QGS estimation example

As an example, consider a genetic region **s** defined by 2 SNPs: rs1 and rs2. For a single sample individual, the dosage vector is given by s = {0.2, 1.2}, where 0.2 is the estimated (imputed) dosage for rs1 and 1.2 is the estimated imputed dosage for rs2. A reference population consisting of 3 individuals is defined as matrix

$$R = \begin{bmatrix} 0.5 & 0.8 \\ 1.1 & 0.1 \\ 1.4 & 0.3 \end{bmatrix}$$

where the top left 0.5 represents the estimated imputed dosage for rs1 for reference individual 1, etc. Following Equation 1 for the QGS, the absolute difference between the sample individual and each reference individual is computed for every SNP:

$$QGS = \sum_{r=1}^{3} \sum_{i=1}^{2} |R_i^r - s_i|$$

$$= |0.5 - 0.2| + |0.8 - 1.2| + |1.1 - 0.2| + \\ |0.1 - 1.2| + |1.4 - 0.2| + |0.3 - 1.2|$$

$$= 0.3 + 0.4 + 0.9 + 1.1 + 1.2 + 0.9 \\ = 4.8$$

Then the summed result is scaled to a value between 0-1:

$$QGS = \frac{\sum_{r=1}^{N_{refs}} \sum_{i=1}^{N_{snps}} |R_i^r - s_i|}{2N_{refs}N_{snps}} = \frac{4.8}{2 \cdot 3 \cdot 2} = \frac{4.8}{12} = 0.4$$

This final result of 0.4 represents the QGS for the sample individual calculated with the provided reference. Note that the QGS does not depend on phenotype and as such does not need to be recalculated for different applications, in contrast with existing genetic aggregation methods such as weighted PRS.

When the reference panel, the included variants, and the QGS value are all known, a limited amount of genetic information from $\vec{s}$ can be inferred. From the above example, one could e.g. deduce that the estimated imputed dosage for rs2 for the sample individual must be ≥1 by solving the below for rs2.

$$|0.5 - rs1| + |0.8 - rs2| + |1.1 - rs1| + \\ |0.1 - rs2| + |1.4 - rs1| + |0.3 - rs2| \\ = 4.8$$

This deduction becomes increasingly less informative with the inclusion of more SNPs in the genetic region, because the number of unknown variables increases.