

SUPPLEMENTAL APPENDIX – GENOTYPE RELIABILITY

BACKGROUND

We tested the reliability of our genotype data by running a subset of 46 of our samples in triplicate PCR runs. The samples selected for this analysis were split across study areas, genotyping batches, and original success rate (the number of loci for which we received data in the first genotyping run). The analysis below compares the results from the first PCR run to the information obtained by the additional runs.

DEFINITIONS

Here we use the term “genotype” to refer to the data for a single locus in a single individual. We had a total of 405 genotypes in this analysis after excluding loci for individuals where a comparison was not possible between the original data and the duplicate runs (i.e. a locus for an individual that either had missing data in the first PCR run, or that had missing data in both of the replicate runs).

In this analysis we classify a “**Confirmed**” genotype as:

- a) Homozygous genotypes that occurred in three replicates
- b) Heterozygous genotypes found in at least two replicates (where the third replicate was either missing, or was homozygous with a likely dropped allele).

Genotypes classified as “**Supported**”:

- a) Homozygous genotypes occurring in two replicates, but no third replicate was available
- b) Heterozygous genotypes where each allele occurred in at least two replicates (in this case across two distinct homozygous replicates and one heterozygous, e.g. AB + AA + BB).

The remaining genotypes were classified as “**Unreliable**”:

- a) When there were three or more different alleles across replicates
- b) If an allele was only found once across three replicates, making it unclear whether it was a likely dropped allele in two replicates, or a genotyping error in one
- c) When additional replicates suggested that the final genotype should be heterozygous, but the original genotype was homozygous.

RESULTS

The majority of our genotypes were confirmed in rePCR replicates. There were exact matches across all three replicates in 261 of the genotypes, with a total of 292 (72%) being classified as confirmed, 53 (13%) supported, and 60 (15%) unreliable. There was a relatively even spread for these categories across different study areas and PCR batches (Fig. 1).

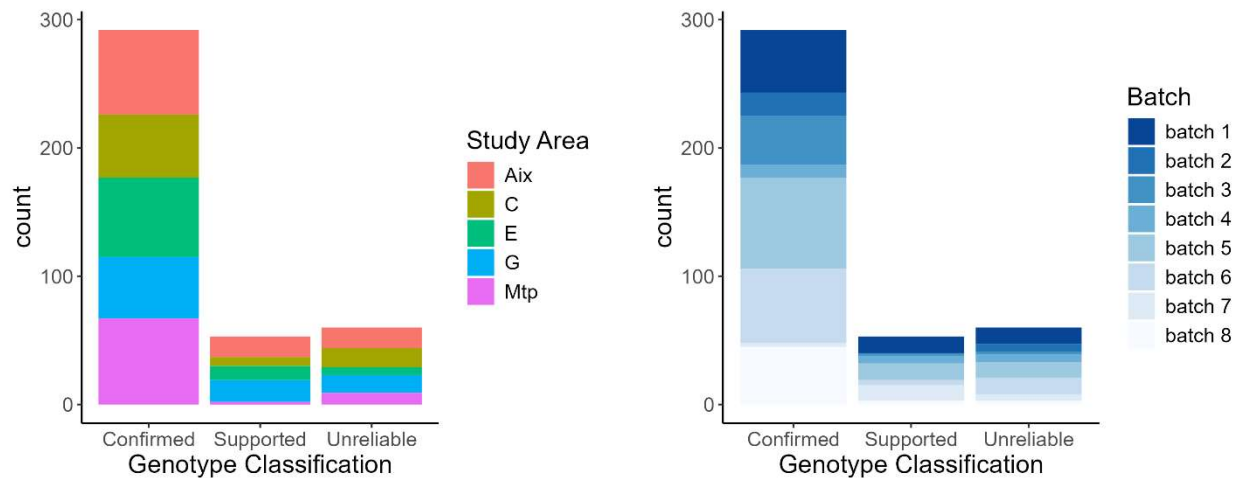


Fig. 1 Classification of genotypes from an original PCR run as confirmed, supported, or unreliable based on information from two replicate runs. Genotypes are split into colors based on study area (left) and PCR batch (right).

There was, however, a pattern where samples that had a higher proportion of missing genotypes tended to have a higher proportion of unreliable loci (Fig. 2-3). There was an especially notable drop-off in unreliable loci in sampled individuals that had missing data at only one locus, or that had no missing data (Fig. 2-3).

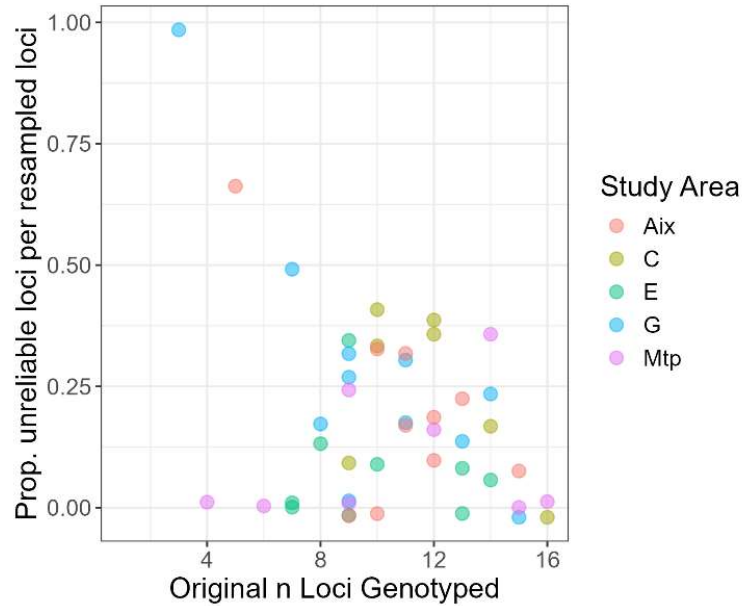


Fig. 2 Plot showing the proportion of loci classified as unreliable *per individual* after rePCR analysis. Calculated as the number of unreliable loci / 16 (the total number of loci genotyped). A small amount of vertical and horizontal jitter was added to points to improve visibility.

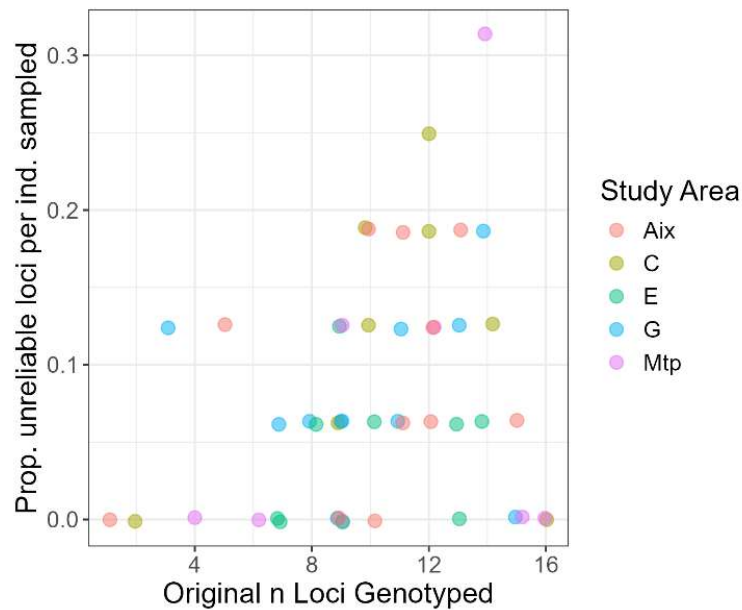


Fig. 3 Plot showing the proportion of loci classified as unreliable out of the total number of loci re-assessed (i.e. loci for which there was data in the original PCR run as well as in at least one re-PCR run) *per individual* after rePCR analysis. A small amount of vertical and horizontal jitter was added to points to improve visibility.

None of the loci were obvious outliers in terms of the proportion of individuals that had unreliable genotypes at that locus (Fig. 4).

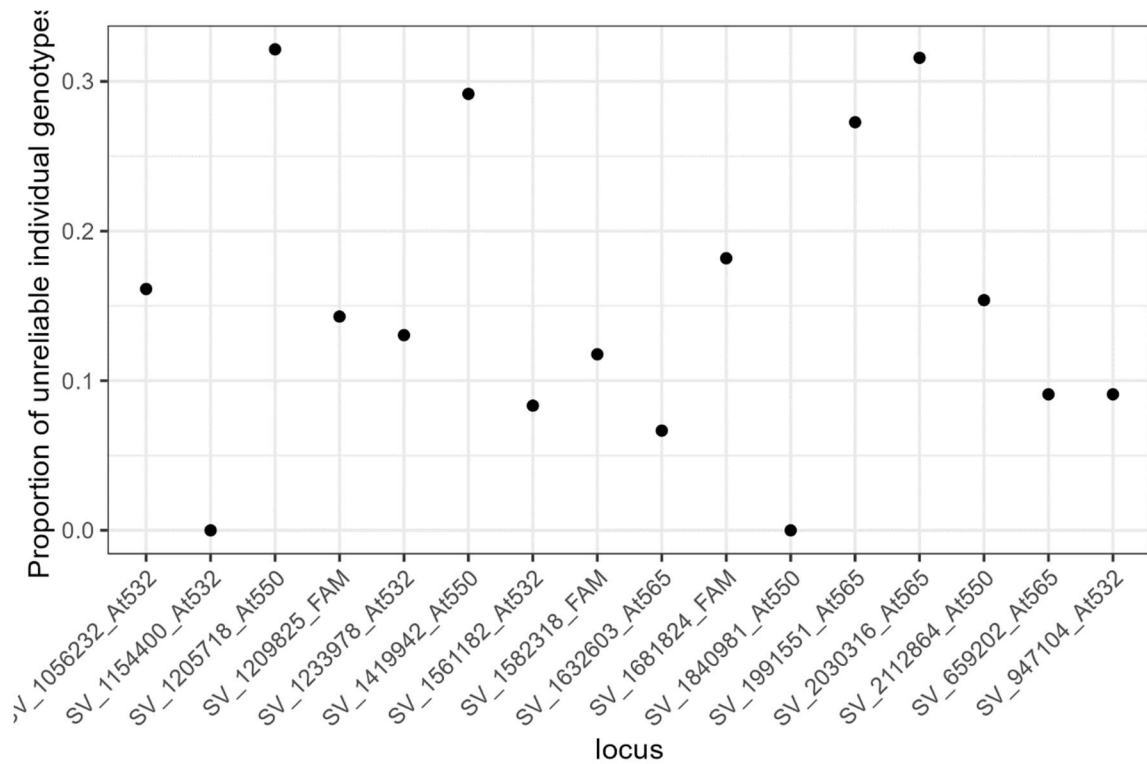


Fig. 4 Plot showing the proportion of samples classified as unreliable per locus out of the total number of samples re-assessed (i.e. samples for which there was genotype data in that locus for the original PCR run as well as in at least one re-PCR run).