A new multipoint method for genome-wide association studies via imputation of genotypes : Supplementary Methods

Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, Peter Donnelly

Imputation of missing genotypes

The Methods section of the paper describes how missing genotypes are inferred through the use of a model of an individual's genotype vector G_i conditional upon a set of N known haplotypes H. A Hidden Markov Model (HMM) is used that has the form

$$Pr(G_i|H) = \sum_{Z_i^{(1)}, Z_i^{(2)}} Pr(G_i|Z_i^{(1)}, Z_i^{(2)}, H) Pr(Z_i^{(1)}, Z_i^{(2)}|H), \qquad (1)$$

where $Z_i^{(1)} = \{Z_{i1}^{(1)}, \dots, Z_{iL}^{(1)}\}$ and $Z_i^{(2)} = \{Z_{i1}^{(2)}, \dots, Z_{iL}^{(2)}\}$ are two sequences of hidden states at the L sites and $Z_{il}^{(j)} \in \{1, \dots, N\}$. At a given locus these hidden states can be thought of as the pair of haplotypes in the set H that are being copied at that locus to form the genotype vector G_i.

Here $Pr(Z_i^{(1)}, Z_i^{(2)}|H)$ defines the prior probability on how the sequences of hidden states, $Z^{(1)}$ and $Z^{(2)}$, change along the sequence. We use a Markov Chain model in which the switching rates depend upon an estimate of the fine-scale recombination map across the genome based upon the HapMap Phase II Data [1]. The initial state of the Markov chain is uniform on the N^2 states

$$Pr(Z_{i1}^{(1)}, Z_{i1}^{(2)}|H) = \frac{1}{N^2}. \qquad (2)$$

The transition probabilities of the chain from site l to $l+1$ are given by

$$Pr(\{Z_{il}^{(1)}, Z_{il}^{(2)}\} \rightarrow \{Z_{i(l+1)}^{(1)}, Z_{i(l+1)}^{(2)}\}|H) = \begin{cases} \left(e^{-\frac{\rho_l}{N}} + \frac{1-e^{-\frac{\rho_l}{N}}}{N}\right)^2 & Z_{il}^{(1)} = Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} = Z_{i(l+1)}^{(2)} \\ \left(e^{-\frac{\rho_l}{N}} + \frac{1-e^{-\frac{\rho_l}{N}}}{N}\right)\left(\frac{1-e^{-\frac{\rho_l}{N}}}{N}\right) & Z_{il}^{(1)} = Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} \neq Z_{i(l+1)}^{(2)} \\ & Z_{il}^{(1)} \neq Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} = Z_{i(l+1)}^{(2)} \\ \left(\frac{1-e^{-\frac{\rho_l}{N}}}{N}\right)^2 & Z_{il}^{(1)} \neq Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} \neq Z_{i(l+1)}^{(2)} \end{cases} \tag{3}$$

where $\rho_l = 4N_e r_l$ and r_l is the per generation genetic distance between sites l and $l+1$. We use the estimate of $N_e = 11,418$ [1]. Overall, the prior distribution on the hidden states can be written as

$$Pr(Z_i^{(1)}, Z_i^{(2)}|H) = Pr(Z_{i1}^{(1)}, Z_{i1}^{(2)}|H) \prod_{l=1}^{L-1} Pr(\{Z_{il}^{(1)}, Z_{il}^{(2)}\} \rightarrow \{Z_{i(l+1)}^{(1)}, Z_{i(l+1)}^{(2)}\}|H) \tag{4}$$

The term $Pr(G_i|Z_i^{(1)}, Z_i^{(2)}, H)$ models how the observed genotypes will be close to but not exactly the same as the haplotypes being copied. This term mimics the effects of mutation in the approximation to the population genetics model. We assume that the mutations are independent across sites and that the two alleles on the haplotype being copied (independently) mutate to their complementary alleles with probability $\lambda = \frac{\theta}{2(\theta+N)}$.

$$Pr(G_i|Z_i^{(1)}, Z_i^{(2)}, H) = \prod_{l=1}^{L} Pr(G_{il}|Z_{il}^{(1)}, Z_{il}^{(2)}, H) = \prod_{l=1}^{L} Pr((H_{Z_{il}^{(1)}l} + H_{Z_{il}^{(2)}l}) \rightarrow G_{il}) \tag{5}$$

where $Pr((H_{Z_{il}^{(1)}l} + H_{Z_{il}^{(2)}l}) \rightarrow G_{il})$ is given by Table 1. Following [2], we use $\theta = \left(\sum_{i=1}^{N-1} \frac{1}{i}\right)^{-1}$.

This model can also be used to infer haplotypes across a region of interest and to deal with uncertainty in the genotype data (to be described elsewhere).

Imputation of completely missing SNPs

To carry out a test of association at a SNP which is completely unobserved in both the set of haplotypes H and the set of sampled data G we simulate M realisations of this SNP in the N observed haplotypes in set H. This sample of SNPs is then

		G_{il}		
		0	1	2
$H_{Z_{il}^{(1)}l} + H_{Z_{il}^{(2)}l}$	0	$(1-\lambda)^2$	$2\lambda(1-\lambda)$	λ^2
	1	$\lambda(1-\lambda)$	$\lambda^2+(1-\lambda)^2$	$\lambda(1-\lambda)$
	2	λ^2	$2\lambda(1-\lambda)$	$(1-\lambda)^2$

Table 1: The probability $Pr((H_{Z_{il}^{(1)}l} + H_{Z_{il}^{(2)}l}) \rightarrow G_{il})$ of mutating from the genotype derived by summing the alleles defined by the two copying states $(H_{Z_{il}^{(1)}l} + H_{Z_{il}^{(2)}l})$ to the observed genotype G_{il}.

treated as a set of known sites in the set H and can be conditioned upon to simulate untyped variants in the set of data G. Frequentist or Bayesian test statistics can then be calculated by averaging over the sample in the appropriate way.

Suppose we wish to simulate a SNP at an unobserved site; let this site be the jth site of the haplotypes in the set H. We assume that H_{ij} is missing for all $i \in \{1, \ldots, N\}$. To simulate this missing data we model the joint distribution of the complete set of haplotypes H. That is,

$$Pr(H) = Pr(H_1)Pr(H_2|H_1)\ldots Pr(H_N|H_1, \ldots, H_{N-1}). \qquad (6)$$

Each of the conditional distributions $Pr(H_i|\cdot)$ is approximated by the Hidden Markov model described in [2] to give

$$Pr(H) \approx \pi(H_1)\pi(H_2|H_1)\ldots \pi(H_N|H_1, \ldots, H_{N-1}). \qquad (7)$$

This is known as a "product of approximate conditionals" (PAC) model. Given an ordering of the haplotypes $H_{(1)}, \ldots, H_{(N)}$ the missing alleles $H_{(1)j}, \ldots, H_{(N)j}$ are simulated sequentially using these approximate conditional distributions. We use this model to make the following approximation of the probability that $H_{(i)j} = k$, where $k \in \{0, 1\}$:

$$\begin{aligned} Pr(H_{(i)j} = k | H_{(i)} \setminus H_{(i)j}; H_{(1)}, \ldots, H_{(i-1)}) &\propto Pr(H_{(i)}, H_{(i)j} = k | H_{(1)}, \ldots, H_{(i-1)}) \qquad (8) \\ &\approx \pi(H_{(i)}, H_{(i)j} = k | H_{(1)}, \ldots, H_{(i-1)}). \qquad (9) \end{aligned}$$

The simulation is initialised by probabilistically choosing a haplotype to which to apply the first mutation. From the model the probability of a mutation to a different allele on the $(r+1)$th haplotype is $\frac{\theta}{2(r+\theta)}$. Therefore, if we use p_r to denote the probability of the first mutation occurring on the $(r+1)$th haplotype in the ordering then we have

$$p_1 = \frac{\lambda_1}{W}, \quad p_r = \frac{1}{W}\prod_{j=1}^{r-1}(1-\lambda_j)\lambda_r, \quad 2 \leq r < N, \tag{10}$$

where $\lambda_j = \frac{\theta}{2(r+\theta)}$ and $W = \sum_{j=1}^{N-1} p_j$.

A draw from this probability distribution determines which of the haplotypes carries the first mutation. If this is the mth haplotype of the ordering then haplotypes $H_{(1)j} = H_{(2)j} = \ldots H_{(m-1)j} = 0$ and $H_{(m)j} = 1$. Subsequent to this the alleles are simulated using higher order approximate conditionals as in (9). For example, $H_{(m+1)j}$ will be simulated from the distribution

$$Pr(H_{(m+1)j} = k) = \frac{1}{B}\pi(H_{(m+1)}, H_{(m+1)j} = k | H_{(1)}, \ldots, H_{(m)}) \quad k \in \{0, 1\} \tag{11}$$

where $B = \sum_{k=0}^{1} \pi(H_{(m+1)}, H_{(m+)j} = k | H_{(1)}, \ldots, H_{(m)})$

We have found that the use of a relatively small set of random haplotype orderings (10-20) is sufficient to provide a stable set of results. This process can be repeated M times to produce a sample of the variation that might exist at a given SNP in the set of haplotypes H. We then condition on this sample to further simulate the variation in the set G. Currently, our implementation of this approach requires the set G to consist of haplotype data. When computing the required probabilities to carry out the simulation we take advantage of the standard recursive calculations that exist for HMMs.

Single SNP disease association

In this section we provide a comprehensive account of the different ways a test of association at a SNP with known genotypes may be carried out. Also, since our methodology carries out tests of association at imputed SNPs that are characterised by a probability distribution on genotypes, we consider how this uncertainty can be taken into account when carrying out a test. For completeness and comparability we present details for both Frequentist and Bayesian tests of association. We focus here on the case where we have a binary phenotype and no other covariates of interest, but our approach can be naturally and simply extended to handle continuous phenotypes and covariates.

Consider a SNP with two alleles coded 0 and 1. Suppose we have genotypes at this SNP in a set of N individuals (N_1 cases and N_2 controls). We use Y_i to denote the binary disease phenotype of individual i (cases have Y_i = 1, controls have Y_i = 0). Let $Z_i \in \{0, 1, 2\}$ denote the genotype of individual i for the given coding of the two alleles at the SNP. The data at the SNP can be summarised in the following table

Z	0	1	2
Cases	s_0	s_1	s_2
Controls	r_0	r_1	r_2

Frequentist Association Tests

The most widely used method of testing association at a SNP employs a model in which the odds of disease change multiplicatively with genotype. This model is specified using a logistic regression framework:

$$L(\theta) = P(Y|Z, \mu, \gamma) = \prod_{i=1}^{N} p_i^{Y_i}(1 - p_i)^{1-Y_i} \qquad (12)$$

where

$$\theta = (\mu, \gamma) \qquad \log \frac{p_i}{1 - p_i} = \mu + \gamma Z_i \quad p_i = \frac{e^{\mu+\gamma Z_i}}{1 + e^{\mu+\gamma Z_i}}. \qquad (13)$$

In this model μ is the baseline log-odds of disease for the 0 genotypes, γ specifies the increase in log-odds due to each copy of the allele coded 1 and p_i is the probability that individual i develops the disease. The odds ratios of disease for individuals with genotypes 1 and 2 (relative to individuals with the 0 genotype) are e^{γ} and $(e^{\gamma})^2$ respectively. This model is multiplicative on the odds scale and additive on the log-odds scale.

This is known as a prospective likelihood, but the natural likelihood for case-control studies is the retrospective likelihood in which genotypes are modelled conditional upon disease status. It has been shown [3] that (in the absence of missing data) the maximum likelihood estimators and asymptotic covariance matrix of the log-odds ratios obtained from the retrospective likelihood are the same as those obtained from the prospective likelihood. This implies that the usual significance tests used in either framework will be equivalent asymptotically and very similar for large enough sample sizes. Intuitively this makes sense, as the main parameters of interest that model the differences in genotype proportions between cases and controls will not be significantly altered by the over-sampling of cases that occurs in a case-control design. We do lose the ability to estimate the prevalence of the disease, but that is never a primary focus of the analysis.

Estimates of the parameters of the model can be obtained by maximising the likelihood. In general, iterative optimisation techniques are required. One such technique is the Newton-Raphson method [4] which has the following updates

$$\theta^{t+1} = \theta^{t} - H^{-1}(\theta^{t})U(\theta^{t}) \qquad (14)$$

where

$$U(\theta) = \frac{d(\theta)}{d\theta} \qquad H(\theta) = \frac{d^2(\theta)}{d\theta^2}. \qquad (15)$$

In the above equations $(\theta) = \log L(\theta)$, $U(\theta)$ is known as the Score and $H(\theta)$ is the Hessian. To test for association we can test the hypothesis

$$H_0 : \gamma = 0 \quad \text{vs} \quad H_1 : \gamma \neq 0 \qquad (16)$$

using a Maximum Likelihood Ratio Test (MLRT) [4] of the form

$$\lambda = \frac{\text{Max}_{H_0} L(\theta)}{\text{Max}_{H_0 \cup H_1} L(\theta)} \quad \text{where} \quad -2\log\lambda \sim \chi^2_1 \,\text{as}\, N_1, N_2 \to \infty. \qquad (17)$$

An alternative test statistic is known as the Score Test [5, 6, 7] and is based on the distribution of the Score under H_0. Intuitively, if the MLE of the parameter of interest is far from (close to) the null then the Score will tend to be far from (close to) 0. More specifically, the test is based on the following asymptotic result

$$U(\theta_0) \sim N(0, I(\theta_0)) \text{ when } H_0 \text{ is true,} \tag{18}$$

where $I(\theta_0) = -H(\theta_0)$ is the Information matrix evaluated at θ_0. This result implies that the Score Statistic, defined as

$$S = U(\theta_0)^T I^{-1}(\theta_0) U(\theta_0) \tag{19}$$

is asymptotically distributed as χ^2_d where $d = \text{dim}(\theta)$ [5]. This test is convenient in that it only requires evaluation of the likelihood under the null hypothesis, not maximisation of the log-likelihood, and thus is more computationally tractable than the MLRT.

If the parameter θ is multivariate i.e $\theta = (\mu, \gamma)$ and the null is of the form $H_0 : \gamma = \gamma_0$, the score and information matrix should be evaluated at the MLE under the null i.e. at $\theta = (\hat{\mu}, \gamma_0)$ where $\hat{\mu}$ is the MLE of μ with γ fixed at γ_0.

For the model above it can be shown that

$$\begin{aligned} U(\theta) &= \sum_{i=1}^{N} (Y_i - p_i)(1 Z_i)^T, & (20) \\ H(\theta) &= -\sum_{i=1}^{N} p_i(1-p_i)(1 Z_i)(1 Z_i)^T. & (21) \end{aligned}$$

The MLE of μ when $\gamma = 0$ is $\hat{\mu} = \log \frac{N_1}{N_2}$ and $p_i = \frac{N_1}{N} \forall i$. This implies

$$\begin{aligned} U(\theta_0) &= (0, \frac{N_2}{N}(s_1 + 2s_2) - \frac{N_1}{N}(r_1 + 2r_2))^T & (22) \\ I(\theta_0) &= \frac{N_1 N_2}{(N)^2} \begin{pmatrix} N & s_1 + r_1 + 2(s_2 + r_2) \\ s_1 + r_1 + 2(s_2 + r_2) & s_1 + r_1 + 4(s_2 + r_2) \end{pmatrix}. & (23) \end{aligned}$$

The Score Test Statistic is $S = U_\gamma^T I^{-1} U_\gamma$ where

$$\begin{aligned} U_\gamma &= U(\theta_0)_\gamma = \frac{N_2}{N}(s_1 + 2s_2) - \frac{N_1}{N}(r_1 + 2r_2) & (24) \\ I_\gamma &= I(\theta_0)_{\gamma\gamma} - I(\theta_0)_{\gamma\mu} I^{-1}(\theta_0)_{\mu\mu} I(\theta_0)_{\mu\gamma} & (25) \\ &= s_1 + r_1 + 4(s_2 + r_2) - (s_1 + r_1 + 2(s_2 + r_2))^2/(N_1 + N_2) & (26) \end{aligned}$$

thus

$$S = \frac{N(N_2(s_1+2s_2) - N_1(r_1+2r_2))}{N_1N_2(s_1+r_1+4(s_2+r_2) - (s_1+r_1+2(s_2+r_2))^2)}. \tag{27}$$

This test statistic is equivalent to the well known Cochran-Armitage Trend Test, which is a popular statistic used to compare genotype frequencies between cases and controls [8].

A dominant or recessive model can be fitted in the same way by changing the coding of the genotype vector Z. For a dominant model the score test statistic becomes

$$S = \frac{N(N_2(s_1+s_2) - N_1(r_1+r_2))}{N_1N_2(s_1+r_1+s_2+r_2 - (s_1+r_1+s_2+r_2)^2)}. \tag{28}$$

For a recessive model the score test statistic becomes

$$S = \frac{N(N_2s_2 - N_1r_2)}{N_1N_2(s_2+r_2 - (s_2+r_2)^2)}. \tag{29}$$

It is important to note that the above formulae assume a specific coding where one allele is coded 1 and the other 0 and that different results may be obtained in these tests if the coding is switched.

A general 3-parameter model can also be considered with the form

$$\log \frac{p_i}{1-p_i} = \mu + \gamma I(Z_i = 1) + \phi I(Z_i = 2) \tag{30}$$

where $I(Z_i = z)$ is the indicator function of the genotype Z_i equalling z where $z \in \{1, 2\}$. Under the null hypothesis $\gamma = \phi = 0$ and a score test can be derived as above, leading to the test statistic

$$S = U^T I^{-1} U \tag{31}$$

where

$$\begin{aligned} U &= \Big(s_1 - \frac{N_1}{N}(s_1+r_1), s_2 - \frac{N_1}{N}(s_2+r_2)\Big)^T && (32) \\ I &= \frac{N_1N_2}{N^3}\begin{pmatrix} N(s_1+r_1) - (s_1+r_1)^2 & -N(s_1+r_1)(s_2+r_2) \\ -N(s_1+r_1)(s_2+r_2) & N(s_2+r_2) - (s_2+r_2)^2 \end{pmatrix} && (33) \end{aligned}$$

Dealing with missing or uncertain genotypes

When some of the genotypes at a SNP are missing or when there is some uncertainty (specified by a probability distribution) as to the correct genotype at a SNP, there are three possible ways in which a test can be carried out. The simplest strategy is to apply a threshold rule to the probability distribution of each SNP. For example, we might choose to use only those genotypes for which a maximum posterior genotype call is greater than some value α. This procedure is simple and quick, and when there is considerable confidence in the genotype calls this method will work well. The problem with this method occurs when there is considerable uncertainty about the genotype calls, which may lead to very little data being used at a given SNP.

An alternative procedure would be to estimate the expected genotype counts at the SNP to produce the following table.

Z	0	1	2
Cases	$\mathbb{E}_{Y_M\|Y_O,\theta}[s_0]$	$\mathbb{E}_{Y_M\|Y_O,\theta}[s_1]$	$\mathbb{E}_{Y_M\|Y_O,\theta}[s_2]$
Controls	$\mathbb{E}_{Y_M\|Y_O,\theta}[r_0]$	$\mathbb{E}_{Y_M\|Y_O,\theta}[r_1]$	$\mathbb{E}_{Y_M\|Y_O,\theta}[r_2]$

This procedure makes use of all of the data at a SNP and, as with the threshold rule described above, will work well when there is high certainty about a SNP. This method makes no allowance for the variability in the genotype counts, so it does not completely account for the uncertainty in the genotypes.

To fully account for the uncertainty in genotypes we need to use well established statistical theory for missing data problems. Suppose we wish to fit a model at a given SNP in a case-control sample but we find that some (or all) of the genotypes are missing at the SNP. Further suppose we have additional non-missing data in other individuals at the SNP and/or data at other SNPs in the same set of individuals. We can partition this data structure into two components, observed data Y_O and missing data Y_M, and we use $Y_F = (Y_O, Y_M)$ to denote the full data. In this situation, the correct likelihood to consider is the observed data likelihood given by

$$l^*(\theta; Y_O) = \log P(Y_O|\theta) = \log \int P(Y_O, Y_M|\theta) dY_M \tag{34}$$

which is the log of the full data likelihood integrated over the missing data [9]. A score test can be computed for this likelihood through the calculation of the score and information matrix of the observed data likelihood

$$U^*(\theta) = \frac{d^*(\theta)}{d\theta} \qquad I^*(\theta) = -\frac{d^{2*}(\theta)}{d\theta^2}. \tag{35}$$

It can be shown [10, 9] that

$$\begin{aligned} U^*(\theta) &= \mathbb{E}_{Y_M|Y_O,\theta}[U(\theta)] & (36) \\ I^*(\theta) &= \mathbb{E}_{Y_M|Y_O,\theta}[I(\theta)] - V_{Y_M|Y_O,\theta}[U(\theta)] & (37) \end{aligned}$$

where $U(\theta)$ and $I(\theta)$ are the full data score and information. In this case, the score statistic would then be $S^* = U^*(\theta_0)^T (I^*(\theta_0))^{-1} U^*(\theta_0)$ where θ_0 is the value of the parameter vector specified by the null hypothesis.

These formulae show that where there is uncertainty in the data at a given SNP, the correct likelihood-based procedure involves using the distribution of the missing data conditional upon both the observed data and the values of the model parameters. This implies that we need to generate a family of distributions for the missing data indexed by θ; however, score tests are based on evaluations of the score and information under the null, so we only need to consider a single distribution specified by the null hypothesis.

Bayesian Association Tests

In a Bayesian framework the Bayes Factor is the alternative to the classical hypothesis tests described above [11]. Given two possible models M_1 and M_0 the Bayes Factor is defined as

$$BF = \frac{P(D|M_1)}{P(D|M_0)} = \frac{\int P(D|\theta_1, M_1) P(\theta_1|M_1) d\theta_1}{\int P(D|\theta_0, M_0) P(\theta_0|M_0) d\theta_0} \tag{38}$$

where D is used to denote the data and θ_1 and θ_0 are the parameters of the models M_1 and M_0. The Bayes Factor should be interpreted as the factor by which the prior odds of association are changed in light of the data to produce the posterior odds of association,

$$\text{Posterior Odds of Association} = BF \times \text{Prior Odds of Association}. \tag{39}$$

It can be clearly seen that instead of maximising the likelihood under the two models the parameters are integrated out of the likelihood with a weighting given by the prior distribution on the parameters. A main advantage of calculating a Bayes Factor as opposed to a classical test statistic is that the approach allows the incorporation of other relevant information through the use of the prior. For example, in our setting we may have good reason to believe that an additive odds ratio of 1.3 is much more plausible for a disease variant than an odds ratio of 50. We can incorporate this information into the prior distribution to improve the inference we obtain just from the use of the data alone. Some evidence is already emerging in the literature that these tests can have more power than their frequentist equivalents[12].

In our setting of testing for association at a given SNP, we use M_1 to denote the model in which the SNP is associated with an additive effect on the log-odds scale and M_0 to denote the 'null' model of no association.
For both models we use a logistic regression model for the likelihood

$$P(D|\theta) = \prod_{i=1}^{N} p_i^{Y_i}(1-p_i)^{1-Y_i} \tag{40}$$

where for model M_1 we have

$$\theta_1 = (\mu, \gamma) \qquad \log\frac{p_i}{1-p_i} = \mu + \gamma Z_i, \tag{41}$$

and for model M_0 we have

$$\theta_0 = (\mu) \qquad \log\frac{p_i}{1-p_i} = \mu. \tag{42}$$

We now need to specify the prior distribution $P(\theta_1|M_1) = P(\mu, \gamma|M_1)$. The parameter μ represents the baseline odds of disease. This parameter will be influenced by the numbers of cases and controls in the dataset. In a case-control design the numbers of cases in the sample have been elevated artificially, which will have a large effect on likely values of μ. For this reason we wish to use a prior distribution that allows flexibility in the prior distribution on μ, so we use a $N(\alpha_1, \beta_1)$ distribution. In practice we have used $\mu \sim N(0, 1)$.

The parameter γ is the increase in log-odds of disease for every copy of the risk allele and e^γ is the additive model odds ratio. We have some good prior information about likely values of this parameter. For example, it is widely believed that the genetic variants underlying common disease will have risk allele odds-ratios in the range 1-2 with substantially more weight on the values between 1-1.5. Note that this implies a protective allele odds-ratio in the range 0.5-1 with substantially more weight on values between 0.67-1. After some experimentation we settled on a flexible prior distribution of $N(\alpha_2, \beta_2)$ for γ. For example, Figure 1 shows a density plot for e^γ from a sample of 1,000,000 draws from the prior $\gamma \sim N(0, 0.2)$.

Overall, the prior distribution on the parameters has the form

$$P(\theta_1|M_1) \propto \frac{1}{\beta_1} e^{-\frac{(\mu-\alpha_1)^2}{2\beta_1^2}} \frac{1}{\beta_2} e^{-\frac{(\gamma-\alpha_2)^2}{2\beta_2^2}}. \tag{43}$$

For $P(\theta_0|M_0) = P(\mu|M_0)$ we use the same prior on μ as in the model M_1. That is,

$$P(\theta_0|M_0) \propto \frac{1}{\beta_1} e^{-\frac{(\mu-\alpha_1)^2}{2\beta_1^2}}. \tag{44}$$

It is well understood that the priors on the parameters of the model can have a non-negligible impact on the value of the Bayes Factor [11] even as the amount of data gets large. In line with this we have found that using different priors on μ for the two models can substantially change the Bayes Factor. We have little strong prior information about μ, and as noted above the case-control ratio will have a large effect on the values that best fit the data. For these reasons we use a reasonably diffuse prior distribution on this parameter that is the same for both models. This acts to focus the comparison between the models on the parameter γ, which is the main parameter of interest.

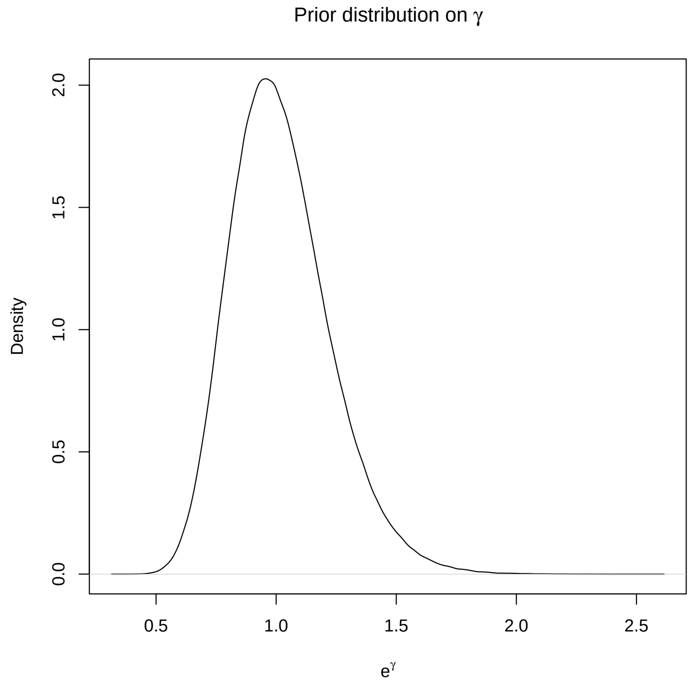

Figure 1: Density plot of the empirical distribution of e^{γ} from a sample of size 10^6 from the distribution $\gamma \sim N(0, 0.2)$

To evaluate the marginal likelihood for the model $P(D|M_1)$ we need to evaluate the integral

$$\int P(D|\theta_1, M_1)P(\theta_1|M_1)d\theta_1. \tag{45}$$

We do this using a Laplace Approximation [13] in which the posterior distribution is approximated using a Gaussian distribution centred on its mode. More specifically, we use

$$\log P(D|M_1) \approx \log P(D|\hat{\theta}_1, M_1) + \log P(\hat{\theta}_1|M_1) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|A| \tag{46}$$

where $\hat{\theta}_1$ is the value of θ_1 that maximises $P(D|\theta_1, M_1)P(\theta_1|M_1)$, and is known as the *maximum a posteriori* (MAP) estimate of θ_1. Also, A is the negative Hessian of $P(D|\theta_1, M_1)P(\theta_1|M_1)$ evaluated at $\hat{\theta}_1$ and d is the dimension of θ_1. We use Newton-Raphson optimisation to find $\hat{\theta}_1$ but if this fails to converge we use a line-search method. Both approaches are numerically efficient for this low-dimensional integral.

In addition, we note that the evaluation of this marginal likelihood will depend upon the way the alleles at the SNP have been coded 0 and 1. Thus, to calculate the marginal likelihood for the additive model we average over the two possible codings with equal weight.

For dominant and recessive models (denoted as M_2 and M_3 respectively) the required marginal likelihoods can be calculated in a similar fashion. Essentially the only difference is that the genotype vector Z is re-coded to indicate the dominant or recessive nature of the SNP. The parameter γ now denotes the increase in log-odds due to the dominant or recessive effect of the risk allele. In practice we use a $N(0, 0.5)$ prior for γ to reflect our beliefs that we might expect a slightly bigger genetic effect from a dominant or recessive model of disease. This results in a prior distribution for the odds-ratio e^γ as shown in Figure 2. As with the additive model above we average the marginal likelihood over the two possible codings for the SNP.

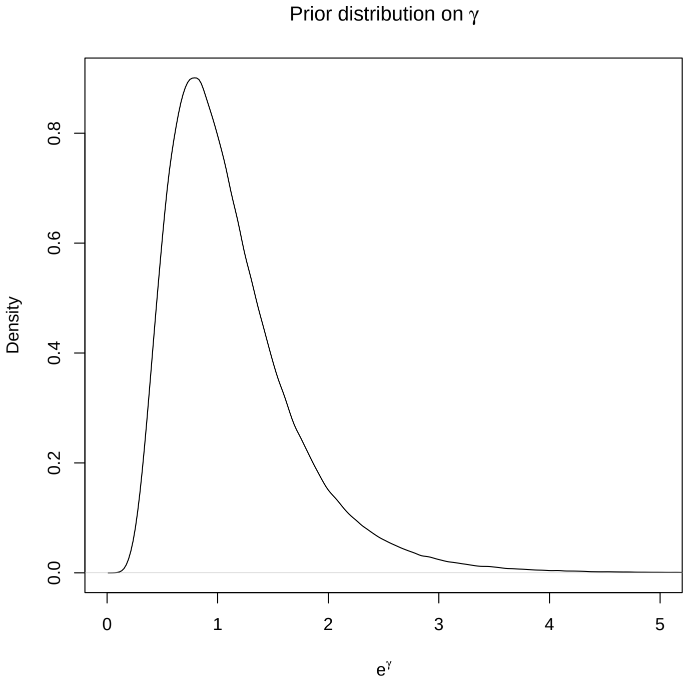

Figure 2: Density plot of the empirical distribution of e^γ from a sample of size 10^6 from the distribution $\gamma \sim N(0, 0.5)$

The general 3-parameter model (denoted M_4) is slightly more complicated in that we require a prior distribution on the additional parameter. We use the following model for the log-odds

$$\log \frac{p_i}{1-p_i} = \mu + \gamma I(Z_i = 1) + 2\phi\gamma I(Z_i = 2) \tag{47}$$

which has an additive genetic effect parametrised by γ and then an additional recessive effect parametrised by ϕ. In this model the additive model occurs when $\phi = 1$. We use a Gaussian prior, $N(\alpha_3, \beta_3)$ for ϕ. In practice we use a $N(1, 1)$ for ϕ which results in a symmetric departure from the additive model, and we use the same prior for γ i.e. $N(0, 0.2)$ as we did above when we considered the additive model. As with the other models above we average the marginal likelihood over the two possible codings for the SNP.

We have also implemented other priors that are more computationally efficient[12]. For the general 3-parameter model if we use the formulation

$$\log \frac{p_i}{1-p_i} = \mu I(Z_i = 0) + \gamma I(Z_i = 1) + \phi I(Z_i = 2) \tag{48}$$

in which each genotype is given its own log-odds parameter then the likelihood can be re-written as

$$P(D|\theta_4, M_4) = p_0^{s_0}(1-p_0)^{r_0} p_1^{s_1}(1-p_1)^{r_1} p_2^{s_2}(1-p_2)^{r_2} \tag{49}$$

where $p_0 = \frac{e^{\mu}}{1+e^{\mu}}$, $p_1 = \frac{e^{\gamma}}{1+e^{\gamma}}$ and $p_2 = \frac{e^{\phi}}{1+e^{\phi}}$. This has the form of an independent Binomial Likelihood for each of the three penetrance parameters p_0, p_1 and p_2. A conjugate Beta prior for these parameters can then be used which facilitates the exact calculation of the integrals. That is if we let

$$P(\theta_4|M_4) = \prod_{g=0}^{2} \frac{1}{\beta(\psi_g, \eta_g)} p_g^{\psi_g - 1}(1-p_g)^{\eta_g - 1} \tag{50}$$

where $\beta(\psi_g, \eta_g) = \frac{\Gamma(\psi_g)\Gamma(\eta_g)}{\Gamma(\psi_g + \eta_g)}$ then

$$P(D|M_4) = \prod_{g=0}^{2} \frac{\beta(s_g + \psi_g, r_g + \eta_g)}{\beta(\psi_g, \eta_g)}. \tag{51}$$

In a similar way the marginal likelihoods for dominant and recessive models using this class of conjugate priors are

$$P(D|M_2) = \frac{\beta(s_0 + \psi_0, r_0 + \eta_0)}{\beta(\psi_0, \eta_0)} \frac{\beta(s_1 + s_2 + \psi_1, r_1 + r_2 + \eta_1)}{\beta(\psi_1, \eta_1)} \tag{52}$$

and

$$P(D|M_3) = \frac{\beta(s_0 + s_1 + \psi_0, r_0 + r_1 + \eta_0)}{\beta(\psi_0, \eta_0)} \frac{\beta(s_2 + \psi_1, r_2 + \eta_1)}{\beta(\psi_1, \eta_1)} \tag{53}$$

respectively, where Beta(ψ_0, η_0) and Beta(ψ_1, η_1) priors are used for the baseline and dominant/recessive effect.

For the null model M_0 of no association we obtain a marginal likelihood of

$$P(D|M_0) = \frac{\beta(s_0 + s_1 + s_2 + \psi_0, r_0 + r_1 + r_2 + \eta_0)}{\beta(\psi_0, \eta_0)} \tag{54}$$

where a Beta(ψ_0, η_0) is used for the baseline penetrance.

It is interesting to consider what the conjugate Beta priors on penetrance actually mean in terms of odds-ratios. It can be shown that a $Beta(a, b)$ prior on a probability p is equivalent to a Generalised Logistic distribution on the log-odds $\log \frac{p}{1-p}$ [14] with mean $\Psi^{(0)}(a) - \Psi^{(0)}(b)$ and variance $\Psi^{(1)}(a) + \Psi^{(1)}(b)$ where $\Psi^{(r)}$ is the polygamma function. For example, a uniform distribution, $p \sim Beta(1, 1)$, results in a distribution for log-odds centred on 0 with a variance of $\pi^2/3$. This implies that the prior on the difference in log-odds between the heterozygote genotype and the baseline homozygote genotype has mean 0 and variance $2\pi^2/3$. This is considerably more diffuse than the $N(0, 0.2)$ prior we use in the additive model above. Using simulation from this prior we found that it corresponds to a prior distribution on the risk-allele odds ratio with a mean of approximately 80, which is rather larger than might be expected for common human diseases. This suggests that for the General, Dominant and Recessive models in which Beta priors are applicable it might be more reasonable to set the hyper-parameters a and b to be greater than 1. This would bring the priors closer to the non-conjugate priors we have suggested above.

To illustrate the sensitivity of Bayes Factors to the priors used on the parameters we analysed the dataset used in [12] (see table) using both sets of conjugate

Z	0	1	2
Cases	$s_0 = 90$	$s_1 = 60$	$s_2 = 20$
Controls	$r_0 = 50$	$r_1 = 70$	$r_2 = 50$

and non-conjugate priors described above. Using our priors we get Bayes Factors for the general 3-parameter model of 12,534 whereas the conjugate priors result in a Bayes Factor of 8,000. [12] took the view that of the order of 1,000 SNPs might be associated with a given disease and are there of the order of 10,000,000 SNPs in the human genome giving a prior odds of association of 1/10,000. So a Bayes Factor of more than 10,000 is required in order for posterior odds of association at a SNP to be greater than 1. The use of our priors results in a posterior odds of 1.25 whereas the conjugate priors result in a posterior odds of just 0.8. For the additive genetic model we get a Bayes Factor of 29,996 which results in a posterior odds of 3.

The Bayes Factors described above are based upon the “prospective” likelihood but the natural likelihood for case-control studies is the “retrospective” likelihood in which genotypes are modelled conditional upon disease status [3]. It will be relatively straightforward to develop Bayes Factors for a retrospective likelihood for single-SNP association along the lines described. The incorporation of covariate information into this framework is an area that needs further work but it seems clear that the prospective likelihood has the distinct advantage that covariates are more easily dealt with, while the development of Bayes Factors in the retrospective likelihood setting may be more challenging.

Dealing with missing or uncertain genotypes

As with the frequentist association tests above missing or uncertain genotypes may be handled by thresholding or using expected genotype counts. These methods will be accurate and equivalent when there is high certainty in the genotypes. The Bayesian solution that correctly accounts for the uncertainty is a little more complicated. In the Bayesian framework we need to calculate marginal likelihoods like $P(D|M_1)$ in which the parameters of the model have been integrated out over the prior. In a logistic regression setting this integral is more correctly written as

$$P(D|M_1) = P(Y|Z; M_1) = \int P(Y|Z; \theta_1, M_1)P(\theta_1|M_1)d\theta_1. \tag{55}$$

If there is missing genotype data then we can partition the genotype data into an observed and missing component, $Z = \{Z_O, Z_M\}$ and should calculate the marginal likelihood for the observed data. That is,

$$P(Y|Z_O; M_1) = \int P(Y|Z_O; \theta_1, M_1)P(\theta_1|M_1)d\theta_1 \tag{56}$$

which can be written as

$$P(Y|Z_O; M_1) = \int\int P(Y|Z_O, Z_M; \theta_1, M_1)P(Z_M|Z_O; \theta_1, M_1)P(\theta_1|M_1)dZ_M d\theta_1. \tag{57}$$

This is a rather complex integral to evaluate. If we make the approximation that $P(Z_M|Z_O; \theta_1, M_1)$ does not depend upon the model or the model parameters θ_1 then we can rearrange the order of integration and integrate out θ_1 to leave an integral of full data marginal likelihoods over the prior of the missing data conditional on the observed data

$$P(Y|Z_O; M_1) = \int P(Y|Z_O, Z_M; M_1)P(Z_M|Z_O)dZ_M. \tag{58}$$

The accuracy of this approximation will depend upon the validity of our assumption that $P(Z_M|Z_O; \theta_1, M_1)$ does not depend upon the model or the model parameters θ_1. In the setting of this paper the missing data are the missing genotypes in the study, the observed data are the genotypes at the genotyped SNPs together with the known HapMap haplotypes and fine-scale recombination rates and θ_1 are the disease model parameters. We think this is a reasonable assumption since we have observed that the missing genotypes are very accurately imputed conditional upon the observed data using a prediction model that is independent of disease model parameters.

In practice the integral in (58) is difficult to evaluate exactly so we use a Monte Carlo approximation

$$P(Y|Z_O;M_1)=\frac{1}{S}\sum_{i=1}^{S}P(Y|Z_O,Z_M^{(i)};M_1) \tag{59}$$

where the $Z_M^{(i)}$ are a sample of size S from the distribution $P(Z_M|Z_O)$. The accuracy of this approximation depends to a large extent upon the variability in the distribution $P(Z_M|Z_O)$ which as we mention above is generally low so we use relatively few samples i.e. $S=100$.

Applying the above approximation to both the numerator and the denominator of the Bayes Factor we get

$$BF=\frac{P(D|M_1)}{P(D|M_0)}=\frac{\sum_{i=1}^{S}P(Y|Z_O,Z_M^{(i)};M_1)}{\sum_{i=1}^{S}P(Y|Z_O,Z_M^{(i)};M_0)}. \tag{60}$$

For our specific M_0 of no association, the marginal likelihood does not depend on the genotype data and is a constant. This means that the Bayes Factor for the SNP can be re-written as the mean of the Bayes Factors applied to the sample of the full data,

$$\begin{aligned}BF &= \frac{\sum_{i=1}^{S}P(Y|Z_O,Z_M^{(i)};M_1)}{\sum_{i=1}^{S}P(Y|M_0)} && (61)\\ &= \frac{1}{S}\sum_{i=1}^{S}\frac{P(Y|Z_O,Z_M^{(i)};M_1)}{P(Y|M_0)} && (62)\\ &= \frac{1}{S}\sum_{i=1}^{S}BF_{M_1 vs M_0}(Z_M^{(i)}). && (63)\end{aligned}$$

Combining Bayes Factors across SNPs and models

In a similar way that we average Bayes Factors across realisations of a given SNP to produce a Bayes Factor for that SNP we can also average Bayes Factors across SNPs to produce a Bayes Factor for a region. This method has been suggested in a similar context as a way of summarizing information across a set of markers [15]. The extra information needed is a prior on SNPs. Suppose we have a region consisting of the set of W SNPs $\{S_1,\ldots,S_W\}$ and a prior $P(S_i)$ on each SNP.

Then the Bayes Factor for the region is given by

$$BF_{\text{region}} = \frac{\sum_{i=1}^{W} P(D|M_1; S_i)P(S_i)}{\sum_{i=1}^{W} P(D|M_0; S_i)P(S_i)}. \tag{64}$$

In our setting, $P(D|M_0; S_i)$ is a constant so this reduces to

$$BF_{\text{region}} = \sum_{i=1}^{W} BF(S_i)P(S_i), \tag{65}$$

which is a weighted sum of Bayes Factors across SNPs. For the analysis in our paper we use a uniform prior across SNPs when calculating the region Bayes Factors for our simulation study. In a similar way it is possible to average across different models at a given SNP by averaging the Bayes Factors for each model weighted by a prior on each model.

References

[1] The International HapMap Consortium. The Phase II Hapmap. *in preparation*, 2007.

[2] N. Li and M. Stephens. Modelling linkage disequilibrium, and identifying recombination hotspots using snp data. *Genetics*, 165:2213–2233, 2003.

[3] R. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.

[4] P. Garthwaite, I. Jolliffe, and B. Jones. *Statistical Inference*. Oxford University Press, 1995.

[5] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall, 1974.

[6] D. J. Schaid, C. M. Rowland, D. E. Tines, R. M. Jackson, and G. A. Poland. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *AMerican Journal of Human Genetics*, 70:425–434, 2002.

[7] Dmitri V Zaykin, Peter H Westfall, S Stanley Young, Maha A Karnoub, Michael J Wagner, and Margaret G Ehm. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered*, 53(2):79–91, 2002.

[8] D. Clayton. Population association. In *Handbook of Statistical Genetics*, pages 519–540. 2001.

[9] J. A. Little and D. B. Rubin. *Statistical Analysis wih missing data*. John Wiley and Sons, New Jersey, 2nd edition edition, 2002.

[10] T. A. Louis. Finding the observed information when using the EM algorithm. *JRSS (B)*, pages 226–233, 1982.

[11] A. O'Hagan and J. Forster. *Bayesian Inference*. Arnold, 2004.

[12] D. J. Balding. A tutorial on statistical methods for population association studies (supplementary note). *Nature Reviews Genetics*, 7:781–791, 2006.

[13] M. I. Jordan, editor. *Learning in Graphical Models*. Kluwer Academic Publishers, 1998.

[14] W. Jong-Wuu, H. Wen-Liang, and H. Lee. Some moments and limit behaviours of the generalized logistic distribution with applications. *Proc. Natl. Sci, Counc. ROC(A)*, 24(1):7–14, 2000.

[15] Nick Patterson, Neil Hattangadi, Barton Lane, Kirk E Lohmueller, David A Hafler, Jorge R Oksenberg, Stephen L Hauser, Michael W Smith, Stephen J O'Brien, David Altshuler, Mark J Daly, and David Reich. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet*, 74(5):979–1000, 2004.