SNP-QC

<Software to find SNPs that are monomorphic, redundant, or have too many missing values or low minor allele frequencies>

Rex Bernardo
University of Minnesota

Background

RR-BLUP, *RR-BLUP2*, *GModel*, and *GModel2* all use a file that indicates the names and chromosomal locations of SNP markers, and a second file with the SNP data. These files are described below. A SNP data file typically has missing values that need to be imputed. Furthermore, when the number of SNP loci is large (few thousands) and the number of individuals genotyped is much smaller (few hundreds), many of the SNP loci could have a high level of redundancy with other SNP loci. This may lead to a marker with a zero or near-zero effect being declared as significant in *GModel* and *GModel2*. In this situation, the effect is likely attributable not to the marker itself but to one or more markers that are highly correlated with the marker in question.

SNP-QC is a one-stop, no-frills software that (1) imputes missing marker data and (2) identifies markers that are monomorphic, have too many missing values, have a low minor allele frequency (MAF), and are redundant with other SNP markers. The user needs to specify the lowest acceptable frequency of missing data (e.g., 15%), lowest acceptable MAF (e.g., 0.05 or 0.10), and maximum acceptable r^2 value (r^2_{Max}, which is a measure of linkage disequilibrium) between SNP markers (e.g., r^2_{Max} = 0.80 to 0.90).

The marker data need to be coded as 1 and –1 for the two homozygotes, 0 for the heterozygote, and 99 (not NA) for missing values. While the SNP-QC software identifies SNP loci with an excessively high rate of missing data, it does not (by design) identify genotyped individuals with an excessively high rate of missing data. Marker data for such individuals need to be identified (e.g., with the `COUNTIF` function in Microsoft Excel) and, if needed, removed prior to running SNP-QC.

Missing data are imputed in one of two ways. For a panel of diverse individuals (population type = `Panel`), imputation is done by identifying the SNP locus that is most highly correlated with the locus with missing data. The SNP genotype at the highly correlated SNP locus is then ascribed to the individuals that are missing data at the SNP. Two conditions are needed for imputation in a biparental cross (population type = `Biparental`): (1) the two parents must be homozygous, and (2) the marker alleles from one parent should be coded as 1 whereas the marker alleles from the other parent should be coded as –1. Imputation is based on the genotypes at the two flanking markers. If the two flanking markers differ in their genotypes, the imputed value for the missing genotype in the middle is 0.

Redundant markers are identified according to the following backwards elimination procedure.

1. For chromosome 1, estimate the correlation (r) between each pair of markers found on that chromosome.
2. If $r^2 > r^2_{Max}$, identify which of the two SNPs has a lower MAF and flag that SNP as *Redundant*. Disregard this redundant marker from further analysis. Keep the first SNP if the MAF values are equal.
3. Repeat step 2 until all of the unflagged markers on chromosome 1 have $r^2 \leq r^2_{Max}$.
4. Repeat the above steps for all of the remaining chromosomes.

Executable File and Input Files

SNP-QC runs as a Windows executable file (`SNP-QC.exe`). It requires two input files plus a parameter file:

1. **Marker names and chromosomes** (Sample: `AMchrom.csv`)
   - This CSV file can be exported from Microsoft Excel or Google Sheets.
   - The marker names should have no spaces or commas, and not be more than 30 characters long.
   - The chromosome names should be numbers, with no letters.
   - The SNPs should be ordered by chromosome (lowest to highest), but the SNPs within a chromosome need not be in map order.

2. **SNP marker data** (Sample: `AMsnp.csv`)
   - The rows correspond to the SNP markers and the columns correspond to the individuals.
   - The order of SNPs in this file should correspond to the order of SNPs in the first file.
   - The SNP genotypes need to be coded as 1 for one homozygote, 0 for the heterozygote, –1 for the other homozygote, and 99 (not NA or some other code) for missing values.

3. **Parameter file** (Sample: `SNPQCParms.csv`)
   - Use the sample file as a template. Change only the information in the second column, and rename the file as needed.

PopulationType	Panel	- Either `Panel` or `Biparental`
Individuals	272	- Number of individuals for which SNP data are available
<MissingValues	0.2	- Threshold for missing SNP data
>MAF	0.1	- Threshold for minor allele frequency
r^2Max	0.8	- r^2 threshold for declaring two SNPs as redundant
ChromFile	AMchrom.csv	
SNPdataFile	AMsnp.csv	
OutputFile	SNPQC_Output.csv	- Summary output file
NewChromFile	AMchromNew.csv	- New ChromFile that has only those SNP loci labeled as `Keep`
NewSNPdataFile	AMsnpNew.csv	- New SNPdataFile that has only those SNP loci labeled as `Keep`

Output Files

1. **Summary output file** (e.g., `SNPQC_Output.csv`)
   - This CSV file indicates the numbers and names of markers that suffer from each of the different issues.
   - These issues are identified under `Status` (column D) as `Monomorphic`, `<MAF`, `>Missing`, or `Redundant`. Pairs of redundant markers detected during backwards elimination and the r^2 value are given. Markers without issues are identified as `Keep`.

2. **Filtered marker names and chromosomes** (e.g., `AMchromNew.csv`)
   - In this CSV file, only those SNP markers identified as `Keep` are retained.
   - This file can be readily used in *RR-BLUP*, *RR-BLUP2*, *GModel*, and *GModel2*.

3. **Filtered SNP marker data** (e.g., `AMsnpNew.csv`)
   - In this CSV file (with values of 1, –1, and 0), only those SNP loci identified as `Keep` are retained.
   - This file can be readily used in *RR-BLUP*, *RR-BLUP2*, *GModel*, and *GModel2*.