Association testing for next-generation sequencing data using score statistics

Skotte, Line; Korneliussen, Thorfinn Sand; Albrechtsen, Anders

Published in:
Genetic Epidemiology

DOI:
10.1002/gepi.21636

Publication date:
2012

Document version
Peer reviewed version

Citation for published version (APA):
Association testing for next-generation sequencing data using score statistics

Line Skotte¹, Thorfinn Sand Korneliussen¹ and Anders Albrechtsen¹

October 6, 2011

¹ Department of Biology, University of Copenhagen, Denmark.

* Corresponding Author: Department of Biology, Ole Målses Vej 5, DK-2200 Copenhagen, Denmark, Europe.

Running title: Association for next-generation sequencing data.
1 Abstract

The advances in sequencing technology have made large scale sequencing studies for large cohorts feasible. Often the primary goal for large scale studies is to identify genetic variants associated with a disease or other phenotypes. Even when deep sequencing is performed there will be many sites where there is not enough data to call genotypes accurately. Ignoring the genotype classification uncertainty by basing subsequent analyses on called genotypes leads to a loss in power. Additionally, using called genotypes can lead to spurious association signals. Some methods taking the uncertainty of genotype calls into account have been proposed, most require numerical optimization which for large scale data is not always computationally feasible.

We show that using a score statistic for the joint likelihood of observed phenotypes and observed sequencing data provides an attractive approach to association testing for next-generation sequencing data. The joint model accounts for the genotype classification uncertainty via the posterior probabilities of the genotypes given the observed sequencing data, which gives the approach higher power than methods based on called genotypes. This strategy remains computationally feasible due to the use of score statistics. As part of the joint likelihood, we model the distribution of the phenotypes using a generalized linear model framework, which works for both quantitative and discrete phenotypes. Thus the method presented here is applicable to case-control studies as well as mapping of quantitative traits. The model allows additional covariates which enables correction for confounding factors such as population stratification or cohort effects.

2 Introduction

Due to rapidly decreasing prices the use of next-generation sequencing technologies is becoming more common in genetic studies. These technologies provide several advantages from previous sequencing and genotyping technologies. However the use of next-generation sequencing technologies in association studies is not unproblematic. Unlike traditional Sanger sequencing and genotyping platforms next-generation sequencing platforms like Solexa, Solid and 454 do not provide genotype calls directly. Sites from pairs of homologous chromosomes are not sequenced in equal proportions but instead the chromosomes are sampled with replacement. Often the depth is low at some positions which means that sometimes only the site on one of the homologous chromosomes is sequenced. On top of that there are non negligible errors in the sequencing data which add another layer of uncertainty to genotype calls. In this paper we directly address the uncertainty of
genotype calls using posterior genotype probabilities based on the observed sequencing data.

So far only partial genome wide data has been generated for a large number of individuals (Li et al., 2010; Durbin et al., 2010) but large studies that will explore genotype phenotype correlations have been undertaken. These studies involve thousands of individuals sequenced at a low depth. A strategy for performing association mapping for these individuals could involve calling genotypes either based solely on the individual reads or based on a multiple sample strategy using information from either the allele frequencies or haplotype frequencies. The later will provide higher accuracy but as we will show calling genotypes using a multiple sample strategy can in some instances cause spurious associations. A much better and more robust choice is to incorporate the genotype classification uncertainty directly in the model used for association testing.

We will take a similar approach to dealing with genotype classification uncertainty as has been taken for haplotype analysis by Schaid et al. (2002). Here the haplotype for each individual is not directly observed but instead posterior probabilities for the possible haplotypes for each individual are estimated based on maximum likelihood estimates of the haplotype frequencies. The uncertainty of the haplotypes are taken into account by considering the joint likelihood of the trait and the observed markers, found by summing a generalized linear model of the phenotype over the possible haplotypes weighted by the posterior probabilities of the haplotypes given the observed markers. A score statistic is then used to evaluate significance. This approach is very attractive because the parameters only need to be estimated under the null hypothesis which means that no additional numeric optimization is needed. This is also one approach used in the testing program for the popular haplotype imputation software IMPUTE (Marchini et al., 2007; Howie et al., 2009).

In this paper we will show that using the joint likelihood of phenotypes and sequencing data provides a more powerful and more robust method for association mapping compared to using standard association methods based on called genotypes. Using a score statistic to evaluate the significance, the method remains computationally feasible for very large samples and whole genomes. We will also show that choosing the generalized linear model framework makes the method flexible enough to correct for confounding factors such a population stratification and cohort effect.
3 Methods

In this section we will first provide an overview of the models on which association mapping can be based, building up to our association method based on the joint likelihood of the observed phenotypes and the observed sequencing data. We then describe the simulated sequencing data and the real sequencing data used to evaluate the performance of the described methods on simulated phenotypes.

3.1 Frequency estimation

To estimate the allele frequencies from the sequencing data we use the maximum likelihood estimate assuming known major and minor alleles as described in Kim et al. (2010) for pooled data. The only difference for this data set is that every individual constitutes its own pool. This frequency estimator assumes that the sites are diallelic and that the two alleles are known.

For a specific site let \( f \) denote the minor allele frequency and let \( D_i \) denote the data for the \( i \)th individual. The likelihood for the data \( D_i \) for the \( i \)th individual given the allele frequency \( f \) can be written as

\[
p(D_i|f) = \sum_{g \in \{0,1,2\}} p(D_i|G_i = g)p(G_i = g|f).
\]

Possible ways to calculate the genotype likelihoods \( p(D_i|G_i = g) \) are described in supplementary material. The probabilities of the genotypes given the minor allele frequency \( p(G_i = g|f) \) are calculated assuming Hardy-Weinberg equilibrium. The maximum likelihood estimator for the frequency is given as

\[
\hat{f} = \arg \max_f \prod_i p(D_i|f).
\]

We use the EM-algorithm for fast optimization, see supplementary material.

3.2 Genotype calling

The easiest way to perform association mapping is to first call the genotypes and then use standard methods and software. There have been different approaches used to call genotypes for next generation sequencing data. Some studies use only the likelihood information in order to call genotypes (Ng et al., 2010), some use a bayesian approach and specify a prior probability for the genotypes (Li et al., 2009b, 2008) or the genotypes can be called using the allele frequency as a prior (McKenna et al., 2010; Kim et al., 2011). Here we compare the genotype calling for two of
these three approaches on real data and compare their performance for association with methods that take the uncertainty of genotype calls into account.

- Maximum likelihood genotype

\[ \hat{g}_i = \arg \max_g p(D_i|\hat{g}) \]

- Maximum posterior genotype with frequency prior

\[ \hat{g}_i = \arg \max_g p(D_i|\hat{g}) p(\hat{g}|\hat{f}) \]

We also explore the possibility of imposing a cutoff based on the posterior probability of the genotype. This cutoff sets the genotype to missing if the posterior probability of the genotype is less than 95%.

3.3 Association testing with uncertain genotypes

The actual genotypes remains unobserved in a next-generation sequencing study. Instead of basing association testing on called genotypes, we therefore suggest using a joint model of the observed sequencing data and the phenotype. This way we take the genotype classification uncertainty into account.

We assume that the phenotype \( y_i \) only depend on the observed data \( D_i \) through the true, but unknown genotype \( g_i \) of the \( i \)th individual. This enable us to rewrite the joint probability of \( y_i \) and \( D_i \) (given any observed additional covariates \( x_i \))

\[ p(y_i, D_i | x_i) = \sum_{g \in \{0,1,2\}} p(y_i | x_i, g) p(D_i, g). \]

We use the Bayesian approach described in section 3.1 to replace \( p(D_i, g) \) with \( p(D_i|g)p(g|\hat{f}) \), where the first term is the genotype likelihood and the second term the genotype prior calculated using the estimated minor allele frequency assuming Hardy-Weinberg equilibrium.

We then model the distribution of the phenotype \( p(y_i | x_i, g) \) in the exact same way as the standard methods for association testing described in the supplement, i.e. equation (1). Thus the full likelihood function for the model that takes genotype classification uncertainty into account
becomes

\[ p_{\alpha,\beta,\phi}(y, D|x) = \prod_i \sum_{g \in \{0, 1, 2\}} p_{\alpha,\beta,\phi}(y_i|x_i, g)p(D_i|g)p(g|\hat{f}). \]

To test for association, the hypothesis \( \beta = 0 \) must be tested. This can be done using the likelihood ratio statistic

\[ Q(x, D) = \frac{p_{\alpha,0,\hat{\phi}}(y, D|x)}{p_{\alpha,\hat{\beta},\hat{\phi}}(y, D|x)}, \]

but finding the maximum likelihood estimates \((\hat{\alpha}, \hat{\beta}, \hat{\phi})\) must be done using numeric optimization, which can be time consuming. Finding the maximum likelihood estimates \((\hat{\alpha}, 0, \hat{\phi})\) under the hypothesis \( \beta = 0 \) is on the other hand not complicated, since the likelihood function becomes much simpler when it is assumed that the genotype has no effect on the trait. If \( \beta = 0 \) then

\[ -2 \log Q(x, D) \]

follows a \( \chi^2 \) distribution.

Alternatively (and asymptotically equivalent) testing of the hypothesis \( \beta = 0 \) can be done using the score statistic,

\[ R(y, D) = s_{y,D}(\hat{\alpha}, 0, \hat{\phi})^\top o_{y,D}(\hat{\alpha}, 0, \hat{\phi})^{-1} s_{y,D}(\hat{\alpha}, 0, \hat{\phi}), \]

where \( s_{y,D}(\alpha, \beta, \phi) \) is the score function, consisting of the first order derivatives of the likelihood function, and \( o_{y,D}(\alpha, \beta, \phi) \) is the observed information matrix, consisting of the second order derivatives of the likelihood function. Both functions are evaluated at the maximum likelihood estimate under the hypothesis \( \beta = 0 \), thus using the score statistic we avoid time consuming numerical optimizations. A general expression for the score statistic of the model described here is given in appendix A.6, along with specific expressions to be used with continuous and discrete traits respectively.

We compared the performance of the score based association method that takes genotype uncertainty into account with the standard association methods based on called genotypes. We also evaluated the importance of basing the genotype prior on the estimated minor allele frequency, by replacing this prior with a uniform prior.
3.4 Simulated sequencing data

When simulating data for quantitative traits, genotypes were simulated assuming a population minor allele frequency, random sampling from this population and Hardy-Weinberg equilibrium (HWE). For case control data we also assumed a minor allele frequency in the population but assumed an uneven sampling from the population based on the relative risk of an allele and the prevalence of the disease in the population. For all case control study simulations we assumed that the prevalence is 10% in the population. To simulate population stratification we used the frequency estimates for the YRI and CEU populations from HapMap phase II for chromosome 21. We used the 37477 overlapping SNPs from the two populations that had a MAF of at least 5%. Using these allele frequencies we simulated 1000 individuals in two populations with a phenotype $\sim N(0,1)$ for population one and $\sim N(0.5,1)$ for population two.

Sequencing data for these simulated genotypes was simulated by first choosing an average depth for all individuals and then sampling the specific depth assuming a Poisson distribution. Quality score for each read was sampled from an empirically obtained distribution (see section below). For simplicity we assume that only two bases exits and sample the reads from these two alleles conditional on the simulated genotype and the error rate from the quality scores – the transformed quality score being interpreted as the probability for a base being an error.

3.5 Low coverage sequencing data

We used the data from the 1000 genomes low sequencing pilot project (Durbin et al., 2010) that overlap with the Hapmap phase I and II individuals (Frazer et al., 2007). 32 unrelated CEPH individuals and 41 Yoruban individuals were present in both HapMap and 1000 Genomes. These individuals had been re-calibrated and aligned to the reference genome using MAQ (Li and Durbin, 2009) and were available in the SAMtools BAM format (Li et al., 2009a).

Genotyping data for these individuals were obtained for the 4936123 available sites for HapMap phase I and II release 24 mapped to hg18. The sequencing data had covered 2722683 of those sites. We obtained an empirical distribution of the quality score for all overlapping sites.

For the association simulations we oversampled the individuals so that each individual were included 100 times. The phenotypes were simulated using a normal distributed quantitative trait for the oversampled individuals. This creates a small dependency between the phenotypes and genotypes that we remove by randomly permuting the phenotypes between individuals for each site within each population.
4 Results

We have summarized the key steps for performing association testing in figure 1. We provide an implementation of the methods for association testing described above. In order to supply flexibility for the users our implementation allow three different entry points. The most basic one is just supplying the aligned reads e.g. sorted SOAP files. If this option is chosen candidate SNP sites will be selected based on the fast frequency estimator used in Li et al. (2010). Sites with a minor allele frequency above a certain threshold will be considered candidate SNP sites. Then genotype likelihoods are calculated using the either the method implemented in GATK (McKenna et al., 2010) or Kim et al. (2011). As a second entry point, our implementation allow for other genotype likelihood estimators e.g. SAMtools (Li et al., 2008), to be used instead. Using the likelihoods we determine the major and minor allele as described in appendix A.2. After the two alleles have been determined we calculate the maximum likelihood estimate of the allele frequency and use this frequency as a prior in order to create the genotype probabilities. As a third entry point these probabilities can also be provided directly from other sources, the most obvious one being haplotype imputation software such as Beagle (Browning, 2008). Additional filters based on the genotypes and allele frequencies can subsequently be applied before performing the association using the score statistics.

4.1 Genotype calling

Association can either be performed by first calling the genotypes from the read data or by incorporating the genotype uncertainty into the association model using the genotype likelihoods. There are many strategies one can use to call genotypes or estimate the probability of the genotypes. These strategies will have different power to detect association but will also have varying degrees of robustness for different biases. The performance of the methods for generating genotype probabilities will be reflected in their performance for genotype calling. In figure S1 we compare called genotypes from the 1000 genomes low coverage pilot project with the genotypes from HapMap. Each genotype calling strategy gives a posterior probability for the genotype being correct. We evaluate their performance for correctly calling a genotypes over a certain probability. The figure shows that using the frequency as a prior is by far the best strategy. However for non-polymorphic sites this will not necessarily be correct. An even more powerful approach is to use haplotype imputation. The principle is the same as using the allele frequency as a prior - the only difference is that instead of using allele frequencies to calculate a prior the haplotype frequencies are used.
4.2 Power

We compare the power of the score statistic to detect an association with methods based on true and called genotypes for various minor allele frequencies and sequencing depths. The results are shown figure 2. In each scenario we assumed (and simulated) a normal distributed trait and an additive genetic model meaning that the mean difference between having no minor alleles and having one minor allele is the same as the difference between having one minor allele and two minor alleles. The largest difference in power is seen when both the minor allele frequency and the sequencing depth is low. In figure A the minor allele frequency is set to 1% and the average depth is set to four. The depth distribution of each individual is sampled from a Poisson distribution. The most powerful approach would be to have genotypes without errors while the least powerful of the shown approaches is to call genotypes using maximum likelihood. Using a score statistic based on a uniform prior on the genotypes performs equally poorly. The score statistic that uses the full information of the genotype likelihoods always performs better or equally well as the methods that rely on calling genotypes. However, if genotypes are called using the allele frequencies as a prior it does perform fairly well especially compared to the methods where genotypes are called independently between individuals. In the simulation with a large minor allele frequency and a average sequencing depth of 8 all the different approached perform almost equally well. However, imposing a cutoff on the posterior probability of the called genotypes leads to a loss in power. Similar simulations for retrospective case-control data can be seen in figure S4 with similar results. For the case-control data without additional covariates the power of the score statistics and the test using differences in allele frequencies (Martin et al., 2010; Kim et al., 2011) have similar power.

4.3 False positive rate

We applied the method to real data from the 1000 Genomes project. We used the individuals that overlapped with the HapMap project Phase I and II and for each of the ca. three millions sites that were polymorphic in HapMap we performed association testing. For the CEU sample there were only 32 individuals. Therefore, we oversampled these individuals so that each individual was included 100 times. For each of the 3200 samples we randomly drew a phenotype from a normal distribution. Because the oversampling causes a small dependency in the individuals we randomly permuted the phenotypes at each sites. We applied some lenient filters on the data because the score tests requires some variation on the genotype likelihoods. We removed sites with a maf less than 10/3200 and sites with less than 10 individuals with a high (0.8) posterior probability of
having a minor allele. After applying the filters the false positive rate is controlled perfectly as seen in quantile-quantile plot (qqplot) in figure S3.

Because the tests are all based on asymptotic assumptions we also investigate the effect of a low number of individuals. The results are shown in figure S5 for sample sizes as low a 100. For the tests that take uncertainty into account the low number of individuals only affects the false positive rate slightly.

To further investigate the robustness of the difference approaches to association testing we simulated read data with an unequal depth distribution. Instead of a quantitative phenotype we simulated case control data with an average read depth of five without including any risk on the genotypes. We varied the average read depth for the case and the control data and applied the different association approaches to the data. The score test is not affected by the unequal depth as seen in figure 3. However if genotypes are called (and even when using a prior based on the allele frequency) then there is an increase in the false positive rate. Garner (2011) found similar effects of differences in sequencing depth basing a collapsing approach to association mapping on called genotypes. We also included the score statistic where the genotype probabilities are calculated without using the frequency as prior. Here we see a rapid increase in the false positive rate. This is because the expected genotypes in the cases and controls no longer is the same when the depth is not the same, and the test that takes genotype uncertainties into account to some extend correspond to testing for difference in genotype expectations.

4.4 Additional covariates

One of the advantages of using a score statistic in a linear regression framework is the ability of including additional covariates. This is very convenient if there are confounding factors that you want to correct for or if there are additional covariates that can explain a large portion of the variation of the trait. Using a simple simulation of two populations with a small difference in mean phenotypes we created the effect of population stratification. We then used principal component analysis (PCA) to detect this structure using either the true unobserved genotypes or using the genotype likelihoods (see figure S2A). Without correction for the population structure there is an increase in the false positive rate figure S2B). However, if the first eigen vectors from the PCA analysis are included as an additional covariate then the population structure is corrected for figure (S2C). The successful correction for population stratification can also be achieved on real data. We used both the CEU and the YRI individuals from the 1000 Genomes
and oversampled each individual 100 times. If we simulate the phenotypes independently of the populations we obtain the correct false positive rate (see figure 4). If the mean phenotype value is different in the two population we see a strong effect of population stratification but if we include the population assignment of each individuals as an additional covariate we can correct for the population stratification. Similar results are shown for a case control scenario in the supplementary figure S6.

4.5 Speed

Once the genotype probabilities have been estimated the score statistic does not require any optimization. This is because the parameters in the linear model only needs to be estimated under the null hypothesis and then these parameters are independent of the genotypes, thus they can be obtain using a set of linear equations. This makes the score statistic very fast. When applying the method on the 7300 genotype likelihood files used in figure 4, reading in the data, estimating the allele frequencies and performing the score test only takes about five hour for the ca. 3 million polymorphic sites. Although our implementation of the score statistic can use multiple CPUs this does not affect the computational speed much since the limiting factor quickly becomes the reading of the data. If the regression coefficient are also estimated under the alternative model the computational speed is decreased by several fold since numeric optimization is required (not shown).

5 Discussion

Taking uncertainty into account using a score statistic in a generalized linear framework is not new. This is the same approach taken by Schaid et al. (2002) for testing haplotypes, where the haplotype uncertainty is incorporated in the model via posterior haplotype probabilities given observed markers. We use an equivalent model to take the uncertainty of genotype calls into account, using the posterior genotype probabilities given the observed sequencing data. However, unlike Schaid et al. (2002) we use the full observed information matrix when calculating the score statistic, see appendix A.6. This makes no difference for the logistic regression model (no variance parameter is estimated) and in practice the difference is negligible for the linear regression model. For simulations performed in this study we found no detectable difference in performance (not shown).
Using simulations and real data we have shown that taking genotype uncertainty into account is a more powerful approach than calling genotypes while still adequately controlling the false positive rate. If a minimum posterior probability is also imposed on the genotype calling the loss of power is even greater. This is in accordance with previous studies of next-generation sequencing data (Kim et al., 2011) however unlike these methods the score statistic framework will also work for quantitative traits and will allow for additional covariates in the model. The latter is very attractive for association studies since population stratification and other confounding factors are a great concern. Additionally fixed effect meta analysis used for combining data from several studies can also be performed using this strategy. We have shown using real data that by including the population assignment in the model we can still control the false positive rate. Note however that because the estimated allele frequency is based on the whole sample it will not represent the individuals from either populations. This might give a slight reduction of power compared to using the population specific allele frequencies. If the population specific allele frequencies are know these can readily be used instead. The major drawback of the score statistic is that it does not provide an estimate of the effect sizes however the slower likelihood ratio test also taking the genotype uncertainties into account, where parameters are estimated both under the null and the alternative hypothesis, can be performed on a smaller subset of the data. Thus giving the opportunity to estimate the effect sizes for the top associations in the data.

Even though using the score statistic is more powerful than performing the association on called genotypes the difference is not so great when the genotypes are called using the estimated frequency as a prior especially when the depth is not low. One could argue that is it not worth performing a more complex analysis if the gain in power is not very significant. However, as we have shown the score statistic is much more robust to small biases in the data. If for example the quality scores or the depth are correlated with the phenotype then calling the genotypes will inflate the false positive rate. Calling the genotype using the inferred allele frequency as a prior is especially sensitive so even though this approach is attractive because of its good performance it is might not be the optimal choice.

We have also shown that using a score statistic it not a guarantee that differences in depth will not increase the false positive rate. If genotype probabilities are estimated independently between individuals we also see a large inflation in the false positive rate. This is because the tests that take the uncertainty info account to some extend corresponds to tests for difference in the expected genotypes. For example if the depth is lower in cases than in controls then the expected minor
allele frequency based on the genotype probabilities will be lower in cases than in the controls. Therefore the genotype prior should always be estimated using the whole sample.

The simulated analyses presented are all based on genotype probabilities estimated from the sample allele frequency assuming HWE. However the assumption of HWE is not necessary for the analysis. For populations where the HWE assumption does not hold it is advisable to base the genotype probabilities on genotype frequencies instead. To achieve higher power it is also possible to base the genotype probabilities on the haplotype frequencies (Browning, 2008). This can be a very powerful approach especially for common variations where adjacent sites provide lots of information through linkage disequilibrium.

The speed of this approach means that testing for associations is computationally feasible for thousands on individuals. This will be crucial for detecting variations involved in complex disease because very large cohort are needed.

6 Acknowledgement

This work has been supported by the Danish Council for Independent Research, Medical Sciences (FSS) and by the Lundbeck Foundation Centre of Applied Medical Genomics for Personalized Disease Prediction, Prevention, and Care (LuCAMP).

7 Web Resources

A software implementation of the method for association testing using score statistics will be made available upon publication.

8 Figures
Figure 1: Flow diagram of the key steps involved in the association test. This diagram is very general and applied to all methods explored in this study. There are three entry points for performing the association testing. (1) Supply the aligned reads and use a simple frequency filter determine candidate SNPs and estimate the genotype likelihoods based on the base counts. (2) Supply the genotype likelihoods and determine the major/minor alleles and their frequency based on the provided likelihoods. (3) Supply the genotype probabilities directly.
Figure 2: All plots are based on simulation of 2000 individuals. The minor allele frequency was set to 1% in the figure A and B and 20% in figure C and D. The average depth was set to 4 in figure A and C and 8 in figure B and D. Note the changed scale on the x axis.
Figure 3: 5000 cases and 5000 controls were simulated with a minor allele frequency of 0.01. The mean depth for all 10000 individuals were held constant at 5x while the ratio of depth between cases and controls were varied. The inflation is the observed false negative rate divided by the expected false negative rate for a significance threshold $\alpha = 0.01$.

Figure 4: QQ plot of the score statistic based on the 1000 Genomes low coverage data. 32 CEPH and 41 YRI individuals were sampled 100 times and normal distributed phenotypes were simulated for each of the 7300 individuals. The score statistic is shown in green for the scenario where the YRI and the CEPH have the same mean phenotype. In the orange and in the pink scenario the mean phenotype is higher for the YRI individuals than the CEPH but in the pink scenario the population affiliation for each individuals is included as an additional covariate.
References


A Supplementary material

A.1 Genotype likelihood estimation

An essential step in the analysis of next generation sequencing data is to determine the probability of the observed data given the 10 different possible genotypes, i.e. to estimate the genotype likelihoods. There are different approaches for calculating the genotype likelihoods (Li et al., 2008, 2009b; McKenna et al., 2010; Kim et al., 2010). For the real sequencing data, from the 1000 genomes low coverage pilot study, we used the likelihood estimator implemented in SAMtools (Li et al., 2008, 2009a).

For the simulated sequencing data we used the simple method presented in (McKenna et al., 2010). Briefly: For a single individual and a single site let \( D \) be the observed data containing \( n \) inferred bases, \( b = \{ b_1, b_2, \ldots, b_n \} \), and their quality scores. The transformed quality scores are interpreted as probabilities, \( P = \{ P_1, P_2, \ldots, P_n \} \), where \( P_j \) is the probability that the base \( b_j \) is an error. The likelihood of the genotype consisting of the alleles \( A_1 \) and \( A_2 \) is then proportional to the probability of the data \( D \) given that the genotype \( G \) of the individual is \( A_1 A_2 \), found by

\[
p(D|G = A_1A_2) = \prod_{j=0}^{n} \left( \frac{1}{2} p(b_j|A_1) + \frac{1}{2} p(b_j|A_2) \right)
\]

where

\[
p(b_j|A_1) = \begin{cases} 
\frac{P_j}{3} & b_j \neq A_1, \\
1 - P_j & b_j = A_1,
\end{cases}
\]

is the probability of observing the base \( b_j \) in a read given that the sequenced allele was \( A_1 \) and equivalently for \( A_2 \) (the three alternative bases splits the probability of a sequencing error equally).

Assuming that the considered site is diallelic it is possible to infer those two alleles directly from the reads as described in Kim et al. (2010) but we infer the alleles from from the genotype likelihoods (see appendix A.2). In the methods described in this paper, we will assume that the two alleles are known, thus only three of the 10 genotypes are possible and we shall use genotype sample space \( \{0, 1, 2\} \) in the following sections.

A.2 Inferring the major and minor allele

Assuming that the considered site is diallelic, we infer those two alleles using the genotype likelihoods. Let \( \{M, m\} \) denote the two possible alleles at the diallelic site, then the maximum likelihood
estimate of this pair is found using the likelihood function

\[
P(D|\{m, M\}) = \prod_i P(D_i|\{m, M\})
\]
\[
= \prod_i \sum_{A_1, A_2 \in \{m, M\}} P(D_i|G = A_1A_2)p(G = A_1A_2|\{m, M\}),
\]

where \(P(D_i|G = A_1A_2)\) is calculated as described in section A.1. We then assume that the two alleles within an individual are independent and randomly drawn from the set \(\{m, M\}\) with equal probability, ignoring the fact that the two alleles at a diallelic site are not observed equally frequent. This gives us

\[
p(G = A_1A_2|\{m, M\}) = 1/4
\]

for all four possible combinations of \(A_1, A_2 \in \{m, M\}\). Therefore we estimate the two possible alleles at the diallelic site by

\[
\arg \max_{\{m, M\}} \prod_i \sum_{A_1, A_2 \in \{m, M\}} P(D_i|G = A_1A_2).
\]

To infer which of these two alleles is the minor allele, we estimate the allele frequencies as described in section 3.1 (only one iteration of the EM algorithm is needed if the starting point is frequencies of 0.5).

### A.3 Em algorithm for frequency estimation

For a given frequency \(f_k\) we calculate the posterior expectation of each \(g_i\) assuming Hardy-Weinberg equilibrium,

\[
E(g_i|D_i, f_k) = \sum_{g \in \{0, 1, 2\}} g p(D_i|G_i = g)p(G_i = g|f) / \sum_{g \in \{0, 1, 2\}} p(D_i|G_i = g)p(G_i = g|f).
\]

Then the frequency is reestimated using

\[
f_{k+1} = \frac{\sum_{i=1}^N E(g_i|D_i, f_k)}{2N}.
\]

This EM-step is repeated until convergence (when successive estimates differ by no more than a small value).
A.4 Principal component analysis for sequencing data

We performed PCA analysis on the genotype data as described in Patterson et al. (2006) where the covariance matrix between all individuals is estimated from the genotypes assuming HWE. Using the estimated frequency as a prior we estimate the covariance matrix from the posterior probability of the genotypes. This is done in the exact same fashion as in eigensoft but instead of genotypes we use genotype expectations $\sum_{g=0}^{2} g p(g|\hat{f})$ and estimate the variance of the posterior expectations using the frequency estimated from the likelihoods. This is the same way eigensoft handles missing data.

A.5 Standard methods for association testing

The framework of generalized linear models provides a unified approach to standard association testing methods for discrete as well as continuous traits. We used the models described below to evaluate the performance of the different genotype callers for association testing. The models also form the basis of the methods for association testing that take the uncertainty of genotype calls into account.

One of the standard methods for performing an association study for a quantitative phenotype is to use a linear regression model (Balding, 2006). Given the genotypes $\{g_1, \ldots, g_N\}$ and possibly also additional covariates $\{x_1, \ldots, x_N\}$ for the $N$ individuals in the study, it is assumed that the quantitative traits $\{y_1, \ldots, y_N\}$ are independent and each follow a normal distribution

$$y_i \sim N(\alpha_0 + \alpha_1 x_i + \beta g_i, \sigma^2).$$

Here $x_i$ can be a vector of multiple observed covariates for individual $i$. To test for an association between the quantitative trait and the genotype, one simply test the hypothesis that $\beta$ is zero. This is usually done using the likelihood ratio statistic.

There are different standard methods for association testing in a case control study. A method that allows correcting for additional covariates is using a logistic regression model (Balding, 2006). Given the genotypes $\{g_1, \ldots, g_N\}$ and possibly also additional covariates $\{x_1, \ldots, x_N\}$ for the $N$ individuals in the study, it is assumed that the disease states $\{y_1, \ldots, y_N\}$ of the individuals are independent and that

$$\frac{p(y_i = 1)}{p(y_i = 0)} = \exp(\alpha_0 + \alpha_1 x_i + \beta g_i).$$

To test for an association between the disease and the genotype, one again test the hypothesis
that \( \beta \) is zero.

Both the linear regression model for the quantitative traits and the logistic regression model for the disease states conforms to the framework of generalized linear models (McCullagh and Nelder, 1989). This simply means that the probability of observing the phenotype \( y_i \) for the \( i \)th individual, can be written on the form

\[
p(y_i|x_i, g_i) = p_{\alpha, \beta, \phi}(y_i|x_i, g_i) = \exp \left( \frac{x_i \eta_i - b(\eta_i)}{a(\phi)} + c(x_i, \phi) \right),
\]

with the effect of the predictors on the phenotype expressed through the linear predictor

\[
\eta_i = \eta_{\alpha, \beta}(x_i, g_i) = \alpha_0 + \alpha_1 x_i + \beta g_i,
\]

with appropriate choices of the functions \( a, b \) and \( c \), depending on whether discrete or continuous phenotypes are considered. For more details see appendix A.6. Note that the distribution of the phenotype depend on the individual predictors (\( x_i \) and \( g_i \)) as well as on common parameters (\( \alpha, \beta \) and \( \phi \)). When convenient this dependence will be understood.

Estimating the parameters of the regression models is done by maximum likelihood estimation, using the likelihood function

\[
p_{\alpha, \beta, \phi}(y|x, g) = \prod_i p_{\alpha, \beta, \phi}(y_i|x_i, g_i).
\]

Testing for association, i.e. testing the hypothesis \( \beta = 0 \), is based on the likelihood ratio statistic

\[
Q(y) = \frac{p_{\hat{\alpha}, 0, \hat{\phi}}(y|x, g)}{p_{\hat{\alpha}, \hat{\beta}, \hat{\phi}}(y|x, g)},
\]

where \((\hat{\alpha}, \hat{\beta}, \hat{\phi})\) denote the maximum likelihood estimate of the parameters and \((\bar{\alpha}, 0, \bar{\phi})\) denote the maximum likelihood estimate assuming that \( \beta = 0 \). If \( \beta = 0 \) then \(-2\log Q(y)\) is approximately \( \chi^2 \) distributed, which gives the classical likelihood ratio test.

### A.6 Score statistic for unknown genotypes

In the following we consider the phenotypes \( y = (y_1, \ldots, y_N) \) of \( N \) independent individuals. The genotypes \( g = (g_1, \ldots, g_N) \) of the individuals are considered to be diallelic. The genotype of an
individual can be recorded as the number of minor alleles,

\[ g_i \in \{0, 1, 2\}. \]

Alternatively genotype can be recorded as \( g_i = (g_{i1}, g_{i2}) \), where \( g_{i1} \) is the indicator of at least one minor allele and \( g_{i2} \) is the indicator of two minor alleles,

\[ g_i \in \{(0, 0), (1, 0), (1, 1)\}. \]

To allow for both alternatives we derive the results below in the general case where \( g_i \) is a vector of length \( d_g \) (with \( d_g = 1 \) possible). We will allow the model to incorporate \( d_x \) additional covariates and let

\[ x_i = (1, x_{i1}, \ldots, x_{id_x}) \]

for the \( i \)th individual, where we have included the constant factor 1 on the 0th position to ease the notation. We let \( x = (x_1, \ldots, x_N) \) denote the intercept and the possible additional covariates of all the \( N \) individuals.

If the considered phenotype is quantitative we model the distribution of the phenotype conditional on the genotype and the other covariates by a standard linear regression model

\[ y_i \sim N(\alpha x_i^\top + \beta g_i^\top, \sigma^2) \]

where \( \alpha = (\alpha_0, \alpha_1, \ldots, \alpha_{d_x}) \) and \( \beta = (\beta_1, \ldots, \beta_{d_g}) \). To ease the notation we introduce the linear predictor \( \eta_i = \alpha x_i^\top + \beta g_i^\top \). Now the density of this distribution can be rewritten

\[

f(y_i | x_i, g_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( \frac{-(y_i - \eta_i)^2}{2\sigma^2} \right) \\
= \exp \left( \frac{y_i \eta_i - \eta_i^2 / 2}{\sigma^2} - \frac{y_i^2 / 2\sigma^2 - \log(2\pi\sigma^2)}{2} \right) \\
= \exp \left( \frac{y_i \eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi) \right)

\]

e. the density is from an exponential family with

\[ \phi = \sigma^2, \quad a(\phi) = \phi, \quad b(\eta) = \eta^2 / 2 \quad \text{and} \quad c(y_i, \phi) = -\frac{y_i^2}{2\phi} - \log(2\pi\phi)/2. \]

Thus the model fits into the framework of generalized linear models, see McCullagh and Nelder.
(1989), and the corresponding link function is the identity function. The distribution of the phenotype depends on the three model parameters $\alpha$, $\beta$ and $\phi$. We will in this appendix use the shorthand notation $\theta = (\alpha, \beta, \phi)$ and make the dependence on the parameters explicit by using $\theta$ as subscript, writing e.g. $f_\theta(y_i|x_i, g_i)$.

If the considered phenotype is e.g. case-control status, we model the distribution of the phenotype conditional on the genotype and the other covariates using a logistic regression model. Let $\pi_i = P(y_i = 1|x_i, g_i)$, then the log odds for a positive outcome is modeled by

$$\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \alpha x_i^\top + \beta g_i^\top.$$

The distribution of the phenotype can be written

$$f(y_i|x_i, g_i) = \pi_i^{y_i}(1 - \pi_i)^{1 - y_i}$$

$$= \exp(y_i \eta_i - \log(1 + \exp(\eta_i)))$$

$$= \exp \left( \frac{y_i \eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi) \right)$$

i.e. the density is from an exponential family with

$$a(\phi) = 1, \quad b(\eta) = \log(1 + \exp(\eta)) \quad \text{and} \quad c(y_i, \phi) = 0.$$

Again the model fits in the framework of generalized linear models and the corresponding link function is the logit function (McCullagh and Nelder, 1989). The framework of generalized linear regression models also includes binomial regression (the logistic regression described above is a simple variant hereof) as well as poisson regression, those models might be of interest for association testing for integer phenotypes.

Since the actual genotypes remains unobserved in next generation sequencing studies, we consider the joint distribution of the phenotype and the observed sequencing data. Assuming that the phenotype is conditionally independent of the sequencing data given the true genotype, the density of the joint distribution factorizes

$$p_\theta(y_i, D_i|x_i) = \sum_{g \in G} f_\theta(y_i|x_i, g)h(g, D_i)$$

where $G$ is the genotype state space and $h$ is shorthand notation for the density for the joint
distribution of the genotype and the observed sequencing data, i.e. \( h(g, D_i) = p(D_i | g) p(g | \hat{f}) \) as described in section 3.3. The log-likelihood function for the model with genotype uncertainties thus becomes

\[
l_{Y,\alpha, \beta, \phi} = \sum_{i=1}^{N} \log \left( \sum_{g \in G} \int_{x_i, \theta \in \Theta_i} f(y_i | x_i, \theta) h(g, D_i) \right).
\]

In this model, testing for association corresponds to testing the hypothesis

\[
H_0 : \beta = 0.
\]

In the following \( \hat{\theta} = (\hat{\alpha}, 0, \hat{\phi}) \) denotes the maximum likelihood estimate of the parameters under the null hypothesis. Under \( H_0 : \beta = 0 \) the density \( f \) no longer depends on the genotype, thus the likelihood simplifies and

\[
l_{Y,\alpha, 0, \phi} \propto \sum_{i=1}^{N} \log \left( f_{\alpha,0,\phi}(y_i | x_i) \right) = \sum_{i=1}^{N} y_i \eta_i - b(\eta_i) + c(y_i, \phi)
\]

Therefore the maximum likelihood estimates under \( H_0 \) are easily found by standard methods for the generalized linear model under consideration.

The null hypothesis is tested using the score test, which is based on the score function

\[
s_{Y,\alpha, \beta, \phi} = \left[ \begin{array}{c}
\partial_{\alpha} l_{Y,\alpha, \beta, \phi} \\
\partial_{\beta} l_{Y,\alpha, \beta, \phi} \\
\partial_{\phi} l_{Y,\alpha, \beta, \phi}
\end{array} \right],
\]

where \( \partial_{\alpha} = \left( \frac{\partial}{\partial \alpha_0}, \frac{\partial}{\partial \alpha_1}, \ldots, \frac{\partial}{\partial \alpha_d} \right)^T \), \( \partial_{\beta} = \left( \frac{\partial}{\partial \beta_1}, \ldots, \frac{\partial}{\partial \beta_g} \right)^T \) and \( \partial_{\phi} = \frac{\partial}{\partial \phi} \) is used as vector notation for the partial derivatives with respect to the parameters.

The observed information

\[
a_{Y,\alpha, \beta, \phi} = - \left[ \begin{array}{ccc}
\partial_{\alpha} \partial_{\alpha} l_{Y,\alpha, \beta, \phi} & \partial_{\alpha} \partial_{\beta} l_{Y,\alpha, \beta, \phi} & \partial_{\alpha} \partial_{\phi} l_{Y,\alpha, \beta, \phi} \\
\partial_{\beta} \partial_{\alpha} l_{Y,\alpha, \beta, \phi} & \partial_{\beta} \partial_{\beta} l_{Y,\alpha, \beta, \phi} & \partial_{\beta} \partial_{\phi} l_{Y,\alpha, \beta, \phi} \\
\partial_{\phi} \partial_{\alpha} l_{Y,\alpha, \beta, \phi} & \partial_{\phi} \partial_{\beta} l_{Y,\alpha, \beta, \phi} & \partial_{\phi} \partial_{\phi} l_{Y,\alpha, \beta, \phi}
\end{array} \right].
\]
serves as an estimate of the variation in the score function. This motivates the score statistic

\[ R(y, D) = s_{y, D}(\tilde{\alpha}, 0, \tilde{\phi})^T o_{y, D}(\tilde{\alpha}, 0, \tilde{\phi})^{-1} s_{y, D}(\tilde{\alpha}, 0, \tilde{\phi}) \]

which for large \( N \) is approximately \( \chi^2 \)-distributed with \( d_g \) degrees of freedom, see Bickel and Doksum (2001, Section 5.4). The approach taken in this appendix is slightly different from that taken by Schaid et al. (2002), where it is assumed that the exponential family under consideration is a single-parameter exponential family, thus ignoring the unknown variance parameter in the normal distribution. Considering the uncertainty of the estimated variance parameter, makes it a two-parameter exponential family, see (McCullagh and Nelder, 1989, Section 2.2.2). All derivations in the following is made in the general case allowing a two-parameter exponential family.

In the following, we derive an explicit expression for the score statistic in our model. Here the score function is given by

\[ s_{y, D}(\alpha, \beta, \phi) = \sum_{i=1}^{N} p_{\theta}(y, D, |x_i) \sum_{g \in G} f_{\theta}(y_i|x_i, g) \left[ \begin{array}{c} \frac{y_i - b'(\eta_i)}{a(\phi)} x_i^T \\ \frac{y_i - b'(\eta_i)}{a(\phi)} g^T \\ -\frac{y_i - b'(\eta_i)}{a(\phi)} a'(\phi) + c'(\phi) \end{array} \right] h(g, D_i) \]

where \( \eta_i = \eta(x_i, g) = \alpha x_i^T + \beta g^T \). The corresponding expression for the observed information is given by the negative second order derivatives of the likelihood, the second order derivatives are given in table S1. When evaluating the score function in \((\tilde{\alpha}, 0, \tilde{\phi})\), the maximum likelihood estimate under the null hypothesis, the expression simplifies to

\[ s_{y, D}(\tilde{\alpha}, 0, \tilde{\phi}) = \sum_{i=1}^{N} \frac{y_i - b'(\tilde{\alpha}x_i^T)}{a(\phi)} E[g^T | D_i] \]

where \( E[g^T | D_i] = \left( \sum_{g \in G} g^T h(g, D_i) \right) / \left( \sum_{g \in G} h(g, D_i) \right) \) is the posterior expectation of the genotype of individual \( i \). Likewise the observed information evaluated in \((\tilde{\alpha}, 0, \tilde{\phi})\) simplifies to

\[ o_{y, D}(\tilde{\alpha}, 0, \tilde{\phi}) = \begin{bmatrix} o_{\alpha \alpha} & o_{\alpha \beta} & 0 \\ o_{\beta \alpha} & o_{\beta \beta} & o_{\beta \phi} \\ 0 & o_{\phi \beta} & o_{\phi \phi} \end{bmatrix} \]
with

\[ o_{\alpha\alpha} = \sum_{i=1}^{N} \frac{b''(\hat{\alpha}x_i^\top)}{a(\hat{\phi})} x_i^\top x_i \]
\[ o_{\alpha\beta} = \sum_{i=1}^{N} \frac{b''(\hat{\alpha}x_i^\top)}{a(\hat{\phi})} x_i^\top E[g|D_i] \]
\[ o_{\beta\beta} = \sum_{i=1}^{N} \left( \frac{b''(\hat{\alpha}x_i^\top)}{a(\hat{\phi})} E[g^\top|D_i]E[g|D_i] - \frac{(y_i - b'(\hat{\alpha}x_i^\top))^2}{a(\hat{\phi})^2} \right) \]
\[ o_{\beta\phi} = \sum_{i=1}^{N} \frac{y_i - b'(\hat{\alpha}x_i^\top)}{a(\hat{\phi})^2} E[g^\top|D_i]a'(\hat{\phi}) \]
\[ o_{\phi\phi} = \sum_{i=1}^{N} (y_i \hat{\alpha}x_i^\top - b(\hat{\alpha}x_i^\top)) \left( \frac{a''(\hat{\phi}_i)}{a(\hat{\phi}_i)^2} - \frac{2a'(\hat{\phi}_i)^2}{a(\hat{\phi}_i)^3} \right) - c''(\hat{\phi}) \]

where \( \text{Cov}[g|D_i] = \left( \sum_{g \in G} g^\top g h(g, D_i) \right) / \left( \sum_{g \in G} h(g, D_i) \right) - E[g^\top|D_i]E[g|D_i] \). Thus an explicit expression of the score statistic in our particular model is given by

\[ R(y, D) = s_\beta^\top o_\beta^{-1} s_\beta \]

where \( s_\beta = \sum_{i=1}^{N} \frac{y_i - b'(\hat{\alpha}x_i^\top)}{a(\hat{\phi})} E[g^\top|D_i] \) and \( o_\beta = o_{\beta\beta} - o_{\beta\alpha} o_{\alpha\alpha}^{-1} o_{\alpha\beta} - o_{\beta\phi} o_{\phi\phi}^{-1} o_{\phi\beta} \). Note that this is identical to the statistic used by Schaid et al. (2002), except the last term, which is due to the unknown ‘second parameter’ of the exponential family.

When the phenotypes under consideration are case control status, modeled by logistic regression, there is no variance parameter in the model, thus the last term should be ignored. The statistic in this case is calculated using \( a(\hat{\phi}) = 1 \), \( b'(\hat{\alpha}x_i^\top) = \exp(\hat{\alpha}x_i^\top) / (1 + \exp(\hat{\alpha}x_i^\top)) \) and \( b''(\hat{\alpha}x_i^\top) = (1 + \exp(\hat{\alpha}x_i^\top))^{-2} \).

For quantitative traits, modeled by linear regression, the model contains an unknown variance parameter and the last term apply. Here the statistic is calculated using \( a(\hat{\phi}) = \hat{\sigma}^2 \), \( a'(\hat{\phi}) = 1 \), \( a''(\hat{\phi}) = 0 \), \( b'(\hat{\alpha}x_i^\top) = \hat{\alpha}x_i^\top \), \( b''(\hat{\alpha}x_i^\top) = 1 \) and \( c''(\hat{\phi}) = -y_i / 2\hat{\sigma}^6 + 1 / 2\hat{\sigma}^4 \).
B Supplemental Figures and Tables

Figure S1: Genotype calling for the CEU HapMap sample
Figure S2: Association study with population stratification. Top figures show the first two principal components of the true genotypes (left) and the posterior probabilities (right) using the eigensoft approach. The two populations are colored in red and black. The middle figures shows the qq plot of the p-values when performing association using either the true genotypes or using the score statistic on the posterior probabilities of the genotypes. The bottom figure also shows a qq plot but in this analysis the first eigen vector was included in the analysis as an additional covariate.
Figure S3: QQplot for the 1000 Genomes CEU population with simulated phenotypes. Normal distributed phenotypes were simulated for the 3200 over-sampled individuals. Association was carried out on all polymorphic sites in the HapMap project. The lambda is the inflation factor from the genomic control.
Figure S4: All plots are based on simulation of 1000 cases and 1000 controls. The minor allele frequency was set to 1% in the figure A and B and 2% in figure C and D. The average depth was set to 4 in figure A and C and 8 in figure B and D. The given frequencies are population frequencies. We assumed that the disease prevalence is 10%. Note the changed scale on the x axis.
Figure S5: All plots are based on 50,000 simulations of different numbers of individual and with equal numbers of cases and controls. The average depth was set to 4 and the population allele frequency to 10% without any genotype effect (RR=1). In figure D we exclude the analysis based on a threshold because of a too low numbers of inferred genotypes.
Figure S6: QQ plot of the score statistic based on the 1000 Genomes low coverage data. 32 CEPH and 41 YRI individuals were sampled 100 times and binary phenotypes were simulated for each of the 7300 individuals. The score statistic is shown in green for the scenario where the YRI and the CEPH have same proportion of cases and controls. In the orange and in the pink scenario the ratio of cases and controls are different for the YRI individuals than the CEPH but in the pink scenario the population affiliation for each individuals is included as an additional covariate. In the legend the number of cases and controls are written as cases:controls.
\[ \begin{align*}
\partial_\alpha \partial_\beta l_{y, D}(\alpha, \beta, \phi) &= \\
&= \sum_{i=1}^{N} \left(-p_{\phi}(y_i, D_i|x_i) \right)^{-2} \left( \sum_{g \in G} f_{0}(y_i, x_i, g) \frac{y_i - b'(\eta_i)}{a(\phi)} - x_i^T h(g, D_i) \right) \left( \sum_{g \in G} f_{0}(y_i, x_i, g) \frac{y_i - b'(\eta_i)}{a(\phi)} - x_i^T h(g, D_i) \right) \\
&+ p_{\phi}(y_i, D_i|x_i)^{-1} \sum_{g \in G} f_{0}(y_i, x_i, g) \left( \frac{(y_i - b'(\eta_i))^2}{a(\phi) - b''(\eta_i)} - b''(\eta_i) \right) x_i^T h(g, D_i) \\
\partial_\alpha \partial_\phi l_{y, D}(\alpha, \beta, \phi) &= \\
&= \sum_{i=1}^{N} \left(-p_{\phi}(y_i, D_i|x_i) \right)^{-2} \left( \sum_{g \in G} f_{0}(y_i, x_i, g) \frac{y_i - b'(\eta_i)}{a(\phi)} - x_i^T h(g, D_i) \right) \left( \sum_{g \in G} f_{0}(y_i, x_i, g) \left( \frac{-y_i b'(\eta_i)}{a(\phi)} - c'_{\phi}(\phi) \right) h(g, D_i) \right) \\
&+ p_{\phi}(y_i, D_i|x_i)^{-1} \sum_{g \in G} f_{0}(y_i, x_i, g) \left( \frac{(y_i - b'(\eta_i))^2}{a(\phi) - b''(\eta_i)} - b''(\eta_i) \right) x_i^T h(g, D_i) \\
\partial_\beta \partial_\phi l_{y, D}(\alpha, \beta, \phi) &= \\
&= \sum_{i=1}^{N} \left(-p_{\phi}(y_i, D_i|x_i) \right)^{-2} \left( \sum_{g \in G} f_{0}(y_i, x_i, g) \frac{y_i - b'(\eta_i)}{a(\phi)} - x_i^T h(g, D_i) \right) \left( \sum_{g \in G} f_{0}(y_i, x_i, g) \left( \frac{-y_i b'(\eta_i)}{a(\phi)} - c'_{\phi}(\phi) \right) h(g, D_i) \right) \\
&+ p_{\phi}(y_i, D_i|x_i)^{-1} \sum_{g \in G} f_{0}(y_i, x_i, g) \left( \frac{(y_i - b'(\eta_i))^2}{a(\phi) - b''(\eta_i)} - b''(\eta_i) \right) x_i^T h(g, D_i) \\
\partial_\alpha \partial_\beta l_{y, D}(\alpha, \beta, \phi) &= \\
&= \sum_{i=1}^{N} \left(-p_{\phi}(y_i, D_i|x_i) \right)^{-2} \left( \sum_{g \in G} f_{0}(y_i, x_i, g) \left( \frac{y_i - b'(\eta_i)}{a(\phi)} - x_i^T h(g, D_i) \right) \left( \sum_{g \in G} f_{0}(y_i, x_i, g) \left( \frac{-y_i b'(\eta_i)}{a(\phi)} - c'_{\phi}(\phi) \right) h(g, D_i) \right) \\
&+ p_{\phi}(y_i, D_i|x_i)^{-1} \sum_{g \in G} f_{0}(y_i, x_i, g) \left( \frac{(y_i - b'(\eta_i))^2}{a(\phi) - b''(\eta_i)} - b''(\eta_i) \right) x_i^T h(g, D_i) \\
\end{align*} \]

Table S1: The second order derivatives of the log likelihood function.