Supplementary information S3 | **Imputation information measures**

In this section we provide detailed information about the information metrics commonly calculated at imputed SNPs. Let $G_{ij} \in \{0, 1, 2\}$ 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, let 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 $X = \sum_{i=1}^{N} G_{ij}$.

The MACH $\hat{r}^2$ measure

This is the ratio of the empirically observed variance of the allele dosage to the expected binomial variance at Hardy-Weinberg equilibrium. At the jth SNP this is defined as

$$\hat{r}^2_j = \begin{cases} \frac{\frac{\sum_{i=1}^{N} e^2_{ij}}{N} - \left(\frac{\sum_{i=1}^{N} e_{ij}}{N}\right)^2}{2\hat{\theta}(1-\hat{\theta})} & \text{when } \hat{\theta} \in (0, 1) \\ 1 & \text{when } \hat{\theta} = 0, \hat{\theta} = 1 \end{cases} \quad (1)$$

When all the genotypes are predicted with high certainty this ratio will be close to 1, although it can go above 1 (Figure 1). As the amount of uncertainty increases the allele dosages will tend to 2θ, the empirical variance will tend to 0 and so $\hat{r}^2$ tends to 0.

The BEAGLE allelic R^2 metric

This metric is derived by approximating the R^2 between the best guess genotype and the true genotype in the case where the genotype is unknown. At the jth SNP this is defined as

$$R^2_j = \frac{[\sum_i z_{ij} e_{ij} - (1/N)(\sum_i z_{ij} \sum_i e_{ij})]^2}{[\sum_i f_{ij} - (1/N)(\sum_i e_{ij})^2][\sum_i z^2_{ij} - (1/N)(\sum_i z_{ij})^2]} \quad (2)$$

where z_{ij} is the most likely imputed genotype in the ith individual at the jth SNP.

The SNPTEST information measure I_S

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

$$U^*_\gamma \sim N(\gamma I^*_\gamma, I^*_\gamma) \quad (3)$$

where U^* and I^* are the observed score and information matrices. This implies that the non-centrality parameter of the Score test is

$$\eta^* = \gamma I^*_\gamma. \quad (4)$$

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

$$\eta = \gamma I_\gamma, \tag{5}$$

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

$$I_S = \frac{\eta^*}{\eta} = \frac{I^*_\gamma}{I_\gamma}. \tag{6}$$

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

Little and Rubin (2002) consider the result

$$I^*(\theta_0) = \mathbb{E}_{Y_M|Y_O,\theta_0}[I(\theta_0)] - V_{Y_M|Y_O,\theta_0}[U(\theta_0)] \tag{7}$$

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

$$i_{obs} = i_{com} - i_{mis} \tag{8}$$

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

$$I_S = \frac{(i_{obs})_\gamma}{(i_{com})_\gamma}. \tag{9}$$

When there is no genotype uncertainty $i_{obs} = i_{com}$ and $I_S = 1$. It is worth noting that this measure depends upon the genetic model of association being tested for at each SNP so that there is no guarantee that the information measure will be similar for different models.

The IMPUTE info measure I_A

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

$$L(\theta_j) = \prod_{i=1}^{N} \theta_j^{G_{ij}} (1-\theta_j)^{2-G_{ij}} \tag{10}$$

For this likelihood the score and information are given by

$$U(\theta_j) = \frac{d \log L(\theta_j)}{d\theta_j} = \frac{X - 2N\theta_j}{\theta_j(1-\theta_j)} \tag{11}$$

$$I(\theta_j) = \frac{-d^2 \log L(\theta_j)}{d\theta_j^2} = \frac{X}{\theta_j^2} + \frac{2N - X}{(1-\theta_j)^2} \tag{12}$$

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.

$$I_A = \frac{\mathbb{E}_{G_{\cdot j}}[I(\hat{\theta})] - V_G[U(\hat{\theta})]}{\mathbb{E}_{G_{\cdot j}}[I(\hat{\theta})]} \tag{13}$$

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

$$\mathbb{E}_{G_{\cdot j}}[I(\hat{\theta})] = \frac{2N}{\hat{\theta}(1-\hat{\theta})} \quad (14)$$

$$V_G[U(\hat{\theta})] = \frac{\sum_{i=1}^{N}(f_{ij}-e_{ij}^2)}{\hat{\theta}^2(1-\hat{\theta})^2} \quad (15)$$

so that

$$I_A = \begin{cases} 1 - \frac{\sum_{i=1}^{N}(f_{ij}-e_{ij}^2)}{2N\hat{\theta}(1-\hat{\theta})} & \text{when } \hat{\theta} \in (0,1) \\ 1 & \text{when } \hat{\theta}=0, \hat{\theta}=1. \end{cases} \quad (16)$$

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 sampled with frequency $\hat{\theta}$.

Comparison of information measures

To compare the different information measures we used HAPGEN[1] to simulate data on 1000 cases and 1000 controls on chromosome 22 based on the CEU HapMap2 haplotypes (release 22) in the interval 14-21Mb. Only genotypes at SNPs on the Affymetrix 500k chip were simulated. IMPUTE was then used to impute all un-genotyped SNPs from the CEU HapMap2 haplotypes. Figure 1 in the main text shows the MACH, BEAGLE and IMPUTE information measures applied to a simulated imputed dataset across the region and shows that the measures are highly correlated, although the MACH $\hat{r}^2$ measure often goes above 1 and the BEAGLE R^2 measure is undefined at almost 3% of SNPs. Supplementary Figure 2 shows bivariate plots of the MACH, BEAGLE and IMPUTE information measures. All 3 measures are highly correlated with the MACH and IMPUTE information measure being most correlated. The BEAGLE measure tend to systematically give a lower estimate of the information than the IMPUTE measure. Supplementary Figure 3 shows the good agreement between the IMPUTE information measure and the SNPTEST information measure when fitting an additive genetic model. Supplementary Figure 4 shows that the IMPUTE information measure and the SNPTEST information measure for a dominant genetic model can be quite different with the SNPTEST measure tending to a smaller estimate of information.

References

1. Spencer, C. C. A., Su, Z., Donnelly, P., and Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* **5**(5), e1000477 (2009).