HS Rats Round 10.2 Genotyping Summary

2023-12-19

This is a summary of genotyping results to accompany the 'Round 10.2' Heterogeneous Stock SNP dataset produced by the lab of Dr. Abraham Palmer at UC San Diego (palmerlab.org).

Data Summary

The dataset comprises 7,358,643 SNP variants for 17,812 individuals. Metadata for these individual samples are available in the accompanying genotyping log `round10_2_genotype_log.csv`. Metadata include sample IDs and their associated projects, sequencing libraries, sequencing flowcells, read mapping statistics, and quality control designations (pass/fail) for various QC steps.

Methods

SNP genotypes were produced using a custom bioinformatic pipeline parallelized to accommodate both double-digest genotyping-by-sequencing data (ddGBS) and low-coverage whole-genome sequencing data (lcWGS). ddGBS libraries were produced following Gileta et al 2020 and lcWGS were produced using the Twist Bioscience 96-Plex Library Prep Kit. The bioinformatic pipeline for the analysis of sequence data is freely available from GitHub at https://github.com/Palmer-Lab-UCSD/HS-Rats-Genotyping-Pipeline (manuscript in prep). Briefly, ddGBS sequencing libraries were demultiplexed using FastX-Toolkit and lcWGS libraries were demultiplexed using fgbio. ddGBS sequences were trimmed for quality using cutadapt and lcWGS sequences were trimmed using cutadapt and bbDuk. All libraries were aligned to the *Rattus norvegicus* mRatBN7.2 reference genome assembly using bwa. SNP genotypes were imputed using STITCH. Imputed genotypes were filtered to keep only those with imputation INFO scores > 0.9. Individual samples missing $> 10\%$ of high-INFO genotypes were removed from the dataset, as were samples with heterozygosity rates > 4 or 5 standard deviations from the mean heterozygosity (per library type), and those samples whose phenotypic sex (when available) did not match their genetic sex (determined by read count ratios on the X and Y chromosomes). This dataset contains all samples that passed all filtering and quality control steps.

Figures

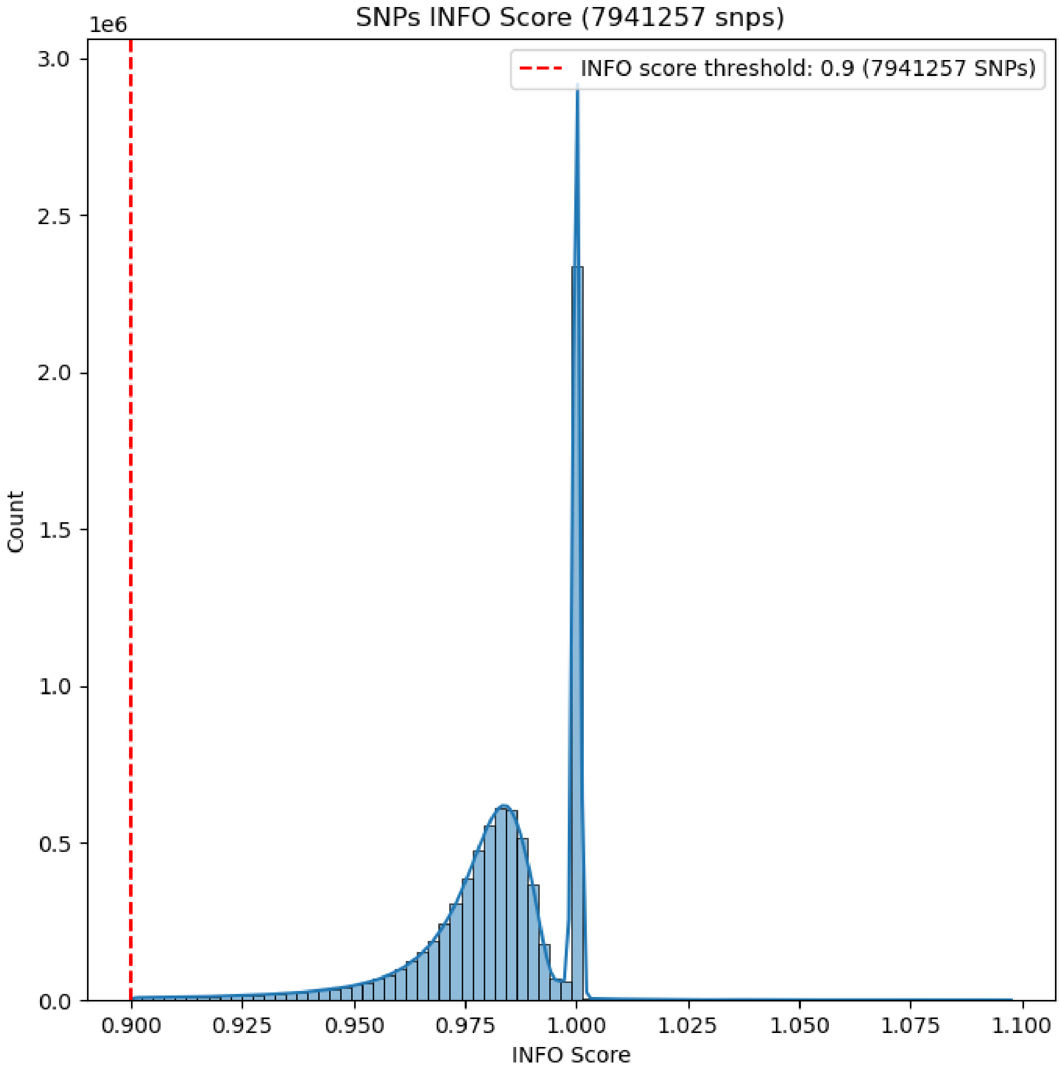

Figure 1: **Fig. 1: INFO score distribution.** All imputed SNP loci are high quality, with INFO scores (a measure of imputation quality) > 0.9.

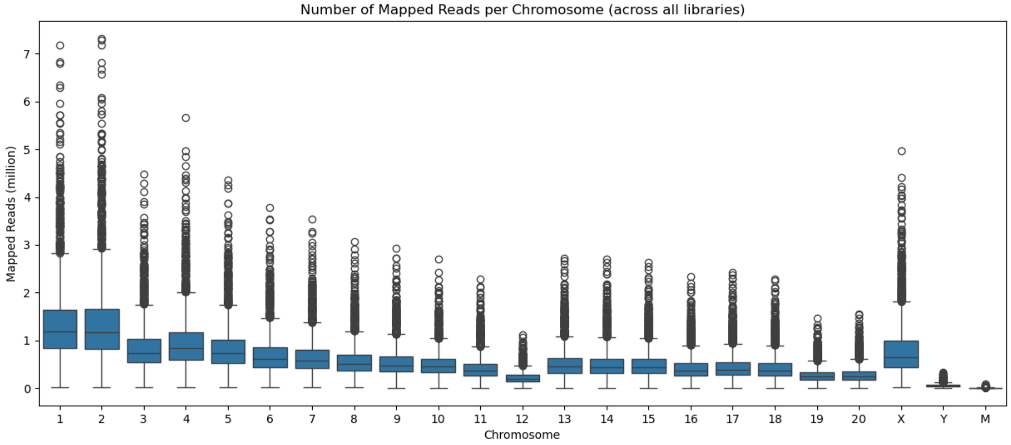

Figure 2: **Fig. 2: Read mapping per chromosome.** All chromosomes are well-represented with millions of reads mapped to each (across all libraries)

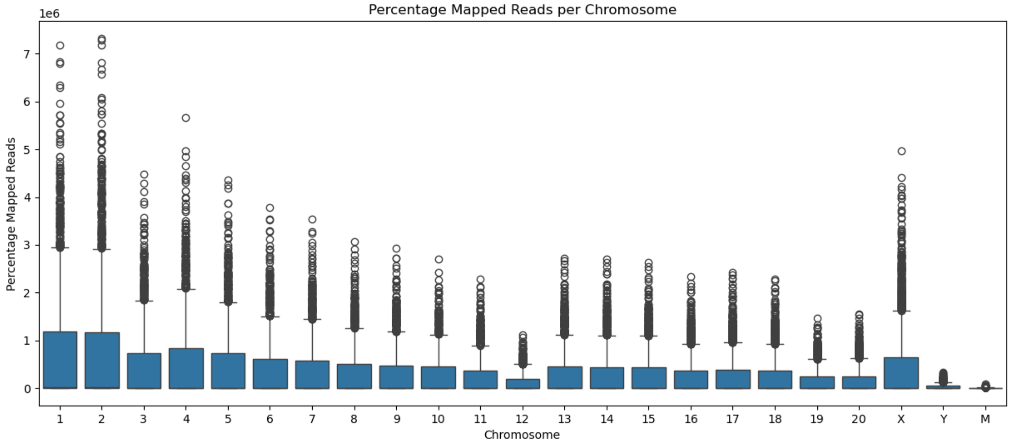

Figure 3: **Fig. 3: Percentage of reads mapped per chromosome.** Read depth is well distributed across the genome (across all libraries)

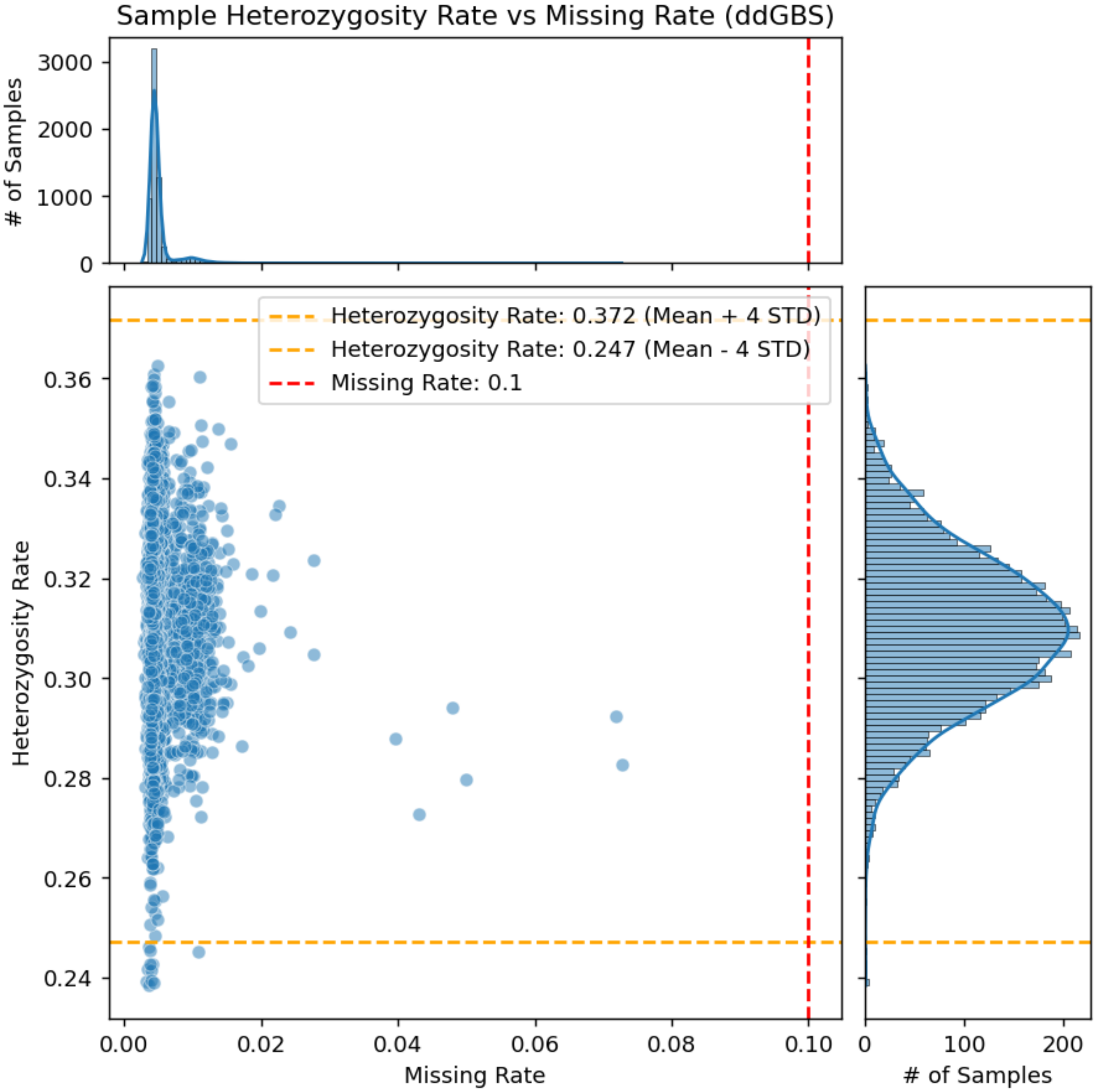

Figure 4: **Fig. 4: ddGBS quality control filters.** Round 10.2 ddGBS samples are only those that fall within filtering cutoffs for heterozygosity (orange lines) and missingness (red line)

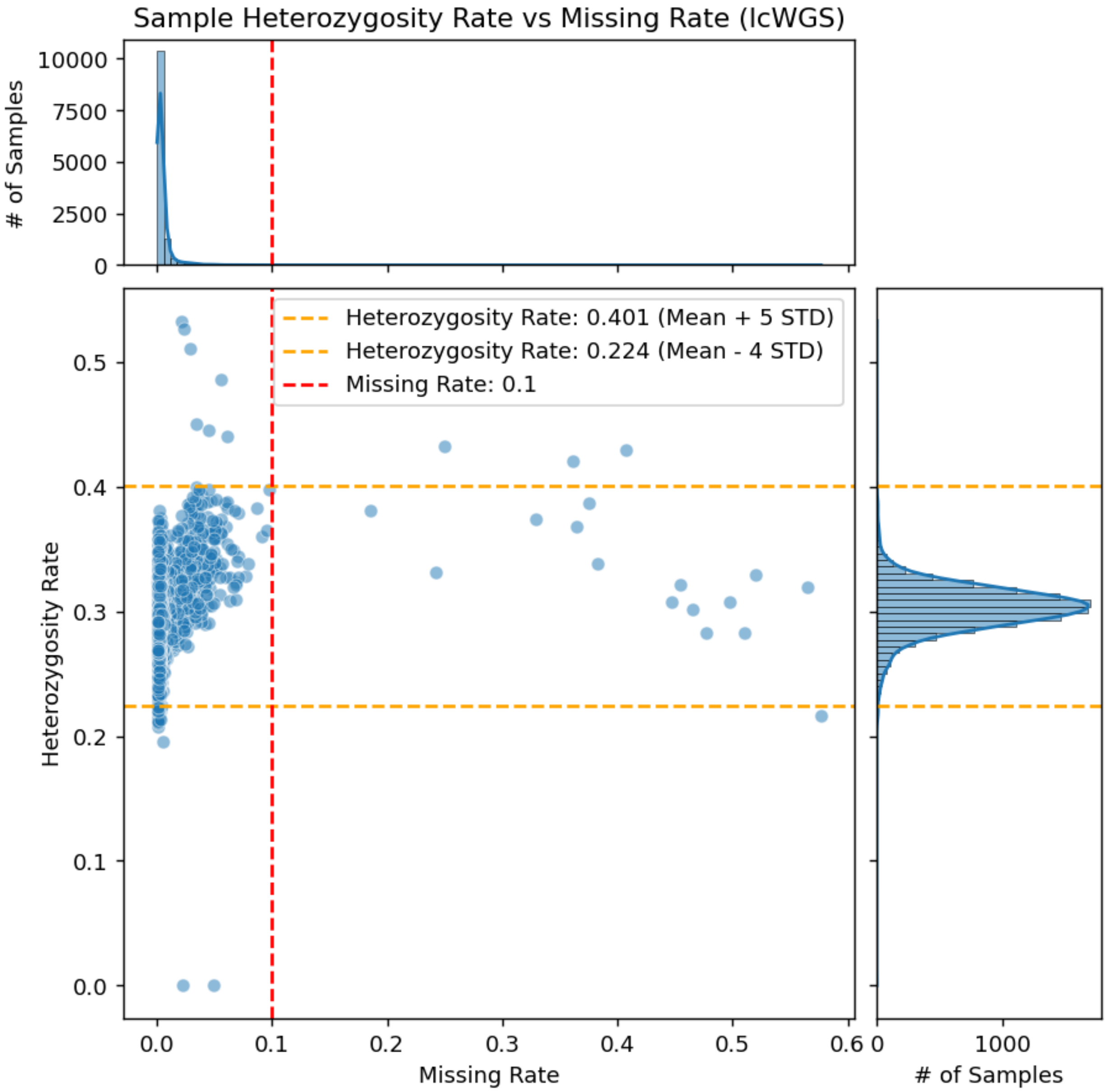

Figure 5: **Fig. 5: lcWGS quality control filters.** Round 10.2 lcWGS samples are only those that fall within filtering cutoffs for heterozygosity (orange lines) and missingness (red line)

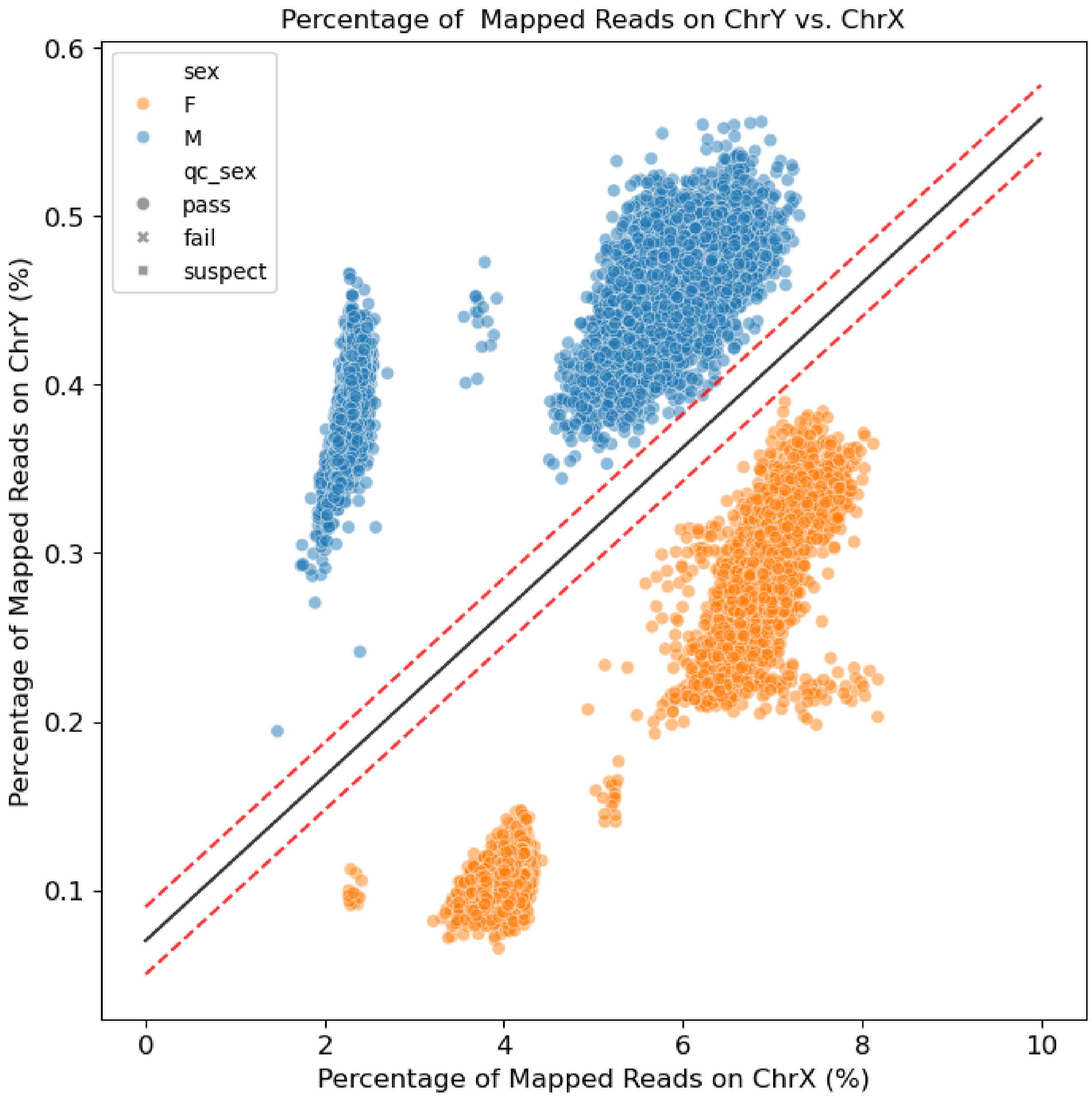

Figure 6: **Fig. 6: Sex quality control filters.** All Round 10.2 samples' genetic sex is distinguishable and consistent with phenotypic sex

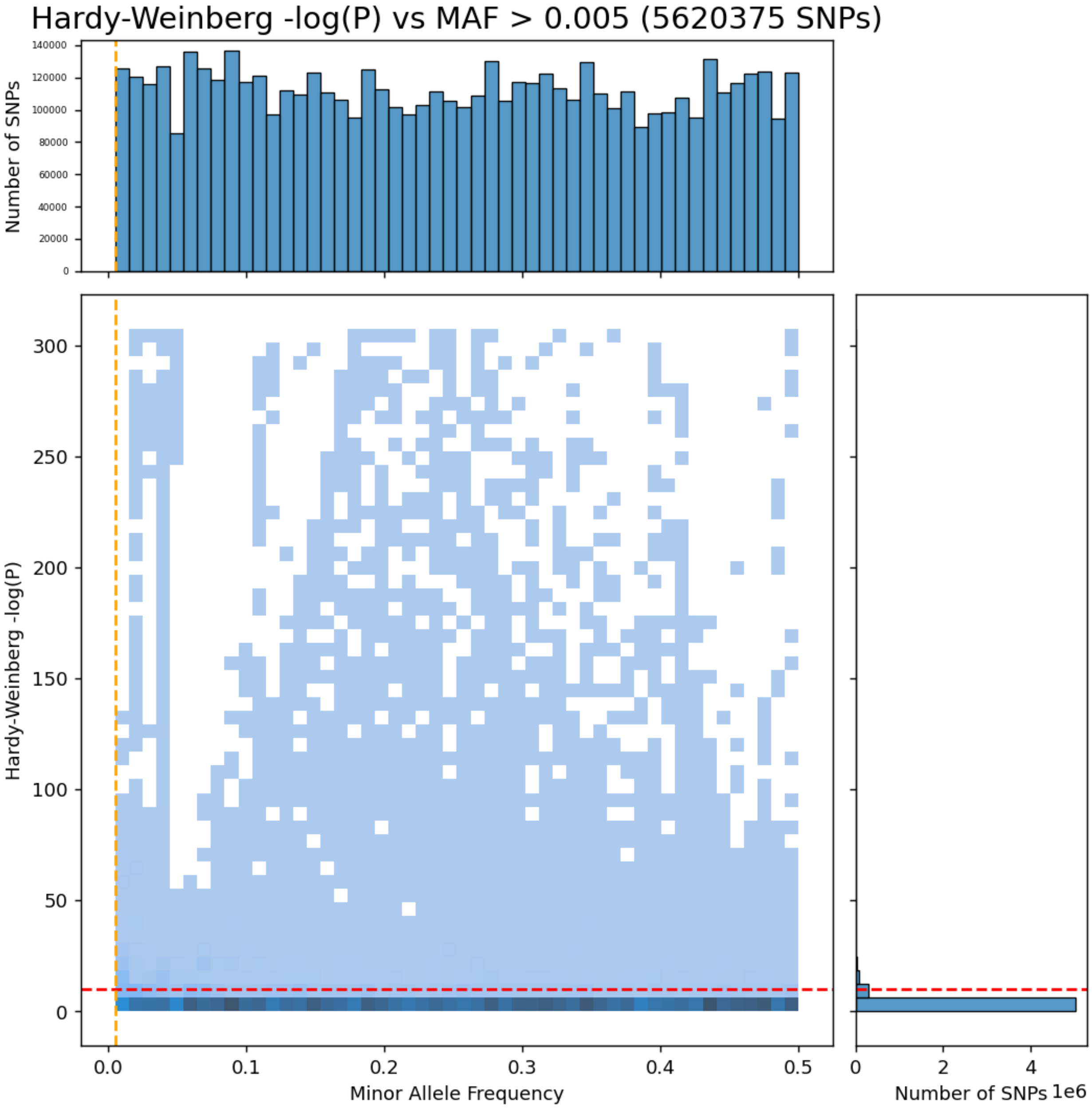

Figure 7: **Fig. 7: GWAS filters.** For conducting GWAS, we recommend conducting further filters on genotypes for minor allele frequency (MAF) and Hardy-Weinberg disequilibrium (following a Hardy-Weinberg equilibrium exact test). Suggested cutoffs of MAF > 0.005 (orange line) and HWE exact test -log(P-value) < 10 (red line) would produce a dataset of 5,620,375 SNPs for association testing. Note: these cutoffs have not already been applied. The set of SNPs displayed here is that which would result by applying the suggested filters on the provided Round 10.2 dataset