Materials

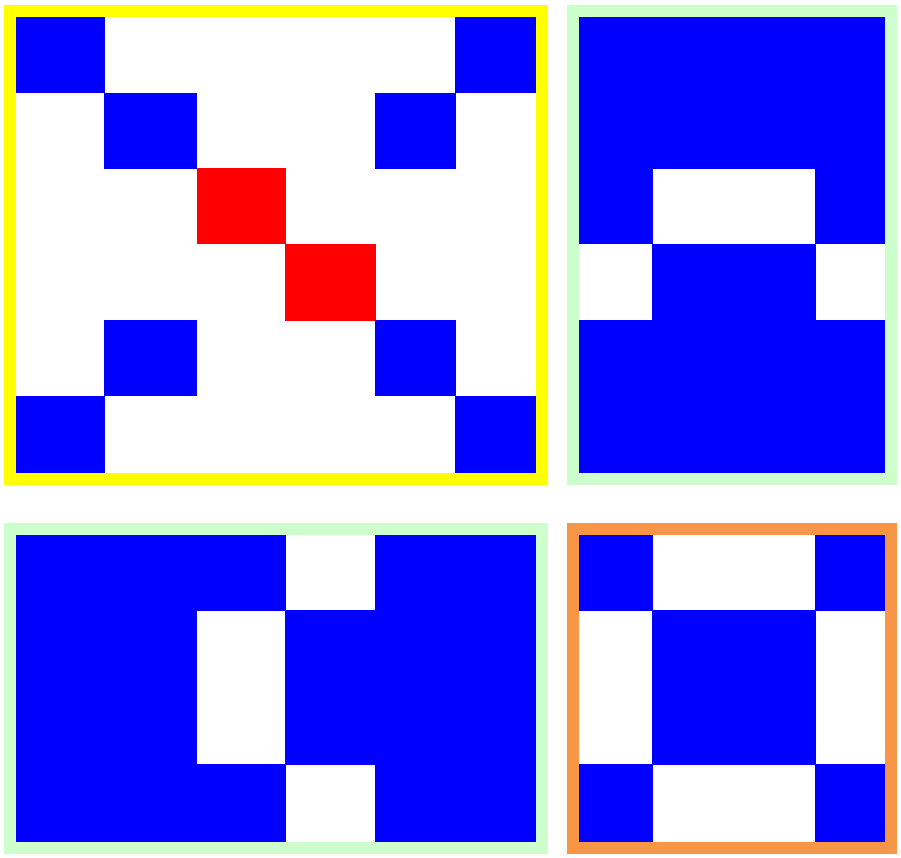
QC I: eliminate loci having more than 2 kinds of alleles

Theoretically, it cannot exclude the possibility, very low though, of more than 2 alleles at a locus, but at the current stage we ascribe the phenomena of more than 2 alleles at a locus attributes to technical limitation and cryptical factors. Given a biallelic locus, its exhaustive scenarios that the alleles at a locus assayed at one sample match its counterparts at another sample are as tabulated below.

For ordered alleles at a biallelic/monomorphic locus, there are $\binom{4}{2} + 4 = 10$ combinations. When two samples assay the same locus independently, it makes up to 100 scenarios. Given biallelic for a locus, it generates 10 possible matched scenarios when the locus is reported to be biallelic in both of the samples, 8 when monomorphic in both, 40 when biallelic in one sample but monomorphic in the other.

Table I: exhaustive enumerations of the scenarios two samples match a locus

SNP Alleles at the testing panel		SNP Alleles at the discovery panel									
		Biallelic						Monomorphic			
		AC	AG	AT	CG	CT	GT	A	C	G	T
Biallelic	AC										
	AG										
	AT										
	CG										
	CT										
	GT										
Monomorphic	A										
	C										
	G										
	T										



Notes: the matched scenarios are highlighted in blue. The yellow squared scenarios represent when the locus is detected to be biallelic at both sample, the green squared scenarios represent the locus is detected to be biallelic in one sample only, the orange squared scenarios represented the locus is detected to be monomorphic in both samples.

Two scenarios, highlighted in red, could cause ambiguity and require other rules to validate the consistency of coding. There are some risk in applying the rule such as minor allele frequency match, the overall accuracy is $p_t\left(\hat{f}_A < 0.5 \middle| f_A < 0.5\right) \times p_d\left(\hat{f}_A < 0.5 \middle| f_A < 0.5\right) + [1 - p_t\left(\hat{f}_A < 0.5 \middle| f_A < 0.5\right)] \times [1 - p_d\left(\hat{f}_A < 0.5 \middle| f_A < 0.5\right)]$. The two terms of probability are interpreted as that the labelled minor alleles are real minor alleles in both samples or in neither.

Which falls to its minimal value of 0.5 when $f_A \to 0.5$ and reaches its maximal value of 1 when $f_A \to 0$ or $f_A \to 1$.

In PGC and WAFSS, no loci had more than 2 alleles. So, each of the 605659 SNP loci matched one of the scenarios listed in Table 1.