Miscellaneous Genomics Notes

- Post-imputation information measure[1]. 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. 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. Define the expected allele dosage for the genotype at the jth SNP of the ith individual be

$$e_{ij} = p_{ij1} + 2p_{ij2}$$

  Note that this equation may define $2e_{ij}$ elsewhere. Let $f_{ij} = p_{ij1} + 4p_{ij2}$, θ_j denote the unknown population allele frequency of the jth SNP with estimate

$$\hat{\theta} = \frac{1}{2N}\sum_{i=1}^{N} e_{ij}$$

  and $X = \sum_{i=1}^{N} G_{ij}$.

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

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

  - The BEAGLE allelic R^2 is derived by approximating the R^2 between the best guess genotype (the most likely imputed genotype in the ith individual at the jth SNP, denoted by z_{ij}) and the allele dosage as an approximation of the true genotype in the case where the genotype is unknown. At the jth SNP this is defined as

$$R_j^2 = \frac{\left[\sum_i z_{ij}e_{ij} - \frac{1}{N}\left(\sum_i z_{ij}\sum_i e_{ij}\right)\right]^2}{\left[\sum_i f_{ij} - \frac{1}{N}\left(\sum_i e_{ij}\right)^2\right]\left[\sum_i z_{ij}^2 - \frac{1}{N}\left(\sum_i z_{ij}\right)^2\right]}$$

  - The IMPUTE info measure is based on measuring the relative statistical information about the population allele frequency, θ_j, given by

$$I_A = \begin{cases} 1 - \dfrac{\sum_{i=1}^{N}(f_{ij} - e_{ij}^2)}{2N(\hat{\theta}(1-\hat{\theta}))} & \hat{\theta} \in (0,1) \\ 1 & \hat{\theta} \in \{0,1\} \end{cases}$$

- The SNPTEST info measure is similar to the IMPUTE info measure when assuming an additive model (but not dominant model) and thus omitted here.

The MACH, BEAGLE and IMPUTE measures seem to be highly correlated with BEAGLE R^2 systemically reporting lower values and undefined at 3% of the SNPs and MACH r^2 often exceeds 1.

[1] http://www.nature.com/articles/nrg2796.