SNPTEST v2

Technical Details

Jonathan Marchini

March 22, 2010

Contents

1 Testing for association at imputed SNPs

1.1 Frequentist tests

1.1.1 Missing data likelihood theory

To fully account for the uncertainty in genotypes we need to use well established statistical theory for missing data problems. This theory partitions the 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

$$l^*(\theta; Y_O) = \log P(Y_O|\theta) = \log \int P(Y_O, Y_M|\theta) dY_M \tag{1}$$

which is the log of the full data likelihood integrated over the missing data. The score and information matrix of the observed data likelihood are given by [1,2]

$$U^*(\theta) = \frac{dl^*(\theta)}{d\theta} = \mathbb{E}_{Y_M|Y_O,\theta}[U(\theta)] \tag{2}$$

$$I^*(\theta) = -\frac{d^2 l^*(\theta)}{d\theta^2} = \mathbb{E}_{Y_M|Y_O,\theta}[I(\theta)] - V_{Y_M|Y_O,\theta}[U(\theta)] \tag{3}$$

where $U(\theta)$ and $I(\theta)$ are the full data score and information.

The full data likelihood is then given by

$$P(\Phi|G, H, \theta) = \prod_{i=1}^{N} \sum_{G_{ij}} P(\Phi|G_{ij}, \theta) P(G_{ij}|G, H). \tag{4}$$

The conditional distribution of the missing data given the observed data and the parameters, $P(Y_M|Y_O, \theta)$, which is needed to calculate the observed data score and information is given by

$$q_{ijk} = P(G_{ij} = k|\Phi, G, H, \theta) \propto P(\Phi_i|G_{ij} = k, \theta) P(G_{ij} = k|G, H). \tag{5}$$

In the context of SNP association if we let Φ_i denote the binary phenotype of the ith individual in a study of N samples with N_1 cases and N_2 controls. Also let $G_i \in \{0, 1, 2\}$ denote the genotype of the ith individual at the jth SNP. Also, let $p_{ijk} = P(G_{ij} = k|H, G)$ be the probability (obtained from imputation) that the genotype at the jth SNP of the ith individual is k, conditional upon the collected genotype data at typed SNPs, G, and the set of haplotype data used in the imputation process, H. Then it is usual to use a logistic regression model of the form

$$P(\Phi|G_{\cdot j}, \theta) = \prod_{i=1}^{N} p_i^{\Phi_i} (1 - p_i)^{1-\Phi_i} \tag{6}$$

where

$$\theta = (\mu, \gamma) \qquad \log \frac{p_i}{1 - p_i} = \mu + \gamma G_{ij} \quad p_i = \frac{e^{\mu + \gamma G_{ij}}}{1 + e^{\mu + \gamma G_{ij}}}. \tag{7}$$

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. The score and information matrix for this model are given by

$$U(\theta) = \sum_{i=1}^{N} (\Phi_i - p_i)(1\, G_{ij})^T, \tag{8}$$

$$I(\theta) = \sum_{i=1}^{N} p_i(1 - p_i)(1\, G_{ij})(1\, G_{ij})^T. \tag{9}$$

Dominant, recessive, heterozygote and general 2-parameter models of association can be dealt with by a small change to the above model and the in what follows.

It may also be the case that the phenotype is quantitative. In this case, an option is to model the phenotype using a Normal distribution,

$$\Phi_i|G_{ij},\theta \sim N(\mu+\gamma G_{ij},\sigma^2). \tag{10}$$

Ofcourse, the distribution of the phenotype for each genotype may not have a Normal distribution so this assumpton should be kept in mind when interpreting results. One option to make the test more robust is to transform the distribution of the phenotypes to a Normal distribution before the application of the test. Although not strictly the correct thing to do, when the effect size is very small, as is often the case in GWAS, this approach should work quite well. One option for handling other types of phenotype is to adopt the generalized linear model (GLM) framework. So for example, Poisson and Gamma regression could be used as ways of handling discrete phenotypes bounded at 0 and phenotypes in which the error distribution is not symmetric.

1.1.2 Score tests

One way of carrying out a test of association is to use a Score Test (option `-method score` in SNPTEST v2), which needs calculations of the observed data score and information matrix only under the null hypothesis, $H_0 : \theta = \theta_0$. For example, for a binary phenotype, if $H_0 : \gamma = 0$ then $\theta_0 = (\hat{\mu}, 0)$ where $\hat{\mu}$ is the MLE of μ with $\gamma = 0$ i.e. $\hat{\mu} = \log \frac{N_1}{N_2}$. Also, in this case, $p_i = \frac{N_1}{N}$ and

$$P(G_{ij} = k|\Phi, G, H, \theta) = P(G_{ij} = k|G, H) = p_{ijk} \tag{11}$$

so that

$$U^*(\theta_0) = (0\ \frac{N_2A - N_1B}{N})^T, \tag{12}$$

$$I^*(\theta_0) = \frac{N_1N_2}{N^2}\begin{pmatrix} N & A+B \\ A+B & C \end{pmatrix} - \frac{1}{N}\begin{pmatrix} 0 & 0 \\ 0 & N_2^2F + N_1^2Q \end{pmatrix} \tag{13}$$

where $e_{ij} = p_{ij1} + 2p_{ij2}$, $f_{ij} = p_{ij1} + 4p_{ij2}$ and $A = \sum_{i:\Phi_i=1} e_{ij}$, $B = \sum_{i:\Phi_i=0} e_{ij}$, $C = \sum_i f_{ij}$, $F = \sum_{i:\Phi_i=1}(f_{ij} - e_{ij}^2)$ and $Q = \sum_{i:\Phi_i=0}(f_{ij} - e_{ij}^2)$.

The Score Test Statistic is $S = \frac{(U_\gamma^*)^2}{I_\gamma^*}$ where

$$U_\gamma^* = U^*(\theta_0)_\gamma = \frac{N_2A - N_1B}{N} \tag{14}$$

$$I_\gamma^* = I^*(\theta_0)_{\gamma\gamma} - I^*(\theta_0)_{\gamma\mu}[I^*(\theta_0)_{\mu\mu}]^{-1}I^*(\theta_0)_{\mu\gamma} \tag{15}$$

$$= \frac{N_1N_2}{N^2}\Big(C - \frac{(A+B)^2}{N} - \frac{N(N_2^2F + N_1^2Q)}{N_1N_2}\Big) \tag{16}$$

So that

$$S = \frac{(N_2A - N_1B)^2}{N_1N_2\Big(C - \frac{(A+B)^2}{N} - \frac{N(N_2^2F+N_1^2Q)}{(N_1N_2)}\Big)}. \tag{17}$$

In the case of an equal number of cases and controls ($N_1 = N_2 = N/2$) this Score test reduces to

$$S = \frac{(A-B)^2}{4\Big(C - \frac{(A+B)^2}{N} - N(F+Q)\Big)}. \tag{18}$$

The Score test relies upon the asymptotic result that $U_\gamma^* \sim N(0, I_\gamma^*)$ under H_0 so that $S \sim \chi_1^2$ under H_0.

When genotypes are imputed with no uncertainty i.e $p_{ijk} = 1$ for some $k \in \{0, 1, 2\}$ then this test statistic reduces to the Armitage Trend Test statistic[3]

$$S = \frac{N(N_2(s_1 + 2s_2) - N_1(r_1 + 2r_2))^2}{N_1N_2(N(r_1 + s_1 + 4(r_2 + s_2)) - (r_1 + s_1 + 2(r_2 + s_2))^2)}, \tag{19}$$

where r_1 and r_2 are the numbers of cases with G_{ij} equal to 1 and 2 respectively and s_1 and s_2 are the numbers of cases with G_{ij} equal to 1 and 2 respectively.

It is important to consider when the Score test will work well. One way to think about the Score test is that it makes an assumption that the log-likelihood curve is quadratic. If so, then the maximum log-likelihood can be found by evaluating the second derivative (the negative information matrix) at any parameter value i.e. under the null. There are 3 factors that will act to degrade the validity of the quadratic assumption (a) small sample size, (b) low allele frequency, (c) increasing genotype uncertainty from imputation. When the assumption isn't valid the score test can behave badly and lead to a spuriously low p-value as has been observed in practice. There can be no correct "threshold" on these factors that determines when it will work well and not work well but in real GWAS studies researchers have used thresholds on information metrics (see Section 2) and allele frequencies to filter out SNPs at which this happens. Since such SNPs are those likely to have very low power to detect effects it is unlikely that has a negative effect on the study. In SNPTEST v2 if the score test does not produce a sensible result then we use the EM algorithm (see below) to fit the model.

1.1.3 Maximum Likelihood Tests

An alternative approach is to try to maximize the likelihood directly (option `-method ml` in SNPTEST v2). This can be done using a Newton-Raphson algorithm which updates the parameter estimates as

$$\theta^{t+1} = \theta^t + [I^*(\theta^t)]^{-1} U^*(\theta^t). \tag{20}$$

If this algorithm converges to $\hat{\theta}$ then estimate of a Maximum Likelihood Ratio Test (MLRT) Statistic can be used which has the form

$$S_{MLRT} = 2 \log \left(\frac{l^*(\hat{\theta}; Y_O)}{l^*(\theta_0; Y_O)} \right) \sim \chi^2_1 \text{ under } H_0. \tag{21}$$

Alternatively, a Wald test can be used which assumes $\hat{\theta} \sim N(\theta_0, I^*(\hat{\theta}))$ under H_0. So that, in the case of the model in Eq. 6, it has the form

$$S_{wald} = \frac{\hat{\theta}^2_\gamma}{I^*(\hat{\theta})_{\gamma\gamma}} \sim \chi^2_1 \text{ under } H_0. \tag{22}$$

Another alternative is to use an EM algorithm (option `-method em` in SNPTEST v2) which in this case iterates between the following 2 steps

E-step Using the current parameter estimate, θ^t, calculate the distribution $P(Y_M|Y_O, \theta)$ given by Eq. 5 and use this to calculate the expected log likelihood as

$$Q(\theta|\theta^t) = \sum_{i=1}^{N} \sum_{k=0}^{2} q_{ijk} \log P(\Phi|G_{ij} = k, \theta). \tag{23}$$

M-step Create the new estimate, θ^{t+1} by maximizing $Q(\theta|\theta^t)$. One option is to calculate the first and second derivatives of this function and use a Newton-Raphson scheme (or any other numerical optimization algorithm) to update θ. Since the Newton-Raphson scheme makes a quadratic assumption the convergence is not guaranteed. In the case of a binary phenotype we've found this approach does tend to have better convergence behaviour than direct maximization. In the case of a quantitative phenotype, analyzed using a Normal model (Eq. 10), the M-step can be done precisely with parameter updates

$$\begin{pmatrix} \mu \\ \gamma \end{pmatrix} = \begin{pmatrix} N & \sum_i d_{ij} \\ \sum_i d_{ij} & \sum_i w_{ij} \end{pmatrix}^{-1} \begin{pmatrix} \sum_i \Phi_i \\ \sum_i \Phi_i d_{ij} \end{pmatrix}, \tag{24}$$

$$\sigma^2 = \frac{1}{N} \sum_i \sum_{k=0}^{2} (\Phi_i - \mu - \gamma k)^2 q_{ijk}, \tag{25}$$

where $d_{ij} = q_{ij1} + 2q_{ij2}$ and $w_{ij} = q_{ij1} + qp_{ij2}$.

Both the Newton-Raphson and EM algorithms are implemented in SNPTEST v2.

1.1.4 Expected count tests

A simpler approach involves using the expected genotype count $e_{ij} = p_{ij1} + 2p_{ij2}$ (also called posterior mean[4] or allele dosage[5]). These expected counts can be used to test for association with a binary or quantitative phenotype, using a standard logistic or linear regression model respectively (option `-method expected` in SNPTEST v2). This method has been shown to provide a good approximation to methods that take the genotype uncertainty into account when the effect size of the risk allele is small[4], which is the case for most of the common variants found in recent GWAS.

1.1.5 Thresholded tests

An even simpler method involves transforming the probabilities of each imputed genotype into a hard genotype call, by taking the genotype with the largest probability (if the probability is larger than some pre-sepcified threshold) and calling a missing genotype otherwise (option `-method threshold` in SNPTEST v2). These called genotypes can then be tested for association in the usual way. This method is not recommended as it can lead to a significant number of genotypes not being called. A variant of this is to use the best guess genotype as the predicted call[6] but since doing so ignores the inherent uncertainty of the imputed genotypes and may tend to produce a set of over-confident genotype calls.

1.2 Bayes Factors

The use of Bayesian methods for analyzing SNP associations have recently been proposed[7,4,8,9]. and have advantages over the use of p-values in power and interpretation. Stephens and Balding (2009)[9] provide an excellent review of this subject and discuss the choice of priors. Within the Bayesian framework focus centers on calculation of the a Bayes factor (BF), which is a ratio of marginal likelihoods between a model of association (M1) and a null model of no association (M0),

$$BF = \frac{P(Data|M_1)}{P(Data|M_0)} \tag{26}$$

where the marginal likelihoods are defined by

$$P(Data|M_l) = \int \Big(\prod_{i=1}^{N} \sum_{k=0}^{2} P(\Phi|G_{ij} = k, \theta) p_{ijk} \Big) P(\theta|M_l) d\theta. \tag{27}$$

and can be approximated using a Laplace approximation. This involves finding the maximum point of the product of the observed data likelihood and the prior on θ,

$$\hat{\theta} = \text{argmax}_\theta \Big(\prod_{i=1}^{N} \sum_{k=0}^{2} P(\Phi|G_{ij} = k, \theta) p_{ijk} \Big) P(\theta|M_l) \tag{28}$$

The missing data likelihood theory (Section 1.1.1) is easily modified to achieve this and the relevant score and information matrices become

$$U^*(\theta) = \mathbb{E}_{Y_M|Y_O,\theta}[U(\theta)] + \frac{d \log P(\theta|M_l)}{d\theta} \tag{29}$$

$$I^*(\theta) = \mathbb{E}_{Y_M|Y_O,\theta}[I(\theta)] - V_{Y_M|Y_O,\theta}[U(\theta)] - \frac{d^2 \log P(\theta|M_l)}{d\theta^2} \tag{30}$$

Newton-Raphson iterations can be used to obtain the *maximum a posteiori* (MAP) estimate, $\hat{\theta}_{M_1}$. In general, we have found that one iteration is often sufficient. The maximisation suffers from less problems than the Score test

since the prior acts to regularize (or penalize) the estimate. In effect, the addition of the prior means that the log posterior is often much closer to a quadratic than the log likelihood would be. This means that the Newton-Raphson algorithm is more stable and converges faster. The EM algorithm can also be used as above.

Consider the case of a binary phenotype model (Eq. 6) and suppose we want to calculate the marginal likelihood for M_1. Then one set of priors that have been used in this scenario are[10] is $P(\theta|M_1) = P(\mu)P(\gamma)$ where $\mu \sim N(0,1)$ and $\gamma \sim N(0,s^2)$ where $s = 0.2$. In this case,

$$\frac{d\log P(\theta|M_1)}{d\theta} = \begin{pmatrix} -\mu & -\gamma s^{-2} \end{pmatrix}^T, \tag{31}$$

$$\frac{d^2\log P(\theta|M_1)}{d\theta^2} = \begin{pmatrix} -1 & 0 \\ 0 & -s^{-2} \end{pmatrix}. \tag{32}$$

A similar set of equations can be derived for the null model, M_0. Once the MAP estimates, $\hat{\theta}_{M_1}$ and $\hat{\theta}_{M_0}$ have been obtained then the marginal likelihoods needed for the Bayes Factor can be approximated as

$$\log P(Data|M_1) \approx \Big(\sum_{i=1}^{N}\sum_{k=0}^{2} P(\Phi|G_{ij}=k,\hat{\theta}_{M_1})p_{ijk}\Big) + \log P(\hat{\theta}_{M_1}|M_1) + \log(2\pi) - \frac{1}{2}\log|I^*(\hat{\theta}_{M_1})|, \tag{33}$$

$$\log P(Data|M_0) \approx \Big(\sum_{i=1}^{N}\sum_{k=0}^{2} P(\Phi|G_{ij}=k,\hat{\theta}_{M_0})p_{ijk}\Big) + \log P(\hat{\theta}_{M_0}|M_0) + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log|I^*(\hat{\theta}_{M_0})|, \tag{34}$$

Stephens and Balding (2009)[9] have pointed out that the tail probabilities of the Normal prior might be to small to reflect realistic beliefs about effect sizes in GWAS. They propose the use of a mixture of Normal distributions as a prior to sufficiently fatten the tails of the prior. Another way to do this is to use a *t*-distribution prior for the effect size parameter γ with density

$$f(\gamma; m, s, d) = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)d^{1/2}\pi^{1/2}s}\Big[1 + \frac{(\gamma-m)^2}{ds^2}\Big]^{-\frac{d+1}{2}} \tag{35}$$

where m is the mean, s^2 is the variance parameter and d is the degrees of freedom (implemented in SNPTEST v2). If we use this as a prior for γ and keep the $N(0,1)$ prior for μ then

$$\frac{d\log P(\theta|M_l)}{d\theta} = \begin{pmatrix} -\mu & -\frac{(d+1)(\gamma-m)}{ds^2-(\gamma-m)^2} \end{pmatrix}^T, \tag{36}$$

$$\frac{d^2\log P(\theta|M_l)}{d\theta^2} = \begin{pmatrix} -1 & 0 \\ 0 & -\frac{(d+1)(ds^2+(\gamma-m)^2)}{(ds^2-(\gamma-m)^2)^2} \end{pmatrix}. \tag{37}$$

Some tail probabilities for the Normal and t priors are given in Table 1. These illustrate how a $t(m=0, s^2=.2^2, d=3)$ prior can give very similar tail probabilities to the mixture of normals prior proposed in Stephens and Balding (2009).

Other options for calculating Bayes factors at imputed SNPs include using the expected genotype counts[4], Wakefield's Approximate Bayes Factor (WABF)[11,4] and retrospective likelihood Bayes Factors[9]. See Stephens and Balding (2009) for a review of these methods. Earlier versions of the program SNPTEST used a Monte Carlo approach in which genotypes were sampled from the imputation distribution, p_{ijk}, and complete data Bayes Factors were averaged over to produce a single Bayes Factor. This was only done at SNPs with low measures of certainty and at SNPs with high certainty the bets guess genotype was used.

Priors for Quantitative Trait models

Setting the priors for a quantitative trait analysis using the Normal model (Eq. 10) is a bit more tricky than when analysing a binary trait as the scale of the phenotype and the size of the expected genetic effect relative to this scale needs to be considered. In SNPTEST v2 there is an option to calculate a Bayes Factor for a quantitative trait using

Table 1: Tail probabilities for different priors on the effect size γ. A prior probability of association $\pi = 10^{-4}$ is assumed in these calculations so as to be comparable with Table 2 in Stephens and Balding (2009). The mixture of normals prior has the form $\gamma \sim 0.9N(0, 0.2) + 0.05N(0, 0.4) + 0.05N(0, 0.8)$.

	$\gamma \sim N(0, .2^2)$	$\gamma \sim N(0, .3^2)$	$\gamma \sim t(m = 0, s^2 = .2^2, d = 3)$	mixture of normals
$P(\|\gamma\| > 0.05)$	8.0×10^{-5}	8.7×10^{-5}	8.2×10^{-5}	8.1×10^{-5}
$P(\|\gamma\| > 0.1)$	6.2×10^{-5}	7.4×10^{-5}	6.5×10^{-5}	6.4×10^{-5}
$P(\|\gamma\| > 0.2)$	3.2×10^{-5}	5.0×10^{-5}	3.9×10^{-5}	3.6×10^{-5}
$P(\|\gamma\| > 0.4)$	4.5×10^{-6}	1.8×10^{-5}	1.4×10^{-5}	8.8×10^{-6}
$P(\|\gamma\| > 1)$	5.7×10^{-11}	8.6×10^{-8}	1.5×10^{-6}	1.1×10^{-6}

the Normal model. The prior used is the conjugate Normal Inverse Gamma (NIG) prior. The way in which this model is formulated is best illustrated though examples. For an additive model the formulation is

$$\phi_i' = \gamma e_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \tag{38}$$

where ϕ_i' is the residual phenotype after a baseline mean and any covariate effects have been estimated and subtracted off (so that the effect we are testing for is conditional upon those estimates), e_{ij} is the expected genotype and σ^2 is the error variance. The NIG prior on the model parameters γ and σ^2 is written as

$$\gamma|\sigma^2 \sim N(m_\gamma, V_\gamma \sigma^2) \tag{39}$$

$$\sigma^2 \sim \text{InverseGamma}(a, b) \tag{40}$$

This results in a marginal prior for γ of

$$\gamma \sim t_{2a}(m_\gamma, 4abV_\gamma/(a-1)) \tag{41}$$

It can be shown that the expected non-centrality parameter for the F-test when fitting (38) is approximately

$$Np(1-p)\frac{2\gamma^2}{\sigma^2} \tag{42}$$

where γ and σ^2 are the true values of the alternative model and $2N$ is the total sample size[12]. This can be usefully compared to the non-centrality parameter for the case-control test which is approximately

$$Np(1-p)\gamma^2 \tag{43}$$

assuming N cases and N controls, and here γ is the log-odds ratio parameter of a logistic regression model. If we believe that the loci underlying quantitative traits are likely to have similar effect sizes to those underlying binary traits then we can equate the priors on γ for a binary trait and $\frac{\sqrt{2}\gamma}{\sigma}$ in model (38). So, the $N(0, 0.2^2)$ prior on γ for a binary trait can bse used for $\frac{\sqrt{2}\gamma}{\sigma}$ in model (38) i.e $\gamma \sim N(0, 0.02\sigma^2)$. In the context of the NIG prior used this would mean setting $V_\gamma = 0.02$. The parameters a and b can be set by ensuring that the total variance of the phenotype lies well within the range of the $IG(a, b)$ distribution which has mean $b/(a-1)$ and variance $b^2/[(a-1)^2(a-2)]$. Extension of this way of setting the priors to dominant, recessive, heterozygote and general models is straightforward.

1.3 Bayes Factors versus p-values

At directly genotped SNPs Bayes factors and p-values can be made equivalent in the sense that they give the same ranking of SNPs[8] but this occurs for a particular choice of prior in which the prior variance of the effect size increases as minor allele frequency decreases (or as the information at the SNP about the effect size parameter

decreases). This prior assumes larger effects at rarer SNPs which may be biologically reasonable. At imputed SNPs the level of uncertainty also influences the amount of information there is about the effect size parameter. To make Bayes factors give the same ranking of SNPs as p-values we would need to allow prior variance to increase as the amount of imputation uncertainty increases which makes no sense[9]. So even when adopting a prior that depends upon allele frequency Bayes factors and p-values will not give the same ranking at imputed SNPs. In practice studies have tended to filter out SNPs with low information so it seems unlikely that a re-analysis of studies using a Bayes factors will result in very different outcomes but as we probe rarer and rarer SNPs based on imputation from the 1000 Genomes data it may become more important to take care of these details.

2 Information measures

SNPTEST produces two different information measures in the output files which are described below. The notation used in the description of these measures is as follows. Let $G_i \in \{0, 1, 2\}$ denote the genotype of the ith individual at the jth SNP in a study cohort of N samples. Also, let $p_{ijk} = P(G_{ij} = k|H, G)$ be the probability (obtained from imputation) that the genotype at the jth SNP of the ith individual is k. Let the expected allele dosage for the genotype at the jth SNP of the ith individual be $e_{ij} = p_{ij1} + 2p_{ij2}$ and define $f_{ij} = p_{ij1} + 4p_{ij2}$. Also, θ_j denote the (unknown) population allele frequency of the jth SNP with estimate $\hat{\theta} = \frac{\sum_{i=1}^{N} e_{ij}}{2N}$. Also, let $X = \sum_{i=1}^{N} G_{ij}$.

2.1 The `info` measure

This is based on measuring the relative statistical information about the population allele frequency, θ_j . If the G_{ij}'s were observed then the full data likelihood is given by

$$L(\theta_j) = \prod_{i=1}^{N} \theta_j^{G_{ij}} (1 - \theta_j)^{2-G_{ij}} \tag{44}$$

For this likelihood the score and information are given by

$$U(\theta_j) = \frac{d \log L(\theta_j)}{d\theta_j} = \frac{X - 2N\theta_j}{\theta_j(1 - \theta_j)} \tag{45}$$

$$I(\theta_j) = \frac{-d^2 \log L(\theta_j)}{d\theta_j^2} = \frac{X}{\theta_j^2} + \frac{2N - X}{(1 - \theta_j)^2} \tag{46}$$

The IMPUTE info measure is based on the same idea used to calculate the SNPTEST information measure i.e. the ratio of the observed and complete information.

$$I_A = \frac{\mathbb{E}_{G_{\cdot j}}[I(\hat{\theta})] - V_G[U(\hat{\theta})]}{\mathbb{E}_{G_{\cdot j}}[I(\hat{\theta})]} \tag{47}$$

where the expectations are taken over the imputed genotype distribution and evaluated at the allele frequency estimate, $\hat{\theta_j}$. The exact terms are given by

$$\mathbb{E}_{G_{\cdot j}}[I(\hat{\theta})] = \frac{2N}{\hat{\theta}(1 - \hat{\theta})} \tag{48}$$

$$V_G[U(\hat{\theta})] = \frac{\sum_{i=1}^{N}(f_{ij} - e_{ij}^2)}{\hat{\theta}^2(1 - \hat{\theta})^2} \tag{49}$$

so that

$$I_A = \begin{cases} 1 - \frac{\sum_{i=1}^{N}(f_{ij} - e_{ij}^2)}{2N\hat{\theta}(1-\hat{\theta})} & \text{when } \hat{\theta} \in (0, 1) \\ 1 & \text{when } \hat{\theta} = 0, \hat{\theta} = 1. \end{cases} \tag{50}$$

So I_A is bounded above at 1 and will equal 0 when the sample mean variance of the imputed genotypes equals the variance you would expect if alleles where samples with frequency $\hat{\theta}$. This measure is not dependent upon any model of association being fitted at a given SNP but in our experience it tends to be very close to the relative information measure for an additive model that is described below.

2.2 The association test information measures

For each of the frequentist association tests SNPTEST will produce a relative information measure about the parameter(s) of the association model being fitted. These measures appear in the output files in columns after the columns that contain the p-values for the test and have column names that end with `_info`. In this way the measures are dependent upon the model being fitted and may vary between models at a given SNP i.e. there may be good relative information to fit an additive model at a SNP but very little information to fit a recessive model at the same SNP. The derivation of these relative information measures is as follows.

The power of the score test is governed by the distribution of the statistic under a specific alternative, say $H_1 : \theta = \theta_1$. Under this alternative, we have the following asymptotic result

$$U^*_\gamma \sim N(\gamma I^*_\gamma, I^*_\gamma) \tag{51}$$

where U^* and I^* are defined in Eqns. (12) and (13). This implies that the non-centrality parameter of the Score test is

$$\eta^* = \gamma I^*_\gamma. \tag{52}$$

If there was no genotype uncertainty then the analogous result would be

$$\eta = \gamma I_\gamma, \tag{53}$$

where I_γ is the marginal full data likelihood information about the parameter γ. Thus the relative information is given by the ratio of the these two non-centrality parameters

$$I_S = \frac{\eta^*}{\eta} = \frac{I^*_\gamma}{I_\gamma}. \tag{54}$$

The term I^*_γ is calculated during the association test but I_γ must be approximated by replacing $I^*(\theta_0)$ with $\mathbb{E}_{Y_M|Y_O,\theta_0}[I(\theta_0)]$ in Eq. (15).

Little and Rubin (2002) consider the result

$$I^*(\theta_0) = \mathbb{E}_{Y_M|Y_O,\theta_0}[I(\theta_0)] - V_{Y_M|Y_O,\theta_0}[U(\theta_0)] \tag{55}$$

and call $i_{com} = \mathbb{E}_{Y_M|Y_O,\theta_0}[I(\theta_0)]$ the complete information, $i_{obs} = I^*(\theta_0)$ the observed information and $i_{mis} = V_{Y_M|Y_O,\theta}[U(\theta)]$ so that

$$i_{obs} = i_{com} - i_{mis} \tag{56}$$

which has the appealing interpretation that the observed information equals the complete information minus the missing information and so

$$I_S = \frac{(i_{obs})_\gamma}{(i_{com})_\gamma}. \tag{57}$$

When there is no genotype uncertainty $i_{obs} = i_{com}$ and $I_S = 1$.

References

1. Louis, T. A. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **44**(2), 226–233.

2. Little, R. J. A. and Rubin, D. B. Statistical analysis with missing data. , 278 (1987).

3. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386 (1955).

4. Guan, Y. and Stephens, M. Practical issues in imputation-based association mapping. *PLoS Genet* **4**(12), e1000279 (2008).

5. Li, Y. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *Presented at the annual meeting of The American Society of Human Genetics, 12 October 2006, New Orleans, Louisiana* (2006).

6. Browning, B. and Browning, S. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**(2), 210–23 (2009).

7. Scott, L., Mohlke, K., Bonnycastle, L., Willer, C., Li, Y., Duren, W., Erdos, M., Stringham, H., Chines, P., Jackson, A., Prokunina-Olsson, L., Ding, C., Swift, A., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X., Conneely, K., Riebow, N., Sprau, A., Tong, M., White, P., Hetrick, K., Barnhart, M., Bark, C., Goldstein, J., Watkins, L., Xiang, F., Saramies, J., Buchanan, T., Watanabe, R., Valle, T., Kinnunen, L., Abecasis, G., Pugh, E., Doheny, K., Bergman, R., Tuomilehto, J., Collins, F., and Boehnke, M. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science* **316**(5829), 1341–5 (2007).

8. Wakefield, J. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology* **33**, 79–86 (2009).

9. Stephens, M. and Balding, D. Bayesian statistical methods for genetic association studies. *Nat Rev Genet* **10**(10), 681–690 (2009).

10. Consortium, T. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**(7145), 661–78 (2007).

11. Wakefield, J. A bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* **81**(2), 208–27 (2007).

12. Searle, S. R. Linear models. .