenotype the individual allele frequency prior also helps control the false positive rate. In the absence of population structure the sample allele frequency prior and the individual allele frequency prior perform similarly. We show through simulations that in certain scenarios the latent variable approach has better statistical power than dosages. Lastly when adding additional covariates to the model our method has more statistical power and provides less biased effect sizes than SNPTEST, while also being much faster than SNPTEST. This makes it possible to properly account

for genotype uncertainty in large scale association studies based on low depth sequencing data.

2 Introduction

2.1 Association with NGS data

Genome-wide association studies (GWAS) performed with low depth next-generation sequencing (NGS) data provide a cost efficient design, where the number of individuals studied can be maximised and therefore this design provides good statistical power to detect associations.

Recent successful GWAS with low depth NGS data have shown the success of this approach, one example of this is Liu et al. [2018]. In Liu et al. [2018] around 140,000 individuals were sequenced to an average sequencing depth of $0.1X$. Despite the low sequencing depth several novel associations were discovered. This shows that when using methods that account for the genotype uncertainty in low depth NGS data, good statistical power for detecting associations can be achieved, despite the modest amount of data.

Using methods that take genotype uncertainty into account have advantages, compared to calling genotypes for low depth NGS data and then doing association analysis with those, as shown in Skotte et al. [2012]. In Skotte et al. [2012] they develop a score test for doing association analysis with low depth NGS data. In that method 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 paper we will introduce a method in a generalised linear model framework that also estimates the effect size of the unobserved genotype, and that in practice can be run almost as fast as the score test. This will be done using a maximum likelihood approach, more specifically we will make use of the EM algorithm to maximise the likelihood, treating the unobserved genotype G as a latent variable. Using a generalised linear model framework enables us to include covariates thereby adjusting for possible confounders, such as population structure. We have implemented an EM algorithm that converges fast plus our method can be run multi-threaded, making the analysis of large data sets possible.

Using the EM algorithm for doing maximum likelihood estimation in a generalised linear model framework using genotype probabilities, has been implemented in SNPTEST [Marchini et al., 2007]. SNPTEST is not designed for the analysis of large scale data sets and is too slow for this. We have designed a much faster implementation that allows for the association analysis of large scale NGS data sets. A common used practice for doing association analysis with genotype data with uncertainty is using genetic dosages. They are easy to implement into most existing methods as the genotype can be directly replaced by the dosage. However dosages do not convey the uncertainty on the genotype as fully as genotype likelihoods or genotype probabilities. In Zheng et al. [2011] they show a gain in power when using genotype probability based

methods compared to dosages, but only for small studies with variants with large effect sizes. However they did not look at how a correlation between the sequencing depth and the phenotype might affect this. This could happen in a case-control study, where a systematic bias in the sequencing depth could be generated if cases and controls were sequenced at different places, or if the data set is merged from other smaller heterogeneous data sets. We will investigate this sequencing depth bias through simulations of a large scale association study with a sequencing depth bias. We will evaluate the performance of our genotype probability based method compared to using dosages.

We will also explore how to take population structure into account when doing association studies with low depth sequencing data. Population structure is a common confounder in association studies if not addressed properly. In low depth sequencing data a sample allele frequency prior is often used when estimating genotype probabilities, however this assumes a homogeneous population without structure. We therefore propose a new method for dealing with structured populations when doing association studies with low depth sequencing data. We will do this by using an individual allele frequency prior for when estimating the genotype probabilities. The individual allele frequency takes both the frequency of the variant and the ancestry of every individual into account. We therefore want to investigate how different priors work in different scenarios. We will look at this both with regards to statistical power and with regards to the false positive rate, if there is correlation between the phenotype and the sequencing depth.

3 Methods

NGS produces short reads that are then mapped to a reference genome. From the aligned reads the probability of observing these reads given a certain genotype can be inferred 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 material. The genotype likelihood can be converted into the probability of the genotype given the data, this 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 analysed in association see Figure 1.

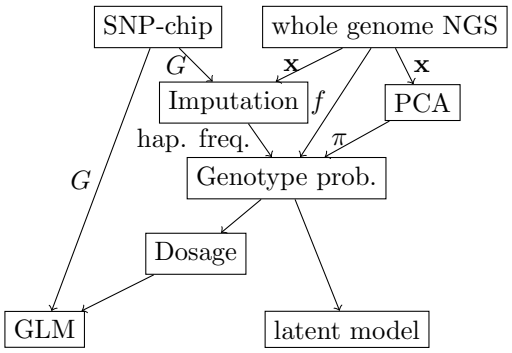


Figure 1: Schematic of workflow for doing association studies with genetic data. Data either gets generated using SNP-chips or doing whole genome NGS. The NGS data can be converted into genotype probabilities assuming no population structure using the sample allele frequency, or assuming population structure and then using PCA to generate genotype probabilities. SNP-chip genotypes can be analysed directly. Both kinds of data can be imputed using haplotype frequencies for generating genotype probabilities. The genotype probabilities can be analysed in a latent model (our model) or converted to dosages and then be analysed with a generalised linear model (GLM). $\mathbf{x}$ is the sequence data that can be converted to genotype likelihoods, G is the genotypes and π is the individual allele frequencies and f is the sample allele frequency.

3.1 EM model

We model the data using a maximum likelihood approach in a generalised linear model framework. This enables us to test for an association without observing the genotype G directly. Rather we observe our NGS data ($\mathbf{x}$), from this we can infer $p(G|\mathbf{x})$ or rather the probability of the genotype given the observed data (reads), this is also referred to as the genotype probability. We write the likelihood for our phenotype data ($\mathbf{y}$) given our sequencing data ($\mathbf{x}$) and covariates (Z)

$$p(\mathbf{y}|\mathbf{x}, Z,) = \prod_i^N p(y_i|x_i, \mathbf{z}_i) = \prod_i^N \sum_{g\in\{0,1,2\}} p(y_i|G=g, z_i)p(G=g|x_i), \qquad (1)$$

where we use the law of total probabilities to introduce the latent variable G. N is the number of individuals, $\mathbf{y} = (y_1, y_2, ..., y_N)$ is a vector of our observed phenotype for each individual, $\mathbf{x} = (x_1, x_2, ..., x_N)$ is a vector of sequencing data for each individual and $Z = (\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_N})$ is a $n \times c$ matrix with the covariates. We see that the trait y_i is conditionally independent of the sequencing data given the genotype (meaning $p(y_i|G=g, x_i, z_i) = p(y_i|G=g, z_i)$). We can calculate the genotype probability only making use of 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 $p(G = g|x_i, \mathbf{z_i}, f) = p(G = g|x_i, f)$), however for simplicity we omit f from the likelihood.

This allows us to write the likelihood, also introducing the parameters of our generalised linear model $\theta = (\alpha, \beta, \gamma)$, again we assume that the genotype is conditionally independent of the covariates given the sequencing data

$$L(\theta) \propto p(\mathbf{y}|\mathbf{x}, Z, \theta) = \prod_i^N p(y_i|x_i, \mathbf{z_i}, \theta) \tag{2}$$

$$= \prod_i^N \sum_{g\in\{0,1,2\}} p(y_i|G = g, \mathbf{z_i}, \theta)p(G = g|x_i) \tag{3}$$

$$= \sum_i^N \log\left(\sum_{g\in\{0,1,2\}} p(y_i|G = g, \mathbf{z_i}, \theta)p(G = g|x_i)\right). \tag{4}$$

Assuming the term $p(y_i|G = g, \mathbf{z_i}, \theta)$ follows a normal distribution, given the genotype G takes the value g, the covariates Z and the linear coefficients θ, the mean is given by

$$\eta_i = \alpha + \beta g + \sum_c \gamma_c z_{ic} + \epsilon_i. \tag{5}$$

Equation 4 is the log-likelihood function that we want to maximise with regards to the parameters θ. We will do this using the EM algorithm where our latent variable is the unobserved genotype G. For the full derivations of this see the supplementary material. We have also implemented logistic and Poisson regression where we have introduced a link function for η_i for eq. 5 and changed the distribution for $p(y_i|G, \mathbf{z_i}, \theta)$ accordingly. For more information on this see the supplementary material.

Furthermore standard errors on the estimated effect sizes are estimated using the observed Fisher information matrix as in Lake et al. [2003] and Skotte et al. [2019].

3.2 Hybrid model - for fast computation

The score test as described in Skotte et al. [2012] only has to estimate the parameters of the null model, where uncertainty on the variables do not have to be taken into account. It is therefore faster than our approach, where we both have to estimate the null and the alternative model. The idea behind the hybrid model is combining the speed of the score test with the desirable properties of the EM algorithm approach, where estimates of the effect size and standard error can be obtained. It works by first running the score test, and then if the site has a P-value below a certain threshold, we additionally run the slower EM algorithm method as well

$$p = \begin{cases} p_{score} < threshold & \Rightarrow \text{return } p_{EM} \\ p_{score} \geq threshold & \Rightarrow \text{return } p_{score}. \end{cases} \tag{6}$$

The threshold can be set by the user in `ANGSD`. The default value is 0.05.

3.3 Dosage model

An easy way to accommodate some of the genotype uncertainty is calculating the expected genotype or the dosage $E[G|\mathbf{x}]$

$$E[G|\mathbf{x}] = p(G = 1|\mathbf{x}) + 2p(G = 2|\mathbf{x}). \tag{7}$$

The genotype probability $p(G|\mathbf{x})$ can be calculated using the genotype likelihood $p(\mathbf{x}|G)$ and the frequency f of the genetic variant using Bayes' formula

$$p(G|\mathbf{x}, f) = \frac{p(\mathbf{x}|G)p(G, f)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|G)p(G, f)}{\sum_{g\in 0,1,2} p(\mathbf{x}|G = g)p(G = g, f)}. \tag{8}$$

Here it is assumed that we have one homogeneous population where f describes the frequency of the genetic variant well across all individuals. Genotype probabilities can also be calculated using haplotype imputation. We have implemented a dosage model in `ANGSD`. We do standard ordinary least squares using $E[G|\mathbf{x}]$ as our explanatory variable

$$y_i = \sum_c \gamma_c z_{ic} + E[G_i|x_i]\beta + \epsilon_i. \tag{9}$$

3.3.1 Implementation

The 3 methods for association analysis are implemented in the `ANGSD` framework [Korneliussen et al., 2014], allowing multi-threaded analysis. `ANGSD` can be downloaded from its github page: `https://github.com/ANGSD/angsd` The EM model is `-doAsso 4`, the hybrid model is `-doAsso 5` and the dosage model is `-doAsso 6`. These methods work on genotype probabilities in the beagle file format, as used in `ANGSD`.

3.4 Individual allele frequency prior

When estimating the genotype probabilities for low depth sequencing data, it is important to have an accurate prior, when dealing with genotype data in a structured population. The sample frequency f of an allele might not describe the occurrence of an allele across individuals very well. This is due to the fact that the frequency of an allele might differ drastically between different ancestries. Therefore a prior based on the sample frequency will not work well in a structured population. If we have a discrete number of ancestral populations then by using a weighted average of the ancestral frequencies we can calculate the individual allele frequency (π_{ji}), for individual i for site j, across k populations

$$\pi_{ji} = \sum_k q_{ki} f_{jk}. \tag{10}$$

Where f_{jk} is the frequency of the jth site in population k and q_{ik} is the admixture proportion of population k for individual i. In order to estimate the individual allele frequencies we will have to first estimate the ancestral frequencies and the admixture proportions. For NGS data this can be done using NGSadmix [Skotte et al., 2013] and for genotypes this can be done using ADMIXTURE [Alexander et al., 2009]. We use the approach from NGSadmix when inferring population frequencies, in our simulations with low depth sequencing data in a structured population, assuming admixture proportions are known.

Another approach is [Hao et al., 2015] or PCAngsd [Meisner and Albrechtsen, 2018], where the population structure between individuals is modelled using principal components rather than a discrete number of ancestral populations. When the individual allele frequencies have been generated we can calculate more accurate genotype probabilities, this can be done using Bayes' formula as laid out in eq. 8 (where we replace f by π). $p(G)$ can be calculated using our individual allele frequency assuming Hardy-Weinberg proportions

$$p(G) = \begin{cases} (1-\pi_{ij})^2 & G = 0 \\ 2\pi_{ij}(1-\pi_{ij}) & G = 1 \\ (\pi_{ij})^2 & G = 2. \end{cases} \tag{11}$$

4 Results

In order to investigate what prior works best for generating the genotype probabilities in different scenarios we simulated data with and without population structure and with and without sequencing depth phenotype correlation. For each scenario we both applied a sample allele frequency prior and an individual allele frequency prior. We wanted to evaluate how both priors work with regards to false positive rates and statistical power to detect an association. Another aspect we wanted to investigate is statistical power in a large scale NGS association study when using dosages versus when using our genotype probabilities based approach. We therefore simulated a large scale association study with low depth sequencing data. We also compared our method with SNPTEST in terms of bias, statistical power and computational speed.

4.1 Evaluation of using different priors

We chose 4 different simulation scenarios, in scenario 1 and 2 there is no population structure. In scenario 1 we looked at the false positive rate when there is sequencing depth and phenotype correlation, under our null hypothesis of no effect of the genotype. In scenario 2 we looked at statistical power simulating under our alternative hypothesis with no sequencing depth and phenotype correlation. Scenario 3 is similar to scenario 1 and 4 is similar to 2, but where there is population structure.

Scenario	freq.	N	population structure	seq. depth and pheno	simulated pheno mean
1	0.45	1000	no	correlated	$\delta \cdot D$
2	0.45	1000	no	not correlated	$\beta \cdot g$
3	$(0.9, 0.1)$	1000	yes	correlated	$q \cdot \gamma + \delta \cdot D$
4	$(0.9, 0.1)$	1000	yes	not correlated	$\beta \cdot g + q \cdot \gamma$

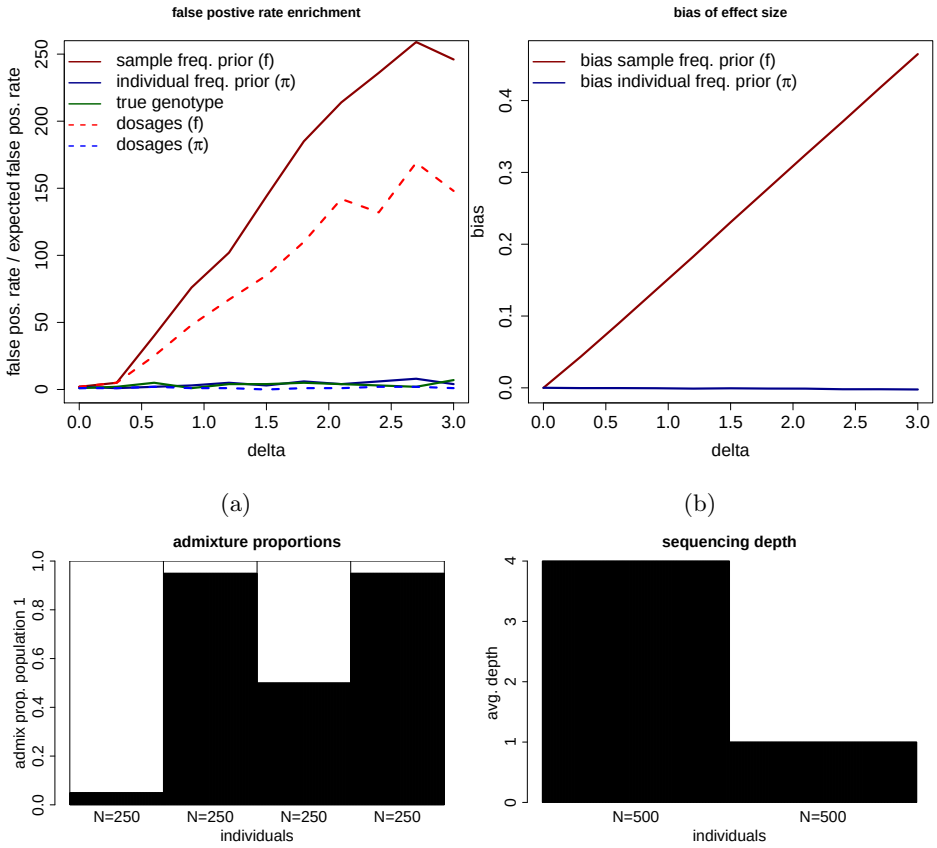
Table 1: All of the four scenarios are run with an additive model and the phenotype is simulated as a quantitative trait, with a mean given in the fifth column, and standard deviation 1. D is the sequencing depth with effect δ, g is the genotype with effect β and q is the ancestry with effect γ. In Scenario 3 and 4 there is population structure, with two ancestral populations. We estimate frequencies from the genotype likelihoods. For the admixed individuals we assume that the admixture proportions are known, we estimate the population frequencies using the approach from Skotte et al. [2013]

.

4.1.1 Using different priors in a homogeneous population

For scenario 1 with no population structure we aim to explore the effect of a sequencing depth phenotype correlation. Supplementary Figure 1 shows an acceptable false positive rate in a population without structure for all approaches. Using an individual allele frequency prior and a sample allele frequency prior yield identical results. This is expected since these priors become identical in the absence of population structure. Supplementary Figure 2 shows that in scenario 2 there is no difference in statistical power between the two priors when there is no population structure.

4.1.2 Using different priors in a structured population



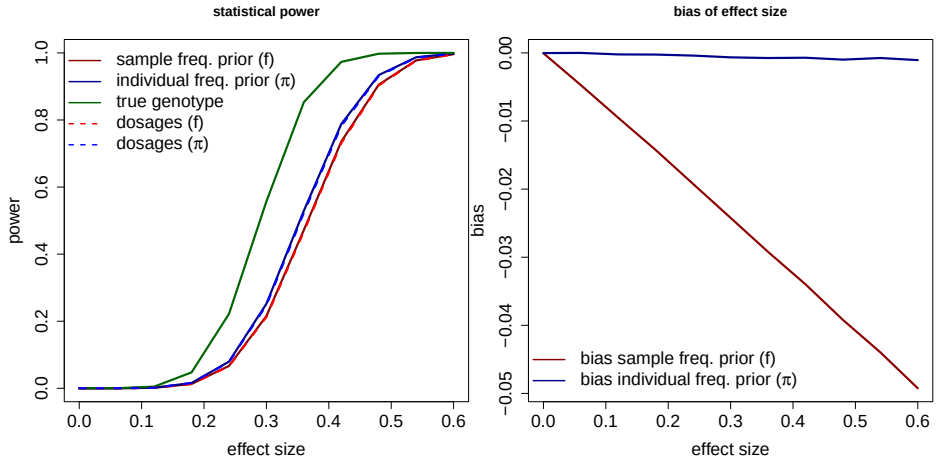
(a) (b)

(c)

Figure 2: This data is simulated according to scenario 3 in Table 1 varying the sequencing depth and phenotype correlation (delta). We have a structured population with 1,000 individuals. There is an effect of ancestry of population 1 ($\gamma = -0.3$). We use a significance threshold of 10^{-5}. The linear model is adjusted for ancestry. Each point is based on 100,000 simulations. **(a)**: We show the false positive rate divided by the expected false positive rate (10^{-5}). **(b)** We show the bias of our estimated effect size of the genotype. **(c)** The simulated admixture proportions and the mean sequencing depth for the simulated individuals.

For scenario 3 Figure 2 shows using the sample allele frequency prior makes us overestimate the effect of the genotype 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 an individual allele frequency prior we do not get biased estimates and have a false positive rate that is the same as when using the true genotype.



(a) (b)

Figure 3: This data is simulated according to scenario 4 in Table 1 varying the effect size of the genotype (beta). We have a structured population with the same admixture proportions and mean sequencing depth as in Figure 2 (c). There is an effect of ancestry of population 1 ($\gamma = -0.3$). We use a significance threshold of 10^{-5}. The linear model is adjusted for ancestry. Each point is based on 100,000 simulations. **(a)**: We show the statistical power to detect a true association. **(b)**: We show the bias of our estimated effect size of the genotype.

For scenario 4 Figure 3 shows using our individual allele frequency approach leads to slightly increased statistical power. For example for an effect size of $\beta = 0.36$ the power is 0.53 compared to 0.47. 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.

4.2 Comparison with dosages in large scale studies

Genotype dosages or the expected genotype is often used in association studies, in order to be able to try and account for the uncertainty on the genotype. However dosages can be very uninformative especially with low depth sequencing data. We did simulations in order to investigate the statistical power to detect an association, when we model the full genotype probabilities instead of using just the genotype dosage. We simulated a scenario with a large case-control study with 100,000 individuals, with low depth sequencing data, where

the cases and controls have been sequenced to different sequencing depths.

program / RR:	1	1.1	1.12	1.14	1.16
True genotype	0	0.587	0.868	0.978	0.999
Dosage	0	0.114	0.300	0.431	0.808
Latent model	0	0.163	0.388	0.659	0.862
R^2 cases/controls	0.91/0.85	0.91/0.85	0.91/0.84	0.90/0.84	0.90/0.84

Table 2: The phenotype is simulated as a binary trait, with different effect sizes or relative risk (RR) of the genotype. We have done 10,000 simulations for each tested effect size. The casual allele of the genetic variant has a frequency of 0.05 and the disease has a prevalence of 0.10 in the population. We have 50,000 controls and cases with an average sequencing depth of $1X$ and $4X$ respectively, effectively we do not have 50,000 controls or cases as some of these individuals will have no data (0 reads). The R^2 values are calculated like the info measure used in the MACH imputation software [Scott et al., 2007]

program / RR:	1	1.1	1.12	1.14	1.16
True genotype	0	0.813	0.974	0.999	1.000
Dosage	0	0.0914	0.262	0.523	0.772
Latent model	0	0.0974	0.273	0.538	0.783
R^2 cases/controls	0.91/0.77	0.90/0.75	0.90/0.75	0.90/0.75	0.90/0.75

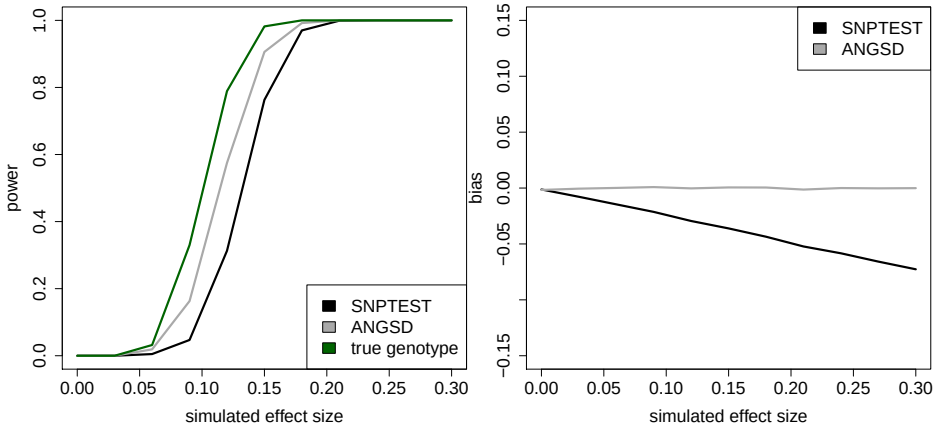
Table 3: This is the same scenario as Table 2, but where we include individuals with 0 reads.

In Table 2 we show how using the full genotype probabilities have increased statistical power compared to when using the 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. 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. To calculate the info measure we used the ratio of observed variance of the dosages to the expected binomial variance at Hardy-Weinberg equilibrium, as used in the imputation software MACH [Scott et al., 2007]. In order to explore if the there is also increased statistical power when analysing a quantitative trait, a version of scenario 1 in Table 1 was simulated but with an effect of the genotype. Supplementary Figure 3 shows that in this scenario there is also increased statistical power when using the full genotype probabilities. In Table 3 we run the analysis from Table 2, but including individuals with no 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, expect for the true genotype, we lose statistical power compared to when we remove individuals without reads.

4.3 Comparison with SNPTEST

SNPTEST [Marchini et al., 2007] also implements an EM algorithm for doing association (using the *-method em*). with genotype probabilities also using a generalisd linear model framework. We compared the estimated P-values for SNPTEST and our method implemented in `ANGSD` in Supplementary Figure 4, which shows a trend of lower P-values using our method. Therefore the bias of the estimated effect sizes was investigated, Figure 4 shows that SNPTEST's effect sizes are downward biased, whereas our method has no bias and that our method has increased statistical power compared to SNPTEST. This is a scenario where covariates are included to adjust for population structure. When not including covariates the estimates of the effect sizes are the same (data not shown). When using the SNPTEST approach for dosages (using the *-method expected*) the effect sizes are the same, also when including covariates (data not shown). We have used the most recent version of SNPTEST (v2.5.4-beta3).



(a) (b)

Figure 4: This data is simulated according to scenario 4 in Table 1, but with 10,000 individuals with an average depth of $0.1, 1, 10, 20X$, 2,500 individuals each. Varying the effect size of the genotype (beta). There is an effect of ancestry of population 1 ($\gamma = -0.3$), admixture proportion are like in Figure 2 (c). We use a significance threshold of 10^{-5}. The linear models are adjusted for ancestry, SNPTEST was run without transforming the phenotype or covariates to make it as comparable to `ANGSD` as possible. Each point is based on 1,000 simulations. **(a)**: We show the statistical power to detect a true association. **(b)**: We show the bias of our estimated effect size of the genotype.

We compared our method to SNPTEST in terms of computational speed and found that our EM algorithm converges faster. Also it can be run multi-threaded resulting in much reduced run times.

program / nr. sites:	100,000	200,000	300,000	400,000	500,000
EM model*	3.93 h	7.94 h	11.52 h	13.81 h	17.17 h
EM model, 20 threads*	0.25 h	0.52 h	0.79 h	1.07 h	1.35 h
SNPTEST	7.16 h	16.85 h	21.19 h	36.58 h	50.29 h
hybrid model*	1.23 h	2.54 h	4.74 h	6.09 h	6.32 h
hybrid model, 20 threads*	0.083 h	0.17 h	0.25 h	0.33 h	0.41 h
score test, 20 threads*	0.078 h	0.17 h	0.25 h	0.33 h	0.41 h
dosage model, 20 threads*	0.12 h	0.24 h	0.37 h	0.42 h	0.57 h

Table 4: Running times for an analysis of a simulated binary trait in 4,474 individuals with 12 covariates (age, gender and the first 10 principal components calculated from the genetic data). The genetic data has an average depth of $1X$. For each point we have run the analysis 3 times and then used the mean running time. * = all run in `ANGSD`.

Our method is many magnitudes faster than SNPTEST, especially for binary data, as shown in Table 4 and in Supplementary Table 1. When running our EM model threaded our method is approximately 30 to 40 times faster. The speed-up is even more dramatic when comparing our hybrid approach, which threaded can handle each of the analyses in less than 1 hour, whereas SNPTEST will takes days to run the largest data set, when running a logistic model.

In order to achieve faster convergence of our EM algorithm, we first do regression on the genotype dosages. We then use the coefficients obtained from the dosage regression as the starting guess for the coefficients for the EM algorithm (we refer to this as priming). As shown in Supplementary Figure 5 this drastically reduces the number of iterations needed for convergence of the EM algorithm.

5 Discussion

5.1 Implementation of model

We have implemented an EM algorithm approach for taking genotype uncertainty into account when doing association studies. The advantage of this approach compared to 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 means we can make use of LD-score regression. It is shown through simulations that our method has increased statistical power compared to SNPTEST as shown in Figure 4, when including covariates in the

model. Including covariates in the linear model is a common way to deal with confounders in association studies.

5.2 Different priors in structured and homogeneous populations

We have shown how using an 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 positives rate when there is sequencing depth phenotype correlation as shown in Figure 2. This correlation might arise if the sequencing is not randomised, for example if cases and controls are being sequenced at different places thereby creating a systematic bias, or if different cohorts have been sequenced at different places. The scenarios from Table 1 are most likely to arise when dealing with non model species where imputation cannot be done. This leads us to recommend using an individual allele frequency prior when doing association studies with NGS data in structured populations, where imputation is not possible.

5.3 Comparison with dosages in large scale studies

In Table 2 and 3 we show through simulations increased statistical power when using genotype probabilities compared to dosages, with a larger gain in power for the scenario from Table 2. In both instances 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 genotypic information for some individuals, could arise doing imputation. As shown in Table 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 are used for cases and controls. A systematic difference in imputation quality is roughly equivalent to having a different average sequencing depth. We also show increased statistical power when using genotype probabilities compared to dosages, when analysing a quantitative trait as shown in Supplementary Figure 3, even though this is a much smaller study in terms of the number of individuals. In this article we have not explored how imputation might affect association in a structured population. Even in a large scale association study with many individuals our method has increased statistical power in some scenarios. With our method implemented in `ANGSD` we have made it possible to do large association studies with low depth sequencing data retaining maximal statistical power, and also estimating effect sizes. In fact our hybrid model is almost as fast as the score test or using dosages as seen in Table 4. SNPTEST is too slow for the analysis of large scale data sets. The speed-up of our method compared to SNPTEST is due to priming for faster convergence of the EM algorithm and threaded analysis using the `ANGSD` [Korneliussen et al., 2014] framework. Our

method makes the analysis of large scale data possible as done in Liu et al. [2018] (141,431 individuals) while retaining maximal statistical power.

References

David H. Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9): 1655–1664, Sep 2009. doi: 10.1101/gr.094052.109.

Wei Hao, Minsun Song, and John D. Storey. Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5):713–721, Nov 2015. doi: 10.1093/bioinformatics/btv641.

T. S. Korneliussen, A. Albrechtsen, and R. Nielsen. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, 15(1):356, Oct 2014.

S. L. Lake, H. Lyon, K. Tantisira, E. K. Silverman, S. T. Weiss, N. M. Laird, and D. J. Schaid. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum. Hered.*, 55(1):56–65, 2003.

Siyang Liu, Shujia Huang, Fang Chen, Lijian Zhao, Yuying Yuan, Stephen Starko Francis, Lin Fang, Zilong Li, Long Lin, Rong Liu, Yong Zhang, Huixin Xu, Shengkang Li, Yuwen Zhou, Robert W. Davies, Qiang Liu, Robin G. Walters, Kuang Lin, Jia Ju, Thorfinn Korneliussen, Melinda A. Yang, Qiaomei Fu, Jun Wang, Lijun Zhou, Anders Krogh, Hongyun Zhang, Wei Wang, Zhengming Chen, Zhiming Cai, Ye Yin, Huanming Yang, Mao Mao, Jay Shendure, Jian Wang, Anders Albrechtsen, Xin Jin, Rasmus Nielsen, and Xun Xu. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and chinese population history. *Cell*, 175(2):347 – 359.e14, Oct 2018.

Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906, 2007.

J. Meisner and A. Albrechtsen. Inferring population structure and admixture proportions in low-depth ngs data. *Genetics*, 210(2):719–731, 2018. doi: 10.1534/genetics.118.301336.

Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443, 2011.

Laura J. Scott, Karen L. Mohlke, Lori L. Bonnycastle, Cristen J. Willer, Yun Li, William L. Duren, Michael R. Erdos, Heather M. Stringham, Peter S. Chines, Anne U. Jackson, Ludmila Prokunina-Olsson, Chia-Jen Ding, Amy J. Swift, Narisu Narisu, Tianle Hu, Randall Pruim, Rui Xiao, Xiao-Yi Li, Karen N. Conneely, Nancy L. Riebow, Andrew G. Sprau, Maurine Tong, Peggy P.

White, Kurt N. Hetrick, Michael W. Barnhart, Craig W. Bark, Janet L. Goldstein, Lee Watkins, Fang Xiang, Jouko Saramies, Thomas A. Buchanan, Richard M. Watanabe, Timo T. Valle, Leena Kinnunen, Gonçalo R. Abecasis, Elizabeth W. Pugh, Kimberly F. Doheny, Richard N. Bergman, Jaakko Tuomilehto, Francis S. Collins, and Michael Boehnke. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, 316(5829):1341–1345, Jun 2007. doi: 10.1126/science.1142382.

L. Skotte, T. S. Korneliussen, and A. Albrechtsen. Association testing for next-generation sequencing data using score statistics. *Genet. Epidemiol.*, 36(5): 430–437, Jul 2012.

L. Skotte, E. Jørsboe, T. S. Korneliussen, I. Moltke, and A. Albrechtsen. Ancestry-specific association mapping in admixed populations. *Genet. Epidemiol.*, 43(5):506–521, Jul 2019.

Line Skotte, Thorfinn Sand Korneliussen, and Anders Albrechtsen. Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3):693–702, Nov 2013. doi: 10.1534/genetics.113.154138.

Jin Zheng, Yun Li, Gonçalo R. Abecasis, and Paul Scheet. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genetic Epidemiology*, 35(2):102–110, Jan 2011. doi: 10.1002/gepi.20552.