Supplementary information S7 | **Testing for association at imputed SNPs**

Score tests

A Score Test needs calculations of the observed data score and information matrix only under the null hypothesis, $H_0 : \theta = \theta_0$. 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

$$P(G_{ij} = k|\Phi, G, H, \theta) = P(G_{ij} = k|G, H) = p_{ijk} \tag{1}$$

so that

$$\begin{aligned} U^*(\theta_0) &= (0\,\frac{N_2A - N_1B}{N})^T, & (2) \\ I^*(\theta_0) &= \frac{N_1N_2}{N^2}\begin{pmatrix} N & A+B \\ A+B & C \end{pmatrix} - \frac{1}{N}\begin{pmatrix} 0 & 0 \\ 0 & N_2^2F + N_1^2Q \end{pmatrix} & (3) \end{aligned}$$

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

$$\begin{aligned} U_\gamma^* &= U^*(\theta_0)_\gamma = \frac{N_2A - N_1B}{N} & (4) \\ I_\gamma^* &= I^*(\theta_0)_{\gamma\gamma} - I^*(\theta_0)_{\gamma\mu}[I^*(\theta_0)_{\mu\mu}]^{-1}I^*(\theta_0)_{\mu\gamma} & (5) \\ &= \frac{N_1N_2}{N^2}\left(C - \frac{(A+B)^2}{N} - \frac{N(N_2^2F + N_1^2Q)}{N_1N_2}\right) & (6) \end{aligned}$$

So that

$$S = \frac{(N_2A - N_1B)^2}{N_1N_2\left(C - \frac{(A+B)^2}{N} - \frac{N(N_2^2F+N_1^2Q)}{(N_1N_2)}\right)}. \tag{7}$$

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

$$S = \frac{(A - B)^2}{4\left(C - \frac{(A+B)^2}{N} - N(F + Q)\right)}. \tag{8}$$

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[1]

$$S = \frac{N(N_2(s_1 + 2s_2) - N_1(r_1 + 2r_2))^2}{N_1N_2(N(r_1 + s_1 + 4(r_2 + s_2)) - (r_1 + s_1 + 2(r_2 + s_2))^2)}, \tag{9}$$

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.

EM algorithm

In the context of testing imputed SNPs for association the EM algorithm iterates between the following 2 steps

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

$$Q(\theta|\theta^t) = \sum_{i=1}^{N}\sum_{k=0}^{2} q_{ijk} \log P(\Phi|G_{ij} = k, \theta). \qquad (10)$$

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, the M-step can be done precisely with parameter updates

$$\begin{aligned} \begin{pmatrix} \mu \\ \gamma \end{pmatrix} &= \begin{pmatrix} N & \sum_i d_{ij} \\ \sum_i d_{ij} & \sum_i w_{ij} \end{pmatrix}^{-1} \begin{pmatrix} \sum_i \Phi_i \\ \sum_i \Phi_i d_{ij} \end{pmatrix}, & (11) \\ \sigma^2 &= \frac{1}{N}\sum_{i}\sum_{k=0}^{2}(\Phi_i - \mu - \gamma k)^2 q_{ijk}, & (12) \end{aligned}$$

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

Calculating Bayes factors at imputed SNPs

The Laplace approximation can be used to estimate the marginal likelihoods ($P(Data|M_1)$ and $P(Data|M_0)$) needed to calculate the Bayes factor. The precise expressions are

$$\begin{aligned} \log P(Data|M_1) &\approx \Big(\sum_{i=1}^{N}\sum_{k=0}^{2} P(\Phi|G_{ij} = k, \hat\theta_{M_1})p_{ijk}\Big) + \log P(\hat\theta_{M_1}|M_1) + \log(2\pi) - \frac{1}{2}\log|I^*(\hat\theta_{M_1})|, \\ \log P(Data|M_0) &\approx \Big(\sum_{i=1}^{N}\sum_{k=0}^{2} P(\Phi|G_{ij} = k, \hat\theta_{M_0})p_{ijk}\Big) + \log P(\hat\theta_{M_0}|M_0) + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log|I^*(\hat\theta_{M_0})|. \end{aligned}$$

These expressions require finding the maximum point of the product of the observed data likelihood and the prior on θ,

$$\hat\theta = \text{argmax}_\theta \Big(\prod_{i=1}^{N}\sum_{k=0}^{2} P(\Phi|G_{ij} = k, \theta)p_{ijk}\Big) P(\theta|M_l) \qquad (13)$$

Then the relevant score and information matrices are

$$\begin{aligned} U^*(\theta) &= \mathbb{E}_{Y_M|Y_O,\theta}[U(\theta)] + \frac{d\log P(\theta|M_l)}{d\theta} && (14) \\ I^*(\theta) &= \mathbb{E}_{Y_M|Y_O,\theta}[I(\theta)] - V_{Y_M|Y_O,\theta}[U(\theta)] - \frac{d^2\log P(\theta|M_l)}{d\theta^2}. && (15) \end{aligned}$$

Newton-Raphson iterations can be used to obtain the *maximum a posteiori* (MAP) estimates, $\hat{\theta}_{M_1}$ and $\hat{\theta}_{M_0}$. 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.

When calculating the marginal likelihood for the alternative model M_1 with a binary phenotype a prior that has been used in this scenario[2] 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,

$$\begin{aligned} \frac{d\log P(\theta|M_1)}{d\theta} &= \begin{pmatrix} -\mu & -\gamma s^{-2} \end{pmatrix}^T, && (16) \\ \frac{d^2\log P(\theta|M_1)}{d\theta^2} &= \begin{pmatrix} -1 & 0 \\ 0 & -s^{-2} \end{pmatrix}. && (17) \end{aligned}$$

A similar set of equations can be derived for the null model, M_0.

t-distribution priors for binary trait Bayes factors

Stephens and Balding (2009)[3] have pointed out that the tail probabilities of the Normal prior might be too 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

$$f(\gamma; m, s, d) = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)d^{1/2}\pi^{1/2}s}\left[1 + \frac{(\gamma - m)^2}{ds^2}\right]^{-\frac{d+1}{2}} \qquad (18)$$

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

$$\begin{aligned} \frac{d\log P(\theta|M_l)}{d\theta} &= \begin{pmatrix} -\mu & -\frac{(d+1)(\gamma - m)}{ds^2 - (\gamma - m)^2} \end{pmatrix}^T, && (19) \\ \frac{d^2\log P(\theta|M_l)}{d\theta^2} &= \begin{pmatrix} -1 & 0 \\ 0 & -\frac{(d+1)(ds^2 + (\gamma - m)^2)}{(ds^2 - (\gamma - m)^2)^2} \end{pmatrix}. && (20) \end{aligned}$$

One clear advantage of using a *t*-distribution prior rather than a mixture of normals is that only one model fit is required rather than three model fits. Some tail probabilities for the Normal and t priors are given in Supplementary Information 11. 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).

Priors for Quantitative Trait models

Setting the priors for a quantitative trait analysis using the Normal model 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 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

$$\phi_i' = \gamma e_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \tag{21}$$

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

$$\gamma|\sigma^2 \sim N(m_\gamma, V_\gamma \sigma^2) \tag{22}$$
$$\sigma^2 \sim \text{InverseGamma}(a, b). \tag{23}$$

This results in a marginal prior for γ of

$$\gamma \sim t_{2a}(m_\gamma, 4abV_\gamma/(a-1)). \tag{24}$$

It can be shown that the expected non-centrality parameter for the F-test when fitting (21) is approximately

$$Np(1-p)\frac{2\gamma^2}{\sigma^2} \tag{25}$$

where γ and σ^2 are the true values of the alternative model and $2N$ is the total sample size[4]. This can be usefully compared to the non-centrality parameter for the case-control test which is approximately

$$Np(1-p)\gamma^2 \tag{26}$$

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 (21). So, the $N(0, 0.2^2)$ prior on γ for a binary trait can be used for $\frac{\sqrt{2}\gamma}{\sigma}$ in model (21) 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.

References

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

2. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**(7145), 661–78 (2007).

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

4. Searle, S. R. *Linear Models*. New York: Wiley, (1997).