Brief explanation of the PPFS2 algorithm

In a PPFS2 inpput file each genome is identified as having a positive pheonotpe, a negative phenotype, or an unknown phenotype. PPFS2 uses a bootstrap algorithm to predict the phenotypes of genomes whose phenotypes are unknown, and to estimate the accuracy of those predictions. PPFS2 performs a user-defined number, typically 100, of replicate runs. For each run half of the genomes whose phenotypes are known are gtemproarily marked as "unknown", then those SNPs that are non-randomly associated with phenotypes in those genomes whose phenotypes are known. For each SNP the probability, that the SNP alleles are randomly associated with the phenotype is determined; then SNPs are sorted from lowest to highest probability that the SNPs are randomly associated with the phenotype. Those SNPs with the lowest probability of being randomly associated with phenotype are the most informative (have the most predictive power). Starting with the most informative SNP the phenotypes of all genomes designated as "unknown phenotype" are predicted and the accuracy of the prediction is determined by comparing the predicted phenotype of the temporarily "unknown" genomes with the actual phenotype of those genomes to generate an accuracy score. The next most informative SNP is added, the phenotypes are predicted on the basis of the two SNPs and the accuracy score is determined. As more SNPs are added the accuracy increases, but the predictive power of each added SNP decreases so that eventually accuracy starts to decrease. At that point the process is terminated and the predicted phenotypes are recorded. That constitutes one run in which a set of predictive SNPs has been identified. After the user-defined number of independent runs a consensus set of "predictive" SNPs is identified and those are used to predict the phenotypes of all genomes. After the set of replicate runs, each genome of known phenotype ahs been used multiple times as a temporary "unknown", thus the ultimate predicted phenotypes of the unknown, and the accuracy of the predictions, is based on information from all of the genomes whose phenotypes are known.

A similar approach is used to identify causal SNPs, those SNPS whose allele state is likely to have caused a change in phenotype. The ancestral state for each internal (ancestral node) is estimated for the phenotype in question, and for the allele of each SNP. A SNP whose allele changed along the same branch as the phenotype changed may have caused that phenotypic change. For each SNP the probability that its state (allele) changed along randomly along branches where the phenotype changed is determined. The lower that probability, the more likely that the change in SNP allele caused the change in phenotype. SNP alleles whose changes were non-synonymous (resulted in an amino acid change) are the most likely candidates for "causal" SNPs. Silent changes are likely to be neutral, but can be non-randomly associated with phenotypic change by having been dragged along when phenotypic changes were selected. In Table S6 Probability is the probability that the change in SNP allele was actually random with respect to the change in phenotype. Probabilities $< 2.3 \times 10^{-308}$ are shown as 0.00 E+00.